US Innovation: The SBIR-NIH-USPTO Axis

Caden Stewart
Halıcıoğlu Data Science Institute
A69031557

Guruprasad Parasnis
Halıcıoğlu Data Science Institute
A69035356

Tamar Schaap
Halıcıoğlu Data Science Institute
A69031567

Background

"Innovation is the multi-stage process whereby organizations transform ideas into new/improved products, service or processes, in order to advance, compete and differentiate themselves successfully in their marketplace"[1] Economists estimate that 50% of Gross Domestic Product (GDP) can be attributed to innovation, which makes it essential for nations to support and cultivate it. [2] A good proxy variable for innovation is patents, as each one represents a novel idea or invention. One way the US seeks to promote innovation is through grants for groups to develop and test ideas. Two key vehicles for this are Small Business Innovation Research (SBIR) grants, and National Institute of Health (NIH) grants.

Challenge

It is in the interest of a nation (in this case the United States) to improve, progress, and grow the maximal amount with minimal investment of current resources. The combination of this interest and the available data leads us to our proposed challenge to solve:

How can we improve the disbursement of SBIR and NIH grants to maximize patent approvals and national innovation levels?

Data Sources

NIH Funding Data

The National Institutes of Health (NIH) maintains a vast repository of biomedical and health-related research funding data. Digitally available since 1985, this dataset includes millions of funded projects. Important fields in NIH records comprise grant ID, funding amount, project abstracts, and principal investigator (PI) details. Researchers use this data to analyze funding trends and scientific advancements in the health sector. The primary source for this data is [NIH RePORTER](https://reporter.nih.gov).

SBIR Award Data

The Small Business Innovation Research (SBIR) program provides funding for small businesses engaged in research and development with commercialization potential. Since its inception in 1982, it has awarded thousands of grants annually. Key data fields from SBIR include award amount, project title, agency, and company details, offering valuable insights into innovation trends and funding distribution. This data is publicly accessible through [SBIR.gov](https://www.sbir.gov).

Patent Data

With more than 13 million patent applications and more than 1 million Patent Cooperation Treaty (PCT) applications, the United States Patent and Trademark Office (USPTO) maintains vast databases on intellectual property. Modern records are well-documented from 1976 onward, covering data as far back as 1790. Patent numbers, assignees, claims, and filing or grant dates are important data fields. This dataset is essential for tracking innovations and conducting research on intellectual property. This information is officially available from [USPTO.gov](https://www.uspto.gov).

Methodology

Preparation of USPTO Data

We made a script that performs XML parsing on patent files. Here's a detailed breakdown of its functionality:

1. **Extracting Patent Data**: The script begins by unzipping all the patent files stored in a specified directory (`2023 patent files`). It uses the `zipfile` and `os` libraries to iterate through all ZIP files in a folder and extract their contents into another directory (`unzipped_files`). It ensures that the extraction directory exists before proceeding.
2. **Processing XML Files**: After extracting the patent files, the script processes the XML files containing patent data. These XML files include metadata about patents, such as application numbers, titles, inventors, abstracts, claims, and classifications.
3. **Parsing XML Content**: The script utilizes Python's built-in `xml.etree.ElementTree` to read and extract specific details from each XML file. It loops through the files, loads their contents, and retrieves structured information.
4. **Data Extraction and Structuring**: The extracted patent data is stored in a structured Pandas DataFrame. The script filters relevant fields like the patent title, abstract, filing date, and assignee.
5. **Output**: After parsing the XML files, the script saves the data into a structured CSV file.

