NIAID Data Science (NDS) Dashboard

Office of Data Science and Emerging Technologies
Anuva Banwasi and Snipta Mallick

CIVIC DIGITAL FELLOWSHIP

NIAID

National Institute of Allergy and Infectious Diseases

Dr. Chris Marcum
Staff Scientist
Mentor

Dr. Steve Tsang
Health Scientist
Mentor

Introductions

Anuva Banwasi

Software Engineering Fellow

Columbia University ‘24

Computer Science

Snipta Mallick

Software Engineering Fellow

University of Texas at Dallas ‘22

Computer Science & Cognitive Science

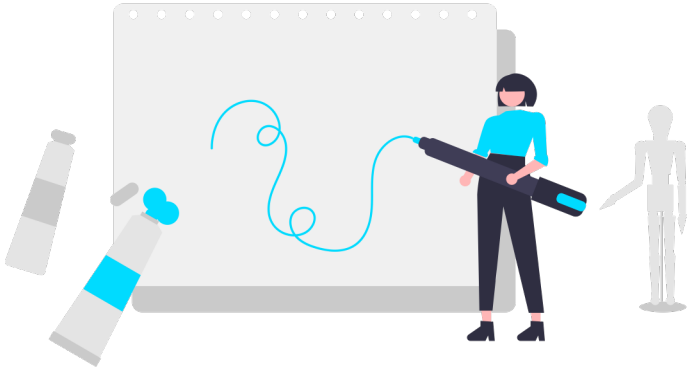
NIH National Institute of Allergy and Infectious Diseases

CIVIC DIGITAL FELLOWSHIP

NIAID

Problem Statement

Absence of an **integrated**, **automated** workflow to analyze, categorize, and visualize NIAID data science projects to streamline the funding application management process

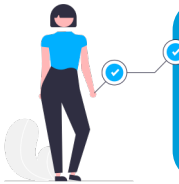


National Institute of Allergy and Infectious Diseases

CIVIC DIGITAL FELLOWSHIP

NIAID

Civic Digital Fellowship Project Goals

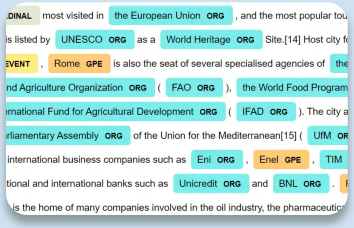


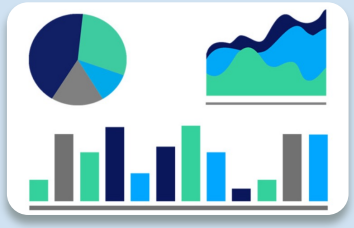
Develop an automated pipeline to characterize and analyze NIAID funded data science projects

Build a platform-independent dashboard to annotate and visualize the rich portfolio

Compile program scripts and documentation in Github repository

Project Components





NIH RePORTER Data Module

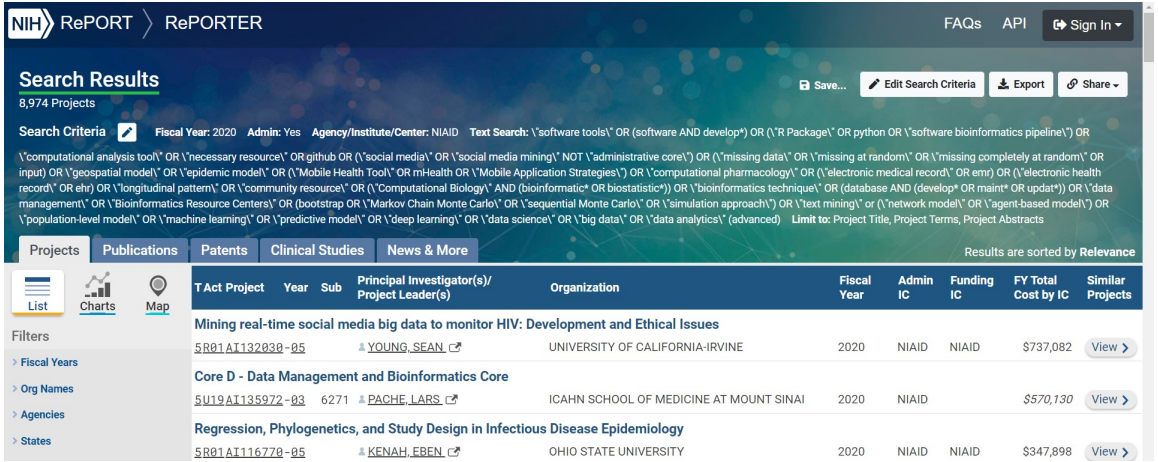
Machine Learning (ML) Classifier Module

Named Entity Recognition (NER) Module

NIAID Data Science (NDS) Dashboard

Data Module

- Extract publicly-available funded project data from NIH RePORTER
  - Title, Abstract, Project Number, Fiscal Year, Funding Institute, etc.
- Pipeline script that downloads data and runs classifier on project abstracts



