ORNL/TM-2017/727

Enhancing Clean Energy Innovation Ecosystem Discovery Tool

Supriya Chinthavali, Sangkeun Lee, Chelsey Dunivan Stahl, Navina Nageswararao

October 2017

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ORNL/TM-2017/727

Computer Science and Mathematics Division

Enhancing Ecosystem Discovery: Measuring Clean Energy Innovation Ecosystems Through Knowledge Discovery and Mapping Techniques

Authors:
Supriya Chinthavali
Sangkeun Lee
Chelsey Dunivan Stahl

Date Published: October 20th, 2017

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee 37831-6283
managed by
UT-BATTELLE, LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

AMO	US DOE's Advanced Manufacturing Office
ARPA-E	Advanced Research Projects Agency - Energy
CBSA	Core-Based Statistical Areas
CEIC	Clean Energy Investment Center
DoD	Department of Defense
EC	Ecosystem Component
EERE	US DOE’s Office of Energy Efficiency and Renewable Energy
EPSA	US DOE’s Office of Energy Policy and System Analysis
IE	Innovation Ecosystem
ERC	Energy Research Centres
MSA	Metropolitan Statistical Areas and Micropolitan Statistical Areas
NAIC	North American Industry Classification
NASA	National Aeronautics and Space Administration
CBSA	Core Based Statistical Areas
DSIRE	Database for Renewable Energy
NLP	Natural Language Processing
NNMI	National Network for Manufacturing Innovation
NSF	National Science Foundation
ORNL	Oak Ridge National Lab
OE	US DOE’s Office of Electricity
FE	Fossil Energy
PV	Photovoltaic
QTR	Quadrennial Technology Review
NE	Nuclear Energy
R&D	Research and development
SBA	Small Business Administration
SBIR	Small Business Innovation Research
STTR	Small Business Technology Transfer
US DOE	US Department of Energy
USPTO	US Patent and Trademark Office

ACKNOWLEDGMENTS

The report was prepared by a small team of data scientists from the Computational Data Analytics group from Oak Ridge National Laboratory. This work was financed by the Office of Program and Innovation Policy Analysis of the Energy Policy and System Analysis Office (EPSA-52) of the US Department of Energy (DOE).

It has been reviewed by Robert Horner (DOE-EPSA). Thanks go to Eric Hsieh (DOE-EPSA) for his conceptualization of this project and to John Jennings (DOE-EPSA) for his leadership and support. Special thanks go to DOE staff for their help in gathering data: Joe Paladino and Debbie Haught from the Office of Electricity Delivery and Energy Reliability, Christine Herbert and Leslie Pezzullo from the Office of Energy Efficiency and Renewable Energy, Andrew Hlasko from the Office of Fossil Energy, and Eddie Christy and Michael Knaggs from the National Energy Technology Laboratory.

Legal Disclaimer

The findings, interpretations and conclusions herein are those of the author(s). They do not necessarily reflect the views of the US DOE or ORNL. The US DOE does not guarantee the accuracy of the data included in this work and accepts no responsibility whatsoever for any consequence of their use.

EXECUTIVE SUMMARY

During March 2016, the DOE's Office of Energy Policy and Systems Analysis (EPSA) sought a methodology to identify and quantify the strength of the existing clean energy innovation ecosystems (IE) in the U.S. and their characteristics. Oak Ridge National Lab (ORNL) did a pilot study (Phase 1) and demonstrated the feasibility of an application comprised of natural language processing, link analysis, and other computational techniques to transform text and numerical data into metrics on clean energy innovation activity and geography. The data-collection, ingest and analytics pipelines were combined with an advanced user interface, together known as the Ecosystem Discovery Tool, to enhance DOE's understanding of existing geographic innovation clusters. During Phase 2 of this project, this tool has been further enhanced by integrating new data sets and through backend software architecture changes, validation and bug fixing, making it a much more robust and powerful application. The tool's user interface is a lot more intuitive, which enables a user to visualize the IE rankings for various geographic regions(CBSA, state level) for different clean energy technologies and seek automated insights. ORNL also created a DOE-specific dashboard that allows the user to visualize and analyze federal funding from various offices including Energy Efficiency and Renewable Energy(EERE), Office of Electricity(OE), Fossil Energy(FE), and Nuclear Energy(NE).

During Phase 1 of this project, EPSA defined a clean energy innovation ecosystem as the overlap of five Ecosystem Components: 1) nascent clean energy Indicators, 2) investors, 3) enabling environment, 4) networking assets and 5) large companies. EPSA and ORNL worked together to collect data for each component: 1) small and medium companies, ARPA-E awardees, SBIR awardees, patents, publications, ERCs for Nascent Clean Energy Indicators; 2) qualified investors; 3) the Clean Edge Policy Index/DSIRE for the Enabling Environment; 4) universities, national laboratories, ERCs and incubators for Networking Assets; and 5) large companies and a subset of the Russell 1000 list for Large Companies.

The ORNL team created a visual tool based on Tableau that integrated ecosystem component data to score, rank and map IEs. The tool was created with the flexibility to allow the user to choose the weights of each of the five ecosystem components and the subcomponents. This flexibility allows the user to visualize different subsets and to use the underlying data for different types of analysis. During Phase 2 of this project, the backend database and computation process have undergone major changes to provide a much more refined user-interface and added new features into the scoring algorithm. This phase also involved addition of newer databases such as a list of Energy Research Centers (ERC) ,DSIRE, and funding data from EERE, FE, NE, and OE to provide more granular information and more accurate/better quantification.

1. INTRODUCTION

1.1. METHODOLOGY OVERVIEW

An integrated database was created in Phase 1, with a schema described in Table 1 using 14 different data sources. However, all the features and numerical metrics that were computed after aggregation of the data for scoring were all implemented within Tableau [2].

Figure 1. Data analytics workflow in Phase 1

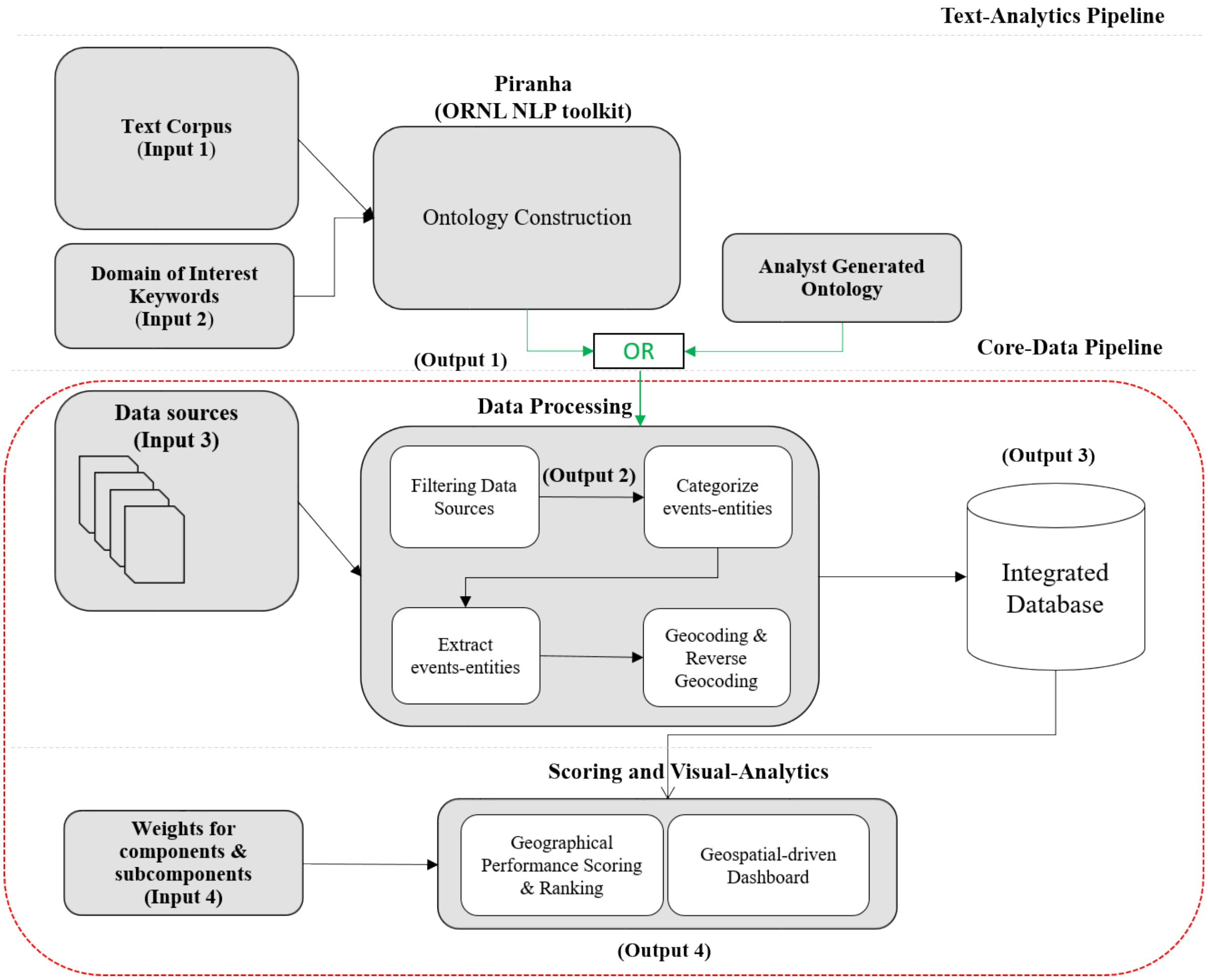

Text-Analytics Pipeline
Text Corpus (Input 1)
Domain of Interest Keywords (Input 2)
Piranha (ORNL NLP toolkit)
Ontology Construction
Analyst Generated Ontology
(Output 1)
OR
Core-Data Pipeline
Data sources (Input 3)
Data Processing
Filtering Data Sources
(Output 2)
Categorize events-entities
Extract events-entities
Geocoding & Reverse Geocoding
(Output 3)
Integrated Database
Scoring and Visual-Analytics
Weights for components & subcomponents (Input 4)
Geographical Performance Scoring & Ranking
Geospatial-driven Dashboard
(Output 4)

Figure 2. Overview of data analytics workflow and tasks in Phase 2 of the project

