Volume 00 (1981), Number 0 pp. 1–26 COMPUTER GRAPHICS *forum*

Viewing Visual Analytics as Model Building

N. Andrienko[1,2], T. Lammarsch[3], G. Andrienko[1,2], G. Fuchs[1], D. Keim[4], S. Miksch[5], and A. Rind[6]

[1]Fraunhofer Institute IAIS, Germany
[2]City University London, UK
[3]Independent Researcher, Austria
[4]University of Konstanz, Germany
[5]Vienna University of Technology, Austria
[6]St. Poelten University of Applied Sciences, Austria

Abstract

To complement the currently existing definitions and conceptual frameworks of visual analytics, which focus mainly on activities performed by analysts and types of techniques they use, we attempt to define the expected results of these activities. We argue that the main goal of doing visual analytics is to build a mental and/or formal model of a certain piece of reality reflected in data. The purpose of the model may be to understand, to forecast, or to control this piece of reality. Based on this model building perspective, we propose a detailed conceptual framework in which the visual analytics process is considered as a goal-oriented workflow producing a model as a result. We demonstrate how this framework can be used for performing an analytical survey of the visual analytics research field and identifying the directions and areas where further research is needed.

Categories and Subject Descriptors (according to ACM CCS): [Human-centered computing → Visual analytics]: Visualization application domains—Visual analytics

1. Introduction

The definition of visual analytics as "the science of analytical reasoning facilitated by interactive visual interfaces" [TC05, p. 4] emphasizes a certain kind of activity (analytical reasoning) and a certain technology (interactive visual interfaces) supporting this activity. The goal of the visual analytics activity is to gain information, insights, and assessments from complex data. Keim et al. [KAF*08, KKEM10] proposed a graphical representation of visual analytics as an iterative process in which knowledge is derived from data by combining visual data exploration with computational processing (Fig. 1, left). A later elaboration of this scheme [SSS*14] focuses on the human cognitive activities through which knowledge is generated. As a complement to this, we focus on the *final product* of visual analytics activities, i.e., on the knowledge that is generated.

The term 'knowledge' as such is very general and does not clearly define the expected product of the visual analytics process. Sacha et al. [SSS*14] define knowledge generated by visual analytics as a trustworthy insight, i.e., an insight sufficiently supported by evidence. An insight, in turn, is defined as an interpreted finding, where a finding is an interesting observation. It can be noted that this chain of definitions focuses on the process rather than the contents. Each definition refers to certain activities of the analyst: observing, interpreting, collecting evidence, and judging the trustworthiness.

This process-oriented definition of knowledge is not instrumental for characterizing visual analytics as a purposeful activity directed to achieving a certain previously stated goal. If we say that the final goal is to gain a trustworthy insight, it would mean that any trustworthy insight is valuable. In the process of data analysis, analysts can make a variety of observations that can be interpreted and supported by evidence. If the goal were just to gain any kind of trustworthy insight, analysts would pay equal attention to all observations, which is usually not the case. Sacha et al. [SSS*14] use the expression 'interesting observation', which means that analysts somehow evaluate observations and judge them as interesting or uninteresting. Hence, an analyst has a specific concept of what needs to be achieved by the analysis, i.e., a goal. An observation is judged as interesting if it is relevant to the goal of the analysis. Consequently, a result of visual analytics activities is not just any kind of trustworthy

This is the accepted version of an article published as:
Andrienko, N., Lammarsch, T., Andrienko, G., Fuchs, G., Keim, D. A., Miksch, S., & Rind, A. (2018). Viewing Visual Analytics as Model Building. Computer Graphics Forum, 37(6), 275–299. doi:10.1111/cgf.13324

The definitive version is available at http://onlinelibrary.wiley.com/doi/10.1111/cgf.13324/abstract.

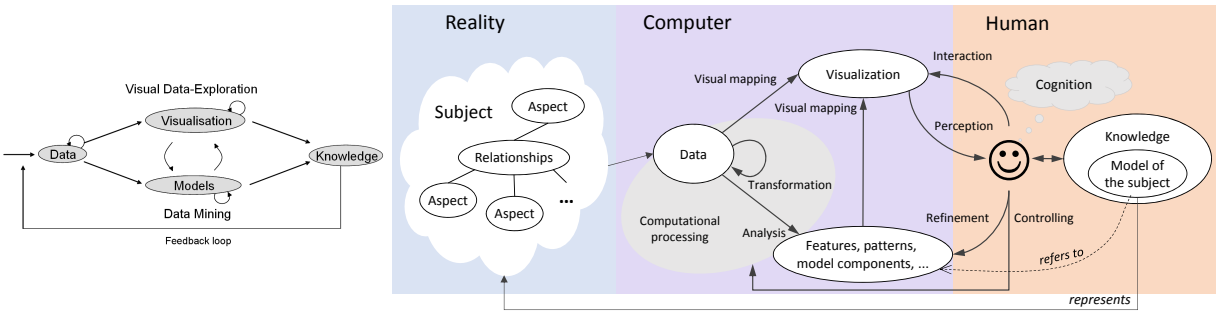

Figure 1: *We build on the commonly adopted representation of the visual analytics activities [KAF*08] shown on the left. To support our reasoning, we have extended it with additional details. The extended scheme, shown on the right, clarifies that data*

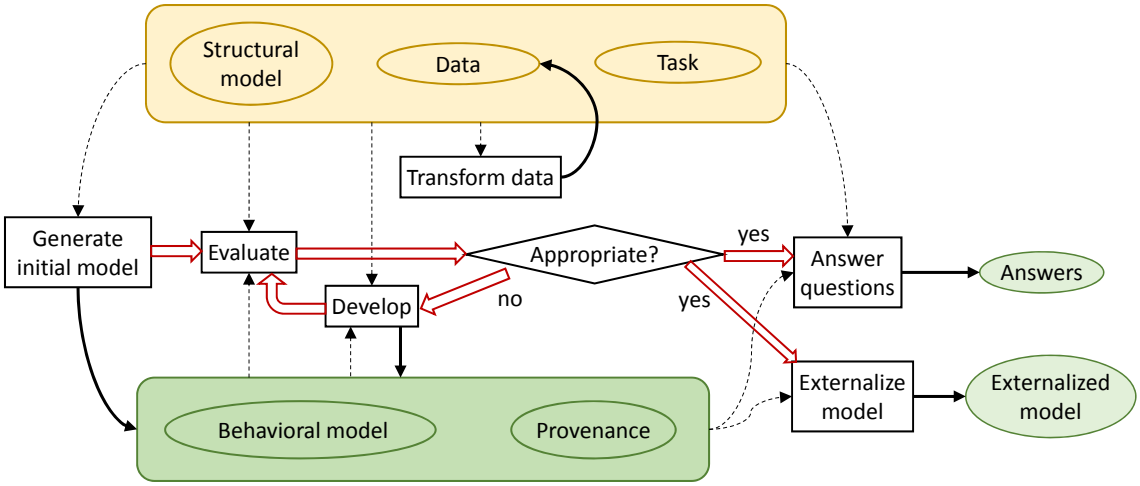

Figure 2: *Our proposed representation of the visual analytics workflow (presented in detail in section 5). The ovals symbolize data, information, and knowledge. At the top is what is given initially, and at the bottom are the primary results of the analysis: a behavioral model of the subject and the provenance of this model. At the right end are secondary results, which may be the behavioral model represented in external media and/or answers to questions concerning the subject. The rectangles stand for activities. The red block arrows show the sequence of the activities. The black arrows represent information flows. The dashed lines show the use of data, information, and knowledge, and the solid lines symbolize generation of knowledge and information.*

insight but a knowledge product satisfying the analysis goal: "Analytical discourse should support the goal of creating a product that articulates a defensible judgment in problems of assessment, forecasting, and planning" [TC05, p. 39].

To give a more specific definition to the product that needs to be created, we propose to see it as a *model* of some piece of reality (real world). We use the ambiguous term 'model' in the sense of "a schematic description or representation of something, especially a system or phenomenon, that accounts for its properties and is used to study its characteristics" [Edi11]. The model is derived from data, which are recorded observations and measurements of a part of the reality. According to the types of visual analytics tasks 'assess', 'forecast', and 'develop options' [TC05, p. 35], analysis may aim at obtaining a *descriptive*, *predictive*, or *decision supporting* model (Fig. 3). A descriptive model describes and explains relationships between aspects of reality. A predictive model describes reality beyond the part reflected in available data. A decision supporting model defines possible actions that can bring the reality to a desired state and assesses the effectiveness and implications of these actions.

According to this view of the visual analytics product, we elaborated the scheme coming from Keim et al. [KAF*08, KKEM10] as is shown in Fig. 1, right. There are

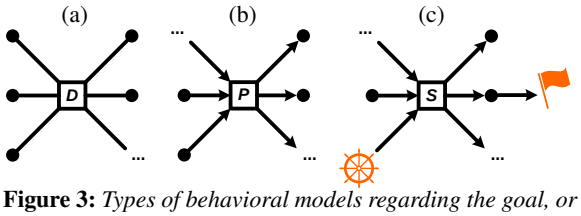

Figure 3: *Types of behavioral models regarding the goal, or task. The dots represent aspects of a subject, and the rectangles linked to the dots stand for combinations of relationships between the aspects. (a) A descriptive model (D), corresponding to the task 'assess', describes the relationships that link aspects of a subject. (b) In a predictive model (P), corresponding to the task 'forecast', aspects are divided into inputs and outputs. The model is capable to tell what outputs will occur for given inputs. (c) A decision supporting model (S), corresponding to the task 'develop options', includes actions (represented by the steering wheel) that can change some aspects. The model is capable to tell what actions can make desired changes. The flag represents a desired state that needs to be achieved.*

two additions to the previous structure. First, the modified scheme shows that the data under analysis reflect some *subject*, which is a part of the reality. Second, it shows that the knowledge that is expected to come out of the analysis process is a *model of the subject*. For simplicity, we do not include a detailed representation of the human cognitive activities [SSS*14] but refer to them through the node 'Cognition'. Compared to the original scheme, we have changed the label of the node initially called 'Models', which was referring to computer-generated models and other computational artifacts. Since we use the term 'model' in a different sense (referring primarily to a model built in the mind of the human analyst), we label computer-generated artifacts as 'Features, patterns, model components, ...'

Based on this scheme, we propose a representation of the visual analytics process as a goal-oriented workflow, shown in Fig. 2, where the primary goal is to create a so-called *behavioral model* of a subject, which can be used for getting answers to questions about the subject. The workflow will be introduced in detail later on. Its main results are represented by a green box at the bottom of Fig. 2. Importantly, the results include not only the behavioral model but also the provenance, i.e., some representation of the way in which the model was obtained. The provenance ensures the traceability and reproducibility of the model derivation process and allows checking if everything was made correctly. The behavioral model and the provenance are, ideally, built in parallel; the workflow steps 'Generate initial model' and 'Develop' are responsible for that.

The general analytical workflow presented in Fig. 2 suggests a perspective from which the research in the visual analytics field can be systematically viewed and analyzed. Existing methods and tools can be characterized and classified based on how they support the model building activities: transform data, generate initial model, evaluate, and develop, along with collecting and representing provenance information. Considering the visual analytics field from this perspective can not only serve as a basis for surveying the field's achievements but also lead to identifying useful directions for further research, as we demonstrate in this paper. Our contribution can be stated as follows:

- We introduce a conceptual framework for considering the visual analytics process as a model building workflow.
- On this basis, we define general requirements to methods and tools needed for supporting the analytical process.
- In the light of these requirements, we survey the research area of visual analytics and identify the existing approaches to supporting different components of the model building workflow.
- We identify the areas of the visual analytics science where further research is needed and the possible directions for advancing the science.

In presenting our argument, we shall refer to a running example, which is introduces in Section 2. In Section 3, we discuss the specifics of visual analytics with regard to other research disciplines concerned with data analysis, namely, information visualization, knowledge discovery in databases, statistics, and machine leaning. Next, we discuss related frameworks and taxonomies in Section 4, introduce our framework in Section 5, and, on this basis, review the state of the art in Section 6. We discuss various aspects of the framework in Section 7 and conclude in Section 8.

2. Running example: VAST Challenge 2011

To illustrate our concept, we use a running example based on the IEEE VAST Challenge 2011, Mini Challenge 1 [GWLN11], requiring analysis of the circumstances of an epidemic outbreak in the fictive city of Vastopolis. The questions mostly refer to the analysis task 'assess', requiring challenge takers to identify the origin of the outbreak and the affected area and to explain the mechanism of infection transmission. There are only some elements of 'forecast' (determine whether the outbreak is contained, i.e., forecast whether it will spread further) and 'develop options' (determine whether it is necessary to deploy treatment resources outside the affected area). We shall extend the original tasks to forecasting how the situation will evolve further and finding suitable actions to fight the epidemic.

Using this example, we can briefly clarify the terms 'behavioral model' and 'structural model' appearing in Fig. 2 and more formally defined in section 5.2.2. A *structural model* defines the structure of the subject, i.e., generic relationships between concepts corresponding to aspects, or components, of the subject. Figure 4 schematically presents a structural model of the analysis subject of the VAST

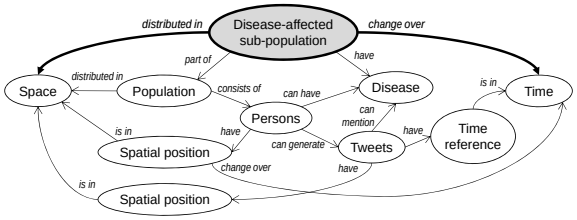

Figure 4: *An example of a structural model describing the analysis subject of the VAST Challenge 2011, Mini Challenge 1 [GWLN11]. The nodes of the scheme represent aspects of the subject; general relationships between them are represented by labeled arrows. The highlighted elements correspond to the stated task of the Challenge: build a behavioral model of the emergence and evolution of the disease-affected sub-population and its spatial distribution.*

Challenge 2011. A *behavioral model* describes a specific instantiation of the concepts and their structural relationships, which may be seen as their *behavior*. Depending on the analysis task, a behavioral model may focus on a subset of aspects and relationships. Thus, the VAST Challenge task requires describing the emergence and evolution of the disease-affected sub-population and its distribution in space. The corresponding parts of the structural model are highlighted in Fig. 4. Please note that the structural model only states that the disease-affected sub-population is distributed in space and changes over time. A behavioral model is expected to describe *how* specifically the affected population is distributed in space and *how* it and its distribution change over time. For example, a behavioral model may say that the disease occurrences were initially concentrated in the city center and then spread eastwards while another concentration appeared in the southwest along the river flow (Fig. 5).

