Semantic Explanation of Interactive Dimensionality Reduction

Yali Bian* Virginia Tech
Chris North† Virginia Tech
Eric Krokos‡ Department of Defense
Sarah Joseph§ Department of Defense

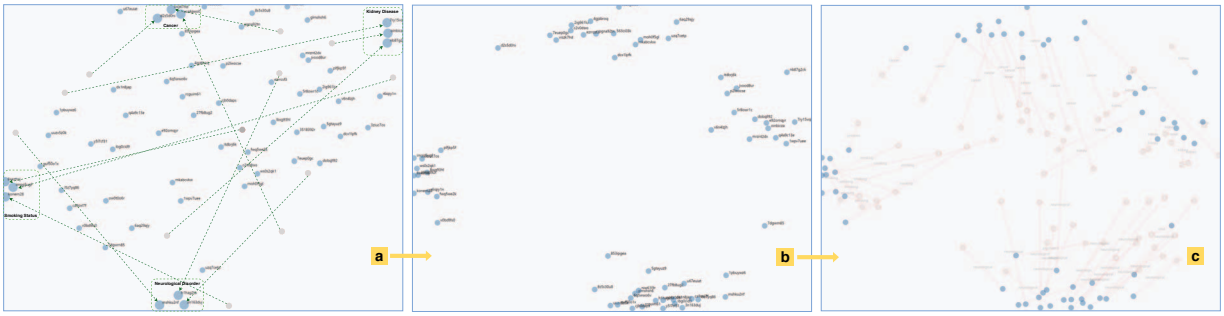

Figure 1: Screenshots during the analysis of COVID-19 research articles about four different risk factors using an interactive DR system powered by deep learning with BERT. (a) With semantic interactions, the analyst can provide visual feedback regarding articles about different risk factors by directly manipulating the initial data projection. These interactions are then exploited to tune the DR model. (b) The resulting projection is updated by the human-steered model. However, the nonlinear model lacks explanations of the meaning of the clusters created. (c) Our proposed semantic explanation solution visualizes how the human-steered model directly manipulated the projection and why.

2021 IEEE Visualization Conference (VIS) | 978-1-6654-3335-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/VIS49827.2021.9623322

Abstract

Interactive dimensionality reduction helps analysts explore the high-dimensional data based on their personal needs and domain-specific problems. Recently, expressive nonlinear models are employed to support these tasks. However, the interpretation of these human-steered nonlinear models during human-in-the-loop analysis has not been explored. To address this problem, we present a new visual explanation design called semantic explanation. Semantic explanation visualizes model behaviors in a manner that is similar to users' direct projection manipulations. This design conforms to the spatial analytic process and enables analysts better understand the updated model in response to their interactions. We propose a pipeline to empower interactive dimensionality reduction with semantic explanation using counterfactuals. Based on the pipeline, we implement a visual text analytics system with nonlinear dimensionality reduction powered by deep learning via the BERT model. We demonstrate the efficacy of semantic explanation with two case studies of academic article exploration and intelligence analysis.

Index Terms: Interactive Dimensionality Reduction, Projection Explanation, Counterfactual Explanation, Human-in-the-loop Analysis

1 Introduction

Interactive dimensionality reduction (DR) [27] is a commonly used human-in-the-loop machine learning technique for exploratory analysis of high-dimensional data. Interactive DR systems enable users to adjust DR parameters (such as feature weights) to incorporate human knowledge and questions into the model during the spatial analytic process [1]. Therefore, users can explore and analyze high-dimensional data based on their own needs and domain-specific problems. However, parameter tuning usually needs particular mathematical knowledge, which is a daunting burden for analysts with cognitively demanding tasks [21]. They must pause the analysis of the data to determine how to adjust model parameters to formally externalize their intents for the next exploration [14].

To solve this bottleneck, *semantic interaction* (SI) was proposed [14]. In interactive DR systems with SI (Fig. 2), analysts can naturally express their intents about the data by directly manipulating the projection. The direct manipulation of visual metaphors is consistent with the analyst's spatial analytic process. With these intuitive interactions, the analyst can remain within the cognitive zone, thereby enhancing the analyst's efficiency in performing analytic tasks [19]. As shown in Fig. 1a, the analyst can move several documents into four clusters to express their preferred data layout, intuitively defining high-level concepts related to COVID-19 risk factors [33]. Subsequently, the system is responsible for learning a new DR model to infer the associated intent behind these interactions and update the projection as visual feedback (Fig. 1b).

