# High-Performance Computing Techniques for Record Linkage

*Peter Christen*, Tim Churches, Markus Hegland, Kim Lim,
Ole M. Nielsen, Stephen Roberts and Justin Zhu

Data Mining Group, Australian National University
Epidemiology and Surveillance Branch, NSW Department of Health

Contact: **peter.christen@anu.edu.au**
Project web page: **http://datamining.anu.edu.au/linkage.html**



## *A Collaborative Research Project*

- ANU Data Mining Group
  - Department of Computer Science
  - Mathematical Sciences Institute
  - Australian Partnership for Advanced Computing (APAC)
- New South Wales Department of Health
  - Epidemiology and Surveillance Branch

*Funded by ANU and NSW Department of Health under an ANU Industry Collaboration Scheme (AICS)*



## *Record Linkage / Data Matching*

- The task of linking together information from one or more data sources representing the same entity (patient, customer, provider, business, etc.)
- If no *unique identifier* is available, *probabilistic linkage techniques* have to be applied
- Applications of record linkage
  - Remove duplicates in a data set (internal linkage)
  - Merge new records into a larger master data set
  - Create patient oriented statistics
  - Compile data for longitudinal studies
  - Clean data sets for data mining projects or mailing lists



## *Record Linkage Example*

- Three records, which are the same person?

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

- *Data cleaning* and *standardisation* is an important first step for successful record linkage
- *Probabilistic linkage* is based on *matching weights*
  - Use available information (names, addresses, dates) (can be missing, wrong, coded differently, outdated, etc.)
  - Compute *matching weights* based on *frequency counts*
  - Classify a pair of records as *link*, *possible-link* or *non-link*

## Why this project?

- Commercial software for record linkage is often expensive and cumbersome to use
- Project aims
  - Allow linkage of larger data sets (high-performance and parallel computing techniques)
  - Reduce the amount of human resources needed (improve linkage quality by using machine learning and data mining techniques)
  - Reduce costs (free open source software)

*Facilitate (epidemiological) research with free and improved tools for record linkage*




## Open Source Software Tools

- Advantages of open source software
  - Can be downloaded for free from the Internet
  - Program code can be modified and improved
  - Often a worldwide supportive user community
  - Examples: *Linux*, *Apache*, *Samba*, *MySQL*, etc.
- Software tools used for this project
  - Programming language *Python* **www.python.org** (efficient, stable, many external modules, good support, easy to extend, available for *Unix*, *Windows* and *Mac*)
  - Parallel extensions and libraries (*MPI*, *OpenMP*, *PyPar* and *PyRO*)

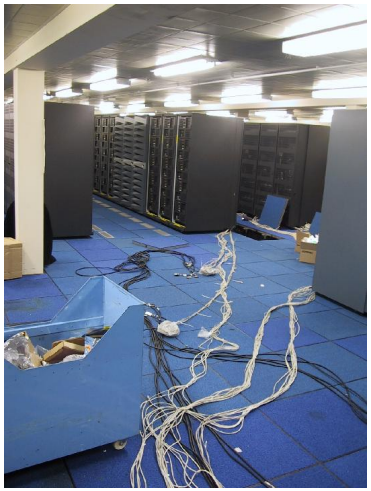



## Target Computing Platforms

- Workstation or PC cluster
  - Commodity PCs connected via local area network
  - Widespread availability, no extra costs
  - Use as virtual parallel computer (over night / weekends)
- Multiprocessor (SMP) servers
  - Example: *Sun Enterprise, HP Superdome*
  - 4 – 30 CPUs, Gigabytes of memory, Terabytes of disk
- High-performance super-cluster
  - Example: *APAC National Facility (Compaq Alphaserver)*
  - >100 CPUs, Gigabytes of memory, mass data storage




