

A Scalable “Exploration” Technique for Hierarchically Indexed Table Data

Natsuki Hosokawa
hosokawa.n.ac@m.titech.ac.jp
School of Computing, Tokyo Institute
of Technology

Kohei Arimoto
khyarnt@gmail.com
Teikoku Databank and The University
of Tokyo

Ken Wakita
wakita@is.titech.ac.jp
Tokyo Institute of Technology

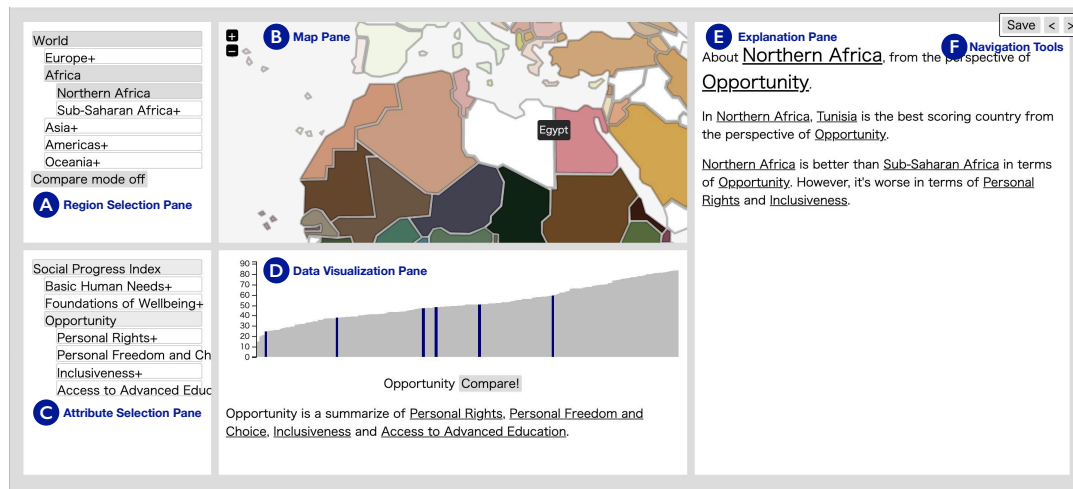


Figure 1: SPIViewer is an “explorative” VA tool for comparing UN member states using the Social Progress Index.[15]

ABSTRACT

Data analytics tools that combine automated text generation and visualization techniques suffer from scalability problems. The amount of the generated text explodes with the increase of items and attributes. This study addresses this problem for table data whose attributes and data items are hierarchically organized. In our approach, the user’s point of view is modeled by the dual focalization axis, which consists of the attribute-based and the data-item-based focal points. The user can refine the two-dimensional focal points to obtain the chart and the text that explain the data facts found in a more focused portion of the dataset. The efficacy of the proposal was assessed through a quantitative and qualitative evaluation using a prototype visual analytics tool that employs the idea.

ACM Reference Format:

Natsuki Hosokawa, Kohei Arimoto, and Ken Wakita. 20XX. A Scalable “Exploration” Technique for Hierarchically Indexed Table Data. In *VINCI ’20: The 13th International Symposium on Visual Information Communication and Interaction*, December 08–10, 2020, Eindhoven, The Netherlands. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/XXXXXXX.XXXXXXX>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

VINCI ’20, December 08–10, 2020, Eindhoven, The Netherlands

© 20XX Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/XXXXXXX.XXXXXXX>

1 INTRODUCTION

In recent years, we see a growing number of visual analytics systems employ the combination of visualization and text synthesis technologies [1, 7–10, 13, 16, 17]. One of the objectives of such systems is to achieve data exploration and data explanations at the same time. The data visualization technology encourages users’ free information-seeking behavior, while text synthesis technology summarizes the facts retrieved from the dataset using data mining techniques. This approach assists people understand the data by presenting visual and textual explanations of the statistical facts that are retrieved from the dataset via data mining techniques. In this way, such techniques can help users who are not trained in visual and statistical analysis to explore the data.

