VALCRI
VISUAL ANALYTICS FOR SENSE-MAKING IN CRIMINAL INTELLIGENCE ANALYSIS

***VALCRI* WHITE PAPER SERIES**
VALCRI-WP-2017-011
1 February 2017
Edited by B.L. William Wong

Applying Visual Interactive Dimensionality Reduction to Criminal Intelligence Analysis

Dominik Sacha[1], Wolfgang Jentner[1], Leishi Zhang[2], Florian Stoffel[1], Geoffrey Ellis[1] and Daniel Keim[1]

[1]University of Konstanz
78457 Konstanz
GERMANY

[2]Middlesex University London
The Burroughs, Hendon
London NW4 4BT
UNITED KINGDOM

Project Coordinator
Middlesex University London
The Burroughs, Hendon
London NW4 4BT
United Kingdom.

Professor B.L. William Wong
Head, Interaction Design Centre
Faculty of Science and Technology
Email: w.wong@mdx.ac.uk

EUROPEAN COMMISSION

Copyright © 2016 The Authors and Project VALCRI. All rights reserved.

I N T E N T I O N A L L Y B L A N K

ABSTRACT

VALCRI provides a challenging and overwhelming high-dimensional dataset that comprises of hundreds of extracted semantic features in addition to the usual spatiotemporal information or metadata. To overcome the curse of dimensionality and to generate low-dimensional representations of these semantic features we apply interactive high-dimensional data analysis techniques with the goal of obtaining clusters of similar crime reports. However, it is still a challenge for crime analysts to make sense of the results and to provide useful interactive feedback to the system. Therefore, we provide several tightly integrated interactive visualizations that allow the analysts to identify clusters of similar crimes from different perspectives and interactively focus their analysis on features or crime records of particular interest.

Keywords

Criminal Intelligence, High-Dimensional Data Analysis, Feature Extraction, Dimensionality Reduction, Visual Analytics

I N T E N T I O N A L L Y B L A N K

INTRODUCTION

A major task for crime analysts is to identify and group crime reports according to their similarity. However, the "similarity" of crimes that is computed based on extracted semantic features from the textual crime report description may be defined in different ways and be specific for a variety of analysis tasks, crime types, as well as geographic, and temporal characteristics. This flexibility requires an exploratory analysis of crime similarities with overviews that can be iteratively refined. In addition, the automatic processing of text documents in VALCRI provides the analyst with a large number of extracted features for each crime report, resulting in a challenging dataset that may contain millions of data records and hundreds of extracted features.

Our approach to tackling this high-dimensional dataset is to apply a dimensionality reduction (DR) pipeline that can be steered by interactive visualizations that are integrated into the VALCRI framework. On the one hand, our solution lets the analyst apply external (e.g., geographic or temporal) filters to the dataset allowing them to investigate and identify feature characteristics of the remaining crimes. On the other hand, our approach provides the analyst with low-dimensional representations of crimes that enable the analyst to investigate and identify clusters/groups of similar crimes in several perspectives. The domain expert is supported to provide feedback to the system at every step of the pipeline.

This paper describes our work in progress. A major challenge of effective user involvement is in translating their feedback in the different components in terms of appropriate DR pipeline adaptations and recomputations. This is a research challenge that has been recognized in the ML and VA communities (Sacha et al., 2016). To this end, we provide the analyst with intuitive visualizations and provide guidance based on automated measures. Our approach hides computational complexity but allows the analyst to identify interesting patterns, such as groups of similar crimes or outliers.

The major steps of our solution are 1) feature extraction from textual crime reports and a weighted similarity model as a pre-processing step, 2) calculation and visualization of feature characteristics, 3) use of different algorithms to produce low-dimensional embeddings of the data, 4) visual interactive representations of similar crime clusters, and 5) adaptive computations of results. The remainder of this paper provides details of the analyzed dataset and high-dimensional data analysis, describes the DR pipeline with its interactive visualizations and explains how it integrates with the table components.

