Rattle and Other Data Mining Tales

Graham J Williams

Togaware Pty Ltd
Graham.Williams@togaware.com

Original publication available from www.springerlink.com.

1 A Voyage to Data Mining

My own voyage to data mining started long before data mining had a name. It started as a curiosity that a young scientist had in searching for interesting patterns in data. In fact, the journey began in 1983 as an artificial intelligence PhD student at the Australian National University, under Professor Robin Stanton.

Data mining emerged in 1989 from the database research community. They had by then established the foundations for the relational database theory that underpinned the storing of masses of data. The research question had now became how to add value to the ever growing amount of data being collected. Data mining was the answer. The question remains with us still today, and many advances have been made in our ability to extract knowledge from data.

It was not long before the machine learners and the statisticians started to become involved in data mining. For those researchers with an interest in the application of research results to real tasks, data mining provided a useful focus. Data mining was born as a discipline where new research in machine learning had application in industry and government.

My early efforts in artificial intelligence focused on knowledge representation. I used the concept of frames to implement the FrameUp language (Williams, 1989). This lead to a simple automated reasoning system called HEFFE—the household electric fault finding expert (Williams, 1989). Rather amusingly the system always thought to first ask "is it turned on at the power point?" Fortunately it then progressed to a more sophisticated dialogue, exposing some complex reasoning to diagnose faults. The research hinted at my interest in helping people and solving real problems.

I soon noticed that the main problem for such a reasoning system was in obtaining the knowledge to reason with. An initial interest in knowledge acquisition grew into a research pathway into machine learning. It was early, yet exciting days for research in machine learning in the 1980's. At the time, decision trees, and in particular the ID3 (Iterative Dichotomiser 3) algorithm

(Quinlan, 1986), were emerging as a popular and successful approach. Every paper published in machine learning seemed to reference Ross Quinlan's seminal work.

The first decision tree induction algorithm was developed by Quinlan whilst visiting Stanford University in 1978. Interestingly, at about the same time and university, a number of statisticians were developing and deploying similar ideas. Their version of the concept became known as classification and regression trees (Breiman et al., 1984). Like many great ideas, it was simple yet powerful.

For my PhD research I implementation the algorithm in the C programming language. This provided a platform for experiments in machine learning. My observations of the mathematics behind the algorithm, lead me to develop the idea of building several decision trees from the same dataset. The PhD developed the idea into an algorithm for multiple inductive learning, which was a simple idea that we might today call an ensemble approach to machine learning.

With the growing interest in data mining in the 1990's, it was not long before decision trees (and classification and regression trees) became a foundational technology for the new research area. Decision trees were an obvious technology for turning data into knowledge.

My research and keen interest in the application of technology navigated my journey towards data mining. Decision tree induction (or the top down induction of decision trees, as we used to call it), and predictive modelling in general, provided the foundation for today's data mining.

In practise we data mine using all the tools we have available. We work on real problems in extracting knowledge from massive collections of data. This is the excitement of data mining in practise. We have the opportunity to apply an arsenal of tools and techniques across multiple disciplines. It opens the door to a lifetime's learning, as we learn to learn.

2 Signposts Along the Way

2.1 Expert Systems

My early research in artificial intelligence involved the development of expert systems. The knowledge embodied in such systems was usually either hand crafted, as rules, or "discovered" through the use of knowledge acquisition tools like ripple down rules (Compton and Jansen, 1988), or through decision tree induction. It was decision tree induction that, as a machine learning researcher, took my interest.

Expert systems were a very practical technology. Whilst undertaking my PhD research I took up an opportunity that sidetracked me from my PhD for a short while, to work on practical applications. I gained many insights and much understanding. Some of these examples below provide a flavour of how practical problems can help form real research questions.

2.2 Bushfires at Kakadu

The first opportunity for some practical expert systems came with a project from the CSIRO (the Commonwealth Scientific and Industrial Research Organisation, which was then Australia's premier government funded research organisation). The Division of Land and Water Research, under Dr Richard Davis, was developing a bushfire prediction expert system for Kakadu National Park.

Kakadu is located in the Northern Territory of Australia, and experienced regular devastating bush fires often started by reckless travellers. Each year the vegetation/fuel grew unchecked and when a fire took hold, large areas of the park would be destroyed. Over centuries though the indigenous population, the Aborigines, had developed sophisticated knowledge of how to manage their environment through fire, and to avoid the devastation of an uncontrolled wildfire.

Through fires being lit at particular times under particular conditions, the Aborigines could control the type and density of vegetation into the future. Through this they then also encouraged animals to graze certain vegetation, providing the Aborigines with their daily food source. The careful management of the vegetation also ensured it maintained a suitable density for habitation and to facilitate hunting and travel.

Our task, while developing a new framework for expert systems, was to capture some of this indigenous expert knowledge within a spatially-oriented expert system. The knowledge base for the expert system was developed in collaboration with Aboriginal elders. The system was designed to predict the extent of a bushfire at any time, through spatial reasoning, and thereby allow a plan to be developed to allow for controlled burning and the appropriate management of these controlled burns.

