Eurographics Conference on Visualization (EuroVis) 2016
E. Bertini, N. Elmqvist, and T. Wischgoll
(Guest Editors)

Short Paper

Clarifying hypotheses by sketching data

Mariana Mărășoiu[†1], Alan F. Blackwell[1], Advait Sarkar[‡1] and Martin Spott[§2]

[1]Computer Laboratory, University of Cambridge, United Kingdom
[2]HTW Berlin, Germany

Abstract

Discussions between data analysts and colleagues or clients with no statistical background are difficult, as the analyst often has to teach and explain their statistical and domain knowledge. We investigate work practices of data analysts who collaborate with non-experts, and report findings regarding types of analysis, collaboration and availability of data. Based on these, we have created a tool to enhance collaboration between data analysts and their clients in the initial stages of the analytical process. Sketching time series data allows analysts to discuss expectations for later analysis. We propose function composition rather than freehand sketching, in order to structure the analyst-client conversation by independently expressing expected features in the data. We evaluate the usability of our prototype through two small studies, and report on user feedback for future iterations.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces and Presentation]: User Interfaces—Interaction styles

1. Introduction

Information visualisation and visual analytics would be much easier if only those who ask for visualisations were more explicit in the analytical requests they make. In practice, they are not, primarily because they are not themselves trained as statisticians or analysts. As a result, the day-to-day work of many professional analysts involves clarifying statistical concepts and helping customers to formulate hypotheses. For example, Kandel et al. observe that analysts have to translate high-level business questions into low-level analysis tasks, and that the process is error prone [KPHH12]. Collaboration between data analysts and their clients could be enhanced with visualisation tools to help such knowledge sharing [HA07] and sensemaking [PC05].

Interviews with data analysts suggest that they often discuss hypotheses with their clients before data is collected. We consider sketching as a strategy for supporting this discussion. Moving beyond traditional pen on paper drawing, our goal is to structure the conversation around expected trends. We draw on the Automatic Statistician approach [DLG*13], using function composition to support structured sketches of time series data. We chose time series visualisation as our interviews suggest it to be one of the most used types of data that analysts work with.

We follow a user-centred design approach as used in the visual languages community, similar to the design study methodology used in information visualisation research [SMM12].

2. Interviews with data analysts

In order to understand more about how data analysts communicate with people requesting analysis work, we conducted semi-structured interviews with four professional data analysts working for BT. Each interview was structured around a recent job done by that analyst, taking into account its typicality, collaboration with other people, software used, iteration cycles, final result (e.g. a chart, a spreadsheet) and its intended use. Two of the analysts worked with collocated colleagues, and two received assignments via phone calls and email from people in other locations. We used Thematic Analysis [BC06] to develop codes and identify themes.

2.1. Results

Types of analyses

Analysts identified several distinct types of request: **Data request**: The analyst is asked only to gather the data and send it to the requester without further analysis; **Insight request**: The analyst is asked to investigate a hypothesis (e.g. *"can you look at the data and see if that's the actual case"* - P4); and **Model request**: The analyst is asked to build a model (a spreadsheet or set of visualisations), to be used by the requester for further analysis (e.g. *"we want a model, can you build a model and give us that"* - P3).

† Mariana is a Vice-Chancellor's Scholar and is supported by an EPSRC industrial CASE studentship co-sponsored by BT. She is also supported by a Qualcomm European Research Studentship in Technology.
‡ Advait is supported by an EPSRC+BT iCASE award, and a Cambridge Computer Laboratory Robert Sansom Scholarship.
§ Martin was with BT Research & Technology during the interviews with BT data analysts.

DOI: 10.2312/eurovisshort.20161173

Questions need to be clarified before analysis

Analysts reported that they often need a conversation with the requester after receiving an email from them in order to clarify and fill in missing details of their high-level question (*"So when I get something like that it will usually be let's have a look at it, let's think about where the gaps are, got to list all the gaps, all of the things that I need answered now, then ask that question, get the information that I actually need to finish it off"*, P3). In guiding the client to refine a question, the analyst is sharing domain and statistical knowledge. Even when the analyst has a good intuition of what is asked for, they still like to double check the requirement (*"Most often it's a quick call saying «is this exactly what you're after, do you need this, this and this and this?»"*, P4).

Data is not always readily available

Analysts often don't have the data needed to answer a request at the time it is made, or during the clarification discussion. Because data is spread across many tables, different users and teams, there are overheads in finding the right data to start analysis. As noted by P4 *"there is no set of rules, no strict process"* about who to contact. Analysts usually create their own databases before they can start work, but this involves obtaining permission to access it, cleaning it, dealing with missing data, and adjusting the schema - whether *"the model of the data is in the right format"* (P1).

The above three key insights into the work practices of data analysts suggest that the clarification call between the analyst and their client could be addressed by a system that allows them to discuss hypotheses in the absence of data. In the remainder of this paper, we discuss a prototype which uses sketching as a strategy for supporting this conversation. We further focus our design on sketching time series data, as we observed from the interviews that much of the data that analysts worked with was time series.

3. Related work

In the context of information visualisation, sketching has been explored in several ways. First, sketching can refer to annotating: writing on, or marking up, a visualisation [WLJ*12]. For example, SketchVis [BLC*11] allows users to annotate a visualisation through hand-drawn sketches. Secondly, sketching can describe gesture-based interfaces used to create and modify visualisations [Cha10, KH15], or to query a larger dataset by sketching the required graph [Wat01] or defining visual constraints [RLL*05]. Finally, sketchy rendering styles can be used to represent uncertainty in a visualisation [BBIF12].

We use the term sketching in the last two senses: creating and modifying a visualisation, as well as suggesting that the data is uncertain. Our approach is also related to data generation tools (e.g. DBGen [dbg] and DBSynth [RDFS15]), which create data tables given a set of column descriptions. However, in our own prototype, it is direct manipulation of the visualisation itself that is the main focus, rather than generating further data.

4. Prototype for sketching data visualisations

We have built a prototype tool allowing data analysts to sketch a time series, and modify it in a conversation structured by analytic hypotheses. The goal is to bridge the gap between analysts and clients without a statistical background, by providing an interactive visualisation that helps the analyst share their domain knowledge and clarify analytic questions. It has been implemented as a web application, using the Dart language and SVG visual rendering.

4.1. The time series chart

The central area in Figure 1(A) represents a time series defined by a set of component functions. We draw on recent probabilistic modelling approaches in which a grammar of base kernels (covariance functions) is added and/or multiplied to model a complex dataset, automatically searching the space of composite kernels for one that best fits the data [DLG*13, LDG*14]. We reverse this principle to take a human-centric perspective. Instead of automatically searching for a set of kernels to describe a known dataset, we observe that if the user has an idea about expected features, synthetic data can be generated by composing base kernels.

Our motivation for this approach is twofold. First, users can interact with each kernel function independently, adjusting specific aspects of the time series. Second, constraining the interaction to a known set of functions allows users to focus on mathematical properties of the visualisation when drawing and discussing it. In contrast to sketching a line on paper, function composition adds hypothesis semantics to the visualisation. Modifying one function has no effect on the others. For example, a linear function with a positive slope might relate to an underlying linear trend. This semantic element should still be available for discussion or adjustment after it has been composed, for example, with a periodic function representing daily or weekly oscillation in the measured value. This allows the analyst to adjust features of the data without having to redraw the entire visualisation.

Through additive composition of the base kernels, the user visually constructs a function $f(x) = \sum_{k \in K} k(x)$, where K is the set of user-selected kernel functions. This function could be rendered precisely as a line graph. However, precise rendering suffers from the effect that users become fixated on manipulating the parameters of the kernel functions in order to achieve an exact trend. To avoid this problem, and to reinforce that the sketch is a provisional, transient conversational aid, we instead render a 'noisy' scatterplot of the function. We choose N equally-spaced x coordinates. For each x coordinate, we render a point on the scatterplot whose y-coordinate is chosen from the Gaussian distribution having $f(x)$ as its mean and a constant variance. Points are rendered as hand-drawn crosses to emphasise uncertainty [WII*12].

4.2. The tool panel

The left panel of Figure 1(B) presents functions and annotations that can be dragged onto the time series. The functions were selected from the grammar of kernels used by the Automatic Statistician [DLG*13, LDG*14]. The annotations were selected from the list of sketch-marks compiled by Eppler and Pfister [EP10]. These are intended to support informal recording of issues to be addressed in later data analysis, and also to externalize implicit expectations for the data [CGS*05].

4.3. Function editor

Every kernel function can be modified independently through two types of controls (Figure 1(C)). Two controls on the x axis (coloured in blue) specify the time range over which the function

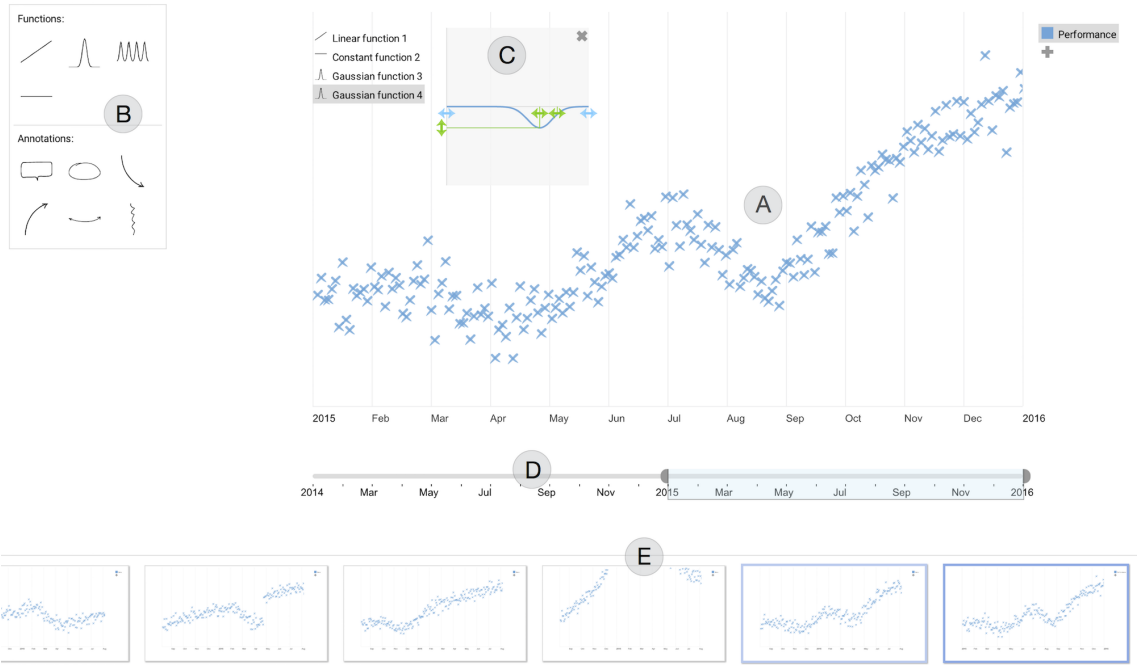

Figure 1: The SelfRaisingData prototype with the main components highlighted. (A) The time series chart with the fictional data points generated around the shape described through function composition, as presented in Section 4.1. (B) The tool panel containing functions and annotations (Section 4.2). (C) The function editor allows interactive modification of the mathematical parameters of the function and the time range for which it applies, as discussed in Section 4.3. (D) The time axis range selector (see Section 4.4). (E) Graphical history using a comic strip metaphor allows branching and visualising previous states (see Section 4.5).

(a) (b) (c)

Figure 2: The controls for the squared exponential function (or Gaussian function), with the range controls shown in blue, and the function parametric controls in green. (a) shows the function with a small width. (b) and (c) are the result of dragging the control for the width towards the right, resulting in a wider bell-shaped curve.

applies. Further, each type of kernel also has specific mathematical parameters. For example, for the squared exponential function, the controls are the height, the centre and the width (see Figure 2).

4.4. Time axis

Adjustment of the time axis is done by selecting the desired interval from the range underneath the chart Figure 1(D). This interaction has been used in a variety of applications that deal with range selection [mov, imo, tab, pub].

4.5. History

The history at the bottom of the display Figure 1(E) provides a graphical record of the analysis conversation using the comic strip metaphor [KF92, NC14, ZBER15]. We took into account previous work on interactive histories in information visualisation systems [HMSA08], for example supporting branches.

4.6. Multiple users

In order to support clarification conversations during remote collaboration as described in Section 2.1, we provide two views of the tool. The analyst view allows full editing functionality of kernel functions and their parameters, while the requester view supports visual annotations for feedback to the analyst. The two views are accessed via different URLs, implemented by transferring changed DOM elements through an App Engine Memcache instance [mem].

5. Usability evaluation study

We conducted two user studies to evaluate the tool, with a design iteration between the two studies. We used mailing lists and social media to recruit 6 participants for each study, selected to have technical background representative of data analysts. All participants had undergraduate level knowledge of mathematics, some experience with statistical tools like Excel, R or Matlab, and an academic background in Computer Science. Ten participants were enrolled as graduate students, one was a recent graduate, while one had more than 10 years of experience in professional visual analytics.

We began each session by describing the tool and the context in which it can be used, after which we let the participants explore the tool for a few minutes. The participants were then given a set of 4 tasks, each describing a visualisation to be sketched, for example:

> Create a visualisation for the traffic load of a broadband line from the beginning of 2014 until the end of 2015. The traffic load is constant for the first half of 2014, but afterwards it slowly increases linearly. Also, in the repair period January – April 2015 there was little traffic load.

Participants were asked to think aloud as they were performing the tasks, followed by an interview addressing their overall experience, as well as individual components of the interface. The session lasted 30-90 minutes.

Results from the first study showed that the interface was easy to learn and that function composition was considered *"intuitive"* (P3). All participants finished all the tasks and were able to understand what each function does, as well as how each parameter control changes the underlying function. The response to the ability to edit each function independently was positive and they found it *"easier to work with [and] to understand"* (P6). However, some participants mentioned that when composing many functions it can be difficult to see the individual components in the final visualisation. In such cases, they also found adjusting the time range of a function somewhat difficult, as the feedback from the movement of the data points wasn't sufficient to accurately indicate whether the desired position on the x axis has been reached. The participants found working in the function editor box somewhat *"fiddly"* (P2), because they had to *"judge what things are, based on what [they] see"* (P3). This led us to add slightly tighter coupling between the function editor and the resulting visualisation, through a vertical guideline on the chart when adjusting horizontal controls (see Figure 3).

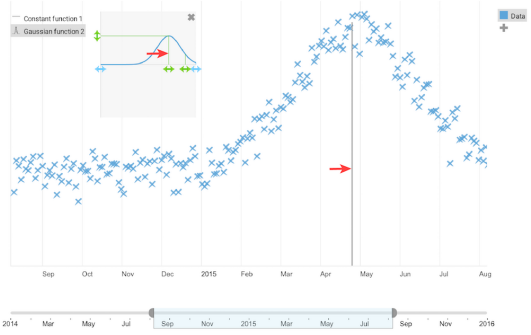

Figure 3: A grey vertical guideline is shown when dragging one of the x axis controls of a function.

After fixing these issues, we ran the second study with a new set of 6 participants. As in the first study, all participants were able to use the tool quickly, describing it as *"easy to use"* (P11) and *"fairly intuitive"*(P12). They understood the function composition principles (*"I liked being able to separate out the functions"* - P8) and mentioned that *"being able to manipulate the dataset - that is a really useful concept of the tool"* (P12).

Whilst most participants understood that the goal of the system is to sketch time series data (e.g. to "show some patterns by drawing" - P10), some participants still wanted the tool to be more precise, e.g.: *"I think the data points, being kinda farther away from the line served as a bit more of a distraction. I feel like it would be handy to be able to manipulate the lines more precisely."* (P8) A few participants suggested that being able to adjust the variance of the data point distribution would be useful, either as a way to show the degree of accuracy of the function composition, or as a way to make the visualisation more realistic, since different datasets have different variations. These observations reflect our initial goal - to create a tool that expresses sketchiness and imprecision.

Moreover, the participant with extensive experience in visual analytics commented on the usefulness of the tool in his work:

> *"One of the biggest challenges I have in communicating to a client is to create dummy data to try and describe what I'm expecting to see, so actually being able to manipulate the dataset like that is a really useful concept."* (P12)

This participant, who had no previous connection with the analysts we originally interviewed at BT, confirmed that time series data is also very frequent in his company's work, and that the beginning of each analytical project is similarly focused around clarifying analytical hypotheses with their clients, often using mock visualisations. These comments are similar to the findings from the initial interviews and they suggest that the system could be used during the clarification call. However, a more rigorous evaluation focused on the collaborative context and hypothesis refining is necessary in order to evaluate the usefulness of the tool in a real-world context.

The qualitative think aloud method used in these studies focused on user understanding, rather than task performance speed, and participants often paused whilst working on a task in order to offer additional explanations or suggestions. However, we note that participants finished each task quickly, usually in a few minutes, indicating that the tool could potentially be applied in its current form as a quick sketching tool.

6. Conclusion and future work

This project set out to investigate ways of improving conversations between data analysts and their customers. Informed by challenges that we observed in the workflow of analysts, and using the principle that time series data can be modelled with composite kernels, we designed a prototype to support sketching of visualisations through interactive function composition. Our tool is intended to be used in a collaborative manner, for defining hypotheses and clarifying questions, before data is available for analysis.

One limitation of the current work is that we only evaluated its usability properties. Whilst the results of both user studies were encouraging, further development would benefit from more discussions and studies of data analysts. We are looking at testing the system in a business environment and studying how it would be used in collaborative analytics practice. Another limitation of the current system is its focus on visualising time series as scatter charts, a decision driven by our analyst interviews. The system could be extended to support other chart types which can be sketched using the function composition strategy, such as bar and line charts.

Future evaluation could include a comparison between our prototype and existing tools that allow creating synthetic data. A different research avenue would be better understanding the boundary between precision and sketchiness, as raised by our user studies.

Acknowledgements

We would like to thank the BT data analysts for their time and very informative conversations, and our user study participants for their time and feedback.

References

[BBIF12] BOUKHELIFA N., BEZERIANOS A., ISENBERG T., FEKETE J. D.: Evaluating Sketchiness as a Visual Variable for the Depiction of Qualitative Uncertainty. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2769–2778. URL: https://hal.inria.fr/hal-00717441/document, doi:10.1109/TVCG.2012.220. 2

[BC06] BRAUN V., CLARKE V.: Using thematic analysis in psychology. *Qualitative Research in Psychology 3* (2006), 77–101. URL: http://eprints.uwe.ac.uk/11735/1/thematic{_}analysis{_}revised{_}{%}2D{_}final.doc, doi:10.1191/1478088706qp063oa. 1

[BLC*11] BROWNE J., LEE B., CARPENDALE S., RICHE N., SHERWOOD T.: Data analysis on interactive whiteboards through sketch-based interaction. *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces* (2011), 154—-157. URL: http://dl.acm.org/citation.cfm?id=2076383. 2

[CGS*05] CONVERTINO G., GANOE C. H., SCHAFER W. A., YOST B., CARROLL J. M.: A multiple view approach to support common ground in distributed and synchronous geo-collaboration. *Proceedings - Third International Conference on Coordinated and Multiple Views in Exploratory Visualization, CMV 2005 2005* (2005), 121–132. doi:10.1109/CMV.2005.2. 2

[Cha10] CHAO W.: NapkinVis: Rapid Pen-Centric Authoring of Improvisational Visualizations. *IEEE Infovis* (2010). URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.156.2717{&}rep=rep1{&}type=pdf. 2

[dbg] DBGen. https://github.com/electrum/tpch-dbgen. Accessed: 10.02.2016. 2

[DLG*13] DUVENAUD D., LLOYD J., GROSSE R., TENENBAUM J., GHAHRAMANI Z.: Structure discovery in nonparametric regression through compositional kernel search. *Proceedings of the International Conference on Machine Learning (ICML) 30* (2013), 1166–1174. URL: http://arxiv.org/abs/1302.4922, arXiv:arXiv:1302.4922v4. 1, 2

[EP10] EPPLER M. J., PFISTER R. A.: Drawing conclusions: Supporting decision making through collaborative graphic annotations. *Proceedings of the International Conference on Information Visualisation*, 1 (2010), 369–374. doi:10.1109/IV.2010.98. 2

[HA07] HEER J., AGRAWALA M.: Design considerations for collaborative visual analytics. *VAST IEEE Symposium on Visual Analytics Science and Technology 2007, Proceedings*, December 2007 (2007), 171–178. doi:10.1109/VAST.2007.4389011. 1

[HMSA08] HEER J., MACKINLAY J. D., STOLTE C., AGRAWALA M.: Graphical histories for visualization: Supporting analysis, communication, and evaluation. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1189–1196. doi:10.1109/TVCG.2008.137. 3

[imo] iMovie for Mac. http://www.apple.com/uk/mac/imovie/. Accessed: 10.02.2016. 3

[KF92] KURLANDER D., FEINER S.: A history-based macro by example system. *Proceedings of the 5th annual ACM symposium on User interface software and technology UIST 92*, November (1992), 99–106. URL: http://portal.acm.org/citation.cfm?doid=142621.142633, doi:10.1145/142621.142633. 3

[KH15] KIM Y.-S., HULLMAN J.: User-driven expectation visualization: Opportunities for personalized feedback. In *Electronic proceedings of the IEEE VIS 2015 workshop "Personal Visualization: Exploring Data in Everyday Life"* (2015), IEEE. 2

[KPHH12] KANDEL S., PAEPCKE A., HELLERSTEIN J. M., HEER J.: Enterprise data analysis and visualization: An interview study. *Visualization and Computer Graphics, IEEE Transactions on 18*, October (2012), 2917–2926. doi:10.1109/TVCG.2012.219. 1

[LDG*14] LLOYD J. R., DUVENAUD D., GROSSE R., TENENBAUM J. B., GHAHRAMANI Z.: Automatic Construction and Natural-Language Description of Nonparametric Regression Models. In *Association for the Advancement of Artificial Intelligence (AAAI)* (2014), pp. 1–11. arXiv:arXiv:1402.4304v2. 2

[mem] Google App Engine Memcache. https://cloud.google.com/appengine/docs/python/memcache/. Accessed: 10.02.2016. 3

[mov] Movie Maker. http://windows.microsoft.com/en-gb/windows/movie-maker. Accessed: 10.02.2016. 3

[NC14] NANCEL M., COCKBURN A.: Causality - A Conceptual Model of Interaction History. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (2014), 1777–1786. URL: http://dl.acm.org/citation.cfm?id=2556288.2556990, doi:10.1145/2556288.2556990. 3

[PC05] PIROLLI P., CARD S.: The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *Proceedings of International Conference on Intelligence Analysis 2005* (2005), 2–4. 1

[pub] Google Public Data. https://www.google.com/publicdata/directory. Accessed: 10.02.2016. 3

[RDFS15] RABL T., DANISCH M., FRANK M., SCHINDLER S.: Just can't get enough - Synthesizing Big Data. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (2015), ACM, pp. 1457–1462. 2

[RLL*05] RYALL K., LESH N., LANNING T., LEIGH D., MIYASHITA H., MAKINO S.: Querylines: approximate query for visual browsing. *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (2005), 1765—-1768. URL: http://doi.acm.org/10.1145/1056808.1057017, doi:10.1145/1056808.1057017. 2

[SMM12] SEDLMAIR M., MEYER M., MUNZNER T.: Design study methodology: Reflections from the trenches and the stacks. *IEEE transactions on visualization and computer graphics 18*, 12 (2012), 2431–2440. 1

[tab] Tableau. http://www.tableau.com/. Accessed: 10.02.2016. 3

[Wat01] WATTENBERG M.: Sketching a Graph to Query a Time-series Database. *CHI '01 Extended Abstracts on Human Factors in Computing Systems* (2001), 381—-382. URL: http://doi.acm.org/10.1145/634067.634292, doi:10.1145/634067.634292. 2

[WII*12] WOOD J., ISENBERG P., ISENBERG T., DYKES J., BOUKHELIFA N., SLINGSBY A.: Sketchy rendering for information visualization. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2749–2758. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6327281, doi:10.1109/TVCG.2012.262. 2

[WLJ*12] WALNY J., LEE B., JOHNS P., HENRY RICHE N., CARPENDALE S.: Understanding pen and touch interaction for data exploration on interactive whiteboards. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2779–2788. 2

[ZBER15] ZHAO Z., BENJAMIN W., ELMQVIST N., RAMANI K.: Sketcholution: Interaction histories for sketching. *International Journal of Human Computer Studies 82*, 301 (2015), 11–20. doi:10.1016/j.ijhcs.2015.04.003. 3