After performing interactions, analysts must understand the changed behavior of the updated model in response to their interactions. With a good understanding, the analysts can make better decisions: (1) *build trust*, intuitively reasoning whether the projection change has sufficiently captured their intents, and (2) *gain knowledge*, formally defining their high-level spatial layout concepts in terms of lower-level input features [13]. In SI systems with linear DR [5, 6, 20, 28], analysts can quickly understand the model behavior through case-based reasoning of the updated projection. Recently, several researches have been done to employ advanced and expressive nonlinear DR models in SI systems [3, 24]. While these nonlinear models offer new opportunities to power interactive DR with more accurate inference, they also make the system challenging to understand. We argue that model explanation is the next

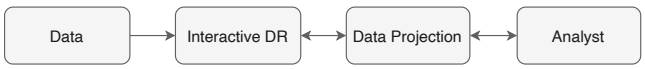

Figure 2: Interactive DR with SI in which analysts directly manipulate the projection to externalize their intents. Adapted from [14, 27].

*e-mail: yali@vt.edu
†e-mail: north@vt.edu
‡e-mail: epkroko@evoforge.org
§e-mail: skjosep@evoforge.org

978-1-6654-3335-8/21/$31.00 ©2021 IEEE
DOI 10.1109/VIS49827.2021.00013

bottleneck preventing analysts from focusing on their analysis with high efficiency.

In this paper, we aim to address this issue by proposing a new visual explanation design, called *semantic explanation* (SE). SE visualizes model behaviors in a manner that is similar to users' direct projection manipulations. As a complement (Fig. 1c), SE explicitly shows how and why the updated DR model changed the data projection in response to the analyst's semantic interactions: the updated model moves the intermixed data points from the center to separated clusters because of the essential features. Similar to the design of SI, SE conforms to the spatial analytic process (Fig. 3). Specifically, as shown in Steps 3-4, analysts can naturally understand the model behaovior and projection updates via visual explanations about how data features support the projection change. SE leverages the cognitive connection formed between analysts and spatial layouts for model understanding, thereby accelerating the analytic process.

We propose a pipeline to empower interactive DR systems with SE using counterfactuals [32]. This is because counterfactual explanation elicits causal thinking between the change of input features and the model prediction updates [32]. It is commonly used in interpreting supervised models, explaining how small changes of input feature values can cause the prediction changes of indivisual instances [32]. Here we utlize a perturbation-based method [26] to generate and select cuonterfactuals to explain the projection changes made by interactive DR systems in an unspervised manner. We apply our SE pipeline to a semantic interaction system for visual text analytics powered by deep learning via the BERT model [12], called DeepSE. We demonstrated the utility of DeepSE via two case studies: academic paper exploration and intelligence analysis. Results demonstrate how SE enables understanding the model behavior and formalizing concepts during the analytic process.

2 Related Work

Three areas of related research support SE: semantic interaction, nonlinear projection explanation, and counterfactual explanation.

Semantic interaction: Interactive DR systems with SI usually adapt linear DR methods [5, 6, 20, 28], because of the easy interpretation property in supporting analysts' process of incremental formalism [14, 29]. Analysts can naturally associate their externalized concepts with the change of projection, which is linearly associated with updates of input feature importance. Recently, several SI systems have employed expressive nonlinear dimensionality reduction to improve the inference ability [3, 24]. DeepSI [3] proposes a general framework to power the interactive DR component with interactive deep learning. However, exiting visual explanations used in SI systems, such as the global feature importance through slider bars [28] and the node-link diagram connected by shared entities [5], only work for linear models. SE can interpret any interactive systems with either linear or nonlinear DR algorithms.

Nonlinear projection explanation: Several visual explantion methods have been proposed to interpret the projection results of nonlinear methods, that can be categorized as: model-specific [9, 30], model-agnostic [7, 16, 17]. These methods offer detailed and valuable guidance in determining the meaning of the data projection. Differently, SE focuses on the interpretation of the updated DR models and the associated projection changes in interactive systems for human-in-the-loop data analysis.

