Analyzing Research and Development Trends Using Administrative Data

Kathryn Linehan, Eric Oh, Joel Thurston, Stephanie Shipp, Sallie Keller (University of Virginia, Biocomplexity Institute, Social and Decision Analytics Division)

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

2020 FCSM Research and Policy Conference

September 21, 2020

UNIVERSITY of VIRGINIA

SDAD & NCSES Partnership

- Social and Decision Analytics Division (SDAD)
  - leading research group in the Biocomplexity Institute and Initiative at the University of Virginia.
  - research has a history of measuring the social condition and economic innovation from multiple perspectives – societal, economic and statistical - with the goal of understanding the bias-precision tradeoffs using new and diverse data sources
- National Center for Science and Engineering Statistics (NCSES)
  - NSF's statistical agency
  - pursuing opportunities to assess the feasibility and ability to use non-survey data flows to supplement or enhance its current efforts in collecting Science, Technology, and Innovation (STI) indicators.

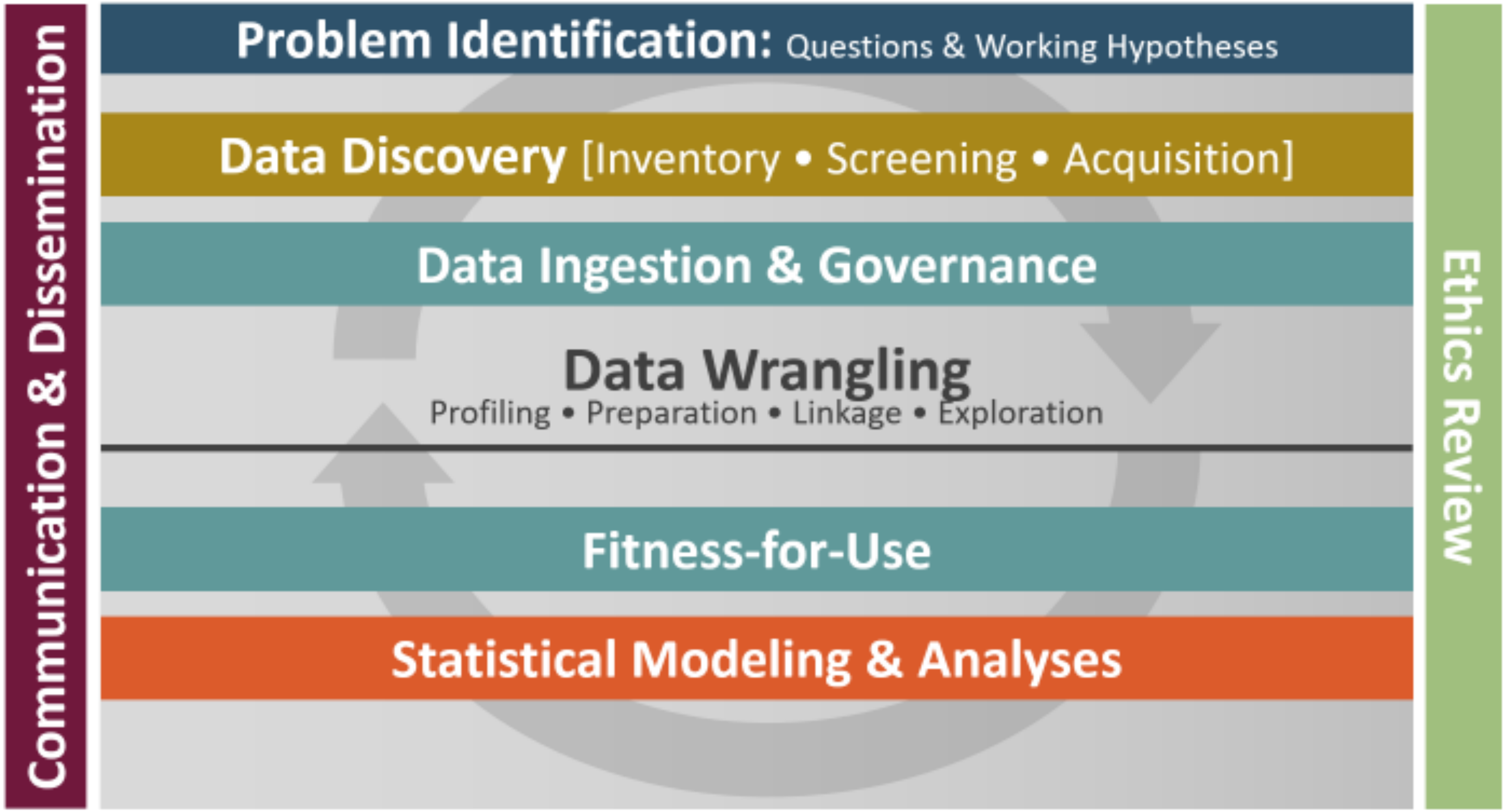
SDAD & NCSES Partnership

We have been collaborating with NCSES since 2016 to explore the feasibility of identifying and collecting data that naturally exists and are emerging for other reasons and repurposing those data to measure innovation and related concepts.

Project Aims

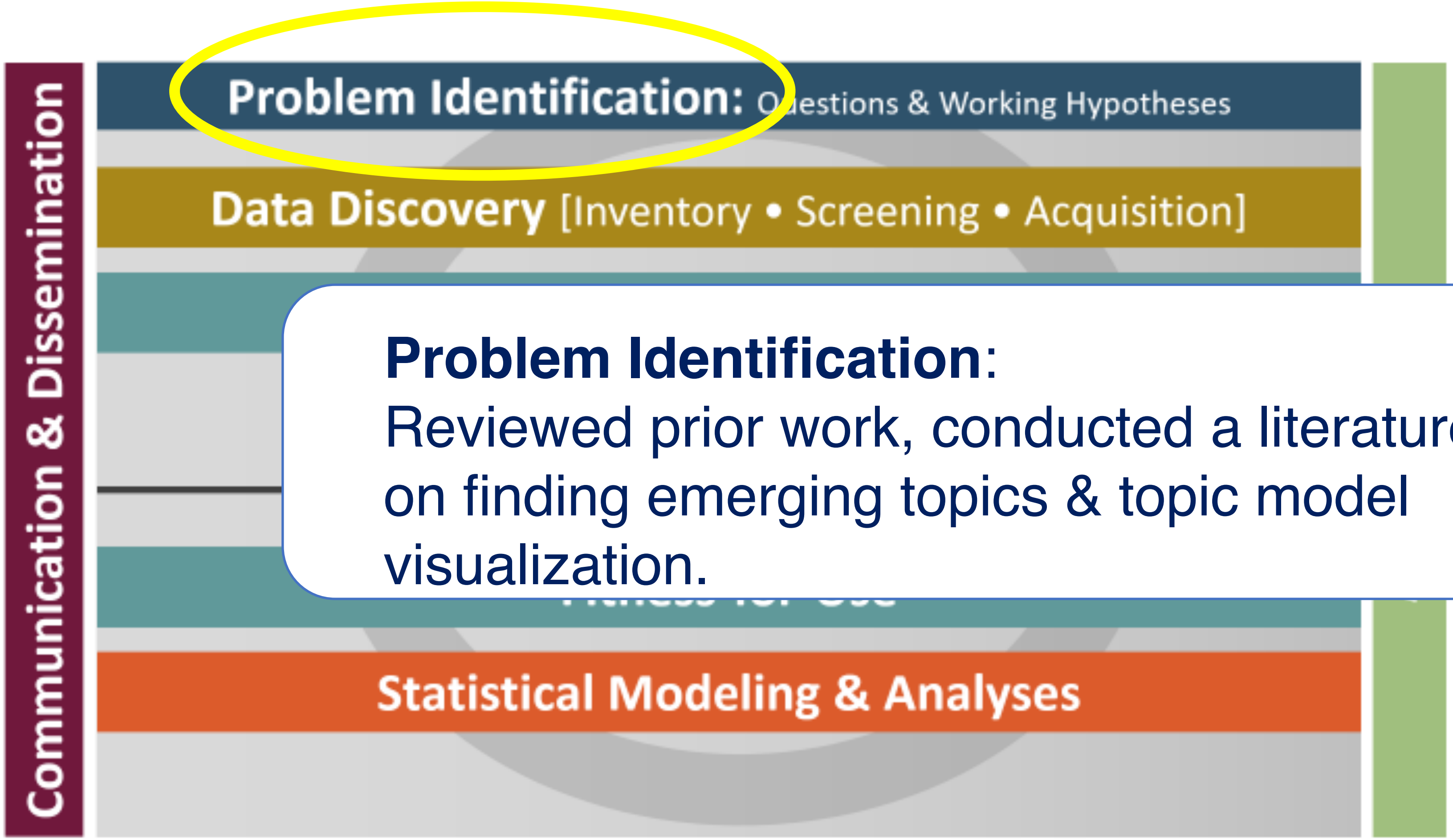
- Examine the use of administrative records to supplement or enhance data collected by NCSES.
- Use **Natural Language Processing (NLP)** and **machine learning techniques** to extract relevant Research and Development (R&D) topics from administrative records to supplement methods based on data collected in the NCSES Federal Funds Survey and Federal Support Survey.

Approach – UVA SDAD Data Science Framework



Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing Data Science: A Framework and Case Study. *Harvard Data Science Review*, *2*(1). https://hdsr.mitpress.mit.edu/pub/hnptx6lq/release/7

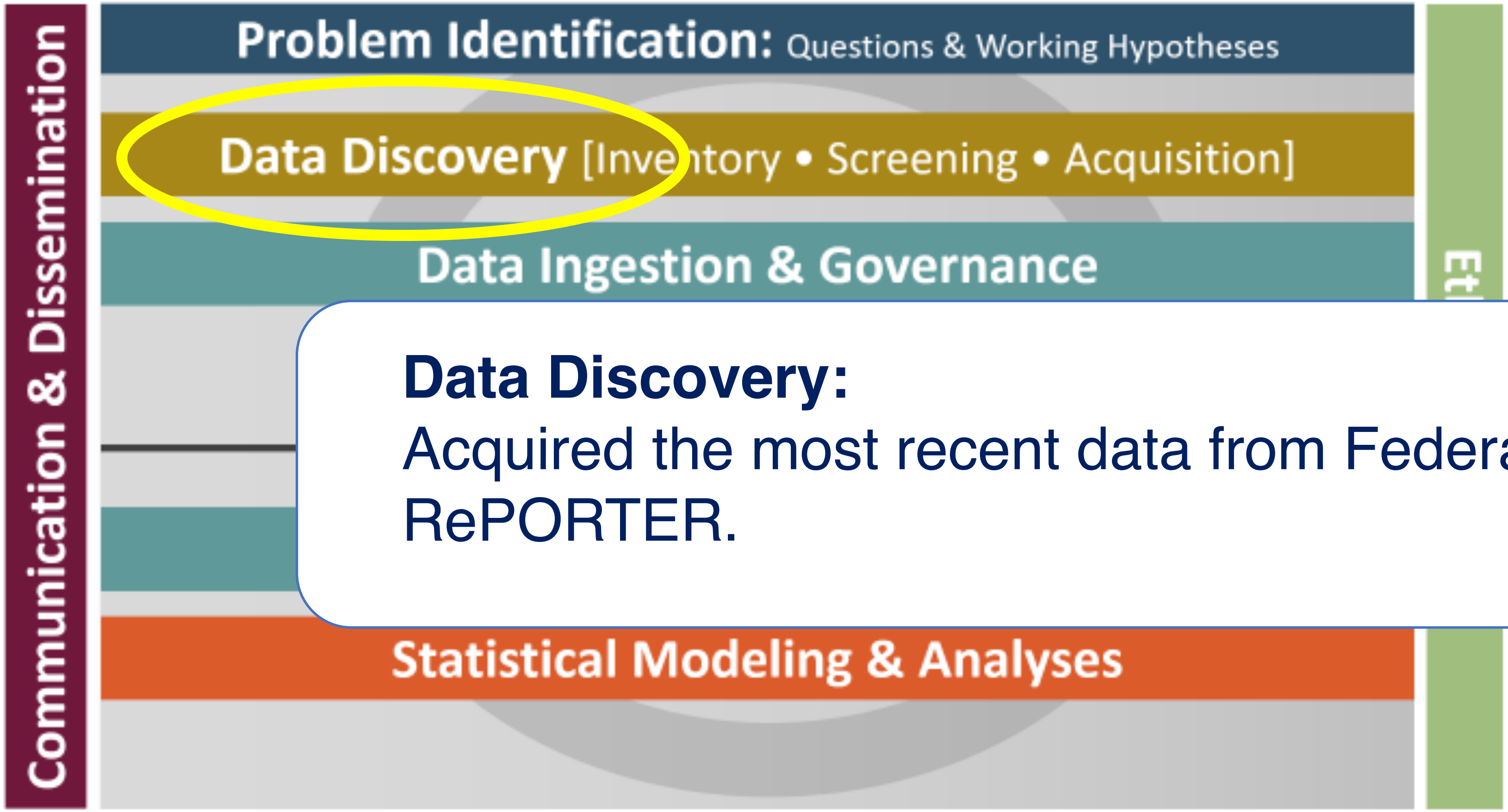
Approach – UVA SDAD Data Science Framework



Problem Identification:
Reviewed prior work, conducted a literature review on finding emerging topics & topic model visualization.

