High Performance Computing and Data Mining

Performance Issues in Data Mining

Peter Christen

Peter.Christen@anu.edu.au

Data Mining Group

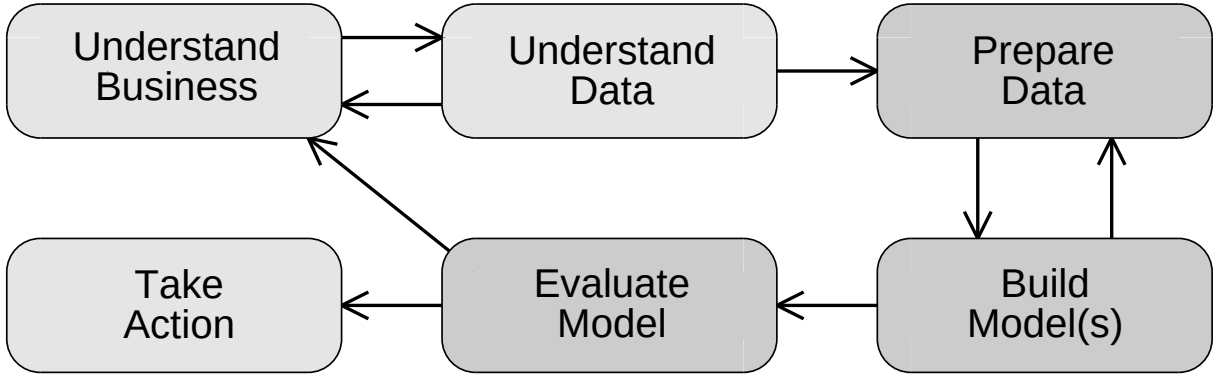
Department of Computer Science, FEIT

Australian National University, Canberra

http://csl.anu.edu.au/ml/dm/

THE AUSTRALIAN NATIONAL UNIVERSITY

The Data Mining Process



- Analysis: Fast data access, large memory, caching
- Preparation: Fast input and output, large memory, fast computing
- Modelling: Fast computing, large memory
- Evaluation: High performance graphics

Why High Performance Computing

- Large data collections $\rightarrow$ Memory and disk space
- Long processing times $\rightarrow$ Processing speed
- Technical limitations
  - Processor speed
  - Input / output bandwidth
  - Memory size and bandwidth
- Many problems are inherently parallel

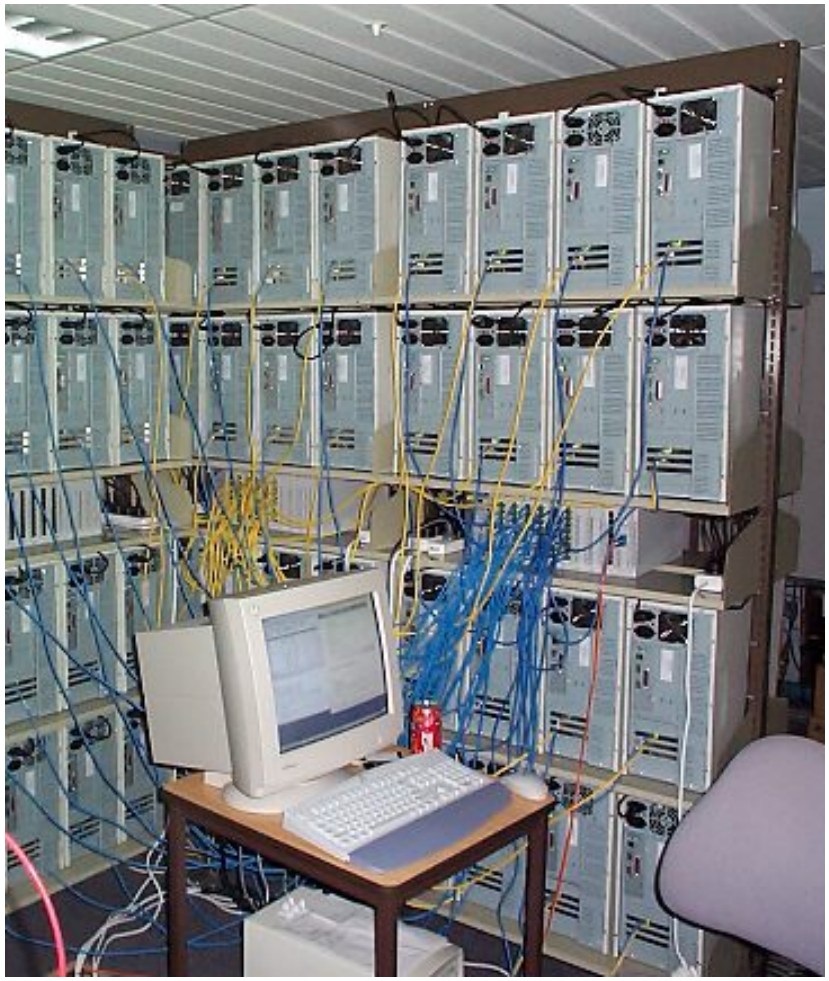
Contemporary high performance computing always involves parallel computing

Parallel Performance

- Goal: Being *P* times faster with *P* processors
  - *Speedup* is usually less than *P*
  - Sequential parts in a program limit speedup
- Parallel scalability
  - Measurement how well speedup scales with increasing number of processors
    (of course scalability with data size is equally important)
- Data distribution and load balancing are critical
- Parallel programs need to be tuned for new architectures

ANU Beowulf Linux Cluster Bunyip

- 96 Dual Pentium III nodes
- 36 Gigabytes main memory
- 1,305 Gigabytes disk space
- Fast-Ethernet network
- Gordon Bell prize winner 2000
- Costs: AU$ 250,000



Australian Partnership for Advanced Computing (APAC)

- ANU Data Mining is 1 of 13 Expertise Programs
  - Conduct research and development projects
  - Provide high-level user support services
- National Facility at ANU opened in May 2001
  - Peak performance close to 1 Tera-Flops
  - 480 Compaq Alpha processors
  - Each with 1 Gigabyte of main memory
  - Connected by a fast, low latency switch
  - Disk capacity around 10,000 Gigabytes
  - Tape storage 300 Terabytes (300,000 Gigabytes)

APAC National Facility

THE AUSTRALIAN NATIONAL UNIVERSITY

Research at ANU Data Mining Group

- *DMtools* facilitate analysis and preprocessing
  - Access to parallel database server
  - Caching for fast retrieval
  - Uniform interface for parallel data mining algorithms
- Parallel scalable data mining algorithms
  - Predictive modelling
  - Clustering and association rules

Aim: Harness the power of high performance computing with a flexible toolbox

Parallel Record Linkage Initiative

- Probabilistic linkage of data sets if no common unique identifier is available
  (Probabilistic measure of how similar two records are)
- Only a few very expensive commercial programs are available (e.g. AutoMatch)
- To reduce the huge number of comparisons, blocking techniques are used
  (e.g. group records with same postcode)
- Blocking allows exploration for parallelism
- Collaboration with NSW Health Department
  (Tim Churches)

Outlook: Current and Future Work

- Parallel record linkage initiative
  (High-performance linkage package, open source software)
- Extension of *DMtools*
  - Integration of parallel data mining algorithms
  - Integration of statistical and graphical packages
- Extension of predictive modelling
  - Sparse grids
  - Complex data types (sets, vectors, etc)

Visit our web site at:

http://csl.anu.edu.au/ml/dm/