Cartograph: Unlocking Thematic Cartography Through Semantic Enhancement

Cartograph Research Team: * **Shilad Sen, Anja Beth Swoap, Qisheng Li**
Brooke Boatman, Ilse Dippenaar, Rebecca Gold, Monica Ngo, Sarah Pujol
Macalester College, St. Paul, United States

Bret Jackson
Macalester College
St. Paul, United States
bjackson@macalester.edu

Brent Hecht
Northwestern University
Evanston, United States
bhecht@northwestern.edu

ABSTRACT

This paper introduces Cartograph, a visualization system that harnesses the vast amount of world knowledge encoded within Wikipedia to create thematic maps of almost any data. Cartograph extends previous systems that visualize non-spatial data using geographic approaches. While these systems required data with an existing semantic structure, Cartograph unlocks spatial visualization for a much larger variety of datasets by enhancing input datasets with semantic information extracted from Wikipedia. Cartograph's map embeddings use neural networks trained on Wikipedia article content and user navigation behavior. Using these embeddings, the system can reveal connections between points that are unrelated in the original data sets, but are related in meaning and therefore embedded close together on the map. We describe the design of the system and key challenges we encountered, and we present findings from an exploratory user study.

Author Keywords

thematic cartography; maps; Wikipedia; neural networks; Wikidata; semantic relatedness

INTRODUCTION

For hundreds of years, humans have leveraged **thematic cartography** as a powerful means to quickly and effectively communicate complex geographic distributions [38]. Thematic cartography helps us understand and explore multifaceted geospatial processes ranging from election results [2, 5] to climate change [4] to sports broadcast availability [1]. As the quantity and diversity of spatial data increases, thematic maps - often of the interactive variety - now frequently appear in news articles, blog posts, educational applications, and in many other contexts.

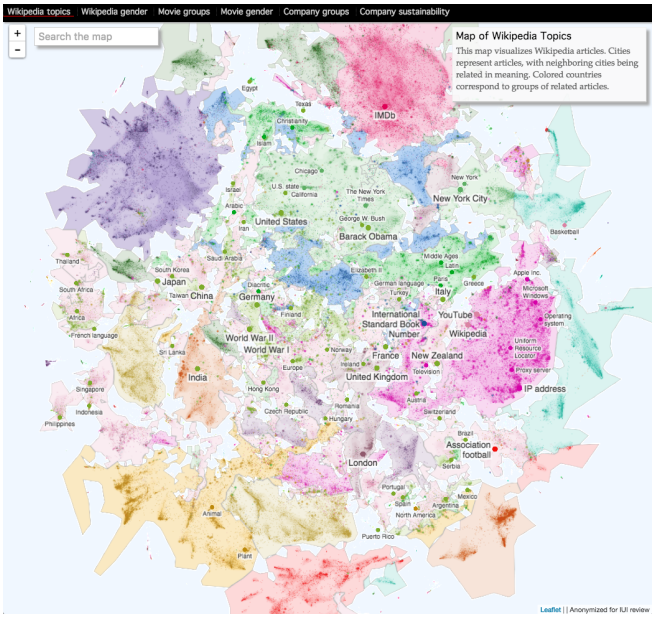


Figure 1: An overview of the Cartograph system webpage. A user can select a map using the title bar at the top and search for a concept using the box in the upper left. In this map, colors represent semantic topics.

One reason thematic cartography has proven so broadly useful is that it offers several widely-established communicative benefits [28, 38]. Most notably, thematic cartography has been shown to be highly effective at simultaneously (1) communicating **specific** values for individual spatial entities (e.g. the vote share in a specific U.S. state), (2) communicating **regional patterns** (e.g. the vote share in the "Great Plains" of the U.S.), and (3) helping people build and reference their **mental maps** (e.g. "I knew the Great Plains had higher church attendance than other areas, so I guess it makes sense that it voted more Republican"). These benefits are often best understood in contrast to other visualization approaches. For example, imagine the challenge of assessing regional patterns

*ssen@macalester.edu anja.beth@gmail.com qli@macalester.edu bboatman1241@gmail.com jdippena@macalester.edu rebeccagold0@gmail.com mngo@macalester.edu spujol@macalester.edu

IUI 2017, March 13 - 16, 2017, Limassol, Cyprus

ISBN 978-1-4503-4348-0/17/03…$15.00

DOI: http://dx.doi.org/10.1145/3025171.3025233

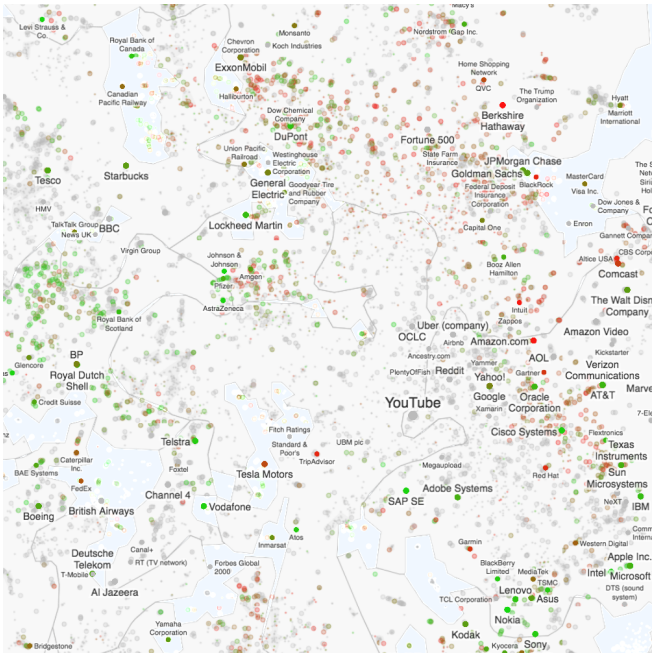


Figure 2: A map of businesses visualizing sustainability corporate sustainability ratings from low (red) to high (green).

or updating one's mental map using only a ranked list of U.S. county election results in contrast to using a thematic map.

Despite its many benefits, however, thematic cartography traditionally has one major limitation: it can only be used to support exploration and understanding in datasets that have explicit **geographic references**. In this paper, we seek to address this limitation by introducing Cartograph,[1] a system that uses Wikipedia-based neural network embeddings to extend the major benefits of thematic cartography to datasets that are not geographic in nature. Specifically, Cartograph uses a novel "base map" defined by low-dimension embeddings of Wikipedia content and Wikipedia navigation behavior to visualize a wide variety of user-defined datasets. This generalizability emerges from applying recent embedding techniques to the vast amounts of available Wikipedia data (and Wikidata [42]), which affords a universal frame of reference on which datasets from many domains can be layered.

Cartograph's approach to thematic cartography is illustrated in Figures 1 and 2. Figure 1 shows the "base map" without any thematic layer. Here, one can see that, through the use of neural network embeddings, related entities have been placed close together and less related entities are further apart. For example, technology-related concepts such as "email", "YouTube" and "web browser" appear nearby each other in the pink region to the "East", while concepts about U.S. culture and politics ("Barack Obama", "Chicago", "NY Times") appear in the middle in green. As we will describe below, the placement of related entities close to one another is an essential precondition to the use of cartography that enables regional exploration and understanding. Cartograph incorporates algorithms that produce maps that effectively maintain these relationships.

