Identifying intelligence links in threat networks through machine learning on explosives chemical data

Simon Crase [a], Benjamin Hall [b] and Suresh Thennadil [a]

[a] Charles Darwin University, College of Engineering, IT & Environment, Darwin, Northern Territory,
[b]Defence Science and Technology Group, Weapons and Combat Systems Division, Edinburgh, South Australia
Email: simon.crase@cdu.edu.au

Abstract: Improvised Explosive Devices (IEDs) and the terrorist or threat networks that employ them pose an ongoing threat in military operations. A significant challenge is identifying the intelligence linkages and relationships between the individuals that form these threat networks. However, this information is essential if these networks are to be disrupted.

This paper presents a novel concept for identifying these network linkages that can complement the threat network understanding generated through traditional military intelligence means.

In searching for opportunities to develop additional intelligence through scientific research, it was identified that the improvised nature of IEDs introduces characteristics that may be unique to each bombmaker. Improvised devices are made by individuals (not a production factory) so their construction, components and characteristics vary based on the maker. Based on the assumption that a bombmaker will regularly make IEDs in the same way (often the way they have been trained to make them), there is the opportunity to identify matching IEDs that have been made by the same maker, creating links between a person and multiple IEDs or attacks. Similarly, there may be common construction characteristics between different bombmakers IEDs enabling linking bombmakers together through their training, construction techniques or materials.

To exploit this opportunity, this research utilises the application of data science and machine learning techniques to analyse chemical test data from recovered samples of explosives, with the aim of identifying matches and relationships between the samples. Previously, forensic chemists have demonstrated the ability to identify matches between explosive samples through detailed chemical analysis. However, this analysis was a manual and time-consuming process using advanced chemical testing techniques and could not be applied at a large scale. The use of data science aims to reduce the need for advanced testing and enable rapid analysis of large data sets.

The methodology presented combines machine learning clustering techniques with traditional chemometric techniques for analysing chemical test data. The process can be summarised as follows:

1. Data pre-processing is used to optimise the data for clustering analysis
2. Principal Component Analysis (PCA) is used to reduce the dimensionality of the data and provides a way of visualising the clustering in 2-dimensions
3. Unsupervised machine learning algorithms then assign the explosive samples into clusters
4. Evaluation (validation) of clustering results and confirmation of the number of clusters is achieved through application of internal and external evaluation indices.

The results presented demonstrate the feasibility of using this machine learning centred approach for matching samples of unknown explosives that could be made by the same bombmaker.

Keywords: *Clustering, unsupervised machine learning, explosives, spectroscopy, intelligence*

1. INTRODUCTION

Improvised Explosive Devices (IEDs) are bombs that are made or used in non-conventional ways, hence the *improvised* nomenclature. While not a new threat, the prominence of IEDs and the threat networks that utilise them has grown in recent decades to pose a significant threat to military and civilian communities across the globe. The 2013 Australian Defence White Paper (Australian Government 2013) recognises the fact that improvised explosive devices are now a part of the future operating environment of the Australian Defence Force (ADF).

Over the last decade, significant work has been done in detecting IEDs with the aim of defeating threats and protecting forces from IEDs. However, it remains an unresolved challenge due to the diversity of the devices and the constant evolution from bombmakers to mitigate counter IED efforts. There is no 'silver bullet' or single piece of technology that has been able to solve it. However, every bit of effort contributed to this domain builds incremental capability and assists overall.

One significant line of effort is countering the threat networks that utilise IEDs and improvised threats. As per the US Joint Improvised Threat Defeat Organisation (2018), "*Lessons learned have shown the importance of taking a wider look than just the device itself to the entire realm of improvised threats as we see and sense who makes and employs them, and how.*" While this focus on the network is not new, it is becoming increasingly important as other lines of effort in technologies to defeat the threat and counter-IED training have proved insufficient to overcome this threat. The Attack the Network line of operations utilises *intelligence* to identify, understand and target the terrorist, insurgent or threat networks.

This paper presents a novel concept for identifying these network linkages that can complement the threat network understanding generated through traditional military intelligence means.

2. CONCEPT

Considerable thought was given to identify opportunities where research and innovation could be applied in this space including discussions with the Defence Science and Technology Group (DST) scientists who conduct IED exploitation and counter IED work. One key point that was identified is the fact that the improvised nature of IEDs introduces characteristics that are unique to each bombmaker. Improvised devices are made by individuals, so their construction, components and characteristics vary based on the maker. Based on the assumption that a bomb-maker will regularly make IEDs in the same way (often the way they have been trained to make them), there is the opportunity to identify matching IEDs that have been made by the same maker, creating links between a person and multiple IEDs or attacks.

