

# Proceedings of the UK Symposium on Knowledge Discovery & Data Mining 2009



**Editor: Sunil Vadera**

ISBN: 978-1-905732-69-2

# The Fifth UK BCS Knowledge Discovery and Data Mining Symposium

## Organising Committee

Sunil Vadera (chair),	University of Salford
Frans Coenen,	University of Liverpool
Trevor Martin,	University of Bristol
George Smith,	University of East Anglia
Alex Freitas,	University of Kent

## Sponsors

The British Computer Society  
BCS Specialist Group on AI (BCS-SGAI)  
SYS Consulting Ltd  
University of Salford

<b>Contents</b>		<b>Page</b>
Sunil Vadera,	<i>Foreword</i>	2
	<i>Brief Presenter Biographies</i>	3
Nello Cristianini,	<i>Scientific Method and Patterns in Data</i>	6
Eduarda Rodrigues,	<i>Mining Online Communities</i>	15
Tony Bagnall,	<i>Time Series Data Mining</i>	21
Gavin Brown,	<i>Feature Selection by Filters: A Unifying Perspective</i>	34
Neil Berry,	<i>The Cook, the Thief, his Wife and her Lover</i>	44
Susan Craw,	<i>Knowledge Discovery from Case Data</i>	58
Sven F. Crone,	<i>Classifying Imbalanced Datasets</i>	66

## **FOREWORD**

Welcome to the 2009 UK Symposium on Knowledge Discovery and Data Mining.

The symposium aims to bring together researchers and practitioners who are interested in advancing the field of data mining and knowledge discovery, sharing their experience of developing state of the art applications and debating some of the scientific, engineering, societal and ethical issues surrounding the field.

The event adopts the successful format of the previous events held at the University of Liverpool, University of East Anglia, University of Kent, and the University of Bristol, where invited speakers present seminars aimed at dissemination of new research, sharing of experience gained in developing state-of-the-art applications and discussion of major issues in data mining.

The topics covered in the symposium represent some of the most important and exciting issues in the field. There are presentations on fundamental research topics such as how data mining is changing the very nature of scientific methods, the challenges of time series data mining, use of social network analysis for classification of messages, knowledge discovery from case data, and development of a unifying framework for feature selection methods. There are also presentations describing the lessons learned from real world case studies in detecting financial crime, profiling electricity usage, image processing, credit scoring, and predicting internet shopping patterns. These exciting topics are, of course, only possible because of the willingness of the invited speakers to share their knowledge and expertise.

I'd like to thank several people, without whom this symposium would not have been possible. I'm most grateful to Frans Coenen, who founded the series, provided encouragement, suggested speakers, promoted the event and is the driving force behind the series. I am also grateful to Trevor Martin who shared his experience of organising the event in 2008, and the previous organisers, George Smith and Alex Freitas for their support. Nathalie Audren-Howarth provided excellent administrative support and Louise Heatley developed a wonderful web site that led to positive comments from several users.

Financial sponsorship for the event was provided by the University of Salford, the British Computer Society and SYS Consulting Ltd.

We hope you find the event useful, get an opportunity to meet others in the field and enjoy the day.

Sunil Vadera

(On behalf of the KDD'09 Organising Committee)

## **BRIEF PRESENTER BIOGRAPHIES**

### **Nello Cristianini, University Of Bristol**

Nello Cristianini is a Professor of Artificial Intelligence at the University of Bristol since March 2006, and a holder of the Royal Society Wolfson Merit Award. He has wide research interests in the area of computational pattern analysis and its application to problems ranging from genomics, to computational linguistics and artificial intelligence systems. He has contributed extensively to the field of kernel methods. Before his appointment at Bristol, he held faculty positions at the University of California, Davis, and visiting positions at the University of California, Berkeley. He has been Action Editor of the Journal of Machine Learning Research (JMLR) since 2001, and Associate Editor of the Journal of Artificial Intelligence Research (JAIR) since 2005. He is co-author of the books 'An Introduction to Support Vector Machines' and 'Kernel Methods for Pattern Analysis' with John Shawe-Taylor, and "Introduction to Computational Genomics" with Matt Hahn (all published by Cambridge University Press).

### **Eduarda Rodrigues, Microsoft, Cambridge**

Eduarda is a Researcher with the Integrated Systems group at Microsoft Research Cambridge. Her research interests lie in the broad areas of data mining and web information retrieval. In particular, her current work is focused on social network analysis, online communities, web link analysis and text mining.

She obtained the *Licenciatura* degree in Electrical and Computer Engineering from the University of Porto, Portugal, in 1998 and a Ph.D. degree in Electronic and Electrical Engineering from University College London, UK in 2005.

Prior to joining Microsoft Research, she was a Research Fellow at the Electronic and Electrical Engineering Department, University College London, working on Web and Grid services for large-scale data analysis. Before initiating her Ph.D. studies, she was a research engineer at the Institute for Systems and Computer Engineering of Porto (INESC Porto), Portugal, working on distributed systems and multimedia applications.

### **Tony Bagnall, University of East Anglia**

Tony joined the School of Computing Sciences at the University of East Anglia as a part time MRes student/part time teaching assistant in 1993. After completing his Masters by research in 1995 he began a PhD titled "Modelling the UK electricity market with autonomous adaptive agents". After a brief period as a research assistant on a data mining project sponsored by Master Foods and Centrica, in 1999 he completed his PhD and was appointed

as a Lecturer in Statistics for Data Mining. In 2007 he was promoted to senior lecturer.

Tony has been involved in researching areas of optimization, machine learning, agent systems, statistics and data mining. Currently he is primarily focused on time series data mining, a topic of his presentation at UK KDD'09.

### **Gavin Brown, University of Manchester**

Gavin is a member of the Machine Learning and Optimization Group at the University of Manchester.

His research interests can be summarised as: feature selection/extraction with information theoretic methods, Markov Blanket algorithms, ensemble learning (aka multiple classifier systems), and online learning. All of the above in application to two domains: Systems Biology and adaptive compiler optimisation. Gavin is a member of the IEEE Technical Committee on Intelligent Systems Applications and on the programme committees of several conferences including the International Conference on Pattern Recognition (2008), Genetic and Evolutionary Computation Conference(2005-2008), Conference on Recent Advances in Soft Computation (2006), and International Joint Conference on Neural Networks (2005-2008).

### **Susan Crow, Robert Gordon University**

Susan Crow joined the School of Computing at the Robert Gordon University in 1983 (then Robert Gordon's Institute of Technology or RGIT) as a lecturer.

In October 1986 she registered for a part-time PhD in Computer Science at Aberdeen University; completing it in early 1991. Later that year, she was seconded for one year as a senior research fellow to the University of Aberdeen where she worked on the MLT ESPRIT project. On her return to the Robert Gordon University in 1992 she was promoted to senior lecturer. She was awarded the title of Reader in 1996, and Professor in 1998. In 1999 she won a Fulbright Scholar award and spent it on sabbatical in ICS at UCI.

She became Head of the School of Computing in 2001 and Head of Research & Graduate Studies for the Faculty of Design & Technology in 2006.

Her research develops machine learning and data mining techniques to discover knowledge for intelligent, decision support and product design software systems. Case-Based Reasoning is a major focus of her research, and she builds automated tools to maintain the case knowledge and to discover knowledge to improve CBR retrieval and reuse.

## **Neil Berry, Deloitte**

Neil Berry is Director of Data practice at Deloitte, the business advisory firm.

Neil's previous role was as Lead Architect within Capgemini's Business Information Management practice where he worked for 6 years. Neil led the SAS (Business Intelligence Software) practice at Capgemini focusing on data management, business intelligence and information architecture.

Neil Berry's specialties include: CIO Services, Architecture, SOA / SOE / SOI, Business Intelligence, Data Warehousing, Business Information Management, Identity Analytics, Fraud & Audit, Business Development, Outsourcing, 3rd party vendor management.

## **Sven F. Crone, Lancaster University**

Sven F. Crone has over 8 years of experience in forecasting. He is currently working as an Assistant Professor (Lecturer) at the department of Management Science at Lancaster University Management School. He also serves as the deputy director of the Lancaster Centre for Forecasting.

Sven received his Dipl.-Kfm. (MBA and BBA equivalent) and PhD (pending) in forecasting with neural networks from Hamburg University, Germany, and was a visiting fellow at Stellenbosch Business School, South Africa, and George Mason University, USA. His research focuses on forecasting and data mining applications, frequently using methods from artificial intelligence. He has authored over 25 articles in international journals and conference proceedings and has given over 45 international presentations in the field.

Some of his recent projects at the Lancaster Centre for Forecasting include event and weather based forecasting with statistical methods for retailer TESCO UK, copper price forecasting with artificial intelligence for producer CODELCO Chile, forecasting advertisement ratings for TV company ITV UK, and implementing international demand planning methods, processes and systems in SAP APO-DP with FMCG producer Beiersdorf AG.

# Scientific Method and Patterns in Data

Nello Cristianini  
University of Bristol

## Abstract

*The way we do science is changing fast, due to large scale data analysis technologies. The automated gathering of massive amounts of data, followed by their automated analysis, is becoming mainstream in many sciences. The knowledge extracted by machines may not even need to be readable by humans, as it can be used by other machines, for example in the design of new experiments. The very foundations of scientific method are undergoing a transformation, and notions like Theory and Model are under discussion. New notions, like Pattern and Predictive Rule, may be taking their place. The Social Sciences may be the next conquest, after Biology and Physics, of this new way of doing science.*

## Introduction

In the summer of 1609, nearly exactly 400 years ago, Galileo Galilei was in Venice, trying to sell his telescope to the Doge, in return for tenure. He had not really invented it, as this was the creation of Dutch spectacle-maker Hans Lippershey, of which he had heard a description. He greatly developed it, and offered it to the Venetian Fleet as an aid to navigation and early detection.

During 1609 Galileo perfected his lens grinding skills, experimenting with methods and designs. He created various models and investigated the principles behind optics. He could have started a business making telescopes, or magnifying glasses, or spectacles. He could have been satisfied with the wage he received from the Republic of Venice.

But Galileo was a scientist, not just a tool maker. Although he did design, create, and test some of the best tools of his time, he was not just concerned with the engineering aspects of his work, and the commercial opportunities. As a true scientist, he was interested in understanding the world around him, something that would get him into trouble more than once. Technical and scientific advances sometimes can have profoundly disruptive effects on society at large.

So in the summer of 1609, at age 45, he turned the telescope to the sky, and started his investigation of the Moon. He discovered mountains and valleys, by observing their changing shadows. Most importantly he discovered that the Moon – contrary to Aristotle's opinion – was not a perfect sphere. Something was wrong with the established model of the Universe.

Later on he discovered with the telescope that Jupiter was orbited by 4 Moons, and this showed that in at least one case, celestial bodies did not revolve just around the Earth. Then with the same tool he discovered that Venus has phases, just like our Moon.

In fact, he realised, Aristotle was wrong, the Earth and Venus and Jupiter orbited the sun, the Moon orbited the Earth, like the 4 Moons of Jupiter orbited their planet. Furthermore our Moon – at least – was not a perfect sphere, but had mountains, and he could infer their height by measuring their shadows, and predict which of them

would come out of the dark first, every month. What he had been taught about the Universe was incorrect.

His work was published very rapidly, in March 1610, in a short booklet entitled “Sidereus Nuncius” (*Starry Messenger*). His work with lenses and telescopes was important not because it had direct implications on how we did things on Earth – although that too – but because it was eventually responsible for a fundamental revolution of our thinking. Its implications were theological, and landed Galileo into trouble with the Church, among other things. His observations forced him to question the received wisdom, and this is always an act of challenge, although one that is expected of scientists. These implications were also philosophical, and methodological.

In fact, that was a very early example of modern systematic scientific investigation: a scientific instrument was developed and used to make observations, mathematical relations were derived for the geometry on the Moon, and predictions were used as a way to validate the models. A key assumption was that the same laws (of geometry for example) must apply on the Moon as on Earth.

For this and many other contributions, Galileo is associated with a major shift in scientific method, although others were thinking along the same lines at the time.

### **Scientific Method.**

The systematic method we use to derive and represent unambiguous knowledge, so that it has predictive and explanatory power over the world, is a major achievement of our culture. Not all cultures focused on a systematic approach to knowledge acquisition and revision, see for example the Romans. There are many ways of knowing the world, and the scientific method is a systematically organised procedure to produce knowledge that is reliable, and remove that which is not.

Over the centuries, we have started gathering knowledge in an organised process, involving a cycle of experimental design and hypothesis generation, representing the results – wherever possible – in unambiguous mathematical terms. In fact, certain branches of mathematics have grown just to accommodate this new role that mathematics had in modelling (while its origins were just in calculation). This has been the accepted way in which we do science for the past few centuries, but is not the only possible way.

In fact, the scientific method has been in constant evolution for a long time. The same can be said of the practices we follow as a research community, with anonymous peer review and publication of results being a crucial part of the current ritual of science. Observations lead to competing models, and this leads to experiments, and their outcomes are used to revise the current models, and this in turn suggests new experiments, and so on, in a cycle. The discovery of the laws of mechanics – for example - can be seen in this light, with competing intuitions about mass, acceleration and friction, leading to key experiments. In most cases, these feedback loops are much more complex and interconnected, but the interactive nature of the modelling process is often very visible.

But things are changing fast. Now the process is going through an “industrial revolution” of its own. Data are gathered automatically, by computers and robots,



effectively acting as massive measurement apparatuses, replacing what were for Galileo the thermometer or the clock. Increased accuracy and the ubiquity of measurement devices result in ever larger repositories of experimental data, stored in dedicated disk farms.

We can look at the examples of Physics, Genomics, Drug Design and Astronomy. They all exemplify the same trend in science.

The Large Hadron Collider at CERN is a machine designed to produce experimental data, potentially 15 Petabytes per year [Duellmann, 2007]. The engineering challenges in producing, storing and managing this amount of information have reached awesome proportions. But it is the analysis of this data that is truly mind boggling. And this experiment can be seen – in a way – as the direct descendent of physical experiments initiated 400 years ago by Galileo: the systematic investigation of the basic laws of nature has led us to this point.

Similar challenges are encountered by today's biology. The direct descendents of Mendel's painstaking collection of genetic inheritance data are experiments aimed at the full sequencing of thousands of genomes at once. Terabytes of data are now produced by each of the new generation of sequencing machines, and the Sanger Centre in Cambridge is now working on the 1000 genomes project [1000genomes.org, 2007]. Hundreds of species have now been fully sequenced, and we are well down the road of comparing multiple complete sequences within the same species.

In drug design, it is standard to test compounds to see if they bind to a given target, by exhaustively testing entire libraries of chemicals, by use of robots, in what is called combinatorial chemistry [DeWitt, 1995]. Hundreds of thousands of compounds can be generated and tested, either by using robotics, or – increasingly – even by computer simulations, in what is essentially a survey of entire regions of chemical space, hunting for compounds with a given set of properties.

Astronomy – another child of Galileo's – is now done by automatic surveys of the night-sky run by computers, and by subsequent automatic analysis of the images and data gathered in this way. One such project is the Sloan Digital Sky Survey (SDSS), which created a 5-wavelength catalogue over 8,000 square degrees of the sky, containing about 200 million objects, described by hundreds of features (data released incrementally to the public [Adelman-McCarthy, 2008]). The SDSS used a dedicated 2.5-meter telescope at Apache Point Observatory, New Mexico. The 120-megapixel camera imaged 1.5 square degrees of sky at a time, about eight times the area of the full moon. A pair of spectrographs fed by fibre-optics measure spectra of (and hence distances to) more than 600 galaxies and quasars in a single observation. The database generated over 8 years by this automated survey is several Terabytes large, presenting serious challenges to data management and mining [Adelman-McCarthy, 2008].

In fact, we should consider this point in all its disruptive implications, that directly challenge normally accepted assumptions. Taken to its extreme consequences, its implications to Epistemology are significant. There is no way that people can analyse the data produced at LHC, or at Sanger Centre, or by sky surveys. They can only be conceived because we can rely on computers to do the analysis of data for us.

And this is the point we are considering: our scientific method has changed. The revolution is not a matter of detail, or even quantity. It is a matter of quality. We have industrialised both the production and the analysis of experimental data. We have industrialised the generation of scientific knowledge. And this will not just lead us to a significant acceleration of knowledge acquisition in the future, by virtue of the automation of the feedback loop, but it also invites us to re-examine what scientific laws and models actually are.

The **automatic analysis of patterns in data**, the automatic generation of hypotheses, is a fundamental part of science. This is how computer science, statistics, and also artificial intelligence, are finding their way to the core of all science, and to the core of how we know our world. This is how automated pattern analysis found itself at the centre of a revolution that will have far reaching consequences.

### **A Newer Method.**

The automatic analysis of data, in search for significant – if elusive - patterns, is now a key part of many scientific experiments, and this is an increasing trend. Statistics and computer science, and the convergence of dozen of smaller disciplines, create a conceptual and technical framework and body of knowledge that we call Pattern Analysis. It includes tools to extract significant information from networks, images, strings, text, bio-sequences, vectors, time series, and any other form of data that scientists routinely analyse and model.

We may think that the process of scientific discovery will not be fully automated until machines will be able to generate complete theories of a domain, with their formalism and equations. This deserves 2 fundamental responses: 1) this is not necessarily out of reach for machines 2) this is not necessary for machines to be doing science.

As for Point 1, I will just point to a line of research, represented by [Schmidt, 2009] where various search algorithms are used to explore the space of mathematical formulae, looking for simple expressions that account for invariants in data gathered from a physical system. Systems of this kind are capable of inferring physical laws from experimental data, either in the form of differential or algebraic equations. The conservation law of angular momentum in a double pendulum, for example, was re-discovered by a fully automated apparatus searching the space of all possible mathematical formulae.

But Point 2 is much more important. We tend to think that the output of a scientific investigation such as Newton's or Einstein's should be a set of equations, and their interpretation, that can be used to work out predictions or models, for specific outcomes and specific experiments. We focus a lot on analytic manipulations of these general equations, as an example of abstract knowledge manipulations.

But this is not strictly necessary to science. The output of the scientific process does not need to be a set of equations – although this is what we have come to expect from Physics. All we ask of a model is to make the right prediction in the right situation. There can be both physical and formal models of physical systems. Different mathematical tools can be used to model the same system.

Calculus is not less arbitrary a representation than others: logical statements or statistical patterns may be used just as well to model some aspects of reality. Calculus simply provides modellers with a language and a technology for computation that was unsurpassed for centuries, and hence was the most natural choice to describe Physics. In fact, the two co-evolved and co-adapted. If one can produce a formal framework that can simulate some aspect of reality, this is sufficient to make it a modelling language.

The detection of subtle, elusive but predictive patterns in vast masses of data may be as useful as the creation of a simple mathematical model to explain them. Typically, the model is used to make predictions anyway, the same predictions that data-patterns can make. What if we had a computer that can make the same predictions without needing to start from a set of high level equations, but instead starting from a set of relations discovered in data? Just as these equations derive their meaning from their use, one could argue that predictive patterns discovered in data could play a similar role.

Besides, it is quite possible for machines to summarise these patterns in compact theories, only to deduce them back when needed from the basic axioms. This is what humans do. But would that be useful for machines?

When was it in history that we started considering ‘explained’ a phenomenon when we had – for example - a few equations describing its dynamics? It surely must have started in mechanics, perhaps with Newton, maybe with Galileo himself. But these equations are ultimately combined together, and with observation of initial conditions, in order to derive predictions. What if we could just derive the very same predictions from initial conditions and knowledge that is represented in a different way, perhaps even as raw data?

Large part of all the scientific knowledge produced by humanity, is not in anyone’s mind, but in some – possibly still unlinked – databases, and will only ever be accessed by machines [Berners-Lee, 2001]. As long as the consumers of this knowledge are other machines, human-readability is not a crucial issue. If the information is used – for example – to design new experiments, or even to design drugs, humans may even be completely out of the knowledge creation / exploitation loop.

### **Designing Experiments**

Recently the function of some yeast genes has been pinpointed by a robotic apparatus, generating hypotheses based on previous observations. This was part of an effort to develop a system capable of performing the full hypothetical-deductive cycle: the design of automatic-scientist systems, which can design and perform experiments based on the outcomes of previous experiments [Bryant, 2004], [King, 2009]. For systems like these, there would really be no reason for the knowledge they produce to be understandable by humans, as it is used only by them, to perform increasingly discriminating experiments.

The system discussed in [King, 2009] “has autonomously generated functional genomics hypotheses about the yeast *Saccharomyces cerevisiae* and experimentally tested these hypotheses by using laboratory automation”. In particular, it was applied “to the identification of genes encoding orphan enzymes in *Saccharomyces cerevisiae*”:

enzymes catalyzing biochemical reactions thought to occur in yeast, but for which the encoding gene(s) are not known”.

The robot “formulated and tested 20 hypotheses concerning genes encoding 13 orphan enzymes. The weight of the experimental evidence for the hypotheses varied (based on observations of differential growth), but 12 hypotheses with no previous evidence were confirmed with  $P < 0.05$  for the null hypothesis.” These hypotheses were later confirmed also by human scientists.

While in this particular system all the knowledge produced is represented as logical statements, it is important to notice that in order for a system to design new experiments and perform hypothesis testing, simple machine-readable knowledge would be sufficient.

### **Pattern Analysis**

Of course we are not announcing “The End of Theory” (although these claims have been made recently, [Anderson, 2008]) but just that we are facing an alternative – and equally valid – scientific method. This will also help us understand better the status of theoretical knowledge produced by science. Patterns extracted from data can reliably be used to make predictions – just ask Amazon.com or Google.com – without the need to formulate the knowledge contained in them in the form of differential equations, or other theoretical constructions, including grand unified theories.

But what matters is that at the centre of this paradigm shift is our capability to gather, store, manage and analyse massive amounts of data automatically. And this is the permanent marriage between statistics and computer science – and many other sub-disciplines – that is represented by computational pattern analysis.

Software tools for data mining, just like Galileo’s telescope, were perhaps not originally created for doing science, but very often for doing business. But just like Galileo, we can turn them and use them to change the way we understand our world.

And the fact that we are using off-the-shelf hardware to produce data, and to manage and store it, and we are using commercial software to analyse it, can only signal that further accelerations are to be expected, as costs are driven down.

### **Social Sciences: Media Content Analysis.**

My research group at the University of Bristol makes extensive use of pattern analysis technologies, which were originated for practical or industrial applications, to answer purely scientific questions. Much like Galileo directed the telescope to the Moon, we are aiming these new tools to another type of “sphere”. The analysis of contents in the Global Media-sphere is becoming accessible to computers, and this means that it can now be done in vast scale and in real time.

The Global Media System (or Mediasphere) is the interconnected system of all newspapers, magazines, broadcast-news outlets, blogs, news-wires, and so on. Every outlet can pick and choose whichever news it wants to carry; each user can choose whichever outlet they want to read; complex dynamics regulates the resulting process of information selection and diffusion; but simple patterns emerge, if we look in the right place. We are interested in observing (and modelling) how “ideas” flow and interact, as they traverse the media system (in the setting of blogs, see [Huberman, 2004] for an example).

In order for machines to access and use the contents of the Global News Media System, it is necessary that they understand (to some extent) human language. And this is a totally new ingredient that we can add to the mix, today: machines can actually read and “understand” certain aspects of text. Our apparatus is machine-translating every day from 22 languages, and reading 1,100 news outlets, obtaining about 20K news items per day. In the resulting vast, machine processed, dataset we have found 450K named entities, for example, exhibiting a perfect power law of popularity, and interesting relations such as a 3-fold extra interest in the Pope found in Spanish-language media over English-language media, over the same period, in the United States.

