JourNetwork

Find the Journal that's Right for You!

Team Members

Brian Philip

Michael Paskett

Taylor Hansen

About the Designers:

Brian Philip
U1138890
brian.philip@utah.edu

Brian Philip, a Biomedical Engineering Ph.D. student studies the way the brain processes information and how it fails for diseases such as Epilepsy. However, this task involves combing through hundred of hours of seismograph-like voltage plots. He took the course to find out if there was a better way?

Michael Paskett
U114400
michael.paskett@utah.edu

Michael Paskett, a second year Ph.D. student faced with the arduous challenge of developing next generation neuroprosthetics, realized he was unequipped with the tools needed to see trends in data. Well known for saying, “what good is data if you can’t present it effectively?” he took the course to present beautiful visuals that tell a story.

Taylor Hansen
U0642850
taylor.c.hansen@utah.edu

Taylor Hansen, a man of few words but when he talks others listen. He aims to merge that philosophy with designs to decode how neurons in the arm communicate - all for better neuroprosthetics. Like the other team members, he is a second year Biomedical Engineering Ph.D. student.

https://github.com/mpaskett/dataviscourse-pr-journetwork

Project Process Book
CS 6630
Data Visualization

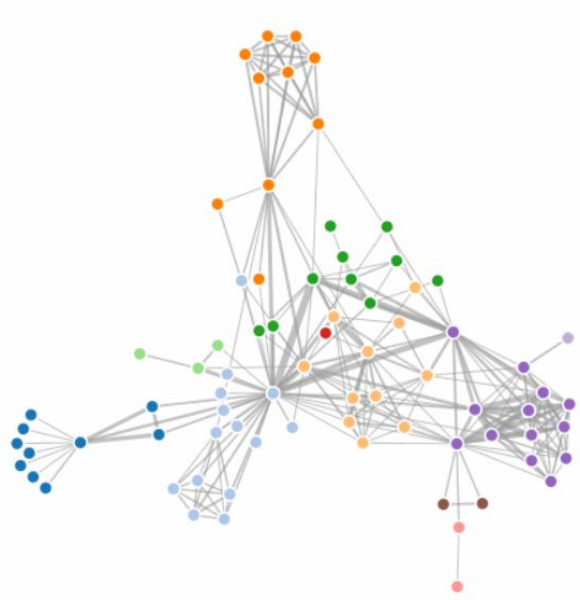
Table of Contents

Overview and Motivation:

PUBLISH! The word no scientist will ever forget. However, the best tool for finding related journals to publish in is the old-fashioned Google search. Here, we provide a visualization approach to quickly assess journals in a specific field, their connections, and their goals. With this, we hope scientists spend less time searching for journals and more time impacting the future of science.

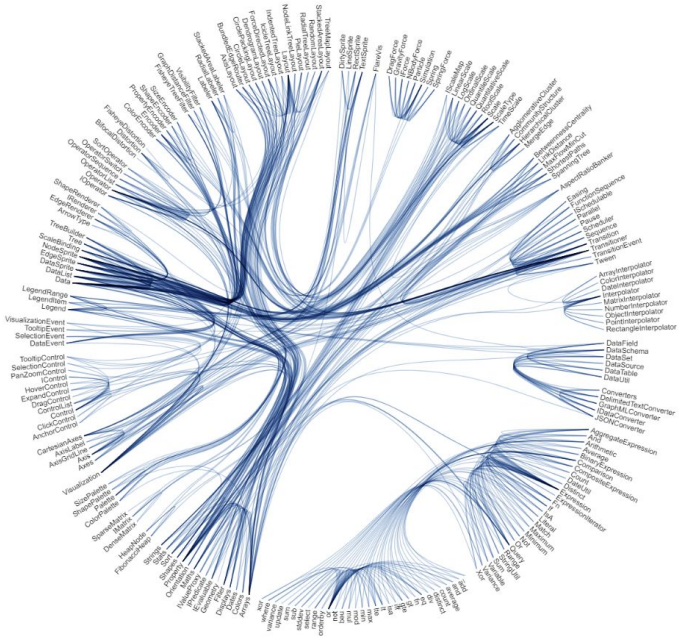
Related Work:

During our PhDs, each of us were tasked with a literature review of our fields. After going on our own, only to find fringe articles in our respective fields, we needed to change our tactics. Instead, oral communication with our PI's, our fellow lab members, and with other instructors enabled us to find parse through the sea of journals and papers. We realized there was a real need for a visualization that explains not just the connections between articles but how related the subspecialties are.

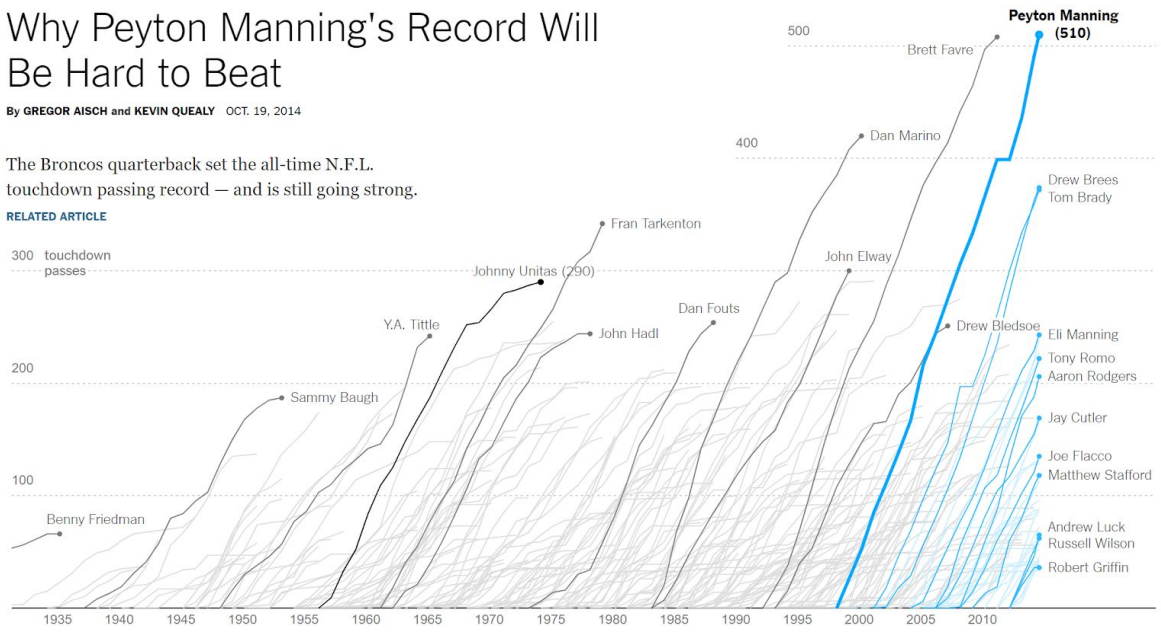


Speaking of relation, when we went over force-directed networks, we thought this was a brilliant solution to highlight subspecialties in our dataset. Let the data organize the network for you based on the push and pull factors. An example of one such network is shown at left.

