# Similarity Search in Non-text Data

Pavel Zezula
Faculty of Informatics
Masaryk University, Brno

# Real-Life Motivation

*The social psychology view*

- Any event in the history of organism is, in a sense, **unique**.

- *Recognition*, *learning*, and *judgment* presuppose an ability to categorize stimuli and classify situations by **similarity**.

- Similarity (*proximity*, *resemblance*, *communality*, *representativeness*, *psychological distance*, etc.) is **fundamental** to theories of *perception*, *learning*, *judgment*, etc.

# Contemporary Networked Media

*The digital data view*

- Almost **everything** that we *see*, *read*, *hear*, *write*, *measure*, or *observe* can be **digital**.
- Users **autonomously** *contribute* to production of global media and the growth is **exponential**.
- Sites like Flickr, YouTube, Facebook host user contributed content for a variety of **events**.
- The elements of networked media are related by numerous multi-facet **links of similarity**.

# Examples

- Does the computer disk of a suspected criminal contain illegal multimedia material?
- What are the stocks with similar price histories?
- Which companies advertise their logos in the direct TV transmission of football match?
- Is it the situation on the web getting close to any of the network attacks which resulted in significant damage in the past?

# Challenge

- Networked media is getting close to the human “fact-bases”.

- **Similarity data management** is needed to *connect*, *search*, *filter*, *merge*, *relate*, *rank*, *cluster*, *classify*, *identify*, or *categorize* objects across various collections.

**WHY?**

**It is the *similarity* which is in the world *revealing*.**

# Limitations: Data Types

## We have

- Attributes
  - Numbers, strings, etc.
- Text (text-based)
  - Documents, annotations

## We need

- Multimedia
  - Image, video, audio
- Security
  - Biometrics
- Medicine
  - EKG, EEG, EMG, EMR, CT, etc.
- Scientific data
  - Biology, chemistry, physics, life sciences, economics
- Others
  - Motion, emotion, events, etc.

# Limitations:
# Models of Similarity

## We have

- Simple geometric models, typically vector spaces

## We need

- More complex model
- Non metric models
- Asymmetric similarity
- Subjective similarity
- Context aware similarity
- Complex similarity
- Etc.

# Limitations: Queries

## We have

- Simple query
  - Nearest neighbor
  - Range

## We need

- More query types
  - Reverse NN, distinct NN, similarity join
- Other similarity-based operations
  - Filtering, classification, event detection, clustering, etc.
- Similarity algebra
  - May become the basis of a “Similarity Data Management System”

# Limitations: Implementation Strategies

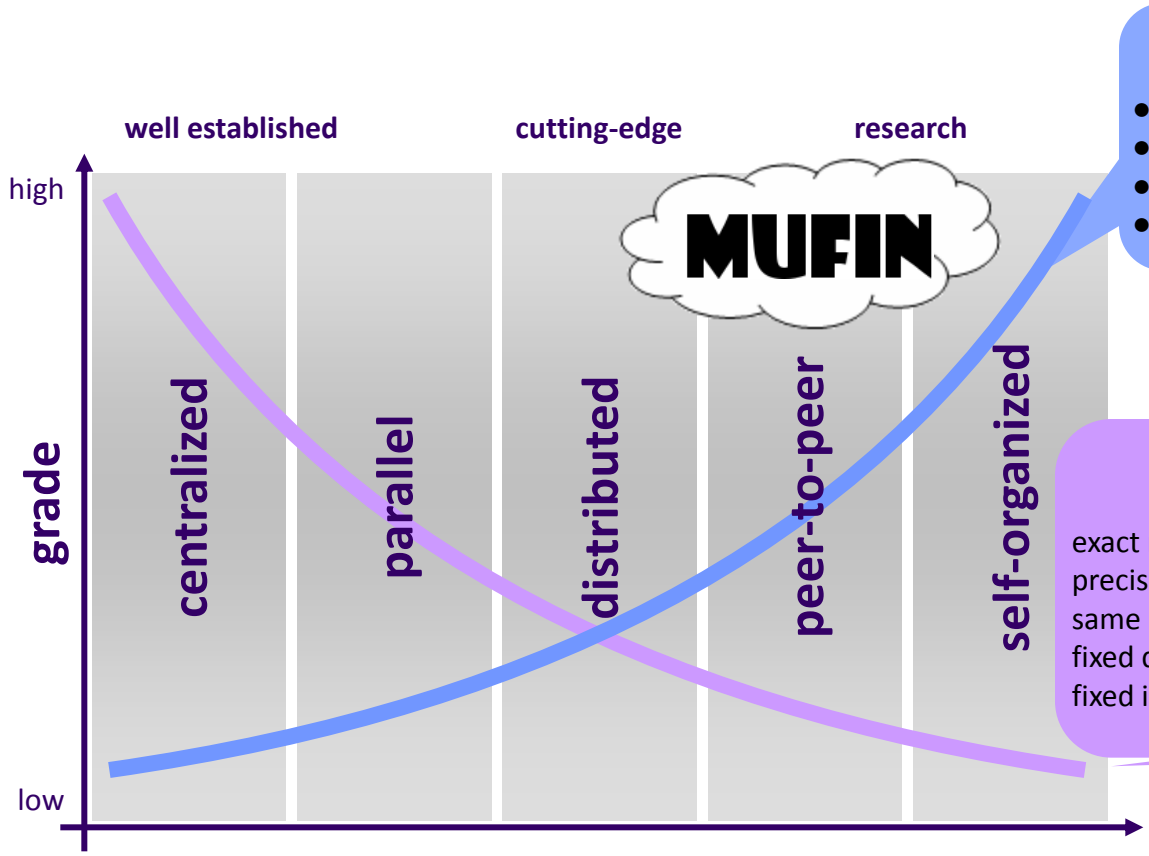
## We have

- Centralized or parallel processing

## We need

- Scalable and distributed architectures
- MapReduce like approaches
- P2P architectures
- Cloud computing
- Self-organized architectures
- Etc.

# Search Strategy Evolution




**Scalability**

- data volume - exponential
- number of users (queries)
- variety of data types
- multi-lingual, -feature –modal queries

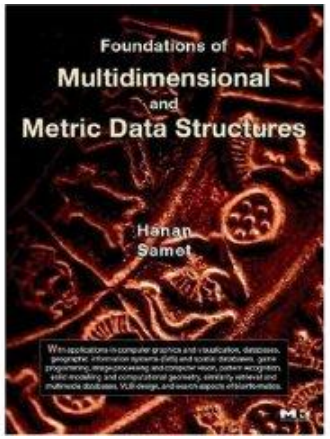
**Determinism**

exact match ▸ similarity
precise ▸ approximate
same answer ▸ good answer; recommendation
fixed query ▸ personalized; context aware
fixed infrastr. ▸ dynamic mapping; mobile dev.