3. Specifics of visual analytics

It can be noted that the workflow in Fig. 2 does not include anything strictly specific to visual analytics. Indeed, it is a generic analytical workflow that visual analytics as a research discipline strives to facilitate. Similar workflows are referred to in defining the scope and research foci of information visualization and knowledge discovery in databases.

In defining information visualization, Card et al. [CMS99] refer to the process called "knowledge crystallization", which involves getting insights about data relative to some task. This usually requires finding some representation (*schema*) for the data that is efficient for the task. The term *schema* corresponds to what we call *behavioral model*, and it can also be matched to the concept of *internal model* used by Spence [Spe01]. According to Spence [Spe07], the underlying philosophy of information visualization is encapsulated in the statement by H.Simon: "solving a problem simply means representing it so as to make the solution transparent" [Sim96]. This statement can be applied to both an internal representation (model, schema) in the human mind and an external representation, e.g., on a computer screen.

Knowledge crystallization [CMS99] is the process of formation of a good representation (schema) for solving a problem. It includes searching for an initial schema, instantiating the schema with data, assessing the residue (data that do not fit the schema), improving the schema to reduce the residue, and searching for a possibly simpler representation. This process corresponds very well to the workflow in Fig. 2.

The goal of information visualization is to facilitate the knowledge crystallization process by visual representations of data. Information visualization is defined as "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition" [CMS99] (p.7). As a field of research, information visualization is mostly concerned with mapping information to graphical representations [CMS99, WGK10]. While the discussion of the knowledge crystallization process defines a possible context in which such representations can be used, information visualization (unlike visual analytics) does not aim at comprehensive support of the entire process but has a much narrower focus. Ward et al. [WGK10] refer to another possible context in which interactive visualizations can be used, namely, the knowledge discovery pipeline [FPSS96b].

Knowledge discovery in databases (KDD) is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [FPSS96a, FPSS96b]. A *pattern* is an expression in some language describing a subset of the data or a model applicable to that subset. Extracting a pattern includes fitting a model to data, finding structure from data, or in general any high-level description of a set of data. The term *process* implies that KDD is comprised of many steps, which can be repeated in multiple iterations. Fayyad et al. [FPSS96b] introduce a schematic representation of the KDD process, known as "the knowledge discovery pipeline" [WGK10]. It includes data selection, preprocessing, subsampling, and transformations, followed by application of data mining methods (algorithms) to extract patterns. Data mining is thus a step in the KDD process. The next step is interpretation and evaluation of the products of data mining to determine which patterns may be considered as new "knowledge". This step can involve visualization of the extracted patterns/models or visualization of the data given the extracted models.

Although Fayyad et al. [FPSS96b] emphasize that the KDD process is performed by a human analyst, who selects appropriate techniques, steers their work, and evaluates the results, KDD as a research discipline focuses on development of computational techniques for data processing and analysis, giving primary attention to data mining. Supporting the activities of the human analyst is out of the scope of KDD. Again, as in information visualization, consideration

of the analytical process defines a broader context in which results of KDD research can be used.

Proposals to consider the overall human-driven analysis process have also been put forward in statistics [Han94] and in machine learning [BS97]. Still, the research in these disciplines concentrates on developing analytical techniques rather than supporting the process as a whole.

Unlike the other disciplines, visual analytics does not focus on particular type of techniques (visualization, data mining, statistical analysis, or machine learning) that can be used within the analytical process. The fundamental goal of visual analytics is *to support the whole analytical process (discourse)* [TC05, p. 40]. To achieve this, it takes an integrative approach leveraging the achievements of the other analysis-centered disciplines for combining the best capabilities of computers and humans [KKEM10].

The schematic representation of the analytical process in Fig. 2 thus can provide a basis for surveying the visual analytics research field in terms of supporting the process.

4. Related work

In this section, we discuss related works proposing conceptual frameworks, taxonomies, or formalizations for visual analytics activities. The relevant works can be organized in categories according to their main focus: the visual analytics process, visual analytics methods, and analysis tasks relevant to visual analytics. After considering these three categories, we overview the works that deal with models in visual analytics. Finally, we briefly state how our framework relates to the previous research, while a more detailed comparison is presented in Section 7.4.

4.1. Defining the visual analytics process

As mentioned earlier, Thomas and Cook introduce three types of analytical tasks: 'assess', 'forecast', and 'develop options' [TC05, p. 35]. 'Assess' means to understand the current world and explain the past, 'forecast' means estimate future capabilities, threats, vulnerabilities, and opportunities, and 'develop options' means establish different optional reactions to potential events and assess their effectiveness and implications. There are no clear definitions of the expected results of these task types. It is stated, without further elaboration, that the product of the task 'assess' is an assessment. For the task 'develop options', a few examples are given in application to homeland security problems.

Further on, various artifacts of analytical reasoning are defined: assumptions, evidence, patterns, arguments, hypotheses, scenarios, etc. The process of analytical reasoning is described as a sense-making loop consisting of activities 'gather information', 're-represent', 'develop insight', and 'produce results' that allow "a defensible judgment in problems of assessment, forecasting, and planning" [TC05, p. 39]. A specific sense-making process of intelligence analysts [PC05] is discussed in more detail.

Keim et al. [KKEM10] extend the scope of the visual analytics science to a much larger area of applications than intelligence analysis. The earlier proposed flow diagram of the visual analytics process [KAF*08] is reused to stress tight coupling between automated analysis methods and interactive visual representations. Lammarsch et al. [LAB*11] present a variant of the visual analytics workflow diagram in which prior domain knowledge is distinguished from insights gained through the analysis. A special block is dedicated to hypotheses, and it is said that validated hypotheses become models. Models are defined as "representations of a system of entities, phenomena, or processes <...> that are validated by comparison to existing data" [LAB*11, p. 10]. This treatment closely corresponds to what we mean by models in our paper; however, Lammarsch et al. [LAB*11] consider models as a means of gaining insights rather than results of analysis.

The aforementioned framework by Sacha et al. [SSS*14] describes the process of knowledge generation by a human analyst using visual analytics tools. The process consists of three loops: exploration, in which observations are made and interpreted, verification, in which hypotheses are formulated and supporting or contradicting evidence is sought, and knowledge generation, in which trustworthiness of insights is assessed. Rind et al. [RAW*16] extended this framework to explicitly focus on users' objectives and their plans to reach these objectives. Another extension is proposed by Ribarsky and Fisher [RF16]. By involving principles from cognitive science, they redefine the human-machine interaction loop as a ring where the ordering of the tasks or actions is not prescribed. The proposed human-computer model is used for deriving interface design principles with the main goal to keep the human "in the cognitive zone" [GRF09, p. 4]. The principles include the support of direct manipulation, search by example, and knowledge externalization through annotation at any point in time.

The models focusing on the human cognitive activities [GRF09, SSS*14, RF16] use the ideas from the van Wijk's model of the visualization process [VW05], in which user's perception P of an image increases the user's knowledge K. The knowledge gain is a function of the image, the user's prior knowledge, and particular properties of the user's perception and cognition. The current knowledge drives the process of interactive exploration E. To assess the value of a visualization method, van Wijk proposes to estimate the costs of the method development and the user's efforts and compare these to the value of the knowledge gained. Green et al. [GRF09] consider P, K, and E as interrelated cognitive processes; thus, P may drive E, and E may feed to K, where K includes cognitive activities that create knowledge, i.e., reasoning and problem solving. Van Wijk's model is also the basis for the knowledge-assisted visual analytics model by

Federico, Wagner et al. [FWR*17], which distinguishes between tacit knowledge within the human and explicit knowledge within the tool and elaborates the processes of knowledge generation, conversion (i.e., internalization and externalization), and exploitation.

Based on the activity theory, which originated from psychology and was adopted in HCI, Edge et al. [EHRLW18] define visual analytics activity as interplay of six elements: personas (types of people using the tools of the activity), products (derive insights, develop options, make arguments, present assessments, manage situations), capabilities (types of task supported by tools), contexts (co-located teams, distributed teams, distributed communities, synchronicity, mobility), rules (types of constraint on the performance of activity, such as relevance, confidence, provenance, access rights, and time pressure), and roles (producers, consumers, responders, decision makers, policy makers). This model is used as a basis for defining guidelines for designers of visual analytics systems. The authors highlight a set of six qualities that need to be supported through system design: portable analysis, presentable analysis, perspectival analysis, proxemic analysis, provisional analysis, and polymorphic analysis. They especially advocate support of workspaces that can be labeled, annotated (with a possibility to represent the work yet to be done), forked, and linked in ways that retail the provenance.

Chen and Jänicke [CJ10] focus on the communication of information contained in data to the user through visualization. The measures defined in information theory, in particular, entropy and mutual information, can be used for quantifying information contents and uncertainty reduction in visualization. Chen and Jänicke note that a visualization system involves three types of information sources: input data, interaction, and prior knowledge. While it is relatively easy to apply the information-theoretic concepts to input data, they may not be readily applicable to the two other information sources, requiring adaptation and extension. The proposed framework deals only with information contained in input data. It is stated that, in principle, the visualization process does not generate more information than what is in the original data. Hence, the framework does not encompass generation of new knowledge by a human interacting with visualization and applying prior knowledge [VW05].

In our work, we elaborate the definitions and models by Thomas and Cook [TC05] and Keim et al. [KAF*08, KKEM10] by proposing a more specific concept of the expected result of the visual analytics process (see 4.5).

4.2. Defining and classifying visual analytics methods

Both books defining visual analytics as a research field [TC05, KKEM10] give much attention to visual analytics methods combining computational processing with interactive visualizations; however, the concept of combining the two types of approaches is much older [SBM92]. Bertini and Lalanne [BL09] describe processes for visual and automated methods, analyze how close they are intertwined, and argue for a tighter integration. Stolper et al. [SPG14] describe a progressive visual analytics workflow, in which the analyst views and interprets partial results of computational processing and, on this basis, focuses the algorithm on subspaces of interest. Mühlbacher et al. [MPG*14] define in a general way the possible strategies of combining computational processing with user involvement.

Several taxonomies of visualization, interaction, and analysis methods exist in the literature. In the task by data type taxonomy [Shn96], methods are organized according to the types of data they can be applied to. Another proposed categorization of methods is into visualization, display modification, data transformation, and computational analysis [AA06]. There are multiple works where a description of the analytical process serves as a basis for systemizing methods [CMS99, Chi00, CR98]. Some taxonomies are analyzed and compared by de Oliveira and Levkowitz [dOL03]. Roth [Rot13] presents a taxonomy of interaction techniques organized according to "three broad user goals motivating use of the visualization": procure, predict, and prescribe. These goals can be considered as corresponding to the task 'assess', 'forecast', and 'develop options' [TC05].

In Section 6, we discuss the existing visual analytics methods from the perspective of supporting the steps and components of the analysis process.

4.3. Characterizing and systematizing analysis tasks

Miksch and Aigner [MA14] link tasks, users, and data of visual analytics methods (which are located in the middle of a triangle) by saying that for users to solve tasks, the methods need to be appropriate, for users to deal with data, the methods need to be effective, and for showing the data according to the tasks, the methods need to be expressive. Amar and Stasko [AS05] describe similar challenges: the worldview gap exists between what is shown and what is actually needed by users; the rationale gap exists between perceiving a relationship and actually being able to explain it and its usefulness.

Aigner et al. [AMST11] refer to the classes of tasks adopted in data mining, which include classification, clustering, search and retrieval, and pattern discovery. A typology of tasks in exploratory data analysis [AA06] defines tasks on the basis of data structure. Data components are categorized into independent and dependent variables, called references and attributes, respectively. Data represent a function that matches references to attributes. The general aim of data analysis is studying the behavior of this function. There are four major classes of analysis tasks:

- Behavior characterization: describe the behavior of one or more attributes.

- Pattern search: locate a particular behavior, i.e., find subsets of references where attributes have this behavior.
- Behavior comparison: identify similarities and differences between two or more behaviors.
- Relation seeking: find subsets of references for which a particular relation ('same', 'different', 'opposite', etc.) exists between the behaviors of two or more attributes.

The taxonomy involves the notion of *pattern*, which is defined as a construct representing essential features of a behavior in a general way. 'To characterize a behavior' means to represent it by one or several patterns; the other classes of tasks are also related to the notion of pattern. The definition of a pattern is similar to that adopted in data mining, where a pattern is defined as an expression in some language describing a subset of facts without enumerating all these facts [FPSS96a]. The definition proposed for exploratory data analysis [AA06] has a broader scope, also including representations in the human analyst's mind. In both definitions, 'pattern' is a representation (constructed by a human or a computer) of something that objectively exists in the studied behavior, i.e., "essential features of a behavior" [AA06]. Accordingly, for a task to be fulfilled by a human with the help of visualization, the visualization must convey these essential features for enabling the human to construct appropriate patterns.