Another way this same information could be depicted is with a chord diagram, where each journal is a node around the circle and the connections are established by the citations between groups. An example of one such chord map developed by Mike Bostock using the Flare toolkit is shown below.



Earlier in the semester, we were introduced to the New York Times [article](), "Why Peyton Manning's Record Will Be Hard to Beat". What would normally be a line plot of footballs caught per quarterback that could be summarized in a single statement was enhanced with interactivity. Line plots, highlights, opacity, and text boxes dynamically updating allowed one to understand differences in players growth, differences between the greats, and what to expect in the future. It was simple, elegant, and invited you to explore.



Questions:

We looked at this project as giving key insights into three themes: journals, interconnectivity, and what's changing. This led us to the following set of questions:

1. What journal should I submit to?
2. Which journals have relevant information?
3. How are these journals growing?
4. Are there subspecialties of journals related to my field?

Data:

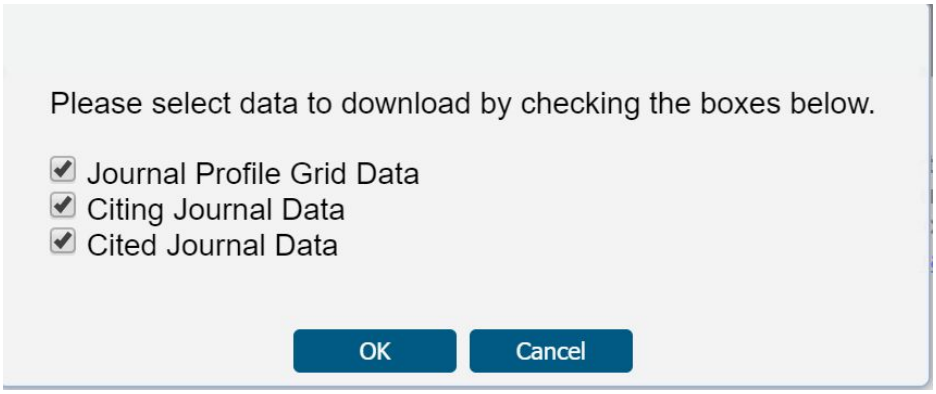
After wandering through the Marriott Library research database for another class, Brian happened to come across InCites Journal Citation Reports. This turned out to be a huge blessing as it contained three sets of .csv datasets that could be easily processed in Python, then visualized with D3. One of our homework assignments rapidly read in .csv files for data, which helped confirm for us that this would be possible by the end of the semester.

"*InCites*™ is a customized, citation-based research evaluation tool that enables one to analyze institutional productivity and benchmark output against peers. It is produced by Thomas Reuters and uses bibliographic record and citation data generated from the *Web of Science* and *Journal Citation Reports*." The University of Utah has an active subscription to *InCites*™, facilitating our use of this resource for our project. The bibliometric indicators with *InCites*™ are divided into various categories.

- Journal Citation Reports Indicators (e.g. Journal Impact Factor, Cited Half-Life, Article Influence®)
- Impact Indicators (e.g. Citation Impact, H-Index)

- Percentile and Percentage Indicators (e.g. Average Percentile, % Documents Cited)
- Collaboration Indicators (e.g. International Collaborators, % of Industry Collaborations)
- ESI Indicators (e.g. Highly Cited Papers, Hot Papers)

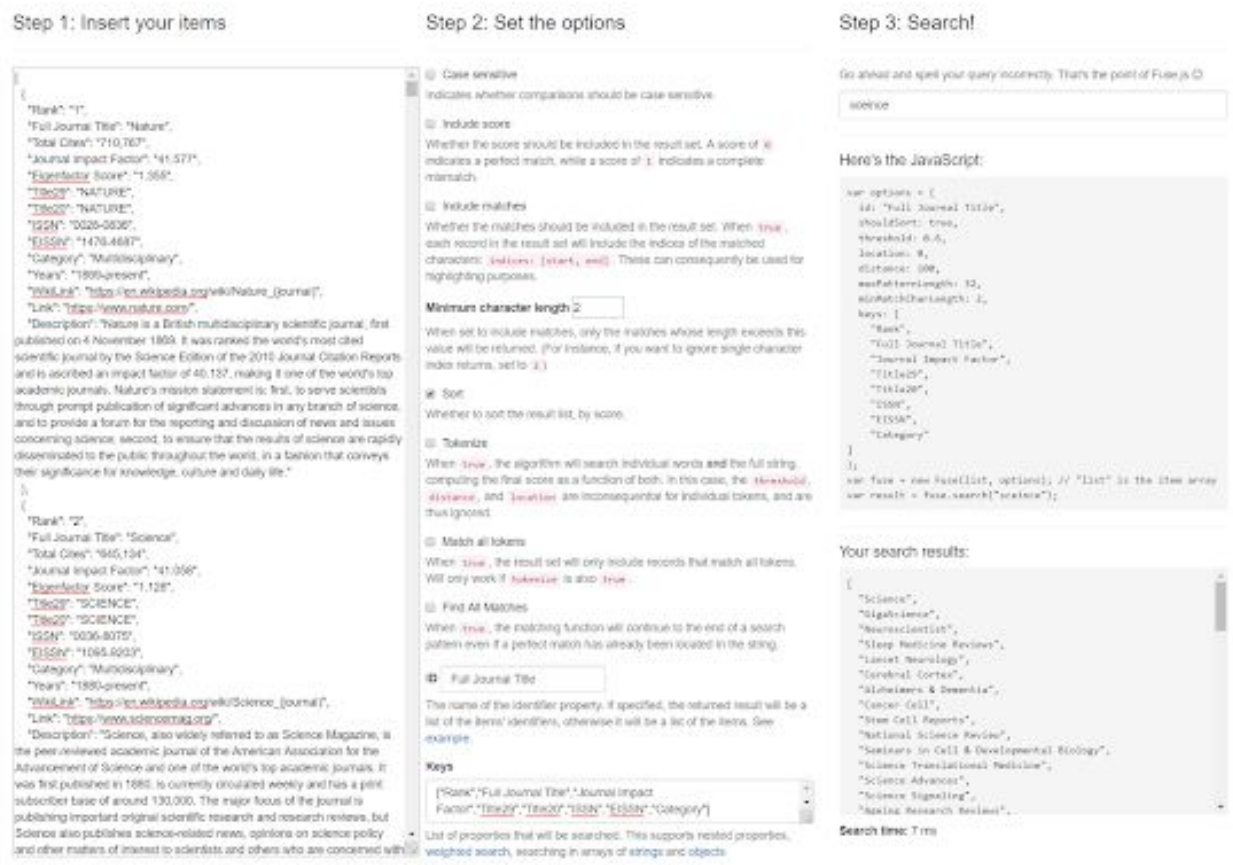
From the additional .csv files, we expect to also glean total number of citations (could be journal specific), total number of cited, and changes in these indicators over time (at a minimum between the years of 2014 to 2017).