We are detecting text re-use, with massive scale implementations of suffix trees, and tracking memes as they spread through the outlets forming the global media sphere. We are recreating social networks, and detecting biases in the choice of topics and words in various types of outlets. We even measure readability.

Social scientists have been interested in understanding the media system for decades, but their investigations could only be performed by hand, on limited numbers of outlets, time spans, and topics. A true constant monitoring of all outlets and all topics in all languages is now within reach, and automatic analysis tools are becoming available.

Similar ideas can apply to the Humanities, with the possibility to analyse millions of books, in an automated – but still partly semantic – way. What is sometimes called “Cyberscholarship” will do for the social sciences and humanities, what has already been the computer revolution in the Life Sciences. Patterns found in text and images can be then used to design more experiments, or to analyse the behaviour of readers, and so on. Also psychology stands to benefit from these advances. There is much more to “data” than numbers, and a data-driven approach to science can cover unexpected fields of knowledge.

### **Publishing Data.**

Making data available in a linked form, a version of the Semantic Web [Berners-Lee, 2001] could one day take the place of publishing a discovery. The data could be made available by a machine, and used by another machine. The notion of scholarly publication, in the form of peer-review report of some experimental findings, is a few centuries old, and is by no means the only possible form of publication of results.

GenBank is a database that contains publicly available nucleotide sequences for more than 300,000 organisms. It has grown exponentially since the early 1980s, and continues to do so with a current doubling time of about 30 months. Currently GenBank contains over 95 billion nucleotide bases from more than 92 million individual sequences, with 16 million new sequences added in the past year. [Benson, 2009]

The examples of Pubmed and Genbank will be followed by other sciences, in the future, with a tight integration of results, data and methods, sharing and globally creating a single unified resource.

### **Conclusions: The Future of Method.**

The scientific method is today evolving faster than ever. The automation, systematisation and industrialisation of information gathering and analysis, are accelerating the rate at which we expand our knowledge of the world. Machines now produce knowledge about our very own biology. The proportions of this transition should not be underestimated, and the science of patterns, information and knowledge is at the centre of this storm. While advances in most other disciplines change the overall body of knowledge we have about the world, advances in Pattern Analysis and Data Mining change the very way in which we acquire that knowledge.

Galileo Galilei could have kept on making hi-tech tools and gadgets, and would certainly have found enough customers to make a comfortable living. But he was a scientist, and so he used those tools to understand the world around him. In the process he used mathematical representations of the laws that he discovered, designed experiments to gather data, and overall deployed the modern methodology. He also got into trouble with the authorities, because he refused to keep his telescope aimed low enough, and refused to ignore what he saw with it.

A new generation of scientists, with a new generation of tools, can now do the same, and gather unprecedented types of data, and draw far reaching conclusions about our world. The automatic collection of data in genomics, chemistry, astronomy, physics and also the social sciences, will revolutionise the way we see our world, and will further an understanding of it as a single interconnected system. Automated Data Analysis is at the centre of a very important revolution in the very way we produce new scientific knowledge.

### **References**

---

- Galileo Galilei, *The Starry Messenger*, 1610 (Padova)
  - [Adelman-McCarthy, 2008] Jennifer Adelman-McCarthy et al.; *The Sixth Data Release of the Sloan Digital Sky Survey*, 2008, *ApJS*, 175, 297-313  
2008ApJS..175..297A; <http://www.sdss.org/>
  - Kevin Kelly, *The Evolution of Scientific Method*; *The Technium*, 2004;  
[http://www.kk.org/thetechnium/archives/2004/12/evolution\\_of\\_th.php](http://www.kk.org/thetechnium/archives/2004/12/evolution_of_th.php)
  - [1000genomes.org, 2007] Meeting Report: A Workshop to Plan a Deep Catalog of Human Genetic Variation; September 17-18, 2007; Cambridge, UK  
[http://www.1000genomes.org/bcms/1000\\_genomes/Documents/1000Genomes-MeetingReport.pdf](http://www.1000genomes.org/bcms/1000_genomes/Documents/1000Genomes-MeetingReport.pdf)
  - [Duellmann, 2007] Dirk Duellmann; *Databases for the Large Hadron Collider at CERN*; *XLDB Workshop @ SLAC*; 25. October 2007; [http://www-conf.slac.stanford.edu/xldb07/xldb\\_lhc.pdf](http://www-conf.slac.stanford.edu/xldb07/xldb_lhc.pdf)
  - [DeWitt, 1995] Sheila H DeWitt and Anthony W Czarnik; *Automated synthesis and combinatorial chemistry*; *Current Opinion in Biotechnology*; Volume 6, Issue 6, 1995, Pages 640-645
  - [Berners-Lee, 2001] Tim Berners-Lee and James Hendler; *Publishing on the semantic web*; *NATURE*|VOL 410 | 26 APRIL 2001 | [www.nature.com](http://www.nature.com)
  - [Schmidt, 2009] Michael Schmidt and Hod Lipson - *Distilling Free-Form Natural Laws from Experimental Data* - *Science* 3 April 2009
-

- [King, 2009] Ross D. King, Jem Rowland, Stephen G. Oliver,<sup>2</sup> Michael Young,<sup>3</sup> Wayne Aubrey,<sup>1</sup> Emma Byrne,<sup>1</sup> Maria Liakata,<sup>1</sup> Magdalena Markham,<sup>1</sup> Pinar Pir,<sup>2</sup> Larisa N. Soldatova,<sup>1</sup> Andrew Sparkes,<sup>1</sup> Kenneth E. Whelan,<sup>1</sup> Amanda Clare<sup>1</sup> - The Automation of Science - Science 3 April 2009:
  - [Anderson, 2008] Anderson, Chris. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired* 16.07 (2008)
  - [Bryant, 2004] Christopher Bryant, Ffion Jones, Douglas Kell, Ross King, Stephen Muggleton, Stephen Oliver, Philip Reiser, Kenneth Whelan; *Functional genomic hypothesis generation and experimentation by a robot scientist*; Nature, Volume 427, January, 2004
  - Cory Doctorow; *Big data: Welcome to the petacentre*; Nature 455, 16-21 (2008)
  - Arms, William and Ronald Larsen (editors). "The Future of Scholarly Communication: Building the Infrastructure for Cyberscholarship". NSF/ JISC workshop, Phoenix, Arizona, April 17 to 19, 2007.
  - Crane, Greg. "What Do You Do with a Million Books?" *D-Lib Magazine* 12.3 (2006).
  - Friedlander, Amy (editor). "Promoting Digital Scholarship: Formulating Research Challenges in the Humanities, Social Sciences and Computation." *Council on Library and Information Resources* (2008).
  - Venter, Craig. "Bigger Faster Better." *Seed*, November 20 (2008).
  - [Huberman, 2004] B.A. Huberman and L.A. Adamic, Information Dynamics in the Networked World, Lect. Notes Phys. 650, 371–398 (2004)
  - [Benson, 2009] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW; GenBank, Nucleic Acids Res. (Jan 2009).
-

# Mining Online Communities

Eduarda Mendes Rodrigues

*Microsoft Research*

*JJ Thomson Avenue, Cambridge CB3 0FB, UK*

eduardamr @ acm.org

## Abstract

*Online communities, such as newsgroups, forums, Q&A services, among others, generate huge amounts of content every day. Such social media often contains information, advice and opinions that are valuable not only for community members but also for Web users in general, who may be searching online for problem-specific information. While some community users are committed to producing quality content, others primarily seek social engagement. Thus, it is important to understand the nature of the users' interactions and the value of individual contributions. In this paper, we discuss how social network analysis can be used to enhance the automatic classification of newsgroup messages and to characterize the nature of social interactions in Q&A communities.*

## 1. Introduction

Newsgroup communities have been around since the early days of the Internet. They are formed around a variety of topics and participants interact with each other through threaded conversations, for sharing information, opinions, providing support, advice, etc. Community question answering services (cQA), such as Yahoo! Answers and Live QnA, have become quite popular in recent years. Their aim is to provide support for users with specific information needs to obtain prompt responses to their questions from other users of the community. Similarly to newsgroups, questions are often requests for advice or opinion, which are unlikely to be obtained through standard Web search. Even though the answers can be submitted by users of all levels of expertise, the quality of answers can compare, or even surpass, the quality of answers given by expert networks and library reference services [11].

The social media content generated by online communities results in a rich knowledge base and valuable resource for other Web users to search and explore. Besides the standard Web users who might come across such content via a search engine, it is known that a

large percentage of users (often over 90% [17]) are *lurkers* who read available content but rarely communicate with others [17, 19]. Thus, providing effective support for search and browsing through community-generated content is of great value to the users. In particular, for finding information it is helpful to understand the structure of discussion threads and quickly 'zoom' onto the 'answer' messages. For those joining in a long discussion it is useful to get a sense of the dynamics and agreement level among participants.

Furthermore, it is also useful to differentiate between threads containing factual information from those where users primarily seek to communicate and connect with each other. Specifically, cQA services while designed primarily to facilitate answering questions, they are based on the premise that their communities are formed, active, and self sustainable. Inevitably, the quality of the cQA services depends on the level of expertise of the community members, the level of responsiveness to questions, and the nature of the users' interactions. Thus, it is important to gain a good understanding of the community dynamics and content contributions in order to provide the right incentives for creating desirable content.

This paper discusses the use of social network analysis for enhancing the automatic classification of newsgroup messages (Section 3) and for characterizing the nature of social interactions in Q&A communities (Section 4). Section 2 presents related work and Section 5 a summary of main findings.

## 2. Related Work

### 2.1 Newsgroup Communities

Discussion groups, blogs, online product reviews, and other community-generated content are rich sources of users' sentiment and opinion and have been a subject of a considerable body of research on opinion polarity and sentiment analysis. Techniques that have been used include text classification methods [3, 18], linguistic



analysis [9, 15], and social network analysis [3, 20]. The properties of the reply-to social network have been used to identify topic polarity of newsgroup participants [3, 13] and to characterize newsgroup types and author roles [7].

Based on the hypothesis that a message response is most likely to disagree with the parent message, Agrawal et al. [3] applied constrained and unconstrained graph partitioning techniques to cluster authors who share similar opinions into two opposing camps. Kelly et al. [13] clustered participants with similar opinions within a newsgroup and found that, regardless of the underlying distribution of participants into the clusters, the ratio of messages on each side of the discussion is balanced. Indeed, the traffic of the minority opinion was found to be larger in order to make up for the smaller number of people.

## 2.2 Community Q&A Services

Community Q&A services have grown in popularity over the last couple of years, greatly due to the success of Yahoo! Answers. The research community has also gained interest in investigating various aspects of this service, leading to a number of studies reported over the past year [1, 2, 11, 12]. There has been a great emphasis on identifying and predicting quality answers [1, 2, 11], and modelling users authority [10, 12].

Adamic et al. [1] analysed the Yahoo! Answers social network, identifying users with similar behaviour to the ‘answer-person’ role found in newsgroup communities [7]. Agichtein et al. [2] provided a classification model for estimating answer quality based on features derived from the content and also authority measures from the social network. Gyongyi et al. [10] and Jurczyk et al. [12] applied variants of Kleinberg’s HITS algorithm to the Q&A social network graph to model user reputation and level of expertise.

## 3. Classification of Newsgroup Messages

We performed a set of experiments with the aim to classify messages posted to two types of newsgroups, political discussion groups and Q&A groups, and to investigate the impact of particular features on the classifiers performance. We applied linear Support Vector Machine (SVM) [5] classifiers to:

- 1) Predict the agreement level between a message and its parent message within discussion newsgroup threads. Messages were classified as ‘agree’, ‘disagree’, or ‘insult’.
- 2) Identify which messages are questions or answers within technical Q&A newsgroup threads. Messages were classified as ‘question’ or ‘answer’.

We represented each message-parent pair by a vector of features and we used a *one-vs-all* multi-class approach

for classifying message pairs. In this section, we describe the dataset, feature sets and present a summary of the classification experiments reported in [8].

### 3.1 Dataset

Our dataset consists of message thread and header information from 4 Usenet newsgroups. The first two newsgroups, *alt.politics.immigration* and *talk.politics.guns*, host mostly political discussions and debates. The other 2 groups, *microsoft.public.internetexplorer.general* and *microsoft.public.windowsxp.general*, host mostly Q&A-type threads. Table 1 contains information about these data sets, hereafter referred to as *immigration*, *guns*, *explorer* and *winxp*. It lists the total number of threads, messages, replies and authors per newsgroup. It also indicates the period of time in which all messages were collected.

**Table 1. Description of the newsgroup data sets.**

Newsgroup	Threads	Messages	Replies	Authors	Collection Period
<i>immigration</i>	1,367	10,095	8,728	463	Aug 31 to
<i>Guns</i>	874	6,776	5,902	844	Oct 19’06
<i>explorer</i>	3,631	10,934	7,303	3,443	Jul 19 to
<i>winxp</i>	10,280	42,052	31,772	8,145	Oct 19’06

For the classification experiments we created training data sets from several samples of threads randomly selected from each newsgroup. The sample messages were annotated by experts with one of the labels listed in Table 2.

**Table 2. Message labels.**

Label	Description
<i>agree</i>	Message agrees with the point of view of the parent message. Adding clarifications or extra info also counts.
<i>disagree</i>	Message disagrees with the point of view of the parent message. Sarcastic comments also count.
<i>insult</i>	Author of the message is purely insulting the author parent message. Insults replying to insults are <i>disagree</i> messages.
<i>question</i>	Message is a question or a clarification of a previously asked question by the same author.
<i>answer</i>	Message is an answer to a question in the parent message or a request for further information about the question.
<i>off-topic</i>	The message has no connection to the parent message and is not a question message.
<i>don't know</i>	If none of the above labels apply.

### 3.2 Feature Sets

We considered a variety of features, both of structural and content nature, to investigate the impact of particular features on the classifiers performance. For content

analysis, we cleaned each message to remove headers and any quoted text from parent messages. Additionally, we derived features from 5 implicit network structures: 3 author networks and 2 thread networks. Past research has used thread-level message features for analysis of newsgroup data [4, 6, 7, 21]. We also ran experiments with such kind of features, but our results did not show much improvement with these features. Thus, here we concentrate on the multi-network features.

#### Author Networks

We captured users’ participation by defining 3 author networks for each newsgroup: *reply-to*, *thread participation*, and *text similarity*. In all of these, the nodes represent authors, but the edges carry distinct semantics:

- A *reply-to network* edge from author A to author B indicates that A has replied to at least one message posted by B.
- A *thread participation network* edge from author A to author B indicates that both authors have participated in the same thread in at least  $k$  occasions.
- A *text similarity network* edge indicates similarity between the content of connected authors’ messages. The messages from each author were summarized by a centroid keyword vector and *author-author* edges were created to indicate cosine similarity of at least  $\eta$ .

We described each *message reply* by a vector of features extracted from the three author networks, **A1**, **A2** and **A3**. Given a message, **M1**, and the message it replies to, **M0**, 3 feature vectors were created for M1 and another 3 for M0 (see Figure 1). Individual features of each vector are associated with nodes in the networks, i.e., authors. A similar author node vector was created for the author of the parent message M0. The final feature set for a *reply message* concatenated the two vectors.

Feature Set	M1 (Author A)	M0 (Author B)	Network
A1	$[a_{11} \ a_{12} \ \dots \ a_{1N}]$	$[b_{11} \ b_{12} \ \dots \ b_{1N}]$	<i>Reply-to</i>
A2	$[a_{21} \ a_{22} \ \dots \ a_{2N}]$	$[b_{21} \ b_{22} \ \dots \ b_{2N}]$	<i>Thread participation</i>
A3	$[a_{31} \ a_{32} \ \dots \ a_{3N}]$	$[b_{31} \ b_{32} \ \dots \ b_{3N}]$	<i>Text similarity</i>

Figure 1. Feature sets from the author networks.

Feature Set	M1 (Thread T)	Network
B1	$[t_{11} \ t_{12} \ \dots \ t_{1M}]$	<i>Common authors</i>
B2	$[t_{21} \ t_{22} \ \dots \ t_{2M}]$	<i>Text similarity</i>

Figure 2. Feature sets from the thread networks.

#### Thread Networks

We captured topic associations by defining 2 types of thread networks for each newsgroup: *common authors network* and *text similarity network*.

The nodes of both networks represent threads but the edges have a different meaning in each case:

- A *common authors network* edge between thread T and Q indicates “thread T has at least  $m$  authors in common with thread Q”.
- A *text similarity network* edge between thread T and Q indicates similarity between the content of their messages. The cosine similarity between centroid keyword vectors was used and an edge between thread T and Q indicates similarity of at least  $\eta$ .

We described each *thread* by a vector of features extracted from the two thread networks, referred to as **B1** and **B2**, respectively. Given a message **M1** belonging to the thread **T**, we created two feature vectors, where individual components were associated with other nodes the networks, i.e. threads – see Figure 2.

### 3.3 Experiments

We conducted a comprehensive set of experiments with the SVM classifiers to investigate the effectiveness of individual feature sets and their combinations in:

- 1) Predicting the level of agreement of messages posted to political discussion newsgroups.
- 2) Identifying *question* and *answer* messages in technical discussion newsgroups.

Given that *reply-to network* features have been used in related work [3, 7, 13], we took the feature vector A1 as the baseline for our analysis of the classification results. For evaluation we used 10-fold cross-validation and the performance of the classifier was measured based on the *break-even-point* (BEP) from the ranked list of messages scored by the classifier. Next, we summarize the main findings.

#### Discussion Newsgroups

To predict the level of agreement between a message and its parent message in discussion threads, we used the relevant training data, i.e. messages labelled as ‘agree’, ‘disagree’ or ‘insult’. We observed increased performance over the baseline when thread network features were introduced. This improvement was particularly evident in the ‘insult’ class, where such messages seem to be strongly predicted through the co-participation in threads (B1): BEP increase from 68% to 74% for *guns* and from 38% to 85% for *immigration*. Using threads text similarity features (B2) gave further boost to the *guns* category: from 74% to 81%.

**Table 3. Classification results for discussion groups**

Feature Sets	<i>Guns</i>			<i>immigration</i>		
	<i>agree</i>	<i>disagree</i>	<i>insult</i>	<i>agree</i>	<i>disagree</i>	<i>insult</i>
F1=A1	61%	80%	62%	65%	75%	37%
F2=A1+A2	69%	82%	72%	66%	76%	45%
F3=F2 + A3	65%	84%	68%	68%	77%	38%
F4=F3 + B1	67%	86%	74%	73%	80%	85%
F5=F4 + B3	66%	85%	81%	72%	80%	85%

**Table 4. Classification results for Q&A groups**

Feature Sets	<i>iexplorer</i>		<i>winxp</i>	
	<i>answer</i>	<i>question</i>	<i>answer</i>	<i>question</i>
F1=A1	70%	59%	93%	78%
F2=A1+A2	71%	64%	94%	80%
F3=F2 + A3	75%	66%	94%	79%
F4=F3 + B1	75%	66%	94%	79%
F5=F4 + B3	75%	65%	94%	77%

#### Technical Q&A Newsgroups

To identify *questions* and *answers* in technical Q&A newsgroups, we used the relevant training data, i.e. messages labelled as ‘question’ or ‘answer’. Unlike the previous case, features derived from the thread networks did not improve the classifier’s performance. Connections among authors that participated in the same threads (A2) were particularly beneficial to predict ‘questions’: BEP increase of 59% to 64% for *iexplorer* and 78% to 80% for *winxp*. Content-based author similarity features (A3) improved the prediction of ‘answers’ for *iexplorer*: from 71% to 75%.

In summary, we found that the co-participation of authors across threads (feature set B1) was a particularly relevant factor for improving the classification of messages in discussion threads. Text similarity features further improved classification. These results hint that authors seem to be consistent in their opinions, when recurring co-participating with other authors across discussion threads. However, thread network features did not enhance the classification performance in the Q&A case. These results are consistent with the observations by Fisher et al. [7] on the distinctive behaviour of core participants of discussion vs. technical newsgroups. The former tend to form fairly closed communities with the most active participants responding to each other often and mostly ignoring newcomers. The latter, on the contrary, tend to be experts who respond primarily to newcomers who ask questions.

## 4. Social Behaviour in cQA Services

The cQA services allow users to freely submit questions and answers on any topic, and provide several mechanisms for self-regulation of the content quality, such as, enabling comments on answers, voting for best answers, reporting abuse, and assigning reputation points

to community members. However, since users need to create a sense of community, it is not surprising that some users seek to communicate and connect with the community by asking questions, such as, ‘*How are you?*’ or ‘*I’m eating a slice of home-made pie. Anyone wants some?*’. This behaviour does not comply with the intended use of the service but aims to engage with and perhaps entertain the community.

We performed extensive analysis of the Live QnA and Yahoo! Answers communities. Although the two services are very similar, they differ on the approach taken to categorize questions. On Yahoo! Answers users assign a label to their questions, by picking a topic category from a fixed taxonomy, while on Live QnA users apply community-generated tags to their questions. In our analysis, we were particularly interested in revealing the implications of the Live QnA question tagging feature on the community dynamics and the observed question types [16]. In this section we present summary findings of this analysis.

### 4.1 Datasets

Our first dataset was obtained from the Live QnA service and spans the first year of its beta release (Sep. 2006 until Sep. 2007). It consists of 488,760 questions and 1,330,819 answers. The questions were submitted by 241,616 unique users, while the answers and comments were contributed by 42,941 and 34,068 unique users, respectively. The second dataset was gathered from the Yahoo! Answers service by seeding a crawler with pages linked to the top level categories that list recent questions with the assigned category and sub-categories. Overall we crawled 309,599 questions, posted by 217,615 distinct users and 1,151,453 answers, given by 202,052 distinct users. Over 95% of the content that we crawled was posted during the 3-month period of March-May 2008.

### 4.2 Analysis of Live QnA Tags Usage

In the Live QnA dataset questions were labelled with  $2(\pm 2.3)$  tags on average. Overall, the community applied 188,468 distinct tags. Some of these tags were used very frequently, possibly due to the automatic recommendation of tags that is provided by the service. Among the 10 most frequently used tags, the technology-related ones (‘Internet’, ‘Technology’, ‘Computers’, ‘Windows’, and ‘Microsoft’, ‘Windows Live’) were applied to questions receiving on average 2.4 answers, whereas the remaining ones (‘Fun’, ‘life’, ‘people’, and ‘Family’) were applied to questions receiving on average 5.1 answers. This indicates that the community members responded more actively to questions on less technical topics.

### 4.3 Analysis of Question Types

Through manual inspection of the random samples of questions from Live QnA and Yahoo! Answers we derived a taxonomy of question types, that includes types such as (a) *information seeking* – requesting information about a fact or a resource that can satisfy the user information need, (b) *opinion seeking* – requesting an opinion about a topic, possibly of personal nature, and (c) *chit-chat* – question instigating community reaction for the purpose of socializing. We observed that information seeking questions were predominant in both datasets: 62.3% on the Live QnA sample and 78.1% on the Yahoo! Answers sample. The percentages of opinion seeking questions were also comparable: 19.0% on Live QnA and 15.3% on Yahoo! Answers. However, we found a substantial difference in the proportion of chit-chat questions: 14.2% on Live QnA and 3.6% on Yahoo! Answers.

Considering the Live QnA questions tagged with one of the top 10 tags we analysed the frequency of question types across tags. Figure 3 shows that tags referring to technology and computer-related topics were predominantly associated with questions of the information-seeking type. In contrast, tags like ‘Fun’, ‘People’ and ‘life’ were mostly associated with chit-chat questions. The ‘Fun’ tag, in particular, is highly correlated with this type of question.

