# Machine-Learning Based Identification of Emerging Research Topics Using Research & Development Administrative Data

Eric J. Oh, Kathryn Linehan, Joel Thurston, Stephanie Shipp
(Social and Decision Analytics Division)

John Jankowski, Audrey Kindlon
(National Center for Science and Engineering Statistics)

FCSM 2021 - November 3, 2021

This work is supported by National Center for Science and Engineering Statistics (49100420C0015)

University of Virginia | Biocomplexity Institute & Initiative

# Motivation

- Around 21% of all research and development (R&D) funding in the United States is provided by the federal government[1]
  - What specific research areas does this public funding support?
- NCSES Federal Funds Survey and Federal Support Survey
  - Provides information on funding by agency and broad research fields
- Can we use other types of data to supplement the surveys?

[1] **Table 3**; **https://ncses.nsf.gov/pubs/nsf21324**; **2021**.

# Federal RePORTER

- Searchable database of scientific awards from agencies
  - HHS, NSF, USDA, NASA, DOD, VA, ED, EPA (most complete for HHS, NSF)
  - $>$ 1 million grants reported in FY 2008-2019
- Includes grant abstracts and metadata
- Agencies submit project information to Federal RePORTER

# Messy administrative data

## General cleaning

- Fill in missing project start date
- De-duplicate abstracts

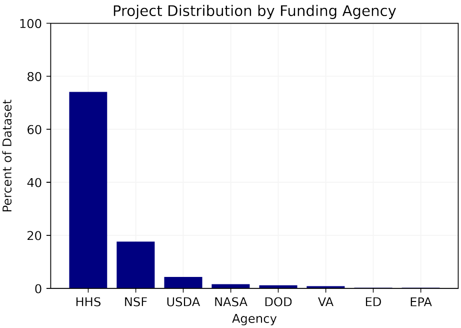
## Modeling purposes

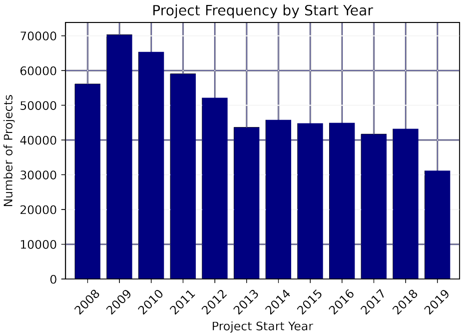
- Remove non project-specific words (e.g., "description", PI names)
- Tokenization, lemmatization, stop word removal, and the addition of bi-grams and tri-grams
- Remove most frequent words[2]

[2] **Schofield**; **A.**; **Magnusson**; **M.**; **Thompson**; **L.**; **& Mimno**; **D. (2017). Understanding text pre-processing for latent Dirichlet allocation**.

# Summary of data

690,814 abstracts for unique projects in FY 2008-19






University of Virginia Biocomplexity Institute & Initiative

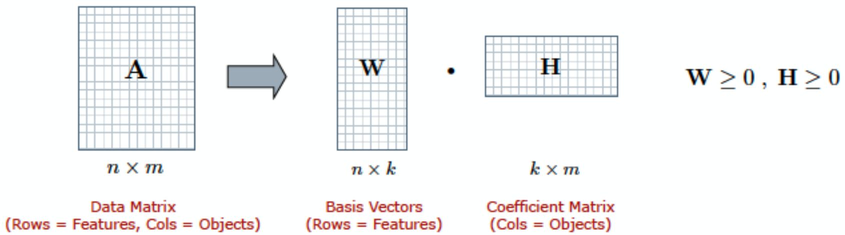
# Topic modeling

- Unsupervised machine learning method for discovering latent "themes" or "topics" in text data
- Topics are defined by a group or cluster of words that share semantic relationships
- Example topics:

| | | |
|---|---|---|
| aging | management | imaging |
| age | soil | image |
| older | crop | mri |
| adult | production | resolution |
| related | agricultural | optical |
| decline | practice | mr |
| life | pest | pet |
| aged | farm | contrast |
| lifespan | economic | microscopy |
| cognitive | farmer | probe |

# Non-negative matrix factorization (NMF)

- Given non-negative data matrix **A**, NMF finds a $k$-dimensional approximation in terms of non-negative factors **W** and **H**




$$\{\text{features}, \text{objects}, \text{basis vectors}\} = \{\text{words}, \text{documents}, \text{topics}\}$$

- Each document is represented by a linear combination of topics and weights (coefficient matrix)
    - $k$, the number of topics, must be chosen by the analyst

Figure source: Dynamic Topic Modeling via Non-negative Matrix Factorization by Derek Greene, slide 3

