The role of data visualization in Railway Big Data Risk Analysis

M. Figueres-Esteban, P. Hughes & C. Van Gulijk
Institute of Railway Research, University of Huddersfield, Huddersfield, UK

ABSTRACT: Big Data Risk Analysis (BDRA) is one of the possible alleys for the further development of risk models in the railway transport. Big Data techniques allow a great quantity of information to be handled from different types of sources (e.g. unstructured text, signaling and train data). The benefits of this approach may lie in improving the understanding of the risk factors involved in railways, detecting possible new threats or assessing the risk levels for rolling stock, rail infrastructure or railway operations. For the efficient use of BDRA, the conversion of huge amounts of data into a simple and effective display is particularly challenging. Especially because it is presented to various specific target audiences. This work reports a literature review of risk communication and visualization in order to find out its applicability to BDRA, and beyond the visual techniques, what human factors have to be considered in the understanding and risk perception of the information when safety analysts and decision-makers start basing their decisions on BDRA analyses. It was found that BDRA requires different visualization strategies than those that have normally been carried out in risk analysis up to now.

1 INTRODUCTION

Big Data technologies currently offer industry and academia opportunities to acquire a detailed understanding of complex processes. Although the origin of big data is not very clear (Gandomi & Haider 2015), big data is considered as the fourth generation of decision support data management (Watson & Marjanovic 2013). IEEE defines big data as:

"*…a collection of very huge data sets from which it is practically impossible to analyze and draw inferences… Big Data usually has a multidimensional structure and can be characterized by the 5V's: Volume, Velocity, Variety, Veracity and Value."* (Attohokine 2014).

Big data for risk analysis (BDRA) might be a great alley for the further development of risk models in the railway transport. Currently railways handle a huge quantity of information from different types of data sources (e.g. unstructured text, signaling or train data streams) that might be used to improve the understanding of risk factors involved in railways, to detect missing relationships and to identify new risks, or even to assess the risk level more accurately. Also, different domains may be targeted for data-analytical techniques such as maintenance, rail and infrastructure, incidents and third party involvement (Van Gulijk et al. 2015). For the efficient use of BDRA, the conversion of huge amounts of data into a simple and effective display is particularly challenging. Key questions include: what is the question required from big data analysis (what do we need to know), how to select and represent information from huge data sources (what the data shows) and what the data actually convey to the audience (what the audience perceives).

This paper focuses on a literature review that investigates the requirements for visualisation for BDRA. It treats both the visual techniques, and what lies beyond: human factors for understanding risk data and risk perception. An understanding of human factors is essential in order to provide risk visualisations that take into account the inherent biases of, and heuristics used by, safety analysts and decision-makers.

The review ends with a set of requirements to help to define the foundations of visualisation in BDRA applied to railway transport.

2 METHOD

Literature search

The literature search started looking for the topics *risk communication*, *visualisation* and *big data* in Scopus, Mendeley and Google Scholar. These main topics were combined with the more frequent keywords found in the first search. In addition, we used the traditional 'snowball method' to find new interesting articles from the identified papers. Since we wanted to gather the maximum points of view, grey literature was also assessed. The type of items selected were articles from journals (including scientific journals), books, proceedings or technical reports.

As a result of this iterative process more than 400 documents were gathered. Figure 1 shows the main areas for BDRA visualisation: data management, data analysis, human-computer interaction, risk communication and visualisation. Each topic is described in this paper.

3 VISUAL ANALYTICS

Keim et al. (2008) define visual analytics as a combination of "*...automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets.*".

Visual analytics is a multidisciplinary area that often attempts to obtain insight from massive, incomplete, inconsistent and conflicting data in order to support assessment and decision-making. It drives the analysis process and supports information communication to target audience but usually requires human judgment (Thomas & Cook 2005; Keim et al. 2008).

Visualisation associated with railway BDRA requires not only representation of a great amount of information but also the ability to support complex analyses, handling data from different types of sources and diverse levels of quality.

Note that railway BDRA is not a "one user type" tool. The potential users will have very different skills and knowledge about railway safety, and depending on their understanding goals, the interaction with the system and the way for representing and visualising the results might be different. Hence, visualisation in railway BDRA is a combination of data management, data analysis, human interaction and risk communication to making decisions, that is, visual analytics.