One of the challenges for “exploration (= exploration + explanation)” technology[13, 21] is the scalability. The number of attributes in the table datasets treated in the past research studies was limited to a few. The problem with the existing proposals is that the size of the generated text is expected to grow in a polynomial order of the number of attributes. For example, the synthesis result for the SPI dataset with 51 basic indicators comparing UN member states is a huge text consisting of 391,273 words.

The goal of this study is to realize an explorative system for large table data with many attributes. Specific challenges include, in addition to the already mentioned scalability, the ability to present an overview and focus on interesting regions, support for exploratory behaviors, identification and comparison of data items, and support for collaborative work.

We focused on the fact that the attributes and data items of large scale table data are often organized in hierarchies. In this research, we took the approach of data and text summarization based on these hierarchical structures. Specifically, the approach is based on the “two axes of focus”, one in the dimension of attributes and the other in data items. We also offer an interaction technique called “two-dimensional drill down,” which allows the user to shift the focal point of analysis along the two axes.

The efficacy of the proposal was assessed quantitatively and qualitatively using a prototype visual analytics tool for the Social Progress Index[15] called SPiViewer¹ system.

2 RELATED WORK

2.1 Automated insight systems and natural language processing

Automated insight systems employ data mining to extract important features called *data facts*. Tang and others [18] extract *k*-important data facts from data; Demiralp and others [3, 4] proposed a visualization technique of data facts. Automated insight systems provide the general user with an overview of large and complex data.

Microsoft Power BI [12] and Google Sheets [5] combine automated insight systems with natural language processing and information visualization. The method generates charts and a textual explanation of the data facts. Also, Quill [2], Wordsmith [11] are Web browser plugins that enhance natural language generation (NLG) capabilities on such systems.

2.2 Natural text generation

To compensate for the shortcoming of the difficulty of grasping the overview of the data through systems based on automated insight systems, Latif et al. have proposed VIS Author Profiles for browsing co-authorship relationships of researchers, the interactive Map Reports for two-attribute datasets representing geographic distributions, and a visual analysis system for data assessed for readability and maintainability with respect to program code [8, 9, 13]. These systems attempt to explain the whole data rather than individual data facts using template based natural language processing technique. There is deep learning based technique to generate high quality natural language text from a dataset[20], however it often generates statements with no factual basis.

2.3 Explorative Visual Analytics

The aim of this study is to build a visual analytics system that integrates data visualization and automatic sentence generation techniques for large, superficial datasets. White and Roth have pointed out eight features of an exploratory search system that is used to analyze large unknown datasets. They include unclear prior goal setting, continuous exploration over the long term, learning and comprehension, collaborative analysis activities, etc. [19]

3 SYSTEM

3.1 Design Considerations

Five considerations were explored in the design for this study.

DC1 (Scalability) The proposal automatically generates charts and text to illustrate the data to aid the visual exploration of the user. Because the amount of noticeable data facts increases combinatorially as the number of attributes increases, a naive application of conventional methods produces a huge amount of charts and explanatory texts as the volume of data increases. The former, i.e., the number of visual elements displayed, is serious but various studies have been conducted in the past, including sophisticated drawing methods, data clustering, multi-layered data, and the adoption of interaction. The latter, i.e., the generation of explanatory texts, however, is less studied.

DC2 (Overview and Focuses) Large-scale data analysis involves *overview tasks*, such as summarization, dissectioning, and discovering patterns. Additionally, it also involves *focused tasks* related to specific areas of the data where the analysts have a prior interest.

DC3 (Exploration) In an exploratory analysis of an unknown data, it is necessary to first grasp the overview and then, according to the findings obtained from the overview, to elaborate the analysis on the focused substructures[14].

DC4 (Identification and Comparison) Data analysis often requires a detailed *examination* of a data item at the focus. It is also necessary to *compare* items with similar or contrasting features.

DC5 (Collaboration and Presentation) The ability to *share* the generated sentences and visualizations with other people is desired. Such feature helps multiple collaborating explorers to break up the search process and share important information with each other.