PostgreSQL Data Model and Schema Overview

Award Data Table

The `award_data` table is designed to store information on SBIR awards. Key aspects include:

- **Company and Award Details:** Fields such as `company_name`, `Award_Title`, and `Agency` capture the identity of the awardee and the nature of the award.
- **Tracking and Financial Information:** Fields like `Agency_Tracking_Number`, `Contract`, `Award_Year`, and `Award_Amount` store tracking identifiers and monetary details.
- **Date Fields:** Although stored as VARCHAR in the current design, fields such as `Proposal_Award_Date`, `Contract_End_Date`, etc., capture key time-related information for subsequent analysis.
- **Ownership and Demographic Indicators:** Boolean flags (e.g., `HUBZone_Owned`, `Socially_and_Economically_Disadvantaged`, `Women_Owned`) alongside `Number_Employees` provide insights into company characteristics.
- **Contact and Address Details:** The schema also includes fields for addresses, contact names, emails, and phone numbers to facilitate follow-up and deeper analysis.

NIH Project Data Table

The `nih_project_data` table is designed to store information on NIH-funded projects. Key aspects include:

- **Project and Funding Details:** Fields such as `Project_Title`, `Abstract_Text`, and `Funding_ICs` capture project objectives and the funding institutions involved.
- **Tracking and Financial Information:** The table includes `Project_Number`, `Application_ID`, `Total_Cost`, and `Direct_Cost` fields to track projects and their funding allocations.
- **Date Fields:** Fields such as `Project_Start_Date` and `Project_End_Date` provide information on the project's timeline.
- **Principal Investigator (PI) Details:** Stores the name, email, and institution of the principal investigator(s) (`PI_Name`, `PI_Email`, `PI_Institution`) to support analysis of researcher activity.
- **Organization and Location Information:** Captures details such as `Organization_Name`, `City`, `State`, and `Country` to analyze geographic distribution of funded research.

Patent Data Table

The `patent_data` table is designed to store information related to granted patents. Key aspects include:

- **Patent Identification and Metadata:** Fields such as `Patent_Number`, `Title`, `Abstract`, and `Patent_Type` capture core patent details.
- **Filing and Grant Dates:** Includes `Filing_Date` and `Grant_Date` to track the patent lifecycle.
- **Inventor and Assignee Information:** Stores names and affiliations of inventors (`Inventor_Name`, `Inventor_City`, `Inventor_Country`) and assignees (`Assignee_Name`, `Assignee_City`, `Assignee_Country`).

Neo4j Data Model

Patent Data Graph

We filtered the data fields in the USPTO data further, as the focus of our Neo4j analysis was to create connections with the SBIR data instead of an in depth analysis of the patent data itself. The final structure of the USPTO graph was as follows:

Nodes - NodeName(symbol) {attributes}:

- Patent(p) {id: publication_doc_number, title: invention_title}
- Organization(o) {name: applicant_organization}
- PatentKind(k) {type: publication_kind}
- Inventor(i) {name: inventor_name}

Relationships - (startNode) - [RELATION] -> (endNode):

- (o) - [FILED] -> (p)
- (p) - [HAS_KIND] -> (k)
- (i) - [INVENTED] -> (p)

SBIR Data Graph + Patent Integration

After running into issues uploading the large USPTO dataset, we decided to limit our integration of SBIR data to only those awards with a PI that is also listed as an inventor somewhere in the patent data.

Nodes:

- SBIR_Award(a) {id:tracking_number}
- Organization(o) {name: company} *merged
- Inventor(i) {name: pi_name} *merged

Relationships:

- (o) - [RECEIVED_AWARD] -> (a)
- (i) - [PI_FOR] -> (a)

Python

To detect connections between patents and SBIR awards based on text descriptions, we mirrored the methodology presented in class to vectorize the patent titles and SBIR abstracts. This was done in a Jupyter notebook using the Python programming language. See our GitHub for detailed code.

Analysis steps:

1. Text lowering, special character removal, and lemmatization of both patent titles and SBIR abstracts.
   (a) Lemmatization involves converting words with very similar meanings into one common expression. This was done using the *spacy* package.
2. Vectorization of text using Bidirectional Encoder Representations from Transformers (BERT) model all-MiniLM-L6-v2. This transformer produces a vector of 384 embeddings for each text entry.
3. Vector normalization (conversion to unit vector) and calculation of cosine similarity using *torch* library.
4. Similarity score plotting (done with *seaborn* library), maximum value extraction and original data linkage, and thresholding.