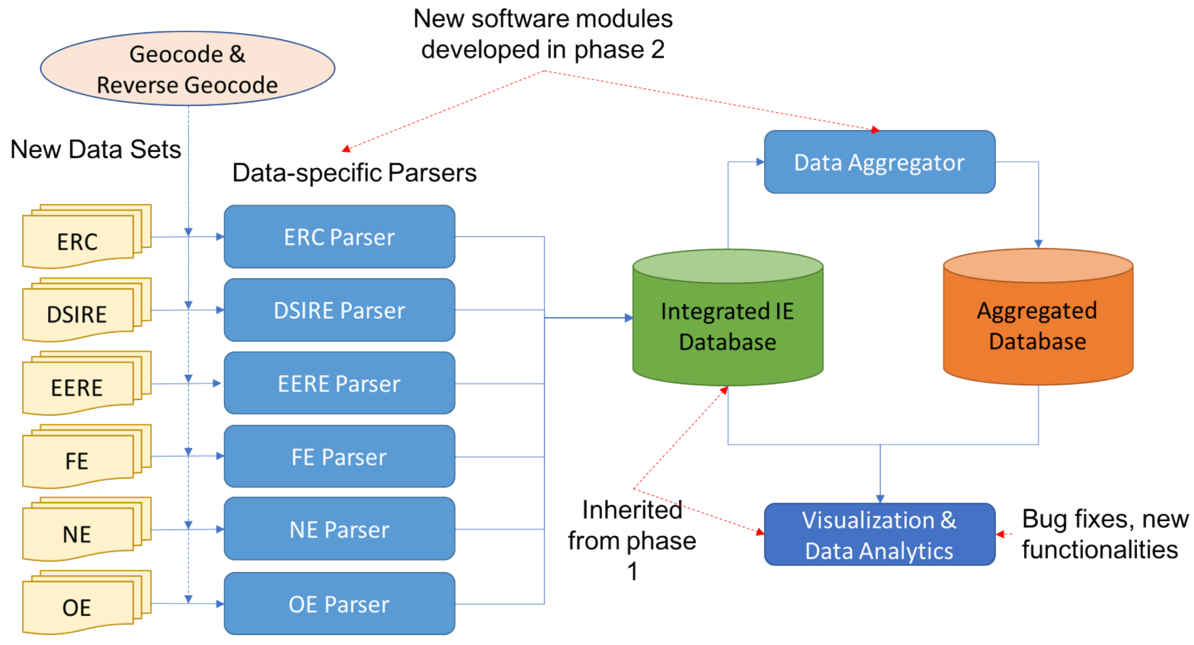


Figure 2 shows a high-level view of the data analytics workflow. In Phase 2 of this project, we incorporated 6 new data sets (ERC, DSIRE, EERE, FE, NE, and OE databases). Similar to the process flow in Phase 1, entities in each dataset needed to be geocoded (i.e., finding latitude and longitude) then reverse geocoded (i.e., finding CBSA codes, states, counties), unless the dataset already had all required geographical information for the entities before ingesting them into custom parsers.

(Task1a/1b) In Phase 2, we developed new software modules to add new data sets into the database. Because every dataset had its own schema and characteristics, we first implemented a data-specific parser for each new dataset that transformed the data into the format suitable for the Integrated IE Database that was inherited from Phase 1. The transformed data was then ingested into the Integrated IE Database. We describe each data set and explain the transformation rules in section 2.1. Second, we implemented a separate Data Aggregator module that uses the integrated IE database as its input and produces an aggregated database. In Phase 1, data aggregation was done by the module internally embedded in the data analytics and visualization dashboard composed using Tableau; however, due to the limited functionalities in terms of aggregation, we separated the module from the visualization dashboard. The new data aggregator is implemented in Python, which is a general-purpose programming language, so it allows much more flexibility and extensibility. The Data Aggregator module is described in section 2.2.

(Task2a/2b) In section 0, we describe the bug fixes and new functionalities added in Phase 2. The bugs fixed were primarily identified in Phase 1 and some bugs were reported with Phase 2 as well after the software architecture changes. The new functionalities mainly include:

- Selection of the enabling environment data source (DSIRE vs CleanEdge Policy Index)

- DOE projects Interactive Dashboard displaying ranking of all CBSA’s using funding information for various technologies
- Incorporating “Number of DOE projects” and “Number of ERC projects” as one of the features for scoring Innovation Ecosystems within the main tool.
- Scoring Algorithm Evaluation for various clean energy technologies

2. Task 1: Incorporating New data sources

In this section, we explain how we incorporated 6 new data sources into the integrated IE database inherited from Phase 1 of the project. The following shows the list of column names in the Integrated IE database.

Table 1. List of attributes in the Integrated IE database.

- EventType
- Name
- Timestamp
- Primary Sector
- Zip CBSA-New
- CBSA
- CBSA Name
- Type
- Population
- In/Out
- Ticker Symbol
- Tags
- Short Description
- Website URL Contact Phone
- Address City
- State
- State for Policy Index
- Country Region
- Institutional Type/Entity Type
- Development Stage
- # of Employees
- Management Team Key Competitors Overview
- Products/Technology Strategy
- Total paid in Capital ($)
- Seeking Funding Seeking Funding Amount ($)
- Seeking Funding Date
- Revenue Range
- Revenue Range Source
- Date Submitted
- SubTechnology
- Participants
- Total Incentives
- CleanTech Policy Index
- Latitude
- Longitude
- Status

- Federal Funding Amount
- Publication_Count
- NewsFeed_Counts Market_Value
- CPC_Group_ID

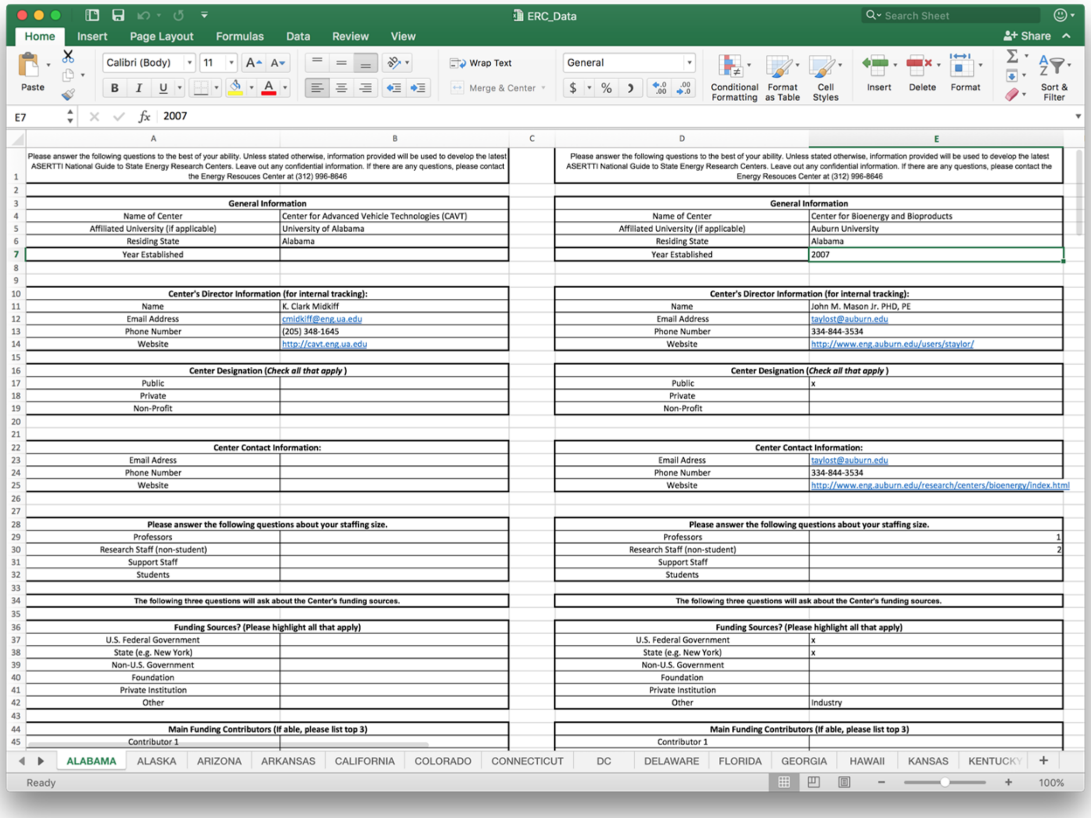
In the following sections, we briefly explain what each new data source is and what information is included in it and how we map attributes of the data source to the attributes of the integrated IE database.

2.1. TASK1A: IMPLEMENTING DATA-SPECIFIC PARSERS

2.1.1. ERC (Energy Research Centers) Database

Since 2012, the energy industry has grown and evolved to an extent that nearly every state contains at least one energy research center, and the number of research centers has grown from 130 to over 200. A report was prepared by the Energy Resources Center at the University of Illinois at Chicago and was conducted under a contract with Argonne National Laboratory. The report provides detailed descriptions of research activities as well as contact information to aid in collaborative efforts of about 213 energy centers. In Phase 2 of this project, we incorporated this list of Energy Research Centers available in a PDF and a spreadsheet format (i.e., xls). The xls format was such that there were 50 sheets corresponding to the US states and within each sheet there are multiple tables populated as key value pairs.

Figure 3. Sample of ERC data set



E7 *fx* 2007

	A	B	C	D	E
1	Please answer the following questions to the best of your ability. Unless stated otherwise, information provided will be used to develop the latest ASERTTI National Guide to State Energy Research Centers. Leave out any confidential information. If there are any questions, please contact the Energy Resouces Center at (312) 996-8646			Please answer the following questions to the best of your ability. Unless stated otherwise, information provided will be used to develop the latest ASERTTI National Guide to State Energy Research Centers. Leave out any confidential information. If there are any questions, please contact the Energy Resouces Center at (312) 996-8646	
2					
3	**General Information**			**General Information**	
4	Name of Center	Center for Advanced Vehicle Technologies (CAVT)		Name of Center	Center for Bioenergy and Bioproducts
5	Affiliated University (if applicable)	University of Alabama		Affiliated University (if applicable)	Auburn University
6	Residing State	Alabama		Residing State	Alabama
7	Year Established			Year Established	2007
8					
9					
10	**Center's Director Information (for internal tracking):**			**Center's Director Information (for internal tracking):**	
11	Name	K. Clark Midkiff		Name	John M. Mason Jr. PHD, PE
12	Email Address	cmidkiff@eng.ua.edu		Email Address	taylost@auburn.edu
13	Phone Number	(205) 348-1645		Phone Number	334-844-3534
14	Website	http://cavt.eng.ua.edu		Website	http://www.eng.auburn.edu/users/staylor/
15					
16	**Center Designation (*Check all that apply*)**			**Center Designation (*Check all that apply*)**	
17	Public			Public	x
18	Private			Private	
19	Non-Profit			Non-Profit	
20					
21					
22	**Center Contact Information:**			**Center Contact Information:**	
23	Email Adress			Email Adress	taylost@auburn.edu
24	Phone Number			Phone Number	334-844-3534
25	Website			Website	http://www.eng.auburn.edu/research/centers/bioenergy/index.html
26					
27					
28	**Please answer the following questions about your staffing size.**			**Please answer the following questions about your staffing size.**	
29	Professors			Professors	1
30	Research Staff (non-student)			Research Staff (non-student)	2
31	Support Staff			Support Staff	
32	Students			Students	
33					
34	**The following three questions will ask about the Center's funding sources.**			**The following three questions will ask about the Center's funding sources.**	
35					
36	**Funding Sources? (Please highlight all that apply)**			**Funding Sources? (Please highlight all that apply)**	
37	U.S. Federal Government			U.S. Federal Government	x
38	State (e.g. New York)			State (e.g. New York)	x
39	Non-U.S. Government			Non-U.S. Government	
40	Foundation			Foundation	
41	Private Institution			Private Institution	
42	Other			Other	Industry
43					
44	**Main Funding Contributors (If able, please list top 3)**			**Main Funding Contributors (If able, please list top 3)**	
45	Contributor 1			Contributor 1	

Data preprocessing

The data-specific parser for this data set first transforms the data into a tabular format having the following columns.

