SightLine: Building on the Web's Visualization Ecosystem

Jordan Sechler
Bucknell University
Lewisburg, PA USA
jls102@bucknell.edu

Lane Harrison
Worcester Polytechnic Institute
Worcester, MA USA
ltharrision@wpi.edu

Evan M. Peck
Bucknell University
Lewisburg, PA USA
evan.peck@bucknell.edu

CHI'17 Extended Abstracts, May 6–11, 2017, Denver, CO, USA.
ACM ISBN 978-1-4503-4656-6/17/05
http://dx.doi.org/10.1145/3027063.3053142

Abstract
The advancement of tools for web designers and developers have transformed the internet into a vibrant ecosystem for data visualization and analysis. However, even when there are hundreds of visualizations exploring the same topic, they are often viewed in isolation with no mechanism to navigate between them. We built SightLine - a web portal that passively collects and organizes visualizations - to explore the design space of the web's scattered visualization ecosystem. By preserving the context of each visualization visit, we enable *personal provenance* through a "visualization history", *discovery and exploration* through trending visualizations, and *targeted search* by querying SightLine's extensive metadata collected for each visualization.

Author Keywords
information visualization; search; web design; provenance

ACM Classification Keywords
H.5.3 [Information interfaces and presentation (e.g., HCI)]: Group and Organization Interfaces

Introduction
In 1970, Ted Nelson proposed the idea of 'ThinkerToys', envisioning an internet-like environment in which information about the 'complexities [people] seek to understand' could be deeply linked, enabling people to shift and study alter-

nate views about the same topic [11]. While the maturation of the internet has provided a wealth of these perspectives, the relationships and coupling of data has not lived up to Nelson's expectations. In this paper, we consider the following question: How can we leverage the web's *existing* ecosystem of rich interactive visualizations to help people discover, compare, and investigate the complexities of the 21st century?

While the research community has explored and evaluated data visualizations in isolation, the rise of the internet has enabled the potential for a "distributed, asynchronous style of collaboration" on data that was previously not possible [6]. For any topic on the internet, there are numerous datasets hosted in a variety of locations (Google Sheets, Github, etc.), each of which may be cleaned or 'wrangled' in a variety of ways. After wrangling, each of those datasets may be visually communicated using any number of representations (bar graph, line chart, scatterplot, etc.) and data visualization tools (d3js [4], Google Sheets [1], Tableau [2]). This pipeline exemplifies the internet's massively-paralleled approach to analysis and data visualization. When considered as a group, they have the potential to create nuanced perspectives, reflecting a diverse set of analyses and design decisions. Yet we often see them alone.

There are very few mechanisms that support navigation, exploration, or search *between* visualizations on the web. While browsers commonly support web history, there is no clear way to retrace interactions with data. While websites often allocate attention by emphasizing trending topics, there is no parallel recommendation system to expose valuable or timely datasets. Finally, while websites such as Google facilitate targeted search, searching for data requires advanced queries with carefully chosen keywords.

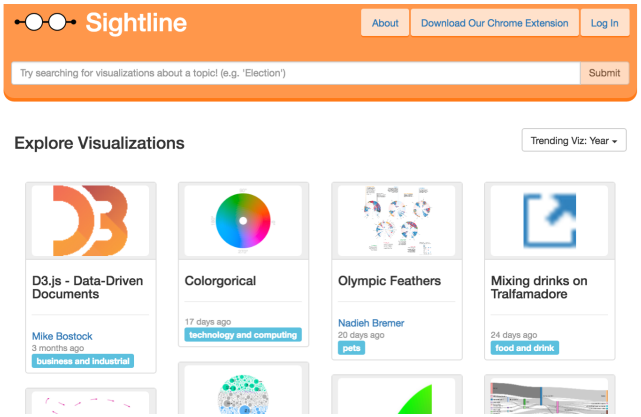

Figure 1: SightLine is an automatic data collection and discovery service that leverages the web's rich visualization ecosystem.