Results

Counting Companies in Each Data Source

- **SBIR Awards:** The query lists the top 10 companies by the number of awards received, providing insight into the most active awardees.

	company	count
1	Physical Optics Corporation	1974
2	PHYSICAL SCIENCES INC.	1639
3	CREARE LLC	1332
4	Intelligent Automation, Inc.	1203
5	LYNNTECH INC.	1084
6	LUNA INNOVATIONS INCORPORATED	1026
7	FOSTER-MILLER, INC.	962
8	RADIATION MONITORING DEVICES, INC.	942
9	TDA RESEARCH, INC.	933
10	CHARLES RIVER ANALYTICS, INC.	921

- **Patents:** A similar aggregation identifies the top 10 companies by patent filings, serving as a proxy for innovation capacity.

	company	count
1	<null>	67236
2	Samsung Electronics Co., Ltd.]	7728
3	QUALCOMM Incorporated]	7002
4	Apple Inc.]	6067
5	SAMSUNG ELECTRONICS CO., LTD.]	5512
6	CANON KABUSHIKI KAISHA]	5177
7	International Business Machines Corporation]	4306
8	LG ELECTRONICS INC.]	4055
9	Intel Corporation]	3789
10	Micron Technology, Inc.]	3775

- **NIH Projects:** The top 10 organizations are ranked by the number of NIH projects, reflecting the prominence of institutions in NIH-funded research.

	company	count
1	JOHNS HOPKINS UNIVERSITY	3564
2	UNIVERSITY OF CALIFORNIA, SAN FRANCISCO	3544
3	UNIVERSITY OF MICHIGAN AT ANN ARBOR	3350
4	UNIVERSITY OF PENNSYLVANIA	3222
5	WASHINGTON UNIVERSITY	2938
6	UNIVERSITY OF PITTSBURGH AT PITTSBURGH	2886
7	STANFORD UNIVERSITY	2832
8	YALE UNIVERSITY	2692
9	COLUMBIA UNIVERSITY HEALTH SCIENCES	2584
10	MASSACHUSETTS GENERAL HOSPITAL	2504

Companies Active in Multiple Innovation Arenas

This analysis:

- Consolidates company names from patents, SBIR awards, and NIH projects.
- Counts the total number of occurrences for each company.

	company	total_records
1	samsung electronics co., ltd.]	13243
2	qualcomm incorporated]	7330
3	apple inc.]	6402
4	international business machines corporation]	6057
5	lg electronics inc.]	5756
6	canon kabushiki kaisha]	5314
7	samsung display co., ltd.]	5155
8	huawei technologies co., ltd.]	5049
9	micron technology, inc.]	4202
10	google llc]	4152

Innovation Footprint by Company Across All Sources

A comprehensive footprint is generated by:

- Aggregating counts of patents, SBIR awards, and NIH projects per company.
- Merging these counts using full outer joins.
- Ranking companies based on the total number of innovation events.

Insight: This analysis helps count the SBIR, NIH awards and the filed patents per company to analyze whether there are companies that are performing well in all fronts

	company	total_patents	total_awards	total_nih_projects
1	samsung electronics co., ltd.]	13243	0	0
2	qualcomm incorporated]	7330	0	0
3	apple inc.]	6402	0	0
4	international business machines corporation]	6057	0	0
5	lg electronics inc.]	5756	0	0
6	canon kabushiki kaisha]	5314	0	0
7	samsung display co., ltd.]	5155	0	0
8	huawei technologies co., ltd.]	5049	0	0
9	micron technology, inc.]	4202	0	0
10	google llc]	4152	0	0

Analyzing NIH Project Duration vs. Funding

This query examines the relationship between project duration and funding by:

- Calculating the difference between project start and end dates.
- Comparing the project duration with the total cost.

