

Determine Trending Research Topics Through Grant Abstracts
Using Grant Data from Dimensions

Sarah Siddiqui and Thomas Durkin
CSC - 440
Data Mining
Prof. Jiebo Luo

ssiddiqui@library.rochester.edu, mdurkin7@ur.rochester.edu

ABSTRACT

This study investigates data extracted from Dimensions in order to produce topic models from grant abstracts. Models were built for the University of Rochester as well as R1 Research Universities. Our approach includes creating models using both Latent Dirichlet Allocation and BERTopic to analyze research topic trends. It was determined that the foremost topics for the University of Rochester consist of Clinical Sciences, Physical Sciences, Cell Biology, and Brain & Cognitive Sciences. R1 institutions had 15 topics which mapped to several categories relating to AI, Systems, Theory and Interdisciplinary Areas in CS Rankings, with an additional topic focused on conferences, education and workshop.

1 INTRODUCTION

1.1 Background and Motivation

For our final project we performed topic modeling with LDA and BERTopic on two datasets from the database Dimensions (<https://app.dimensions.ai/discover/grant>). Dimensions was launched in 2018 by Digital Science and is well-known for its collection of grants, most of which is directly provided by the funders or aggregated from public sources [1].

The first dataset contained all the grants received by the University of Rochester and its affiliates (including River Campus, Medical Center, Laboratory for Laser Energetics, etc.) for the period of 2000-2020. The second dataset included grants in the area of Computer Science received by 131 R1 or “Very High Research” institutions across the United States in the same timeframe. Both of these would help us understand the research landscape in the context of funding; awarding the grants would imply that funding agencies are interested in these topics. Most bibliometric studies are done at the article level and focus on citations and related impact factors [2] and we hope to gain more insight into the funding that in many cases leads to the articles, since studies have shown that “publications from grant sponsored research exhibit higher impacts in terms of both journal ranking and citation counts than research that is not grant sponsored” [3].

1.2 Problem Statement

This study had two facets and the question we pursued can be described as:

“Considering the entire University of Rochester as well as Computer Science in the United States, how has associated research evolved during the past 20 years?”

1.3 Data Description

Grant and article records in Dimensions are categorized into various Fields of Research (FoR) based on the Australian and New Zealand Standard Research Classification (ANZSRC) [4][5]. Some of these fields include Mathematical Sciences, Physical Sciences, Engineering, and Technology. A more detailed illustration of the FoR can be seen in Appendix A. The models for the University of Rochester will be built using data from all research categories.

Data for the Research I Universities will only include subcategories of Information and Computing Sciences such as Computer Software and Distributed Computer. See Appendix B for a full breakdown of the subcategories.

We limited the data attributes for our analysis to include only **Abstract** and **Start Year**. A more detailed explanation is covered in section 2.1 Data Collection and Preprocessing.

1.4 Related Work

There are some instances of similar research of analyzing grants available in Dimensions [6], but these documents tend to either have a broad focus or are used for internal purposes or specific institutes [7] [8]. With this project, one of our goals was also institutional analysis, specifically in the context of the University of Rochester. With the R1 data analysis, we are digging deeper into Computer Science topics, since these are relevant to our class/field.

Other studies looking at grants typically focus on funding acknowledgment statements in the database Web of Science [9]. Since these are based on published articles supported by funded research, they can be more influenced by publications and citations, and the analyses are performed using techniques such as logistic regression.

While a lot of the topic modeling techniques use probabilistic techniques such as LDA (section 2.2), recently neural techniques such as BERT have been used for topic modeling of scientific articles [10]. Therefore, we decided to use both methods to compare our topics. Our implementation of BERT is described in section 2.5 of Methodology.

2 METHODOLOGY

2.1 Data Collection and Preprocessing

The data for this project was extracted from the Dimensions database and analyzed using Python. We started with the University of Rochester, looking at grants active in the period 2000-2020, which contains over 3800 records. The database has a limit of downloading a maximum of 5000 records so this could be extracted in one query.

For grants awarded to Computer Science, the database showed over 100,000 results within the United States - with academic, corporate, and federal organizations - for the 20 year period under the category *08 Information and Computing Sciences*. As the data had to be downloaded in batches, we decided to focus on grants awarded to Research I (R1) Doctoral Universities, i.e. those classified as having “Very High Research” by the Carnegie Classification of Institutions of Higher Education. These 131 institutions, listed in [12] received over 51000 grants in the period, amounting to over \$35.6 billion out of the total \$71.5 billion, and will provide good representation of the funding topics.

To build the topic models, only the Abstract, a short summary of the research project, column was used. If the grant did not have an abstract then it was removed from the overall data selection. The Start Date data was included to construct the Topics Over Time graphs using BERT [10], shown in the Results section.

Preprocessing the data drastically helped us in creating unique topics. All punctuation was removed from the Abstract. A list of English stopwords was pulled from Natural Language Toolkit [13]. Additional words listed in Figure 1 were added based on the initial models.

```
'study', 'project', 'provided', 'utilized', 'also', 'whether',
'much', 'based', 'used', 'using', 'many', 'abstract', 'program',
'university', 'rochester', 'research', 'develop', 'evaluate', 'caused',
'allows', 'requires', 'includes', 'recent'
```

Fig. 1 Stop words added to the NLTK list

These words were removed from our abstract data as well as words that were less than a length of three characters. To improve interpretability of the resulting topics, bigrams and trigrams were constructed using the Gensim Phrases library. The rules for these word pairs include a scoring “threshold” for building the phrases and a “minimum count” for the number of times the

combination of words should occur together to be considered in the n-grams. We tried different values for these parameters so the model is able to identify the phrases. Lastly, lemmatization was applied to our vocabulary so that all common words were reverted to their root word. Only words whose part of speech was a noun, adjective, verb, or adverb were kept.

Once the data was lemmatized, we created our dictionary and corpus. We further cleaned the corpus by removing terms with a high TF-IDF. Term frequency-inverse document frequency tells us how relevant a document is for a certain term relative to the corpus [14]. Thus, terms that were too frequent across all documents would not be useful for distinguishing between topics. We used the TfidfVectorizer from the Scikit learn library to have our dictionary and corpus ready for LDA.

2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation or LDA is a “flexible generative probabilistic model for collections of discrete data” [15]. LDA classifies documents into clusters known as topics, which contains a set of words that contribute to that cluster. Each word is paired with a weight to indicate how much of a presence that word has within the topic. Since each topic is composed of a set of words, the user has to manually assign an overall label. Two essential parts to this algorithm are the dictionary and corpus. The dictionary is our preprocessed vocabulary used for assigning specific words to a unique id, while the corpus is our set of documents represented by the word id and its document frequency.

2.3 Procedure

The LDA models were built using functionality from the Gensim library [16]. The parameters for the function include the dictionary, corpus, the number of topics and hyperparameters alpha and eta. Alpha and eta are based on the Dirichlet prior and give us the per-document topic and per-topic word distributions, respectively. In Gensim, these values are set to symmetric by default [16], but we set them to auto so the model would learn the best values for both. This was also seen in the Gensim documentation, although the models take slightly longer to run.

To further determine the best number of topics for each model, a comparison of coherence scores between models was made by plotting the scores against the number of topics.

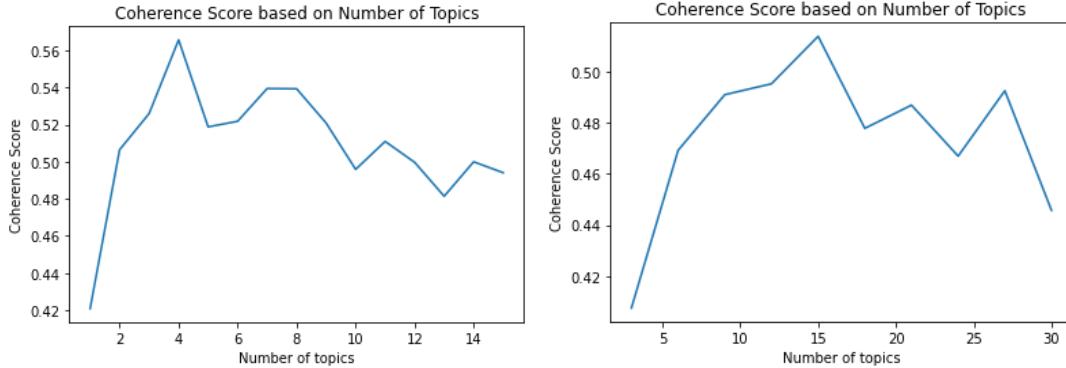


Fig. 2 Coherence Scores from UR (left) and R1 (right) models

As demonstrated in Figure 2, a model with four and fifteen topics proved to be the best for the UR and R1 data respectively since they had the highest coherence scores. This measure is important with LDA topic modeling because it is the degree of semantic similarity between high scoring words in the topic [17]. For R1, we wanted to create a model that has the least amount of overlapping between topics and a model with fifteen topics provided us with that outcome.

