Some issues in data mining research
Một số vấn đề trong nghiên cứu về khai phá dữ liệu

Hồ Tú Bảo

Institute of Information Technology, CNST, Vietnam

Japan Advanced Institute of Science and Technology, Japan

(invited talk for the author's group B.H. Khang, L.C. Mai, H.T. Bao)

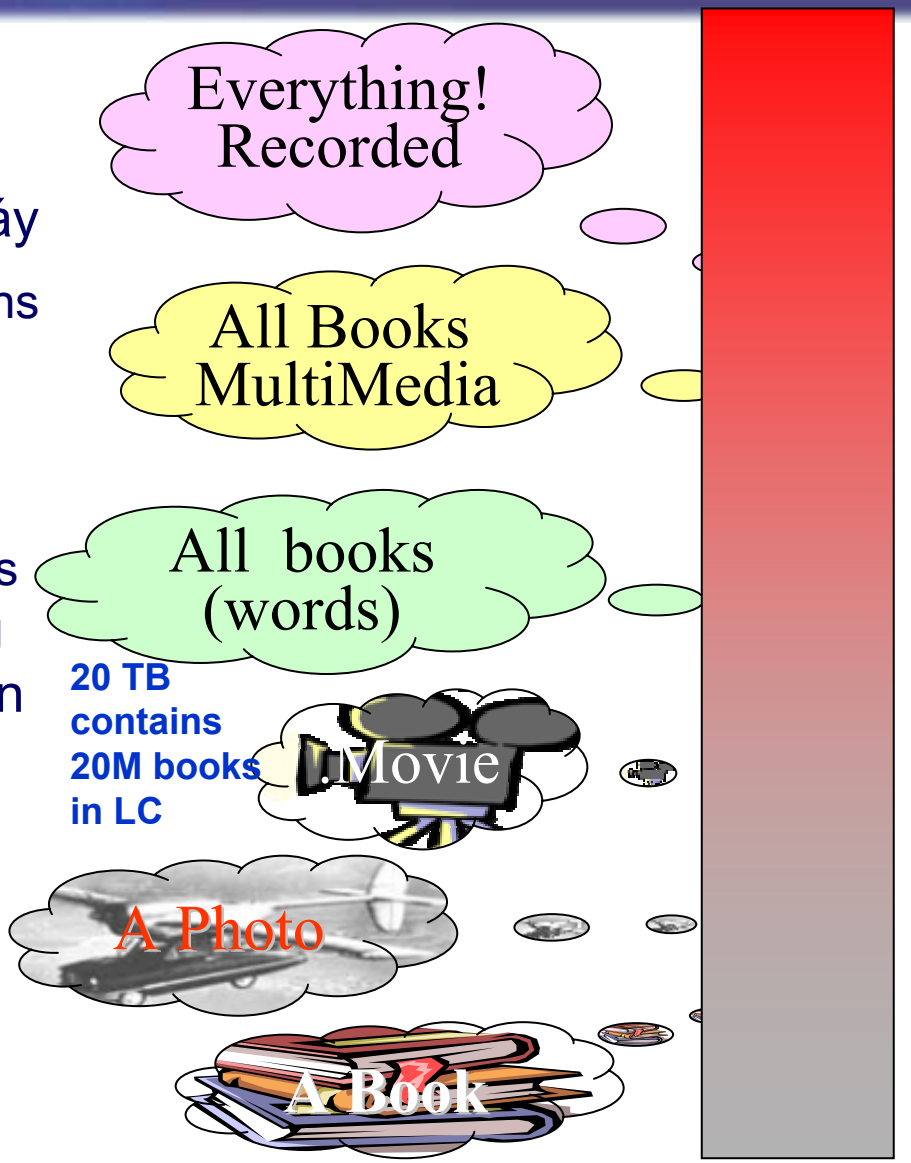
Outline

Notes on data mining

Some research issues

How much information is there?

- Soon everything can be recorded and indexed -- Mọi thứ sẽ sớm được lưu và chỉ số hóa trên máy
- Most bytes will never be seen by humans Hầu hết dữ liệu sẽ chẳng bao giờ được con người ngó ngàng
- Data summarization, trend detection anomaly detection are key technologies Tóm tắt dữ liệu, phát hiện xu hướng và bất thường là các công nghệ then chốt



See Mike Lesk:

How much information is there:
http://www.lesk.com/mlesk/ksg97/ksg.html

See Lyman & Varian:

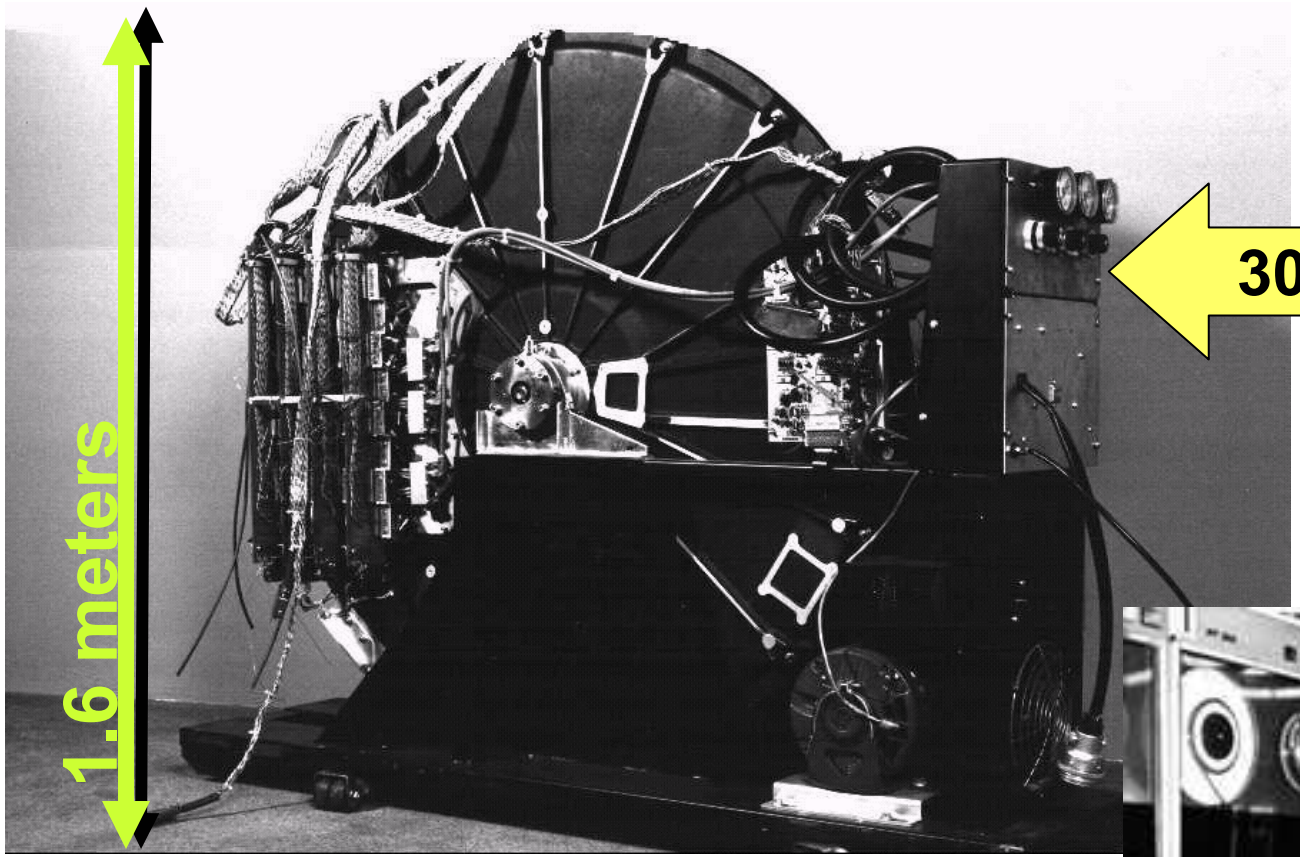
How much information

http://www.sims.berkeley.edu/research/projects/how-much-info/

First Disk 1956

- IBM 305 RAMAC
- 4 MB
- 50x24” disks
- 1200 rpm
- 100 ms access
- 35k$/y rent
- Included computer & accounting software
  (tubes not transistors)