Working as part of the team on this project was very satisfying. The project was successful, developing new technology in the form of a Prolog-based spatial expert system (Davis et al., 1985). As a young researcher I had the opportunity to publish my first paper and to present it at an international conference in France. It was awarded as the best student paper (Williams et al., 1986) at the conference.

2.3 Decision Trees and 4GLs

Continuing with this interest in the practical application of the research, I had an opportunity to lead a team of developers in building a knowledge acquisition tool into a fourth generation environment. A Melbourne-based company, BBJ Computers, had a successful fourth generation language (4GL) called Today. The Today product allowed database systems to be rapidly developed.

My task, leading a small team, was to implement the decision tree induction algorithm, building on the ideas I was developing for my PhD. It was to

be implemented as an integrated component of Today, so that customers were able to not only build their database systems, but had a tool that allowed them to discover knowledge from the data thus collected in their databases. The concept had attracted the interest of the managing director, after seeing the Australian press coverage of my award for the bushfire expert system.

After a year (1987-1988) we had a system implemented and being sold to customers in Australia and Europe. This development was an early example, perhaps one of the earliest, of incorporating knowledge discovery or data mining (though called machine learning at that time) into a database context.

2.4 Car Loans

In 1989 I had another opportunity to use my growing expertise in decision tree induction to build an expert system for industry. Through a consultancy with Esanda Finance (a subsidiary of the ANZ Banking Group), lead by Vish Vishwanathan, our task was to automate much of the decision making for granting loans for the purchase of motor vehicles.

This was a time when desktop computers were just starting to be used by the financial controllers of car yards. Previously, when a customer was asked to take out a loan to purchase a vehicle, a lending clerk would seek information on the phone. Using further information offline a decision was made. That decision was then fed back to the sales yard, perhaps that same day or within a few days.

Over a number of years, Esanda had collected data about the profitability of each of the car loans. In particular there was quite a good record of those clients who had defaulted on their payments, and those who were very reliable in their repayments. Such data provided an excellent source for building a decision tree model, which we did.

The resulting decision tree model was implemented into their loans system. This then allowed the car yard's financial controller to simply enter the clients details and obtain a decision. That decision might be yes, no, or offline. An offline decision was one where further consideration was required, and perhaps a decision by a human expert was required. The more complex cases remained the domain of the experts, and they no longer had to deal with the often less interesting, simple decisions.

We deployed the system successfully. Over many years, and into the early years of this century, I had the occasional report that the rules developed for the decision making system back in 1989 were still in operation. No doubt some of the characteristics will have changed over time, but the foundations were solid. Like many expert systems, it had become part of the furniture.

2.5 A Framework for Learning Models

Moving on from expert systems, the bigger challenge was to allow machines to automatically learn as they interact with their world. From an artificial intelligence perspective, this meant machine learning.

Machine learning can provide a framework which describes how we build models of the world. Much of what we do as computer programmers can be seen in the light of building models. Our stylised models help us to understand the world, and to facilitate our automated reasoning. Whether we write a program that implements a spreadsheet, or a program to predict medical outcomes, we are modelling what happens in the world. The models often help us identify the key elements as the foundation of our understand. Machine learning is about automatically building models of the world—automatic programming.

A formal framework that helps me picture each new machine learning algorithm (in the context that they are all about building models) has three elements: a language for representing a model; a measure that is used to assist in building a good model; and a search algorithm that seeks for a good model in a respectable amount of time. All data mining algorithms can be characterised within such a framework, and doing so allows us to compare and contrast them.

We have then a search-oriented framework for understanding data mining algorithms. The language describes how we express a model. This might be as a decision tree, or as a neural network, or as a linear formula. Once we have a language, we can often allow for an infinite number of sentences to be written in that language (i.e., each possible sentence is a candidate model expressed in that language). How then do we identify the best (or perhaps just a good) sentence/model, from all of these possibilities. That is where a heuristic search algorithms comes into the picture. Using a measure of goodness that can be applied to a model, we can heuristically search through this infinite search space for the "best" model.

This framework, and explaining a number of data mining algorithms within the framework, is covered in more detail in a recent book (Williams, 2011). The framework has served well over many years as a way to understand data mining (and originally machine learning) as the task of searching for the right model.

2.6 Ensembles of Decision Trees

In parallel to having a number of excursions into the practical development of expert systems, my PhD research (Williams, 1991) was developing the then new idea of building multiple decision trees. At the time, we were familiar with an issue that the knowledge acquisition community was exploring—the issue of multiple experts and how to capture knowledge from them and unify it into a single system. Reasoning Systems and formal systems of logic were being developed to handle conflicting assertions and deductions. However, the idea that we might build multiple decision trees was yet to have its day.

Back in the 1980's our "large" datasets were quite small compared to today's data rich world. Indeed, today, research based on such small datasets would hardly be credible. For my PhD research I had access to a dataset of

106 observations. The observations were from a collection of remote sensing data combined with local observational data. It was this small collection of data that allowed me to explore the mathematics of the algorithm in more detail, and to watch the calculations as they were being performed.

The data was provided by colleagues at CSIRO. The Australian continent was divided into 106 regions. For each of these regions a number of observations of different characteristics were made, including soil types, vegetation type, and vegetation coverage (with some data from remote sensing satellite imagery). Also recorded for each region was a measure of the feasibility for grazing cattle. It was this feasibility that we were attempting to model as our target variable.