We divvied up the sections of the project into three topics for the Process Book.

First, the interactive search bar. If we had to implement our own algorithm, we knew it would be too difficult of a task to do by the end of the semester. However, Taylor managed to find 3rd party tools to aid us. Fortunately, InCites has a publicly available master journal list in .xlsx format. While this provided journal names, abbreviations, ISSNs, ESSNs, and category, it did not provide any sort of ranking for the journals or other metrics for use with D3. Taylor merged the top 100 journals list (described below) with the matching journals from the master list, thus adding columns for impact factor, rank, and total cites. In order to display relevant journal information, Taylor had to manually search for active years, journal website, and journal description for each row in the .csv above. This was a laborious task, but it couldn't be avoided. The resulting single .csv is concise and easily applicable to D3.

The .csv created above with journal names, abbreviations, impact factors, ISSNs, ESSNs, category, active years, website, and journal description was converted to a .json for use with fuse.js: the fuzzy search javascript library we used because creating our own algorithm would take too long. The next step is for us to implement fuse.js in our workspace to create the interactive search bar for the journals. It's implementation will be somewhat like this live demo using the .json discussed above.



Second, taking the large .csv data and preprocessing them into the right format for easy implementation of our visualizations. This was split into two parts. Michael developed the skeletal backbone of our project which had folders for unprocessed and processed data, folders to hold Javascript files, and folders that allow our HTML file to read in those folders. One dataset was manually processed in order to give Michael some data to start working on developing the force-directed network. Meanwhile, an automatic script to download the rest of the .csv files in the correct format was being created. This enabled Brian to develop

two key Python scripts: ImportCSVRename.py and FindingImpactFactors.py. We tried to load in the .csv files, but there were issues. The first line and last two lines of each .csv were copyrights and journal names but not headers. On top of that, the names of the files were all a generic name, so downloading multiple files at once would overwrite each other. To solve this, we removed the first line and last two lines and renamed the recently saved files based on the first line that we removed.

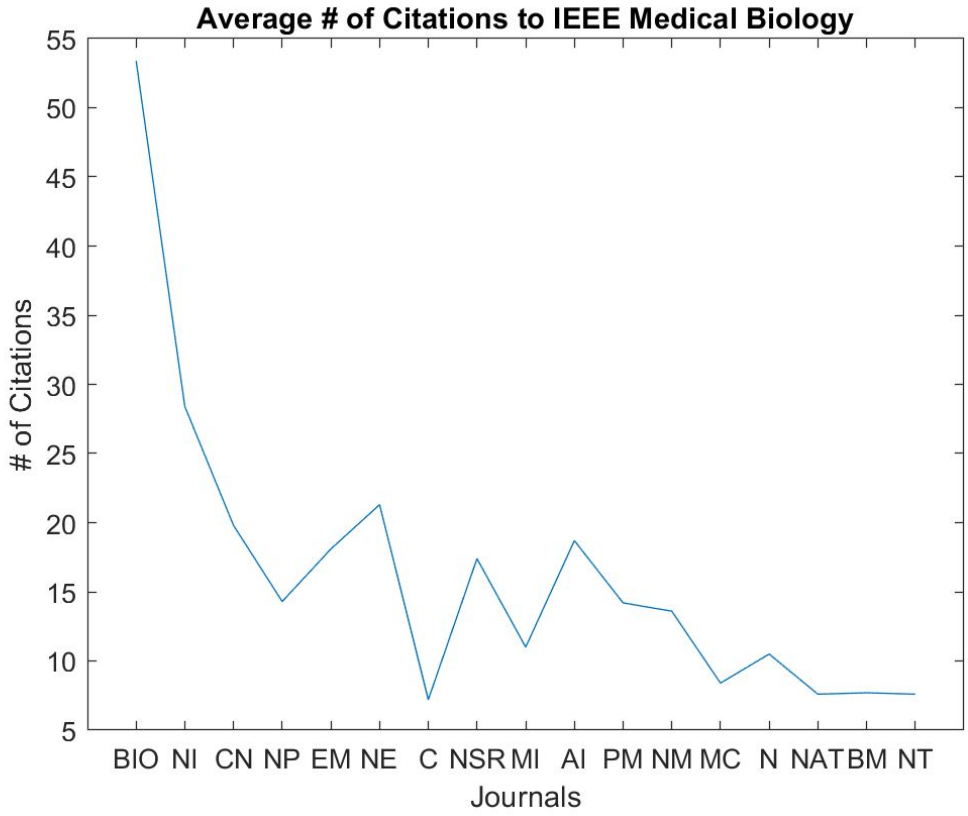
However, the problems didn't stop there. Our datasets were taking our original three .csv files, naming them appropriately, then saving them as three copies of the first of the three .csv files. Apparently, one of our if statements was not logically correct and would constantly be true for the first case. After this, we loaded in our parsed, correctly valued .csv files only to find out that every other line contained white spaces which would ruin our D3 code. Turns out, Python .csv writer requires a new line remover flag to remove the excessive \n. Yay, we were able to automatically load in data for several datasheets. Or so we thought. We quickly realized that only the file name contained the journal they were from and none of the headers referenced it. We remembered from one of our previous homeworks that there were columns that had the same information repeated throughout. So we crudely appended a large column containing the Journal name.