# Word Cloud of Applications

# Metric Search Grows in Popularity

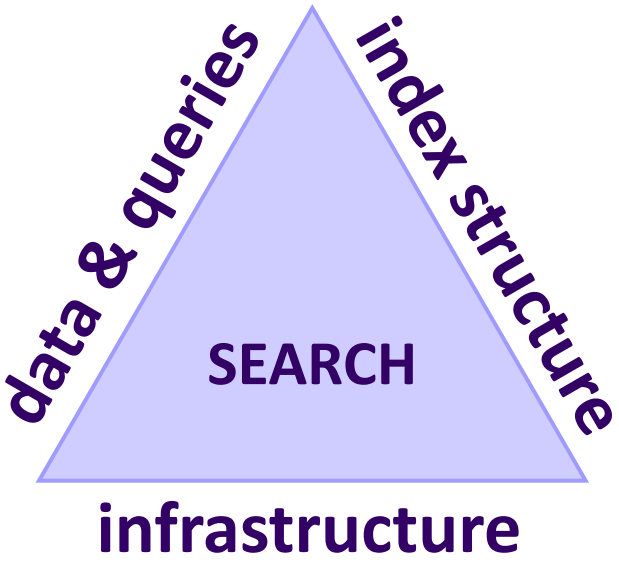
Hanan Samet
**Foundation of Multidimensional and Metric Data Structures**
*Morgan Kaufmann, 2006*




P. Zezula, G. Amato, V. Dohnal, and M. Batko
**Similarity Search: The Metric Space Approach**
*Springer, 2006*

# The MUFIN Approach

MUFIN: MUlti-Feature Indexing Network




**Extensibility**
metric space

**Scalability**
P2P structure

**Tuning** of performance
Internet / GRID / LAN
network independence

# Metric Space
# an Abstraction of Similarity

- Metric space: $\mathcal{M} = (\mathcal{D}, d)$
  - $\mathcal{D}$ – domain
  - distance function $d(x,y)$

    $\forall x,y,z \in \mathcal{D}$
    - $d(x,y) > 0$ - *non-negativity*
    - $d(x,y) = 0 \Leftrightarrow x = y$ - *identity*
    - $d(x,y) = d(y,x)$ - *symmetry*
    - $d(x,y) \leq d(x,z) + d(z,y)$ - *triangle inequality*

# Examples of Distance Functions

- $L_p$ metric functions (for vectors)
  - $L_1$ – city-block distance
  - $L_2$ – Euclidean distance
  - $L_\infty$ – infinity

$$L_1(x,y) = \sum_{i=1}^{n} | x_i - y_i |$$

$$L_2(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$L_\infty(x,y) = \max_{i=1}^{n} |x_i - y_i|$$

- edit distance (for strings)
  - minimal number of insertions, deletions and substitutions
  - d('application', 'applet') = 6
- Jaccard's coefficient (for sets A,B)

$$d(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

# Examples of Distance Functions

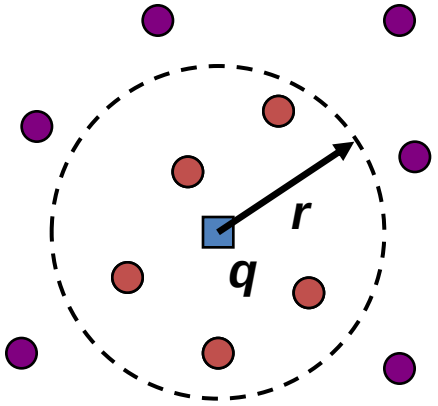
- Quadratic-form distance
  - for vectors with correlated dimensions
- Hausdorff distance
  - for sets with elements related by another distance
- Earth movers distance
  - primarily for histograms (sets of weighted features)
- and many others

# Similarity Search Problem

- For $\mathcal{X} \subseteq \mathcal{D}$ in metric space $\mathcal{M}$, pre-process $\mathcal{X}$ so that the similarity queries are executed efficiently.

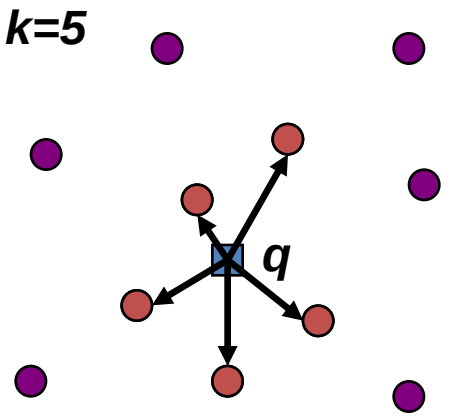
- similarity queries
  - range search
  - $R(q,r) = \{ x \in \mathcal{X} \mid d(q,x) \leq r \}$

    $q \in \mathcal{D}, r \geq 0$

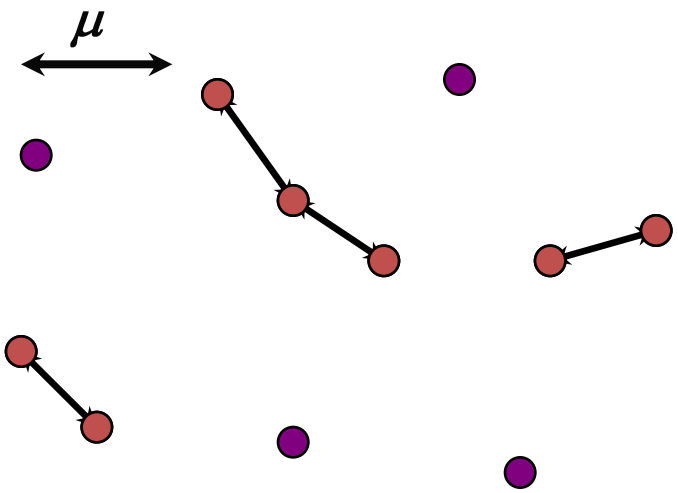


# Similarity Queries

- *k*-nearest neighbours
  - $NN(q,k) = A,\ q \in \mathcal{D},\ k > 0$
  - $A \subseteq \mathcal{X},\ |A| = k$
  - $\forall x \in A,\ y \in \mathcal{X} - A,\ d(q,x) < d(q,y)$

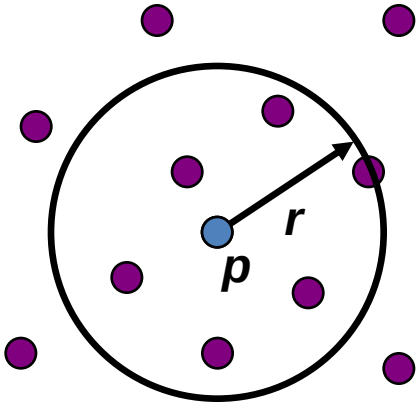



- similarity join
  - $X = \{x_1, x_2, \ldots x_N\},\ Y = \{y_1, y_2, \ldots y_M\}$
  - $\{(x_i,y_j) \mid d(x_i,y_j) < \mu\}$
  - similarity „self“ join ⇔ X = Y

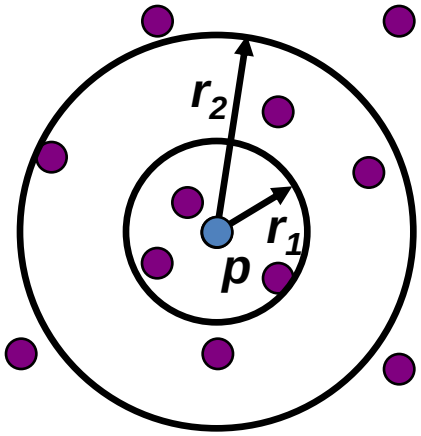


# Basic Partitioning Principles

- ball partitioning
  - $\{ x \in \mathcal{X} \mid d(p,x) \leq r \}$
  - $\{ x \in \mathcal{X} \mid d(p,x) \geq r \}$

