Mining the Data Stream

Graham Williams

CSIRO Data Mining
GPO Box 664, Canberra 2601, Australia
datamining.csiro.au
Graham.Williams@csiro.au

Plenary Presentation to Hybrid Intelligent Systems
16 December 2003

CSIRO Data Mining

- CSIRO - Australian Government research organisation
- Applied research delivering to industry
- CSIRO Mathematical and Information Science
  Mathematicians, Statisticians, Computer Scientists
- CSIRO Data Mining established in 1994
  Machine learning, database, visualisation, statistics
- Industry
  - Government
  - Insurance
  - Fraud
  - Security
  - Science
  - Health

Data Streams

- Data Supply
- Mining The Stream

Data Streams

- Data Supply
- Mining The Stream

Data Mining

- Doing Data Mining
- Technology for Intelligent Data Mining
- Transactions to Systems

Data Streams

- Data Supply
- Mining The Stream

Data Mining

- Doing Data Mining
- Technology for Intelligent Data Mining
- Transactions to Systems

Moving Forward

- Temporal Nature of Data Streams
- Ready, Efficient, Communicating Agents

More Data Than We Know What To Do With

- Australian Health Transactions: 700 thousand per day
- Financial Transactions: 1 million per day?
- Telecom Transactions: 20 million per day?

- Health Insurance Commission: terabyte data storage
- Australian Taxation Office: 24 databases, multiple tables

Drowning in the Data Stream

Stream Data:

- Continuously becoming available
- Sequentially ordered arrival of data
- Generally arriving in quick succession
- Large amounts and perhaps limited storage
- Possibly distributed sources (e.g., sensor data)
- Foundation: database queries for data streams

Examples

- Telecommunications
- Financial transactions
- Web clicks
- Passenger arrivals
- Electronic health records
  - Infectious outbreaks - SARS
  - Adverse drug reactions - interactions
  - Insurance cost blowouts - pharmaceuticals

Data Mining

- The non-trivial extraction of novel, implicit, and actionable knowledge from large databases

Data Mining

- The non-trivial extraction of novel, implicit, and actionable knowledge from large databases and in a *timely* manner.

Data Mining

- The non-trivial extraction of novel, implicit, and actionable knowledge from large databases and in a *timely* manner.
- Anytime Data Mining - *give me the answer now*.
  - Data exploration
  - Data visualisation
  - Very large data extractions
  - Data analysis
  - Hypothesis generation
  - Model building.

Data Characteristics

- Administrative, transaction data
- 10 to 200 features
- Multiple entities
- Typical Examples:
  - Pathology (Health Insurance)
    37 million transactions (3 years of data)
    80 pathology laboratories
    20 thousand doctors
    3 million patients
    50 features
  - Queensland Linked Health Dataset
    5 years of data
    100 million (MBS), 60 million (PBS), 3 million (Hos)
    1 million patients, $50 billion
    80 features

How to Data Mine?

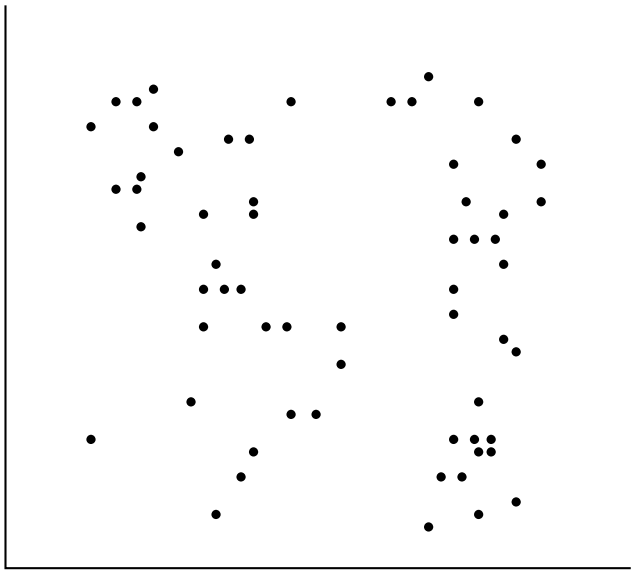
- Are we currently mining data in the best possible way?
- Using tools and techniques developed for different tasks?
- So far the track record is quite good, but . . . ?
- Should we take a fresh look at data mining . . . ?

Motor Vehicle Insurance Premium Setting

- Several million transactions annually
- Consider more than the traditional small number of factors
- Hot Spots: *Cluster* $\Rightarrow$ *Rule Induction* $\Rightarrow$ *Interestingness* (Williams, Huang, AI97)

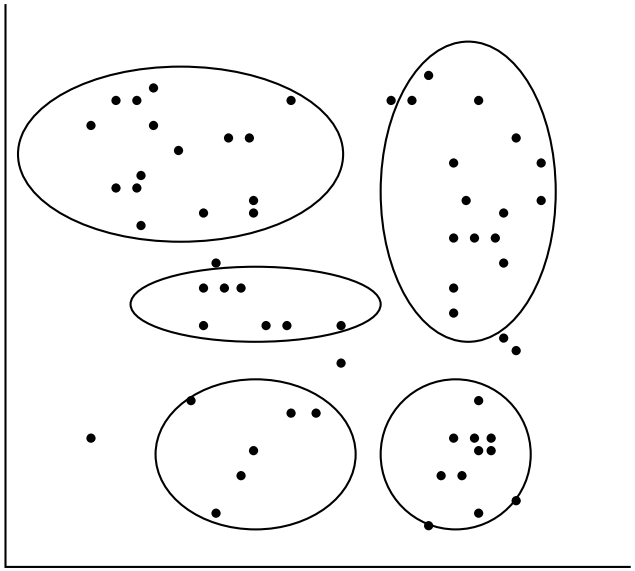
Motor Vehicle Insurance Premium Setting

- Several million transactions annually
- Consider more than the traditional small number of factors
- Hot Spots: *Cluster* $\Rightarrow$ *Rule Induction* $\Rightarrow$ *Interestingness* (Williams, Huang, AI97)



