IS 525: Data Warehouse and BI

Professor Michael Wonderlich

Submitted By:

Arundhati Raj (raj9)

Abstract

The scenario for this project is focused on developing an interactive, data-driven dashboard to explore two key datasets from the Author-ity 2009 collection. The goal is to offer comprehensive visual insights into the research and innovation landscape, particularly related to NIH/NSF grants, academic publications, patents, and researcher demographics.

Problem Statement

The main problem this project aims to solve is to provide a unified research analytics platform that can illuminate hidden patterns in research funding, institutional collaboration, and demographic representation across the academic sector. The dashboards will serve multiple stakeholders, including policymakers, academic institutions, and funding agencies, by delivering actionable insights to support data-driven decision making.

Intended Audience

The primary intended audience for this project includes:

- Policymakers - To provide evidence-based insights that can inform program development and resource allocation decisions
- Academic Institutions - To enable performance assessment and strategic planning through data-driven analytics
- Funding Agencies - To enhance portfolio analysis and impact measurement capabilities

Data Sources

The project utilizes two key datasets from the Author-ity 2009 collection:

- https://databank.illinois.edu/datasets/IDB-4370459

Dataset connecting researchers to records from NIH/NSF grants, USPTO patents, and academic publications

- https://databank.illinois.edu/datasets/IDB-9087546

Client And Perspective

There is no specific client for this project. The project is an independent effort to develop a comprehensive, interactive visualization platform that can serve the needs of multiple stakeholders in the academic ecosystem, including policymakers, academic institutions, and funding agencies. The goal is to enhance the understanding of the academic research and innovation landscape by providing data-driven insights.

Methods & Analysis

- Data Collection: The primary dataset were sourced from different datasets namely authorlink_nih.csv, authorlink_nsf.csv, authorlink_uspto.csv, uiuc_uspto.csv and genni-ethnea-authority2009.csv supplemented with NIH (National Institutes of Health) and NSF (National Science Foundation) grant data, USPTO patent linkage data, Inventor data (detailed patent records) and EthnicSeer demographic predictions dataset.
- Data Cleaning and Transformation: Joins were used to produce a unified table. This consolidated dataset enabled efficient analysis and visualization. Missing and null values were removed or transformed to maintain consistency. Outliers were handled to ensure reliability.

Dataset Details

1. authorlink_nih.tsv

This dataset contains information about NIH grants and their linkage to authors in the Author-ity 2009 database. The detailed description of all fields in this dataset is shown in the following figure:

Field Name	Description	Data Type
`app_id`	Application ID for the grant.	Number (#)
`nih_full_proj_nbr`	Full project number of the grant.	Text (Abc)
`nih_subproj_nbr`	Sub-project number, if applicable (null for main projects).	Number (#)
`fiscal_year`	The fiscal year of the grant.	Number (#)
`pi_position`	Position of the principal investigator (PI).	Number (#)
`nih_pi_names`	Full names of the principal investigator(s).	Text (Abc)
`org_name`	Name of the organization receiving the grant.	Text (Abc)
`org_city_name`	City of the organization.	Text (Geographic)
`org_bodypolitic_code`	State or region of the organization.	Text (Abc)
`age`	Number of years since the investigator's first paper was published.	Number (#)
`prob`	Probability that the author matches the Author-ity 2009 database (`> 0.5` threshold).	Number (#)
`au_id`	Unique identifier for the author in the Author-ity 2009	Text (Abc)

2. authorlink_nsf.tsv

This dataset contains information about NSF grants and their linkage to authors in the Author-ity 2009 database. The detailed description of all fields in this dataset is shown in the following figure:

