Explorative Visualization of Food Data to Raise Awareness of Nutritional Value

Anna Lunterova(✉), Ondrej Spetko, and George Palamas

Aalborg University, A. C. Meyers Vaenge 15, 2450 Copenhagen, Denmark
`alunte14@student.aau.dk`

Abstract. The following research proposes a web based, interactive data visualization of a large food dataset, with an aim to raise awareness of their nutritional value. The system consists of an automatic dimensionality reduction scheme, based on t-SNE algorithm, and an exploratory user interface based on dynamic views of the data. Our approach allows users to effectively compare and relate visual representations of nutritional value in both micro and macro levels. The communicativeness of the application was evaluated regarding user experience and the storytelling functionality. Results show that the participants experienced a shift in perception from one mainly focused on taste to one aware of nutritional aspects. Additionally, the results suggest that a narrative experience can be generated, along with an effective long-term engagement.

Keywords: Big data visualization · Dimensionality reduction · Exploratory data visualization · Interactive data storytelling

1 Introduction

Over the past decades great amounts of data have been accumulated and stored from a variety of sources. These data hold qualities that are not always obvious and comprehensible at first sight. Due to data complexity and high-dimensionality, there is a need to generate a model that can present the data in a simpler and more comprehensible way. Moreover, due to the curse of dimensionality [10], a phenomenon occurring in high dimensional spaces, additional problems such as incoherence in the perception of data might arise. Dimensionality reduction can be performed by selecting the primary components of the dataset and together with using dimensionality reduction algorithm t-SNE, the data can be turned into a 2D visualization. Data visualizations (DV) are commonly used for the representation and presentation of the data. They provide an effective way to transform a data-driven information, such in form of numbers stored in tables, into easily understandable and aesthetically pleasing visuals of various forms [2,6,8,17,19,20]. Additionally, complementing data visualization with storytelling and interactivity elements has been found to enhance and promote user-interaction and further exploration of the presented information [11,18]. From a general audience point of view, this can turn the plain data visualization into a more effective and memorable narrative experience [9,12,26].

The USDA National Nutrient Dataset of 2017 includes 8400 different foods with each of these holding 41 dimensions. The high density and amount of the information might be one of the reason of generally low nutritional awareness [25]. Combination of above mentioned elements (dimensionality reduction, data visualization, and interactive storytelling) can be used for the design and development of a web-based application, (accessible at https://tinyurl.com/y5lwevp9) offered to the public to explore the food dataset. Furthermore, this combination suggests a way to tackle the problem of effectively communicating complex, high dimensional datasets.

2 Background Research

Even though the roots of DV in forms of maps, charts and graphs go all the way into 17th century, interactive DV are only recently being explored. Moreover, they are mostly used for communicating low dimensional datasets. Visualizing high dimensional datasets in an interactive way is a more advanced topic, with only a few solutions proposed so far [18]. *The infinite drum machine* experiment done by Google Creative Labs [23] is an example of this approach. This experiment use the t-SNE dimensionality reduction algorithm [15] to organize a big dataset of everyday sounds by similar pitch characteristics. The Google interface can be used to create both random and customized beats using the drum sequencer, as well as to explore different sound regions in the interactive DV (Fig. 1).
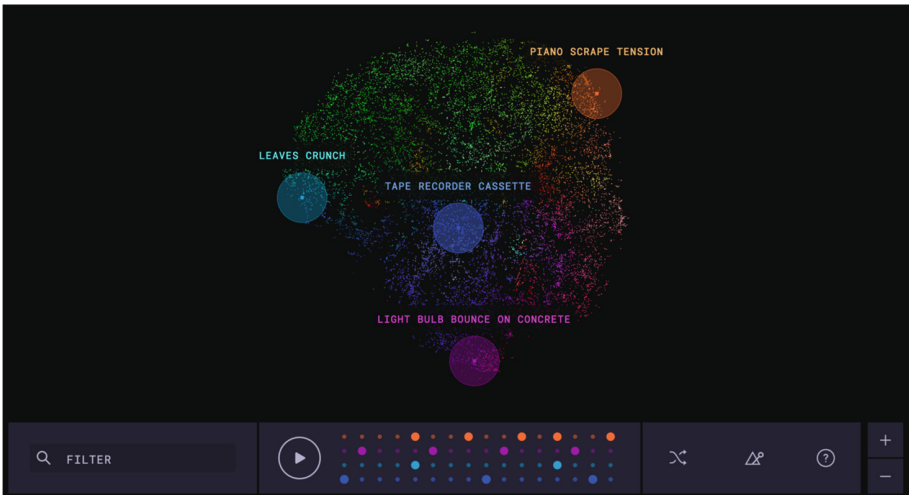