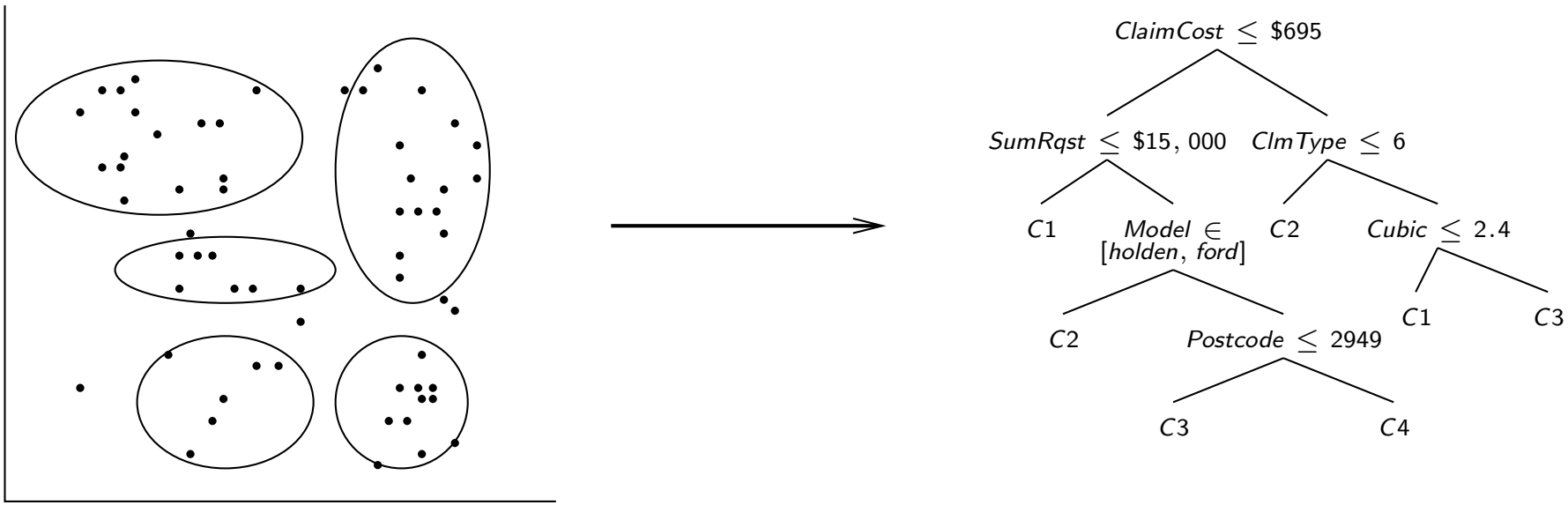
Motor Vehicle Insurance Premium Setting

- Several million transactions annually
- Consider more than the traditional small number of factors
- Hot Spots: *Cluster* $\Rightarrow$ *Rule Induction* $\Rightarrow$ *Interestingness* (Williams, Huang, AI97)



