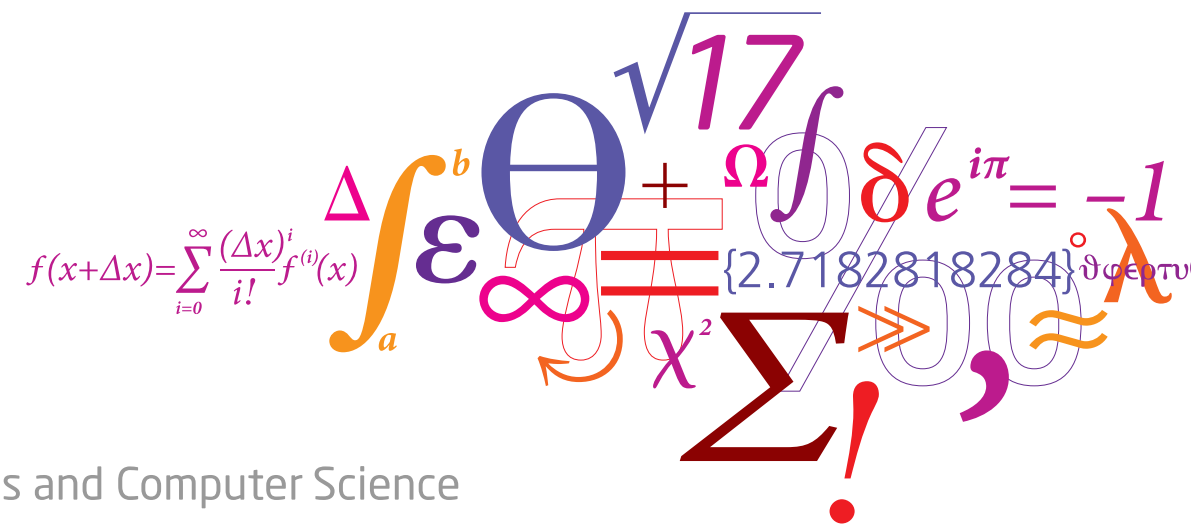
# 02450: Introduction to Machine Learning and Data Mining

Association mining

Mikkel N. Schmidt

DTU Compute, Technical University of Denmark (DTU)




DTU Compute
Department of Applied Mathematics and Computer Science

# Lecture Schedule

1. Introduction
   7 October: C1

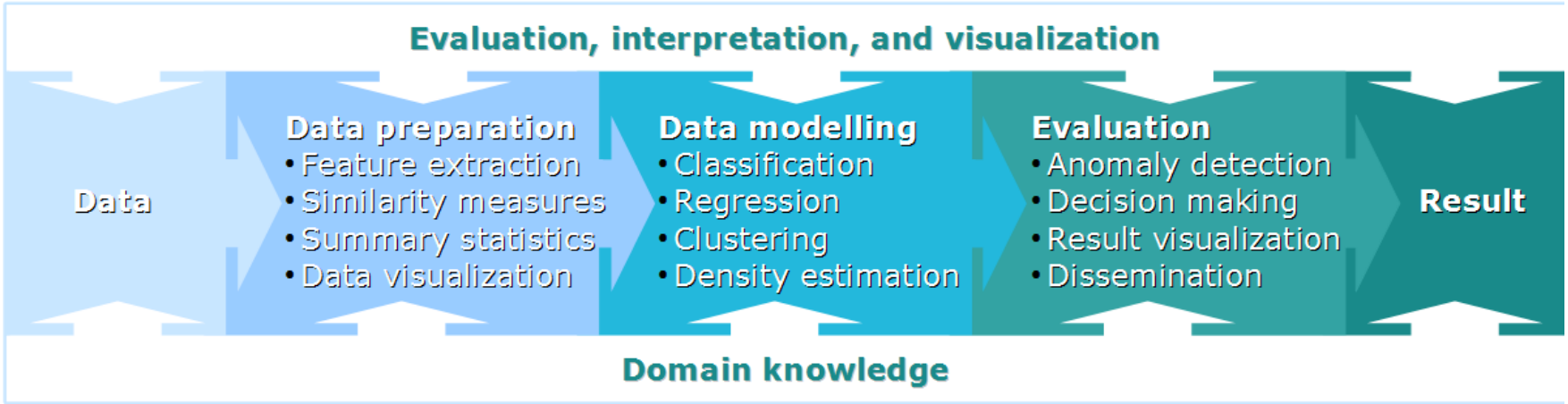
   Data: Feature extraction, and visualization
2. Data, feature extraction and PCA
   7 October: C2, C3
3. Measures of similarity, summary statistics and probabilities
   7 October: C4, C5
4. Probability densities and data Visualization
   7 October: C6, C7