# NMF results - increasing weights

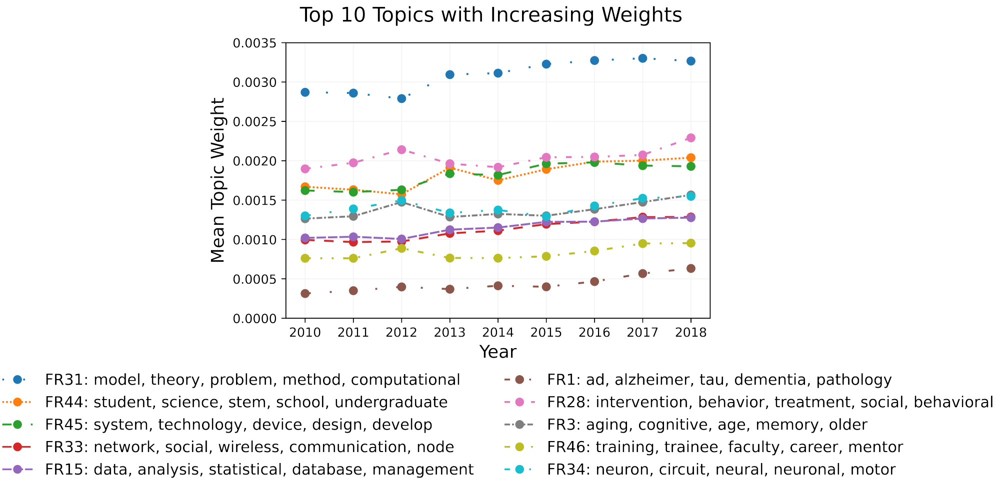



Figure: Federal RePORTER, FY 2008-19; University of Virginia, Social and Decision Analytics Division computations. Due to incomplete data from 2019, results include only 2010-18 abstracts

# NMF results - decreasing weights

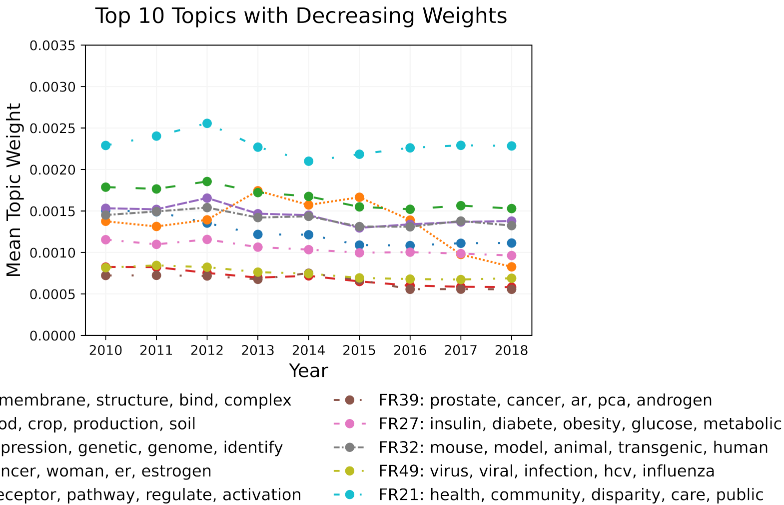



Figure: Federal RePORTER, FY 2008-19; University of Virginia, Social and Decision Analytics Division computations. Due to incomplete data from 2019, results include only 2010-18 abstracts

# Identifying topics within research areas

- We want to be able to discover topics within specific research areas of interest (e.g. coronavirus, artificial intelligence)
- Fitting topic models on all Federal RePORTER abstracts is too broad - need to narrow down somehow

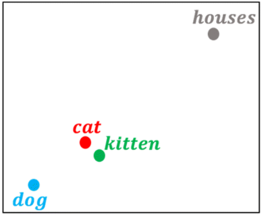
**How can we select the abstracts from Federal RePORTER that relate to a given research area?**

# Embeddings

Embeddings are representations of words as vectors, with semantically similar words resulting in similar vectors

| | living being | feline | human | gender | royalty | verb | plural |
|---|---|---|---|---|---|---|---|
| *cat* → | 0.6 | 0.9 | 0.1 | 0.4 | −0.7 | −0.3 | −0.2 |
| *kitten* → | 0.5 | 0.8 | −0.1 | 0.2 | −0.6 | −0.5 | −0.1 |
| *dog* → | 0.7 | −0.1 | 0.4 | 0.3 | −0.4 | −0.1 | −0.3 |
| *houses* → | −0.8 | −0.4 | −0.5 | 0.1 | −0.9 | 0.3 | 0.8 |

