Multimedia Similarity Search (Tutorial)

Prof. Dr. Thomas Seidl[1], Dr. Christian Beecks[2], Dr. Seran Uysal[2]

[1] LMU München, Lehrstuhl für Datenbanksysteme und Data Mining
[2] RWTH Aachen, Lehrstuhl für Informatik 9 (Datenmanagement und –exploration)

06.03.2017, BTW 2017, Stuttgart

What is this tutorial about?

1. **Object representations**
   - *How to model and represent multimedia data?*

2. **Fundamental similarity models for multimedia data**
   - *How do distance-based similarity models look like?*

3. **Efficient query processing**
   - *How to process distance-based similarity queries efficiently?*

4. **Indexing**
   - *How to index spatial and high-dimensional multimedia data?*
   - *What are the principles behind metric and Ptolemaic indexing approaches?*

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Tutorial Outline

1) Object Representation
   - Feature Extraction and Representation
   - Feature Aggregation

2) Fundamental Similarity Models
   - Dissimilarity Measures
   - Distance Functions for Feature Histograms
   - Distance Functions for Feature Signatures

3) Efficient Query Processing
   - Similarity Queries
   - Lower-Bounding: 2 examples

4) Indexing
   - Spatial Indexing
   - Metric and Ptolemaic Indexing

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Explosive Growth of Multimedia Data

- **4.5 million photos** are uploaded to Flickr **every day**
  http://advertising.yahoo.com/article/flickr.html

- **300 million images** are uploaded to Facebook **every day**
  https://developers.facebook.com/blog/post/2012/07/17/capturing-growth--photo-apps-and-open-graph/

- **100 hours of video** are uploaded to YouTube **every minute**
  http://www.youtube.com/yt/press/statistics.html

- **50 million tweets** are uploaded to Twitter **every day**
  https://blog.twitter.com/2011/numbers

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Content-based Information Access

- **Multimedia retrieval** is about the search for knowledge in all its forms, everywhere [Lew et al., 2006].
- **Goal** is to find multimedia objects of interest

→

- **Content-based multimedia retrieval**:
  - Utilization of automatically extracted features
  - Additionally, meta data can be taken into account
  - ➢ **Necessity of efficient similarity search techniques**

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Many Application Areas

- **Retrieval**
  - image, video, audio, music, tweet, text search
- **Content analysis**
  - copy, duplicate, near-duplicate detection
- **Mining**
  - classification, clustering, associations
- **Browsing** and **exploring**
- **etc.**

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Multimedia Information Retrieval – Similarity Search

- Task
  - given a query image, retrieve the most similar objects

- Variants
  - ε-range query
  - k-nearest neighbor query
  - Ranking enumeration query (give-me-more)

- Evaluation measures
  - Recall: how many of the desired objects in the database are retrieved (fraction)?

$$recall = \frac{retrieved\ and\ desired}{desired}$$

  - Precision: how many of the retrieved objects are desired ones (fraction)?

$$precision = \frac{retrieved\ and\ desired}{retrieved}$$

  - F1-Measure: harmonic mean of recall and precision

$$f_1 = \frac{1}{\frac{1}{2}\left(\frac{1}{recall} + \frac{1}{precision}\right)} = \frac{2 \cdot recall \cdot precision}{recall + precision}$$

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Clustering

- Task
  - Group similar objects while separating dissimilar ones

- Variants, algorithms
  - Partitioning clustering: (k-means), k-medoid
  - Hierarchical clustering: agglomerative (single link, complete link), divisive
  - Density-based clustering: DBSCAN

- Evaluation measures
  - External measures: try to retrieve known clusters
    - Require „Ground truth“, or expert knowledge
  - Internal measures: measure cluster coherence and separation
    - Example: Silhouette coefficient

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Classification

- Task
  - Training phase: Based on a given training data set with class labels, learn a classifier
  - Application phase: predict class label for unknown query objects

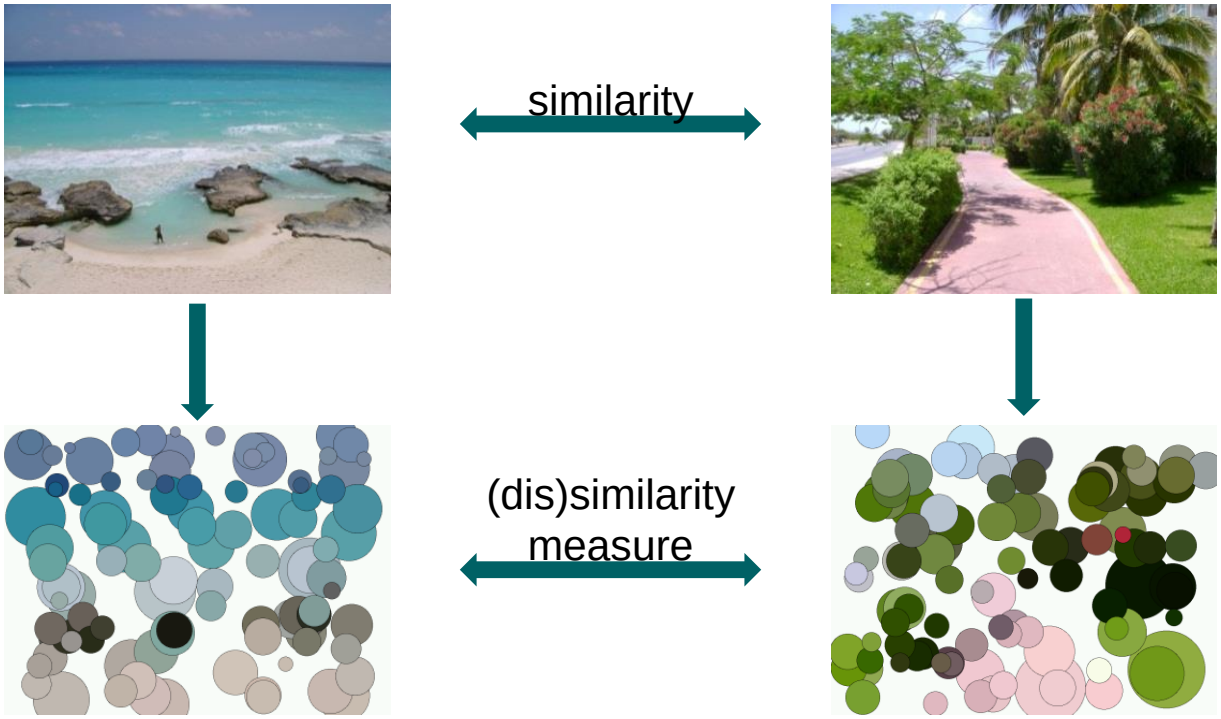
- Variants, algorithms
  - Some require just a similarity measure
    - k-nearest neighbor classifier
    - Bayesian classifiers
  - Others require more, e.g. attribute structure
    - Kernel SVM, neural networks, decision trees

- Evaluation measures
  - Classification accuracy or classification error (complement each other)
  - Over-generalization vs. Overfitting – e.g., model size

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

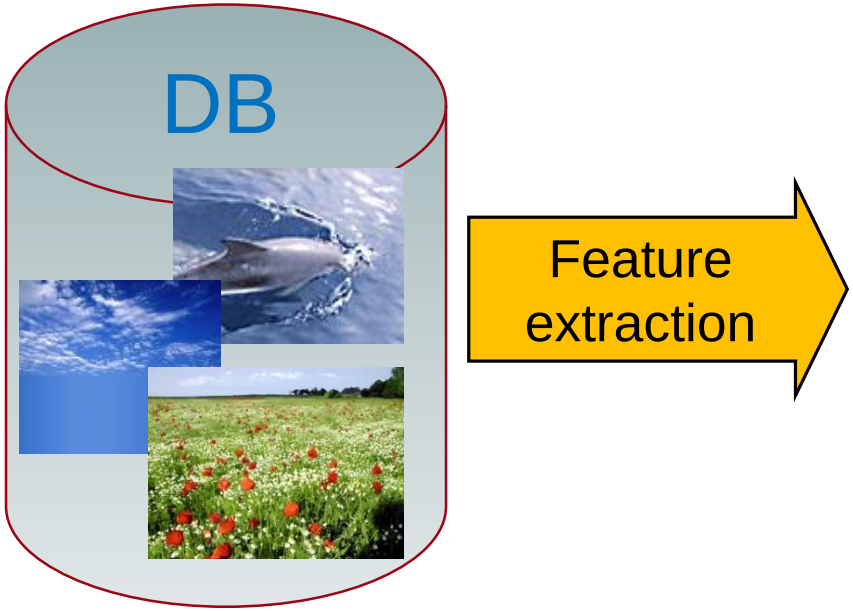
Similarity Model

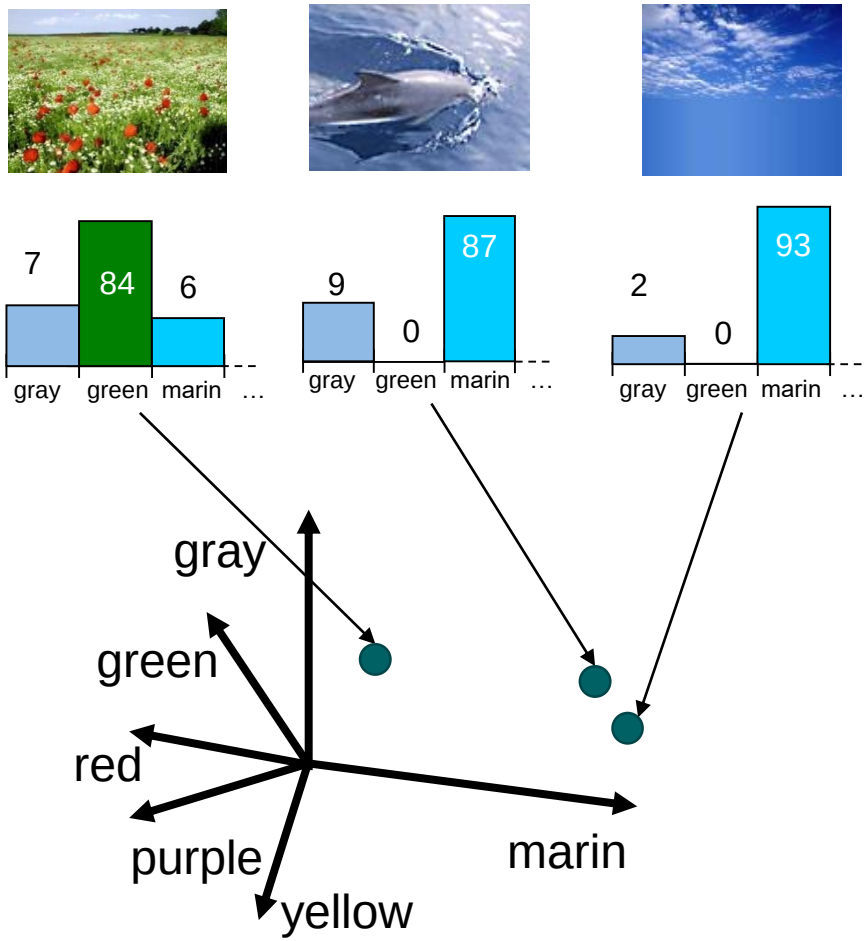
- A **similarity model** formalizes the notion of (dis)similarity
- The models are (regularly) tightly bound to specific object **representations**
  - In this tutorial, we focus on *feature histograms* and *feature signatures*



Feature-based Similarity

- Feature-space embedding





- Feature-based similarity
  - Points represent objects
  - Distance corresponds to (dis-)similarity

Tutorial Outline

1) **Object Representation**
   - Feature Extraction and Representation
   - Feature Aggregation

2) Fundamental Similarity Models
   - Dissimilarity Measures
   - Distance Functions for Feature Histograms
   - Distance Functions for Feature Signatures

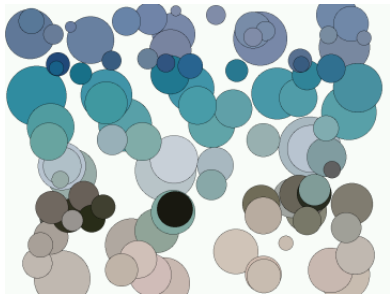
3) Efficient Query Processing
   - Similarity Queries
   - Lower-Bounding: 2 examples

4) Indexing
   - Spatial Indexing
   - Metric and Ptolemaic Indexing

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Feature-based similarity

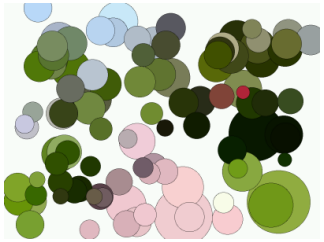
- Process of extracting features from multimedia data objects
  - A **feature** is a mathematical description of an inherent property of a multimedia data object, often in a Euclidean space $\mathbb{R}^d$



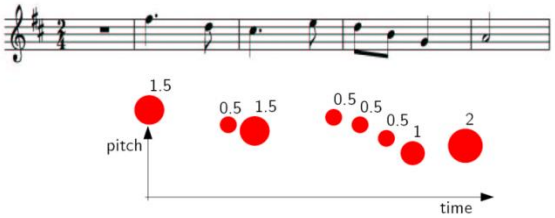
- Different types of features:
  - **Global features** describe a multimedia data object as a whole
  - **Local features** describe parts of a multimedia data object
- Different semantics of features:
  - **High-level features** such as concepts, tags, etc.
  - **Low-level features** such as color, texture, …, shapes, etc.

Features

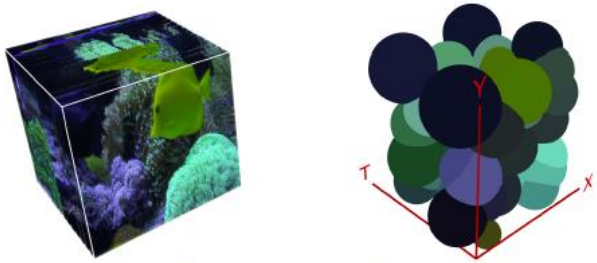
- **Features** describe inherent properties of multimedia data objects
  - Image features [DJL+08], e.g.:
    - Color
    - Texture
    - Shape



  - Audio features [MZB10, CVG+08], e.g.:
    - Pitch
    - Loudness



  - Video features [HXL+11], e.g.:
    - key-frame features
    - object features
    - motion features



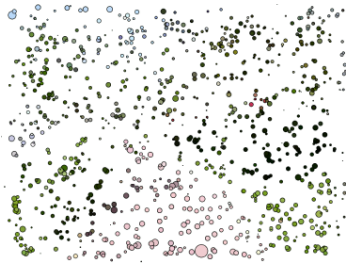
(a) video clip (b) video signature

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Feature Extraction

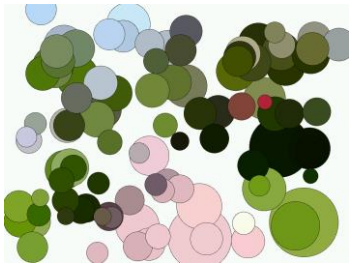
multimedia data object

feature extraction



features $f_1, \ldots, f_n \in \mathbb{F}$

feature aggregation



feature representation

- Feature extraction
  - Multimedia objects are represented by means of features $f_1, \ldots, f_n \in \mathbb{F}$ in a feature space $\mathbb{F}$
  - Example: SIFT features: $\mathbb{F} = \mathbb{R}^{128}$
- Feature aggregation
  - The features $f_1, \ldots, f_n$ of an object are aggregated to a compact feature representation
  - Obtained by clustering algorithms: k-means, expectation maximization, …
- Feature representations objects O may be defined as functions $O: \mathbb{F} \rightarrow \mathbb{R}$

Tutorial Outline

1) **Object Representation**
   - Feature Extraction and Representation
   - Feature Aggregation

2) Fundamental Similarity Models
   - Dissimilarity Measures
   - Distance Functions for Feature Histograms
   - Distance Functions for Feature Signatures

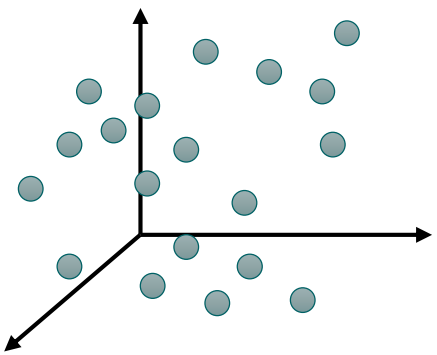
3) Efficient Query Processing
   - Similarity Queries
   - Lower-Bounding: 2 examples

4) Indexing
   - Spatial Indexing
   - Metric and Ptolemaic Indexing

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Feature-based representations as functions [Beecks13]

- Consider multimedia objects O as distributions in the feature space $\mathbb{F}$:



- Aggregation of features leads to weighted distributions, i.e. real-valued, non-binary functions of features:

$$O: \mathbb{F} \rightarrow \mathbb{R}$$

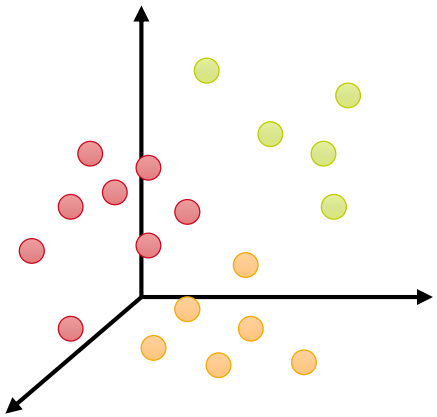
- The non-vanishing features are called the representatives $R_O \subseteq \mathbb{F}$ of a feature representation O:

$$R_O = \{f \in \mathbb{F} \mid O(f) \neq 0\}$$

- The weight of a single feature $f \in \mathbb{F}$ is defined as $O(f) \in \mathbb{R}$

Feature Aggregation by Clustering

- Feature representations may be derived by means of clustering algorithms, e.g.:
  - K-Means algorithm [MacQueen67]
  - Expectation Maximization algorithm [DLR77]



feature space $\mathbb{F}$

clustering algorithm

feature signatures ($\mathbb{S}$)

feature histograms ($\mathbb{H}_{\mathrm{R}}$)

probabilistic feature signatures ($\mathbb{S}^{\mathrm{Pr}}$)

Feature Signatures and Feature Histograms

- **Feature signatures**
  - Multimedia objects O are described by a (finite) number of features:

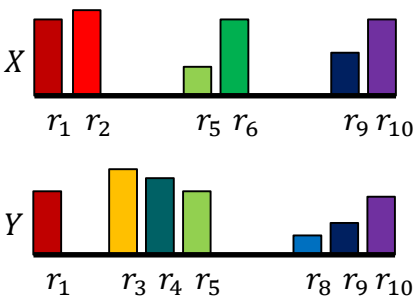
$$S(O): \mathbb{F} \rightarrow \mathbb{R}$$