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing Data Science: A Framework and Case Study. *Harvard Data Science Review*, *2*(1). https://hdsr.mitpress.mit.edu/pub/hnptx6lq/release/7

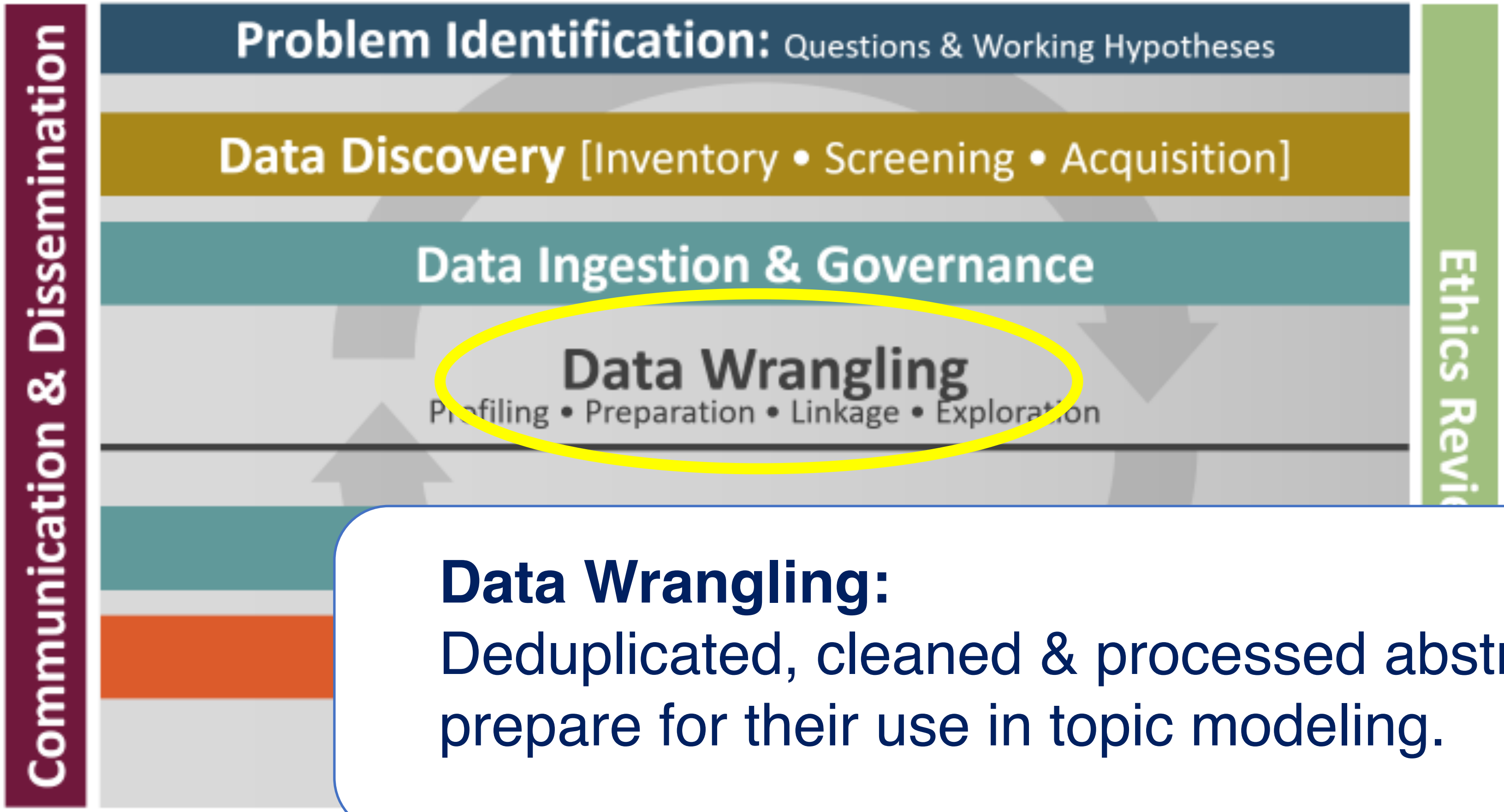
Approach – UVA SDAD Data Science Framework



Data Discovery:
Acquired the most recent data from Federal RePORTER.

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing Data Science: A Framework and Case Study. *Harvard Data Science Review*, *2*(1). https://hdsr.mitpress.mit.edu/pub/hnptx6lq/release/7

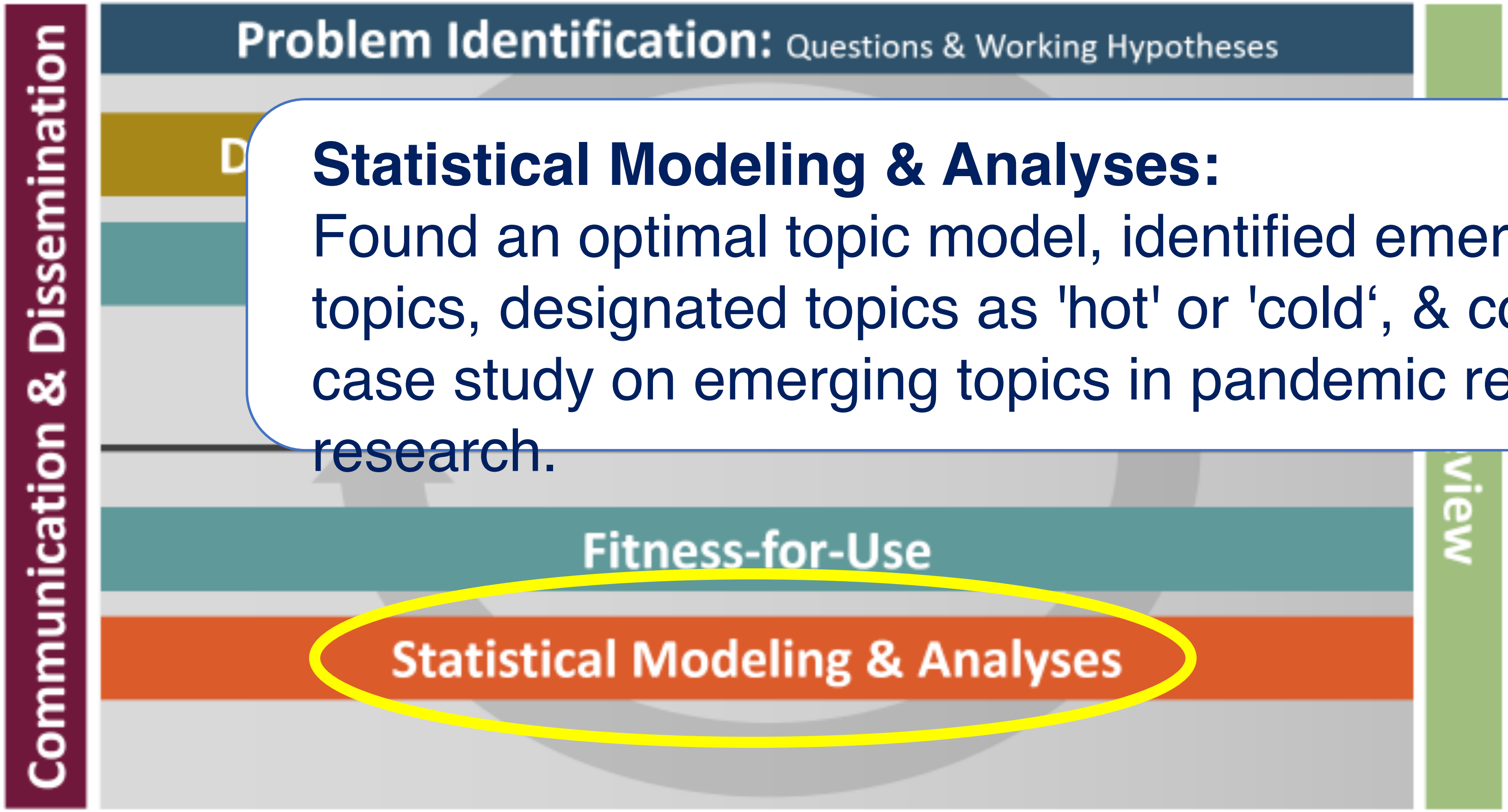
Approach – UVA SDAD Data Science Framework



Data Wrangling:
Deduplicated, cleaned & processed abstracts to prepare for their use in topic modeling.

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing Data Science: A Framework and Case Study. *Harvard Data Science Review*, *2*(1). https://hdsr.mitpress.mit.edu/pub/hnptx6lq/release/7

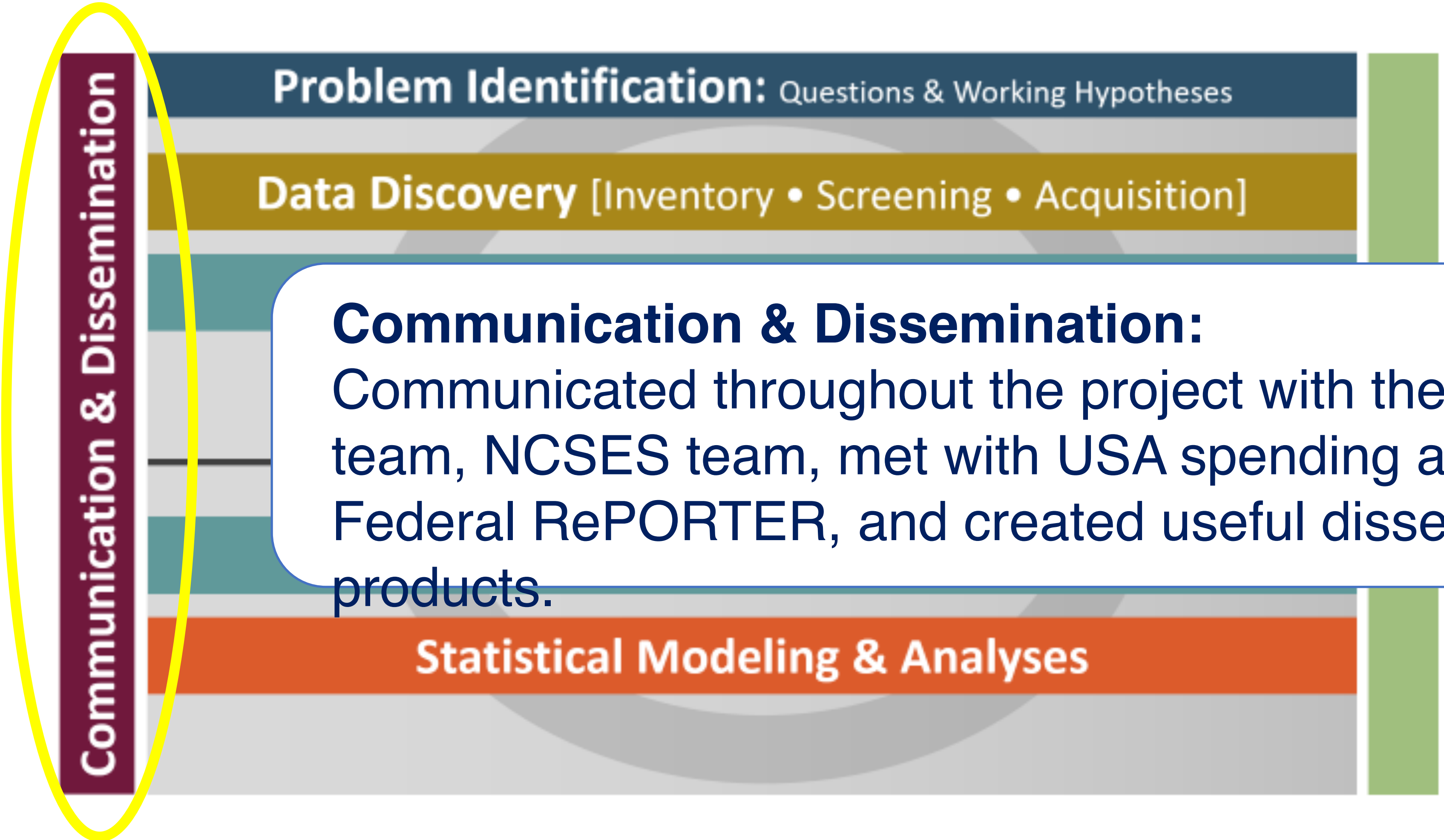
Approach – UVA SDAD Data Science Framework



Statistical Modeling & Analyses:
Found an optimal topic model, identified emerging topics, designated topics as 'hot' or 'cold', & conducted a case study on emerging topics in pandemic related research.

