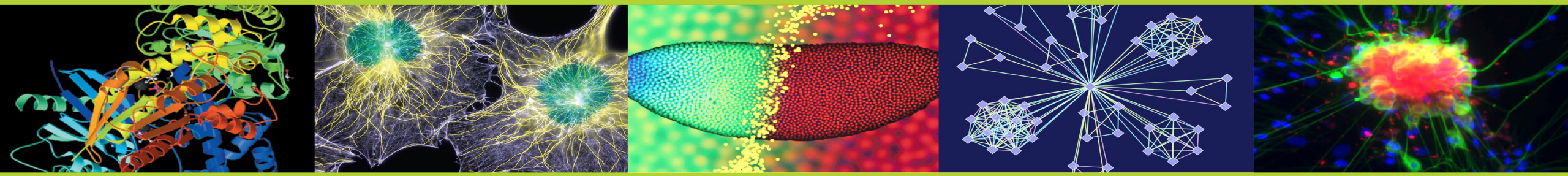
Eshaan Agrawal
and
Jordan Jomsky

Coding it Forward Fellows 2021
NIGMS Division of Data Integration, Modeling, and Analytics (DIMA)

National Institute of General Medical Sciences

Coding it Forward Presentation



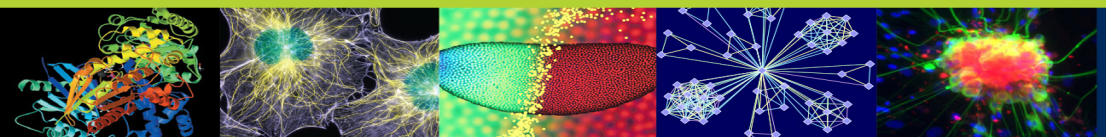
Introductions

Eshaan Agrawal
Sophomore at the University of Georgia
Majoring in Cognitive Science and Computer Science
Hobbies include playing tennis, running, reading, and watching Atlanta sports teams

Jordan Jomsky
Graduate from the University of California, Berkeley
B.A. in Data Science and Molecular and Cell Biology
Hobbies include swimming, cycling, and making Spotify playlists

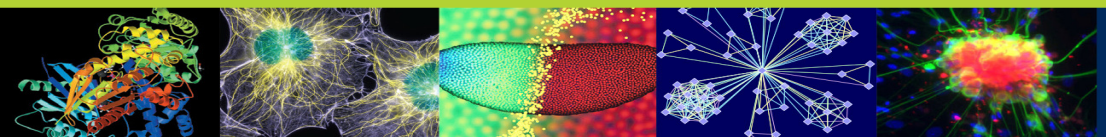
Background on UNITE

- **Goal:** identify and address structural racism within the NIH-supported and the greater scientific community.
- Launched in March 2021
- **Our committee:** N committee
  - **Goal:** Identify and support health disparity, minority health, and health equity research and ensure NIH-wide transparency, accountability, and sustainability in doing so.



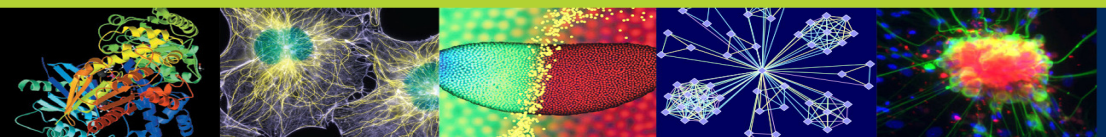
Problem and Goals

- **Goal**: to better identify grants that fall under the Health Disparities (HD) and/or Minority Health (MH) categories
- **Product**: several classifiers that identify HD/MH grant applications
  - Looking at a variety of methods including supervised and unsupervised models
- **Product**: a dashboard displaying the results of the classifiers
  - Shows the performance of the classifiers in comparison to each other
  - Highlights what grants are being misidentified or missed entirely by manual coding of grant applications
  - Explore patterns in grant application identification to determine