  - The mappings $\mathbb{F} \rightarrow \mathbb{R}$ form a vector space, $\mathbb{R}^{\mathbb{F}}$, so feature signatures are represented by linear combinations of base signatures b_f

$$S(O) = \sum_{f \in \mathbb{F}} w_{O,f} b_f$$

  - Storage and processing requires finite representations:

$$\left|\{w_{O,f}, f \in \mathbb{F}, O \in DB\}\right| < \infty$$

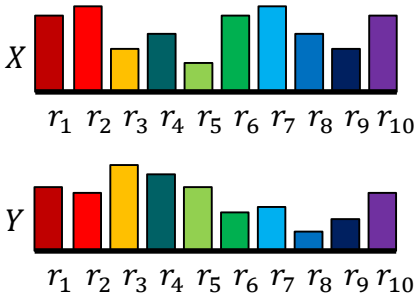


$$X = w_1 r_1 + \cdots + w_{10} r_{10}$$

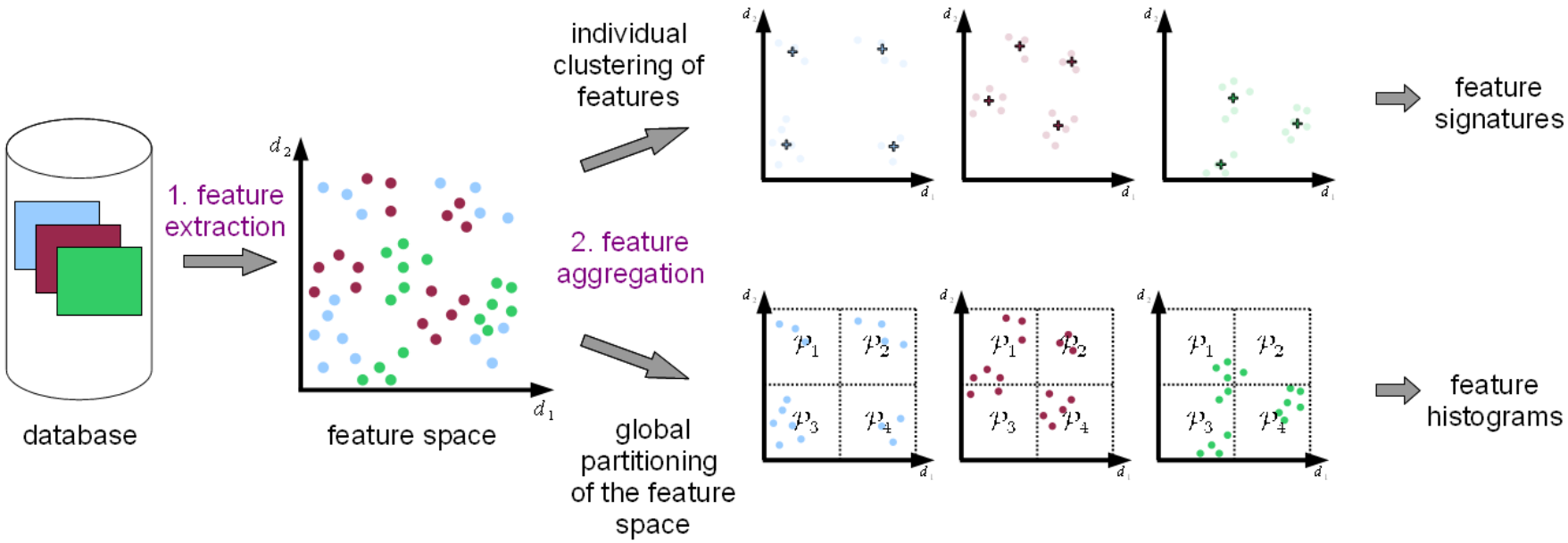
- **Feature histograms**
  - All signatures share the same (restricted) base features:

$$H_{\mathrm{R}}: \mathbb{F} \rightarrow \mathbb{R} \text{ subject to } H_{\mathrm{R}}(\mathbb{F} \backslash R) = \{0\}$$

  - Store feature base once only, not for each object individually



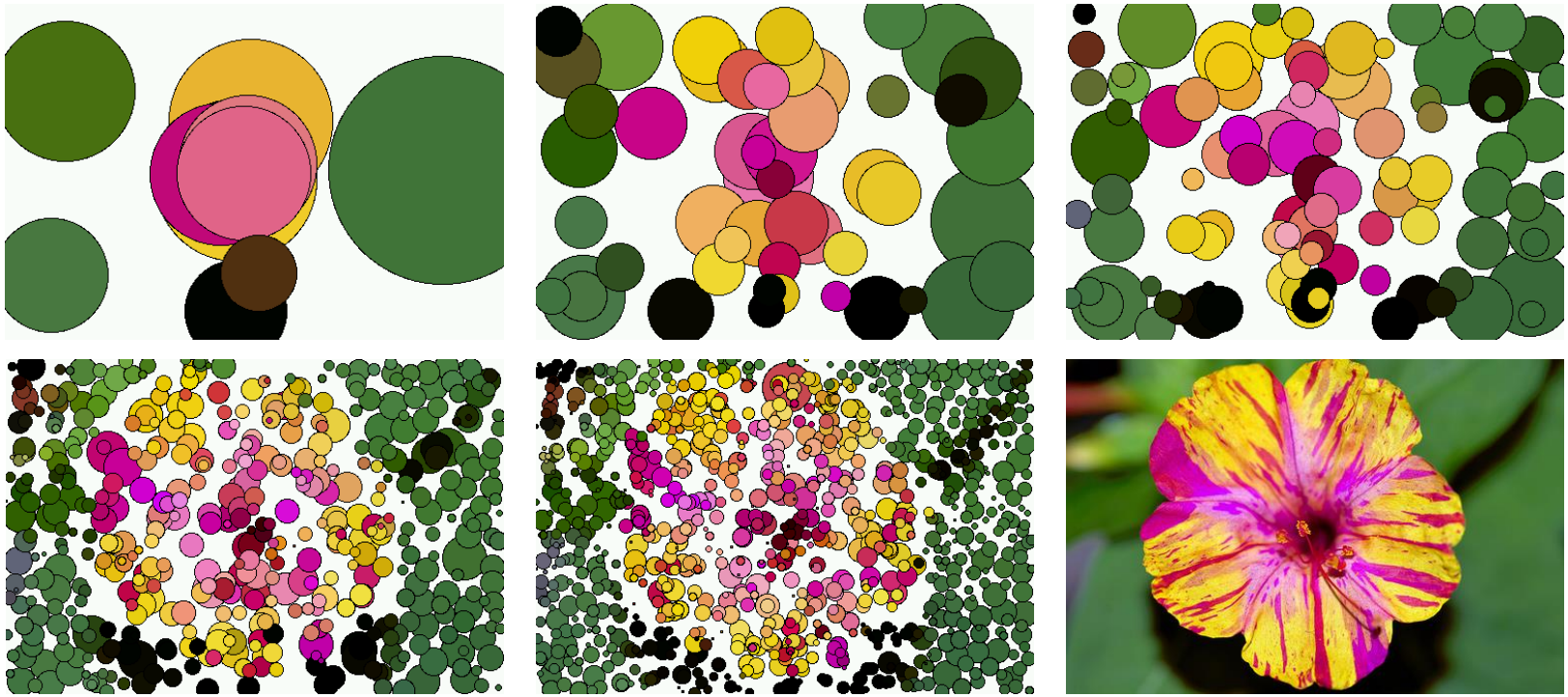
Feature Extraction and Aggregation



- Different means of feature aggregation:
  - **Feature Histogram**: features are summarized according to a global partitioning which is fixed for all multimedia data objects
  - **Feature Signature**: features are summarized individually (per object)

Example: Feature Signatures

- 7-dimensional features: position, color, coarseness, and contrast
- Random sampling of 40.000 image pixels
- Increasing the number of representatives from 10 to 1000:



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Summary

- Features reflect characteristic properties of multimedia objects
- The characteristic properties are summarized by feature representations
- Feature representations assign each feature a weight
- Finite feature representations:
  - Feature histograms
  - Feature signatures
  - (Probabilistic feature signatures [BIK+11a])
- We compute our feature representations by using clustering algorithms

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Tutorial Outline

1) Object Representation
   - Feature Extraction and Representation
   - Clustering-based Computation

2) **Fundamental Similarity Models**
   - Dissimilarity Measures
   - Distance Functions for Feature Histograms
   - Distance Functions for Feature Signatures

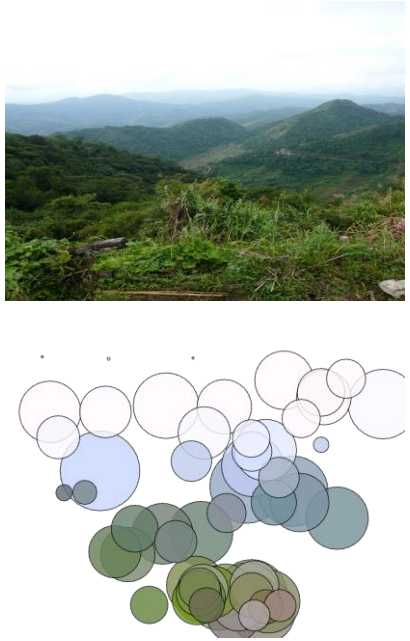
3) Efficient Query Processing
   - Similarity Queries
   - Lower-Bounding: 2 examples

4) Indexing
   - Spatial Indexing
   - Metric and Ptolemaic Indexing

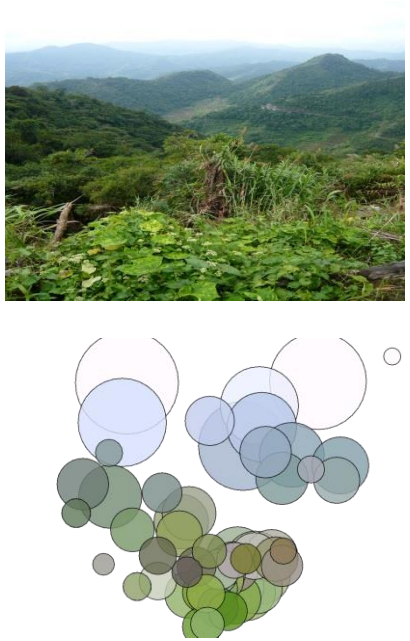
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Similarity vs. Dissimilarity

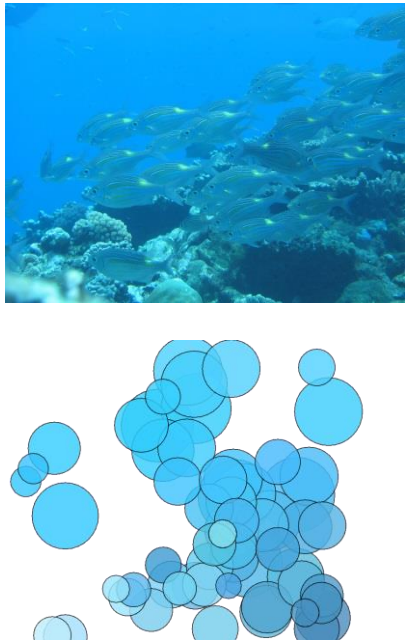
object o_1



object o_2



object o_3



- Similarity measures assign high values to similar objects: $sim(o_1, o_2) \geq sim(o_1, o_3)$
- Dissimilarity measures assign low values to similar objects: $\delta(o_1, o_2) \leq \delta(o_1, o_3)$

Dissimilarity Measures

- Dissimilarity measures may follow the idea of a **geometric approach**, which is common, preferable and influential [AP88, Shepard57, JSW08, SJ99]

- Multimedia objects are defined by their **perceptual representations** in a **perceptual space**
  - Perceptual representations = features or feature representations
  - Perceptual space = feature or feature representation space

- **Geometric distance** between the perceptual representations defines dissimilarity of multimedia objects
  - geometric distance = distance function

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Distance Functions

- Distance functions $\delta: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}^{\geq 0}$ are a specific type of dissimilarity measures

- **Distance functions** satisfy the following properties [DD09]:
  i. Reflexivity: $\forall x \in \mathbb{X}: \ \delta(x,x) = 0$
  ii. Non-negativity: $\forall x, y \in \mathbb{X}: \ \delta(x,y) \geq 0$
  iii. Symmetry: $\forall x, y \in \mathbb{X}: \ \delta(x,y) = \delta(y,x)$

- **Metric distance functions** satisfy the following properties [DD09]:
  i. Identity of indiscernibles: $\forall x, y \in \mathbb{X}: \ \delta(x,y) = 0 \Leftrightarrow x = y$
  ii. Symmetry: $\forall x, y \in \mathbb{X}: \ \delta(x,y) = \delta(y,x)$
  iii. Triangle inequality: $\forall x, y, z \in \mathbb{X}: \ \delta(x,y) \leq \delta(x,z) + \delta(z,y)$

  Non-negativity is derived: $\forall x, y \in \mathbb{X}: \ 0 = \delta(x,x) \leq \delta(x,y) + \delta(y,x) = 2 \cdot \delta(x,y)$

- We call the tuple $(\mathbb{X}, \delta)$ a **distance space** or a **metric space**, respectively.

Multitude of Literature

- Many distance functions for feature histograms have been investigated analytically and empirically …
  - 1979: An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems [McGill79]
  - 1999 & 2001: Empirical Evaluation of Dissimilarity Measures for Color and Texture [PRT+99, RPT+01]
  - 2003: Evaluation of similarity measurement for image retrieval [ZL03]
  - 2008: Dissimilarity measures for content-based image retrieval [HRS+08]
  - 2009: Encyclopedia of distances [DD09]
  - etc.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Encyclopedia of Distances

- Very exhaustive book by M. M. Deza and E. Deza [DD09]:

Contents

Part I Mathematics of Distances

1 General Definitions ... 3
1.1 Basic definitions ... 3
1.2 Main distance-related notions ... 11
1.3 Metric numerical invariants ... 19
1.4 Metric mappings ... 30
1.5 General distances ... 40

2 Topological Spaces ... 59

3 Generalizations of Metric Spaces ... 67
3.1 m-metrics ... 67
3.2 Indefinite metrics ... 68
3.3 Topological generalizations ... 69
3.4 Beyond numbers ... 72

4 Metric Transforms ... 79
4.1 Metrics on the same set ... 79
4.2 Metrics on set extensions ... 82
4.3 Metrics on other sets ... 85

5 Metrics on Normed Structures ... 89

Part II Geometry and Distances

6 Distances in Geometry ... 103
6.1 Geodesic Geometry ... 103
6.2 Projective Geometry ... 110
6.3 Affine Geometry ... 115
6.4 Non-Euclidean Geometry ... 117

xi

xii Contents

7 Riemannian and Hermitian Metrics ... 127
7.1 Riemannian metrics and generalizations ... 128
7.2 Riemannian metrics in Information Theory ... 143
7.3 Hermitian metrics and generalizations ... 148

8 Distances on Surfaces and Knots ... 161
8.1 General surface metrics ... 161
8.2 Intrinsic metrics on surfaces ... 167
8.3 Distances on knots ... 170

9 Distances on Convex Bodies, Cones, and Simplicial Complexes ... 173
9.1 Distances on convex bodies ... 173
9.2 Distances on cones ... 179
9.3 Distances on simplicial complexes ... 181

Part III Distances in Classical Mathematics

10 Distances in Algebra ... 187
10.1 Group metrics ... 187
10.2 Metrics on binary relations ... 196
10.3 Metrics on lattices ... 197

11 Distances on Strings and Permutations ... 201
11.1 Distances on general strings ... 202
11.2 Distances on permutations ... 210

12 Distances on Numbers, Polynomials, and Matrices ... 215
12.1 Metrics on numbers ... 215
12.2 Metrics on polynomials ... 220
12.3 Metrics on matrices ... 221

13 Distances in Functional Analysis ... 229
13.1 Metrics on function spaces ... 229
13.2 Metrics on linear operators ... 236

14 Distances in Probability Theory ... 241
14.1 Distances on random variables ... 242
14.2 Distances on distribution laws ... 243

Part IV Distances in Applied Mathematics

15 Distances in Graph Theory ... 257
15.1 Distances on vertices of a graph ... 258
15.2 Distance-defined graphs ... 264

Contents xiii

15.3 Distances on graphs ... 271
15.4 Distances on trees ... 278

16 Distances in Coding Theory ... 285
16.1 Minimum distance and relatives ... 286
16.2 Main coding distances ... 289

17 Distances and Similarities in Data Analysis ... 297
17.1 Similarities and distances for numerical data ... 298
17.2 Relatives of Euclidean distance ... 301
17.3 Similarities and distances for binary data ... 304
17.4 Correlation similarities and distances ... 307

18 Distances in Mathematical Engineering ... 311
18.1 Motion planning distances ... 311
18.2 Cellular automata distances ... 315
18.3 Distances in Control Theory ... 316
18.4 MOEA distances ... 318

Part V Computer-related Distances

19 Distances on Real and Digital Planes ... 323
19.1 Metrics on real plane ... 323
19.2 Digital metrics ... 332

20 Voronoi Diagram Distances ... 339
20.1 Classical Voronoi generation distances ... 340
20.2 Plane Voronoi generation distances ... 342
20.3 Other Voronoi generation distances ... 346

21 Image and Audio Distances ... 349
21.1 Image distances ... 349
21.2 Audio distances ... 362

22 Distances in Internet and Similar Networks ... 371
22.1 Scale-free networks ... 371
22.2 Network-based semantic distances ... 374
22.3 Distances in Internet and Web ... 376

Part VI Distances in Natural Sciences

23 Distances in Biology ... 385
23.1 Genetic distances for gene-frequency data ... 387
23.2 Distances for DNA/RNA data ... 391

xiv Contents

23.3 Distances for protein data ... 400
23.4 Other biological distances ... 404

24 Distances in Physics and Chemistry ... 419
24.1 Distances in Physics ... 419
24.2 Distances in Chemistry and Crystallography ... 436

25 Distances in Geography, Geophysics, and Astronomy ... 445
25.1 Distances in Geography and Geophysics ... 445
25.2 Distances in Astronomy ... 456

26 Distances in Cosmology and Theory of Relativity ... 465
26.1 Distances in Cosmology ... 465
26.2 Distances in Theory of Relativity ... 473

Part VII Real-world Distances

27 Length Measures and Scales ... 495
27.1 Length scales ... 495
27.2 Orders of magnitude for length ... 503

28 Distances in Applied Social Sciences ... 509
28.1 Distances in Psychology and Sociology ... 509
28.2 Distances in Economics and Human Geography ... 514
28.3 Distances in Perception, Cognition and Language ... 522
28.4 Distances in Philosophy, Religion and Art ... 530

29 Other distances ... 537
29.1 Distances in Medicine, Antropometry and Sport ... 537
29.2 Equipment distances ... 546
29.3 Miscellany ... 554

References ... 563

Index ... 571

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Tutorial Outline

1) Object Representation
   - Feature Extraction and Representation
   - Clustering-based Computation

