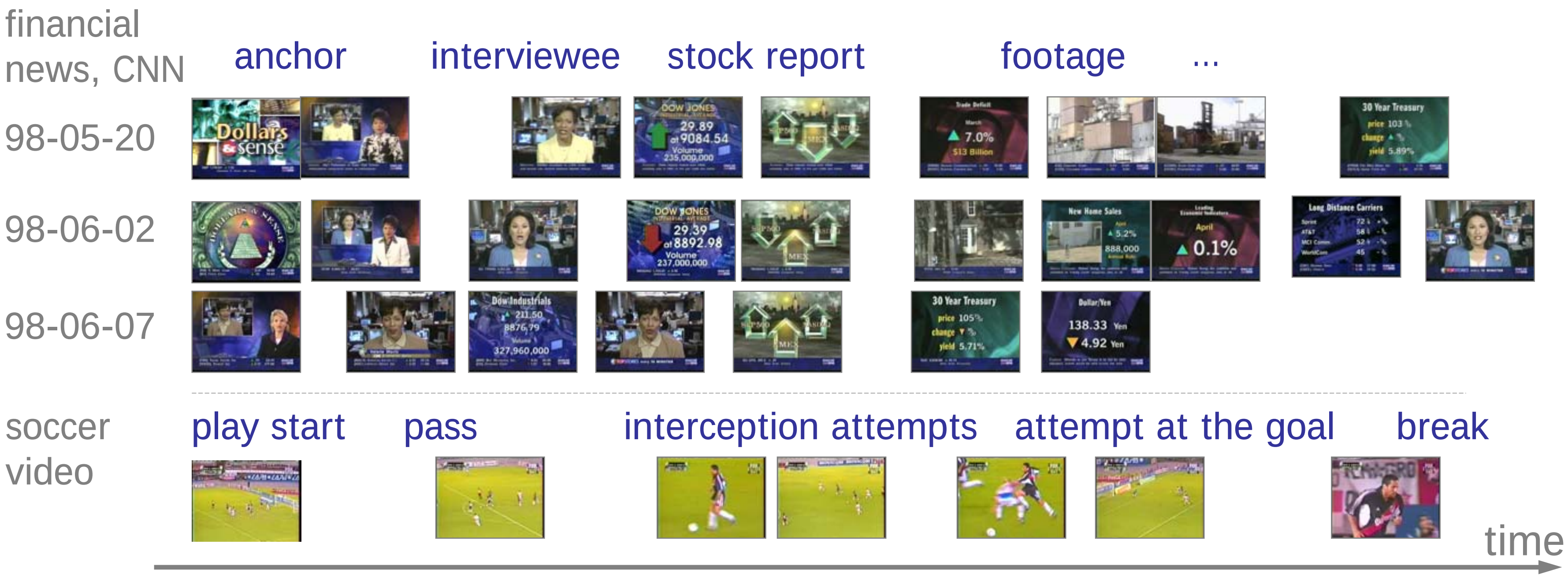
Unsupervised Pattern Discovery for Multimedia Sequences

Lexing Xie, Shih-Fu Chang

Digital Video Multimedia Lab
Columbia University

In collaboration with Dr. Ajay Divakaran and Dr. Huifang Sun (MERL)

The Problem



financial news, CNN

anchor

interviewee

stock report

footage

...

98-05-20

98-06-02

98-06-07

soccer video

play start

pass

interception attempts

attempt at the goal

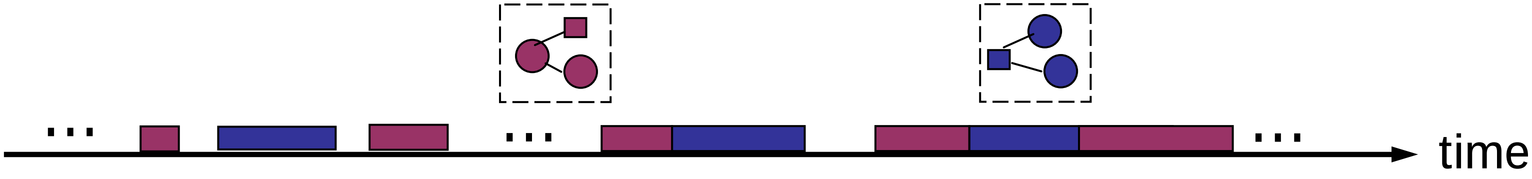
break

time

- Unsupervised pattern discovery: capturing distinct temporal patterns in diverse domains
  - → Suitable computational models
  - → Appropriate content features.

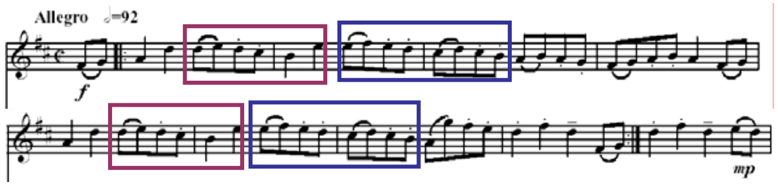
Unsupervised Pattern Discovery

- Recurrent segments with consistent characteristics.
- Find an appropriate model of the temporal pattern;
- Locate segments that match the model.

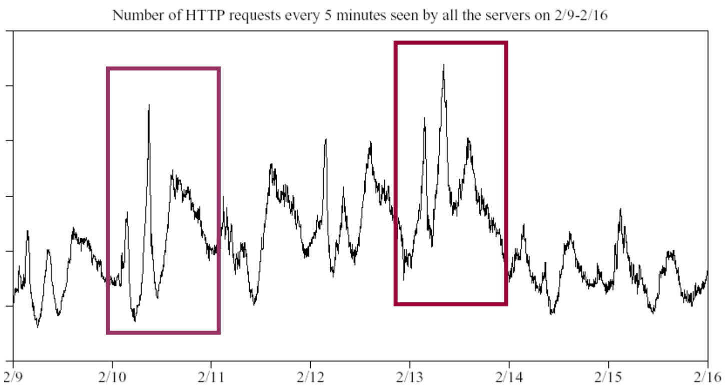


Unsupervised Pattern Discovery

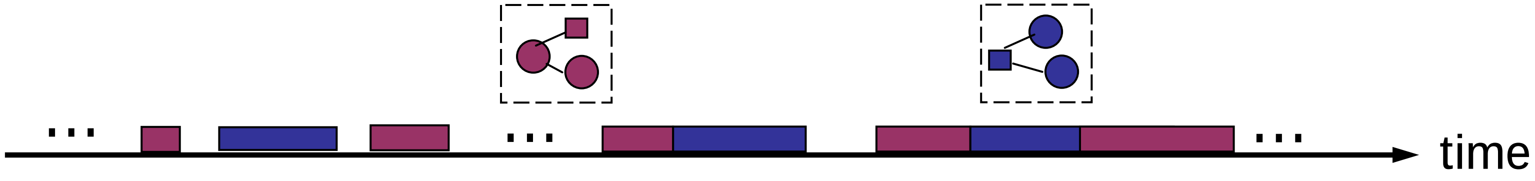
- “Temporal” patterns exist in many different domains.



