Evolution of similarity search in databases

… from trivial applications to real problems

Tomáš Skopal
Siret Research Group, KSI MFF UK

3/2012

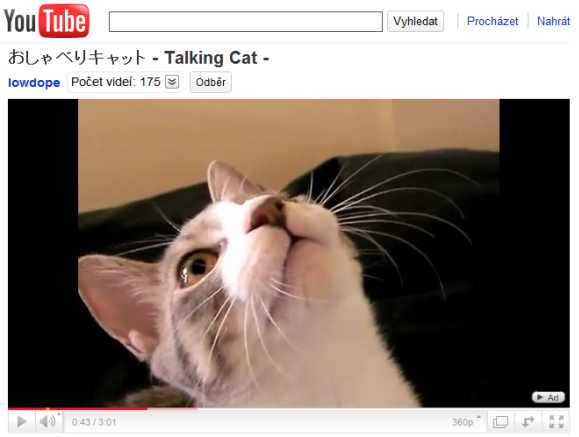
siret
research group

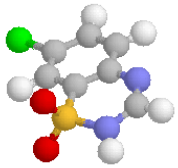
Talk outline

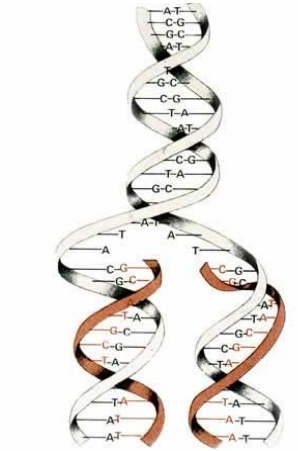
- the problem of similarity search
- similarity models
  - early models (vector space)
  - current models (metric space)
  - future models (non-metric models)
- applications
  - data-centric (multimedia search engines)
  - service-centric ("enterprise" applications)

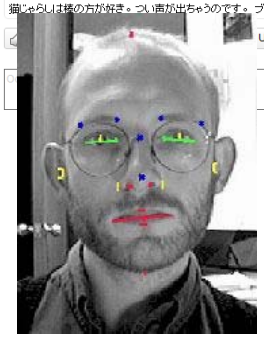
Similarity search

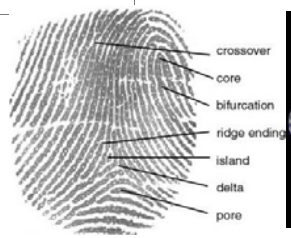
- task
  - how to search non-structured data based on content?
    - we cannot use traditional approaches (e.g., linear order, SQL)

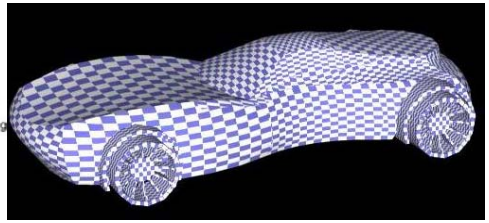


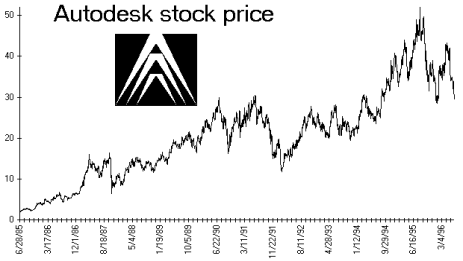








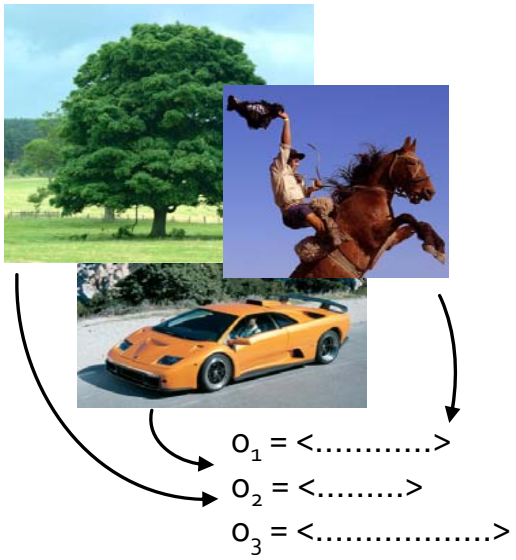


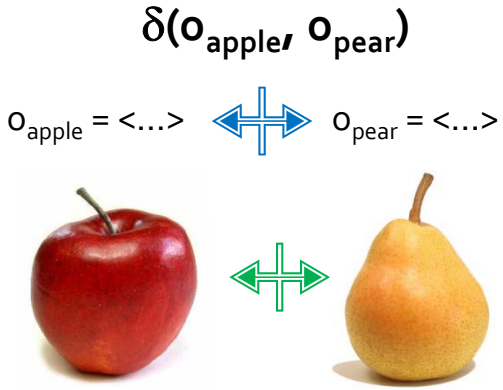


T.Skopal, Evolution of similarity search in databases

Similarity search

- task
  - how to search non-structured data based on content?
    - we cannot use traditional approaches (e.g., linear order, SQL)
  - similarity model
    - descriptor universe (feature extraction) $\mathbf{o_i \in U}$
    - pair-wise similarity/distance function $\boldsymbol{\delta(x,y)}$



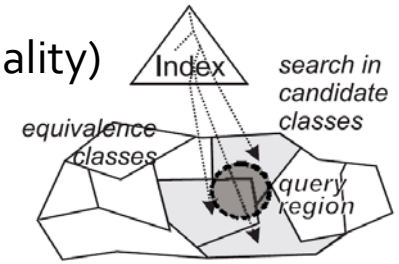


Similarity search

- task
  - how to search non-structured data based on content?
    - we cannot use traditional approaches (e.g., linear order, SQL)
  - similarity model
    - descriptor universe (feature extraction) $\mathbf{o_i \in U}$
    - pair-wise similarity/distance function $\boldsymbol{\delta(x,y)}$
  - retrieval
    - querying (range query, kNN query)
    - exploration (and other modern means of retrieval)