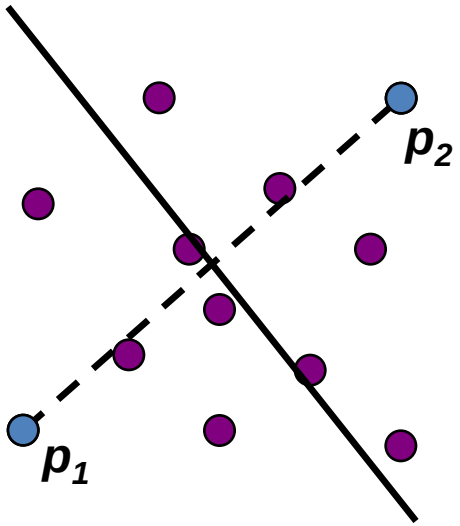



- multiple ball partitioning
  - $\{ x \in \mathcal{X} \mid d(p,x) \leq r_1 \}$
  - $\{ x \in \mathcal{X} \mid d(p,x) > r_1$ and $d(p,x) \leq r_2 \}$
  - $\{ x \in \mathcal{X} \mid d(p,x) > r_2 \}$



# Basic Partitioning Principles

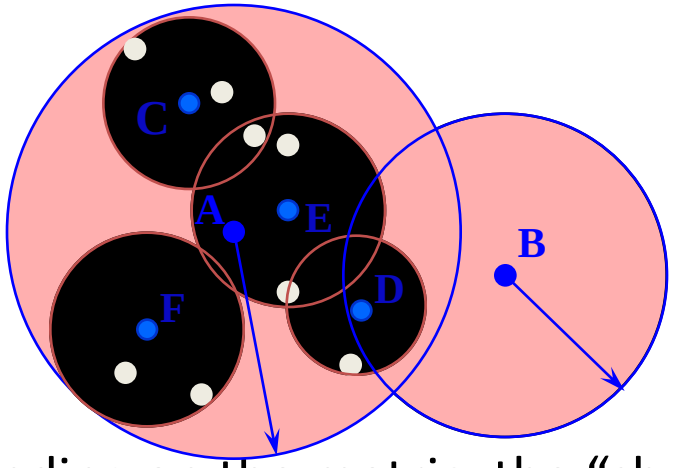
- generalised hyperplane
  - $\{ x \in \mathcal{X} \mid d(p_1,x) \leq d(p_2,x) \}$
  - $\{ x \in \mathcal{X} \mid d(p_1,x) > d(p_2,x) \}$



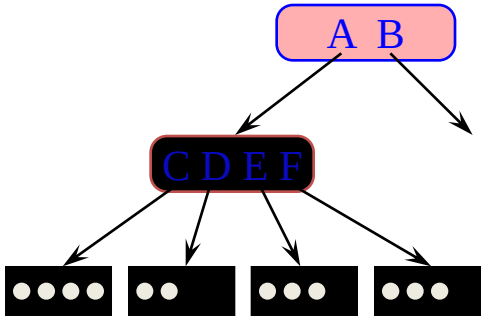
# The M-tree [Ciaccia, Patella, Zezula, VLDB 1997]

1) Paged organization
2) Dynamic
3) Suitable for arbitrary metric spaces
4) I/O and CPU optimization - computing *d* can be time-consuming

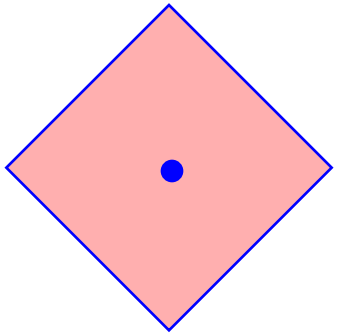
# The M-tree Idea



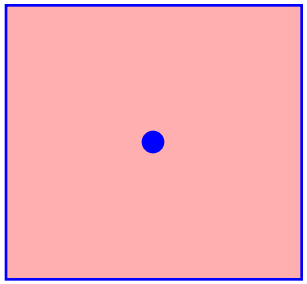

Metric: $L_2$ (Euclidean)



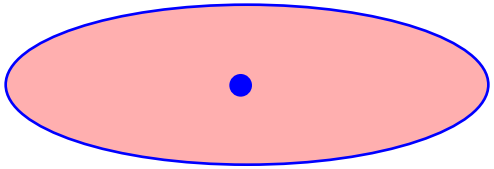

- Depending on the metric, the "shape" of index regions changes



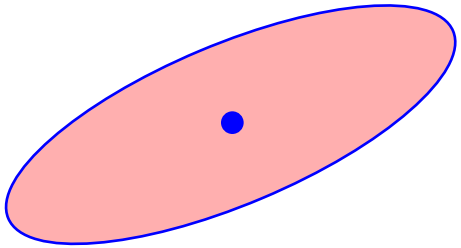
$L_1$ (city-block)



$L_\infty$ (max-metric)



weighted-Euclidean



quadratic form

# The M-tree on the Web

- Home page: http://www-db.deis.unibo.it/Mtree/


M-tree software can be freely downloaded

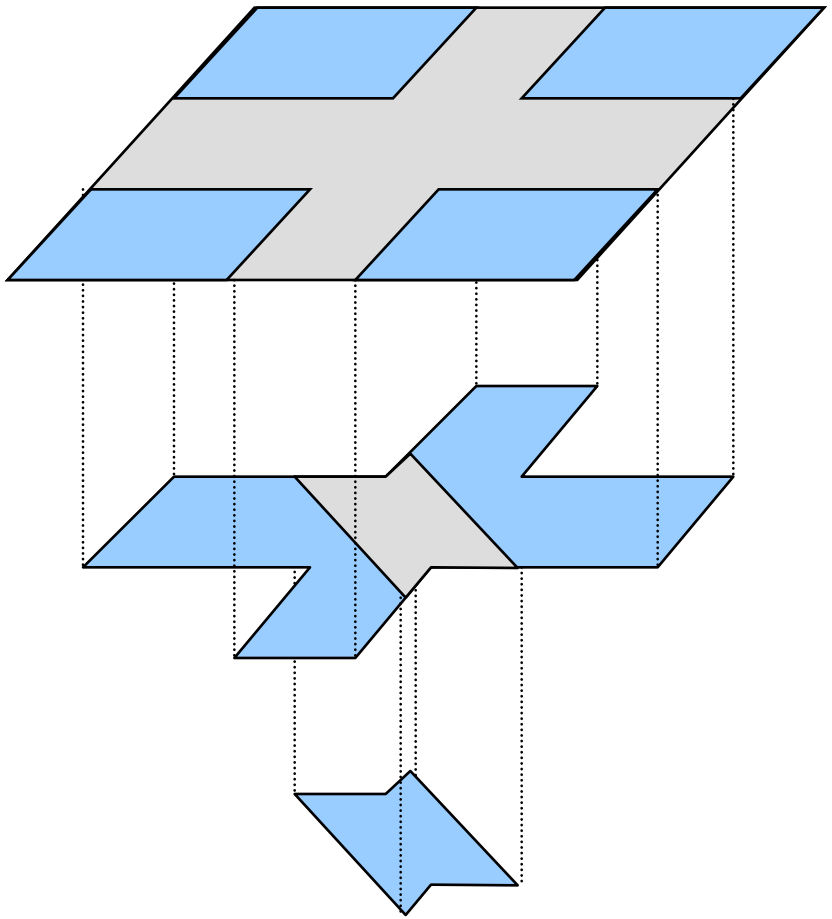
- Based on GiST package (Berkeley Univ.)