Insight: This analysis helps identify trends, such as whether longer projects tend to receive higher funding, which can inform funding strategies and project management practices.

	full_project_num	project_title	project_start	project_end	total_cost	project_duration
1	5T32MH013043-52	Research Training Program in Psychiatric Epid…	1972-07-01	2027-06-30	423907	
2	5T32MH013043-52	Research Training Program in Psychiatric Epid…	1972-07-01	2027-06-30	423907	
3	2T32EY007001-46A1	Research Training in Visual Sciences	1975-07-01	2028-07-31	229707	
4	2T32EY007001-46A1	Research Training in Visual Sciences	1975-07-01	2028-07-31	229707	
5	2T32HL007028-46	Nutrition, Obesity and Atherosclerosis Traini…	1975-07-01	2028-06-30	353205	
6	2T32HL007028-46	Nutrition, Obesity and Atherosclerosis Traini…	1975-07-01	2028-06-30	353205	
7	2T32DK007056-47A1	Training Grant in Academic Gastroenterology	1975-07-01	2028-06-30	345108	
8	2T32DK007056-47A1	Training Grant in Academic Gastroenterology	1975-07-01	2028-06-30	345108	
9	2T32HD007009-48	Ruth L. Kirschstein National Research Service…	1975-07-01	2028-04-30	581246	
10	2T32HD007009-48	Ruth L. Kirschstein National Research Service…	1975-07-01	2028-04-30	581246	

Correlating Records Across Datasets (Companies)

This analysis consolidates records from patents, SBIR awards, and NIH projects to:

- Identify companies that appear in more than one dataset.
- Generate a JSON aggregation of the records associated with each company.

Insight: This approach reveals companies with multi-domain activity, highlighting their strategic emphasis on innovation across various channels.

	company	record_count	records
1	samsung electronics co., ltd.]	13243	[{"source" : "patents", "record_id" : "712310", "title" : "Wa
2	qualcomm incorporated]	7330	[{"source" : "patents", "record_id" : "713589", "title" : "St
3	apple inc.]	6402	[{"source" : "patents", "record_id" : "712880", "title" : "Ro
4	international business machines corporation]	6057	[{"source" : "patents", "record_id" : "713131", "title" : "Im
5	lg electronics inc.]	5756	[{"source" : "patents", "record_id" : "712314", "title" : "Dr
6	canon kabushiki kaisha]	5314	[{"source" : "patents", "record_id" : "712003", "title" : "Aq
7	samsung display co., ltd.]	5155	[{"source" : "patents", "record_id" : "712002", "title" : "In
8	huawei technologies co., ltd.]	5049	[{"source" : "patents", "record_id" : "712637", "title" : "Vi
9	micron technology, inc.]	4202	[{"source" : "patents", "record_id" : "719694", "title" : "Bu
10	google llc]	4152	[{"source" : "patents", "record_id" : "713238", "title" : "Op

Correlating Records Across Datasets (People)

Two variations of this analysis are provided:

- **Simple Aggregation:** Combines names of inventors, contacts, and PIs from the three datasets and aggregates their records.
- **Handling Multiple Names:** Utilizes functions to split comma-separated names into individual entries, ensuring accurate counting.

Insight: These queries help identify individuals active across multiple innovation platforms, highlighting key cross-domain influencers, which can be a major factor in influencing the decision we plan to interpret in this project.

