

MineBench : A Benchmark Suite for Data Mining Applications

Ramanathan Narayanan Berkin Ozisikyilmaz Gokhan Memik Alok Choudhary

Department of EECS
Northwestern University
Evanston, IL 60208

{ran310, boz283, memik, choudhar} @eecs.northwestern.edu

Joseph Zambreno

Electrical and Computer Engineering
Iowa State University

Acknowledgements

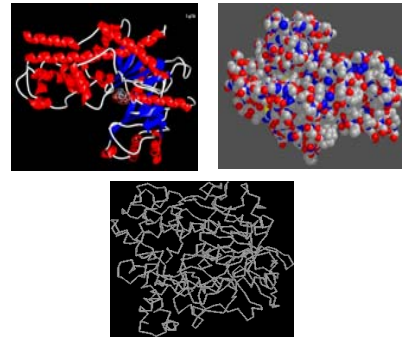
National Science Foundation (grants CCF-0444405, CNS-0406341, CCR-0325207)
Department of Energy (grant DE-FC02-01ER25485)
Intel Corporation

Overview

- ❑ Introduction
- ❑ Motivation
- ❑ Benchmark Suite Overview
- ❑ Using MineBench
- ❑ Conclusions

Explosion of Data

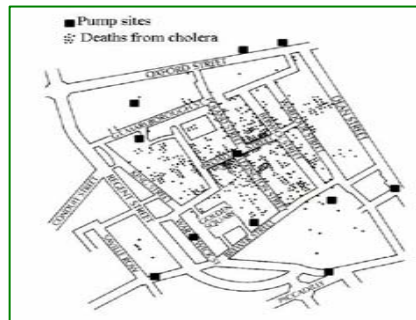
- Today's digital society has seen enormous data growth in both commercial and scientific databases
- Data scales at a high rate, exceeding Moore's Law
- Data Mining is becoming a commonly used tool to extract information from large and complex datasets



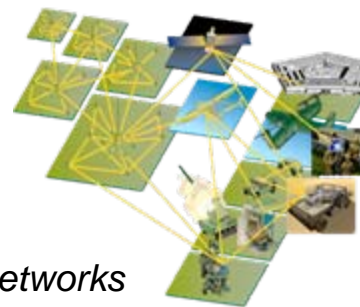
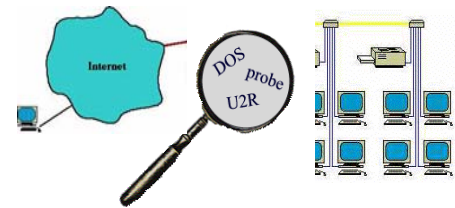
Biomedical Data



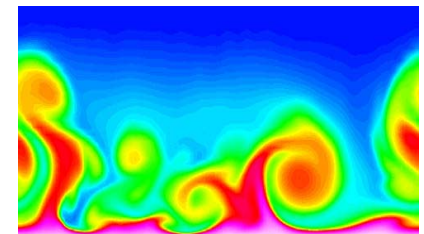
Homeland Security



Geo-spatial data



Sensor Networks

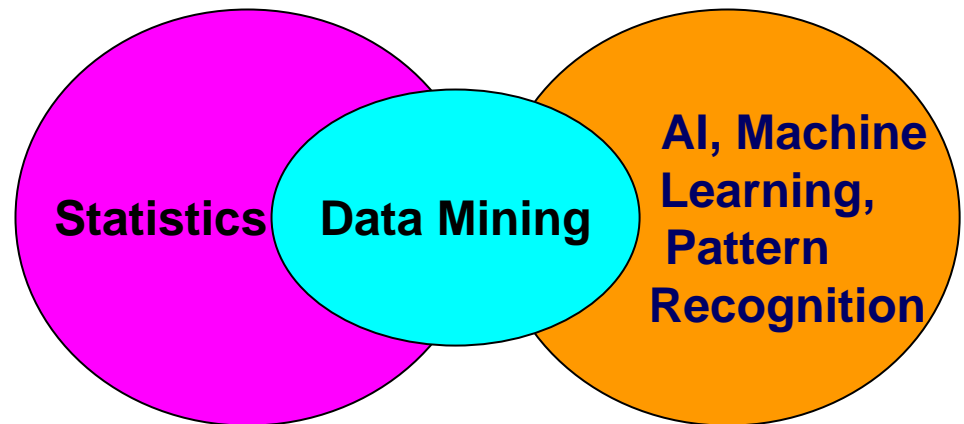


Computational Simulations

What is Data Mining ?

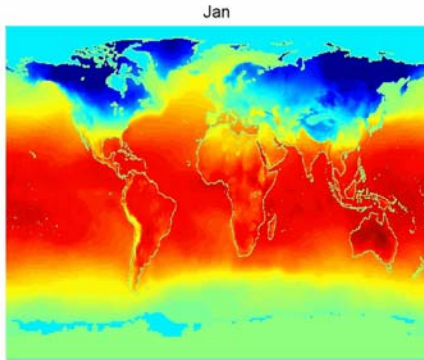
❑ Data Mining has been defined as “ the non-trivial extraction of implicit, previously unknown and potentially useful information from data ”

- Non-Trivial
- Previously unknown
- Automated
- Predictive



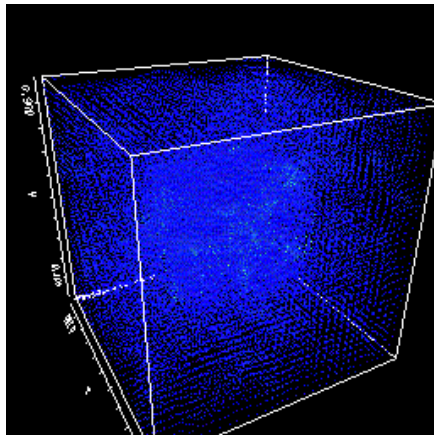
Database Technology, Parallel & Distributed Computing

Data Mining in Scientific Applications



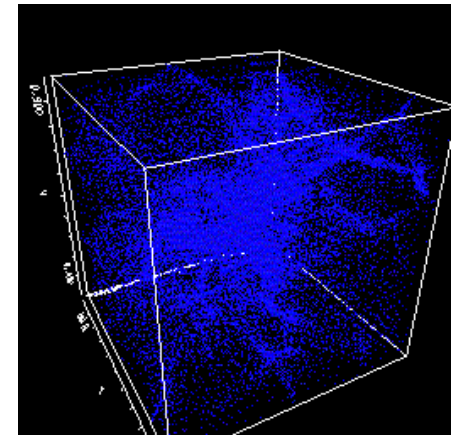
Climate Modeling

- What are the primary factors influencing climate ?
- How do natural and human-induced changes affect climate ?
- How well can we predict changes?



