


# StoryFacets: A design study on storytelling with visualizations for collaborative data analysis

Information Visualization  
2022, Vol. 21(1) 3–16  
© The Author(s) 2021  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/14738716211032653  
journals.sagepub.com/home/ivi  


Deokgun Park<sup>1</sup>, Mohamed Suhail<sup>2</sup>, Minsheng Zheng<sup>3</sup>,  
Cody Dunne<sup>4</sup>, Eric Ragan<sup>5</sup> and Niklas Elmqvist<sup>6</sup>

## Abstract

Tracking the sensemaking process is a well-established practice in many data analysis tools, and many visualization tools facilitate overview and recall during and after exploration. However, the resulting communication materials such as presentations or infographics often omit provenance information for the sake of simplicity. This unfortunately limits later viewers from engaging in further collaborative sensemaking or discussion about the analysis. We present a design study where we introduced visual provenance and analytics to urban transportation planning. Maintaining the provenance of all analyses was critical to support collaborative sensemaking among the many and diverse stakeholders. Our system, STORYFACETS, exposes several different views of the same analysis session, each view designed for a specific audience: (1) the *trail view* provides a data flow canvas that supports in-depth exploration + provenance (expert analysts); (2) the *dashboard view* organizes visualizations and other content into a space-filling layout to support high-level analysis (managers); and (3) the *slideshow view* supports linear storytelling via interactive step-by-step presentations (laypersons). Views are linked so that when one is changed, provenance is maintained. Visual provenance is available on demand to support iterative sensemaking for any team member.

## Keywords

Exploratory analysis, visualization, provenance, communication, storytelling, presentation, narrative visualization

## Introduction

Recording and visualizing the history of an analysis process – its *analytic provenance* – can support sensemaking in many different ways.<sup>1,2</sup> For instance, Ragan et al.<sup>3</sup> discusses several purposes for collecting and displaying provenance information for not only sensemaking but also collaboration and presentation purposes. Provenance can be presented in many different formats, and visual representations in particular can help to establish a **common ground**<sup>4</sup> for collaborative communication among participants in the same analysis, as well as support sharing and communication of data, findings, and processes. For those not directly involved in the analysis, provenance can aid in

understanding analyses and findings, and is closely related to narrative visualization and storytelling.<sup>5,6</sup>

However, we argue that collaborative communication and presentation are not necessarily dichotomous

<sup>1</sup>University of Texas at Arlington, Arlington, TX, USA

<sup>2</sup>Texas A&M University, College Station, TX, USA

<sup>3</sup>OCAD University, TO, CA

<sup>4</sup>Northeastern University, Boston, MA, USA

<sup>5</sup>University of Florida, Gainesville, FL, USA

<sup>6</sup>University of Maryland, College Park, MD, USA

## Corresponding author:

Deokgun Park Computer Science and Engineering University of Texas at Arlington 701 S. Nedderman Drive Arlington, TX 76019.  
Email: deokgun.park@uta.edu

purposes, and that through clever visual application design we can support visualization consumers in becoming participating analysts in their own right. In practice, many complex analyses must eventually be communicated to stakeholders anyway – such as colleagues, managers, customers, or the general public – in order to be useful and actionable.<sup>7,8</sup> But many current analysis presentations suffer from the so-called “PowerPoint gap”: expert analysts often end up copying and pasting screenshots from tools into a Microsoft PowerPoint slideshow to present to a stakeholder audience.<sup>9,10</sup> This process is time-consuming, error-prone, and prevents easy updates to the presentation when mistakes are found or new data is added. While some commercial visual analytic tools with built-in presentation features such as Story Points in Tableau<sup>9</sup> mitigates these limitations, still viewers cannot easily engage with the analysis or interrogate the data except through the curated materials and, if they are in attendance, the analyst’s memory.

In this paper, we demonstrate how visual provenance can be used to mitigate the barrier between the analysis phase and the presentation phase, allowing even the audience to participate in the iterative sensemaking process. The main idea is to capture all analysis actions as abstract provenance operations, and then visualizing the analysis history using multiple visual formats so that different types of users can work together. We base our work on results from a design study<sup>11</sup> on supporting a data-driven urban transportation planning project in the City of Toronto, Canada.<sup>12</sup> *Urban transportation planning* is a form of transportation planning concerned with establishing goals, policies, and investments to prepare for future means of moving people and goods from one place to another in an urban environment. Given the importance of transportation in modern cities, urban transportation planning is a key component of most *smart city* initiatives that aim to take advantage of modern information and computing technology to optimize the efficiency, sustainability, and social well-being of a city. In such settings, transportation planning becomes a highly data-driven activity where multiple and heterogeneous data sources are collected and fused to enable elected and career officials to make informed decisions about highway networks, mass transit, street infrastructure, etc. Due to the vast scale of the data as well as the wide range of disciplines involved, transportation planning is a highly collaborative analysis process involving not just expert analysts, but also managers, politicians, and local residents. With such a wide array of stakeholders, it is important that the provenance of the data analysis is maintained, and that the data is presented in a format suitable for the audience.

During the requirements gathering process of the project, our key finding was that while visual methods

are ideal for this purpose, the collaborative nature requires maintaining data provenance, and the wide range of audiences requires multiple different representations. Provenance was primarily required to eliminate the need for reflecting on past work and avoid reinventing the same analysis repeatedly. As for varying audiences, while expert analysts are well-versed in visual analytics and data science workflows and would benefit from a full-fledged data-flow analysis system,<sup>13–15</sup> managers may prefer only a high-level interactive dashboard with the overall findings, and politicians and citizens may just want a slideshow or an infographic summarizing important outcomes.

As part of our multi-phase design process, we designed, implemented, and evaluated STORYFACETS, a data exploration system that maintains full analysis provenance and allows users to generate multiple linked representations of the data and user analysis process from a single source (Figure 1). The STORYFACETS workflow typically begins with an analyst exploring data in a full-fledged analysis view, called the *trail view*. In the background, STORYFACETS automatically maintains multiple different views, or *facets*, of the same data-driven story:

- *Trail view*: a data-flow view<sup>13–15</sup> where the analyst can load datasets, apply data transformations, run statistics, and interact with visualizations.
- *Dashboard view*: a space-filling information dashboard that organizes all or selected parts of the analysis and allows direct interaction and exploration.
- *Slideshow view*: a traditional slideshow format where each visualization is shown as an interactive slide.