Figure 2 shows how Cartograph can visualize a non-spatial dataset, in this case business sustainability ratings from CSRHub.[2] Here, we utilize well-known cartographic techniques like graduated symbol mapping and standards-based variation in hue to indicate the domains in which companies are sustainable and those in which they are not. Zooming into the map shows several surprising and semantically-grounded regional patterns. While large European energy companies, banks and conglomerates in the southwest region such as Credit Suisse and Royal Dutch Shell show high sustainability ratings, similar U.S. corporations in the northeast region (Berkshire Hathaway, ExxonMobil) generally do not.

However, Cartograph extends more than just the regional communicative benefits of thematic cartography. Cartograph is interactive and supports (semantic) zoom, allowing people to see patterns at various semantic/spatial scales. This interactivity also supports details-on-demand through pop-ups that show additional information about each entity, reinforcing thematic cartography's ability to communicate information about specific entities. Similarly, because Cartograph uses a persistent base map, users can correlate what they learn about company sustainability with all the other datasets they visualize on this reference system. In other words, through its persistent, universal base map, Cartograph reinforces the indexing and updating of a mental map.

Cartograph can be understood as a **spatialization** system, a family of technologies that seek to represent large corpora (usually text documents) in 2D or 3D spaces. While spatialization systems often adapt techniques from cartography, they have limitations that have prevented them from taking advantage of several of the key benefits of cartography listed above. Specifically, existing spatialization systems either (1) cannot utilize a consistent base map, eliminating the mental map benefits of thematic cartography [11, 25, 37] or (2) are limited to a small family of non-geographic visualizations [6, 19]. Through its Wikipedia embeddings-based approach, Cartograph creates a persistent environment in which a large variety of datasets can be visualized, addressing both of these well-known limitations. Additionally, existing spatializations systems face well-known scaling challenges [19]. Cartograph addresses these challenges through its use of large-scale neural network embedding algorithms and recent advances in web mapping technologies. This enables Cartograph to offer users fluid web interaction for datasets containing millions of points, an order of magnitude larger than existing systems.

Cartograph requires two data characteristics that are common in exploratory analyses. First, dataset records must be **associable** with Wikipedia. However, as we note later, NLP techniques can be used to associate "tail concepts" that are not notable enough for inclusion in Wikipedia with related Wikipedia entities. For example, IUI researchers do not typically have Wikipedia articles about them. However, we could use algorithms to identify Wikipedia concepts mentioned in

[1] http://cartograph.info

[2] https://www.csrhub.com/

a researcher's publications or homepage. Second, the data layers must exhibit **semantic alignment** with Wikipedia. If the patterns to be visualized (e.g. sustainability ratings) do not correlate with Wikipedia's semantic structure (e.g. the link and text patterns among companies), Cartograph's approach will be less effective. Exploratory tasks (the focus of Cartograph) are likely to obey the alignment property because they seek to augment "human understanding" to produce data insights, an approach which explicitly leverages the relationship between data and semantics [43].

Below, we describe work that motivated Cartograph, highlighting the well-known limitations of existing systems that Cartograph directly addresses. Next, we overview the numerous design choices that went into Cartograph and their motivation. We then present several case studies to demonstrate a series of use cases for Cartograph. We close the paper by presenting an exploratory user study that provides insights into the strengths, weaknesses, and usage patterns of the system.

Lastly, while we include screenshots of the system throughout this paper, we encourage the reader to explore Cartograph online to experience these interaction techniques firsthand.[3]

RELATED WORK

Our work builds upon prior research that also visualizes data using cartographic metaphors created by embedding higher-dimensional data into a two or three dimensional map.

The goal of data visualization is to create a mapping from data to visuals that is insightful, communicates necessary information, and is aesthetically pleasing. This mapping process is sometimes called a "digital visual metaphor" for its similarities to linguistic metaphors, which map from one domain of information onto another [12]. Spatialization is a specific form of digital visual metaphor that maps non-spatial data onto cartographic maps [11, 25].

Cartographic maps make use of Tobler's *First Law of Geography*, which states that "Everything is related to everything else, but near things are more related than distant things" [39]. This distance-similarity relationship is one of the founding principles of geographic analysis [36], and it has been shown to hold for spatializations representing non-spatial data as dots placed in a 2D or 3D space [14, 31]. In thematic cartography, the distance-similarity metaphor is critical to supporting one of the three key benefits of thematic cartography listed above: regional analysis. If similar places were not related in some way — e.g. if "western Europe" or "the (American) South" did not share characteristics that bind it as a region — regional analysis would be futile. Ensuring distance-similarity is thus critical to any application of thematic cartography in a non-geographic domain.

The most prominent work in this area, like ours, recognizes the valuable role this distance-similarity relationship plays in sense-making and data analysis. This can be traced back to early efforts to display search results of document collections [8, 9, 23] or the world wide web [34] by extracting semantic similarity information and using dimensional reduction techniques such as multi-dimensional scaling (MDS), principal component analysis (PCA), or self-organizing maps (SOM) to place them in a 2D space. Our visualization system builds on these systems and others that integrate additional spatial metaphors such as: (1) network links between data points representing roads, (2) regions representing countries, and (3) other geographic boundaries such as contours or lakes (e.g. Gronemann et al. [18]).

Of particular note is the GMap system [17, 21], which presents an algorithm to produce cartographic maps from graphs using clusters as country regions. Although GMap is based on graph data, while our starting point is a set of vectors that are embedded as 2D points, the concepts are very similar. Like GMap, Cartograph embeds and clusters the dataset and then draws country borders based on those clusters. However, our border-generation algorithm has been refined to create more realistic internal and external boundaries, which enhances the map metaphor and makes it easier for novice users to understand and navigate.

A few spatialization systems provide inspiration for integrating additional data beyond similarity into the visualization. In this style, Gansner et al. [16] show how recommendations can be displayed using a heatmap overlayed on a cartographic visualization of movies and TV shows. Additional features such as the amount of time spent watching a movie are charted using label color and font size. Cartograph utilizes a similar approach, augmenting the map with a thematic layer that visualizes how the input data varies across geographic area.

Cartograph is also interactive. In addition to traditional query based searching (e.g. Fluit et al. [15]), it enables users to pan and zoom in to a focused detailed view or out to see the larger context. This multi-scale zooming approach has two key advantages: First, it allows Cartograph to run interactively in a web browser with millions of data points (most spatializations to date are limited to a few thousand points [24]). Second, it promotes exploration, allowing for serendipitous discovery and new insight generation. Cartograph "hints" at the points visible on the next zoom levels, a technique that has been successful in graph-based map visualizations [32].

As noted above, perhaps the most important distinction between Cartograph and previous work is that Cartograph works with almost any data. While research suggests that traditional spatialization visualizations promote discovery of similarities, clusters, and outliers (important criteria for any exploratory visualization) [10, 37, 40], traditional approaches require semantic relatedness features to be present *within* the dataset to be visualized. This limits the types of data that can be visualized using these systems. Cartograph instead applies semantic relatedness (SR) estimates extracted from Wikipedia for any lexically expressed concepts in the data, enabling use of spatialization techniques.

Along the same lines, Cartograph's use of SR does bear similarity to Hecht et al.'s Atlasify system [19] and related systems like Frankenplace [6]. Atlasify uses SR data for "explicit spatialization" to map data onto various spatial reference systems,

[3] http://cartograph.info

including a periodic table, a US map, and a map of Congress. These maps serve a different purpose from Cartograph. Rather than using the data to generate entirely new spatial reference systems of an information space, Atlasify overlays data on pre-existing spatial reference systems. While this has the benefit of leveraging existing mental maps of these reference systems, it significantly limits the types of visualizations systems like Atlasify can support. Indeed, Hecht et al. write that the Atlasify approach could be extended to arbitrary domains through an approach like Cartograph.