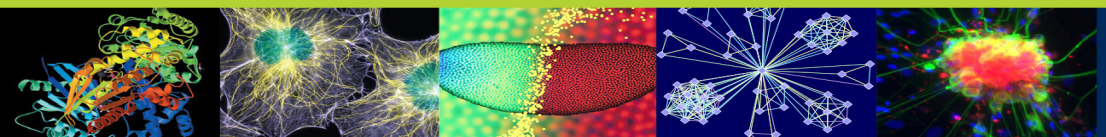
Current Coding Background

- We have a goal of improving classification of HD/MH grants
- NIH currently has an automated classification algorithm that uses a lexical based approach involving term/phrase weights as well as document weights.
- NIH also has a manual HD/MH classification system for HD/MH grants
- In the results, the classifiers we compare to are:
  - **HDMH_IC** – NIH's current automated classifier of HD/MH grants
  - **HDMH_R** – NIH's current manual (human) classifier of HD/MH grants



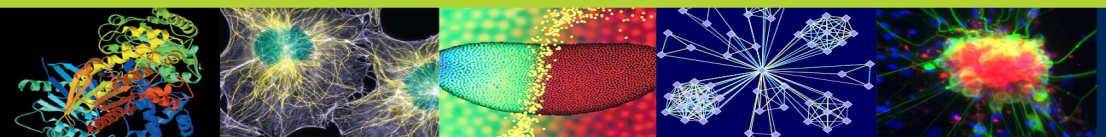
Automated Methods

- **Two main categories**: Non-ML and ML modeling
  - Non-ML
    - HDMH_IC (baseline for comparison)
    - Keyword search (TANS Model)
  - ML
    - Naïve Bayes, Decision Trees, Random Forest, etc.
    - Document Similarity
    - BERT Model



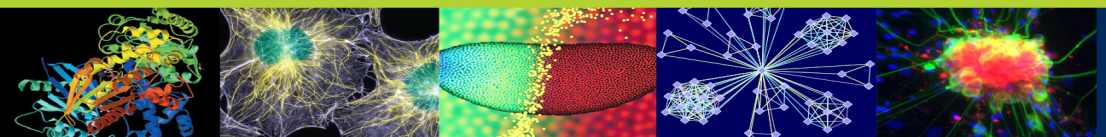
Keywords (in TASN Model)

AAPI | Access to Care | African American | African Ancestry | African Descent | AI/AN | Alaska Native | American Indian | Asian American | Asian Ancestry | Asian Community | Asian Descent | Asian Population | Asians | Bisexual | Black American | Black Ancestry | Black Community | Black Population | Blacks | Care Access | Chicano | Chinese American | Community-Based | Cultural | Disabled | Disadvantaged | Disparit | Equalit | Equit | Ethnic | Gay | Hispanic | Homeless | Homosexual | Immigrant | Indian American | Indigent | Inner-City | Intersex | Japanese American | Korean American | Latina | Latino | Latinx | Lesbian | LGBT | Lower-Income | Low-Income | Mexican | Minorit | MSM | Native American | Native Hawaiian | Pacific Islander | Poverty | Prisoner | Puerto Rican | Queer | Race | Races | Racial | Refugee | Rural | SDOH | SES | Sexuality | SGM| Social Determinants | Social Status | Socioeconomic | Socio-economic | Transgender | Transsexual | Underrepresented | Under-served | Underserved | Uninsured | Urban | Veteran | Vulnerable Population | With Disabilit