   Supervised learning: Classification and regression
5. Decision trees and linear regression
   8 October: C8, C9
6. Overfitting, cross-validation and Nearest Neighbor
   8 October: C10, C12
7. Performance evaluation, Bayes, and Naive Bayes
   9 October: C11, C13
8. Artificial Neural Networks and Bias/Variance
   9 October: C14, C15
9. AUC and ensemble methods
   10 October: C16, C17

   Unsupervised learning: Clustering and density estimation
10. K-means and hierarchical clustering
    10 October: C18
11. Mixture models and density estimation
    11 October: C19, C20
12. **Association mining**
    **11 October: C21**

    Recap
13. Recap
    11 October: C1-C21

Piazza online help: `https://piazza.com/dtu.dk/fall2019/october2019`



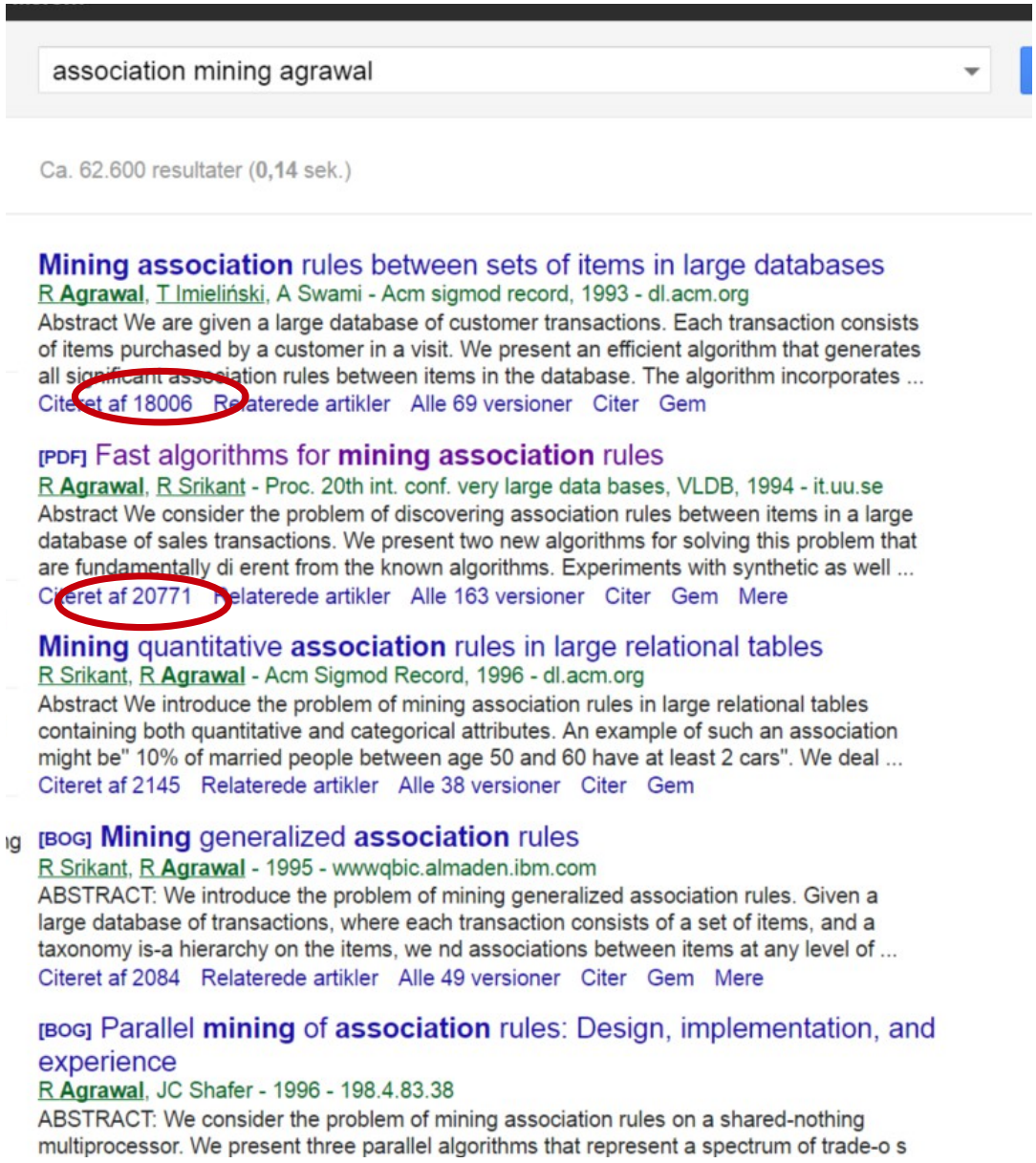

## Learning Objectives

- Calculate support and confidence of association rules
- Describe the Apriori algorithm for association mining and how it is used for efficient estimation of association rules