2.5 BERTopic

BERT (Bidirectional Encoder Representations from Transformers) is a language representation model introduced by Google. For our model, we applied the BERTopic technique by Maarten Grootendorst [18]. BERTopic uses BERT embeddings for documents which is followed by dimensionality reduction with UMAP. Next, clusters of documents with similar topics are created with the density based HDBSCAN technique.

This technique is useful because there was less data preprocessing to be done. We were able to modify some parameters and added the stop words used in LDA to avoid some frequent words from appearing in the topic. As we will see in the following sections the results were very meaningful. The only downside of BERTopic is that it takes significant time to run: for the R1 data the runtime was over 3.5 hours for building the model, double the time it took for LDA.

3 RESULTS AND ANALYSIS

3.1 University of Rochester

For the University of Rochester, we performed LDA and BERTopic for the full period of 2000-2020, which is discussed in section 3.1.1. To get a better look at the University of Rochester, section 3.1.2 covers LDA performed in 5 year periods for UofR.

3.1.1 University of Rochester 2000-2020

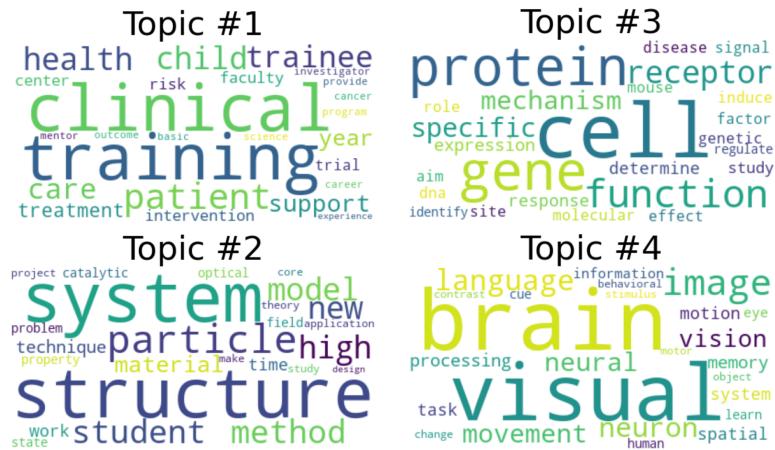


Fig. 3 Word clouds for UR LDA topics 2000-2020

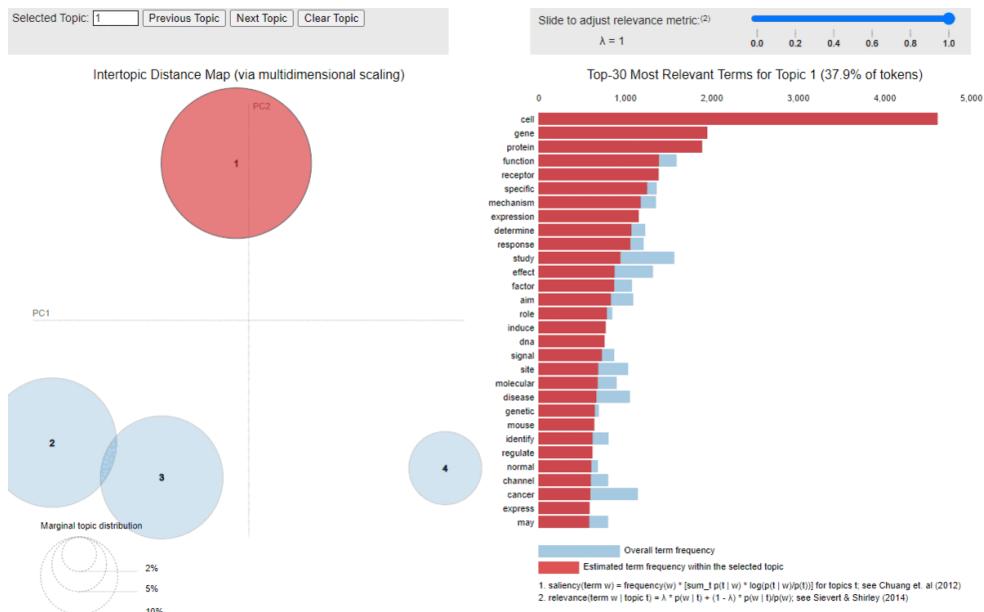


Fig 4. Representing the topic clusters for UR with pyLDAvis

One way of evaluating our model was to see the categories that show up in the database Web of Science for University of Rochester publications that are associated with a funding agency, implying that the authors received a grant. It is worth noting that Web of Science returned 30163 publications [19] compared to the 3825 grants in Dimensions. Since one grant can result in multiple publications we recognize that these are not exactly the same things, but can give a rough idea of the University's focus of research. Figure 4 is the visualization available in Web of Science for the period between 2000-2020 for the top 5 topics. We see that our LDA topics are fairly similar. The only category that was not present in our word clouds (Figure 4) is Astronomy; instead we have Healthcare represented by Topic 1.

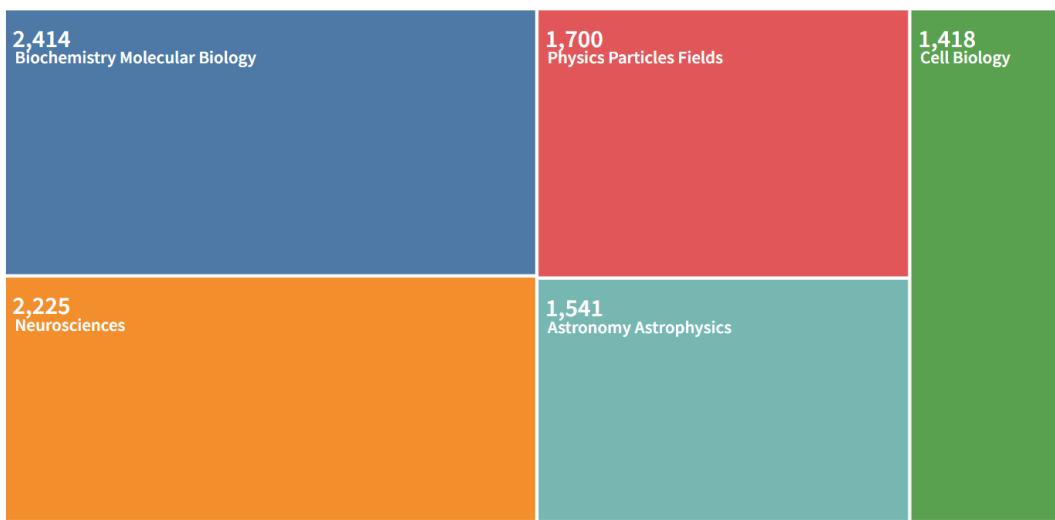


Fig. 5 Publications at the University of Rochester 2000-2020 from Web of Science

3.1.2 Evaluating 5 Year LDA Models

Looking at the word clouds for University of Rochester 2020 (Appendix C), Topic 3 stands out from all the other topic breakdowns because words such as *laser* and *field* have become more prominent. We can see through Appendix C (2005-2009), *field* is first present in Topic 3 with a decent contribution to the overall topic. This word plays a role throughout the following yearly models with *lasers* being introduced 2015-2020 (Appendix C). The gradual increase of *laser* and *field* contributing to overall topics demonstrates that the University of Rochester's Laboratory for Laser Energetics is receiving more funding within recent years allowing the lab to conduct more research projects.

Comparing the rest of the yearly models we can see similar results. Although, *patient care* has become prominent within 2020, which can be as a result of the hospitals wanting to focus more on overall patient health or an overload of patients within hospitals due to COVID-19.

