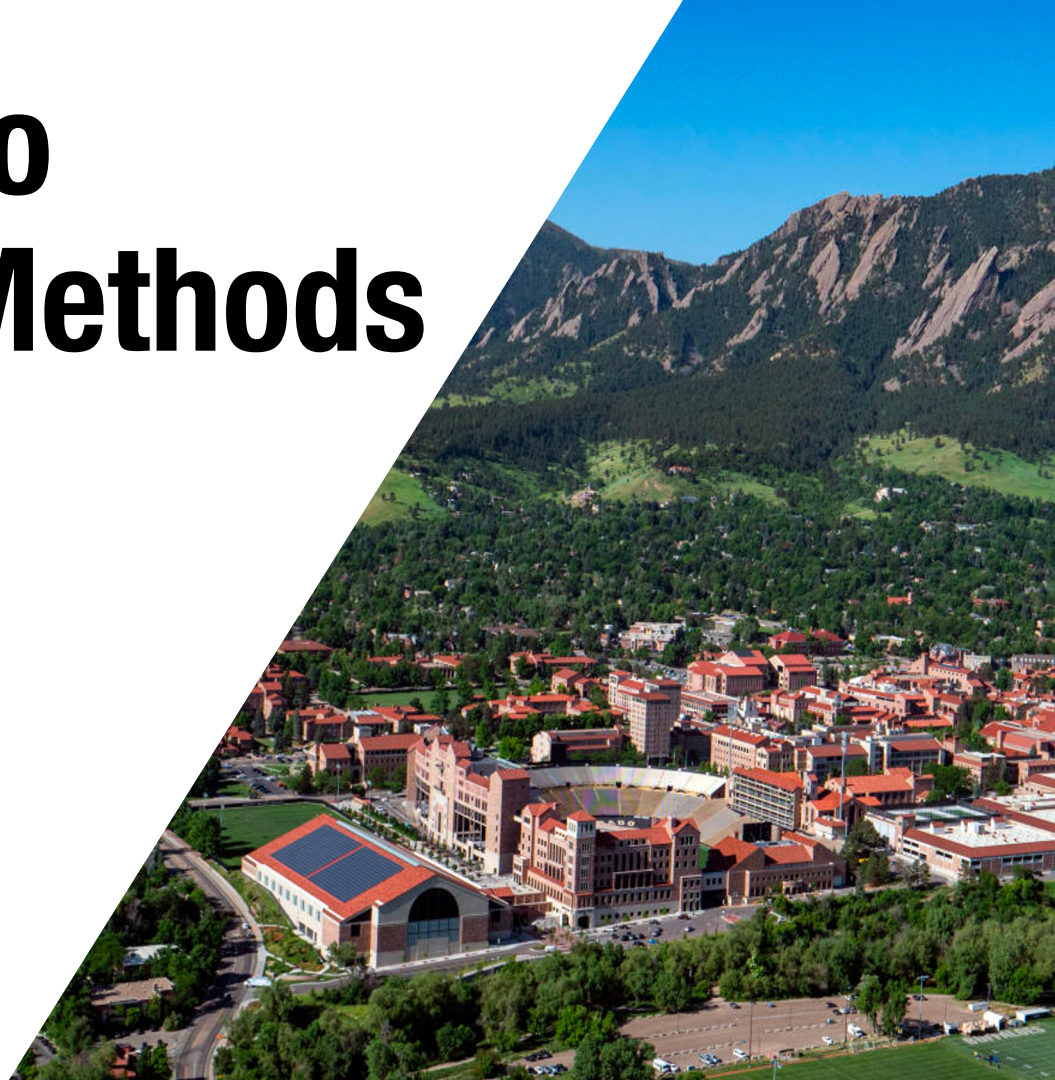
Introduction to Data Mining Methods

Data Mining: Data Mining Methods

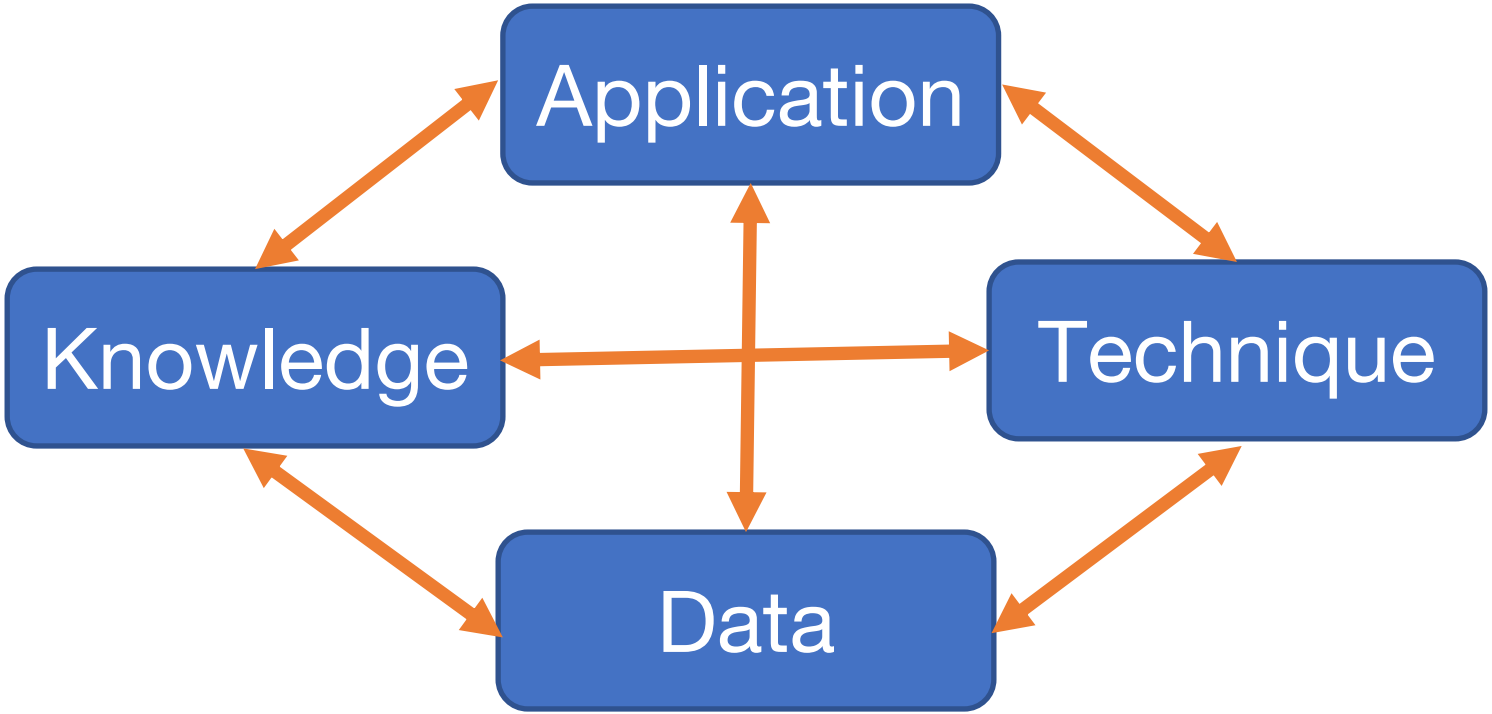