Field Name	Description	Data Type
`AwardId`	Unique award ID for the NSF grant.	Number (#)
`fiscal_year`	Fiscal year range of the grant (e.g., `1986–1986`).	Text (Abc)
`pi_position`	Position of the principal investigator (PI).	Number (#)
`PrincipalInvestigators`	Names of the principal investigator(s) (can include multiple names separated by `;`).	Text (Abc)
`Institution`	Name of the institution receiving the grant.	Text (Abc)
`InstitutionCity`	City where the institution is located.	Text (Abc)
`InstitutionState`	State where the institution is located.	Text (Abc)
`age`	Number of years since the investigator's first paper was published.	Number (#)
`prob`	Probability that the author matches the Author-ity 2009 database (`> 0.5` threshold).	Number (#)
`au_id`	Unique identifier for the author in the Author-ity 2009 database.	Text (Abc)

3. authorlink_uspto.tsv

This dataset links authors in the Author-ity 2009 database to inventors in the USPTO (United States Patent and Trademark Office). The detailed description of all fields in this dataset is shown in the following figure:

Field Name	Description	Data Type
`au_id`	Unique identifier for the author in the Author-ity 2009 database.	Text (Abc)
`inv_id`	Unique identifier for the inventor in the USPTO database.	Number (#)
`prob`	Probability that the author matches the inventor in the USPTO database (`> 0.5` threshold).	Number (#)

4. uiuc_uspto.tsv

This dataset contains information about disambiguated inventors in the USPTO database. The detailed description of all fields in this dataset is shown in the following figure:

Field Name	Description	Data Type
`inv_id`	Unique identifier for the inventor in the USPTO database.	Number (#)
`is_lower`	Binary flag indicating whether the inventor's name is lowercase in the database.	Number (#)
`is_upper`	Binary flag indicating whether the inventor's name is uppercase in the database.	Number (#)
`fullnames`	Full name of the inventor.	Text (Abc)
`patents`	List of patents associated with the inventor (separated by `	`).
`first_app_yr`	Year of the inventor's first patent application.	Number (#)
`last_app_yr`	Year of the inventor's last patent application.	Number (#)

5. genni-ethnea-authority2009.tsv

This dataset provides demographic information about authors, including ethnicity and gender predictions. The detailed description of all fields in this dataset is shown in the following figure:

Field Name	Description	Data Type
`auid`	Unique identifier for authors in the Author-ity 2009 database.	Text (Abc)
`name`	Full name of the author, used as input for ethnicity/gender predictions.	Text (Abc)
`EthnicSeer`	Predicted ethnicity from the EthnicSeer tool.	Text (Abc)
`prop`	Confidence score of the EthnicSeer prediction.	Number (#)
`lastname`	Last name of the author (used as input for Ethnea+Genni).	Text (Abc)
`firstname`	First name of the author (used as input for Ethnea+Genni).	Text (Abc)
`Ethnea`	Predicted ethnicity from Ethnea+Genni (detailed, e.g., `ENGLISH`, `SLAV-ENGLISH`).	Text (Abc)
`Genni`	Predicted gender (`M` for male, `F` for female).	Text (Abc)
`SexMac`	Predicted gender using a third-party tool (`female`, `male`, `mostly_female`, etc.).	Text (Abc)
`SSNgender`	Predicted gender based on US Social Security Name data (`F`, `M`, or `-`).	Text (Abc)

Pre-processing:

To prepare the data for analysis, a structured approach was taken to clean and integrate multiple datasets seamlessly:

Step 1: Import Data. Loaded the four files (authorlink_nih.tsv, authorlink_nsf.tsv, authorlink_uspto.tsv, uiuc_uspto.tsv) into Tableau Prep.

Step 2: Perform Joins

1. NIH and NSF Tables

- Join Type: Outer Join to ensure all grant-related records are included.
- Join Condition: au_id (NIH) = au_id (NSF).
- Result: A combined table containing grant information from both NIH and NSF.

2. Add USPTO Linking Data

- Join Type: Inner Join to include only authors linked to inventors.
- Join Condition: au_id (from the previous join) = au_id (USPTO linking dataset).
- Result: Merges grant and patent data.

3. Merge with Inventor Data

- Join Type: Left Join to retain all grant and patent data while adding inventor details.
- Join Condition: inv_id (USPTO linking dataset) = inv_id (uiuc_uspto.tsv).
- Result: Adds detailed patent information to the dataset.