Machine Learning Classifier

Determine probability that a project is data-science related

1

Load and Prepare Dataset

Bag of words model

2020 RePORTER projects within NIAID

Manually coded:
200 positives
200 negatives

word	weight
lorem	17
ipsum	16
amet	13
sit	13
vitae	13
eget	12
quis	11
sem	11
massa	10
nec	10
sed	10
semper	9
egestas	8
libero	8
nisl	8
odio	8
scelerisqu	8

Figures Designed By Dr. Chris Marcum

NIH National Institute of Allergy and Infectious Diseases

CIVIC DIGITAL FELLOWSHIP

NIAID

Machine Learning Classifier

2

Train and test supervised ML models

2020 RePORTER NIAID manually coded data

- Linear SVC, Ridge classifier, Perceptron, etc.
- Outputs probability that project is data science

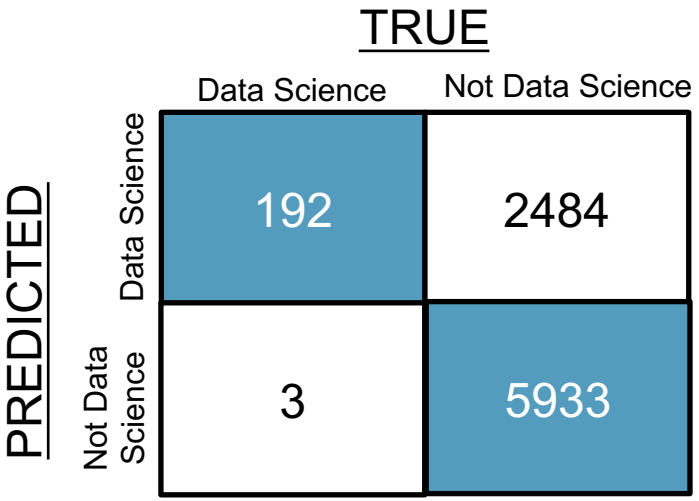
3

Second test on 2019 RCDC data

Research, Condition, and Disease Categorization

Classifier Validation

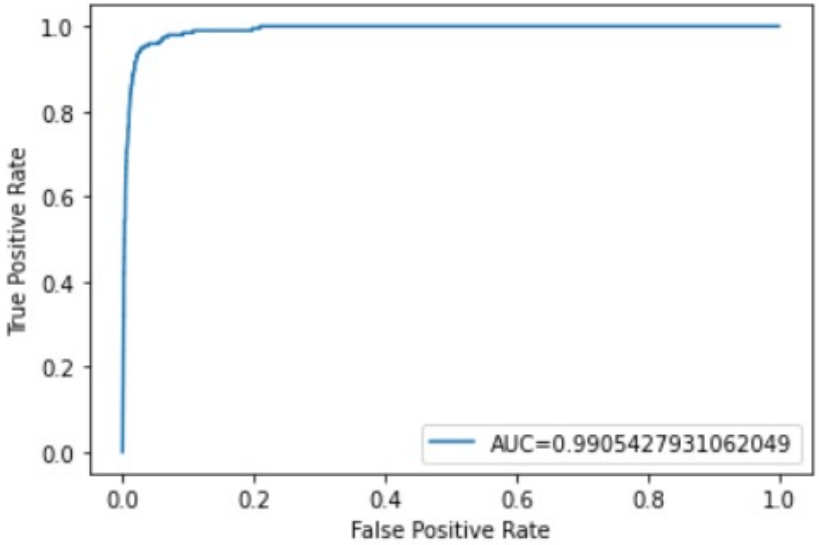
2019 RCDC Data Science



2020 NIAID RePORTER

ML Classifier Results and Insights

ROC Curve for 2020 Grants



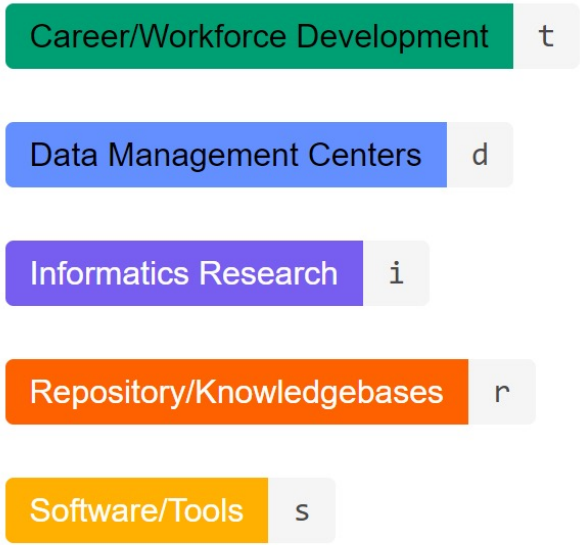
Ridge classifier

F1 score: 0.94
Weighted average: 0.92

Named Entity Recognition Module

" Biostatistics / Epidemiology Training Grants in AIDS. Summary/Abstract This is a request for an extension of the Biostatistics and Epidemiology in AIDS Training Program at the Harvard T. H. Chan School of Public Health (HSPH). The program prepares pre-doctoral and postdoctoral fellows in the application of biostatistics and epidemiology to HIV research. The program is an active collaboration between the Departments of Biostatistics and Epidemiology. Trainees receive high-quality instruction in basic biostatistical theory and methods, such as probability, statistical inference, computing and data analysis. The program also provides training in specialized topics of particular relevance for HIV/AIDS applications, such as survival, longitudinal and multivariate data analysis, causal inference and statistical genetics. Options are available for training in

Sample NER Annotation



ODSET Data Science Subcategories

National Institute of Allergy and Infectious Diseases

CIVIC DIGITAL FELLOWSHIP

NIAID

Named Entity Recognition Model

1. **Create Training Set using Positively Labeled Data**
   Doccano: open-source text annotation tool
2. **Build Natural Language Processing Model**
   spaCy: open-source NLP python library
3. **Display Named Entity Recognition Results**
   spacy-streamlit: library to visualize spacy models

