Technical Report - Detecting Federally Funded Research and Development Trends Using Machine Learning and Information Retrieval Methods

Kathryn Linehan, Eric Oh, Joel Thurston, Guy Leonel Siwe, Madeline Garrett, Sallie Keller, Stephanie Shipp,[1] Audrey Kindlon and John Jankowksi[2]

Abstract

Federal RePORTER, a recently retired federally funded research and development (R&D) grant database, contained a vast amount of information on federally funded R&D and was utilized by researchers, citizens, and policymakers alike to uncover insights. In this report, we provide a classification of research topics contained within Federal RePORTER project abstracts, as well as trends in these topics over time, using natural language processing (NLP) and machine learning techniques. In collaboration with the National Center for Science and Engineering Statistics (NCSES), we examined how the topics and their trends change as a result of the number of topics produced by the model. In addition, we utilized information retrieval techniques to find theme-related topics and their trends over time. This is realized through two case-studies, the first using the theme of pandemics and the second using the theme of coronavirus.

[1]University of Virginia, Biocomplexity Institute, Social and Decision Analytics Division
[2]National Center for Science and Engineering Statistics

Contents

Technical Report - Detecting Federally Funded Research and Development Trends Using Machine Learning and Information Retrieval Methods

Kathryn Linehan, Eric Oh, Joel Thurston, Guy Leonel Siwe, Madeline Garrett, Sallie Keller, Stephanie Shipp,[3] Audrey Kindlon and John Jankowksi[4]

1 Introduction

Automatically extracting topics from large text corpora is crucial for researchers whose collection of documents is too large to feasibly complete this task manually. They may be interested in all corpus topics or those that only relate to a specific theme. For example, one might seek to identify the range of diseases studied by pandemic researchers across a 20-year time span whether the topic involves Spanish Flu, Ebola, Zika, SARS; or one might wish to isolate pandemic research as it relates solely to COVID-19. In this work, we demonstrate the use of two types of topic models, latent Dirichlet allocation (LDA) and non-negative matrix factorization (NMF), to discover federally funded research and development (R&D) topics within a corpus of publicly available grant abstracts from Federal RePORTER, a database of federally funded R&D grants. We also use the information retrieval techniques of term matching and latent semantic indexing (LSI) to discover topics in this corpus related to a specific theme. We then analyze the discovered topics over time using a linear trend analysis.

This work contributes through a novel use of the Federal RePORTER dataset to discover R&D topics and identify R&D topic trends over time. The Organisation for Economic Co-operation and Development (OECD, 2015) defines R&D as "creative and systematic work undertaken in order to increase the stock of knowledge – including knowledge of humankind, culture and society – and to devise new applications of available knowledge". We also discuss the results with respect to the prevalence of topics in federally funded R&D and the stability of these topic models over time. Our work demonstrates the value of applying machine learning and information retrieval techniques to organize and interpret large data sets, highlighted by two case studies - one focusing on pandemic-related topics and the other identifying coronavirus-related topics.

The technical report is organized as follows. Section 2 provides information on the Federal RePORTER data, and Section 3 describes the wrangling steps required to clean and process these data. Next, we review topic modeling (Section 4) and analyze trends in topics across time (Section 5) on our processed Federal RePORTER dataset. Section 6 covers filtering a corpus for a theme (e.g., pandemics, coronavirus) and topic trend analyses through the use of the two case studies. The paper finishes with Sections 7, and 8: conclusion and future work, and acknowledgements. Related work is presented and discussed in Appendix E.

2 Federal RePORTER

Federal RePORTER was a "collaborative effort led by STAR METRICS to create a searchable database of scientific awards from [federal] agencies. This database promoted "transparency and

[3]University of Virginia, Biocomplexity Institute, Social and Decision Analytics Division
[4]National Center for Science and Engineering Statistics

engages the public, the research community, and agencies to describe federal science research investments and provide empirical data for science policy" (U.S. Department of Health and Human Services [HHS], 2020, March 6). STAR METRICS (Science and Technology for America's Reinvestment—Measuring the EffecTs of Research on Innovation, Competitiveness, and Science) was a "federal effort to create a repository of data and tools that will be useful to assess the impact of federal R&D investments" (HHS, 2020, March 6). The Federal Research Portfolio Online Reporting Tools (i.e., Federal RePORTER) was a major component of STAR METRICS. The National Institutes of Health (NIH) and the National Science Foundation (NSF), under the auspices of Office of Science and Technology Policy (OSTP), led this project and funding was provided by NIH, NSF, and other agencies (HHS, 2020, March 6). The effort began as part of the 2009 American Recovery and Reinvestment Act's (ARRA's) Science of Science Policy initiative.

Federal RePORTER was retired on March 1st, 2022, although archived data through fiscal year (FY) 2020 are available at https://federalreporter.nih.gov/. For a future project, we plan to asses the feasibility of pulling data directly from agency award databases to recreate similar data.

Federal RePORTER data consisted of project abstracts and information for more than 1 million federally funded R&D grants from science and technology federal agencies beginning in FY 2000. Project information included title, funding department, funding agency, principal investigator (PI), organization, start date, and FY total cost. Federal RePORTER data were submitted by the individual agencies with most agencies providing this information. In the last few years, Federal RePORTER downloaded some agency data directly from Research.gov.[5]

Previous research using Federal RePORTER. STAR METRICS and Federal RePORTER have been used by others studying similar topics. These data sources first gained popularity in 2011 with a new upswing in use beginning in 2020.

Topic Modeling. In 2011, NIH funded about 80,000 awards. At that time, there was no comprehensive scheme to characterize NIH research. Talley et al. (2011) created the NIH Map ViewerTopic, a database using text mining to extract latent categories and clusters from NIH grant titles and abstracts. It became part of the STAR METRICS system. Using topic modeling and a graph-based clustering method, the authors produced a two-dimensional visualized output. Grants were grouped based on their overall topic-and word-based similarity to one another. This approach allowed for quick and reproducible retrieval of meaningful categorical information, compared to time-consuming searches of NIH websites using keywords. The approach also provided contextual information to ensure transparent and accurate representations of the algorithm-derived topics. Boyd-Graber et al. (2014) described the STAR METRICS NIH Map ViewerTopic (as well as the similar NSF Portfolio Explorer) as a noteworthy example of topic modeling.

Newman et al. (2010) compared human scoring with automatic scoring of topic models using a broad collection of books, news articles, and NIH abstracts. Over 70 coders scored the documents across multiple genres and domains. The authors found that the pointwise mutual information (PMI) of word pairs based on Wikipedia, Google and Medline data sources can predict human scores. The PMI method also has the advantage of identifying junk topics that may be related but not useful to the analyses.

Mimno et al. (2011) conducted an analysis about why topics can be flawed. They used NIH grant abstracts from Federal RePORTER. Based on review by experts, they found that as the number of topics increases, the smallest topics are generally poor quality. They then described an automated evaluation metric to identify topics that do not rely on expert input. Using this automatic evaluation metric, they created a model that improves the topic quality when using NIH grant abstracts.

[5]email from Cindy Danielson (cindy.danielson@nih.gov) on November 10, 2021.

To gain a deeper understanding of R&D investments, Craigin et al. (2012) used STAR METRICS Federal RePORTER data to conduct two portfolio analysis projects. The first project used statistical topic modeling to identify latent concepts in the NSF award portfolio on environmental sustainability and their social components to assess the feasibility of the NSF Portfolio Explorer. The second project created a conceptual framework to describe and quantify the NSF portfolio and contributions to the bioeconomy.

Funding Calculations. Boyack et al. (2020) conducted a portfolio analysis of PubMed literature using a hybrid citation analysis + text similarity approach to identify document relatedness. STAR METRICS Federal RePORTER data were used to proportionally assign funding amounts to the clusters containing these papers. The funding amounts were then summed and used to calculate mean funding from NIH and NSF per paper.

Identifying Research Gaps. The next two articles used Federal RePORTER to identify gaps in research. In a meta-analysis of articles on drug and alcohol use of older adults, Rosen et al. (2019) initiated their research by examining research funded by US federal agencies. They retrieved this information from Federal RePORTER. They found 12 relevant studies between 2011 and 2017 by searching relevant Medical Subject Heading (MeSH) terms related to older adults and substance abuse. Their discussion noted that this small number of projects receiving federal funding to study substance-abusing older adults reflects the lack of researchers in this area. They recommended that funding agencies promote this area of research (Rosen et al., 2019).

In another article, Cunningham et al. (2019) used Federal RePORTER data to identify gaps in federal funding for research related to firearm death and injury to children ages 1-18. This age group's leading causes of death were motor vehicle crashes, firearm injuries, and cancer. Research dollars per death were respectively $26K, $0.5K, and $196K. To calculate these estimates, the authors used Federal RePORTER data. They identified relevant research using MeSH terms supplemented with additional keywords identified from the literature review, a review of funded R&D grants, and expert opinion. They reviewed ten seminal research articles for each injury cause of death to determine related keywords. The keyword search terms were then applied to Federal RePORTER titles and abstracts to select relevant grants (Cunningham et al., 2019). Thus, Federal RePORTER can be used to identify gaps in funding as well as trends.

Grant Publications and Tools. Powell (2019) searched NIH RePORTER data (along with PubMed and Web of Science) to find publications associated with grants. For this article, the focus was on a single federally-funded grant number. PubMed found 980 of the 986 publications, followed by NIH RePORTER with 860, and 787 in Web of Science. Librarians used this information to build a publication list for a specific grant. Researchers can use this information to understand the diffusion of funded research.

Zeng and Acuna (2020) selected NIH abstracts from Federal RePORTER and PubMed to create a grant recommendation system based on publications related to grants. They described their approach as 'learning to rank' to improve how scientists find grants based on their research interests. They built a recommendation system called GotFunding, that learns from the history of publication-grant relationships. Using Federal RePORTER data, they filtered and cleaned the data, removing duplicates and publications without links in Federal RePORTER. They also removed outliers in which a grant had more than ten publications and publications funded by more than three grants. They then created an automatic ranking system using 32 features and tested the performance. These features were then used to test the importance of grant abstract, grant title, and grant agency The grant abstract contributed the most to the selection of grants.

Training Evaluation. The NIH-funded National Research Mentoring Network implemented a Grant Writing Coaching Program (GCP) to provide diverse cohorts of early-career investigators

with coaching throughout the proposal writing and development process. The program goal was to attract more diverse proposal applicants and to increase their success in obtaining grant funding. Weber-Main et al. (2020) evaluated the program's national reach and short-term impact on participants' proposal submissions and funding outcomes. They conducted surveys at 6, 12, and 18 months after the program and verified and supplemented these data through searches of public repositories, including NIH Reporter, Federal Reporter, and Grantome at 6-month intervals. Of the 545 participants in the program over four years (2015-2019), almost 60 percent submitted at least one proposal, and of those who submitted, 40 percent received funding (Weber-Main et al., 2020).

Jacob et al. (2020) compared the publication and funding outcomes of the recipients (fellows) of the R25 Mentored Training for Dissemination and Implementation (D&I) Research in Cancer (MT-DIRC) between 2014 and 2017 with unsuccessful applicants (nonfellows). They extracted publications from SCOPUS and funding information from NIH RePORTER and Federal RePORTER tools. To extract data from the RePORTER tools, the R package "fedreporter" was used. The data were deidentified, coded for D&I research, and aggregated to the applicant level for analysis. The authors used logistic regression models to compute the odds of (1) a D&I publication and (2) US federal grant funding post year of application for fellows (N = 55) and nonfellows (N =47). Fellows were three times more likely to receive grant funding and four times more likely to publish D&I research after receiving the MT-DIRC grant (Jacob et al., 2020).

Other Applications. George Mason University (GMU) created the Science and Technology Campus (SciTech) in 1997 and today there are six centers on the campus. Mahapasuthanon and Hoffman (2019, April 6) used citations from Federal RePORTER and Web of Science to create funding visualizations based on researchers associated with centers at the SciTech campus. They divided GMU publications into two timespans - before 1997 and after 1997. Using a keyword-based analysis, the authors found that the research trends at the SciTech Campus have shifted towards applied health and biological medicine. They used the findings to detect trends by five of the six centers and by individual researchers. (The products from the sixth center, the Virginia Serious Games Institute (VSGI), cannot be visualized by bibliometrics.) The GMU Libraries used the findings to tailor their services to each of the SciTec centers (Mahapasuthanon & Hoffman, 2019, April 6).

Bruce et al. (2019) examined the differences in the use of cooperative agreements and grants when working with private sector firms. They use three sets of data – USASpending.gov, the US Patent and Trademark Office (USPTO) PatentsView.org, and US Office of Personnel Management's (OPM) Central Personnel Data Files (CPDF). These data provide information on the funding characteristics, the patents generated, and the CPDF to identify the number of federal researchers dedicated to early-stage research, development research, and application. Data from NIH RePORTER and the Federal Procurement System were used to fill in missing project information and match to government-supported patents. The authors found that cooperative agreements were more likely to be used for early-stage projects where federal personnel have relevant technical expertise and patents are more likely to be produced from these projects (Bruce et al., 2019).

3 Data Wrangling

To find and analyze federally funded R&D topics, we began with a data discovery process to choose a suitable data source. After exploring USAspending (U.S. Government Accountability Office [GAO], 2021), Federal RePORTER, and separate agency databases (e.g. NSF Award Search), we chose to use Federal RePORTER because it contains project abstracts and contains data from most science and technology federal agencies.

We assumed that Federal RePORTER contained mainly R&D projects. As mentioned in section 2, "STAR METRICS is a federal effort to create a repository of data and tools that will be useful to assess the impact of federal R&D investments" (HHS, 2020, March 6) and Federal RePORTER is a part of STAR METRICS.

However, we recognize that not all federally funded R&D projects appear in Federal RePORTER and that some projects in Federal RePORTER may be the broader category of science and engineering (S&E) and not R&D. S&E includes R&D as well as fellowships, traineeships, and training grants. The Survey of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions: Fiscal Year 2019, Table 1 (National Center for Science and Engineering Statistics [NCSES], 2021) shows that in the years 2008-2019, on average, 89.6% of dollars obligated to universities and colleges for S&E are R&D. We therefore expect the percentage of projects that are S&E but not R&D to be small. In Federal RePORTER we estimate that at least 74.1% of grants are to institutions of higher education; we came to this conclusion by counting the number of organization names that included any of the terms "university", "college", "univ", "school", "institute of technology" or "polytechnic institute".

We utilized Federal RePORTER data reported in FYs 2008-2019. Federal RePORTER started large scale reporting in 2008. When we accessed these data in July 2020 (HHS, 2020, March 6), 2020 data were not yet available for all agencies. In addition, 2019 data were not available for NSF awards. We ingested the Federal RePORTER data utilizing XML and CSV data formats from Federal ExPORTER,[6] and linked project abstracts with corresponding project information. We explored this data for issues that could affect topic model results and/or topic trend analyses, for example missing abstracts and project start dates, duplicate projects, and phrases in abstracts that did not contribute to abstract meaning. These issues were addressed by the wrangling steps described below.

In total, there were 1,156,137 projects in our raw dataset. We removed 42,380 projects with a null (missing) abstract from the dataset. We associated each project with the year given in the project start date. This allowed us to associate the topics assigned to a project (by the topic model) to a given year as well. Any given topic could be assigned to multiple projects and thus be associated with multiple years. By looking across topics assigned to all projects, we could then track topics over time. For projects that were missing a start date, we used the budget start date (if available) to assign a year. Otherwise, projects were assigned a year based on the FY in which they were added to the Federal RePORTER database. See Table 1 for detailed information on start date missingness.

Number of projects with non-null abstract	1,113,757
Percent missing project start date	13.5%
Percent filled in with budget start date	9.7%
Percent filled in with FY reported	3.8%

Table 1: Federal RePORTER, projects reported in FY 2008-2019: Project start date missingness.

Because our goal was to identify the proportion of novel projects associated with a topic in any given year, we deduplicated (i.e., removed all but one entry) projects that shared the same title, abstract, and start date. For example, multi-institutional projects (projects associated with different investigators across two or more universities) with an entry in Federal RePORTER for

[6]Specifically, we used the XML project data and CSV abstract data. We ran into parsing errors when using the CSV project data that did not occur when using the XML data.

each organization would be considered duplicate entries for the same project. We identified and removed all but one instance of each duplicate entry from the dataset to avoid double counting the same project in a given year, thus preventing multi-institutional projects from artificially inflating the number of projects associated with a topic. There were 414,105 duplicate entries removed from the dataset.

For projects that have their funding extended, a new entry is added to Federal RePORTER (with a new start date) for each extension. In this case the project title and abstract were the same as the initial project entry, but the start date was different. By keeping both project entries in the dataset we noted that the project (and the topics contained within) appeared again in the renewal year. We do not see renewals as duplicate data in the same vein as a multi-institutional project logged multiple times during the same start year for the purpose of our trend analysis. In tracking topic trends across time, renewing a project's funding serves as a reaffirmation of interest in the underlying topic on both the part of the researchers and the funding agency.

Prior to their removal, duplicate entries made up 37% of the set of projects with non-null abstracts. To better understand the nature of the projects with duplicate entries in our data set, we examined the duplicate entries (i.e., the information for projects that we removed) and classified them based on their degree of similarity to one another. This allowed us to calculate the "true" number of duplicates by accounting for the fact that some projects are associated with more than two duplicate entries (e.g., one project appeared in the original data set 79 times). Once the number of duplicate entries per individual project is taken into account, we find that the overall percentage of projects with duplicate entries in the set of projects with non-null abstracts is 28% (a reduction of 9%). We note that the maximum number of duplicate entries associated with one project is 79, and 75% of projects with duplicate entries have been duplicated 4 times. We constructed an index to analyze similarity among duplicated project entries looking across variables. We found that on average, projects with duplicate entries for the same start year also had the same PI contact information, Catalog of Federal Domestic Assistance (CFDA) number, NIH Institute and Centers (IC) funding source, and awardee organization name. Areas where they differed were typically FY (of award), budget start and end date.

Some project abstracts were not abstracts, but short phrases such as 'Abstract not provided' and 'No abstract provided'. To remove these types of abstracts from the dataset, we removed projects with abstracts that were less than 150 characters, where the 150 character cutoff was chosen through data exploration. There were 3,520 abstracts that satisfied this criterion. We also removed three projects with abstracts composed entirely of non-alphanumeric characters and one project composed of entirely non-English symbols. An overview of the number of entries removed for these types of reasons is given in Table 2.

Number of projects in raw dataset	1,156,137
Project Removal Reason	**Number Removed**
Projects with a null (missing) abstract	42,380
Duplicate projects	414,105
Projects with an abstract of less than 150 characters	3520
Projects with an abstract of non-alphanumeric characters/symbols	4
Number of projects remaining	696,128

Table 2: Federal RePORTER, projects reported in FY 2008-2019: Data wrangling and project removal overview

3.1 Text Data Preparation

To prepare the abstracts for the topic modeling analysis, we further cleaned the remaining abstracts by removing elements that were not relevant to the specifics of the project (e.g. generic phrases such as 'description (provided by applicant)' and 'end of abstract') and phrases specific to groups of projects (e.g., 'This subproject represents an estimate of the percentage of the Clinical and Translational Science Awards (CTSA) Program funding that is being utilized for a broad area of research (AIDS research, pediatric research, or clinical trials. The Total Cost listed is only an estimate of the amount of CTSA infrastructure going towards this area of research, not direct funding provided by the [National Center for Research Resources] NCRR grant to the subproject or subproject staff.'). These phrases were discovered by manual inspection of the data and were generally found at the beginning or end of many abstracts.

After cleaning the abstracts, we used standard natural language processing (NLP) techniques of tokenization and lemmatization. We converted all tokens to lower-case and removed tokens of generic phrases commonly found at the beginning of abstracts such as ['project', 'summary', 'abstract']. Stop word removal was performed next; this removed one abstract in its entirety, as it was composed exclusively of stop words. Next, we added bi-grams and tri-grams to the abstracts. These steps produced a list of tokens for each abstract. We then standardized hyphens, removed non-alphanumeric characters and leading/trailing underscores in tokens, removed single character tokens and numeric tokens that were not length four (e.g. years). This final token cleaning process created 34 empty abstracts (those with no tokens) that were then removed from the dataset.

3.2 Final Dataset

Our final dataset prepared for analysis included 696,093 projects. Figure 1 shows the distribution of projects by agency and start year for the years 2008-2019 (inclusive). While our dataset (and therefore the topic model) includes projects from the years 1965 to 2019, we limit the figure to 2008-2019, because only 13.3% (92,611) of projects have a start year before 2008.

Most projects in the dataset are funded by HHS (73.5%) and NSF (17.5%). HHS houses NIH, which is responsible for funding 503,425 out of the 511,923 (98.3%) HHS projects. Also of note, the increase in the number of projects in 2009 and 2010 can be attributed to the increased science and science-related funding spurred by the American Recovery and Reinvestment Act of 2009 (ARRA).

4 Topic Modeling

To automatically discover broad categories of federally funded R&D topics, we used LDA and NMF, two popular topic modeling algorithms, to model the project abstracts of our processed dataset. These widely used algorithms can assign multiple topics to a document and are soft clustering, meaning that the same term can appear in multiple topics.

LDA is a probabilistic algorithm that sorts words based on their likelihood of appearing in the same document as one another and reports these common word-association patterns as the most probable topics in the corpus (Blei et al., 2003). The input to LDA includes term frequencies by document. NMF is an approximate matrix decomposition that in the context of topic modeling, finds the document-topic and topic-term weights through iterative optimization (D. D. Lee & Seung, 1999). The input to NMF includes term frequency-inverse document frequency (TFIDF) weights by document.

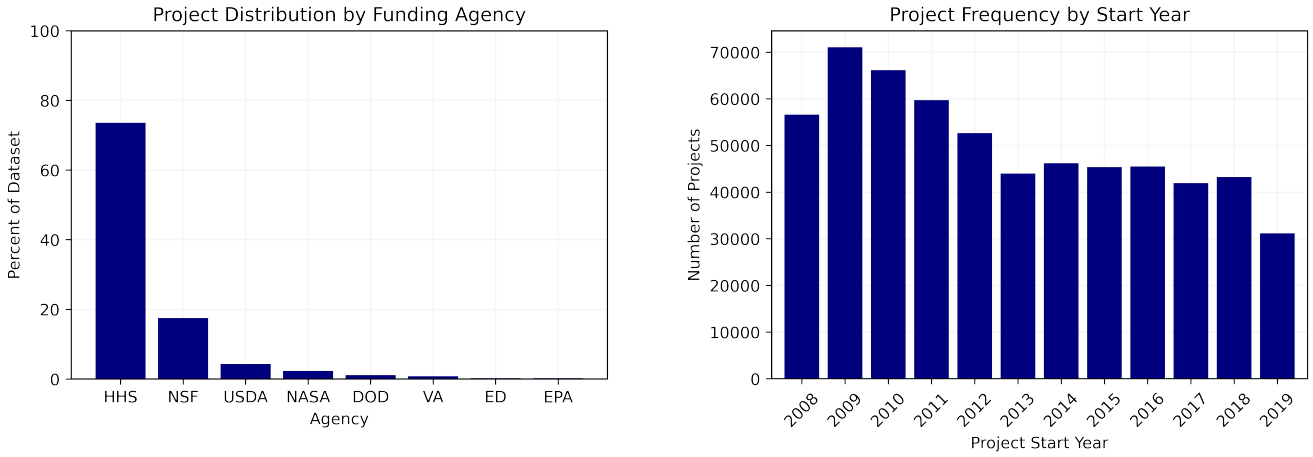

Figure 1: Federal RePORTER, projects reported in FY 2008-2019: Cleaned and Processed dataset. Distributions by project funding agency and start year for the years 2008-2019.

Before running each topic model, we filtered the available terms for the model by excluding terms that appear in less than twenty abstracts or more than 60% of abstracts. Filtering extremes removes terms that are not frequent enough to become a high ranking word in a topic, and terms so common to the corpus that they would not contribute to topic meaning. Based on recommendations in Schofield et al. (2017), we also filtered out three of the four most frequent (remaining) terms in the corpus, 'research', 'study', and 'project', which could be relevant to all topics but would not contribute to topic meaning, since our corpus is comprised of scientific grant abstracts. The most frequent remaining word was 'cell' which we did not filter out since it could contribute to topic meaning.

LDA and NMF are stochastic algorithms and the number of topics present in the corpus is not known in advance so we tested LDA and NMF at varying numbers of topics by performing ten runs of each model for each number of topics. We reported the model C_V topic coherence (Röder et al., 2015) score for each model that we ran. For each topic, C_V topic coherence encodes how often the top n topic words appear together in close proximity within the documents as well as semantic information. We calculated the C_V coherence score for each topic using $n = 10$ and averaged these scores to provide the score for the model as a whole. This measure takes values between 0 and 1 with a higher score indicating a better model with more coherent topics. It is also the coherence measure most correlated with human interpretation of topics (Röder et al., 2015).

The results of our topic model runs are given in Figure 2. These results were computed on the University of Virginia's High-Performance Computing system with Intel Xeon Gold processors of at least 2.10GHz and using 256GB of RAM. In addition, we used a parallel LDA implementation on 40 cores; NMF ran serially. The LDA and NMF models were tested at 5, 10, 15, ..., 130, 140, 150, 175, and 200 topics, and the LDA model parameters for the document-topic and topic-term distribution priors were $\alpha = 1/N$, where N was the number of topics, and $\eta = 0.1$.