2) **Fundamental Similarity Models**
   - Dissimilarity Measures
   - Distance Functions for Feature Histograms
   - Distance Functions for Feature Signatures

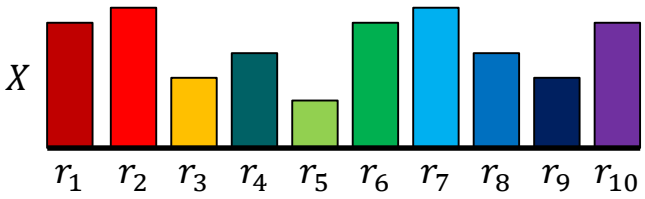
3) Efficient Query Processing
   - Similarity Queries
   - Lower-Bounding: 2 examples

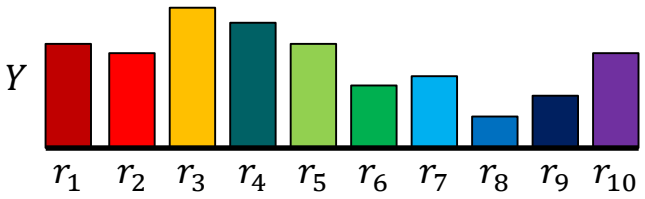
4) Indexing
   - Spatial Indexing
   - Metric and Ptolemaic Indexing

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Distance Functions for Feature Histograms

- Given two feature histograms $X, Y \in \mathbb{H}_{\mathrm{R}}$, how can we define a distance between them?
- Consider the following color histograms for $\mathrm{R} = \{r_1, \dots, r_{10}\}$:



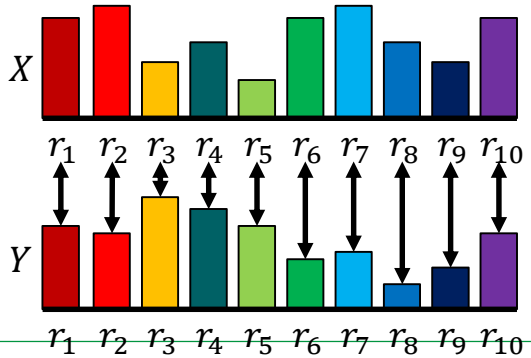


Minkowski Distances

- Idea: Measure the dissimilarity by summing up the differences in all dimensions, i.e. for all representatives $f \in \mathrm{R} \subseteq \mathbb{F}$
- Given two feature histograms $X, Y \in \mathbb{H}_{\mathrm{R}}$, the **Minkowski Distance** is defined for $p \in \mathbb{R}^{\geq 0} \cup \{\infty\}$ as:

$$\mathrm{L}_p(X,Y) = \left(\sum_{f \in \mathrm{R}} |X(f) - Y(f)|^p \right)^{\frac{1}{p}}$$

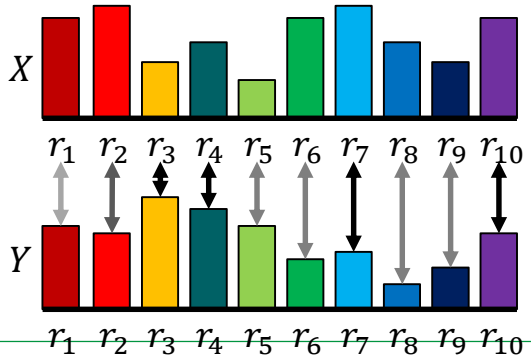
- This corresponds to taking into account all pairwise differences:



Weighted Minkowski Distances

- Idea: Model the influence of the shared representatives $\mathrm{R} \subseteq \mathbb{F}$ by a **weighting function** $w: \mathbb{F} \rightarrow \mathbb{R}^{\geq 0}$
- Given two feature histograms $X, Y \in \mathbb{H}_{\mathrm{R}}$, the **Weighted Minkowski Distance** is defined for $p \in \mathbb{R}^{\geq 0} \cup \{\infty\}$ and a weighting function w as:

$$\mathrm{L}_{p,w}(X,Y) = \left(\sum_{f \in \mathrm{R}} w(f) \cdot |X(f) - Y(f)|^p \right)^{\frac{1}{p}}$$



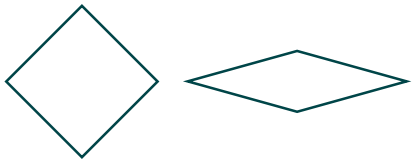
Examples of Minkowski distances

- General (weighted) formula

$$\mathrm{L}_{p,w}(X,Y) = \left(\sum_{f \in \mathrm{R}} w(f) \cdot |X(f) - Y(f)|^p \right)^{\frac{1}{p}}$$

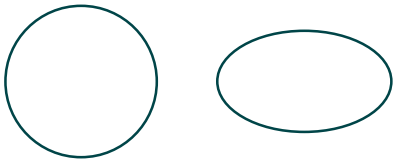
- Prevalent instances (illustrated by iso-surfaces, unweighted and weighted case)

$p = 1$



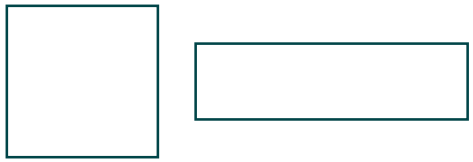
Manhattan distance,
city-block distance,
based on sum norm

$p = 2$



Euclidean distance,
aerial distance

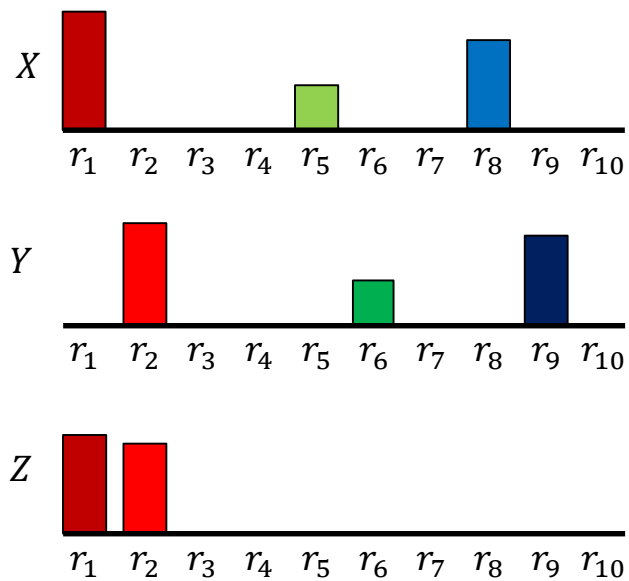
$p \to \infty$



maximum distance
based on max norm

Issues of bin-by-bin distances

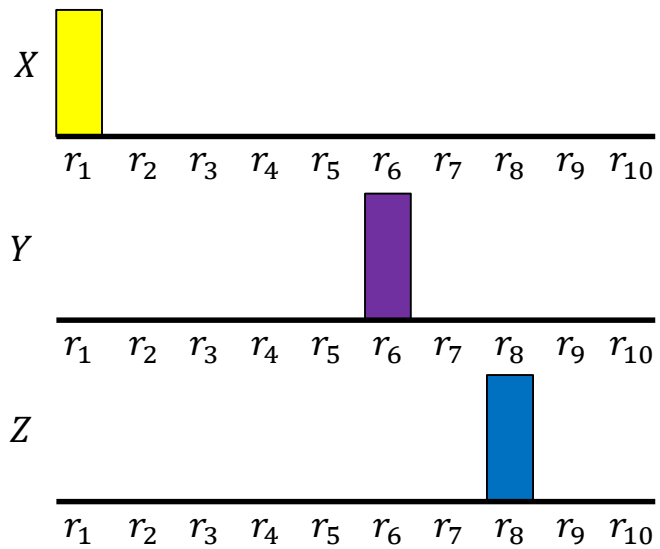
- Bin-by-bin distance functions neglect cross-similarities of representatives
- Consider the following color histograms $X, Y, Z \in \mathbb{H}_{\mathrm{R}}$ with $R = \{r_i\}_{i=1}^{10}$:



- Despite X and Y seem to be more similar, we find $\mathrm{L}_p(X, Y) \geq \mathrm{L}_p(X, Z)$

Issues of bin-by-bin distance functions cont'd

- Consider the following color histograms $X, Y, Z \in \mathbb{H}_{\mathrm{R}}$ with $R = \{r_i\}_{i=1}^{10}$:



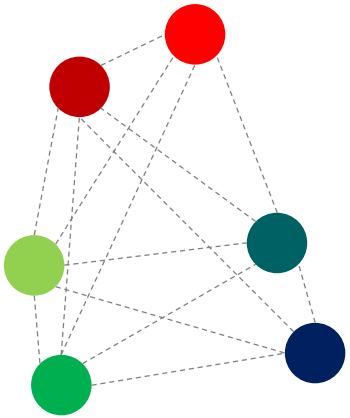
- These color histograms $X, Y, Z \in \mathbb{H}_{\mathrm{R}}$ share the same pairwise Minkowsi Distance:

$$\mathrm{L}_p(X, Y) = \mathrm{L}_p(X, Z) = \mathrm{L}_p(Y, Z)$$

- As an inherent shortcoming, bin-by-bin distances neglect the higher similarity of purple to blue than to yellow

Cross-bin Distance Functions

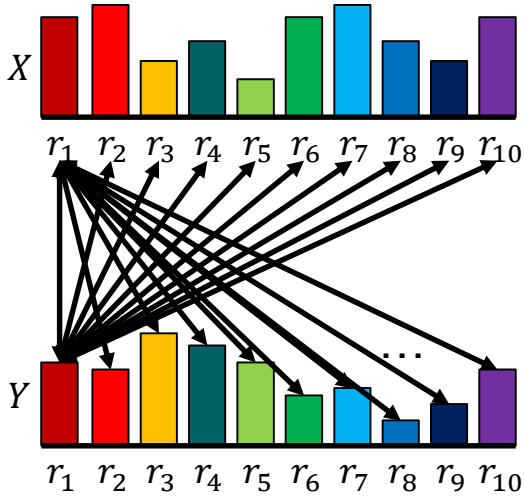
- More flexible than bin-by-bin distance functions
- **Basic ideas**:
  - Replace the weighting of single representatives by a weighting of pairs of representatives
  - Model the influence not only for each single representative, but also among different representatives
  - This influence is often defined in terms of a similarity relation
  - Thus, we utilize **similarity functions** $s: \mathbb{F} \times \mathbb{F} \rightarrow \mathbb{R}$ on the feature level which provide similarity values for all pairs of features



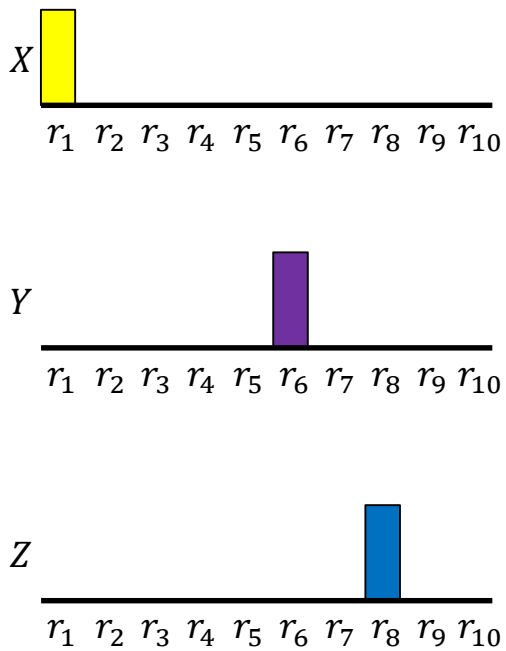
Quadratic Form Distance

- The **Quadratic Form Distance** [Ioka89, NBE+93, FBF+94, HSE+95] is a cross-bin distance function that takes into account all pair-wise similarities
- Given two feature histograms $X, Y \in \mathbb{H}_{\mathrm{R}}$, the Quadratic Form Distance w.r.t. to a similarity function $s: \mathbb{F} \times \mathbb{F} \rightarrow \mathbb{R}$ is defined as:

$$\mathrm{QFD}_s(X,Y) = \sqrt{\sum_{f \in \mathbb{F}} \sum_{g \in \mathbb{F}} \left(X_f - Y_f\right) \cdot s(f,g) \cdot \left(X_g - Y_g\right)}$$



Quadratic Form Distance: Example



- The Quadratic form distances are as follows:
  - $\mathrm{QFD}_s(X,Y) = \sqrt{1.6} \approx 1.265$
  - $\mathrm{QFD}_s(X,Z) = \sqrt{1.6} \approx 1.265$
  - $\mathrm{QFD}_s(Y,Z) = \sqrt{0.8} \approx 0.894$
- Better fits our intuition of dissimilarity
- Adaptation to different users' needs by modifying or changing the underlying feature similarity function

Distance Functions for Feature Histograms

- Distance functions for feature histograms (= shared representatives)
- Weighted Minkowski Distances have limited adaptability (weights only) but show *linear* computation time complexity
- Quadratic Form Distances are very adaptable (ground feature similarity) but show *quadratic* computation time complexity
- Other distance functions (e.g., [RPT+01, ZL03, HRS+08]):
  - geometric measures such as cosine distance
  - information theoretic measures such as Kullback-Leibler [KL51]
  - statistic measures such as χ^2-statistics [PHB97]

Tutorial Outline

1) Object Representation
   - Feature Extraction and Representation
   - Clustering-based Computation

2) **Fundamental Similarity Models**
   - Dissimilarity Measures
   - Distance Functions for Feature Histograms
   - Distance Functions for Feature Signatures

3) Efficient Query Processing
   - Similarity Queries
   - Lower-Bounding: 2 examples

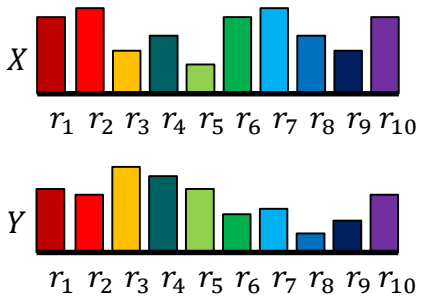
4) Indexing
   - Spatial Indexing
   - Metric and Ptolemaic Indexing

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Conceptual Differences of Feature Representations

Feature histograms $\mathbb{H}_R$

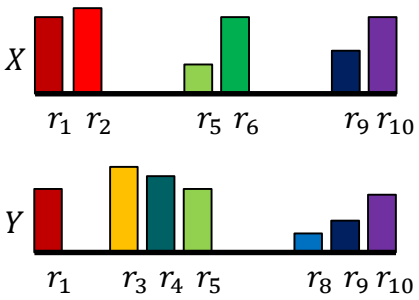
- Multimedia objects share representatives:
  - Sufficient to store weights only
  - Feature histograms have the same cardinality
  - Equivalent to Euclidean vectors (representatives = dimensions)



  - Distance computation by means of differences in each dimension

Feature signatures $\mathbb{S}$

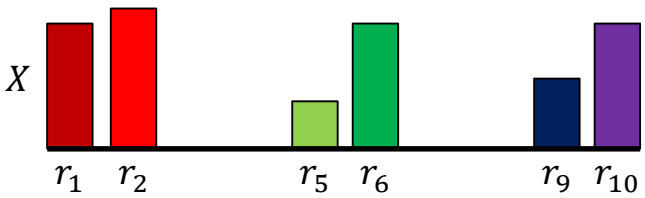
- Multimedia objects use individual representatives:
  - Store weights and representatives
  - Feature signatures have different cardinalities



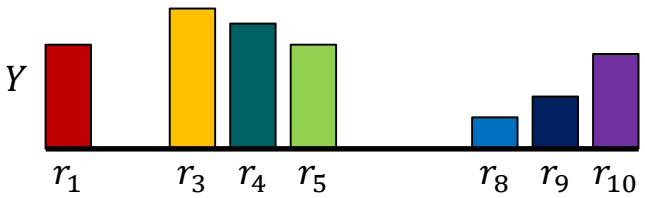
  - Distance computation along single dimensions not meaningful

Distance Functions for Feature Signatures

- How to define a distance values for feature signatures?
- Consider the following color signatures $X, Y \in \mathbb{S}$:



$$\mathrm{R}_X = \{r_1, r_2, r_5, r_6, r_9, r_{10}\}$$

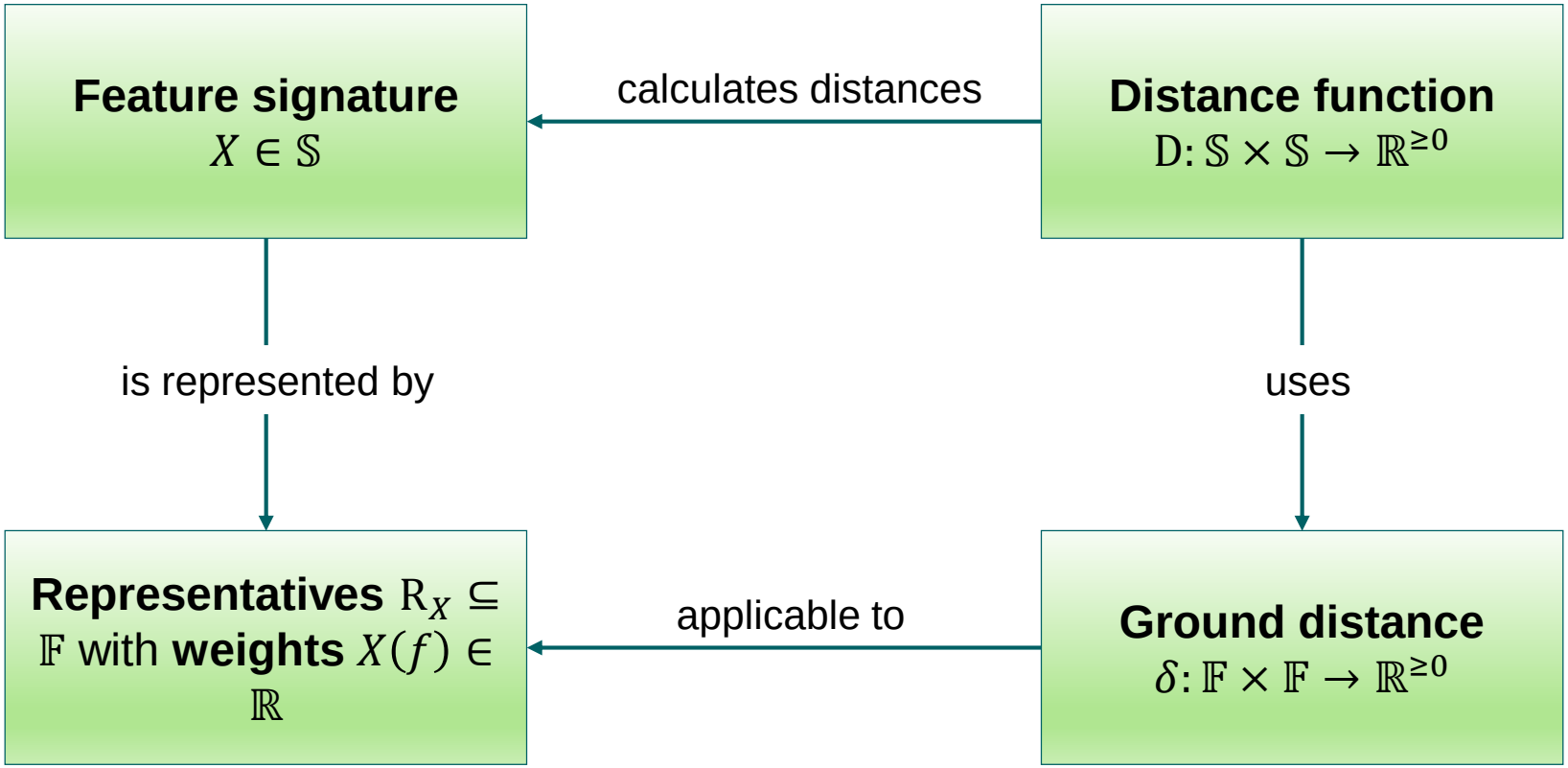


$$\mathrm{R}_Y = \{r_1, r_3, r_4, r_5, r_8, r_9, r_{10}\}$$

- Question: How to relate different representatives arising from different feature signatures to each other?
  - Utilization of a **ground distance** $\delta: \mathbb{F} \times \mathbb{F} \rightarrow \mathbb{R}^{\geq 0}$