4. Add Demographics from EthnicSeer

- Join Type: Inner Join to retain only records with demographic predictions.
- Join Condition: au_id (from the combined dataset) = auid (EthnicSeer).
- Result: Final unified dataset with grant, patent, and demographic information.

Step 3: Clean Data

- Filter Records:
  - Removed rows where prob (match probability) < 0.5.
- Standardize Field Names:
  - Renamed all fields for consistency (e.g., nih_full_proj_nbr → Project Number).
- Aggregate Fields:
  - Combine fields like patents into (e.g., count(Number of Patents)).
- Export Final Dataset:
  - Saved the cleaned dataset as a .csv file for use in Tableau.
- Filtered and cleaned data in Tableau using calculated fields to replace nulls with placeholders like 'Name Not Available'.
- Created calculated groups for Funding Agency to categorize records based on field conditions.

Analysis Overview:

Key Findings:

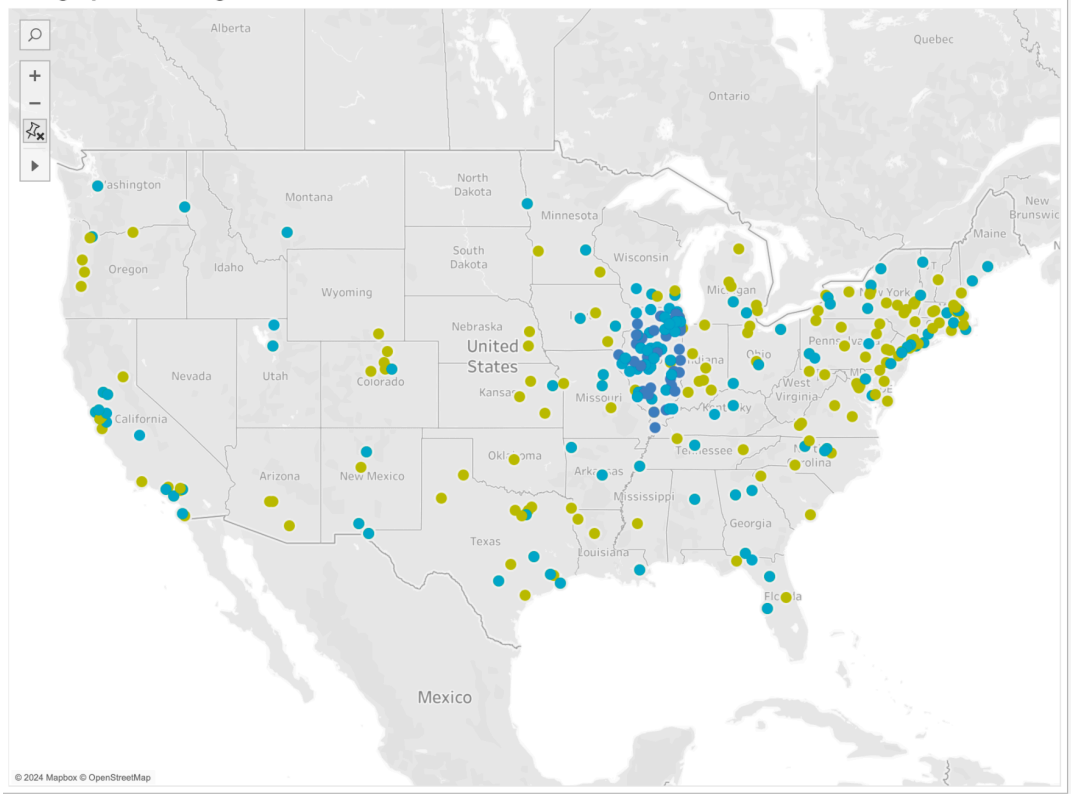
- Age and Funding Connections:
    - Award distributions varied across different age groups, with the 20-25 and 25-30 age groups consistently receiving the highest number of awards over the years.
    - The award success probability also differed across age groups, with the 'Early-Career' and 'Mid-Career' stages having the highest probabilities.
- Geographic Trends:
    - Award achievements were concentrated in a few key cities, with New York, Cambridge, and La Jolla leading the top 50 cities.
    - The institution distribution by state showed funding was primarily focused in a few states, with California, New York, and Massachusetts being the top recipients.
- Institutional Performance:
    - The institution rankings based on awards revealed a clear hierarchy, with top universities like the University of Wisconsin, University of California, and University of Michigan consistently receiving the most awards.

