Chapter 8

Case Studies

These case studies are the results of the Information Visualization MOOC 2013 client projects. The students were asked to form groups of four to five and select a real-world project from a list of potential client projects. The clients made their data available to the students, who worked to conduct a requirement analysis, develop an early sketch, conduct a literature review, preprocess and clean the data, and finally perform analysis and visualization. The students then submitted their visualizations to the client for validation. The following chapter highlights six of these projects, including feedback and insights from the clients.

Case Study #1

Understanding the Diffusion of Non-Emergency Call Systems

CLIENT:
John C. O'Byrne [jobyrne4@gmail.com]
Virginia Tech

TEAM MEMBERS:
Bonnie L. Layton [bllayton@indiana.edu]
Steve C. Layton [stlayton@indiana.edu]
James S. True [jitrue@iu.edu]
Indiana University

PROJECT DETAILS

Local governments throughout the United States began adopting non-emergency call systems in the late 1990s. Known commonly as "311 systems," they free 911 emergency systems from being overloaded, provide citizens with a single number to call for requests, increase bureaucratic efficiency, and give citizens the opportunity to participate in local government. Virginia Tech Center for Public Administration and Policy doctoral candidate John O'Byrne's dissertation studied the factors influencing the adoption of 311 systems. His data included the years 1996 to 2012 when large U.S. cities implemented 311. He examined various factors that encouraged adoption: the size of the cities' populations, their forms of government, and their crime rates. A team of Indiana University (IU) informatics graduate students designed four versions of geospatial visualizations showing the 311 diffusion across the United States. They created the versions to explore differences between the level of user engagement, sense making, and retention with touchscreen interactivity to that of print and online visualizations.

REQUIREMENTS ANALYSIS

O'Byrne, the project client, requested a visualization that would indicate when the rate of 311 adoption reached a "critical mass" large enough to encourage other cities to follow suit. He wanted to show how closely the 311 data compared with Everett Rogers's Diffusion of Innovation curve, which was modeled on hundreds of innovation-focused studies.[1] These studies have tracked acceptance of innovations in many disciplines, including science, medicine, technology, and sociology. The team proposed combining two datasets in a year-by-year bar graph indicating the cumulative adoption of cities with proportional symbols (circles) below each bar to represent how many had adopted that year.

O'Byrne had created an extensive database of cities' populations, crime rates, and forms of government. The team proposed three maps that would visualize the researcher's hypotheses that

1. cities with larger populations would adopt 311 faster,
2. cities with higher crime rates would adopt it faster, and that
3. cities with mayor-council governments would adopt it faster.

The team created static, animated, and interactive versions of the maps to compare their effectiveness.

RELATED WORK

The visualization team incorporated visual-processing principles and theories postulated by Stuart Card, Jock Mackinlay, and Ben Shneiderman. Card and Mackinlay define visualization attributes as being constructed from (1) marks, such as points, lines, surface, area, and volume; (2) their graphical properties; and (3) elements requiring human-controlled processing, such as text.[2] They delineate two kinds of human visual processing as "automatic," in which users identify properties such as color and position using highly parallel processing but are limited in power, in contrast with "controlled processing" in tasks such as reading that have powerful operations but are limited in capacity. These classifications were useful in breaking down the task of visualizing the geospatial and temporal data of the U.S. map and bar chart (automatic) in combination with the trend data consisting of text explaining each of the adoption factors. Shneiderman's basic-principles mantra of "overview first, zoom and filter, then details on demand" guided the team in creating an adoption overview first (the S-curve and proportional-symbol map) and allowing users to click (filter) to isolate each adoption factor.[3]

[1] Rogers, E.M. 2003. *Diffusion of Innovations*, 34, 344–347. New York: Free Press.

[2] Card, S.K., and J. Mackinlay. 1997. "The Structure of the Information Visualization Design Space." In *Proceedings of the IEEE Symposium on Information Visualization*, 92–99. Washington, DC: IEEE Computer Society.

[3] Shneiderman, B. 1996. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations." In *Proceedings of the IEEE Symposium on Visual Languages*, 336–343. Washington, DC: IEEE Computer Society.

THE DIFFUSION OF 311

In 1996, U.S. cities began adopting 311 call systems to lower the number of non-emergency calls being made to 911. Local governments also wanted to provide a more efficient way for citizens to contact various city departments to handle simple complaints like potholes.

Researchers studied the adoption process of cities through the "diffusion of innovation" theory. "Diffusion" is a process by which an invention or a new way of doing things is communicated and then adopted.

The years they examined began with the first year 311 systems were introduced until the "critical mass point" of 311 adoptions. Based on the diffusion of innovation theory, the critical mass point — when enough city governments adopted systems so that the further rate of adoption became self-sustaining — occurred in 2003.

The following graphics explain what researchers gleaned through the process:

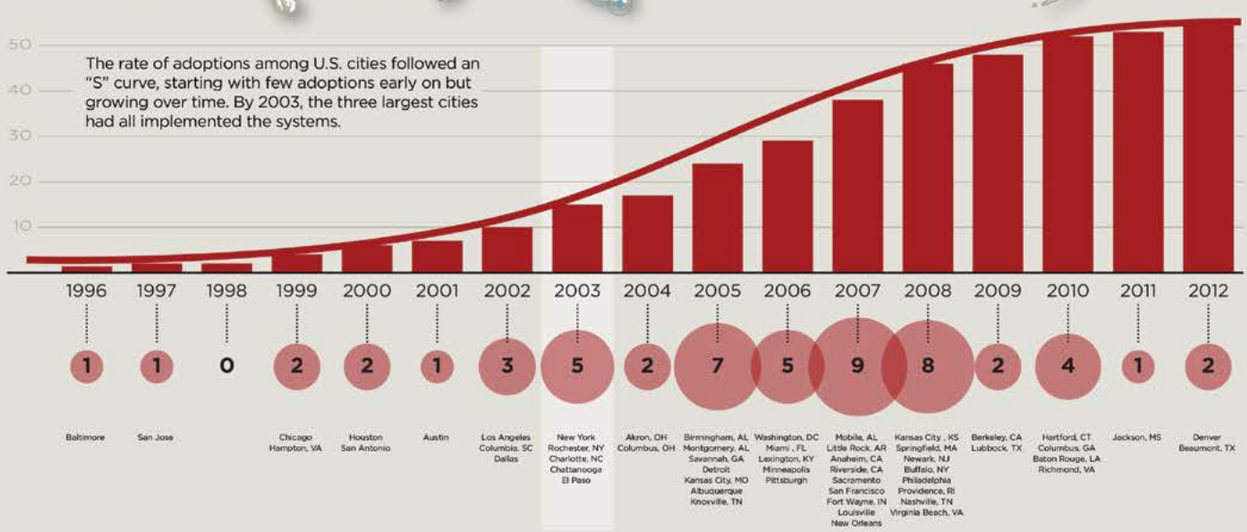
BY POPULATION

Researchers accurately predicted that larger cities would adopt 311 systems earlier. They assumed they would want to reduce the burden of 911 calls they were receiving. For example, after adopting 311, Baltimore reduced 911 calls by almost 5,000 a week.

Adoptions: **55** Non-adoptions: **183**

(11 cities use 311 at the county level)

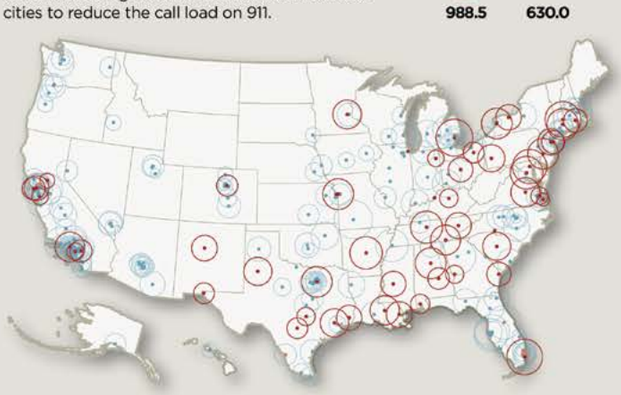
Average city populations: **685,160** **215,173**



BY VIOLENT CRIME RATE

Researchers also accurately predicted that cities with higher crime rates would adopt earlier. They assumed the higher crime rates would motivate cities to reduce the call load on 911.

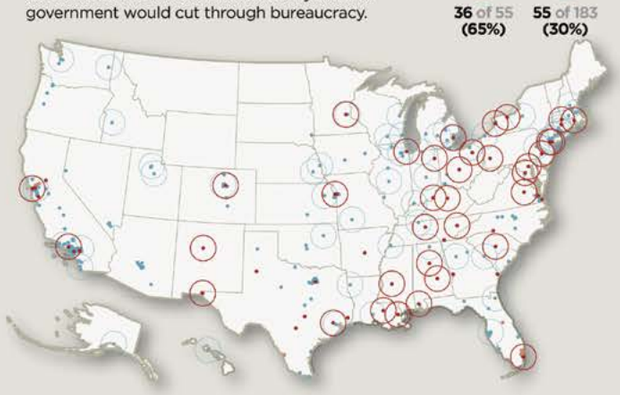
Average violent crime rate:* **988.5** **630.0**



BY MAYORAL SYSTEM

Researchers predicted that mayoral-council forms of government would adopt 311 systems earlier, because a more centralized-authority form of government would cut through bureaucracy.

Mayor-council systems **36 of 55 (65%)** **55 of 183 (30%)**



SOURCES: U.S. Census Bureau 2003 population estimates, Federal Bureau of Investigation Uniform Crime Reporting Statistics, International City/County Management Association, 911dispatch.com, CNS (cns.iu.edu)

Figure 8.1 The Diffusion of 311 (http://cns.iu.edu/ivmoocbook14/8.1.jpg)

DATA COLLECTION AND PREPARATION

The visualization team used O'Byrne's database of cities as the basis for maps and adoption curve visualization. The team generated latitude and longitude data to plot the cities, incorporated census figures, and generated vector files of the municipal data from the proportion symbol map tool in the Sci2 visualization software.[4] The files were brought into Adobe Illustrator to customize the colors and lines and exported into Adobe InDesign. By using InDesign's Digital Publishing Suite overlays, the team produced animations showing each city as they adopted 311 systems. The interactive online visualization that incorporated buttons was created using HTML and the JQuery JavaScript framework. For the touchscreen visualization, the team created an iPad version using InDesign.

ANALYSIS AND VISUALIZATION

The team recognized the complexity of the data and determined it would be useful to compare a static visualization with animated versions using the same data. They separated each factor (population, crime rate, government form) into discrete temporal visualizations. Furthermore, the team recognized that understanding the visualization would require both "automatic" and "controlled" processing, which they assumed would be more efficiently processed through a self-paced, interactive experience. The four visualizations consisted of the mouse-controlled, self-paced version, the non-interactive animated version, the touchscreen-controlled, self-paced version, and the static version (Figure 8.1).