By this point we hadn't decided to reduce the data. How many journals could be cited by a single journal, right? Surely it should only be a few hundred. Nope; we had .csv files containing thousands of cited journals. Because we don't have access to the InCites API, which allows dynamic journal dataset loading on-the-fly, we chose to limit the total number of journals for this project to 100. To create this subselection, we chose the top 100 journals of 2017 by impact factor that fell into one or more of the following categories: cell and tissue engineering, cell biology, clinical neurology, biomedical engineering, multidisciplinary sciences, neuroimaging, and neurosciences. We chose these categories because they were the most relevant to our own field of biomedical engineering.

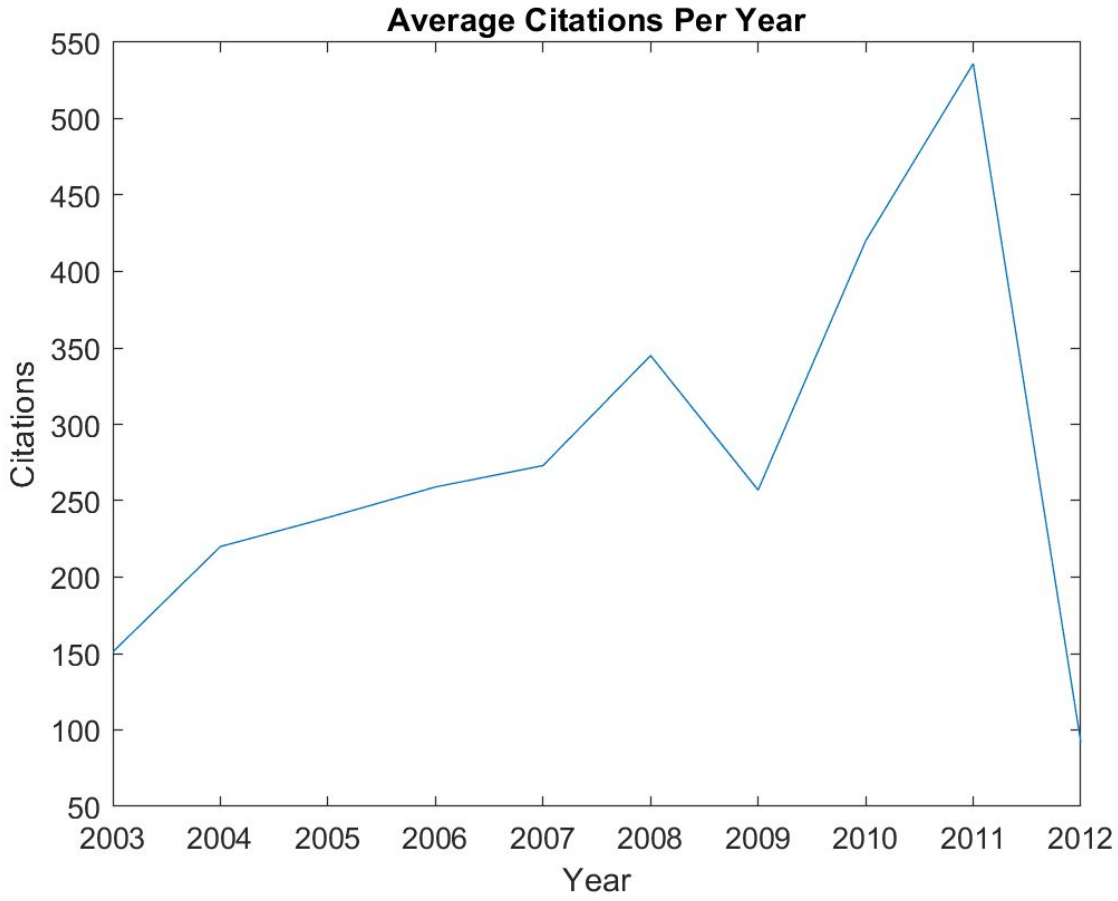
Third, visualizing the data. We realized that the first two steps would be very time-consuming. So, to balance between making sure our visuals worked and debugging our data processing steps, we hand processed one dataset. This allowed Michael to make headway on the force-directed network. These initial results are presented below in the Implementation section.

Exploratory Data Analysis:

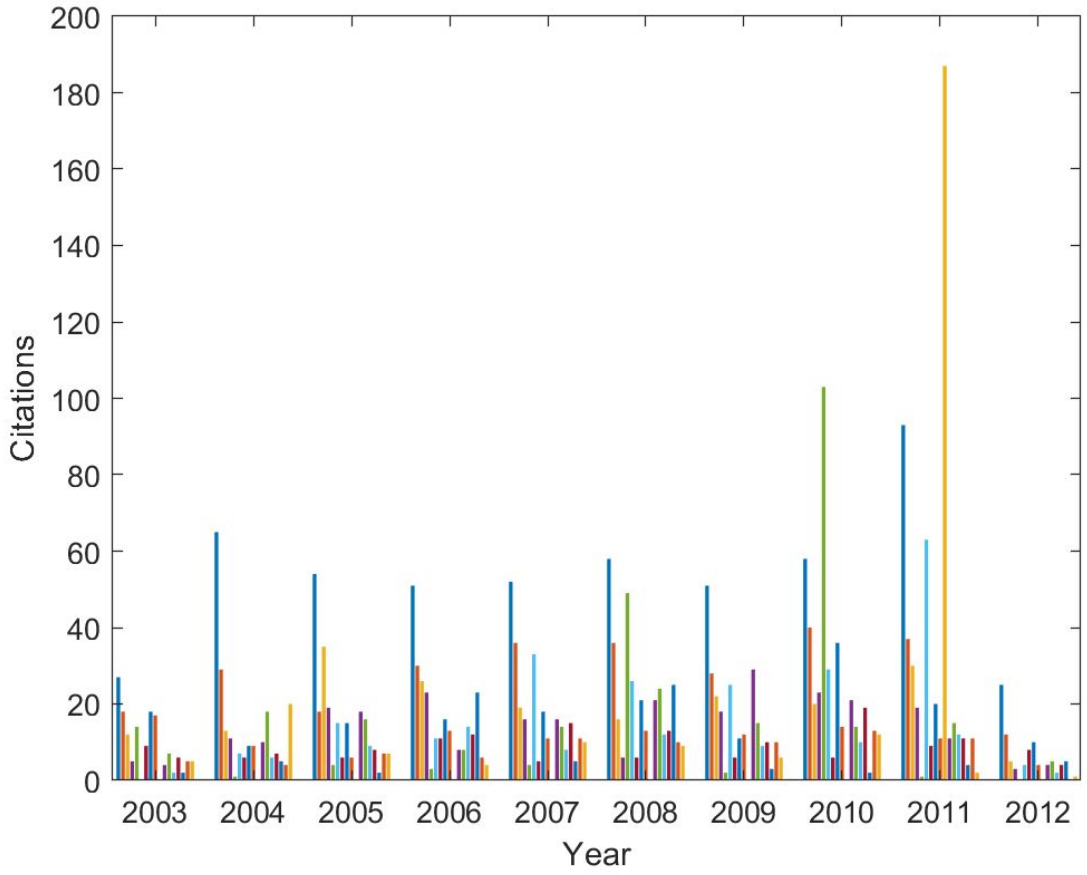
As all three of us are engineers, we utilized MATLAB to generate three prototype images exploring three different features.



