Multimodal Fusion for Content-based Retrieval and Management of Multimedia Data

Stéphane Marchand-Maillet
Viper group,
University of Geneva
http://viper.unige.ch
marchand@cui.unige.ch

Outline

- Introduction
  - ⇒ Facets of Multimedia Information Retrieval and Management

I. Multimedia document management

II. Temporal document processing

III. Multimodal video indexing and retrieval

IV. Multimedia Information Exploration

Objectives

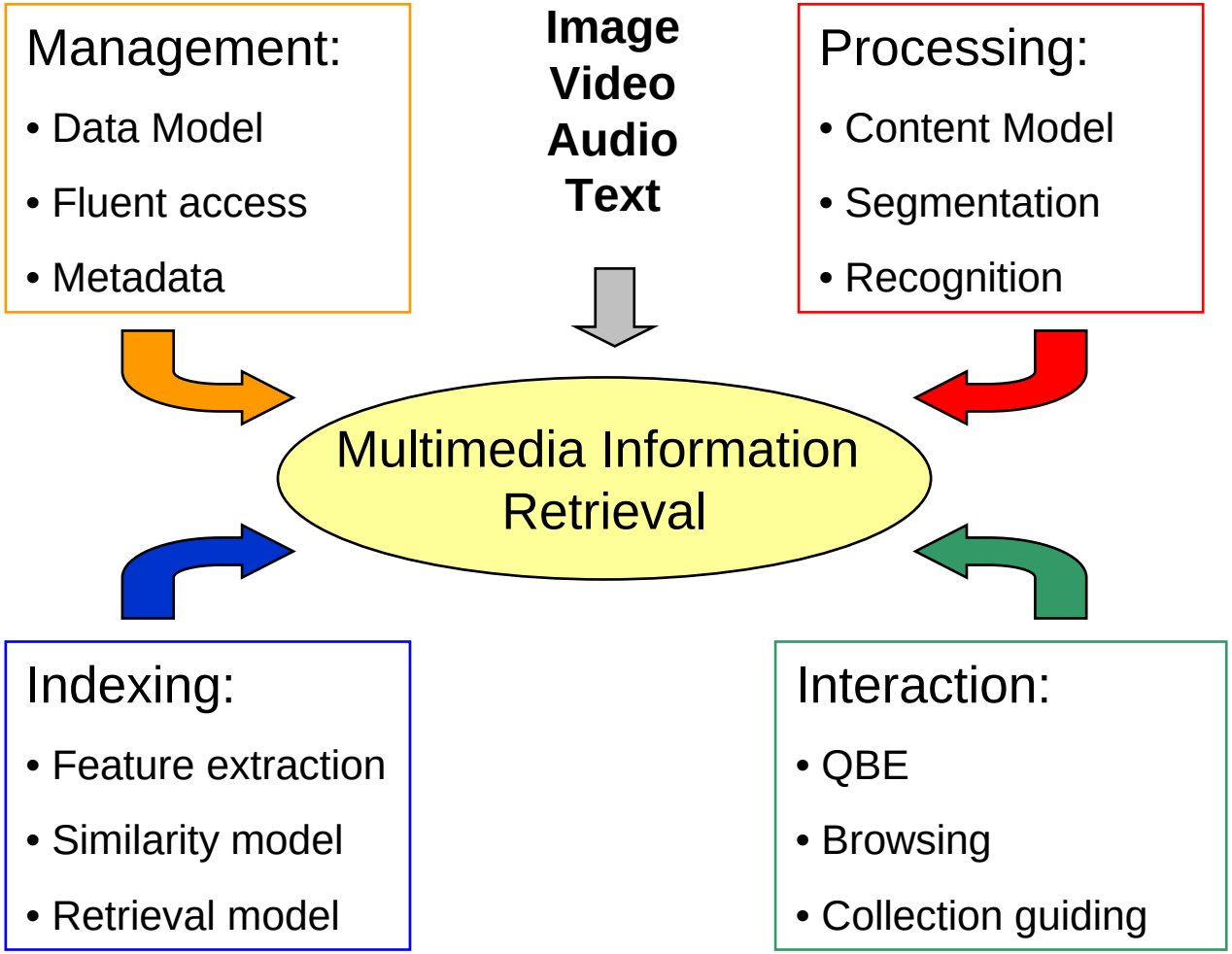
This course aims at emphasizing:

- The need for a formal management model
- The complexity of multimodal fusion
- The *complementarity* of several approaches
- The need for new interaction paradigms

Definitions

- Multimedia
  - Any document composed of heterogeneous media
- Medium
  - Support for a base stream: visual, audio, text,...
- Multimodal
  - Any information addressing many modes of perception
- Metadata
  - Data about data: data that one can hardly derive from the document itself without very high-level knowledge (author, date, place)
- Annotation
  - Data describing the document: object types, color, size, location,...

Multimedia Information Management

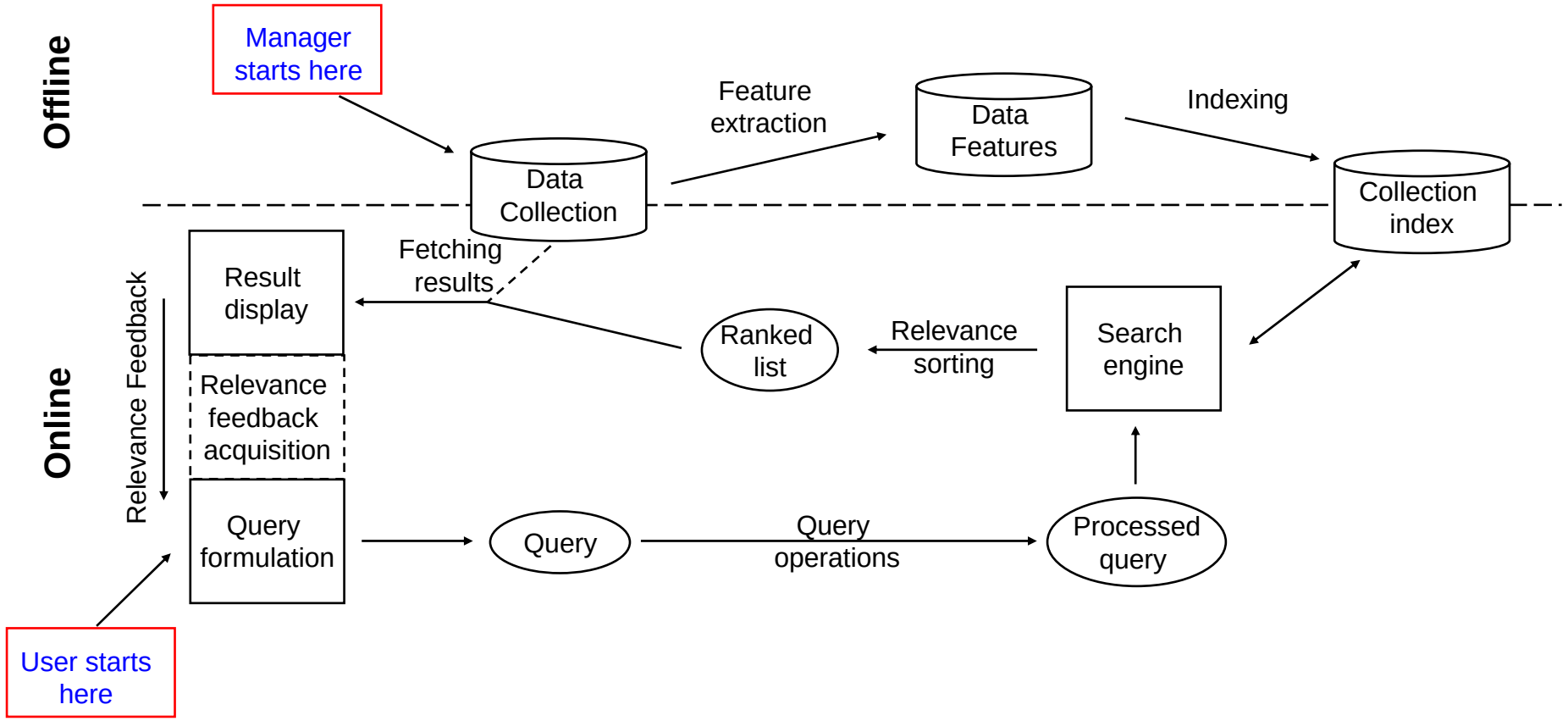


Part I: Management

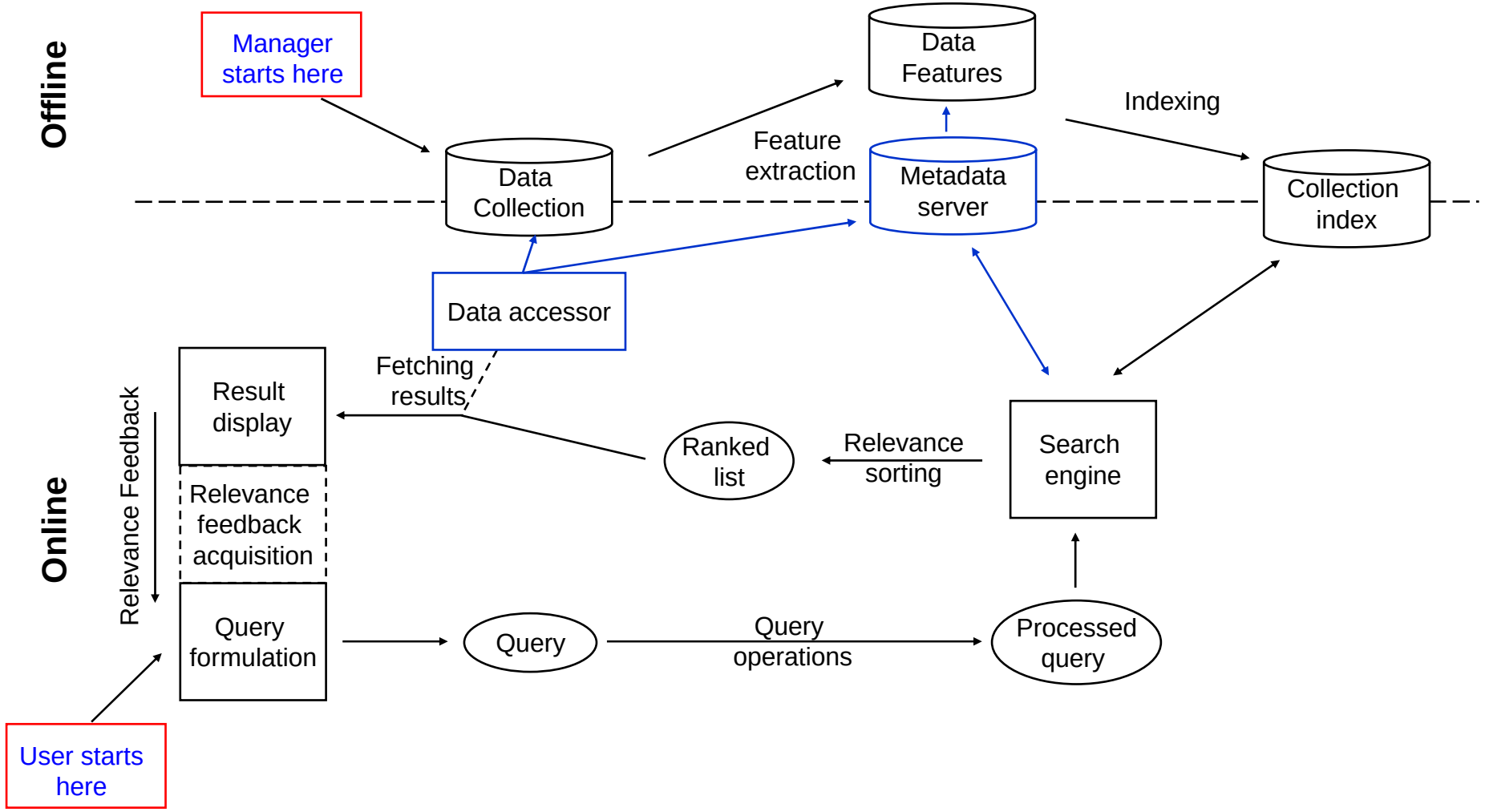
Management:

- Data Model
  - Required if the data is complex and multipart
- Fluent access
  - Required for anything else than static data
- Metadata
  - Model needed to manage the "Data about data" and preserve synchronization

Structure of a classical IR system…

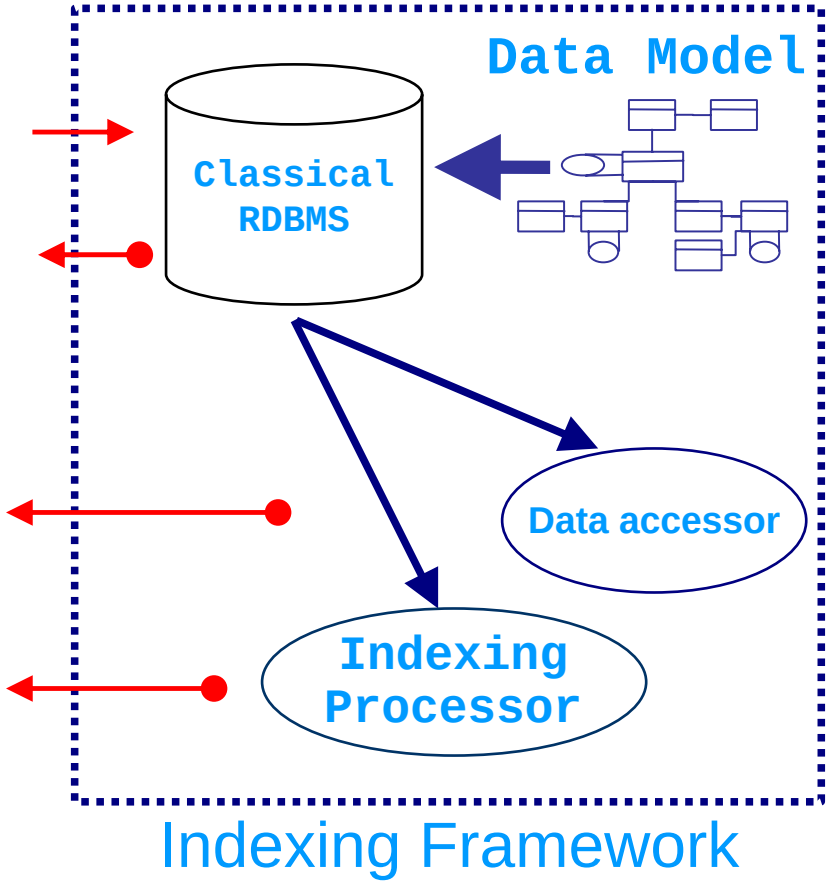


... adapted to temporal information



Indexing framework

- Generic Data Model
- Relational *DBMS*
- Raw Data Access
- External Indexing & Retrieval Processor



Data accessor: image

Image queries

- Global image
- Image region
- Image channel

An image has no temporal dimension

An image may be multimodal

- If annotation is provided (text+image)
- If segmentation is provided (background, foreground)

Synchronization: region ←→ annotation

Corpus: Corel collection, TRECVid keyframes

Data accessor: text

Text queries

- Keyword
- Named entity
- Structure (paragraph,...)

A text has a *weak* temporal dimension

- Its structure can be captured instantaneously
- Its content is captured temporally

A text is generally mono-modal

- Even if queries may be made on its structure (different mode)

Synchronization: no real need

Corpus: TREC

Data accessor: audio

Audio queries

- Depends on base type (speech, music, generic)

⇒ Similarity, humming, keyword spotting, speaker,…

The audio stream bears a temporal dimension

- Its content is captured temporally

An audio stream is generally mono-modal

- Difficult to perceive several audio streams simultaneously
- Text transcription may be considered as an attached mode

Synchronization: Audio + transcripts

Corpus: Linguistic data (LDC), TRECVid audio

Data accessor: video

Video queries

- May be based on its constituting streams

⇒ See image (+motion), audio, text,...

The video stream bears a temporal dimension

- Its content is captured temporally

A video stream is multimodal

- Classical: visual + audio
- Augmented: visual + audio + text (transcripts)
- ... any combination of streams

Synchronization: Find all mode values at a given temporal point

Video data (examples)

- Movie (A/V + transcripts + summary)
  - ⇒ Queries: character occurrences, scene, summary
- News broadcast (A/V+transcripts+stories+documents)
  - ⇒ Queries: topic (story), character, event, summary
- Sport broadcast (A/V + player list + score sheet)
  - ⇒ Queries: events (score), player / team / object (car)
- Video surveillance (multi-visual, audio)
  - ⇒ Queries: character occurrences, events, summary
- Meeting record (multi-A/V, transcripts, documents)
  - ⇒ Queries: topic, speaker, event
- Home video (A/V, attached photos,...) ...

Corpus: TRECVid, AMI, CAVIAR

A note on Web pages

A web page is a composite document:

- Text related to inserts (images, audio, video,...)

⇒ Multimodal document

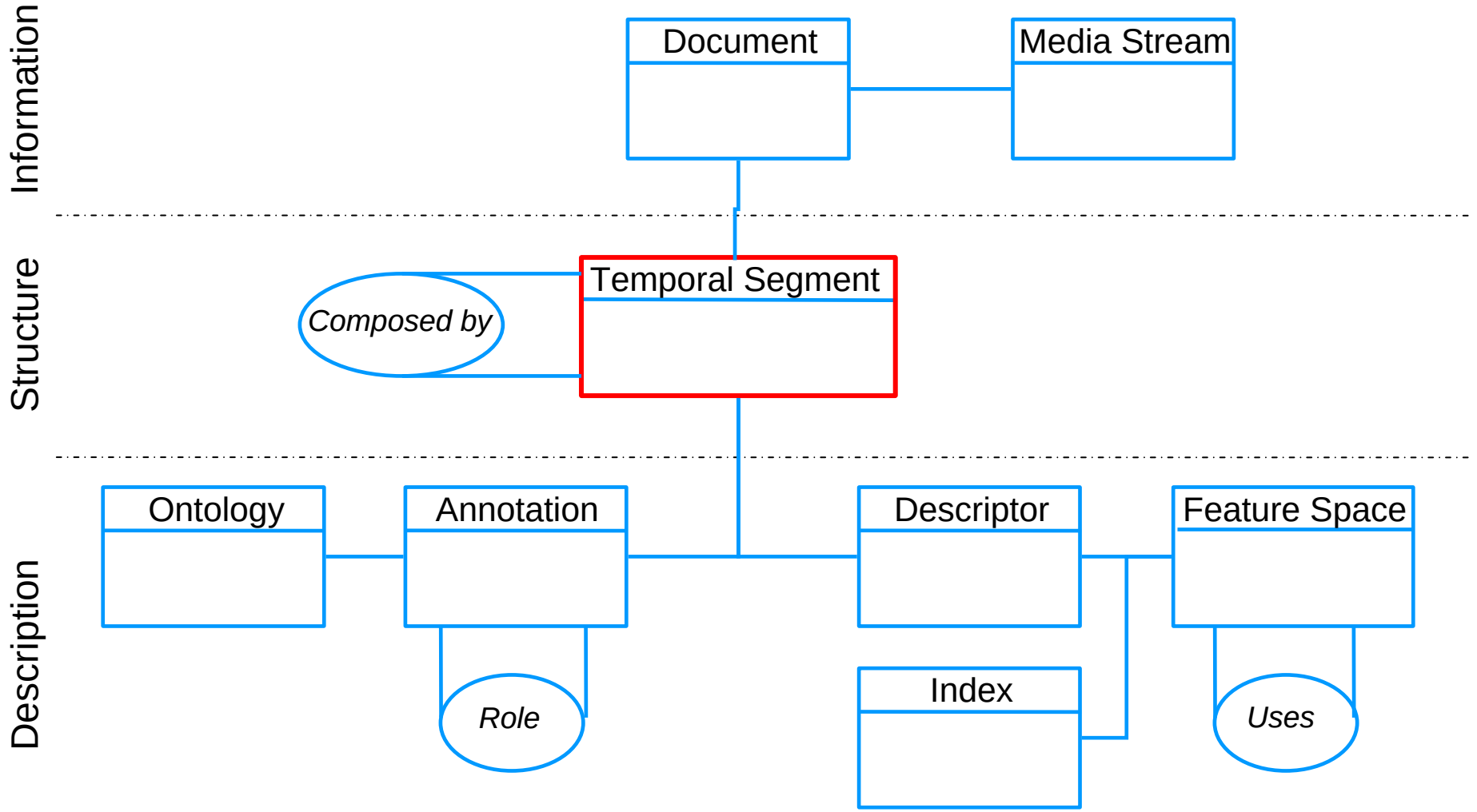
Web page retrieval is a combination of:

- Document retrieval (principally text)
- Hyperlink structure exploration (PageRank, HITS)

⇒ This context does not really fit within the current discussion

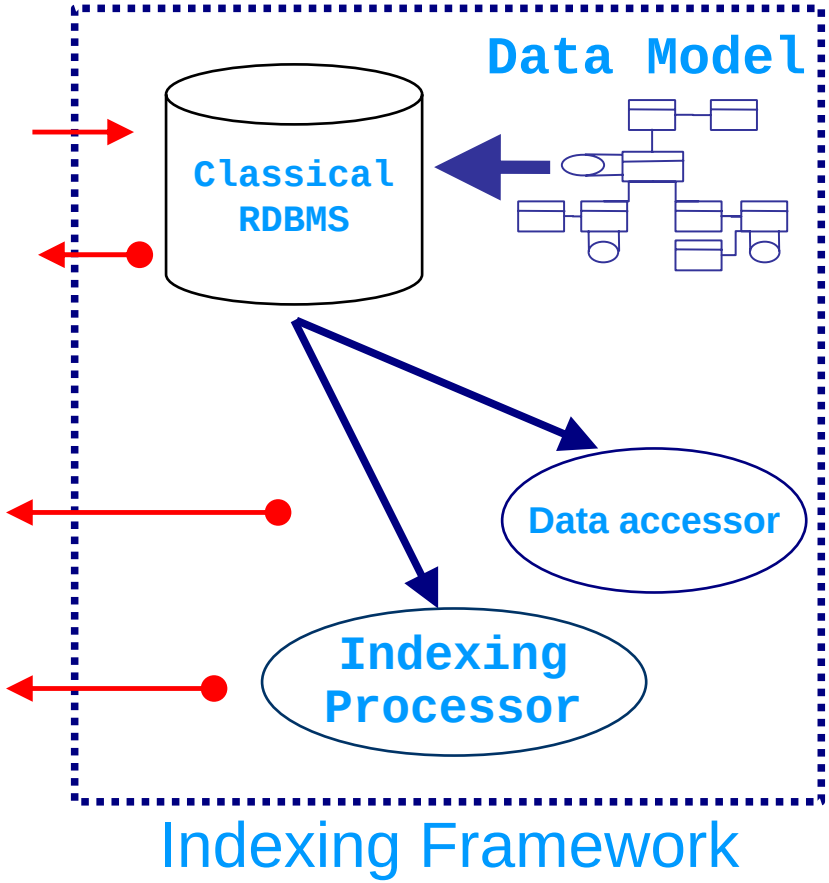
⇒ Other lectures (IR, text analysis,...)

