**Ludwig-Maximilians-Universität München**
**Institut für Informatik**
Prof. Dr. Thomas Seidl
Max Berrendorf, Julian Busch

## Knowledge Discovery and Data Mining I

WS 2018/19

## Exercise 2: Feature Extraction, Similarity Search, Data Reduction

### Exercise 2-1 Feature Extraction and Similarity Search

Consider the 5 images in Figure 1 with 36 pixels each.

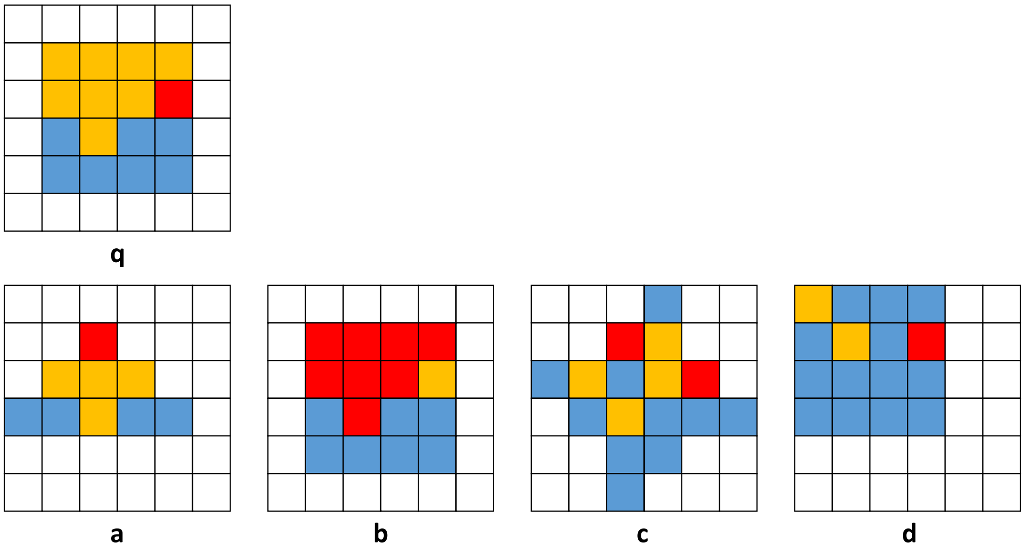

Abbildung 1: 6 × 6 Pixel Images.

(a) Extract the following features from all images (the white pixels can be ignored):

- Color features: Use a color histogram with bins *red, orange* and *blue*.

(b) Which images are most similar to $q$ w.r.t color when the Euclidean distance is used? Provide a ranking.

(c) If you want to find only the top-$k$ most similar images to $q$ in a database, is there a more efficient way than computing the distance $d$ between $q$ and all images in the database? Name and describe two different approaches to this problem.

(d) The query result might not be fully satisfying. Can you think of different ways of modifying the feature extraction or distance function to obtain possibly better results? Provide at least one alternative for both components.

## Exercise 2-2 Incremental Aggregation

Given a Data Warehouse with e.g. 10 million entries, additional 1000 entries arrive each day. Rather than recomputing the desired aggregates, an incremental adaptation to the new data should be supported. In order to accelerate the (re-)computation, precomputed intermediate results shall be stored and intermediate results for the new entries shall be computed. What (and how many) values suffice when considering the following aggregates? For each measure note whether it is an algebraic, holistic or distributive measure.

(a) Product.

(b) Mean.

(c) Variance.

(d) Median.

## Exercise 2-3 Histograms

(a) Given are the following data points:

$$1; 1; 4; 4; 5; 5; 5; 5; 6; 7; 7; 9$$

Using three bins, compute

(i) the equi-width histogram

(ii) the equi-height histogram

(b) Assume there is an additional data point 29. What changes in the histograms compared to (a)?