**Human-centred Machine Learning through Interactive Visualizations: Reflections on the Design of a Visual Analytics Tool for Criminal Intelligence Analysis**

**Keywords:** Machine Learning, Human Computer Interaction, Criminal Intelligence Analysis

# Introduction

With the increasing popularity of Machine Learning (ML) in many application fields such as health care, environmental protection and criminal investigation, more and more application analysts are keen on incorporating ML to their existing analytical approaches in order to improve the efficiency and effectiveness [1]. Given an application problem, often tailor-made ML software is developed to help addressing the existing challenge in data processing and knowledge extraction. Since most of the application analysts are not ML experts, the challenge is to design the tool in such a way that the analyst can easily understand the underline principle behind the algorithms and feel comfortable to use it. Further challenge lies in the design of semantic interactions that can seamlessly translate the user input as new knowledge and activate the backend re-computation of the model. In this abstract we introduce our journey through the design of a visual analytics tool as part of the EU-funded project “Visual Analytics for Sense-making and Criminal Intelligence Analysis (VALCRI)” [2], with the aim of making ML more approachable by domain analysts.

# Background

The aim of the VALCRI project is to develop a Visual Analytics (VA) system to improve the effectiveness of current criminal intelligence analysis solutions. The task we focus on in this abstract is the so called Comparative Case Analysis (CCA) that is commonly used by police forces. Given a collection of crime reports, the idea of CCA is to analyze the commonalities between crimes in order to support reasoning and decision making. For instance, examining solved crimes that have similar characteristics as an unsolved crime may help the analyst generate a new hypothesis during a criminal investigation, and understanding the uneven distribution of crimes in terms of spaces, types of offenders and victims may help the police to allocate police resources more effectively [3].

CCA starts with the extraction of relevant headings (concepts or features) that are considered to be useful for the understanding of the crime case. Information is then collated under the headings, resulting in a CCA table where each row is a crime case. Typically the CCA table headings include names of structured fields in the crime report such as time and location of the crime, as well as concepts or features extracted from the Modus Operandi (MO) of the crime. For instance, given the MO text below, concepts such as entry type (door), premise type (school), stolen item type (money, jewellery) are often extracted as CCA table headers.

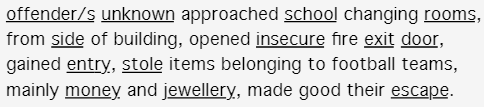


Figure 1 an example Modus Operandi

CCA is traditionally carried out manually on a spreadsheet. Once the CCA table is generated, the analyst scans through the table to identify commonality between crimes. The tasks become increasingly difficult due to the growing volume and complexity of today’s crime data. We aim at developing a VA tool can automatically generate the CCA table and supports similarity analysis via interactive visualizations.

We iteratively refine our tool based on several rounds of expert feedback and evaluation. We worked in close collaboration with three groups of experts from different police forces across Europe. In the early phases, we collect feedback after demonstration of our prototypes. Later, we evaluate the tool by asking the expert to perform particular tasks with given datasets.

# Initial Design

Our initial prototype takes crime reports as input, automatically extract relevant features in the text using a series of text mining approaches such as tokenization, POS tagging, concept extraction, and term frequency extraction [4]. The result is used to form the CCA table. A similarity metric is then applied to calculate the similarity between crimes based on the CCA table. To map the similarity into visual displays, a Dimensionality Reduction (DR) technique is applied, result in a scatterplot-like visualization where each point represents a crime and the distances between points represent similarities. The user interface also includes visualization panels to support geo-temporal analysis and commonality analysis on the data. Figure 1 shows the VA workflow of the prototype.