Temporal Data Model

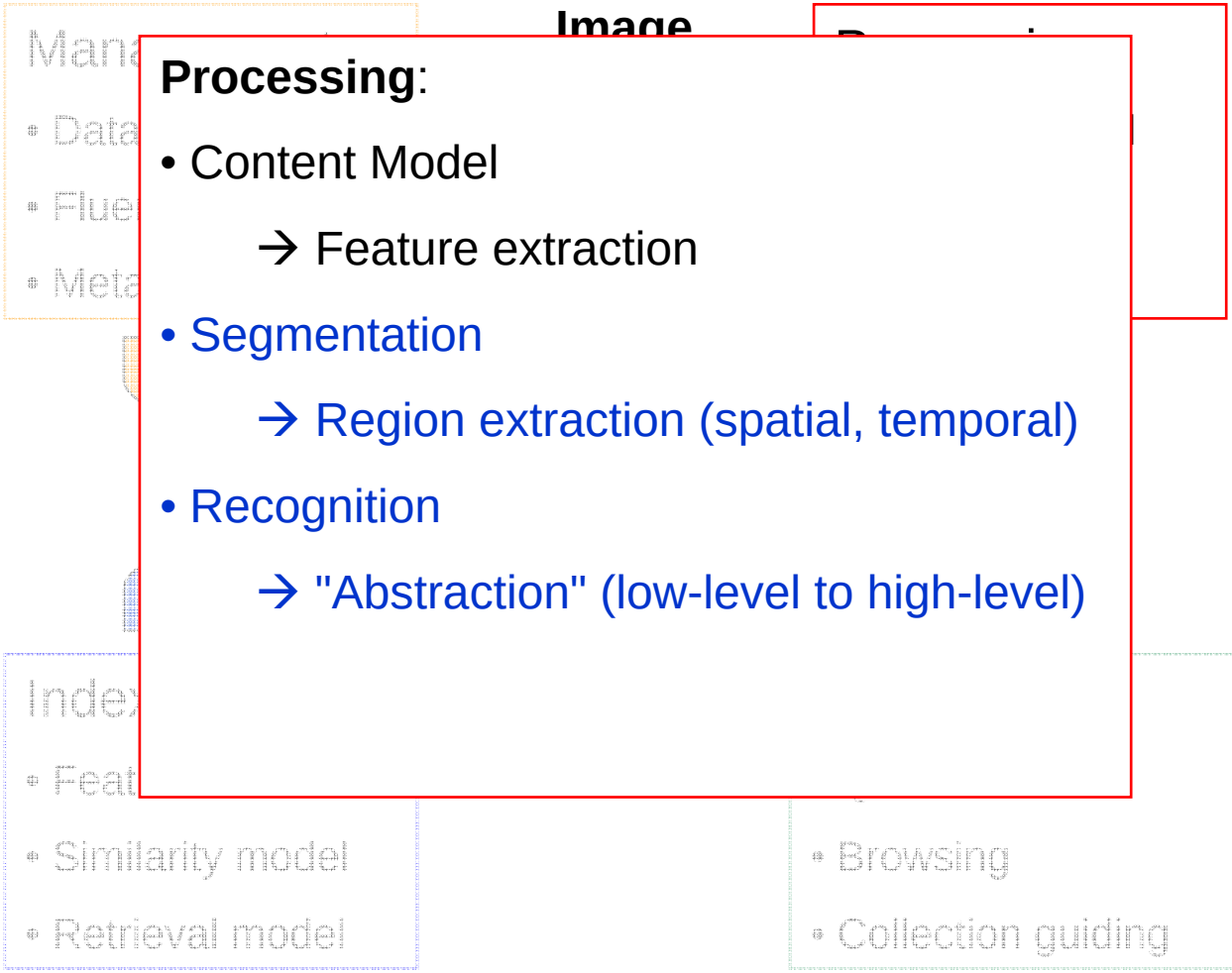


Indexing framework

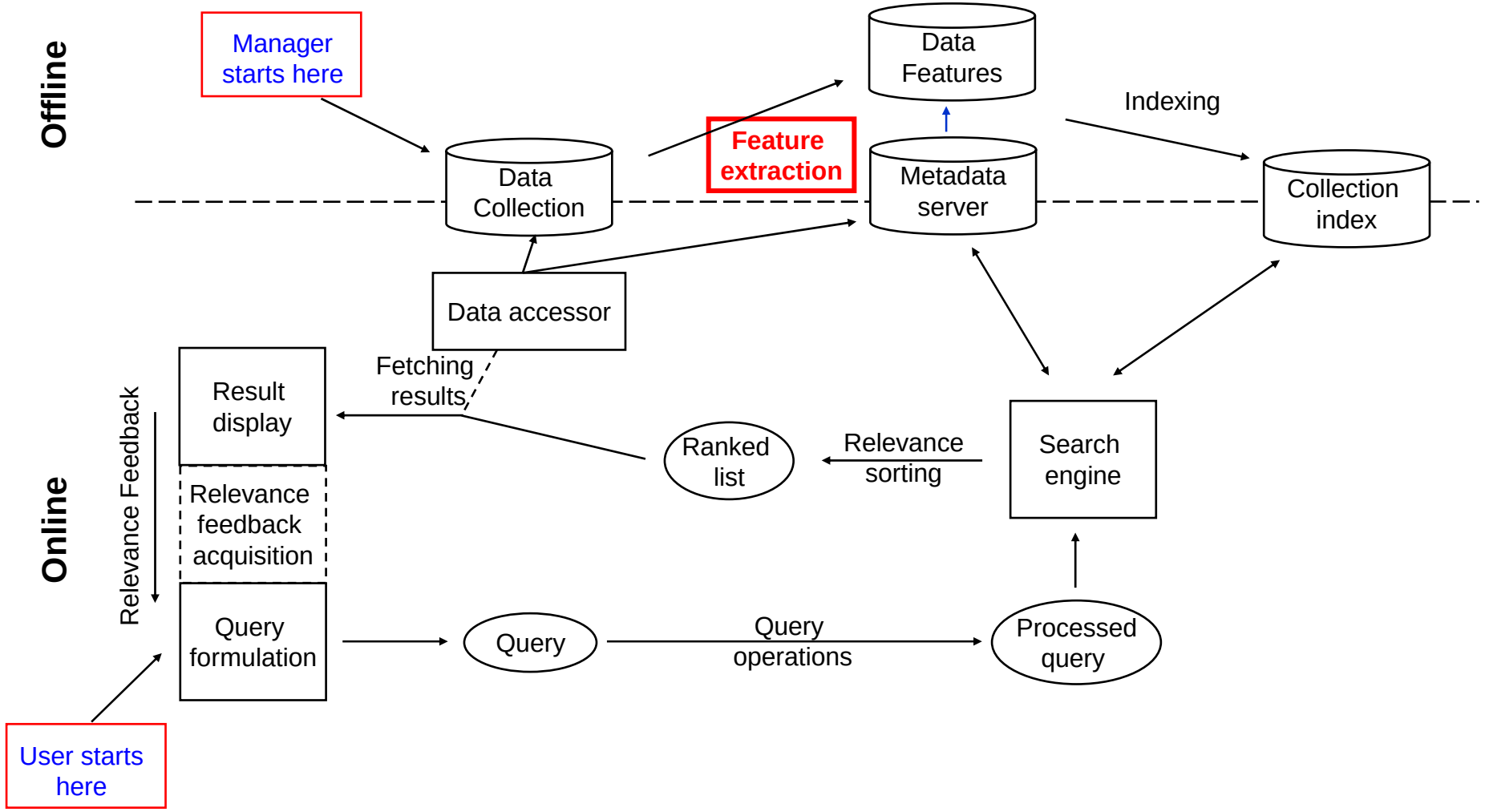
- Generic Data Model
- Relational *DBMS*
- Raw Data Access
- External Indexing & Retrieval Processor



Part II: Processing



Temporal IR


Offline
Manager starts here
Data Features
Indexing
Feature extraction
Data Collection
Metadata server
Collection index
Data accessor
Fetching results
Result display
Online
Relevance Feedback
Relevance feedback acquisition
Ranked list
Relevance sorting
Search engine
Query formulation
Query
Query operations
Processed query
User starts here

Temporal data processing

Temporal data cannot be considered globally

⇒ Need for a temporal structure of logical units before indexing

⇒ Each logical unit will be considered as a whole and represented by a static sample (icon, keyframe,...)

Text:

- *Topical segmentation*

Audio:

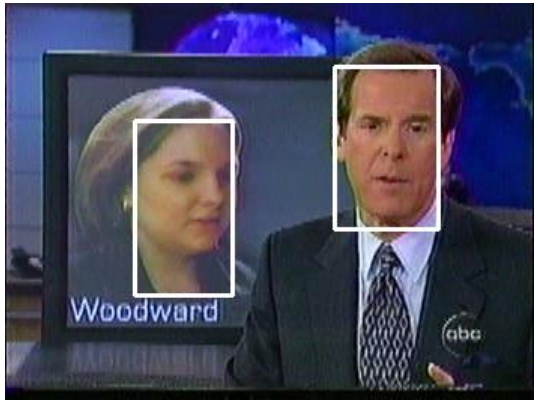
- Partition according to audio classes (speech, music,...)
- Speech: *eg* speaker-based segmentation

Video:

- Movie: scene
- News: Story
- Surveillance: ?? Event ??

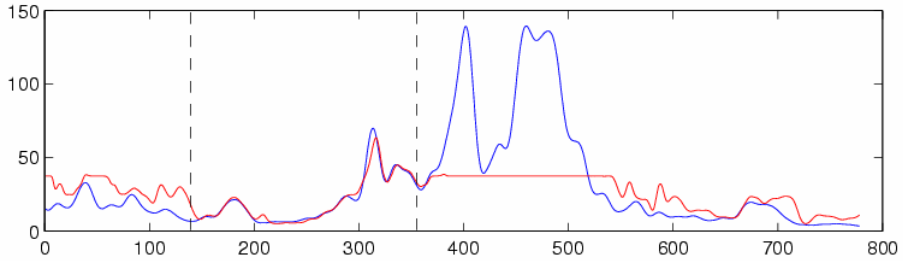
Visual content analysis

- Color and texture descriptors
- Salient object detection
- Background mosaicking
- Face detection



Temporal models and activity analysis

- Optical flow characterization/camera motion estimation
- Nonlinear models of time-dependent parameters
- Event detection and tracking



Temporal video segmentation

Base video structure is required

- Shot:
  Group of video frames with visual consistency
  ⇒ Model for a continuous scene recording

- Story:
  Group of video frames with topical consistency
  ⇒ Model for a scene, report
  ⇒ From multimodal information

Here, we look at **multimodal story segmentation**

Temporal (high-level) segmentation

Classical assumption for story segmentation:

Every story is composed of several (visual) shots

1. Generate shot boundaries with high recall
   - ⇒ Do not miss any

2. For each shot boundary, decide whether it is a story boundary or not
   - ⇒ Look at all available streams (local multimodality)
   - ⇒ Look at context (temporal consistency)

Reminder: shot boundary detection

Frame difference:

- Pixel-based (I_i and I_j are channels of frames i and j , respectively)

$$D(i,j) = \frac{1}{N} \sum_{(x,y) \in I} |I_i(x,y) - I_j(x,y)|$$

- Histogram-based (H_i and H_j are histograms of I_i and I_j , respectively)
  - L1 norm:

$$D(i,j) = \sum_{k=1}^{N} |H_i(k) - H_j(k)|$$

  - Kullback-Leibler:

$$D(i,j) = \sum_{k=1}^{N} H_i(k) \log \left(\frac{H_i(k)}{H_j(k)} \right)$$

  - Jeffrey divergence:

$$D(i,j) = \sum_{k=1}^{N} (H_i(k) \log \left(\frac{H_i(k)}{m(k)} \right) + H_j(k) \log \left(\frac{H_j(k)}{m(k)} \right))$$

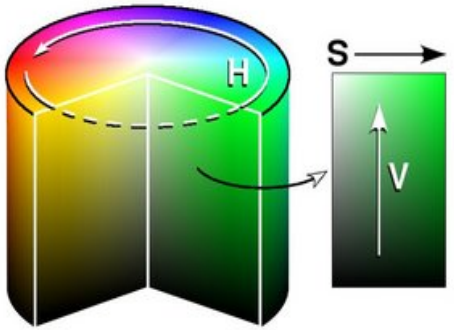
with:

$$m(k) = \frac{H_i(k) + H_j(k)}{2}.$$

Low-level visual feature representation

Global Colour Histogram data in HSV space:

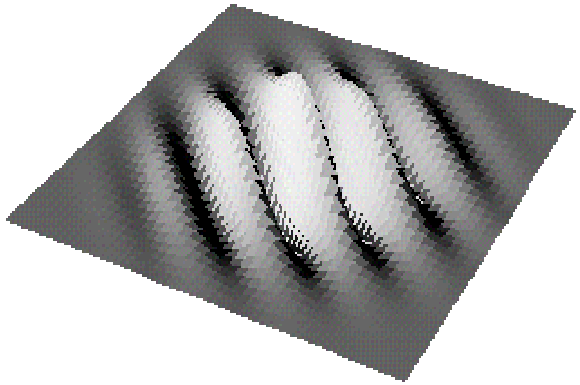
- 18 hue, 3 saturation, 3 value levels, together with additional 4 grey levels
- 18*3*3+4 = 166 colour descriptors



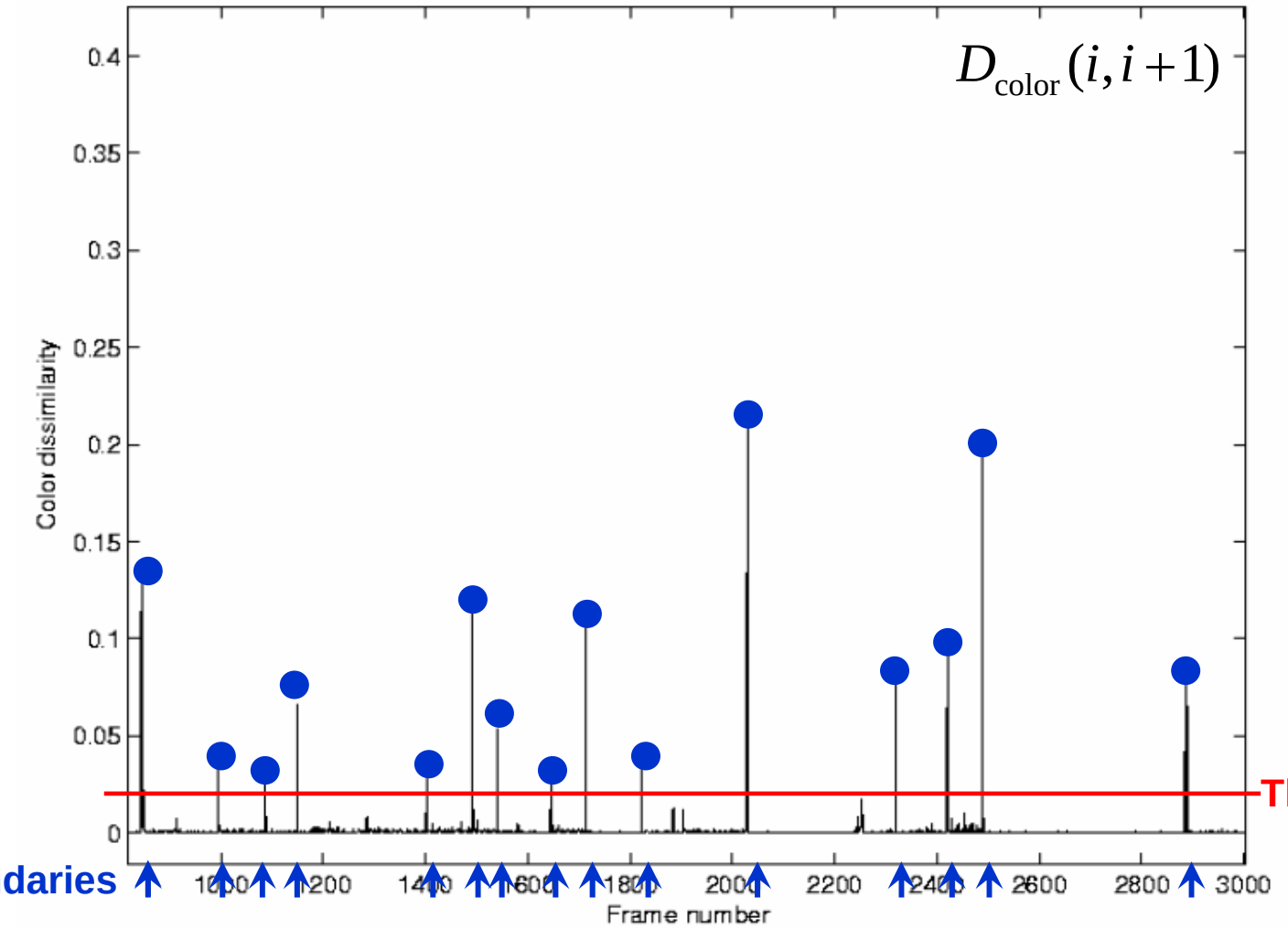
Global Texture data from Gabor filter banks:

$$g(x,y) = \left(\frac{1}{2\pi\sigma_x\sigma_y}\right)\exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)+2\pi jWx\right)$$

- 3 scales, 4 orientations, 10 bands
- 3*4*10 = 120 texture descriptors



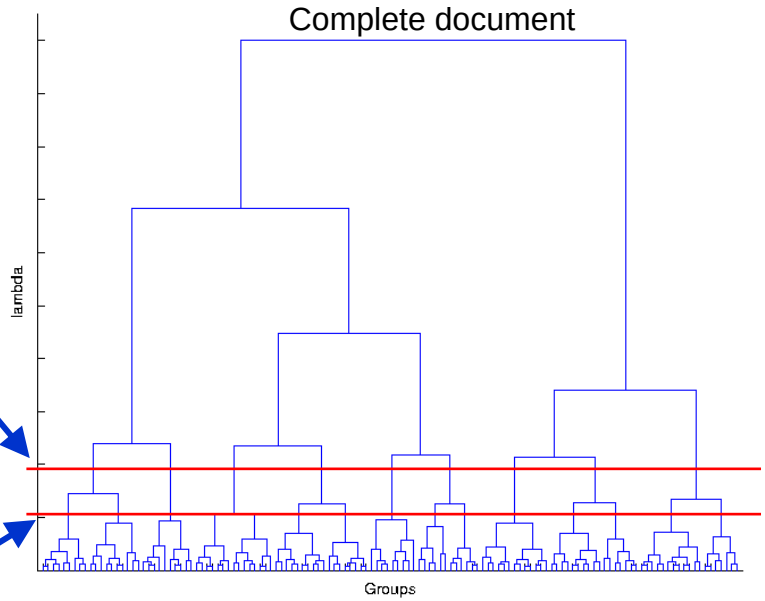
Frame dissimilarity profile



Shot detection: Example

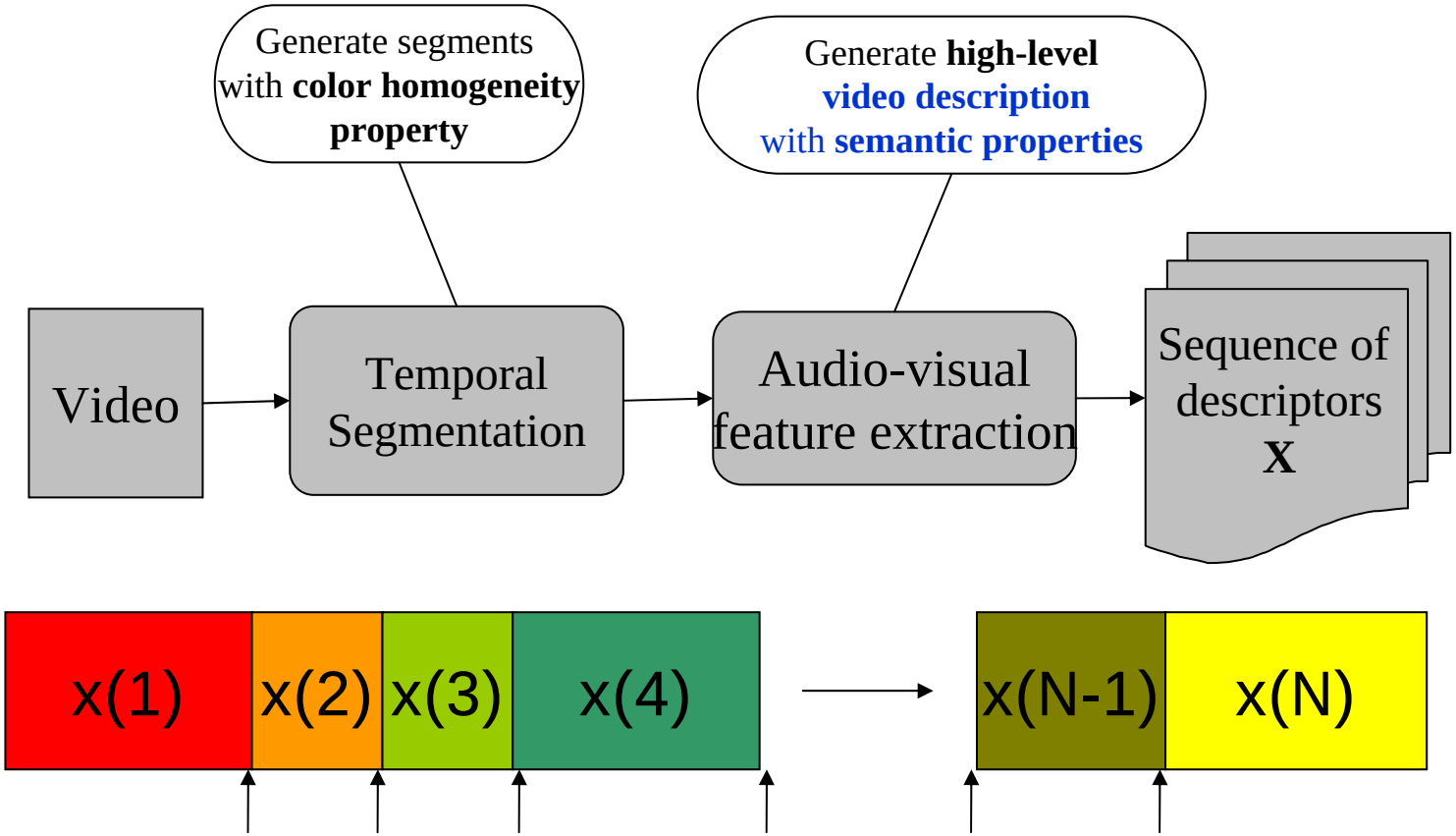
Storyboard representation

- The keyframe is the middle frame



Shot merging from varying the threshold

From shots to stories



Set of candidate points **C** to decide if there is a story boundary or not

From low-level to semantic description

Need for domain knowledge to understand and model the content

The case of news video content:

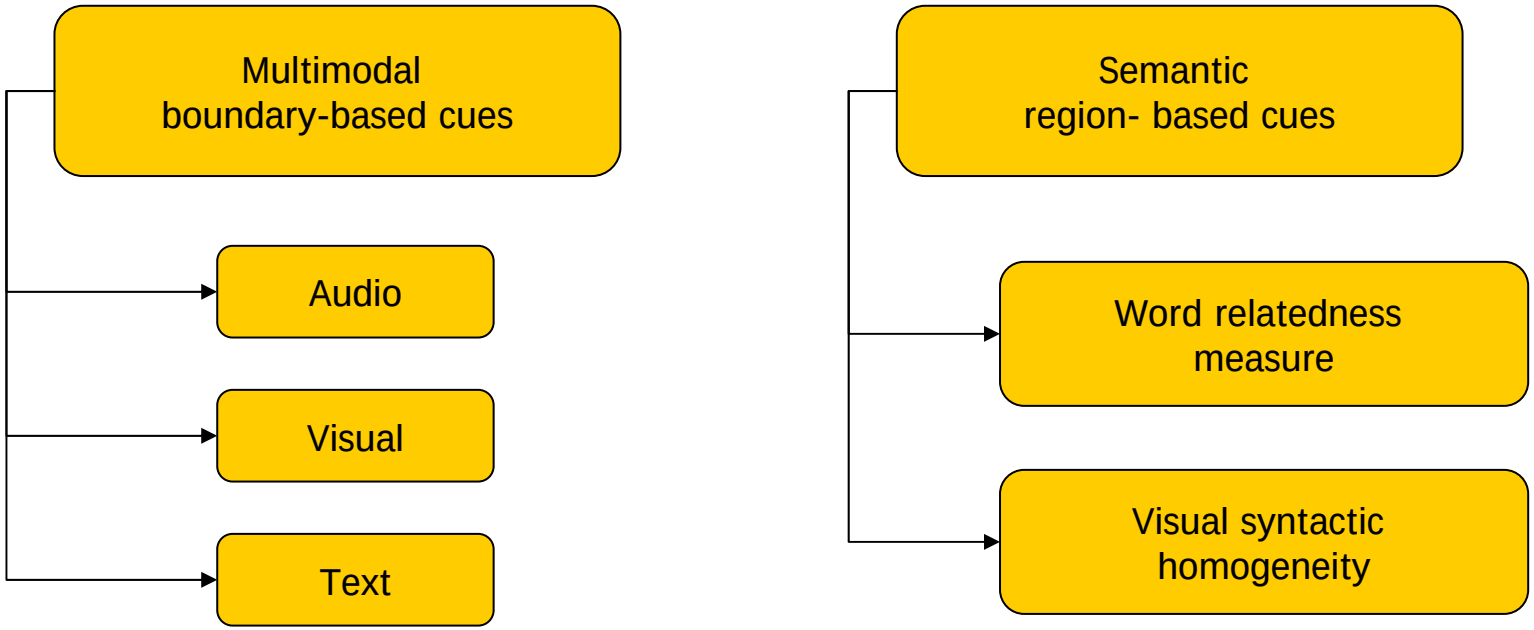
- Video classification
  - news subject monologue, studio-settings, outdoors, man-made scene, cartoon, weather news, text overlay, sport scene, non news
- Audio classification
  - Speech, music, noise, speech+music, speech+noise, other sound
- Text classification
  - *N* classes inferred using unsupervised clustering

⇒ Importance to design an expressive feature pool

Video features

Over temporal modalities:

- Segmentation (boundary) cues
- Categorization (region) cues



Improving classification by using multimodality

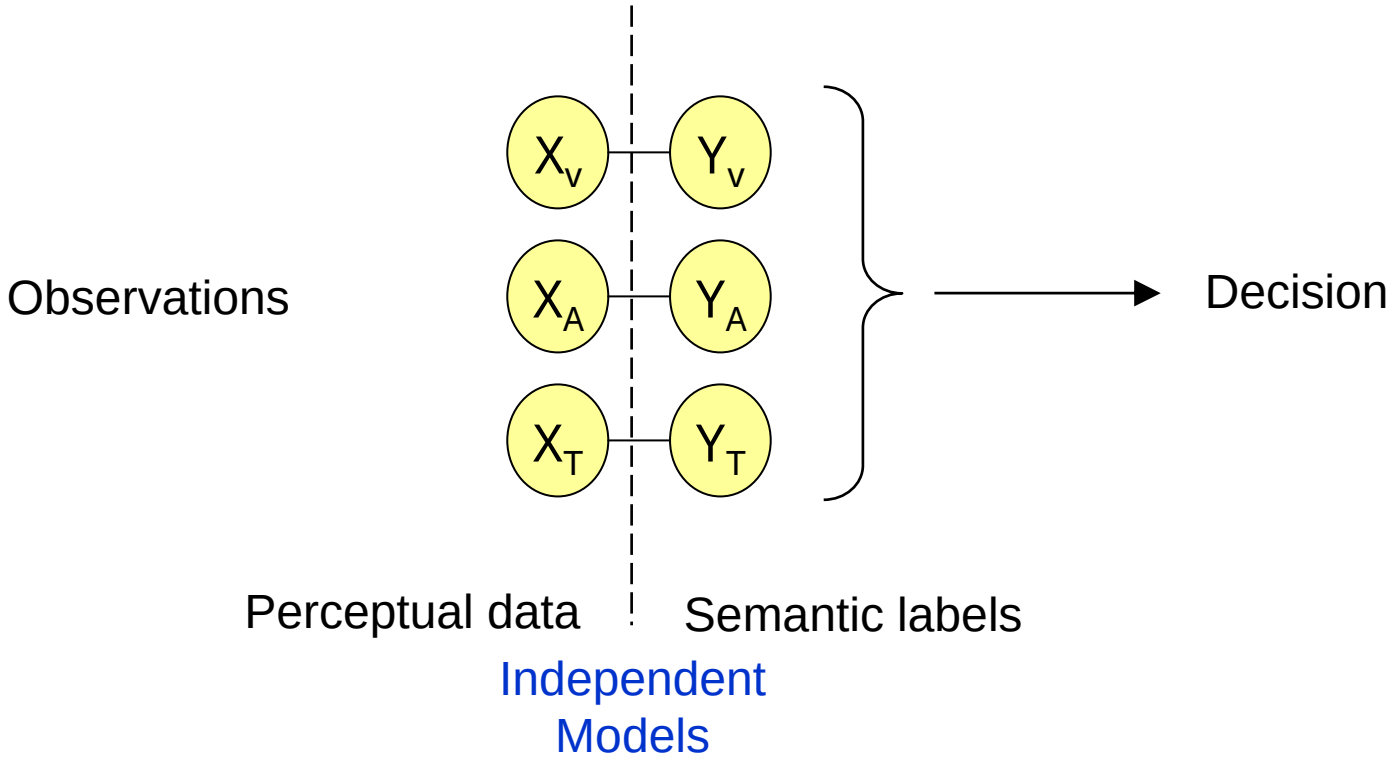
Need to use the features altogether :

- at multiple scales around the observations
  - what happened **locally** and at a **global level**
- of different modalities
  - **Visual** : color statistics, motion vector statistics, multiscale structural features (Haar filters)
  - **Audio** : volume, band width, subbands energy, cepstral flux, 4Hz frequency energy, ...
  - **Text** : word *N*-grams number of appearance

Improving classification by using multimodality

Mono-modal approach:

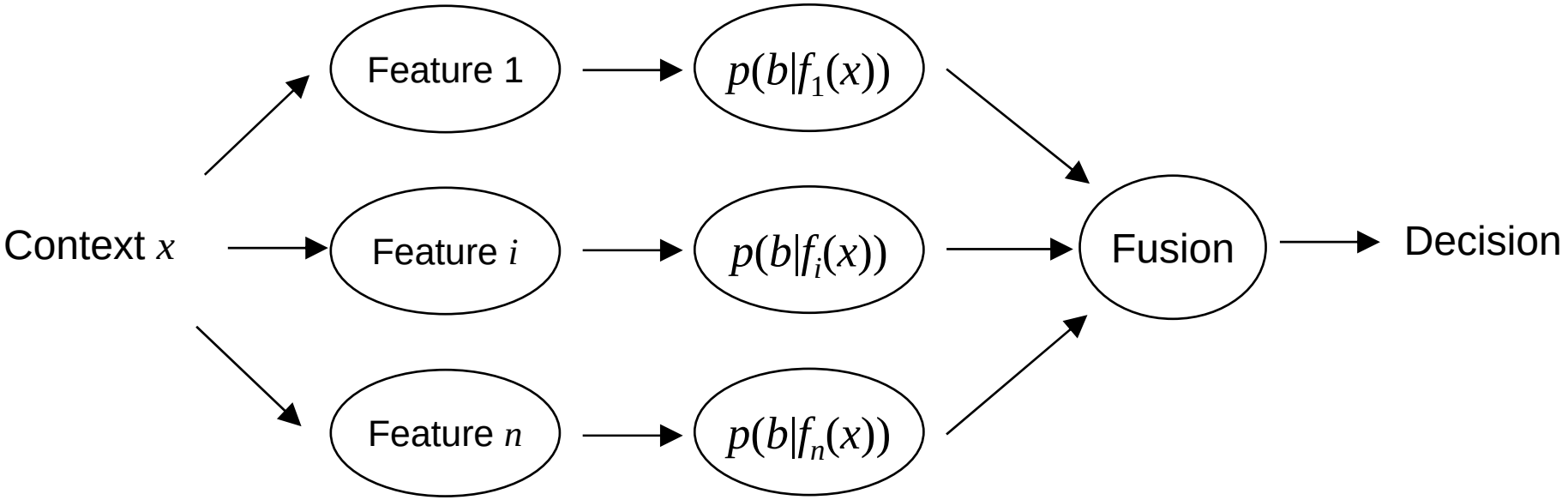
- Each stream is processed separately and independently



Max entropy classifier

The probability that a candidate point corresponds to a story boundary given a context x and a set of features $\{f_i\}$ as :

$$p_\lambda(b|x) = \frac{1}{Z} exp(\sum_i \lambda_i p(b|f_i(x)))$$



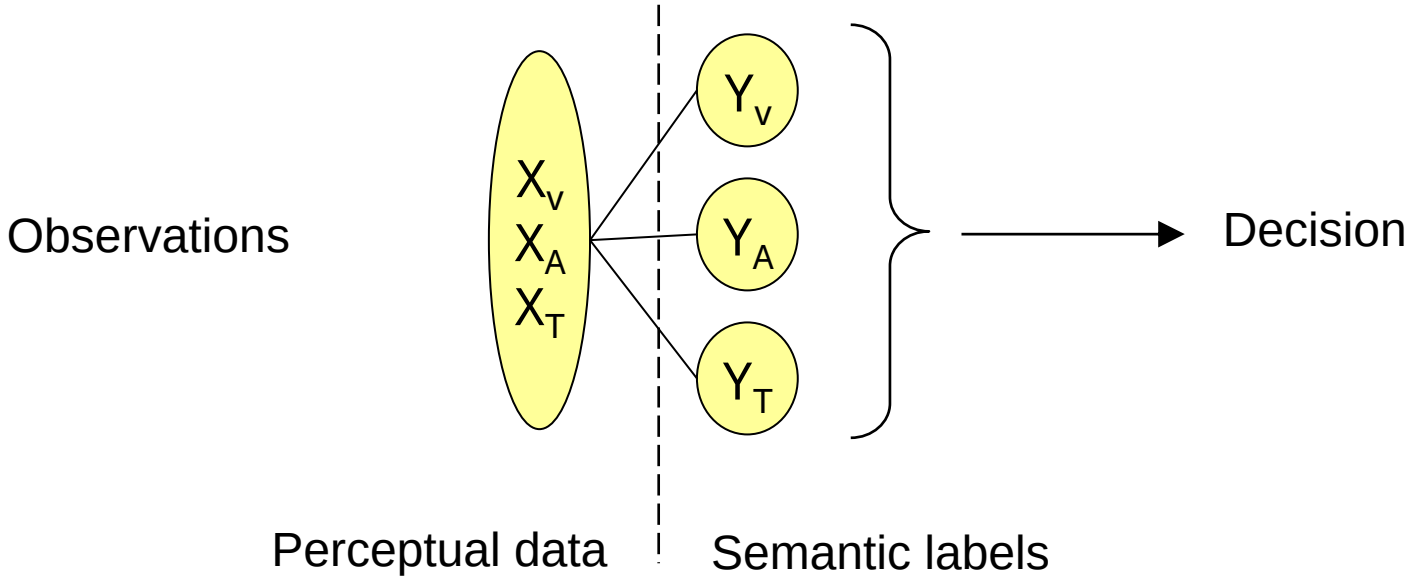
The decision is taken on the basis of local models $p(b \mid .)$

Improving classification by modeling context

Multi-modal approach:

- Streams are taken jointly
- Semantic labels are still supposed to be independent

⇒ Decision taken over potentially incompatible labels



Joint models

Example

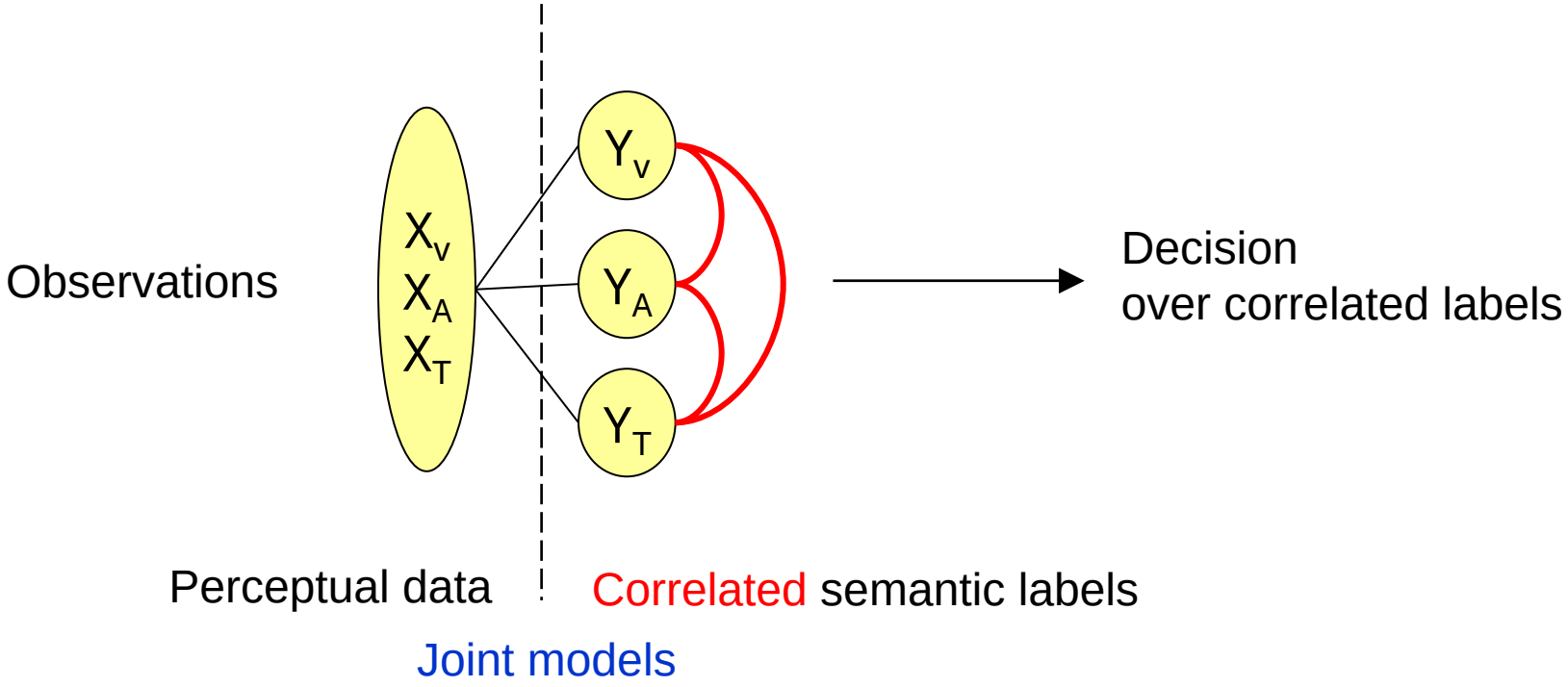
- Video classification
  - news subject monologue, studio-settings, outdoors, man-made scene, cartoon, weather news, text overlay, sport scene, non news
- Audio classification
  - Speech, music, noise, speech+music, speech+noise, other sound
- Text classification
  - Topic 1, Topic 2, Topic 3, …, Default class

⇒ Decision for that context?

Improving classification by modeling context

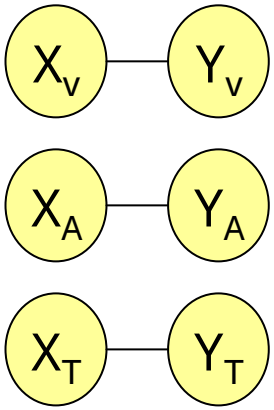
Contextual and multi-modal approach:

- Streams are taken jointly
- Semantic labels are supposed to be correlated

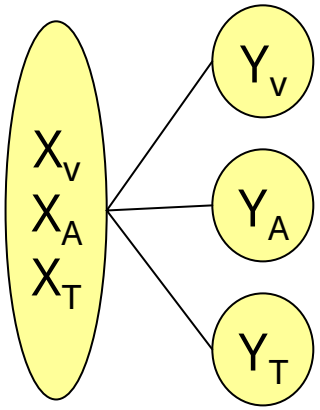


Improving classification by modeling context

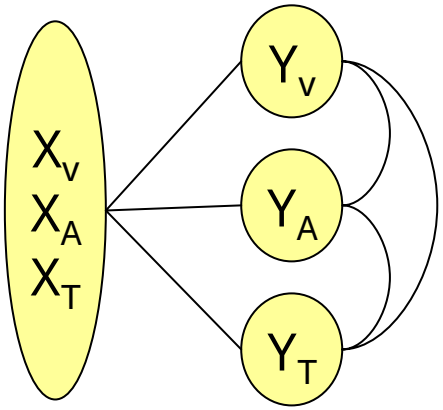
Mono-modal approach



Multimodal classification



Contextual and Multimodal classification



- Context can improve classification results by using :
  - correlations in the label's time sequence
  - correlation between labels over different modalities

Challenges

- Feature selection
  - From the design of a large feature pool, select those that are the most useful (informative, discriminative,...)
  - ⇒ Cannot just try them all (curse of dimensionality, dimension incompatibility,...)
  - ⇒ Principle of boosting
- Label compatibility
  - How to model in a tractable way compatibility between labels
  - ⇒ Face combinatorial exploration
  - ⇒ Conditional Random Fields (CRFs)

Boosting

Principle: Analyse training errors to combine additively a set of weak learners f_m into a strong one F

$\Rightarrow$ Most used algorithm: AdaBoost

- AdaBoost fits an additive model : $F(x) = \sum_{m=1}^{M} c_m f_m(x)$
- Adaboost minimizes the criterion : $J(F) = E(e^{-yF(x)})$
  - Exponential criteria are monotone and smooth
- Iterative minimization via Newton-like updates
- Adaboost
  - includes feature selection
  - focuses on hard examples
  - generalizes well

Conditional Random Fields

Conditional Random Fields (CRFs) a natural way to model correlations between labels

- $\phi_i(Y_i)$ are local evidence potentials
- $\psi_i(Y_i, Y_j)$ are compatibility potentials

$$p(Y|x) = \frac{1}{Z} \prod_i \phi_i(Y_i) \prod_{j \in N_i} \psi_{i,j}(Y_i, Y_j)$$

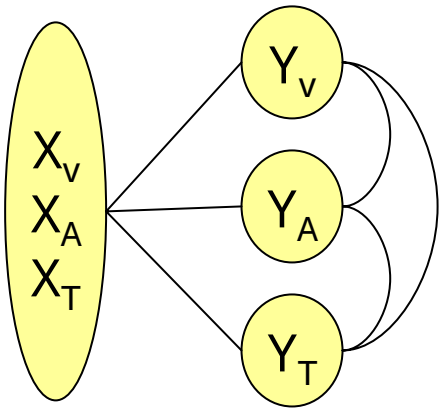
$$Y_i \in \{-1, +1\}$$

Boosted random fields

- Boosted Random Fields
  - The belief of a node i at iteration t is $b_i^t(Y_i) = \phi_i^t(Y_i) M_i^t(Y_i)$
  - M is the product of all messages coming from neighbors

$$M_i^{t+1}(Y_i) = \prod_{k \in N_i} \mu_{k \to i}^{t+1}(Y_i)$$

$$\mu_{k \to i}^{t+1}(Y_i) = \sum_{y_k \in L} \psi_{k,i}(y_k, Y_i) \frac{b_k^t(y_k)}{\mu_{i \to k}(y_k)}$$



  - The cost function to be minimized is the per label loss :

$$J = -\prod_m \prod_i b_{i,m}^t(+1)^{\frac{(Y_{i,m}+1)}{2}} b_{i,m}^t(-1)^{1-\frac{(Y_{i,m}+1)}{2}}$$

m = training instance index

Boosted random fields (cont'd)

- If $\phi_i^t = e^{Y_i F_i}$ it can be shown that :

$$b_i^t = \frac{1}{1 + e^{-(F_i^t + G_i^t)}}$$
$$G_i^t = \log M_i^t(+1) - \log M_i^t(-1)$$

- The cost function simplifies to :

$$\log J_i^t = \sum_m \log(1 + e^{-Y_{i,m}(F_{i,m}^t + G_{i,m}^t)})$$

- where :

$$F_{i,m}^t = \sum_{n=1}^{t} f_i^t(x_{i,m})$$
$$G_{i,m}^t = \sum_{n=1}^{t} g_i^n(b_m^t)$$

Boosted random fields (algo)

- Algorithm
  - For t=1...T
    - Learn weak classifier f_i^t for local evidence potentials
    - Learn weak classifier g_i^t for compatibility potentials
    - Update local evidence potentials $F_{i,m}^t = F_{i,m}^{t-1} + f_i^t(x_{i,m})$
    - Update compatibility potentials $G_{i,m}^t = \sum_{n=1}^{t} g_i^n(b_{N_i,m}^{t-1})$
    - Update beliefs using approximate belief propagation
    - Update weights on training examples

Results: Text Clusters (TRECVid 2003)

	Most frequent words
Cluster 1	rain weather with storm continue today across temperature forecast new california coast
Cluster 2	med won gold olymp when team with game five two today hi second sport
Cluster 3	presid clinton today south with say hi africa nate first visit lead talk
Cluster 4	day with look go what plain like all wait make get again off well first win season
Cluster 5	presid with house clinton white about lawyer investigation jury grand said case sexual
Cluster 6	point nasdaq dow wall street gain back market stock industrial today jones eighty seven close
...	...

Evaluation

- Contextual *vs* multimodal approaches

Visual	Anchor	Outdoor	Financial	Weather	Ads
	57/**67**	44/**53**	65/**71**	82/**83**	56/**55**
Audio	Speech	Music	Noise		
	67/**78**	60/**64**	30/**32**		
Text	Cluster1	Cluster2	Cluster3	Cluster4	...
	80/**83**	72/**76**	75/**70**	74/**81**	...

Segmentation: Example

HIS SECOND ATTEMPT NATHAN UNDER THE WEIGHT DESPITE LAST YOU DID SHORT OF HIS GOAL TO BE THE FIRST BALLOONISTS TO FLY AROUND THE WORLD NONSTOP SHEILA MACVICAR A.B.C. NEWS LONDON

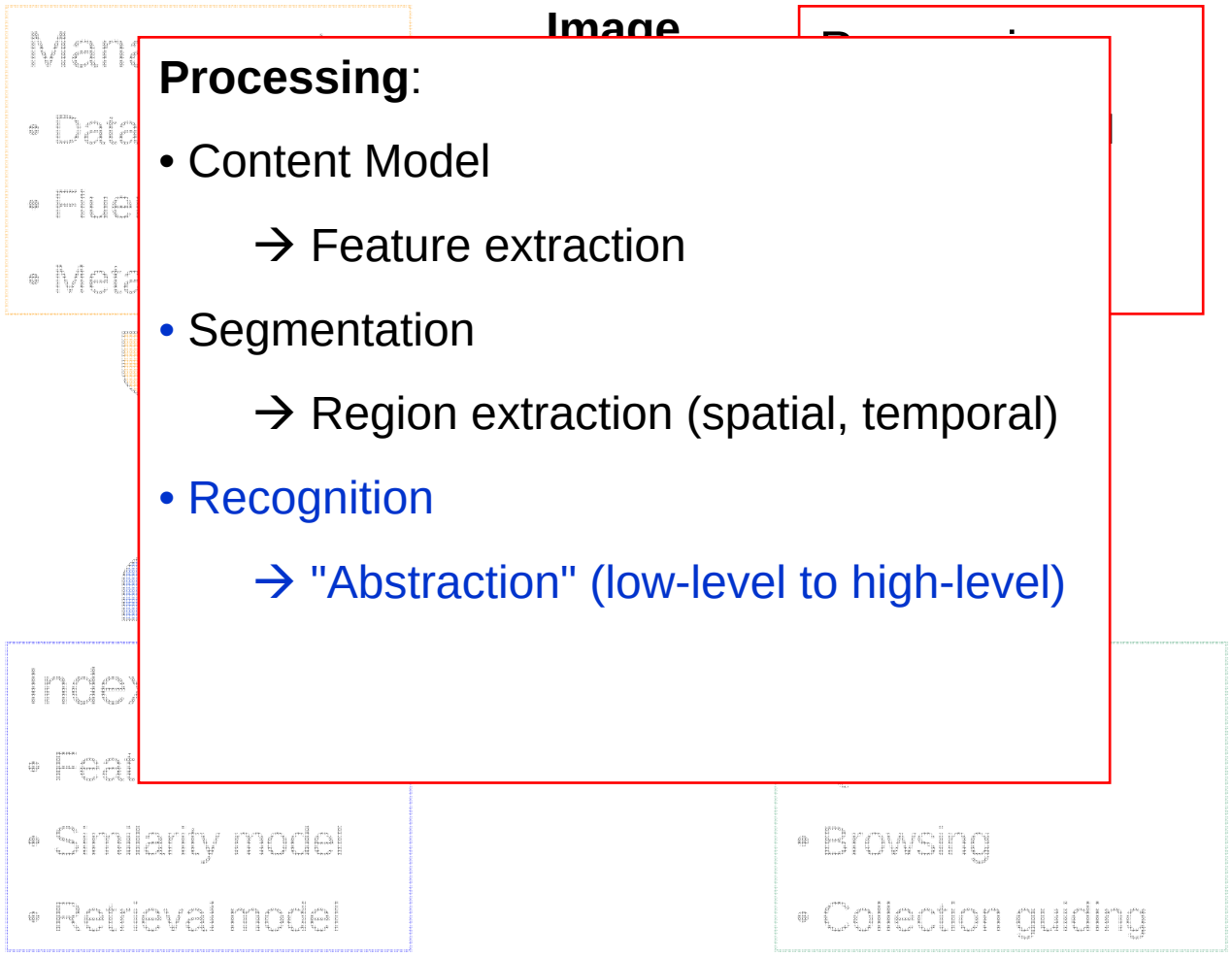
THIS WAS THE DEADLIEST WEEKEND IN MEMORY IN THE ROCKY MOUNTAINS A SERIES OF AVALANCHES STRUCK ON BOTH SIDES OF THE CANADIAN U.S. BORDER IN MONTANA TWO PEOPLE DIED IN SEPARATE INCIDENTS

News story segmentation: useful cues

- Boosting tells us which were the most efficient features to derive the final results
- In our case, labels associated to streams (V, A, T)

1	News subject monologue after (V)
2	Graphics and text after (V)
3	Graphics and text before (V)
4	Non studio settings after (V)
5	Sport event after (V)
6	Text cluster 4 after (T)
7	Text cluster 13 after (T)
8	Text cluster 16 before (T)
...	...

Part II: Processing



Content abstraction

- From low-level (perceptual) content and external knowledge, infer high-abstraction (semantics)
  - *Recognition*
  - *Labeling*
  - *Cross-modal transform*

- Examples
  - Speech recognition
  - Face recognition
  - Object recognition
  - Video summarization
  - Auto-annotation

Image auto-annotation

- Essential challenge in modern IR:

  Ability to deduce high-level semantics from low-level perceptual features of multimedia.

- Terminology:
  - Semantic categorization
  - Keyword prediction
  - Auto-annotation
  - Automatic linguistic indexing

Approaches

- Contributing domains:
  - Generative probabilistic models
    (Barnard *et al.*)
  - LSI, cross-language extension
    (Zhao and Grosky, Praks *et al.*, Monay and Gatica-Perez)
  - Multiple-category classification
    (E. Chang *et al.*, K. Goh *et al.*, Li and Goh, Li and Wang)
  - Hierarchical Semantic Ensembles
    (S. Kosinov *et al.*)

- General assumption to be challenged:
  - The semantic categories (i.e., target classes) are
    (a) independent,
    (b) non-overlapping,
    (c) exhaustive.

Main principle

The semantic categories (i.e., target classes) should be considered as:

- Dependent:
  - Can antagonist labels appear together?
  - ⇒ "sun" and "night"
- Overlapping:
  - Labels may partly bring the same information
  - ⇒ "Dolphin" and "sea"
- Non-exhaustive:
  - A scale should be chosen
  - ⇒ "Person" and "Body", "Arms", "Legs",...

