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

Authors:

Jessica Lin (ORISE Fellow for the US Department of Energy)

Supriya Chinthavali (Oak Ridge National Laboratory)

Chelsey Dunivan Stahl (Oak Ridge National Laboratory)

Christopher Stahl (Oak Ridge National Laboratory)

Sangkeun Lee (Oak Ridge National Laboratory)

Mallikarjun Shankar (Oak Ridge National Laboratory)

Date Published: August 25, 2016

Prepared by
Computational Sciences & Engineering Division
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

GLOSSARY

ARPA-E	Advanced Research Projects Agency - Energy
CBSA	Core-Based Statistical Areas
CEIC	Clean Energy Investment Center
EC	Ecosystem Component
EPSA	US DOE's Office of Energy Policy and System Analysis
IE	Innovation Ecosystem
MSA	Metropolitan Statistical Areas and Micropolitan Statistical Areas
NIICE	National Incubator Initiative for Clean Energy
NLP	Natural Language Processing
ORNL	Oak Ridge National Lab
PV	Photovoltaic
QTR	Quadrennial Technology Review
SBIR	Small Business Innovation Research

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.

ABSTRACT

While the term “innovation ecosystem” is often utilized, the concept is rarely quantified. Oak Ridge National Lab (ORNL) conducted a ground-breaking application 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 for the US DOE’s Office of Energy Policy and System Analysis (EPSA). Large and varied datasets were combined with an advanced user interface, together known as the Ecosystem Discovery tool, to score, rank and map existing clean energy innovation ecosystems (IEs).

EPSA advanced a novel definition for 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. 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 of data as well as the composite IE rank. In an independent parallel effort, a DOE analyst in EPSA developed a short list of 22 top US clean energy innovation ecosystems; the Ecosystem Discovery tool was able to identify over 90% of the analyst-reported ecosystems. Full validation and calibration remain outstanding tasks.

The tool and the underlying datasets have the potential to address a number of important policy questions. The initial broad list of U.S. clean energy innovation ecosystems, with geographic area, technology focus, and list and types of involved organizations can help describe regional technology activities and capabilities. The implementation of knowledge discovery techniques also revealed both the potential and limitations of an automatic machine extraction methodology to gather ecosystem component data. The project demonstrates that a machine-assisted methodology gives the user a replicable method to rapidly identify, quantify, and characterize clean energy innovation ecosystems.

Keywords: clean energy innovation ecosystems, natural language processing, mapping

1 INTRODUCTION

1.1 BACKGROUND

On November 30, 2015, as part of the Paris Conference of Parties 21, the US Government along with 19 other countries announced that they would double their R&D budgets for clean energy innovation over the next 5 years under the umbrella of Mission Innovation. In this context, EPSA required a broad list of current clean energy IEs in the U.S. and their characteristics. Manual IE compilation can be labor-intensive and is subject to poorly characterized completeness. EPSA and ORNL jointly tested whether knowledge discovery methods would work to find information about clean energy IEs.

While there is much competing literature on the topic of their effectiveness, our intent was not to study whether or not an ecosystem is “effective.” The purpose of this exercise was to demonstrate the feasibility of an automatic data ingest pipeline to perform text analysis, natural language processing, and link analysis to identify clean energy IEs in the U.S. in a replicable manner and if possible, offer insights into their data characteristics. Far less literature is available with empirical data on the characteristics and composition of IEs, so this work was intended to fill this gap.

2 THEORY

Per the project’s objective, we were looking for clean energy IEs that already exist and that have the ability to share resources within a close physical proximity. In addition to the human capital resource, clean energy innovations face barriers to finance due to the high capital intensity and long lead times compared to other sectors [3], so we deemed it important to include the five most critical components of clean energy innovation in our definition of ecosystems - nascent clean energy indicators, investors, a commercialization-enabling policy environment, networking assets and large companies. At the inception of the project, we were not sure how heavily each factor would weigh in the ecosystem, but did want to be able to have measurable factors. Therefore, we defined a clean energy IE as the intersection of nascent clean energy assets with financing mechanisms and related companies located in a geographically concentrated area that has an enabling environment, or supportive clean energy policies, which encourages commercialization and networking assets to increase communication and collaboration opportunities as illustrated in the diagram below in Figure 1.

Figure 1: Clean Energy Innovation Ecosystem

Our working definition of an innovation ecosystem is derived from the cluster theory of competitive advantage [4], a highly regional strategy of economic development introduced by Michael Porter in 1990. Porter's framework defines clusters as geographically concentrated collections of firms, suppliers and related industries. Today, economists still define a cluster as a concentration of interconnected companies that both compete and collaborate. [5] Brookings has recently expanded the definition of innovation ecosystems as a synergistic relationship between people, firms and place that facilitates idea generation and accelerates commercialization. [6] Physical proximity is a key factor in this framework because proximity assists in idea formulation, refinement, and implementation by creating conditions conducive to the formation of teams, and connecting those teams to resources that support transition to market.