Table 2. List of attributes in the transformed ERC database

- STATE
- Name of Center
- Affiliated University (if applicable)
- Residing State
- Year Established
- State Abbr
- Name
- Email Address
- Phone Number
- Website
- Public
- Private
- Non-Profit
- Email Address
- Phone Number
- Website
- Professors
- Research Staff (non-student)
- Support Staff
- Students
- U.S. Federal Government
- State (e.g. New York)
- Non-U.S. Government
- Foundation
- Private Institution
- Other
- Contributor 1
- Contributor 2
- Contributor 3
- Yes
- No
- Collaborator 1
- Collaborator 2
- Collaborator 3
- Collaborator 4
- Collaborator 5
- Perform R&D in--house
- Issue R&D Grants/Contracts

- Own Intellectual Property (IP)Portfolio
- Perform Demonstrations
- Spin-off Companies
- Set/Verify Standards
- Make Equity Investments
- Provide Analysis and/or Data
- Provide Shared Facilities
- Provide Education & Training
- Provide Project Management / Professional Services
- Provide Preferential Access to Research
- Advanced Electronics
- Bioenergy and Biofuels
- Carbon Capture and Sequestration
- Climate Research
- Economic Modeling and Analysis
- Energy Efficiency and Sustainable Building Design
- Energy Storage and Fuel Cells
- Environmental and Emissions Technologies
- Manufacturing
- Nuclear power
- Fossil Fuels and Advanced Plant Technologies
- Policy
- Solar Wind Geothermal Hydropower and Marine and Hydrokinetic Power
- Smart Grid and Transmission and Distribution
- Materials
- Transportation Technology
- Water Technology and Water Use Efficiency
- Other
- Yes
- No
- Total Annual Funding for the Center (Approximate):
- Please describe the model of your collaboration
- What is your center's mission statement? (375 characters)
- Please describe some notable projects or research assignments that you would like to be highlighted. (1300 characters)

Geocoding & Reverse Geocoding

In order to use this data in the tool, we first needed to obtain location data. We were able to get coordinate data for each entry in this dataset using the *Affiliated University* attribute. We began by downloading a comprehensive list of accredited universities and their addresses from the U.S. Department of Education [3]. Next, we matched the *Affiliated University* name to its address using the VLOOKUP function in excel. Once we had obtained an address for each row, we used the Google Geocoding API [4] to get the latitude and longitude.

To get the CBSA data, we used a custom script developed in Phase 1 that takes as input a list of zip codes and outputs the associated CBSA data for each entry in the list. Using the zip codes we obtained in the previous step, we were able to get both the CBSA and the CBSA Name.

At the conclusion of this step, we had added four additional columns (CBSA, CBSA NAME, lat, long).

Data Transformation

The data-specific parser for ERC data transforms each row of the preprocessed & geocoded data set into 1 or more data entries for the integrated IE database. The following summarizes the transformation mapping.

Table 3. Basic information mapping for ERC dataset

- EventType ←”ERC Center”
- Name ← Name of Center
- CBSA←CBSA
- CBSA Name←CBSA NAME
- Latitude←lat
- Longitude←long

The ERC database also captured information about the main research activities which needed to be utilized to assign the appropriate ecosystem category (nascent clean tech, investors, networking etc.). Depending on what major research activities the ERC are contributing towards, we account them as entities for scoring various ecosystem components as follows.

Table 4. Ecosystem Component Category Mapping for ERCs (Achieving values for Short Description column)

- (If value of the column) Perform R&D in--house (is YES) ⇒Nascent CT
- Issue R&D Grants/Contracts ⇒Nascent CT, Investors
- Own Intellectual Property (IP)Portfolio ⇒Nascent CT
- Perform Demonstrations⇒Nascent CT
- Spin-off Companies⇒Nascent CT
- Set/Verify Standards / Policy?⇒Enabling Environment
- Make Equity Investments⇒Investors
- Provide Analysis and/or Data⇒Nascent
- Provide Shared Facilities⇒Networking
- Provide Education & Training⇒Networking
- Provide Project Management / Professional Services⇒Networking
- Provide Preferential Access to Research⇒Networking

Each data entry in ERC data set is to be mapped to 1 or more technology categories as shown below. If an ERC does research on multiple clean energy technologies, then each ERC data entry has been duplicated in the integrated database, one entry for each technology that it performs research on. This allows the technology filter to work on the Events map as well(the drill down view). However, note that this leads to heavy weighting of the ERCs that support multiple technologies while scoring, when multiple technologies are selected.

If, for example, an ERC focuses on wind, solar and geothermal research activities, selection of wind, solar and geothermal using the Technology filter will take into account the ERC as count 3, although it is a single center.

Table 5. Clean Energy Technology Mapping for ERCs (Achieving values for Primary Sector column)

- (If value of the column) Advanced Electronics (is YES) ⇒Others
- Bioenergy and Biofuels⇒Biopower
- Carbon Capture and Sequestration⇒Carbon Capture and Storage
- Climate Research⇒Others
- Economic Modeling and Analysis⇒Others
- Energy Efficiency and Sustainable Building Design⇒Energy Efficiency
- Energy Storage and Fuel Cells⇒Fuel Cells, Energy Storage
- Environmental and Emissions Technologies⇒Others
- Manufacturing⇒Others
- Nuclear power⇒Nuclear
- Fossil Fuels and Advanced Plant Technologies⇒Advanced Plant Technologies
- Policy⇒Others
- Solar Wind Geothermal Hydropower and Marine and Hydrokinetic Power⇒Solar, Wind, Geothermal, Hydropower,Marine and Hydrokinetic Power
- Smart Grid and Transmission and Distribution⇒Smart Grid
- Materials⇒Others
- Transportation Technology⇒Advanced Clean Transportation and Vehicle System
- Water Technology and Water Use Efficiency⇒Others
- Other⇒Others

2.1.2. DSIRE (Database of State Incentives for Renewables & Efficiency)

In order to help citizens find financial incentive programs offered by US governments for clean energy technologies and solutions, the U.S. Department of Energy established the Database of State Incentives for Renewables & Efficiency (DSIRE) in 1995. The database is operated by the N.C. Clean Energy Technology Center at North Carolina State University and funded by the U.S. Department of Energy. The DSIRE website provides various ways to access the database including a search tool, dynamic maps, charts, and tables. The website also provides an Application Program Interface (API) freely available for download that contains all of the data on DSIRE in an easy-to-read format such as CSV or JSON. [5]

We began the process of incorporating DSIRE data by first downloading the data using the API. We then developed code that iterated through the dataset and grouped the policies by category and state. This was accomplished using a bucketing system in which we divided the programs by technology category and then by state. The results were then fed into another function that processed every CBSA in the country and assigned it the appropriate policy counts for each category. The number of policies for each state/category pair was calculated by adding the number of policies for the state the CBSA is located in plus the number of all Federal policies. This data was then output to a CSV.

The following shows the list of attributes in the CSV file generated by pre-processing the DSIRE database, which we used as an input file for DSIRE parser.

Table 6. List of attributes in the preprocessed DSIRE database

- CBSA
- Technology
- State
- Number of Policies

- Policy Names

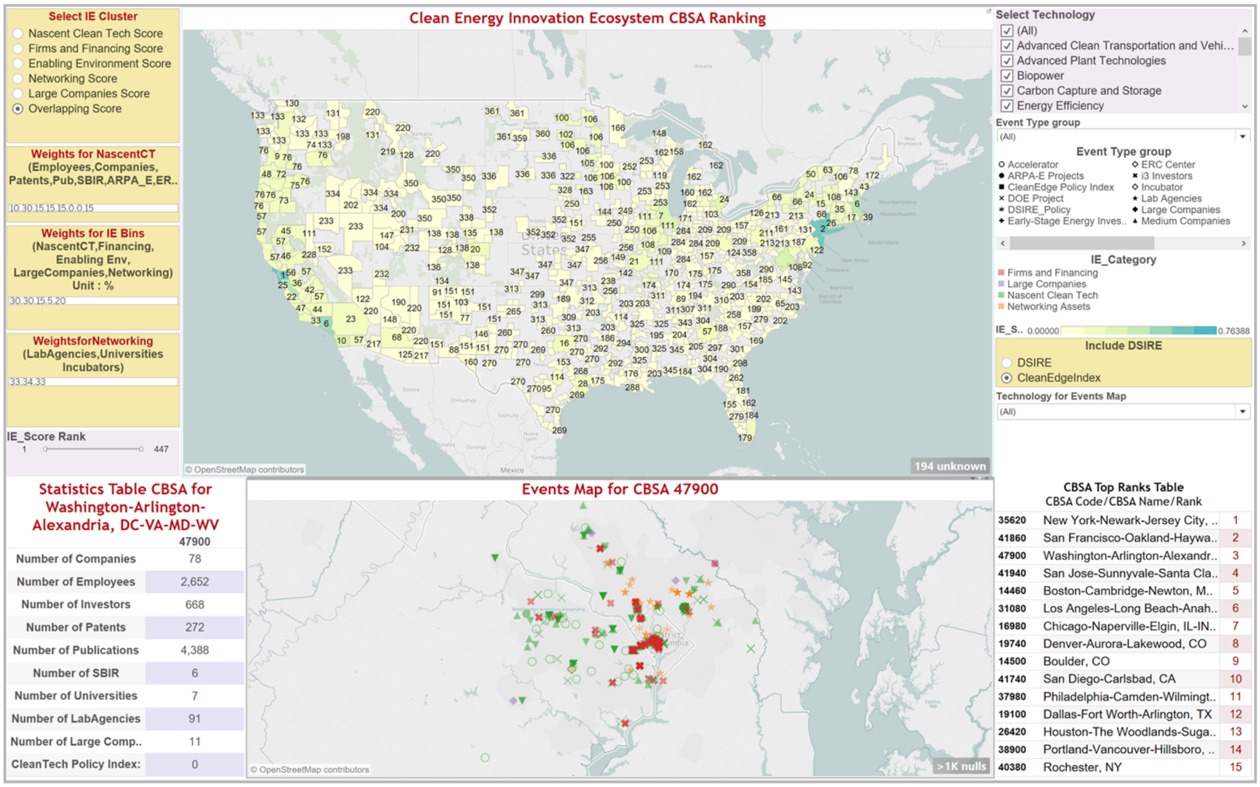
Next, The DSIRE parser transforms this file for IE database. The following summarizes the transformation mapping.

Table 7. Basic information mapping for DSIRE dataset

- EventType ←"DSIRE_Policy"
- Name ← Policy Names
- Primary Sector ←Technology
- CBSA←CBSA
- State←State
- CleanTech Policy Index←Number of Policies

The IE CBSA dashboard of the main tool, now has an option of ranking the CBSA using either the Clean Edge Policy Index for the state or the DSIRE data(Figure below).