Requirement 1: Railway BDRA visualisation should be addressed as a visual analytic problem. Visualisation has to support railway risk analysis in data management, data analysis, human interaction and risk communication processes.

4 DATA MANAGEMENT

Type of data is one of the dimensions to consider in Big Data Analysis. Zhang (2013) defines five

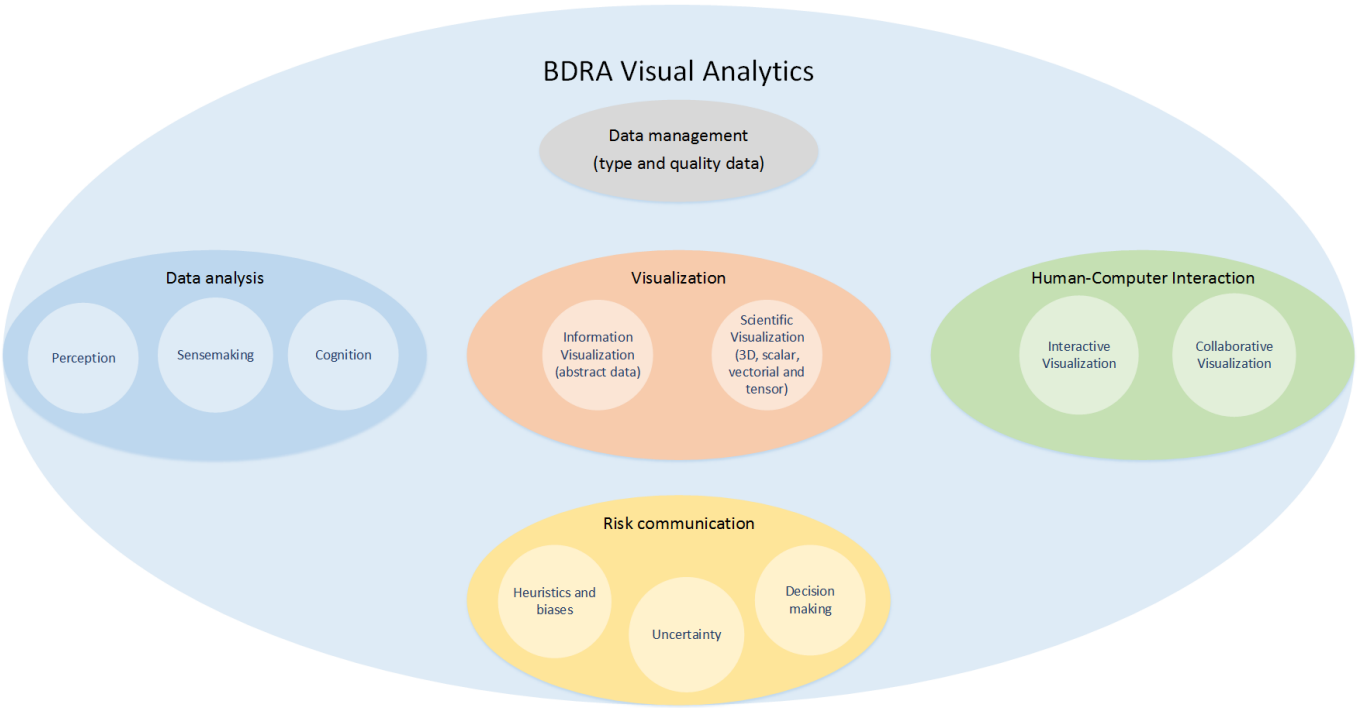

Figure 1. Main areas involved in BDRA visualisation.

sources of data: image, video, audio, alpha-numeric, semi-structured and unstructured. The combination and transformation of all these types of data sources in a suitable form is the first step to start visual analysis. However, different types of inconsistencies and uncertainties arise depending on the type of data and data fusion strategies (Naumann & Bleiholder 2006; Dong et al. 2009; Zhang 2013). Combining data management, data analysis and visualisation in order to visualise and assess these inconsistencies and uncertainties as part of the visual analysis by human judgment might support the data analysis to achieve best results (Keim et al. 2008).