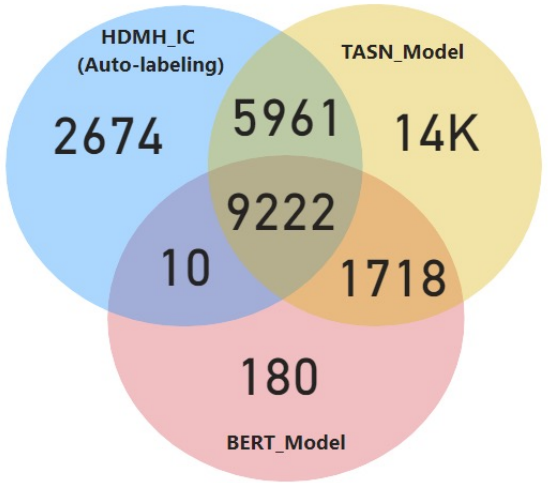


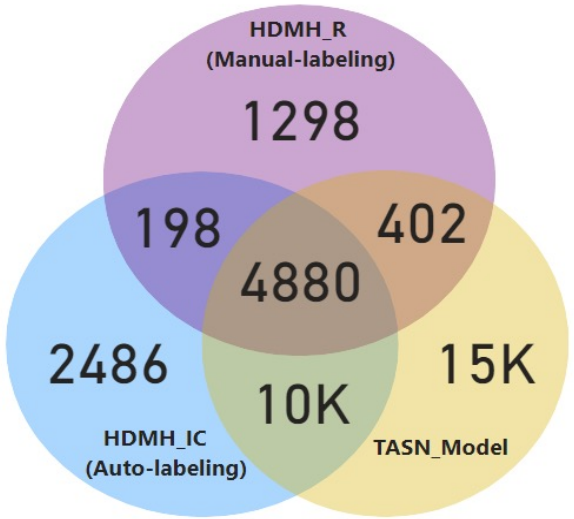
Discussion

- Supervised ML models (Naïve Bayes, Random Forests, Document Similarity) struggle to effectively classify HD/MH grants
  - **Hypothesized reasons:** inconsistent data labeling, no perfect labels, imbalanced data, limits on NLP models
- BERT, our deep learning model, is effective in classifying grants and, in conjunction with keyword analysis, identified some grants that may have been missed by HDMH_IC and HDMH_R.

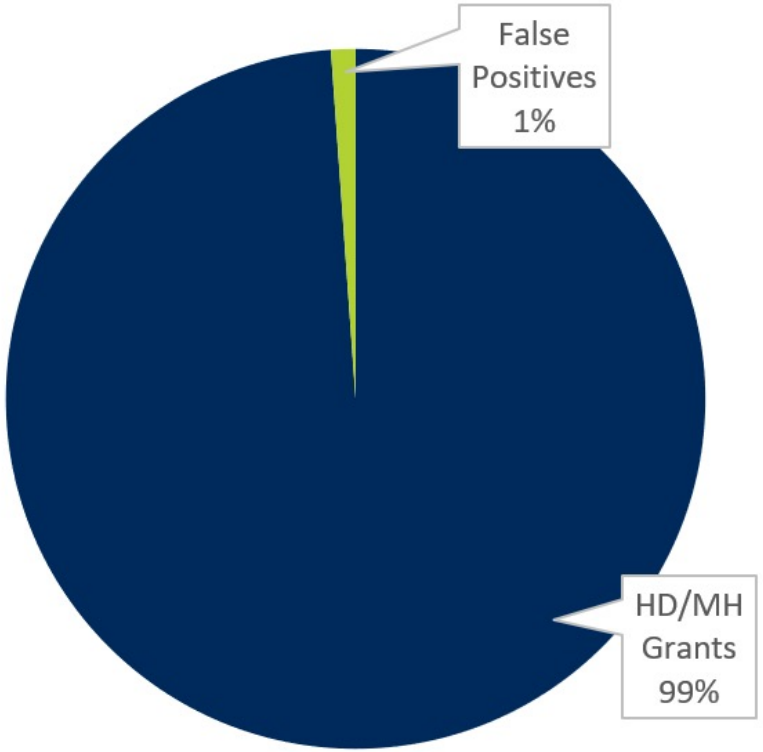


Dashboard/Results



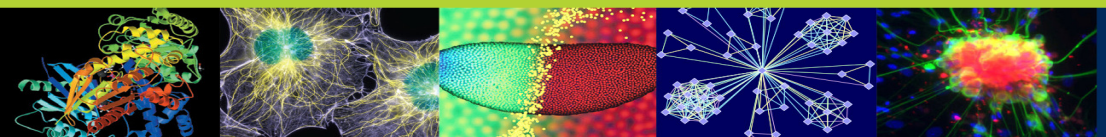


- We created a dashboard and are using it to compare/analyze our models and find insights
  - The pie chart to the right shows grants labeled "Yes" under the BERT model and contains a keyword in the title but is labeled "No" for HD/MH_IC and HD/MH_R
  - False positives contained topics like "dementia in older adults" and "pregnant women"