DATA PROCESSING AND FUNDAMENTALS ON HIGH-DIMENSIONAL DATA ANALYSIS

Semantic Crime Case Analysis

The VALCRI text processing pipeline extracts a large number of features from crime reports based on semantic word lists or ontologies about known and domain specific concepts. For example, all terms that describe materials (copper, steel, wood, etc.) can be identified and extracted as features that belong to a common "materials" concept category. Similarly, physical parts that belong to a house (door, floor, kitchen, etc.) can be identified and extracted as features that belong to a common "building parts" category. Hence, an analysis question could be to distinguish burglaries that identify doors according to the associated materials (e.g., "wooden door" vs. "steel door") or whether the doors were "locked" or "unlocked". The text processing pipeline provides numerous of these semantic features that can be transferred to a binary feature vector for every crime record (a vector that describes whether a specific concept was identified or not).

Similarity Mapping and Feature Selection

For comparative case analysis, analysts want to find out the similarities between crime cases to support reasoning and sense making. Given a collection of crimes, it is possible to apply a metric (for example, *Euclidean distance* that takes into consideration the number of common concepts/features in two crime cases) to measure the distance/dissimilarity between pairwise crime cases, resulting in a so-called distance matrix (see Table 1 as an example for US-cities). In VALCRI we make use of a weighted similarity metric that is computed based on a feature importance. Feature selection can be interactively defined based on semantic relations between features (e.g., select all the building parts and materials), but also semi-automatically using correlation analysis or subspace clustering.

Table 1. Distance matrix between seven US cities

	BOSTON	*NY*	*DC*	*CHICAGO*	*SEATTLE*	*SF*	*LA*
BOSTON	0	206	429	963	2976	3095	2979
NY	206	0	223	802	2815	2934	2786
DC	429	223	0	671	2684	2799	2631
CHICAGO	963	802	671	0	2013	2142	2954
SEATTLE	2976	2815	2684	2013	0	808	1131
SF	3095	2934	2799	2142	808	0	379
LA	2979	2786	2631	2954	1131	379	0

Visual Embedding Techniques (Dimensionality Reduction)

A common approach to analyzing the similarity is to embed them into a visual display as a scatter plot where each point represents a crime and distances between pairwise points indicates their dissimilarities, i.e., similar crimes are placed close to each other, and dissimilar ones are far apart. This allows the analyst to see the patterns in the data. Figure 1 shows some example visual embeddings.

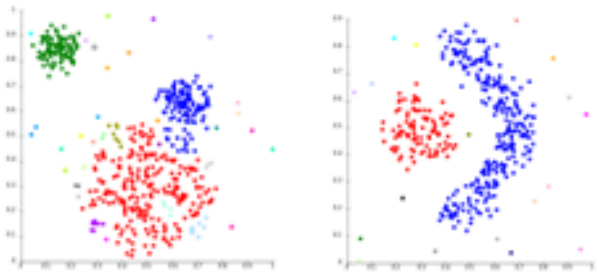

Figure 1. Visual embeddings that show similarity and patterns in the data. Colors indicate different class labels.

Various dimensionality reduction algorithms exist for computing the visual embedding of a dataset (Maaten et al. 2009). The idea is to *best approximate* the dissimilarity between objects in the data space to the Euclidean distance between points in the visual display so that groups of similar object and outliers, can be easily identified. However, it is impossible for any algorithm to guarantee a 100% accurate mapping as different algorithms have different strengths in showing patterns in data. For VALCRI we chose three algorithms: *PCA*, *MDS,* and *t-SNE*.