The diverse actors involved in railways (train and freight operators, infrastructure managers and national/international transport and safety bodies *inter alia*) handle different types of data sources independently: e.g. free text from close call events/near misses or accident investigation reports; videos and images from security organisations; data feeds from train movements, signaling systems or train recorders; or semi-structured data from accidents or incident databases. Unstructured data sources (e.g. close call text descriptions) present different type of ambiguities on a lexical, syntactic and semantic level (Hughes et al. 2015). Therefore, matching objects such as places, codes or time from this database to other data sources might be problematic.

Efficiently integrating and sharing diverse types of data, visualising the inconsistencies and uncertainties derived from this process is a challenge to solve in railway BDRA.

Requirement 2: Introduce visual analysis in the data management techniques in order to support the uncertainty of data analysis.

5 DATA ANALYSIS

Data analysis or data mining "*is the investigative process used to extract knowledge, information and insight about reality by examining data*" (Grolemund & Wickham 2014). Data analysis research in visual analytics has mainly been focused on developing automatic statistical techniques and machine learning techniques to extract valuable information or reveal complex relationships, even from large volumes of data (Keim et al. 2008). Nevertheless, the use of complex computer software to analyse large data sets might affect the level of control and understanding of users, reducing their performance. Moreover, representing large data sets might introduce limitations such as occlusion of data, disorientation or misinterpretation (Shneiderman 2002). Thus technology and large data sets might not guarantee useful results (Grolemund & Wickham 2014).

Although cognitive science has been present since the early literature of exploratory data analysis for improving data analysis, new scientific models that explain the data analysis process are proposed. Grolemund & Wickham (2014) point out that data analysis is a sensemaking task, which has the goal of creating "*reliable ideas of reality from observed data*". Sensemaking is subjective analysing results, it is common to see that people with different expertise achieve different conclusions. However, if sensemaking is augmented, it would be possible to obtain more objective outcomes. In this regards, visualisation may help data analysts increase their perceptual abilities, extend the work memory and assist in retrieval from long-term memory (Patterson et al. 2014).

Data analysis in railway BDRA should be a combination of visualisation, automated data analysis and human interaction which could support the exploratory analysis and the choice of statistical tests under supervision of human judgment. A first assisted visual analysis from the system should help the data analyst to determine the analysis strategy. The user should be able to choose any type of statistical test, independently of the type of data, allowing the user control to see what the system does. Results should be presented suitably to increase sensemaking and improve understanding from data. Thus cognitive aspects should be considered in visual data analysis process.

Suitable visualisation techniques under human supervision might support BDRA data analysis, for example, supporting clustering techniques applied to unstructured data in order to determine or detect new categories of possible events that might help to explain and understand safety issues.

Requirement 3. Data analysis should be a combination of automated data analysis, human interaction and visualisation. Visualisation has to support cognitive processes in data analysis.

6 HUMAN-COMPUTER INTERACTION

It is challenging to find a clear definition of "*interactive visualisation*". Yi et al. (2007) define the interaction component of an information visualisation systems as "*the dialog between the user and the system as the user explores the data set to uncover insight*". The authors point out that the interaction with visual representations has played a secondary role in the information visualisation research field, with the new types of representations being the area that has received the most part of attention. Nevertheless, the dialog between the data analyst and the representations in order to interpret and make sense from the results, thinking the next question to request the system considering new factors or varia-

bles, is essential to visual analytics (Thomas & Cook 2005).

Different attempts to develop task taxonomies related to visualisation has been developed (Yi et al. 2007). Shneiderman (1996) summarized his task taxonomy as the Visual Information Seeking Mantra "*overview first, zoom and filter, then details-on-demand*". More recent studies have focused on "*what the user want to achieve*" or "*insight provenance*" (Yi et al. 2007; Gotz & Zhou 2009).

On the other hand, the real world problems that analysts and decision-makers have to face are increasing in complexity, size and uncertainty. Realistic problems from huge amount of data might require broad experience, diverse knowledge domains and a number of dedicated people to solve them. Although traditional visual analytics tools have been developed for interacting with a single user, collaborative visualisation is an emerging field intended to consider relationships detected by users with different knowledge domains in an interactive and collaborative data analysis, transferring that expertise into a computational system (Keel 2006; Isenberg et al. 2011).

Human-computer interaction (HCI) in Railway BDRA should be a combination of interactive and collaborative visualisation. The first to support and improve the understanding of the data analysis process, providing the maximum level of transparency in the risk analysis. The second to make easier the introduction of human knowledge to the system and the interpretation of results. Railway safety analysts may belong to different organisations (e.g. train operators, infrastructure managers or safety bodies) and handle different safety information (e.g. safety measures applied in railways). A collaborative visual analysis in railways might be particularly effective in the interpretation and understanding of results, for example easily detecting temporal effects derived from new measures applied in some organization.