with **Dr. Qin Lv**



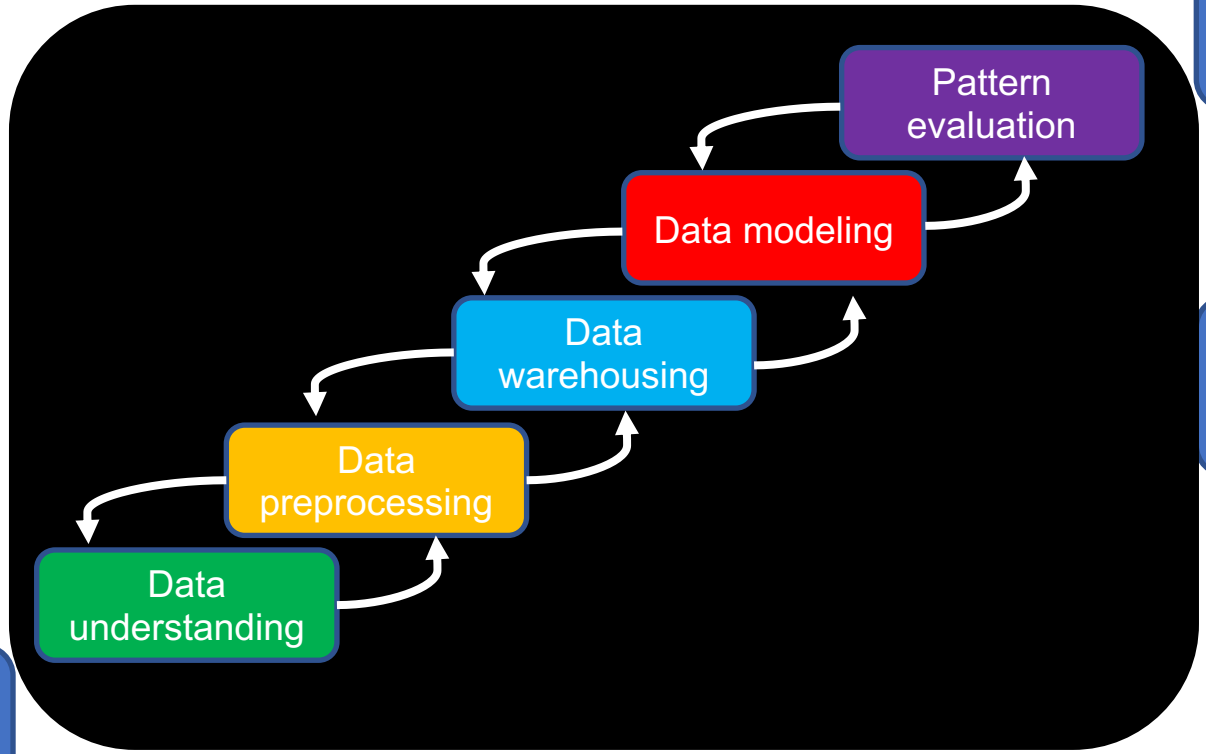
Master of Science in Data Science
UNIVERSITY OF COLORADO **BOULDER**

Learning objective: Identify the core functionalities of data modeling in the data mining pipeline. Apply the Apriori algorithm for frequent itemset mining.