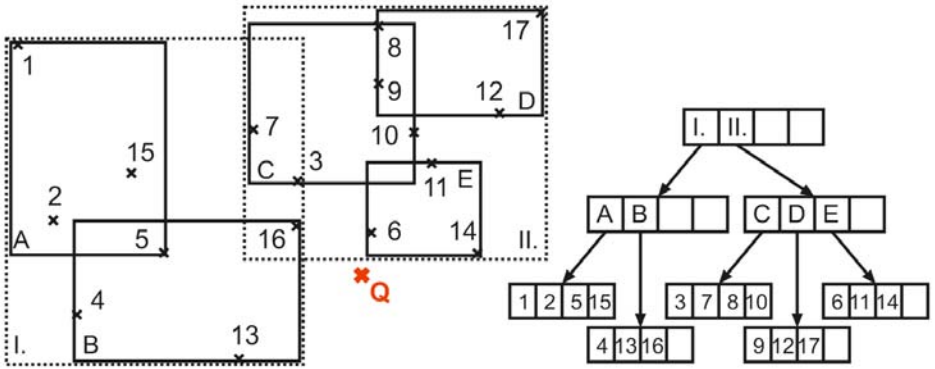
Similarity search

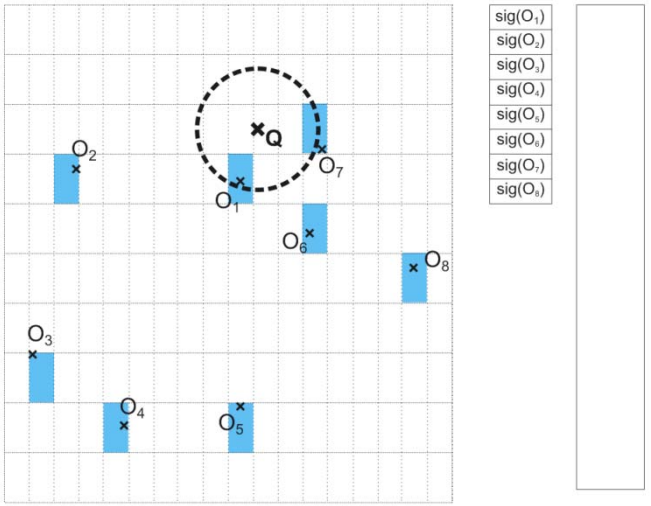
- task
  - how to search non-structured data based on content?
    - we cannot use traditional approaches (e.g., linear order, SQL)
  - similarity model
    - descriptor universe (feature extraction) $\mathbf{o_i} \in \mathbf{U}$
    - pair-wise similarity/distance function $\delta(\mathbf{x},\mathbf{y})$
  - retrieval
    - querying (range query, kNN query)
    - exploration (and other modern means of retrieval)
- database solution
  - efficiency vs. effectiveness (performance vs. quality)
    - assumption: sequential scan is too slow
      I/O cost but also (mainly) CPU cost of $\delta(\mathbf{x},\mathbf{y})$



Early models – vector models

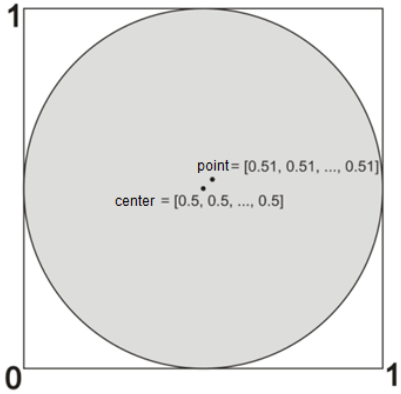
- similarity modeled in vector space
  - reuse of spatial indexes designed for GIS, CAD/CAM, etc.
  - R-tree, X-tree, SS-tree, Grid file, VA-file, inverted index, etc.

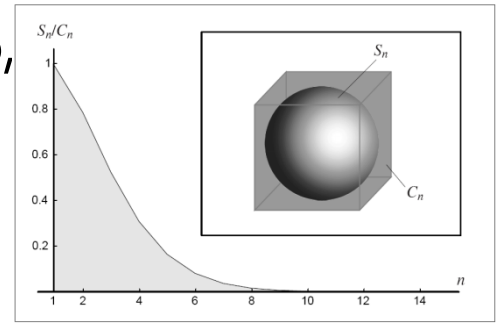




Early models – vector models

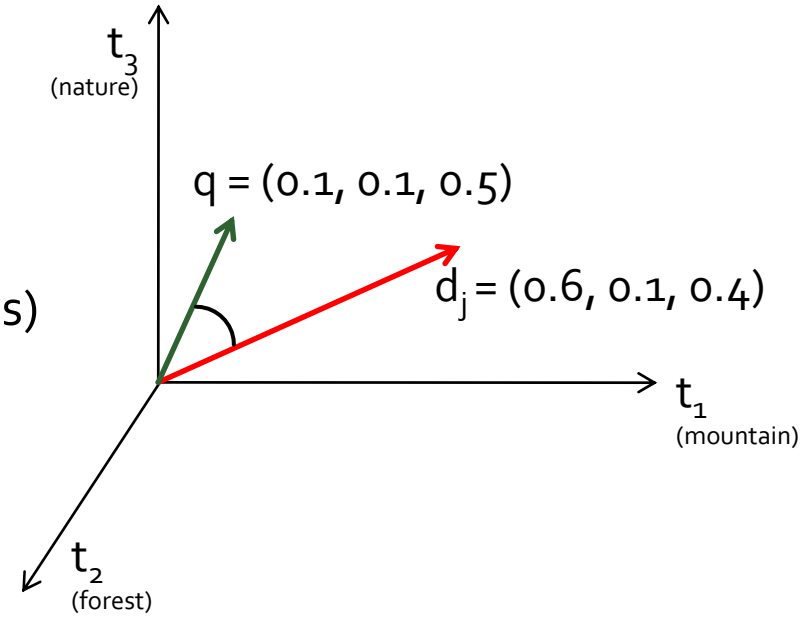
- similarity modeled in vector space
  - reuse of spatial indexes designed for GIS, CAD/CAM, etc.
  - R-tree, X-tree, SS-tree, Grid file, VA-file, inverted index, etc.
- cons
  - similarity modeling limited to L_p spaces
    - descriptors = numeric vectors, similarity = L_p metric
    - independent dimensions („smart" descriptors, „stupid" similarity)
  - curse of dimensionality
    - inefficient for high-dimensional data, say dim>10, (not the case of original use in GIS/CAD, i.e., 2D, 3D data!)