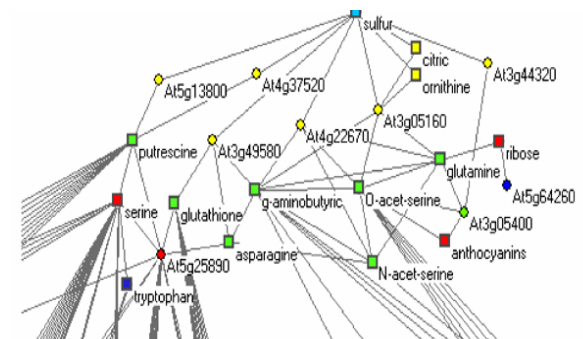
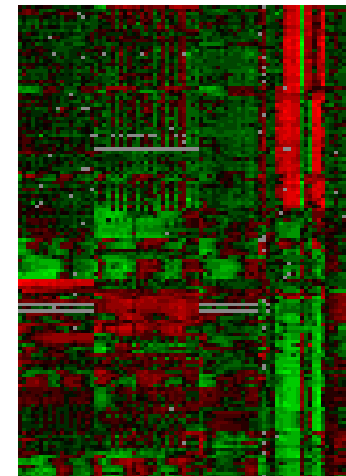
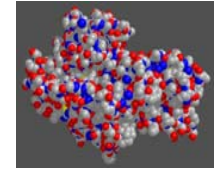
Cosmological Simulations

- Simulate formation and evolution of galaxies
- Adaptive mesh refinement



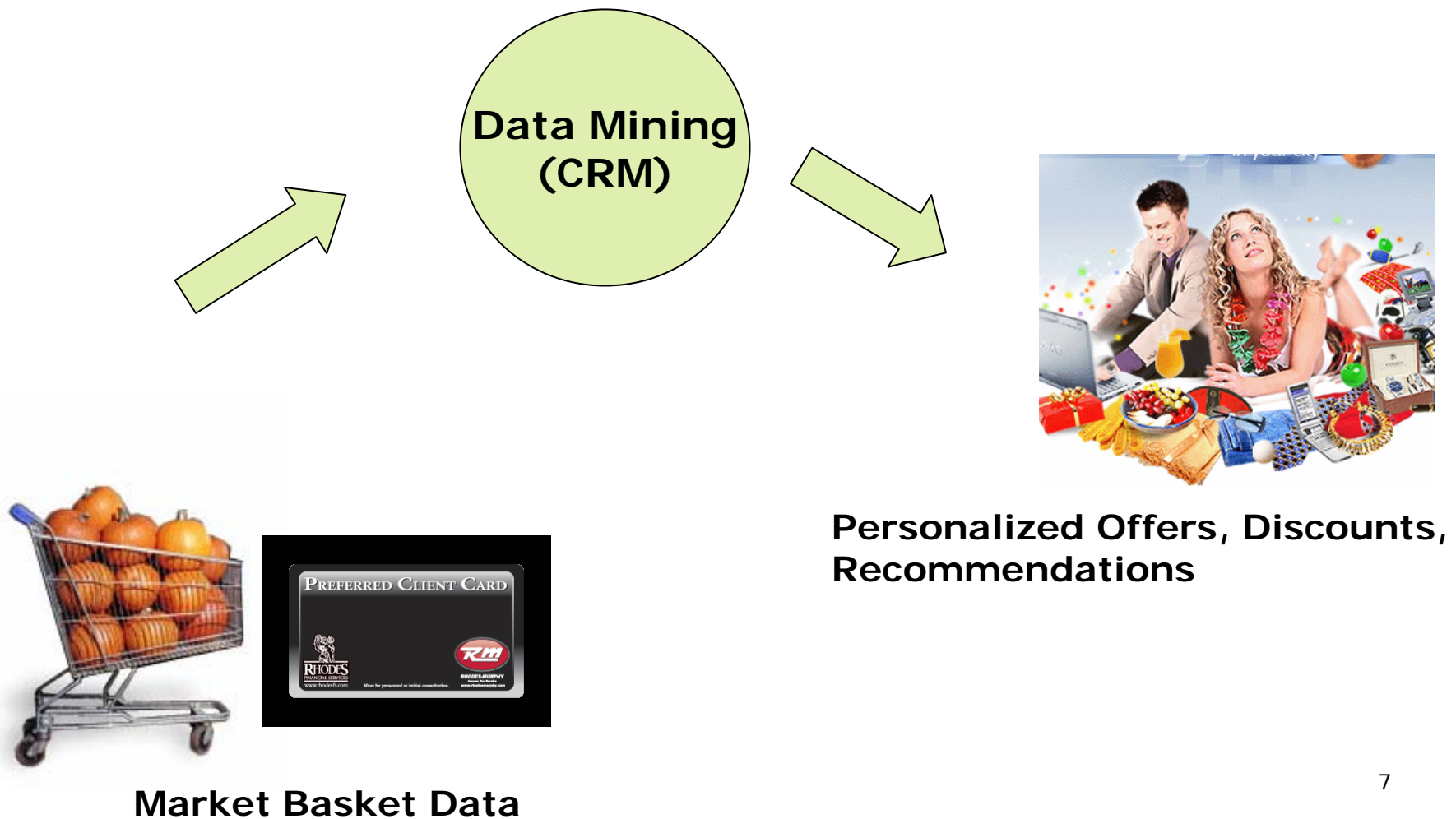
Data Mining for Bioinformatics

- Recent technological advances are helping to generate large amounts of both medical and genomic data
 - High-throughput experiments/techniques
 - Gene and protein sequences
 - Gene-expression data
 - Biological networks and phylogenetic profiles
- Data mining offers potential solution for analysis of large-scale data
 - Automated analysis of patients history for customized treatment
 - Design of drugs/chemicals
 - Prediction of the functions of anonymous genes