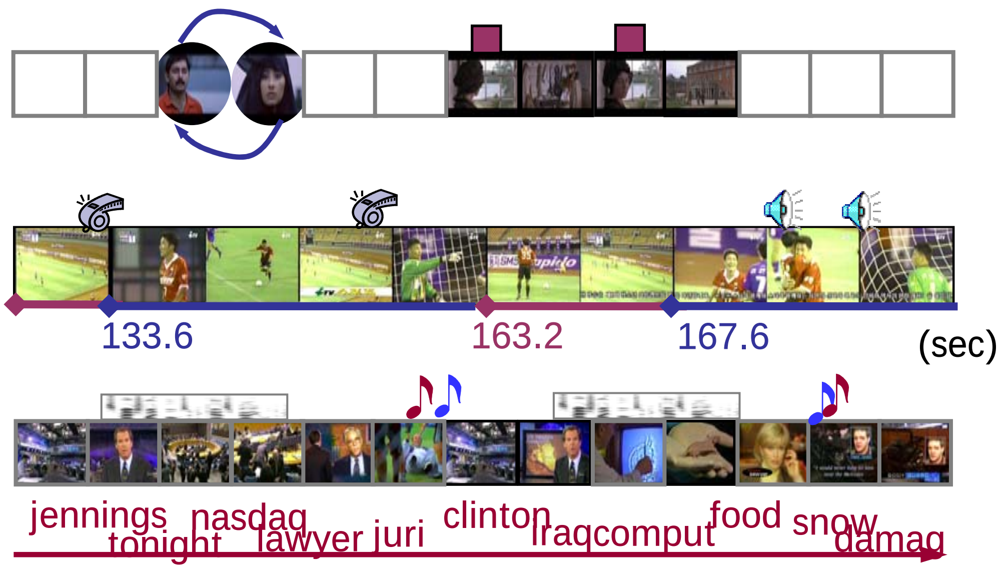
Fis Protein	QTRAALMMGINRGTLRKKLK
λ Rep	QESVADKMGMGQSGVGALFN
λ Cro	QTKTAKDLGVYQSAINKAIH
434 Cro	QTELATKAGVKQQSIQLIEA



[Iyengar99]
[www.music-scores.com][Purdue Stat490B]



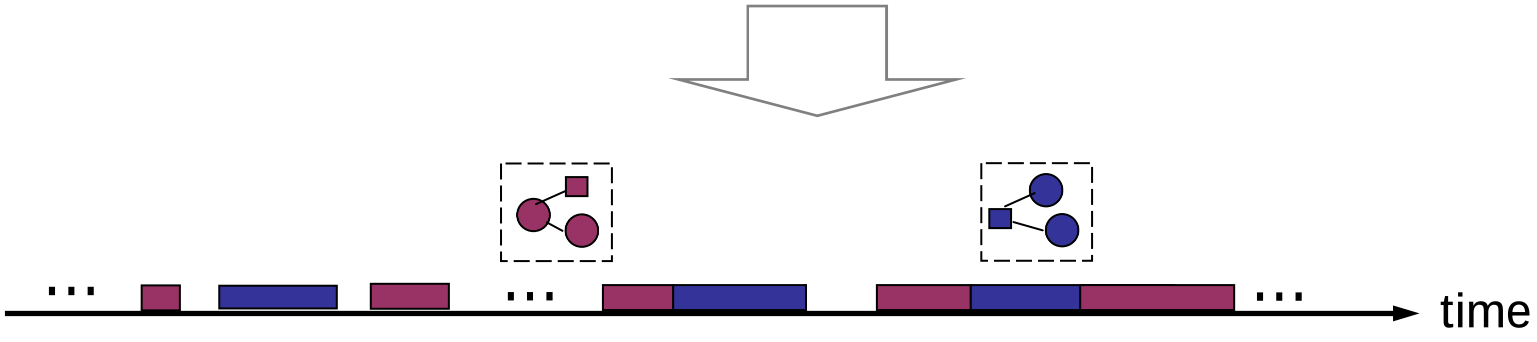
Multimedia Patterns



Need unsupervised discovery:

! Patterns/events unknown a priori

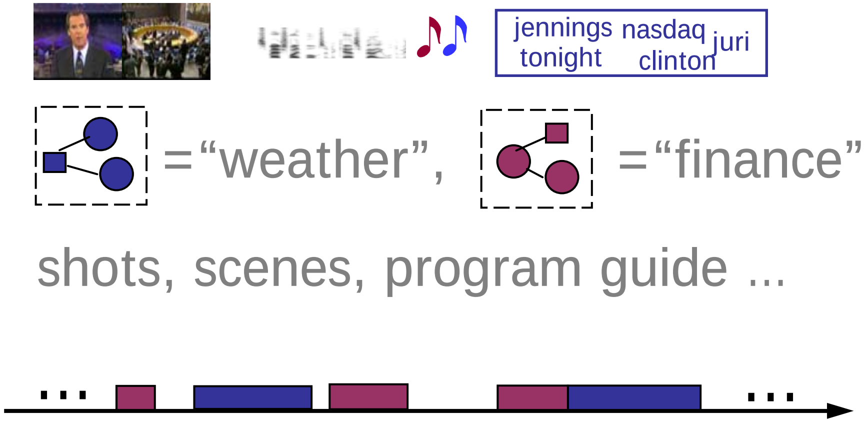
! Annotation very costly (~10x real-time)



Multimedia Pattern Discovery

- Discover meaningful patterns in diverse domains
  - Incomplete domain knowledge
  - Unsupervised non-interactive analysis
- Desired properties
  - Versatile
  - Multi-modal
  - Meaningful
  - Knowledge-adaptive

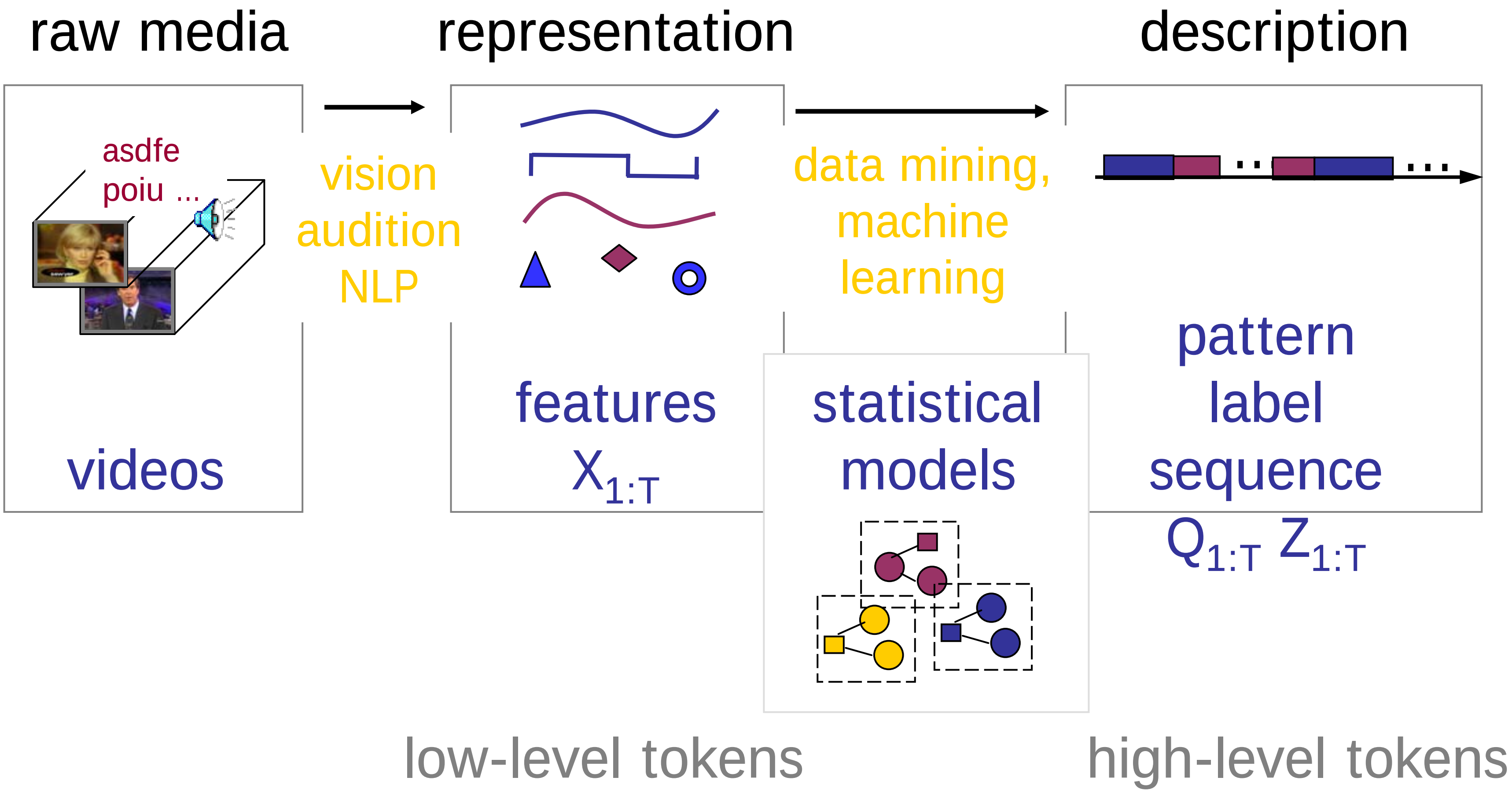
news, sports, surveillance, ...