Figure 4. Clean Energy Innovation Ecosystem CBSA Ranking



Select IE Cluster

○ Nascent Clean Tech Score
○ Firms and Financing Score
○ Enabling Environment Score
○ Networking Score
○ Large Companies Score
◉ Overlapping Score

Weights for NascentCT (Employees,Companies, Patents,Pub,SBIR,ARPA_E,ER..: 10.30.15.15.15.0.0.15

Weights for IE Bins (NascentCT,Financing, Enabling Env, LargeCompanies,Networking) Unit : %: 30.30.15.5.20

WeightsforNetworking (LabAgencies,Universities Incubators): 33.34.33

IE_Score Rank: 1 – 447

Clean Energy Innovation Ecosystem CBSA Ranking

Select Technology

☑ (All)
☑ Advanced Clean Transportation and Vehi...
☑ Advanced Plant Technologies
☑ Biopower
☑ Carbon Capture and Storage
☑ Energy Efficiency

Event Type group: (All)

Event Type group

○ Accelerator	◇ ERC Center
● ARPA-E Projects	✖ i3 Investors
■ CleanEdge Policy Index	◇ Incubator
× DOE Project	★ Lab Agencies
✱ DSIRE_Policy	◆ Large Companies
✚ Early-Stage Energy Inves..	▲ Medium Companies

IE_Category

- 🟥 Firms and Financing
- 🟪 Large Companies
- 🟩 Nascent Clean Tech
- 🟧 Networking Assets

IE_S.. 0.00000 – 0.76388

Include DSIRE

○ DSIRE
◉ CleanEdgeIndex

Technology for Events Map: (All)

Statistics Table CBSA for Washington-Arlington-Alexandria, DC-VA-MD-WV

	47900
Number of Companies	78
Number of Employees	2,652
Number of Investors	668
Number of Patents	272
Number of Publications	4,388
Number of SBIR	6
Number of Universities	7
Number of LabAgencies	91
Number of Large Comp..	11
CleanTech Policy Index:	0

Events Map for CBSA 47900

CBSA Top Ranks Table
CBSA Code/CBSA Name/Rank

35620	New York-Newark-Jersey City, ..	1
41860	San Francisco-Oakland-Haywa..	2
47900	Washington-Arlington-Alexandr..	3
41940	San Jose-Sunnyvale-Santa Cla..	4
14460	Boston-Cambridge-Newton, M..	5
31080	Los Angeles-Long Beach-Anah..	6
16980	Chicago-Naperville-Elgin, IL-IN..	7
19740	Denver-Aurora-Lakewood, CO	8
14500	Boulder, CO	9
41740	San Diego-Carlsbad, CA	10
37980	Philadelphia-Camden-Wilmingt..	11
19100	Dallas-Fort Worth-Arlington, TX	12
26420	Houston-The Woodlands-Suga..	13
38900	Portland-Vancouver-Hillsboro, ..	14
40380	Rochester, NY	15

2.1.3. EERE (Energy Efficiency & Renewable Energy) DOE Project Database

The Office of Energy Efficiency and Renewable Energy (EERE) provided 3 separate excel spreadsheets that contain lists of DOE projects funded by the office. To pre-process these files, we converted them into the CSV file format, then performed geocoding and CBSA mapping for each entry.

We obtained the coordinate data by passing the *Street Address, City, State* and *Zip* attributes into the Google Geocoding API to get the latitude and longitude. The *Zip* attribute was also used as input for the CBSA mapping script developed in Phase 1 to get the CBSA information.

At the conclusion of this step, we had added four additional columns (lat, long, CBSA, and ST). We show the attributes of each preprocessed file in the following.

File 1 includes a list of DOE projects funded by EERE.

Table 8. List of attributes in the preprocessed EERE data - File 1(EERE_activities.csv)

- TECHNOLOGY OFFICE ABBRV
- CID
- CONTRACT/ AWARD DATE
- PROJECT NAME
- VENDOR NAME
- STREET ADDRESS
- STREET ADDRESS 2
- CITY
- STATE
- ZIP
- ITD OBLIGATIONS
- CY OBLIGATIONS
- YTD UNCOSTED OBS
- INSTITUTION TYPE
- DEVELOPMENT STAGE/TRL
- MATCHING FUNDS
- NOTES
- Lat
- Long
- CBSA
- ST

File 2 includes list of DOE projects that focus on Wind technology funded by EERE. Note that the projects included in this file can also exist in File 1, so we had to take care of the duplicates to avoid including multiple entities representing the same project in the IE database. CID was used to identify the duplicated entries.

Table 9. List of attributes in the preprocessed EERE data - File 2(Mater Wind Projects Database.csv)

- CID
- Project_Title
- Awardee
- Program_Area
- DOE_Funding_Amount
- Recipient_Type
- Award_Type,State
- FOA_Name
- FOA_NodeID

- Fiscal_Year_Awarded
- Subprogram_Node
- Project_Description
- Status
- Contact_email
- Contact_url
- Lat
- Long
- CBSA
- ST

File 3 includes list of DOE projects that focus on Water technology funded by EERE. Note that the projects included in this file can also exist in the File 1, so we had to take care of the duplicates to avoid including multiple entities representing the same project in the IE database.

Table 10. List of attributes in the preprocessed EERE data - File 3(water_projects_data_2017_01_17_3 - forRSKS.csv)

- CID
- Award_Type
- FOA_Name
- FOA_NodeID
- Fiscal_Year_Awarded
- Project_Title
- Awardee
- Recipient_Type
- Program_Area
- Subprogram_Node
- State
- DOE_Funding_Amount
- Project_Description
- Status
- Principal_Investigator
- Lat
- Long
- CBSA
- ST

Next, The EERE parser transformed the three files for IE database. The following summarizes the transformation mappings.

Table 11. Basic information mapping for File 1 to IE database

- EventType ←"DOE Project"
- Name ← PROJECT NAME
- State ←State
- Timestamp ← parse_year(CONTRACT/ AWARD DATE)
- CBSA←CBSA

- Latitude←Lat
- Longitude←Long
- Total paid in Capital ($)←max(parse_fund_str(ITD OBLIGATIONS), clean_fund_str(CY OBLIGATIONS))
- Primary Sectors←mapping_office(TECHNOLOGY OFFICE ABBRV)

The functions used in Table 11 are described in the followings.

- parse_year(year_str) returns a string formatted YYYY parsing the given *year_str*
- parse_fund_str(fund_str) returns a string after getting rid of '$', ',' in the string *fund_str*.
- max(n1,n2) returns a larger value between n1 and n2
- mapping_office(office_name) returns a primary sector value based on the following mapping

Table 12. Mapping of office name to primary sector

Office name	**Return value**
AMO	Advanced Plant Technologies
BETO	Biomass
BTO	Energy Efficiency
FCTO	Fuel Cells
GTO	Geothermal
SETO	Solar
VTO	Transportation
WWPTO-Water	Hydropower
WWPTO-Wind	Wind Power
WIPO, OSP, F&I, FEMP, others	*N/A (ignore this data entry and do not include into IE database)*

Table 13. Basic information mapping for File 2 to IE database

- EventType ←"DOE Project"
- Name ← Project_Title
- State ←ST
- Timestamp ← parse_year(Fiscal_Year_Awarded)
- CBSA←CBSA
- Latitude←Lat
- Longitude←Long
- Total paid in Capital ($)←parse_fund_str(DOE_Funding_Amount)
- Primary Sectors←'Wind Power'

The functions used in Table 13 are described in the followings.

- parse_year(year_str) returns a string formatted YYYY parsing the given *year_str*
- parse_fund_str(fund_str) returns a string after getting rid of ‘$’, ‘,’ in the string *fund_str*.

Table 14. Basic information mapping for File 3 to IE database

- EventType ←”DOE Project”
- Name ← Project_Title
- State ←ST
- Timestamp ← parse_year(Fiscal_Year_Awarded)
- CBSA←CBSA
- Latitude←Lat
- Longitude←Long
- Total paid in Capital ($)←parse_fund_str(DOE_Funding_Amount)
- Primary Sectors←HydroPower’

The functions used in Table 14 are described in the followings.

- parse_year(year_str) returns a string formatted YYYY parsing the given *year_str*
- parse_fund_str(fund_str) returns a string after getting rid of ‘$’, ‘,’ in the string *fund_str*.

In order to identify duplicate entries across files, we used the CID as an identifier. However, we modified CID values in the File 1 to match the CID format in Files 2 and 3. (In File 1, CID values starts with “DE-”, but not in Files 2 and 3). All non-duplicated entries are automatically processed by EERE but for the duplicated entries, in order to keep the accurate information from multiple files as much as possible, which is inevitably an ad-hoc process, we manually processed the data.

2.1.4. FE(FOSSIL ENERGY DATABASE)

The Office of Fossil Energy (FE) provided an excel spreadsheet that contains a list of DOE projects funded by the office. To pre-process these files, we converted them into the CSV file format, then geocoded & reverse geocoded and performed the CBSA mapping.

We used the *Performer City* and *Performer State* attributes to collect the coordinate information by passing them into the Google Geocoding API. In order to perform the CBSA mapping, we had to obtain zip code data for each entry. To do so, we ran the coordinates in the previous step through the Google Reverse Geocoding API. Once we had the zip code data we were able to run it through the CBSA mapping script.

At the conclusion of this step, we had added three additional columns (lat, long, and CBSA). We show the attributes of each preprocessed file in the following.

Table 15. List of attributes in the preprocessed FE data

- Agreement Number
- Cost Plan DOE Share
- Cost Plan Performer Share
- Cost Plan Total Award Value
- Directorate
- Performer
- Performer City

- Performer State
- Prime/Sub Indicator
- Program Area
- Project Status
- Project Title
- Subprogram
- Technology Area
- Year of Completion Date
- Year of Start Date
- Latitude (generated)
- Longitude (generated)
- CBSA

The preprocessed file is transformed into the format that can be ingested into the IE database as follows.

Table 16. Basic information mapping for FE dataset to IE database

- EventType ←”DOE Project”
- Name ← Project Title
- State ←Performer State
- Timestamp ← Year of Start Data
- CBSA←CBSA
- Latitude←Latitude (generated)
- Longitude←Longitude (generated)
- Total paid in Capital ($)←Cost Plan Total Award Value
- Primary Sectors←’Carbon Capture and Storage’

2.1.5. NE (Nuclear Energy) DOE Project Database

The Office of Nuclear Energy (NE) provided an excel spreadsheet that contains a list of DOE projects funded by the office. To pre-process these files, we converted them into the CSV file format, then geocoded and performed CBSA mapping.

We collected the coordinate data for each entry in this dataset using the *Institution* attribute. We again used the list of accredited universities and their addresses from the U.S. Department of Education to match the *Institution* name to its address using the VLOOKUP function in excel. Once we had obtained an address for each row, we used the Google Geocoding API to get the latitude and longitude and CBSA mapping script to get the CBSA data.