Counterfactual explanation: Counterfactual explanations describes the causal relationships between the change to data features and the change to model results, for what-if analysis [32]. Recently, counterfactual explanations have been widely applied to interpret machine learning, particularly classification models [23, 32]. A variety of VA techniques have been developed for counterfactual explanations, including ViCE [34], What-If Tool [11], and DECE [11]. In this paper, we implement and intergrate SE to interactive DR systems through the usage of counterfactuals.

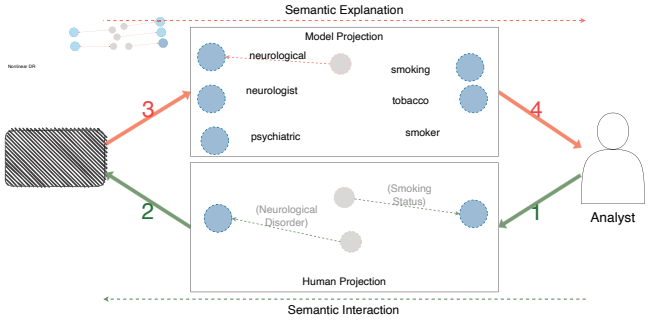

Figure 3: Design rationale of SE.

3 Semantic Explanation

3.1 Design Rationale

The cognitive connection between the analyst and the spatial layout, through the visual *proximity* ≈ *similarity* metaphor, makes SI an effective interaction methodology for analysts to express their intent, as it does not require analysts to formalize their analytic reasoning [22]. Similarly, the design rationale behind SE is also the connection between the analyst and the spatial layout for model explanations. These explanations explicitly correlate the abstract projection to input features (words) and help the analyst formally define semantic concepts with high efficiency. Fig. 3 shows the detailed description of the rationale.

With SI, the analyst performs spatial analytic interactions to externalize their semantics of information about concepts in the projection [14] (Step 1). For example, the analyst moves two articles apart because these articles have different semantic meanings: *smoking status* vs. *neurological disorder*. Alternatively, they might create a spatial cluster of several similar articles related to *neurological disorders* as a COVID-19 risk factor (as shown in Fig. 1a). Using cases (informal relationships) rather than formalized feature descriptions, the analyst can express high-level semantics on the fly. It is the system's responsibility to infer the semantics behind interactions, formalize the concept (Step 2), and update the model projection.

With SE, the system explicitly formalizes and contextualizes the associated analytical (causal) reasoning in updating the projection (Step 3). Specifically, the system explicates how essential features drive the underlying model to update all observations in a way similar to the analyst's spatial analytic interaction. In the model projection in Fig. 3, for example, all the data points are moved into two separate clusters by the underlying model. Furthermore, all the movements made by the model are interpreted with the essential features that contribute most to these movements. These visual explanations can be correlated with the analytical reasoning behind the analysts' interactions. The explanations *neurological*, *neurologic*, and *psychiatric* support the analyst's internal decision to search for *neurological disorder*. The explanations *smoking*, *tobacco*, and *smoker* support the analyst's internal choice of *smoking status*. Therefore, the analyst gains a deeper, more formalized understanding of the meaning of the clusters they have created (Step 4).

3.2 Interactive DR with SE Using Counterfactuals

Counterfactual explanation elicits causal thinking regarding the change of input features and the prediction updates for individual data instances [32]. The use of counterfactuals can generate explanations that meet the design requirements described in the previous subsection. We propose a new pipeline to power interactive DR systems with SE using counterfactuals. As shown in Fig. 4, the pipeline consists of two components: the counterfactual engine and the counterfactual projection.

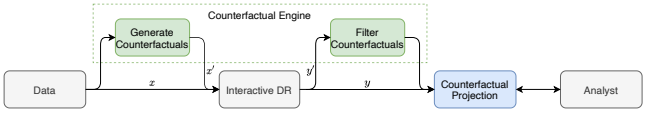

Figure 4: Interactive DR pipeline with SE.

3.2.1 Counterfactual Engine: Generate Explanations

The counterfactual engine interprets nonlinear DR models by creating representative counterfactual explanations for all individual observations. As shown in Fig. 4, it has two subprocesses: generate counterfactuals and filter counterfactuals. First, the engine generates a group of counterfactual candidates, x', by making small changes to the features compared to the input instance x. For simplicity, we only change one independent feature dimension at a time when generating counterfactuals. For a given instance with M features, M counterfactuals will be generated. Then, the high-dimensional data instances and their counterfactual candidates are projected into the low-dimensional space ($x \rightarrow y, x' \rightarrow y'$). This is a perturbation-based method [26] which has been commonly used in existing counterfactual visualization systems [11, 18, 34].