Data Mining: Four Views



Data Mining Pipeline



Application

Knowledge

Technique

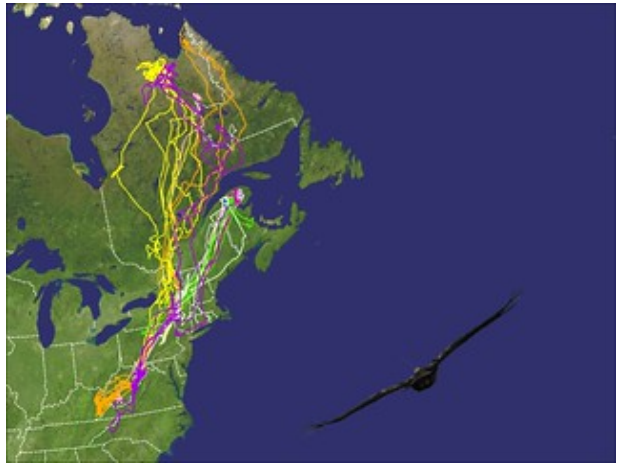
Data

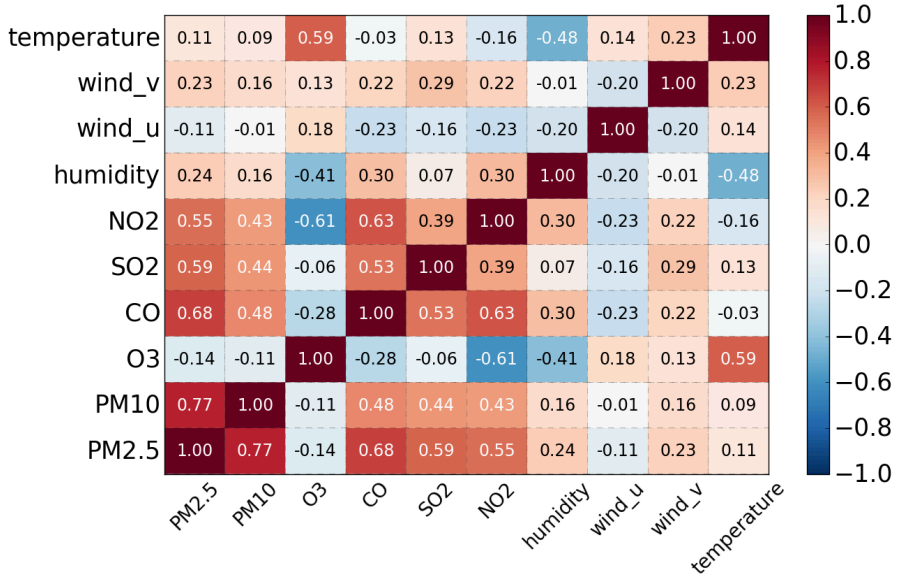
Technique View

- Frequent pattern analysis
- Classification, prediction
- Clustering
- Anomaly detection
- Trend and evolution analysis

Frequent Pattern Analysis

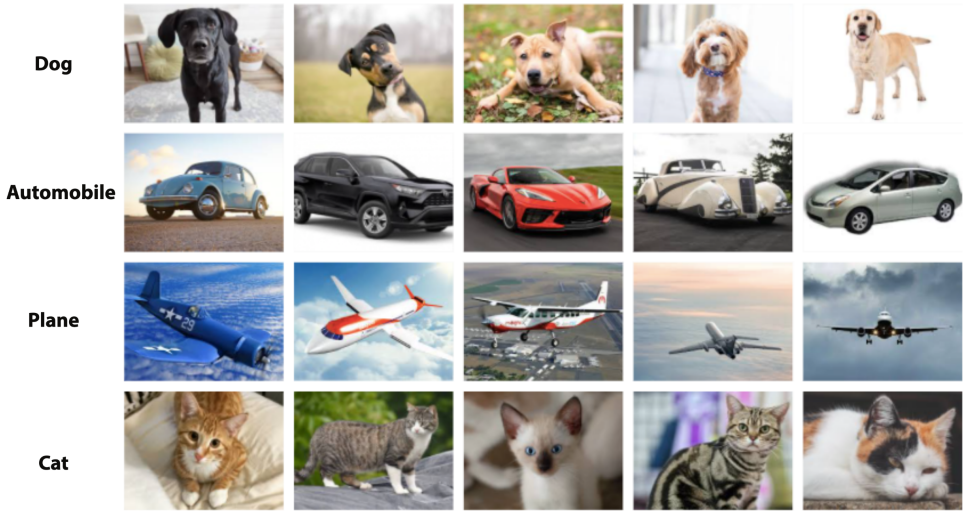
- Frequent itemset
- Frequent sequence
- Frequent structure
- Association rules
- Correlation analysis





