The Potential for Using Big Data Analytics to Predict Safety Risks by Analysing Rail Accidents.

Dr Howard Parkinson CEng FIMechE MIRSE
Rail System Engineering Limited
Dr Gary Bamford FBCS, CEng, CITP
NTTX Advisory Limited

RESEARCH ABSTRACT
The hallmark of modern safety analysis in the railway is that we attempt to predict accidents before they happen to make sure that the correct mitigations are in place. The industry has gone from a reactive approach to safety in the early days of rail to this proactive mind-set [1].

With modern powerful computing and the explosion in data availability, from ever expanding sources, there may be opportunities to use a Big Data (BD) approach to flag up high risk scenarios on the railway before accidents occur. For example, in the UK there are already BD initiatives for railway risk assessment research [5] that are taking into account data available from signalling systems, asset condition monitoring, safety incident reporting and even social media. There is a lot of hype around the topics of BD and ‘analytics’. This paper presents an approach that is grounded in the knowledge of railway operations and systems, together with a study of these new data initiatives, with an aim of developing an approach to using them to help in accident prevention.

To begin to understand how this area of research might take shape, this paper will review several serious accidents from around the world that have happened in the last twenty years including those at Santiago de Compostela and Platja de Casteldefels in Spain, Ladbroke grove, Potters Bars and Hatfield in England, Whenzou in China, Waterfall in Australia, Eschede in Germany, and finally Fox River Grove and Chatsworth in the USA. These accidents have been chosen because they are high profile with multiple causes that potentially might have been flagged up as high risk by the analysis of data.

The analysis of accidents takes into account the importance of human factors, management and systems, social media, news, issues and the increasing complexity resulting from, for example, the increased use of software and the disaggregation of railway system management. Failings in the design of processes, standards, rules and procedures, human factors, failings of equipment and degraded modes could also be identified.

It will be necessary to envisage that the current level or potential level of data availability, computing power and intelligent algorithms were around when the accidents occurred. The accidents will be reviewed in the light of the data situation now and in the next few years. It is known that accidents very rarely have one simple cause and usually have multiple causes such as management, environment and design, to name a few [4]. The accidents will be reviewed and critically appraised by analysing their causes in depth using a systems analysis approach [3][2]. The approach will look at;

1. What would have needed to be known to avoid the accident,
2. What type of analysis would have been needed to identify the hazards,
3. What data would be needed to support this analysis?
27-May-15
It will then be assessed what possible data might be relevant to the causes of accident and a matrix developed linking the potential data source to the accident. This matrix will give the bounds to a BD analytics approach to identifying high safety risk in terms of potential data sources.

The typical identifiers of a BD approach are the 5 Vs, Volume, Velocity, Variety, Veracity and Value [6]. The approach discussed here and the intended use of data satisfies all these BD criteria. There are various places to look for current thinking in BD and the UK Government are sponsoring research [7], but companies like Google and online retailers such as Amazon are leading the way in the state of the art in BD. A good example is how Google assessed data in the flu outbreak analysis and predictions [8] of 2009.

The crucial requirement is bounding the available data and then turning the vast data into safety information taking into account the fact that correlations and causations will sometime be counterintuitive. It will be necessary to envisage a machine learning algorithm which combines data mining, artificial intelligence,, pattern recognition, neural networks, clustering, and data visualization. This will be one of the major challenges; however there are commercially available systems such as Mahout and Rapid Miner [8] which may be suitable. Interconnected computer connectivity is available for the sheer processing power with, for example, the HADOOP system [8].

At recent BD risk assessment symposium the following quote was made “Instead of constructing a pre-determined structure for risk calculations, such as fault trees, different sources of safety-relevant information are brought together to find answers to specific safety questions, to identify new trends and to identify new threats to safety,” [5]. It is not intended to develop a BD system that will actually predict an accident but to present an approach that will flag up high risk scenarios based upon the flow of data from the sources identified in the matrix.

The deliverables from the research will lead to a greater understanding or the safety data that are available, the methods that are needed to make sense of it and the unforeseen correlation and causations that may be lurking within the data that we can learn from. The main output will be an understanding of the potential data sources linked to the multiple causes of serious accidents in a usable matrix that will facilitate BD analytics for major rail accident scenarios. This will provide the potential to identify near misses in the future and also potentially generate real time risk contour maps for certain hazards.

References
7. UK Government, The Parliamentary Office of Science and Technology Number 472 July 2014, Big and Open Data in Transport
8. Big Data Analytics (S183) RSSB November 2014 Version 2