A Research Agenda for Managing Uncertainty in Visual Analytics

Karsten Seipp[1], Xavier Ochoa[2], Francisco Gutiérrez[1], Katrien Verbert[1]

Department of Computer Science, KU Leuven[1]
Faculty of Electric and Computing Engineering, Escuela Superior Politécnica del Litoral[2]

Abstract

This paper proposes a research agenda to tackle the challenges of uncertainty in Visual Analytics. We propose a four-pronged strategy that addresses the need of both entities involved, human and computer. It comprises the communication of uncertainty in data and model, the communication of uncertainty regarding user intent and knowledge, human to human information interchange, as well as the definition of suitable, interactive means for conducting this dialogue with precision and ease.

1 Introduction

The combination of information visualisation with information mining techniques to support discovery of new knowledge has been researched for more than a decade. Shneiderman (2002) proposed this combination in so called 'discovery tools' to preserve user control, enable more effective exploration, and promote responsibility. In 2004, the term 'Visual Analytics' was coined. The Visual Analytics process combines automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data (Keim et al. 2010). The overarching driving vision of Visual Analytics is to turn the information overload into an opportunity: just as information visualisation has changed our view on databases, the goal of Visual Analytics is to make our way of processing data and information transparent for an analytic discourse (Keim et al. 2008).

Although several interesting prototypes have been developed, little work has been done to enable non-experts in data processing to conduct and steer analysis tasks (Keim et al. 2010). There is an increased interest of such users to steer data analysis. Examples include researchers in the humanities who want to apply analysis techniques to large text corpora, or teachers who want to analyse student data. While these users are experts in their domains, they usually have little expertise in data processing. This lack of skills leads to a common pattern where the analytical task is shared by two user roles (Bernard et al. 2012): in a first step, the domain expert without data processing skills (e.g. a humanities scholar or teacher)

Veröffentlicht durch die Gesellschaft für Informatik e.V. 2016 in
B. Weyers, A. Dittmar (Hrsg.):
Mensch und Computer 2016 – Workshopbeiträge, 4. - 7. September 2016, Aachen.

http://dx.doi.org/10.18420/muc2016-ws11-0007

defines criteria for useful data as well as requirements for the analysis. In a second step, a data processing expert, usually ignorant of the domain, is responsible to choose, modify and integrate automatic and visual methods of analysis. Whereas such an approach is feasible and has resulted in a number of applications that can be used by non-technical users (as in Rudolph et al. (2009)), it is also constrained by several limitations.

The main issue with decoupling the data processing design from the data processing use is the loss of vital information in the visual translation. The most pernicious omission in this process is the suppression of prediction uncertainty and data quality. Dealing with uncertainty and trust in Visual Analytics is nontrivial because of the large amount of noise and missing values originating from heterogeneous data sources and bias introduced by automatic analysis methods, as well as human perception and intent (Thomas and Kielman 2009). To address this, data quality and algorithm confidence need to be appropriately presented. Users need to be aware of uncertainty and be able to read quality properties at any stage of the data analysis. Whereas several approaches have been researched under the umbrella of 'data wrangling' (Kandel et al. 2011), most of this work concentrates on preprocessing data (i.e. data entry, data (re)formatting, data cleaning, etc.). Little work has been done to visualise data quality and confidence in parallel to the outcomes of the analysis. Yet, this visualisation is vital to supporting the decision-making: as what constitutes an error is often context-dependent and requires human judgment of domain experts (Kandel et al. 2012), there is a need to research how such variables can be represented in parallel to outcomes of an iterative analysis process.

Visualizing data and process quality is not a topic that has been explored extensively in the visualisation literature (Kennedy et al. 2009). There has been research into specific techniques for uncertainty visualisation (Sanyal et al. 2009) and what uncertainty itself entails (Skeels et al. 2010), but it tends to concentrate on preprocessing data before such data is used in applications (Kandel et al. 2011), rather than looking at how to communicate quality indicators to support decision-making by domain experts.