="weather", ="finance"

shots, scenes, program guide ...

Multimedia Pattern Discovery



raw media

representation

description

videos

vision
audition
NLP

features
$X_{1:T}$

data mining,
machine
learning

statistical
models

pattern
label
sequence
$Q_{1:T}$ $Z_{1:T}$

low-level tokens

high-level tokens

Outline

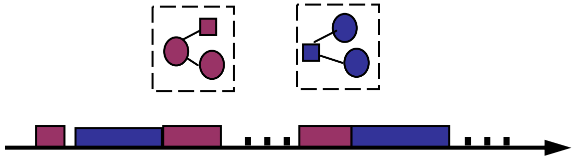
- The problem
- Unsupervised pattern discovery
  → temporal token generation
  - HHMM
  - Automatic feature grouping
- Finding meaningful patterns
  → multi-modal token fusion
- Summary

± Versatile

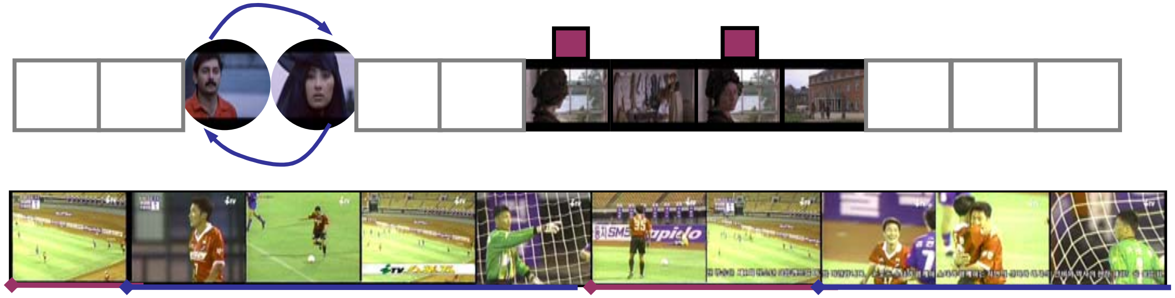
± Multi-modal

± Meaningful

? Knowledge-adaptive



Modeling Video Patterns



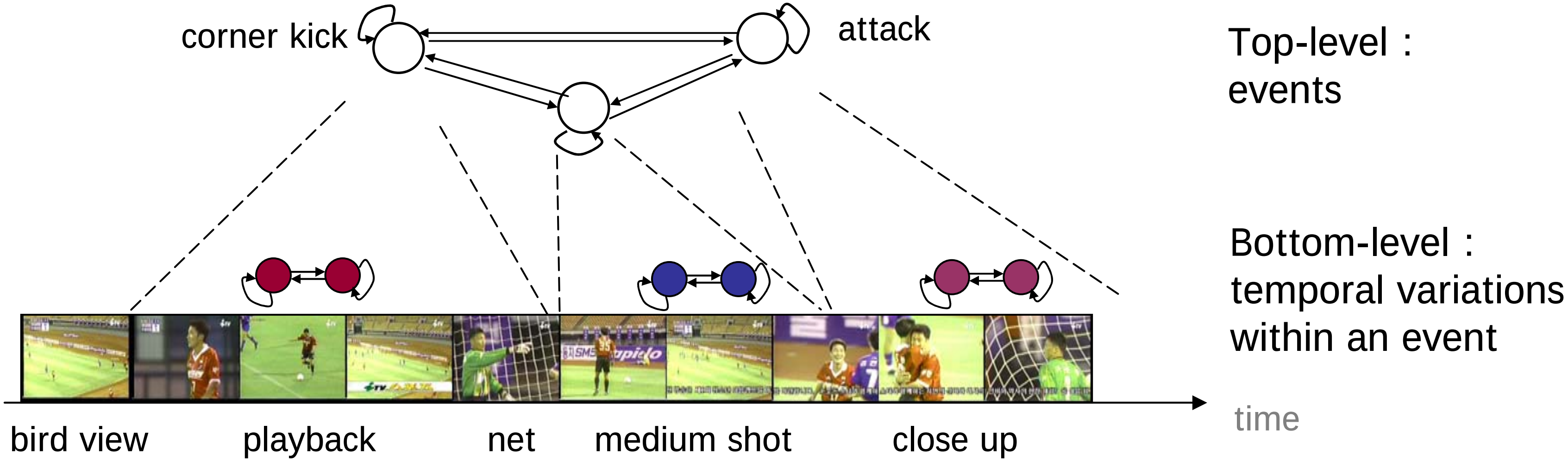
sparse deterministic

dense stochastic √

Modeling Video Patterns with HHMM

Supervised	Tracking, speech, DNA sequence recognition ...	[Fine, Singer, Tishby'98] [Zweig 1997], [Ivanov'00]...
Left-right	Video clustering	[Clarkson'99][Naphade'02]

- Unsupervised: (1) the model? (2) its size? (3) the features?



corner kick

attack

Top-level : events

Bottom-level : temporal variations within an event

time

bird view

playback

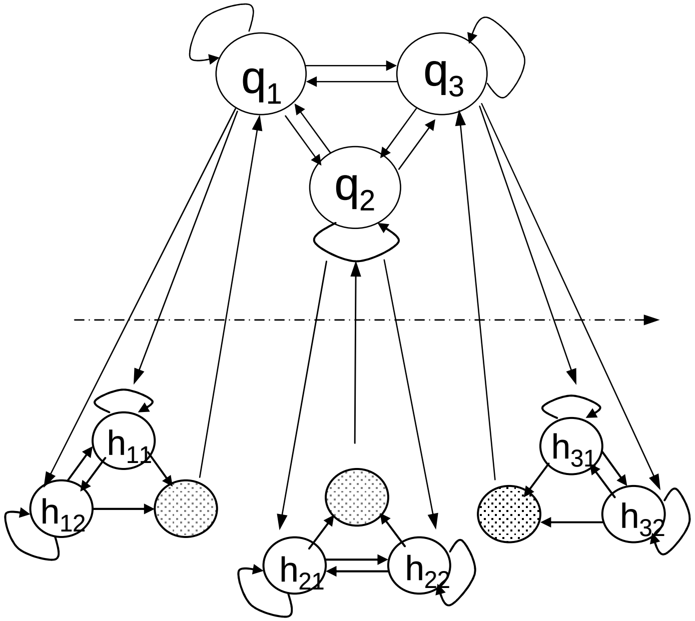
net

medium shot

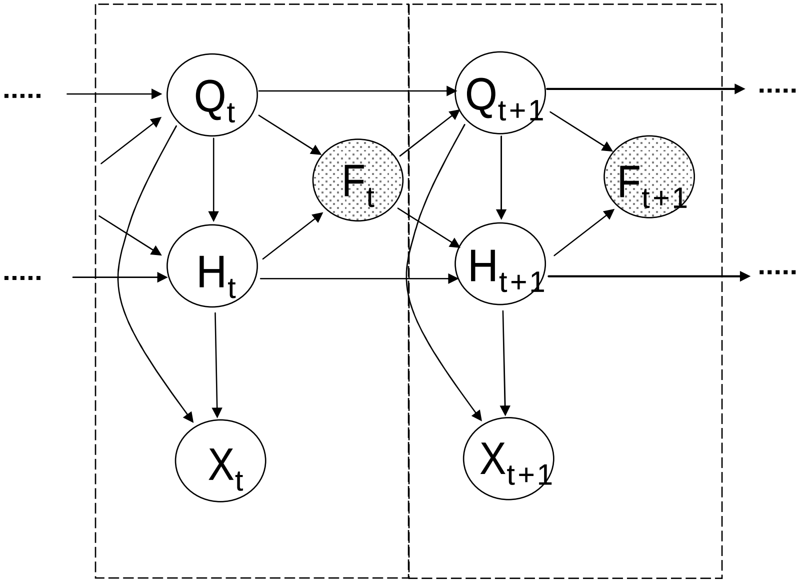
close up

Hierarchical HMM

[Fine, Singer, Tishby'98]
[Murphy'01] [Xie et al. ICME03]



