Parallel Computing Techniques for High-Performance Probabilistic Record Linkage

Peter Christen, Markus Hegland, Stephen Roberts and Ole M. Nielsen

Tim Churches and Kim Lim

Data Mining Group, Australian National University

Epidemiology and Surveillance Branch, NSW Health Department

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

Project web page: **http://datamining.anu.edu.au/linkage.html**

THE AUSTRALIAN
NATIONAL UNIVERSITY

Project Description and Aim

- A collaborative research project between the ANU and the NSW Health Department
- Use of (free) open-source software tools to develop open-source record linkage software
  - Step 1: Develop high-performance techniques for record linkage ⇒ **Faster linkage of larger data sets**
  - Step 2: Explore machine learning and data mining techniques for record linkage ⇒ **Better linkage quality**

The aim is to facilitate (epidemiological) research with free and improved tools for record linkage

A Collaborative Research Project

- Australian National University Data Mining Group
  - **Peter Christen**, Department of Computer Science
  - **Markus Hegland**, School for Mathematical Sciences
  - **Stephen Roberts**, School for Mathematical Sciences
  - **Ole M. Nielsen**, School for Mathematical Sciences and Australian Partnership for Advanced Computing (APAC)
  - **Justin Zhu**, Computer Science Honours student
- New South Wales Health Department
  - **Tim Churches**, Epidemiology and Surveillance Branch
  - **Kim Lim**, Epidemiology and Surveillance Branch

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

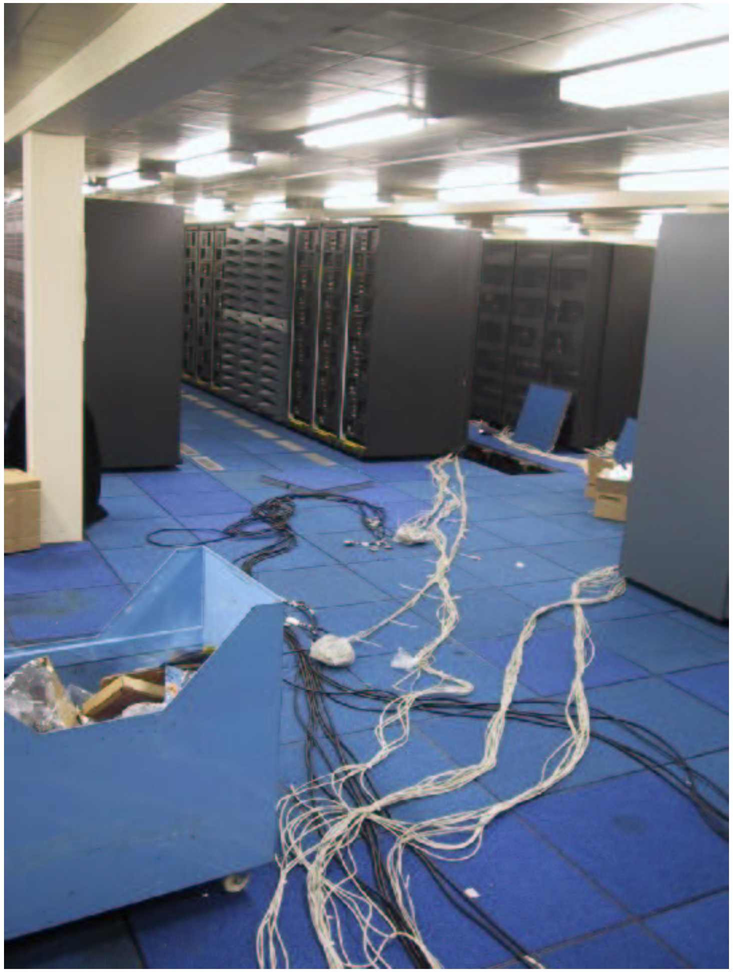
Open Source Software Tools

- Scripting language *Python*
  - Easy and rapid prototype software development
  - Provides lists and dictionaries (lookup tables)
  - Can handle large data sets stable and efficiently
  - Many external modules, easy to extend
  - Available from **www.python.org** (Windows, Unix, Mac)
- Parallel libraries *MPI* and *OpenMP*
  - For communication between processes
  - Widespread use in high-performance computing (quasi standards) ⇒ Portability and availability

Target Computing Platforms

- Workstation or PC cluster
  - Commodity PCs connected via local area network
  - Widespread availability, no extra costs
  - Use as virtual parallel computer (nights / 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