A multitude of task typologies have been introduced in the research areas of visualization, human-computer interaction, and information retrieval. The TaskCube concept by Rind et al. [RAW*16] surveys such task typologies according to the dimensions 'perspective', 'abstraction', and 'composition', and discusses how different notions of 'task' fit into various design and evaluation scenarios. Brehmer and Munzner [BM13] cite and discuss about 30 existing typologies in the context of presenting their own typology that unites the previously existing low-level and high-level classifications. The typology organizes the tasks according to three questions: why the task is performed, how the task is performed, and what the task inputs and outputs are. This is similar to the questions considered by Aigner et al. [AMST11]: what is presented, why is it presented, and how is it presented? Similarly, Schulz et al. present a taxonomy that deals with questions 'why', 'what', 'where', 'when', and 'how' [SNHS13]. With 'what' they mean patterns in the data, with 'how' the methods. The existing task typologies, including the multi-level taxonomy from Brehmer and Munzner [BM13], either do not consider the overall goals and expected final results of analytical activities or refer to them using very abstract terms, such as 'present', 'discover', 'enjoy', and 'produce'. Thus, the term 'produce' means generation of any artifacts, including derived data, annotations, recorded interactions, or screen shots.

Gotz and Zhou [GZ09] propose a multi-tier framework representing visual analytics activity. At the highest level, there are tasks, i.e., the overall goals of the analysis. At the next level, tasks are decomposed into simpler subtasks. At the third level, subtasks are translated into actions, which are accomplished using elementary events (operations), such as mouse clicks and selections in a menu. The authors focus on the tier of actions, for which they propose a typology similar to the task typologies of other authors. The framework serves as a basis for semi-automated capturing of insight provenance during the analysis process, which is an important aspect of support to model building and externalization.

Our framework mainly focuses on the types of analytical tasks 'assess', 'forecast', and 'develop options' [TC05] and use concepts from the typology of exploratory tasks [AA06]. An important part of our framework relates to the work by Gotz and Zhou [GZ09] on collecting insight provenance.

4.4. Considering models

Tory and Möller [TM04] propose a taxonomy of visualization techniques applied in information visualization and scientific visualization, which are categorized based on the type of data model they use. 'Continuous model' corresponds largely to scientific visualization and 'discrete model' to information visualization. The term 'model' refers to data structure, unlike 'model of reality' in our paper.

Sedlmair et al. [SHB*14] propose a conceptual framework for exploration of the behavior of simulation models. It includes a general data flow model, four navigation strategies, and six typical tasks pertaining to analysis of simulation models. From the perspective of our framework, a simulation model is a piece of reality that is studied, and the overall analysis goal is to understand the relationships between the parameter settings and corresponding outputs, i.e., to build a model of these relationships. Nevertheless, some concepts discussed by Sedlmair et al., such as uncertainty and sensitivity, are not specific to only simulation models but are relevant to various kinds of models.

There are works proposing models of human cognitive processes involved in the use of visual displays and interaction techniques [GRF09, RFP09, RF16]. Liu and Stasko [LS10] discuss the construction of mental models. Visualization is considered as a means that helps users to internalize, process, and augment mental models. They also talk about externalization of models, but do not go deeper into building of models which are not mental.

In our framework, we consider mental models as essential results of visual analytics activities but do not focus on the cognitive processes involved in constructing such models.

4.5. Positioning of our work

Our paper elaborates and complements the existing frameworks focusing on the visual analytics process [TC05, KAF*08, SSS*14]. We begin with specifying the expected result of the process and then represent the process as a

directed workflow leading to this result. We also elaborate the knowledge crystallization model [CMS99] considered in information visualization. It views the analytical process as iterative construction, evaluation, and improvement of a schema representing a problem (Section 3). The concept of schema is equivalent to our concept of model of the subject. We define in more detail the starting point, goal, steps, components, and artifacts of the analytical process (Section 5.2). We deem this specification useful for visual analytics research and practice, since visual analytics aims at facilitating the whole analytical process in a comprehensive manner.

In this paper, we use our framework as a basis for a systematic survey of the research that has been done in visual analytics and as an instrument helping us to identify the directions and areas that need further research. The possible practical use of the framework is discussed in Section 7.1.2.

5. Our conceptual framework

5.1. Summary

The underlying idea is that the overall goal of analysis is to build an appropriate representation, called 'model', of some piece of reality, called 'subject' (of the analysis); see Fig. 1, right. Our use of the term 'subject' corresponds to the term 'object of study' used by Tory and Möller [TM04]. 'Appropriate' means congruous to the reality and fitting to the purpose. The possible purposes correspond to the tasks 'assess', 'forecast', and 'develop options' [TC05]; accordingly, a model can be descriptive, predictive, or decision supporting (Fig. 3).

The subject is considered as a system composed of aspects linked by relationships. A model needs to represent (some of) these relationships. The subject usually does not allow direct perception and analysis. Models have to be built by analyzing available data on the subject, i.e., recorded observations and measurements of its aspects. However, we deem appropriate to emphasize that *the primary interest of the analyst is not the data per se but the reality reflected in the data*.

Data can reflect only a part of the subject, while a model needs to represent not just this part but the subject as a whole. Hence, a model is a generalization from data. It is also a simplification of the subject: it may represent only a subset of the aspects and relationships and may omit details. However, simplification is not the goal but the means of analysis.

A simple model can fully reside in the mind of the analyst, but more complex models may be hard to fully keep in mind. Parts of such models may need to be offloaded to external representations, such as formulas, graphics, and texts, while the human mind keeps an overall frame model containing references to these external representations. Models may include component parts intended for performing calculations in computers. These parts are represented in computer-readable form and reside in computers. Still, the mind of the human analyst contains a frame model comprising a high-level representation of the subject and a sufficient representation of the computer-resident components allowing appropriate use of these components, e.g., for obtaining forecasts or making decisions.

A model of the subject, either fully contained in the analyst's mind or distributed over several media, is the knowledge that is gained through the process of analysis.

Figure 2 schematically represents the analysis process in accord with the model building perspective of visual analytics. It explicitly includes the step 'Evaluate' and shows that an initially created model must be checked for appropriateness and, if not yet appropriate, developed further. The steps from 'Evaluate' to 'Develop' may need to be performed several times. This is an elaboration of the 'Feedback loop' from the original framework we build on [KAF*08]. Unlike the previous schemes [KAF*08, SSS*14], Figure 2 shows that the analysis process eventually terminates, and that the termination condition is that the model is appropriate. It will be discussed later on what model appropriateness means. Another important addition to the previous frameworks is explicit inclusion of the model provenance as a main result, apart from the model itself.

5.2. Basic definitions

Here we present our framework in detail through a set of definitions, in which we use relevant concepts from the areas of entity-relationship modeling [Che76], object-oriented analysis and design [BME*07], and systems sciences [Kli85]. The research in these areas forms a suitable basis for high-level modeling of the analytical process.

5.2.1. Subject

Definition S.1. A *subject* (of analysis, reasoning, etc.) is a piece of reality (real world). A subject can be seen as a system of components and their properties, jointly called *aspects*, that are linked by *relationships*. Any component can, in turn, be a system composed of other aspects linked by relationships.

Comment. The reality objectively exists. It can be observed and/or measured, but it exists independently of the observations and measurements. The entire reality is too complex to describe in full. Analysis focuses on some piece of the reality, called 'subject'.

Examples. The subject in the VAST Challenge is a system composed of the geographic space (specifically, the territory of Vastopolis) with its properties, population of the city, tweets posted by the people, disease with its properties (symptoms, mechanism of spreading, etc.), and weather conditions. Furthermore, almost all these aspects change over

time or happen in time; hence, time is also an aspect of this subject. These aspects are linked into the system by various relationships (Fig. 4): people live on the territory of Vastopolis, they move, i.e., their spatial locations change over time, they post tweets, the tweets have locations in space and times of appearance, some people get the disease, the disease cases have locations in space and times of appearance, and many others. Almost all aspects listed above are systems on their own. Thus, the territory of Vastopolis is composed of the land, river, streets, buildings, and other geographical objects linked by spatial relationships. The population consists of people. The disease outbreak is a system that includes the origin, the symptoms, the set of disease cases and its spatial distribution, the temporal evolution of the spatial distribution of the disease cases, etc.

Definition S.2. Aspects can be categorized into *entities* and *attributes*. Entities exist as separate and distinct things, i.e., they can be separated and distinguished from others. Attributes are characteristics of entities; they do not exist separately from entities.

Examples. The people and the tweets are entities. The health condition is an attribute of people, and the message texts are an attribute of tweets. Locations in space (i.e., on the territory of Vastopolis) are entities; space can be seen as a continuous set the elements of which are distinct locations. Spatial locations have attributes, such as land cover or land use (river, street, building, etc.). Moments in time are also entities. Time is a continuous set composed of linearly ordered moments. Time moments have attributes, such as time of the day (day or night) and day of the week. Some relationships between entities may be treated similarly to attributes. Thus, people, tweets, and disease occurrences are located in space, i.e., linked to certain spatial locations. This relationship to a spatial location (which may change over time) can be seen as an attribute of the people, tweets, and other kinds of entities. Tweets and disease occurrences are linked to certain moments in time when they appeared. This relationship to time moments can also be seen as an attribute of the tweets and disease occurrences.

Definition S.3. Sets and subsets of entities are entities.

Comment. Any set of entities can be considered in its entirety as a separate and distinct entity on its own. Such a composite entity is also an aspect of the subject. Generally, aspects may consist of other (simpler) aspects. Composite entities as wholes may have their attributes differing from attributes of the simpler entities they are composed of.

Examples. All people considered together are the population of Vastopolis. The population as an entity has attributes 'number of people' and 'spatial distribution'. The disease outbreak can be seen as an entity consisting of multiple disease cases. As an entity, it has attributes 'source', 'number of cases', and 'spatial extent', the latter two changing over time.

Definition S.4. *Structural relationships* between aspects are abstract generic relationships by which the aspects as classes are arranged in a subject or in a more complex aspect.

Definition S.5. *Instances of relationships* between aspects (shortly, *relationship instances*) are specific associations and interactions that actually happen.

Comment. Structural relationships are abstractions referring to classes of entities and types of attributes. Relationship instances are specific realizations of these abstractions. A relationship instance may have a limited time of existence.

Examples. 'People move in space' is a structural relationship between people in general, space in general, and time in general. Similar considerations refer to 'people have health condition' (i.e., the attribute 'health condition') and 'people produce tweets'. The relationships shown in Fig. 4 are structural relationships. An instance of the relationship 'people move in space' is a specific person being in a specific location at some time. Such an instance exists for a limited time and then is replaced by another instance, in which the same person is linked to another spatial location. A specific health condition of a specific person at a specific time moment is an instance of the relationship 'people have health condition'. The structural relationship 'people produce tweets' is realized in multiple relationship instances between actually produced tweets and their specific authors.

Definition S.6. The *behavior* of a structural relationship within a subject (shortly, *relationship behavior*) is the realization of this relationship in actual instances.

Comment. The term 'behavior' refers to the relationship instances that actually happened and also to those that can potentially happen. It is a general manner in which a structural relationship is realized in various instances. The term 'behavior' may also refer to a particular subset of relationship instances. There may be variations in the realization of a structural relationship between subsets of instances. This can be considered as a complex behavior consisting of several partial behaviors.

We use the term 'behavior' as an umbrella term embracing several more specific terms used in the literature for denoting particular classes of behaviors: *distribution* (of an attribute over a set of entities, of set of entities over space), *variation* (of an aspect over space or over time), *correlation* (between attributes, between appearances of different entities), *evolution* (of an aspect or another behavior over time), *influence* (of one aspect upon another), and *interaction* (between aspects). This is mainly consistent with the previous usage of the term 'behavior' [AA06], but here we apply this term to aspects of the objectively existing reality while previously it was applied to components of data.

Examples. The behavior of the structural relationship

'people move in space' within the VAST Challenge scenario consists of the spatial positions and movements of all people during the time span of the scenario. In this overall behavior, several partial behaviors may be specially considered, such as the movements of all people before the disease outbreak, the movements during the outbreak, appearance of people in contaminated areas, and the movements of infected people.

The behavior of the structural relationship 'people have health condition' consists of the health conditions of all people of Vastopolis at different times. This includes the distribution of different health conditions over the population and the evolution of this distribution over time. The behavior of the structural relationship 'people produce tweets' consists of all instances of twittering.

5.2.2. Model

Definition M.1. A *model* is any representation of aspects of a subject and relationships between them.

Comment. The form and medium of the representation are not specified. It may be text in any language, either natural or formal, graphics, formulas, computer code, etc. It may exist in human mind, be written or drawn on paper or in a digital document, encoded in internal computer structures, or it may be a physical model made of some material.

A model may represent not all aspects and not all relationships of a subject. For practical reasons, a model needs to be simpler than the subject it represents.

Definition M.2. A *structural model* is a representation of structural relationships between aspects of a subject.

Comment. A structural model often exists in the mind of the analyst as a part of the prior knowledge, or it is given in a problem definition, as in the case of VAST Challenge. Figure 4 gives an example of a structural model represented graphically.

Definition M.3. A *behavioral model* of a subject is a generalized representation of the behavior of one or more structural relationships within the subject. The *goal of analyzing a subject* is to obtain a behavioral model of this subject.

Comment. A behavioral model is a generalized representation in the sense that it does not refer to particular relationship instances but represents the general manner of the realization of the relationships; see Definition S.6 and the following comment.

Definition M.4. A *focus behavior* is a relationship behavior that needs to be represented in a behavioral model, according to the analysis task. A *focus relationship* is a structural relationship whose behavior need to be represented in the model.

Examples. In the VAST Challenge, the goal is to obtain a model representing the behavior of the relationships between the disease outbreak and the population of Vastopolis, space, and time. The model must also specify what aspects and in what ways affect the disease spread, i.e., change the behavior of the disease relationships to the population and space.

Definition M.5. A *descriptive model* is a behavioral model representing focus behaviors in a descriptive, passive manner and used for explanation and understanding.

Definition M.6. A *predictive model* is a behavioral model representing the behavior of relationships between one subset of aspects, called 'inputs', and another subset, called 'outputs', in a functional manner allowing to determine which specific outputs will actually happen for given specific inputs.

