Some emerging trends in machine learning and data mining research

Ho Tu Bao

School of Knowledge Science
Japan Advanced Institute of Science and Technology

Institute of Information Technology
Vietnamese Academy of Science and Technology

John E. Hopcroft (Hanoi, 10.8.2004)

vietnamnet
WWW.VNN.VN

g: hotnews@vasc.com.vn Tel HN: (04) 7722729 / 84 - 9

ENGLISH | Toà soạn và trị sự

TRANG NHẤT
Chính trị
Xã hội
Đại dịch cúm
Kinh tế
Quốc tế
Văn hoá - Giải trí

GS John Hopcroft: Cần đào tạo các chuyên gia lãnh đạo CNTT

16:17' 10/08/2004 (GMT+7)

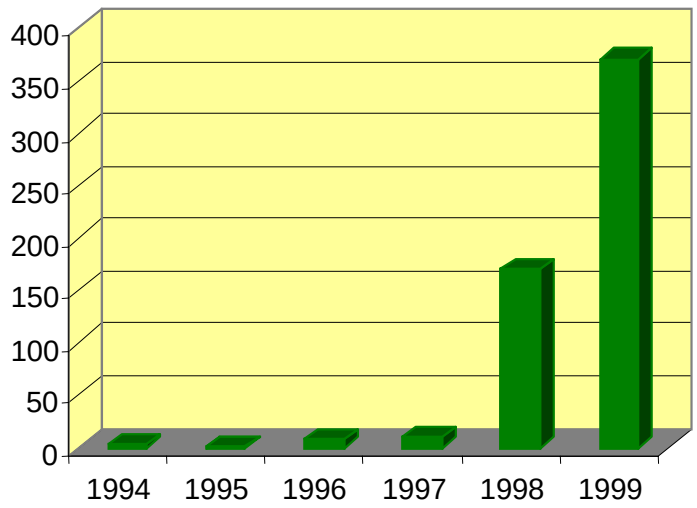
- Công trình nghiên cứu của giáo sư hiện nay là gì? (Chu Quang Chấp, 41 tuổi, Hải Phòng)

- Hiện nay, tôi đang nghiên cứu một vài chương trình cực kỳ phức tạp có liên quan đến việc ***tìm kiếm và xử lý tin tức trong mạng Internet***. Một trong những chương trình đó là phương pháp tìm kiếm và phân tích tất cả hơn 300.000 bài nghiên cứu trong tất cả các lĩnh vực khoa học và công nghệ trên toàn thế giới. Qua dự án này, chúng tôi có thể biết được những mũi nhọn nghiên cứu chính hiện nay trên thế giới, đây là phương cách ***tìm kiếm được biên cương mới nhất trong lĩnh vực nghiên cứu khoa học***.

(Aho, Hopcroft, Ulmann, *Data Structures and Algorithms*,1983)

What are emerging trends?

- Our research environment greatly changed with computer networks, Web, digital library, etc.
- Difficult to being up-to-date
  - ➔ JAIST has 4700 online journals, 282,000 papers per year, reading 1%/year = reading 2820 papers/yr = reading 8 papers/day.



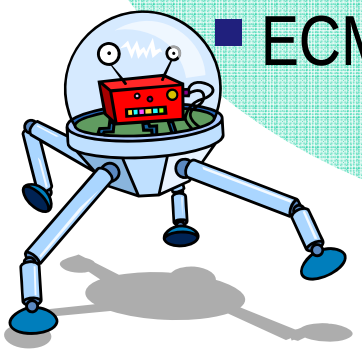
Keyword "XML" in INSPEC® database

Emerging Trend: a topic that is growing in interest and utility overtime.

Machine learning and data mining

- **Machine learning**
  To build computer systems that learn as well as humans do (learning from data).
- ICML since 1982 (23th ICML in 2006), ECML since 1989
- ECML/PKDD since 2001

- **Data mining**
  To find new and useful knowledge from large datasets
- ACM SIGKDD since 1995, PKDD and PAKDD since 1997 (PAKDD'05 in Hanoi), IEEE ICDM and SIAM DM since 2000, etc.



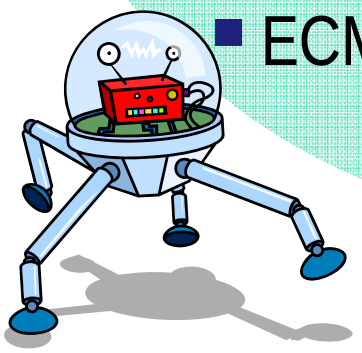
Machine learning and data mining

- **Machine lear**
  To build cor
  systems th
  well as h
  (learning
- ICML sinc
  (23th ICM
  ECML since
- ECML/PKDD since 2001

ning
nd
ge
ts
ce
D and
ce 1997
D 05 in Hanoi),
EE ICDM and SIAM
DM since 2000, etc.

What are emerging trends in machine learning and data mining research?

(topics that are growing in interest and utility overtime)



Outline

- Discriminative random fields
- Kernel methods

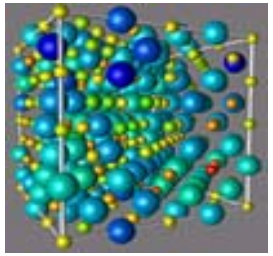
Some emerging trends

Our recent work in these trends

Data schemas vs. learning/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

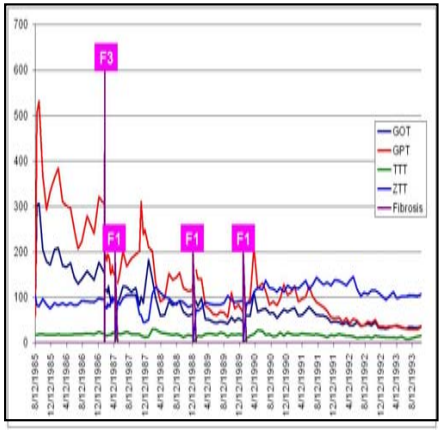
- **Classification/Prediction**
  - ➔ Decision trees
  - ➔ Neural network
  - ➔ Rule induction
  - ➔ Support vector machines
  - ➔ Hidden Markov Model
  - ➔ etc.
- **Description**
  - ➔ Association analysis
  - ➔ Clustering
  - ➔ Summarization
  - ➔ etc.

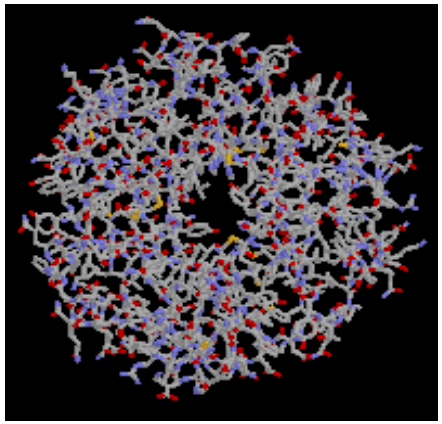
10 challenging problems in data mining
(ICDM'05)

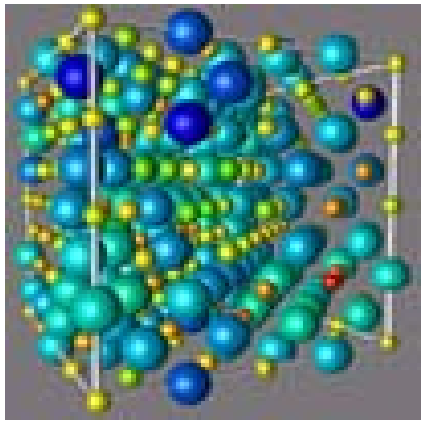
1. Developing a unifying theory of data mining
2. Scaling up for high dimensional data/high speed streams
3. Mining sequence data and time series data
4. Mining complex knowledge from complex data
5. Data mining in a network setting
6. Distributed data mining and mining multi-agent data
7. Data mining for biological and environmental problems
8. Data-mining-process related problems
9. Security, privacy and data integrity
10. Dealing with non-static, unbalanced and cost-sensitive data

Complexly structured data

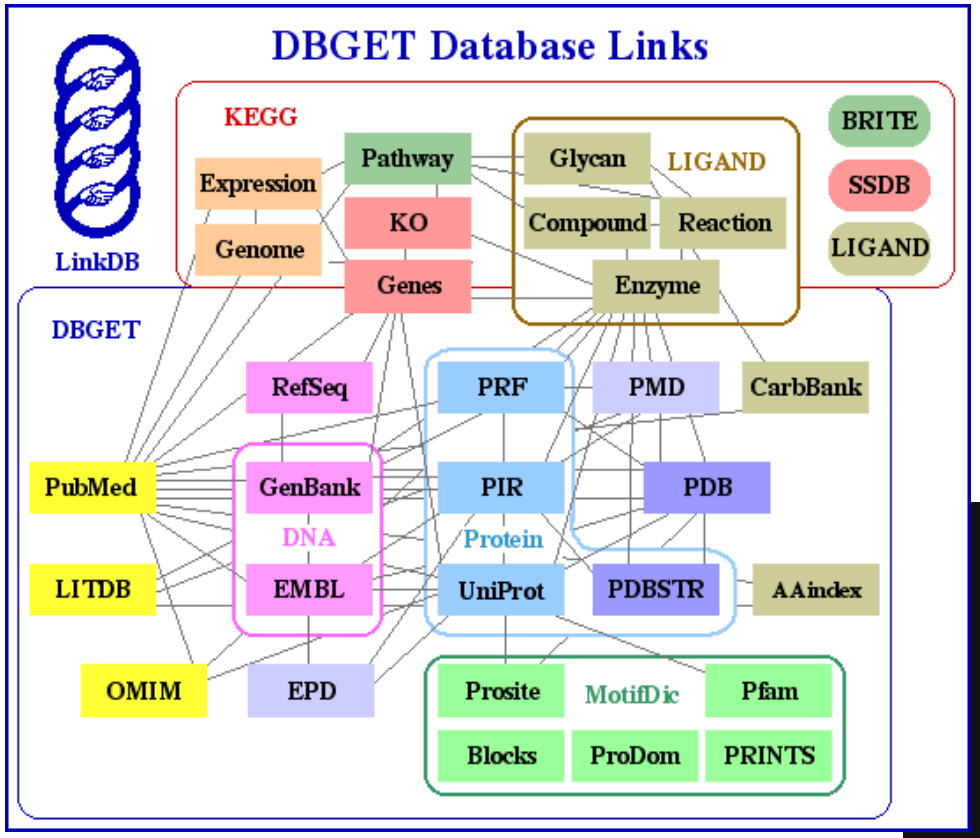
- Relational databases
- XML databases
- Sequences, molecules,
- Graphs and trees
- others

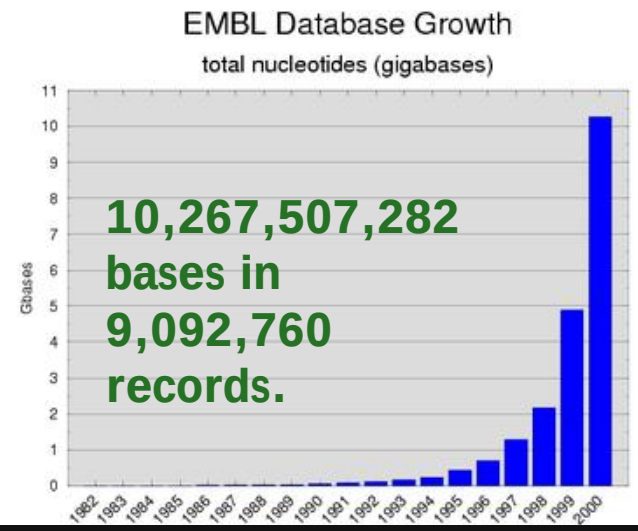


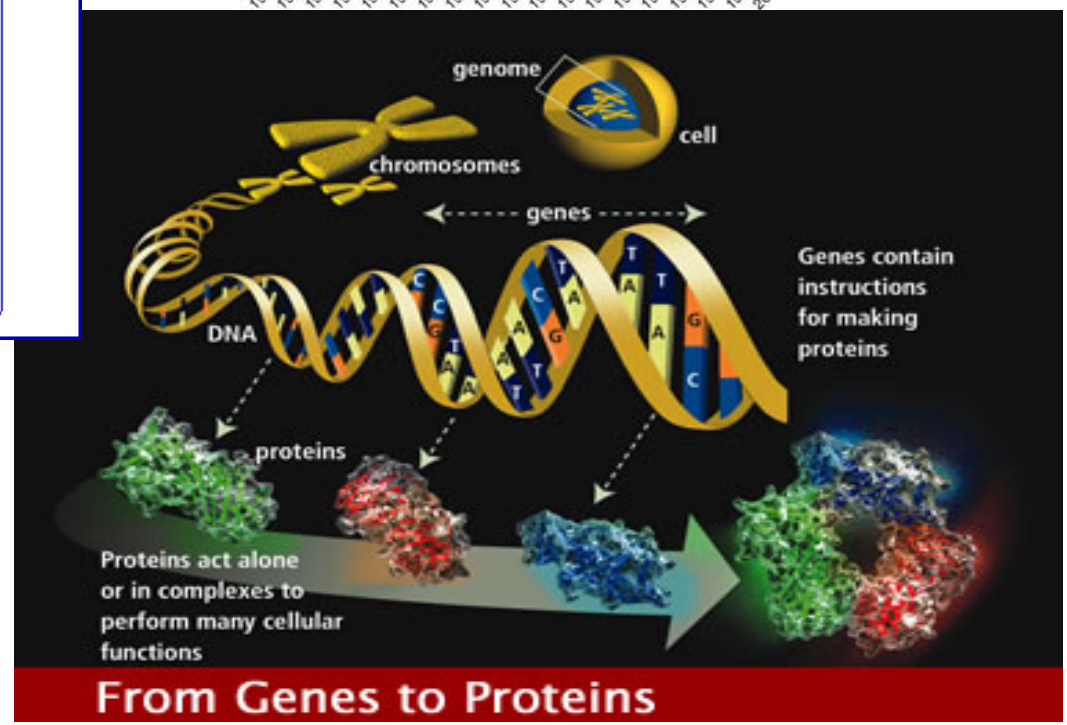




Explosion of biological data







Bioinformatics problems:

- **Sequence analysis**
- **Genomics**
- **Proteomics**
- **Others (e.g., systems biology)**

How biological data look like?