Google Scholar: 1 300 citations in Dec. 2011

# M-tree family

- Bulk loading
- Slim-tree
- Multi-way insertion
- PM-tree
- $M^2$-tree
- etc.

# D-Index [Dohnal, Gennaro, Zezula, MTA 2002]

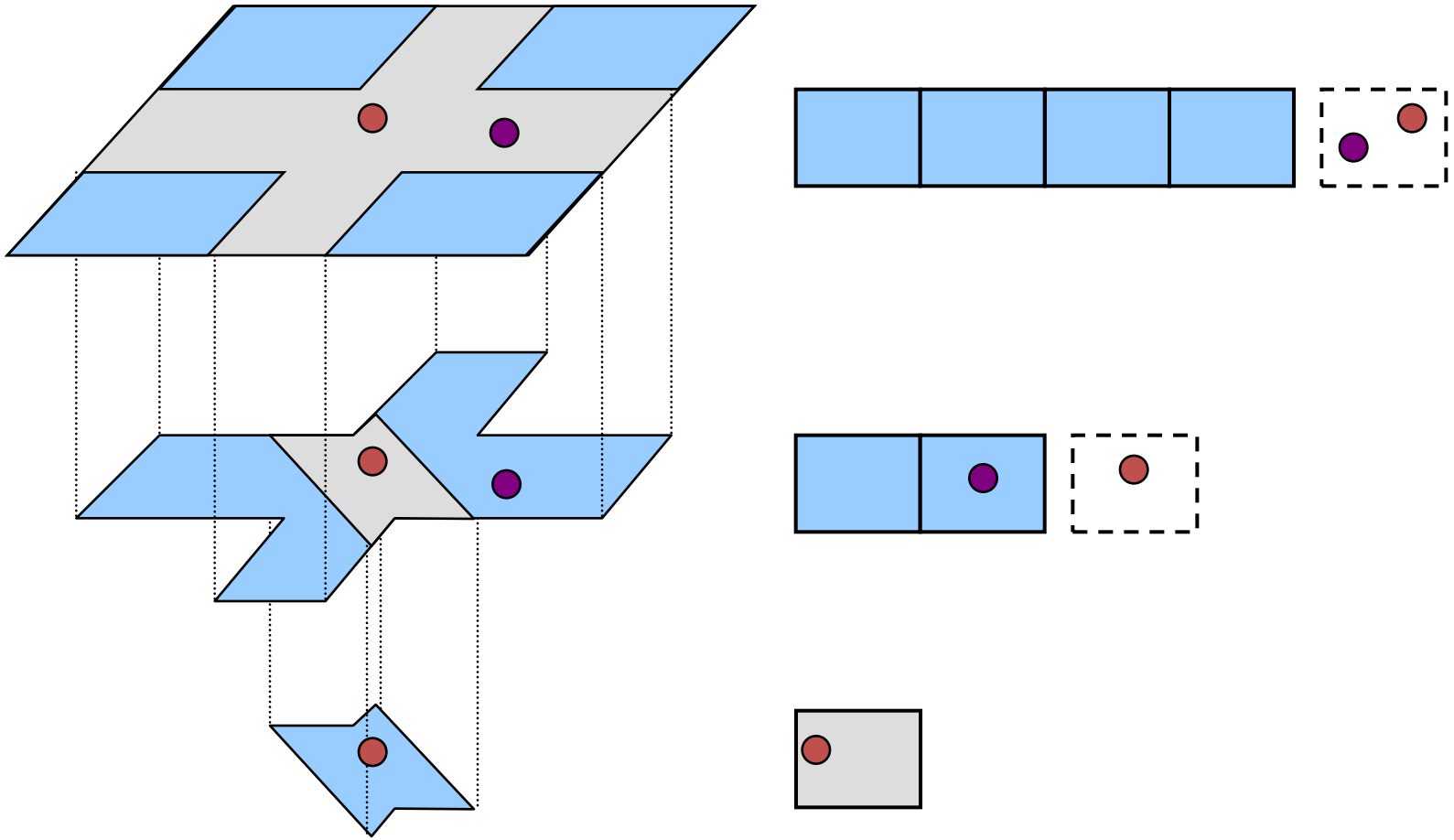


4 separable buckets at the first level

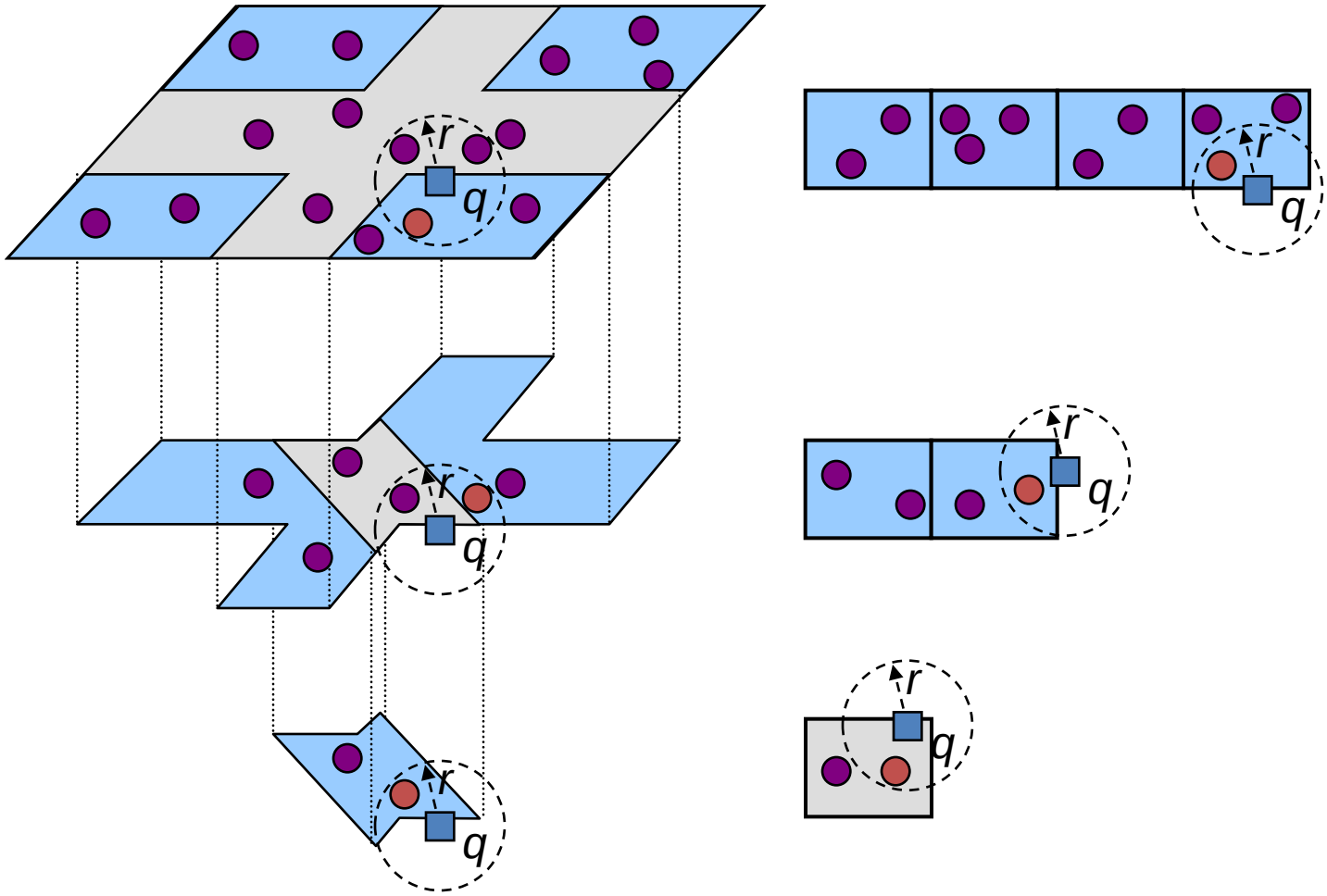
2 separable buckets at the second level

exclusion bucket of the whole structure