10 years later

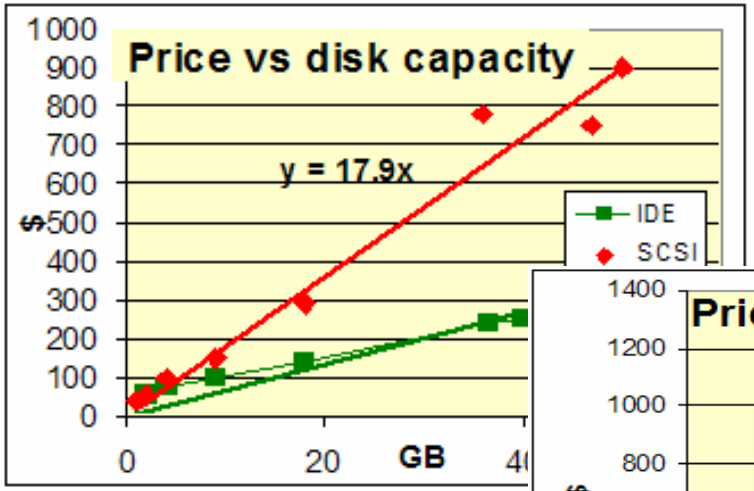


30 MB

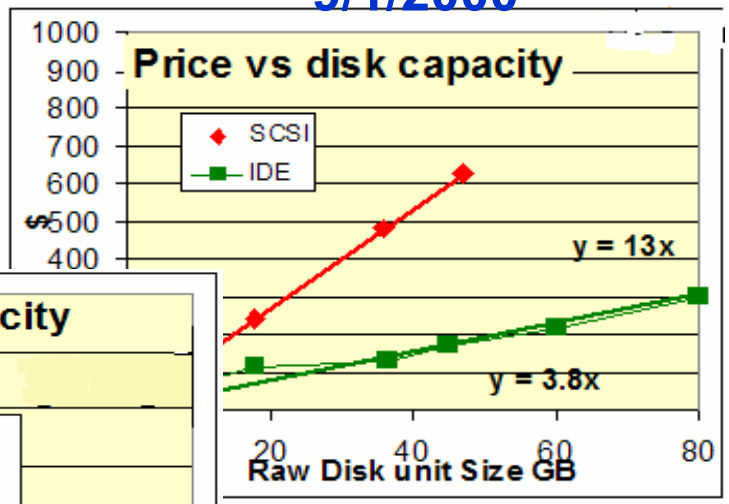
ODRA 1304

Price vs. Disk Capacity

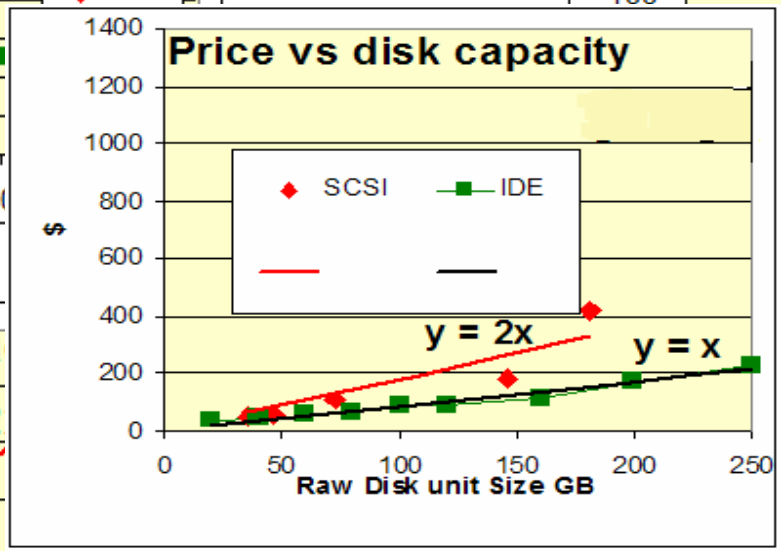
12/1/1999



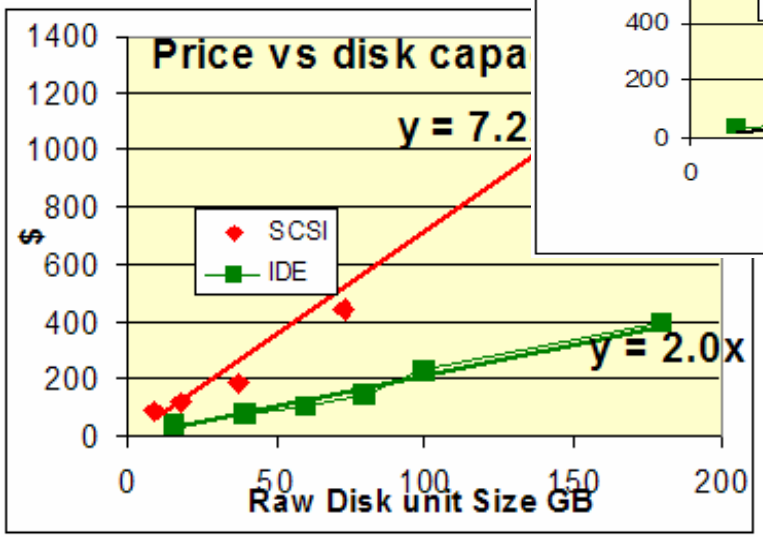
9/1/2000



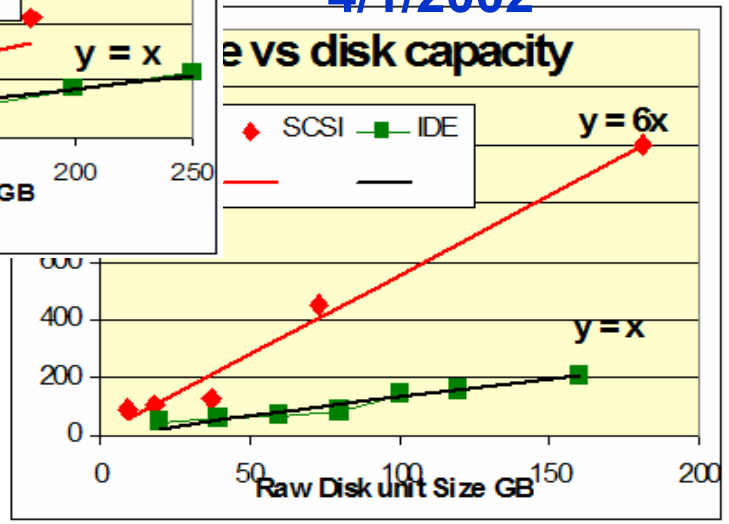
22/9/2003



9/1/2001



4/1/2002



Disk Storage Cheaper Than Paper

- File Cabinet:

Cabinet (4 drawer)	250$
Paper (24,000 sheets)	250$
Space (2x3 @ 10€/ft2)	180$
Total	700$

0.03 $/sheet

3 pennies per page

- Disk:

disk (250 GB =) 250$

ASCII: 100 m pages

2e-6 $/sheet(10,000x cheaper)

micro-dollar per page

Image: 1 m photos

3e-4 $/photo (100x cheaper)

milli-dollar per photo

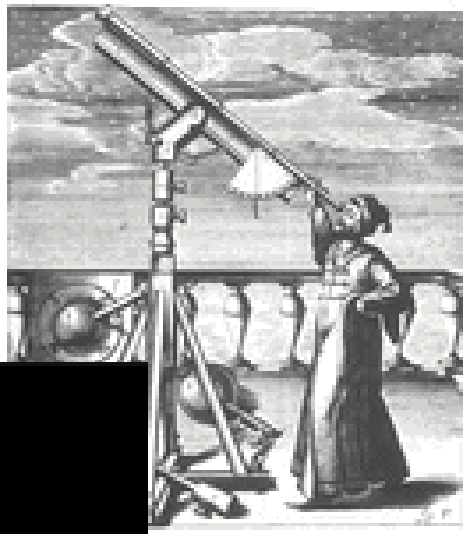
- Store everything on disk

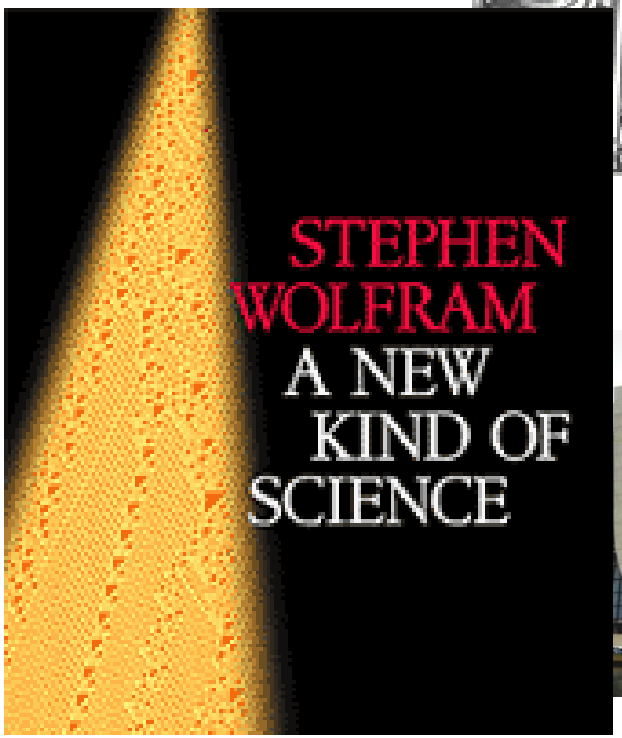
Note: Disk is 100x to 1000x cheaper than RAM

The Evolution of Science

- Observational Science Khoa học quan sát
  - ➔ Scientist gathers data by direct observation
  - ➔ Scientist analyzes data
- Analytical Science Khoa học phân tích
  - ➔ Scientist builds analytical model
  - ➔ Makes predictions.
- Computational Science Khoa học
  - ➔ Simulate analytical model
  - ➔ Validate model and makes predictions
- Data Exploration Science
  Khoa học khai thác dữ liệu

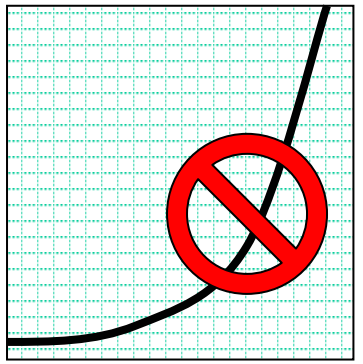
  Data captured by instruments
  Or data generated by simulator
  - ➔ Processed by software
  - ➔ Placed in a database / files
  - ➔ Scientist analyzes database / files



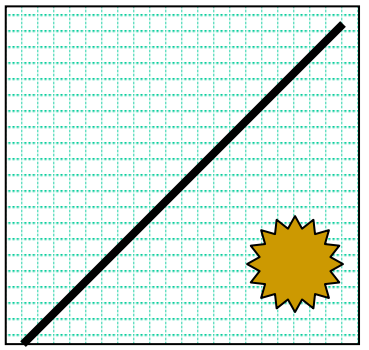


Organization & Algorithms

- Fast, approximate heuristic algorithms – Thuật toán heuristic xấp xỉ và nhanh
  - ➔ No need to be more accurate than data variance
  - ➔ Fast CMB analysis by Szapudi et al (2001)
    - ⇨ *NlogN* instead of N^3 ➔ 1 day instead of 10 million years



- Take cost of computation into account – Giá tính toán
  - ➔ Controlled level of accuracy
  - ➔ Best result in a given time, given our computing resources



- Use parallelism Dùng tính toán song song
  - ➔ Many disks
  - ➔ Many cpus

Polynomial time algorithms do not always work!

Historical Context: Statistics

- Gauss, Fisher, and all that
  - ➔ least-squares, maximum likelihood
  - ➔ development of fundamental principles
- The Mathematical Era Kỷ nguyên toán học
  - ➔ 1950's: The mathematicians take over
- The Computational Era Kỷ nguyên tính toán
  - ➔ steadily growing since the 1960's
  - ➔ 1970's: Exploratory Data Analysis, Bayesian estimation, flexible models, EM, etc.
  - ➔ a growing awareness of the computing power & role in data analysis