A portion of the DNA sequence, consisting of 1.6 million characters, is given as follows (about 350 characters, 4570 times smaller):

...TACATTAGTTATTACATTGAGAAACTTTATAATTAAAAAAGATTCATGT
AAATTTCTTATTTGTTTATTTAGAGGTTTTAAATTTAATTTCTAAGGGTT
TGCTGGTTTCATTGTT
TGAAAATTAGGATTAA
GTTAAATTTTTAAATTT
GAAAGAAAGATTTAAA
CTTAGAAAAATATGGT
TATTATGT...

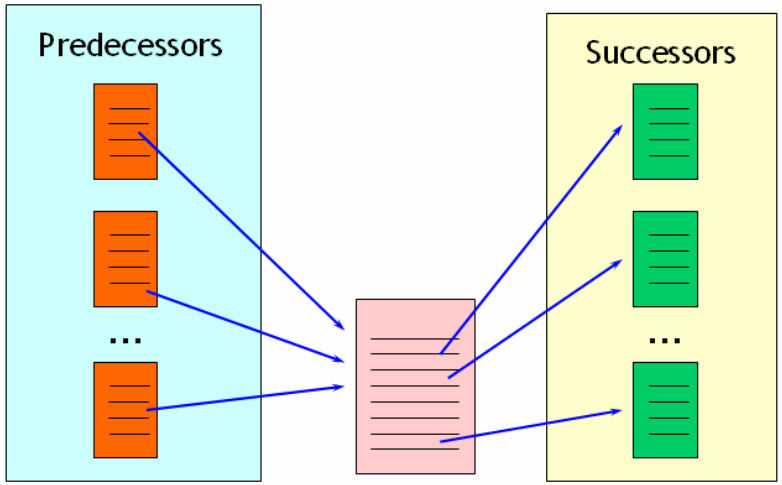
View - ecoli_functions.pl[1]

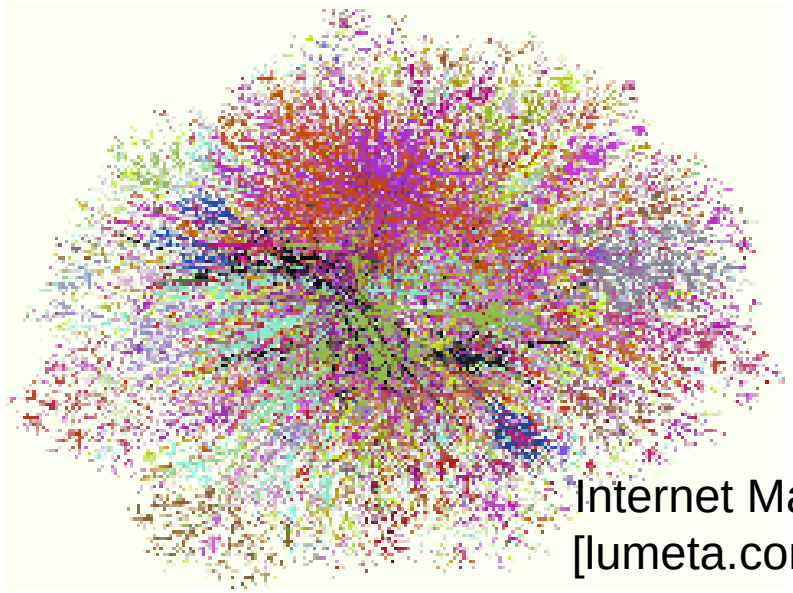
File Edit View Help

```
function(ecoli962,5,1,4,'insB_4','IS1 protein InsB').
function(ecoli963,7,0,0,'b0989','cold shock-like protein').
function(ecoli964,7,0,0,'cspG','low-temperature-responsive gene analog of CspA and CspB homolog of Salmonella
cold shock protein').
function(ecoli965,7,0,0,'sfa','suppresses fabA and ts growth mutation').
function(ecoli966,0,0,0,'b0992','orf').
function(ecoli967,3,5,2,'torS','sensor protein tors  (3rd module transmitter domain (?kinase) interacts with torr)').
function(ecoli968,3,5,2,'torT','part of regulation of tor operon periplasmic(1st module)').
function(ecoli969,3,5,2,'torR','response transcriptional regulator for torA  (sensor TorS)(1st module)').
function(ecoli970,3,5,2,'torC','trimethylamine N-oxide reductase cytochrome c-type subunit also has activity as
negativer regulator of tor operon(1st module)').
function(ecoli971,3,5,2,'torA','trimethylamine N-oxide reductase subunit').
function(ecoli972,3,5,2,'torD','part of trimethylamine-N-oxide oxidoreductase').
function(ecoli973,0,0,0,'yccD','orf').
```

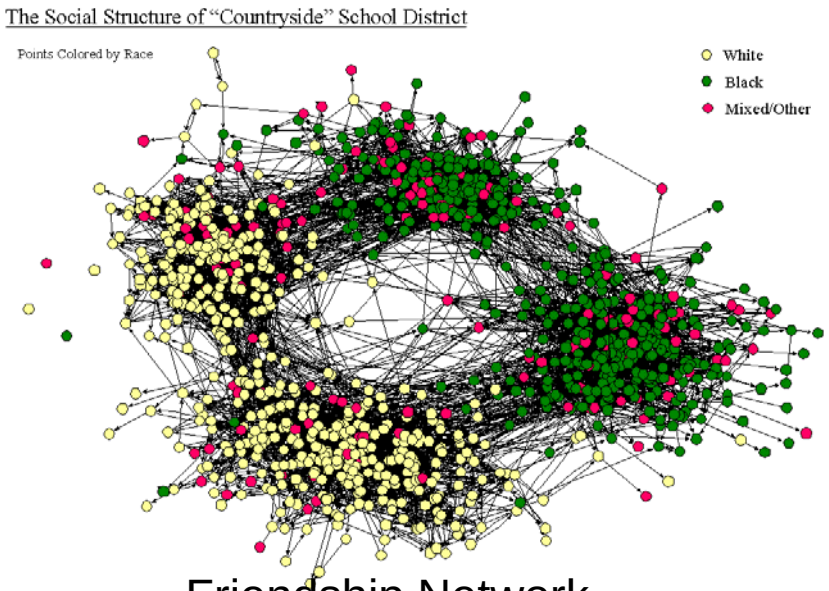
324,027 bytes

Web link data

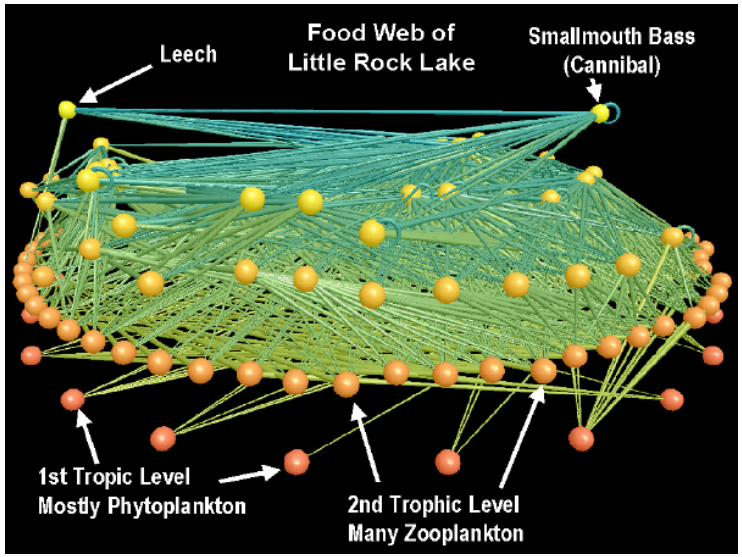




Internet Map
[lumeta.com]



Friendship Network
[Moody '01]

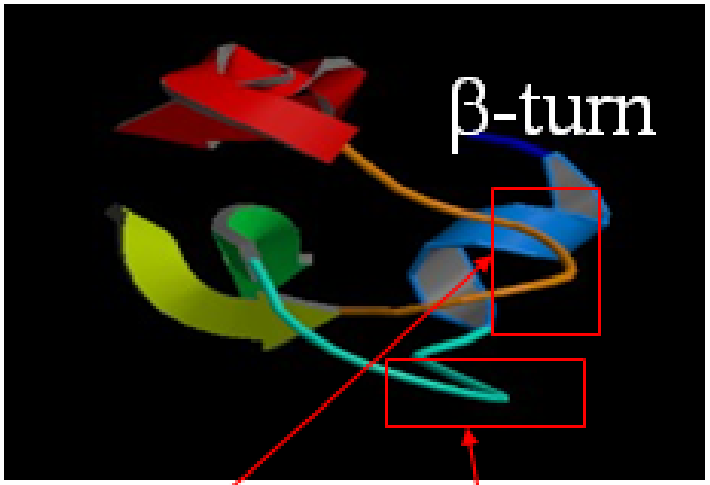


Food Web
[Martinez '91]

Labeling sequence data problem

- **X** is a random variable over data sequences
- **Y** is a random variable over label sequences whose labels are assumed to range over a finite label alphabet A
- *Problem*: Learn how to give labels from a closed set **Y** to a data sequence **X**

	x_1	x_2	x_3
X:	Thinking	is	being
Y:	noun	verb	noun
	y_1	y_2	y_3



X KARIIRYFYNAKAGLCQTFCRAKRNNFKSAED
Y nnnnnnnnnTTtttnnnnnnnnTtttnnnnnn

- POS tagging, phrase types, etc. (NLP),
- Named entity recognition (IE)
- Modeling protein sequences (CB)
- Image segmentation, object recognition (PR)
- etc.

Generative vs. discriminative methods

Training classifiers involves estimating **f: X → Y**, or **P(Y|X).**
Examples: P(apple | red ∧ round), P(noun | "cá")

Generative classifiers

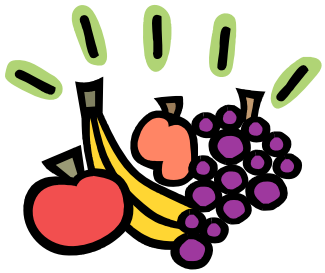
- Assume some functional form for P(X|Y), P(Y)
- Estimate parameters of **P(X|Y), P(Y)** directly from training data, and use Bayes rule to calculate **P(Y|X = x_i)**
- HMM, Markov random fiel
  Bayesian networks,
  Gaussians, Naïve Bayes, e

Discriminative classifiers

- Assume some functional form for P(Y|X)
- Estimate parameters of

s,
g,
n

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$



$$P(apple \mid red \wedge round) = \frac{P(red \wedge round \mid apple)P(apple)}{P(red \wedge round)}$$

(cá: fish, to bet)

Generative vs. discriminative methods

Generative approach

- Try to build models for the underlying patterns
- Can be learned, adapted, and generalized with small data.

Discriminative approach

- Try to learn to minimize an utility function (e.g. classification error) but not to model, represent, or "understand" the pattern explicitly (detect 99.99% faces in real images and do not "know" that a face has two eyes).
- Often need large training data, say 100,000 labeled examples, and can hardly be generalized.

Finite state machines (FSM): a bit of history

Archeology of computational linguistics

- **1990s–2000s: Statistical learning**
  - ➔ algorithms, evaluation, corpora
- **1980s: Standard resources and tasks**
  - ➔ Penn Treebank, WordNet, MUC
- **1970s: Kernel (vector) spaces**
  - ➔ clustering, information retrieval (IR)
- **1960s: Representation Transformation**
  - ➔ Finite state machines (FSM) and Augmented transition networks (ATNs)
- **1960s: Representation**—beyond the word level
  - ➔ lexical features, tree structures, networks

Trainable parsers

Trainable FSMs

(E. Hovy, COLING 2004)

Finite state machines (FSM): a bit of history

Archeology of computational linguistics

- **1990s–2000s: Statistical learning**
  - ➔ algorithms,
- **1980s**
  - ➔ F
- **1**
- **19**
  - ➔ ugmented tra
- **1960s: Representati** beyo the word level
  - ➔ lexical features, tree structures, networks

Trainable parsers

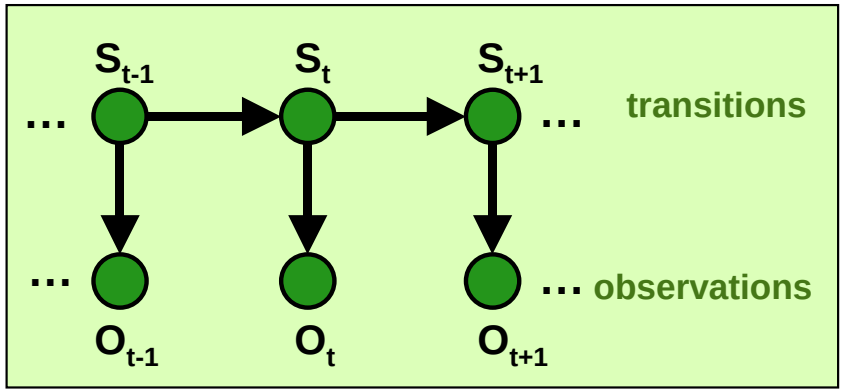
Trainable FSMs

- Hidden Markov Models (HMM, 1960s)
- Maximum Entropy Markov Models (MEMM, 2000)
- Conditional Random Fields (CRF, 2001)

(E. Hovy, COLING 2004)

Hidden Markov models [Baum et al., 60s]

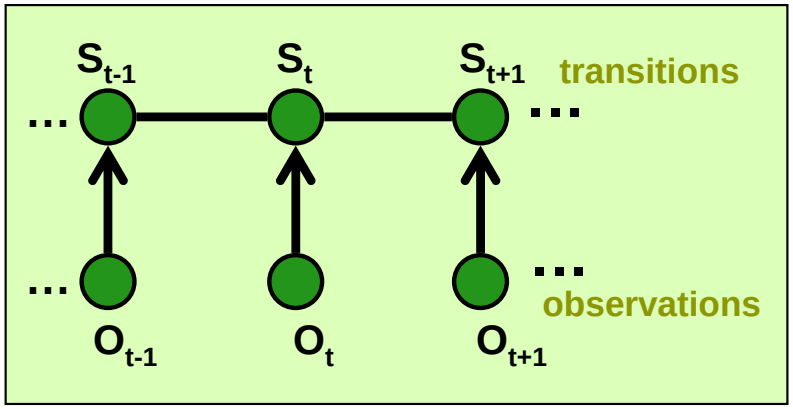
- An HMM is a 5-tuple (O, S, A, B, π)
  - ➔ Observations $O = \{o_1, o_2, \ldots, o_N\}$
  - ➔ States $S = \{s_1, s_2, \ldots, s_M\}$
  - ➔ Transition probability $P(s_t | s_{t-1})$
  - ➔ Emission probability $P(o_t | s_t)$
  - ➔ Start state probabilities $P(s_t)$



- Characteristics
  - ➔ A direct graphical model, generative
  - ➔ Locally normalized at each state
  - ➔ Applied to a wide variety of problems in speech & text processing, biology, etc.