To prototype tasks in this space, we present **SightLine** – a web application that automatically collects, indexes, and organizes data visualizations that are visited by its users. **We demonstrate how data sensemaking tasks can apply not only within a visualization, but more broadly to the web's visualization ecosystem**. SightLine supports three such functions: *(1) Personal provenance:* SightLine maintains a personal web history of interactive visualizations visited by each person, *(2) Discovery and Exploration:* Using the aggregated histories of all users, SightLine suggests current trending interactive data visualizations. In addition, it allows exploration by author and topic (*e.g.* sports). And *(3) Targeted Search:* Using natural language processing, SightLine extracts a series of metadata (category, keywords, authors, titles, etc.) that is used to identify relevant data visualizations during more targeted queries.

Background

While advances in authoring tools have simplified the process of creating data visualizations [2, 16], notable recent advances have contributed to an increase in data visualizations available on the web. Bostock, Ogievetsky, and Heer developed the d3js library to create visualizations using web browsers' built-in Document-Object Model (DOM) and Scalable Vector Graphics (SVG) support [4]. Since its 2011 release, the d3js library has been used to create tens of thousands of data visualizations on the web, and news organizations such as the *Washington Post* and *The New York Times* make extensive use of it in data-driven stories.

This new web-based technologies has enabled research to leverage the fact that people are naturally engaging with data and data visualizations. Harper and Agrawala, for example, built a Chrome browser extension called *D3 Deconstructor* that extracts the underlying data from existing d3 visualizations, making it possible to re-analyze and re-visualize [5]. Recently Kim, Hullman, and Agrawala introduced a Chrome extension - *Atlas of Me* - that generates personalized spatial analogies for distances and areas naturally occurring in internet articles [7]. Advances in authoring tools and adaptation techniques have enabled a new visualization ecosystem, bringing unique research opportunities for supporting people's discovery and engagement with data.

Visualization Ecosystems

McKeon discusses the value of visualization's web ecosystem in the context of *Dashiki* - a website where users collaboratively built visualization dashboards from distributed data sources [10]. McKeon states that, "No visualization application on the web exists in a vacuum; instead, every visualization is a potential participant in a diverse information ecology". In this model, IBM's *Many Eyes*, an online platform where people uploaded datasets, created visualizations, and viewed designs built by other people, allowed users to collaborate, remix, and revise across a broad set of visualizations within the scope of the platform [17].

In contrast to the centralized nature of *Many Eyes*, other research has taken a decentralized approach to facilitating data analysis and visualization. Lee *et al.* created *Viziometrics*, an online database that searches through figures and data visualizations from scientific papers. Academic papers and figures are published using many different authoring tools and in many different venues. Viziometrics collects and processes figures from these venues to facilitate search, discovery, and analysis. Their work has unified the broader ecosystem of nearly 7 million figures in scientific literature [8]. Given the steady growth of data visualization on the web, **SightLine explores the emerging need for provenance and management in the distributed visualization ecosystem on the web.**

SightLine Design and Use

SightLine's goal is to situate familiar information-seeking design patterns that already exist on the web specifically in a visualization context - personal provenance, casual discovery and exploration, and targeted search. To accomplish these goals, SightLine uses three core components: (1) a Chrome browser extension that detects visited data visualizations, (2) a backend that collects and stores context and metadata for each visualization, and (3) a website that curates and displays the collected visualizations.

SightLine leverages the dominance of d3js as a programming language for interactive visualizations. The Chrome extension uses a simple detection algorithm in which it searches each website for the presence of d3 and an SVG. If both conditions are met, SightLine captures the largest