Fig. 1. The infinite drum machine [23]

The exploratory data visualization process consists of the following steps: dimensionality reduction, visualization, user interface and narrative structure. The following sections presents the elements separately.

2.1   Dimensionality Reduction

Dimensionality reduction is a pre-processing stage that is part of the data analysis and data mining phase. By reducing the amount of features in the original dataset, the derived values become more informative, non-redundant, and facilitate comprehension. Two of the most common and most effective dimensionality reduction algorithms are principal component analysis (PCA) and T-distributed Stochastic Neighbor Embedding (t-SNE).

Principal Component Analysis PCA, one of the most common and old (1933) algorithms, uses an orthogonal transformation to manipulate a set of observations of likely similar variables into principal components (a set of values of linearly uncorrelated variables). One downside with linear dimension reduction algorithms is that linear algorithms are not able to describe complex polynomial relations between features. The dissimilar data points are being placed far apart in lower dimensional representations. T-SNE solves these issues by using probability distribution with elements of randomness instead. This allows t-SNE to model distances between points in the low-dimensional map, resulting in an alteration of the Kullback-Leibler divergence curve [7].

T-SNE is a relatively new technique (2008) developed by Laurens van der Maaten [15], with aim to reduce the amount of dimensions by assigning relative distance values. It is based on relative correlations between each datapoint of the raw dataset, resulting in a two or three-dimensional map. The strength of t-SNE lies mainly in its preservation properties of the local distances of the high-dimensional data while mapping them to low-dimensional data. Main tunable parameters of the algorithm are perplexity (sets the number of effective nearest neighbors) and number of iterations. The choice of those parameters has a complex effect on the final picture and depends on the density of the dataset. By exploring its behaviour with different parameters it results in more meaningful information. However, original dimensions are mapped to lower dimensions without preserving any link to their original values, therefore definite assumptions based only on t-SNE output can not be made [20]. Also, the t-SNE algorithm is quite heavy on the system resources as it compares the relations pairwise with the goal of minimizing the sum of the difference of the probabilities in higher and lower dimensions. However, the output of the t-SNE can be used in the process of classification or clustering as an input for further classification or clustering algorithms. The visualization produced by t-SNE are found to be significantly more accurate and informational compared to other dimensionality reduction techniques such as PCA, Sammon Mapping, Isomap, Locally Linear Embedding (LLE) and other.

2.2   Data Visualization

Visualization Tools. The simplest examples of DV includes graphs, maps, 3D models, plots and others [16]. Several online tools and systems are being developed to make the creation of visualizations accessible to non experts and people with low programming skills. (E.g., Many Eyes, Tableau, Power Map, Flourish, iNZight, RAWgraphs, QGIS, Gephi, NodeBox, etc.) However powerful, these solutions are based on predefined templates. The shortcomings are poor customization, visualization of only low dimensional data, mostly static or very little interaction, and exclusively author-driven solutions [11].

Data Visualization Principles. Based on data expert Alberto Cairo's recent book "The Truthful Art" [6], and pioneer in the DV Edward Tufte's "The Visual Display of Quantitative Information" [24], qualities of great visualizations are summarized into five principles. While the name of these principles differ in between these two authors, their characteristics are the same. The order of these design principles are organized by their level of importance, from highest to lowest:

– Truthfulness, or "the lie factor", meaning data are based on thorough and honest research.
– Functionality and low data-ink ratio. Graphic providing accurate and concise depiction of data, that help the viewer to understand and think about the shown information (rather than the design).
– Should be aesthetically pleasing, in the sense of being intriguing and attractive for its audience.
– Insightfulness, revealing otherwise harder to be revealed evidence.
– Enlightening, in a way that if the audience grasps and understand the shown evidence, it changes their understanding or behaviour for the better.