As Visual Analytics is a complex process, providing insight and richer interaction capabilities for users to help them steer this process is a difficult task. Visual Analytics usually oversimplifies the complexity of analysing large amounts of real-world data. Important information about data quality and algorithm trade-offs are frequently hidden to reduce cognitive stress. However, if the analyst does not have access to this information, the use of the resulting visualisation is diminished. It is therefore essential to provide well-implemented representations of uncertainty that are meaningful and accurate, while being easy to visually process and understand.

But just as Visual Analytics is a bilateral affair with human and computer solving a problem together, so too is the notion of and reason for uncertainty: not only is the representation of uncertainty concerning data quality and prediction of great importance to the user, but the indication (and detection) of uncertainty regarding a user's intent and ultimate goal may be equally important to the computer. The user may have an agenda that the computer is unaware of, in addition to varying degrees of domain knowledge (Chapman & Chapman 1969), graph literacy (Pinker 1990), and visual bias (Hollands & Dyne 2000).

If the above factors were known to the machine, help could dynamically be offered and implemented into all parts of the analysis. For example: if the computer was aware of one's habits, agenda, and circumstances, it could dynamically change the weighting of various factors in a calculation, adapt the interface, or automatically highlight parts in the data relating to one's interests and goals. Only if both entities are clear about meaning, accuracy, and intent, a well-informed and accurate insight into a given problem may be gained. As a result, the determination of user intent and user ability – and therefore the reduction of "goal uncertainty" from the point of view of the computer – is likely to require at least three strands of research: First, the development of mechanisms that unobtrusively track user behaviour, before, during, and after the interaction with the application. Second, the efficient and precise extraction of user goals and biases from this information. Third, the implementation of the determined intent and bias into the application in a manner that supports both human and machine in the problem-solving process.

In addition to addressing the detection and representation of uncertainty between human and computer, we propose another aspect for inclusion in this research agenda: the communication of uncertainty between humans, in particular between data processing experts and domain experts. Whereas the former may be aware of certain problems or trends in the data based on its structure and quality, the latter may be more apt at interpreting the data and attributing results and observations to a certain problem and therefore at an advantage in the sense-making stage of the analysis. Uncertainties may exist for both groups not only in their own domain, but also in that of others. It is therefore essential to investigate how users can annotate and highlight different parts of the data and the analysis to communicate possible interpretative or structural uncertainties in a collaborative environment.

Finally, it is necessary to investigate which interaction modalities and strategies might be most suitable for supporting the above steps. This part of the agenda should explore how notions of uncertainty can be communicated between all entities. Although having been researched for decades, the communication between human and computer may still be regarded as inefficient, with input and output often encoded into a set of graphical elements, operated via peripherals or direct touch. Despite promising advances in conducting the human-computer dialogue (Jacob et al. 2008, Weigel et al. 2015, Seipp and Verbert 2016), it remains unclear how these can be utilised in the Visual Analytics process, especially with regards to the communication of uncertainty. If a visual language could be developed to comprehensively express uncertainty, which interaction modality would be most suitable for defining a consistent and meaningful dialogue between all stakeholders?

To help define the steps necessary to address the communication of uncertainty, the following section will provide more details on our proposed four-pronged agenda.

2 Research Agenda

Most previous and ongoing research is focused on developing advanced solutions for people with highly specialized skills (e.g. statistics, molecular biology, micro-economics) (Kandel et al. 2011). Little attention has been given to interface issues and support for users with limited knowledge of data processing, data analysis, information visualisation, or computer-aided problem solving in general. As a result, a research agenda should – at the very least – include challenges related to confidence and trust, user-system interaction, user-user interaction, and the inclusion of these into the analysis. We therefore propose four main strands of research:

1. Communicating uncertainty from the computer to the user.
2. Communicating uncertainty from the user to the computer.
3. Communicating uncertainty between data processing experts and domain experts.
4. Providing interaction mechanisms to allow both human and computer to express uncertainty and to react upon it.

