**Visualization of the Occupational Therapy Researcher Database**

# For Assignment #1­10

***1****.* ***Official Project Title:***

**Visualization of the Occupational Therapy Researcher Database**

***2****.* ***Title of planned visualizations:***

Temporal analysis and project analysis where there are bursts of activity.

Geospatial analysis and map of collaboration links between institutions.

Topical analysis using text associated with projects and researchers and map the evolving topic space.

Network analysis of experts and their expertise areas, projects and diagnosis visualized over time.

***3. Authors:***

Peter Annable, Roshamiliza Rahman, Shahzad Saleem, Mahesh Suravajjala & Yelena Yezerets

@ Indiana University, 2016

***4. Description of visualizations****:*

The American Occupational Therapy Foundation, in partnership with the American Occupational Therapy Association, has built up a researcher database to track research on Occupational Therapy techniques and practices. We intended to create a comprehensive understanding of the scientific community for building scientific networks, to identify scientific leaders for specific initiatives, and to summarize capacity to external stakeholder groups.

We created temporal visualizations and burst analysis using SCi2 and Gephi Fruchterman Reingold algorithm to identify and create bipartite network of 51 researchers working on 29 diagnoses. In addition, geospatial analysis of research institution funding and diagnosis timeline were created using Tableau.

Research Project table in conjunction with Project ID and Profile ID were used to calculate Total Funding Amount by summing up NIH Funding Amount, Fed Funding Amount, and Non Fed Amount.

Burst analysis was performed based on 29 diagnosis descriptions using burst detection algorithm via Sci2.

***5. Discussion***

## What problems surfaced during validation and how does your redesign resolve them?

Burst analysis of the various diagnosis areas gets skewed as more and more research areas were funded over time. Burst analysis had to be re­created multiple times by adjusting Gamma & Density in order to make sure all the important diagnosis areas have been highlighted and persistence of funding is appropriately visualized. In order to further substantiate various diagnosis areas that were funded, and highlighted, a word cloud by decade has been created, using Tableau. This word cloud along with the burst analysis should clearly articulate how various various research areas have been funded over the decade, and also highlights how research areas evolved over time.

## Discussion of challenges and opportunities. Address complexity and scaling issues, desirable modifications and extensions of your work. You have only four weeks for this project restricting the amount of work that can be done considerably. I would like to know what promising avenues you see for future work.

Challenges:

* 1. Most of the research projects covered multiple diagnosis areas. There are no specific fields to highlight how much of the funding went towards which area. So, it was difficult to use funding amount in temporal analysis.
  2. Tokenization and Normalization of health care jargons is a big complex. Different set of stopwords had to be created.

Opportunities:

1. Results of temporal analysis should be overlayed with various diseases that primarily hit during this period. Overlaying this disease information with AOTF research areas could explain if AOTF research projects were reactive or proactive in nature. Also, it will highlight how relevant research areas have been with the conditions/health care issues prevalent during those times.
2. Peer comparison could be interesting. Comparing AOTF research areas and funding amounts to other organizational funded research areas will explain if there is funding overload in some areas, or lack of funding. Such analysis will also highlight if the research areas are complementary or supplementary in nature with the industry.

**Topical Analysis (lead: Shahzad Saleem)**

Goal: Complete a topical analysis using text associated with projects and researchers and map the evolving topic space.

Need: Topical analysis will extract the set of unique words and their frequency from text associated with projects and researchers in occupational therapy research dataset. We will be creating a word co­occurrence network visualization for this dataset. Extract Word Co‐Occurrence Network will create a weighted network where each node is a word and edges will connect words to each other, where the strength of an edge will represent how often two words occur in the same body of text together. Stop words, such as 'the' and 'of' are removed, and stemming will be applied. Once done this visualization will highlight the important words/text for occupational therapy research.

Related work: There has not been any Topical analysis done on Text associated with projects and researchers involved in occupational therapy research.

First and 2nd Iteration of Visualizations

In order to do the Topical analysis on Keywords associated with projects and researchers in occupational therapy research dataset. Excluded word “Null” and combined all three keyword fields (USR\_KEYWORD\_1, USR\_KEYWORD\_2, USR\_KEYWORD\_3) in to one called “ALL\_KEYWORDS”. Converted xls file into csv and loaded in Sci2 tool.

Applied Pre­Processing step “Lowercase, Tokenize, Stem, and Stopword” to standardize text. Once standardized explored data and found out there are total of 470 unique keywords, which got referred from 1 to 59 times.

Breakdown is shown in following table,

|  |  |  |  |
| --- | --- | --- | --- |
| Words Count  228 | Count of appearance  1 | Words Count  1 | Count of appearance  16 |
| 81 | 2 | 1 | 17 |
| 42 | 3 | 2 | 18 |
| 30 | 4 | 2 | 19 |
| 13 | 5 | 3 | 20 |
| 15 | 6 | 1 | 23 |
| 6 | 7 | 2 | 24 |
| 11 | 8 | 1 | 25 |
| 4 | 9 | 2 | 26 |
| 3 | 10 | 1 | 37 |
| 2 | 11 | 1 | 45 |
| 3 | 12 | 1 | 46 |
| 6 | 13 | 1 | 48 |
| 1 | 14 | 1 (Stroke) | 59 |
| 5 | 15 |  |  |

From above can be seen the most frequent word was Stroke which was encountered 59 times in keywords field.

Applied “Extract Word Co­Occurrence Network” algorithm with “ALL\_KEYWORDS” as input parameter. Using “Extract Nodes Above or Below Value” step, selected only nodes (Keywords) which got referred at least 6 times in the text, Used “Extract Edges Above or Below Value” algorithm to select edges (relationships among keywords) with the value at­least 3. Executed “Network Analysis Toolkit” and “Remove Isolates” procedures to get rid of 15 isolates.

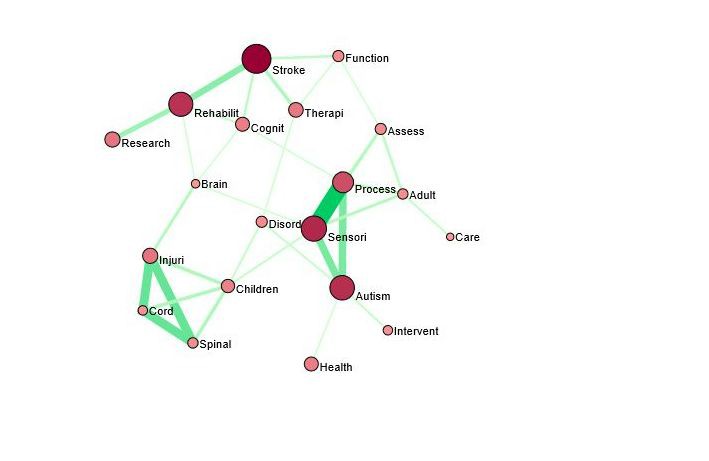
For Visualizations, Tried multiple visualizations and selected following

1: TreeMap, This visualization is showing Keyword names and number of times they got referred in the keywords field, selected only keywords which got referred at­ least 6 times. Color scheme lighter to darker is based on number of references for those keywords. (Used Tableau)Geo



2: Second Visualization is Kambada Kawai Layout using GUESS, through which we can see Keywords and their relationships, For better understanding and readability Picked top twenty keywords based on their reference count (Ranging from reference count 16 to 59), node size and color is reflecting reference count.