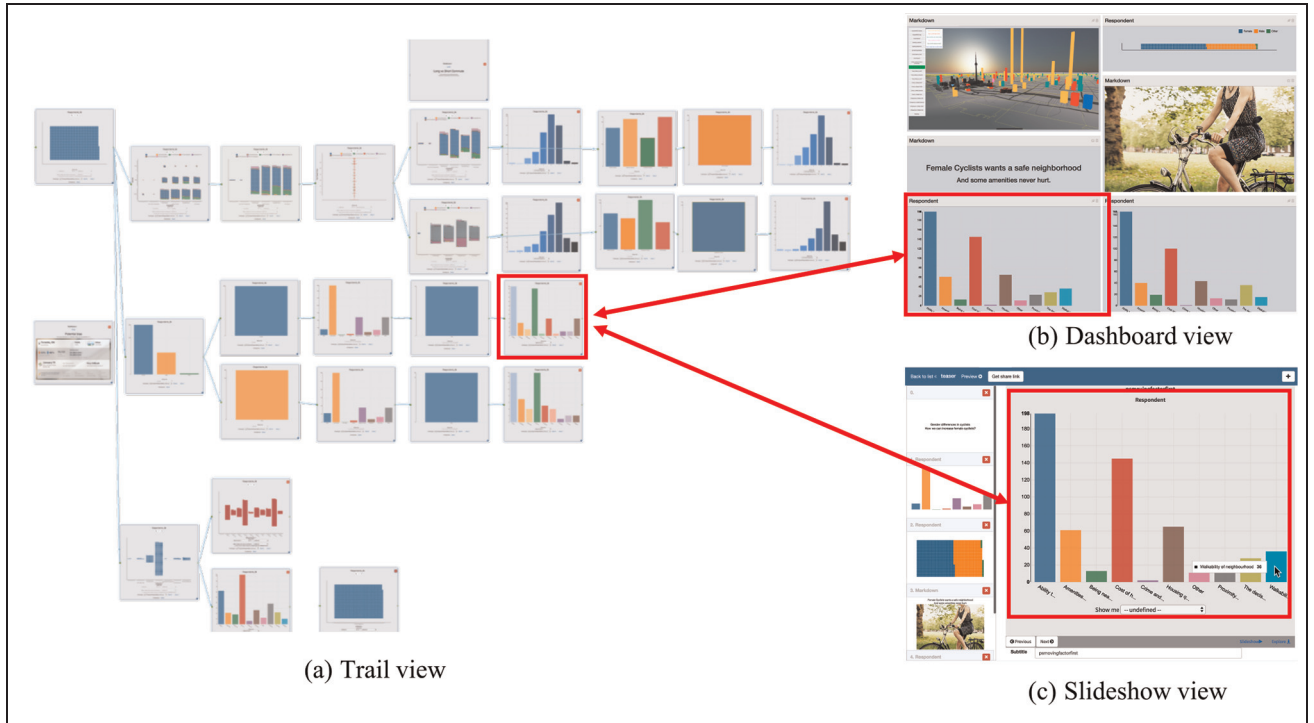
With these different presentation formats, STORYFACETS supports a wide variety of communication scenarios – such as formal presentations, informal analysis reviews, and collaborative analysis – as well as a wide audience, from experts to beginners, all drawing on the same provenance-tracked analysis session.

## Background

Here we review the literature on communication, provenance visualization, and storytelling.

### Communication for visualization

Presenting the outcome of a visual exploration process has always been a priority in visualization research; in fact, it can be argued that the easy accessibility and familiarity of visual communication is one of the primary reasons that visualization is useful. Viégas and



**Figure 1.** The STORYFACETS system maintains provenance of the same analysis session and creates multiple linked formats – called *facets* – for both exploration and presentation. The trail view (a) shows multiple analysis steps and visual provenance links between them on a data flow canvas. The result of one step (red box) is also used in a dashboard (b) and slideshow (c). Users can jump from a visualization in a dashboard or slideshow to its representation on the trail view, so as to see its provenance and exploration context.

Wattenberg<sup>8</sup> unified these ideas into the concept of *communication-minded visualization* (CMV): the notion that useful visualizations are part of a greater ecosystem where viewers also participate in a collaborative data analysis process facilitated by the representation.

Unfortunately, most early visualization systems were designed for expert users and thus provided few visualizations suitable for novice stakeholders such as managers, policy makers, or the general public. In 2007, Pousman et al.<sup>16</sup> captured the grassroots effort to democratize visualization for the masses by reducing barriers as the concept of *casual visualization*. However, even today, many visualization tools still lack an easy path from analysis to presentation.

**Take-away:** The optimal data-driven storytelling method depends on the context and audience of the presentation. Thus, supporting a single presentation format is generally not sufficient. We are aware of no visual analytics tools that support multiple presentation formats.

### Provenance for visualization

Recording the history of an analysis process is referred to as maintaining its *analytic provenance*, and is important for overview as well as recall during the analysis

itself.<sup>3</sup> However, it is also useful for communication to stakeholders; history and provenance can be used to construct stories about a visual exploration such as by serializing the exploration into a slideshow.<sup>17,18</sup> For both of these reasons – improved analysis and improved communication – provenance for visualization has long been an important research topic, and multiple avenues have been explored.<sup>1,3</sup>

Graphical histories is perhaps the most straightforward provenance mechanism. Heer et al.<sup>19</sup> propose different types of graphical histories to save intermediate visualization states during the analysis process. Similarly, Dou et al.<sup>20</sup> demonstrate how interaction logs can help users understand the history of financial data analysis. Sarvghad and Tory<sup>21</sup> study several representations (including sequence diagrams, treemaps, and radial diagrams) to summarize the history of analysis coverage of different dimensions of data sets. In another example, Matejka et al.<sup>22</sup> represent interaction history by augmenting an interface with a heatmap of frequency of button clicks.

Facilitating the user's mental model for provenance is important for both overview and recall. The sense-making loop proposed by Pirolli and Card<sup>23</sup> explains the iterative nature of the analysis process. For

example, once an analyst finds an interesting insight, she might go back to the search and filter process to validate the insight by changing parameters in search of more examples of the same principle.

However, the iterative process changes when multiple analysts are involved. Information sharing on a team often takes place in the final stage of the loop: *presentation*. For this reason, many provenance-tracking visualizations use a spatial analysis workspace where elements are organized in a semi-structured manner. Maintaining spatial persistence in such representations promotes recall<sup>24</sup>; in fact, Ragan et al.<sup>25</sup> found that merely showing the final visual state of a spatial analysis workspace was a sufficient memory aid to significantly help analysts remember the analysis.

Many representations – pioneered by the GRASPARC<sup>13</sup> system – are therefore based on branching *exploration trees* that can be deterministically arranged on a spatial workspace. Similarly, Derthick and Roth<sup>26</sup> show how this form of “branching time model” supports memory off-loading and comparison across time and exploration paths. One of the best-known provenance-tracking systems with a branching exploration tree is VisTrails,<sup>27</sup> which uses a tree diagram to represent sequences of actions, function calls, and resulting visualizations during computational data analysis. Similarly, Shrinivasan and van Wijk<sup>28</sup> use a branching timeline to let the user navigate in time for a complex analysis.

Finally, *data flow systems* replace a strict hierarchy with a directed acyclic graph representation of intermediate states wired together to form a flexible pipeline. The Sandbox<sup>29</sup> is one of the early examples of data flow systems; it provides a semi-structured analysis canvas for intelligence analysis. DataMeadow<sup>15</sup> allows the user to create branching chains of visualization glyphs for multidimensional analysis. Similarly, LARK<sup>30</sup> exposes the full visualization pipeline as a data flow chain on a collaborative space, allowing users to branch and modify the pipeline at different stages. ExPlates<sup>18</sup> automatically generates new nodes in an exploration graph in response to interaction in multidimensional data, such as filtering, selection, or brushing. GraphTrail<sup>14</sup> applies the data flow model to graph and network data, providing chains – or *trails* – of connected charts to visualize, filter, and drill into a dataset.

