8 Evaluation

8.1 Motivation

Visual analytics is a promising and ambitious concept. The aims are to enable people to get insight in large amounts of heterogeneous data, understand the underlying phenomena described by the data, to smoothly integrate multiple data analysis methodologies, and to offer support for the complete knowledge discovery process. These aims are very challenging. For many practical instances, it is unknown how to reach these; for existing solutions it is often unknown how well they realise these aims; and overall, there is a lack of solid findings, models and theories. As a result, visual analytics still has a long way to go before it can be considered a mature technology.

In making progress towards meeting these aims, evaluation will play a crucial role, but the characteristics of visual analytics presents difficult problems for effective evaluation. In this chapter, we elaborate on this by examining particular problems, then give an overview of the state of the art in evaluation, and finally present some recommendations for the research roadmap.

Evaluation include techniques, methods, modes and theories as well as software tools

Evaluation concerns here the assessment of the quality of artefacts related to visual analytics. Both *quality* and *artefacts* should be considered as broad container terms. Artefacts are not limited to software tools, but also include, for example, techniques, methods, models and theories. As visual analytics is both a science and a technology, the key aspects of quality are *effectiveness*, *efficiency*, and *user satisfaction*. In other words, artefacts should be evaluated on whether they fulfil their aims, on the resources required, and whether they meet needs and expectations of users. Taking a broad view, this includes aspects such as degree of fit in current workflows, performance, and ease of use. As argued in the previous chapter, users are central in all this, and awareness of their importance is still increasing, not only in visual analytics, but also in related fields. One example from geovisualisation is that the International Cartographic Association (ICA) has established a committee on Use and User Issues[1]

Stakeholders include developers and end users

The results of evaluation are important for all stakeholders. Integrators and end-users of visual analytics need to know about the quality of artefacts. Put practically, the developer of a new system that takes advantage of visual analytics techniques needs to know which techniques to choose for the problem at hand; users who have to select a system, a method, or even a parameter-setting need information to make the best decision, in order to save time and to prevent themselves from the use of inappropriate techniques, leading to

[1] http://www.univie.ac.at/icacomuse

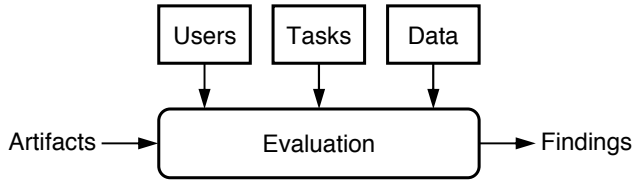

Figure 8.1: The main ingredients of evaluation

wrong results. Furthermore, good evaluation results are important to convince integrators and end-users to adopt novel techniques. Hence, evaluation is an important ingredient in the innovation process, from research to application on larger scales.

The task of researchers and developers is not just to develop new ideas and techniques; assessment of the quality, scope, and applicability of those innovations is equally important. As mentioned, those results are vital for end-users, but also for research itself. Evaluation can show which problems have and have not been solved, it provides benchmarks, against which new results can be compared.

However, for several reasons proper evaluation of visual analytics is not easy. First, visual analytics encompasses many different aspects and disciplines, which makes it hard to make generic statements; second, in visual analytics humans play a central role, in contrast to, say, statistics.

Evaluation involves users, tasks and data

Figure 8.1 shows a schematic overview of evaluation in visual analytics. Evaluation leads to findings on the quality of artefacts. Such findings are never absolute, but depend on *users*, *tasks*, and *data*, which taken together define the scope of the findings. To give a simple example, a finding could be that the use of scatterplots (artefact) is helpful to find clusters (task) in records with a limited number of real-valued attributes (data), provided that observers have had training in the proper interpretation (users). Such findings can be produced using relatively simple lab experiments, as all aspects are well-defined. Much more challenging is to obtain generic findings, such as when to use automated techniques instead of techniques with a human in the loop, for broad classes of users, tasks, and data. Another challenge is to obtain precise, quantitative findings, for instance on how much time is saved by adopting a technique. Again, solid findings would be highly useful, and to produce such findings is a major challenge for the field. However, an even more daunting challenge is to obtain findings that characterise the knowledge discovery process: the rationale behind the decisions taken by the user and the type (and quality and quantity) of insight being obtained.