Historical Context: Statistics

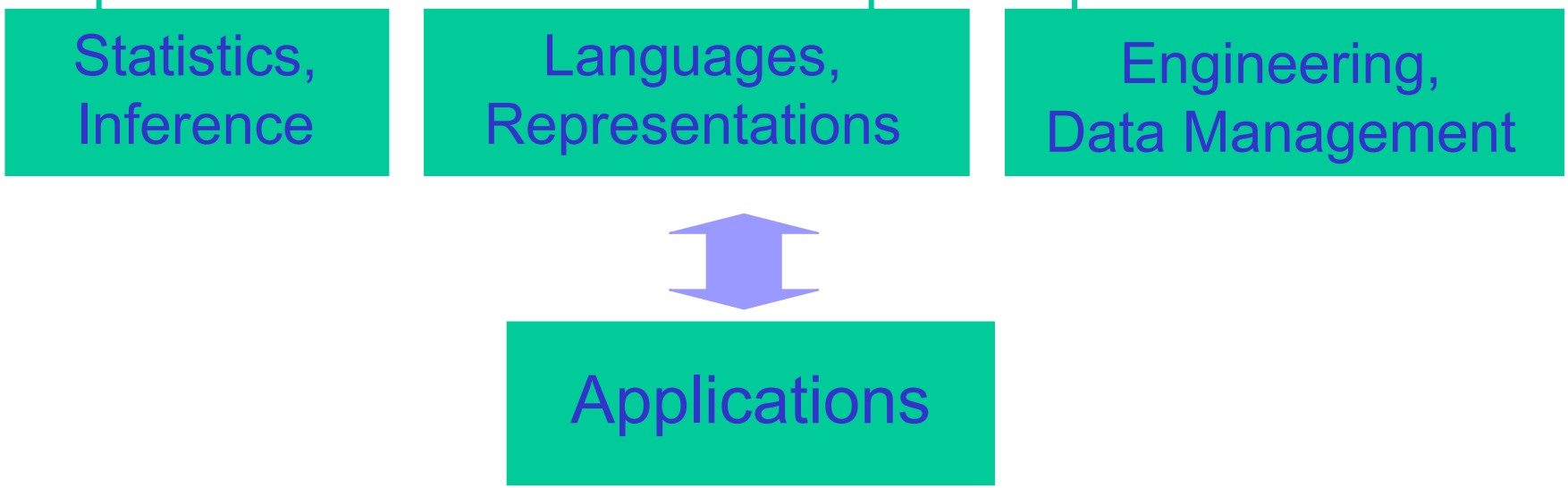
Objective and subjective probability XS chủ quan-khách quan

- **Frequentist view** (probability = limiting proportion of times that the event would occur in repetitions)
  - ➔ the dominant perspective throughout most of the last century, primarily of theoretical interest
  - ➔ it restricts our application of probability (cannot access the probability that Bùi Thị Nhung will jump 1.88m in Sea games 22)
- **Subjective view** (probability = individual degree of belief that a given event will occur)
  - ➔ Acquired increasing importance since last decade for data analysis
  - ➔ referred to as **Bayesian statistics**. A central tenet of Bayesian statistics is the explicit characterization of **all** forms of uncertainty, e.g., uncertainty about any parameters we estimate from the data.

What is Data Mining?

"Data-driven discovery of models and patterns from massive observational data sets"

Phát hiện các mô hình và mẫu dạng do khai phá các tập dữ liệu rất lớn

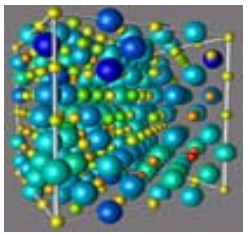


Data types vs. Mining methods

Types of data

- Flat data tables
- Relational database
- Temporal & Spatial
- Transactional databases
- Multimedia data
- Genome databases
- Materials science data
- Textual data
- Web data
- etc.

Different data schemas



Mining tasks and methods

- Classification/Prediction
  - ✓ Decision trees
  - ✓ Neural network
  - ✓ Rule induction
  - ✓ Support vector machine
  - ✓ Hidden Markov Model
  - ✓ etc.
- Description
  - ✓ Association analysis
  - ✓ Clustering
  - ✓ Summarization
  - ✓ etc.

Outline

Notes on data mining

Some research issues

Topics to address

- **Heterogeneity – Không đồng chủng**
  - ➔ Mixed data and multimedia data
  - ➔ Independent component analysis (ICA)
  - ➔ Some others (Kernel methods, Level sets, etc.)
- **Scaling up – Hợp với mọi kích cỡ (khả cỡ)**
  - ➔ Power search heuristics (e.g., K-means clustering)
  - ➔ Parallel computing (e.g., PC clusters)
- **Bioinformatics – Sinh tin học**
  - ➔ Protein structure prediction (SVM and HMM)
- **Materials structure analysis – Phân tích cấu trúc vật chất**
  - ➔ Crystal structure prediction
  - ➔ Mining structured data
- **Text and Web mining -- Khai phá dữ liệu văn bản và Web**
  - ➔ Mining structurally non-identical data

Numerical vs. Categorical Data

Combinatorial search in hypothesis spaces (machine learning)

Attribute	Numerical	Symbolic	
No structure $= \neq$		Places, Color	Nominal (categorical)
Ordinal structure $= \neq \geq$	Age, Temperature, Taste,	Rank, Resemblance	Ordinal
Ring structure $= \neq \geq + \times$	Income, Length		Measurable

Matrix-based computation (multivariate data analysis)

Example of a Scalable Algorithm

- Mixed Similarity Measures (MSM): **Độ đo sự giống nhau cho dữ liệu hỗn hợp**
  - ➔ Goodall (1966) time $O(n^3)$, Diday and Gowda (1992),
  - ➔ Ichino and Yaguchi (1994),
  - ➔ Li & Biswas (1997) Time $O(n^2 \log n^2)$, Space $O(n^2)$: $\hat{P}_{ij}$
- New and Efficient MSM (Binh & Bao, 2000): $\hat{P}^*_{ij}$
  - ➔ Time and Space O(n): $\hat{P}_{ij} = 1 - \hat{P}^*_{ij}$

Comparative Results

US Census database 33 sym + 8 num attributes, Alpha 21264, 500 MHz, RAM 2 GB, Solaris OS (Nguyen N.B. & Ho T.B., PKDD 2000)

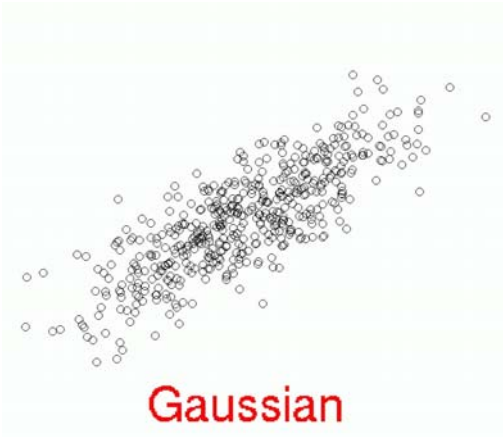
# cases	500 (0.2M)	1.000 (0.5M)	1.500 (0.9M)	2.000 (1.1M)	5.000 (2.6M)	10.000 (5.2M)	199.523 (102M)
# values	497	992	1.486	1.973	4.858	9.651	97.799
time of LiBis $O(n^2 \log n^2)$	67.3s	26m6.2	1h46m31s	6h59m45s	>60h	not app	not app
Time of OURS O(n)	0.1s	0.2s	0.3s	0.5s	2.8s	9.2s	36m26s
Memory of LiBis $O(n^2)$	5.3M	20.0M	44.0M	77.0M	455.0M	not app	not app
Memory of OURS O(n)	0.5 M	0.7M	0.9M	1.1M	2.1M	3.4M	64.0M
Preprocessing	0.1s	0.1s	0.2s	0.5s	0.9s	6.2s	127.2s

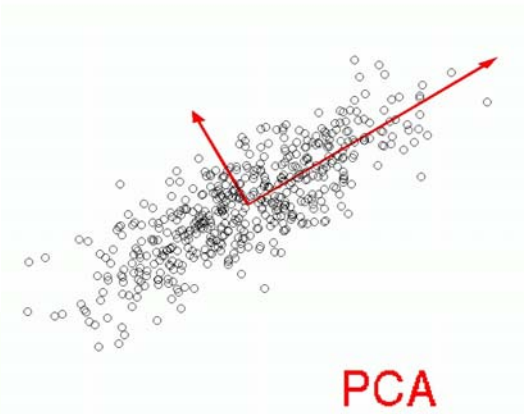
ICA vs. PCA

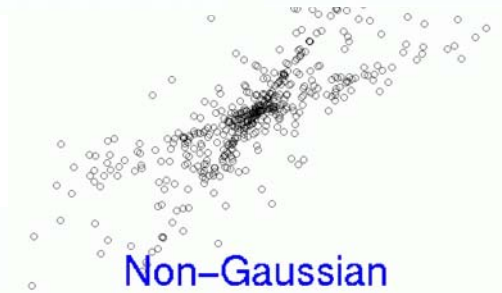
- Principal Component Analysis (PCA) finds directions of maximal variance (khác biệt cực đại) in Gaussian data (second-order statistics).

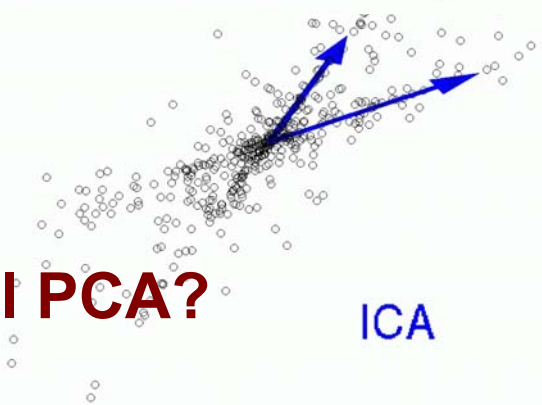
- Independent Component Analysis (ICA) finds directions of maximal independence (độc lập cực đại) in non-Gaussian data (higher-order statistics).

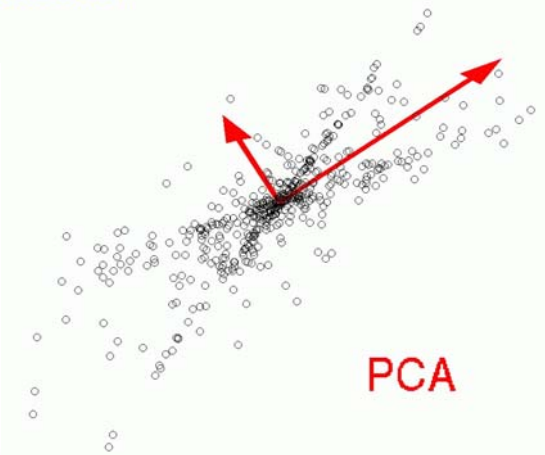
Challenge: Categorical PCA?