Having coded the ID3 algorithm in C, many models were built with this tiny dataset (as well as with the other commercial datasets I had access to through my expert systems developments). A fundamental operation of the decision tree algorithm is to choose from amongst the available variables one that "best" partitions the dataset, with respect to the target variable. It had become quite clear that often, when the algorithm had to make this choice, the mathematics for making that choice gave ambiguous answers. That is, two or more variables were almost (or some times precisely) equally good according to the measure being used.

Without a clear answer, I wondered about the rationale for choosing one variable over another. The choice seemed rather arbitrary. So instead, I started to build all the equally (or nearly equally) good decision trees. The task then was to understand how best to combine them into a single model. Some structural integration was developed, as well as the idea of having the models vote for their outcome. The approach ended up being quite similar to, though rather primitive compared to, what emerged later as ensemble approaches to building models.

2.7 Next Port of Call: CSIRO

After completing my PhD I had the opportunity to join the CSIRO. In 1995, under the insightful guidance of Peter Milne, we started the first data mining research group, as such, in Australia. At the same time, another colleague, Warwick Graco, had been delivering data mining projects for one of the Australian Government departments (the Health Insurance Commission). I believe he had set up the first applied data mining team in Australia. It wasn't long though before data mining was sweeping through industry in Australia. The finance industry, the insurance industry, and government were all starting to investigate this new technology that promised to deliver knowledge from their rapidly bulging warehouses of data.

CSIRO, working collaboratively with colleagues from other institutions and industry, provided the opportunity to explore a variety of approaches for discovering knowledge from data. We explored many quite different approaches, including evolutionary systems (Riessen et al., 1997), a B-spline

based approach to multivariate adaptive regression splines (Bakin et al., 2000), and finite mixtures (Yamanishi et al., 2004), to name some.

We also developed a successful unsupervised technique that we still use today, called hot spots (Williams and Huang, 1997). In this approach to data mining of extremely large datasets we cluster our data, and then use decision tree induction to symbolically describe the clusters. This was then combined with the idea from evolutionary computing of measuring the fitness of a sub population. Our sub populations were described by symbolic rules generated from the decision trees. We referred to each sub population or cluster that scored particularly high, according to the measure, a hot spot.

CSIRO provided an excellent environment for applied research. The opportunity to work with industry drove much of the research. It also provided much satisfaction in solving real problems using data mining. We worked in finance, in insurance, and with government.

One quite strong and pioneering body of work was in the area of health. Working with the Australian Federal government and the State governments we brought together, for the first time, data from the two levels of government. The Federal government maintained data about a patient's visit to a doctor, and their prescriptions. The State governments maintained data about episodes in hospital. Bringing such data together for over 20 million people, proved to be a powerful tool for discovering, for example, hitherto unknown adverse reactions to pharmaceuticals that lead to patients ending up in hospital.

2.8 Data Treasures: Australian Taxation Office

In 2004 the Australian Taxation Office was in the early stages of a major project to substantially replace its aging processing systems. A forward looking component of the project was the formal introduction of Analytics into the system. The organisation set up a data mining team which became the largest data mining team in Australia, with 18 new staff as a corporate resource. Its aims including improved fraud identification and to assist taxpayers in meeting their obligations.

I had the opportunity to join early on to lead the deployment of the technology across the organisation. This included setting up the state-of-the-art infrastructure for data mining, based around a collection of GNU/Linux servers running mostly free and open source software (including the R Statistical Software).

My basic principle of Analytics and data mining, in any industry, is that the focus needs to be on the data miners, not on the tools. Any relevant set of tools will do, and I have found that the open source offerings are more extensive than the commercial offerings.

Part of my role was then to introduce data mining technology to a broader community of up to 150 data analysts in the organisation. This was delivered through the shared data mining infrastructure and a regular weekly meeting

of the Analytics Community of Practise, where topics covered new technology, infrastructure training, project developments, and much more.

The outcomes and benefits of just a small number of projects have been made publicly available through press releases of the Commissioner of Taxation. These have included projects that identify non-lodged tax returns (more difficult to accurately determine than one might imagine). This project, in one year alone, identified over $100 million of additional liabilities. Another publicly announced project described the benefits of an analytical system that reviews each tax return lodged (e.g., through electronic lodgments). In the context of the population of lodgments the models developed are able to identify the risk of the lodgment being fraudulent.

Most projects in such an organisation has significant impact. By ensuring taxpayers are all meeting their obligations, the overall system is fairer to everyone. Any taxpayer illegally reducing their tax liability only results in an increased burden for the rest of the population.

A range of technology is used within the Analytics capability of the Australian Taxation Office. Traditional linear regression modelling can go a long way. This is augmented by decision trees and ensembles including random forests and boosting. Neural networks and support vector machines also make a contribution in descriptive and predictive modelling. Self organising maps (SOMs), a variety of approaches to clustering, and hot spots analysis all contribute to the insights gained from the models built.

One of the more reassuring observations in working in the area of fraud detection and non-compliance, is that most people most of the time are doing the right thing. Society relies on people working together, and organisations like the Tax Office work to assist the tax payer in doing the right thing. Whenever patterns extracted from the data begin to indicate the emergence of systemic misunderstandings, then appropriate action can be undertaken to rectify the issue. This can be through advertising campaigns or by providing letters to tax payers that can better explain the tax payer's obligations. Such transparency provides positive support to the community.