Overall, the NMF models have higher C_V topic coherence than the LDA models at each number of topics, but take longer to compute after about 20 topics. As the number of topics increases, the time to compute NMF has more variation and becomes much larger than that of LDA (due to the parallel implementation of LDA). Based on these results, we chose to use NMF as our topic model algorithm for the remainder of the work. We explore models with 50, 100, 150, and 200 topics. In this case, we did not choose the model with the highest mean coherence as an "optimal" model

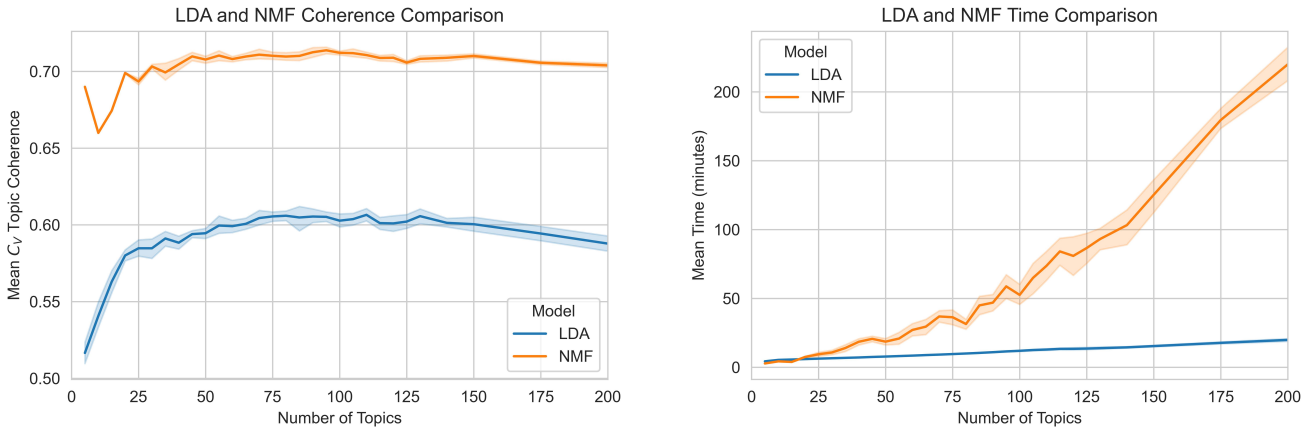

Figure 2: Topic model mean C_V coherence score and run time on processed Federal RePORTER project abstracts reported in FY 2008-2019. The shaded region gives a 95% confidence interval on the measure. Models were tested with 10 runs each at 5, 10, 15, ..., 130, 140, 150, 175, and 200 topics. Additional LDA model parameters used were $\alpha = 1/N$, where N was the number of topics, and $\eta = 0.1$.

since almost all NMF models have a mean coherence of at least 0.70. The choice for the number of topics can be based on the user's need; for example, broad topics or more specific topics.

The topic coherence for each of these models is given in Table 3. Topics from the 50-topic NMF model are given in Table 4 and present broad R&D research areas. The five highest weighted terms per topic are listed, in order of weight starting with the highest weighted term. Most of the topics produced by the model are health related, which is to be expected since grant abstracts from HHS comprise 73.5% of our dataset. We also see a number of general science topics such as FR13 and FR38. In Table 5, topics are presented for a 100-topic NMF model. Some topics have the same top five terms as in the 50-topic model, e.g. X25/FR15 (data analysis) and X68/FR39 (prostate cancer), while others have almost the same top five terms such as X1/FR1 (Alzheimer's disease) and X57/FR31 (computational modeling). Some topics from the 50-topic model appear to get split into more specific topics, e.g. FR6 (neurology) potentially splits into topics X10 (neural disorders) and X87 (traumatic brain injuries), and FR49 (viral infections) potentially splits into topics X35 (hepatitis C virus) and X96 (influenza virus). In the 100-topic model we still see a majority of health-related topics, but other scientific fields such as chemistry (X16), quantum physics (X89), and linguistics (X49) also appear.

Number of Topics	C_V **Topic Coherence**
50	0.716
100	0.714
150	0.709
200	0.701

Table 3: C_V topic coherence scores for NMF topic models on processed Federal RePORTER project abstracts reported in FY 2008-2019.

The topics for the 150-topic and 200-topic NMF models are in Appendix A. As the number of topics in the model increases, topics generally become more focused and specific, and other topics are introduced. For example, in the 200-topic model topic Z168 is about star and galaxy formation;

Label	Top Five Terms
FR1	ad, alzheimer, tau, dementia, pathology
FR2	administrative, core, scientific, meeting, coordinate
FR3	aging, cognitive, age, memory, older
FR4	alcohol, ethanol, drinking, consumption, abuse
FR5	bone, tissue, fracture, osteoporosis, osteoblast
FR6	brain, tbi, injury, neural, mri
FR7	breast, cancer, woman, er, estrogen
FR8	cancer, ovarian, nci, pancreatic, member
FR9	cell, stem, differentiation, tissue, progenitor
FR10	center, resource, support, investigator, facility
FR11	child, parent, language, family, childhood
FR12	clinical, trial, protocol, translational, phase
FR13	conference, meeting, workshop, researcher, international
FR14	core, investigator, provide, service, analysis
FR15	data, analysis, statistical, database, management
FR16	disease, human, kidney, infectious, pd
FR17	dna, repair, damage, replication, genome
FR18	dr, career, mentor, award, director
FR19	drug, compound, screen, target, inhibitor
FR20	gene, expression, genetic, genome, identify
FR21	health, community, disparity, care, public
FR22	heart, cardiac, vascular, injury, mitochondrial
FR23	hiv, aids, infect, infection, antiretroviral
FR24	imaging, image, mri, resolution, tissue
FR25	immune, response, il, cytokine, inflammation

Label	Top Five Terms
FR26	infection, host, pathogen, bacterial, antibiotic
FR27	insulin, diabete, obesity, glucose, metabolic
FR28	intervention, behavior, treatment, social, behavioral
FR29	lung, airway, pulmonary, asthma, injury
FR30	material, chemical, property, chemistry, energy
FR31	model, theory, problem, method, computational
FR32	mouse, model, animal, transgenic, human
FR33	network, social, wireless, communication, node
FR34	neuron, circuit, neural, neuronal, motor
FR35	pain, chronic, opioid, treatment, analgesic
FR36	patient, care, treatment, outcome, therapy
FR37	plant, food, crop, production, soil
FR38	program, member, funding, support, grant
FR39	prostate, cancer, ar, pca, androgen
FR40	protein, membrane, structure, bind, complex
FR41	risk, exposure, factor, woman, environmental
FR42	rna, mirna, expression, translation, micro
FR43	signal, receptor, pathway, regulate, activation
FR44	student, science, stem, school, undergraduate
FR45	system, technology, device, design, develop
FR46	training, trainee, faculty, career, mentor
FR47	tumor, therapy, target, metastasis, growth
FR48	vaccine, antibody, antigen, vaccination, protection
FR49	virus, viral, infection, hcv, influenza
FR50	water, climate, change, ecosystem, forest

Table 4: Top five topic terms from NMF model with 50 topics on processed Federal RePORTER project abstracts reported in FY 2008-2019. Topics are listed and labeled in alphabetical order by the most important topic term.

this topic does not appear in the other three models. On the other hand, there are still topics that appear in all four topic models with similar top five terms, e.g. FR1/X1/Y2/Z2 (Alzheimer's disease) and FR23/X38/Y55/Z77 (HIV/AIDS).

We can also analyze which topics appear the most and which are the most predominant in the abstracts. As an example we use the 50-topic NMF model and present Figures 3 and 4: the ten topics appearing in the highest and lowest percentage of abstracts, and the ten topics that have the highest and lowest percentage of being the predominant topic, where the predominant topic for an abstract is defined as the highest weighted topic for that abstract. In both figures, topics are given by their labels and five most important terms.

In Figure 3, the ten topics appearing in the highest percentage of abstracts are generally broader than the ten topics that appear in the lowest percentage of abstracts. For example, topic FR15 is a general data analysis topic whereas topic FR27 is specifically about diabetes. Intuitively, this behavior in the model results is expected. Comparing Figures 3 and 4, we see that there is not a particular pattern relating topics that appear in a high percentage of abstracts and topics that are predominant. While the most appearing topics are in about 50% of abstracts, the most predominant topics are predominant for about 3-6% of abstracts. It is interesting to note that FR16 (disease), is the topic appearing in the second highest percentage of abstracts (about 55%) and is the topic with the second highest percentage of being the predominant topic (about 4%), signaling its importance in the corpus. Topic FR15 (data analysis) is the topic appearing in the highest percentage of abstracts yet is the fourth least predominant topic implying that data analysis occurs in many abstracts yet is not the main focus of the abstract. While we can draw some insights from these results, we are also ignoring the information of 'how much' a topic appears in each abstract. We

Label	Top Five Terms
X1	ad, alzheimer, tau, dementia, cognitive
X2	administrative, scientific, meeting, coordinate, management
X3	adolescent, youth, family, behavior, substance
X4	aging, age, older, adult, muscle
X5	alcohol, ethanol, drinking, consumption, alcoholism
X6	animal, model, rat, small, colony
X7	antibody, antigen, peptide, epitope, bind
X8	asthma, airway, allergic, asthmatic, allergen
X9	bone, fracture, osteoporosis, osteoblast, skeletal
X10	brain, neural, mri, disorder, region
X11	breast, cancer, er, metastasis, estrogen
X12	cancer, nci, pancreatic, ovarian, member
X13	care, quality, provider, practice, medical
X14	cell, differentiation, type, culture, line
X15	center, university, director, resource, leadership
X16	chemical, chemistry, reaction, synthesis, metal
X17	child, parent, family, pediatric, childhood
X18	climate, change, ecosystem, forest, species
X19	clinical, trial, protocol, translational, phase
X20	cocaine, addiction, nicotine, self, abuse
X21	community, partnership, outreach, partner, education
X22	conference, meeting, researcher, hold, field
X23	core, investigator, provide, analysis, expertise
X24	crop, soil, production, management, agricultural
X25	data, analysis, statistical, database, management
X26	disease, cause, infectious, progression, pd
X27	dna, repair, damage, genome, replication
X28	dr, director, career, mentor, award
X29	drug, abuse, target, delivery, discovery
X30	engineering, design, education, engineer, technology
X31	exposure, environmental, effect, chemical, pregnancy
X32	facility, instrument, equipment, laboratory, user
X33	food, safety, consumer, agriculture, intake
X34	gene, expression, genetic, genome, identify
X35	hcv, infection, hepatitis_c, chronic, viral
X36	health, disparity, public, population, mental
X37	heart, cardiac, failure, muscle, cardiomyocyte
X38	hiv, aids, infect, infection, antiretroviral
X39	human, genetic, genome, skin, model
X40	il, cytokine, cd4, th2, treg
X41	imaging, image, mri, resolution, optical
X42	immune, response, innate, antigen, dc
X43	infection, host, pathogen, bacterial, bacteria
X44	inflammation, macrophage, inflammatory, induce, activation
X45	insulin, diabete, glucose, type, islet
X46	intervention, randomize, control, group, base
X47	investigator, support, fund, pilot, grant
X48	kidney, renal, ckd, chronic, transplant
X49	language, speech, linguistic, word, processing
X50	liver, hepatic, hepatocyte, hcc, fibrosis
X51	lung, pulmonary, copd, airway, fibrosis
X52	malaria, parasite, vector, transmission, control
X53	material, property, energy, polymer, device
X54	mechanism, aim, regulate, role, function
X55	membrane, lipid, channel, fusion, transport
X56	memory, cognitive, learning, task, impairment
X57	method, model, computational, develop, simulation
X58	mitochondrial, mitochondria, pd, ros, mt
X59	mouse, model, transgenic, mutant, strain
X60	network, wireless, communication, node, connectivity
X61	neuron, neuronal, circuit, motor, synaptic
X62	obesity, weight, metabolic, diet, energy
X63	pain, chronic, opioid, analgesic, neuropathic_pain
X64	patient, outcome, therapy, surgery, improve
X65	plant, species, crop, pathogen, trait
X66	product, contract, testing, development, infectious
X67	program, member, department, year, phd
X68	prostate, cancer, ar, pca, androgen
X69	protein, interaction, bind, proteomic, peptide
X70	receptor, ligand, bind, agonist, gpcr
X71	resistance, antibiotic, mutation, resistant, inhibitor
X72	risk, factor, genetic, population, variant
X73	rna, mirna, translation, expression, micro
X74	sample, biomarker, assay, analysis, detection
X75	science, scientific, scientist, policy, collaboration
X76	screen, compound, assay, inhibitor, molecule
X77	service, resource, provide, share, support
X78	signal, pathway, kinase, activation, wnt
X79	sleep, circadian, disorder, insomnia, sleep_disturbance
X80	social, behavior, asd, behavioral, individual
X81	spore, translational, developmental, drp, career
X82	stem, hsc, hematopoietic, progenitor, differentiation
X83	stress, response, er, depression, anxiety
X84	structure, structural, complex, crystal, bind
X85	student, undergraduate, graduate, college, faculty
X86	system, technology, device, power, sensor
X87	tbi, injury, traumatic, recovery, trauma
X88	teacher, school, learning, classroom, mathematics
X89	theory, problem, quantum, physic, pi
X90	tissue, specimen, pathology, adipose, organ
X91	training, trainee, faculty, career, mentor
X92	treatment, therapy, efficacy, dose, effect
X93	tumor, metastasis, growth, therapy, metastatic
X94	vaccine, vaccination, protection, adjuvant, protective
X95	vascular, endothelial, blood, flow, vessel
X96	virus, viral, influenza, replication, infection
X97	visual, vision, eye, retinal, neural
X98	water, surface, irrigation, quality, energy
X99	woman, pregnancy, hpv, maternal, female
X100	workshop, participant, researcher, international, hold

Table 5: Top five topic terms from NMF model with 100 topics on processed Federal RePORTER project abstracts reported in FY 2008-2019. Topics are listed and labeled in alphabetical order by the most important topic term.

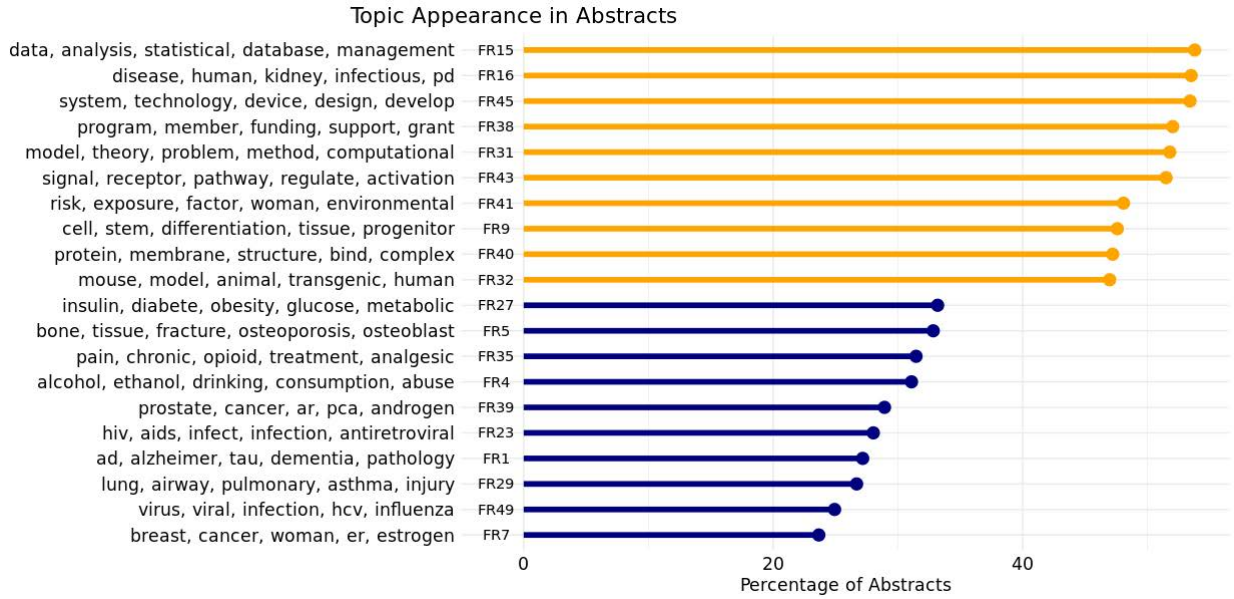

Figure 3: The ten topics from the 50-topic NMF model on processed Federal RePORTER project abstracts reported in FY 2008-2019 appearing in the highest (orange) and lowest (blue) percentage of documents. Topics are given by their five highest weighted terms and labels.

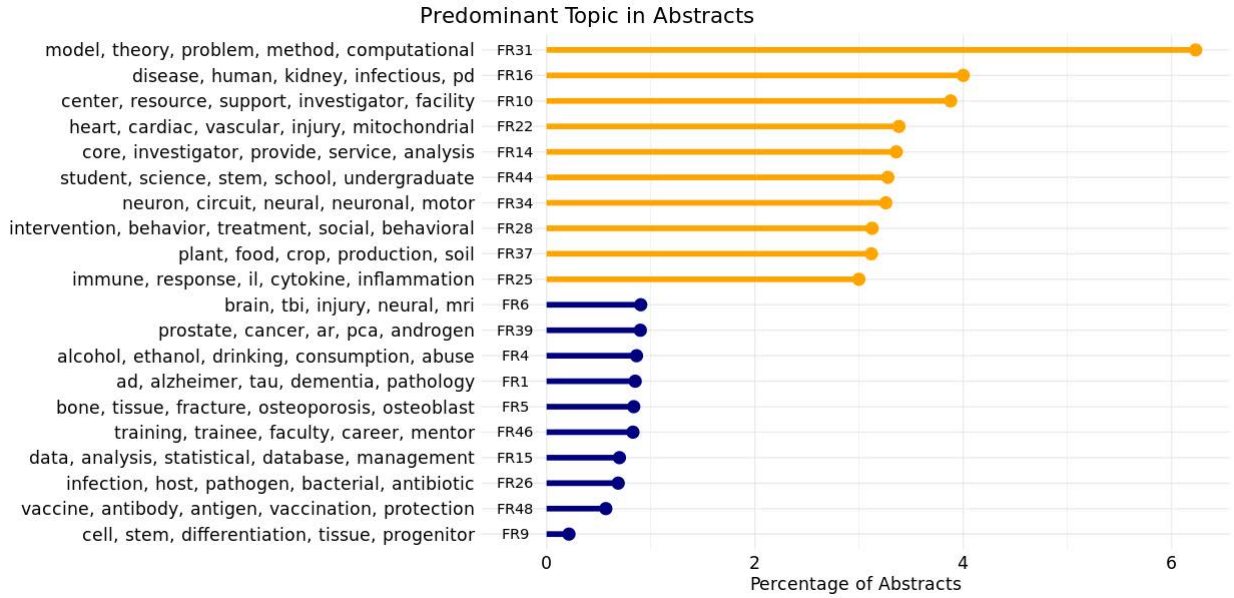

Figure 4: The ten topics from the 50-topic NMF model on processed Federal RePORTER project abstracts reported in FY 2008-2019 that have the highest (orange) and lowest (blue) percentage of being the predominant topic. Topics are given by their five highest weighted terms and labels. A predominant topic for an abstract is defined as the highest weighted topic for that abstract.

use this information to analyze topic trends over time in Section 5.

Lastly we assess the stability of each of these models with respect to the topics that are produced across various model runs. One known, yet often ignored aspect of topic models in practice is that different runs of the same model on the same data can produce different topics. This instability results from the initialization required to run the optimization to find a local solution. It manifests as different terms associated with topics and different documents associated with topics across each

initialization. To quantify the extent of this instability, we computed three measures proposed in (Belford et al., 2018), Descriptor Set Difference (DSD), Topic-Term Stability (TS), and Partition Stability (PS). Broadly, DSD, TS, and PS measure the stability of the set of top terms across all topics, the top terms for matched individual topics, and the predominant topic for each document, respectively, for two models with different seed initializations. These measures are then averaged across pairwise comparisons of r runs of the model. Values for the average DSD, TS, and PS take the range $[0, 1]$ where DSD values closer to 0 represent more stability and TS and PS values closer to 1 represent more stability. Stability results for each topic model are given in Table 6 and indicate that the topic models are relatively stable as determined by the DSD and TS measures. The PS measure, however, decreases significantly as the number of topics increases. This can be attributed to the fact that the document-topic weights become more uniform as the number of topics increases, yielding different predominant topics for each document across runs.

Number of Topics	DSD	TS	PS
50	0.14	0.77	0.70
100	0.12	0.75	0.59
150	0.12	0.70	0.48
200	0.11	0.70	0.39

Table 6: Stability measures for NMF topic models on processed Federal RePORTER project abstracts reported in FY 2008-2019. DSD, TS, and PS are given as average measures across $r = 10$ runs utilizing 10 terms to describe the topics.

5 Topic Trend Analysis

We then analyzed the resultant topics from our topic models from Section 4 by examining their weights over time. This allowed us to roughly characterize the relative prevalence of each topic in the corpus of Federal RePORTER abstracts over time. We limited our analysis of R&D trends in Federal RePORTER to be between the years 2010-2019.

We used the document-topic distribution to obtain the topic weights for each abstract. Then, for each year the means of the weights of the topics are calculated for the project abstracts that have the given start year. The relationship between mean weight and year for each topic was modeled using linear regression, thus capturing the trend of the topic weights over time. In the following discussion, we use the size and sign of the regression slope to characterize the prevalence of each topic over the time period considered; however, we do not focus on the p-value, or resulting (non)-significance, of the test statistic nor do we dichotomize topics into "hot" or "cold" topics. We present results for the 50 topic model in this section and give other results including those for the 100, 150, and 200 topic models in Appendix B.

5.1 Results

We present the ten topics with the largest positive slopes in Figure 5 for the 50 topic model. Many of the topics are medically related, including topics related to neurodegenerative diseases (FR1), the cognitive changes related to aging (FR3), chronic pain treatment (FR35), clinical trials (FR12), kidney disease (FR16), and cardiovascular disease (FR22). In addition, there are some more general topics on behavioral treatments (FR28), patient care (FR36), and grant awards (FR18). Lastly,

there is a basic science related topic on neurons and circuits (FR34). The ten topics with largest negative trend are shown in Figure 6. There are fewer HHS specific topics compared to the topics with the largest positive trends. Topics with large negative slopes include crop production (FR37), climate change (FR50), protein structures (FR40), computational models (FR31), science, technology, engineering, and math (STEM) education (FR44), conferences (FR13), chemical engineering (FR30), gene expression (FR20), breast cancer (FR7), and health disparities (FR21).

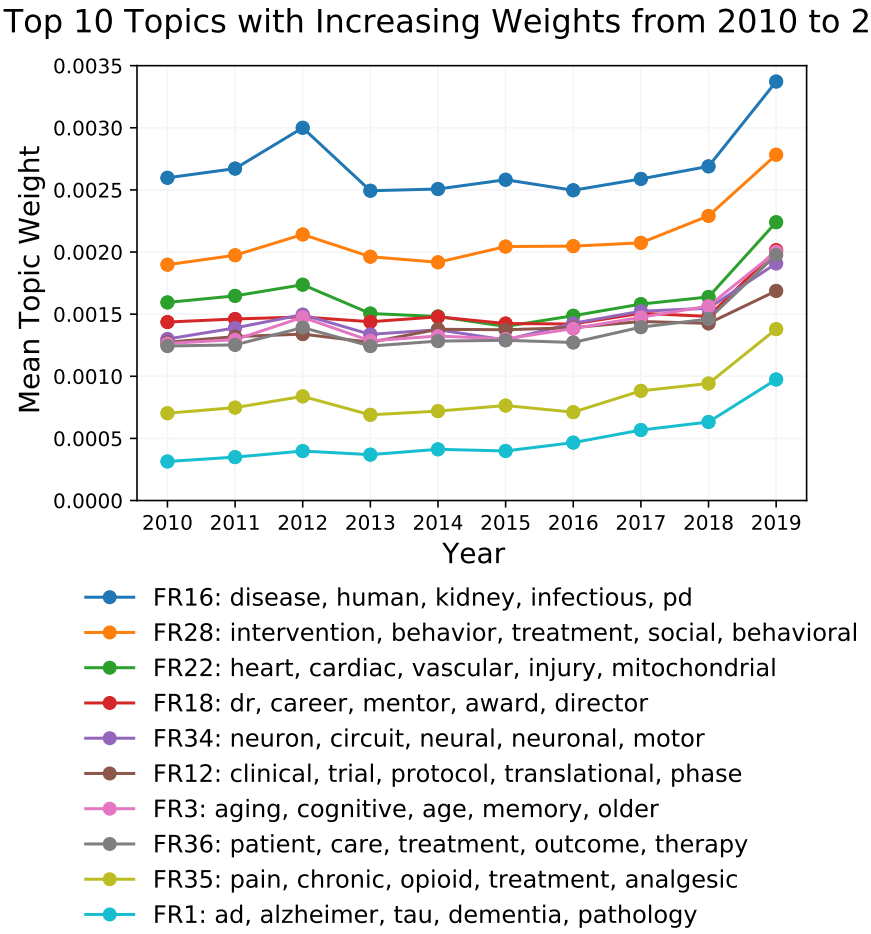

Figure 5: Ten topics with largest positive regression line slopes from the 50 topic model. The slopes are calculated using the weights from 2010 through 2019.