Although team members sent all four versions to O'Byrne, they concentrated their questions on the print version, since he planned to insert it into his dissertation.

O'Byrne's feedback overall was extremely positive. He said the organization and hierarchy of the visuals reflected their relative importance and encouraged proper eyeflow from the center to the right column leading down and below to the left. When asked whether he would have preferred using the city population bar chart at bottom right as the dominant element instead of the map, he said he thought New York's extremely tall bar would create awkward trapped space, so he thought it should stay as a secondary element.

In regard to the color palette, the client preferred the three-color version, although he said he would be receptive to using a map in which the blue "non-adopters" were shown in green, the complement of red.

The client recognized the diffusion of innovation adoption rate bar chart in the lower left would require deeper reading concentration (Shneiderman's "controlled processing") because of its multivariate nature. But he said pairing the proportion symbols below the bars and labeling the cities would not complicate the information too much.

[4] Sci2 Team. 2009. "Science of Science (Sci2) Tool." Indiana University and SciTech Strategies. http://sci2.cns.iu.edu.

Since the client's main objective is to use the visualization in his dissertation, he said that he would prefer the maps and charts broken into separate figures for inclusion, but would use the design intact for use at a conference poster session. The dissertation version will have to be in black and white, so the team agreed to send him each map separately.

DISCUSSION

The maps in conjunction with the proportion symbol comparison clarify and support the client's three hypotheses. In the first map, especially, it's clear at a glance that the larger populations are colored red (adopters). If readers study the proportion symbol comparison labels, they see that the average city population of adopters is 685,160 compared to a much smaller average population average of 215,173 for non-adopters. The actual adoption rate of cities aligns closely with Rogers's Diffusion of Innovation curve, which the client also hypothesized. The bar chart at bottom right supports visually the hypothesis that the largest cities adopted earlier.

The team was pleased with the client's reaction. The complexity of the data presents a large challenge in creating a static visualization. In order to create a readable, aesthetically pleasing print version, the team needed to use at minimum a tabloid size, although a poster version of 2′ × 3′ (.609 × .914 meters) allowed viewers to see details with more clarity. The print version also lacks the opportunity to layer information through rollovers and taps the way the online/touchscreen versions do. The team feels the scalability of the interactive model would not be constrained by geography (e.g., creating a world map of diffusions versus limiting it to the United States). However, it's still necessary to assess the processing demands on users. After evaluating client feedback during the initial prototyping process, the team realized the complexity of some information and attempted to simplify further. In a future iteration, the team would like to layer more data on the interactive version to encourage more interactivity (e.g., a graphic that could allow users to zoom into any municipality and see a choropleth map of population breakdown with rollover information about the city's adoption). Shneiderman's principles underline the importance of providing an overview while allowing users to drill down. In trying to visualize geospatial/temporal/quantitative 311 data most efficiently, the interactive versions seemed better suited.

ACKNOWLEDGMENTS

The authors would like to recognize the help and assistance of Dr. Katy Börner, director of the Cyberinfrastructure for Network Science Center at Indiana University, School of Informatics and Computing, Department of Information and Library Science doctoral student Scott E. Weingart, and CNS research and editorial assistant David E. Polley.

Case Study #2

Examining the Success of *World of Warcraft* Game Player Activity

TEAM MEMBER:
Arul Jeyaseelan [ajeyasee@indiana.edu]
Indiana University

TEAM MEMBER/CLIENT:
Isaac Knowles [iknowles@indiana.edu]
Indiana University

PROJECT DETAILS

In the video game development and publication industries, analytics and visualization are increasingly used to drive design and marketing decisions. Especially for games that are based in large virtual worlds, a game's "virtual geography" is a matter of significant importance. The costs and benefits of moving to one place or another bear enormously on player decisions. Specialized tools are needed to help analysts and developers understand how players respond to their virtual surroundings, and how well those surroundings engage players.

Case in point: Activision Blizzard's *World of Warcraft* (WoW). WoW is a massively multiplayer online role-playing game based in an immense virtual world. In the game, players may traverse four continents, each with its own unique geographical features, challenges, transit routes, and economic centers. Some parts of the virtual world are visited more than others, but every place must meet strict quality standards. Underused areas represent losses on investments, while overcrowded regions can cause server or client crashes. Thus geospatial analyses are vital to identifying and generating solutions to these and other issues.

To that end, we built a tool that helps users understand the movements of players through the World of Warcraft. Using a map of that world, we created a geographical information system for the game, which allows users to explore the travel behavior of players in an underlying dataset. The user can pull up and compare player trends across many geographical locations in the game. The tool helps answer questions like:

1. What are the major travel hubs?
2. How do travel habits change over time?
3. How do location and travel habits vary with server population and in-game zone population density?

The clients for this project were Isaac Knowles (also a participant) and Edward Castronova, two economists at Indiana University whose research focuses on the economies of virtual worlds. In an effort to better understand how people work together in groups, our clients had collected many thousands of "snapshots" of player activities inside WoW. These players were all members of competitive raiding guilds. In WoW, guilds are formally recognized player groups. Some of these groups take part in "raids," which are comprised of a series of battles against computer-controlled monsters, and which require the coordination of a large number of players to complete successfully. The world's highest ranked raiding guilds were chosen by Knowles and Castronova for their study.

REQUIREMENTS ANALYSIS

The initial client requirement was very broad, emphasizing only that our end visualization should help discern patterns and trends in the data. Consequently, our idea of building a customized visualization tool, rather than a single visualization, was met with the clients' considerable interest and approval. During the design phase, we were fortunate in that the client-participant was able to react to the group's ideas and suggest amplifications or changes on the fly. He also had a great deal of first-hand knowledge about the data we were using, which he had already cleaned and analyzed prior to beginning the project. This was a significant advantage and time saver.

RELATED WORK

The bulk of our work consisted of designing and deploying a new and unique tool for data exploration. In doing this, we created several visualizations to help users understand some of the basic geospatial and demographic patterns that are in the data. These visualizations are similar to those found in several previous works on *World of Warcraft*.

Christian Thurau and Christian Bauckhage investigated 1.4 million teams in *World of Warcraft*, spanning over four years.[1] They identified some of the social behaviors that distinguished guilds and analyzed how different behaviors affected how quickly guild members leveled up. They used histograms and bar charts to indicate the development of guilds both in the United States and European Union. In addition, they used network analysis to demonstrate how U.S. and EU guilds evolved over time.

[1] Thurau, C., and C. Bauckhage. 2010. "Analyzing the Evolution of Social Groups in World of Warcraft." Paper presented at the *IEEE Conference on Computational Intelligence and Games*, IT University of Copenhagen.

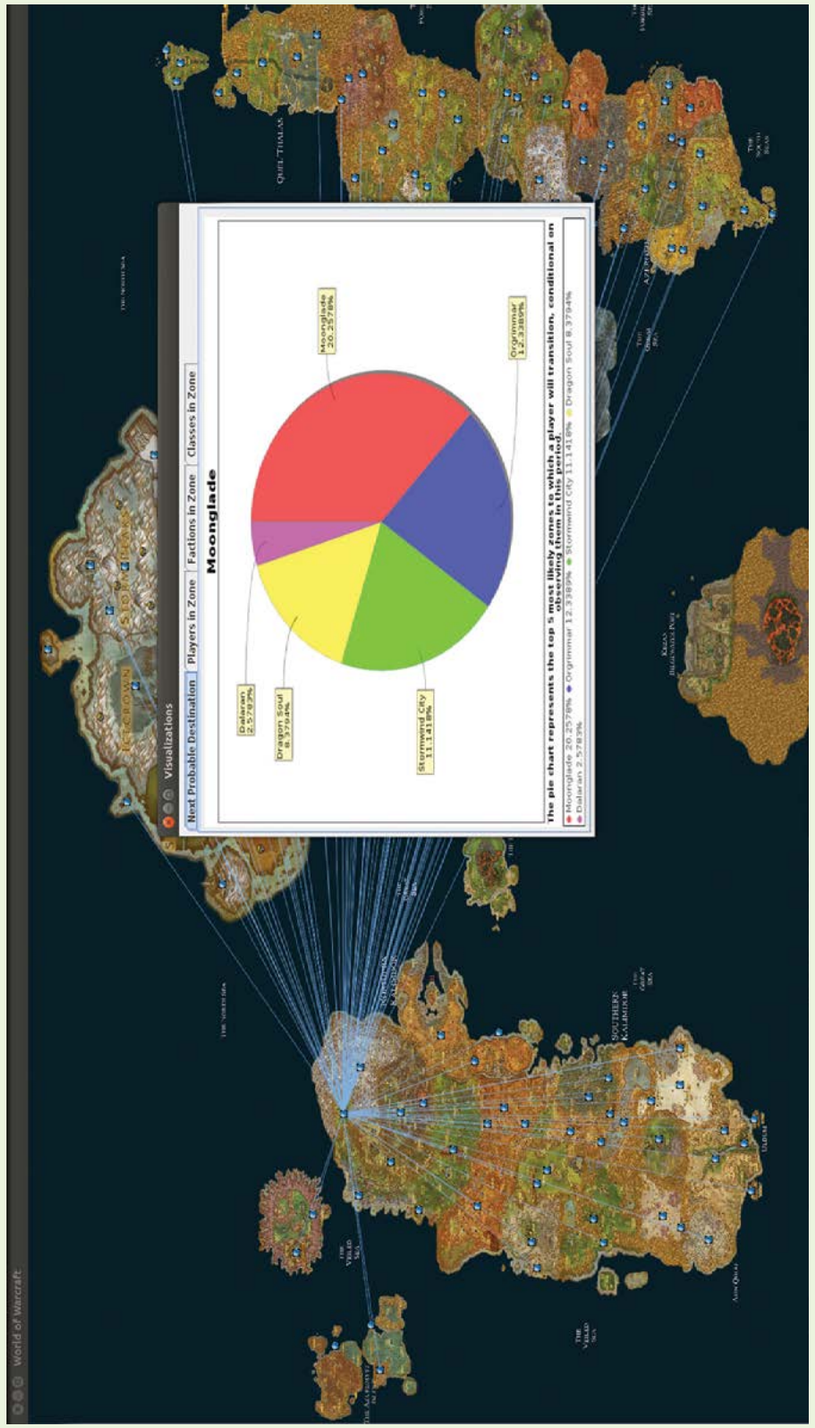


Figure 8.2 Final visualization showing lines of transit and next-location probabilities displayed for the Moonglade region of WoW (http://cns.iu.edu/ivmoocbook14/8.2.jpg)

Nicolas Ducheneaut et al discuss the relationship between online social networks and the real-world behavior in organizations in more depth.[2] They use histograms, bar charts, and line charts to do this. Besides the network analysis, they also use statistics to show the extent and nature of players' behavior in the virtual world. In another article, Ducheneaut et al analyze many of the broader trends of class, faction, and race choice; zone choice; and transit using data very similar to the one used here.[3] While their goal was to "reverse engineer" design patterns in WoW, our project aims to reveal patterns of behavior in WoW players.

