“Multimedia Information Retrieval”

Stephane Marchand-Maillet

University of Geneva
Viper Group
http://viper.unige.ch
Stephane.Marchand-Maillet@unige.ch

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Agenda

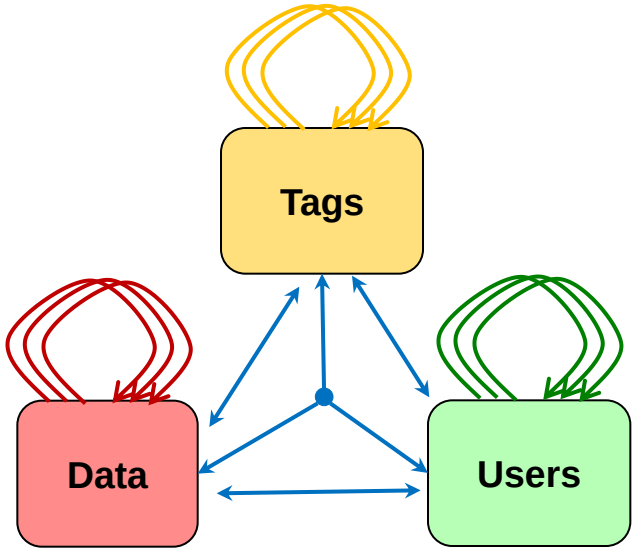
- What is multimedia?
  - Networked Media
  - Use cases of Multimedia Information Retrieval
- Challenges in Multimedia IR
  - Volume of data
  - Interpretation of content
- Multimodal Fusion for Multimedia IR
  - Understanding Multimodal Fusion
  - Multimodal Fusion Models
  - Interactive Multimodal Fusion for Multimedia Retrieval
- Wisdom of crowds
  - Mining search logs
- Look Ahead

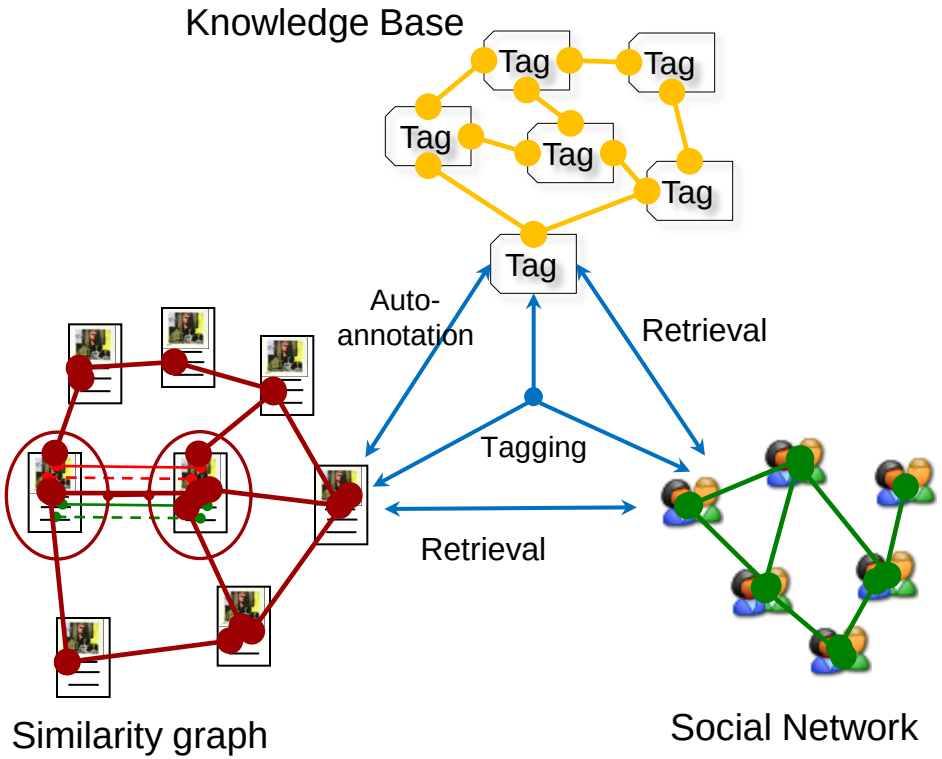
What is multimedia?

- Text
  - Plain text (any language)
  - Structured text (XML-like, code,…)
- Visual
  - Images (Photo)
  - Sketches (drawing, map,..)
- Audio
  - Music
  - Speech
  - Sound
- Misc
  - 3D object
  - Video
  - … (software, playlist,…)

+ Any combination
  - PDF, Web pages and alike

Social media context and model





Multimedia content analysis

1 media item

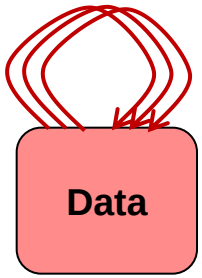
- Content abstraction: Giving more semantic to content
  - Object/person/scene detection/localisation/recognition
  - Audio/speech segmentation/recognition
  - …

2+ media items

- Content similarity: how much can an item be taken for another
  - Feature-based distances
  - Feature selection, learning
  - …

Complex media items

- Multimodal fusion
- Scalability



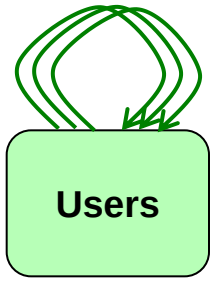
Add-on: Social Network Analysis

"Limited" group of users: *eg*, meeting

- Mining patterns of interaction
  - Role recognition
  - …

"Large" groups of users

- Analysing connectivity
  - Identifying communities
  - "Small world" effect
  - …
- Analysing organisation
  - Authority, thrustworthiness, …



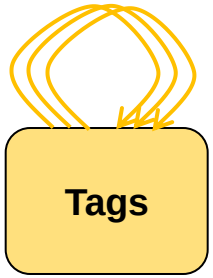
Prior: Information management

Lexical database

- Sense disambiguation, management of synonymy, polysemy
- Word distance computation

Ontology

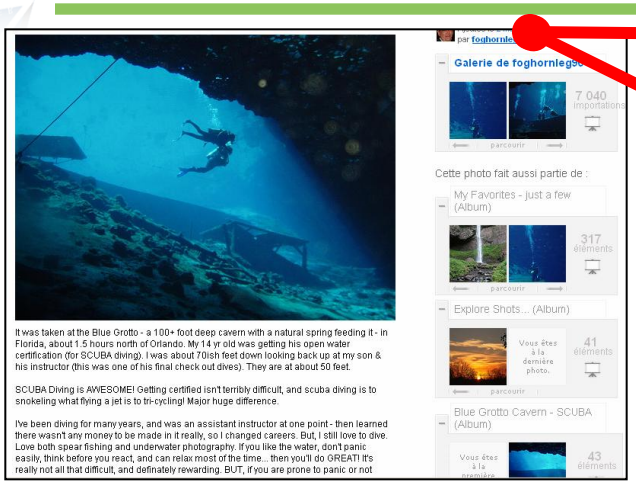
- Knowledge representation and inference



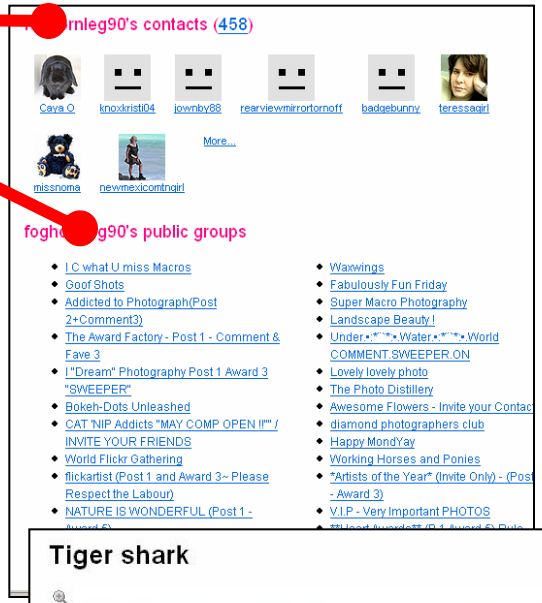
Folksonomy

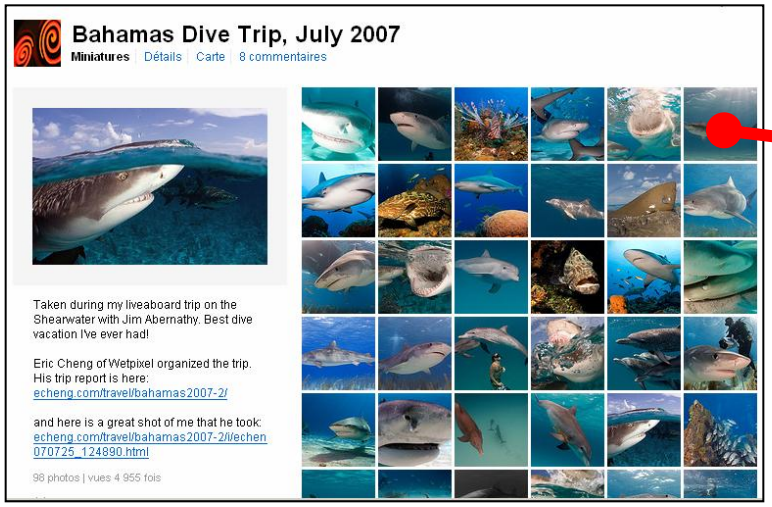
- Decentralised knowledge management

Networked multimodal data

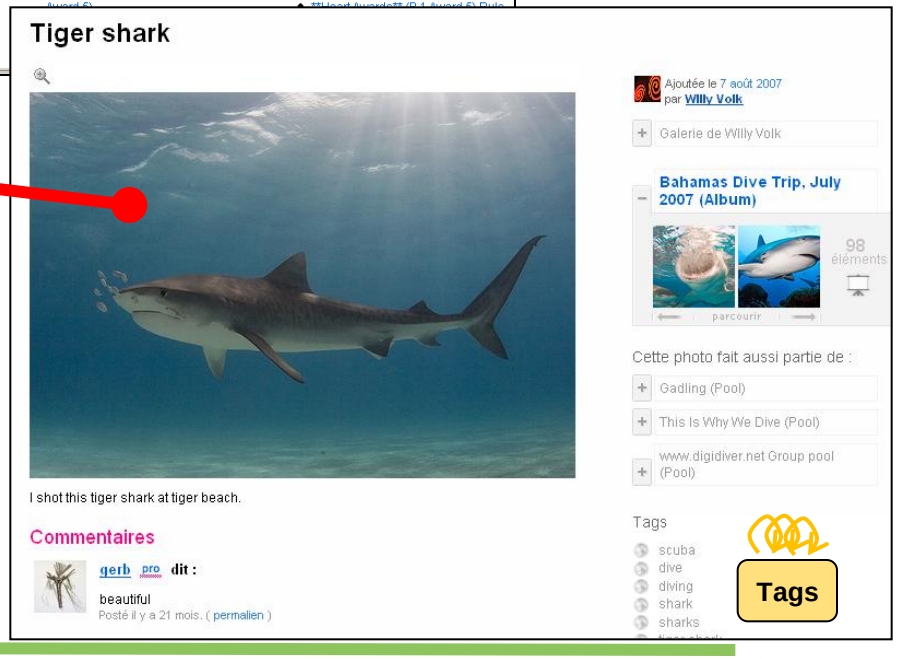


Users





Data



Tags

Multimedia Information Retrieval: Use Cases

- Image
  - Find look-alike pictures
- Video
  - Find specific actions
- Music
  - Find inspirational music
- Medical
  - Find similar cases
- Patent
  - Find similar proposals

➔ Not always doable with text

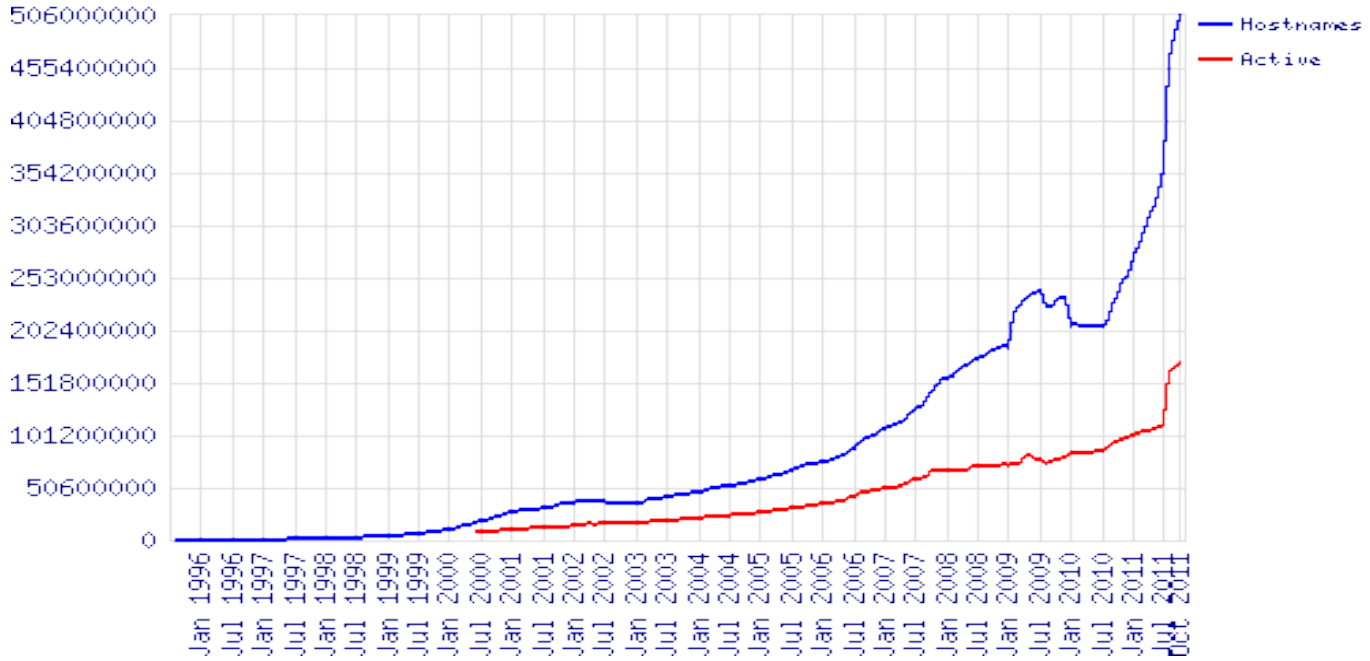
Agenda

- What is multimedia?
  - Networked Media
  - Use cases of Multimedia Information Retrieval
- Challenges in Multimedia IR
  - Volume of data
  - Interpretation of content
- Multimodal Fusion for Multimedia IR
  - Understanding Multimodal Fusion
  - Multimodal Fusion Models
  - Interactive Multimodal Fusion for Multimedia Retrieval
- Wisdom of crowds
  - Mining search logs
- Look Ahead

Volume of multimedia data

Sources:
R. Baeze-Yates – RussIR 2010
http://news.netcraft.com/

- 2 billion persons connected to internet
  - Mostly via mobile devices (phones)
- 1.8 billions active mobile phones
- 180 millions active web servers



Volume of multimedia data

- Web scale:
  - Google:
    - Early 2004: 4.3 billions indexed pages
    - Early 2005: 8 billions indexed pages
    - Today: 20 billion indexed pages?
  - http://www.archive.org
    - Growth: 20 Tb/month

- Multimedia collection:
  - Facebook: 140 billions photos
    ➔180 years at 25 fps
  - YouTube (2008): 83.4 million videos

- Curated collections:
  - AOL Video library
    - 410 millions views in October 2011
  - Institut National de l'Audiovisuel (INA-France):
    - 700'000h audio (radio)
    - 400'000h video (television)
    - 2 millions documents

From data to information

Interpreting the content

- Data is physically measured
- Information is interpreted semantically

  ⇒**Semantic gap**

- Fusion is a way to reduce the semantic gap
- Examples
  - Looking at a movie w/o audio
  - Listening to a story w/o visual
  - Voice conversation vs video conversation
  - Disambiguation (eg „jaguar“)

Information Retrieval

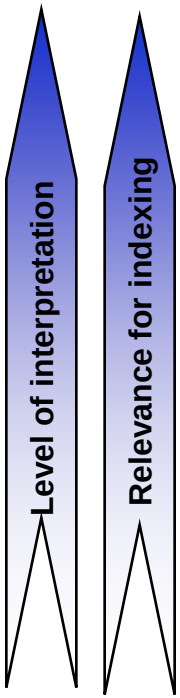
- IR tells us that we do not (really) need to have a complete absolute understanding of documents to respond queries, we (just) need to be able to compare 2 documents
- We merely need appropriate distance measurements
  - To infer similarity
  - Based on document features (space)
  - and a notion of vicinity (distance)

→ Everything is about inspecting neighborhoods
  - Provided the distance is semantically relevant

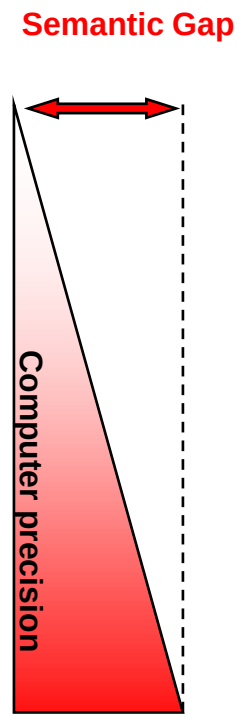
From data to information

Semantic gap:

"Discrepancy between the level of analysis of a computer and the level of perception of the same data by a human user."

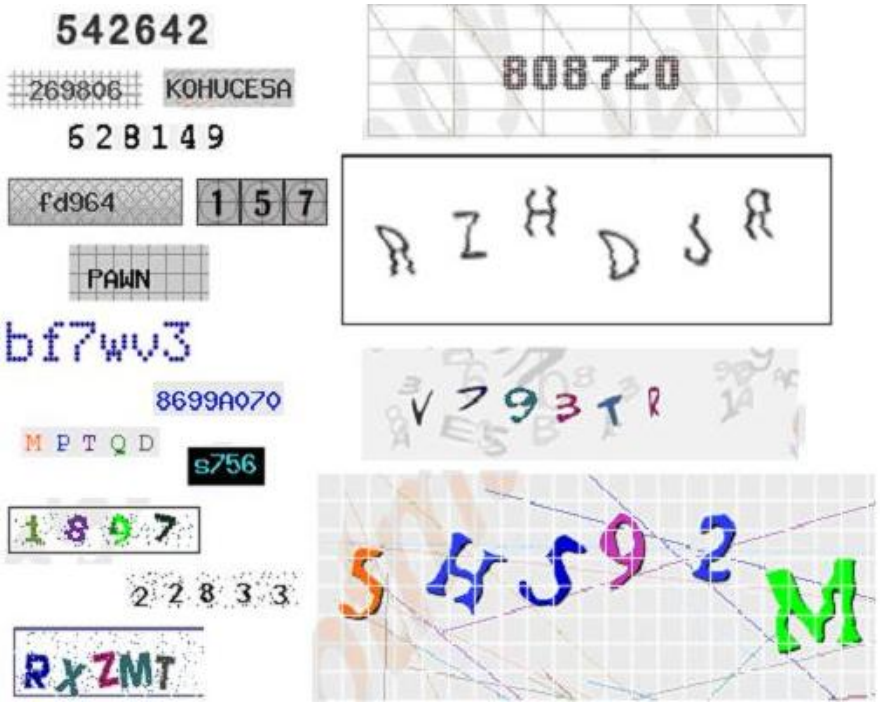


Level	Text	Visual
Interpretation	Story	"Action"
Semantic segmentation	Paragraph	Scene
Group	Sentence	Object
Physical segmentation	Word	Homogeneous region
Physical inspection	Character	Color, texture



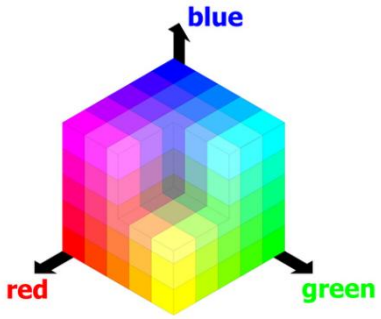
Semantic gap

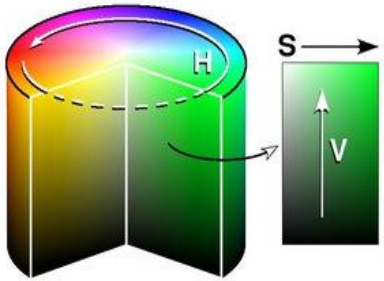
- Exploited by Captchas

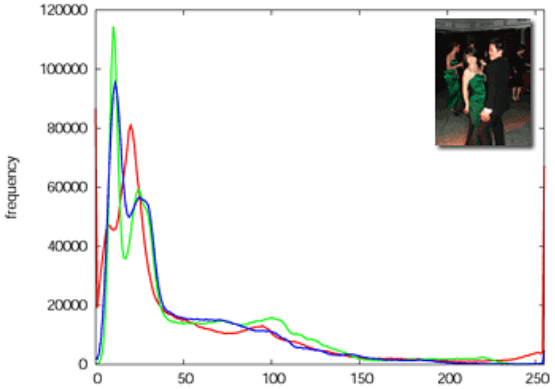


Low-level image representation

Global Color Histogram data in color spaces:

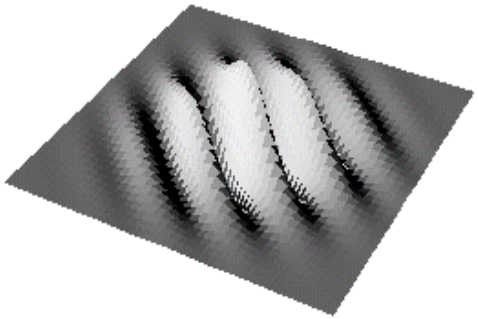


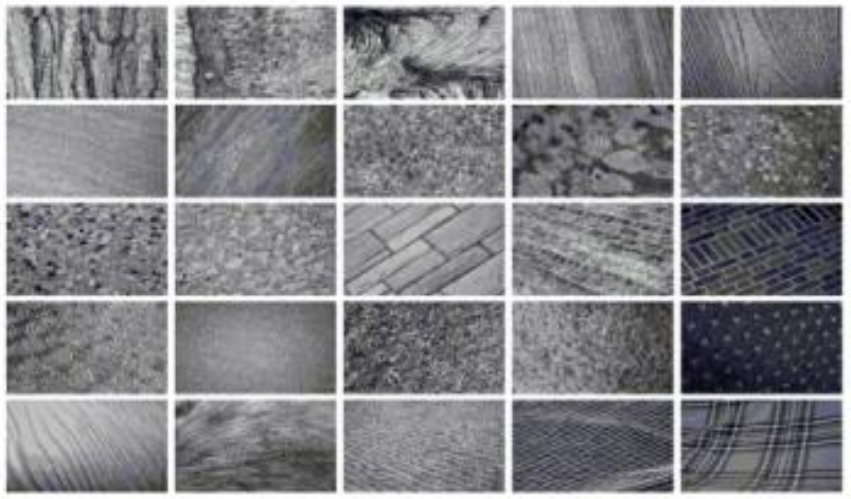




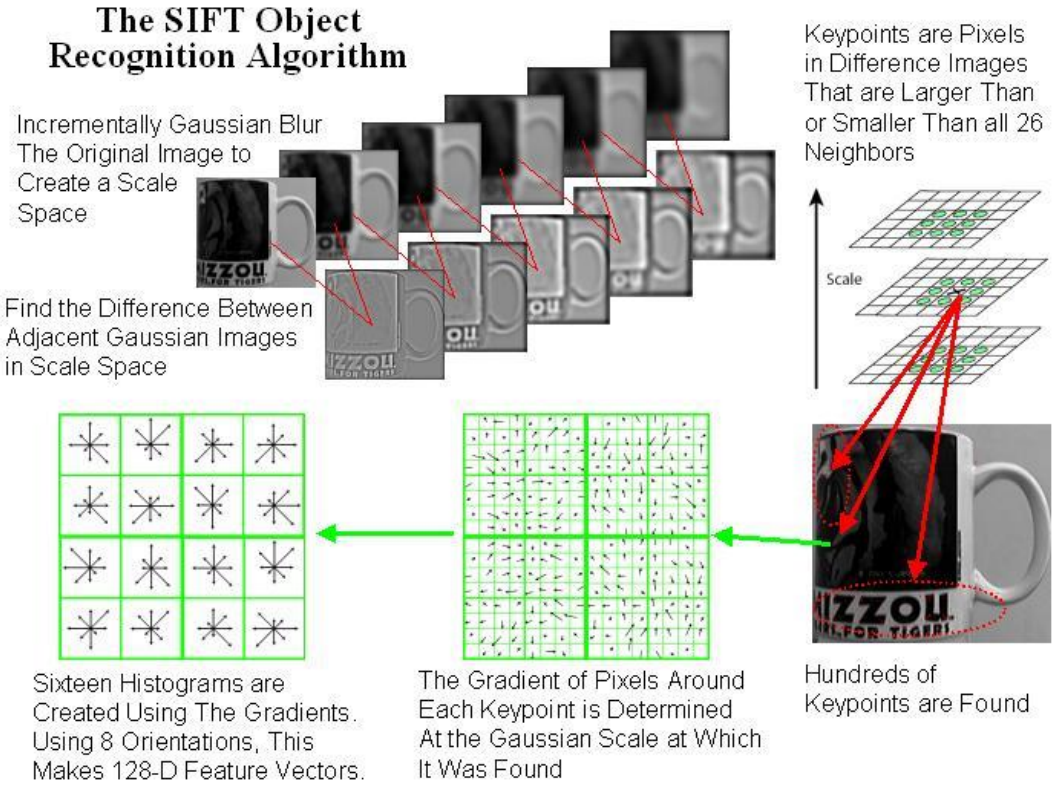
Global Texture data from Gabor filter banks:

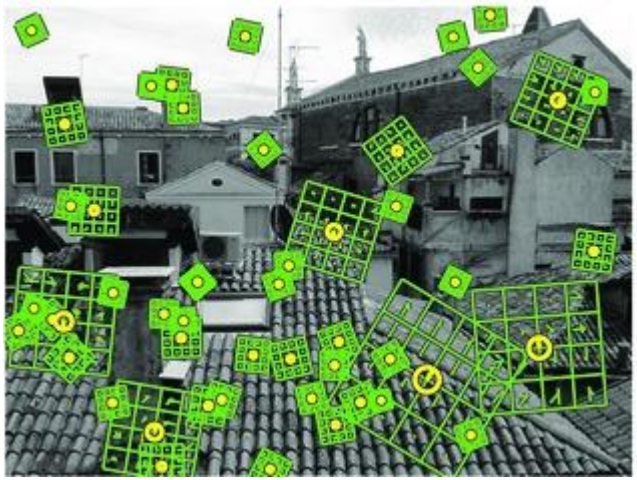
$$g(x,y) = \left(\frac{1}{2\pi\sigma_x\sigma_y}\right)\exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)+2\pi jWx\right)$$

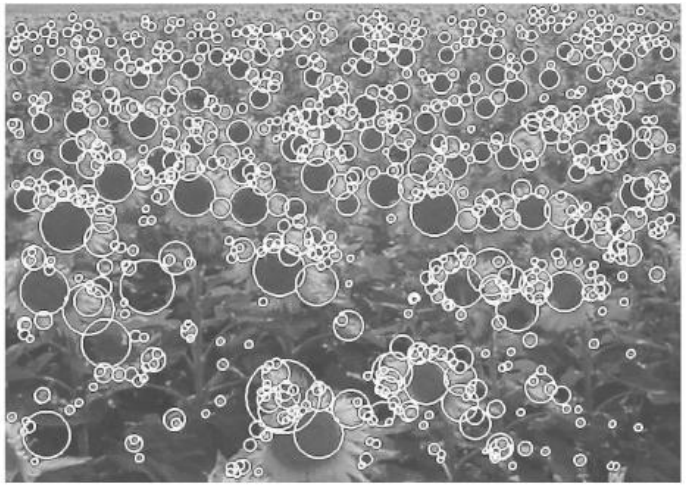




Local features: SIFT[Lowe 1999] and SURF[Bay *et al* 2006]

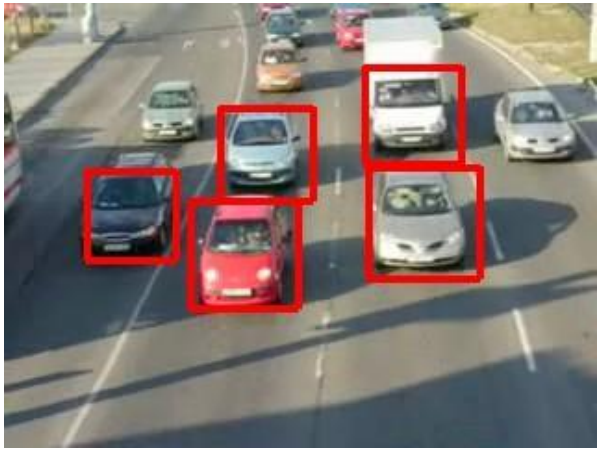


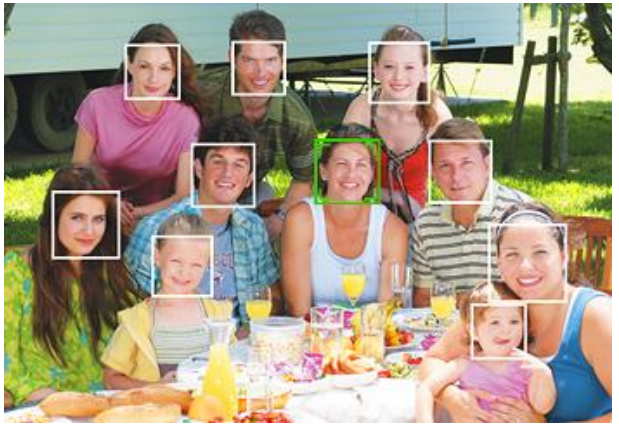




Region-based image understanding

Images are compared w.r.t the regions (hopefully objects/concepts) they contain

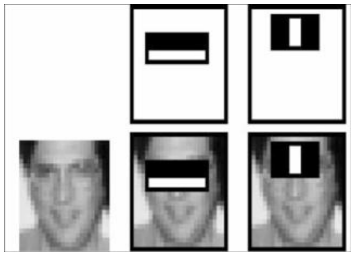




➔Object detection *eg,* Face detection

Object detection

- Train the machine to recognize specific groups of patterns



Faces
[Viola, Jones, 2001]

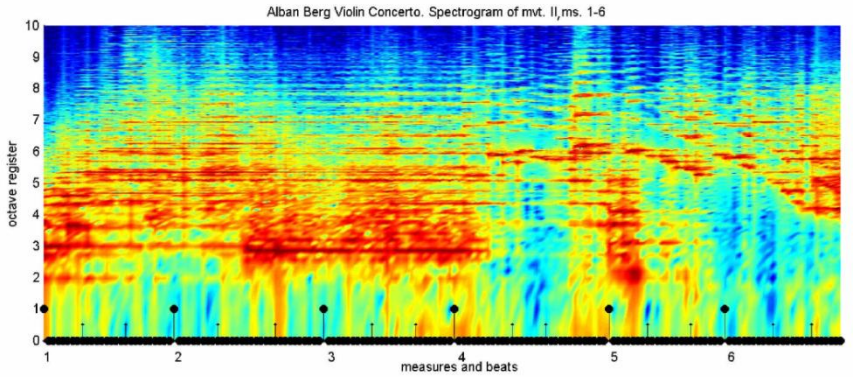
Objects

Naming faces

Specific characteristics of a face are extracted (eg eigenfaces)

Audio features

- Low-level-audio representation
  - Spectrogram
  - MFCC



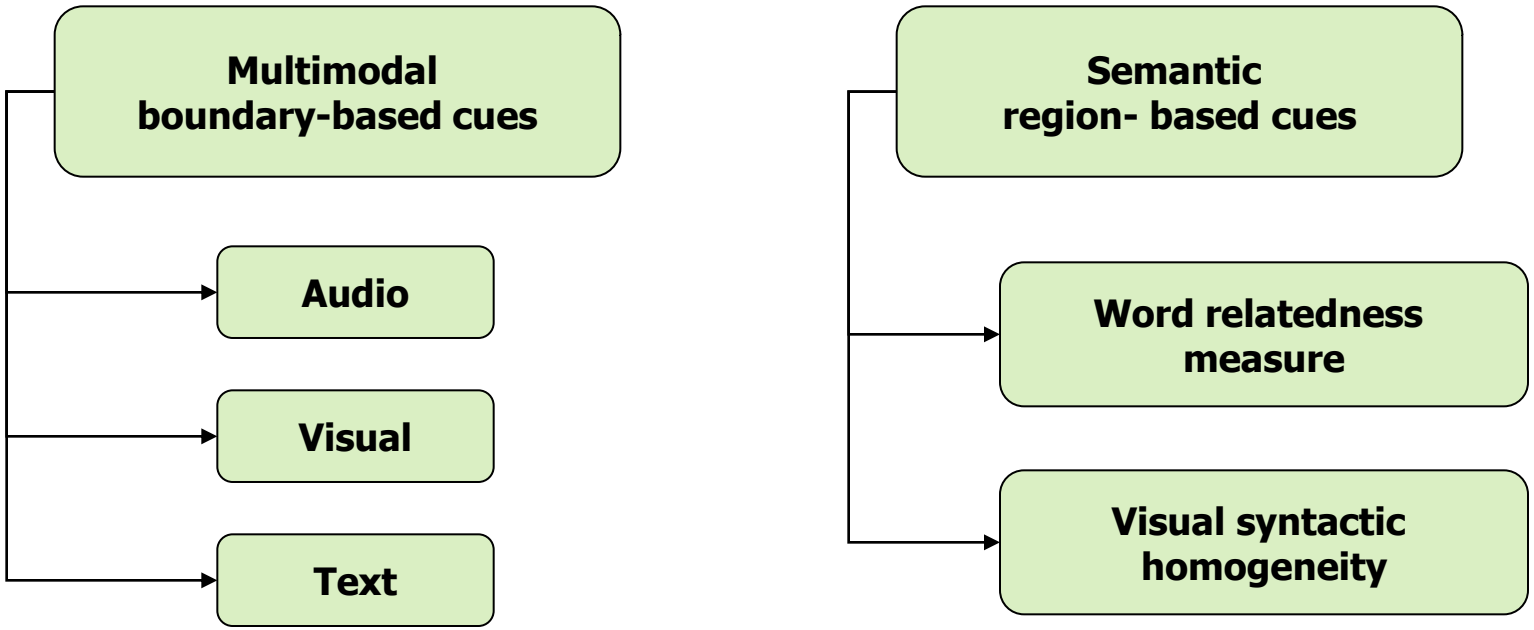
- Mid-level representations
  - MIDI strings
  - LPC

- High-level representation
  - Automatic speech recognition (ASR)

Multimodal video features

Over temporal modalities:

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



Adding context to interpretation

Video Story Segmentation:

Using **fusion** between the visual and audio modalities, high-level segmentation may be achieved

Agenda

- What is multimedia?
  - Networked Media
  - Use cases of Multimedia Information Retrieval
- Challenges in Multimedia IR
  - Volume of data
  - Interpretation of content
- Multimodal Fusion for Multimedia IR
  - Understanding Multimodal Fusion
  - Multimodal Fusion Models
  - Interactive Multimodal Fusion for Multimedia Retrieval
- Wisdom of crowds
  - Mining search logs
- Look Ahead

Main goals of fusion

Multimodal information fusion aims at interpreting jointly multiple sources of information representing the same underlying „concept“

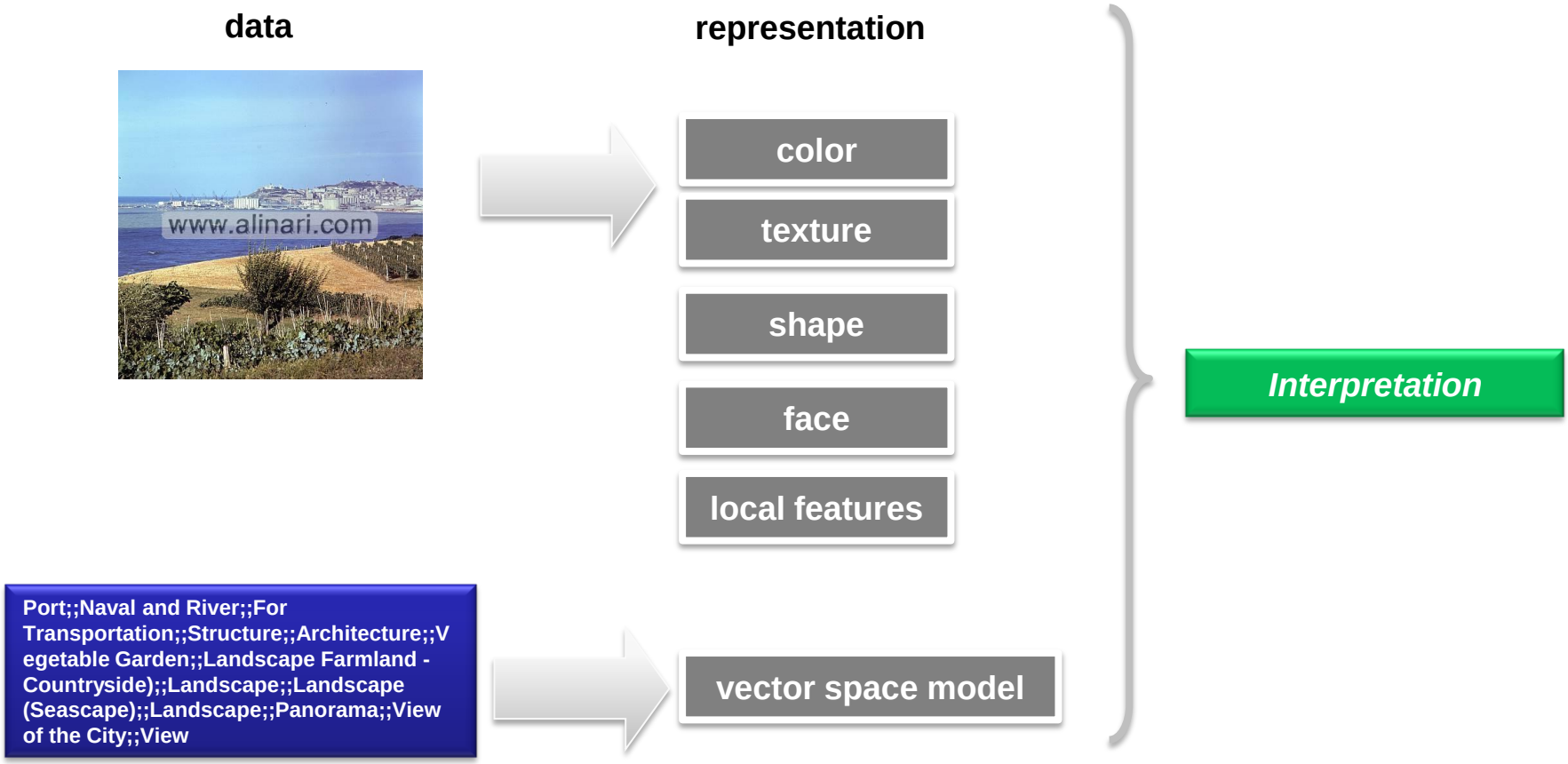
⇒The main goal is the extraction of information

By fusing information, one aims at:

- Being more accurate in the discovery of the „concept“
  - Each individual stream may be incomplete
- Being more robust in the discovery of the „concept“
  - Each individual stream may be distorted (*eg,* noisy)

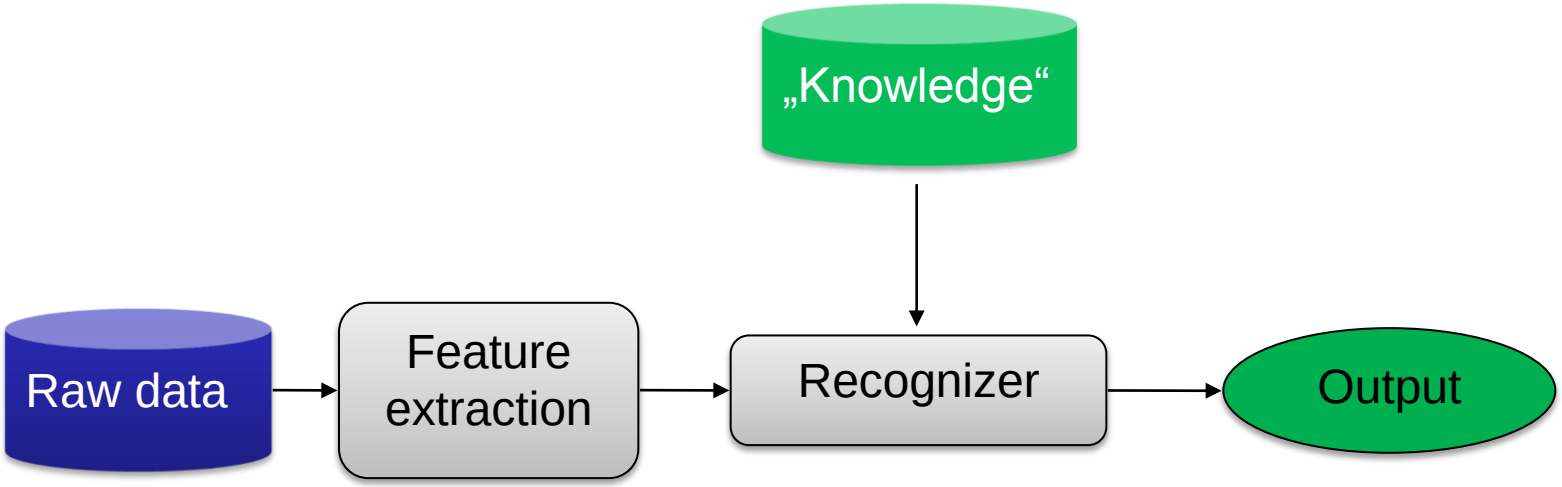
Multimodal fusion

- From many sources of information and context, how to make our best to “interpret” the data

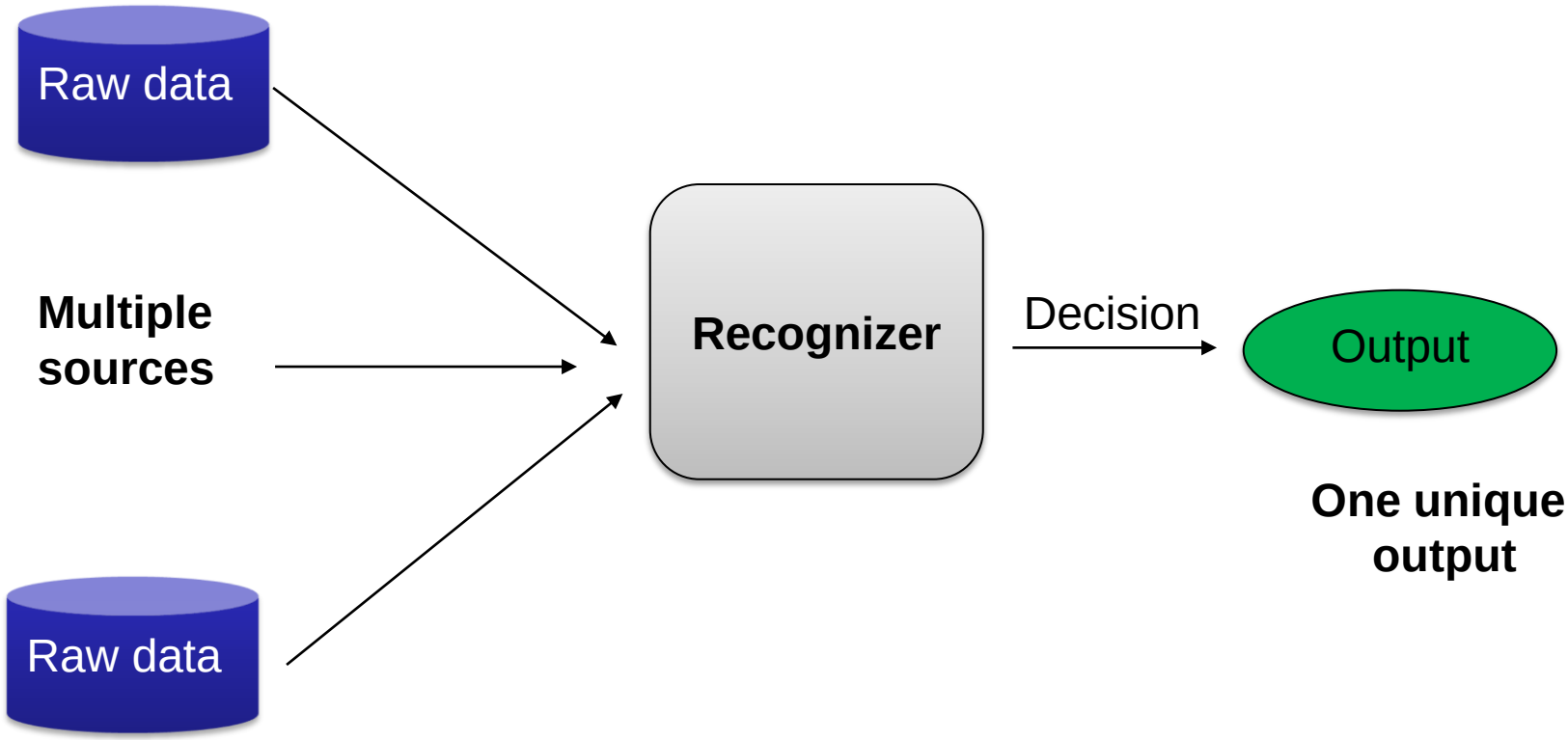


Levels of fusion

How to organise fusion from classical „unimodal“ interpretation?

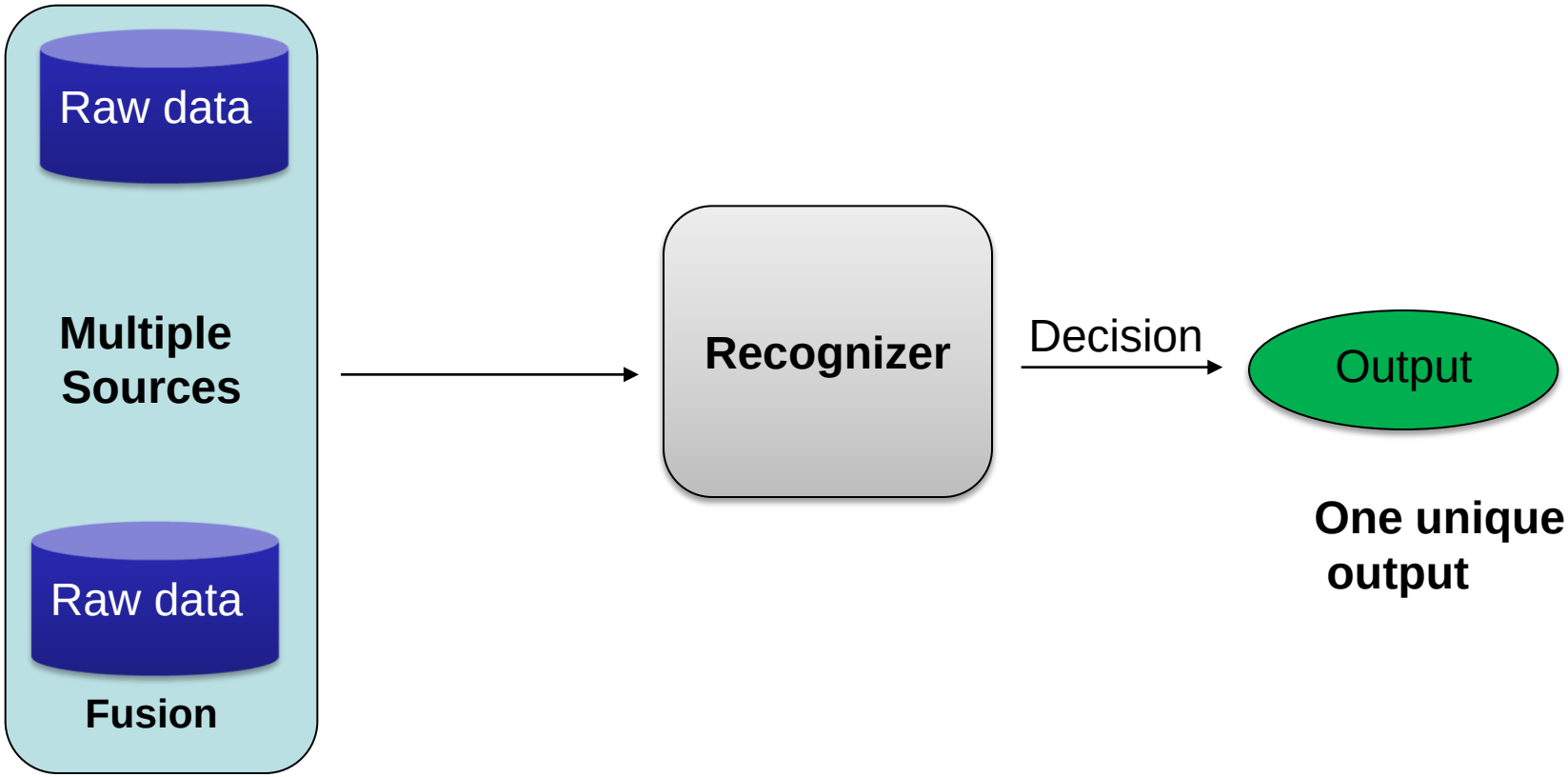


Levels of fusion



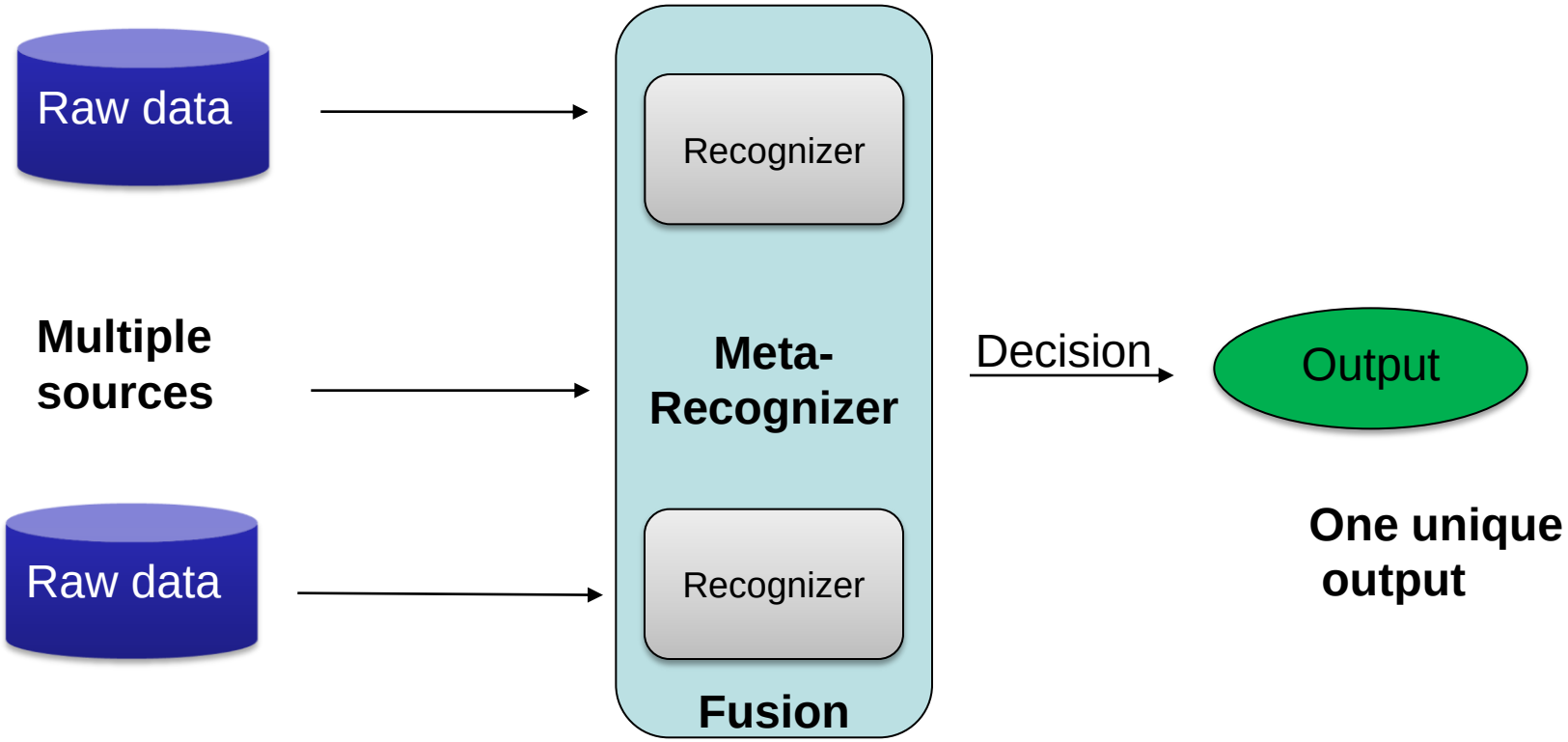
Early fusion strategy

- Acts over features
- All modalities are „concatenated into one“
- Only one decision is taken over the concatenated input



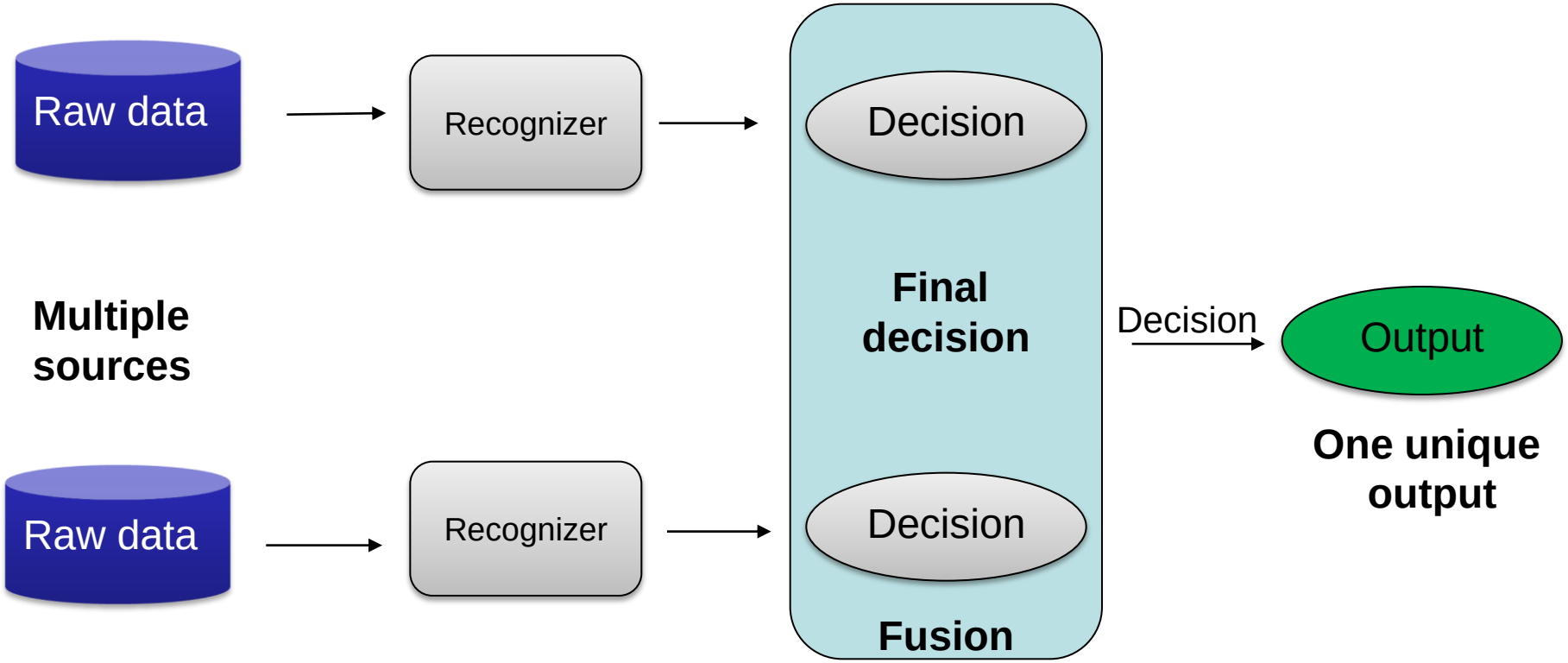
Intermediate fusion strategy

- Each source acts as an input for a specific decision
- All recognizers are coupled in some way to form a meta-recognizer
- One decision is taken



Late fusion strategy

- Each source is processed individually by a specific recognizer
- Multiple independent initial decisions are taken, possibly associated with confidence scores
- A final decision is taken based on this output



Extra levels

- Low-level: feature enhancement, denoising
  - Cross analysis on data may help cleaning, smoothing, or removing infeasible values
- High-level: injection of knowledge (*eg* semantic)
  - Knowledge is viewed as a permanent, ubiquitous modality
  - This may help in disambiguating features, or contextualising the problem

Agenda

- What is multimedia?
  - Networked Media
  - Use cases of Multimedia Information Retrieval
- Challenges in Multimedia IR
  - Volume of data
  - Interpretation of content
- Multimodal Fusion for Multimedia IR
  - Understanding Multimodal Fusion
  - Multimodal Fusion Models
  - Interactive Multimodal Fusion for Multimedia Retrieval
- Wisdom of crowds
  - Mining search logs
- Look Ahead

Joint understanding of modalities

Modalities should be combined for better content understanding

⇒ How to combine them?

⇒ How to make sure we will gain (reliable) information?

⇒ Can we estimate how much information we will gain?

Theoretical frameworks for early fusion

- Bayesian inference theory:

  ⇒fusion of redundant information (Kalman)

- *Information Theory:*

⇒Mutual information (Kullback-Leiber divergence as measure)

- *Belief Theory*

⇒Dempster-Shafer evidential reasoning at a symbolic level

- Other frameworks

⇒*Fuzzy Reasoning, possibility theory,…*

Feature Information Interaction

Multivariate, information-theoretic dependence measure for feature interaction detection [Jakulin, Bratko, 2003], [Matsuda, 2000]

$$I_n(X_1,\ldots,X_m) = \sum_{k=1\ldots n} (-1)^{k+1} \sum_{s=\{i_1,\ldots,i_k\}} H(X_{i_1},\ldots,X_{i_k})$$

Information shared exclusively by k random variables

2-way: mutual information I(A;B) = H(A) + H(B) – H(A,B)

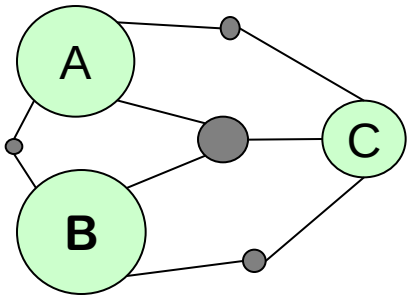
3-way: I(A;B;C) = H(A)+H(B)+H(C)–H(A,B)–H(A,C)-H(B,C)+H(A,B,C)

N-way …

- Finds irreducible and unexpected patterns in data
- Local, stable, unambiguous, symmetric, undirected

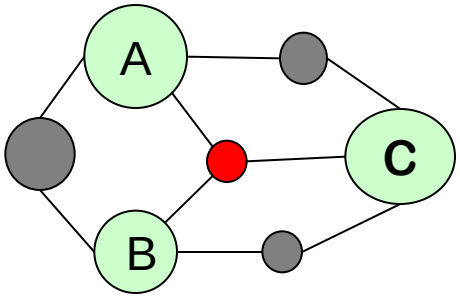
Synergy *vs* Redundancy

I(A;B;C) < 0	I(A;B;C) > 0
A,B statistically independent, but get dependent in context of C ⇒(C = A + B, XOR problem)	**A and B carry the same/similar evidence in context of C**
A,B: relevant, non-redundant towards C	A,B: relevant towards C, but redundant
under fitting: benefit by complicating the model	over fitting: redundancy, possible to simplify the model by feature selection

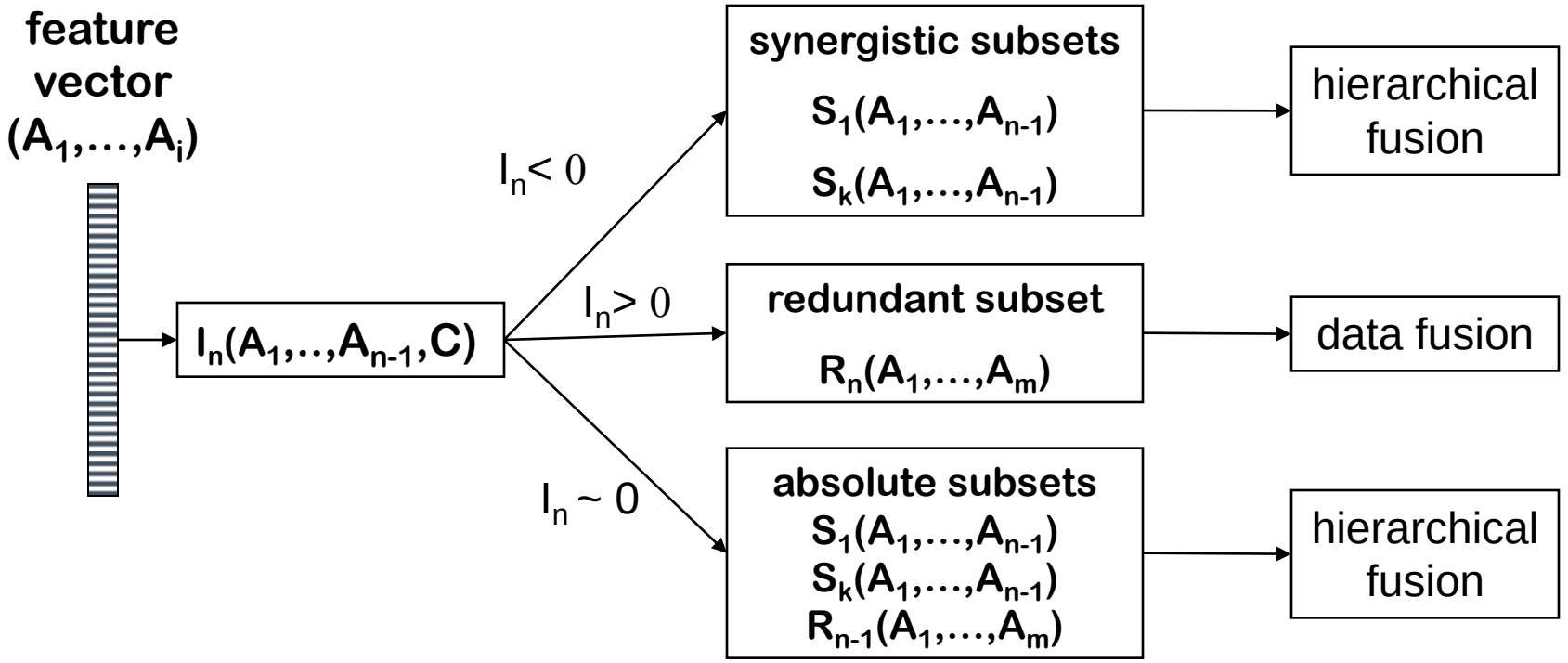


Interaction diagram:

- **ordinary graph**
- **surface = amount of uncertainty**



Feature construction



C … class label, i … number of features

n … order of interactions, k … number of synergistic subsets found

m … size of redundant subset, all redundant features found are concatenated

How to fuse modalities?

Before fusing information, one should look at:

- Potential gain
  - Information theory proposes several measures for upper or lower bounds

  (often problems in getting valid estimations of these bounds)

Gain *vs* Complexity

- Complexity
  - Computation: Feature aggregation may lead to « heavy » data
  - Combinatorial exploration for feature selection/construction
  - Dimensionality

Multidimensionality

Simple feature concatenation may be associated with several issues

- Normalization: how to mix sources of various meaning though same representation
  - Color histogram, Curvature Scale Space signature

  $\Rightarrow$ Both histograms but…

  $\Rightarrow$ Solution comes from "later" fusion

- Heterogeneity: how to mix sources with incompatible representations
  - Color histogram, Covariance matrix

  $\Rightarrow$ Not the same "dimension"

- Problems related to the estimation of:
  - Distance measurements
  - Probability distributions

  $\Rightarrow$ So-called "Curse of dimensionality" [Bellman, 1961]

Data sparsity

Sampling with the same precision the unit disc demands exponentially many points as dimension increases:

- *d=1*: sampling at 0.1 intervals requires 10 points
- *d=D*: sampling at 0.1 intervals requires 10^D points

  $\Rightarrow$ Sampling coarsely a 9-dim space still requires a billion samples!

Conversely:

- *N* samples allow for uniformly sampling the *d*-dimensional unit disc with precision $N^{-1/d}$

$\Rightarrow$ Precision cost in estimation increases drastically!

Parameterization

- Reduce the number of required samples by estimating parameters rather than the actual distribution (histogram)

- Gaussian distribution in D dimensions:
  - D mean values
  - $D(D+1)/2$ co/variance values

  $\Rightarrow D+D(D+1)/2$ parameters ~ $O(D^2)$ samples

  $\Rightarrow$(instead of $O(N^D)$)

Distance measurement

- Learning is often about looking at the vicinity, hence nearest neighbor problems are important

- However, it is known that the difference between the minimum distance and the maximum distance from a given point does not increase as fast as the nearest distance to this point

$\Rightarrow$The relative distance within the neighborhood shrinks to 0

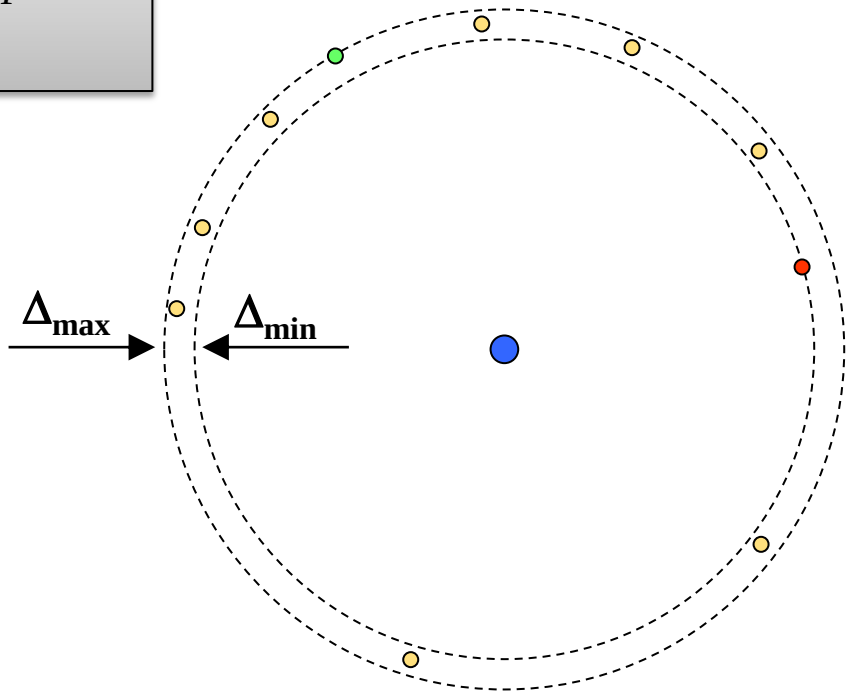
$\Rightarrow$Neighborhood structures may not be meaningful in high dimensions

k-NN issue in high dimensional spaces

Under (fairly generic) conditions:

$$\frac{(\Delta_{\max} - \Delta_{\min})}{\Delta_{\min}} \xrightarrow[d \to \infty]{P} 0$$

On the Surprising Behavior of Distance Metrics in High Dimensional Spaces, Aggarwal *et al*, Lecture Notes in Computer Science, 2001



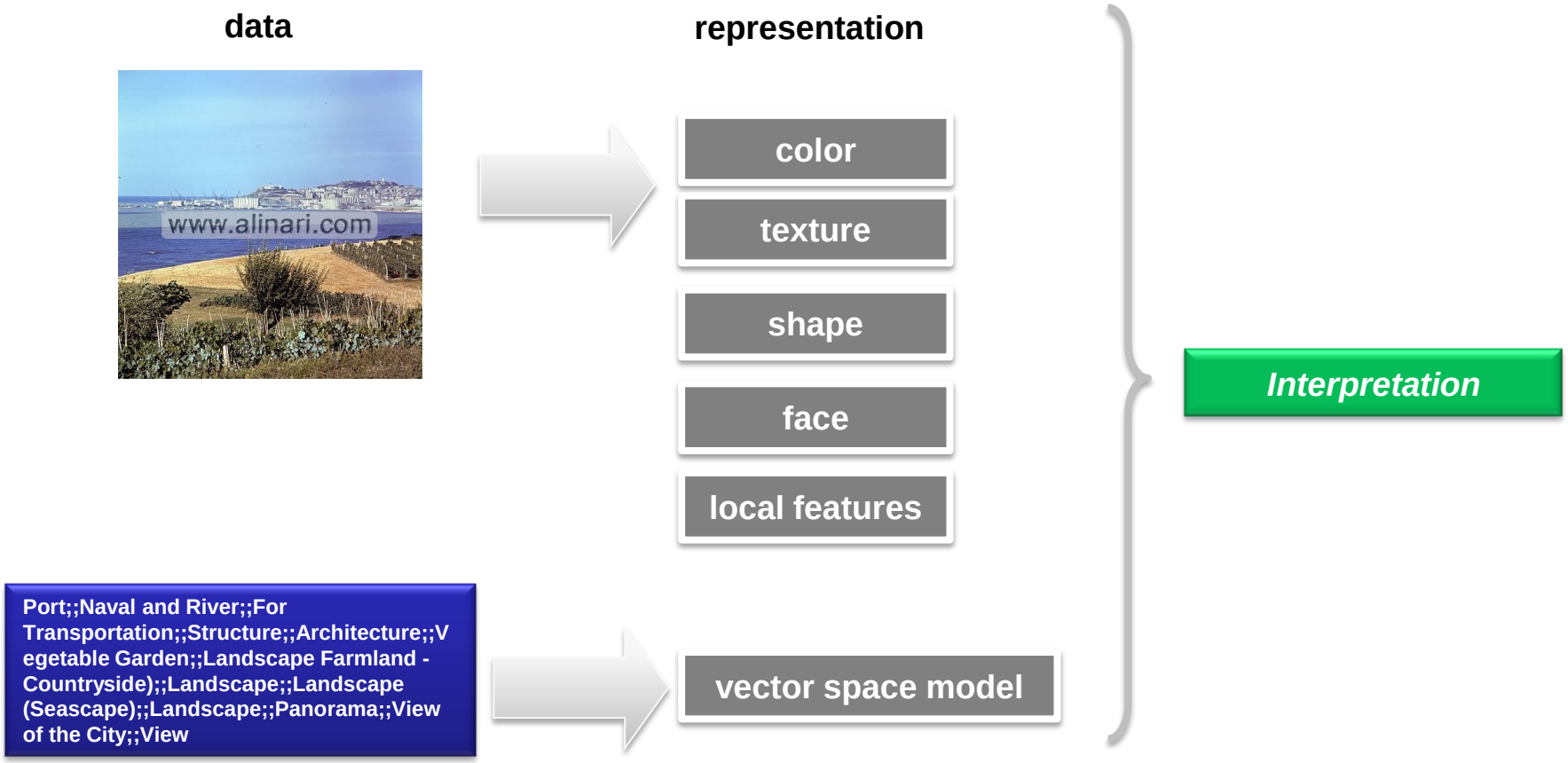
⇒ Neighbourhood-based structures may not be so meaningful in high dimensions

Agenda

- What is multimedia?
  - Networked Media
  - Use cases of Multimedia Information Retrieval
- Challenges in Multimedia IR
  - Volume of data
  - Interpretation of content
- Multimodal Fusion for Multimedia IR
  - Understanding Multimodal Fusion
  - Multimodal Fusion Models
  - Interactive Multimodal Fusion for Multimedia Retrieval
- Wisdom of crowds
  - Mining search logs
- Look Ahead

Multimodal fusion

- From many sources of information and context, how to make our best to “interpret” the data



Multidimensionality

Simple feature concatenation may be associated with several issues

- Normalization: how to mix sources of various meaning though same representation
  - Color histogram, Curvature Scale Space signature
  - ⇒ Both histograms but…
  - ⇒ Solution comes from “later” fusion
- Heterogeneity: how to mix sources with incompatible representations
  - Color histogram, Covariance matrix
  - ⇒ Not the same “dimension”
- Problems related to the estimation of:
  - Distance measurements
  - Probability distributions
  - ⇒ So-called “Curse of dimensionality” [Bellman, 1961]

Principle of pairwise data representation

Instead of being absolute (relative to an „origin“), features are made relative:

- to each other (pairwise)
- with respect to selected items (*eg* for indexing)
- with respect to a group (*eg* from clustering)
- …

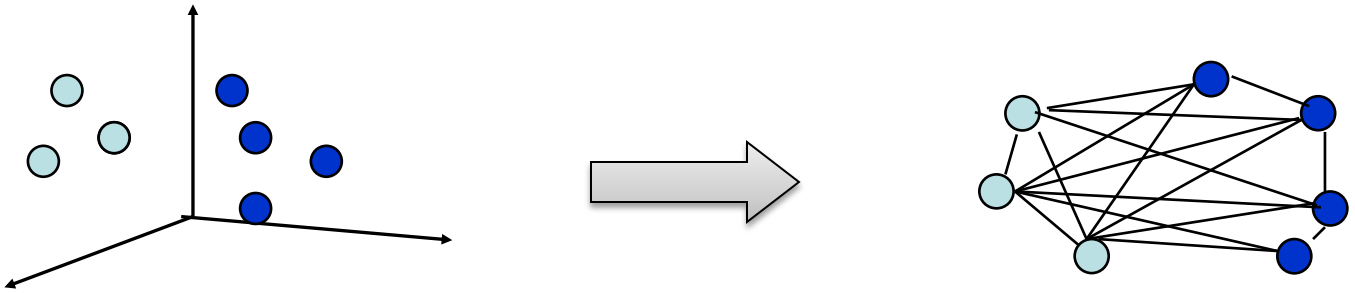
⇒ All values become distance values

- Positive real number
- Scale is not resolved
- Space simplification (*eg* pivot point)

The dissimilarity space [Pekalska et. al, JMLR, 2002]

Data is no longer considered from the point of view of the features but from the point of view of **(dis)similarities**

- Absolute feature measurements replaced by relative similarity relationships



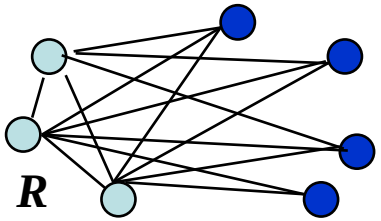
- Abstraction of the nature of the features/modalities used

Monomodal dissimilarity space

- Given collection Ω of elements, $d(p_i, p_j)$ is a dissimilarity measure between elements p_i and $p_j \in \Omega$
- Given a subset $R = \{p_1, p_2, \ldots, p_N\}$ of Ω, the dissimilarity space is defined as:

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

- R is called the **representation set**
  - "new features" are dissimilarities of an element to the representation set
  - the more R is large, the best $\mathbf{d}$ approximates the initial feature space (Classical Scaling Projection)



- Choosing R must be a trade-off between space dimensionality and learning efficiency

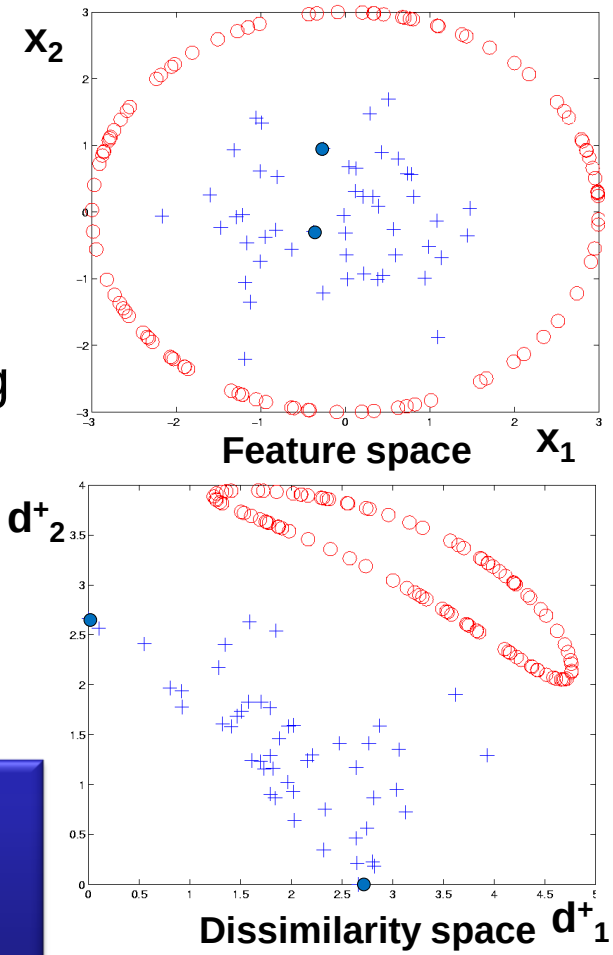
Task-dependent dissimilarity space

- S^+ positive and S^- negative classes
  - Positives belong to one class
  - Negatives belong to x class

- Choosing S^+ as representative , and assuming

$$\overline{\left\|\mathbf{d}\left(S^+, S^+\right)\right\|} < \overline{\left\|\mathbf{d}\left(S^-, S^+\right)\right\|}$$

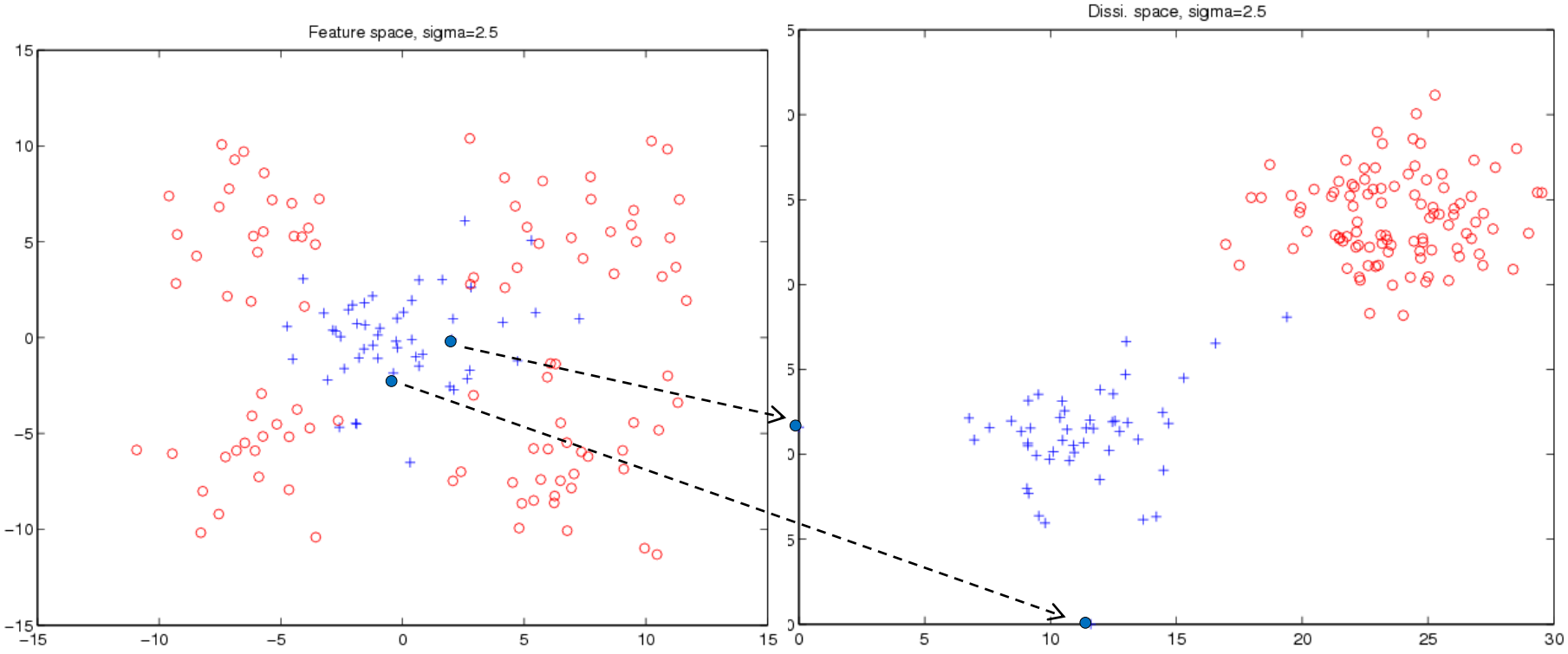
  - 1+x to 1+1 classification
  - Low dimensional space = card(S^+)=N



Bruno, E., & Marchand-Maillet, S. (2009). Multimodal Preference Aggregation for Multimedia Information Retrieval. Journal of Multimedia, 4(5), 321-329

Bruno, E., Moënne-Loccoz, N., & Marchand-Maillet, S. (2008). Design of multimodal dissimilarity spaces for retrieval of multimedia documents. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(9), 1520-1533.

Example



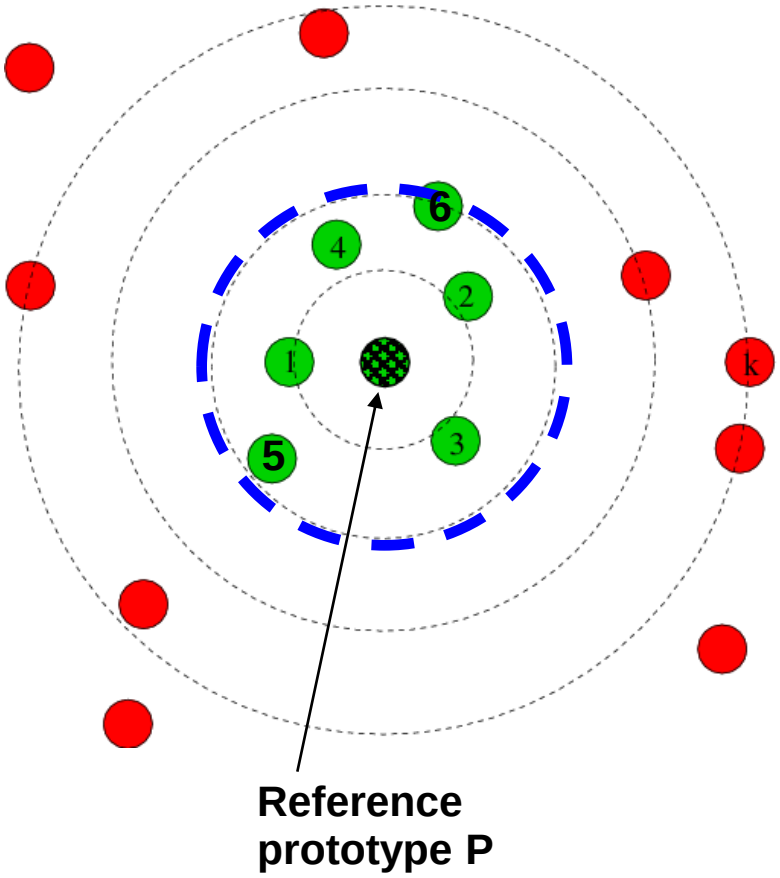
Application to learning

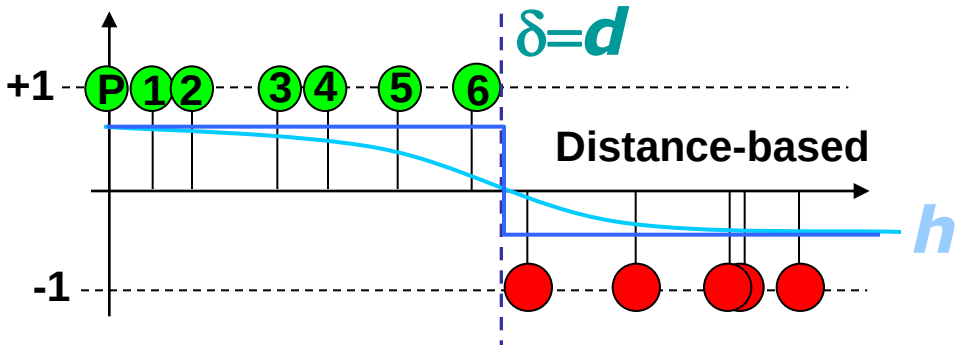
From training data, select a decision function that characterises class separation present within the training set

- Avoid over-fitting (learn only the training set)
- Ensure generalisation (« smooth » decisions)

In this context, the representative set is chosen from the training set…

Prototype-based representation

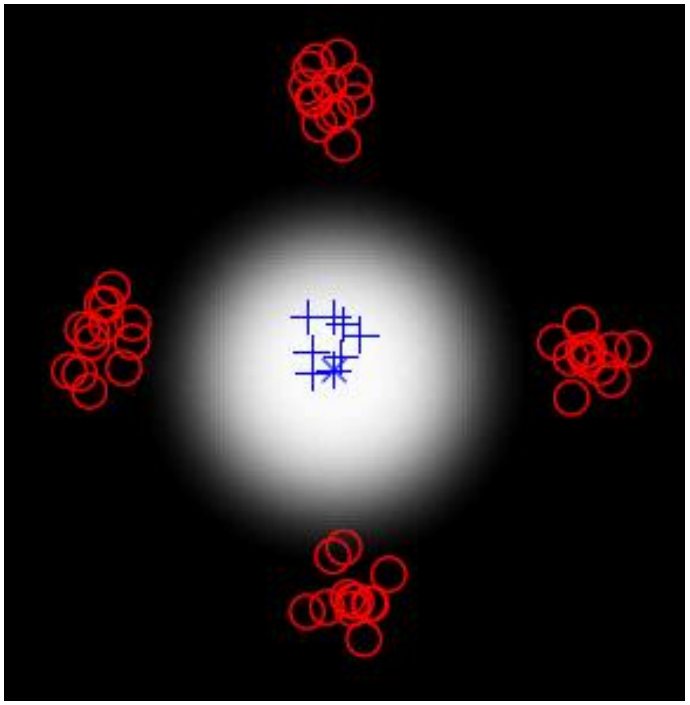


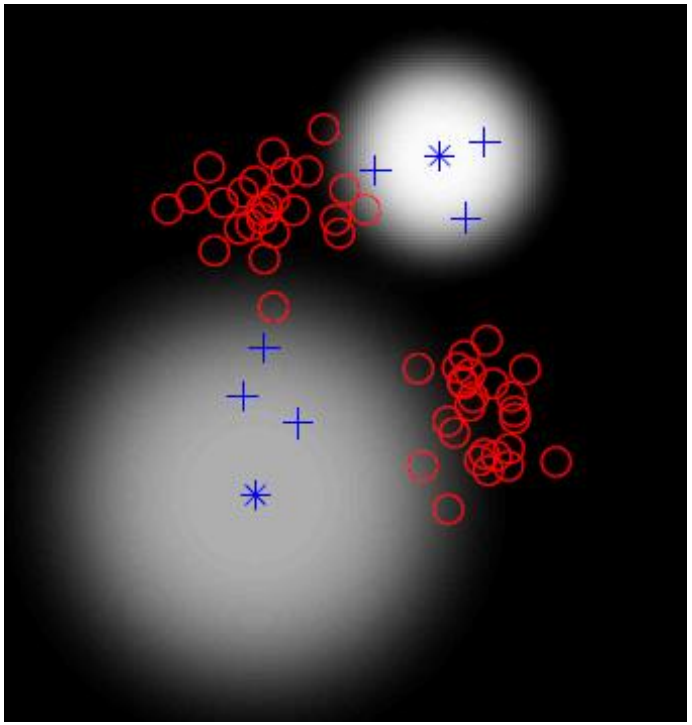


- hard function allow for a decision
- soft function enables ordering

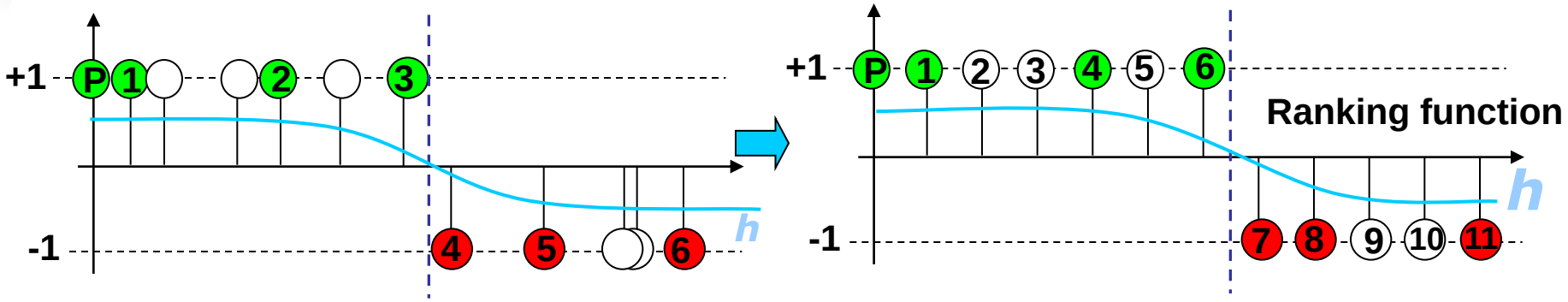
The « green » class is described by P and *h*

Examples





Preference-based learning



- Simplifies the estimation of h_i
- Conforms to the boosting principle (RankBoost):

$$H(x) = \sum_{i=1}^{M} \alpha_i h_i(x) = \langle \alpha, h \rangle \qquad M = \#\text{examples x } \#\text{modalities}$$

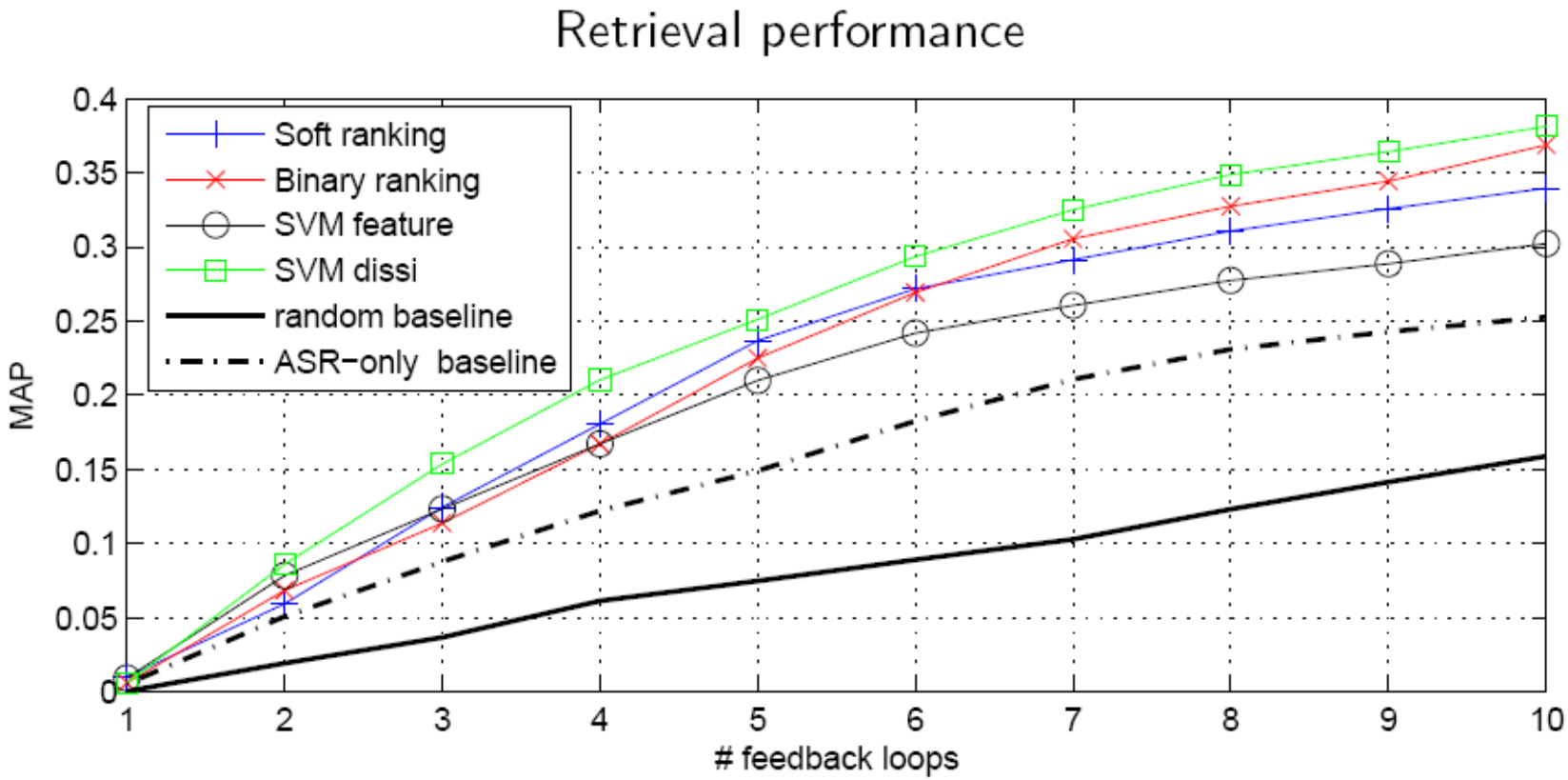
$$h_i(x) = 2\exp(-\gamma \pi_i^2(x)) - 1 \qquad \textbf{(weak learner)}$$

$\pi_i(x)$ is the rank of document x in a given modality w.r.t to a given positive example document, taken as prototype

Prototype-based representation

$$x \mapsto \boldsymbol{\pi}(x) = \left[\cdots \quad ,2,7,\pi_p^{(\mathrm{SIFT})}(x), \quad \cdots\right]$$

TRECVid 2005 corpus, search task



Bruno, E., & Marchand-Maillet, S. (2009). Multimodal Preference Aggregation for Multimedia Information Retrieval. Journal of Multimedia, 4(5), 321-329

Bruno, E., Moënne-Loccoz, N., & Marchand-Maillet, S. (2008). Design of multimodal dissimilarity spaces for retrieval of multimedia documents. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(9), 1520-1533.

Agenda

- What is multimedia?
  - Networked Media
  - Use cases of Multimedia Information Retrieval
- Challenges in Multimedia IR
  - Volume of data
  - Interpretation of content
- Multimodal Fusion for Multimedia IR
  - Understanding Multimodal Fusion
  - Multimodal Fusion Models
  - Interactive Multimodal Fusion for Multimedia Retrieval
- Wisdom of crowds
  - Mining search logs
- Look Ahead

Mining search logs

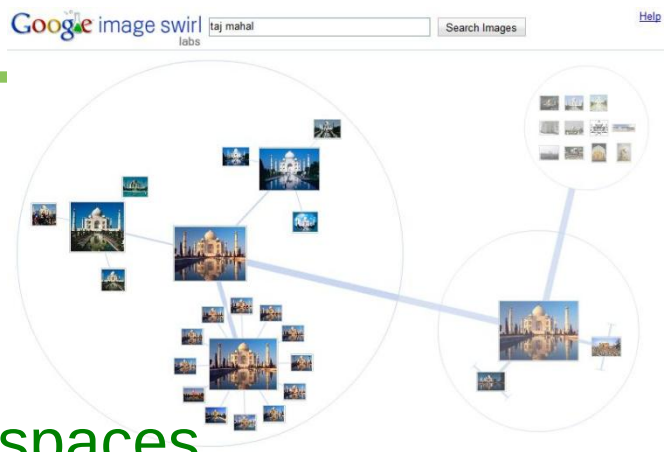
- Long-term learning
  - Transfer of semantic knowledge onto the data itself for improving further retrieval

- User profiling / recommendation
  - By knowing better the user profile from preceding actions, the system may better anticipate future actions
  - By learning from search history, the system may simulate the user behavior and discover (and recommend) relevant items

Agenda

- What is multimedia?
  - Networked Media
  - Use cases of Multimedia Information Retrieval
- Challenges in Multimedia IR
  - Volume of data
  - Interpretation of content
- Multimodal Fusion for Multimedia IR
  - Understanding Multimodal Fusion
  - Multimodal Fusion Models
  - Interactive Multimodal Fusion for Multimedia Retrieval
- Wisdom of crowds
  - Mining search logs
- Look Ahead

Look ahead



- Search and browse
  - Google Swirl
  - Collection browsers
  - → Mining multimedia representation spaces
- Scalable Multimedia Information Retrieval
  - Large-scale **interactive** MIR
- **Mobile** Multimedia Information Retrieval
  - Light-weight efficient process
    - Processing and communications

Agenda

- Complement: Benchmarking MIR
  - Some pointers for you to keep
  - See also lectures on Evaluation

ImageCLEF Wikipedia task 2010

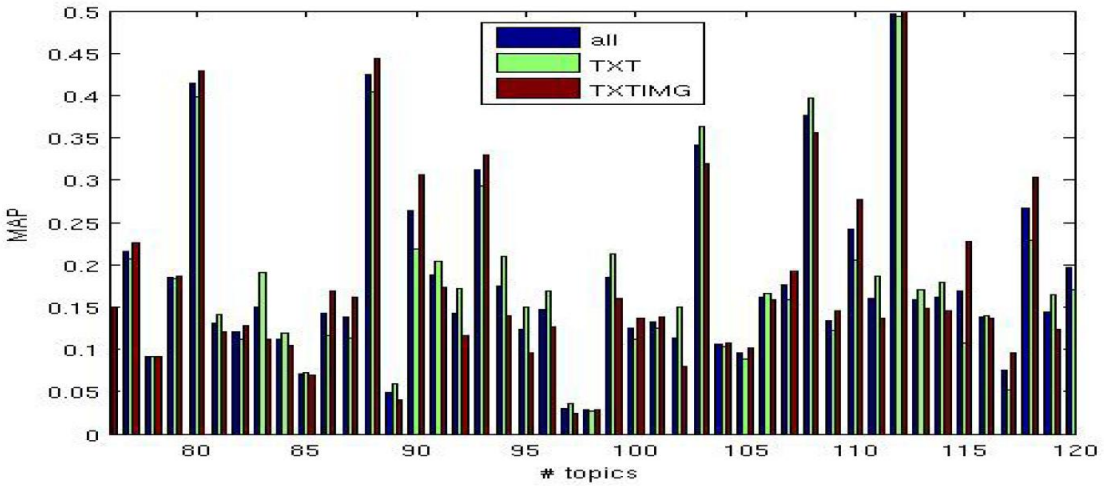
Images + text (Overall: 237,434)

- English only: 70,127
- German only: 50,291
- French only: 28,461
- English and German: 26,880
- English and French: 20,747
- German and French: 9,646
- English, German and French: 22,899
- Language undetermined: 8,144
- No textual annotation: 239

```
<?xml version="1.0" encoding="UTF-8" ?>
<image id="8120" file="images/1/8120.jpg">
    <name>Kanzler21a.jpg</name>
    <text xml:lang="en">
        <description>German Federal Chancellery, Berlin. View
        from the east of the main entrance.</description>
        <comment />
        <caption article="text/en/1/301543">Chancellery in
        Berlin, since 2001</caption>
        <caption article="text/en/1/307163">The Chancellery in
        Berlin is the seat of the Chancellor</caption>
        <caption article="text/en/3/328103"> Bundeskanzleramt
        </caption>
    </text>
    <text xml:lang="de">
        <description>Bundeskanzleramt, Berlin. Blick von Osten
        (Haupteingang)</description>
        <comment />
        <caption article="text/de/1/400134">
        Bundeskanzleramtsgebäude in Berlin </caption>
        <caption article="text/de/1/405148">Kanzleramtsgebäude
        in Berlin</caption>
    </text>
    <text xml:lang="fr">
        <description />
        <comment />
        <caption article="text/fr/1/500997">La chancellerie
        </caption>
    </text>
    <comment>(contrast)</comment>
    <license>GFDL</license>
</image>
```

ImageCLEF 2009 Results [Tsikrika & Kludas, 2010]

Modality	MAP		P@20		R-prec.	
	Mean	SD	Mean	SD	Mean	SD
All top 90% runs (46 runs)	0.1751	0.0302	0.2356	0.0624	0.2076	0.0572
TXT in top 90% runs (23 runs)	0.1726	0.0326	0.2278	0.0427	0.2038	0.0328
TXTIMG in top 90% runs (23 runs)	0.1775	0.0281	0.2433	0.0364	0.2115	0.0307

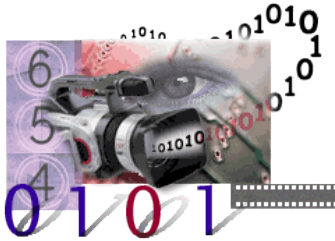


http://www.imageclef.org/

Fig. 2. Average topic performance over all, text-only and text/visual runs

easy	medium	hard	very hard
(112) hot air balloons	(118) coral reef underwater	(120) yellow flower	(105) snowy street
(88) madonna portrait	(90) satellite image of river	(91) landline telephone	(78) sculpture of an animal
(80) orthodox icons	(110) desert landscape	(99) flowers on trees	(117) earth from space
(108) bird nest	(77) real rainbow	(79) stamp human face	(85) aerial ph. of landscapes
(103) palm trees		(107) red fruit	(89) people laughing
(93) close up antenna		(94) people with dogs	(97) woman in pink dress
6	4	28	7

TRECVid



DIGITAL VIDEO RETRIEVAL at NIST

http://trecvid.nist.gov/

The goal of the conference series is to encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results.

MIREX

http://www.music-ir.org

- The Music Information Retrieval Evaluation eXchange (MIREX) is an annual evaluation campaign for Music Information Retrieval (MIR) algorithms, coupled to the International Society (and Conference) for Music Information Retrieval (ISMIR)

Other

- 3D Retrieval
  - SHREC: Shape Retrieval Contest
  - http://www.aimatshape.net/event/SHREC
- XML retrieval
  - INitiative for the Evaluation of XML retrieval
  - https://inex.mmci.uni-saarland.de/
- Many dedicated corpora for specific tasks
  - Generic (Image-Net)
  - Medical (ImageCLEF)
  - Satellite imagery
  - …

Thank you!

Questions?

Some References

Bruno, E., & Marchand-Maillet, S. (2009). Multimodal Preference Aggregation for Multimedia Information Retrieval. Journal of Multimedia, 4(5), 321-329

Bruno, E., & Marchand-Maillet, S. (2009). Multiview clustering: a late fusion approach using latent models. In Proceedings of the 32nd ACM Special Interest Group on Information Retrieval Conference, SIGIR 09, Boston, USA.

Kludas, J., Marchand-Maillet, S., & Bruno, E. (2008). Exploiting document feature interactions for efficient information fusion in high dimensional spaces. In Proceedings of the First International Workshops on Image Processing Theory, Tools and Applications (IPTA'2008), Sousse, Tunisia.

Kludas, J., Bruno, E., & Marchand-Maillet, S. (2008). Can Feature Information Interaction help for Information Fusion in Multimedia Problems?. *To appear in Multimedia Tools and Applications Journal special issue on "Metadata Mining for Image Understanding"*.

Bruno, E., Moënne-Loccoz, N., & Marchand-Maillet, S. (2008). Design of multimodal dissimilarity spaces for retrieval of multimedia documents. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(9), 1520-1533.

Bruno, E., Kludas, J., & Marchand-Maillet, S. (2007). Combining Multimodal Preferences for Multimedia Information Retrieval. In Proc. of International Workshop on Multimedia Information Retrieval, Augsburg, Germany.

Janvier, B., Bruno, E., Marchand-Maillet, S., & Pun, T. (2005). A contextual model for semantic video structuring. In Proceedings of the 13th European Signal Processing Conference (Eusipco2005), Antalya, Turkey.

Kosinov, S., & Marchand-Maillet, S. (2004). Multimedia autoannotation via hierarchical semantic ensembles. In Proceedings of the Int. Workshop on Learning for Adaptable Visual Systems (LAVS 2004), Cambridge, UK.

Kosinov, S., Marchand-Maillet, S., Kozintsev, I., Dulong, C., & Pun, T. (2006). Dual diffusion model of spreading activation for content-based imageretrieval. In 8th ACM SIGMM International Workshop on Multimedia Information Retrieval, Santa Barbara, CA, USA.

Tsikrika, T., & Kludas, J. (2010). Overview of the wikipediaMM task at ImageCLEF 2009. LNCS, Springer.