Obtaining findings which can be applied generically is a daunting task

The complexity and diversity of users, tasks and data is high

The complexity and size of evaluation in visual analytics can be understood further by considering the ingredients (users, tasks, artefacts and data) in more detail. All these are complex in themselves. They are hierarchical, because different levels of abstraction can be distinguished; multivariate, because different properties can be distinguished; and heterogeneous, because in real-world scenarios, combinations of data, tasks, etc. usually have to be dealt with. This

complexity is within the core of the mission of visual analytics. Whereas other fields in visualisation often focus on specific user groups with well-defined tasks and standardised, homogeneous datasets, visual analytics aims at much more diversity. In the following, this diversity and complexity is discussed in more detail for users, tasks, artefacts, and data.

Users. The user community targeted at is large. In the ideal case, findings apply to the general user, but for specific problems specific users have to be targeted, and their capabilities, interests, and needs have to be taken into account (for more on this, see Chapter 7). At various levels of detail, a distinction can be made between professionals and a lay-audience; professionals can be split up into, for instance, scientists, data-analysts, managers, etc.; and of course, all these categories can be subdivided further, down to, for example, experts in European patents on laser-optics technology or bioinformatics researchers dealing with crop diseases. Furthermore, aspects like age, country, culture, gender, training, perceptual and cognitive skill levels, or motivation can have an influence on the performance obtained when an artefact is used.

Obtaining appropriate expert users is difficult; results from using students may not be representative

This leads to interesting problems for evaluation. For example, dealing with experts requires a thorough understanding of their needs and wishes, such that the appropriate aspects are evaluated; also, such experts are often scarce and have limited time available. One often used escape route is to replace the experts with undergraduate students and have them evaluate new methods and techniques, but it is unclear to what extent the results found carry over to real-world users.

Tasks. Users apply visual analytics to fulfil tasks, and here again complexity strikes. In information visualisation, often just low-level tasks are considered, such as spotting trends, clusters, and outliers. However, people that use visual analytics have to carry out tasks like protecting the safety of a computer network or a transportation system, manage a company, or decide on a policy. There are many levels between such complex responsibilities and the elementary tasks; and, given the ambition of visual analytics these fall within the scope. A practical and important issue here is that such more complex tasks do not lend themselves well to standard lab-experiments. They can require from days to months to complete, require in-depth expertise of the subjects, and these tasks are too important to allow wrong decisions to be made. In the following section, current approaches to handle this are discussed.

Complex and extended tasks are often not suitable for laboratory experiments

Artefacts. The artefacts of visual analytics can also be considered at various levels of detail. On a very detailed scale, one can study the effectiveness of, say, graphical representations or a specific technique. On a higher level are the software tools, to be compared with other tools. On a still higher level, one can study the suitability of such technologies in general. This implies that one also has to study aspects such as the available tutorial material, coupling with other systems, and the costs involved. Besides these levels, the scope of the artefacts varies greatly. Artefacts can relate to visualisation, automated analysis, knowledge management, presentation, data cleansing, etc., and in a full-blown

Artefacts for evaluation range from graphical representations to the suitability of particular technologies

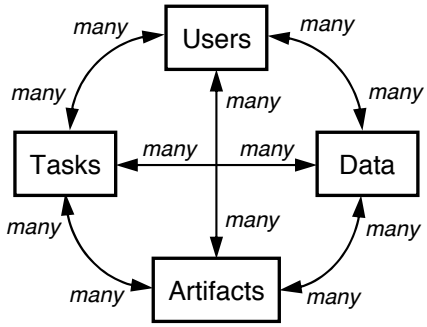

Figure 8.2: Relations between users, tasks, data, and artefacts

environment for visual analytics, all these issues have to be addressed in one way or another.

Data. The data to be considered is also complex (see Chapter 3 for a detailed discussion). Whereas a standard visualisation usually deals with homogeneous, single datasets (which is often difficult enough), visual analytics has to deal with combinations of heterogeneous data (for example, weather data, multi-media, written reports), huge amounts of data, requiring reduction via automated methods; and new data can arrive or is sought during the analysis.