These principles serve as inspiration for evaluation method of an effective data visualization.

2.3   Narratives and Interactive Storytelling

Storified visualizations are DVs that contain story components (structures, elements, and concepts) and are framed within elements that mediate the telling of the story "(people, tools, and channels)" [13,14] . The process of creating data stories in research is mapped out in three phases by author Bongshin Lee and Nathalie Henry Riche in their work entitled "More Than Telling a Story: Transforming Data into Visually Shared Stories" (see the process on Fig. 2).

Firstly the chosen data is explored and analyzed (data analysis). Secondly, after seeing patterns, trends, differences etc., the chosen facts, (story nodes) are connected (story transitions) in meaningful order (plot), to form a story. Lastly, the story is told using a chosen medium and narrative technique. Those "story nodes" are visualized in forms of graphs, charts, maps etc., with the use of clarifying annotations, or narrations [13]. After having created story pieces, that are
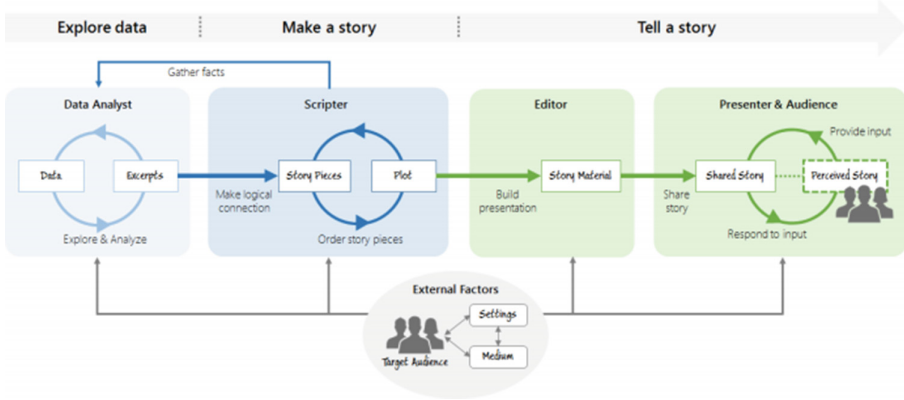
Fig. 2. Storytelling process in DV, taken from *More Than Telling a Story: Transforming Data into Visually Shared Stories* [13]. The process is composed from three main phases: exploration of the data, making of the story, and telling of the story.

the core of the information to be communicated, they need to be transformed into a physical form of a design. The layout of the visual data story, (defined during editor's phase) characterized as "design space dimensions" in research "Narrative visualization: Telling stories with data", where they analyzed 58 visualization, is composed from three characteristics: genre, visual narrative tactics, and narrative structure tactics [22]. Those characteristics are chosen by the author, based on the kind of data (or story nodes) he has, and his rhetorical strategies (story, plot, narration) [1]. This can serve as a foundation for connecting DV within a narrative context.

To further enhance the audience relatedness of the data story, the level of abstractedness or didascalicity, also called author-audience distance, serves as a function of narrative intelligibility and closure that the system achieve or aim to achieve [5].

In relation to the problem of author-audience distance, Hulman & Diakopoulos and Segel & Heer suggest that in data visualizations there needs to be certain balance between author-driven and reader-driven scenarios. This is described as an optimal interplay between an unrestricted exploration from the reader side and a clear story from the author side. A balanced narrative should be a combination of persuasive-rhetorical strategies to transmit an intended information to users, as well as exploratory-dialectic strategies, to give the reader a certain control over the insights. This is specifically called a "Martini glass" structure presenting narrow fully author driven beginnings, and reader driven exploration at the end of the interface [11,22].

Compared to static visualizations, exploratory visualizations give an ability to interact, inviting the reader to be part of the story creation and offers a variety of perspectives on the same information [4,19]. Interactive storytelling is defined as unique user story experiences that go beyond typical clicking, hovering, scrolling and linear structures of the story [3]. Users can choose the new

story nodes in an unrestricted way while remaining within the overall story's beginning and ending.