## Linux Cluster 'Bunyip' and APAC National Facility

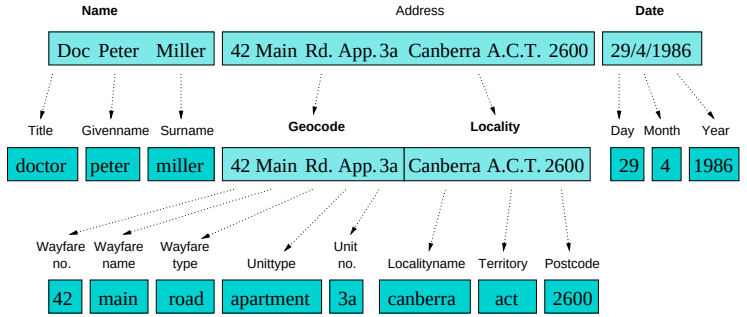


# Project Plan and Status

- Project plan
  - Data standardisation *(January - July)*
  - Probabilistic record linkage *(August - September)*
  - Parallelisation *(October - November)*
  - Data mining and machine learning *(December - March)*
- Prototype software will be available soon
  - Data standardisation routines for *names* and *addresses*
  - Lookup-tables (*names*, *titles*, *localities*, *countries*, etc.)
  - Various other routines (e.g. *name encodings* and *approximate string comparisons*)



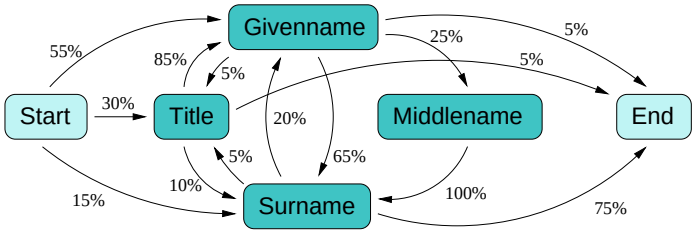

# Data Standardisation Approach




- Two different approaches
  - Rules based (traditional) [e.g. *AutoStan* software]
  - Probabilistic *Hidden Markov Models (HMM)* [new]
    (standardised training data needed)



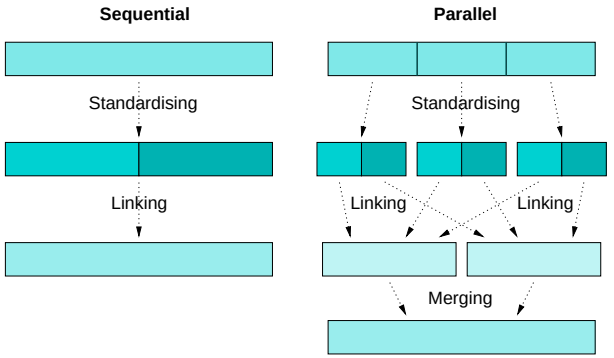

# HMM Standardisation




- First experiments with *Midwifes Data Collection*
  - Ten years of data (1990 - 2000), with 962,776 records
    (no medical information, names and addresses only)
  - Average $95\% \pm 5\%$ accuracy with addresses
    (10-fold cross-validation using 900 training records and 100 test records per fold)




# Parallelisation Approach




- Each record can be standardised independently
- Linkage is done using *blocking*
  (each block can be processed independently)

## Data Mining Approach

- *Data mining* and *machine learning* techniques to learn data characteristics
  - *Clustering* (as alternative for blocking)
  - *Predictive modelling*
  - *Decision trees* and *rules* (for matches / non-matches)
- *Training data* needed to build model
  (example pairs of known matches and non-matches)

*ANU Data Mining group has several years of experience in predictive modelling, handling of health data sets, data processing, etc.*




## Outlook

- A new approach to probabilistic record linkage
  - High-performance and parallel computing techniques
  - Data mining and machine learning techniques
  - Free open source software
- Future extension of this project likely
  - ARC Linkage grant for 2003
- Further collaborations are welcome
- Free prototype software available online soon:

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