The remainder of this section will briefly discuss scope and challenges of these aspects (Figure 1).

2.1 Communicating uncertainty from the computer to the user

This part oft he agenda comprises two essential areas oft the process in which uncertainty needs to be conveyed from the computer to the user: Uncertainties in the data and uncertainties in the model built using this data.

2.1.1 Visualising Data Quality

A central issue in Visual Analytics is the avoidance of misinterpretations due to uncertainty and errors in the input data. Therefore, data quality needs to be appropriately represented and the user be aware of these at any stage of the sense-making process. The work of Thomson et al. (2004) established a detailed typology for the limitations of data that affect certainty in predictive models: accuracy, precision, completeness, consistency, lineage, currency, credibility, subjectivity and interrelatedness. These types of limitations are usually defined at the dataset level and their uncertainty is usually propagated to the model built with that dataset.

We suggest this part of the strategy to encompass two main steps: The exploration of uncertainty in other domains and the determination of the correct visual representation of uncertainty for a particular problem. As for the choice of stylistic means to represent uncertainty, we intend to explore the approaches taken by fields with a much longer standing history than the field of Visual Analytics, such as geospatial visualisation or even gaming. We aim to investigate what can be learned and appropriated from these. As the field of Visual Analytics spans across a wide array of domains, we need to seek out knowledge that can be transferred to help create a widely applicable representation of uncertainty.

Regarding the determination of the correct representation of uncertainty for a certain problem, we need to look back at the rich history of information visualisation and investigate what visual means might work when and for what. Following this, aspects of domain specificity and user bias also need to be considered – different users have different expectations and graph literacy (Pinker 1990).

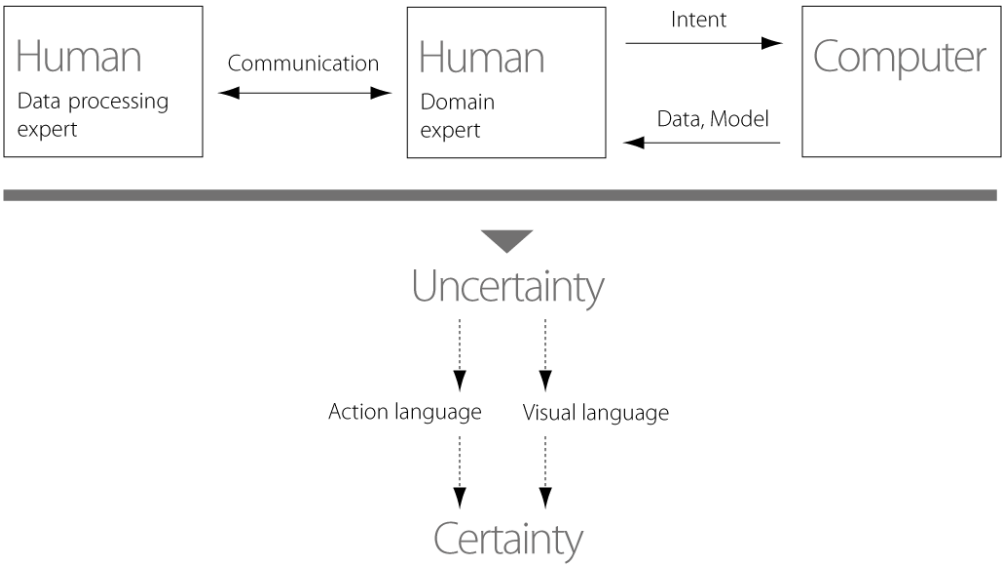

Figure 1: Types of uncertainty (communication uncertainty, intent uncertainty, data and model uncertainty) between the entities (data processing expert, domain expert, computer). Using a visual language supplemented by matching actions, uncertainty can be managed and its impact on decision time and decision quality can be reduced.