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing Data Science: A Framework and Case Study. *Harvard Data Science Review*, *2*(1). https://hdsr.mitpress.mit.edu/pub/hnptx6lq/release/7

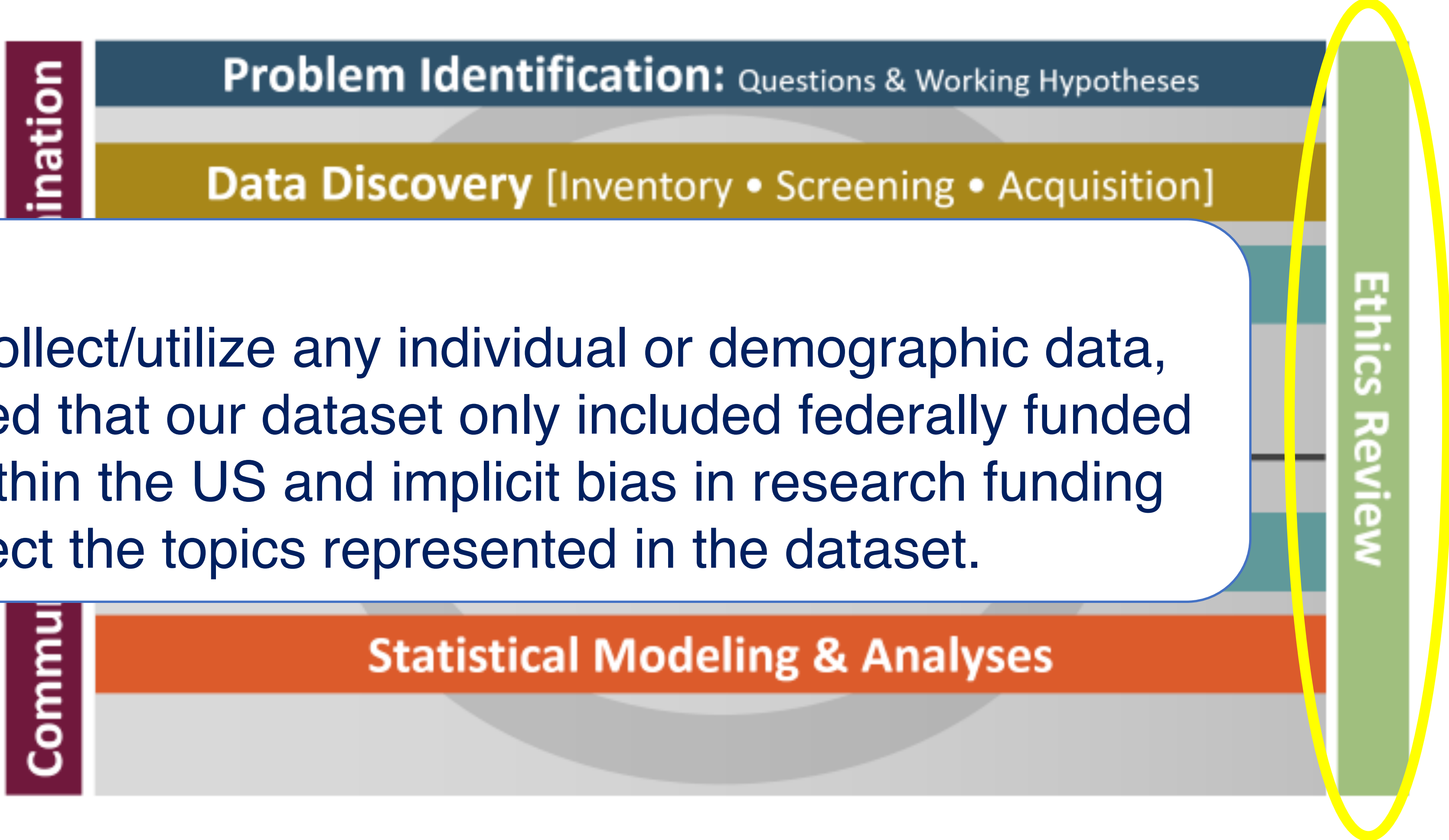
Approach – UVA SDAD Data Science Framework



Communication & Dissemination:
Communicated throughout the project with the SDAD team, NCSES team, met with USA spending and Federal RePORTER, and created useful dissemination products.

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing Data Science: A Framework and Case Study. *Harvard Data Science Review*, *2*(1). https://hdsr.mitpress.mit.edu/pub/hnptx6lq/release/7

Approach – UVA SDAD Data Science Framework



Ethics:
Did not collect/utilize any individual or demographic data, recognized that our dataset only included federally funded grants within the US and implicit bias in research funding could affect the topics represented in the dataset.

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing Data Science: A Framework and Case Study. *Harvard Data Science Review*, *2*(1). https://hdsr.mitpress.mit.edu/pub/hnptx6lq/release/7

Data Source

Federal RePORTER

- 2008-2019
- Over 1 million grants from multiple agencies, i.e., DOD, ED, EPA, HHS, NASA, NSF, USDA, VA
- Includes grant abstracts and metadata

Topic Modeling

- Unsupervised machine learning technique for grouping text data into themes
- A topic is a list of words clustered together by the algorithm that should 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

Data Preparation for Emerging Topics

- We needed to:
  - fill in missing information for project start date
  - decide on a deduplication strategy for dealing with duplicate abstracts in the corpus, and
  - clean and prepare the abstracts for use in topic modeling.

Data Preparation for Emerging Topics

- We needed to:
  - **fill in missing information for project start date**
  - decide on a deduplication strategy for dealing with duplicate abstracts in the corpus, and
  - clean and prepare the abstracts for use in topic modeling.

Raw Abstract	Project Start Date	Budget Start Date	FY entered into Federal RePORTER
Waldmann co-discovered the cytokine IL-15 and ...	NaN	NaN	2016
PROJECT SUMMARY (See instructions): Protocol S...	NaN	12/1/2015	2016
PROJECT SUMMARY (See instructions): The objec...	NaN	6/1/2016	2016

Data Preparation for Emerging Topics

- We needed to:
  - **fill in missing information for project start date**
  - decide on a deduplication strategy for dealing with duplicate abstracts in the corpus, and
  - clean and prepare the abstracts for use in topic modeling.

Raw Abstract	Project Start Date	Budget Start Date	FY entered into Federal RePORTER
Waldmann co-discovered the cytokine IL-15 and ...	2016 NaN	NaN	2016
PROJECT SUMMARY (See instructions): Protocol S...	12/1/2015 NaN	12/1/2015	2016
PROJECT SUMMARY (See instructions): The objec...	6/1/2016 NaN	6/1/2016	2016

Data Preparation for Emerging Topics

- We needed to:
  - fill in missing information for project start date
  - decide on a deduplication strategy for dealing with duplicate abstracts in the corpus, and
  - **clean and prepare the abstracts for use in topic modeling**.

DESCRIPTION (provided by applicant): The objective of this research is to understand the biophysical basis (thermodynamics, kinetics) for multivalency. Multivalency is concerned with biologically important interactions in which multiple receptors and multiple ligands interact simultaneously. The first focus of this work is to explore the synthesis and properties of groups used to join monovalent ligands (linkers) in the synthesis of multivalent ligands.

Data Preparation for Emerging Topics

- We needed to:
  - fill in missing information for project start date
  - decide on a deduplication strategy for dealing with duplicate abstracts in the corpus, and
  - **clean and prepare the abstracts for use in topic modeling**.

description (provided by applicant): the objective of this research is to understand the biophysical basis (thermodynamics, kinetics) for multivalency. multivalency is concerned with biologically important interactions in which multiple receptors and multiple ligands interact simultaneously. the first focus of this work is to explore the synthesis and properties of groups used to join monovalent ligands (linkers) in the synthesis of multivalent ligands.

Data Preparation for Emerging Topics

- We needed to:
  - fill in missing information for project start date
  - decide on a deduplication strategy for dealing with duplicate abstracts in the corpus, and
  - **clean and prepare the abstracts for use in topic modeling**.

~~description (provided by applicant):~~ ~~the~~ objective ~~of this~~ research ~~is to~~ understand ~~the~~ biophysical basis ~~(~~thermodynamics~~,~~ kinetics~~)~~ ~~for~~ multivalency~~.~~ multivalency ~~is~~ concerned ~~with~~ biologically important interactions ~~in which~~ multiple receptors ~~and~~ multiple ligands interact simultaneously~~.~~ ~~the first~~ focus ~~of this~~ work ~~is to~~ explore ~~the~~ synthesis ~~and~~ properties ~~of~~ groups ~~used to~~ join monovalent ligands ~~(~~linkers~~)~~ ~~in the~~ synthesis ~~of~~ multivalent ligands~~.~~

Data Preparation for Emerging Topics

- We needed to:
  - fill in missing information for project start date
  - decide on a deduplication strategy for dealing with duplicate abstracts in the corpus, and
  - **clean and prepare the abstracts for use in topic modeling**. [1]

objective ~~research~~ ~~understand~~ biophysical basis thermodynamics kinetics multivalency multivalency concerned biologically important interactions multiple receptors multiple ligands interact simultaneously focus work explore synthesis properties groups join monovalent ligands linkers synthesis multivalent ligands

Data Preparation for Emerging Topics

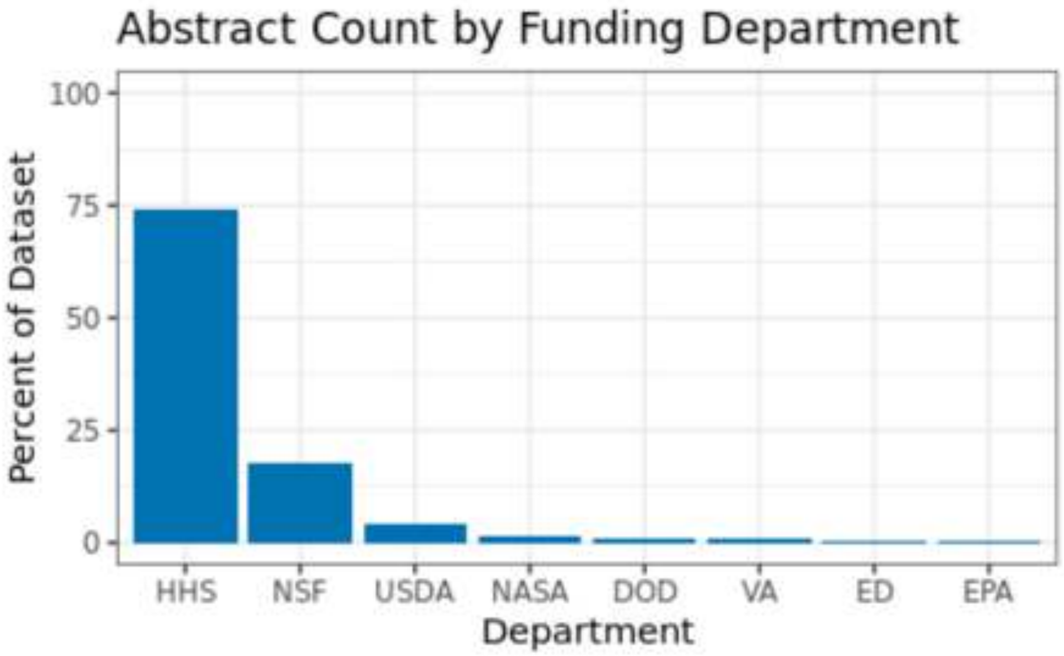
- We needed to:
  - fill in missing information for project start date
  - decide on a deduplication strategy for dealing with duplicate abstracts in the corpus, and
  - **clean and prepare the abstracts for use in topic modeling**.