Critics to Porter argue that his theory and subsequent work pays insufficient attention to innovation and knowledge as inputs to cluster formation. For example, the US Cluster Mapping website, a national initiative that provides open data on regional clusters and economies to support U.S. business, innovation and policy, co-sponsored by Harvard Business School and the US Economic Development Agency, have based their analysis on North American Industry Classification (NAIC) codes which rely on standard industrial classifications which were originally developed in the 1930s to promote the comparability of data describing various facets of the US economy. This information is too broad and dated to characterize clean energy innovation ecosystems. To address this criticism, we focused on metrics that could potentially capture early stage, technology specific activities in each component of our framework in Figure 1.

The definitions of the five components of a clean energy IE are below.

Nascent clean energy indicators are the drivers of innovation. Depending upon the stage of the idea, these assets could range from very early stage research ideas formed in the lab or a university to a fledgling startup to an idea that was conceived in the research lab of an incumbent company. These assets are also determined by the presence of individuals who can build technologies to scale such as engineers, research scientists, and business people. Nascent clean energy indicators could be measured by the presence of startup companies, patents coming out of research labs or universities or even the movement of individuals who are cited in cutting edge research reports or management teams of successful startups. Research also supports the idea that a disproportionate amount of high technology startups in the United States are founded by highly skilled immigrant populations. [7], [8]

Investors and financing mechanisms are key to the growth and commercialization of nascent clean energy assets. Many early stage finance companies have a geographical bias, investing locally, because their work tends to be very hands-on. Access to finance is particularly important in the clean energy space where the amount of funding needed as well as the time horizons to commercialization are typically larger than other sectors, such as IT or biotech. The funding types and levels could vary based on the size of the nascent clean energy indicators and could range from grants such as those offered by Small Business Innovation Research (SBIR) and the Advanced Research Projects Agency - Energy (ARPA-E) programs, to venture capital to crowdsourced funds. In general, these will be more risk tolerant investing mechanisms.

The enabling environment encompasses the regulatory regime and/or the physical electricity generating characteristics of a region that would favor the growth of one sector of clean energy over another. There is evidence throughout the world that local market incentives have contributed to the diffusion of renewable technologies, and have likely enabled smaller companies to perform research in these fields. [9]. For example, a robust net metering program could incent distributed generating options such as solar photovoltaic technologies. The physical topology of an area could also give rise to demand for certain technologies. For example, the wind-swept plains of the Midwest or the mountain-bracketed waterways of the Pacific Northwest could create corresponding demand for wind or hydro technologies, respectively, that utilize local resources. Most of the enabling environment characteristics will be at the local or state level, rather than national and include things such as general business incentives, state and local clean energy subsidies, and other clean energy incentives.

Networking assets are events or entities that facilitate communication and collaboration among the various actors of the ecosystem. There is a growing body of research which asserts that networking assets are increasingly valuable within innovation driven economic clusters. [6] Brookings characterizes an IE as the overlap of physical assets, economic assets and networking assets where networking assets are the relationships between actors¹. Given the recent proliferation of the “open innovation” mode of R&D, universities and research labs, for example, are now more porous and networked, and therefore these have become networking assets despite previous criticism that they were not open institutions.

Large companies could play a number of different roles within an IE, so they were deemed important enough to have their own category. The project intended to analyze the influence of large companies on varying technologies upon completion of the data ingest.

3 COMPLETE DATA ANALYTICS WORKFLOW

The methodology of the project was divided into three phases:

- 1) Collect and categorize US clean energy IE data;
- 2) Visualize the data in a user friendly format; and

¹ In fact, Brookings specifies that networking assets that build strong ties focus on strengthening relationships within similar fields whereas networking assets that build weak ties focus on building cross sector relationships.

- 3) Create a tool that could analyze and score the data to discover and rank a broad list of IEs for a number of technologies including, but not limited to, solar, storage & wind power.