2.1.2 Visualising Model Quality

Perhaps the most obvious source of uncertainty introduced in any type of prediction is the imperfection of the predictive model. Such a model is built to take as input a group of predictor variables to produce a value. Given that models are only an approximation and simplification of reality, it is expected that the predicted values differ from real values to varying degrees. A whole area of statistics is devoted to measure the predictive power of different types of models. A good example of the measure of predictive power is the R-squared statistic used to score regression models. This measurement establishes what percentage of the variance in the real values of the predicted quantity is explained by the model. Different models usually have different predictive power depending on the predictor variables used, the type of algorithm and the amount and quality of data used to build them. Yet, just how well the model may fit the actual data can be hard to interpret. Kay et al. (2015) rightfully asked “How good is 85%”? Understanding the implications of the degree of the goodness-of-fit of a model to one’s data is a crucial point in the sense-making process. It is therefore necessary to determine a visual presentation of model uncertainty that is easy to

perceive and accurate in its depiction and interpretation for domain-experts. At the same time, the representation needs to be in-line with those used in representing data uncertainty in order to allow a continuous and coherent analytic process. The crucial task thus is to explore which visual means are suitable for representing uncertainty in both data and model visualisation, in isolation or in unison. It is our goal to create a comprehensive visual language that addresses these concerns. Therefore, a detailed review and analysis of user needs and expectations will be necessary to allow an optimal, task-oriented representation of uncertainty in data and model visualisations.

2.2 Communicating uncertainty from the user to the computer

When a user explores a data set, they may have a certain degree of domain knowledge (Chapman & Chapman 1969), graph literacy (Pinker 1990), or a predefined agenda. These aspects, however, are unknown to the computer, therefore representing a high degree of uncertainty concerning a user's capabilities and goals from the point of view of the machine. If the computer was aware of these, interface and calculation could be adapted and the experience and outcome improved. As a result, this part of the agenda deals with the detection of user context and bias and its implementation into the problem-solving process. Whereas the word 'bias' may be interpreted as having a negative connotation, we propose to see it as a chance to improve the bilateral data exploration and sense-making process. We suggest approaching this part of the agenda in three steps:

2.2.3 Logging user interaction and context

To better understand a user's background and intent, user behaviour may be tracked before, during, and after the operation of the analytics application. The tracking before the application operation may be done by harnessing location data, calendar data, or even email content. During the application operation, interactions are tracked by recording input from all devices. Post-application operation, further user behaviour may be tracked to detect possible actions taken as a result of the analytics session to validate potential assumptions about a user's goal. Yet, privacy concerns are paramount and need to be respected. An ethical evaluation may be as important as the evaluation of the technical feasibility.

2.2.4 Extracting and determining bias

This part represents the analytical process necessary to interpret the data of step one. It may comprise the determination of the correct algorithms and technical approaches as well as qualitative user studies to confirm assumptions and aid in building flexible user models.

2.2.5 Feeding the results into the analytical process

The results of logging and interpretation are fed back into the application and analytical process. It is important to explore how this can be done in an unobtrusive, yet controllable manner. Research regarding this part needs to consider aspects of visual representation and user acceptance, as well as the impact on any models built in the analysis.

2.3 Communicating uncertainty between data processing experts and domain experts

While the representation and detection of uncertainty is important in the human-machine dialogue, the human-human dialogue will also have to be considered, especially in a collaborative work setting. Further, this type of uncertainty communication should also encompass the dialogue between the designers of Visual Analytics tools and their end-users. The feedback loop between these types of actors could generate a continuous improvement process that will optimise the content and purpose of the application.

Research regarding this part of the agenda should focus on (but not be limited to) the creation and exchange of annotations or highlights using a variety of techniques. Following Buxton's (1986) example, rigorous studies are required to find intuitive and easy-to-use methods that facilitate this aspect.