# D-index: Insertion



# D-index: Range Search

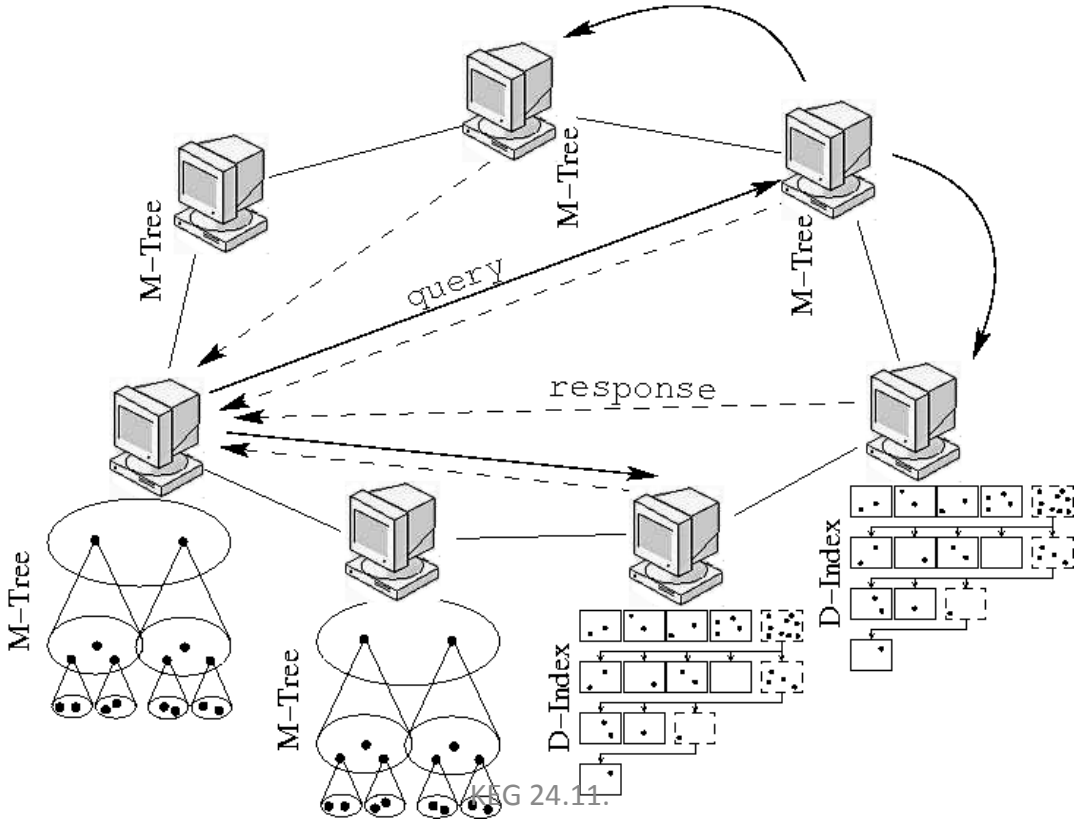


# Implementation Postulates of Distributed Indexes

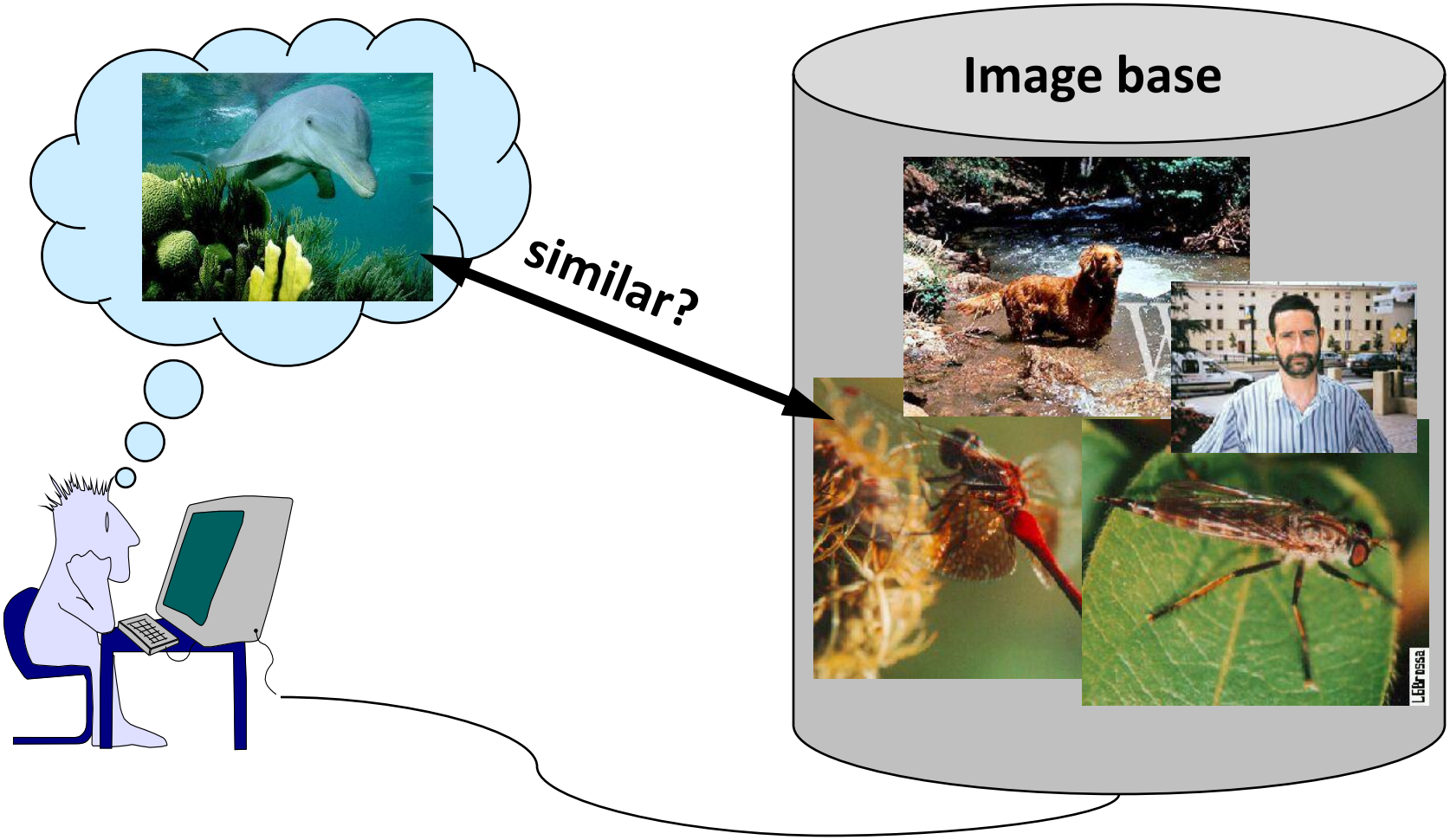
- **scalability** – nodes (computers) can be added (removed)
- **no hot-spots** – no centralized nodes, no flooding by messages
- **update independence** – network update at one site does not require an immediate change propagation to all the other sites