3.1 CLEAN ENERGY TECHNOLOGY TYPES AND SUBCATEGORIES

The ecosystem discovery project covered all of the technologies that are included in the Quadrennial Technology Review (QTR). In order to improve the quality of our datasets, we needed search terms not only for the technology type, but also sub-categories of those technologies. To begin, we created an initial technology list composed of the following fourteen major technology categories from the QTR:

- Advanced Clean Transportation and Vehicle System Technologies
- Advanced Plant Technologies
- Biopower
- Carbon Capture and Storage
- Energy Efficiency
- Fuel Cells
- Geothermal
- Hydropower
- Marine and Hydrokinetic Power
- Nuclear
- Smart grid
- Solar
- Storage
- Wind Power

We took a sample of abstracts published in 2013-2015 related to those fourteen technologies and ran them through the ORNL tool Piranha to discover publication clusters. The software processes large volumes of unrelated free-text documents and shows relationships amongst them. The results are then presented in clusters of prioritized relevance, showing relationships among documents so that users can quickly recognize connections (See Figure 2). The software creates hierarchical groups or clusters based on document similarity calculated using term frequency weightings and using unsupervised learning methods on the full content of each document (See Figure 3). [10] Doing so allowed us to see cross cutting technologies, but because of the broad scope of the initial topic areas searched, the clusters formed around basic concepts i.e. climate, power, wastewater, etc. as opposed to identifying technologies.

In order to narrow the scope of technologies further data was collected from “New Energy Congress”, the 2011 and 2015 QTR reports. The data was then translated into a hierarchical listing of technologies and their associated sub-technologies. For each sub-technology a topic search was done across Web of Science to identify publications over the last three years. Using the Piranha toolset to cluster the resulting publication set, significant keywords (or technologies) were identified for each sub technology.

Figure 2: Clustering of Clean Energy Topic Search Abstracts

Figure 3: Drill Down View on Piranha Clustering Output

3.2 DATA SOURCES OVERVIEW

The first step in working with each of the datasets was to weed out data that was not relevant to the project. In most cases we were able to reduce our datasets by filtering on criteria such as keywords, date ranges, geographic ranges, etc. We began the project by downsizing the data from the i3 database. From the 24,000+ companies in the i3 database, we were able to create an initial list of 3,859 companies by filtering on the primary sectors of interest, geographic location within the United States, and a status of private, public, or acquired. Although Web of Science contains over 90 million records², we were able to reduce the set to 35,000+ publications by filtering only those since 2012 that contained a key technology. Similarly, we reduced the number of news feeds from approximately 16 million to 10,000+. Our first pass at collecting patents used a filter to limit results to the years 2013 - present that had an assignee located within the United States and were related to the sub-technologies list. From the one million+ patents granted in the US from 2013-2015, our search returned 35,000+ results; we applied additional filters on the CPC codes to come out with 13,939 patents.

Once we had a reduced corpus we needed to categorize each event separately in order to be able to identify IEs by technology. Due to data availability, we focused on giving a technology categorization to the datasets in the Nascent Clean Energy Indicator category, rather than all five of the Ecosystem Components³. This information came with the data in the case of the i3 companies, and was linked during

² "Web of Knowledge - Real Facts - IP & Science - Thomson Reuters." 2013. 22 Aug. 2016

<http://wokinfo.com/citationconnection/realfacts/>

³ Enabling Environment, Networking Assets, and some Investor and Large companies datasets are not categorized by technology.

the filtering process for patents, publications and news feeds. The SBIR and ARPA-E sets, however, posed more of a challenge. For these data sets we tried to match each project with a technology by doing a keyword search on the project title, which resulted in the categorization of about 60% of these two data sets.

Each event also needed to have geographic information such as state, zip code, latitude, and longitude coordinates in order to be used in the scoring and visualization. Since most of our datasets were missing some or all of this data, we obtained it using the Google Maps Geocoding API [11] and the Google Places API [12] to perform both geocoding and reverse geocoding. This was successful for approximately 88% of the data. Using the geographic data, we were able to link events to their Core-Based Statistical Areas (CBSA)⁴ in order for them to be included in the scoring process. We accomplished this by cross referencing the event geographic data against ZIPList5 CBSA data [13].

3.3 INTEGRATED DATABASE

The quality of the data chosen was extremely important to the feasibility of this methodology to avoid the common garbage in garbage out phenomena which would ultimately skew results. The data sources shown in Table 1 are not exhaustive, but are indicative of how much information we could reliably access and collect within the time frame of the study. Future iterations of this project would benefit from more granular clean energy innovation data.

The integrated database that underpins the Ecosystem Discovery tool contains all of the data that was collected, even though only the geocoded data is used for scoring / mapping purposes. The list of data sets (which also illustrates which datasets have technological characteristics and which used knowledge discovery tools to derive) is shown in Table 1.

Table 1. Ecosystem Component Data Sources

Ecosystem Component	Dataset Source	Rows of Data Events	Data Range	Technology specific	Data Mining/ NLP/ Machine Learning
Nascent Clean Energy Indicator	I3 small and medium companies	3,859	1900-present	X	
Nascent Clean Energy Indicator	ARPA-E 2016 projects	484	2016	X	X

⁴ U.S. government *Core-Based Statistical Areas* (CBSAs) are defined by the Office of Management and Budget. They are geographic locations neighboring urban areas of at least 10,000 people and/or are socioeconomically tied to the urban center by commuting. Prior to year 2000, these were collectively known as Metropolitan Statistical Areas and Micropolitan Statistical Areas (MSAs). The previous MSA codes covered about 80 percent of the country, but the new codes cover about 93 percent. An exclusive feature of Great Data CBSA Data is that the population statistics are calculated annually. Therefore, the CBSA rank is based on a more recent population than typical census data classifications.