The smooth running of governments tasked with the delivery of services to its citizens depends on having the finances to deliver those services. Budgets can be seriously affected by non-compliance and fraudulent activity. When data mining identifies fraud, then more serious action needs to be undertaken. This can (and does) result in the perpetrators facing the legal system.

The Tax Office actively works toward bring together all of its knowledge into a shared knowledge-based framework, to improve its deliver of service to people and the government. Like any large data-oriented organisation, and most organisations are data oriented, there remains many opportunities for the deployment of data mining. This will continue to require the skilled knowledge of the data miners, who remain a scarce resource, together with the new technology emerging from current research in data mining and machine learning.

2.9 Reflections

Over the years, the application of research to real problems and real data has facilitated the delivery of research that has real impact. It does have a cost though, in an academic context focused on counting papers. There does remain, for the practising data miner, opportunities to write research papers and to contribute to the research community. I continue to contribute as co-chair or member of a number of data mining and artificial intelligence conference steering committees, for example.

The impact of ones research and development can be measured in terms of how one changes an organisation (in addition to how our work changes the research directions). To see a single data mining project prevent significant fraud, or to make business operate more efficiently, or to save lives by identifying adverse outcomes from drug combinations, or to reduce the danger and consequences of out of control wild fires, are all important outcomes from the application of our research. Data mining has contributed strongly to each of these.

Another impact one can have is on sharing the technology we use to deliver outcomes from data mining. We have the opportunity to share with many colleagues around the world our advances through open source software. The trend of doing so is growing, but perhaps too slowly. More researchers need to see the benefits of sharing the results of their research in this way. The positive feedback and recognition from this should not be under estimated.

Also, of course, the opportunity to share experience and guide research arises with the supervision of research, often through our PhD students. More are needed to fill the growing need for data miners out there in real world projects. Data mining research, through a PhD for example, is a great passage through the straights to a fulfilling career in data mining.

In the end though, the most satisfying is to know that we are contributing positively to society, to improve and advance society, and to better understand our world. The technology of data mining is becoming accessible to a larger user base. We will increasingly see many more people able to discover new knowledge through the assistance of data mining tools.

3 Rough Waters

As a rather young and less experienced student and researcher, exploring and developing new ideas, I gradually learnt a couple of useful lessons. With a passion, one should explore their new ideas until they either deliver, or else gain the insights that tell us why they won't. Don't let others discourage the journey, though be aware of their wisdom. Also, new ideas need to be communicated with clarity and conviction. Sometimes ideas may be rejected because they have not been communicated well.

I learnt these early on.

I've described above my PhD research where I had the idea to build multiple decision trees, rather than relying on a single tree. I found that combining the decision from multiple trees gave better results, surprisingly. After quite a bit of experimentation and developing the idea, I was quite sure this was a significant new finding for the world of machine learning: combining multiple models, just like getting a team of experts to make a combined decision.

I wrote a journal paper describing some experiments in combining decision trees (Williams, 1987), and then a conference paper on what I called multiple inductive learning, or the MIL algorithm (Williams, 1988). The paper was accepted for presentation at the very first Australian Joint Artificial Intelligence Conference, held in Sydney in 1987.

I remember the session well. It was Tuesday, 3 November, 1987. Australians will know that the first Tuesday of November is when "the nation stops" for a horse race in Melbourne—the Melbourne Cup. The artificial intelligence community must not have been so enamoured with horse racing. My paper was scheduled to coincide with the running of the Melbourne Cup.

Being scheduled to clash with the most famous horse race in Australia was just the beginnings of some trepidations. Professor J. Ross Quinlan , the pioneer of decision tree induction, was to be the session chair. As a PhD student I was looking forward to the opportunity to present my research on decision tree induction to him. It is not often we have such an opportunity.

I began my presentation. It started with a review of how decision tree induction works. (I know the session chair already knew that bit, and probably most of the audience knew that bit too, but the slides had already been prepared, so I pushed on.) I then presented the idea of building a number of decision trees, and combining them into a single model. That was the MIL algorithm. The results presented in the slides clearly demonstrated the improved performance one obtained when multiple trees were combined to work together. Phew—that seemed to go okay.

The session chair took the option to ask the first question, but not before announcing the results of the Melbourne cup. (I guess some might have been following the horse race on their little transistor radios during the presentation, back in those days.) I forget the actual wording of the question, and it might actually have been a comment rather than a question. The point of building more than a single decision tree to obtain a model had not been well communicated. It seemed like a rather odd thing to do. Surely we should aim to get the simplest, best, single model.

That took something of the wind out of my sails. It was an awkward moment whilst I came to realise that either it was a poor idea or I was not particularly convincing in presenting the evidence. Maybe the interest in the Melbourne Cup was too much of a distraction.

Though a little demoralised, I stuck with the concept of building multiple models, though not with the vigour I could have. I wrote a PhD thesis around the topic (Williams, 1991), obtained my Doctorate, and moved on. Nonetheless, as I developed my career as both a researcher and consultant, over and

over again I found that the concept of ensembles was always with me and delivering results.