MEMMs [McCallum et al., 2000]

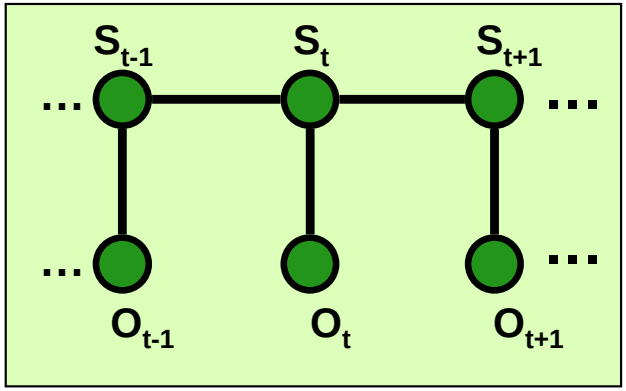
- Features
  - ➔ Represents the probability of reaching a state given an observation and the previous state.
  - ➔ These conditional probabilities are specified by exponential models based on arbitrary observation features.



- Characteristics
  - ➔ A direct graphical model
  - ➔ Discriminative
  - ➔ Locally normalized at each state
  - ➔ Can represent an array of highly dependent observation features (of different levels of granularity)

CRFs [Lafferty et al., 2001]

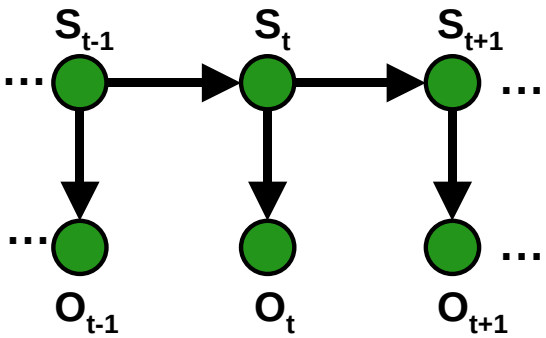
- Solve label bias problem: CRFs drop the requirement of local normalization
- Globally normalized
  (HMM & MEMM: locally normalized)



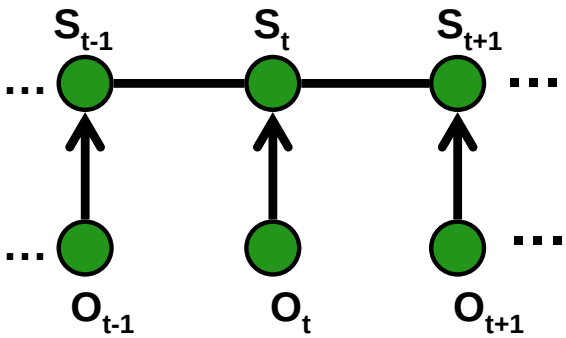
$$P(\vec{s} \mid \vec{o}) = \frac{1}{Z(\vec{o})} \prod_{t=1}^{|\vec{o}|} \Phi_s(s_t, s_{t-1}) \Phi_o(o_t, s_t)$$

- Characteristics
  - ➔ Undirected graphical model, discriminative
  - ➔ Globally normalized
  - ➔ Consider the state sequence as a whole (not separated unit)
  - ➔ Can represent rich-dependent features of training data
  - ➔ Parameters of a CRF are the feature weight vetor $\lambda = \{\lambda_1, \ldots, \lambda_K\}$

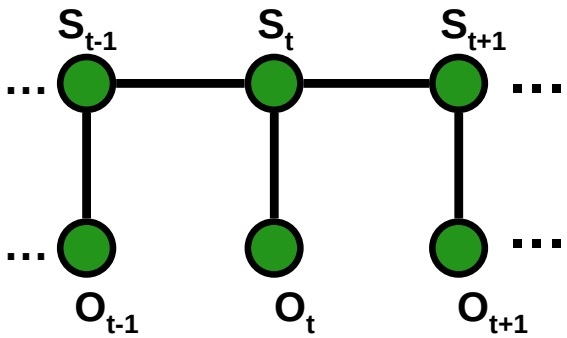
Trainable finite state machines



HMMs (directed graph, joint, generative)



MEMMs (directed graph, conditional, discriminative)



CRFs (undirected graph, conditional, discriminative)

Using advanced statistical and graphical methods:

Maximum Entropy (Maxent), Hidden Markov Models (HMM), Maximum Entropy Markov Models (MEMM), Conditional Random Fields (CRF), Markov Random Fields (MRF), Relational Probabilistic Models (RPM), etc.

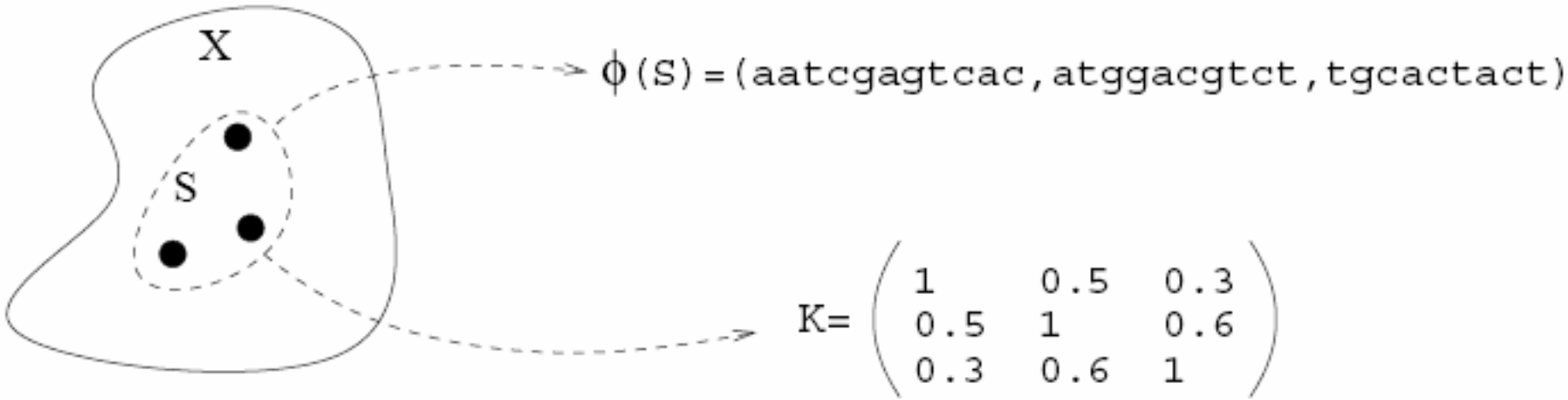
The issue of data representation

- Let $\mathbb{S} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ a set of *n* objects to be analyzed.
- Suppose that each object $\mathbf{x}_i$ is an element of a set $\mathcal{X}$, which may be images, molecules, texts, etc.
- Majority of data analysis methods represent $\mathbb{S}$ by:
  - ➔ defining a representation $\phi(\mathbf{x}) \in \mathcal{F}$ for each object $\mathbf{x} \in \mathcal{X}$, where $\mathcal{F}$ can be real-valued vector ($\mathcal{F} = \Re^p$) or finite-length strings, or more complex representation.
  - ➔ representing the objects by a set of their representations, $\phi(\mathbb{S}) = (\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n))$

Kernel representation: idea

- Data are not represented individually anymore, but only through a set of pairwise comparisons.
- Instead of using a mapping $\phi: \mathcal{X} \rightarrow \mathcal{F}$ to represent each object $x \in \mathcal{X}$ by $\phi(\mathbf{x}) \in \mathcal{F}$, a real-valued "comparison function" $k: \mathcal{X} \times \mathcal{X} \rightarrow \Re$ is used (kernel), and the data set $\mathcal{S}$ is represented by the $n \times n$ matrix of pairwise comparisons $k_{i,j} = k(x_i, x_j)$.
- In the new space, the problem solving is easier (e.g. linear)
- All kernel methods are designed to process such square matrices.

Kernel representation: idea



- $\mathcal{X}$ is the set of all oligonucleotides, $\mathcal{S}$ consists of three oligonucleoides.
- Traditionally, each oligonucleotide is represented by a sequence of letters.
- In kernel methods, $\mathcal{S}$ is represented as a matrix of pairwise similarity between its elements.

More about kernels

A kernel k(x,y)

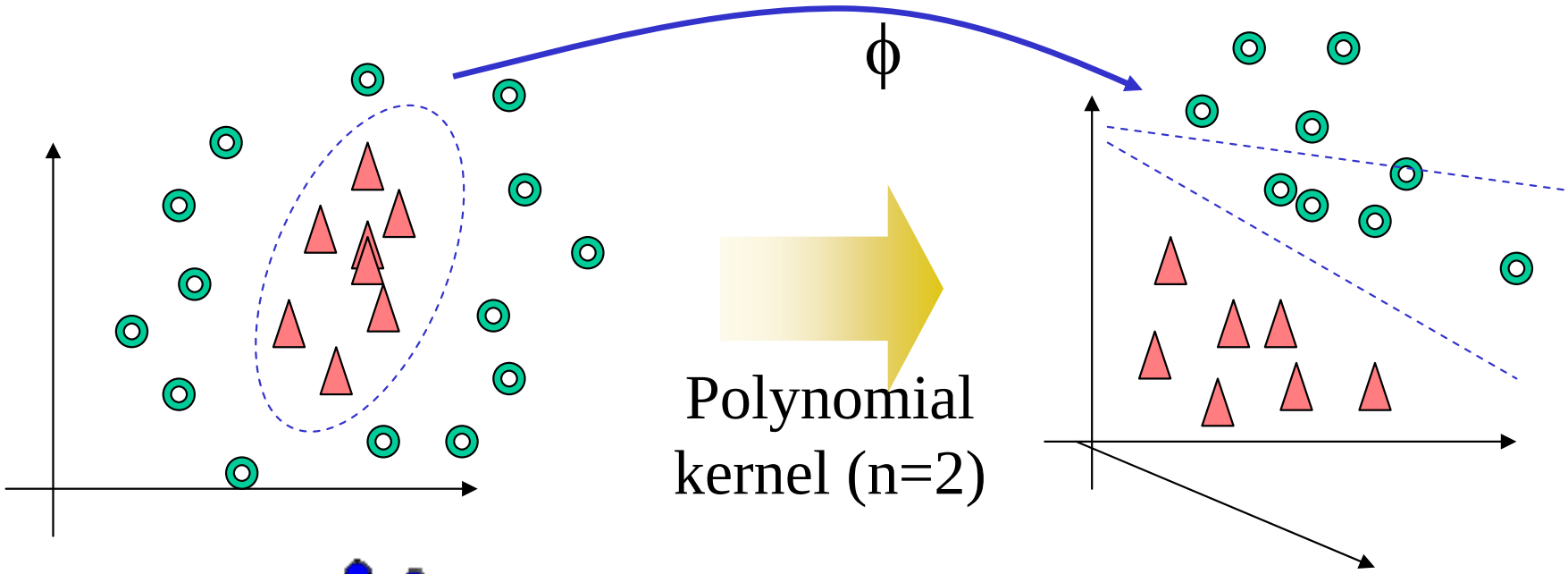
- ➔ is a similarity measure
- ➔ defined by an implicit mapping ϕ
- ➔ from the original space to a vector space (feature space)

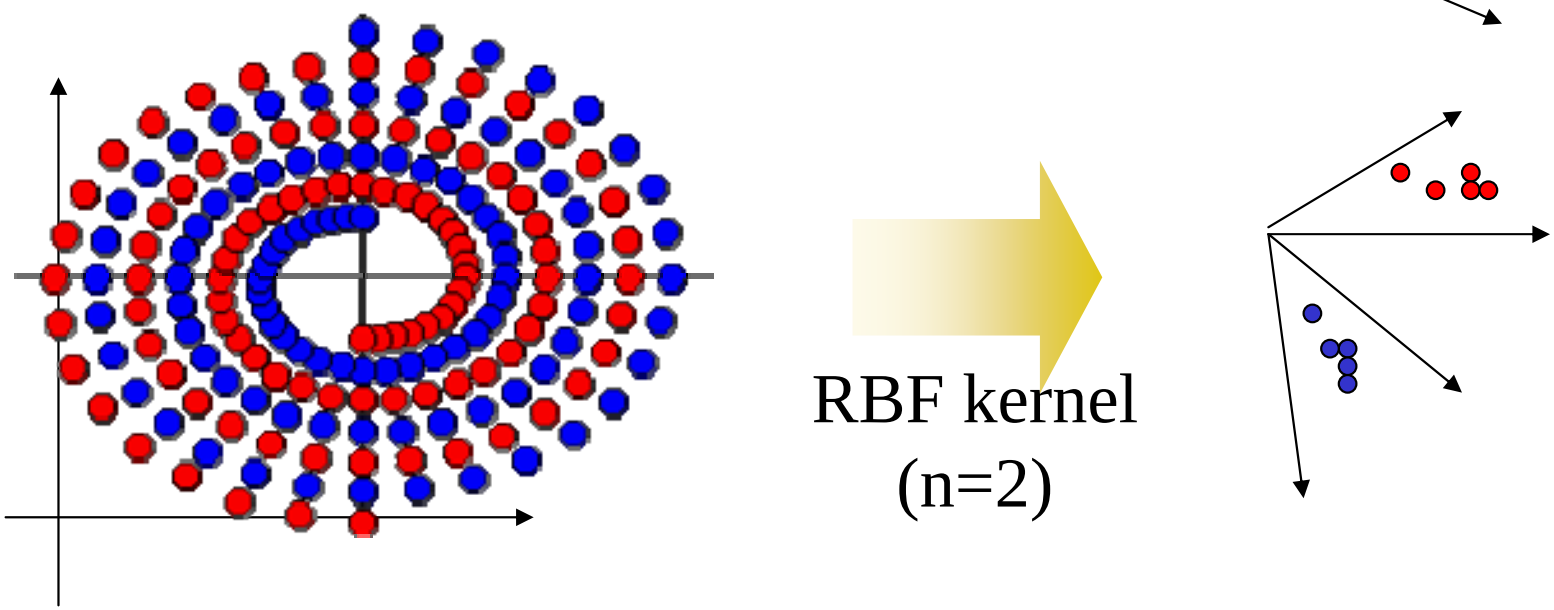
  such that: $k(x,y)=\phi(x)\bullet\phi(y)$
- ➔ Different kernels for different data types
  - ⇨ Vector space kernel for text
  - ⇨ Spectrum kernel for sequential data
  - ⇨ Diffusion kernels for graph
  - ⇨ etc.

