End to End Data Mining: The Next Challenge

Frans Coenen

Department of Computer Science

The University of Liverpool

Liverpool L69 3BX

Tuesday 21 April 2009

ASC, Imperial

Presentation Overview

- Motivation ("Where I'm coming from").
- Some specific applications.
- A generic application (but with lots of different elements).
- Multi-Agent Data Mining (MADM), a potential solution.
- Conclusions.

Motivation

Applied data mining.

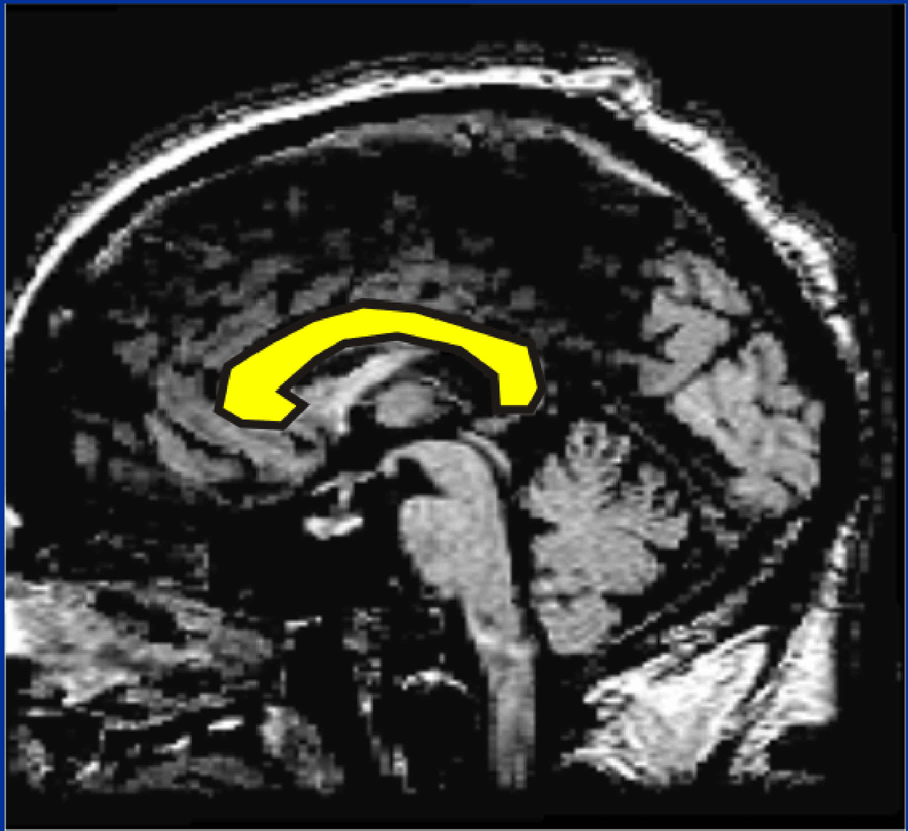
- As a community we have produced a rich and successful range of data mining tools and techniques.
- However, many applications of our knowledge provide new and interesting challenges, often unique to the application under consideration.
- The main issue is the process of putting all the constituent parts together to address a given real world data mining task, i.e. the **end to end data mining** process.
- This presentation focuses on a number of sample real-world applications so as to highlight the challenge, and then presents a potential solution.