Next, the engine filters out all the representative counterfactual candidates based on an objective function. Existing counterfactual methods are often used in interpreting supervised models [31, 32], in which class labels are used to define objectives. However, the explicit ground truth is not available in interactive DR systems. For simplicity of exposition, we propose a new optimization function consisting of two objectives: *proximity* and *validity*. The proximity objective seeks to create a set of counterfactual examples, x', from the original instance, x, with small feature perturbations by minimizing the distance between x and x'. The validity objective seeks to select important counterfactual examples, x', that contribute most to the changes in the projection space, precisely the distance between y and y'. Taken together, we seek counterfactuals where small data changes cause significant representation changes. By default, the counterfactual projection displays the top one counterfactual for each instance, and the detailed explanation view displays the top 5 for one selected instance. An open area of research is how to design valid counterfactual constraints to extract representative counterfactuals based on human interactions during the analytic process.

3.2.2 Counterfactual Projection: Contextualize Explanations

The counterfactual projection is a novel visualization to contextualize counterfactual explanations into the data projection layout. Inspired by the flow-based scatterplot [8] and Praxis [7], we propose a counterfactual projection. As shown in Fig. 3 (model projection), we integrate counterfactuals for all data points into the traditional scatterplot with proline-like visual metaphors [7] to highlight both local and global patterns of explanations for the DR model and projection changes. As in the original scatterplot, data points are rendered as solid blue dots and positioned based on their similarities. In addition to data points, counterfactual examples are also visualized as dots but in a smaller size and with a light gray color so that the analysts can easily distinguish between data instances and explanations. The dotted red line between the data instances and the counterfactual examples shows the projection changes induced by the input feature changes.

In addition, the counterfactual projection provides a series of interactions to explore explanations and avoid visual clutter. Extra detailed views can also be employed to help analysts interpret the feature influence within the data points under investigation, as in StarSPIRE [5]. We describe the detailed visualization and interaction design for the visual text analytics system in the following section.

These two components enable model explanations with the beneficial properties described in the previous subsection. It is worth noting that the counterfactual examples do not represent where the data instances were before the interaction and re-projection. Instead, the counterfactual explanation shows the projected position of counterfactual instances with the perturbations of these essential features from the original data instances in the current, updated model.

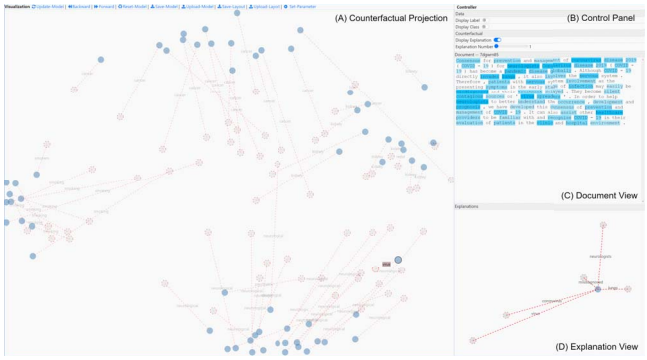

Figure 5: The visual interface of DeepSE, and Case Study 1.

4 DeepSE System for Visual Text Analytics

Following the proposed pipeline described in Sec. 3.2, we build an interactive DR system with SE for visual text analytics, called DeepSE. The interactive DR component of the system is powered by deep learning with the BERT model, as is explained in more detail in DeepSI [3]. It is worth noting that DeepSI learns from user interactions, while DeepSE adds the counterfactual engine and the counterfactual projection to provide visual explanations of what was learned from the user interactions.

The counterfactual projection is the main view and entry point for analysts. This view is designed based on the model projection in Fig. 3 to provide visualizations for both predicted data points and their relevant, representative counterfactuals. The text label overlying the counterfactual line shows the keyword removed from the original instance to create the counterfactual example at the other end of the line. Analysts can intuitively understand that removing the keyword from the data instances leads to the instance moving to the counterfactual position. In reverse, these counterfactual lines allow analysts to obtain a sense of the data flow from the counterfactual examples to the instances because of the keyword. Therefore, analysts gain an overall comprehension of how the projection is influenced by important words in the documents (Fig. 5). The visualization also provides relevant interactions for analysts to explore explanations in detail as needed, such as selecting a data point and highlighting relevant counterfactuals.

