The State of the Art in Integrating Machine Learning into Visual Analytics

A. Endert, W. Ribarsky, C. Turkay, W Wong, I. Nabney, I Díaz Blanco, Fabrice Rossi

Volume xx (200y), Number z, pp. 1–29

The State of the Art in Integrating Machine Learning into Visual Analytics

A. Endert[1], W. Ribarsky[2], C. Turkay[3], W. Wong[4], I. Nabney[5], I. Díaz Blanco[6], F. Rossi[7]

[1]Georgia Tech, USA
[2]University of North Carolina, Charlotte, USA
[3]City University of London, UK
[4]Middlesex University, UK
[5]Aston University, UK
[6]University of Oviedo, Spain
[7]Paris 1 Panthéon Sorbonne University, Paris

Abstract

Visual analytics systems combine machine learning or other analytic techniques with interactive data visualization to promote sensemaking and analytical reasoning. It is through such techniques that people can make sense of large, complex data. While progress has been made, the tactful combination of machine learning and data visualization is still under-explored. This state-of-the-art report presents a summary of the progress that has been made by highlighting and synthesizing select research advances. Further, it presents opportunities and challenges to enhance the synergy between machine learning and visual analytics for impactful future research directions.

Categories and Subject Descriptors (according to ACM CCS): Human-centered computing - Visualization, Visual analytics

1. Introduction

We are in a data-driven era. Increasingly more domains generate and consume data. People have the potential to understand phenomena in more depth using new data analysis techniques. Additionally, new phenomena can be uncovered in domains where data is becoming available. Thus, making sense of data is becoming increasingly important, and this is driving the need for systems that enable people to analyze and understand data.

However, this opportunity to discover also presents challenges. Reasoning about data is becoming more complicated and difficult as data scales and complexities increase. People require powerful tools to draw valid conclusions from data, while maintaining trustworthy and interpretable results.

We claim that visual analytics (VA) and machine learning (ML) have complementing strengths and weaknesses to address these challenges. Visual analytics (VA) is a multi-disciplinary domain that combines data visualization with machine learning (ML) and other automated techniques to create systems that help people make sense of data [TC05, KSF*08, Kei02, KMSZ06]. Over the years, much work has been done to establish the foundations of this area, create research advances in select topics, and form a community of researchers to continue to evolve the state of the art.

Currently, VA techniques exist that make use of select ML models or algorithms. However, there are additional techniques that can apply to the broader visual data analysis process. Doing so reveals opportunities for how to couple user tasks and activities with such models. Similarly, opportunities exist to advance ML models based on the cognitive tasks invoked by interactive VA techniques.

This state-of-the-art report briefly summarizes the advances made at the intersection of ML and VA. It describes the extent to which machine learning methods are utilized in visual analytics to date. Further, it illuminates the opportunities within both disciplines that can drive important research directions in the future. Much of the content and inspiration for this paper originated during a Dagstuhl Seminar titled, "Bridging Machine Learning with Information Visualization (15101)" [KMRV15].

1.1. Report organization

This report is organized as follows. Section 2 of the report discusses three categories of models: human reasoning, visual analytics and information visualization, and machine learning. The models describing the cognitive activity of sensemaking and analytical reasoning characterize the processes that humans engage in cognitively to gain understanding of data. The models and frameworks for visual analytics depict systematic descriptions of how computation and analytics can be incorporated in the systematic construction and design of visual analytic applications. Finally, the machine learning community has several models that illustrate how models are trained, used, and interactively steered.

Section 3 categorizes the integration of machine learning techniques into visual analytic systems. Section 4 discusses how such systems have been used in specific domains to solve real-world challenges. Section 5 discusses a research direction for integrating steerable dimension reduction techniques into visual analytics. Finally, Section 6 discusses open challenges and opportunities for ML and VA. While the current work shows how some progress has been made in bringing these two communities closer together, there are several open challenges.

2. Models and Frameworks

To ground the discussion of embedding ML techniques into VA systems for data analysis and knowledge discovery, we describe three categories of models and frameworks below. First, we discuss existing models meant to describe the cognitive stages people progress through while analyzing data. These models show the complex processes people go through to gain insight from data, which developed systems must support. Second, we discuss existing models and frameworks that describe interaction and information design of visual analytic applications. These models illustrate how data transformation and analytic computation are involved in generating the visual representations of data in tools. User interaction is critical in tuning and steering the parameters of these models. Finally, we show select ML frameworks that emphasize the importance of training data and ground truth for generating accurate and effective computational models. In addition, we describe the main techniques developed in the ML field to integrate user feedback in the training process.

2.1. Models of Sensemaking and Knowledge Discovery

One should emphasize that a primary purpose of data analytics is for people to understand, and gain insights into, their data [CMS99, Chr06]. Thus, it is important to understand the cognitive processes of people as they reason about data. It is from such an understanding that "human-in-the-loop" application designs are realized. Prior work exists that provides models and design guidelines for visual analytics.

Sense-making is the process of "structuring the unknown"

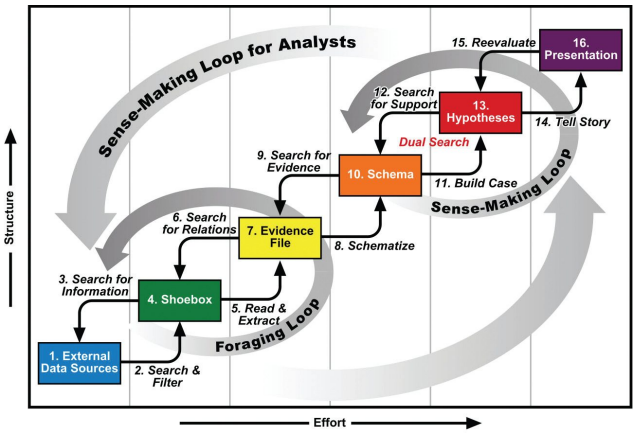


Figure 1: *The "sensemaking loop" (from [PC05]) illustrating the cognitive stages people go through to gain insight from data.*

by organising data into a framework that enables us "to comprehend, understand, explain, attribute, extrapolate, and predict" [Anc12]. It is this activity of structuring–the finding and assembly of data into meaningful explanatory sequences [LI57]–that enables us to turn ever more complex observations of the world into findings we can understand "explicitly in words and that serves as a springboard into action" [WSO05]. By attempting to articulate the unknown, we are driven more by "plausibility rather than accuracy" [Wei95] as we create plausible explanations that can be used to evolve and test our understanding of the situation or the data. Decision makers are often faced with inaccurate representations of the world [EPT*05] and have to fill-in the gaps with strategies such as "story-telling" to create stories that explain the situation.

One of the earliest models to describe the iterative process of data analysis as "sensemaking" [RSPC93] is presented in Figure 1 and illustrates the well-known (and probably the most frequently cited) Pirolli and Card sensemaking model [PC05]. Proposed in the context of intelligence analysis, it is useful for showing how information is handled through the process of searching and retrieving relevant information, organizing, indexing and storing the information for later use, structuring the information to create a schema or a way to explain what has been observed, the formulation and testing of hypotheses, which then leads to the determination of a conclusion, and a sharing of that conclusion. This notional model depicts the cognitive stages of people as they use visual analytic tools to gain understanding of their data.

From Pirolli and Card's perspective, sensemaking can be categorized into two primary phases: foraging and synthesis. Foraging refers to the stages of the process where models filter and users gather collections of interesting or relevant information. This phase emphasizes the computational ability of models, as the datasets are typically much larger than what a user can handle. Then, using that foraged information, users advance through the synthesis stages of the process, where they construct and test hypotheses about how the

foraged information may relate to the larger plot. In contrast to foraging, synthesis is more "cognitively intensive", as much of the insights stem from the user's intuition and domain expertise. Most existing visualization tools focus on either foraging or synthesis, separating these two phases.

As with all models of cognitive processes, there have been criticisms. For instance, while there are feedback loops and repeat loops, and cycles within cycles, it still is somewhat a linear model. It describes the data transaction and information handling and transformation processes, "... rather than how analysts work and how they transition" [KS11]. Human analysts carry out their work within this framework, but their thinking and reasoning processes are much less linear and structured. For example, although recognised as a part of the sense-making loop, there is little explanation about the thinking and reasoning strategies that are invoked to formulate hypotheses. This is a critical aspect of the sense-making process: how are explanations of the situation or data formed in the mind of the human in order that the explanation can be used to test one's understanding of the data or situation? Later in this section, we report on work that is attempting to unravel this aspect of how analysts think.

Another useful model that can be employed to describe the human-centered sense-making process is the "data-frame model" by Klein et al. [KMH06b, KMH06a]. Their model (Figure 2) depicts an exchange of information between the human and the data in terms of frames. People make sense of a situation by interpreting the data they are presented with in relation to what they already know to create a new understanding. A user has an internal "frame" that represents her current understanding of the world. The data connects with the frame. As she continues to explore a particular dataset, her frames of the world are mapped against the information she uncovers. If the information supports a specific frame, that frame is thought to strengthen in a process they call elaboration. As she understands the situation better, she searches for more relevant information, learning that there may be other factors to the problem than originally thought or known, therefore driving the demand for more information, and building her frame. However, when evidence is discovered through exploration that contradicts or refutes the existence of such a mental frame, the frame can either be augmented or a new one created. This is the important process that leads her to question her earlier conclusions or assumptions made to arrive at these conclusions. Additionally new frames can also be created to reframe the problem. In situations where data is missing or ambiguous or unknown, reframing enables her to articulate the problem in different ways that may allow her to change her information search strategy and perhaps even her goals. One of the key benefits of the Data-Frame Model is that it points to the importance of designing visual analytics in a way that encourages analysts to question their data and their understanding, and to facilitate visualizations and transformations that enable reframing of their understanding of the situation.

Recently a set of knowledge generation and synthesis models have been proposed that comprehensively attack a central issue of visual analytics: developing a human-computer system that enables analytic reasoning to produce actionable knowledge. The first of these models was proposed by Sacha et. al. [SSS*14] and is shown in Figure 3. One sees looping structures and components familiar from Pirolli and Card's sensemaking model, as depicted in Figure 1 above. However, the computer and human regions of the model, and their relationship with each other, are now explicitly expressed, and the paper shows a clear relationship, via interaction, between the human and both the visualization and the model. The paper also describes detailed steps for the data-visualization and data-model pipelines (the latter in terms of KDD processes that couple, for example, to machine learning algorithms). Whereas the sensemaking model was conceptual, this model is concrete and shows, better than other models, where to put computing and (via interactive interfaces) human-in-the-loop steps in order to build an actual system.

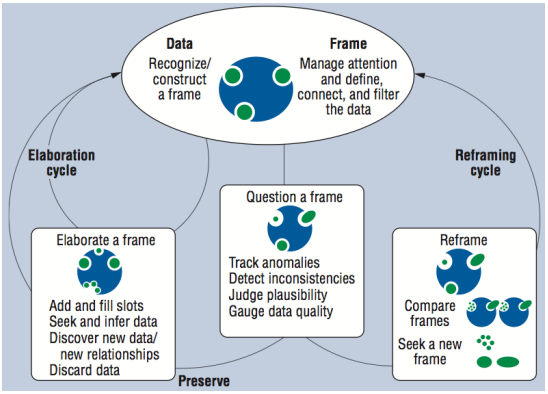


Figure 2: *The Data-Frame Model of Sensemaking [KMH06b].*

The Sacha et al. model has recently been generalized to produce a more complete knowledge generation and synthesis (KGS) model [RF16]. The KGS model explicityly accounts for both Prior Knowledge (placed between Data, Visualization, and Model in Figure 3) and User Knowledge (placed between Action and Finding). Prior Knowledge is quite important for any exploration involving experts or based on expertise; experts will want to know immediately the relationship of new knowledge to existing domain knowledge. User knowledge is built up during complex reasoning, where it can then be the basis for generating additional knowledge or can be synthesized with Prior Knowledge to produce more general truths. The KGS model posits an iterative process that addresses high level reasoning, such as inductive, deductive, and abductive reasoning, in the knowledge generation and exploration loops. It is based on a framework by Gahegan et al. [GWHR01] that was developed for GIScience but is generalizable.

These models provide a roadmap for visualization and analytics processes, and for the role of human-computer interaction. In particular, they illuminate the relationships among machine learning, visualization, and analytics rea-

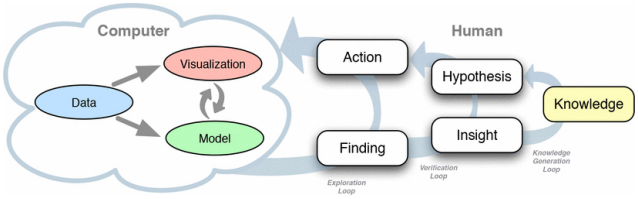


Figure 3: *Human-Computer knowledge generation model of Sacha et al. [SSS*14].*

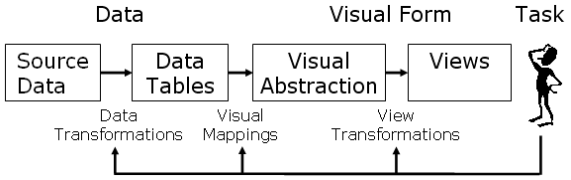


Figure 4: *The information visualization pipeline [Hee06] depicting the data transformation and visual mapping process for constructing visualizations.*

soning processes including exploration and knowledge generation. For example, Klein's data frame model, discussed above, would fit in this structure, providing a focus for ML components while the models in Figure 3 would show how to connect the data frame model with interactive visualization and hypothesis-building. There are no VA systems that embody all the components of the Sacha and KGS models, but there are some (e.g., VAiRoma [CDW*16]) that include parts of the model. Typically in these systems, ML is a static pre-processing step applied to the data at the beginning. For example, in VAiRoma time-dependent, hierarchical topic modeling is applied to large text collections [CDW*16]. However, the KGS model shows how interactive ML can be placed in the human-computer process and how it relates to interactive visualization and reasoning. There is further discussion of interactivity in VAML systems below. The discussion in Sacha et al. [SSS*14] implies two main roles for ML; one is to transform unstructured or semi-structured data into a form more meaningful for human exploration and insight discovery. The other is to use unsupervised or semi-supervised ML to guide the analysis itself by suggesting the best visualizations, sequences of steps in the exploration, verification, or knowledge generation processes, guarding against cognitive bias, etc. In addition, since the KGS model was derived with reference to cognitive science principles [GRF09], there is a possibility for merging ML with cognitive models to produce even more powerful human-machine models. To illustrate, one could explore Fu and Pirolli's SNIF-ACT cognitive architecture model [FP07], which connects human exploration and information foraging in a sensemaking context. This could be married with ML approaches to refine and focus the parameters of the ML approach for particular exploration strategies.

2.2. Models of Interactivity in Visual Analytics

Frameworks or pipelines for information visualization have been previously developed [Hee06, Van05]. For example, the information visualization pipeline depicted in Figure 5 shows how data characteristics are extracted and assigned visual attributes or encodings, ultimately creating a visualization. The designs of visualizations adhering to this pipeline exhibit two primary components of the visual interface: the visualization showing the information, and a graphical user interface (GUI) consisting of graphical controls or widgets. The graphical controls in the GUI (e.g., sliders, knobs, etc.) allow users to directly manipulate the parameters they control. For example, "direct manipulation" [Shn83] user interfaces for information visualizations enable users to directly augment the values of data and visualization parameters to see the corresponding change in the visualization (e.g., using a slider to set the range of home prices and observing the filtering of results in a map showing homes for sale). This model is a successful user interaction framework for information visualizations.

Visual analytic systems have adopted this method for user interaction, but with the distinct difference of including analytic models or algorithms, as discussed earlier in this section. For example, in addition to filtering the data by selecting ranges for home prices, users could be given graphical controls over model parameters such as weighting the mixture of eigenvectors of a principal component analysis (PCA) dimension reduction (DR) model to produce two-dimensional views showing pairwise similarity of homes across all of the available dimensions. To users who lack expertise in such models, this may pose fundamental usability challenges.

In contrast, prior work has proposed frameworks to perform model steering via machine learning techniques applied to the user interactions performed during visual data analysis, called semantic interaction [EFN12b]. Semantic interaction is an approach to user interaction for visual data exploration in which analytical reasoning of the user is inferred and in turn used to steer the underlying models implicitly (illustrated in Figure 5). The goal of this approach to user interaction is to enable co-reasoning between the human and the analytic model (or models) used to create the visualization (coupling cognition and computation) without requiring the user to directly control the models.

The approach of semantic interaction is to overload the visual metaphor through which the insights are obtained (i.e., the visualization of information created by computational models) and the interaction metaphor through which hypotheses and assertions are communicated (i.e., interaction occurs within the visual metaphor). Semantic interaction enables users to directly manipulate data within visualizations, from which tacit knowledge of the user is captured, and the underlying analytic models are steered. The analytic models can be incrementally adapted based on the user's sensemaking process and domain expertise explicated via the user interactions with the system (as described in the models of Section 2.1).

The semantic interaction pipeline (shown in Figure 5) takes an approach of directly binding model steering techniques to the interactive affordances created by the visualization.

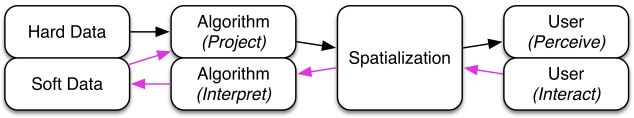


Figure 5: *The semantic interaction pipeline [EFN12b] showing how the user interactions in a spatial visualization can be incorporated into the computation of a visual analytic system.*

For example, a distance function used to determine the relative similarity between two data points (often visually depicted using distance in a spatial layout), can serve as the interactive affordance to allow users to explore that relationship. Therefore, the user interaction is directly in the visual metaphor, creating a bi-directional medium between the user and the analytic models [LHM*11].