State-space representation

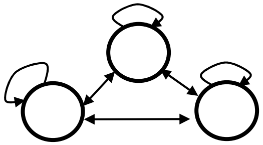


DBN representation, unrolled in time

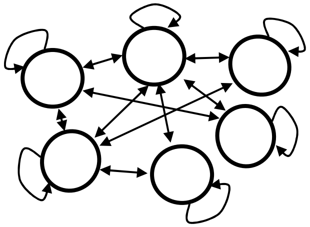
- Emission and transition parameters, $\{\Theta_{top}, \Theta_{bot}\}$
- Inference and estimation in O(T)

The Need for Model Selection

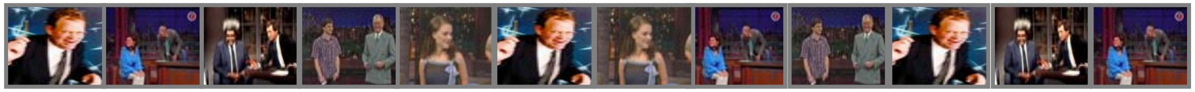
soccer

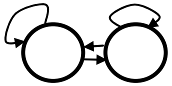


news



talk show

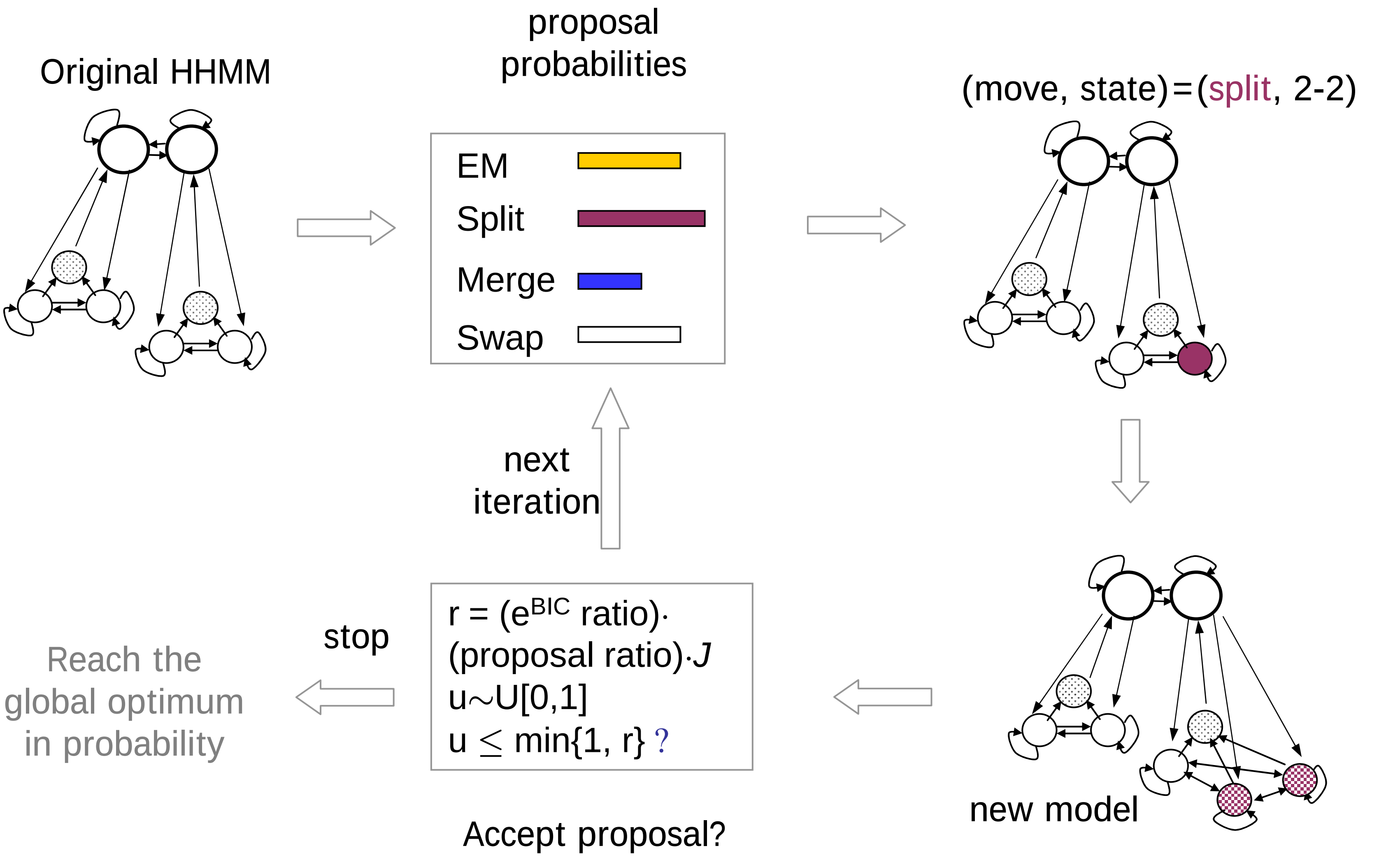




- Different domains have different descriptive complexities.

Model Selection with RJ-MCMC

[Green95]
[Andrieu99]
[Xie ICME03]



Original HHMM

proposal probabilities

EM

Split

Merge

Swap

(move, state)=(split, 2-2)

next iteration

new model

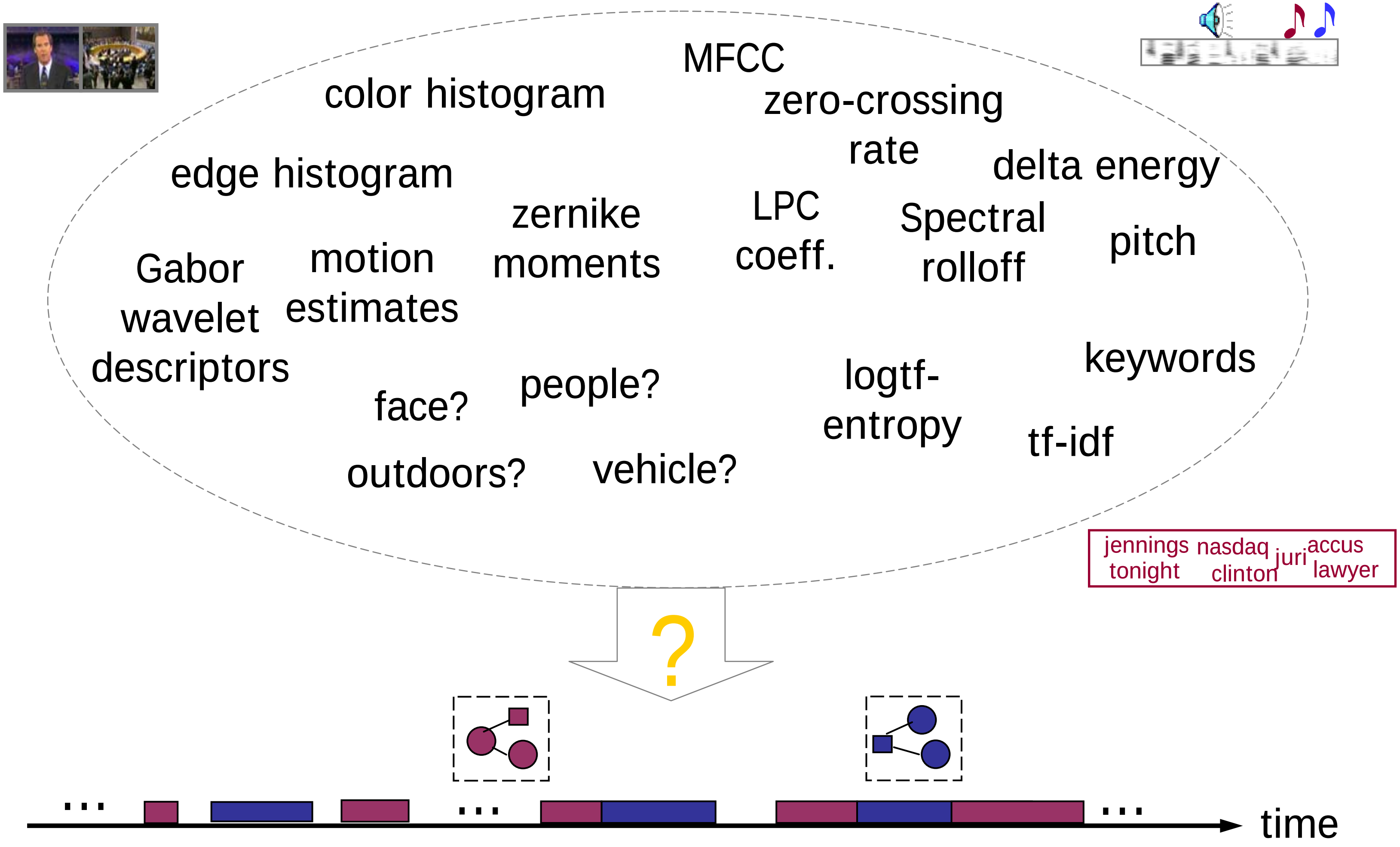
Accept proposal?

r = (e^{BIC} ratio)·(proposal ratio)·*J*
u∼U[0,1]
u ≤ min{1, r} ?

stop

Reach the global optimum in probability

Which Features Shall We Use?



color histogram

MFCC

zero-crossing rate

delta energy

edge histogram

zernike moments

LPC coeff.