In summary, we argued that users, tasks, artefacts, and data in visual analytics are complex and heterogeneous. In reality, it is even more complex, as all this complexity multiplies, as shown in Figure 8.2. In a simple laboratory experiment, one standard user evaluates a few variations of an artefact, for a small number of well-defined tasks, using similar datasets. In the real world, people use several tools simultaneously, have a variety of tasks, use many different datasets, cooperate with other people, and all this over extended periods of time in a flexible and dynamic setting. All this makes evaluation a difficult task, and shows that it is not easy to find generic and reliable answers to the question of which artefacts to use and when.

In the next sections, we describe the state of the art of evaluation methodologies in visual analytics and present recommendations for improvements of current approaches.

8.2 State of the Art

Visual analytics artefacts should be evaluated in terms of effectiveness, efficiency, and user satisfactions to assess their quality. This requires evaluation methodologies, covering a wide range of algorithmic performance measures to real-world technology adoption and utility metrics. Chapter 6 of Thomas and Cook's book[111] outlines evaluation approaches for visual analytics on three levels: *component*, *system*, and *environment*. With respect to components, there

exists a proliferation of isolated evaluations. On the system level, success is hard to quantify and difficult to trace back to individual components or computations. Here, it is important to track the history of investigation, e. g., in analytic workflows. Metrics are needed to address the learnability and utility of systems. Quantification of insights is essential (examples in bioinformatics have recently appeared[95]). On the environment level, evaluation needs to consider technology adoption. Across all levels, one needs to measure the benefit of the technology in producing an improved product.

Metrics are needed to measure usability, learnability, quantification of insights and technology benefits

Visual analytics technology is used by people who carry out their tasks with visualisation tools, sometimes over long periods of time, searching for information in various ways[88]. This means that, in addition to measures of performance and efficiency, there is a need to evaluate the interaction of people with the visualisation tools in order to understand their usability, usefulness, and effectiveness. Such aspects can be addressed by empirical evaluation methodologies, as often used in the fields of human-computer interaction (HCI) and computer-supported collaborative work (CSCW).

This section gives an overview of the state of the art of such methods for evaluating visual analytics.

8.2.1 Empirical Evaluation Methodologies

A range of evaluation methods exist for examining interactive techniques[26]. These include quantitative methods, qualitative methods, mixed method approaches, usability studies, and informal evaluation techniques (the classes not being mutually distinct). Depending on the chosen method one can, for example, examine in a controlled environment (such as a laboratory) very specific questions for which a testable hypothesis can be formulated, and this can lead to conclusions with high confidence. Another type of evaluation can look at broader questions using qualitative methods. Here, the focus is on data acquisition through observation and interviewing. A wide range of specific techniques exists in this area that can be used depending on the types of questions to be answered. In addition, there are also mixed-method techniques that combine aspects from both qualitative and quantitative evaluation. A separate category within evaluation techniques is usability evaluation, which deals specifically with the ease of use of interactive tools. Here, a combination of quantitative and qualitative evaluation can be employed. Finally, informal evaluations can be used. These involve fewer people who give feedback on a visualisation or the interactive system used to create them, providing anecdotal evidence for its usefulness or effectiveness, which can be useful for techniques that mainly focus on a technical contribution.

Range of evaluation methods include qualitative, quantitative, combined and informal

A number of papers discuss evaluation methodology in general. In his seminal paper, McGrath[78] identifies important factors that are all desired but not simultaneously realisable in evaluation studies: *generalisability*, *precision*, and *realism*. He also classifies specific evaluation approaches with respect to their abstractness and obtrusiveness and, on this continuum, indicates the respective position of the three aforementioned factors.