Comment. In a descriptive model (Fig. 3a), the directionality of relationships is not prescribed, and all aspects are treated equally. A predictive mode (Fig. 3b) distinguishes between inputs and outputs. Inputs are represented on the left of Fig. 3b and outputs on the right. The relationships are directed from the inputs to the outputs.

Examples. A descriptive model of the disease outbreak in the VAST Challenge scenario may say that the epidemic started on May 18, 2011 in the city center. There were two diseases with differing symptoms. One was conveyed by wind and spread in the eastern direction, and the other was conveyed by the river and spread in the southwestern direction.

A predictive model of the outbreak must be capable of forecasting for the future time. Here, time is input, and outputs are the people that will get infected and the spatial spread of the new disease cases. For a given moment in time, the model is expected to tell how many people will be infected and where in space they will be located.

Definition M.7. An *action* is purposeful modification of some relationship behaviors.

Comment. An action may affect a relationship behavior so that certain (desired) relationship instances will occur or get a higher probability of occurrence while occurrences of other (unwanted) relationship instances are precluded or become less probable. Actions are purposefully performed by some agents, in particular, by people. Actions are a particular kind of aspect of a subject.

Examples. Actions in the VAST Challenge scenario may include setting restrictions to movements of people (affecting the behavior of the relationship 'people move in space'), giving medical treatment to sick persons (affecting the behavior of the relationship 'people have health condition'), cleaning of the affected area, decontamination of the water in the river, isolation of the infected people, evacuation of people from the affected area, informing and instructing people through mass media, acquisition and deployment of additional medical resources, etc.

Definition M.8. A *decision supporting model* is a behavioral

model representing the behavior of relationships between inputs, actions, and outputs in a procedural manner allowing to determine which actions will bring about desired modifications of outputs for given specific inputs.

Comment. In Fig. 3c, actions are symbolically represented by a steering wheel. A decision supporting model is used for choosing suitable actions (possibly, combinations of several actions) by which desirable outputs (represented by a flag in Fig. 3c) can be achieved. However, the model does not prescribe that certain actions must be fulfilled. Instead, it describes what would happen based on given decisions. Moreover, the decision maker, while taking the model result into account, may also apply tacit knowledge, criteria, and preferences that have not been explicitly included in the model. The final decision may thus differ from what the model suggests.

Examples. In the VAST challenge scenario, the desired condition is that the epidemic stops (no further people get infected) and sick people recover. A decision supporting model should allow the analyst to choose suitable actions for achieving this condition depending on the expected evolution of the disease.

5.2.3. Data

Definition D.1. *Data* are recorded observations or measurements describing some relationship instances.

Comment. Data include references to entities and specifications of attributes of the entities, i.e., data describe, in particular, instances of relationships between entities and values of attributes. Data cannot describe all relationship instances that ever occurred; hence, data are always incomplete. Besides, there can be no data describing instances that will occur in the future. Recorded observations or measurements can also be erroneous, i.e., describe relationship instances that did not happen instead of those that actually happened.

Example. For the VAST Challenge, there are the following sets of data: (1) data describing the territory of Vastopolis, given in a form of a map; (2) data describing the tweets by specifying their authors, times of posting, spatial locations, and message texts; (3) data describing the wind direction and speed on different days.

Definition D.2. A relationship behavior is *directly reflected* in data if the data contain records describing instances of this relationship.

Definition D.3. Let $R(A,B,C,\ldots)$ be a structural relationship between aspects A, B, C, $\ldots$ such that its behavior is not directly reflected in data. Let A' be an aspect that may serve (possibly, under some assumptions) as a proxy for one of the aspects (say, A) in R. If the behavior of $R(A',B,C,\ldots)$ is directly reflected in data, the relationship $R(A,B,C,\ldots)$ is said to be *surrogated* in the data.

Comment. When some focus behavior is not directly reflected in available data, it is necessary to acquire additional data describing instances of this relationship. One possibility is to derive the necessary data from available data that reflect the behaviors of other relationships based on the knowledge (model) of the structural relationships between the aspects. When there are no structural relationships allowing such a transformation, it may be possible to derive data surrogating the focus relationship under reasonable assumptions. If this is also not possible, the necessary data need to be obtained from other sources.

Example. Behaviors of the disease in relation to the population, space, and time are not directly reflected in the VAST Challenge data. However, the following structural relationships of the disease to other aspects are known: people may contract the disease, and people may write about their health condition in tweets, i.e., the message texts of tweets may contain indications of the disease. The available data directly reflect the behaviors of the relationships between the tweets and people (the tweet authors) and between the tweets and disease indications (occurrences of particular keywords in the message texts). The known structural relationships of the disease to people and tweets suggest the following transformations of the available data: (1) select the tweets where the message texts contain disease indications; (2) take the authors of the selected tweets as the sub-population affected by the disease; (3) take the locations of the earliest selected tweets of the affected people as the locations of the disease occurrences. The so derived data surrogate instances of the relationship between the disease, people, space, and time under the assumption that infected people post messages containing disease indications as soon as they get infected. These data may be inaccurate or even erroneous if the assumption does not hold. Besides, the data are definitely incomplete: since not all people may tweet about their health condition, not all disease cases are reflected in the data.

5.2.4. Analysis

Definition A.1. *Analysis* is the process of deriving a behavioral model of a subject from data on this subject.

Definition A.2. A model is considered *appropriate* if it meets the following requirements:

- Correctness, i.e., consistency with the data;
- Fitness to the purpose, or task, i.e., capability to describe and explain the subject ('assess'), to determine outputs corresponding to given inputs ('forecast'), or to find suitable actions for achieving a desired state of the subject ('develop options').
- Comprehensiveness, i.e., representation of all focus behaviors;
- Sufficient scope:
  - Sufficient coverage of the available data, i.e., consistency with all instances of the focus relationships that are available in the data;
  - For predictive models, sufficient extension beyond the

available data, i.e., the capability to produce forecasts for all required combinations of inputs;
- Generalization, i.e., representation of all instances by a much smaller number of model components;
- Specificity, i.e., representation of significant distinctions among behavior instances;
- Parsimony, i.e., involvement of a minimal number of components;
- Resource efficiency, i.e., avoiding or reducing the use of excessively costly components (e.g., heavy calculations).

Comment. An initial model of the subject built at an early stage of analysis may not yet be fully appropriate. In the following analysis, the initial model is developed into an appropriate model. The development process involves repeated evaluation of the current state of the model against the requirements of Definition A.2. As these requirements are partly contradicting (e.g., generalization vs. specificity), an appropriate trade-off may need to be found.

Example. A model of the disease outbreak in Vastopolis is supposed to represent two focus behaviors: (1) the distribution of the disease cases over the population, space, and time and (2) the influences of the wind, river, and contacts between people on the evolution of behavior (1) (which is treated as a composite aspect according to Definition S.1). The model must be consistent with the data, i.e., with the observed distribution of the disease cases reflected (or surrogated) in the data. A descriptive model must explain the observed distribution by specifying the disease origin and the mechanism of the disease spread. A predictive model must forecast how the disease will evolve further, and a decision supporting model must support the choice of mitigation and recovery actions.

The model must be general, i.e., describe the outbreak as a whole and not the individual disease cases. However, the model must also be sufficiently specific. Thus, there are two distinct groups of disease symptoms (flu and diarrhea) and two distinct spreading behaviors (in the eastern and in the southwestern directions); see Fig. 5. The model needs to represent these distinctions but avoid representing unnecessary distinctions, to satisfy the parsimony criterion.

Definition A.3. *Model evaluation* is checking whether a model is appropriate.

Definition A.4. *Model development* is the process of modifying a model to make it more appropriate.

Comment. Model development may include the following operations:

- Rectify: decrease or eliminate the discrepancies between the model and the data;
- Expand the scope: modify the model so that it becomes consistent with a larger part of the data;
- Increase inclusiveness: add representation of missing focus behaviors in case of insufficient comprehensiveness;

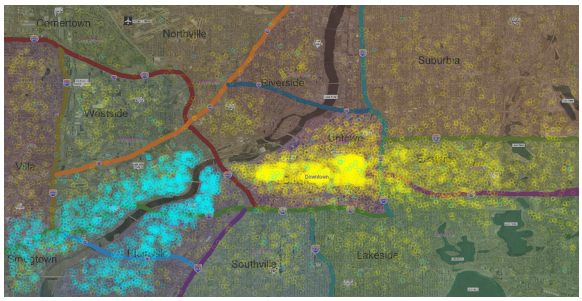

Figure 5: *A map of Vastopolis with the locations of the tweets mentioning flu and diarrhea symptoms represented by dots in yellow and cyan, respectively.*

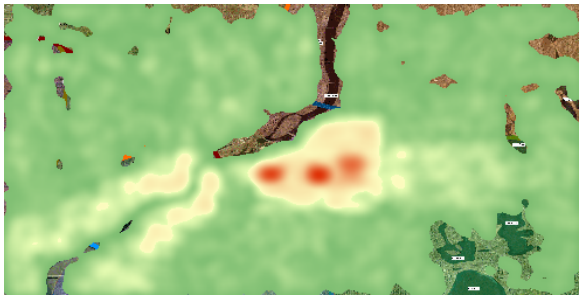

Figure 6: *Application of spatial smoothing to the positions of the tweets.*

- Simplify: decrease the number of model components, e.g., by merging some of the existing components or by subdividing the data into a smaller number of parts;
- Reduce cost: determine excessively costly components of the model and find cheaper substitutes.

The latter operation may require finding of a suitable trade-off between the cost and the accuracy of the model.

Definition A.5. A *mental model* of a subject is a model that is fully kept in the mind of the analyst.

Definition A.6. A *distributed model* of a subject is a model that is partly kept in the mind of the analyst and partly in external media, so that the part residing in the analyst's mind contains references to the external components. The mental part is called *mental frame model*.

Comment. We posit that analysis always results in a model that is at least partly contained in the mind of the analyst. When the analyst's mind does not contain the full model, it contains a frame model, which defines how all model components fit together and how the external components are used. It is possible that parts of a distributed model are contained in the minds of several analysts or experts. However, even in this case, there should be at least one analyst that holds a mental frame model, which allows the full model to be composed from the distributed parts.

Definition A.7. A *formal model* is a distributed model in-

cluding one or more components represented in computer-readable form, residing in computers, and intended for performing calculations in computers.

Comment. A model that fully resides inside a computer cannot be used without any knowledge of it by a human, i.e., without a mental model of the model residing in the computer and a structural mental model of the subject it represents. Therefore, we define a computer-oriented (formal) model as a special kind of distributed model.

Definition A.8. An *externalized mental model* is a representation of a mental model (in particular, a frame model) in external media.

Comment. Since a model resulting from analysis resides fully or partly in the analyst's mind, others will not be able to use this model unless the analyst represents the mental component of the model in external media, so that this representation can be communicated to others and internalized by them.

Definition A.9. *Provenance of a model* or model component is a representation of the way in which it was created.

Comment. In communicating a model to others, the analyst often needs also to explain how the model was obtained, so that it can be trusted by others. Provenance is also required in communicating not the model itself but results of using it, i.e., answers to questions, forecasts, or recommended actions. The analyst needs to refer the presented results to the parts of the model by which they have been obtained and also to explain where these model parts come from.

Example. In presenting answers to the VAST Challenge, it is required to substantiate each statement, i.e., to communicate the parts of the model from which these statements follow and to provide the provenance of these parts of the model.

5.3. A model as a subject or data source

There are analysis tasks in which analysts need to explore the behavior of a given computational model, such as a simulation model [SHB*14]. The analysis goal is to understand the relationships between the parameter settings and corresponding outputs,that is, in our terms, to build a (mental) behavioral model of these relationships. In such analysis tasks, the given computational model is the subject of the analysis and the model of the relationships between the parameters and the outputs is the result of the analysis.

Another possible case is when analysts deal with a computational model representing a certain subject rather than with data reflecting this subject. The role of the computational model is to produce data for the analysis. For example, the analysis goal may be to study the connectedness between different parts of the city by public transport and how it is affected by delays in the vehicle circulation with respect to the timetables. There is a computational model that takes the existing timetables and the actual tracks of the public transport vehicles as inputs and calculates the ideal (timetable-based) and real (actual track-based) trajectories of trips by public transport for given origin-destination pairs and departure times. It also computes the trip durations and waiting times required for changing from one public transport route to another. In this example, the analysis subject is not the computational model producing the trajectories and their characteristics but the public transport system. The role of the model is to produce data reflecting the behavior of the system. More precisely, it transforms the available data (timetables and vehicle tracks) that do not reflect the focus relationships (connectedness between places) into data reflecting these relationships. The expected result of the analysis is a completely different model, which identifies the parts of the city that are poorly connected to the others by the existing public transport routes and those parts for which the connectedness is greatly affected by fluctuations in the public transport circulation.

A recent trend in visual analytics is the involvement of interactive visual interfaces in derivation of predictive models using techniques of machine learning [ERT*17], particularly, deep learning (e.g., [AJY*18]). In these works, computer-generated models often appear as an additional subject of analysis (apart from the reality that is being modeled): the analyst needs to understand how they work for being able to improve (develop) them. Visual analytics aims to help the analyst "to open the black box" and consciously steer model development.

5.4. Analysis workflow

Based on the concepts introduced, the analysis process is schematically represented in Fig. 2. The yellow block at the top represents what is initially given or known. This includes

- a structural model of the subject (D M.2; here and further on, 'D' stands for 'Definition'), which may be a part of the prior knowledge of the analyst or a part of a problem statement,
- available data on the subject (D D.1),
- a task, including specification of the focus relationships (D M.4) and questions that need to be answered.

The original data may need to be transformed in order to reflect or surrogate the focus relationships (D D.2, D D.3).

Based on the data and structural model and taking into account the task, the analyst generates an initial behavioral model (D M.3) and the provenance of it (D A.9) and then evaluates the model (D A.3, D A.2). If the model is not yet appropriate, the analyst further develops the model (D A.4), thereby collecting the provenance. The steps of the model evaluation and further development are repeated until the model is judged as appropriate. Finally, depending on the task, the analyst externalizes the mental component of the

model (D A.5–A.8) for communicating to others and/or answers the questions of the task.

Note. The specifics of model building from streaming data is that a model that was built using data available at a certain moment needs to be regularly evaluated against newly appearing data. When the model becomes inappropriate, it needs to be developed further.

6. Surveying the visual analytics research

An obvious implication from the model building perspective on the analysis process is that visual analytics researchers and tool developers need to care about appropriate support for the model building activities and for the representation of analysis results, i.e., of the model and its provenance. Currently it is not done in a systematic way, although elements of such support are present in all visual analytics papers describing methods, systems, or applications. In this section, we review the literature for discussing these elements. We cannot survey all existing papers, but we have selected a representative subset of papers that covers all components of our framework and shows the existence of different approaches to supporting the activities and representations involved in the analysis process.

6.1. Focus relationships and behaviors

Visual analytics papers describing techniques or applications usually focus on certain focus relationships and present methods or procedures by which representations (models) of the behaviors of these relationships are derived from data. There is a relatively small set of generic structural relationships (D S.4) that repeatedly appear in papers as focus relationships. These include:

- entities have an attribute or multiple attributes;
- entities appear and disappear over time;
- entities have locations in space;
- entities change their locations in space;
- attributes change over time;
- attributes vary over space;
- entities interact with other entities;
- entities consist of other entities or contain other entities;
- entities are arranged with respect to each other.

These common structural relationships are realized in various subjects (D S.6). Thus, the relationships ‘entities appear and disappear over time’ and ‘entities are arranged with respect to each other’ are present in collections of electronic health records [MLL*13], social media posts [WLY*14], people’s daily behaviors [VJC09], and many other subjects. While the behavior of a structural relationship may differ from subject to subject, there may be common types of features pertaining to different behaviors. Thus, for the relationship ‘entities appear and disappear over time’, pertinent behavior features occurring in various subjects are randomness, constancy, temporal trends, or periodicity regarding the temporal frequency of the appearance of entities, or the lifetimes of the entities, or re-appearances of the same entities (or the same categories of entities). Similarly, common behavioral features could be listed also for the other structural relationships.

Directions for visual analytics research. It might be useful for the development of visual analytics as a science to create a taxonomy of common structural relationships and their pertinent behavioral features. If visual analytics researchers explicitly refer the approaches they propose to elements of this taxonomy, it may be easier for other researchers and for practitioners to find existing approaches oriented to particular structural relationships and behavioral features. Thereby, approaches could be transferred from subject to subject, and the applicability of the approaches could be extended. Another possible use of the relationship and behavior taxonomy is systematic cataloging of the existing approaches and discovery of gaps in the coverage of the relationship space.

6.2. Data transformations

Data transformations are widely applied in visual analytics tools and workflows. Thus, it is very typical to transform unstructured data (images, videos, and texts) into structured records. The goal is not only to make data more suitable for machine processing but also to have the behaviors of certain focus relationships reflected or at least surrogated in the data (D D.2, D D.3). For example, Matković et al. [MGS*14] are interested in the distribution of image attributes, such as lightness and colorfulness, over a set of images. To obtain data that reflect this behavior, Matković et al. derive the attributes from the sets of image pixels. For Gu et al. [GWM*15], the focus behavior is the distribution of contents over a set of images and texts. However, the contents are not directly represented in the data, and, moreover, extracting contents from unstructured data is a complex problem for which no general solution exists. Therefore, Gu at al. create a surrogate for the focus behavior based on an assumption that documents with similar contents have similar characteristics in terms of specific attributes, such as grayscale content, power spectrum, and color histogram of an image. These attributes are derived from the original data. Based on attribute similarities and distinctions, the documents are arranged on a plane, to allow the user to perceive and inspect the distribution of the contents.

Oelke et al. [OSSK12] are interested what attributes of a text affect its readability. In our terms, the focus behavior is the influence of attributes on another attribute. To reveal and represent this behavior, Oelke et al. derive various attributes of texts (e.g., word length, sentence length, sentence structure complexity, etc.) and compare these attributes for easily readable documents and those that are difficult to read. Many other examples of derivation of various secondary data from

texts, images, and video records (e.g., [HHHW13]) exist and continue to appear in the literature.

Andrienko et al. [AAB*13] present in a systematic way the possible transformations of movement data that change the data structure for adapting it to different tasks. For example, trajectories of moving objects can be transformed into time series of situations (i.e., spatial distributions of the objects and their movement) if the focus behavior is the distribution of the overall movement over space and its evolution in time. Another example is transformation of geographic coordinates into relative positions with respect to the center and the movement direction of a group of coherently moving objects. The transformed data reflect the behaviors of the group members with respect to the group (e.g., leader, follower, explorer, wanderer, etc.).

Directions for visual analytics research. Similarly to the inventory of the possible transformations of movement data [AAB*13], it would be useful to systemize the known transformations of other data types and to explicitly specify for each transformation (including those described in [AAB*13]) what kinds of relationships and behaviors are reflected in the resulting transformed data. This information would be valuable for researchers and practitioners.

6.3. Generation of an initial model

Generation of an initial model can be supported either by visual representation of available data allowing the analyst to perceive the focus behaviors or by computational derivation of features characterizing these behaviors. Since the model being built needs to represent the focus behaviors in a generalized way (D M.3), the visual representation must involve or promote generalization and abstraction from individual relationship instances to a holistic view of the relationship behavior. For this purpose, the visualization may involve data aggregation and/or smoothing, or it may represent individual instances in a way enabling an overall view of the whole relationship behavior. For example, a scatter plot represents individual instances but allows the user to see the presence or absence of an overall or partial correlation between two aspects. Similarly, the map in Fig. 5 represents the locations of individual tweets but shows the overall spatial distribution not less clearly than the map in Fig. 6, where spatial smoothing is applied. Van Wijk and van Selow [vWvS99] support holistic perception of a large set of numeric time series by drawing them one next to another in a 3D display.

Spatialization [SF08] supports abstraction by involving spatial metaphors and relying on the human inherent capability to perceptually unite spatially close items and interpret them as similar or strongly related. The main idea of spatialization is to arrange visual objects within the display space in such a way that the distances between them reflect the degree of similarity or relatedness between the data items they represent. This is achieved by using various projection techniques [ELP*16]. Spatialization can be applied to data of various kinds: entities described by multiple numeric attributes [TMN03], text documents [PNML08], images [ENP*09], states of evolving networks [vdEHBvW16], and others. Bach [BSH*16] introduced Time Curve, a general technique involving spatialization for visual representation of temporal evolution of an arbitrary object; examples include evolving texts, video recordings, geographic phenomena, and brain connectivity networks. Different states of an object are represented by dots in the screen space, and the Time Curve connects these dots following the chronological order of the states. The curve is perceived as a single object, and characteristics of the temporal behavior can be understood from the shape of the curve. The authors explain how to interpret different patterns that can be observed in a curve. Elzen et al. [vdEHBvW16] use a similar technique for representing the evolution of a network.

For complex data, generalization can be supported by applying clustering (i.e., grouping by similarity) to relationship instances and presenting the resulting clusters so that the analyst can holistically perceive them. For example, clustering can be applied to multiple time series, and each resulting cluster can be represented by a time series of average values [vWvS99]. Höferlin et al. [HHHW13] apply clustering to trajectories of moving objects extracted from surveillance video records and represent the clusters in a summarized form as flows. Time intervals can be clustered according to the similarity of time-related instances, e.g., traffic situations in [AAB*13] or hourly time series [vWvS99]. The result is represented in a calendar view [vWvS99] or time mosaic [AAB*13] by coloring display elements corresponding to different time intervals according to their cluster membership. Such a visualization supports holistic perception of a behavior over time.

A well-known example of computational characterization of relationship behaviors is scagnostics [WAG05, WW08], which computes various measures for a relationship between two numeric attributes: outlying, skewed, clumpy, convex, skinny, striated, monotonic, stringy, straight, and monotonic. The measures are derived from a scatterplot representing relationship instances. Another well-known example is the rank-by-feature framework [SS04], which derives another set of measures for a relationship between two numeric attributes: correlation coefficient, least square error for curvilinear regression, quadracity, the number of potential outliers, and uniformity. It also characterizes the distribution of a numeric attribute over a set of entities (represented visually by a histogram) by such features as normality, uniformity, the number of potential outliers, the number of unique values, and the size of the biggest gap.

Another kind of computational support is discovery of frequent patterns, or motifs (e.g., [HMJ*12]). This does not represent a behavior as a whole but reveals groups of similar

sub-behaviors and thereby gives a general idea of how diverse and how self-repeating the overall behavior is. An appropriate visual representation of the results of the frequent pattern discovery allows the analyst to see the frequency and regularity of the pattern occurrence as well as the frequency and amount of unique or infrequent sub-behaviors.

When the analysis aims at building a formal model (D A.7) using statistical or data mining techniques, the analyst nevertheless needs to have an initial mental model of the subject for choosing a suitable modelling method, providing appropriate input data to the modeling software, and setting the method parameters. In visual analytics literature, there are many examples of supporting the analyst in obtaining an initial mental model followed by choosing and setting a modelling method. The approaches often combine visualizations with computational derivation of behavior features, as discussed earlier in this section. There are approaches oriented to different classes of formal models, in particular, classification [Gle13], regression [MP13], time series models [BAF*13], and spatio-temporal models [AA13].

Directions for visual analytics research. Various methods of visual and computational support to creation of an (initial) mental model of a subject are currently dispersed among many application-specific works. Visual analytics research and tool development would benefit from generalizing and cataloguing the existing approaches, which would promote the transfer of the methods to new applications. This would be especially useful for methods dealing with complex and/or very large data, for which there are no obvious ways for holistic representation of behaviors.

6.4. Model evaluation

6.4.1. Formal models

For formal models (D A.7), which are built with the use of statistical and machine learning methods, evaluation is an essential part of model building. A typical practice is to derive a model from a subset of available data and then test it on the remaining data. The model is expected to generate data that are very close to the test data. Model accuracy is checked by comparing the model-generated data with the test data. There exist statistical measures of model accuracy, such as root mean squared error, mean absolute percentage error, and mean absolute scaled error. It is typical to perform multiple runs of model evaluation with different partitioning of the available data into training and test data. This process is called cross-validation. The error measures from the multiple runs are averaged.

Statistical software packages and libraries (e.g., R [RC17]) often not only provide various modelling methods but also automatically perform cross-validation and calculate the error measures. Moreover, model parameters can be automatically tuned for minimizing the errors. The model evaluation functionality provided by existing statistical software can be exploited in visual analytics tools (e.g., [BAF*13]). However, it is not sufficient to calculate the numeric measures of model errors. It is necessary to look at the behavior of the model residuals, i.e., the deviations of the model-generated data from actual data. A model perfectly represents a focus behavior if the behavior (distribution) of the residuals is random. Visual analytics tools visualize the distributions of model residuals allowing the analyst to check for apparent randomness or non-randomness, i.e., for absence or presence of visible patterns [AA13, BAF*13, MP13].

Classification models that deal with qualitative data are evaluated based on the proportions of correctly and wrongly classified data items (e.g., [MW12, SHJ*15]. Apart from the numeric measures, it is useful to visualize the distribution of the correct and wrong classification results over the data space and to show the decision boundaries of the model [MW12].

The numeric quality measures introduced for formal models are only suitable for the assessment of the model correctness and scope (D A.2). Formal models usually satisfy the criterion of generalization. Regarding the purpose, formal models are typically predictive (D M.6). It is their predictive capability that is exploited in the testing and cross-validation. Many types of predictive models can also be used as descriptive. For this purpose, they need to be representable in a human-interpretable form. This does not hold, for example, for models based on neural networks. Such models can be used for forecasting but not for describing and explaining the behaviors they represent, i.e., they are not appropriate for the analysis task 'assess'. Formal models alone are not decision supporting (D M.8) as they do not represent actions (D M.7).

Checking formal models against the remaining criteria of model appropriateness (D A.2) is under the responsibility of the analyst. Visualizations can help the analyst to assess model specificity by showing what parts of the behavior are represented well enough and where the model is incorrect. In a case of an inhomogeneous behavior, it is reasonable to represent it by a combination of several partial models (i.e., covering subsets of relationship instances) rather than by a single global model. An example is representing street traffic over a territory by a set of time series models [AA13].

6.4.2. Mental models

An analyst can evaluate a mental model using visual displays and interactive techniques provided by visual analytics software. Possible operations are re-aggregation (using a different set of bins, or smoothing with a different kernel), re-clustering (changing clustering parameters), and taking a random sample of the data. From visual displays of the modified data, the analyst judges whether what is seen is consistent with the mental model. In principle, the analyst can apply the same approach as in statistical cross-validation: use

a subset of randomly selected data items for creation of an initial mental model and then evaluate the model using the remaining data.

Although many visual analytics systems and toolkits include interactive facilities for the operations mentioned above, it has to be a decision of the analyst to apply these operations. The existing software neither informs/reminds the analyst about the possible use of the available interactive techniques for mental model evaluation nor encourages the analyst to even concern about such an evaluation. Moreover, not every software supports comparisons between the results of different aggregations or applications of clustering.

In principle, it is possible to gently engage the analyst into evaluation activities and to help in performing these activities. Thus, at the beginning of the work, the software could propose the analyst to use a randomly selected subset of data for initial overview and preliminary analysis and to reserve the remaining data for checking the validity of observed patterns and trends. Whenever the analyst performs data aggregation, smoothing, or clustering, it may be useful to produce several results through automatic variation of the analyst-chosen settings and enable visual comparison between the variants. These or other approaches to engaging analysts into model checking need to be tested for adoptability and effectiveness, i.e., whether analysts can develop better models and /or gain more confidence in the analysis results.

Directions for visual analytics research. Currently, there are not many visual analytics methods and tools explicitly supporting model evaluation, except the involvement of statistical techniques and measures for the evaluation of formal model accuracy. There is a wide space for research on how to support model evaluation beyond computing model errors and, especially, how to explicitly support and stimulate evaluation of mental models.

6.5. Model development

According to the results of model evaluation, the analyst may need to perform certain model development operations (D A.4). Formal models can sometimes be improved by modification of model parameters, which can be supported by interactive visual interfaces (e.g., [BAF*13]). However, parameter adjustment alone may be ineffective when the focus behavior is inhomogeneous. A reasonable approach to reducing discrepancies between a model and available data is to partition the data according to behavior variations and apply the modeling tool separately to each partition [AA13,MP13]. Visualization of the distribution of the model errors over the dataset, which is made at the evaluation stage, may suggest the analyst how the data should be partitioned. Appropriate partitioning not only can rectify the model but may also lead to model simplification [AAR*16a]. Performance of classification and regression models can also be improved by interactive modification of the set of features that are used for modeling [MP13].

Partitioning is, in fact, a general approach applicable to building of both formal and mental models. When the generation of an initial model is done with the use of clustering (Section 6.3), it naturally leads to partition-wise modeling. In this case, expanding the model scope means building a sub-model for a partition that is not yet covered. Model rectification, when necessary, can be achieved by subdividing the existing partitions [AA13].

Partitioning can also be done interactively. For example, one may visually detect two distinct sub-behaviors in the spatial distribution of the disease cases in the Vastopolis outbreak (Figs. 5 and 6) and interactively divide the data into three subsets: the cases located along the river, the cases concentrated in the city center, and the remaining cases that appear to be uniformly spread over the territory. Wang and Mueller [WM17] describe a system for derivation of causal models, where the user interactively refines a model derived by a causal inference algorithm by defining meaningful data subdivisions. In system TimeNotes [WBJ16], parts of long time series can be interactively selected for detailed exploration, compared with others, grouped by similarity, and labeled as instances of specific behaviors.

In building a model of a complex behavior involving multiple aspects of diverse nature, such as space, time, and entities or attributes, the analyst may begin with a simpler behavior involving fewer aspects. For example, in investigating the Vastopolis outbreak, the analyst may initially focus on the temporal distribution of the tweets mentioning disease symptoms and identify the time when the outbreak began. Next, the analyst may explore the spatial distribution of the relevant tweets posted after the identified time. After detecting two distinct spatial clusters, the analyst may seek for explanation by comparing the symptom occurrences between the clusters. Then, the analyst may investigate the temporal evolution of each cluster. This example (presented in [AAB*13]) demonstrates gradual increase of model inclusiveness regarding the focus relationships and focus behaviors.

Directions for visual analytics research. Many examples of gradual interactive model development exist in the literature, often in case study descriptions. There is a need for generalization and systematization of the approaches. Another need is to consider the problem of model cost-effectiveness, which has not yet been sufficiently addressed in the literature. Thus, a model involving time-consuming calculations may be inappropriate for practical use. While possible approaches to choosing suitable substitutes for expensive models or model components may be case-dependent and not easily generalizable, it may be possible to develop general methods for assessing the losses and gains due to such substitutions.

6.6. Model externalization and provenance collection

Recommendation 2.1 in "Illuminating the Path" [TC05, p. 42] calls for the development of "knowledge representations to capture, store, and reuse the knowledge generated throughout the entire analytic process". In our terms, it refers to externalized representations of the mental models built (D A.8). As we stated in Section 5.1, even when a model is built in a computer, the analyst needs to have a mental frame model allowing appropriate use of the computer model.

Capturing and explicitly representing human's knowledge (mental models) is a very hard problem, which was earlier addressed in the research field of artificial intelligence [SBF98]. Many knowledge representation methods were proposed, as well as some techniques for expert knowledge elicitation [Coo94]. Still, it remains a hard and tedious job, usually performed by trained knowledge engineers, to convert knowledge stored in the expert's mind into an explicit form.

The task in visual analytics is more specific: to elicit the new knowledge emerging during the analysis and the relevant prior knowledge used by the analyst. Both the new and prior knowledge are related to what the analyst is viewing and doing. The existing visual analytics systems and platforms thus support knowledge externalization by enabling the analyst to annotate and comment on visualizations [WSP*06, EKHW08, SvW08, HVW09, WBJ16]. Ribarsky and Fisher [RF16] stress the importance of enabling the analyst to externalize his knowledge at any time, not only at the end. The annotation module must be callable whenever and wherever needed during the analysis process. What can be captured in this way is fragments of the analyst's mental model (which get an explicit representation in the form of texts) and, simultaneously, the provenance of these fragments, due to the maintained links of the analyst's notes to the visualizations (and, moreover, to specific parts of these visualizations) that motivated making these notes.

Fragmentary notes associated with different views do not form an adequate representation of the model the analyst has in mind. For a more complete and systematic external representation of the model, analysts may compose a "story" [EKHW08] with hypertext links to display snapshots, or organize their notes in a graph [SvW08], or arrange and link them within a workspace [WSP*06], or explicitly define and manage a system of concepts and instances of these concepts [GZA06]. Zhao et al. [ZGI*18] describe how a concept map, called Knowledge Transfer Graph, facilitates asynchronous collaboration by supporting the externalization of the analytic process through dedicated graph elements. The graph is used for transferring knowledge from one analyst to another. The described system enables interactive playback of graph creation for seeing how the concepts and links were derived. Concept maps can also be automatically derived from text notes [WSP*06].

Despite the provided support, externalization of a mental model is a laborious and time-consuming job. Systems oriented to specific types of data and analysis tasks may reduce the effort of the analyst by automatically detecting certain types of behaviors and patterns in data and creating draft annotations [EKHW08] or by providing templates for the analysts to organize their findings [WSP*06]. In our terms, a template is a representation of the structural model (D M.2) of a subject or an aspect. A template explicitly presents relevant aspects and structural relationships to the analyst thus giving a direction to the analysis. By filling a template with notes describing observed behaviors, the analyst builds a behavioral model (D M.3). Obviously, templates need to be previously loaded in the system. The possibilities for automatic behavior discovery and for the use of templates are briefly mentioned but not elaborated in the respective papers [WSP*06, EKHW08].

In the systems supporting knowledge externalization, the provenance is represented in the form of links from concepts or descriptive notes to visual displays or display areas containing relevant evidence. Additionally, the analysis history, i.e., the displays viewed and operations performed by the analyst can be automatically tracked and visualized, allowing the user to return to earlier steps and try alternative analysis paths (e.g., [SvW08]). Gotz and Zhou [GZ09] describe how a taxonomy of the user's actions can be used for automatic capture of semantically meaningful and logically organized provenance. This requires the system to have a "semantic" user interface organized according to the action taxonomy.

So far, research on knowledge externalization and provenance collection did not consider analysis processes in which formal models (D A.7) are built. There are types of formal models that allow representation in a human-readable form; thus, causal models are representable in the form of causal networks [WM17]. In other cases, externalization is required for the mental frame models (D A.6) that allow proper use of the computer-based components. This problem has not been addressed yet, while there are examples of provenance collection for computer models [AAR*09].

Directions for visual analytics research. In the existing systems, the level of facilitation of the knowledge externalization job is insufficient; it still requires much effort of the analyst beside the analysis itself. The model building view of the analytical process may be helpful for finding new approaches to ease the analyst's burden. The ideas of using knowledge templates [WSP*06] and automated detection of potentially interesting patterns [EKHW08] can be re-considered and further developed. Thus, a system could help the analyst to externalize a structural model of the analysis subject, which gives a basis for automatic construction of knowledge templates.

Being represented in a computer-interpretable form, a structural model of a subject could be used for automated data transformation or pattern extraction. Computers could also create draft annotations and/or fill in knowledge tem-

plates. The use of structural models may be a direction for the further development of the ideas of creating "semantic" user interfaces and semantics-aware provenance capture [GZ09]. In short, the computer can better assist the human if the former "understands" what the latter is going to do and why. A suitably represented structural model may provide such an "understanding". Some of the knowledge representation methods from artificial intelligence may be applicable for representing structural models.

In provenance collection, it might be useful to distinguish between initial model building, evaluation, and development, and to reflect the operations performed for model evaluation and their results, including formal measures and the analyst's judgements. It would be also good to track the modifications of the model under development and capture the reasons for these modifications.

Research is also needed on externalizing mental frame models of distributed models (D A.6) and capturing their provenance.

6.7. Specific research on predictive and decision supporting models

The discussion in Sections 6.1–6.6 refers to all model types, including predictive and decision supporting models (D M.5–M.8). As mentioned in Section 6.4, formal models are typically predictive, but predictive mental models are also possible. Having grasped a trend or regularity in a focus behavior, the analyst can extrapolate this behavior beyond the part that is reflected in the available data. Thus, in the VAST Challenge scenario, the trend in the temporal behavior of the disease is that the number of new cases declines, which allows the analyst to forecast that the epidemic spread will stop soon.

A specific requirement to the support of predictive modeling is to make sure that a model can correctly predict outputs for inputs that were not used for model creation, as discussed in Section 6.4.

Simulation models are a special subtype of predictive models. Building of simulation models has been mostly out of the scope of visual analytics (with some exceptions [AAR16b]), whereas analysis of behaviors of existing simulation models is quite a popular topic [SHB*14]. From our perspective, a simulation model in such an analysis is the subject, and the task is to build a descriptive model of its behavior.

There are visual analytics papers on supporting decision making with the use of simulation models, which forecast the development of, e.g., a pandemic [AME11] or a flood [WFR*10, KWS*14]. Interactive tools enable analysts to assess the forecasts, imitate implementation of possible actions (D M.7), observe their effects (as predicted by the simulation models), and compare the expected results of different actions. An action plan can be generated automatically and visually presented to the analyst, including justifications for the decisions taken [KWS*14].

These examples demonstrate the use of decision supporting models (D M.8). The specifics of the latter compared to the other model classes is the involvement of actions, a.k.a. decision measures [AME11]. A decision supporting model is a procedural representation in the sense that it can simulate the execution of actions and action plans. In current visual analytics systems for decision support, action representations are either in-built [AME11, KWS*14] or imitated by the analyst through making changes in the simulation settings [WFR*10, AAR16b]. In-built actions may have parameters to be set by the analyst [AME11], such as when, where, how long, how much, etc. The system translates these settings into inputs or parameters for the base simulation model.

The research on decision modeling and decision support in visual analytics has not yet reached maturity. In particular, the possibilities for representing actions are quite limited. Thus, in real world applications, it may be necessary to consider not only the action capabilities for attaining desired effects but also their costs, which may include monetary costs, resource consumption, negative impacts on the population or environment, or other unwanted consequences.

Directions for visual analytics research. Further research towards comprehensive support to decision making is required. It is necessary to take care for proper representation of possible actions, including their positive and negative consequences, costs, and applicability conditions regarding the inputs, in particular, those that are not under the control of a decision maker. It is necessary to support cost-benefit analysis of the action execution and to facilitate the development of cost effective action plans.

The existing works in visual analytics deal with the application of decision supporting models rather than the development of such models. It may not always be possible to create a complete visual analytics embedding for the development of a decision supporting model, in particular, when a simulation model is required for imitating and testing action execution. Simulation models are often not derived from data but developed on the basis of theories or analogies. Assuming that a suitable simulation model exists, the task for visual analytics research is to support analysts in defining possible actions and to devise methods for translating the action definitions into parameter settings or inputs of the simulation model.

7. Discussion

Here, we discuss the intended use of our framework, its properties, and relationships to the antecedent frameworks.

7.1. Use of our framework

7.1.1. Systematic view of the research field

Based on the model building perspective, we propose a systematic approach to reviewing the research in visual analytics. Rather than organizing the work done according to data types or techniques involved, we arranged the state of the art according to the components of the visual analytics process (Section 6). This arrangement allowed us to identify several promising directions for theoretical research or meta-research in visual analytics (Sections 6.1, 6.2, 6.3, 6.5) and several areas where insufficient research has been made so far. The latter refers, in particular, to support for model evaluation (6.4), externalization and provenance capture (6.6), and decision modeling (6.7). Adopting the model building perspective gives a better understanding of what needs to be supported. We hope that this will promote new ideas concerning how to provide the required support.

7.1.2. Practical use

By adopting the idea that the visual analytics process is directed to building of a behavioral model of a piece of the real world, a visual analytics researcher or tool developer gains a certain practical guidance in designing methods, procedures, and tools. Thus, what is expected from visualization is representing not data per se but the behaviors of focus relationships. So, it is necessary to identify the focus relationships and think how they can be represented. It is also necessary to check whether these relationships are reflected in the data and, if not, find or devise suitable transformations to derive the required data. The understanding that behaviors need to be represented in a generalized way motivates looking for visual representations and data transformations facilitating generalization. Since the model under construction must be repeatedly evaluated and developed, researchers and tool developers need to care about proper support to model evaluation and development. The understanding that an essential part of the model, if not the entire model, resides in the mind of the analyst calls for methods and tools facilitating model externalization for later recall and for communication. The need to communicate the model to others requires proper support for collection of the model provenance along the analysis process. To summarize, a researcher or developer needs to perform the following actions:

1. Define the subject (D S.1) that needs to be analyzed and the purpose of the analysis (D A.1), i.e., the type of the model that needs to be built: descriptive, predictive, or decision-supporting (D M.5, M.6, M.8).
2. Identify the essential aspects of the subject (D S.2) and the kinds of relationships between these aspects that need to be in focus (D M.4). For building a decision-supporting model (D M.8), the essential aspects include actions capable to change other aspects (D M.7).
3. Study the structure of the available data: are the essential aspects directly reflected (D D.2) in data components?
   - 3.1. If not, determine what data components indirectly reflect the essential aspects and find ways to transform the data for suitable reflection of the essential aspects.
   - 3.2. For aspects reflected neither directly nor indirectly, assume that they are represented in the prior knowledge of the analyst and find a way in which the analyst can inject the relevant knowledge in the analysis.
4. For the identified focus relationships, find methods of computational extraction and/or visual representation. The methods must involve or support generalization (D M.3), e.g., through smoothing or aggregation.
5. Find methods for checking the validity of the extracted or human-perceived relationships, i.e., conformity to the data (D A.2–A.3).
   - In particular, in building predictive models, different portions of data need to be used for deriving and for checking relationships.
6. Find methods to support offloading of model parts to external media (D A.8).
7. Find or develop tools for capturing and representing the model provenance (D A.9) along the analysis process.