2.4 Providing interaction mechanisms to allow both human and computer to express uncertainty and to react upon it

Finally, we need to explore interaction techniques that allow us to tie the above strands together and support the proposed visual language with a continuous and intuitive interaction design. Following the work concerning the communication of human-human expressions of uncertainty, this step will require an in-depth exploration of a wide array of input and output methods that help to smoothen the dialogue between the entities. Existing methods are largely non-intuitive and require significant expertise (Keim et al. 2010). Which methods, then, are suitable for the task at hand? Following a user-centred design approach we hope to define a set of techniques and methods that will bring together all steps of the process in harmony with is visual and non-visual counterparts.

3 Conclusion

This paper has discussed the need for researching the management of uncertainty in Visual Analytics. To do so, we propose an agenda that addresses the topic on four levels:

1. The communication of uncertainty from the computer to the user: this comprises the research of visual means for representing uncertainty in data and model accuracy.
2. The communication of uncertainty from the user to the computer: the computer tracks user behaviour before, during, and after the use of the application to gain certainty about the user's intent and abilities.
3. The communication of uncertainty between data processing experts and domain experts: this step will explore mechanisms for creating and displaying annotations in a collaborative environment as well as enabling the dialogue between creator and user of an application.

4. The provision of interaction mechanisms to allow both human and computer to express uncertainty and to react upon it: this step aims to define a set of intuitive interaction techniques that will combine the visual language with one of action to amalgamate all elements into a well-defined system.

Visual Analytics is a bilateral process between human and machine. By addressing the management of uncertainty from both perspectives, we hope to find a comprehensive solution that satisfies both entities and improves the decision-making process.

Acknowledgments

Part of this work has been supported by the Research Foundation Flanders (FWO), grant agreement no. G0C9515N, and the KU Leuven Research Council, grant agreement no. STG/14/019.

Bibliography

Bastin, L., Cornford, D., Jones, R., Heuvelink, G., Pebesma, E., Stasch, C., Nativi, S., Mazzetti, P. and Williams, M. (2013). Managing uncertainty in integrated environmental modelling: The UncertWeb framework. *Environmental Modelling & Software*, *39*, 116–134.

Bernard, J., Ruppert, T., Goroll, O., May, T., and Kohlhammer, J. (2012). Visual-interactive preprocessing of time series data. In Kerren, A. and Seipel, S., editors. *Proceedings of SIGRAD 2012, Interactive Visual Analysis of Data*, *volume 81 of Linköping Electronic Conference Proceedings*, Linköping University Electronic Press, 39–48.

Buxton, W. (1986). There's more to interaction than meets the eye: Some issues in manual input. *User centered system design: New perspectives on human-computer interaction*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 319–337.

Chapman, Loren J., and Jean P. Chapman (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of abnormal psychology* 74(3), 271.

Collins, A., Joseph, D. and Bielaczyc, K. (2004) Design Research: Theoretical and Methodological Issues. *Journal of the Learning Sciences*, 13(1), 15–42.

Duval, E. (2011, February). Attention please!: learning analytics for visualization and recommendation. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. New York: ACM, 9–17.

Hollands, J. G., & Dyre, B. P. (2000). Bias in proportion judgments: The cyclical power model. *Psychological Review*, *107*(3), 500–524.

Jacob, R. J., Girouard, A., Hirshfield, L. M., Horn, M. S., Shaer, O., Solovey, E. T., & Zigelbaum, J. (2008). Reality-based interaction: a framework for post-WIMP interfaces. *Proceedings of the SIGCHI conference on Human factors in computing systems*. New York: ACM. 201–210.

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., Weaver, C., Lee, B., Brodbeck, D. & Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4), 271–288.

Kandel, S., Parikh, R., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012). Profiler: Integrated statistical analysis and visualization for data quality assessment. *Proceedings of the International Working Conference on Advanced Visual Interfaces*. New York: ACM. 547–554.

Kay, M., Patel, S. N., & Kientz, J. A. (2015). How Good is 85%?: A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. New York: ACM. 347–356.

Keim, D., Andrienko, G., Fekete, J. D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). *Visual analytics: Definition, process, and challenges*. Springer Berlin Heidelberg. 154–175.