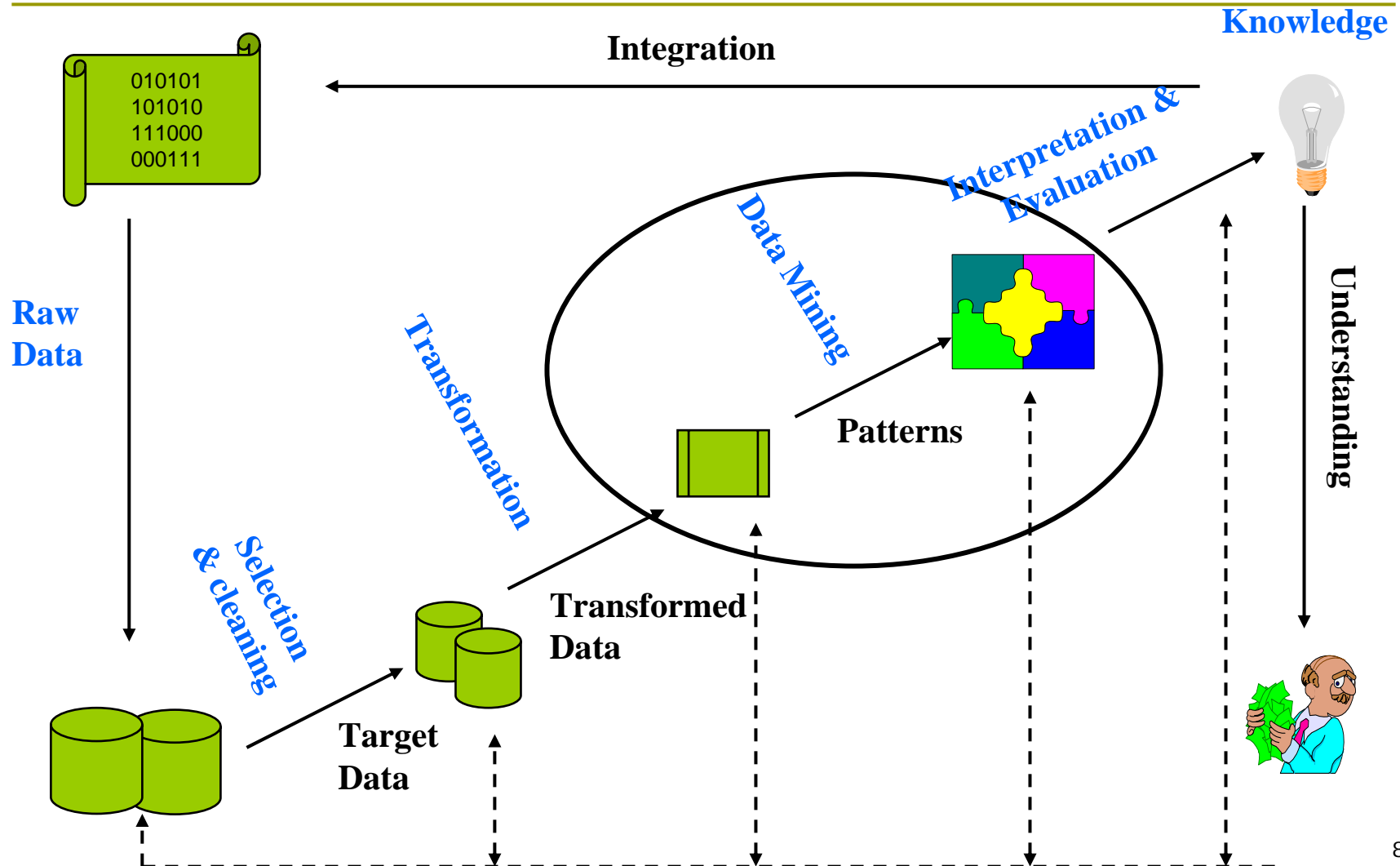


Protein Interaction Network

Data Mining in Business

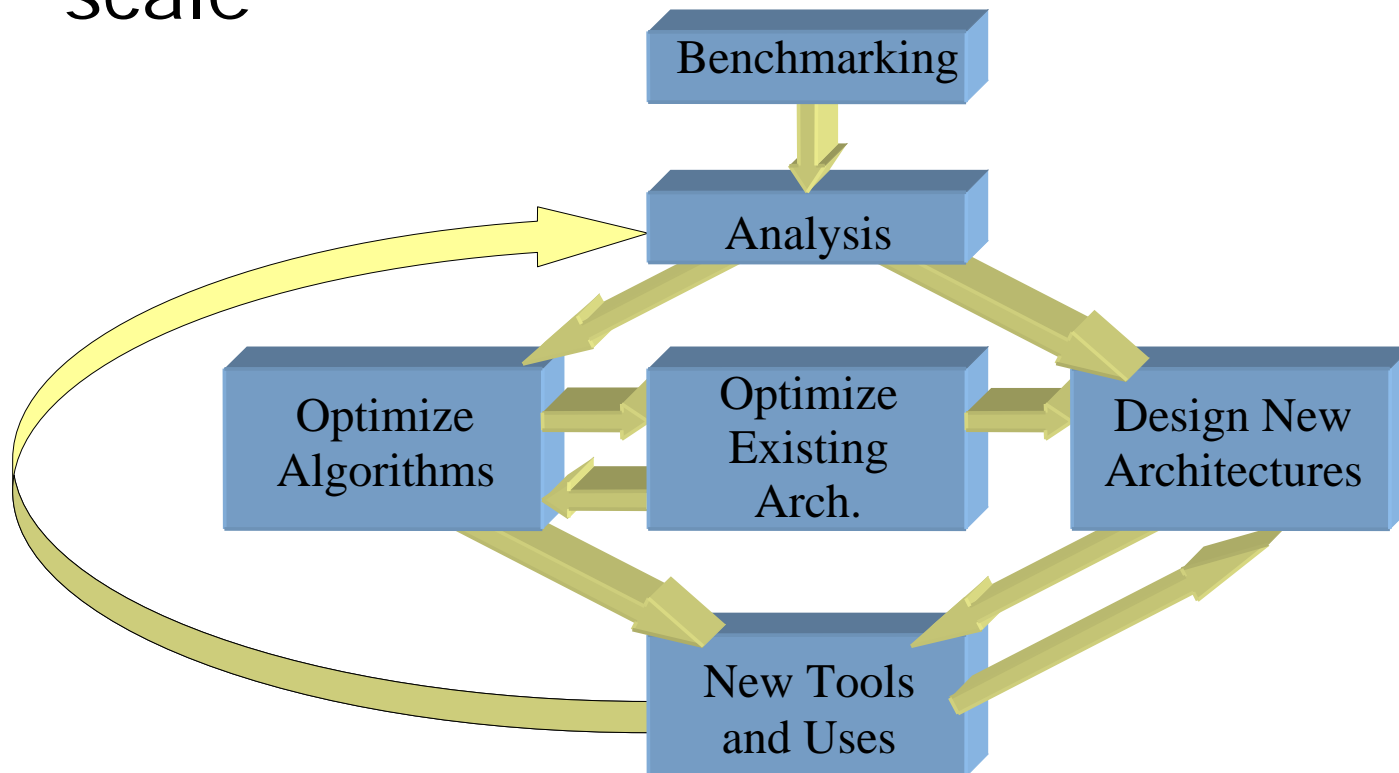


Data Mining Process



Motivation for Benchmarking

- ❑ Explosion of Data
- ❑ Existing systems, algorithms unable to scale



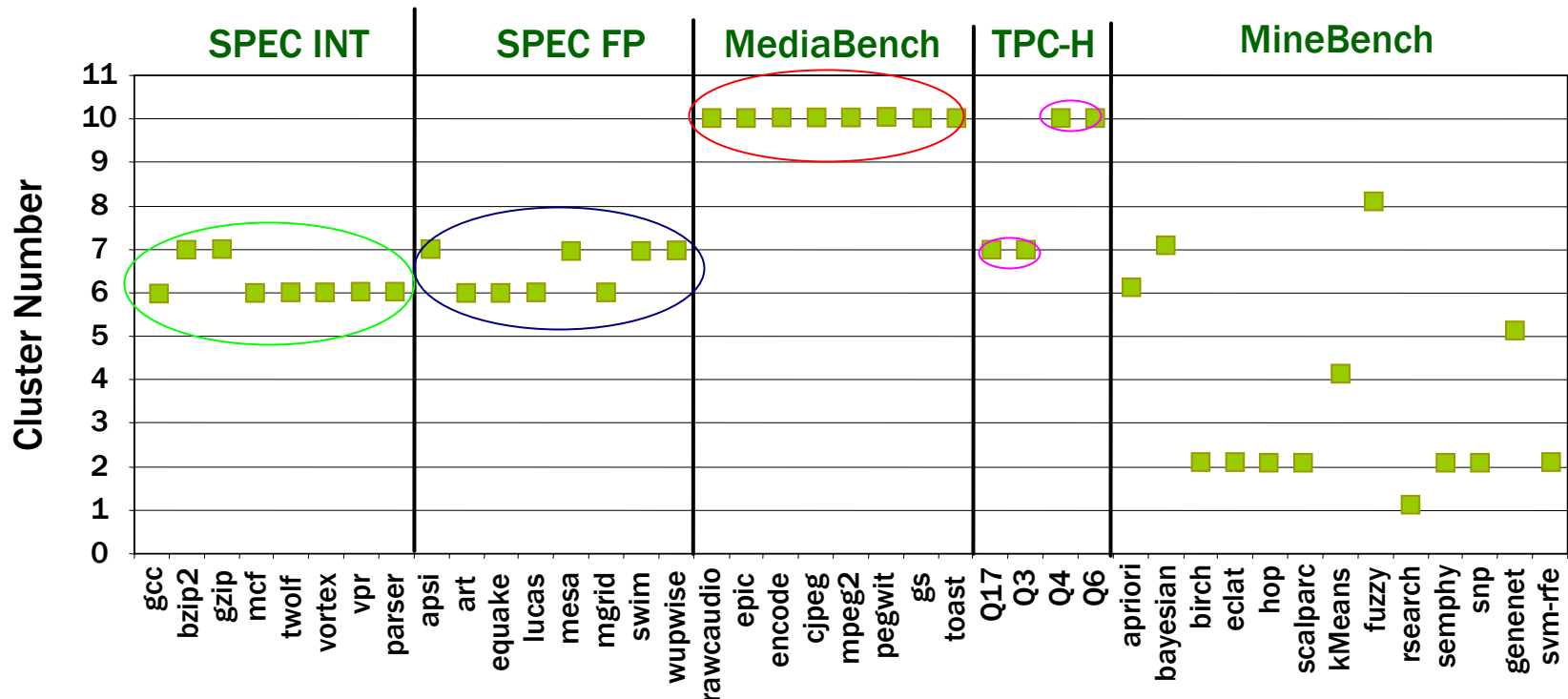
Role of Benchmarks in Systems Design

- For better or worse, benchmarks shape the corresponding industry segment
- Benchmarks
 - guide the development of new processor architectures
 - used to measure the relative performance of different systems
 - Architectures/tools are tailored for them
- **SPEC**: General purpose architecture
“Advances in the microprocessor industry would not have been possible without the SPEC benchmarks” - David Patterson