Figures 7 and 8 display the mean weights for each year and the corresponding regression lines for all 50 topics for the years 2010 to 2019. Table 15 in Appendix B presents the top 5 words, count of abstracts with weights greater than 0, slope of the regression line, standard error (SE), and p-value for all topics. Many of the 30 topics not captured by the ten largest positive or negative slopes have relatively stable mean weights across the years considered. Some topics, however, maintain higher weights than some of the topics with largest positive slopes. For example in Figure 7, topics on RNA expression (FR42) and environmental risk factors (FR41) maintain higher weights than those of neurodegenerative diseases (FR1), which has one of the highest slopes. Similarly in Figure 8, topics on MRI imaging (FR24) and diabetes (FR27) maintain relatively high mean weights. Also from Figure 8, we see that topics on clinical trials (FR12), ovarian cancer (FR8), environmental risk factors (FR41), and kidney disease (FR16) all maintain relatively high weights, indicating that these topics are present across many projects.

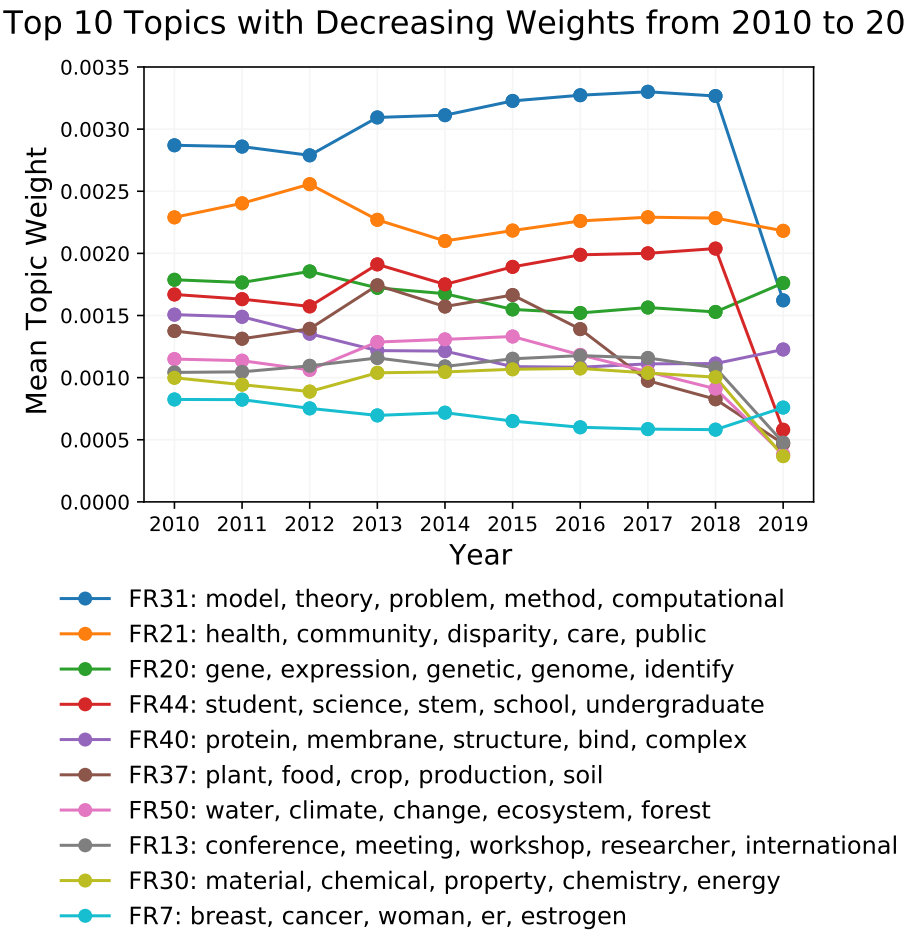

Figure 6: Ten topics with largest negative regression line slopes from the 50 topic model. The slopes are calculated using the weights from 2010 through 2019.

Many all of the topics, however, have weights in 2019 that are much higher than the trends through 2018 would suggest. Given that Federal RePORTER did not contain NSF project data for 2019 when we accessed it, it is possible that the medicine and health topics, which are likely funded by HHS, have higher weights than they would if the NSF project data were present. Thus, we ran the same trend analysis limited to the years 2010 through 2018 and presented the results in Appendix Sections B.2 - B.3.

6 Topic Trends Related to a Specific Theme

We now address the problem of how to find topics and trends related to a specific theme within the corpus. For example, in light of recent events, we mined Federal RePORTER for the theme of pandemics. Our goal was to find Federal RePORTER project abstracts related to pandemics and then perform a topic trend analysis on the relevant abstract subset. We also performed a second case study using the theme of coronavirus. For a theme that has low signal in the corpus, the results of this process can identify topics that generally would not occur in a topic model on the entire corpus.

To find abstracts relevant to a specific theme we utilized the information retrieval methods of term matching and LSI. Term matching is a common technique of identifying documents relevant

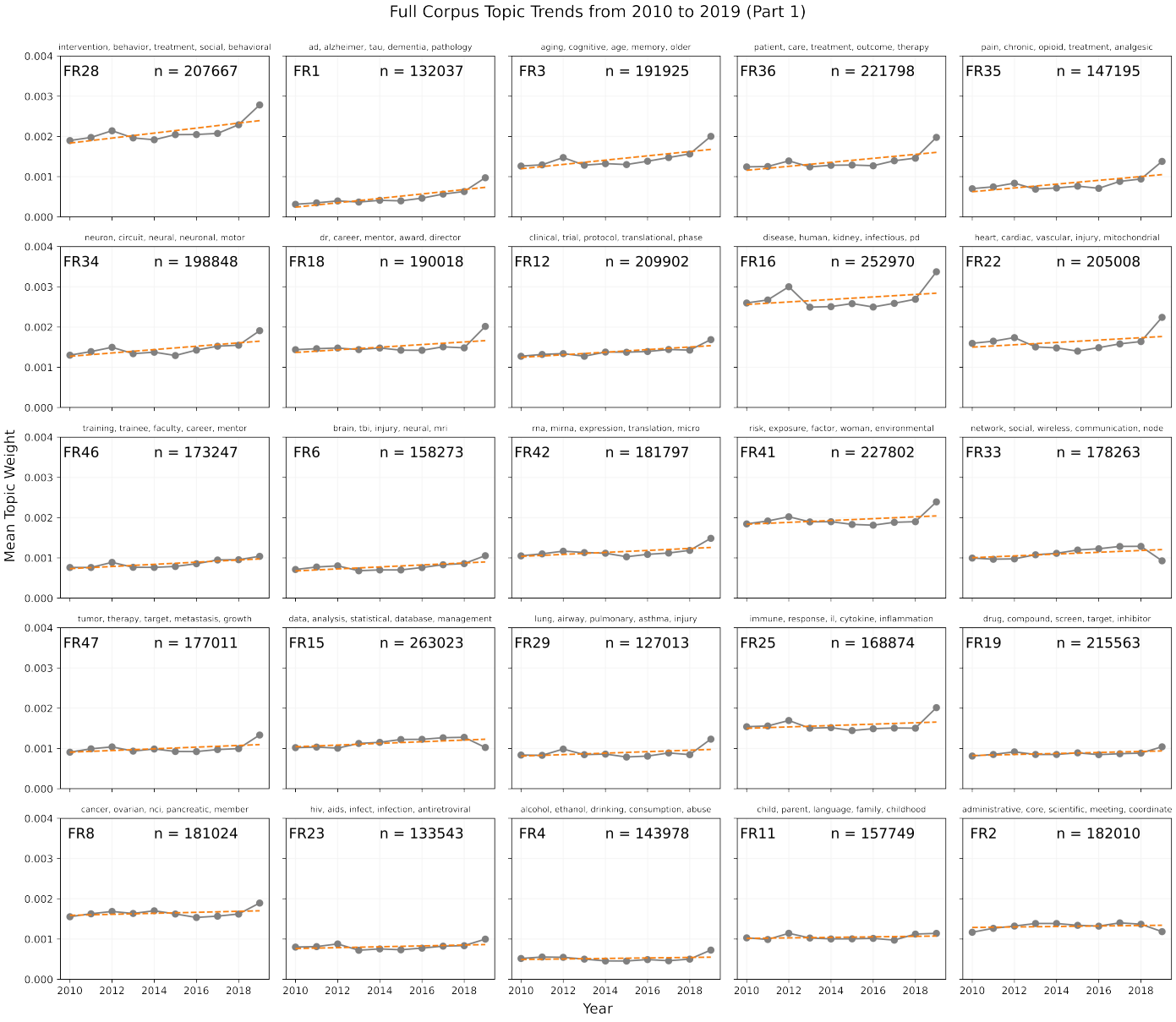

Figure 7: Topic trend results for topics 1 to 25 of 50 topics produced by an NMF model on the full corpus. Trends in topic prevalence are captured between 2010-2019 and topics are ordered from largest positive to largest negative regression line slopes. Topics with positive slopes have orange regression lines and topics with negative slopes have blue regression lines. Standard errors on the means are represented on each plot using error bars.

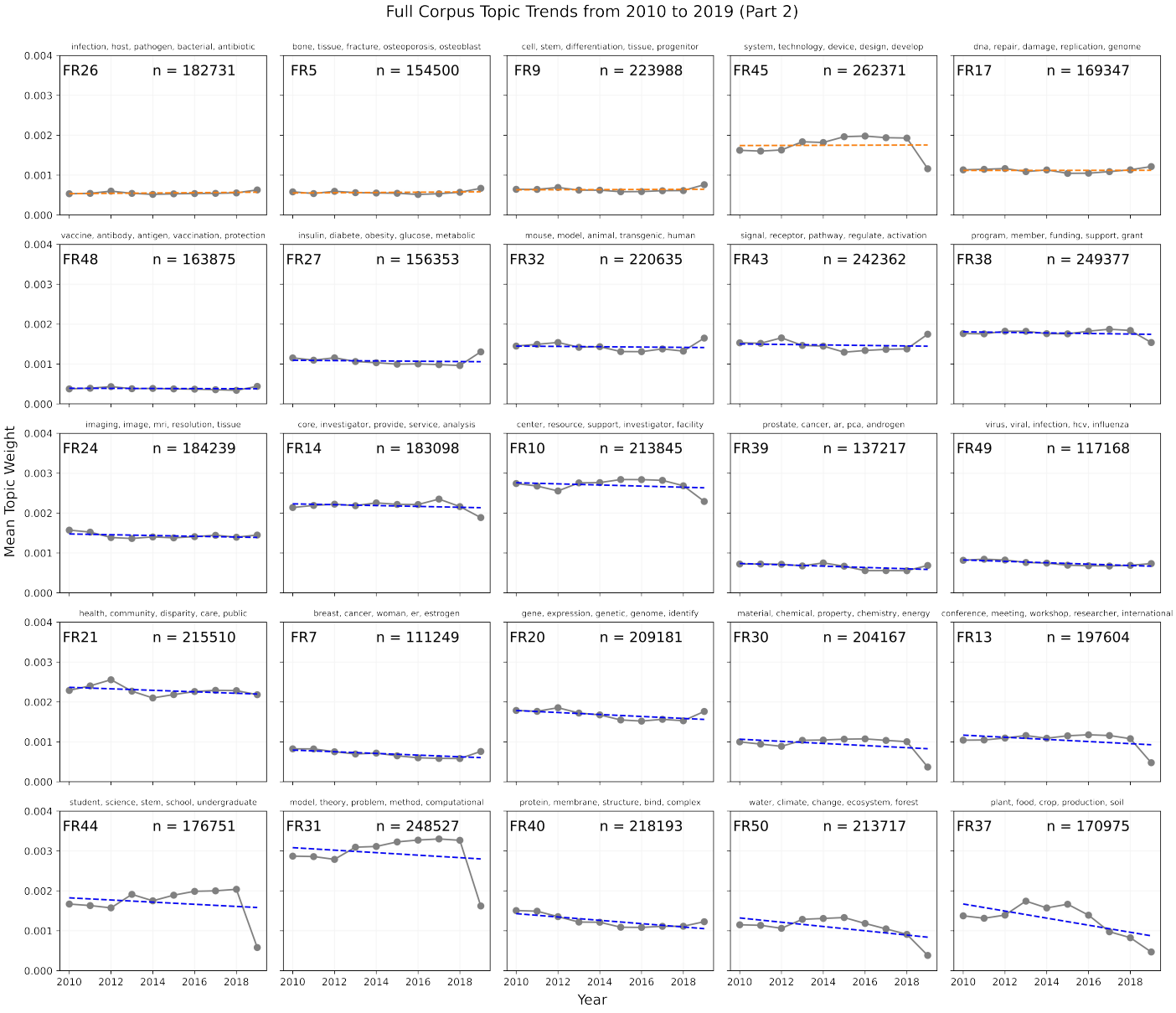

Figure 8: Topic trend results for topics 26 to 50 of 50 topics produced by an NMF model on the full corpus. Trends in topic prevalence are captured between 2010-2019 and topics are ordered from largest positive to largest negative regression line slopes. Topics with positive slopes have orange regression lines and topics with negative slopes have blue regression lines. Standard errors on the means are represented on each plot using error bars.

to a theme by marking a document as relevant if it contains a particular keyword or keywords. A challenge with term matching is the construction of a keyword list that fully and non-ambiguously describes the theme. It is common for expert input to be used in this construction (Eads et al., 2021; OECD, 2019).

One of the pitfalls of term matching is that it will not identify theme-relevant documents that use terms other than those in the keyword list to discuss the theme. So in addition to term matching, we utilized LSI (Deerwester et al., 1990) to address this as it can identify relevant documents that may not necessarily contain the keyword(s). LSI computes a relevance score for each document to the list of keywords, or query as it is commonly called in information retrieval. A higher score corresponds to higher relevance to the search query and generally the top scoring documents are considered relevant to the query. For information about how the relevance score is calculated, see Appendix C. Term matching and LSI do not necessarily produce the same information retrieval results and it was suggested by Deerwester et al. (1990) that LSI be "regarded as a potential component of a retrieval system, rather than a complete retrieval system".

6.1 Case Study 1: Pandemics

To create a themed pandemic corpus, we used term matching and LSI on the final abstract tokens in our processed Federal RePORTER dataset. Our initial term matching keyword list only contained one word: pandemic. However, our final abstract tokens include bi-grams and tri-grams that include the word pandemic, so we included all tokens that contained 'pandemic' in the keyword list.[7] Any abstract that included at least one of these keywords was included in the themed corpus. These keywords also served as the query for LSI. We used a rank-50 truncated singular value decomposition (SVD) as the matrix approximation for LSI and included the 2,000 abstracts with the highest relevance scores to the query in the themed corpus. The rank and number of abstracts to keep were chosen by a trial and error process that included manual inspection of abstracts and results from a themed corpus topic model. We note that the matrix spectrum and relevance score distribution did not provide any clear guidance when choosing these parameters. There was some overlap in the abstracts chosen to be included in the themed corpus by term matching and LSI.[8] See Table 7 for details.

To identify topic trends in Federal RePORTER within the area of pandemics, we utilized an NMF topic model of 30 topics on the themed pandemics corpus where we excluded the terms

[7]Pandemic Keywords: 1918_influenza_pandemic, 1918_pandemic, 1957_1968_pandemics, 2009_pandemic_h1n1, aidspandemic, andpandemic, andpandemics, apandemic, bothpandemic, causedpandemics, co_pandemic, detetermrminineififtththesuprragenomeoffttthepandemicicclonesiis, devastating_pandemics, ebolapandemic, epidemics_occasional_pandemics, epidemics_pandemics, escalatingpandemic, establishingpandemic, flu_pandemic, futurepandemics, globalpandemic, greatpandemic, growingpandemic, h1n1_influenza_pandemic, h1n1_pandemic, h5n1_pandemic, hivpandemic, humanpandemics, increasedpandemic, influenza_pandemics, influenzapandemic, influenzapandemics, inpandemic, inter_pandemic, interpandemic, majorpandemic, non_pandemic, occasional_pandemics, occasionalpandemics, ofpandemic, ofpandemics, pandemic, pandemic57499, pandemic_1918, pandemic_flu, pandemic_h1_n1, pandemic_h1n1, pandemic_h2n2, pandemic_influenza, pandemic_non_pandemic, pandemic_preparedness, pandemically, pandemicand, pandemiccompare, pandemicdisease, pandemicemergence, pandemicflu, pandemichuman, pandemicin, pandemicinfection, pandemicinfluenza, pandemicpreparedness, pandemics, pandemics_1918, pandemicsettings, pandemicsthat, pandemicstrain, pandemicthreat, pandemicvaccine, possiblepandemics, post_pandemic, pre_pandemic, prepandemic, recurrentpandemics, seasonal_pandemic, thepandemic, thispandemic, threepandemics, understandpandemic, withpandemic, worldwide_pandemics_1957, yearly_epidemics_occasional_pandemics

[8]Every document receives a score in the LSI process. If we had used more documents than the 2,000 with the highest score, this overlap would have been larger. In fact, 75% of the documents marked as relevant by term matching are in the top 12.29% of LSI relevance scores.

TM	LSI	TM & LSI	Total
1839	1531	469	3839

Table 7: Pandemics themed corpus contribution by information retrieval method. The units for each column are number of abstracts. TM: term matching, TM and LSI: overlap of abstracts returned by both methods.

'research', 'study', and 'project' and those that appeared in less than three abstracts. The number of topics for the model was chosen by inspection based on topic specificity. We computed the three stability measures for $r = 20$ runs and 10 terms to describe the topics. The average DSD, TS, and PS were 0.13, 0.81, and 0.83, indicating that our pandemics focused topic model is quite stable. Topics are given by their label and top five words in Table 8. Several viruses appear in these topics including influenza (P7, P13, P14, P15), HIV/AIDS (P10), Zika (P30), West Nile (P29), hepatitis (P9), and tuberculosis (P24). A number of topics mention vaccines (P2, P15, P21, P27) and there are also some general topics such as P8 (manufacturing and facilities) and P28 (human viruses).

Label	Top Five Terms
P1	antibody, neutralization, bind, human, neutralize
P2	attenuate, vaccine, virus, live, candidate
P3	cell, response, memory, infection, cd4
P4	core, diagnostic, support, technology, poc
P5	dengue, virus, serotype, denv, den
P6	drug, inhibitor, compound, resistance, antiviral
P7	epitope, influenza, ha, conserve, strain
P8	facility, product, manufacturing, material, raw
P9	hcv, hepatitis, chimpanzee, genotype, liver
P10	hiv, aids, infect, env, prevention
P11	host, viral, rna, antiviral, replication
P12	hsv, genital_herpes, dl5_29, herpes_simplex_virus, infection
P13	iav, lung, sp, response, evolution
P14	influenza, animal, ecologic, cross_protection, immune
P15	influenza, vaccination, strain, child, virus

Label	Top Five Terms
P16	mucus, igg, trap, vaginal, trapping
P17	obesity, cancer, insulin, obese, diabete
P18	organism, gene, sequence, ortholog, genome
P19	patient, clinical, trial, dose, care
P20	protein, bind, fusion, structure, membrane
P21	rsv, child, respiratory_syncytial, vaccine, mtase
P22	siv, challenge, mucosal, transmit_founder, transmission
P23	swine, prrsv, prrs, pig, porcine
P24	tb, mtb, co_infection, infection, treatment
P25	training, program, trainee, student, university
P26	transmission, intervention, disease, model, health
P27	vaccine, adjuvant, protection, antigen, immune
P28	virus, human, cause, infect, vector
P29	wnv, flaviviruse, denv, flavivirus, infection
P30	zikv, zika_virus_zikv, microcephaly, infection, fetal

Table 8: Top five topic terms from NMF model with 30 topics on themed pandemics corpus. Topics are listed and labeled in alphabetical order by the most important topic term.

We present the trend of each topic over the years 2010-2019 in Figure 9 and the topic label, top five words, and slope and p-value of the topic mean weight regression line in Table 9. Topics are plotted in Figure 9 from largest to smallest regression line slope, with regression lines colored orange for positive slopes and blue for negative slopes. Topic label is given in the upper left corner of each plot and the number of abstracts containing the topic (where the topic weight is greater than zero), n, is given in the upper right corner of each plot.

We also show project frequency by start date and funding agency in Figure 10 for projects with start dates between 2010-2019 to remain consistent with the topic trend analysis that is limited to the same range. HHS projects dominate this themed corpus, which can be expected, and the distribution by start year looks very similar to the distribution by start year for the entire corpus (Figure 1). The total number of projects with start years in the pandemics themed corpus between 2010-2019 is 2,598.

We see a notable trend in topic P30 (Zika virus); specifically there is a large increase in the prevalence of work in this topic between 2015 to 2017, which follow the 2015-2016 Zika outbreak in North and South America (Division of Vector-Borne Diseases [DVBD], n.d.). This topic has the largest regression line slope, signaling that it experienced the largest increase in prevalence

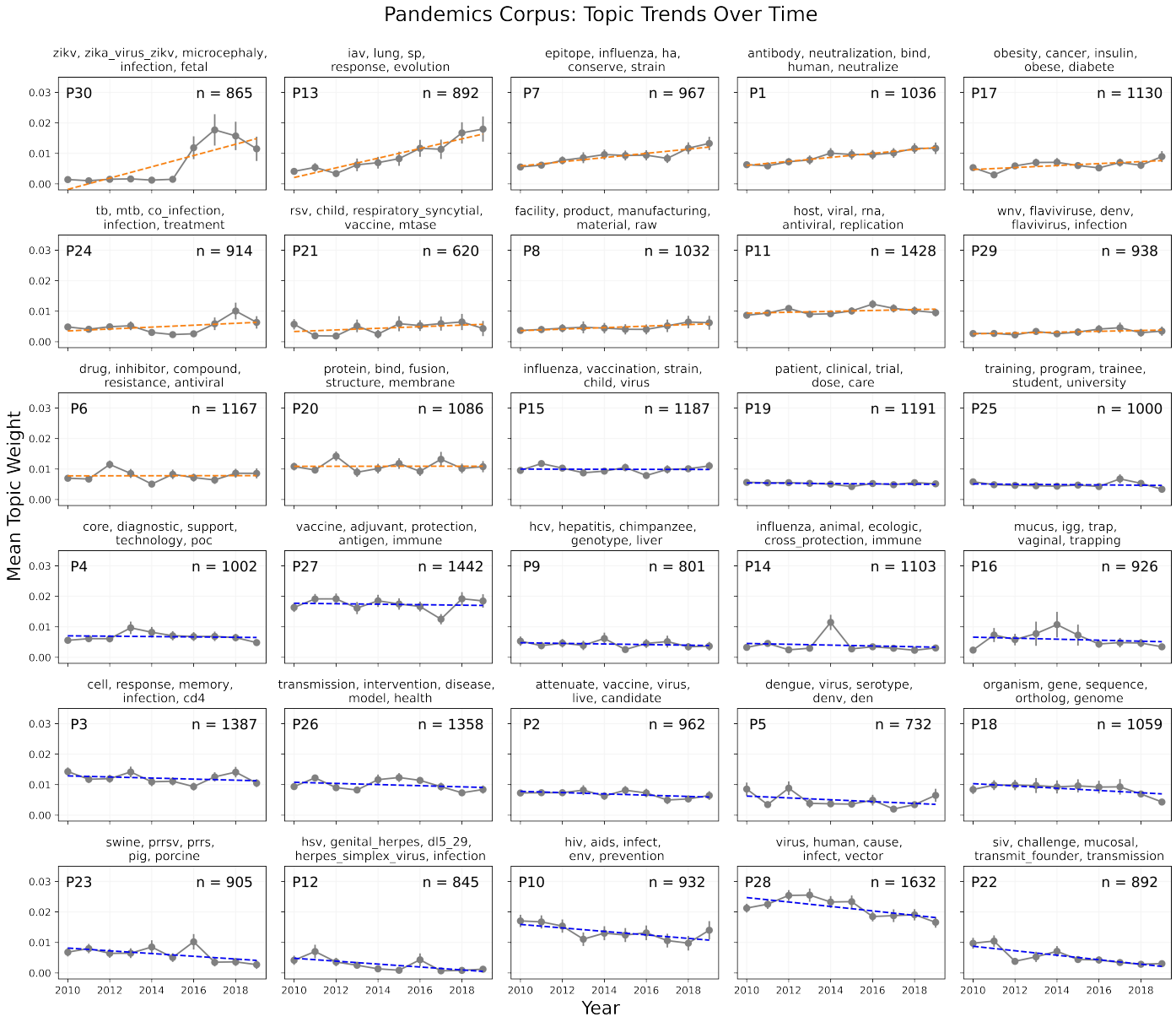

Figure 9: Topic trend results from 2010-2019 for each of 30 topics produced by an NMF model on the pandemics corpus. Topic labels and the number of abstracts containing the topics (where the topic weight is greater than zero), n, are given in the upper left and right plot corners respectively. Plots are ordered from largest to smallest regression line slope; orange lines have a positive slope and blue lines have a negative slope. Standard errors on the means are represented on each plot using error bars.