It was interesting to watch similar ideas emerge, from other directions. Today, of course, we have random forests (Breiman, 2001) and boosting algorithms (Freund and Schapire, 1995) that deliver excellent results for many situations in data mining. Ensembles continue to make a lot of sense—ask Netflix (Bell et al., 2009).

It's obvious now, but not then—new ideas need work. Others may take time to come along the journey with you.

4 Steaming Ahead

A key point that we come to understand in data mining and machine learning, is that the different algorithms perform their tasks admirably, but often almost equally well, or at least similarly well. If there are nuggets of knowledge to be discovered in our data, then various algorithms will give similar results. Often, from a machine learning practitioner's point of view, the difference may sometimes come down to what knowledge is discovered, rather than the accuracy of a model. We know full well from computer science and artificial intelligence, that a change in how we represent a problem or represent our knowledge, may be the difference between solving a problem, or not, or gaining insights from the knowledge, or not.

The key trick, in all of data mining—whether it be text mining, predictive analytics, or social network analysis—the key to successful data mining is to live and breathe the data. Grasp the data and turn it into a form that can be data mined (generally flat tables). Then apply the depth and breadth of our tools to the data to build new models of the world. Once we have the right set of features, irrespective of the structure of the original data (database, text, links, audio, video), building models is simple, and any tool might do.

How we represent our discovered knowledge, and how we combine discovered knowledge into universal knowledge bases that can reason about the world, continues to be a goal for a lot of research. It is a worthy goal.

On a more concrete footing, researchers will have their pet ideas about the most interesting areas for research in the next few years. I won't labour the near future too much, because the interesting research around data mining is, I think, for the longer term. Over the next few years we will see much the same from data mining research as we have for the past few years. New tweaks on old algorithms will continue to be implemented. And new areas of application will lead to some variety in how we think about the algorithms we are using. We might make some steps also toward better representations of our learned knowledge.

Specific areas that I see continuing to grow strongly include social network analysis and text mining, along with mining of other media such as video and audio. But data mining should become more personal. The personal mining

of podcasts, for example, to find those that might be of interest to us, may be a key example of the kind of challenge. It will continue to replicate how we mine spatial, temporal, relational and text data, where we extract a textual representation from the original representation, and turn that into a flat structure which we then data mine. This continues to work quite well and the sophistication is in how to actually extract the "data" from these different representations. I look forward to further research advances around how we do this most effectively.

Another area of growing interest is in the sharing and deployment of models. A very early meeting in Newport Beach, California (1997), at the Third International Conference on Knowledge Discovery and Data Mining (KDD97), introduced me to the concept of a new standard for exchanging predictive models amongst different tools. PMML, or the Predictive Modelling Markup Language (Guazzelli et al., 2009, 2010), has developed over a the years to become, now, a mature standard for the interchange of models. Active research continues to ensure the PMML standard captures not only the models, but also the transforms required on our data to be used by the model.

PMML is important because it allows us to build models, and have them deployed on other platforms. These other platforms may be carefully tuned deployment platforms that can score large volumes efficiently, as demonstrated by the ADAPA real-time PMML-based scoring tool. Open standards that allow the interchange of models between closed source and open source software will be increasingly important.

There is also a growing appreciation of the Analyst First movement. The movement recognises that the first most priority is with the analyst or the data miner, not the tools being used. A corollary of the movement might be that the freely available open source tools, in the hands of the most skilled analysts, will deliver more than the most expensive data mining tools, in the hands of the less skilled analyst. Thus, a focus on open source data mining has brought me on quite a journey, to the development of Rattle (Williams, 2011).

5 Charting the Route: Open Source Tools

Data mining is becoming a technology that is freely available to anyone who wishes to make use of it. Over the years the technology has only been generally available to researchers and through large statistical software vendors. The vendors have been able to command significant prices for what is essentially a collection of easily implemented algorithms. But large organisations (the customers of the software vendors) have not understood what data mining is. These customers have been sold on the idea that software, rather than the skills of an analyst, is what makes a difference. It would be nice if that were true (see Section 6), but we have a long way to go yet.

The algorithms which are commonly implemented in a data mining suite have included kmeans clustering, association rules, decision trees, regression, neural networks, and support vector machines. Each of these basic data mining algorithms is actually quite simple to implement (for a software engineer) and I (and many others) have implemented, over the years, versions of each of them.

All of the data mining algorithms are available freely as open source software. Indeed, newly developed algorithms, emerging from the research laboratories, are now often available freely as open source software for many years before the vendors are able to catch up. The problem, though, is the lack of skilled people to be able to make use of the tools. Many organisations are instead quite happy to spend millions on the commercial products to deliver the mythical silver bullet, rather than investing in the people who can more effectively deliver with the technology.

With the growing popularity of the free and open source R statistical software (R, 1993), widely tested and used implementations of all of these data mining algorithms have been available for many years. Other free and open source offerings include Weka (Witten and Frank, 2005), written in Java, KNIME (Berthold et al., 2007), and RapidMiner (Mierswa et al., 2006). However, the problem remains that we need to have a high level of skill to use many of these tools.

