Toward Universal Spatialization Through Wikipedia-Based Semantic Enhancement

SHILAD SEN, ANJA BETH SWOAP, QISHENG LI, ILSE DIPPENAAR, MONICA NGO, SARAH PUJOL, REBECCA GOLD, and BROOKE BOATMAN, Macalester College
BRENT HECHT, Northwestern University
BRET JACKSON, Macalester College

This article introduces Cartograph, a visualization system that harnesses the vast 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. Although 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 datasets 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. We present findings from two user studies exploring design choices and use of the system.

CCS Concepts: • **Human-centered computing → Visualization systems and tools**; *Wikis*;

Additional Key Words and Phrases: Thematic cartography, maps, Wikipedia, neural networks, Wikidata, semantic relatedness, user studies

ACM Reference format:
Shilad Sen, Anja Beth Swoap, Qisheng Li, Ilse Dippenaar, Monica Ngo, Sarah Pujol, Rebecca Gold, Brooke Boatman, Brent Hecht, and Bret Jackson. 2019. Toward Universal Spatialization Through Wikipedia-Based Semantic Enhancement. *ACM Trans. Interact. Intell. Syst.* 9, 2–3, Article 12 (April 2019), 29 pages.
https://doi.org/10.1145/3213769

1 INTRODUCTION

For hundreds of years, humans have leveraged *thematic cartography* as a powerful means to quickly and effectively communicate complex geographic distributions [55]. Thematic cartography helps us understand and explore multi-faceted spatial processes ranging from election results [2, 5] to

The reviewing of this article was managed by special issue associate editors Liang Gou and Anbang Xu.
This research was 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.
Authors' addresses: S. Sen, A. Beth Swoap, Q. Li, I. Dippenaar, M. Ngo, S. Pujol, R. Gold, B. Boatman, and B. Jackson, Macalester College, 1600 Grand Ave. St. Paul, MN, 55105; emails: ssen@macalester.edu, anja.beth@gmail.com, liqs@uw.edu, {ilsedipp, ngonmonica94, rebeccagold0, bboatman1241}@gmail.com, bjackson@macalester.edu; B. Hecht, Northwestern University, 633 Clark St, Evanston, IL 60208; email: bhecht@northwestern.edu.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
2160-6455/2019/04-ART12 $15.00
https://doi.org/10.1145/3213769

climate change [4] to sports broadcast availability [1]. As the quantity and diversity of geographically referenced data increases, thematic maps—often of the interactive variety—now frequently appear in news articles, blog posts, educational applications, and many other contexts.

One reason thematic cartography has proven so useful is that it offers several widely established communicative benefits [39, 55]. Most notably [21], 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. county), (2) communicating *regional patterns* (e.g., the vote share in the Great Plains of the United States), 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 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 article, 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 [61]), 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," whereas concepts about U.S. culture and politics ("Barack Obama," "Chicago," "NY Times") appear in the middle in green. As we will describe In the following, 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 in to the map shows several surprising and semantically grounded regional patterns. Although 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, ExxonMobile) 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.

[1] http://cartograph.info.
[2] https://www.csrhub.com/.

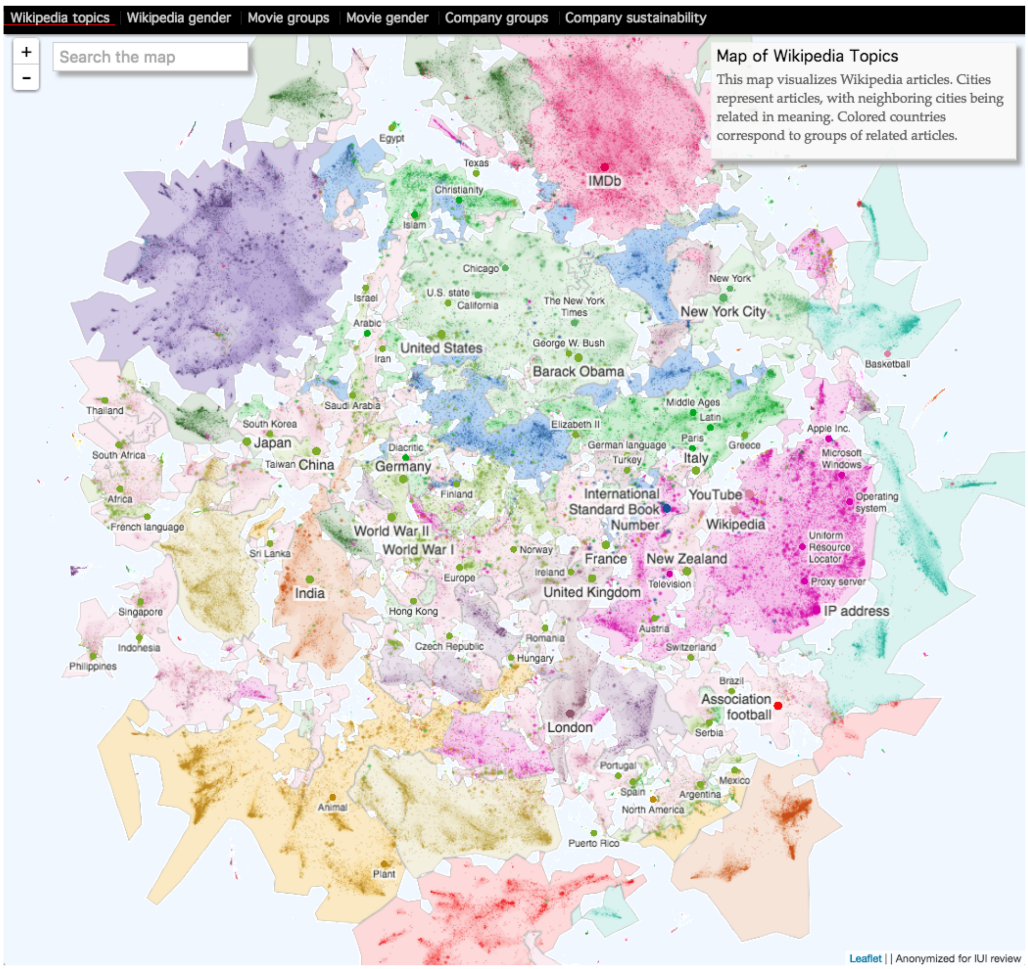


Fig. 1. An overview of the Cartograph system webpage. A user can select a map using the title bar at the top and search for a concept using the box in the upper left. The map supports fluid zoom and pan interactions for large datasets with millions of points. In this map, colors represent semantic topics.

Similarly, because a persistent base map is innate to the Cartograph approach, users can correlate what they learn about company sustainability with all the other datasets they visualize using Cartograph. In other words, through the use of a persistent base map, Cartograph reinforces the indexing and updating of a mental map, a key benefit of cartographic approaches. This work describes the Cartograph spatialization system [51], detailing two methods for generating a base map: one that uses Wikipedia content and another that uses Wikipedia navigation behavior. Since choosing a single base map is critical to gaining the mental maps benefits of cartography, we ran a study with 331 participants who compared the navigation base map to the content base map. Our results suggest that for most users, the navigation-based approach is preferable because its layout of concepts appears to better align with user understandings of conceptual regions.

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. Although 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 earlier. Specifically,