Nascent Clean Energy Indicator	SBIR projects	395	Last 5 years	X	X
Nascent Clean Energy Indicator	Patent Ingest	13,939	Last 3 years	X	X
Nascent Clean Energy Indicator	Publication Ingest	35,347	Last 3 years	X	X
Nascent Clean Energy Indicator	News Ingest	10,041	No time constraint	X	X
Investors	I3 Rounds of financing	5,082	1900-present	X	
Investors	Clean Energy Investment Center (CEIC) list	83	1 year		
Enabling Environment	Clean Edge (US Clean Tech Policy Index 2016)	945	2016		
Networking Assets	Carnegie Classification list of higher and highest universities	222	2016		
Networking Assets	Labs (DOE & other agencies)	343	2016		
Networking Assets	Incubators (NIICE list)	32	2015		
Large Companies	I3 Large companies	67	1900-present		
Large Companies	Subset of Russel 1000 list	329	2015		

3.4 DATA ANALYSIS AND VISUALIZATION

Tableau Desktop was used to geographically display the results because Tableau gives users the ability to perform complex data visualization in an intuitive, drag and drop manner. Its data visualization is highly interactive and the user can highlight sections and drill down into other visualization charts with ease. The following dashboards were developed.

Figure 4. Companies Map Dashboard	Figure 5. Investor Map Dashboard
--	---

3.4.1 Companies Map Dashboard

The Companies Map dashboard (figure 4) graphically display the 3,625 small and medium sized company data from i3 that were able to be geolocated on a map of the US. It allows users to select any subset of the 14 different technologies. The Companies map visually represents the technology characteristic by color where, for example, pink represents energy efficiency and yellow represents solar. The intensity or size of the representative dot can also vary based on the following attributes: number of employees or total paid in capital. In the map view, this information can also be viewed as time series data where one can use a scroll to change the year and watch the companies grow over time. The tree map view of the companies in the bottom of the Companies Map dashboard is where the user can select any number of technologies and get a view of how “large” each company is in graphical form, represented by the size of the box, measured either by number of employees or paid in capital. This representation helps the user compare how much capital or how many employees a given technology sector has in comparison to another.

3.4.2 Investors Map Dashboard

The Investors Map Dashboard (figure 5) is composed of the i3 Companies Map, the Investors Map, the Investments bar chart and the Number of companies. By using the lasso feature, the user can select the

number of companies from the Companies map it wants investor information about. The investor information for the companies selected will appear in the Investor Map in the bottom Right side of the Dashboard. On the bottom left side of the dashboard, the investment bar chart displays the total amount of funding for the investors displayed in the Investor map in the bottom left side. The number of companies for which information is being given is displayed in the top right corner.

3.4.3 Events Map Dashboard

The Events Map worksheet (Figure 6) aggregates all of the events of all of the other worksheets in one place and assigns a different color for each ecosystem component and different shape for each type of event. The benefit of this view is that the user could select various types of entities and visually see its relationship to one another. For example, one could select a university and energy storage patents to see how much a university has patented this technology.

This dashboard offers two other visualization options of the data. On the bottom right side of the dashboard, the tree map displays the entities chosen as a hierarchical structure. It is color coded based on the IE category and the size of the rectangles represents the number of employees if the event type is small and medium companies, Investment amounts if the event type is i3 investors. On the bottom left side of the dashboard, the table view displays a table with the underlying raw data. This data can also be exported to excel for further use by the user.

3.4.4 IE State Dashboard

The IE State Dashboard (Figure 7) aggregates the results of the CBSA Dashboard to the state level to rank which state has the most IEs, for any chosen technology or combination of technologies. Much like the CBSA map, the user can select one or more of the 14 primary technologies (right panel), or could focus on the strength of any one of the five ecosystem components (Nascent Clean Energy Indicator, Financing Mechanisms, Enabling environment⁵, Networking Assets or Large Companies)⁶ or the overlapping score, and the map will dynamically update itself with the IE rank displayed on each state with a filled color. This worksheet maps the states ranked by the number of IEs (which range from highest activity in blue to lowest activity in yellow) as a filled map with a diverging color scale.