Evaluation period	Number of topics	Topic Names (assigned by us)	Coherence Score
2000-2020	4	Clinical Sciences & patient care; Particle & applied Physics; Genomics; Brain & Imaging Sciences	0.565
2020	4	Patient care; Stem learning; Lasers & Applied Physics; Genomics	0.44
2015-2020	4	Clinical Sciences; Genomics; Stem education; Lasers & materials	0.56
2010-2014	3	Clinical Sciences; Applied Physics; Genomics	0.52
2005-2009	4	Infections, genes & viruses; Applied Sciences; Clinical Sciences; Genomics & Cancer	0.49
2000-2004	3	Cell & genomes; Clinical Sciences; Applied Sciences	0.55
2000	3	Genomics; Clinical Sciences; Applied Sciences (optical, image, brain)	0.51

Table 1: Model evaluation (comparing coherence scores)

3.1.3 University of Rochester BERTopic Model

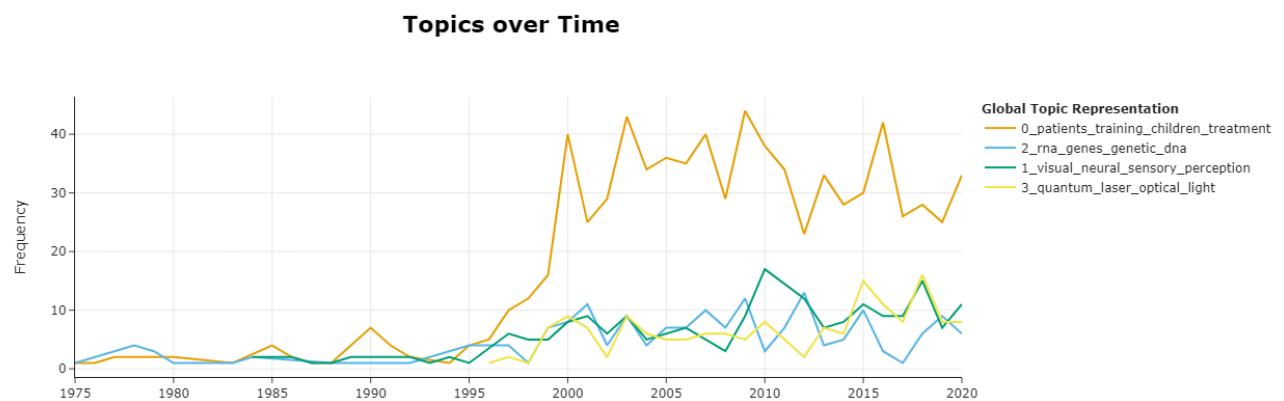


Fig 6. University of Rochester Topics over Time

The BERT model for the 2000-2020 period gave 33 topics in all. Most of these were from the medical fields followed by the physical sciences. There was also one topic (number 7) about hardware, memory, systems and computing. When we look at the top 4 topics, the results are similar to our LDA models. We can see that there is a significant increase in patient treatment over the other topics. The other three topics follow a similar increase over time.

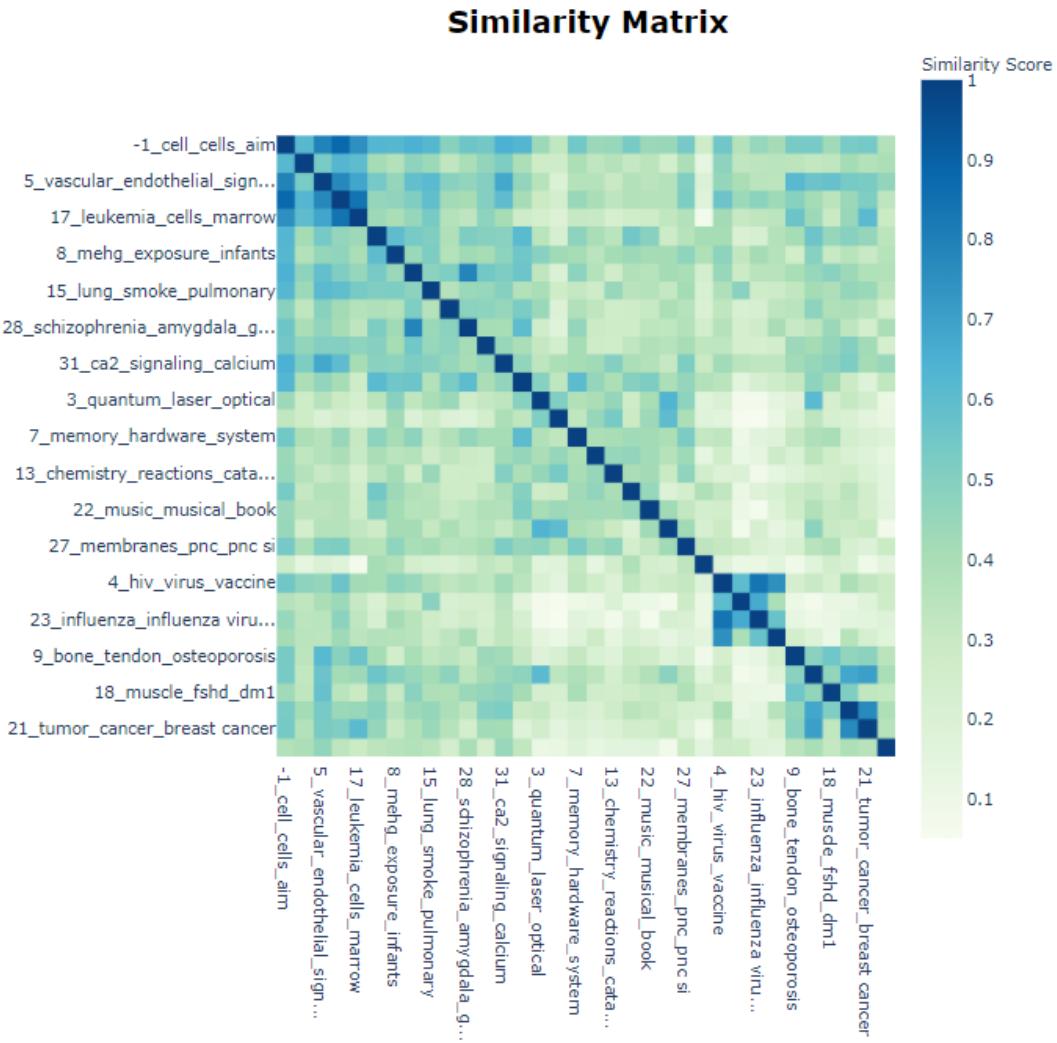


Fig 7. Similarity Matrix for all 33 topics generated by BERTopic.

The above figure represents the similarity matrix for the topics generated by BERTopic. Most of the topics are fairly dissimilar, although there are 3-4 topics with a relatively high similarity. These refer to areas relating to HIV, influenza, viruses, etc. and the algorithm may have separated them because of slightly different keywords. On the whole, this appears to represent a good chunk of research at the university, at least in the STEM areas.

3.2 Computer Science at R1 Institutions

For the data about the R1 institutions, we focused on topics for the entire 20 year period and performed LDA and BERTopic on them. LDA gave us 15 topics as the best based on coherence scores (figure 8), while BERT identified 35 topics. For comparison, figures 9 and 10 on the next page show word clouds for both models.