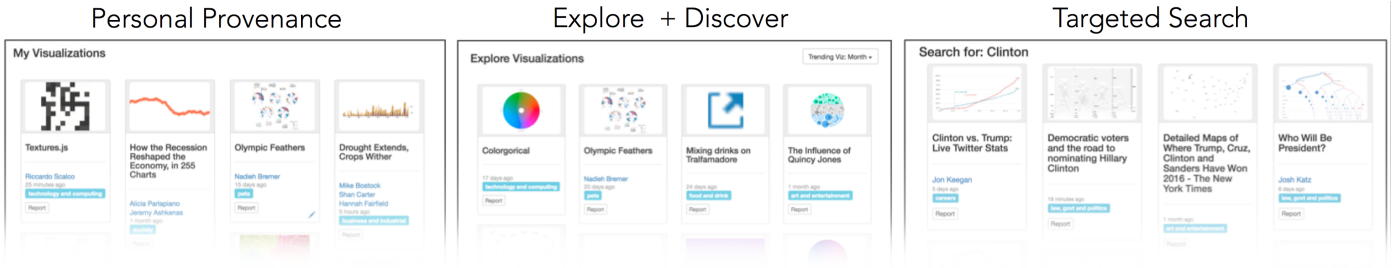

Figure 2: SightLine supports three user functions– *personal provenance (left):* Visualization history is presented in order of a user's most recent visit, *explore + discover (middle):* Visualizations from all users are sorted by those most frequently visited within the past month. *targeted search (right):* Search results for "Clinton" yields more than 15 recent visualizations about the current US presidential election.

SVG on the page and sends the url to the server. While imperfect, we find that this heuristic allows us to explore and prototype in the design space of a system that builds on the web's visualization ecosystem. We look to ongoing research, such as Savva *et al.*'s work on automatically detecting the presence of visualizations in images, for the possibility to achieve much greater coverage of the visualization ecosystem [15]. SightLine was publicly launched on July 26, 2016. The website and extension were shared on social media with no further incentive to participate. From July 26, 2016 through September 11, 2016, around 50 users of SightLine's Chrome extension passively submitted over **1300 unique visualizations** to its servers. We have received informal feedback suggesting that this group primarily consists of visualization practitioners.

To further unify and build relationships between the discovered visualizations, SightLine extracts and stores each visualization's context using *AlchemyLanguage* - an API on *IBM Watson Developer Cloud* that offers text analysis through natural language processing. For each vis, we store its *url*, *title*, *authors*, *keywords*, *keyword sentiment*, *keyword relevance*, *taxonomy categories*, *concepts*, *concept relevance*, *sentiment (positive or negative)*, and *emotions (anger, joy, etc.)*. To preserve visual information, we also store a thumbnail image of the largest SVG detected on the page. SightLine currently maintains this data for the 1300 aforementioned visualizations.

Visualization Tasks Across the Web

Each data visualization stored by SightLine is presented in a **viz card** on the website https://sightlinevis.com (see Figure 1). Cards consist of a thumbnail image, title, author names, top-level taxonomy of the visualization, and an indicator of when it was first collected. The organization of these cards is dependent on three views that correspond to dominant use-cases: *personal provenance*, *untargeted search (or exploration and discovery)*, and *targeted search* (searching for specific information or keywords).

Personal Provenance

To support personal provenance, SightLine acts as a visualization history for the user. Rather than store interaction history within a visualization or visitation history

SIGHTLINE TOP SEARCH RESULTS

Key (variety of search):

direct	author
shared link	taxonomy

1	brexit
2	Election
3	election
4	Clinton
5	science
6	sports
7	Trump
8	law,govt, and politics
9	gooey
10	bar
11	education
12	Olympics
13	gin
14	Map
15	unit
16	music
17	olympics
18	President
19	Krist Wongsuphasawat
20	Olympic

Figure 3: Top search queries on SightLine as of Sep. 2016. Uncolored terms are direct searches, red terms are searches shared on social media, blue terms are search pages generated author links, and green terms are search pages generated by taxonomy (category) links.

across visualizations on a single platform (for example, IBM's ManyEyes), we explore the idea of simple visitation history across the web's visualization ecosystem. Located under a "My Visualizations" tab, viz cards are listed in the order in which the user most recently viewed them. (Figure 2:left)

Aside from the user-side functionality, recording history lends insight into visit patterns of users who engage with data visualizations. We find that on average, 67% of the total visualizations visited were unique (or just 33% of all views were revisits). However, a small subset of users highlighted in Figure 4 revisit visualizations over and over again. In the future, analyzing behavioral patterns may highlight different user groups (such as practitioners and casual users, for example).