DATA COLLECTION AND PREPARATION

The data were comprised of periodic snapshots of players' characters, including information about their in-game location, real-world location, guild membership, class, race, and language spoken. Players' chosen faction—which can be Horde or Alliance, and which limits player communication to that faction—was also recorded. Approximately 48,155 *characters* appear in the data; the number of *players* represented by those characters could be considerably smaller (30,000 to 40,000), as one player may take on multiple characters. Only players that were online at the time of a snapshot can appear in that snapshot. Snapshots were taken at roughly half-hour intervals, from December 2011 to May 2012, with some gaps due to game server outages or scraping program bugs. Table 8.1 contains basic information about the data, which required very little cleaning prior to loading into our server.

Table 8.1 Basic Data Facts

Region	Servers	Players	Guilds	Total Records
US	58	31,995	382	2,966,757
EU	59	18,520	107	2,272,390

VISUALIZATION

The final visualization consisted of an interactive, high-resolution map of the world. Owing to the size of the high-resolution map image file (9000 × 6710, 13MB) and the short duration of time available to produce a prototype, a web application format was abandoned in favor of a desktop client model that would have all resources prepackaged. The desktop client was developed as a Java application with the data on a MySQL server as the backend.

---

2 Ducheneaut, N., N. Yee, E. Nickell, and R.J. Moore. 2006. "'Alone Together?': Exploring the Social Dynamics of Massively Multiplayer Online Games." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 406–417. New York: ACM.

3 Ducheneaut, N., N. Yee, E. Nickell, and R.J. Moore. 2006. "Building an MMO with Mass Appeal: A Look at Gameplay in World of Warcraft." *Games and Culture* 1, 4: 281–317.

For the user, clicking on a location button resulted in two major events. First, it caused lines to be drawn from the chosen flag to every other flag on the map at which players were next seen in the data. Second, a new window popped up containing several tabs, which provided different visual breakdowns of the player-base that was found at a particular location. We used pie charts to display probabilistic information, line graphs for temporal population data, and bar graphs for comparing static player information.

The positions of all the zones with respect to the original full-scale image were maintained in a CSV file that the application would access at startup. For example, the city Orgrimmar can be found at pixel (2205, 3679) of the map. With this information, the map could be scaled to fit various monitor resolutions and still have each zone's position updated correctly. As we were dealing with more than 5 million records (see Table 8.1), performance was a major concern. Though optimized when possible, the application did face slowdowns when fetching information for locations that had heavy player traffic. Once fetched, the data from the various queries were used to generate the transit lines, and charts were generated using JFreeChart.[4]

DISCUSSION

Work continues on making a more user-friendly version of the tool. A web equivalent of the desktop application is currently being developed using Javascript with D3 for the front end and PHP with MySQL for the backend. The new version of the application will try to address some of the issues mentioned by our classmates and validators. Among these are the non-interactivity of links between zones, the difficulty of comparing information between zones, and the quality of the visualizations themselves; for example, the use of a pie chart to display information on where players travel. The primary challenge remains: finding an efficient way to serve the map image to the client.

Though our team was fortunate to have several skilled programmers, the interface was difficult to produce and complete. We thus faced limitations on the functionality of the interface at the time it was presented to our classmates and our clients. There are a few improvements to the tool that we continue to work on.

First, the information that is served by the tool is currently static; it provides no way to look at or compare data from different time periods. It is also not possible to look at smaller cross-sections of the data. For example, we could not separately visualize the data for particular servers, guilds, factions, players, or other groups within the data. To correct

[4] http://www.jfree.org/jfreechart

this, we will add interface elements to the visualization, as well as the capacity to generate suitable queries to get the information from the server.

Second, and more important, although making the tool more dynamic is a relatively minor technical task, making it easy to work with will require considerably more work. In its current form, for example, the only way to compare information at two locations or more is to call up their information windows and compare them side by side. While this is better than having no tool at all, small differences will be difficult to detect without visuals that more effectively afford comparison.

Third, though we do not specifically deal with the fact that our data are comprised of competitive guild members, in the future our interface could be modified to analyze guild success. For example, zone choice and transit speed may bear on the efficiency of a guild, and therefore its final place in the competition.

ACKNOWLEDGMENTS

The authors wish to thank their other team members (in alphabetical order): Shreyasee Chand, Zhichao Huo, Sameer Ravi, and Gabriel Zhou. Thanks also to Edward Castronova for his comments and for his support in producing the data we used. We also appreciate the advice and encouragement we received from Katy Börner, Scott E. Weingart, David E. Polley, and all of our Information Visualization classmates.

Case Study #3

Using Point of View Cameras to Study Student-Teacher Interactions

CLIENTS:
Adam V. Maltese [amalteste@indiana.edu]
Joshua Danish [jdanish@indiana.edu]
Indiana University

TEAM MEMBERS:
Michael P. Ginda [mginda@indiana.edu]
Tassie Gniady [ctgniady@indiana.edu]
Michael J. Boyles [mjboyles@iu.edu]
Indiana University

Laura E. Ridenour [ridenour@uwm.edu]
University of Wisconsin-Milwaukee

PROJECT DETAILS

The goal of this project and the resulting visualizations and analysis is to help researchers identify how they might understand student engagement within science, technology, engineering, and mathematics (STEM) classes in the future. More specifically, this project aims to see if there is a correlation between the activity of the instructor in large science lectures and the corresponding activities of the students by using cameras that record student actions from Point of View (POV) cameras instead of self-report or external observation in real time. The educational research clients did not provide a concrete research hypothesis but were interested to see and understand their data in new ways by means of data visualizations.

The team consisted of four individuals affiliated at the time with Indiana University. Michael Ginda, Tassie Gniady, and Laura Ridenour were graduate students within the School of Library and Information Science. Michael Boyles was also an Informatics graduate student and manager of the Advanced Visualization Lab at Indiana University. Our client was Dr. Adam Maltese from Indiana University's School of Education and his colleague, Dr. Joshua Danish.

REQUIREMENTS ANALYSIS

The researchers had used three methods to collect the data: video of the instructor teaching the class, POV cameras mounted on baseball caps and worn by students, and Livescribe Pens with audio.[1] Our group set out to determine what data is most useful and what should be collected and preprocessed, along with which methods of visual analysis are best suited for this problem. The clients wanted a sustainable workflow leading to visualizations their team could recreate, as well as a way to highlight student actions as they corresponded with instructor actions.

RELATED WORK

Student attrition within STEM programs is greatest within the first academic year of study. High attrition rates within lecture formats have been linked to a lack of student engagement with materials, instructors, and other students.[2,3] Large lecture settings can create impediments to involvement by limiting student conceptualization through active engagement. Conversely, lectures that utilize a conversational tone and a self-critical method that explain thought processes behind ideas and that examine mistakes have been shown to positively impact student engagement.[4,5]

Efforts to visualize student engagement with instructional materials have utilized student activity data derived from online courses using course management software and social networks.[6] Data visualizations derived from course management software have allowed instructors to temporally track student engagement with course materials, discussions, and performance, which can help instructors improve course materials and track student behavior and learning outcomes.[7]

1 Livescribe. 2007. Livescribe Echo Pen. http://www.livescribe.com/en-us/.

2 Gasiewski, J.A., M.K. Eagan, G.A. Garcia, S. Hurtado, and M.J. Chang. 2011. "From Gatekeeping to Engagement: A Multicontextual, Mixed Method Study of Student Academic Engagement in Introductory STEM Courses." *Research in Higher Education* 53, 2: 229–261. doi:10.1007/s11162-011-9247-y.

3 Gill, R. 2011. "Effective Strategies for Engaging Students in Large-Lecture, Nonmajors Science Courses." *Journal of College Science Teaching* 41, 2: 14–21.

4 Long, H.E., and J.T. Coldren,. 2006. "Interpersonal Influences in Large Lecture-Based Classes: A Socioinstructional Perspective." *College Teaching* 54, 2: 237–243

5 Milne, I. 2010. "A Sense of Wonder, Arising from Aesthetic Experiences, Should Be the Starting Point for Inquiry in Primary Science." *Science Education International* 21, 2: 102–115.

6 Badge, J.L., N.F.W. Saunders, and A.J. Cann. 2012. "Beyond Marks: New Tools to Visualise Student Engagement via Social Networks." *Research in Learning Technology* 20, 16283). doi:10.3402/rlt.v20i0/16283.

7 Mazza, R., and V. Dimitrova. 2004. "Visualising Student Tracking Data to Support Instructors in Web-Based Distance Education." *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters*, 154. New York: ACM Press. doi:10.1145/1013367.1013393.

In the area of semantic web research, semantic browsers have offered various fields robust tools to organize and visualize temporal categorical data. Semantic browsing visualizations have been applied to video conference recordings, such as distance learning lectures, through the indexing of participant activities to help user information retrieval needs.[8] Implementing this approach in a dynamic visualization environment, such as that provided by the Tableau Dashboard,[9] allows for both efficient and effective visualization.

DATA COLLECTION AND PREPARATION

In an effort to examine introductory science students' (overt) cognitive actions and how they adjust their attention in response to classroom activities, POV video data and pen recordings were collected from about 50 students over multiple class sessions in an organic chemistry course and an introductory biology course. The data we discuss in this paper come from a subset of data including videos from six students, pen data from seven students, and classroom video recordings from a single organic chemistry lecture. The screen captures below demonstrate the type of POV data collected from the participants. Student POV video and pen data were coded using a grounded, iterative approach to identify emerging trends. Instruction was coded with specific behaviors defined in the revised Teacher Dimensions Observation Protocol (TDOP).[10] Several deviations from the revised TDOP were used to more closely capture instructor behavior–student response dynamics.