Requirement 4: Visualisation of railway safety has to support interaction with data and simultaneous collaborative analysis with several users.

7 RISK COMMUNICATION

Visualisation has performed an important role in the communication of risk. Mass media and health communication researchers have often used visual representations to enhance the understanding of numerical risk information, especially in small probability events, although few experimental researches have tested how visualisation risk affects perceived risk, decision-making and behavior of people (Lipkus & Hollands 1999).

A vast amount of literature has been written about particular tasks of risk communication, focused primarily on risk communication strategies, assessment and maximization of the effectiveness of graphs in the risk communication process, individual differences and psychological principles that influence risk perception and decision-making and the communication and visualisation of uncertainty (Tversky & Kahneman 1974; Tversky & Kahneman 1981; Fischhoff 1995; Lipkus & Hollands 1999; Bier 2001b; Gigerenzer 2003a; Gigerenzer 2003b; Frewer 2004; Peters 2008; Visschers et al. 2009; Spiegelhalter et al. 2011; Xie et al. 2011; Peters 2012)

The book *Illuminating the Path: The Research and Development Agenda for Visual Analytics* breaks down the communication tasks related to visual analytics into production, presentation and dissemination, with the aim of conveying "*analytical results in meaningful ways to a wide variety of audiences, including peers, decision makers,…"(p. 137).* However, little attention in risk communication to regulators and decision-makers have been received in comparison with the general public (Bier 2001a)

Railway BDRA should support the risk communication process at all levels and safety audiences. Depending on the type of audience (e.g. decision-makers, data analysts, railway staff or general public), the information should be depicted in a specific manner. For example, for decision-makers and data analysts, the "risk dashboard" would include additional and detailed representation of information related to uncertainty of data; for railway staff and the public, the information represented should improve the understanding of hazards (e.g. improving the way to communicate low frequency events).

In addition, Railway BDRA might be effective in the management of emergency situations or giving public warnings to the security bodies.

Requirement 5: The risk communication process should be tailored to each type of audience involved in safety. Suitable visualisation techniques to represent risk data and uncertainty should be used to reduce the biases and improve the understanding of information.

8 VISUALISATION

Since the research in visualisation has been developed from different disciplines (e.g. computer science, engineering, psychology or management sciences), its definition has been expressed in terms that might indicate different levels of abstraction and understanding depending on the discipline, generating conflicts and inconsistencies (Chen et al. 2009). Although we can find references related to the understanding and insight of data by means of visual perception to support the cognitive process in data analysis (Tukey & Wilk 1966; Card et al. 1999; Heer

et al. 2010; Van Wijk 2005), visualisation has historically been divided into *information visualisation* and *scientific visualisation*. *Scientific visualisation* would develop visual techniques to depict scientific and spatial data, whilst *information visualisation* represents abstract and non-spatial data (Tory & Möller 2004).

In both areas important attempts have been carried out to define cognitive models related with graphical perception, principles of design for graphical excellence, taxonomies to classify the "zoo of visual representations" or even classify the visualisation problems (Bertin 1983; Tufte 1983; Cleveland & McGill 1984; Wehrend & Lewis 1990; Shneiderman 1996; Lohse 1997; Card et al. 1999; Tversky, B., Morrison, J. B., & Betrancourt 2002; Heer et al. 2010).

Systems that receive data and convey them effectively and automatically ("automatic presentation systems") have been developed (Mattis & Roth 1990; Mackinlay 1986) with a great influence from the Stevens data classification (nominal, ordinal, interval and ratio), that at the same time indicated what type of statistical analysis was permissible (Stevens 1946). However, criticism of Stevens' work has been focused on the restriction of the choice of statistical methods, its adaptation to the real-world data and the degradation of data (Velleman & Wilkinson 1993).

Visualisation is the core of visual analytics and is one of the corner stones of railway BDRA. As we have shown above it has a strong relationship with data management, human computer-interaction, data analysis and risk communication.

Requirement 6: Develop an automated visualisation system to support visual analytics involved in railway BDRA (data management, human computer-interaction, data analysis and risk communication).