At the conclusion of this step, we had added four additional columns (lat, long, CBSA, ST). We show the attributes of each preprocessed file in the following.

Table 17. List of attributes in the preprocessed NE data

- Title

- Institution
- Estimated Funding*
- Project Description
- Final Report
- Award Program
- Tech Area
- Year
- Lat
- Lng
- CBSA
- ST

The preprocessed file is transformed into the format that can be ingested into the IE database as follows.

Table 18. Basic information mapping for NE dataset to IE database

- EventType ←"DOE Project"
- Name ← Title
- State ←ST
- Timestamp ← parse_year(Year)
- CBSA←CBSA
- Latitude←Lat
- Longitude←Lng
- Total paid in Capital ($)←parse_fund_str(Estimated Funding*)
- Primary Sectors←'Nuclear'

The functions used in

Table 18 are described in the followings.

- parse_year(year_str) returns a string formatted YYYY parsing the given *year_str*
- parse_fund_str(fund_str) returns a string after getting rid of '$', ',' in the string *fund_str*.

2.1.6. OE (Office of Electricity Delivery and Energy Reliability) DOE Project Database

The Office of Electricity Delivery and Energy Reliability (OE) provided an excel spreadsheet that contains a list of DOE projects funded by the office. To pre-process these files, we converted them into the CSV file format, then performed geocoding and CBSA mapping.

The coordinate data was obtained by passing the *Street Address, City, State* and *Zip Code* attributes into the Google Geocoding API. The *Zip Code* attribute was also used as input for the CBSA mapping script.

At the conclusion of this step, we had added three additional columns (lat, long, and CBSA). We show the attributes of each preprocessed file in the following.

Table 19. List of attributes in the preprocessed OE data

- CID
- Internal/External
- Project Title
- Technology Area
- Performer
- Award Govt Share
- Award Cost Share
- Award Date
- Award Year
- Award End Date
- Street Address
- City
- State
- Zip Code
- FIPS
- Business Type
- Project ID
- Project Description
- Lat
- Lng
- CBSA

The preprocessed file is transformed into the format that can be ingested into the IE database as follows.

Table 20. Basic information mapping for OE dataset to IE database

- EventType ←”DOE Project”
- Name ← Project Title
- State ←State
- Timestamp ← parse_year(Award Date)
- CBSA←CBSA
- Latitude←Lat
- Longitude←Lng
- Total paid in Capital ($)←Award Govt Share
- Primary Sectors←’Smart Grid’

The functions used in Table 20 are described in the followings.

- parse_year(year_str) returns a string formatted YYYY parsing the given *year_str*

2.2. TASK1B: DATA AGGREGATOR MODULE

For ranking CBSAs, it is necessary to perform data aggregations to achieve features that can be used for scoring. In Phase 1, after the IE database was constructed from data sources, it was then imported into Tableau to convert the data into numerical metrics and score and rank the clean energy innovation ecosystem.

However, since the entire scoring algorithm(weighted summation of 5 normalized ecosystems component scores) was implemented within Tableau, there were limitations that we ran into in terms of how we handle various event type data points. For example, entities such as universities, lab agencies, incubators, etc. that were applicable to all technologies had to be assigned to a new technology category "NA" and only if the user selects this technology category will these assets be utilized in computing the networking scores. Similarly the enabling environment score is computed using "clean edge policy index" event type data points which were mapped to "NA" as well. So if a user fails to select "NA" within the technology filter, then the enabling environment score will not be computed at all. The Russell large companies event type also runs into the same issue since they are applicable to multiple technologies.

To avoid such a scenario, an additional software layer has been implemented in Python that can take the integrated database as input and generate all the numerical metrics needed for scoring as another output file. This output table will have all the numerical metrics as columns such as number of employees, companies, incubators, patents, publications etc. computed and aggregated for each CBSA and all the rows correspond to CBSAs. Note that there can be repeated CBSAs corresponding to different technologies.

Table 21. List of attributes in the output spreadsheet generated by the data aggregator module

- CBSA
- Primary Technology
- # of Employees
- ARPA-E Projects COUNT
- Company COUNT
- DESIRE_Policy_Num
- DOE Project COUNT
- DOE Project Funding Amount
- ERC Center (Enabling Environment) COUNT
- ERC Center (Investors) COUNT
- ERC Center (Nascent) COUNT
- ERC Center (Networking) COUNT
- Patent COUNT
- Publication_Count
- SBIR COUNT
- i3 Investors COUNT
- i3 Large Companies
- Accelerator COUNT
- CBSA Name
- CleanEdge Policy Index
- Early-Stage Energy Investors COUNT
- Incubator COUNT
- Lab Agencies COUNT
- Population
- Russell Large Companies COUNT
- University COUNT

Here are some details. For a region specified CBSA and a particular technology (Primary Technology),

- The module sums up the # of entities (e.g., # of publications) or amount of value (e.g., DOE Project Funding Amount) associated with the region and the technology
- Company formation with #employee >500 considered to be an i3 large company
- Population is not an aggregated value but we included for the convenience of use in Tableau

3. TASK 2: BUG FIXES AND NEW CAPABILITIES

3.1. TASK2A: BUG FIXES

Table 22. Identified Bugs

Bug	Bug details	Level of difficulty	Fixed	Notes
ARPA-E data	No time stamp on ARPA-E map	Minimal	Yes	
SBIR data	No time stamp on SBIR map. SBIR company type says private or null, but all SBIR applicants must be private companies so we can not show null or non-profit as other company types.	Minimal	Yes	
University data	Some universities show up on the aggregated state map, but not the CBSA map. It is computed in the scoring algorithm to calculate the ecosystem, but the underlying data does not consistently appear in the CBSA map.	Medium – geo-coding issue	Yes	
Statistics table on IE CBSA Dashboard	The # of SBIRs & Clean Edge policy index score does not seem to populate in the statistics table on the left, but it does on the hover drop down list.	Minimal	Yes	
CBSA level data	South Carolina Transportation CBSA: CBSA 161 has 0 events, but 170 has many events. Santa Fe, NM Solar IEs: CBSA # 116 has less events than CBSA #119.	Calibration: What are the weights for events?	Yes	ISNULL Tableau function needed to be used

CCS Data	Ecosystems above rank #50 have blank data in the IE. How did they get ranked with no data?	Calibration -	Yes	They get a rank due to cleanedge policy index and networking assets
CBSA Top Ranks Table on the IE CBSA Dashboard	CBSA top rank table does not conform to ranks on map for specific technologies like solar - Rochester solar and Stamford CT solar both rank 10 on map and table respectively.	Some cbsas can have the same rank	Yes	Not sure if we want to fix this. CBSAs with same score must get the same rank ideally.
Nuclear ranking	There are patents and publications in Idaho, but the CBSAs in the state are raking higher (250+) than CBSAs that have no metrics.	Medium – look into individual datasets	Yes	(Idaho falls, Boise city have some events)
Aravaipa Ventures	Check geolocation of investors – Aravaipa comes up in Phoenix, AZ, but it should be based in Boulder, CO		Yes	
Large Companies	Russell List of companies currently assigned to “No data” technology, so needs to be mapped to a category “NA” like clean edge policy enabling environment.	Medium- Design change.	Yes	
Incubators	Incubators currently assigned to “No data” technology, so needs to be mapped to a category like “NA” similar to clean edge policy enabling environment.	Medium- Design change.	Yes	

Delete NewsFeeds from the Integrated database.	Could affect scoring if the eventtype filter is not selected properly.	Minimal	Yes	
CBSA map disappears if “NA” not selected	When “NA” is not selected as a technology, the number of cleanedge policy datapoints become 0,so the formula that computes “Enabling Env Score” encounters a “divide by 0” error	Minimal	Yes	No more divide by 0 error, so even if “NA” is not selected map shows up, but Enabling Env Score is not considered and displayed as a 0.
Rename “NA” and “No Data” categories.	Rename “NA” with “Applicable to All Technologies(Always check this)” Rename “No Data” with “Could not map technology”	Minimal	Yes	
CBSA 47900 DC-VA-MD-WV has no clean edge policy index	We used the first state name for getting the policy index number. Since DC is not a state, it was assigned N/A.	Minimal	Yes	Average policy index after combining the policy index of VA-MD-WV=37.5 has been assigned.
Ranking CBSA’s	When CBSA’s have same ranks assigned to them when they have same scores, the next rank to be assigned should skip as many as the number of shared ranks.	Minimal	Yes	Changed the RANK_DENSE() call within Tableau to RANK()

3.1.1. University data geocoding Bug

Several universities that were mapped to states did not show up on the IE CBSA Dashboard page especially on the West Coast, only on the east. It is computed in the scoring algorithm to calculate the ecosystem, but the underlying data does not consistently appear in the CBSA map.

The following observations were made:

Total Universities : 222
21 - No State name, No CBSA name
30 - No CBSA, Has state name (This might have caused the universities to not show-up on the CBSA map but only the state map).

The following process was used to update the universities geocoding information.

- Downloaded Data for All Accredited Universities
- Got full address for each University using the Excel VLOOKUP function. This required some editing of the university names
- Obtain coordinate data by geocoding the addresses.
- Converted zip code to CBSA using the CBSA mapping tool developed in Phase 1.

3.1.2. CBSA mapping Bug

It was noted that occasionally entries assigned to a CBSA were mapped outside of the US. Upon further inspection, this was caused by the geocoding/cbsa mapping process. The CBSA mapping tool developed in Phase 1 takes as input a list of zip codes to obtain the CBSA data for. The issue occurs when the dataset contains postal code information for a location outside the US. The CBSA mapping tool assumes that everything it's given is a valid US zip code and looks for the data that matches it. If it does not find a match, it returns nothing. Otherwise, we get the CBSA data associated with that zip code. In these instances, we were able to match the foreign postal code to a US zipcode which resulted in the inaccuracy.

For example, the dataset that we received for company information contained *Address, City, State, Zip,* and *Country* attributes. The entry for AMC ETEC can be seen in Table 23.

Table 23. Example entry for company information

Address	City	State	Zip	Country
37 Avenue Arlucs	Cannes		6150	United States

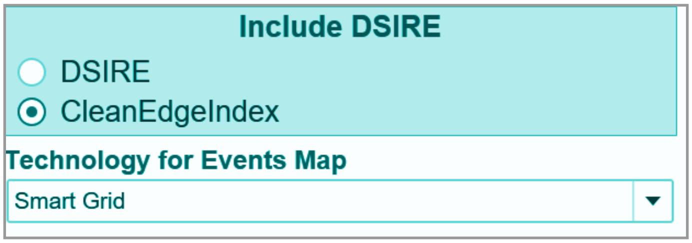
When we passed the address, city, state, and zip through the geocoding API, it correctly returned the coordinates in Cannes, France. However, when we passed the "zip code" into the CBSA mapper it matched this to the US zip code and returned a CBSA of 25540 which is for the Hartford, CT area. We corrected these errors by removing the incorrect CBSA.

3.2. Task2b: New Capabilities

3.2.1. Selection of the enabling environment data source for scoring

The dashboard provides an option to choose between the DSIRE or CleanEdge Policy Index numbers as input features for the “Enabling Environment” component of the Innovation Ecosystem.