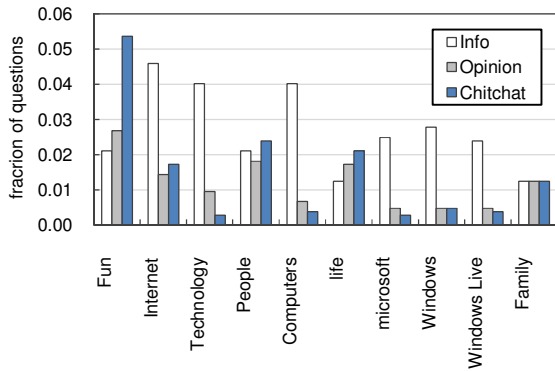


Figure 3. Distribution of labelled questions across the top 10 most frequent Live QnA tags.

### 4.4 Community Tags & Social Network Activity

With the new insights about the question types and community tags, we investigated the properties of the social network that emerges from answering questions with specific tags. We performed an analysis of the *answer-to* social network derived from the Live QnA data set. In such network the nodes correspond to active users and the directed edges indicate that, for example, a user A has responded to a question of the user B.

For each community tag we considered the associated sub-graph of the answer-to network and calculated the *density of the sub-graph* to assess the strength of the social ties among the involved users. The graph density measures how close a subset of vertices is to forming a clique (i.e., to include the maximal number of edges):

*Definition.* For a directed graph with  $|E|$  edges and  $|V|$  vertices, the graph density is defined as  $D = \frac{|E|}{|V| \cdot (|V| - 1)}$ .

We examined sub-graphs consisting of the 100 most active ‘answerers’ and 100 most active ‘questioners’ for each of the top ten Live QnA tags. In Table 5 we show for each tag sub-graph the overlap between the top questioners and top answerer ( $V_Q \cap V_A$ ) and the density of the sub-graphs associated with answerers ( $D_A$ ) and questioners ( $D_Q$ ).

Table 5. Density of the social network resulting from answer-to interactions between top answerers (DA) and questioners (DQ), on the specified tag.

Tags	Questions	$D_A$	$D_Q$	$V_Q \cap V_A$
Fun	41,259	0.588	0.613	52%
Internet	34,005	0.243	0.255	31%
People	26,583	0.450	0.459	42%
Technology	25,116	0.092	0.089	17%
Computers	24,633	0.092	0.088	21%
Life	21,739	0.365	0.357	38%
Windows	18,499	0.067	0.066	19%
Microsoft	18,343	0.066	0.069	16%
Windows Live	17,644	0.107	0.120	27%
Family	17,498	0.307	0.326	40%

We observe that the community users exhibit different behaviour across tags. For tags like ‘Fun’, ‘People’ and ‘Family’, a high percentage of users who post questions also engage very actively in giving answers to other users, indicating that there are sub-communities of active users formed around such tags.

Furthermore, the density of the tag-induced sub-graphs hints that specific type of questions may be predominant for a given tag. For example, the density values for the ‘Fun’ tag indicate that highly active users interact with a large proportion of other highly active users and thus support our hypothesis that ‘Fun’ tag is associated predominantly with chit-chat questions. We can contrast that with density values for tags like ‘Microsoft’ or ‘Windows’, which are significantly lower. The low overlap between top answerers and top questioners for these tags is more typical of information-seeking communities where expert users provide the most answers [15].

## 5. Concluding Remarks

In this paper we discussed the use of social network analysis to enhance the automatic classification of newsgroup messages and to characterize the social tagging behaviour in cQA services. We developed robust message classifiers to detect messages of selected response types, including agreement and disagreement in newsgroup discussion threads. We have found that with well selected author and thread network features we can achieve very good classification results for any topic being discussed. The results clearly demonstrate the superiority of the thread network features over the standard reply-to network alone. Our findings offer the foundation for the design of ranking functions for newsgroup search that take into account the types of messages, given a search goal, such as, finding answers to a question, finding a similar question, or finding strong positive and negative opinions about a topic.

Through the analysis of the Live Q&A community we found that community-generated tags reflect both the social interactions among users and the topic of the questions. In fact, we hypothesise that the freedom to contribute with new tags has led to the possibility of disseminating questions that are of social nature, and vice versa, that the variety of social interactions have influence the evolution of the community taxonomy.

## 6. Acknowledgment

This paper was prepared for the invited talk at the 2009 UK KDD Symposium and is based on previous joint work with Natasa Milic-Frayling and Blaz Fortuna [8,16].

## 7. References

- [1] Adamic, L. A., Zhang, J., Bakshy, E., and Ackerman, M. S., Knowledge sharing and yahoo answers: everyone knows something. In *Proc. of WWW '08*, pp. 665-674.
- [2] Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G., Finding high-quality content in social media. In *Proc. of WSDM '08*, pp.183-193, 2008.
- [3] Agrawal, R., Rajagopalan, S., Srikant, R., Xu, Y., Mining Newsgroups Using Networks Arising from Social Behavior, In *Proc. of WWW'03*, pp. 529-535, 2003.
- [4] Borgs, C., Chayes, J., Mahdian, M. and Saberi, A., Exploring the Community Structure of Newsgroups, In *Proc. of KDD'04*, 2004.
- [5] Cortes, C. and Vapnik, V. Support Vector Networks. *Machine Learning*, vol. 20, pp. 273-297, 1995.
- [6] Fiore, A., Teirnan, S.L., Smith, M., Observed Behavior and Perceived Value of Authors in Usenet Newsgroups: Bridging the Gap, In *Proc. of CHI'02*, pp. 323-330, 2002.
- [7] Fisher, D., Smith, M., Welsler, H., You Are Who You Talk To: Detecting Roles in Usenet Newsgroups, In *Proc. of the 39th HICSS*, 2006.
- [8] Fortuna, B., Mendes Rodrigues, E. and Milic-Frayling, N., Improving the Classification of Newsgroup Messages through Social Network Analysis, In *Proc. of CIKM'07*, 2007.
- [9] Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R., and Tomokiyo, T., Deriving Marketing Intelligence from Online Discussion, In *Proc. of KDD'05*, pp. 419-428, 2005.
- [10] Gyongyi, Z., Koutrika, G., Pedersen, J., Garcia-Molina, H., Questioning Yahoo! Answers. In *Proc. First Workshop on Question Answering on the Web*, held at WWW '08.
- [11] Harper, F. M., Raban, D., Rafaeili, S., and Konstan, J. A., Predictors of answer quality in online Q&A sites. In *Proc. of CHI '08*.
- [12] Jurczyk, P. and Agichtein, E., Discovering authorities in question answer communities by using link analysis. In *Proc. of CIKM '07*, pp. 919-922.
- [13] Kelly, J.W., Fisher, D., and Smith, M., Friends, Foes, and Fringe: Norms and Structure in Political Discussion Networks, *Proc. of the Int. Conf. on Digital Government Research '06*, pp. 412-417, 2006.
- [14] Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., and Riedl, J., Applying Collaborative Filtering to Usenet News, *Communications of the ACM*, vol. 40, no. 3, pp. 77-87, 1997.
- [15] B. Liu, M. Hu, , J. Cheng, Opinion Observer: Analyzing and Com-paring Opinions on the Web, In *Proc. of WWW'05*, pp.342-351, 2005.
- [16] Mendes Rodrigues, E., Milic-Frayling and N., Fortuna, B. Social Tagging Behaviour in Community-driven Question Answering, *Proc. of IEEE/WIC/ACM WI 2008*, Dec. 2008.
- [17] Nonnecke, B. and Preece, J., Lurker demographics: counting the silent, In *Proc. of CHI '00*, pp.73-80, 2000.
- [18] Pang, B., Lee, L. and Vaithyanathan, S., Thumbs up? Sentiment Classification using Machine Learning Techniques, In *Proc. of EMNLP'02*, pp. 79-86, 2002.
- [19] Soroka, V. and Rafaeili, S., Invisible participants: how cultural capital relates to lurking behavior, In *Proc. of WWW'06*, pp. 163-172, 2006.
- [20] Tuulos, V. and Tirri, H., Combining topic models and social networks for chat data mining, In *Proc. of the IEEE/WIC/ACM WI'04*, pp. 206-213, 2004.
- [21] W. Xi, J. Lind and E. Brill, Learning effective ranking functions for newsgroup search," In *Proc. of SIGIR'04*, pp 394-401, 2004.

**Tony Bagnall**  
**University of East Anglia**

## **Time Series Data Mining**

### **Abstract**

This paper highlights the unique challenges of time series data mining from a statistical, machine learning and data mining perspective with examples taken from three very different problem domains.

Time series data is increasingly prevalent and as the information age matures the problems and opportunities offered by vast inter-dependent data sets will be one of the defining features of future data mining research. In statistics, the traditional approach to time series has been to model the auto-correlation structure with the focus usually on forecasting. In machine learning, the usual methodology is to derive a set of features from time-dependent data, then most interest lies in clustering and classification. In data mining, the major concern is compression and similarity measures and the majority of research is concerned with query by content. In this paper an overview of all three approaches is given and their similarities and differences for the wide range of problems that arise in the field are highlighted.

Many different application areas can be treated as time series, and we show this with examples derived from electricity usage profiling, image processing and RNA analysis.

# Time Series Data Mining

A personal perspective on time series data mining

**Tony Bagnall**  
**School of Computing Sciences**  
**University of East Anglia**  
**Norwich**

UEA, Norwich

# Time Series

**Definition:** Time series is a stochastic process where the time index takes on a finite or countably infinite set of values.

the ordering of the variables does not have to be time dependent. What differentiates time series from a normal vector of observations is the presence of autocorrelation



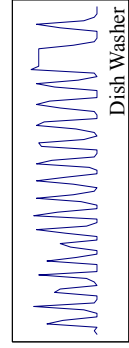
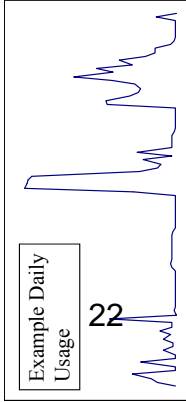
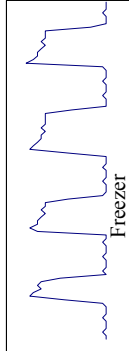
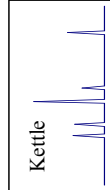
$$r_k = \frac{\sum_{i=1}^{n-k} (Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{(Y_i - \bar{Y})^2}$$

$$r_{10} = 0.82$$

UEA, Norwich

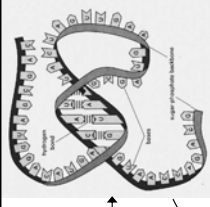
## Traditional Example

Electricity Usage



UEA, Norwich

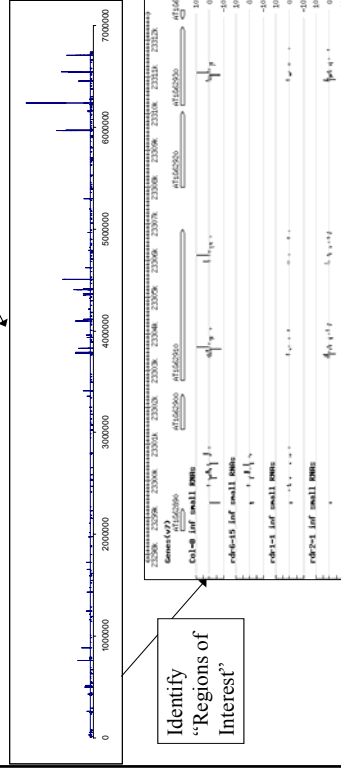
## short RNA (sRNA)



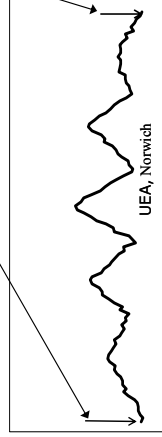
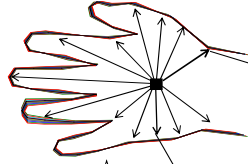
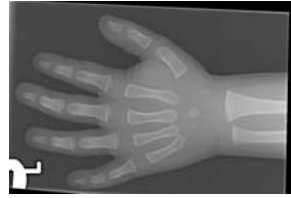
Query Set  
 ACGU...  
 GGGA...  
 CCUU...

Count occurrences along RNA

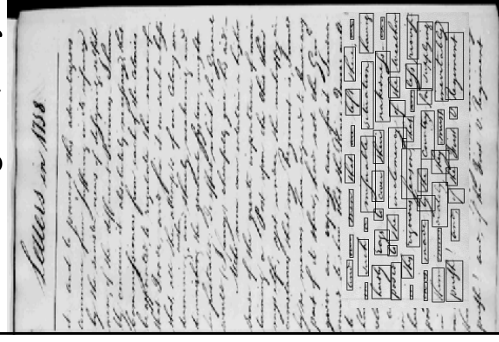
Form "pile up" time series of observations



## Hand Images

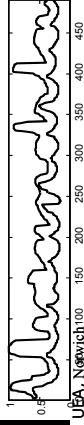


Handwriting data, may best be thought of as time series...



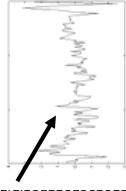
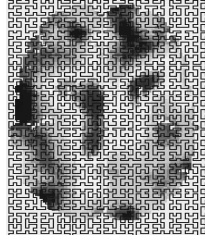
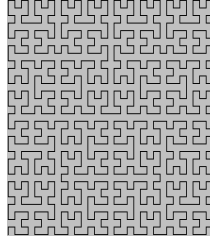
George Washington  
1732-1799

Alexandria



George Washington Manuscript

Brain scans (3D voxels), may be thought of as time series...



## Modelling Time Series

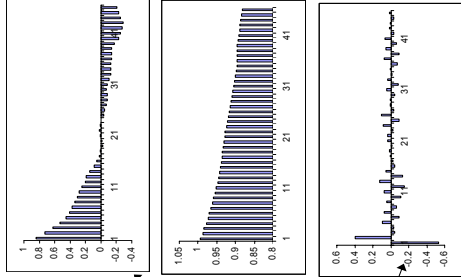
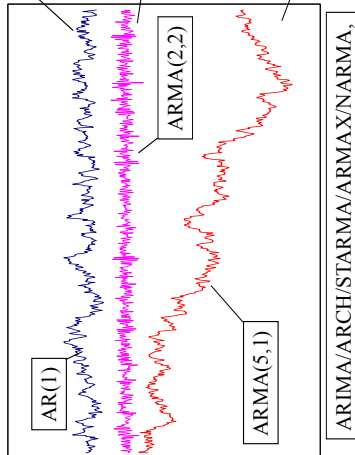
- Any exploratory analysis with time series involves two steps
  1. Modelling the series
  2. Using the model to perform some task
- The aim of modelling is to explain variation and reduce dimensionality/compress the data
- There are two basic approaches to modelling time series
  1. Model based on autocorrelation function
  2. Model based on fitted curves



# Auto-Correlation Based Models

Correlograms used to determine structure, algorithm used to estimate parameters

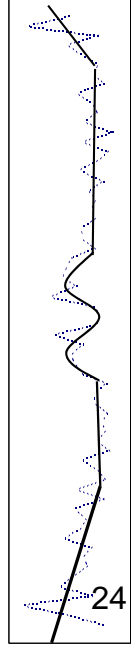
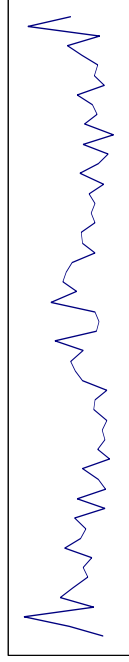
$$X_t = \sum_{i=1}^p \phi_i X_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t$$



ARIMA/ARCH/STARMA/ARMAX/NARMA,

# Shape based models

1. Local mode/decomposition  
fit piecewise functions



Splines/Wavelets/

UEA, Norwich

# Shape based models

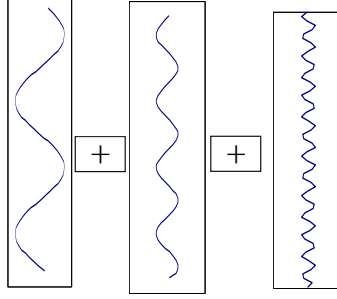
Functional form of the data assumed, some algorithm used to estimate parameters

## 1. Global model/decomposition

For example we could model data as a composition of sine waves through Fourier transform



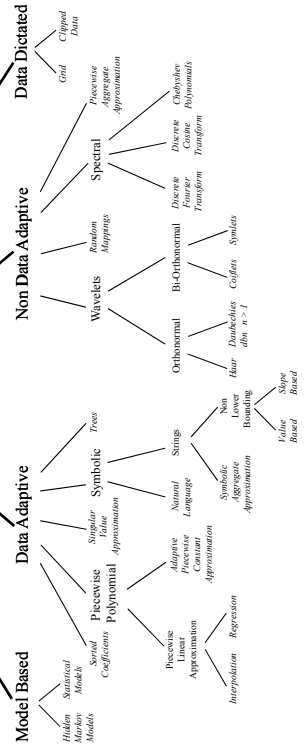
=



Note there is a direct relationship between the periodogram and the correlogram

UEA, Norwich

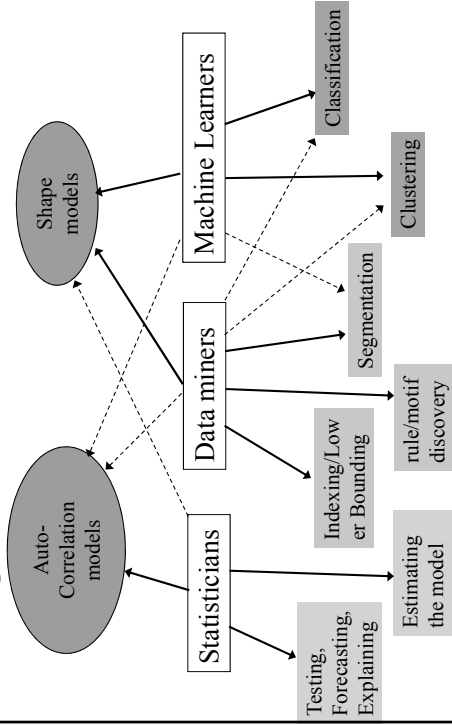
# Time Series Representations



UEA, Norwich

# Using Time Series

## Gross Generalisation:



UEA, Norwich

# Statistics Journals



2005-present  
179 papers



2005-present  
papers 251

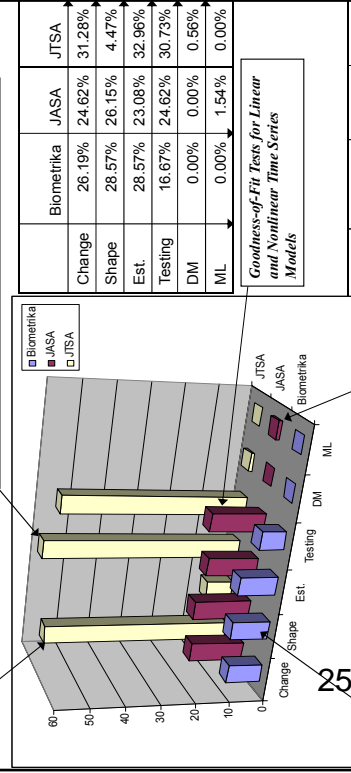


2005-present  
papers 518

UEA, Norwich

# Statistical Journals, 2005-present

Stability of nonlinear AR-GARCH models



Maximum likelihood estimation of higher-order integer-valued autoregressive processes

	Biometrika	JASA	J TSA
Change	26.19%	24.62%	31.28%
Shape	28.57%	26.15%	4.47%
Est.	28.57%	23.08%	32.96%
Testing	16.67%	24.62%	30.73%
DM	0.00%	0.00%	0.56%
ML	0.00%	1.54%	0.00%

Goodness-of-Fit Tests for Linear and Nonlinear Time Series Models

Spatially adaptive smoothing splines

Hidden Markov Models for Longitudinal Comparisons

	Biometrika	JASA	J TSA
Total Papers	251	518	179
Time Series	42	65	179
	16.73%	12.55%	100%

25

# Data Mining Journals/Conference



2005-present  
184 papers



2004-present  
papers 561



2005-present  
papers 648

UEA, Norwich

# Data Mining Papers

*Clustering time series from ARMA models with clipped data*

*A unifying framework for detecting outliers and change points from time series*

*Mining, Indexing, and Querying Historical Spatiotemporal Data*

*Characteristic-Based Clustering for Time Series Data*

Category	JMLR	ML
Change	5	0
Shape	35	43
Est.	0	6
Testing	5	6
DM	40	31
ML	15	12

*Experiencing SAX: a novel symbolic representation of time series*

Time Series	20	16	28
Total	184	648	561
%age	10.87%	2.47%	4.98%

UEA, Norwich

# Machine Learning Journals

2005-present  
177 papers

2005-present  
413 papers

Couldn't find any in PAMI

UEA, Norwich

# Machine Learning

*Evolutionary Rule Mining in Time Series Databases*

*Automatic Feature Extraction for Classifying Audio Data*

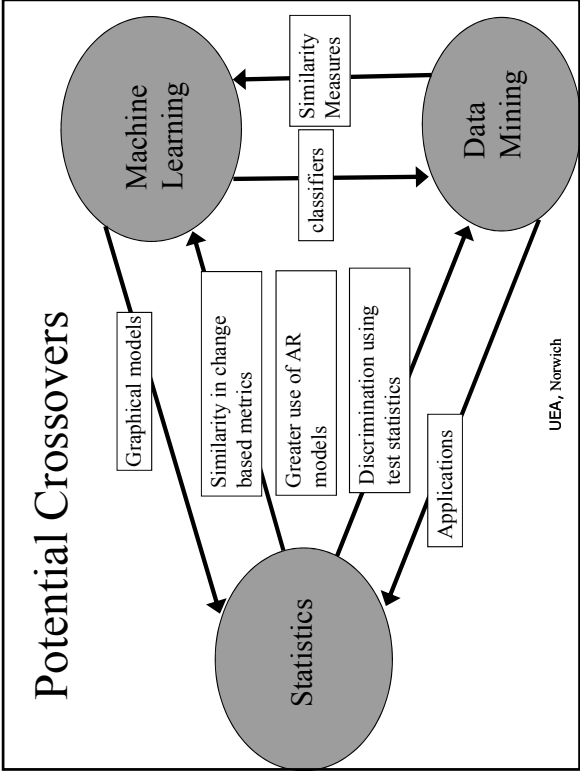
Category	JMLR	ML
Change	0	0
Shape	100	23
Est.	0	0
Testing	0	0
DM	0	17
ML	0	58

*Search for Additive Nonlinear Time Series Causal Models*

Time Series	7	17
Total	413	177
%age	1.69%	9.60%

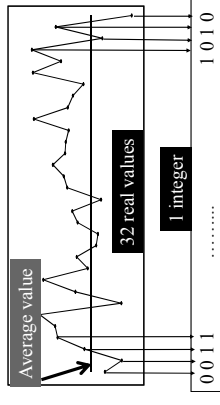
UEA, Norwich

# Potential Crossovers



## Using ARMA models in data mining

Clustering time series from ARMA models with clipped data (SIGKDD 2004)



**Transform and compress by clipping**

**Clipping is a very simple transformation**

- Reduces the memory requirements massively
- Can allow for faster clustering algorithms
- Can help detect outliers and model misspecification
- For long series, is as accurate as clustering with the whole data**

UEA, Norwich

Machine Learning

Using ARMA models in data mining

Statistics

Data Mining

UEA, Norwich

## ARMA: Auto-Regressive Moving Average

- Model ARMA(p,q)

$$X_t = \phi_1 X_{t-1} + \dots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}$$

$$\varepsilon_i \sim N(0, \sigma)$$

- Assume series in the same cluster are from the same ARMA model
- Fit an ARMA model to each series then cluster based on similarity of the fitted parameters

UEA, Norwich

## Clustering ARMA Data

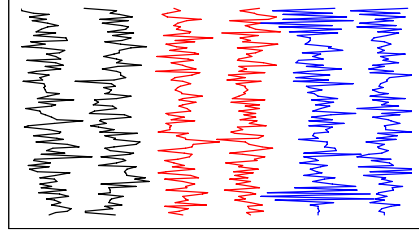
Find partial autocorrelations using Durbin-Levinson recursions

-0.655	0.000	0.000	0.000	0.000
-0.423	0.354	0.000	0.000	0.000
-0.462	0.400	0.109	0.000	0.000
-0.460	0.404	0.104	-0.011	0.000
-0.460	0.402	0.096	-0.002	0.020

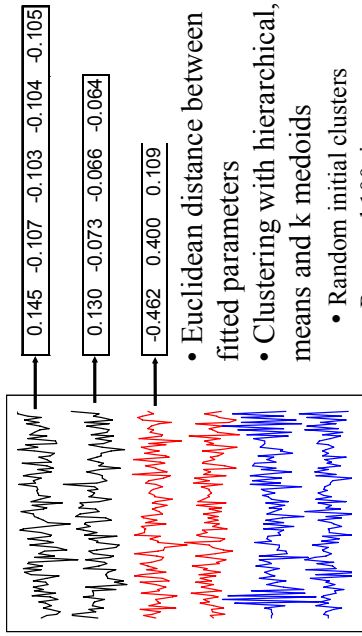
Choose model that minimizes AIC

$$x_t = -0.462x_{t-1} + 0.4x_{t-2} + 0.109x_{t-3}$$

UEA, Norwich



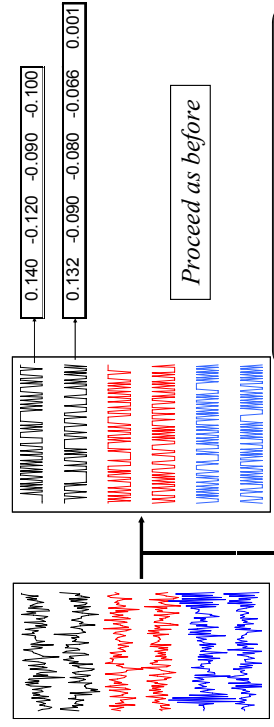
# Clustering ARMA Data



- Euclidean distance between fitted parameters
- Clustering with hierarchical, k means and k medoids
  - Random initial clusters
  - Restarted 100 times

UEA, Norwich

# Clustering Clipped ARMA Data

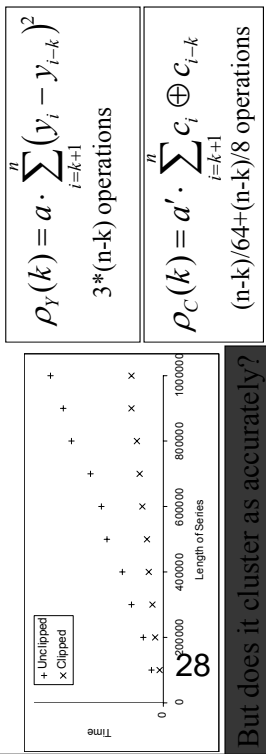


- Why do it?
- What effect does it have on performance?

