

The background of the slide is a complex, abstract composition. It features a network of thin, reddish-brown lines connecting various points, some of which are green dots. There are also faint, light-colored geometric shapes and patterns, including a grid of small plus signs in the top-left and bottom-left corners. A large, white, angular shape is positioned behind the main text, creating a sense of depth. The overall color palette is muted, with earthy tones and soft pastels.

# **Session 5: Mining Compressed Patterns**

# Mining Compressed Patterns

Pat-ID	Item-Sets	Support
P1	{38,16,18,12}	205227
P2	{38,16,18,12,17}	205211
P3	{39,38,16,18,12,17}	101758
P4	{39,16,18,12,17}	161563
P5	{39,16,18,12}	161576

- ❑ Closed patterns
  - ❑ P1, P2, P3, P4, P5
  - ❑ Emphasizes too much on support
  - ❑ There is no compression
- ❑ Max-patterns
  - ❑ P3: information loss
- ❑ Desired output (a good balance):
  - ❑ P2, P3, P4

❑ Why mining compressed patterns?

- ❑ Too many scattered patterns but not so meaningful

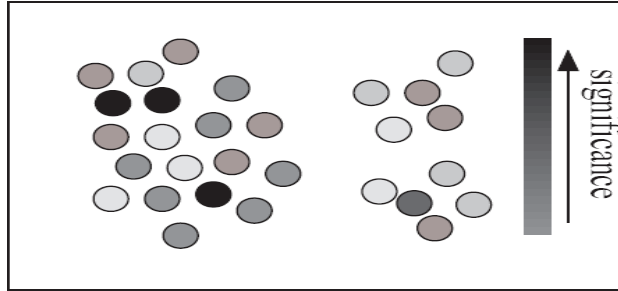
❑ Pattern distance measure

$$Dist(P_1, P_2) = 1 - \frac{|T(P_1) \cap T(P_2)|}{|T(P_1) \cup T(P_2)|}$$

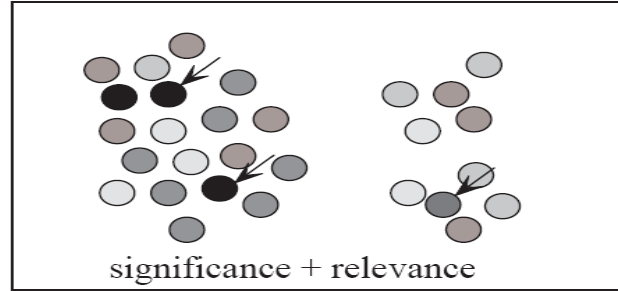
- ❑  $\delta$ -clustering: For each pattern P, find all patterns which can be expressed by P and whose distance to P is within  $\delta$  ( $\delta$ -cover)
- ❑ All patterns in the cluster can be represented by P
- ❑ Method for efficient, direct mining of compressed frequent patterns (e.g., Xin et al., VLDB'05)

# Redundancy-Aware Top-k Patterns

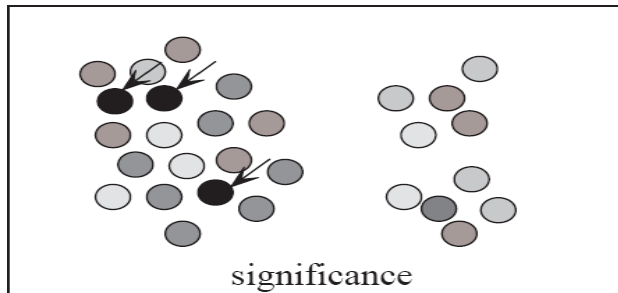
- Desired patterns: high significance & low redundancy



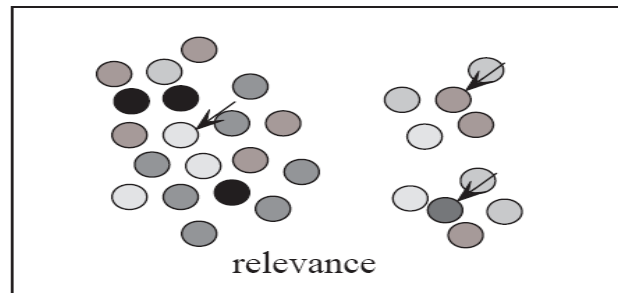
(a) a set of patterns



(b) redundancy-aware top- $k$



(c) traditional top- $k$



(d) summarization

- Method: Use MMS (Maximal Marginal Significance) for measuring the combined significance of a pattern set
- Xin et al., Extracting Redundancy-Aware Top-K Patterns, KDD'06