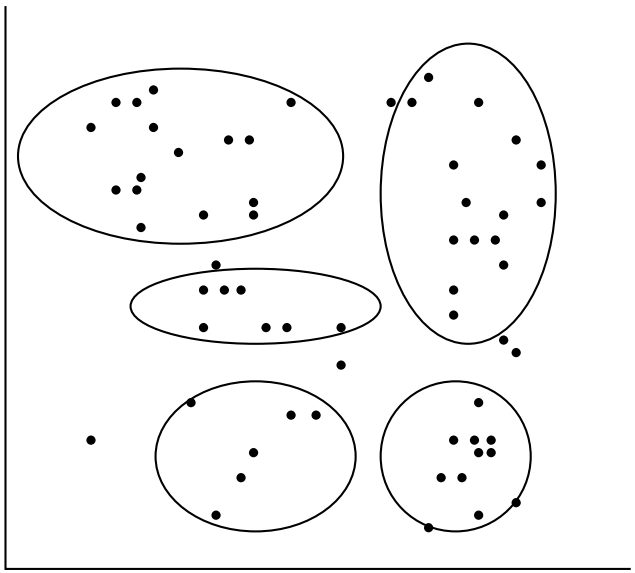
Motor Vehicle Insurance Premium Setting

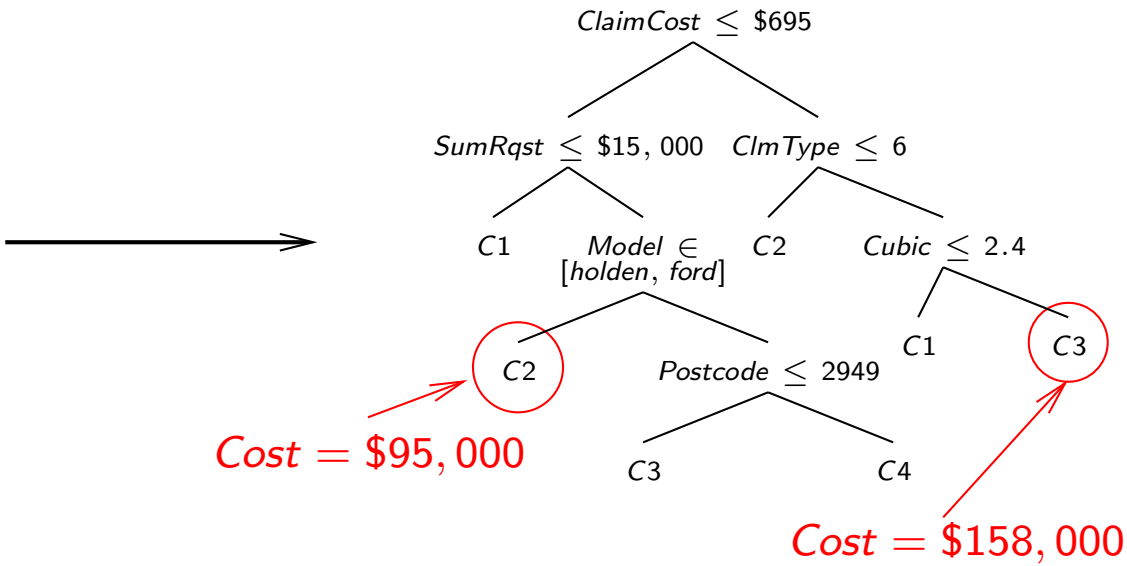
- Several million transactions annually
- Consider more than the traditional small number of factors
- Hot Spots: *Cluster* $\Rightarrow$ *Rule Induction* $\Rightarrow$ *Interestingness* (Williams, Huang, AI97)



Motor Vehicle Insurance Premium Setting

- Several million transactions annually
- Consider more than the traditional small number of factors
- Hot Spots: *Cluster* $\Rightarrow$ *Rule Induction* $\Rightarrow$ *Interestingness* (Williams, Huang, AI97)



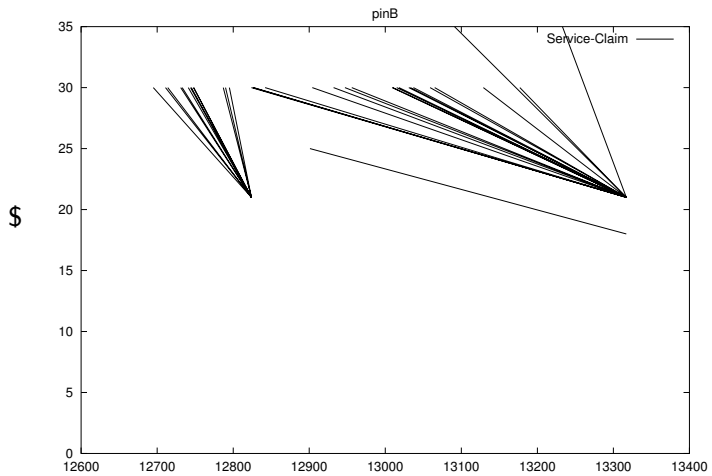


Medicare *Health Insurance Commission*

- Universal health coverage since 1975
- Terabytes of patient claims in storage in Canberra
- Inappropriate Provider Practices (neural networks)
- *Public Fraud* (including doctor shoppers)
- Hot Spots: *Cluster* $\Rightarrow$ *Rule Induction* $\Rightarrow$ *Interestingness*

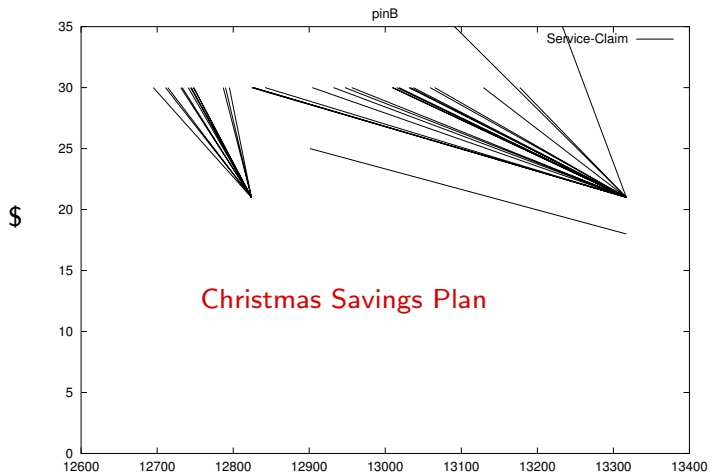
Medicare *Health Insurance Commission*

- Universal health coverage since 1975
- Terabytes of patient claims in storage in Canberra
- Inappropriate Provider Practices (neural networks)
- *Public Fraud* (including doctor shoppers)
- Hot Spots: *Cluster* ⇒ *Rule Induction* ⇒ *Interestingness*