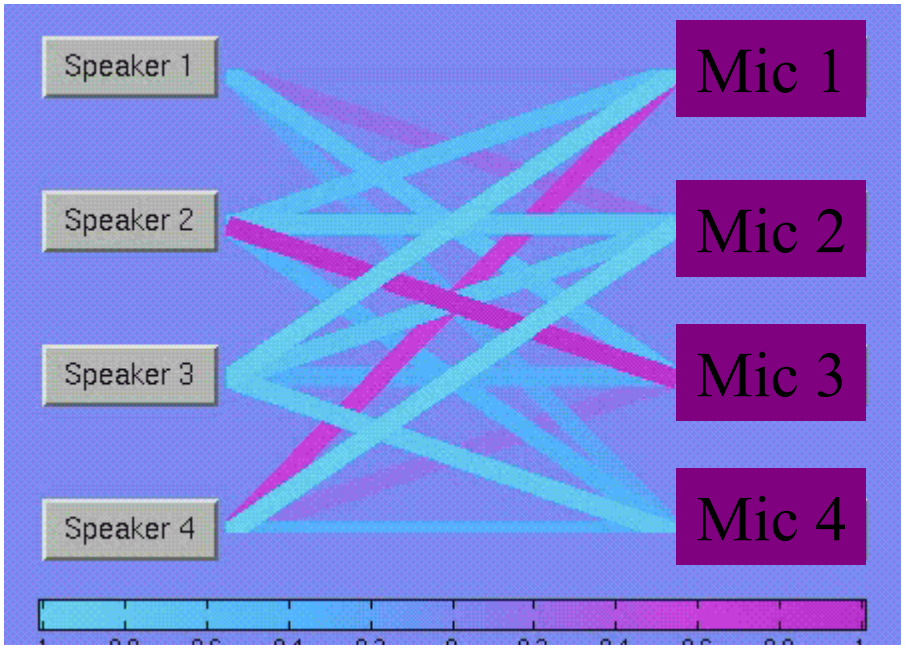




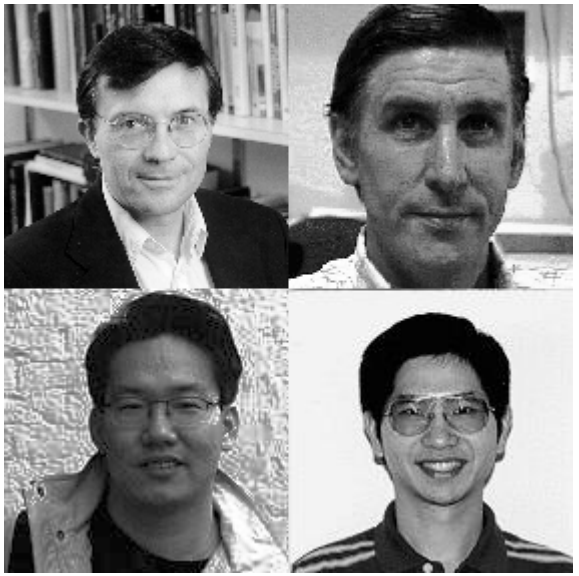


ICA: Example of Audio Decomposition

Perform ICA



Play Mixtures

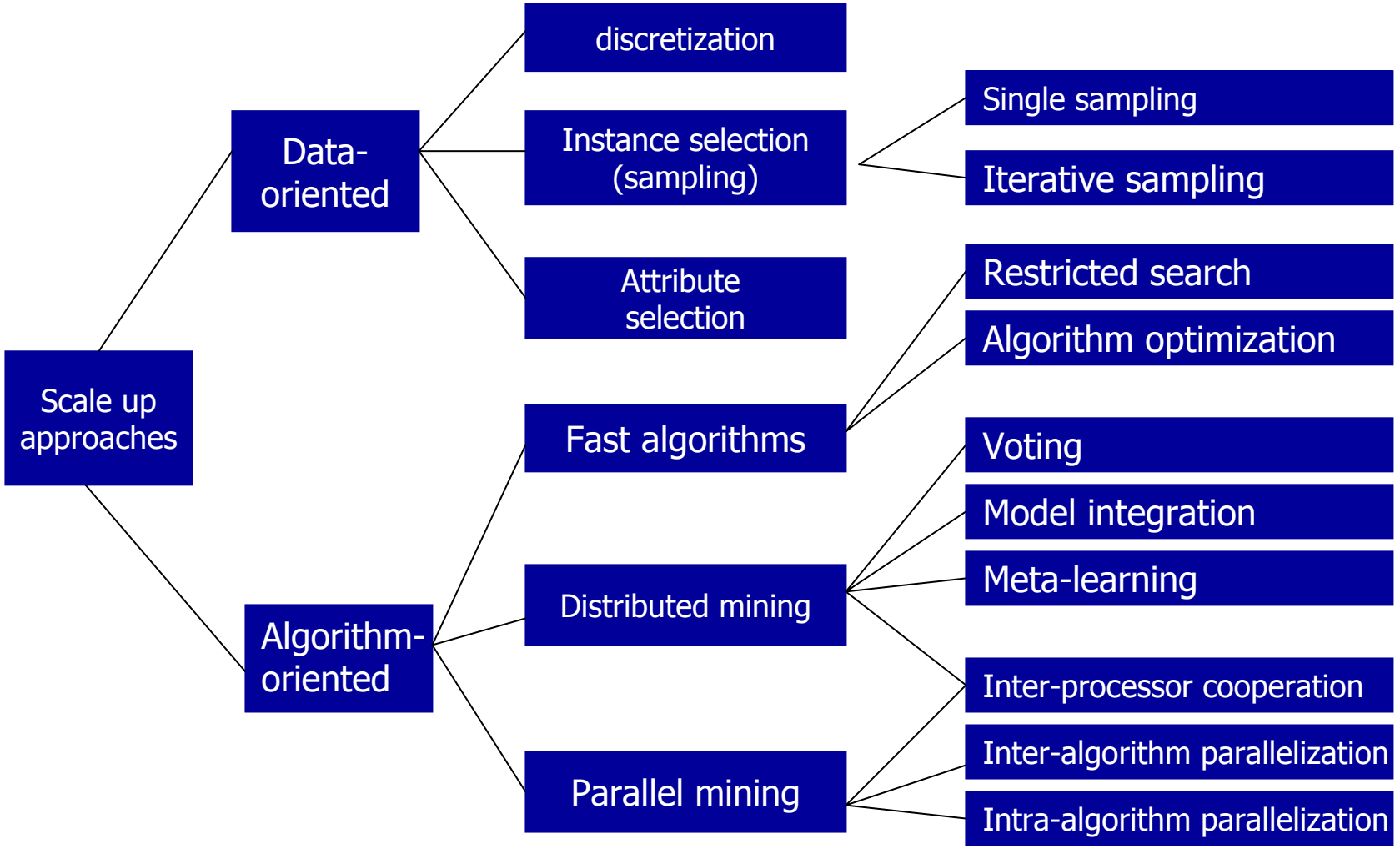


Terry Scott

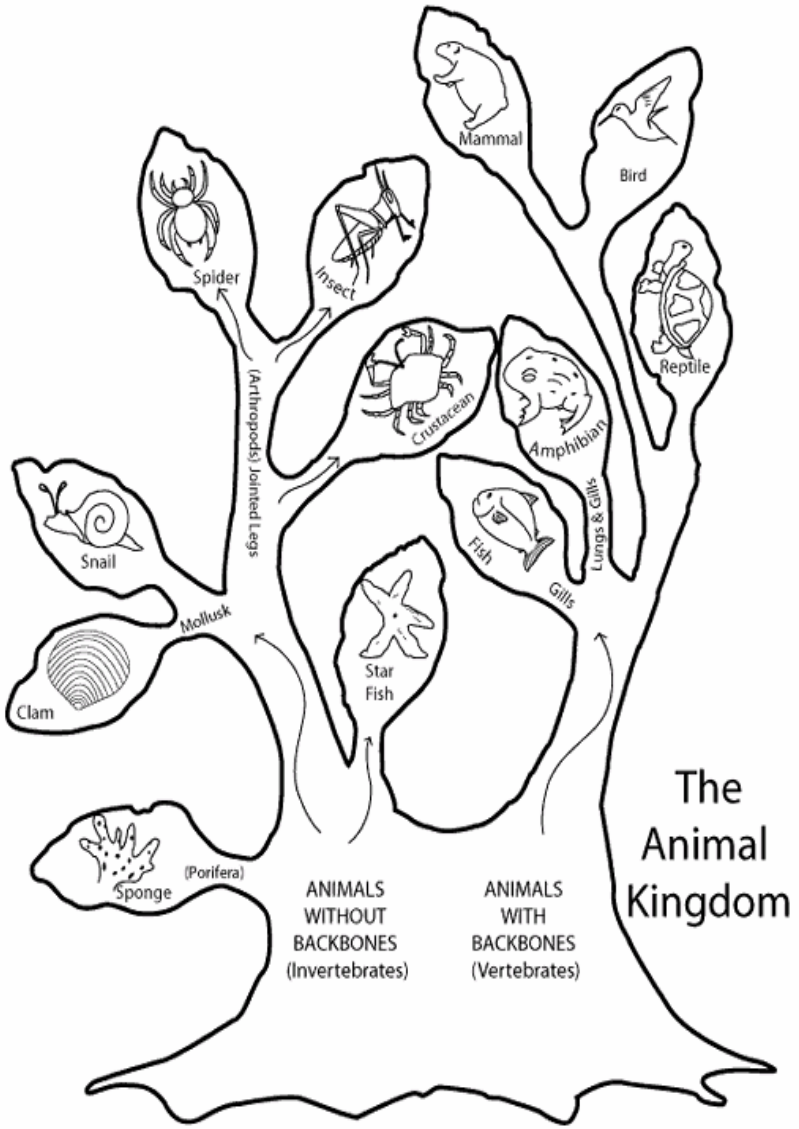
Te-Won Tzyy-Ping

Play Components

Scaling Up Approaches



Clustering Methods

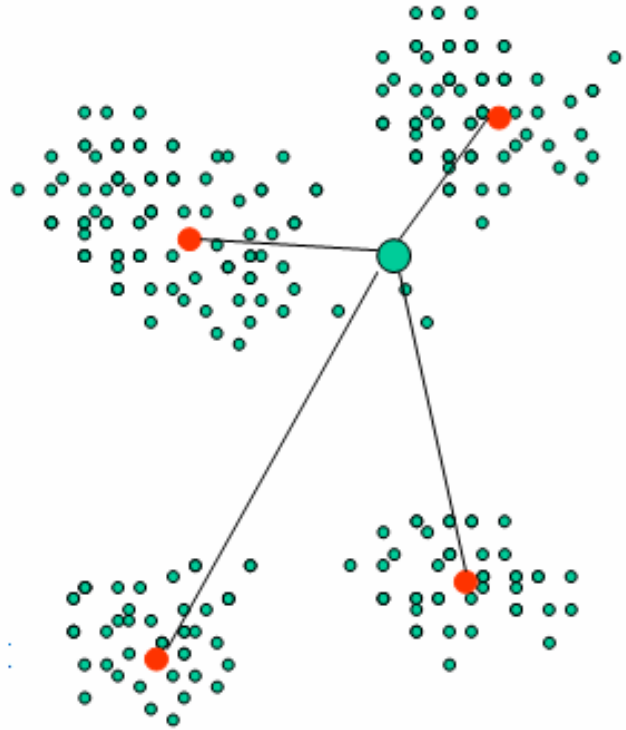


- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Model-Based Methods

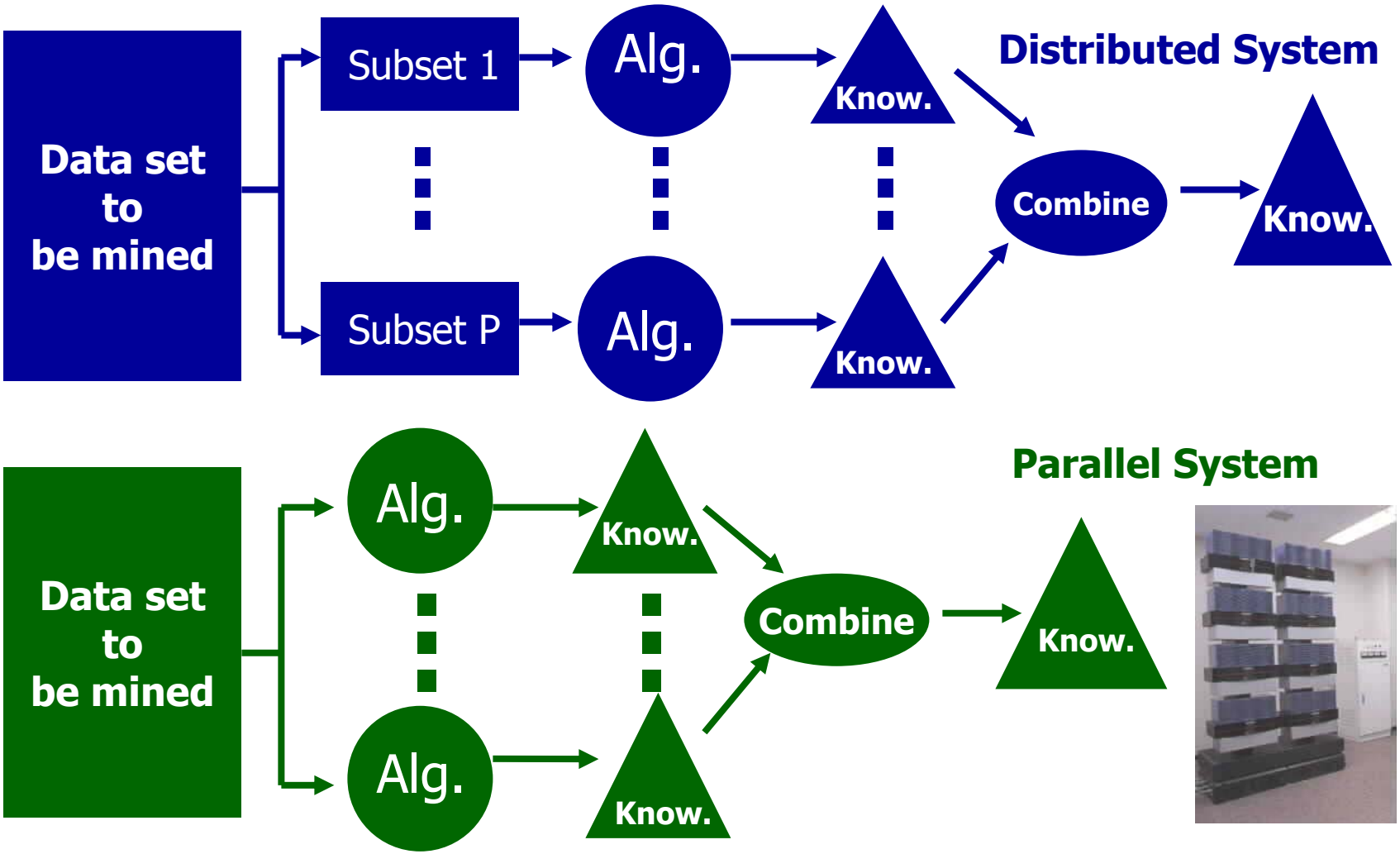
A key problem: Similarity between objects represented by non-standard data?

k-means: fast, faster, and fastest

- Work of Charles Elkan, ICML'03, 20-24/8/2003, "k-means: fast, faster, fastest"
- K-means đòi hỏi tính khoảng cách từ mỗi đói tượng đến **tất cả** tâm của các clusters ở mỗi bước lặp.
- Key idea: Các đối tượng chỉ có thể được phân vào một trong các tâm gần chúng → kiểm tra tính xa gần bằng **bất đẳng thức tam giác.**
- Greatly scaling up, says, when #instances = 10^6 and k = 10^3.