3.4.5 Publications, Patents, SBIR and ARPA-E Project Maps

The tool allows creation of multiple maps such as the Publication, Patents, SBIR and ARPA-E maps geographically maps these data events on a US Map. The view allows the user to choose specific technologies (which are color coded) and the year of interest. The timestamp slider can be used to

⁵ Since the enabling environment, networking bin features has no technology categorization, the overlapping score and the enabling environment scoring and ranking works only when the primary sector “NA” is selected.

⁶ For all the event data points that do not have the primary technology identified have been mapped with a “No data” primary sector name.

visualize the publications as a time animation. By hovering over each event, the user can see specific attributes about the data event points.

3.5 SCORING CLEAN ENERGY INNOVATION ECOSYSTEMS

The team developed an algorithm that could capture the existence and strength of the five ecosystem components. The result is an IE score for a specific region r , where r is one of the 945 Core-Based Statistical Areas (CBSA) codes in the US. The program computes an Ecosystem Component (EC) score for each region based on the sum of the number of events in that EC (See Table 1), except the enabling environment⁷; the program computes an EC score for *nascent clean energy indicators, investors and financing mechanisms, networking assets and large companies*. The weightings of the subcomponents of each EC can be set by the user, but must sum to 1. Once the five EC scores have been computed for each region, the EC scores are used to compute the IE score. The final IE score of a region r is computed from the weighted summation of the five EC scores, where the weighting can be determined by the user. The final score is normalized to 1.

3.6 CBSA MAP DASHBOARD

The results of the IE scoring algorithm are graphically displayed in the CBSA Map dashboard. The worksheet maps all the IEs ranked in the US for each CBSA region as a filled map with the diverging color scale to reflect its intensity measure of how much activity the IE has (i.e. which has the most data events). The user can select one or more of the 14 primary technologies (right panel), or could focus on the strength of any one of the five ecosystem components or the overlapping score, and the map dynamically updates itself with the IE rank displayed on each CBSA region with a filled color.

Figure 8: CBSA Map Dashboard

⁷ The enabling environment score is determined by the Clean Energy Policy index score and is measured at the state level so all CBSA belonging to a state receive the same score as the state.

The user can choose the range of ecosystems from the top 10 ecosystems to IEs that are ranked between 100 and 200. The IE Ranking sliding filter allows the user to filter the range of interest. On the bottom right corner of the screen there is a table with the top 15 IEs as well as their CBSA Codes and ranking.

The dashboard offers more granular information about each ecosystem in a number of other ways. By hovering over the ecosystem in the US Map in the top center, a popup display appears which gives the following features and scores: CBSA Rank, CBSA Code, CBSA Name, Avg. Population of the CBSA, Number of Companies, Employees, SBIR, Patents, Publications, ARPA-E Projects, Investors, Large Companies, Lab Agencies and Universities and, Clean Edge Policy Index Score. In addition, when a user clicks on an IE in the US map in the top center of the screen, an Event Map of the chosen IE is displayed in the bottom half of the screen to display the Events for the chosen IE. Selection of any CBSA on the map provides a drill down view of the location and details of the events that contributed to the computed overlapping score. On the bottom left side of the screen there is a statistics tables which provides a list of the description of the events that are viewable in the Events Map. All of the information that underlies any image can also be exported to the user in the form of an excel sheet.

4 RESULTS

This section will provide examples of output from the Ecosystem Discovery tool. As a caveat, the results of the output will differ based on the assumptions made for the weightings of the various categories. The

goal of the study was to see whether this methodology would work, but not to definitively determine where each ecosystem is in the US. Therefore, there is no assumption that the following examples are correct, but are simply used to illustrate the capabilities of the tool.

The weightings of each component can be changed, but for this exercise, the weights were kept as close to equal as possible. There are three categories where the weightings can be changed - *nascent clean energy indicators*, *networking assets* and the weightings of each of the *Ecosystem Components*. For the three categories where weights could be chosen, the following assumptions were made to test the tool:

Table 2: Weightings Used for Scoring Algorithm

Category	Dataset	Weighting
Nascent Clean Energy Indicators	Employee	10
Nascent Clean Energy Indicators	Company	30
Nascent Clean Energy Indicators	Patent	15
Nascent Clean Energy Indicators	Publication	15
Nascent Clean Energy Indicators	SBIR	15
Nascent Clean Energy Indicators	ARPA-E	15
	Nascent Clean Energy Indicators Total	100%
Networking Assets	Lab agencies	33
Networking Assets	Universities	34
Networking Assets	Incubators	33
	Networking Assets Total	100%
Ecosystem Components	Nascent Clean Energy	30
Ecosystem Components	Investors and Financing Mechanisms	30
Ecosystem Components	Enabling Environment	15
Ecosystem Components	Large Companies	5
Ecosystem Components	Networking Assets	20
	Ecosystem Components Total	100%

With the weights described above and with the “no data” option⁸ chosen, the results for the top 15 clean energy IEs for the technology foci of this project (See Figure 9) were consistent with the literature and the results one would expect from such an exercise. Silicon Valley is well known for its high rate of commercialization and consistently ranks high across all technologies through the IE tool. The tristate area in NY, given its population density and the density of financial firms also scores high⁹. The Boston-Cambridge-Newton, MA-NH CBSA around MIT and Route 128 consistently rank high. It is encouraging that the tool is functioning at base level to be able to detect the “well-known” ecosystems. Surprisingly, despite their small size, the Boulder and Denver CBSAs near the National Renewable Energy Lab also tended to score in the top 15 across a number of renewable technologies.