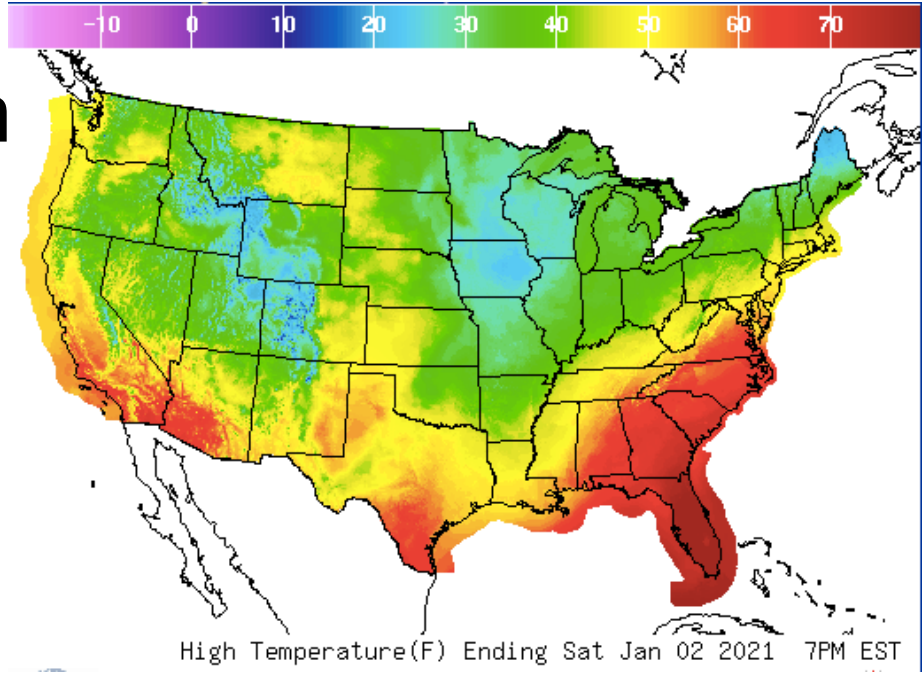
Classification

- Pre-defined classes
- Need training data
- Build model to distinguish classes



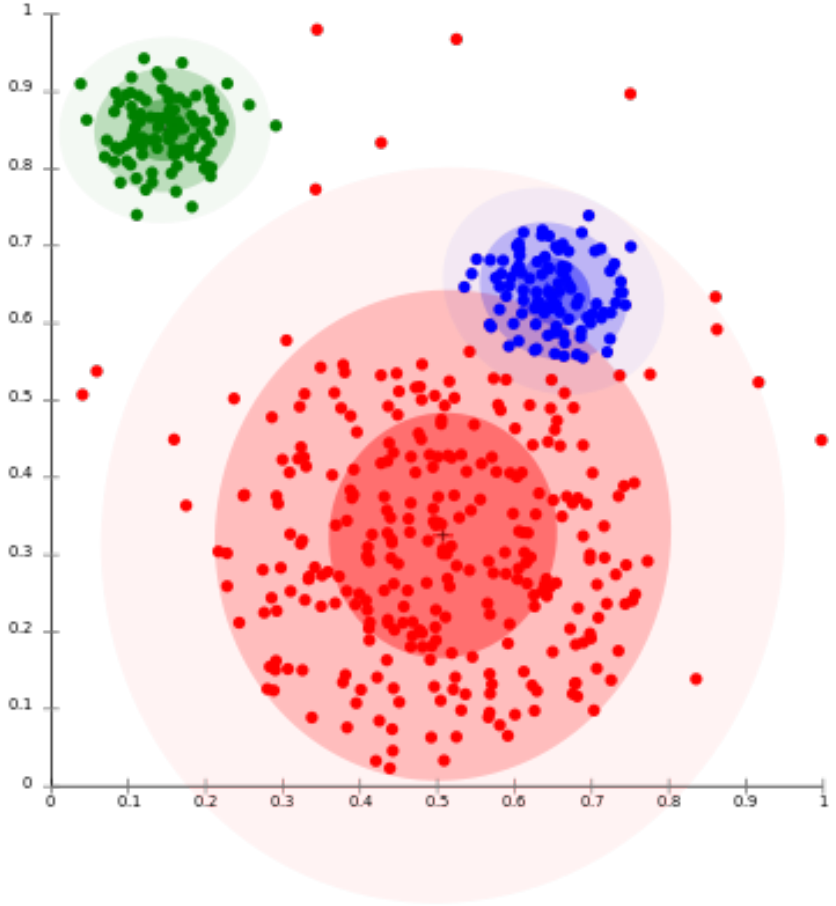
Prediction

- Numerical prediction (continuous value)
  - E.g., weather
  - E.g., stock price
  - E.g., traffic