- **TPC**: Database Systems, **SPLASH**: Parallel machine architectures, **Mediabench**: Media and Communication Processors, **NetBench**: Network/Embedded processors

What makes data mining different ?

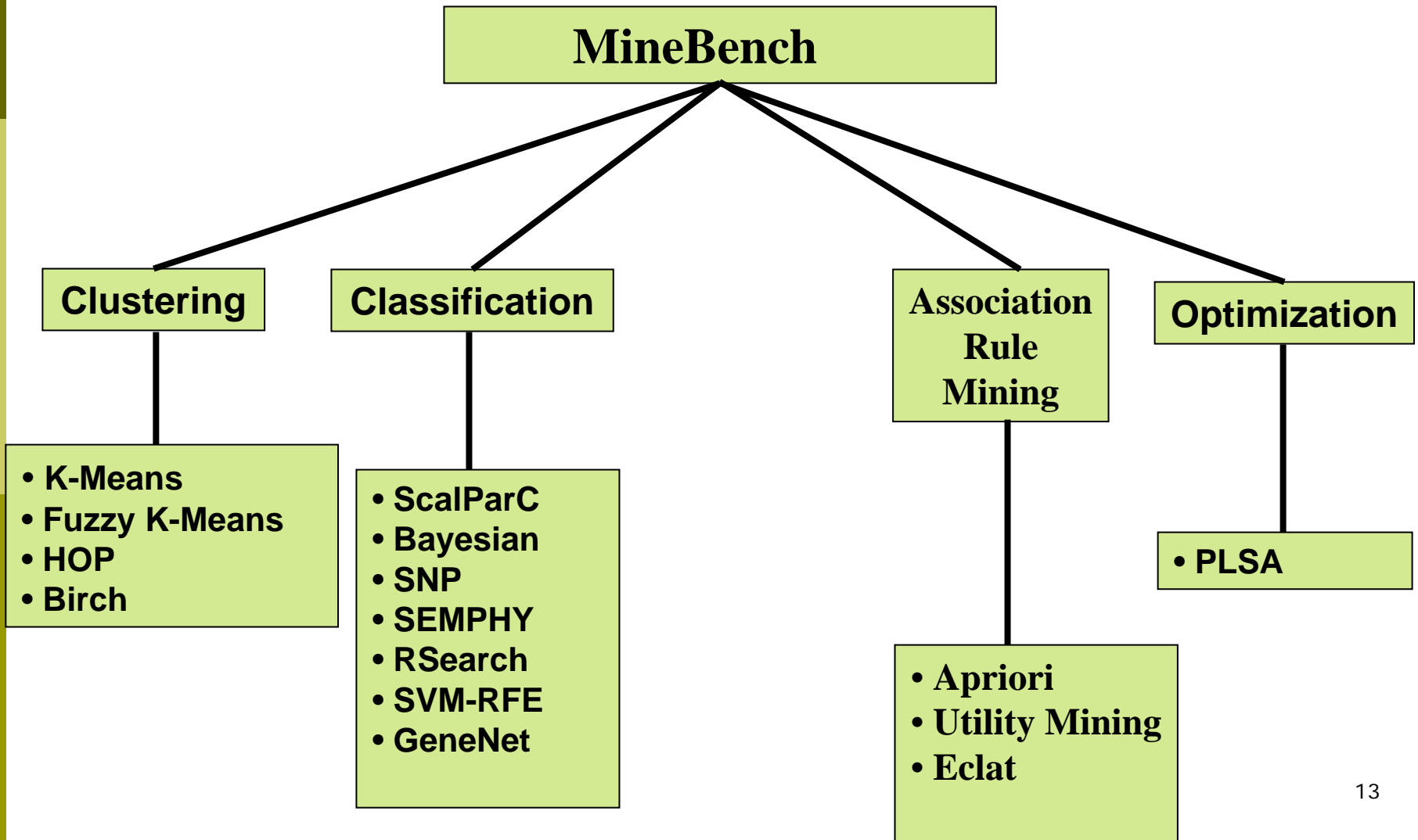
- Are data mining applications different ?
- If so, what characteristics do they exhibit ?

Need for a Data mining Benchmark



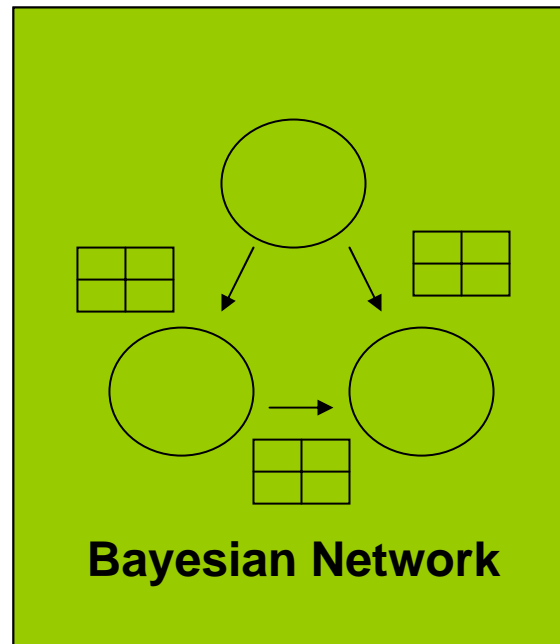
Reference: [Ozisikyilmaz B., Narayanan R., Zambreno J., Memik G., Choudhary A., IISWC 2006]

Benchmark Suite Overview

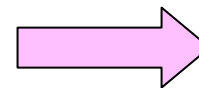
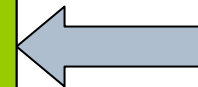


Classification Workloads

<i>Rid</i>	<i>Age</i>	<i>Income</i>	<i>Location</i>	<i>...</i>	<i>Offer</i>
A1	35	\$50000	CA	...	NO
A2	34	\$90000	IL	...	YES
...	