Principles governing kernel design

- ➔ Invariance or other a priori knowledge
- ➔ Simpler structure (linear representation of the data)
- ➔ The class of functions the solution is taken from
- ➔ Possibly infinite dimension (hypothesis space for learning)
- ➔ … but still computational efficiency when computing $k(x,y)$

Examples of kernels (III)



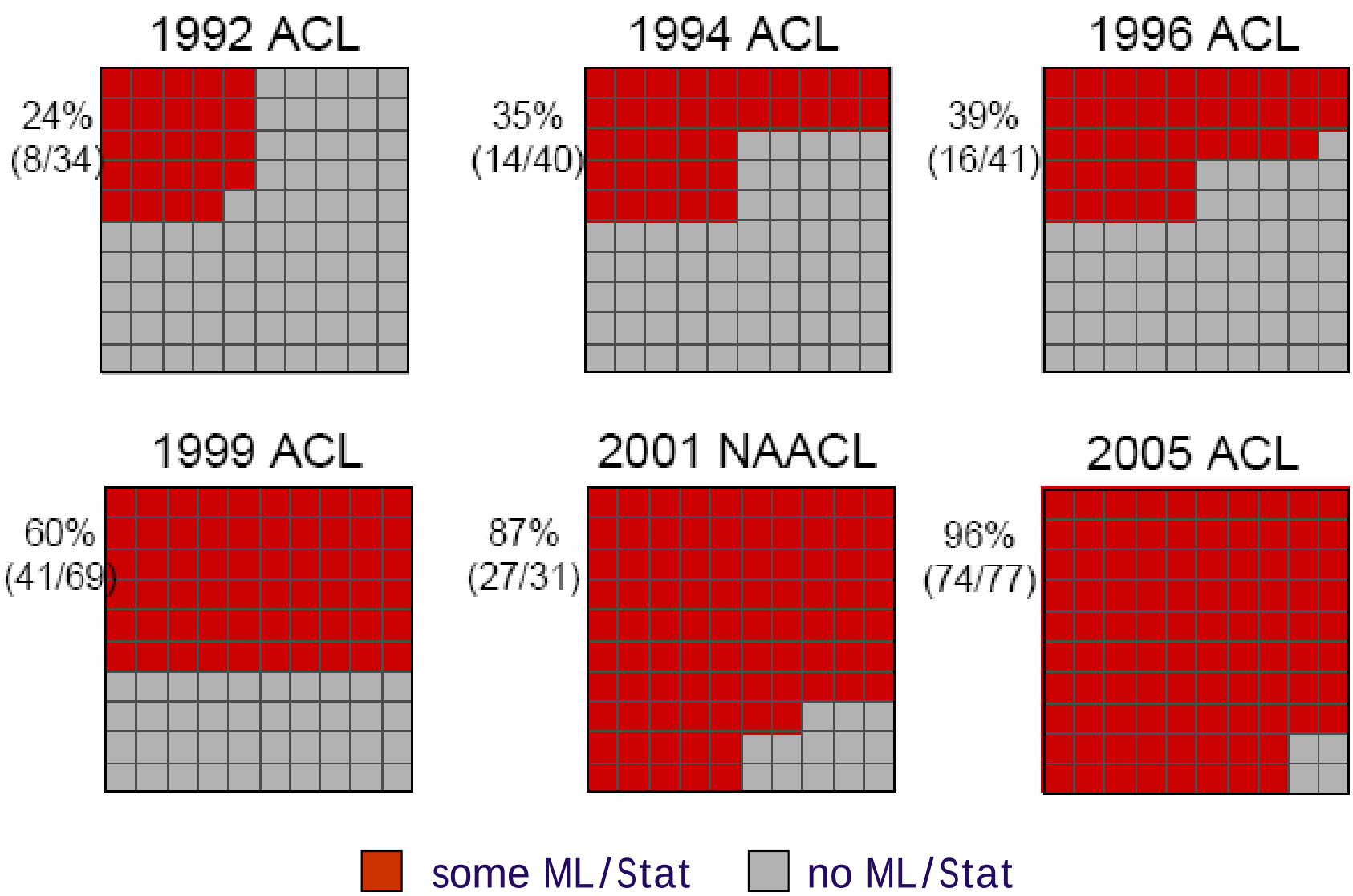


(Jean-Michel Renders, ICFT'04)

Kernel methods: a bit of history

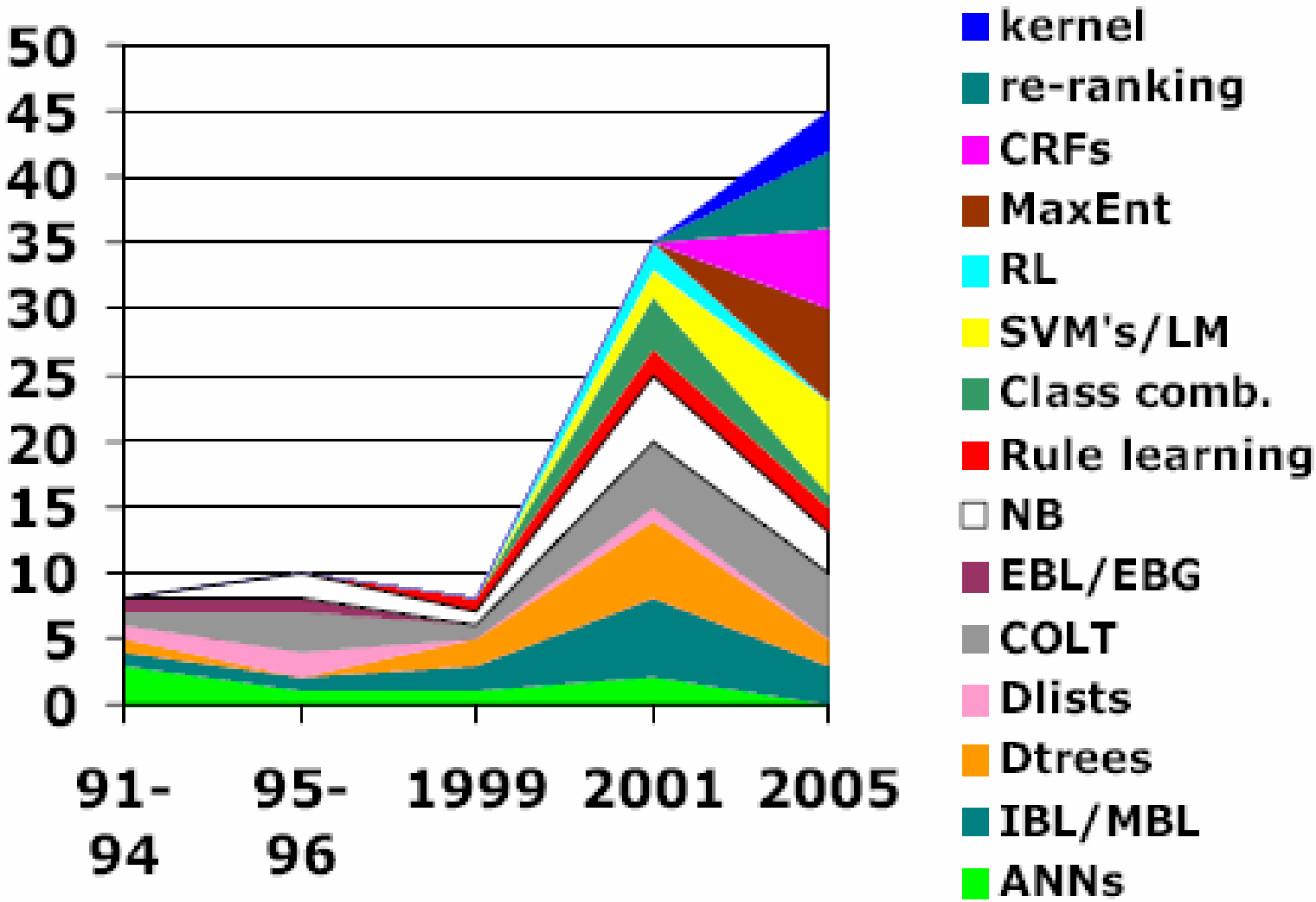
- Aronszajn (1950) and Parzen (1962): employ positive finite kernels in statistics.
- Aizerman et al. (1964): positive definite kernel is identical to a dot product in another space in which their algorithm reduced to perceptron algorithm.
- Boser et al. (1992): optimal hyperplane algorithm (SVM).
- Schölkopf (1997): work with nonvectorial data; (1999): kernels can be used to construct generalizations of any algorithm that can be carried out in terms of dot products.
- Haussler (1999) and Watkins (2000): first examples of nontrivial kernels defined on nonvectorial data.
- Since 2000: large number of “kernelizations" of various algorithms, various applications.

ML and statistical methods in NLP



(Marie Claire, ECML/PKDD 2005)

Why they can be viewed as emerging trends?



(Marie Claire, ECML/PKDD 2005)

Observed from ICML 2004

Honorable Mention for Outstanding Paper Award

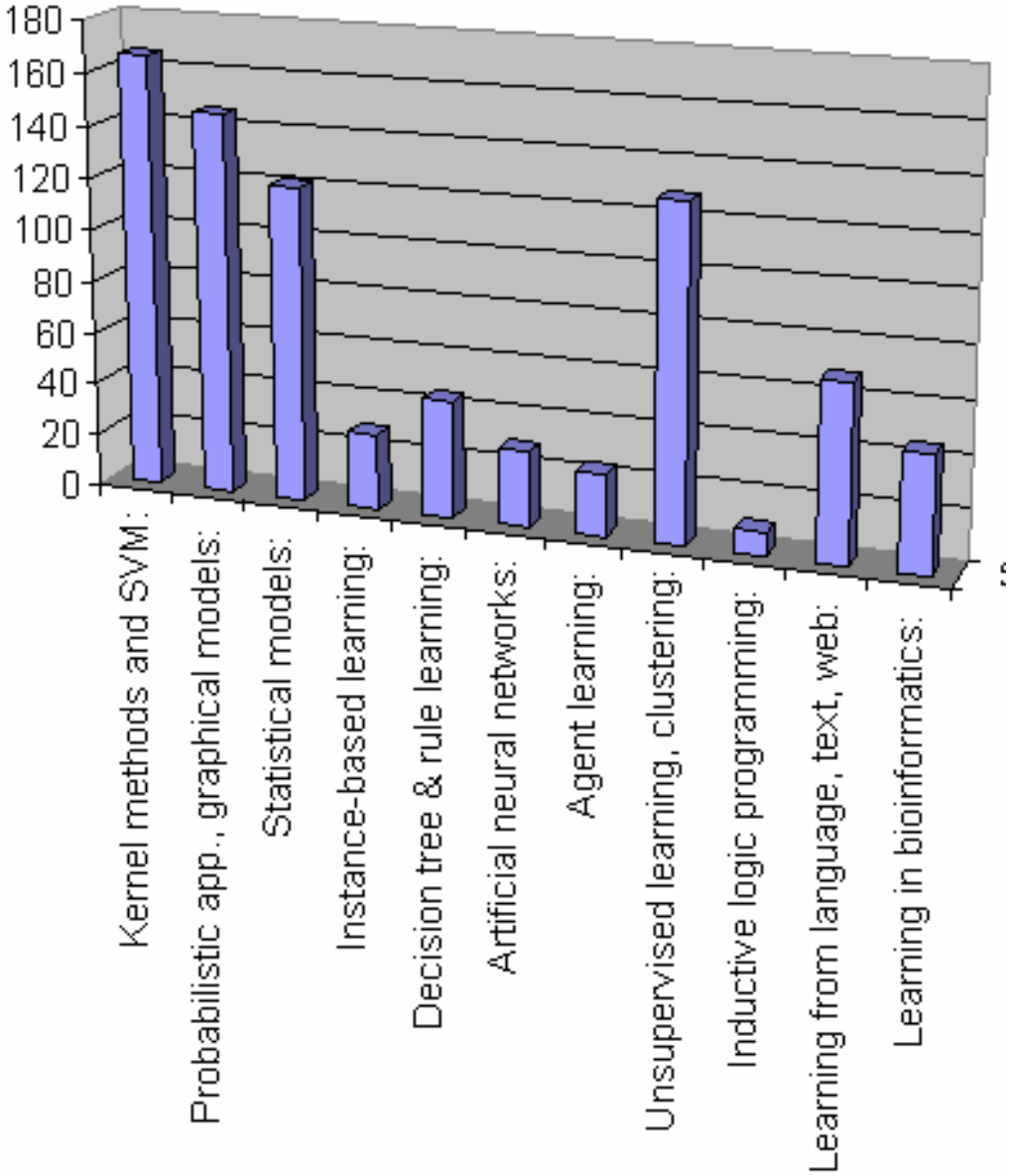
- **Multiple Kernel Learning, Conic Duality, and the SMO Algorithm**
  - *Francis Bach, Gert Lanckriet, Michael Jordan*
- **Efficient Hierarchical MCMC for Policy Search**
  - *Malcolm Strens*
- **Authorship Verification as a One-Class Classification Problem**
  - *Moshe Koppel, Jonathan Schler*