Label	Slope (x100)	p-Value	Top Five Terms
P1	0.065277	0.000016	antibody, neutralization, bind, human, neutralize
P2	-0.020972	0.080225	attenuate, vaccine, virus, live, candidate
P3	-0.017544	0.37679	cell, response, memory, infection, cd4
P4	-0.006033	0.7124	core, diagnostic, support, technology, poc
P5	-0.030112	0.259389	dengue, virus, serotype, denv, den
P6	0.000825	0.968662	drug, inhibitor, compound, resistance, antiviral
P7	0.069025	0.00039	epitope, influenza, ha, conserve, strain
P8	0.02509	0.005549	facility, product, manufacturing, material, raw
P9	-0.010022	0.42036	hcv, hepatitis, chimpanzee, genotype, liver
P10	-0.057207	0.026361	hiv, aids, infect, env, prevention
P11	0.014838	0.24947	host, viral, rna, antiviral, replication
P12	-0.047957	0.0248	hsv, genital_herpes, dl5_29, herpes_simplex_virus, infection
P13	0.158493	0.000044	iav, lung, sp, response, evolution
P14	-0.013477	0.678932	influenza, animal, ecologic, cross_protection, immune
P15	-0.001413	0.91637	influenza, vaccination, strain, child, virus
P16	-0.0165	0.572417	mucus, igg, trap, vaginal, trapping
P17	0.032532	0.049451	obesity, cancer, insulin, obese, diabete
P18	-0.037193	0.045762	organism, gene, sequence, ortholog, genome
P19	-0.00509	0.293815	patient, clinical, trial, dose, care
P20	0.000431	0.983284	protein, bind, fusion, structure, membrane
P21	0.026301	0.190765	rsv, child, respiratory_syncytial, vaccine, mtase
P22	-0.073417	0.004626	siv, challenge, mucosal, transmit_founder, transmission
P23	-0.044908	0.089521	swine, prrsv, prrs, pig, porcine
P24	0.03131	0.227215	tb, mtb, co_infection, infection, treatment
P25	-0.005374	0.628176	training, program, trainee, student, university
P26	-0.018817	0.375968	transmission, intervention, disease, model, health
P27	-0.00794	0.750543	vaccine, adjuvant, protection, antigen, immune
P28	-0.072886	0.019793	virus, human, cause, infect, vector
P29	0.013982	0.076246	wnv, flaviviruse, denv, flavivirus, infection
P30	0.185024	0.003739	zikv, zika_virus_zikv, microcephaly, infection, fetal

Table 9: Pandemics corpus topic trend line results limited to projects with start years between 2010-2019. Slope refers to the slope of the regression line relating project start year and mean topic weight. Slopes are multiplied by 100 for easier viewing.

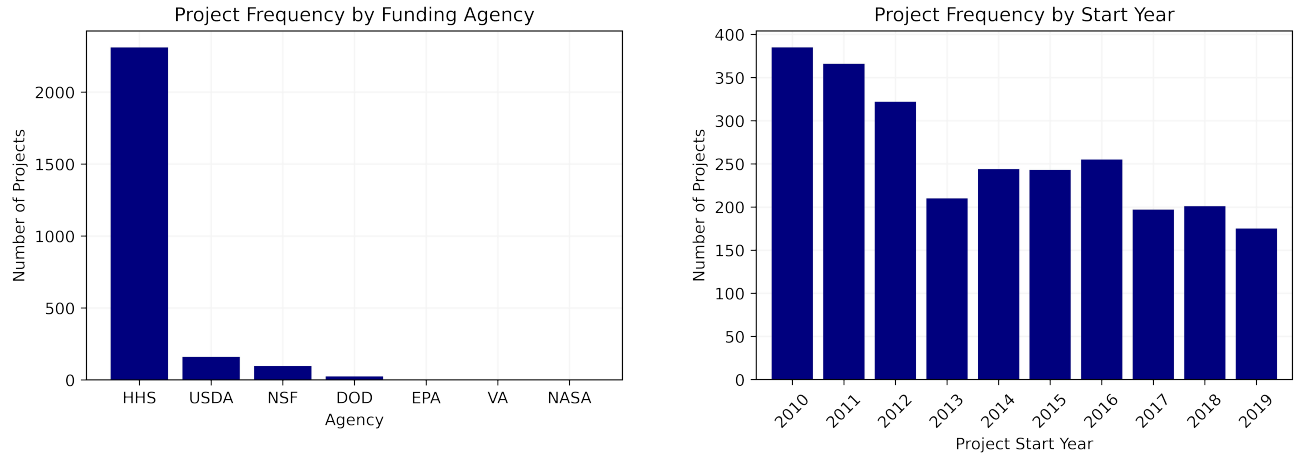

Figure 10: Pandemics corpus projects with start dates between 2010-2019. Distributions by project funding agency and start year.

compared to the other topics over the given time period. Topic P14 (ecological influenza) exhibits a steady trend in most years except for a considerable increase in 2014. This could be connected to the H7N9 (avian influenza, "bird flu") outbreak in China in 2013, in which the first human case of

H7N9 was reported (Poovorawan et al., 2013). While we cannot be certain that these past events caused the trends in P30 and P14, there does seem to be at least a reasonable connection.

Influenza topics P13 and P7 show considerable increases over time, and specifically, topic P13 (influenza A virus, IAV) rises to a high mean topic weight of almost 0.02 by 2019. HIV/AIDS topic P10 shows a considerable decrease although it has a fairly high mean topic weight. Topic P28 (human viruses) experiences a considerable decrease as well; however, it is present in the highest number of abstracts ($n = 1632$) and has a very high mean topic weight in all years compared to other topics, signaling that it is still a well researched topic. Topics with decreasing trends can still be very prevalent research areas. Topic P27 (vaccine adjuvants) technically has a decreasing trend although to the eye, the trend looks fairly steady. However, its mean topic weight is very high compared to other topics and it is included in 1,442 abstracts (second highest value for n), signaling this is also an important topic. As a final note, Topic P25 (student training) is included in 1,000 abstracts but has a fairly steady, low mean topic weight. This could be attributed to many abstracts mentioning student training but not being focused on student training.

6.2 Case Study 2: Coronavirus

Our second case study focused on identify emerging topics in Federal RePORTER within the area of coronavirus. We paralleled our approach from the first case study with the exceptions of using a coronavirus keyword list,[9] including the 500 abstracts with the highest LSI relevance scores to the query (again the relevance score distribution did not provide any clear guidance to choosing this parameter), and using a 25 topic NMF model. The coronavirus focused topic model is also quite stable; for $r = 20$ runs and 10 topic terms, the average DSD, TS, and PS were 0.19, 0.72, and 0.77. Table 10 presents the breakdown of the themed coronavirus corpus by information retrieval method,[10] and Table 11 lists the 25 topics discovered by the NMF model including topic labels and the top five terms for each topic. There are not as many abstracts containing the term 'coronavirus' as there are that contain the term 'pandemics' in our Federal RePORTER corpus. Our coronavirus corpus is smaller than the pandemics corpus which intuitively makes sense as coronavirus is a more specific theme than pandemics.

TM	**LSI**	**TM & LSI**	**Total**
522	481	19	1022

Table 10: Coronavirus themed corpus contribution by information retrieval method. The units for each column are number of abstracts. TM: term matching, TM and LSI: overlap of abstracts returned by both methods.

[9]Coronavirus Keywords: abetacoronavirus, acoronavirus, andcoronaviruse, anycoronavirus, arecoronavirus, ascoronaviruse, bat_coronaviruse, beta_coronavirus, betacoronavirus, betacoronaviruse, carriedcoronavirus, coronavirus, coronavirus_cov, coronavirus_hcov, coronavirus_hcov_emc, coronavirus_mers_cov, coronavirus_nl63, coronavirus_papain, coronavirus_sar, coronavirus_sar_cov, coronavirus_sars, coronavirus_sars_cov, coronavirus_spike, coronavirusand, coronaviruse, coronaviruse_cov, coronaviruse_hcov, coronaviruses, coronavirusesinteract, forcoronavirus, gammacoronavirus, manycoronavirus, mers_coronavirus, neurotropic_coronavirus, ofcoronaviruse, othercoronaviruse, pan_anticoronavirus, pan_coronavirus, peritonitiscoronavirus, respiratorycoronavirus, sar_coronavirus, sar_coronavirus_sar_cov, sars_coronavirus, sars_coronavirus_sars_cov, syndrome_coronavirus_mers, syndromecoronavirus, thatcoronaviruse, thecoronavirus, tractable_sar_coronavirus

[10]If we had used more documents than the 500 with the highest LSI relevance score, the overlap between methods would have been larger. In fact, 97% of the documents marked as relevant by term matching are in the top 9.88% of LSI relevance scores.

Label	Top Five Terms
C1	ace2, epithelial, sar_cov, airway, lung
C2	aged, mouse, response, cell, severe
C3	animal, influenza, bird, surveillance, contact
C4	assembly, virus, capsid, hcv, particle
C5	cns, mhv, cell, cn, type
C6	compound, fidelity, activity, cov, vivo
C7	core, hrv, stock, virus, recombinant
C8	disease, infectious, respiratory, develop, health
C9	entry, cell, gene, virus, cellular
C10	fusion, peptide, protein, dv, membrane
C11	gene, uncharacterize, orf, encode, rp
C12	iav, evolution, ha, transmission, influenza
C13	il, injury, te, lung, inflammation

Label	Top Five Terms
C14	immune, polygenic, trait, regulate, response
C15	influenza, virus, 1918, pandemic, human
C16	inhibitor, structure, protease, crystal, enzyme
C17	mers_cov, mers, dpp4, cov, mouse
C18	novel, virus, identify, thesedisease, gastroenteritis
C19	ns1, ifn, rig, trim25, influenza
C20	receptor, rbd, bind, spike, antibody
C21	replication, rna, protein, viral, host
C22	swine, influenza, iaa, relatedness, quantification
C23	vaccine, attenuate, sars_cov, sar_cov, vector
C24	virus, host, transmission, viral, interaction
C25	zoonotic, emerge, bat_cov, movement, species

Table 11: Top five topic terms from NMF model with 25 topics on themed coronavirus corpus. Topics are listed and labeled in alphabetical order by the most important topic term.

We note two specific coronaviruses that appear in the 25 topics: SARS-CoV (topics C1 and C23) and MERS-CoV (topic C17). The data from 2010-2019 is pre-COVID-19 so this topic is not represented. The zoonotic nature of coronaviruses is given in topic C25, and general respiratory/lung issues are presented in topics C8 and C13. A number of influenza topics appear in the model (C3, C12, C15, C19, and C22) which came about due to the use of LSI. Influenza is a virus, thus related to coronavirus. See Appendix D for further information on using term matching and LSI versus only using term matching to created themed corpora.

Topic trend results over the years 2010-2019 are presented in Figure 11 and Table 12. Coronavirus corpus project frequency by start year and funding agency is given in Figure 12 for projects with start years between 2010-2019. There are 651 projects in this limited coronavirus corpus and 590 (90.6%) are funded by HHS. The distribution by start year is fairly similar to that of the entire Federal RePORTER corpus and the pandemics themed corpus (Figures 1 and 10) with the exceptions of a fairly steady project frequency trend beginning in 2012 rather than 2013, and a slight drop off beginning in 2017.

We mainly focus our discussion of the results on two specific coronaviruses, MERS-CoV and SARS-CoV, as each of these viruses garnered international attention over the last decade and recently with the 2020 COVID-19 outbreak (Li et al., 2020). Topic C17 (MERS-CoV) shows the largest prevalence increase over 2010-2019 with higher mean topic weights beginning in 2014. The first (and only) cases of MERS in the United States occurred in 2014 (Division of Viral Diseases, 2019, August 2), and there were also MERS outbreaks in 2015 (South Korea) and in 2018 (Saudi Arabia) (Li et al., 2020). While we cannot claim that these events caused the topic C17 trend, there could be a connection.

Topic C23 (SARS-CoV vaccines) has a relatively steady prevalence trend with somewhat low mean topic weights. However, it is present in $n = 253$ abstracts in the corpus. Topic C1 (SARS-CoV, specifically the role of the enzyme ACE-2) appears in the lowest number of abstracts ($n = 188$) and experiences a decreasing trend in prevalence over time with somewhat low to low mean topic weights. While these trends contrast those of topic C17 (MERS-CoV) there is still research interest in SARS-CoV, especially around vaccination. These trends could be a result of the SARS epidemic happening in 2003 and that no cases of SARS have been reported after 2004 (Division of Viral Diseases, n.d.); these events fell outside of our analysis window of 2010-2019.

The general topic C24 (viruses) exhibits a fairly steady trend of high mean topic weights and appears in the highest number of abstracts ($n = 377$) signaling its importance in the corpus. Another general topic, C8 (respiratory disease), exhibits a slightly increasing trend of fairly high

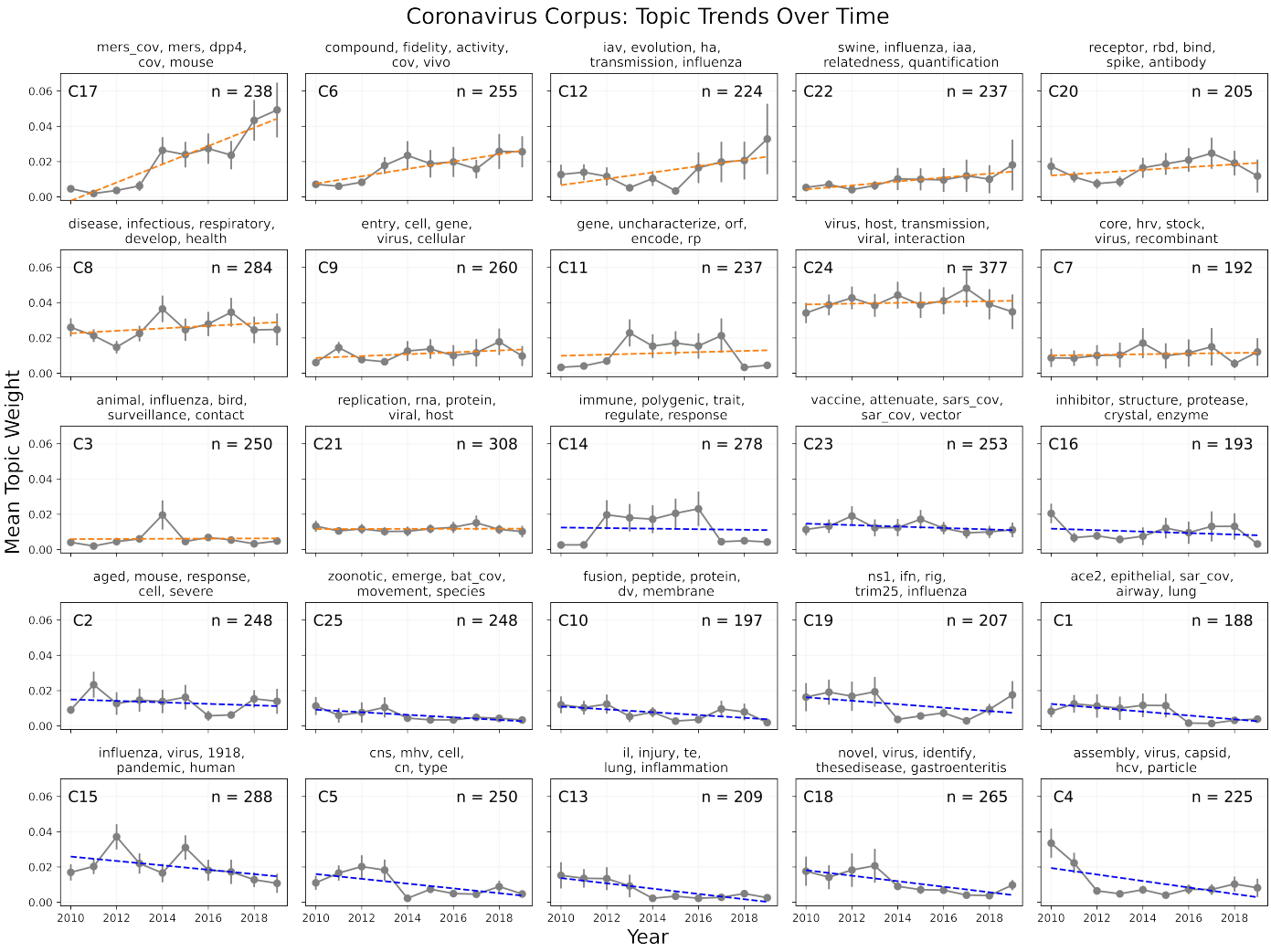

Figure 11: Topic trend results from 2010-2019 for each of 25 topics produced by an NMF model on the coronavirus corpus. Topic labels and the number of abstracts containing the topics (where the topic weight is greater than zero), n, are given in the upper left and right plot corners respectively. Plots are ordered from largest to smallest regression line slope; orange lines have a positive slope and blue lines have a negative slope. Standard errors on the means are represented on each plot using error bars.

Label	Slope (x100)	p-Value	Top Five Terms
C1	-0.109132	0.016467	ace2, epithelial, sar_cov, airway, lung
C2	-0.041694	0.49809	aged, mouse, response, cell, severe
C3	0.004698	0.936858	animal, influenza, bird, surveillance, contact
C4	-0.183775	0.070699	assembly, virus, capsid, hcv, particle
C5	-0.135659	0.045218	cns, mhv, cell, cn, type
C6	0.207738	0.002061	compound, fidelity, activity, cov, vivo
C7	0.018024	0.649416	core, hrv, stock, virus, recombinant
C8	0.069691	0.339927	disease, infectious, respiratory, develop, health
C9	0.052778	0.220847	entry, cell, gene, virus, cellular
C10	-0.079847	0.048466	fusion, peptide, protein, dv, membrane
C11	0.034471	0.712011	gene, uncharacterize, orf, encode, rp
C12	0.179079	0.045725	iav, evolution, ha, transmission, influenza
C13	-0.14988	0.001835	il, injury, te, lung, inflammation
C14	-0.015944	0.877048	immune, polygenic, trait, regulate, response
C15	-0.124103	0.175778	influenza, virus, 1918, pandemic, human
C16	-0.041417	0.48025	inhibitor, structure, protease, crystal, enzyme
C17	0.517138	0.000085	mers_cov, mers, dpp4, cov, mouse
C18	-0.156366	0.00884	novel, virus, identify, thesedisease, gastroenteritis
C19	-0.097763	0.197948	ns1, ifn, rig, trim25, influenza
C20	0.07947	0.218588	receptor, rbd, bind, spike, antibody
C21	0.002583	0.892694	replication, rna, protein, viral, host
C22	0.111803	0.001835	swine, influenza, iaa, relatedness, quantification
C23	-0.041203	0.235605	vaccine, attenuate, sars_cov, sar_cov, vector
C24	0.02318	0.6457	virus, host, transmission, viral, interaction
C25	-0.072686	0.013187	zoonotic, emerge, bat_cov, movement, species

Table 12: Coronavirus corpus topic trend line results limited to projects with start years between 2010-2019. Slope refers to the slope of the regression line relating project start year and mean topic weight. Slopes are multiplied by 100 for easier viewing.

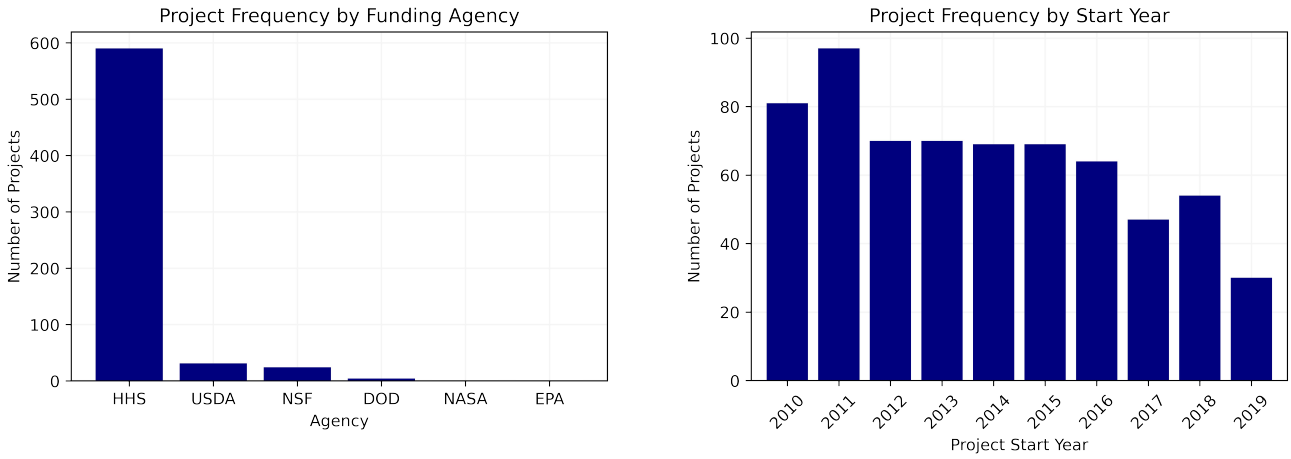

Figure 12: Coronavirus corpus projects with start dates between 2010-2019. Distributions by project funding agency and start year.

mean topic weights while appearing in a large number of abstracts ($n = 284$). Intuitively it makes sense that general topics have more signal in the corpus than more specific topics, for example the SARS-CoV topics C23 and C1. Of note, Topic C3 (animal influenza) exhibits a similar trend to topic P14 (ecological influenza) from the pandemics themed corpus.

7 Conclusion and Future Work

Federally funded R&D topics are identified through the use of NLP and NMF topic modeling using Federal RePORTER project abstracts. In addition, topics related to pandemics and coronavirus are presented, which we found using information retrieval and NMF topic modeling. Topic trends over time are also shown.

Our approach was informed by the Data Science Framework (Keller et al., 2020), a core component which involved reviewing the ethical impact of our work. We did not collect or utilize any individual data, which minimizes the potential harm to individuals. In considering the larger implications of the project, we recognize that our data only included federally funded grants within the United States. It contains no data on federally funded contracts, nor on R&D performed by Federal agencies. Therefore it does not capture the full scope of R&D within the United States nor around the world. We also recognize that implicit bias in research funding may affect the representation of topics within our data and, while not addressed within the scope of this project, could serve as a focus for future analysis.

We plan to continue this work by extending our approach to create themed corpora for themes that are complex, multi-faceted, and difficult to define, such as "artificial intelligence". This could include extending the list of theme keywords or using expert input. Another approach we may utilize is comparing project abstracts to a themed Wikipedia page (for example, the artificial intelligence page) and scoring abstracts for inclusion in the themed corpus based on their similarity to the page. We will also research other existing methods to create themed corpora such as the methods of Eads et al. (2021) and OECD (2019). Performance of these methods will be measured, for example using precision and recall.

For detecting topic trends, we are exploring dynamic topic models as an alternative to the current Griffiths and Steyvers (2004) method. We will test current and new approaches and themes, using abstracts pulled directly from agency award databases, since the Federal RePORTER database will now no longer include data beyond 2020. We will continue to analyze ongoing themes of artificial intelligence, pandemics, and coronavirus funded R&D projects. We will also explore new themes such as the bioeconomy. We believe that the methods described in this technical report show promise to supplement the information currently collected in NCSES surveys by providing information that the surveys do not collect.

8 Acknowledgements

This research was funded by the National Center for Science and Engineering Statistics contract 49100420C0015. We extend our thanks to the students who contributed to this project in the 2020 and 2021 Data Science for the Public Good programs through the Social and Decision Analytics Division (SDAD), Biocomplexity Institute, University of Virginia. These students are (listed in alphabetical order) Martha Czernuszenko, Lara Haase, Elizabeth Miller, Cierra Oliveira, Sean Pietrowicz, Haleigh Tomlin, and Crystal Zang. We also thank Madeline Garrett for her contributions as a student SDAD Project Fellow in Fall 2021 and Spring 2022 and Samantha Cohen, a former SDAD Postdoctoral Research Associate, for contributions to the text cleaning and processing work.

References

Ankam, S., Dou, W., Strumsky, D., Wang, D. X., Rabinowitz, T., & Zadrozny, W. (2012). Exploring emerging technologies using patent data and patent classification. *Proc. IEEE VIS Workshop Interactive Vis. Text Anal.*

Belford, M., Mac Namee, B., & Greene, D. (2018). Stability of topic modeling via matrix factorization. *Expert Systems with Applications*, *91*, 159–169.

Berg, S., Wustmans, M., & Bröring, S. (2019). Identifying first signals of emerging dominance in a technological innovation system: A novel approach based on patents. *Technological Forecasting and Social Change*, *146*, 706–722.

Bishop, C. M., Svensén, M., & Williams, C. K. (1998). GTM: The generative topographic mapping. *Neural computation*, *10*(1), 215–234.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, *3*(null), 993–1022.

Boyack, K. W., Smith, C., & Klavans, R. (2020). A detailed open access model of the PubMed literature. *Scientific Data*, *7*(1), 1–16. https://www.nature.com/articles/s41597-020-00749-y.pdf?origin=ppub

Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements. In E. M. Airoldi, D. Blei, E. A. Erosheva, & S. E. Fienberg (Eds.), *Handbook of Mixed Membership Models and Their Applications* (1st). CRC Press.

Bruce, J. R., de Figueiredo, J. M., & Silverman, B. S. (2019). Public contracting for private innovation: Government capabilities, decision rights, and performance outcomes. *Strategic Management Journal*, *40*(4), 533–555.

Choi, J., Jang, J., Kim, D. H., & Yoon, J. (2019). Identifying interdisciplinary trends of humanities, sociology, science and technology research in Korea using topic modeling and network analysis. *Journal of the Society of Korea Industrial and Systems Engineering*, *42*(1), 74–86.

Craigin, M., Nichols, L., Simon, M., & Watts, S. (2012). Measuring science: Emerging tools for analysis of federal R&D investments. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–4. https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/meet.14504901160

Cunningham, R. M., Ranney, M. L., Goldstick, J. E., Kamat, S. V., Roche, J. S., & Carter, P. M. (2019). Federal funding for research on the leading causes of death among children and adolescents. *Health Affairs*, *38*(10), 1653–1661.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6⟨391::AID-ASI1⟩3.0.CO;2-9

Division of Vector-Borne Diseases. (n.d.). *Zika Virus - Statistics and Maps*. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Emerging and Zoonotic Infectious Diseases (NCEZID). Retrieved June 3, 2021 from https://www.cdc.gov/zika/reporting/index.html.