# Peer-to-Peer Indexing

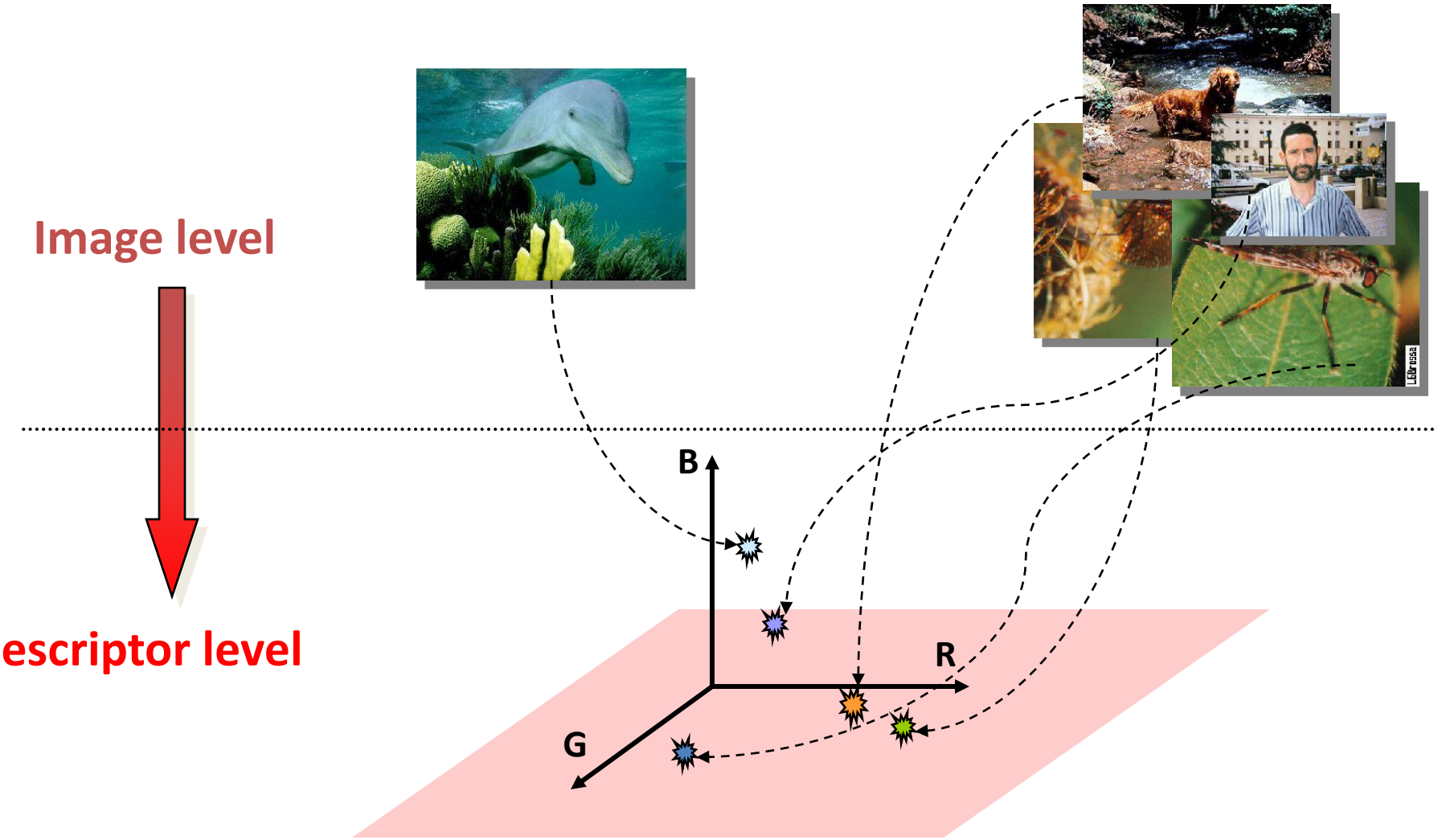
- Native metric techniques: **GHT*, VPT***
- Transformation techniques: **M-CAN, M-Chord**



# Image search



# Images and their Descriptors



# CoPhIR: Content-based Photo Image Retrieval

**100M**

**images + metadata + MPEG-7 VDs**

http://cophir.isti.cnr.it/

- Largest publicly available collection of high-quality images metadata: **106 million images**
- Each image contains:
  - Five MPEG-7 VDs: Scalable Color, Color Structure, Color Layout, Edge Histogram, Homogeneous Texture
  - Other textual information: title, tags, comments, etc.
- Photos have been crawled from the **Flickr** photo-sharing site.

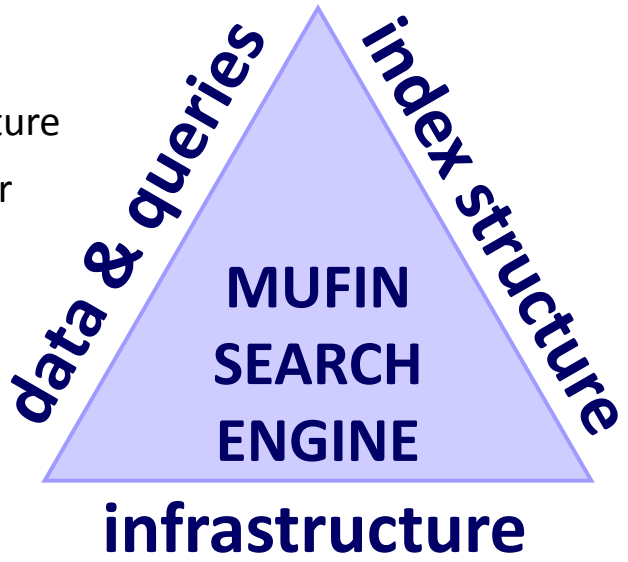
# Image Search Demo

**http://mufin.fi.muni.cz/imgsearch/**

## Extensibility

COPHIR

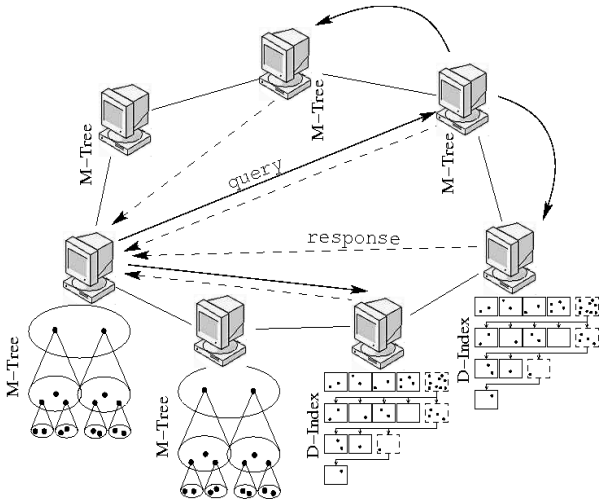
- color structure
- scalable color
- color layout
- edge histogram
- homogeneous texture