38

(Russ Greiner, ICML'04 PC co-chair)

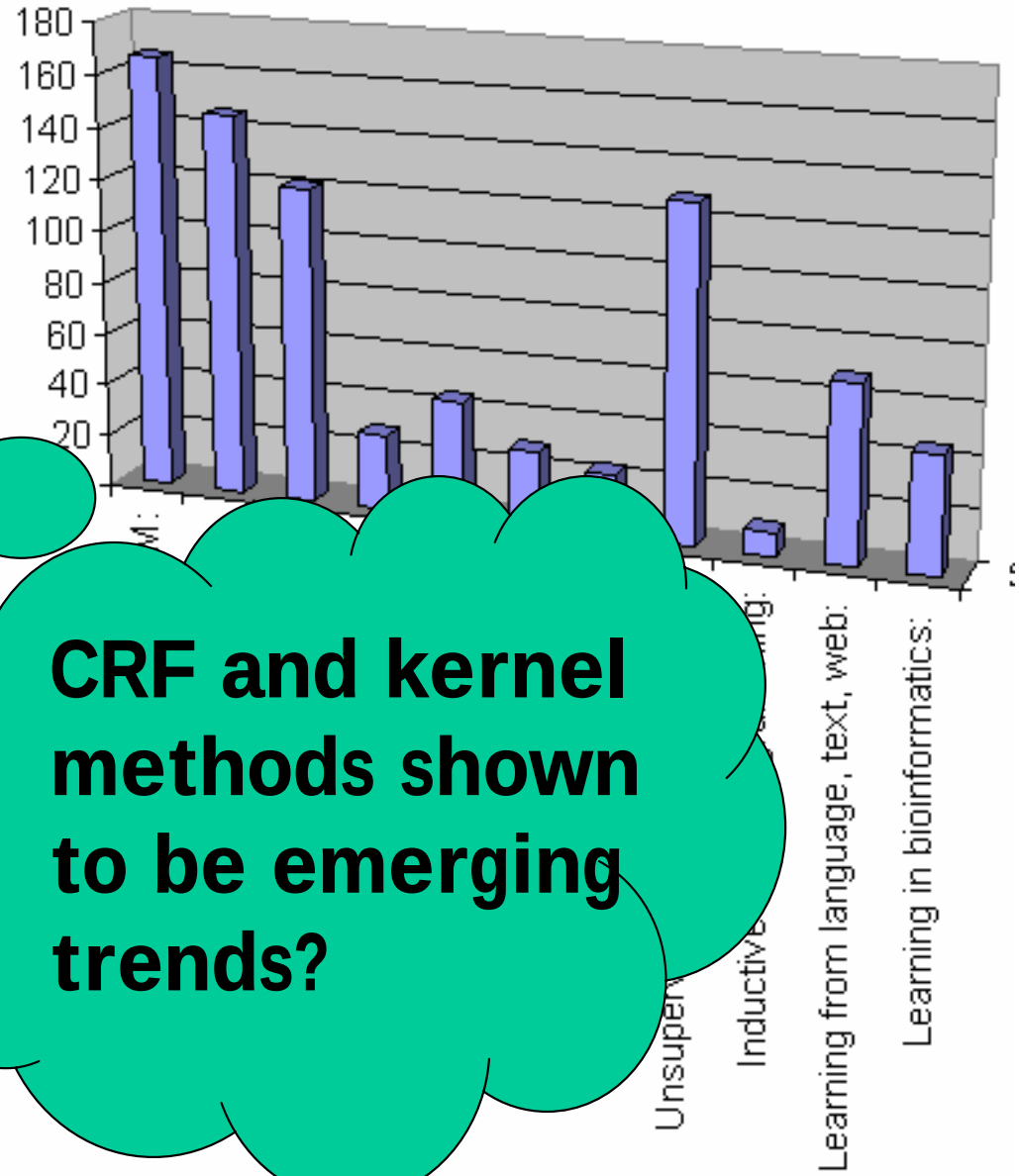
ICML 2006 (720 abstracts)

Kernel methods & SVM	**166**
Probabilistic, graphical models	146
Unsupervised learning, clustering	128
Statistical models	121
Language, Text & web	68
Learning in bioinformatics	45
ANN	29
ILP	9
CRF	**13**



ICML 2006 (720 abstracts)

Kernel methods & SVM	**166**
Probabilistic, graphical models	146
Unsupervised learning, clustering	128
Statistical models	121
Language, Text & web	68
Learning in bioinformatics	45
ANN	29
ILP	9
CRF	**13**



CRF and kernel methods shown to be emerging trends?

Outline

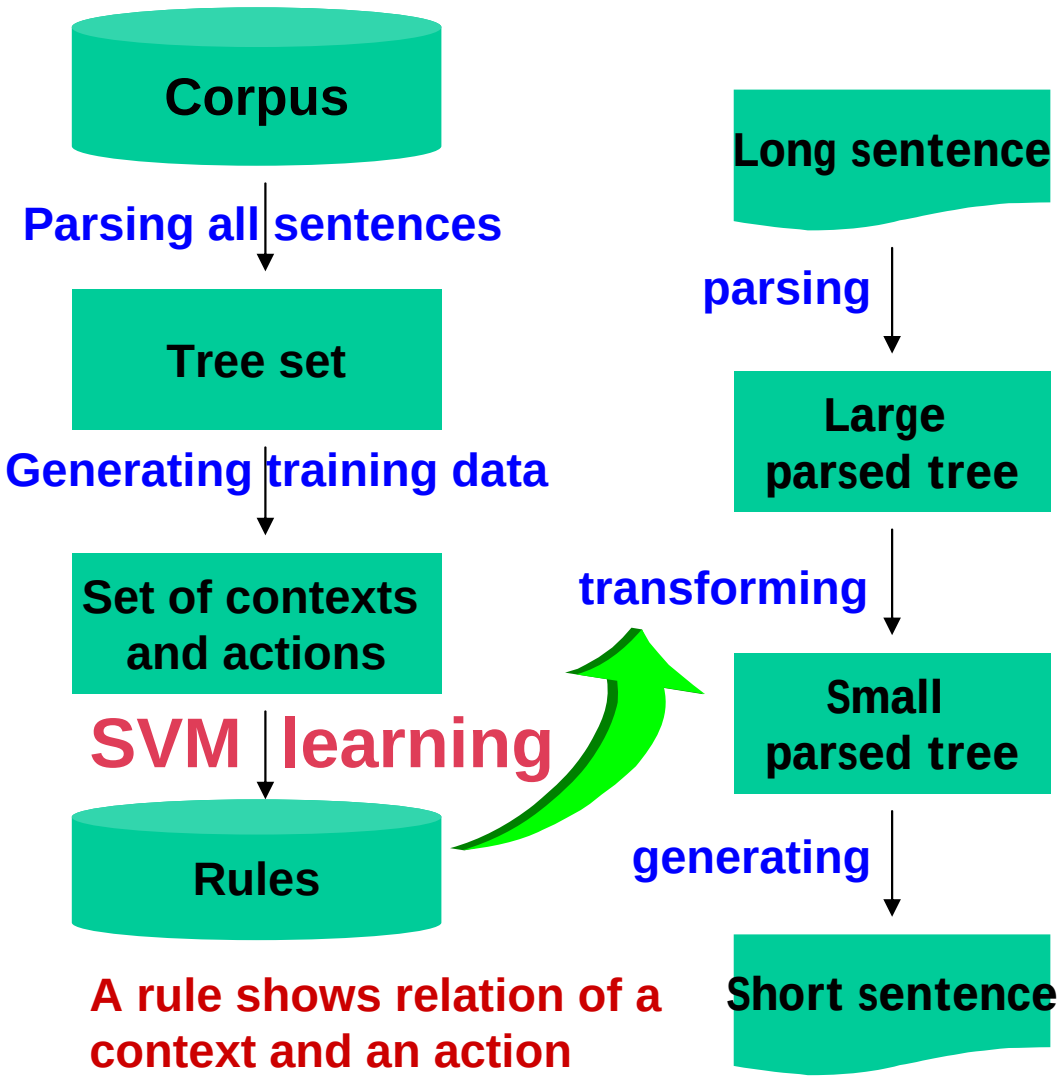
Some emerging trends

Our recent work in these trends

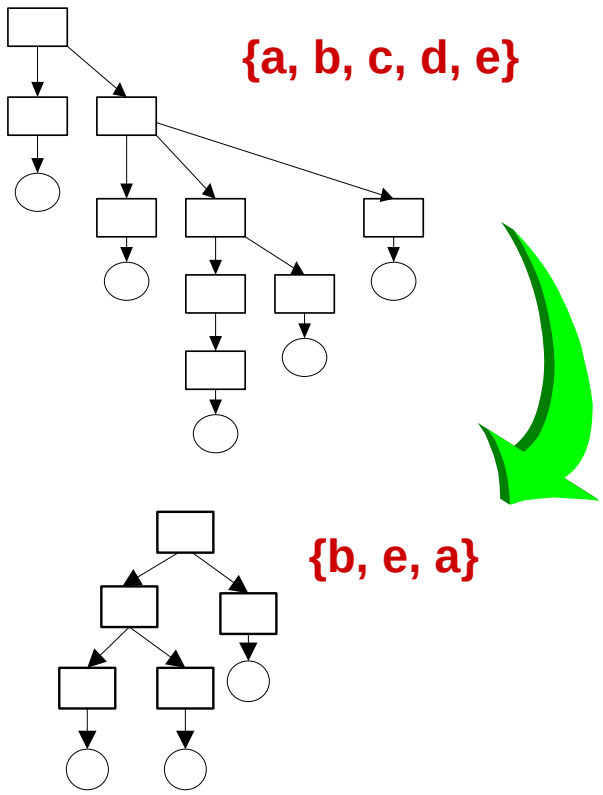
Summary

- Improving prediction performance of CRFs (**KDD'05**)
- High-performance training of CRFs for large-scale applications (**HPCS'06, ICML'06**)
- Sentence reduction (in text summarization) by SVM (**COLING'04**)
- Simplifying support vector machines (**ICML'05, IEEE Trans. Neural Network**)
- Prediction and analysis of β-turns in protein structures (**GIW'03, JBCB'05**) and histone modifications by SVM (**GIW'05**) and CRFs (**ICMLB'06**)
- Manifolds in imbalanced data learning **(ICML'06)**
- Model for emerging trend detection (**PAKDD'06, KSSJ**)

Sentence reduction by SVM
(Minh et al., COLING'04)



- Input list, CSTACK, RSTACK
- Actions: SHIFT, REDUCE, DROP, ASSIGN TYPE, RESTORE
- Transforming tree is a sequence of actions



Prediction of β-turns & γ-turns by SVM
(Hoan et al., GIW'03, JBCB'05)

Protein sequence

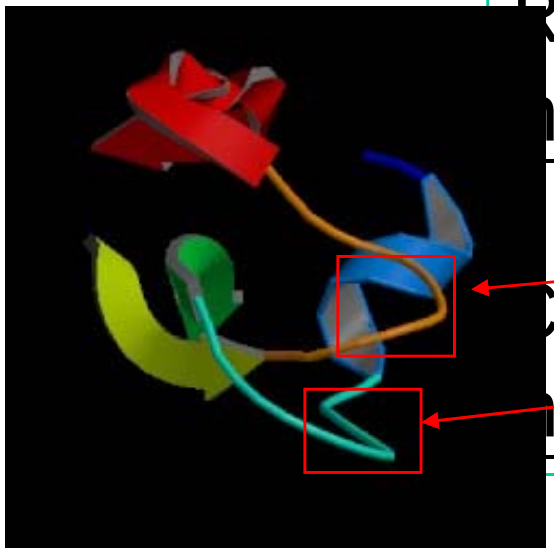
RPDFCLEPPYTGPCKARIIRYFYNAKAGL
CQTFVYGGCRAKRNNFKSAEDCMRTCGGA

Predict

β-turns

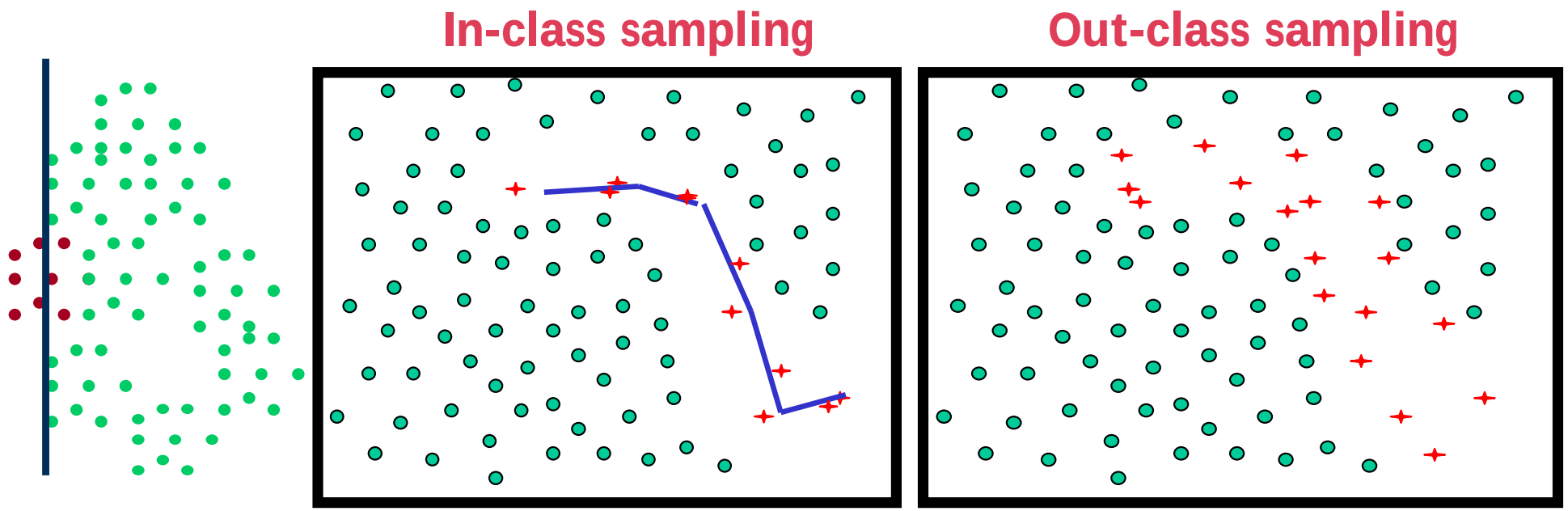
RPDFCLEPPYTGPCKARIIRYFYNAKAGL
nnnnnnnnnnnnnnnnnnnnnnnnTTtttn

CQTFVYGGCRAKRNNFKSAEDCMRTCGGA
nnnnnnnnnnnTtttnnnnnnnnnnnnnn



Manifold for imbalanced data learning
(Hao & Bao, ICML'06)

- Flexible assumption: Data having manifold structures.
- Up sampling data to make it exhibit manifold structures → give rise to patterns of interest.
- Our algorithms outperform SVMs and SMOTE (Chawla et al, JAIR'02).



