



The background of the slide is a complex, abstract composition. It features a network of thin, light-colored lines forming a web-like structure. Overlaid on this are various data points and clusters. In the upper left, there's a grid of small, light-colored plus signs. In the lower left, there's a cluster of orange and red dots, some of which are larger and more prominent. The overall color palette is muted, with shades of brown, beige, and light blue. The text is centered in a large, bold, black font.

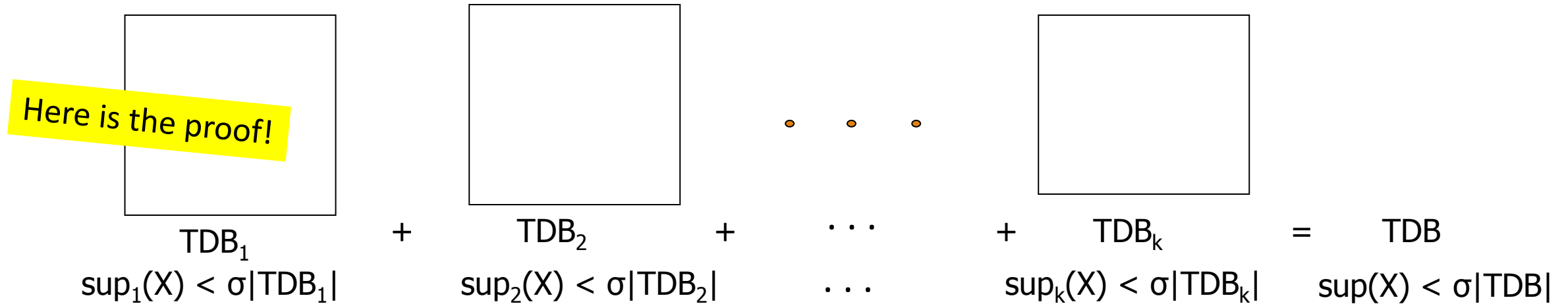
Session 3: Extensions or Improvements of Apriori

Apriori: Improvements and Alternatives

- ❑ Reduce passes of transaction database scans
 - ❑ Partitioning (e.g., Savasere, et al., 1995)  To be discussed in subsequent slides
 - ❑ Dynamic itemset counting (Brin, et al., 1997)
- ❑ Shrink the number of candidates
 - ❑ Hashing (e.g., DHP: Park, et al., 1995)  To be discussed in subsequent slides
 - ❑ Pruning by support lower bounding (e.g., Bayardo 1998)
 - ❑ Sampling (e.g., Toivonen, 1996)
- ❑ Exploring special data structures
 - ❑ Tree projection (Aggarwal, et al., 2001)
 - ❑ H-miner (Pei, et al., 2001)
 - ❑ Hypercube decomposition (e.g., LCM: Uno, et al., 2004)

Partitioning: Scan Database Only Twice

- Theorem: *Any itemset that is potentially frequent in TDB must be frequent in at least one of the partitions of TDB*



- Method: (A. Savasere, E. Omiecinski and S. Navathe, *VLDB'95*)
 - Scan 1: Partition database (how?) and find local frequent patterns
 - Scan 2: Consolidate global frequent patterns (how to?)
- Why does this method guarantee to scan TDB only twice?

Direct Hashing and Pruning (DHP)

- ❑ DHP (Direct Hashing and Pruning): Reduce the number of candidates (J. Park, M. Chen, and P. Yu, SIGMOD'95)
- ❑ Observation: A k -itemset whose corresponding hashing bucket count is below the threshold cannot be frequent

- ❑ Candidates: a, b, c, d, e

- ❑ Hash entries

- ❑ {ab, ad, ae}

- ❑ {bd, be, de}

- ❑ ...

Itemsets	Count
{ab, ad, ae}	35
{bd, be, de}	298
.....	...
{yz, qs, wt}	58

Hash Table

- ❑ Frequent 1-itemset: a, b, d, e

- ❑ ab is not a candidate 2-itemset if the sum of count of {ab, ad, ae} is below support threshold