Edge (relationship among nodes) is reflected with color and relationship strength (ranging from count 3 to 33).



## What problems surfaced during validation and how does your redesign resolve them?

In order to do topical analysis, multiple visualizations were tested: Word Cloud, Tree Map, Network with many different layouts, GEM, Circular, Radial Tree, Kambada Kawai. Preprocessing step on keywords caused issues because stemming caused use to lose meaning in some of the terms. For example, the word “injury” became “injuri”, “Early Intervention” was one word but after tokenization became two words: “earli” and “intervent”. Similarly, “Spinal Cord” became Spinal and Cord . Due to this a lot of edges arose between the nodes which should not have been if we would have preserved the space between these kinds of words. I manually modified network file after executing Word Co Occurrence network so that word reflection is appropriate as encountered in the input file.

## Discussion of challenges and opportunities.

THe word cloud was prefect where I was able to show all 470 keywords and frequency of words was being displayed through font size. However, it was challenging to show all those keywords in any other layouts. As a trade off­ for

the network analysis, I picked only keywords which showed up in the text at least 6 times for the Kambada Kawai layout.

I gound a lot of spelling mistakes and related issues in data due to which Word Co­occurrence algorithm was not producing correct results and had to manually review all keywords. Use of existing stop words list along with standard tokenization technique with this healthcare dataset was a challenge so edited the network file multiple times

Not having an Undo capability in GUESS made things much more time consuming too, one single mistake made me redo all again. As a consolation, it was kind of a blessing in disguise in that I discovered a lot of its features. Down the road I see a need for industry specific stop words list and tokenization technique, GUESS tool improvement.

# Geospatial Analysis (lead: Rose Rahman)

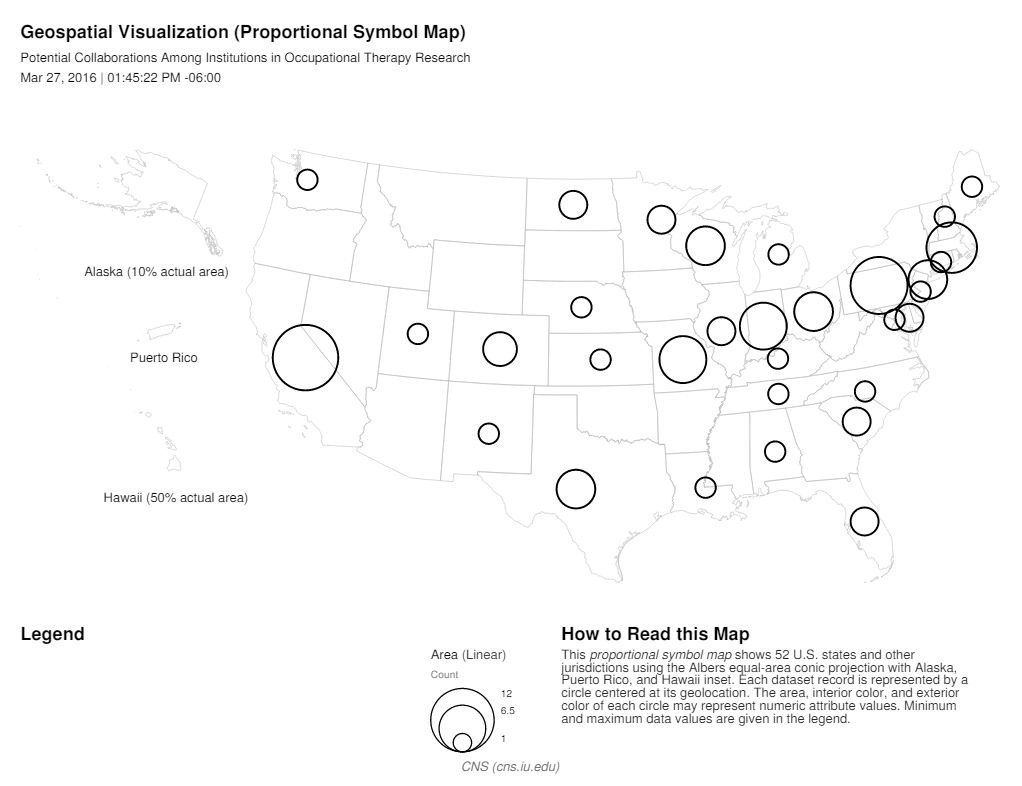
Goal: Complete a geospatial analysis and map to identify potential collaboration links between institutions.

Title of planned visualization: Proportional Symbol Map of Potential Collaboration Links Among Institutions in Occupational Therapy Research, YYYY – YYYY

The proportional symbol map is necessary to map out the institutions with scientific leaders in Occupational Therapy research as to visualize the potential collaboration links among them. The nodes may represent the institution mapped on the U.S. map based on states/zip codes. The size of the nodes may represent the total amount of funding received by the institutions. The color of the nodes may represent the number of researchers within each institution. Discussion will be held among group members to reach decisions on mapping details.

Related work: There has not been any geospatial analysis done on institutions conducting research occupational therapy, particularly proportional symbol map. There has been some analysis done on mapping the literature of occupational therapy however, there is no actual geospatial visualization.

Mock­Up of Visualization



The cleaned data set is aggregated by zip code (unique to institutions) and sum the total funding amount, number of researchers and number of projects. The resulting data set contains 29 rows of unique institutions. Five existing attributes were retained to create the data set for geospatial analysis. An additional of seven attributes were created in geocoding process to extract longitude and latitude markers.

The geospatial analysis was run using Congressional District Geocoder. There were five zip codes which could not been given a congressional district. The longitude and latitude markers were located using GPS Visualizer online tool for the five zip codes. The data was then visualized using proportional symbol map with the size of the edges represents the total amount of funding and the color of the circles represents the number of researchers in sample 1, while in sample 2, the size of the edges represents the number of projects and the color of the circles represent the number of researchers.

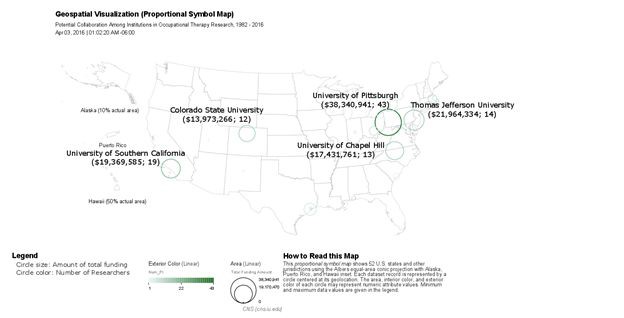
The proportional symbol maps clearly show the top five players in the occupational therapy research – The University of Pittsburgh, Thomas Jefferson University, University of Southern California, University of Chapel Hill, and Colorado State University. These institutions have the most researchers who have conducted the most research

projects from 1982 until recent year (based on the earliest funded project). The two maps showing different attributes did not differ as much. The map, however, will not be a good visualization to show the topics researched by these institutions.