2.3. Machine Learning Models and Frameworks

There is not as much work in machine learning models and frameworks. Most of the proposals correspond to some form of *de facto* industrial standards, such as the SEMMA (Sample, Explore, Modify, Model, and Assess) methodology advertised by SAS Institute Inc. Among those, a vendor neutral framework, CRISP-DM [She00], is somewhat comparable to knowledge discovery and visual analytics frameworks. There are six phases in the framework: business (or problem) understanding; data understanding (developed through exploration of the data and discussion with data owners); data preparation (including feature extraction, noise removal, and transformation); modeling; evaluation (testing the quality of the model, and particularly its generalization performance); deployment (embedding the model in practice). In some versions of this framework, there is an additional link from deployment back to business understanding - this represents the fact that the underlying data generator may change over time. The model needs continuous evaluation in deployment and when performance degrades, the process starts again. Perhaps more importantly, all the steps of the framework are embedded in a general loop comparable to the ones observed in other frameworks. This emphasize the feedback from the latter stage of the process (evaluation in numerous machine learning applications) to the early stages (e.g. data preparation in CRISP-DM).

As pointed out in e.g. [ACKK14], the traditional implementation of the machine learning workflow leads to long development cycles where end users (who are also domain experts) are asked to give feedback on the modeling results. This feedback is used by machine learning experts to tune the whole processing chain, especially at the data preparation stage. Ideally, this feedback should take the form of specific and formal user inputs, for example positive and negative feedback on exemplars (such as "those two objects should not belong to the same cluster" or "this object is misclassified").

User feedback in this formal, expressive form lends itself very well to steering and training machine learning models, for example via *interactive machine learning* approaches [PTH13]. Figure 6 shows an early model of interactive machine learning that emphasizes the feedback that users give to train classifiers [FOJ03]. Through multiple iterations of feedback, the classifier gets more training examples, and is thus able to more closely approximate the phenomena or concept being classified in the data.

To further establish an ML framework, we note the following. Machine learning tasks are traditionally divided into two broad categories, supervised tasks and unsupervised tasks. In supervised learning, the goal is to construct a model that maps an input to an output, using a set of examples of this mapping, the training set. The quality of the model is evaluated via a fixed loss criterion. Up till recently, it has generally been considered that human input is not needed in the model construction phase. On the contrary, it could lead to undetected overfitting. Indeed the expected quality of the model on future data (its so-called generalization ability) is generally estimated via an independent set of examples, the test set. Allowing the user (or a program) to tune the model using this set will generally reduce the generalization ability of the model and prevent any sound evaluation of this ability (unless yet another set of examples is available).

Supervision via examples can be seen as a direct form of user control over the training process. Allowing the user to modify the training set interactively provides an indirect way of integrating user inputs into the model construction phase. In addition, opportunities for user feedback and control are available before and after this modeling step (e.g., using the CRISP-DM phases). For instance, user feedback can be utilized at the feature selection, error preferences, and other steps. Leveraging those opportunities (including training set modification) has been the main focus of interactive machine learning approaches. For instance, tools such as the Crayons system from [FOJ03] allow the user to add new training data by specifying in a visual way positive and negative examples. This specific type of user feedback in the form of labelling new examples is exactly the focus of the *active learning* framework [Set09] in machine learning. This learning paradigm is a variation over supervised learning in which ML algorithms are able to determine interesting inputs for which they do not know the desired outputs (in the training set), in such a way that given those outputs the predictive performances of the model would greatly improve. Interestingly active learning is not the paradigm used in e.g. [FOJ03]. It seems indeed that in real world applications, active learning algorithms tend to ask too many questions and possibly to similar ones, as reported in e.g., [GB11]. More generally, the need for specific and formal user inputs can create usability issues with regards to people and their tasks, as pointed out in e.g., [ACKK14,EHR*14]. That is, the actions taken by the user to train the systems are often not the actions native to the exploratory data analysis described in the previously mentioned frameworks. This is starting to become more commonly used in the ML community, as exemplified by [BH12]. In this paper the authors consider additional questions a system can ask a user, beyond just labelling. They focus in particular on *class conditional queries*

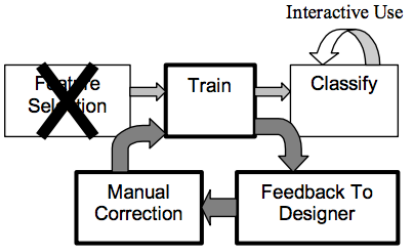


Figure 6: *A model for interactive machine [FOJ03] learning depicting user feedback for model training.*

– the system shows the user unlabeled examples and asks him or her to select one that belongs to a given class (if one exists).

In unsupervised learning, the data has no input/output structure and the general goal is to summarize the data in some way. For instance, as discussed further below, dimension reduction techniques build low dimensional approximations of the data from their high dimensional initial representation; clustering groups data into similar objects; etc. Unsupervised learning is generally considered ill posed in the ML field in the following sense: most of the tasks of unsupervised learning (clustering, dimensionality reduction, etc.) have only an informal description to which numerous formal models can be related. Those models are very difficult to compare on a theoretical point of view as well as on a practical one. In unsupervised learning, the need for user input, steering and control is therefore broadly accepted and techniques to include user feedback into e.g., clustering have been studied for some time. Variations over unsupervised methods that take explicitly into account some form of additional information are generally called semi-supervised methods. The supervision is frequently provided by external data in an automated way, but those methods can lead to principled ways of integrating user feedback.

It should be noted however that most of the methodological development in machine learning that can be used to integrate user feedback, from active learning to triplet based constraints [vdMW12], are seldom evaluated in the context of visualization systems. In general, the feedback process is either simulated or obtained via off line and slow process (e.g. Amazon's Mechanical Turk for triplet in [WKKB15]). Thus while specific frameworks that enable user feedback have been defined by the ML community, the practical relevance of the recent ones in the context of interactive visualization remains untested.

2.4. Comparison to another classification framework

A recent paper by Sacha et al. [SZS*16] overlaps with this STAR Report. It focuses on the specific area of dimensionality reduction and how these techniques integrate with interactive visualization in visual analytics systems. The paper builds around a systematic analysis of visualization literature, which reveals seven common interaction scenarios. The evaluation leads to the identification of future research opportunities.

The current paper provides a significantly broader survey of machine learning methods coupled with interaction, while Sacha et al. [SZS*16] probe deeper in one important area. In addition to dimension reduction, the current paper deals with ML methods for clustering, classification, and regression. There is some overlap in the literature covered in the two papers. However, the literature reviewed in the current paper cites ML methods that are already coupled with interactive visualization systems plus those that are not yet (but it would be beneficial if they were); Sacha et al. deal mostly with ML methods that are already coupled with interactive visualizations.

The two papers complement each other with Sacha's deeper analysis in DR strengthening the wider analysis in the current paper, and vice versa. The human-in-the-loop process model in [SZS*16] has similarities with the use of the human-machine interaction loop in the current paper; they also share a common origin. The classifications used in Sacha et al's structured analysis are different than those in the current paper's taxonomy, although one could be mapped into the other, with modifications. However, there are also multiple similarities; in particular, classification according to "modify parameters and computation domain" and "define analytical expectations" in Sections 3.2 and 3.3 of the current paper map to various interaction scenarios in Sacha et al. [SZS*16]. For example, the first classification maps to data manipulation, DR parameter tuning, and DR type selection scenarios in Sacha et al's model. The second classification, in permitting the user to tell the system (based on results it gives) expectations that are consistent with domain knowledge, maps to feature selection and emphasis and defining constraints scenarios. The current paper then goes beyond DR, including for each classification a discussion of clustering, classification, and regression methods. This broadens and strengthens the discussion from Sacha et al. [SZS*16].

3. Categorization of Machine Learning Techniques Currently used in Visual Analytics

The visual analytic community has developed systems that leverage specific machine learning techniques. In this section, we give an overview of the existing ways that machine learning has been integrated into VA applications from two transversal perspectives: the *types of ML algorithms* and the so-called *interaction intent*. We pay special attention to the "interaction intent" as described below, because this focuses on human-in-the-loop aspects that are central to VA systems. There are also other papers where the main role of visualization is on communicating the results of computations to improve comprehension [TJHH14] that are not directly covered in this section. Some of the most significant of these papers, referring to VA systems, are described in Section 4.

Along the first perspective, we consider the different *types of ML algorithms* that have been considered within visual analytics literature. Although one might think of several other

possible ways to categorize the algorithms [Alp14, FHT01], here we adopt a high-level task-oriented taxonomy and categorize the algorithms under the following headings: *dimension reduction*, *clustering*, *classification*, *regression/correlation analysis*. We observe that ML algorithms to tackle these tasks are frequently adopted in visual analytics applications since these analytical tasks often require the joint capabilities of computation and user expertise. To briefly summarize: i) *dimension reduction* methods help analysts to distill the information in high-dimensional data so that conventional visualization methods can be employed and important features are identified ii) *clustering* methods help to identify groups of similar instances which can be done both in a supervised or unsupervised manner iii) *classification* methods are often supervised and help to build models to associate labels to data instances, and finally iv) *regression/correlation* analysis methods help to investigate relations between features in the data and to understand/generate causal links to explain phenomena.

Along the second perspective, we focus on the user side of the process. We name this aspect as *interaction intent* and categorize the actions taken by users within visual analysis in terms of the methods through which the analyst tries to improve the ML result.

This perspective of our taxonomy resonates with the "*user intent*" categories suggested by Yi et al. [YaKSJ07] for low-level interactions within InfoVis applications. Our focus, however, is targeted on higher-level analytical intents within the narrower scope of visual analytics applications that involve ML methods. With this motivation in mind, we suggest two broad categories for "*intents*": *modify parameters and computation domain* and *define analytical expectations*. Table 1 shows the organization of literature along the dimensions of algorithm type vs the two categories of user intent. Here we survey the existing literature within the scope of this characterization.

3.1. Review Methodology

The literature summarized and categorized in this section are taken from impactful ML and visualization conferences and journals. They were chosen and categorized based on discussions the authors had at the Dagstuhl Seminar titled, "Bridging Machine Learning with Information Visualization (15101)" [KMRV15], and later refined through a more extensive literature review.

Within this report, we review existing literature on the integration of machine learning and visualisation from three different perspectives – models and frameworks, techniques, and application areas. When identifying the relevant works in these domains, we follow a structured methodology and identified the different scopes of investigation for these three different perspectives. One important note to make is, due to our focus on the integration of the two fields, we scanned resources from both the visualisation and machine learning domain.

Within the domain of visualisation, we initiated our survey starting with publications from the following resources:

Journals: IEEE Transactions on Visualization and Computer Graphics, Computer Graphics Forum, IEEE Computer Graphics and Applications, Information Visualization

Conferences: IEEE Visual Analytics Science and Technology (partially published as a special issue of IEEE TVCG), IEEE Symposium on Information Visualization (InfoVis) (published as a special issue of IEEE TVCG since 2006), IEEE Pacific Visualization Symposium (PacificVis), EuroVis workshop on Visual Analytics (EuroVA)

Within the domain of machine learning, we initiated our survey starting with publications from the following resources:

Journals: Journal of Machine Learning Research, Neurocomputing, IEEE Transactions on Knowledge and Data Engineering

Conferences: International Conference on Machine Learning (ICML), ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)

We then scanned the relevant papers identified in the above resources and performed a backward and forward literature investigation using Google Scholar. In producing the taxonomy of works within Section 3, we labelled the publication both in terms of the analytical task and the integration strategy incorporated.

3.2. Modify parameters and computation domain

Here we list techniques where interaction has been instrumental in modifying the parameters of an algorithm, defining the measures used in the computations, or even changing the algorithm used. Another common form of interaction here is to enable users to modify the computational domain to which the algorithm is applied. Such operations are often facilitated through interactive visual representations of data points and data variables where analysts can select subsets of data and run the algorithms on these selections within the visual analysis cycle to observe the changes in the results and to refine the models iteratively. The types of techniques described in this section can be considered as following a "direct manipulation" [Shn83] approach where the analysts explicitly interact with the algorithm before or during the computation and observe how results change through visualization.

Dimension Reduction One class of algorithms that is widely incorporated in such explicit modification strategy is dimension reduction. Since high-dimensional spaces are often cognitively challenging to comprehend, combinations of visualization and dimension reduction methods have demonstrated several benefits. Johansson and Johansson [JJ09] enable the user to interactively reduce the dimensionality

	Modify Parameters & Computation Domain	Define Analytical Expectations
Dimension Reduction	[JJ09], [FJA*11], [FWG09], [SDMT16], [WM04], [NM13], [TFH11], [TLLH12], [JBS08], [ADT*13], [JZF*09]	[EHM*11], [EBN13], [BLBC12], [HBM*13], [GNRM08], [IHG13], [KP11], [PZS*15], [KCPE16], [KKW*16]
Clustering	[Kan12], [RPN*08], [SBTK08], [RK04], [SS02], [LSS*12], [LSP*10], [TLS*14], [TPRH11a], [AW12], [RPN*08], [HSCW13], [TPRH11b], [PTRV13], [HHE*13], [WTP*99], [YNM*13], [SGG*14]	[HOG*12], [CP13], [BDW08], [CCM08], [BBM04], [ABV14], [KKP05], [KK08]
Classification	[PES*06], [MK08], [MBD*11], [vdEvW11], [CLKP10], [KPB14], [AAB*10], [AAR*09], [KGL*15]	[Set09], [SK10], [BKSS14], [PSPM15]
Regression	[PBK10], [MP13], [MME*12], [TLLH12], [KLG*16]	[MGJH08], [MGS*14] [LKT*14] [YKJ16]

Table 1: *In Section 3, we review the existing literature in visual analytics following a 2D categorization that organizes the literature along two perspectives: Algorithm Type (rows) and Interaction Intent (columns).*

of a data set with the help of quality metrics. The visually guided variable ordering and filtering reduces the complexity of the data and provides the user a comprehensive control over the whole process. The authors later use this methodology in the analysis of high-dimensional data sets involving microbial populations [FJA*11]. An earlier work that merges visualization and machine learning approaches is by Fuchs et al. [FWG09]. The authors utilize machine learning techniques within the visual analysis process to interactively narrow down the search space and assist the user in identifying plausible hypotheses. In a recent paper, Stahnke et al. [SDMT16] devised a probing technique using interactive methods through which analysts can modify the parameters of a multi-dimensional scaling projection. The visualization plays a key role here to display the different dimension contributions to the projections and to communicate the underlying relations that make up the clusters displayed on top of the projection results.

In MDSteer [WM04], an embedding is guided by user interaction leading to an adapted multidimensional scaling of multivariate data sets. Such a mechanism enables the analyst to steer the computational resources accordingly to areas where more precision is needed. This technique is an early and good example of how a deep involvement of the user within the computational process has the potential to lead to more precise results. Nam and Mueller [NM13] provide the user with an interface where a high-dimensional projection method can be steered according to user input. They provide "key" computational results to guide the user to other relevant results through visual guidance and interaction. Turkay et al. introduce the dual-analysis approach [TFH11] to support analysis processes where computational methods such as dimension reduction [TLLH12] are used. The authors incorporate several statistical measures to inform analysts on the relevance and importance of variables. They provide several perspectives on the characteristics of the dimensions that can be interactively recomputed so that analysts are able to make multi-criteria decisions whilst using computational methods. Jänicke et al. [JBS08] utilize a two-dimensional projection method where the analysis is performed on a projected 2D space called the attribute cloud. The resulting point cloud is then used as the medium for interaction where the user is able to brush and link the selections to other views of the data. In these last group of examples, the capability to run the algorithms on user-defined subsets of the data through visually represented rich information is the key mechanism to facilitate better-informed, more reliable data analysis processes.

Clustering Clustering is one of the most popular algorithms that have been integrated within visual analytics applications. Since visual representations are highly critical in interpreting and comprehending the characteristics of clusters produced by the algorithms, direct modification of clustering algorithms are often facilitated through interactive interfaces that display new results "on-demand". gCluto [RK04] is an interactive clustering and visualization system where the authors incorporate a wide range of clustering algorithms. This is an early example where multiple clustering algorithms can be run on-the-fly with varying parameters and results can be visually inspected. In *Hierarchical Clustering Explorer* [SS02], Seo and Shneiderman describe the use of an interactive dendogram coupled with a colored heatmap to represent clustering information within a coordinated multiple view system.

Other examples include work accomplished using the Caleydo software for pathway analysis and associated experimental data by Lex et al. [LSS*12, LSP*10]. In their techniques, the authors enable analysts to investigate multiple runs of clustering algorithms and utilize linked, integrated visualizations to support the interpretation and validation

of clusters. Along the same lines, Turkay et al. present an interactive system that addresses both the generation and evaluation stages within the clustering process and provides interactive control to users to refine grouping criteria through investigations of measures of clustering quality [TPRH11a]. In a follow-up work [TLS*14], within the domain of clustering high-dimensional data sets, integrated statistical computations are shown to be useful to characterize the complex groupings that analysts encounter in such data sets. Figure 7 demonstrates how the authors incorporated statistical analysis results to indicate important features for data groups. In this work, the most discriminative features (indicated with red dots as opposed to blue ones that are less important) for the clustering result of a high-dimensional data set are represented as integrated linked views. The user is able to select these features in one clustering result (e.g., within the clustering result in the right-most column in Figure 7) and observe whether the same features are represented in others, e.g., in the left-most column.

Schreck et al. [SBTK08] propose a framework to interactively monitor and control Kohonen maps to cluster trajectory data. The authors state the importance of integrating the expert within the clustering process for achieving good results. Kandogan [Kan12] discusses how clusters can be found and annotated through an image-based technique. His technique involves the use of "just-in-time" clustering and annotation, and the principal role for visualisation and interaction is to aid the interpretation of the structures observed, and provide a deeper insight into why and how particular structures are formed.