Concept Overview

- Idea: Utilization of a ground distance $\delta: \mathbb{F} \times \mathbb{F} \rightarrow \mathbb{R}^{\geq 0}$ on the representatives $\mathrm{R}_X, \mathrm{R}_Y \subseteq \mathbb{F}$ of two feature signatures $X, Y \in \mathbb{S}$



Different Approaches for Feature Signatures

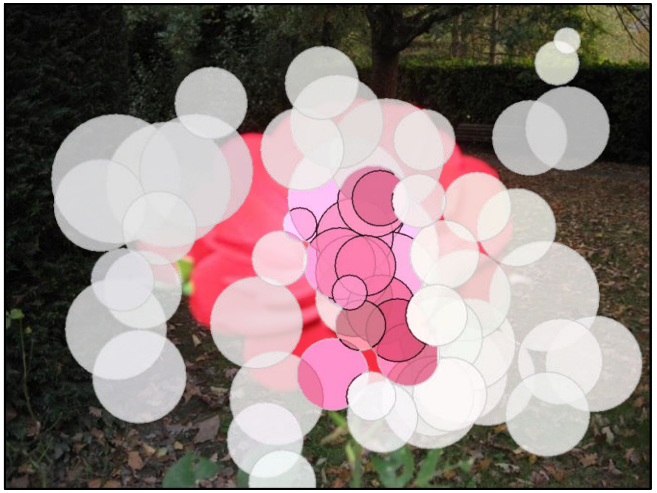
- **Matching-based measures:**
  - Distance functions are defined by matching similar representatives and by determining the cost of a matching
  - Examples: Hausdorff Distance [H14] and variants [HKR93, PLL08], Signature Matching Distance [BKS13]

- **Transformation-based measures**:
  - Distance functions are defined by measuring the costs of transforming one feature signature into another one
  - Example: Earth Mover's Distance [RTG00]

- **Correlation-based measures**:
  - Distance functions are defined by means of the correlation of the representatives of the feature signatures
  - Examples: Signature Quadratic Form Distance [BUS10], Weighted Correlation Distance [LL04]

Matching-based Measures

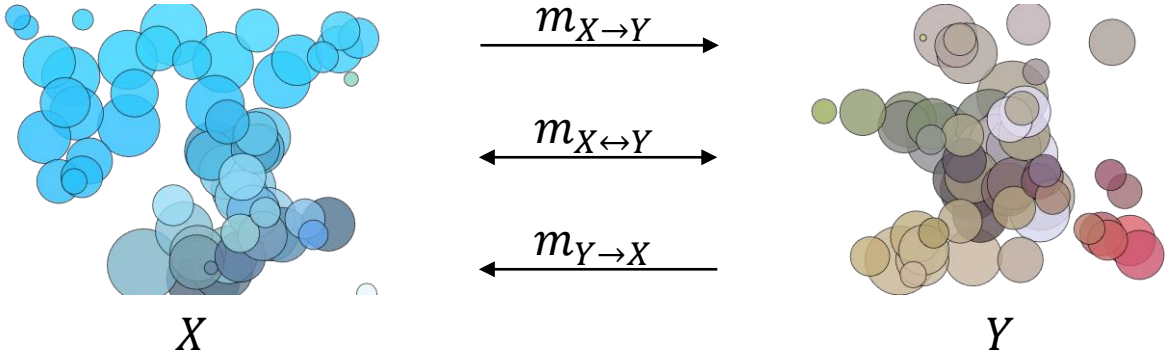
- The conceptual idea consists in defining a distance value between two feature signatures based on coincident similar parts of their representatives
- Approach:
  1. Computation of a **matching**
  2. Computation of a **cost function** that evaluates the matching quality
- Example: Distance is attributed to the most visually similar parts



Signature Matching Distance

- **Idea**: Define distance by the cost of the symmetric difference of matching representatives of the feature signatures
- Given two feature signatures $X, Y \in \mathbb{S}$ over a feature space $\mathbb{F}$, the **Signature Matching Distance** $\mathrm{SMD}_\delta: \mathbb{S} \times \mathbb{S} \rightarrow \mathbb{R}$ of X and Y for some ground distance δ and parameter λ is defined as:

$$\mathrm{SMD}_\delta(X, Y) = c(m_{X \rightarrow Y}) + c(m_{Y \rightarrow X}) - 2 \cdot \lambda \cdot c(m_{X \leftrightarrow Y})$$



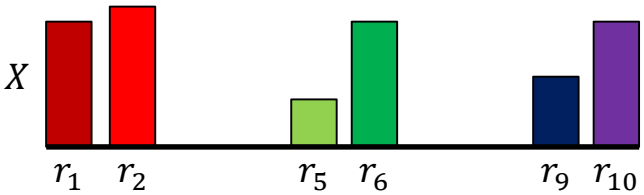
Transformation-based Measures

- The conceptual idea consists in transforming one feature signature into another one and treating the transformation costs as distance
- Examples:
  - **Levenshtein/Edit Distance** on discrete structures [L66]
    - Edit operations: insertion, deletion, substitution
    - Distance is defined as the minimum number of edit operations
  - **Dynamic Time Warping Distance** on times series [I75,SC78]
    - Warping operation: replication
    - Distance is defined based on the minimum number of replications

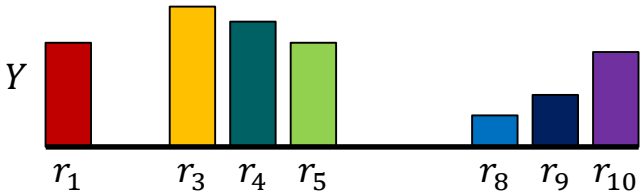
- **Earth Mover's Distance** on feature signatures [RTG00]

Earth Mover's Distance: Principle

- Given two color signatures $X, Y \in \mathbb{S}$:



$$\mathrm{R}_X = \{r_1, r_2, r_5, r_6, r_9, r_{10}\}$$



$$\mathrm{R}_Y = \{r_1, r_3, r_4, r_5, r_8, r_9, r_{10}\}$$

- The transportation (earth moving) problem is formalized by:
  - **Earth hills** R_X with capacities $X(r_i)$ for $r_i \in \mathrm{R}_X$
  - **Earth holes** R_Y with capacities $Y(r_i)$ for $r_i \in \mathrm{R}_Y$
  - **Cost** (ground distance) $\delta: \mathbb{F} \times \mathbb{F} \to \mathbb{R}$ for moving earth
  - **All possible flows** $F = \{f \mid f: \mathbb{F} \times \mathbb{F} \to \mathbb{R}\}$
- Solution: flow $f_{min} \in F$ that minimizes cost flow $\sum_{\substack{g \in \mathrm{R}_X \\ h \in \mathrm{R}_Y}} f_{min}(g,h) \cdot \delta(g,h)$

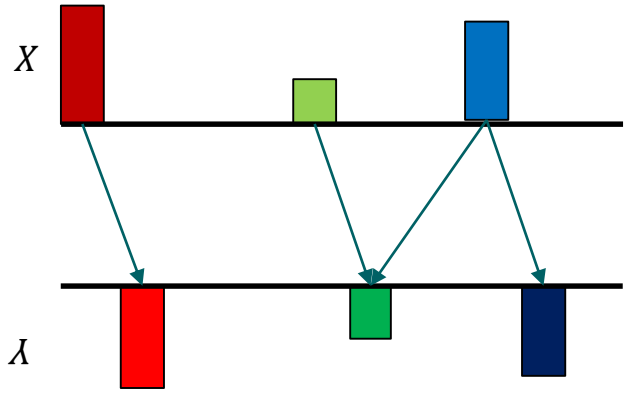
Earth Mover's Distance: Definition

- Given two feature signatures $X, Y \in \mathbb{S}$ over a feature space $\mathbb{F}$, the **Earth Mover's Distance** $\mathrm{EMD}_\delta: \mathbb{S} \times \mathbb{S} \to \mathbb{R}$ between X and Y is defined as:

$$\mathrm{EMD}_\delta(X,Y) = \min_{\{f \mid f:\mathbb{F}\times\mathbb{F}\to\mathbb{R}\}} \left\{ \frac{\sum_{g\in \mathrm{R}_X} \sum_{h\in \mathrm{R}_Y} f(g,h)\cdot\delta(g,h)}{\min\left\{\sum_{g\in \mathrm{R}_X} X(g), \sum_{h\in \mathrm{R}_Y} Y(h)\right\}} \right\}$$

subject to the following sets of constraints:

- CNNeg: $\forall g \in \mathrm{R}_X, \forall h \in \mathrm{R}_Y: f(g,h) \geq 0$
- CSource: $\forall g \in \mathrm{R}_X: \sum_{h\in \mathrm{R}_Y} f(g,h) \leq X(g)$
- CTarget: $\forall h \in \mathrm{R}_Y: \sum_{g\in \mathrm{R}_X} f(g,h) \leq Y(h)$
- CMaxFlow: $\sum_{\substack{g\in \mathrm{R}_X \\ h\in \mathrm{R}_Y}} f(g,h) = \min\left\{\sum_{g\in \mathrm{R}_X} X(g), \sum_{h\in \mathrm{R}_Y} Y(h)\right\}$



Earth Mover's Distance: Properties

- The Earth Mover's Distance is defined as a linear optimization problem
- Finding an optimal solution can be computed based on a specific variant of the simplex algorithm [HL90]
- Exponential computation time complexity in the worst case
- Average empirical computation time complexity between $\mathcal{O}(|\mathrm{R}_X|^3)$ and $\mathcal{O}(|\mathrm{R}_X|^4)$ for $|\mathrm{R}_X| \geq |\mathrm{R}_Y|$ [SJ08]
- Earth Mover's Distance is metric if and only if
  - feature signatures are normalized, i.e. $\sum_{f \in \mathrm{R}_X} X(f) = \sum_{f \in \mathrm{R}_Y} Y(f)$
  - ground distance δ is metric

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

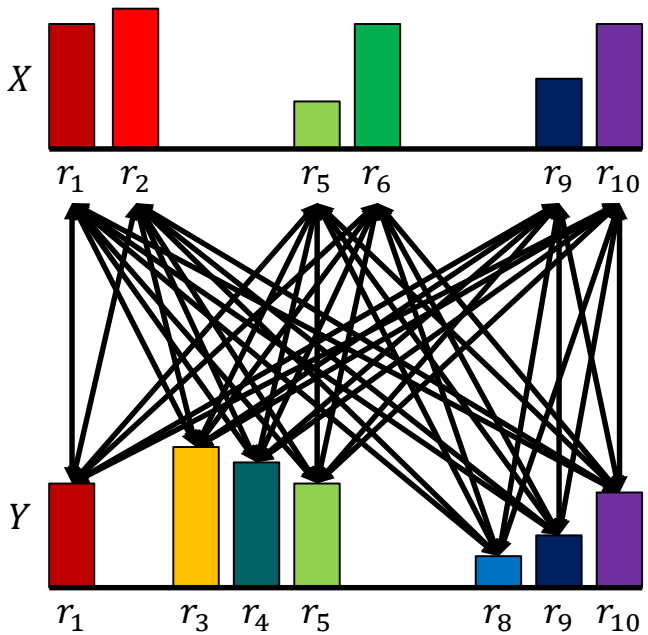
Correlation-based Measures

- Basic idea: define a distance value by means of the correlation of the representatives of the feature signatures
- All representatives and weights are compared with each other
- Given two feature signatures $X, Y \in \mathbb{S}$ over a feature space $\mathbb{F}$ and a similarity function $s: \mathbb{F} \times \mathbb{F} \rightarrow \mathbb{R}$, the **similarity correlation** $\langle \cdot, \cdot \rangle_s: \mathbb{S} \times \mathbb{S} \rightarrow \mathbb{R}$ between X and Y is defined as:

$$\langle X, Y \rangle_s = \sum_{f \in \mathbb{F}} \sum_{g \in \mathbb{F}} X(f) \cdot Y(g) \cdot s(f, g)$$

Signature Quadratic Form Distance

- **Idea**: Utilization of the similarity correlation on the difference signature in order to correlate all representatives of two feature signatures $X, Y \in \mathbb{S}$ with each other



- The resulting Signature Quadratic Form Distance [BUS10] can be thought of as a generalization of the Quadratic Form Distance

Signature Quadratic Form Distance: Definition

- Given two feature signatures $X, Y \in \mathbb{S}$ over a feature space $\mathbb{F}$, the **Signature Quadratic Form Distance** $\mathrm{SQFD}_s: \mathbb{S} \times \mathbb{S} \to \mathbb{R}$ of X and Y for a similarity function $s: \mathbb{F} \times \mathbb{F} \to \mathbb{R}$ is defined as:

$$\mathrm{SQFD}_s(X, Y) = \sqrt{\langle X - Y, X - Y \rangle_s}$$

- Simple decomposition yields:

$$\langle X - Y, X - Y \rangle_s = \underbrace{\langle X, X \rangle_s}_{\substack{\text{Similarity} \\ \text{correlation of } X}} - \underbrace{\langle X, Y \rangle_s - \langle Y, X \rangle_s}_{\substack{\text{Similarity correlation} \\ \text{between } X \text{ and } Y}} + \underbrace{\langle Y, Y \rangle_s}_{\substack{\text{Similarity} \\ \text{correlation of } Y}}$$

Signature Quadratic Form Distance: Properties

- Signature Quadratic Form Distance is implied by an inner product norm:

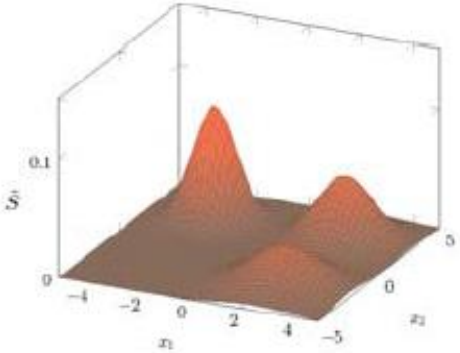
$$\mathrm{SQFD}_s(X,Y) = \sqrt{\langle X-Y, X-Y\rangle_s} = \|X-Y\|_s$$

- Signature Quadratic Form Distance can be thought of as the **length of the difference of two feature signatures**
- Computation time complexity in $\mathcal{O}\big((|\mathrm{R}_X| + |\mathrm{R}_Y|)^2 \cdot \gamma\big)$, where γ denotes the computation time complexity of similarity function s
- Different mathematical models available
- Metric dependent on the similarity function

Further Aspects of SQFD

- Widespread applicability
  - For instance to Gaussian mixture models [BIK+11, BIK+11a]

(a) Picturesque sunset



(b) Continuous distribution

- Promising indexability
  - Steerable by adapting the similarity function [BLS+11]
  - Beyond metric indexing: Ptolemaic indexing [LHS+11, HSL+12]
- Impressive parallelization
  - For instance on multi-core CPUs and many-core GPUs [KLB+11, KSL+12]

Performance Evaluation (Holidays & UKBench databases)

- In [BKS13] feature signatures were evaluated based on complex descriptors for the Holidays database [JDS08] and UKBench database [NS06]:

		Holidays		UKBench		
		MAP	size	MAP	score	size
EMD	PCT	0.720	90	0.741	2.78	50
	SIFT	0.678	70	0.536	2.05	90
	CSIFT	0.749	40	0.605	2.31	30
PMHD	PCT	0.804	80	**0.866**	3.30	90
	SIFT	0.673	70	0.531	2.03	90
	CSIFT	0.755	40	0.594	2.27	30
SQFD	PCT	0.761	40	0.766	2.86	60
	SIFT	0.690	80	0.585	2.23	100
	CSIFT	0.756	20	0.494	2.25	20
SMD	PCT	**0.810**	100	0.845	3.20	60
	SIFT	0.653	40	0.463	1.75	20
	CSIFT	0.735	20	0.531	2.00	20

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Summary

- A similarity model comprises two components:
  - Feature representations
  - (Dis)similarity measure
- Multitude of distance functions available
- Minkowski Distance and Quadratic Form Distance for feature histograms
- Various distance-based similarity measures for feature signatures:
  - Signature Matching Distance
  - Earth Mover's Distance
  - Signature Quadratic Form Distance

Tutorial Outline

1) Object Representation
   - Feature Extraction and Representation
   - Clustering-based Computation

2) Fundamental Similarity Models
   - Dissimilarity Measures
   - Distance Functions for Feature Histograms
   - Distance Functions for Feature Signatures