DESIGN OF CARTOGRAPH

This section describes the Cartograph system and the way in which it creates its map. The section that follows relies on two definitions: we use the term *domain concept* to refer to the external data points that are mapped, and *Wikipedia article* to refer to structured article content within Wikipedia.

Overview of System

Cartograph combines a four stage offline batch data pipeline with an online map server. We summarize the stages below and describe each stage in detail in the sections that follow.

1. **Concept definition:** Domain concepts broadly define the inputs to the Cartograph system. At a minimum, Cartograph requires the names of the domain concepts that should be mapped (e.g. "IBM", "Abraham Lincoln"). Cartograph associates each domain concept to Wikipedia and mines other key attributes such as popularity estimates and semantic vectors from Wikipedia itself.

2. **X,Y embedding:** Concept embedding produces (x,y) coordinates for each named concept. We note the distinction between the **high-dimensional vector space** used by Cartograph for semantic interpretation and the **two-dimensional x, y coordinate space** used for visualization. The high-dimensional space (typically 100 to 600 dimensional dense vectors [22]) supports semantic needs, such as neighbor extraction and clustering. The two-dimensional x-y space provides latitude and longitude for the spatial visualization.

3. **Country formation**. Next, "countries" are formed by clustering points in the high-dimensional space. Areas in the coordinate space associated with the same cluster form a portion of that cluster's country. Borders are then generated around each country and topological contours are created. These visual elements serve as landmarks that enable users to quickly identify meaningful semantic structures in the map.

4. **Domain-specific data layers**. During this step, Cartograph produces thematic cartography for any domain-specific data layer using GIS techniques. As with the definition of the concept space, domain specific metrics need only include concept names and quantitative metrics (for example, corporation names and sustainability indicators). GIS approaches such as choropleth maps, dot density visualizations, and heat maps can be used to visualize this data.

5. **Map Server**. Cartograph visualizes the concept data as a zoomable web-based map. It combines vector-based and raster-based approaches along with hardware-accelerated browser technologies to deliver a fluid online map of the data. By leveraging NLP algorithms trained on Wikipedia, Cartograph also supports natural language search, even for content not specified in the source concept space.

Stage 1: Concept Definition

The domain concept definition stage produces the raw inputs for the Cartograph system. Throughout the concept definition stage, Cartograph uses the WikiBrain system [35] to extract information from Wikipedia including textual content, article pagerank, page views, and content-based vector embeddings.

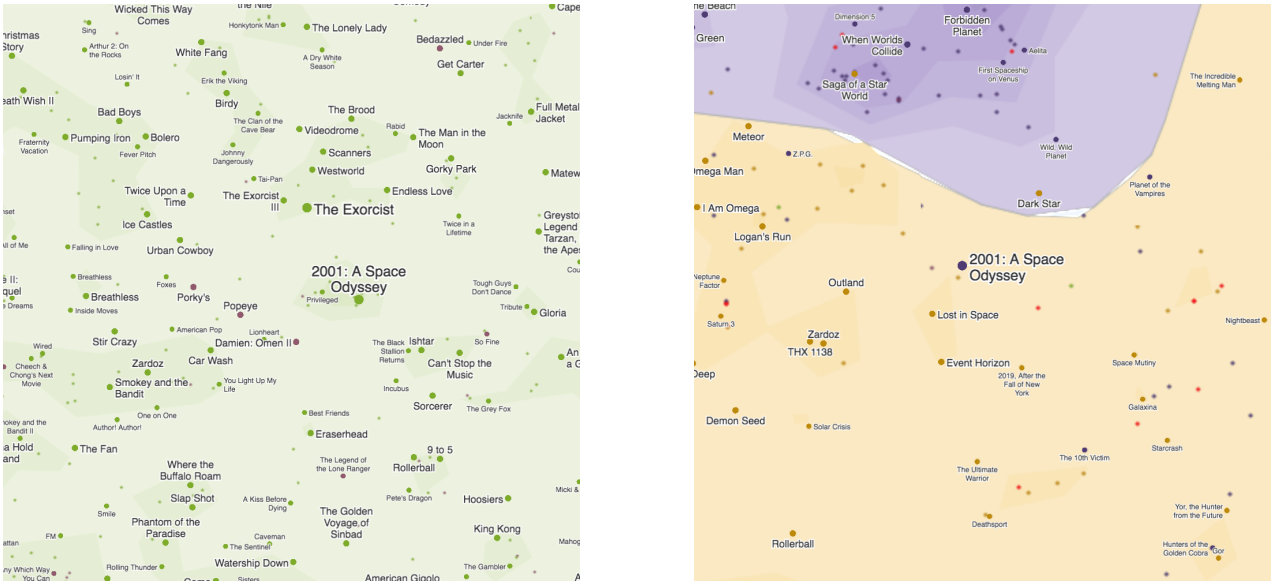
Concept identification: As an external input, Cartograph must know the *domain concepts* it should map and the *relationship* between those domain concepts and Wikipedia articles. The domain of concepts can be represented using Wikipedia article identifiers (titles or page ids), free text names and phrases (e.g. "PC", "Mac", "Linux", "notebook", "tablet"), or a query that can be run against Wikipedia (the articles broadly within the category "Movies").

The relationship between domain concepts and Wikipedia articles is most commonly a one-to-one relationship (phrase "PC" $\rightarrow$ article "Personal Computer"). However, more expressive relationships are possible. For example, unstructured textual phrases can be modeled directly, enabling maps to visualize the approximately 60 million words that appear with regularity in any language edition of Wikipedia. Cartograph uses standard NLP techniques such as named-entity disambiguation to algorithmically map domain concepts to phrases. Additionally, some domain concepts may not appear in Wikipedia explicitly at all and must be modeled as a "bag of articles." In these cases, Cartograph can apply *Wikification* algorithms [33] that take as input unstructured text describing domain concepts and produce as output mentions of Wikipedia articles.

Wikipedia articles (and therefore Cartograph concepts) are designed to be unambiguous. For example, the term "beetle" might be represented by an articles about the insect, the Volkswagen car, and 19 other meanings of beetle. As mentioned above, Cartograph uses algorithms to create these associative mappings. In the film and corporation case studies in this paper we use named-entity entity detection to define this mapping. The case studies of Wikipedia articles require no additional mapping.

While we refer to the structured Wikipedia data as "Wikipedia articles" we note that Cartograph internally uses language-independent representations of articles from the Wikidata project [42], a human-editable database of facts about Wikipedia articles. As we mention later, this enables Cartograph to draw upon both unstructured text related to Wikipedia articles, as well as structured ontologies and attributes related to those entities that are mapped.