An important role for visualization is to get the user engaged in *progressive* and *iterative* generation of clusters [RPN*08]. In such approaches, the user is presented with content that is built step-by-step and gains additional insight in each iteration to decide whether to continue, alter, or terminate the current calculations. Such levels of interactivity, of course, require the solutions to be responsive and capable of returning results within acceptable delays. Ahmed and Weaver [AW12] address this problem through forward-caching expected interaction possibilities and providing users with clustering results without breaking the responsive analytical flow.

Visual analytics applications that involve clustering algorithms within the analysis of complex dynamic networks have also been developed [HSCW13]. The use of visualisation is in particular critical with such dynamic relational data sets due to the limitations in interpreting the algorithmic results; well-designed combinations of visual summaries can assist analysts in this respect. In the domain of molecular dynamics simulation, there are some examples of tight integrations of interactive visualizations, clustering algorithms, and statistics to support the validity of the resulting structures [TPRH11b], [PTRV13].

Classification Being a relevant and widely utilized technique, classification algorithms have also found their place within visual analytics applications. Common roles for interactive visualization are filtering the feature space, iteratively observing and fixing problems, and when the classification tasks involve multiple mediums such as space, time and abstract features, providing multiple perspectives to the algorithmic results.

A conceptual framework on how classification tasks can be supported by interactive visualizations is presented by May and Kohlhammer [MK08]. Their approach improved the classification of data using decision trees in an interactive manner. They proposed the use of a technique called KVMaps to inform users on classification quality thus enabling the iterative refinement of the results. The authors later proposed a technique called SmartStripes [MBD*11] where they investigated the relations between different subsets of features and entities. Interactive visual representations have been used to help create and understand the underlying structures within decision trees [vdEvW11]. The authors not only presented the overall structure of decision trees, but also provided intuitive visual representations of attribute importance within the different levels of the tree. Such interactive visualizations are critical in unraveling the computed information hidden within the layers and can be quite instrumental in increasing the trust in such computational models. Similar insights can be gained on other models (additive ones, e.g. naive Bayes, in [PES*06] and more general ones in [SK10]) by *explaining* individual classification. In these papers, the authors display the contribution of features to the classification made by the model and enable what-if scenarios, such "how would the classification change if this particular feature was set to another value?"

In iVisClassifier by Choo et al. [CLKP10], the authors improve classification performance through interactive visualizations. Their technique supports a user-driven classification process by reducing the search space, e.g., through recomputing Latent Dirichlet Allocation (LDA) [BNJ03] with a user-selected subset of data defined through filtering in additional coordinated views. Klemm et al. [KGL*15] investigate the use of interactive visualisation to compare multiple decision trees in investigating relations within non-image and image based features for a medical application. They visualise the quality aspects of classifiers to infer observations on the predictive power of the features.

Krause et al. [KPB14] address the important process of feature selection within model building for classification purposes. Through visual representations of cross-validation runs for feature ranking with various algorithms, their method supports the decisions made while including or excluding particular features from a classification model (see Figure 8). Their approach enables users to be part of the predictive model building process and, as also demonstrated by the authors, leads to better performing/easier to interpret models. Their methodology is based on producing glyphs for the features of a data set to represent how important each one is within a number of classification models. In addition, the glyphs are also used as elements for visual selections and

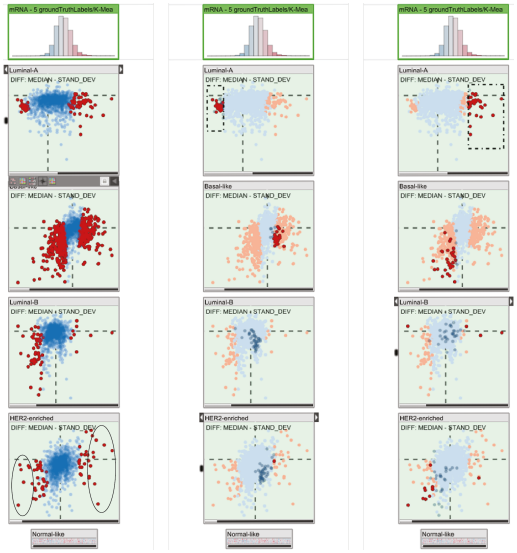


Figure 7: *Visualization of clustering results, together with associated on-the-fly computations to identify discriminating features of groups, are used here to aid analysts in interpreting the clusters and refining them further [TLS*14].*

enable analysts to interactively apply modelling on subsets of features.

Classification of spatio-temporal patterns is one of the complex tasks that requires the involvement of user input and efficient algorithms due to the complex nature of structures found in such data sets. Andrienko et al. [AAB*10] investigate how self organizing maps (SOMs) are integrated into the visual analysis process. They integrate a SOM matrix where the user can interactively modify the parameters and observe the changes in the results in various visual representations, e.g., where space is represented in time, and the time is represented in space. Again involving spatio-temporal data, an interactive process where a clustering algorithm assists users to pick relevant subsets in building classifiers has shown to be effective in categorizing large collections of trajectories [AAR*09].

Regression Identifying the multivariate relations within data variables, in particular when their numbers are high, is one of the critical tasks in most data analysis routines. In order to evaluate to what degree observed relations can be attributed to underlying phenomena and to build causal interpretations, visual analytics approaches have shown good potential. Visualization has shown to be effective in validating predictive models through interactive means [PBK10]. The authors visually relate several n-dimensional functions to known models through integrated visualizations within a model building process. They observed that such a visualization-powered approach not only speeds up model building but also increases the trust and confidence in the results. Mühlbacher and Piringer [MP13] discuss how the process of building regression models can benefit from integrating domain knowledge. Berger et al. [BPFG11] introduce

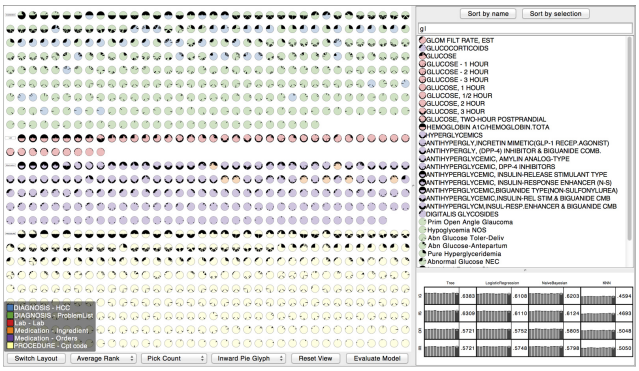


Figure 8: *Visual summaries to indicate the relevance of features over cross-validation runs support analysts in making informed decisions whilst selecting features for a classification model [KPB14].*

an interactive approach that enables the investigation of the parameter space with respect to multiple target values. Malik et al. [MME*12] describe a framework for interactive auto-correlation. This is an example where the correlation analysis is tightly coupled with the interactive elements in the visualization solution. Correlation analysis has been integrated as an internal mechanism to investigate how well lower-dimensional projections relate to the data that they represent [TLLH12]. The use of relational representations here supports analysts to evaluate how local projection models behave in preserving the correlative structures in the data. In a recent paper, Klemm et al. [KLG*16] demonstrates the use of visualisation to show all combinations of several independent features with a specific target feature. The authors demonstrate how the use of template regression models, interactively modifiable formulas and according visual representations help experts to derive plausible statistical explanations for different target diseases in epidemiological studies.

3.3. Define analytical expectations

Unlike the papers in the previous category where the user explicitly modifies the parameters and the settings of an algorithm, the works we review under this section follow a different strategy and involve users in communicating *expected results* to the computational method. In these types of interactive methods, the user often observes the output of an algorithm and tell the machine which aspect of the output is inconsistent with the existing knowledge, i.e., correcting the algorithm. Furthermore, analysts can also communicate examples of relevant, domain-knowledge informed relations to be preserved in the final result. Since this is a relatively recent approach to facilitate the interaction between the user and the algorithms, the number of works in this category is not as high as the previous section. In the following, we review such works again under a categorization of different ML algorithm types involved. Notice that integrating user knowledge in this way in unsupervised learning contexts falls into the general semi-supervised framework, which is a

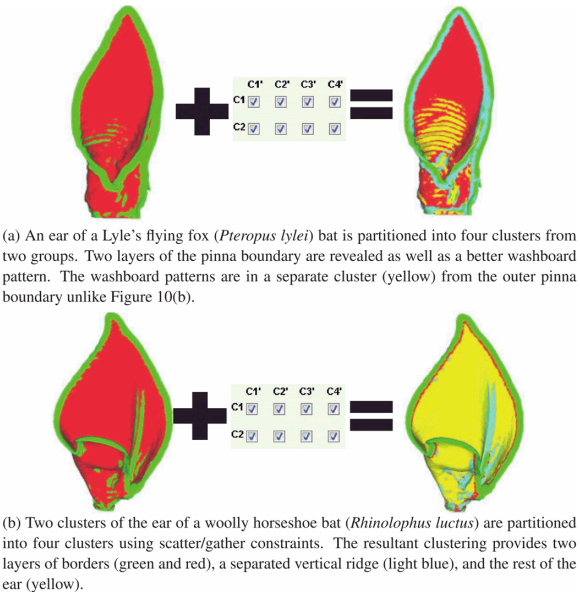
principled way in ML for making unsupervised problems less ill-posed.

Dimension Reduction Dimension reduction algorithms are suitable candidates for such approaches due to the often "unsupervised" nature of the algorithms and the possibility that errors and losses within the reduction phase are high, in particular with datasets with high numbers of dimensions. As one of the early works along these lines, Endert et al. [EHM*11] introduce observation level interactions to assist computational analysis tools to deliver more interpretable/reliable results. The authors describe such operations as enabling the *direct manipulation* for visual analytics [EBN13]. In this line of work, the underlying idea is to provide mechanisms to users to reflect their knowledge about the data through interactions that directly modify computational results. One typical interaction is through *moving* observations in a projection such that the modified version is more similar to the *expectation* of the analyst [EHM*11, BLBC12]. This line of research has been expanded to focus on the interpretability of linear [KCPE16] and non-linear DR models [KKW*16]. Hu et al. [HBM*13] complemented such visualization level interaction methods with further interaction mechanisms. The authors aim to understand users' interaction intent better and give them mechanisms to also highlight preferences on *unmoved* points.

In their Model-Driven Visual Analytics system, Garg et al. [GNRM08] suggest the use of a "pattern painting" mechanism that enables analysts to paint interesting structures in the visualization which are then turned into logical rules that can be fed into a projection algorithm to build an effective model.

An interesting supervised point of view has been proposed in [IHG13] on the dimension reduction steering. The main idea is to introduce an information theoretic criterion that evaluates the uncertainty in the representation, considering that the original high dimensional points are noisy. Given this criterion, the authors apply an active learning approach to select points that are maximally informative: if the user can move one of those points to its desired position, the uncertainty of the representation will be maximally reduced (compared to the reduction expected with other points). The experimental evaluation shows that the optimal points tend to be more uniformly distributed over the projected data set than with other selection methods, possibly reducing some of the drawbacks of active learning summarized in e.g. [ACKK14].

Clustering There are a number of works where user knowledge is incorporated to feed a clustering algorithm with expected results. Hossain et al. makes use of a scattergather technique to iteratively break up or merge clusters to generate groupings that meet analysts' expectations [HOG*12]. (See Figure 9.) In their technique, the expert iteratively introduces constraints on a number of required relations and the algorithms take these constraints into consideration to generate more effective groupings. The users state whether clusters



(a) An ear of a Lyle's flying fox (*Pteropus lylei*) bat is partitioned into four clusters from two groups. Two layers of the pinna boundary are revealed as well as a better washboard pattern. The washboard patterns are in a separate cluster (yellow) from the outer pinna boundary unlike Figure 10(b).

(b) Two clusters of the ear of a woolly horseshoe bat (*Rhinolophus luctus*) are partitioned into four clusters using scatter/gather constraints. The resultant clustering provides two layers of borders (green and red), a separated vertical ridge (light blue), and the rest of the ear (yellow).

Figure 9: *Scatter Gather [HOG*12] is a technique to interactively gather feedback from analysts in response to algorithmic output and refine user-generated constraints to improve the clustering.*

in the current segmentation should be broken up further or brought back together. Upon inspection of a clustering result, the user interactively constructs a scatter gather constraint matrix which represents a preferred clustering setting from her perspective. The algorithm then considers this input along with the clustering result to come up with an "optimized" result. In a number of papers, the user has been involved even further to modify clustering results. In order to support a topic modeling task through clustering, Choo et al. [CP13] enable users to interactively work on topic clusters through operations such as splitting, merging and also refining clusters by pointing to example instances or keywords.

More generally, clustering is one of the first tasks of machine learning to include ways to take into account expert knowledge, originally in the form of contiguity constraints (see [Mur85] for an early survey): the expert specifies a prior neighborhood structure on data points (for instance related to geographical proximity) and the clusters are supposed to respect this structure (according to some notion of agreement). While the original methodology falls typically into the offline slow steering category, it has been extended to more general and possibly online steering based on two main paradigms for constraints clustering [BDW08]: the pairwise paradigm (with *must-link*/*cannot-link* constraints) and the triplet paradigm (with constraints of the form x *must be closer to* y *than to* z).

An early example of the pairwise paradigm is provided by [CCM08]. The authors describe a document clustering method that takes into account feedback of the form: this document should not belong to this cluster, this document

should be in this cluster, those two documents should be (or should not be) in the same cluster (this mixes pointwise constraints, with pairwise ones). Active learning has been integrated into this paradigm in [BBM04]. A variation over the pairwise approach which consists in issuing merge and/or split requests at the cluster level has been proposed and studied in [ABV14].

Constraints based on triplet are more recent and were proposed in the context of clustering by [KKP05, KK08]. The main advantage of specifying triplet based constraints over pairwise ones is that they allow relative qualitative feedback rather than binary ones. They are also known to be more stable than pairwise comparisons [KG90].

Classification Classification tasks are suitable for methods where users communicate known/expected/wrong classification results back to the algorithm. The ideas employed under this section show parallels to the Active Learning methodologies develop in the ML literature [Set09] where the algorithms have capabilities to query the user for intermediate guidance during the learning process. In their visual classification methodology, Paiva et al. [PSPM15] demonstrates that effective classification models can be built when users' interactive input, for instance, to select wrongly labeled instances, can be employed to update the classification model. Along the similar lines, Behrisch et al. [BKSS14] demonstrate how users' feedback on the relevance of features in classification tasks can be incorporated into decision making processes. They model their process in an iterative dialogue between the user and the algorithm and name these stages as *relevance feedback* and *model learning*. This work serves as a good example of how user feedback might lead to better performing, fit-for-purpose classification models.

Regression Although examples in this category are limited in numbers, defining the "expected" has shown great potential to support interactive visual steering within the context of ensemble simulation analysis [MGJH08, MGS*14]. In their steerable computational simulation approach, Matkovic et al. [MGJH08] demonstrate how a domain expert (an engineer) can interactively define and refine desired simulation outputs while designing an injection system. Their three-level steering process enables the expert to define desired output values through selections in multiple views of simulation outputs. The expert then moves on to visually explore the control variables of the simulation and assess whether they are feasible and refine/re-run the simulation models accordingly. The authors went on to incorporate a regression model within this process to further optimise the simulation results based on users' interactive inputs [MGS*14]. With this addition to the workflow, the experts again indicate desired output characteristics visually and a regression model followed by an optimization supports the process to quickly converge to effective simulation parameters. The critical role that the users play in these examples is to express their expert knowledge to identify and communicate suitable solutions to the algorithmic processes which in turn try and optimize for those.

4. Application Domains

The integration of ML techniques into VA systems has been exemplified in different domains, described below. Each of these domains present unique and important challenges, thus different combinations of interactive visualizations and ML techniques are used. Some of these techniques are related to, but go beyond the classifications in Section 3. For instance, dimension reduction, clustering, etc. since they must be closely embedded in the VA system and can be attached to higher level meanings. However, most are relevant to the Define Analytical Expectations category in Table 1. The examples given in this section generally make use of one or more technique categories in Section 3, depending on the particular domain for which the applications are designed for.

4.1. Text Analytics and Topic Modeling

Text corpora are frequently analyzed using visual analytic systems. Text is a data format that lends itself nicely to specific computational processes, as well as human reasoning. Various text analytics methods have seen a lot of use in visual analytics systems over the past 6-7 years. A main reason is that these methods have proved useful in organizing large, unstructured text collections around meaningful topics or concepts. The text collections considered have been diverse including research publications, Wikipedia entries, streaming social media such as Twitter, Facebook entries, patents, technical reports, and other types.

Visual analytic tools have been used to support information foraging by representing high-dimensional information, such as text, in an easily comprehensible two-dimensional view. In such views, the primary representation is one where information that is relatively closer to other information is more similar (a visualization method borrowed from cartography [Sku02]). These applications allow users to find relevant information and gain new insights into topics or trends within the data. An early example of combining machine learning with visual analytics for analyzing text is a system called *IN-SPIRE* [WTP*99]. One of the views of the system, the *Galaxy View* shown in Figure 10, displays documents clustered by similarity. Using dimension reduction techniques, this view encodes relative similarity as distance (documents near each other are more similar). The high-dimensional representation of the text documents is created by keyword extraction from each document (defining a dimension), and weightings on the keywords determined computationally using popular methods such as TF-IDF, etc. [RECC10].