3.2 System Overview

The data items and attributes of many large-scale table statistical data, are often hierarchically organized. In dealing with scalability (DC1), we employed the notion *dual focalization axis* in which the analyst sets the focus for each hierarchy of items and attributes, and avoids the explosion of text volume by generating concise text with the focus in the hierarchical structure as the context. To handle the dual focalization axis, the system’s hierarchical discovery function was designed to support the *two-dimensional drill-down* mechanism (DC2, DC3). To aid the analysis work, where an analyst’s interest shifts from one scope of data items and attribute set to another, the text and illustrations that describe the relevant scope is reflected in the display. The system provides the ability to compare pairs of data items and attributes (DC4). Finally, the ability to save and share the status of the analysis at any point in time is offered (DC5).

3.3 Extraction of Data Facts

Statistical features extracted from a dataset, such as extremums and outliers from the overall distribution, are called *data facts* [18]. The automated insight system generates sentences that explain the dataset by describing the *k*-most prominent data facts. In contrast, this study generates natural language texts based on three kinds of statistics: extremums and outliers, following [8].

For extraction of one-dimensional outliers, the range of the first and third quartiles is taken [6]. For the extraction of two-dimensional outliers, we select attribute pairs that has a sufficiently large (> 0.7) absolute values of correlation coefficients and the second PCA-component being a one-dimensional outlier.

¹SPiViewer: <http://smartnova.github.io/spi/>

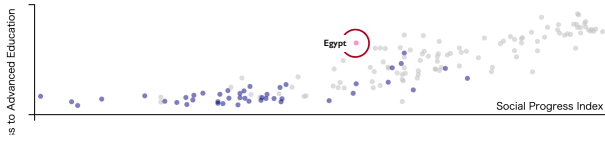


Figure 2: Regions and Attributes compared on the data visualization view (Northern Africa vs. Africa and Access to Advanced Education vs. Overall SPI)

3.4 Data Visualization

SPIViewer is a prototype system of the proposed idea. It is a visual analytics tool for the SPI data. In the map view (Figure 1-Ⓐ) it depicts the map of the regional focus. It illustrates the UN member states in the regional focus (DC1). The colors illustrate the their ratings in the three main SPI components, e.g, **Basic Human Needs**, **Foundations of Wellbeing**, and **Opportunity** (DC2). For example, from the characteristic reddish color given to Egypt, its higher level of **Basic Human Needs** is observed (Figure 1). The prominence of this country in comparison with others is recognized, too.

The bar chart displays the attribute values of all the member states and highlights bars of in-focus region (northern Africa; Figure 1-Ⓓ; DC4). When two attributes are compared, a scatterplot illustrates the data distribution. When two regions are compared simultaneously blue, pink, and gray colors are used to depict the 1st region, 2nd region, and the rest (DC4; Figure 2).

3.5 Text Generation

The system converts the extracted data facts, into textual explanation by a text template. According to the kind of the selected data items, the system employ three kinds of templates.

Case 1 (when a single data item is selected): The path starting from the topmost data item in the hierarchy down to the selected item is regarded as a series of gradually narrowing the dataset. For example, when Germany is selected, the path starts from the topmost element (all UN member states), European countries, Western European countries, and finally to Germany.

By comparing the selected item with a group of higher-level items along such path, the characteristics of the selected item can be described in the broader context. For example, we compare Germany with all UN member states, then with European countries, and finally with Western European countries, to generate a series of descriptions on Germany situated in different geographical scopes.

A simple application of this technique would generate redundant text. For example, Germany is a world leader in 10 indicators. Naturally, Germany also beats other European countries in these indicators. To avoid repeated generation of redundancy explanation due to those self-evident data facts, SPIViewer suppresses repeated references to the same attributes ².

Case 2 (when an element in the middle tier in the hierarchy is selected): Firstly, among the selected data items, one that outperforms others is introduced. For example, suppose (Southern Europe, Basic Human Needs) are selected for regional and attribute focuses, respectively. Here, the generated text introduces Portugal as it marks the highest Basic Human Needs value among Southern

Europe countries ³. Then the text continues with comparison of the selected element with its siblings. With the SPI example, the system conducts a comparison of Southern Europe with others (i.e., Northern-, Eastern-, and Western European countries, respectively).