	person_name	dataset_count	records
1	Haytham Elhawary	2	[{"source" : "award_data", "record_id" : "2030327"}, {"source
2	Wei Sun	2	[{"source" : "award_data", "record_id" : "DOT-23-FH2-005"}, {
3	Maximilian Liese	2	[{"source" : "award_data", "record_id" : "FX224-OCSO1-0183"},
4	Joshua Martin	2	[{"source" : "award_data", "record_id" : "N231-063-0537"}, {"
5	Joseph Batta-Mpouma	2	[{"source" : "award_data", "record_id" : "2022-01401"}, {"sou
6	John Campbell	2	[{"source" : "award_data", "record_id" : "F141-093-1759"}, {"
7	Roei Ganzarski	2	[{"source" : "award_data", "record_id" : "F2D-4881"}, {"sourc
8	Surya Moganty	2	[{"source" : "award_data", "record_id" : "1248692"}, {"source
9	David Levitt	2	[{"source" : "award_data", "record_id" : "104376"}, {"source"
10	David Rozema	2	[{"source" : "award_data", "record_id" : "2019-00992"}, {"sou

Word Frequency Analysis Across Datasets

A Common Table Expression (CTE) is used for the purpose of this query to:

- Combine text from invention titles, award titles, and project titles.
- Split these text fields into individual words and convert them to lowercase.
- Aggregate and rank the words by frequency.

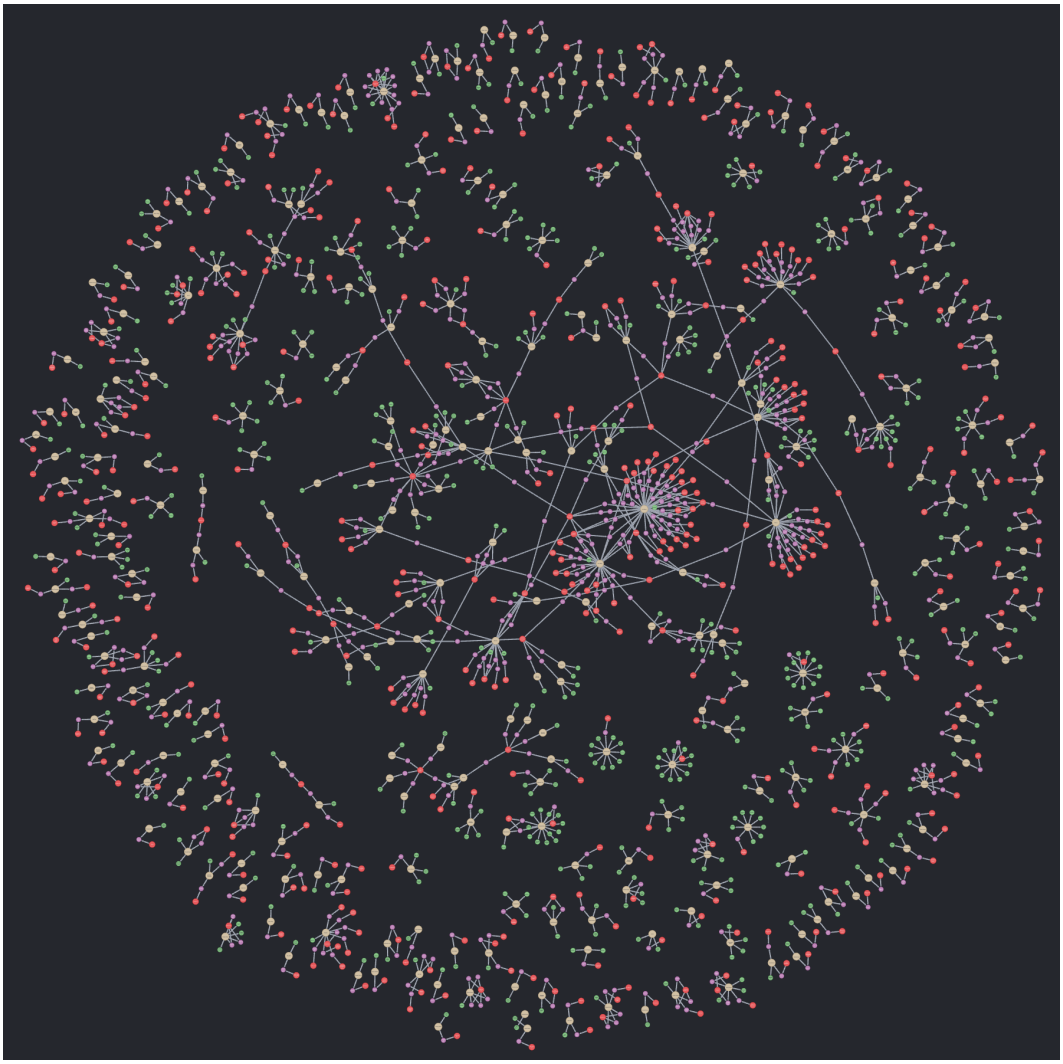
Insight: This analysis identifies recurring keywords and themes across the datasets, providing insights into potential words that can help identify any features that might be common across data sources associated with innovation and awards.