High-performance CRFs
(Hieu et al., HPCS'06)

- Training CRFs (i.e., estimating parameters for CRFs on the training data $\mathcal{D} = \{(\mathbf{o}^{(j)}, \mathbf{l}^{(j)})\}_{j=1}^{N}$ is to maximize the log-likelihood function:

$$L = \sum_{j=1}^{N} \log\left(p_\theta(\mathbf{l}^{(j)}|\mathbf{o}^{(j)})\right) - \sum_k \frac{\lambda_k^2}{2\sigma^2},$$

where

$$p_\theta(\mathbf{s}|\mathbf{o}) = \frac{1}{Z(\mathbf{o})}\exp\left(\sum_{t=1}^{T} \mathbf{F}(\mathbf{s}, \mathbf{o}, t)\right)$$

- Problem: very expensive due to the computation of partition function Z(o)
- Solution: **Training CRFs on massively parallel computers**

High-performance CRFs
(Hieu et al., HPCS'06)

- Experimental environment:
  - ➔ Massively parallel computer (Cray XT3): 90 nodes, each node has four 2.4GHz processors, 32GB RAM (total: 90 x 4 x 2.4GHz processors, 2.88TB RAM)
  - ➔ Linux OS and MPI library

- Experimental data:
  - ➔ Wall Street Journal corpus of Penn TreeBank
  - ➔ Text chunking and Part-of-speech tagging

Computational linguistics

text

Lexical / Morphological Analysis

Tagging

Chunking

Syntactic Analysis

Grammatical Relation Finding

Named Entity Recognition

Word Sense Disambiguation

Semantic Analysis

Reference Resolution

Discourse Analysis

meaning

Shallow parsing

The woman will give Mary a book

POS tagging

The/Det woman/NN will/MD give/VB
Mary/NNP a/Det book/NN

chunking

[The/Det woman/NN]$_{NP}$ [will/MD give/VB]$_{VP}$
[Mary/NNP]$_{NP}$ [a/Det book/NN]$_{NP}$

relation finding

subject

[The woman] [will give] [Mary] [a book]

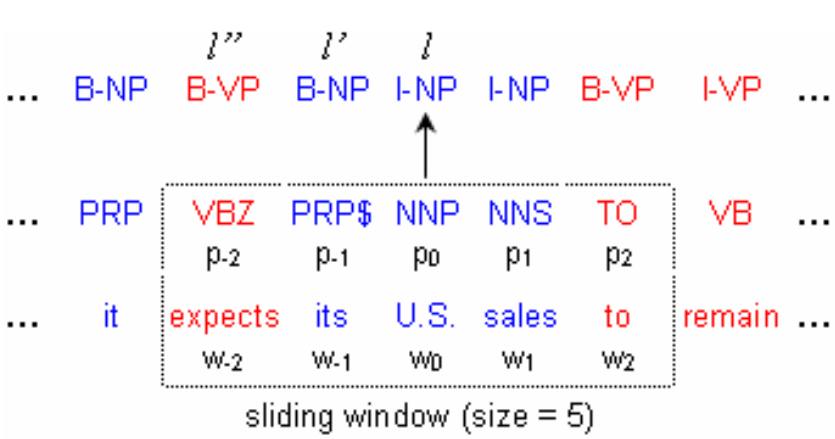
i-object

object

High-performance CRFs
(Hieu et al., HPCS'06)

- Contributions:
  - ➔ Investigate the learning power of CRFs on large-scale dataset
  - ➔ Reduce the training time dramatically
- Text chunking result:
  - ➔ Training: 39,832 sentences of the sections
  - ➔ Testing: 1,921 sentences of the section 00

state-of-the-art accuracy
22.93% error reduction rate in comparison with the previous work



Methods	NP $F_{\beta=1}$	All $F_{\beta=1}$
Ours (majority voting among 16 CRFs)	96.74	96.33
Ours (CRFs, about 1.3M - 1.5M features)	96.59	96.18
Kudo & Matsumoto 2001 (voting SVMs)	95.77	–
Kudo & Matsumoto 2001 (SVMs)	95.34	–
Sang 2000 (system combination)	94.90	–

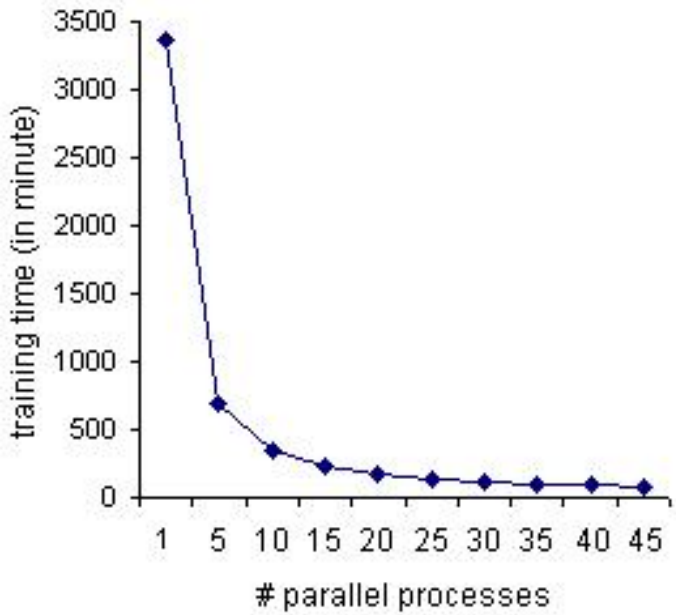
High-performance CRFs
(Hieu et al., HPCS'06)

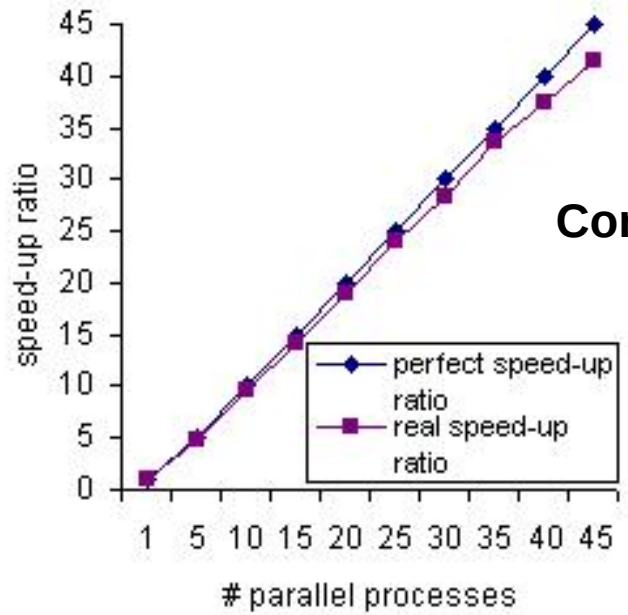
- Experiments for POS tagging
  - ➔ 24 sections of WSJ (Penn TreeBank): about 1,000,000 words (more than 40,000 English sentences)
  - ➔ Achieved competitive results

Comparative results

Methods	Devel. Acc.%	Test Acc.%
Toutanova et al. 2003 (Dependency Net.)	97.15	97.24
Ours (Second-order CRFs (3D Rep.))	**97.05**	**97.16**
Collins 2002 (Discriminative HMMs)	97.07	97.11
Ours (First-order CRFs)	**96.92**	**96.92**

Computational time reduction



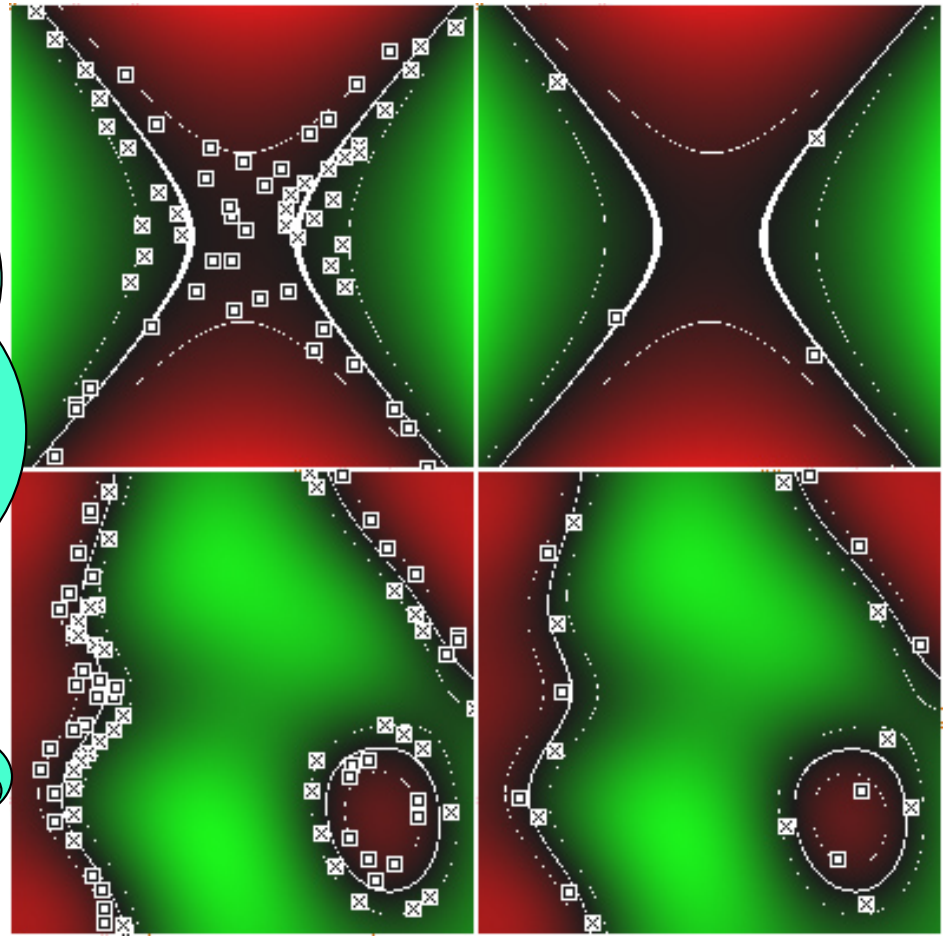


Computational speed-up

SVMs simplification
(DucDung & Bao, ICML'05, IEEE Trans. NN)

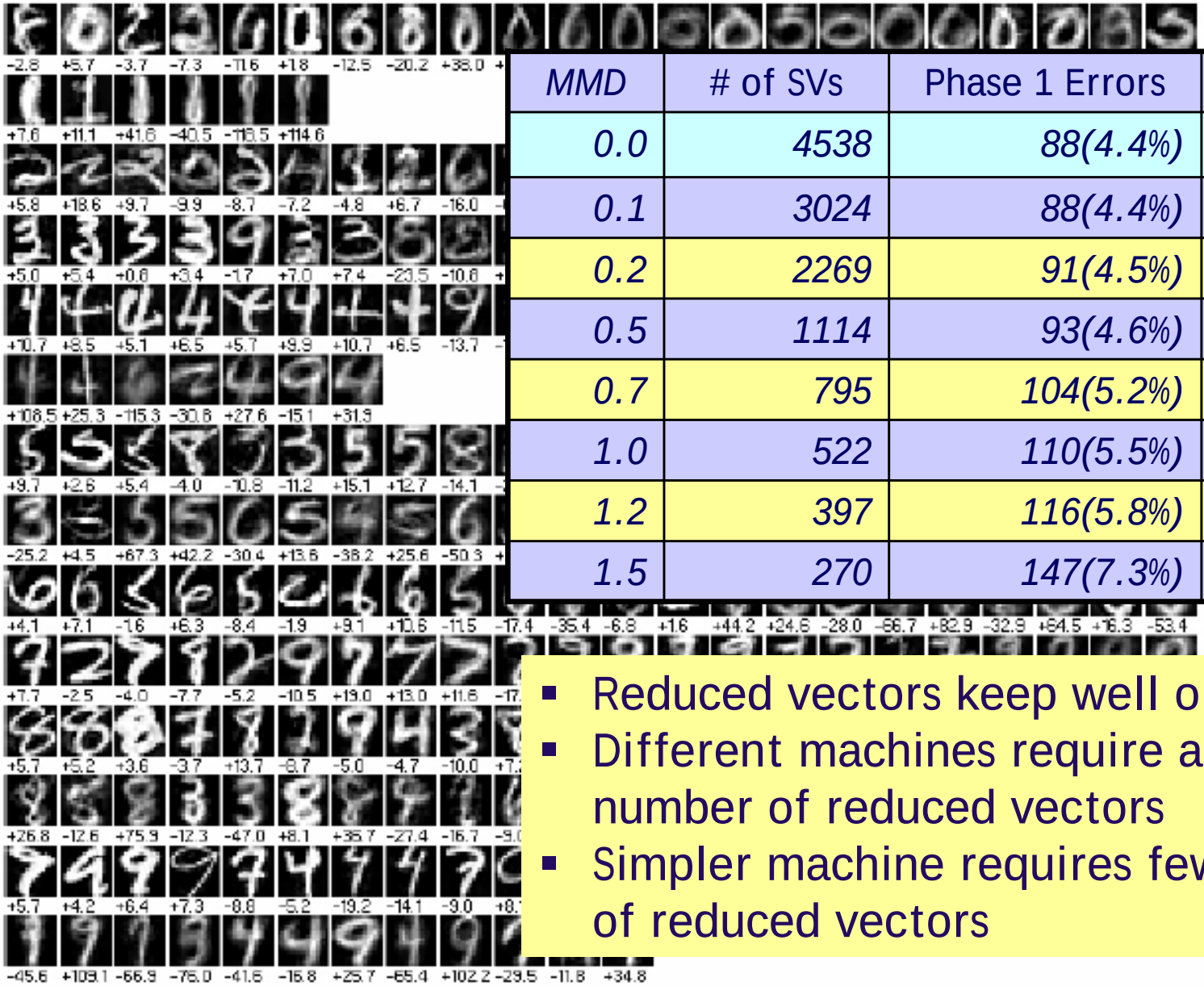
To replace original machine

Bottom-up approach that finds solution in a univariable function instead of multivariable ones in previous methods



