

## **QCIF Technology Diffusion Project Proposal**

### **Exploring the Benefits of Data Mining on Juvenile Justice Data**

#### **Chief Investigators:**

Dr Brett Gray  
Assoc Prof Anna Stewart  
Dr Troy Allard  
Dr Andrew Lewis

#### **Contact:**

Brett Gray  
Justice Modelling @ Griffith  
Griffith University  
[b.gray@griffith.edu.au](mailto:b.gray@griffith.edu.au)

#### **External Client:**

Department of Communities  
111 George Street  
Brisbane Qld 4001

#### **Business Case**

Knowledge discovery and data mining (KDD) is an area of analysis that has emerged from computer science. In this area the emphasis is on the semi-automated extraction of relationships from large datasets rather than the traditional statistical framework of testing relationships to provide evidence for or against a given hypothesis. KDD has proven to be a useful analytical methodology and has been applied in a range of commercial applications such as credit risk assessment, fraud detection and medical survival analysis. Additionally, KDD has had several scientific applications and been applied in biomedical, geospatial and climatic research.

Although data mining has proven to be a useful methodology in a number of fields, its use in criminal justice applications and research is in its infancy. Given this, the purpose of this project is to apply a number of data mining techniques to data of Queensland juvenile court appearances to extract relevant, potentially unforeseen, relationships. An analysis procedure allowing significance tests of extracted relationships will be incorporated into the validation process. Extracted relationships may have implications for policies related to juvenile crime prevention such as the target allocation of resources to appropriate intervention programs.

The work will be used to test the effectiveness of the method and to develop data mining expertise within JMAG. This expertise will be of value to criminal justice organisations responsible for devising justice policies that may have the effect of reducing crime. JMAG currently has strong relationships with Queensland criminal justice agencies. A significant outcome of this work will be a report on discovered relationships for the Department of Communities. In addition opportunities will be used to inform and educate criminal justice agencies on the benefits of KDD. It is

anticipated that further consulting/research opportunities in the area may be established.

### Justification of Budget

This project has been budgeted for a grant of \$40,000. This funding will be primarily used to cover the salary of a research assistant (SRA1.1) to work on the project three days each week for a year (.6 FTE). The salary expense will be \$39,410, inclusive of on-costs (\$6,026). The research assistant will contribute to the project by undertaking a literature review, assisting with data cleaning, formatting and analyses (including validation of the techniques), and report writing. It is anticipated that this person will have a background in computer science and /or statistics.

The remainder of the grant (\$510) will be used to cover general administrative and research costs such as photocopying and printing. It is anticipated that the project will take 12 months to complete:

Tasks	Month												
	1	2	3	4	5	6	7	8	9	10	11	12	
Liaising with DoC to obtain final approval to use the dataset	PR/RA												
Conduct literature review	RA	RA											
Data analyses and validation			PR/RA	PR/RA	PR/RA	PR/RA	PR/RA	PR/RA					
Write interim report					RA	PR/RA							
Write final report									PR/RA	RA	RA		PR/RA

Note: DoC = Department of Communities, PR = Principal Researcher and RA = Research Assistant

In addition to the funding from the grant, in kind resources to a value of approximately \$50,000 will be provided by Griffith University and Queensland Department of Communities. These resources are primarily drawn from Griffith University in the form of the involvement of two of the principal researchers in the data analysis and validation process (Brett Gray and Andrew Lewis) and all principal researchers in the production of reports and meeting with the Department of Communities for the education process. Additional resources involved include the use of high performance computing and storage infrastructure. Data mining techniques can be computationally intensive and the availability of super computing power will facilitate a faster analysis process that will enable more iterations of the analysis methodology to be performed. Support from the Department of Communities is in the form of staff time for involvement in the education process and the communication of discovered relationships as well as the provision of data and any support that may be required by Griffith University for the understanding of the data provided.

### Products and Benefits

This project will perform a data mining analysis of Queensland juvenile court data in order to establish a data mining methodology suitable to criminology research and test

the effectiveness of this method. The analysis will provide a comparative study of different data mining techniques and adopt an analysis methodology that allows testing for significance of the relationships extracted. The outputs from the project will include a technical report designed to train criminological researchers and policy practitioners in the benefits of the data mining methodology and the expertise that can be provided by Griffith University to apply this method. It is anticipated that, in future projects, these methods will be applied to other criminal justice databases. These analytical methods have the potential to assist targeted resource allocation across the criminal justices system in areas such as in depth risk needs assessments and criminal justice interventions aimed at reducing crime. The development of Griffith University expertise to apply these techniques should have flow on benefits for Queensland government departments that may gain benefit from data mining to uncover relationships that may support evidence-based policy decisions relating to the justice system.

### *Data*

The dataset to be analysed is data on Queensland juvenile court appearances obtained from the Queensland Department of Communities. The data has been transformed so that each record represents a single finalised court appearance with the most serious finalised offence at the appearance represented. The data includes offender demographic information (Indigenous status, gender, age, locality etc.), information on the offence committed, the number of matters dealt with at a given appearance, the resulting court order and sentencing information.

### *Analysis*

The analysis will include models of:

1. The types of offences committed by given offender demographics.
2. The court order assigned to offenders based on factors such as demographic characteristics, the offence type and the number of matters heard at a given appearance.
3. Factors that may influence desistance from further crime
4. Modelling time to re-offence for repeat offenders. This will be broken down into time between first and second appearance and time between second and third appearance based on time to second appearance. A relationship between successive times between appearances could serve to quantify a notion of high risk and low risk offenders or offender propensity.

Part of the project will involve adopting a method to test the validity of relationships discovered. Data mining techniques do not typically result in a p-value that indicates the significance of the relationship discovered. In order to address this, the data will be separated into an estimation sample set and a validation sample set. Analysis methods will be estimated on the estimation set and applied to the validation set. We intend to adopt statistical hypothesis tests on the predictive performance observed in the validation set. In the case of a categorical dependent variable, this could be based on a chi-squared test of independence between the prediction and the dependent variable. For regression applications, this could involve an analysis of variance between the prediction and the observed data, accounting for the lack of degrees of

freedom in the analysis method (given that it was estimated on the estimation dataset). As a result of this procedure, we should gain the benefit of an exploratory, data-driven process for model forming combined with a validation procedure that can test the significance of the results obtained.

In developing these models a number of data mining techniques will be adopted including neural networks, decision trees, support vector machines and association rule discovery methods. It is expected that the analysis will involve an iterative process in which the data is transformed to establish appropriate independent variables, models are estimated and validation performance measured. This process may then provide insight into other possible transformations that may improve predictive performance and the process is then repeated.

### **Project Outputs**

This project will serve as a case study for the application of data mining to crime data. A progress report will be provided at the end of six months outlining the progress and expected deliverables. A final report will be submitted at the end of the project. We will also produce a technical report to educate criminological researchers and people in the justice system about data mining, some of the more prevalent techniques that are adopted, the methodology for a data mining analysis, the results we have obtained in applying these techniques to crime data. Additionally, documentation will emphasise the ability for Griffith University to provide expertise in the application of the method for Government projects related to the use of data analysis to support evidence-based policy decisions.