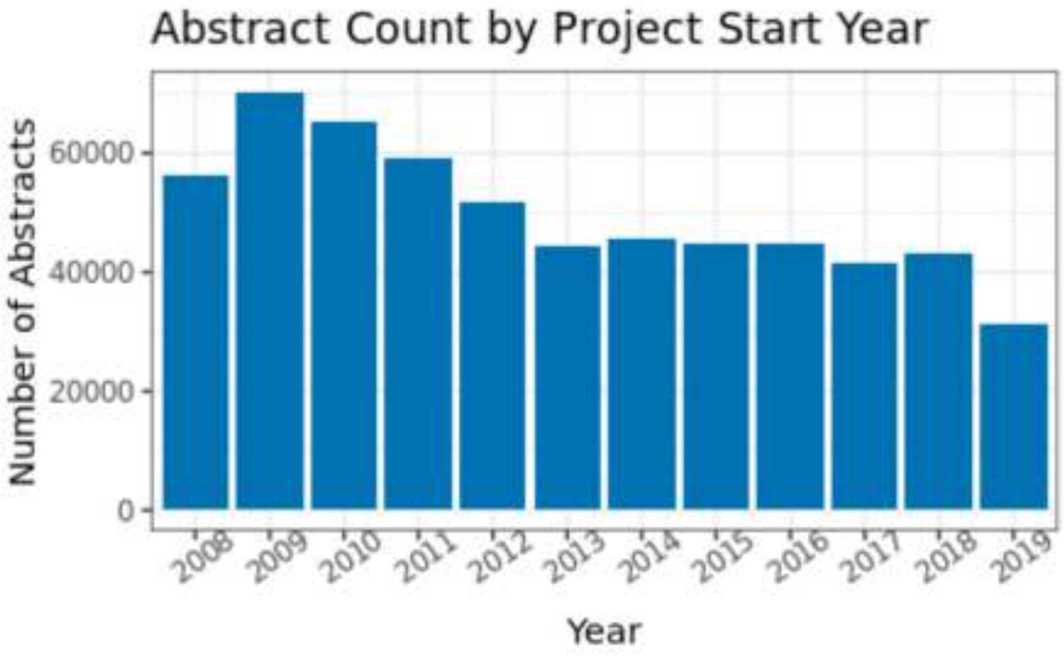
objective biophysical basis thermodynamics kinetics multivalency multivalency concerned biologically important interactions multiple receptors multiple ligands interact simultaneously focus work explore synthesis properties groups join monovalent ligands linkers synthesis multivalent ligands

	basis	behavior	biologically	brain	…	multivalency	…
Doc 1	1	0	1	0		2	
…							

Dataset after Deduplication

- 690,814 abstracts for unique projects



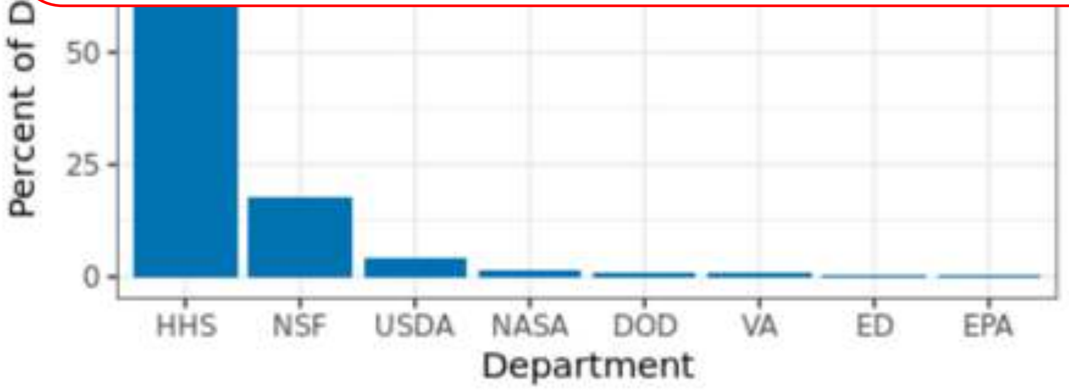


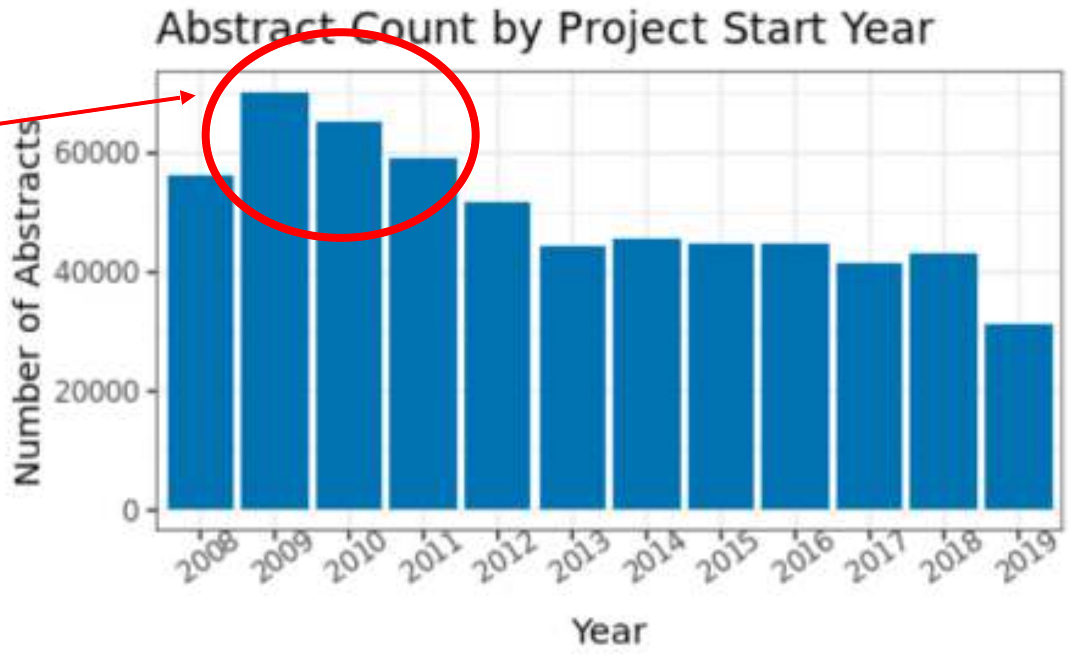
Data Source: [Federal RePORTER](), 2008-2019,
University of Virginia, Social and Decision Analytics Division computations

Dataset after Deduplication

- 690,814 abstracts for unique projects

American Recovery and Reinvestment Act of 2009 (ARRA) – affected 2010 and 2011 funding

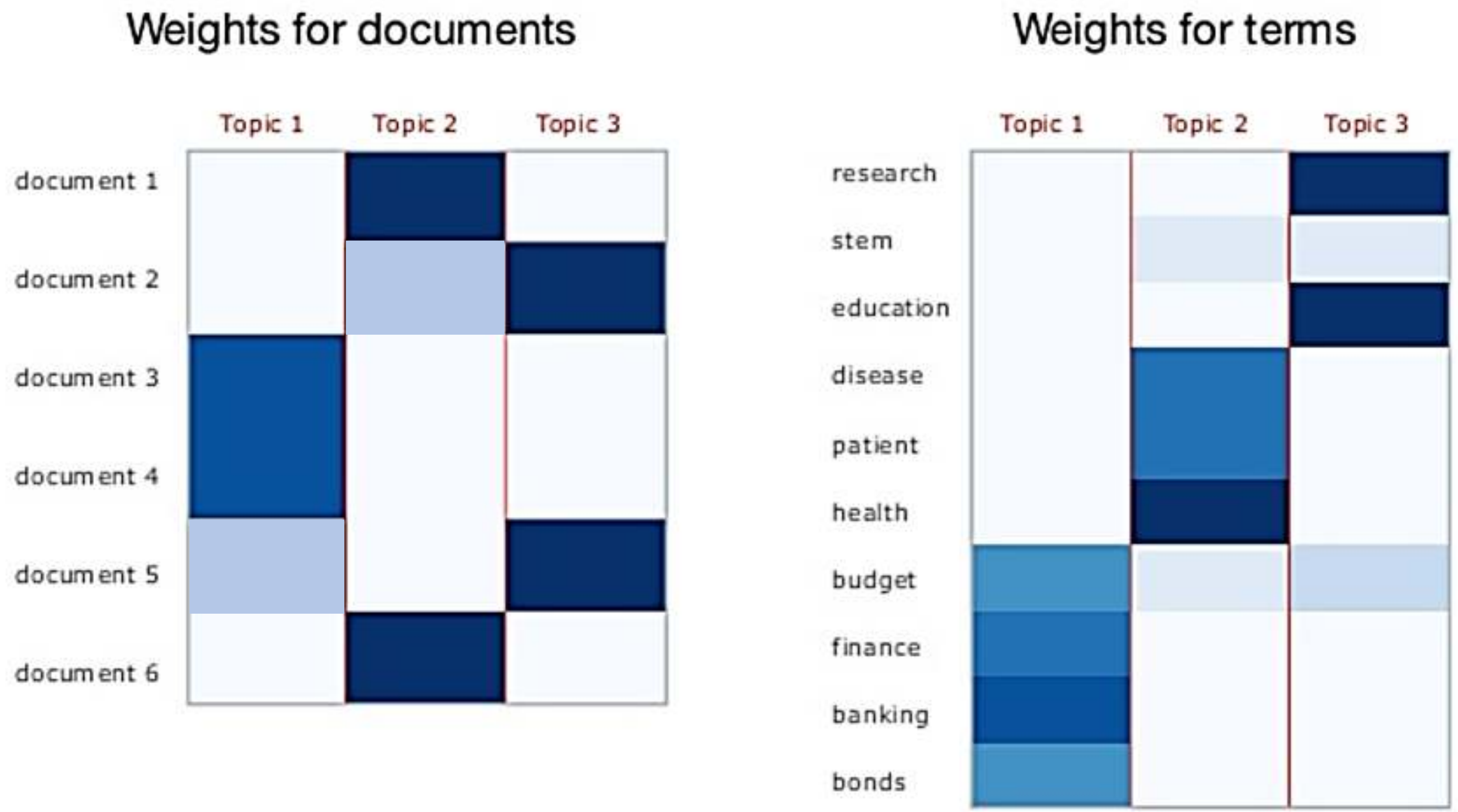




Data Source: Federal RePORTER, 2008-2019,
University of Virginia, Social and Decision Analytics Division computations

Topic Modeling using Non-negative Matrix Factorization (NMF)

- Linear algebra technique that can be used for clustering groups of words into topics