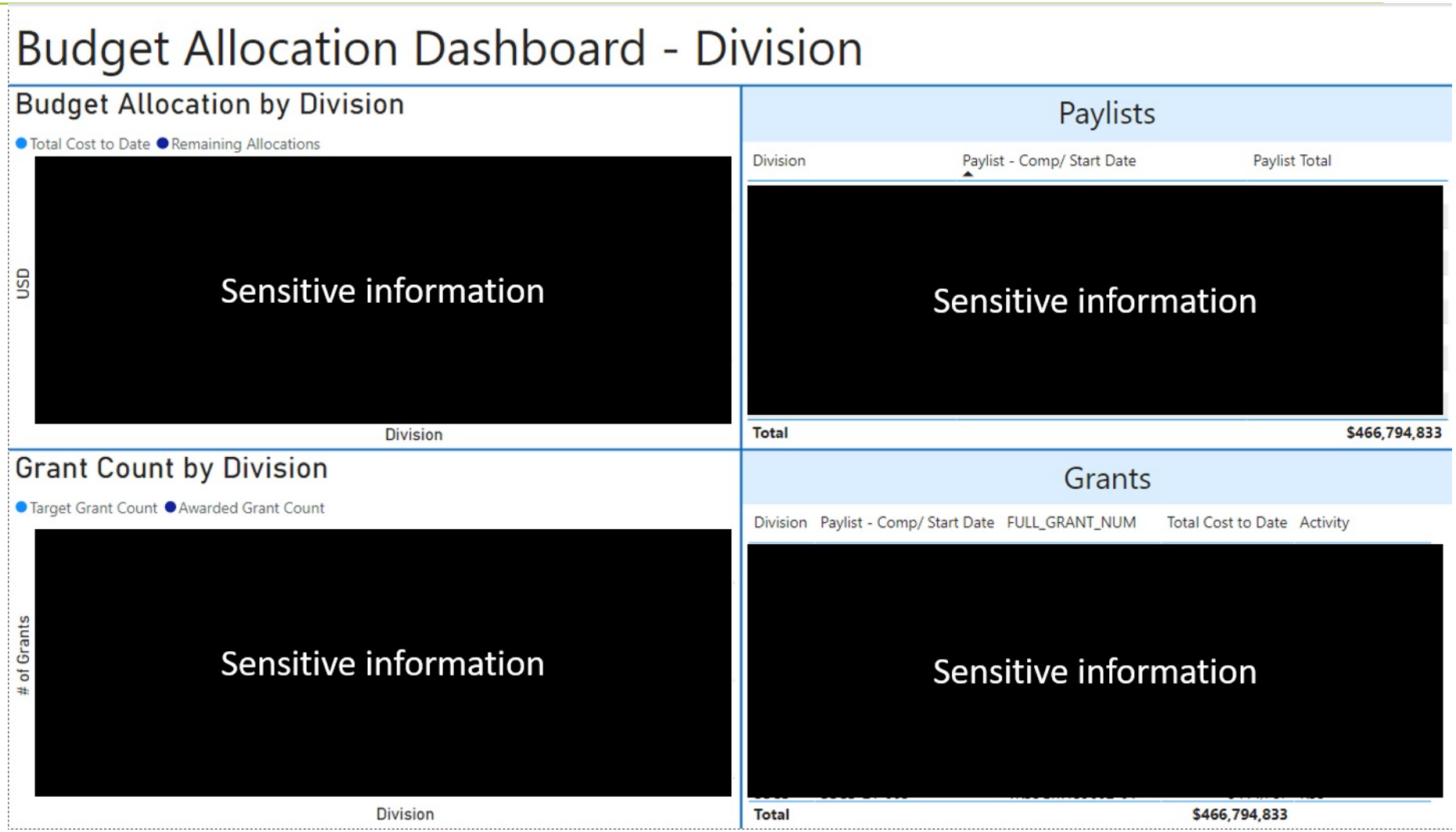
Our Other Projects

We worked on 3 other projects with NIGMS

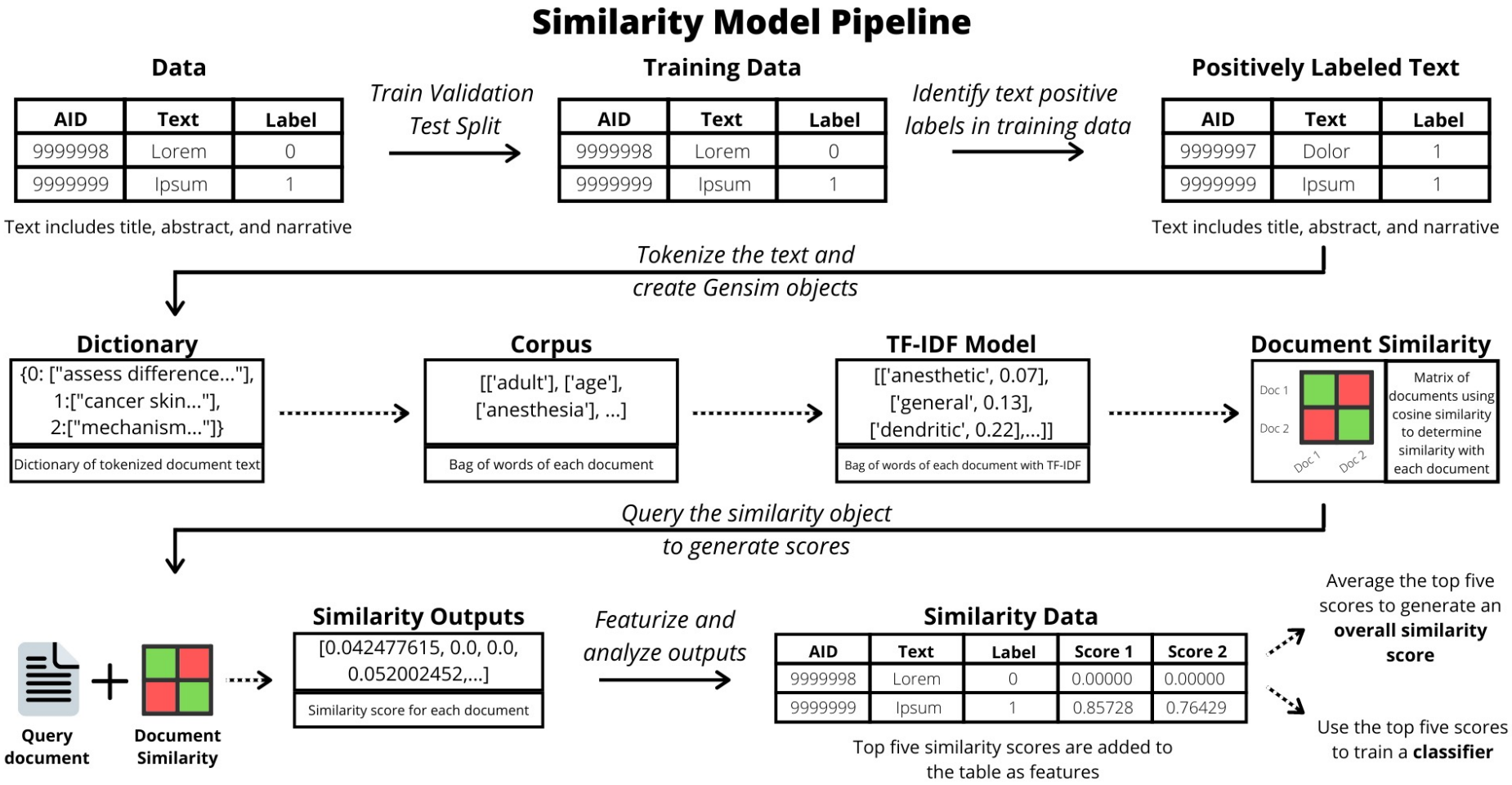


Budget Dashboard

- The screenshot to the right is one of 6 pages on the live dashboard which includes various budget breakdowns and a change tracker.
- Dashboard allows division and institute leaders to easily understand their remaining budget and reduces budget office workload by automating information distribution.