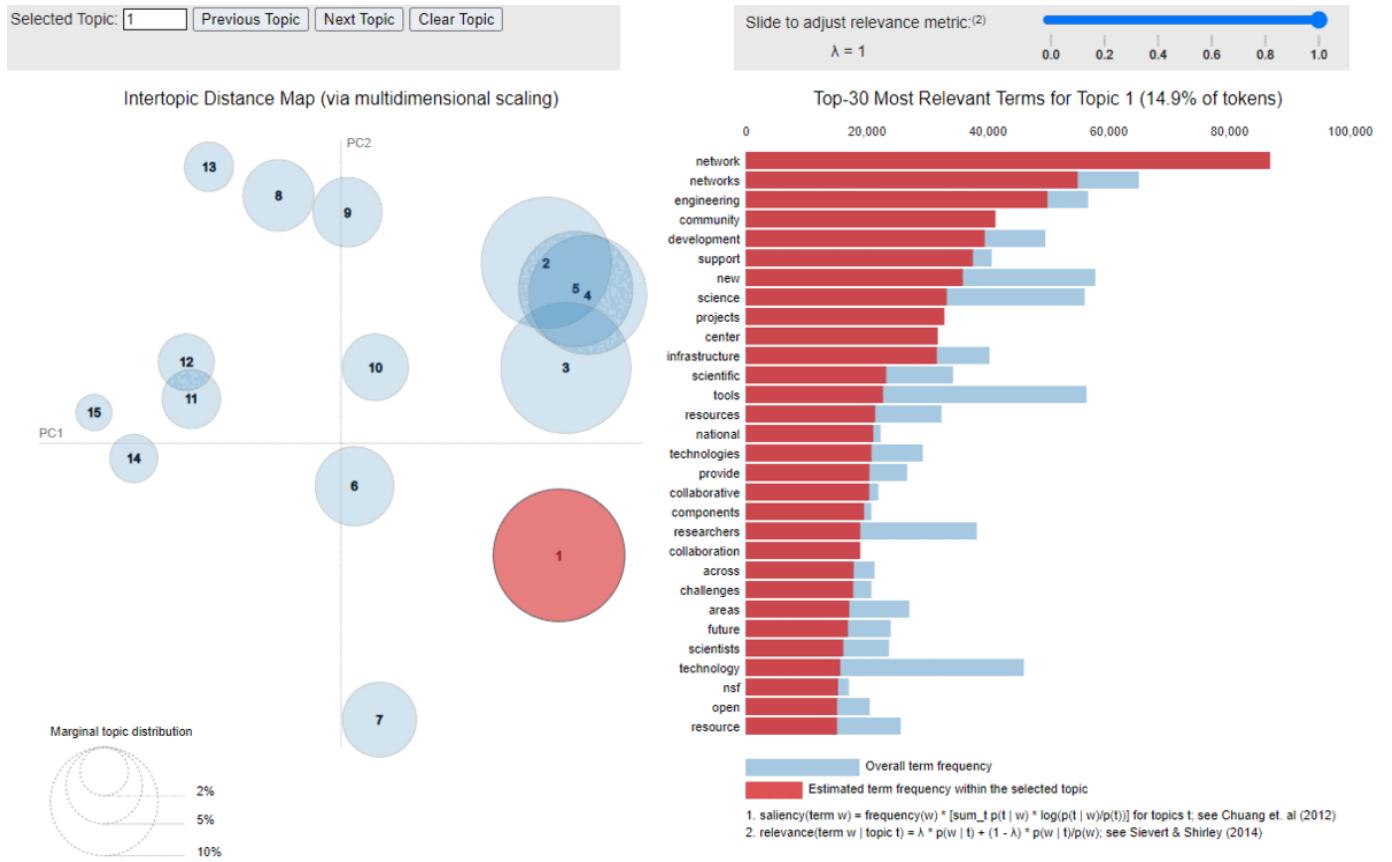


Fig 8. Visualizing the Computer Science LDA topics at R1 institutions

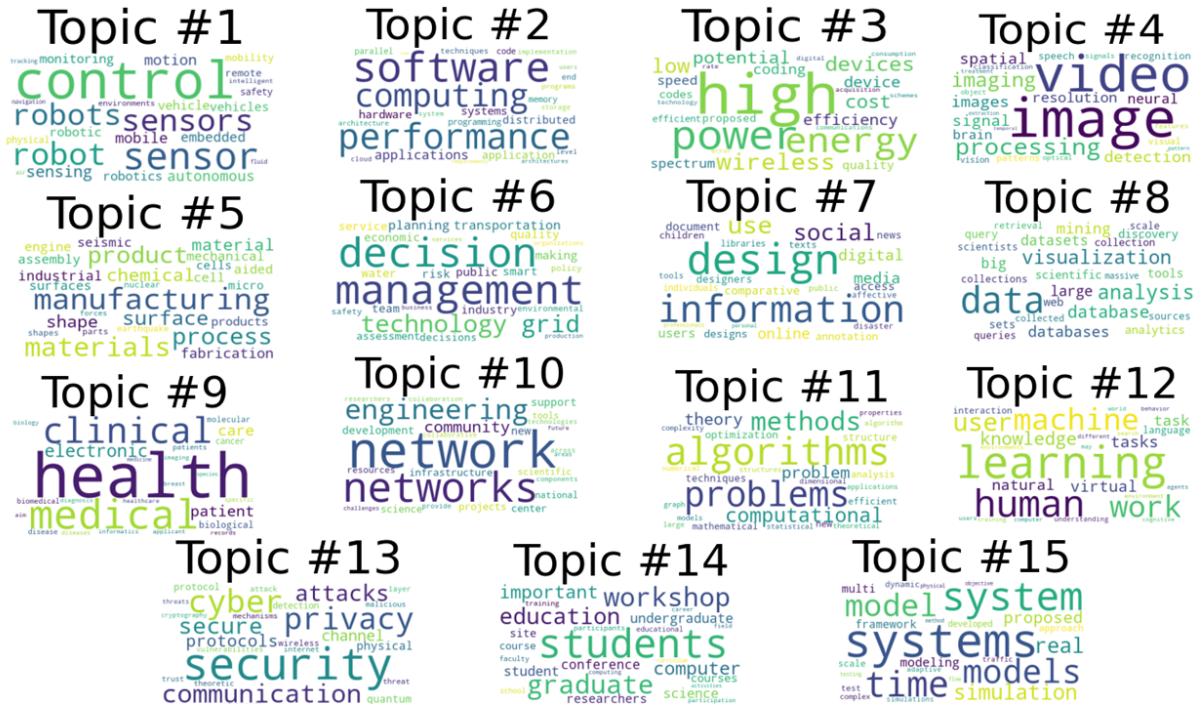


Fig. 9 Word cloud for R1 data with LDA

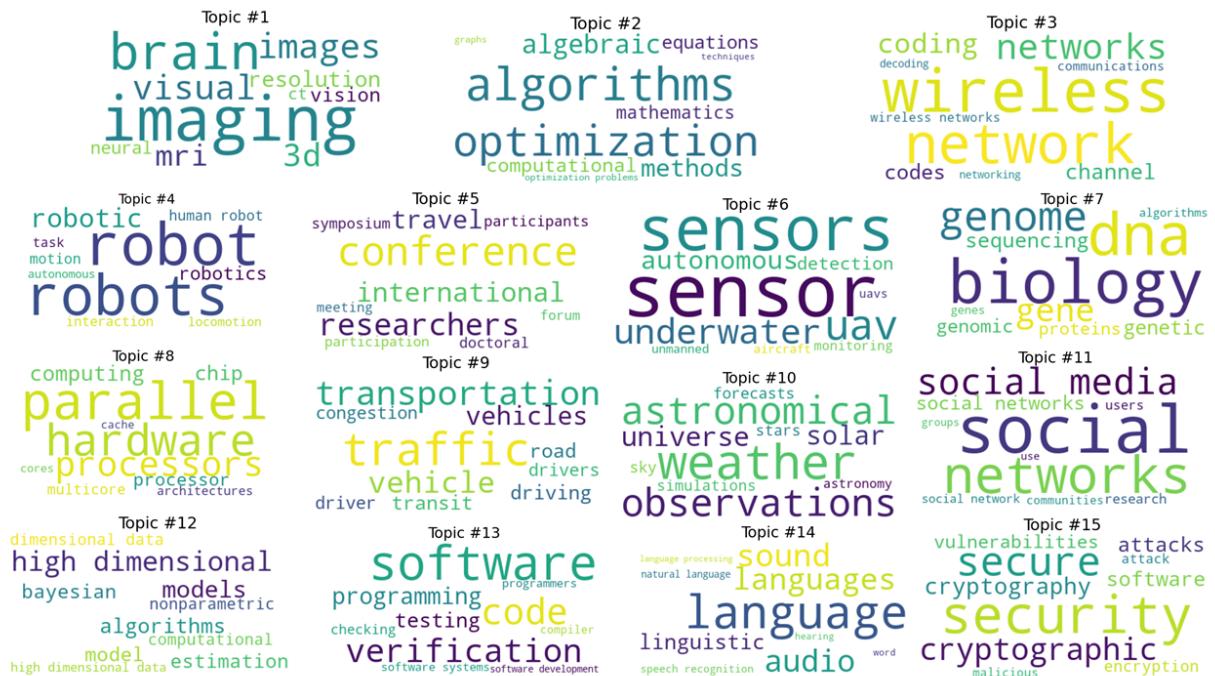


Fig. 10 Word cloud of top 15 topics for R1 data with BERTopic

To get a better understanding of how well our models identified the topics, we compared them with the categories listed in CS rankings [20]. The below table is tagged based on our best efforts to map the topics within our models. The numbers next to the round tags - red for LDA and blue for BERTopic - indicate the topic number in the models. Note: LDA numbers correspond to the pyLDAvis bubbles shown in figure 8.

AI 24	Systems	Theory	Interdisciplinary Areas
Artificial intelligence 6 25	Computer architecture 5 8 30	Algorithms & complexity 4 2 26 12	Comp. bio & bioinformatics 8 12 7 16
Computer vision 4 1	Computer networks 3 21	Cryptography 13 15	Computer graphics 29
Machine learning & data mining 14 11 32	Computer security 13 15 35	Logic & verification	Economics & computation 6 18
Natural language processing 14	Databases		Human-computer interaction 3 29
The Web & information retrieval 10 28	Design automation 15 9		Robotics 11 4
	Embedded & real-time systems 19		Visualization
	High-performance computing 20		
	Mobile computing 9		
	Measurement & perf. analysis 2 22		
	Operating systems		
	Programming languages		
	Software Engineering 13		

Key: ● In LDA ● In BERTopic (number indicates topic number in models)

Table 2: Mapping CSRankings categories with modes.

Since our topics were slightly different, some of them were mapped to multiple categories. For instance, topic 13 in LDA (with terms: cyber, security, protocols, privacy, cryptography, threats) and topic 15 in BERTopic (terms: security, secure, cryptography), seem to perfectly match with “Computer security” under Systems and “Cryptography” under Theory. Another example is “Algorithms & Complexity” in the table which is tagged with three BERT topics: #2: algorithms_optimization_algebraic_methods; #12: high_dimensional_models_algorithms_model; #26: quantum_computation_algorithms_theory. These represent different facets of complex algorithms so it seems reasonable that the model would differentiate them. Further, topic #24 from BERTopic, relating to Big Data can potentially match with most of the AI and Interdisciplinary areas. For our table, we tagged the entire AI column with that topic.

On the other hand, certain topics generated from our models did not fall under any of the categories. These are listed below with some of their relevant terms:

- LDA topic 1*: network, community, projects, collaborative
- LDA topic 7*: students, graduate, workshops, education
- BERTopic 5*: Conference, researchers, international, travel, participants
- BERTTopic10: weather, astronomical, solar, universe
- BERTTopic 17*: learning, students, teachers
- BERTTopic 23: disaster, emergency, hurricane
- BERTTopic 27*: university, industry, cooperative
- BERTTopic 31: geoscience, community, cyberinfrastructure
- BERTTopic 33*: reu, students, undergraduates
- BERTTopic 34*: corps-project, impact-commercial, broader-impact

We observe that some of these are relating to applications of Computer Science to different disciplines such as Astrophysics and Geosciences. These could be additional topics under Interdisciplinary, maybe something we even see on the website in the future.

However, the *most interesting* topics not covered in CS Rankings which both our models gave were relating to a broader theme of “sharing research”, indicated with an asterisk. With terms relating to students, conferences, workshops and collaboration, we did not expect to see it in grant abstracts, but the terms showed up despite tweaking with the stop words and other model parameters. Some of these topics are also rather similar, as we see in the similarity matrix in figure 11. Overall though, they highlight the importance of sharing the work done with the community in the field. This does corroborate with a recent study about big data needs by the organization ITHAKA S+R [21], wherein “Sharing Knowledge” and “Communicating Research Outputs” was a broad theme among various disciplines involved in big data research across several universities in the US.

We also looked at the change in research topics over the 20 year period, shown in figure 12. While most topics are fairly close to each other, the topic “imaging, brain, neural, 3d, visual” (topic 1 in table _ and topic 0 in graph) has consistently shown a higher frequency relative to other topics. This is followed by the topic relating to “algorithms, algebra, optimization, math”, however since 2010 the topic relating to robotics has overtaken as the second most frequent topic while the former has slightly dropped. The topic relating to “sensors, unmanned, aerial, uav”, which we categorized under Artificial Intelligence has also shown a steady rise overall, although there was a slight drop corresponding to 2015 followed by another increase. It is worth pointing out though that the “conferences, workshops”, etc. topic was one of the few that dropped in 2020, perhaps because in-person conferences couldn’t happen due to COVID (although there were online shifts). This might be worth investigating in future.

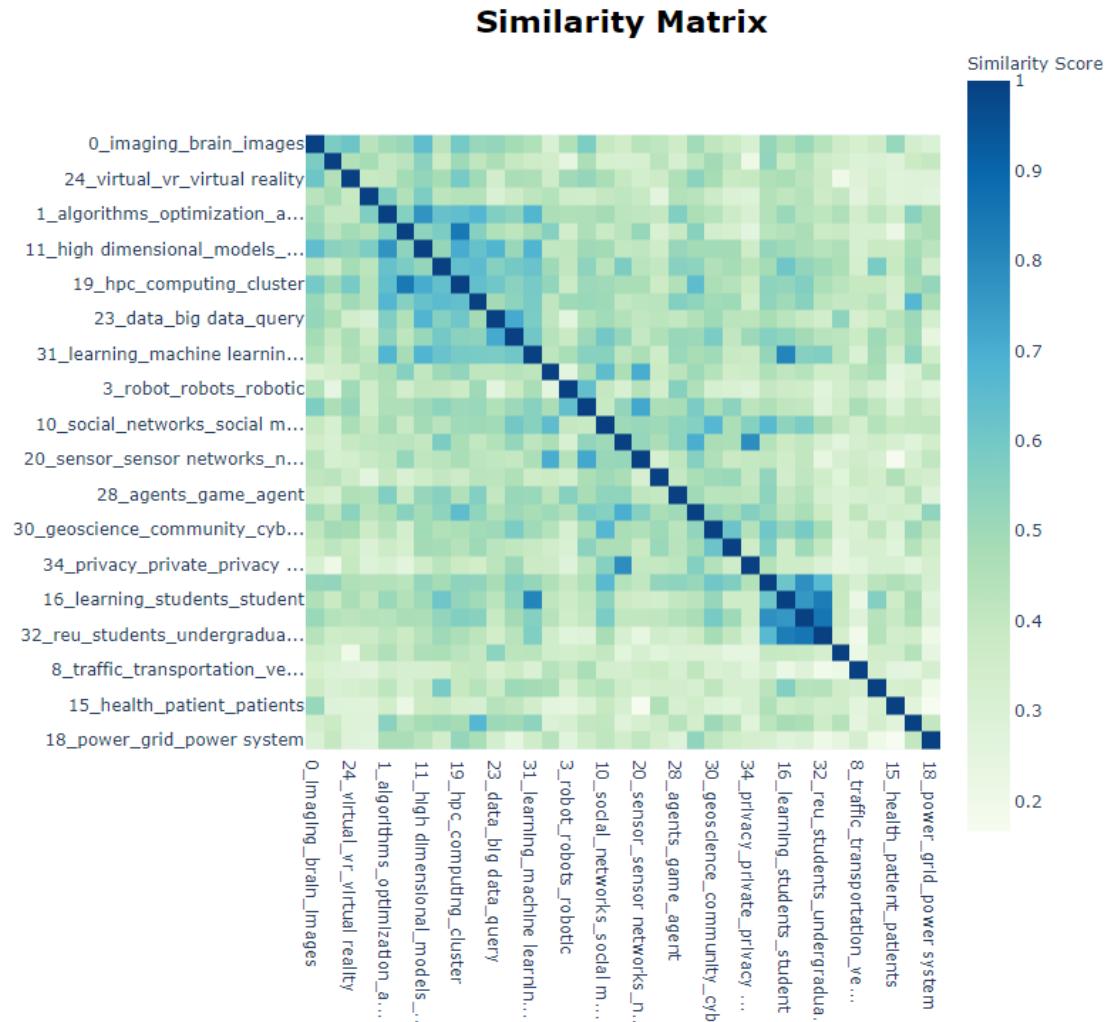


Fig. 11 Similarity Matrix for all 35 topics with BERTopic

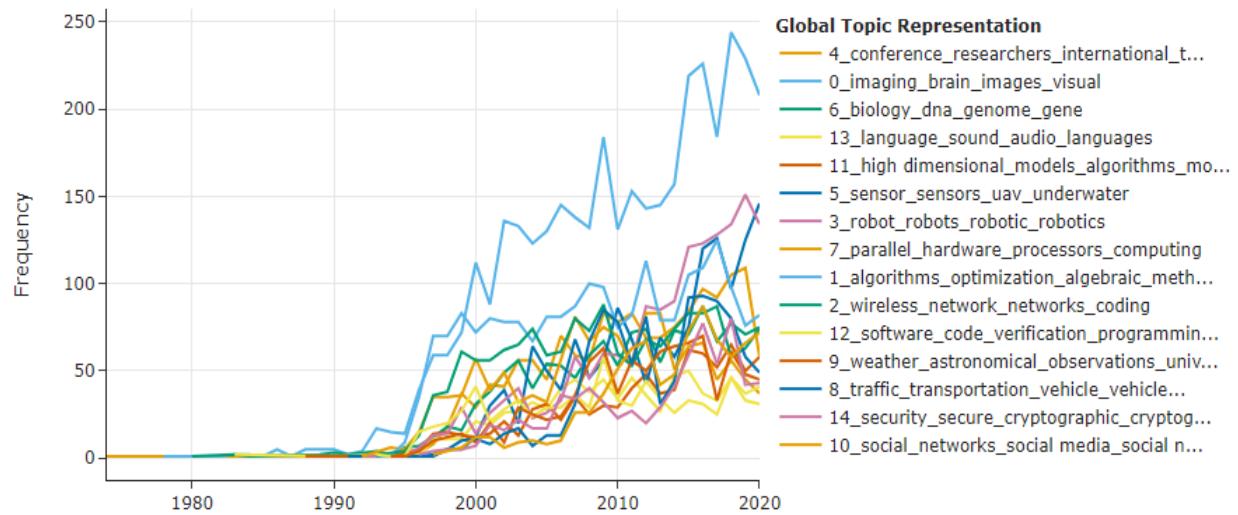


Fig. 12 Topics over time using BERTopic, focusing on top 15

4. FUTURE WORK

This analysis gave us very interesting insights into the research areas being funded over the years. We believe that increasing the scope of our data could further help enhance our understanding of the research landscape. Therefore for the future, it will be useful to include grants received in the 20th century by the University of Rochester to see the bigger picture. Similarly the Computer Science study can be done by including abstracts from more institutions. It could also be interesting to compare if the research topics at non-R1 institutions and private corporations are vastly different. Finally, this study could also be expanded into the publications emerging out of the grants, and doing analyses of the relationships between funding and research impact, similar to the work done by Wang and Shapira [3] mentioned earlier.

5. CONCLUSION

For the University of Rochester, we looked at the major topics that received grants over the identified 20 year period. As expected, topics relating to the medical field consistently came up in the model. For the overall period, we identified four major topics: i. Clinical Sciences & patient care; ii. Particle & applied Physics; iii. Genomics; iv. Brain & Imaging Sciences. Diving deeper into the topics over increments of 5 years the model identified 3-4 major topics of research for each period. Terms relating to optics and lasers (LLE/Optics areas of research) were more pronounced in the latter models instead of general Physical Sciences. There was also a bit of a shift to more patient care topics between 2010-2020, with the addition of Psychology/Psychiatry topics. BERTopic identified 33 topics, nearly all of which were from the

STEM fields, majority in medicine, followed by the Physical Sciences. The topics on the top of the list closely matched our LDA results.

When it comes to the Computer Science topics for R1 institutions, our LDA model identified 15 topics from different CS fields, while BERTopic generated 35 topics. We compared these topics with the categories in CS Rankings and found a lot of overlap. Naturally, the categories didn't exactly match so some topics like cryptography and security were mapped to multiple categories. On the whole, the models identified the topics fairly well. What surprised us about this data was the presence of topics relating to conferences, students, researchers and education in both models. This was not a category in CS Rankings but points towards the values of collaboration and sharing knowledge - which makes sense in the field of Computer Science as research evolves quickly with new solutions and updated hardware that increase computational capability, as was also discussed during our class lectures.

REFERENCES

- [1] How is grants data incorporated into Dimensions? *Dimensions*. Retrieved December 12, 2021 from <https://plus.dimensions.ai/support/solutions/articles/23000012993-how-is-grants-data-incorporated-into-dimensions>
- [2] I. Diane Cooper. 2015. Bibliometrics basics. *J Med Libr Assoc* 103, 4 (October 2015), 217–218. DOI:<https://doi.org/10.3163/1536-5050.103.4.013>
- [3] Jue Wang and Philip Shapira. 2015. Is There a Relationship between Research Sponsorship and Publication Impact? An Analysis of Funding Acknowledgments in Nanotechnology Papers. *PLoS One* 10, 2 (February 2015), e0117727. DOI:<https://doi.org/10.1371/journal.pone.0117727>
- [4] Which research categories and classification schemes are available in Dimensions? Dimensions. Retrieved December 7, 2021 from <https://dimensions.freshdesk.com/support/solutions/articles/23000018820-which-research-categories-and-classification-schemes-are-available-in-dimensions>
- [5] Commonwealth of Australia Australian Bureau of Statistics (ABS). 2008. Chapter - DIVISION 08 INFORMATION AND COMPUTING SCIENCES. Retrieved December 7, 2021 from <https://www.abs.gov.au/ausstats/abs@.nsf/0/4C3249439D325D6CA257418000470E3>
- [6] 2021. Discovering Funding Sources with Dimensions at University of Colorado, Boulder. *Dimensions*. Retrieved October 18, 2021 from <https://www.dimensions.ai/resources/discovering-funding-sources-with-dimensions-at-university-of-colorado-boulder/>
- [7] 2019. Dimensions - addressing analytical needs across campus at UC San Diego. *Dimensions*. Retrieved October 18, 2021 from <https://www.dimensions.ai/resources/dimensions-addressing-analytical-needs-across-campus-at-uc-san-diego/>
- [8] 2021. Students at Carnegie Mellon University Use Dimensions to Create Research Funding Dashboards. *Dimensions*. Retrieved December 11, 2021 from <https://www.dimensions.ai/resources/students-at-carnegie-mellon-university-use-dimensions-to-create-research-funding-dashboards/>
- [9] Abdullah Gök, John Rigby, and Philip Shapira. 2016. The impact of research funding on scientific outputs: Evidence from six smaller European countries. *Journal of the Association for Information Science and Technology* 67, 3 (2016), 715–730. DOI: <https://doi.org/10.1002/asi.23406>
- [10] Anna Glazkova. 2021. Identifying Topics of Scientific Articles with BERT-Based Approaches and Topic Modeling. In *Trends and Applications in Knowledge Discovery and Data Mining*, Springer International Publishing, Cham, 98–105.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (May 2019). Retrieved December 12, 2021 from <http://arxiv.org/abs/1810.04805>
- [12] List of research universities in the United States - Wikipedia. Retrieved December 4, 2021 from https://en.wikipedia.org/wiki/List_of_research_universities_in_the_United_States
- [13] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. (May 2002). Retrieved December 4, 2021 from <https://arxiv.org/abs/cs/0205028v1>

- [14] William Mattingly. 2021. *Introduction to Topic Modeling and Text Classification*. Retrieved December 10, 2021 from <http://topic-modeling.pythonhumanities.com/>
- [15] Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3, 993--1022.
- [16] Radim Rehurek. 2021. Gensim: topic modelling for humans. Retrieved December 10, 2021 from https://radimrehurek.com/gensim/auto_examples/tutorials/run_lda.html
- [17] Shashank Kapadia. 2020. Evaluate Topic Models: Latent Dirichlet Allocation (LDA). *Medium*. Retrieved December 12, 2021 from <https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0>
- [18] Maarten Grootendorst. 2021. MaartenGr/BERTopic: Major Release v0.7. DOI:<https://doi.org/10.5281/zenodo.4719700>
- [19] Analyze Results. Retrieved December 12, 2021 from <https://www.webofscience.com/wos/woscc/analyze-results/981a44d8-e112-400e-80e1-8cc0e142e6d0-151f5d71>
- [20] Emery D. Berger. 2020. CSRankings. 2020. Retrieved December 10, 2021 from <https://csrankings.org>
- [21] Big Data Infrastructure at the Crossroads. 2021. *Ithaka S+R*. Retrieved December 10, 2021 from <https://doi.org/10.18665/sr.316121>

APPENDICES

APPENDIX A: DIMENSIONS' FIELDS OF RESEARCH

Categorization based on the Australian and New Zealand Standard Research Classification (ANZSRC). The original FOR system has three levels (2-, 4- and 6-digit codes). The table below shows the implementation in Dimensions from the 2- and 4-digit code categories.

01 Mathematical Sciences 0101 Pure Mathematics 0102 Applied Mathematics 0103 Numerical and Computational Mathematics 0104 Statistics 0105 Mathematical Physics	02 Physical Sciences 0201 Astronomical and Space Sciences 0202 Atomic, Molecular, Nuclear, Particle and Plasma Physics 0203 Classical Physics 0204 Condensed Matter Physics 0205 Optical Physics 0206 Quantum Physics 0299 Other Physical Sciences	03 Chemical Sciences 0301 Analytical Chemistry 0302 Inorganic Chemistry 0303 Macromolecular and Materials Chemistry 0304 Medicinal and Biomolecular Chemistry 0305 Organic Chemistry 0306 Physical Chemistry (incl. Structural) 0307 Theoretical and Computational Chemistry 0399 Other Chemical Sciences	04 Earth Sciences 0401 Atmospheric Sciences 0402 Geochemistry 0403 Geology 0404 Geophysics 0405 Oceanography 0406 Physical Geography and Environmental Geoscience 0499 Other Earth Sciences
05 Environmental Sciences 0501 Ecological Applications 0502 Environmental Science and Management 0503 Soil Sciences 0599 Other Environmental Sciences	06 Biological Sciences 0601 Biochemistry and Cell Biology 0602 Ecology 0603 Evolutionary Biology 0604 Genetics 0605 Microbiology 0606 Physiology 0607 Plant Biology 0608 Zoology 0699 Other Biological Sciences	07 Agricultural and Veterinary Sciences 0701 Agriculture, Land and Farm Management 0702 Animal Production 0703 Crop and Pasture Production 0704 Fisheries Sciences 0705 Forestry Sciences 0706 Horticultural Production 0707 Veterinary Sciences 0799 Other Agricultural and Veterinary Sciences	08 Information and Computing Sciences 0801 Artificial Intelligence and Image Processing 0802 Computation Theory and Mathematics 0803 Computer Software 0804 Data Format 0805 Distributed Computing 0806 Information Systems 0807 Library and Information Studies 0899 Other Information and Computing Sciences
09 Engineering 0901 Aerospace Engineering 0902 Automotive Engineering 0903 Biomedical Engineering 0904 Chemical Engineering 0905 Civil Engineering 0906 Electrical and Electronic Engineering 0907 Environmental Engineering 0908 Food Sciences 0909 Geomatic Engineering 0910 Manufacturing	10 Technology 1001 Agricultural Biotechnology 1002 Environmental Biotechnology 1003 Industrial Biotechnology 1004 Medical Biotechnology 1005 Communications Technologies 1006 Computer Hardware 1007 Nanotechnology 1099 Other Technology	11 Medical and Health Sciences 1101 Medical Biochemistry and Metabolomics 1102 Cardiorespiratory Medicine and Haematology 1103 Clinical Sciences 1104 Complementary and Alternative Medicine 1105 Dentistry 1106 Human Movement and Sports Science 1107 Immunology 1108 Medical Microbiology 1109 Neurosciences	12 Built Environment and Design 1201 Architecture 1202 Building 1203 Design Practice and Management 1204 Engineering Design 1205 Urban and Regional Planning 1299 Other Built Environment and Design

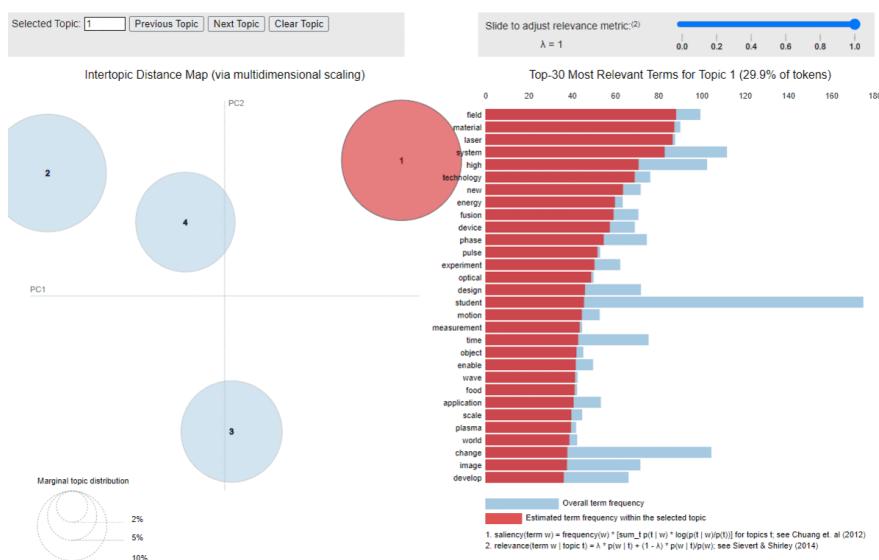
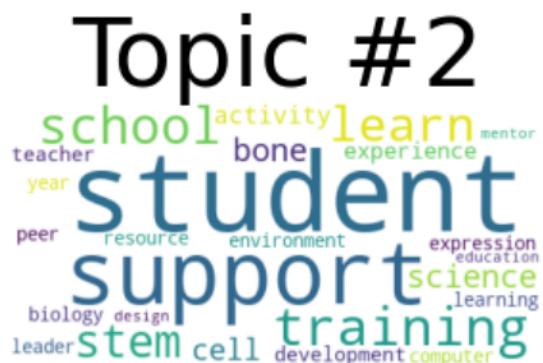
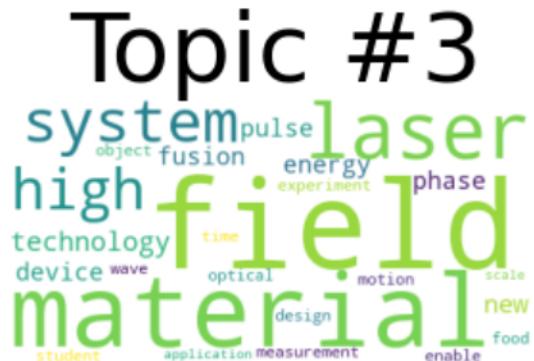
Engineering 0911 Maritime Engineering 0912 Materials Engineering 0913 Mechanical Engineering 0914 Resources Engineering and Extractive Metallurgy 0915 Interdisciplinary Engineering 0999 Other Engineering		1110 Nursing 1111 Nutrition and Dietetics 1112 Oncology and Carcinogenesis 1113 Ophthalmology and Optometry 1114 Paediatrics and Reproductive Medicine 1115 Pharmacology and Pharmaceutical Sciences 1116 Medical Physiology 1117 Public Health and Health Services 1199 Other Medical and Health Sciences	
13 Education 1301 Education Systems 1302 Curriculum and Pedagogy 1303 Specialist Studies In Education 1399 Other Education	14 Economics 1401 Economic Theory 1402 Applied Economics 1403 Econometrics 1499 Other Economics	15 Commerce, Management, Tourism and Services 1501 Accounting, Auditing and Accountability 1502 Banking, Finance and Investment 1503 Business and Management 1504 Commercial Services 1505 Marketing 1506 Tourism 1507 Transportation and Freight Services	16 Studies in Human Society 1601 Anthropology 1602 Criminology 1603 Demography 1604 Human Geography 1605 Policy and Administration 1606 Political Science 1607 Social Work 1608 Sociology 1699 Other Studies In Human Society
17 Psychology and Cognitive Sciences 1701 Psychology 1702 Cognitive Sciences 1799 Other Psychology and Cognitive Sciences	18 Law and Legal Studies 1801 Law 1899 Other Law and Legal Studies	19 Studies in Creative Arts and Writing 1901 Art Theory and Criticism 1902 Film, Television and Digital Media 1903 Journalism and Professional Writing 1904 Performing Arts and Creative Writing 1905 Visual Arts and Crafts 1999 Other Studies In Creative Arts and Writing	20 Language, Communication and Culture 2001 Communication and Media Studies 2002 Cultural Studies 2003 Language Studies 2004 Linguistics 2005 Literary Studies 2099 Other Language, Communication and Culture
21 History and Archaeology 2101 Archaeology 2102 Curatorial and Related Studies 2103 Historical Studies 2199 Other History and Archaeology	22 Philosophy and Religious Studies 2201 Applied Ethics 2202 History and Philosophy of Specific Fields 2203 Philosophy 2204 Religion and Religious Studies 2299 Other Philosophy and Religious Studies		

APPENDIX B: INFORMATION AND COMPUTING SCIENCES SUBCATEGORIES

08 Information and Computing Sciences	
0801	Artificial Intelligence and Image Processing
0802	Computation Theory and Mathematics
0803	Computer Software
0804	Data Format
0805	Distributed Computing
0806	Information Systems
0807	Library and Information Studies
0899	Other Information and Computing Sciences

APPENDIX C: WORD CLOUDS AND TOPIC CLUSTERS FOR UNIVERSITY OF ROCHESTER LDA IN DIFFERENT YEARS

2020 word clouds



2015-2020

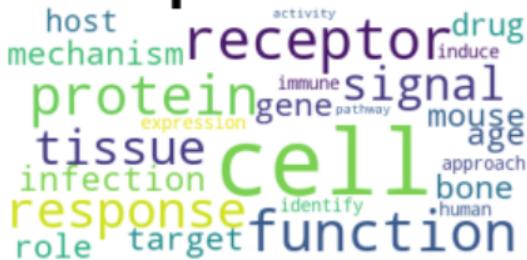
Topic #1



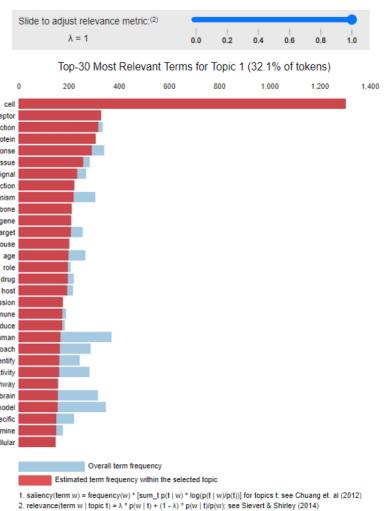
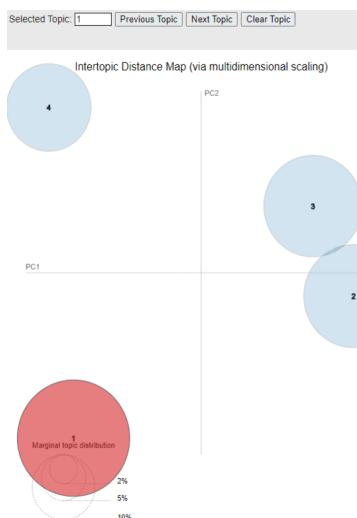
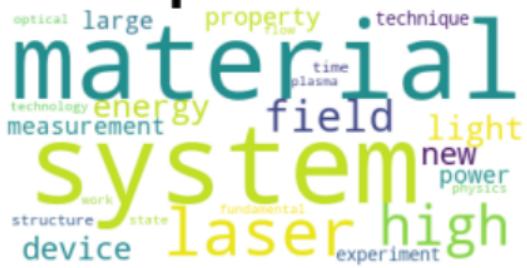
Topic #2