In this study, we coded each instructor action independently and without regard to time durations. This is a deviation from the prescribed method for the use of TDOP codes, which recommends coding behaviors concurrently within two-minute time segments. We felt that in order to flesh out how the individual actions of the instructor induced a student response, it was important to monitor the class as a continuum. In this manner, we can more easily pinpoint the results of specific actions, as opposed to observing a grouped generalized course of behavior. It is important to note that this method of coding only allows for the use of one type of behavior code at a given instance. This results in small segments of action that if viewed as raw data may appear as if the instructor spent the majority of the class period randomly jumping from topic to topic. However, when a full timeline of the codes are compiled and presented as a continuum it becomes clearer that the instructor may be frequently switching between pedagogical techniques

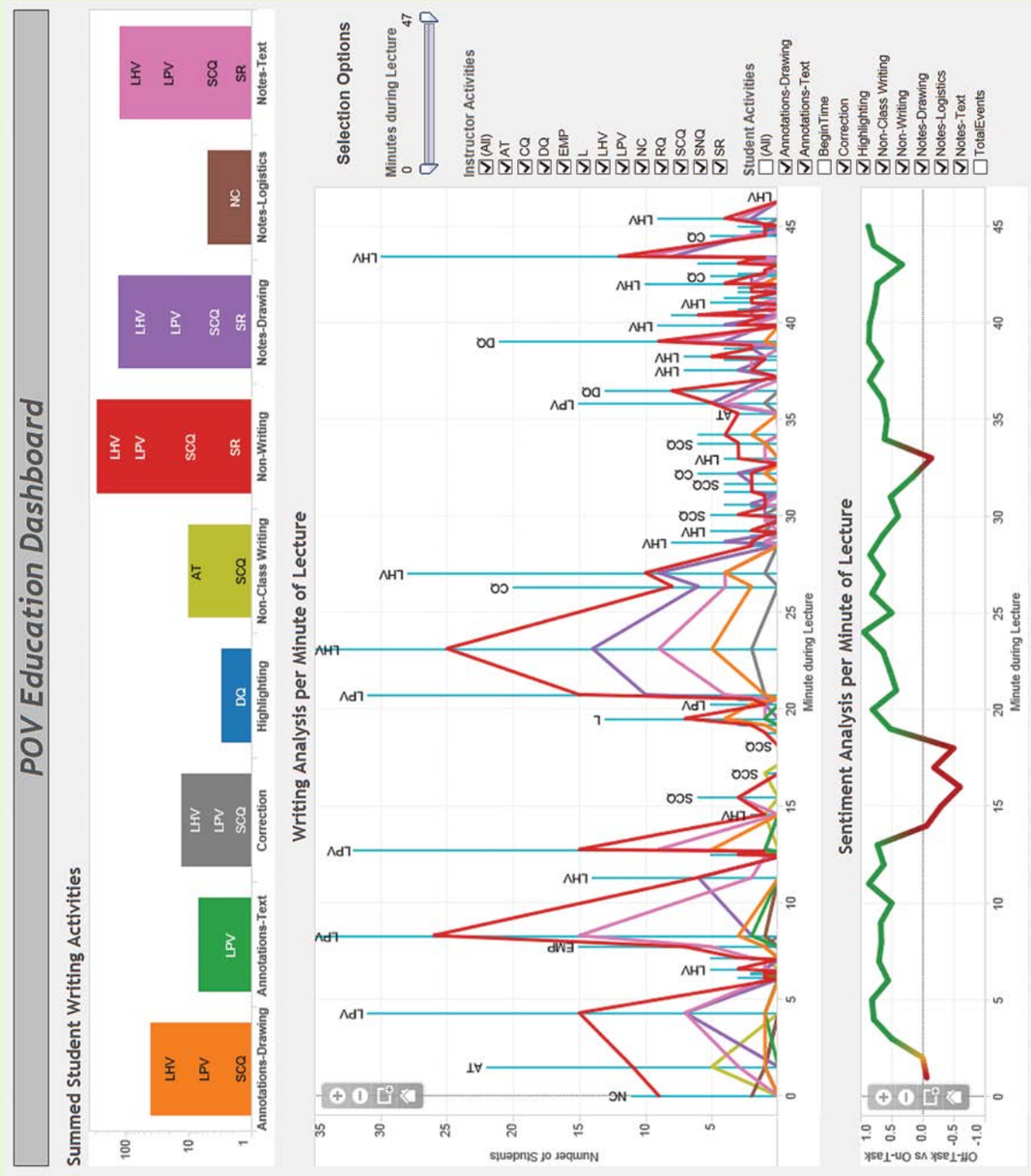


Figure 8.3 POV Education Dashboard created with Tableau Public (http://cns.iu.edu/ivmoocbook14/8.3.jpg)

[8] Martins, D.S., and M. da G.C. Pimentel. 2012. "Browsing Interaction Events in Recordings of Small Group Activities via Multimedia Operators." *Proceedings of the 18th Brazilian Symposium on Multimedia and the Web*, 245. New York: ACM Press. doi:10.1145/2382636.2382689.

[9] Tableau Software. 2013. Tableau Public. http://www.tableausoftware.com/public.

[10] Hora, M., and J. Ferrare. 2010. "The Teaching Dimensions Observation Protocol (TDOP)." Madison: University of Wisconsin-Madison, Wisconsin Center for Education Research.

that symbiotically flow together. Data coding was conducted using the ELAN qualitative software package.[11] Each different stream of data had the common element of an audio track. We were able to use the audio tracks to sync up all the video and pen recordings so that comparison across students and across data sources was reasonable.

Table 8.2 Instructor and Student-Writing Codes

Teacher Code	Meaning	Student Code	Meaning
AT	Administrative Task	1	Non-writing
CQ	Instructor Comprehension Question	2	Notes-Text
DQ	Instructor Display Question	3	Notes-Drawing
EMP	Emphasis	4	Notes-Logistics
L	Lecture	5	Annotations-Text
LHV	Lecture with Handwritten Visuals	6	Annotations-Drawing
LPV	Lecture with Pre-made Visuals	7	Correction
NC	No Code		
RQ	Instructor Rhetorical Question		
SCQ	Student Comprehension Question		
SNQ	Student Novel Question		
SR	Student Response		

All data were time stamped and coded using a coding system for student and instructor activities (see Table 8.2). The original coding scheme included 46 possible instructor actions of which 10 were used. Student codes were numerical, from 1 to 9, each with its own assigned meaning.

A variety of data preprocessing and reorganization was performed using Python and Excel. Instructor actions and individual student actions were programmatically combined into a single event timeline. Sentiment analysis for the six POV cameras was aggregated by minute of instruction and averaged to give an overall sentiment for each minute. Each of these are appropriate for analysis using the same 45-minute timeline.

ANALYSIS/VISUALIZATION

Using Tableau Public[12] an interactive aggregation of the POV data was created. Summed Student Writing Activities, Writing Analysis per Minute of Lecture, and Sentiment Analysis per Minute of Lecture are all displayed and may be manipulated to focus on any of the "Selection Options" on the right-hand side of the visualization via a list of checkboxes (Figure 8.3). The core of the dashboard facilitates both a collective overall view of student-writing activities as well as a time-based view of student writing and sentiments in relation to instructor activities.

Using the POV Education Dashboard, a researcher can see the interactions between student and instructor data by hovering the mouse over the different times in the Writing Analysis graph. More detailed exploration is enabled using the zoom-and-pan feature available in all three views. Instructor codes are linked between the student-writing overview and time-based views. Summarily, since select instructor and student-writing activities can be filtered out, the dashboard provides an extremely flexible environment for the quick exploration of a number of scenarios.

DISCUSSION

The project team identified and preprocessed both the instructor activities and student-writing codes into the following three categories: (1) no interactions between the instructor and students, (2) student-led interactions, and (3) instructor-led interactions. Coded student actions were aggregated for the duration of each set of instructor actions. The client can easily adjust these groupings as their needs and research ideas evolve while utilizing similar visualization techniques by choosing other criteria to group interactions.

Student engagement measured through PEN actions is shown to be higher when instructors pose questions broadly to students during the lecture. Conversely, student engagement is shown to be lower when their peers pose questions to the instructor during the lecture.

Further work with this data should explore the capabilities of Tableau for overlaying and comparing different types of STEM lectures and different instructor approaches. For example, a logical follow-up to our work would extend the existing dashboard to include tabs that link to the same data but provide many different views.

ACKNOWLEDGMENTS

The project team would like to thank Scott E. Weingart and David E. Polley for their technical assistance and expertise and Drs. Adam Maltese and Joshua Danish for the initial project idea and their willingness to review and work through many iterations as we converged on a final solution.

[11] Brugman, H. and A. Russel. 2004. "Annotating Multimedia/Multi-modal resources with ELAN." *Proceedings of LREC 2004*, 2065–2068. Fourth International Conference on Language Resources and Evaluation.

[12] http://bit.ly/YC7eY0

Case Study #4

Phylet: An Interactive Tree of Life Visualization

CLIENT:
Stephen Smith [blackrim@gmail.com]
University of Michigan

TEAM MEMBERS:
Gabriel Harp [gabrielharp@gmail.com]
Genocarta, San Francisco, CA

Mariano Cecowski [marianocecowski@gmail.com]
Ljubljana, Slovenia

Stephanie Poppe [spoppe@umail.iu.edu]
Shruthi Jeganathan [sjeganathan@indiana.edu]
Indiana University

Cid Freitag [cjfreitag@gmail.com]
University of Wisconsin

PROJECT DETAILS

Phylet visualizes a subset of Open Tree of Life (OToL[1]) data using a graph network visualization (spring forced directed acyclic graph using D3 Javascript library[2]). The Phylet web application employs an incremental graph API (HTTP JSON requests), a client-server data implementation (Neo4j, Python, py2neo, web cache), database and local searches (JavaScript, Python), a Smart Undo/Session interface (Javascript), and a browser-based UI (HTML5, Bootstrap.js).

Phylet was developed in six weeks—synchronously and asynchronously—by a global team. The project is an open-source project under the Apache 2.0 license hosted at the Phylet code repository[3] and continues in affiliation with the Smith Lab at the University of Michigan, the National Center for Evolutionary Synthesis (NESCent), and a growing network of at-large contributors.[4]

1 http://blog.opentreeoflife.org
2 http://d3js.org
3 Phylet code repository: http://bitbucket.org/phylet/phylet
4 http://www.onezoom.org and http://tolweb.org/tree

REQUIREMENTS ANALYSIS

The Open Tree of Life is a large-scale cyberinfrastructure project aimed at assembling, analyzing, visualizing, and extending the use of all available phylogenetic data about global extant (living) species. We found that variation in the collection and analysis of phylogenetic data results in evolutionary tree visualizations that obscure the extent of (1) social disagreement among scientists about species lineages and classifications, (2) sources and ubiquity of the evidence used to support relationships, and (3) the biological and genetic messiness that exists within diverse species assemblages.

After exploring zoomable trees for informal public educational purposes (OneZoom and Tree of Life Web Project),[5] our focus evolved into creating a visualization to assist researchers and specialists in the areas of phylogenetics, biology, evolutionary biology, and others in identifying unresolved regions of conflict among OToL data. This tool is to encourage research exploration, hypothesis generation, and data conflict resolution. The resulting visualization of these conflicts can illuminate the as-yet-unresolved areas of important biological and methodological questions.

During the initial phase of the project, we studied and implemented several visualization methods, including dynamically loaded trees, circular trees, and circle packing within D3, as well as complete networks within Gephi. Our final product uses a forced-based dynamically loaded graph using D3.