Evaluation strives to be generalisable, precise and realistic

Qualitative and longitudinal studies are particularly suitable for information visualisation

In her overview on evaluation methodologies for information visualisation[26], Carpendale carefully discusses the various approaches in quantitative and qualitative evaluation, following the terminology of McGrath: field study, field experiment, laboratory experiment, experimental simulation, judgement study, sample survey, formal theory, and computer simulation. In particular, she emphasises the importance of qualitative approaches as a valid group of evaluation techniques. Plaisant[86] discusses the challenges of evaluation, also in the context of information visualisation. In addition to controlled experiments, the need for longitudinal studies is stressed. Recommended steps to improve evaluation and facilitate adoption are: repositories (data and tasks), collecting case studies and success stories, and strengthening the role of toolkits. Chen and Yu[28] report on a meta-analysis of empirical studies in information visualisation. They included only studies on tree or network visualisations and restricted themselves to studies with information retrieval tasks. Due to the very strict requirements, of the original 35 studies selected only 6 remained for the final analysis. They found that due to the diversity of studies it is very difficult to apply meta-analysis methods. They conclude that the measurement of higher cognitive abilities is especially hard and more task standardisation in cognitive ability testing is required.

Repositories of data, tasks, case studies are useful

Task based evaluation is generally not suitable to measure insight

Zhu[129] focuses on the definition of effectiveness of visualisation and how to measure it. Current definitions of effectiveness are reviewed and a more comprehensive definition of effectiveness is introduced, which comprises accuracy, utility, and efficiency. As one aspect of efficiency of evaluation, North in his Visualisation Viewpoints paper[82] focuses on the question of how to measure insight, the ultimate goal of all of visualisation. One of his main observations is that task-based evaluation is too narrow. What works well for a given task might not work at all for tasks not studied. Generalisation from simple to difficult tasks is hard. Either more complex tasks are needed, or one may eliminate benchmark tasks completely and put the emphasis on more qualitative insights. Involving learning processes as in traditional education will be helpful. Finally, Munzner[80] presents a nested model for visualisation design and evaluation. She subdivides the process of creating visualisations into four nested levels: domain problem characterisation, data/operation abstraction design, encoding/interaction technique design, and algorithm design. She then argues that distinct evaluation methodologies should be used for each of these levels because each of the levels has different threats to its validity.

Generalisation is difficult

8.2.2 Examples of Evaluation

We now discuss a number of specific approaches to evaluation of visual analytics that are used in practice.

Program understanding and software visualisation. In the field of program understanding and software visualisation, evaluation of combined analysis and visualisation techniques and tools already has a rich history.

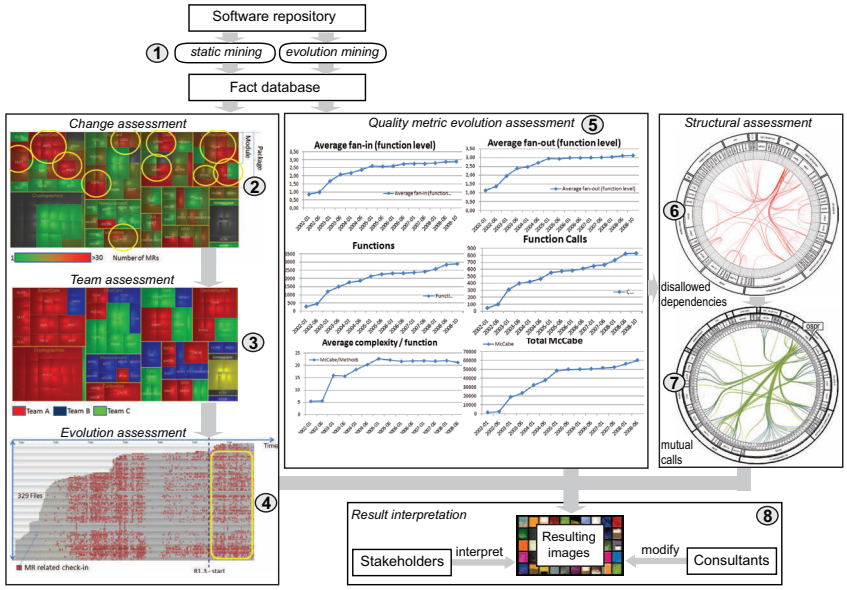

Figure 8.3: Visual analytics for software product and process assessment. A large software repository is mined (1), a variety of aspects are visualised and assessed (2)-(7), findings are discussed with stakeholders and consultants (8)

Several particular difficulties for evaluation exist in this field. The diversity of tasks and questions in software analysis requires analysts to easily combine and customise their analysis and visualisation tools in ways, which often go beyond what these tools were designed to support. Many such tasks require high precision analyses, such as extracting exact call, control flow, and data flow graphs in static analysis or the exact detection of structural or execution patterns in program verification. This is a challenge both for the analysis tools, but also for the creation of visualisations able to convey exact insights at fine-grained levels. Scale is also a problem. Modern software projects have millions of lines of code, structured in tens of thousands of entities, which are modified continuously by hundreds of developers over many years. Although most software systems are hierarchically organised, which enables level-of-detail and aggregation techniques to simplify visualisations, many analyses such as impact and correctness assessment involve concerns, which cut across boundaries and make localised examination difficult.