- Lesson: Các giải pháp hiệu quả thường đơn giản (và độc đáo)!



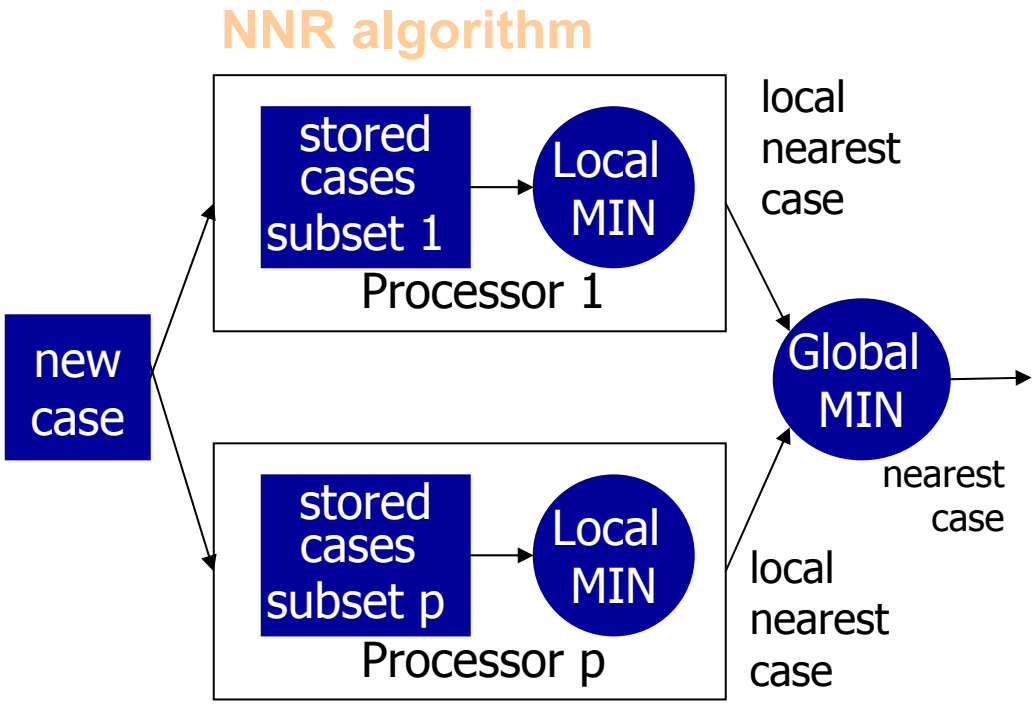
Distributed & Parallel Data Mining



Parallel Data Mining

My lab PC cluster

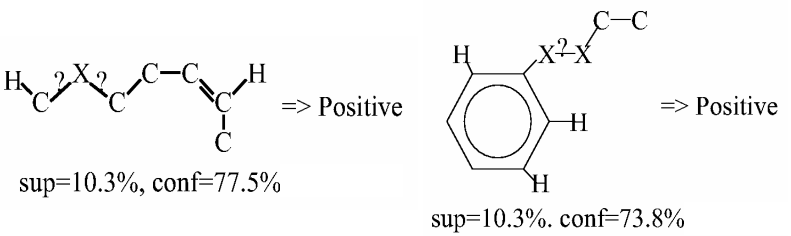
- 16 dual CPU nodes
  Intel Xeon 2.4 GHz
- About 1 billion VND



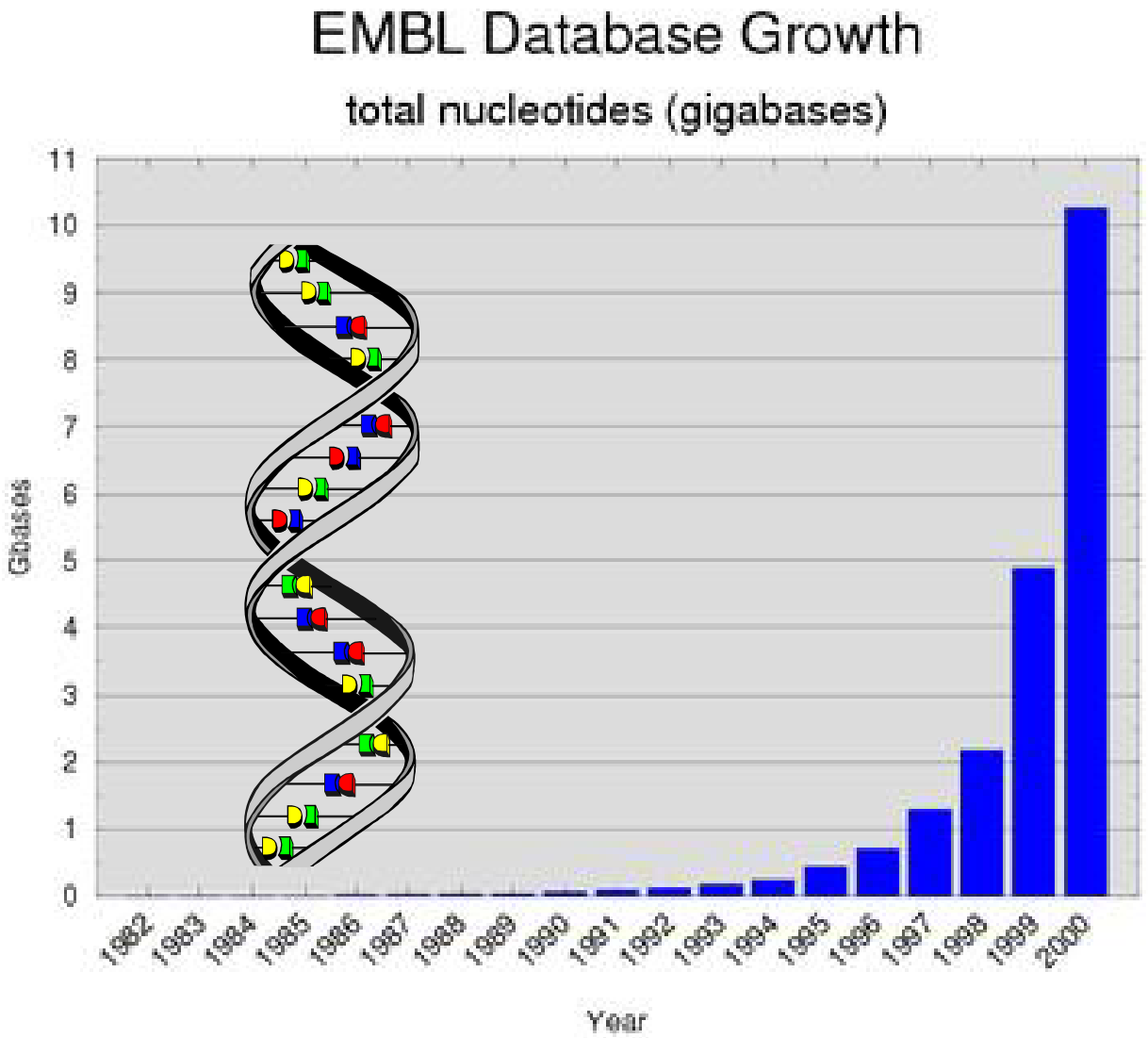
Example of exploiting data parallelism in instance-based learning

Mining Scientific Data

- **Data Mining in Bioinformatics**
  - ➔ β-turns prediction by SVM (P.T. Hoan)
  - ➔ Plant (rice) growth modeling (with L.M. Hoang): Alife + Genome data
- **Mining Physical and Chemical Data**
  - ➔ Crystal structure prediction (with D.H. Chi)
  - ➔ Molecular structure analysis (with N.T. Tai)
- **Mining Medical Data**
  - ➔ Stomach cancer and hepatitis
  - ➔ Temporal abstraction (with N.T. Dung, S. Kawasaki, L.S. Quang, N.D. Dung)



Base Pairs in GenBank



10,267,507,282 bases in 9,092,760 records.