Early models – vector models

- suitable application in similarity search – vector query (originated in Information retrieval)
  - inverted index
  - specific requirements
    - cosine measure/distance
    - sparse vectors (query and data)
  - applications
    - text retrieval (dimensions are terms)
    - **bag-of-words models** in multimedia retrieval

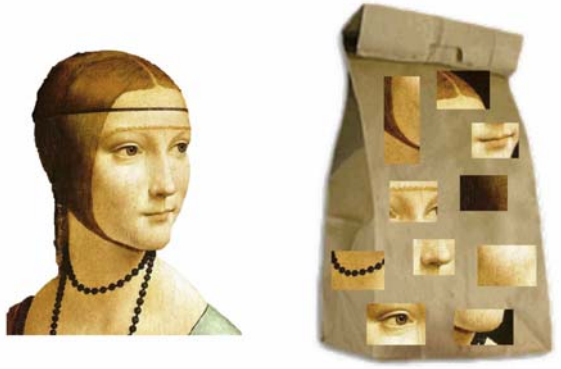


Early models – vector models

- bag-of-words in image retrieval

1. extract features (from many images)

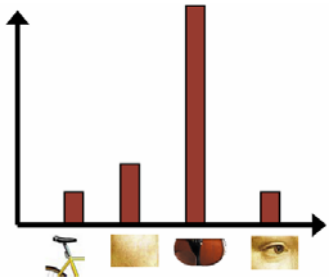
2. quantize/cluster features



credit: Fei-Fei Li at al., 2005

3. learn visual vocabulary

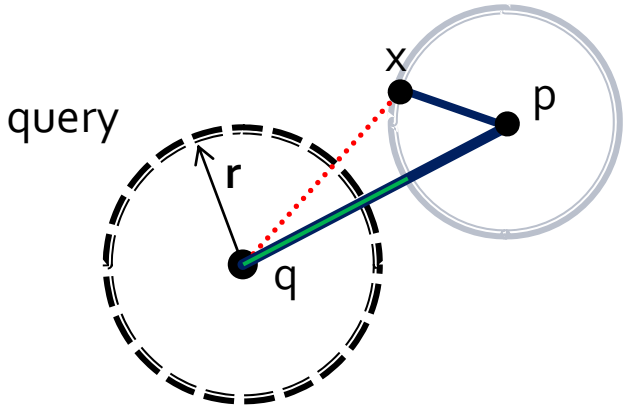
4. represent images by histograms (vectors)



Metric space model

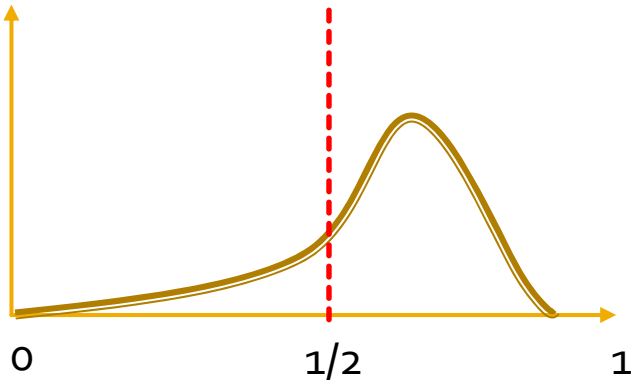
- more general than vector models
  - any metric space
    - black-box descriptor and black-box distance function δ
    - **the distance must be a metric (axioms)**
    - "lowerbounding" using pivots (triangle inequality)

$$\delta(x, y) + \delta(y, z) \geq \delta(x, z),$$



Metric space model

- more general than vector models
  - any metric space
    - black-box descriptor and black-box distance function δ
    - **the distance must be a metric (axioms)**
    - "lowerbounding" using pivots (triangle inequality)
  - curse of dimensionality reduced to some extent
    - generalized problem of **intrinsic dimensionality** (dist. distrib. matters)



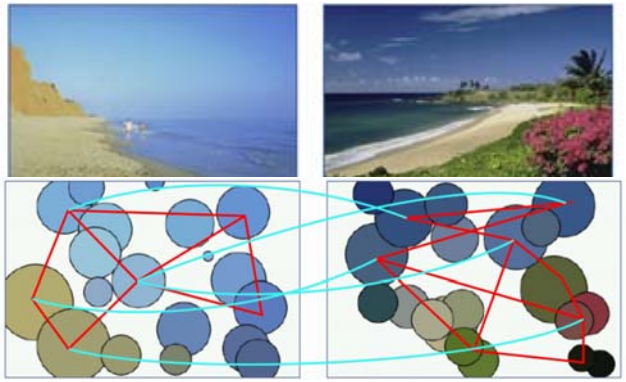
Metric space model

- more general than vector models
  - any metric space
    - black-box descriptor and black-box distance function δ
    - **the distance must be a metric (axioms)**
    - "lowerbounding" using pivots (triangle inequality)
  - curse of dimensionality reduced to some extent
    - generalized problem of **intrinsic dimensionality** (dist. distrib. matters)
  - many metric indexes proposed
    - main-memory/persistent, serial/parallel/distributed, static/dynamic, etc.
    - attempts extending SQL by similarity predicates + implementations
- for real problems often still not sufficient
  - metric axioms limit the modeling capabilities