	word	frequency
1	and	487205
2	for	393440
3	of	287361
4	method	176873
5	a	172392
6	in	136822
7	system	131403
8	device	125362
9	the	97293
10	with	86266
11	to	75105
12	methods	68637
13	apparatus	68264

USPTO and SBIR Data Linkage

The merging of the USPTO and SBIR datasets was done to demonstrate the overlap and network behavior of these two distinct communities. The figure below shows just 2000 of the almost 700,000 we successfully integrated into our Neo4j database. As a reminder, we only uploaded about 1/4 of the possible patent records, and only those SBIR records with authors who had also invented patented inventions.

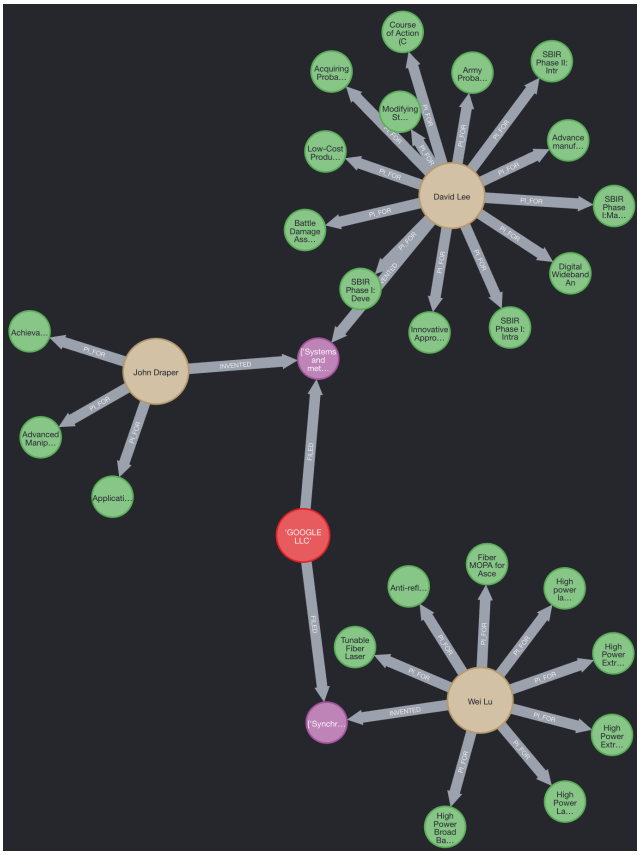
The node counts of this network were as follows:
Inventor: 430,105
Patent: 192,307
Organization: 49,230
SBIR_Award: 13,034
PatentKind: 6



Insight: This figure clearly demonstrates powerful connectivity between these communities, with a large central group and several smaller groups in the periphery.

