Machine Learning from Interactions in Multi-Model Visual Analytics

John Wenskovitch
Chris North
Virginia Tech
Blacksburg, VA, USA
{jw87,north}@cs.vt.edu

ABSTRACT

We discuss visualization and interaction challenges that arise in human-in-the-loop visual analytics systems, with focus on those with multiple computational models in the pipeline between data and visualization. A fundamental challenge in such systems lies in disambiguating interactions, mapping interactions to the appropriate model, and determining how systems learn this disambiguation from user interactions. Similar challenges exist in determining the visual style of a system, determining the tasks that the system supports, and in the algorithms selected for the visualization system.

CCS CONCEPTS

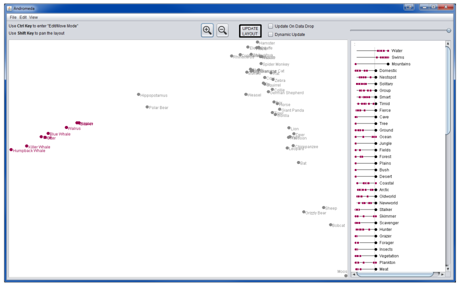
• **Human-centered computing → Visual analytics**; *Interaction techniques*; *Interaction design*;

KEYWORDS

Dimension reduction; clustering; visual analytics; interaction; human-centered machine learning

ACM Reference Format:
John Wenskovitch Chris North. 2019. Machine Learning from Interactions in Multi-Model Visual Analytics. In *Proceedings of ACM CHI Conference on Human Factors in Computing Systems, "Human-Centered Machine Learning Perspectives" Workshop (CHI'19).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.000/000_0

CHI'19, May 2019, Glasgow, UK



Use Ctrl Key to enter "Edit/Move Mode"
Use Shift Key to pan the layout

☐ Update On Data Drop
☐ Dynamic Update

Figure 1: Andromeda, an interactive visual analytics tool that enables the exploration of high-dimensional quantitative data projections.

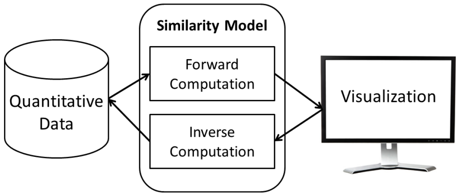


Figure 2: A pipeline representation of Andromeda, with a Similarity Model connecting the dataset and the visualization.

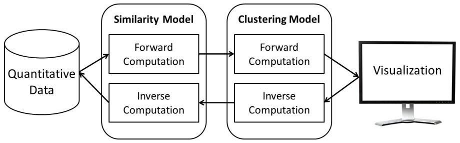


Figure 3: A multi-model pipeline representation of a system that incorporates both dimension reduction (similarity) and clustering in a single system.

INTRODUCTION

In recent years, analysts have worked to explore and draw conclusions from increasingly larger datasets, growing both in cardinality and dimensionality. Visual metaphors for exploring high-dimensional datasets come in a variety of forms, each with their own strengths and weaknesses in both visualization and interaction [7, 11]. One frequently-used method of visual abstraction is to reduce a high-dimensional dataset into a low-dimensional space while preserving properties of the high-dimensional structure (e.g., clusters and outliers) [8]. Because many dimension reduction algorithms rely on a "proximity ≈ similarity" metaphor, **implicit** clusters of similar items naturally begin to form in the projection. This stands in contrast to **explicit** clusters as identified by clustering algorithms. An example of these implicit clusters is shown in Figure 1, a visual analytics tool presented at the previous CHI HCML workshop by Self et al [13]. This tool, Andromeda (Figure 1), provides analysts with the ability to interactively explore projections of high-dimensional data by learning weights on the dimensions of the dataset. Of particular interest to HCML researchers is the process of learning weights in response to user interactions, termed *observation-level interaction* [5].

Using the pipeline representation from Dowling et al. [4], Andromeda can be represented as a single Similarity Model that bridges the dataset and visualization (Figure 2). In this representation, the Forward Computation of the Similarity Model handles the data projection, while the Inverse Computation supports the weight learning stage. Despite its computational simplicity, a significant number of visualization and interaction challenges exist in such a system, as further detailed by Self et al [14]. One notable challenge is the disambiguation of the interaction, or in other words, when an analyst repositions an observation in the projection, what are they moving it with respect to?

Say that we wish to update Andromeda to support **explicit** clustering in the projection, identifying clusters and allowing analysts to interact with them. Naturally, a pipeline that incorporates multiple models (or a multi-model pipeline) will result in a broader set of challenges, expanding the "with respect to what" problem and introducing further issues. Indeed, the visualization and interaction challenges involved in such a system expand greatly as models are added to the computational pipeline. Such multi-model systems are becoming prevalent in visual analytics research, and include but are not limited to other combinations such as relevance and similarity [2] and sampling and projection [10]. In this work, we detail multi-model system challenges, specifically using a dimension reduction and clustering pipeline (Figure 3) and system (Figure 4) as a representative example [15].