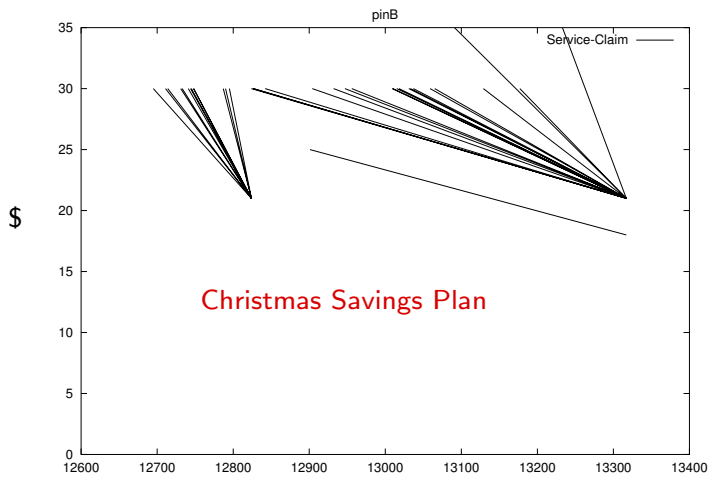
Medicare *Health Insurance Commission*

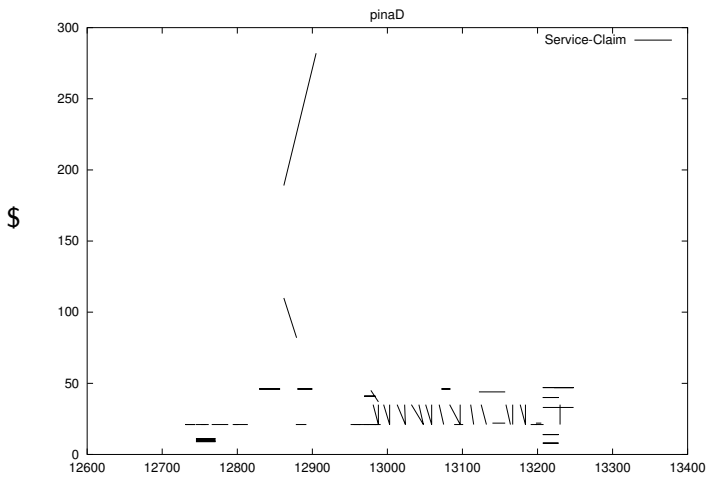
- Universal health coverage since 1975
- Terabytes of patient claims in storage in Canberra
- Inappropriate Provider Practices (neural networks)
- *Public Fraud* (including doctor shoppers)
- Hot Spots: *Cluster* ⇒ *Rule Induction* ⇒ *Interestingness*



Medicare *Health Insurance Commission*

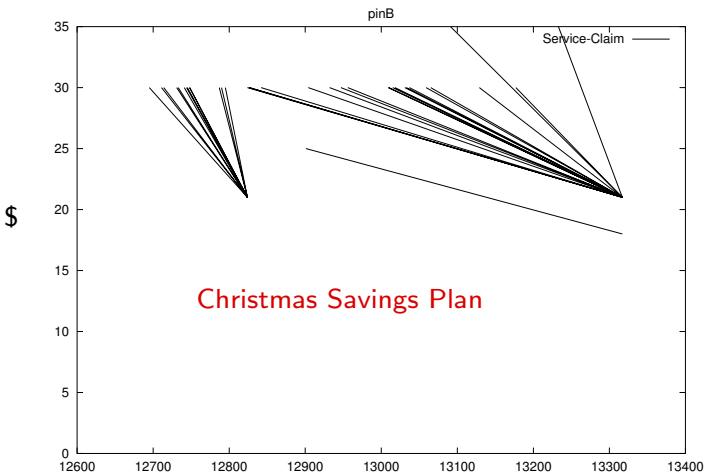
- Universal health coverage since 1975
- Terabytes of patient claims in storage in Canberra
- Inappropriate Provider Practices (neural networks)
- *Public Fraud* (including doctor shoppers)
- Hot Spots: *Cluster* ⇒ *Rule Induction* ⇒ *Interestingness*