Dimensionality reduction of word embeddings from 7D to 2D →




| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *man* → | 0.6 | −0.2 | 0.8 | 0.9 | −0.1 | −0.9 | −0.7 |
| *woman* → | 0.7 | 0.3 | 0.9 | −0.7 | 0.1 | −0.5 | −0.4 |
| *king* → | 0.5 | −0.4 | 0.7 | 0.8 | 0.9 | −0.7 | −0.6 |
| *queen* → | 0.8 | −0.1 | 0.8 | −0.9 | 0.8 | −0.5 | −0.9 |

Dimensionality reduction of word embeddings from 7D to 2D →

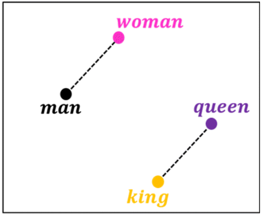



Figure source: https://medium.com/@hari4om/word-embedding-d816f643140

# Bidirectional Encoder Representations from Transformers (BERT)

- State of the art model for creating context aware embeddings
  - Vectors for the word "running" in "They are running a company" and "They are running a marathon" are different
- BERT provides pre-trained embeddings for use in various tasks
  - We can download the model and immediately calculate the embeddings for our corpus
- Standard BERT used Wikipedia corpus to train the model
  - Many extensions of BERT for task-specific purposes (e.g., trained on scientific articles to get better embeddings for STEM words)

# Cosine similarity

- Embeddings are constructed such that semantic similarity can be captured with the cosine similarity

$$\text{similarity}(\mathbf{A},\mathbf{B}) = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\cdot\|\mathbf{B}\|}$$

  where **A** and **B** are vectors.

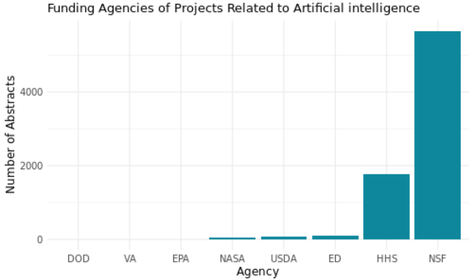
- We want to compare the semantic similarity of longer text (e.g., abstracts and search queries)
- Sentence BERT accomplishes this by constructing context-aware sentence embeddings
- Using BERT embeddings gives us a score to rank relevance/similarity to a research area
  - What query/document do we use to compare to the abstracts?

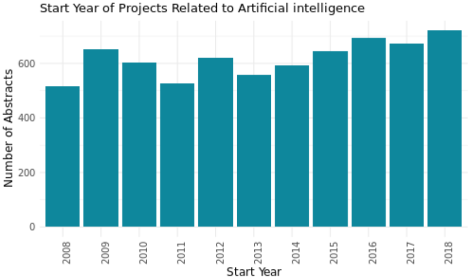
# Approach for artificial intelligence (AI)

- Scrape and extract the text from the Wikipedia page for "artificial intelligence"
- Construct embeddings for each sentence in AI Wiki and Federal RePORTER abstracts
- Calculate the cosine similarity for each pairwise combination of sentences
- Take the mean of the top 10 scores for each sentence in an abstract: average the scores across sentences to obtain the similarity score for an abstract
- Use a cutoff of 2.5 SDs above the mean abstract similarity score to classify as AI

# AI corpus summary

7,658 abstracts identified as AI





University of Virginia | Biocomplexity Institute & Initiative

# Initial AI topic modeling results

| Topic | Top 5 words |
|---|---|
| 1 | algorithm, optimization, computational, solution, complexity |
| 2 | biomedical, computational, phenotype, biology, biological |
| 3 | brain, neural, neuroscience, circuit, neuron |
| 4 | child, cognitive, social, developmental, mathematical |
| 5 | decision, choice, agent, uncertainty, value |
| 6 | engineering, engineer, education, nsf, student |
| 7 | language, word, processing, linguistic, text |
| 8 | learning, learn, learner, deep, machine_learning |
| 9 | network, social, agent, deep, dynamics |
| 10 | robot, human, robotics, task, robotic |
| 11 | science, social, scientific, workshop, innovation |
| 12 | software, user, code, developer, computing |
| 13 | statistical, dimensional, inference, variable, estimation |
| 14 | student, stem, teacher, thinking, skill |
| 15 | visual, object, image, vision, recognition |

# Discussion

- Federal RePORTER can yield very interesting results about research topics that have been funded over time
- Topic Modeling is able to discover those latent topics from just the grant abstracts
- Wikipedia and BERT embeddings can extract the relevant abstracts for artificial intelligence - hopefully, this can apply for other research areas

# References


1. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
2. Lee, D. D., Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.
3. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
4. Reimers, N., Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.