Following and expanding on the popular tree-style designs for workflow representation, researchers have developed a wide variety of visualization tools. For instance, SensePath<sup>31</sup> focuses on tracking sensemaking and page hopping in web browsing, while AVOCADO<sup>32</sup> provides provenance-graph visualization for biomedical research. Beyond capture and representation of provenance information, other research

seeks to support use of that information. For example, Stitz et al.<sup>33</sup> apply information retrieval approaches to improve effectiveness of search and use of provenance states in visual analysis. Crouser et al.<sup>34</sup> use analytic provenance as a means of studying different user profiles of intelligence analysts and how individual differences influences the analysis process. In other work, Mathisen et al.<sup>35</sup> study the utilization of provenance data as a basis for information sharing for collaborative analysis and intermediate reporting.

**Take-away:** *Visual provenance* facilitates overview and recall both during exploration and afterwards, and *data flow systems* provide a flexible method of explicitly representing provenance. The STORYFACETS system uses a data flow model where each component can be automatically reformatted into multiple communication-oriented representations.

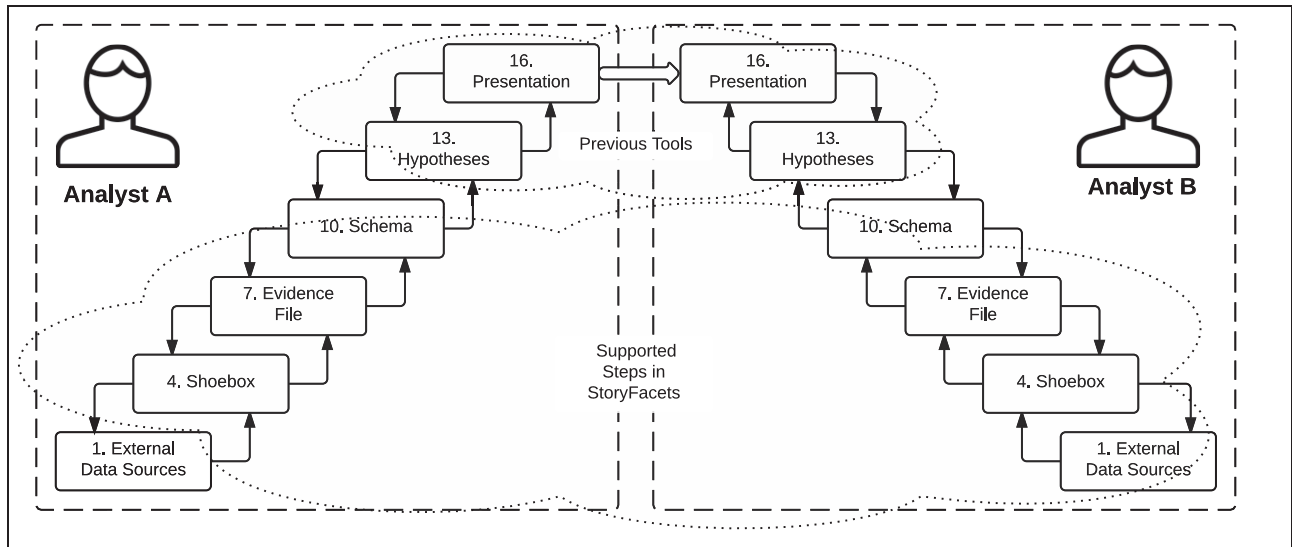
## Storytelling in visualization

*Storytelling* conveys sequences of events using plot, locations, actions, and characters, and visual storytelling is the use of visual communication for storytelling. Already in 2001, Gershon and Page<sup>36</sup> suggested that the combination of storytelling and visualization could become a powerful one, and drew on multiple media such as comics, film, and visual metaphors to argue this point. However, it was not until 2010 that these ideas fully bore fruit, yielding two workshops at the IEEE VisWeek conference in quick succession (2010<sup>37</sup> and 2011,<sup>38</sup> respectively), a survey by Segel and Heer,<sup>5</sup> and a Dagstuhl scientific workshop in 2016. This Dagstuhl seminar also led to the eventual publication of *Data-Driven Storytelling*<sup>6</sup> in 2018.

Segel and Heer’s work is particularly interesting because it identifies seven distinct genres for presenting data narratives: magazine style, annotated chart, partitioned poster, flow chart, comic strip, slide show, and film/video/animation. Taking data narratives a step further, Kosara and Mackinlay<sup>39</sup> argue that storytelling may in fact be the next grand challenge for visualization research, and they go on to survey the history of visual communication and its core mechanisms, such as annotations, highlights, textual descriptions, etc. Several specific narrative visualization techniques have since been proposed, such as the use of sequence,<sup>40,41</sup> geographic stories,<sup>42</sup> spatiotemporal events,<sup>43</sup> sketch-based presentations,<sup>44</sup> and narrative annotations.<sup>45</sup>

Some commercial tools provide support for this activity – for example, Story Points in Tableau<sup>9</sup> and dashboards in Spotfire.<sup>46</sup> Still data analysis and storytelling processes are handled separately in those tools. In comparison, our focus is on connecting the storytelling process and the analysis process organically. The main difference is that a specific chart in a slide





**Figure 2.** The sensemaking loop by Pirolli and Card<sup>23</sup> extended to beyond the presentation step to include multiple users. The findings or hypothesis from analyst A is delivered to analyst B in communication materials such as presentation slides, reports, or infographics. Analyst B should be able to participate in the analysis to examine a hypothesis or test their own alternative ideas. Previous tools such as Story Points in Tableau<sup>9</sup> usually do not allow this. STORYFACETS proposes how combining storytelling and provenance can result in collaborative analysis.

or dashboards is connected to the original exploration in STORYFACETS, preserving the provenance of that chart. Therefore, readers can participate in the exploration of the data with all histories of operations maintained as shown in Figure 2. This leads to the combination of the storytelling with the provenance features.

Our approach leverages design considerations explored by others aiming to support presentation through provenance. For example, Wohlfart and Hauser<sup>47</sup> provided authoring support for visual stories from volume visualization. As another example, Chen et al.<sup>48</sup> presents a method to assist story generation by using topic modeling to generate clusters that can be selected as the starting point for story slides. Also highly relevant, the CLUE<sup>49</sup> approach similarly allows saving and annotating a visual state of a visualization to create presentation slides. While CLUE records the provenance trail in a separate history view for generalizability for different base visualization software, the STORYFACETS design embeds the history record as part of the analysis environment. Our research emphasizes direct coupling between the provenance format of the analysis space and representative snapshots communicated through the presentation space.

**Take-away:** *Visualization and storytelling* is a powerful combination, but many current tools generally lack the support for collaborative analysis beyond presentations.

## Overview: Data analytics for teams

Our research of provenance-supporting visualization is grounded in the need for multiple visual formats that allow accessibility for different types of users to share analysis insights. By automatically maintaining linked provenance views through interaction logs, we designed a system to support iterative analysis and discussion for urban transportation planners. The design study focuses on the iCity smart cities project, a collaboration between the City of Toronto, University of Toronto, OCAD University, the University of Waterloo, Esri Canada, and IBM. Thus, we worked closely with the planners through a multi-phase design process that was inspired by the design study methodology proposed by Sedlmair et al.<sup>11</sup> Our study quickly turned from focusing on the specifics of urban transportation planning to the more general challenge of supporting *data analysis for diverse teams* (i.e. where team members and stakeholders have varying levels of expertise, knowledge, and motivation) through visual provenance formats.

## Phase I: Domain characterization

The analysis domain establishes the foundation of our study of linked provenance formats. As part of the winnowing and discovery stages of a design study,<sup>11</sup> we worked with our urban transportation planning users through interviews, discussions, brainstorming sessions, formative design sketching, and wireframing.

We identified three generalizable roles – analyst, data consumer/manager, and client – and a variety of problems faced by urban transportation planners.

### *User and task analysis*

Our discussions with transportation planning and civil engineering experts revealed three main audiences for us to support. These categories are not disjoint; for example, if given easy access to the necessary tools and data, motivated clients can become analysts in their own right.

**Analysts:** These are the expert users who are conducting the data analysis, either individually or in collaboration with others. Analysts have motivation and capability to learn complex interfaces as well as invest in long analyses. *Unique tasks* include creating, modifying, and presenting a visual exploration. Specific domain users include city planners (municipal workers who designs streets or approves designs), transportation services engineers (municipal workers who works on overall transportation issues), and consultants contracted to design streetscapes. Analyses are triggered by specific events, such as the city deciding to modify the streetscape, advocates calling for an evaluation, and changes to adjacent land use. The goals of the analyst generally include:

- Determine the best allocation of space in a corridor;
- Evaluate consistency of street design w.r.t. demands;
- Identify deficient corridors per transportation mode;
- Compare multiple alternatives w.r.t. costs and benefits;
- Survey clients to determine priorities and feedback;
- Iteratively refine street designs; and
- Convey designs to stakeholders, collect feedback.

**Data consumer/manager:** While not directly involved in data analysis, these users are deeply invested in the outcome of a data exploration. *Unique tasks* include interpreting and presenting the outcome from a visual data exploration. They can apply analysis to new data or may be capable of rudimentary analysis, but generally do not have time or skills for this. Specific domain users include city planners and transportation services engineers (when consultants are the primary analysts), municipal boards, police and emergency service agencies, maintenance providers, transit agencies, and advocacy groups. Their goals include:

- Understand analysis outcomes;
- Evaluate the costs and benefits of alternatives;

- Inform deliberation and negotiations;
- Improve transparency between stakeholders; and
- Understand an analysis process on demand.

**Client:** The end-user or stakeholder of an analysis, the client is a visualization consumer mainly interested in understanding findings. *Unique tasks* include following narratives and validating results. Clients are non-experts in analysis, and generally do not have the resources to engage in analyses themselves. Specific domain users include city councillors, residents and businesses in the study area, people and services that use transportation within the study area, and advocacy groups. Goals of clients include:

- Understand a design proposal;
- Understand analysis outcome and its implications;
- Provide feedback on a design; and
- Understand an analysis process on demand.

### *Design rationale*

The user and task analysis gave rise to several requirements:

R1 *Provenance*: All three of our user groups expressed the need to understand where the data came from, what analysis had already been conducted, and by which analyst.

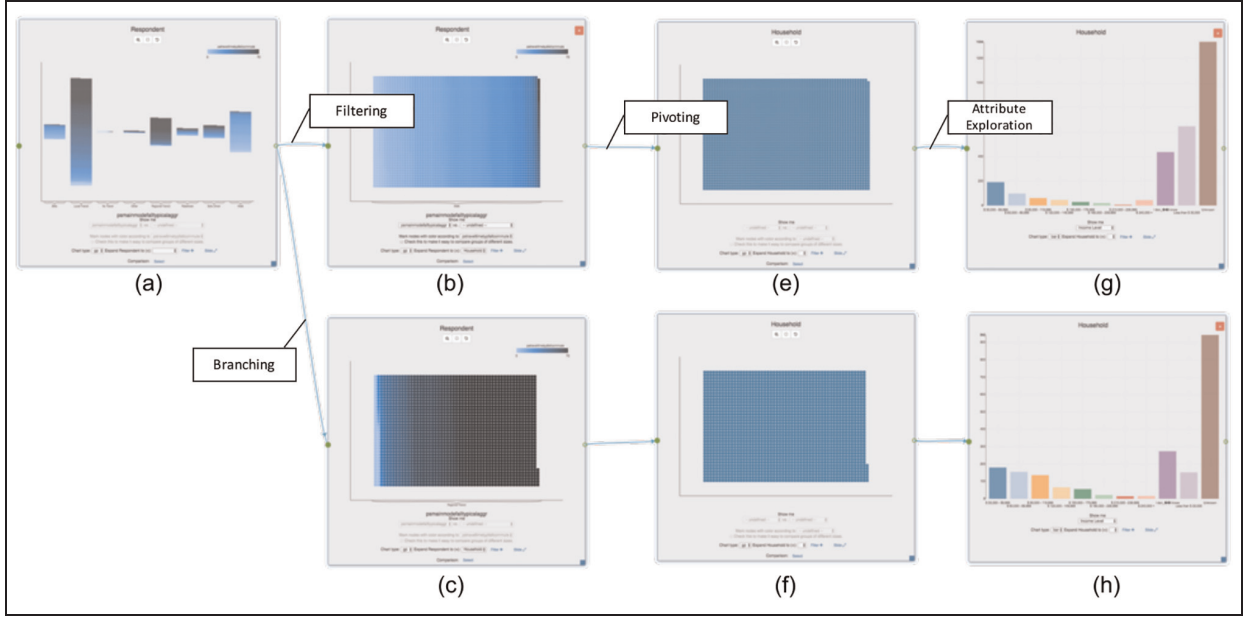
R2 *Data flow system*: Visually representing data provenance promotes overview and recall, which is important for iterative refinement by multiple users.

R3 *One source*: Separating presentation from analysis is time-consuming and error-prone; presentations should update as the analysis changes.

R4 *Multiple media*: Effective visualization design depends on the intended audience and the nature of the presentation; each use case requires its own design.

## **Phase II: Initial tool design**

STORYFACETS is a provenance-tracking visual analytics system for network data where nodes and edges have attributes (i.e. multivariate), and there are multiple types of nodes and edges (i.e. heterogeneous or multimodal). STORYFACETS's unique characteristic is that it not only maintains the provenance of the data exploration, but it also provides a multi-format representation of its progress. Each such representation is called a *facet*. Users can chain together *cards* to visualize data or show annotations. The system automatically updates the different facet views that serve as different formats for presentation. Below, we discuss the rationale behind our design choices, the data model, and the user interface.



**Figure 3.** Visualization cards in an example trail view: (a) shows students by commute method. Node color represents individual commute times. To compare those groups, (b) shows the result of *filtering* to students who walk while (c) is filtered to students using regional transit. As (b, c) have the same parent card (a), we conceptually have a *branching* exploration. Through *pivots*, all households each student belongs to are shown in cards (e, f). We examine household income in (g, h). Students using regional transit tend to be from households with more income than students who walk. Perhaps this reflects the trend that students who live with wealthy parents tend to live in suburbs far from campus. .

### Data model

The STORYFACETS data model is based on multimodal networks. The main operations needed are linking, filtering, and pivoting. Linking is the process by which users create links between data tables they have loaded, building the network. Filtering is reducing the number of elements displayed through interaction. Pivoting is when users select a set of nodes in the network and then traverse edges to select a connected set of nodes. For example, we can do a many-to-many pivot from several students to the households they are part of by pivoting on the “student-household” edge type.

We also use a directed acyclic graph to track the analytic provenance of an exploration. Acyclic graphs have been frequently used in the previous work<sup>14,19,27</sup> to maintain provenance for visual analytics. Nodes represent either (1) a source data set, (2) a subset of data resulting from a query along with an associated visualization and parameterization of that visualization, (3) a data transformation, or (4) Markdown content. Edges represent specific user operations.

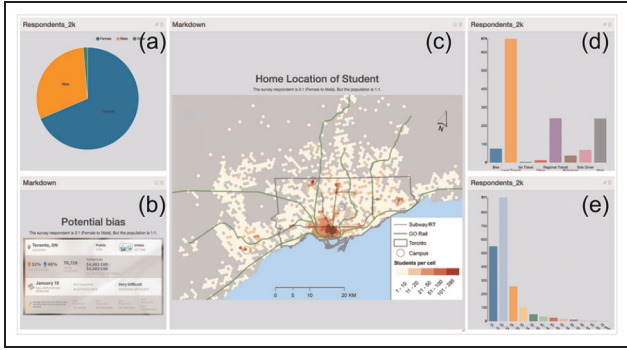
### Cards

The basic visual element used in STORYFACETS is the *card*, which contains content along with interactive

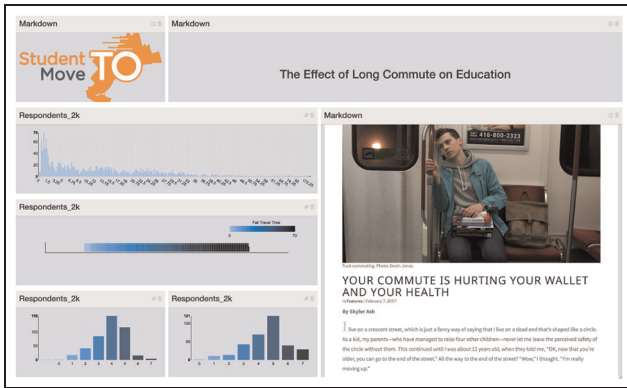
widgets for editing. A key feature of a card is that they are persistent across all views in the system. This supports the *provenance, primarily* (R1) requirement. Each card has UI widgets that allow cards to be selected, resized, scrolled, zoomed, reset (zoom and scroll), deleted, or made full screen. *Visualization cards* (Figure 1) include a visualization of a subset of the data from the *data flow system* (R2) as well as parameterization widgets and a caption. Visualization cards include additional widgets for filtering, pivoting, selection, axis, and color variable selectors, and choosing relative or absolute sizing. Filtering and pivoting create new *child* cards that are linked to their *parent* card in the provenance graph. For example, see the labeled links between cards in Figure 3. Our modular architecture allows easy integration of new visualization types.

### Facet views

Cards can be arranged in *facet* views to present a subset of the provenance graph and underlying data. To support the *multiple media* (R4) design rationale, we provide different views for each of our target audiences: *trail view*, *dashboard view*, and *slideshow view*. Following previous work,<sup>14,19,27</sup> the *trail view* is constructed as the user explores the data providing the



**Figure 4.** Dashboard view showing Markdown and visualization cards in a customizable, space-filling orthogonal layout: (a) shows the gender composition of respondents, (b) warns of potential bias by comparing student enrollment statistics to survey results, (c) shows student home locations and transportation infrastructure, (d) shows the main commuting methods, and (e) shows the age distribution of respondents.



**Figure 5.** The dashboard view supports infographic elements such as text, images, videos, and mashups to tell a story.

graphical history or sense-making log. The *trail view* (Figure 3) is designed for *analysts* and provides a visual *data flow system* (R2) that renders the directed acyclic provenance graph of the entire exploration history as a node-link diagram on a zoomable 2D canvas. Large rectangular nodes contain visualization cards, and directed edges show user operations that create new subsets of the data. New cards created by user interaction are automatically placed, and can then be rearranged. By placing visualization cards from parallel explorations beside each other, we support *direct comparison* through juxtaposition. Cards can be locked, preventing dependent cards from being affected.

The core novel method is maintaining the visual provenance of the visualization cards from the *trail view* so they are *reused* to construct the communication



**Figure 6.** Slideshow view with Markdown and visualization cards presented linearly. Each card is interactive and visualization parameters can be changed to explore the various attributes. Thumbnails for the other cards/slides are seen on the left.

medium such as infographics or slideshows. The *dashboard view* (Figures 4 and 5) for data consumers or managers organizes the cards in a compact layout. Users can choose which cards to include, and they can move and resize cards to achieve the desired layout. Again, *direct comparison* between alternatives is achieved by juxtaposing cards.

The *slideshow view* (Figure 6) for *clients* shows each card as a slide. Users choose which cards to include as well as their order of appearance using an interface similar to Microsoft PowerPoint. Users can also include an existing dashboard view as a slide. Slides can be presented in full-screen mode with one card/dashboard per screen.

By *reusing* cards, the benefit is that we can maintain the history of analysis beyond the analysis stage into presentation stage. For example, the readers can verify how the conclusion is derived by looking at the analysis process from the raw data to the visualizations. Or readers can explore different aspects of data or *branch* the analysis to test different hypothesis. Cards remain interactive both for editing and presentation. Live views support *one source* (R3), as edits are immediately updated in other views. Read-only views do not preserve edits so as to prevent unintentional alterations, but a read-only view can be forked as a new live view.

### Implementation details

We implemented StoryFacets as a web-based client/server framework\*. The client uses the AngularJS, D3,<sup>50</sup> NVD3,<sup>51</sup> and JQuery libraries. The server uses

\*The source is available as an open source project at <https://gitlab.com/intuinno/graphtrail/>



Meteor.js and MongoDB. The provenance/data flow graph is stored in a MongoDB NoSQL database where each node contains the relevant subset of the data, which can be large. Meteor.js syncs the database contents from the client and server. This somewhat limits scaling to larger datasets because of the communication overhead but provides the benefit of making the creation of shareable views straightforward because each card is self-contained. Future implementations could simply store a snapshot of the entire dataset, pre-computed SVGs for each visualization card, and a graph of operations which could be re-applied when it is necessary to rebuild a subset of the data, for example, for branching from an existing trail. Because all the data and cards are stored on the server side, users can simply send a URL to share the slides or dashboards.

For ease of development, the visualizations are implemented in SVG. However, this may lead to rendering scalability issues for particularly large explorations or those with many marks in a visualization. Future implementations may wish to use the more scalable Canvas, for example, converting from SVG automatically (<https://www.ssvg.io/>) or reimplementing, or WebGL.

### Phase III: Feedback from experts

We conducted a structured interview with experts<sup>52</sup> with the initial STORYFACETS design to collect insights about appropriateness of different views for different purposes. Because we were looking for actionable guidelines on how to best improve the tool, we asked three experienced visual analytics researchers from industry.

The participants were recruited based on their experience in the field. Given the demands on their time, we only requested 1 h for the study (actual times ranged between 59–65 min). Because our experts were geographically distributed, the expert reviews were conducted remotely using video conferencing with screen sharing. The experts were also briefed on the general problem statement that STORYFACETS aims to address prior to beginning the study. We were most interested in their feedback about the use of multiple views for different purposes or audiences. Thus, we specifically told the experts we were interested in their feedback about the design.

Since our experts were all external to the project, we decided to use Star Wars movie data from the Star Wars API (<https://swapi.dev/>, originally at <https://swapi.co/>) rather than specific urban transportation planning data, which would have required lengthy instructions and training. This dataset contains various types of information about characters, places, and

vehicles in the fictional Star Wars universe. As with the types of data commonly used for urban transportation planning, the Star Wars data also involves multidimensional information and entity relationships, which makes it sufficient to demonstrate the true design focus – the multiple linked facets of the STORYFACETS tool combined with visual provenance.

The procedure called for our experts to explore the data in a free-form fashion while speaking their observations and thoughts aloud. To guide their exploration, we seeded the experts with a list of questions.

### Findings

Here, we focus on qualitative feedback rather than task performance. The main finding is that the experts all saw different and specific usages for the various views.

All three experts stated that the trail view was better suited for a technical audience. E3 further thought this view would be the best for sharing results, at least with other analysts. However, E2 noted that a complicated analysis with many branching paths may cause the trail view to grow out of control. Nevertheless, the trail view was collectively lauded across all three experts; this was not surprising, as they were representatives of the intended users of this view.

The dashboard view was seen by E1 and E2 as most suitable for presentation to a less technical audience (e.g. management). E1 thought that it represents a good tradeoff between clarity and flexibility, and can even be used to explain complex data analysis with many branches. Simple branching can be shown as parallel rows or columns, especially with annotations. It could also support off-the-cuff presentations of an analysis currently in progress when the optimal order of presentation has not yet been established. E2 particularly enjoyed the animated data transitions in the slideshow view.

Of the three views, the slideshow view was the most controversial; all experts agreed that its utility was limited to presentations to novice stakeholders, but that it was highly useful for this specific purpose. E1 noted that creating a slideshow requires knowing the correct order of presentation, which is not always known in the midst of data analysis, but E2 stated this was the very aspect of the slideshow view that made it appealing once such an order is established. One compelling scenario E1 suggested was that when preparing a routine presentation for management, a traditional presentation can be quickly and easily created. However with STORYFACETS, a presenter could switch to the trail view and retarget the presentation for an unengaged audience.

The experts all gave suggestions for future improvement. E1 noted a Prezi-style interactive tour of the

workspace in the trail view could be a good alternative to present the state of a visual exploration to other analysts.<sup>53</sup> Both E2 and E3 agreed that the trail view may also be useful for communication, and suggested adding the ability to add annotations directly. This was unexpected because we designed the trail view to be primarily an exploration space and the others as presentation formats. However, in many cases, this boundary is not strict, and the trail view can also function as an effective communication medium. This leads to a fundamental trade-off between provenance and presentation. For representing data provenance, a complex analysis trail should be preserved as a reminder of previous actions. But for the purpose of presentation to non-experts, complex exploration processes with branching analysis pathways and dead-end results may be irrelevant, redundant, or unrelated to the intended message. E2 noted a data flow system such as our trail view could easily become visually complex, and suggested simplifying the workspace with mechanisms such as editing exploration paths, collapsing or expanding branches, or eliminating fruitless paths – though removing fruitless paths may lead to later redundant effort.

E3 suggested version management for cases when the original author shares the exploration and colleagues build on the exploration in the original space. This use case raises important questions about how to facilitate modifications, notify the original author, and visualize differences across versions of the same exploration.

### *Outcome: Modifications needed*

Since the intention of this study was to guide the design of STORYFACETS, a key aspect was identifying actionable modifications. Here are the changes needed based on the study:

- Adding annotation capabilities to each view;
- Adding standard visualization types;
- Making cards and views responsive and resizable;
- Rectangular, individual item, and modifier key interactions for selecting items and aggregates;
- Maintaining consistent color scales across cards;
- Using a natural sort order for labels that sorts string and numeric components separately; and
- Fixing label overplotting.

### **Phase IV: Iterative tool refinement**

We refined STORYFACETS iteratively based on the results of the expert review as well as informal usability tests.

To support annotations – as well as to integrate qualitative data into an exploration – we developed Markdown cards. *Markdown cards* allow annotations and qualitative content to be added, such as text captions, bullet points, hyperlinks, images, video, and even interactive webpages. These can be used for integrating the results of analyses conducted in other tools, including embedded web pages, images, and video. Examples are shown in Figure 5, where they are used to create an infographic-style interactive dashboard suitable for novice users. Markdown cards can also take the place of non-visualization slides in the slideshow. This can help build a strong narrative about the data.

Several changes we made were targeted at increasing consistency and readability. This included using top-level color scales for each attribute for all views as well as an improved label ordering algorithm. We also implemented responsive resizing of cards in the trail and dashboard view to support exploring elements in more detail.

While our project was canceled at this stage, preventing us from running an in-depth followup study with our group of urban transportation planners, informal feedback on the new version of our tool was very positive.

### **Usage scenario**

Our design study with STORYFACETS explores the automatic generation of presentation formats from provenance logs and demonstrates the benefits of using multiple linked visual representations for different types of users. Here, we present a usage scenario to demonstrate how our approach facilitates collaboration and discussion through linked visual formats generated via interaction logs. For this scenario, consider Jane, a municipal worker in charge of transportation service policy. She wants to improve the mobility of students around the four campuses in Toronto without burdening the existing transportation infrastructure by adopting bike and car sharing platforms. Jane recruits John, a contractor data scientist, to assist her project. John analyzes the recently collected *StudentMoveTO* survey data (<http://www.studentmoveto.ca/>) about university student behavior and travel.

John begins by loading all respondents into STORYFACETS. Since the survey is voluntary and the response rate is low, the distribution of respondents might not reflect the target population. He finds that female students responded twice as often as male students, despite an enrollment gender ratio of 48:52. This might be a source of bias, and he adds a Markdown note with this insight as shown in Figure 4.

He continues the exploration comparing the differences between the long-commute and short-commute groups. He visualizes the primary motivator for

selecting housing, and finds that housing cost and the ability to walk or bike are key factors. He also finds that longer commutes are correlated with students attending campus less frequently. Unexpectedly he found that students with longer commutes on regional transit tend to be from wealthier households as shown in Figure 3. To summarize his findings, John prepares two reports in separate formats. The first is a one-page executive summary dashboard consisting of key findings and visualization cards. He selects a few key cards and organizes the layout to align with his understanding of commute time and wealth. Next, he needs to share progress report with Jane, so he composes a slideshow by choosing cards from his exploration for a presentation in the next project meeting.

Later, while reviewing the presentation at the project meeting, Jane asks a question about a specific visualization card showing the relative income level between the bicyclist and car sharing groups. STORYFACETS allows John to answer the question by interactively changing from the slideshow view to the trail view, where the provenance of the data operations help communicate exactly how John's inference was derived. During the meeting, they decide to prepare a press release to inform parents that a long commute has a detrimental effect on on-campus time – therefore promoting independent living near campus as a better alternative. John uses the dashboard to create the base infographic to share the results as shown in Figure 5. He publishes the dashboard online where students and parents can see it.

After the meeting, John shares a link to his slides in STORYFACETS. Another colleague, Kate, would like to explore her hypothesis about the different factors for selecting housing between the female and male bicyclist groups. She clicks “show trail” on the relevant card, and the system shows the exploration and the location of the specific cards on the exploration canvas. Using STORYFACETS to take advantage of the analysis history from John's analysis, she continues from John's gender analysis and finds that the female students are half as likely to choose bicycling compared to male students. She compares the motivation for house selection between male and female students and finds that females prioritize walkable safe environments more than male students. She concludes that by promoting safer clean neighborhoods around campus, she can increase female bicycle use. She prepares another infographic and slides to share this result as shown in Figure 1(b) and (c).

There are two signature interaction patterns that we would like to emphasize here. First, the provenance capture capability with linked formats allowed John to easily share the results of his analysis as a presentation, trail view, and dashboard/infographic from the same

base exploration for distinctly different audiences. That is, he (1) presented his main findings during the meeting, (2) allowed a colleague to continue a branching analysis from his analytic provenance, and (3) directly created an infographic for public dissemination from the same application. This demonstrates the core value of the approach: by integrating analytic provenance from the analysis with multiple visual formats for different purposes, it is possible to maintain the record of analytic activity for the benefit of vary different types of audiences. And a specific benefit is narrowing the analyst-client gulf for data communications – a major challenge that relates to trust and understanding of data reporting – through the application of provenance.

## Discussion

We developed STORYFACETS, a visual analytics system for supporting multiple user roles in collaborative data analysis using urban transportation planning as context. We made a few design choices that are different from existing literature. In this section, we summarize what we think are generalizable lessons from our work, which we hope can lead to better collaborative analysis tools in the future.

### *L1: Sensemaking beyond the powerpoint gap*

The existing sensemaking loop<sup>23</sup> illustrates that exploratory analysis is conducted in an iterative loop. However, this iteration breaks after the presentation step; viewing the presentation is outside of the loop, making it difficult for consumers to participate in the analysis. We claim that this barrier should be overcome by adding provenance capability to the presentation material. This way, consumers can see how this visualization has been derived from original data and a sequence of transformations. This leads to better analysis for two reasons: First, provenance allows consumers to verify the validity of the analysis. Second, it facilitates consumers building on top of the existing analysis, thus contributing with new analysis and resulting insights. This was raised by E2 as well as highlighted in the usage scenario when another colleague, Kate, could conduct her own analysis from the slides that John shared. To the best of our knowledge, this functionality is novel in the literature, as well as among commercial products such as Tableau, QlikView, and Spotfire.

### *L2: Multiple formats from one analysis*

Different audiences and contexts require different formats for sharing insights such as infographics, slides,

or trail views. The previous usage scenario demonstrated how different views have different uses for supporting provenance in collaborative settings with multi-role teams. The expert review in Phase III suggested that the trail view was preferred for data analysis, whereas the slideshow view was favored for formal presentations. This confirms that the basic rationale behind STORYFACETS is sound: data exploration can be viewed through several radically different lenses – each with a unique and valuable *raison d'être*. In other words, this is a validation of our “*one source, multiple media*” motto that arose in the early stages of this design study.

### *L3: Tighter integration of analysis and publication*

Historically data analysis and sharing results have been two separate activities. For example, analysts conduct analysis using a spreadsheet program and paste the resulting charts into other software to share the findings. Novel commercial products such as Tableau or Microsoft Power BI changed this practice by incorporating the capability to compose sharing materials. Integration between the analysis and publication steps will also be helpful in implementing the L1 and L2 objectives.

### *Limitations and future work*

We propose three generalizable lessons above. However, it remains an open question how each of these lessons contribute to better collaborative analysis. To evaluate this, a large-scale followup study with a large number of teams of different users roles would be required. Unfortunately, our overall research project on urban transportation planning was canceled before we could deploy this work with the original intended users. However, while much more work clearly remains to be done, the findings from this design study will be formative and point to the overall utility of the ideas we uncovered in the project.

One specific goal for future improvement includes media and mechanisms for sharing the provenance of exploratory analysis. Another improvement would be to automatically organize visualizations and findings based on the content. For example, if two adjacent exploration branches show two related data subsets, the dashboards or slide layout algorithm should be able to position them intelligently side-by-side to enable easy comparison. Similarly, the animation for the filtering in the slide deck can be improved so the items maintain their identity over the slides. Animation support could show entities appearing and disappearing in

response to filtering and pivoting operations to maintain object identity.

## Conclusion

Exploratory data analysis involves much more than the initial data exploration that generates findings; the analysis and findings must often be shared with colleagues, discussed with managers, and eventually presented to stakeholders or the general public.<sup>7</sup> In this paper, we reported on an in-depth design study with urban transportation planners, which yielded a common theme about the need for multiple stakeholders of varying expertise to have access to the outcomes of data analysis. As a result, we designed STORYFACETS, a communication-minded visualization system that maintains the provenance of all data exploration and provides multiple, linked visual formats for analysis and presentation. Our mixed-method validations support this design rationale and provide new insights for designing multi-faceted approaches.

## Acknowledgements

We would like to thank our expert review participants and reviewers for their feedback on this work, and the iCity team for their assistance with this project.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially funded by the Ontario Research Fund, IBM Corporation, DARPA XAI program N66001-17-2-4032, and the U.S. National Science Foundation grant 1539534 and 1565725. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the funding agencies.

## ORCID iDs

Deokgun Park  <https://orcid.org/0000-0003-0054-9944>

Niklas Elmqvist  <https://orcid.org/0000-0001-5805-5301>

## Supplemental material

Supplemental material for this article is available online.