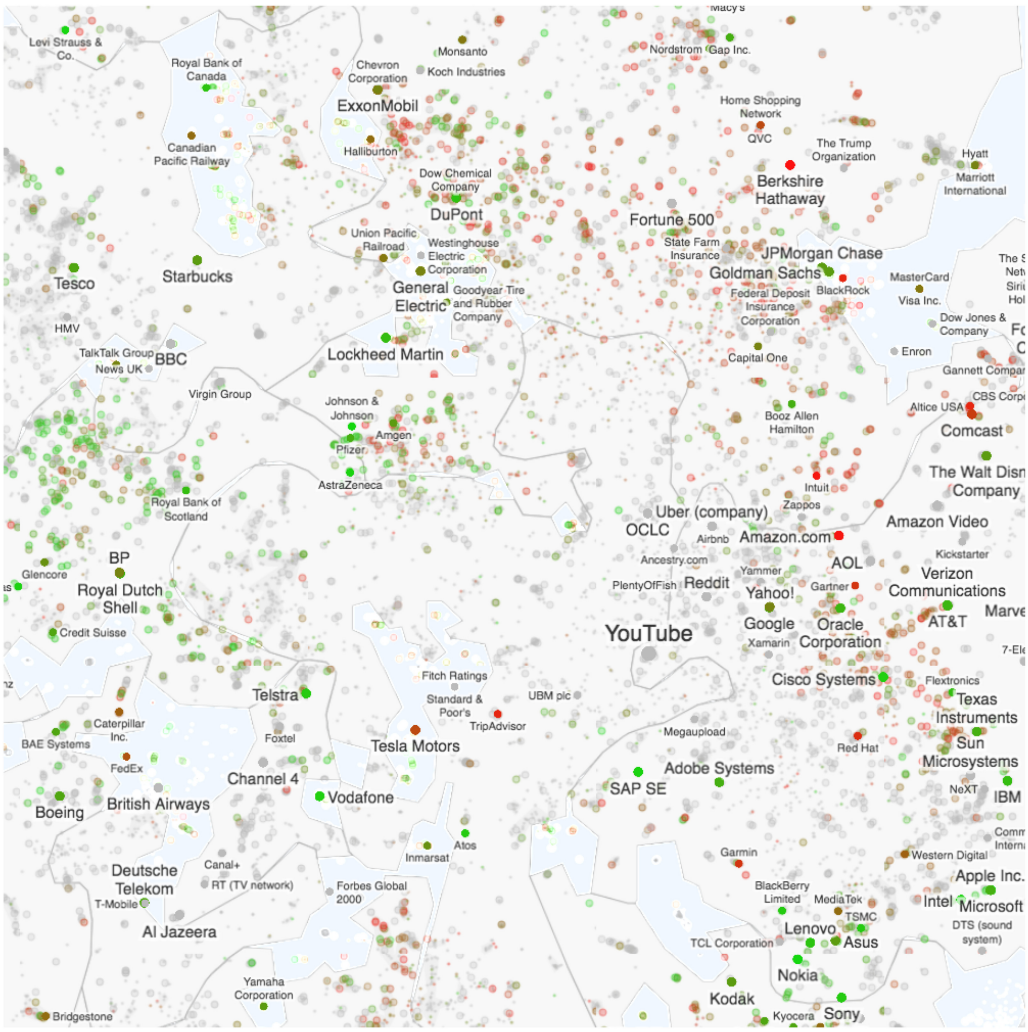


Fig. 2. A map of businesses visualizing sustainability corporate sustainability ratings from low (red) to high (green).

existing spatialization systems either (1) cannot utilize a consistent base map, eliminating the mental map benefits of thematic cartography [13, 34, 54] (as discussed earlier) or (2) are limited to a small family of non-geographic visualizations [6, 21]. 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 [21]. 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, TiiS researchers do not typically have Wikipedia articles

about them. However, we could use algorithms to identify Wikipedia concepts mentioned in 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 that explicitly leverages the relationship between data and semantics [62].

In the following, 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, followed by a user study conducted to identify the best method for creating the base map. We close the article by presenting an exploratory user study that provides insights into the strengths, weaknesses, and usage patterns of the system.

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

2 RELATED WORK

Our work builds on prior research that also visualizes data using cartographic metaphors created by embedding higher-dimensional data into a 2D or 3D map.

2.1 Cartographic Map Visualization

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 [14]. Spatialization is a specific form of digital visual metaphor that maps non-spatial data onto cartographic maps [13, 34].

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" [58]. This distance-similarity relationship is one of the founding principles of geographic analysis [53], and it has been shown to hold for spatializations representing non-spatial data as dots placed in a 2D or 3D space [17, 43]. In thematic cartography, the distance-similarity metaphor is critical to supporting one of the three key benefits of thematic cartography listed earlier: 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), 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 [10, 11, 31] or the World Wide Web [47] 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 like contours or lakes (e.g., Gronemann and Jünger [20]).

Of particular note is the GMap system [19, 24], which presents an algorithm to produce cartographic maps from graphs using clusters as country regions. Although GMap is based on graph

[3] http://cartograph.info.

data, whereas 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. [18] show how recommendations can be displayed using a heat map 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.

As noted earlier, perhaps the most important distinction between Cartograph and previous work is that Cartograph works with almost any data. Although research suggests that traditional spatialization visualizations promote discovery of similarities, clusters, and outliers (important criteria for any exploratory visualization) [12, 54, 59], traditional approaches require semantic-relatedness (SR) 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 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 the Atlasify system of Hecht et al. [21] and related systems like Frankenplace [6]. Atlasify uses SR data for "explicit spatialization" to map data onto various spatial reference systems, including a periodic table, a U.S. 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. Although 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. [21] write that the Atlasify approach could be extended to arbitrary domains through an approach like Cartograph.

2.2 Evaluating Spatialization Systems

Although many spatialization systems exist, there is little agreement on how to evaluate the 2D embeddings produced by the dimensional reduction techniques [37]. One approach is to use metrics to algorithmically score the quality of these visualizations. Albuquerque et al. [7] propose the *class density metric* using image processing techniques to measure variance and density to find embeddings that show correlation or clearly separate the data into clusters. The *class consistency metric* proposed by Sips et al. [52] also supports finding embeddings that maintain good separation of data clusters, but rather than using image processing techniques, they propose measuring centroid distance and distribution consistency of the data during embedding. Martins et al. [40] focus on metrics to identify neighborhood errors, where the nearest neighbors of a point are different in the original and projected space. Recognizing that each of this individual metrics are optimized for preserving specific types of structures, like clusters or outliers, through dimensional reduction, Johansson and Johansson [27] propose a method of user-defined weighting of multiple metrics to preserve as many structures as possible.

Despite the metrics that have been presented in prior work, almost none report on user evaluation [9]. There is an assumption that algorithmic rankings match user-defined rankings; however, this is not universally true. A study comparing the rankings of the class density metric [7] and class consistency metric [52] with user rankings found that the class consistency metric more closely matched the user's perceptions and that occasionally they both differed significantly from

human perception judgments [57]. Notably, Sedlmair et al. [49] visually evaluated more than 800 scatterplots and found that in more than two-thirds of the cases using real data, the metrics failed to identify plausible results [49]. Given these findings, we are hesitant to use algorithmic metrics to evaluate embedding design options until further studies have clarified where they sufficiently match human perception judgments for quality embeddings.

Empirical user evaluations with perceptual judgments are another way to evaluate embedding methods. Perceptual judgment evaluations typically present visualizations of embeddings from two different techniques side-by-side and invite participants to conclude that the preferred embedding is superior to rival embeddings. Often, these comparisons are simply presented in the published work rather than as empirical studies [26, 35], such as the comparison of t-SNE to prior embedding techniques [38] or indeed our own initial comparison of content vs. navigation vectors for Cartograph [51]. In this extension to Cartograph, we expanded our evaluation of the design options to specifically include an empirical study.

Work that includes empirical perceptual studies frequently focuses on specific judgments of visualization features (e.g., class separation [48]). Etemadpour et al. [16] present a taxonomy of possible study tasks beyond class separation, including cluster identification, similarity seeking, cluster ranking, comparison, and counting objects. Our work proposes an approach to evaluating embedding alternatives for Cartograph using label prediction accuracy, as described in Section 5.

3 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.

3.1 Overview of System

Figure 1 shows the Cartograph system. This map visualizes 1.3 million Wikipedia articles, with the most prominent articles ("Barack Obama," "World War II," etc.) labeled as "cities" on the map. Colors correspond to thematic groups of articles that are semantically related. For example, the pink area at the top of the map primarily covers "Film." Users can focus on a specific article using the search box in the upper left. The map supports standard zoom and pan interactions used on modern web maps to see a focused detailed view or 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 [33]). 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 [44]. Users can also click on a particular "city" to get more information about an article.

Cartograph combines a four-stage offline batch data pipeline with an online map server. We summarize the stages next 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 2D x,y coordinate space used for visualization. The high-dimensional space (typically 100D to 600D dense vectors [28]) supports semantic needs, such as neighbor extraction and clustering. The 2D 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 (e.g., 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 both vector- 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.