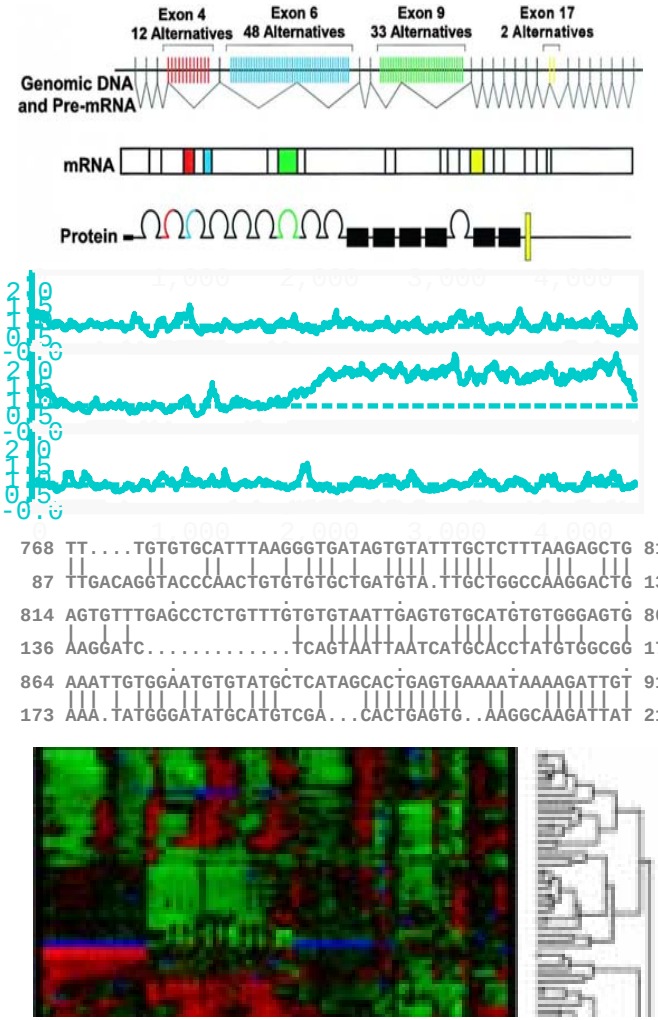
Problems in Bioinformatics

Structure analysis
- Protein structure comparison
- Protein structure prediction
- RNA structure modeling

Pathway analysis
- Metabolic pathway
- Regulatory networks

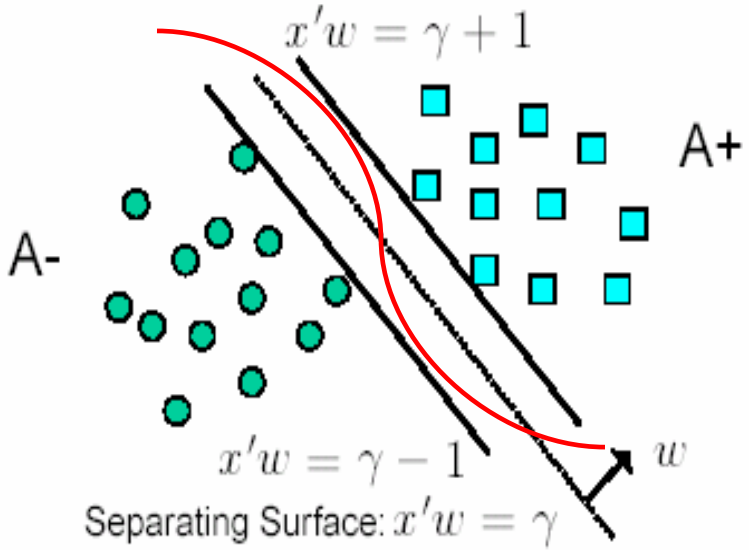
Sequence analysis
- Sequence alignment
- Structure and function prediction
- Gene finding

Expression analysis
- Gene expression analysis
- Gene clustering



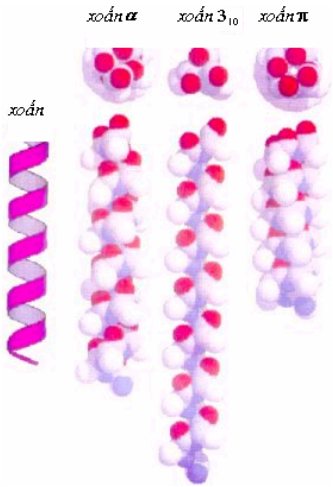
Support Vector Machines

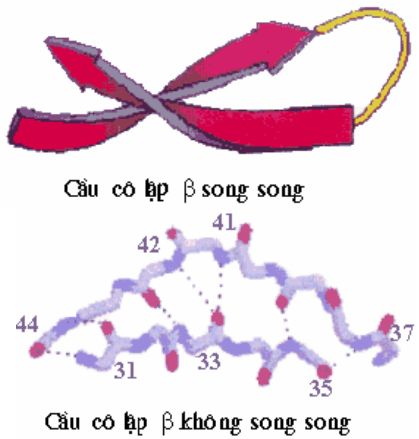
- Machine learning technique based on statistical learning theory (Vapnik, 1995)
- Find the separating surface that discriminates class A+ from class A- (**binary classifier**)
- Idea: The best learning can be achieved with the surface that maximizes "**margin**" determined by "**support vectors**".
- Data that are non-separable in N-dimensions have a higher chance of being separable if mapped into a space of higher dimension.

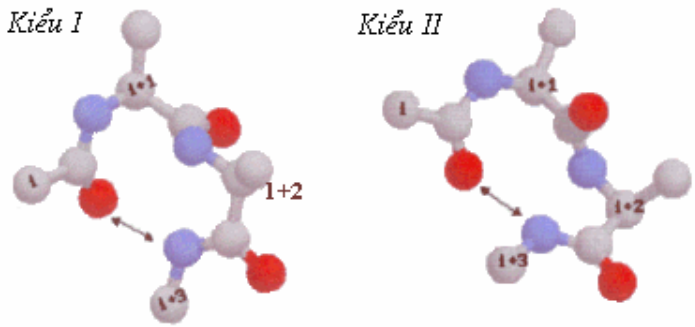


β-turns Prediction with SVM (P.T. Hoan)

Methods	Q_{total}	Q_{pred}	Q_{obs}	MCC
Chou-Fasman	74.9	46.1	16.9	0.16
Thornton	74.5	44.0	16.7	0.15
1-4 & 2-3 correlation model	63.2	35.3	60.4	0.21
Sequence couple model	50.5	31.7	88.4	0.23
BTPRED	73.5	47.2	64.3	0.37
SVM	**78.4**	**55.9**	**58.6**	**0.43**