Visual analytic tools have also been used to support synthesis by enabling users to externalize their insights during an investigation. In a spatial workspace where users can manually manipulate the location of information, users build spatial structures to capture their synthesis of the information over time - a process referred to as "incremental formalism" [SM99, SHA*01]. Andrews et al. found that intelligence analysts can make use of such spatial structures as

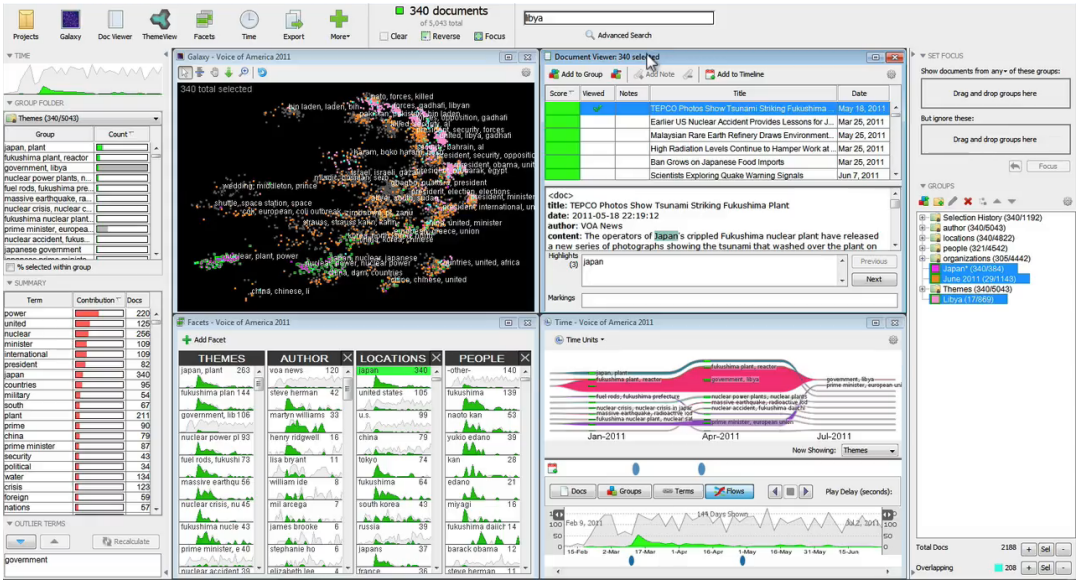


Figure 10: *IN-SPIRE [WTP*99], a VA system for text corpora. IN-SPIRE combines computational metrics with interactive visualizations.*

a means to externalize insights during sensemaking, manually placing relevant documents in clusters on a large, high-resolution display [AEN10]. Additionally, they found that the large display workspace promoted a more spatially-oriented analysis. Tools, such as I2 Analyst's Notebook [i2], Jigsaw's "Tablet view" [SGL08], nSpace2 [EKHW08, WSP*06], Analyst's Workspace [AN12], and others have also found it helpful to provide users with a workspace where spatial representations of information can be manually organized.

More recently, researchers have developed techniques such as Latent Semantic Analysis (LSA) for extracting and representing the contextual meaning of words [LD97]. LSA produces a concept space that could then be used for document classification and clustering. Also, probabilistic topic models have emerged as a powerful technique for finding semantically meaningful topics in an unstructured text collection [BL09]. Researchers from the knowledge discovery and visualization communities have developed tools and techniques to support visualization and exploration of large text corpora based on both LSA (e.g., [DWS*12, CDS09]) and topic models (e.g., [IYU08, LZP*09, WLS*10, OST*10]).

The Latent Dirichlet Allocation (LDA) model of Blei et al. [BNJ03], which represents documents as combinations of topics that are generated, in the unsupervised case, automatically has proved particularly useful when integrated in a visual analytics system. The LDA model postulates a latent topical structure in which each document is characterized as a distribution over topics and most prominent words for each topic are determined based on this distribution. Each topic is then described by a list of leading keywords in ranked order. When combined with VA techniques, LDA provides meaningful, usable topics in a variety of situations (e.g., [GS04, ZC07, DWCR11]). Recent developments in the ML community provide ways to refine and improve topic models by integrating user feedback, e.g. moving words from one topic to another [HBGSS14].

There have been extensions of LDA-based techniques and other text analytics by investigating texts in the combination ¡*topic, time, location, people*¿. This permits the analysis of the ebb and flow of topics in time and according to location [DWCR11, DWS*12, LYK*12]. Time-sensitivity is revealed not only in topics but in keyword distributions [DWS*12]. Lately there has been work to add people and demographic analysis as well [DCE*15]. Combining topic, time, and location analysis leads to identification of events, defined as "meaningful occurrences in space and time" [KBK11, DWS*12, CDW*16, LYK*12]. Here the topic analysis can greatly help in pinpointing the meaning. In addition, combining topic modeling with named entity extraction methods, such as lingpipe [20008], can greatly enhance the time, location, and even people structure since these quantities can be automatically extracted from the text content [MJR*11, CDW*16].

At this point, it is worthwhile to describe a visual analytics system that combines all these characteristics. VAiRoma [CDW*16] (shown in Figure 11) creates a narrative that tells the whole 3,000 year history of Rome, the Empire, and the state of Italy derived from a collection of 189,000 Wikipedia articles. The articles are selected from the nearly 5M English language article collection in Wikipedia using a short list of keyword, but otherwise the initial topic modeling and named entity extraction are done automatically. The interface for VAiRoma is displayed in Figure 11. The individual topics are depicted as color-coded streams in the timeline view (A). The circular topic view in (C) provides a compact way of depicting topics, the weights of their contributions for a given time range, and topic keywords. The navigable map view in (B) provides immediate updates

of geographic distribution of articles (based on locating the geographic entities in the text) in terms of hotspots for a selected time range and topic. The window (f) lists article titles for selected geographic view, time range, and topic. In Figure 11, one can clearly see event peaks for selected topics having to do with Roman government and military battles in the period from 500 BC to 500 AD. The interlinked windows in the interface plus key topics and event peaks permit a user to quickly peruse the main events in ancient Roman history, including the rise of Christianity and the Catholic church, trade with India and the Far East, and other events that one might not find in looking narrowly at, say, just the history of the Roman Empire. In this case, the user can focus from thousands of articles to a few hundred articles overall, which she can then quickly peruse. See the VAiRoma article for more details.

VAiRoma shows the power of the overall model depicted in Figure 3. Though it is not complete w.r.t. this model (no current VA system is), it provides an integrated approach to data handling, interactive visualization, ML (in this case topic modeling) combined with other techniques, and exploration and knowledge building techniques. It shows the power of an integrated approach. The approach is general and is now being applied to large, heterogeneous collections of climate change documents. In addition, full text journal article collections are being analyzed using extensions of the topic modeling and entity extraction methods. This shows that once ¡topic, time, location, people¿ features and event signatures can be extracted, analyses based on these analytics products can integrate a wide range of heterogeneous collections.

4.2. Multimedia Visual Analytics

Visual analytic applications have also been developed to allow people to explore multimedia (i.e., images, video, audio). For example, iVisClassifier shows how facial expression features can be incrementally explored and classified by a combination of image feature-detection algorithms and user feedback [CLKP10]. Through interactively adding and removing images from classifiers, the model learns the facial expressions that are interesting (and similar) to the user. It combines analytic models such as feature extraction and classification with visual analytic approaches. *MultiFacet* is another example of visually analyzing multimedia data [HHE*13]. MultiFacet presents facets of each data type to users as interactive filters. Thus, the process of interactively selecting attributes of different data types helps create groups of conceptually interesting and related information.

As image and video data is often combined with text data (or textual metadata attached to the images or videos), fusing the feature space between these datatypes is an open challenge. Automated approaches are error-prone, and often require user intervention and guidance when semantic concepts and relationship need to maintained across data types [CBN*12]. Similarly, an example of a much more specific application is given in [BM13] where the authors present a steering mechanism for source separation in a single monophonic recording. The user can annotate a standard time-frequency display to roughly define the different sources. Errors made by the algorithm can be annotated to improve further the separation.

4.3. Streaming Data: Finance, Cyber Security, Social Media

Streaming data is a growing area of interest for visual analytics. Data are no longer isolated and static, but instead are part of a sensor-laden ecosystem that senses and stores data at increasing frequencies. Thus, visual analytic systems that integrate machine learning models have great potential. Examples of domains that generate streaming data include the financial industry, cyber security, social media, and others.

In finance, for example, *FinVis* is a visual analytics system that helps people view and plan their personal finance portfolio [RSE09]. The system incorporates uncertainty and risk models to compute metrics about a person's portfolio, and uses interactive visualizations to show these results to users. Similarly, Ziegler et al. presented a visual analytic system to help model a user's individual preferences for short, medium, and long-term stock performance [ZNK08] and later extended their approach to real-time market data [ZJGK10]. Figure 12 is an example of how visualisations can provide an in-depth understanding of the groupings (clusterings) of financial time series. Here, financial market data for assets in 3 countries and 28 market sectors from 2006 and 2009 are depicted. The red bars indicate the crash of the stock market in 2008 and the visualisation enables the user to identify the overall changes but also notice subtle variations such as the lack of a response in some countries for particular sectors.

Cyber security is a domain fraught with fast data streams and alerts. Examples of machine learning techniques often incorporated into systems that support this domain include sequence and pattern-based modeling, rule-based alerting, and others [BEK14]. People in charge of the safety and reliability of large networks analyze large amounts of streaming data and alerts throughout their day, thus the temporal component of making a decision from the analysis is emphasized. For example, Fisher et al. presented *Event Browser*, a visual analytic system for analyzing and monitoring network events [FMK12]. Their work emphasizes how different tasks of the analyst have to happen at different time scales. That is, some tasks are "real-time", while others can be taken "offline" and performed for a longer duration of time. The persistent updating of new data into the offline tasks presents challenges.

Social media data can also be analyzed using visual analytic systems. For example, *Storylines* [ZC07] and *EventRiver* [LYK*12] are two examples of how visual analytic applications can help people understand the evolution of events, topics, and themes from news sources and social media feeds. In these systems, similar machine learning techniques are used as for text. However, the temporality of the data is more directly emphasized and taken into account.

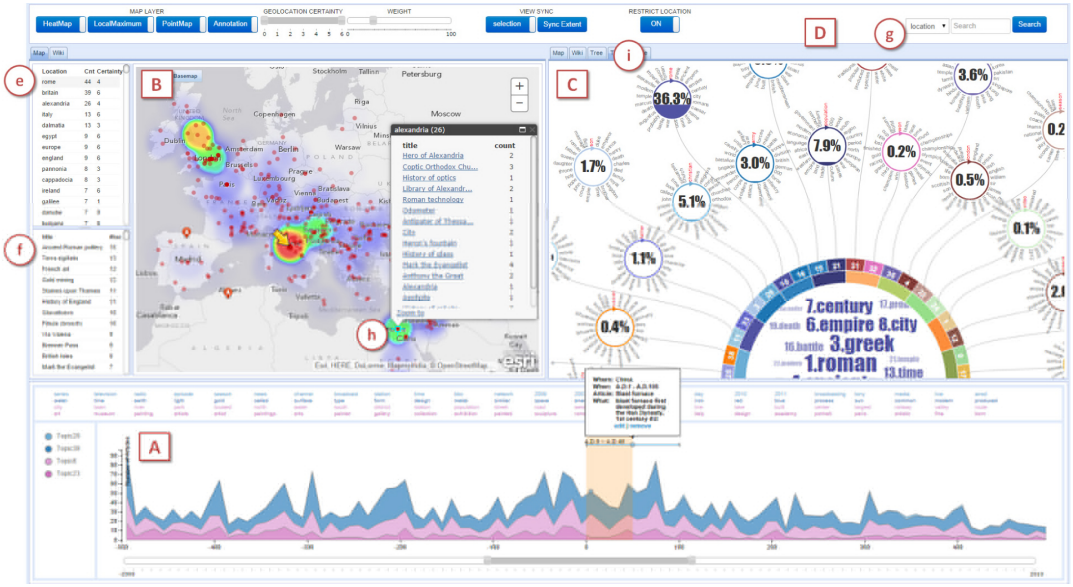


Figure 11: *Overview of VAiRoma Interface. The interface has three main views: Timeline view (A), Geographic view (B) and Topic view (C). A user-generated annotation is shown in the Timeline view.*

Lu et al. [LKT*14] showed how appropriate social media analysis could have predictive power, in their case predicting movie box office grosses from early word of mouth discussion on Twitter, YouTube, and IMDB. A dictionary-based sentiment analysis was used along with analytics from the R statistical computing environment and the Weka machine learning workbench. This permitted a choice of modeling in terms of multivariate regression, support vector machines, and neural networks. The paper promoted an integrated visual analytics approach where the interactive visualizations, based on D3, permitted users to investigate comments and sentiment, classify similar movies, and follow trends and identify features. The user could then improve a base line regression model based on trends and features identified in the visaulizations. Results of the use cases were positive with several of the non-expert participants being able to outperform experts in predicting opening weekend grosses for 4 films, according to the criteria set up by the authors. The paper has the usual limitation of supervised learning approaches in that a training dataset must first be collected and analyzed as a preliminary step, but it does successfully allow for improvement of the analytic model within the VA environment. Also, like many papers dealing with more complex analysis, it defines a process for best use of the system; this appears to be an important and effective approach for VA + ML systems.

Yeon et al. [YKJ16] covered similar ground in their identification and analysis of interesting past abnormal events as a precursor for predicting future events. Here, as in Lu et al. and in other papers using ML, context and analytic power is obtained from combining multiple sources (in this case social media and news media). Yeon et al. identify contextual pattern in these past events, which permit them to make predictions for future events in similar contexts. An interactive interface involving spatio-temporal depiction of events plus identification of other features permits the choosing of interesting events and specification of their contexts. Trends for the unfolding of future events and possible unfolding story lines can then be created. The authors evaluated their VA system with three use cases.

Figure 12: *Aggregated visual representations and clustering have been used in supporting the real-time analysis of temporal sector-based market data [ZJGK10].*

4.4. Biological Data

Biology, and in particular, bio-informatics are fields that are increasingly becoming data-rich and the use of visualisation empowered analysis methods are proving highly useful and effective [GOB*10]. Although most computational analysis solutions only incorporate visualization as a communication medium and do not make use of interaction, there are a number of examples where VA and ML approaches operate in integration. Within the context of epigenomic data analysis, Younesy et al. [YNM*13] present how a number of ill-defined patterns and characteristics within the data can be identified and analysed through the help of interactive visualizations

and integrated clustering modules. They demonstrate how user-defined constraints can be utilised to steer clustering algorithms where the results are compared visually. Grottel et al. [GRVE07] discuss how interactive visual representations can be instrumental in interpreting dynamic clusters within molecular simulations. In addition to these, interactive visualisations have been shown to support bi-cluster analysis [SGG*14]. The authors utilize an interactive layout where fuzzy bi-clusters are investigated for multi-tissue type analysis. Biclustering is an algorithmic technique to solve for coordinated relationships computed from high-dimensional data representations [MO04], and has been used in other domains, including text analysis [SNR14, SMNR16, FSB*13].

In addition to the above methods where the focus is mainly on investigating clusters, there are also works where interactively specified high-dimensional data projections are utilised to characterize and compare different cancer subtypes [ADT*13]. In their tool called viSNE, the authors demonstrate how user-driven, locally applied projections preserve particular relations and they argue that such methods are instrumental in interpreting any multi-dimensional single-cell technology generated data.

5. Embedding Steerable ML Algorithms into Visual Analytics

As discussed above at several points and categorized in Section 3, one area of research that has been recently attracting much interest in the machine learning and data visualization communities is the development of interactive approaches binding visualizations to steerable ML algorithms. This goes beyond typical interactive ML methods in that it places interaction at the same level as visualization and ML, thus producing a powerful extension of visual analytics. As explained in [Van05], [PSCO09], interaction provides feedback in the visualization process, allowing the user to manipulate the parameters that define a visualization on the basis of the knowledge acquired in previous iterations. In particular, low latency interaction with large update rates of the visual display provides higher levels of user involvement in the analysis [EMJ*11], triggering low level attention and processing mechanisms (such as tracking moving items), where the user's senso-motor actions have immediate effects in the displayed information. Despite interaction mechanisms having extensively been discussed in the visualization literature [Van05], [PSCO09], the relationships between these parameters and the resulting visualization are in most cases of a simple nature, including changes of scale, displacements, brushing, etc., specially for low latency interaction. As pointed out in [VL13], hardly ever are complex interactions or transformations based on intelligent data analysis undertaken at this level. This fact is certainly surprising, especially considering that ML is a mature discipline and the power of today's hardware, as well as programming languages and libraries make it possible to use algorithms (or adapted versions of them) as intermediates between the user actions and the visualization, even at low latency levels.

The DR algorithms discussed in Section 3, which construct a mapping from a high dimensional input space onto a typically 2D or 3D visualization space, would be particularly useful for extended VA approaches. To build such mappings, DR algorithms seek to preserve neighborhood relationships among the items in both spaces, resulting in representations that follow the so called "spatialization principle" (based on the cartographic principle where closeness $\approx$ similarity [Sku02]). Placing similar items in close positions results in highly intuitive arrangements of items in a visual map that serves as a basis for developing insightful visualizations of high dimensional elements [Ves99, KP11, EBN13]. Moreover, the connection that DR mappings make between something that can be "seen" and a high dimensional feature space suggests using the visual map as a canvas where classical interaction mechanisms (zoom, pan, brushing & linking, etc.) can be used to explore high dimensional data.