9 CONCLUSION

Nowadays the railway industry has available different types of vast data sources that might be used in railway safety to develop new risk analysis models. Nonetheless there is no way that humans are able to handle and comprehend these vast amounts of data, automated intelligence is necessary.

Railway Big Data Risk Analysis (BDRA) is a decision support system with the purpose of supporting users of different knowledge domains to select suitable data and data sources, conduct risk data analysis and achieve best results and comprehension from them. Visualisation techniques are key to capturing the right lessons from railway BDRA. In this paper we have described a first approach of the applicability of visualisation in railway BDRA.

A literature review has been carried out to identify relevant documents that guide us to define the foundations of visualisation in railway BDRA. The first outcome obtained is that railway BDRA supported by visualisation is a visual analytics problem: determine the visual requirements of railway BDRA means to determine the visual requirements of each process involved in visual analytics (data management, data analysis, human-computer interaction and risk communication).

From relevant documents found we have proposed different requirements for railway BDRA visualisation. Since railway BDRA deals with diverse range of railway users it has to be tailored to those users' needs.

Each type of user would need a different a visual environment to interact with the data sources and gain insight from railways. Moreover, railways would need a risk analysis tool bespoke to their own organizational structure. All safety data analysts from the diverse railway actors should be able to participate or supervise a visual risk analysis, adding value to the process.

Finally, we highlight that it is necessary to visualise and communicate risk information suitably to all the safety audiences in order to involve them in the safety process to improve the quality of data analysis. As we described above, the risk communication process should reach out to railway staff and members of the public to improve their comprehension of railway hazards. That might imply a better feedback from these audiences, for example, increasing the number and quality of close call/near misses reported.

10 REFERENCES

Attoh-okine, N., 2014. Big Data Challenges in Railway Engineering. *2014 IEEE International Conference on Big Data*, pp.7–9.

Bertin, J., 1983. Semiology of graphics: diagrams, networks, maps. University of Wisconsin Press

Bier, V.M., 2001a. On the state of the art: Risk communication to decision-makers. *Reliability Engineering and System Safety*, 71(2), pp.151–157.

Bier, V.M., 2001b. On the state of the art: Risk communication to the public. *Reliability Engineering and System Safety*, 71(2), pp.139–150.

Card, S.K., Mackinlay, J.D. & Shneiderman, B., 1999. Readings in Information Visualization: Using Vision to Think. In *Information Display*. p. 686.

Chen, M. et al., 2009. Data, Information, and Knowledge in Visualization. *Computer Graphics and Applications, IEEE*, 29, pp.12–19.

Cleveland, W.S. & McGill, R., 1984. Graphical Perception: Theory, Experimentation, and Application to the

Development of Graphical Methods. *Journal of the American Statistical Association*, 79, pp.531–554.

Dong, X.L., Halevy, A. & Yu, C., 2009. Data integration with uncertainty. *VLDB Journal*, 18, pp.469–500.

Fischhoff, B., 1995. Risk perception and communication unplugged: Twenty years of process. *Risk Analysis*, 15(2), pp.137–145.

Frewer, L., 2004. The public and effective risk communication. *Toxicology Letters*, 149(1-3), pp.391–397.

Gandomi, A. & Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), pp.137–144.

Gigerenzer, G., 2003a. *Reckoning with risk: learning to live with uncertainty*, Penguin UK.

Gigerenzer, G., 2003b. Why Does Framing Influence Judgment? *Journal of General Internal Medicine*, 18(11), pp.960–961.

Gotz, D. & Zhou, M.X., 2009. Characterizing users visual analytic activity for insight provenance. *Information Visualization*, 8(1), pp.42–55.

Grolemund, G. & Wickham, H., 2014. A Cognitive Interpretation of Data Analysis. *International Statistical Review*, 82(2), pp.184–204.

Heer, J., Bostock, M. & Ogievetsky, V., 2010. A tour through the visualization zoo. *Communications of the ACM*, 53, p.59.

Hughes, P., Van Gulijk, C. & Figueres-Esteban, M., 2015. Learning from close call. *SPARK*.

Isenberg, P. et al., 2011. Collaborative Visualization: Definition, Challenges, and Research Agenda. *Information Visualization*, 10, pp.310–326.