Concept prominence: Spatial maps with large datasets must decide which landmarks to show at a particular scale, and how those landmarks should be sized. While geographic maps rely on features such as population to do so, Cartograph extracts information about each concept's prominence from Wikipedia. We experimented with a variety of features related to concept



(a) Content-based movie embedding

(b) Navigation-based movie embedding

Figure 3: Embedding for movies surrounding "2001: A Space Odyssey" using vectors mined from Wikipedia content (left), and user navigation logs (right). Notice that the navigation-based vectors (right) are surrounded by space-oriented movies such as "Lost in Space" and "Event Horizon" while the content-based neighborhood (left) appears more scattered thematically.

prominence. The two most effective features we explored were the Pagerank of articles as measured using the Wikipedia link graph [7], and the number of times each page was viewed.[4]

The Pagerank of articles tended to favor highly interlinked concepts such as "United States", "1997", and "International Standard Book Number." While the generality of these concepts was appealing, the metric seemed to place too much importance on the number of pages that link to a concept. Pageviews, on the other hand, favored popular concepts that trended during the period in which pageviews are counted, such as movies ("Star Wars - The Force Awakens"), politicians ("Donald Trump"), and athletes ("Kobe Bryant.") To mitigate the volatile distribution of page views, we selected the median views for each page from a sample of 100 hours over a one-year period (the median was far more robust than the mean to spikes in interest). We additionally log-transformed the page views to normalize the long-tailed distribution of interest in Wikipedia articles.

Once we computed both page rank and page views, we found that formula below effectively balanced between concept generality and viewer interest, where $P(a)$ calculates a prominence score for article a.

$$P(a) = pageRank(a) * log(median(pageviews(a)))$$

We chose to multiply the two terms described above before we found they had similar importance and variability. The most prominent concepts using this formulation included countries ("United States", "United Kingdom," and many others) prominent figures ("Barack Obama"), internet companies ("Google", "Facebook", "YouTube"), and other similarly broad and notable concepts.

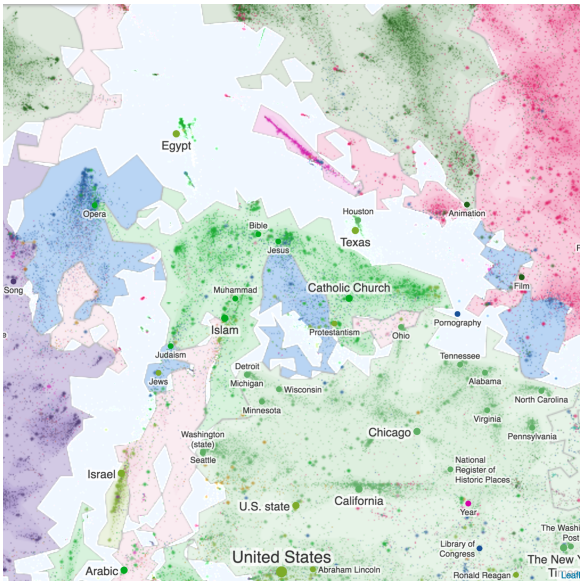
Semantic vectors: Cartograph uses vectors representing each concept to reason about relationships. We experimented with two types of vectors learned from Wikipedia, both based on the Word2Vec algorithm of Mikolov et al. [30], which mines co-occurrence patterns in words within sentences.

The first vector embedding approach analyzed the **content** within Wikipedia pages. To generate these 200-dimensional content-based vectors we applied the Word2Vec algorithm to the entire Wikipedia corpus, with two enhancements to strengthen the vector representations of articles. First, we incorporated the doc2vec algorithm [13] to produce vectors for every article. Second, we used wikification [33] to extract each mention of an article within Wikipedia — whether or not it was hyperlinked. This ensured that each article's representation captured not just the content within the article, but also the context in which it was mentioned throughout the encyclopedia.

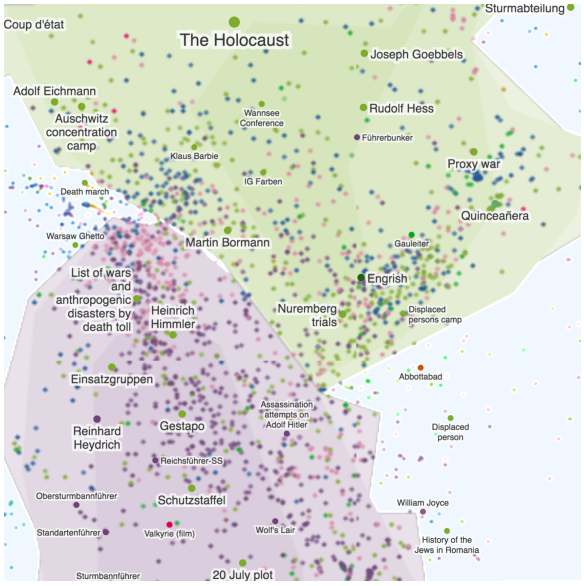
The second vector embedding approach analyzed **navigation** logs within Wikipedia developed by Wulczyn [44] to create 100-dimensional vectors. This model treats user web sessions as sentences, with words corresponding to the articles that were viewed in each session. Correspondingly, this approach mines co-occurrence patterns in visits to article pages. Wulczyn's vectors are trained using approximately 1.6 billion user sessions containing 6.2 billion page views.

We also experimented with vectors that combine the content and navigation approaches via concatenation, but our initial experiments suggested they were not effective. As we mention in our discussion, an open area for future research uses deep learning approaches to combine these techniques.

[4]Pageview statistics for Wikipedia are publicly available from https://dumps.wikimedia.org/other/pagecounts-raw/.



(a) Homogeneous thematic cluster areas



(b) Heterogeneous thematic cluster areas

Figure 4: Two areas within the Cartograph map of Wikipedia articles showing homogeneous (left) and heterogeneous (right) clustering results. Points are colored by their topical group. On the left points are generally colored similarly to their cluster. On the right points show greater variation. The heterogeneous areas are relatively rare within Cartograph.

Figures 3a and 3b compares the navigation and content approaches for vector-creation in movies. A detailed description of the embedding process is described in the next section; here we evaluate the output embeddings resulting from the content-based and navigation-based approaches. In general the output embeddings coming from the two approaches seem similar in quality. However, we noted that in the movie embedding (Figures 3a and 3b), the navigation-based embeddings appeared noticeably superior to the content-based embeddings. While our hypothesis requires more evaluation, our intuition is that the humans largely perceive relationships between two companies based on information that is encoded in Wikipedia, such as the company's industry, its size, and its location. However the relationship humans perceive between two movies is not; a movie's genre, actors, year, plot-line, etc. are not sufficient to capture human semantic understanding of movies. As a result, we use the navigation-based embeddings throughout the remainder of this paper.

Stage 2: X,Y Embedding

Next, Cartograph embeds the domain concepts into the (x, y) plane. The high-dimensional Word2Vec vectors served as the starting point for these (x, y) embeddings. Our goals in the embedding were two-fold: 1) to ensure that neighboring (related) points in the high-dimensional space were also neighbors in the low-dimensional space, and 2) to produce embeddings that appeared "land-like", with variations in density and shape.

We experimented with a variety of embedding algorithms and found that the t-SNE algorithm, which is known to generally produce high-quality embeddings [27], performed well. t-SNE also seemed to yield "natural point formations". At a high level t-SNE embeddings exhibited a dense center region that resembled continents with decreasing density at the edges of the map that that resembled island nations (Figure 1). At a low level t-SNE also exhibited localized variations in density that approximated rural and urban areas (Figure 2).

We found that even the highly optimized t-SNE algorithm described in [41] required 24 hours to produce an embedding for 500,000 points. This stage was, by far, the most time consuming stage in our data pipeline. Therefore, we limited the running time by sampling 500,000 points for the initial embedding. The remaining out-of-sample points were placed by interpolating the locations of each point's in-sample neighbors. Given a point p, we found that p's neighbors in high-dimensional space often spanned vast regions of the (x,y) space. Therefore, we only used points in p's densest (x, y) neighborhood during interpolation.