The control panel allows the analyst to change the display of the counterfactual projection. When a document in the projection is selected, the document view displays the full content of the selected document in the form of a text heatmap visualization. The word importance is calculated based on the internal attention maps of the BERT model [12]. The explanation view enables analysts to inspect the full explanations for the selected document.

4.1 Case Study 1: Academic Articles on COVID-19

This case study shows how SE assists a medical researcher in reasoning whether the model has sufficiently captured her intent of four different risk factors from the COVID-19 Open Research Dataset (CORD-19) [33]. Fig. 1 shows the process of communication between the analyst and the DeepSE system. Fig. 1-a shows the projection initialized by the default DR model, in which all the articles are spatially intermixed. The analyst then performs SI to provide visual feedback on 12 articles about different risk factors

of interest to her (cancer, chronic kidney disease, neurological disorders, and smoking status) to reflect the perceived connections between articles. Fig. 1-b shows the model projection updated by the underlying model to capture these new risk factor concepts. The overall projection shows four clear clusters, but does not show how consistent the new clusters are to user expectations. Without SE, the analyst must check all the articles to ensure they are appropriately grouped to learn about related topics. Therefore, the analyst turns on the *Display Explanation* to show explanations and understand the model projection with the help of SE.

As shown in Fig. 1-c and Fig. 5, the counterfactual projection shows all the data points and their relevant counterfactual explanations. Almost all the counterfactual points are towards the center of the scatterplot, compared with the data points. The global trends of the counterfactual lines show the data flow from the center to the margins. There are essential words that contribute most to the forming of these four clusters. This indicates the updated model has pushed all the instances away from the center because of some essential words related to COVID-19 risk factors to form these four, clear clusters. For example, counterfactual explanations show that the cluster *smoking status* is formed by documents with the keyword *smoking* or *smoker*. The underlying model updated documents from the center of the projection to the left because of these keywords.

Based on SE, the analyst concludes that the model-constructed clusters are indeed consistent with her intent as expressed via her SI. She learns that there are numerous additional documents to support her hypothesis about these four key risk factors, which guides her continued investigation. She also finds an outlier document about the virus on the right side that does not focus on these factors.

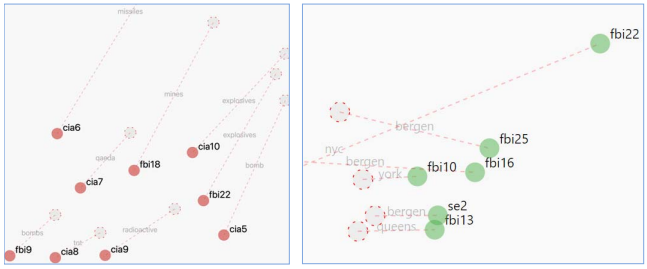

Figure 6: Explanations of two clusters formed in Case Study 2.

4.2 Case Study 2: Intelligence Reports

This case study shows how explanations evolve during iterative interaction. In the analytic process, analysts progressively refine the spatial structure of the projection to gain new insights over time. This is often the situation in early stages of the analytic process before clear concept definitions emerge. For example, documents that belong to one cluster may be relocated to another cluster for a different meaning in a later phase of the analytic process [15]. To demonstrate the versatility of SE in different stages of the analytic process, we applied DeepSE to an intelligence analysis training dataset containing 42 fictional intelligence reports regarding a coordinated terrorist plot and international transportation of explosives.

At the beginning of the study, the analyst looks for reports mentioning *explosive* to find potentially suspicious and interesting facts [1]. Fig. 6a shows the updated counterfactual projection after the analyst groups several reports related to this concept. More reports gather close to the grouped reports and form a cluster. Keywords that have similar meanings to *explosive* (such as *missiles*, *mines*, *bombs*, and *radioactive*), have a dominant effect on forming this cluster. This is consistent with the intended semantics from the analyst, and helps to broaden their understanding of the plot to other types of explosives and other useful documents.