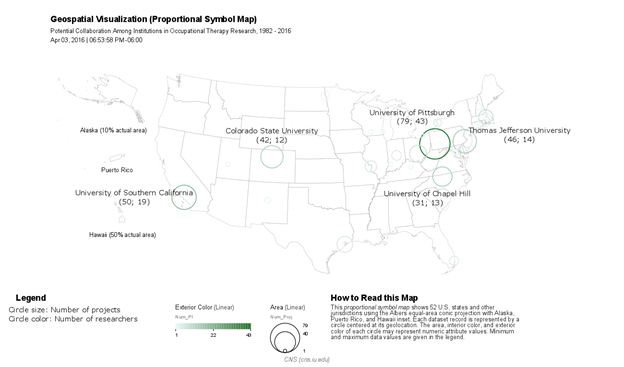
Further work on geospatial analysis and map:

The proportional symbol map function in Sci2 (on my computer) did not show the option to assign the edge color, which will be useful to make the map more readable, visually. I will explore this issue and hoping that the final map will have edge color.

Sample 1: Proportional symbol map showing total amount of funding and number of researchers



Sample 2: Proportional symbol map showing number of research projects and number of researchers



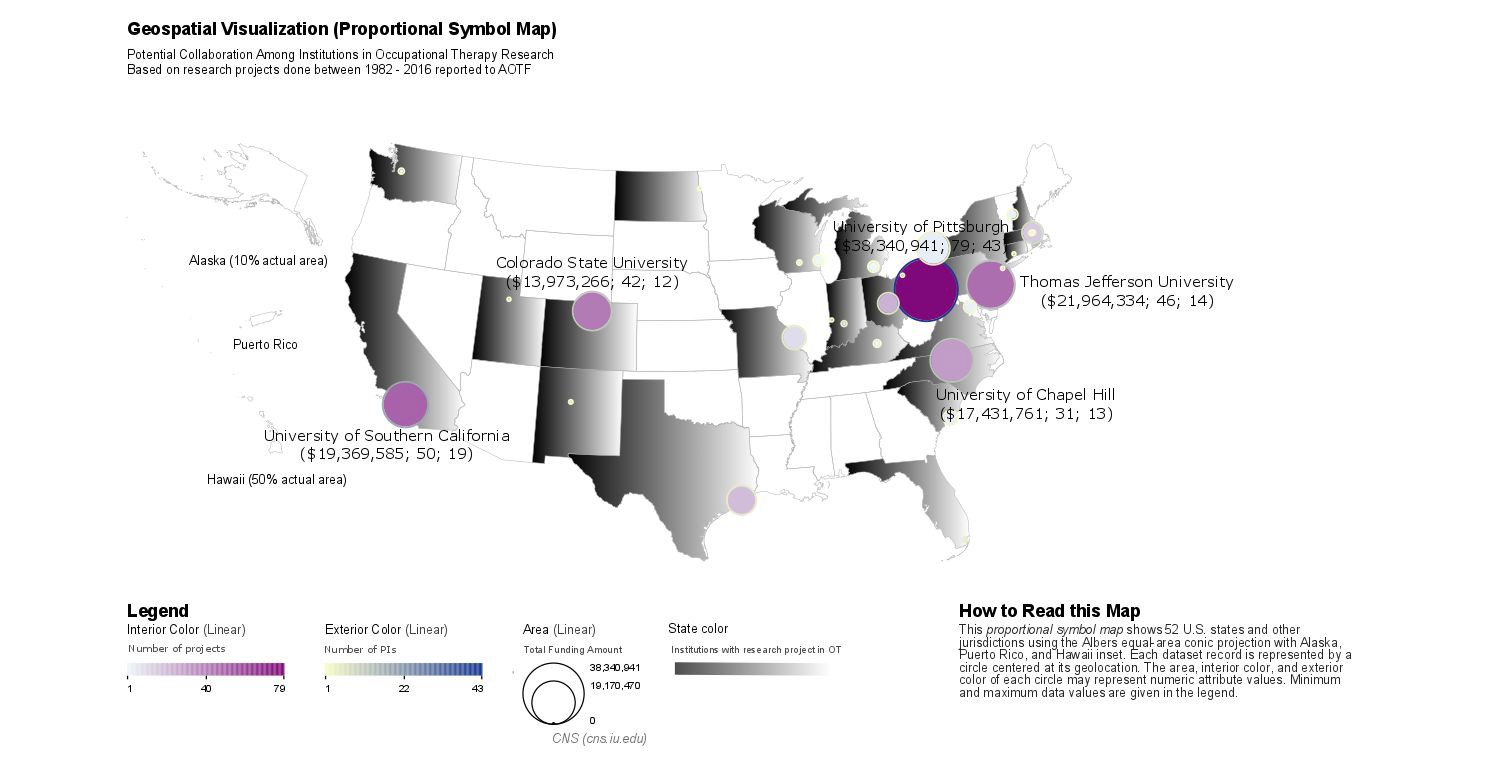
.

1. **What problems surfaced during validation and how does your redesign resolve them?**

There were some discrepancies in the locations of the nodes when comparing the geospatial map created using the parameters described earlier in Sci2 and Tableau. The issue seems to stem out from differences in zip codes since zip codes were not provided in the original data set from AOTF. A thorough google search was done to find the best zip codes. Geocoding was rerun on the updated zip codes using the Generic Geocoder. Longitude and latitude information was found on all 29 zip codes. The map was updated to show total funding amount, number of projects and number of researchers.

1. **Discussion of challenges and opportunities.**

Since the map shows the potential collaboration links among the institutions, it would be helpful to show the network connections based on the keywords. However, trying to create the networks where there are three keyword columns with multiple rows per institution poses some complexities. In addition, the need to use other software like Photoshop, Adobe Illustrator and/or GIMP to overlay the network onto the map also adds to the complexity and is time consuming. The network overlay will certainly be an informational piece to add on the geospatial map for future visualization.



[Proportional Symbol Map\_2016Apr15.pdf](https://iu.instructure.com/files/61621878/download?download_frd=1&amp;verifier=h25Pg48hHm1ueCBzIwra52oDf500ZG4pE2aZigeS)

**Network Analysis (lead: Helen Yezerets)**

Goal: Complete a network analysis and identify networks of experts and their institutions/expertise areas, projects and diagnosis/ICF/categories then visualize and animate these over time.