PCA is the ancestor of all embedding algorithms and is still the most widely used methods across different disciplines and applications. The method has the strength of highlighting linear patterns in the data (see the left embedding in Figure 1 as an example). The visualization is based on the top two principal components, each of which is determined by a set of features with a different level of influence. These components reflect linear combinations of the largest variation of the original features. Although classic MDS (Kruskal and Wish, 1978) is similar to PCA, non-metric **MDS** allows the analysts to find the nonlinear patterns in the data (see the right embedding in Figure 1 as an example). Distance-based techniques, such as MDS, aim to preserve the distances (e.g., Figure 1) in the embedding. The principle of non-metric MDS is a nonparametric and monotonic distance transformation that is identified with either isotonic optimization or semi-definite programming. **t-SNE** is one of the state-of-the-art embedding algorithms that have the strength of highlighting the nonlinear patterns in the data and is an example of a neighborhood preserving technique. The algorithm outperforms many other methods in the visual quality in terms of clearer separation between clusters (Maaten and Hinton 2008). In VALCRI we implemented all these different methods to allow the analyst to visualize the data from different perspectives (linear, distance-based, and neighborhood-preserving) in order to validate and verify clusters and outliers across techniques in an exploratory manner.

S^3 – SIMILARITY SPACE SELECTOR

We implemented the described techniques in an interactive DR pipeline and give the analyst a way to interact in several visualizations. We first provide an overview of the pipeline steps and subsequently describe the different visualizations in more detail.

Pipeline Overview

Our visual analytics pipeline, shown in Figure 2, illustrates the computational steps (top), accompanying visualizations (middle) and interactive feedback of the analyst (bottom). In the very first step (Feature Extraction) the features are extracted from crime report text documents and transferred to binary feature vectors.

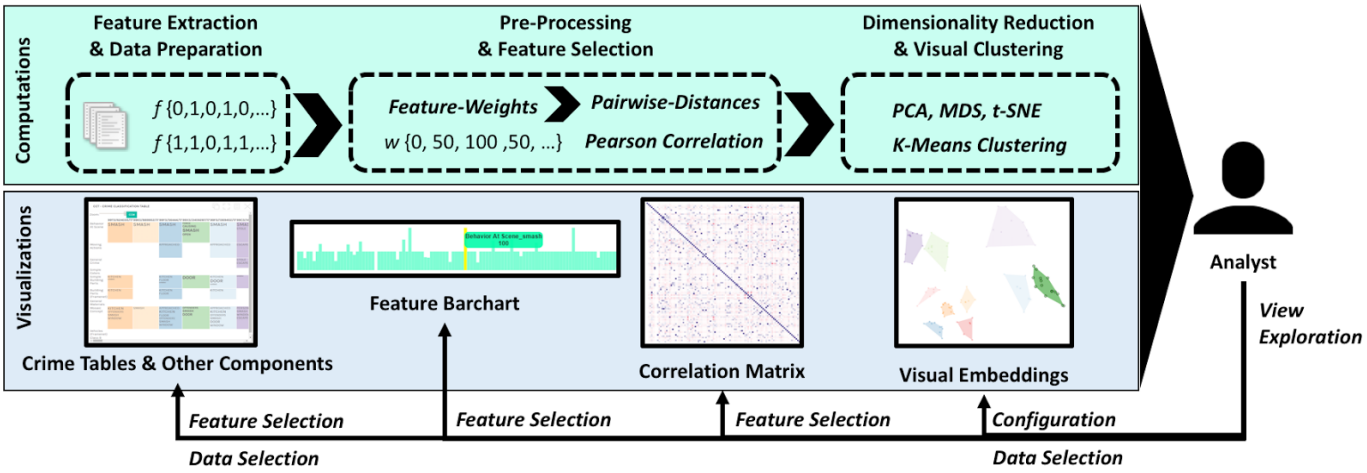

Figure 2. DR pipeline overview: Computation steps are shown on top, with their respective visualizations shown below. The analyst is able to adapt the pipeline during the analysis, as shown at the bottom.