These links could provide information (intelligence) that is new, may not be obtainable from other forms of intelligence and can contribute to the understanding of the links within a threat network. Of specific interest here is the network and relationship between people, objects, actions, places and events. Examples include bombmakers, suppliers, communications, attacks, cache finds, and materials. Understanding the linkages between these enables understanding of the threat network and potential opportunities to disrupt it.

2.1. Potential Outcomes

There are multiple types of linkages or intelligence that could be gained through analysing and matching the construction characteristics of IEDs.

Direct Attribution

One output from matching of IED samples is the direct attribution of an individual to multiple IED events. If multiple recovered IED samples are an exact match (batch matching), then it is reasonable to assume that they were made by the same bombmaker. This allows linking of a single individual to multiple IEDs. If the identity of the individual responsible for one of these events is known, then they can be linked to all of them. Expanding on this direct attribution to an individual; if there are direct matches to different types of IEDs within the data set, this can imply that there are multiple bombmakers operating and the number of bombmakers within the dataset. If further metadata about events is included such as location and time, this may enable understanding of the bombmakers areas of operation, or the temporal evolution of a bombmakers operations.

Group Attribution

One level up from direct matching of IEDs is when common characteristics can be identified but there are enough differences in the IED construction to assume they were made by different bombmakers. These common characteristics may enable linking those bombmakers together. Common characteristics could be the same recipe for the explosives or the same design for the IED trigger switch. This may indicate the bombmakers were trained together or trained by the same person, linking them to the same threat group. Similarly, the use of the same (unique) materials in an IED may link the bombmakers to a common supplier of materials.

Source Attribution

The broadest level of attribution would be the identification of a distinct ingredient or component across a significant number of samples (or all samples). This would indicate a common use of a material across a large portion of the threat network and a potential vulnerability of that network. If that ingredient or component can be distinctly identified, the source of that material may be able to be identified and disrupted.

2.2. Sample Matching Opportunities

Data on IEDs is obtained through an activity called IED exploitation. Based on the opportunity identified above to match IED samples, applications may include:

- Explosives matching: Matching the chemical composition in the main charge or detonator
- Electronic circuitry matching: Matching the design, construction, layout and materials used in electrical and electronic circuitry typically incorporated into the IED switch.
- Communications configuration matching: Radio controlled IEDs contain many configurable components that can be decided or adjusted by the bombmaker.
- Machining and materials matching: The matching of tooling marks on housings or components, the machinery used to make the IED, and the materials used in the construction.
- A combination of all the above.

In reviewing these opportunities, explosives matching was chosen for this research due to the availability and unclassified nature (when metadata is removed) of the test data resulting from the IED exploitation.

3. EXPLOSIVES EXPLOITATION AND DATA MATCHING

The explosives used for IEDs are predominantly homemade using commercial precursors such as fertilisers, household chemicals and supplies, or industrial chemicals and materials (National Academies of Sciences Engineering and Medicine 2018). During IED exploitation, there is a suite of tests that could be conducted depending on the explosives used and the detail required about the chemical composition. These include spectroscopy and spectrometry, chemical reaction tests, physical characterisation, and explosive behaviour.

In reviewing literature on matching explosive samples, the only directly relevant work identified was previous research conducted at Flinders university by McCurry (2015) in collaboration with DST. This research focused on the chemical testing techniques that could be applied to batch matching and source matching of homemade explosives. It demonstrated that chemical testing techniques are available that allows matching of samples on a one to one basis. However, this is a labour-intensive process and it would not be feasible to manually identify matches across a large data set. Hence, this problem is well suited to data science and machine learning where analysis of larger data sets is feasible. This is the focus of the research presented in this paper.

3.1. Machine Learning Applied to Spectroscopy and Chemometrics

There has been significant growth in the application of machine learning to chemometrics in the last decades (Butler et al. 2018). The vast majority of machine learning applications are for classification of chemicals or prediction of concentrations by regression. For example, test equipment that identified explosives at an airport is testing whether any of the samples are classified as those within its library of explosives. Within machine learning, these use supervised learning techniques where a defined library is used for training and classification. The IED sample matching problem is different to these. In this application, the ‘library’ of potential future homemade explosives is unknown. A library could be developed from analysing existing samples and then used for assessing future explosive samples against to find matches. However, this approach is undesirable as it limits the potential explosives that could be matched to a defined set. Additionally, the homemade nature of the explosives of interest are varied and constantly evolving. For this applied problem, the requirement is to match IED samples to each other, regardless of their classification.