Stage 3: Country formation

Our country formation procedure roughly follows procedures used by previous spatialization projects [18, 21]:

Clustering: We identify the main groups of semantic topics within a domain concept space by using the kmeans++ algorithm to cluster the high-dimensional vectors. While we acknowledge that some domain areas may have an existing category structure that can be used (e.g. movies have genres), this is not always the case. Cartograph can incorporate existing categories, but this work focuses on inferring topics for data where none is available.

Clusters in the high dimensional space correspond to surprisingly homogeneous areas once embedded in (x,y) space, as shown by the consistent coloring between dots and background colors in the overall view of the Wikipedia map in Figure 1 and the Figure 4a. However,heterogeneous topical areas still remain, as shown in Figure 4b. These areas often correspond

to multi-faceted articles that intuitively lie at the intersection of two topics. For example, the area on the right shows the border between a cluster related to the Holocaust (top, in green), and War (below, in purple). Many articles in the area, such as "Heinrich Himmler", lie at the intersection of these two topics.

Water modeling: We add random "water points" throughout the map, with more points appearing toward the edges of the graph. These points help identify the regions that are dominated by domain concepts versus those that have lower point densities and more "open space." The areas with open space are turned into water regions in later processing stages.

Denoising: We identify areas in the low-dimensional x, y space that are dominated by a single cluster or water. This is achieved using signal-processing techniques from [20]. We temporarily remove outliers that are not members of the area's primary cluster for this phase of processing.

Country borders: We construct borders using a Delaunay triangulation procedure with noising, following the procedure described in [21].

Topological contours: Cartograph produces topological contours for each country. We experimented with both density-based and centrality-based contours. Density-based contours are commonly used for relief maps in spatialization systems, with higher-density areas associated with higher contours. Centrality-based contours reflect the the similarity between each point's vector and the centroid for the country as a whole. We found that the information shown by density contours was already conveyed by the points visualized by Cartograph. Centrality contours, on the other hand, highlighted the areas that were most "representative" of each country.

Stage 4: Domain-specific Data Visualization

Once Cartograph has produced the map features, GIS data visualization techniques can be used to show the relationship between semantic space and any "domain-specific" dataset. For example, in the case studies that follow, we show graphs of Wikipedia article quality ratings and sustainability ratings for companies. GIS approaches such as choropleth maps, dot density visualizations, and heat maps can be used to visualize this data. We also note that the interactive, zoomable nature of the map lends itself well to dot-density visualizations. These visualizations allow one to identify high-level patterns and then zoom in to understand the individual data points contributing to those patterns.

We note that it is not required that the visualization dataset "cover" every domain-specific concept; some domain specific concepts can have missing values, as shown in our case studies. Including missing domain concepts allows users to extrapolate values for a point without data based on patterns in the semantic region.

Stage 5: Map Server

We implemented a custom web-based framework to serve maps that leverages recent advances in map rendering technologies. On the browser side, we used the Tangram javascript

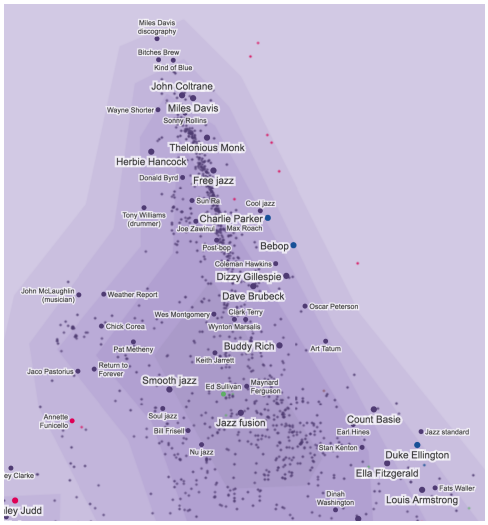


Figure 5: A zoomed in version of the Wikipedia map focused on jazz music.

open source framework[5] to render maps using WebGL,[6] a hardware accelerated rendering engine supported by 92% of browsers as of October 2016.[7]. We implemented a custom map server that serves raster layers for background topology and points, and vector layers for foreground points. To speed up spatial queries, the map server loads data into memory, uses optimized spatial indices, and precomputes and caches both vector and raster layers. While this approach may not be practical for a site that serves block-level imagery of the entire earth, the one-time caching of five million data points only took a few minutes on the Cartograph server. As far as we know, Cartograph is the first spatialization system to make use of the combination of vector, raster, and WebGL technologies that has been effectively used by products such as Mapbox, Bing Maps, and Google Maps.

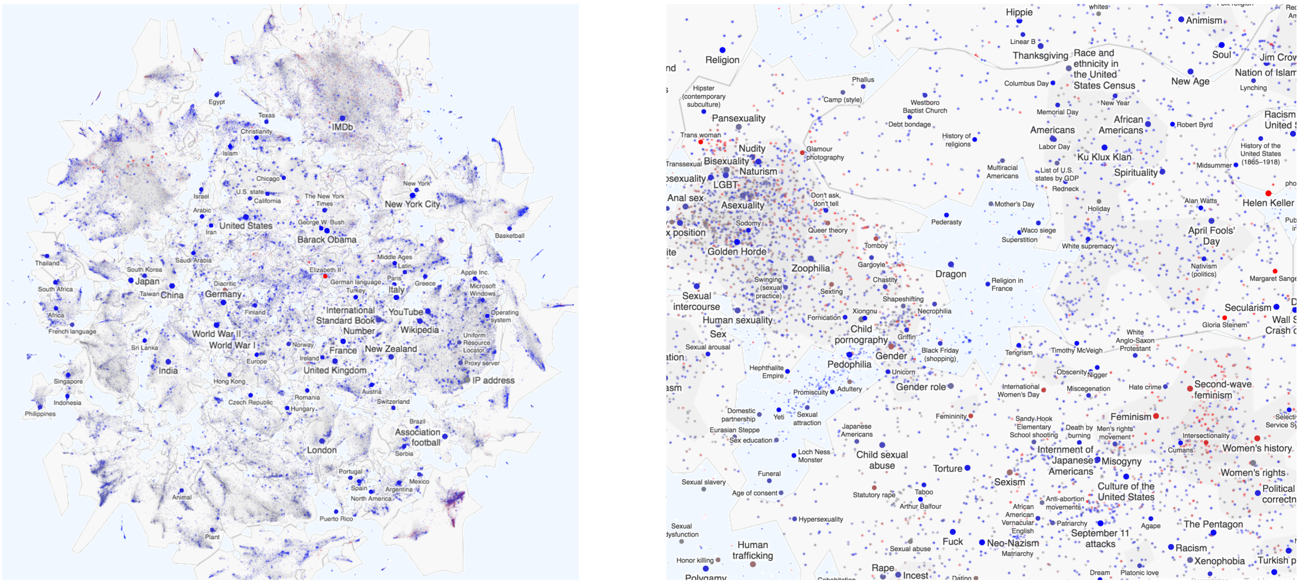
CASE STUDIES

In this section we present case studies of Cartograph maps for three sets of domain concepts.