6 x IBM server x3400

## Scalability

M-Chord + M-Tree

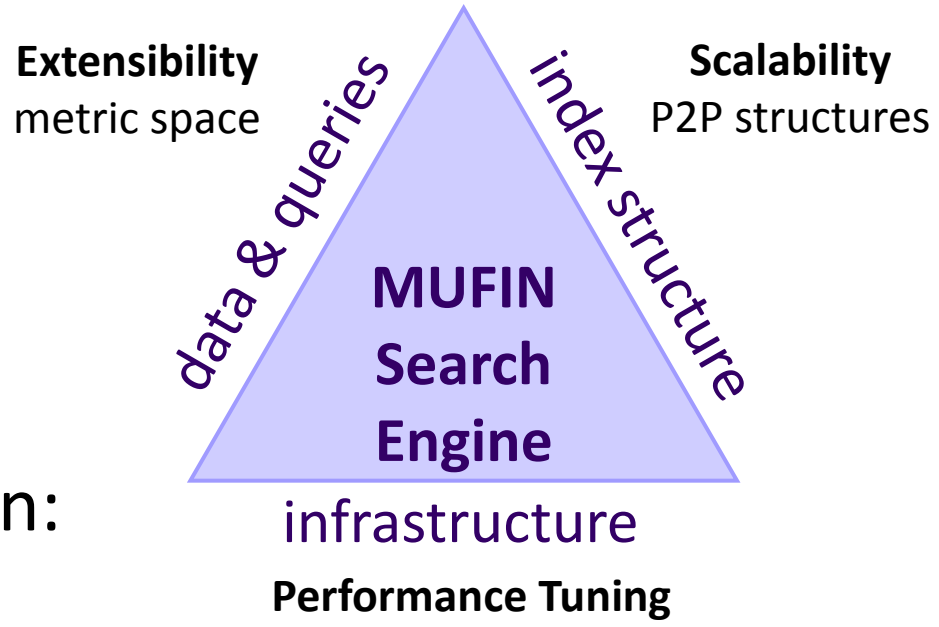


# demos

- http://mufin.fi.muni.cz/apps.html

# MUFIN Trends Summary

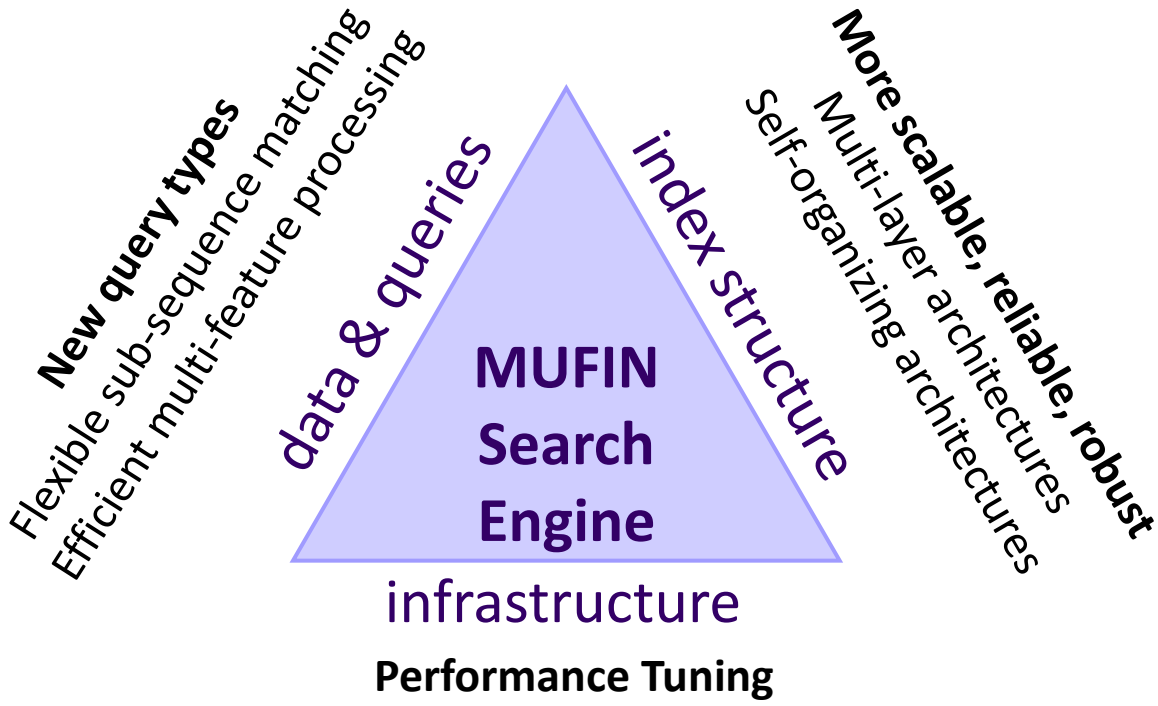
- MUFIN - a universal similarity search technology




- Research directions in:
  - Core technology
  - Applications
  - A style of computing

# Core Technology

- Development of the MUFIN core technology



# Applications

- Images:
  - Sub-image retrieval
  - Ranking
  - Annotation
  - Categorization
  - Benchmarking
- Biometrics:
  - Face recognition
  - Fingerprint recognition
  - Gait recognition
- Signals:
  - Audio recognition
  - Time series similarity
- Videos:
  - Event detection

# A New Style of Computing

- From the project-oriented approach towards similarity cloud
- Advantages:
  - Cloud makes similarity search accessible to common users
  - Computational resources are shared – users don't need to maintain any hardware infrastructure
  - Users don't need to care for the OS, security, software platform, etc.

# Current Research Activities

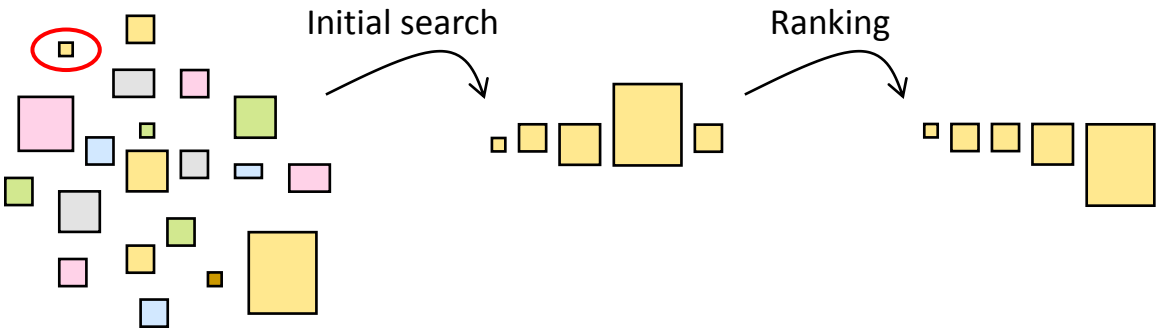
- Image Query Postprocessing
- Sub-image Searching
- Remote Biometrics
- Event Detection in Video
- Signal Processing