The representation of evolutionary histories using tree-like graphs began long before Charles Darwin wrote *The Origin of Species*, and since Darwin, visual explanations of phylogenetic relationships have diversified.[6] Despite these advances, the visual grammar and notational precision of evolutionary biology and phylogenetics remains fairly fuzzy.

This is important because psychologist Mihaly Csikszentmihalyi described how more precise notational systems make it easier to detect change and evaluate whether or not individuals or groups have made original, creative contributions to a particular domain.[7] In music, for example, there are a variety of different use contexts and each of these different use contexts is supported by the precision of music's notational system, which includes different semantic marks. When a domain like evolutionary biology and phylogenetics employs an increasingly precise notational system, it means that creative contributions can be detected, shared, and rewarded more easily, making the domain more flexible, responsive, and innovative. This can potentially open a field up to insights from other domains using at least two distinct leverage

5 Phylet is migrating to http://phylet.herokuapp.com or http://phylet.com. At the time of this writing the Phylet development site can be reached at http://mariano.gmajna.net/gol/index.html.
6 Pietsch, T. W. 2012. *Trees of Life: A Visual History of Evolution*. Baltimore: Johns Hopkins University Press.
7 Csikszentmihalyi, M. 1988. "Society, Culture, and Person: A Systems View of Creativity." In *The Nature of Creativity: Contemporary Psychological Perspectives*, edited by R.J. Sternberg, 429–440. Cambridge: Cambridge University Press.

points for signaling to other communities, public engagement, and broader impact: the level of *social agreement* within the field and the *threshold for meaningful contributions* to that field.

Our project then consisted of two distinct efforts: a *visual grammar* for the dataset and a *web application* for browsing and navigating the dataset. This fit well with our client's goal of using the visualization as a preliminary map of the massive OToL dataset.

RELATED WORK

Given the task of representing loosely hierarchical data, we investigated several visualizations of phylogenetic data, as well as hierarchical and network information in general.

We evaluated several web applications available to visualize small phylogenies (a few hundred species), which included basic bifurcating trees in either a traditional or circular dendrogram format, including PhyloBox, Archaeopteryx, OneZoom, DeepTree, and Dendroscope, all freely available online. Nevertheless, some are dynamic in nature, but all are focused in a strictly tree-like taxonomy structure, and in the best cases, such as Mesquite, it allowed for comparison of two different and conflicting taxonomy trees.

Other non-strictly hierarchical visualizations such as the network-based GNOME suffer from scalability problems, relying on the complete dataset to form a visualization, and can only handle small sets or predefined subsets.

In terms of workflows and task analysis, requirements around the Encyclopedia of Life provided some articulation for specific requirements to engaging their communities of practice.[8]

DATA COLLECTION AND PREPARATION

The original OToL data comprised approximately 1.9 million species and included 2,159,861 nodes, 9,037,190 properties, and 6,505,797 relationships. Phylet visualization data were drawn from a subset and included 120,461 nodes, 504,866 properties, and 367,334 relationships. OToL investigator Stephen Smith supplied the data in the Neo4j database format, which we used to build the visualization along with the visualization libraries and the Python-based API.

For the visualization, we were unable to make use of all the data features and chose to simplify the fields to include child-parent relationships, data source, and the presence or absence of any conflicts in child-parent relationships. We also created additional properties: the number of children and the number of parents. This allowed sizing of collapsed nodes according to number of children, providing an additional navigational cue. We were primarily interested in a visual grammar that would orient the viewer towards regions of conflict, while providing secondary visual cues around species diversity.

[8] http://wiki.eol.org/display/public

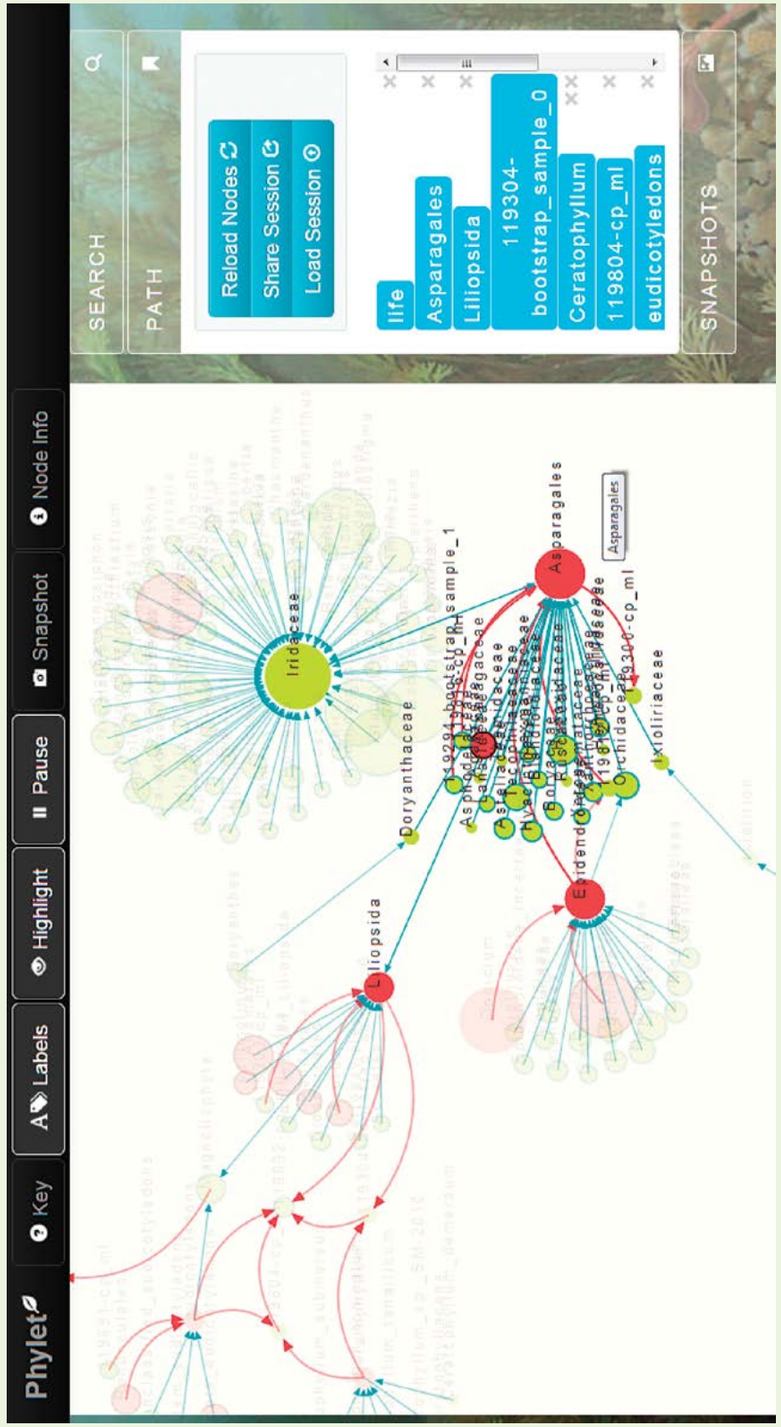


Figure 8.4 Phylet web application view with visualization network graph (left), session tools (right), and information toggles (top navigation) (http://cns.iu.edu/ivmoocbook14/8.4.jpg)

ANALYSIS AND VISUALIZATION

Based on the observation that evolutionary trees and similar representations often function as *boundary objects* that connect the technical and meaning-making practices of different communities—from the scientific to the non-scientific,[9] we considered tasks a user might want to do: share a relationship, their working session, or specific images or information.

We also aimed for a visual grammar that would scale well. The properties of our data visualization can be seen in Figure 8.4 and are summarized as follows:

- Children and parent nodes: colored green for resolved if there is only one parent node or red for unresolved if there are two or more parent nodes
- Nodes sized according to number of children nodes
- Presence of black, stroked node indicates more relationships to be explored
- Blue edges for resolved nodes
- Red curved edges for unresolved nodes using "cp_ml," "bootstrap," and other identification methods

We rapidly prototyped a variety of sketches, workflow tasks, and visualizations to include a handful of interaction tasks:

- Drag-pan-zoom the visualization
- Toggle on/off node labels
- Display the node info and metadata about parents, children, etc.
- Highlight node path on hover to highlight the node's parents and children
- Search for a node in displayed nodes/remote data including wildcard entries
- Path info track the clicks made to reach a node
- Remove actions from path
- Share, save, and load work sessions
- Snapshots of current graph saved as an SVG
- Gravity toggle to stop the movement of nodes

User research and prototyping revealed several important insights, including the need for clear narration, definitions, and storytelling around the goals of the visualization, supported tasks, workflows, and opportunities for creative appropriation of Phylet's functionality.

[9] Star, S.L., and J.R. Griesemer. 1989. "Institutional Ecology, Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39." *Social Studies of Science* 19, 3: 387–420.

Users did not enjoy switching tasks (e.g., switching between the visualization and the legend). This suggests that the interface should contain all relevant information, while resources for learning about use (e.g., node-click interactions) should be embedded into the visualization itself. Additionally, users wanted to selectively disclose waypoints during work sessions to assist dataset navigation.

Additional work remains to be done on the UI, interaction, and graphics, but these will develop as our skill with the software and data improves. Improvements to the backend data processing and visualization architecture will support better frontend usability, real-time dynamic views, improved usability, and shareability. In our next steps, we plan to add a tree-view visualization option that will fit many users' prior expectations for this context.

DISCUSSION

Although our technological fluency limited the rate at which we could work, our biggest challenges were social. Distributed collaboration is still a relatively new experience for any team. Computer and network-supported collaborative tools like Skype, Google+ Hangouts, instant messaging, email, large file transfer services, project management dashboards, and others lower the barriers to cooperative effort, but replicable strategies are needed for teams to formulate goals, uncover best practices, provide feedback, resolve conflicts, identify complementary skills, and coordinate roles. We know of no toolset or resource for collaboratively carrying out many of the specialized tasks associated with data and information visualizations. This would be a fantastic area for future research.

Our team benefitted from multiple and frequent face-to-face conferences, and in general, we made quick, provisional decisions. Our big leaps came with each prototype, demonstration, write-up, or contribution submitted by team members. In this sense, the project was built largely from the initiative and iterative contributions of each member. Using a version control system (e.g., Mercurial/Git) to host and track changes during software code and documentation development helped the team progress. However, our choices of technology significantly excluded some members of the team from participating in the technical implementation—although they remained influential in design, strategy, and research.

Ultimately, the collaboration and mutual reinforcement helped team members develop new skill sets and capabilities. The work is not complete, but the important lesson is that we have created a visualization that asks as many more questions than it answers. This generative characteristic is critical for the designer and the user and will be instrumental in guiding the Phylet project forward.

ACKNOWLEDGMENTS

The authors wish to acknowledge and thank Katy Börner, David E. Polley, Scott E. Weingart, Karen Cranston, and Chanda Phelan for their cooperation, guidance, and contributions.

Case Study #5

Isis: Mapping the Geospatial and Topical Distribution of a History of Science Journal

CLIENT:
Dr. Robert J. Malone [jay@hssonline.org]
History of Science Society

TEAM MEMBERS:
David E. Hubbard [hubbardd@library.tamu.edu]
Texas A&M University

Anouk Lang [anouk@cantab.net]
University of Strathclyde

Kathleen Reed [reed.kathleen@gmail.com]
Vancouver Island University

Anelise Hanson Shrout [anshrout@davidson.edu]
Davidson College

Lyndsay D. Troyer [ld.Troyer@gmail.com]
Colorado State University

PROJECT DETAILS

This project explores the history of science through the geospatial distribution of authors and trending topics in the journal *Isis* from 1913 to 2012. The project was requested by Dr. Robert J. Malone, Executive Director of the History of Science Society, and the analysis was completed by a project team comprised of a chemist, two academic librarians, and two digital humanists. Major insights include shifts in author contributions from Europe to the United States, as well as more geographically dispersed authorship within the United States over the last century. In addition to changes in authorship, contributed articles shifted from the study of individuals to collective endeavors with greater social context.

REQUIREMENTS ANALYSIS

Isis, an official publication of the History of Science Society, was launched by Belgian mathematician George Sarton in 1913. Dr. Malone sought a visual representation of *Isis* contributors and their locales over the past 100 years—one that would provide a dynamic picture of how scholarship in the history of science has shifted over the last century. The client also expressed interest in displaying the visualization as a poster, so the visualization needed to be static and possess sufficient resolution for a large display. The challenge for the project team was to represent temporal changes in the geospatial distribution of authors within a static visualization. While not specifically requested by the client, the project team also conducted a topical analysis of the journal article titles to explore major publication themes over the last 100 years.

RELATED WORK

Most studies visualizing the geospatial aspects of authorship utilize proportional symbols,[1,2,3] though choropleth maps can and have been used.[4] In the studies cited, the absolute number of publications is encoded onto a geographic map for a specified date range. The present study extends this approach by mapping the difference in publication activity between two historical periods. Another aspect of the current study employs Kleinberg's burst detection algorithm to explore topical trends.[5] The approach is similar to Mane and Börner's exploration of topic bursts and word co-occurrence in the *Proceedings of the National Academy of Sciences,*[6] though the current study is limited to article titles. To date, there are no known studies of *Isis*, or any other history of science publication, using these types of visualizations.

DATA COLLECTION AND PREPARATION

The spreadsheet obtained from the client contained 2,133 entries for articles published in *Isis* from 1913 to 2012. The main attributes used for the analysis were publication year, article title, and geographic location of the first author for each entry. The article titles were cleaned up, foreign language article titles were translated, and the geographic location and date formats normalized. The topical analysis (i.e., burst detection) was then

1 Batty, M. 2003. "The Geography of Scientific Citation." *Environment and Planning A* 35, 5: 761–765.

2 LaRowe, G., Ambre, S., Burgoon, J., Ke, W., Börner, K. 2009. "The Scholarly Database and Its Utility for Scientometrics Research." *Scientometrics* 79, 2: 219–234.

3 Lin, J.M. , Bohland, J.W., Andrews, P., Burns, G.A.P.C., Allen, C.B., Mitra, P.P. 2008. "An Analysis of the Abstracts Presented at the Annual Meetings of the Society for Neuroscience from 2001 to 2006." *PLoS ONE* 3, 4: 2052.

4 Mothe, J., Chrisment, C., Dkaki, T., Dousset, B., Karouach, S. 2006. "Combining Mining and Visualization Tools to Discover the Geographic Structure of a Domain." *Computers, Environment and Urban Systems* 30, 4: 460–484.

5 Kleinberg, J. 2002. "Bursty and Hierarchical Structure in Streams." *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 91–101. New York: ACM.

6 Mane, K.K., Börner, K. 2004. "Mapping Topics and Topic Bursts in PNAS." *Proceedings of the National Academy of Sciences of the United States of America* 101, Suppl. no.1: 5287–5290.

One Hundred Years of *Isis*

David E. Hubbard (Texas A&M University), Anouk Lang (University of Strathclyde), Kathleen Reed (Vancouver Island University), Anelise Hanson Shrout (Davidson College) and Lyndsay D. Troyer (Colorado State University)

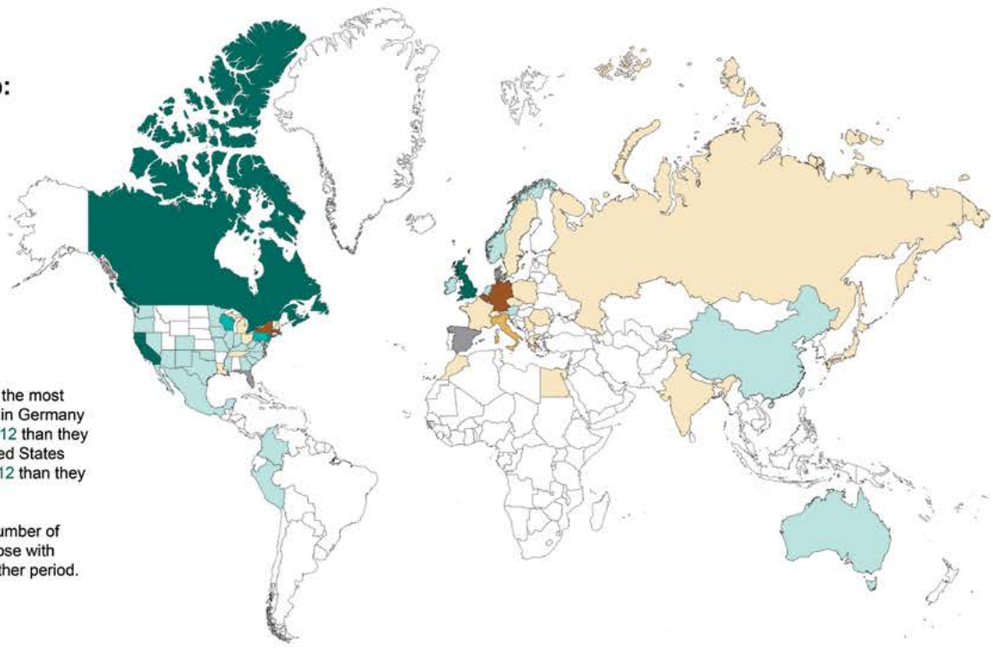
Changes in *Isis* Authorship: 1913-1937 to 1988-2012.

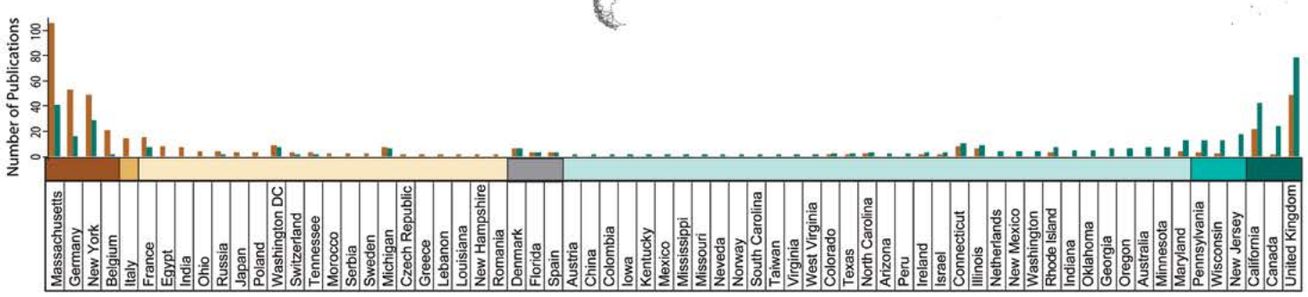
This map and chart show geographic shifts in *Isis* authors' locations across two time periods of 25 years each: 1913-1937 and 1988-2012.

Countries and states tinted brown had a greater number of articles published in the 1913-1937 period, while those tinted green had a greater number published in the 1988-2012 period.

Germany and the United States experienced the most extreme shifts in authorship. Authors located in Germany published 37 fewer articles between 1988-2012 than they did in 1913-1937. Authors located in the United States published 31 more articles between 1988-2012 than they did in 1913-1937.

Countries and states in gray had the same number of of published articles in both periods, while those with no color (i.e., white) had no publications in either period.





Thematic trends over time

In *Isis*'s early years, the journal showed considerable interest in non-Western cultures. The appearance of *Arab* and *Egyptian* in the 1920s reflect Sarton's interests in that decade. In the 1970s, *Isis*'s interest shifted focus to the West. *German* emerged as scholars began to pay attention to 19th-century German universities, which supported Wissenschaft (the study of science that involves systematic research and teaching), and from which grew many of the American research universities that became the primary locus for American scientific research for many years.

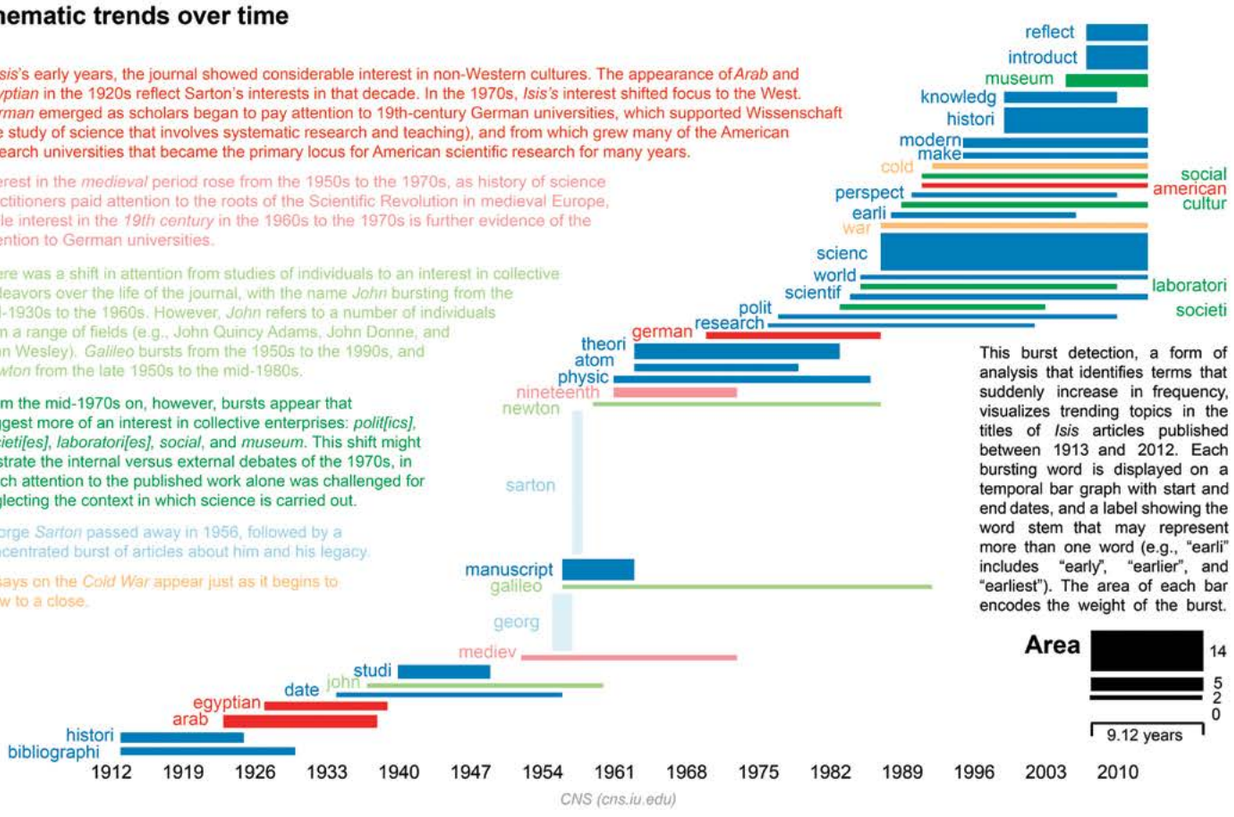
Interest in the *medieval* period rose from the 1950s to the 1970s, as history of science practitioners paid attention to the roots of the Scientific Revolution in medieval Europe, while interest in the *19th century* in the 1960s to the 1970s is further evidence of the attention to German universities.