Figure 9: Top 15 IEs for all technologies

As a precursor to this work, an energy analyst in EPSA was asked to develop a short list of top US clean energy IEs. Through secondary research, primarily literature reviews and interviews with experts, she was able to identify 22 IEs of varying sizes and technology foci over the course of eight weeks of work. Our work sought to understand over the course of three months of work whether data characterization and manipulations could allow a computer to reliably choose as many or more IE’s by relying solely on a data ingest pipeline of relevant sources. The Ecosystem Discovery tool picked up at least 91% of the ecosystems that the human analyst was able to report. There were 2 out of 22 that were not detected. One was not detected because the human analysts’ definition of an ecosystem was too broad to be measured at the CBSA level (so was not truly an error of the tool, in this regard); if the tool is considered more precise than the human analyst than the hit rate would actually be over 95% for the Ecosystem

⁸ The tool give the option of choosing “No data” which means that Nascent Clean Energy Indicators and Investors that could not be categorized according to technology are included in the scoring algorithm.

⁹ Demographic characteristics which may skew the data were not normalized for. For example, the question of whether the NY-NJ-CT tristate area is truly innovative or whether it simply has the most investment activity and people and therefore has an absolutely high level of activity vs. a relatively high level has not been addressed. However, because the team chose to use CBSA’s as their underlying method of ecosystem discovery, this question can be addressed external to the tool by simply taking the IE results and dividing by relevant demographic information such as population, educational attainment levels and access to finance measures.

Discovery tool. A potential next step could be to have a few analysts look at some of these sectors and verify the location of specific IEs, so that we could control for intercoder variation.

Next, the team met with experts in solar, storage and nuclear technology to ask which IEs they would use to check the tool. The system could be evaluated on both precision and recall using such a framework by finding Toledo, OH, Austin, TX and St. Louis, MO, respectively, for these three technologies. The results from the solar example is below and offers a number of insights into the tool's capabilities.

5 DISCUSSION

The tool uses data to determine where ecosystems exist and the underlying scoring algorithm “ranks” the strength of the ecosystem. The user could choose the range of ecosystems that they want to see such as the top 15 in solar (See Figure 10) or the top 125 (See Figure 11). The user could also drill down on a certain region and get a more granular view of the surrounding CBSAs (See Figure 12).

From Figure 10, the top ecosystems are those that would be expected from such an exercise - 1) San Francisco, CA, 2) NY-Newark-Jersey City, NY-NJ, 3) San Jose-Sunnyvale-Santa Clara-CA, 4) Washington-Arlington-Alexandria, DC-VA-MD, 5) Boston-Cambridge-Newtown, MA-NH - as discussed above. However, in Figure 11, once the search is opened up to the top 125, the individual CBSA level ecosystems start to form their own clusters in various parts of the United States, not only in the East and West coast. In fact, a drill down on the CBSA's in OH in Figure 12 reveals that #107 Cleveland-Elyria, OH, #124 Cincinnati, OH-KY-IN, and #125 Toledo, OH are all relatively close to one another.

Figure 12: Solar IEs around OH region

Figure 13: IE Example: Toledo, OH

Toledo, OH is not necessarily well known for solar technology, but as shown in Figure 13, it is home to 3 small-medium solar companies, 30 solar employees, 49 patents, 187 Publications, 3 investors, 2 large companies and 2 universities. The CBSA does not have any SBIRs, ARPA-E awards, lab agencies or other US DOE interventions, but its clean edge policy index is 46.4.

Toledo's earlier prominence as the glass capital of the world enabled subsequent expertise in photovoltaic (PV) related technology.¹⁰ Northwestern Ohio's automotive glass industry generated employees with transferable skills. Ohio was also home to the visionary innovator Harold McMaster.¹¹ Dr. McMaster founded Solar Cells, Inc., to develop thin-film cadmium telluride technology. Solar Cells was later bought and renamed First Solar, currently a world leader in thin-film PV. Investments in manufacturing and basic research at the University of Toledo and other institutions which catalyzed much of the research expertise that characterizes the region today. In fact, for two decades, the Wright Center for Photovoltaic Innovation and Commercialization at the University of Toledo has been funding basic energy research.