Map of Wikipedia

The map of all of Wikipedia serves as a test case for a large dataset. In this case, the domain of articles is the approximately 5 million concepts in Wikipedia, limited to the 1.4 million articles that have sufficient pageviews to warrant vectors in the navigation data set. Figure 1 shows the overview of the basic view of the map, with countries colored according to their topical clusters. Sports broadly appear in contiguous regions at the edges of the graph. American football, baseball, and basketball appear in the teal region on the east edge of the map, while soccer appears in red in the bottom. We found these groupings to robustly appear across repeated randomized map recreations. They also always appeared on the outskirts of the map. This suggests that a sport such as baseball exhibits a high degree of local similarity in its articles, but fewer "long

[5]https://mapzen.com/blog/tangram-a-mapping-library/
[6]https://www.khronos.org/registry/webgl/specs/1.0/
[7]http://webglstats.com/



(a) The Wikipedia map of gender focus

(b) Area of the Wikipedia map of gender focus related to feminism and sexuality

Figure 6: The Wikipedia map of gender focus. Blue and red dots correspond to articles that focus on men and women respectively.

distance" similarities to other clusters. Across map iterations, we also found that consistent topical groups appeared for technology (in fuchsia, to the east), movies (in pink, to the north), music (purple, to the north east), and Bollywood (in orange, to the southeast). Figure 5 provides a focused view of the region in the map related to jazz music. Surprising local relationships emerge, with bebop music (John Coltrane, Miles Davis, Thelonious Monk) appearing towards the top, big band music and vocal music appearing in the lower right (Duke Ellington, Count Basie), and more contemporary jazz fusion appearing in the lower left (Return to Forever, Chick Corea).

Figures 6a and 6b show a domain-specific thematic map of Wikipedia, with points colored by the gender focus of the article. Blue articles refer primarily to women, red articles refer primarily to men, and purple articles are more balanced. To collect the gender focus dataset, we used the Wikidata project to identify articles about men and women, and connected people to articles using the Wikipedia link graph. The overview of the map in Figure 6a shows a striking focus on men (blue) throughout Wikipedia. However, some areas of red emerge. Figure 6b focuses on one such area, related to sexuality and feminism. Other areas with a strong female focus include modeling and womens' sports. Areas related to entertainment (musicians, actors, and television personalities) and Greek mythology display a balance of focus on men and women. We return to this map in the pilot study described later.

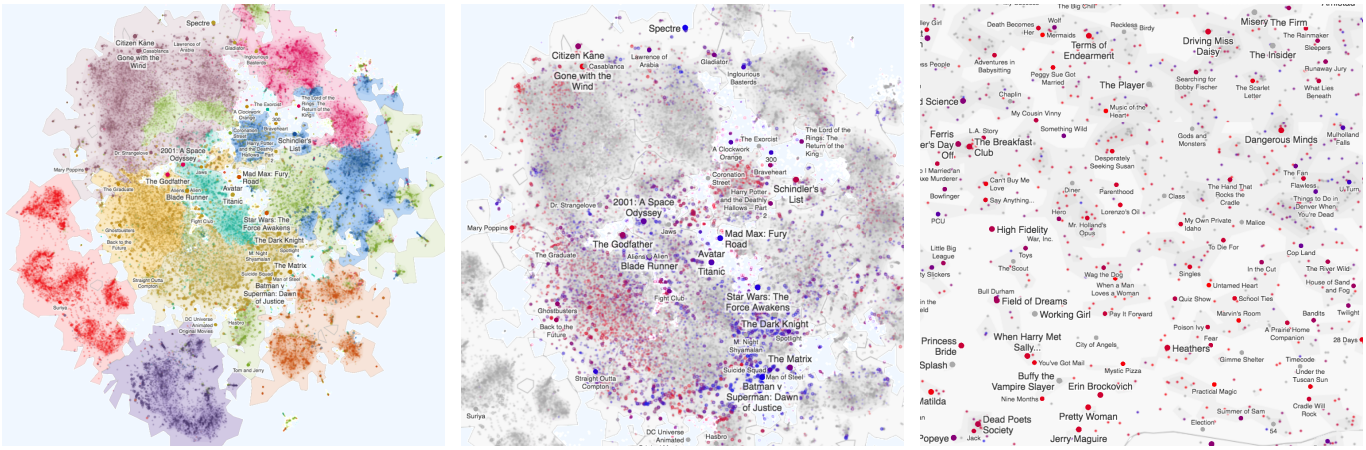
Map of Films

The second case study visualizes the map of films. Since the Wikidata attribute "film" was consistently used to describe movies, this map includes all Wikipedia articles that are marked as film and have a navigation-based vector, representing 72,229 movies. Figure 7a shows the basic thematic map with cities colored by cluster. Bollywood movies appear in the southwest, colored red. Films connected to Asian culture, including anime and martial arts films appear in the southeast in orange. Independent, foreign and art films ("Bicycle Thieves", "Cinema Paradiso") appear in pink in the northeast, and older critically acclaimed movies ("classics") appear in dark purple in the north ("Dr. Strangelove", "Easy Rider"). The middle of the map exhibits more thematic overlap, but yellow is broadly action and comedy ("Ghostbusters", "Rocky", "Platoon").

Figure 7b shows a domain-specific movie layer that visualizes the "gender" of each movie. Movies of more interest to men and women are blue and red respectively. This data was collected from the MovieLens recommender system. Following the procedure of [26] we used the number of times each movie was rated by men and women to assign a "gender score" to each movie, and included all movies that had been rated by at least 20 users with known gender (MovieLens does not require users to specify their gender). While many areas in the top-right image show balanced interest from men and women, several homogenous areas emerge. In particular, the south of the map, showing action movies such as "Deadpool", "Batman v Superman" and "Furious 7" appears predominantly of interest to men. The diagonal red patch in the southwest of the map, shown at a high zoom in Figure 7c features many movies generally referred to as "chick flicks", such as "Sleepless in Seattle" and "Pretty in Pink."

EVALUATIVE FEEDBACK FROM USERS

To better understand how the Cartograph system would work in practice, we deployed a version using the map of Wikipedia articles colored by gender focus shown in Figure 6a and solicited user feedback from members of several groups of Wikipedia editors who contribute to projects related to gender. In this exploratory study, we were less interested in algorithmic performance or user performance with time and error metrics. Instead, our goal was to learn how domain experts interpret the cartographic embedding, how the gender focus



(a) Map of movies colored by topical cluster

(b) Map of movies colored by gender interest

(c) Area of gender map focused on movies of interest to women

Figure 7: The Cartograph map of films. The first image shows topical clusters, while the second two images show movies that exhibit more interest from men (blue) and women (red).

information overlaid on top of the map would help them analyze Wikipedia data, and what we can learn about the design of spatialization tools to support analytical tasks.

Participants

Participants were recruited through postings to the discussion pages for three WikiProjects related to gender: the "Gender-Gap task force", "Women and Red", and "Feminism" . Each of these task forces contribute to Wikipedia in ways that address systemic gender bias in Wikipedia articles. Six participants (three female) completed the study successfully. A seventh user attempted the study but did not compete it successfully. He did not provide any feedback on the tasks and notified the authors that he had not realized the system could zoom. His results were removed before analysis. Although six is too low a number of participants to draw any statistical conclusions, from users' qualitative feedback we are able to identify common patterns in usage. Participants' ages range between 24–53 (median: 40 years). All of the participants edit Wikipedia articles at least yearly, with three out of the six editing monthly.

Methodology and Tasks

Each of the participants performed three tasks, structured to model specific exploratory visualization tasks:

1. Locate – Identify Wikipedia articles with the highest women's gender focus.
2. Identify Distribution, Associate, and Correlate – Describe the common characteristics of articles with a high women's gender focus and how they are related to other articles nearby in the map that have a higher male focus.
3. Browsing – Explore the map while noting observations.