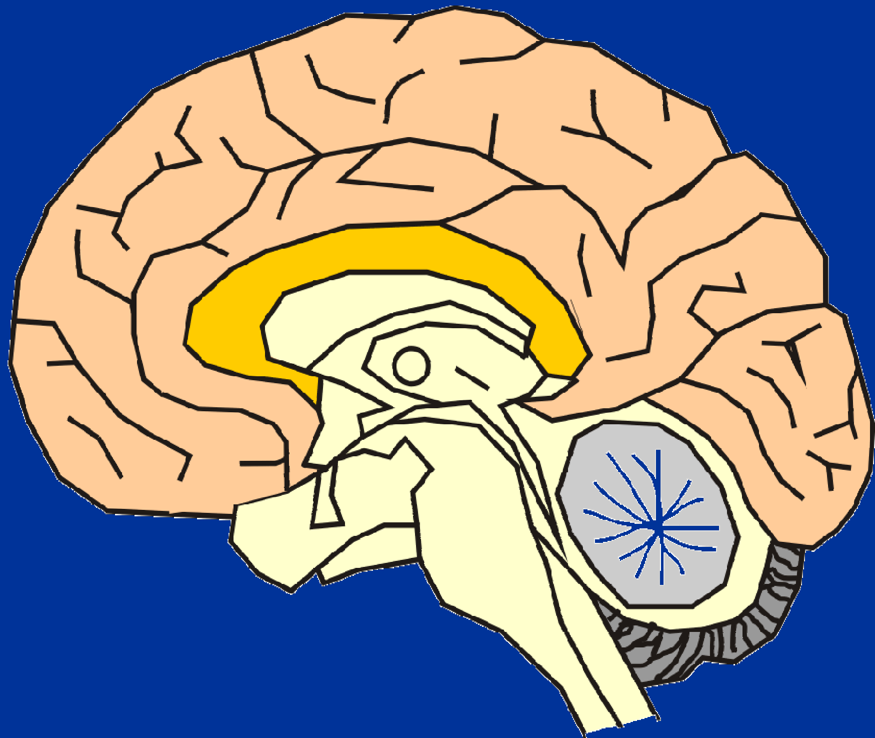
Mining MRI Scan Data

- (A particular problem example.)
- We wish to classify MRI scan data collections, for medical research purposes, according to a particular feature in these scans called the Corpus Callosum (CC)
- The conjecture is that the shape and size of the CC serves to distinguish, for example, musicians and non-musicians. It is also suspected that the shape and size of the CC plays a role in the identification of medical conditions such as epilepsy, schitsophrenia, autism, etc. The size and shape is also effected by age.

Mining MRI Scan Data

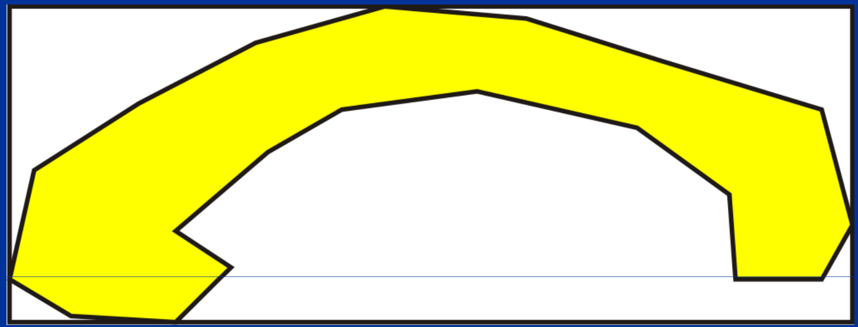
- Example images:





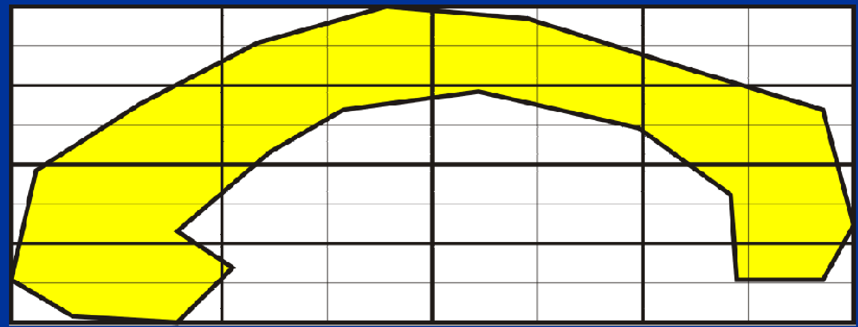
Mining MRI Scan Data

(1)



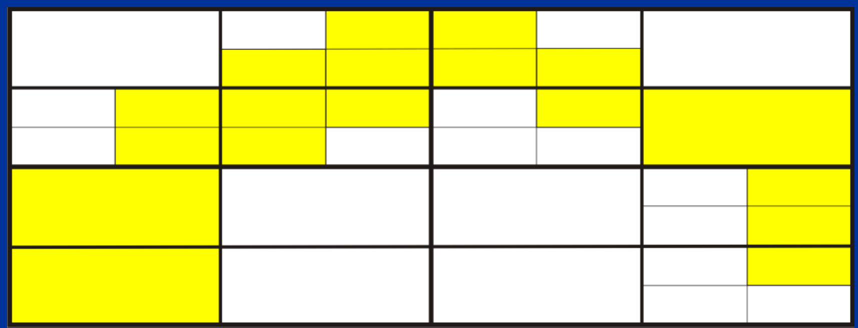
Minimum Bounding Rectangle (MBR)