After learning additional information and developing a more precise understanding, the analyst narrows the focus to reports related to a specific location near *New York City* because of a potential attack plot. As shown in Fig. 6b, six reports gather together in the updated projection to capture this concept. To check the semantics of the cluster in detail, the intelligence analyst uses SE. The location keywords (*bergen*, *york* and *queens*) are the main reason in forming the cluster. For example, document "fbi22" was originally clustered based on *explosives*, but is now clustered based on *nyc* location. The model has captured the analyst's process of incremental formalism [29], forming an updated schema focused on attack location.

5 Discussion and Conclusion

Generality and Applicability: Our prototype DeepSE currently interprets nonlinear DR models powered with deep learning for text data. As a model-agnostic explanation method, SE can be generalized to different DR models and applied to other data types. For example, SE can be integrated into SI systems with Weighted MDS, such as Andromeda for numerical data analysis [28] or DeepVA for visual concept analysis [4].

Performance: The most significant issue with the counterfactual explanation method is the time complexity. The time complexity of the counterfactual engine depends on data size N and feature dimension M. It needs to generate and sort M counterfactual explanations for all N instances (totally $N*M$ counterfactuals). In DeepSE, we use a heuristic method based on BERT's internal attention-maps to reduce M to a small constant amount by selecting the top m words ($m \ll M$) with high attention scores. The total number of counterfactuals is limited to $N*m$. This reduces the counterfactual engine response to near real time (0.83s for 61 articles in Case Study 1).

Contextual Explanation: It is worth noting that SE can be used independently to explain DR models using counterfactuals. In DR systems without SI, SE can still assist users in understanding the structure of data projection. However, the design motivation of SE is to provide a contextual explanation of interactively updated DR models, which is similar to SI. During human-in-the-loop analysis, the visual explanation of models should be contextual to the human-AI conversation and consistent with user interactions [25]. We hope SE will encourage more future designs of the contextual property of visual explanation for interactive machine learning systems.

Future Work: There are some limitations in our current design of the SE pipeline. More advanced counterfactual methods could be explored to provide more accurate explanations, such as hierarchical explanations [10] and complex constraints [31]. In addition, we do not take projection distortion introduced by counterfactual instances into consideration. Out-of-sample extension methods [2] could be applied to the system to alleviate this problem by projecting counterfactual instances into the pre-computed data projection [7].

Conclusion: We proposed SE to generate and contextualize explanations for visual analytics systems with nonlinear DR. SE is designed as an output analogy to semantic interaction input. The SE pipeline contains two main components: the counterfactual engine and the counterfactual projection. These key elements forge natural connections between cognitive projections and computational projections, thereby yielding contextual explanations. We implemented DeepSE, a visual text analysis system with nonlinear DR powered by deep learning. We demonstrated the utility of DeepSE via two case studies. With SE, analysts can better understand the updated DR model while maintaining focus on the analytic task.

Acknowledgments

This work was supported in part by NSF I/UCRC CNS-1822080 via the NSF Center for Space, High-performance, and Resilient Computing (SHREC).

REFERENCES

[1] C. Andrews, A. Endert, and C. North. Space to think: large high-resolution displays for sensemaking. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 55–64, 2010.

[2] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. L. Roux, and M. Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, p. 177–184. MIT Press, Cambridge, MA, USA, 2003.

[3] Y. Bian and C. North. Deepsi: Interactive deep learning for semantic interaction. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, IUI '21, p. 177–188. Association for Computing Machinery, College Station, TX, USA, 2021. doi: 10.1145/3397481.3450670

[4] Y. Bian, J. Wenskovitch, and C. North. Deepva: Bridging cognition and computation through semantic interaction and deep learning. In *Proceedings of the IEEE VIS Workshop MLUI 2019: Machine Learning from User Interactions for Visualization and Analytics. VIS'19.*, 10/2019 2019.

[5] L. Bradel, C. North, L. House, and S. Leman. Multi-model semantic interaction for text analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 163–172. IEEE.