THE AUSTRALIAN
NATIONAL UNIVERSITY

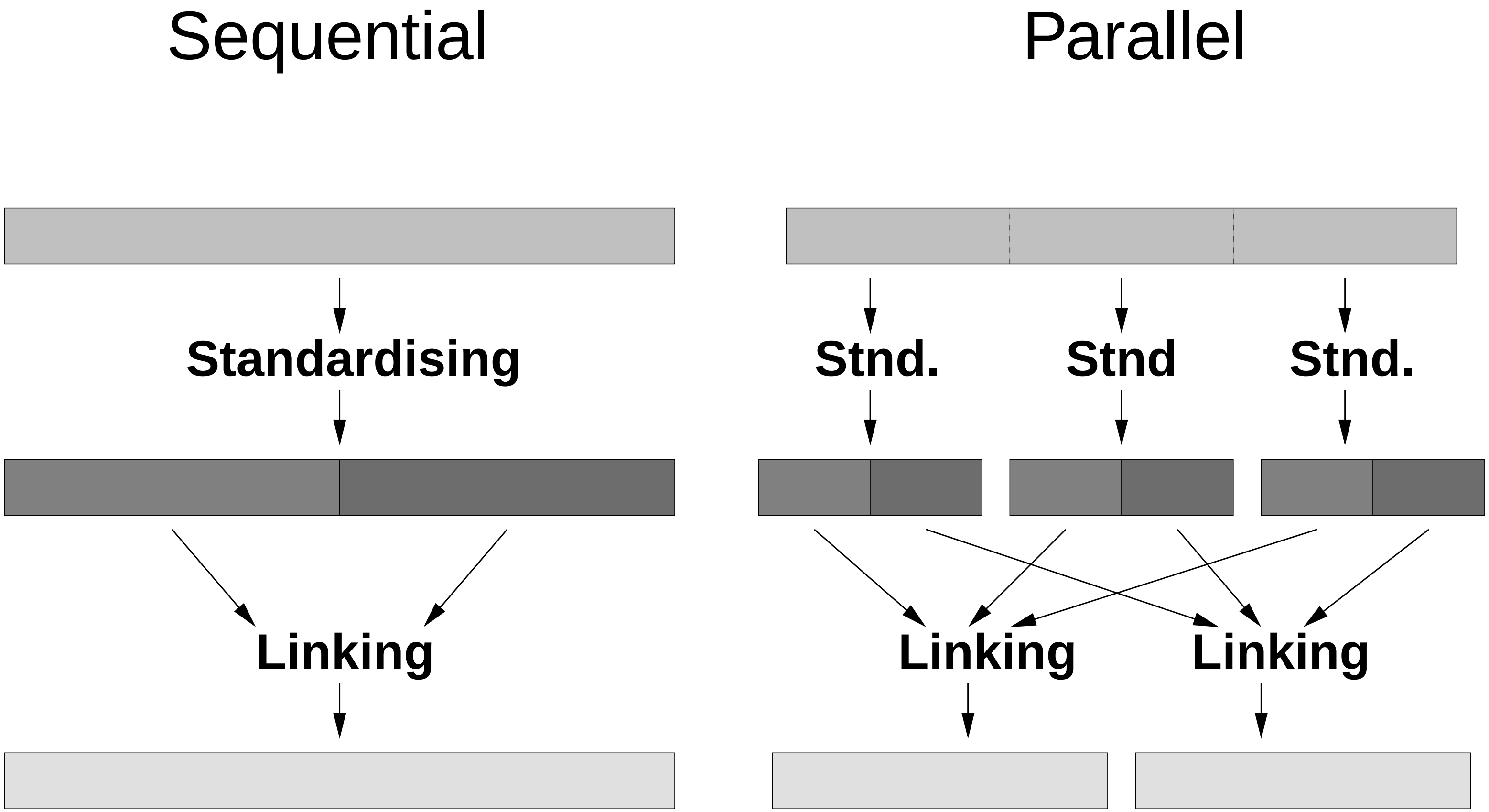
Status and Ongoing Work

- Project started in January 2002 (officially March)
- Implemented and tested *Python* modules
    - Name encodings: *Soundex*, *NYSIIS*, *Double-Metaphone*
    - String comparators: *Jaro*, *Winkler*, *Bigram*, *Edit distance*
- Currently working on standardisation routines
    - *NAME* (almost finished), *GEOCODE* and *LOCALITY*
- Students
    - Justin Zhu (Honours) *Hidden Markov Models*
    - Agnes Boskovitz (PhD) *Inductive Logic Data Cleaning*

Standardisation Approach

- Routines for *NAME*, *GEOCODE* and *LOCALITY*
- *NAME* standardisation:
  - Remove unwanted characters, replace certain characters by others. Example: Replace *[*, *{*, < with *(*
  - Split into a list of words and separators
    Example: *['ms', 'monica', '(', 'mon', ')', 'meyer', '-', 'miller']*
  - Assume and use sequence structure
  - Extract titles from beginning of list (use lookup tables)
  - Handle easier names first (e.g. if only two words left)
- Similar for *GEOCODE* and *LOCALITY*

Parallelisation Approach



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
  (pairs of known matches and known 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
  - Open-Source software
  - Data mining and machine learning techniques
- Future extension of this project likely
  - ARC Linkage grant for 2003
- Further collaborations are welcome
- Prototype software available in second half of 2002

Project web page: **http://datamining.anu.edu.au/linkage.html**