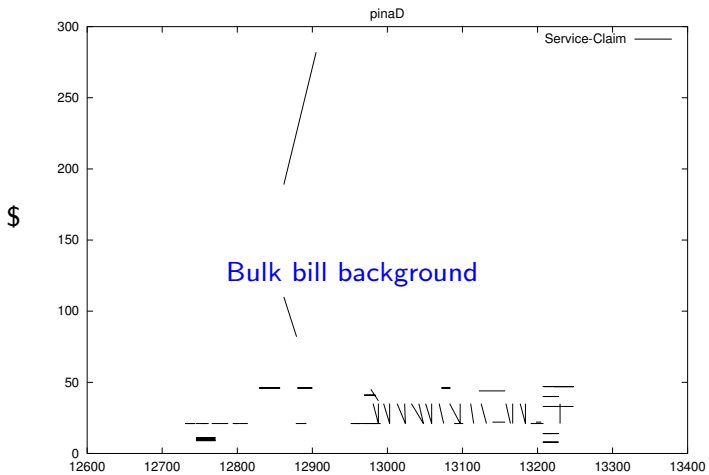




Medicare *Health Insurance Commission*

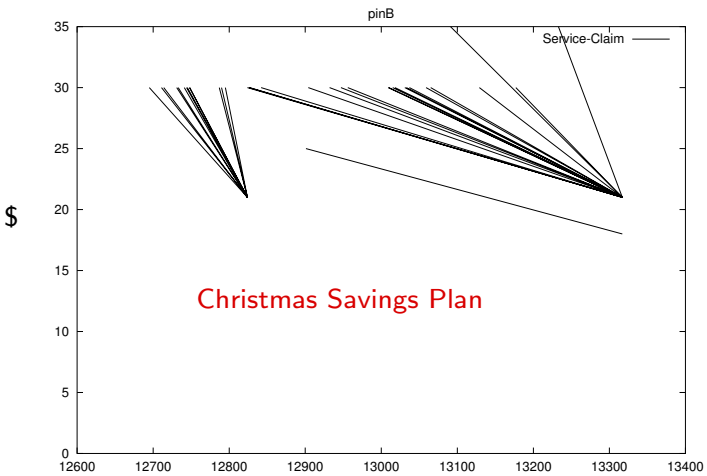
- Universal health coverage since 1975
- Terabytes of patient claims in storage in Canberra
- Inappropriate Provider Practices (neural networks)
- *Public Fraud* (including doctor shoppers)
- Hot Spots: *Cluster* ⇒ *Rule Induction* ⇒ *Interestingness*

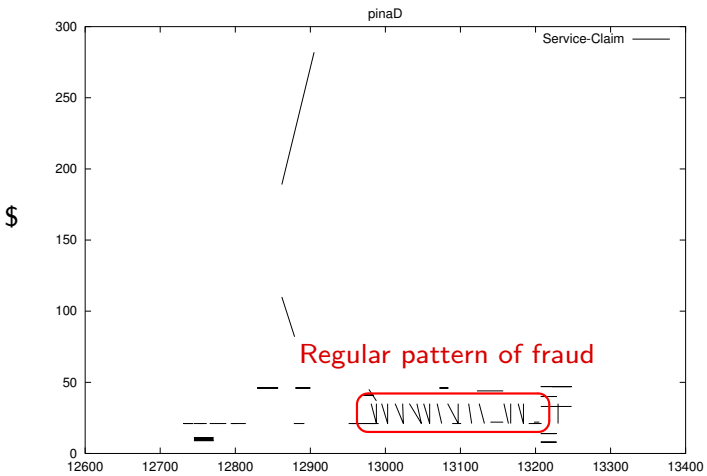




Medicare *Health Insurance Commission*

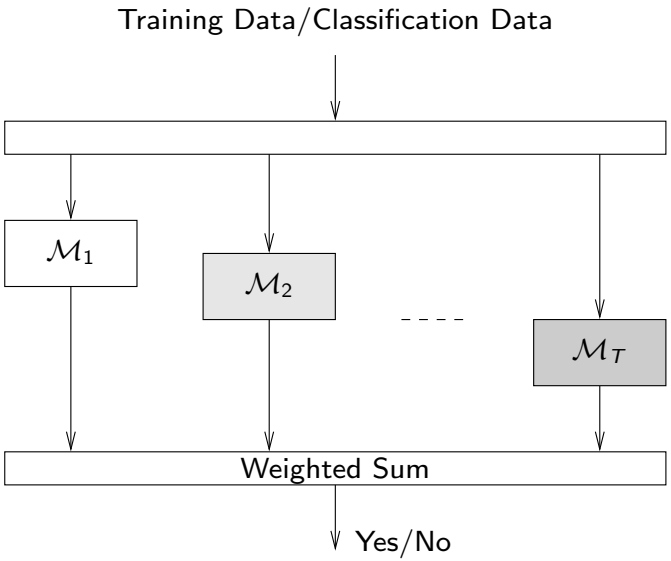
- Universal health coverage since 1975
- Terabytes of patient claims in storage in Canberra
- Inappropriate Provider Practices (neural networks)
- *Public Fraud* (including doctor shoppers)
- Hot Spots: *Cluster* ⇒ *Rule Induction* ⇒ *Interestingness*





Internal Fraud and Compliance

- Investigation of life factors as determinants of compliance
- *Boosted Tree Stumps*
  (Bartlett, Baxter, Milne, Williams, 1999)