(2)



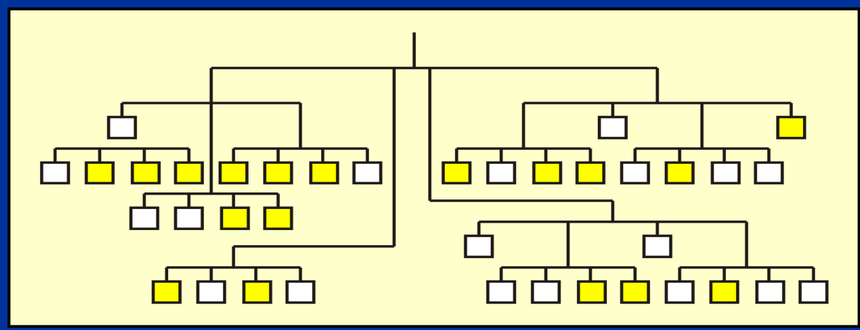
Tesselate

(3)



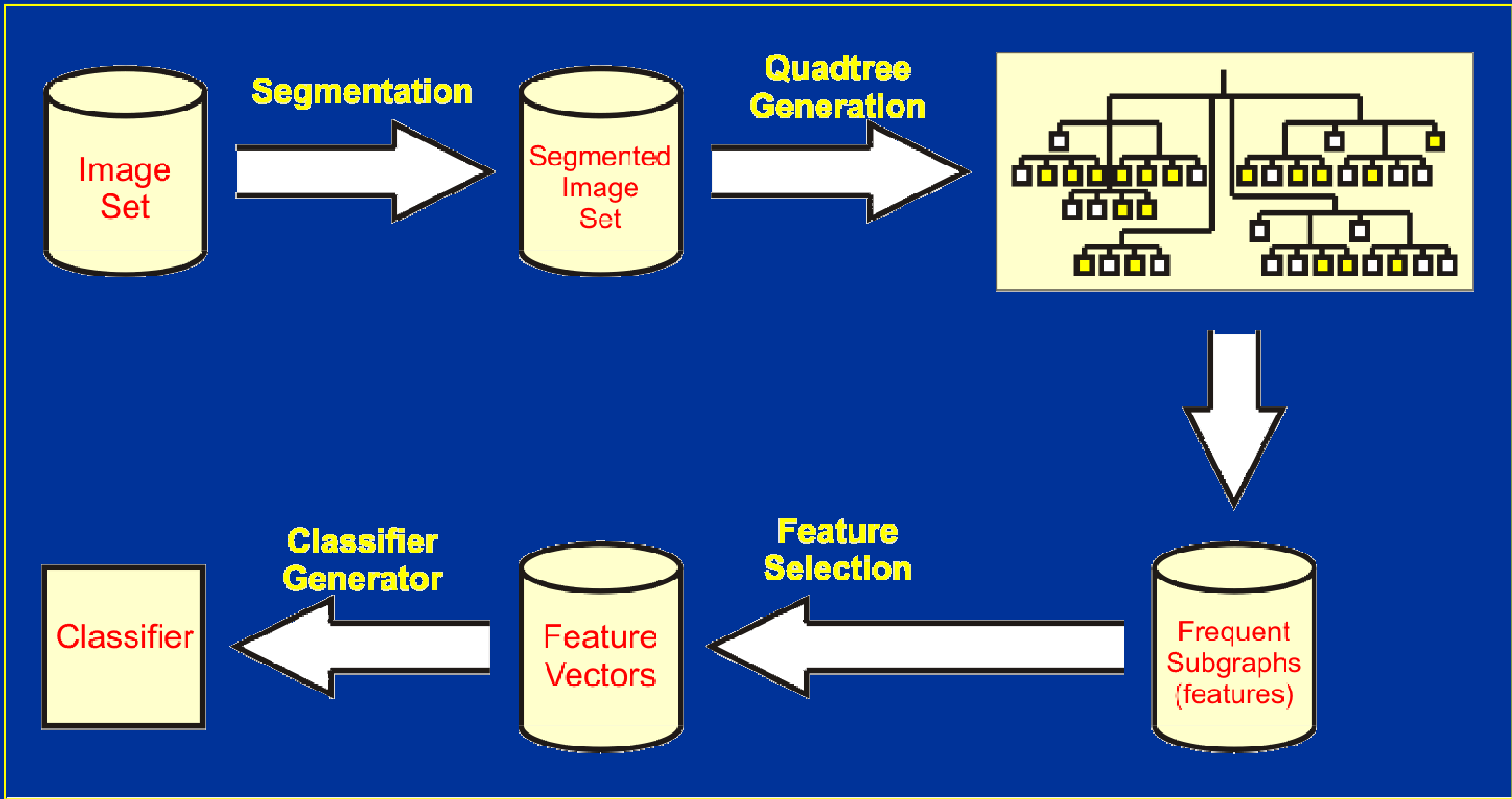
ID Colour Blocks

(4)



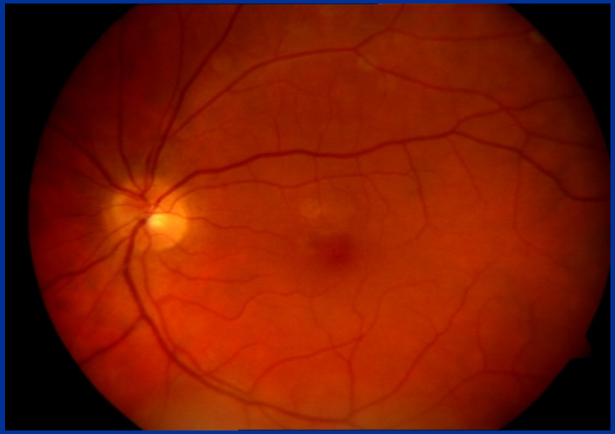
Quad Tree

Mining MRI Scan Data

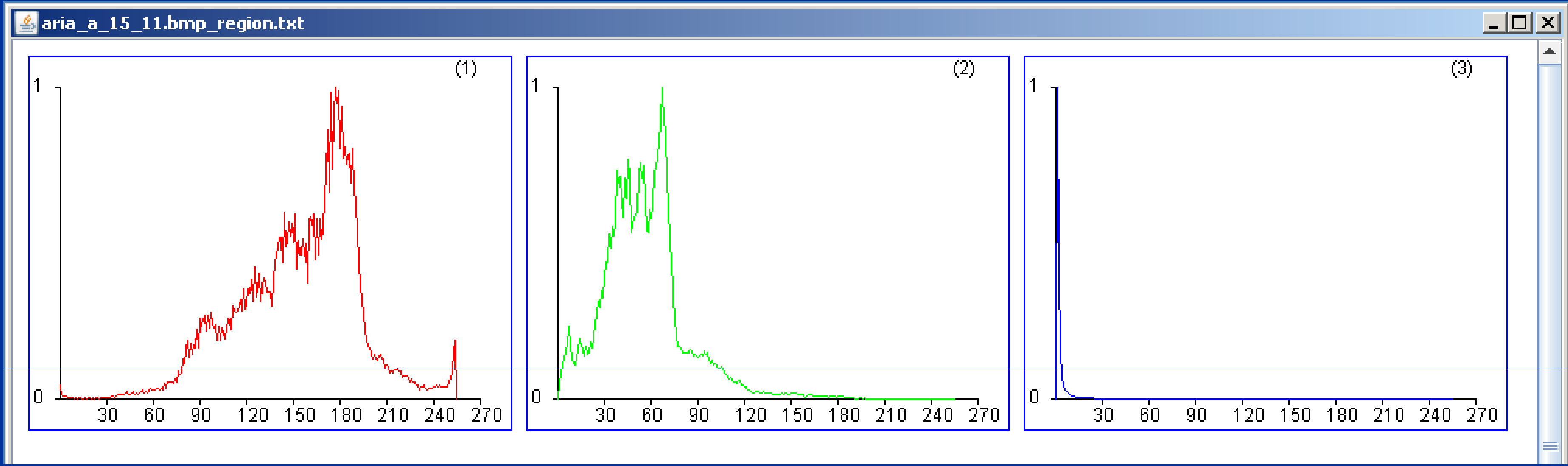


AMD (<u>A</u>ge related <u>M</u>acular <u>D</u>egeneration) Example (1)

- (Another specific problem example.)
- We wish to provide screening support for the early diagnosis of AMD.
- A common (standard) mechanism for doing this is by identifying "drusen" in retina scans.