Often the extracted data facts are related with many attributes. Enumerating them all may result in a long sentence. In this study, lower-level attributes are hidden but a toggle unhide/hide switch is offered for the user to see them all.

Case 3 (when two sets of data items (group A and group B) are selected): From the comparison between them, we can split the set of attributes into three groups: a set of attributes that group A (B) is noticeably better than B (A), and the set of other attributes. We provide five text templates according to the ratio of these three attribute groups to improve readability (DC1,2). The details are omitted due to the space limitation.

3.6 Interaction

SPIViewer is a Web-based VA tool for the SPI dataset (Figure 1). The *region selection pane* Ⓐ presents the regional hierarchy in a tree view (DC3). The *map pane* Ⓑ automatically zooms in/out on the selected region in response to the selection of the current region. The analyst can also click on and select a UN member state from the map pane. Selection of different geographical region updates the textual description in the *explanation pane* Ⓔ. Using “Compare mode” toggle switch in the region selection pane, the user can see and read the comparison results in the chart and the text (DC4).

The *attribute selection pane* Ⓒ provides a tree view similar to the region selection pane, which allows the analyst to browse the attribute hierarchy and refine the selection of the attributes. The SPI dataset consists of 51 statistical attributes and 15 aggregate attributes that group together statistics of lower-level attributes. The value distribution of the selected attribute for UN member states are depicted in a bar chart in the *data visualization pane* Ⓓ. Below the figure, a brief description of the respective statistics and a link to the source of the information are displayed. The data visualization pane offers similar comparison capabilities to the region selection pane. By using the toggle switch, the analyst can compare two different statistics in a scatterplot and see the correlation among them.

Tools are provided in the upper right corner to assist in the browsing process Ⓕ. The system records changes in two-dimensional focus for bookkeeping purpose. The “<” and “>” icons instruct the system to move backward and forward, respectively, along with the browsing history. The “Save” button saves the snapshot of the analysis by assignment of a separate URL to the two-dimensional focus settings. The user can duplicate the view by opening a new browser tab and use the URL for bookmarking and sharing/posting.







4 EVALUATION



4.1 Use Case


Suppose at the original globe view ⁴, we are interested in a northern African country with a characteristic reddish color. Clicking it, the two-dimensional focus moves to (the selected country, all attributes). As a result, the contents of the panes are updated to reflected the selection ⁵. With the help of the hover tool, we learn the selected country is Egypt and its SPI value is 61.71. The the


²Click on ⁶ and your Web browser shows the SPIViewer.

Table 1: The size of the explanation text generated for each selection of the member state and the SPI index

Rank	Member state	Index	#words
1	 Egypt	Social Progress Index	122
1,956	 Argentina	Access to Basic Knowledge	31
3,556	 Denmark	Biome protection	30
6,159	 Italy	Internet users/Population	22
9,629	 Ireland	Homicide rate/Population	21
12,261	 Greenland	Social Progress Index	11

explanation pane explains that “Egypt scores highest in Africa on **Access to Advanced Education**, while worst on **Outdoor air pollution attributable deaths**. Clicking the latter in the explanation pane, the two-dimensional focus shifts to (Egypt, Outdoor air pollution attributable deaths). The updated data visualization pane  shows that the severity of the situation for Egypt is outstanding in comparison with other countries. We use the “<” button to move backward, then click on **Access to Advanced Education** to find that Egypt ranks highest in Africa, and 39th worldwide . Using the comparison function, we see the indicator and the SPI overall index are linearly correlated but that Egypt stands out in this global trend in offering a superior level of higher education (Figure 2).

Here, our interest shifts to trends in Africa rather than just Egypt. Selecting Africa in the item selection pane, the two-dimensional focus moves to (Africa, Access to Advanced Education). The bar chart tells  general poor performance of African countries on this indicator but that a few excellent countries including Egypt.

From the above explorations, we learned that basic living conditions and advanced education are good in Egypt but it is challenged with respect to welfare, most seriously by air pollution. For the moment we go back to the point where this exploration started (Egypt, all attributes) and used the “Save” button to bookmark this exploration . We send this URL to a colleague and ask for further research, using our research result as a starting point.