Figure 5. Selection of enabling environment data source



Include DSIRE

○ DSIRE

● CleanEdgeIndex

Technology for Events Map

Smart Grid ▼

3.2.2. Addition of new features into scoring algorithm

New features such as “Number of DOE projects” and “Number of ERC projects” were added as subcomponents of the “Nascent Cleantech” bin which is one of the 5 main components of an Innovation Ecosystems within the main tool.

3.2.3. DOE Projects Interactive Dashboard

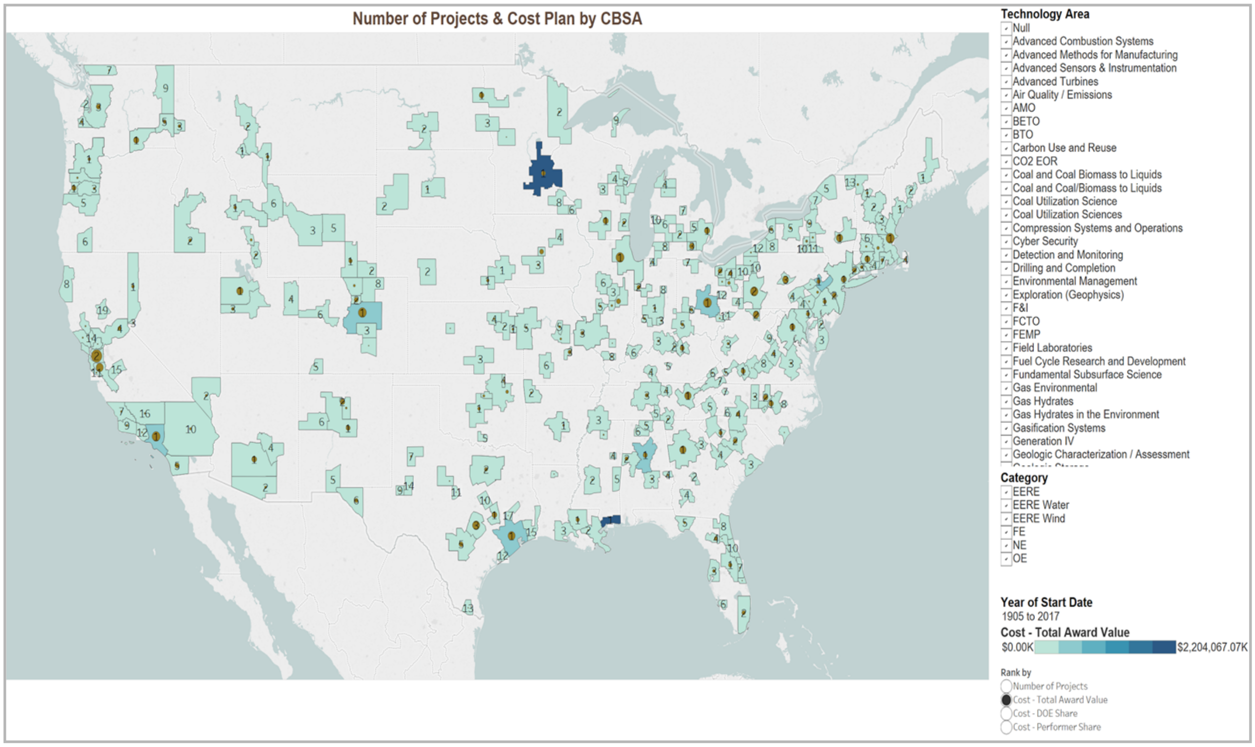
In addition to the main tool that was developed to identify clean energy innovation ecosystems at the CBSA and state level, a new visualization capability called “DOE Projects dashboard” was developed using Tableau Desktop software. This dashboard specifically focuses on visualizing and identifying top N CBSA’s based on the number of DOE projects and their funding amounts. A common database schema was first designed to accommodate integration of datasets from multiple DOE offices(such as EERE, OE, FE and NE). The schema details are given below:

Table 24. List of attributes in the input spreadsheet for Tableau to generate the DOE Projects Dashboard

- Agreement Number
- Cost Plan DOE Share
- Cost Plan Performer Share
- Cost Plan Total Award Value
- Directorate
- Performer
- Performer City
- Performer State
- Prime/Sub Indicator
- Program Area
- Project Status
- Project Title
- Subprogram
- Technology Area
- End Date
- Year of Completion Date
- Start Date
- Year of Start Date
- Category

Datasets provided by OE, NE, FE and EERE were used to populate the above schema table. Since most of these datasets were already formatted as csv files and had most of the attributes, these datasets were manually integrated to create the combined input spreadsheet for Tableau desktop. Finally, we provide two main dashboards for the domain expert to support decision making. The CBSA DOE Projects Dashboard Figure below allows the user to select a category of interest and any sub technology areas, and the map automatically updates the ranking and the color of each CBSA based on the Total Award Value assigned to that CBSA. Each CBSA can also be ranked based on the number of DOE projects, Total Award Value, DOE Cost Share or Cost Performer share.

Figure 6. DOE Projects CBSA Dashboard



Number of Projects & Cost Plan by CBSA

Technology Area
☑ Null
☑ Advanced Combustion Systems
☑ Advanced Methods for Manufacturing
☑ Advanced Sensors & Instrumentation
☑ Advanced Turbines
☑ Air Quality / Emissions
☑ AMO
☑ BETO
☑ BTO
☑ Carbon Use and Reuse
☑ CO2 EOR
☑ Coal and Coal Biomass to Liquids
☑ Coal and Coal/Biomass to Liquids
☑ Coal Utilization Science
☑ Coal Utilization Sciences
☑ Compression Systems and Operations
☑ Cyber Security
☑ Detection and Monitoring
☑ Drilling and Completion
☑ Environmental Management
☑ Exploration (Geophysics)
☑ F&I
☑ FCTO
☑ FEMP
☑ Field Laboratories
☑ Fuel Cycle Research and Development
☑ Fundamental Subsurface Science
☑ Gas Environmental
☑ Gas Hydrates
☑ Gas Hydrates in the Environment
☑ Gasification Systems
☑ Generation IV
☑ Geologic Characterization / Assessment

Category
☑ EERE
☑ EERE Water
☑ EERE Wind
☑ FE
☑ NE
☑ OE

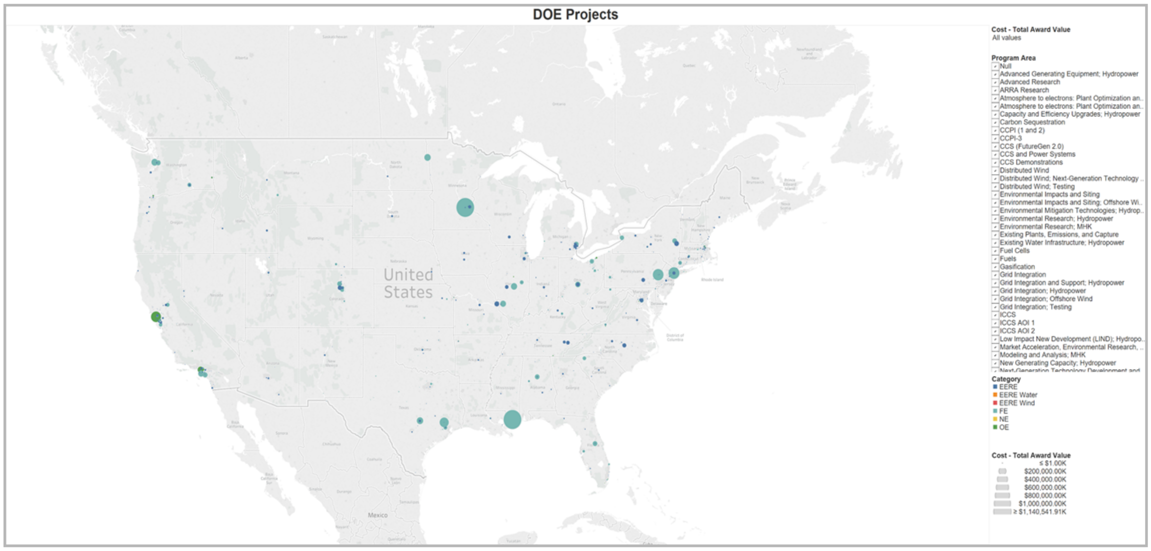
Year of Start Date
1905 to 2017

Cost - Total Award Value
$0.00K $2,204,067.07K

Rank by
○ Number of Projects
● Cost - Total Award Value
○ Cost - DOE Share
○ Cost - Performer Share

The DOE Projects Dashboard Figure below allows the user to select a program area of interest and the map automatically updates the size of the dots(each dot represents a project) based on the Total Award Value assigned for that project. The color of the dot represents one of the categories such as EERE, EERE Water, EERE Wind, FE , NE and OE.

Figure 7. DOE Projects Dashboard



DOE Projects

Cost - Total Award Value
All values

Program Area
☑ Null
☑ Advanced Generating Equipment; Hydropower
☑ Advanced Research
☑ ARRA Research
☑ Atmosphere to electrons: Plant Optimization an..
☑ Atmosphere to electrons: Plant Optimization an..
☑ Capacity and Efficiency Upgrades; Hydropower
☑ Carbon Sequestration
☑ CCPI (1 and 2)
☑ CCPI-3
☑ CCS (FutureGen 2.0)
☑ CCS and Power Systems
☑ CCS Demonstrations
☑ Distributed Wind
☑ Distributed Wind; Next-Generation Technology ..
☑ Distributed Wind; Testing
☑ Environmental Impacts and Siting
☑ Environmental Impacts and Siting; Offshore Wi..
☑ Environmental Mitigation Technologies; Hydrop..
☑ Environmental Research; Hydropower
☑ Environmental Research; MHK
☑ Existing Plants, Emissions, and Capture
☑ Existing Water Infrastructure; Hydropower
☑ Fuel Cells
☑ Fuels
☑ Gasification
☑ Grid Integration
☑ Grid Integration and Support; Hydropower
☑ Grid Integration; Hydropower
☑ Grid Integration; Offshore Wind
☑ Grid Integration; Testing
☑ ICCS
☑ ICCS AOI 1
☑ ICCS AOI 2
☑ Low Impact New Development (LIND); Hydropo..
☑ Market Acceleration, Environmental Research, ..
☑ Modeling and Analysis; MHK
☑ New Generating Capacity; Hydropower