On joining the Australian Taxation Office, to lead the roll out of data mining technology across a very large data rich organisation, I quickly realised the issue facing many such organisations. There was a large population of quite skilled data analysts, quite happy with extracting data from large data warehouses, but not familiar with the advances in data mining that could quite quickly add considerable value to that data. Expensive data mining software could not be most effectively used because of a lack of understanding of how data mining worked. The first project or two delivered quite good results, but as the technology began to be employed seriously, the limited depth of knowledge about data mining began to inhibit progress.

I also began questioning the use of the commercial and closed source software when the same functionality was freely available. Using R I was able to replicate all of the modelling performed by the expensive tools. But there is no question that R is a programming language, for what I call programming the data analyses, and requires a high level of skill. Data mining is about living and breathing our data, and not about pushing buttons to get results.

With the goal of providing more data miners with access to the most powerful suite of data mining algorithms, and providing an approach to facilitate an understanding of what is being done, I began work on Rattle (Williams, 2009, 2011).

5.1 A Beacon to Light the Way: Rattle

Rattle provides a very simple and easy to use graphical interface for data mining. Whilst others have delivered quite sophisticated, attractive interfaces for data mining, including the very common process diagram interfaces, Rattle continues to provide a basic but readily usable interface. The real focus is on migrating the user from the basics of data mining to the full power of programming with data. The goal is to turn a data analyst into a data miner, augmenting an SQL programming skill with the power of a fully fledged statistical programming language like R.

Over 6 years of development, Rattle has become a mature product, freely available. (It is also available as a plug-in for a commercial business intelligence product from Information Builders). It is used for teaching data mining and widely used by consultants for delivering data mining projects, without the overhead of expensive software.

Rattle can be used by anyone! It is a simple installation. With a few clicks (after we have the right data) we can have our first models. But of course, that is simply the beginning of a long journey to becoming a skilled data miner programming our analyses in R. As we gradually allow these skills to become more readily accessible, the algorithms will become "second nature" and part of the common toolbox.

5.2 Are We There Yet: Freely Receive and Freely Give

Data mining is essentially a practical application of research to real world analysis of data. We will continue to see a growing number of research papers published, incrementally increasing our global knowledge. But the real impacts are when we apply the technology. This can be where we can make significant impacts and change how organisations do things. It can be very satisfying. But the real key to this, I believe, is in freely making available the fruits of your research to all.

Scientific research has always held high the principle of repeatability and sharing of results. Peer review is important for any scientific endeavour, and advances are made by freely discussing and critiquing our work with others. Today's focus on commercialising everything leads us to hide and protect everything we do, just in case we have the big winner that will ensure our personal financial security.

Financial security is important, but so is the need for our society as a whole to benefit from all the good that we can offer. Too often I have seen students, perhaps encouraged by their advisers or their institutions, to not make their implementations of their research available, just in case they can make some significant money from it. The majority of this code then simply disappears forever. What a waste. Even simply hiding the implementations for a year or two can waste other peoples time and effort that could be better

spent on pushing forward the science rather than unnecessarily replicating the implementations.

For the good of all of us, do consider making your software implementations of new algorithms available. We can then try out your algorithms on our data and repeat your experiments to see how well your results generalise to other problems. We can more efficiently compare your approach to other approaches, in unbiased experiments, and share the results of these experiments, to help improve results all round.

My recommendation is to package up your algorithm to make it available, for example, in R or in Weka, or simply out there.

6 Rowing in Unison: A Bright Future

Perhaps the most interesting thing to do every now and again is to sit back and wonder where the world is heading. Many writers and researchers do this and it is instructive, though not necessarily accurate, to do so. As Alan Kay (inventor of the computer desktop-window-mouse, paradigm we are still using today) said in 1971, "The best way to predict the future is to invent it." Science fiction provides a most fruitful avenue for exploring possible futures, and developing an understanding of the consequences. Some predictions come true, some don't, but the fact of exploring the ideas and possibilities, influences the very future we are heading towards.

For me, I see the longer term future of data mining leading towards delivering on some of the goals of machine learning and artificial intelligence—goals that have been around since the 1950's: for our computers to behave intelligently by learning from their interactions with the world. But the world is going to be quite different. We need to put data mining out there as a technology for accessible by all who wish to do so. And where we obtain the data to be mined will be radically different to where it is today (centralised versus distributed).

Data mining technology must become common place and even disappear into the fabric of our daily life. The story was the same for expert systems of the 1980's. Expert systems are no longer specifically talked abut, but are ever present. They underpin many decision making systems in use today, from supporting doctors in the interpretation of medical images, to deciding whether you are a good customer for the bank.

The research for this is not necessarily about better algorithms for machine learning. We will, nonetheless, see further incremental improvements in our algorithms over time. We may see some significant conceptual shifts in our paradigms that suddenly make significant advances in collecting and using knowledge.

The direction I see that is needed for data mining is more about how we freely make the technology available to anyone. Rattle is a small, but practical, attempt to move in this direction, providing a free and open source

tool. Others include KNIME (Berthold et al., 2007) and RapidMiner (Mierswa et al., 2006). The aim is to make it simple for the less statistically and computationally sophisticated to do the right thing in analysing data.