4.2 Scalability

The SPI dataset was used to assess the scalability of the proposal. First, we naively enumerated meaningful data facts from the combination of all countries and attributes, then converted them all into a textual explanation, and measured its volume. The text consists of 367,476 words and is beyond quick comprehension for the human.

With our text reduction techniques, the size of the text generated is expected proportional to the product of the number of data facts and attributes. The longest text is expected to be generated from the combination of data items and attributes that has the largest number of child elements in the hierarchies. This means that the text size is independent from the height of the hierarchy.

The average number of sentences generated was 24 (Table 1), with the longest being (Egypt, SPI). 3,234 focuses resulted in the shortest sentences of length 11 words. Their significantly shorter lengths are due to missing data in the dataset.

4.3 Task-based Analysis

We conducted experiments using the Japanese translation of SPIViewer. All the participants are Japanese, computer science students and

knowledgeable of visual analytics. We performed a comparison between SPIViewer and Google Spreadsheets with eight tasks (<https://j.mp/3lbuL3Q>). The tasks starts from a simple identification and comparison task to more complex one: “Which of the two distributions of opportunities for social advancement scored higher overall, Africa or Asia?”. Participants were asked to learn from three tutorial videos (SPI data - 2 minutes 3 seconds, SPIViewer - 6 minutes 26 seconds, and Google Spreadsheet and its automated insight function - 5 minutes 36 seconds) before the experiment.

Average working time was 15.4 minutes on SPIViewer and 35.3 minutes on Google Spreadsheet. A strong correlation in the distribution of working time was observed: participants completed the tasks in half the time of using Google Spreadsheets. The experiment and the short interview session were recorded on video. A qualitative evaluation is now underway. In the interviews, the participants provided the following positive comments about SPIViewer: “The structure of the screen is easy to understand and operate - (X2, X3),” “SPIViewer was much easier than a spreadsheet because it does not require me a formula - (All),” “Spreadsheets are a pain in the ass to visualize,” and “SPIViewer is Fun! - (X4).” Some of the respondents asked for improvements in the software: “I want the ability to search for a country (All),” “I wish the text was more detailed - (X4, X7),” “I want the scatterplots to have the same aspect ratio - (X3, X6, X7),” and “I need a better color scheme - (X2).” More detailed qualitative analysis is undergoing.

5 DISCUSSION

The text generated by the proposal explains outstanding attributes for each data item. It is easy to understand the characteristics of the data items. However, comprehending attributes-based facts is difficult. This issue should be addressed by a text generation where attributes are explained by outstanding data items.

We have observed that certain interesting facts were overlooked by the outlier detection techniques employed in this research. For example, **Egypt’s outstanding color** among neighboring countries visually the human but the system failed to suggest it.

Mumtaz and others stress the importance of *consistent linking* across the triad of data visualization, text, and word-sized charts embedded in the text [13]. From their argument, the SPIViewer can be improved upon in a few directions.

6 SUMMARY

A growing number of data analytics system incorporates the combination of data visualization and automated document generation. One of the contributions of this work is the identification of challenges for such systems to deal with large scalable datasets. We have shown that the challenges can be solved by exploiting the hierarchical structure often found in large table datasets. The effectiveness of our method was demonstrated through SPIViewer. Future work includes more detailed user studies, application to other datasets, support for time series analysis, and provision of a richer user interface for exploratory analysis.

Acknowledgments The authors would like to acknowledge Teikoku Databank, Ltd’s Center for TDB Advanced Data Analysis and Modeling for providing financial support.