We implemented a tabular and matrix-based visualization that shows the semantic features of selected crime reports (see next section for more details). In the second step (Pre-Processing) the feature weight (or importance) model is created and visualized as bar charts. Based on these feature-weights, the pairwise distances are prepared for the subsequent DR computations. The feature selection in this stage is supported by Pearson correlation analysis or semantic concept relationships. In the DR step, three different visual embeddings can be computed (PCA, MDS, t-SNE) and a k-means clustering is computed in the visual space (obtained x and y coordinates). The identified clusters are broadcast to the other components in the VALCRI framework (e.g., crime tables, maps). The analyst can select and highlight clusters of interest and adapt the similarity model. Further external changes to the analyzed data (setting geographic, temporal or concept-based filters) will cause an immediate recomputation of the pipeline.
Once the results are computed we update all the visualizations accordingly and animate between the previous and the new states. This allows the analyst to track changes, investigate the similarities of the crimes, and to fine-tune the DR-pipeline by feature and data selections as well as parameter tuning.

Data Preparation
The data to be analyzed can be changed by external filters using the search functionality, temporal selections, or geographic filters. Any changes to the input data will cause a recalculation of the pipeline.

Weighted Similarity Model & Feature Selection
A foundation for adapting the feature importance of the pipeline is a feature weight model in the pre-processing step. The model covers an importance (or weight) value between 0-100 for each feature (with a default value of 50). This weight vector can be changed based on user feedback or automated feature selection algorithms and depicts the basis for calculating the pairwise distances. These distances are calculated using a weighted Euclidean distance measure and serve as an input for the subsequent DR and clustering step. The feature weights are visualized as a bar chart (Figure 3) below each scatter plot (Figure 5). Hovering over a bar reveals the feature names in tooltips whereas dragging a bar changes the weight (between 0 and 100) of that feature and when finished, will cause a recomputation of the DR. The feature weights can also be changed by other (external) components, such as clicking on a concept in the table (see “Crime Tables” Section for more details).

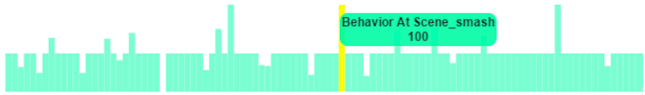

Figure 3. Bar chart: The feature weights are shown as bars. Hovering will reveal information at the tooltip (as illustrated), whereas dragging up and down will change a feature weight.

Feature selection is further supported by computing feature correlations in the analyzed dataset. Therefore, we compute Pearson correlation coefficients and visualize them in a matrix visualization (Figure 4). Positive correlations color the cells blue whereas negative correlations color them red. This allows the analysts to understand feature relationships and to spot co-occurring (e.g., “smash” occurs likely with “window”) as well as those which do not co-occur (e.g., either “door” or “window” occurs, but not both). Hovering will reveal the feature names and the correlation value. Clicking on a cell will set the respective feature weight to zero and cause a recomputation of the pipeline. In the future, we plan to apply reorganization algorithms to reveal patterns in the matrix.

Figure 4. Correlation matrix view: Hovering over a cell will reveal the feature names and correlation strength. Clicking on a cell will exclude the respective feature from the computations.