In a search for suitable approaches to accomplishing these actions, the researcher or developer can use the proposed survey (Section 6), which is conveniently structured according to the action list. The results of research or tool development can then be evaluated by checking how well the process represented in Fig. 2 is supported. It may be hard to cover the whole process in a single research work. In case of addressing only a subset of activities, the evaluation means checking how well these activities are supported.

7.2. Self-applicability of our conceptual framework

Let us try to characterize the proposed conceptual framework using its own concepts. In fact, it is a model (D M.1) of a visual analytics process, which is the subject of our study (D S.1). Figure 1 shows a structural model of the subject (D M.2), i.e., the main aspects (D S.1,S.2) of the visual analytics process and the structural relationships (D S.4) between them. On this basis, we have built a behavioral model (D M.3) of the process of building a model of a subject (Fig. 2 and Section 5.4). Our focus relationships (D M.4) are:

- data may reflect behaviors of relationships between aspects of a subject;
- a model represents behaviors of relationships;
- a model has attribute ‘appropriateness’;
- a model is obtained using visual analytics techniques: data transformations, visualizations, interactive operations, and computational analysis techniques.

In Section 5.2, we described in a general way the behaviors (D S.6) of these focus relationships: how data reflect relationship behaviors and what can be done if a focus relationship is not reflected in data, how a model represents a behavior, and how this representation can be achieved using visual

analytics techniques. We decomposed the attribute 'appropriateness' (D A.2) into simpler components and characterized the possible relationships of these components to visual analytics techniques. Our descriptions of these relationship behaviors are general, i.e., apply to many instances (concrete realizations; D S.5) of the visual analytics process. In Section 6, we corroborated our general statements by discussing numerous research works in visual analytics in which the focus relationships have been instantiated. This can serve as a kind of provenance (D A.9) of our model.

7.3. Explanatory capabilities of our framework

Our framework does not only describe the visual analytics process and its outcome but also explains some common opinions and adopted practices in visual analytics:

- Why a human analyst is an essential actor in data analysis: because, first, a human has (as prior knowledge) or is capable to acquire (e.g., from a task description) a structural model (D M.2) of a subject, which needs to be used in the analysis (Fig. 2), and, second, an essential part of the behavioral model that is built resides in the mind of the human (D A.5–A.6).
- Why the visual analytics process needs to be done in an iterative way: because, in order to make a model fully appropriate (D A.2), it needs to be repeatedly evaluated against the appropriateness criteria (D A.3) and further developed (D A.4) when some criteria are not satisfied.
- Why data transformations may be necessary (thus, the research agenda for visual analytics [TC05] gives much attention to data transformations): because available data may not originally reflect a focus behavior (D D.2). Data are transformed to make this behavior reflected or at least surrogated (D D.3).
- Why "Overview first" [Shn96]: because the behavioral model being built must represent behaviors in a generalized way (D M.3), abstracted from individual relationship instances (D S.5). Visual representations and data transformations facilitating abstraction and generalization are therefore very important.

7.4. Relation to the previous frameworks

Our framework does not differ in essence from the previously proposed frameworks representing the analytical process, and it accommodates well their key concepts. Moreover, our framework elaborates the previous ones by defining many of the concepts more explicitly and/or in more detail. Table 1 shows how the concepts from the previous frameworks can be translated into ours.

The previous frameworks do not say explicitly when the analysis finishes. According to our framework, the analysis terminates when the whole behavioral model under construction is marked as correct and the other appropriateness criteria (D A.2) are also met. Unlike the previous frameworks, which do not consider the purpose of generating the knowledge, we specify, in accord with [TC05], that the final product of the visual analytics process must fit to the purpose, i.e., to one or more of the tasks 'assess', 'forecast', and 'develop options'. Our framework differs from the others also by considering model provenance as an essential part of the result of the analytical process (Fig. 2).

Hence, our representation of the visual analytics process details and extends the previous frameworks that represent the analytical process. In our terms, by developing the previous models (D A.4), we create a more comprehensive and more specific model (D A.2). A note needs to be made concerning the representation of the cognitive aspects of the analytical process. Our framework represents the analysis process as a cognitive activity by which the analyst builds a mental model (D A.5) or a distributed model with a mental frame model residing in the analyst's mind (D A.6). The framework also encompasses externalization of a mental model (D A.8) and representation of the model provenance (D A.9). Our framework represents these concepts referring to human cognition in a generalized way, leaving space for their elaboration. The latter can be achieved by including relevant concepts from the frameworks that predominantly focus on human cognitive activities [GRF09, RFP09, LS10, SSS*14, RF16, FWR*17].

8. Conclusion

We started our research from trying to answer the question: What is the aim of visual analytics activities? It is commonly believed that the aim is gaining new knowledge, but it is rarely discussed what kind of new knowledge is expected and how it should relate to the tasks 'assess', 'forecast', and 'develop options' [TC05]. Our answer is that the aim is to obtain knowledge that can be used for accomplishing one of these tasks. Since the tasks come from the real world, the knowledge must represent a task-relevant piece of the real world and not just the available data. We call such a representation 'appropriate model'.

Hence, the goal of visual analytics activities is not to analyze data per se but to build an appropriate model of a piece of the world. Data are needed for building the model. Accordingly, the goal of the visual analytics science is to develop methods, tools, and procedures enabling analysts to obtain appropriate models of various subjects from various kinds of data.

Having adopted this view, we defined and described the necessary components of the analysis process and conducted a structured analytical review of the state of the art in the visual analytics science by arranging the research achievements according to these components and considering the ways in which these components have been addressed and supported by researchers. We also identified the areas of vi-

Card et al. [CMS99]	Pirolli & Card [PC05]	Thomas & Cook [TC05]	Sacha et al. [SSS*14]	Ribarsky & Fisher [RF16]	Lammarsch et al. [LAB*11]	This paper
		issue	problem domain			subject: S.1, S.2, S.4-S.6
relevant data	relevant data	relevant information	data	data	data	data on subject: D.1-D.3
task/problem	task	task/problem				goal/task; focus behaviour: M.4
		task types: assess, forecast, develop options				task types and model types: M.5-M.8
				prior knowledge	domain knowledge	structural model; pre-existing behavioural model(s) for some aspects of the subject: M.2-M.3
forage for data	foraging loop	gather information	data preparation			obtain data reflecting or surrogating focus behaviour: D.2-D.3
		transform data	transform data			transform data
		pattern	finding; pattern	finding		observation referring to multiple instances of a focus relationship: S.4-S.5, M.4
			insight (interpreted finding)			finding linked to the structural model
schema	schema	structure		concept/schema	model	structural model; behavioural model (tentative): M.1-M.3
search for schema	search for relations; schematize	generate candidate explanations (hypotheses)	exploration loop	abduction	take models from domain knowledge; create models from data	generate initial model
			model (computer-generated)			formal model: A.7
	evidence	evidence	evidence	evidence		data consistent with tentative model
	hypothesis	hypothesis; candidate explanation	hypothesis	hypothesis	hypothesis	part of a tentative model
instantiate schema with data	search for evidence; search for support; re-evaluate	evaluate alternative explanations	verification loop	hypothesis confirmation	validate hypotheses	evaluate model: A.2-A.3
improve schema or search for a better schema	sense making loop; build case	consider other explanations		hypothesis development	form models by validating hypotheses	develop model: A.4
	theory or case	model; scenario	new knowledge (verified hypotheses)	model	insights	appropriate behavioural model: M.3, A.2
problem-solve	tell story	articulate defendable judgement			application	answer questions
decide or act					application	use model results
package the patterns in some output product		capture and represent knowledge		externalization/ presentation	externalize hypotheses	externalize model: A.8
		track the analytical process		attach annotations to reasoning steps		keep provenance: A.9

Table 1: *Mapping of concepts from different models and frameworks to our conceptual framework.*

sual analytics where further research is needed for supporting practitioners and revealed the research directions for advancing the visual analytics science. Being consistent with previously proposed frameworks, our framework extends them and makes them more specific.

Acknowledgements. This work was partially funded by German Science Foundation (DFG) in priority research program SPP 1894 on VGI and by Austrian Science Fund (FWF) in projects #P22883, #P25489-N23, and #P28363; CVAST #822746.

References

[AA06] ANDRIENKO N., ANDRIENKO G.: *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer Verlag, 2006. 6, 7, 9

[AA13] ANDRIENKO N., ANDRIENKO G.: A visual analytics framework for spatio-temporal analysis and modelling. *Data Mining and Knowledge Discovery 27*, 1 (2013), 55–83. 16, 17

[AAB*13] ANDRIENKO G., ANDRIENKO N., BAK P., KEIM D., WROBEL S.: *Visual Analytics of Movement*. Springer-Verlag Berlin Heidelberg, 2013. 15, 17

[AAR*09] ANDRIENKO G., ANDRIENKO N., RINZIVILLO S., NANNI M., PEDRESCHI D., GIANNOTTI F.: Interactive visual clustering of large collections of trajectories. In *2009 IEEE Symposium on Visual Analytics Science and Technology* (Oct 2009), pp. 3–10. 18

[AAR*16a] ANDRIENKO G., ANDRIENKO N., RYUMKIN A., RYUMKIN V., KRAVCHENKO G., TYABAEV E., KHLOPTSOV D., TROFIMOVA S.: Exploration and refinement of regression tree models with interactive maps and spatial data transformations. *International Journal of Cartography 2*, 1 (2016), 59–76. 17

[AAR16b] ANDRIENKO N., ANDRIENKO G., RINZIVILLO S.: Leveraging spatial abstraction in traffic analysis and forecasting with visual analytics. *Information Systems 57* (2016), 172–194. 19

[AJY*18] ALSALLAKH B., JOURABLOO A., YE M., LIU X., REN L.: Do convolutional neural networks learn class hierarchy? *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 152–162. doi:10.1109/TVCG.2017.2744683. 13

[AME11] AFZAL S., MACIEJEWSKI R., EBERT D. S.: Visual analytics decision support environment for epidemic modeling and response evaluation. In *IEEE Conference on Visual Analytics Science and Technology (VAST)* (2011), IEEE, pp. 191–200. 19

[AMST11] AIGNER W., MIKSCH S., SCHUMANN H., TOMINSKI C.: *Visualization of Time-Oriented Data*. Springer, 2011. 6, 7

[AS05] AMAR R. A., STASKO J. T.: Knowledge precepts for design and evaluation of information visualizations. *IEEE Transactions on Visualization and Computer Graphics 11*, 4 (2005), 432–442. 6

[BAF*13] BÖGL M., AIGNER W., FILZMOSER P., LAMMARSCH T., MIKSCH S., RIND A.: Visual analytics for model selection in time series analysis. *IEEE Transactions on Visualization and Computer Graphics, Special Issue "VIS 2013" 19* (12 2013), 2237–2246. 16, 17

[BL09] BERTINI E., LALANNE D.: Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: integrating Automated Analysis with interactive Exploration, VAKD* (2009), ACM, pp. 12–20. 6

[BM13] BREHMER M., MUNZNER T.: A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2376–2385. 7

[BME*07] BOOCH G., MAKSIMCHUK R., ENGLE M., YOUNG B., CONALLEN J., HOUSTON K.: *Object-oriented Analysis and Design with Applications, Third Edition*, third ed. Addison-Wesley Professional, 2007. 8

[BS97] BRODLEY C. E., SMYTH P.: Applying classification algorithms in practice. *Statistics and Computing 7*, 1 (Jan. 1997), 45–56. 5

[BSH*16] BACH B., SHI C., HEULOT N., MADHYASTHA T., GRABOWSKI T., DRAGICEVIC P.: Time curves: Folding time to visualize patterns of temporal evolution in data. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (Jan. 2016), 559–568. doi:10.1109/TVCG.2015.2467851. 15

[Che76] CHEN P. P.-S.: The entity-relationship model—toward a unified view of data. *ACM Trans. Database Syst. 1*, 1 (Mar. 1976), 9–36. 8

[Chi00] CHI E. H.: A taxonomy of visualization techniques using the data state reference model. In *IEEE Symposium on Information Visualization (InfoVis)* (2000), pp. 69–75. 6

[CJ10] CHEN M., JAENICKE H.: An information-theoretic framework for visualization. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (Nov 2010), 1206–1215. 6

[CMS99] CARD S., MACKINLAY J., SHNEIDERMAN B.: *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers, San Francisco, 1999. 4, 6, 8, 22

[Coo94] COOKE N. J.: Varieties of knowledge elicitation techniques. *Int. J. Hum.-Comput. Stud. 41*, 6 (Dec. 1994), 801–849. 18

[CR98] CHI E. H., RIEDL J. T.: An operator interaction framework for visualization systems. In *Proceedings IEEE Symposium on Information Visualization* (1998), pp. 63–70. 6

[dOL03] DE OLIVEIRA M. C. F., LEVKOWITZ H.: From visual data exploration to visual data mining: a survey. *IEEE Transactions on Visualization and Computer Graphics 9*, 3 (July 2003), 378–394. 6

[Edi11] EDITORS OF THE AMERICAN HERITAGE DICTIONARIES: *The American Heritage Dictionary of the English Language, Fifth Edition*. Houghton Mifflin Harcourt Trade, 2011. 2

[EHRLW18] EDGE D., HENRY RICHE N., LARSON J., WHITE C.: Beyond tasks: An activity typology for visual analytics. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 267–277. 6

[EKHW08] ECCLES R., KAPLER T., HARPER R., WRIGHT W.: Stories in GeoTime. *Information Visualization 7*, 1 (Oct 2008), 3–17. 18

[ELP*16] ETEMADPOUR R., LINSEN L., PAIVA J. G., CRICK C., FORBES A. G.: *Choosing Visualization Techniques for Multidimensional Data Projection Tasks: A Guideline with Examples*. Springer International Publishing, Cham, 2016, pp. 166–186. doi:10.1007/978-3-319-29971-6_9. 15

[ENP*09] ELER D. M., NAKAZAKI M. Y., PAULOVICH F. V., SANTOS D. P., ANDERY G. F., OLIVEIRA M. C. F., BATISTA NETO J., MINGHIM R.: Visual analysis of image collections. *The Visual Computer 25*, 10 (Oct 2009), 923–937. doi:10.1007/s00371-009-0368-7. 15

[ERT*17] ENDERT A., RIBARSKY W., TURKAY C., WONG B. W., NABNEY I., BLANCO I. D., ROSSI F.: The state of the

art in integrating machine learning into visual analytics. *Computer Graphics Forum* (2017), n/a–n/a. doi:10.1111/cgf.13092. 13

[FPSS96a] FAYYAD U., PIATETSKY-SHAPIRO G., SMYTH P.: From data mining to knowledge discovery in databases. *AI magazine 17*, 3 (1996), 37. 4, 7

[FPSS96b] FAYYAD U., PIATETSKY-SHAPIRO G., SMYTH P.: Knowledge discovery and data mining: Towards a unifying framework. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (1996), KDD'96, AAAI Press, pp. 82–88. 4

[FWR*17] FEDERICO P., WAGNER M., RIND A., AMOR-AMORÓS A., MIKSCH S., AIGNER W.: The role of explicit knowledge: A conceptual model of knowledge-assisted visual analytics. In *Proc. IEEE Conference on Visual Analytics Science and Technology (VAST)* (2017), IEEE. In press. 6, 21

[Gle13] GLEICHER M.: Explainers: Expert explorations with crafted projections. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec 2013), 2042–2051. 16

[GRF09] GREEN T., RIBARSKY W., FISHER B.: Building and Applying a Human Cognition Model for Visual Analytics. *Information Visualization 8*, 1 (2009), 1–13. 5, 7, 21

[GWLN11] GRINSTEIN G., WHITING M., LIGGETT K., NEBESH D.: IEEE VAST Challenge 2011. http://hcil.cs.umd.edu/localphp/hcil/vast11/, last accessed 01/29/2014, 2011. 3, 4

[GWM*15] GU Y., WANG C., MA J., NEMIROFF R. J., KAO D. L.: iGraph: a graph-based technique for visual analytics of image and text collections. In *IS&T Electronic Imaging Conference on Visualization and Data Analysis* (2015), vol. 9397, pp. 939708–1–939708–15. 14

[GZ09] GOTZ D., ZHOU M. X.: Characterizing users' visual analytic activity for insight provenance. *Information Visualization 8*, 1 (2009), 42–55. 7, 18, 19

[GZA06] GOTZ D., ZHOU M. X., AGGARWAL V.: Interactive visual synthesis of analytic knowledge. In *2006 IEEE Symposium On Visual Analytics Science And Technology* (Oct 2006), pp. 51–58. 18

[Han94] HAND D. J.: Deconstructing statistical questions. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 157*, 3 (1994), 317–356. 5

[HHHW13] HÖFERLIN M., HÖFERLIN B., HEIDEMANN G., WEISKOPF D.: Interactive schematic summaries for faceted exploration of surveillance video. *IEEE Transactions on Multimedia 15*, 4 (June 2013), 908–920. 15

[HMJ*12] HAO M. C., MARWAH M., JANETZKO H., DAYAL U., KEIM D. A., PATNAIK D., RAMAKRISHNAN N., SHARMA R. K.: Visual exploration of frequent patterns in multivariate time series. *Information Visualization 11*, 1 (Jan. 2012), 71–83. 15

[HVW09] HEER J., VIÉGAS F., WATTENBERG M.: Voyagers and Voyeurs: Supporting Asynchronous Collaborative Visualization. *Communications of the ACM 52*, 1 (2009), 87–97. 18

[KAF*08] KEIM D., ANDRIENKO G., FEKETE J.-D., GÖRG C., KOHLHAMMER J., MELANÇON G.: Visual analytics: Definition, process, and challenges. In *Information Visualization: Human-Centered Issues and Perspectives*, Kerren A., Stasko J. T., Fekete J.-D., North C., (Eds.). Springer, Berlin, 2008, pp. 154–175. 1, 2, 5, 6, 7, 8

[KKEM10] KEIM D. A., KOHLHAMMER J., ELLIS G., MANSMANN F. (Eds.): *Mastering The Information Age – Solving Problems with Visual Analytics*. Eurographics Association, Goslar, 2010. 1, 2, 5, 6

[Kli85] KLIR G.: *Architecture of Systems Problem Solving*. Springer Dordrecht, 1985. 8

[KWS*14] KONEV A., WASER J., SADRANSKY B., CORNEL D., PERDIGAO R. A., HORVÁTH Z., GRÖLLER E.: Run watchers: Automatic simulation-based decision support in flood management. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1873–1882. 19

[LAB*11] LAMMARSCH T., AIGNER W., BERTONE A., MIKSCH S., RIND A.: Towards a concept how the structure of time can support the visual analytics process. In *Proceedings of the Second International Workshop on Visual Analytics held in Europe (EuroVA 2011)* (2011), Miksch S., Santucci G., (Eds.), Eurographics Publications, pp. 9–12. 5, 22

[LS10] LIU Z., STASKO J. T.: Mental models, visual reasoning and interaction in information visualization: A top-down perspective. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 999–1008. 7, 21

[MA14] MIKSCH S., AIGNER W.: A matter of time: Applying a data-users-tasks design triangle to visual analytics of time-oriented data. *Computers & Graphics 38* (2014), 286–290. 6

[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. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1803–1812. 14

[MLL*13] MONROE M., LAN R., LEE H., PLAISANT C., SHNEIDERMAN B.: Temporal event sequence simplification. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2227–2236. 14

[MP13] MÜHLBACHER T., PIRINGER H.: A partition-based framework for building and validating regression models. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec 2013), 1962–1971. 16, 17

[MPG*14] MÜHLBACHER T., PIRINGER H., GRATZL S., SEDLMAIR M., STREIT M.: Opening the black box: Strategies for increased user involvement in existing algorithm implementations. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec 2014), 1643–1652. 6

[MW12] MIGUT G., WORRING M.: Visual exploration of classification models for various data types in risk assessment. *Information Visualization 11*, 3 (2012), 237–251. 16

[OSSK12] OELKE D., SPRETKE D., STOFFEL A., KEIM D. A.: Visual readability analysis: How to make your writings easier to read. *IEEE Transactions on Visualization and Computer Graphics 18*, 5 (2012), 662–674. 14

[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). 5, 22

[PNML08] PAULOVICH F. V., NONATO L. G., MINGHIM R., LEVKOWITZ H.: Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics 14*, 3 (May 2008), 564–575. doi:10.1109/TVCG.2007.70443. 15

[R C17] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL: https://www.R-project.org. 16

[RAW*16] RIND A., AIGNER W., WAGNER M., MIKSCH S., LAMMARSCH T.: Task cube: A three-dimensional conceptual space of user tasks in visualization design and evaluation. *Information Visualization 15*, 4 (2016), 288–300. 5, 7

[RF16] RIBARSKY W., FISHER B.: The human-computer system: Towards an operational model for problem solving. In *Proc. Hawaii Int. Conf. on System Sciences (HICSS)* (Jan. 2016), pp. 1446–1455. 5, 7, 18, 21, 22

[RFP09] RIBARSKY W., FISHER B., POTTENGER W. M.: Science of Analytical Reasoning. *Information Visualization 8* (2009), 254–262. 7, 21

[Rot13] ROTH R. E.: An empirically-derived taxonomy of interaction primitives for interactive cartography and geovisualization. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2356–2365. 6

[SBF98] STUDER R., BENJAMINS V., FENSEL D.: Knowledge engineering: Principles and methods. *Data & Knowledge Engineering 25*, 1 (1998), 161–197. 18

[SBM92] SPRINGMEYER R. R., BLATTNER M. M., MAX N. L.: A characterization of the scientific data analysis process. In *Proceedings IEEE Conference on Visualization* (1992), pp. 235–242. 6

[SF08] SKUPIN A., FABRIKANT S. I.: *Spatialization*. Blackwell Publishing Ltd, 2008, pp. 61–79. doi:10.1002/9780470690819.ch4. 15

[SHB*14] SEDLMAIR M., HEINZL C., BRUCKNER S., PIRINGER H., MÖLLER T.: Visual parameter space analysis: A conceptual framework. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 2161–2170. 7, 13, 19

[SHJ*15] STEIN M., HÄUSSLER J., JÄCKLE D., JANETZKO H., SCHRECK T., KEIM D. A.: Visual soccer analytics: Understanding the characteristics of collective team movement based on feature-driven analysis and abstraction. *ISPRS International Journal of Geo-Information 4*, 4 (2015), 2159. 16

[Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings IEEE Symposium on Visual Languages* (1996), pp. 336–343. 6, 21

[Sim96] SIMON H. A.: *The Sciences of the Artificial (3rd Ed.)*. MIT Press, Cambridge, MA, USA, 1996. 4

[SNHS13] SCHULZ H.-J., NOCKE T., HEITZLER M., SCHUMANN H.: A design space of visualization tasks. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2366–2375. 7

[Spe01] SPENCE R.: *Information visualization*. ACM Press books. Addison-Wesley, Harlow, England, 2001. 4

[Spe07] SPENCE R.: *Information Visualization: Design for Interaction (2nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2007. 4

[SPG14] STOLPER C. D., PERER A., GOTZ D.: Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec 2014), 1653–1662. 6

[SS04] SEO J., SHNEIDERMAN B.: A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *IEEE Symposium on Information Visualization* (2004), pp. 65–72. 15

[SSS*14] SACHA D., STOFFEL A., STOFFEL F., KWON BUM CHUL K., ELLIS G., KEIM D. A.: Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec. 2014), 1604–1613. 1, 3, 5, 7, 8, 21, 22

[SvW08] SHRINIVASAN Y., VAN WIJK J.: Supporting the Analytical Reasoning Process in Information Visualization. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2008), ACM New York, NY, USA, pp. 1237–1246. 18

[TC05] THOMAS J., COOK K.: *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE, 2005. 1, 2, 5, 6, 7, 8, 18, 21, 22

[TM04] TORY M., MÖLLER T.: Rethinking visualization: A high-level taxonomy. In *Proceedings IEEE Symposium on Information Visualization (INFOVIS 2004)* (2004), pp. 151–158. 7, 8

[TMN03] TEJADA E., MINGHIM R., NONATO L. G.: On improved projection techniques to support visual exploration of multi-dimensional data sets. *Information Visualization 2*, 4 (2003), 218–231. doi:10.1057/palgrave.ivs.9500054. 15

[vdEHBvW16] VAN DEN ELZEN S., HOLTEN D., BLAAS J., VAN WIJK J. J.: Reducing snapshots to points: A visual analytics approach to dynamic network exploration. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (Jan 2016), 1–10. doi:10.1109/TVCG.2015.2468078. 15

[VJC09] VROTSOU K., JOHANSSON J., COOPER M.: Activitree: interactive visual exploration of sequences in event-based data using graph similarity. *IEEE Transactions on Visualization and Computer Graphics 15* (2009), 945–952. 14

[VW05] VAN WIJK J. J.: The value of visualization. In *Proc. IEEE Visualization* (2005), pp. 79–86. 5, 6

[vWvS99] VAN WIJK J. J., VAN SELOW E. R.: Cluster and Calendar Based Visualization of Time Series Data. In *Proceedings of the IEEE Symposium on Information Visualization 1999 (InfoVis99)* (1999), pp. 4–9. 15

[WAG05] WILKINSON L., ANAND A., GROSSMAN R.: Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* (Oct 2005), pp. 157–164. 15

[WBJ16] WALKER J., BORGO R., JONES M. W.: Timenotes: A study on effective chart visualization and interaction techniques for time-series data. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (Jan 2016), 549–558. doi:10.1109/TVCG.2015.2467751. 17, 18

[WFR*10] WASER J., FUCHS R., RIBICIC H., SCHINDLER B., BLOSCHL G., GROLLER E.: World lines. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (Nov 2010), 1458–1467. 19

[WGK10] WARD M., GRINSTEIN G., KEIM D.: *Interactive Data Visualization: Foundations, Techniques, and Applications*. A. K. Peters, Ltd., Natick, MA, USA, 2010. 4

[WLY*14] WU Y., LIU S., YAN K., LIU M., WU F.: Opinionflow: Visual analysis of opinion diffusion on social media. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec 2014), 1763–1772. 14

[WM17] WANG J., MUELLER K.: Visual causality analysis made practical. In *Proc. IEEE Conference on Visual Analytics Science and Technology (VAST)* (2017), IEEE. In press. 17, 18

[WSP*06] WRIGHT W., SCHROH D., PROULX P., SKABURSKIS A., CORT B.: The sandbox for analysis: Concepts and methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2006), CHI '06, ACM, pp. 801–810. 18

[WW08] WILKINSON L., WILLS G.: Scagnostics distributions. *Journal of Computational and Graphical Statistics 17*, 2 (2008), 473–491. 15

[ZGI*18] ZHAO J., GLUECK M., ISENBERG P., CHEVALIER F., KHAN A.: Supporting handoff in asynchronous collaborative sensemaking using knowledge-transfer graphs. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 340–350. doi:10.1109/TVCG.2017.2745279. 18