To address this problem, it is proposed that machine learning clustering techniques are used. Clustering is the approach of identifying similarities and grouping together objects that are more similar than others. In this application, clustered items would be samples of identical, or similar explosives depending on how tight a cluster is. The actual type of explosive within a cluster does not need to be known or defined beforehand. This is an unsupervised learning technique as the data does not need to be labelled beforehand for learning. There are many clustering techniques. The metrics they use for clustering and the mechanisms they use for finding them vary greatly. A technique's suitability depends on the characteristics of the data being analysed and the desired outcomes from the clustering (Jain and Dubes 1988). Hence, consideration is now given to the desired outcomes of the clustering analysis and the data available on the explosives.

3.2. Data Characteristics and the Desired Clustering Outcomes

The primary desired outcome from the clustering is the identification of small tightly clustered groups of samples that can be considered close enough to a match for the implication that they are made by the same bombmaker. There is the expectation that there will be multiple tight clusters that need to be found in a dataset (multiple bombmakers) and the number of clusters is not known in advance. The ability to identify wider groupings that include samples that are not identical but have some similar characteristics is desirable for identifying linkages between potentially related bombmakers. Relationships may relate to similar training and techniques in their IED construction or use of the same recipe or materials. It is not expected that all data points will form a cluster. There may be significant numbers of unmatched data samples due to the variance in how homemade explosives are made, significant levels of contamination, or explosives made as a 'one-off'. These characteristics will need to be considered when selecting or developing appropriate clustering techniques (Bailey 1975) for this IED matching problem.

The total number of IEDs in recent conflicts like Iraq and Afghanistan are likely to be over 100,000, of which a percentage are collected and analysed (ABC News 2013). Within a region of interest and over a timeframe of interest, the number of analysed samples may be in the hundreds or thousands. Hence, to make this research useful, it must be feasible for results to be achieved from data sets with hundreds of samples. This is a relatively small number of samples compared to what is often used in machine learning applications (Brereton 2015). DST has agreed to provide hundreds of unclassified results from testing of explosives. For our study, Fourier Transform Infrared (FTIR) spectroscopy test data is being analysed. This data was selected as FTIR testing is commonly conducted on all explosive samples and FTIR data is available for this study. While FTIR testing is not the most sophisticated or advanced form of testing, it is hoped that any shortcomings in this testing can be overcome through the application of advanced data analysis techniques. Additional test data sets can be introduced and 'fused' if it is found that more fidelity is needed.

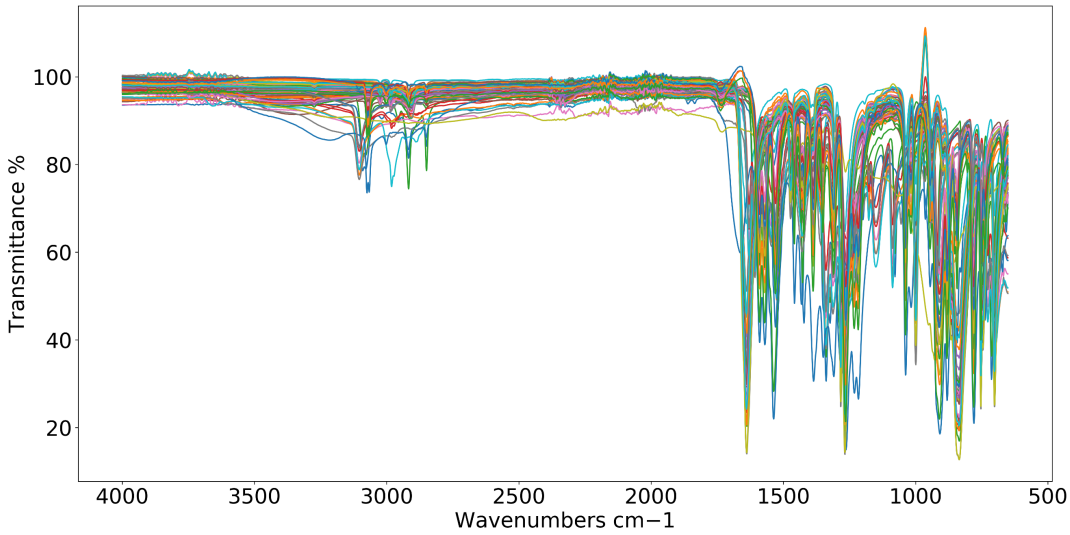

Figure 1. FTIR spectral data from the booster stage of 71 improvised detonators

The data from FTIR testing is a spectral plot (as shown in Figure 1) with measurements taken over a wide spectral range of wavelengths. This results in approximately 3500 data points per test (3500-dimensions).