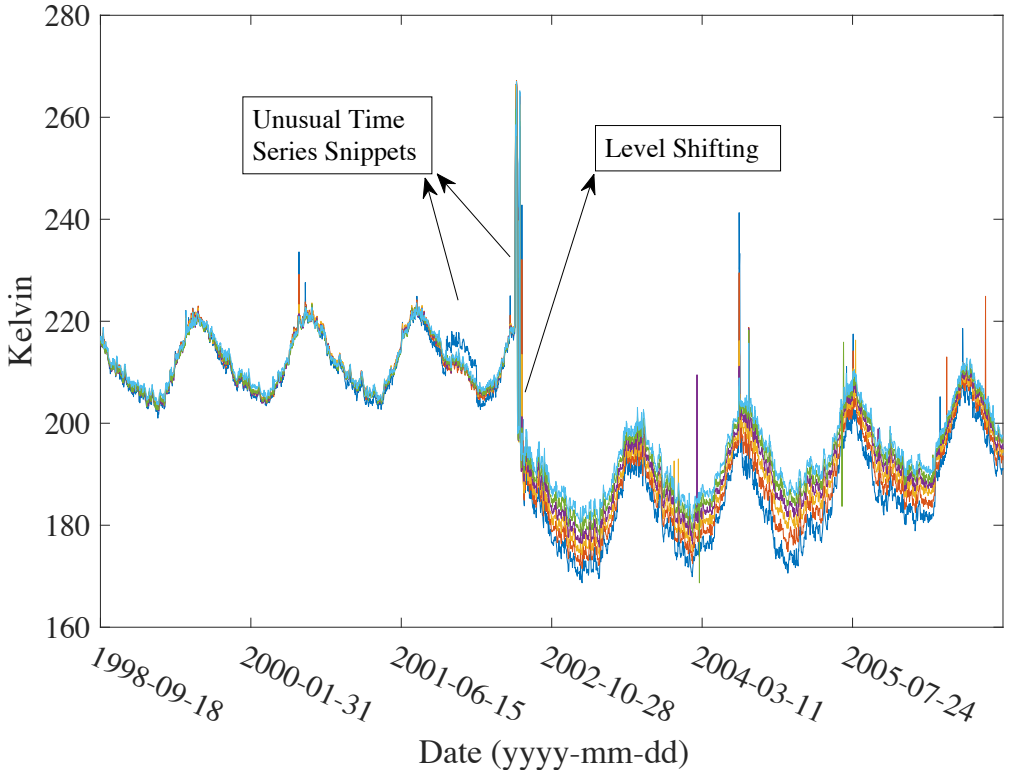
Clustering

- No predefined classes
- Intra-cluster similarity
- Inter-cluster dissimilarity



Anomaly Detection

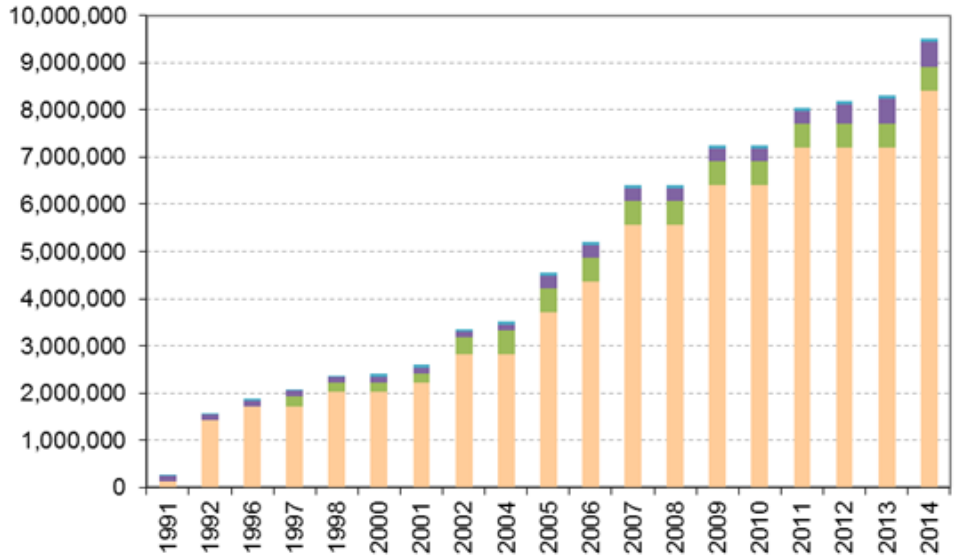
- Anomaly/outlier
  - Differ from the “norm”
  - E.g., error, noise
  - E.g., fraud
  - E.g., extreme events



Trend and Evolution Analysis

- Changes over time
  - Overall trend
  - Periodical patterns
  - Anomalies
  - E.g.,

Google Trends



Data Mining Methods

- Frequent pattern analysis
- Classification
- Clustering
- Outlier analysis

Market Basket Analysis

- List of transactions
  - Each T_i contains multiple items
- (Frequent) itemset
  - $X = \{x_1, x_2, \ldots, x_k\}$
- (Minimum) support
  - Probability of T_i containing X

Tid	Items
1	A, B, C, E
2	A, D, E
3	B, C, E
4	B, C, D, E
5	B, D, E

Frequent Pattern Mining

- Brute force approach (e.g., 100 items)

$$\binom{100}{1} + \binom{100}{2} + \cdots + \binom{100}{100} = 2^{100} - 1 \approx 1.27 \times 10^{30}$$

- Closed pattern X: no super-pattern Y ⊃ X w/ the same support
- Max-pattern X: no super-pattern Y ⊃ X

Closed & Max Pattern Example