In addition, the investor owned utilities in the area are among the most expensive in the nation which increased an interest in alternative sources and made renewable energy more price competitive with grid power. Other enabling factors include the state policy environment and networking assets. Ohio's advanced energy portfolio standard, which mandates that at least 25 percent of Ohio's electricity come from clean and renewable sources by 2025, helped to advance a portfolio of clean energy technologies. In the Toledo IE, universities, industry and government worked closely together. The Ohio Advanced Energy business trade association has actively promoted solar statewide. The Ecosystem Discovery tool did not uncover any of these reasons explicitly, but by detecting that this ecosystem does exist, it allows the user to compare its origins to other more and less well known ecosystems for further research. [14]

Toledo, OH ranks low at 125, perhaps because its strength in the solar space declined after First Solar

¹⁰ Pioneering Toledo firms included Edward Ford Plate Glass Company (1899-1930), Toledo Glass Company (1895-1931), and Libbey-Owens Glass Company (1916-1933).

¹¹ Harold McMaster (1916-2003), became the first research physicist ever employed by Libbey Owens Ford Glass in Toledo in 1939 and went on to found four glass companies. Dr. McMaster was once called "The Glass Genius" by Fortune magazine.

moved to Arizona, but its detection by the tool speaks to the importance of the user case or user benefit of this tool. This tool allows the user to discover IEs that could otherwise go under the radar but that have the potential to be top performers for investors or programmatic institutions.

6 CONCLUSION

The solar case study was based on an expert recommendation of a known ecosystem. The Ecosystem Discovery tool was initially validated by being able to find the known ecosystems, Toledo, OH for solar, Austin, TX for storage and St. Louis, MO for nuclear. While Austin, TX and even St. Louis, MO may have been otherwise detected by traditional means, the Toledo, OH example best illustrates the importance and use case of this tool. The Ecosystem Discovery tool can detect lesser known ecosystems with the potential to be top performers that could otherwise go under the radar.

In addition, by allowing the user to see the components of the ecosystem, the policy analyst using the tool can begin to understand what aspects of an ecosystem's composition contributed to its success. There are a number of uses that can be exploited by this flexible tool in addition to the ones that we originally intended to meet. This work can also support the understanding the effect of federal funding or private sector investments on IE activity.

The tool gives the user a replicable method to identify clean energy IEs and their characteristics. This proof of concept demonstrates that a machine-assisted methodology can help rapidly identify, quantify, and characterize clean energy ecosystems. The initial research which led to the definition of a clean energy IE as the composite of five broad ecosystem components has contributed to a generalized framework of for energy innovation clusters which drove the success of the methodology.

The ingest pipelines also revealed both the potential and limitations of an automatic machine extraction methodology to gather ecosystem component data. One of the major challenges of this work was access to relevant data. For example, the tool has some investor information, primarily from the i3 database, but additional data would help to refine results. With more time and resources, there is also an opportunity to use new NLP techniques such as document summarization and supervised machine learning on the news feed ingest to glean high quality, relevant information from public sources such as where technology activity happens vs. where a headquarter is located.

The tool can be refined by calibrating the weightings of ECs to match known entities. In addition, surprise results generated by the tool could be validated through other means. With further calibration and validation of known and previously unknown ecosystems, this tool will allow the user to rapidly and dynamically identify hundreds of ecosystems for any given technology.

ACKNOWLEDGEMENTS

We would like to thank Eric Hsieh (US DOE), Hugh Chen (US DOE), Hugh Ho (US DOE), John Jennings (US DOE), Jeff Dowd (US DOE), and Jeff Alexander (RTI) for their review and input. 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.