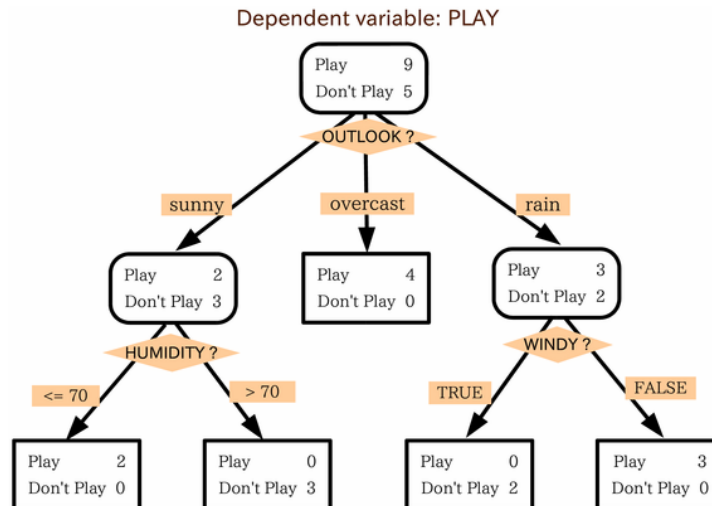
A3 39 \$60000 NY ...



Yes !

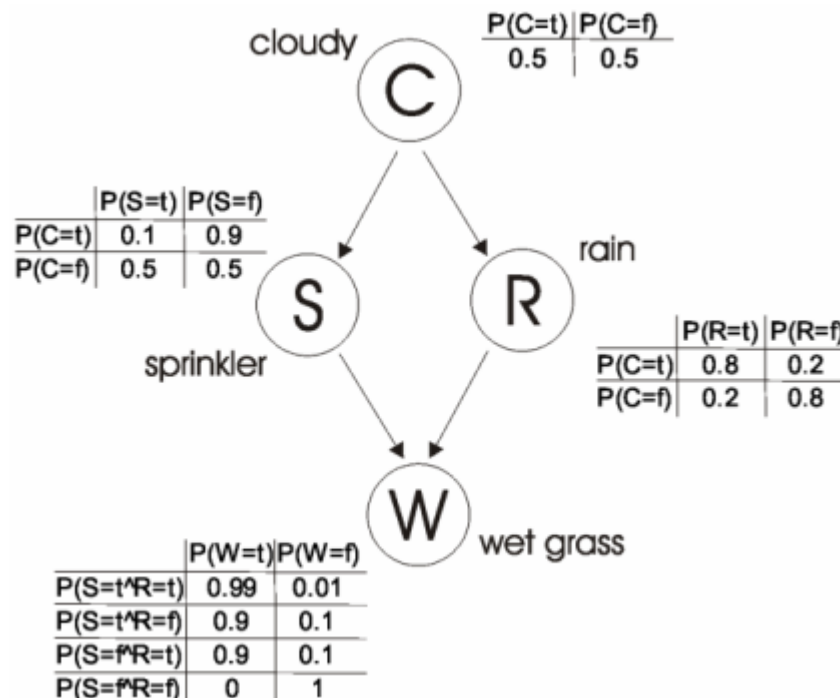
Classification Workloads

- ❑ A Classification algorithm uses a *training set* of records to build a *model*, which can be used to assign unclassified records into pre-defined classes
- ❑ ScalParC: Efficient and Scalable implementation of Decision Tree Classification
- ❑ Dataset: Synthetic dataset generated by IBM Quest Data Generator



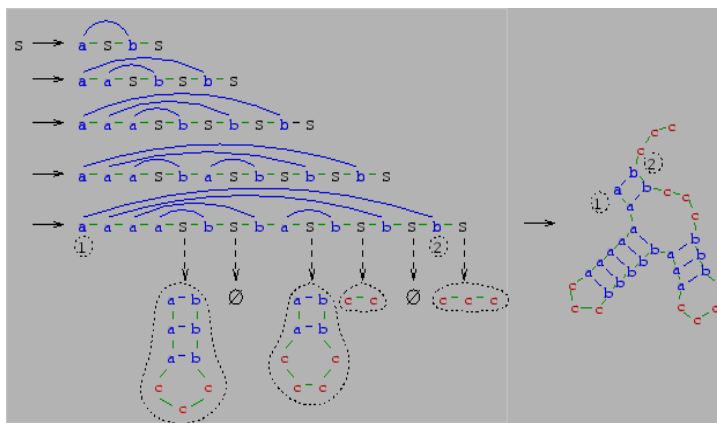
Classification Workloads

- Naïve Bayesian: Statistical Classifier, used in e-mail filters
- Dataset: Synthetic Dataset generated by IBM Quest Data Generator

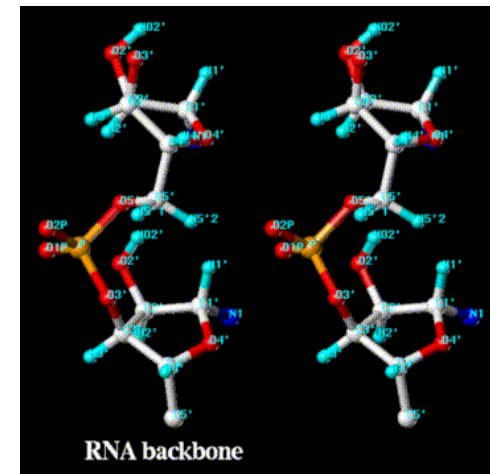


Classification Workloads

- ❑ Rsearch: Uses stochastic context free grammars to search gene data base for homologous RNA sequences
- ❑ Dataset: RNA sequence of length 97 on a database provided by UW



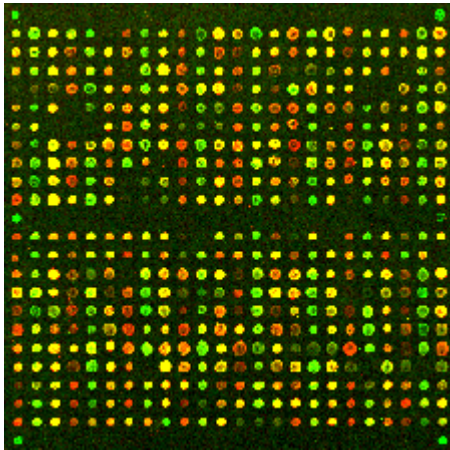
SCFG



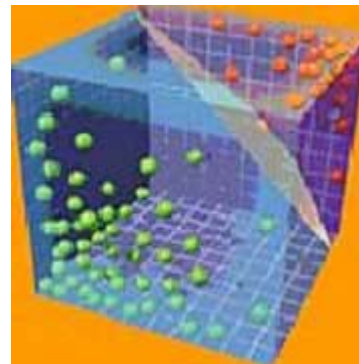
RNA

Classification Workloads

- ❑ SVM-RFE: is a feature selection method extensively used in disease finding
- ❑ Dataset: Microarray dataset on ovarian cancer [253(tissue samples)X15154(gene expression values)]