Some notations

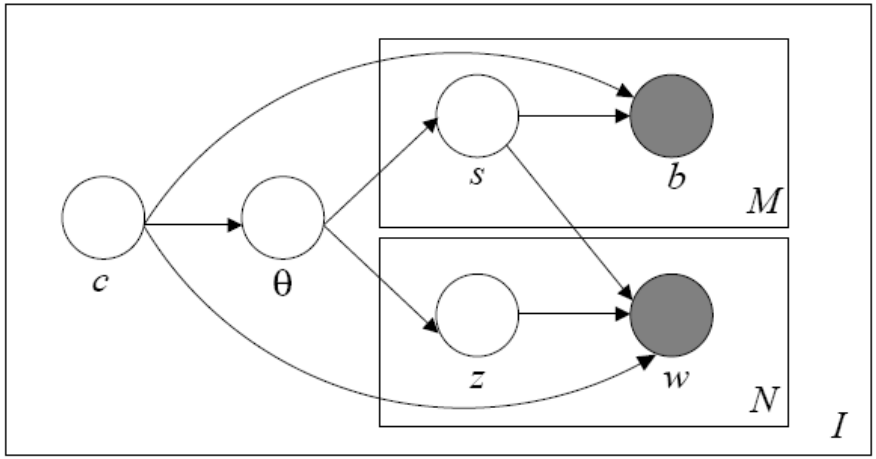
- $X = \{I_t, K_t\}$ - Training set of annotated images
- I_t – feature vector representation of t^{th} image
- K_t – associated keyword set of t^{th} image
-

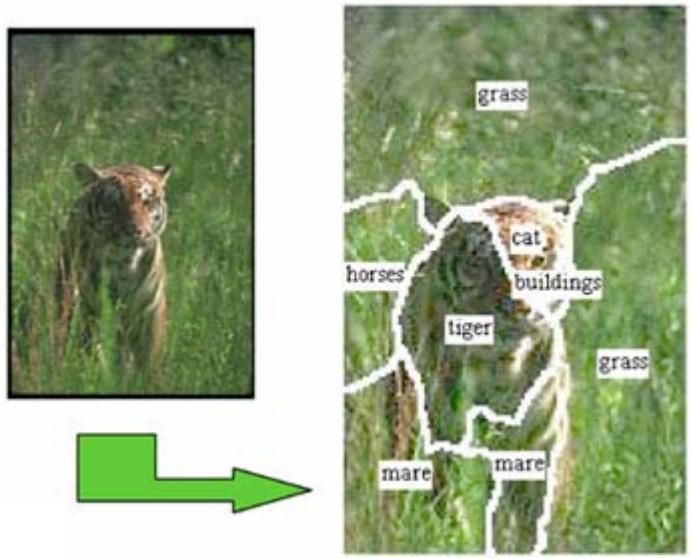
$V = \cup K_t$ – all unique nouns of the annotation vocabulary

- $H = \{C_i\}$ – a concept hierarchy comprising V and hypernyms
- Φ_i – a binary baseline classifier associated with C_i
- $L(C_i)$ – set of leaf concepts subsumed by C_i
-

- I_u – a query image feature vector to be autoannotated

Matching words and pictures [Barnard, 2003]



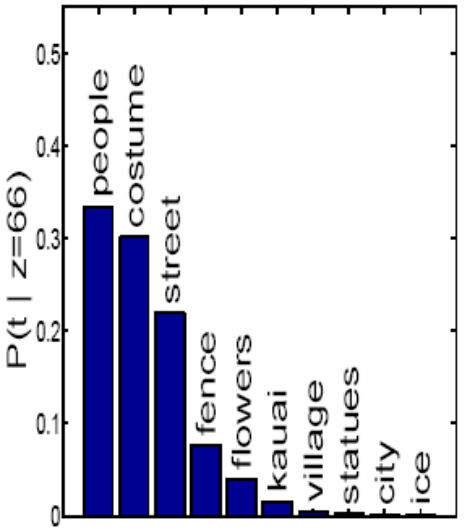


$$p(w|b) = \sum_{c=1}^{J} p(c|\phi) \sum_{z=1}^{K} p(w|z) \int p(z|\theta) p(\theta|\gamma_c) d\theta$$

Goal: labelling image regions (blobs *b*) with words (*w*)

Latent Dirichlet Allocation (LDA) model combined with the Mixture of Multimodal models (MoM) principle
⇒ Computing posteriors over mixtures and distributions
⇒ Creating a word prediction for regions

PLSA-based annotation [Monay, 2004]



Goal: Annotate full images with a given vocabulary

Principle: Words (*w*) are conditional independent from queries images (*q*), provided a marginalization over latent concepts (*z*)

$$P(w_j \mid q) = \sum_{k=1}^{K} P(w_j \mid z_k) P(z_k \mid q)$$

Use of a lexical database

Since at one end we have text, we can model correlations using external available knowledge

⇒ WordNet: http://wordnet.princeton.edu/

- Models relationships within the English vocabulary
  - Synonyms, Hyponyms, Hypernyms
  - Meronyms, antonyms,Troponyms,...
- Mostly used concept: Synset:
  "A synonym set; a set of words that are interchangeable in some context."
- Each word has some Senses:
  "*A meaning of a word in WordNet. Each sense of a word is in a different synset."*

⇒ WordNet is becoming an essential resource for multimedia indexing

WordNet Hypernym Query Example

Hypernyms (Ordered by Estimated Frequency) of noun tree

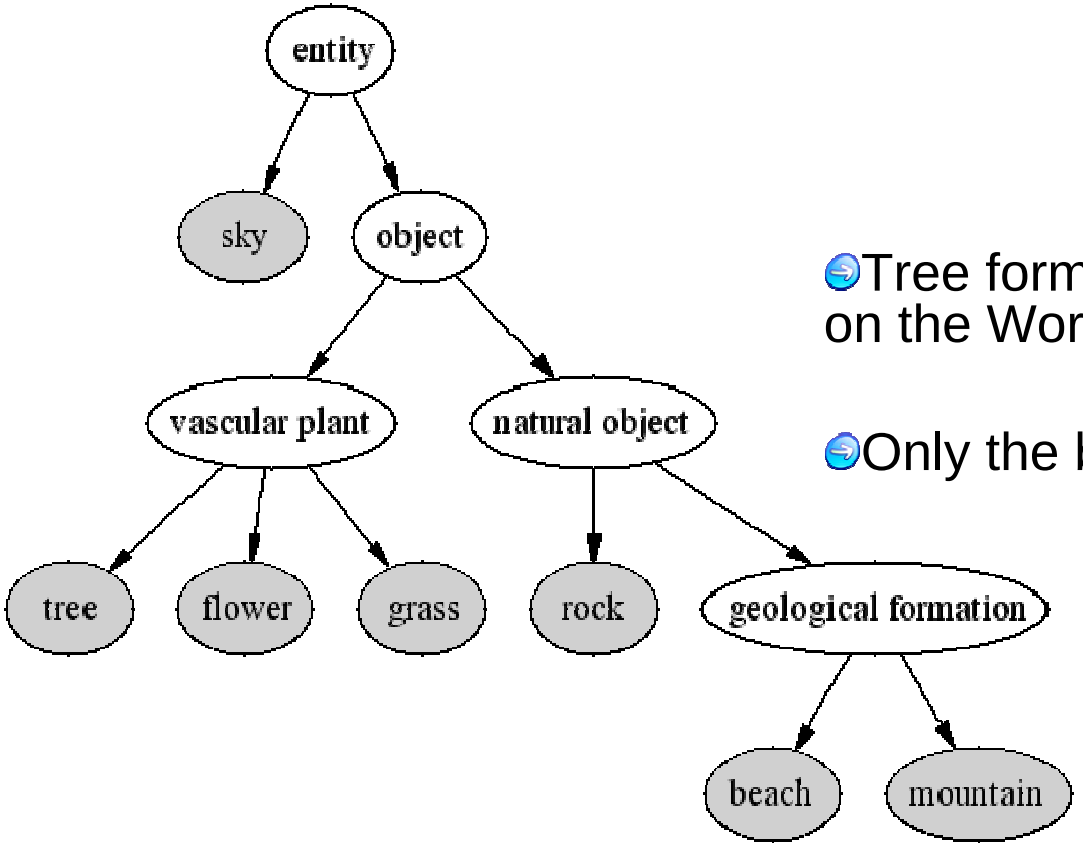
3 senses of tree

```
Sense 1
tree
       => woody plant, ligneous plant
           => vascular plant, tracheophyte
               => plant, flora, plant life
                   => organism, being
                       => living thing, animate thing
                           => object, physical object
                               => entity, physical thing

Sense 2
...
```

Image Auto-annotation

$\boldsymbol{V}$ = {beach, flower, grass, mountain, rock, sky, tree}
Shaded nodes (leaves) denote $\boldsymbol{C_i} \in \boldsymbol{V}$



- Tree formed out of keywords aligned on the WordNet hierarchy
- Only the branching nodes are kept

Image Auto-annotation

Premise:

Each concept C_i is to be assessed as a potential candidate

Main question:

How to ensure comparability of different $C_i \in H$?

"*Goodness of fit* vs. *Concept specificity*" trade-off:

- $P(C_i \mid I_u)$ – how well concept C_i is fit by the feature data from the classification accuracy point of view (concept posterior given data);
- $P(C_i \mid k)$ – how specific or unambiguous the candidate set of keywords $L(C_i)$ is (concept posterior assuming a particular keyword k from the set of hyponyms of C_i is selected correctly).

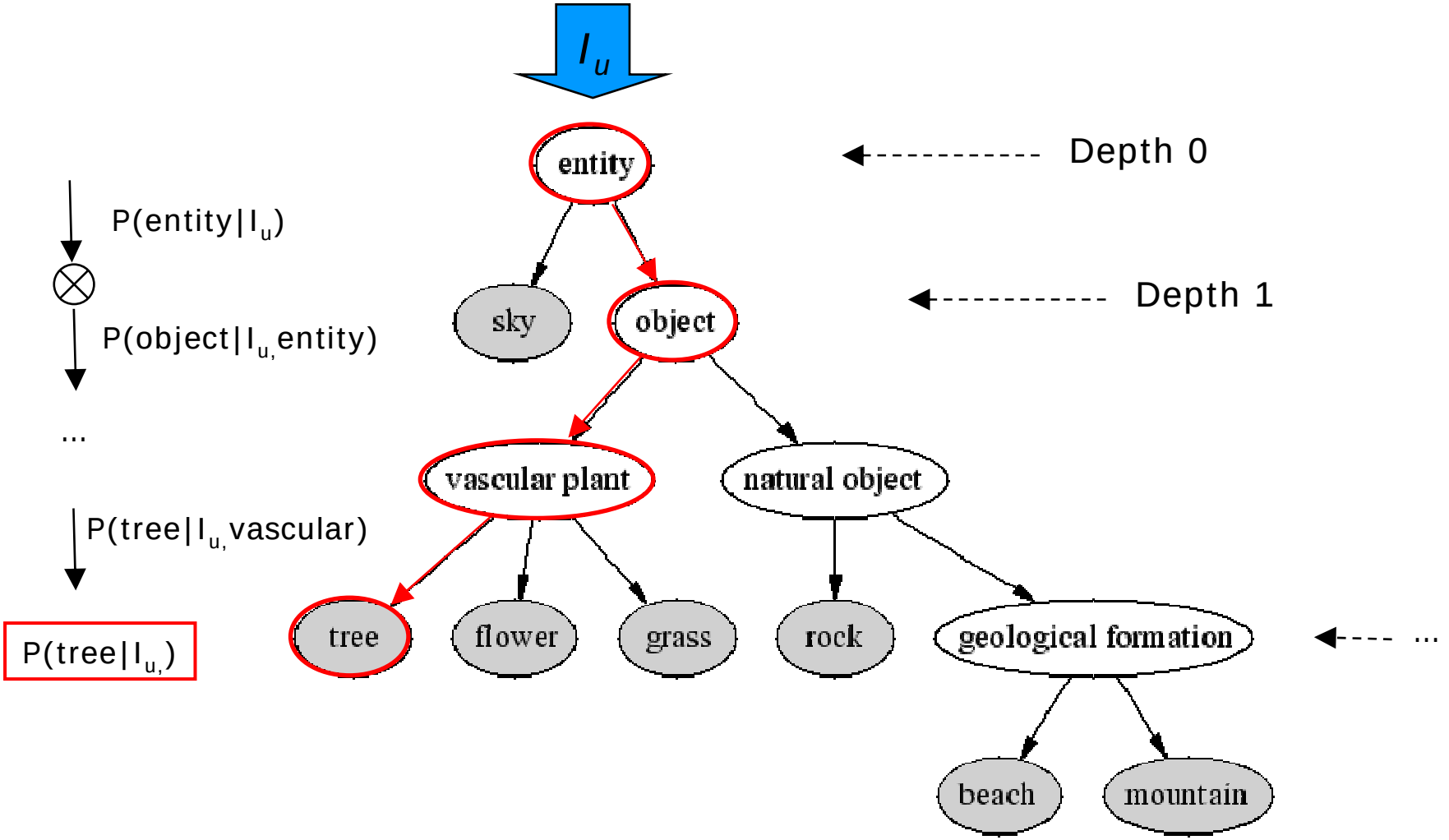
Goodness of Fit: $P(C_i|I_u)$

Theorem 1, (Kumar *et al.*, 2002). *The posterior probability $P(C_i|I_U)$ for any input I_U is the product of the posterior probabilities of all the internal classifiers along a unique path from the root node to C_i, i.e.*

$$P(C_i|I_U) = \prod_{l=0}^{\mathscr{D}(C_i)-1} P(C_i^{(l+1)}|I_U, C_i^{(l)}), \quad (1)$$

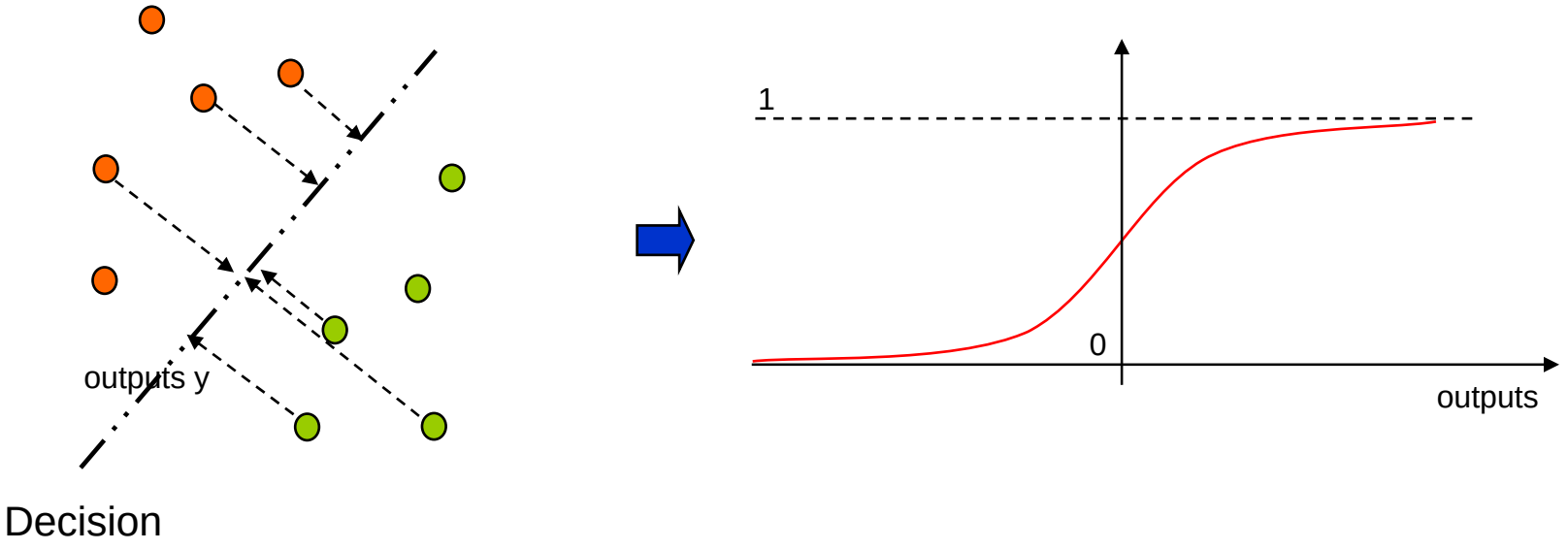
where $\mathscr{D}(C_i)$ is the depth of C_i (the depth of the root concept C_1 is 0), $C_i^{(l)}$ is the concept at depth on the path from the root node to C_i, such that $C_i^{(\mathscr{D}(C_i))} \equiv C_i$ and $C_i^{(0)} \equiv C_1$.

Conditional certainty



From outputs to probabilities

Caveat: To utilize baseline classifiers Φ_i with non-probabilistic outputs, a sigmoid is fit to raw output values y_i, (Platt, 1999).



Concept Specificity, P($C_i|k$)

Bayes Theorem:

$$P(C_i|k) = \frac{P(k|C_i)P(C_i)}{\sum_{C_i \in H} P(k|C_i)P(C_i)},$$

where:

$$P(C_i) = \frac{\sum_{C \in \mathbf{L}(C_i)} freq^{(T)}(C)}{\sum_{C \in V} freq^{(T)}(C)}, \quad P(k|C_i) = \frac{\min_{C \in \mathbf{L}(C_i)} freq^{(W)}(C)}{freq^{(W)}(C_i)}$$

$freq^{(w)}(C_i)$ – the cardinality of the WordNet hyponym set of C_i

$freq^{(T)}(C_i)$ – the frequency of a given concept in the training data annotation corpus

Concept Relevance, $P(C_i|k, I_u)$

Assumption:

likelihood of the input data I_u given C_i is not dependent on the correctness of a particular choice of k from the hyponym set of C_i

Concept relevance:

$$P(C_i|I_U, k) \propto P(C_i|I_U)P(C_i|k) = \rho,$$

Example application

Image collection: Corel

- Directories of 100 images for a class
- 4 keywords per image (flat annotation)
- High intra class similarity
- Depending on the subset, high inter-class dissimilarity

- Example:

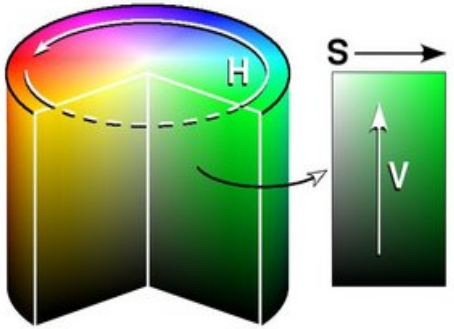
Supplied annotation:

{*flowers, path, grass, trees*}

Low-level visual feature representation

Global Color Histogram data in HSV space:

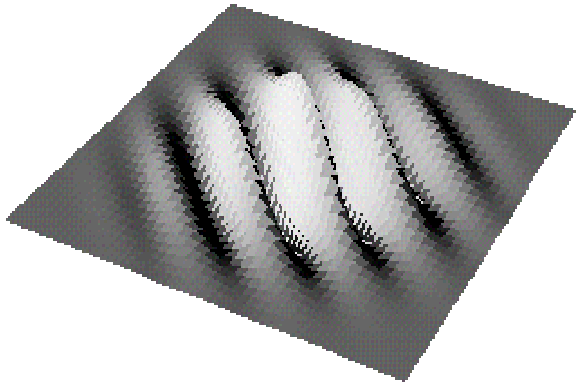
- 18 hue, 3 saturation, 3 value levels, together with additional 4 grey levels
- 18*3*3+4 = 166 color descriptors



Global Texture data from Gabor filter banks:

$$g(x,y) = \left(\frac{1}{2\pi\sigma_x\sigma_y}\right)\exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)+2\pi jWx\right)$$

- 3 scales, 4 orientations, 10 bands
- 3*4*10 = 120 texture descriptors

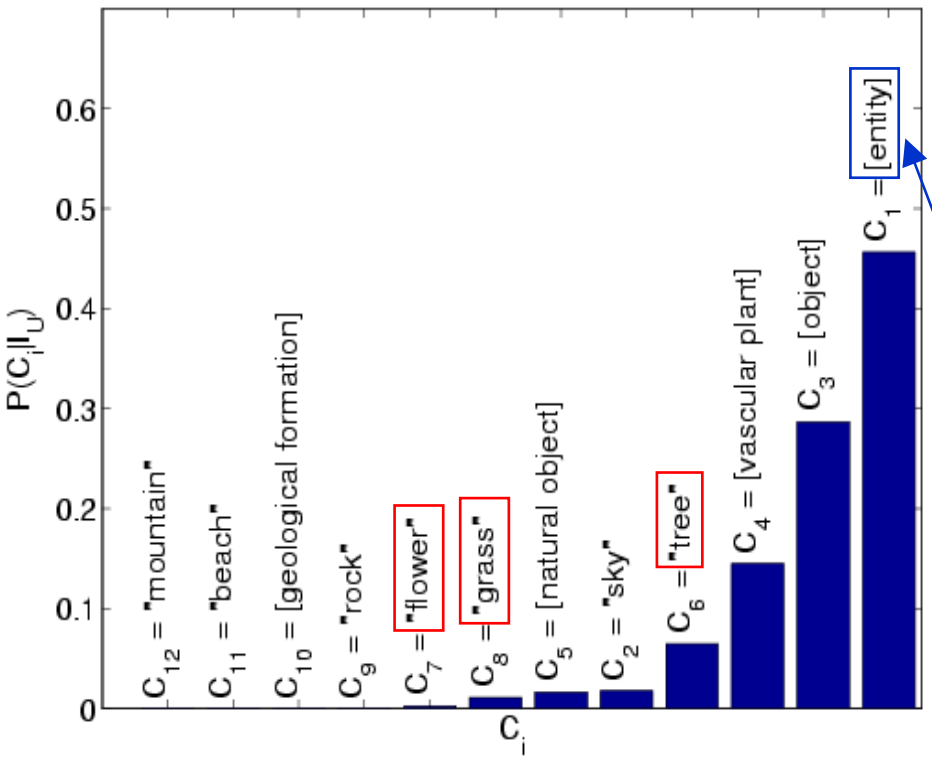


Example

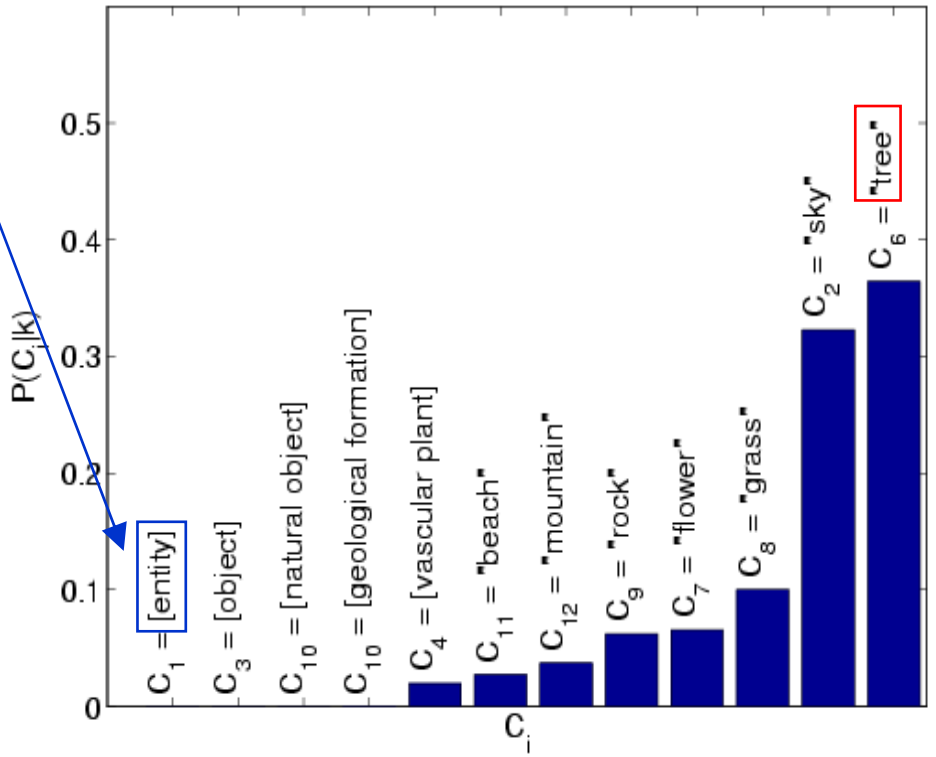
Supplied annotation:

{*flowers, path, grass, trees*}

Goodness of fit $P(C_i|I_u)$



Concept specificity $P(C_i|k)$



Example

Supplied annotation:

{*flowers, path, grass, trees***}**

TABLE 1: CANDIDATE CONCEPTS RANKED BY RELEVANCE

Rank	$-\log_2 \rho(C_i)$	Concept C_i
1	5.41	C_6 = tree
2	7.46	C_2 = sky
3	8.44	C_4 = vascular plant (flower, grass, tree)
4	9.84	C_8 = grass
5	12.64	C_7 = flower
6	**17.26**	**C_1 = entity**
7	17.87	C_3 = object
8	19.42	C_5 = natural object
9	21.00	C_9 = rock
10	44.32	C_{10} = geological formation
11	55.97	C_{12} = mountain
12	56.35	C_{11} = beach

Baseline classifiers Φ_i

Support Vector Machines (SVM)

$$f(\mathbf{x}) = h(\mathbf{x}) + b,$$

$$h(\mathbf{x}) = \sum_i y_i \alpha_i k(\mathbf{x}_i, \mathbf{x})$$

minimization:

$$C \sum_i (1 - y_i f_i)_+ + \frac{1}{2} \|h\|_{\mathcal{F}}$$

Discriminant Analysis (DDA)

minimization:

$$J(T) = \frac{\left(\prod_{i<j}^{N_X} \Psi \left(d_{ij}^W(T) \right) \right)^{\frac{2}{N_X(N_X-1)}}}{\left(\prod_{i=1}^{N_X} \prod_{j=1}^{N_Y} d_{ij}^B(T) \right)^{\frac{1}{N_X N_Y}}}$$