$\{(z_i,\beta_j)\}_{j=1,\ldots,Nz}$ – reduced vectors

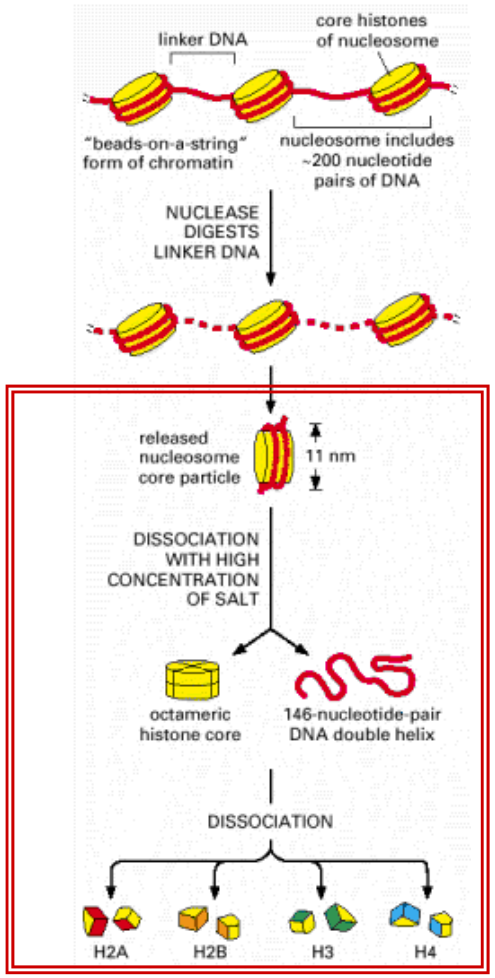
SVMs simplification: evaluation



MMD	# of SVs	Phase 1 Errors	Phase 2 Errors
0.0	4538	88(4.4%)	88(4.4%)
0.1	3024	88(4.4%)	88(4.4%)
0.2	2269	91(4.5%)	88(4.4%)
0.5	1114	93(4.6%)	89(4.4%)
0.7	795	104(5.2%)	89(4.4%)
1.0	522	110(5.5%)	91(4.5%)
1.2	397	116(5.8%)	93(4.6%)
1.5	270	147(7.3%)	95(4.7%)

- Reduced vectors keep well original shape
- Different machines require a different number of reduced vectors
- Simpler machine requires fewer number of reduced vectors

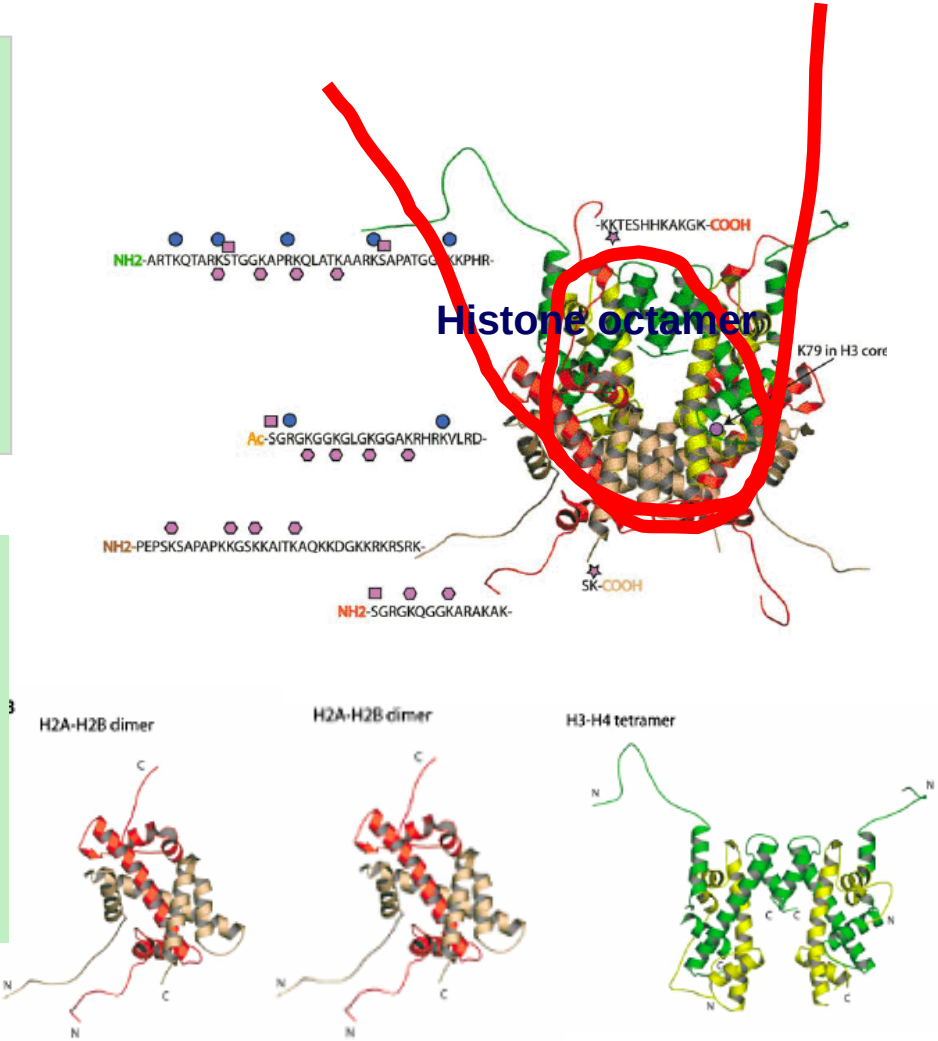
Prediction of histone modifications
(Hoan et al., GIW'05)



From Molecular biology of the Cell.

146 pairs **of DNA** in nucleosomes are wrapped around a core of histone proteins,

Histone octamer consists of 8 proteins: a H3-H4 tetramer and two H2A-H2B dimers



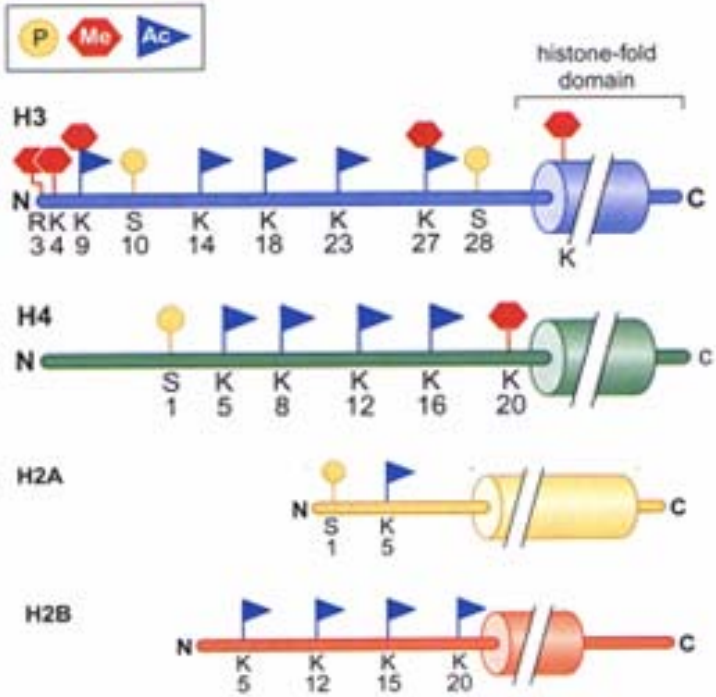
Prediction of histone modifications

About histone modification

Histone modifications: Some amino acids of histone proteins (H3, H4, H2A, H2B) in nucleosomes are modified by added methyl group (methylation), acetyl group (acetylation), or other chemical groups. Most of them are in N-terminal tails that are highly conserved.

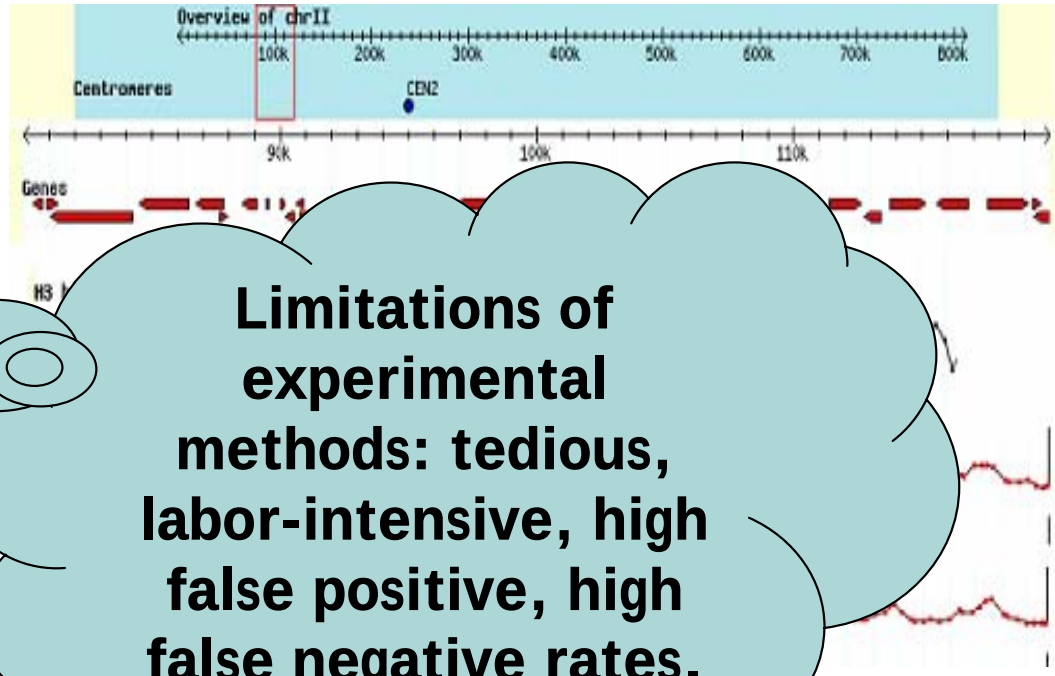
Nucleosome modifications

- **Methylation** → **H3**:R3,K4,K9,K27 **H4**:K20;
  Adding one, two or three $-CH_3$'s to a.a.
- **Acetylation** → **H3**:K9,K14,K16, K23,K27 **H4**:K5,K8,K12; etc.
  Adding acetyl $(CH_3C)O$
- **Phosphory-lation** → **H3**:S10,S25 **H4**:S1; **H2A**:S1
  Adding $-PO_3$ to a.a.



Prediction of histone modifications
Experimental approach

- Recent **experimental studies** on relative occupancy and modification state of nucleosomes (Bernstein, 2002; Kurdistani, 2004; Humphrey, 2004; Lee, 2004; Kurdistani, 2004; Pokholok et al., Cell, 8.2005)
- **Quantitative measurements** of histone occupancy and modifications at positions in DNA sequences with different resolution levels.



Limitations of experimental methods: tedious, labor-intensive, high false positive, high false negative rates.

Fro
in vivo
by experimental method

Prediction of histone modifications

Research objective: Computational solution?

From DNA sequences

CACTACGGGCCTGTGTACATTCTGCGCGACATTCACCCAGTGTGCAGTGTGAGAGGTACAGGTGGCGCATGTGGTGTGCGCCACACACGTTGGCACC

To **computationally** predict:
- H3, H4 occupancy
- Acetylation state
- Methylation state

To find characteristics of areas at which H3, H4 occupancy, histone acetylation and methylation are at high and low levels.

In our work, we define two state classes of H3, H4 occupancy, acetylation, methylation: **high and low**.