[6] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. Dis-function: Learning distance functions interactively. *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 83–92, 2012. doi: 10.1109/VAST.2012.6400486

[7] M. Cavallo and Ç. Demiralp. A visual interaction framework for dimensionality reduction based data exploration. In R. L. Mandryk, M. Hancock, M. Perry, and A. L. Cox, eds., *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*. ACM, 2018. doi: 10.1145/3170427.3186508

[8] Y. Chan, C. D. Correa, and K. Ma. Flow-based scatterplots for sensitivity analysis. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pp. 43–50, Oct 2010. doi: 10.1109/VAST.2010.5652460

[9] A. Chatzimparmpas, R. M. Martins, and A. Kerren. t-visne: Interactive assessment and interpretation of t-sne projections. *IEEE Transactions on Visualization and Computer Graphics*, 26(8):2696–2714, 2020. doi: 10.1109/TVCG.2020.2986996

[10] H. Chen, G. Zheng, and Y. Ji. Generating hierarchical explanations on text classification via feature interaction detection. *arXiv preprint arXiv:2004.02015*, 2020.

[11] F. Cheng, Y. Ming, and H. Qu. Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1438–1447, 2021. doi: 10.1109/TVCG.2020.3030342

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[14] A. Endert, P. Fiaux, and C. North. Semantic interaction for sensemaking: Inferring analytical reasoning for model steering. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2879–2888, Dec 2012. doi: 10.1109/TVCG.2012.260

[15] A. Endert, S. Fox, D. Maiti, S. Leman, and C. North. The semantics of clustering: Analysis of user-generated spatializations of text documents. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, p. 555–562. Association for Computing Machinery, New York, NY, USA, 2012. doi: 10.1145/2254556.2254660

[16] R. Faust, D. Glickenstein, and C. Scheidegger. DimReader: Axis lines that explain non-linear projections. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):481–490, 2018. doi: 10.1109/tvcg.2018.2865194

[17] A. Ghosh, M. Nashaat, J. Miller, and S. Quader. Interpretation of structural preservation in low-dimensional embeddings. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2020. doi: 10.1109/TKDE.2020.3005878

[18] O. Gomez, S. Holter, J. Yuan, and E. Bertini. Vice: Visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, p. 531–535. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3377325.3377536

[19] T. M. Green, W. Ribarsky, and B. Fisher. Building and applying a human cognition model for visual analytics. *Information visualization*, 8(1):1–13, 2009.

[20] L. House, S. Leman, and C. Han. Bayesian visual analytics: Bava. *Stat. Anal. Data Min.*, 8(1):1–13, Feb. 2015. doi: 10.1002/sam.11253

[21] Y.-a. Kang and J. Stasko. Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 21–30, 2011. doi: 10.1109/VAST.2011.6102438

[22] C. C. Marshall, F. M. Shipman III, and J. H. Coombs. Viki: Spatial hypertext supporting emergent structure. In *Proceedings of the 1994 ACM European conference on Hypermedia technology*, pp. 13–23, 1994.

[23] D. Martens and F. Provost. Explaining data-driven document classifications. *MIS Q.*, 38(1):73–100, Mar. 2014. doi: 10.25300/MISQ/2014/38.1.04

[24] A. G. Martínez, B. T. Wooton, N. Kirshenbaum, D. Kobayashi, and J. Leigh. Exploring Collections of research publications with Human Steerable AI. pp. 339–348, 2020. doi: 10.1145/3311790.3396646

[25] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2018. doi: 10.1016/j.artint.2018.07.007

[26] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144. ACM, New York, NY, USA, 2016. doi: 10.1145/2939672.2939778

[27] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim. Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE transactions on visualization and computer graphics*, 23(1):241–250, 2016.

[28] J. Z. Self, M. Dowling, J. Wenskovitch, I. Crandell, M. Wang, L. House, S. Leman, and C. North. Observation-level and parametric interaction for high-dimensional data analysis. *ACM Trans. Interact. Intell. Syst.*, 8(2):15:1–15:36, June 2018. doi: 10.1145/3158230

[29] F. M. Shipman III and R. McCall. Supporting knowledge-base evolution with incremental formalization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 285–291. ACM, 1994.

[30] J. Stahnke, M. Dörk, B. Müller, and A. Thom. Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):629–638, 2016. doi: 10.1109/TVCG.2015.2467717

[31] S. Verma, J. Dickerson, and K. Hines. Counterfactual explanations for machine learning: A review, 2020.

[32] S. Wachter, B. D. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017.

[33] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, et al. Cord-19: The covid-19 open research dataset. *ArXiv*, 2020.

[34] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, Jan 2020. doi: 10.1109/TVCG.2019.2934619