WordStream: Interactive Visualization for Topic Evolution

T. Dang, H.N. Nguyen, and V. Pham

Texas Tech University, Lubbock, USA

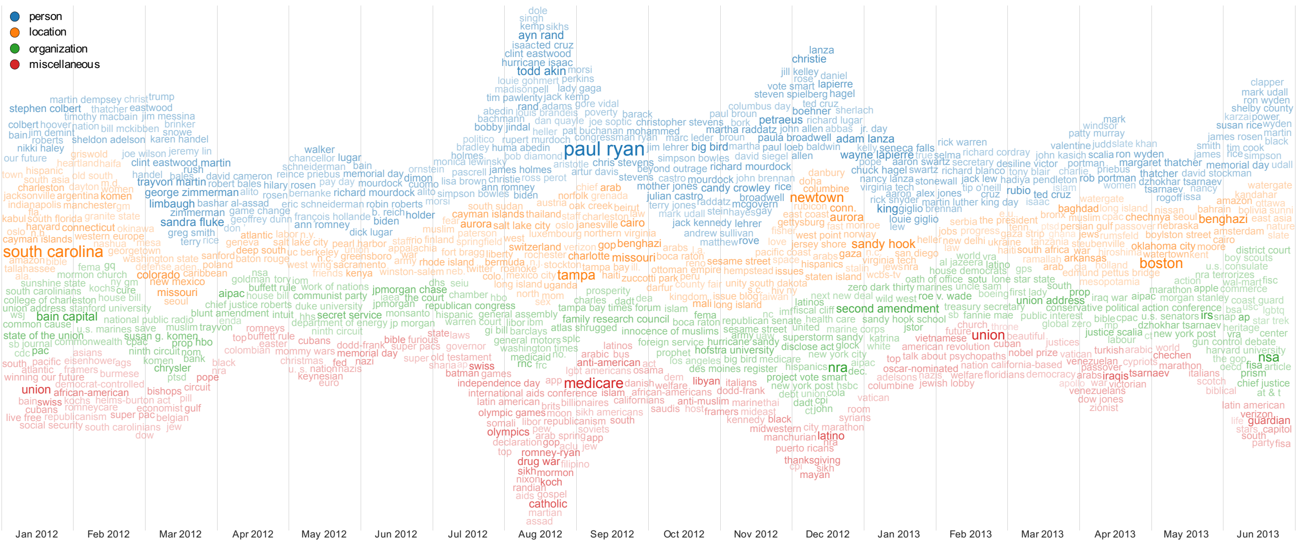

Figure 1: *WordStream visualization for the Huffington Post data, from January 2012 to June 2013. Terms are color-coded by category.*

Abstract

This paper introduces WordStream, an interactive visual tool for the demonstration of topic evolution. Our approach utilizes the two popular techniques. Word clouds are designed to give an engaging visualization of text via font sizes and colors, while stacked graphs are a common method for visualizing topic evolution. In particular, WordStream emphasizes essential terms chronologically and spatially. To show the usefulness of WordStream, we demonstrate its applications on various data sets, including the Huffington Post and IEEE VIS publications.

1. Introduction

The illustration of topic evolution has a long history. In 1931, *HistoMap* [Spa31] was created by John Spark, showcasing the power of civilizations along four thousand years of world history. In more recent efforts, *ThemeRiver* [HHN00] and, later, *StreamGraph* [BW08] expand the idea of the stacked graph to convey the evolution of topic [DGWC10]. Limited screen display and a large number of layers lead to the small area allocated for each term; therefore, the task of fitting terms into topic streams becomes more challenging. *Word clouds* [VWF09] are designed for optimizing space usage [Fei10]. However, temporal information has not been considered properly in this type of text presentation.

The combination of word cloud and stacked graph models has been recently studied [SWL*14]. However, there is still room for improvement and optimization, especially when the topic streams highly fluctuate. In this paper, we introduce a hybrid visualization to fill this gap. Our contributions in this paper are:

- We propose a synthesized approach to visualize information by means of word cloud within a stream layer.
- We implement an interactive text visualization prototype, named *WordStream*, to represent the evolution of topics to convey both spatial and temporal information.
- We evaluate the usefulness of the *WordStream* on various data sets (e.g., political blogs and other application domains).

DOI: 10.2312/evs.20191178

2. Related Work

2.1. Word cloud models

Wordle [VWF09] uses a randomized greedy algorithm to place words. It is greedy since it prioritizes the more frequent words. In addition, Wordle is aesthetically and visually appealing [Fei10]. In the past few years, there many efforts to optimize the *Wordle* layout. *ManiWordle* [KLKS10] provides more flexible control over how to form the word layout with the interaction from the users. *Rolled-out Wordles* [SSS*12] makes use of Linear Sorting (*RWordle-L*) and Concentric Sorting (*RWordle-C*) to place the words more compactly and preserve the orthogonal ordering and topology. *Wordle-Plus* [JLS15] extends the idea of *ManiWordle* to provide some further natural interaction supported for pen- and touch-enabled tablets while controlling the overall *Wordle* layout such as resizing, adding, deleting elements. A recent work, called *EdWordle* [WCB*18], allows editing and preserving the word cloud layout. These work mostly focus on extending/optimizing the *Wordle* layout and discard the time element.

There are attempts to integrate temporal constraints into *Wordle*. *Parallel Tag Clouds* [CVW09] utilize the parallel coordinates to represent time constraint. At each time step terms are placed in alphabetical order or order of importance, based on term frequency. This technique also has a feature that is implemented in *WordStream*, which is to display a stream when a term is selected. However, *Parallel Tag Clouds* is not space-efficient and cannot show the topic or overall evolution across time.

2.2. Time Series Visualizations

Stacked graph [Har00] has a baseline representing time constraint, and each layer serves as a topic of interest. The layers are stacked on one another starting from the baseline, and the change of the width of each layer represents the evolution over time. *ThemeRiver* [HHN00] is an optimization of the stacked graph, which provides a symmetric layout by setting the baseline at the center of the overall graph which then helps to smooth the transition across time. *StreamGraph* [BW08], an evolution of *ThemeRiver*, focuses on minimizing the wiggle per layer, which is sum of squares of the slopes at each time point, to avoid legibility issue in the previous stacked graph and *ThemeRiver* versions, and also improve the aesthetic aspect of the overall graph. All these versions of the stacked graph have a common issue: the limitation in layer width when the number of layers increases [CSYP18]. With limited space for the stream layer, the process of locating terms of the topic within the stream layer [Wat05] may encounter difficulties. Therefore, the stream layer contains typically only a few or no terms. This makes it difficult to view the evolution at a finer granularity [WSK*13]. Our *WordStream* places terms as close to its *time step* as possible by utilizing space sharing approach between adjacent time steps.

2.3. Combined models

TIARA [LZP*12] utilized the combination of the word cloud and stacked graph to demonstrate visual representation from abstract and complex text summarization. The *TIARA* visualization includes keyword clouds embedded in the layers of a stacked graph, whose layers depict the different topics. However, TIARA only shows a few word clouds into the stream boxes where space is sufficient. Therefore, maximizing the space usage within the stream layer was not the priority of this technique. *TextFlow* [CLT*11] demonstrates evolution and relationships among topics and their critical events: birth, death, split, merge. A word cloud at a particular timestamp can be only displayed on request. Our *WordStream* embeds terms directly into the stream layer. By using the space sharing technique between consecutive word cloud and color-encoding for terms in the same layer, we present the global evolution of topic streams using their text elements.

3. Design Decisions for *WordStream* Visualization

This section presents the goals for designing *WordStream* and decisions made during the implementation of *WordStream* to meet these goals. The aim of *WordStream* is to communicate the global patterns of the text corpus across time. The design goals are largely drawn from the related work reviewed in Section 2.

- **G1**. Display the evolution of stream topics [CLT*11].
- **G2**. Emphasize important terms in their corresponding topic layers at their corresponding time steps [PD18].
- **G3**. Maximize the space usage and place as many terms within each topic stream layer as possible [FFB18, WSK*13].

The following decisions were made during the implementation of *WordStream*:

- **D1**: Apply *StreamGraph* to represent the topic evolution (**G1**). The time direction is from left to right [DW13, DAW13] while the height of each stream layer at a time step is relative to the total term frequencies [LZT09].
- **D2**: Utilize a word cloud algorithm to display terms from each topic within its corresponding stream layer. This helps to satisfy **G2** [VWF09, Fei10].
- **D3**: Place each term into its corresponding horizontal position according to its timestamp [CVW09]. However, this constraint is loosened to *as close as possible to its timestamp* so some terms may be placed outside of their time boxes to meet **G3**.
- **D4**: Arrange texts with regard to their stream flows. Each word may have its own orientation (not always horizontal) to form the stream flows (**G1**).

Figure 2 shows a schematic overview of our approach. Topics are color-coded by using *displaCy Named Entity Visualizer* [Mon17].

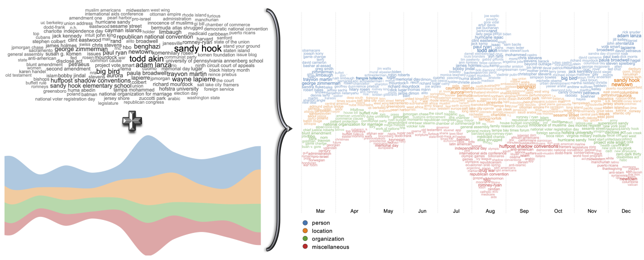

Figure 2: *A standard word cloud and a stream graph are synthesized into a single snapshot (on the right).*

On the control panel, users can customize the *WordStream* layout by adjusting the visual settings such as the font scale, the maximum number of words in each box, and the dimensions of the overall layout. Regarding term orientation, we provide the following two options.

- **Flow**: The orientation of the words corresponds to its streams orientation at the time step that they belong to.
- **Angle variance**: The angle of rotation of the words varies within a fixed range with regard to the medium line of the stream flow.

4. Computing *WordStream* Visualization

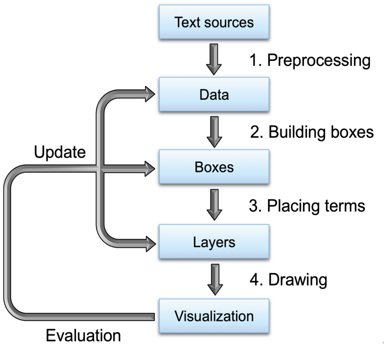

Figure 3: *The main components of our WordStream visualization: Preprocessing data, building boxes, placing terms, and drawing.*

Preprocessing data: The input text documents are preprocessed into entities and further classified into different categories [Mon17]. In many cases, the term frequency might not convey user interests [DPF16]. For example, the term "Obama" repeated numerous times in political blogs and news might not draw a lot of attention or interest [SMR08]. To focus on the more significant terms, we use the *sudden attention* measure, referring to a sharp increase in frequency [DN18]. Let $F_1, F_2, \ldots, F_n$ be the frequency of an entity at n different time points. The sudden attention series $(S_1, S_2, \ldots, S_n)$ is computed by $S_t = \frac{(F_t+1)}{(F_{t-1}+1)}$.

Building boxes: As for **D1** and **D3**, our approach places terms inside their corresponding stream and close to their time steps. Invisible boxes are created for this purpose as depicted in Figure 4. *WordStream* scans along a spiral pattern centered at the box to find the first available space to place terms.

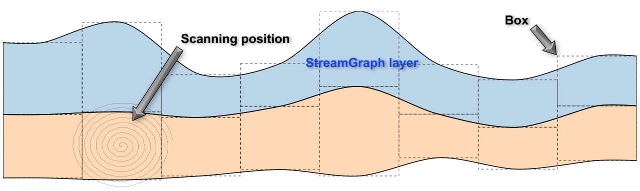

Figure 4: *Building boxes and placing terms into stream layers.*

Placing terms: *Mask-based, pixel-perfect* collision detection algorithm is used to detect collisions. The *Mask-based* algorithm uses a board to represent the stream layer with all the terms that have been placed on the board at the checking time. This board is used to check for collision against a new term. In the *pixel-perfect* approach, the terms and the board are represented in terms of pixels. To reduce the memory space, only the red component (instead of all red, green, blue, and alpha components) is used to represent a pixel; this data is stored into a variable called *sprite* of the board or the term. The *sprite* value of a pixel i is computed from the pixel data using the following formula: $sprite[i] = pixels[i << 2]$. The collision detection checks the positions of all pixels on the sprite of the term and the corresponding positions on the board. Similarly, the placement of the new term onto the current board adds these values from the sprite of the term to the corresponding positions of the sprite of the board. As depicted in Figure 4, this spiral starts at the center of each box as calculated for its corresponding time step, and its maximum deviation from the center (*dx, dy*) is smaller than the diagonal of the box.

Drawing: The filtered terms are sequentially located to formulate the stream layers. Notice that the layers can be ordered vertical for quantitative data, such as in increasing order of security levels or user ratings. In addition, interactive features are supported to highlight individual term evolution (see Figure 7).

5. Evaluation

5.1. Quantitative Evaluation

For each test data set, the combinations of two options (**flow** and **angle variance**) are evaluated on the *Compactness* metric as depicted in Figure 5. *Compactness* is defined by the area of all displayed words divided by area of the stream [BKP14], indicating the level of coverage over the stream layer. *Compactness* has the range from 0 to 1; a higher value means the words are closer to forming the full shape of the stream layers.

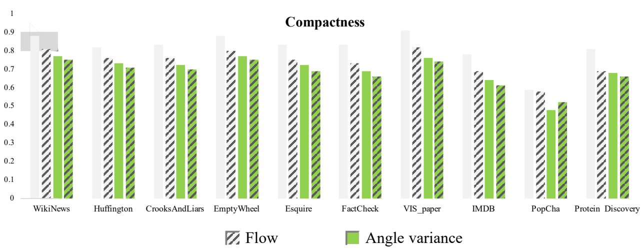

Figure 5: *Comparisons of the Compactness measure for 10 test data sets (from left to right).*

As seen in Figure 5, surprisingly, the combination that yields the best result is **disabling both features**. One example of this arrangement can be seen at the top panel of Figure 6. The conflicts in placing terms can be explained: If the variation is allowed, besides the conflicts from each sprite to one another, there are also conflicts from constraints in direction. In contrast, the combination of enabling both options produces the lowest *Compactness* scores on most test data sets.

WordStream is implemented using D3.js [BOH11]. The demo video, online prototype, and more examples can be found at the Github page `https://idatavisualizationlab.github.io/WordStream/`.

5.2. Exploring IEEE VIS author contribution

The *IEEE Visualization Publications* data set [IHK*17] contains 2,867 entries with attributes such as *Conference* (*InfoVis*, *VAST*, and *SciVis*), *Year* (from 1990 to 2016), *Paper Title*, *Link*, and *Author names*. This data set is particularly useful for finding appropriate authors for reviewing paper/proposal submissions and exploring the contributions of researchers over time. Figure 6 presents popular IEEE VIS authors over a period of 10 years. From top down, we show the different generated layouts by combining the two options: **Flow** and **Angle variance**. As depicted, the top panel is the most *compact* layout as it can fit more author names than the others. This confirms our observation in the previous section. Moreover, the top panel also achieves the best readability as all terms are horizontal.

Figure 6: *Popular IEEE VIS authors over 10 years from 2007 to 2016: author names are colored by their first publication venue.*

5.3. Informal User Study

We conducted informal user studies to gather qualitative responses about *WordStream* from two experts, one researcher in political science and one professor in data science. The study began with a brief description of *WordStream* to familiarize them with the usage of the analytic tool. We also adapted the implementation of *TimeArcs* [DPF16] for the same political blogs as a reference. Then the experts were free to explore the visual interfaces for a specific task: *What are the top political events in the past ten years*. Both of them agreed that, in comparison to *TimeArcs*, the *WordStream* is useful to convey the global trend and can be applied to visualize the emerging topics in various domains.

For the overall presentation, both of the users commented that they can quickly understand the idea of the layout. The data science professor stated that he had known word cloud before, and *WordStream* "allows you to do the longitudinal analysis easily". He chose a term and scrolled through the entire timeline to see the fluctuation of its occurrences as depicted in Figure 7. Hence, the visualization is helpful in an exploratory analysis. However, he commented that the layout might be cluttered when the number of layers increases to about ten. On the other hand, the political expert at first found it "intimidating," but shortly after the description, he found the interface easy to use. Furthermore, he found the brushing and linking are efficient for highlighting the temporal patterns of terms, along with supporting content analysis.

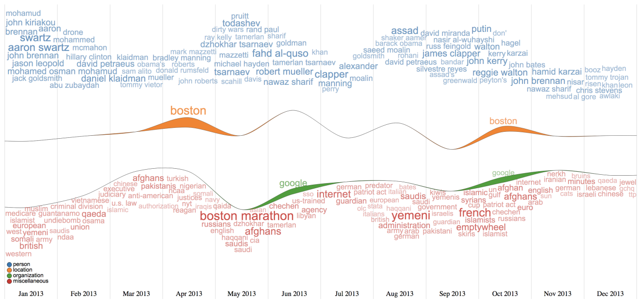

Figure 7: *Term selection in our WordStream layout: boston (location) and google (organization).*

Besides the positive feedback, the experts also mentioned some limitations of *WordStream*, in which the related words are not shown in clusters, unlike *TimeArcs*. One of them suggested that the relationships being drawn explicitly among terms would be more useful than the proximity of terms as in the current visualization.

6. Conclusion

This paper presents a hybrid text visualization technique. *WordStream* aims to communicate the global trends of the underlying topic evolution while preserving the presentation-oriented criteria of the visualization solution. We demonstrate the applications on various data sets, showing that *WordStream* could quickly highlight important terms and could assist users in exploring term evolution at a finer granularity. Future work will focus on algorithms to cluster the terms within and across topic streams. Also, more interactive features should be supported to optimize *WordStream* layout.

References

[BKP14] BARTH L., KOBOUROV S. G., PUPYREV S.: Experimental comparison of semantic word clouds. In *Experimental Algorithms* (Cham, 2014), Gudmundsson J., Katajainen J., (Eds.), Springer International Publishing, pp. 247–258. 3

[BOH11] BOSTOCK M., OGIEVETSKY V., HEER J.: D^3 data-driven documents. *IEEE Transactions on Visualization & Computer Graphics*, 12 (2011), 2301–2309. 3

[BW08] BYRON L., WATTENBERG M.: Stacked Graphs - Geometry & Aesthetics. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1245–1252. 1, 2

[CLT*11] CUI W., LIU S., TAN L., SHI C., SONG Y., GAO Z., QU H., TONG X.: TextFlow: Towards better understanding of evolving topics in text. *IEEE Trans. Vis. Comput. Graph. 17*, 12 (2011), 2412–2421. 2

[CSYP18] CUENCA E., SALLABERRY A., YING WANG F., PONCELET P.: MultiStream: A Multiresolution Streamgraph Approach to Explore Hierarchical Time Series. *IEEE Transactions on Visualization and Computer Graphics* (2018). 2

[CVW09] COLLINS C., VIÉGAS F. B., WATTENBERG M.: Parallel tag clouds to explore and analyze faceted text corpora. *Proceedings of the VAST 09 - IEEE Symposium on Visual Analytics Science and Technology* (2009), 91–98. 2

[DAW13] DANG T. N., ANAND A., WILKINSON L.: TimeSeer: Scagnostics for high-dimensional time series. *IEEE Trans. Vis. Comput. Graph. 19*, 3 (March 2013), 470–483. 2

[DGWC10] DÖRK M., GRUEN D., WILLIAMSON C., CARPENDALE S.: A Visual backchannel for large-scale events. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1129–1138. 1

[DN18] DANG T., NGUYEN V. T.: ComModeler: Topic Modeling Using Community Detection. In *EuroVis Workshop on Visual Analytics (EuroVA)* (2018), Tominski C., von Landesberger T., (Eds.), The Eurographics Association. 3

[DPF16] DANG T. N., PENDAR N., FORBES A. G.: Timearcs: Visualizing fluctuations in dynamic networks. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 61–69. 3, 4

[DW13] DANG T., WILKINSON L.: TimeExplorer: Similarity search time series by their signatures. In *Proc. International Symp. on Visual Computing* (2013), pp. 280–289. 2

[Fei10] FEINBERG J.: Wordle. *Beautiful Visualization: Looking at data through the eyes of experts.* (2010), 37–58. 1, 2

[FFB18] FELIX C., FRANCONERI S., BERTINI E.: Taking word clouds apart: An empirical investigation of the design space for keyword summaries. *IEEE transactions on visualization and computer graphics 24*, 1 (2018), 657–666. 2

[Har00] HARRIS R. L.: *Information Graphics: A Comprehensive Illustrated Reference*. Oxford University Press, 2000. 2

[HHN00] HAVRE S., HETZLER B., NOWELL L.: ThemeRiver: visualizing theme changes over time. *Proceedings of IEEE Symposium on Information Visualization 2000. INFOVIS 2000 2000* (2000), 115–123. 1, 2

[IHK*17] ISENBERG P., HEIMERL F., KOCH S., ISENBERG T., XU P., STOLPER C., SEDLMAIR M., CHEN J., MÖLLER T., STASKO J.: vispubdata.org: A metadata collection about IEEE visualization (VIS) publications. *IEEE Transactions on Visualization and Computer Graphics 23*, 9 (Sept. 2017), 2199–2206. 4

[JLS15] JO J., LEE B., SEO J.: WordlePlus: Expanding Wordle's Use through Natural Interaction and Animation. *IEEE Computer Graphics and Applications 35*, 6 (2015), 20–28. 2

[KLKS10] KOH K., LEE B., KIM B., SEO J.: ManiWordle: Providing flexible control over Wordle. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1190–1197. 2

[LZP*12] LIU S., ZHOU M. X., PAN S., SONG Y., QIAN W., CAI W., LIAN X.: TIARA: interactive, topic-based visual text summarization and analysis. *ACM TIST 3*, 2 (2012), 25:1–25:28. 2

[LZT09] LOHMANN S., ZIEGLER J., TETZLAFF L.: Comparison of tag cloud layouts: Task-related performance and visual exploration. In *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part I* (Berlin, Heidelberg, 2009), INTERACT '09, Springer-Verlag, pp. 392–404. 2

[Mon17] MONTANI I.: An open-source named entity visualiser for the modern web, 2017. https://explosion.ai/blog/displacy-ent-named-entity-visualizer [Accessed date: April 10, 2019]. 2, 3

[PD18] PHAM V., DANG T.: Cvexplorer: Multidimensional visualization for common vulnerabilities and exposures. In *2018 IEEE International Conference on Big Data (Big Data)* (Dec 2018), pp. 1296–1301. 2

[SMR08] SCHÜTZE H., MANNING C. D., RAGHAVAN P.: *Introduction to information retrieval*, vol. 39. Cambridge University Press, 2008. 3

[Spa31] SPARK J.: Time chart of world history: A histomap of peoples and nations for 4,000 years, 1931. http://www.worldhistorycharts.com/the-histomap-by-john-sparks/ [Accessed date: Feb 20, 2019]. 1

[SSS*12] STROBELT H., SPICKER M., STOFFEL A., KEIM D., DEUSSEN O.: Rolled-out Wordles: A Heuristic Method for Overlap Removal of 2D Data Representatives. *Computer Graphics Forum 31*, 3pt3 (2012), 1135–1144. 2

[SWL*14] SUN G., WU Y., LIU S., PENG T., ZHU J. J. H., LIANG R.: Evoriver: Visual analysis of topic coopetition on social media. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec 2014), 1753–1762. 1

[VWF09] VIEGAS F. B., WATTENBERG M., FEINBERG J.: Participatory visualization with Wordle. *IEEE transactions on visualization and computer graphics 15*, 6 (2009). 1, 2

[Wat05] WATTENBERG M.: Baby names, visualization, and social data analysis. In *Proc. IEEE Symp. on Information Visualization* (2005), pp. 1–7. 2

[WCB*18] WANG Y., CHU X., BAO C., ZHU L., DEUSSEN O., CHEN B., SEDLMAIR M.: EdWordle: Consistency-Preserving Word Cloud Editing. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 647–656. 2

[WSK*13] WALDNER M., SCHRAMMEL J., KLEIN M., KRISTJÁNSDÓTTIR K., UNGER D., TSCHELIGI M.: Facetclouds: Exploring tag clouds for multi-dimensional data. In *Proceedings of Graphics Interface 2013* (Toronto, Ont., Canada, Canada, 2013), GI '13, Canadian Information Processing Society, pp. 17–24. 2