Division of Viral Diseases. (n.d.). *Severe Acute Respiratory Syndrome (SARS)*. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases. Retrieved June 7, 2021 from https://www.cdc.gov/sars/index.html.

Division of Viral Diseases. (2019, August 2). *Middle East Respiratory Syndrome (MERS) - About MERS*. U.S. Department of Health and Human Services, Centers for Disease Control and

Prevention, National Center for Immunization and Respiratory Diseases (NCIRD). Retrieved June 3, 2021 from https://www.cdc.gov/coronavirus/mers/about/index.html.

Doanvo, A., Qian, X., Ramjee, D., Piontkivska, H., Desai, A., & Majumder, M. (2020). Machine learning maps research needs in COVID-19 literature. *Patterns*, *1*(9), 100123.

Eads, A., Schofield, A., Mahootian, F., Mimno, D., & Wilderom, R. (2021). Separating the wheat from the chaff: A topic and keyword-based procedure for identifying research-relevant text*. *Poetics*, *86*, 101527.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Jacob, R. R., Gacad, A., Padek, M., Colditz, G. A., Emmons, K. M., Kerner, J. F., Chambers, D. A., & Brownson, R. C. (2020). Mentored training and its association with dissemination and implementation research output: a quasi-experimental evaluation. *Implementation Science*, *15*, 1–8.

Jeong, Y., Park, I., & Yoon, B. (2019). Identifying emerging Research and Business Development (R&BD) areas based on topic modeling and visualization with intellectual property right data. *Technological Forecasting and Social Change*, *146*, 655–672.

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing Data Science: A Framework and Case Study [https://hdsr.mitpress.mit.edu/pub/hnptx6lq]. *Harvard Data Science Review*, *2*(1). https://doi.org/10.1162/99608f92.2d83f7f5

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–791. https://doi.org/10.1038/44565

Lee, H., & Kang, P. (2018). Identifying core topics in technology and innovation management studies: A topic model approach. *The Journal of Technology Transfer*, *43*(5), 1291–1317.

Li, Y.-D., Chi, W.-Y., Su, J.-H., Ferrall, L., Hung, C.-F., & Wu, T.-C. (2020). Coronavirus Vaccine Development: from SARS and MERS to COVID-19. *Journal of Biomedical Science*, *27*(1 104). https://jbiomedsci.biomedcentral.com/articles/10.1186/s12929-020-00695-2

Mahapasuthanon, P., & Hoffman, K. (2019, April 6). *Beyond Bibliometrics: Understanding Library Services in Multidisciplinary Research* [Poster presentation], Mason Graduate Interdisciplinary Conference, George Mason University campus, Arlington, VA, United States. http://mars.gmu.edu/handle/1920/11419

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the 2011 conference on empirical methods in natural language processing*, 262–272. https://aclanthology.org/D11-1024.pdf

Mohr, J. W., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. *Poetics*, *41*(6), 545–569. https://doi.org/10.1016/j.poetic.2013.10.001

National Center for Science and Engineering Statistics. (2021). *Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions: Fiscal Year 2019*, NSF 21–333. Alexandria, VA: National Science Foundation, https://ncses.nsf.gov/pubs/nsf21333/.

Newman, D., Noh, Y., Talley, E., Karimi, S., & Baldwin, T. (2010). Evaluating topic models for digital libraries. *Proceedings of the 10th annual joint conference on Digital libraries*, 215–224. https://doi.org/10.1145/1816123.1816156

Office of the Director of National Intelligence. (2011, September 27). *IARPA Launches New Program To Enable The Rapid Discovery Of Emerging Technical Capabilities* [Press release]. https://www.odni.gov/index.php/newsroom/press-releases/press-releases-2011/item/327-iarpa-launches-new-program-to-enable-the-rapid-discovery-of-emerging-technical-capabilities

Organisation for Economic Co-operation and Development. (2015). *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*, The

Measurement of Scientific, Technological and Innovation Activities, Organisation for Economic Co–operation and Development (Publishing, Paris, https://doi.org/10.1787/7b43b038–en.

Organisation for Economic Co-operation and Development. (2019). *Identifying government funding of AI-related R&D projects - An initial exploration based on US NIH and NSF project funding data* (DSTI/STP/NESTI(2019)1), Directorate for Science, Technology and Innovation and Committee for Scientific and Technological Policy, Organisation for Economic Co–operation and Development, Paris.

Organisation for Economic Co-operation and Development. (2021). Working Party of National Experts on Science and Technology Indicators Establishment of an OECD Expert Group on the Management and Analysis of R&D and Innovation Administrative Data (MARIAD), Organisation for Economic Co–operation and Development, Directorate for Science, Technology, and Innovation, Committee for Scientific and Technological Policy.

Park, J. S., Hong, S.-G., & Kim, J.-W. (2017). A study on science technology trend and prediction using topic modeling. *Journal of the Korea Industrial Information Systems Research*, *22*(4), 19–28.

Poovorawan, Y., Pyungporn, S., Prachayangprecha, S., & Makkoch, J. (2013). Global Alert to Avian Influenza Virus Infection: From H5N1 to H7N9. *Pathogens and Global Health*, *107*(5), 217–223. https://doi.org/10.1179/2047773213Y.0000000103

Porter, A. L., Garner, J., Carley, S. F., & Newman, N. C. (2019). Emergence scoring to identify frontier R&D topics and key players. *Technological Forecasting and Social Change*, *146*, 628–643.

Powell, K. (2019). Searching by grant number: comparison of funding acknowledgments in NIH RePORTER, PubMed, and Web of Science. *Journal of the Medical Library Association: JMLA*, *107*(2), 172.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Rosen, D., Engel, R. J., Beaugard, C., Davis, N., & Cochran, G. (2019). Baby boomer's substance abuse and researcher indifference. *Journal of gerontological social work*, *62*(1), 16–28.

Schofield, A., Magnusson, M., Thompson, L., & Mimno, D. (2017). Understanding Text Pre-Processing for Latent Dirichlet Allocation. *ACL Workshop for Women in NLP (WiNLP)*.

Suominen, A., Peng, H., & Ranaei, S. (2019). Examining the dynamics of an emerging research network using the case of triboelectric nanogenerators. *Technological Forecasting and Social Change*, *146*, 820–830.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, *67*(10), 2464–2476.

Talley, E. M., Newman, D., Mimno, D., Herr, B. W., Wallach, H. M., Burns, G. A., Leenders, A. M., & McCallum, A. (2011). Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods*, *8*(6), 443–444. doi:10.1038/nmeth.1619

U.S. Department of Health and Human Services. (2020, March 6). *STAR METRICS Federal RePORTER FAQS*. Retrieved September 19, 2021 from https://federalreporter.nih.gov/Home/FAQ.

U.S. Government Accountability Office. (2021). *Federal Spending Transparency: Opportunities Exist to Further Improve the Information Available on USAspending.gov* (GAO-22-104702). https://www.gao.gov/assets/gao-22-104702.pdf

Wang, J., Fan, Y., Feng, L., Ye, Z., & Zhang, H. (2019). Research Hotspot Prediction and Regular Evolutionary Pattern Identification Based on NSFC Grants Using NMF and Semantic Retrieval. *IEEE Access*, *7*, 123776–123787.

Weber-Main, A. M., McGee, R., Eide Boman, K., Hemming, J., Hall, M., Unold, T., Harwood, E. M., Risner, L. E., Smith, A., Lawson, K., et al. (2020). Grant application outcomes for biomedical researchers who participated in the National Research Mentoring Network's Grant Writing Coaching Programs. *PloS one*, *15*(11), e0241851.

Winnink, J., Tijssen, R. J., & Van Raan, A. (2019). Searching for new breakthroughs in science: How effective are computerised detection algorithms? *Technological Forecasting and Social Change*, *146*, 673–686.

Yamashita, I., Murakami, A., Cairns, S., & Galindo-Rueda, F. (2021). Measuring the AI content of government-funded R&D projects: A proof of concept for the OECD Fundstat initiative. *OECD Science, Technology and Industry Working Papers*, (No. 2021/09), Organisation for Economic Co–operation and Development, Paris, https://doi.org/10.1787/7b43b038–en.

Zeng, T., & Acuna, D. E. (2020). GotFunding: A grant recommendation system based on scientific articles. *Proceedings of the Association for Information Science and Technology*, *57*(1), e323.

Zhang, Y., Huang, Y., Porter, A. L., Zhang, G., & Lu, J. (2019). Discovering and forecasting interactions in big data research: A learning-enhanced bibliometric study. *Technological Forecasting and Social Change*, *146*, 795–807.

Zhang, Y., Porter, A., Chiavetta, D., Newman, N. C., Guo, Y., et al. (2019). Forecasting technical emergence: An introduction. *Technological Forecasting and Social Change*, *146*(100), 626–627.

Zhou, X., Huang, L., Porter, A., & Vicente-Gomila, J. M. (2019). Tracing the system transformations and innovation pathways of an emerging technology: Solid lipid nanoparticles. *Technological Forecasting and Social Change*, *146*, 785–794.

Appendices

Appendix A NMF 150- and 200-Topic Models

Label	Top Five Terms
Y1	activity, physical, outreach, coordinate, active
Y2	ad, alzheimer, tau, dementia, pathology
Y3	administrative, meeting, scientific, coordinate, communication
Y4	adolescent, youth, family, substance, parent
Y5	aging, age, older, adult, related
Y6	aim, test, hypothesis, determine, propose
Y7	alcohol, drinking, consumption, alcoholism, abuse
Y8	analysis, statistical, design, bioinformatic, biostatistic
Y9	animal, rat, colony, small, veterinary
Y10	antibody, epitope, bind, antigen, monoclonal_antibody
Y11	antigen, dc, treg, tolerance, cd4
Y12	asd, autism, autism_spectrum, disorder, developmental
Y13	asthma, airway, allergic, asthmatic, allergen
Y14	autophagy, p_53, apoptosis, pathway, death
Y15	behavior, behavioral, neural, circuit, reward
Y16	bone, fracture, osteoporosis, osteoblast, skeletal
Y17	brain, neural, mri, region, fmri
Y18	breast, cancer, er, metastasis, estrogen
Y19	cancer, nci, pancreatic, ovarian, member
Y20	care, quality, provider, medical, practice
Y21	cell, differentiation, type, culture, line
Y22	center, director, resource, leadership, member
Y23	channel, ca2, calcium, ion_channel, release
Y24	chemical, chemistry, reaction, synthesis, catalyst
Y25	child, parent, pediatric, childhood, family
Y26	climate, change, ocean, global, ecosystem
Y27	clinical, translational, basic, medicine, medical
Y28	cocaine, addiction, abuse, self, relapse
Y29	cognitive, impairment, function, decline, cognition
Y30	community, partnership, outreach, partner, engagement
Y31	compound, lead, synthesis, library, natural
Y32	conference, meeting, researcher, hold, field
Y33	core, provide, personnel, ppg, ensure
Y34	data, database, collection, collect, set
Y35	diabete, insulin, glucose, type, islet
Y36	disease, infectious, cause, progression, alzheimer
Y37	disorder, depression, anxiety, ptsd, symptom
Y38	dna, repair, damage, methylation, replication
Y39	dr, director, career, mentor, award
Y40	drug, abuse, delivery, addiction, discovery
Y41	effect, determine, increase, level, dose
Y42	effort, molecule, general, small, medicinal_chemistry_campaign
Y43	energy, power, solar, efficiency, fuel
Y44	engineering, design, education, engineer, mechanical
Y45	ethanol, withdrawal, induce, chronic, consumption
Y46	exposure, environmental, chemical, arsenic, expose
Y47	facility, laboratory, equipment, biology, space
Y48	food, safety, consumer, intake, agriculture
Y49	gene, expression, transcription, identify, promoter
Y50	genetic, variant, genome, variation, association

Label	Top Five Terms
Y51	hcv, hepatitis_c, chronic, antiviral, infect
Y52	health, disparity, public, population, mental
Y53	hearing, auditory, speech, sound, loss
Y54	heart, cardiac, failure, cardiomyocyte, hf
Y55	hiv, aids, infect, prevention, antiretroviral
Y56	host, pathogen, bacterial, bacteria, virulence
Y57	hpv, cervical, type, oral, vaccination
Y58	hsc, transplantation, transplant, donor, gvhd
Y59	human, subject, robot, primate, derive
Y60	il, cytokine, th2, nk, th17
Y61	imaging, image, mri, resolution, mr
Y62	immune, response, innate, immunity, adaptive
Y63	infant, pregnancy, maternal, fetal, mother
Y64	infection, infect, cause, acute, chronic
Y65	inflammation, macrophage, inflammatory, activation, monocyte
Y66	information, provide, decision, database, communication
Y67	injury, repair, radiation, damage, regeneration
Y68	instrument, user, instrumentation, capability, ms
Y69	intervention, randomize, control, base, outcome
Y70	investigator, principal, junior, expertise, provide
Y71	iron, heme, metal, deficiency, transport
Y72	kidney, renal, ckd, hypertension, chronic
Y73	language, linguistic, word, speech, speaker
Y74	learning, learn, course, skill, practice
Y75	liver, hepatic, hepatocyte, hcc, fibrosis
Y76	lung, pulmonary, copd, airway, fibrosis
Y77	malaria, parasite, vector, transmission, control
Y78	management, resource, self, manage, forest
Y79	material, property, polymer, mechanical, composite
Y80	mechanism, regulate, role, function, molecular
Y81	melanoma, skin, melanocyte, uv, cutaneous
Y82	membrane, lipid, fusion, transport, plasma
Y83	memory, hippocampus, hippocampal, cd8_t, long
Y84	metabolic, metabolism, lipid, enzyme, glucose
Y85	method, develop, new, apply, approach
Y86	mitochondrial, mitochondria, ros, mt, dysfunction
Y87	model, modeling, develop, prediction, simulation
Y88	mouse, transgenic, strain, knockout, line
Y89	muscle, skeletal_muscle, exercise, fiber, motor
Y90	mutation, mutant, cause, defect, gene
Y91	network, wireless, node, communication, connectivity
Y92	neuron, neuronal, circuit, synaptic, motor
Y93	nicotine, tobacco, smoking, smoker, smoking_cessation
Y94	obesity, weight, obese, diet, fat
Y95	pain, chronic, opioid, analgesic, neuropathic_pain
Y96	particle, flow, physic, quantum, field
Y97	patient, therapy, outcome, surgery, improve
Y98	pd, parkinson, motor, lrrk2, da
Y99	peptide, ms, mhc, bind, sequence
Y100	pilot, fund, grant, funding, faculty

Label	Top Five Terms
Y101	plant, crop, root, growth, seed
Y102	policy, decision, economic, impact, survey
Y103	product, contract, testing, infectious, development
Y104	production, crop, agricultural, farm, pest
Y105	program, member, department, phd, leader
Y106	prostate, cancer, ar, pca, androgen
Y107	protein, interaction, bind, proteomic, folding
Y108	protocol, review, committee, monitoring, scientific
Y109	receptor, ligand, bind, gpcr, agonist
Y110	resistance, antibiotic, insulin, resistant, mechanism
Y111	risk, factor, cvd, cohort, population
Y112	rna, mirna, translation, micro, expression
Y113	sample, biomarker, blood, specimen, collect
Y114	science, scientific, scientist, career, discipline
Y115	screen, assay, throughput, crc, high
Y116	service, resource, provide, share, access
Y117	signal, pathway, activation, kinase, wnt
Y118	sleep, circadian, insomnia, sleep_disturbance, osa
Y119	social, individual, people, interaction, media
Y120	software, computational, algorithm, tool, computer
Y121	soil, carbon, forest, ecosystem, nutrient
Y122	species, evolutionary, evolution, population, diversity
Y123	spore, translational, developmental, drp, career
Y124	stem, college, differentiation, progenitor, mathematics
Y125	stress, er, response, stressor, induce

Label	Top Five Terms
Y126	stroke, motor, recovery, rehabilitation, ischemic_stroke
Y127	structure, structural, complex, crystal, domain
Y128	student, undergraduate, graduate, college, experience
Y129	support, provide, infrastructure, development, national
Y130	system, control, complex, nervous, component
Y131	target, inhibitor, agent, therapeutic, therapy
Y132	tb, mtb, tuberculosis, tuberculosis_tb, m_tuberculosis
Y133	tbi, traumatic, injury, ptsd, outcome
Y134	teacher, school, mathematics, classroom, high
Y135	technology, device, sensor, cost, phase
Y136	theory, problem, mathematical, mathematics, pi
Y137	tissue, specimen, pathology, adipose, organ
Y138	training, trainee, faculty, career, mentor
Y139	treatment, therapy, outcome, efficacy, effective
Y140	trial, phase, clinical, ii, conduct
Y141	tumor, metastasis, growth, metastatic, microenvironment
Y142	university, state, institution, faculty, education
Y143	vaccine, vaccination, protection, adjuvant, protective
Y144	vascular, endothelial, blood, vessel, flow
Y145	viral, replication, vector, hsv, host
Y146	virus, influenza, host, antiviral, infect
Y147	visual, vision, eye, retinal, retina
Y148	water, surface, quality, irrigation, watershed
Y149	woman, ovarian, man, female, reproductive
Y150	workshop, participant, researcher, hold, international

Table 13: Top five topic terms from NMF model with 150 topics on processed Federal RePORTER project abstracts reported in FY 2008-2019. Topics are listed and labeled in alphabetical order by the most important topic term.

Label	Top Five Terms
Z1	activity, physical, outreach, increase, coordinate
Z2	ad, alzheimer, tau, dementia, pathology
Z3	administrative, scientific, coordinate, communication, oversight
Z4	adolescent, youth, family, substance, parent
Z5	adult, older, age, young, life
Z6	aging, age, related, lifespan, decline
Z7	aim, test, hypothesis, determine, propose
Z8	alcohol, drinking, consumption, alcoholism, abuse
Z9	analysis, statistical, design, bioinformatic, biostatistic
Z10	animal, rat, colony, veterinary, laboratory
Z11	antibody, epitope, antigen, monoclonal_antibody, mab
Z12	ar, pca, androgen, crpc, progression
Z13	asd, autism, autism_spectrum, disorder, developmental
Z14	asthma, airway, asthmatic, allergic, allergen
Z15	autophagy, pathway, death, autophagic, degradation
Z16	behavior, behavioral, sexual, self, change
Z17	bind, domain, interaction, ligand, molecule
Z18	biology, molecular, biological, cellular, genetics
Z19	biomarker, detection, early, marker, assay
Z20	blood, pressure, platelet, hypertension, vessel
Z21	bone, fracture, osteoporosis, osteoblast, skeletal
Z22	brain, region, neuronal, connectivity, fmri
Z23	breast, cancer, er, metastasis, estrogen
Z24	cancer, nci, pancreatic, ovarian, member
Z25	care, quality, provider, medical, hospital
Z26	career, mentor, development, mentoring, skill
Z27	cell, differentiation, type, culture, line
Z28	center, excellence, mission, medical, member
Z29	channel, ca2, calcium, ion_channel, release
Z30	chemical, chemistry, reaction, synthesis, catalyst
Z31	child, parent, pediatric, family, childhood
Z32	climate, change, global, ecosystem, variability
Z33	clinical, translational, basic, medical, medicine
Z34	cocaine, addiction, self, dopamine, relapse
Z35	cognitive, impairment, decline, function, cognition
Z36	community, outreach, partner, partnership, engagement
Z37	compound, assay, library, synthesis, chemical
Z38	computational, simulation, modeling, experimental, dynamics
Z39	conference, researcher, international, hold, field
Z40	contract, testing, development, infectious, candidate
Z41	control, quality, controls, pest, robot
Z42	core, provide, personnel, ppg, ensure
Z43	crop, pest, agricultural, farmer, grower
Z44	data, database, collection, collect, set
Z45	decision, process, choice, decision_making, task
Z46	device, sensor, power, design, wireless
Z47	diabete, type, diabetic, complication, diabetes
Z48	director, leadership, senior, associate, phd
Z49	disease, cause, alzheimer, infectious, progression
Z50	disorder, depression, anxiety, schizophrenia, mental
Z51	disparity, minority, health, partnership, education
Z52	dna, repair, damage, replication, methylation
Z53	dr, drs, expert, award, mentor
Z54	drug, abuse, delivery, discovery, addiction
Z55	east_asia_summer, location_east_asia, foreign_counterpart, us, pacific
Z56	effect, determine, level, examine, increase
Z57	energy, power, solar, efficiency, storage
Z58	engineering, design, education, engineer, mechanical
Z59	environmental, environment, factor, toxicology, chemical
Z60	ethanol, withdrawal, induce, chronic, consumption
Z61	exercise, physical, rehabilitation, week, improve
Z62	exposure, arsenic, expose, chemical, prenatal
Z63	facility, equipment, laboratory, space, staff
Z64	faculty, junior, biomedical, institution, member
Z65	flow, fluid, dynamics, transport, numerical
Z66	food, safety, consumer, agriculture, intake
Z67	forest, ecosystem, tree, land, fire
Z68	function, mechanism, role, regulate, regulation
Z69	gene, expression, identify, vector, candidate
Z70	genetic, trait, variation, genetics, phenotype
Z71	genome, sequence, sequencing, genomic, genomics
Z72	group, work, underrepresented, member, focus
Z73	hcv, hepatitis_c, chronic, viral, antiviral
Z74	health, public, mental, relevance, national
Z75	hearing, auditory, speech, sound, noise
Z76	heart, cardiac, failure, cardiomyocyte, hf
Z77	hiv, aids, infect, antiretroviral, prevention
Z78	host, pathogen, bacterial, bacteria, virulence
Z79	hpv, cervical, type, oral, vaccination
Z80	human, subject, robot, primate, derive
Z81	il, cytokine, cd4, th2, nk
Z82	image, resolution, 3d, processing, reconstruction
Z83	imaging, optical, pet, resolution, vivo
Z84	immune, response, innate, immunity, antigen
Z85	infant, mother, maternal, early, neonatal
Z86	infection, infect, viral, cause, siv
Z87	inflammation, inflammatory, intestinal, chronic, induce
Z88	information, provide, database, communication, web
Z89	inhibitor, target, agent, therapeutic, therapy
Z90	injury, repair, regeneration, acute, damage
Z91	instrument, user, instrumentation, laser, capability
Z92	insulin, glucose, metabolic, secretion, beta
Z93	intervention, base, randomize, outcome, prevention
Z94	investigator, principal, junior, provide, cobre
Z95	iron, heme, metal, deficiency, transport
Z96	kidney, renal, ckd, chronic, aki
Z97	language, linguistic, word, speech, speaker
Z98	learning, learn, course, skill, practice
Z99	liver, hepatic, hepatocyte, hcc, fibrosis
Z100	lung, pulmonary, copd, airway, fibrosis

Label	Top Five Terms
Z101	macrophage, monocyte, atherosclerosis, lipid, cholesterol
Z102	malaria, parasite, vector, transmission, mosquito
Z103	management, self, manage, practice, coordination
Z104	material, property, polymer, mechanical, composite
Z105	meeting, annual, scientist, scientific, symposium
Z106	melanoma, metastatic, braf, melanocyte, metastasis
Z107	membrane, lipid, fusion, plasma, transport
Z108	memory, hippocampus, cd8_t, hippocampal, long
Z109	method, develop, apply, approach, measurement
Z110	mirna, expression, mi, target, microrna_mi
Z111	mitochondrial, mitochondria, ros, mt, dysfunction
Z112	model, develop, modeling, prediction, parameter
Z113	motor, movement, als, microtubule, transport
Z114	mouse, transgenic, strain, knockout, line
Z115	mri, mr, functional, magnetic_resonance, measure
Z116	ms, proteomic, mass_spectrometry, eae, metabolite
Z117	muscle, skeletal_muscle, fiber, mass, muscular_dystrophy
Z118	mutation, mutant, cause, defect, somatic
Z119	network, wireless, node, communication, connectivity
Z120	neural, circuit, sensory, cortical, neuroscience
Z121	neuron, neuronal, synaptic, synapse, axon
Z122	new, development, develop, discovery, approach
Z123	nicotine, tobacco, smoking, smoker, smoking_cessation
Z124	obesity, metabolic, fat, obese, diet
Z125	ocean, earth, temperature, marine, measurement
Z126	opioid, abuse, addiction, morphine, dependence
Z127	p_53, apoptosis, induce, damage, mdm2
Z128	pain, chronic, neuropathic_pain, analgesic, sensory
Z129	particle, nanoparticle, size, physic, delivery
Z130	patient, therapy, outcome, surgery, improve
Z131	pd, parkinson, lrrk2, da, dopamine
Z132	peptide, mhc, antigen, tcr, epitope
Z133	phase, ii, iii, sbir, prototype
Z134	pi, co, work, propose, undergraduate
Z135	pilot, fund, grant, funding, developmental
Z136	plant, growth, root, crop, arabidopsis
Z137	policy, economic, state, practice, impact
Z138	population, individual, evaluate, general, consortium
Z139	pregnancy, maternal, fetal, growth, placental
Z140	problem, algorithm, solution, optimization, application
Z141	product, natural, manufacturing, evaluate, market
Z142	production, increase, cost, produce, quality
Z143	program, member, department, theme, evaluation
Z144	prostate, cancer, man, androgen, psa
Z145	protein, folding, proteomic, complex, interaction
Z146	protocol, review, committee, monitoring, scientific
Z147	ptsd, symptom, trauma, veterans, veteran
Z148	quantum, physic, electron, state, spin
Z149	radiation, dose, therapy, irradiation, damage
Z150	receptor, ligand, gpcr, agonist, nmda
Z151	resistance, antibiotic, resistant, antimicrobial, mechanism
Z152	resource, share, access, expertise, provide
Z153	risk, factor, cvd, high, cardiovascular
Z154	rna, translation, micro, splicing, mrna
Z155	sample, specimen, assay, collection, collect
Z156	science, scientific, scientist, computer, discipline
Z157	screen, throughput, assay, molecule, high
Z158	service, provide, member, access, consultation
Z159	signal, pathway, activation, kinase, wnt
Z160	site, reu, enzyme, active, substrate
Z161	skin, wound, cutaneous, keratinocyty, epidermal
Z162	sleep, circadian, insomnia, sleep_disturbance, osa
Z163	social, people, media, individual, relationship
Z164	software, tool, user, computing, computer
Z165	soil, carbon, nutrient, microbial, ecosystem
Z166	species, evolutionary, evolution, diversity, ecological
Z167	spore, translational, developmental, drp, biostatistic
Z168	star, galaxy, formation, mass, gas
Z169	stem, hsc, hematopoietic, progenitor, college
Z170	stress, er, response, stressor, induce
Z171	stroke, recovery, ischemic_stroke, rehabilitation, acute
Z172	structure, structural, complex, crystal, resolution
Z173	student, undergraduate, graduate, college, summer
Z174	support, provide, infrastructure, continue, request
Z175	surface, biofilm, adhesion, coating, formation
Z176	system, complex, nervous, component, integrate
Z177	tb, mtb, tuberculosis, tuberculosis_tb, m_tuberculosis
Z178	tbi, traumatic, injury, outcome, blast
Z179	teacher, school, mathematics, classroom, teaching
Z180	technology, platform, develop, enable, industry
Z181	theory, mathematics, mathematical, space, geometric
Z182	tissue, pathology, specimen, adipose, organ
Z183	tolerance, gvhd, transplantation, treg, transplant
Z184	trainee, year, phd, mentor, postdoctoral
Z185	training, train, fellow, graduate, skill
Z186	transcription, chromatin, factor, epigenetic, transcriptional
Z187	treatment, therapy, effective, efficacy, outcome
Z188	trial, clinical, conduct, randomize, design
Z189	tumor, growth, metastasis, microenvironment, metastatic
Z190	university, state, institution, college, education
Z191	vaccine, vaccination, protection, adjuvant, immunity
Z192	variant, association, snp, gwas, identify
Z193	vascular, endothelial, vessel, angiogenesis, vegf
Z194	virus, viral, influenza, replication, host
Z195	visual, vision, eye, retinal, retina
Z196	water, quality, irrigation, watershed, groundwater
Z197	weight, loss, body, gain, diet
Z198	woman, ovarian, man, female, reproductive
Z199	workshop, researcher, participant, international, hold
Z200	year, subject, participant, measure, follow

Table 14: Top five topic terms from NMF model with 200 topics on processed Federal RePORTER project abstracts reported in FY 2008-2019. Topics are listed and labeled in alphabetical order by the most important topic term.