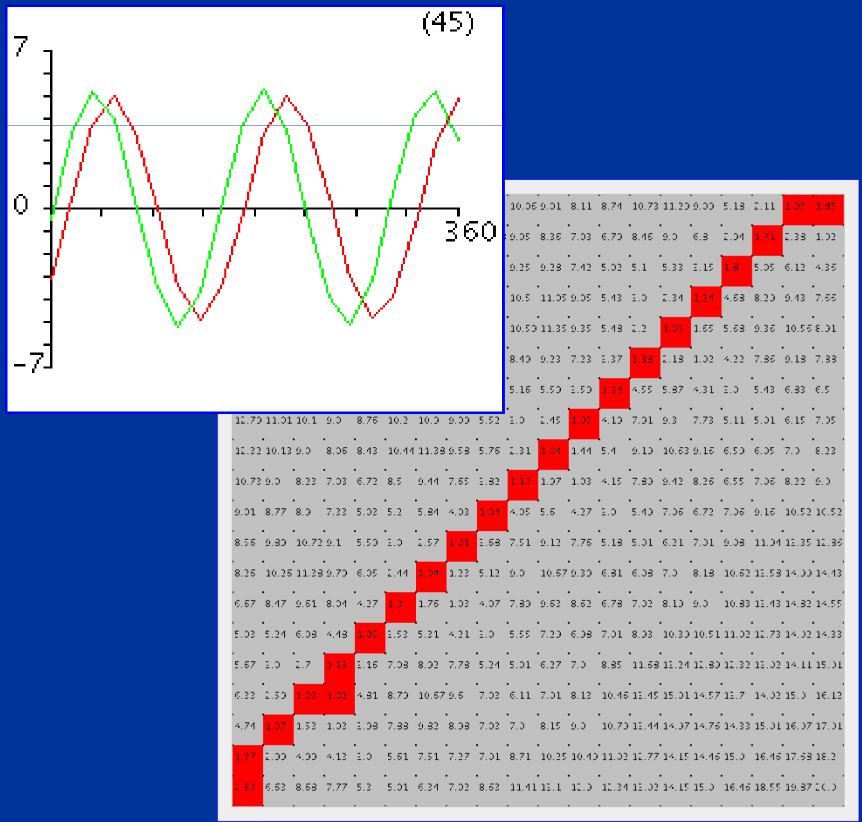
AMD Example (2)