3   Design

The process of design includes the four following topics: dimensionality reduction, visualization, interactivity and narrative structure. The chosen dataset, (USDA National Nutrient Dataset of 2017) reported on 8400 different foods with each of these holding 41 dimensions. These dimensions hold attributes including food category, amount of calories, protein, carbohydrate, fat, fiber values, amount of different minerals and vitamins, and recommended daily amounts of each of the item.

3.1   Dimensionality Reduction

The aim of the reduction of the information is to automatically extract the nutritional composition of each food and organize them in a way that retain nutritional similarities across the data-set. The prioritization of what food attributes should be communicated, was based upon a brief online research of what matters most from a physiological point of view. Moreover, and from the same point of view, a discussion was held regarding the USDA database with two medical students. It was important to be aware of the tuning parameters that are influencing the output of the t-SNE algorithm. The best visual representation was achieved
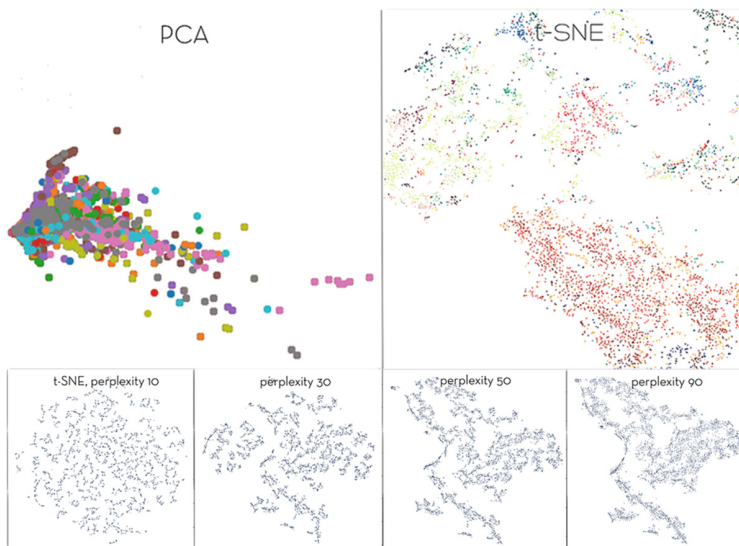

Fig. 3. Dimensionality reduction, visualizations comparison of outputs of PCA and t-SNE algorithm with comparison of t-SNEs maps with perplexity of 10, 30, 50 and 90.

with the number of iterations 4000 and perplexity of 90. The t-SNE performed better than PCA algorithm which resulted in a single cluster formation. Figure 3 shows the comparison of the visualization output of PCA and t-SNE algorithm of perplexity 10, 30 50, 90).

Furthermore, based on the general population selection criteria, a decision was made to color code all 23 food groups. The final t-SNE map (Fig. 4) shows all the points organized in nutritionally similar groups in the 2D space.
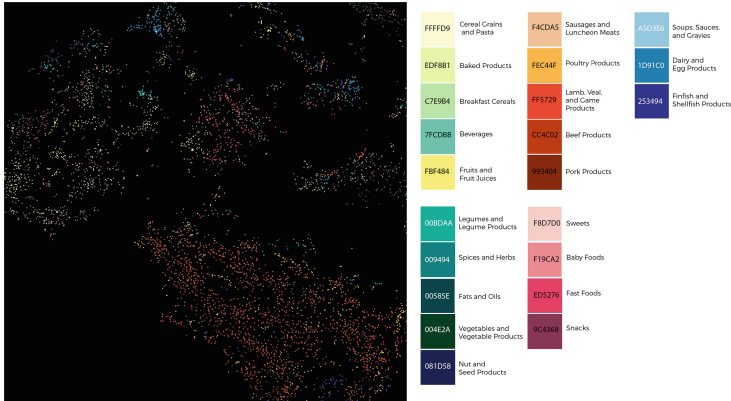

Fig. 4. 2D representation of food data set in nutritionally neighboring positions (Macro level). Color represents food category.