Appendix B Federal RePORTER - Complete Topic Trend Results

B.1 50-Topic Model, Trends Calculated Over 2010-2019

Topic Words	n	Slope	SE	p-value
ad, alzheimer, tau, dementia, pathology	132037	0.000055	0.000012	0.002233
administrative, core, scientific, meeting, coordinate	182010	0.000006	0.000009	0.537303
aging, cognitive, age, memory, older	191925	0.000053	0.000018	0.017458
alcohol, ethanol, drinking, consumption, abuse	143978	0.000006	0.000009	0.50091
bone, tissue, fracture, osteoporosis, osteoblast	154500	0.000003	0.000005	0.506452
brain, tbi, injury, neural, mri	158273	0.000025	0.00001	0.033721
breast, cancer, woman, er, estrogen	111249	-0.000021	0.000008	0.027553
cancer, ovarian, nci, pancreatic, member	181024	0.000013	0.000011	0.294593
cell, stem, differentiation, tissue, progenitor	223988	0.000002	0.000006	0.802127
center, resource, support, investigator, facility	213845	-0.000014	0.000019	0.475676
child, parent, language, family, childhood	157749	0.000006	0.000007	0.403363
clinical, trial, protocol, translational, phase	209902	0.000032	0.000008	0.003528
conference, meeting, workshop, researcher, international	197604	-0.000027	0.000022	0.260734
core, investigator, provide, service, analysis	183098	-0.000011	0.000013	0.443685
data, analysis, statistical, database, management	263023	0.000021	0.00001	0.080438
disease, human, kidney, infectious, pd	252970	0.000031	0.000031	0.340591
dna, repair, damage, replication, genome	169347	0	0.000006	0.952566
dr, career, mentor, award, director	190018	0.000033	0.000017	0.095661
drug, compound, screen, target, inhibitor	215563	0.000013	0.000006	0.064165
gene, expression, genetic, genome, identify	209181	-0.000025	0.000011	0.062219
health, community, disparity, care, public	215510	-0.000019	0.000013	0.195198
heart, cardiac, vascular, injury, mitochondrial	205008	0.000029	0.000025	0.282512
hiv, aids, infect, infection, antiretroviral	133543	0.000011	0.000009	0.239594
imaging, image, mri, resolution, tissue	184239	-0.00001	0.000007	0.207642
immune, response, il, cytokine, inflammation	168874	0.000017	0.000019	0.379898
infection, host, pathogen, bacterial, antibiotic	182731	0.000004	0.000004	0.3069
insulin, diabete, obesity, glucose, metabolic	156353	-0.000004	0.000012	0.752676
intervention, behavior, treatment, social, behavioral	207667	0.000062	0.000021	0.019869
lung, airway, pulmonary, asthma, injury	127013	0.000018	0.000014	0.222027
material, chemical, property, chemistry, energy	204167	-0.000027	0.000023	0.278029
model, theory, problem, method, computational	248527	-0.000031	0.000057	0.599787
mouse, model, animal, transgenic, human	220635	-0.000004	0.000013	0.757703
network, social, wireless, communication, node	178263	0.000022	0.000014	0.143424
neuron, circuit, neural, neuronal, motor	198848	0.000042	0.000015	0.024665
pain, chronic, opioid, treatment, analgesic	147195	0.000047	0.000018	0.028489
patient, care, treatment, outcome, therapy	221798	0.000049	0.000019	0.032914
plant, food, crop, production, soil	170975	-0.000089	0.000035	0.033758
program, member, funding, support, grant	249377	-0.000007	0.00001	0.516806
prostate, cancer, ar, pca, androgen	137217	-0.000017	0.000007	0.038355
protein, membrane, structure, bind, complex	218193	-0.000042	0.000011	0.005952
risk, exposure, factor, woman, environmental	227802	0.000023	0.000018	0.237313
rna, mirna, expression, translation, micro	181797	0.000025	0.000012	0.077899
signal, receptor, pathway, regulate, activation	242362	-0.000006	0.000017	0.715707
student, science, stem, school, undergraduate	176751	-0.000027	0.000049	0.598052
system, technology, device, design, develop	262371	0.000002	0.00003	0.960564
training, trainee, faculty, career, mentor	173247	0.000027	0.000007	0.004955
tumor, therapy, target, metastasis, growth	177011	0.000021	0.000012	0.13207
vaccine, antibody, antigen, vaccination, protection	163875	-0.000001	0.000004	0.689183
virus, viral, infection, hcv, influenza	117168	-0.000017	0.000004	0.003164
water, climate, change, ecosystem, forest	213717	-0.000054	0.000026	0.075666

Table 15: Full corpus topic trend results.

B.2 50-Topic Model, Trends Calculated Over 2010-2018

We ran the same trend analysis from Section 5.1 limited to the years 2010 through 2018. Results are presented in Figures 13-16. From Figure 13 we see that while there are a few topics from the 2010-2019 analysis still present (FR1, FR3, FR34, FR28), many of the ten topics with largest positive slopes are different and tend to be topics that would be funded by NSF. Topics related to computational models (FR31), systems and device design (FR45), social networks (FR33), and statistical analysis (FR15) are present as well as some other broad topics. Interestingly, we see that one of the topics with the ten largest slopes from 2010 to 2018 (FR31, computational models) has one of the largest negative slopes from 2010 to 2019 due to the weight in 2019 being much lower. Many of the topics with the largest negative slopes through 2018 are similar to those through 2019, including FR40, FR37, FR20, FR7, and FR21. Other topics include signal transduction (FR43), prostate cancer (FR39), diabetes (FR27), mouse models (FR32), and influenza virus (FR49).

Similar to Figure 8 in Section 5.1, from Figure 15 we see that topics on clinical trials (FR12), ovarian cancer (FR8), environmental risk factors (FR41), and kidney disease (FR16) all maintain relatively high weights, indicating that these topics are present across many projects.

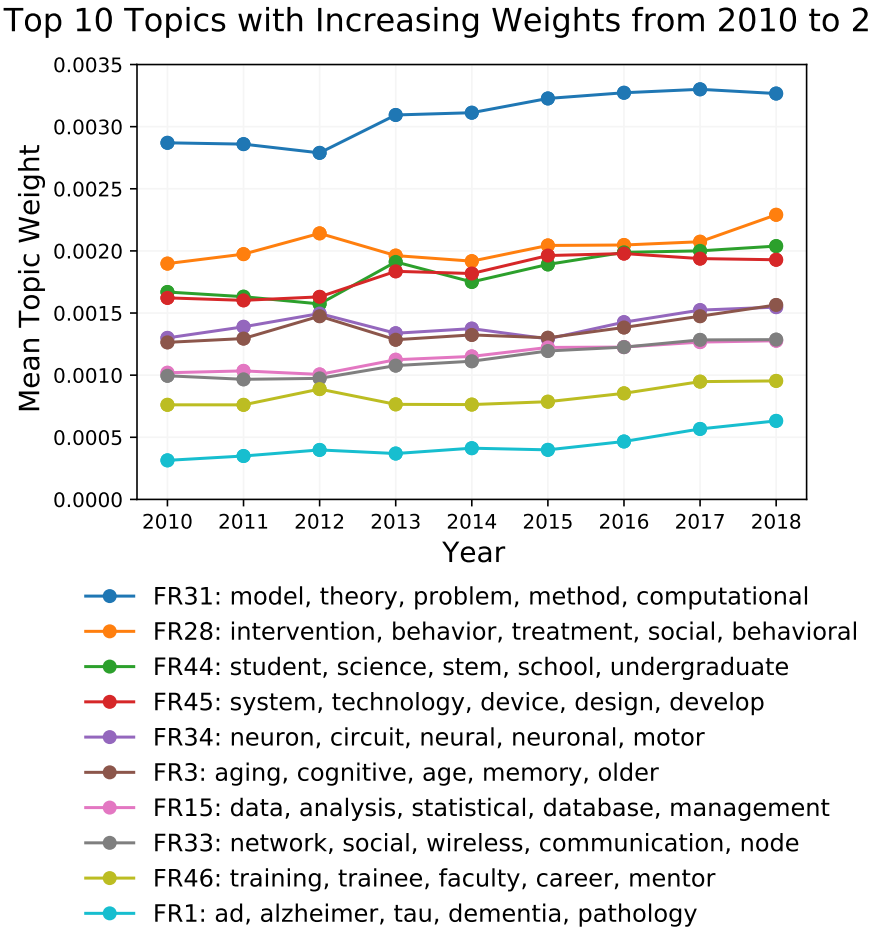

Figure 13: Ten topics with largest positive regression line slopes from the 50 topic model. The slopes are calculated using the weights from 2010 through 2018.

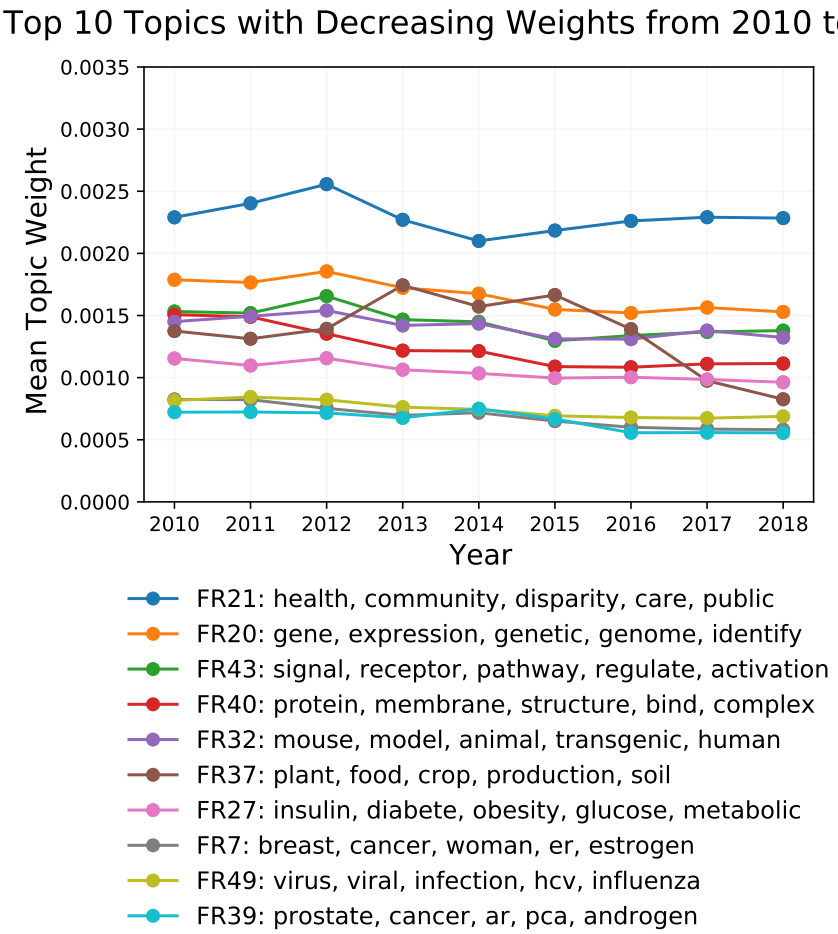

Figure 14: Ten topics with largest negative regression line slopes from the 50 topic model. The slopes are calculated using the weights from 2010 through 2018.

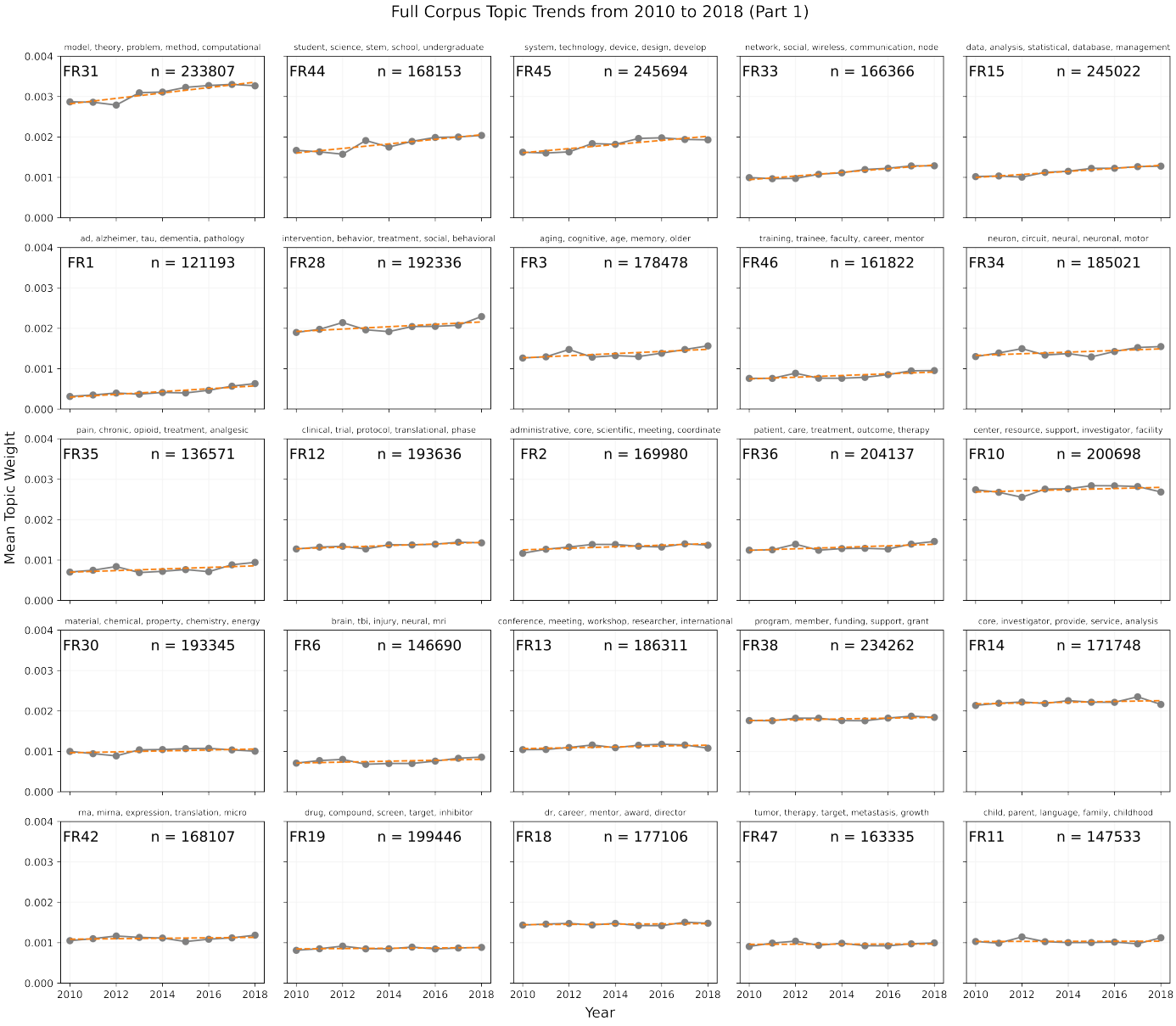

Figure 15: Topic trend results for topics 1 to 25 of 50 topics produced by an NMF model on the full corpus. Trends in topic prevalence are captured between 2010-2018 and topics are ordered from largest positive to largest negative regression line slopes. Topics with positive slopes have orange regression lines and topics with negative slopes have blue regression lines. Standard errors on the means are represented on each plot using error bars.

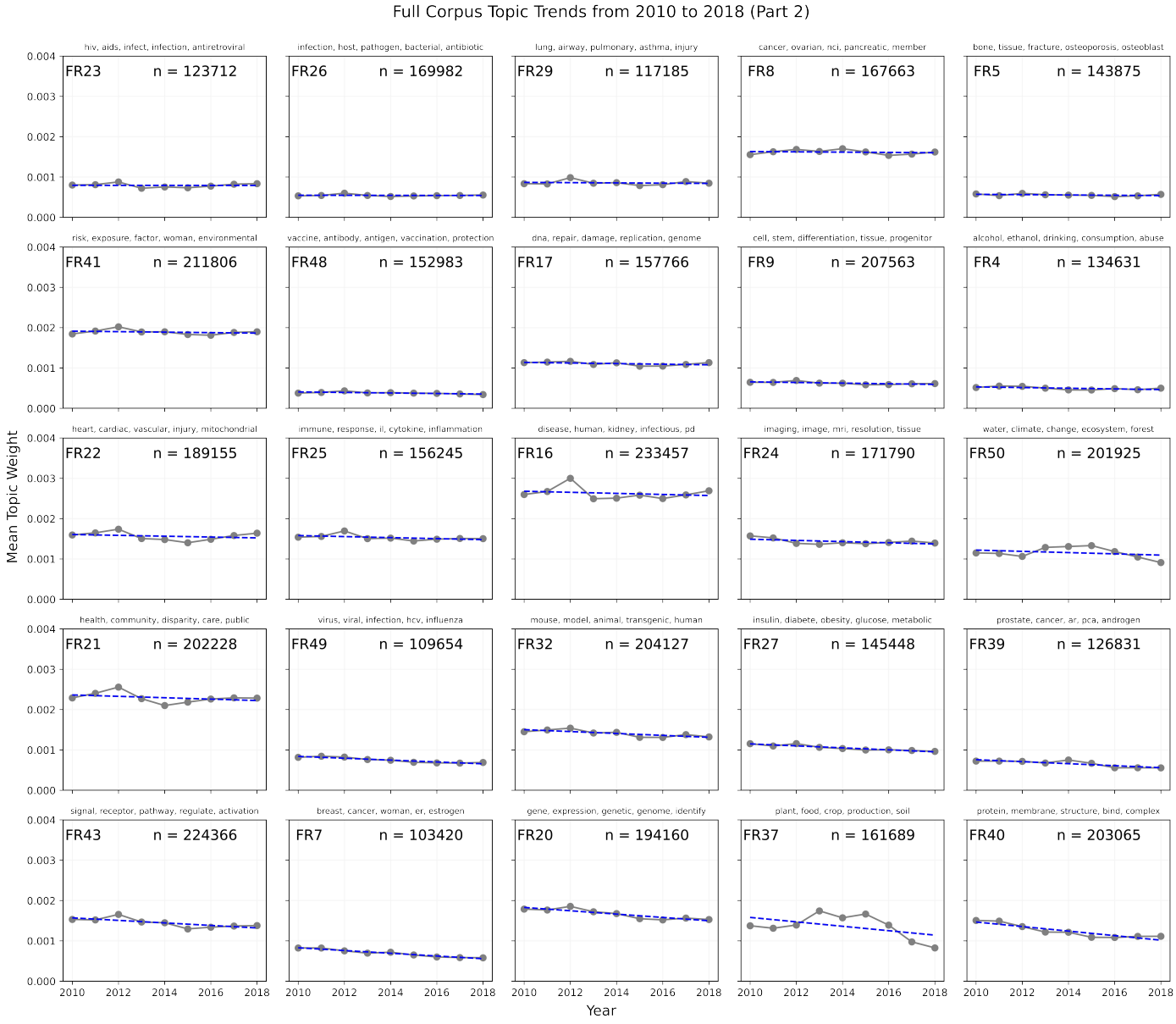

Figure 16: Topic trend results for topics 26 to 50 of 50 topics produced by an NMF model on the full corpus. Trends in topic prevalence are captured between 2010-2018 and topics are ordered from largest positive to largest negative regression line slopes. Topics with positive slopes have orange regression lines and topics with negative slopes have blue regression lines. Standard errors on the means are represented on each plot using error bars.

B.3 100-, 150-, and 200-Topic Models: Trends Calculated Over 2010-2019 and 2010-2018

The ten topics with the largest positive and negative slopes from the 100 topic model for the years 2010 to 2019 and 2010 to 2018 are shown in Figures 17 to 20. Similar figures for the 150 topic model and 200 topic model are shown in Figures 21 to 24 and 25 to 28 respectively. The ten topics with the largest positive slopes for the 100, 150, and 200 topic models from 2010 to 2019 include more HHS related topics such as chronic pain, autism disorder, tuberculosis, PTSD, and RNA sequencing. When examining the ten topics with the largest positive slopes from 2010 to 2018 for the same topic models, we again see that more NSF related topics appear, including those on computational methods, language processing, prediction modeling, software tools, and mathematical theory/geometry. The ten topics with the largest negative slopes for the 100, 150, and 200 topic models from 2010 to 2019 have considerable overlap with those from the 50 topic model. Some of the additional topics include animal models, particle physics, solar energy, and astronomy. When examining the ten topics with the largest negative slopes from 2010 to 2018 for the same topic models, many of the topics are similar to those found from 2010 to 2019.

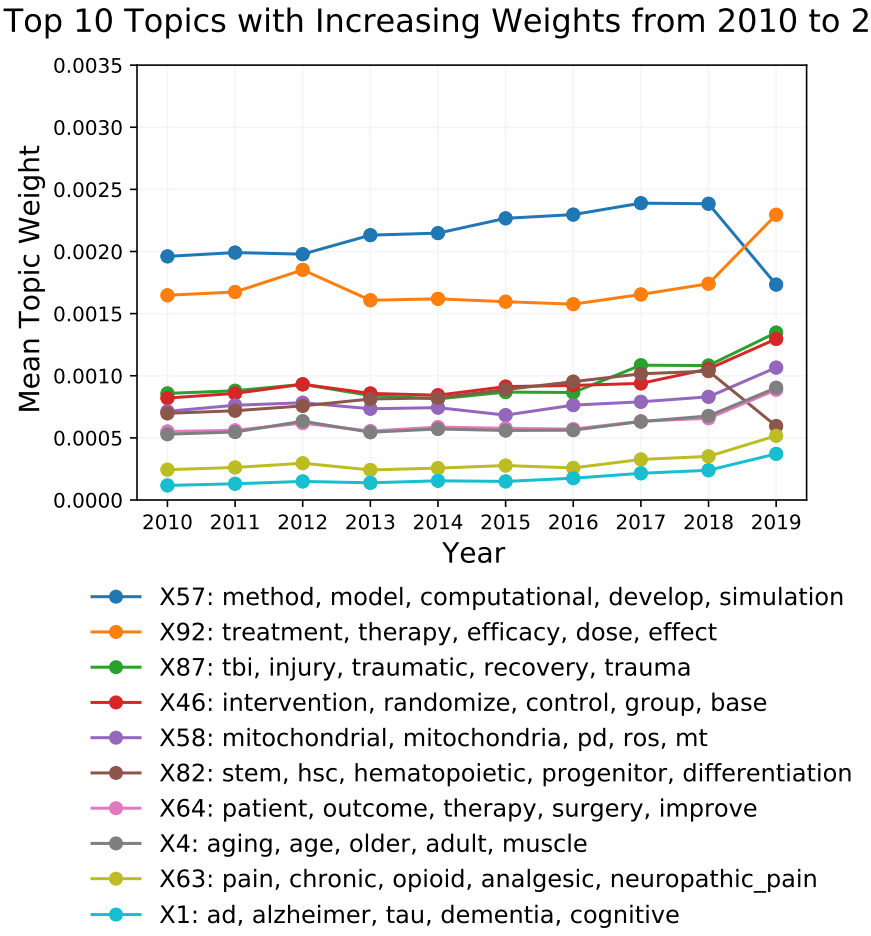

Figure 17: Ten topics with largest positive regression line slopes from the 100 topic model. The slopes are calculated using the weights from 2010 through 2019.