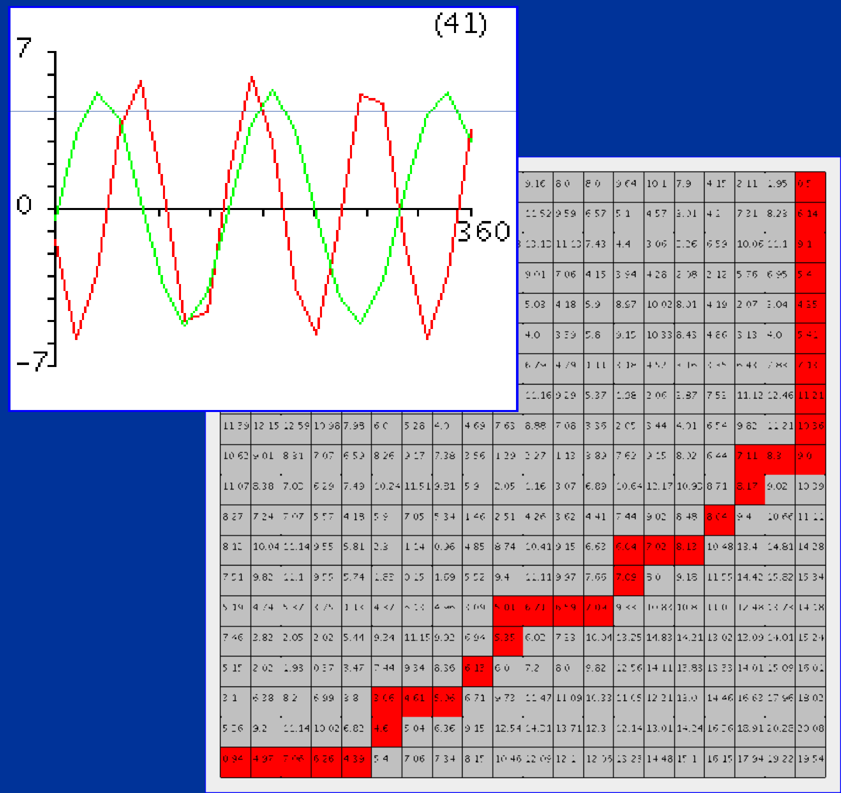


- Histogram approach
- Histogram is in effect a time series.
- Consequently we can use time series analysis techniques, in this case dynamic time warping.

AMD Example (3)

- Dynamic Time Warping (DTW).
- Case base of known time series which new case is compared to.





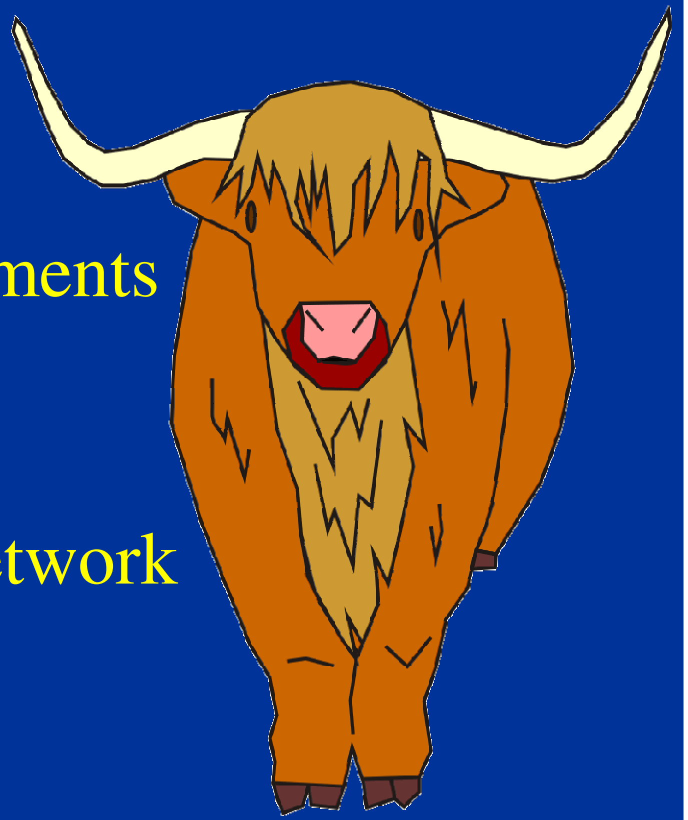
Trend Mining

- (A more generic problem example.)
- Many institutions and commercial enterprises are interested in trends.
- The technique adopting (in various forms) is emerging and/or jumping pattern (EPs and JPs) mining.
- This is an extension of established Association Rule Mining (ARM) technology that looks at how the significance (support) of identified patterns (itemsets) changes over time.
- Number of collaborations in this area.

Trend Mining in Customer Bases

- Particular case is in collaboration with a freight forwarder who wish to identify groups of customers (may be very small groups) whose behaviour changes.
- Patterns here are made of attributes from the customer base: location of sender, destination, weight, size, price, route, etc. Data all requires pre-processing.
- Once emerging/jumping patterns have been identified need to trace patterns back to customer IDs.

Trend Mining in Social Networks



- Particular case is The UKs cattle movement DB.
- Large DB recording all cattle movements between locations in the UK (administered by DEFRA).
- Represents a time stamped social network (social network mining).
- Using the EP and JP idea to identify changes in behaviour.
- Aim is to determine the effect that changes in government policy and working practices might have (or not have).