However, interaction can go far beyond this point by allowing the user to steer the DR algorithm through the visualization by direct modification of its parameters or by making transformations on the input data. As discussed in Section 3, this idea has been explicitly formulated in [CP13] as *iteration-level interactive visualization*, which aims at visualizing intermediate results at various iterations and letting the users interact with those results in real time. In a slightly more formal way, as shown in [DCV16], an interactive DR algorithm –the argument can be extended to other ML algorithms– can be considered as a dynamically evolving system, driven by a *context* that includes the input data and the algorithm's parameters

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, \mathbf{u}), \qquad \mathbf{v} = \mathbf{g}(\mathbf{y}) \tag{1}$$

where $\mathbf{y}$ is the *internal state* of the algorithm, $\mathbf{v}$ is the *outcome* of the algorithm (e.g. a visualization), which depends on the internal state, and $\mathbf{u} = \{\mathbf{x}, \mathbf{w}\}$ is a *context vector* that contains the *input data* $\mathbf{x}$ and the *algorithm parameters* $\mathbf{w}$. In a general framework, the user will steer the algorithm by manipulating $\mathbf{w}$ based on his/her knowledge acquired from the visualization $\mathbf{v}$. Under a fixed context $\mathbf{u}^0$ –i.e. no changes in the input data or the algorithm parameters–, the internal state $\mathbf{y}$ in model (1) will keep on changing until it reaches convergence to a steady state condition $\mathbf{0} = \mathbf{f}(\mathbf{y}^0, \mathbf{u}^0)$. Changes in the algorithms parameters $\mathbf{w}$ or in the input data $\mathbf{x}$ will make the internal state evolve to a new steady state condition $\mathbf{0} = \mathbf{f}(\mathbf{y}^1, \mathbf{u}^1)$, and hence result in a new visualization $\mathbf{v}^1$. For a continuous $\mathbf{f}(\cdot)$ –typically for non-convex algorithms, based on gradient descent approaches– the representation $\mathbf{v}(t)$ will smoothly change, resulting in animated transitions that provide a continuous feedback to the user. Despite the fact that this behavior opens a broad spectrum of novel and advanced user interaction modes and applications, this is still a rather unexplored topic.

Many possibilities may arise from this approach, all based on changes in different elements of the context vector $\mathbf{u}$:

- One fundamental subset of parameters that conveys a great deal of user insight are the *input data metrics*, which

can be expressed as a weight matrix $\Omega = (\omega_{rs})$ being $\|\mathbf{a}\|_\Omega = \sum_r \sum_s a_r \omega_{rs} a_s$, whose parameters are included in $\mathbf{w}$. Prior knowledge on the *relevance of features* can be easily considered allowing user-driven modifications in the diagonal elements of $\omega_{ii} \subset \mathbf{w}$. An example related to this idea is the iPCA [JZF*09], an interactive tool that visualizes the results of PCA analysis using multiple coordinated views and a rich set of user interactions, including modification of dimension contributions. A similar idea on the stochastic neighbor embedding algorithm (SNE) was also proposed in [DCP*14].

- The user might also have insight on the *similarities between items*. In [BLBC12], a system called dis-function was developed, featuring DR visualization that allows the user to modify the distance matrix $D_{ij} = \|\mathbf{a}_i - \mathbf{a}_j\|_\Omega$ between items i, j, by moving points in the visualization based on his/her understanding of their similarity, and see new results after a recomputation of the projections with the new metrics.
- Also, *prior knowledge on class information* can be inserted by the user, suggesting techniques to increase the similarity of items belonging to the same class. In [PZS*15] a method is proposed to allow the user to include prior class knowledge in the DR projections by extending the original dataset with transformations of the original feature space based on his existing class knowledge.
- Finally, the input data $\mathbf{x}$ in model (1) may change with time ($\mathbf{x} = \mathbf{x}(t)$), suggesting the use of iDR on streaming data to provide live visualizations $\mathbf{v}(t)$ that convey *time varying information*; in this case, user interaction is possible through timeline sliders, making it possible to explore how input data items and their relationships evolve in time by moving back and forth in time.

These cases imply a substantially more advanced kind of feedback to the user than traditional interaction mechanisms. Placing these capabilities in a visual analytics framework greatly empowers them. As described in Figures 2 and 3, such a framework supports analytic reasoning, the discovery of much deeper insights, and the creation of actionable knowledge. The mere fact of being part of sensemaking and knowledge feedback loops (a virtuous cycle) suggests that there is huge potential and a broad spectrum of possibilities in the integration of ML algorithms discussed in this paper, where even the simplest ones may have multiplicative effects. For certain types of analysis, such as following animated transitions, this sort of interaction mechanism must be achieved in a fluid manner, with low latencies and fast update rates. However, this is not necessarily required for all knowledge generation and synthesis activities, as discussed next.

Levels of Interactive Response A long-recognized upper threshold for latency in WIMP and mobile interfaces is 0.1 second. Faced with higher latencies, users start to lose the connection between their actions and the visual response, commit more typing or selection errors, and become frustrated [HB11]. This limit has also been discussed as an upper threshold for coherent animations (though completely smooth animations would require a lower latency) and for a range of interactions in immersive VR. However, the detailed effects of particular latency thresholds depend on the task. For embedded analytics tools in VA systems, such as steerable ML methods, it is useful to define a wider range of interactive responses [RF16]:

- *Real-time Regime*: ¡ 0.1 second. Interactions such as moving a time slider to control an animation of time-dependent behavior or changing the weighting factors of leading dimensions in an interactive PCA tool [JZF*09] to reveal changes in the projected surface fall into this regime. Such interactions can be employed for rapid exploration and spotting of trends.
- *Direct Manipulation Regime*: 0.1 to 2-3 seconds. Analytic reasoning tends to involve more complicated interlinking of rich visualizations with ML methods. For example, the VAiRoma geographic window shows multiple hierarchical hotspot clusters (Figure 11) when a time range and topic are selected, but there is a delay of 2-3 seconds before the result is displayed. The user must peruse this distribution and its areas of concentration, which can take several seconds or more. During interface evaluation the delay was not noted and does not seem to hinder the user's reasoning process [CDW*16], perhaps because the user is thinking about the selection when it is made, and what it may mean, which then flows into her reasoning process once the result appears. The same seems to be true when the user makes a selection of a geographic region or a topic and experiences a similar delay until updates in the timeline or other linked windows appear.
- *Batch Regime*: 10 seconds or more. Here the cognitive flow of human reasoning is interrupted. To minimize effects of this interruption, the best analytics at this level of response might be those that launch a new reasoning direction (e.g., recalculation of textual topics based on a revised set of keywords).

These levels of response are related to performance timings from enactive cognition [GSFS06], suggesting that this model can be applied here. An important conclusion of this discussion is that it is not necessary to have real-time response for certain interactive ML algorithms; delays up to 2-3 seconds and perhaps more might be digestible by the user. This could substantially reduce the burden of interactive response for ML algorithms. Of course, further user studies of these algorithms in action should be carried out.

6. Open Challenges and Opportunities for ML and VA

Collaboration between ML and VA can benefit and drive innovation in both disciplines. Advances in ML can be used by VA researchers to create more advanced applications for data analysis. This includes the optimization of currently integrated techniques, but also the discovery of additional techniques that fit into the broad range of analytic tasks covered by visual analytic applications [AES05, LPP*06]. Similarly, as advances are made in VA applications, the user

requirements and needs can drive new ML algorithms and techniques.

Below, we list a collection of current challenge and opportunities at the intersection of ML and VA.

6.1. Creating and Training Models from User Interaction Data

ML models are typically built and modified based on ample training data that contain positive and negative ground truth examples. While many domains and tasks can be solved with ample training data, there exist scenarios, as discussed in this paper, where not enough training data is available. For these cases, it becomes important to incorporate user feedback into the computation in order to guide and parametrize the computational model being used. This raises the challenges of *how to incorporate user feedback into computation in an effective and expressive, yet usable manner?*

The concept of interactive machine learning has taken into account user feedback to steer and train these models. For example, users can provide positive or negative feedback to give support for or against suggestions or classifications made by the model. The models adjust over time based on this input.

However, there is the ability to look beyond labeling, or confirming and refuting suggestions as way to incorporate user feedback [ECNZ15] - what about the remaining user interaction that people perform during visual data exploration? User interaction logs contain rich information about the process and interests of the user. Examples of the kinds of inferences that can be made from the user interaction logs are shown in more detail earlier in the report. Thus, the opportunity exists for ML techniques to leverage the real-time user interaction data generated from the analysts using the system to steer the computation.

Systems that take into account a broader set of user interactions enable people more expressivity in conveying their mental model, preferences, and subject matter expertise. Further, taking into account the broader set of user interaction allows users of the system to stay more engaged in the act of visual data exploration, as opposed to actively training the model and system.

Figure 13 shows a model for how multiple types of user input can be incorporated into the machine learning models driving visual analytic techniques. As is shown in this model, two broad types of models can be created from user interaction: Data models and User Models. In general, data models refer to weighted data items and attributes. These can be weighted computationally, or via user feedback. Further, these weights can be computed based on inferences on the user interaction (i.e., to approximate user interest of focus). User models typically refer to computational approximations of the state of the user (e.g., cognitive load, personality traits [BOZ*14], etc.)

In addition to steering existing models (such as dimension

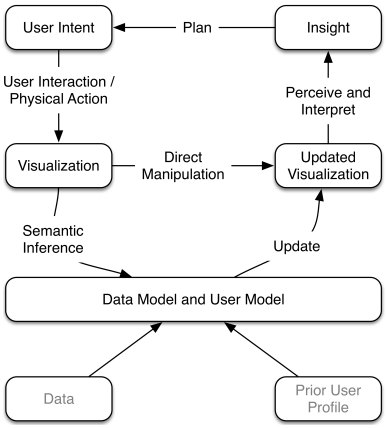


Figure 13: *A model from [ECNZ15] showing how multiple types of user input can be used to steer machine learning models in VA.*

reduction models, topic models, etc.), such user feedback can indicate the need for novel models to be created. By focusing on the user interaction, new discoveries can be made about the processes and analytic tasks of people during data analysis. This continued study, or *science of*, interaction [PSCO09] can lead to advances in the machine learning community in the way of new algorithms or techniques that model analytic tasks or processes of people.

6.2. Balancing Human and Machine Effort, Responsibility, and Tasks

For mixed-initiative systems, it is a common notion that there exists a balance of effort between the user and the machine [Hor99]. This effort can be divided by decomposing the larger task into sub-tasks that are either better suited to the person, or more quickly performed by the system. Similarly, these tasks often break down into being more well-defined and quantitative (i.e., solved by computation), or subjective and less formally defined (and thus needing input from the user). For example, a mixed-initiative visual analytic system for grouping and clustering can take into account the exemplar data items that are grouped by the user, generate a data model from those examples, and organize the remaining data points [DFB11].

However, there remains the need for generalizable empirical evidence to inform researchers about how to balance this effort between the user and the machine. It is not clear the extent to which tasks should be divided, or co-completed. Typical data analysis sessions involve many user tasks and sub-tasks [AES05], and dividing the effort of these tasks between the user and the system is challenging.

It is also unclear exactly how to measure the amount of effort expended by both the user and the system. For example, in a visual analytic system that helps people cluster documents, Endert et al. used a measure of how many documents were moved and grouped by the user and how many were automatically grouped by the system [EFN12a]. However,

there exist opportunities to consider additional metrics for the balance of effort in mixed-initiative systems that can drive the possibility of novel evaluations of effectiveness.

6.3. Complex Computation Systems can lead to Automation Surprise

By coupling machine learning with visual analytics systems, we can develop complex systems made up of many inter-related and inter-dependent "black boxes" of automated components for data analysis, knowledge discovery and extraction. Complex systems will typically comprise many instances of known and hidden inter-dependencies between components and yield outputs that are emergent where the interactions among agents and individual units may be deterministic. The global behaviour of the system as a whole may conform with rules that are only sometimes deducible from knowledge of the interactions and topology of the system. This makes it difficult to know exactly which inputs contribute to an observed output, and the extent of each factor's contributions [SS11, Orm]. Sarter and Woods [SWB97] observed that interactions between these tightly coupled automated "black boxes" can create consequences and automation surprises that arise from a lack of awareness of system state and the state of the world. This creates potential for error, complacency from trusting the technology, placing new demands on attention, coordination and workload.

At the risk of saying the obvious, an approach proposed by Norman [Nor86] to address some of the problems of controlling complex systems is based on observability and feedback. They are crucial for figuring out how a system works, and they help us affirm the mental models that drive our thinking and analysis of a problem or a device. Poor observability of automated advanced intelligent processes makes it difficult to evaluate if outcomes from the automated computations are within the bounds of normal or acceptable behavior, or whether our instructions to the system were correctly executed or what else was included in the execution that was not intended. Good mapping between designed action and desired action helps us anticipate and learn how to interact with the system. Good mapping also helps us see the connection between what the system was instructed to do, and the outcome of carrying out that instruction.

One of the major challenges then, is for visual analytics designers to create designs that "... facilitate the discovery of meaningfulness of the situation ... not as a property of the mind, but rather as a property of the situation or functional problems that operators are trying to solve ... [by] developing representations that specify the meaningful properties of a work domain ... so that operators can discover these meaningful properties and can guide their actions appropriately" [BF11].

To create such a design, there is a need to have a conception of the analytical thinking and reasoning process that extends beyond the information handling and manipulation aspects that are frequently described. A focus group study with 20

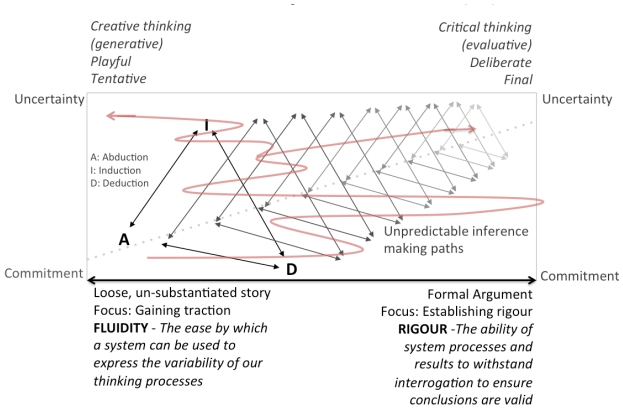


Figure 14: *Characterizing the thinking terrain of analysts [Won14].*

intelligence analysts [WV12], think-aloud studies with 6 analysts performing a simulated intelligence task [RAWC14], and think-aloud studies with 6 librarians carrying out a surrogate task of creating explanations from a literature review task [KAW*13] provide insight into this analytical thinking and reasoning process. The results of these studies indicate that analyst make use of the various inference making strategies described in Section 2.1 - induction, deduction and adduction - depending upon what data they have, the rules for interpreting the data, and premise they are starting with and the conclusions they would make or would like to make. Furthermore, very often they would test the validity of the propositions they arrive at by practicing critical thinking - where they attempt to assess the quality and validity of their thinking and the data they use, the criteria they use for forming judgments, and so forth. In fact, critical thinking is so important that many intelligence analysis training schools have introduced it into their training.

One thing else that is observed to happen alongside all of this is somewhat more subtle: Analysts are constantly trying to explain the situation, sometimes re-constructing the situation from pieces of data and from inferential claims; and then carrying out searches or further analysis to find necessary data back the claims. This process of explanation is crucial to making sense and how it is used to link data, context and inferences. It often starts off as a highly tentative explanation that is based on very weak data or hunches. The analyst then explores this possibility, making conjectures, suppositions and inferential claims, from which they then connect with further data (testing their relevance and significance), elaborate, question, and often reframe and discard, their ideas, and eventually building up the story so that it eventually becomes robust enough to withstand interrogation.

We see a progression - not necessarily in a linear manner - where explanations reflect tentative, creative and playful, and generative thinking, and then transitions towards thinking strategies that are more critical, evaluative, deliberate and final (see Figure 14 for an illustration depicting this

discussion). One can assume a continuum where at one end we have a tentative explanation we call a "loose story" that accounts for the data, and at the other end the loose story has evolved into a strong and more formal argument such that it is rigorous and able to withstand interrogation, say, in a court of law.

At the "formal argument" end of the continuum, there is much lower uncertainty. The analyst is more definite about what the data and their relationships mean, and very likely has become more committed to a particular path of investigation. At this end, the emphasis is on verifying that the data used to construct the conclusions, the claims being made based on the data, and the conclusions themselves, are valid.

The combined machine learning and visual analytics tools to be built should fluidly link the generative, creative, playful and tentative exploration activities that encourage the exploration of alternatives, appreciation of the context, and the avoidance of pre-mature commitment, with the more evaluative, critical inquiry that leads to a deliberate, final and rigorous explanation. This is the notion of the design principle of fluidity and rigour.

6.4. Visualizing Intermediate Results and Computational Process

Many kinds of ML algorithms undergo a continuous convergence process towards the final solution. In general, only this final solution is rendered into a visualization, which may incorporate classical interaction mechanisms (zoom, pan, brushing, focus&context, etc.). This convergence is often done within a fixed context, that includes the training set, the algorithm parameters and the cost function. These elements often convey a large amount of insight for the user, but since they remain fixed during convergence users are deprived of the benefits of interaction. What if the user could steer these fixed elements "during" convergence?.

A promising topic, involving innovation by both VA and ML communities, is rendering visualizations of the intermediate results during convergence, allowing the user to tunesteer the ML algorithms by changing these elements. Designing *ad hoc* ML algorithms with this approach in mind that pave the way for new and useful kinds of interaction mechanisms opens new and exciting research paths. There has been some prior work on this topic. For example, Stolper et al. developed a system for *progressive visual analytics*, where intermediate results of a sequence-mining algorithm running on medical treatment events can be shown to clinicians [SPG14]. Their work gave analysts the ability to see broader results sooner to help decide if the entire computation needed to be executed. Similarly, systems to show partial query results of large datasets [FPDs12] and partial dimension reduction and clustering results [TKBH17] have been recently developed.. These works raise important questions about the tradeoff between accuracy and execution time of these algorithms, and also about how to incorporate user feedback into computation during runtime.