UEA, N

# Benefits of Clipping

- Reduced memory requirements (64 reals to 1 integer)
- Potential for faster algorithms
  - Autocorrelation calculation made upto 5 times faster with bitwise operators



But does it cluster as accurately?

UEA, Norwich

# Properties of Clipped Series

- Kedem [1980,81] showed that the autocorrelation function of clipped data is related to that of the unclipped series

$$\rho_C(k) = \frac{2}{\pi} \sin^{-1} \rho_Y(k)$$

$\rho_C(k)$  Autocorrelation for Clipped  
 $\rho_Y(k)$  Autocorrelation for Unclipped

- He also demonstrated that the ML estimators from the clipped series tend toward the true values from the underlying generating series

# Properties of Clipped Series

- We extended these results by showing that if the original series can be written in autoregressive form

$$Y(t) + \phi_1 Y(t-1) + \phi_2 Y(t-2) + \dots + \phi_p Y(t-p) = \varepsilon(t)$$

- the clipped series has the form

$$C(t) + \phi_1 C(t-1) + \phi_2 C(t-2) + \dots + \phi_p C(t-p)$$

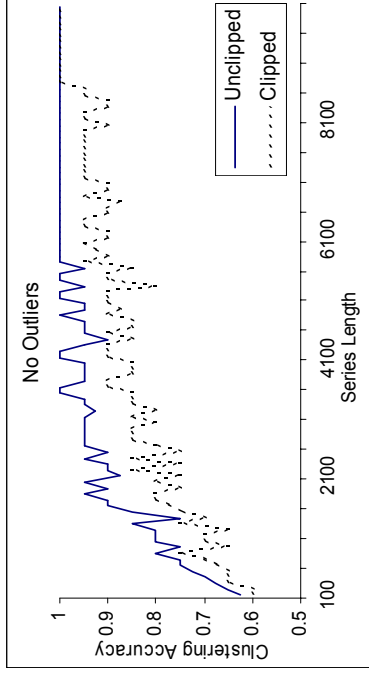
$$= \frac{\varepsilon(t)}{\sqrt{2\pi}} + e(t) - \phi_1 e(t-1) - \dots - \phi_p e(t-p)$$

$$e(t) = C(t) - \frac{1}{\sqrt{2\pi}} Y(t)$$

A linear ARMA model for the original series implies a linear ARMA model for the clipped series

UEA, Norwich

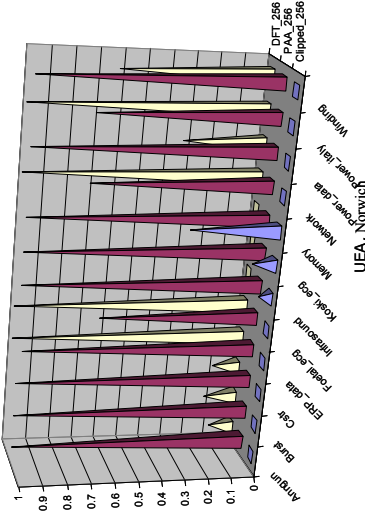
UEA, Norwich



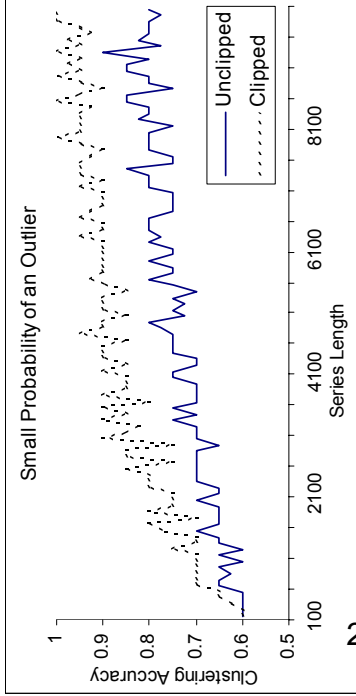
UEA, Norwich

# Clipping with query by content

For similarity in time problems from the UCR repository, queries with clipped data require significantly fewer disk accesses

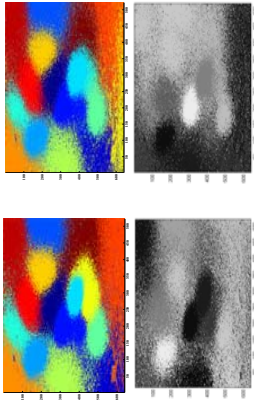


UEA, Norwich



# Clipping

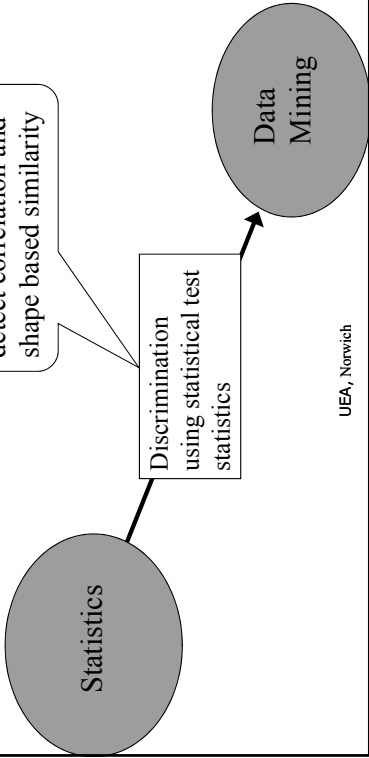
- Using clipped data does not decrease clustering accuracy if the series are long enough
- It can help in the detection of outliers and model misspecification



Clipping reduces memory requirements and increases clustering speed

UEA, Norwich

measures that can detect correlation and shape based similarity



UEA, Norwich

# Similarity in Time

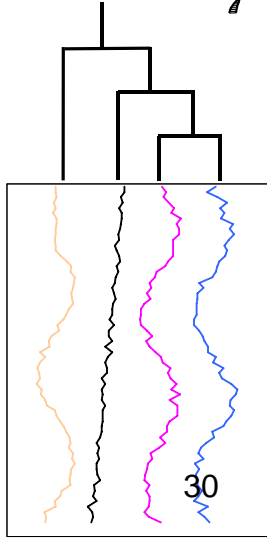
Measure of correlation between series. In TSDB most people use Euclidean distance

Given two time series:

$$Q = q_1 \dots q_n$$

$$C = c_1 \dots c_n$$

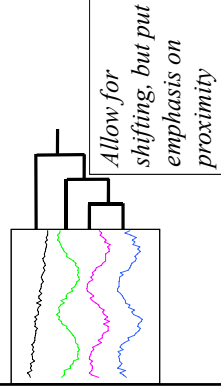
$$D(Q, C) \equiv \sqrt{\sum_{i=1}^n (q_i - c_i)^2}$$



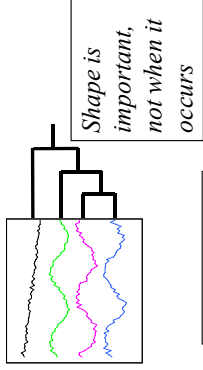
UEA, Norwich

# Similarity in Shape

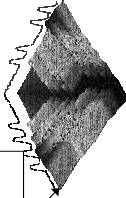
Weak Time Independence



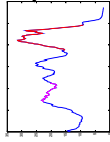
Strong Time Independence



Dynamic Time Warping



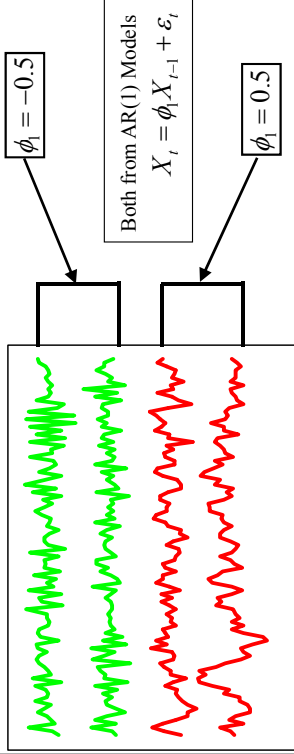
Shapelets and Motifs



UEA, Norwich

# Similarity in Change

Measure of similarity in auto-correlation structure



UEA, Norwich

# Fourier Transforms

The spectral density function is the Fourier transform of the autocovariance function

**Fourier Transforms are hence unique in TSDM as they can be used to measure all three types of similarity**

However, this is not true if you use the Euclidean distance between transformed variables

$$x_1, x_2, \dots, x_n \quad y_1, y_2, \dots, y_n$$

$$x_j = \langle (p_0, q_0), (p_1, q_1), \dots, (p_{n-1}, q_{n-1}) \rangle \quad y_j = \langle (r_0, s_0), (r_1, s_1), \dots, (r_{n-1}, s_{n-1}) \rangle$$

$$d_E(x_j, y_j) = \sum_{i=0}^{n-1} (p_i - r_i)^2 + (q_i - s_i)^2$$

Not phase independent

UEA, Norwich

# Likelihood Ratio Based Distance

The periodogram of a series is the terms

$$a_i = p_i^2 + q_i^2 \quad b_i = r_i^2 + s_i^2$$

Each term is an observation i.i.d. r.v. with exponential density

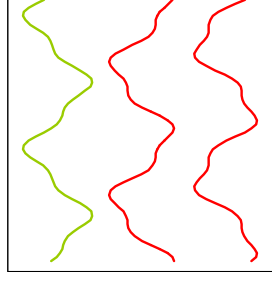
This means we can derive the likelihood ratio for the test that the periodograms are from the same distribution

$$\lambda = \prod_{i=1}^{n-1} \frac{2a_i}{a_i + b_i} \cdot \frac{2b_i}{a_i + b_i}$$

$$d(x_j, y_j) = 2 \sum_{i=1}^{n-1} 2 \log(a_i + b_i) - \log a_i - \log b_i$$

UEA, Norwich

# LR Distance is phase independent



Euclidean

	A	B	C
A	0%		
B	35%	0%	
C	40%	100%	0%

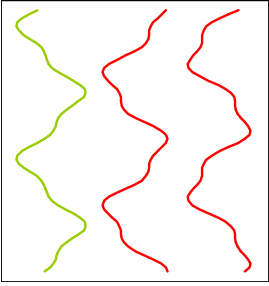
Likelihood Ratio

	A	B	C
A	0.0000%		
B	100.0000%	0.0000%	
C	99.9999%	99.9999%	0.0000%

UEA, Norwich



## LR Distance can be used for decisions



Cannot reject the null hypothesis that series are from the same process

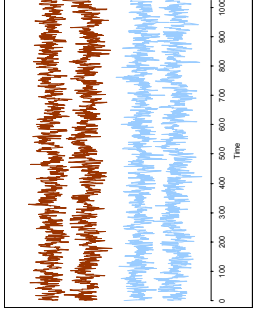


- Put in same cluster
- Use in classification scoring
- Ignore for query

UEA, Norwich

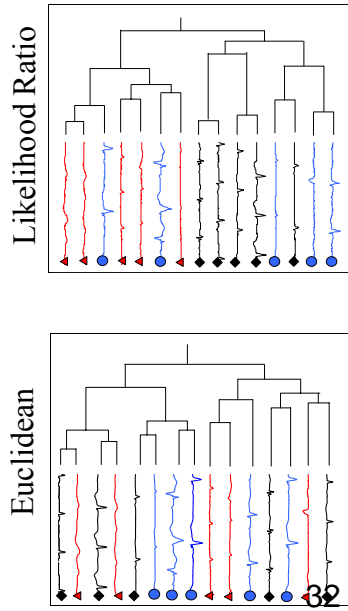
## LR Distance better than Euclidean for ...

- Random walk data
- AR(1) data
- AR(1) + sinusoidal data

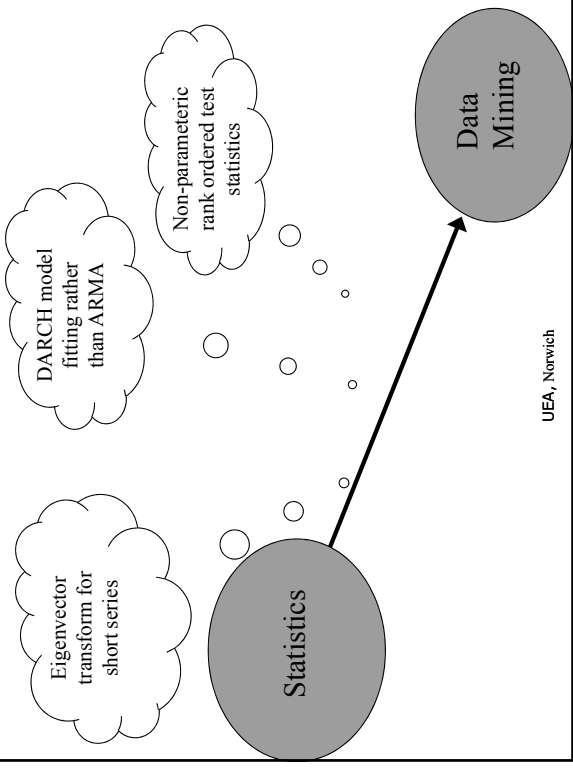


UEA, Norwich

## LR Distance better than Euclidean for ECG Data

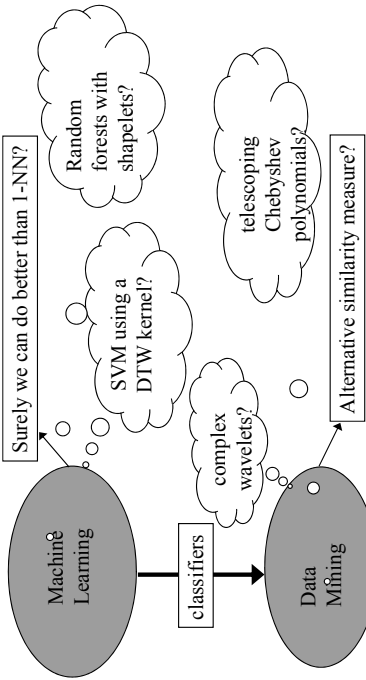


UEA, Norwich



UEA, Norwich

“While there has been work on classifying (shape-bases) time series with decision trees, neural networks, bayesian classifiers etc. None of these approaches is competitive with 1-nearest neighbor with DTW” (2004).



UEA, Norwich

## Conclusions

As the information age matures, grab set data will be time dependent

Even data not traditionally thought of as time series has autocorrelation structure

There is meta data to be mined in those stats journals!

## Questions?

UEA, Norwich

**Gavin Brown**  
**University of Manchester**

**Feature Selection by Filters: A Unifying Perspective**

**Abstract**

Feature Selection is an essential component of modern data mining.

The principle is to eliminate irrelevant or redundant variables from a dataset, given the requirement to predict a target. This has the dual advantage of reducing computation time, and increasing interpretability.

Datasets with thousands to millions of variables require fast methods for selection---these are known as "filters". The last 15 years has seen a huge publication surge of candidate filter methods, with no common way to relate them or pick the right one for the right task.

We focus on filters based on mutual information. This talk will give an overview of information theoretic methods, and present a recent unifying framework that shows the existence of a continuous space of filters. Each paper over the last 15 years corresponds to a point in the space, most of which has never been explored.

## Feature Selection by Filters: A Unifying Perspective

Gavin Brown, University of Manchester

UK Symposium on Knowledge Discovery and Data Mining 2009, Salford

### Outline

We will:

- review the need for feature selection
- review general feature selection methods
- review **mutual information** based methods
- show the literature is confusing

Then we will:

- provide a novel perspective
- unify 12 separate methods over 2 decades
- show a **continuous space** of existing & new methods
- do some experiments

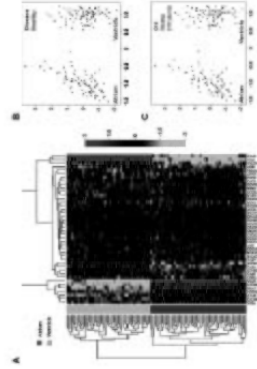
### Outline

We will:

- review the need for feature selection
- review general feature selection methods
- review **mutual information** based methods
- show the literature is confusing

### High-dimensional data : Gene expression

- high-throughput technology produces 1000s of measurements
- typical: > 10,000 features (columns), < 50 patients (rows)
- very very easy to overfit



## High-dimensional data : Netflix

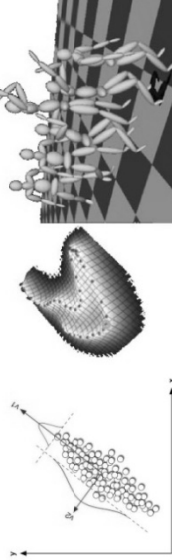
- 480,000 users rated 18,000 movies over several years
- $\approx 100,000,000$  ratings (i.e. sparse matrix)
- hugely expensive on time & memory



Plus: **Million dollar prize!**

## Handling Hi-D data?

- Feature extraction - PCA, ICA, GP-LVM, etc
- $S = f(\Omega)$ , usually  $S = w^T \Omega$ .



## Handling Hi-D data?

### Feature Extraction

$$S = f(\Omega), \text{ usually } S = w^T \Omega.$$

### Feature Selection

$$S \subseteq \Omega.$$

With Feature Extraction, we lose the *meaning* of original features.

With Feature Selection, we identify meaningful *subsets* of originals. Combinatorial optimization over space of possible feature subsets.

## Feature Selection: Wrappers

- Input:** large feature set  $\Omega$
- 10 Identify candidate subset  $S \subseteq \Omega$
  - 20 While !stop\_criterion()
    - Evaluate error of a classifier using  $S$ .
    - Adapt subset  $S$ .
  - 30 Return  $S$ .

- Pros: excellent performance for the chosen classifier
- Cons: computationally and memory-intensive

## Feature Selection: Filters

- Input:** large feature set  $\Omega$
- 10** Identify candidate subset  $S \subseteq \Omega$
- 20** While `!stop_criterion()`  
    Evaluate utility function  $J$  using  $S$ .  
    Adapt subset  $S$ .
- 30** Return  $S$ .
- Pros: fast, provides generically useful feature set
  - Cons: generally higher error than wrappers

## Common Filter Criteria

- Utility function  $J$  is some kind of statistical dependence measure.  
Rank features by the value of  $J$ .
- Pearson's correlation coefficient.
  - Chi-squared coefficient
  - Bhattacharyya distance
  - Mutual Information
- Pearson can only detect linear relationships.  
Chi-squared and Bhattacharyya assumes Gaussian variables.

## Common Filter Criteria

- Utility function  $J$  is some kind of statistical dependence measure.  
Rank features by the value of  $J$ .
- Pearson's correlation coefficient.
  - Chi-squared coefficient
  - Bhattacharyya distance
  - Mutual Information

## Common Filter Criteria

- Utility function  $J$  is some kind of statistical dependence measure.  
Rank features by the value of  $J$ .
- Pearson's correlation coefficient.
  - Chi-squared coefficient
  - Bhattacharyya distance
  - Mutual Information
- Pearson can only detect linear relationships.  
Chi-squared and Bhattacharyya assumes Gaussian variables.
- Mutual information is non-parametric, and can detect arbitrary nonlinear relationships between  $X$  and  $Y$  :-)

## Mutual Information

$I(X; Y)$  - measure of dependence between feature  $X$  and target  $Y$ .

Zero when  $X$  is independent of  $Y$ .  
Increases as  $X$  and  $Y$  become dependent.

## Filters using Mutual Information

Rank features  $X_i, \forall i$  by their values of  $J = I(X_i; Y)$ .  
Retain the highest ranked features, discard the lowest ranked.

$i$	$J(X_i)$
35	0.846
42	0.811
10	0.810
654	0.611
22	0.443
59	0.388
...	...
212	0.09
39	0.05

Cut-off point decided by user, e.g.  $|S| = 5$ , so  $S = \{35, 42, 10, 654, 22\}$ .

## Mutual Information

$I(X; Y)$  - measure of dependence between feature  $X$  and target  $Y$ .

Zero when  $X$  is independent of  $Y$ .  
Increases as  $X$  and  $Y$  become dependent.

Defined as KL divergence:

$$I(X; Y) = \sum_X \sum_Y p(xy) \log \frac{p(xy)}{p(x)p(y)}$$

## Filters using Mutual Information

**Problem:** Highly correlated features are no use!

Suppose, from the previous slide,  $S = \{35, 42, 10, 654, 22\}$ , but 42 and 10 are almost identical!