Trend Mining in Medical Applications

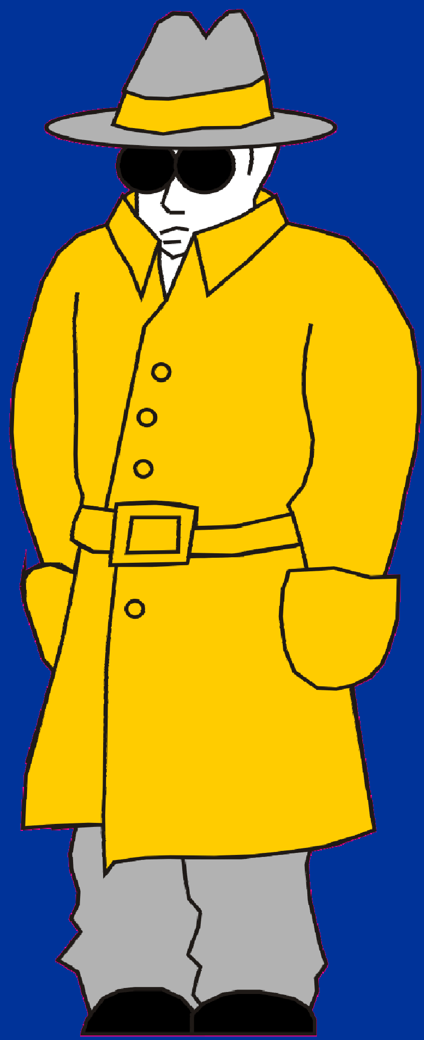
- Longitudinal data sets are common in medical applications (patient records).
- Work with diabetes unit at The Royal Liverpool Hospital.
- Royal Liverpool Hospital has the largest collection of diabetes data records in the UK (actually four DBs).
- Patients have regular consultations.
- Problems with: (a) missing data, (b) heterogeneity
- Interested in changes in patient data (but lack of change is also interesting).

Trend Mining in Web Usage Mining

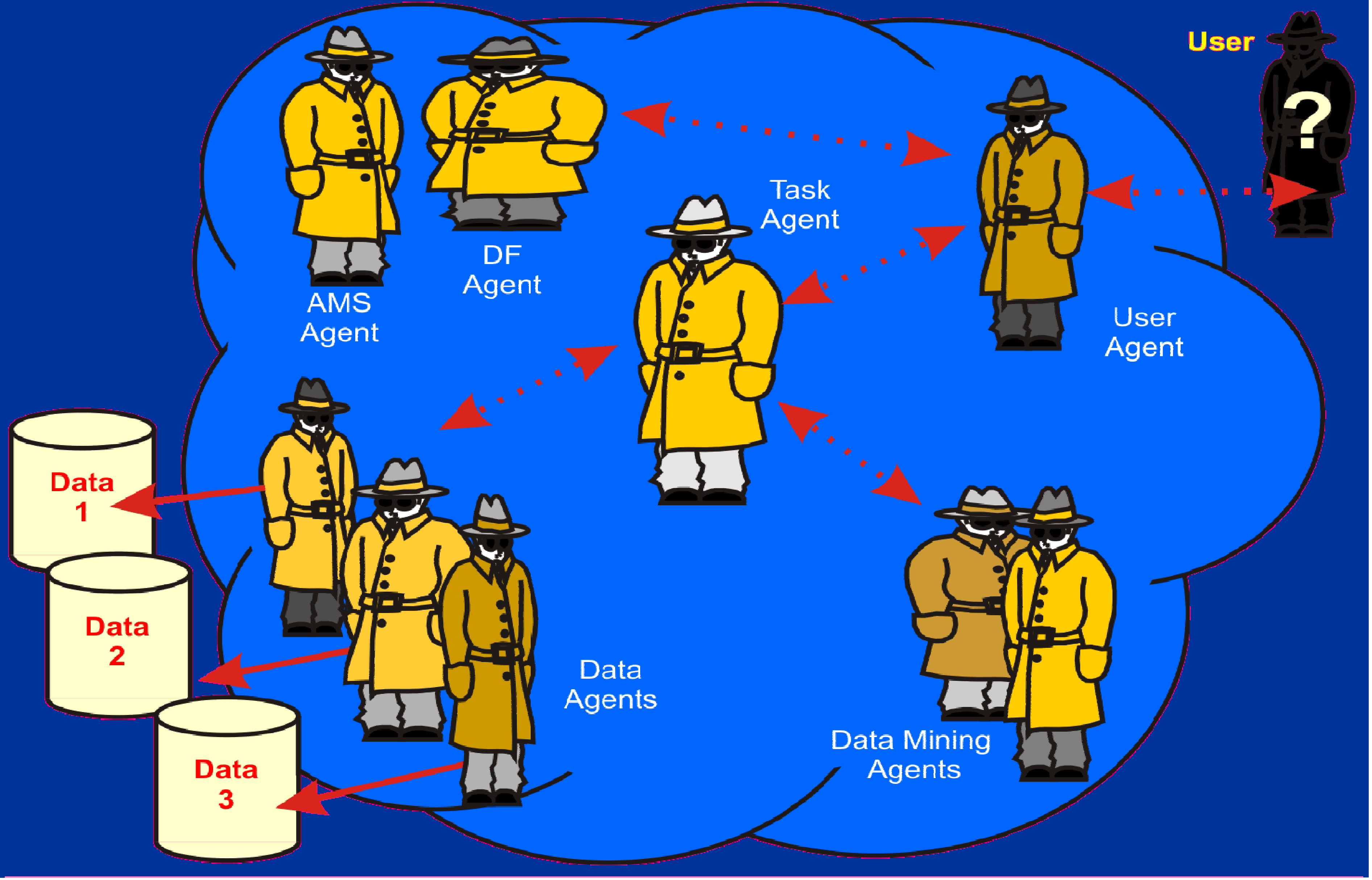
- Web usage mining is a popular KDD application.
- Learn Higher initiative.
- We wish to identify changes in WWW site usage behaviour.
- This is expected to provide information which will in turn provide evidence for restructuring of the site.
- Input is WWW log data time stamped at weekly intervals.