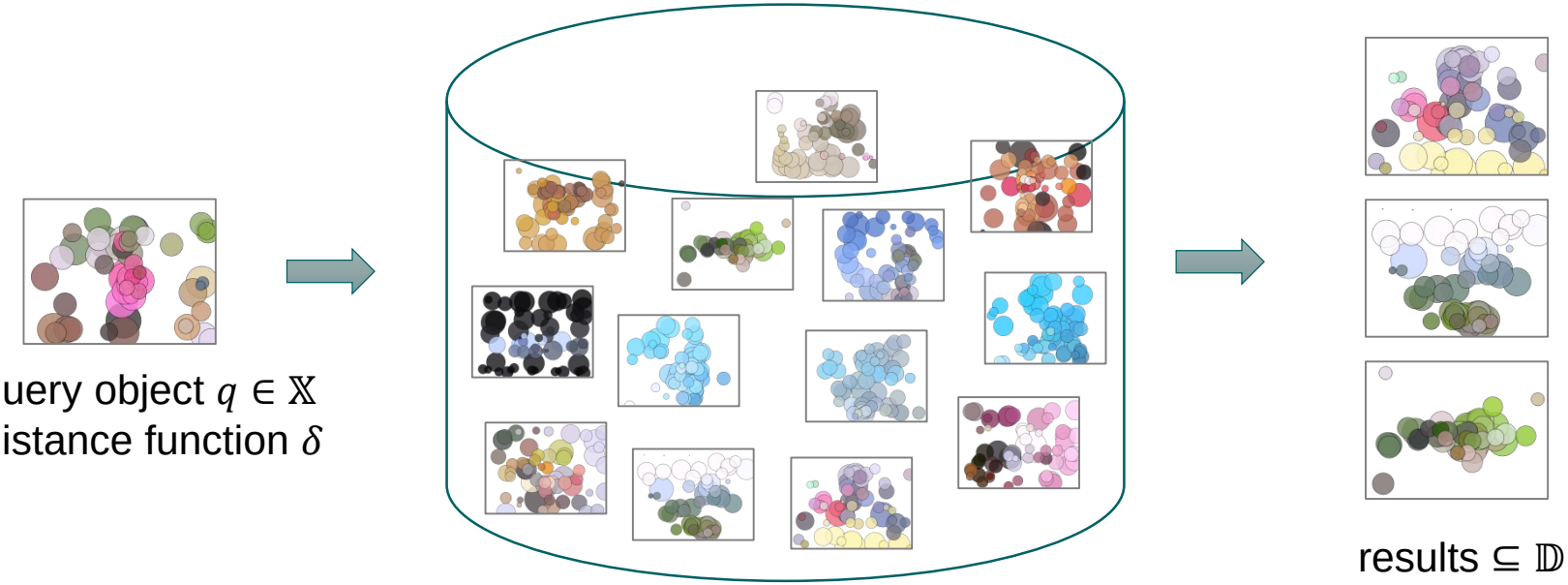
3) **Efficient Query Processing**
   - Similarity Queries
   - Lower-Bounding: 2 examples

4) Indexing
   - Spatial Indexing
   - Metric and Ptolemaic Indexing

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Similarity Queries

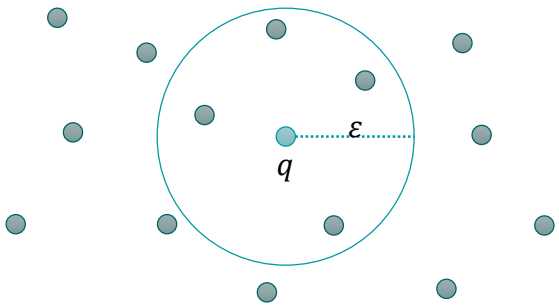
- A **query** $q \in \mathbb{X}$ formalizes an information need
  - **Similarity query**: Given a finite database $\mathbb{D} \subseteq \mathbb{X}$, retrieve multimedia data objects that are similar to a query q
  - **Distance-based similarity query**: use dissimilarity function $\delta: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$



Range Query

- **Range queries** return database objects whose distances to the query object do not exceed the query threshold ε
- Formally: Given
  - a database $\mathbb{D} \subseteq \mathbb{X}$ from a multimedia universe $\mathbb{X}$,
  - a dissimilarity function $\delta: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$,
  - a query object $q \in \mathbb{X}$, and
  - a query range $\varepsilon \in \mathbb{R}^{\geq 0}$,

  the according range query returns the following result set:

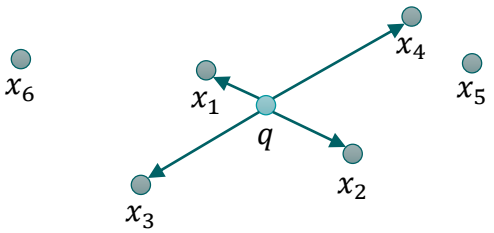


$$\mathrm{range}_{\varepsilon}(q, \delta, \mathbb{D}) = \{x \in \mathbb{D} \mid \delta(q, x) \leq \varepsilon\}$$

K-Nearest-Neighbor Query

- **K-nearest-neighbor queries** return database objects up to the kth-smallest distance to the query object
- Formally: Given
  - a database $\mathbb{D} \subseteq \mathbb{X}$ from a multimedia universe $\mathbb{X}$,
  - a dissimilarity function $\delta: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$,
  - a query object $q \in \mathbb{X}$, and
  - a number $k \in \mathbb{N}$ of desired answers,

  the according k-NN query returns the smallest set $\mathrm{NN}_k(q, \delta, \mathbb{D}) \subseteq \mathbb{D}$ with $|\mathrm{NN}_k(q, \delta, \mathbb{D})| \geq k$ such that:

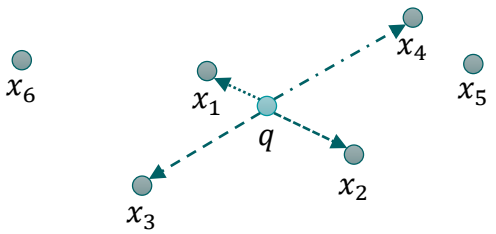


$$\forall x \in \mathrm{NN}_k(q, \delta, \mathbb{D}), \forall x' \in \left(\mathbb{D} - \mathrm{NN}_k(q, \delta, \mathbb{D})\right): \delta(q, x) < \delta(q, x')$$

Ranking Query

- **Ranking queries** return database objects on (individual) request in ascending order w.r.t. the distance to a query object
- Formally: Given
  - a database $\mathbb{D} \subseteq \mathbb{X}$ from a multimedia universe $\mathbb{X}$,
  - a dissimilarity function $\delta: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$,
  - a query object $q \in \mathbb{X}$, and
  - a stream of object requests ("give-me-more"),

  the ranking query returns the sequence



$$\text{ranking}(q, \delta, \mathbb{D}) = \langle x_1, \dots, x_{|\mathbb{D}|} \rangle$$

where $\delta(q, x_i) \leq \delta(q, x_j)$ for all $x_i, x_j \in \mathbb{D}$ and $1 \leq i \leq j \leq |\mathbb{D}|$

Efficient Query Processing

- Given a distance-based similarity model
  - Feature representation, e.g. feature signature, feature histogram
  - Distance function, e.g. SQFD, EMD

- **How to process distance-based similarity queries efficiently?**
  - **How to avoid time-intensive sequential scan?**

- Filter-refine architectures (Orenstein et al., Faloutsos et al.; etc.)
  - Multi-step range query architecture
  - Optimal Multi-step kNN query architecture
  - Additionally supported by index structures

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Complexity of Distance-based Similarity Queries

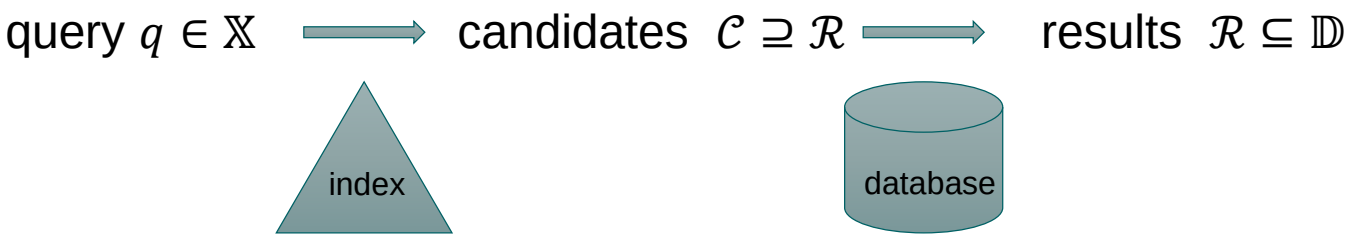
- **Problem**: quality determines complexity
  - High dimensionality (high-resolution partitioning of color space) → better quality
  - Complex distance measure (e.g. Earth Mover's Distance) → better quality
  - But: both require much computing time

- **Solution**: Filter step for reduction of expensive computations
  - Consider a range query $\mathrm{range}_{\varepsilon}(q, \delta, \mathbb{D})$
  - Choose a filter distance δ_{filter} with small computational effort
  - Discard all objects with $\delta_{filter} > \varepsilon$

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Multi-Step Query Architecture

- Processing of distance-based similarity queries in multiple steps:
  - **Filter step** is applied to the *entire database*
    - Use of approximations
    - Efficient generation of candidates
  - **Refinement step** is applied to *candidates only*
    - Use of exact distances
    - **Correctness**: do not return wrong objects
    - **Completeness**: do not discard correct objects
    - **Efficiency**: short response times



Quality of filters in filter-refine processing: ICES Criteria

I

Indexable

- Filter is well supported by indexing techniques

C

Complete

- No correct answers are dismissed in the filter step
- Limited completeness e.g. in PAC-NN [CP00]

E

Efficient

- Filter distances are calculated fast, e.g. linear wrt. Dimensionality

S

Selective

- Filter generates small candidate set only

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Lower-bounding property

- Let $\mathbb{X}$ be a set and $\delta: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ be a distance function. A function $\delta_{LB}: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ is a **lower bound** of δ if it holds that:

$$\forall x, y \in \mathbb{X}: \delta_{LB}(x, y) \leq \delta(x, y)$$

- **Theorem**: The range query and k-nn query algorithms below are **complete** if the filter distance δ_{LB} is a lower bound of the exact object distance δ
  - Proof: Assume a desired object o, $\delta(q, o) \leq \varepsilon$, is missing from the result as it failed to pass the filter due to $\delta_{LB}(q, o) > \varepsilon$. This immediately contradicts the lower-bounding property $\delta_{LB}(q, o) \leq \delta(q, o)$, q.e.d.

- Two approaches of deriving a lower bound:
  - **Model-specific approaches**
    - exploit the inner workings of a distance function
  - **Generic approaches**
    - exploit the properties of the corresponding (metric/ptolemaic) distance space $(\mathbb{X}, \delta)$

Multi-Step Range Query [FRM94]

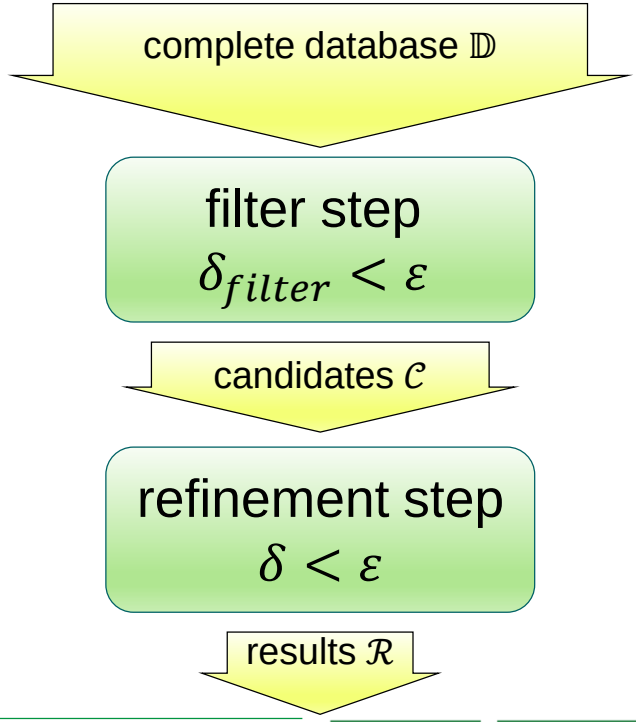
- Given a set $\mathbb{X}$, a database $\mathbb{D} \subseteq \mathbb{X}$, and a distance function $\delta: \mathbb{X} \times \mathbb{X} \to \mathbb{R}$
- Given a lower bound $\delta_{LB}: \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ of δ, how to process a query $\mathrm{range}_\varepsilon(q, \delta, \mathbb{D}) = \{x \in \mathbb{D} \mid \delta(q, x) \leq \varepsilon\}$ efficiently?
- Process:
  - Filter step: evaluate range query with the same $\varepsilon \in \mathbb{R}$ but cheaper filter distance δ_{LB} to generate the candidates

$$\mathcal{C} = \{x \in \mathbb{D} \mid \delta_{LB}(q, x) \leq \varepsilon\}$$

  - Refinement step: refine candidates with the exact distance δ to obtain the results

$$\mathcal{R} = \{x \in \mathcal{C} \mid \delta(q, x) \leq \varepsilon\}$$

- It holds that $\mathcal{R} = \mathrm{range}_\varepsilon(q, \delta, \mathbb{D})$ iff $\delta_{LB} \leq \delta$



Multi-Step Range Query: Pseudo Code

procedure $\text{range}_\varepsilon(q, \delta, \mathbb{D})$:

$candidates\ \mathcal{C} \leftarrow \text{range}_\varepsilon(q, \delta_{LB}, \mathbb{D})$

$results\ \mathcal{R} \leftarrow \emptyset$

for $x \in \mathcal{C}$ **do**

if $\delta(q, x) \leq \varepsilon$ **then**

$\mathcal{R} \leftarrow \mathcal{R} \cup \{x\}$

return $\mathcal{R}$

- Complexity depends on search time within the index structure and number of candidates, i.e. it depends on $\text{range}_\varepsilon(q, \delta_{LB}, \mathbb{D})$
- This algorithm is complete, i.e. it holds that

$$\text{range}_\varepsilon(q, \delta, \mathbb{D}) = \text{range}_\varepsilon\big(q, \delta, \text{range}_\varepsilon(q, \delta_{LB}, \mathbb{D})\big)$$

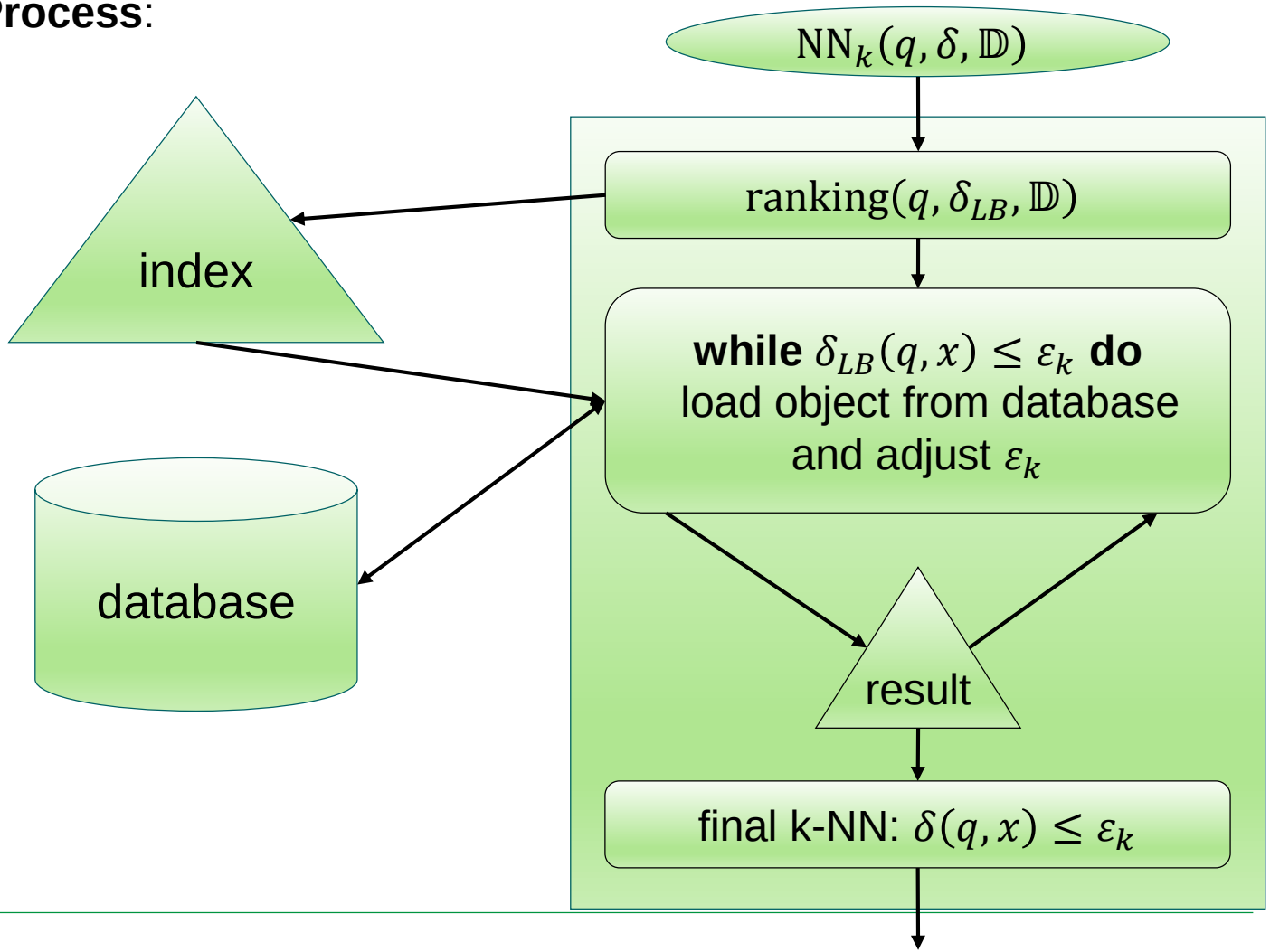
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Optimal Multi-Step k-NN Query [SK98]

- Given a set $\mathbb{X}$, a database $\mathbb{D} \subseteq \mathbb{X}$, and a distance function $\delta: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$
- How to process a query $\mathrm{NN}_k(q, \delta, \mathbb{D})$ efficiently by means of a lower bound $\delta_{LB}: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ and an optimal number of candidates?
- **Idea**:
  - Utilization of an (incremental) ranking query
  - Adaptation of k-distance ε_k after each object
- **Properties**:
  - It can be shown that the resulting algorithm is complete
  - It can be shown that the number of candidates is optimal (minimal)

Optimal Multi-Step k-NN Query cont'd

- **Process**:



Optimal Multi-Step k-NN Query: Pseudo Code

procedure $\mathrm{NN}_k(q, \delta, \mathbb{D})$:

$results\ \mathcal{R} \leftarrow \emptyset$

$filterRanking \leftarrow \mathrm{ranking}(q, \delta_{LB}, \mathbb{D})$

$x \leftarrow filterRanking.getnext()$

$\varepsilon_k \leftarrow \infty$

while $\delta_{LB}(q, x) \leq \varepsilon_k$ **do**

  if $|\mathcal{R}| < k$ **then**

    $\mathcal{R} \leftarrow \mathcal{R} \cup \{x\}$

  else if $\delta(q, x) \leq \varepsilon_k$ **then**

    $\mathcal{R} \leftarrow \mathcal{R} \cup \{x\}$

    $\mathcal{R} \leftarrow \mathcal{R} - \left\{\underset{r \in \mathcal{R}}{\mathrm{argmax}}\, \delta(q, y)\right\}$

    $\varepsilon_k \leftarrow \max_{y \in \mathcal{R}} \delta(q, y)$

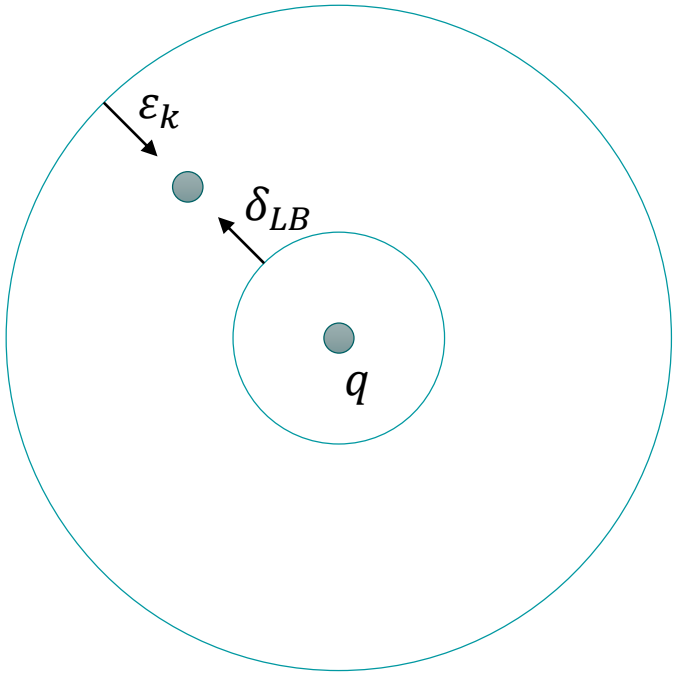
  $x \leftarrow filterRanking.getnext()$

return $\mathcal{R}$

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Optimal Multi-Step k-NN Query: Properties

- Observation:
  - k-distance ε_k decreases
  - Filter distance δ_{LB} increases
  - Algorithm terminates when $\delta_{LB} \geq \varepsilon_k$



Multi-Step Query Processing: Summary

- Lower bound (filter distance) δ_{LB} can be utilized to process similarity queries efficiently without the need of an index structure
  - $\mathrm{range}_{\varepsilon}(q, \delta, \mathbb{D})$
  - $\mathrm{NN}_{k}(q, \delta, \mathbb{D})$
- They can, however, be supported by index structures
- The multi-step query processing approach has recently been further investigated for instance by
  - Kriegel et al. [KKK+07]: Generalizing the Optimality of Multi-step k-Nearest Neighbor Query Processing
  - Houle et al. [HMN+12]: Dimensional Testing for Multi-step Similarity Search
- Question: How to define appropriate lower bounds?