Fitting posterior probabilities

Parameterized sigmoid representation of the posterior probability of a concept, Platt (1999):

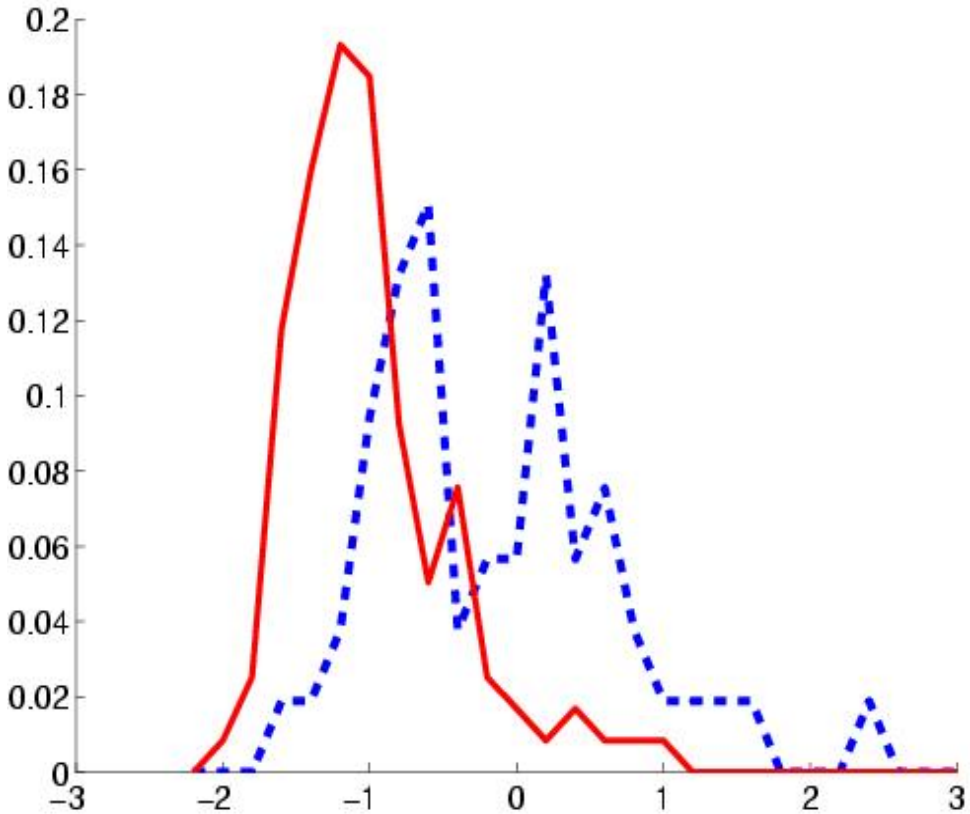
$$P(C|f) = \frac{1}{1+\exp(Af+B)}$$

$f = f(I_u)$ – raw output of a baseline classifier

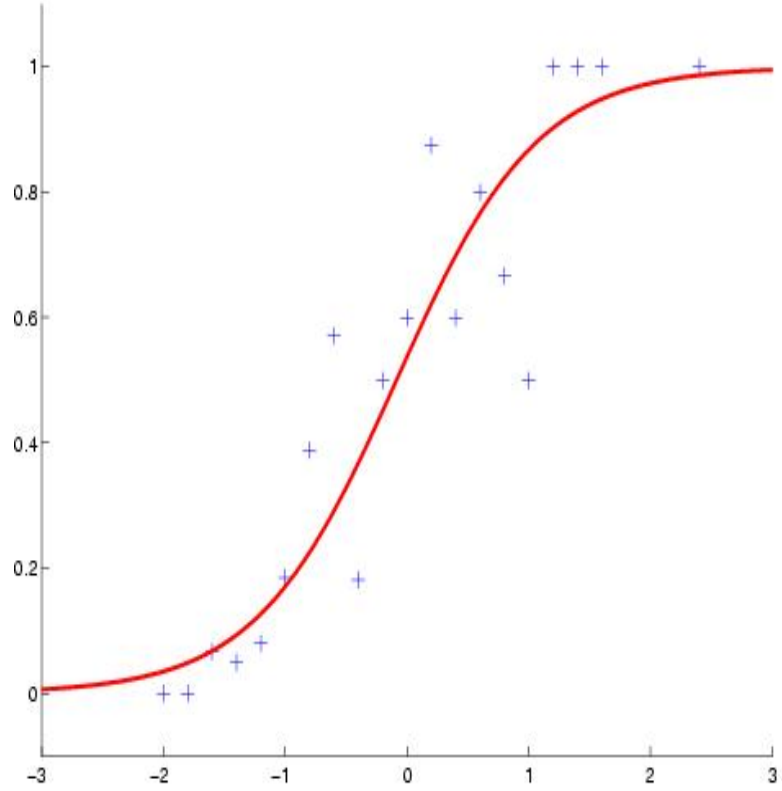
Parameters A,B are fit using MLE:

$$\min_{A,B} -\sum_i t_i \log(p_i) + (1-t_i)\log(1-p_i)$$

Illustration for concept **C** = "**trees**"



Class-conditional densities of raw Gaussian SVM outputs for a concept (dashed) and its complement (solid).



Comparison between the fitted sigmoid and the actual posterior probabilities.

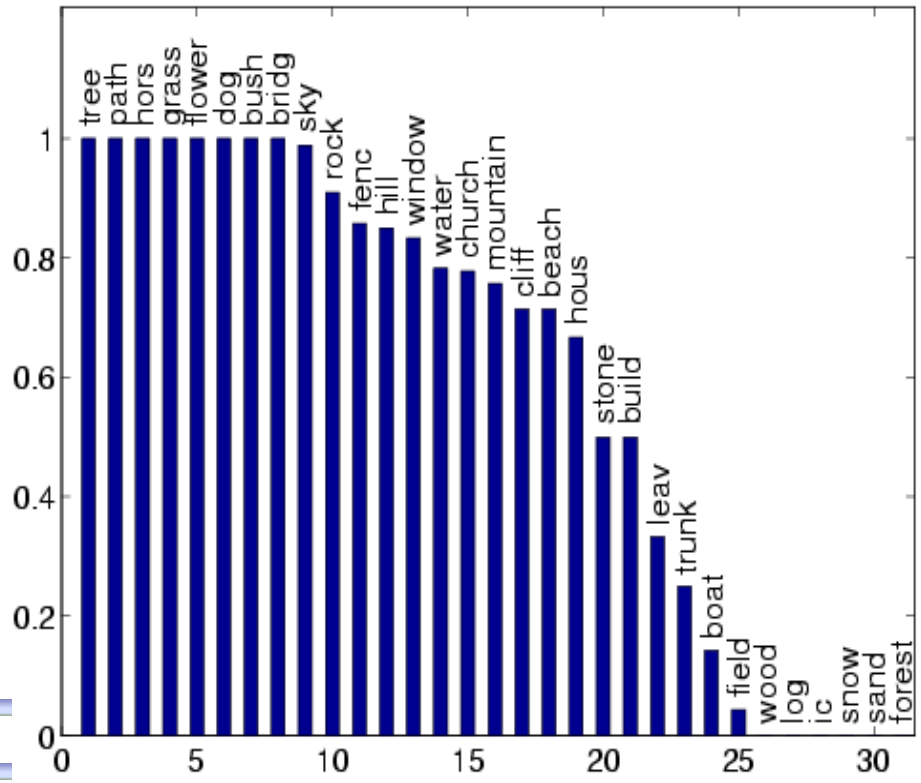
Example results

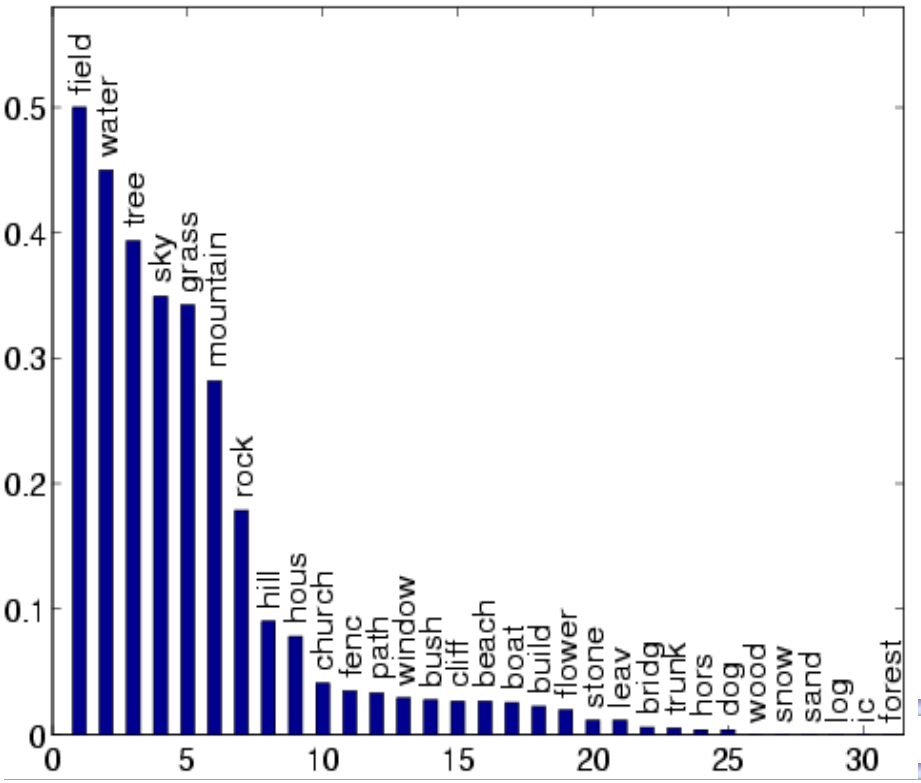
Training data: ~600 images, Washington collection

Testing data: ~250 images, COREL

204 unique tokens, $\text{freq}^{(T)}(C_i) < 2$ removed, $|H| = 60$

$C_i \notin V$, $L(C_i)$ is predicted, e.g. [vessel,watercraft] → {boat, sailboat, ferryboat}





True annotation:
sky, street, buildings, town
Autoannotation:
sky, construction,
natural object, artefact

True annotation:
sky, castle, water, tree
Autoannotation:
sky, rock, tree

True annotation:
cows, road, trees, grass
Autoannotation:
bush, tree, grass, vascular
plant, woody plant, organism

True annotation:
sky, water, mountain, trees,
Autoannotation:
sky, water, geological formation,
natural object, artefact

Some classification figures

TABLE 2: CLASSIFIER ENSEMBLE PERFORMANCE RESTRICTED TO TOP 5 KEYWORDS

Ensemble	Baseline classifier	% Recall	% Precision
Empirical	none	16.13	5.04
Max Wins	SVM, polynomial	8.14	3.83
Max Wins	SVM, gaussian	10.61	4.47
OPC	SVM, polynomial	20.31	7.85
OPC	SVM, gaussian	21.27	10.19
HSE	DDA	21.22	10.20
HSE+siblings	DDA	28.42	26.88

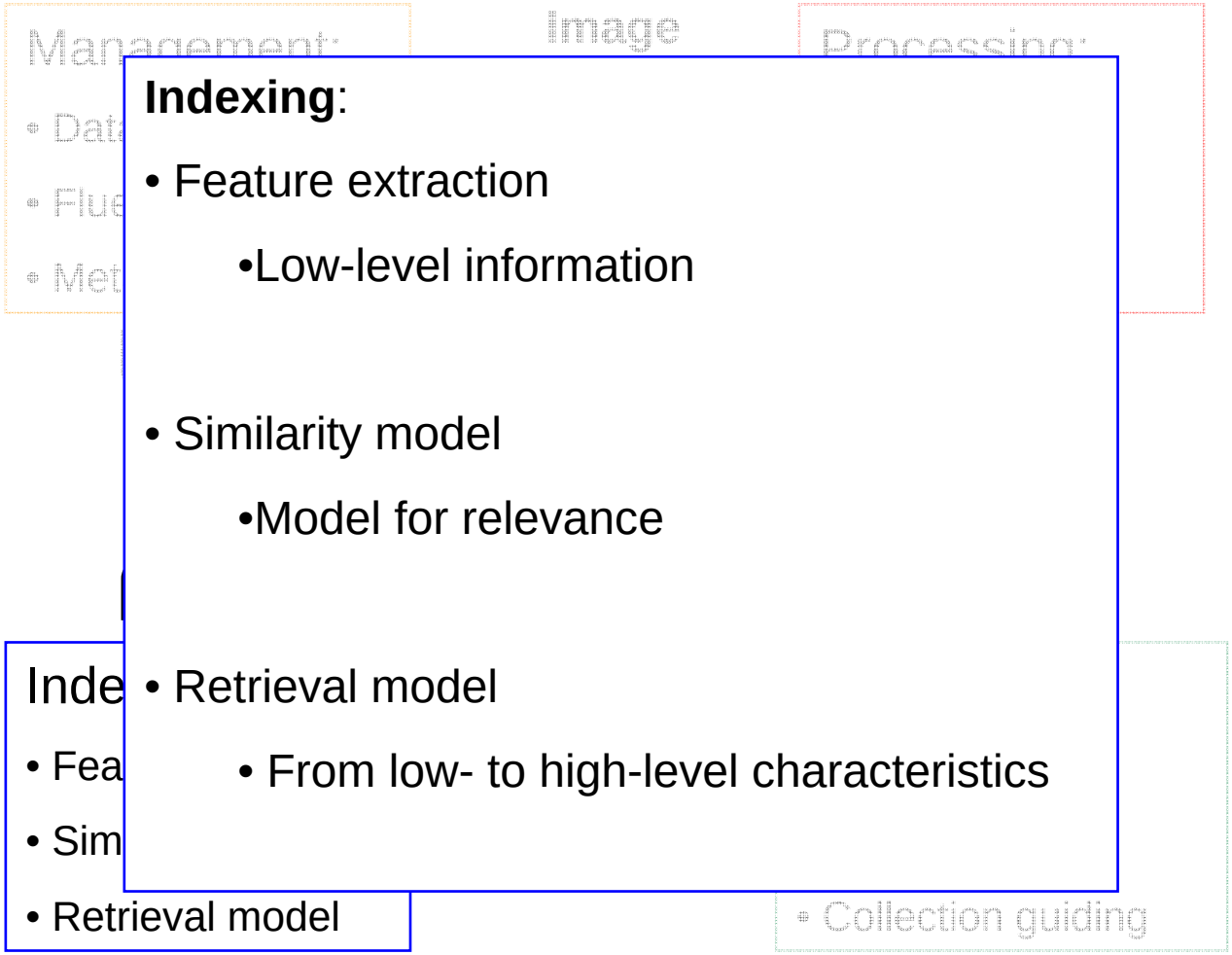
TABLE 3: INFLUENCE OF BASELINE CLASSIFIER ON HSE PERFORMANCE

Baseline classifier	% Recall	% Precision
SVM, linear	18.12	5.28
SVM, polynomial	18.34	5.67
SVM, gaussian	18.62	6.05
DDA	21.22	10.20

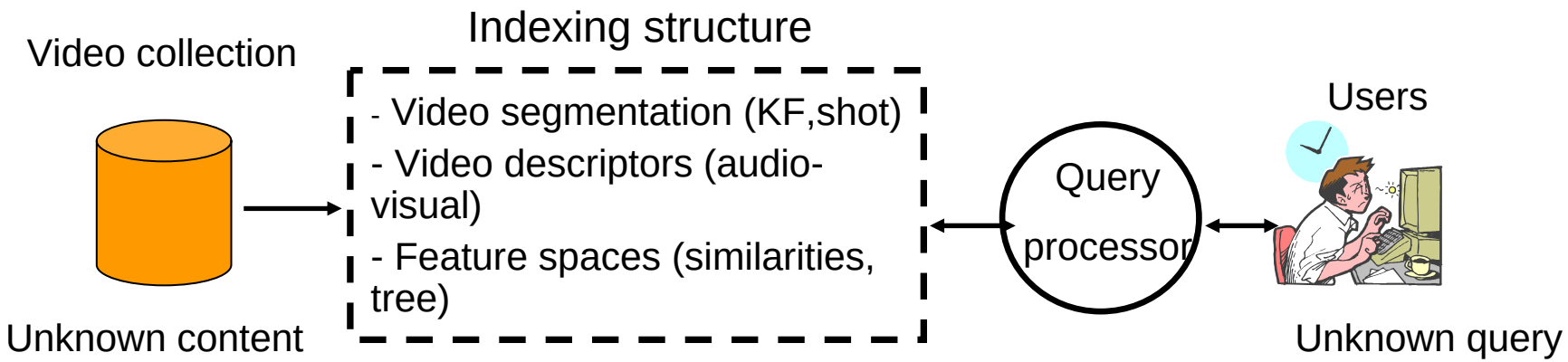
Still a long way to go...

- Recall is around 20%
- Only considers IS_A relationships
- Image is seen as a whole
  - Segmentation
  - Scale selection

Part III: Indexing and retrieval

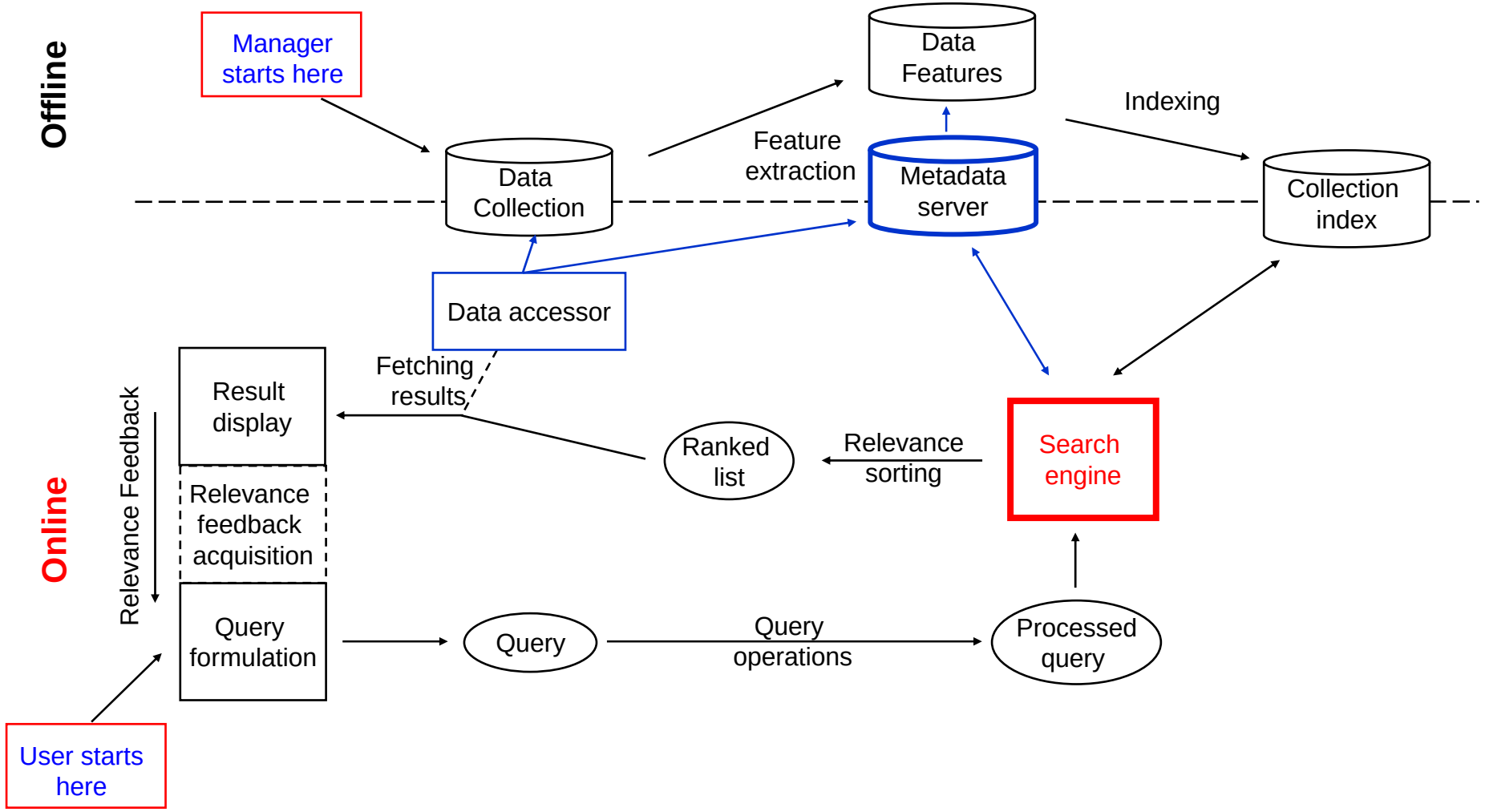


Video indexing and retrieval



- Queries can hardly anticipated (no prior query distribution)
  - ⇒ Little information for indexing…
  - ⇒… over the data
  - ⇒… over the query type (modality)

Temporal multimedia IR



Reference corpus: TRECVid

- TRECVid2006 Collection
- 170 h of news data (English, Chinese and Arabic)
- Misc data (NASA, BBC rushes)
- MPEG-1 movies grabbed from analogic streams
- ~200GB of raw data (video+audio)
- Comes with manual annotations/descriptions
  - Speech transcripts + ASR
  - Shot segmentation, Story segmentation
  - Salient features (faces, text, ...)
  - Temporal annotation against limited ontology
  - ...

http://www-nlpir.nist.gov/projects/trecvid/

Problem to handle data...

- Video document storage and access
- Multiple and overlapping temporal segmentations
- Annotation and ontology
- Heterogeneous audio-visual descriptors/feature-spaces

⇒ Keep all data synchronized!

⇒ Perform information fusion (visual+audio+text+...)