We need **relevant** features, but not **redundant** features.

## Filters using Mutual Information

**Problem:** Highly correlated features are no use!

Suppose, from the previous slide,  $S = \{35, 42, 10, 654, 22\}$ , but 42 and 10 are almost identical!

We need **relevant** features, but not **redundant** features.

**Solution:** Penalize correlations.

$$J(X_n) = I(X_n; Y) - \sum_{X_i \in S} I(X_n; X_i)$$

(Batitti, IEEE TNN 1994)

## Filters using Mutual Information

**Another Solution:** 'Joint Mutual Information' (Yang & Moody, NIPS 1999)

$$J(X_n) = \sum_{X_i \in S} I(X_n; X_i; Y)$$

"How useful is  $X_n$  when paired with each of the existing features?"

## Filters using Mutual Information

**Another Solution:** 'Max Relevance Min Redundancy' (Peng et al, IEEE PAMI 2005)

$$J(X_n) = I(X_n; Y) - \frac{1}{|S|} \sum_{X_i \in S} I(X_n; X_i)$$

Smooths out noise by averaging across  $S$ !

## The Confusing Literature

Why should we use any of these? How do they relate?

Criterion	Full name	Authors
MIFS	Mutual Information Feature Selection	Batitti 1994
KS	Koller-Sahami metric	Koller & Sahami 1996
JMI	Joint Mutual Information	Yang & Moody 1999
MIFS-U	MIFS-'Uniform'	Kwak & Choi 2002
IF	Informative Fragments	Vidali-Naquet 2003
FCBF	Fast Correlation Based Filter	Yu et al 2004
CMIM	Conditional Mutual Info Maximisation	Fleuret 2004
MIRM	Max-Relevance Min-Redundancy	Peng et al 2005
ICAP	Interaction Capping	Jakulin 2005
CIFE	Conditional Informax Feature Extraction	Lin et al 2006
DISR	Double Input Symmetrical Relevance	Meyer 2006
MINRED	Minimum Redundancy	Duch 2006
IGFS	Interaction Gain Feature Selection	Akadi 2008



## Let's start from scratch

Define relevance of a feature **set**  $S = \{X_1, X_2, \dots, X_n\}$  as

$$I(X_{1:n}; Y).$$

Overall we know we want to maximize this.

## Let's start from scratch

Define relevance of a feature **set**  $S = \{X_1, X_2, \dots, X_n\}$  as

$$I(X_{1:n}; Y).$$

Overall we know we want to maximize this.

We know that we need to:

- have large values of  $I(X_i; Y)$
- have small(ish?) values of  $I(X_j; X_i)$
- measure information properties of  $N$  variables

Shannon's Mutual Info  $I(X_1; X_2)$  is a function of two variables.

Not able to measure properties of multiple ( $N$ ) variables.

## Let's start from scratch

Define relevance of a feature **set**  $S = \{X_1, X_2, \dots, X_n\}$  as

$$I(X_{1:n}; Y).$$

Overall we know we want to maximize this.

We know that we need to:

- have large values of  $I(X_i; Y)$
- have small(ish?) values of  $I(X_j; X_i)$
- measure information properties of  $N$  variables

## Background: Multivariate Information Theory

Shannon's Mutual Info  $I(X_1; X_2)$  is a function of two variables.

Generalized by **Interaction Information** (McGill 1954),

$$I(\{X_1, X_2, X_3\}) = I(\{X_1, X_2\} | X_3) - I(\{X_1, X_2\})$$

- Takes set argument  $S$
- Reduces to Shannon's case with  $|S| = 2$
- Conditional form obtained by marginalising over  $X_3$
- General form for arbitrary size  $S$  defined recursively

$$I(\{S \cup X\}) = I(S|X) - I(S)$$

## A Novel Expansion Theorem

### Theorem 1

Given a set of input features  $S = \{X_1, \dots, X_n\}$ , and a target  $Y$ , their Shannon mutual information can be expanded as

$$I(X_{1:n}; Y) = \sum_{T \subseteq S} I(\{T \cup Y\}), \quad |T| \geq 1.$$

The Shannon Mutual Information between  $X_{1:n}$  and  $Y$  expands into a sum of Interaction Information terms. Note that  $\sum_{T \subseteq S} I(\{T \cup Y\})$  should be read, "sum over all possible subsets  $T$  drawn from  $S$ ".

### Proof

Brown, AI-STATS 2009

## A Novel Expansion Theorem

### Intuition

$$I(X_{1:n}; Y) = \text{0th order interactions} \\
 + \text{1st order} \\
 + \text{2nd order} \\
 + \dots \\
 + \text{nth order}$$

### Example

$$I(X_{1:3}; Y) = I(\{X_1, Y\}) + I(\{X_2, Y\}) + I(\{X_3, Y\}) \\
 + I(\{X_1, X_2, Y\}) + I(\{X_1, X_3, Y\}) + I(\{X_2, X_3, Y\}) \\
 + I(\{X_1, X_2, X_3, Y\}).$$

Order 0 means  $X_i$  does not interact with other features.

Order-1 means  $X_i$  interacts with one other feature.

Order-2 means  $X_i$  interacts simultaneously with two other features.

## A Novel Perspective

A more precise statement of the problem, we want to maximise:

$$J = I(X_{1:n}; Y) - I(X_{1:n-1}; Y)$$

That is the difference in information, **with** and **without**  $X_n$ . Involves high-dimensional probabilities,  $p(x_1, x_2, \dots, x_n, y)$ .

## A Novel Perspective

A more precise statement of the problem, we want to maximise:

$$J = I(X_{1:n}; Y) - I(X_{1:n-1}; Y)$$

That is the difference in information, **with** and **without**  $X_n$ . Involves high-dimensional probabilities,  $p(x_1, x_2, \dots, x_n, y)$ . Objective:

$$J = I(X_{1:n}; Y) - I(X_{1:n-1}; Y) \\
 = \text{approx} - \text{approx} \\
 = I(X_n; Y) - \sum_{k=1}^{n-1} I(X_n; X_k) + \sum_{k=1}^{n-1} I(X_n; X_k | Y).$$

## Existing Filters can be Re-written

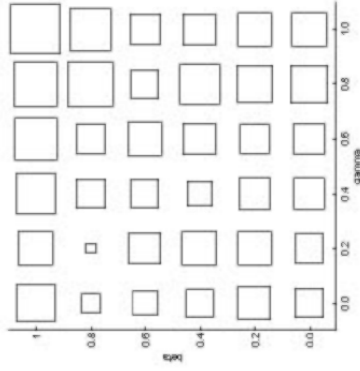
Joint Mutual Info, (Yang & Moody 1999)

$$J_{jms} = \sum_{X_i \in \mathcal{S}} I(X_{r_0}; X_i; Y)$$

With some funky information calculus...

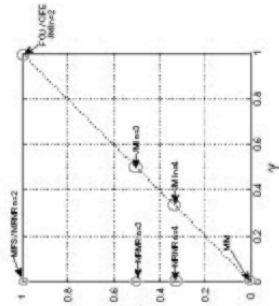
$$J_{jms} = I(X_{r_0}; Y) - \frac{1}{r-1} \left\{ \sum_{k=1}^{r-1} I(X_{r_0}; X_k) - \sum_{k=1}^{r-1} I(X_{r_0}; X_k | Y) \right\}$$

## Exploring the Space



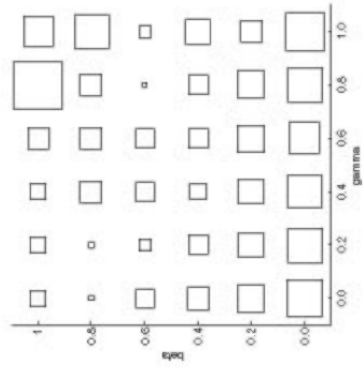
Lung Cancer data

## A Space of Feature Filters



$$J = I(X_{r_0}; Y) - \beta \sum_{k=1}^{r-1} I(X_{r_0}; X_k) + \gamma \sum_{k=1}^{r-1} I(X_{r_0}; X_k | Y)$$

## Exploring the Space



SRBCT tumor data

## Conclusion

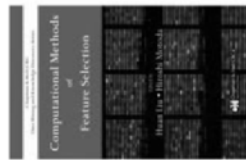
### Done:

- information theoretic clouds cleared, multivariate sunshine emerging.
- 12 separate methods united
- new space of novel criteria defined and explored

### Ongoing work:

- Determining  $\beta/\gamma$  automatically from data
- Markov blanket methods
- links to ICA and information bottleneck
- Renyi entropy

## Feature Selection Literature



**Neil Berry**  
**Deloitte**

**Real world applications of data mining technology:**  
**"The cook, the thief, his wife and her lover" a financial services case study.**

**Abstract**

Financial crime is something that impacts each and every one of us. The average general insurance policy in the UK costs £40 more than it should due to fraud, and overall fraud is estimated to cost the insurance industry over £1.6bn a year. Extrapolate this across other financial services products, and link it with other areas of concern e.g. Anti Money Laundering and Sanctions Compliance, and you have the makings of a huge problem.

Data mining techniques are being increasingly applied to these issues in ever more innovative ways. Given the scale of potential fines (up to £250,000 per transaction) and potential reputational damage, not to mention huge financial losses, institutions have a large vested interest in getting this right! This presentation will use real examples from the world of financial crime to illustrate different techniques and methods that are currently deployed in the market to tackle these problems. It will also look to the future, to examine how the industry is changing, and what some of the challenges may be as technology tries to advance to keep pace with the criminals.

**Deloitte.**

# The Cook, the Thief, his Wife and her Lover.

Real world applications for data mining technology: A financial services case study

Dr. Neil Berry  
Director  
Enterprise Risk Services

April 09

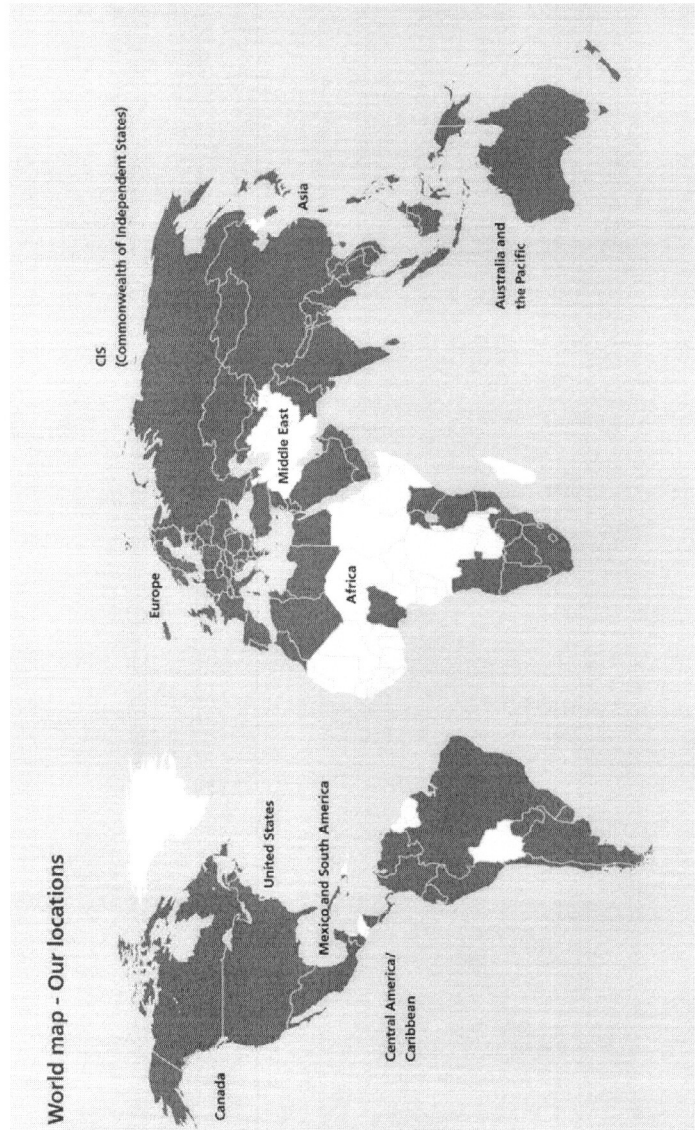
# Contents

1. Introduction	3
2. The Cook	4
3. The Thief	6
4. His Wife	8
5. Her Lover	10
6. Conclusions	12

# 1. Introduction

Deloitte is a global organisation with a presence in over 110 countries throughout the world. Our significant data management practice is recognised as the pre-eminent data team and has an ever growing portfolio of high profile, multi-national clients, offering a broad range of technical skills and supporting industry standard toolkits.

We have specialist teams of analytics and data mining experts often recruited from academia, and have partnerships with some of the largest data mining and analytics companies in the world e.g. SAS, SPSS, Oracle, IBM, SAP etc



- Asia Pacific: Singapore, Australia (includes Papua New Guinea), Thailand, Vietnam, Canada, EMEA: Albania, Algeria, Azerbaijan, Bahrain, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Egypt, India, Indonesia, Japan, Korea, Malaysia, New Zealand (includes Fiji and the Cook Islands), Philippines
- Australia and the Pacific: Cayman Islands, Chile, Colombia, Costa Rica, Ecuador, Guatemala, Jamaica, Mexico, Nicaragua, Paraguay, Peru, Uruguay, Venezuela, United States
- Europe: Estonia, Finland, France, FYR Macedonia, Gaza Strip/West Bank, Germany, Gibraltar, Greece, Hungary, Iceland, Ireland, Israel, Jordan, Kazakhstan, Kenya (includes Tanzania and Uganda), Kyrgyzstan, Kuwait, Latvia, Lebanon
- Middle East: Libya, Lithuania, Luxembourg, Malta, Moldova, Morocco, Netherlands, Nigeria, Norway, Oman, Poland, Portugal, Qatar, Romania, Russia, Saudi Arabia, Serbia and Montenegro, Slovak Republic, Slovenia, South Africa (includes Angola, Botswana, Malawi, Mozambique, Namibia)
- Africa: Swaziland and Zimbabwe, Spain, Sweden, Switzerland, Syria, Tunisia, Turkey, United Arab Emirates, United Kingdom, Uzbekistan, Yemen
- United States: Argentina, Aruba/Netherlands Antilles, Bahamas, Barbados, Bermuda, Brazil, British Virgin Islands
- CIS (Commonwealth of Independent States): Swaziland and Zimbabwe, Lithuania, Luxembourg, Malta, Moldova, Morocco, Netherlands, Nigeria, Norway, Oman, Poland, Portugal, Qatar, Romania, Russia, Saudi Arabia, Serbia and Montenegro, Slovak Republic, Slovenia, South Africa (includes Angola, Botswana, Malawi, Mozambique, Namibia)
- Asia: Singapore, Taiwan, Vietnam, Canada, EMEA: Albania, Algeria, Azerbaijan, Bahrain, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Egypt, India, Indonesia, Japan, Korea, Malaysia, New Zealand (includes Fiji and the Cook Islands), Philippines



## 2. The Cook

Financial crime touches everyone in many and varied ways. The impact is not solely financial, and organised crime is heavily involved. All financial institutions are heavily regulated but ultimately we pay the price in a very real way!

The “Cook” works for a bank in the back office, and is up to no good. He has decided that he wants to “earn” some extra cash and is supplying details of customers accounts to people outside the bank. He works as part of a payments group for private clients in the banks wealth management department and processes high value payments internationally. To cover up his activity he tries to “cook the books” by making his activities as difficult to trace as possible.

Remember frauds are always perpetrated by people (either as individuals or acting as part of a group either from inside or from outside an organisation)



### What are the problems that he might cause?

- Account breaches / takeovers
- Stolen account and credit card details
- Loss of insurance policies
- Creates a bad debt profile & poor credit history
- Breaches of US sanctions compliance
- Reputational & operational risk to the bank

### What are the things that will help us (or not!)

- Application and systems controls
- Segregation of duties
- Business processes
- Data
- Technology
- Insider threat monitoring

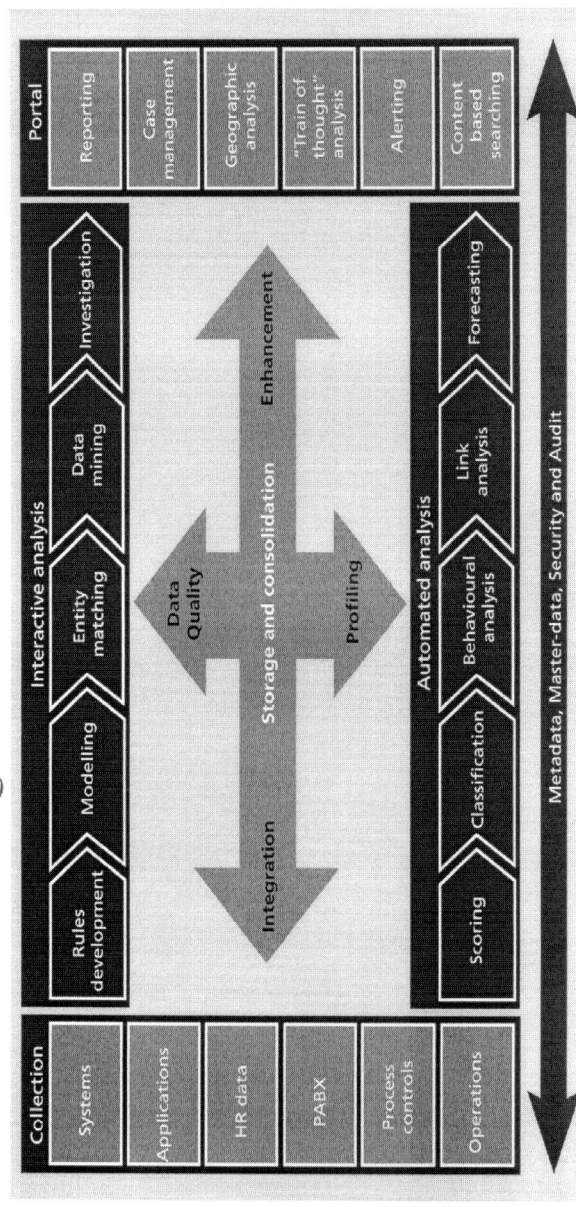
## 2. What can we do to stop him?

Insider Threat Monitoring – behavioural analytics

Watch list Compliance – fuzzy matching of names, banks etc

Account Transaction Profiling

<p><b>Watch List Matching</b></p> <p><b>What results do you get?</b></p> <ul style="list-style-type: none"> <li>•No technique performs better than all others</li> <li>•Pattern matching methods clearly outperform phonetic encoding methods</li> <li>•Simple phonetic encoding methods perform better than more complex ones</li> <li>•Combined techniques do not perform as well as expected</li> <li>•Surnames are harder to match than given names (due to complete name changes)</li> <li>•Dictionary &amp; Transliteration techniques are vital for common matches e.g. Elizabeth, Liz, Lizzy etc</li> </ul>
---



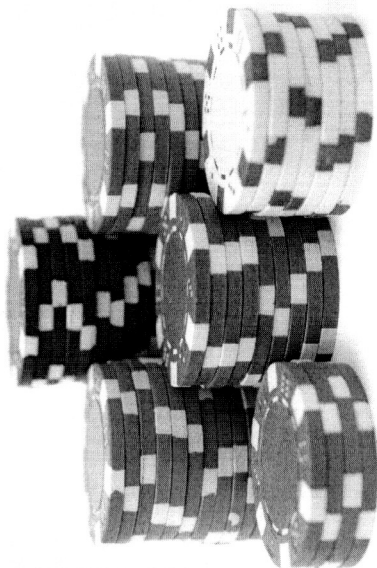
### Account Transaction Profiling

### 3. The Thief

The Thief is the person controlling The Cook on the outside. The Thief has connections into local organised crime, and pays the Cook for the information he supplies and the payments that he makes via his bank connections. He runs a number of “Cooks”.

The “Thief” is the first step on the chain upwards. He collates the financial information that his “Cooks” sell him, and he re-sells this on a number of websites to other Thieves. He has connections throughout the world thanks to the internet, and runs various ‘phishing’ scams from countries where banking controls are more lax. He believes that the internet gives him complete anonymity.

On average he can sell an individuals credit card and security details for £8.



#### What are the things we need to understand about him?

- Motive** – Behaviours and background
- Network** – Is he connected to anyone else?
- Previous** – Does he have a history?
- Money** – How has he hidden his ill gotten gains?

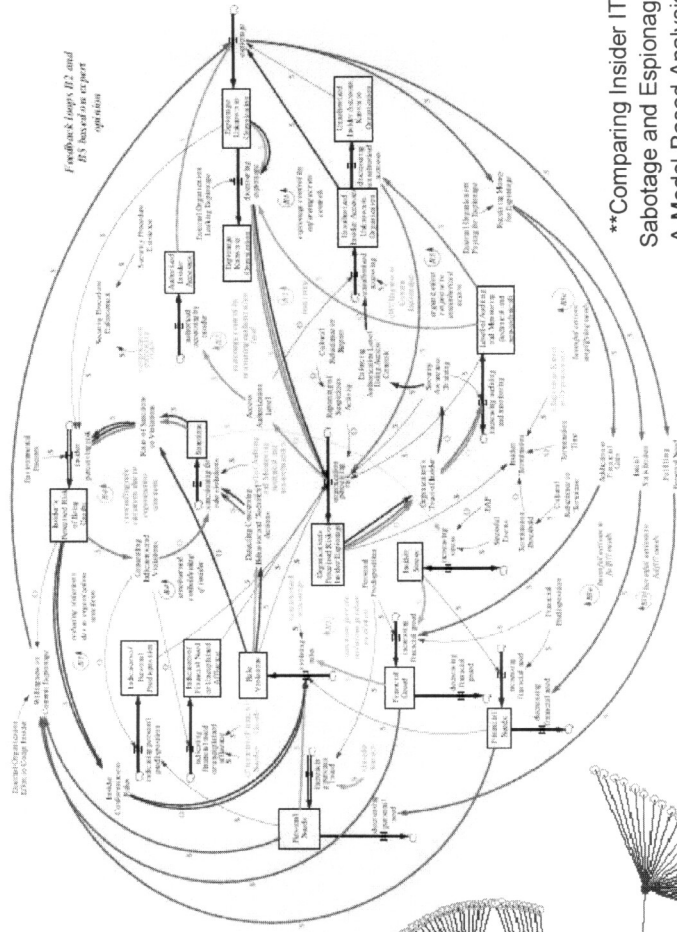
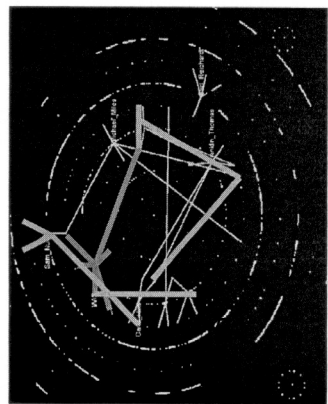
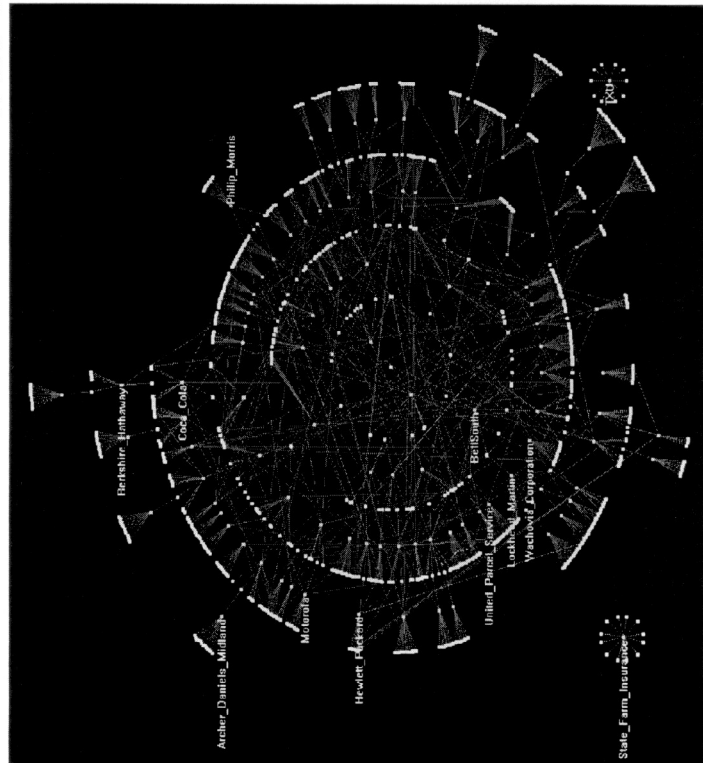
#### What are the things that help us (or not!)