Document Similarity Pipeline



Similarity Model Pipeline

Data

AID	Text	Label
9999998	Lorem	0
9999999	Ipsum	1

Text includes title, abstract, and narrative

Train Validation Test Split

Training Data

AID	Text	Label
9999998	Lorem	0
9999999	Ipsum	1

Identify text positive labels in training data

Positively Labeled Text

AID	Text	Label
9999997	Dolor	1
9999999	Ipsum	1

Text includes title, abstract, and narrative

Tokenize the text and create Gensim objects

Dictionary

{0: ["assess difference..."], 1:["cancer skin..."], 2:["mechanism..."]}

Dictionary of tokenized document text

Corpus

[['adult'], ['age'], ['anesthesia'], ...]

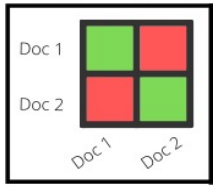
Bag of words of each document

TF-IDF Model

[['anesthetic', 0.07], ['general', 0.13], ['dendritic', 0.22],...]]

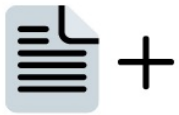
Bag of words of each document with TF-IDF

Document Similarity



Matrix of documents using cosine similarity to determine similarity with each document

Query the similarity object to generate scores



Query document

Document Similarity

Similarity Outputs

[0.042477615, 0.0, 0.0, 0.052002452,...]

Similarity score for each document

Featurize and analyze outputs

Similarity Data

AID	Text	Label	Score 1	Score 2
9999998	Lorem	0	0.00000	0.00000
9999999	Ipsum	1	0.85728	0.76429

Top five similarity scores are added to the table as features

Average the top five scores to generate an **overall similarity score**

Use the top five scores to train a **classifier**

Similarity Model Pipeline

Data

AID	Text	Label
9999998	Lorem	0
9999999	Ipsum	1

Text includes title, abstract, and narrative

Train Validation Test Split

Training Data

AID	Text	Label
9999998	Lorem	0
9999999	Ipsum	1

Identify text positive labels in training data

Positively Labeled Text

AID	Text	Label
9999997	Dolor	1
9999999	Ipsum	1

Text includes title, abstract, and narrative

Tokenize the text and create Gensim objects

Dictionary

{0: ["assess difference..."], 1:["cancer skin..."], 2:["mechanism..."]}

Dictionary of tokenized document text

Corpus

[['adult'], ['age'], ['anesthesia'], ...]

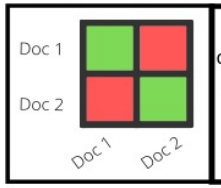
Bag of words of each document

TF-IDF Model

[['anesthetic', 0.07], ['general', 0.13], ['dendritic', 0.22],...]]

Bag of words of each document with TF-IDF

Document Similarity



Matrix of documents using cosine similarity to determine similarity with each document

Query the similarity object to generate scores

Query document

Document Similarity

Similarity Outputs

[0.042477615, 0.0, 0.0, 0.052002452,...]

Similarity score for each document

Featurize and analyze outputs

Similarity Data

AID	Text	Label	Score 1	Score 2
9999998	Lorem	0	0.00000	0.00000
9999999	Ipsum	1	0.85728	0.76429

Top five similarity scores are added to the table as features

Average the top five scores to generate an **overall similarity score**

Use the top five scores to train a **classifier**

Similarity Model Pipeline

Data

AID	Text	Label
9999998	Lorem	0
9999999	Ipsum	1

Text includes title, abstract, and narrative

Train Validation Test Split

Training Data

AID	Text	Label
9999998	Lorem	0
9999999	Ipsum	1

Identify text positive labels in training data

Positively Labeled Text

AID	Text	Label
9999997	Dolor	1
9999999	Ipsum	1

Text includes title, abstract, and narrative

Tokenize the text and create Gensim objects

Dictionary

{0: ["assess difference..."], 1:["cancer skin..."], 2:["mechanism..."]}

Dictionary of tokenized document text

Corpus

[['adult'], ['age'], ['anesthesia'], ...]