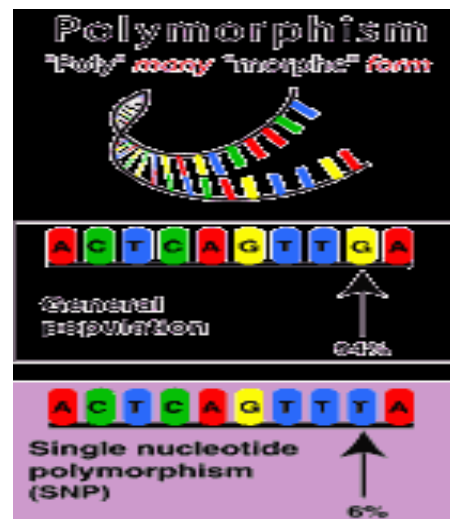
Microarray Sample



Support Vector Machine

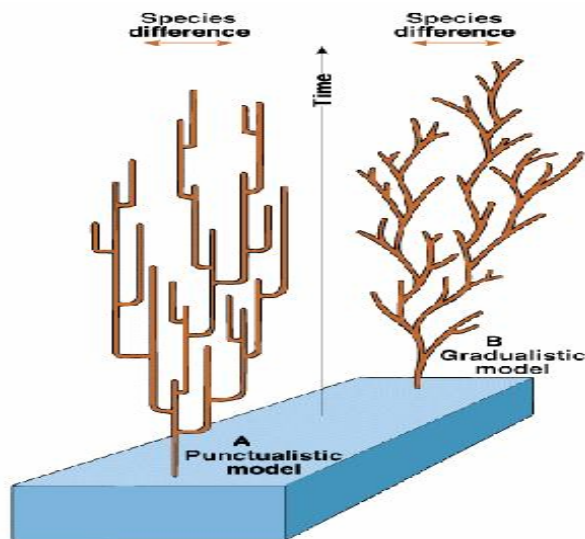
Classification Workloads

- ❑ SNP: uses Hill climbing method and Bayesian networks to find Single nucleotide polymorphisms in DNA sequences
- ❑ Dataset: Human Genic Bi-Allelic (HGBASE) database containing 616,179 SNP's



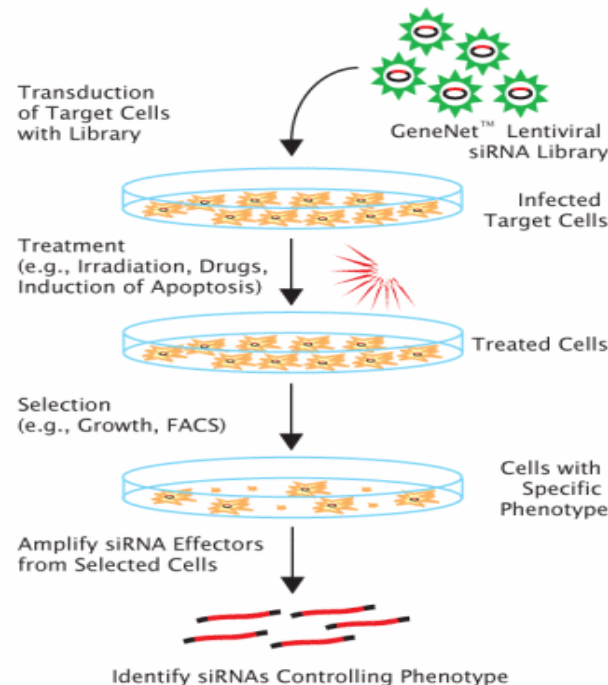
Classification Workloads

- Structure learning workloads
- SEMPHY: Uses Expectation Maximization method on phylogenetic trees to find genetic distance between neighboring species
- Dataset: 3 different datasets from Pfam database



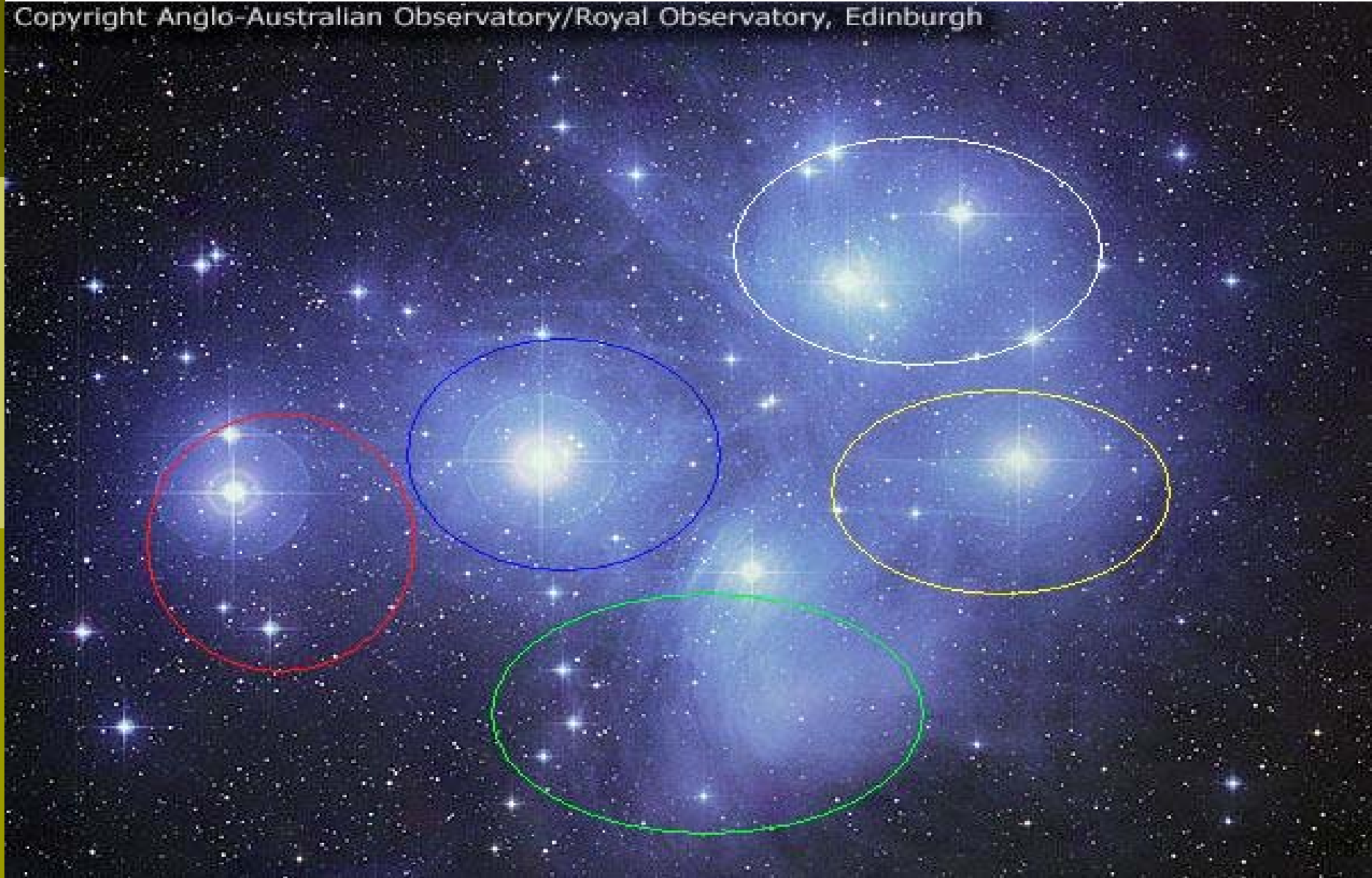
Classification Workloads

- GeneNet: Analysis of gene expressions using Bayesian networks and hill climbing(like SNP), handles thousands of variables with few training records
- Dataset: Yeast Microarray data