Metric space model

- suitable application
  - image feature signatures + SQFD

$$\mathrm{SQFD}_{f_s}(S^q, S^p) = \sqrt{(w_q \mid -w_p) \cdot A_{f_s} \cdot (w_q \mid -w_p)^T}$$

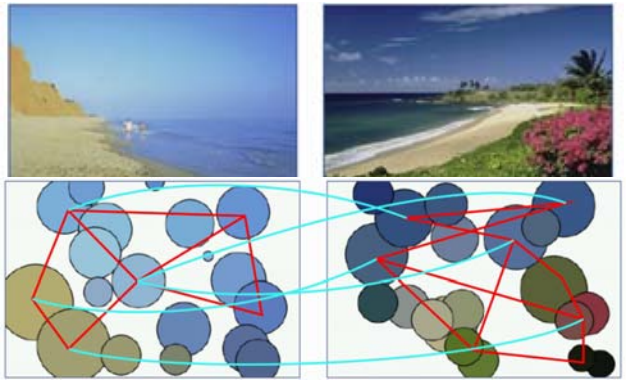


M. Kruliš, T. Skopal. J. Lokoč, Ch. Beecks. Combining CPU and GPU Architectures for Fast Similarity Search, **Distributed and Parallel Databases**, Springer, 2012
M. Kruliš, J. Lokoč, Ch. Beecks, T. Skopal, T. Seidl. Processing Signature Quadratic Form Distance on Many-Core GPU Architecture, **ACM CIKM 2011**, Glasgow, UK
Ch. Beecks, J. Lokoč, T. Seidl, T. Skopal. Indexing the Signature Quadratic Form Distance for Efficient Content-Based Multimedia Retrieval, **ACM ICMR 2011**, Trento, Italy

Metric space model

- suitable application
  - image feature signatures + SQFD

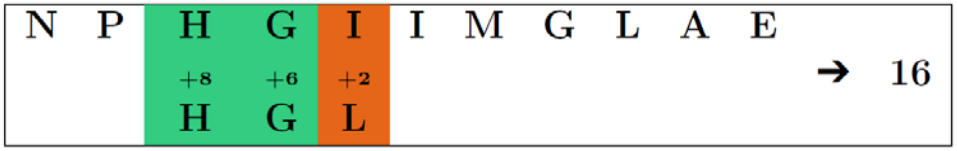
$$\mathrm{SQFD}_{f_s}(S^q, S^p) = \sqrt{(w_q \mid -w_p) \cdot A_{f_s} \cdot (w_q \mid -w_p)^T}$$



  - „mistyping vocabulary" using edit distance

δ_{edit}('bank','bar') = 2 δ_{edit}('tank','bank') = 1

- bad application
  - indexing of protein sequences using edit distance
    - edit distance too simple for biologic applications
    - extension needed like scoring matrices, gap penalties, local alignment
      - e.g., Smith-Waterman



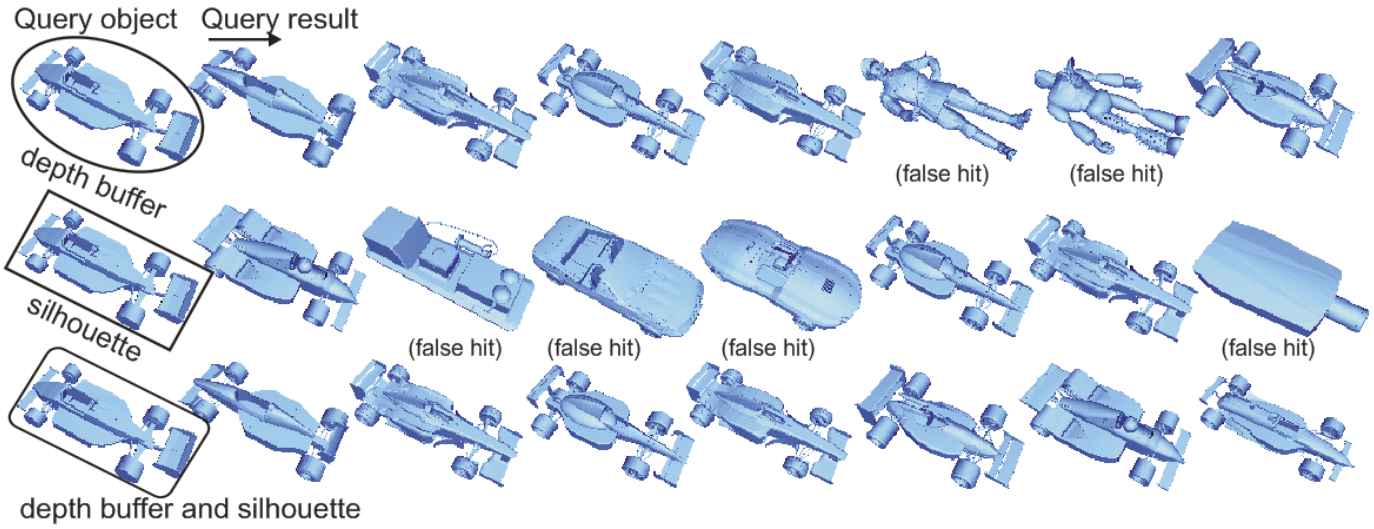
Non-metric similarity models

- absence of some metric axiom = non-metric function
- similarity parameterization
  - inclusion of user preferences/context at query time
- non-metric similarities "by design"
  - robust behavior (ignoring „noisy" parts)
  - local behavior (favoring similarity)
  - comfort of black-box approach (domain experts outside CS)

T. Skopal, B. Bustos. On Nonmetric Similarity Search Problems in Complex Domains, **ACM Computing Surveys**, 43(4):34:1-34:50, 2011
B. Bustos, T. Skopal. Nonmetric Similarity Search Problems in Very Large Collections, **ICDE 2011**, Hannover, Germany, IEEE

