Data Quality Aspects in Data Mining, Data Linkage and Geocoding

Peter Christen

Department of Computer Science,
Faculty of Engineering and Information Technology,
ANU College of Engineering and Computer Science,
The Australian National University

Contact: **peter.christen@anu.edu.au**

Project Web site: **http://datamining.anu.edu.au/linkage.html**

Funded by the Australian National University, the NSW Department of Health,

and the Australian Research Council (ARC) under Linkage Project 0453463.

Outline

- A short introduction to data mining
  - Applications, techniques and challenges
  - The data mining process
- Short introductions to data linkage and geocoding
- Data quality aspects
  - Data size, complexity, sources, types and formats
  - Data processing issues, techniques and measures
- Our project: *Febrl*
  (Freely extensible biomedical record linkage)
- Conclusions and outlook

Data collections in the real world

- Many companies, organisations and research projects collect massive amounts of data
  - The ten largest *transaction-processing* databases range from 3 to 18 Terabytes, and the ten largest *decision support* databases range from 10 to 29 Terabytes
  - Sizes have doubled / tripled between 2001 and 2003
- Questions arise:
  - Is there any new, unexpected and potentially useful information contained in this data?
  - Can we use historical data to predict future outcomes (e.g. customer behaviour, fraud detection, etc.)

Example data mining applications (1)

- Health
  - Data collected at GPs, specialists, hospitals, Medicare, private health insurances, pharmacies, etc.
  - Questions include: *Are people committing fraud? Are doctors following the procedures? Are there adverse drug reactions?*
- Telecommunication
  - Data collected about transactions (phone calls, Internet access), customers (billing, addresses), networks, etc.
  - Questions include: *Which customer group is highly profitable, which one is not? How do customer profiles change over time? Can we detect fraud?*

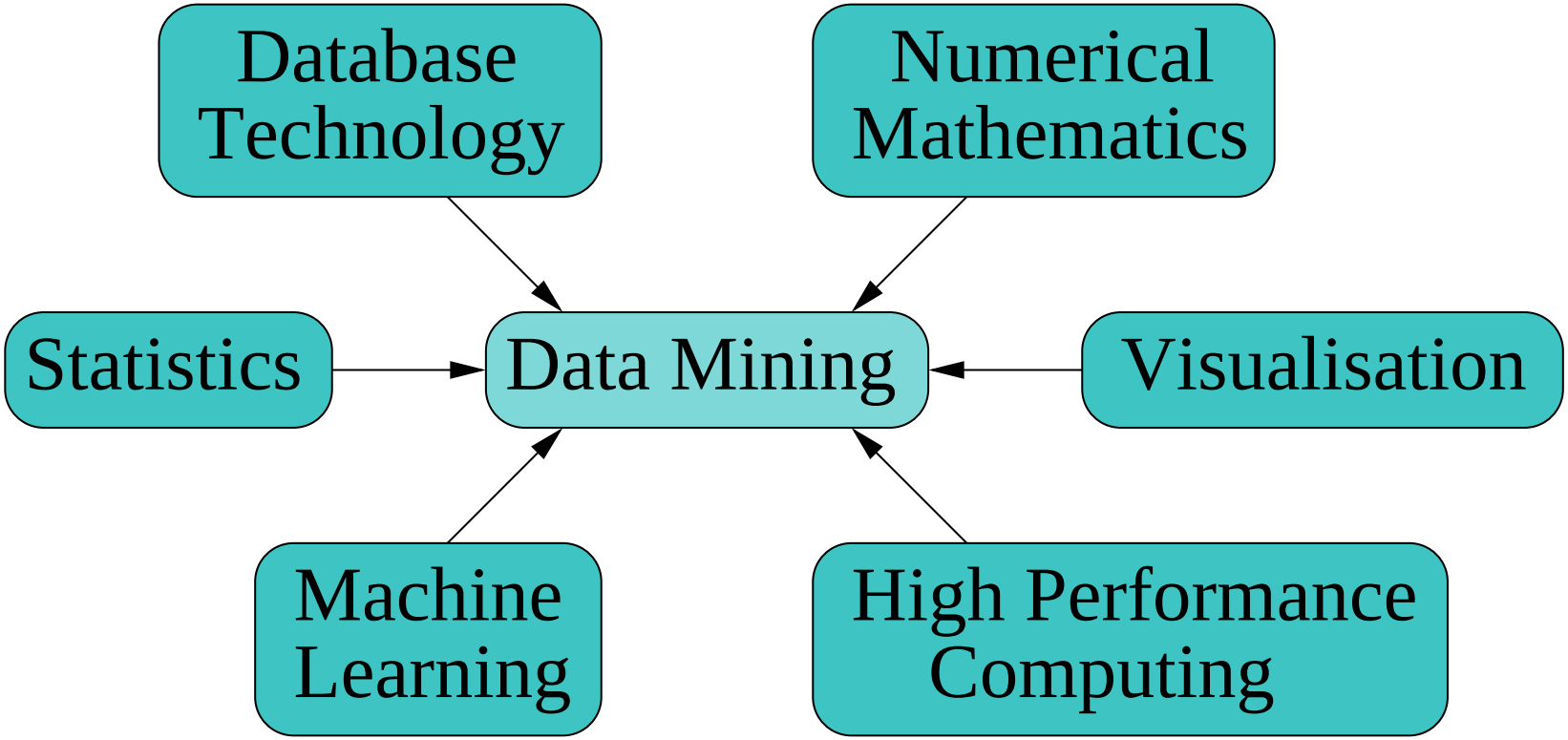
Example data mining applications (2)