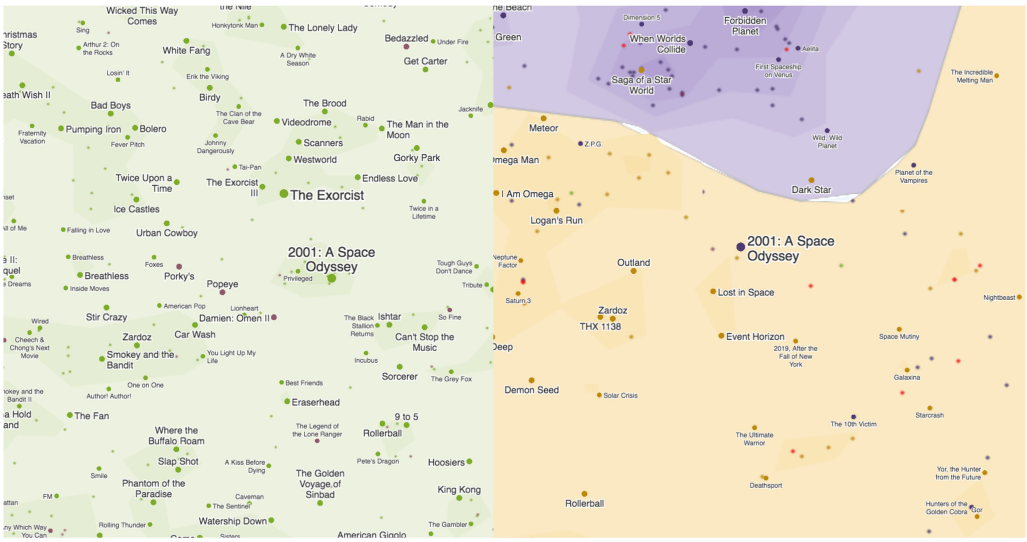
3.2 Stage 1: Concept Definition

The domain concept definition stage produces the raw inputs for the Cartograph system. Throughout the concept definition stage, Cartograph uses the WikiBrain system [50] 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* that 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" → 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 [45] 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 earlier, Cartograph uses algorithms to create these associative mappings. In the absence of any context, ambiguous phrases can be resolved to the most common form using named-entity-detection algorithms (e.g., in Wikipedia intra-article anchor texts, "beetle" most often refers to the insect). However, typically datasets will have some context. In our case studies, we have one dataset about "films" and a second about "corporations." We use the dataset category (e.g., "film") as context to the named entity detection algorithm. Although it



(a) Content-based movie embedding **(b)** Navigation-based movie embedding

Fig. 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," whereas the content-based neighborhood (left) appears more scattered thematically.

was not necessary for our case studies, we have also experimented with a technique for resolving ambiguous concepts that preferentially selects meanings that are closely related to other unambiguous concepts in the dataset. This effectively "bootstraps" a dataset under the assumption that concepts within a dataset are generally more likely to be related than those not within the dataset.

Although 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 [61], a human-editable database of facts about Wikipedia articles. As we mention later, this enables Cartograph to draw on 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. Although geographic maps rely on features such as population to do so, Cartograph extracts information about each concept's prominence from Wikipedia. Inspired by cartographic approaches, we experimented with a variety of approaches for calculating prominence based on the popularity of a concept (as measured by viewer interest) and generality of a concept (as measured by the structure of content within Wikipedia articles). We incorporated two metrics for prominence focused on each of these goals.

We used the PageRank of articles and the number of times each page was viewed as prominence metrics in our current prototype.[4] A concept's PageRank, as measured using the Wikipedia link graph [8], favored general concepts. PageRank's ability to measure importance within a noisy network fit our needs, and it has been used as a measure of prominence in the past [56]. As shown in Table 1, articles with a high PageRank tended to be highly linked concepts such as "IP Address," "ISBN," and "United States." Page views, however, favored popular concepts that trended during

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

Table 1. Concept Prominence According to Four Approaches

	PageRank	Mean Views
1.	IP address	Main Page
2.	ISBN	XHamster
3.	United States	Darth Vader
4.	Network Address Translation	Google
5.	Tilde	Test
6.	Geographic Coordinate System	Deaths in 2017
7.	Wikipedia	XXX
8.	Virtual International Authority File	Comparison of Feed Aggregators
9.	Digital object identifier	Angle of View
10.	Diacritic	Baahubali 2: The Conclusion

	Median Views	Hybrid
1.	Main Page	Main Page
2.	XHamster	United States
3.	Darth Vader	Wikipedia
4.	Deaths in 2017	IP Address
5.	XXX	ISBN
6.	Comparson of Feed Aggregators	United Kingdom
7.	Earth	Australia
8.	Wikipedia	Canada
9.	Donald Trump	Germany
10.	Test	India

The PageRank approaches favor highly linked concepts. The view-based approaches cater to users' time-specific information needs. The Hybrid approach balances between the two.

the period in which page views 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 1-year period. Our current prototype uses the median value because we expect it to be more robust than the mean to short-term spikes in interest in news-related articles, but future work should carefully compare different representative metrics for page views. We additionally log-transformed the page views to normalize the long-tailed distribution of interest in Wikipedia articles.

Once we computed both PageRank and page views, we found that the following formula 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 multiplied the two terms described previously because we found that 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.

Concept prominence affects several dimensions of the Cartograph visualization: the sizing of dots and labels for landmarks, the choice of landmarks for a particular zoom level, and the ordering for autocomplete search results. Although the preceding analyses of prominence suggest that it

aligns with human intuition of concept prominence better than more simple measures such as mean views per article, this area deserves further experimental study.

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. [42], which mines co-occurrence patterns in words within sentences.

The first vector embedding approach analyzed the *content* within Wikipedia pages. To generate these 200D 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 [15] to produce vectors for every article. Second, we used Wikification [45] to extract each mention of an article within Wikipedia—whether or not it was hyperlinked. This ensured that each article's representation captured not only 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 [63] to create 100D vectors. This model uses data browsing behavior as input. Navigation logs are mined for individual users' browsing sessions.[5] For each individual browser session, the list of viewed articles is constructed. The algorithm treats this as a corpus, where a "sentence" corresponds to a browser session and a "word" within a sentence corresponds to a single viewed article within a session. We then apply the standard Word2Vec algorithm to this corpus. Correspondingly, this approach mines co-occurrence patterns in visits to article pages. The vectors of Wulczyn [63] are trained using approximately 1.6 billion user sessions containing 6.2 billion page views.

We explored vectors that combine the content and navigation approaches via concatenation, but this produced maps with counter-intuitive neighborhoods. As we mention in our discussion, an open area for future research uses deep learning approaches to combine these techniques.

Figure 3(a) and (b) compare 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- 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 (Figure 3(a) and (b)), the navigation-based embeddings appeared noticeably superior to the content-based embeddings. 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 that humans perceive between two movies is not; a movie's genre, actors, year, plot line, and so forth, are not sufficient to capture human semantic understanding of movies. In Section 5, we present a novel grounded evaluation approach that compares the content vectors and navigation vectors and finds evidence supporting the use of navigation vectors. As a result, we use the navigation-based embeddings throughout the remainder of this article.

3.3 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 twofold: (1) to ensure that neighboring (related) points in the high-dimensional space

[5] A unique user is identified by the user IP, user-agent string, and several other characteristics. A session is defined as activity without a 30-minute gap. More details are available at https://meta.wikimedia.org/wiki/Research:Wikipedia_Navigation_Vectors.

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 including PCA, local-linear embedding, IsoMap, and several others. We found that the t-SNE algorithm, which is known to generally produce high-quality embeddings [38], performed well. Consistent with previous research, the t-SNE (x,y) embeddings preserved a higher proportion of neighbors from the high-dimensional space. t-SNE's preservation of high-dimensional densities in the low-dimensional space 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 (see Figure 1). At a low level, t-SNE also exhibited localized variations in density that approximated rural and urban areas (see Figure 2). Many of the dense areas corresponded to large groups of closely related articles, such as articles about Bollywood movies or soccer players.

We found that even the highly optimized t-SNE algorithm described in van der Maaten [60] 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.

3.4 Stage 3: Country Formation

Our country formation procedure roughly follows procedures used by previous spatialization projects [20, 24].

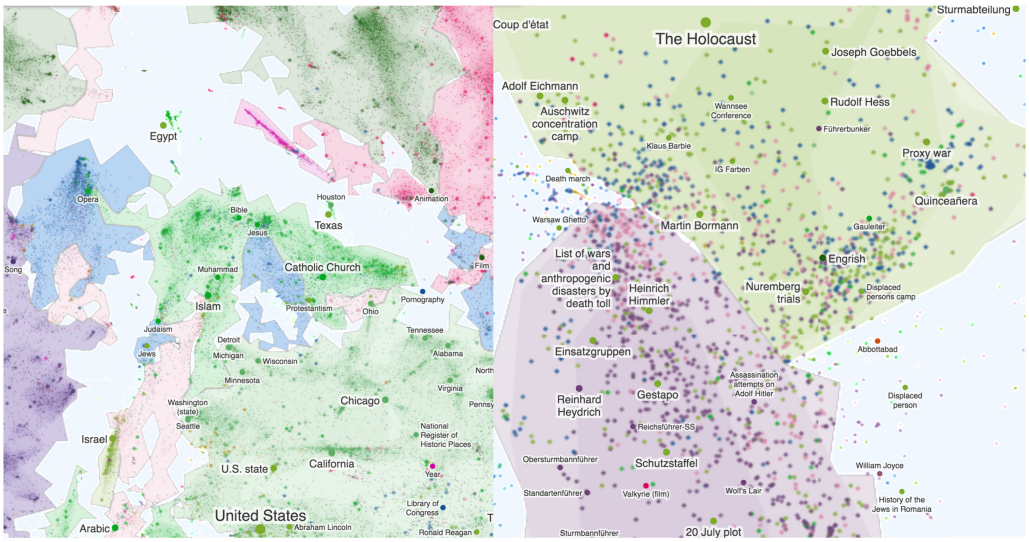
Clustering. We identify the main groups of semantic topics within a domain concept space by using the kmeans++ algorithm to cluster the high-dimensional vectors. Although we acknowledge that some domain areas may have an existing category structure that can be used (e.g., movies have genres), this is 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 Figure 4(a). However,heterogeneous topical areas still remain, as shown in Figure 4(b). 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 (bottom, 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 vs. 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 Hein and Maier [23]. 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 Hu et al. [24].