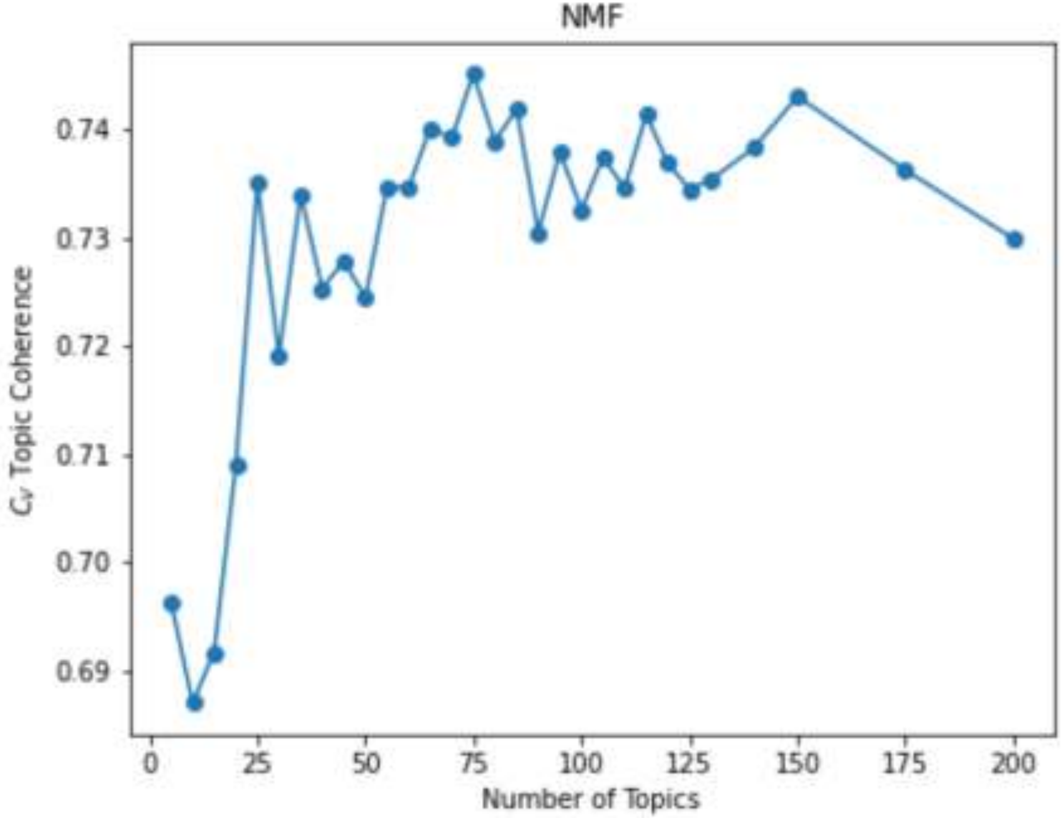
Graphic Source: [Dynamic Topic Modeling via Non-negative Matrix Factorization by Dr. Derek Greene, slide 6].

Topic Modeling using Non-negative Matrix Factorization (NMF)

- We do not know the number of topics in advance, so we tested a number of topic models in order to find an optimal model.
- Measure for “goodness of fit” for topic models: C_v Topic Coherence

Tuning the Number of Topics

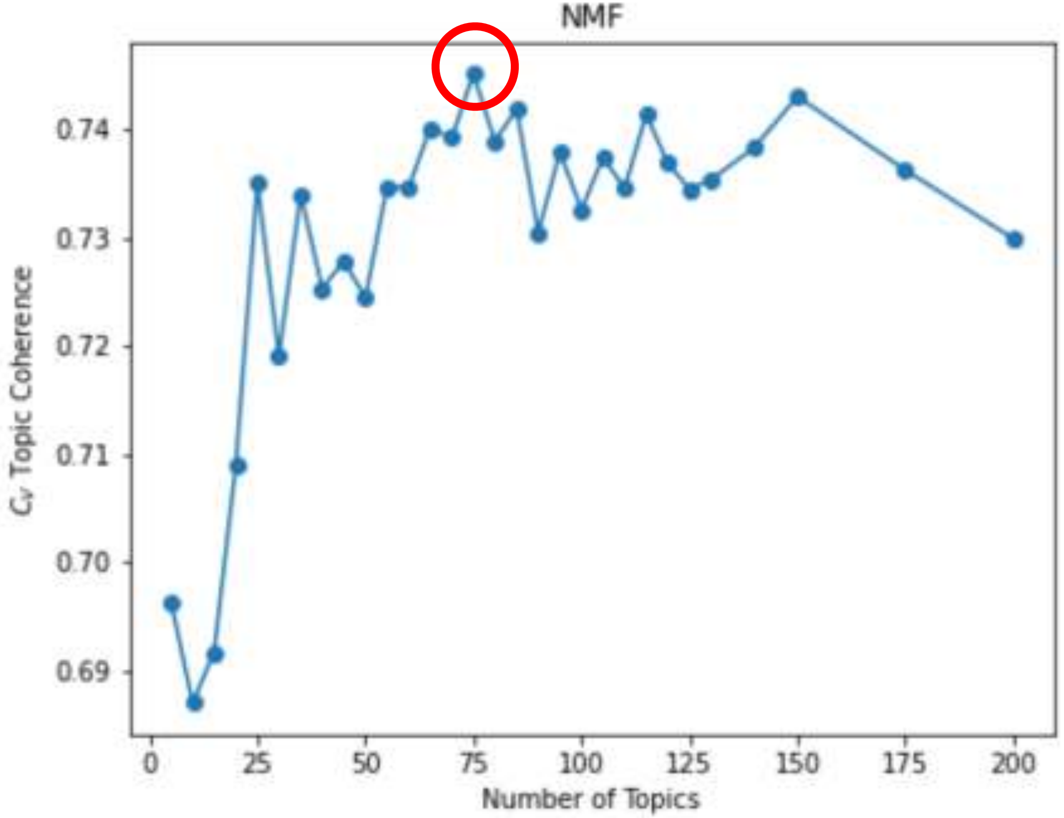
- Single topic model runs on the full dataset – sample results
- The best model of these is an NMF topic model with 75 topics.



Data Source: [Federal RePORTER](), 2008-2019,
University of Virginia, Social and Decision Analytics Division computations

Tuning the Number of Topics

- Single topic model runs on the full dataset – sample results
- The best model of these is an NMF topic model with 75 topics.

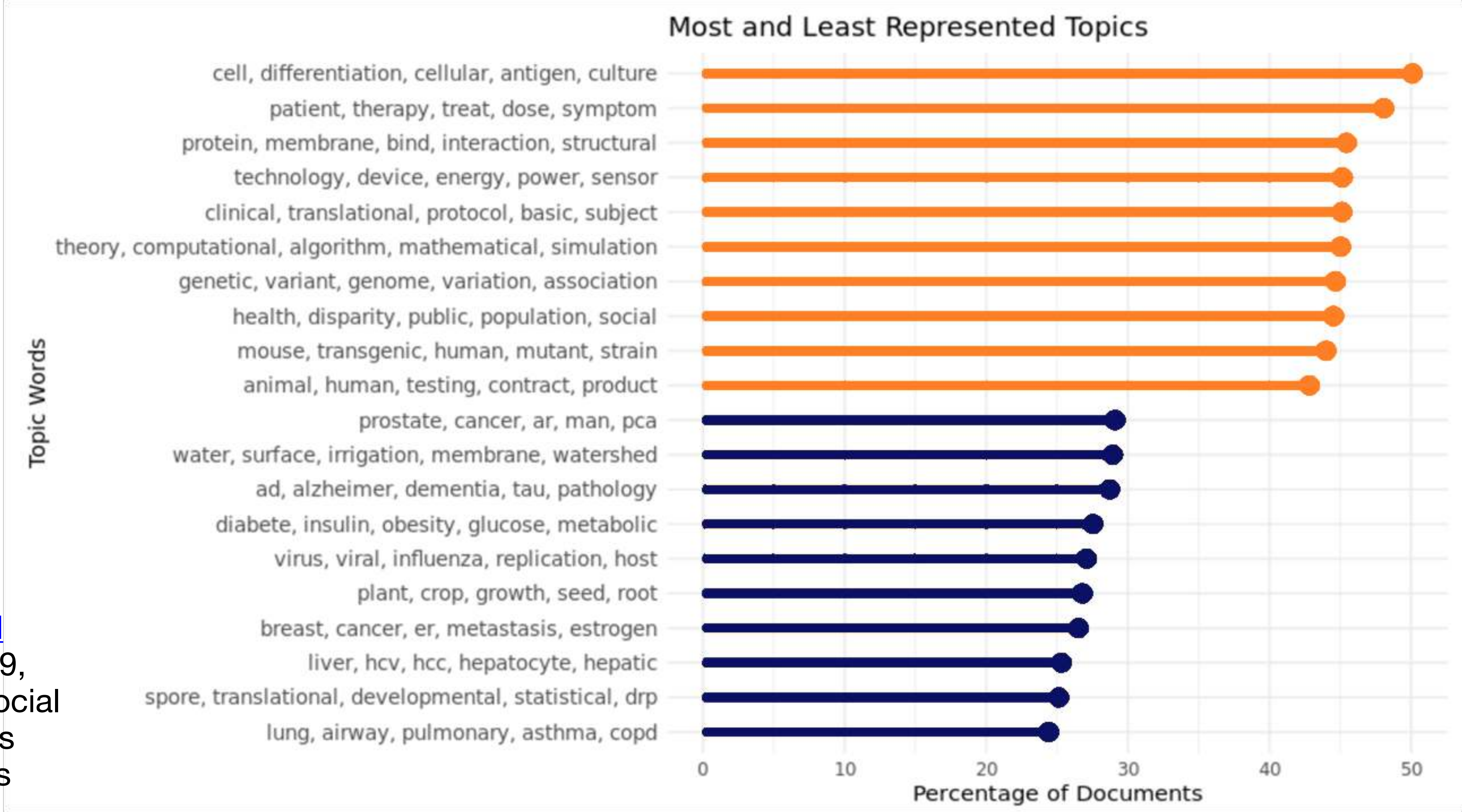


Data Source: [Federal RePORTER](), 2008-2019,
University of Virginia, Social and Decision Analytics Division computations