Tutorial Outline

1) Object Representation
   - Feature Extraction and Representation
   - Clustering-based Computation

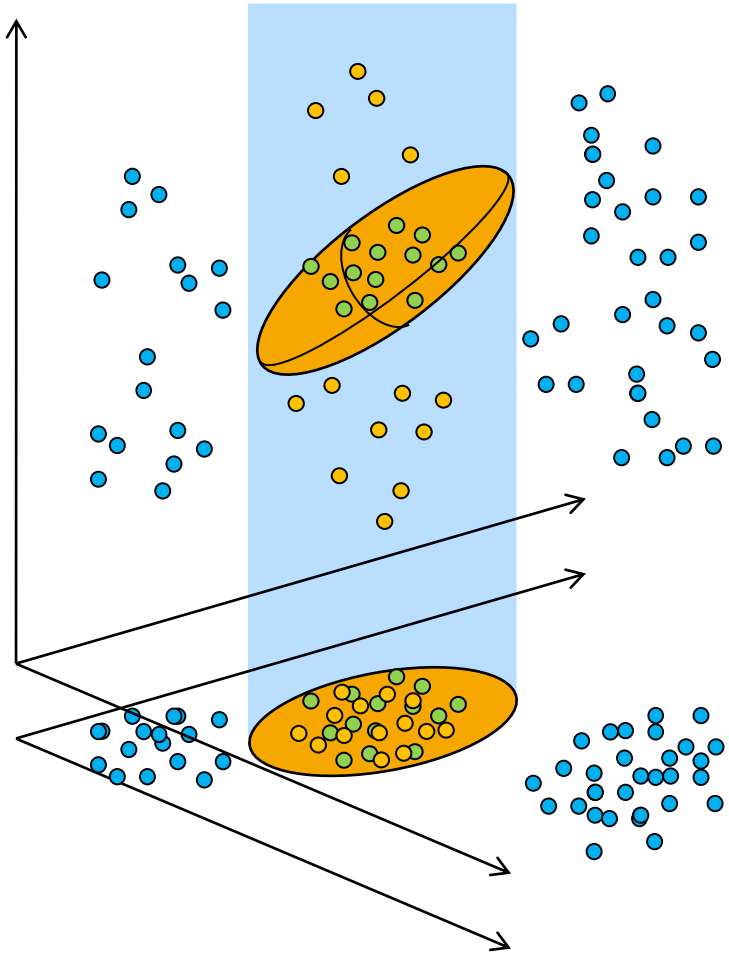
2) Fundamental Similarity Models
   - Dissimilarity Measures
   - Distance Functions for Feature Histograms
   - Distance Functions for Feature Signatures

3) **Efficient Query Processing**
   - Similarity Queries
   - Lower-Bounding: 2 examples

4) Indexing
   - Spatial Indexing
   - Metric and Ptolemaic Indexing

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Filters to accelerate similarity search



Sequential search may run too long

- Multi-step query processing
- Filter for fast candidate selection
- Refinement step for exact evaluation

[GEMINI: Faloutsos 1996; KNOP: Seidl&Kriegel 1998]

- Example: 2D filter for 3D query

ICES criteria for filter quality

I	ndexable	– Index enabled
C	omplete	– No false dismissals
E	fficient	– Fast individual calculation
S	elective	– Small candidate set

[Assent, Wenning, Seidl: ICDE 2006]

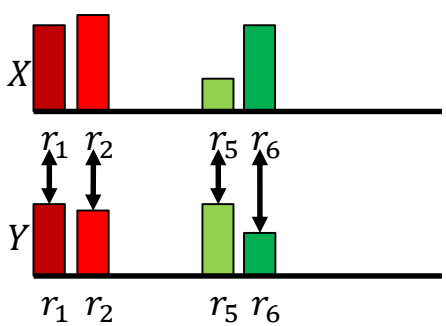
Lower Bound of Minkowski Distance

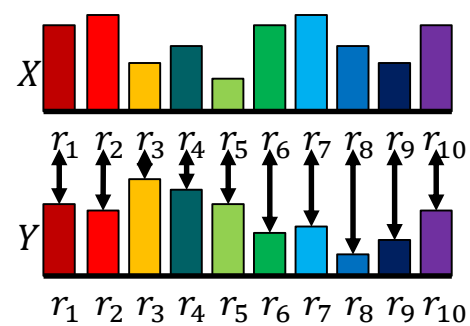
- Given two feature histograms $X, Y \in \mathbb{H}_{\mathrm{R}}$ and the Minkowski Distance

$$\mathrm{L}_p(X,Y) = \left(\sum_{f \in \mathrm{R}} |X(f) - Y(f)|^p \right)^{\frac{1}{p}}$$

- Any subset $\mathrm{R}' \subseteq \mathrm{R}$ defines a lower bound, i.e. it holds for all $X, Y \in \mathbb{H}_{\mathrm{R}}$:

$$\mathrm{L}_p(X|_{\mathrm{R}'}, Y|_{\mathrm{R}'}) = \left(\sum_{f \in \mathrm{R}'} |X(f) - Y(f)|^p \right)^{\frac{1}{p}} \leq \left(\sum_{f \in \mathrm{R}} |X(f) - Y(f)|^p \right)^{\frac{1}{p}} = \mathrm{L}_p(X,Y)$$



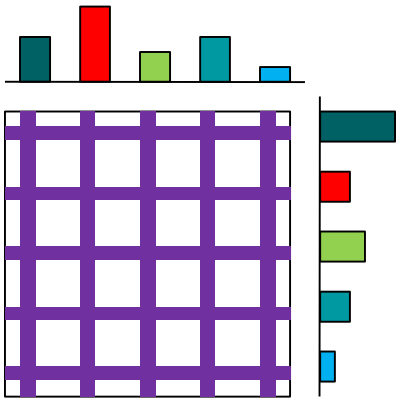


A filter for the Earth Mover's Distance

EMD 2005-10

[Assent, Wenning, Seidl: ICDE 2006] und [Uysal, Beecks, Schmücking, Seidl: CIKM 2014]

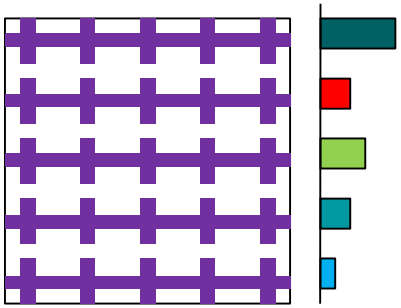
$$EMD_C(p,q) = \min_{(f_{ij})} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} f_{ij} \;\middle|\; \begin{array}{l} \forall i : \forall j : \quad f_{ij} \geq 0 \\ \forall i : \sum_{j=1}^{n} f_{ij} = p_i \\ \forall j : \sum_{i=1}^{n} f_{ij} = q_j \end{array} \right\}$$



$\geq$

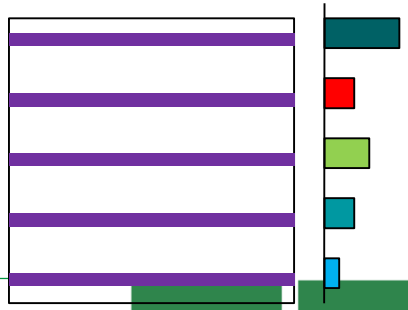
$\subseteq$ Constraint Relaxation

$$LB_{IM,C}(p,q) = \min_{(f_{ij})} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} f_{ij} \;\middle|\; \begin{array}{l} \forall i : \forall j : \quad f_{ij} \geq 0 \\ \forall i : \sum_{j=1}^{n} f_{ij} = p_i \\ \forall j : \forall i : \quad f_{ij} \leq q_j \end{array} \right\}$$



Distributivity

$$= \sum_{i=1}^{n} \min_{(f_{ij})} \left\{ \sum_{j=1}^{n} c_{ij} f_{ij} \;\middle|\; \begin{array}{l} \forall j : \quad f_{ij} \geq 0 \\ \sum_{j=1}^{n} f_{ij} = p_i \\ \forall j : \quad f_{ij} \leq q_j \end{array} \right\}$$



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Tutorial Outline

1) Object Representation
   - Feature Extraction and Representation
   - Clustering-based Computation

2) Fundamental Similarity Models
   - Dissimilarity Measures
   - Distance Functions for Feature Histograms
   - Distance Functions for Feature Signatures

3) Efficient Query Processing
   - Similarity Queries
   - Lower-Bounding: 2 examples

4) **Indexing**
   - Spatial Indexing
   - Metric and Ptolemaic Indexing

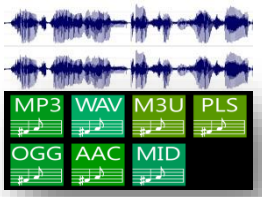
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Principle

- Given a multimedia database, how to organize and store multimedia objects on hard disk?
- Assume multimedia objects are points in a small-to-moderate dimensional Euclidean space
- Organization of multimedia objects according to their spatial proximity
- Features of multimedia objects are turned into spatial properties

text

image

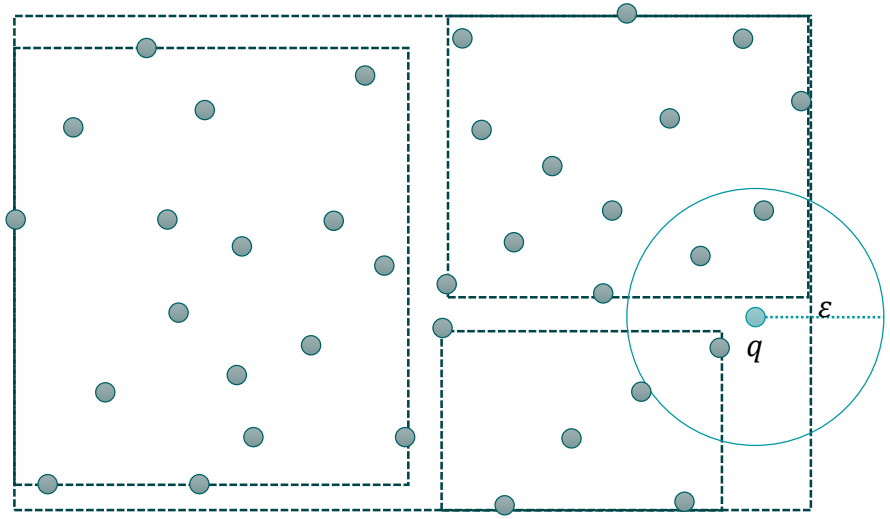


audio

video

Spatial Access Methods

- **Spatial Access Methods** (SAM) frequently organize objects in a hierarchical way
- Distance-based similarity queries are processed by taking into account the hierarchical structure



Spatial Access Methods: Approaches

- Many approaches available which differ in the way of how the space is partitioned
- Some prominent examples:
  - Main-memory access methods: KD-Tree, BSP-Tree, …
  - Space-filling curves: Z-order, Hilbert-curves, Gray-codes, …
  - Quadtree family: region, point, edge, polygonal map, …
  - Grid files
  - R-Tree and variants

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Tutorial Outline

1) Object Representation
   - Feature Extraction and Representation
   - Clustering-based Computation

2) Fundamental Similarity Models
   - Dissimilarity Measures
   - Distance Functions for Feature Histograms
   - Distance Functions for Feature Signatures

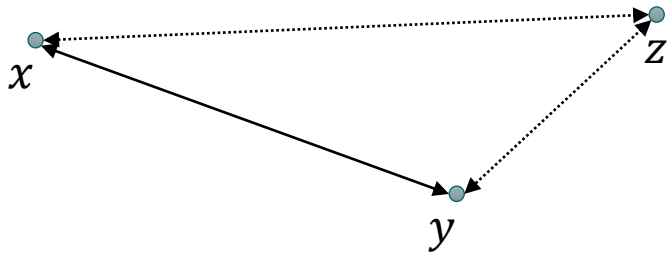
3) Efficient Query Processing
   - Similarity Queries
   - Lower-Bounding: 2 examples

4) **Indexing**
   - Spatial Indexing
   - Metric and Ptolemaic Indexing

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Metric Space Properties

- Given a metric space $(\mathbb{X}, \delta)$ how to estimate the distance $\delta: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ between two objects $x, y \in \mathbb{X}$?
  - identity of indiscernibles: $\delta(x, y) \neq 0$
  - non-negativity: $\delta(x, y) \geq 0$
  - symmetry: $\delta(x, y) = \delta(y, x)$
  - triangle inequality: $\delta(x, y) \leq \delta(x, z) + \delta(z, y)$



- Triangle inequality puts into relation three objects
- Triangle inequality is the only means that allows to estimate the distance between two objects by using another additional object

Triangle Lower Bound

- Multiple lower bounds $\delta_{z_1}^{\Delta}, \dots, \delta_{z_k}^{\Delta}$ w.r.t. objects $\{z_1, \dots, z_k\} \subseteq \mathbb{X}$ are combined to a single lower bound by using their maximum
- Let $(\mathbb{X}, \delta)$ be a metric space and $\mathbb{P} \subseteq \mathbb{X}$ be a finite set of **pivot elements**, the **triangle lower bound** $\delta_{\mathbb{P}}^{\Delta}: \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ w.r.t. $\mathbb{P}$ is defined for all $x, y \in \mathbb{X}$ as follows:

$$\delta_{\mathbb{P}}^{\Delta}(x, y) = \max_{p \in \mathbb{P}} |\delta(x, p) - \delta(p, y)|$$

- Triangle lower bound $\delta_{\mathbb{P}}^{\Delta}$ can be utilized directly in the multi-step query processing algorithm
- Direct utilization not meaningful since a single lower bound computation requires $2 \cdot |\mathbb{P}|$ distance evaluations
- Solution: precomputation of distances

Pivot Table

- The idea of a **pivot table** [N09] consists of storing the distances between each database object and each pivot element
- Originally introduced as LAESA by Micó et al. [MOV94]
- **Approach**:
  - Given a database $\mathrm{DB} = \{\mathrm{o_i}\}_{i=1}^{n}$ and a set of pivot elements $\mathbb{P} = \{p_i\}_{i=1}^{k}$
  - Pivot table $\mathcal{T} \in \mathbb{R}^{n \times k}$ stores distances between all pairs of database objects $o_i \in \mathrm{DB}$ and pivot elements $p_i \in \mathbb{P}$:

$\mathcal{T}$	$\delta(\cdot, p_1)$	...	$\delta(\cdot, p_k)$
o_1			
$\vdots$			
o_n			

  - $|\mathrm{DB}| \cdot |\mathbb{P}| = n \cdot k$ distance computations necessary prior to query processing

Pivot Table: Query Processing & Properties

- A query $q \in \mathbb{X}$ is processed as follows:
  1. Distances $\delta(q, p_i)$ are computed for all $p_i \in \mathbb{P}$
  2. Linear scan of the pivot table $\mathcal{T}$ with L_∞ to generate candidates
  3. Refinement of candidates with original distance δ

- **Properties**:
  - Pivot table is regarded as one of the most simplistic yet effective metric access method
  - It applies caching of distances
  - Due to the linear behavior, a pivot table scales for small-to-moderate size databases

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Basic Principles of Metric Indexing [H09]

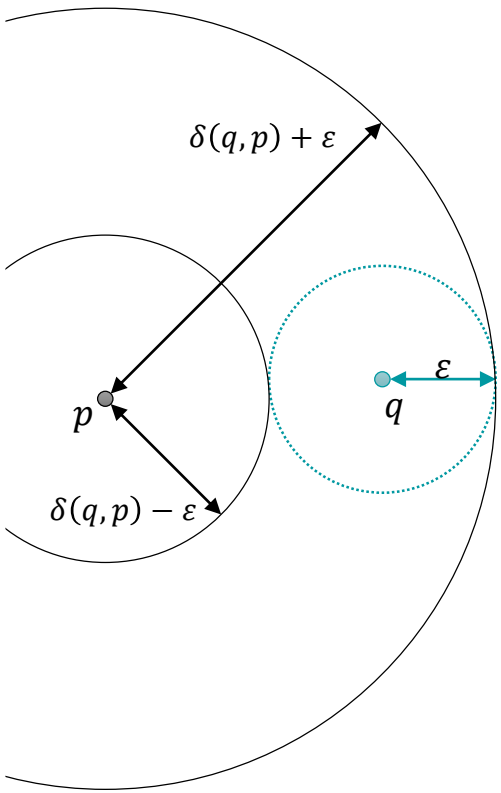
- In general, there are three main approaches to metric indexing:
  - **Pivoting**
    - Idea: Utilization of exact distances to pivot elements
  - **Ball partitioning**
    - Idea: Aggregation of information about exact distances within regions in the metric space
  - **Generalized hyperplane partitioning**
    - Idea: Utilization of generalized hyperplanes