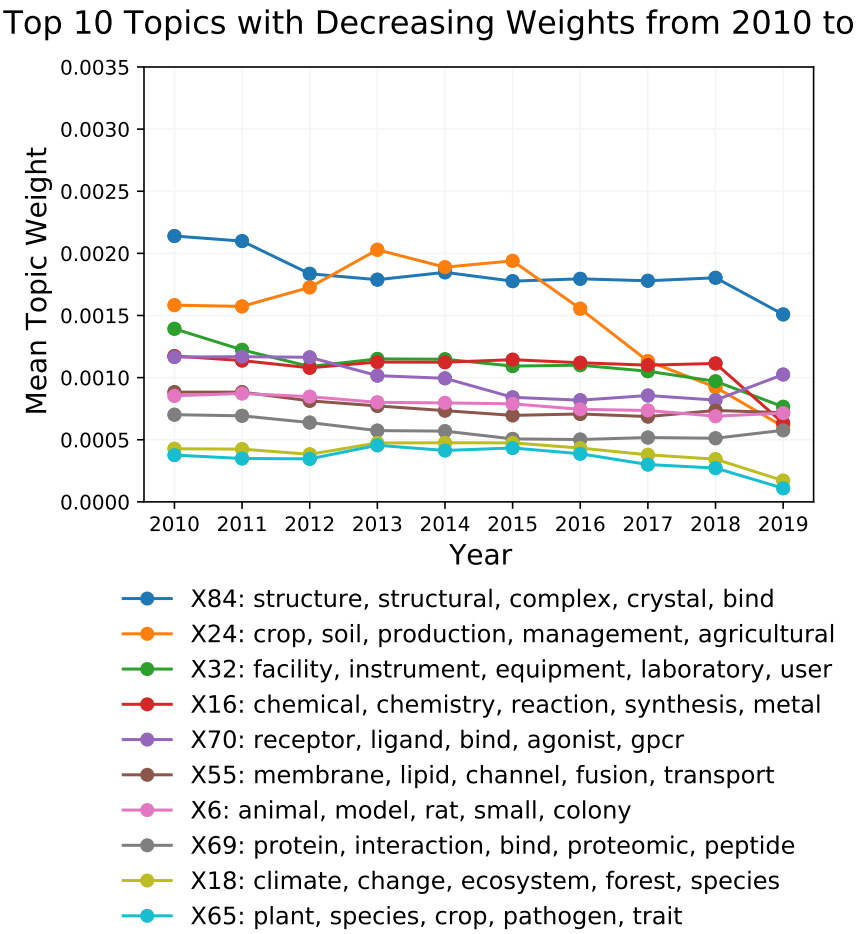

Figure 18: Ten topics with largest negative regression line slopes from the 100 topic model. The slopes are calculated using the weights from 2010 through 2019.

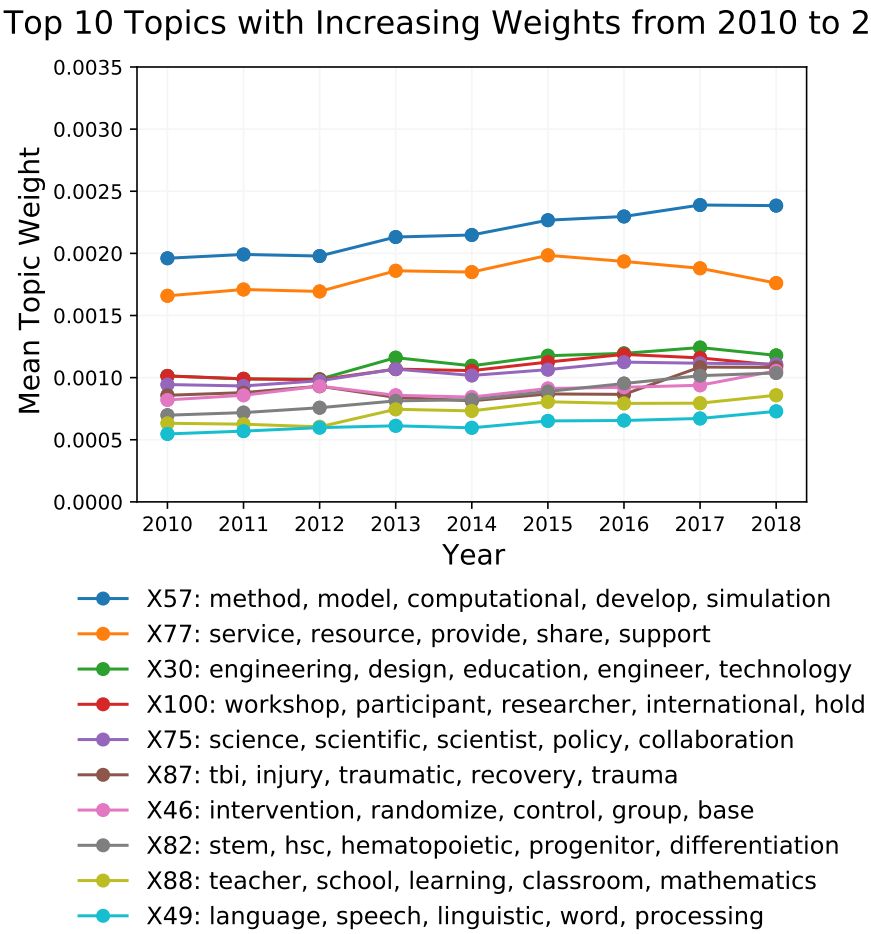

Figure 19: Ten topics with largest positive regression line slopes from the 100 topic model. The slopes are calculated using the weights from 2010 through 2018.

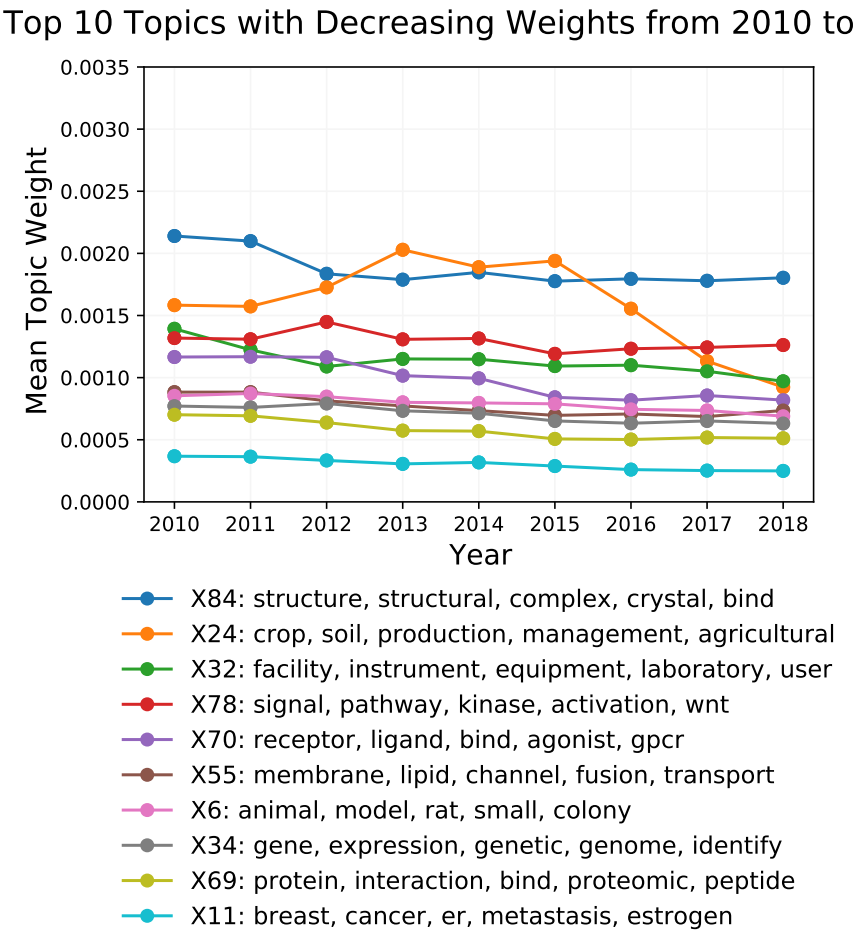

Figure 20: Ten topics with largest negative regression line slopes from the 100 topic model. The slopes are calculated using the weights from 2010 through 2018.

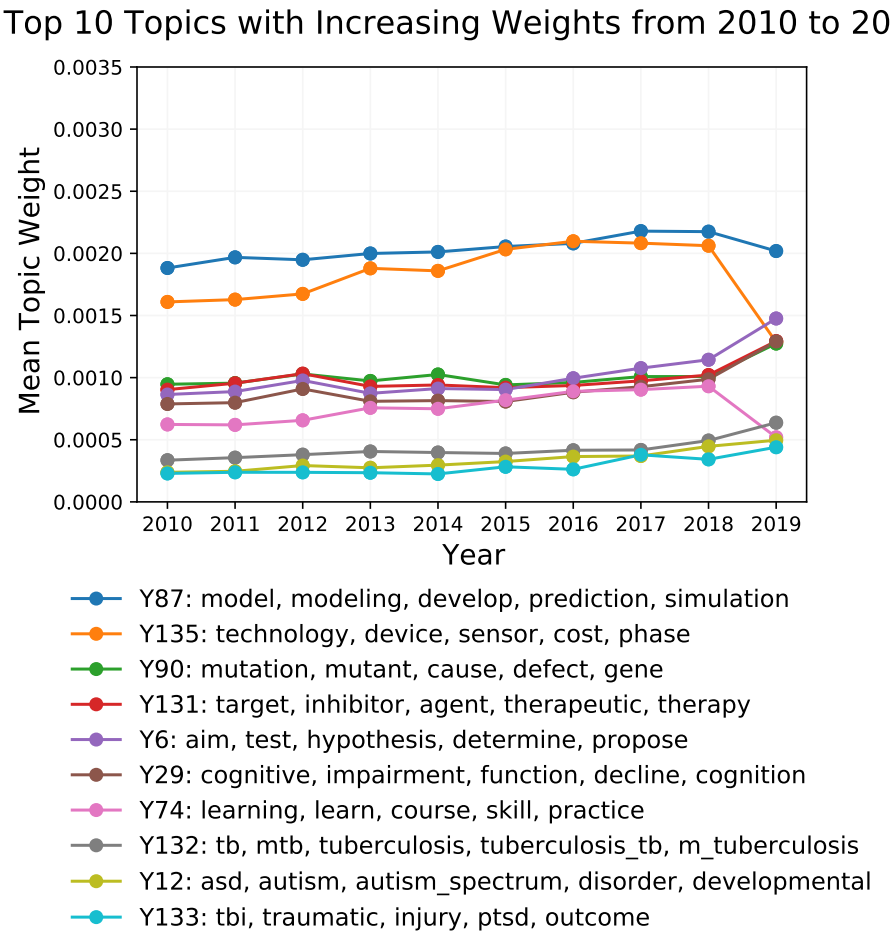

Figure 21: Ten topics with largest positive regression line slopes from the 150 topic model. The slopes are calculated using the weights from 2010 through 2019.

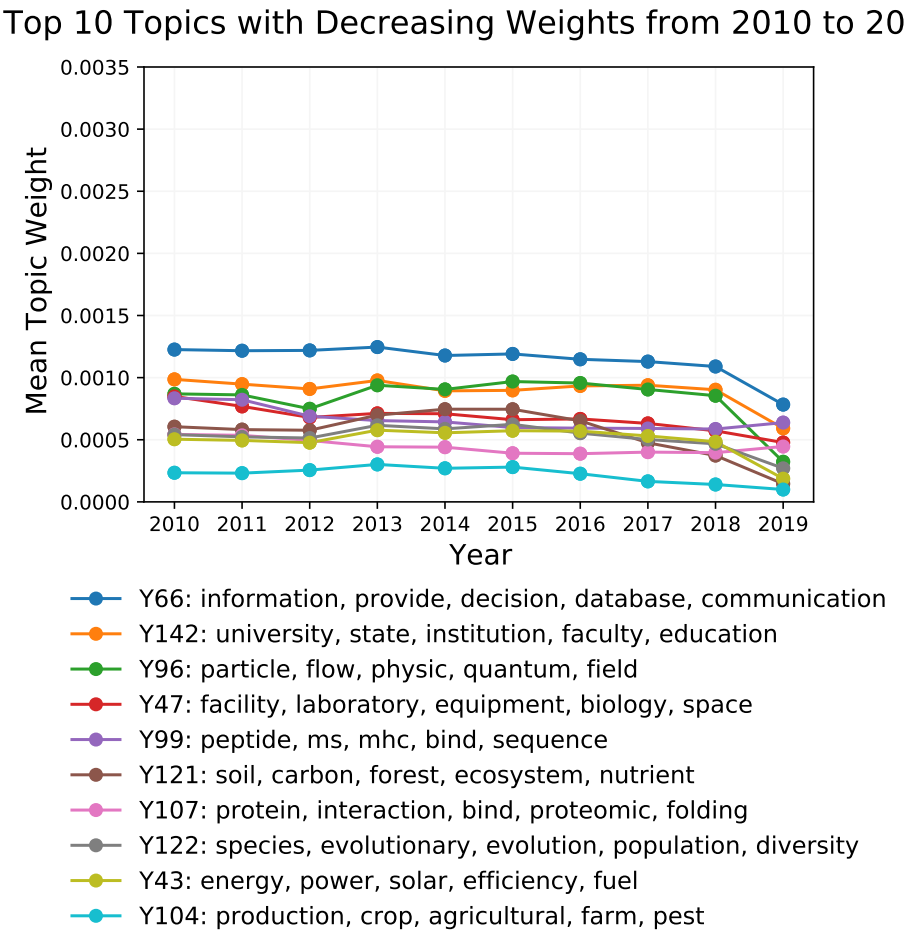

Figure 22: Ten topics with largest negative regression line slopes from the 150 topic model. The slopes are calculated using the weights from 2010 through 2019.

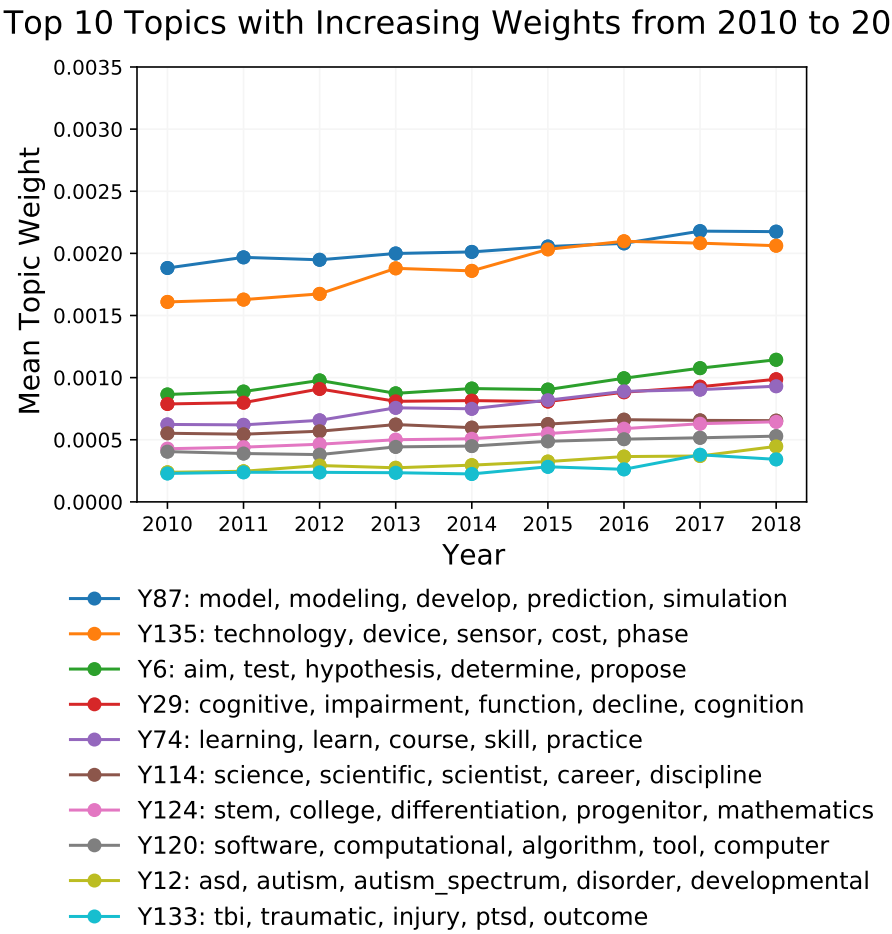

Figure 23: Ten topics with largest positive regression line slopes from the 150 topic model. The slopes are calculated using the weights from 2010 through 2018.

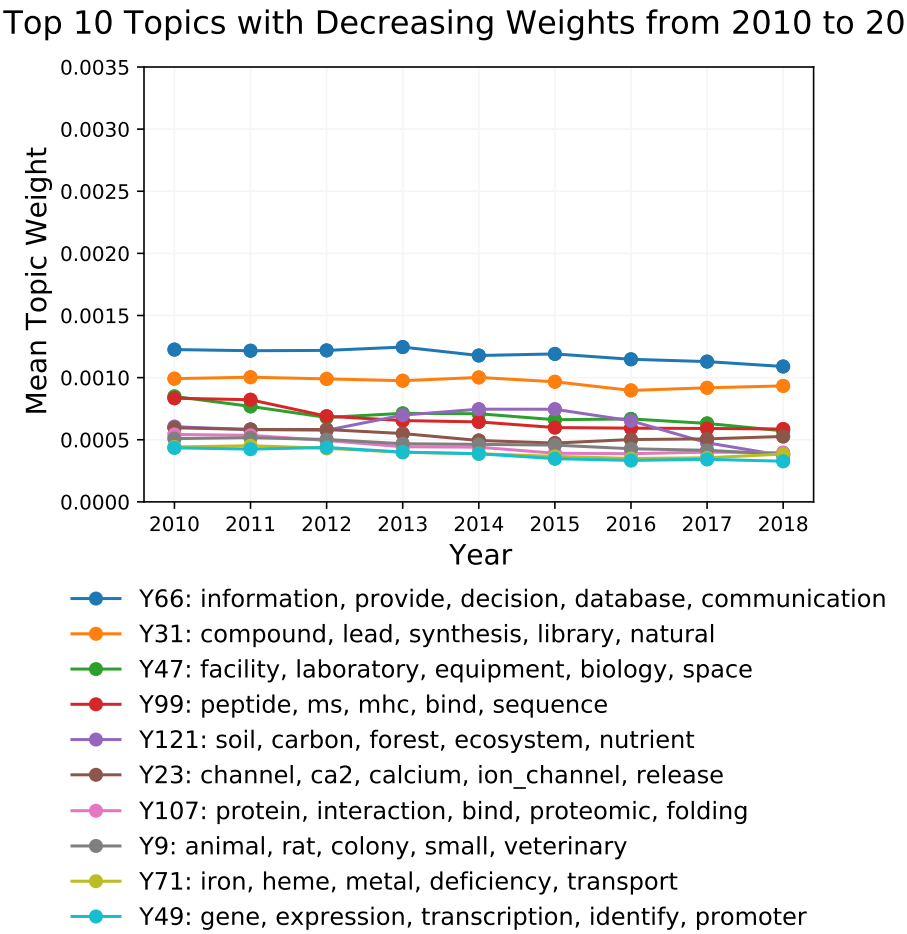

Figure 24: Ten topics with largest negative regression line slopes from the 150 topic model. The slopes are calculated using the weights from 2010 through 2018.

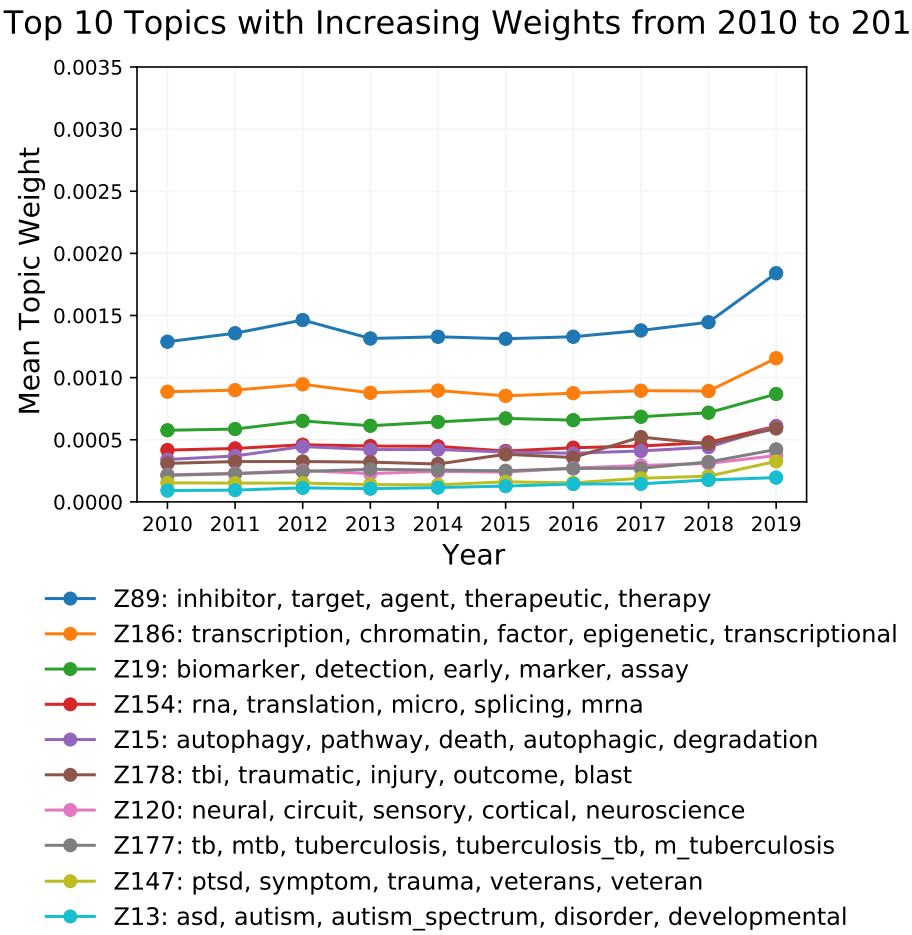

Figure 25: Ten topics with largest positive regression line slopes from the 200 topic model. The slopes are calculated using the weights from 2010 through 2019.

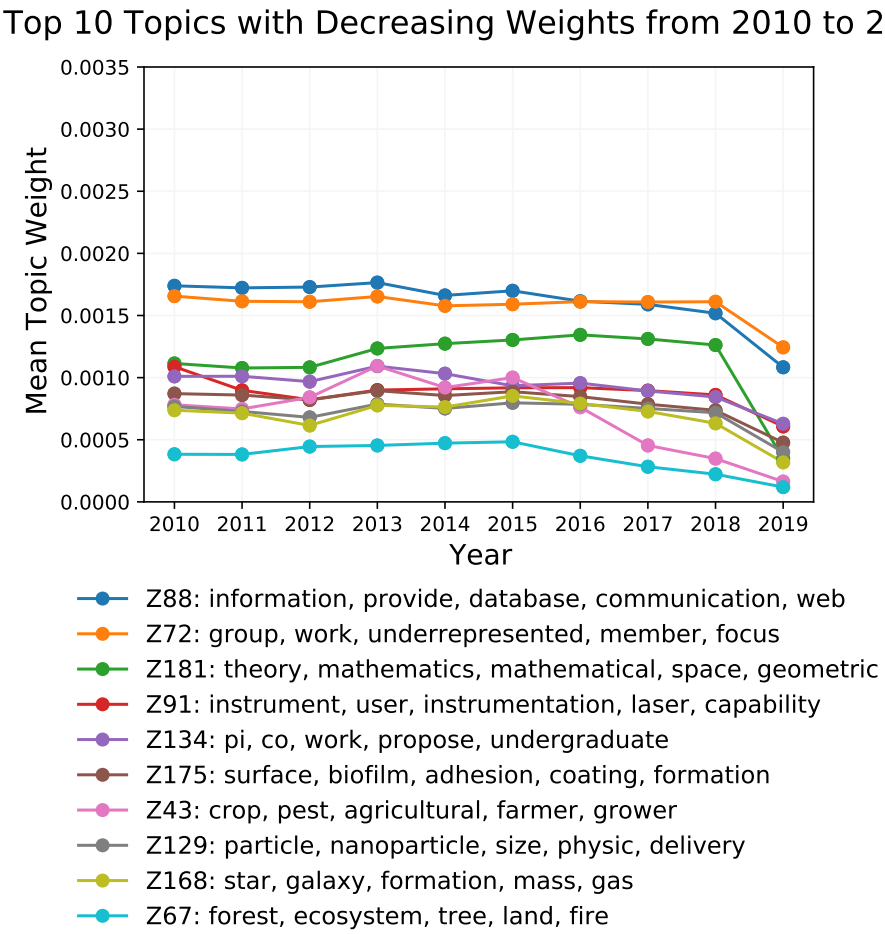

Figure 26: Ten topics with largest negative regression line slopes from the 200 topic model. The slopes are calculated using the weights from 2010 through 2019.