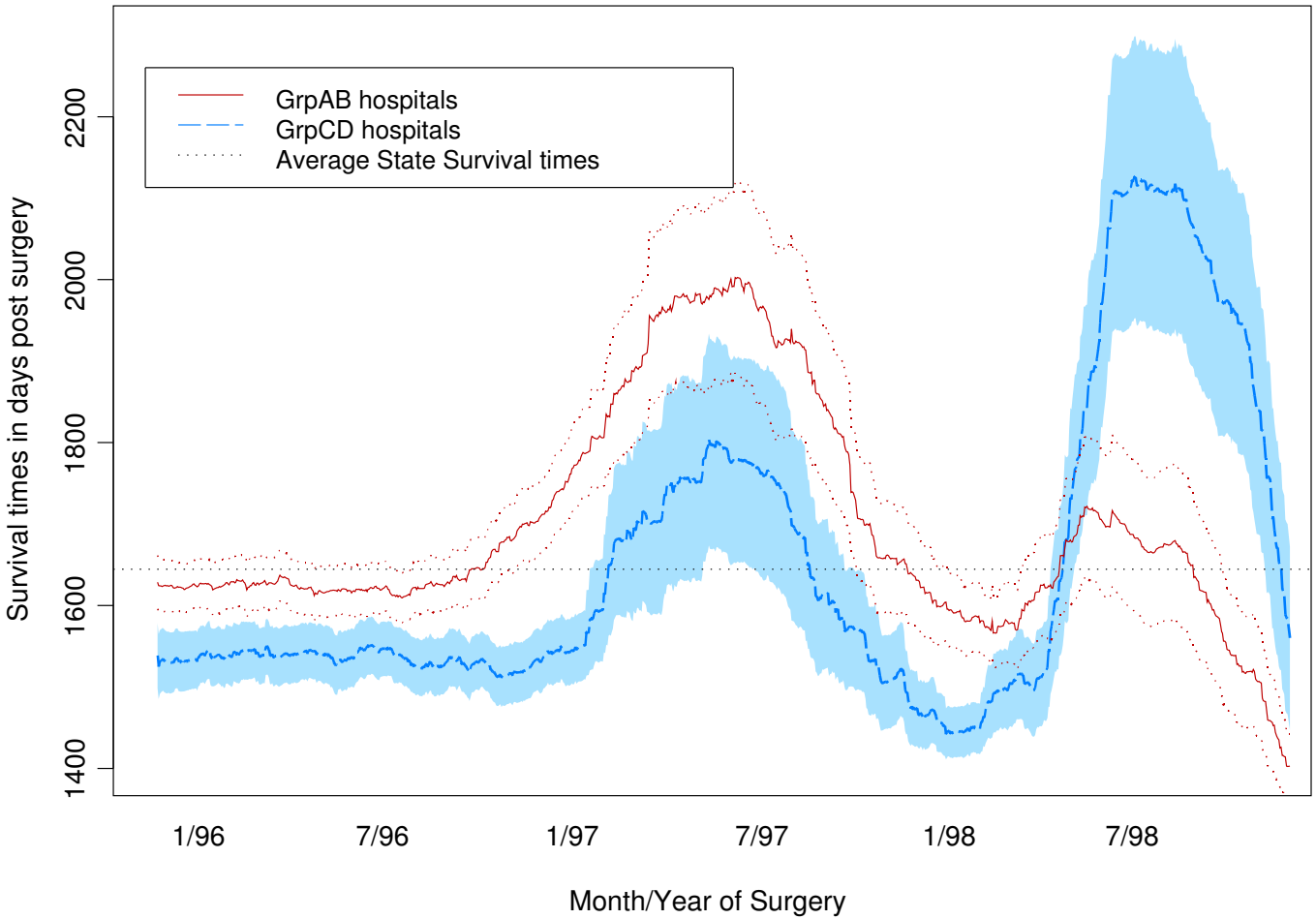
If Age $>$ 40 *Then Non Compliant (0.3)*

If MStatus Separated Then Non Compliant (0.8)

- Multiple agents
- Simple agents
- Working together!

Australian Department of Health and Ageing

Variation in Clinical Outcomes



Evolving Interesting Rules

- HotSpots searches for *interesting* regions described by rules
- So, evolve $\mathcal{R}$ using *interestingness* as measure of fitness
- Unlike rule induction, accuracy and coverage are not the goal
- *Interestingness* components: size, statistical profile, density.

Revolver

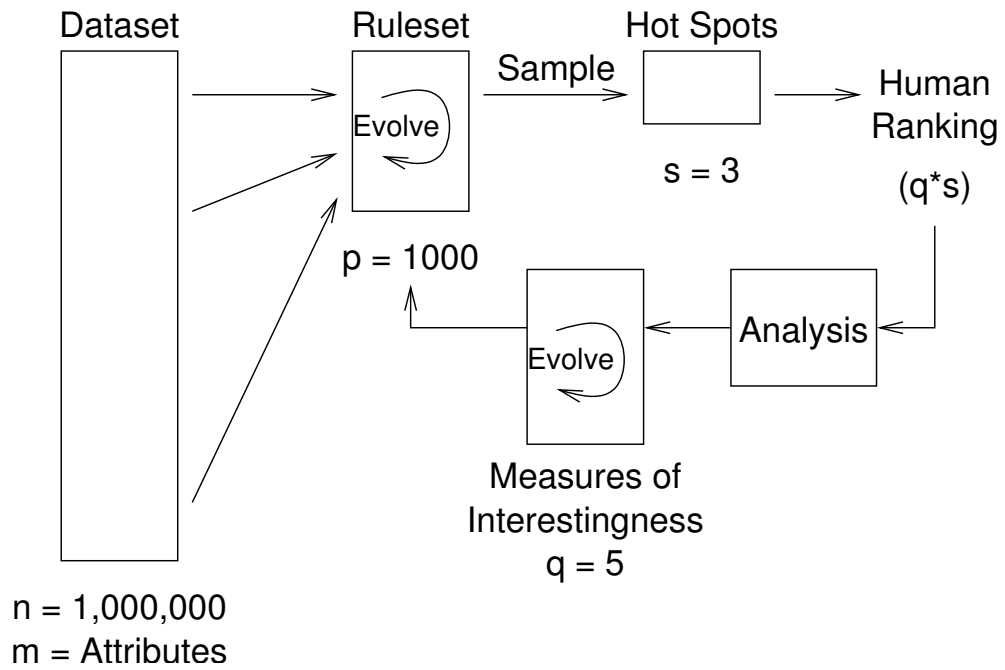
- Crossover, mutation, islands, migration strategies
- Fitness:

$$fit(r_i) = \frac{Max\{\mu_{r_i}(A_3), \mu_{r_i}(A_{12}), \mu_{r_i}(A_{43})\}}{Max\{\mu(A_3), \mu(A_{12}), \mu(A_{43}\}} + \frac{\Omega_{r_i}(A_{65} = True)}{\Omega_{r_i}} + \frac{\Omega_{r_i}(A_{39} = Yes)}{\Omega_{r_i}} + \frac{\mu_{r_i}(A_{54})}{\mu(A_{54})} + \frac{1}{\chi(r_i)}$$