Classical IR models (*eg* vector space) do not work well

Multimodal video indexing

- Large amounts of high-dimensional descriptors associated to various modalities
- Video segments need to be compared to each others according to their features
- The “bag of features” model is not really valid here
  - Fusion as additive operation
  - Non-homogeneity of features

⇒ The index consists in **dissimilarities matrices computed off-line**

  - Fast retrieval
  - Homogeneity of the index whatever the features used

Interactive multimodal retrieval

- Query-by-example (QBE) paradigm associated to relevance feedback
  - Set of positive & negative examples : $\{S^+, S^-\} \subset S$
- Multiple feature spaces available through dissimilarity matrices
  - Set of *M* feature space distances : $\{d^{f_1}, d^{f_2}, \cdots, d^{f_M}\}$

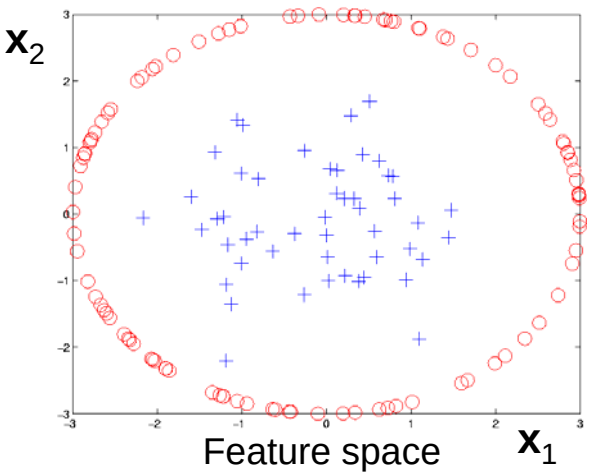
➔ Constraint: real-time interactions

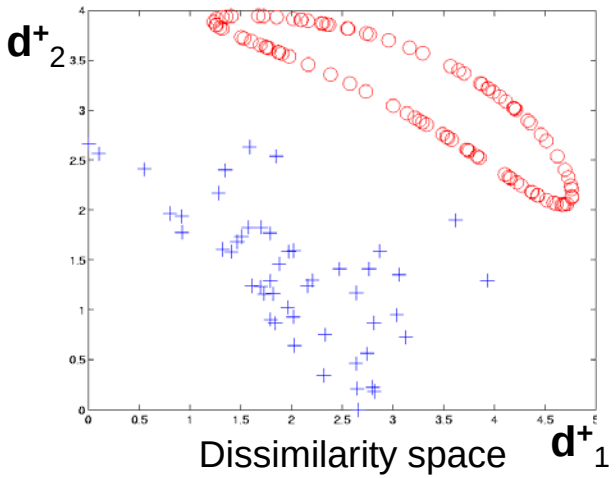
⇒ Dissimilarity-based learning

Dissimilarity space

- Pair-wise dissimilarities replace features
- Be $R=\{p_1,\ldots,p_N\}$ the representation set, the dissimilarity space is:

$$\mathbf{d}(z,R)=[d(z,p_1),d(z,p_2),\cdots,d(z,p_N)]$$



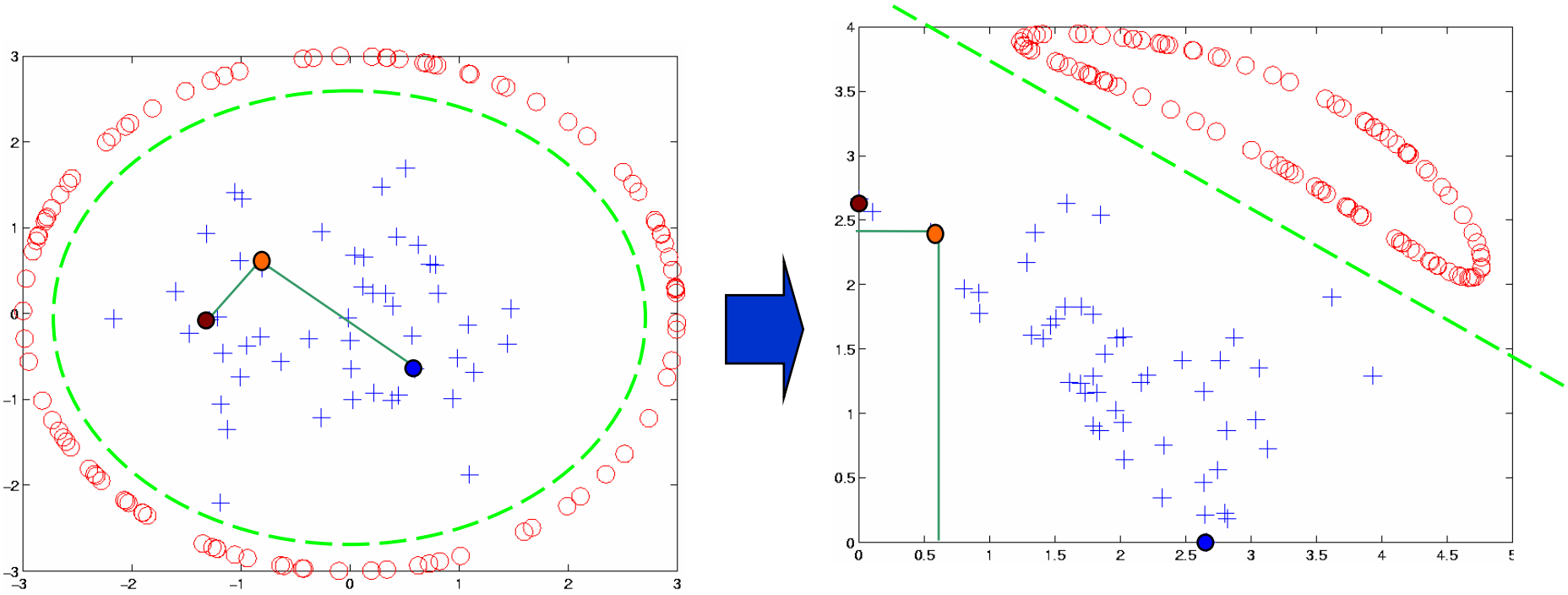


- If $R=S^+$,

  → Low dimensional space (size N)

Dissimilarity representation

Example:



Feature space

Dissimilarity space

Kernel Discriminant Analysis

- Relevance feedback: user gives
  - S^+ positive examples $p_i^+ \rightarrow \mathbf{d}_i^+$
  - S^- negative examples $p_i^- \rightarrow \mathbf{d}_i^-$

- Estimate a ranking function that places positives on the top and pushes negatives to the end

$$\tilde{D} = \arg\max_{D} \frac{\sum_i D(\mathbf{d}_i^-)}{\sum_i D(\mathbf{d}_i^+)} \text{ with } D(\mathbf{d}) = \sum_i \alpha_i k(\mathbf{d}, \mathbf{d}_i^{\pm})$$

➔ Solution is an expansion of kernel functions centered on training vectors

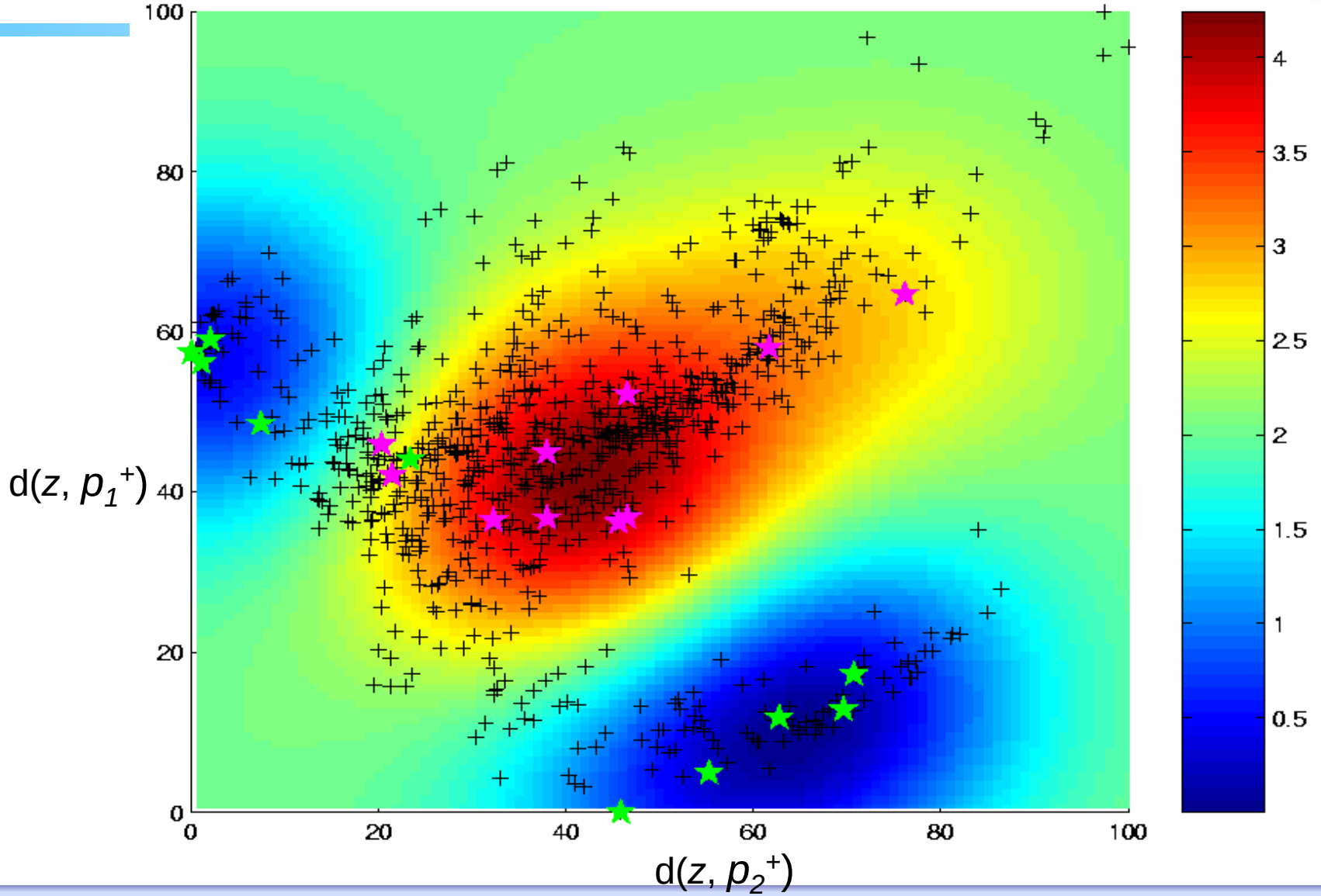
Multimodal Analysis

- Dissimilarities are known for *M* features $d^{f_1}, d^{f_2}, \cdots, d^{f_M}$
  - Multiple dissimilarity spaces $\mathbf{d}^{f_1}, \mathbf{d}^{f_2}, \cdots, \mathbf{d}^{f_M}$
- Concatenation $\mathbf{d} = \left[\mathbf{d}^{f_1}, \mathbf{d}^{f_2}, \cdots, \mathbf{d}^{f_M}\right]$
- Multimodal RBF kernel

$$k(\mathbf{d}_1, \mathbf{d}_2) = \exp\left(-0.5 \cdot (\mathbf{d}_1 - \mathbf{d}_2)^T \Sigma^{-1} (\mathbf{d}_1 - \mathbf{d}_2)\right)$$

$$\Sigma = \mathrm{diag}\left(\underbrace{\sigma_{f_1}^2, \sigma_{f_1}^2, \ldots,}_{|S^+|} \sigma_{f_2}^2, \cdots, \sigma_{f_M}^2\right)$$

Example



Evaluation

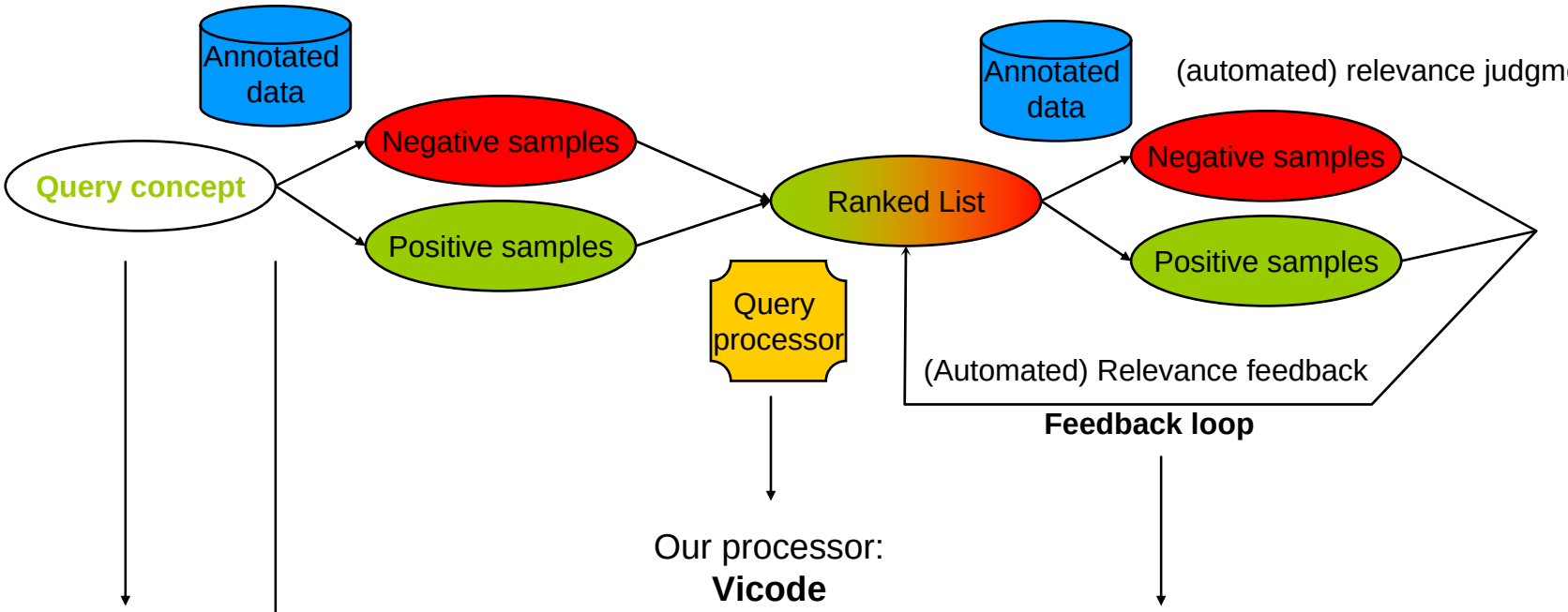
- TRECVid 2003 corpus → around 120 hrs of annotated videos
- **37'000** shots indexed by low-level features
  - Global color histogram
  - Global motion histogram (MPEG motion vectors)
  - ASR histogram (word occurrences and co-occurrences)
- **Euclidean distance** between histograms

$$\text{Recall} = R = \frac{\#\text{ of relevant docs retrieved}}{\text{Total }\#\text{ of relevant docs}} = \frac{TP}{TP + FN}$$

$$\text{Precision} = P = \frac{\#\text{ of relevant docs retrieved}}{\text{Total }\#\text{ of docs retrieved}} = \frac{TP}{TP + FP}$$

We use the Precision, averaged over a number of queries (random feedback effect)

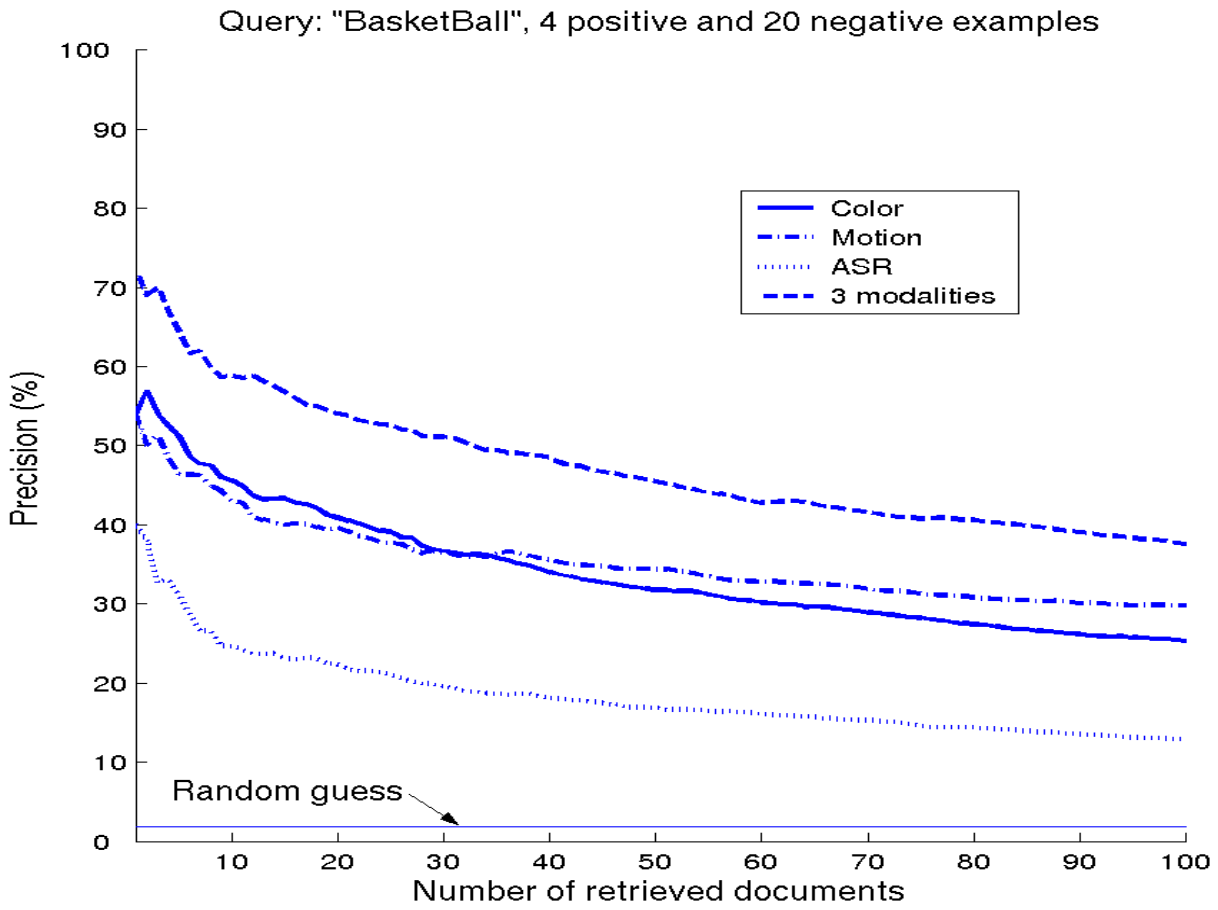
Protocol



- Variation of precision/features used?
- Variation of precision/retrieved set?
- Variation of precision/feedback set size?
- Variation of precision/number of positives?
- Variation of precision/number of negatives?
- …

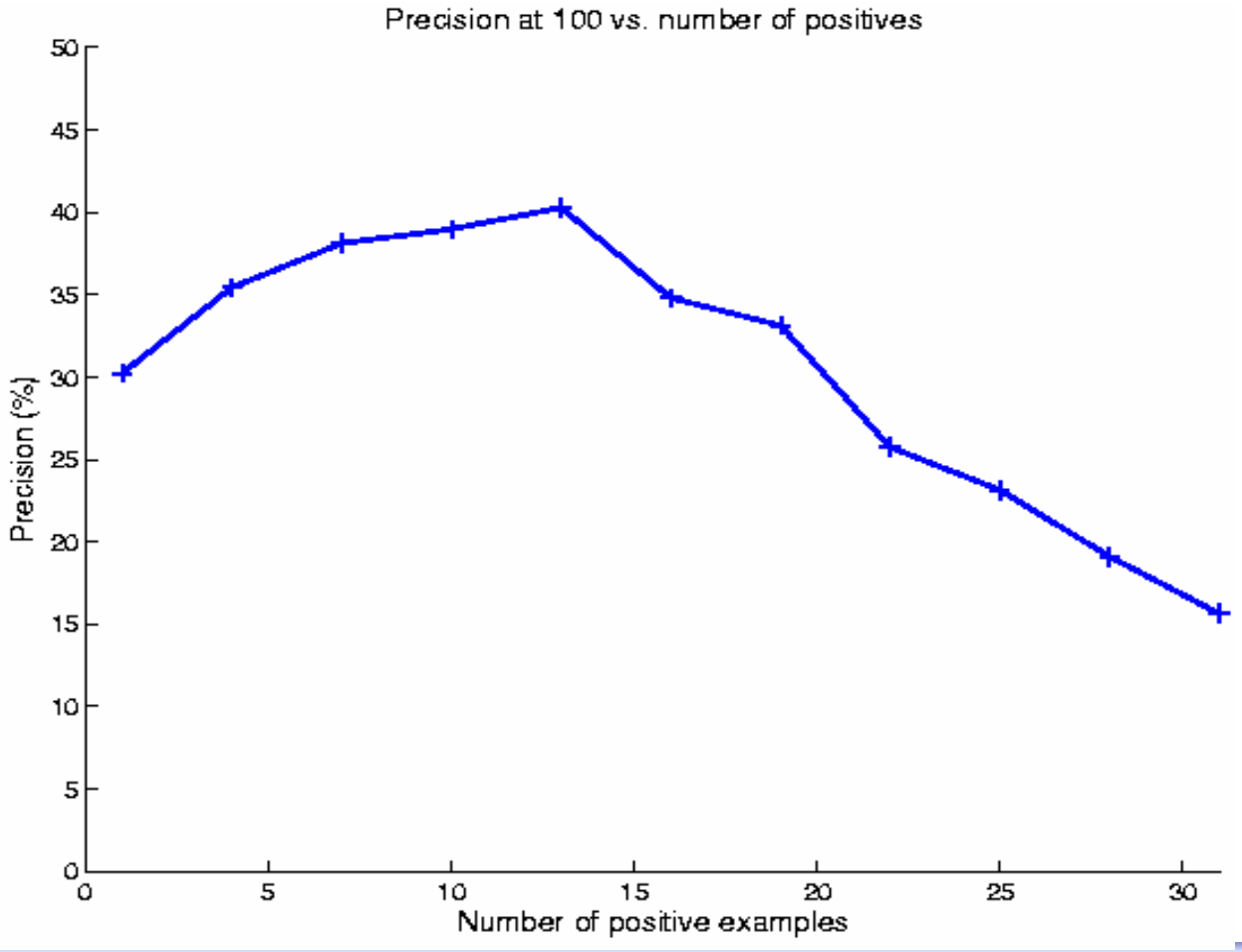
Adding modalities

- Average **Precision** (100 instances of the query « Basketball »)



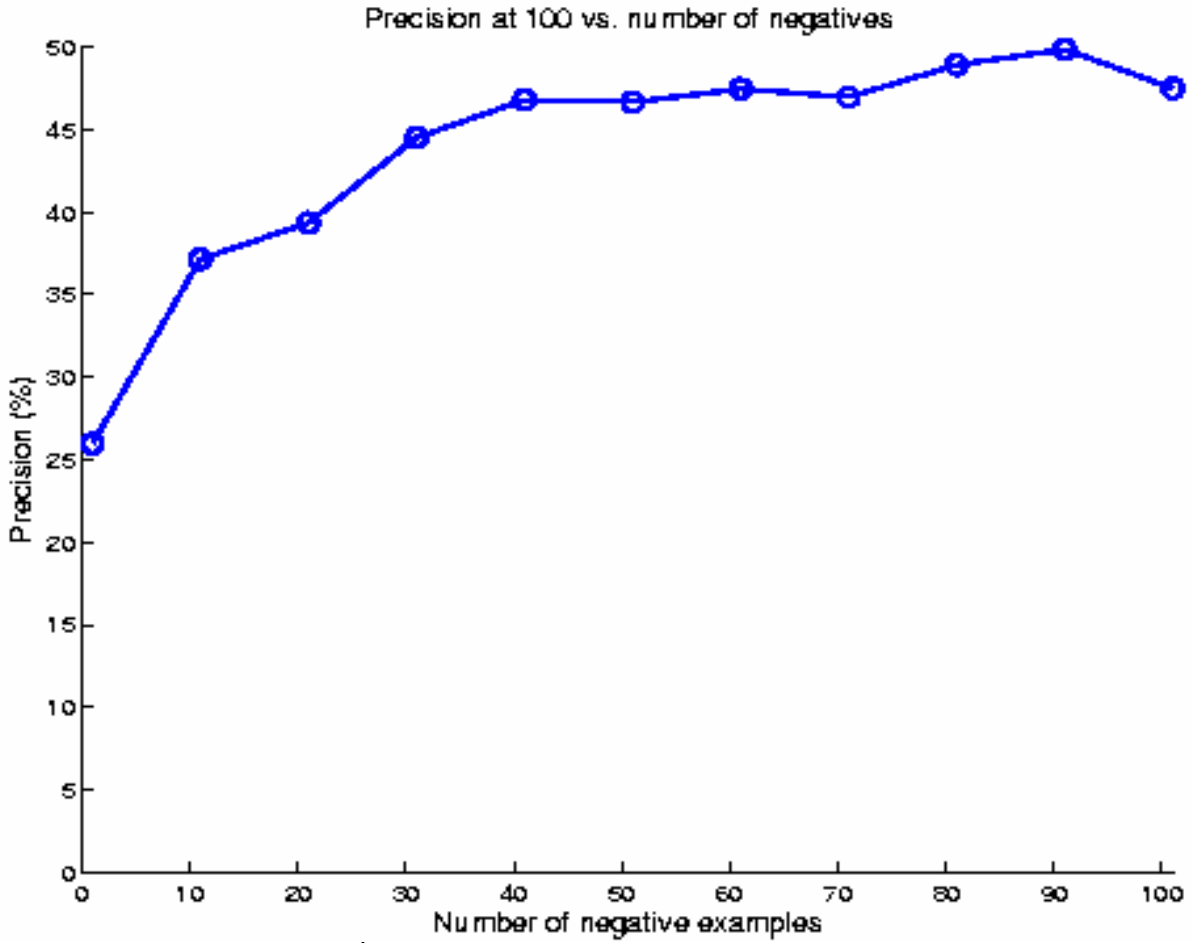
Varying the size of the training set

- Average **Precision** (100 instances of the query « Basketball »)



Cont'd

- Average **Precision** (100 instances of the query « Basketball »)



Part IV: Integrating the user

Interaction:

- Query-by-Example
  - → Relevance Feedback acquisition
- Browsing
  - →User preference
- Serendipity
  - → Facilitate navigation

• Retrieval model

• Collection guiding

Searching for data

- Current systems are query-based
  - QBE, Browsing
  - Implicit goal (target) to reach (retrieve)

⇒ The user is a customer to the system

- Typical limitations:
  - “Page zero” problem
  - Semantic gap
  - Interaction protocol
  - ....