# Query Postprocessing

- The understanding of similarity is:
  - subjective
  - context-dependent
  - multi-modal
- Semantic gap
- Overcoming semantic gap by combining aspects
  - semantics-learning
  - result postprocessing
  - relevance feedback & iterative search
- Our objectives
  - Large general data collections with various quality of metadata
  - Online searching response times

# Query Postprocessing by Ranking

- Two-phase query evaluation model
  - Search the whole collection by some aspects => candidate set
  - Rank the candidate set – sort by other aspects




☺ Advantages

- Fast, enables to combine more similarity measures
- Enables cooperation with user

☹ Disadvantages

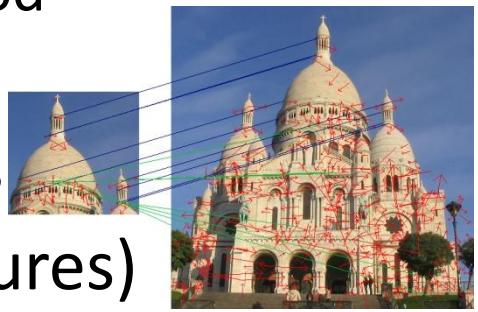
- Only a subset of the whole dataset is used in the ranking phase

# Sub-image Searching

- Retrieves all images containing the query image

Query Answer:

- Based on local image descriptors
  - Scale Invariant Feature Transform (SIFT):
    - Descriptor – content of a small neighborhood
    - Locator – coordinates of the neighborhood
    - Scale – importance of the descriptor
  - Image $\Rightarrow$ a set of features, descriptors
  - Task: Find matching pairs (similar features)



# Remote Biometrics: Motivation

- Most biometrics require the subject's cooperation
  - Fingerprint, iris, palmprint, handwriting, voice recognition

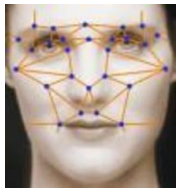


- Challenge – recognizing people at a distance
  - Capture devices do not require a close contact with the subject (e.g., surveillance cameras)
    - It can be applied unobtrusively
  - Face and gait recognition at a distance
  - Problems – camera view, lighting, pose
  - Applications – surveillance, security

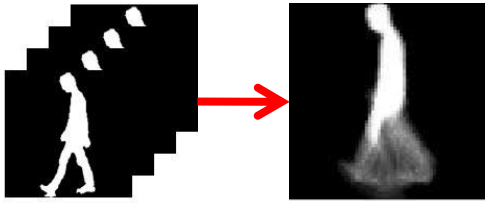
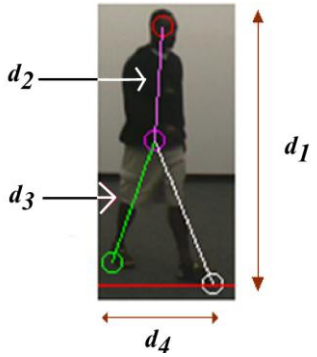

# Remote Biometrics: Approaches

- Detection, normalization, extraction, recognition
- Face recognition
  - Methods:
    - Appearance-based – analyze the face as a whole
    - Model-based – compare individual features (e.g., eyes, mouth)
  - MUFIN face recognition demo: http://mufin.fi.muni.cz/faces-feret/

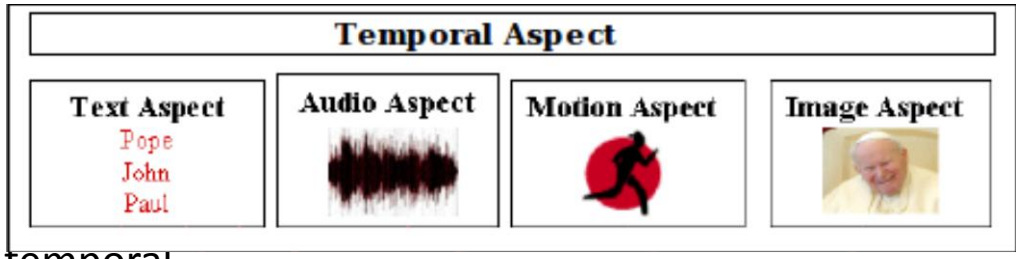




- Gait recognition
  - Less likely to be obscured, low resolution suffices
  - Methods are based on shape or dynamics of the person:
    - Appearance-based – analyze person's silhouettes
    - Model-based – compare features (e.g., trajectory, angular velocity)

# Event Detection in Video

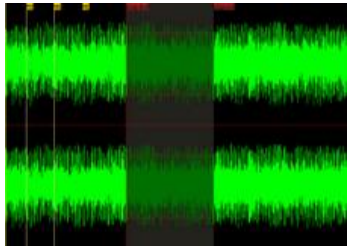



- Video
  - continuous data
  - several aspects
    - image, sound, text, motion, temporal
- Event
  - defined aspects occurring in given time interval
    - definition of a sample aspect by example or value
    - definition is imprecise – looking for “similar” aspects
  - combination of aspects
    - aggregation function
- Current approaches
  - annotation-based, learning-based (classifiers)
  - specific domains

**Example**
TV news (by image) AND about IRAQ (by text) AND burning vehicles (by image) AND time interval < 1 minute (by temporal)

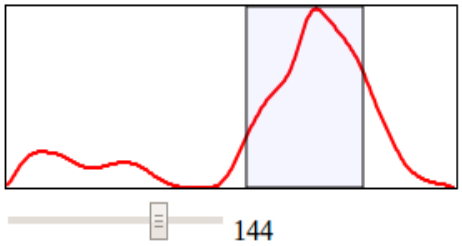
# Signal Processing

- Vast amount of signals produced:
  - Biomedicine data – ECG, CT
  - Biometric data – personal identification
  - Audio data – audio similarity, recognition
  - Sub-image searching
  - Financial time series – analysis, forecasting
  - Time series streams
- Demand for
  - a graceful handling of this data
  - flexible reactions to new application needs

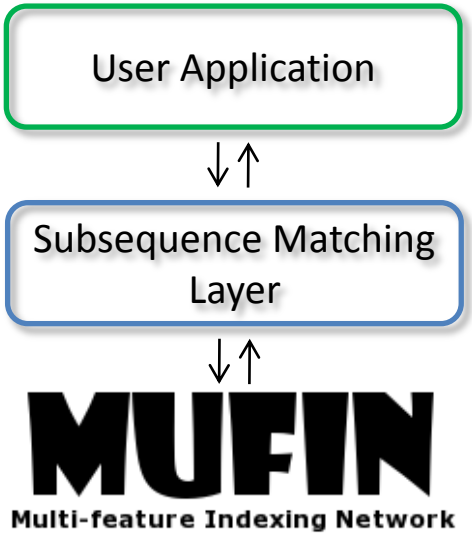


# Flexible Subsequence Matching

- Generic engine for rapid development of subsequence matching applications
  - can be used for any class of one-dimensional signals
  - Implementation of various subsequence matching approaches
  - Demo web application







# Demo application

# Face Retrieval Application

- 10,000 images with people
- 14,000 faces
- Face detection – MPEG7