Non-metric similarity models

- need to allow explicit preferences/user profile at query time
  - dynamic function δ(**x, y,** *other parameters*)
    - i.e., the same descriptors **x,y** lead to different similarity values (so it is not even a function, not yet metric distance)



Non-metric similarity models

- possible solution
  - multi-metric approach, linear combination of metrics
  - query time weighting

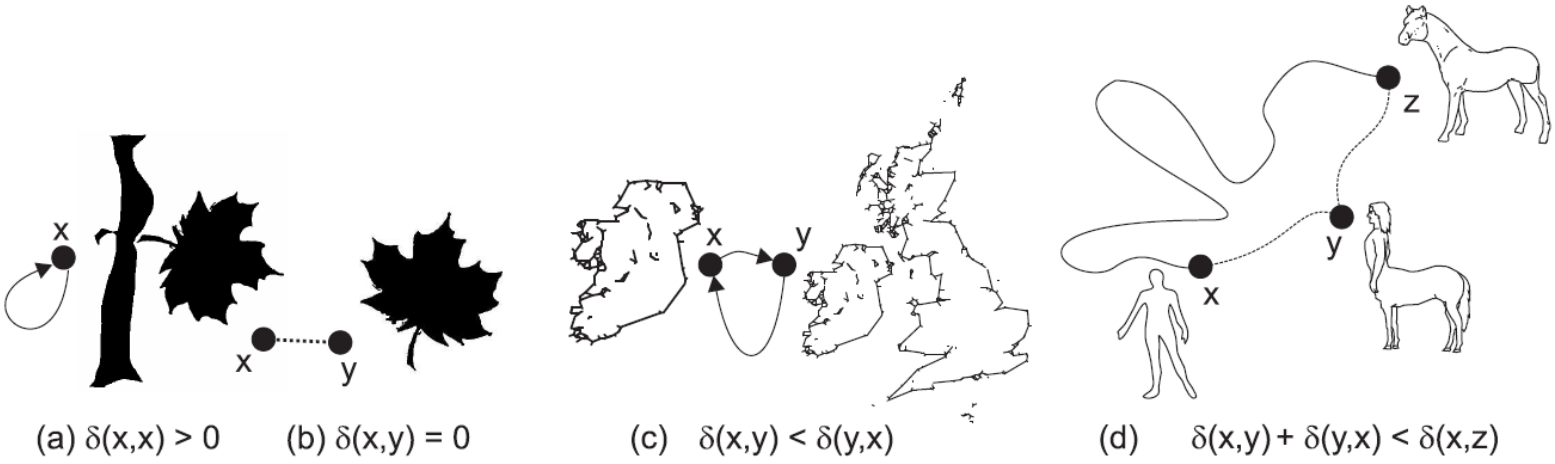
$$\Delta_{\mathbb{W}}(O_1, O_2) = \sum_{i=1}^{m} w_i \cdot \delta_i(O_1, O_2)$$

  - multi-metric indexes (adaptation of metric ones, e.g., M³-tree)
    - storage of distance components,
      i.e., final aggregation with weights at query time

B. Bustos, S. Kreft, T. Skopal. Adapting Metric Indexes for Searching in Multi-Metric Spaces, **Multimedia Tools and Applications**, Springer, 2012
B. Bustos, T. Skopal. Dynamic Similarity Search in Multi-Metric Spaces, **ACM MIR 2006**, Santa Barbara, CA, USA

Non-metric similarity models

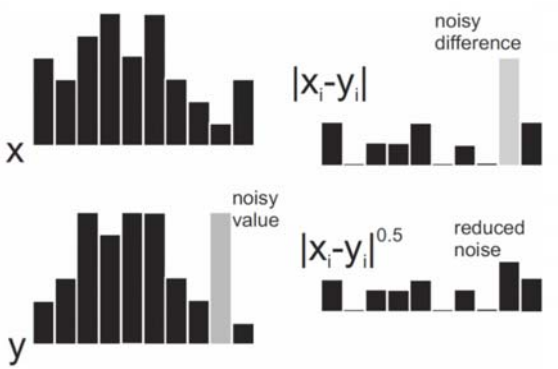
- arguments against metric axioms in theory (psychology, cognitive sciences)

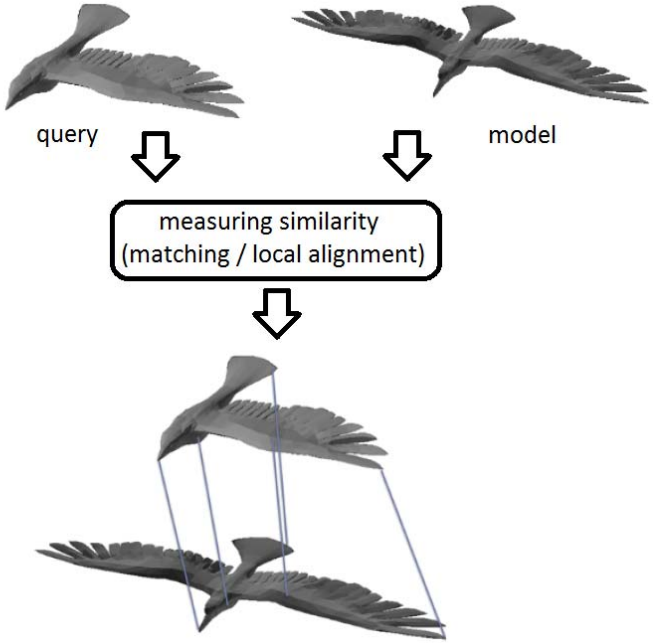


(a) $\delta(x,x) > 0$ (b) $\delta(x,y) = 0$ (c) $\delta(x,y) < \delta(y,x)$ (d) $\delta(x,y) + \delta(y,x) < \delta(x,z)$