- {<a_1, ..., a_{100}>, <a_1, ..., a_{50}>} min_sup = 0.5
- Frequent pattern? all item combinations
- Closed pattern?
  - <a_1, ..., a_{100}>: 1; <a_1, ..., a_{50}>: 2
- Max-pattern?
  - <a_1, ..., a_{100}>: 1

Apriori Algorithm

- Apriori pruning: if X is infrequent, then any of its superset cannot be frequent
- Procedure

- Scan dataset to get freq. 1-itemsets
- Generate candidate (k+1)-itemsets from freq. k-itemsets
- Scan dataset to remove infreq. candidate (k+1)-itemsets
- Stop when no more freq. or candidate itemsets

Apriori Algorithm Example

- min_sup = 0.6

Tid	Items
1	A, B, C, E
2	A, D, E
3	B, C, E
4	B, C, D, E
5	B, D, E

What about {B, D, E} or {C, D, E}?

Itemset	#
{A}	2
{B}	4
{C}	3
{D}	3
{E}	5

Itemset	#
{B, C}	3
{B, D}	2
{B, E}	4
{C, D}	1
{C, E}	3
{D, E}	3

Itemset	#
{B, C, E}	3

Important Details

- ➢ Self-joining of k-itemsets => (k+1)-itemsets
  - Only join if their first (k-1) items are the same
- ➢ Pruning: remove if subset is not frequent
- ➢ Example: L3 = {abc, abd, acd, ace, bcd}
  - abc and abd => abcd and bcd is in L3 => valid candidate
  - acd and ace => acde but ade is not in L3 => pruned