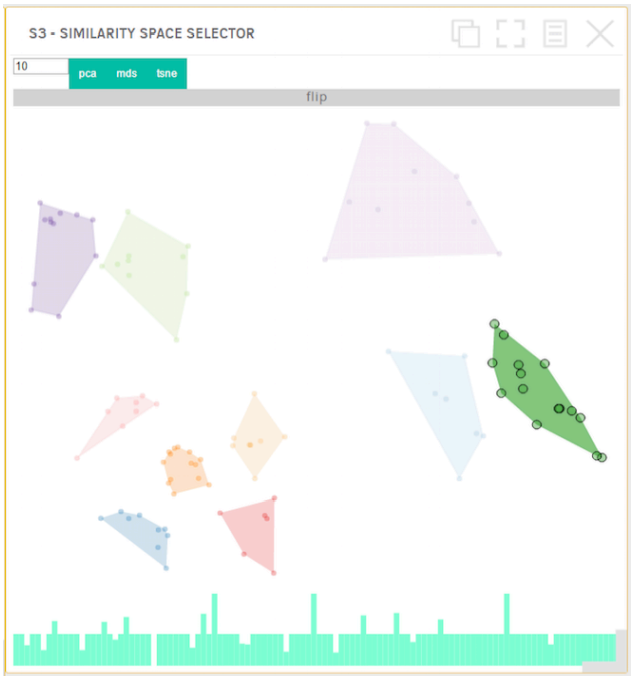

Figure 5. Scatter plot and bar chart: The bar chart shows the feature importance model for the respective embedding above. The scatter plot visualizes the crimes in a two-dimensional embedding together with the clustering result (color and cluster outline).

Dimensionality Reduction & Visual Clustering

The final step in the pipeline is the computation of the visual embedding and an additional clustering within the resulting two-dimensional space. The analyst can choose between three different algorithms for different purposes (*PCA* - feature-based linear embedding, *MDS* - (large) distance preserving, *t-SNE* - neighborhood preserving). For clustering, we implemented the *k*-Means algorithm that lets the analyst set the number of desired clusters. The obtained clusters are visualized as colored dots within the convex hull of all the cluster members in a scatter plot (Figure 5). Hovering over the cluster area will highlight the cluster, clicking will select it. Other components are notified of any selections in order, for instance, to reveal the respective crimes (linking & brushing). Similarly, external components can send highlight-selections to S^3 to reveal the desired items in the scatter plot. This allows the analyst to see the selected clusters in the other views, which can help them understand characteristics of the cluster (e.g., find out which concepts are important and distinguish between different clusters). Interactions, such as changing the feature weights or choosing the DR algorithm will cause a recomputation of the pipeline. Once the computation is finished, we animate the crimes moving to their new positions in the scatter plot. This allows the analyst to keep track of changes and to perform "what-if" testing. For example, an analyst may remove a feature and observe how the visualization (the distances and clusters) changes. Clicking on the "flip" area on top of the visualization (scatter plot or correlation matrix) allows the analyst to switch between the views.

CRIME TABLE

The design of the crime table component is inspired by one of the tools available to crime investigators. Its main purpose is to provide a quick overview of concepts and their underlying terms that are associated with each crime. Currently, this task is mainly solved with large tabular representations using standard software tools, such as Microsoft Excel. Our component offers two different modes that can be viewed at any time. The Concept Classification Table (CCT) provides a detailed view of terms organized by their semantic concept affiliation. The second mode, called Crime Concept Matrix (CCM), features a matrix-based overview visualization of concept-term combinations.

Concept Classification Table (CCT)

The default setting of the CCT view is shown in Figure 6. Each concept (or feature) that occurs at least once in the selected set of crimes is represented as a row in the table. The crimes are listed column-wise and denoted with their crime number on top. When a concept-term occurs in a crime it is shown in the respective cell. For example, the terms "left" and "walked" belonging to the concept "Moving Actions" are shown in the first colored cell of the table. The colors are provided by the S^3 component and represent clusters of similar crimes.