Spectral rolloff

pitch

Gabor wavelet descriptors

motion estimates

keywords

face?

people?

logtf-entropy

tf-idf

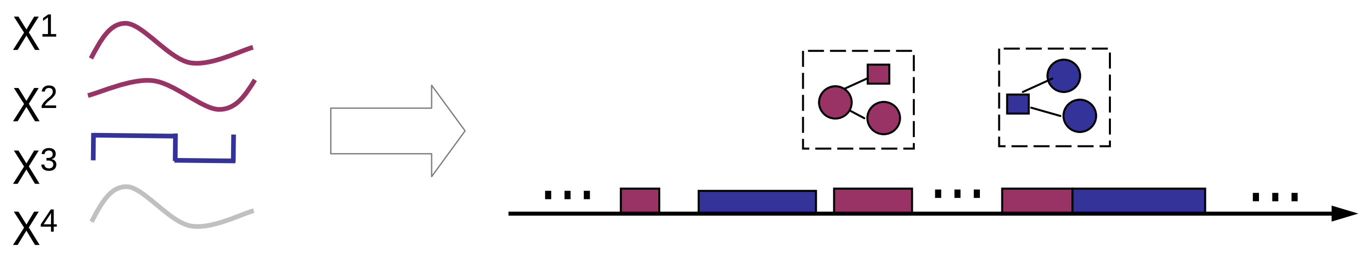
outdoors?

vehicle?

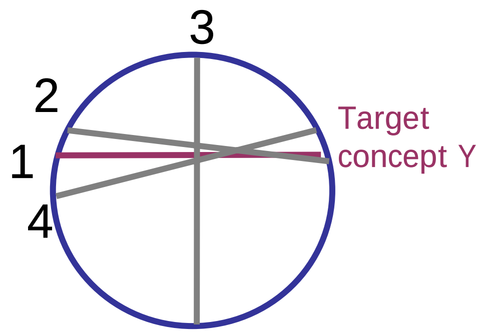
Feature Selection

[Koller,Sahami'96] [Zhu et.al.'97]
[Xing, Jordan'01] [Ellis,Bilmes'00]...

Goal: To identify a good subset of measurements in order to improve generalization and reduce computation.



Criteria: irrelevance and redundancy between the features X and the *target label* Y.



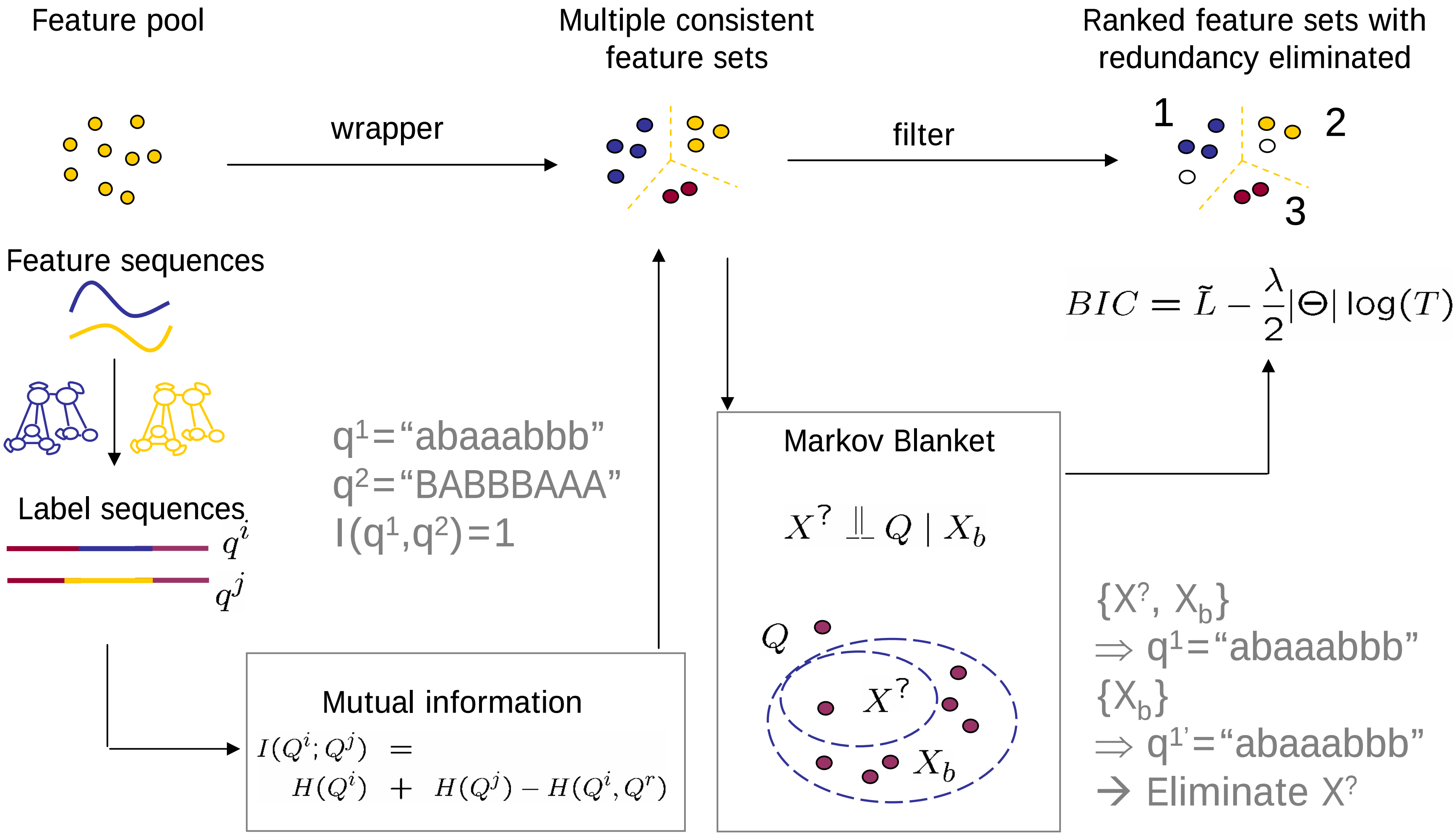
Our problem:

Unsupervised → No canonical target concept

Temporal sequence → Samples not i.i.d.

Feature Selection

[Koller'96] [Xing'01]
[Xie et al. ICIP'03]



Feature pool

wrapper

Multiple consistent feature sets

filter

Ranked feature sets with redundancy eliminated

$$BIC = \tilde{L} - \frac{\lambda}{2}|\Theta| \log(T)$$

Feature sequences

Label sequences

q^i

q^j

q¹="abaaabbb"
q²="BABBBAAA"
I(q¹,q²)=1

Mutual information

$$I(Q^i;Q^j) = H(Q^i) + H(Q^j) - H(Q^i,Q^r)$$

Markov Blanket

$$X^? \perp\!\!\!\perp Q \mid X_b$$

$\{X^?, X_b\}$
⇒ q¹="abaaabbb"
$\{X_b\}$
⇒ q¹'="abaaabbb"
→ Eliminate $X^?$

Results: on Sports Videos

videos

baseball

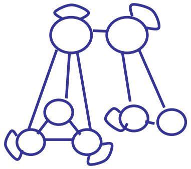
soccer

features

visual: dominant color ratio, camera translation estimates

audio: energies, zero-crossing rate, spectral rolloff

HHMM



patterns

HHMM top-level label sequence

$q^*_{1:T}$

vs. play/break?