Keel, P.E., 2006. Collaborative visual analytics: Inferring from the spatial organization and collaborative use of information. In *IEEE Symposium on Visual Analytics Science and Technology 2006, VAST 2006*. Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, pp. 137–144.

Keim, D. et al., 2008. Visual analytics: Definition, process, and challenges. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 154–175.

Lipkus, I.M. & Hollands, J.G., 1999. The visual communication of risk. *Journal of the National Cancer Institute. Monographs*, 27701, pp.149–163.

Lohse, G.L., 1997. Models of graphical perception. *Handbook of Human-Computer Interaction*, 2, pp.107–135.

Mackinlay, J., 1986. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5, pp.110–141.

Mattis, J. & Roth, S.F., 1990. Data Characterization for Intelligent Graphics Presentation. *In Proceedings of the Conference on Human Factors in Computing Systems (SIGCHI '90*, pp.193–200.

Naumann, F. & Bleiholder, J., 2006. Data Fusion in Three Steps : Resolving Schema , Tuple , and Value Inconsistencies. *IEEE Data Engineering Bulletin*, 29, pp.21–31.

Patterson, R.E. et al., 2014. A human cognition framework for information visualization. *Computers & Graphics*, 42, pp.42–58.

Peters, E., 2012. Beyond Comprehension: The Role of Numeracy in Judgments and Decisions. *Current Directions in Psychological Science*, 21(1), pp.31–35.

Peters, E., 2008. Numeracy and the perception and communication of risk. In *Annals of the New York Academy of Sciences*. pp. 1–7.

Shneiderman, B., 2002. Inventing Discovery Tools: Combining Information Visualization with Data Mining. *Information Visualization*, 1, pp.5–12.

Shneiderman, B., 1996. The Eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*. Univ of Maryland, College Park, United States: IEEE, pp. 336–343.

Spiegelhalter, D., Pearson, M. & Short, I., 2011. Visualizing uncertainty about the future. *Science*, 333(6048), pp.1393–1400.

Stevens, S.S., 1946. On the Theory of Scales of Measurement. *Science*, 103(2684), pp.677–680.

Thomas, J.J. & Cook, K.A., 2005. *Illuminating the path: The research and development agenda for visual analytics*. IEEE press.

Tory, M. & Möller, T., 2004. Rethinking visualization: A high-level taxonomy. In *Proceedings - IEEE Symposium on Information Visualization, INFO VIS*. pp. 151–158.

Tufte, E.R., 1983. The visual display of quantitative information. Graphics press

Tukey, J.W. & Wilk, M.B., 1966. Data analysis and statistics: an expository overview. In *Proceedings of the November 7-10, 1966, fall joint computer conference*. ACM, pp. 695–709.

Tversky, a & Kahneman, D., 1981. The framing of decisions and the psychology of choice. *Science (New York, N.Y.)*, 211, pp.453–458.

Tversky, A. & Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157), pp.1124–1131.

Tversky, B., Morrison, J. B., & Betrancourt, M., 2002. Animation : can it facilitate ? *International journal of human-computer studies*, pp.247–262.

Velleman, P.F. & Wilkinson, L., 1993. Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading. *The American Statistician*, 47, pp.65–72.

Visschers, V.H.M. et al., 2009. Probability information in risk communication: A review of the research literature. *Risk Analysis*, 29(2), pp.267–287.

Watson, H.J. & Marjanovic, O., 2013. Big Data: The Fourth Data Management Generation. *Business Intelligence Journal*, 18(3), pp.4–8.

Wehrend, S. & Lewis, C., 1990. A problem-oriented classification of visualization techniques. *Proceedings of the First IEEE Conference on Visualization: Visualization `90*.

Van Gulijk, C. et al., 2015. Big Data Risk Analysis for Rail Safety? In *Proceedings of the 25th European Safety and Reliability Conference, ESREL 2015*.

Van Wijk, J.J., 2005. The value of visualization. In *Proceedings of the IEEE Visualization Conference*. p. 11.

Xie, X.-F. et al., 2011. The Role of Emotions in Risk Communication. *Risk Analysis*, 31(3), pp.450–465.

Yi, J.S.Y.J.S. et al., 2007. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(August).

Zhang, D., 2013. Inconsistencies in big data. In *Proceedings of the 12th IEEE International Conference on Cognitive Informatics and Cognitive Computing, ICCI*CC 2013*. pp. 61–67.