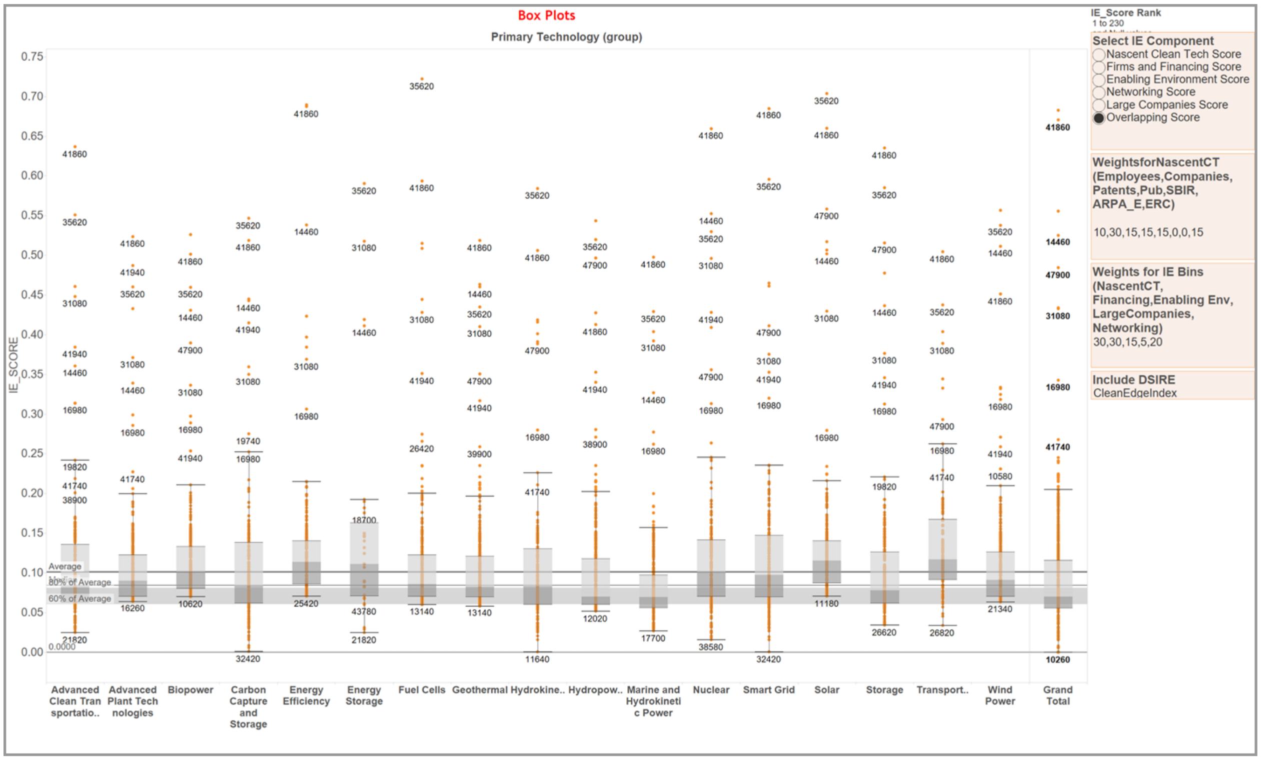
Category
■ EERE
■ EERE Water
■ EERE Wind
■ FE
■ NE
■ OE

Cost - Total Award Value
≤ $1.00K
$200,000.00K
$400,000.00K
$600,000.00K
$800,000.00K
$1,000,000.00K
≥ $1,140,541.91K

3.2.4. Scoring methodology evaluation for all 14 clean energy technologies

The data aggregator module, enabled the development of the Boxplot visualization, where the X-axis represents the various clean energy technologies and the Y-axis maps the IE score assigned to each CBSA. This plot not only allows the user to identify the outliers right away, but also allows us to assess our scoring mechanism. As an example, the below figure clearly shows how the IE score for a majority of the CBSA's ranges between a small window of 0.05 and 0.15 for all the technologies. Hence this clearly provides a means to compare with other scoring methods such as different ways of normalizing the features for example to widen the spread of the scoring and clearly binning the CBSA's into different bins in terms of their innovation strength.

Figure 8. Boxplots Dashboard



IE_Score Rank
1 to 230

Select IE Component
○ Nascent Clean Tech Score
○ Firms and Financing Score
○ Enabling Environment Score
○ Networking Score
○ Large Companies Score
● Overlapping Score

WeightsforNascentCT (Employees,Companies, Patents,Pub,SBIR, ARPA_E,ERC)
10,30,15,15,15,0,0,15

Weights for IE Bins (NascentCT, Financing,Enabling Env, LargeCompanies, Networking)
30,30,15,5,20

Include DSIRE
CleanEdgeIndex

4. SUMMARY

This document summarizes all the enhancements made to the innovation ecosystem discovery tool that was developed in Phase 1 last year with DOE-EPSA. The underlying backend architecture was refined to allow for easy ingestion of new and diverse datasets of various formats. All the major bugs identified in Phase 1 have been fixed and new capabilities such as DOE projects dashboard, selection of enabling environment data source options, and new features for scoring were added. Finally, the scoring methodology was visually evaluated using the box plots dashboards. These plots indicate an opportunity for further improving the scoring methodology in the future by implementing different normalization techniques for various features. This would allow us to clearly distinguish/rank geographic areas in terms of clean energy innovation activity.

REFERENCES

[1] "Regional Energy Technology Innovation," [Online]. Available: https://energy.gov/mission-innovation/regional-energy-technology-innovation.

[2] Tableau, [Online]. Available: https://www.tableau.com/ .

[3] U.S. Department of Education Office of Postsecondary Education, "The Database of Accredited Postsecondary Institutions and Programs," April 2017. [Online]. Available: https://ope.ed.gov/accreditation/GetDownLoadFile.aspx.

[4] Google, "Geocoding API," August 2017. [Online]. Available: https://developers.google.com/maps/documentation/geocoding/start.

[5] NC Clean Energy Technology Center, "Database of State Incentives for Renewables & Efficiency," [Online]. Available: http://www.dsireusa.org/ .

APPENDIX A. EXCERPT FROM STATEMENT OF WORK

- Project Lead: Robert Horner
- EPSA Information
    - Office: 52
    - Project:
    - CPS Agreement:
- Performer: Oak Ridge National Laboratory
- Performer POC: Supriya Chinthavali

A.1. SUMMARY

ORNL and EPSA worked jointly to develop the Ecosystem Discovery tool that can compile a list of current clean energy innovation ecosystems (IE) in the U.S. and their characteristics with the support of an automatic machine extraction methodology. The data collection supported DOE's understanding of existing clusters of innovation institutions at the local, state, and regional levels.

The pilot study demonstrated the feasibility of an automatic data ingest pipelines to perform text analysis, natural language processing, and link analysis to identify innovation ecosystems. The team now proposes to build out the tool's capabilities by fixing identified bugs and integrating new data sets.

A.2. TASK SCHEDULE

Two tasks are included in this work plan: adding data sets and fixing bugs in the existing tool.

A.2.1. TASK 1: ADDING DATA SETS

The performer will add data to the tool from the following sources, in the order of priority listed and in consultation with the DOE sponsor, within the budget constraints of the project. Additional or alternative data may be identified and included by mutual agreement. Data originating within the Department of Energy (DOE) will be collected by the EPSA project lead and provided to the performer. Data external to DOE will be collected by the performer. Any DOE data provided to the performer will be handled according to the stipulations agreed to when the data is transmitted. The datasets will be further curated, geocoded and mapped to specific technology as needed. The backend software that accomplishes this task with be refined, generalized and pushed into the Git repository.

Data sources:

- Department of Energy
    - Fossil Energy
        - Demo studies for CCS
    - Nuclear Energy
        - PICS – project integration control systems
        - Nuclear competitive awards
        - Nuclear voucher program
        - NSUF – Nuclear Science User Facilities
        - NEET – nuclear energy enabling technology database
        - GAIN – nuclear private companies
    - Energy Efficiency and Renewable Energy
        - List of startups from incubator network

        - Small business CRADAs
    - Clean Energy Investment Center
        - DOE spend data from individuals in program offices
        - DSIRE database
    - Office of Technology Transitions
        - Data on small, medium, and large companies
        - Venture development organizations from EDA
    - Office of Electricity
        - AMI – EIA data sets
        - ARRA smart-grid data
    - Loan Program Office
        - Project data
            - Large companies
            - Intermediaries
            - Law firms
            - Engineers
- Small Business Innovation Research program
    - SBIR.gov projects
- National Science Foundation
    - Award information
- Environmental Protection Agency
    - Energy Research facilities
- Bureau of Labor Investigation and Department of Education
    - Graduate Education

A.2.2. Task 2: Bug fixing

Fix bugs identified during Phase I which are listed below:

- University data may not be properly geocoded. Some universities show up on the aggregated state map, but not the CBSA map. It is computed in the scoring algorithm to calculate the ecosystem, but the underlying data does not consistently appear in the CBSA map.
- The # of SBIRs & Clean Edge policy index score does not seem to populate in the statistics table on the left, but it does on the hover drop down list.
- South Carolina Transportation CBSA: CBSA 161 has 0 events, but 170 has many events.
- Santa Fe, NM Solar IEs: CBSA # 116 has less events than CBSA #119.
- CBSA top rank table does not conform to ranks on map for specific technologies like solar - Rochester solar and Stamford CT solar both rank 10 on map and table respectively. Large companies score was not computed correctly. A quick release was made to fix this issue in the last phase but was not tested exhaustively and hence needs to be revisited.
- Others identified during task performance, subject to budget availability.

A.2.3. Task 3: Optional refinements and Documentation

As budget allows, develop an alternative to strict ranking of innovation clusters, allowing for categorization of cluster scores and/or descriptions of distribution. Documents the changes made to the tool.

A.3. Milestone/Deliverables

Table 25. Milestone/Deliverables Schedule

Milestone/Deliverable	Date
Bi-weekly update	Every two weeks
Updated tool	4 months from project start
Updated documentation	4 months from project start

A.4. Estimated Cost

Task 1 : Adding Datasets – 60K
Task 2: Bug Fixing – 20K
Task 3: Optional Refinements and Documentation – 20 K

APPENDIX B. DETAILED INFORMATION ON DATA SOURCES

B.1. QTR LIST COMPARISON TO ERC CATEGORIES

Table 26. QTR List Comparison to ERC Categories

QTR List	ERC Categories
1.Biopower	Bioenergy and Biofuels
2.Hydropower, 4. Marine and Hydrokinetic Power	Hydropower, and Marine and Hydrokinetic Power
3.Geothermal	Geothermal
5.Nuclear	Nuclear power
6.Solar	solar
7.Wind Power	wind
8.Carbon Capture and Storage	Carbon Capture and Sequestration
9.Energy Efficiency	Energy Efficiency and Sustainable Building Design
10.Smart grid	Smart Grid and Transmission and Distribution
11.Advanced Plant Technologies	Fossil Fuels and Advanced Plant Technologies
12.Storage 13.Fuel Cells	Energy Storage and Fuel Cells
14.Advanced Clean Transportation and Vehicle System Technologies	Transportation Technology
Others	Water Technology and Water Use Efficiency
	conventional fuels**
	recycling and waste**

*There is not a clear delineation between hydro and marine and hydrokinetic power in ERC. The one category applies to two in QTR list.