The above image checks 17 different journals against the number of citations each had to the Journal IEEE Medical Biology. By doing so, we can explore if there are any subspecialties within these journals. A subspecialty can be seen by a high number of citations between some journals and low citations between others.



The above figure illustrates how one journal, IEEE Medical Biology, has changed over the years when considering the number of citations made. Prior to 2012, there was a general upward trend in these citations which is sharply contrasted by the rapid decline from 2011 to 2012. This is to be expected as the journal ceased official publication in 2010. It has since been renamed IEEE Pulse.



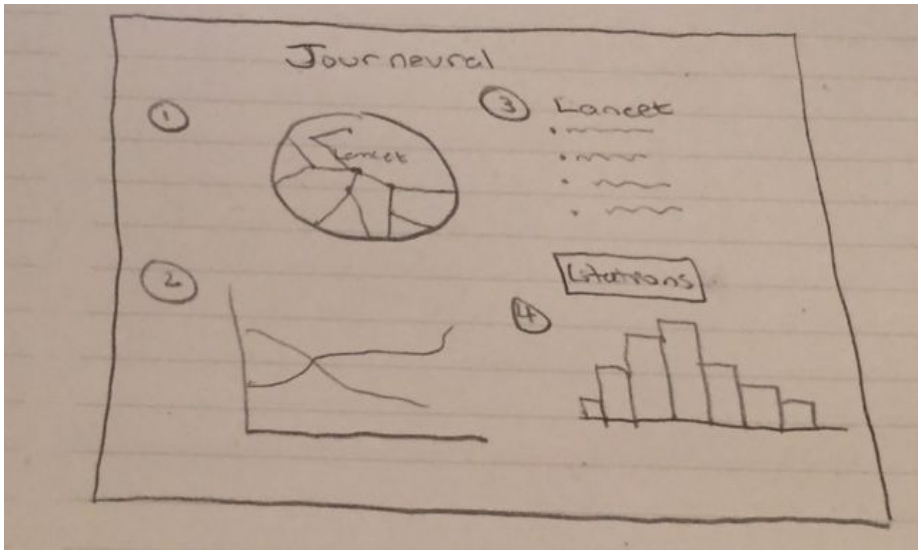
The above figure shows how IEEE Medical Biology has cited different journals over several years. Through the figure, we can pick out specific journals which seem to grow in number of citations. This could indicate a trend in which similar journals tend to become more connected over time.

Design Evolution:

Individual members were tasked with creating initial ideas for the visualization design. After forming and sketching these ideas, our group met and evaluated ideas, merged concepts, improved on individual designs, and molded the different ideas into a single draft with which we could move forward with the visualization design process.

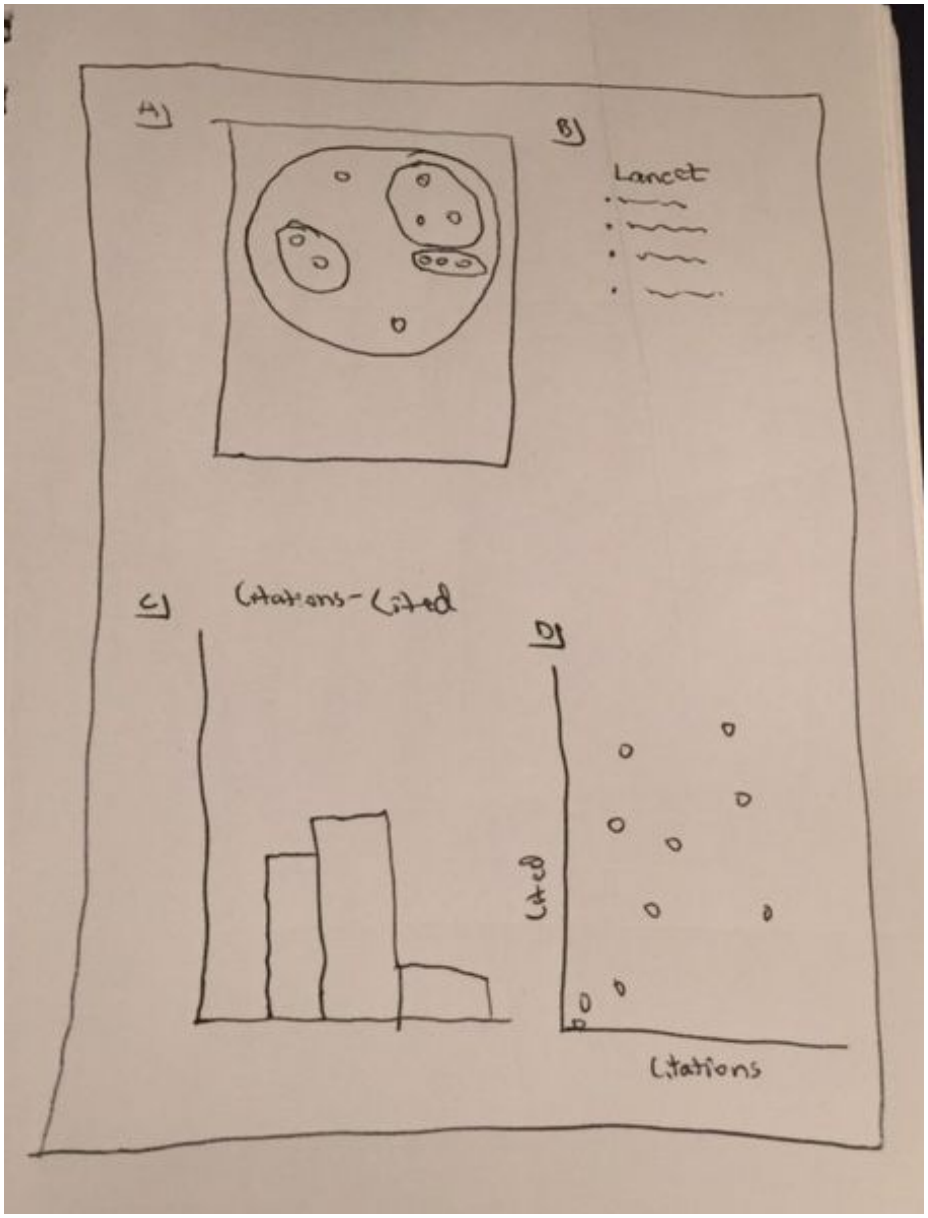
Brainstorm Design:

This design contains a force-directed network, a journal information box, a time-series graph showing trends in a journal citing another vs. being cited, and a histogram showing citations between journals for a specific year.



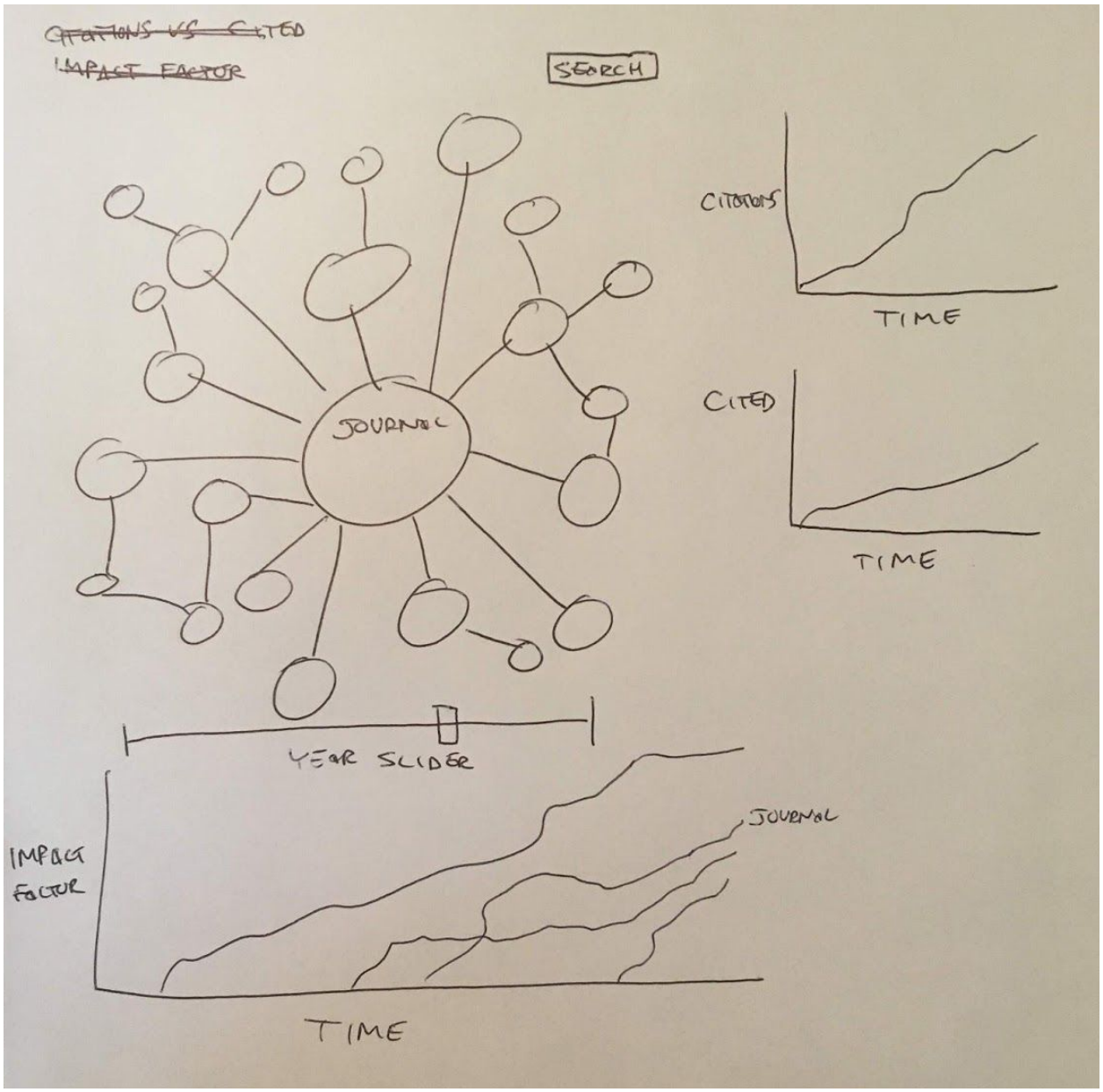
Second Design:

Our second design contained 4 main visualizations: a force-directed network that would encircle similar journals, a journal information box, a visualization showing citations over time, and a plot with the number of times a journal was cited vs. the number of times the journal cited another journal.