Scale is a problem

Shneiderman and Plaisant[101] argue for the need for longitudinal studies as an effective means for understanding the issues involved in tool adoption. A useful taxonomy of evaluation procedures is given by Kraemeret al.[84], who considers surveys, interviews, observational studies and think-aloud studies. These studies form valuable input in organising effective evaluations in the larger context of visual analytics.

There are many types of evaluation including, longitudinal studies, surveys, interviews, observation and think-aloud studies

Several evaluations of the effectiveness of visual analytics solutions in software

A large study of visual analytics in software maintenance suggest a tight integration of analysis and visualisation tools

maintenance are presented by Voinea et al.[120]. They used software visualisation, repository data mining, and static analysis tools to support assessments of software product quality attributes such as maintainability, modularity and complexity, and to elicit process patterns such as implicit developer networks and workflow structure. These techniques have been used on a variety of repositories such as open-source code bases with millions of lines of code, but also on commercial projects. Insight was collected from a variety of users including over sixty master students and over twenty professional software developers in the industry, over a period of several years. Usage patterns involved using tools during education, open-source development, longitudinal product development, and short assessment sessions. Evaluation results pointed strongly to the need of using simple visualisation metaphors, tight integration of analysis and visualisation tools within the accepted workflow by the targeted stakeholder group, and visualisation design, which closely reflects the concepts and values seen as important by the users.

The usage of visual analytics in software maintenance is a good illustration of the particular mix of heterogeneous data sources, combined analysis and visualisation, and the overall hypothesis forming, refinement, validation, and presentation of results for decision making support.

Illustrative example. An example of visual analytics for software product and process assessment (Voinea et al.[121]) is shown in Figure 8.3. In this application, an industrial automotive company developed a large software stack over a period of eight years. Towards the end of the process, it was seen that the product could not be completed within the time, and an assessment of the causes of the problem was required within one week in order to decide on the future of the project. Following static code and software evolution data mining (1), several hypotheses were generated for the problem causes, and several visualisations such as tree-maps, metric charts, compound graphs, and time-lines were used to assess the team and code evolution (2–4), quality evolution (5), and system architecture (6–7). The findings were combined and discussed with the stakeholders (i. e., project managers and team leaders), who obtained extremely valuable insights to guide their decision making.

The whole evaluation process needs to be considered from hypothesis forming to presentation of the results

This example outlines several interesting aspects related to visual analytics evaluation. The tools and techniques involved in analysis were used by a user group (the consultants), which was separate from the actual stakeholders (the project owners). Images and insights were presented by the tool users to the stakeholders, but result interpretation and decision making was left entirely to the latter group. The usage of simple business graphics, as opposed to more sophisticated visualisations, was seen as crucial for acceptance. As such, what was evaluated as successful was the entire process of hypothesis forming, refinement, validation, and presentation, rather than specific tool usability.

Collaborative visual analytics adds a further level of difficulty to its evaluation

Collaborative visual analytics. Visual analytics often involves a highly collaborative analysis process. Hence, evaluation plays an important role to determine how successful collaborative visual analytics systems can support the

reasoning processes in teams, something that is often difficult to evaluate in a controlled manner. Only few approaches have addressed this issue specifically, one of them presented by Isenberg and Fisher[63]. In their paper, the authors describe their Cambiera system that has dedicated support for awareness for collaborative visual analytics. In their informal evaluation of Cambiera, the authors presented two pairs of researchers with a data set in which they asked the participants to identify a story line. In particular, the authors paid attention to the use of awareness features provided by their tool and found that these were used by the participants in a variety of different ways.

