Thesis Project Planning:

**Data collection / cleaning / processing:**

* Get SBIR/STT data from SBIR.gov for 2008-2018
  + Join the list of 1000 record JSON files
  + Geocode recipients with Geocod.io
  + Clean geocod.io outputs
* Pre-process SBIR/STT data using NLPre
  + Edit NLPre to include additional vocabularies, linked data taxonomies for keyword extraction
    - MeSH, Economics and Social Sciences, Technologies, Energy, AGROVOC, SCIVOC
    - Need to extract pure labels from different sources (RDF, XLSX, TTL, NT, etc) and create two column TERM/REPLACEMENT dictionary; run the replace\_from\_dictionary.py function from NLPre and examine where in the basic NLPre pipeline it usually runs (does it handle punctuation or expect it’s been cleaned of punctuation already?—if the former then I can run it iteratively, appending a prefix for each vocab set which will in each case have the same special character; the special character can be used later to reconcile whether a term is tagged with multiple dictionaries and to ensure it’s not double counted in term freq data afterwards)
  + Run scripts to clean up keyword columns
  + Check the Topics Dictionary to see if that is a tenable alternative to doing full keyword extraction pipelines
* Investigate data collection / survey methodology for ACS industry data
* Pre-process industry data
  + Rolling averages, % changes etc.
* Run spatial analysis scripts
  + Identify clusters for ACS industry data for each period
  + Identify clusters for Funding data – perhaps disaggregate for agency / period
* **Outputs from processing and analysis:**
  + Process / code:
    - Github fork of edited NLPre
    - Documentation of Vocabularies used
    - Documentation of variables from datasets used
    - Github repository of processing scripts, python for NLP and R for spatial analysis
  + Data:
    - Gecoded Grants Data
      * Geocoded Grants + Term Document Matrix (limit # of terms and freq measures; use json)
    - SHP file with cluster + ACS stats

**Visualization / Web App**

* Design data flow chart between modules
  + Understand interactivity between charts, id key that will coordinate
  + Include GEOID, State, Region on all datasets
  + Include GRANTID List on Geographic/ACS data
* Design app layout
* Panel 1: Parallel coordinates + Table
  + Add html for axis level summary stats (amount of funding + industry stats)
  + Enable axis hiding on parallel coordinates plot
    - <https://stackoverflow.com/questions/27132902/removing-an-axis-in-d3js-parallel-coordinates>
  + Add search functionality for both location and keywords info
    - https://stackoverflow.com/questions/9463285/slickgrid-searching-multiple-column
  + Edit table to include alternate data other than the data displayed in the table
    - Add full “District Name, State, Region” to make filtering more straightfoward
    - Tab to switch from industry data to auxiliary thematic data? Note: it’s useful to be able to sort on industry…
* Panel 2: Leaflet map
  + Recreate shiny app with Leaflet.js
    - <https://leafletjs.com/examples/choropleth/>
    - UI elements:
      * district/state/region/
      * funding agency
      * cluster variable
      * time range slider
* Panel 3: Keyword data
  + Small Cleveland dot plot with top 5-10 increasing and decreasing terms for region
* Auxiliary 1: Summary Text + stats at state/district/region level
* Auxiliary 2: Summary by agency, $, grants, themes, grantee characteristics
* Print layout generator
  + Allow user to select a district, get the summary funding stats, ACS stats, thematic stats in a ready-to-print PDF
  + <https://github.com/akulmehta/SaveSVG>
  + <https://stackoverflow.com/questions/18191893/generate-pdf-from-html-in-div-using-javascript>
  + <http://pdfkit.org/index.html>
  + <http://pdfmake.org/#/> -- this will probably work well if we can just save the current state of the selections and load them into a file (see: <https://pdfmake.github.io/docs/0.1/> )
    - <https://stackoverflow.com/questions/39243149/how-to-handle-svg-and-canvas-in-pdfmake-html2pdfmake>
    - <https://www.npmjs.com/package/pdfmake-wrapper>
    - <https://github.com/bpampuch/pdfmake/blob/0.1/examples/pdfs/svgs.pdf>

**Analysis**

* Brief background on economic geography + scientometric research
* Perform some basic analysis including clusters + funds + themes
* Include it as an About/Analysis page

**Outputs**

* Whitepaper
* Github
* Analysis + Interactive Viz