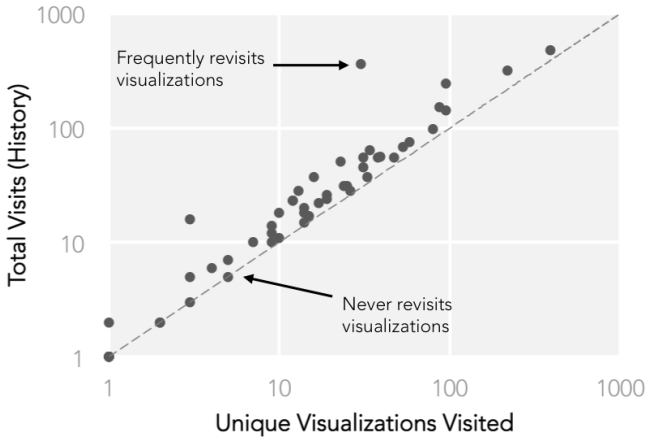

Figure 4: For each SightLine user, we show the relationship between the total number of times they visit a visualization and the number of unique visualizations they have viewed.

Exploration and Discovery

Although SightLine has a small group of core users that take advantage of its history features, the large majority of SightLine's activity comes from non-users discovering and exploring visualizations on its website. To facilitate these activities, SightLine crowdsources and displays all of the data visualizations visited by all of its users (Figure 2:center). On the "Explore Visualizations" tab (see Figure 5), viz cards are ordered by the frequency in which they have been visited by users of the Chrome extension. Users can discover emerging visualizations by filtering for time intervals of "Today" and "Week", or discover popular, highly-visited visualizations by filtering by "Month" or "Year". Although currently in a very simple state, this kind of prioritizing mirrors recommendation work that can highlight, for example, relevant news articles [9].

Discovery is further facilitated by each viz card's author and taxonomy meta-data. Clicking on an author name, for example, will show other visualizations created by that author. Clicking on taxonomy information (*e.g.* law, govt, and politics) will display other visualizations that fall under the same categorization. In Figure 3, many of these links are used to search for additional visualizations by topic (highlighted in green) or by author (highlighted in blue). We believe that enabling users to pivot between related visualizations moves us closer to Nelson's vision of deeply linked perspectives of 'complexities people seek to understand' [11].

Targeted Search

Finally, SightLine provides basic mechanisms for directed queries through search. Search terms access all the collected metadata stored by SightLine to list, for example, many data visualizations about the upcoming presidential election from a single term. (in the case of Figure 2:right, 'Clinton') Through September 2016, we have recorded

over 500 unique search queries, a majority of which can be attributed to targeted search. As shown Figure 3, these terms tend to focus predominantly around current events ('Olympics' and 'election'), as well as visualization-related queries ('treemap' and 'sankey'). In addition, because queries create unique URLs, we have seen search results turn into opportunities for discovery as these search pages are shared on social media. The frequencies of SightLine's queries exhibit the typical long tail of low-traffic searches.

Figure 5: Frequency of the top 50 search results on SightLine exhibits the long tail often seen in search

Limitations

SightLine is still a prototype. It does not detect static visualizations on the web or ones generated by tools such as Tableau [2]. In addition, its trending views reflect the users that have installed the Chrome Extension - a group that currently consists of primarily visualization practitioners and enthusiasts. There is still significant room to explore to the capture, linking, and searching of visualizations on the web.

Discussion

The goal of SightLine is to consider the web's visualization ecosystem as a valuable resource that can be used to help reasoning and understanding through data. Due to the web's active, decentralized characteristics, people organically contribute thousands of data visualizations from many data sources across the internet, seeking to explain emerging complexities of the 21st century. Rather than build new, centralized visualization platforms, we ask: *how can research empower people to take advantage of this rich ecosystem?*

One useful model for reasoning about SightLine's support for sensemaking is Pirolli and Card's *Information Foraging Theory* [12]. In the current visualization ecosystem on the web, it is difficult for people to forage for relevant visualizations given a particular starting topic. In many cases, SightLine can reduce the expected cost of accessing such information - aligning with the Information Foraging model.