DIMENSION REDUCTION AND CLUSTERING AS A COMPLEX, MULTI-MODEL EXAMPLE

Though dimension reduction and clustering algorithms serve different cognitive purposes (spatializing and grouping, respectively), we noted in the previous section that they can often take on similar effects within a projection – groups of points that are similar as a result of the dimension reduction

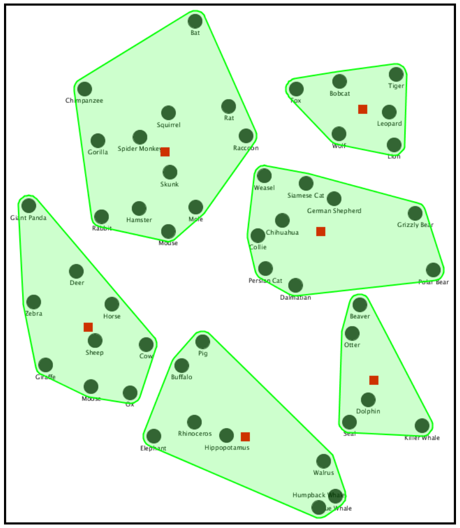


Figure 4: A system incorporating dimension reduction and clustering in a single system using convex hulls to represent clusters [17].

distance function will naturally begin to form clusters. However, clusters are inherently subjective structures, making their identification by both humans and machines a challenging process. Previous research has shown that humans use a variety of organizational principles to cluster information [3], even when addressing the same task [1]. In order to computationally identify clusters, hundreds of clustering algorithms have been implemented, each with strengths and weaknesses. Indeed, there is no universally-optimal clustering algorithm. Instead, the best clustering algorithm to solve a problem is often determined experimentally [6]. Therefore, introducing a clustering algorithm to explicitly define these implicit clusters (Figure 4) in a dimension-reduced projection presents challenges for visualization developers and designers.

Though the underlying cognitive actions of grouping and spatializing are different, the result of the corresponding algorithms in a projection can be similar. As a result, the combination of dimension reduction and clustering algorithms into a single computational pipeline results in interaction ambiguity that must be resolved.

INTERACTION CHALLENGES

Disambiguating Interactions on Observations

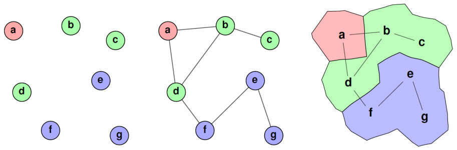
The introduction of a clustering algorithm and explicit clusters into an interactive dimensionally-reduced projection further complicates the “With respect to what” problem identified by Self et al [13]. Instead of simply identifying a corresponding observation or set of observations to complete the interpretation of the interaction, manipulation of cluster membership must be taken into account. Our previous work [17] attempts to resolve this issue by only treating reclassification interactions as meaningful to the computational backend, but the general challenge to disambiguate such interactions is much broader [16]. We suggest the following dimensions to consider when interpreting the intent of an interaction:

- **Interaction Target:** The interaction could be applied to the observations, the clusters, or both.
- **Cardinality:** The interaction could be applied to the nearest observation, the nearest *n* observations, a cluster, or all observations in the projection, among other possibilities.
- **With Respect to What:** Is the important relationship relative to other observations in the projection at the source of the interaction, the destination of the interaction, or both?
- **Level of Thinking:** When performing the interaction, is the analyst thinking high- or low-dimensionally? In other words, is the analyst merely altering the projection, or are they considering all properties of a group of observations?
- **Visual Design:** Is the intent of the interaction influenced by the way that observations and clusters are encoded in the visualization?

With these dimensions in mind, a designer can better strive to better map interactions to intent.

Table 1: A set of cluster-specific interactions that include some inherent ambiguity.

Interaction
Cluster Change in Membership
Move cluster into cluster
Move cluster out of cluster
Move cluster between clusters
Move cluster external to clusters
Move cluster within a cluster
Join/Split Clusters
Join Clusters
Split Clusters
Create/Remove Clusters
Create Cluster
Remove Cluster



(a) N (b) NL (c) NLG

Figure 5: Three options for encoding cluster membership, as studied by Saket et al [12].

Disambiguating Cluster-Specific Interactions

The previous research challenge is focused on interactions with observations and the responses that address those interactions, but the introduction of explicit clustering further leads to a set of cluster-specific interactions that should be supported. Some of these are listed in Table 1. Disambiguating the intent of such interactions is a related but separate challenge, especially in hierarchical clustering systems. For example, if an analyst drags one cluster into another, is their intention to join the clusters into a single structure or is it to make one cluster a child of the other?

VISUALIZATION CHALLENGES

The Effect of Clustering Representation

Some previous work has addressed options for encoding cluster membership. Saket et al. [12] evaluate three options for encoding cluster membership (node, node-link, and node-link-group), relating each to the effectiveness of performing node- and group-based tasks (Figure 5). Similarly, Jianu et al. [9] evaluate the Linesets, GMap, and BubbleSets encoding techniques, along with a more traditional node-link diagram. However, these studies resulted in some conflicting findings. In particular, Saket found that the addition of group encodings does not negatively impact time and accuracy on node-based tasks, but Jianu saw a 25% time penalty on such tasks. The fact that such conflicting results appear in a single-model clustering system will lead to larger complications in a multi-model system.

Ordering the Computational Models

The order in which models appear in the computational pipeline can also have effects on both the visualization and the supported interactions, even if the same visual interface can support multiple backend pipelines. Figure 6 shows six possibilities for ordering just these two models in a computational pipeline. For example, the Independent Algorithms pipeline computes both the position in the projection and the cluster membership of each observation using only the high-dimensional data. The Dimension Reduction Preprocessing for Clustering pipeline, in contrast, projects the high-dimensional data into its low-dimensional form, and then clusters the data based on the reduced-dimensionality representation. This can enable the clustering algorithm to run more quickly and may produce more compact clusters visually, though this comes with a slight loss in clustering accuracy. As a third option, the clustering algorithm can run on the high-dimensional data and assign each observation to a cluster, with the cluster centroids then being the primary layout mechanism in the low-dimensional projection. This can enable more rapid projections while providing accurate cluster membership assignments, though with possible inaccuracies in precise observation positions in the projection.

The variety of computational pipelines further leads to difficulties in inferring the intent of an analyst interaction. For example, repositioning an observation from one cluster to another in a Dimension

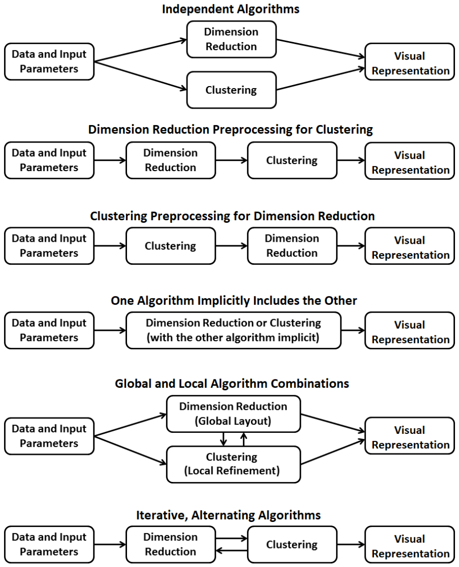

Independent Algorithms
Data and Input Parameters
Dimension Reduction
Clustering
Visual Representation
Dimension Reduction Preprocessing for Clustering
Data and Input Parameters
Dimension Reduction
Clustering
Visual Representation
Clustering Preprocessing for Dimension Reduction
Data and Input Parameters
Clustering
Dimension Reduction
Visual Representation
One Algorithm Implicitly Includes the Other
Data and Input Parameters
Dimension Reduction or Clustering (with the other algorithm implicit)
Visual Representation
Global and Local Algorithm Combinations
Data and Input Parameters
Dimension Reduction (Global Layout)
Clustering (Local Refinement)
Visual Representation
Iterative, Alternating Algorithms
Data and Input Parameters
Dimension Reduction
Clustering
Visual Representation

Figure 6: Six methods for ordering dimension reduction and clustering algorithms in a computational pipeline for visual analytics [15].

Reduction Preprocessing for Clustering system may imply an inaccuracy in the low-dimensional assignment, a result of the clustering computation running on the low-dimensional data. The result of such an interaction may be an update to the projection weights to correct for this misclassification. However, the same reclassification interaction in a Clustering Preprocessing for Dimension Reduction system would imply an actual misclassification of that observation in the high-dimensional space, leading to the need to reweight the dimensions applied at the clustering stage to make this correction.

RESEARCH AGENDA

Beyond the challenges articulated above, much more research is needed in the area of human-in-the-loop visual analytics, regardless of whether the systems are single- or multi-model. Indeed, space limitations here prevented the discussion of challenges that exist in which algorithms are selected, which tasks are supported, and how to best evaluate such systems. At a high level, the overarching research question is focused on capturing the cognitive intent of an analyst's interactions and responding appropriately. In particular, we note some open research questions here:

- What kinds of interactions can best provide feedback to machine learning algorithms?
- What can machine learning algorithms learn from interactions?
- How can machine learning algorithms be best designed to enable user interaction and feedback?
- How can visualizations and interactions be designed to best exploit machine learning algorithms?
- How can visualization system architectures be best designed to support machine learning?
- How should a designer manage conflicts between the analyst's intent and the data or machine learning algorithm capabilities?
- How can we evaluate systems that incorporate both machine learning algorithms and user interaction training together?
- How can machine learning and user interaction together make both computation and user cognition more efficient?
- How can a designer best support the sensemaking process by learning from user interaction?

ACKNOWLEDGMENTS

This work was funded by NSF grant IIS-1447416. Members of the Virginia Tech InfoVis@VT and BaVA research groups provided useful feedback towards developing and structuring these ideas.

REFERENCES

[1] Christopher Andrews, Alex Endert, and Chris North. 2010. Space to Think: Large High-resolution Displays for Sensemaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 55–64. https://doi.org/10.1145/1753326.1753336

[2] L. Bradel, C. North, L. House, and S. Leman. 2014. Multi-model semantic interaction for text analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 163–172. https://doi.org/10.1109/VAST.2014.7042492

[3] Paul Dourish, John Lamping, and Tom Rodden. 1999. Building Bridges: Customisation and Mutual Intelligibility in Shared Category Management. In *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work (GROUP '99)*. ACM, New York, NY, USA, 11–20. https://doi.org/10.1145/320297.320299

[4] M. Dowling, J. Wenskovitch, P. Hauck, A. Binford, N. Polys, and C. North. 2018. A Bidirectional Pipeline for Semantic Interaction. In *Proceedings of the Workshop on Machine Learning from User Interaction for Visualization and Analytics*. 11.

[5] A. Endert, C. Han, D. Maiti, L. House, S. Leman, and C. North. 2011. Observation-level interaction with statistical models for visual analytics. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 121–130. https://doi.org/10.1109/VAST.2011.6102449

[6] Vladimir Estivill-Castro. 2002. Why So Many Clustering Algorithms: A Position Paper. *SIGKDD Explor. Newsl.* 4, 1 (June 2002), 65–75. https://doi.org/10.1145/568574.568575

[7] Usama M Fayyad, Andreas Wierse, and Georges G Grinstein. 2002. *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann.

[8] I K Fodor. 2002. *A Survey of Dimension Reduction Techniques*. https://doi.org/10.2172/15002155

[9] R. Jianu, C. Demiralp, and D. Laidlaw. 2009. Exploring 3D DTI Fiber Tracts with Linked 2D Representations. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (Nov 2009), 1449–1456. https://doi.org/10.1109/TVCG.2009.141

[10] G. M. H. Mamani, F. M. Fatore, L. G. Nonato, and F. V. Paulovich. 2013. User-driven Feature Space Transformation. *Computer Graphics Forum* 32, 3.3 (2013), 291–299. https://doi.org/10.1111/cgf.12116

[11] Tamara Munzner. 2014. *Visualization Analysis and Design*. CRC Press.

[12] B. Saket, P. Simonetto, S. Kobourov, and K. BÃűrner. 2014. Node, Node-Link, and Node-Link-Group Diagrams: An Evaluation. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec 2014), 2231–2240. https://doi.org/10.1109/TVCG.2014.2346422

[13] Jessica Zeitz Self, Xinran Hu, Leanna House, Scotland Leman, and Chris North. 2016. Designing Usable Interactive Visual Analytics Tools for Dimension Reduction. In *CHI 2016 Workshop on Human-Centered Machine Learning (HCML)*. 7.

[14] Jessica Zeitz Self, Radha Krishnan Vinayagam, J. T. Fry, and Chris North. 2016. Bridging the Gap Between User Intention and Model Parameters for Human-in-the-loop Data Analytics. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics (HILDA '16)*. ACM, New York, NY, USA, Article 3, 6 pages. https://doi.org/10.1145/2939502.2939505

[15] John Wenskovitch, Ian Crandell, Naren Ramakrishnan, Leanna House, Scotland Leman, and Chris North. 2018. Towards a Systematic Combination of Dimension Reduction and Clustering in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics Proceedings of the Visual Analytics Science and Technology 2017* 24, 1 (Jan 2018), 131–141. https://doi.org/10.1109/TVCG.2017.2745258

[16] John Wenskovitch, Michelle Dowling, and Chris North. 2019. Simultaneous Interaction with Dimension Reduction and Clustering Projections. In *24th International Conference on Intelligent User Interfaces (IUI '19 companion)*. ACM, New York, NY, USA, 2. https://doi.org/10.1145.3308557.3308718

[17] John Wenskovitch and Chris North. 2017. Observation-Level Interaction with Clustering and Dimension Reduction Algorithms. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics (HILDA'17)*. ACM, New York, NY, USA, Article 14, 6 pages. https://doi.org/10.1145/3077257.3077259