The tasks were presented in a panel on the right side of the map visualization that contained a text box for participants to enter their feedback. The order of the three tasks was randomized to avoid any learning effects. Prior to starting the first task, participants were given an interactive tutorial of the system using IntroJS [3] that walked them step-by-step through the user interface. After the third trial was complete participants answered a short survey consisting of demographic information and the Likert scale questions shown in Table 1.

Analytic Strategies Using Cartograph

The results indicate that Cartograph enables users to identify overall patterns within the data and dive deeper to identify more complex relationships. One participant mentioned that her "first reaction ... is 'wow that's a small number of red dots' - but beyond that, it's a UNIFORMLY[sic] small number. I'd expect a higher proportion of women and women-focused articles in areas traditionally considered more 'feminised'. And it may well be that there are but they don't make it to the top view (and are hard to find) because those areas are themselves underrepresented and underlinked." By looking more closely at individual articles, several participants found that the articles with a high female gender focus are about actual women who existed as opposed to topic and idea articles which are male dominated. This finding may serve the WikiProjects as they select articles to focus on.

In general, the relationships between article regions representing countries were clear. One participant identified that "the links surrounding articles on women, or articles on places about women, seem to be education-related". Another said "Janet Jackson and Beyoncé [articles] seem to be clustered with music related nodes. Elizabeth II is clustered with other world leaders and European countries. Diacritic seems to be clustered with northern European countries and languages".

To identify patterns and relationships in the map participants used one of two general strategies: (1) They zoomed in to a high level of detail and then panned around the map, or (2) They zoomed in on specific areas, explored, and then zoomed back out to get additional context before zooming in again on other areas of the map. This can be seen in the logging data as participants completed the task.

Table 1: Mean survey question responses on a seven point Likert Scale (higher values indicate positive agreement).

How quickly could you achieve your tasks?	4.8
The tool required a lot of explanation to use.	3.5
It was unclear why specific articles were grouped together.	4.6
I learned new information about the data.	5.6
The tool was easy to use.	5.8
The tool was fun to use.	6
How successful were you in accomplishing what you were asked to do?	5.5

Table 2: Results for exploring gender focus in Wikipedia articles.

Participant	Time Spent	Pan Operations	Number of Searches	Article Clicks
Participant 0	6:22	64	0	26
Participant 1	13.23	0	0	1
Participant 2	6:00	31	6	3
Participant 3	14.26	87	1	7
Participant 4	37:24:23	116	0	5
Participant 5	12:30	80	1	54

Feedback on Specific Features

Table 2 shows the number of times participants panned from one area of the map to another, the number of searches they performed, and the number of articles that they clicked on to receive further information. It also shows the total time that the visualization was open in participants' browsers. Note that participants 4 took a long break in the middle of completing the tasks before returning to finish.

Participants primarily panned and zoomed around the map without frequent use of the search feature. This is consistent with the exploratory nature of the tasks rather than more focused searching. User feedback also shows that an increased ability to see more information about article relationships is needed. Cartograph enables users to find new relationships, but it does not directly explain what those relationships are beyond a high-level idea of similarity. All of the participants spent time clicking on articles to read descriptions in an attempt to gain more insight into their similarity.

We are enthusiastic about the potential Cartograph has for visualizing data that does not originally contain any spatial or semantic relatedness components, perhaps through additional data layers. One participant acknowledged how an article's gender focus could serve as one layer, while other demographic statistics could be represented at the same time on other layers. These data overlays could include additional visual glyphs or heatmaps to visualize more data that may help with analysis.

CONCLUSION

This paper introduces Cartograph, a system that unlocks thematic cartography for diverse data. Cartograph supports nearly universal spatialization, transforming a dataset with no existing semantic information - for example, business names and sustainability ratings - into an interactive thematic map grounded in semantic information mined from Wikipedia. Its use of recent advances in mapping technology affords multi-scale analyses that support datasets containing millions of points in a web browser, an order of magnitude larger than previous efforts.

While our exploratory study of the initial Cartograph system yielded generally positive feedback, it also suggested a variety of areas for future research.

Several users stated that they were confused by the relationship between neighboring cities. This shortcoming might be addressed in a variety of different ways. First, the 2D embeddings could be directly improved. This might be accomplished by creating a hybrid vector representation that combines the content-based and navigation-based vectors using a deep learning. Second, the embeddings and the clustering could be jointly constructed in a way that encourages more homogeneous clusters. This might increase the effectiveness of the "country" metaphor and help delineate boundaries neighboring between points that are in between semantic clusters. Third, and perhaps most promisingly, Cartograph could be enhanced so that it goes beyond displaying semantic neighbors to *explaining* semantic neighbors. With this goal in mind, we have been experimenting with adding labeled "roads" to the map to describe relationships between points. To understand these questions, we plan to use Cartograph to conduct larger-scale studies of more varied datasets and tasks. In particular, we would like to understand whether answers to the question above vary depending on a user's task.

Cartograph, as currently designed, visualizes a static set of concepts. It creates a single initial map that does not reflect longitudinal changes in a particular domain's information space. Ideally, Cartograph would use incremental forms of clustering and embedding algorithms that start with an initial map and incrementally adapt as the information changes. Alternately Cartograph could draw inspiration from recent research that has experimented with alternative interaction patterns, including visualizing dynamic data [29] and surfacing personalized recommendations [16]. Cartograph's interactive and scalable design makes it a good fit for experimenting with interactions including these and others.

Cartograph could be enhanced to provide services to end users with no programming skills. While Cartograph's source code is publicly available[8], mapping new datasets requires command line expertise that limits the potential audience for our approach. In the future, we hope to extend the system with a web-based map-creator interface and API so that end users, third party websites and data analysis tools can incorporate the research advances in this paper into their own work.

ACKNOWLEDGMENTS

This research is generally supported through grants from the Clare Boothe Luce Foundation, the National Science Foundation (IIS 1526988, III 1421655, and IIS 1527173), and a Wallace Scholarly Activities Grant from Macalester College. The authors would like to thank David Shuman for suggesting the graph-based denoising algorithm we used and Ashley Nepp for her general feedback about spatial visualization. Finally we thank Wikipedians who participated in the user study for their feedback and Heather Ford and Aaron Halfaker for helping to identify these participants.