- Data
- Technology
- Intelligence
- Good “Old Fashioned” Policing

# 3. What can we do to stop him?

Who is he and where is he? - Device Fingerprinting

Who does he do business with? - Social Network Analysis



\*\*Comparing Insider IT Sabotage and Espionage: A Model-Based Analysis, Band et al, Dec 2006

## 4. His Wife

The 'Wife' is an intermediary, and launders money through the international markets by using local contacts like the Thief. She has contacts into international organised crime, and may see many tens of millions of pounds going through her hands.



The 'Wife' controls many individuals through her connections within organised crime. She may act for many such groups, but will never act alone.

She uses her influence (bribery, corruption, threats, drugs, violence) to coerce people to provide financial "services" for her clients to help them to try to evade international legislation.

Her prime focus is on laundering money through the system to turn her clients ill gotten gains into "clean" financial instruments that can be used in the open market.

### What are the things she needs to do to be successful?

**Placement** – successful deposits of large quantities of cash.

**Layering** – moving the money through different banks / accounts

**Extraction** – how to get "clean" instruments out of the system at the end?

**Evade KYC** – false documents, account take overs, alternative remittance systems etc

### What do we know about her that might help us?

Her associations with others

Links to fake documents

Existing criminal record

Existing financial crime alerts

Present on watch lists

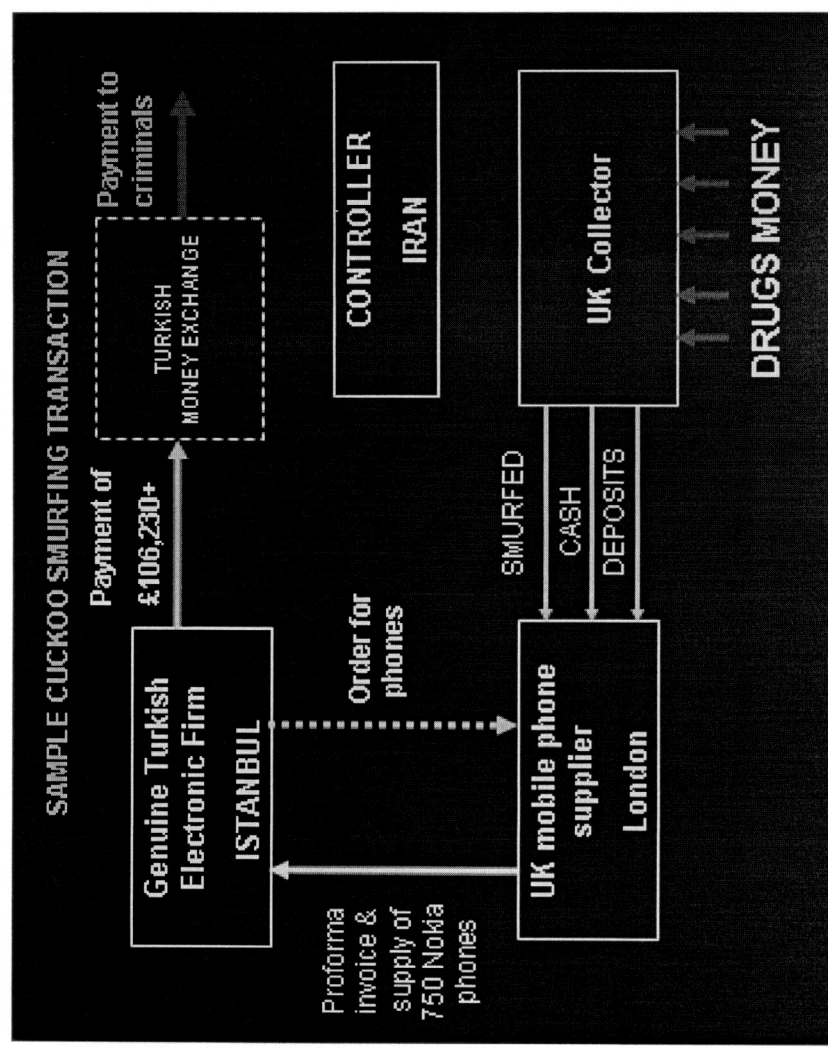
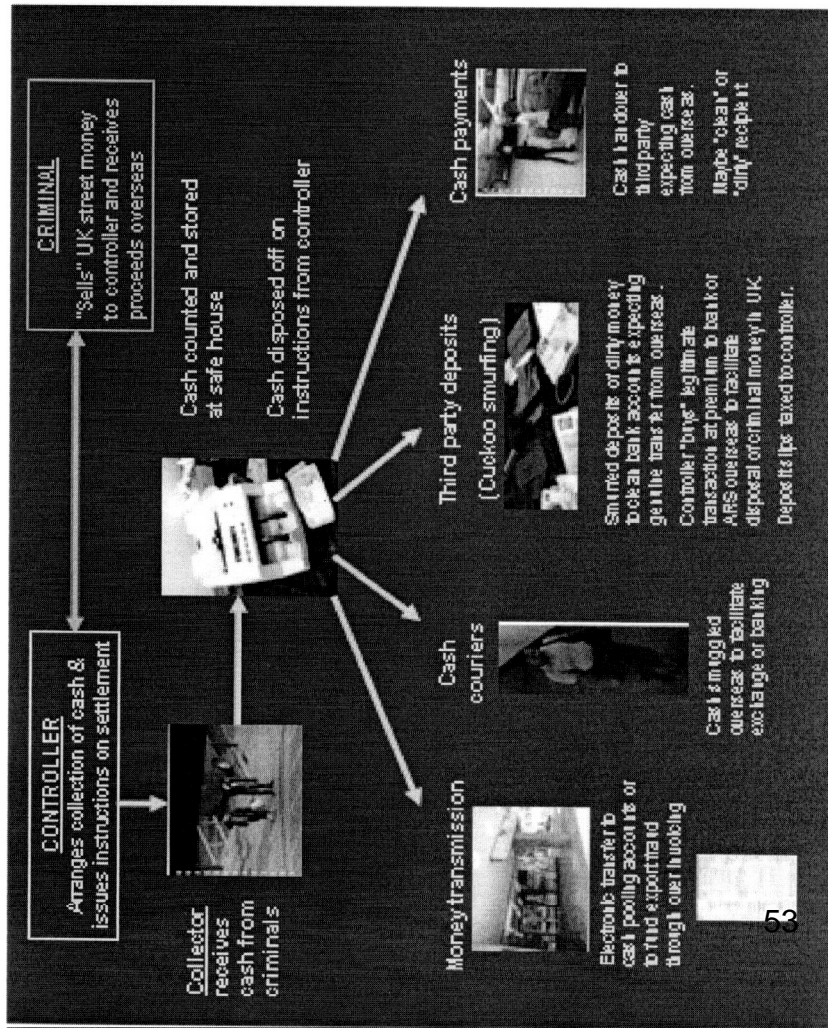
Existing banking relationships

Flows of cash

Mobile phone records

## 2. What can we do to stop her?

KYC, EDD, Alternative remittance systems & Payments monitoring  
 Terrorist financing typologies – What does she do and how?  
 Analysis of mobile phone call records – Social networks  
 Layering & Smurfing – She is the controller

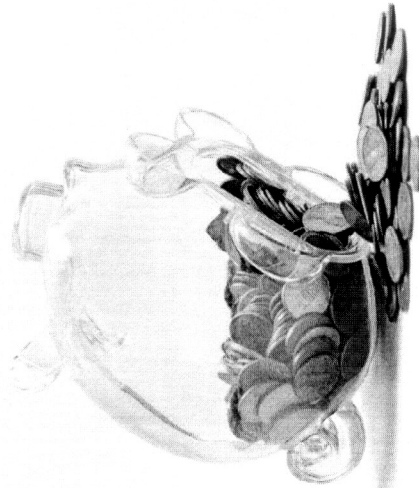


## 5. Her Lover

Her 'Lover' is in fact an international arms dealer who is suspected of terrorist financing as well as having links to South American drug cartels.

Her 'Lover' has a string of legitimate businesses that he uses to mask his real activities, including arms trading, heavy manufacturing goods, casinos, money exchange and transfer businesses. He also has a number of illegal operations and activities and links to other illegal and sanctioned organisations and individuals.

He is one of the key individuals in the global web of money laundering and illegal trade activities thought to be in excess of \$1.5 trillion per annum!



### Why should we be interested?

Who's responsibility is this?  
Who should fund the work / investigation etc?  
How can data mining help here?  
Which agencies might use which techniques?

### What are the things that might help?

Voice records – satellite phones  
International payments  
Video evidence  
Intelligence reports

But how do we mine these types of data?

## 2. What can we do to stop him?

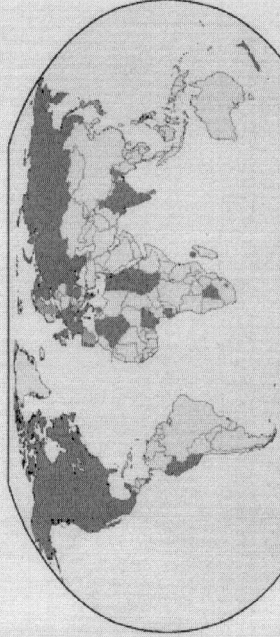
Sanctions compliance – financial & trade.

Customs and border control - eBorders & shipping

International multi agency intelligence sharing

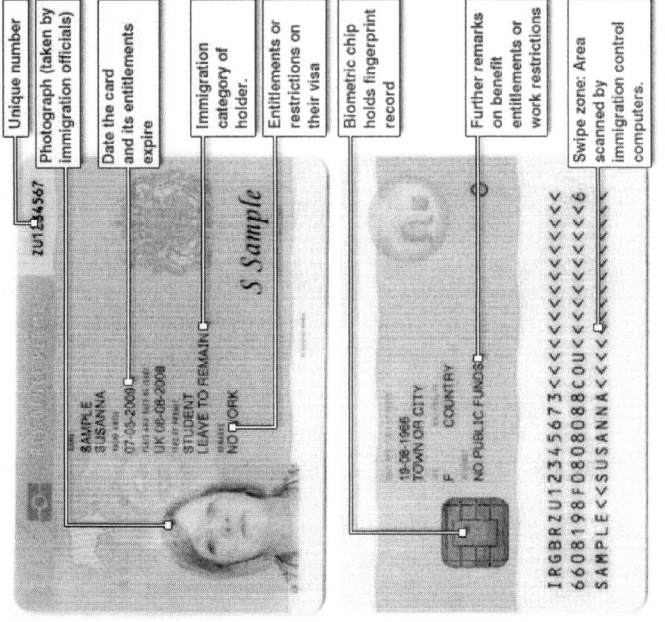
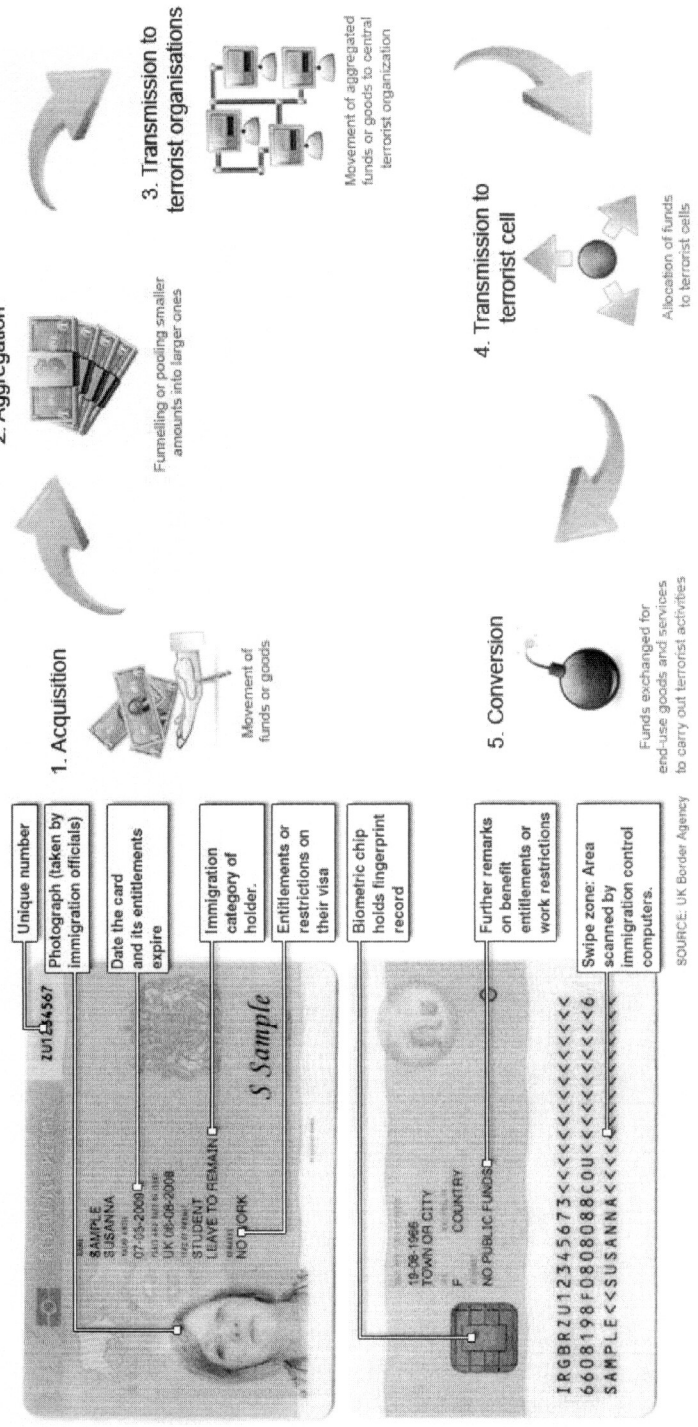
Mobile phone & internet data tracking

Signatories to the International Convention for the Suppression of Terrorist Financing



- |   |  |   |
|---|--|---|
| <b>North America</b><br>Canada<br>United States                 | <b>Europe and Eurasia</b><br>Czech Republic<br>Estonia<br>Finland<br>France<br>Georgia<br>Germany<br>Greece<br>Italy<br>Macedonia, The<br>Former Yugoslav<br>Republic of<br>Malta<br>Netherlands<br>Portugal<br>Romania<br>Russia<br>San Marino<br>Ukraine<br>United Kingdom | <b>Asia/Oceania</b><br>India<br>New Zealand<br>Sri Lanka<br>Japan |
| <b>Latin America</b><br>Costa Rica<br>Ecuador<br>Mexico<br>Peru | <b>Africa and the Middle East</b><br>Algeria<br>Botswana<br>Comoros<br>Egypt<br>Gabon<br>Israel<br>Lesotho<br>Nigeria<br>Sudan   |   |

States that have signed and ratified EU and G-8 states that have NOT signed and/or ratified  
 \*Signed but not ratified



SOURCE: UK Border Agency



# 6. Conclusions

Leveraging the value of data and analytics to help solve real world problems is an increasingly important skill in today's markets.

But what of the future?



Key challenges for the future of data mining and analytics	
Data Volumes	Some corporate data sets now exceed 100Tb – mobile call switching data (1Pb)?
Real time analytics	Need further development in the area of analytic performance enhancement
Analytics on demand	How do we create an architecture that allows large numbers of people to intelligently query data – green issues!
Visualisation	How do we help people to understand increasingly complex outcomes?
New methods and approaches including unstructured Data: –Text –Speech –Video	What can be created in terms of new algorithms to improve and expand the range of possible analyses?

This document is confidential and prepared solely for your information. Therefore you should not, without our prior written consent, refer to or use our name or this document for any other purpose, disclose them or refer to them in any prospectus or other document, or make them available or communicate them to any other party. No other party is entitled to rely on our document for any purpose whatsoever and thus we accept no liability to any other party who is shown or gains access to this document.

Deloitte LLP is a limited liability partnership registered in England and Wales with registered number OC303675 and its registered office at 2 New Street Square, London EC4A 3BZ, United Kingdom. Deloitte LLP is the United Kingdom member firm of Deloitte Touche Tohmatsu ('DTT'), a Swiss Verein, whose member firms are legally separate and independent entities. Please see [www.deloitte.co.uk/about](http://www.deloitte.co.uk/about) for a detailed description of the legal structure of DTT and its member firms.

**Susan Crow**  
**Robert Gordon University**

**Knowledge Discovery from Case Data**

Case-based reasoning systems solve problems by retrieving and reusing similar experiences from the case base as the fundamental knowledge source. However, the cases can be used for more than solving problems, and the knowledge available in a collection of cases may be exploited to improve the system's problem-solving. The ability to use the cases to identify and understand regular and complex regions of the problem-solving landscape offers the potential for data selection, pre-processing, data cleaning and knowledge maintenance for case-based reasoning systems. The cases also capture implicit knowledge that may be learned to improve the retrieval of suitable cases and to enable effective adaptation of the retrieved solution to suit the new problem.

This talk explores introspection of the case knowledge and some attractive prospects to exploit its implicit knowledge.



# Knowledge Discovery from Case Data

Susan Crow  
The Robert Gordon University  
Aberdeen

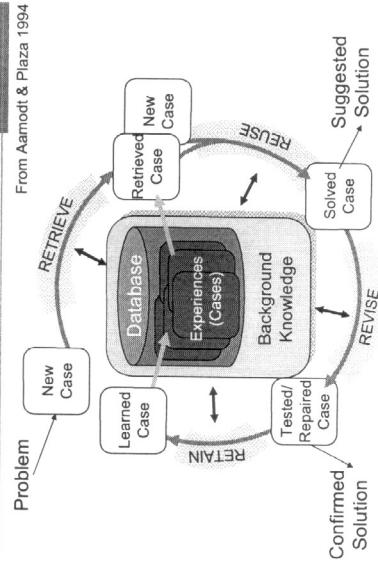
Thanks to Nirmala Wiratunga, Stewart Massie, PhD Students



## Outline

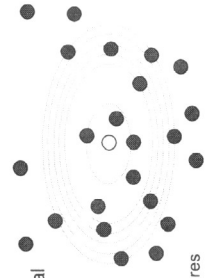
- Case Based Reasoning
- Mining the Case Base
  - CBR Knowledge Containers
  - Self Adaptation of the Case Base
- Agile CBR
  - Implicit knowledge for agility
- Conclusions

## Case Based Reasoning



## Retrieval is Key

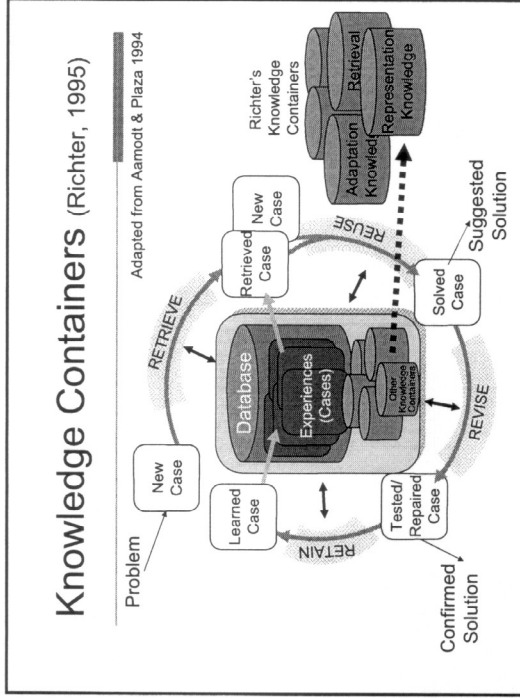
- Retrieve most similar
- k-nearest neighbour retrieval
  - like scoring in bowls or curling
- Distance is key
  - distance of pairs of individual feature values
  - mathematical distance
    - Manhattan
    - Euclidean
  - weighted sum of distance
    - relative importance of features





## Outline

- Case Based Reasoning
- Mining the Case Base
  - CBR Knowledge Containers
  - Self Adaptation of the Case Base
- Agile CBR
  - Implicit knowledge for agility
- Conclusions



## Tablet Formulation

- Drug
  - active ingredient (~25%)
- Filler
  - provides bulk (~65%)
- Binder
  - cohesiveness
- Lubricant
  - eject from die
- Disintegrant
  - break down
- Surfactant
  - aids mixing

## FormuCase

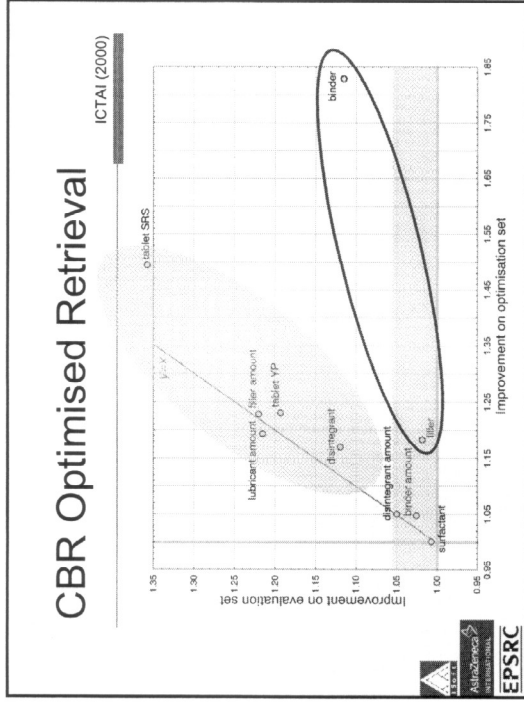
problem		case		solution		extra
physical properties	drug	chemical properties	form	amount	amount	info
1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	5	5	5	5	5	5
6	6	6	6	6	6	6
7	7	7	7	7	7	7
8	8	8	8	8	8	8
9	9	9	9	9	9	9
10	10	10	10	10	10	10
11	11	11	11	11	11	11
12	12	12	12	12	12	12
13	13	13	13	13	13	13
14	14	14	14	14	14	14
15	15	15	15	15	15	15
16	16	16	16	16	16	16
17	17	17	17	17	17	17
18	18	18	18	18	18	18
19	19	19	19	19	19	19
20	20	20	20	20	20	20
21	21	21	21	21	21	21
22	22	22	22	22	22	22
23	23	23	23	23	23	23
24	24	24	24	24	24	24
25	25	25	25	25	25	25
26	26	26	26	26	26	26
27	27	27	27	27	27	27
28	28	28	28	28	28	28
29	29	29	29	29	29	29
30	30	30	30	30	30	30
31	31	31	31	31	31	31
32	32	32	32	32	32	32
33	33	33	33	33	33	33
34	34	34	34	34	34	34
35	35	35	35	35	35	35
36	36	36	36	36	36	36
37	37	37	37	37	37	37

- Problem: physical & chemical drug properties and dose
- Solution: filler,binder,lubricant,disintegrant,surfactant & amounts
- Extra tablet properties
- CBR retrieval produces competitive formulations
- but further knowledge improves formulation accuracy
- Formulation is a challenging CBR domain



## CBR Retrieval Knowledge

- Genetic Algorithm
  - feature selections for index
  - feature importances for k-NN retrieval



## CBR Adaptation Knowledge

## CBR Adaptation Knowledge

- Adaptation Training Data
  - to learn adaptation rules
- Feature-based adaptation experts
  - Solubility, YP, SRS, Bonding, Stability
    - "If drug is insoluble and retrieved filler is insoluble then select another filler with increased YP"
- Ensemble of adaptation experts
  - Balances
    - design constraints adaptation
    - compatibility requirements



## CBR Ensemble Adaptation

- Binder improvement is dramatic
  - 34% to 74%!