There was a shift in attention from studies of individuals to an interest in collective endeavors over the life of the journal, with the name *John* bursting from the mid-1930s to the 1960s. However, *John* refers to a number of individuals from a range of fields (e.g., John Quincy Adams, John Donne, and John Wesley). *Galileo* bursts from the 1950s to the 1990s, and *Newton* from the late 1950s to the mid-1980s.

From the mid-1970s on, however, bursts appear that suggest more of an interest in collective enterprises: *polit[fics], societ[es], laboratori[es], social*, and *museum*. This shift might illustrate the internal versus external debates of the 1970s, in which attention to the published work alone was challenged for neglecting the context in which science is carried out.

George *Sarton* passed away in 1956, followed by a concentrated burst of articles about him and his legacy.

Essays on the *Cold War* appear just as it begins to draw to a close.



This burst detection, a form of analysis that identifies terms that suddenly increase in frequency, visualizes trending topics in the titles of *Isis* articles published between 1913 and 2012. Each bursting word is displayed on a temporal bar graph with start and end dates, and a label showing the word stem that may represent more than one word (e.g., "earli" includes "early", "earlier", and "earliest"). The area of each bar encodes the weight of the burst.

Figure 8.5 Final visualization: geospatial and topical analysis of *Isis* (http://cns.iu.edu/ivmoocbook14/8.5.jpg)

performed on all 2,133 article titles using burst detection in Sci2.[7] Due to absence of many geographic locations for the authors in the client data, the project team decided to compare the first 25 years (1913–1937) to the most recent 25 years (1988–2012) for the geospatial analysis. Once limited to the two 25-year periods, the spreadsheet contained 438 entries for 1913 to 1937 and 430 entries for 1988 to 2012.

ANALYSIS AND VISUALIZATION

A number of approaches were explored to visualize the data prior to the final visualization. The initial topical analyses brought to light a number of functional words (e.g., *volume*, *index*, and *preface*) in the article titles whose inclusion was distorting the dataset and potentially crowding out other more interesting terms. The visualization was improved by removing additional common terms, identified with the help of AntConc,[8] and categorizing words that experienced a sudden increase in their usage frequency as identified in the burst detection. The topical bursts, displayed as weighted horizontal bars, were color-coded to match accompanying explanatory text in the final visualization.

The initial geospatial visualizations utilized proportional symbols to encode the number of authored articles onto geographic maps for the two historical periods, but there was considerable crowding of the proportional symbols in some geographic locations. Using a choropleth map rather than proportional symbols offered an alternative. By calculating the difference between the number of articles published for each country in the 1913–1937 range and the number published in the 1988–2012 range, the difference in the number of articles authored between the historical periods could be encoded onto a single choropleth map. Since a large number of authors from the United States dominated both historical periods, each U.S. state was encoded in the same manner as a country. The geospatial visualization used a divergent brown-green color scheme to indicate decreases and increases in publication activity. The final visualization combined the geospatial and topical visualizations, as well as a bar graph to display the absolute number of publications for each country (Figure 8.5).

DISCUSSION

Sarton moved from Belgium to Cambridge, Massachusetts, after the outbreak of World War I and re-launched *Isis*.[9] This may explain the concentration of contributors from the Northeast

7 Sci2 Team. 2009. "Science of Science (Sci2) Tool." Indiana University and SciTech Strategies. http://sci2.cns.iu.edu.

8 Anthony, L. "AntConc." http://www.antlab.sci.waseda.ac.jp/software.html (accessed February 27, 2013).

9 McClellan, J.E., III. 1999. "Sarton, George Alfred Léon." In *American National Biography*, vol. 19, edited by J.A. Garraty and M.C. Carnes, 295–297. New York: Oxford University Press.

during the early period (1913–1937), which after Sarton's death in the 1950s became more dispersed throughout the United States in 1988 to 2012. In terms of changes in author locales, Germany and the United States experienced the most extreme shifts in authorship.

In relation to the burst detection analysis, there appears to be an editorial shift in the 1950s because the burst patterns that precede this decade, and those that follow it, are markedly different. This shift coincided with the death of Sarton in 1956. After the 1950s, attention shifted to different geographic regions and time periods and from studies of individuals to those of larger groups. For example, interest in the medieval period rises from the 1950s to the 1970s as history of science practitioners paid attention to the roots of the Scientific Revolution in medieval Europe. There was also a shift in attention from studies of individuals to an interest in collective endeavors. The name John, for example, bursts from the mid-1930s to the 1960s, representing figures such as John Quincy Adams, John Donne, and John Wesley. Galileo bursts from the 1950s to the 1990s, and Newton from the late 1950s to the mid-1980s. From the mid-1970s on, however, the bursty words that appear suggest more of an interest in collective enterprises: *polit*[*ics*], *societi*[*es*], *laboratori*[*es*], *social*, and *museum*. This shift might illustrate the internal versus external debates of the 1970s within the history of science field, in which attention to the published work alone was challenged for neglecting the context in which science is carried out. The project team put their hypotheses about the significance of these patterns to the client during the validation process, and he and a colleague in the discipline provided contextual information that fleshed out these hypotheses.

Using all 100 years might provide a fuller picture and the opportunity to study changes decade-by-decade, as would using other standard bibliometric approaches (e.g., co-citation analysis). The client asked about Charles Darwin, but Darwin did not appear as one of the "bursty words." Further refining of the burst detection could also be performed on the named persons from the titles of articles (e.g., Charles Darwin). The dataset only contained 2,133 articles and three main attributes (year, article title, and location), but contained sufficient complexity to justify and benefit from geospatial and topical visualizations. The approaches utilized in this study could be used for other publications and scaled to larger projects.

ACKNOWLEDGMENTS

We would like to thank Dr. Robert J. Malone for the opportunity to work on this project, assistance securing the data, and feedback. We also extend our appreciation to the anonymous History of Science Society member who answered our questions and provided insights into the history of science.

Case Study #6

Visualizing the Impact of the Hive NYC Learning Network

CLIENT:
Rafi Santo
Indiana University

TEAM MEMBERS:
Simon Duff [simon.duff@gmail.com]
John Patterson [jono.patterson@googlemail.com]
Camaal Moten [camaal@gmail.com]
Sarah Webber [sarahwebber@gmail.com]

PROJECT DETAILS

Hive NYC Learning Network, an out-of-school education network made up of 56 organizations, aims to develop a range of techniques, programs, and initiatives to create connected learning opportunities for youth, building on twenty-first-century learning approaches.[1] The network is being studied and supported by a group of researchers from Indiana University and New York University called Hive Research Lab.

Learning networks like Hive NYC have great potential to act as an infrastructure for improving the skills and abilities of local populations, the young in particular. They can also open up opportunities for greater collaboration and innovation amongst educators and learning organizations. Understanding how individual learning networks operate and grow may provide valuable insight into opportunities for development and improvement of other learning networks.

REQUIREMENTS ANALYSIS

The client in this instance was Rafi Santo, project lead of Hive Research Lab and a doctoral candidate at Indiana University. The dataset obtained by the client provided information on organizations who received funding, the project name, year and date of award, type and amount of grant funding, and the number of youth the project reached.

The brief given by our client was very broad and summarized as "[provide] substantive insight into various patterns within the network, most importantly the patterns of collaborations between organizations over time and the numbers of youth reached for the amount of resources used in a project."

1 Hive NYC Mission Statement: http://explorecreateshare.org/about/ (accessed August 2013).

Our team sought to answer three questions with the visualization. First, who had received most funding and who had worked with whom and when? Second, what was the return on investment in terms of project funding versus amount of youth reached? Third, what does the urban geography of Hive NYC look like and how might this have influenced the project? To answer these questions, three different visualizations were designed:

1. A network diagram showing patterns of collaboration over time.
2. A graph depicting the funding awards and number of youth reached.
3. A map supporting understanding of the geography of the Hive.

RELATED WORK

For visualizing information sharing and collaboration over time, a display static visualization such as the history flow interface, used to show the evolution of Wikipedia entries over time,[2] can be used. However, Reda et al. suggest that for temporal visualization of the emergence, evolution, and fading of communities or dynamic social networks, graph-based representations are not ideal and they argue for employing interactive visualizations. Animation of network graphs were implemented by Leydesdorff et al. to communicate the evolution of scholarly communities.[3,4] Harrer et al. argued for a three-dimensional visualization that places the network structure within the first two dimensions, and representing change over time on the third axis—moving through "time slices" enables viewers to explore the dynamics of the network.[5] Falkowsi and Bartelheimer propose two approaches to displaying social network dynamics: one is aimed at visualizing stable subgroups, the other more transient communities where members join in and out. Both use temporal displays to show the communities at each point in time.[6] Finally, Börner presents a suitable workflow for the production of network "with whom" analysis which was used in the project.[7]

2 Viégas, F. B., M. Wattenberg, and K. Dave. 2004. "Studying Cooperation and Conflict between Authors with History Flow Visualizations." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 575–582. New York: ACM.

3 Reda, K., C. Tantipathananandh, A. Johnson, J. Leigh, and T. Berger Wolf. 2011. "Visualizing the Evolution of Community Structures in Dynamic Social Networks" *Computer Graphics Forum* 30, 3: 1061–1070.

4 Leydesdorff, L., Schank, T. 2008. Dynamic Animations of Journal Maps: Indicators of Structural Change and Interdisciplinary Developments, *Journal of the American Society for Information Science and Technology* 59.