4. ANALYSIS AND PRELIMINARY RESULTS

Analysis has been conducted on explosive samples to demonstrate the feasibility of this concept of matching samples of explosives using machine learning and data analysis. The data used for this demonstration is a selection of 71 explosive samples extracted from the booster or output energetic stage of improvised detonators. This component of improvised detonators is typically made of a limited set of commonly available explosives. Hence the data is likely to be clearly clustered for this demonstration. A brief summary of the analysis workflow and results is as follows:

4.1. Data Pre-processing

Data pre-processing is used to remove unwanted variation in the data and emphasise the variation of interest, hence improving the overall analysis results. There are many common pre-processing techniques that can applied to spectral data. However, consideration must be given for the characteristics of the data and the goals of the analysis to determine appropriate techniques (Engel et al. 2013).

For identifying matching explosive samples, the information about the composition of the explosive samples is largely held in the location and the shape of the peaks in the spectrum. Hence, selected pre-processing techniques were trialled that minimised any offset between spectra, largely leaving the peaks and spikes in the spectra for further analysis. Trialled techniques included centring, scaling, autoscaling, baseline correction, and Savitzky-Golay smoothing. It was found that axis 1 autoscaling (where each spectra is centred and then scaled based on its standard deviation) (Jackson 2005) had the most significant positive effect on clustering results (as shown in Figure 2) and is used in the remainder of this paper. Our researching is continuing to further improve pre-processing techniques.

4.2. Principal Component Analysis

Principal Component Analysis (PCA) is a commonly applied to spectral data. It is valuable in reducing the dimensionality (and complexity) of data and visualising the data for exploratory analysis.

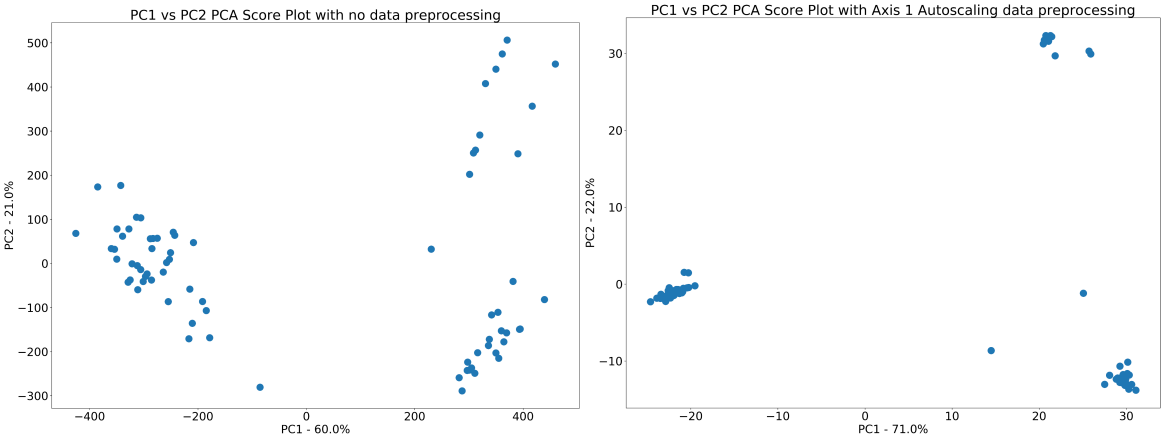

Figure 2. PCA score plots for unprocessed data (left) and pre-processed data (right)

As can be seen in Figure 2, our application of PCA provides a way of visualising the data in 2 dimensions and visually identifying clusters. This example highlights the improvement in clustering that is achieved through the application of Axis 1 Autoscaling as pre-processing.

4.3. Clustering

Unsupervised machine learning algorithms are now applied to identify and label clusters within the data. This is applied based on the principle that the tightest groupings or clustered data points will be the closest explosive matches. However, there are many types of clustering utilising different mechanisms or algorithms and can deliver differing results. While our research is continuing to identify or develop the most suitable clustering algorithms for our application, two have been selected and extensively applied to date with positive results:

Hierarchical clustering was chosen for application as the hierarchical aspect may be desirable for identifying group or source attribution. *Agglomerative* hierarchical clustering was selected as the bottom up approach ensures the cluster starting point is the closest pairs of points. If the lowest levels of grouping with the hierarchy are direct matches, then levels above that may show groupings and groups with a common characteristic.