- Economics and commerce (analysis and prediction of stock market)
- Market basket analysis (rules like *"if a customer buys beer he also buys chips with a likelihood of 80%"*)
- Bioinformatics (predict diseases based on genome sequences, find similar sequences)
- Governments (statistics, census, taxation, social security, immigration) (prevent fraud, improve outcomes)
- Credit card and insurance companies (segment customers for targeted marketing, detect fraud)
- Terror, crime and fraud detection (detect unusual events and suspicious individuals)

Definitions of data mining

- Fayyad, Piatetsky-Shapiro and Smyth, 1996
  Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.
- Two Crows (www.twocrows.com/glossary.htm)
  An information extraction activity whose goal is to discover hidden facts contained in databases. Using a combination of machine learning, statistical analysis, modelling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results. Typical applications include market segmentation, customer profiling, fraud detection, evaluation of retail promotions, and credit risk analysis.

Data mining is multi disciplinary



Data mining techniques (1)

- What they do: Detect patterns in data
  (rules, classes, clusters, associations and functional dependencies, outliers, data distributions)
- How they do it: Search through data and pattern space
  (non-parametric modelling, filtering, aggregation, supervised or non-supervised learning)
- How well they do it
  (errors and biases, over-fitting, confounding effects, speed, scalability, computational resources needed)

Data mining techniques (2)