Trial	RetrieveOnly	1-NN-Ensemble	C4.5-Ensemble	RISE-Ensemble
1	30%	30%	30%	30%
5	35%	35%	35%	35%
10	40%	40%	40%	40%
15	45%	45%	45%	45%
20	50%	50%	50%	50%
25	55%	55%	55%	74%

Accuracy

Trial

RetrieveOnly 1-NN-Ensemble C4.5-Ensemble RISE-Ensemble

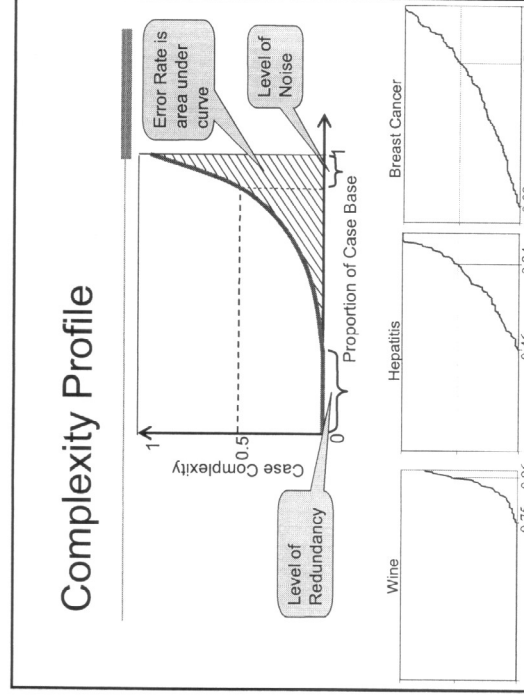
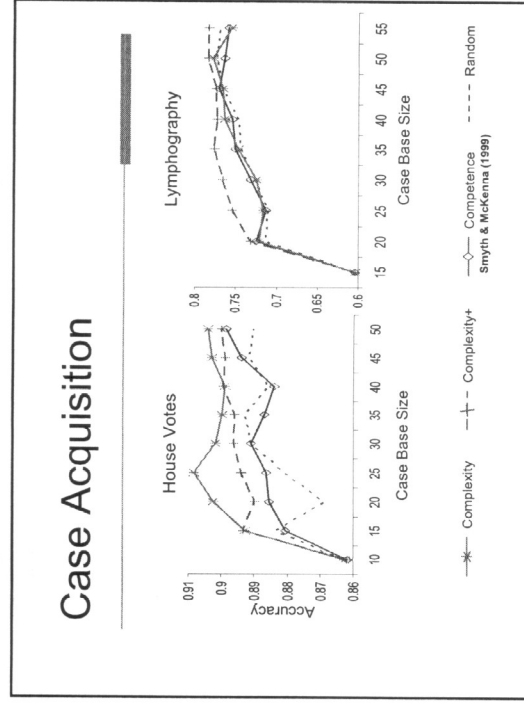
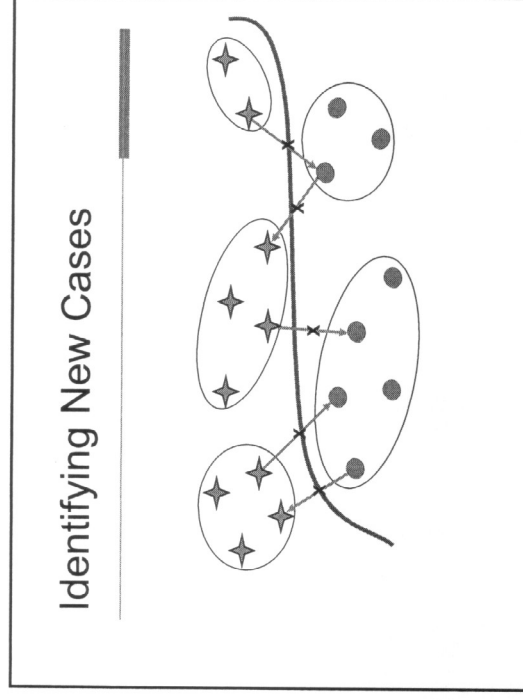
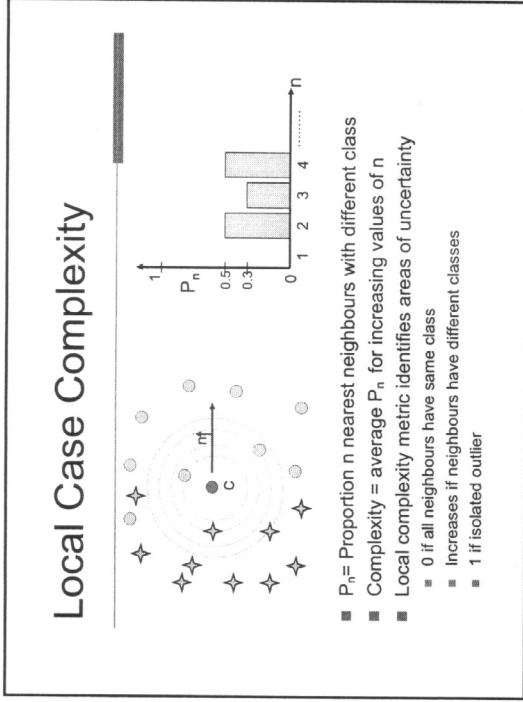
## CBR Knowledge Discovery

- Retrieval Knowledge
  - feature selection and weighting
    - improves all quantities
    - improves most predictions
- Adaptation Knowledge
  - Ensembles of rules/decision trees
    - improves all quantities & predictions
      - improves filler substantially
      - doubles binder accuracy

## Outline

- Case Based Reasoning
- Mining the Case Base
  - CBR Knowledge Containers
  - Self Adaptation of the Case Base
- Agile CBR
  - Implicit knowledge for agility
- Conclusions

## Similar Problems ~ Similar Solutions?

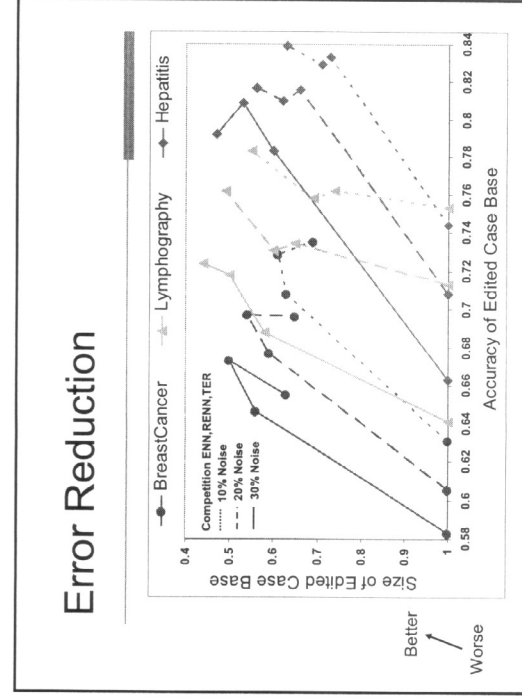
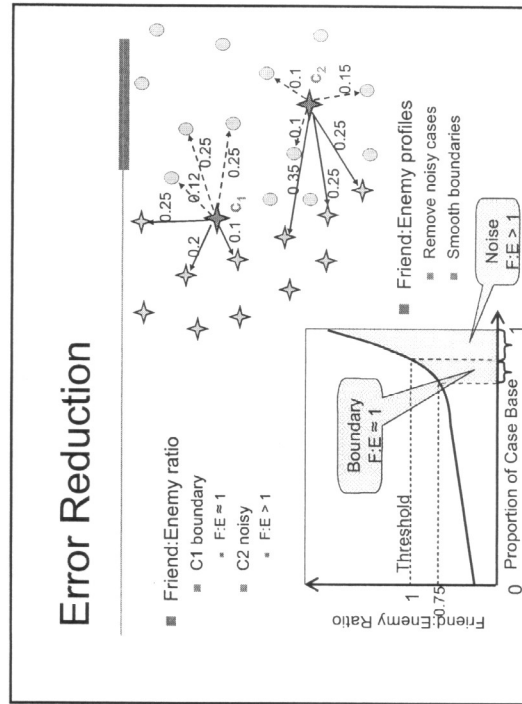
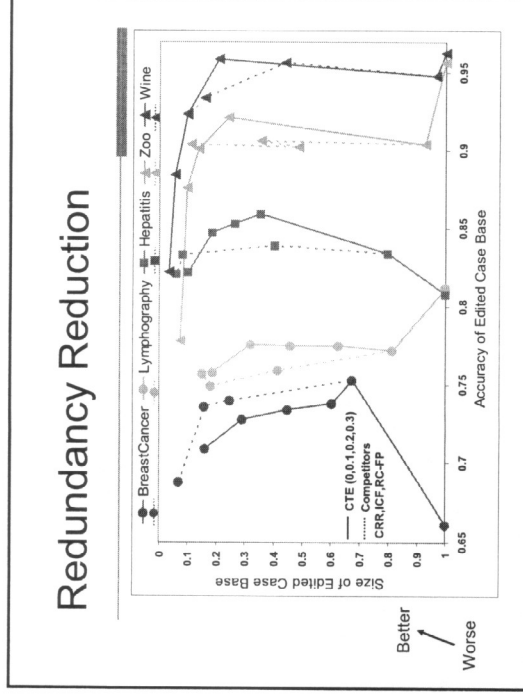






## Redundancy Reduction

- **Utility Problem**
  - benefit of extra case knowledge
  - cost of finding most similar cases for retrieval
- **Redundant cases are far from boundaries**
  - Remove cases with complexity below threshold
- **Friend:Enemy ratio breaks ties**
  - Remove low ratio cases
- **Recalculate complexities**
  - To avoid removing all cases in a cluster





## Self-Adapting Case Bases

- Complexity-based profiling
  - Enables experimentation and compromise
  - Distinguishes between
    - regularity (same class)
    - irregularity (near boundaries or isolated cases)
- Complexity-based data analysis (cases) leads to knowledge refinement (case base)
  - Case deletion, but also case discovery
- Neighbourhoods may be important for more than k-NN classifiers

## Outline

- Case Based Reasoning
- Mining the Case Base
  - CBR Knowledge Containers
  - Self Adaptation of the Case Base
- Agile CBR
  - Implicit knowledge for agility
- Conclusions

## The Agile Manifesto

<http://agilemanifesto.org>

- We are uncovering better ways of developing software by doing it and helping others do it
- Through this work we have come to value:
  - **Individuals and interactions** over processes and tools
  - **Working software** over comprehensive documentation
  - **Customer collaboration** over contract negotiation
  - **Responding to change** over following a plan

## Agile CBR

- Agility
  - A "just in time" instantiation of solution parts
    - *incrementally* extends a partial solution
    - but *opportunistically* rather than in a predetermined order
  - **Commitment** to parts of the solution are able to
    - influence the choices for others
    - and informs different levels of re-design and revision
- **Features of Agile CBR**
  - **Opportunism** – the ability to select which part of the solution to tackle next and how to extend the evolving solution incrementally
  - **Commitment** – the degree to which parts of the solution have growing support as the solution evolves, or are unchangeable because they contain essential requirements
  - **Flexibility** – the facility to refine, revise or backtrack from parts of the solution



**Sven F. Crone**  
**Lancaster University Management School**

**Classifying Imbalanced Datasets –Evidence from case studies in Business Data Mining**

Data Mining methods and procedures are routinely employed in business, but often neglect the specific properties of the dataset. For many corporate applications the actual class of interest, e.g. those responding to a direct mailing or defaulting on a loan, is often an underrepresented minority, which should be either targeted or avoided to ensure profitability. But how important is the data in the majority class of lesser interest? Is it required at all, or can we discard parts of it? And if so, is there some 'golden ratio' of negative to positive examples? A variety of simple to sophisticated sampling strategies are now available to under- or over-sample the existing data. This talk will demonstrate how different approaches of basic data sampling can enhance or impair predictive accuracy, using case studies from company projects in database marketing and direct mailing, credit and behavioural scoring, and predicting internet shopping adoption to distinguish customers between online-shoppers, browsers and offline shoppers.

# Classifying Imbalanced Datasets

Evidence from case studies in Business Data Mining



**Directors**  
 Prof. Robert Fildes  
 Prof. Peter Young  
 Dr. Sven F. Crone

**Researchers**  
 Dr. Steve Finlay  
 Dr. Alastair Robertson  
 Dr. Didier Soopramanien  
 Dr. Kostas Nikolopoulos

**Prof. Stephen Taylor**  
 Dr. Wlodek Tych  
 Prof. David Peel  
 Prof. Peter Pope

**Associated Experts**  
 Prof. Paul Goodwin  
 Dr. Andrew Eaves

**Research & PhD students**  
 Heiko Kausch, RA  
 Stavros Asimakopoulos  
 Xi Chen  
 Bruce Havel  
 Suzi Ismail  
 Nikolaos Kourentzes  
 Ioannis Stamatopoulos  
 Andrey Davidenko  
 Charlotte Brown  
 Hong Juan Liu  
 T Hu  
 John Prest  
 Huang Tao

**Visiting Researchers**  
 Prof. Geoff Allen  
 Dr. Yukun Bao  
 Young-Sang Cho

**Your Take Aways ...?**  
*Winston Churchill*  
 CHURCHILL

“Take away this pudding, it has no theme.”  
 Sir Winston Churchill (1915)

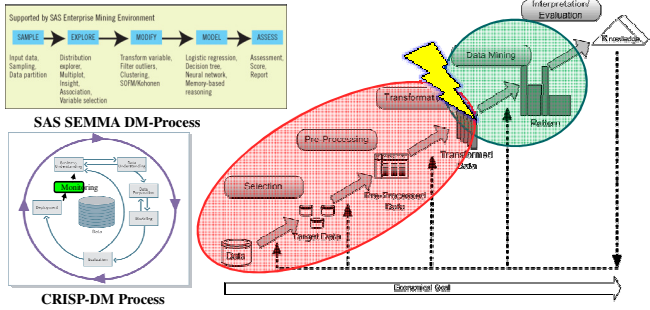
- Data Mining is preoccupied with Algorithms
- Data Preprocessing is equally important
- Most industry datasets are imbalanced
- Oversampling increases accuracy!
- Simple tricks exist – but are rarely used!
- Works for credit scoring, direct marketing ...

**Agenda**

- Sampling issues in Data Mining
- Case study 1: Direct Marketing
  - Cross-selling of Magazine subscriptions
  - Effect of data preprocessing: Sampling
  - Interaction of Sampling with Scaling & Coding
- Case study 2: Credit & Behavioral Scoring
  - Predicting consumer credit default
  - Effects of sample size
  - Effects of sample distribution
- Case study 3: Online Shopping Behaviour
  - Predicting consumer shopping channel choice
  - Sample distribution & multiple classes
- Conclusion & Take-aways

## Why (Under/Over) Sampling?

- Knowledge Discovery (KDD) = non-trivial **process** of identifying valid, novel, useful **patterns** in large data sets
  - Data Mining = only **one single** step in the KDD process
  - Data sample determines the whole process! (→ GiGO)
  - "Research seems **preoccupied** with algorithms"

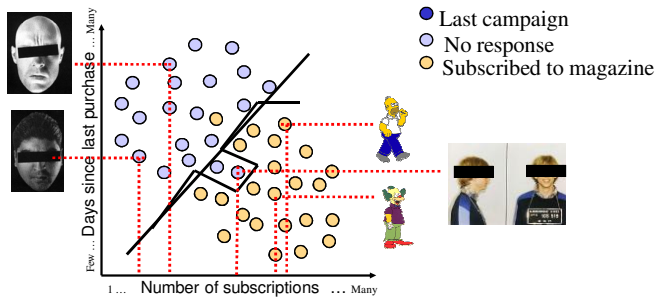


## Sampling in Direct Marketing Literature?

Input type*	Methods***	Parameter tuning	Data reduction**		Data projection		
			Feature Selection	Re-sampling	Continuous attributes	Categories	
					Standardisation	Discretisation	Coding
[2]	BMLP, LR, LDA, ODA	X				X	
[42]	MLP, LR, CHAID	X				X	
[43]	MLP, RBF, LR, GP, CHAID	X				X	
[44]	MLP, LR, LDA	X				X	
[4]	CHAID, CART	X				X	
[6]	MLP, LR	X				X	
[9]	LVO, RBF, 22 DT, 9 SC	X				X	
[45]	LDA, LR, KNN, KDE, CART, MLP, RBF, MOE, FAR, LVO	X				X	
[3]	MLP		X			X	
[7]		X	X			X	
[11]	LR, LS, SVM, KNN, NB, DT	X				X	
[10]	LDA, ODA, LR, BMLP, DT, SVM	X				X	
[46]	LSSVM, TAN, LP, KNN	X				X	
[47]	LR, MLP, BMLP	X	X				
[47]	LSSVM, SVM, DT, RL, LDA, ODA	X				X	
[48]	LR, NB, IBL	X				X	
[49]	DT, MLP, LR, FC	X				X	
[49]	FC	X				X	

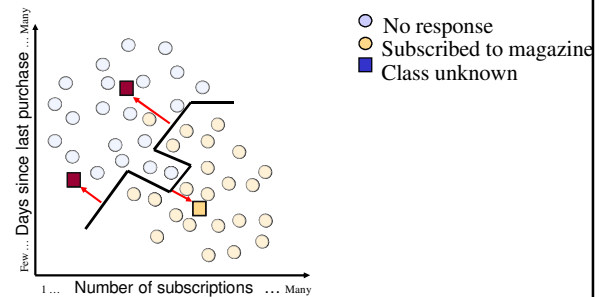
- Majority of direct marketing papers focus on algorithm tuning
- Only 3 papers consider Resampling / Instance Selection
- No analysis of the interaction with Sampling & Projection & ...

## Classification



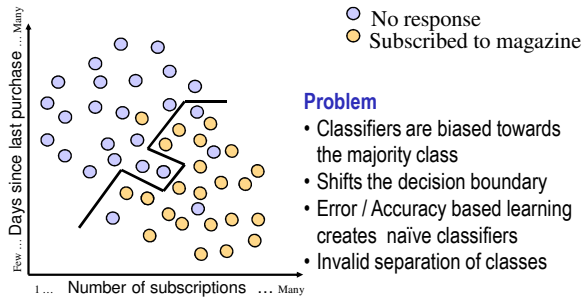
- Database of customers (instances)
- Known attributes for all customers (age, gender, existing subscriptions, ...)
- Known response (class membership) of buyers & non-buyers from past mailings
- Build a model to separate classes → decision boundary of different complexity

## Classification



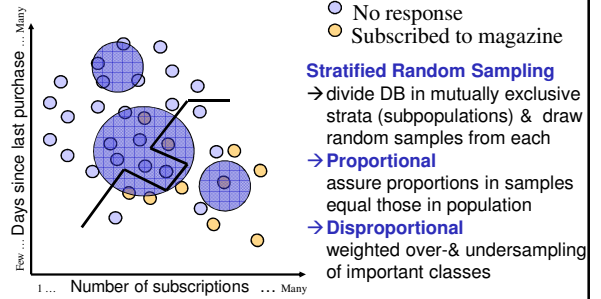
- Use the decision boundary to classify unseen instances
- Calculate on which side of hyperplane the instances lie (or distance)
- Assign class to unseen instances

## Reality Check: Imbalanced classes



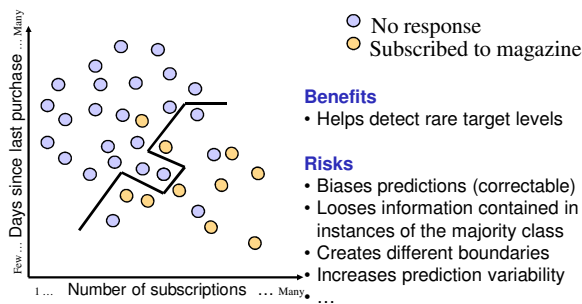
- **Balanced dataset = class distributions are equal**  $P(x|y=A)=P(x|y=B)$ 
  - proportional sampling or stratified sampling feasible
- **Imbalanced dataset = class distributions unequal**  $P(x|y=A) \gg P(x|y=B)$ 
  - The class of interest is often the minority (in most business applications)

## Imbalanced Data Sampling



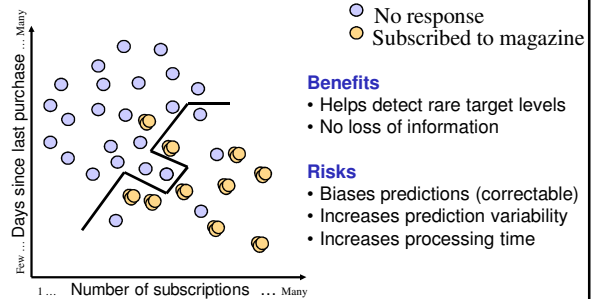
- **Size of the sample?**
- **Distribution / location of the sample?**

## Random Undersampling



- **Exclude random instances of the majority class**
- **Retain all instances of the minority class**
- **Establish a balanced class distribution**

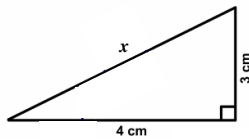
## Random Oversampling



- **Retain all instances of the majority class in the sample**
- **Duplicate identical instances of the minority class**
- **Establish a balanced class distribution**

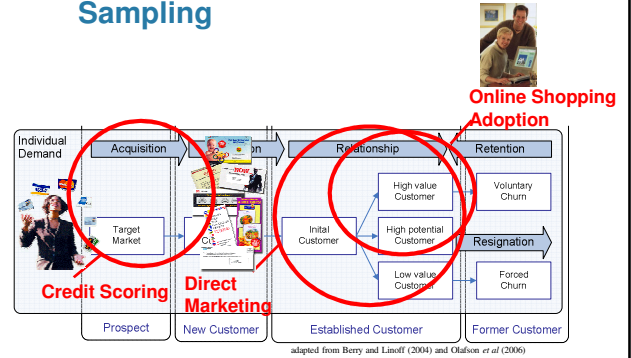
Ready for more theory...?

3. Find  $x$ .



→ rather some case studies ...!

## Case studies on Sampling



→ Evidence from 3 case studies using industry datasets

### Agenda

- Sampling issues in Data Mining
- Case study 1: Direct Marketing
  - Cross-selling of Magazine subscriptions
  - Effect of data preprocessing: Sampling
  - Interaction of Sampling with Scaling & Coding
- Case study 2: Credit & Behavioral Scoring
  - Predicting consumer credit default
  - Effects of sample size
  - Effects of sample distribution
- Case study 3: Online Shopping Behaviour
  - Predicting consumer shopping channel choice
  - Sample distribution & multiple classes
- Conclusion & Take-aways

## Business Case: Direct Marketing/Response Optimization

- Sell a magazine subscription to existing customers




- Whom to send mail to? (Which customers are most likely to respond?)
- How many customers to contact? (What is the optimal mailing size?)