6.5. Enhancing Trust and Interpretability

A key element of the visualization approach is its ability to generate trust in the user. Unlike pure machine learning techniques, in a data visualization the user "sees" the data and information as a part of the analysis. When the visualization is interactive, the user will be part of the loop and involved in driving the visualization. In such a context, the development of a mental model goes hand in hand with the visualization, as everything is part of the process. This tight involvement of the user in the development of the visualization based on the results of previous iterations, along with the highly visual component of human thinking, can make this approach generate a great amount of trust in the user. However, such "trust" can have different meanings at different levels of cognition. An apparently trustable result at an intuitive level can arouse suspicions at a higher cognitive level, demanding methods for statistical confirmation of the results. On a broad view, two different levels can be identified:

1. A "qualitative level", that would make heavy use of perception visualization principles along with interaction mechanisms to present data in an intuitive way. The communication in both senses (from and to the interface) will typically seek to: a) adapt to individual's perception mechanism so that the information throughput and knowledge increment on the user is maximized; and b) in a higher level, to adapt to the human cognitive process so that data and information is presented in a way that is intuitive to the user. The means to carry out this approach would rely on classical visualization methods (adequate use of visual encodings and spatial layouts) and on interaction techniques, including brushing, linking, coordinated views, animated transitions, etc., but also in much more powerful approaches such as user-driven steering of ML algorithms (such as DR, clustering, etc.) resulting in the reconfiguration of the visualization on the basis of changes in the context such as time varying data or changes in the user focus on different types of analysis.
2. A "quantitative level" is, however, needed to provide sound statistical validation of the former visualization results. Taken in an isolated way, this level would lack insight. However, its outcomes are supposed to be trustworthy so the user can consider them as definite validations. Quantitative approaches –mainly belonging to the realm of ML– are in essence deterministic, which makes them less prone to human errors and reproducible. This helps to standardize decisions and provides congruence, accurateness, uniformity and coherence in the results.
   However, quantitative approaches tend to avoid the need for user intervention by trying to automate the process. In general they do not look for human feedback but undertake as many human tasks as possible in the process, automating it to the maximum possible extent, aiming to avoid any kind of human subjectivity and seeking rigor (statistical, mathematical). But many problems in real life are built on sparse bits of knowledge coming from diverse domains. Moreover, such knowledge is often made

of vague or imprecise mental models. Purely quantitative approaches cannot operate with such small, diverse and "fuzzy" bricks; they need solid foundations to be operative.

The previous division is only conceptual. Both approaches can (and should) be combined. For instance, a statistical validation of one or more facts can be displayed on top of the qualitative visualization by making use of visual encodings and text labels. We encourage visual analytics designers to seek efficient combinations between qualitative and quantitative approaches, looking for concurrent visualization of actual problem data and sophisticated computed features, both coexisting in the same representation. The mere fact of representing statistical validations sharing the same layout and structure as the original data in a same visualization allows the user to internalize that quantitative information allowing her to connect it to its domain knowledge, with an unquestionable positive effect on trust and confidence in the results.

6.6. Beyond Current Methods

Currently, many of the applications of machine learning in visual analytics relate to dimensionality reduction. In addition, as discussed in Section 4, there are a different sort of ML methods based on Bayesian inferencing and including topic modeling and textual analytics approaches. These are becoming more prominent. While these applications are undeniably an important use of machine learning, we contend that consideration of the role of the user opens up several new fields of study where machine learning can play an important role. First amongst these is the role of machine learning in creating a computational model for the user's analytical process. This complements cognitive task analysis and aims to model how domain expert users use visual analytics to tackle important tasks, and how they reason about the problem. This will enable better system design to support expert strategies and provide support to less-trained users.

Every user interaction has two primary functions: i) to communicate a direct *explicit* intent from the user to the analytical system and receive an appropriate response (e.g. if the user requests a zoom into a particular area, the system should create that zoomed-in visual display), and ii) to carry out an indirect *implicit* piece of analytical reasoning.

The point is that every user choice in the visual analytics frame is equivalent to a statistical choice in the mathematical frame: we need users to make appropriate choices that do not invalidate the (implied) statistical analysis that they are carrying out. Motivated by the analysis of how users carry out visual analytics, particularly the concepts of sense-making and knowledge generation, the first step to understanding the details of this process is to compile a complete log of users' analytical process and the information that they record. This is the base dataset that can be used for traceability, responsibility and provenance: providing an argued case for others (such as collaborators or managers) to critique and use to make decisions. However, beyond this use, the database is also a resource to mine in order to clarify the decisions that are made in the course of visual analytics, leading to the potential to develop adaptable interfaces and a greater depth of understanding of users' mental models, which can then be used to guide other, perhaps less skilled or experienced, users.

It would not be feasible (nor practically useful) to track every single change in a visualisation. It is essential that the process involves minimal interruption to cognitive flow (so as to avoid damaging the very process we are trying to understand). However, it would be helpful to prompt the user for feedback (preferably in visual ways), in the form of annotations, at certain key points of the analysis. We propose using machine learning (e.g. to look for breakpoints in the way information is displayed) as cues for these prompts. The process model can also learn from user interaction (with appropriate additional guidance). For example, if the user 'undoes' a particular action, it could mean "I don't want this: my choice was wrong" or "The visualisation is useful, but it is a dead end and I need to back-track". Other simple user interactions that can connect to reasoning processes include brushing data points (which corresponds to selecting and labelling a subset of data) and linking (which corresponds to hypothesising correlations between variables and data points).

As a complement to this database of successful analytic practice, what many users need is a way of avoiding bad practice (or errors). A catalogue of 'typical'errors that is searchable (using case-based reasoning tools) could be crowd-sourced from teachers (and their students!) or training courses.

How can machine learning aid the understanding of user processes? At the simplest level, user interactions are a linear sequence of actions: discovering the underlying sequence and the transitions between items is relatively straight-forward, since a Markov (or hidden Markov) model can easily be trained to uncover this structure. However, an unstructured and unannotated sequential list does not contain enough structure to infer the analytical process. Firstly, we need to understand the reasons why a user has made choices (which requires annotations). Secondly, it is clear that the analytical process is not a simple sequence of logical choices leading inexorably to a goal. Instead, the process involves exploratory analysis – trying a range of options and assessing which is the most successful – and back-tracking when results show that a particular line of inquiry is fruitless. These transform what is, in terms of a graphical model, a one-dimensional structure, into a tree or directed acyclic graph.

The theory of Bayesian belief networks (BBNs) is relevant here. There are two aspects of the model that can be learned: the *conditional probability tables* (CPTs) for the links from all the parents of a particular node; and the *structure* of the network (the presence or absence of directed links) which represents the conditional (in)dependence of variables. Learning the CPTs for a given network structure is straightforward: with suitably chosen Bayesian priors (a Dirichlet

distribution), it is a matter of counting co-occurrences of value pairs in a dataset [SDLC93]. Learning the structure of a BBN is much more complex: in fact, the general case is NP-hard [Chi96]. Some special cases (such as trees) are tractable, but in this domain it is preferable to fix the structure based on our understanding of the users' analytical process. Models for this process, such as CRISP-DM [WH00] (used in data mining) or those drawn from the infovis community (such as the semantic interaction pipeline), are currently rather high-level, and a more detailed task analysis is necessary before the requisite level of detail for a full computational model can be achieved.

Once a computational user model for the analytic process is established, there are a number of other ways machine learning and visual analytics can be brought into dialogue.

1. Semi-automated report generation. Machine learning can be used to infer links and relations between concepts, data, and analytical results, while frequentist or Bayesian statistical analysis can be used to attach a statistical significance to each finding. This could be presented to the user as a checklist of automatically discovered analytical findings (or hints) that the user can accept or reject.
2. Annotations can be categorised using automated topic analysis (for example by Natural Language Processing that uses probabilistic graphical models [LHE10]). The value of this is to link annotations and find common approaches to tasks.
3. Model-based layout. The goal is to provide a semi-automated way of modifying the layout of visual information. One aspect of this is related to the steerable DR discussed in Section 5. This can be extended to learning the criteria that analysts use: for example, how the user selects principal components.
4. Extreme value theory [DHF07] to identify low-frequency (but potentially high-value) data points or variables. Recent research in this area supports the automated identification of outliers even in the multivariate case.
5. Integrated prior knowledge and data. Often the expert user will have a great deal of prior cognitive knowledge embodied in a computational model of a physical system (e.g. geochemists supporting hydrocarbon exploration; meteorologists). Machine learning can be used to generate an *emulator*, a technique for model reduction that reduces the exceptionally high computational burden imposed by many physical models, while retaining the key features of the original model and allowing much greater user interaction for tasks such as sensitivity analysis and control [CO10].

It is clear from the discussion throughout this paper that there are barriers to the closer integration of machine learning and visual analytics. One of the main technical barriers is that the current software tools are strongly divided between the research communities. Visualization tools are strong at close control over the form and layout of information, and user interaction: Some tend to be written as bespoke integrated tools, such as Tableau (http://www.tableau.com), Orange (orange.biolab.si) and JMP (www.jpm.com). On the other hand, the most advanced machine-learning tools are often written as libraries in numerical or statistical languages (such as Matlab, e.g. [Nab02] and R), as well, as in high level general purpose languages, like Java (with Weka, a widely used collection of ML algorithms for data mining tasks, or the Stanford NLP tools with advanced ML algorithms for natural language processing) or Python (with powerful and widely adopted data analysis and ML libraries like scipy, scikit-learn, pandas, etc.); all of them focus on supporting the (often) challenging task of learning complex models from data but provide limited graphical display and interaction. The best solution to this problem, short of reimplementing large toolkits in other languages is to take a client-server approach: a backend server running a good mathematical package for the machine-learning components complemented by web services and html+js clients, able to take advantage of the huge and growing spectrum of javascript libraries and frameworks (such as d3js) to provide interactive information visualisation.

7. Conclusions

This paper provides a comprehensive survey of machine learning methods, and visual analytics systems that effectively integrate machine learning. Based on this survey, we present a set of opportunities that offer a rich set of ideas to further the integration between these two scientific areas. Among these are formalizing and establishing steerable ML, generally providing coupled interaction and visualization methods that offer substantially more advanced user feedback. There is the opportunity to better determine how tasks should be divided between humans and machines, perhaps in a dynamic manner, including determining metrics for a balance of effort between these two components. The paper shows how recent models and frameworks could be used to develop considerably more powerful visual analytic systems with integrated machine learning. The summary and discussion presented in this paper seeks to excite and challenge researchers from the two disciplines to work together to tackle the challenges raised, ultimately creating more impactful systems to help people gain insight into data.

8. Acknowledgments

The authors would like to acknowledge that much of the content and inspiration for this paper originated during a Dagstuhl Seminar titled, "Bridging Machine Learning with Information Visualization (15101)" [KMRV15]. In addition, funding for the authors was provided in part by the Analysis in Motion Initiative at PNNL, Spanish Ministry of Economy & Competitivity and FEDER funds, under grant DPI2015-69891-C2-2-R

9. Author Bios

Alex Endert is an Assistant Professor in the School of Interactive Computing at Georgia Tech, where he directs the Visual Analytics Lab. In 2013, his work on Semantic Interaction was awarded the IEEE VGTC VPG Pioneers

Group Doctoral Dissertation Award, and the Virginia Tech CS Best Dissertation Award.

William Ribarsky is the Bank of America Endowed Chair and Director of the Charlotte Visualization Center at UNC Charlotte. He is one of the founders of the field of visual analytics and was Chair of the IEEE Visualization Analytics Science and Technology (VAST) Steering Committee until October, 2015.

Cagatay Turkay is a Lecturer (Assistant Prof.) at the Department of Computer Science at City, University of London. He carries out his research at the giCentre and develops methods where interactive visualisations and computational tools are used in tandem for informed analysis processes.

William Wong is a professor of Human-Computer Interaction at Middlesex University, London. He is the head of the Interaction Design Centre. His research interest include investigating the problems of visual analytics in sense-making domains with high information density, such as intelligence analysis, financial systemic risk analysis, and low literacy users.

Ian Nabney is a Professor at Aston University. He is Director of the System Analytics Research Institute. His research interests are in machine learning, particularly in data visualisation, time series, and Bayesian methods. His application interests are broad, and include condition monitoring, biomedical engineering, and urban science.

Ignacio Díaz Blanco is an associate professor at the Department of Electrical Engineering at the University of Oviedo. He researches in intelligent data analysis, data visualization, control and signal processing to understand, diagnose and optimize processes and complex systems.

Fabrice Rossi is a Professor at Paris 1 Panthéon Sorbonne University. He is a member of the SAMM research group and the head of the statistical learning team of this group. His research interests include machine learning and data analysis. He is particularly interested in interpretable systems.

References

[20008] Alias-i. 2008. LingPipe 4.0.1, 2008. URL: http://alias-i.com/lingpipe. 13

[AAB*10] Andrienko G., Andrienko N., Bremm S., Schreck T., Von Landesberger T., Bak P., Keim D.: Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns. In *Computer Graphics Forum* (2010), vol. 29, Wiley Online Library, pp. 913–922. 8, 10

[AAR*09] Andrienko G., Andrienko N., Rinzivillo S., Nanni M., Pedreschi D., Giannotti F.: Interactive visual clustering of large collections of trajectories. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on* (2009), IEEE, pp. 3–10. 8, 10

[ABV14] Awasthi P., Balcan M., Voevodski K.: Local algorithms for interactive clustering. In *Proceedings of The 31st International Conference on Machine Learning* (2014), Xing E. P., Jebara T., (Eds.), vol. 32 of *JMLR Proceedings*, JMLR.org, pp. 550–558. 8, 12

[ACKK14] Amershi S., Cakmak M., Knox W. B., Kulesza T.: Power to the people: The role of humans in interactive machine learning. *AI Magazine 35*, 4 (2014), 105–120. 5, 11

[ADT*13] Amir E.-a. D., Davis K. L., Tadmor M. D., Simonds E. F., Levine J. H., Bendall S. C., Shenfeld D. K., Krishnaswamy S., Nolan G. P., Pe'er D.: visne enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature biotechnology 31*, 6 (2013), 545–552. 8, 16

[AEN10] Andrews C., Endert A., North C.: Space to think: large high-resolution displays for sensemaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), ACM, pp. 55–64. 13

[AES05] Amar R., Eagan J., Stasko J.: Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005* (2005), pp. 111–117. doi:10.1109/INFVIS.2005.1532136. 17, 18

[Alp14] Alpaydin E.: *Introduction to machine learning*. MIT press, 2014. 7

[AN12] Andrews C., North C.: Analyst's Workspace: An embodied sensemaking environment for large, high-resolution displays. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2012), pp. 123–131. doi:10.1109/VAST.2012.6400559. 13

[Anc12] Ancona D.: Framing and acting in the unknown. *S. Snook, N. Nohria, & R. Khurana, The Handbook for Teaching Leadership* (2012), 3–19. 2

[AW12] Ahmed Z., Weaver C.: An Adaptive Parameter Space-Filling Algorithm for Highly Interactive Cluster Exploration. In *Procedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)* (2012). 8, 9

[BBM04] Basu S., Banerjee A., Mooney R. J.: Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM International Conference on Data Mining* (2004), pp. 333–344. URL: http://epubs.siam.org/doi/abs/10.1137/1.9781611972740.31, arXiv:http://epubs.siam.org/doi/pdf/10.1137/1.9781611972740.31, doi:10.1137/1.9781611972740.31. 8, 12

[BDW08] Basu S., Davidson I., Wagstaff K.: *Constrained clustering: Advances in algorithms, theory, and applications*. CRC Press, 2008. 8, 11

[BEK14] Best D. M., Endert A., Kidwell D.: 7 Key Challenges for Visualization in Cyber Network Defense. In *Proceedings of the Eleventh Workshop on Visualization for Cyber Security* (New York, NY, USA, 2014), VizSec '14, ACM, pp. 33–40. doi:10.1145/2671491.2671497. 14

[BF11] Bennett K. B., Flach J. M.: *Display and interface design: Subtle science, exact art*. CRC Press, 2011. 19

[BH12] Balcan M. F., Hanneke S.: Robust interactive learning. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)* (Edinburgh, Scotland, June 2012), vol. 23 of *JMLR Workshop and Conference Proceedings*. 5

[BKSS14] Behrisch M., Korkmaz F., Shao L., Schreck T.: Feedback-driven interactive exploration of large multidimensional data supported by visual classifier. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on* (2014), IEEE, pp. 43–52. 8, 12

[BL09] Blei D., Lafferty J.: Text mining: Theory and applications, chapter topic models, 2009. 13

[BLBC12] Brown E. T., Liu J., Brodley C. E., Chang R.: Dis-function: Learning distance functions interactively. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on* (2012), IEEE, pp. 83–92. 8, 11, 17

[BM13] Bryan N. J., Mysore G. J.: An efficient posterior regularized latent variable model for interactive sound source separation. In *Proceedings of The 30th International Conference on Machine Learning (ICML)* (Atlanta, Georgia, USA, 2013), Dasgupta S., McAllester D., (Eds.), vol. 28 of *JMLR Workshop and Conference Proceedings*. 14

[BNJ03] Blei D. M., Ng A. Y., Jordan M. I.: Latent dirichlet allocation. *the Journal of machine Learning research 3* (2003), 993–1022. 9, 13

[BOZ*14] Brown E. T., Ottley A., Zhao H., Lin Q., Souvenir R., Endert A., Chang R.: Finding Waldo: Learning about Users from their Interactions. 18

[BPFG11] Berger W., Piringer H., Filzmoser P., Gröller E.: Uncertainty-aware exploration of continuous parameter spaces using multivariate prediction. *Computer Graphics Forum 30*, 3 (2011), 911–920. 10