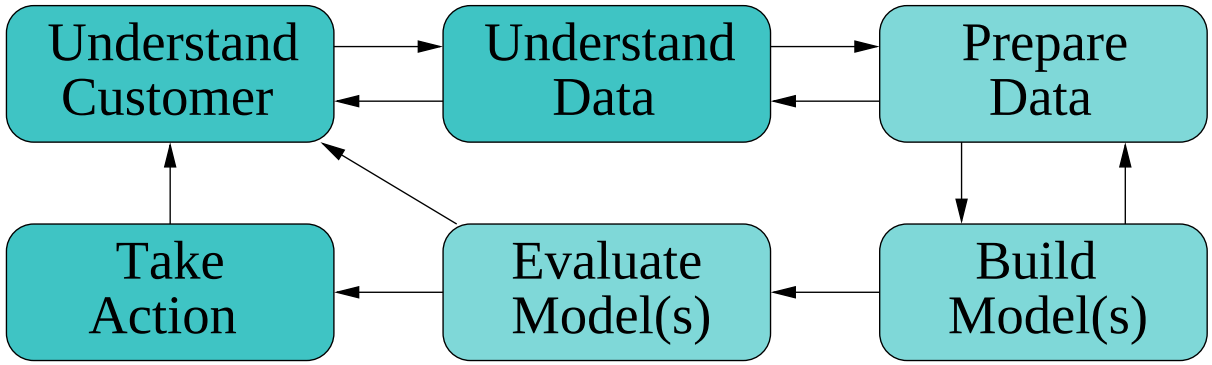
- Cluster analysis
  Group data to form classes, maximise intra-cluster similarity and minimise similarity between clusters
- Association rules discovery
  Find frequent rules in the data; popular with *market basket analysis*
- Classifi cation (e.g. decision trees)
  Build (binary) tree where each node corresponds to a split of attribute values, e.g. *"if the weather is sunny play golf else don't play golf.¨*
- Predictive modelling
  Build mathematical models (functions) of the data in order to predict unknown or missing values, or future outcomes

Data mining techniques (3)

- Outlier detection
  Find unusual, rare events (often regarded as noise, these can be the most interesting objects or events in the data), used for fraud detection, network intrusion detection, etc.
- Sequence / time series mining
  Find patterns over time (e.g. episodes, clusters)
- Spatial mining (geographical data analysis)
- Stream mining
  Where access to the data is limited to once (e.g. network data, telecommunications data, etc.), special algorithms are necessary
- Multimedia mining (images, audio, video)

The data mining / KDD process

- Data mining is an interactive process



- Typically up to 90% of time and efforts are spent in the fi rst three steps
- Data mining corresponds to "Build Model(s)" step
- Data mining is also called *Knowledge discovery in databases* (KDD) (some say data mining is only one essential step in the KDD process)

Major challenges in data mining (1)

- Data size
  - Size of data collections grows more than linear, doubling every 18 months (similar to Moore's law of CPU speed)
  - Scalable algorithms are needed
- Data complexity
  - Different types of data (semi- or unstructured like free text, HTML, XML, multimedia, etc.)
  - Dimensionality of the data increases (more attributes)
  - The *curse of dimensionality* affects many algorithms (for example find nearest neighbours in high dimensions)
- Data quality *(discussed later)*

Major challenges in data mining (2)

- Interestingness
  - Mining large data collections often results in many rules and patterns – which are new, useful and interesting?
  - How to measure novelty and interestingness?
- Embedding data mining algorithms and solutions within operational systems (mining on the desktop)
- Privacy and confi dentiality
  - The public is increasingly worried about their data being shared, matched, analysed and mined
  - Privacy-preserving techniques are needed that allow data mining without compromising privacy of individuals

Outline: Data linkage and geocoding

- A short introduction to data mining
  - Applications, techniques and challenges
  - The data mining process
- Short introductions to data linkage and geocoding
- Data quality aspects
  - Data size, complexity, sources, types and formats
  - Data processing issues, techniques and measures
- Our project: *Febrl*
  (Freely extensible biomedical record linkage)
- Conclusions and outlook

What is data (or record) linkage?

- The process of linking and aggregating records from one or more data sources representing the same entity (patient, customer, business name, etc.)
  - Also called *data matching*, *data integration*, *data scrubbing*, *ETL (extraction, transformation and loading)*, *object identification*, *merge-purge*, etc.
- Challenging if no unique entity identifi ers available

E.g., which of these records represent the same person?