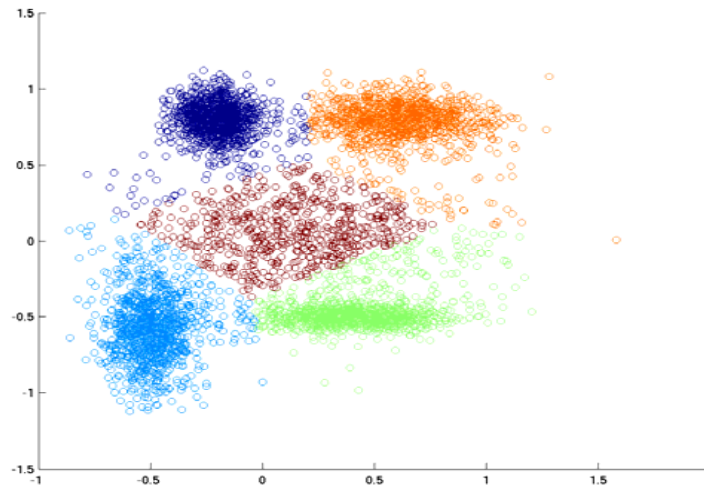
Clustering Workloads

Copyright Anglo-Australian Observatory/Royal Observatory, Edinburgh



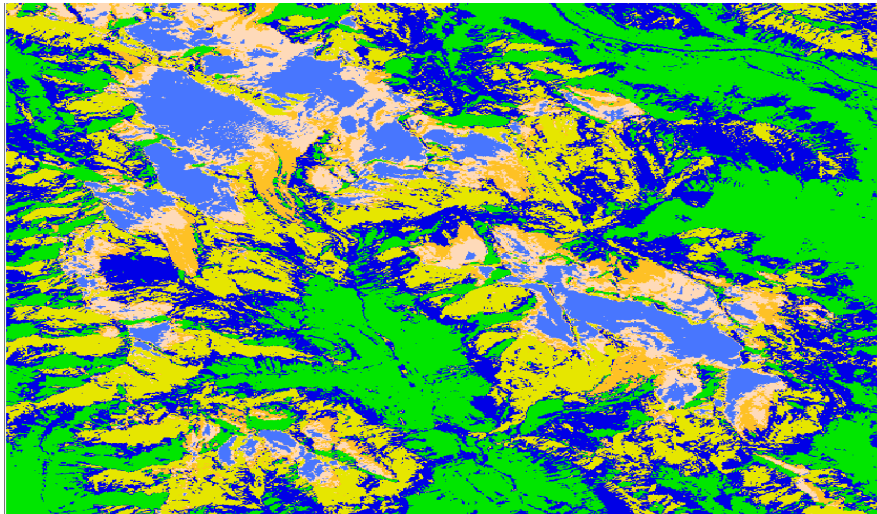
Clustering Workloads

- ❑ Clustering is a form of unsupervised learning that aims to find groups of similar objects from a database
- ❑ K-means: Assigns objects to clusters based on a similarity function, iteratively refines the cluster till convergence criterion is met
- ❑ Dataset: Real Image Database



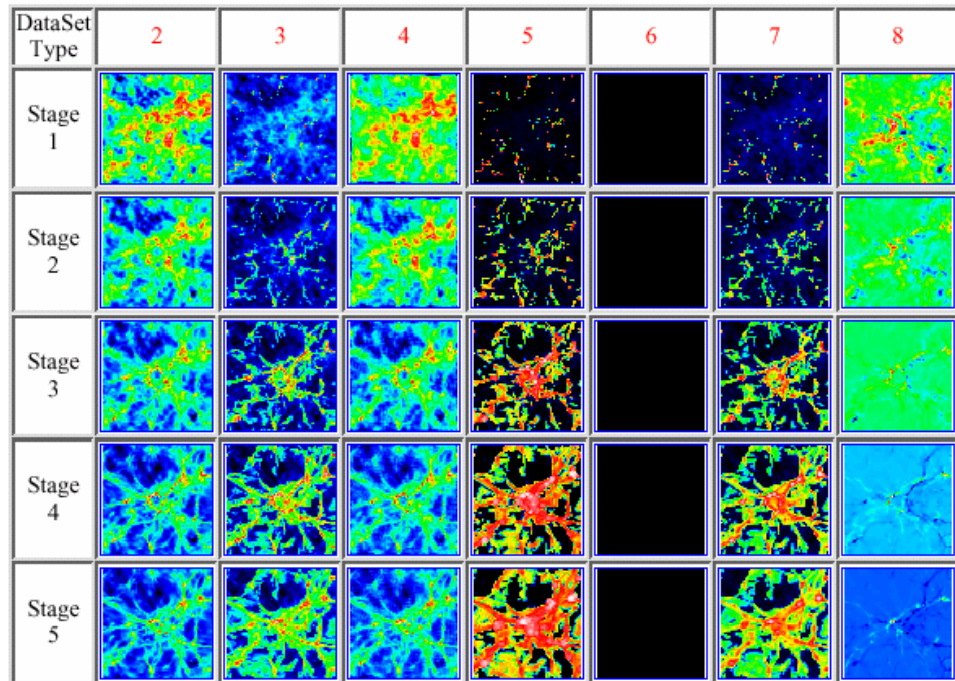
Clustering Workloads

- ❑ Fuzzy K-means: Similar to k-means, except that now an object may have degrees of membership in multiple clusters
- ❑ Dataset: Real Image Database



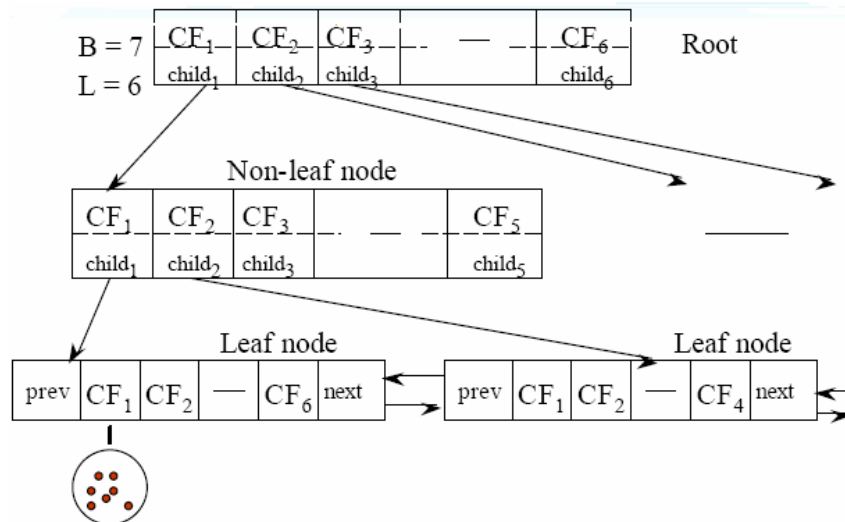
Clustering Workloads

- ❑ HOP: Density-based clustering method used in Astrophysics. Spatially adaptive, co-ordinate free, uses KD tree for load distribution
- ❑ Dataset: Cosmology simulation data using ENZO



Clustering Workloads

- ❑ Birch: Incremental and hierarchical clustering algorithm
- ❑ Based on the notion of Clustering Feature(CF) and Clustering Tree
- ❑ Works with different distance metrics, Effective with outliers
- ❑ Dataset: Cosmology data from ENZO




Association Rule Mining

1. Market-basket transactions

<i>TID</i>	<i>Items</i>
1	Bread, Diaper, Milk
2	Beer, Diaper, Bread, Eggs
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Bread, Diaper, Milk

2. Find item combinations (itemsets) that occur frequently in data



Item Combination	Count
Bread	4
Coke	2
Milk	4
...	...
Bread & Coke	1
Bread & Milk	3
...	...
Bread & Milk & Diaper	3
...	...