Results: on Baseball Videos

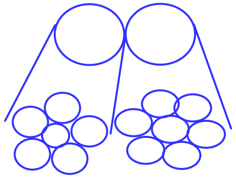
videos

baseball

BIC score

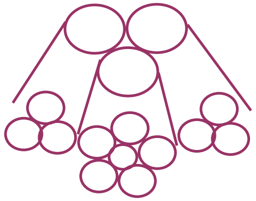
HHMM + feature selection

(1)



dominant color ratio
horizontal motion
(vertical motion)

(2)



audio volume
low-band energy

(3) ... ...

patterns

82.3%

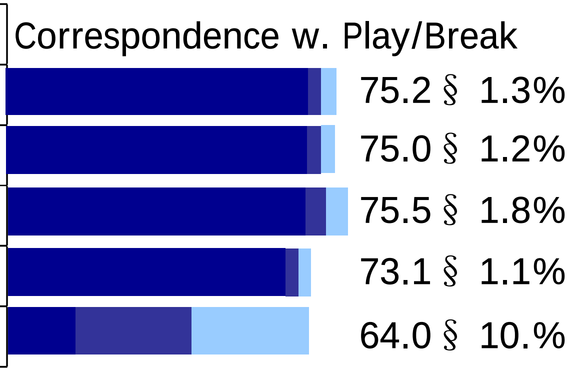
52.4%

correspondence
with play/break

Results: Comparison

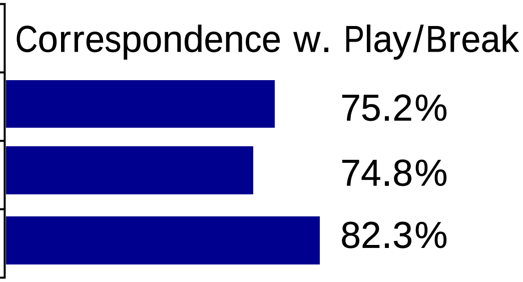
Fixed features {DCR, MI}, MPEG-7 Korean Soccer video

Model	Supervised?	Model Selection
HHMM	N	Y
HHMM	N	N
HMM	Y	N
LR-HHMM	N	N
K-Means	N	N



Automatic selection of both model and features

Test clip	Feature Set	# "events"
Korea	DCR,Mx	2∼4
Spain	DCR,Volume	2∼3
Baseball	DCR,Mx	2



* DCR='dominant-color-ratio', MI='motion-intensity', Mx='horizontal-camera-pan'

Outline

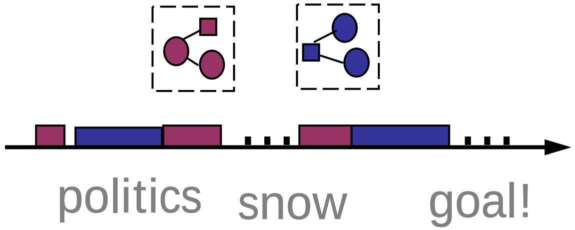
- The problem
- Unsupervised pattern discovery with HHMM
  → audio-visual token generation
- Finding meaningful patterns
  → token fusion
  - With text association
  - By multi-modal fusion
- Summary

± Versatile

± Multi-modal

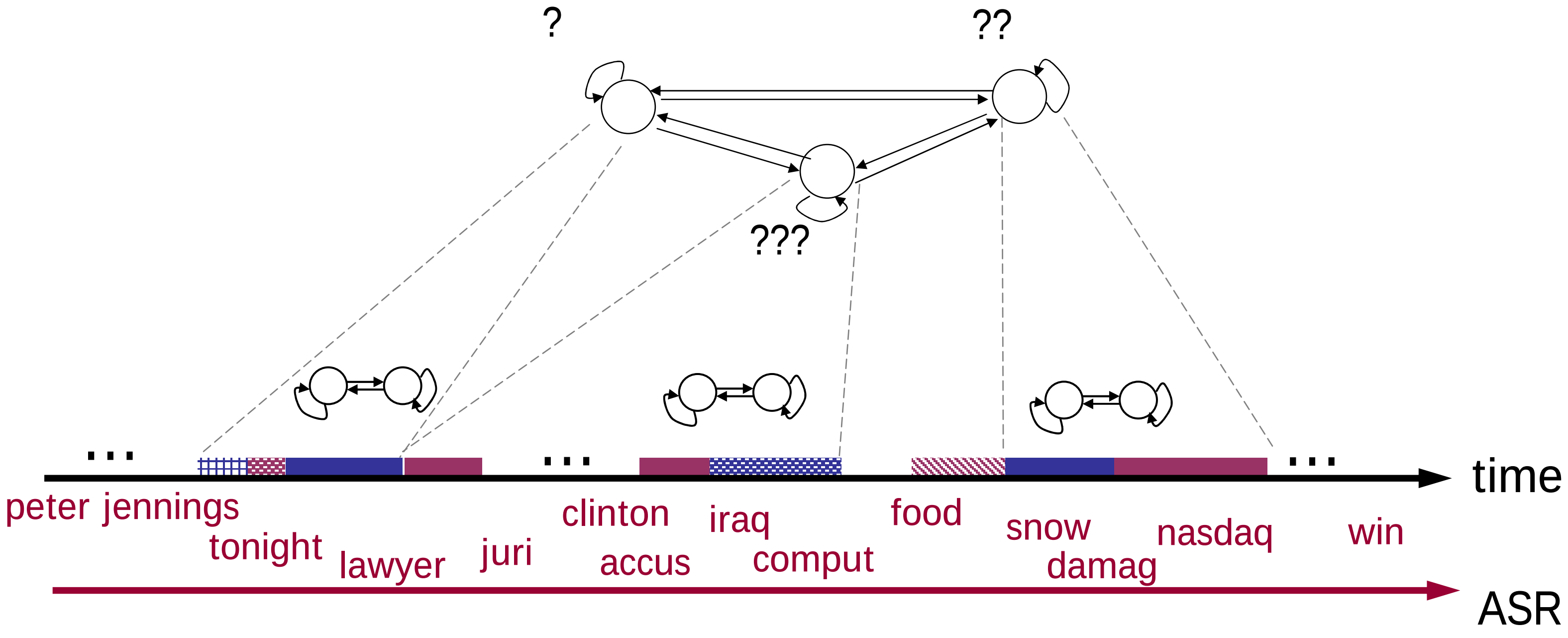
± Meaningful

? Knowledge-adaptive

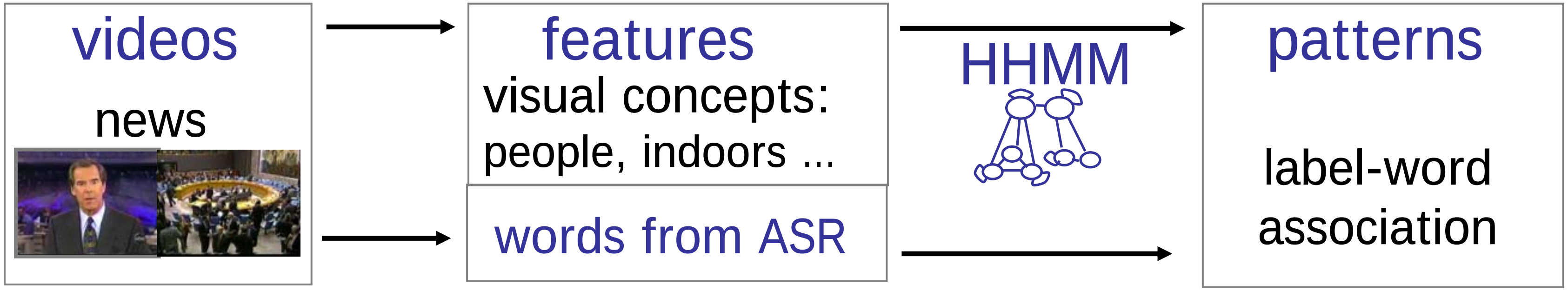


Towards Meaningful Patterns

- Manual association feasible only if meanings are *few* and *known.*
- Metadata come to the rescue.