Dr Smith, Peter	*42 Miller Street 2602 O'Connor*
Pete Smith	*42 Miller St 2600 Canberra A.C.T.*
P. Smithers	*24 Mill Street 2600 Canberra ACT*

Traditional data linkage

- For each compared record pair a vector containing *matching weights* is calculated

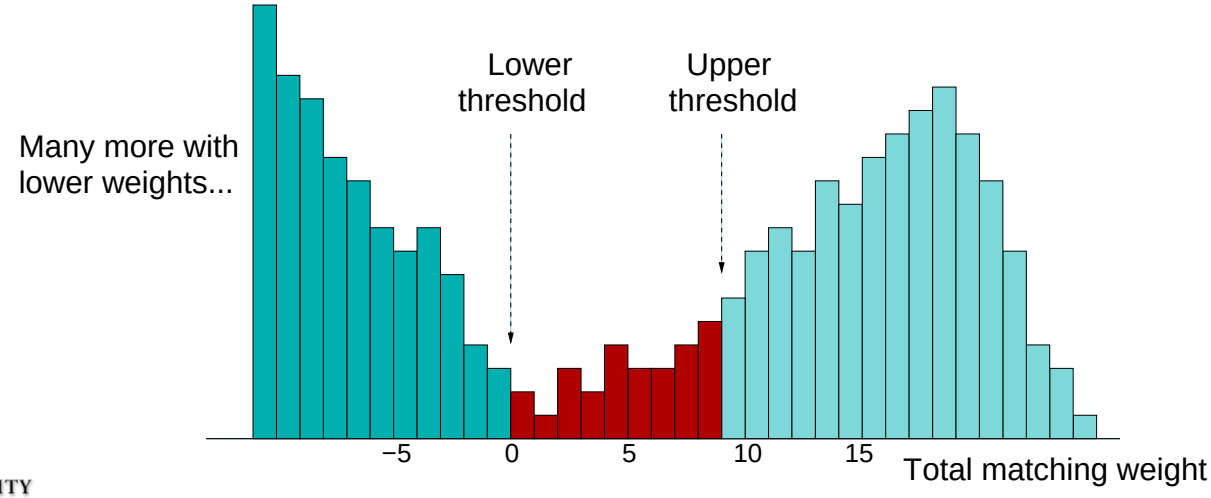
  Record A: `['dr', 'peter', 'paul', 'miller']`

  Record B: `['mr', 'john', '', 'miller']`

  Matching weights: `[0.2, -3.2, 0.0, 2.4 ]`

- Traditional approach sums all weights

  (then classifies record pairs into *matches*, *non-matches*, and *possible matches*) [Fellegi and Sunter 1969]



Modern data linkage approaches

- Summing of weights results in loss of information
  (like *same name but different address*, or *different address but same name*)
- View record pair classifi cation as a *multi-dimensional binary classifi cation* problem
  (use weight vector to classify record pairs a *matches* or *non-matches*, but no *possible matches*)
- Many machine learning techniques can be used
  - Supervised: Decision trees, neural networks, learnable string comparisons, active learning, etc.
  - Un-supervised: Various clustering algorithms
- Major issue: Lack of training data

Challenges in data linkage (1)

- Increasingly large data collections
  - Number of possible record pairs to compare equals the product of the sizes of the two data sets
  - Performance bottleneck is usually the (expensive) comparison of attribute values between record pairs
- Manual clerical review process
  - Traditionally, *possible matches* are manually looked at to decide their linkage status
  - With larger data collections, the number of possible matches also increases
  - Very time consuming and tedious, but also hard to make correct and consistent decisions