Non-metric similarity models

- and practically…






```
if (condition) {
  statements;
}
else if (condition) {
  statements;
}
else {
  statements;
}
if (x == 3)   // curly braces not needed
  flag = 1;   // when if statement is
else          // followed by only one
  flag = 0;   // statement
```

Repetition (while)

```
while (expression) { // loop until
  statements;      // expression is false
}
```

Repetition (do-while)

```
do {            // perform the statements
  statements; // as long as condition
} while (condition); // is true
```

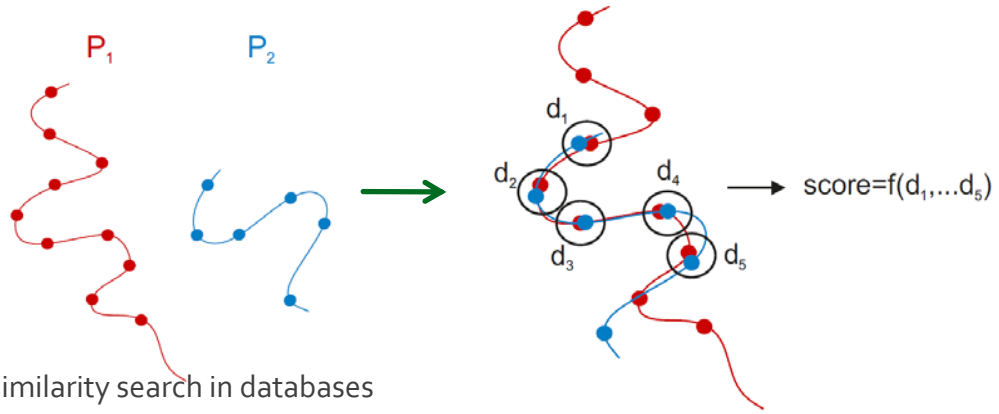
Repetition (for)

init - initial value for loop control variable
condition - stay in the loop as long as condition is true
increment - change the loop control variable

```
for(init; condition; increment) {
  statements;
}
```

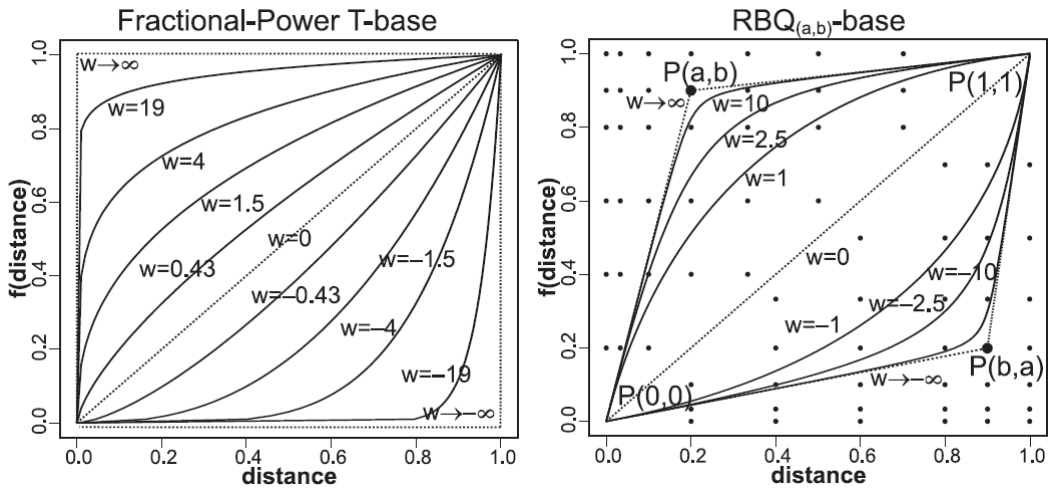
Bifurcation (break, continue, goto, exit)

```
break; // ends a loop
continue; // stops executing statements
// in current iteration of
```



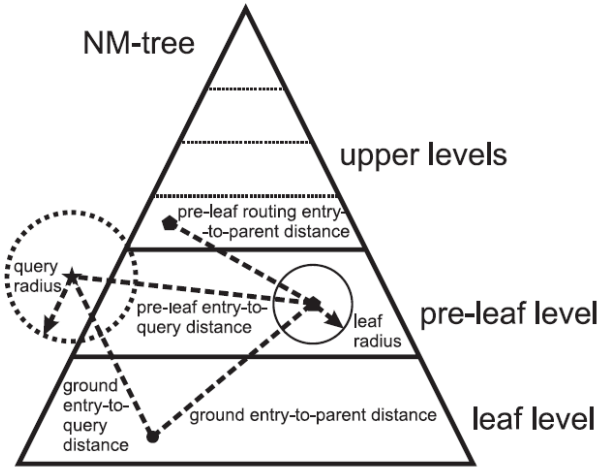
Non-metric similarity models

- general indexing of non-metric distances
  - TriGen algorithm
    - modification of semimetric distance δ to (partially) fulfill triangle inequality (T-error)
    - the lower T-error, the more precise but slower search, and vice versa



Non-metric similarity models

- general indexing of non-metric distances
  - TriGen algorithm
    - modification of semimetric distance δ to (partially) fulfill triangle inequality (T-error)
    - the lower T-error, the more precise but slower search, and vice versa
  - NM-tree (index+TriGen, multiple precisions at query time)



T.Skopal, Evolution of similarity search in databases

Non-metric similarity models

- general indexing of non-metric distances
  - TriGen algorithm
    - modification of semimetric distance δ to (partially) fulfill triangle inequality (T-error)
    - the lower T-error, the more precise but slower search, and vice versa
  - NM-tree (index+TriGen, multiple precisions at query time)
- pros – universal non-metric index
- cons – might lead to large intrinsic dimensionality

T. Skopal, J. Lokoč. NM-tree: Flexible Approximate Similarity Search in Metric and Non-metric Spaces, **DEXA 2008**, Turin, Italy, LNCS 5181, Springer
T. Skopal. Unified Framework for Fast Exact and Approximate Search in Dissimilarity Spaces, **ACM Transactions on Database Systems**, 32(4):1-47, 2007
T. Skopal. On Fast Non-Metric Similarity Search by Metric Access Methods, **EDBT 2006**, Munich, Germany, LNCS 3896, Springer