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Pivoting

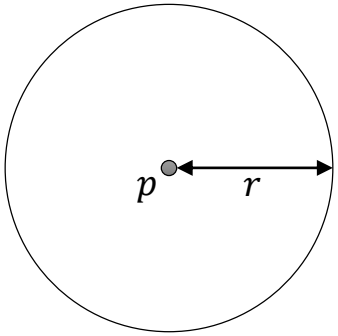
- Searching by means of precomputed distances to pivot elements $\mathbb{P}$ and the triangle lower bound $\delta_{\mathbb{P}}^{\Delta}$

- **Filtering Principle** for $\mathbb{P} = \{p\}$ and range query with $\varepsilon \in \mathbb{R}^{+}$:
  - Objects o inside the inner ball around p are filtered out because it holds that $\delta(q,p) - \delta(p,o) > \varepsilon$
  - Objects o outside the outer ball around p are filtered out because it holds that $\delta(p,o) - \delta(q,p) > \varepsilon$
  - Thus only objects o inside the shell between the two balls are candidates because it holds that $\delta_{\mathbb{P}}^{\Delta} = |\delta(q,p) - \delta(p,o)| \leq \varepsilon$



Ball Partitioning

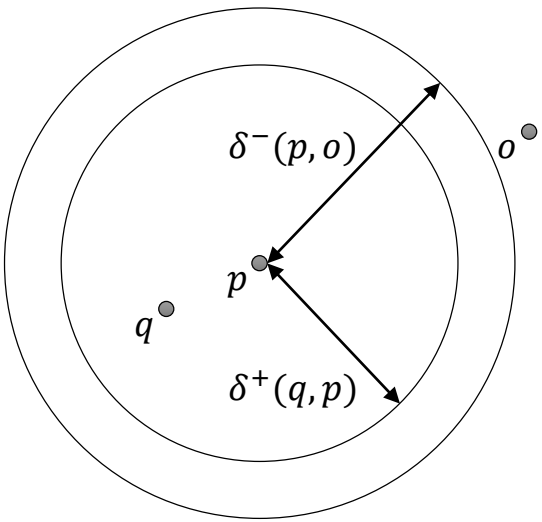
- Reduction of information available about each object
- Instead of storing the exact distance from an object o to all pivot elements $\mathbb{P}$, the object is placed in a **region** that is defined by a pivot element
- Each pivot element $p \in \mathbb{P}$ defines a **metric ball** around p with **covering radius** $r \in \mathbb{R}^+$
- Each single metric ball corresponds to two regions:



Ball Lemma

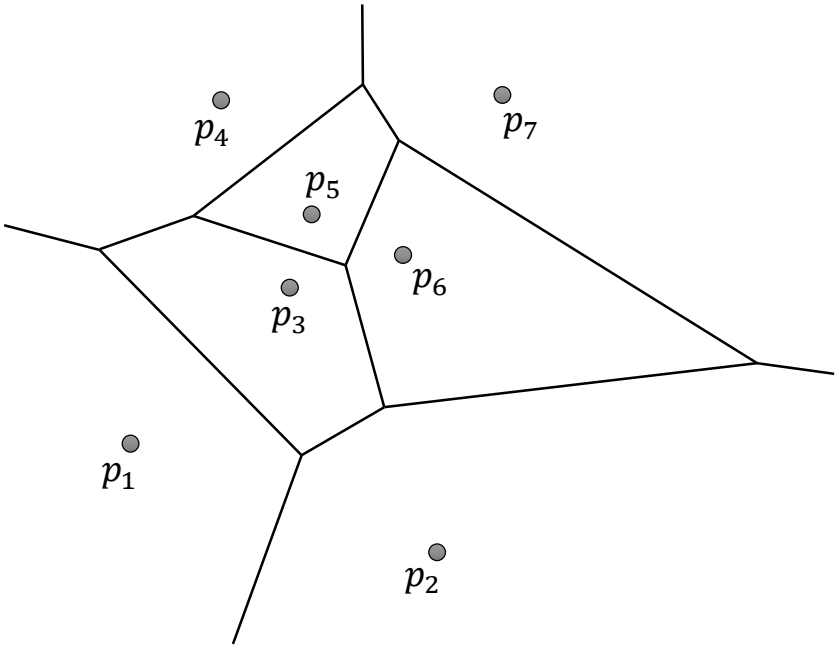
- Let $(\mathbb{X}, \delta)$ be a metric space and let the distance $\delta: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ between any two objects $x, y \in \mathbb{X}$ be in the interval $\delta(x, y) \in [\delta^-(x, y), \delta^+(x, y)]$
- The distance $\delta(q, o)$ can be bounded for any $q, o, p \in \mathbb{X}$ as follows:

$$\max\{0, \delta^-(p, o) - \delta^+(p, q), \delta^-(q, p) - \delta^+(o, p)\} \leq \delta(q, o) \leq \delta^+(q, p) + \delta^+(p, o)$$



Generalized Hyperplane Partitioning

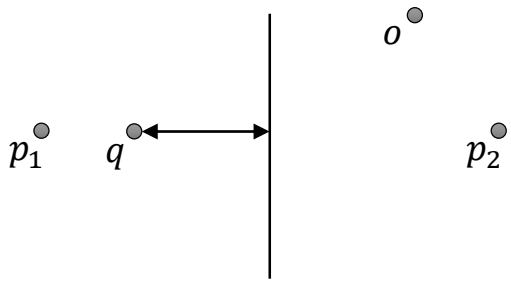
- A **generalized hyperplane** is a set of objects for which the pivot elements are equidistant
- Each object o is assigned to its closest pivot element $p_i \in \mathbb{P}$
- No distances and cover radii are maintained or used for searching
- Example for a multi-way partitioning with $\mathbb{P} = \{p_i\}_{i=1}^{7}$:
  - Objects belong to Voronoi cells
  - $\delta(q, o)$ can be lower bounded by means of the **plane lemma** (see next slide)



Plane Lemma

- Let $(\mathbb{X}, \delta)$ be a metric space, let $o, q \in \mathbb{X}$ be objects, and let $p_1, p_2 \in \mathbb{P}$ be two pivot elements
- Let further $\delta(o, p_2) \leq \delta(p_1, o)$
- The distance $\delta(q, o)$ can be bounded as follows:

$$\max\left\{\frac{\delta(q, p_2) - \delta(q, p_1)}{2}, 0\right\} \leq \delta(q, o)$$



Overview of Metric Indexing Methods (1) [H09]

		Structure					Functionality		
Method		P	BS	VP	GH	MW	D	CPU	I/O
AESA	40–42	●							
AHC	43	●		●				●	
Antipole Tree	44	●	●		○	○		●	
BS-Tree	45		●		○		●	●	
BU-Tree	46		●		○			●	
CM-Tree	47	●	●		○	●	●	●	●
DBM-Tree	48, 49	○	●		○	●	●	●	●
DF-Tree	37	●	●		○	●	●	●	●
D-Index	50–53	●		●		●	○	●	●

Overview of Metric Indexing Methods (2) [H09]

		Structure					Functionality		
Method		P	BS	VP	GH	MW	D	CPU	I/O
DSA-Tree	54–56				●	●	●	●	○
DSAP-Tree	57–59	●			●	●	●	●	○
DVP-Tree	60			●		●	●	●	●
EGNAT	2	○		●	○	●	●	●	●
EHC	43			●		○		●	
EM-VP-Forest	61			●				●	
GH-Tree	62, 63				●		●	●	
GNAT	64			●	○	●		●	
HC	43, 65			●				●	
iAESA	66	●							
iDistance	67	○	●		○	●	●	●	●

Overview of Metric Indexing Methods (3) [H09]

		Structure					Functionality		
Method		P	BS	VP	GH	MW	D	CPU	I/O
LAESA	8, 40, 68, 69	●					●		○
LC	70, 71			●				●	○
Linear Scan	72						●		
MB$^+$-Tree	73			●		●	●	●	●
MB*-Tree	74, 75		●		○			●	
M*-Tree	76	●	●		○	●	●	●	●
M-Tree	77–80	○	●		○	●	●	●	●
MVP-Tree	81, 82	●		●		●	●	●	●
Omni	1	●					●	●	●
Opt-VP-Tree	83			●		●	●	●	●

Overview of Metric Indexing Methods (4) [H09]

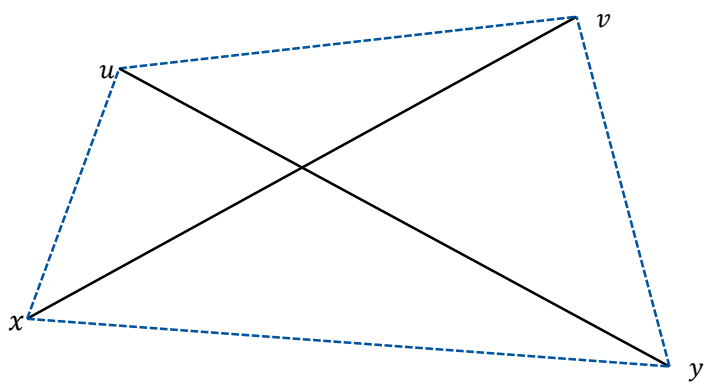
		Structure					Functionality		
Method		P	BS	VP	GH	MW	D	CPU	I/O
PM*-Tree	76	●	●		○	●	●	●	●
PM-Tree	84	●	●		○	●	●	●	●
ROAESA	85	●						●	
SA-Tree	86, 87				●	●		●	
Slim-Tree	36, 88	○	●		○	●	●	●	●
Spaghettis	89	●					●	●	
SSS-Tree	90		●		●	●		●	
SSS-LC	91	●		●			○	●	●
TLAESA	92, 93	●					●	●	
Voronoi-Tree	94		●		○	○	●	●	
VP-Tree	62, 63, 95			●				●	

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Ptolemaic Indexing [H09b, HSL+13]

- A **Ptolemaic metric distance function** $\delta: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}^{\geq 0}$ over a set $\mathbb{X}$ satisfies the following properties:
  - identity of indiscernibles: $\forall x, y \in \mathbb{X}: \delta(x, y) = 0 \Leftrightarrow x = y$
  - non-negativity: $\forall x, y \in \mathbb{X}: \delta(x, y) \geq 0$
  - symmetry: $\forall x, y \in \mathbb{X}: \delta(x, y) = \delta(y, x)$
  - triangle inequality: $\forall x, y, z \in \mathbb{X}: \delta(x, y) \leq \delta(x, z) + \delta(z, y)$
  - Ptolemy's inequality: $\forall u, v, x, y \in \mathbb{X}$:

$$\delta(x, v) \cdot \delta(y, u) \leq \delta(x, y) \cdot \delta(u, v) + \delta(x, u) \cdot \delta(y, v)$$



Ptolemaic Lower Bound

- Let $(\mathbb{X}, \delta)$ be a metric space and $\mathbb{P} \subseteq \mathbb{X}$ be a finite set of **pivot elements**, the **Ptolemaic lower bound** $\delta_{\mathbb{P}}^{\mathrm{Pto}}: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ w.r.t. $\mathbb{P}$ is defined for all $x, y \in \mathbb{X}$ as follows:

$$\delta_{\mathbb{P}}^{\mathrm{Pto}}(x, y) = \max_{p_i, p_j \in \mathbb{P}} \frac{\left|\delta(x, p_i) \cdot \delta(y, p_j) - \delta(x, p_j) \cdot \delta(y, p_i)\right|}{\delta(p_i, p_j)}$$

- $\delta_{\mathbb{P}}^{\mathrm{Pto}}$ involves all pairs of pivot elements
- Each computation of $\delta_{\mathbb{P}}^{\mathrm{Pto}}$ entails $5 \cdot \binom{|\mathbb{P}|}{2}$ distance computations
- Problem of distance caching becomes more apparent

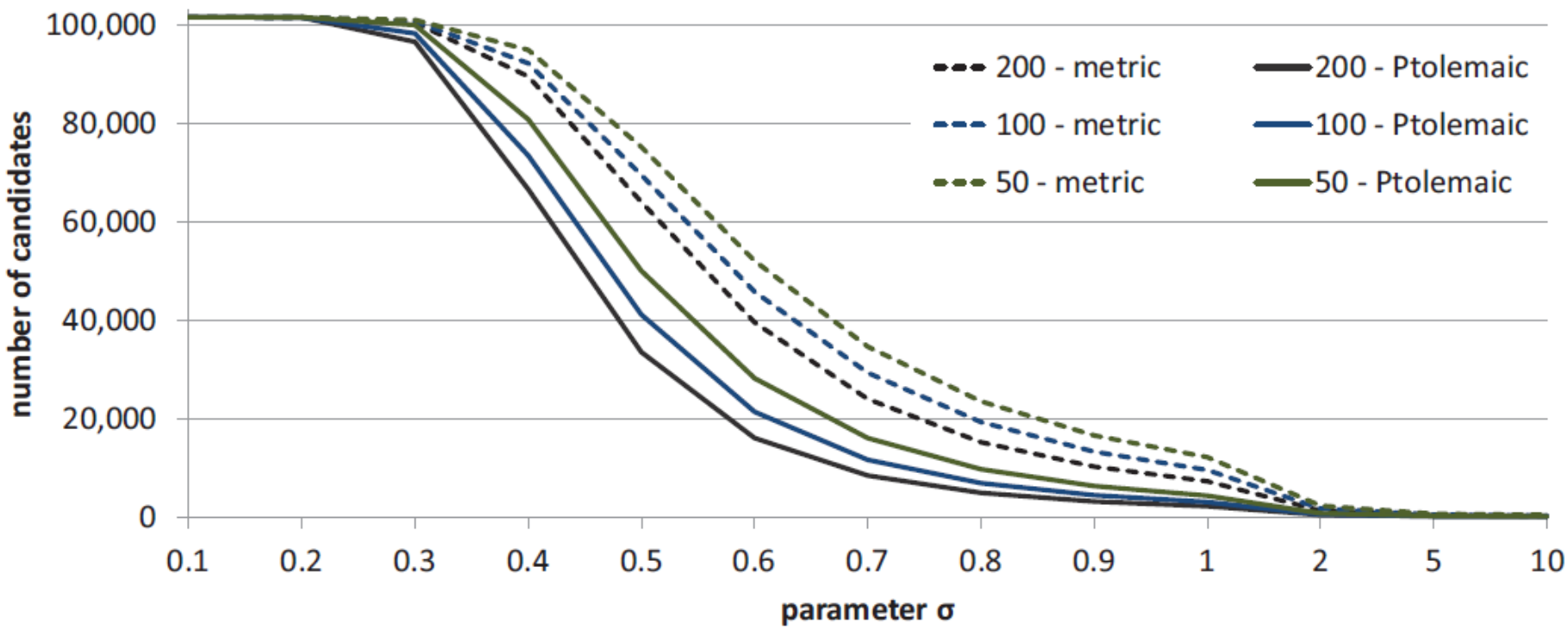
Ptolemaic Lower Bound: Properties

- The examination of all pivot pairs is too inefficient
- Different pivot evaluation heuristics which follow the idea of minimizing $\delta(x, p_j) \cdot \delta(y, p_i)$ in the numerator [LHS+11, HSL+13]:
  - **Unbalanced heuristic**
    - Examining those pivots $p_i, p_j \in \mathbb{P}$ which are either close to x or to y
  - **Balanced heuristic**
    - Examining those pivots $p_i, p_j \in \mathbb{P}$ which are close to both x and y
  - Both heuristics rely on storing the corresponding pivot permutations for each database object in order to approximate $\delta_{\mathbb{P}}^{\mathrm{Pto}}$ efficienty
- Ptolemaic lower bound can be integrated in many metric access methods
  - Ptolemaic Pivot Table
  - Ptolemaic PM-Tree
  - Ptolemaic M-Index

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

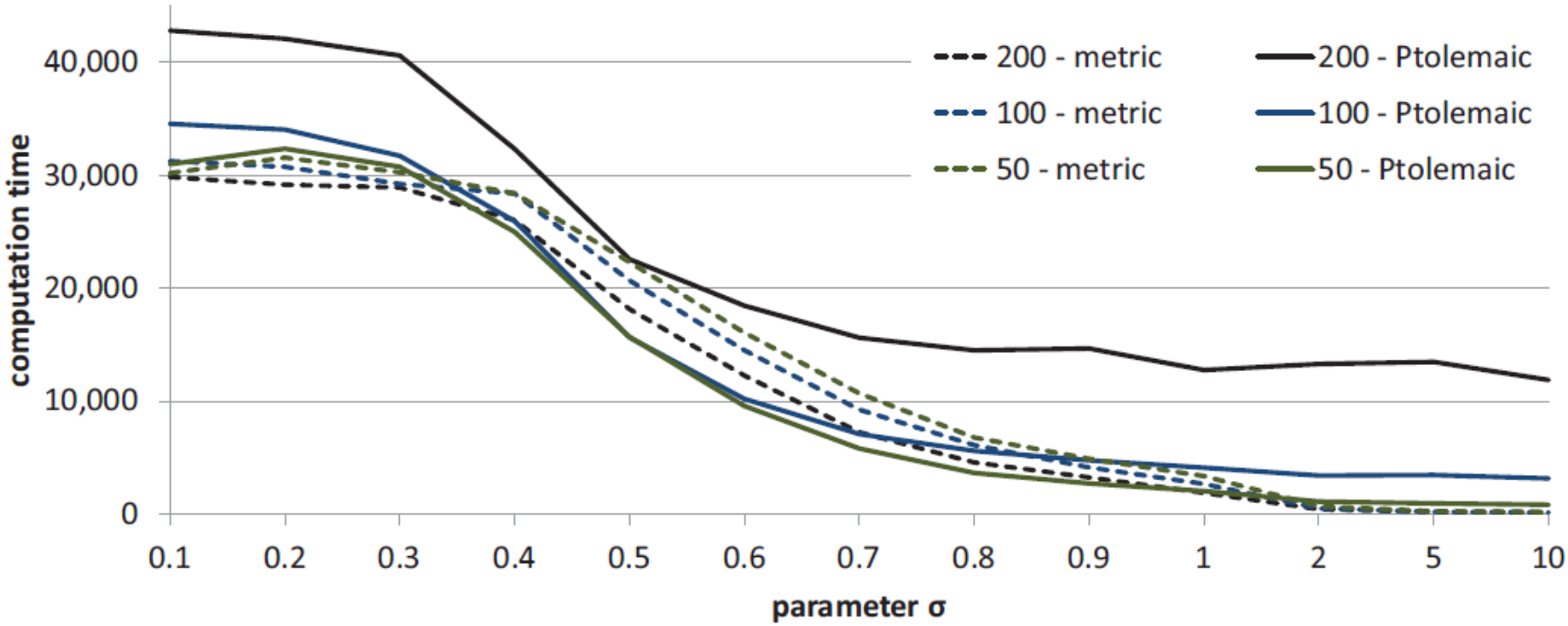
Lower Bounds: Performance

- Comparison of lower bounds $\delta_{\mathbb{P}}^{\Delta}$ and $\delta_{\mathbb{P}}^{\mathrm{Pto}}$ with respect to the $\mathrm{SQFD}_{k_{Gaussian(\sigma)}}$ on an image database comprising $\sim 100k$ feature signatures of cardinality 40



Lower Bounds: Performance cont'd

- Comparison of lower bounds $\delta_{\mathbb{P}}^{\Delta}$ and $\delta_{\mathbb{P}}^{\mathrm{Pto}}$ with respect to the $\mathrm{SQFD}_{k_{Gaussian(\sigma)}}$ on an image database comprising $\sim 100k$ feature signatures of cardinality 40



Summary

- Depending on multimedia objects, different indexing approaches are feasible
- Spatial access methods are useful for multimedia objects, whose properties can be expressed in a low-dimensional Euclidean space
- Metric access methods can deal with “non-dimensional” data
- Earth Mover’s Distance and Signature Quadratic Form Distance satisfy the metric properties
- Signature Quadratic Form Distance additionally satisfies the Ptolemy inequality

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

What was this tutorial about?