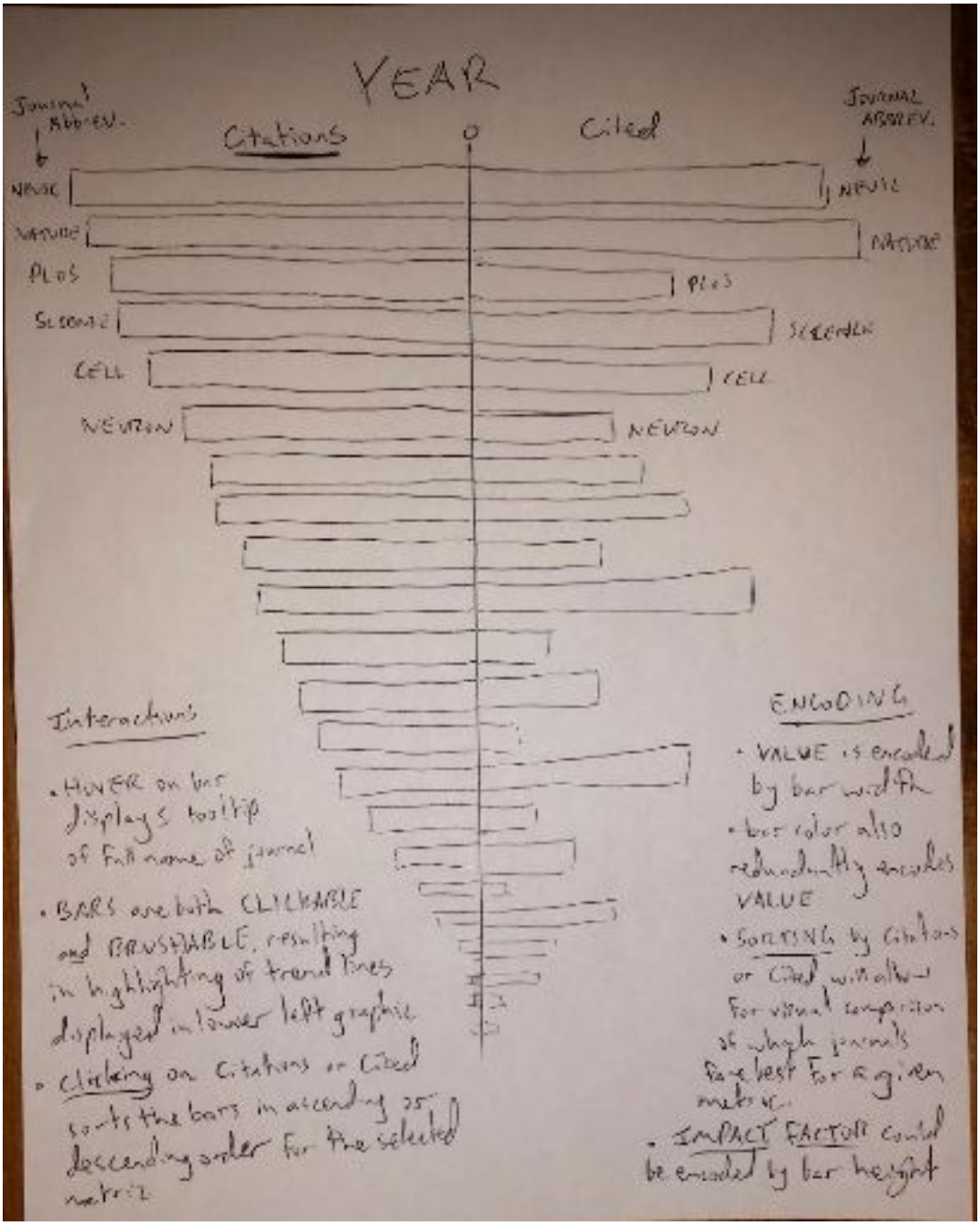
Third Design:

We added two main features to our third design: a search bar and a time-series plot showing the impact factor of journals over time (influenced by the New York Times article). We also chose to separate our citations vs. cited graph.