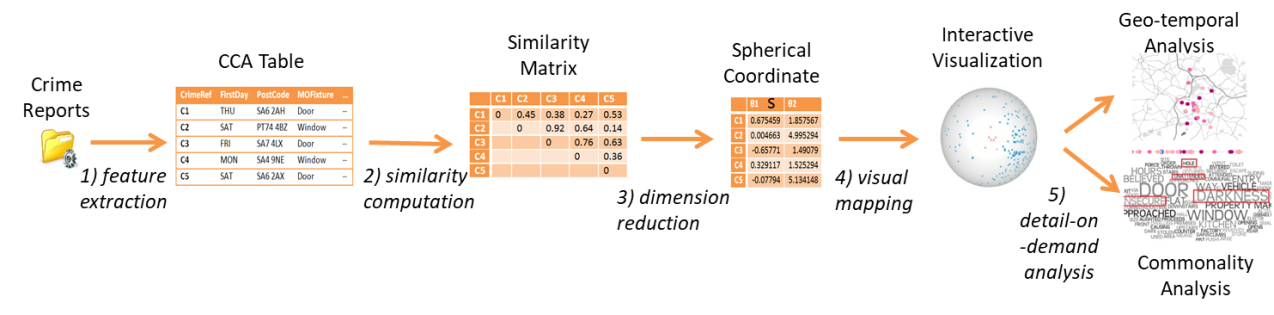


Figure 2 Visual Analytics Workflow of the Initial Prototype

# Feedback from end user

We demonstrated the prototype to our end user experts. They welcome the fact that the data can be processed quickly and the similarity between crimes can be examined more easily through a visual representation with quantitative measures. But the majority of them found the ML part of the system “scary”, mainly due to the lack of understanding and control of the algorithms. For instance they find it hard to understand the meaning of axes in the scatterplot visualization and hard to interpret the similarity. “What are the key features that characterise a group of similar crimes?” they asked. Another complain came from the loss of mental map. The prototype projects the crime onto a visual display without showing the CCA table. Interestingly they also identified some new requirements after the demonstration such as clustering (“Can I group the crimes according to the similarity?”) and feature weighting ((“Can I adjust the weight of the headings?”)

# Our Reflection

With growing size and complexity of data; more and more application analysts who are not necessarily ML experts want the help of ML. The challenge is to make ML more approachable to them. This involves the design of intuitive visualizations that matches the mental map of the analysts, as well as intuitive interactions that can be seamlessly translated into backend computations. For instance a CCA table is a better choice for police analysts than the 3D projection of crimes. A scrolling bar that allows the analyst to adjust the weight of the headings and as consequences triggers the re-computation of similarity and update of the visualization makes the tool more attractive to use. Furthermore, ML approaches such as feature selection and weighting provides the analyst with more control over the algorithms as well as better understanding of the causal effect of different parameter settings. The iterative design process also helped us to identify hidden user requirements such as grouping crimes such that appropriate ML approaches (clustering) can be implemented to support their reasoning and sense making.

# Design Refinements

In [1] we proposed a conceptual framework for human centred machine learning. The aim is to identify possible “handles” that could be implemented at each stage of the analysis and ML process for the analysts to interact with the data and the model. We refined our design by adapting the model for CCA. The design incorporates the feedback from our end users. The new framework (see Figure 3) first processes the crime data to generate a CCA table, and then allows the analyst to select subset of data for analysis. The analyst can analyze the correlations between features and select the interesting ones for further investigation. They can define the “interestingness” of features by adjust their weights. The framework includes a series of DR algorithms and clustering algorithms to project the crime as 2D scatterplot and group the crimes according to their similarities. The result of the projection and grouping is visualized in a set of UI components, where the analyst can interact with the data and the visualizations.

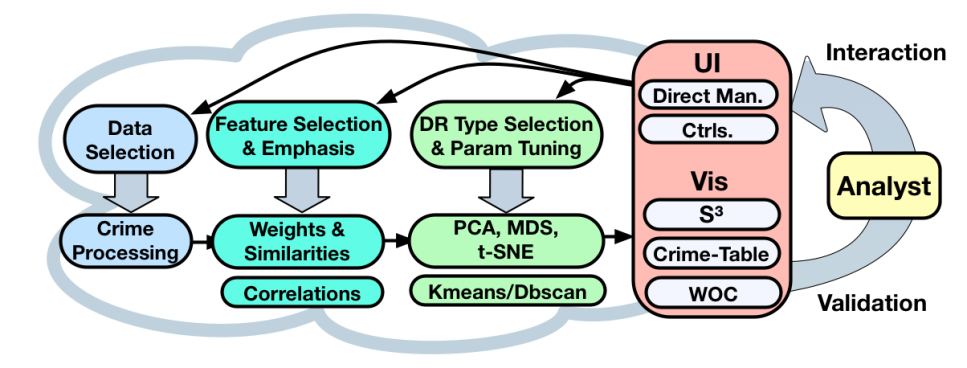


Figure 3 the framework of the new prototype, it embeds a DR pipeline (bottom) into an iterative exploration process (right) with several user interactions (top).