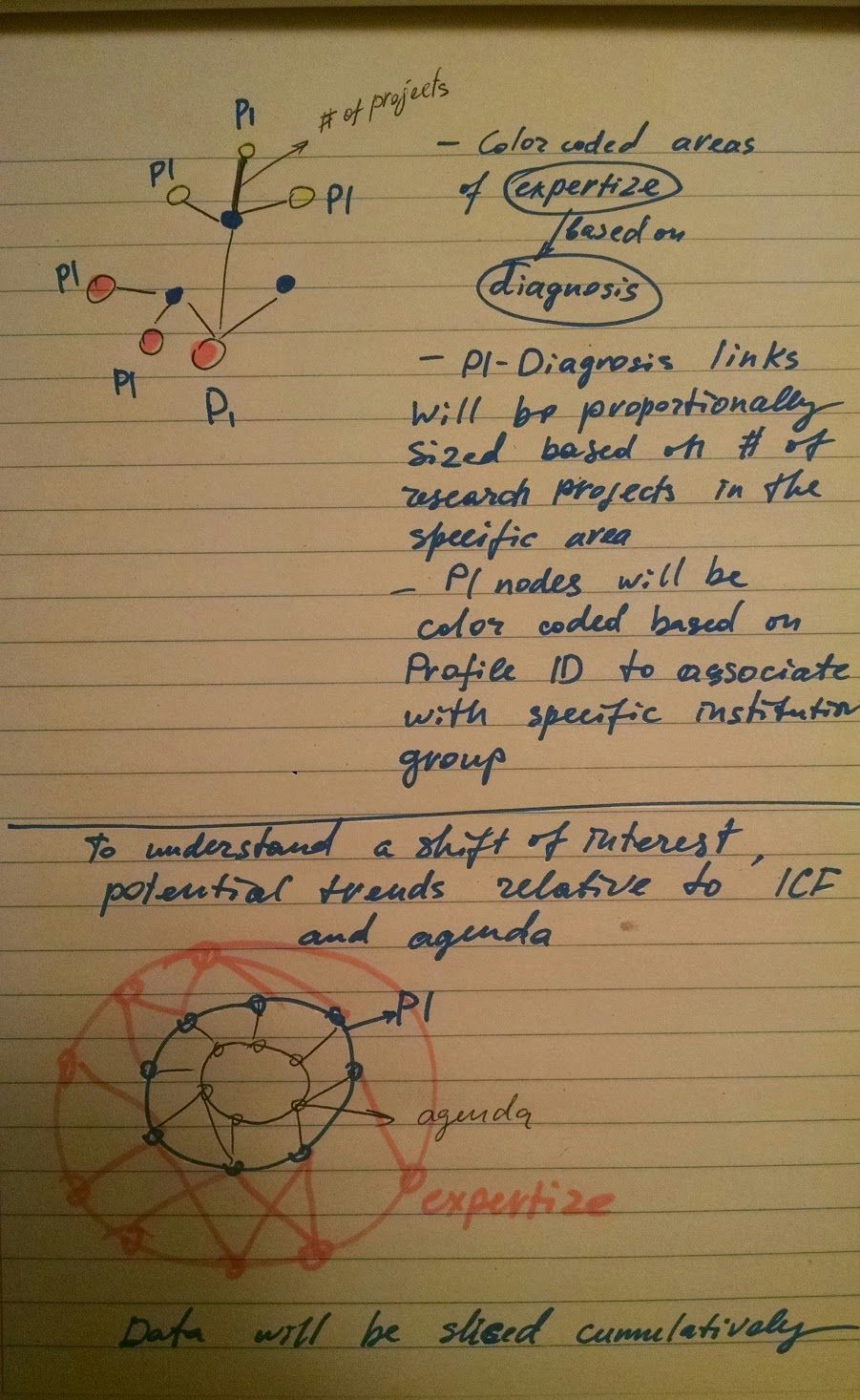
Significance: the network analysis will elicit insights into the existing scientific networks and help to identify scientific leaders for specific expertise areas based on diagnosis/ICF/categories relevant to particular research project.

The following features will be implemented:

* + color­coded areas of expertise based on diagnosis description
  + PI nodes will be color­coded based on the profile id to associate research projects with specific organizations
  + PI­diagnosis links will be proportionally sized based on # of research projects in the specific area.

To understand shift of interest, potential research trend relative to ICF and agenda data will be cumulatively sliced according to the funding date from 1982 to 2016.

Mock­up of visualization



Network Visualizations

#6. Simple statistics of the data sets used, e.g., number of entities, major entity attributes, etc. The following clean up was performed based on the Julie Bass’s feedback:

­Identified projects with duplicate values and deleted projects with highest Project ID number (314, 92, 99, 90, 88, 86, 89, 85, 87, 91)

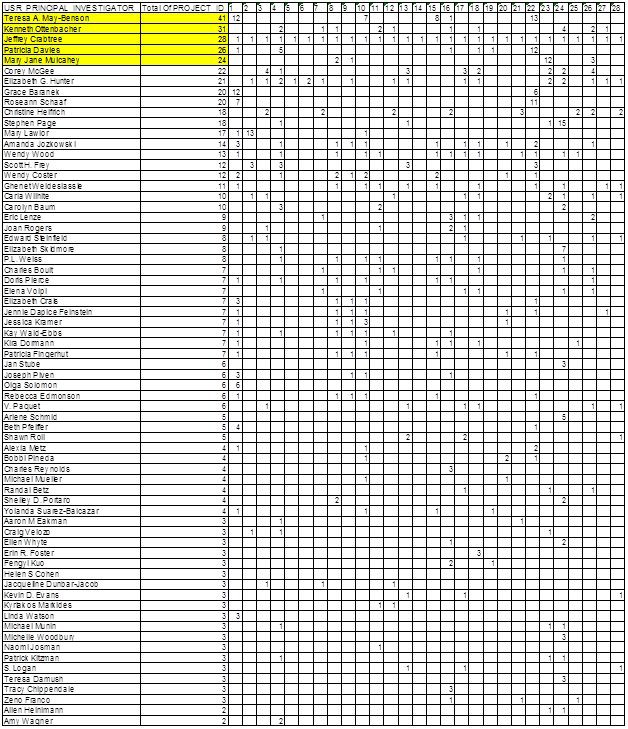
* assigned ProjectID 9999 to Project ID 1
* reassigned Ellen S Cohn projects to Helen S Cohen
* Updated blank USR\_FUNDING\_DATE\_END with the current date.

Performed crosstab analysis between USR\_PRINCIPAL\_INVESTIGATOR and number of associated parameters:

USR\_PRINCIPAL\_INVESTIGATOR (193 unique) vs USR\_AGENDA\_CATEGORY (7) (excerpt): top 5 PIs depicted in yellow. [Complete dataset](https://iu.instructure.com/groups/253971/files/folder/Diagnosis%20ICF%20Agenda?preview=61355596)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| USR PRINCIPAL INVESTIGATOR | Total Of PROJECT ID | 1 | 2 | 3 | 4 | 5 | 6 | 8 |
| Steohen Paae | 26 | 3 12 | | 1 | 8 | 1 | 1 |  |
| Kenneth Ottenbacher | 25 | 2 | 1 | 1 | 9 | 7 | 3 | 2 |
| MarvJane Mulcahey | 20 | 11 | 3 |  | 1 | 1 |  | 4 |
| Patricia Davies | 20 | 9 | 3 |  | 8 |  |  |  |
| Marv Lawlor | 17 | 9 | 1 | 5 |  | 1 | 1 |  |
| Scott H.Frey | 17 | 3 | 2 | 3 | 3 | 3 | 3 |  |
| Roseann Schaaf | 15 |  | 2 |  | 10 | 1 |  | 2 |
| Tere sa A. May-Benson | 15 | 8 | 5 |  | 2 |  |  |  |
| Grace Baranek | 14 | 2 | 2 |  | 9 |  |  | 1 |
| Amanda Jozkowski | 12 | 3 | 2 | 2 |  | 2 | 3 |  |
| Carolvn Baum | 11 | 1 | 2 | 1 | 4 | 2 |  | 1 |
| Corey McGee | 11 | 6 | 2 |  | 3 |  |  |  |
| Helen S Cohen | 10 | 2 | 1 | 3 | 2 | 2 |  |  |
| Joan Roaers | 10 | 1 | 2 |  | 7 |  |  |  |
| Christine Helfrich | 8 |  | 2 | 3 |  | 3 |  |  |
| Elizabeth Skidmore | 7 |  | 5 |  | 2 |  |  |  |
| FenaviKuo | 7 |  | 3 | 2 |  | 2 |  |  |
| Michae IMuelier | 7 | 1 | 1 | 1 | 1 | 1 | 2 |  |
| Oloa Solomon | 7 |  | 2 | 1 |  | 2 |  | 2 |
| Ariene Schmid | 6 |  | 1 |  |  | 1 |  | 4 |
| Elizabeth Cra is | 6 |  | 2 |  | 1 |  | 3 |  |
| Elizabeth G.Hunter | 6 |  | 1 | 1 |  | 3 | 1 |  |
| Erin R.Foster | 6 |  | 3 | 3 |  |  |  |  |
| Bobbi Pineda | 5 | 1 | 1 | 1 |  | 2 |  |  |
| Eric Lenze | 5 | 1 | 3 |  | 1 |  |  |  |
| Ghenet We Ide slassie | 5 | 1 | 1 | 1 |  | 1 | 1 |  |
| Jan Stube | 5 |  | 3 |  |  |  |  | 2 |
| Shelley D.Portaro | 5 | 2 | 1 |  | 2 |  |  |  |
| Beth Pfeiffer | 4 |  | 3 |  | 1 |  |  |  |
| Charles Reynolds | 4 |  | 1 |  | 2 |  | 1 |  |
| Elicia Dunn Cruz | 4 | 1 | 1 |  |  | 1 | 1 |  |
| Joseoh Piv en | 4 | 1 |  |  | 2 |  | 1 |  |
| Miehelie Woodburv | 4 | 1 | 3 |  |  |  |  |  |
| Naomi Josman | 4 | 3 |  |  |  | 1 |  |  |
| Christine Arenson | 3 |  | 1 |  |  | 1 |  | 1 |
| Dorothv Edwards | 3 |  | 2 |  | 1 |  |  |  |
| Eileen S.Auerbach | 3 | 1 | 1 |  | 1 |  |  |  |
| Jacob J Bloombero | 3 | 1 | 1 | 1 |  |  |  |  |
| Jessica Kramer | 3 | 1 | 2 |  |  |  |  |  |
| Joseoh Maraolick | 3 | 1 |  | 1 | 1 |  |  |  |
| Kira Dormann | 3 |  |  |  | 1 | 1 | 1 |  |
| Linda Watson | 3 |  | 2 |  | 1 |  |  |  |
| Michae IBoninaer | 3 |  |  |  | 1 |  | 2 |  |
| Sharon A. Cermak | 3 | 1 | 1 | 1 |  |  |  |  |
| Shawn Roll | 3 | 1 | 1 |  | 1 |  |  |  |
| Stev e Cramer | 3 | 1 | 1 | 1 |  |  |  |  |
| Teresa Damush | 3 |  |  |  |  |  |  | 3 |
| T *racv* Chinnendale | 3 |  | 2 |  | 1 |  |  |  |
| Wendy Coster | 3 | 3 |  |  |  |  |  |  |
| A. Peters | 2 |  | 1 |  |  | 1 |  |  |
| Aaron M Eakman | 2 |  | 1 | 1 |  |  |  |  |
| Aaron Ste infeId | 2 |  |  | 1 |  |  | 1 |  |
| Alexia Metz | 2 |  |  |  | 2 |  |  |  |
| AlixSle io ht | 2 |  |  | 1 |  | 1 |  |  |
| Amy Waaner | 2 |  | 2 |  |  |  |  |  |
| Barb Hoooer | 2 |  |  |  | 2 |  |  |  |
| **Rrhr Rnkhn11r** | ? |  | 1 |  |  | 1 |  |  |

USR\_PRINCIPAL\_INVESTIGATOR vs USR\_DIAGNOSIS\_AREA(29) (excerpt): top 5 PIs depicted in yellow [Complete dataset](https://iu.instructure.com/groups/253971/files/folder/Diagnosis%20ICF%20Agenda?preview=61355615)



USR\_PRINCIPAL\_INVESTIGATOR (193) vs ICF Categories(7) (excerpt) [Complete dataset](https://iu.instructure.com/groups/253971/files/folder/Diagnosis%20ICF%20Agenda?preview=61355616)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| USR PRINCIPAL INVESTIGATOR | Total OfPROJECT ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Kenneth Ottenbacher | 42 | 6 | 5 | 6 | 9 | 6 | 1 | 9 |
| Teresa A. May-Benson | 41 | 13 | 11 |  | 4 13 | |  |  |
| Patricia Dav ies | 34 |  | 13 | 13 | 1 | 1 | 1 | 5 |
| Stephen Paoe | 29 | 1 | 17 | 3 | 6 | 2 |  |  |
| Grace Baranek | 26 |  | 6 | 2 | 5 | 5 | 3 | 5 |
| MarvJane Mulcahey | 24 | 1 | 3 | 7 | 3 | 2 |  | 8 |
| Marv Lawlor | 20 | 17 |  |  |  | 3 |  |  |
| Scott H.Frey | 18 | 3 | 3 | 3 | 3 | 3 | 3 |  |
| Roseann Schaaf | 16 |  | 11 |  | 2 |  |  | 3 |
| Amanda Jozkowski | 15 | 2 | 2 | 2 | 3 | 3 | 3 |  |
| CoreY McGee | 15 | 4 | 4 | 2 | 2 | 1 | 1 | 1 |
| Elizabeth Skidmore | 15 |  | 7 | 1 | 4 | 3 |  |  |
| Jan Stube | 15 |  | 1 | 1 | 5 | 4 | 4 |  |
| Joan Rooers | 14 | 1 | 3 | 1 | 3 | 3 |  | 3 |
| Oloa Solomon | 14 | 4 |  |  | 4 | 6 |  |  |
| Carolyn Baum | 13 | 1 | 2 |  | 2 | 4 |  | 4 |
| Erin R.Foster | 12 |  | 3 |  | 3 | 3 | 3 |  |
| Naomi Josman | 11 | 2 | 3 | 2 | 2 | 2 |  |  |
| Helen S Cohen | 10 | 3 | 3 | 3 | 1 |  |  |  |
| Christine Helfrich | 9 | 3 |  |  | 2 | 3 | 1 |  |
| TraCY Chinnendale | 9 | 3 | 2 |  | 2 | 2 |  |  |
| MarkT .Heoel | 8 | 2 |  |  | 2 | 2 | 2 |  |
| MichaeIMuelier | 8 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| Ariene Schmid | 7 | 1 |  |  | 1 | 1 |  | 4 |
| Bobbi Pineda | 7 | 2 | 1 |  | 1 | 2 | 1 |  |
| FenqyiKuo | 7 | 2 | 1 |  | 1 | 3 |  |  |
| Barbara Bokhour | 6 | 1 | 1 | 1 | 1 | 1 | 1 |  |
| Carla Wilhite | 6 | 2 |  | 1 | 1 | 1 | 1 |  |
| Charles Reynolds | 6 | 1 | 2 |  | 1 | 1 |  | 1 |
| Elizabeth G.Hunter | 6 |  |  |  | 1 | 2 | 1 | 2 |
| Ellen Whvte | 6 |  | 2 | 2 | 1 | 1 |  |  |
| Jessica Kramer | 6 |  |  |  | 1 | 3 | 2 |  |
| Miehelie Woodburv | 6 |  | 3 |  | 3 |  |  |  |
| Shawn Roll | 6 |  | 1 | 2 | 2 | 1 |  |  |
| Shelley D.Portaro | 6 |  | 2 |  | 2 | 2 |  |  |
| Wendv Coster | 6 |  |  |  | 2 | 3 | 1 |  |
| Alexia Metz | 5 |  | 1 | 1 | 1 | 1 | 1 |  |
| Chinohui "Jean" Hsieh | 5 | 1 | 1 | 1 | 1 | 1 |  |  |
| Elena Volpi | 5 | 1 | 1 | 1 | 1 | 1 |  |  |
| Elizabeth Schienk | 5 | 1 | 1 | 1 | 1 | 1 |  |  |
| Ghenet WeIde slassie | 5 | 1 | 1 |  | 1 | 1 | 1 |  |
| Leeanne Carey | 5 |  | 1 | 1 | 1 | 1 | 1 |  |
| Matthew Bair | 5 |  | 1 | 1 | 1 | 1 |  | 1 |
| Rachel KizonY | 5 | 1 | 1 | 1 |  | 1 | 1 |  |
| Aaron M Eakman | 4 | 1 | 1 |  | 1 | 1 |  |  |
| AmY Waoner | 4 |  | 2 |  | 1 | 1 |  |  |
| Beth Pfeiffer | 4 |  |  |  |  | 4 |  |  |
| Christine Arenson | 4 | 1 |  |  | 1 | 1 | 1 |  |
| Diane Corman Lev y | 4 | 1 |  |  | 1 | 1 | 1 |  |
| Dorothy Edwards | 4 |  | 2 |  |  | 2 |  |  |
| Elicia Dunn Cruz | 4 | 1 |  |  | 1 | 1 | 1 |  |
| Elizabeth Vanderlaan | 4 |  | 1 | 1 | 1 |  | 1 |  |
| Eric Lenze | 4 |  | 2 |  |  | 1 |  | 1 |
| GaeIOrsmond | 4 |  |  |  | 1 | 2 | 1 |  |
| Janet Poole | 4 | 1 | 1 |  | 1 | 1 |  |  |
| Kathleen LYons | 4 | 1 |  |  | 1 | 1 | 1 |  |
| **Kvrikn r\11 rkirl*P*** | 4 | 1 | 1 |  | 1 | 1 |  |  |