A longer term goal in many areas of research relating to modelling and computer science has been in intelligent tool selectors (or intelligent model selection). In the data mining context, given a dataset for analysis, the user needs to be guided accurately and with statistical validity, in the right direction. The intelligent guidance provided by the data mining platform of the future will consult, collaborate, and co-design[1] the analysis with the user. The ensemble of the data mining tool and the expert user will work to deliver new knowledge from the ever growing supplies of data.

7 New Horizons: Private Distributed Data

Reflecting on how we are moving forward with the management of data into the future, we might start to see some interesting trends, and postulate some interesting directions. Today, we have moved into a phase of the deployment of technology which is very much oriented around putting commercial interests well and truly before individual or societal interests. The pendulum is about to start swinging the other way, we hope, for the good of society.

The concepts of cloud computing and our willingness to (perhaps unwittingly) hand over so much personal data to be controlled by social network hubs, have considerable currency. Facebook, as an example, probably makes a claim to the ownership of all of the data that hundreds of millions of users are freely providing to it. The owners of these massive data stores may, at their own discretion, or under duress from well intentioned (or not) authorities, decide to do with that data whatever they like, without recourse or reference to those who supplied the data. The Wikileaks episode of 2010-2011 made this very clear, with US authorities requiring social networking sites to hand over related data.

We now see, though, a growing interest and concern for privacy amongst the general users of the Internet. Consequently we are also seeing the emergence of alternative, privacy preserving, cloud and social networking applications. There is starting to emerge new technology that allows a user, or a household, to retain all of their data locally within a tiny plug device, or perhaps within a personal smartphone. The technology allows these same users to continue to participate actively in the networked social communities as they do now. Plug devices (or smartphones), and the Debian-based Freedom Box,[2] are technology pointing the way to a possible future.

Such a device will be low powered with many days of battery charge in case of outages, and will connect through wireless or Ethernet to the Internet.

[1] A motto borrowed from the Australian Taxation Office: `http://www.ato.gov.au/corporate/content.asp?doc=/content/78950.htm`

[2] `https://freedomboxfoundation.org/`

The network connection will normally be via an ISP over the phone line, as now, but with backup through the mobile network (3G and 4G), and further backup through local wireless mesh networks. The mesh network operates by connecting to your neighbours device, which connects to their neighbours, and so on.

Under this scenario, data will be distributed. All of my emails, all of my social networking comments and photos, all of my documents, all of my music and video, all of my medical records, and more, will live on the small, high capacity, but very energy efficient, device. All of this will be securely available to myself, wherever I am anywhere on the network, whether through a desktop computer or my smartphone. The data will also be available to those to whom I give access—perhaps to share social network type interactions, or for my doctor to review my complete medical history. I retain ownership of my own data, and provide it to others explicitly, at my discretion.

There will also be intelligent agents developed that will look after the security and backup of the data stored on the device. The intelligent agents will monitor and report to me who is accessing the data, to keep me aware of such things, when I want to be.

A new economy will develop. My intelligent agent, looking after my interests, will negotiate with centralised services, such as with Google, perhaps, to share some of my data. A service like Google may then use this data for large scale analysis, perhaps for targeted advertising, or improving searching and spam filtering, etc. Those negotiations may involve financial rewards (or enticements) when I make some of my data available. But I have the ultimate control over what data I make available and when I make it available.

As this scenario begins to develop over the coming decade, how we do data mining will fundamentally have to change. Some advocates of the distributed data model rally against the prospect of data mining, seeing the distributed data as a mechanism to defeat data mining. However, we should not allow data mining to become the enemy, but instead, as I describe in the previous section, allow all to understand what data mining is about, and even facilitate many more to easily access the technology. We can then choose to opt in to have our data analysed as part of the world wide population, for benefits that we clearly understand and agree to, but to be in control of when we make it so available. We may receive rewards or enticements to do so, or other benefits, but we emphasise that it is under our personal control.

The data mining of distributed data in the new distributed personal server world will present many interesting challenges. Can we, for example, perform distributed data mining at this fine grained level whist not removing the raw data from a users personal server? Or can we guarantee anonymity in the analysis of the data as it is scooped from a million personal servers? There may need to be developments around how our intelligent agents will collaborate with several million other agents to allow patterns of interest to be discovered, using distributed and privacy preserving approaches to data mining.

In summary, we will see a movement of individuals, slowly but surely, moving from centralised services to distributed, personal services. Individuals will again own their personal data, and can make that data available on a per request basis, accepting payment or other benefits for making the data available. The data remains on their own personal server, but it is queried by the data mining tools across the network, with the network acting as a large distributed database. This early phase of the movement provides an opportunity for some pioneering work in distributed, fine grained, privacy preserving, data mining, linked also to the concept of intelligent agents looking after the interests of the individual personal servers.

8 Summary

We continue to see an increasing demand for data miners and data mining tools. There is today a lot of active research in data mining, though it is an increasingly crowded space. Advances in data mining research deliver very small steps, one step at a time.

Pondering over where computers and technology are heading, after watching the incredible growth of social networking, we begin to see the amazing uptake of smart mobile devices (smartphones). We are now beginning to see an increasing concern for privacy and questions being raised about the ownership of data. New technology will arise over the coming years which will deliver personal servers (plug devices and smartphones) to individuals and households, where data remains with the individual. The data of the world will be very fragmented and finely distributed. Significant new challenges for data mining arise in this scenario, and little research has focused on data mining of such fine grained sources of data.