Optimal Topic Model Results

Measuring the percentage of documents in which each topic occurs

- Full dataset
- NMF - 75 topics

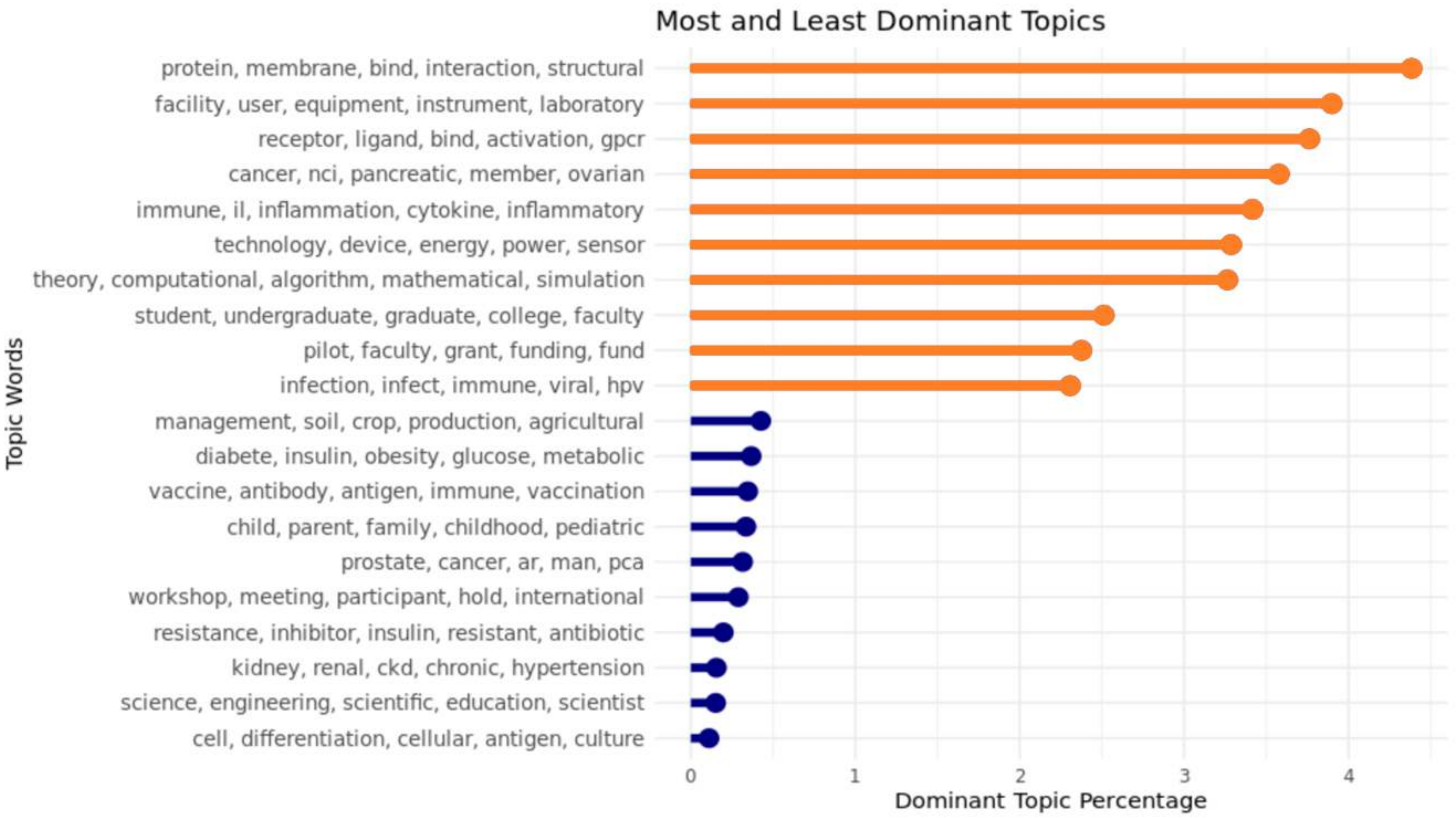


Data Source: [Federal RePORTER](), 2008-2019, University of Virginia, Social and Decision Analytics Division computations

Optimal Topic Model Results

Measuring the percentage of times each topic was “dominant,” i.e., the topic with the highest weight for a document.

- Full dataset
- NMF - 75 topics



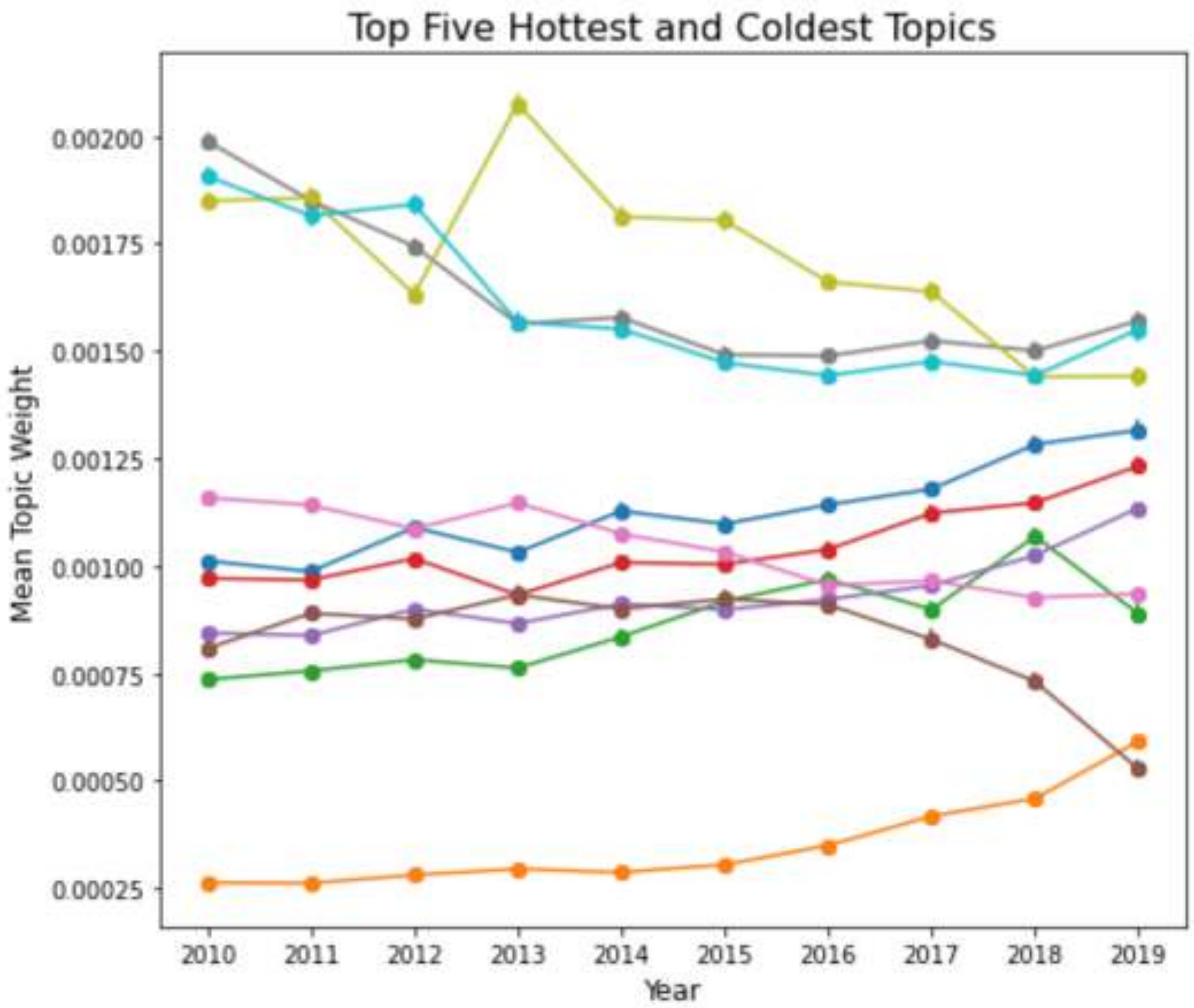
Data Source: Federal RePORTER, 2008-2019, University of Virginia, Social and Decision Analytics Division computations

Emerging Topics Method

- Using our optimal topic model, an NMF model with 75 topics, we analyze its results to discover and characterize 'hot' and 'cold' topics.
- We follow the approach of [2] using our optimal NMF topic model. We also use the work in [3] as a reference.
- To categorize a topic as “hot” or “cold”, we
  1. Find the average weight of each topic in each year between 2010-2019.
  2. Model the relationship between the average weights and years for each topic using linear regression.
  3. Topics that have regression lines with positive slopes are considered ‘hot' and those that have regression lines with negative slopes are considered ‘cold'.

Emerging Topics Results

Full dataset, NMF - 75 topics



Top Five Hottest Topics

- care, provider, practice, medical, cost
- ad, alzheimer, dementia, tau, pathology
- stem, hsc, hematopoietic, progenitor, college
- brain, neural, mri, region, fmri
- patient, therapy, treat, dose, symptom

Top Five Coldest Topics

- water, surface, irrigation, membrane, watershed
- animal, human, testing, contract, product
- protein, membrane, bind, interaction, structural
- facility, user, equipment, instrument, laboratory
- receptor, ligand, bind, activation, gpcr

Note: “cold” topics can still be popular, but just trending downward in prevalence

Data Source: [Federal RePORTER](), 2008-2019, University of Virginia, Social and Decision Analytics Division computations

Pandemics Case Study Approach

- We explore emerging topics around the research areas of pandemics and coronavirus.
  1. We use information retrieval techniques to create two smaller corpora: one that focuses on pandemics, and one that focuses on coronavirus.

Smaller Corpora Sizes

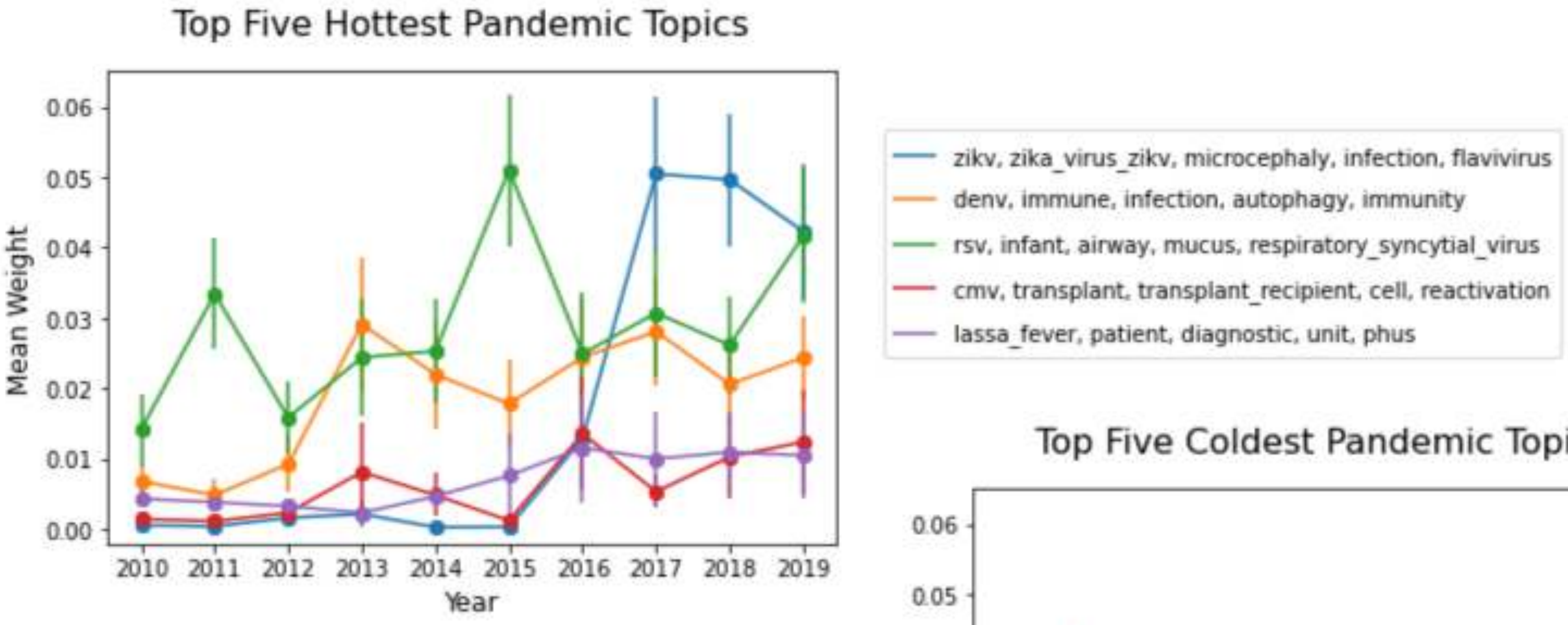
Corpus	Number of Projects
Pandemics	1137
Coronavirus	1012