* Data was sliced by the USR\_FUNDING\_DATE\_START 1982 through 2017 by seven years cumulatively Number of project was proportionally (times 3) increasing from 18 to 54 to 149 to 406 to 755 in 2016 Properties used:

node.countProjectsperResearcher = PROJECT\_ID.count edge.countProjectsPerDiagnosis=USR\_DIAGNOSIS\_AREA.count

Network Analysis Toolkit (NAT) was selected. This graph claims to be directed.

Nodes: 18

Isolated nodes: 0

Node attributes present: label, countProjectsperResearcher, bipartiteType Edges: 16

No self loops were discovered.

No parallel edges were discovered. Edge attributes:

Did not detect any nonnumeric attributes. Numeric attributes:

|  |  |  |  |
| --- | --- | --- | --- |
|  | min | max | mean |
| countPr... | 1 | 2 | 1.125 |

This network seems to be valued.

Average total degree: 1.7778

Average in degree: 0.8889 Average out degree: 0.8889

This graph is not weakly connected.

There are 3 weakly connected components. (0 isolates) The largest connected component consists of 9 nodes. This graph is not strongly connected.

There are 18 strongly connected components.

The largest strongly connected component consists of 1 nodes. Density (disregarding weights): 0.0523

Additional Densities by Numeric Attribute

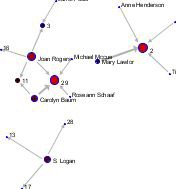
Initial analysis was performed on USR\_PRINCIPAL\_INVESTIGATOR vs USR\_DIAGNOSIS\_AREA

#7. Data analysis/visualization (algorithms) applied and resulting visualizations

Two visualization models were created and analyzed based on Network with directed edges from USR\_PRINCIPAL\_INVESTIGATOR to USR\_DIAGNOSIS\_AREA.2 visualized using

1. GUESS GEM

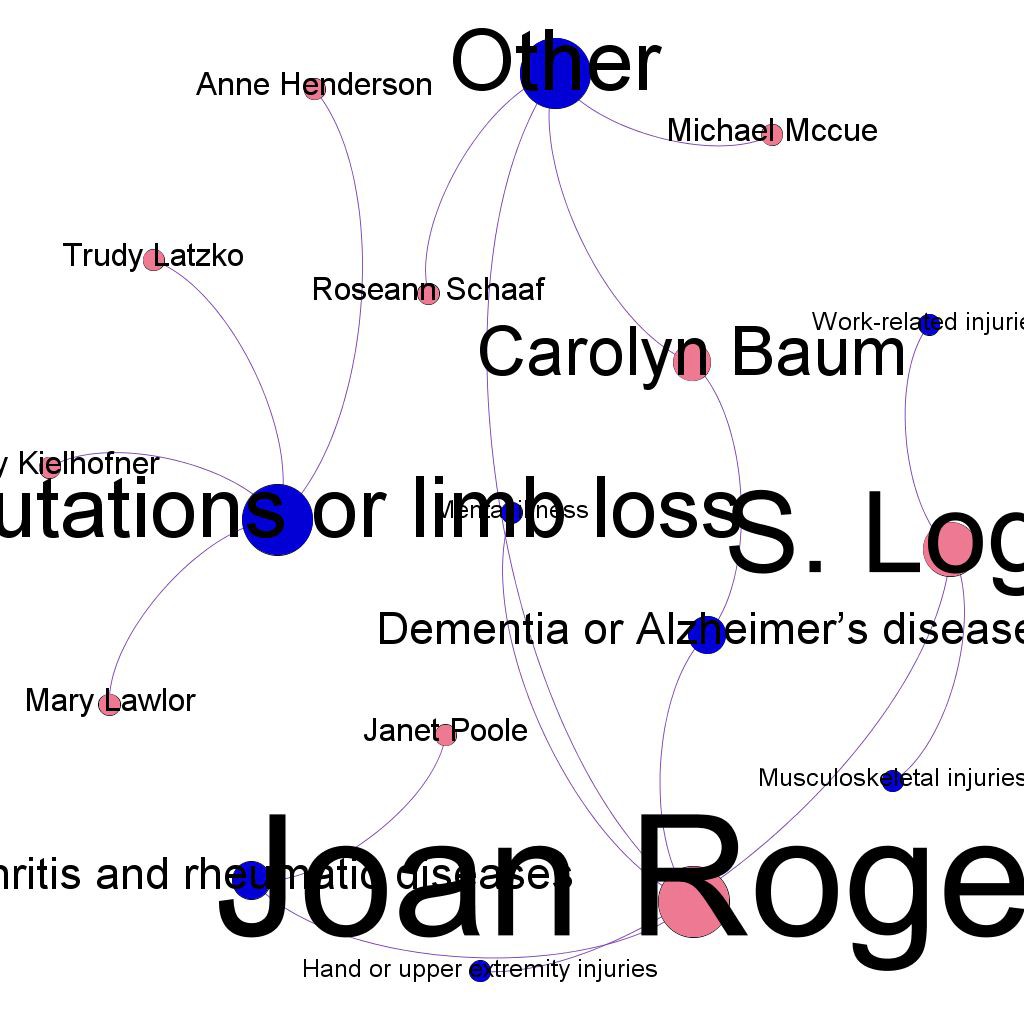
Analyzed [time slice 1982 through 1989](https://iu.instructure.com/files/61364211/download?download_frd=1)



Numbers represent the Diagnosis Area ID(1 through 29). Nodes were proportionally sized based on the number of projects per PI. Edges were proportionally sized by number of projects associated with the Diagnosis area

Encountered limitations of the GUESS visualization: cannot easily distinguish between PIs and Diagnosis area.