3. Generate association rules

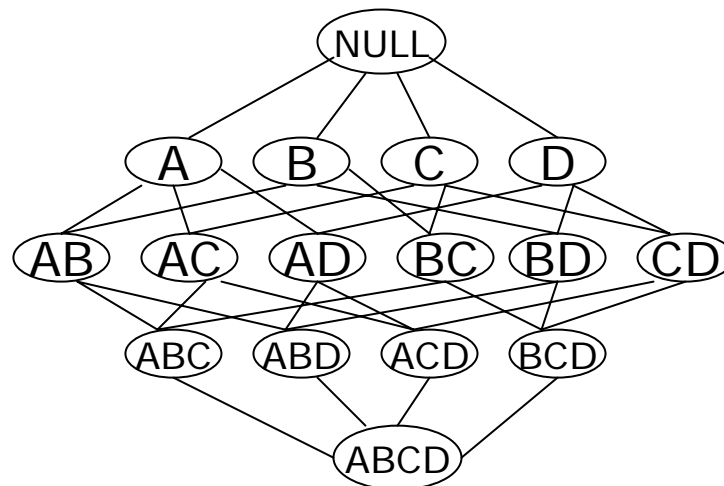


$\{\text{Diaper, Milk}\} \Rightarrow \{\text{Beer}\}$

$\{\text{Bread}\} \Rightarrow \{\text{Diaper}\}$

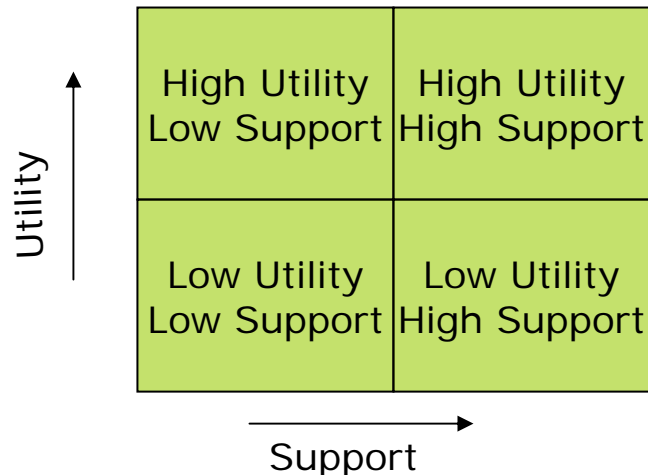
Association Rule Mining Workloads

- ❑ Apriori: Mines commonly occurring subsets of items(itemsets) level wise by using support-based pruning to systematically reduce the search space
- ❑ Dataset: IBM Quest data generator used to create datasets with varying number of transactions, average transaction size and maximum size of frequent itemsets



Association Rule Mining Workloads

- ❑ Utility Mining assumption of uniformity among items is discarded. Uses a 2 phase algorithm to find “utility” itemsets by considering different values of different items
- ❑ Certain items may be more important than others, allows organizations to segment customers into profit-based groups
- ❑ Eclat mines Association rules using a vertical database format
- ❑ Dataset: Supermarket transaction data



Optimization Workloads

- Sequence mining is used to find functional and evolutionary differences among sequences in bioinformatics
- PLSA uses a dynamic programming approach to find optimal alignments between RNA, DNA or protein primary sequences
- PLSA: Nucleotide sequences of varying length are used as input sequences (30K to 900K)

$$S(x, x) = 0$$

$$S(x, y) = -2$$

$$S(x, _) = -1$$

$$S(_, x) = -1$$

A	C	T	T	G	T	A	G	G	A
A	I	G	G	A	G	A	G	A	A

Score = -8

Alignment

Score Matrix

Using MineBench

- ▣ Applications written using C/C++
- ▣ Parallelized using OpenMP
- ▣ Current release uses GCC v2.96 for serial workloads and ICC v7.0 for parallel workloads
- ▣ Compatible with several compilers by making minor modifications
- ▣ Runtimes obtained using a dual processor Intel Xeon 2.8GHz system on medium sized datasets

Compiler Compatibility

Application	ICC 7.0	ICC 8.1	ICC 9.1	GCC 2.96	GCC 3.2	GCC 3.4.5
ScalParC	55.49	56.30	56.54	68.07	52.21	52.60
Naïve Bayesian	26.50	69.85	25.59	24.48	72.63	24.60
K-means	29.70	81.10	29.24	29.4	30.46	32.70
Fuzzy K-means	962.40	1375.42	625.86	949.68	938.21	937.88
HOP	26.47	29.53	27.11	33.16	31.85	31.00
Birch	C.E.*	C.E.	C.E.	15.86	C.E	C.E
Eclat	C.E.*	92.63	30.50	33.96	113.59	33.86
Apriori	55.53	98.53	54.64	53.42	74.24	51.64

Runtimes of MineBench applications with different compilers (*in seconds*)

32

***C.E.: Compilation Error**

Compiler Compatibility

Application	ICC 7.0	ICC 8.1	ICC 9.1	GCC 2.96	GCC 3.2	GCC 3.4.5
Utility	8.65	9.14	6.16	9.52	8.47	8.37
SNP	855.41	995.98	C.E.*	C.E.	C.E.	C.E.
GeneNet	1166.29	698.23	C.E.	C.E.	C.E.	663.5**
SEMPHY	880.40	1422.31	1423.50	C.E.	1356.27	1234.52
Rsearch	742.81	714.65	835.28	1473.70	1546.06	1831.15
SVM-RFE	51.69	44.57	40.74	C.E.	45.65	40.37
PLSA	1648.6	2295.68	3035.40	3045.05	1716.62	1733.29

Runtimes of MineBench applications with different compilers (*in seconds*)

33

***C.E.: Compilation Error**

****Compiled using gcc 3.4.4**

Conclusions

- ❑ As data sizes exponentially increase, it is essential to apply data mining techniques to extract knowledge
- ❑ Current data mining systems and algorithms do not scale → Need for design and evaluation of systems optimized for data mining
- ❑ Uniqueness of data mining applications necessitates a new benchmark
- ❑ MineBench: A data mining benchmark containing several representative applications
- ❑ MineBench is freely available on <http://cucis.ece.northwestern.edu/projects/DMS>

MineBench Project Homepage

<http://cucis.ece.northwestern.edu/projects/DMS>

**CENTER FOR
ULTRA-SCALE
COMPUTING AND
INFORMATION
SECURITY**

[contact](#)

[publications](#)

[projects](#)

[members](#)

Sponsors:

- [National Science Foundation](#)
(grants CCF-0444405,
CNS-0406341,
CCR-0325207)
- [Department of Energy](#) (grant
DE-FC02-01ER25485)
- [Intel Corporation](#)

Duration:

January 2004 - Present

Project Team Members:

Northwestern University

- [Jay Pisharath](#)
- [Ying Liu](#)
- [Wei-keng Liao](#)
- [Gokhan Memik](#)
- [Alok Choudhary](#)

- [Project Goals](#) • [Methodology](#) • [Current Accomplishments](#) • [Publications](#) • [Talk Slides](#)
- [Downloads](#) •

Design, Development and Evaluation of High Performance Data Mining Systems

Project Goals:

With the enhanced features in recent computer systems, increasingly larger amounts of data are being accumulated in various fields. The collected data is growing exponentially every year, and it becomes increasingly necessary to use automated tools in order to extract precise and useful information from the collected data. Data mining is a powerful tool that enables one to achieve this. Data mining programs have become essential tools in many domains including business (marketing, customer relationship management, scoring and risk management, fraud detection), science (astrophysics, climate modeling, particle physics), biotechnology (understanding diseases, protein identification, drug discovery, personalized

Thank You

Email: ran310@eecs.northwestern.edu

Web: <http://www.ece.northwestern.edu/~ran310>

Project Web Page:

<http://cucis.ece.northwestern.edu/projects/DMS>