Figure 4 shows the two main components of the design. The **Similarity Space Selector** provides a simple interface for the analysts to understand the relations and similarities among multiple crimes across different DR and clustering results. It represents the two-dimensional data space with crimes arranged according to feature similarities (i.e., if they contain similar crime patterns). The **Crime Cluster Table** adopts the mental models of the traditional CCA tables recorded on spreadsheets. The analyst is presented with an aggregated cluster representation that encodes feature frequencies in each cell of the table (clusters are represented as rows and features in columns). Analysts can expand any cluster to reveal the detailed crimes as columns listing the contained concept terms. Feature weights are mapped to font size and the user can directly adjust them within the table (by clicking on a term and changing the weight using a slider).

Other components such as **Similarity Space Selector, Weight Observer Component** and **Pattern Selector** are also added at later stage to support different types of analysis. More details of the final product can be found in [5]. The feedback from the end user is positive. The tool turned out to be a good alternative to their traditional spreadsheet approach and provides more functionality that makes their daily job easier. With the help of ML techniques, the tool improves the effectiveness and efficiency of their analysis.

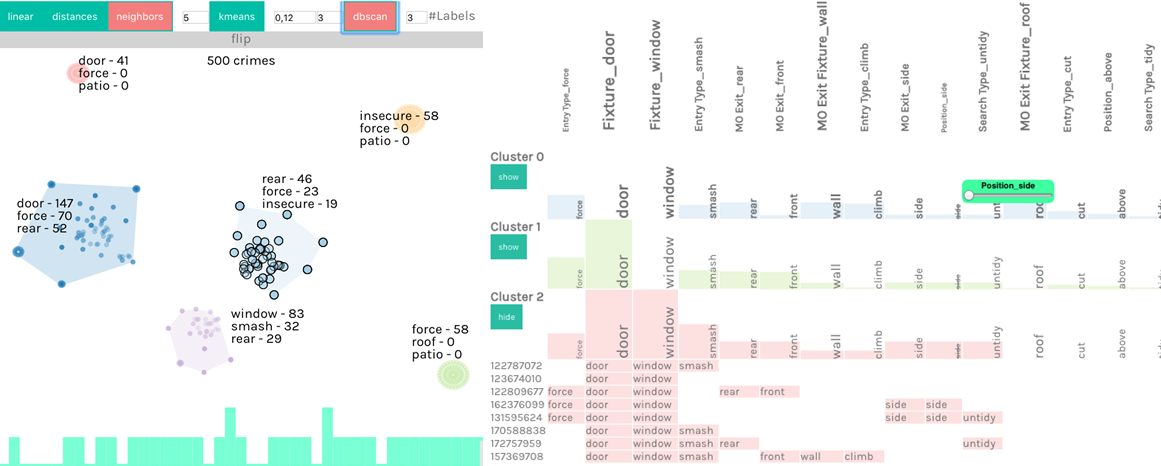


Figure 4 Left: Similarity Space Selector, Right: Crime Cluster Table

# Conclusion

In this abstract we presented our journey in the design of an interactive CCA system. The current system provides a powerful tool using a hybrid approach to simultaneously analyze and explore the data and an automatically generated feature space. DR techniques are utilized in a similar fashion to visualize the similarity spaces. The hybrid view aids the users in drawing conclusions on the effects of features in the data space. The tight coupling of multiple components allows access to the data from different perspectives. Our DR pipeline implementation supports a variety of interactions but we observed and learnt that analysts may be overwhelmed by an excessive number of visual alternatives and configuration options. To tackle this problem we allow the users to interpret the results and interact directly with them in the crime table (the tool that they are familiar with). This helped them to understand and importantly, build trust in the computations. Our visual interaction design is generalizable to other data types and applications.

# References

[1] D Sacha, M Sedlmair, L Zhang, JA Lee, J Peltonen, D Weiskopf, What you see is what you can change: Human-centred machine learning by interactive visualization, SC North, Neurocomputing 268, 164-175

[2] VALCRI, http://www.valcri.org/, last retrieved 14/08/2018

[3] N. Cope. Intelligence led policing or policing led intelligence?: Integrating volume crime analysis into policing. Br. J. Criminol., 2004.

[4] L Zhang, C Rooney et al., Spherical similarity explorer for comparative case analysis, In: IS&T Electronic Imaging 2016 Conference on Visualization and Data Analysis 2016, 16-18 Feb 2016, San Francisco, CA, USA.

[5] D Sacha, W Jentner, L Zhang, F Stoffel, G Ellis (2017) Visual comparative case analytics. In: EuroVis Workshop on Visual Analytics, 12-13 June 2017, Barcelona, Spain.