(a) Homogeneous thematic cluster areas **(b)** Heterogeneous thematic cluster areas

Fig. 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.

Topological contours. Cartograph produces topological contours for each country. We experimented with both density- 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, however, highlighted the areas that were most "representative" of each country.

3.5 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.

3.6 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 open source

framework[6] to render maps using WebGL,[7] a hardware accelerated rendering engine supported by 92% of browsers as of October 2016.[8]

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. Although this approach may not be practical for a site that serves block-level imagery of the entire earth, the one-time caching of 5 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.

Cartograph applies these technologies to visualize the massive amount of semantic information available within Wikipedia using recent scalable neural network embeddings algorithms. This combination of increased capacity of algorithms, improved technologies, and larger knowledge bases allows Cartograph to provide fluid interactive visualizations for 5 million data points, the largest dataset we had available. To our knowledge, this is an order of magnitude beyond prior spatialization systems.

4 CASE STUDIES

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

4.1 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 page views to warrant vectors in the navigation dataset. 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, whereas 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 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 toward 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 6 and 7 show a domain-specific thematic map of Wikipedia, with points colored by the gender focus of the article. Blue articles refer primarily to men, red articles refer primarily to women, 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 6 shows a striking focus on men (blue) throughout Wikipedia. However, some areas of red emerge. Figure 7 focuses on one such area, related to sexuality and feminism. Other areas with a strong female focus include modeling and

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

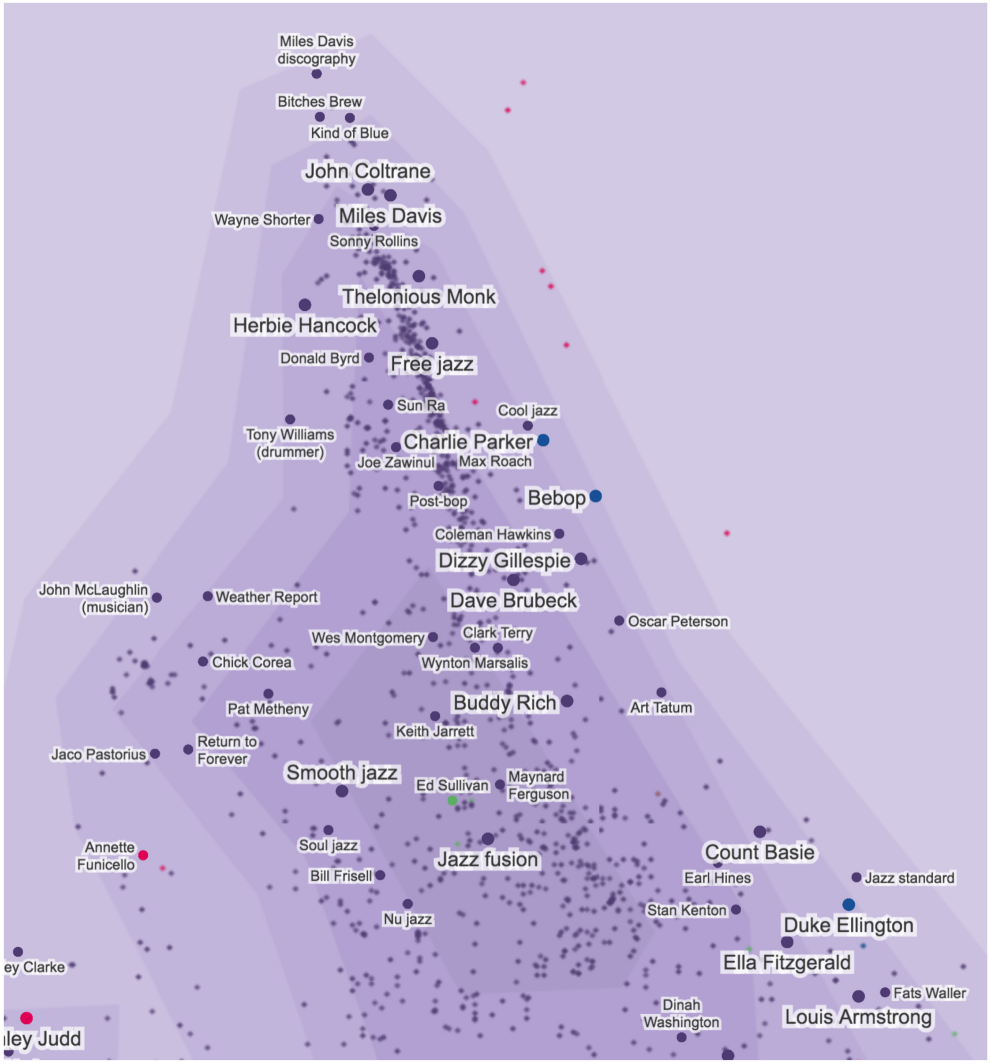


Fig. 5. A zoomed-in version of the Wikipedia map focused on jazz music.

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.

4.2 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 8 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

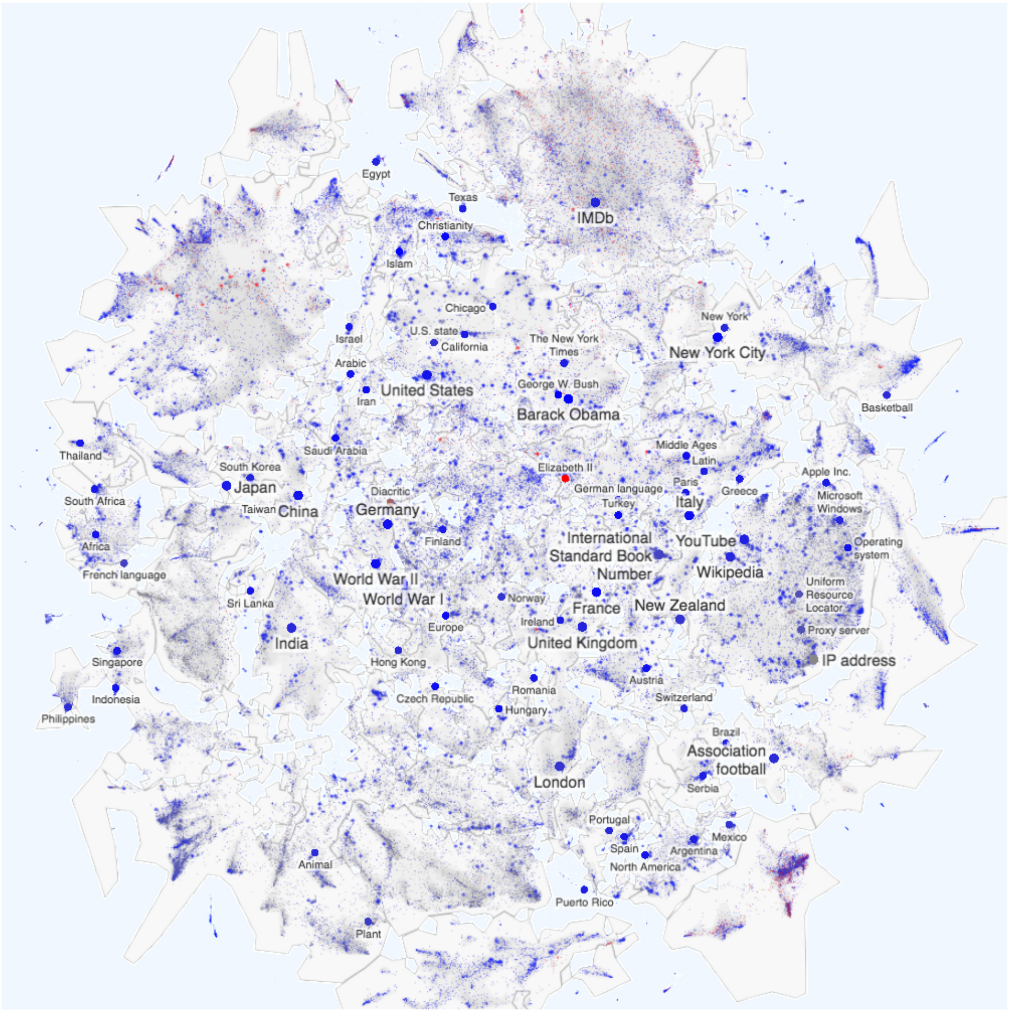


Fig. 6. The Wikipedia map of gender focus. Blue and red dots correspond to articles that focus on men and women, respectively.

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 9 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 Lam et al. [36] 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 10, features many movies generally referred to as "chick flicks," such as "Sleepless in Seattle" and "Pretty in Pink.'