[CBN*12] Choo J., Bohn S., Nakamura G., White A. M., Park H.: Heterogeneous data fusion via space alignment using nonmetric multidimensional scaling. In *SDM* (2012), SIAM, pp. 177–188. 14

[CCM08] Cohn D., Caruana R., McCallum A.: Semi-supervised clustering with user feedback. In *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, Basu S., Davidson I., Wagstaff K., (Eds.). CRC Press, 2008, ch. 2, pp. 17–32. 8, 11

[CDS09] Crossno P. J., Dunlavy D. M., Shead T. M.: Lsaview: a tool for visual exploration of latent semantic modeling. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on* (2009), IEEE, pp. 83–90. 13

[CDW*16] Cho I., Dou W., Wang D. X., Sauda E., Ribarsky W.: Vairoma: A visual analytics system for making sense of places, times, and events in roman history. *Visualization and Computer Graphics, IEEE Transactions on 22*, 1 (2016), 210–219. 4, 13, 17

[Chi96] Chickering D. M.: Learning bayesian networks is np-complete. In *Learning from data*. Springer, 1996, pp. 121–130. 22

[Chr06] Chris N.: Toward Measuring Visualization Insight. 6–9. doi:10.1109/MCG.2006.70. 2

[CLKP10] Choo J., Lee H., Kihm J., Park H.: ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on* (2010), IEEE, pp. 27–34. 8, 9, 14

[CMS99] Card S. K., Mackinlay J. D., Shneiderman B.: *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., 1999. 2

[CO10] Conti S., OHagan A.: Bayesian emulation of complex multi-output and dynamic computer models. *Journal of statistical planning and inference 140*, 3 (2010), 640–651. 22

[CP13] Choo J., Park H.: Customizing Computational Methods for Visual Analytics with Big Data. *Computer Graphics and Applications, IEEE 33*, 4 (2013), 22–28. doi:10.1109/MCG.2013.39. 8, 11, 16

[DCE*15] Dou W., Cho I., ElTayeby O., Choo J., Wang X., Ribarsky W.: Demographicvis: Analyzing demographic information based on user generated content. In *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on* (2015), IEEE, pp. 57–64. 13

[DCP*14] Díaz I., Cuadrado A. A., Pérez D., García F. J., Verleysen M.: Interactive Dimensionality Reduction for Visual Analytics. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (Bruges, Belgium, 2014). 17

[DCV16] Díaz I., Cuadrado A. A., Verleysen M.: A state-space model on interactive dimensionality reduction. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (Bruges, Belgium, April 2016), Verleysen M., (Ed.), pp. 647–652. 16

[DFB11] Drucker S. M., Fisher D., Basu S.: Helping users sort faster with adaptive machine learning recommendations. Springer-Verlag, pp. 187–203. 18

[DHF07] De Haan L., Ferreira A.: *Extreme value theory: an introduction*. Springer Science & Business Media, 2007. 22

[DWCR11] Dou W., Wang X., Chang R., Ribarsky W.: Paralleltopics: A probabilistic approach to exploring document collections. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (2011), IEEE, pp. 231–240. 13

[DWS*12] Dou W., Wang X., Skau D., Ribarsky W., Zhou M. X.: Leadline: Interactive visual analysis of text data through event identification and exploration. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on* (2012), IEEE, pp. 93–102. 13

[EBN13] Endert A., Bradel L., North C.: Beyond Control Panels: Direct Manipulation for Visual Analytics. *IEEE Computer Graphics and Applications 33*, 4 (2013), 6–13. doi:10.1109/MCG.2013.53. 8, 11, 16

[ECNZ15] Endert A., Chang R., North C., Zhou M.: Semantic Interaction: Coupling Cognition and Computation through Usable Interactive Analytics. *IEEE Computer Graphics and Applications 35*, 4 (July 2015), 94–99. doi:10.1109/MCG.2015.91. 18

[EFN12a] Endert A., Fiaux P., North C.: Semantic Interaction for Sensemaking: Inferring Analytical Reasoning for Model Steering. *Visualization and Computer Graphics, IEEE Transactions on 18* (2012), 2879–2888. 12. doi:10.1109/tvcg.2012.260. 18

[EFN12b] Endert A., Fiaux P., North C.: Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), ACM, pp. 473–482. 4, 5

[EHM*11] Endert A., Han C., Maiti D., House L., Leman S. C., North C.: Observation-level Interaction with Statistical Models for Visual Analytics. In *IEEE VAST* (2011), pp. 121–130. 8, 11

[EHR*14] Endert A., Hossain M. S., Ramakrishnan N., North C., Fiaux P., Andrews C.: The human is the loop: new directions for visual analytics. *Journal of Intelligent Information Systems* (Jan. 2014), 1–25. doi:10.1007/s10844-014-0304-9. 5

[EKHW08] Eccles R., Kapler T., Harper R., Wright W.: Stories in GeoTime. *Information Visualization 7* (2008), 3–17. 1. doi:10.1145/1391107.1391109. 13

[EMJ*11] Elmqvist N., Moere A. V., Jetter H.-C., Cernea D., Reiterer H., Jankun-Kelly T.: Fluid interaction for information visualization. *Information Visualization 10*, 4 (2011), 327–340. 16

[EPT*05] Elm W., Potter S., Tittle J., Woods D., Grossman J., Patterson E.: Finding decision support requirements for effective intelligence analysis tools. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (2005), vol. 49, SAGE Publications, pp. 297–301. 2

[FHT01] Friedman J., Hastie T., Tibshirani R.: *The elements of statistical learning*, vol. 1. Springer series in statistics Springer, Berlin, 2001. 7

[FJA*11] Fernstad S., Johansson J., Adams S., Shaw J., Taylor D.: Visual exploration of microbial populations. In *Biological Data Visualization (BioVis), 2011 IEEE Symposium on* (2011), pp. 127 –134. 8

[FMK12] Fischer F., Mansmann F., Keim D. A.: Real-time visual analytics for event data streams. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing* (New York, NY, USA, 2012), SAC '12, ACM, pp. 801–806. URL: http://doi.acm.org/10.1145/2245276.2245432, doi:10.1145/2245276.2245432. 14

[FOJ03] Fails J. A., Olsen Jr D. R.: Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces* (2003), ACM, pp. 39–45. 5, 6

[FP07] Fu W.-T., Pirolli P.: Snif-act: A cognitive model of user navigation on the world wide web. *Human–Computer Interaction 22*, 4 (2007), 355–412. 4

[FPDs12] Fisher D., Popov I., Drucker S., schraefel m.: Trust Me, I'M Partially Right: Incremental Visualization Lets Analysts Explore Large Datasets Faster. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2012), CHI '12, ACM, pp. 1673–1682. doi:10.1145/2207676.2208294. 20

[FSB*13] Fiaux P., Sun M., Bradel L., North C., Ramakrishnan N., Endert A.: Bixplorer: Visual Analytics with Biclusters. *Computer 46*, 8 (2013), 90–94. 16

[FWG09] Fuchs R., Waser J., Gröller M. E.: Visual human+machine learning. *IEEE TVCG 15*, 6 (2009), 1327–1334. 8

[GB11] Guillory A., Bilmes J.: Simultaneous learning and covering with adversarial noise. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (New York, NY, USA, June 2011), Getoor L., Scheffer T., (Eds.), ICML '11, ACM, pp. 369–376. 5

[GNRM08] Garg S., Nam J. E., Ramakrishnan I., Mueller K.: Model-driven visual analytics. In *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on* (2008), IEEE, pp. 19–26. 8, 11

[GOB*10] Gehlenborg N., O'Donoghue S., Baliga N., Goesmann A., Hibbs M., Kitano H., Kohlbacher O., Neuweger H., Schneider R., Tenenbaum D., et al.: Visualization of omics data for systems biology. *Nature methods 7* (2010), S56–S68. 15

[GRF09] Green T. M., Ribarsky W., Fisher B.: Building and applying a human cognition model for visual analytics. *Information Visualization 8* (2009), 1–13. 1. doi:10.1057/palgrave.ivs.2008.28. 4

[GRVE07] Grottel S., Reina G., Vrabec J., Ertl T.: Visual verification and analysis of cluster detection for molecular dynamics. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1624–1631. doi:http://dx.doi.org/10.1109/TVCG.2007.70614. 16

[GS04] Griffiths T. L., Steyvers M.: Finding scientific topics. *Proceedings of the National Academy of Sciences 101*, suppl 1 (2004), 5228–5235. 13

[GSFS06] Gray W. D., Sims C. R., Fu W.-T., Schoelles M. J.: The soft constraints hypothesis: a rational analysis approach to resource allocation for interactive behavior. *Psychological review 113*, 3 (2006), 461. 17

[GWHR01] Gahegan M., Wachowicz M., Harrower M., Rhyne T.-M.: The integration of geographic visualization with knowledge discovery in databases and geocomputation. *Cartography and Geographic Information Science 28*, 1 (2001), 29–44. 3

[HB11] Hoober S., Berkman E.: *Designing mobile interfaces.* " O'Reilly Media, Inc.", 2011. 17

[HBGSS14] Hu Y., Boyd-Graber J., Satinoff B., Smith A.: Interactive topic modeling. *Machine Learning 95*, 3 (2014), 423–469. doi:10.1007/s10994-013-5413-0. 13

[HBM*13] Hu X., Bradel L., Maiti D., House L., North C., Leman S.: Semantics of directly manipulating spatializations. *Visualization and Computer Graphics, IEEE Transactions on 19*, 12 (2013), 2052–2059. 8, 11

[Hee06] Heer J.: prefuse manual, 2006. URL: http://prefuse.org. 4

[HHE*13] Henry M. J., Hampton S., Endert A., Roberts I., Payne D.: MultiFacet: A Faceted Interface for Browsing Large Multimedia Collections. In *2013 IEEE International Symposium on Multimedia (ISM)* (Dec. 2013), pp. 347–350. doi:10.1109/ISM.2013.66. 8, 14

[HOG*12] Hossain M. S., Ojili P. K. R., Grimm C., Müller R., Watson L. T., Ramakrishnan N.: Scatter/gather clustering: Flexibly incorporating user feedback to steer clustering results. *Visualization and Computer Graphics, IEEE Transactions on 18*, 12 (2012), 2829–2838. 8, 11

[Hor99] Horvitz E.: Principles of Mixed-initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 1999), CHI '99, ACM, pp. 159–166. doi:10.1145/302979.303030. 18

[HSCW13] Hadlak S., Schumann H., Cap C. H., Wollenberg T.: Supporting the visual analysis of dynamic networks by clustering associated temporal attributes. *Visualization and Computer Graphics, IEEE Transactions on 19*, 12 (2013), 2267–2276. 8, 9

[i2] i2 Analyst's Notebook. URL: http://www.i2inc.com/products/analysts_notebook/. 13

[IHG13] Iwata T., Houlsby N., Ghahramani Z.: Active learning for interactive visualization. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013* (Scottsdale, AZ, USA, April 2013), vol. 31 of *JMLR Proceedings*, JMLR.org, pp. 342–350. URL: http://dblp.uni-trier.de/db/conf/aistats/aistats2013.html#IwataHG13. 8, 11

[IYU08] Iwata T., Yamada T., Ueda N.: Probabilistic latent semantic visualization: topic model for visualizing documents. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (2008), ACM, pp. 363–371. 13

[JBS08] Jänicke H., Böttinger M., Scheuermann G.: Brushing of attribute clouds for the visualization of multivariate data. *IEEE Transactions on Visualization and Computer Graphics* (2008), 1459–1466. 8

[JJ09] Johansson S., Johansson J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. *Visualization and Computer Graphics, IEEE Transactions on 15*, 6 (2009), 993–1000. 7, 8

[JZF*09] Jeong D. H., Ziemkiewicz C., Fisher B., Ribarsky W., Chang R.: iPCA: An Interactive System for PCA-based Visual Analytics. *Computer Graphics Forum 28*, 3 (2009), 767–774. doi:10.1111/j.1467-8659.2009.01475.x. 8, 17

[Kan12] Kandogan E.: Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on* (2012), IEEE, pp. 73–82. 8, 9

[KAW*13] Kodagoda N., Attfield S., Wong B., Rooney C., Choudhury S.: Using interactive visual reasoning to support sense-making: Implications for design. *Visualization and Computer Graphics, IEEE Transactions on 19*, 12 (2013), 2217–2226. 19

[KBK11] Krstajić M., Bertini E., Keim D. A.: Cloudlines: Compact display of event episodes in multiple time-series. *Visualization and Computer Graphics, IEEE Transactions on 17*, 12 (2011), 2432–2439. 13

[KCPE16] Kim H., Choo J., Park H., Endert A.: Interaxis: Steering scatterplot axes via observation-level interaction. *Visualization and Computer Graphics, IEEE Transactions on 22*, 1 (Jan 2016), 131–140. doi:10.1109/TVCG.2015.2467615. 8, 11

[Kei02] Keim D. A.: Information visualization and visual data mining. *Visualization and Computer Graphics, IEEE Transactions on 8* (2002), 1–8. 1. 1

[KG90] Kendall M., Gibbons J. D.: *Rank correlation methods*. Oxford University Press, 1990. 12

[KGL*15] Klemm P., Glaer S., Lawonn K., Rak M., Vlzke H., Hegenscheid K., Preim B.: Interactive visual analysis of lumbar back pain - what the lumbar spine tells about your life. In *Proceedings of the 6th International Conference on Information Visualization Theory and Applications (VISIGRAPP 2015)* (2015), pp. 85–92. doi:10.5220/0005235500850092. 8, 9

[KK08] Kumar N., Kummamuru K.: Semisupervised clustering with metric learning using relative comparisons. *IEEE Transactions on Knowledge and Data Engineering 20*, 4 (April 2008), 496–503. doi:10.1109/TKDE.2007.190715. 8, 12

[KKP05] Kumar N., Kummamuru K., Paranjpe D.: Semi-supervised clustering with metric learning using relative comparisons. In *Fifth IEEE International Conference on Data Mining (ICDM'05)* (Nov 2005). doi:10.1109/ICDM.2005.128. 8, 12

[KKW*16] Kwon B. C., Kim H., Wall E., Choo J., Park H., Endert A.: Axisketcher: Interactive nonlinear axis mapping of visualizations through user drawings. *IEEE transactions on visualization and computer graphics* (2016). 8, 11

[KLG*16] Klemm P., Lawonn K., Glasser S., Niemann U., Hegenscheid K., Volzke H., Preim B.: 3d regression heat map analysis of population study data. *Visualization and Computer Graphics, IEEE Transactions on 22*, 1 (2016), 81–90. 8, 10

[KMH06a] Klein G., Moon B., Hoffman R.: Making Sense of Sensemaking 1: Alternative Perspectives. *IEEE Intelligent Systems 21*, 4 (2006), 70–73. doi:10.1109/MIS.2006.75. 3

[KMH06b] Klein G., Moon B., Hoffman R.: Making Sense of Sensemaking 2: A Macrocognitive Model. *IEEE Intelligent Systems 21*, 5 (2006), 88–92. doi:10.1109/MIS.2006.100. 3

[KMRV15] Keim D. A., Munzner T., Rossi F., Verleysen M. (Eds.):. *Bridging Information Visualization with Machine Learning (Dagstuhl Seminar 15101)* (Dagstuhl, Germany, 2015), vol. 5, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. URL: http://drops.dagstuhl.de/opus/volltexte/2015/5266, doi:http://dx.doi.org/10.4230/DagRep.5.3.1. 1, 7, 22

[KMSZ06] Keim D. A., Mansmann F., Schneidewind J., Ziegler H.: Challenges in Visual Data Analysis. IEEE Computer Society. doi:10.1109/iv.2006.31. 1

[KP11] Kaski S., Peltonen J.: Dimensionality reduction for data visualization. *IEEE Signal Processing Magazine 28*, 2 (2011), 100–104. doi:10.1109/MSP.2010.940003. 8, 16

[KPB14] Krause J., Perer A., Bertini E.: Infuse: interactive feature selection for predictive modeling of high dimensional data. *Visualization and Computer Graphics, IEEE Transactions on 20*, 12 (2014), 1614–1623. 8, 9, 10

[KS11] Kang Y.-a., Stasko J.: Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (2011), IEEE, pp. 21–30. 3

[KSF*08] Kerren A., Stasko J., Fekete J.-D., North C., Keim D., Andrienko G., Görg C., Kohlhammer J., Melançon G.: Visual Analytics: Definition, Process, and Challenges. In *Information Visualization*, vol. 4950 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2008, pp. 154–175. doi:10.1007/978-3-540-70956-5_7. 1

[LD97] Landauer T. K., Dumais S. T.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review 104*, 2 (1997), 211. 13

[LHE10] Lin C., He Y., Everson R.: A comparative study of bayesian models for unsupervised sentiment detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning* (2010), Association for computational linguistics, pp. 144–152. 22

[LHM*11] Leman S. C., House L., Maiti D., Endert A., North C.: *A Bi-directional Visualization Pipeline that Enables Visual to Parametric Interation (V2PI)*. Tech. rep., 2011. FODAVA-10-41. 5

[LI57] Lonergan B. J., Insight A.: A study of human understanding. *New York 298* (1957). 2

[LKT*14] Lu Y., Krüger R., Thom D., Wang F., Koch S., Ertl T., Maciejewski R.: Integrating predictive analytics and social media. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct 2014), pp. 193–202. doi:10.1109/VAST.2014.7042495. 8, 15

[LPP*06] Lee B., Plaisant C., Parr C. S., Fekete J.-D., Henry N.: Task taxonomy for graph visualization. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization* (2006), ACM, pp. 1–5. 17