APPENDIX C. FUTURE WORK

C.1. Calibration and Validation of the Current Tool

Typically, in computer science, precision and recall are used as tools to determine whether an approach has been successful or not. However, precision, or the positive predictive value, is the fraction of retrieved instances that are relevant, whereas recall, or sensitivity, is the fraction of relevant instances that are retrieved. Both assume that there is a quantifiable measure of relevance which is not necessarily the case with the study of innovation ecosystems. It took three months to develop the tool, another 4-6 months for tool enhancement and the data validation will take at least another 3 months, or longer, depending upon the availability of information on innovation ecosystems for various technology types.

The scoring algorithm for clean energy innovation ecosystems is flexible; it allows the user to choose different weighting for each of the five components as well as the weighting of the subcomponents of each component of the ecosystem as well. This flexibility will be useful for the next phase of analysis which will involve validating the results of the tool with experiences from subject matter experts in the space. For those who have studied innovation ecosystems and who know exactly where certain clean energy technology specific ecosystems are, this tool can be refined by calibrating the weightings to match the known entities. In addition, the tool will hopefully generate some surprise results which could be validated through other means. With further calibration and validation of known and previously unknown ecosystems, this tool will allow the user to find more unknown ecosystems because it could leverage the weightings chosen and the underlying datasets to identify hundreds of ecosystems for any given technology, depending upon data availability.

However, it remains to be seen whether it will be possible to definitively calibrate the tool. As discussed in Section 2, the US Cluster Mapping project's reliance on industry wide NAIC codes as an input makes their results too broad to characterize clean energy technology specific innovation. Other analysis by the Small Business Administration and others have focused on institutions and their ability to foster growth of a cluster, but not necessarily on using data to discover where technology specific ecosystems exist.

We may find that there is no binary assumption for an innovation ecosystem. There are a plethora of reasons why ecosystems form and even more reasons why individuals would want to study these phenomena, so future analysis may prove that there is no one calibration of the tool that could be used for all purposes, but that the flexibility of the weighting could help the user utilize the tool for various analyses in the field of innovation, such as the success of universities, the rate of publications if certain types of entities are in the ecosystems, the role of large companies in ecosystems, etc.

C.2. Data Improvements

We relied heavily on i3 subscription level data for our baseline given the challenges of retrieving DOE data. Even with a paid subscription, there are many quirks to the i3 data. In particular, much of the corresponding financial data such as revenue, paid in capital and round of finance to size the company were not available. Even for employment data, there was a suspiciously high level of companies with 0 employees which implies that the data was likely not correct. For example nearly half of our set of companies had 0 employees. In addition, many databases tend to list the headquarter of the organization, rather than the place where the company or university is actually doing research or has specific activities that the ecosystem would be most interested in.

The tool has i3 data for investors which is a very solid start for CEIC analysis, however, the i3 database was missing almost half of the investor information for their covered companies. This is difficult information to find because private companies do not need to report their earnings so this information in 10ks or other such reports that public companies are required to file. For future analysis on investor information, either the Pitchbook subscription could be purchased or, with more time, the ORNL could do further NLP work on News feeds to try to glean this data directly.

Pitchbook is a subscription database that has proprietary information about venture capital, private equity and merger and acquisition deals for private companies. The company uses artificial intelligence to crawl the web to find their data and then their analysts verify the data (http://pitchbook.com/). Although we do not have a subscription, it seems like this site has more comprehensive private company finance data than others we have reviewed. CB Insights (https://www.cbinsights.com) is another alternative which could be explored to provide additional financial data for the companies selected in the i3 database.

We tried as much as possible to limit our data to energy related technologies with a focus on what role the DOE plays on innovation ecosystems. For example, we deferred to the PAMS database to search for SBIR awards specific to the US DOE. However, there are many SBIR awards from other agencies and a search on the Small Business Administration's (SBA) Tech-NET, may have provided a more robust list of technologies. However, we had challenges associated with choosing which SBIR was relevant even within an energy specific database, so for the pilot, did not want to open the search to the larger corpus of Tech-NET, but this is a possible improvement for the future.

On the contrary, we did use the full list of nationals labs, even those from other agencies even though we recognize that not all of the labs are specifically energy related. In this case, we included all of the national labs because we were looking for a proxy for locations that would likely improve the flow of human capital and ideas, which research institutions, such as labs tend to do. Cooperative Research and Development Agreement (CRADA)information for clean energy who have small businesses as a notable share of their total active CRADAs, relative to other labs that do not would be an interesting metric to help target which labs are most active in clean energy commercialization for future analysis. In addition, some national labs are also now participating in new lab initiatives to better facilitate technology commercialization by linking lab scientists and lab resources to innovators/ entrepreneurs. Examples of these lab initiatives are DOE EERE's Lab-Corps, Small Business Vouchers, and Lab-Embedded Entrepreneurship Programs (LEEP). Other federal agencies have new lab initiatives underway that have technology commercialization and networking objectives as well which could be included to refine our list of labs.

In future iterations, it would be interesting to expand the Networking Assets to include entities funded by the US DOE, National Science Foundation (NSF) and National Institute of Science and Technology (NIST) such as the National Network for Manufacturing Innovation (NNMI) and others. In late 2014, Congress passed the Revitalize American Manufacturing and Innovation Act (RAMI), which established the National Network for Manufacturing Innovation (NNMI). Several federal agencies are contributing to the national NNMI. These include the DOE, DoD, NIST, NASA, and NSF. NNMI brings together industry, academia and federal partners to increase U.S. manufacturing competitiveness. In addition, the DOE's Advanced Manufacturing Office (AMO) has established advanced manufacturing facilities. AMO Facilities are collaborative communities that provide participants with affordable access to physical and virtual tools and enable demonstration in targeted technical areas of manufacturing[1]. Specific DOE AMO facilities such as the Manufacturing Demonstration Facility (MDF), Critical Materials Hub, Institute for Advanced Composites Manufacturing Innovation (IACMI), and others could be added to the Networking

[1] See http://energy.gov/eere/amo/facilities

Assets component. The NSF also has many centers (>50 since 1985) intended to serve as hubs for regional innovation clusters[2].

Our methodology for choosing patent data could be revisited. In particular, there is competing literature about how long clean energy technologies take from patent to commercialization. We used three years, but other literature finds time lines as long as eight to ten years as a potential range. In addition, there is a delay from the patent application to the time of grant, so we used the grant date, but the application date may have been closer to the actual "start date" of the innovation. Finally, much like the i3 headquarter issue, we took the headquarters location of the assignee rather than the inventor's address, which was not immediately available. This may not be a problem for small firms, but for large firms, the headquarter may not actually point to where the research actually happens.

In fact, there is a persistent geo-coding issue across many categories, such as the investor component. Given our time constraints, we chose the first investor, when in fact, the first entity listed may not be the most influential actor. In addition, we did not have the means within the scope of this project to pinpoint where the actual activity came from vs. headquarters. For example, an investor could be listed by their headquarter as opposed to the regional office that did the due diligence and had the relationship with the investee. It would be helpful to explore whether the use of NLP could help to discover the actual location of the work being done because we acknowledge there are issues with the results if relying on the address of headquarters.

Currently the technology focus is primarily determined by the information in the Nascent Clean Energy Indicator component. That said, it would be helpful to collect and aggregate technology specific data for each category. For example, to collect information about University program/dollars/graduates/theses in technology X, investors in technology X, large company divisions/sales/R&D/staff in technology X, and incubators in technology X. It would also be helpful to categorize an institution's funding/ideas/technologies by innovation stages, activity type, and success metrics. Success metrics, however, are subjective and would need to be researched and decided upon in advance of the collection period.

It would be helpful for funding organizations to be able to track the impact of its investments by following the flow of capital and return on investment through private capital flows to sponsored ideas or entities. It would also be helpful to collect and insert state and local data such as NYSERDA awards and others. In fact, right now, we currently only have relationships between firms and investors, but we should explore every permutation of relationships between various organizations, if possible. This analysis could include:

- Federal government's relationships with other entities to understand federal government's chain of influence
- State and local government relationships with other entities
- Umbrella or coordination organizations (consortia, center, NGOs, trade group, chambers of commerce)

With this addition of data and analysis on relationships, it would be possible to test hypotheses of important relationships, such as the relationships between universities and small/medium firms, universities and large firms, shared patents between organizations, etc.

[2] Peterson, Thomas, "Creating an Innovation Ecosystem": https://www.nsf.gov/od/oia/programs/epscor/natcon/presentations/PetersonPresentationEPSCoR.pdf

C.3. Analytical Improvements

Once the tool is complete and has been calibrated, there are a number of analyses which could be explored. For example, it would be possible to analyze the underlying metrics of ecosystems and to answer questions such as, the x% of [technology specific] ecosystems tend to have a university and/ or a lab. Such assertions would help policymakers understand the relative importance of such networking assets to the success of ecosystems.

By the same token, with additional time series data, it would be possible to analyze the growth and or decline of ecosystems over time to understand if there are common trends or factors which contribute to these phenomena. For example, if a net metering policy changes will solar investors immediately pull out? Or if an SBIR is awarded, do other areas with high patent rates start to commercialize as well?

The current tool offers a lot of flexibility, but with more time, the tool could also be refined to allow for further analysis. One possible improvement is to allow the user to choose the exact number of subcomponents to define what they would like to see in an ecosystem; for example, choose an ecosystem that has at least one university, 3 small firms and 10 patents for a particular technology. It would also be interesting to see how the composition of an ecosystem would change if the underlying region of interest was not a CBSA, but a county or other statistical method of cluster identification not based on a socio-political unit. Practically speaking the latter is less of a priority because the benefit of a CBSA is the commuter shed which typically demarcates potential flow of human capital.

C.4. Function and User Experience of the Tool

The Ecosystem Discovery tool was created in Tableau as a pilot product because Tableau offered a number of the visualization capabilities that we wanted to exploit in the context of being able to view regional activity and capabilities. However, the analytical questions that the underlying integrated database could address were revealed through the creation of the product. While Tableau is a flexible software package that allows the user to choose many different filters, it could also be seen as a complicated tool for the non-technical user. Upon completion of the pilot, the underlying integrated database could be used to create hypothesis driven tools to address specific policy questions for the non-technical user.

For example, if the Office of Technology Transitions would like to use the tool to do impact evaluation exclusively, perhaps a text base search, much like the Google experience, would be better suited for program managers without a computer science background. Google has been able to figure out in a much broader, and therefore, more complex context how to take in text based searches and return very relevant, prioritized, comprehensive results without the user having to choose filters, data sets, priority of innovation criteria or any other detail. They have also created knowledge graphs which may be another tool that could be used to analyze regional IEs, rather than weighed sets of criteria. Ultimately, we would want the same experience for the user of the Ecosystem Discovery tool. Once the tool has been calibrated and the most important options for analysis have been undertaken on the back end, there is a level of work on the front end that would be needed to make the experience easier and therefore more accessible to a broader audience. The concern with not doing the work to create a user experience that is widely usable to the non-technical audience is that the tool will sit on the shelf and not be used.

This tool is as powerful as the data sets that underpin it, so a maintenance plan will need to be put in place to update the data on a regular basis. A maintenance plan will improve the quality of results and the constant refresh will also create the time series needed to begin to do future trend analysis.