## References

1. Freire J, Koop D, Santos E, et al. Provenance for computational tasks: a survey. *Comput Sci Eng* 2008; 10(3): 11–21.



2. Groth DP and Streefkerk K. Provenance and annotation for visual exploration systems. *IEEE Trans Vis Comput Graph* 2006; 12(6): 1500–1510.
3. Ragan ED, Endert A, Sanyal J, et al. Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes. *IEEE Trans Vis Comput Graph* 2016; 22(1): 31–40.
4. Clark HH and Brennan SE. Grounding in communication. In: Resnick LB, Levine JM and Teasley SD (ed.) *Perspectives on Socially Shared Cognition*. Washington, DC: American Psychological Association, 1991, pp.1270–149.
5. Segel E and Heer J. Narrative visualization: telling stories with data. *IEEE Trans Vis Comput Graph* 2010; 16(6): 1139–1148.
6. Henry Riche N, Hurter C, Diakopoulos N, et al. (eds). *Data-driven storytelling*. New York, NY: AK Peters/CRC Press, 2018.
7. Madanagopal K, Ragan ED and Benjamin P. Analytic provenance in practice: the role of provenance in real-world visualization and data analysis environments. *IEEE Comput Graph Appl* 2019; 39: 30–45.
8. Viegas FB and Wattenberg M. Communication-minded visualization: a call to action [technical forum]. *IBM Syst J* 2006; 45(4): 801–812.
9. Kosara R. Story points in tableau software. In: *Keynote at tableau customer conference*, Washington, DC, USA 2013.
10. Thomas JJ and Cook KA (eds). *Illuminating the path: the research and development agenda for visual analytics*. Washington, DC: IEEE Computer Society Press, 2005.
11. Sedlmair M, Meyer M and Munzner T. Design study methodology: reflections from the trenches and the stacks. *IEEE Trans Vis Comput Graph* 2012; 18(12): 2431–2440.
12. Toronto Centre for Active Transportation. Complete streets for Canada. <http://completestreetsforcanada.ca>. (2012, accessed 6 September 2018).
13. Brodlie K, Poon A, Wright H, et al. GRASPARC: a problem solving environment integrating computation and visualization. In: *Proceedings Visualization '93*, San Jose, CA, USA, 1993, pp. 102–109.
14. Dunne C, Riche NH, Lee B, et al. GraphTrail: analyzing large multivariate, heterogeneous networks while supporting exploration history. In: *Proc. ACM Conference on Human Factors in Computer Systems*, Austin, TX, USA 2012, pp. 1663–1672.
15. Elmqvist N, Stasko J and Tsigas P. DataMeadow: a visual canvas for analysis of large-scale multivariate data. In: *IEEE symposium on visual analytics science and technology*, Sacramento, CA, USA, 2007, pp. 187–194.
16. Pousman Z, Stasko J and Mateas M. Casual information visualization: depictions of data in everyday life. *IEEE Trans Vis Comput Graph* 2007; 13(6): 1145–1152.
17. Javed W and Elmqvist N. Stack zooming for multi-focus interaction in time-series data visualization. In: *IEEE Pacific visualization symposium (PacificVis)*, Taipei, Taiwan, 2010, pp. 33–40.
18. Javed W and Elmqvist N. ExPlates: spatializing interactive analysis to scaffold visual exploration. *Comput Graph Forum* 2013; 32(3pt4): 441–450.
19. Heer J, Mackinlay J, Stolte C, et al. Graphical histories for visualization: supporting analysis, communication, and evaluation. *IEEE Trans Vis Comput Graph* 2008; 14(6): 1189–1196.
20. Dou W, Jeong DH, Stukes F, et al. Recovering reasoning processes from user interactions. *IEEE Comput Graph Appl* 2009; 29(3): 52–61.
21. Sarvghad A and Tory M. Exploiting analysis history to support collaborative data analysis. In: *Graphics interface conference*, Halifax, NS, CA, 2015, pp. 123–130.
22. Matejka J, Grossman T and Fitzmaurice G. Patina: dynamic heatmaps for visualizing application usage. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, Paris, FR, EU, 2013, pp. 3227–3236.
23. Pirolli P and Card S. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In: *Proceedings of the international conference on intelligence analysis*, 2005, volume 5, pp. 2–4.
24. Ragan ED, Endert A, Bowman DA, et al. How spatial layout, interactivity, and persistent visibility affect learning with large displays. In: *Proceedings of the international working conference on advanced visual interfaces - AVI '12*, Capri, Island, IT, 2012, pp. 91–98.
25. Ragan ED, Goodall JR and Tung A. Evaluating how level of detail of visual history affects process memory. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, Seoul, KR, 2015, pp. 2711–2720.
26. Derthick M and Roth SF. Enhancing data exploration with a branching history of user operations. *Knowl Based Syst* 2001; 14(1-2): 65–74.
27. Bavoil L, Callahan SP, Scheidegger CE, et al. VisTrails: Enabling Interactive Multiple-View Visualizations. In: *VIS 05. IEEE visualization*, Minneapolis, MN, USA, 23–28 October 2005, pp. 135–142. IEEE.
28. Shrinivasan Y and van Wijk J. Supporting the analytical reasoning process in information visualization. In: *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, Florence, IT, 2008, pp. 1237–1246.
29. Wright W, Schroh D, Proulx P, et al. The sandbox for analysis: concepts and evaluation. In: *Proceedings of the ACM conference on human factors in computing systems*, 2006, San Jose, CA, USA, pp. 801–810.
30. Tobiasz M, Isenberg P and Carpendale S. Lark: coordinating co-located collaboration with information visualization. *IEEE Trans Vis Comput Graph* 2009; 15(6): 1065–1072.
31. Nguyen PH, Xu K, Wheat A, et al. Sensepath: understanding the sensemaking process through analytic provenance. *IEEE Trans Vis Comput Graph* 2016; 22(1): 41–50.
32. Stitz H, Luger S, Streit M, et al. Avocado: visualization of workflow-derived data provenance for reproducible

- biomedical research. *Comput Graph Forum* 2016; 35(3): 481–490.
33. Stitz H, Gratzl S, Piringer H, et al. KnowledgePearls: provenance-based visualization retrieval. *IEEE Trans Vis Comput Graph* 2018; 25(1): 120–130.
34. Crouser RJ, Ottley A, Swanson K, et al. Investigating the role of locus of control in moderating complex analytic workflows. *EuroVis 2020 Short Papers*, 2020. DOI:10.2312/evs.20201050.
35. Mathisen A, Horak T, Klokmoose CN, et al. InsideInsights: integrating data-driven reporting in collaborative visual analytics. *Comput Graph Forum* 2019; 38(3): 649–661.
36. Gershon N and Page W. What storytelling can do for information visualization. *Commun ACM* 2001; 44(8): 31–37.
37. DiMicco J, McKeon M and Karahalios K. Telling stories with Data—a VisWeek 2010 workshop, 2010.
38. Diakopoulos N, DiMicco J, Hullman J, et al. Telling stories with data: the next Chapter—a VisWeek 2011 workshop, 2011.
39. Kosara R and Mackinlay J. Storytelling: the next step for visualization. *IEEE Computer* 2013; 46(5): 44–50.
40. Hullman J and Diakopoulos N. Visualization rhetoric: framing effects in narrative visualization. *IEEE Trans Vis Comput Graph* 2011; 17(12): 2231–2240.
41. Hullman J, Drucker S, Henry Riche N, et al. A deeper understanding of sequence in narrative visualization. *IEEE Trans Vis Comput Graph* 2013; 19(12): 2406–2415.
42. Gao T, Hullman J, Adar E, et al. NewsViews: an automated pipeline for creating custom geovisualizations for news. In: *Proceedings of the ACM conference on human factors in computing systems*, Toronto, CA, 2014, pp. 3005–3014.
43. Eccles R, Kapler T, Harper R, et al. Stories in GeoTime. *Inf Vis* 2008; 7(1): 3–17.
44. Lee B, Kazi RH and Smith G. SketchStory: telling more engaging stories with data through freeform sketching. *IEEE Trans Vis Comput Graph* 2013; 19(12): 2416–2425.
45. Satyanarayan A and Heer J. Authoring narrative visualizations with Ellipsis. *Comput Graph Forum* 2014; 33(3): 361–370.
46. Ahlberg C. Spotfire: an information exploration environment. *SIGMOD Record* 1996; 25(4): 25–29.
47. Wohlfart M and Hauser H. Story telling for presentation in volume visualization. In: *Proceedings of the 9th joint Eurographics/IEEE VGTC conference on visualization*, Norrköping Sweden, 2007, pp. 91–98.
48. Chen S, Li J, Andrienko G, et al. Supporting story synthesis: bridging the gap between visual analytics and storytelling. *IEEE Trans Vis Comput Graph* 2020; 26: 2499–2516.
49. Gratzl S, Lex A, Gehlenborg N, et al. From visual exploration to storytelling and back again. *Comput Graph Forum* 2016; 35(3): 491–500.
50. Bostock M, Ogievetsky V and Heer J. D<sup>2</sup>: data-driven documents. *IEEE Trans Vis Comput Graph* 2011; 17(12): 2301–2309.
51. Novus Partners. NVD3 re-usable charts for d3.js. <http://nvd3.org/>, 2014.
52. Tory M and Möller T. Evaluating visualizations: do expert reviews work? *IEEE Comput Graph Appl* 2005; 25(5): 8–11.
53. Laufer L, Halácsy P and Somlai-Fischer A. Prezi meeting: collaboration in a zoomable canvas based environment. In: *Proceedings of the 2011 annual conference extended abstracts on human factors in computing systems - CHI EA '11*, 2011, pp. 749–752.