$\mu(Att)$	mean value of the attribute over data set
$\mu_{r_i}(Att)$	mean value over the subset r_i
Ω_{r_i}	number of records in r_i
$\Omega_{r_i}(Cond)$	number of records in r_i satisfying $Cond$
$\chi(r_i)$	number of conditions in the rule r_i

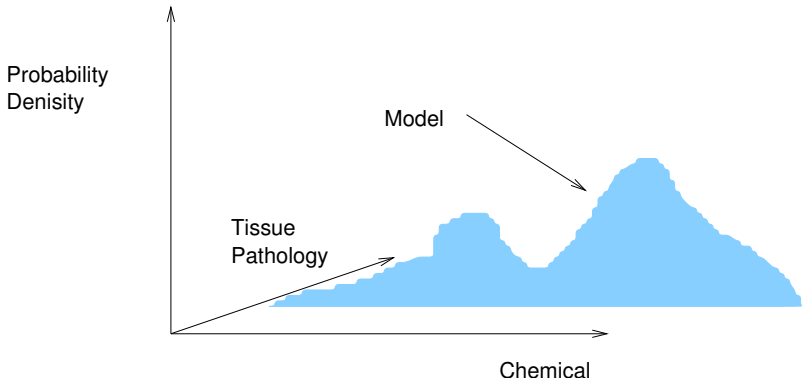
Evolving Interestingness

- Interestingness difficult to formulate!
- Evolve measure of interestingness with fitness guided by expert selection:



Emergent Outliers

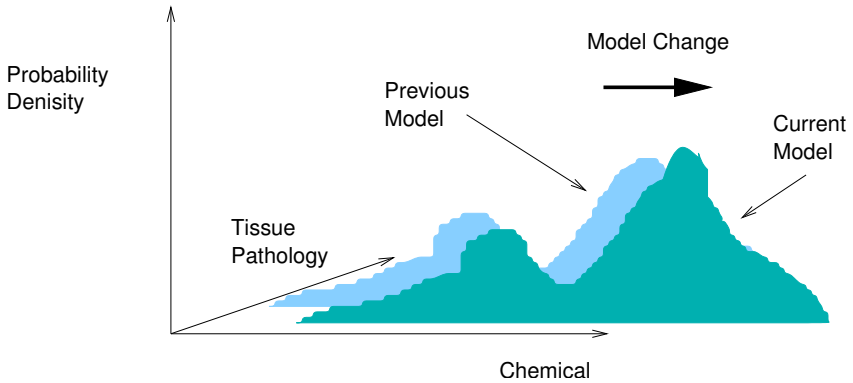
- Statistical Outliers with SmartSifter (Yamanishi, Takeuchi, Williams, Milne, 2002, 2004).



- Represent patient patterns using a statistical model
- On-line unsupervised learning of the model each time a new transaction is input using a discounting learning algorithm
- Score the transaction according to how much it changes the model.

Emergent Outliers

- Statistical Outliers with SmartSifter (Yamanishi, Takeuchi, Williams, Milne, 2002, 2004).



- Represent patient patterns using a statistical model
- On-line unsupervised learning of the model each time a new transaction is input using a discounting learning algorithm
- Score the transaction according to how much it changes the model.

Common Characteristics

- Transaction data, administrative data
- Patterns emerge as the data streams in
- Simple, conceptually independent, models
- Identify interesting patterns at any time, as they emerge.

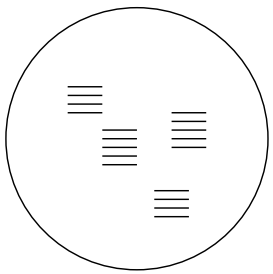
Transactions → Entities → System

- Health Insurance: Typical transaction processing
- Desire to identify fraud as it emerges
- Levels of checking:
  - Transaction level: claim matches age and gender
  - Patient level: childbirth less than 6 months apart
  - System level: collusion between patients and doctor

Transactions → Entities → System

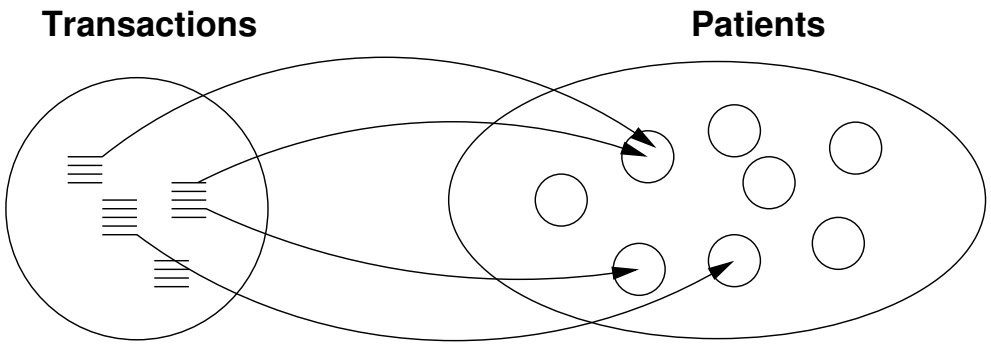
- Health Insurance: Typical transaction processing
- Desire to identify fraud as it emerges
- Levels of checking:
  - Transaction level: claim matches age and gender
  - Patient level: childbirth less than 6 months apart
  - System level: collusion between patients and doctor