Non-metric similarity models

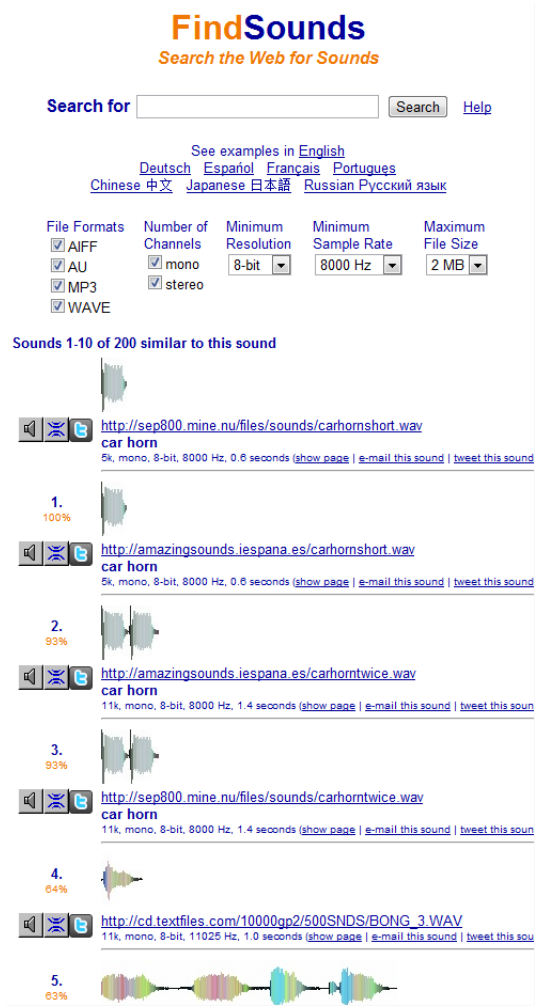
- applications for non-metric similarity search in bioinformatics
  - mass spectra interpretation (protein identification)
  - protein identification based on (3D) structure
  - RNA identification based on (3D) structure

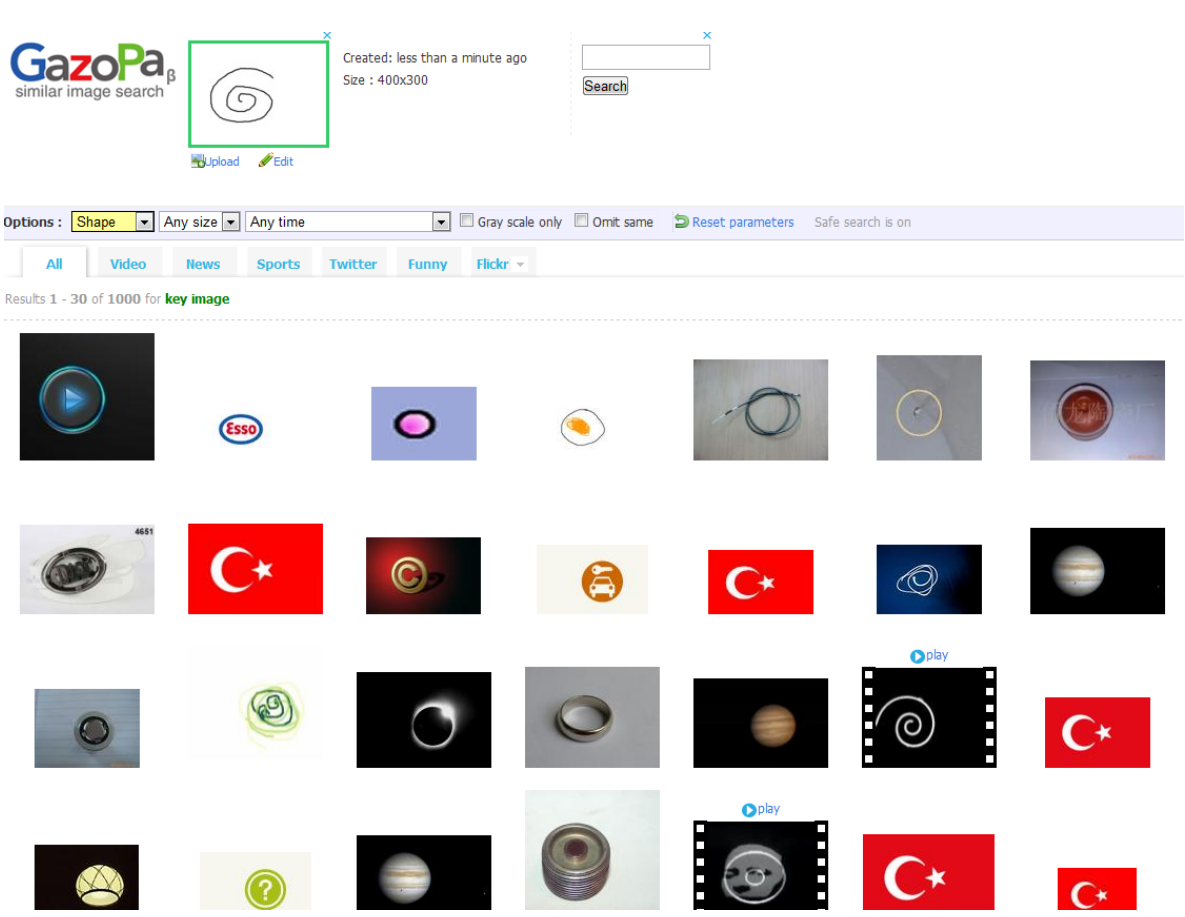
J. Novák, T. Skopal, D. Hoksza, J. Lokoč. Non-metric Similarity Search of Tandem Mass Spectra Including Posttranslational Modifications, **Journal of Discrete Algorithms**, Elsevier, 2012
J. Galgonek, D. Hoksza, T. Skopal. SProt: sphere-based protein structure similarity algorithm, **Proteome Science**, 9(Suppl 1):S20, 2011
D. Hoksza, D.Svozil. SETTER - RNA SEcondary sTructure-based TERtiary Structure Similarity Algorithm, **ISBRA 2011**, Changsha, China, Springer