Multi-Agent Data Mining (1)

- A potential generic solution is using a MAS (Multi-Agent System) approach.
- Vision is that of an anarchic collection of software agents; contributed by various participants, and cooperating to address a rich range of KDD tasks.
- The challenge is that the technical domain of KDD and (as already illustrated) the variety of applications is extensive.
- Propose EMADS, The Extendible Multi-Agent Data Mining System.

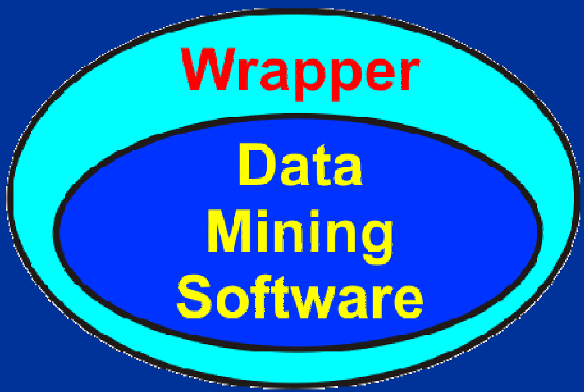


Multi-Agent Data Mining (2)



Multi-Agent Data Mining (3)

- EMADS supports extendibility through a number of predefine generic (**Data** and **DM**) wrappers.
- Wrappers are in effect EMADS agents in there own right that merge with whatever they are used to wrap to become data mining or data agents.
- Data mining wrappers require some programming knowledge.
- In case of Data wrappers, usage is facilitated by a GUI.
- (Creation of task agents requires more extensive knowledge, but not excessively so.)



Summary

- Motivation ("Where I'm coming from").
- Some specific applications: MRI scan and Retina image mining.
- A generic application (but with lots of different elements): Freight forwarding, cattle movement social network, longitudinal patient data sets and WWW usage mining.
- A potential solution: Multi-Agent Data Mining (MADM).


Questions!
Frans Coenen

Credits

- **MRI Scan Inage Mining**: Ashraf El Sayed, Martha van der Hoek[1], Vanessa Slumming[2], Chuntao (Geof) Jiang.
- **AMD**: Hanafi Hijazi, Yalinn Zhang[3].
- **Trend Mining**:

  Freight Forwarding Customer Base: Reshma Patel[4], Lawson Archer[4].

  Cattle Movement: Puteri Nohuddin, Christian Setzkorn[5], Bob Christie[5], Suzy Robinson[5].

  Patient data: Vassiliki Somaraki, Simon Harding[3], Deborah Broadbent[3].

  Learn Higher: Mohammad Khan[6], David Read[6].
- **EMADS**: Kamal Ali Albashiria, Paul Leng, Santhana Chaimontree, Katie Atkinson.

[1]UoL Dept. Medical Statistics, [2]UoL Dept. Public Health, [3]Royal Liverpool Hospital, [4]Transglobal Express Ltd., [5]UoL Vet school, [6]Liverpool Hope University. All other contributors are from the Department of Computer Science at The University of Liverpool