→ Corporate project with leading German Publishing House  
 → Provided data set of past mailing campaigns  
 → Benchmark novel methods against in-house SPSS Clementine  
 → Explore Neural Networks (NN) and Support Vector Machines (SVM)



## Benefits of Direct Marketing



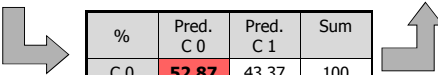
	Simple	With data mining
Addressees	100.000	Top 40% = 40.000
Cost	2€/mail = 200.000€	2,5€/mail = 100.000€
Response rate	0,5% = 500	1,0% = 400
Ø Sales volume	300€	300€
Sales volume	150.000€	120.000€
Revenue	<b>-50.000€</b>	<b>20.000€</b>

- Smaller mailing (number of letters sent) → lower costs (Euro 1.- per letter)
- Higher response rate → higher revenue
- More specific mailing → lower cost
- More relevant information → higher customer satisfaction

## NN get worse with learning ...

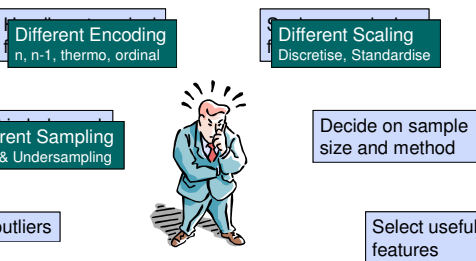
- **Wish to implement Neural Networks for next campaign**
  - In-house team (with no NN knowledge) outperformed us EVERY TIME!
  - Analyzed software, training parameters, etc. → internal competition
  - Observed expert in building models ... !

%	Pred C 0	Pred C 1	Sum	%	Pred. C 0	Pred. C 1	Sum
C 0	<b>61.86</b>	38.14	100	C 0	<b>72.96</b>	27.04	100
C 1	55.09	<b>44.81</b>	100	C 1	62.02	<b>37.98</b>	100
	116.95	82.95	<b>54.26</b>		134.98	65.02	<b>55.47</b>



%	Pred. C 0	Pred. C 1	Sum
C 0	<b>52.87</b>	43.37	100
C 1	47.13	<b>56.63</b>	100
	100	100	<b>54.75</b>

## Experimental Design: Different data pre-processing



Evaluate across 3 algorithms:  
→ Neural Networks (MLPs), Support Vector Machines & Decision Trees

## Dataset Structure

### Data set size

- 300,000 customer records
- 4,019 subscriptions sold
- Response rate of 1.3%

### Data set structure

- 18 categorical features
- 35 numerical features
- Binary target variable

### → Evaluated the Impact of Data Preprocessing

- Data **Sampling** (over sampling vs. undersampling)
- Categorical attribute **Encoding** (N, N-1, thermo, ordinal)
- Continuous attribute **Projection** (Binning vs. Normalisation)
- Continuous attribute **Scaling** ([0,+1] vs. [-1,+1] range)

### → Multifactorial design to evaluate impact across multiple methods

- Neural Networks (NN)
- Support Vector Machines (SVM)
- Decision Trees (CART)



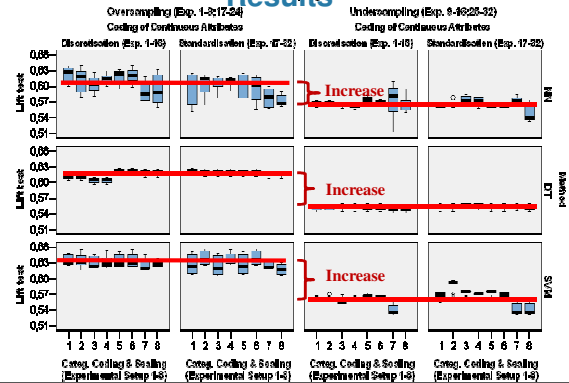
## Sampling

→ Created 2 Dataset Sampling candidates

Data subset	Data partition (number of records)			
	Oversampling		Undersampling	
	Class 1	Class -1	Class 1	Class -1
Training set	20,000	20,000	2,072	2,072
Validation set	10,000	10,000	1,035	1,035
SUM	30,000	30,000	3,107	3,107
Test (hold-out) set	912	64,088	912	64,088

→ Different balancing in the training data  
→ Original distribution in the test data (65,000 instances)

## Results



→ Oversampling outperforms undersampling consistently!  
→ Gain in Lift depends on method (different sensitivity)  
→ Oversampling has higher impact than data coding & scaling

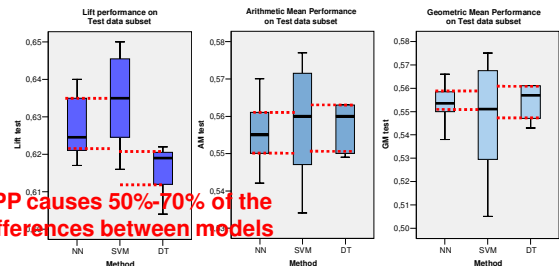
## Recommendations from Case Study

- **Sampling**
  - Oversampling outperforms undersampling for all methods
  - Undersampling: better in-sample results & worse out of sample
- **Choice of method**
  - NN & SVM better than CART
- **Encoding & Projection**
  - SVM: avoid Ordinal coding (e.g. 1,2,3) all other similar (incl. N !)
  - NN: avoid standardization & ordinal encoding
  - DT / CART: use temperature, all others similar (incl. ordinal)

→ Binning & Scaling of continuous attributes irrelevant for all methods!  
→ Use Undersampling & N-1 encoding with SVM & NN  
→ Best preprocessed SVM → lift of 0.645 on test set ... BUT ...

## Results across Pre-processing

- **Preprocessing: higher impact than method selection**
  - Lift-variation per method from Sampling/Scaling/Coding > Difference of Lift between competing methods!



→ Results are consistent across error measures  
→ Experiments allow identification of 'best practices' to model methods  
→ Best-practice preprocessing varies between methods

### Agenda

- Case study 1: Direct Marketing
  - Cross-selling of Magazine subscriptions
  - Effect of data preprocessing: Sampling
  - Interaction of Sampling with Scaling & Coding
- Case study 2: Credit & Behavioral Scoring
  - Predicting consumer credit default
  - Effects of sample size
  - Effects of sample distribution
- Case study 3: Online Shopping Behaviour
  - Predicting consumer shopping channel choice
  - Sample distribution & multiple classes
- Conclusion & Take-aways

### Business Case: Credit scoring

**“Bad” customer (uncreditworthy)**  
**Declined (credit withheld)**

**“Good” customer (creditworthy)**  
**Accepted (credit provided)**

**“Bad” customer (uncreditworthy)**    **“Good” customer (creditworthy)**

→ Definitions of ‘good’ and ‘bad’ based on repayment behavior  
 → Default, e.g. if customer is 3 months in arrears

### Sampling issues in Credit Scoring

#### Sample size

- Very large customer populations
- Millions of customer records (e.g. Barclaycard >10 mio cards & 300,000 new in 2007)
- Requires sampling to be cost & time efficient in model building

→ Draw suitably large sample to have discriminatory power

#### Sample Distribution

- Highly imbalanced datasets
- Datasets skewed to majority class of “good” customers (e.g. credit scoring from 2:1 for subprime portfolios to over 100:1 for high quality mortgages)

→ Lenders ask similar questions & use industry data sources  
 → Datasets across lenders are very homogeneous  
 → Wide acceptance of heuristic rules of thumb (Lewis 1992, Siddiqi 2005)  
 → 1500–2000 cases of each class is sufficient (incl. validation)  
 → in each class 10 \* number of predictors (Harrell, Lee et al. 1996).

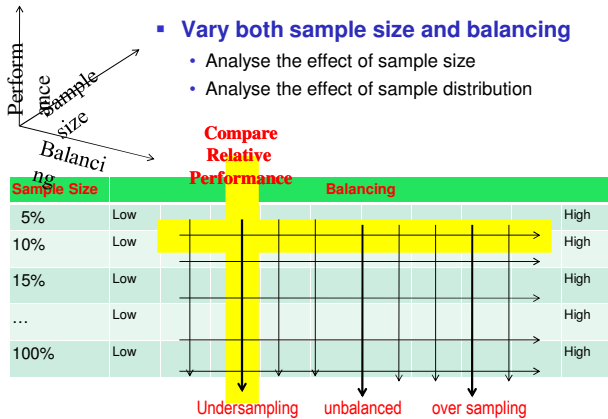
→ Use small datasets & undersampling  
 → Issues of sample size and sample distribution have been neglected

### Datasets in Literature

Study	Methods						Dataset & Samples				
	LDA	LR	NN	KNN	CART	other	# data sets	good cases	bad cases	independ. variables	
Boyle et al. 1992	X					hyb. LDA	3	1	???	139	7 to 24
Henley 1995	X	X		X	X	PP PR	6	1	???	4,132	16
Desai et al. 1997	X	X	X			GA	4	1 <sup>4</sup>	714	293	18
Arminger et al. 1997	X	X	X				3	1	1,390	1,294	21
West 2000	X	X	X	X	X	KD	6	2	360	270	24
Baesens et al. 2003	X	X	X	X	X	QDA	9	8	466	200	20
						BC			455	205	14
						SVM			1,056	264	19
						LP			2,376	264	19
									1,388	694	33
			3,555	1,438	33						
			4,680	1,560	16						
			6,240	1,560	16						
Ong et al. 2005	X	X	X			GP RS	5	2	246	306	26
									560	240	31

→ All but Baesens (2003) & Henley (1995) use small datasets → Reliability?  
 → All but Arminger (1997) use imbalanced dataset → Validity?

## Experimental Set-Up

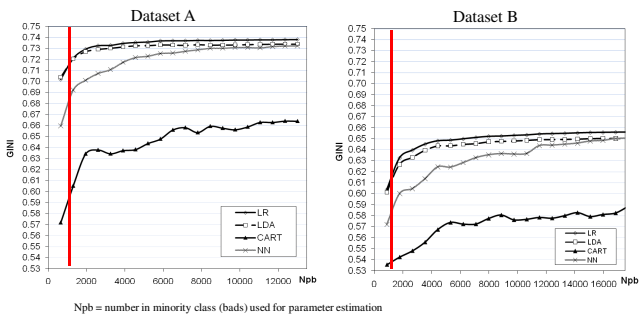


## Experimental Set-Up

- **Two Industry Data Sets**
  - A. Application scoring data set (~89K observations. ~14K bad)
  - B. Behavioural scoring data set (~121K observations. ~18K bad)
- **Four methods**
  - Logistic regression
  - Linear discriminant analysis
  - CART ~ c4.5
  - Neural networks
- **Data pre-processed using binary dummy variables.**
  - A standard practice applied to credit scoring problems
  - Preliminary stepwise procedure used for variable selection
    - 81 dummy variables for data set A.
    - 113 dummy variables for data set B.
- **Validation**
  - 50 fold cross validation for all sampling combinations

## Results of Sample Size

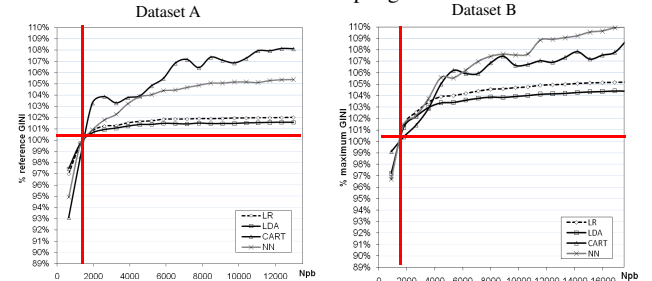
Absolute Performance - AUC measure (GINI coefficient)



- Results of sample size for Undersampling – robust across dataset A & B
- LR outperforms all methods across both datasets
  - All methods increase performance with larger samples
  - NN increases performance most with additional data (up to LR)

## Results of Sample Size

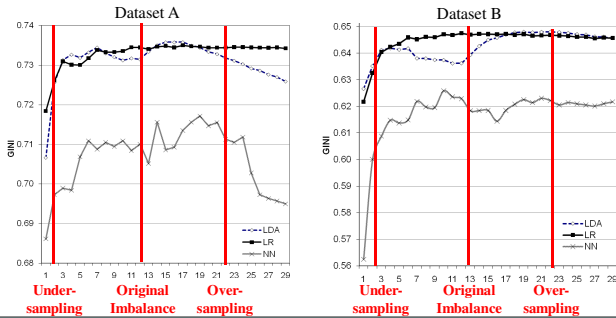
Relative Performance in % of undersampling with 1500 bads - AUC measure



- Results of sample size for Undersampling
- Performance increases of 1% to 8% through larger sample size
  - LR most robust regarding sample size

## Results of Sample Distribution

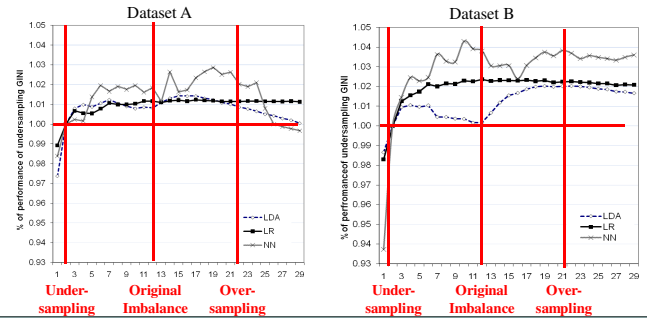
Absolute Performance - AUC measure (GINI coefficient)



- Results of Sample Distribution for small sample size (1500 bads)
- Oversampling on average outperforms undersampling
- LR and LDA outperform each other based upon distribution
- Methods show different sensitivity to sampling balances

## Results of Sample Distribution

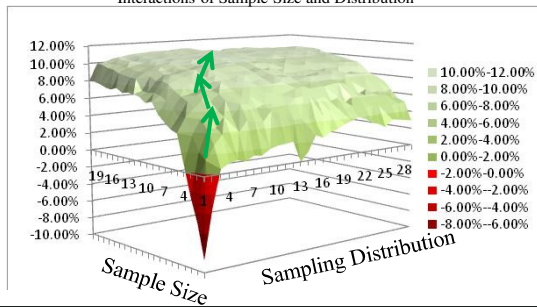
Relative Performance - AUC measure (GINI coefficient)



- Results of Sample Distribution for small sample size (1500 bads)
- Improvements of 1%-2% for LR, 1%-4% for NN feasible
- Original & Oversampling outperform Undersampling
- LDA most sensitive / LR most robust to sampling distribution

## Results - Interactions

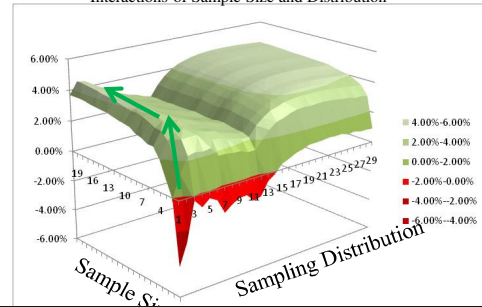
Dataset B - Neural Network  
Interactions of Sample Size and Distribution



- Results of relative performance (undersampling & 1500 bads)
- Improvements of up to 10% of NN performance possible
- Additional data more helpful than increasing (over-)sampling
- No improvement beyond oversampling

## Results - Interactions

Dataset B - Linear Discriminant Analysis  
Interactions of Sample Size and Distribution



- Interaction (base upon benchmark) varies substantially by method
- Additional data more helpful than increasing sampling
- Under- and oversampling outperform imbalanced data

## Results

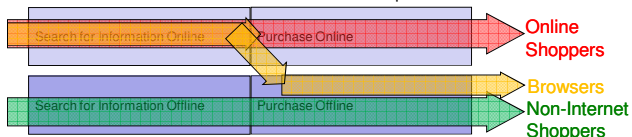
- **Sample size**
  - Taking larger samples than those commonly quoted can lead to significant performance gains.
  - >2% improvement for all methods considered
  - >5 % for CART and NNs
- **Balanced data sets better than unbalanced ones**
  - Balanced sampling outperforms imbalanced classes
  - Over sampling out performs undersampling
    - But even over sampling is not necessarily optimal
- **Some methods much more sensitive to balancing** than others
  - Logistic regression very insensitive
  - CART very sensitive

## Agenda

- Sampling issues in Data Mining
- Case study 1: Direct Marketing
  - Cross-selling of Magazine subscriptions
  - Effect of data preprocessing: Sampling
  - Interaction of Sampling with Scaling & Coding
- Case study 2: Credit & Behavioral Scoring
  - Predicting consumer credit default
  - Effects of sample size
  - Effects of sample distribution
- Case study 3: Online Shopping Behaviour
  - Predicting consumer shopping channel choice
  - Sample distribution & multiple classes
- Conclusion & Take-aways

## Business Case: Predicting Customer Online Shopping Adoption

- **Traditional buying process is offline & simultaneous** → “bricks” store
- **Introduction of the Internet changes consumer behaviour**
  - Seek information online & offline
  - Purchasing online & offline
  - Changing purchasing behaviour through internet adoption
  - Changing purchasing behaviour through Technology Acceptance
- **Development of heterogeneous Purchasing Behaviour**
  - Example: Purchasing electronic durable consumer goods
  - Search for product info (e.g. video cameras) online
    - test product in-store
    - search for best deal on internet & purchase



## Stages of Internet Adoption

**1. OFFLINE BUYERS**  
Information gathering & purchasing in Stores



**2. BROWERS**

Information gathering online & purchasing in stores



**3. ONLINE BUYERS**  
Information gathering & purchasing online



## Motivation

### DIDIER: Marketing Modelling

- Econometric / Marketing Domain
- Seeks to **explain** how customers behave in online shopping
- Use of "black-box" logistic regression models

→ Models class membership to identify causal variables that **explain** choices

→ **Descriptive & Normative Modelling**

#### Best practices

- balance datasets for distribution representative of population
- Use ordinal variables & nominal variables without recoding
- Do not normalise / scale data

### SVEN: Data Mining Perspective

- IS/OR/MS Domain → Data Mining
- Seeks to accurately **predict** regardless of explanation why customers buy
- Use of "black-box" methods from computational intelligence

→ Models class membership to accurately **classify** unseen instances

→ **Predictive Modelling**

#### Best practices

- Rebalance datasets for equal distribution of target variables
- Recode ordinal → binary scale
- Rescale & normalise data to facilitate learning speed etc.

→ same dataset & same objectives & similar methods  
 → Conflicting "best practice" approaches to modelling  
 → Outside of most software simulators!!! Implicit knowledge?  
 → ... WHO IS "CORRECT"? WHAT IS THE IMPACT?

## Dataset

### Survey on Internet Shopping Behaviour

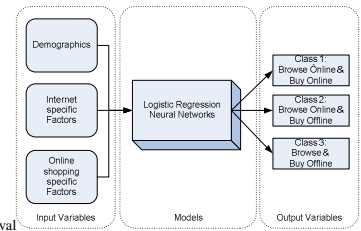
- 5500 UK households → 685 respondents
- Adjusted for age, income etc. of customers (older less likely to buy)
- Adjusted for product specific risk of online shopping for branded durable consumer goods (inspection required to some extent)
- 73 questions on factors related to internet shopping, products etc.

Online Shopping Factors:  
 "Going to the shops is as convenient as Internet shopping"  
 "I would buy online if products are branded" etc. [1=strongly agree; ...]

Demographic Factors  
 Age, Gender, Income

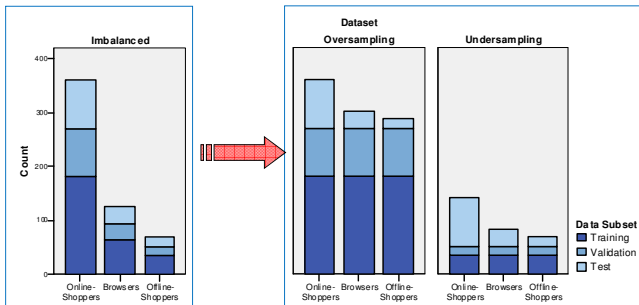
Internet Utility Factors  
 Score from 6 correlated variables

→ Mixed scale of nominal, ordinal, interval



## Imbalanced Classification problem

- Split of Dataset for Training, Validation and Test (50%;25%;25%)
- Distribution of target classes is skewed (65% online buyers; 22.5% browsers; 12.5% offline shoppers)
- Rebalancing of data sets through over- & undersampling)



## Results without Discretisation

Logist.Reg. Dataset	True Value	Training Data			Test Data			MCR <sub>train</sub>	MCR <sub>test</sub>
		Online	Browse	Offline	Online	Browse	Offline		
Original	Online	93.36	5.17	1.48	88.89	7.78	3.33	54.3%	48.9%
Imbalanced	Browser	62.77	23.40	13.83	49.39	22.58	29.03	55.8%	41.8%
	Offline	36.54	17.31	46.15	35.29	29.41	35.29	58.4%	48.2%

Neural Net Dataset	True Value	Training Data			Test Data			MCR <sub>train</sub>	MCR <sub>test</sub>
		Online	Browse	Offline	Online	Browse	Offline		
Original	Online	86.19	12.71	1.10	86.67	8.89	4.44	54.9%	35.7%
Imbalanced	Browser	53.13	31.25	15.63	41.94	25.48	22.58	54.9%	35.7%
	Offline	25.17	28.57	45.71	29.41	35.29	35.29	88.0%	75.6%

Mean Classification Rate (%)

## Results with Discretisation of Ordinal

Logist.Reg. Dataset	True Value	Training Data			Test Data			MCR <sub>train</sub>	MCR <sub>test</sub>
		Online	Browse	Offline	Online	Browse	Offline		
Original	Online	91.51	6.64	1.85	85.56	7.78	6.67	61.15%	
Imbalanced	Browse	54.26	36.12	9.57	48.39	32.26	19.35	45.1%	
	Offline	26.92	17.31	55.77	58.82	47.62	17.65		

MCR<sub>train</sub>=69.9%  
MCR<sub>test</sub>=34.4%

MCR<sub>train</sub>=66.0%  
MCR<sub>test</sub>=62.3%

Neural Net Dataset	True Value	Training Data			Test Data			MCR <sub>train</sub>	MCR <sub>test</sub>
		Online	Browse	Offline	Online	Browse	Offline		
Original	Online	96.13	3.87	0.00	84.44	11.11	4.44	56.5%	
Imbalanced	Browse	68.75	28.13	3.13	64.52	22.58	12.90	45.5%	
	Offline	40.00	14.28	45.17	58.82	11.76	29.41		

MCR<sub>train</sub>=55.2%  
MCR<sub>test</sub>=28.0%

MCR<sub>train</sub>=99.5%  
MCR<sub>test</sub>=79.0%

Mean Classification Rate (%)

## Summary

### Oversampling outperforms other samplings

- Across Different Datasets
- Across various data preprocessing

### Methods show different sensitivity to Sampling

- More variation from sampling, coding & scaling than between methods
- Using different preprocessing variants is important in modeling

### Various sophisticated extensions exist

- SMOTE (Synthetic Minority Oversampling Technique)
- K-nearest Neighbor sampling (removal / creation)
- One-class learning etc. ...

### Extend your bad of tricks ...

- ... and experiment with imbalanced sampling!

## Questions?



Sven F. Crone

Lancaster University Management School  
Centre for Forecasting  
Lancaster, LA1 4YX  
email s.crone@lancaster.ac.uk

$$SY_t = \alpha Y_{t-1} + (1 - \alpha)SY_{t-1} +$$