In their later work on the *sensemaking loop* [13], Pirolli and Card define bottom-up and top-down processes for information collection and sensemaking, both of which are supported through SightLine. For example, the bottom-up process of *Search and Filter* is explicitly supported: users gather information passively via browsing on the web, and actively via targeted search within SightLine. Importantly, SightLine also supports the top-down process of *search for support*, where users may easily revisit visualizations to recover previously gathered information. SightLine also supports *search for relations*, in which SightLine's extensive metadata is accessed in any user queries. This allows people to iteratively refine search terms (*e.g.* transitioning from 'election' to 'Clinton') to find additional perspectives on their topic of interest.

Overall, it is important to consider that everyday users may have widely different motivations for engaging with a given data visualization. Recent research, for example, indicates that some users engage with data visualizations for enjoyment and immersion/reflection moreso than for specific analysis goals or hypothesis support [3, 14]. SightLine supports some of these use cases, and as more is learned about how and why people engage with data visualizations, more tools can be developed that further leverage the ecosystem of visualizations on the web.

References

[1] 2016 (accessed Sept 1, 2016). *Google Spreadsheets*. http://spreadsheets.google.com

[2] 2016 (accessed Sept 1, 2016). *Tableau Software*. http://tableau.com

[3] Scott Bateman, Regan L. Mandryk, Carl Gutwin, Aaron Genest, David McDine, and Christopher Brooks. 2010. Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. 2573–2582.

[4] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2301–2309.

[5] Jonathan Harper and Maneesh Agrawala. 2014. Deconstructing and restyling D3 visualizations. *Proceedings of the ACM symposium on User Interface Software and Technology (UIST '14)* (2014), 253–262.

[6] Jeffrey Heer and Maneesh Agrawala. 2007. Design considerations for collaborative visual analytics. *Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST '07)* (2007), 171–178.

[7] Yea-Seul Kim, Jessica Hullman, and Maneesh Agrawala. 2016. Generating Personalized Spatial Analogies for Distances and Areas. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '16)*. 38–48.

[8] Po-shen Lee, Jevin D. West, and Bill Howe. 2016. VizioMetrix: A Platform for Analyzing the Visual Information in Big Scholarly Data. In *Proceedings of the International Conference Companion on World Wide Web (WWW '16)*. 413–418.

[9] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized news recommendation based on click behavior. In *Proceedings of Intelligent User Interfaces*. ACM, 31–40.

[10] Matt McKeon. 2009. Harnessing the web information ecosystem with wiki-based visualization dashboards. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1081–1088.

[11] Theodor H. Nelson. 1974. *Computer Lib/Dream Machines*. Microsoft Press, Redmond, WA.

[12] Peter Pirolli and Stuart Card. 1995. Information foraging in information access environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95)*. 51–58.

[13] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, Vol. 5. 2–4.

[14] Bahador Saket, Carlos Scheidegger, and Stephen G. Kobourov. 2016. Comparing Node-Link and Node-Link-Group Visualizations From An Enjoyment Perspective. *Computer Graphics Forum* 35 (2016), 41–50.

[15] Manolis Savva, Nicholas Kong, Arti Chhajta, Li Fei-Fei, Maneesh Agrawala, and Jeffrey Heer. 2011. Revision: Automated classification, analysis and redesign of chart images. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 393–402.

[16] Will J. Schroeder, Bill Lorensen, and Ken Martin. 2004. *The Visualization Toolkit: an object-oriented approach to 3D graphics; 3rd ed.* Kitware.

[17] Fernanda B. Viegas, Martin Wattenberg, Frank Van Ham, Jesse Kriss, and Matt McKeon. 2007. Many Eyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (2007), 1121–1128.