Associating Patterns with Text



videos

news

features

visual concepts:

people, indoors ...

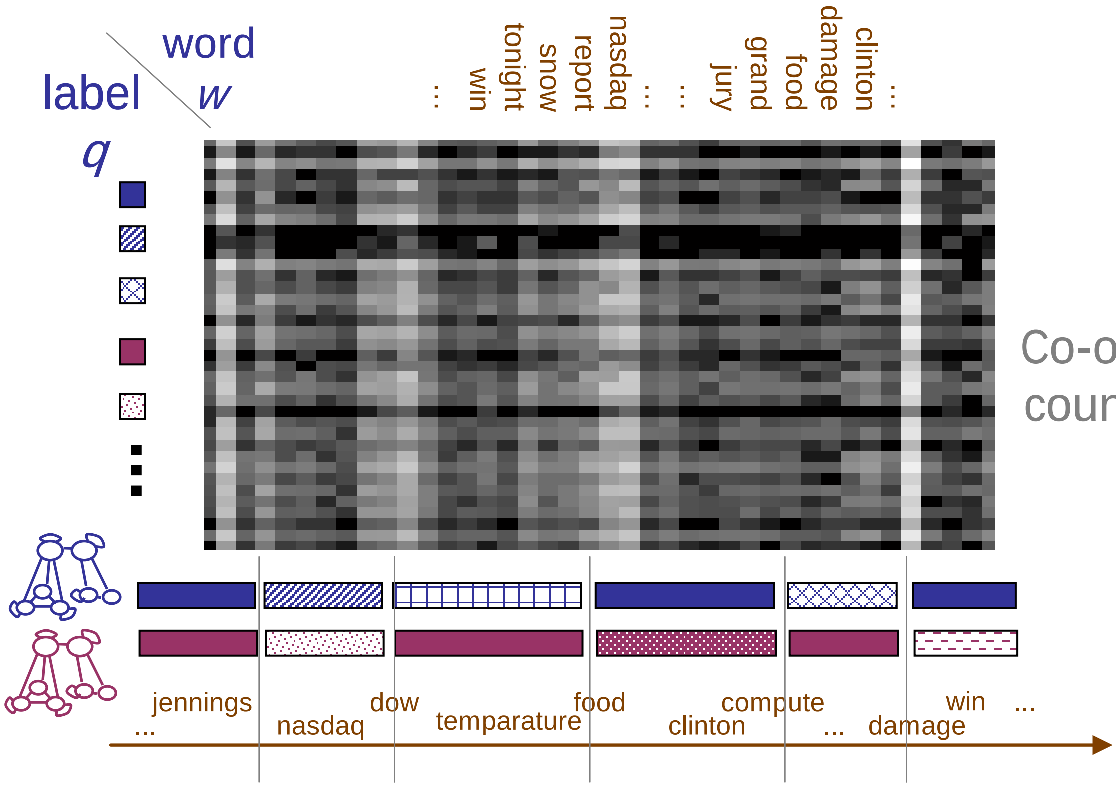
words from ASR

HHMM

patterns

label-word association

HHMM Labels and Words



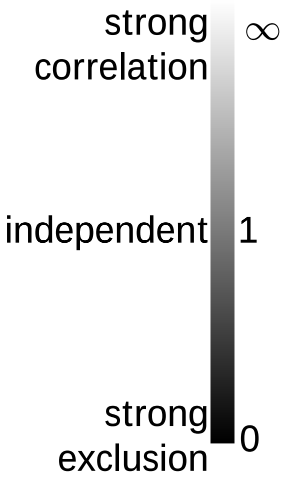
Co-occurrence counts C(q,w)

Language “X”

English words

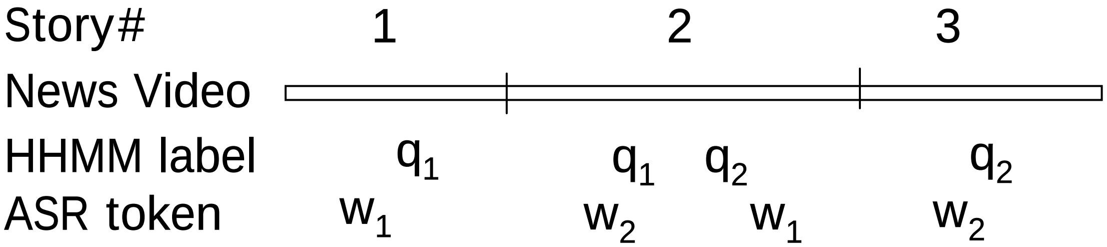
Likelihood ratio:

$$L^c(q,w) = \frac{C(q,w)}{C(q,\cdot)C(\cdot,w)}$$



“*Translation*” between HHMM labels and words
→ co-occurrence counts.

Refining the Co-occurrence Statistics



Story# 1 2 3

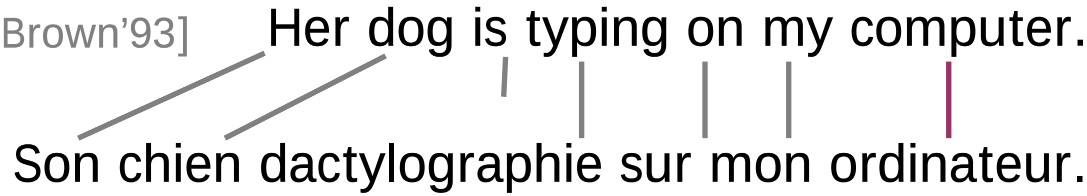
News Video

HHMM label q_1 q_1 q_2 q_2

ASR token w_1 w_2 w_1 w_2

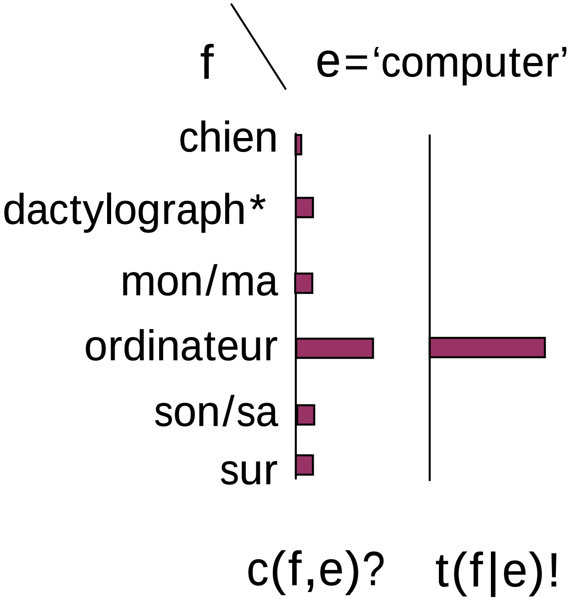
“true” cooc. $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ “smoothed” $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

MT [Brown’93]



Her dog is typing on my computer.

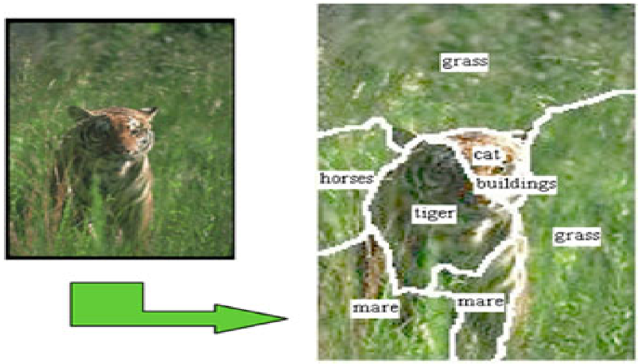
Son chien dactylographie sur mon ordinateur.



c(f,e)? t(f|e)!