5 Harrer, A., S. Zeini, S. Ziebarth, and D. Münter. 2007. "Visualisation of the Dynamics of Computer-Mediated Community Networks."

6 Falkowski, T., and J. Bartelheimer. 2006. "Mining and Visualizing the Evolution of Subgroups in Social Networks." *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 52–58. New York: ACM.

7 See Figure 6.10 in this book.

Visualising Hive NYC

The following visualizations map the connections between Hive NYC members and community projects using various perspectives. The dataset was obtained from the Hive Fund Projects Database, which consisted of 54 projects and 47 members from 2011-2013.

TEAM MEMBERS
SIMON DUFF | JOHN PATTERSON | CAMAAL MOTEN | SARAH WEBBER

GEOSPATIAL NETWORK

ABOUT THIS VISUALIZATION

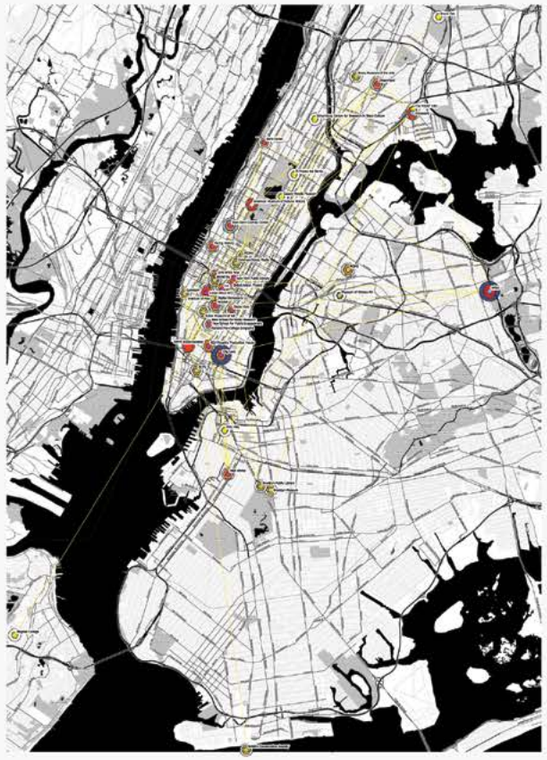
INNER CIRCLE - HIGHEST GRANT ACHIEVED
- Catalyst
- Lever
- Spark
- Unknown

HEXAGON - FUNDING AMOUNT SECURED
- 0 - 7500
- 7500 - 50000
- 50000 - 140000
- 140000 - 275000
- 275000 - 375000

OUTER CIRCLE - TOTAL YOUTH REACHED
- 0 - 124
- 124 - 248
- 248 - 372
- 372 - 496
- 496 - 620

EDGES - RELATIONSHIPS

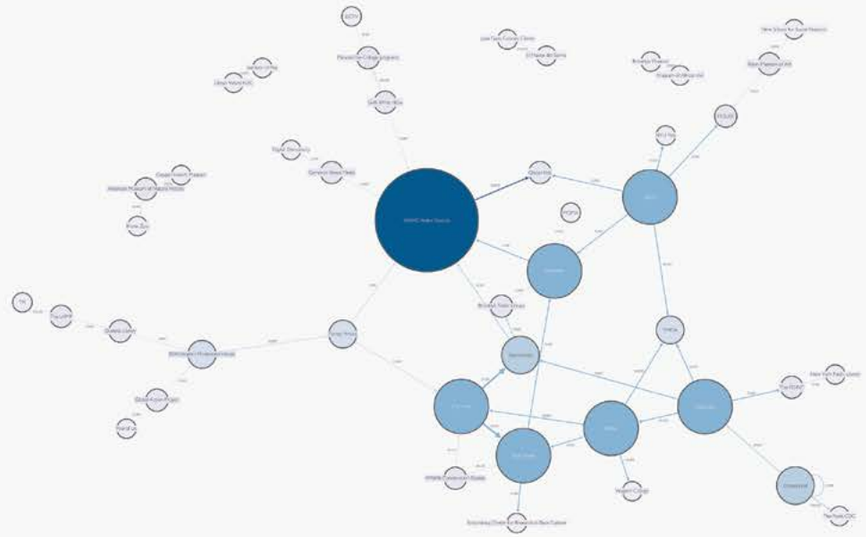
STAMEN TONER/OSM



COLLABORATION NETWORK

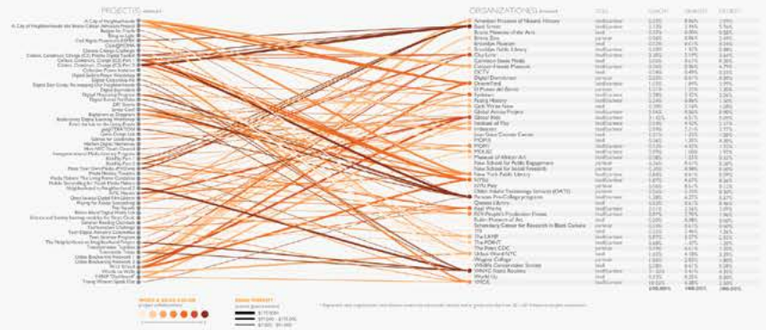
ABOUT THIS VISUALIZATION

NODE SIZE & COLOR

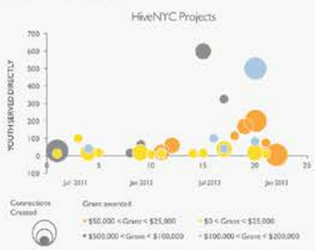


BIPARTITE NETWORK

ABOUT THIS VISUALIZATION



ROI BUBBLE PLOT



TEMPORAL NETWORK

ABOUT THIS VISUALIZATION

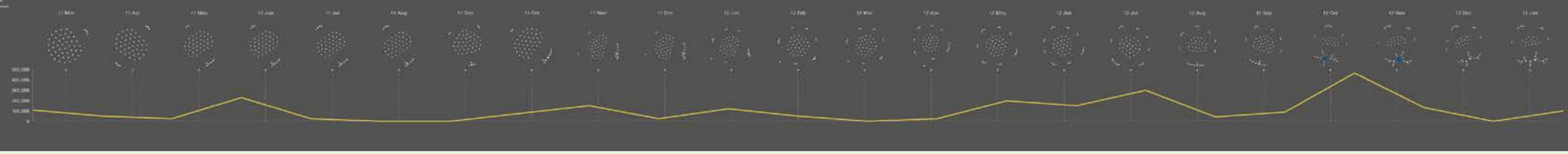


Figure 8.6 Hive NYC geospatial map (top), collaboration network (top right), bipartite network (middle), ROI Bubble Plot (middle right), and temporal network evolution (bottom) (http://cns.iu.edu/ivmoocbook14/8.6.jpg)

DATA COLLECTION AND PREPARATION

The dataset provided by Hive Research Lab consisted of 54 projects that 47 member organizations engaged in between 2011 and 2013. Data preprocessing involved correcting typos in organization names, looking for duplicate records and filling in missing values, and normalizing data to the same formats (e.g., dates in U.S. calendar notation). We enriched the data by appending any new data we needed to create our proposed visualizations—for example, geocoding organization addresses to append the latitude and longitude coordinates to the dataset. We also gathered a range of contextual background information on Hive NYC. Next, range summary statistics were calculated (see Table 8.3).

Table 8.3 Basic Statistics

Total # of Projects	54
Total # of Organizations	47
Average $ per Award	$60,200
Total Awarded (all projects)	$3,300,000
Average Youth Reached per Project	76
Total Youth Reached	3,449
Average Partners per Project	2.3
Average Cost per Youth Engaged	$1,697

ANALYSIS AND VISUALIZATION

The outputs generated from the analysis were five distinct visualizations that were combined with narrative to form the final visualization:

1. A geospatial network visualization to show Hive NYC member organizations, project links, and encode funding, youth reached, and funding type as points on the map. This provided a sense of the geographical distribution and collaborative links of partner organizations. Generally, organizations are tightly geographically clustered, but there are clearly strong links being forged between more distant partners, such as Bronx Zoo and NYSci.
2. A network showing 47 Hive NYC partner organizations as nodes, and the collaborative relationships between them as edges, directed from project leads towards secondary project participants. Nodes are labeled with organization names, while the link between any two nodes is labeled with the cumulative grant amount of the projects.
3. A bipartite network analysis was conducted to illustrate the participation of organizations (listed on the right) according to projects (listed on the left). Each record was represented using a labeled circle encoding project connections and edges weighted by grant amount. The goal was to illustrate each organization's contribution to the overall impact of the Hive NYC learning network. At a glance, one can identify organizations that helped secure the most participation, grant awards, or shared the workload.
4. The team also created a ROI Bubble Plot, using a traditional bubble plot that highlights projects funded through time and how each project returned on investment in terms of youth reached (*y*-axis) and connections created (bubble size). It demonstrates that the number of youth reached does not appear to be correlated with amount invested, but that projects after July 2012 are proving more successful than earlier projects. This might indicate a maturing of Hive NYC.

5. A temporal network visualization, presented as a small multiples graph, shows collaboration during a specific timeframe. It illustrates how the connections between Hive NYC members change from March 2011 to January 2013. The total grants awarded are plotted below each project start date. The visualization clearly highlights how relationships form around grants in fall and spring periods. A small number of organizations collaborate at the same time.

Finally, a large-format visualization was created by combining all of these visualizations (see Figure 8.6).

DISCUSSION

A number of insights can be gained from the visualizations:

- The Hive NYC learning network has grown rapidly, and funding has been key to this growth. The use of a three- to four-stage funding cycle by the learning network helps retain partners over the long-term.
- There was no correlation between funding amount and youth reached, where one might otherwise expect that more funding results in more youth reached. This is unsurprising given that the network sees itself as innovation oriented and thus takes higher risks.
- The bulk of Hive NYC organizations are clustered in the central boroughs of NYC with some organizations further out where there is a strategic link (e.g., zoos, universities, and other more unique institution types).

Initially, Hive NYC was viewed as a constantly evolving and growing network with more and more "live" relationships developing and being sustained. The temporal network suggests an alternative view: formal relationships are created around projects funding and then disband into informal relationships afterward; successful projects might then lead to future collaborations fueling a positive feedback cycle that is propelled by projects that make a difference.

Throughout the project, a number of challenges, including poor comparability between grants and projects data and limited documentation for the dataset. The six week time frame also posed a significant challenge.

ACKNOWLEDGMENTS

We would like to thank our client Rafi Santo, the Hive Research Lab, and Indiana University professors and staff for feedback, information, and encouragement; the Hive NYC Learning Network member organizations and the Mozilla Foundation for performing and supporting all the projects that we visualized; and Stamen Design for making awesome map tiles.