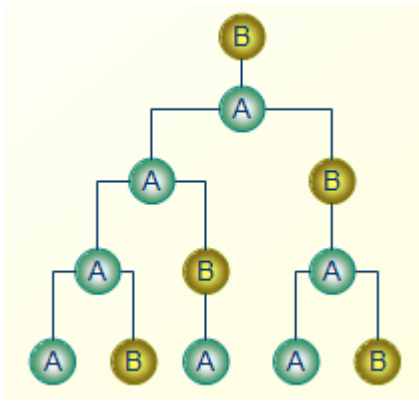


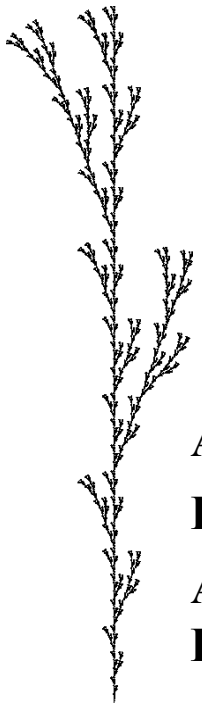


Artificial Life and L-system

a L-system (Lindenmayer, 1968) consists of (1) axioms, and (2) a set of rules

Axiom: B
Rules:
$B \rightarrow A$
$A \rightarrow AB$

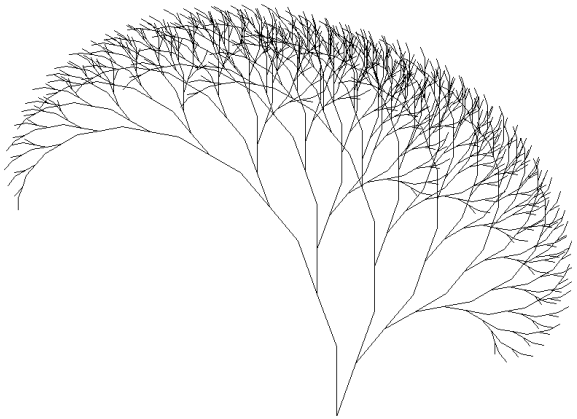




Axiom: F

Rule: $F = F\ [-F]\ F\ [+F]\ F$

Angle: 20
Depth: 7

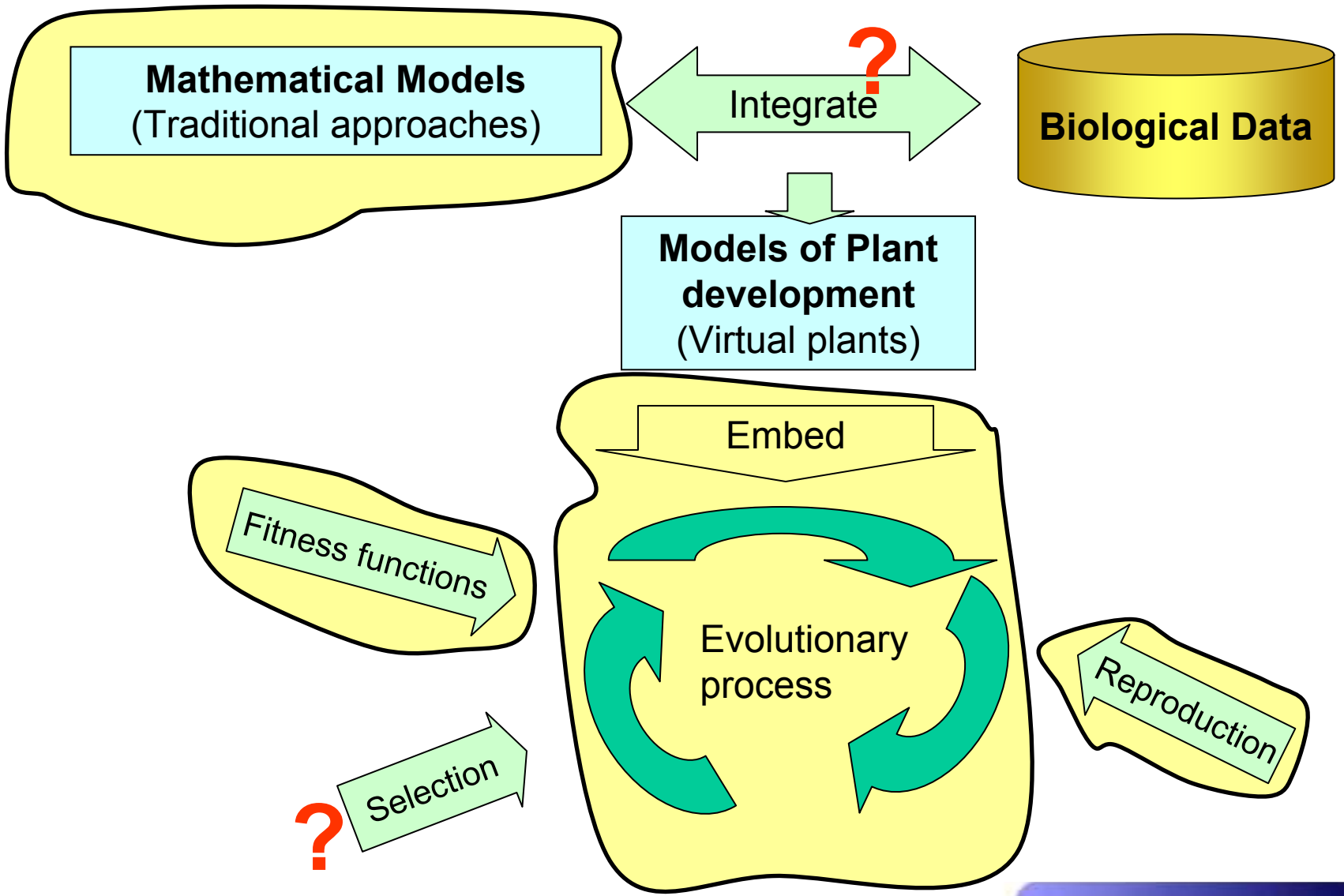


Axiom: F

Rule: $F = |\ [+F]\ |\ [-F]\ +F$

Angle: 20
Depth: 9

Rice Plant Growth Model? (L.M. Hoang)



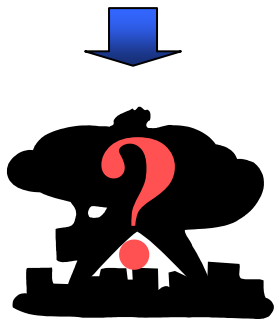
Discovery in Physics and Materials?

Discover the knowledge of electron

Conventional approach

Experimental data

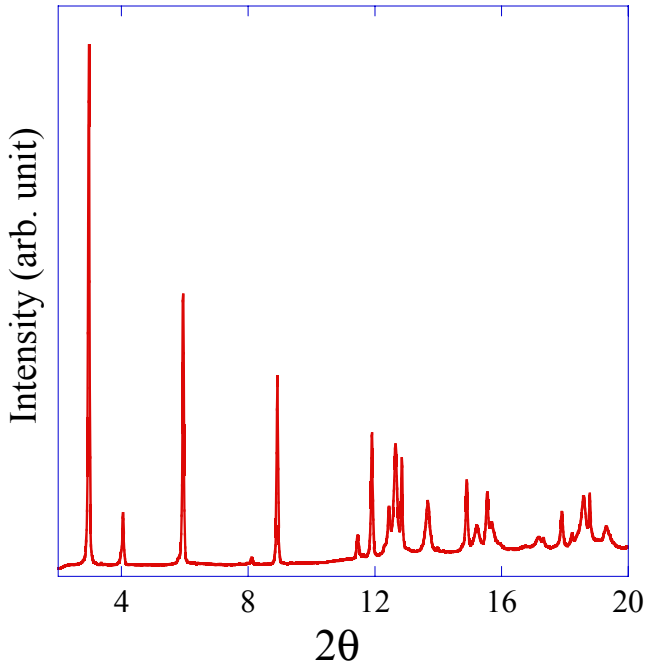
- Faraday law
- Coulomb law
- Current of electric
- Cathodic rays
- ß rays
- ß scattering
- Emission of H atoms
- Milliken measurement ($e=1.6\times10^{-19}C$)
- Photoelectric effect
- e/m_e measurement
- Electron diffraction
- etc.



Model construction	Model Revision	Final model
- Particle model - Wave model With their fitness to experimental data	Human Intelligence - De Broglie - Heisenberg - Schödinger	Quantum theory Wave packets
Automatically generate reasonably assumed models and accumulate their fitness to the experiments as data	Knowledge discovery and data mining: Automatic extraction of non-obvious, hidden knowledge New trial models	A challenge to discoveries in physics with computers

Discover the rules to create new assumed model that can fit to the experimental data

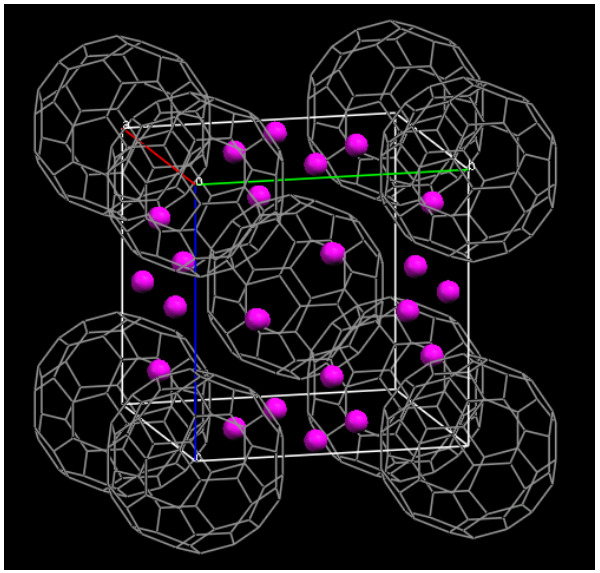
Crystal Structure Analysis (D.H. Chi)



Simulation problem

Fourier transformation

Prediction problem (limited data)