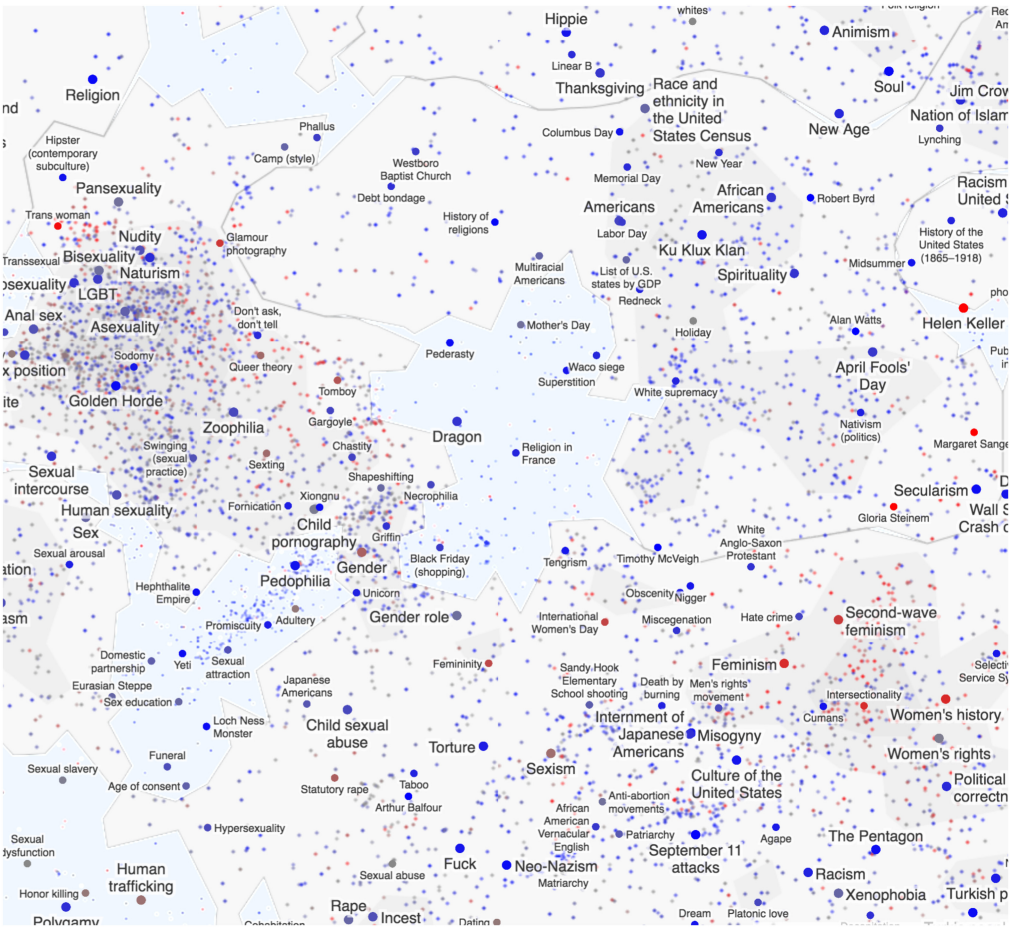


Fig. 7. Area of the Wikipedia map of gender focus related to feminism and sexuality. Blue and red dots correspond to articles that focus on men and women, respectively.

4.3 Map of Businesses

Figure 2 from Section 1 showed Cartograph's map of businesses with a thematic layer showing sustainability. Although we described the narrative of the map in Section 1, we describe details of its creation here. We identified all articles whose Wikidata type was a descendant of "business enterprise" or one of the synonyms listed in its Wikidata entry. Of the articles in this set, 19,522 had navigation vectors and were included in our map. We used corporate sustainability ratings from CSRHub,[9] which provides overall sustainability and social responsibility scores for more than 6,000 companies derived from 480 different datasets.

5 STUDY COMPARING VECTOR OPTIONS

When designing Cartograph, the choice of initial vectors (content based or navigation based) stood out as a key design decision. We desired an evaluation procedure that could empirically compare the base maps for both methods of creating vectors. Historically, similar scatterplot-like

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

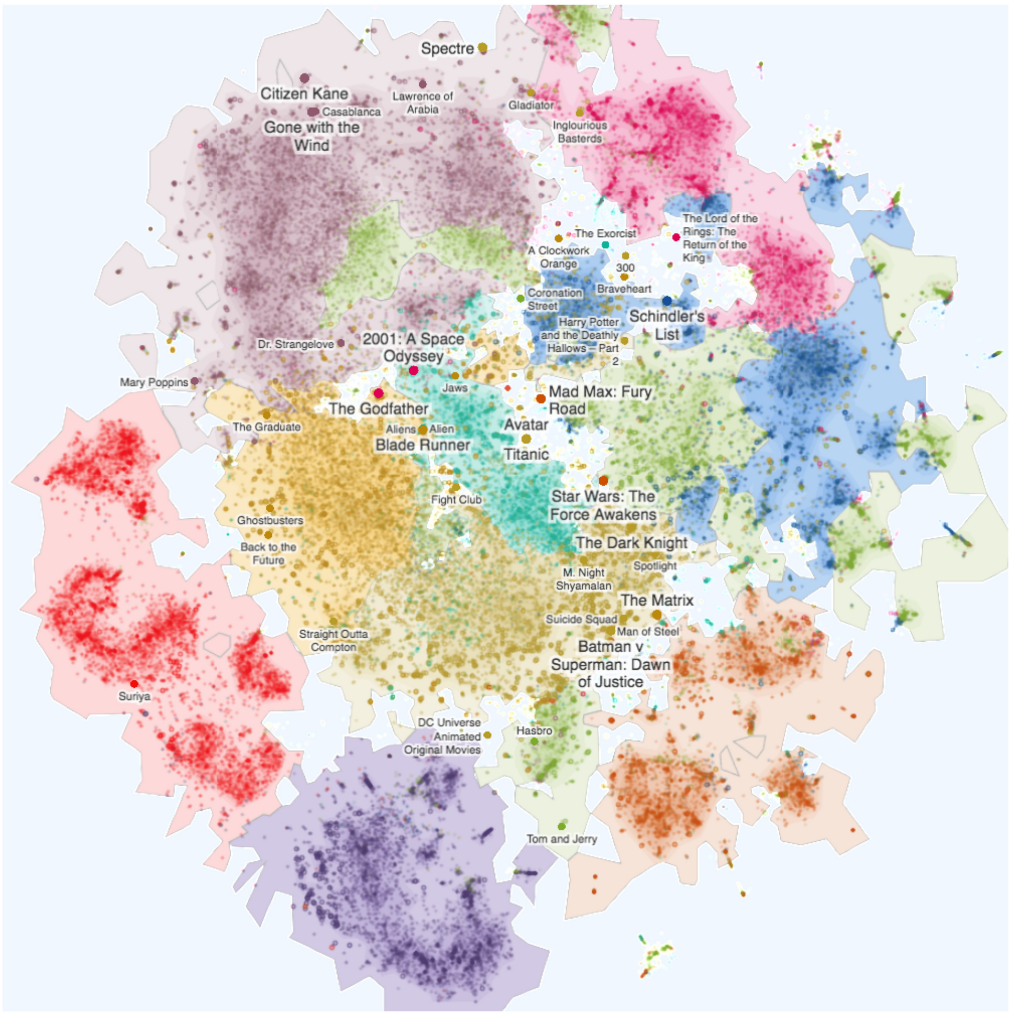


Fig. 8. The Cartograph map of films colored by topical clusters.

visualization systems are evaluated using synthetic metrics (e.g., class separation or correlation) or through the publication of exemplar visualization (e.g., a single image visualizing the MNIST digit handwriting database). However, previous research by Sedlmair et al. [49] has found that synthetic metrics do not typically reflect human perception.

5.1 Vector Study Protocol

Rather than rely on intuition from exemplar images or synthetic metrics, we devised a human-centered evaluation procedure to measure the effectiveness of Cartograph maps. Our exploration of maps generated using each type of vector (see Section 3.2) suggested that in some areas, Wikipedia's content (and therefore the content-based vectors) did not easily encode the human taste space. For example, two movies with identical directors and release years need not be closely related. Therefore, our initial evaluation uses the Map of Films described in the previous section.