Keim, D. A., Kohlhammer, J., Ellis, G., & Mansmann, F. (Eds.). (2010). *Mastering The Information Age-Solving Problems with Visual Analytics*. Goslar: Eurographics Association.

Kennedy, J., Graham, M., Paterson, T., & Law, A. (2013, October). Visual cleaning of genotype data. *Biological Data Visualization (BioVis), 2013 IEEE Symposium on*. IEEE. 105–112.

Méndez, G., Ochoa, X., & Chiluiza, K. (2014, March). Techniques for data-driven curriculum analysis. *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*. New York: ACM. 148–157.

Pinker, Steven. (1990). A theory of graph comprehension. *Artificial intelligence and the future of testing*. Hillsdale, NJ: Lawrence Erlbaum Associates. 73–126.

Rudolph, S., Savikhin, A., & Ebert, D. S. (2009). FinVis: Applied visual analytics for personal financial planning. *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*. IEEE. 195–202.

Sanyal, J., Zhang, S., Bhattacharya, G., Amburn, P. and Moorhead, R.J. (2009) A User Study to Compare Four Uncertainty Visualization Methods for 1D and 2D Datasets. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), 1209–1218.

Shneiderman, B. (2002) Inventing discovery tools: combining information visualization with data mining. *Information visualization*, 1(1), 5–12.

Seipp, K., & Verbert, K. (2016). From inaction to interaction: concept and application of the null gesture. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, New York: ACM. 525–540.

Skeels, M., Lee, B., Smith, G., and Robertson, G.G. (2010) Revealing uncertainty for information visualization. *Information Visualization*, 9(1), 70–81.

Thomas, J, and Kielman, J. (2009). Challenges for visual analytics. *Information Visualization* 8(4), 309–314.

Verbert, K., Parra, D., Brusilovsky, P. and Duval, E. Visualizing recommendations to support exploration, transparency and controllability. *Proceedings of the 17th International Conference on Intelligent User Interfaces (IUI'13), IUI'13*, New York: ACM. 351–362.

Weigel, M., Lu, T., Bailly, G., Oulasvirta, A., Majidi, C., & Steimle, J. (2015, April). Iskin: flexible, stretchable and visually customizable on-body touch sensors for mobile computing. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, New York: ACM. 2991–3000.

Authors

Karsten Seipp

Karsten is a Postdoctoral Research Fellow at KU Leuven. He has been awarded his PhD by Goldsmiths, University of London for his research into mobile interaction design and techniques. He has won multiple international awards for his work at Imperial College London as Senior Digital Designer/Developer and is currently conducting research on Visual Analytics, affective computing, and embodied interaction.

Xavier Ochoa

Xavier Ochoa is a Principal Professor at the Faculty of Electrical and Computer Engineering at Escuela Superior Politécnica del Litoral (ESPOL) in Guayaquil, Ecuador. He is the coordinator of the Research Group on Teaching and Learning Technologies (TEA). He obtained his Ph.D at the University of Leuven in 2008 for his work on Learnometrics and has served in many coordination bodies in the field. More information at: http://ariadne.cti.espol.edu.ec/xavier

Francisco Gutiérrez

Francisco Gutiérrez is a researcher and doctoral candidate in computer science at the KU Leuven, Belgium. He is working towards a PhD degree in the field of Human-Computer Interaction. His interests include Visual Analytics in the decision-making process, and the visualisation of prediction systems with inherent uncertainty.

Katrien Verbert

Katrien Verbert is an assistant professor at the HCI research group of KU Leuven. She obtained a doctoral degree in Computer Science in 2008 at KU Leuven, Belgium. Her research interests include visualisation techniques, visual analytics, user interfaces for recommender systems, learning analytics and digital humanities.

Contact

Karsten Seipp, karsten.seipp@cs.kuleuven.be

Xavier Ochoa, xavier@cti.espol.edu.ec

Francisco Gutiérrez, francisco.gutierrez@cs.kuleuven.be

Katrien Verbert, katrien.verbert@cs.kuleuven.be