Data-centric applications

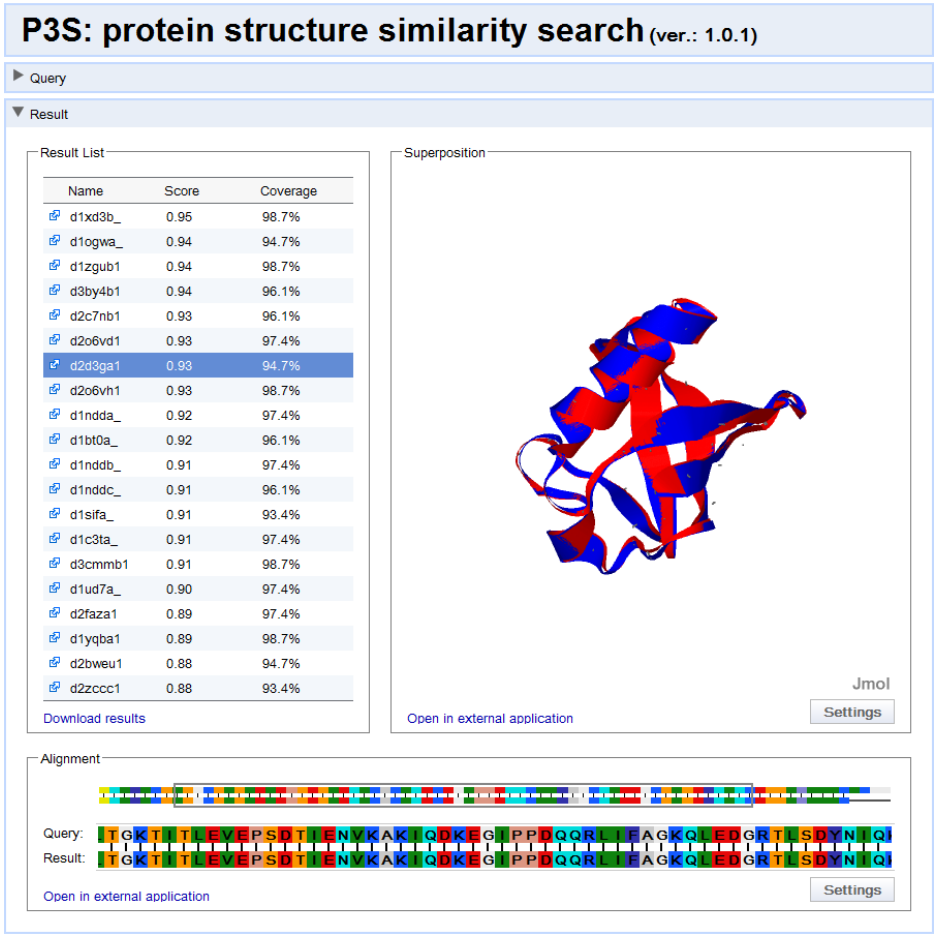
- similarity search as "exposed technology"
  - querying mechanism is the main functionality
  - the user must cooperate
    (e.g., formulate a query, or browse/explore)
  - today mostly desktop applications
- data is the final product
  - is this the "killer application" for similarity search?
    …depends

Data-centric applications





Data-centric applications



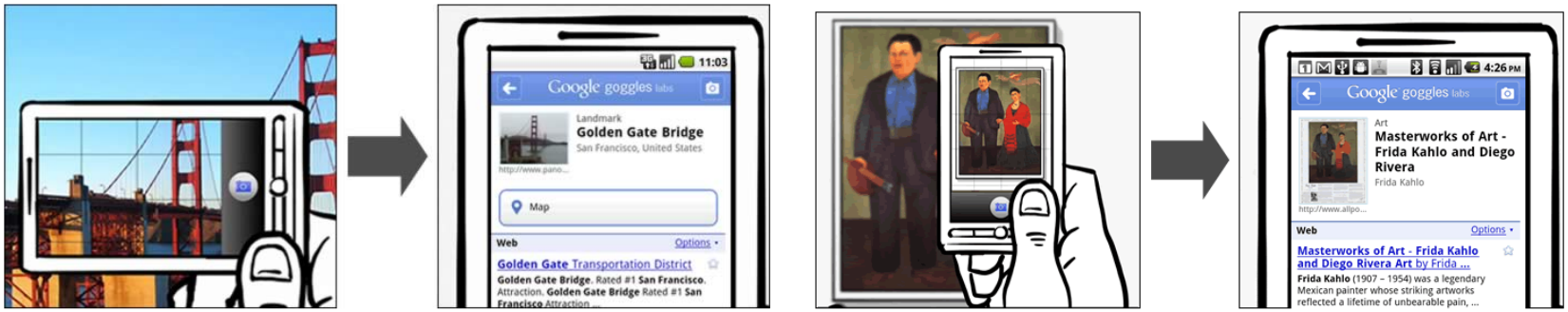
T.Skopal, Evolution of similarity search in databases

Service-centric applications

- similarity search as "hidden technology"
  - querying mechanism and even the data type hidden to the user
  - data is just means of search
- service is the final product
  - much wider applicability of similarity search?
  - e.g., e-commerce, "audiovisual wikipedia", etc.

Service-centric applications

- Google Goggles – mobile version of Google
  - combination of OCR and pattern matching (similarity-based) techniques and content-based image retrieval
  - Android, iOS (iPhone, iPad)
- nowadays provides identification of:



Service-centric applications

- augmented reality
  - very near future