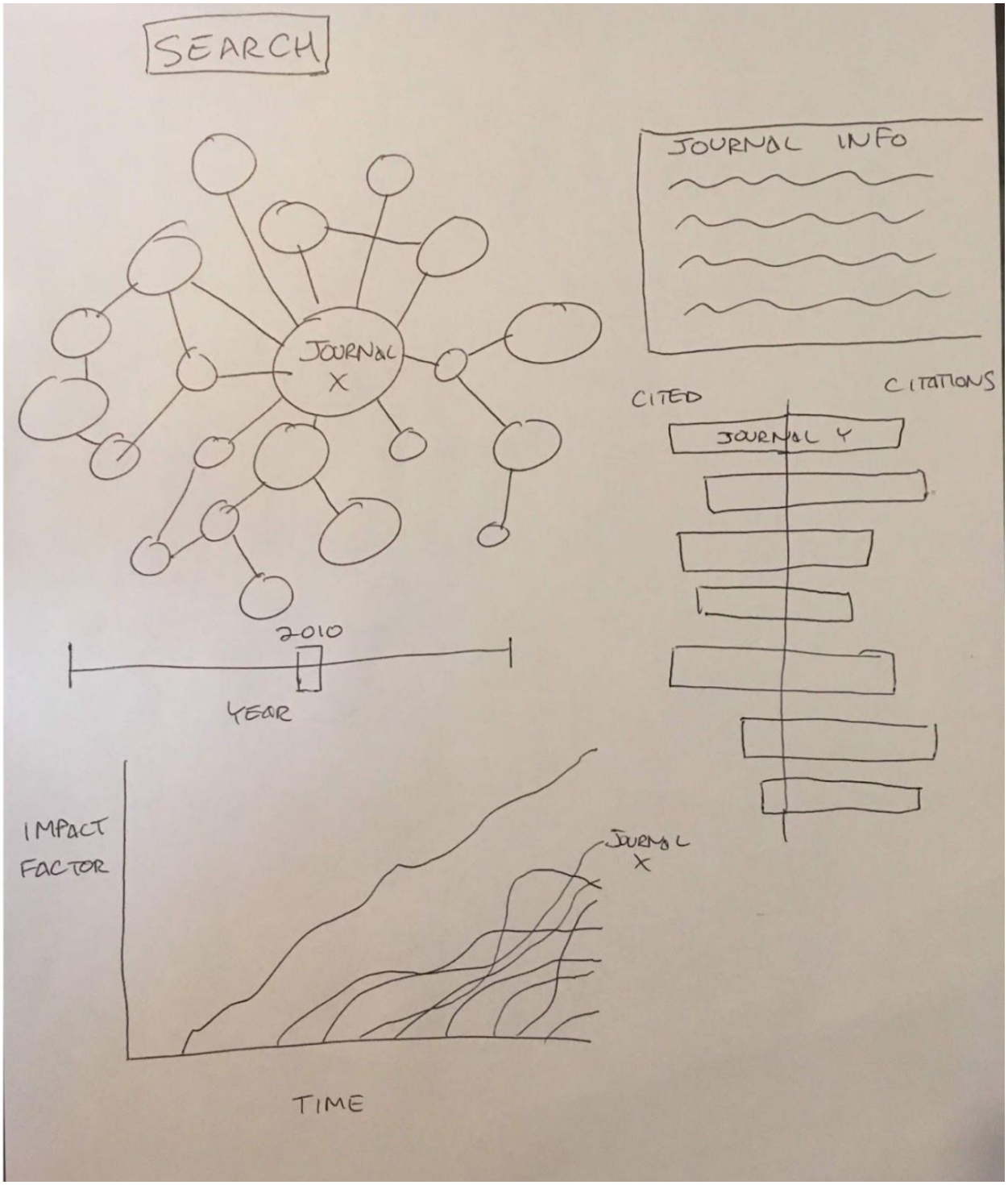


Fourth Design:

Our fourth design was similar to the third with one added graph: a horizontal bar chart showing the number of citations vs. number of times cited by a single journal. The addition to the design is shown below.



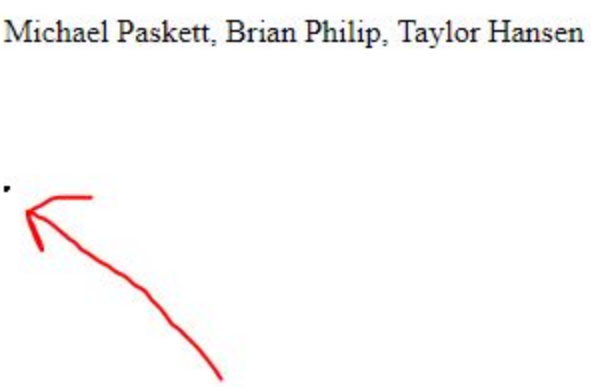
Final Proposal Design:



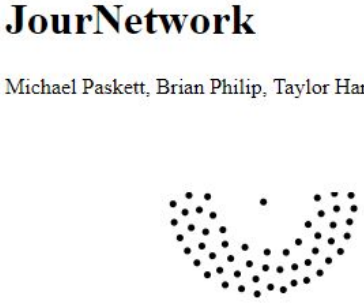
Before handing in our proposal, we unanimously agreed our final design would appear like above. While it is crude, we designed it according to the nested model. Our Domain Problem Characterization was when we realized that there

were no good visuals for understanding how journals are interconnected and where we can find journals to find relevant research or publish to. Once we knew what we had to do, we needed to manipulate the data. Luckily for us, the downloaded data is in a form that is compatible with our visualization. Lastly, our design has “4 Quadrants”. The main attraction in the top left is a force-directed network conveying three pieces of information: number of journals that cited the core journal, number of citations the core journal has for other journals, and the Journal Impact Factor. Our goal was to show how journals with high citations and is cited a lot by others are pulled close to generate a small distance between them. By this approach, the size of the network would be the size of the radius of the circle. It follows that one could see interactions between small and large impact journals and maybe parse subspecialties within the network. To do all of this, we learned about code in D3 that can compute the force-directed network.

Implementation:



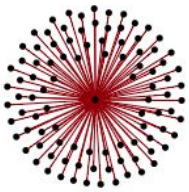
Initial attempt at force directed network. The circles appended to the svg, but not in the correct position.



Getting a little further… with a lot of help thanks to Mike Bostock.

JourNetwork

Michael Paskett, Brian Philip, Taylor Hansen

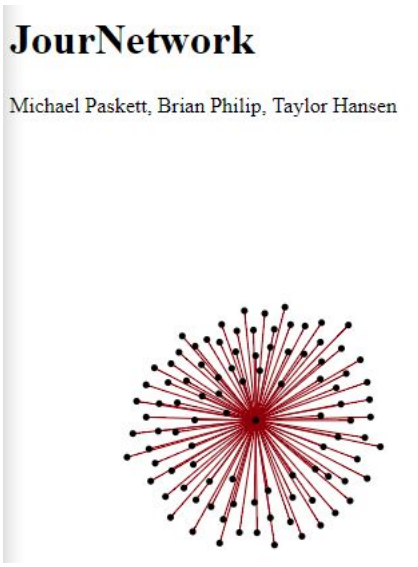


Got links to show up.

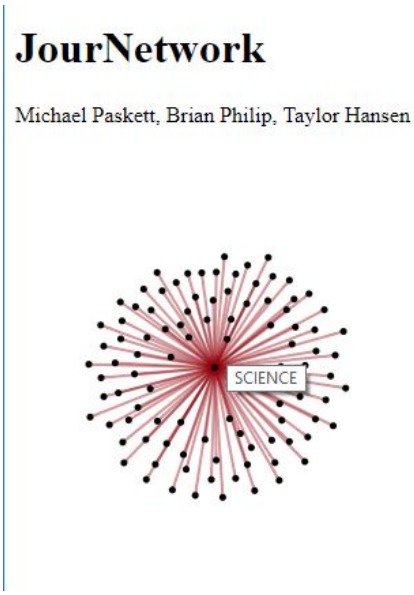
JourNetwork

Michael Paskett, Brian Philip, Taylor Hansen

Links now vary according to impact factor… working toward making the distance according to citations between the journals.



Distances are now being correctly being changed by number of citations. We chose to do this on a log scale due to some journals having a much higher number of citations than others.



We reduced the opacity of our links, which helps aesthetically. Hovering over a node displays the journal name.

Evaluation:

At this point the evaluation of our visualization and data is somewhat limited. We can see some general relations through the number of citations for a single journal, but we still need to load the relations between the journals other than the main journal. This will greatly increase our ability to see relationships between journals.

Our data presented some unexpected challenges that we are going to have to continue working through.

At this point, we aren't really able to answer our initial questions, but we have made a lot of ground towards being able to answer them. We still have a lot of improvements to make to our visualization, but the remainder should be easier, as they are mostly visualizations we have worked on previously.