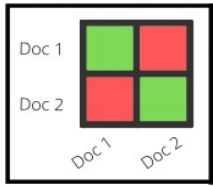
Bag of words of each document

TF-IDF Model

[['anesthetic', 0.07], ['general', 0.13], ['dendritic', 0.22],...]]

Bag of words of each document with TF-IDF

Document Similarity



Matrix of documents using cosine similarity to determine similarity with each document

Query the similarity object to generate scores

Query document + **Document Similarity**

Similarity Outputs

[0.042477615, 0.0, 0.0, 0.052002452,...]

Similarity score for each document

Featurize and analyze outputs

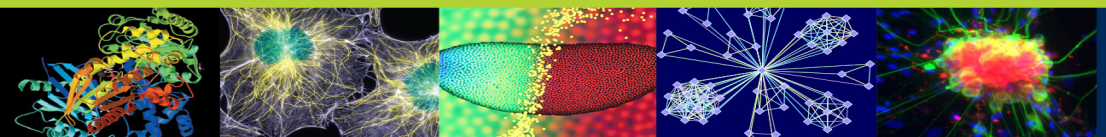
Similarity Data

AID	Text	Label	Score 1	Score 2
9999998	Lorem	0	0.00000	0.00000
9999999	Ipsum	1	0.85728	0.76429

Top five similarity scores are added to the table as features

Average the top five scores to generate an **overall similarity score**

Use the top five scores to train a **classifier**



Generative Modeling

'Abstract Comprehensive High Impact Prevention (CHIP) services have reduced new HIV infections in major HIV epicenters, but many CHIP goals have not been met. A fundamental limitation is the failure to test 90% of at risk persons. HIV testing is critical to the success of CHIP. Currently, the HIV testing system fails to reach segments of the men-who-have-sex-with men (MSM) population. Moreover, because CHIP is costly to sustain, expansion of the current venue-based testing system (i.e., clinical & non-clinical sites) poses a financial burden for departments of public health (DPHs) with limited budgets. Further, HIV testing costs increase as HIV prevalence decreases. Consequently, DPHs in urban areas with lower HIV prevalence face an additional financial burden. Lower prevalence areas are important because they account for a large proportion of new HIV infections. Identifying cost-efficient methods to increase HIV testing in these communities is key to reducing the spread of HIV. The proposed R34 will conduct formative research to develop a low-cost dissemination intervention designed to increase the uptake of HIV testing in an urban community (Portland, Oregon) with a large MSM population and lower HIV prevalence. Using a community-engaged approach, we will develop a culturally sensitive intervention to increase the uptake of no-cost Oral-Self-Implemented HIV testing (Oral-SIT) and facilitate Oral-SIT distribution. Guided by the Push-Pull Infrastructure Model (PPIM), we will develop intervention core components related to the Push and Infrastructure factors of the PPIM: 1) an Information-Motivation (IM) component (Push factor) to inform and motivate uptake and correct use of Oral- SIT; and 2) a Distribution-network component to build community-based infrastructure by organizing a network (i.e., local DPH, LGBTQ businesses & associated cultural events) to distribute Oral-SIT kits (Infrastructure factor). We will conduct a formative test of the intervention over a period of six months (i.e., deliver the IM component, distribute 3000 Oral-SIT kits). Process evaluations at mid-point and the end of the formative intervention (e.g., focus groups, street intercepts) will assess intervention awareness, acceptability, implementation feasibility, and potential sustainability, and, specific to the IM component, potential reach. We will assess Oral-SIT uptake over the course of the intervention (weekly). Lastly, we will develop and evaluate a supplemental component to the local public health surveillance system that may be used to document new HIV infections that occur as a function of increased Oral-SIT use. Overall, the results of this study will provide critical information for the construction and evaluation of a clinical trial of a multi-component dissemination intervention in lower prevalence cities that will identify new HIV cases and link them to care (R01). 1'

```
In [36]: generate_summary(unite.sample(1)['Abstract'].item(), top_n=2)
```

```
Summarized Text:
 The proposed R34 will conduct formative research to develop a low-cost dissemination intervention designed to increase the uptake of HIV testing in an urban community (Portland, Oregon) with a large MSM population and lower HIV prevalence. HIV testing is critical to the success of CHIP
```