Show me the money!:
A roadmap to biomedical grants

Lo-Yu Chang Ding Ding Hui Yao Linda Ye
Team Quarto_quartet.qmd

Research Question and Relevance

Federal funding is the leading source of scientific research, accounting for nearly 40% of research and development expenditures. In the current sociopolitical circumstances, securing research funding is becoming less certain and even less predictable in the United States. This differentially impacts academic institutions and researchers, particularly students and junior to mid-level faculty at large, elite research institutions such as Johns Hopkins, which receives the most federal grant dollars in the country. As we near the end of 2025, it is apt to look back on the year to evaluate how funding dynamics have changed in order to prepare for the future.

We aim to aggregate and analyze available public data from the separate federal funding agencies on currently and previously awarded biomedical, clinical, and human research topics and describe the trends in allocation of federal research grant awards over the past five years (2020-2025). We will use this information to generate an aggregated, digestible data visualization dashboard and create an interactive analytic tool incorporating machine learning methods to predict the most appropriate funding mechanism based on the user's proposed research domain and subdomain, and university affiliation, among other details. The intended audience is students, junior faculty, and any researcher interested in current trends in research activity and funding allocation.

Previous Evidence and Work in the Field

Compared to 2024, federal spending on scientific research has decreased by 53% across multiple agencies since the new administration took office in January 2025, and at least $5 billion has been frozen at some point at multiple elite research institutions (Malakoff and Brainard 2025). These federal funding agencies have publicly available search tools to identify active projects and previously awarded grants, however these websites are maintained separately by each

department, and the advanced search criteria are too clunky, specific, and impractical for the average user to synthesize efficiently.

Project Outline

1. **Identify Funding Agencies & Data Sources**
   Compile relevant agencies funding biomedical research, grant information, and available APIs and datasets. We decided to focus on the largest funding sources of biomedical research: National Institute of Health (NIH) and National Science Foundation (NSF). Within the NIH RePORTER database, we will also include data from the Agency for Healthcare Research and Quality (AHRQ), Centers for Disease Control and Prevention (CDC), Food and Drug Administration (FDA), and Veterans Affairs (VA).

2. **Data Scraping**
   Use a combination of APIs and direct export of search results to collect raw data from each funding source. NIH, NSF, and HHS provide publicly available API and batch data.

3. **Database Building**
   Organize scraped data into a structured database with a clear schema and consistent variables.

4. **Data Cleaning**
   Standardize variables, remove errors/duplicates, and build a data dictionary.

5. **Data Aggregation**
   Combine and summarize data across years, agencies, institutions, research domains and subdomains, and award amounts. Identify key words in available funding proposals and abstracts.

6. **Exploratory Analysis**
   Generate descriptive statistics and data visualization to identify patterns and trends over a 10 years span. Compile a word cloud using key words from abstracts and titles to better identify funded research topics.

7. **Creating an Interactive Dashboard**
   Build an interactive dashboard in R Shiny for users to explore funding data.

8. **Building an Analytic Tool**
   Explore supervised ML models to construct scoring metrics for features that result in successful funding or predict funding trends.

9. **Presentation**
   Prepare a presentation and website summarizing methods, data visualization, and descriptive results. The analytic tool will be incorporated into the website as an interactive feature.

10. **Final Write-up and Project Deployment**
    We will compile the final report integrating all findings, and the website will be hosted on an interactive dashboard as mentioned above to allow access from external users.

Data and Database Accessibility

The project will draw from two primary data sources: NSF Award Search (U.S. National Science Foundation 2025a) and NIH RePORTER (U.S. National Institutes of Health 2025a), which also houses award information from the Agency for Healthcare Research and Quality (AHRQ), Centers for Disease Control and Prevention (CDC), Food and Drug Administration (FDA), and Veterans Affairs (VA). Both APIs are publicly accessible and do not require API keys. The links to the APIs are: NIH RePORTER (https://api.reporter.nih.gov/) (U.S. National Institutes of Health 2025b) and NSF Award Search (https://www.nsf.gov/digital/developer)(U.S. National Science Foundation 2025b).

For NIH RePORTER and NSF Award Search, the data provides project-level information such as titles, abstracts, funding organizations, institutions, fiscal years, research categories, grant amounts. In addition, the NIH provides research grant success rates. We will collect data from projects that fall under biomedical and human health–related projects using direct CSV downloads from html and automated API calls. The JSON files will be parsed using the `jsonlite` package in R and converted into CSV files for cleaning and analysis.

Programming Paradigms

For this project, we will rely on four main programming paradigms that align with both our data sources and analytic goals: a version-control and collaboration paradigm (Git + GitHub), a functional programming paradigm in R, a data collection paradigm based on APIs and web scraping, and a machine learning workflow paradigm.

First, we will follow a **version-control and collaboration paradigm** using Git and GitHub. All code, Quarto documents, and configuration files will reside in a shared repository. Branching, pull requests, and commit history will allow us to experiment, review each other's changes, and maintain a reproducible record of the analytic tool's evolution over the semester.

Second, our R code will adopt a **functional programming paradigm**, particularly through the `purrr` "map" family of functions. We will write small, reusable functions for tasks such as calling an API for a given agency and year, cleaning a raw JSON response, and constructing features for a single modeling run. These functions will then be mapped over lists of agencies, years, or endpoints, reducing repetitive code and simplifying extensions to new data sources.

Third, we will employ a **data collection paradigm centered on APIs and web scraping**. Whenever possible, we will query official APIs (e.g., NIH RePORTER and NSF Award Search)

to obtain structured JSON data. When key information is only available on web pages, we will use web scraping tools (e.g., rvest and SelectorGadget) with clearly defined selectors. Treating data collection as its own paradigm ensures a systematic, documented approach for acquiring and updating funding data.

Finally, we will organize our predictive models within a **machine learning workflow paradigm**. We plan to treat the problem as a supervised learning task (for example, predicting funding amounts or the probability of award given project characteristics) and to implement a consistent workflow for data splitting, preprocessing, model fitting, tuning, and evaluation. This ML workflow paradigm will make it easier to compare models and integrate the final model into our analytic tool.

Packages and Software

Software

`VScode` and `Positron` will be our main IDEs for the project. We will use `Quarto` as the main markdown language. A shared data repository will be hosted on `GitHub` for collaboration and version control. The final website and interactive dashboard will be deployed using `shinyapps.io`.

Language and Packages

The main language will be `R`. We plan to use the packages `httr` and `jsonlite` to retrieve data from publicly available APIs, `tidyverse` (e.g. `tidyr`, `dplyr`, `purrr`, `ggplot2`) for data wrangling, analysis, and visualization, `tidymodels` for machine learning analyses, and `shiny` for deploying the final interactive dashboard.

Proposed Data Analytic Product

After obtaining grant and project information from both the NIH and NSF and completing data wrangling, we plan to implement the data in three major ways, which will also be showcased in the final `Shiny` dashboard. A brief overview is provided below:

Tab 1. Interactive data explorer

In this section, we will use aggregated grant data from both the NIH and NSF to create an interactive dashboard that allows users to search across all available fields in the dataset. For example, users can look up a particular principal investigator's total grant amount for a given year or an institution's total awarded funding across both agencies. Although the NSF and NIH each provide their own interactive search tools, their data are not aggregated. Our platform aims to offer a more comprehensive view of the biomedical funding landscape. This resource would be valuable for prospective graduate students seeking an overview of a potential PI's funding situation and current research direction, as grant activity often reflects emerging research trends more accurately than published work, which typically lags behind.

Tab 2. Analysis and data visualization

In this section, we will take advantage of the aggregated dataset to analyze annual changes in grant amounts, shifts in funding agencies, and the top awarded institutions. In addition to data visualization, we also plan to include a brief discussion offering our perspective on the observed trends in biomedical funding.

Tab 3. Grant matcher and statistical modeling

In this section, we plan to use the aggregated dataset along with abstract keywords, funding agencies and their subdivisions, awarded amounts, and historical funding success rates per project type to train a machine learning model that predicts the most suitable funding agency and potential award amount based on a user-provided keyword. We hope this matcher will help users navigate different funding agency divisions and identify better fits based on their research topics.

Tentative Timeline

Nov 10 - Nov 16	Write up the proposal
	Explore the APIs
Nov 17 - Nov 23	Download the data from both agencies
	Data cleaning and wrangling
	Exploratory data analysis to see what analytical angles are feasible
Nov 24 - Nov 30	Data analysis and visualization
	Machine learning model design and training
Dec 1 - Dec 7	`Shiny` interactive dashboard implementation
	Website deployment and troubleshoot

Dec 8 - Dec 14	Final project presentation Write up the analytical results in the final report
Dec 15 - Dec 21	Write up the analytical results in the final report

Proposed Tasks for Team Members

Lo-Yu Chang

Data scraping and cleaning, design of analytic tool, webpage building

Ding Ding

Data scraping and cleaning, data cleaning, statistical analysis, webpage building

Hui Yao

Data cleaning, webpage building, statistical analysis

Linda Ye

Data scraping, statistical analysis, webpage design and building, presentation

References

Malakoff, David, and Jeffrey Brainard. 2025. "How Trump Upended Science." *Science*, 577.

U.S. National Institutes of Health. 2025a. *NIH RePORTER - Grants and Awards Portal*. https://reporter.nih.gov/.

———. 2025b. *NIH RePORTER API Documentation*. https://api.reporter.nih.gov/.

U.S. National Science Foundation. 2025a. *NSF Award Search*. https://www.nsf.gov/awardsearch/simple-search/.

———. 2025b. *NSF Developer Resources - Open Government*. https://www.nsf.gov/digital/developer.