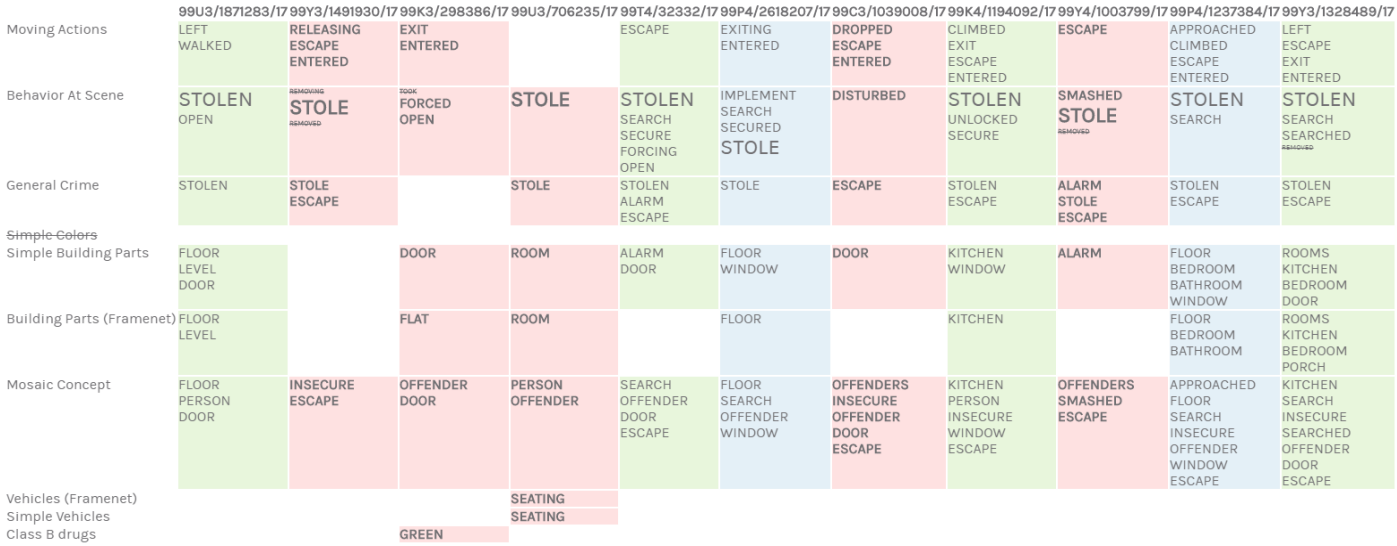

Figure 6. Table Component (CCT-Mode): The crimes are listed in the columns. Each row represents a concept. When a concept term occurs in a specific crime it is written into the cell.

Crime Concept Matrix (CCM)

The CCT can be morphed into the CCM where each concept-term combination is represented as its own row (see Figure 7). In contrast to the previous example in the CCT section, the terms "left" and "walked" are now represented as separate rows and are prefixed with their affiliated concept. As in the CCT, the colored cells depict that the respective concept-term combination occurs in this crime. The color represents the cluster membership as defined by S^3.

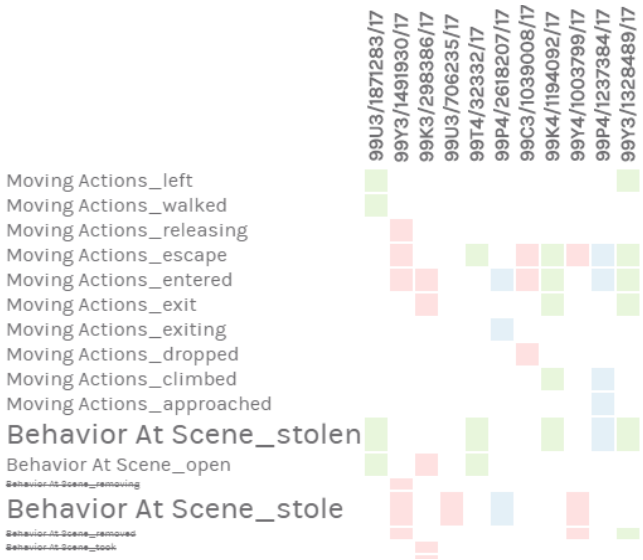

Figure 7. CCM: Each concept-term combination is displayed row-wise.

Both, the table as well as the matrix can be zoomed. Zooming out provides an overview of a large number of crimes or concepts can help detect larger patterns. The terms are faded out in CCT mode when the zoom level decreases to a certain degree in order to provide the same look and feel as the CCM. However, in contrast to the CCM, only the concepts are displayed in combination with the colored cells.