[8] http://cartograph.info

REFERENCES

1. 2016. 506 Sports: NFL TV Schedule and Maps. http://506sports.com/nfl.php?yr=2016&wk=6. (2016). Accessed: 2016-10-14.
2. 2016. FiveThirdyEight 2016 Election Forecast. http://projects.fivethirtyeight.com/2016-election-forecast/. (2016). Accessed: 2016-10-14.
3. 2016. IntroJS. http://introjs.com/. (2016). Accessed: 2016-10-14.
4. 2016. NASA Earth Observatory. http://earthobservatory.nasa.gov/Features/WorldOfChange/decadaltemp.php. (2016). Accessed: 2016-10-14.
5. 2016. The Upshot 50 Years of Electoral College Maps: How the U.S. Turned Red and Blue. http://nyti.ms/2bwoJsf. (2016). Accessed: 2016-10-14.
6. Benjamin Adams, Grant McKenzie, and Mark Gahegan. 2015. Frankenplace: Interactive Thematic Mapping for Ad Hoc Exploratory Search. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. ACM, New York, NY, USA, 12–22.
7. Francesco Bellomi and Roberto Bonato. 2005. Network analysis for Wikipedia. In *Proceedings of the 1st International Wikimedia Conference, Wikimania 2005*. Wikimedia Foundation.
8. U. Brandes and T. Willhalm. 2002. Visualization of Bibliographic Networks with a Reshaped Landscape Metaphor. In *Proceedings of the Symposium on Data Visualization (VISSYM '02)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 159–ff.
9. Matthew Chalmers. 1993. Using a landscape metaphor to represent a corpus of documents. In *Proceedings of Spatial Information Theory A Theoretical Basis for GIS: European Conference*, Andrew U. Frank and Irene Campari (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 377–390.
10. Chaomei Chen and Timothy Cribbin. 2001. A Study of Navigation Strategies in Spatial-semantic Visualizations. In *Proceedings of the 9th International Conference on Human-Computer Interaction*. 948–952.
11. H. Couclelis. 1998. Worlds of Information: The Geographic Metaphor in the Visualization of Complex Information. *Cartography and Geographic Information Systems* 25 (1998), 209–220.
12. 'Donna Cox. 2006. Metaphoric mappings: The art of visualization. In *Aesthetic computing*, P. Fishwick (Ed.). MIT Press, Cambridge, MA, 89–114.
13. Andrew M Dai, Christopher Olah, and Quoc V Le. 2015. Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998* (2015).
14. S. I. Fabrikant, D. R. Monteilo, and D. M. Mark. 2006. The distance-similarity metaphor in region-display spatializations. *IEEE Computer Graphics and Applications* 26, 4 (July 2006), 34–44.
15. Christiaan Fluit, Marta Sabou, and Frank van Harmelen. 2006. Ontology-Based Information Visualization: Toward Semantic Web Applications. In *Visualizing the Semantic Web: XML-Based Internet and Information Visualization*, Vladimir Geroimenko and Chaomei Chen (Eds.). Springer London, London, 45–58.
16. Emden Gansner, Yifan Hu, Stephen Kobourov, and Chris Volinsky. 2009. Putting Recommendations on the Map: Visualizing Clusters and Relations. In *Proceedings of the Third ACM Conference on Recommender Systems (RecSys '09)*. ACM, New York, NY, USA, 345–348.
17. Emden R. Gansner, Yifan Hu, and Stephen G. Kobourov. 2010. GMap: Visualizing graphs and clusters as maps. In *IEEE Pacific Visualization Symposium PacificVis 2010, Taipei, Taiwan, March 2-5, 2010*. 201–208.
18. Martin Gronemann and Michael Jünger. 2013. Drawing Clustered Graphs As Topographic Maps. In *Proceedings of the 20th International Conference on Graph Drawing (GD'12)*. Springer-Verlag, Berlin, Heidelberg, 426–438.
19. Brent Hecht, Samuel H. Carton, Mahmood Quaderi, Johannes Schöning, Martin Raubal, Darren Gergle, and Doug Downey. 2012. Explanatory Semantic Relatedness and Explicit Spatialization for Exploratory Search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, New York, NY, USA, 415–424.
20. Matthias Hein and Markus Maier. 2006. Manifold denoising. In *Advances in neural information processing systems*. 561–568.
21. Yifan Hu, Stephen Kobourov, and Emden R. Gansner. 2010. Visualizing Graphs and Clusters as Maps. *IEEE Computer Graphics and Applications* 30 (2010), 54–66.
22. Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410* (2016).
23. Robert R. Korfhage. 1991. To See, or Not to See&Mdash; is That the Query?. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '91)*. ACM, New York, NY, USA, 134–141.
24. Sherry Koshman. 2006. Visualization-based information retrieval on the Web. *Library & Information Science Research* 28, 2 (2006), 192–207.
25. Werner Kuhn and Brad Blumenthal. 1996. Spatialization: Spatial Metaphors for User Interfaces. In *Conference Companion on Human Factors in Computing Systems (CHI '96)*. ACM, New York, NY, USA, 346–347.
26. Shyong Tony K Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R Musicant, Loren Terveen, and John Riedl. 2011. WP: clubhouse?: an exploration of Wikipedia's gender imbalance. In *Proceedings of the 7th international symposium on Wikis and open collaboration*. ACM, 1–10.

27. Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

28. Alan M. MacEachren. 1982. The Role of Complexity and Symbolization Method in Thematic Map Effectiveness. *Annals of the Association of American Geographers* 72, 4 (1982), 495–513.

29. Daisuke Mashima, Stephen Kobourov, and Yifan Hu. 2012. Visualizing dynamic data with maps. *IEEE Transactions on Visualization and Computer Graphics* 18, 9 (2012), 1424–1437.

30. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3111–3119.

31. Daniel R. Montello, Sara Irina Fabrikant, Marco Ruocco, and Richard S. Middleton. 2003. Testing the First Law of Cognitive Geography on Point-Display Spatializations. In *Proceedings of Proceedings, Conference on Spatial Information Theory (COSIT '03), Lecture Notes in Computer Science 2825, Ittingen*. Springer, 24–28.

32. Lev Nachmanson, Roman Prutkin, Bongshin Lee, Nathalie Henry Riche, Alexander E. Holroyd, and Xiaoji Chen. 2015. GraphMaps: Browsing Large Graphs as Interactive Maps. In *Graph Drawing and Network Visualization: 23rd International Symposium*, Emilio Di Giacomo and Anna Lubiw (Eds.). Springer International Publishing, 3–15.

33. Thanapon Noraset, Chandra Bhagavatula, and Doug Downey. 2014. Adding High-Precision Links to Wikipedia.. In *EMNLP*. Citeseer, 651–656.

34. Randall M. Rohrer and Edward Swing. 1997. Web-Based Information Visualization. *IEEE Compututer Graphics Applications* 17, 4 (July 1997), 52–59.

35. Shilad Sen, Toby Jia-Jun Li, WikiBrain Team, and Brent Hecht. 2014. Wikibrain: democratizing computation on wikipedia. In *Proceedings of The International Symposium on Open Collaboration*. ACM, 27.

36. André Skupin. 2000. From Metaphor to Method: Cartographic Perspectives on Information Visualization. In *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS '00)*. IEEE Computer Society, Washington, DC, USA, 91–.

37. André Skupin and Sara Irina Fabrikant. 2003. Spatialization Methods: A Cartographic Research Agenda for Non-geographic Information Visualization. *Cartography and Geographic Information Science* 30, 2 (2003), 99–119.

38. Terry A. Slocum, Robert B. McMaster, Fritz C. Kessler, and Hugh H. Howard. 2009. *Thematic Cartography and Geovisualization* (3 ed.). Prentice Hall, Saddle River, NJ, USA.

39. Walter R. Tobler. 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 46 (1970), 234–240.

40. M. Tory, C. Swindells, and R. Dreezer. 2009. Comparing Dot and Landscape Spatializations for Visual Memory Differences. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (Nov 2009), 1033–1040.

41. Laurens Van Der Maaten. 2014. Accelerating t-SNE using tree-based algorithms. *Journal of machine learning research* 15, 1 (2014), 3221–3245.

42. Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

43. Ryen W White and Resa A Roth. 2009. Exploratory search: beyond the query-response paradigm (Synthesis lectures on information concepts, retrieval & services). (2009).

44. Ellery Wulczyn. 2016. Wikipedia Navigation Vectors. (10 2016). DOI: http://dx.doi.org/10.6084/m9.figshare.3146878.v3