[Dyugulu et. al. 2002]

image
$\approx \{b_1, \ldots, b_n\}$
$\approx \{w_1, \ldots, w_n\}$

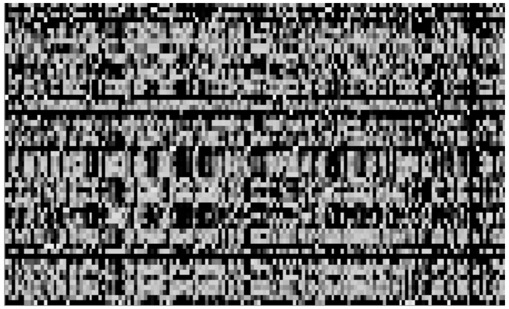


Translation between AV and Words

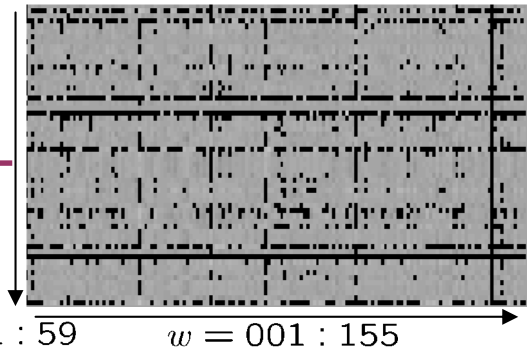
The problem:
Co-occurrence “un-smoothing”.
know: $C(q,w)$;
seek: $t(w|q)$, $t(q|w)$.

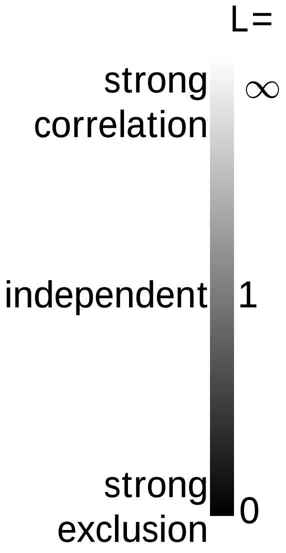
Solve with EM [Brown’93]

$L^t_q(q,w)$

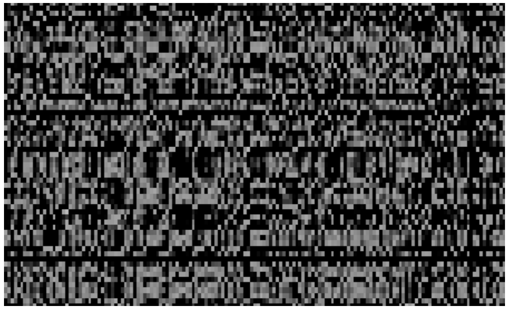


$$L^c(q,w) = \frac{C(q,w)}{C(q,\cdot)C(\cdot,w)}$$





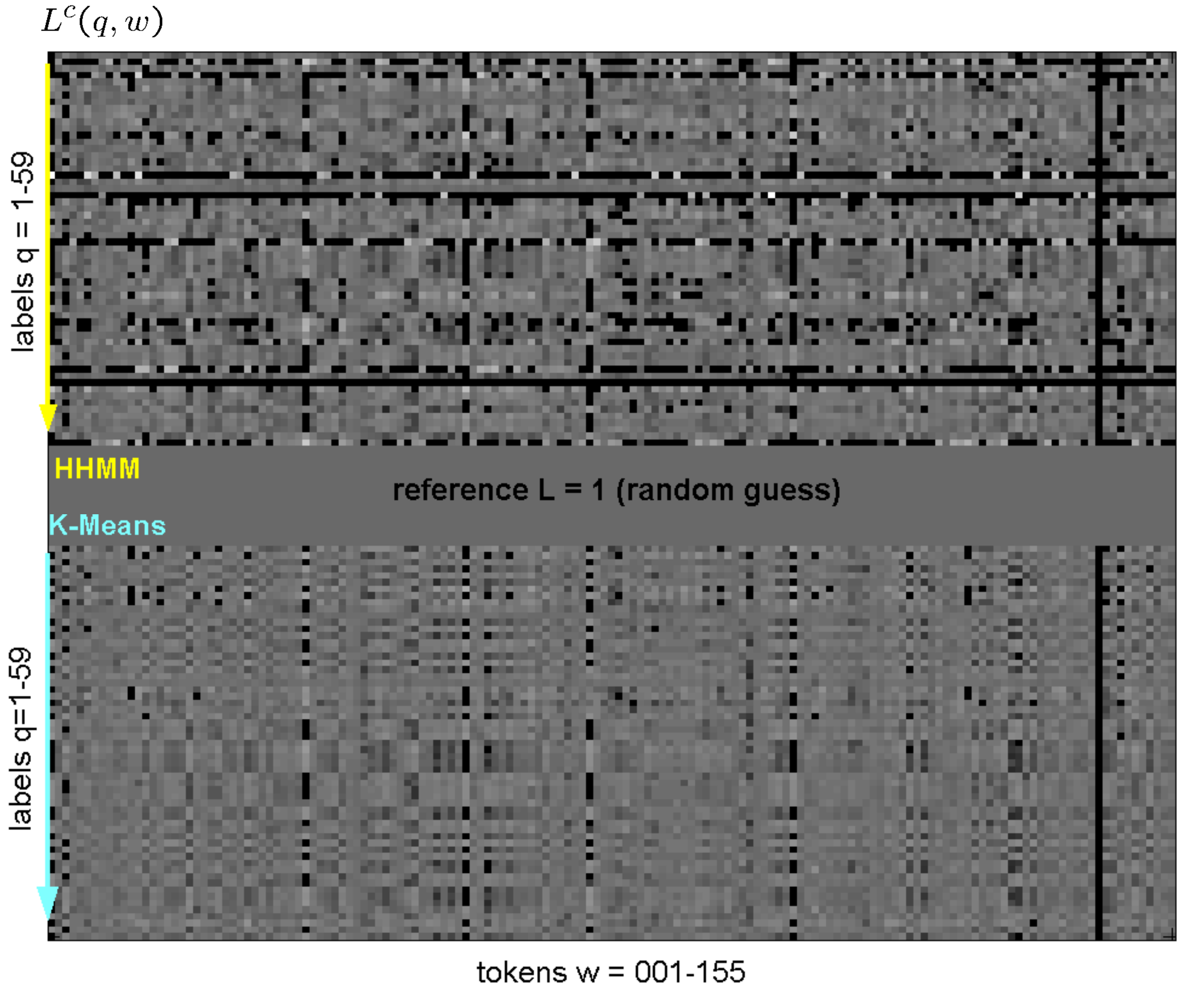
$L^t_w(q,w)$



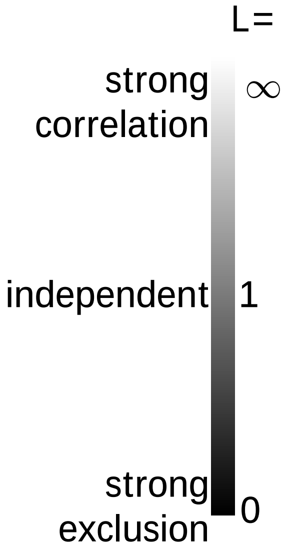
Experiments

- TRECVID2003 news
  - 44 30-min videos, ABC/CNN
  - 12 visual concepts for each shot [IBM-TREC'03]
  - ASR transcript
- HHMM on concept confidence scores
  - 10 models from hierarchical clustering in feature selection, size automatically determined
  - Co-occurrence with story boundaries

HHMM vs. Kmeans



HHMM:
more meaningful associations, less randomness.



Example Correspondences

[Xie et al. ICIP'04]

HHMM label	Visual Concept	Words	Topic groundtruth
(6,3)	people, non-studio-setting	storm, rain, forecast, flood, coast, el, nino, administer, water, cost, weather, protect, starr, north, plane, ...	El-nino'98
(9,1)	outdoors, news-subject-face, building	murder, lewinski, congress, allege, jury, judge, clinton, preside, politics, saddam, lawyer, accuse, independent, monica, charge, ...	Clinton-Jones (Recall 45%, Precision 15%) Iraqi-weapon (Recall 25%, Precision 15%)
(m, q): model # m state # q	Obtained with SVM classifiers [IBM'03]	Lexicon obtained by shallow parsing of keywords from speech recognition output.	

Summary

- Statistical models for pattern discovery
  - Unsupervised learning of temporal patterns with hierarchical HMM
  - Multi-modal fusion with statistical association and layered mixture models
- Open issues
  - Multi-modal fusion: when, why, how
    - Early fusion vs. late fusion
    - Single-modal tokens vs. multi-modal tokens
    - Bottom-up fusion vs. bi-directional propagation
  - Model selection and validation
  - Evaluation metric for multimedia patterns