DR Pipeline Integration

The S^3 and the crime table components are tightly integrated. Besides the brushing of the clusters represented by their colors, the components are also interactively linked. Hovering over a cluster or single point highlights the crime(s) as bold text in the table (see Figure 6, red cluster). Clicking on a term will change the font size and will make the term in all cells appear larger and thus, more important. This is shown in Figure 6 with the terms "stole" and "stolen". Furthermore, an event is broadcasted updating the weights for the current data set. The event is intercepted in S^3 where the bar chart is updated and the currently selected projection method is used to recalculate the new clusters with the new weight settings. Each recalculation may update the clusters and thus, might change the colors in the table component. A second click on the term will decrease the font-size to a minimum and strike-through the term implying that the weight of this term is set to zero. In Figure 6 this true for the terms "removed", "removing", and "took". A third click puts back the term to the default state (50%). The same interaction is also possible in the CCM-mode when the user clicks on a concept-term combination (see Figure 7). In CCT-mode, the user may also interact with a concept. This will hide the whole row in the table and, therefore, set all weights of the terms belonging to this concept to zero (see Figure 6 with concept "Simple Colors"). A second click will make the row reappear, putting the weights back into the default state. As described in the previous chapter, the user may also define the weights using the bar chart causing the same weight event to be broadcasted from the S^3 component. The weights are again mapped to the font-size and are set as strike-through when the weight is set to zero.

CONCLUSIONS

This paper describes our work in progress towards applying high-dimensional data analysis techniques to crime case analysis with the goal to explore semantic similarities. In the future, we will focus on supporting and guiding the analyst further in configuring the pipeline and the visualizations. Therefore, we plan to investigate automatic feature selection algorithms (such as subspace clustering), matrix sorting algorithms, and visual quality metrics to reveal patterns in the visualizations and to provide the analyst with recommendations.

ACKNOWLEDGEMENTS

The research leading to the results reported here has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) through Project VALCRI, European Commission Grant Agreement N° FP7-IP-608142, awarded to Middlesex University and partners.

REFERENCES

J. B. Kruskal and M. Wish. Multidimensional scaling, vol. 11. Sage, 1978

L. van der Maaten, .J.P.; Hinton, G.E. (2008). Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research. 9: 2579–2605.

L. van der Maaten, E. Postma, and H. van den Herik (2009). Dimensionality reduction: A comparative review. Technical report, Tilburg Centre for Creative Computing, Tilburg University.

D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North and D. A. Keim (2016). Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis. IEEE Transactions on Visualization and Computer Graphics (TVCG), DOI: 10.1109/TVCG.2016.2598495.

The research leading to the results reported here has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) through Project VALCRI, European Commission Grant Agreement Number FP7-IP-608142, awarded to Middlesex University and partners.

	VALCRI Partners	**Country**
1	Middlesex University London Professor B.L. William Wong, Project Coordinator Professor Ifan Shepherd, Deputy Project Coordinator	United Kingdom
2	Space Applications Services NV Mr Rani Pinchuck	Belgium
3	Universitat Konstanz Professor Daniel Keim	Germany
4	Linkopings Universitet Professor Henrik Eriksson	Sweden
5	City University of London Professor Jason Dykes	United Kingdom
6	Katholieke Universiteit Leuven Professor Frank Verbruggen	Belgium
7	A E Solutions (BI) Limited Dr Rick Adderley	United Kingdom
8	Technische Universitaet Graz Professor Dietrich Albert	Austria
9	Fraunhofer-Gesellschaft Zur Foerderung Der Angewandten Forschung E.V. Mr. Patrick Aichroft	Germany
10	Technische Universitaet Wien Assoc. Prof. Margit Pohl	Austria
11	ObjectSecurity Ltd Mr Rudolf Schriener	United Kingdom
12	Unabhaengiges Landeszentrum fuer Datenschutz Dr Marit Hansen	Germany
13	i-Intelligence Mr Chris Pallaris	Switzerland
14	Exipple Studio SL Mr German Leon	Spain
15	Lokale Politie Antwerpen	Belgium
16	Belgian Federal Police	Belgium
17	West Midlands Police	United Kingdom