8.2.3 Contests

One aim of evaluation is to find out what is the best approach to solve a certain problem. An interesting alternative way of evaluation is to compete: present a problem to the community and challenge researchers and developers to show that their solution is best. Competitions and contests have shown their value for advancing a field quickly.

Information retrieval has a long history of providing test collections

Early developments. In some research communities large scale, competitive evaluation efforts have a long history and developed into a central focus. For example, in text retrieval, large test collections are provided, aiming to encourage research and communication among industry, academia and governments. Different research areas are addressed with different tracks, which study issues such as text retrieval in blogs, legal documents, cross-language documents, Web pages etc.

Besides the development of text source material, it led to the development of standards for evaluation, e.g., the adoption of metrics like precision and recall.

An advantage in these cases is that the 'ground truth', i. e., what constitute good results, can be established objectively, even though generation of this ground truth often requires human experts. For exploratory data analysis and visualisation this is much harder to establish. A common format for contest in these communities is therefore to visualise a given data set and to report on findings.

The first graph drawing contest was in 1994

Graph drawing community. The graph drawing community focuses on the development of methods and techniques for producing diagrams of graphs that are aesthetically pleasing and follow layout conventions of an application domain. Annual contests have been held in conjunction with the Symposium of Graph Drawing since 1994. The categories have varied over the years, including free-style (all type of drawings for arbitrary datasets, judged on artistic merit and relevance), evolving graphs, interactive graph analysis, and social networks. Particularly interesting and exciting for participants is the on-site challenge format, where teams are presented a collection of graph data and have approximately one hour to submit their best drawings.

Information visualisation community. The information visualisation community started in 2003 with a contest at the yearly IEEE InfoVis Conference.

Information visualisation contests have run successfully since 2003

Catherine Plaisant, Jean-Daniel Fekete, and Georges Grinstein have been involved in the first three contests, and have given a thorough report on their experiences and lessons learned[87]. Participants were provided with a large dataset, and had typically four months to prepare their submissions, in the form of a two page summary, a video, and a Web page. Examples of datasets used are tree-structured data (from various sources), citation networks, multivariate data of technology companies, and data on Hollywood movies. Results were judged for the quality of the data analysis (what interesting insights have been found); quality of the techniques used (visual representation, interaction, flexibility); and quality of the written case study. Judges reviewed submissions, and they reported that this was difficult and time-consuming, as there was no ground truth available and because processes and results were difficult to compare. Participants in the contest worked hard and were motivated. Students could use the contest to test their PhD research and small companies reported that they appreciated the exposure.

Contest datasets are a valuable resource and should be made available in repositories

Some other recommendations given to organisers of contests are to facilitate student participation, to provide examples, and to use datasets with established ground truth. They emphasise that the contest is only a first step, and that the prepared datasets and the submissions are valuable material. They argue that an infrastructure is needed to facilitate the use of the datasets, leading to a repository of benchmarks. Also, given that the efforts required are above what can be expected from volunteers, they argue for support by funding agencies to plan and build up long term coordinated evaluation programs and infrastructures, to increase the impact of such contests.

Software visualisation community. Challenges involving evaluation of combined visualisation and analysis tools and methods are well established in software visualisation and several conferences have organised challenge tracks, specifically for software visualisation techniques.

Datasets are prepared and offered to participants for investigation several months in advance, with a number of questions being asked. Questions and tasks range from generic software understanding, such as investigating the evolution of large-scale software repositories in order to identify trends of interest for typical software maintenance tasks, up to precisely defined tasks such as assessing the modularity or presence of certain design and coding patterns for a given software system. Participants typically use visualisation and analysis tools of their own design to answer these questions. Contest entries are typically published as short papers in the conference proceedings.

Apart from challenges focusing on pre-selected datasets, software visualisation conferences also encourage the submission of short tool demonstration papers. In contrast to challenges, which focus on the insight gained when analysing a given data set, tool demo papers focus on a more general set of aspects that make a tool efficient and effective, such as scalability, genericity, integration with accepted workflows, and ease of use.