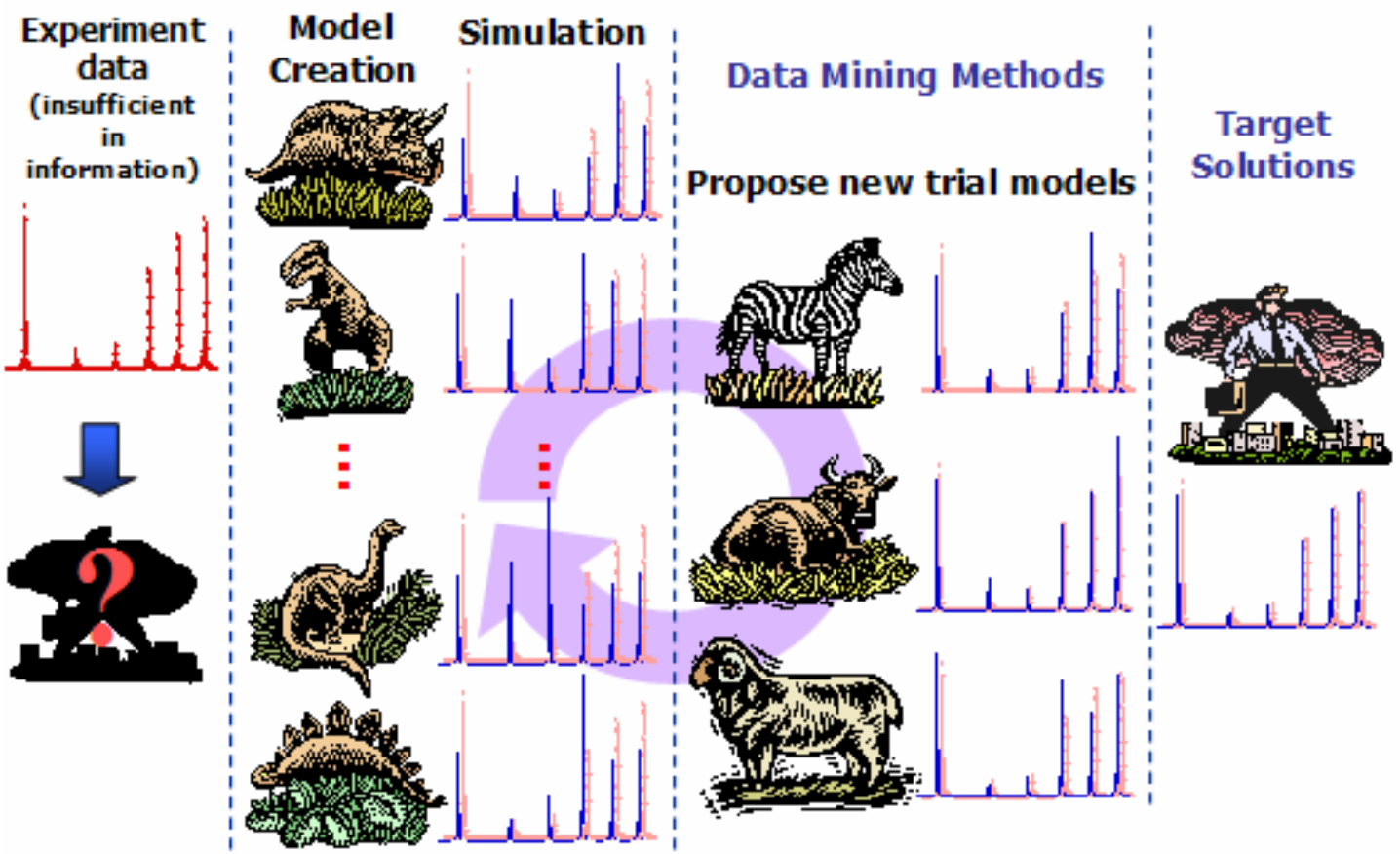


(ad-hoc)
Human knowledge on
Geometry
Physics
Chemistry

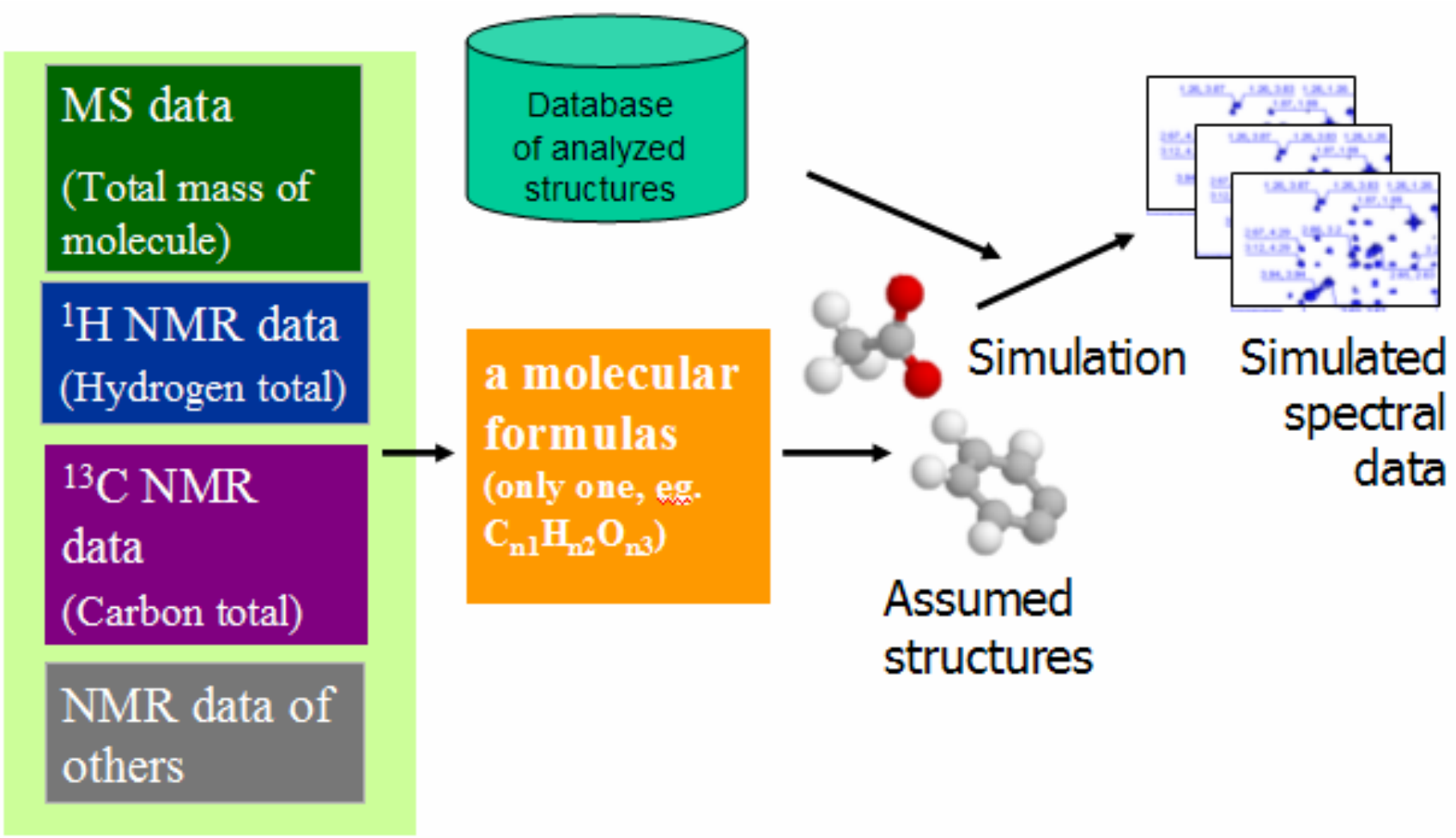
9.2003 XXX chuyển phase problem về bài toán quy hoạch nguyên

Comic: Data Mining in Structural Analysis

Quá trình lặp: (1) Xây dựng nhiều mô hình và mô phỏng để tạo dữ liệu; (2) phân tích các dữ liệu này nhằm phát hiện ra các quy luật có thể dùng được để tiếp tục tạo ra các mô hình (phổ) gần với mô hình cần dự đoán (phổ gốc)

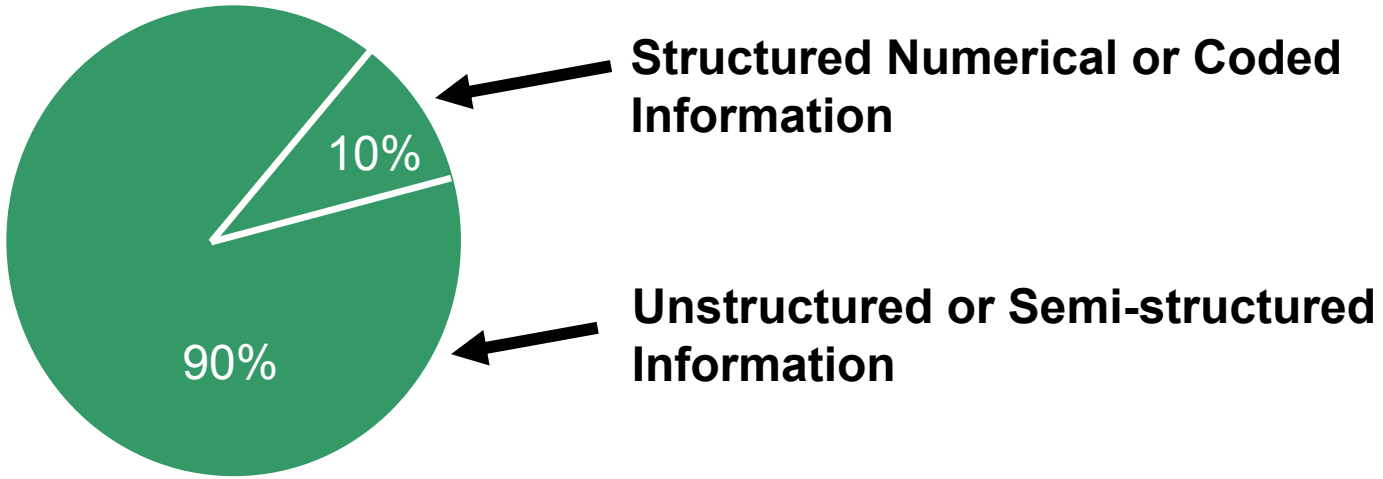


Molecular Structure Analysis (N.T. Tai)



Motivation for Text Mining

- Approximately 90% of the world's data is held in unstructured formats (source: Oracle Corporation)
- Information intensive business processes demand that we transcend from simple document retrieval to "knowledge" discovery.



Challenge of Text Mining

- Very high number of possible “dimensions” – Rất nhiều “chiều”
  - ➔ All possible word and phrase types in the language!!
- Unlike data mining – không giống khai phá dữ liệu
  - ➔ records (= docs) are not structurally identical
  - ➔ records are not statistically independent
- Complex and subtle relationships between concepts in text – Các quan hệ phức tạp và khó thấy giữa các khái niệm
  - ➔ “AOL merges with Time-Warner”
  - ➔ “Time-Warner is bought by AOL”
- Ambiguity and context sensitivity – Nhập nhằng và cảm ngữ cảnh
  - ➔ automobile = car = vehicle = Toyota
  - ➔ Apple (the company) or apple (the fruit)

Về nghiên cứu cơ bản trong CNTT ở Việt nam

- Theo Bùi Duy Hiển (Tạp chí Tia sáng): Viện thông tin khoa học Mỹ thống kê 9.000 tạp chí
- Trong 1998-2002, Việt Nam có gần 1.500 bài báo trên các tạp chí quốc tế (ngang Thái-lan 10 năm trước, 6.4K người vs. 21 K người), mỗi năm chừng 340 bài.
- Cần ít nhất 116 K$ để ra được một công trình, cần 39 M$/năm cho 340 công trình (???)
- Ta nên làm nghiên cứu cơ bản ở lĩnh vực nào và ở mức độ nào?

Hình 2. Số lượng công trình R&D về khoa học tự nhiên của Việt Nam công bố trên các tạp chí quốc tế trong 5 năm gần đây (1998 - 2002).

BẢNG 2

ĐẦU TƯ CHO CÔNG TRÌNH KHOA HỌC Ở MỘT SỐ NƯỚC CÓ NỀN KHOA HỌC TIÊN TIẾN (1991-2000)

	Số công trình trong một năm	Số lần trích dẫn trung bình	GNP (tỉ USD)	Đầu tư cho R&D (tỉ USD)	Đầu tư cho một công trình (nghìn USD)
Mỹ	175773	5.12	6737	168	958
Nhật	41065	2.99	4321	121	2946
Hà Lan	10233	3.99	338	6.76	661
Trung Quốc	5434	0.97	630	0.63	116
Singapore	642	1.62	65.8	0.72	1128
Nga	30504	0.66	393	4.32	142
Ấn Độ	9736	1.09	279	1.67	172

Summary

- Khoa học đang rất tập trung vào khai thác dữ liệu (data intensive). Khả năng phân tích các tập dữ liệu cực lớn là cốt yếu và thách thức trong phát triển CNTT
- Khai phá dữ liệu liên quan đến các tiến bộ cơ bản của databases, algorithmics, statistics, machine learning, visualization, etc.
- Hai vấn đề then chốt của khai thác dữ liệu
  - ➔ Các lược đồ dữ liệu khác nhau.
  - ➔ Tìm các thuật toán có độ phức tạp nlogn là thách thức chủ yếu trong khai phá dữ liệu
- My personal view: Applied research should be the main focus of scientific research in Vietnam

Acknowledgments

- Some slides were adapted from those of Jim Gray (Microsoft), Padhraic Smyth (Univ. California Irvine)
- Projects KC01-03, NCCB, Tokyo Cancer Center, Active Mining, Hợp tác khoa học với Việt Nam, etc.
- Setsuo Ohsuga, Hiroshi Motoda, Phòng Nhận dạng & CNTT, H. Nakamori, Nguyen Ngoc Binh, Nguyen Trong Dung, A. Saitou, S. Kawasaki, Nguyen Duc Dung, Le Si Quang, Huynh Van Nam, Nguyen Tien Tai, Dam Hieu Chi, Nguyen Phu Chien, H. Zhang, A. Hassine, H. Yokoi, T. Takabayashi, A. Yamaguchi, Pham Tho Hoan, Le Minh Hoang, …