1. performed viz in GEPHI using Fruchterman Reingold view



* node color by multimode

­label for Researcher sized by out­degree (number of Diagnosis per researcher)

* node sized by degree (number of projects per Diagnosis)
* edges colored by numberProjects perDiagnosisperInvestigator Preview:default curved

#8. Discussion of key insights gained from the analysis/visualization

For the particular time slice (1982­1989) both types of visualization clearly depicted top players in the OT field being Joan Rogers, S. Logan and Carolyn Baum (nodes and/or labels were sized proportionally to the number of respective projects). In addition, the Diagnosis areas such as 2 ­ Amputations or limb loss, 29­Other, 3­ Arthritis and rheumatic diseases and 11­ Dementia or Alzheimer’s disease were associated with the most number of research projects.

Even though both algorithms correctly represented the data and the insights, the Gephi algorithm was the most flexible in capturing the types of the nodes used in the bipartite network and allowed to color code PI nodes and diagnosis areas that significantly simplified data analysis. The Gephi algorithm will be used to implement further network visualizations.

1. **What problems surfaced during validation and how does your redesign resolve them?**

After initial visualization of PI vs Diagnosis in Gephi we realized that coloring nodes based on the bipartite network only depicts major PIs in the field and helps to identify the core types of the diagnoses. At the same time, the essential linkages between individual nodes (PIs and respective diagnoses) become quite obscure. In order to overcome this issue, we decided to provide a Blondel Community Detection analysis that helped to identify additional connections between various diagnosis and researchers.

In order to connect diagnosis descriptions with agenda we tested out several different visualization types including Fruchterman-Reingold with Annotation and Circular Hierarchy, however the Bipartite Network Graph proved to be the most precise and straightforward. The same type of time slicing on the

Agenda-Diagnosis data helped to compare PI-Diagnosis with the similar time slice of the Agenda-Diagnosis data and to provide more insights into the nature of OT research during respective time slice.

1. **Discussion of challenges and opportunities.**

Due to creation of five time slices per network the main challenge was presentation of significant part of the details without compromising the clarity of the images. Since number of the nodes (number of PIs and diagnoses) increased proportionally with every cumulative time slice, in order to minimize crowding of the nodes the last two slices for PI-Diagnosis were restricted to edges with weight >1. Respective isolated nodes were subsequently removed. In addition, we encountered that Gephi cannot provide persistent coloring of the blondel communities by modifying the color based on the percentage of the specific community size in relation to the other communities per time slice. We will continue investigating possibility to use specific colors per community to ensure their color consistency.

Finally, we will work on identifying links between diagnoses and ICF items and will add respective time slices to the previously created network maps.

[Network­ppt­slides](https://iu.instructure.com/groups/253971/files/folder/First%20Draft?preview=61627308)

# Online Data Explorer using Tableau (lead: Peter Annable)

Goal: The intent of this visualization is to provide a simple way for researchers to explore the information available today about Occupational Therapy Research. This can be used in conjunction with the detailed analyses described above to potentially find other connections in the data, or to simply understand the details of research in a given area.

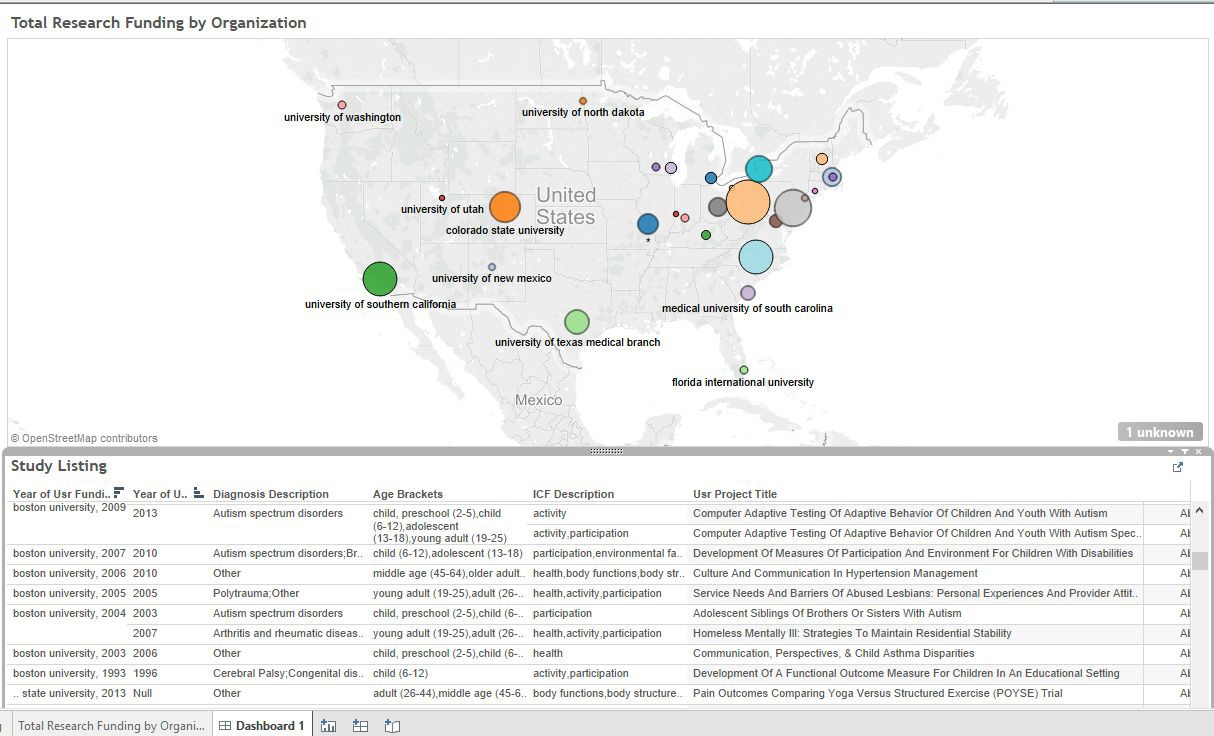
The anticipated visualizations include:

* Use of a tree map and bar graph to understand research concentrations by topic keywords, and ability to click and see what studies make up the included topic.
* Use of a U.S. map to understand research by location
* Research topics organized by amount of funding. Use to understand if certain topics are potentially under represented.
* Note: We are assuming it is permissible to publish these data using Tableau ­ will verify with the Client

Related work: The AOTF Report to Donors from 2013 and 2014 provide some additional context and examples that the data explorer should connect to. There are no existing visualizations to re­apply.