3.2 Visualization

Taking the resulting position of food points from the t-SNE, a custom built DV using the JavaScript open source library D3.js was created. This solution allows the creation of any kind of customized visualization by manipulating HTML DOM elements [19]. For visually encoding data into visual attributes the hierarchy of elementary perceptual tasks was followed [6]. Based on the information being communicated, a cyclical area chart was chosen as a DV type [21]. The customized cyclical area chart, on Fig. 5 shows the final zoomed visualization of a data point with an added legend. The color of the lines represents the food group, while the length represents the percentage of recommended dietary allowance of specific minerals and vitamins. The maximal amount is marked with the short line representing invisible circle around the chart. The color of the small circles at the end of the lines represents the macro-nutrients of the food, with orange for carbohydrates, green for proteins, and red for fat. The intensity of the color is linked with the percentage of the specific macro-nutrient.

3.3 Interactivity

Differences and similarities between food and their associated food groups distribution can be been explored by moving in this 2D space and by zooming in
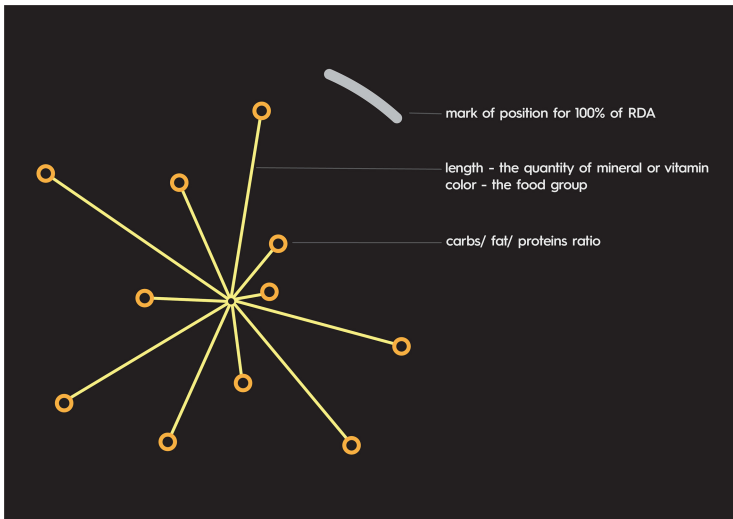
Fig. 5. Graphical representation of the nutritional value of a single food point (Micro level)

and out. Specific foods or food groups can be picked up in the map, customarily searched by name, or selected through the navigation bar. Tool tips and details-on-demand were added to complement the experience and facilitate the understanding of the DV. The aim of this is to strengthen the ability of mental map formation for the audiences, as well as enhance the awareness of how our bodies perceive food.

3.4 Narratives

The goal of the narrative was to acquire narrative closure with the audience, in terms of how they consider the purpose of food and their relationship to it. The chosen plot consisted of three parts: (a) A shortly animated slideshow narrated through gradually appearing text introducing the topic of people's relationship with food. (b) A demo introduction to the food world presented, including a demonstration of the navigation tools. In this phase the user is also invited to choose a character/category to identify with. The choice to be made addresses the user's perspective on food. This was assumed to lead the user to question how and why he chooses his food, and/or to be reminded of what matters particularly to him about a meal. (c) Pure reader-driven exploration of DV, with selective interactions (Fig. 4). This is the conclusion of the story, from a user perspective, as the information presented is large and complex. The points of possible inquiry are in the form of (a) data point stories representing food nutrients, calories, and food groups, (b) comparison of visualizations with other food points, or other food groups, (c) random recommendations of complimentary foods by the system that get highlighted after choosing specific points in the map. The Fig. 6 shows

(a) slide from a short animation introducing the topic of people's relationship to food; (b) A demo introduction to the food Data Visualization, including a tutorial and an invitation to choose a character/category to identify with.

Fig. 6. A user driven narrative (a) with selective interactions (b).