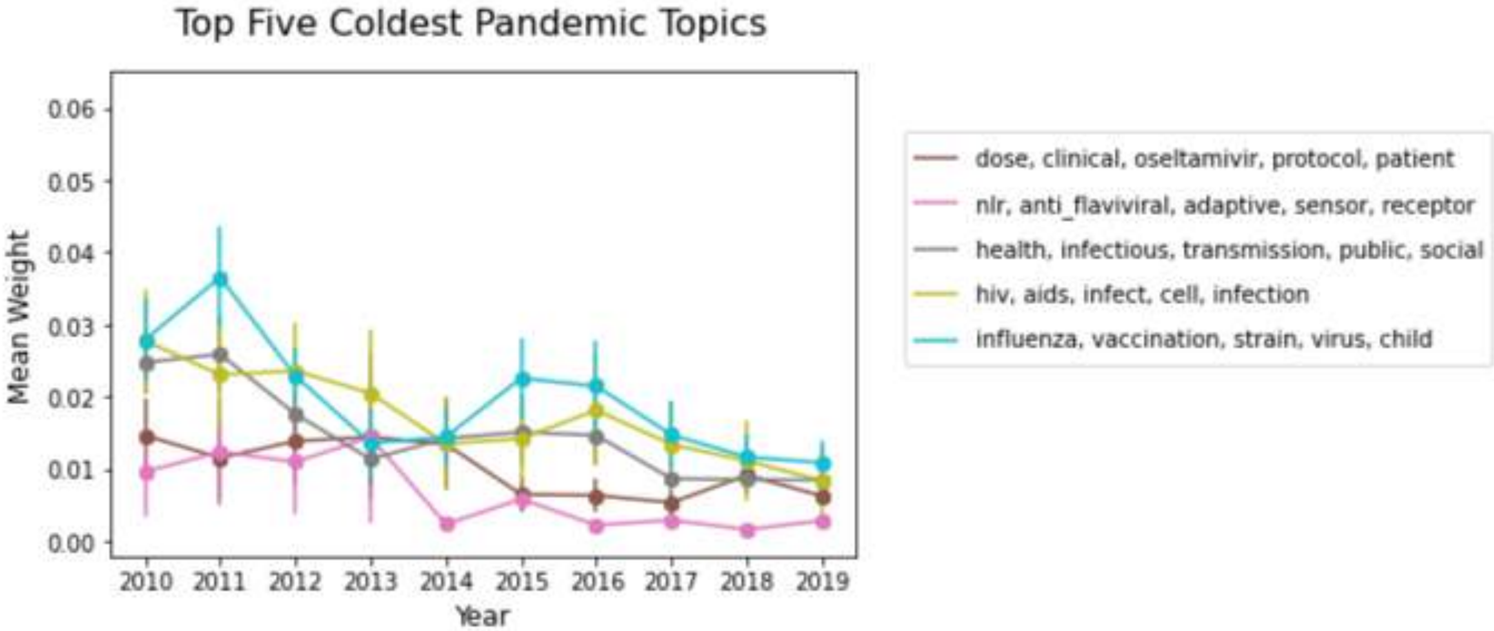
Data Source: Federal RePORTER, 2008-2019,
University of Virginia, Social and Decision Analytics Division computations

  2. We use an NMF topic model of 30 topics on each smaller corpus and conduct the emerging topics analysis.

Case Study Results – “Pandemic” Corpus

NMF - 30 topics

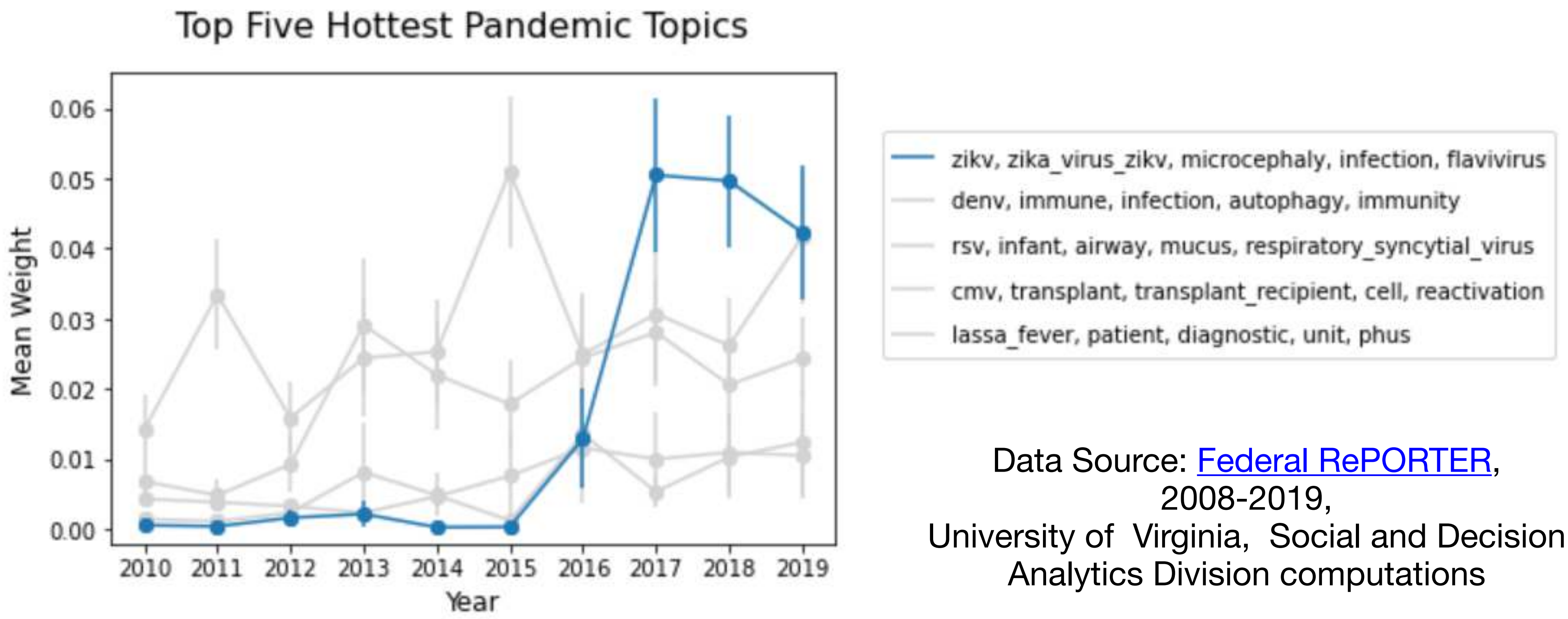




Data Source: [Federal RePORTER](), 2008-2019,
University of Virginia, Social and Decision Analytics Division computations

Case Study Results – "Pandemic" Corpus

NMF - 30 topics

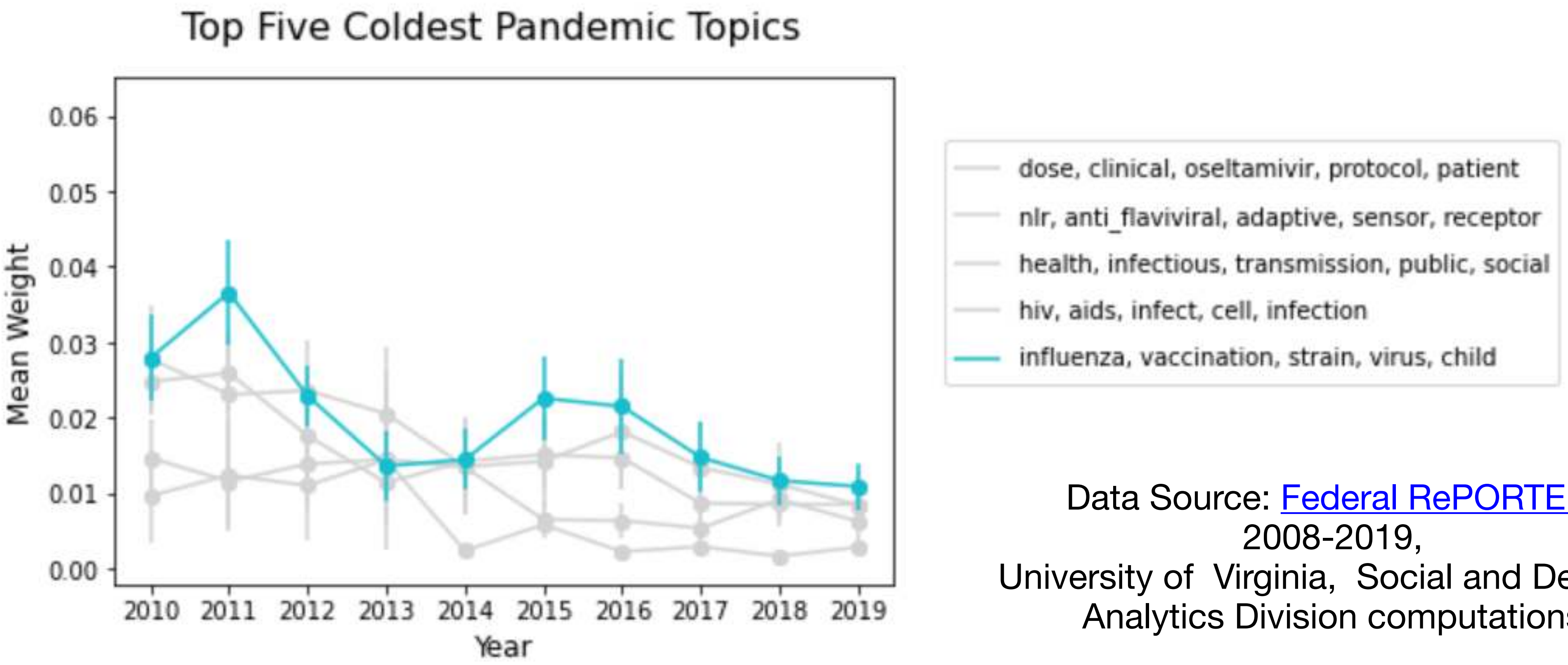


Data Source: Federal RePORTER, 2008-2019, University of Virginia, Social and Decision Analytics Division computations

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
n	105	87	90	55	76	74	59	68	84	77