- **Object representations**
  - *How to model and represent multimedia data?*
- **Fundamental similarity models for multimedia data**
  - *What is a distance-based similarity model?*
- **Efficient query processing**
  - *How to process distance-based similarity queries efficiently?*
- **Indexing**
  - *How to index spatial and high-dimensional multimedia data?*
  - *What are the principles behind metric and Ptolemaic indexing approaches?*

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

References (1)

[AF06] A. E. Abdel-Hakim, A. A. Farag: CSIFT: A SIFT Descriptor with Color Invariant Characteristics. CVPR (2), 1978-1983 (2006)

[AP88] Ashby, F. G. & Perrin, N. A. Toward a unified theory of similarity and recognition. Psychological Review, 1988, 95, 124-150

[BET+08] H. Bay, A. Ess, T. Tuytelaars, L. J. Van Gool: Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding 110(3), 346-359 (2008)

[Beecks13] Beecks, C.: Distance-based similarity models for content-based multimedia retrieval. PhD thesis, RWTH Aachen University, 2013.

[BIK+11] Beecks, C.; Ivanescu, A. M.; Kirchhoff, S. & Seidl, T. Modeling Image Similarity by Gaussian Mixture Models and the Signature Quadratic Form Distance. Proceedings of the IEEE International Conference on Computer Vision, 2011, 1754-1761

[BIK+11a] Beecks, C.; Ivanescu, A. M.; Kirchhoff, S. & Seidl, T. Modeling Multimedia Contents through Probabilistic Feature Signatures. Proceedings of the ACM International Conference on Multimedia, 2011, 1433-1436

[BKS13] Beecks, C.; Kirchhoff, S. & Seidl, T. Signature Matching Distance for Content-based Image Retrieval. Proceedings of the ACM International Conference on Multimedia Retrieval, 2013, 41-48

[BLS+11] Beecks, C.; Lokoč, J.; Seidl, T. & Skopal, T. Indexing the signature quadratic form distance for efficient content-based multimedia retrieval. Proceedings of the ACM International Conference on Multimedia Retrieval, 2011, 24:1-8

[BUS10] Beecks, C.; Uysal, M. S. & Seidl, T. Signature Quadratic Form Distance. Proceedings of the ACM International Conference on Image and Video Retrieval, 2010, 438-445

[BCK08] Boriah, S.; Chandola, V. & Kumar, V. Similarity Measures for Categorical Data: A Comparative Evaluation. Proceedings of the SIAM International Conference on Data Mining, 2008, 243-254

[BTB05] Boughorbel, S.; Tarel, J.-P. & Boujemaa, N. Conditionally Positive Definite Kernels for SVM Based Image Recognition. Proceedings of the IEEE International Conference on Multimedia and Expo, 2005, 113-116

[CLÖ12] M. Calonder, V. Lepetit, M. Özuysal, T. Trzcinski, C. Strecha, P. Fua: BRIEF: Computing a Local Binary Descriptor Very Fast. IEEE Trans. Pattern Anal. Mach. Intell. 34(7), 1281-1298 (2012)

[CVG+08] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, M. Slaney: Content-based music information retrieval: Current directions and future challenges. Proceedings of the IEEE 96.4: 668-696 (2008)

[CTC+12] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, Y. A. Reznik, R. Grzeszczuk, B. Girod: Compressed Histogram of Gradients: A Low-Bitrate Descriptor. International Journal of Computer Vision 96(3), 384-399 (2012)

[CHV99] Chapelle, O.; Haffner, P. & Vapnik, V. Support vector machines for histogram-based image classification. IEEE Transactions on Neural Networks, 1999, 10, 1055-1064

[DJL+08] R. Datta, D. Joshi, J. Li, J. Z. Wang: Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40(2) (2008)

References (2)

[DLR77] A. P. Dempster, N. M. Laird, D. B. Rubin: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1), 1-38 (1977)

[DD09] Deza, M. & Deza, E. Encyclopedia of distances. Springer Verlag, 2009

[FBF+94] Faloutsos, C.; Barber, R.; Flickner, M.; Hafner, J.; Niblack, W.; Petkovic, D. & Equitz, W. Efficient and Effective Querying by Image Content. Journal of Intelligent Information Systems, 1994, 3, 231-262

[Gärtner03] Gärtner, T. A survey of kernels for structured data. SIGKDD Explorations, 2003, 5, 49-58

[GW02] R. C. Gonzalez and R. E. Woods: Digital image processing. Prentice Hall (2002)

[HSE+95] Hafner, J. L.; Sawhney, H. S.; Equitz, W.; Flickner, M. & Niblack, W. Efficient Color Histogram Indexing for Quadratic Form Distance Functions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995, 17, 729-736

[H14] Hausdorff, F. Grundzüge der Mengenlehre. Von Veit, 1914

[HSL+12] Hetland, M. L.; Skopal, T.; Lokoč, J. & Beecks, C. Ptolemaic access methods: Challenging the reign of the metric space model. Information Systems, 2013, 38, 989 - 1006

[HL90] Hillier, F. & Lieberman, G. Introduction to Linear Programming. McGraw-Hill, 1990

[HSS08] Hofmann, T.; Schölkopf, B. & Smola, A. Kernel methods in machine learning. The annals of statistics, JSTOR, 2008, 1171-1220

[HRS+08] Hu, R.; Rüger, S. M.; Song, D.; Liu, H. & Huang, Z. Dissimilarity measures for content-based image retrieval. Proceedings of the IEEE International Conference on Multimedia and Expo, 2008, 1365-1368

[HXL+11] W. Hu, N. Xie, L. Li, X. Zeng, S. J. Maybank: A Survey on Visual Content-Based Video Indexing and Retrieval. IEEE Transactions on Systems, Man, and Cybernetics, Part C 41(6): 797-819 (2011)

[HKR93] Huttenlocher, D. P.; Klanderman, G. A. & Rucklidge, W. Comparing Images Using the Hausdorff Distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1993, 15, 850-863

[Ioka89] Ioka, M. A Method of Defining the Similarity of Images on the Basis of Color Information. IBM Research, Tokyo Research Laboratory, 1989

[I75] Itakura, F. Minimum prediction residual principle applied to speech recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, IEEE, 1975, 23, 67-72

[JSW08] Jäkel, F.; Schölkopf, B. & Wichmann, F. A. Similarity, kernels, and the triangle inequality. Journal of Mathematical Psychology, 2008, 52, 297 – 303

References (3)

[JDS08] Jegou, H.; Douze, M. & Schmid, C. Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search. Proceedings of the European Conference on Computer Vision, 2008, 304-317

[JF87] Jones, W. P. & Furnas, G. W. Pictures of relevance: a geometric analysis of similarity measures. Journal of the American Society for Information Science, John Wiley & Sons, Inc., 1987, 38, 420-442

[KS04] Y. Ke, R. Sukthankar: PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. CVPR (2), 506-513 (2004)

[KLB+11] Kruliš, M.; Lokoč, J.; Beecks, C.; Skopal, T. & Seidl, T. Processing the signature quadratic form distance on many-core GPU architectures. Proceedings of the ACM Conference on Information and Knowledge Management, 2011, 2373-2376

[KSL+12] Kruliš, M.; Skopal, T.; Lokoč, J. & Beecks, C. Combining CPU and GPU architectures for fast similarity search. Distributed and Parallel Databases, 2012, 30, 179-207

[Krumhansl78] Krumhansl, C. Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. Psychological Review, Elsevier, 1978, 85, 445-463

[KL51] Kullback, S. & Leibler, R. A. On Information and Sufficiency. The Annals of Mathematical Statistics, Institute of Mathematical Statistics, 1951, 22, 79-86

[LL04] Leow, W. K. & Li, R. The analysis and applications of adaptive-binning color histograms. Computer Vision and Image Understanding, 2004, 94, 67-91

[L66] Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals .Soviet Physics Doklady, 1966, 10, 707-710

[LCS11] S. Leutenegger, M. Chli, R. Siegwart: BRISK: Binary Robust invariant scalable keypoints. ICCV, 2548-2555 (2011)

[LHS+11] Lokoč, J.; Hetland, M. L.; Skopal, T. & Beecks, C. Ptolemaic indexing of the signature quadratic form distance. Proceedings of the International Conference on Similarity Search and Applications, 2011, 9-16

[Lowe04] D. G. Lowe: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91-110 (2004)

[MacAdam42] D. L. MacAdam: Visual sensitivities to color differences in daylight. *JOSA* 32.5, 247-273 (1942)

[MacQueen67] J. B. MacQueen: Some Methods for Classification and Analysis of MultiVariate Observations. Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, 281-297 (1967)

[MRS08] C. D. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval, Cambridge University Press (2008)

[McGill79] McGill, M. An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems. 1979

[MZB10] D. Mitrovic, M. Zeppelzauer, C. Breiteneder: Features for Content-Based Audio Retrieval. Advances in Computers 78: 71-150 (2010)

References (4)

[NBE+93] Niblack, W.; Barber, R.; Equitz, W.; Flickner, M.; Glasman, E.; Petkovic, D.; Yanker, P.; Faloutsos, C. & Taubin, G. QBIC project: querying images by content, using color, texture, and shape. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 1993, 1908, 173-187

[NS06] Nistér, D. & Stewénius, H. Scalable Recognition with a Vocabulary Tree. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, 2161-2168

[OD12] I. R. Otero, M. Delbracio: The Anatomy of the SIFT Method. Image Processing On Line, 2012

[PLL08] Park, B. G.; Lee, K. M. & Lee, S. U. Color-Based Image Retrieval Using Perceptually Modified Hausdorff Distance. EURASIP Journal of Image and Video Processing, 2008

[PHB97] Puzicha, J.; Hofmann, T. & Buhmann, J. M. Non-parametric Similarity Measures for Unsupervised Texture Segmentation and Image Retrieval. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 1997, 267-272

[PRT+99] Puzicha, J.; Rubner, Y.; Tomasi, C. & Buhmann, J. M. Empirical Evaluation of Dissimilarity Measures for Color and Texture. Proceedings of the IEEE International Conference on Computer Vision, 1999, 1165-1172

[RRK+11] E. Rublee, V. Rabaud, K. Konolige, G. R. Bradski: ORB: An efficient alternative to SIFT or SURF. ICCV, 2564-2571 (2011)

[RPT+01] Rubner, Y.; Puzicha, J.; Tomasi, C. & Buhmann, J. M. Empirical Evaluation of Dissimilarity Measures for Color and Texture. Computer Vision and Image Understanding, 2001, 84, 25-43

[RTG00] Rubner, Y.; Tomasi, C. & Guibas, L. J. The Earth Mover's Distance as a Metric for Image Retrieval. International Journal of Computer Vision, 2000, 40, 99-121

[SC78] Sakoe, H. & Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, IEEE, 1978, 26, 43-49

[SJ99] Santini, S. & Jain, R. Similarity Measures. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21, 871-883

[Schölkopf01] Schölkopf, B. The kernel trick for distances. Advances in neural information processing systems, MIT; 1998, 2001, 301-307

[Shepard57] Shepard, R. N. Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. Psychometrika, Springer New York, 1957, 22, 325-345

[SJ08] Shirdhonkar, S. & Jacobs, D. W. Approximate earth mover's distance in linear time. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2008

[TMY78] H. Tamura, S. Mori, T. Yamawaki: Textural Features Corresponding to Visual Perception. IEEE Transactions on Systems, Man and Cybernetics, vol.8, no.6, 460-473 (1978)

[Tversky77] Tversky, A. Features of similarity. Psychological review, American Psychological Association, 1977, 84, 327-352

[ZL03] Zhang, D. & Lu, G. Evaluation of similarity measurement for image retrieval. Proceedings of the IEEE International Conference on Neural Networks and Signal Processing, 2003, 2, 928 - 931

References (5)

[ABK+98] Ankerst, M.; Braunmüller, B.; Kriegel, H.-P. & Seidl, T. Improving Adaptable Similarity Query Processing by Using Approximations. Proceedings of the International Conference on Very Large Data Bases, 1998, 206-217

[BKS01] Börzsönyi S., D. Kossmann, and K. Stocker. The Skyline Operator. In Proceedings of the 17th International Conference on Data Engineering. 421-430. 2001.

[BNC03] B. Bustos, G. Navarro, E. Chávez: Pivot selection techniques for proximity searching in metric spaces. Pattern Recognition Letters 24(14): 2357-2366 (2003)

[CP00] Ciaccia, P., and M. Patella. PAC nearest neighbor queries: Approximate and controlled search in high-dimensional and metric spaces. In Proceedings of 16th International Conference on Data Engineering. 244-255. 2000

[FRM94] Faloutsos C., Ranganathan M., Manolopoulos Y. Fast Subsequence Matching in Time-Series Databases. Proc. ACM SIGMOD Int. Conf. on Management of Data, 1994, 419-429

[HSE+95] Hafner, J. L.; Sawhney, H. S.; Equitz, W.; Flickner, M. & Niblack, W. Efficient Color Histogram Indexing for Quadratic Form Distance Functions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995, 17, 729-736

[H09] Hetland. The Basic Principles of Metric Indexing. Swarm Intelligence for Multi-objective Problems in Data Mining, Springer Berlin Heidelberg, 2009, 242, 199-232

[H09b] Hetland, M. L. Ptolemaic Indexing CoRR, 2009, abs/0905.4627

[HSL+13] Hetland, M. L.; Skopal, T.; Lokoč, J. & Beecks, C. Ptolemaic access methods: Challenging the reign of the metric space model. Information Systems, 2013, 38, 989 – 1006

[HMN+12] Houle, M. E.; Ma, X.; Nett, M. & Oria, V. Dimensional Testing for Multi-step Similarity Search. Proceedings of the IEEE International Conference on Data Mining, 2012, 299-308

[IBS08] Ilyas, I. F., G. Beskales, and M. A. Soliman. A survey of top-k query processing techniques in relational database systems. ACM Computing Surveys (CSUR) 40.4 (2008): 11.

[KKK+07] Kriegel, H.-P.; Kröger, P.; Kunath, P. & Renz, M. Generalizing the Optimality of Multi-step it k -Nearest Neighbor Query Processing. Proceedings of the International Symposium on Spatial and Temporal Databases, 2007, 75-92

[KM00] Korn, F., and S. Muthukrishnan. Influence sets based on reverse nearest neighbor queries. ACM SIGMOD Record 29.2 (2000): 201-212.

[LHS+11] Lokoč, J.; Hetland, M. L.; Skopal, T. & Beecks, C. Ptolemaic indexing of the signature quadratic form distance. Proceedings of the International Conference on Similarity Search and Applications, 2011, 9-16

[MOV94] Micó, L.; Oncina, J. & Vidal, E. A new version of the nearest-neighbour approximating and eliminating search algorithm (AESA) with linear preprocessing time and memory requirements. Pattern Recognition Letters, 1994, 15, 9-17

[N09] Navarro, G. Analyzing Metric Space Indexes: What For?. Proceedings of the International Workshop on Similarity Search and Applications, 2009, 3-10

[RSC+13] Ruiz G., Santoyo F., Chávez E., Figueroa K., Tellez E. S.: Extreme Pivots for Faster Metric Indexes. SISAP 2013: 115-126

[SH94] Sawhney H., Hafner J.: Efficient Color Histogram Indexing. Proc. Int. Conf. on Image Processing (ICIP): 66-70, 1994

[SK97] Seidl, T. & Kriegel, H.-P. Efficient User-Adaptable Similarity Search in Large Multimedia Databases. Proceedings of the International Conference on Very Large Data Bases, 1997, 506-515

[SK98] Seidl T., Kriegel H.-P. Optimal Multi-Step k-Nearest Neighbor Search. ACM SIGMOD 1998, 154-165

[SBL11] Skopal, T.; Bartos, T. & Lokoč, J. On (not) indexing quadratic form distance by metric access methods. Proceedings of the International Conference on Extending Database Technology, 2011, 249-258

Thanks to my PhD students

- Anna Beer
- Janina Bleicher
- Julian Busch
- Daniyal Kazempour
- Yifeng Lu
- Florian Richter
- Sebastian Schmoll

Further and former members

- Prof. Dr. Hans-Peter Kriegel (i.R.)
- Prof. Dr. Christian Böhm
- Prof. Dr. Volker Tresp (Hon.)
- Prof. Dr. Peer Kröger (apl.)
- Prof. Dr. Matthias Schubert (apl.)
- Dr. Tobias Emrich (DSLab)

… plus their PhD students

… and many Bachelor and Master students!

- Dr.-Ing. M. Seran **Uysal** (vsl. 2016), RWTH Aachen U
- Dr.-Ing. Roland Assam (2015), G&D, Munich
- Prof. Dr.-Ing. Marwan Hassani (2015), TU Eindhoven, NL
- Dr. Ines Färber (2014), P3 group, Aachen
- Dr. Sergej Fries (2014), P3 group, Aachen
- Dr. Brigitte Boden (2014), DLR, Cologne
- Dr. Anca Zimmer (2013), Heidenhain, Traunreut
- Dr. Hardy Kremer (2013), Deloitte, Berlin
- Dr. Christian **Beecks** (2013), RWTH Aachen U
- Prof. Dr. Stephan Günnemann (2012), TUM, Munich
- Dr. Philipp Kranen (2011), Microsoft, Munich
- Dr. Marc Wichterich (2010), Amazon, USA
- Prof. Dr. Emmanuel Müller (2010), U Potsdam
- Dr. Ralph Krieger (2008), Avanade
- Dr. Christoph Brochhaus (2008), Bosch/Samsung
- Prof. Dr. Ira Assent (2008), U Aarhus, Denmark

Questions?

wordle.net

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN