Wang XM, Zhang TY, Ma YX *et al.* A survey of visual analytic pipelines. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 31(4): 787–804 July 2016. DOI 10.1007/s11390-016-1663-1

A Survey of Visual Analytic Pipelines

Xu-Meng Wang, Tian-Ye Zhang, Yu-Xin Ma, Jing Xia, and Wei Chen *, *Senior Member, IEEE*

State Key Laboratory of Computer Aided Design and Computer Graphics, Zhejiang University, Hangzhou 310058, China
Innovation Joint Research Center for Cyber-Physical-Society System, Zhejiang University, Hangzhou 310058, China

E-mail: {wangxumeng, zhangtianye1026, mayuxin}@zju.edu.cn; jjane.summer@gmail.com
chenwei@cad.zju.edu.cn

Received April 17, 2016; revised May 31, 2016.

Abstract Visual analytics has been widely studied in the past decade. One key to make visual analytics practical for both research and industrial applications is the appropriate definition and implementation of the visual analytics pipeline which provides effective abstractions for designing and implementing visual analytics systems. In this paper we review the previous work on visual analytics pipelines and individual modules from multiple perspectives: data, visualization, model and knowledge. In each module we discuss various representations and descriptions of pipelines inside the module, and compare the commonalities and the differences among them.

Keywords visual analytics, pipeline, visualization, model, knowledge

1 Introduction

Nowadays the increasing availability of massive datasets has raised a revolution of data gathering, storage and analysis. It becomes difficult and gradually infeasible to apply standard tools for data analysis, which are widely utilized during the past decades by business analysts, scientists and government employees for insight gaining and decision making.

In many fields such as biological computation, business intelligence and online transaction analysis, automated data analysis approaches such as machine learning and data mining are commonly deployed to extract patterns from existing data. The patterns are represented as the high-level abstraction of insights from the data and then transformed into knowledge[1]. Visualization, from the perspective of human vision, provides another scheme for analysts to enhance the ability of understanding and exploring datasets. Usually visualization methods employ visual channels to represent and transform raw datasets into various visual representation forms, and thereby human intelligence is incorporated into the data analysis process via intuitive interactive interface. In the past decade, the theory of "visual analytics" (or visual analysis) has been widely studied by combining automated data mining techniques and visualization methods. Visual analytics "integrates the capability of computer and the abilities of the human analyst"[2] to empower the control of the entire analysis and decision-making process. In fact, pioneers provide a few valuable literatures. Keim *et al.*[2-3] gave a general introduction of visual analytics. In addition, some scholars summarize the state-of-the-art part in the field of visual analytics, for instance, Zhang *et al.*[4] focused on advanced commercial systems and Sun *et al.*[5] generalized cutting-edge research and future challenges from the perspective of analytics space.

The purpose of this paper is to bring visual analytics into the limelight. We review a set of literatures on visual analytics and propose a summarization of visual analytics pipelines that cover automated data process-

Survey

The work was supported by the National Basic Research 973 Program of China under Grant No. 2015CB352503, the Major Program of National Natural Science Foundation of China under Grant No. 61232012, the National Natural Science Foundation of China under Grant Nos. 61422211, u1536118, and u1536119, Zhejiang Provincial Natural Science Foundation of China under Grant No. LR13F020001, and Fundamental Research Funds for the Central Universities of China.

*Corresponding Author

ing, visualization and human interactions. The main methodology in this survey is a categorization of multiple levels that are included in the visual analytics loop which is considered as a global picture of the entire framework, and the detailed analysis of each module to present specific techniques in the processing steps. Furthermore, for specific modules in the visual analytics loop, we compare several different representations of pipelines proposed in history, especially the entire evolutionary process and the shared steps exist in all pipelines. The comparison is specifically designed to clarify the commonalities and differences among multiple pipeline presentations.

The paper is organized as follows. Section 2 introduces a conventional visual analytics pipeline that is widely adopted in the community. Sections 3∼6 expand the modules described in Fig.1. Finally, we give a conclusion in Section 7.

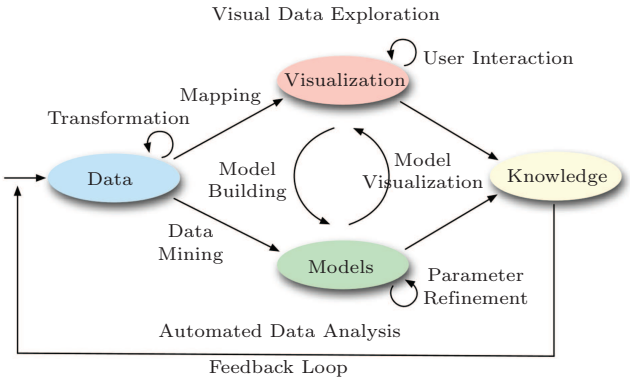


Fig.1. Visual analytics pipeline proposed by Keim *et al.*[2]

2 Conventional Visual Analytics Pipeline

The visualization pipeline illustrates the process of transforming data to visual representations, which comprises the key structure extracted from massive visualization systems[6]. A visual analytic process emphasizes on analytical reasoning as well as decision making with interactive visual interfaces. Countless research studies in the field of visual analytics have been performed, and most of them follow a conventional visual analytics pipeline presented by Keim *et al.*[2] as shown in Fig.1. We describe several other pipelines in Sections 3∼6 about their focuses. The representative pipeline from Keim *et al.* is introduced in this section. This conventional pipeline guides visual analytics processes as an abstract outline, including four major stages and significant relationships between them.

Beginning with raw data, either structured or unstructured, it is essential to employ a series of data pre-processing steps, like data transformation. Raw data may contain a variety of errors and invalid items. The data pre-process procedure is used not only to remove the redundancy, errors and invalid items of the raw data, but also to provide analysts with data in the exact form they need.

After the data pre-process procedure, both visual exploration and automatic analysis methods are available. Visual data exploration methods offer an interactive visual interface to display transformed data in a visual mapping way, while automatic analysis methods require analysts to generate or select appropriate models based on data features, by applying different data mining methods. However, models generated in the automatic analysis process may contain their own drawbacks and therefore need to be validated and refined. Fortunately, visualization allows analysts to participate in the model generation and modification process by refining parameters or selecting another model. Actually, many analysis missions are too complex for computers to complete on their own, making the position of visualization irreplaceable. As we can figure out, visualization is not only expected to help the model building process with generated hypotheses and insights, but also expected to evaluate the results and the findings of models by model visualization.

It is natural to draw the conclusion that analysts are able to gain knowledge using visual data exploration or automatic analysis methods, or even a combination suite of both of them. Knowledge generation, however, does not necessarily put an end to this conventional visual analytics pipeline. For those complicated visual analytics missions, the knowledge gained at the first time may not be adequate for the analysts and thus it is fundamental to refine the data pre-process methods in order to carry out the whole process again and again to receive satisfying results. This is the way in which the feedback loop works.

As previously stated, this conventional pipeline only shows an abstract overview of the entire visual analytics process which contains more specific steps in each stage. Based on Keim *et al.*'s work[2] and more recent literatures, we drill down to each specific stage so as to give a more detailed and practical description of the conventional pipeline.

3 Data

Everyday, massive raw data is collected through sensors, experiments, questionnaires, network, etc. With the increasing variety of available data and the facilitation of data collection, the application domain of visual analysis keeps expanding. We collect some literatures and classify them by data category. As a matter of fact, a majority of visual analytic systems are task-driven[7-9]. However, tasks are specific and distinctive from each system in most instances, thus classifying visual analytic systems depending on tasks is quite challenging. Rather than classifying methods with tasks, data-based classification[10] can not only be general but also reflect the applications[5], which is shown in Table 1 and the related applications are listed in the last column. With the obtained raw data, data pre-processing needs to be done. Routinely, there are many differences between the raw data and the data used in the visualization module and the model module in that raw data could be incomplete, noisy or inconsistent[52]. In order to eliminate these differences and meet the needs for the next steps, some procedures called data pre-processing have to be executed. Data pre-processing is a flexible process, which depends on the raw data. The common data pre-processing procedures comprise of[53]:

Data Integration: combine data from different sources based on a global schema and provide a uniform interface of these data[54-55].

Data Cleaning: detect the data quality problems by data profiling and data mining, and resolve them[56].

Data Transformation: convert data from one format to another format. A representative transformation is the data normalization.

Data Reduction: remove useless data to improve the efficiency of operations.

In addition, some analysis missions involve several heterogeneous data sources or confidential data. For the purposes of facilitation and security, they also need a systematic and effective data management.

4 Visualization

Human eyes can not only quickly accept a huge amount of visual signals, but also process information in parallel[57]. Taking advantage of visual channels, visualization accelerates the human acceptance of information and enhances the efficiency of the analysis. As shown in Fig.2, a variety of studies about visualization pipeline have been done. In this section, we compare four typical visualization pipelines in Subsection 4.1. We give the specific pipeline of the visualization module in accordance with the common parts of those pipelines and some practical examples of visualization, as shown in Fig.3. The description of our visualization pipeline can be found in Subsection 4.2.

4.1 Typical Visualization Pipelines and Comparison

Fig.2(a) is an overview of data state model proposed by Chi and Riedl in 1998[58]. Its taxonomy can be found in [59]. The process contains four stages (dark parallelograms in Fig.2(a)). The operator for each stage represents a kind of interaction. When the results obtained are unsatisfactory or not able to meet the needs of the next stages, visualization can provide users chances to interactively re-select parameters or approaches and re-run this stage. Meanwhile, all steps between two stages are regarded as a kind of transformation. In the entire process, visualization transformation connects two parts: the data space controlled by system (the half above the dotted line) and the view space controlled by users (the half below the dotted line) via converting analytical abstraction like metadata into visualizable information, and visualization abstraction.

Table 1. Data Classification and Related Applications

Data Type	Data	Main Application
One-dimensional data	Signal record data[11], comments amount[12]	Trend research, prediction
Two-dimensional data	Vector data[13-14]	Physical
Multi-dimensional data	Movement data and trajectory data[15-20], utility services[9], mobility data[21], boundary changes[22], social media[23], personal data[24-28], car data[29], movie genres[30], OECD countries data[31], climate data[32], weather data[33], resource data[34-36], eye tracking data[37]	Environmental protection, behavior analysis, city planning, evaluation and intelligence analysis
Text data	Document data[38-39], news[40-42], wikipedia articles[43], poem[44], literature, dictionary[45], opinion[46]	Sociology, journalism, literature
Networks	Social network data[47], biological data[48-49], molecular structure[50], network[51]	Supervision, psychology, biochemistry, sociology

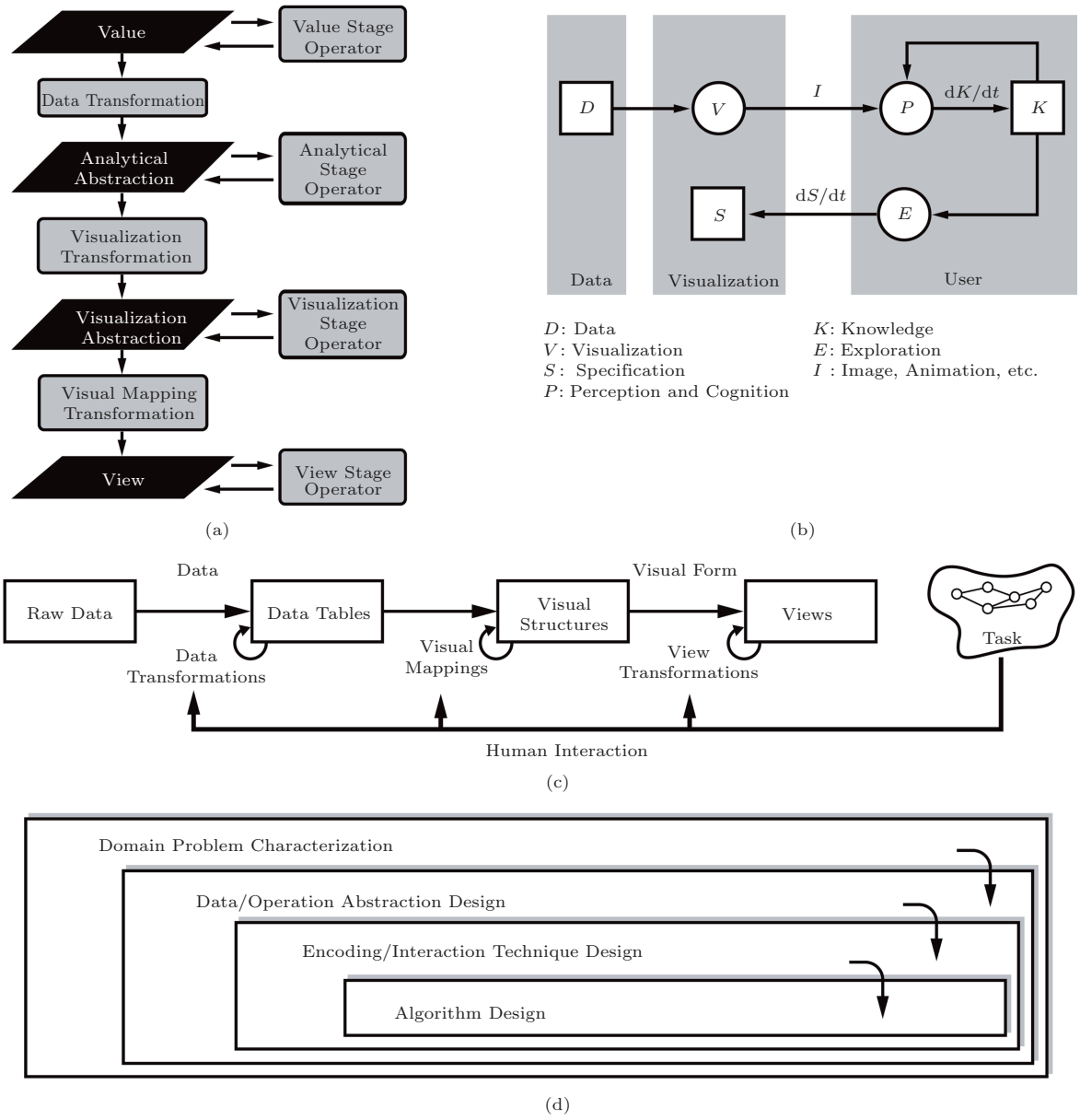


Fig.2. Typical visualization pipelines. (a) Information visualization data state reference model[58-59] proposed by Chi and Riedl in 2000. (b) Generic visualization model[61] proposed by Van Wijk in 2005. (c) Reference model[60] presented by Card *et al.* in 1999. (d) Nested model of visualization creation[62] presented by Munzner in 2009.

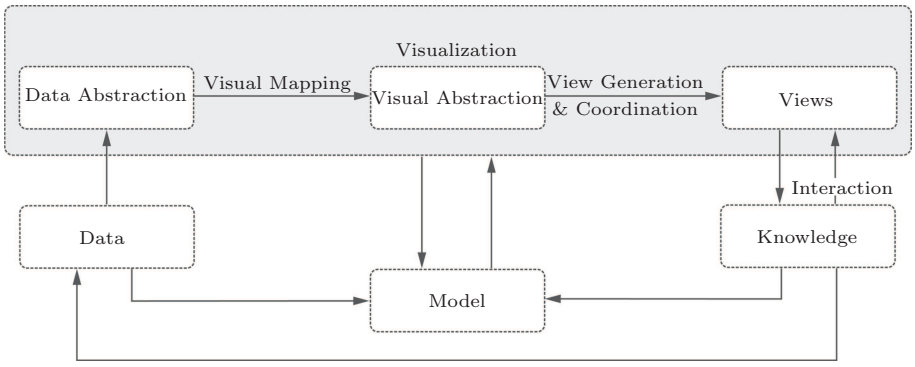


Fig.3. Specific visualization pipeline.

The pipeline from Card *et al.*[60] has two parts likewise (see Fig.2(c)). One part is data and the other part is visual form. The main difference between these two pipelines is that the one in [60] emphasizes on human interactions. From Fig.2(b), we can see that all the interactions are carried out according to the task. Compared with Chi and Riedl's data state model[58-59], this one emphasizes the purposes of the visualization and the role of human.

In 2005, Van Wijk gave a model[61] with precise description using merely mathematical notation. Besides, there are two circulations in the model. They represent the output and the input of visualization respectively. The circulation of $P \rightarrow K \rightarrow P$ is about knowledge accumulation, and the circulation of $V \rightarrow P \rightarrow K \rightarrow E \rightarrow S \rightarrow V$ means interactions. Users perceive the information from visualization and convert the information into knowledge. If the knowledge is not adequate, users will explore more based on knowledge acquired through changing the specification including hardware and algorithms. The resultant specification leads to a new visualization, from which the knowledge is gained.

The nested model[62] in Fig.2(d) was presented by Munzner in 2009 by taking a different perspective to visualization design and creation. The first stage of this model is different from the above ones. This process starts from domain problem characterization rather than data. Designers must clearly understand the system requirements including tasks and data, and every other step is based on the output of the previous one. Although a nested model implies a design order, the temporal sequence is not always carried out strictly and the refinements should not be limited in the current stage[62]. When designers have better understanding about a previous step, the process may restart from it, yielding a so-called iterative refinements process.

4.2 Common Steps

Visualization transforms data abstraction to visual abstraction and combines visual abstractions into a group of views. To help human explore those views, visualization responds to human requests. This process is called interaction shown in Fig.3.

4.2.1 Visual Mapping

Visual mapping is a kind of transformation that converts the data abstraction (output of data module) into the visual abstraction. Visual abstraction refers to the elements that are displayed on the screen and convey information to human by the sense of sight. Common visual channels, which can be used to encode information, include position, size, shape, direction, hue, saturation, brightness, etc. A record of data may have many attributes, and similarly, each visual abstraction can have several visual channels as well. The relationship between data attributes and visual channels can be not only one-to-one but also one-to-many. Every visual channel has a threshold value, and only when the difference exceeds the threshold value can most people make a distinction. Some important and accurate attributes should be represented with more than one visual channel in order to allow users to get information easily and clearly. Unfortunately, the total number of visual channels is limited. Besides, using many visual channels can lead to mutual interferences so that users can hardly get meaningful information. Therefore, it is unwise to use too many visual channels simultaneously (for high-dimensional data, data pre-processing always includes dimension reduction). Some recommendations about how to select visual channels can be found in Fig.4.

Visualization scholars have designed numerous classical visual mapping methods, such as parallel coordinates, force-directed graph, chord graph, scatter matrix and so on. Besides, numerous literatures give assessments[64-67] and improvements[68-72] to existing methods.

4.2.2 View Generation and Coordination

Views deliver information between users and the system, and hence they play a significant role in visualization. In this stage, visual abstractions are rendered systematically on the screen. Except for the visual abstractions, menus and specifications like captions and legends can also be found on some views. The function of menus is offering diverse selections of interaction, and thereby users can explore more information on the views. In addition, necessary specifications can introduce visual mapping designs to users directly so that they will not feel strange or confused.

Using a single complex view may be stressful for users to cognize information, and multiple views can be employed in a "divide and conquer" fashion[73]. The system in Fig.5 is an appropriate example to prove that[74]. Thus, it is inevitable to use multiple distinct views when the dataset contains a variety of data or when data is complex. However, the multiple views increase the users' learning costs and dis-

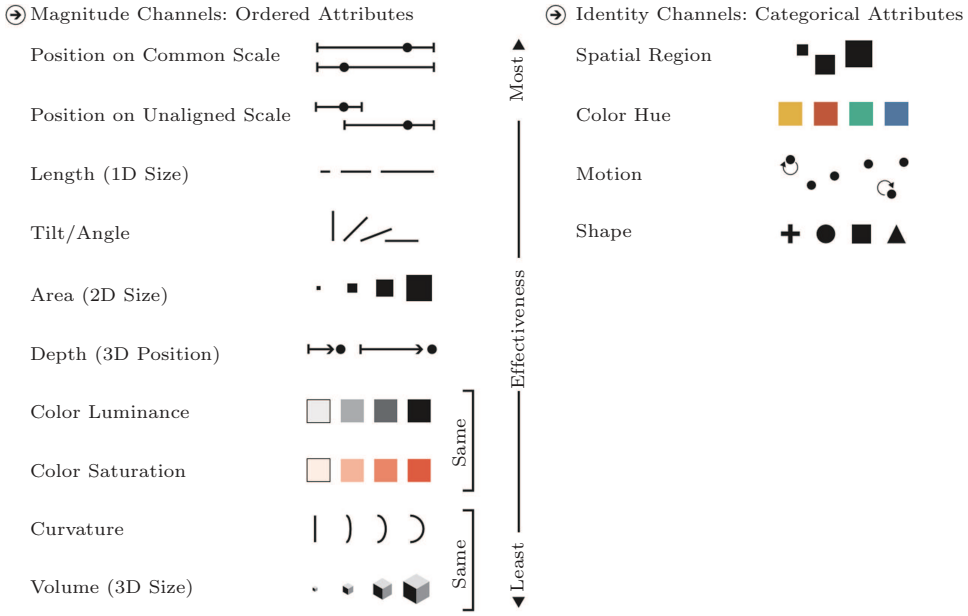


Fig.4. Visual channels: expressiveness types and effectiveness ranks[63].

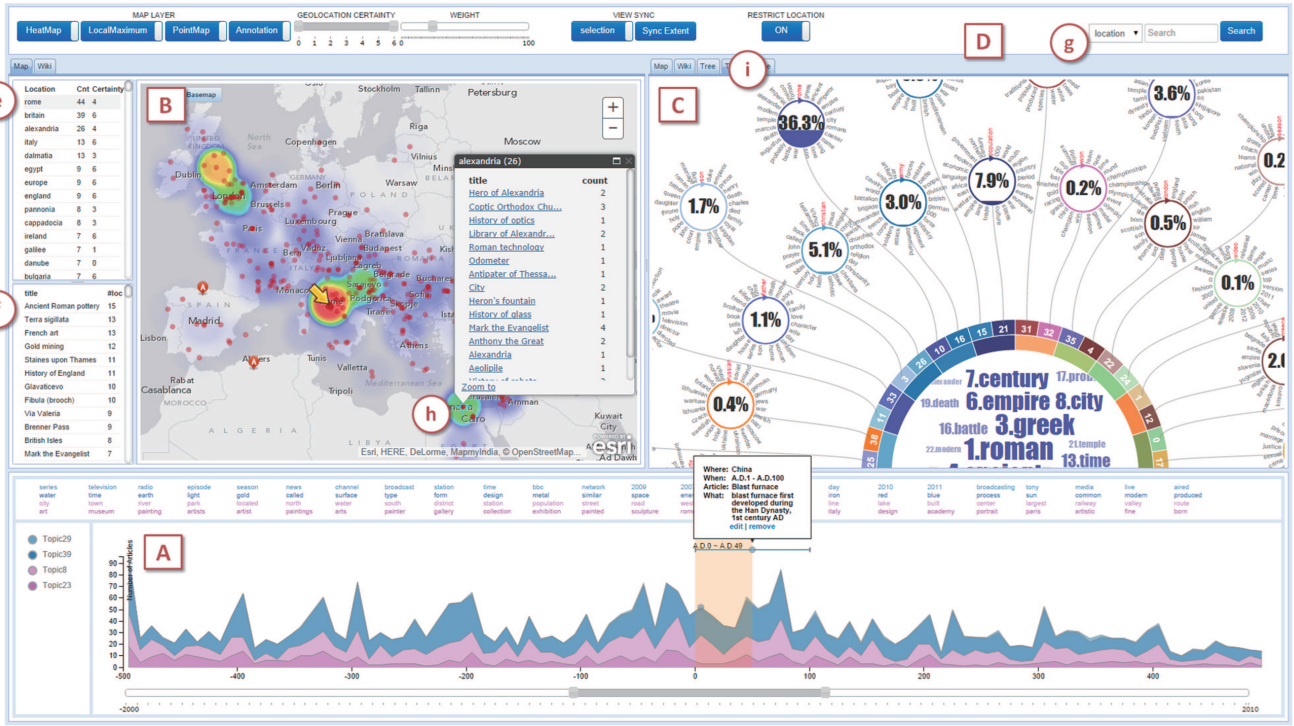


Fig.5. Visualization system with multiple views[74].

tract users' attention, and thus applying multiple views blindly is not desirable. Coordinating views is the key of using multiple views effectively. Above all, designers must clearly understand the relationship between views. Roberts *et al.*[75] presented six variants of side-by-side views. Among them, overview + detail[76-77] and small-multiples[26,36] are frequently used. Then, the problem to be solved is how to respond to the update of other views. The types of responses are summarized as replacements, replications and overlays[78]. Note that the view will discard the original things if replacements occur. Replications often mean that a new

view will be generated and used to show new contents, and overlays add new information to the original contents. Designers usually choose one response from them according to analytic tasks or let users decide which response should be used.

4.2.3 Interaction

Interaction brings vitality to visualization. Without interaction, visual representations are merely some static images[79]. To complete tasks, users need to explore the visualization system based on existing knowledge so that interaction is indispensable to the exploration process. The interaction of visualization is a kind of human-computer interaction. We may consider interaction as a process that a computer responds to the information input by a user. Chuan and Roth[80] classified basic visualization interaction (BVI) by output states, while Yi *et al.*[79] proposed a more detailed classification of interaction from the perspective of users' demands. In the work of Chuah and Roth, the categories comprise graphical operations, data operations and set operations[80]. The three operations affect graphical representations, data state and control state respectively. Actually, graphical representations also change correspondingly when the other two are changed. On the other hand, with the classification of Yi *et al.*, interaction includes selecting, exploring, reconfiguring, encoding, abstracting/elaborating, filtering and connecting[79]. There is no doubt that diverse interaction allows users to operate freely. But interaction designs may also have to pay[81]. For instance, as the amount of interactive selections increases, users need to spend more time on understanding the entire system. Furthermore, too many interactive selections always lead to confusion.

5 Model

Once data are pre-processed or transformed, analysts are supposed to decide to apply either visual methods or automatic analysis methods. Automatic analysis allows analysts to apply data mining methods to generate models[2], and therefore achieves the goal of transferring information from pre-processed data and helping users gain knowledge in distinct areas. Witten and Frank[82] classified machine learning models into eight basic kinds, according to the simple structures exhibited by datasets. For example, statistical models, such as Bayesian models for document classification, are more suitable for the datasets whose attributes might contribute independently and equally to the final outcome, while unsupervised clustering models are used to divide instances into natural groups without providing class values.

There are situations, however, when automation processes are not able to satisfy users. Since the existence of the users' needs to visually understand, explore and optimize the datasets and the computation process, visualization-driven data mining is gaining increasingly attention nowadays[83-86]. Unlike automatic analysis methods mentioned above, it is a semi-automatic method, which combines visualization with data mining in order to create a win-win situation for visual analytics processes. The pattern-searching model presented by Palomo *et al.*[19] is based on interactive visualization, which clearly identifies spatio-temporal patterns for transportation schedules. And Klemm *et al.*[87] employed an adjustable regression model to build a 3D heat map visualization, in order to support hypothesis generation of epidemiological studies.

Most analytics processes, no matter automatic or semi-automatic, are carried out in relatively conventional ways. Many studies have been done to reveal a generalized pipeline, displaying biased results with similar steps shared. We discuss two pipelines and then present a conventional pipeline generalized from them.

5.1 Existing Pipelines

Han *et al.*[88] presented a complete pipeline of the knowledge discovery in database (KDD) process as shown in Fig.6, in which data mining acts as an intermediate step to apply intelligent methods for the purpose of extracting data patterns. The entire framework can be divided into three major parts: data preprocessing, data mining and pattern evaluation, and knowledge presentation. In this subsection, we focus on the data mining and pattern evaluation stage. Han *et al.*[88] suggested that data characterization and discrimination can be useful in data mining, by summarizing and comparing the general features of the target class of data objects. With features characterized, intelligent methods can be applied to reveal the patterns of data. However, patterns generated by a data mining system are not always potentially useful and need to be evaluated by interactively filtering. Han *et al.*[88] noted that the evaluation process ought to be carried out during data mining in order to make the data mining process more sufficient.

A more specific pipeline has been presented by Lu *et al.*[89] targeting at predictive visual analytics as shown

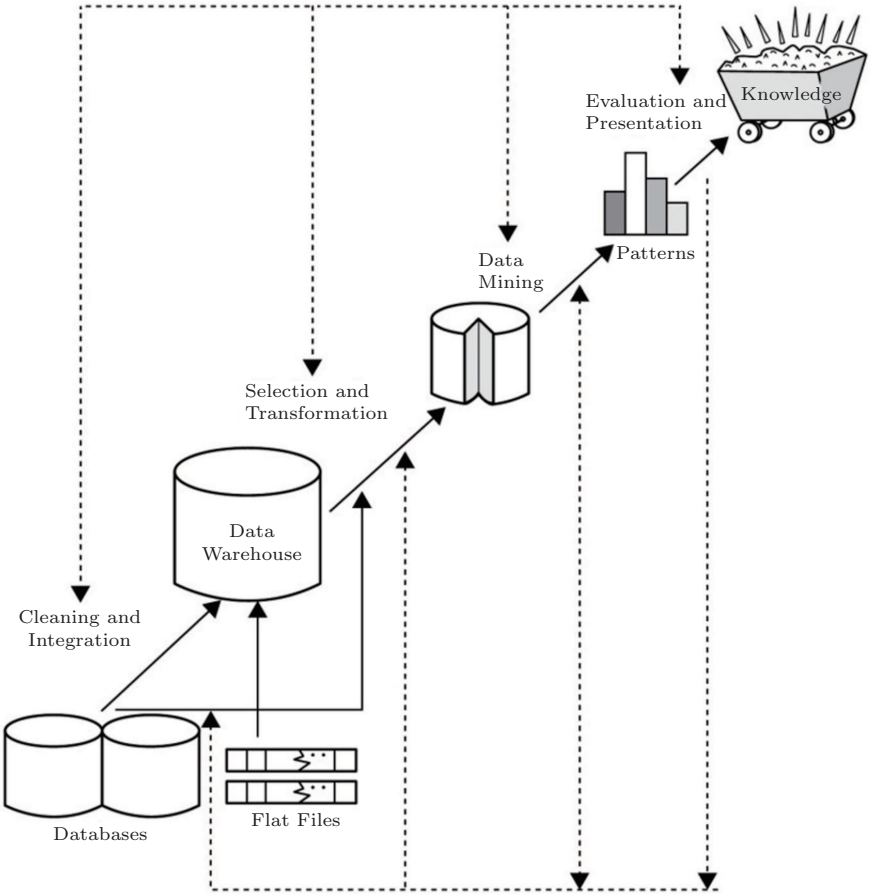


Fig.6. Data mining as a step in the process of knowledge discovery[88].

in Fig.7. Different form Han *et al.*'s work[88], they explained the data mining process in a detailed way by emphasizing the significant roles played by model, visualization and adjustment in the entire pipeline. They are fully convinced of the necessity of model selection, training and validation in knowledge generation when features are selected. Visualization is optional during each of these phases to enhance comprehensibility as well as effectiveness, and the adjustment loop ensures the involvement of human knowledge in all stages of predictive analysis.

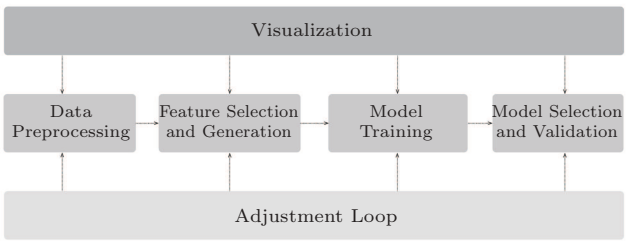


Fig.7. Predictive visual analytics pipeline[89].

Though these two pipelines may seem different in form, they share the same idea which includes data mining in the analytics process and both mention the importance of feature selection and result validation. As a result, it alleviates the difficulty in our work for generalizing a conventional pipeline in analytics processes.

5.2 Generalized Pipeline

In accordance with the pipelines described previously, we present a general pipeline in Fig.8. It mainly consists of three major steps: feature selection and generation, model building and selection, and model validation. Each step is strongly connected with each other.

5.2.1 Feature Selection and Generation

Served as an essential stage in model design, feature selection aims at reducing the cost of recognition by only keeping the most expressive features of data that are used to build models[90]. Dash and Liu[91] gave

a definition of feature selection as an attempt to select the minimally-sized subset of features, by following two criteria: not significantly decreasing the accuracy of classification, and maintaining the original class distribution as much as possible.

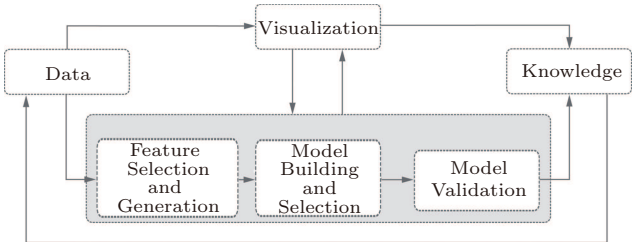


Fig.8. Our proposed model design driven pipeline.

Classical techniques such as clustering, ranking and sorting work well for feature selection. Dy and Brodley[92] introduced a wrapper framework for performing feature subset selection, using EM clustering method. The rank-by-feature framework provided by Seo and Shneiderman[93] applies user-chosen ranking criteria to present low dimensional projections in an ordered manner. The INFUSE system developed by Krause *et al.*[94] eases the feature selection process in a visualization way of grouping and ranking the results of various feature selection algorithms.

Apart from the above classical techniques, analysts have also proposed algorithms for automatic feature generation. These algorithms use limited low-level features to build new features, which significantly increases the classification accuracy[95]. Schuller *et al.*[96] employed feature generation methods to expand the feature space while doing speech emotion recognition researches. In the work of Zahálka and Worring[97], two pipelines for high-level extraction of semantics from multimedia data are described.

Unfortunately, we have to admit that automatic feature selection and generation methods sometimes come up with inevitable subtle mistakes of feature identification or classification. As a result, analysts have been trying to develop integrated feature selection and generation tools that involve both human knowledge and automatic algorithms. Users are expected to interact with the developed tool. Interactions make it possible for users to examine the accuracy as well as the rationality of the results obtained from the computer (and refine them if possible), or even to guide the entire selection and generation process. For example, Lu *et al.*[12] built predictive models of social media data with the implementation of interactive components, which allow users to modify and explore various features in order to receive better box-office predictions. Janetzko *et al.*[98] incorporated a user-configurable classifier to interactively assist the feature selection process when dealing with soccer data, and therefore make it possible to discover further events of potential interests. FeatureInsight, presented by Brooks *et al.*[45], is another example of visual tools designed to support feature selection and generation. Users are enabled to interactively add new features and modify or delete automatically generated ones, which makes the results more interpretable.

5.2.2 Model Building, Selection and Validation

Feature selection and generation identify the most expressive features with which analysts are supposed to generate appropriate models. They attempt to select the most salient model from various existing ones, including statistical models, physical models, data mining models, etc. There are times, however, when existing models are not able to meet the needs of the analysts due to the specificity of their task and they have to build novel models based on existing relevant ones. Once models are selected or built, prediction results are acquired from the trained model, accompanying with validation processes to examine the performance of the generated model.

Indeed, model validation is embedded in model building and selection in most situations, enhanced by interactive visual tools. For example, PEARL is an interactive visual analytic tool that aids users in revealing personal emotion style from social media text data, developed by Zhao *et al.*[99] It adopts two well-known psychological models to capture emotion styles and makes further progress by deploying an improved lexicon-based model in multi-dimensional emotion analytics. Another example is the model created by Kay and Heer[100], used to rank the precision of visualizations for estimating correlation. Taking the model of Harrison *et al.*[101] as a beginning point, they took four more steps to refine it: incorporation of individual differences, log transformation, censoring and Bayesian modeling. And each refinement is presented under the guidance of visual assistance.

Besides the model building methods mentioned above, analysts have also been developing visualization tools for model selection. Bögl *et al.*[102] implemented model selection processes that deal with time series data by using a visualization system called TiMoVA which combines automated computation with human

intelligence. Their work presents a description of visual analytics process for model selection, as shown in Fig.9. Based on input data and prior analyses, users interact with the visual interface to analyze the resulting model using their domain knowledge in order to eventually get a most adequate model for the given dataset. Alexander and Gleicher[7] explored task-centric topic models comparison with a novel visualization called buddy plot for distance comparing. In this way of validating models interactively, users may have a better understanding of the advantages as well as the shortcomings of each topic model when tackling different tasks.

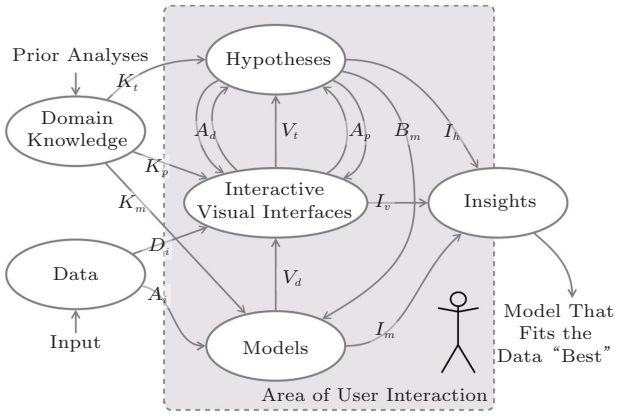


Fig.9. Visual analytics process for model selection presented by Bögl *et al.*[102]

Insights can be gained through the process of model building, selection and validation, and in the meantime they provide the guidance for the improvement of the feature selection and generation process. The looping back makes sure that the entire analysis process is able to refine itself and provide analysts with the most adequate model to gain knowledge.

6 Knowledge

Knowledge is the externalized awareness or understanding of something, which is acquired from and in return applied to the process of knowledge generation. In this paper, we focus on knowledge generation instead of knowledge because it is the ultimate goal of visual analytics. Knowledge generation is the process of generating a conclusion which either accepts or denies the hypothesis. In the literature of visual analytics and cognitive science, alternative terms of knowledge generation include intelligence-gaining, sense-making, decision-making and concept-building[103]. Ware[103] gave two approaches of forming knowledge: the Bayesian approach and the physicalist theory. The Bayesian approach takes knowledge as "repeated associations" or "repeated co-occurrences", and to generate knowledge is to build connections between events or things. On the other hand, the physicalist theory takes knowledge as "the sensory modality of the formative experiences", and the generation process is "an approximate modeling based on everyday physics". Both approaches describe the process as the generalization of findings, either from repeated associations or from everyday physics.

6.1 Knowledge Generation Pipelines

Knowledge can be gained from both computational models mentioned in Section 5 and visual models. In the scope of this paper, we focus on knowledge from visual models. Several pipelines[61,104-107] that describe the knowledge process have been proposed. We analyze two of them in details and briefly introduce others.

6.1.1 Sense-Making Pipeline

The sense-making process[104] (see Fig.10) is a cognitive task analysis diagram including six analytical notions and the transformations among them: external data sources, shoebox, evidence file, schema, hypothesis and presentation. The transformations for data processing construct a foraging loop while the transformations for reasoning construct a sense-making loop. In the foraging loop, the external data sources contain complete and raw data collections and serve as data provenance sources. The shoebox collects processed information from the external data sources. The data in the shoebox is cleaned, organized and ready for use. An example might be an organized (by time, topic, etc.) collection of newspapers. The evidence files are useful fragments extracted from shoebox, such as clippings of news related to a certain topic. The foraging loop describes the flow of step-by-step refinement of raw data to evidence snippets, while the sense-making loop builds concepts based on the evidence snippets. In this loop, the schemas are organized evidences, from which hypotheses are proposed and verified. Finally, a presentation which accepts or denies the hypotheses is delivered to the analysts or to a wider audience. It is also noteworthy that the presentation not only draws the conclusion but also describes the evidences and the reasoning logics that help reach the conclusion.

Knowledge generation with the sense-making model can be carried out via two processes: a bottom-up process and a top-down process. Following the bottom-up

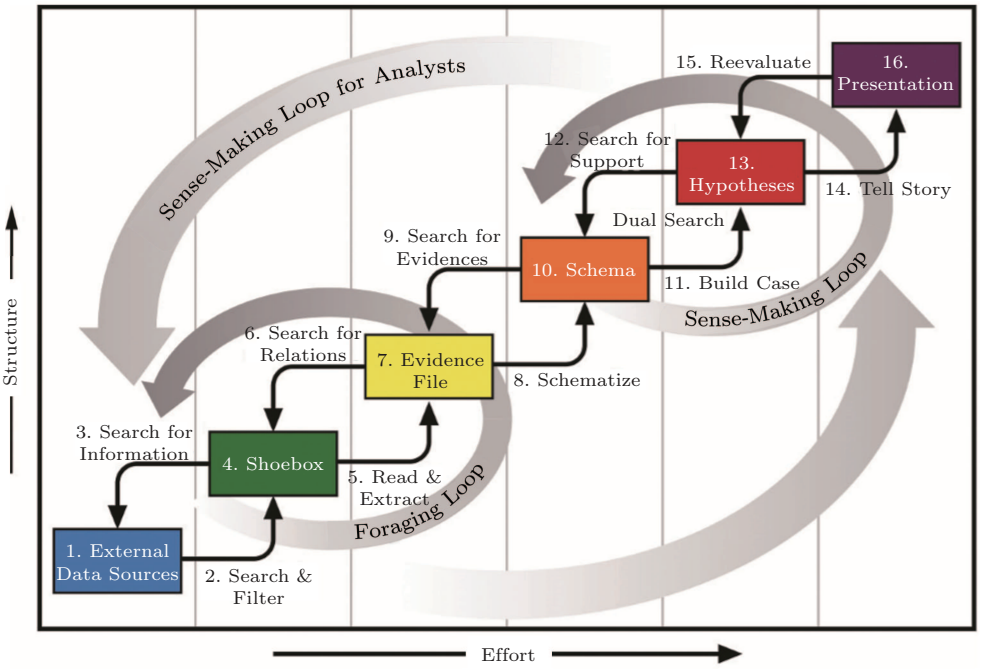


Fig.10. Sense-making model[104].

process, the analysts first put cleaned raw data into a shoebox and extract the evidence from the shoebox. They then organize the evidence into schemas and build a hypothesis. Decisions made for the hypothesis are finally presented. Following the top-down process, the analysts first propose a hypothesis (mostly based on prior knowledge) and search for schemas that support the hypothesis and then the evidences that support the schemas. To consolidate the conclusion, the analysts can look for the data provenance from the shoebox and further the raw data contributes the foundation of reasoning. Although not explicitly described in the paper[104], we think the bottom-up process and the top-down process correspond to the two reasoning principles: inductive reasoning and deductive reasoning, respectively. Similar to the bottom-up process, inductive reasoning builds generalized concepts (conclusions) from cases (evidences). Likewise, similar to the top-down process, deductive reasoning starts with statements (hypotheses) and searches for evidences that either support or oppose the statements.

6.1.2 Knowledge Generation Pipeline

The knowledge generation model for visual analytics[105] (see Fig.11) is composed of two parts: a system part that consists of data, model and visualization; and a human part that consists of action, finding, hypothesis, insight and knowledge. Also starting from data and ending with knowledge, this pipeline describes the knowledge generation process from the visual analytics perspective. Models are often automatic models, but sometimes can also be descrip-

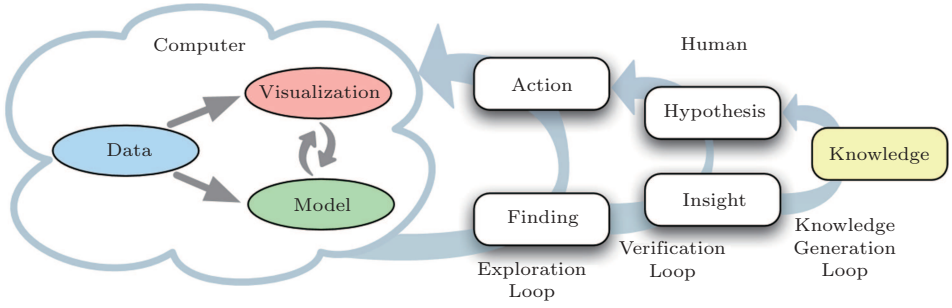


Fig.11. Knowledge generation model[105].

tive statistics, configurable models, or visual models. Visualization is built on the models or built to explain the models, e.g., an open-box visualization of support vector machine[106]. Actions cover all interactions in visualization (referring to Subsection 4.2.3). Findings are the observed snippets or schemas extracted from the visualized models and data. Insights are collected when the analysts are able to interpret the findings, and further form the foundation for hypotheses. Insights are accumulated until it is reliable enough to wrap up as knowledge.

The human part is further decomposed into three loops as shown in Fig.11: the exploration loop, the verification loop and the knowledge generation loop, corresponding to the following three levels of analytical operations respectively. The exploration loop, containing action and finding, describes the process of the analysts manipulating the visualization via interactions. Moreover, the verification loop is accomplished based on the basis of the exploration loop and describes the process of the analysts proposing and verifying hypotheses. Finally, the analysts extract the findings and hypotheses from the verification loop and generate knowledge.

Sacha *et al.* applied and enriched the model to evaluate the role of uncertainty and trust in visual analytics[107]. Uncertainty is propagated through the visualization pipeline in the system part, while trust is built progressively when the analysts explore the system, extract insights from visualization and gain knowledge.

6.1.3 *Other Pipelines*

Other knowledge models[61,108-110] elaborate the knowledge generation process in various perspectives. The human cognition model (see Fig.12) proposed by Green *et al.*[108] describes the process revolving around human discovery. Computer and human collaborate in the model to explore examples and patterns from visualization. Hypotheses are generated and analyzed via the analytical process of competing hypotheses: generating hypotheses — listing evidences — proving/disproving — creating the matrix of hypotheses and evidences — drawing conclusions — reanalyzing conclusion based on evidences. Dykes *et al.*[109] prac-

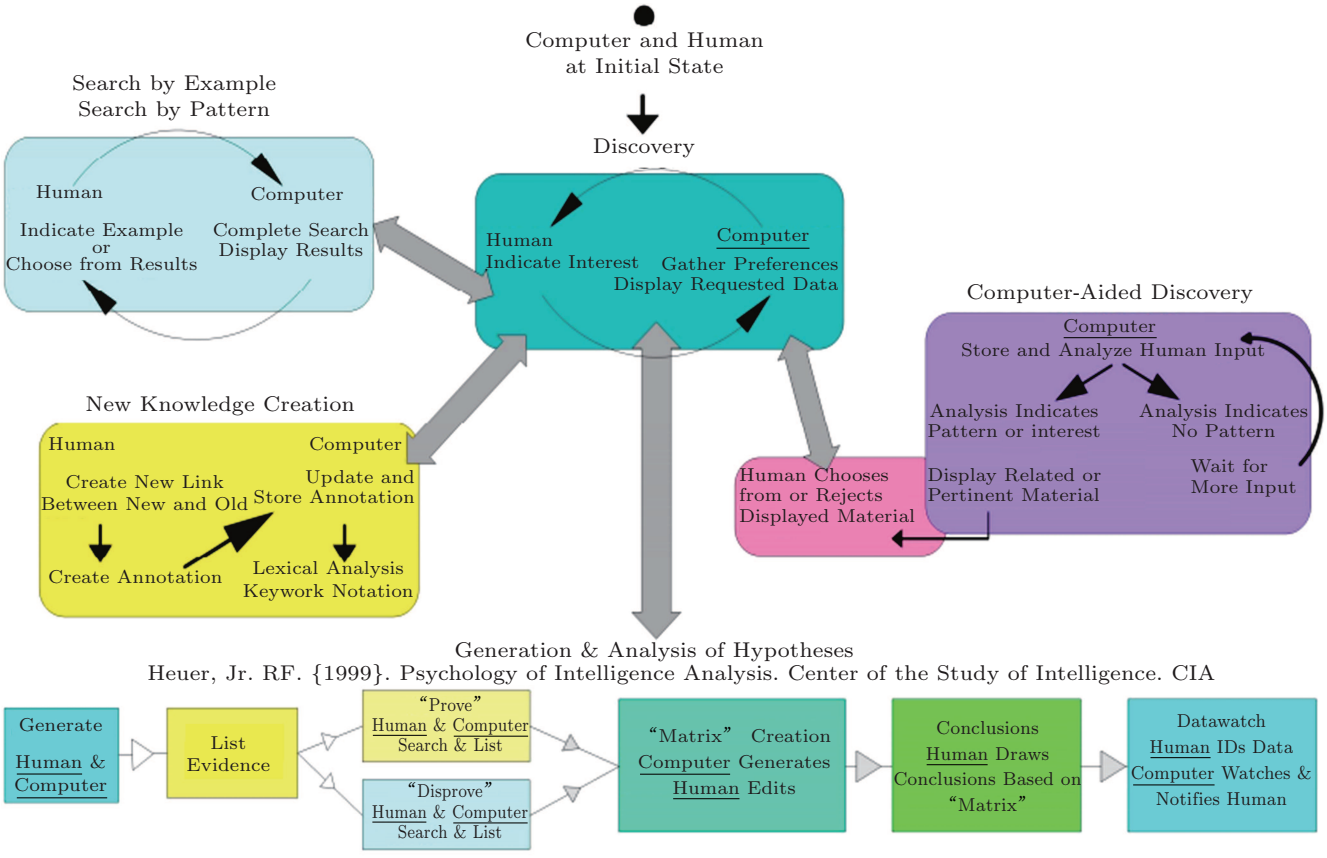


Fig.12. Human cognition model[108]. Underling in the figure indicates the process' initiator.

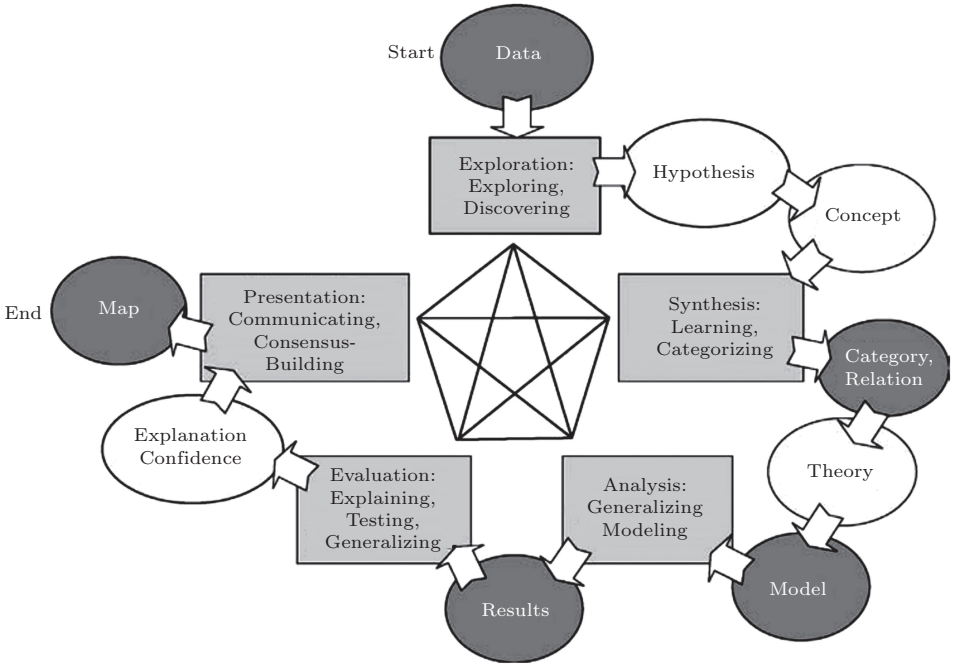


Fig.13. Knowledge generation model[107].

ticed several paths of the nexus of activities comprising the scientific process. Fig.13 depicts a path: data — hypothesis — theory — explanation, linking the core activities (in medium grey). Dark grey entries indicate inputs and outputs, while light grey entries indicate the knowledge provided by the analysts. Van Wijk integrated knowledge generation in his visualization model[61] (see Fig.2). Knowledge (K) is gained through the perception (P) of visualization (V) images, the exploration (E) and specification (S) of visualization (V). The knowledge gained is quantified by the function $\mathrm{d}K/\mathrm{d}t = P(I, K) = P(V(D, S, t), K)$, which indicates that gaining knowledge with visualization over time depends on data (D), algorithm and hardware specification (S) as well as the prior knowledge (K) of the analysts. Klein *et al.*'s data/frame theory of sense-making[110] elaborates the sense-making process with two components: data and frame (see Fig.14). The analysts explore the data with a frame: a metaphor representing the "perspective, viewpoint, or framework". Starting with a frame, the analysts either preserve and elaborate the frame, or decline and rebuild the frame.

6.2 Induction and Deduction

Although there are several knowledge generation pipelines, the reasoning logics behind are the same: new knowledge is transformed from either inductive reasoning or deductive reasoning[111]. In inductive reasoning, the analysts build concepts from observations or schemas. However, the analysts with various prior knowledge would have different concepts based on the same observations or schemas. The bottom-up process in the sense-making model[104], the exploration-verification-knowledge path from the loop in the knowledge generation model[105], the generation of hypotheses in the human cognition model[108], the data to concept to theory path in the knowledge model[109], the exploration to knowledge process in the visualization model[61] and the preserving-elaborating path from the data/frame theory of sense-making[110] are the practices of inductive reasoning. On the other hand, in deductive reasoning, the analysts search for evidences that either confirm or deny the initialized hypotheses. And starting with the same hypotheses, the analysts normally reach the same conclusion. Similarly, the top-down process in the sense-making model[104], the knowledge-verification-exploration path from the loop in the knowledge generation model[105], the analysis of hypotheses process in the human cognition model[108], the knowledge to exploration process in the visualization model[61] and the reframing path of the data/frame theory of sense-making[110] are the practices of deductive reasoning. The two reasoning processes can also be performed in an alternating way. Knowledge generated

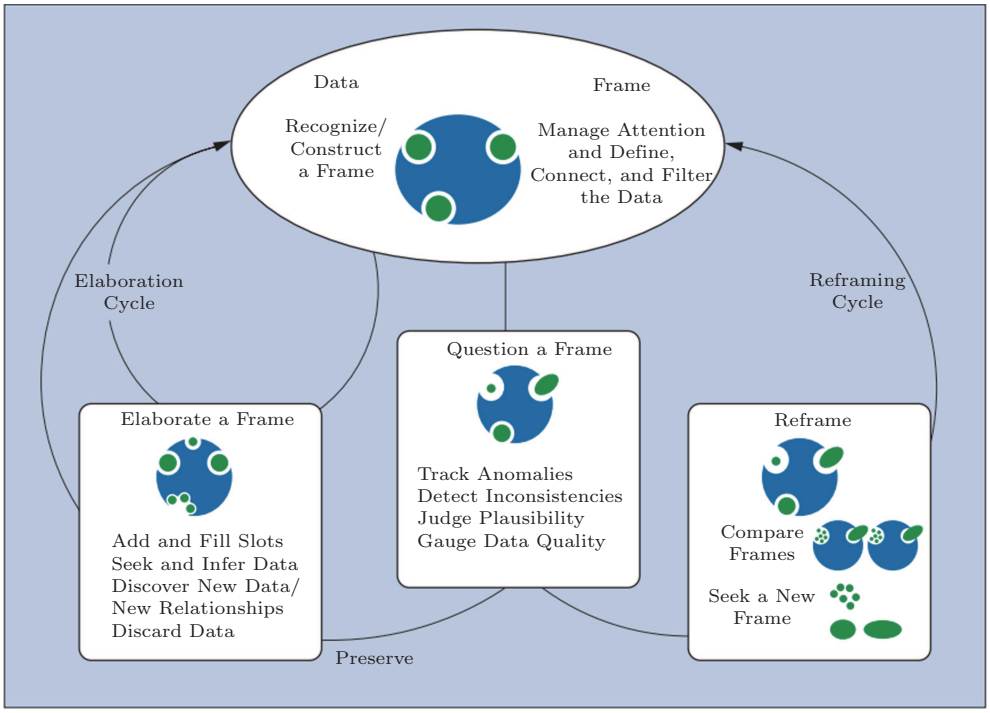


Fig.14. Data/frame theory of sense-making[110].

in the inductive reasoning is applied to the deductive reasoning. Alternatively, the verification of knowledge may result in new hypotheses. That is how knowledge is accumulated.

6.3 Guidelines

Based on the discussion of the knowledge generation pipelines, we propose four guidelines for good analytical system design.

Enable Induction and Deduction. To enable both induction and deduction is to enable both bottom-up reasoning and top-down reasoning. The analysts should be able to generate knowledge from data or verify the hypotheses with data.

Enable Knowledge Externalization. Because the analysts proceed inductive reasoning and deductive reasoning iteratively in the visual analytical process, knowledge is better to be externalized (e.g., as attached notes[112]) so that they can be referred to anytime.

Enable Data Provenance. To enable data provenance is another aspect of enabling deductive reasoning because it facilitates bottom-up reasoning. VisTrails[113] captures data provenance with a history management interface.

Enable Uncertainty-Aware Knowledge Generation. In the knowledge generation process, uncertainties are propagated from data to visualization[109]. The analysts need to be aware of the uncertainty before they can generate reliable knowledge.

7 Conclusions

In this survey we presented a comprehensive summarization of visual analytics pipelines. We reviewed the classic visual analytics feedback loop proposed by Keim *et al.*[2] Furthermore, we described individual stages in the loop and discussed detailed pipelines. Additionally, we discussed the commonality and the difference of various pipeline representations in each stage. For most of the visual analytics system, all the stages are contained in the analysis pipeline with different functionalities and implementations, and thus in our summarization we took several existing visual analytics systems as examples and assigned their analysis pipelines into corresponding stages. The visual analytics pipeline can be used as a guideline for structuring and developing visual analytics systems in real life.

References

[1] Fayyad U M, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, Fayyad U M, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds.),

American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996, pp.1-34.

[2] Keim D, Kohlhammer J, Ellis G, Mansmann F (eds.). Mastering the information age: Solving problems with visual analytics. http://www.vismaster.eu/wp-content/uploads/2010/11/title-page-to-chapter-1.pdf, June 2016.

[3] Keim D, Andrienko G, Fekete J D *et al.* Visual analytics: Definition, process, and challenges. In *Lecture Notes in Computer Science 4950*, Kerren A, Stasko J T, Fekete J D *et al.* (eds.), Springer Berlin Heidelberg, 2008, pp.154-175.

[4] Zhang L, Stoffel A, Behrisch M *et al.* Visual analytics for the big data era — A comparative review of state-of-the-art commercial systems. In *Proc. IEEE Conference on Visual Analytics Science and Technology*, Oct. 2012, pp.173-182.

[5] Sun G D, Wu Y C, Liang R H, Liu S X. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology*, 2013, 28(5): 852-867.

[6] Moreland K. A survey of visualization pipelines. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 19(3): 367-378.

[7] Alexander E, Gleicher M. Task-driven comparison of topic models. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 320-329.

[8] Sun M, North C, Ramakrishnan N. A five-level design framework for bicluster visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(12): 1713-1722.

[9] Zhang J, E Y, Ma J *et al.* Visual analysis of public utility service problems in a metropolis. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(12): 1843-1852.

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

[11] 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*, 2016, 22(1): 549-558.

[12] Lu Y, Kruger R, Thom D, Wang F, Koch S, Ertl T, Maciejewski R. Integrating predictive analytics and social media. In *Proc. the IEEE Conference on Visual Analytics Science and Technology*, Nov. 2014, pp.193-202.

[13] Ferstl F, Burger K, Westermann R. Streamline variability plots for characterizing the uncertainty in vector field ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 767-776.

[14] Skraba P, Wang B, Chen G, Rosen P. Robustness-based simplification of 2D steady and unsteady vector fields. *IEEE Transactions on Visualization and Computer Graphics*, 2015, 21(8): 930-944.

[15] Wang Z, Ye T, Lu M *et al.* Visual exploration of sparse traffic trajectory data. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(12): 1813-1822.

[16] Wang F, Chen W, Wu F *et al.* A visual reasoning approach for data-driven transport assessment on urban roads. In *Proc. the IEEE Conference on Visual Analytics Science and Technology*, Oct. 2014, pp.103-112.

[17] Huang X, Zhao Y, Ma C *et al.* TrajGraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 160-169.

[18] Vrotsou K, Janetzko H, Navarra C *et al.* SimpliFly: A methodology for simplification and thematic enhancement of trajectories. *IEEE Transactions on Visualization and Computer Graphics*, 2015, 21(1): 107-121.

[19] Palomo C, Guo Z, Silva C T, Freire J. Visually exploring transportation schedules. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 170-179.

[20] Scheepens R, Hurter C, Van De Wetering H, Van Wijk J J. Visualization, selection, and analysis of traffic flows. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 379-388.

[21] Di Lorenzo G, Sbodio M, Calabrese F *et al.* AllAboard: Visual exploration of cellphone mobility data to optimise public transport. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(2): 1036-1050.

[22] Wu W, Xu J, Zeng H *et al.* TelCoVis: Visual exploration of co-occurrence in urban human mobility based on Telco data. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 935-944.

[23] Zhao J, Cao N, Wen Z *et al.* #FluxFlow: Visual analysis of anomalous information spreading on social media. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(12): 1773-1782.

[24] Huang D, Tory M, Aseniero B A *et al.* Personal visualization and personal visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 2015, 21(3): 420-433.

[25] Janicke S, Focht J, Scheuermann G. Interactive visual profiling of musicians. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 200-209.

[26] Glueck M, Hamilton P, Chevalier F *et al.* PhenoBlocks: Phenotype comparison visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 101-110.

[27] Chen H, Zhang S, Chen W *et al.* Uncertainty-aware multidimensional ensemble data visualization and exploration. *IEEE Transactions on Visualization and Computer Graphics*, 2015, 21(9): 1072-1086.

[28] Thudt A, Baur D, Huron S, Carpendale S. Visual mementos: Reflecting memories with personal data. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 369-378.

[29] Wongsuphasawat K, Moritz D, Anand A *et al.* Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 649-658.

[30] Lex A, Gehlenborg N, Strobelt H *et al.* UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(12): 1983-1992.

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

[32] Dasgupta A, Poco J, Wei Y *et al.* Bridging theory with practice: An exploratory study of visualization use and design for climate model comparison. *IEEE Transactions on Visualization and Computer Graphics*, 2015, 21(9): 996-1014.

[33] Quinan P S, Meyer M. Visually comparing weather features in forecasts. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 389-398.

[34] Accorsi P, Lalande N, Fabregue M *et al.* HydroQual: Visual analysis of river water quality. In *Proc. the IEEE Conference on Visual Analytics Science and Technology*, Oct. 2014, pp.123-132.

[35] Crnovrsanin T, Muelder C, Ma K L. A system for visual analysis of radio signal data. In *Proc. the IEEE Conference on Visual Analytics Science and Technology*, Oct. 2014, pp.33-42.

[36] Goodwin S, Dykes J, Slingsby A, Turkay C. Visualizing multiple variables across scale and geography. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 599-608.

[37] Kurzhals K, Hlawatsch M, Heimerl F *et al.* Gaze stripes: Image-based visualization of eye tracking data. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 1005-1014.

[38] Etemadpour R, Motta R, de Souza Paiva J G *et al.* Perception-based evaluation of projection methods for multidimensional data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2015, 21(1): 81-94.

[39] Sun M, Mi P, North C, Ramakrishnan N. BiSet: Semantic edge bundling with biclusters for sensemaking. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 310-319.

[40] Brehmer M, Ingram S, Stray J, Munzner T. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(12): 2271-2280.

[41] Gad S, Javed W, Ghani S *et al.* ThemeDelta: Dynamic segmentations over temporal topic models. *IEEE Transactions on Visualization and Computer Graphics*, 2015, 21(5): 672-685.

[42] Fulda J, Brehmer M, Munzner T. TimeLineCurator: Interactive authoring of visual timelines from unstructured text. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 300-309.

[43] Bach B, Shi C, Heulot N *et al.* Time curves: Folding time to visualize patterns of temporal evolution in data. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 559-568.

[44] McCurdy N, Lein J, Coles K *et al.* Poemage: Visualizing the sonic topology of a poem. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 439-448.

[45] Brooks M, Amershi S, Lee B *et al.* FeatureInsight: Visual support for error-driven feature ideation in text classification. In *Proc. the IEEE Conference on Visual Analytics Science and Technology*, Oct. 2015, pp.105-112.

[46] Wu Y, Liu S, Yan K *et al.* OpinionFlow: Visual analysis of opinion diffusion on social media. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(12): 1763-1772.

[47] Gomez S R, Guo H, Ziemkiewicz C, Laidlaw D H. An insight- and task-based methodology for evaluating spatiotemporal visual analytics. In *Proc. the IEEE Conference on Visual Analytics Science and Technology*, Oct. 2014, pp. 63-72.

[48] Yu B, Doraiswamy H, Chen X *et al.* Genotet: An interactive web-based visual exploration framework to support validation of gene regulatory networks. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(12): 1903-1912.

[49] Lenz O, Keul F, Bremm S *et al.* Visual analysis of patterns in multiple amino acid mutation graphs. In *Proc. the IEEE Conference on Visual Analytics Science and Technology*, Oct. 2014, pp.93-102.

[50] Skanberg R, Vazquez P P, Guallar V, Ropinski T. Real-time molecular visualization supporting diffuse interreflections and ambient occlusion. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 718-727.

[51] Shi L, Wang C, Wen Z *et al.* 1.5 D egocentric dynamic network visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2015, 21(5): 624-637.

[52] Janikow C Z. Fuzzy decision trees: Issues and methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 1998, 28(1): 1-14.

[53] Liu M, Wang X, Huang Y. Data preprocessing in data mining. *Scientific Journal of Computer Science*, 2000, 27(4): 54-57. (in Chinese)

[54] Friedman M, Levy A Y, Millstein T D. Navigational plans for data integration. In *Proc. the 16th National Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence,* July 1999, pp.67-73.

[55] Lenzerini M. Data integration: A theoretical perspective. In *Proc. the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, June 2002, pp.233-246.

[56] Rahm E, Do H H. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 2000, 23(4): 3-13.

[57] Chen W, Shen Z, Tao Y. Data Visualization. Publishing House of Electronics Industry, 2013. (in Chinese)

[58] Chi E H h, Riedl J T. An operator interaction framework for visualization systems. In *Proc. the IEEE Symposium on Information Visualization*, Oct. 1998, pp.63-70.

[59] Chi E H. A taxonomy of visualization techniques using the data state reference model. In *Proc. the IEEE Symposium on Information Visualization*, Oct. 2000, pp.69-75.

[60] Card S K, Mackinlay J D, Shneiderman B. Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann, 1999.

[61] Van Wijk J J. The value of visualization. In *Proc. the 16th IEEE Visualization Conference*, Oct. 2005, pp.79-86.

[62] Munzner T. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 2009, 15(6): 921-928.

[63] Munzner T. Visualization Analysis and Design. CRC Press, 2014.

[64] Albo Y, Lanir J, Bak P, Rafaeli S. Off the radar: Comparative evaluation of radial visualization solutions for composite indicators. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 569-578.

[65] Gschwandtner T, Bogl M, Federico P, Miksch S. Visual encodings of temporal uncertainty: A comparative user study. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 539-548.

[66] Johansson J, Forsell C. Evaluation of parallel coordinates: Overview, categorization and guidelines for future research. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 579-588.

[67] Jianu R, Rusu A, Hu Y, Taggart D. How to display group information on node-link diagrams: An evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(11): 1530-1541.

[68] Lee J H, McDonnell K T, Zelenyuk A, Imre D, Mueller K. A structure-based distance metric for high-dimensional space exploration with multidimensional scaling. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(3): 351-364.

[69] Kieffer S, Dwyer T, Marriott K, Wybrow M. HOLA: Human-like orthogonal network layout. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 349-358.

[70] Raidou R G, Eisemann M, Breeuwer M, Eisemann E, Vilanova A. Orientation-enhanced parallel coordinate plots. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 589-598.

[71] Lehmann D J, Theisel H. Optimal sets of projections of high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 609-618.

[72] Yoghourdjian V, Dwyer T, Gange G *et al*. High-quality ultra-compact grid layout of grouped networks. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 339-348.

[73] Wang Baldonado M Q, Woodruff A, Kuchinsky A. Guidelines for using multiple views in information visualization. In *Proc. the Working Conference on Advanced Visual Interfaces*, May 2000, pp.110-119.

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

[75] Roberts J C. State of the art: Coordinated & multiple views in exploratory visualization. In *Proc. the 5th International Conference on Coordinated and Multiple Views in Exploratory Visualization*, July 2007, pp.61-71.

[76] Papadopoulos C, Gutenko I, Kaufman A. VEEVVIE: Visual explorer for empirical visualization, VR and interaction experiments. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 111-120.

[77] Wang Y, Shen Q, Archambault D, Zhou Z, Zhu M, Yang S, Qu H. AmbiguityVis: Visualization of ambiguity in graph layouts. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 359-368.

[78] Roberts J C. Display models: Ways to classify visual representations. In *Proc. IEEE Conference on Information Visualization*, July 1999.

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

[80] Chuah M C, Roth S F. On the semantics of interactive visualizations. In *Proc. the IEEE Symposium on Information Visualization*, Oct. 1996, pp.29-36.

[81] Lam H. A framework of interaction costs in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2008, 14(6): 1149-1156.

[82] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2005.

[83] Ma Y, Cao Z, Wei C. A survey of visualization-driven interactive data mining approaches. *Journal of Computer-Aided Design & Computer Graphics*, 2016, 28(1): 1-8. (in Chinese)

[84] De Oliveira M C F, Levkowitz H. From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 2003, 9(3): 378-394.

[85] Ma K L. Machine learning to boost the next generation of visualization technology. *IEEE Transactions on Computer Graphics and Applications*, 2007, 27(5): 6-9.

[86] Bertini E, Lalanne D. Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In *Proc. the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*, June 2009, pp.12-20.

[87] Klemm P, Lawonn K, Glaβer S *et al*. 3D regression heat map analysis of population study data. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 81-90.

[88] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques (3rd edition). Morgan Kaufmann, 2011.

[89] Lu J, Ma Y, Chen W *et al*. Recent progress and trends in predictive visual analytics. *Frontiers of Computer Science*, 2016. (accepted)

[90] Jain A, Zongker D. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997, 19(2): 153-158.

[91] Dash M, Liu H. Feature selection for classification. *Intelligent Data Analysis*, 1997, 1(1/2/3/4): 131-156.

[92] Dy J G, Brodley C E. Feature selection for unsupervised learning. *The Journal of Machine Learning Research*, 2004, 5: 845-889.

[93] Seo J, Shneiderman B. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 2005, 4(2): 96-113.

[94] Krause J, Perer A, Bertini E. INFUSE: Interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(12): 1614-1623.

[95] Markovitch S, Rosenstein D. Feature generation using general constructor functions. *Machine Learning*, 2002, 49(1): 59-98.

[96] Schuller B, Reiter S, Rigoll G. Evolutionary feature generation in speech emotion recognition. In *Proc. the IEEE International Conference on Multimedia and Expo*, July 2006, pp.5-8.

[97] Zahalka J, Worring M. Towards interactive, intelligent, and integrated multimedia analytics. In *Proc. the IEEE Conference on Visual Analytics Science and Technology*, Oct. 2014, pp.3-12.

[98] Janetzko H, Sacha D, Stein M *et al*. Feature-driven visual analytics of soccer data. In *Proc. the IEEE Conference on Visual Analytics Science and Technology*, Oct. 2014, pp.13-22.

[99] Zhao J, Gou L, Wang F, Zhou M. Pearl: An interactive visual analytic tool for understanding personal emotion style derived from social media. In *Proc. the IEEE Conference on Visual Analytics Science and Technology*, Oct. 2014, pp.203-212.

[100] Kay M, Heer J. Beyond Weber's law: A second look at ranking visualizations of correlation. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 469-478.

[101] Harrison L, Yang F, Franconeri S, Chang R. Ranking visualizations of correlation using Weber's law. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(12): 1943-1952.

[102] Bogl M, Aigner W, Filzmoser P *et al*. Visual analytics for model selection in time series analysis. *IEEE Transactions on Visualization and Computer Graphics*, 2013, 19(12): 2237-2246.

[103] Ware C. Information Visualization: Perception for Design (3rd edition). Morgan Kaufmann, 2012, pp.388-391.

[104] Pirolli P, Card S. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proc. the International Conference on Intelligence Analysis*, May 2005, pp.2-4.

[105] Sacha D, Stoffel A, Stoffel F *et al*. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(12): 1604-1613.

[106] Ma Y, Chen W, Ma X *et al*. EasySVM: A visual analysis approach for open-box support vector machines. In *Proc. IEEE VIS Workshop on Visualization for Predictive Analytics*, Nov. 2014.

[107] Sacha D, Senaratne H, Kwon B C, Ellis G, Keim D A. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 2016, 22(1): 240-249.

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

[109] Dykes J, MacEachren A, Kraak M. Beyond tools: Visual support for the entire process of GIScience. In *Exploring Geovisualization*, Dykes J, MacEachren A M, Kraak M J (eds.), Elsevier Ltd., 2005, pp.83-99.

[110] Klein G, Moon B, Hoffman R R. Making sense of sensemaking 2: A macrocognitive model. *IEEE Transactions on Intelligent Systems*, 2006, 21(5): 88-92.

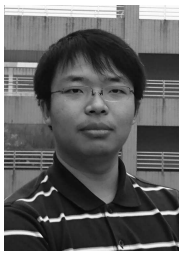
[111] Legrenzi P, Girotto V, Johnson-Laird P N. Focussing in reasoning and decision making. *Cognition*, 1993, 49(1/2): 37-66.

[112] Andrews C, North C. The impact of physical navigation on spatial organization for sensemaking. *IEEE Transactions on Visualization and Computer Graphics*, 2013, 19(12): 2207-2216.

[113] Callahan S P, Freire J, Santos E *et al*. VisTrails: Visualization meets data management. In *Proc. the 2006 ACM SIGMOD International Conference on Management of Data*, June 2006, pp.745-747.

Xu-Meng Wang is a Ph.D. student in the State Key Laboratory of Computer Aided Design (CAD) and Computer Graphics (CG) at Zhejiang University, Hangzhou. She earned her B.S. degree in information and computing science from Zhejiang University in 2016. Her research interest is visual analytics.

Tian-Ye Zhang is a Ph.D. student in the State Key Lab of CAD&CG at Zhejiang University, Hangzhou. She earned her B.S. degree in information and computing science from Zhejiang University in 2016. Her research interest is visual analytics.



Yu-Xin Ma is a Ph.D. student in the State Key Lab of CAD&CG at Zhejiang University, Hangzhou. He earned his B.S. degree in software engineering from Zhejiang University, in 2012. His research interests are visual analytics and information visualization.

Jing Xia is a Ph.D. student at the State Key Lab of CAD&CG, Zhejiang University, Hangzhou. She earned her B.S. degree in software engineering from Zhejiang University, in 2011. Her research interests include information visualization and visual analytics, especially visual analytics of temporal data and spatial-temporal data.

Wei Chen is a professor at the State Key Lab of CAD&CG, Zhejiang University, Hangzhou. Professor Chen received his Ph.D. degree in applied mathematics from Zhejiang University in 2002. His research interests include visualization, visual analytics, and biomedical image computing. He has published more than 60 papers in international journals and conferences. Professor Chen served as papers co-chair of IEEE PacificVis 2013 and conference chair of IEEE PacificVis 2015. He presently serves on the steering committee of IEEE PacificVis.