NDS Dashboard

Select Data Parameters

Year Range

2019 2020

2019 2021

IC Name

NIAID

Data Science

All

Data Science Category

All

Core Project Number

All

Funding Mechanism

All

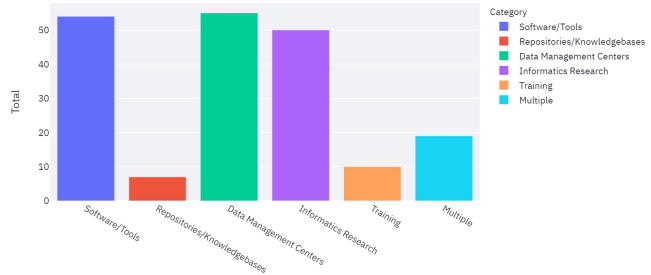
Activity Code

All

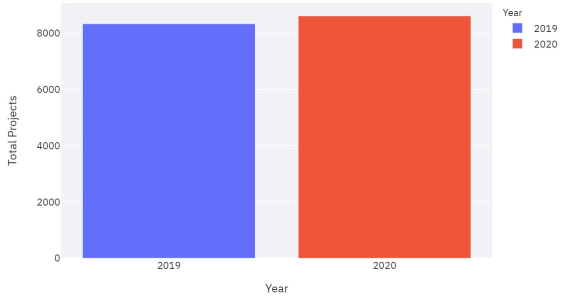
Application Type

All

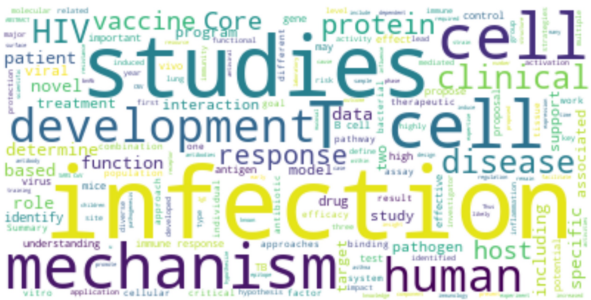
Total Projects by Data Science Category



Total Number of Projects by Fiscal Year



Word Cloud



Model View

NDS Grants Table

APPLICATION_ID	ACTIVITY	ADMINISTERING_IC	APPLICATION_TYPE	ARRA_FUNDED	AWARD_NOTICE_DATE	BUDGET_START	BUDGET_END	CFDA_CODE	CORE_PROJECT_NUM	ED_INST_TYPE	FOA_NUMBER	FULL_PROJECT_NUM
9428897	K22	AI	1	N	12/7/2018	12/7/2018	11/30/2019	855	K22AI134993	SCHOOLS OF ARTS AND SC...	PAR-16-434	1K22AI
9433315	R03	AI	1	N	2/22/2019	2/25/2019	1/31/2020	855	R03AI135322	SCHOOLS OF PUBLIC HEAL...	PA-16-163	1R03AI
9463225	U19	AI	2	N	1/10/2019	2/1/2019	1/31/2020	855	U19AI104204	SCHOOLS OF MEDICINE	RFA-AI-16-065	2U19AI
9463226	U19	AI	2	N	1/10/2019	2/1/2019	1/31/2020	NaN	U19AI104209	nan	RFA-AI-16-065	2U19AI
9463227	U19	AI	2	N	1/10/2019	2/1/2019	1/31/2020	NaN	U19AI104206	nan	RFA-AI-16-065	2U19AI
9463228	U19	AI	2	N	1/10/2019	2/1/2019	1/31/2020	NaN	U19AI104208	nan	RFA-AI-16-065	2U19AI
9463229	U19	AI	2	N	1/10/2019	2/1/2019	1/31/2020	NaN	U19AI104205	nan	RFA-AI-16-065	2U19AI
9463230	U19	AI	2	N	1/10/2019	2/1/2019	1/31/2020	NaN	U19AI104207	nan	RFA-AI-16-065	2U19AI