Challenges and tool demo contests in software visualisation share a number of particular aspects. Several de facto standard datasets have emerged from the

research community, such as the Mozilla Firefox, KDE, and ArgoUML code bases. Compared to some other sub-domains of visual analytics, generation and acquisition of realistic, challenging, data is not seen as a problem in software visualisation. Many open source repositories exist, which contain large and complex systems. These repositories cover a wide range of aspects, such as long-term evolving code, multiple designs, architecture, programming languages and patterns, and access to specific questions and challenges of the developers, present in design documents and commit logs. Furthermore, tools, technologies and data interchange formats are relatively well standardised across the field.

Many open source repositories exist within the software visualisation domain

Visual analytics community. In 2006 a highly relevant contest emerged: the VAST Contest[89], renamed to VAST Challenge in 2008, held in conjunction with the Visual Analytics Software and Technology symposium.

VAST Challenge

In several respects, the challenges in visual analytics contests are close to perfect. The data provided are large. Each year a new challenge is addressed, typically with a security or intelligence aspect. Several different datasets are provided for a challenge, each giving different cues and different aspects. For instance, the 2008 challenge scenario concerned a fictitious, controversial socio-political movement; and the data consisted of cell phone records, a chronicle of boat journeys with passenger lists, a catalogue of Wiki edits, and geo-spatial data of an evacuation after a bomb attack. The datasets are carefully generated by the National Visualisation and Analytics Center (NVAC) Threat Stream Generator project team at PNNL, and a ground truth (as well as false trails) is hidden in these.

High quality, complex data is at the heart of the successful VAST Challenge

In many respects, the VAST Challenge is highly successful. It encourages and stimulates researchers and students, and it has led to a repository of large heterogeneous datasets with ground truths. These datasets are used now by researchers to test new methods, but also in education. A large part of its success can be attributed to the high quality and high motivation of the organisers. Another important ingredient is the support by government agencies (NIST and PNNL), especially for constructing the datasets and for judging the results.

8.3 Next Steps

Overall, visual analytics is a highly promising concept for increasing the effectiveness of obtaining new insights. It helps solving actual data-related problems in many application fields where multivariate, highly dimensional, and complex datasets are involved. However, there are several challenges to the adoption of visual analytics in actual application areas, and evaluation is a crucial one of these. In the preceding sections, we have discussed why evaluation is highly important, why evaluation is hard in visual analytics, and that despite the efforts so far, there is a lack of solid findings. We now put forward recommendations to improve this situation, which we hope

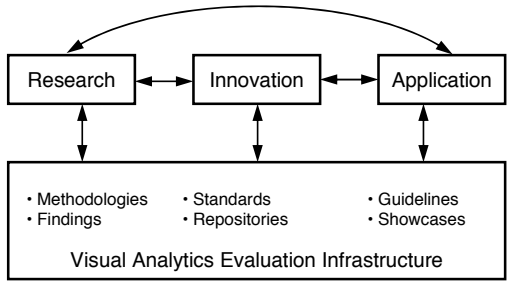

Figure 8.4: Overview of stakeholders and recommendations

will ultimately lead to successful adoption of visual analytics on a large scale.

We formulate our recommendations by urging the need for a solid evaluation infrastructure for visual analytics, consisting of a number of components. An overview of this evaluation infrastructure is shown in Figure 8.4. The main categories of stakeholders are researchers, innovators (translating new ideas into useful tools), and, most importantly, the users of these tools. To enable the latter to take advantage of the opportunities offered by visual analytics, various aspects of the innovation process have to be strengthened, leading to the following set of recommendations.

Stimulate research on evaluation methodologies for visual analytics. In the preceding sections, we have seen that visual analytics methods can be evaluated in a variety of ways, ranging from informal user studies to studies of the adoption in practise. All these have their own strengths and weaknesses. Given the complexity of the topic, it is highly unlikely that the current approaches used are already fully developed, and we are convinced that there is much room for improvement. A promising area is, for instance, the use of eye-tracking techniques and measurement of physical signals in general, as these give a wealth of information on what the user is actually doing, although the interpretation of this information is very hard. Instrumentation of software gives another detailed view on the actions of users, but here again correct interpretation is difficult. Experiments with professional, overburdened users is notoriously difficult, and often students are taken as replacements. The extent to which the results obtained through the use of non-experts can be translated to professional use is unclear. In short, we expect that methodologies for evaluation can be improved significantly, and that this will provide means to obtain more insight with less costs into the quality of visual analytics artefacts.