Challenges Encountered:

- Data Quality Issues:
    - The dataset contained null values and inconsistencies that required careful data cleaning and preparation.
    - Solution: Used Tableau Prep to identify and address data quality issues, such as replacing null values and harmonizing data types.
- Visualization Complexity:
    - The large number of data points and multiple dimensions made creating effective visualizations a challenge.
    - Solution: Leveraged Tableau's features like calculated fields, filters, and grouping to simplify the data and enhance the storytelling.

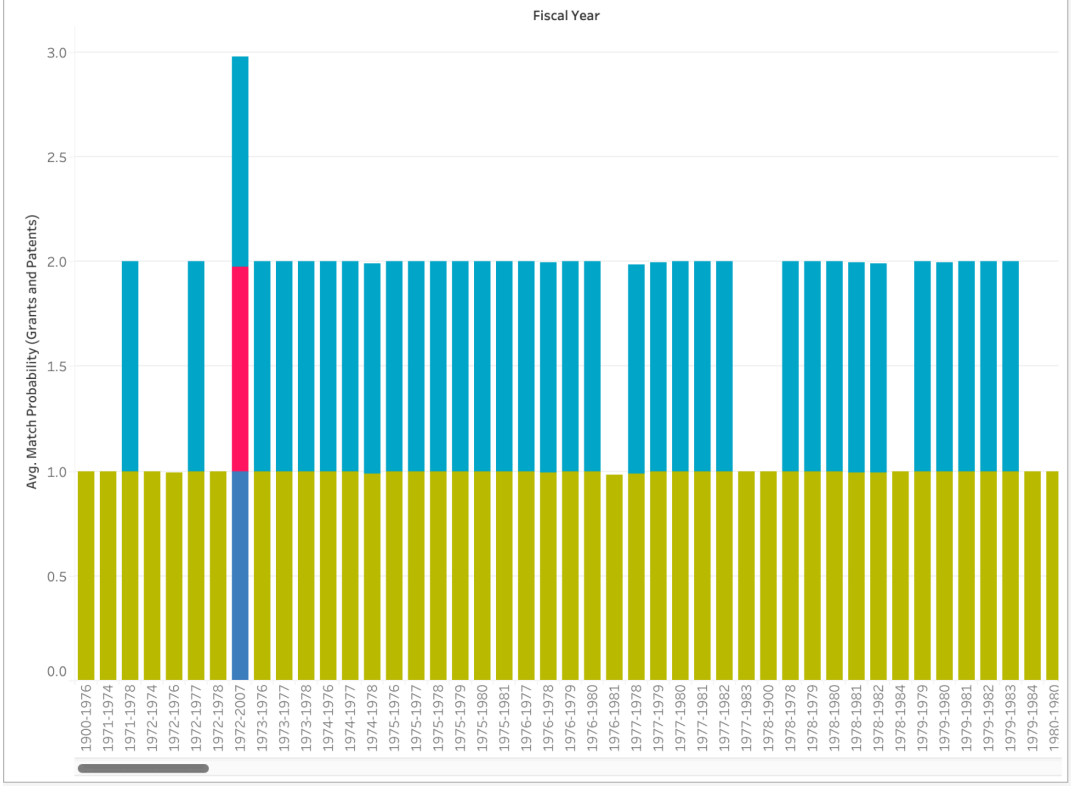
Final Outcome:

Geographic Funding Trends



Funding Agency
- NIH Award
- NSF Award
- Unknown Funding A..
- USPTO

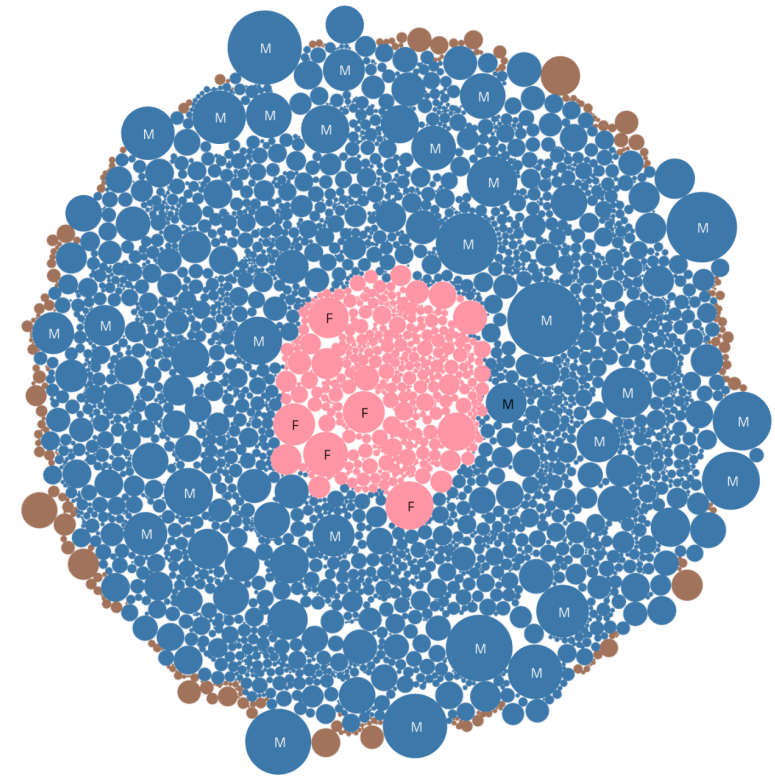
Grant and Patent Probability Analysis



Funding Agency

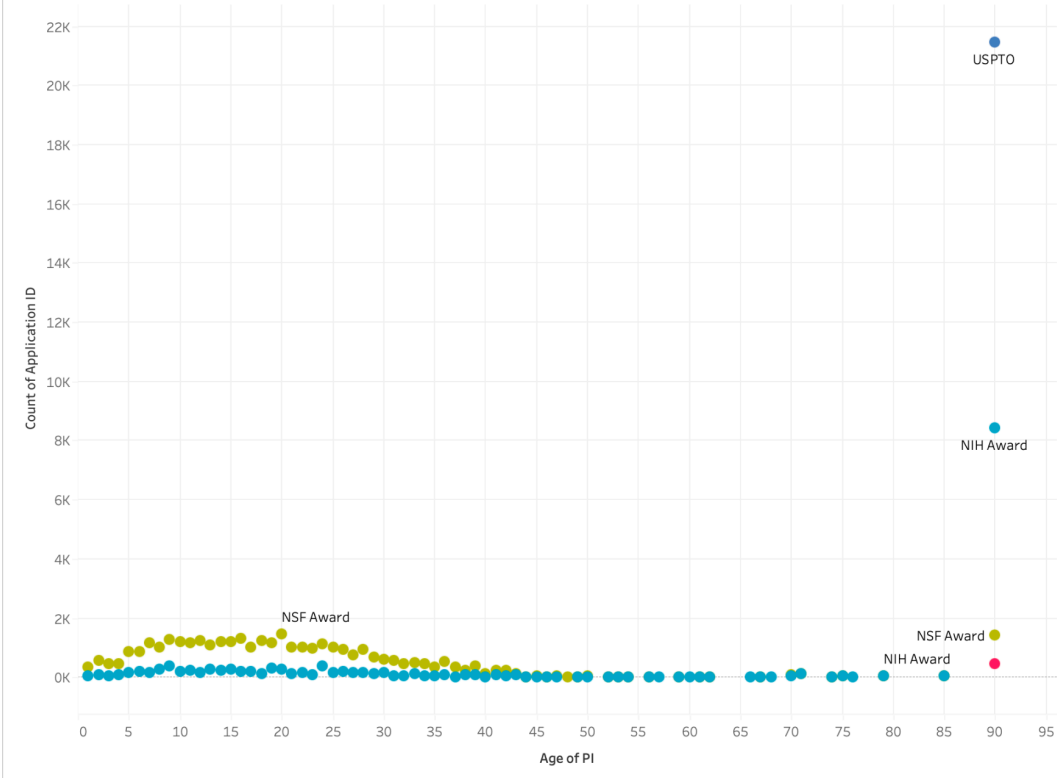
Funding Agency
- NIH Award
- NSF Award
- Unknown Funding A..
- USPTO

Gender vs. Patent Success Rate



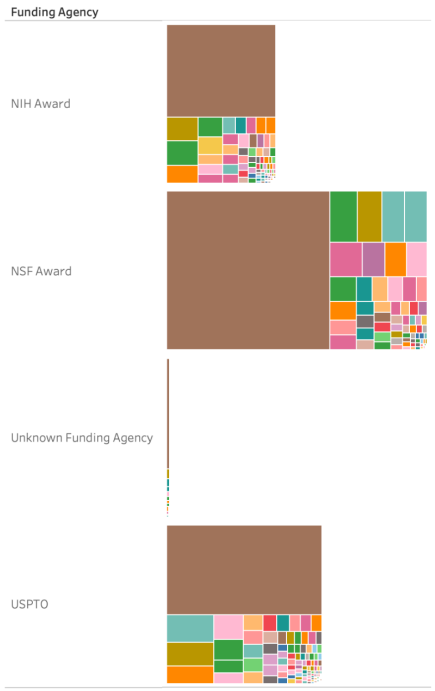
Gender
- F
- M
- Others

Patent and Grant Connections by Age

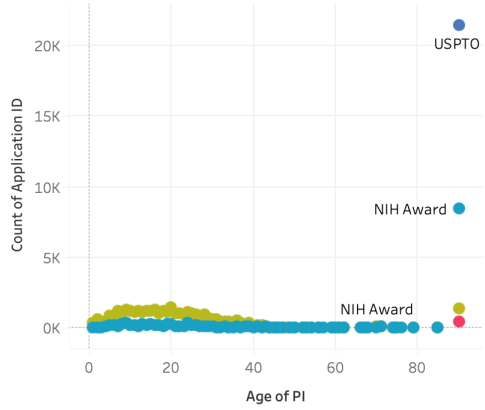


Funding Agency
- NIH Award
- NSF Award
- Unknown Funding A..
- USPTO

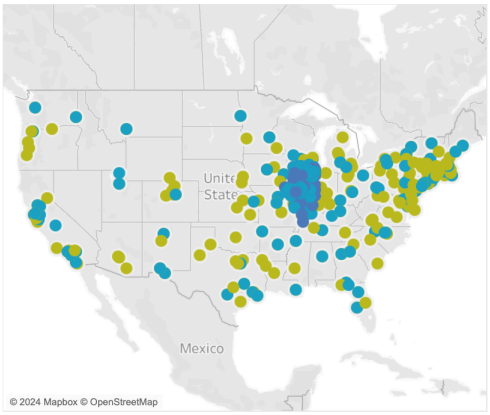
Ethnic Distribution by Funding Agency



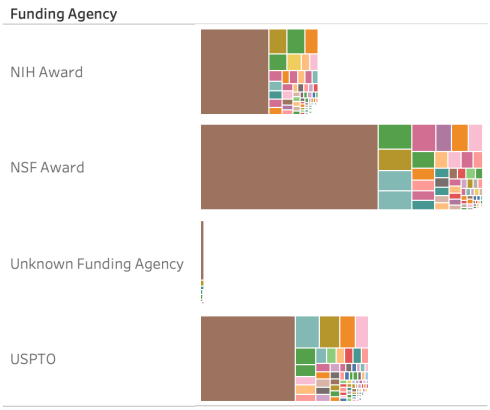
Patent and Grant Connections by Age



Geographic Funding Trends



Ethnic Distribution by Funding Agency



Gender vs. Patent Success Rate