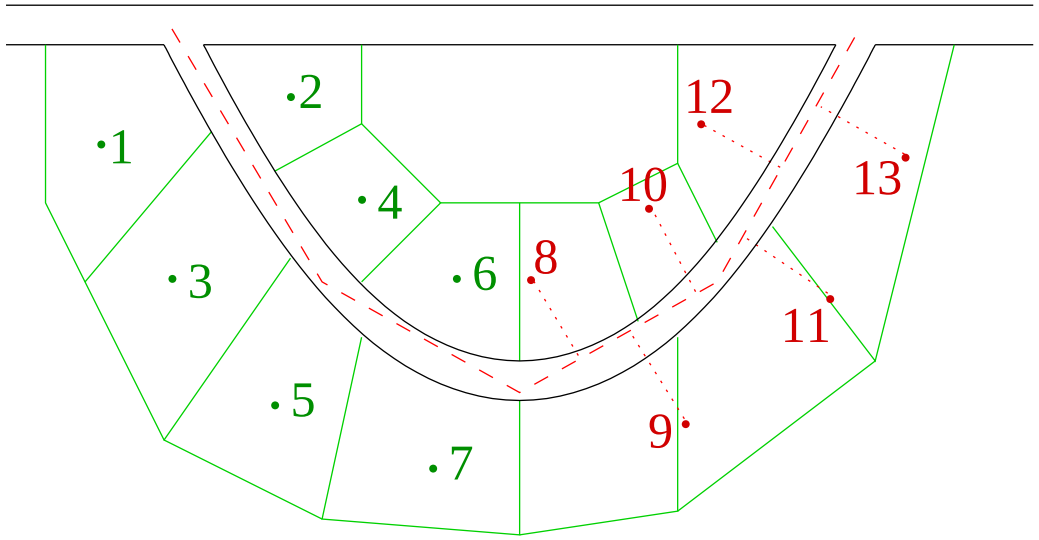
Challenges in data linkage (2)

- General public is worried about their information being linked and shared between organisations
  - Good: health and social research; statistics, crime and fraud detection (taxation, social security, etc.)
  - Scary: intelligence, surveillance, commercial data mining (not much information from businesses, no regulation)
  - Bad: identity fraud, re-identification
- Traditionally, *identifi ed data* has to be given to the person or organisation performing the linkage
  - Privacy of individuals in data sets is invaded
  - Consent of individuals involved is needed (often not possible, so seek approval from ethics committees)

What is geocoding?

- The process of matching addresses to their geographic locations (longitude and latitude)
  - Large reference database of standardised addresses needed (in Australia: *G-NAF*)
  - Accurate matching is important
  - Addresses often contain typographical and other errors, are incomplete or out-of-date (discussed later)
- It is estimated that 80% to 90% of governmental and business data contain address information *(US Federal Geographic Data Committee)*
- Useful in many application areas
  - Visualisation, spatial data analysis and mining

Geocoding techniques



- Street centreline based (many commercial systems)
- Property parcel centre based (e.g. *G-NAF* based)
- A recent study found substantial differences
  (specially in rural areas)
  Cayo and Talbot; Int. Journal of Health Geographics, 2003

Geocoding examples (GPs)

- Red dots: *Febrl* geocoding (*G-NAF* based)
- Blue dots: Street centreline based geocoding

Outline: Data quality

- A short introduction to data mining
  - Applications, techniques and challenges
  - The data mining process
- Short introductions to data linkage and geocoding
- Data quality aspects
  - Data size, complexity, sources, types and formats
  - Data processing issues, techniques and measures
- Our project: *Febrl*
  (Freely extensible biomedical record linkage)
- Conclusions and outlook

Data size and complexity

- *We are drowning in data, but starving of knowledge* (Jiawei Han)
- Automated data collection and mature database technology (allows data to be stored efficiently, cheap, persistent; in databases, data warehouses and other repositories)
- Large and massive data collections
  - Tens to thousands of attributes (or variables)
  - Millions to billions of records
  - **Data is rarely collected for data mining purposes** (rather mainly for online transaction processing)
  - A lot of data is *write only* (or *read once* only)

Sources of data

- Relational databases (transactional data, mostly normalised into many tables, with keys between them, continuous and frequent updates)
- Data warehouses (decision support data, processed and cleaned, historical data, aggregated, updated at certain intervals – daily, weekly, monthly)
- Internet (click-stream data, log files, HTML, XML, e-mails)
- Simple fi les (portable text (e.g. comma separated values) or non-portable, proprietary binary files)
- Scientifi c instruments and experiments (astronomy, genomics, seismology, physics, chemistry, meteorology, etc.)