4   Experimental Procedure

The Web-based DV application was tested with 25 participants. Self reported data, regarding the communicativeness and the usability of the solution, were collected from a pre and post-questionnaire. The pre-questionnaire served as an introduction to the exploration and included demographic questions, questions about their interest in the topic of food, and reasons for their interest in participating in the experiment. In the post-questionnaire eight Likert scale questions served for self-evaluating (a) the usability of the interface, (b) the perceived attentiveness, (c) and the degree of relatedness to the information presented. Additional questions for evaluating the insightfulness, the aesthetic level, and the information obtained were given. At the very end of the post-questionnaire, two open questions served for further open reflections on the interest in food and experiment overall. Twenty-five participant's answers were collected, in the span of three days. The experiment took approximately 5 to 10 min, with identical questionnaires and procedures given to all participants.

The results from the questionnaire measuring the effectiveness of the scheme, regarding the communicativeness (attentiveness, relatedness, insightfulness) and the usability of the solution are showed in a radar chart in the Fig. 7(a). The 7(b) shows keywords analysis from two same but differently asked questions regarding what the user cares about the food from before and after-experiment. The questions were: *What do you care about food?* and afterwards *How do you define healthy relationship with food?*
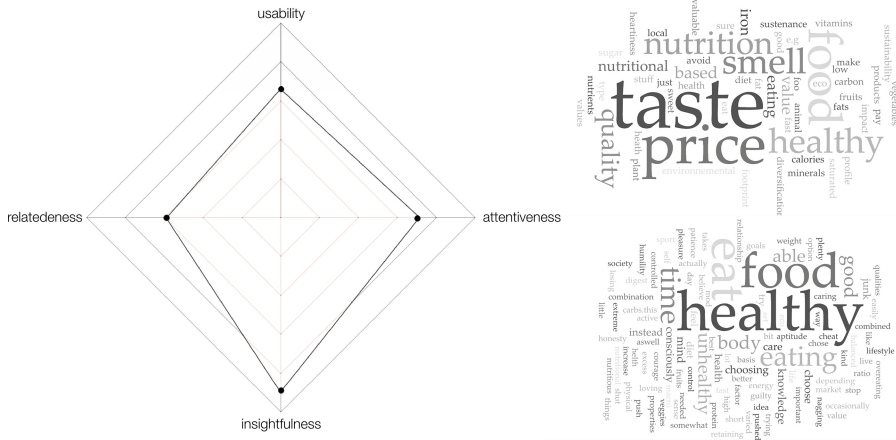
Fig. 7. Usability testing (a) pre and post-questionnaire responses (b).

5   Discussion

From the analysis of over 250 keywords from the pre-questionnaire it appeared that, initially, the prioritization of taste was a common drive for choosing what food to eat (Fig. 7(b)). Post-questionnaire responses, however, show that users were now more concerned about the health side of different foods, even though the definition of "healthy food" here varied for most of the participants and might not be clear as a whole. This suggests that the goal of the narrative was partly fulfilled. For properly evaluating the communicativeness of the visualization and the usability of the interface, more data are needed to confirm the results.

6   Conclusions

The aim of this research was to communicate effectively a large data-set of over 8400 food data points with a trans-disciplinary approach. There were two main focus points of this research, the communicativeness and usability of the visualization. The approach consists of a combination of dimensionality reduction, data visualization and interactive storytelling. Even though further evaluation is needed to confirm the results, the effectiveness of the approach was validated as positive and led us to conclude that there is a big potential that deserves further investigation and study.

References

1. Abbott, H.P.: Story plot, and narration. In: Herman, D. (ed.) The Cambridge Companion to Narrative, pp. 39–51. Cambridge University Press. https://doi.org/10.1017/ccol0521856965.003

2. Billeskov, J.A., Møller, T.N., Triantafyllidis, G., Palamas, G.: Using motion expressiveness and human pose estimation for collaborative surveillance art. In: Brooks, A.L., Brooks, E., Sylla, C. (eds.) ArtsIT/DLI 2018. LNICST, vol. 265, pp. 111–120. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-06134-0_12

3. Bostan, B., Marsh, T.: Fundamentals of interactive storytelling. AJIT-e **3**(8), 19 (2012)

4. Boy, J., Detienne, F., Fekete, J.D.: Storytelling in information visualizations: does it engage users to explore data? In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM Press (2015). https://doi.org/10.1145/2702123.2702452