Stimulate evaluation of visual analytics methods and techniques. Despite enthusiastic efforts of the research community in visual analytics, there is still a great lack of solid results on the quality and scope of visual analytics artefacts. Given the novelty, size and complexity of the field, this is very understandable, but significant effort is required to improve this situation. There is a strong awareness now that evaluation is important, and in many research projects much

effort is already expended on evaluation. To stimulate this further, new research programs should emphasise and encourage evaluation, including efforts aimed at evaluating existing methods.

A particularly important driver for evaluation is setting the right frame of reference. In many cases, visualisation tools and techniques, which have successfully passed evaluations related to ease of use, interaction, and perception, are still not adopted by practitioners in the field. One major reason for this is that such tools are not seen as bringing essential value for their targeted users. This situation is even more important for visual analytics technologies which, by definition, combine many aspects, data sources and tools. As such, one promising direction is to evaluate the effectiveness of an entire visual analytics solution, or workflow, rather than to focus only on its separate components. However, to summarise the quality in a simple measure (e.g., time on task) is too simplistic, and leads to a 'credit assignment' problem (e.g., which features of the systems helped or not). This is crucial for formative evaluation. Ideally we should evaluate a complete workflow, but with carefully designed measurements that are able to tap into parts of the process, including interpretative feedback from users, maybe after the event to prevent interference effects.

Stimulate standardisation. Standardisation is a vital aspect in building up a body of knowledge. Visual analytics subsumes a wide variety of types of data, techniques, applications and users. In order to make evaluation results comparable and retrievable, standard definitions and taxonomies of all these aspects are important. In other fields, such as statistics, techniques can be accurately classified based on the characteristics of the data analysed, and the results have a clear and precise meaning. In visual analytics, such a precision cannot be attained, almost by definition. A fundamental assumption is that the data to be analysed is so complex that the human in the loop, with their own great strengths but also their complexity and variety, is essential. However, this does not mean that we should not try to reach higher levels. For example, the use of standard measures and tests to assess the perceptual and cognitive skills of participants would be a good step forward. In general, standardisation enables researchers, innovators, and users to exchange results, where the meaning of the various aspects is as clear and unambiguous as possible. We think that efforts in this direction are important and should be stimulated, as it helps to build a foundation and infrastructure that many will benefit from.

Stimulate repositories. Standardisation is one aspect to enable and stimulate exchange of results; the development of repositories is its natural complement. Evaluation of visual analytics can be performed much more effectively and efficiently if central repositories are set up and maintained that provide relevant material. Such repositories could provide:

- Datasets for a variety of applications and at various levels of detail. Preferably they should also include information on results to be found, that can act as a ground truth.

- Data generation tools to generate benchmark datasets, again of many different types and ranges of complexity, where the information to be found can be inserted on request.
- Analysis tools and libraries to perform automated analysis and evaluation, whenever possible.
- Standardised questionnaires to assess users experience of the artefacts tested;
- Detailed results of previous evaluation studies, to enable further analysis and comparison.

In certain fields, such as software visualisation, the emergence of lively open-source communities provides a good, low-cost, solution. Datasets such as software repositories are the prime vehicle of information interchange in such fields, and are open to everyone for examination. Given the focus on software technologies, this field also sees a strong development and sharing of analysis and visualisation tools, and strong interaction between researchers, industry practitioners and individual developers.

Collect showcases. For the adoption of visual analytics technology, the outside world has to become more aware of its possibilities and advantages. Potential users, include software and system developers, which could take advantage of integration of visual analytics technology in their products, as well as end-users. Collection and dissemination of showcases, including successful evaluation approaches, is important in this respect. Such showcases provide an overview of the possibilities, and should clearly show the benefits, in terms of novel and valuable insights obtained as well as reduced costs for data analysis. Here, we can exploit the particularities of each field to stimulate dissemination and create awareness.

Stimulate development of guidelines. Potential users need guidance on what technology to adopt, how to apply it in order to solve their problems, and how to evaluate the effectiveness. Development of guidelines, tutorials, handbooks deserves attention. These should be useful and understandable for the target audience, and be grounded in results from the scientific community as well as real world practise and experience.