Prediction of histone modifications

Convert a DNA sequence into vector using k-grams

AATTTTTATAGGTCGACCAATCTGTCG

No.	Fea.	Occ.
1	**AAT**	**2**
2	**ATT**	**1**
3	**TTT**	**3**
4	**TTA**	**1**
…	…	…

```
4:2 5:2 7:1 13:1 17:17 18:1 19:1 20:5 21:8 23:1 28:1 29:1 31:1
0:3 1:1 2:2 3:2 5:1 8:2 9:1 12:3 14:1 15:3 18:1 19:2 21:1 22:2
0:17 1:1 2:4 3:2 4:2 5:1 7:2 8:1 10:4 11:2 12:4 13:1 14:1 15:2
0:7 1:1 2:3 3:4 7:2 8:2 9:1 10:1 12:3 13:1 14:1 15:3 16:1 27:1
0:2 1:1 2:1 3:2 4:1 5:2 6:1 3:2 15:1 16:1
0:8 1:3 2:3 3:4 4:3 13:1 14:3 1
0:4 1:1 2:2 3:3 4: 15:3 16:2 1
0:7 1:1 2:2 3:5 4: 4:1 15:3 1
0:5 1:2 2:2 3:3 17:2 18:1
1:1 4:1 6:1 7: 1 26:1 30
3:1 4:2 7:1 11 1:2 34:2
1:1 2:2 7:1 8:2 :2 25:1 27
0:1 2:3 6:1 10: 27:1 28:1
0:6 2:3 3:2 4:1 16:1 18:1 19:
0:4 1:2 2:1 3:2 5. 14:2 15:2 18:1
0:13 1:2 2:4 3:4 4:1 11:2 12:2 14:1 15:3
0:6 1:1 2:1 3:4 5:1 7:1 12:3 15:3 16:1 17:2 20:1
0:3 1:1 3:3 4:1 10:1 12:3 14:3 15:2 16:2 20:1 28:1 31:1 3
0:1 1:1 2:1 3:2 6:2 8:3 10:1 11:2 12:2 14:1 19:1 24:1 26:1 27:
0:4 1:1 2:4 4:1 7:1 8:1 9:1 10:2 11:1 12:1 14:1 17:9 18:2 19:3
0:2 3:2 4:1 5:1 8:1 10:3 11:1 14:2 15:3 16:1 18:2 21:1 23:2 25
0:1 2:1 3:1 4:1 6:1 7:1 8:1 10:4 11:1 14:1 15:1 16:1 17:1 18:2
1:3 3:1 4:1 6:2 7:2 8:1 9:1 10:3 13:1 14:1 15:1 16:3 18:1 22:1
```

Unsupervied data

Prediction of histone modifications

Prediction results (K) by SVM and CRF

Dataset	k=3		k=4		k=5		k=6	
	acc	cc	acc	cc	acc	cc	acc	cc
H3 occupancy	84.93	0.70	*85.88*	*0.72*	85.50	0.71	85.10	0.70
H4 occupancy	85.91	0.71	87.14	0.74	*87.77*	*0.75*	85.95	0.75
H3K9ac	71.04	0.41	73.64	0.47	75.58	0.51	77.27	0.54
H3K14ac								
H4ac								
H3K4me1								
H3K4me2								
H3K4me3								
H3K36me3								
H3K79me3	78.25	0.56	79.91	0.60	80.87	0.61	82.15	0.64

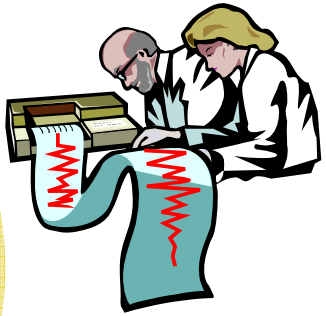
- The accuracy and correlation coefficient of qualitative prediction are consistent with experimental approach.
- The highest prediction of H3 and H4 occupancy achieved when using 4 or 5-gram features.

Sequence length L = 500

A model for emerging trend detection
(Hoang & Bao, KSS journal)

ETD: Detecting **topics** that are **growing** in **interest** and **utility overtime** from a corpus

Topic verification
How to define interest and utility functions and evaluate their increase overtime?



M = (D, E, T, TR, TI, TV, f, g)

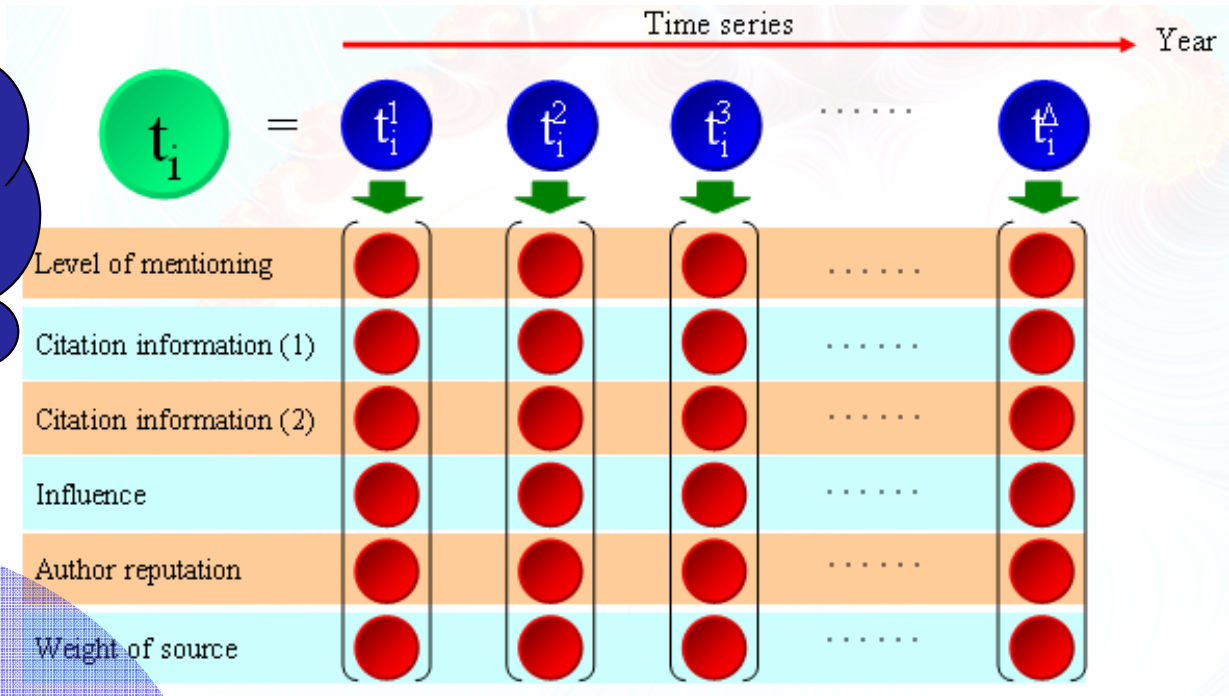
Topic representation
Which features are necessary to characterize topics (interest and utility overtime)?

Topic identification
How to extract these features from the corpus for each topic?

ETD: Topic representation

ETD: Detecting **topics** that are **growing** in **interest** and **utility overtime** from a corpus

Define 6 types of citation



Topic representation
Which features are necessary to characterize topics (interest and utility overtime)?

neural network

t_i=NNs	1998	1999	2000	2001	2002	2003
t_i^k (1)	0.06	0.10	0.08	0.10	0.09	0.06
t_i^k (2)	0.20	0.33	0.28	0.06	0.11	0.04
t_i^k (3)	0.41	0.40	0.50	0.12	0.07	0.32
t_i^k (4)	0.17	0.40	0.06	0.12	0.33	0.02
t_i^k (5)	0.65	0.55	0.13	0.24	0.67	0.11
t_i^k (6)	0.33	0.44	0.22	0.33	0.44	0.56

ETD: Topic identification

ETD: Detecting **topics** that are **growing** in **interest** and **utility overtime** from a corpus

- Build 6 models corresponding to 6 types of citation
- Using HMM, MEMM, an CRF to extract features

$$P_O(\lambda_i) = \frac{P(O|\lambda_i)}{\sum_j P(O|\lambda_j)}$$

$$P(O \mid \lambda_i) = \max_s P(s, o|\lambda_i)$$

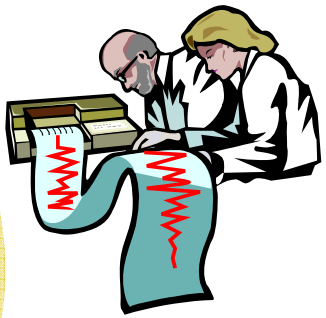
$$H_O(\lambda_i) = -\sum_i P_O(\lambda_i).\log(P_O(\lambda_i))$$

Topic identification
How to extract these features from the corpus for each topic?

ETD: Topic verification

ETD: Detecting **topics** that are **growing** in **interest** and **utility overtime** from a corpus

Topic verification
How to define interest and utility functions and evaluate their increase overtime?

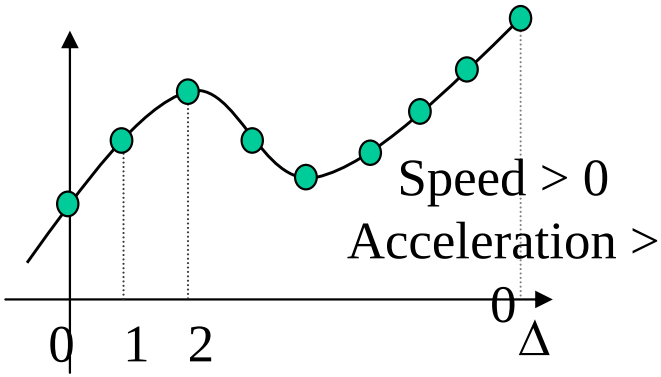


$$Growth(t_i, j) = growth\ of\ time\text{-}series\ \{t_i^k(j)\}_k\ along\ the\ time\ axis$$

$$Interest\ \ f(t_i) = \frac{1}{4}\sum_{j\in\{1,3,5,6\}} Growth(t_i, j),\ \ Utility\ g(t_i) = \frac{1}{4}\sum_{j\in\{2,4,5,6\}} Growth(t_i, j)$$

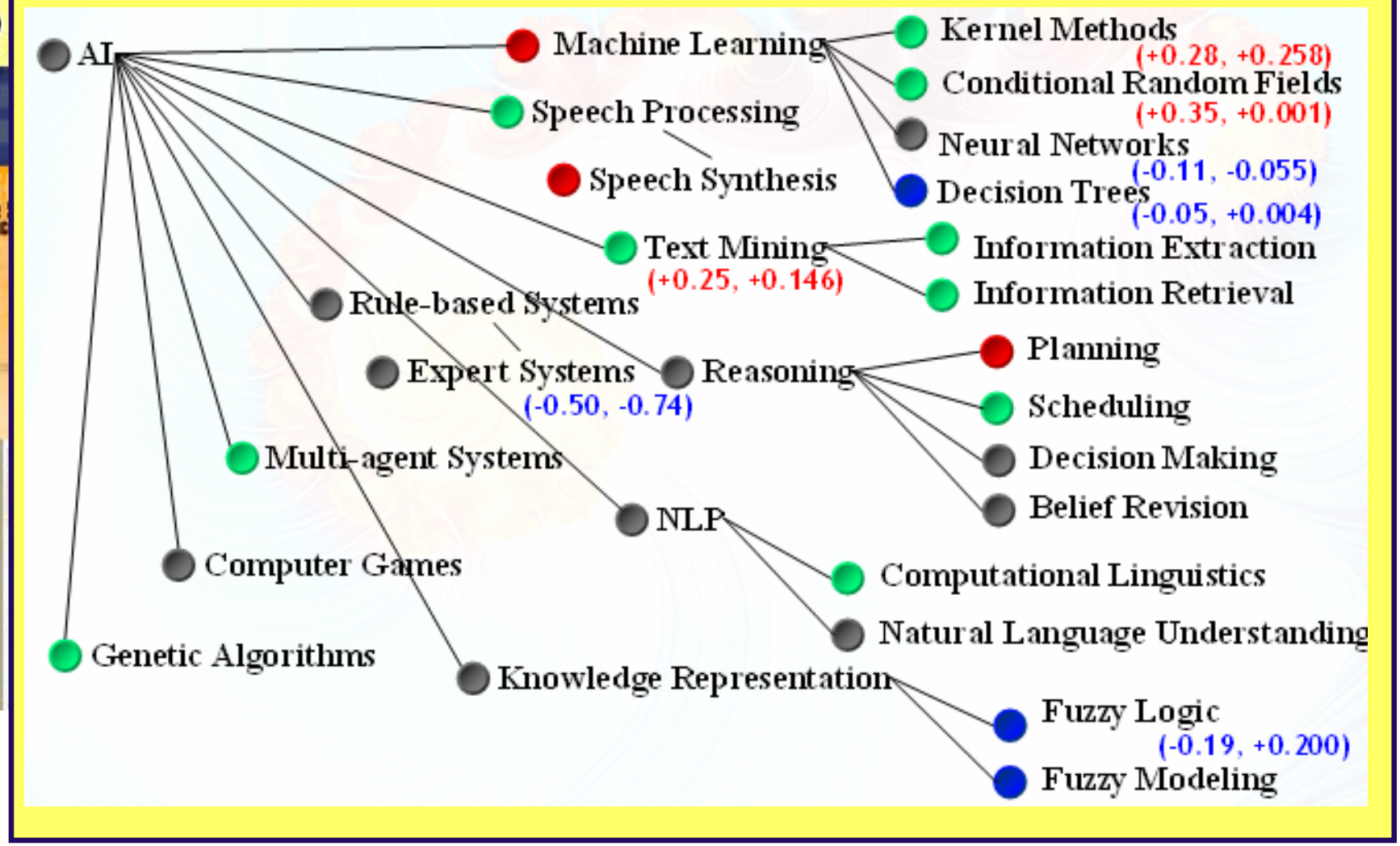
$$f'(k) = \frac{\partial f}{\partial x}(k)$$ the speed of growing at x = k

$$f''(k) = \frac{\partial^2 f}{\partial x^2}(k)$$ the acceleration of growing at x = k



ETD: Evaluation

ETD: Detecting **topics** that are **growing** in **interest** and **utility overtime** from a corpus



Conclusion

- Much complexly structured data.
- Two emerging trends in machine learning and data mining fields: discriminative random fields and kernel methods.
- And some of our works relating to them.
- Good to deal with complexly structured data and are today and future technologies.
- Many open and challenging problems.

Acknowledgments

- Hiroshi Motoda, Yoshiteru Nakamori, Kenji Satou, Phạm Thọ Hoàn, Nguyễn Đức Dũng, Nguyễn Lê Minh, Phan Xuân Hiếu, Lê Minh Hoàng, Nguyễn Cảnh Hào, Saori Kawasaki, Đàm Hiếu Chí, Lê Anh Cường, Huỳnh Văn Nam, Trần Tuấn Nam, Lê Sĩ Quang, Nguyễn Thanh Phương, Trần Đăng Hưng, …
- Projects:
  - ➔ Realization of Active Mining in the Era of Information Flood
  - ➔ Genome Information Science
  - ➔ COE "Technology Creation Based on Knowledge Science"
  - ➔ Multi-Sources Data Mining in Hepatitis Study
  - ➔ Computational Materials Science
  - ➔ Advanced Technology for Cross Language Processing
- Organizers of RIVF'06

See you in RIVF'07 in Hanoi

- Thank you for your attention
- Merci beaucoup pour votre attention
- Cám ơn các anh chị đã chú ý

http://www.jaist.ac.jp/~bao/talks