To finish this journey, let me recount something I once heard Marvin Minsky (one of the pioneers of artificial intelligence) say: when a research area gets crowded, its time to move on. Look for new pastures where you can make a major splash, rather than staying within the crowded stadium with much competition and only little progress for each step forward. Invent the future and develop the technology that is needed to work there.

Finally, ensure that our future is free—make your research and algorithms and implementations freely available for all to benefit. Allow all the choice to participate.

References

Bakin, S., Hegland, M. and Williams, G. J. (2000), 'Mining taxation data with parallel bmars.', *Parallel Algorithms Appl.* pp. 37–55.

Bell, R. M., Bennett, J., Koren, Y. and Volinsky, C. (2009), 'The million dollar programming prize', *IEEE Spectrum* **46**, 28–33.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K. and Wiswedel, B. (2007), KNIME: The Konstanz Information Miner, in *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*, Springer.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification and Regression Trees*, Wadsworth and Brooks, Monterey, CA.

Compton, P. and Jansen, R. (1988), Knowledge in context: a strategy for expert system maintenance, in *Proceedings of the 2nd Australian Joint Conference on Artificial Intelligence*, pp. 292–306.

Davis, J. R., Nanninga, P. M. and Williams, G. J. (1985), Geographic expert systems for resource management, in *Proceedings of the First Australian Conference on Applications of Expert Systems*, Sydney, Australia.

Freund, Y. and Schapire, R. E. (1995), A decision-theoretic generalization of on-line learning and an application to boosting, in *Proceedings of the Second European Conference on Computational Learning Theory*, Springer-Verlag, London, UK, pp. 23–37.

Guazzelli, A., Lin, W.-C. and Jena, T. (2010), *PMML in Action*, CreateSpace.

Guazzelli, A., Zeller, M., Lin, W.-C. and Williams, G. (2009), 'Pmml: An open standard for sharing models', *The R Journal* **1**(1), 60–65. `http://journal.r-project.org/2009-1/RJournal_2009-1_Guazzelli+et+al.pdf`.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M. and Euler, T. (2006), Yale: Rapid prototyping for complex data mining tasks, *in* L. Ungar, M. Craven, D. Gunopulos and T. Eliassi-Rad, eds, *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia, PA, USA, ACM, New York, NY, USA, pp. 935–940.

Quinlan, J. R. (1986), 'Induction of decision trees', *Machine Learning* **1**(1), 81–106.

R (1993), *A Language and Environment for Statistical Computing*, Open Source. `http://www.R-project.org`.

Riessen, G. A., Williams, G. J. and Yao, X. (1997), Pepnet: Parallel evolutionary programming for constructing artificial neural networks, *in* P. J. Angeline, R. G. Reynolds, J. R. McDonnell and R. Eberhart, eds, *Evolutionary Programming VI*, Vol. 1213 of *Lecture Notes in Computer Science*, Springer Verlag, Indianapolis, pp. 35–46.

Williams, G. J. (1987), 'Some experiments in decision tree induction', *Australian Computer Journal* **19**(2), 84–91. `http://togaware.com/papers/acj87_dtrees.pdf`.

Williams, G. J. (1988), Combining decision trees: Initial results from the MIL algorithm, *in* J. S. Gero and R. B. Stanton, eds, *Artificial Intelligence Developments and Applications: Selected papers from the first Australian Joint Artificial Intelligence Conference, Sydney, Australia, 2-4 November, 1987*, Elsevier Science Publishers B.V. (North-Holland), pp. 273–289.

Williams, G. J. (1989), 'Frameup: A frames formalism for expert systems', *Australian Computer Journal* **21**(1), 33–40. http://togaware.com/papers/acj89_heffe.pdf.

Williams, G. J. (1991), Inducing and combining decision structures for expert systems, PhD thesis, Australian National University. http://togaware.com/papers/gjwthesis.pdf.

Williams, G. J. (2009), 'Rattle: A Data Mining GUI for R', *The R Journal* **1**(2), 45–55. http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Williams.pdf.

Williams, G. J. (2011), *Data Mining with Rattle and R: The art of excavating data for knowledge discovery.*, Use R!, Springer, New York.

Williams, G. J., Davis, J. R. and Nanninga, P. M. (1986), Gem: A microcomputer based expert system for geographic domains, in *roceedings of the Sixth International Workshop and Conference on Expert Systems and Their Applications*, Avignon, France. Winner of the best student paper award.

Williams, G. J. and Huang, Z. (1997), Mining the knowledge mine: The hot spots methodology for mining large real world databases, in *Advanced Topics in Artificial Intelligence*, Springer-Verlag, London, UK, pp. 340–348.

Witten, I. H. and Frank, E. (2005), *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn, Morgan Kaufmann, San Francisco. http://www.cs.waikato.ac.nz/~ml/weka/book.html.

Yamanishi, K., Takeuchi, J.-i., Williams, G. J. and Milne, P. (2004), 'Online unsupervised outlier detection using finite mixtures with discounting learning algorithms', *Data Mining and Knowledge Discovery* **8**, 275–300.

Index