DBSCAN: Density-Based Spatial Clustering of Applications with Noise was proposed by Ester et al. (1996) to discover arbitrarily shaped clusters amongst noise. As it finds clusters based on density, it does not need to know the number of clusters at initialisation time. Hence, DBSCAN was chosen as a potentially relevant technique to explore as it allows for specification of the tightness of clusters and allows noise or unclustered data points.

4.4. Internal Evaluation

Internal evaluation is a form of validation based on the clustering itself, without a known true set of labels for comparison. Internal evaluation algorithms score clustering results based on whether the sets of clusters are compact and well separated from each other. There are multiple potential metrics for the tightness of a cluster and separations of clusters resulting in multiple potential evaluation schemes. We have applied the commonly used Davies-Bouldin Index (Davies and Bouldin 1979) and Silhouette Coefficient (Rousseeuw 1987) to evaluate pre-processing techniques, clustering techniques and identify the correct number of clusters within a dataset. These indices can be used as an ongoing tool to evaluate the goodness of clustering.

4.5. External Evaluation

External evaluation is a form of validation of clusters against a known ground truth such as known class labels. For the data set used in this exemplar analysis, Figure 3 shows the comparison of the clustering generated by the machine learning algorithms (left) with the results generated by DST explosive specialists (chemists) on the right. In this relatively simple data set, the Hierarchical Clustering and DBSCAN clustering algorithms correctly separated all the data points into the same clustering as the DST scientists.

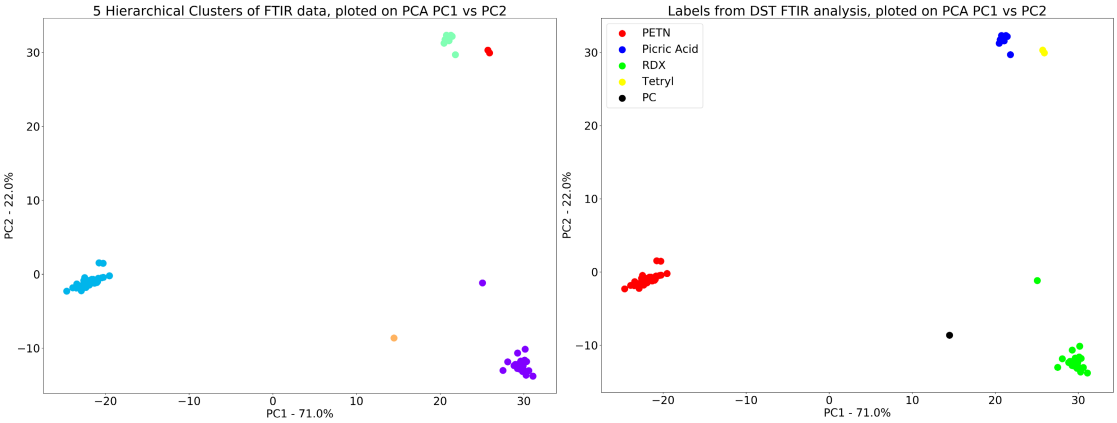

Figure 3. Clustering from machine learning algorithms (left) and from DST chemical testing (right)

A significantly more challenging data set and its associated clustering results are presented in Figure 4. This data is from the middle stage transition energetic from the improvised detonators which contain more varied homemade explosive materials. Here, the machine learning algorithms labelled and clustered 59 of the 69 samples correctly. While this is a good result, research is continuing to refine and develop pre-processing and machine learning algorithms to improve this result.