Organization Networks

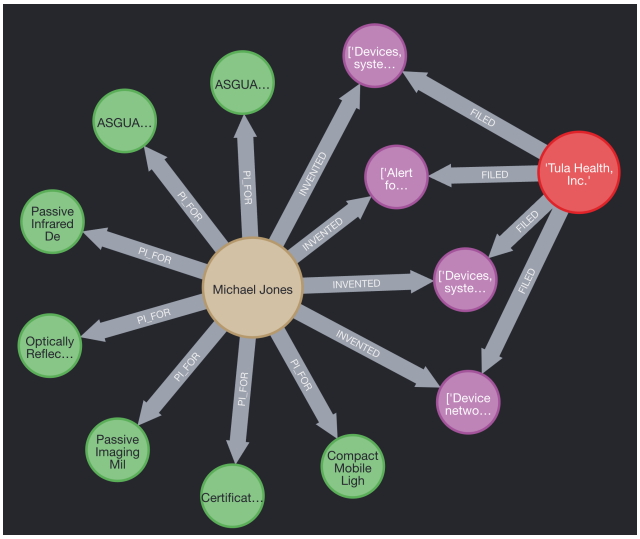
Organizations, especially large ones, tend to be important in connecting projects together. They may have employees that focus in several disparate fields, and would otherwise have no connection. our network effectively surfaces these connections and allows for analysis of organization networks.



Insight: Organization data labels provide important information on the connections between authors and projects, and some organizations may tend to be more productive in the use of federal funding.

Individual Networks

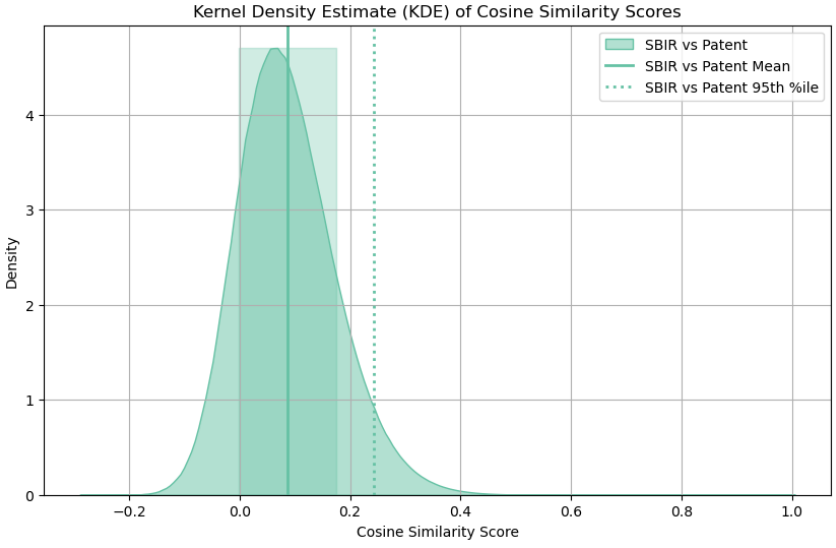
The most fundamental connection between patent records and SBIR awards in the inventor/PI name. In the case shown below, the name is common, so there may be an issue of multiple scientists with the same name. However, seeing that several of the patents also were filed by the same company, we can be confident that the author for those patents is truly the same.



Insight: The emergence of inventor/PI centered networks demonstrates the success of our data processing during the matching and merging process. These networks allow us to determine individual productivity.

Connection of Invention Titles and Award Abstracts

The figure below shows the distribution of 188,060,457 cosine similarity scores. These scores were calculated between patent titles (typically one line) and SBIR abstracts (typically 1-2 paragraphs). We would expect very few of the records to match well.



While the vast majority of these scores were low, there were several that crossed the 0.7 threshold, a commonly used threshold for determining if two texts are linked. The top matches are described below.

Top 3 Matches:

1. Cosine similarity: 0.77
   SBIR Abstract:
   LM Group Holdings Inc. (LMGH) partnering with Fabrisonic LLC is proposing a program to investigate manufacturing of amorphous metal alloy laminate composites and **cladding of metallic surfaces** by using **ultrasonic additive manufacturing** (UAM), a solid-state 3D metal printing technology.
   Patent:
   Applicant Organization: LM Group Holdings, Inc., Fabrisonics LLC
   Title: Ultrasonic additive manufacturing of cladded amorphous metal products
2. Cosine similarity: 0.77
   SBIR Abstract:
   nLight proposes to develop and commercialize an innovative kilo-watt class fiber laser pump combiner which is especially suited for counter-pumping architectures that is indispensable for various coherent combining techniques.... nLight will demonstrate kilowatt-**class amplifier** using this pump combiner and its **proprietary active fiber.**
   Patent:
   Applicant Organization: NLIGHT INC.
   Title: Cladless fiber for fiber laser pump and combiner
3. Cosine similarity: 0.74
   SBIR Abstract:
   Flexible silica aerogel composites, a class of super-insulation material recently developed by **Aspen Aerogels**, has not been utilized before in **high temperature** TPS designs. Thermo-physical characterization data will be collected... for **high-temperature** durable, oxidatively stable, flexible **aerogel composites** at different densities, pressures and temperatures.... The aerogels will be compatible with all **high temperature** capable face-skin materials.
   Patent:
   Applicant Organization: Aspen Aerogels
   Title: Aerogel compositions for high temperature applications

Insight: It works! Using vector embeddings and cosine similarity, we were able to successfully link SBIR awards to their corresponding patents.

Summary and Implications

The set of PostgreSQL queries and the underlying data models offer some unique insights into the databases we experimented on. Some of the key outcomes include:

- **Cross-Dataset Linkages:** The joins between awards, NIH projects, and patents reveal relationships where research funding, award-winning projects, and patentable inventions overlap.
- **Geographic and Temporal Trends:** Aggregations by country and year provide insight into where and when innovation is most active.
- **Activity by Company and Person:** Ranking companies and individuals helps identify leaders and key players in the innovation ecosystem.
- **Thematic Analysis:** Word frequency analysis sheds light on dominant themes and technological areas across the datasets.
- **Comprehensive Footprint:** Combining all data sources to assess a company's overall innovation output delivers a holistic view of their impact.

The Neo4j graph database we successfully constructed added another level of nuance and analysis capability to our work in PostgreSQL. Some key contributions:

- **Evidence of Network Existence:** SBIR awards and patents are clearly linked by authors and organizations, and should be analyzed together.
- **Localized Pattern Emergence:** The performance of individuals and organizations can be clearly visualized and evaluated.

Finally, the culmination of our analysis was a concept-based connection of patents and SBIR awards in Python, which was ultimately successful. This combined with author and organization matching provides us with a robust and automated way to follow the flow of funding and ideas.

Limitations

- **Computational resources:** all of the data was housed an analyzed on personal computing machines, which limited the speed at which we could perform processing and therfore the depth of analysis we were able to complete.
- **Data inconsistencies:** Since our datasets were collected by different groups and in different formats, many fields could not be matched perfectly. A good example is organization/company name. Samsung used over 15 different variations of its name when filing patents or submitting award applications, it is impossible to account for all possible variations of all possible companies.
- **Incongruent text comparison:** While our text analysis and comparison ultimately produced successful cases, an abstract is not equivalent to a title, so it is possible that our method would produce a significant amount of false negatives, where the title simply did not contain enough information to link it to a more detailed abstract.

Future Directions/Extensions

There is much than can be done to build on the work we have done here. For example, more detailed text/string processing could be done to ensure names and organizations are matched effectively. The most obvious and likely the most fruitful next step would be to use the calculated similarity scores as a node relationship in the Neo4j database, and extract communities that clearly turned SBIR awards into patents or vice versa. We could then extract detailed data on these communities (how many coauthors? what region are they from? what fields are most productive?) and use this to advise funding decisions.

References

[1]Baregheh, A., Rowley, J., & Sambrook, S. (2009). Towards a multidisciplinary definition of innovation. *Management Decision*, *47*(8), 1323–1339. https://doi.org/10.1108/00251740910984578

[2]McKinney, P. (2023, January 1). Innovation Fuels 50% of GDP Growth per Economists. *The Innovators Network*. https://theinnovators.network/innovation-fuels-50-of-gdp-growth-per-economists