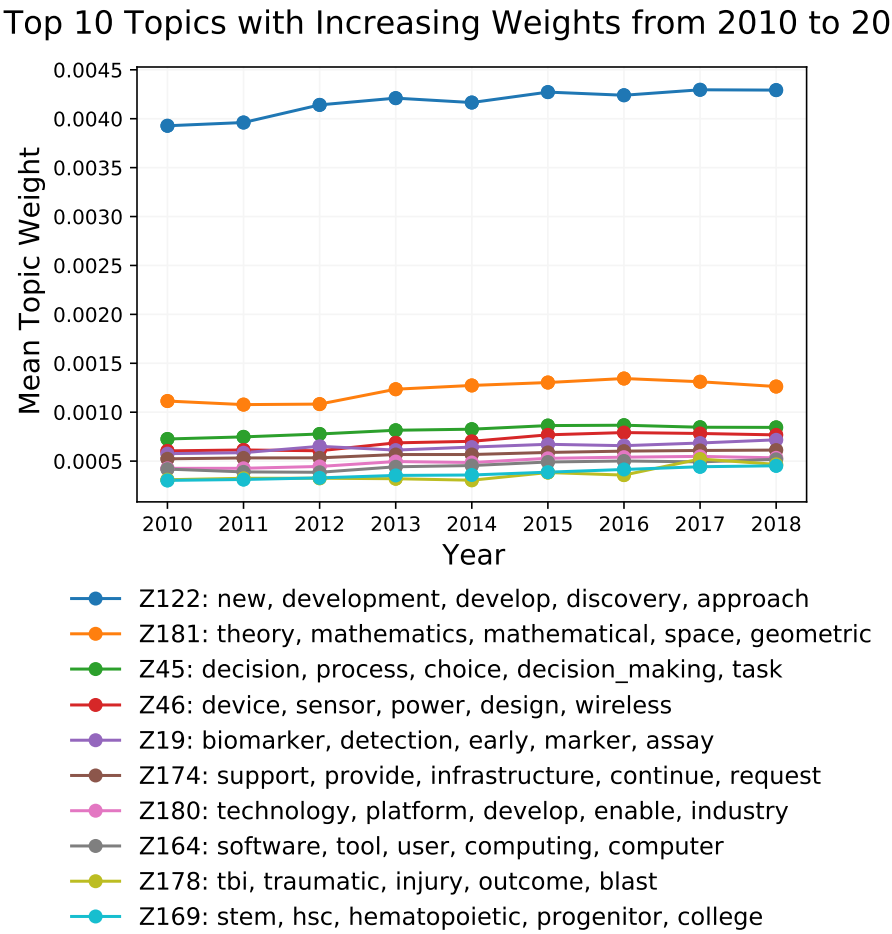

Figure 27: Ten topics with largest positive regression line slopes from the 200 topic model. The slopes are calculated using the weights from 2010 through 2018.

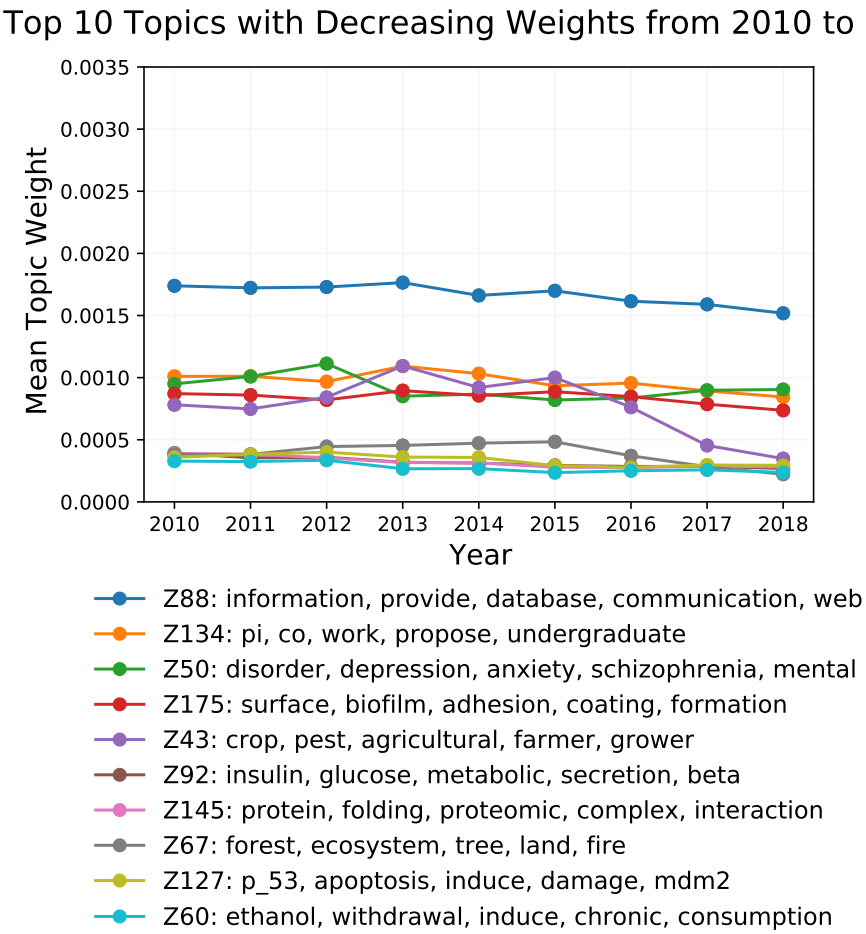

Figure 28: Ten topics with largest negative regression line slopes from the 200 topic model. The slopes are calculated using the weights from 2010 through 2018.

Appendix C LSI Relevance Score Calculation

Assume that A is an $m \times n$ document term matrix where entries are weighted using term frequency-inverse document frequency (TFIDF), and that q is a $n \times 1$ binary query vector with a 1 in entries corresponding to search words and 0 otherwise. The rank-k truncated SVD of A is given by $A = U_k \Sigma_k V_k^T$. The documents and query are transformed through multiplication by V_k: $AV_k = U_k \Sigma_k$ and $q^T V_k$, respectively. The relevance score of each document is the cosine similarity between the transformed document (row of AV_k) and transformed query. For more information on LSI, see Deerwester et al. (1990).

Appendix D Term Matching and LSI versus Term Matching

We provide a comparison of topics found using themed corpora created using term matching and LSI, and those created using only term matching. Results for the pandemics and coronavirus themed corpora are given in Tables 16 and 17 respectively.

Label	Top Five Terms
P1	antibody, neutralization, bind, human, neutralize
P2	attenuate, vaccine, virus, live, candidate
P3	cell, response, memory, infection, cd4
P4	core, diagnostic, support, technology, poc
P5	dengue, virus, serotype, denv, den
P6	drug, inhibitor, compound, resistance, antiviral
P7	epitope, influenza, ha, conserve, strain
P8	facility, product, manufacturing, material, raw
P9	hcv, hepatitis, chimpanzee, genotype, liver
P10	hiv, aids, infect, env, prevention
P11	host, viral, rna, antiviral, replication
P12	hsv, genital_herpes, dl5_29, herpes_simplex_virus, infection
P13	iav, lung, sp, response, evolution
P14	influenza, animal, ecologic, cross_protection, immune
P15	influenza, vaccination, strain, child, virus
P16	mucus, igg, trap, vaginal, trapping
P17	obesity, cancer, insulin, obese, diabete
P18	organism, gene, sequence, ortholog, genome
P19	patient, clinical, trial, dose, care
P20	protein, bind, fusion, structure, membrane
P21	rsv, child, respiratory_syncytial, vaccine, mtase
P22	siv, challenge, mucosal, transmit_founder, transmission
P23	swine, prrsv, prrs, pig, porcine
P24	tb, mtb, co_infection, infection, treatment
P25	training, program, trainee, student, university
P26	transmission, intervention, disease, model, health
P27	vaccine, adjuvant, protection, antigen, immune
P28	virus, human, cause, infect, vector
P29	wnv, flaviviruse, denv, flavivirus, infection
P30	zikv, zika_virus_zikv, microcephaly, infection, fetal

'Pandemic' Term Matching - Top Five Terms
animal, bird, surveillance, close, contact
antibody, ha, epitope, bind, human
cancer, aoic, co_infection, aids, associate
cell, memory, cd4, subset, selectin
core, support, investigator, method, technology
diagnostic, poc, detection, lrs, technology
dna, vaccine, boost, testing, vrc
drug, resistance, inhibitor, compound, antiviral
env, trimer, clade, hiv, gag
facility, product, manufacturing, material, vaccine
hiv, aids, infect, infection, transmission
iav, response, sp, cd1c_dc, htbe
influenza, ecologic, cross_protection, immune, environmental
influenza, strain, virus, vaccination, year
inhibitor, entry, hlv, molecule, fusion
intervention, model, social, health, disease
lung, infection, injury, bacterial, ifn
meeting, scientist, conference, biology, university
obesity, insulin, diabete, metabolic, obese
organism, gene, sequence, ortholog, genome
patient, clinical, trial, dose, protocol
protein, membrane, ns1, viral, interaction
response, immune, age, vaccine, laiv
rna, np, viral, polymerase, virus
tb, mtb, tuberculosis_tb, treatment, tuberculosis
training, program, trainee, faculty, student
v_cholerae, virulence, cholera, colonization, gene
vaccine, adjuvant, antigen, protective, protection
viral, host, evolution, pathogen, evolutionary
virus, human, influenza, 1918, swine

Table 16: Left - table of 30 topics from the NMF model on the pandemics themed corpus. Right - table of 30 topics from an NMF model on the processed project abstracts from Federal RePORTER reported in FY 2008-2019 that contain the term 'pandemic'.

Label	Top Five Terms
C1	ace2, epithelial, sar_cov, airway, lung
C2	aged, mouse, response, cell, severe
C3	animal, influenza, bird, surveillance, contact
C4	assembly, virus, capsid, hcv, particle
C5	cns, mhv, cell, cn, type
C6	compound, fidelity, activity, cov, vivo
C7	core, hrv, stock, virus, recombinant
C8	disease, infectious, respiratory, develop, health
C9	entry, cell, gene, virus, cellular
C10	fusion, peptide, protein, dv, membrane
C11	gene, uncharacterize, orf, encode, rp
C12	iav, evolution, ha, transmission, influenza
C13	il, injury, te, lung, inflammation
C14	immune, polygenic, trait, regulate, response
C15	influenza, virus, 1918, pandemic, human
C16	inhibitor, structure, protease, crystal, enzyme
C17	mers_cov, mers, dpp4, cov, mouse
C18	novel, virus, identify, thesedisease, gastroenteritis
C19	ns1, ifn, rig, trim25, influenza
C20	receptor, rbd, bind, spike, antibody
C21	replication, rna, protein, viral, host
C22	swine, influenza, iaa, relatedness, quantification
C23	vaccine, attenuate, sars_cov, sar_cov, vector
C24	virus, host, transmission, viral, interaction
C25	zoonotic, emerge, bat_cov, movement, species

'Coronavirus' Term Matching - Top Five Terms
ace2, epithelial, sar_cov, nl63_cov, nl63
aged, response, mouse, cell, anti_virus
airway, aav, glycan, tropism, cell
cd8_t, cell, ifn, interferon, type
core, fidelity, compound, cov, exon
disease, lung, develop, prognostic_indicator, severity
entry, gene, virus, share, cell
fusion, peptide, protein, spike, membrane
gene, uncharacterize, rp, encode, lav_ebov
hrv, asthma, assay, core, ppg
immune, polygenic, trait, regulate, protective
inhibitor, structure, protease, crystal, enzyme
mali, lassa_virus, uganda, health, rodent
mers_cov, mers, dpp4, mouse, disease
mhv, cns, chemokine, ifn, ms
novel, virus, thesedisease, gastroenteritis, identify
oligodendrocyte, cn, persistent, cell, demyelination
orf, expression, rna, vector, 2012
rbd, neutralize_antibody, receptor, subunit, vaccine
rna, protein, replication, coronavirus, viral
te, il, injury, inflammation, lung
training, student, program, virology, epidemiology
vaccine, protein, attenuate, sars_cov, rsar_cov
vector, vaccine, ndv, hpiv3, dose
zoonotic, emerge, bat_cov, movement, species

Table 17: Left - table of 25 topics from the NMF model on the coronavirus themed corpus. Right - table of 25 topics from an NMF model on the processed project abstracts from Federal RePORTER reported in FY 2008-2019 that contain the term 'coronavirus'.

Appendix E Related Work

Detecting trends in science and technology is of broad interest to researchers and policymakers alike. Researchers are interested in exploring and learning from new areas, and policymakers want to determine priorities and maintain competitive advantage. Meanwhile, there is growing interest in using machine learning and NLP tools to detect trends from unstructured text by identifying meaning (latent structures) from observed data (e.g., words in text) (Griffiths & Steyvers, 2004). Specifically, many researchers have used LDA and NMF, two popular topic modeling algorithms, to organize textual information and detect trends.

Griffiths and Steyvers (2004) used LDA to identify topics from a set of Proceedings of the National Academy of Sciences (PNAS) abstracts from 1991-2011. They analyzed the dynamics of these topics to characterize "hot" and "cold" topics that rise and fall in popularity using a linear trend analysis. The hottest and coldest topics were selected based on the size of the linear trend test statistic.

H. Lee and Kang (2018) applied LDA to nearly 12,000 articles published between 1997 to 2016 in 11 technology and innovation management (TIM) journals and reviewed the top 10 most popular topics in this area of research. Topics were ranked in descending order of their proportions in the collection of articles, and the rankings were compared over four time periods between 1997 and 2016 (i.e., 1997-2001, 2002-2005, 2006-2010, and 2010-2016). They applied the Griffiths and Steyvers (2004) approach of estimating linear regression to find hot and cold topics, then ranked increasing and decreasing linear trends (identified as statistically significant regression slopes) of the mean topic weight over time and visualized the trends.

To define trends in hotspots (popular topics), Wang et al. (2019) combined four methods (1) segmentation information statistics to determine the details (complexities and granularities) of the grants project titles, (2) word clouds to display hot words, (3) NMF methods to analyze the distributions of hot topics and the corresponding keywords over time, and (4) hotspot evolution analysis based on semantic computing and keyword scoring using semantic retrieval. They then compared NMF to other methods (principal component analysis, SVD, and LDA). They found that NMF was a better method for detecting hotspots based on two different evaluation metrics. The authors concluded that the use of project titles was sufficient for the analysis as they are "increasingly comprehensive and meticulous" (Wang et al., 2019). Other authors such as Doanvo et al. (2020), used principal component analysis and LDA to draw conclusions about research needs highlighted in the COVID-19 literature.

Automated versus human classification. Some authors noted that automation is the key to successful detection of trends stating that unsupervised learning methods used to classify scientific knowledge "eliminate the need to fit new-to-the-world knowledge into known-to-the-world definitions" (Suominen & Toivanen, 2016). Others, however, used a mix of automated and human-involved methods. One such approach created integrated frameworks that include topic modeling and technical expertise (Zhou et al., 2019). Others used expert knowledge to assess if the results are "reasonable" (Zhang, Porter, et al., 2019) or questionnaires to ask researchers why they participated in a specific research area to assess how community dynamics evolve during the emergence phase of a technology (Suominen et al., 2019).

Eads et al. (2021) created a structured procedure for filtering large amounts of text. This is a semiautomatic method for finding documents containing a complex multi-faceted concept with boundaries that are not well defined. This systematic approach requires some human intervention in developing and refining keyword lists initially derived using topic modeling. This iterative approach is guided by a document taxonomy that classified text into relevant types of text and non-relevant

ones and has been applied by other researchers to analyze trends in other topic areas such as artificial intelligence (Eads et al., 2021).

To accelerate research on detecting and analyzing trends in science and technology, the U.S. Intelligence Advanced Research Projects Activity (IARPA) created the Foresight and Understanding from Scientific Exposition (FUSE) program. Their goal was to partner with the scientific community "to develop validated indicators and theories of technical emergence detection" (Office of the Director of National Intelligence [ODNI], 2011, September 27). In a special issue of Technological Forecasting and Social Change in 2019, selected FUSE projects were highlighted, some of which are highlighted here. The FUSE program created a lot of excitement in the growing community.

Describing Trends. Topic trends across time are described both conceptually and by methods used to measure the trends. The descriptive language includes references to "scientific breakthrough" (Winnink et al., 2019), "potential innovation pathways" (Zhou et al., 2019), and "emerging research leading to commercialization" (Jeong et al., 2019) and (Zhang, Porter, et al., 2019). Winnink et al. (2019) classified breakthroughs by the prominence of the publication or patent, such as Nobel Prize research papers, papers occurring in Nature's Top-100 Most Cited Papers Ever, papers still highly cited by review papers or patents, or those frequently mentioned in today's social media. Zhou et al. (2019) examined technology trends by combining text mining approaches and Technology Roadmapping (TRM) to explore how technologies develop and identify innovation pathways. Trends in research and business development areas are Bishop et al. (1998) used LDA and network analysis to predict the structure of relationships among social entities as well as Generative Topographic Mapping (GTM) to group, visualize, and interpret the data.

Other approaches to identify trends include tracing how technologies evolved over a life cycle (Berg et al., 2019); using a scientific evolutionary pathways (SEP) approach (Zhang, Porter, et al., 2019); measuring how research participation and community dynamics evolved (Suominen et al., 2019); and using trends analysis (Griffiths & Steyvers, 2004; Suominen & Toivanen, 2016). For a technology life-cycle measurement approach, patents were used to define three eras: (1) ferment, (2) dominant design, and (3) incremental change. Berg et al. (2019) created two indicators to measure patent trajectories and category concentration that anticipated emerging category dominance early signals. The method was tested on one facet of the bioeconomy (algae) and has not yet been examined for the entire life cycle.

Zhang, Huang, et al. (2019) adopted a SEP approach to detect and visualize technological changes in big data research from 2000 to 2015. This method used text mining and bibliometric techniques to detect and visualize changes in research. They tracked the interactive relationships between topics in sequential time slices. The model identified technological evolution and death by identifying predecessors and descendants of big data topics.

Suominen et al. (2019) examined the growth and persistence of a research community, measured by the number of unique authors who are active, new, or leaving, as a trend indicator. Park et al. (2017) derived topics using LDA and then divided technologies into hot and cold based on trend analysis. Suominen and Toivanen (2016) also used LDA and then compared how topics grew yearly against the overall growth in scientific publishing. They defined increasing trends as topics that grew faster than science publishing overall.

Porter et al. (2019) created indicators of technological emergence for R&D priorities. They implemented an algorithm to calculate an R&D emergence indicator that captured novelty, persistence, growth, and community. They calculated scores to identify recent surges in R&D activity in a field of study and, using these scores, created indicators of new terms and leading players in a community. Primary emergence indicators identified hot topic terms, which were then used to produce secondary indicators to identify active, cutting-edge organizations, countries, and authors

in the selected R&D domain.

Narrow vs. broad area focus. When using topic modeling approaches, some researchers focused on single fields of study. Zhou et al. (2019) examined the evolving field of solid lipid nanoparticles, which they described as emerging from within the field of nano-enabled drug delivery. Berg et al. (2019) used patents to study the changes in algae research, one facet of the bioeconomy. Jeong et al. (2019) focused on organic light-emitting diode (OLED) technology to examine research and business development changes. Suominen et al. (2019) examined the triboelectric nanogenerator (TENG) technology field to investigate the growth and decline in researcher communities to measure emerging trends. Other researchers leveraged topic modeling to identify trends within broader areas of research. Park et al. (2017) utilized artificial intelligence to understand and predict science and technology trends. Wang et al. (2019) examined grant titles for one of eight departments in the National Natural Science Foundation of China (NSFC). They noted that their approach and results for the Department of Information Science may or may not apply to the other seven departments. Suominen and Toivanen (2016) used publications to create a science map of Finland over the 1995-2011 time period. Others took a broader tract examining overarching topics such as research and development (Choi et al., 2019) and big data (Zhang, Huang, et al., 2019).

A narrow topic may be examined if there exists a focused set of text on a specific topic. Alternatively, information retrieval techniques, such as term matching and latent semantic indexing (LSI) can be used to identify relevant documents. Term matching retrieves relevant text using specific keywords. For example, Doanvo et al. (2020), (2020) used search terms such as "COVID-19", "COVID", "2019-nCOV" and "SARS-CoV-2" (case sensitive) to search for coronavirus topics that related to the pandemic that emerged in 2019. This method can be combined with LSI to identify theme-relevant documents that may not necessarily contain the keywords used in term matching (Deerwester et al., 1990).

OECD (2019) examined artificial intelligence (AI), a single field covering many broad areas, such as health care, banking and finance, surveillance, space exploration, and almost every area that touches our lives. They developed keywords to identify relevant abstracts from NSF and NIH funded projects. To create keywords, they created an operational definition of AI using the NIH taxonomy defined by medical subject headings (MeSH), input from an OECD Advisory Expert Groups, a literature search using Scopus, Web of Science, and Google Scholar, and information from other US Federal agencies that conduct or fund AI research. They divided their keywords into core and non-core terms. A document was determined to focus on AI if it contained at least one core term within its title or abstract or two or more different non-core words. They have used this method to identify AI-related R&D projects in 13 funding databases from eight OECD countries (Yamashita et al., 2021).

Validating findings. Researchers have used a variety of methods to validate their findings, including expert input, questionnaires, and classification manuals. Griffiths and Steyvers (2004) validated their findings through comparisons with Nobel Prizes. They validated popular topic trends that occurred during earlier years that were later recognized by Nobel Prizes. For example, Nobel Prizes were awarded for work on immunology in 1989 and sequencing in 1993 for research conducted in the 1980s.

Zhang, Huang, et al. (2019) consulted an expert panel about their analytic results to determine if the results were reasonable and, if not, how to modify their approach. Jeong et al. (2019) tried to implement a systematic approach to reduce the influence of subjective opinions; however they ended up seeking expert input to achieve more accurate forecasts. Suominen et al. (2019) invited researchers to answer a questionnaire to ask why they participated in a specific research area to assess how community dynamics evolve during the emergence phase of a technology. Ankam et

al. (2012) measured novelty by listing, counting, and then comparing the words in the abstract to the description of the topic area in the US Patent classification manual. A star graph visually represented the overlap and differences between the patent classification manual's broad descriptions and the submitted applications.

OECD (2019) validated their keyword approach to identify AI government-funded projects by more closely examining a sample of 400 documents. They discovered that core terms provide "reasonably unambiguous predictors" for AI relevance (e.g., "machine learning," "natural language processing," and "deep learning.") Non-core terms, on the other hand, often required checking the context in which they were used in a document. Using Lorenz concentration graphs, they measured the degree of AI intensity of funding at NIH and NSF. To test for bias, they examined the incidence of false positives and negatives. To test for robustness, they compared their keywords to other lists found in the literature.

Researchers used coherence measures to assess the interpretability of topic model results. Röder et al. (2015) developed a framework to evaluate coherence measures using publicly available topic data sets as a benchmark. These data sets record human judgments of the interpretability of topics. The authors compared seven component measures and evaluated them for performance of their data sets and models. Out of the seven, they found that the coherence measure (C_V) outperformed the other measures. C_V combines the indirect cosine measure with the NPMI (normalized pointwise mutual information) coherence measure (one of the six measures) and the boolean sliding window that captures word counts and proximity between word tokens (Röder et al., 2015). The authors found that using various combinations of measures provides more insights than any one single measure. They validated the performance measures with human ratings.

Limitations of studies. Most articles reviewed provided insights for potential approaches to avoid in future studies based on the limitations they identified. Zhang, Huang, et al. (2019) noted that most studies used one or two data sources such as academic journal articles or patents. By not including popular, business, and regulatory articles within the data, the results may miss part of the technology life-cycle, such as applying basic research to applications and commercialization. The time frame covered can be another limitation, especially if the time frame is too narrow. The timing of the data used can influence results (Suominen et al., 2019). For example, Winnink et al. (2019) used short time frames to identify breakthroughs. They examined papers two to three years after their initial publication, potentially missing out on "sleeping beauties" or research that receives delayed recognition. Additionally, such a limited time frame does not account for breakthroughs that do not achieve success or article retractions that can take up to four or more years to occur. Another challenge is related to changes that occur in data sources. In the 1990s, Web of Science increased indexing of conference proceedings and abstracts. This change was responsible for much of the publication growth observed in the late 1990s (Suominen & Toivanen, 2016). Finally, lags in updating databases over time can lead to using incomplete data.

Mohr and Bogdanov (2013) also identified limitations. They noted that topic models can be informative for projects that use text to measure meaning but are less suitable for studying narratives, such as autobiographies, theater plays, or other stories. They expressed concern that it can be challenging to assess interpretability. They proposed that topic modeling is a lens to view text in different light and scale and not as an automatic text analysis program. Topic modeling can provide early insights into an area of research and guide the acquisition of deeper knowledge and context (Mohr & Bogdanov, 2013).

To summarize, topic modeling can provide early insights into an area of research and guide acquiring deeper knowledge and context (Mohr & Bogdanov, 2013). The benefits and uses of this research include the ability to identify directionality of research, inform R&D and innovation activities, and tailor advice to policymakers about science, technology, and innovation funding priorities (OECD, 2021).