REFERENCES

- [1] C. Bryan, K. Ma, and J. Woodring. 2017. Temporal Summary Images: An Approach to Narrative Visualization via Interactive Annotation Generation and Placement. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (Jan 2017), 511–520. <https://doi.org/10.1109/TVCG.2016.2598876>
- [2] Jason Chen. [n.d.]. Quill - Your powerful rich text editor. <https://quilljs.com/>
- [3] Çağatay Demiralp, Peter Haas, Srinivasan Parthasarathy, and Tejaswini Pedapati. 2017. Foresight: Rapid Data Exploration Through Guideposts. In *Workshop on Data Systems for Interactive Analysis (DSIA) at IEEE VIS 2017* (Phoenix, AZ).
- [4] Çağatay Demiralp, Peter J. Haas, Srinivasan Parthasarathy, and Tejaswini Pedapati. 2017. Foresight: Recommending Visual Insights. *Proc. VLDB Endow.* 10, 12 (8 2017), 1937–1940. <https://doi.org/10.14778/3137765.3137813>
- [5] Google. [n.d.]. Google Sheets: Free Online Spreadsheets for Personal Use. <https://www.google.com/intl/en/sheets/about/>. <https://www.google.com/intl/en/sheets/about/>
- [6] David C. Hoaglin, Frederick Mosteller, and John W. Tukey. 2000. *Understanding Robust and Exploratory Data Analysis*. Wiley.
- [7] Nicholas Kong, Marti A. Hearst, and Maneesh Agrawala. 2014. Extracting References Between Text and Charts via Crowdsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) (*CHI '14*). ACM, New York, NY, USA, 31–40. <https://doi.org/10.1145/2556288.2557241>
- [8] S. Latif and F. Beck. 2019. Interactive map reports summarizing bivariate geographic data. *Visual Informatics* (2019). <https://doi.org/10.1016/j.visinf.2019.03.004>
- [9] S. Latif and F. Beck. 2019. VIS Author Profiles: Interactive Descriptions of Publication Records Combining Text and Visualization. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan 2019), 152–161. <https://doi.org/10.1109/TVCG.2018.2865022>
- [10] Shahid Latif, Diao Liu, and Fabian Beck. 2018. Exploring Interactive Linking Between Text and Visualization. In *EuroVis 2018 - Short Papers*, Jimmy Johansson, Filip Sadlo, and Tobias Schreck (Eds.). The Eurographics Association. <https://doi.org/10.2312/eurovisshort.20181084>
- [11] Lexical Analysis Software Ltd. [n.d.]. WordSmith Tools home page. <https://www.lexically.net/wordsmith/>
- [12] Microsoft. 2020. Data Visualization | Microsoft PowerBI. <https://powerbi.microsoft.com/en-us/>. <https://powerbi.microsoft.com/en-us/>
- [13] H. Mumtaz, S. Latif, F. Beck, and D. Weiskopf. 2020. Explorantative Code Quality Documents. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (Jan 2020), 1129–1139. <https://doi.org/10.1109/TVCG.2019.2934669>
- [14] Ben Shneiderman. 1996. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*. IEEE Computer Society, Boulder, Colorado, 336–343.
- [15] Social Progress Imperative. [n.d.]. Social Progress Imperative. <https://www.socialprogress.org/>. Accessed: 2020-08-24. <https://www.socialprogress.org/>
- [16] A. Srinivasan, S. M. Drucker, A. Endert, and J. Stasko. 2019. Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 672–681. <https://doi.org/10.1109/TVCG.2018.2865145>
- [17] H. Strobelt, D. Oelke, B. C. Kwon, T. Schreck, and H. Pfister. 2016. Guidelines for Effective Usage of Text Highlighting Techniques. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan 2016), 489–498. <https://doi.org/10.1109/TVCG.2015.2467759>
- [18] Bo Tang, Shi Han, Man Lung Yiu, Rui Ding, and Dongmei Zhang. 2017. Extracting Top-K Insights from Multi-Dimensional Data. In *Proceedings of the 2017 ACM International Conference on Management of Data* (Chicago, Illinois, USA) (*SIGMOD '17*). Association for Computing Machinery, New York, NY, USA, 1509–1524. <https://doi.org/10.1145/3035918.3035922>
- [19] Ryan W White and Resa A Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009), 1–98.
- [20] Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in Data-to-Document Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2253–2263. <https://doi.org/10.18653/v1/D17-1239>
- [21] A. Ynnerman, J. Löwgren, and L. Tibell. 2018. Explorantation: A New Science Communication Paradigm. *IEEE Computer Graphics and Applications* 38, 3 (May 2018), 13–20. <https://doi.org/10.1109/MCG.2018.032421649>