Visaulization Iterations

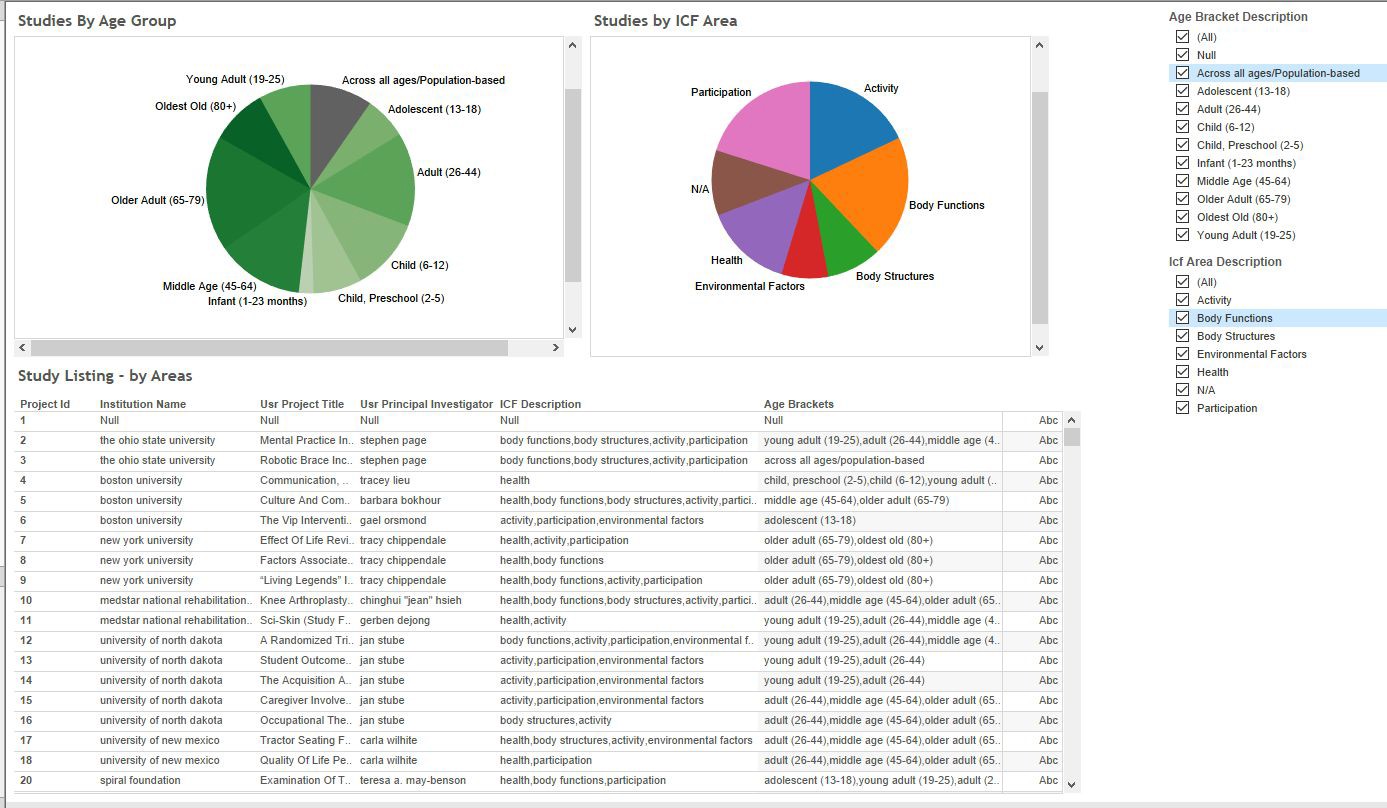
For the first visualization, I will use funding amount to show the total research by each organization. Then allow the user to select from the organization map to see a listing of research projects using Tableau:



### For the 2nd Visualization:

This visual allows the user to select and filter a complete list of studies by age group and/or ICF research area. The pie chart shows the relative percentage of studies for each. To produce these charts, I used two versions of the data: in the study listing chart, the Age Brackets and ICF Descriptions are merged value into one cell,this was already done by Mahesh. I then joined this table with the original Age and ICF values listing, which were provided as separate data tables. This allow me to relate the two together.

Both visuals allow the user to interactively make selections and view resulting data.



1. **What problems surfaced during validation and how does your redesign resolve them?**

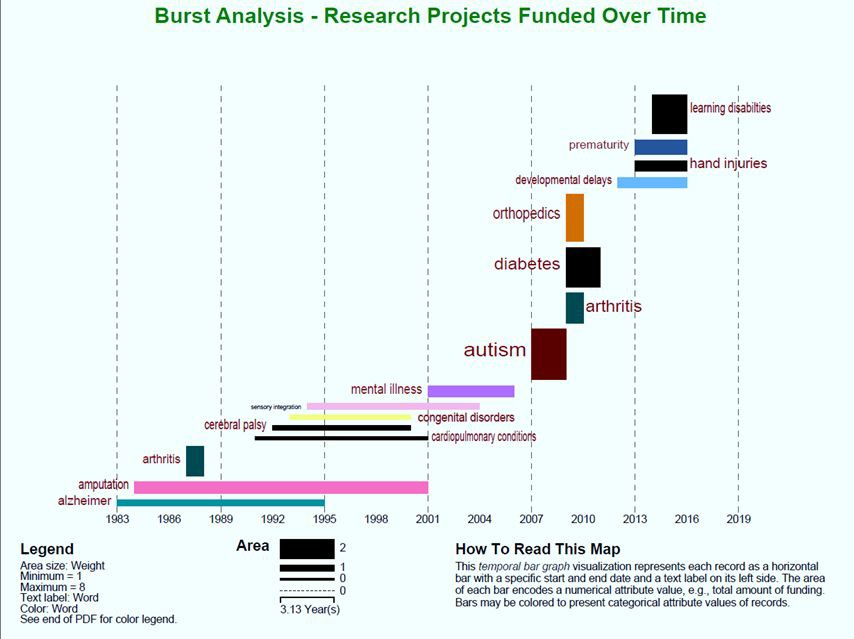
The first issue was getting client approval since the easiest way to share online is via Tableau public. This permission was obtained but with the condition that we not include and funding amounts. Therefore, my data and visuals were updated to use the number of projects to size the proportional symbol map, and not total funding.

The second issue was setting the data up in a way that allowed filtering of multiple data items having multiple values. In our combined data set, we had combined multi-values into one column, which made for easier display of a study details. In order allow for filtering, I added the original mapping tables that mapped each value to its study. Then it took a few attempts to get this relationship set up correctly in Tableau. For example, to preserve all studies and combine with each meta-data item, a LEFT Join is required, not the default INNER join.

1. **Discussion of challenges and opportunities.**

In Tableau it took quite a while to get the layout of my dashboard in a way I felt was easy to use, and made the most efficient use of screen space. This took several hours. Now that I am more fluent in Tableau, I can more quickly add additional options, and will likely do this before submitting the final version.

Longer term, it would be useful to connect the Tableau template directly to the AOTA database, so that it can be used on an ongoing basis to easily examine submitted research data.

A general overall challenge was coordinating data and visual presentation across the team members. For instance, our Temporal analysis uses time slices by decade, and the network visualizations use 7 year time slices. We didn’t realize this issue until it was too late to fix for this submission. This is something we will try to harmonize better for the final presentation. We were able to coordinate a common powerpoint template and color scheme so the final product presentation looks quite professional and should make it easy for our client to understand key i  nsights from our work.