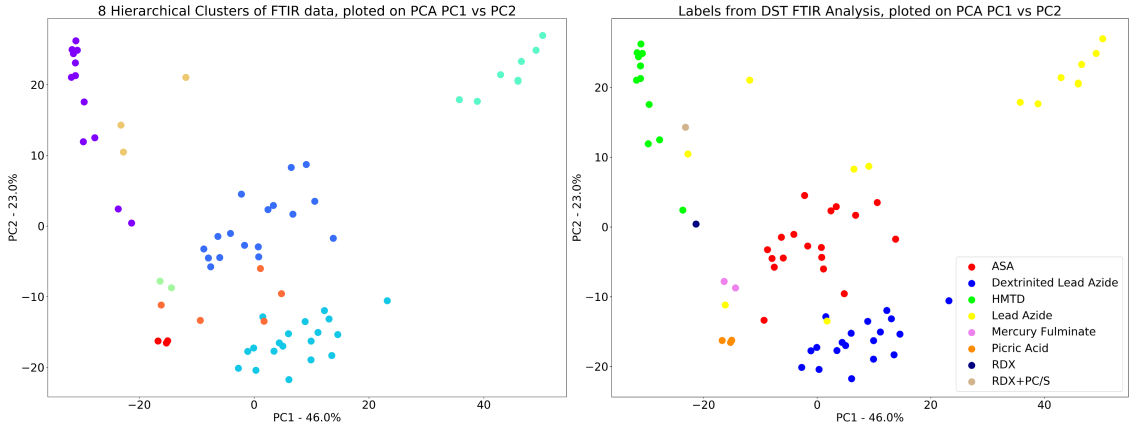

Figure 4. Clustering from machine learning algorithms (left) and from DST chemical testing (right)

5. CONCLUSION

This paper presents a novel concept for identifying links in threat networks. Links between IEDs and bombmakers are generated through matching specific unique characteristics in the construction of IEDs. In the example application presented in this paper, it is the chemical composition of the IEDs homemade explosives that are being matched. While the feasibility of this has previously been demonstrated through one to one matching of the chemical composition of explosive samples through advanced chemical analysis techniques (McCurry 2015), our research expands on this capability through the use of data science and clustering machine learning. This aims to enable the use of less sophisticated chemical tests (in this case, FTIR testing), rapid application across large data sets, and adaptability to threat evolution and unknown types of future explosives.

The data analysis process as presented can correctly cluster matching explosive samples in a relatively simple real-world dataset. When presented with a challenging and complex real-world dataset, the algorithms were able to correctly cluster the vast majority of explosive samples. These matches may be considered to be explosives made by the same bombmaker or manufacturer. If there are multiple bombmakers utilising the same type of explosive, additional subdivision of the clusters may be required. Research is continuing to develop and refine these algorithms to improve these results with the end goal of having confidence in the results when applied to future test results from unknown explosives.

REFERENCES

ABC News. 2013. 'Exclusive: Enter America's Repository of Pain, 100,000 Weapons of War - ABC News', Accessed 19/1/2019. https://abcnews.go.com/Blotter/enter-americas-repository-pain-100000-weapons-war/story?id=20978554.

Australian Government. 2013. "Defence White Paper 2013." In, edited by Department of Defence. Australian Government.

Bailey, Kenneth D. 1975. 'Cluster Analysis', *Sociological Methodology*, 6: 59-128.

Brereton, R. G. 2015. 'Pattern recognition in chemometrics', *Chemometrics and Intelligent Laboratory Systems*, 149: 90-96.

Butler, K. T., D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh. 2018. 'Machine learning for molecular and materials science', *Nature*, 559: 547-55.

Davies, D. L., and D. W. Bouldin. 1979. 'A cluster separation measure', *IEEE Trans Pattern Anal Mach Intell*, 1: 224-7.

Engel, Jasper, Jan Gerretzen, Ewa Szymańska, Jeroen J. Jansen, Gerard Downey, Lionel Blanchet, and Lutgarde M. C. Buydens. 2013. 'Breaking with trends in pre-processing?', *TrAC Trends in Analytical Chemistry*, 50: 96-106.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. "A density-based algorithm for discovering clusters in large spatial databases with noise." In *Kdd*, 226-31.

Jackson, J Edward. 2005. *A user's guide to principal components* (John Wiley & Sons).

Jain, Anil K., and Richard C. Dubes. 1988. *Algorithms for clustering data* (Prentice-Hall, Inc.).

Joint Improvised-Threat Defeat Organisation. 2018. 'About JIDO', Accessed 7 May 2018. https://www.jieddo.mil/about.htm.

McCurry, Paul Matthew. 2015. 'The Use of Advanced Analytical Techniques to Enable Batch and Source Matching of Homemade Explosives', Flinders University.

National Academies of Sciences Engineering and Medicine. 2018. *Reducing the Threat of Improvised Explosive Device Attacks by Restricting Access to Explosive Precursor Chemicals* (The National Academies Press: Washington, DC).

Rousseeuw, Peter J. 1987. 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics*, 20: 53-65.