The main task of our evaluation procedure is shown in Figure 11. Participants were asked to guess the identify of a hidden "mystery film" shown as a rainbow-colored dot on the map. We

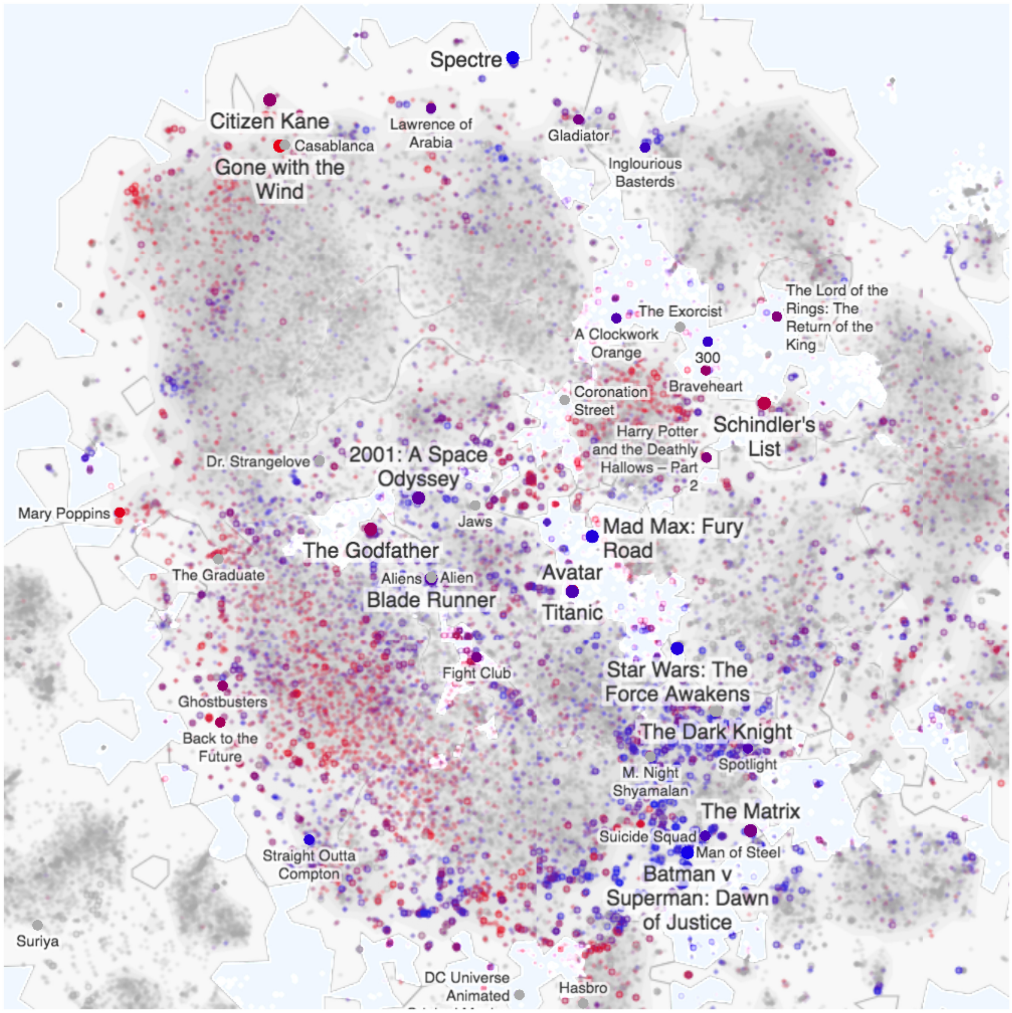


Fig. 9. Map of movies colored by gender interest. Movies that exhibit more interest from men are colored blue, whereas those interesting to women are colored red.

evaluated the two sets of maps by measuring the percentage of mystery film guesses that were correct.

For the user tasks, we created maps centered and zoomed in around a fixed set of 40 films. Our selection process for the 40 films follows:

(1) We noticed that concepts that were part of a a multi-part series led to trivial tasks. For example, if a map prominently featured the movie "Die Hard" near the rainbow dot, "Die Hard II" would be an obvious answer for the mystery city. Therefore, we removed sequels and prequels from the candidate pool.
(2) We wanted to increase the likelihood that respondents would know films. Therefore, we considered candidates ordered by their "predominance" as measured by the PageRank and page view metrics described earlier.

Fig. 10. Area of gender map focused on movies of interest to women.

(3) We wanted the selected 40 concepts to have sufficient distance separation in maps to decrease the likelihood that a "good" map had two plausible options for a mystery city. Thus, we iteratively chose the 40 films (ordered by prominence) with sufficient distance separation from other selected films.

Our experiment was designed as a web-based, crowdsourced study using Amazon's Mechanical Turk[10] (MTurk). This methodology has been shown to substantially reduce cost and time to results compared to traditional laboratory studies [29] while reaching a wider demographic of participants [25, 32]. Furthermore, MTurk studies replicating classical lab experiments have found them reliable for evaluating graphical perception in visualizations [22], analyzing performance of user interfaces [30], and experiments related to judgment and decision making [46]. However, like any evaluation technique, crowdsourced studies are not without their limitations. In exchange

[10] http://www.mturk.com/.

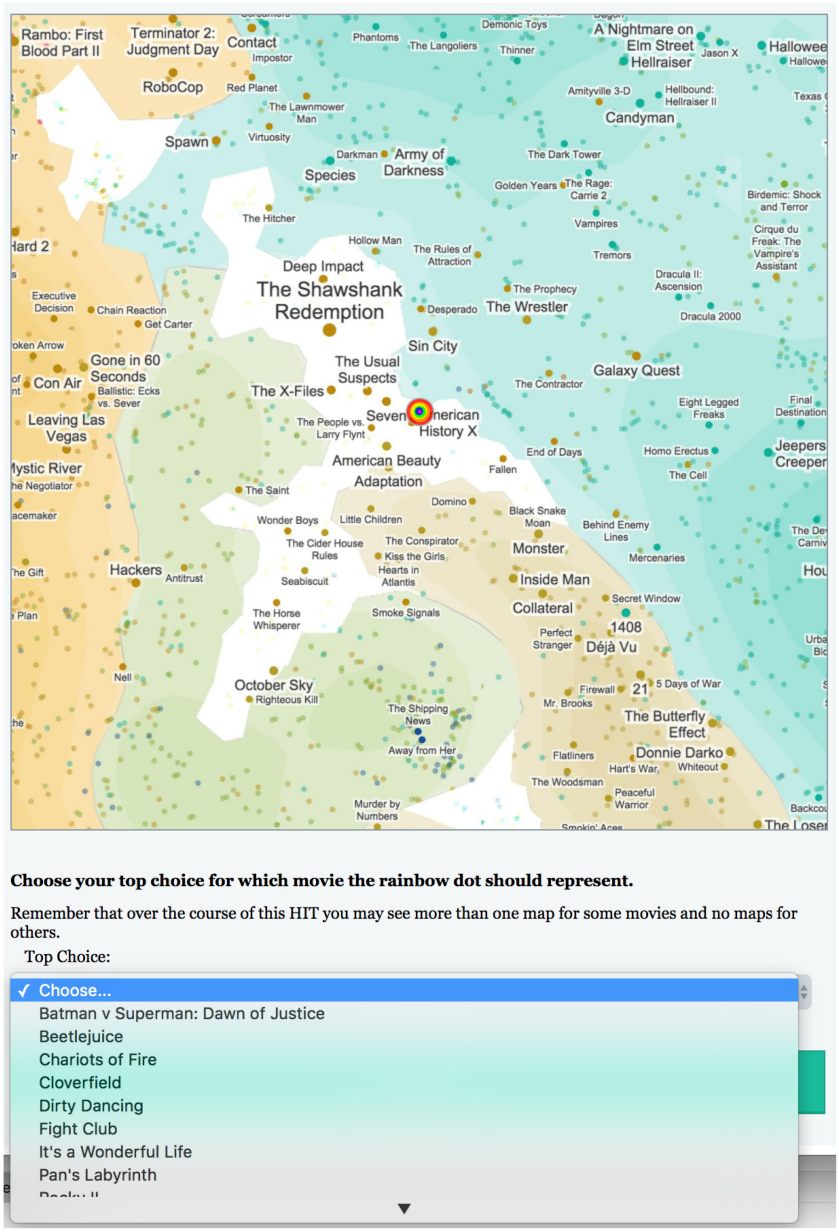


Fig. 11. Main screen of Mechanical Turk user study. Users are asked to guess the hidden mystery city located at the rainbow dot from a set of 12 choices.

for reaching a wider participant pool, experimenters give up control of many experimental conditions, including display type and environment, completion rates, and consistent quality of responses. These limitations can be mitigated by verification questions and qualification tasks to produce high-quality results [22]. As described in the following, our study makes use of both of these techniques.

We hired 331 crowdworkers on Amazon's Mechanical Turk to complete the study. As a qualification, all workers were required to have completed at least 5,000 previous MTurk tasks with a 98% approval rate and seen at least 12 of the 40 candidate concept films. The majority of respondents were male (59%) with ages that ranged from 20 to 71 years (median 35 years). In addition, 49% of respondents held a bachelor's degree or higher, and 80% indicated that they watched films on a weekly or daily basis. Median completion time for the survey was 9 minutes 27 seconds. Workers received $2.50 USD for the survey, which represented an hourly pay rate of $17.75 USD. Workers indicated they had seen a median of 17 of the 40 movies.

Respondents first consented to the study, provided basic demographic information, and indicated which of the 40 movies they had seen. Next, we sampled 10 seen films for each worker for map regions; 5 of these were used as mystery dots on maps, and all 10 were available as options in the drop-down as guesses for the mystery film. We then showed 12 randomly shuffled images of map regions: five content-based maps centered around each sampled film, five navigation-based maps centered around each film, and two "validation" maps (as described in the following). For each map, the center film "city" was replaced with a rainbow-colored dot, and the user was asked to guess the mystery film from 12 choices (the 10 sampled films and the 2 verification films).

To ensure that the users completed the survey carefully and in good faith, we also introduced two verification maps. These maps made use of obvious prequel/sequel pairs: "Rocky" vs. "Rocky II" and "The Lord of the Rings: The Fellowship of the Ring" vs. "The Lord of the Rings: Return of the King." One of each pair of concepts was featured prominently nearby the rainbow dot, whereas the second was available as a choice for the mystery city. As such, 296 of the 331 participants correctly answered both verification questions. We only include the 296 respondents in the following results and do not include the two verification questions in the results.

5.2 Vector Study Results

Respondents exhibited much higher accuracy for navigation maps (61.2%) than content-based maps (36.8%) ($p < 0.0001, N = 1480, \chi^2 = 176$). Although both maps far exceeded the random guessing rate of 8.3%, respondents' higher accuracies in the navigation-based maps indicates that they better correspond to users' understandings of conceptual regions. On a film-by-film basis, the navigation embeddings outperformed content embeddings on 25 of the 40 films. Surprisingly, on most films, the performance of the two embeddings was similar. However, for a small subset of films, the navigation vectors dramatically outperformed the content vectors. For example, on the mystery map for "The Night of the Living Dead," 5 of 51 content responses were correct compared to 32 of 51 navigation responses. On the mystery map for "Ferris Bueller's Day Off," 12 of 70 content responses were correct compared to 45 of 70 navigation responses. On the mystery map for "Dirty Dancing," 3 of 68 content responses were correct compared to 56 of 68 navigation responses.

Although the exact reasons behind the differences in performance on a movie-by-movie basis require a larger randomized experiment, by definition the variability reflects differences between the semantic structure observed in Wikipedia articles themselves and human navigation patterns around those articles. To better understand these differences, consider Table 2, which shows the 10 closest semantic neighbors of "Dirty Dancing" derived using three different models. The first model uses movie ratings from MovieLens.[11] Movie ratings form a more explicit signal of user interest, and we therefore introduce it as an informal "gold standard." As a note, we cannot generally use the MovieLens dataset for Cartograph because it only serves the movies domain. The next two methods display neighbors using content and navigation embeddings.

[11]Derived from the MovieLens 10M dataset https://grouplens.org/datasets/movielens/10m/ using item-based rating similarity.

Table 2. Movies Most Similar to "Dirty Dancing" Using Three Approaches

	MovieLens Ratings	Content Vectors	Navigation Vectors
1.	Pretty Woman	Dirty Rotten Scoundrels	Dirty Dancing: Havana Nights
2.	Flashdance	Dirty Harry	One Last Dance
3.	Ghost	Flashdance	Ghost
4	The Bodyguard	Dirty Pretty Things	Fame!
5.	Grease	Footloose	Flashdance
6.	Sister Act	Fame!	Grease
7.	Top Gun	The Dirty Dozen	Footlose
8.	Free Willy	Twins	The Breakfast Club
9.	Runaway Bride	Ironweed	Ferris Bueller's Day Off
10.	Grease 2	Arachnophobia	Sixteen Candles

The MovieLens neighbors reflect a combination of classic late 1980s and early 1990s romance movies, several of which have a strong dance theme ("Flashdance," "Grease"). The navigation neighbors show similar patterns, with a stronger thematic dancing influence ("Fame," "One Last Dance," "Dirty Dancing: Havana Nights"). The content vectors appear markedly different. "Dirty Harry," "Dirty Rotten Scoundrels," and "Dirty Pretty Things" all include the word "dirty," reflecting superficial textual patterns. The focus on these simple textual elements is not surprising. The Wikipedia article about "Dirty Dancing" is divided into main sections of "Plot," "Cast," "Soundtrack," "Production," "Reception," "Music," "Legacy," "Alternate versions," and "Remake." These sections focus almost exclusively on objective characteristics of the movie. They do not focus on thematic elements that shape the other two lists.

Overall, navigation vectors outperform content vectors in most cases and show marked advantages in a few cases. However, many questions remain. In particular, does the relative effectiveness of content and navigation vectors change across domain areas? Do the results hold across maps of Biology, Computer Science, Books, and Sports? Can the the observed variations in vector performance be modeled?

This method represents a human-centered approach to evaluation for the fields of algorithmic embedding, network visualization, and spatialization systems. Research in these areas almost universally relies on metrics that have been shown to deviate substantially from human perception [49]. Our approach overcomes this flaw by grounding measurements in human perception tasks. Moreover, our approach is relatively lightweight and engaging. Respondents required less than 30 seconds per map, and although respondents were not required to submit comments, 32 free-text comments used either the word "fun" or "interesting" to describe the task.

6 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 6 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 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.

Table 3. 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

6.1 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 and 53 years (median 40 years). All of the participants edit Wikipedia articles at least yearly, with three of the six editing monthly.

6.2 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 3.

6.3 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

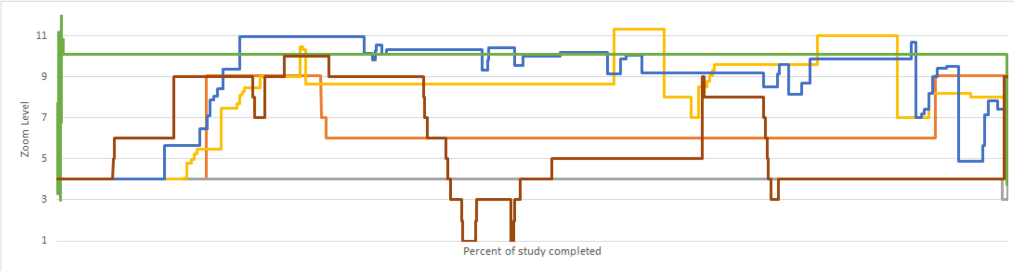


Fig. 12. Zoom levels for each participant over the course of the study. Zoom level 4 is the initial level, shown in Figure 1. Higher zoom levels contain a greater level of detail.

Table 4. 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

high female gender focus are about actual women who existed as opposed to topic and idea articles that 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."

Although these general patterns were identified, the visualization did prompt questions about the similarity between specific articles that were closely spaced on the map (e.g., "why does Madonna segue into ancient Persia?"). This may suggest that the visualization spurs new insights about relationships in the gender data that may not have been visible until augmented with relationships present in the high-dimensional semantic space. This finding is further supported by the responses in Table 3 that indicate that participants learned new information about the data.

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 Figure 12 generated from the logging data as participants completed the task. The participants represented by the green and blue lines used the first strategy, whereas the participants represented by the brown and yellow lines used the second.

6.4 Feedback on Specific Features

Table 4 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 Participant 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.

Although we are enthusiastic about the potential that Cartograph has for visualizing data that does not originally contain any spatial or SR components, we must be mindful that this type of visualization can confuse users. One participant mentioned, "I find the display confusing as it does not relate to geography. No obvious links to women."

Cartograph can potentially visualize data that does not originally contain any spatial or SR components, perhaps through additional data layers. One participant acknowledged how an article's gender focus could serve as one layer, whereas other demographic statistics could be represented at the same time on other layers. These data overlays could include additional visual glyphs or heat maps to visualize more data that may help with analysis.

7 CONCLUSION

This article introduces Cartograph, a system that unlocks thematic cartography for diverse data. Cartograph supports nearly universal spatialization, transforming a dataset with no existing semantic information (e.g., 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.

The human-centered evaluation we used to compare Cartograph content vectors and navigation vectors could provide human-centered evaluations from the field of algorithmic embedding, network visualization, and spatialization systems. Respondents indicated the approach was "interesting" and "fun," and reasonably fast. Thus, our approach offers a scalable method to collect measurements about the quality of many scatterplot-based visualizations.

Although 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- 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 preceding question 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 [41] and surfacing personalized recommendations [18]. 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. Although Cartograph's source code is publicly available,[12] 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 article into their own work.

ACKNOWLEDGMENTS

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. We also thank Wikipedians and Mechanical Turk Crowdworkers who participated in the user study for their feedback and Heather Ford and Aaron Halfaker for helping to identify the Wikipedia participants.

REFERENCES

[1] 506 Sports. 2016. NFL TV Schedule and Maps: Week 6, 2016. Retrieved February 23, 2019 from http://506sports.com/nfl.php?yr=2016&wk=6.

[2] FiveThirtyEight. 2016. Who Will Win the Presidency? Retrieved February 23, 2019 from http://projects.fivethirtyeight.com/2016-election-forecast/.

[3] 2016. Home Page. Retrieved February 23, 2019 from http://introjs.com/.

[4] NASA. 2016. Earth Observatory. Retrieved February 23, 2019 from http://earthobservatory.nasa.gov/Features/WorldOfChange/decadaltemp.php.

[5] New York Times. 2016. 50 Years of Electoral College Maps: How the U.S. Turned Red and Blue. Retrieved February 23, 2019 from http://nyti.ms/2bwoJsf.

[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, NY, 12–22.

[7] Georgia Albuquerque, Martin Eisemann, Jorn Schneidewind, Holger Theisel, Marcus Magnork, and Daniel Keim. 2009. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST'09)*. 59–66.

[8] Francesco Bellomi and Roberto Bonato. 2005. Network analysis for Wikipedia. In *Proceedings of the 1st International Wikimedia Conference (Wikimania'05)*.

[9] E. Bertini, A. Tatu, and D. Keim. 2011. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (Dec. 2011), 2203–2212.

[10] 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)*. 159–ff.

[11] Matthew Chalmers. 1993. Using a landscape metaphor to represent a corpus of documents. In *Spatial Information Theory: A Theoretical Basis for GIS*. Lecture Notes in Computer Science, Vol. 715. Springer, 377–390.

[12] 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.

[13] H. Couclelis. 1998. Worlds of information: The geographic metaphor in the visualization of complex information. *Cartography and Geographic Information Systems* 25 (1998), 209–220.

[14] Donna Cox. 2006. Metaphoric mappings: The art of visualization. In *Aesthetic Computing*, P. Fishwick (Ed.). MIT Press, Cambridge, MA, 89–114.

[15] Andrew M. Dai, Christopher Olah, and Quoc V. Le. 2015. Document embedding with paragraph vectors. arXiv:1507.07998,

[12] http://cartograph.info.

[16] Ronak Etemadpour, Lars Linsen, Christopher Crick, and Angus Forbes. 2015. A user-centric taxonomy for multidimensional data projection tasks. In *Proceedings of the 6th International Conference on Information Visualization Theory and Applications*. 51–62.

[17] 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.

[18] Emden Gansner, Yifan Hu, Stephen Kobourov, and Chris Volinsky. 2009. Putting recommendations on the map: Visualizing clusters and relations. In *Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys'09)*. ACM, New York, NY, 345–348.

[19] Emden R. Gansner, Yifan Hu, and Stephen G. Kobourov. 2010. GMap: Visualizing graphs and clusters as maps. In *Proceedings of the IEEE Pacific Visualization Symposium (PacificVis'10)*. 201–208.

[20] Martin Gronemann and Michael Jünger. 2013. Drawing clustered graphs as topographic maps. In *Proceedings of the 20th International Conference on Graph Drawing (GD'12)*. 426–438.

[21] 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, NY, 415–424.

[22] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. 203–212.

[23] Matthias Hein and Markus Maier. 2006. Manifold denoising. In *Advances in Neural Information Processing Systems*. 561–568.

[24] Yifan Hu, Stephen Kobourov, and Emden R. Gansner. 2010. Visualizing graphs and clusters as maps. *IEEE Computer Graphics and Applications* 30 (2010), 54–66.

[25] Panagiotis G. Ipeirotis. 2010. *Demographics of Mechanical Turk*. Technical Report. New York University. http://hdl.handle.net/2451/29585.

[26] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. 2013. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (Dec. 2013), 2818–2827.

[27] Sara Johansson and Jimmy Johansson. 2009. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (Nov. 2009), 993–1000.

[28] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. arXiv:1602.02410.

[29] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08)*. ACM, NY, 453–456.

[30] Steven Komarov, Katharina Reinecke, and Krzysztof Z. Gajos. 2013. Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*. ACM, New York, NY, 207–216.

[31] Robert R. Korfhage. 1991. To see, or not to see—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, 134–141.

[32] Robert Kosara and Caroline Ziemkiewicz. 2010. Do Mechanical Turks dream of square pie charts? In *Proceedings of the 3rd BELIV'10 Workshop: Beyond Time and Errors: Novel Evaluation Methods for Information Visualization (BELIV'10)*. ACM, NY, 63–70.

[33] Sherry Koshman. 2006. Visualization-based information retrieval on the Web. *Library and Information Science Research* 28, 2 (2006), 192–207.

[34] Werner Kuhn and Brad Blumenthal. 1996. Spatialization: Spatial metaphors for user interfaces. In *the Conference Companion on Human Factors in Computing Systems (CHI'96)*. ACM, New York, NY, 346–347.

[35] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. 2012. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics* 18, 9 (2012), 1520–1536.

[36] 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, New York, NY, 1–10.

[37] Joshua M. Lewis, Laurens van der Maaten, and Virginia R. de Sa. 2012. A behavioral investigation of dimensionality reduction. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*. 671–676.

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

[39] 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.

[40] Rafael Messias Martins, Rosane Minghim, and Alexandru C. Telea. 2015. Explaining neighborhood preservation for multidimensional projections. In *Computer Graphics and Visual Computing (CGVC)*, R. Borgo and C. Turkay (Eds.). Eurographics Association.

[41] 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.
[42] 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, 3111–3119.
[43] 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 *Spatial Information Theory*. Lecture Notes in Computer Science, Vol. 2825. Springer, 24–28.
[44] 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*, E. Di Giacomo and A. Lubiw (Eds.). Springer International Publishing, 3–15.
[45] Thanapon Noraset, Chandra Bhagavatula, and Doug Downey. 2014. Adding high-precision links to Wikipedia. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language (EMNLP'14)*. 651–656.
[46] Gabriele Paolacci, Jesse Chandler, and Panos Ipeirotis. 2010. Running experiments using Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (Aug. 2010), 411–419.
[47] Randall M. Rohrer and Edward Swing. 1997. Web-based information visualization. *IEEE Computer Graphics and Applications* 17, 4 (July 1997), 52–59.
[48] M. Sedlmair, T. Munzner, and M. Tory. 2013. Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (Dec. 2013), 2634–2643.
[49] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. 2012. A taxonomy of visual cluster separation factors. *Computer Graphics Forum* 31, 3 (June 2012), 1335–1344.
[50] 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, New York, NY, 27.
[51] Shilad Sen, Anja Beth Swoap, Qisheng Li, Brooke Boatman, Ilse Dippenaar, Rebecca Gold, Monica Ngo, et al. 2017. Cartograph: Unlocking spatial visualization through semantic enhancement. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI'17)*. ACM, NY, 179–190.
[52] Mike Sips, Boris Neubert, John P. Lewis, and Pat Hanrahan. 2009. Selecting good views of high-dimensional data using class consistency. In *Proceedings of the 11th Eurographics/IEEE VGTC Conference on Visualization (EuroVis'09)*. 831–838.
[53] André Skupin. 2000. From metaphor to method: Cartographic perspectives on information visualization. In *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'00)*. IEEE, Los Alamitos, CA, 91.
[54] 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.
[55] Terry A. Slocum, Robert B. McMaster, Fritz C. Kessler, and Hugh H. Howard. 2009. *Thematic Cartography and Geovisualization* (3rd ed.). Prentice Hall, Upper Saddle River, NJ.
[56] Yuku Takahashi, Hiroaki Ohshima, Mitsuo Yamamoto, Hirotoshi Iwasaki, Satoshi Oyama, and Katsumi Tanaka. 2011. Evaluating significance of historical entities based on tempo-spatial impacts analysis using Wikipedia link structure. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*. ACM, New York, NY, 83–92.
[57] Andrada Tatu, Peter Bak, Enrico Bertini, Daniel Keim, and Joern Schneidewind. 2010. Visual quality metrics and human perception: An initial study on 2D projections of large multidimensional data. In *Proceedings of the International Conference on Advanced Visual Interfaces (AVI'10)*. ACM, New York, NY, 49–56.
[58] Walter R. Tobler. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46 (1970), 234–240.
[59] 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.
[60] Laurens Van Der Maaten. 2014. Accelerating t-SNE using tree-based algorithms.*Journal of Machine Learning Research* 15, 1 (2014), 3221–3245.
[61] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM* 57, 10 (2014), 78–85.
[62] Ryen W. White and Resa A. Roth. 2009. *Exploratory Search: Beyond the Query-Response Paradigm*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool.
[63] Ellery Wulczyn. 2016. Wikipedia Navigation Vectors. Retrieved February 23, 2019 from DOI: https://doi.org/10.6084/m9.figshare.3146878.v3.

Received June 2017; revised March 2018; accepted May 2018