[LSP*10] Lex A., Streit M., Partl C., Kashofer K., Schmalstieg D.: Comparative analysis of multidimensional, quantitative data. *IEEE Transactions on Visualization and Computer Graphics (Proceedings Visualization / Information Visualization 2010) 16*, 6 (2010), 1027–1035. 8

[LSS*12] Lex A., Streit M., Schulz H.-J., Partl C., Schmalstieg D., Park P. J., Gehlenborg N.: StratomeX: visual analysis of large-scale heterogeneous genomics data for cancer subtype characterization. *Computer Graphics Forum (EuroVis '12) 31*, 3 (2012), 1175?1184. doi:10.1111/j.1467-8659.2012.03110.x. 8

[LYK*12] Luo D., Yang J., Krstajic M., Ribarsky W., Keim D.: Eventriver: Visually exploring text collections with temporal references. *Visualization and Computer Graphics, IEEE Transactions on 18*, 1 (2012), 93–105. 13, 14

[LZP*09] Liu S., Zhou M. X., Pan S., Qian W., Cai W., Lian X.: Interactive, topic-based visual text summarization and analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management* (2009), ACM, pp. 543–552. 13

[MBD*11] May T., Bannach A., Davey J., Ruppert T., Kohlhammer J.: Guiding feature subset selection with an interactive visualization. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (2011), IEEE, pp. 111–120. 8, 9

[MGJH08] Matković K., Gračanin D., Jelović M., Hauser H.: Interactive visual steering-rapid visual prototyping of a common rail injection system. *Visualization and Computer Graphics, IEEE Transactions on 14*, 6 (2008), 1699–1706. 8, 12

[MGS*14] Matkovic K., Gracanin D., Splechtna R., Jelovic M., Stehno B., Hauser H., Purgathofer W.: Visual analytics for complex engineering systems: Hybrid visual steering of simulation ensembles. *Visualization and Computer Graphics, IEEE Transactions on 20*, 12 (2014), 1803–1812. 8, 12

[MJR*11] MacEachren A. M., Jaiswal A., Robinson A. C., Pezanowski S., Savelyev A., Mitra P., Zhang X., Blanford J.: Senseplace2: Geotwitter analytics support for situational awareness. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (2011), IEEE, pp. 181–190. 13

[MK08] May T., Kohlhammer J.: Towards closing the analysis gap: Visual generation of decision supporting schemes from raw data. In *Computer Graphics Forum* (2008), vol. 27, Wiley Online Library, pp. 911–918. 8, 9

[MME*12] Malik A., Maciejewski R., Elmqvist N., Jang Y., Ebert D. S., Huang W.: A correlative analysis process in a visual analytics environment. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on* (2012), IEEE, pp. 33–42. 8, 10

[MO04] Madeira S. C., Oliveira A. L.: Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 1*, 1 (2004), 24–45. 16

[MP13] Muhlbacher T., Piringer H.: A partition-based framework for building and validating regression models. *Visualization and Computer Graphics, IEEE Transactions on 19*, 12 (2013), 1962–1971. 8, 10

[Mur85] Murtagh F.: A survey of algorithms for contiguity-constrained clustering and related problems. *The computer journal 28*, 1 (1985), 82–88. 11

[NM13] Nam J., Mueller K.: Tripadvisorn-d: A tourism-inspired high-dimensional space exploration framework with overview and detail. *Visualization and Computer Graphics, IEEE Transactions on 19*, 2 (2013), 291–305. 8

[Nor86] Norman D. A.: Cognitive engineering. *User centered system design: New perspectives on human-computer interaction 3161* (1986). 19

[Orm] Ormand C.: What constitutes a complex system? (http://serc.carleton.edu/nagtworkshops/complexsystems/introduction.html). 19

[OST*10] Oesterling P., Scheuermann G., Teresniak S., Heyer G., Koch S., Ertl T., Weber G. H.: Two-stage framework for a topology-based projection and visualization of classified document collections. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on* (2010), IEEE, pp. 91–98. 13

[PBK10] Piringer H., Berger W., Krasser J.: Hypermoval: Interactive visual validation of regression models for real-time simulation. In *Proceedings of the 12th Eurographics / IEEE - VGTC Conference on Visualization* (Aire-la-Ville, Switzerland, Switzerland, 2010), EuroVis'10, Eurographics Association, pp. 983–992. URL: http://dx.doi.org/10.1111/j.1467-8659.2009.01684.x, doi:10.1111/j.1467-8659.2009.01684.x. 8, 10

[PC05] Pirolli P., Card S.: The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis* (2005), vol. 5, pp. 2–4. 2

[PES*06] Poulin B., Eisner R., Szafron D., Lu P., Greiner R., Wishart D. S., Fyshe A., Pearcy B., Macdonell C., Anvik J.: Visual explanation of evidence with additive classifiers. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006,* (Boston, Massachusetts, USA, July 2006), AAAI Press, pp. 1822–1829. URL: http://www.aaai.org/Library/AAAI/2006/aaai06-301.php. 8, 9

[PSCO09] Pike W. A., Stasko J., Chang R., O'Connell T. A.: The science of interaction. *Information Visualization 8*, 4 (2009), 263–274. 16, 18

[PSPM15] Paiva J. G. S., Schwartz W. R., Pedrini H., Minghim R.: An approach to supporting incremental visual data classification. *Visualization and Computer Graphics, IEEE Transactions on 21*, 1 (2015), 4–17. 8, 12

[PTH13] Porter R., Theiler J., Hush D.: Interactive machine learning in data exploitation. *Computing in Science and Engineering 15*, 5 (2013), 12–20. doi:http://doi.ieeecomputersociety.org/10.1109/MCSE.2013.74. 5

[PTRV13] Parulek J., Turkay C., Reuter N., Viola I.: Visual cavity analysis in molecular simulations. *BMC Bioinformatics 14*, 19 (2013), 1–15. 8, 9

[PZS*15] Pérez D., Zhang L., Schaefer M., Schreck T., Keim D., Díaz I.: Interactive feature space extension for multidimensional data projection. *Neurocomputing 150, Part* (2015), 611–626. URL: http://www.sciencedirect.com/science/article/pii/S0925231214012879, doi:http://dx.doi.org/10.1016/j.neucom.2014.09.061. 8, 17

[RAWC14] Rooney C., Attfield S., Wong B. W., Choudhury S.: Invisque as a tool for intelligence analysis: the construction of explanatory narratives. *International Journal of Human-Computer Interaction 30*, 9 (2014), 703–717. 19

[RECC10] Rose S., Engel D., Cramer N., Cowley W.: Automatic Keyword Extraction from Individual Documents. John Wiley & Sons, Ltd, pp. 1–20. doi:10.1002/9780470689646.ch1. 12

[RF16] Ribarsky W., Fisher B.: The human-computer system: Towards an operational model for problem-solving. In *Hawaii International Conference on Systems Science (HICSS 2016)* (2016). 3, 17

[RK04] Rasmussen M., Karypis G.: *gCLUTO–An Interactive Clustering, Visualization, and Analysis System., University of Minnesota, Department of Computer Science and Engineering, CSE.* Tech. rep., UMN Technical Report: TR, 2004. 8

[RPN*08] Rinzivillo S., Pedreschi D., Nanni M., Giannotti F., Andrienko N., Andrienko G.: Visually driven analysis of movement data by progressive clustering. *Information Visualization 7*, 3 (2008), 225–239. 8, 9

[RSE09] Rudolph S., Savikhin A., Ebert D. S.: Finvis: Applied visual analytics for personal financial planning. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on* (2009), IEEE, pp. 195–202. 14

[RSPC93] Russell D. M., Stefik M. J., Pirolli P., Card S. K.: The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (1993), ACM, pp. 269–276. 2

[SBTK08] Schreck T., Bernard J., Tekusova T., Kohlhammer J.: Visual cluster analysis of trajectory data with interactive Kohonen Maps. In *IEEE Symposium on Visual Analytics Science and Technology, 2008. VAST'08* (2008), pp. 3–10. 8, 9

[SDLC93] Spiegelhalter D. J., Dawid A. P., Lauritzen S. L., Cowell R. G.: Bayesian analysis in expert systems. *Statistical science* (1993), 219–247. 22

[SDMT16] Stahnke J., Dork M., Muller B., Thom A.: Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. *Visualization and Computer Graphics, IEEE Transactions on 22*, 1 (2016), 629–638. 8

[Set09] Settles B.: *Active Learning Literature Survey.* Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. 5, 8, 12

[SGG*14] Streit M., Gratzl S., Gillhofer M., Mayr A., Mitterecker A., Hochreiter S.: Furby: fuzzy force-directed bicluster visualization. *BMC bioinformatics 15*, Suppl 6 (2014), S4. 8, 16

[SGL08] Stasko J., Goerg C., Liu Z.: Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization 7* (2008), 118–132. 2. doi:10.1145/1466620.1466622. 13

[SHA*01] Shipman F., Hsieh H., Airhart R., Maloor P., Moore J. M., Shah D.: Emergent Structure in Analytic Workspaces: Design and Use of the Visual Knowledge Builder. pp. 132–139. 12

[She00] Shearer C.: The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing 5*, 4 (2000), 13–22. 5

[Shn83] Shneiderman B.: Direct Manipulation: A Step Beyond Programming Languages. *Computer 16*, 8 (1983), 57–69. doi:10.1109/MC.1983.1654471. 4, 7

[SK10] Strumbelj E., Kononenko I.: An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research 11* (2010), 1–18. URL: http://doi.acm.org/10.1145/1756006.1756007, doi:10.1145/1756006.1756007. 8, 9

[Sku02] Skupin A.: A Cartographic Approach to Visualizing Conference Abstracts. *IEEE Computer Graphics and Applications 22* (2002), 50–58. doi:10.1109/38.974518. 12, 16

[SM99] Shipman F., Marshall C.: Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems. *Comput. Supported Coop. Work 8* (1999), 333–352. 4. doi:10.1023/A:1008716330212. 12

[SMNR16] Sun M., Mi P., North C., Ramakrishnan N.: Biset: Semantic edge bundling with biclusters for sensemaking. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 310–319. 16

[SNR14] Sun M., North C., Ramakrishnan N.: A five-level design framework for bicluster visualizations. *IEEE transactions on visualization and computer graphics 20*, 12 (2014), 1713–1722. 16

[SPG14] Stolper C. D., Perer A., Gotz D.: Progressive visual analytics: User-driven visual exploration of in-progress analytics. *Visualization and Computer Graphics, IEEE Transactions on 20*, 12 (2014), 1653–1662. 20

[SS02] Seo J., Shneiderman B.: Interactively exploring hierarchical clustering results. *IEEE Computer 35*, 7 (2002), 80–86. 8

[SS11] Satinover J., Sornette D.: Taming manias: On the origins, inevitability, prediction and regulation of bubbles and crashes, chapter of the book "governance and control of financial systems: A resilience engineering perspective,". *published by Ashgate Publishing Group in their Resilience Engineering Perspectives series* (2011). 19

[SSS*14] Sacha D., Stoffel A., Stoffel F., Kwon B. C., Ellis G., Keim D. A.: Knowledge generation model for visual analytics. *Visualization and Computer Graphics, IEEE Transactions on 20*, 12 (2014), 1604–1613. 3, 4

[SWB97] Sarter N. B., Woods D. D., Billings C. E.: Automation surprises. *Handbook of human factors and ergonomics 2* (1997), 1926–1943. 19

[SZS*16] Sacha D., Zhang L., Sedlmair M., Lee J. A., Peltonen J., Weiskopf D., North S. C., Keim D. A.: Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE Trans. on Visualization and Computer Graphics* (2016). 6

[TC05] Thomas J. J., Cook K. A.: *Illuminating the path: The research and development agenda for visual analytics*. IEEE Computer Society Press, 2005. 1

[TFH11] Turkay C., Filzmoser P., Hauser H.: Brushing dimensions-a dual visual analysis model for high-dimensional data. *Visualization and Computer Graphics, IEEE Transactions on 17*, 12 (2011), 2591–2599. 8

[TJHH14] Turkay C., Jeanquartier F., Holzinger A., Hauser H.: On computationally-enhanced visual analysis of heterogeneous data and its application in biomedical informatics. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*. Springer, 2014, pp. 117–140. 6

[TKBH17] Turkay C., Kaya E., Balcisoy S., Hauser H.: Designing progressive and interactive analytics processes for high-dimensional data analysis. *IEEE Transactions on Visualization & Computer Graphics 23*, 1 (2017), 131–140. 20

[TLLH12] Turkay C., Lundervold A., Lundervold A., Hauser H.: Representative factor generation for the interactive visual analysis of high-dimensional data. *Visualization and Computer Graphics, IEEE Transactions on 18*, 12 (2012), 2621–2630. 8, 10

[TLS*14] Turkay C., Lex A., Streit M., Pfister H., Hauser H.: Characterizing cancer subtypes using dual analysis in caleydo stratomex. *IEEE Computer Graphics and Applications 34*, 2 (2014), 38–47. doi:http://doi.ieeecomputersociety.org/10.1109/MCG.2014.1. 8, 9, 10

[TPRH11a] Turkay C., Parulek J., Reuter N., Hauser H.: Integrating cluster formation and cluster evaluation in interactive visual analysis. In *Proceedings of the 27th Spring Conference on Computer Graphics* (2011), ACM, pp. 77–86. 8, 9

[TPRH11b] Turkay C., Parulek J., Reuter N., Hauser H.: Interactive visual analysis of temporal cluster structures. *Computer Graphics Forum 30*, 3 (2011), 711–720. URL: http://dx.doi.org/10.1111/j.1467-8659.2011.01920.x. 8, 9

[Van05] Van Wijk J. J.: The value of visualization. In *16th IEEE Visualization 2005 (VIS 2005)* (2005), IEEE Computer Society, p. 11. doi:10.1109/VIS.2005.102. 4, 16

[vdEvW11] van den Elzen S., van Wijk J. J.: Baobabview: Interactive construction and analysis of decision trees. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (2011), IEEE, pp. 151–160. 8, 9

[vdMW12] van der Maaten L., Weinberger K.: Stochastic triplet embedding. In *2012 IEEE International Workshop on Machine Learning for Signal Processing* (Sept 2012), pp. 1–6. doi:10.1109/MLSP.2012.6349720. 6

[Ves99] Vesanto J.: SOM-based data visualization methods. *Intelligent Data Analysis 3*, 2 (1999), 111–126. doi:10.3233/IDA-1999-3203. 16

[VL13] Verleysen M., Lee J. A.: Nonlinear Dimensionality Reduction for Visualization. In *Neural Information Processing* (2013), Springer, pp. 617–622. 16

[Wei95] Weick K. E.: *Sensemaking in organizations*, vol. 3. Sage, 1995. 2

[WH00] Wirth R., Hipp J.: Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (2000), Citeseer, pp. 29–39. 22

[WKKB15] Wilber M. J., Kwak I. S., Kriegman D., Belongie S.: Learning concept embeddings with combined human-machine expertise. In *2015 IEEE International Conference on Computer Vision (ICCV)* (Dec 2015), pp. 981–989. doi:10.1109/ICCV.2015.118. 6

[WLS*10] Wei F., Liu S., Song Y., Pan S., Zhou M. X., Qian W., Shi L., Tan L., Zhang Q.: Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (2010), ACM, pp. 153–162. 13

[WM04] Williams M., Munzner T.: Steerable, progressive multidimensional scaling. In *Proceedings of the IEEE Symposium on Information Visualization* (Washington, DC, USA, 2004), IEEE Computer Society, pp. 57–64. 8

[Won14] Wong B.: How analysts think (?): Early observations. In *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint* (2014), IEEE, pp. 296–299. 19

[WSO05] Weick K. E., Sutcliffe K. M., Obstfeld D.: Organizing and the process of sensemaking. *Organization science 16*, 4 (2005), 409–421. 2

[WSP*06] Wright W., Schroh D., Proulx P., Skaburskis A., Cort B.: The Sandbox for analysis: concepts and methods. ACM, pp. 801–810. doi:10.1145/1124772.1124890. 13

[WTP*99] Wise J. A., Thomas J. J., Pennock K., Lantrip D., Pottier M., Schur A., Crow V.: Visualizing the non-visual: spatial analysis and interaction with information for text documents. Morgan Kaufmann Publishers Inc., pp. 442–450. 8, 12, 13

[WV12] Wong B. W., Varga M.: Black holes, keyholes and brown worms: Challenges in sense making. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (2012), vol. 56, SAGE Publications, pp. 287–291. 19

[YaKSJ07] Yi J. S., ah Kang Y., Stasko J. T., Jacko J. A.: Toward a deeper understanding of the role of interaction in information visualization. *Visualization and Computer Graphics, IEEE Transactions on 13*, 6 (2007), 1224–1231. 7

[YKJ16] Yeon H., Kim S., Jang Y.: Predictive visual analytics of event evolution for user-created context. *Journal of Visualization* (2016), 1–16. 8, 15

[YNM*13] Younesy H., Nielsen C. B., Möller T., Alder O., Cullum R., Lorincz M. C., Karimi M. M., Jones S. J.: An interactive analysis and exploration tool for epigenomic data. In *Computer Graphics Forum* (2013), vol. 32, Wiley Online Library, pp. 91–100. 8, 15

[ZC07] Zhu W., Chen C.: Storylines: Visual exploration and analysis in latent semantic spaces. *Computers & Graphics 31*, 3 (2007), 338–349. 13, 14

[ZJGK10] Ziegler H., Jenny M., Gruse T., Keim D. A.: Visual market sector analysis for financial time series data. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on* (2010), IEEE, pp. 83–90. 14, 15

[ZNK08] Ziegler H., Nietzschmann T., Keim D. A.: Visual analytics on the financial market: Pixel-based analysis and comparison of long-term investments. In *Information Visualisation, 2008. IV'08. 12th International Conference* (2008), IEEE, pp. 287–295. 14