Transactions



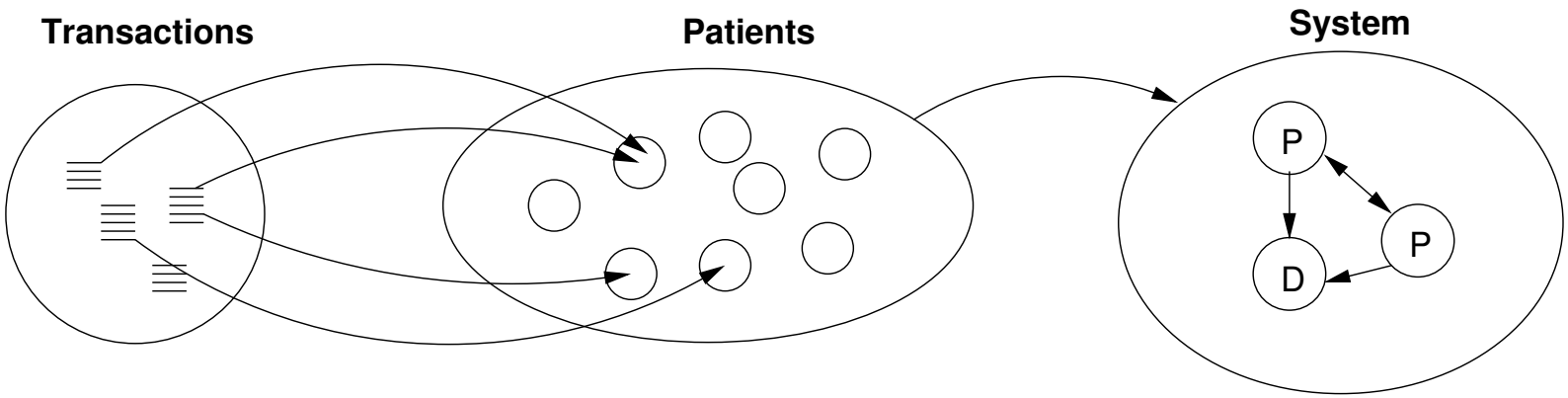
Transactions → Entities → System

- Health Insurance: Typical transaction processing
- Desire to identify fraud as it emerges
- Levels of checking:
  - Transaction level: claim matches age and gender
  - Patient level: childbirth less than 6 months apart
  - System level: collusion between patients and doctor



Transactions → Entities → System

- Health Insurance: Typical transaction processing
- Desire to identify fraud as it emerges
- Levels of checking:
  - Transaction level: claim matches age and gender
  - Patient level: childbirth less than 6 months apart
  - System level: collusion between patients and doctor



Challenges for Data Mining

For fraud, security, improving business processes:

- Source data generally arrives as transactions
- Thing of interest occurs at the entity and system level
- Thing of interest is usually not self revealing
- Patterns of interest are constantly changing
- Patterns of interest are generally unusual
- Service guarantee: rapid turnaround of service

Challenges for Data Mining

For fraud, security, improving business processes:

- Source data generally arrives as transactions
- Thing of interest occurs at the entity and system level
- Thing of interest is usually not self revealing
- Patterns of interest are constantly changing
- Patterns of interest are generally unusual
- Service guarantee: rapid turnaround of service
- Delivering results as soon as possible

Temporal Nature of Data

- Adverse Drug Reactions
  Ace Inhibitor Usage $\longrightarrow$ Angioedema
- Ambulatory Care Sensitive Conditions
  Improved Primary Care $\longrightarrow$ Decrease Hospitalisation
- Hormone Replacement Therapy
  Increased Risk of Breast Cancer
  Decreased Risk of Colorectal Cancer

Hybrid Intelligent Systems

- Data mining systems based around agents
- Agents actively pursuing the streaming data
- Building their own models, monitoring change, communicating discoveries

Mining the Data Stream

- Intelligent data mining is fundamentally hybrid
- Streaming data, emergent patterns, dynamic monitoring
- Streams are fundamentally temporal in nature
- Requirement for good enough answers, quickly
- Foundational developments in data mining are there
- Goal is a population of simple, communicating, data mining agents, monitoring data streams for trigger events, which need to be brought to the attention of the analysts.

Mining the Data Stream

- Intelligent data mining is fundamentally hybrid
- Streaming data, emergent patterns, dynamic monitoring
- Streams are fundamentally temporal in nature
- Requirement for good enough answers, quickly
- Foundational developments in data mining are there
- Goal is a population of simple, communicating, data mining agents, monitoring data streams for trigger events, which need to be brought to the attention of the analysts.

THANK YOU

References

- **On-line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms.** Data Mining and Knowledge Discovery, Volume 8, 2004.
- **Temporal Sequence Associations for Rare Events.** (submitted)
- **A Framework for Comparison in Temporal Sequence Clustering.** (submitted)
- **Temporal Event Mining of Linked Medical Claims Data.** (DMAK-2003)
- **Evolutionary Hot Spots Data Mining: An architecture for exploring for interesting discoveries.** In Methodologies for Knowledge Discovery and Data Mining. Lecture Notes in Artificial Intelligence, Volume 1574, Springer-Verlag, 1999, (PAKDD-99).
- **Combining decision trees: Initial results from the MIL algorithm.** Artificial Intelligence Developments and Applications, Elsevier Science Publishers, Pages 273–289.