5. Bruni, L.E., Baceviciute, S.: Narrative intelligibility and closure in interactive systems. In: Koenitz, H., Sezen, T.I., Ferri, G., Haahr, M., Sezen, D., Catak, G. (eds.) ICIDS 2013. LNCS, vol. 8230, pp. 13–24. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-02756-2_2

6. Cairo, A.: The Truthful Art: Data, Charts, and Maps for Communication. New Riders (2016)

7. Canning, J.P., et al.: Network classification and categorization. arXiv preprint http://arxiv.org/abs/1709.04481.pdf

8. Friendly, M., Denis, D.J.: Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization. Springer (2001). http://www.datavis.ca/papers/gfkl.pdf

9. Gershon, N., Page, W.: What storytelling can do for information visualization. Commun. ACM **44**(8), 31–37 (2001). https://doi.org/10.1145/381641.381653. Association for Computing Machinery

10. Houle, M.E., Kriegel, H.-P., Kröger, P., Schubert, E., Zimek, A.: Can shared-neighbor distances defeat the curse of dimensionality? In: Gertz, M., Ludäscher, B. (eds.) SSDBM 2010. LNCS, vol. 6187, pp. 482–500. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13818-8_34

11. Hullman, J., Diakopoulos, N.: Visualization rhetoric: framing effects in narrative visualization. IEEE Trans. Visual Comput. Graphics **17**(12), 2231–2240 (2011)

12. Kosara, R., Mackinlay, J.: Storytelling: the next step for visualization. Computer **46**(5), 44–50 (2013). https://doi.org/10.1109/mc.2013.36

13. Lee, B., Riche, N.H., Isenberg, P., Carpendale, S.: More than telling a story: transforming data into visually shared stories. IEEE Comput. Graph. Appl. **35**(5), 84–90 (2015). https://doi.org/10.1109/mcg.2015.99

14. Ma, K.L., Liao, I., Frazier, J., Hauser, H., Kostis, H.N.: Scientific storytelling using visualization. IEEE Comput. Graph. Appl. **32**(1), 12–19 (2012). https://doi.org/10.1109/mcg.2012.24

15. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008). http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

16. Mallon, M.: Data visualization. Public Serv. Q. **11**(3), 183–192 (2015)

17. Moere, A.V.: Aesthetic data visualization as a resource for educating creative design. In: Dong, A., Moere, A.V., Gero, J.S. (eds.) Computer-Aided Architectural Design Futures (CAADFutures), pp. 71–84. Springer, Netherlands (2007). https://doi.org/10.1007/978-1-4020-6528-6

18. Murray, S.: Engaging Audiences with Data Visualization. O'Reilly Media (2014). https://www.oreilly.com/pub/e/2584

19. Murray, S.: Interactive Data Visualization for the Web: An Introduction to Designing with D3, vol. 2. O'Reilly Media, Inc. (2017). http://opencarts.org/sachlaptrinh/pdf/28380.pdf

20. Nasser, A., Hamad, D., Nasr, C.: Visualization methods for exploratory data analysis. In: 2006 2nd International Conference on Information & Communication Technologies. IEEE (2006). https://doi.org/10.1109/ictta.2006.1684582
21. Raghavan, S.: How to begin your own data science journey (2017). https://towardsdatascience.com/how-to-begin-your-own-data-science-journey-2223caad8cee
22. Segel, E., Heer, J.: Narrative visualization: Telling stories with data. IEEE Trans. Visual Comput. Graphics **16**(6), 1139–1148 (2010). https://doi.org/10.1109/tvcg.2010.179
23. Tan, M., McDonald, K.: The Infinite Drum Machine (2017). https://experiments.withgoogle.com/drum-machine
24. Tufte, E.R.: The Visual Display of Quantitative Information, vol. 2. Graphics Press, Cheshire (2001)
25. Wojcicki, J.M., Heyman, M.B.: Adolescent nutritional awareness and use of food labels: results from the national nutrition health and examination survey. BMC Pediatr. **12**(1), 55 (2012)
26. Yi, J.S., ah Kang, Y., Stasko, J.: Toward a deeper understanding of the role of interaction in information visualization. IEEE Trans. Visual Comput. Graphics **13**(6), 1224–1231 (2007). https://doi.org/10.1109/tvcg.2007.70515