Using Machine Learning and Information Retrieval to Identify Federally Funded Research and Development Trends

Authors

1. Kathryn Linehan: University of Virginia (UVA), Biocomplexity Institute (BI), Social and Decision Analytics Division (SDAD).

   Corresponding author at: Biocomplexity Institute, 29th Floor, 1100 Wilson Blvd., Arlington, VA 22209. Email address: kjl5t@virginia.edu

2. Eric Oh: UVA, BI, SDAD

3. Joel Thurston: UVA, BI, SDAD

4. Guy Leonel Siwe: UVA, BI, SDAD

5. Audrey Kindlon: National Center for Science and Engineering Statistics (NCSES)

6. John Jankowksi: NCSES

7. Stephanie Shipp: UVA, BI, SDAD

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 (BI), 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 thank Madeline Garrett for her contributions in the 2021-2022 academic year as a student Project Fellow, and Samantha Cohen, a former SDAD Postdoctoral Research Associate, for contributions to the text cleaning and processing work. We thank Sallie Keller for helpful discussions and suggestions throughout the project. Dr. Keller is the Distinguished Professor in Biocomplexity, Division Director, SDAD, BI, Professor of Public Health Sciences, University of Virginia.

Declarations of Interest

none

Using Machine Learning and Information Retrieval to Identify Federally Funded Research and Development Trends

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

Abstract

A vast amount of information on federally funded research and development (R&D) is available and can be utilized by researchers, policymakers, and the public to uncover insights on the directionality and extent of government R&D funding. In this work, we use natural language processing (NLP), machine learning, and information retrieval techniques to classify broad research topics and pandemic-related research topics contained within Federal RePORTER grant abstracts, a typical example of a scientific award database. In collaboration with the National Center for Science and Engineering Statistics (NCSES), we examine these topics, their trends over time, and how the topics and their trends change as a result of the number of topics produced by the model. The methods described in this paper show promise to supplement the information currently collected through the NCSES Federal Survey of Funds for Research and Development (FFS) and Survey of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions (FSS) by providing information that the surveys do not collect.

1 Introduction

Research and development (R&D) is defined by the Organisation for Economic Co-operation and Development (OECD, 2015) 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". R&D is critical to a country's development. In the United States (U.S.), R&D spending increased yearly from 2010 to 2017, averaging $21 billion annually. In 2018 and 2019, there were even greater increases in R&D spending from the previous years, totaling about an additional $50 billion in expenditures. In addition, the ratio of national R&D spending to national gross domestic product (GDP), a common metric of a nation's R&D effort, has steadily risen in the U.S. since the mid-1990's. Currently, nearly 21% of all U.S. R&D funding is provided by the federal government (National Center for Science and Engineering Statistics [NCSES], 2021b).

Administrative data exist that provide information about federal funding of R&D i.e., scientific activity focused on basic and applied research and technological developments, as well as more general science and engineering (S&E)[3]. This work is focused on assessing the feasibility of using administrative data to supplement data collected through the NCSES Federal Survey of Funds for Research and Development (FFS) and Survey of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions (FSS) (Pece, 2016). While the FFS presents federal R&D funding information by broad disciplines (e.g. mathematics, computer science, civil engineering, oceanography, economics, or sociology), neither survey presents information on the specific R&D areas that the funding supports (e.g. Alzheimer's disease, HIV/AIDS, food safety, or sleep disorders)[4].

[1] University of Virginia, Biocomplexity Institute, Social and Decision Analytics Division

[2] National Center for Science and Engineering Statistics

[3] S&E includes R&D as well as fellowships, traineeships, and training grants.

[4] An exception to this is that data on federal funding for COVID-19 related R&D was collected in the FFS for FY 2020-21 (NCSES, 2022).

We use Federal RePORTER, administrative data and a typical example of a scientific award database, to characterize federally funded R&D research areas. Given the large collection of grants in the database, we use the text analysis method of topic modeling to discover the latent topics from the grant abstract text in Federal RePORTER. This results in a broad categorization of topics; however, the researcher may also be interested in more granular information about latent topics related 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, or SARS. These pandemic themes may not appear as frequently in the text as the themes identified by the topic model, so may not appear in the broad classification of topics given by the model. To find theme-related topics we use information retrieval in conjuction with topic modeling. In this work, we demonstrate the use of two types of topic models, latent Dirichlet allocation (LDA) and non-negative matrix factorization (NMF), two popular approaches to characterize latent topics in text data. We also use the information retrieval techniques of term matching and latent semantic indexing (LSI) to discover pandemic-related topics in this corpus. We then analyze the discovered topics over time using a linear trend analysis.

Our contributions in this paper are threefold. We present a novel use of scientific award data to discover R&D topics and identify topic trends over time using machine learning. 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. Lastly, we demonstrate the use of information retrieval techniques to organize and interpret large data sets, highlighted by a case study focusing on pandemic-related topics. The methods described in this paper show promise to supplement the information currently collected in the NCSES FFS and FSS surveys by providing information that the surveys do not collect. The benefits 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).

The paper is organized as follows. Section 2 presents and discusses related work. Section 3 provides information on the Federal RePORTER data, and Section 3.1 describes the steps required to process these data. Section 4 reviews topic modeling and analyzes trends in topics across time on our processed Federal RePORTER dataset. Section 5 covers filtering our corpus for the pandemics theme and a topic trend analyses through the use of a case study. The paper finishes with Sections 6 and 7: conclusion & future work and acknowledgements.

2 Related Work

There is accelerating interest in using machine learning and natural language processing (NLP) tools to detect trends from unstructured text (Griffiths & Steyvers, 2004). Many researchers have used LDA and NMF to organize textual information and detect trends. For example, 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 linear regression. The hottest and coldest topics were selected based on the size of the regression line slope.

Wang et al. (2019) compared NMF to other methods (principal component analysis, singular value decomposition, and LDA). They found that NMF was a better method for detecting emerging topics based on two different evaluation metrics. The authors also concluded that the use of project titles (rather than abstracts, for example) as input to a topic model were sufficient for the analysis as they are "increasingly comprehensive and meticulous" (Wang et al., 2019). While some authors advocate that automation is the key to successfully detecting trends using unsupervised learning methods to classify scientific knowledge (Jeong et al., 2019; Suominen & Toivanen, 2016), others 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).

Topic trend classification techniques. Topic trends have been classified and described differently across time. For example, Winnink et al. (2019) classified breakthroughs by the prominence of the publication, Suominen et al. (2019) examined the number of authors active, new, or leaving, and Suominen and Toivanen (2016) used LDA to compare yearly against the overall growth. Porter et al. (2019) created indicators of technological emergence for R&D priorities by implementing an algorithm to calculate an R&D emergence indicator.

Narrow vs. broad area focus. Topic modeling can provide early insights into an area of research and guide acquiring deeper knowledge and context (Mohr & Bogdanov, 2013). Some researchers focus on single fields of study when using topic modeling approaches. For example, Zhou et al. (2019) examined the evolving field of solid lipid nanoparticles, Berg et al. (2019) studied the changes in algae research, and Suominen et al. (2019) examined the triboelectric nanogenerator technology field. Doanvo et al. (2020) used principal component analysis and LDA to investigate COVID-19 research hotspots and areas warranting exploration.

A narrow theme may be examined if there exists a focused set of text on a specific theme. Information retrieval techniques such as term matching can be used to identify thematically relevant documents in a more general corpus. Term matching retrieves relevant documents using specific keywords. For example, Doanvo et al. (2020), used search terms such as “COVID-19”, “COVID”, “2019-nCOV” and “SARS-CoV-2” (case sensitive) to search for documents that related to the pandemic that emerged in 2019.

In contrast to narrow themes discussed, OECD (2019) examined artificial intelligence (AI), a broad field that touches on many other areas, such as health care, banking and finance, surveillance, space exploration, and almost every area that touches our lives. In their early work, they developed a keyword list to identify relevant abstracts of U.S. National Science Foundation (NSF) and National Institutes of Health (NIH) funded projects. To construct this list, they created an operational definition of AI, drew on expert input, and used word2vec (Mikolov et al., 2013), a neural network approach. They divided their keywords into core and non-core terms. A document was determined to focus on AI if its title or abstract contained at least one core term or two or more distinct non-core terms. They have used this method to identify AI-related R&D projects in 13 funding databases from eight OECD countries (Yamashita et al., 2021). Similarly, Eads et al. (2021) created a structured procedure for identifying theme relevant documents in a corpus utilizing topic modeling, keyword list creation, and human intervention.

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 results through comparisons with Nobel Prizes. Zhang et al. (2019) consulted an expert panel to determine if their analytical results were reasonable and, if not, how to modify their approach. Suominen et al. (2019) invited researchers to answer a questionnaire to assess how dynamics evolve during the emergence phase of a technology. OECD (2019) researchers validated their keyword approach to identify AI-related government-funded projects by examining a sample of 400 documents more closely. Researchers have also used coherence measures to assess the understanding of topic model results (Röder et al., 2015).

3 Data

The data source used for the research is Federal RePORTER, a recently retired publicly available and searchable database of scientific awards from federal agencies[5]. Federal RePORTER is a typical example of a scientific award repository that includes information such as award description and funding agency, amount, and fiscal year. We used Federal RePORTER to analyze federally funded R&D topics rather than sources such as USAspending (U.S. Government Accountability Office [GAO], 2021) or separate agency databases (e.g. NSF Award Search) because Federal RePORTER contained the raw project abstracts and included data from most science and technology federal agencies.

As part of the 2009 American Recovery and Reinvestment Act's Science of Science Policy initiative, STAR METRICS® (Science and Technology for America's Reinvestment—Measuring the Effects of Research on Innovation, Competitiveness, and Science) led the effort to create Federal RePORTER (Federal Research Portfolio Online Reporting Tools). This database was designed to be "a repository of data and tools" that "promote[d] transparency and engage[d] 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-a, 2020, March 6-b).

Federal RePORTER contained project abstracts and related metadata for more than 1 million federally funded grants from science and technology federal agencies beginning in fiscal year (FY) 2000; however large scale reporting began in FY 2008. Project metadata included the title, funding agency, FY, and other information such as the principal investigator, organization, start date, and FY total cost. Historically, all of the project data was submitted to Federal RePORTER by the individual agencies themselves; however, this was recently changed and Federal RePORTER downloaded some agency data directly from Research.gov[6]. We recognize that not all federally funded R&D projects may have appeared in Federal RePORTER and that some projects in Federal RePORTER may have been categorized in the broader class of S&E and not R&D. However we expect the percentage of projects that are S&E but not R&D to be small[7] and thus assumed that all Federal RePORTER projects could be classified as R&D.

3.1 Data Processing

We filtered Federal RePORTER data from FYs 2008-2020, resulting in a total of 1,262,655 projects. We decided to use project abstracts to identify R&D trends, so removed 42,536 projects with a null (missing) abstract from the dataset. We linked each project with the FY in which it was awarded, allowing us to eventually connect R&D topics with specific projects and years and track these topics over time. In addition, we deduplicated (i.e., removed all but one entry) projects that shared the same title, abstract, and FY to be able to identify the proportion of novel projects associated with a topic in any given year. For example, multi-institutional projects (projects associated with

[5] Federal RePORTER was retired on March 1st, 2022, although archived data through fiscal year (FY) 2020 is available at https://federalreporter.nih.gov/. We completed our analysis before the retirement was announced.

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

[7] In Federal RePORTER we estimate that at least 75.4% of grants are to institutions of higher education (see footnote 3 for definition of S&E). 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". The Survey of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions: Fiscal Year 2019, Table 1 (NCSES, 2021a) shows that in the years 2008-2019, on average, 89.6% of dollars obligated to universities and colleges for S&E are R&D.

different investigators across two or more universities) with an entry in Federal RePORTER for each organization were considered duplicate entries for the same project and were deduplicated. There were 71,902 duplicate entries removed from the dataset. Furthermore, we removed projects with abstracts that were short phrases such as 'Abstract not provided' and 'No abstract provided'.

For the remaining projects in the dataset, we cleaned the abstract text by removing phrases that were not relevant to the specifics of the projects (e.g. generic phrases such as 'description (provided by applicant)' and 'end of abstract'), which were generally found at the beginning or end of many abstracts. We then used standard NLP techniques of tokenization, lemmatization, stop word removal, and the addition of bi-grams and tri-grams on the abstracts, and removed single character tokens and numeric tokens that were not length four (e.g. years) from the abstracts. For more details on the processing of the data, please see the code.[8]

The resulting processed dataset contained 1,143,869 projects. Figure 1 shows the distribution of these projects by funding agency and FY. Most projects in the processed dataset are funded by HHS (81.1%) and the NSF (12.5%). HHS houses NIH, which is responsible for funding 98.0% of the HHS projects. 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).

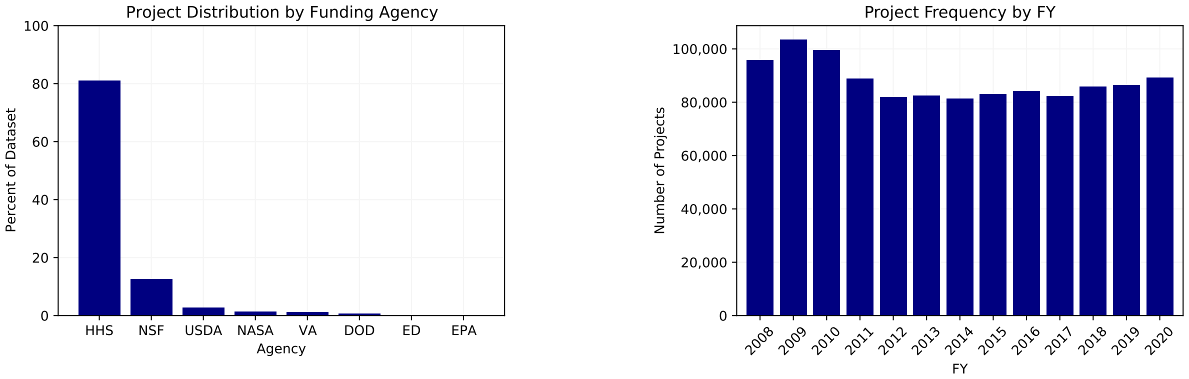

Figure 1: Federal RePORTER projects funded in FY 2008-2020: Processed dataset. Distributions by funding agency and FY.

4 Broad Topics and Trends

4.1 Topic Modeling

To automatically discover broad topics of federally funded R&D, we utilized topic modeling on the project abstracts from the processed Federal RePORTER dataset. Given a set of documents, topic models produce lists of terms that should be thematically related (i.e. the topics) and also the distribution of topics within each document. Topic term lists are ordered by term importance to the topic. There are different topic modeling algorithms available; we chose two widely-used stochastic algorithms—LDA (Blei et al., 2003) and NMF (Lee & Seung, 1999)—to find the best model for reporting broad topics of federally funded R&D. Both LDA and NMF can assign multiple topics to a document and can assign the same term to multiple topics; these are reasonable assumptions

[8] For more details see the 01-Cleaning, 02-lemmatize, and 03-Processing code at https://github.com/uva-bi-sdad/publicrd/tree/master/src/Paper/01_wrangling.

given our data source of project abstracts. The number of topics to be modeled is a user-chosen parameter for both algorithms.

LDA is a generative probabilistic model that assumes that documents are mixtures of latent topics, where each topic is represented by a multinomial distribution over the set of words (Blei et al., 2003). The multinomial distribution for the topics is parameterized by a Dirichlet random variable that is parameterized by α, a tuning parameter. In addition, each word is sampled from a multinomial distribution conditioned on the topic and parameterized by η, another tuning parameter. Larger values of α and η indicate that each document and topic contains a mixture of most of the topics and words, respectively, and vice versa. The input to LDA is an $m \times n$ document-term matrix for a corpus where entry (i, j) contains the frequency of term j in document i. LDA then uses the document-term matrix to perform inference using an online variational Bayes algorithm and estimates the parameters to characterize the latent topics.

NMF utilizes iterative optimization to approximately factor a matrix $\mathbf{A}$ as

$$\underset{m\times n}{\mathbf{A}} \approx \underset{m\times k}{\mathbf{W}}\ \underset{k\times n}{\mathbf{H}},$$

where all entries of $\mathbf{A}$, $\mathbf{W}$, and $\mathbf{H}$ are nonnegative. In the context of topic modeling, $\mathbf{A}$, the input to the algorithm, is the term frequency-inverse document frequency (TFIDF) weighting of the document-term matrix. Entries of $\mathbf{A}$ are computed using the formula below and then the vector 2-norm is used to normalize each row of $\mathbf{A}$.

$$\mathbf{A}(i, j) = f_{ij} \times \left[\log\left(\frac{1+m}{1+d_j}\right) + 1\right],$$

where $f_{i,j}$ is the frequency of term j in document i (i.e., the (i, j) element of the document-term matrix), m is the count of documents in the corpus, and d_j is the count of documents in the corpus in which term j is used (scikit-learn developers, n.d.). The TFIDF weighting has the effect of penalizing terms that occur very frequently in many documents of the corpus. The matrices $\mathbf{W}$ and $\mathbf{H}$ give document-topic and topic-term relationships, respectively. Specifically, $\mathbf{W}(i, j)$ contains the weight of topic j in document i, and $\mathbf{H}(i, j)$ contains the weight of term j in topic i. The parameter k is the number of topics chosen for the model.

We compared LDA and NMF at varying numbers of topics by performing ten runs of LDA and NMF for each number of topics. We reported the model C_V topic coherence (Röder et al., 2015) score for each run. For each topic, C_V topic coherence encodes how often the top w topic words appear together in close proximity within the documents as well as semantic information. The topic coherences are averaged to provide a score for the model. 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). We used $w = 10$ in our calculations. In addition, we filtered the available terms for each 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', 'aim', 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.

The results of our topic model comparisons 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 implementation of LDA that ran on 40 cores; NMF ran serially.[9] We computed the LDA and NMF models with 5 to 50 topics at 5 topic intervals (e.g., 5, 10, 15, ..., 50), and chose common values for the LDA model parameters for the document-topic and topic-term distribution priors: $\alpha = 1/N$, where N was the number of topics, and $\eta = 0.1$.

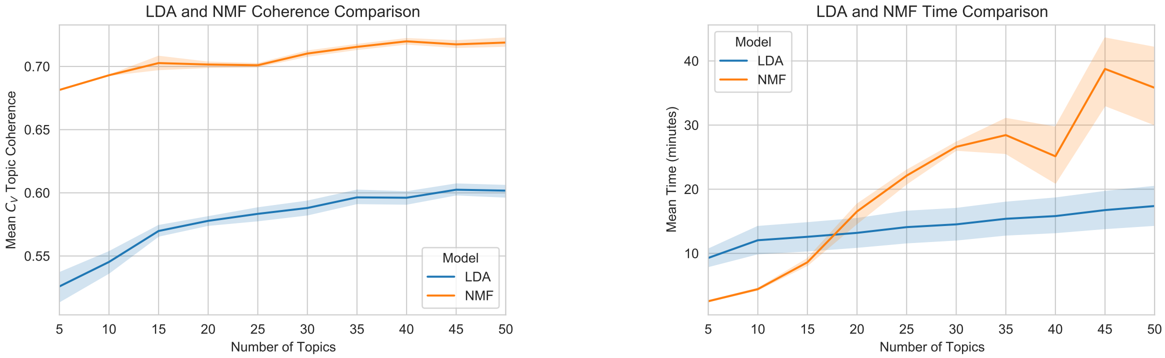

Figure 2: Federal RePORTER projects funded in FY 2008-2020: Processed dataset. Topic model mean C_V coherence score and run time. The shaded region gives a 95% confidence interval on the mean. LDA and NMF were tested with 10 runs each at 5, 10, 15, ..., 45 and 50 topics.

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 and explore models with 20 and 50 topics. In this case, we did not choose the number of topics that maximized mean coherence 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 or more specific topics.

Topics from a 20-topic and a 50-topic NMF model are given in Tables 2 and 3, respectively. The five highest weighted terms per topic are listed, in order of weight starting with the highest weighted term. The 20-topic model results are presented with lower case fr topic labels and the 50-topic model results are presented with uppercase FR topic labels. These models each show a broad categorization of R&D topics present in Federal RePORTER with the 50-topic model obviously showing more specific topics than the 20-topic model. The topic coherence for each of these models is given in Table 1.

Number of Topics	C_V **Topic Coherence**
20	0.70
50	0.72

Table 1: Federal RePORTER projects funded in FY 2008-2020: Processed dataset. C_V topic coherence scores for NMF topic models.

[9] Throughout this work we predominantly used the Python library scikit-learn (Pedregosa et al., 2011) for NLP, machine learning, and information retrieval methods. To a lesser extent we used the libraries Gensim (Řehůřek & Sojka, 2010) and spaCy (Honnibal et al., 2020) for some of the NLP tasks. We note that the C_V coherence scores were calculated using Gensim.

Label	Top Five Terms
fr1	ad, cognitive, disease, alzheimer, brain
fr2	alcohol, ethanol, drinking, use, consumption
fr3	brain, neuron, neural, circuit, synaptic
fr4	cancer, breast, prostate, woman, risk
fr5	cell, stem, differentiation, tissue, signal
fr6	child, risk, intervention, age, parent
fr7	core, administrative, center, investigator, support
fr8	drug, compound, target, inhibitor, resistance
fr9	gene, genetic, genome, dna, expression
fr10	health, community, care, disparity, intervention
fr11	hiv, aids, infect, infection, prevention
fr12	infection, vaccine, virus, immune, response
fr13	lung, asthma, airway, pulmonary, injury
fr14	mouse, signal, insulin, stress, mechanism
fr15	patient, clinical, trial, treatment, care
fr16	protein, membrane, structure, bind, complex
fr17	student, science, program, graduate, school
fr18	system, datum, method, develop, technology
fr19	training, program, trainee, faculty, mentor
fr20	tumor, metastasis, therapy, metastatic, anti

Table 2: Federal RePORTER projects funded in FY 2008-2020: Processed dataset. Top five topic terms from NMF model with 20 topics. Topics are listed and labeled in alphabetical order by the highest weighted topic term.

Label	Top Five Terms
FR1	ad, alzheimer, tau, amyloid, dementia
FR2	age, cognitive, aging, old, memory
FR3	alcohol, ethanol, drinking, use, consumption
FR4	bone, fracture, osteoporosis, osteoblast, skeletal
FR5	brain, injury, tbi, stroke, neural
FR6	breast, cancer, woman, metastasis, estrogen
FR7	cancer, pancreatic, ovarian, nci, colorectal
FR8	cell, antigen, differentiation, type, cd4
FR9	center, support, resource, investigator, pilot
FR10	child, parent, language, family, asd
FR11	clinical, trial, phase, translational, protocol
FR12	conference, meeting, workshop, researcher, field
FR13	core, administrative, provide, investigator, program
FR14	datum, analysis, data, statistical, method
FR15	disease, kidney, renal, liver, progression
FR16	dna, repair, damage, replication, methylation
FR17	dr, career, mentor, award, independent
FR18	drug, compound, target, inhibitor, cocaine
FR19	food, safety, product, animal, fda
FR20	gene, expression, transcription, rna, regulatory
FR21	genetic, variant, genome, variation, genomic
FR22	health, community, disparity, care, public
FR23	heart, cardiac, failure, vascular, cardiovascular
FR24	hiv, aids, infect, infection, prevention
FR25	imaging, image, mri, resolution, pet
FR26	infection, virus, viral, host, hiv_1
FR27	insulin, glucose, resistance, diabetes, diabete
FR28	intervention, behavior, treatment, adolescent, youth
FR29	lung, asthma, airway, pulmonary, injury
FR30	material, device, energy, technology, chemical
FR31	mitochondrial, mitochondria, ros, dysfunction, oxidative
FR32	mouse, model, animal, human, transgenic
FR33	network, model, system, problem, computational
FR34	neuron, synaptic, circuit, neural, neuronal
FR35	obesity, weight, metabolic, diet, fat
FR36	pain, chronic, opioid, analgesic, treatment
FR37	patient, care, treatment, outcome, quality
FR38	plant, water, climate, change, soil
FR39	prostate, cancer, ar, pca, androgen
FR40	protein, membrane, structure, bind, complex
FR41	risk, woman, exposure, factor, pregnancy
FR42	signal, receptor, pathway, activation, regulate
FR43	sleep, circadian, disorder, insomnia, disturbance
FR44	stem, hsc, progenitor, teacher, differentiation
FR45	stress, response, oxidative, ptsd, er
FR46	student, science, program, school, graduate
FR47	tissue, muscle, injury, liver, regeneration
FR48	training, program, trainee, faculty, year
FR49	tumor, therapy, metastasis, anti, metastatic
FR50	vaccine, immune, response, antigen, antibody

Table 3: Federal RePORTER projects funded in FY 2008-2020: Processed dataset. Top five topic terms from NMF model with 50 topics. Topics are listed and labeled in alphabetical order by the highest weighted topic term.

Most of the topics produced by both models are health related, which is to be expected since grant abstracts from HHS comprise 81.1% of our processed dataset. We also see some topics about student or other training programs, conferences, and administrative tasks (fr7, fr17, fr19, FR9, FR12, FR13, FR46, and FR48). Some general topics from the 20-topic model appear as multiple, distinct topics in the 50-topic model, e.g. while there is a breast and prostate cancer topic (fr4) in the 20-topic model, there are separate breast (FR6) and prostate cancer (FR39) topics in the 50-topic model. In the 50-topic model we also see computational topics such as data analysis (FR14), medical imaging (FR25), and network models (FR33), and agriculture/environmental topics: food safety (FR19) and plant care (FR38). It is interesting to note that in the 20-topic model, Alzheimer's disease (fr1), cancer (fr4, fr20), and HIV/AIDS (fr11) each appear as distinct topics signaling the prominence of these topics in R&D projects reported in Federal RePORTER.

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 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 different runs of the model. To quantify the extent of this instability, we computed three measures proposed by 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 and TS and PS values closer to 1 represent more stability. Stability results for each topic model are given in Table 4 and indicate that the topic models are relatively stable.

Number of Topics	**DSD**	**TS**	**PS**
20	0.06	0.92	0.91
50	0.16	0.74	0.70

Table 4: Stability measures for NMF topic models on processed Federal RePORTER project abstracts funded in FY 2008-2020. DSD, TS, and PS are given as average measures across $r = 10$ runs utilizing 10 terms to describe the topics.

4.2 Topic Trends

To analyze topic trends and more readily compare topic prevalence, we examined document-topic weights over time, modeling our approach after Griffiths and Steyvers (2004). Specifically, for each NMF model we used the matrix $\mathbf{W}$ to obtain the topic weights for each abstract and calculated mean topic weight per topic per FY. We calculated mean topic weight for each FY using abstracts funded in that particular FY. The relationship between the variables mean weight and FY for each topic was modeled using linear regression, thus capturing the trends of the topic weights over time.

In the following discussion, we use the size and sign of the estimated slope of the regression line to characterize the prevalence of each topic. However, we do not dichotomize topics into "hot" or "cold" topics as is done by Griffiths and Steyvers (2004). Topic trends for the 20-topic NMF model are presented in Figure 3 and information per topic is presented in Table 5, namely the top five words, count of abstracts that contain the topic, and the estimated slope of the regression line relating mean weight and FY and its associated p-value. Trend results are set up similarly for the 50-topic NMF model and are presented in Appendix C, Table 11 and Figures 7 and 8.

Figure 3 displays multiple measures of topic prevalence, namely mean topic weight, the number of abstracts a topic appears in (n), and the slope of the regression line relating mean topic weight and FY. We can use all of these measures to form a more complete picture of topic trends. For example, the Alzheimer's disease topic (fr1) has the largest positive regression line slope meaning that it is the topic that is increasing in mean topic weight the fastest over 2008-2020. However, its mean topic weight is relatively low compared to most other topics and it only appears in $386,955$ abstracts, the third lowest value of n. Another observation is that topics with a negatively sloped

Label	n	Slope (x100)	p-value	Top Five Terms
fr1	386,955	0.0067	0.0000	ad, cognitive, disease, alzheimer, brain
fr2	412,892	-0.0004	0.1601	alcohol, ethanol, drinking, use, consumption
fr3	557,838	0.0016	0.0002	brain, neuron, neural, circuit, synaptic
fr4	473,102	-0.0007	0.4587	cancer, breast, prostate, woman, risk
fr5	591,781	-0.0010	0.0033	cell, stem, differentiation, tissue, signal
fr6	536,165	0.0001	0.8229	child, risk, intervention, age, parent
fr7	520,675	0.0051	0.0038	core, administrative, center, investigator, support
fr8	582,112	0.0004	0.4603	drug, compound, target, inhibitor, resistance
fr9	606,450	-0.0036	0.0000	gene, genetic, genome, dna, expression
fr10	587,322	0.0019	0.1853	health, community, care, disparity, intervention
fr11	368,173	0.0014	0.0008	hiv, aids, infect, infection, prevention
fr12	489,113	-0.0002	0.7581	infection, vaccine, virus, immune, response
fr13	362,964	0.0006	0.0383	lung, asthma, airway, pulmonary, injury
fr14	687,591	-0.0040	0.0005	mouse, signal, insulin, stress, mechanism
fr15	611,375	0.0061	0.0000	patient, clinical, trial, treatment, care
fr16	610,846	-0.0107	0.0000	protein, membrane, structure, bind, complex
fr17	478,818	-0.0010	0.3087	student, science, program, graduate, school
fr18	755,750	0.0017	0.0551	system, datum, method, develop, technology
fr19	537,548	0.0017	0.1853	training, program, trainee, faculty, mentor
fr20	422,008	0.0005	0.0995	tumor, metastasis, therapy, metastatic, anti

Table 5: Federal RePORTER projects funded in FY 2008-2020: Processed dataset. Topic trend results produced by a 20-topic NMF model. Topics are listed and labeled in alphabetical order by the most important topic term. The number of abstracts that a topic appears in is given by n and Slope and p-value refer to the regression line relating FY and mean topic weight. Slopes are multiplied by 100 for easier viewing.

regression line can still be very prevalent and a topic of high importance in the corpus. For example, the insulin signaling topic (fr14) has a negatively sloped trend line, but has the highest mean topic weights over all years analyzed and is present in $687,591$ abstracts, the second highest value of n. Additionally, the two other topics with larger negative trend line slopes, genetics (fr9) and membrane proteins (fr16) also exhibit fairly high mean topic weights and high values of n.

The data systems topic (fr18) appears in the highest number of abstracts ($755,750$) and has higher mean topic weights than most other topics for all years 2008-2020, signaling its importance as a topic in the corpus. Other trends of note: training and student programs (fr19 and fr17) are well represented in the corpus, and more than half of the topics have steady mean topic weights over 2008-2020 as with HIV/AIDS (fr11), pulmonary issues (fr13), and alcohol use (fr2). Specifically, these three topics experience steady, low mean topic weights.

5 Pandemic-Related Topics and Trends

We now address how to find topics and trends related to pandemics within Federal RePORTER. Given how broad the topics discovered in Section 4 were, we were not able to use the entire Federal RePORTER corpus to characterize pandemic-specific topics because this was too small a focal area to detect within the larger corpus. Thus, we used two information retrieval techniques—term matching and latent semantic indexing (LSI) (Deerwester et al., 1990)—to select Federal RePORTER project abstracts related to pandemics and then ran a topic trend analysis on these relevant abstracts. Term matching is a common technique for identifying documents relevant to a theme by marking a document as relevant if it contains a particular keyword or set of 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.,

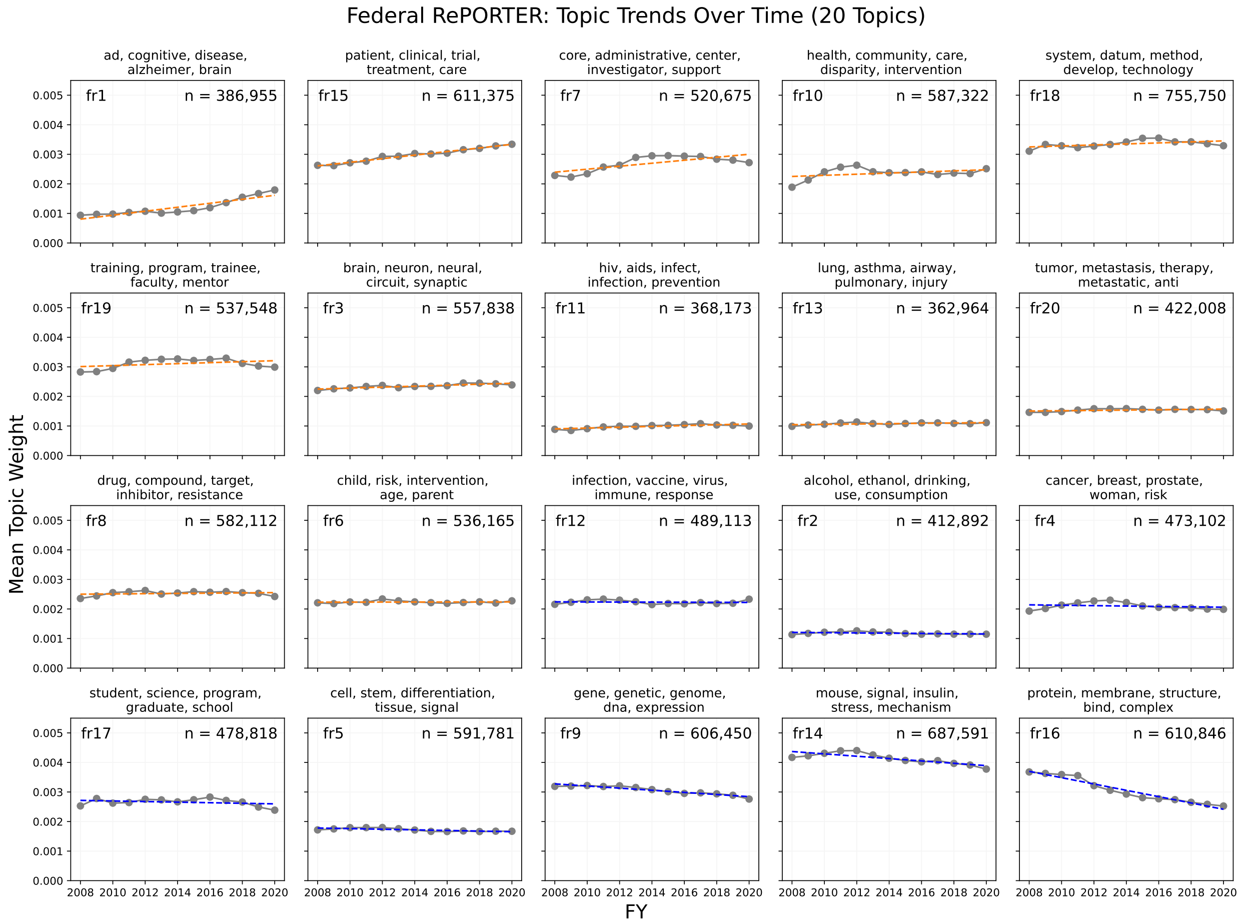

Figure 3: Federal RePORTER projects funded in FY 2008-2020: Processed dataset. Topic trend results produced by a 20-topic NMF model. Topic labels and the number of abstracts containing the topics, n, are given in the upper left and right plot corners respectively. Plots are ordered from largest to smallest 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, but are too small to be visible.

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 used LSI as it can identify theme-relevant documents that may not necessarily contain the keyword(s). This is realized through calculating 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 B. 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".

To create a pandemics-themed corpus, we used term matching and LSI on the abstracts in our processed Federal RePORTER dataset. Our term matching keyword list included any abstract token that contained the word 'pandemic.' The list of 115 keywords can be found in Appendix A. 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 documents scoring 0.7 or above were counted as relevant to the query. The rank and score threshold were chosen by a trial and error process. We inspected abstracts and their scores and created themed corpora for different values of rank and score threshold. We ran a topic model on each themed corpus and chose values for the rank and score threshold based on visual inspection of the topic results for relevance to pandemics, specificity of the topics, and size of the themed corpus. We note that the matrix spectrum and abstract relevance score distributions for various values of the rank 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[10]. See Table 6 for details.

TM	**LSI**	**TM & LSI**	**Total**
4,936	1,788	847	7,571

Table 6: Pandemics Corpus, contribution by information retrieval method. The units for each column are number of abstracts. TM: term matching, LSI: latent sematic indexing, TM & LSI: overlap of abstracts returned by both methods.

The project distributions by funding agency and FY for the pandemics corpus are shown in Figure 4. HHS projects dominate this themed corpus and the distribution by agency looks very similar to the distribution by agency for the entire Federal RePORTER corpus (Figure 1). In 2020, the number of projects is about four times as many projects in each of the years 2008-2019. There are slightly more projects in 2009-2011 than in 2008 and 2012-2019; this is similar to the FY distribution for the entire Federal RePORTER corpus (Figure 1).

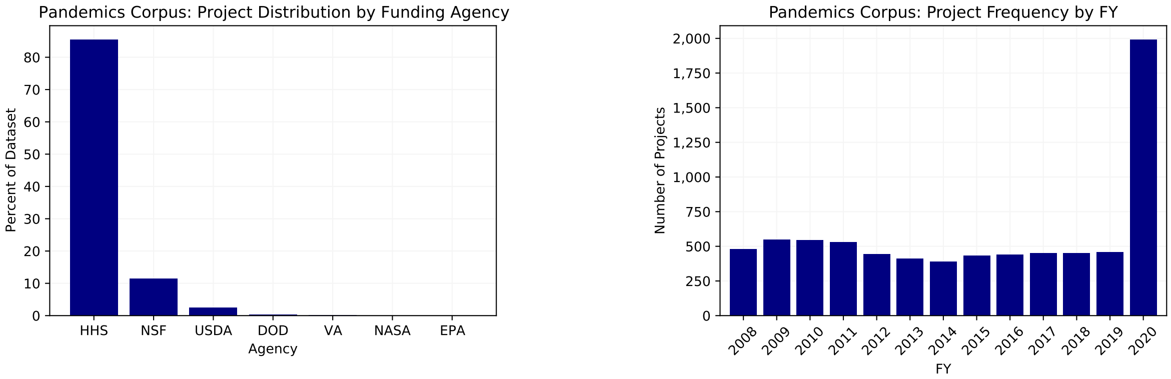

Figure 4: Pandemics corpus, distributions by funding agency and FY.

To identify topic trends in Federal RePORTER within the area of pandemics, we fit a 5-topic NMF model and a 20-topic NMF model on the pandemics corpus. We excluded terms from the models that appeared in less than three abstracts, as well as the terms ‘research’, ‘aim’, and ‘project’ as we had done with the topic models on the entire Federal RePORTER corpus. The 5-topic model presents a broad view of pandemic topics while the 20-topic model presents more specific topics. Both models are quite stable as evident in the measures provided in Table 7.

Results for the 5-topic model and its corresponding topic trend analysis are given in Table 8 and Figure 5. Topic labels are given with a lower case p for this model. The five topics are COVID-19

[10] Every abstract receives a score in the LSI process. If we had lowered the score threshold, we would have included more abstracts in the themed corpus and this overlap would have been larger. In fact, 75% of the documents marked as relevant by term matching are in the top 10.40% of LSI relevance scores.

Model	DSD	TS	PS
5-topic NMF	0	1	1
20-topic NMF	0.11	0.84	0.89

Table 7: Pandemics Corpus, model stability. DSD, TS, and PS are average measures computed using $r = 20$ runs and 10 terms to describe the topics.

(p1), HIV/AIDS (p2), influenza (p3), and general topics on vaccines (p4) and viruses (p5). We notice a large increase in mean topic weight from 2019-2020 for the COVID-19 topic (p1), but yet a fairly large decrease from 2019-2020 for each of the other four topics. The trend from 2019-2020 also has a large effect on the regression line slope for each topic; for example the COVID-19 (p1) and influenza (p3) topics have fairly steady mean topic weights between 2008-2019, but their regression lines reflect their respective increase or decrease in mean topic weight from 2019-2020.

Label	n	Slope (x100)	p-value	Top Five Terms
p1	5,766	0.1038	0.1733	covid_19, health, datum, disease, community
p2	3,825	-0.0598	0.0110	hiv, aids, hiv_1, drug, cell
p3	4,563	-0.0493	0.1694	influenza, virus, vaccination, protection, immune
p4	5,120	-0.0570	0.0506	vaccine, antibody, virus, protective, response
p5	5,922	-0.0561	0.0328	virus, viral, cell, host, infection

Table 8: Pandemics corpus topic trend results produced by a 5-topic NMF model. Topics are listed and labeled in alphabetical order by the most important topic term. The number of abstracts that a topic appears in is given by n and Slope and p-value refer to the regression line relating FY and mean topic weight. Slopes are multiplied by 100 for easier viewing.

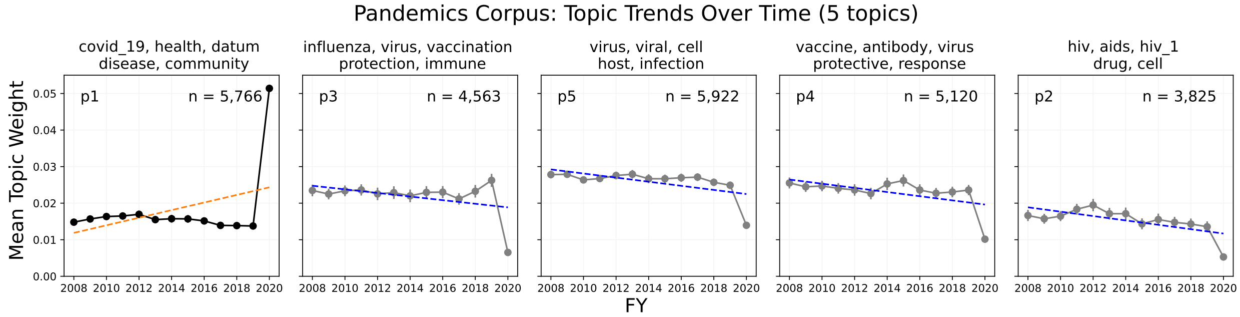

Figure 5: Pandemics corpus topic trend results produced by a 5-topic NMF model. Topic labels and the number of abstracts containing the topics, n, are given in the upper left and right plot corners respectively. Plots are ordered from largest to smallest regression line slope; orange lines have a positive slope and blue lines have a negative slope. Standard errors on the means are represented on each plot using error bars. Topics plotted with a black line are those that experienced a mean topic weight increase from 2019-2020; those that experienced a decrease are plotted in gray.

We see that the viruses topic (p5) generally has the largest mean topic weight from 2008-2019, and the influenza (p3) and vaccines (p4) topics also have higher mean topics weights during this time compared to those of the COVID-19 (p1) and HIV/AIDS topics (p2). This signals the importance of the topics of influenza (p3), vaccines (p4), and viruses (p5) in the corpus during 2008-2019. The HIV/AIDS topic (p2) appears in the least number of documents (3,825), but has a fairly similar mean topic weight to that of the COVID-19 topic (p1) until 2020. While in 2008-2019 the viruses topic (p5) achieves about double the mean topic weight of the COVID-19 topic (p1), the COVID-19 topic (p1) appears in 5,766 abstracts which is almost as many abstracts as the viruses topic (p5) (5,922). The general trends that are present in each topic from 2008-2019 change in 2020 with the large increase in research focusing on COVID-19, when COVID-19 is arguably the most important topic in the corpus.

Results for the 20-topic model and its corresponding topic trend analysis are given in Table 9 and Figure 6, and topic labels are given with an upper case P. These topics are more specific than those of the 5-topic model and include influenza (P10, P11), HIV/AIDS (P8, P9), and COVID-19 (P14, P15) as well as other viruses such as Ebola (P7) and Zika (P20), and the obesity pandemic (P12). Similar to the 5-topic model results, we see that only a few topics had increasing mean topic weights from 2019-2020; these are COVID-19 infections (P14), COVID-19 social implications (P15), cancer and Kaposi's sarcoma-associated herpes virus (P3), student training (P16), and assay detection technology (P5), where assay detection technology (P5) likely also has a relation to COVID-19 testing. The increases are very large for the COVID-19 infections (P14) and COVID-19 social implications (P15) topics. All other topics had decreasing mean topic weights from 2019-2020 with the influenza vaccination topic (P11) having the steepest decline. Despite being a far more specific topic, COVID-19 social implications (P15) appears in 4,345 abstracts in the themed corpus which is almost as many as the general viruses topic (P19) (4,368).

Label	n	Slope (x100)	p-Value	Top Five Terms
P1	3,471	-0.0121	0.0995	animal, bird, contact, surveillance, close
P2	3,230	0.0267	0.0969	antibody, epitope, ha, immunogen, conserve
P3	2,706	0.0333	0.0039	cancer, patient, care, kshv, treatment
P4	3,815	-0.0327	0.0188	cell, response, memory, cd4, immunity
P5	4,090	-0.0206	0.1241	diagnostic, detection, technology, sample, assay
P6	3,579	-0.0281	0.0019	drug, inhibitor, resistance, compound, antiviral
P7	2,661	0.0169	0.3402	ebola, virus, outbreak, gp, filovirus
P8	3,185	-0.0423	0.0139	hiv, aids, prevention, trial, infection
P9	2,526	-0.0490	0.0000	hiv_1, env, subtype, transmission, shiv
P10	2,483	0.0586	0.0086	iav, lung, host, sp, evolution
P11	3,431	-0.0504	0.0822	influenza, vaccination, strain, effectiveness, virus
P12	3,918	-0.0006	0.9672	obesity, disease, infection, mtb, tb
P13	3,442	0.0035	0.6808	protection, ecologic, immune, evolution, pathogenicity
P14	2,913	0.0744	0.1244	sars_cov_2, covid_19, patient, infection, coronavirus
P15	4,345	0.0617	0.1773	social, covid_19, health, datum, community
P16	3,489	-0.0015	0.8971	training, program, student, trainee, faculty
P17	4,217	-0.0656	0.0057	vaccine, candidate, adjuvant, efficacy, protection
P18	3,811	-0.0179	0.1627	viral, protein, rna, host, interaction
P19	4,368	-0.0719	0.0000	virus, human, infection, genetic, transmission
P20	2,267	0.0414	0.1073	zikv, dengue, zika, mosquito, flavivirus

Table 9: Pandemics corpus topic trend results produced by a 20-topic NMF model. Topics are listed and labeled in alphabetical order by the most important topic term. The number of abstracts that a topic appears in is given by n and Slope and p-value refer to the regression line relating FY and mean topic weight. Slopes are multiplied by 100 for easier viewing.

As in the 5-topic model results, the general trends present in most topics from 2008-2019 change drastically in 2020 with the large increase in COVID-19 research. For example the influenza A virus topic (P10) research was steadily increasing until 2020 when it significantly decreased. In addition, from 2008-2019 the HIV/AIDS topic (P8) has a mean topic weight that is larger than that of the COVID-19 infections (P14) and COVID-19 social implications (P15) topics; however, in 2020 the mean topic weight of the HIV/AIDS topic (P8) is much lower than that of the COVID-19 infections (P14) or COVID-19 social implications (P15) topics. Of note is that the COVID-19 infections topic (P14) has the lowest mean topic weight of all topics from 2008-2019, before experiencing the

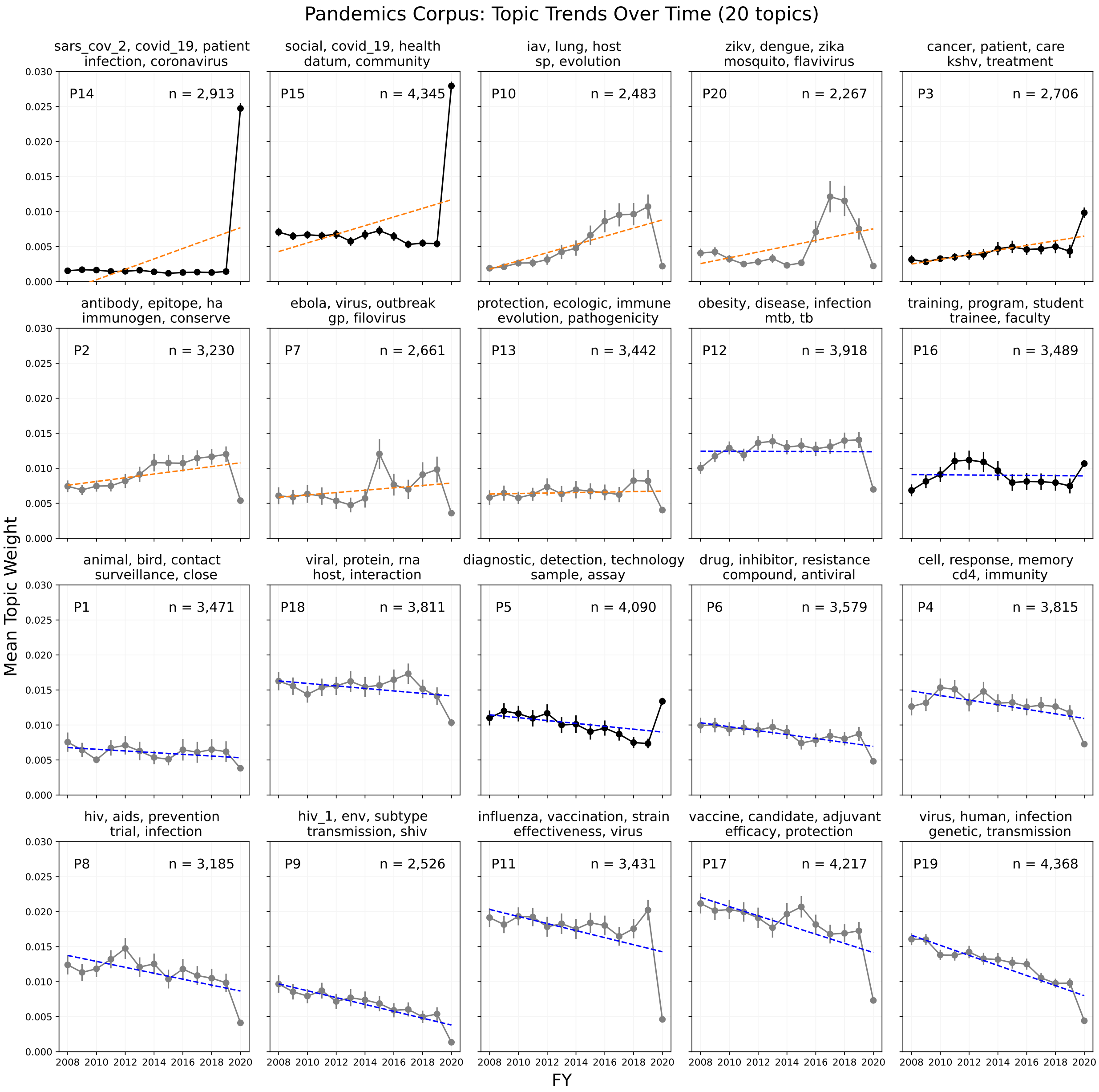

Figure 6: Pandemics corpus topic trend results produced by a 20-topic NMF model. Topic labels and the number of abstracts containing the topics, n, are given in the upper left and right plot corners respectively. Plots are ordered from largest to smallest regression line slope; orange lines have a positive slope and blue lines have a negative slope. Standard errors on the means are represented on each plot using error bars. Topics plotted with a black line are those that experienced a mean topic weight increase from 2019-2020; those that experienced a decrease are plotted in gray.

steepest increase and second highest mean topic weight in 2020. (The highest mean topic weight in 2020 is the COVID-19 social implications topic (P15).)

While COVID-19 related topics dominate recent years, other topics exhibit trends that likely reflect similar outbreaks, although at a smaller scale. For example, the Zika virus topic (P20) experiences a large increase in mean topic weight from 2015 to 2017, which follows after the 2015-2016 Zika outbreak in North and South America (Division of Vector-Borne Diseases [DVBD], n.d.), and the Ebola virus topic (P7) experiences a spike in mean topic weight in 2015, following the beginning of the Ebola outbreak in West Africa in 2014-2016 (Viral Special Pathogens Branch [VSPB], n.d.). While we cannot be certain that these past events caused the trends in the Zika virus (P20) and Ebola virus (P7) topics, there does seem to be a reasonable connection.

Our findings that federally funded COVID-19 research greatly increased in prevalence in 2020 while most other research topics experienced a significant decrease in prevalence is consistent with findings by Raynaud et al. (2021). They manually investigated "high-impact medical journals" and discovered that "the dramatic rise in COVID-19 publications was accompanied by a substantial decrease of non-COVID-19 research." Our findings are also consistent with a statement describing 2020 research by Callaway et al. (2020): "In many fields not directly related to the pandemic, projects and progress slowed to a crawl." Through the use of machine learning and information retrieval we were able to visualize these trends.

6 Conclusion and Future Work

Federally funded R&D topics are identified through the use of NLP and machine learning, specifically NMF topic modeling on grant abstracts found in Federal RePORTER. In addition, topics related to pandemics are presented, which we found using information retrieval and NMF topic modeling. Topic trends over time are also shown. Since Federal RePORTER is a typical example of a scientific award database, our methodology can be applied to other scientific award databases as well.

In considering the larger implications of the project, we recognize that our data included the majority but not all federally funded grants within the U.S. Nor does it capture the full scope of non-government funded R&D within the U.S., much less R&D funding 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 themed topic trend analyses approach to themes that are complex, multi-faceted, and difficult to define, such as "artificial intelligence" or "bioeconomy." This could include extending the list of theme keywords, using expert input, or employing methods such as word2vec. 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 can also 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 also be focusing on gathering new data from agency specific sources such as NIH RePORTER and NSF Award Search with data from FYs after 2020. We will test current and future approaches on this new dataset. We have begun working on the theme of artificial intelligence and will continue to analyze this as well as pandemic funded R&D projects. We will also explore new themes such as the bioeconomy. We believe that the methods described in this paper show promise to supplement the information currently collected through the NCSES FFS and FSS by providing information that the surveys do not collect.

7 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 (BI), 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 Samantha Cohen, a former SDAD Postdoctoral Research Associate, for contributions to the text cleaning and processing work. We thank Sallie Keller for helpful discussions and suggestions throughout the project. Dr. Keller is the Distinguished Professor in Biocomplexity, Division Director, SDAD, BI, Professor of Public Health Sciences, University of Virginia. She is currently on an IPA with the US Census Bureau.

References

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.

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

Callaway, E., Ledford, H., Viglione, G., Watson, T., & Witze, A. (2020). COVID and 2020: An extraordinary year for science. *Nature*, *588*, 550–552. https://doi.org/10.1038/d41586-020-03437-4

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.

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

Honnibal, M., Montani, I., Van Landeghem, S., & Adriane, B. (2020). spaCy: Industrial-strength Natural Language Processing in Python. doi: 10.5281/zenodo.1212303.

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.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, 3111–3119.

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. (2021a). *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/.

National Center for Science and Engineering Statistics. (2021b). U.S. R&D Increased by $51 Billion, to $606 Billion, in 2018; Estimate for 2019 Indicates a Further Rise to $656 Billion, National Center for Science and Engineering Statistics (NCSES).

National Center for Science and Engineering Statistics. (2022). *Federal Funds for Research and Development: Fiscal Years 2020–21*, NSF 22–323. Alexandria, VA: National Science Foundation, https://ncses.nsf.gov/pubs/nsf22323/.

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, OECD 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 (OECD), Directorate for Science, Technology, and Innovation, Committee for Scientific and Technological Policy.

Pece, C. V. (2016). Putting the Cart Before a Lame Horse: A Case Study for Future Initiatives to Automate the Use of Administrative Records for Reporting Government R&D, National Center for Science and Engineering Statistics, National Science Foundation.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

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.

Raynaud, M., Goutaudier, V., Louis, K., Al-Awadhi, S., Dubourg, Q., Truchot, A., Brousse, R., Saleh, N., Giarraputo, A., Debiais, C., Demir, Z., Certain, A., Tacafred, F., Cortes-Garcia, E., Yanes, S., Dagobert, J., Naser, S., Robin, B., Bailly, E., ... Loupy, A. (2021). Impact of the COVID-19 pandemic on publication dynamics and non-COVID-19 research production. *BMC Medical Research Methodology*, *21*. https://doi.org/10.1186/s12874-021-01404-9

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora [http://is.muni.cz/publication/884893/en]. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

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

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

scikit-learn developers. (n.d.). *6.2. Feature extraction*. Retrieved March 18, 2022 from https://scikit-learn.org/stable/modules/feature_extraction.html.

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.

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

U.S. Department of Health and Human Services. (2020, March 6-b). *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

Viral Special Pathogens Branch. (n.d.). *2014-2016 Ebola Outbreak in West Africa*. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), Division of High-Consequence Pathogens and Pathology (DHCPP). Retrieved February 3, 2022 from https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak/index.html.

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.

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.

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.

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 Pandemic Keywords

There are 115 pandemic keywords.

1918_influenza_pandemic
1918_pandemic
1957_1968_pandemic
2009_pandemic_h1n1
aidspandemic
andpandemic
anotherpandemic
apandemic
assesspandemic
betweenpandemic
bothpandemic
causedpandemic
chikvpandemic
cov_2pandemic
covid19_pandemic
covid19pandemic
covid_19_pandemic
covid_19pandemic
covid_19pandemic5
covid_pandemic
covidpandemic
criticalpandemic
currentpandemic
detetermrminineif*
devastatingpandemic
ebolapandemic
emergingpandemic
escalatingpandemic
establishingpandemic
forpandemic
frompandemic
futurepandemic
globalpandemic
greatpandemic2
growingpandemic
h1n1_pandemic
h1n1pandemic
hivpandemic
howpandemic
humanpandemic
increasedpandemic
influenzapandemic
inpandemic
interpandemic
interpandemic_pandemic
majorpandemic
multiclade_recombinant_pandemic
newpandemic
non_pandemic
occasional_pandemic
occasionalpandemic
ofpandemic
ofseasonal_pandemic
ongoingpandemic
pandemic
pandemic1
pandemic2
pandemic2009
pandemic57499
pandemic_1918
pandemic_1918_1919
pandemic_1957
pandemic_flu
pandemic_h1n1
pandemic_preparedness
pandemic_sobering
pandemically
pandemicand
pandemiccompare
pandemiccovid_19
pandemicdisease
pandemicemergence
pandemicflu
pandemich3n2
pandemicha
pandemichas
pandemichave
pandemichuman
pandemicin
pandemicinfection
pandemicinfluence
pandemicinfluenza
pandemicinvolve
pandemiclike
pandemicon
pandemicpose
pandemicpreparedness
pandemicprogresse
pandemicproportion
pandemicremain
pandemics
pandemics18
pandemicsabstractthe
pandemicsetting
pandemicsh1n1
pandemicstrain
pandemicthat
pandemicthis
pandemicthreat
pandemicvaccine
pandemicwill
pandemicwith
possiblepandemic
prepandemic
prepandemic_vaccination
recurrentpandemic
seasonal_pandemic
severepandemic
thecovid_19_pandemic
thepandemic
thispandemic
threepandemic
understandpandemic
withpandemic
worldwide_pandemic_1957

*term was too long to fit in the table. The full term is detetermrminineififththesuprragenomeofftthepandemicicclonesiis.

Table 10: Pandemics keywords used in the term matching and LSI processes to create the pandemics corpus.

Appendix B LSI Relevance Score Calculation

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

Appendix C Results for 50-topic NMF Model

Label	n	Slope (x100)	p-value	Top Five Terms
FR1	291,024	0.0028	0.0000	ad, alzheimer, tau, amyloid, dementia
FR2	475,598	0.0022	0.0000	age, cognitive, aging, old, memory
FR3	367,059	-0.0003	0.0733	alcohol, ethanol, drinking, use, consumption
FR4	400,762	-0.0006	0.0008	bone, fracture, osteoporosis, osteoblast, skeletal
FR5	399,250	0.0020	0.0000	brain, injury, tbi, stroke, neural
FR6	268,463	-0.0021	0.0000	breast, cancer, woman, metastasis, estrogen
FR7	458,431	0.0005	0.4120	cancer, pancreatic, ovarian, nci, colorectal
FR8	595,139	-0.0005	0.0016	cell, antigen, differentiation, type, cd4
FR9	517,063	0.0022	0.0066	center, support, resource, investigator, pilot
FR10	366,758	-0.0004	0.0019	child, parent, language, family, asd
FR11	524,822	0.0024	0.0000	clinical, trial, phase, translational, protocol
FR12	476,699	-0.0001	0.8135	conference, meeting, workshop, researcher, field
FR13	454,467	0.0036	0.0056	core, administrative, provide, investigator, program
FR14	606,259	0.0030	0.0000	datum, analysis, data, statistical, method
FR15	586,438	0.0005	0.6162	disease, kidney, renal, liver, progression
FR16	407,927	-0.0009	0.0001	dna, repair, damage, replication, methylation
FR17	464,272	0.0029	0.0000	dr, career, mentor, award, independent
FR18	510,914	0.0002	0.5022	drug, compound, target, inhibitor, cocaine
FR19	420,317	0.0002	0.8210	food, safety, product, animal, fda
FR20	510,413	-0.0048	0.0000	gene, expression, transcription, rna, regulatory
FR21	522,732	0.0008	0.0086	genetic, variant, genome, variation, genomic
FR22	510,583	0.0000	0.9704	health, community, disparity, care, public
FR23	400,150	-0.0014	0.0065	heart, cardiac, failure, vascular, cardiovascular
FR24	325,533	0.0008	0.0013	hiv, aids, infect, infection, prevention
FR25	474,396	-0.0002	0.1535	imaging, image, mri, resolution, pet
FR26	375,919	0.0002	0.6354	infection, virus, viral, host, hiv_1
FR27	346,995	-0.0015	0.0000	insulin, glucose, resistance, diabetes, diabete
FR28	521,049	0.0025	0.0002	intervention, behavior, treatment, adolescent, youth
FR29	313,854	0.0003	0.0525	lung, asthma, airway, pulmonary, injury
FR30	561,205	-0.0005	0.1554	material, device, energy, technology, chemical
FR31	379,045	0.0011	0.0000	mitochondrial, mitochondria, ros, dysfunction, oxidative
FR32	576,549	-0.0006	0.0004	mouse, model, animal, human, transgenic
FR33	606,146	0.0015	0.0000	network, model, system, problem, computational
FR34	463,695	0.0007	0.0003	neuron, synaptic, circuit, neural, neuronal
FR35	384,442	0.0002	0.2830	obesity, weight, metabolic, diet, fat
FR36	365,548	0.0011	0.0000	pain, chronic, opioid, analgesic, treatment
FR37	505,545	0.0028	0.0000	patient, care, treatment, outcome, quality
FR38	504,999	-0.0027	0.0050	plant, water, climate, change, soil
FR39	332,450	-0.0022	0.0000	prostate, cancer, ar, pca, androgen
FR40	562,142	-0.0054	0.0000	protein, membrane, structure, bind, complex
FR41	540,128	0.0007	0.0093	risk, woman, exposure, factor, pregnancy
FR42	616,862	-0.0027	0.0000	signal, receptor, pathway, activation, regulate
FR43	388,414	0.0013	0.0000	sleep, circadian, disorder, insomnia, disturbance
FR44	397,926	0.0031	0.0000	stem, hsc, progenitor, teacher, differentiation
FR45	424,792	-0.0005	0.0037	stress, response, oxidative, ptsd, er
FR46	405,801	-0.0006	0.0706	student, science, program, school, graduate
FR47	460,926	0.0003	0.0457	tissue, muscle, injury, liver, regeneration
FR48	447,303	0.0003	0.7385	training, program, trainee, faculty, year
FR49	429,974	0.0006	0.0112	tumor, therapy, metastasis, anti, metastatic
FR50	397,545	-0.0001	0.4112	vaccine, immune, response, antigen, antibody

Table 11: Federal RePORTER projects funded in FY 2008-2020: Processed dataset. Topic trend results produced by a 50-topic NMF model. Topics are listed and labeled in alphabetical order by the most important topic term. The number of abstracts that a topic appears in is given by n and Slope and p-value refer to the regression line relating FY and mean topic weight. Slopes are multiplied by 100 for easier viewing.

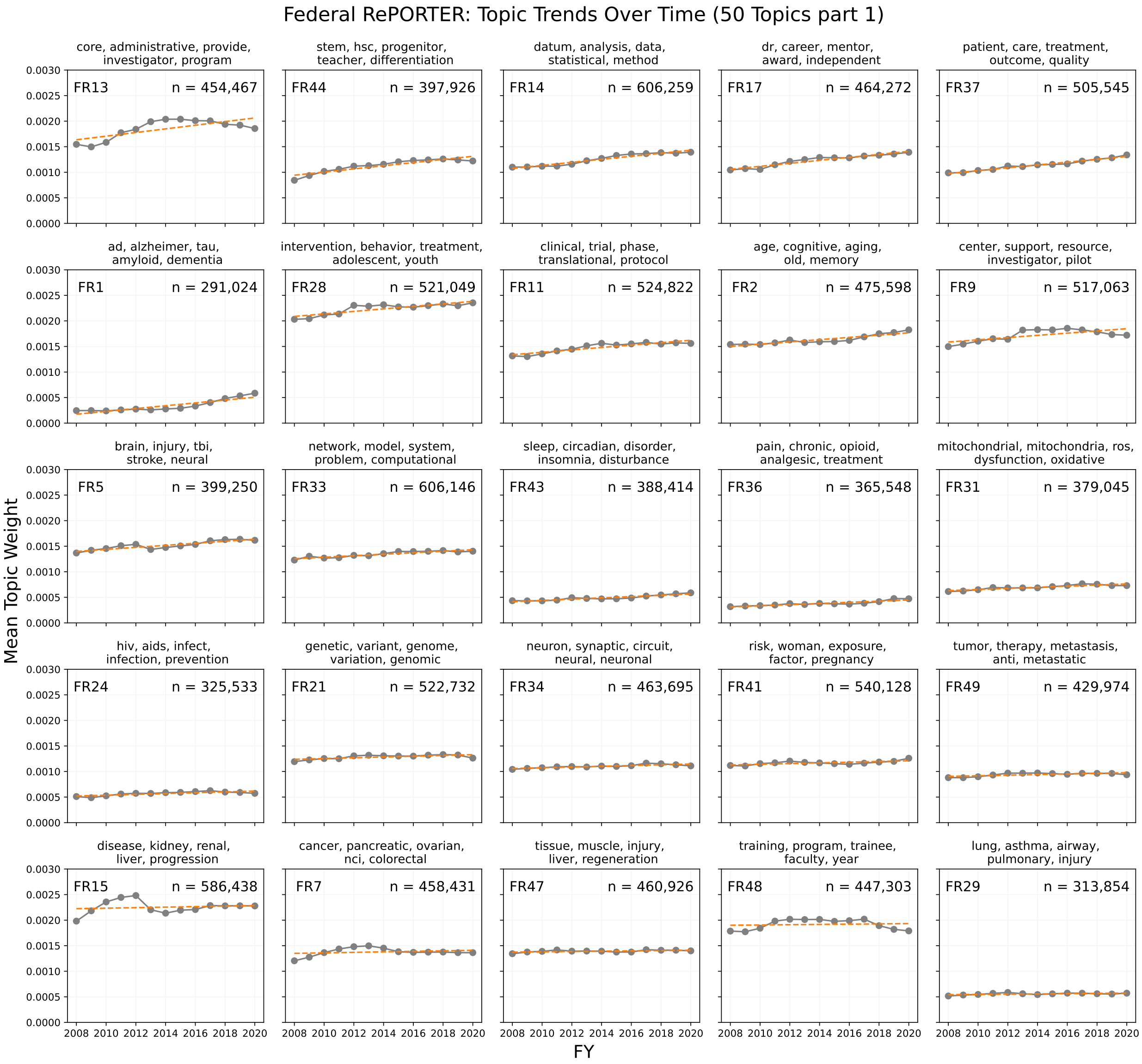

Figure 7: Federal RePORTER projects funded in FY 2008-2020: Processed dataset. Topic trend results produced by a 50-topic NMF model. Topic labels and the number of abstracts containing the topics, n, are given in the upper left and right plot corners respectively. Plots are ordered from largest to smallest 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, but are too small to be visible.

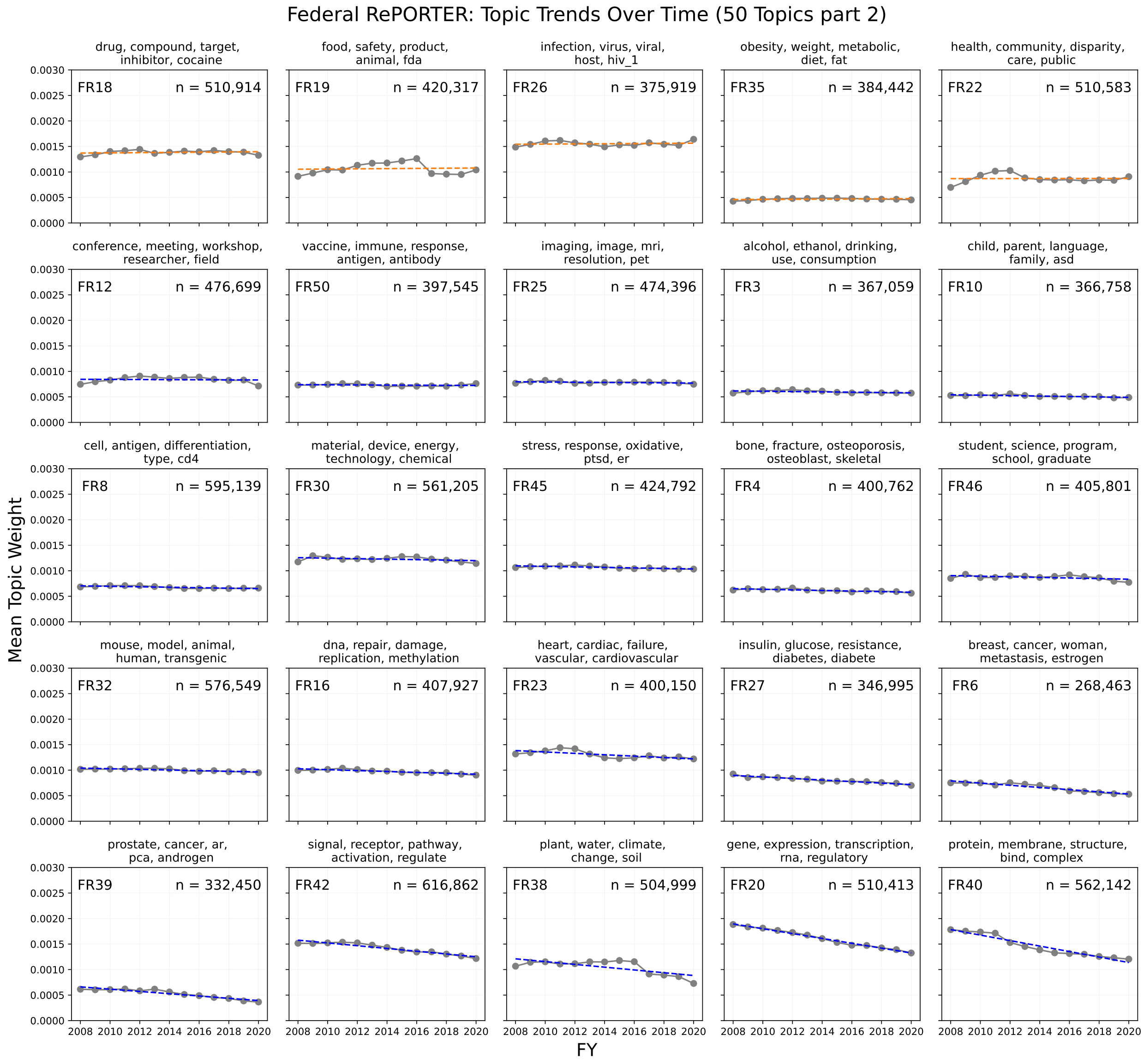

Figure 8: Federal RePORTER projects funded in FY 2008-2020: Processed dataset. Topic trend results produced by a 50-topic NMF model. Topic labels and the number of abstracts containing the topics, n, are given in the upper left and right plot corners respectively. Plots are ordered from largest to smallest 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, but are too small to be visible.

Author Bios

Kathryn Linehan is a Research Scientist at the Social and Decision Analytics Division, Biocomplexity Institute, University of Virginia. She is also a Ph.D. candidate at the University of Maryland, College Park in the Applied Mathematics & Statistics, and Scientific Computation program. Her primary research interests include randomized numerical linear algebra, low-rank matrix approximations and applications, and natural language processing. She is currently the Chair of the MD-DC-VA Section of the Mathematical Association of America.

Eric Oh is a Data Scientist at Reify Health, Inc. He holds a Ph.D. in Biostatistics from University of Pennsylvania. His interests span Bayesian statistics, causal inference, and clinical research.

Joel Thurston is a Senior Scientist at the Social and Decision Analytics Division, Biocomplexity Institute, University of Virginia. Joel received his Ph.D. in Social Psychology from the University of California Santa Barbara (UCSB). His research interests include the interface of group perception and group dynamics, conceptualizing and measuring emergent group properties, and the science of team science. He seeks to apply social science research methodologies to improve the use of administrative data in data science research.

Guy Leonel Siwe is a Postdoctoral Research Associate at the Social and Decision Analytics Division, Biocomplexity Institute, University of Virginia. He received his Ph.D. in Economics from University of Montreal and holds a master's degree in Statistics. His primary research areas are macroeconomics, firm heterogeneity, and international trade. He worked as a consultant at the World Bank.

Audrey Kindlon is a survey statistician at the National Center for Science & Engineering Statistics (NCSES) within the National Science Foundation. Her areas of expertise include microbusiness research and development and business innovation data.

John Jankowski is the Senior Economic Advisor for the National Center for Science & Engineering Statistics (NCSES) within the National Science Foundation. During his 30 years at NCSES Mr. Jankowski has been responsible for measurement of the Nation's financial and physical resources for R&D and innovation. Mr. Jankowski serves as Chair of the OECD's Working Party of National Experts on Science and Technology Indicators (NESTI), which is responsible for the development of internationally comparable statistics, indicators and quantitative analyses on science, technology, and innovation. He holds degrees from Georgetown University and the Johns Hopkins' School for Advanced International Studies.

Stephanie Shipp is the Acting Director for the Social and Decision Analytics Division, Biocomplexity Institute, University of Virginia. Dr. Shipp's work spans topics related to the use of all data to advance policy, the science of data science, community analytics, and innovation. She is leading and engaging in projects at the local, state, and federal level to assess data quality and the ethical use of new and traditional sources of data. She is a member of the American Statistical Association's Committee on Professional Ethics.

R&D – Supplementary Material

Link to dataset: https://doi.org/10.18130/V3/ATJOZW

Link to code: https://github.com/uva-bi-sdad/RnD_trends