Data types and formats

- Numerical data (integer, floating-point, non-scalar)
- Non-numerical data (nominal, categorical, ordinal)
- Series data (ordering is an important feature, often time)
- Multimedia data (images, video, audio)
- Data can be structured, semi-structured or free-format (database tables, HTML, e-mails, simple text)
- Different mappings and conversions between data types are possible and often needed
- Different data mining techniques can handle different types of data (or are restricted to certain types of data)

Sources of errors in data

- *Real world data is dirty* (M. Hernandez / S. Stolfo)
- Various sources of errors in data
  - Errors during data entry or data collection
  - Missing data
  - Out-of-date data
  - Misinterpretations
  - Equipment malfunctioning
- Personal information (names and addresses) are especially prone to data entry errors
- A great effort is often needed to *standardise* and *clean* the raw data (data pre-processing)

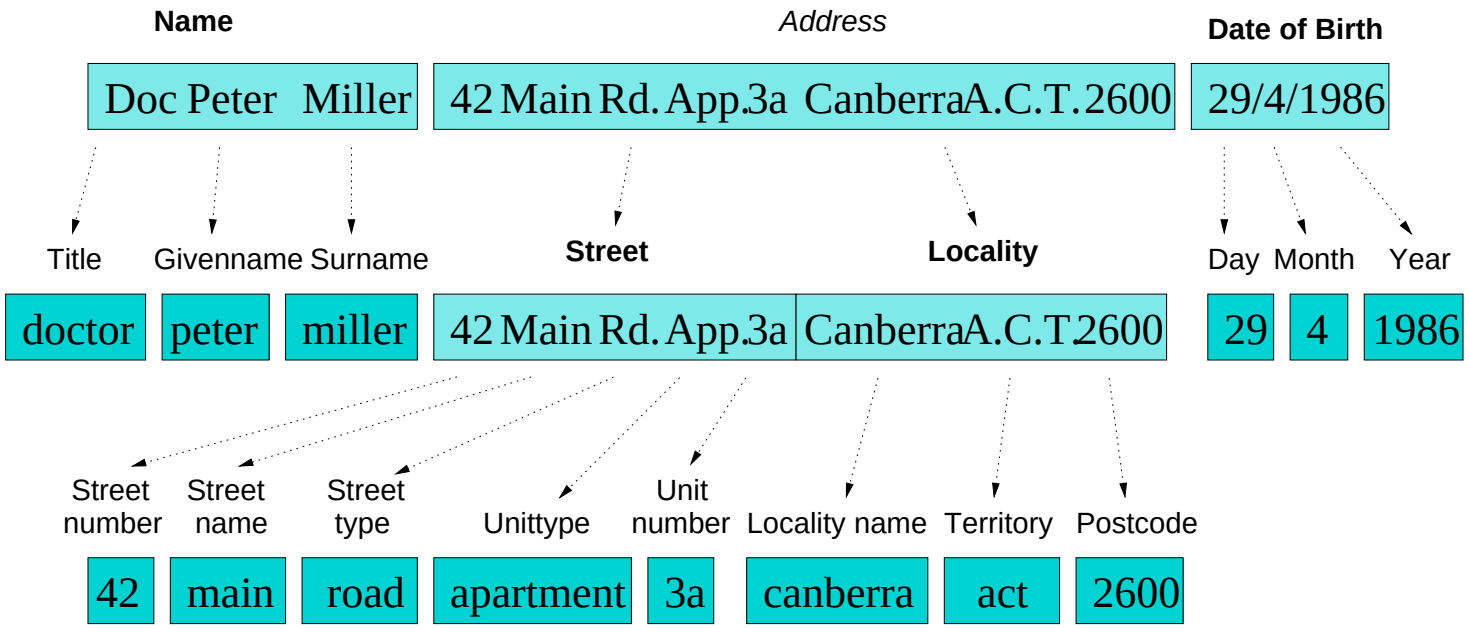
Dirty data

- What does *dirty data* mean?
  - Incomplete data (missing attributes, missing attribute values, only aggregated data, etc.)
  - Inconsistent data (different coding schemes and formats, impossible values or out-of-range values)
  - Noisy data (containing errors and typographical variations, outliers, not accurate values)
- Good quality data is needed for good quality data mining and data linkage (*garbage in – garbage out*)
- Transactional databases systems should be designed with data quality in mind (data entry, validity, duplication and consistency checks)

Issues with names and addresses

- Names have different characteristics to normal text
  - Many names have several valid forms (like *Gail*, *Gale* and *Gayle*)
  - Nicknames are commonly used (like *Liz* and *Bob*)
  - Names are influenced by culture and language
  - People change names over time
- Data entry errors on names and addresses
  - Handwritten forms (manually typed or optical character recognition), character substitutions (like $q \leftrightarrow g$, $l \leftrightarrow 1$)
  - Over the telephone (phonetic mistakes)
  - Limitations of input fields (maximum length)

Name and address parsing



- Remove unwanted characters and words
- Expand abbreviations and correct misspellings
- Segment free format text into well defined attributes
- Check validity and consistency using look-up tables

Data pre-processing tasks

- Data cleaning (fill in missing values, smooth noisy data, identify/remove outliers, detect and resolve inconsistencies)