9463231	U19	AI	2	N	1/10/2019	2/1/2019	1/31/2020	NaN	U19AI104209	nan	RFA-AI-16-065	2U19AI
9463232	U19	AI	2	N	1/10/2019	2/1/2019	1/31/2020	NaN	U19AI104209	nan	RFA-AI-16-065	2U19AI
9483499	U01	AI	1	N	3/17/2019	3/17/2019	4/30/2020	855	U01AI135708	SCHOOLS OF PUBLIC HEAL...	PAR-16-270	1U01AI

Choose Application ID

9428897

Machine Learning Classifer Module

Data Science Probability

0.465763

Current Data Science Label

No

Manual Data Science Labels

No Label

NIH National Institute of Allergy and Infectious Diseases

CIVIC DIGITAL FELLOWSHIP

NIAID

Potential Users

Researchers

Strategy Analysts

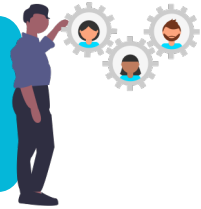
Program Officers

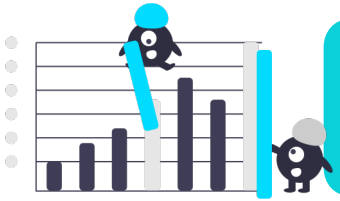
Portfolio Analysts

Future Directions

Support data analysis initiatives for additional institutes and projects across the NIH

Visualize the downstream impacts of grant funding by displaying network analysis models

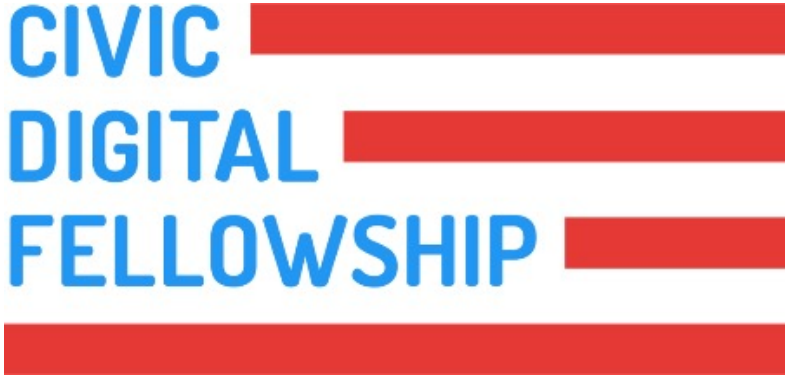




Add containerization to improve the FAIRness and reproducibility

Acknowledgements

- **NIAID/ODSET**
  - Dr. Chris Marcum
  - Dr. Steve Tsang
  - Sydney Foote
  - Sara Jones
- **Coding it Forward**
  - Rachel Dodell
  - Ariana Soto



Thank you!

Any questions?

Slide Image Sources

- Figures on Slide 6
  - Google Images
- Figures on Slides 4, 5, and 15
  - Katerina Limpitsouni, Undraw.co
- Figures on Slide 8
  - Dr. Chris Marcum