Information search (retrieval)

General assumptions:

- There exists an information need
  - Punctual: one document
  - Broad: category, ensemble
- The user can formulate a description of this need
- The information repository is finite and static

Managing data

The user is a manager of the system

- Large data collection at hand
- No specific needs
- Just wants to keep things tidy:
  - Summarizing, Filtering
  - Sorting, Organizing
  - Annotation, Description

Examples

- Visual content provider
  - Need to know well the content of the collection (overview)
  - Need to create catalogs (summaries)
- Home imagery
  - Content classification
  - Content overview
  - Content annotation
- Massive (Blind) Web Harvester
  - Filtering
  - Classification

Challenge

To create a tool (or context or framework) that would allow a (naive) user to grasp the content of a data collection as quickly as possible

Baselines:

- “Linear visit” of the collection
- Random sampling of the collection

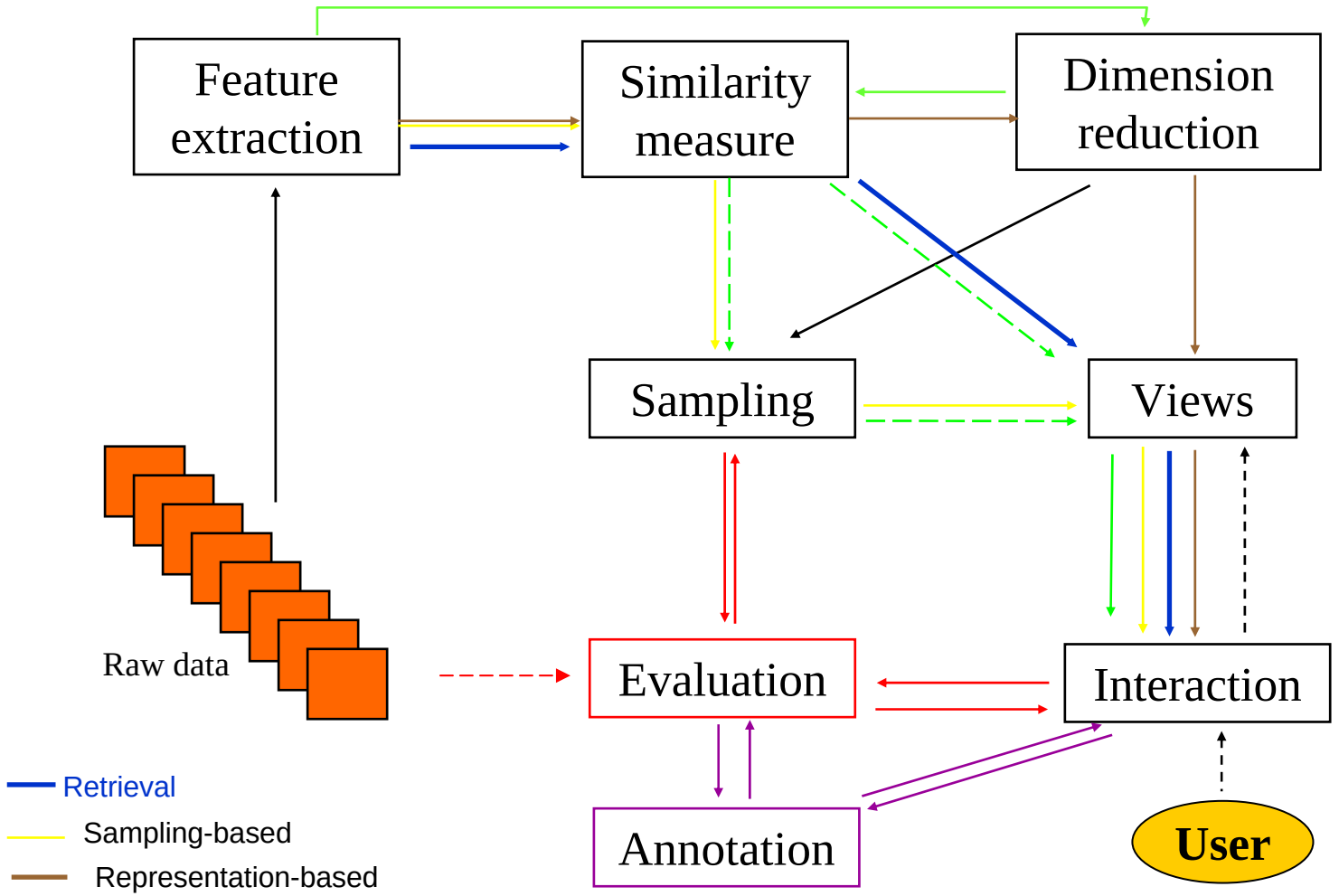
Main principles

- Intelligent sampling
  - Select a subset of the image collection that represent it well
  - Show *n* items (*n* given)
- Hierarchical visit
  - Develop when necessary only
- Organized visit
  - Follow a coherent path within the collection
  - Show *all* items

Several proposals

- D. Cutting *et al.* (1992)
  - Scatter/Gather for text (clustering)
- Y. Rubner (PhD, 1999)
  - EMD, MDS (2D, 3D)
- S. Craver *et al.* (1999)
  - Multi-linearization, space-filling curves
- I. Cox et al. (2000)
  - `PicHunter`, Bayesian Browsing
- K. Barnard (2001)
  - MDS, Clustering

Tools at hand



Statistical framework

- Statistical clustering (*k*-means,...)
- Hierarchical clustering (simple-link,...)
- Classification context (*x*DA, kernel-based,...)
- Information-theoretic processing
  - Coding theory
  - Sampling theory

Essentially, view the collection of items as a set of realizations of a random variable with given (estimated) pdf

- Requires a model (prior)
- Smooth out peculiarities

Statistical framework: clustering

- *k*-means clustering
- View *k* centers
- Develop interactively

\+ Simple, fast

\- Requires a model (prior)

\- Rough approximation, smooth out peculiarities

\- Not intuitive (no overlap of clusters)

\- Not adaptive

Statistical framework: agglomeration

Hierarchical clustering

- Complete-/Single-link clustering

⇒ Dendrogram

\+ Choice of the coarseness

\+ Helps exhibiting global structures of the collection

\- Computationally costly

\- May not behave robustly

Statistical framework: classification

- Learn *n* classes
- Sample according to classification (SV,...)

\+ Powerful (semantic)

\- Requires (consistent/large) learning data

\- Requires tuning of the technique

Statistical framework: Information

Preserve the collection entropy

- Similar to data coding

\+ Theoretically sound

\+ Implicit evaluation

\- Smooth out outliers

\- Approximation in practice (P(I)=?)

Discrete framework

The collection is a set of **discrete** instances in a given (possibly **non-metric**) high-dimensional feature space

⇒ Use optimal structures to characterize properties of the collection

- Minimum spanning tree:
  - Indicator of minimal proximity relationship
  - Mappable in 2D
- Set cover
  - Indicator of global item span
- Tours (eg TSP)
  - From nD to 1D
- ...

Discrete sampling

Optimal cover of image feature space

Given a "coverage power" for each image, find the minimum set of images that "covers" the complete collection

+ Highlights outliers

- Coverage power (zone of influence)?

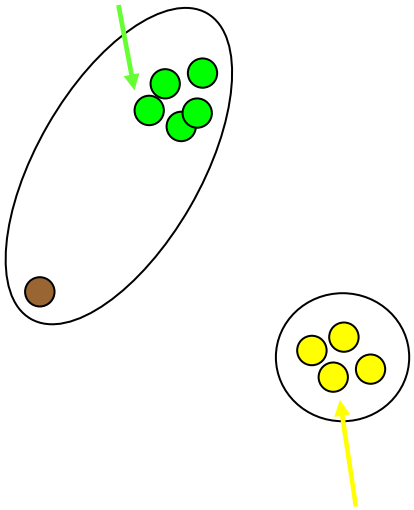
  ⇒ Way of interaction !

- Computational load

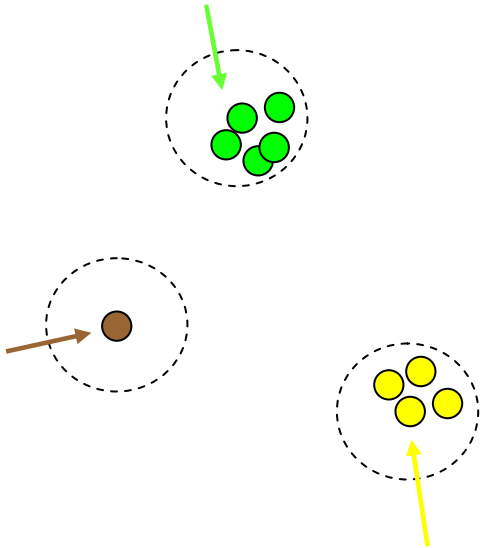
  - Set cover has one nice approximation: greedy cover!

Discrete sampling (example)

_k_-Means (_k_=2)



Set cover



Collection *visit*: formulation

To map the collection on a “near 1D” path so as to visualize it coherently

(like a guide in a museum!)

Visit a number of sites so as to minimize the sum of inter-distances

⇒ Euclidean TSP (NP-complete)

⇒ Suboptimal solution from MST+DFS

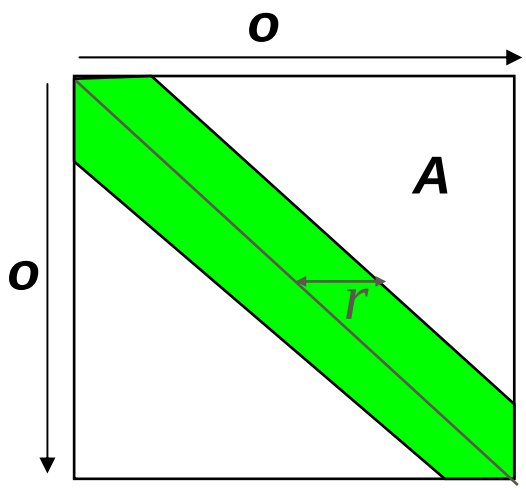
⇒ Incremental (interactive) solution from Lin-Kernigan heuristic

Collection visit: Formulation

Alternative formulation: “Wide band” TSP

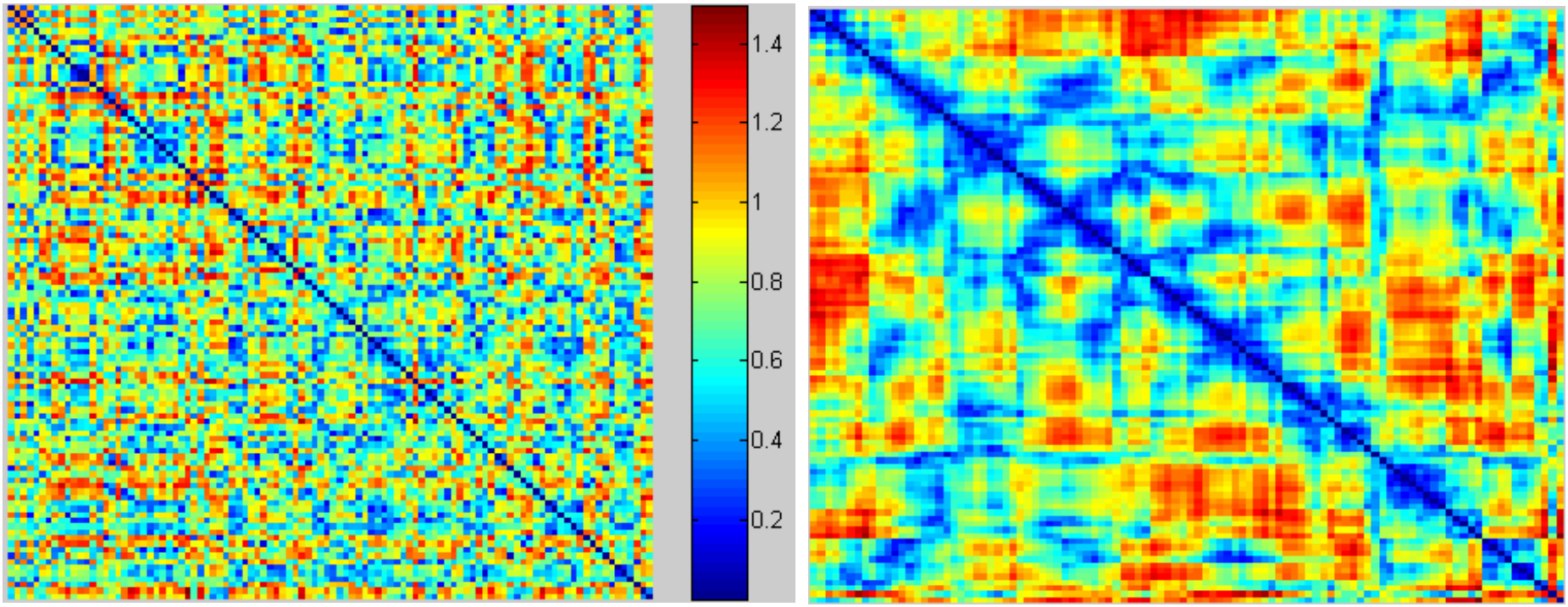
Reorganize the matrix of inter-distances so that the sum of r diagonals is minimum

$$o^* = \underset{o}{\operatorname{argmin}} \Sigma_{i,|k<r|} A_{ii+k}$$



- Empirically, if $r=1$ (classical TSP) we obtain a good approximation
- Solution for $r>1$ desirable!

Example



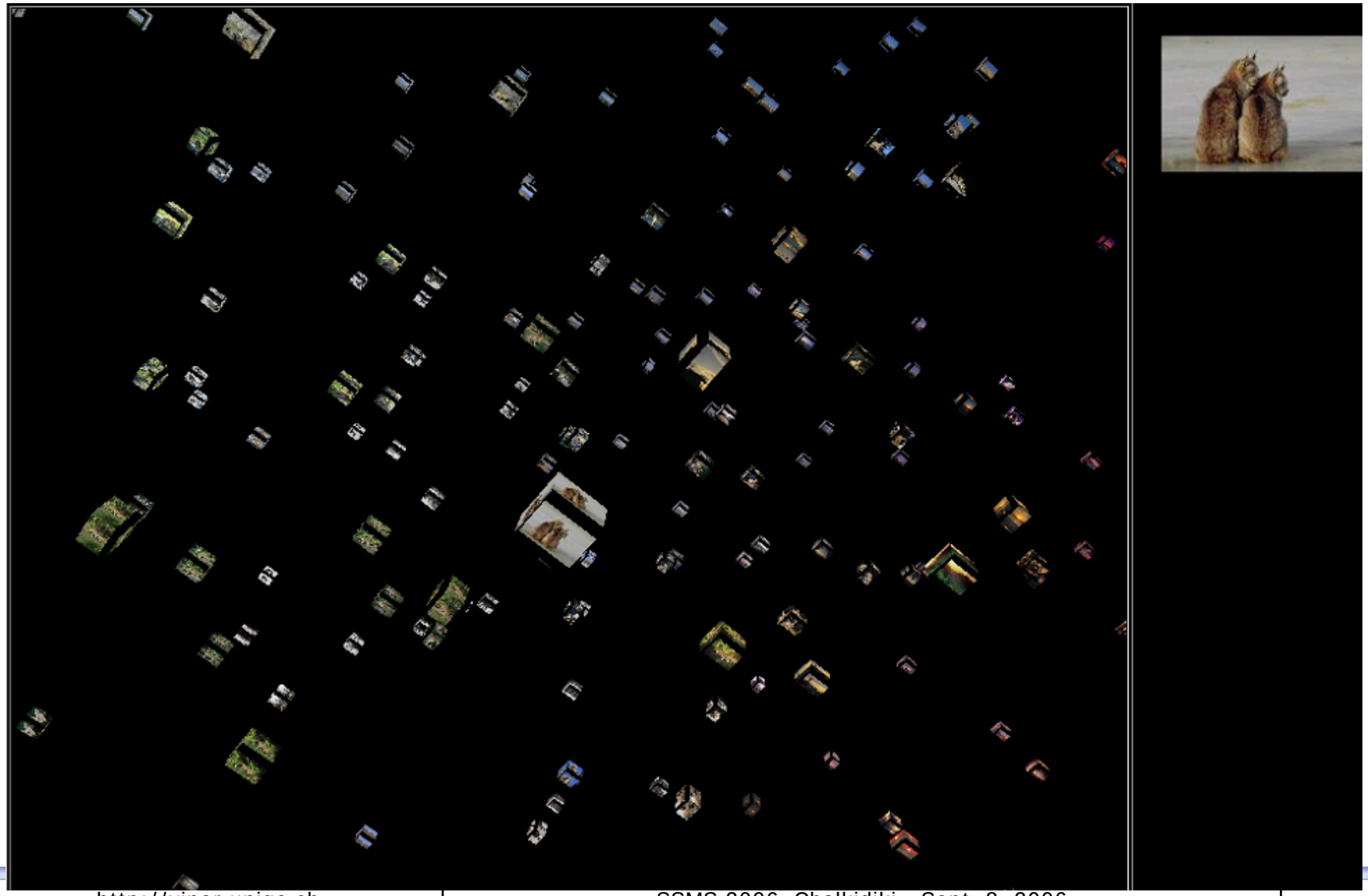
Dimension reduction

Idea: Preserve inter-item distances in both spaces

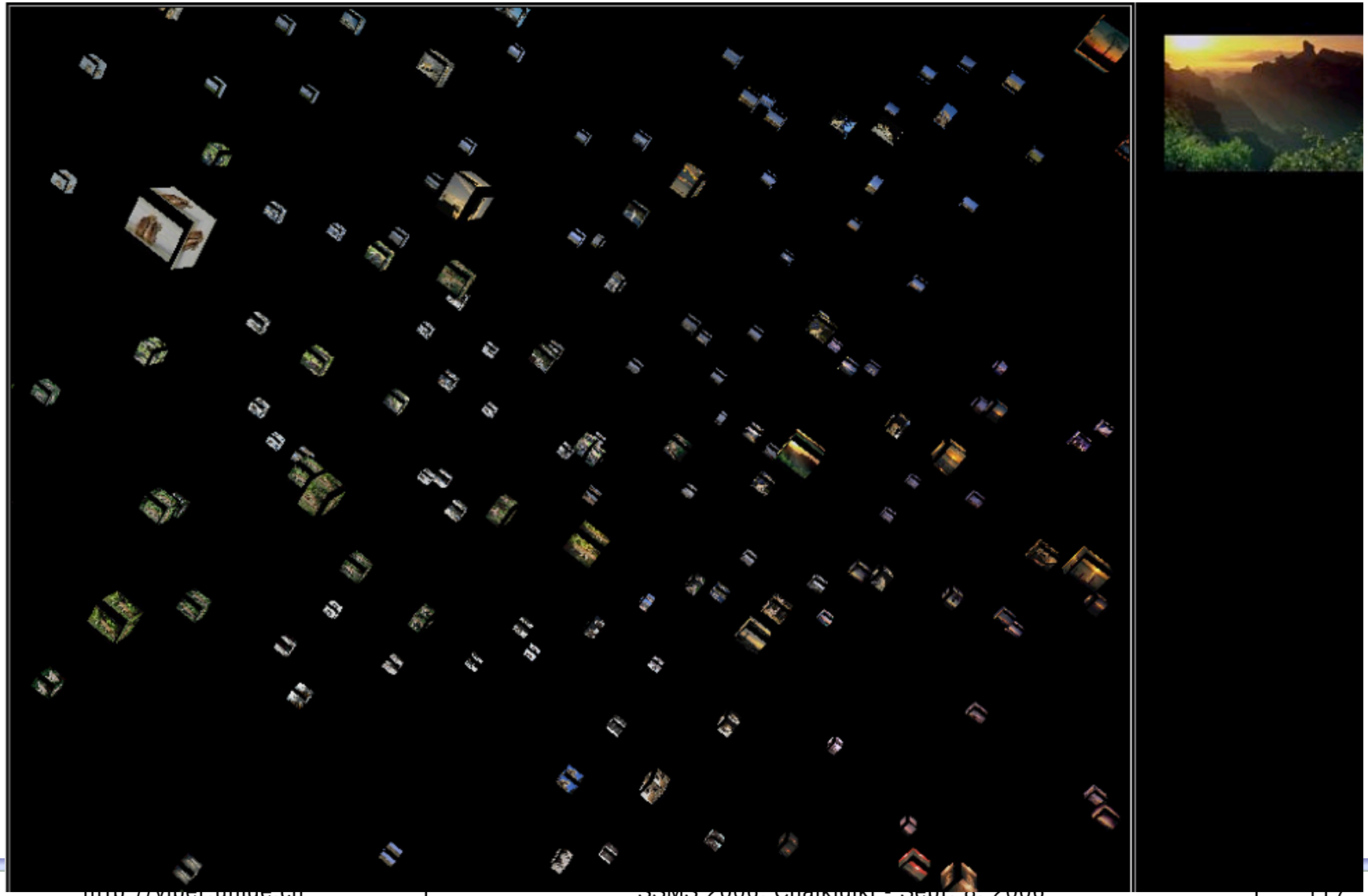
- Typical solutions
  - Multi-Dimensional Scaling
  - Sammon (NLM) mapping
  - CCA (from Self Org. Maps)
  - IsoMap (Geodesic distances)
  - LLE (local linear approximation)
  - FDP (physical model)
  - Relational Perspective Map (geometrical model)

Visual Collection management (raw)

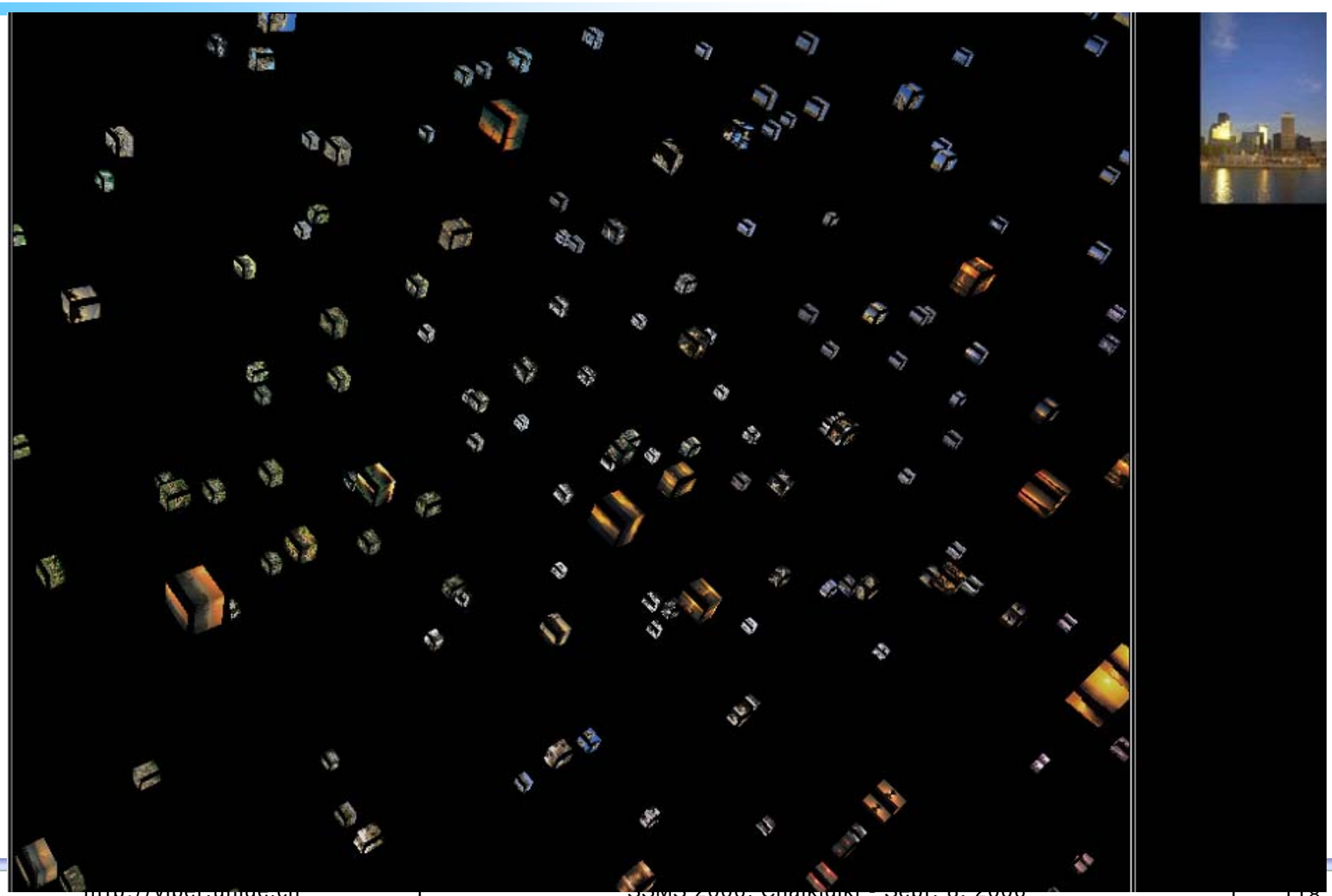
Visual Collection management (3D)



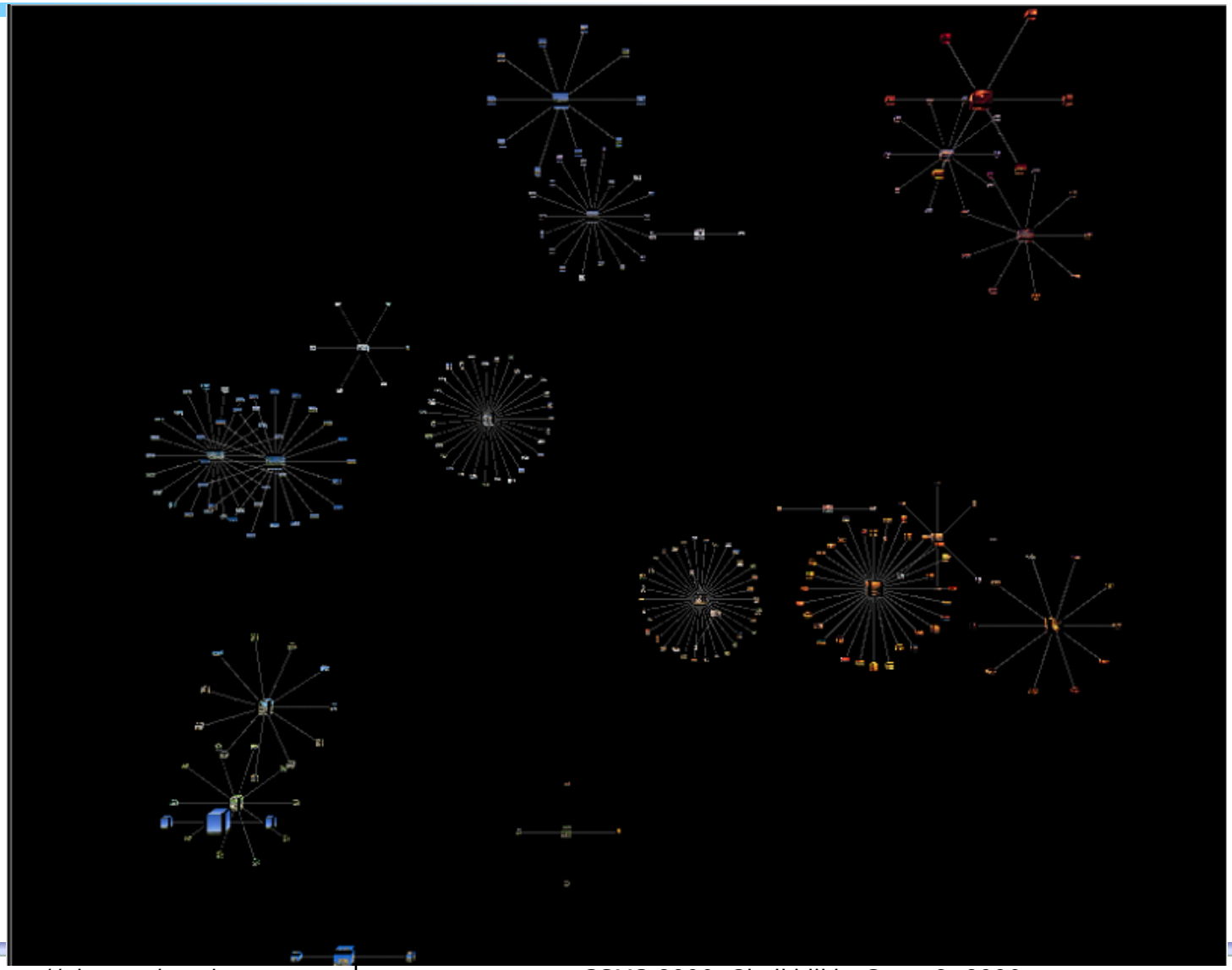
Visual Collection management (3D)



Visual Collection management (3D)

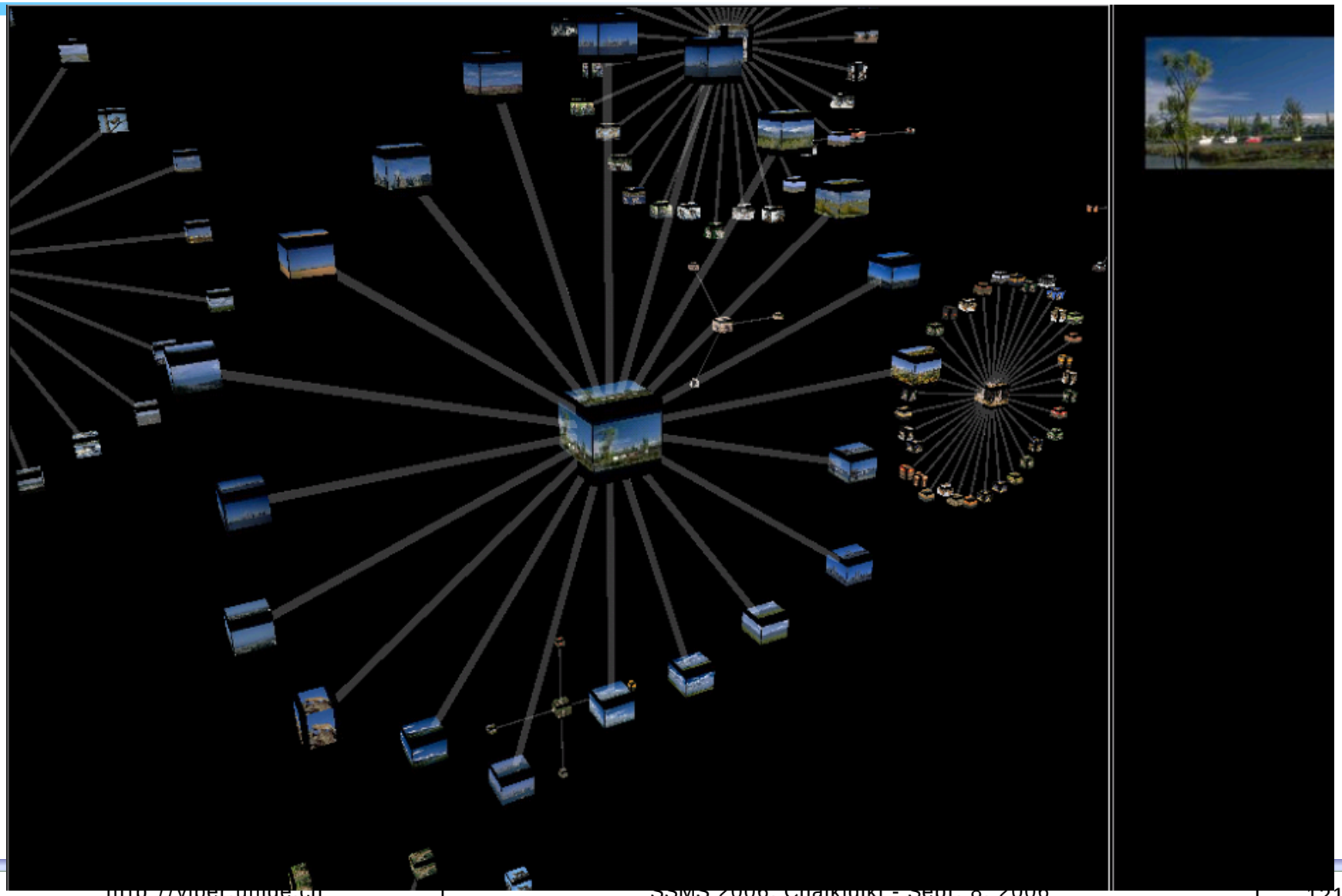


Visual Collection management (clusters)

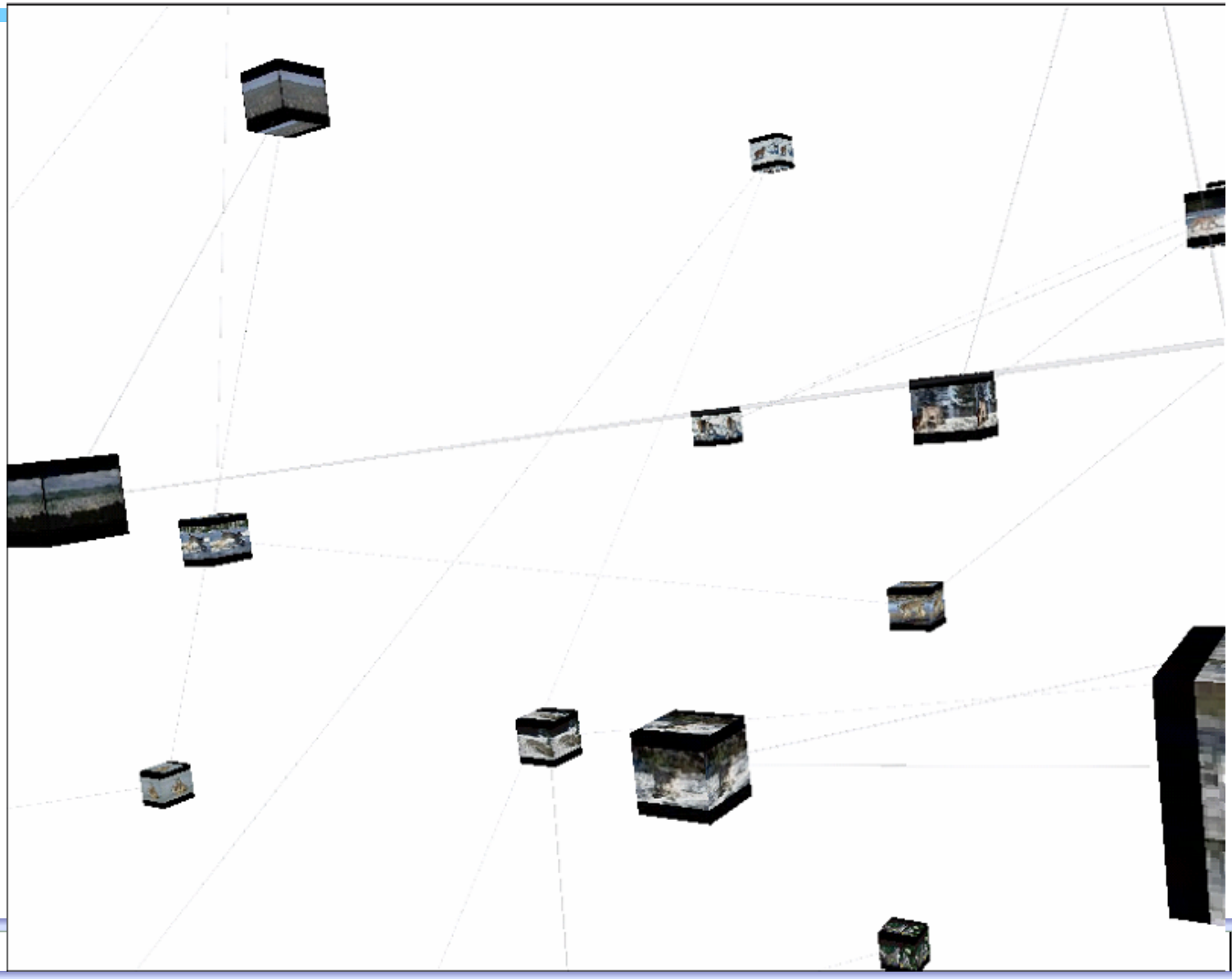


Visual Collection management (clusters)

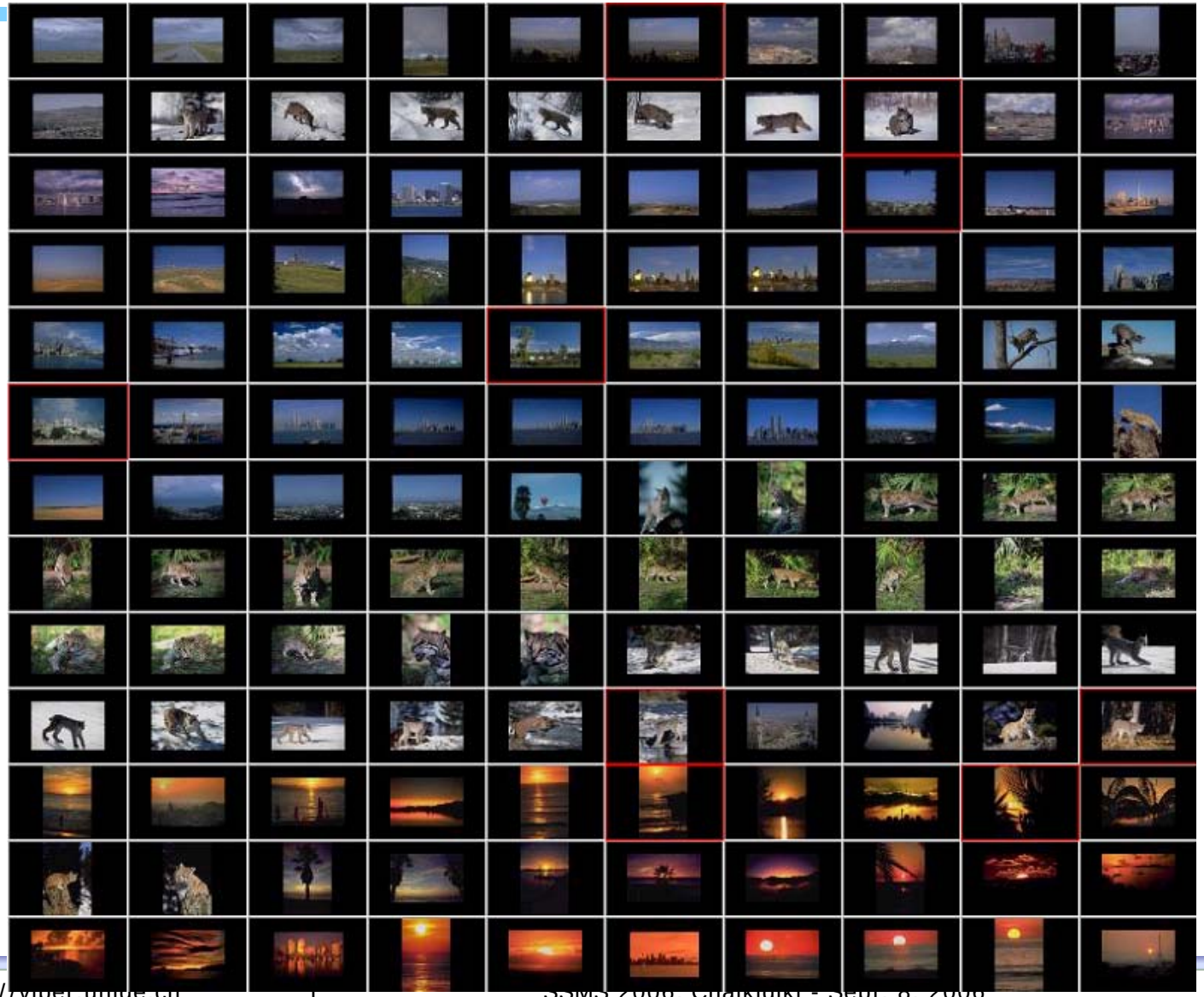
Visual Collection management (clusters)



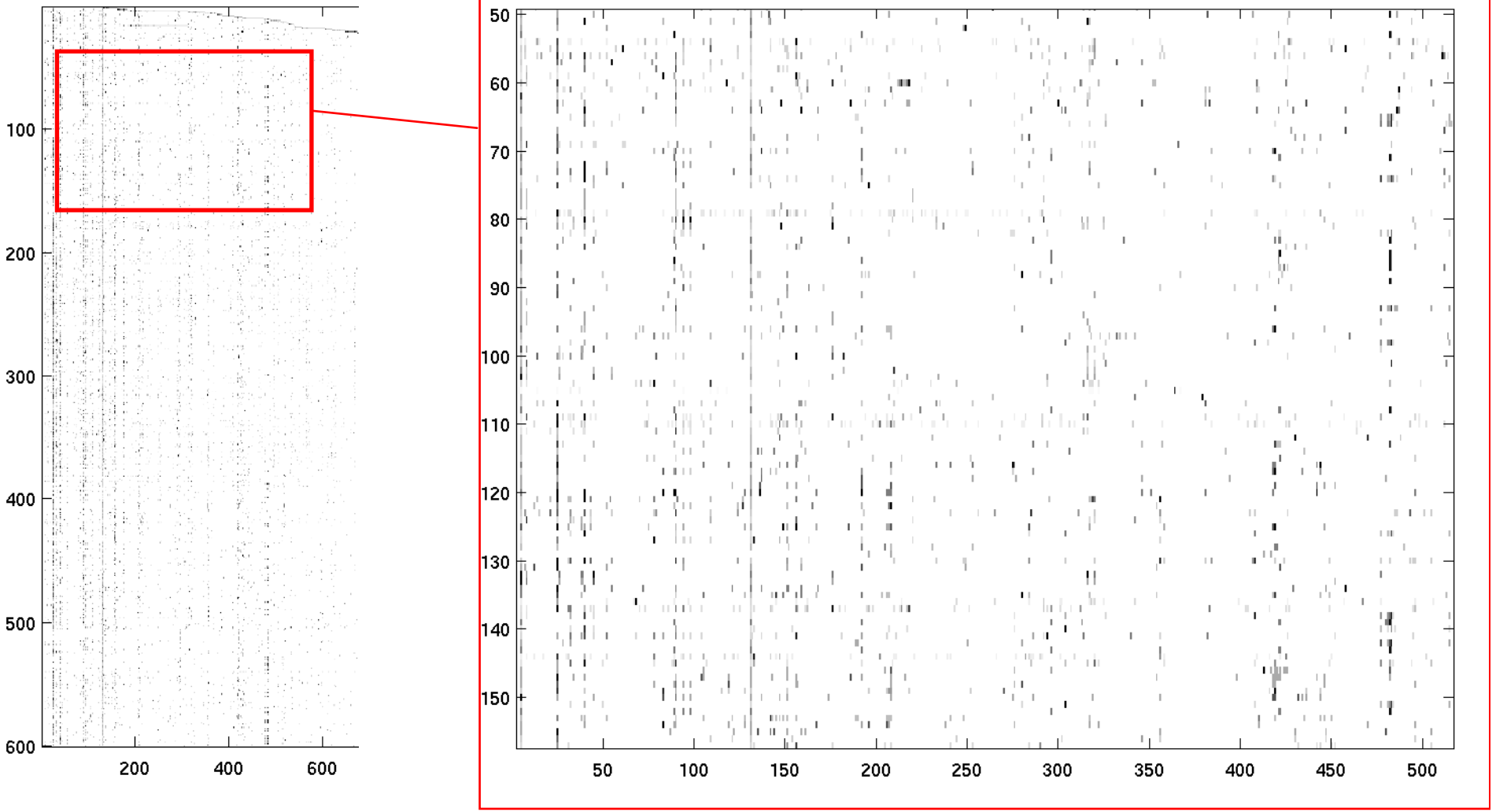
Visual Collection management (exploration)



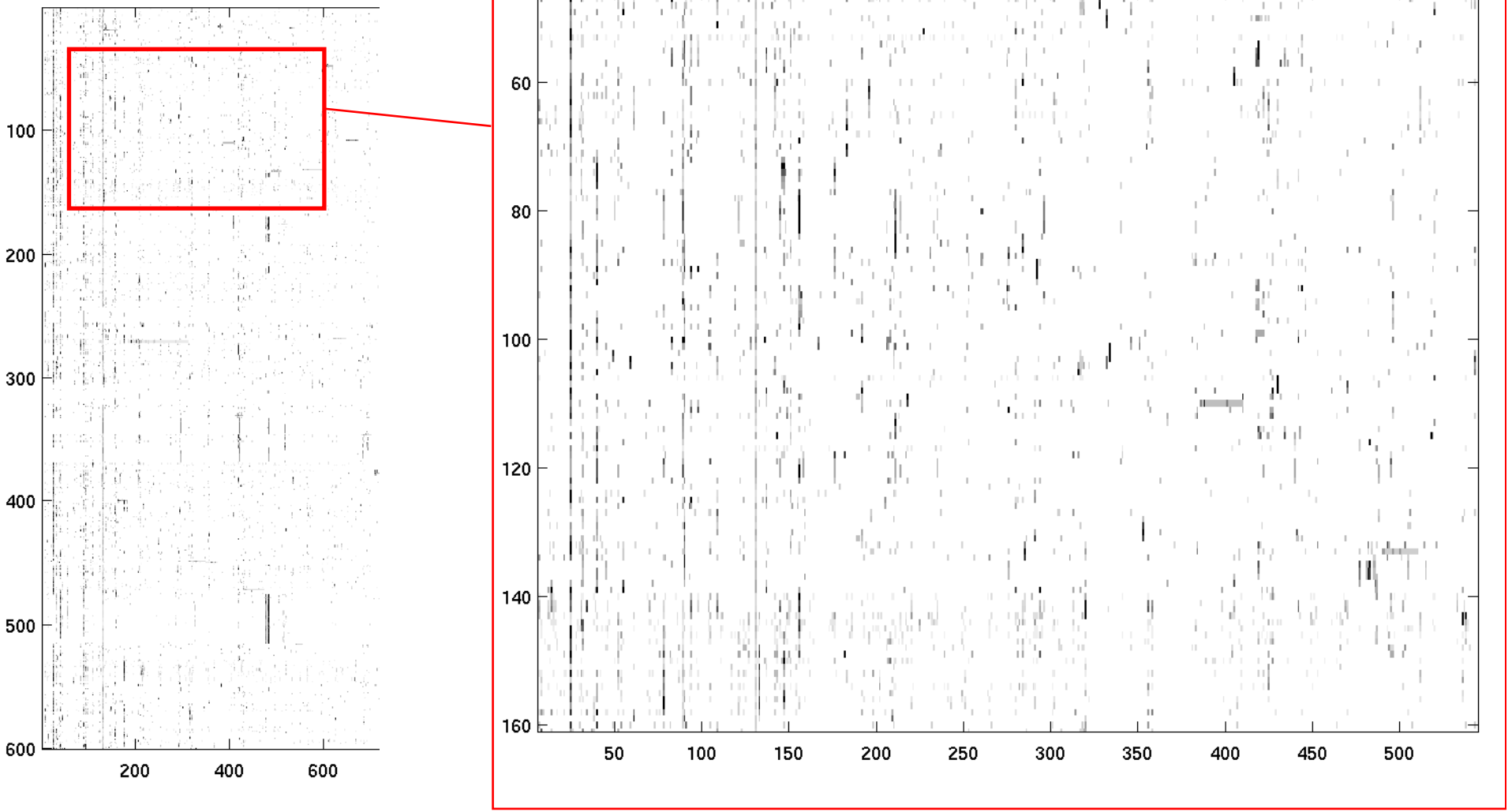
Visual Collection management (organised)



Text collection management



Text collection management



A note on evaluation

- Evaluating the interactive aspect of a system is complex
  - Subjectivity of perception
  - Lack of objective generic criteria
  - Lack of automated measure
  - ⇒ User tests

However, it is critical!

- Some automated evaluation made be performed
  - Fit to the model
    - Statistical properties
    - Legal approximations of complex models
  - Cross model measurement
    1. Defining a solution to a model based of principle A
    2. Testing this solution against principle B
    3. Find a tradeoff between optimality w.r.t A and B

Objectives (fulfilled?)

This course aims at emphasizing:

- The need for a formal management model
- The complexity of multimodal fusion
- The *complementarity* of several approaches
- The need for new interaction paradigms

Summary

- Multimodal information is complex as
  - A large volume of data to handle
  - Several temporal streams to keep synchronized
  - Several modes, each carrying its own (reliable) information
  - ...
- Multimodal information is rich as carrier
  - Semantic categorization
  - Story segmentation
  - ...
- Information retrieval models need to be adapted
  - Classical text-based IR models are not sufficient
  - Fusion may be performed in several ways and in many places
  - New trend: learning-based IR

References

(Barnard, 2003) Kobus Barnard, Pinar Duygulu, Nando de Freitas, David Forsyth, David Blei, and Michael I. Jordan, "Matching Words and Pictures," Journal of Machine Learning Research, Vol 3, pp 1107-1135, 2003

(Bruno, 2006) E Bruno and N. Moenne-Loccoz and S. Marchand-Maillet, Asymmetric Learning and Dissimilarity Spaces for Content-based Retrieval, in International Conference on Image and Video Retrieval (CIVR 2006), Tempe, AZ, 2006.

(Cox, 1996) I. J. Cox, M. L. Miller, S. M. Omohundro and P. N. Yianilos, PicHunter: Bayesian Relevance Feedback for Image Retrieval, Proc. Int. Conf. on Pattern Recognition, Vienna, Austria, C:361-369, August 1996.

(Craver, 1999) S.A. Craver, B-L. Yeo, and M. Yeung, "Multi-Linearization Data Structure for Browsing." In Proceedings of SPIE, Storage and Retrieval of Image and Video Databases, San Jose: CA: January 1999. 155-166.

(Cutting, 1992) Douglass R. Cutting, David R. Karger, Jan O. Pedersen, John W. Tukey , Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. ACM SIGIR Conference on Research and Development in Information Retrieval, 1992.

(Kosinov, 2004) S. Kosinov and S. Marchand-Maillet, Hierarchical ensemble learning for multimedia categorisation and auto-annotation, in IEEE Machine Learning for Signal Processing, Sao Luis, Brazil, September, 2004.

(Kumar, 2002) Kumar, S., Ghosh, J., and Crawford, M. M. Hierarchical fusion of multiple classifiers for hyperspectral data analysis. Pattern Analysis and Applications 5 (2002), 210–220.

(Janvier, 2005) B. Janvier and N. Moenne-Loccoz and S. Marchand Maillet and T. Pun, A contextual model for semantic video structuring, in 13th European Signal Processing Conference, EUSIPCO'05, Antalya, Turkey, September, 2005.

(Marchand-Maillet, 2005) S. Marchand-Maillet and E. Bruno, Collection Guiding: A new framework for handling large multimedia collections, in First Workshop on Audio-visual Content And Information Visualization In Digital Librairies, AVIVDiLib05, Cortona, Italy, May, 2005,

(Moenne-Loocoz, 2005) N. Moenne-Loccoz and E. Bruno and S. Marchand-Maillet, Knowledge-based Detection of Events in Video Streams from Salient Regions of Activities, in Pattern Analysis and Applications (PAA), special issue Video Based Event Detection, 2005.

(Moenne-Loocoz, 2004) N. Moenne-Loccoz and B. Janvier and E. Bruno and S. Marchand-Maillet, Managing Video at Large, in ACM SIGMOD Workshop on Computer Vision meets Databases, Paris, France, June, 2004.

(Monay, 2004) F. Monay and D. Gatica-Perez, " PLSA-based Image Auto-Annotation: Constraining the Latent Space", in Proc. ACM Int. Conf. on Multimedia (ACM MM), New York, Oct. 2004.

(Platt, 1999) Platt, J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In Advances in Large Margin Classifiers, A. Smola, P. Bartlett, B. Scholkopf, and D. Schurmans, Eds. MIT Press, 1999.

(Rubner, 2000) Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision, 40(2):99-121, November 2000

More information

- Viper group: http://viper.unige.ch
- Video retrieval framework (inc Demo):
- http://viper.unige.ch/vicode
- IR course at University of Geneva

http://dokeos.unige.ch/courses/1843/