- Data transformation (normalisation, generalisation and aggregation, consolidate into a form suitable for data mining)
- Feature construction (based on existing attributes, can help to discover missing information about the relationships between data attributes)
- Data reduction and discretisation (reduce volume of data, but still produce same or similar analytical result, discretisation in particular for numerical data)
- Important to save data pre-processing steps performed (in meta-data repository)

Data quality measures

- Accuracy
- Completeness
- Consistency
- Timeliness
- Believability
- Interpretability
- Accessibility

Outline: Our project

- A short introduction to data mining
  - Applications, techniques and challenges
  - The data mining process
- Short introductions to data linkage and geocoding
- Data quality aspects
  - Data size, complexity, sources, types and formats
  - Data processing issues, techniques and measures
- Our project: *Febrl*
  (Freely extensible biomedical record linkage)
- Conclusions and outlook

Our project: Febrl

- We aim to develop new and improved techniques for parallel large scale data linkage
- Main research areas
  - Probabilistic techniques for automated data cleaning and standardisation (mainly on addresses, using *G-NAF*)
  - New and improved blocking and indexing techniques
  - Improved record pair classification using (un-supervised) machine learning techniques (reduce clerical review)
  - Improved performance (scalability and parallelism)
- Project Web site:

  http://datamining.anu.edu.au/linkage.html

Febrl prototype software

- An experimental platform for new and improved data linkage algorithms
- Modules for data cleaning and standardisation, data linkage, deduplication, geocoding, and generation of synthetic data sets
- Open source **https://sourceforge.net/projects/febrl/**
- Implemented in *Python* **http://www.python.org**
  - Easy and rapid prototype software development
  - Object-oriented and cross-platform *(Unix, Win, Mac)*
  - Can handle large data sets stable and efficiently
  - Many external modules, easy to extend, large community

Conclusions and outlook

- Good quality data is important for data mining, data linkage and geocoding
  - *Data cleaning* and *standardisation* are important first steps in any data mining or linkage project
  - Names and addresses are especially prone to data entry errors (telephone, handwritten, OCR)
- New ANU master level course *Algorithms and Techniques for Data Mining* (COMP8400)
  - Starts February 2007, lecturer: Dr Peter Christen
- For more information on our project please visit:
  http://datamining.anu.edu.au/linkage.html