Case Study Results – “Pandemic” Corpus

NMF - 30 topics

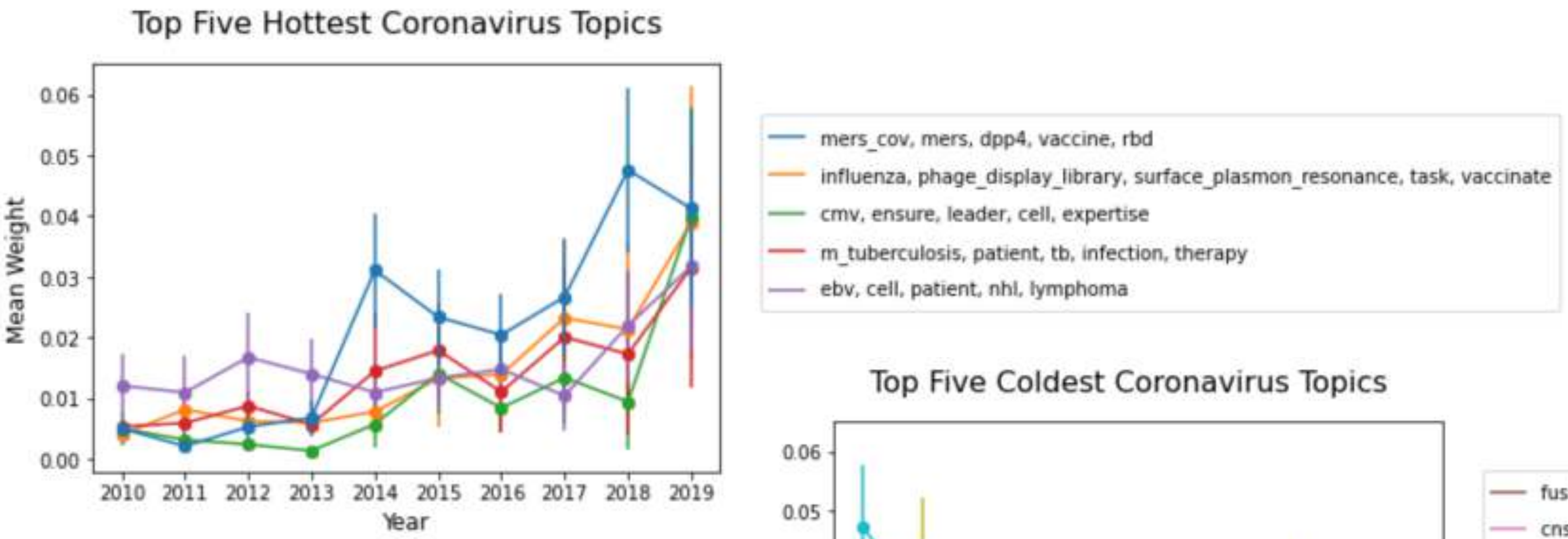


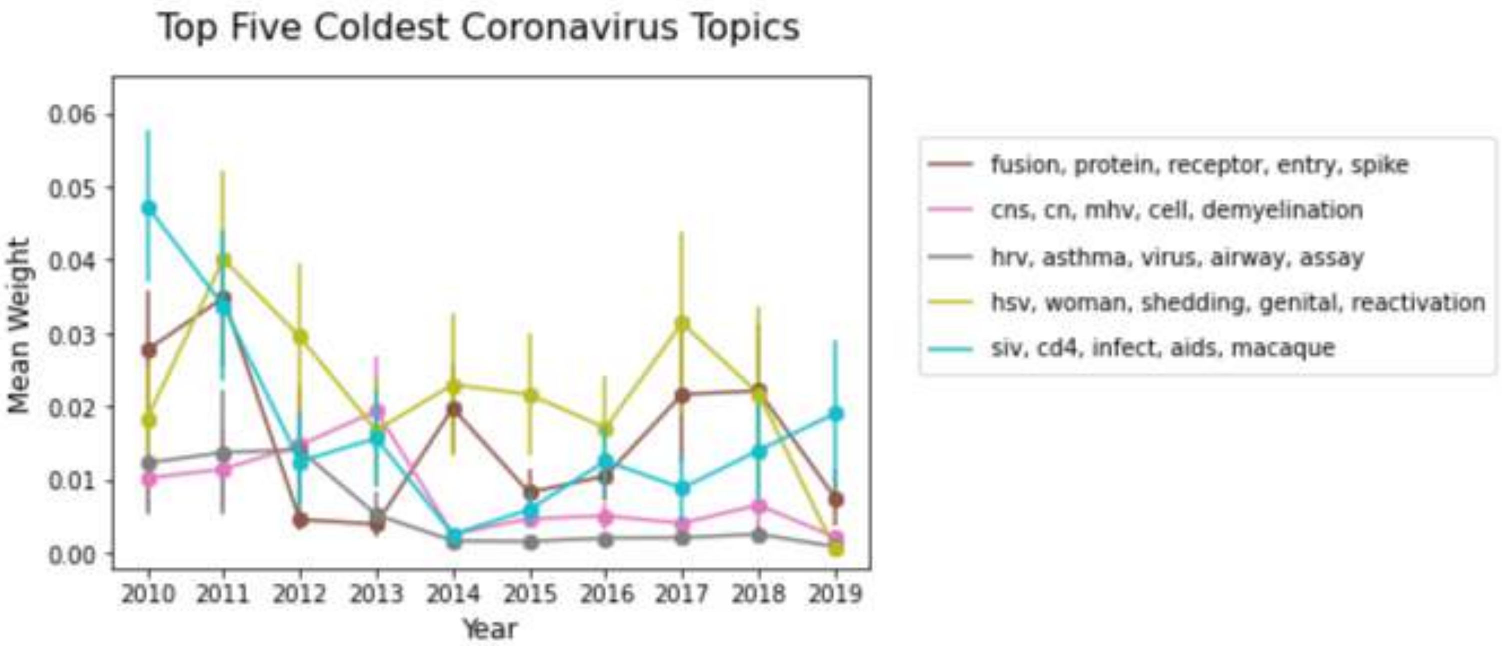
Data Source: [Federal RePORTER](), 2008-2019,
University of Virginia, Social and Decision Analytics Division computations

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
n	105	87	90	55	76	74	59	68	84	77

Case Study Results – “Coronavirus” Corpus

NMF - 30 Topics

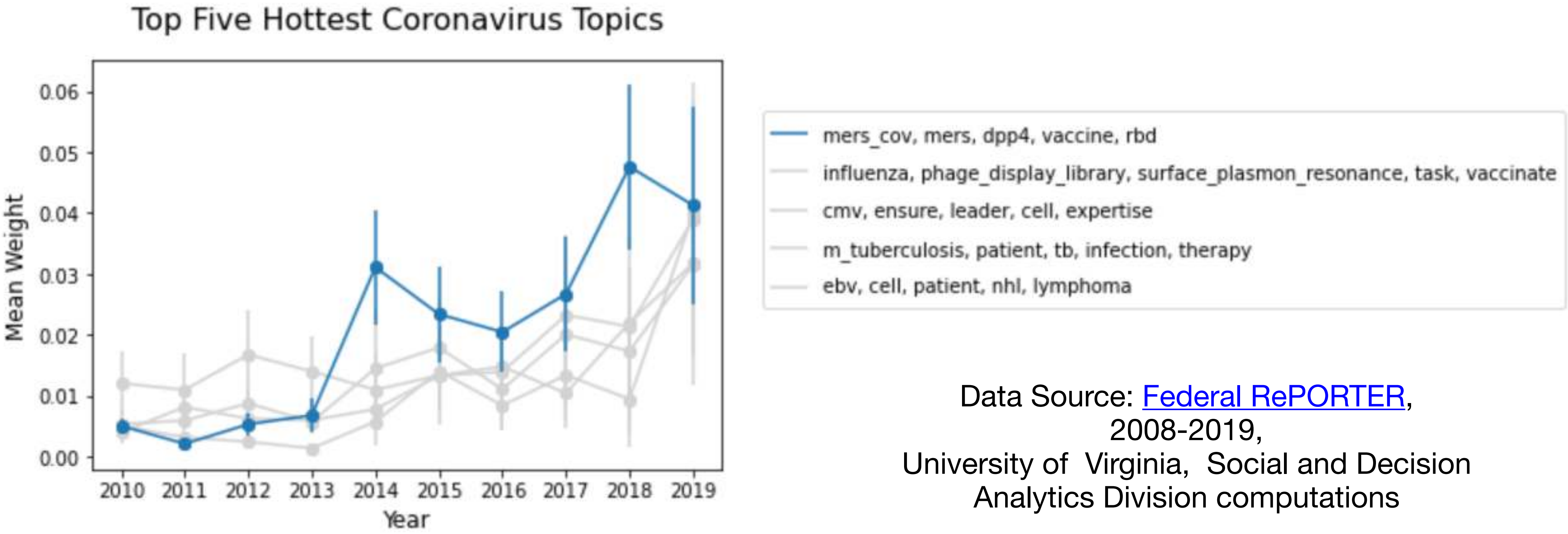




Data Source: Federal RePORTER, 2008-2019,
University of Virginia, Social and Decision Analytics Division computations

Case Study Results – “Coronavirus” Corpus

NMF - 30 topics

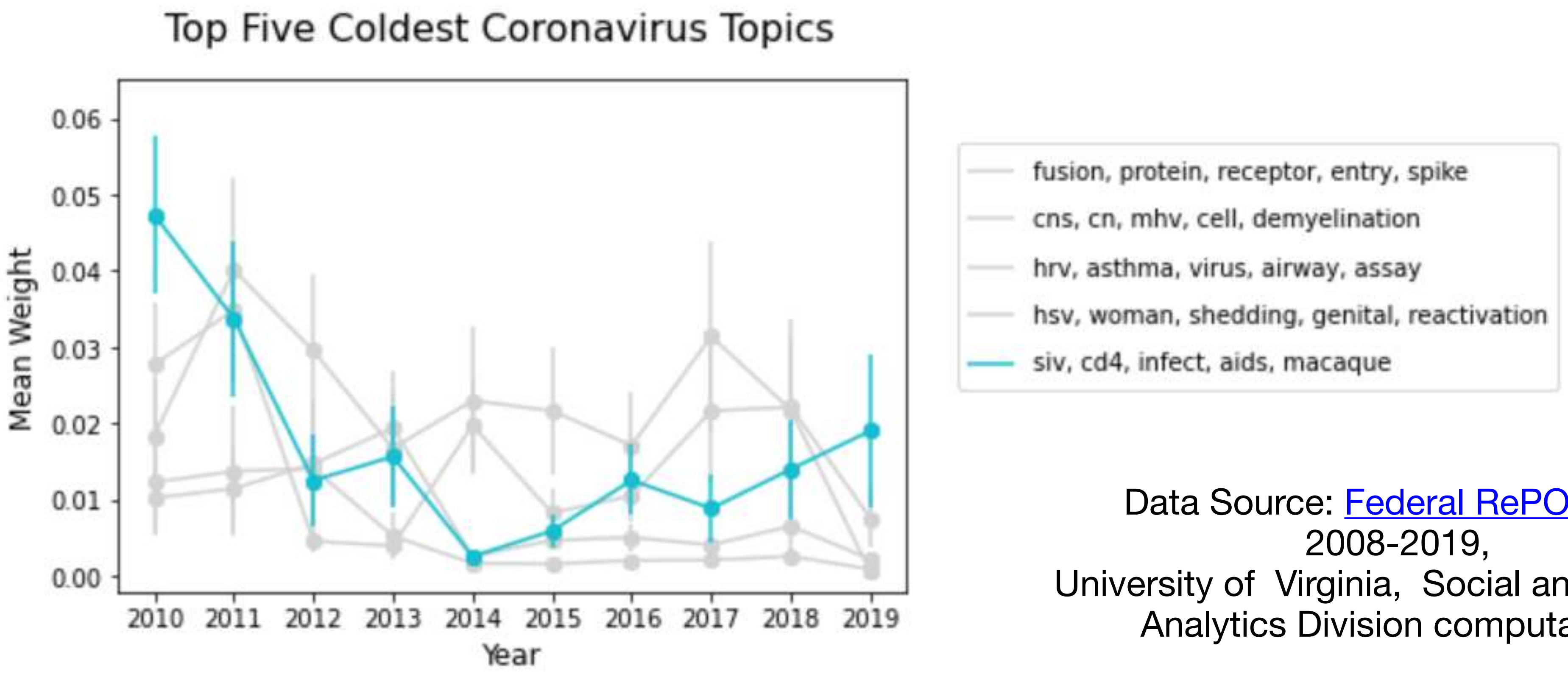


Data Source: [Federal RePORTER](), 2008-2019,
University of Virginia, Social and Decision Analytics Division computations

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
n	89	74	71	69	62	80	89	51	45	30

Case Study Results – “Coronavirus” Corpus

NMF - 30 Topics



Data Source: Federal RePORTER, 2008-2019, University of Virginia, Social and Decision Analytics Division computations

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
n	89	74	71	69	62	80	89	51	45	30

Takeaways

- We discovered that emerging topics methods can give the user valuable information about the popularity of topics over time and current research trends.
- We implemented topic modeling for specific areas of interest:
  - We combined a search engine strategy and topic model
  - This can quickly give users a “deeper dive” into a specific area
  - Allows for finer grained topics
- Our dashboard provides results and documentation for the project: http://rnd.policy-analytics.net/ (temporary link)

Next Steps

- Assess whether the slopes of each topic trend line are significantly different from 0.
- Quantify uncertainty around topic model results.
- Investigate hierarchical models and how we may integrate them into our research.
- Begin researching methods for statistically aggregating results from multiple topic model runs.

Acknowledge our
Data Science for the Public Good Students

Lara Haase

Graduate Fellow

Lara is pursing a Masters of Science in Public Policy & Management - Data Analytics at Carnegie Mellon.

Martha Czernuszenko

Intern

Martha recently graduated from The University of Texas where she studied Information Systems & Business Honors.

Liz Miller

Intern

Liz is an incoming senior at William and Mary where she studies International Relations & History.

Sean Pietrowicz

Intern

Sean recently graduated from Notre Dame where he studied Applied Computational Math & Statistics

References

[1] Schofield, A., Magnusson, M., Thompson, L., & Mimno, D. (2017). Understanding text pre-processing for latent Dirichlet allocation. Proceedings of the 1st Workshop for Women and Underrepresented Minorities in Natural Language Processing. https://www.cs.cornell.edu/~xanda/winlp2017.pdf.

[2] Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, USA, 101(1), 5228-35. https://doi.org/10.1073/pnas.0307752101.

[3] Lee, H., & Kang, P. (2018). Identifying core topics in technology and innovation management studies: A topic model approach. *Journal of Technology Transfer,* 43, 1291-1317. https://doi.org/10.1007/s10961-017-9561-4.

Other references are on the dashboard