Topic #3



Topic #4



2010-14

Topic #1

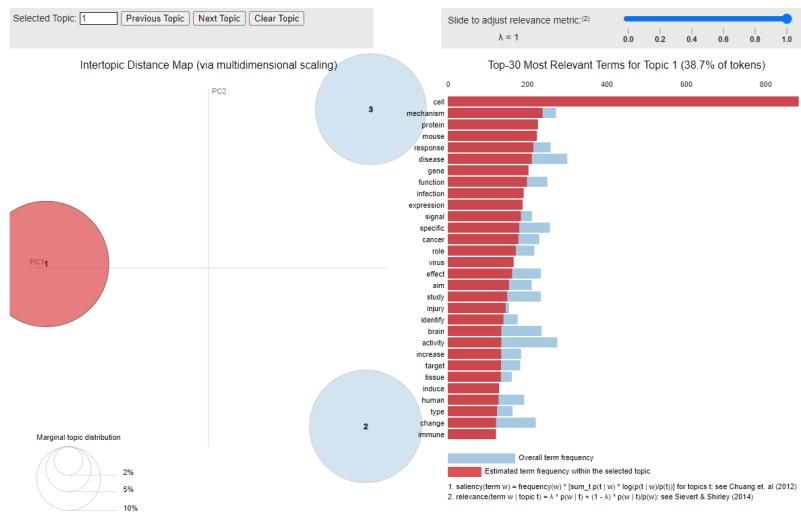
disorder quality human cognitive trial group
 child health outcome year people training
 clinical patient care disease
 population treatment risk
 intervention control social behavior community improve

Topic #2

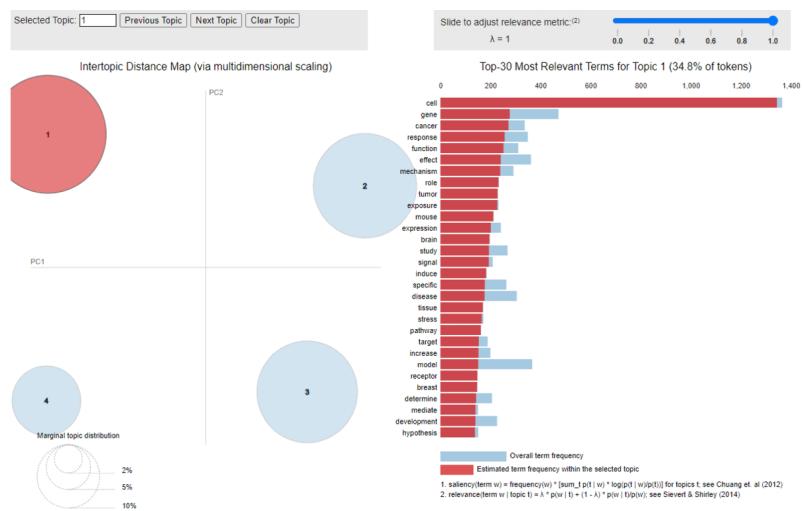
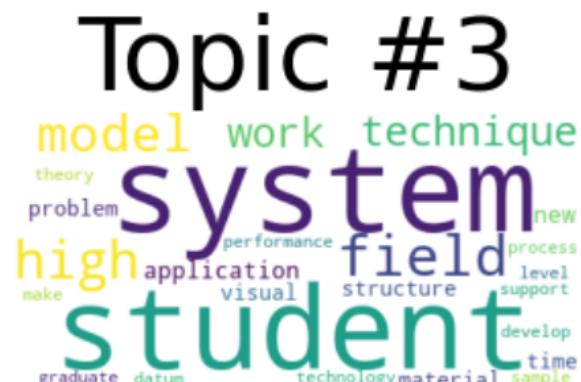
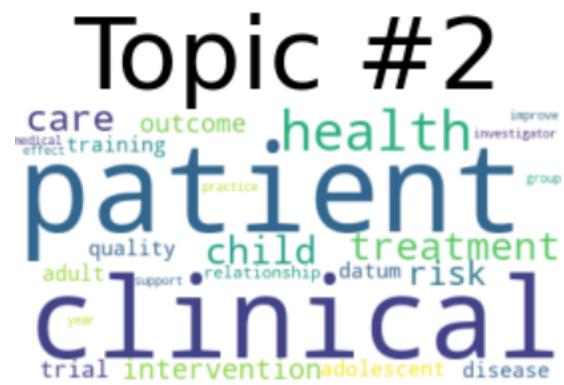
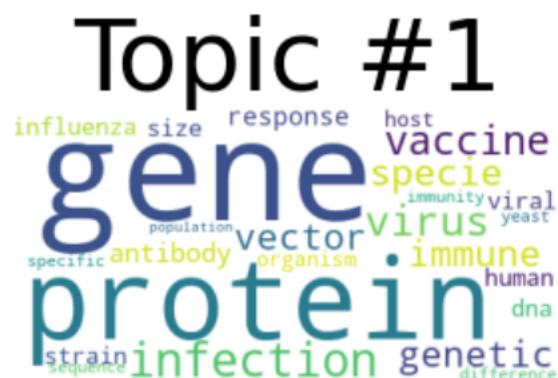
memory optical performance
 technique visual core image time design
 student science
 system new application energy high
 method analysis brain work field
 technology structure information

Topic #3

protein gene specific
 mouse activity virus target
 response cancer identify cell
 function increase brain tissue disease
 mechanism signal effect injury
 aim expression study infection



2005-09



2000-04

Topic #1

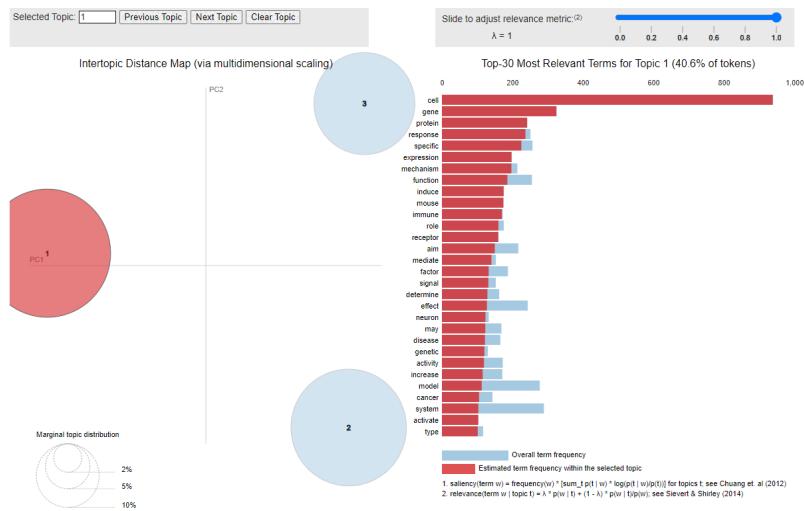
intervention
model
patient
health
care
trial
risk
depression
adolescent
treatment
datum
outcome
effect
family
population
cost
clinical
life
symptom
method
quality
child
suicide
measure

Topic #2

signal
factor
function
mouse
mediate
mechanism
effect
specific
neuron
receptor
response
immune
aim
may
increase
disease
genetic
role
activity
determine
protein
expression

Topic #3

new
training
time
state
area
science
language
support
work
field
group
use
high
optical
material
visual
student
high
optical
material
visual
provide
technique
property
brain
application
use
high
optical
material
visual
student
high
optical
material
visual
provide
technique
property
brain
application



2000

Topic #1

genetic system gene specific type role
induce protein activity
may response expression factor effect
cancer increase factor effect immune mouse
receptor model target function mechanism
cell determine

Topic #2

health intervention cost training age treatment
patient care effect risk quality family
adolescent year plan depression symptom model
trial outcome measure

Topic #3

work optical language science application
system new process
use area high develop material
model visual field brain property
student technique provide group support time