References

- [1] Wiley, "Too Big to Ignore: The Business Case for Big Data - Phil Simon," <http://www.wiley.com/WileyCDA/WileyTitle/productCd-1118638174.html>, 2012.
- [2] O. R. N. L. ORNL. [Online]. Available: <http://cda.ornl.gov/piranha.shtml>. [Accessed July 2016].
- [3] S. Howell, "Financing Constraints as Barriers to Innovation: Evidence from R&D Grants to Energy Startups," Harvard University, Cambridge, 2015.
- [4] "Harvard Business School: Institute for Strategy and Competitiveness," [Online]. Available: <http://www.isc.hbs.edu/competitiveness-economic-development/frameworks-and-key-concepts/Pages/clusters.aspx>. [Accessed 24 March 2016].
- [5] A. Regalado, "The Next Silicon Valley," *MIT Technology Review*, vol. 116, no. No 5, September / October 2013.
- [6] B. a. J. W. Katz, "The Rise of Innovation Districts: A New Geography of Innovation in America," Brookings, Washington, DC, May 2014.
- [7] C. Jackson, "Lessons from the Kauffman Index: Immigrants are Infused with Entrepreneurial Energy," Kauffman Foundation, June 2015.
- [8] "Immigration and America's High Tech Industry: The jobs machine," *Economist*, April 2013.
- [9] T. Corsatea, "Localized Knowledge, Local Priorities and Regional Innovation Activity for Renewable Energy Technologies: Evidence from Italy," *Regional Science*, 2014.
- [10] "Piranha (software)," Wikipedia, the free encyclopedia, [Online]. Available: [https://en.wikipedia.org/wiki/Piranha_\(software\)](https://en.wikipedia.org/wiki/Piranha_(software)). [Accessed July 20 2016].
- [11] D. Google, "Developer's Guide: Google Maps Geocoding API," [Online]. Available: <https://developers.google.com/maps/documentation/geocoding/intro>. [Accessed 21 July 2016].
- [12] D. Google, "Google Places API," [Online]. Available: <https://developers.google.com/places/>. [Accessed 21 July 2016].
- [13] "ZIPLIST5 CBSA TM (August 2016 data)," [Online]. Available: <https://www.zipinfo.com/products/z5cbsa/z5cbsa.htm>. [Accessed 22 July 2016].
- [14] C. W. Wessner, "Committee on Competing in the 21st Century: Best Practice in State and Regional Innovation Initiatives," in *The Future of Photovoltaics Manufacturing in the United States: Summary of Two Symposia (2011)*, Washington, DC.

AUTHOR BIOS

Jessica Lin is a Senior ORISE Fellow to the US Department of Energy in the Innovation Department of the Office of Energy Policy and System Analysis. Prior to her current role, she worked on international energy projects with the World Bank Group, primarily focusing on what role the Bank could play to commercialize clean energy technologies. She started her career in the private sector, working in finance and strategic management consulting, before leading a wireless mesh technology social enterprise startup. She continues to serve as a thought leader on issues of early stage finance and clean energy commercialization for the World Bank's Climate Technology Program. She received her Master's in Public Policy from the Harvard Kennedy School and her Bachelors of Science in Electrical Engineering and Computer Science from MIT.

Supriya Chinthavali has been working as a computer scientist in ORNL since March 2010. She has multiple years of experience working on US-DOE and DOD sponsored projects like VERDE (Visualizing Energy Resources Dynamically on Earth), SPIDERS, PFMT, etc. in the Energy domain at ORNL. Prior to ORN, she worked for Delphi Automotive systems as an Advanced Software Engineer and developed embedded software systems for Toyota and GM. Her research interests include data analysis and information visualization of electric grid and complex infrastructure systems and machine learning. She graduated with a B.S. in Electronics and Communication, M.S. in Automotive Embedded Systems from India (2008) and a M.S. in CSE from Georgia Institute of Technology (2015).

Chelsey Dunivan Stahl graduated from Morningside College in 2013 with a double major in Computer Science and Business Administration. She is currently a Data Analytics Software Engineer for the Computational Data Analytics Group at Oak Ridge National Laboratory focusing primarily in data mining and knowledge discovery.

Christopher Stahl graduated from Florida Southern College with a Bachelor of Science in 2011 and has been working with the Computational Data Analytics group at Oak Ridge National Laboratory ever since. He currently is a Data Analytics Software Engineer who works on multiple projects with a focus on data mining, data analytics, and knowledge discovery.

Sangkeun Matt Lee received the PhD degree in computer science and engineering from Seoul National University in 2012. He joined Oak Ridge National Laboratory as a post-doctoral research associate in 2013 and became a research staff member in 2015. His research interests include large-scale graph mining and analytics, big data systems and architectures, information retrieval and recommender systems, and their applications. He has experience of working on ORNL's Lab-directed Research & Development (LDRD), US-DOE and NIH sponsored projects and developing various software tools like gm-sparql (Graph Mining Using SPARQL) and ORiGAMI (Oak Ridge Graph Analytics for Medical Innovation).

Dr. Mallikarjun (Arjun) Shankar's research focuses on the interdisciplinary bridge between computer science and the basic and applied sciences. He is the Director of ORNL's Compute and Data Environment for Science (CADES) institute which hosts joint scientific initiatives between the computing directorate and ORNL's science focus areas including environmental and biological sciences, materials and neutron science, nuclear science, and the physical sciences.

Arjun received his Ph.D. in computer science from the University of Illinois, Urbana-Champaign, following which he worked in industry designing and building next-generation data management infrastructures. His research in the national laboratory setting has involved designing national scale sensor networking systems, energy infrastructure monitoring and control frameworks, and mechanisms to perform integrated data analytics and modeling and simulation. He is a member of the ACM and the IEEE Computer Society.