# Association rule discovery: Definition

- Given a set of **records**
  - Each containing a number of **items from a set**
- **Goal:** Produce dependency rules
  - Predict the occurence of an item based on occurences of other items

# Association Mining



# Association rule discovery: Definition

- Given a set of **records**
  - Each containing a number of **items from a set**
- **Goal:** Produce dependency rules
  - Predict the occurence of an item based on occurences of other items

# Association rule discovery: Example

## Market basket analysis

**Training set**

1. {Bread, Soda, Milk}
2. {Beer, Bread}
3. {Beer, Soda, Diaper, Milk}
4. {Beer, Bread, Diaper, Milk}
5. {Soda, Diaper, Milk}

**Rules discovered**

{Milk} ▸ {Soda}
{Diaper, Milk} ▸ {Beer}

# Market basket data

- Representation as

**Transaction table**

| ID | Items |
|---|---|
| 1 | Bread, Soda, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Soda, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Soda, Diaper, Milk |

**Data matrix**

| ID | Bread | Soda | Milk | Beer | Diaper |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 | 1 | 0 |
| 3 | 0 | 1 | 1 | 1 | 1 |
| 4 | 1 | 0 | 1 | 1 | 1 |
| 5 | 0 | 1 | 1 | 0 | 1 |

# Association analysis, rules and support

- **Itemset**
  - For example {Bread, Soda, Milk}, {Milk, Diaper}, {}
- **Support** for an itemset **X**
  - Percentage of transactions that contain **X**
- **Association rule**
  - Expression of the form: **X** ▸ **Y**
    where **X** and **Y** are disjoint item sets
- **Support** for an association rule **X** ▸ **Y**
  - Percentage of transactions that contain **X** ∪ **Y**

$$s(X \to Y) = \frac{\sigma(X \cup Y)}{N} = P(X, Y)$$

# Piazza quiz 1: Support (Spring 2018)

| | $x_1^L$ | $x_1^H$ | $x_2^L$ | $x_2^H$ | $x_3^L$ | $x_3^H$ | $x_4^L$ | $x_4^H$ | $x_5^L$ | $x_5^H$ | $x_6^L$ | $x_6^H$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| O1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| O2 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| O3 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| O4 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| O5 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| O6 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| O7 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| O8 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| O9 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| O10 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |

Table 1: The ten first observations of the airline safety dataset binarized considering the attribute $x_1$–$x_6$.

We consider a dataset of airline safety binarized according to the median value. Values below median is referred to with the superscript $L$ and above the median value using the superscript $H$. In Table 1 is given the first 10 observations O1–O10. Consider the association rule:

$$\{x_2^H, x_3^H, x_4^H, x_5^H\} \rightarrow \{x_6^H\}.$$

What is the support of the rule?

A. 0.0 %

B. 20.0 %

C. 66.7 %

D. 100.0 %

E. Don't know.

# Solution:

| | $x_1^L$ | $x_1^H$ | $x_2^L$ | $x_2^H$ | $x_3^L$ | $x_3^H$ | $x_4^L$ | $x_4^H$ | $x_5^L$ | $x_5^H$ | $x_6^L$ | $x_6^H$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| O1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| O2 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| O3 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| O4 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| O5 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| O6 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| O7 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| O8 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| O9 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| O10 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |

Table 1: The ten first observations of the airline safety dataset binarized considering the attribute $x_1$–$x_6$.

The support of $\{x_2^H, x_3^H, x_4^H, x_5^H\} \rightarrow \{x_6^H\}$ is given by the number of times out of the total number of customers that customers have relatively high values of both $x_2$, $x_3$, $x_4$, $x_5$, and $x_6$, i.e., given by the support of the itemset $\{x_2^H, x_3^H, x_4^H, x_5^H, x_6^H\}$. Only customer O2 and O6 have this property out of the 10 customers, thus the support is 2/10.

# Association analysis, confidence

- **Itemset**
  - For example {Bread, Soda, Milk}, {Milk, Diaper}, {}
- **Support** for an itemset **X**
  - Percentage of transactions that contain **X**
- **Association rule**
  - Expression of the form: **X** ▸ **Y**
    where **X** and **Y** are disjoint item sets
- **Support** for an association rule **X** ▸ **Y**
  - Percentage of transactions that contain **X** ∪ **Y**

$$s(X \to Y) = \frac{\sigma(X \cup Y)}{N} = P(X, Y)$$

- **Confidence** for an association rule **X** ▸ **Y**
  - Percentage of transactions containing **X** that also contain **Y**

$$c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{P(Y, X)}{P(X)} = P(Y|X)$$

# Piazza quiz 2: Confidence (Spring 2018)

| | $x_1^L$ | $x_1^H$ | $x_2^L$ | $x_2^H$ | $x_3^L$ | $x_3^H$ | $x_4^L$ | $x_4^H$ | $x_5^L$ | $x_5^H$ | $x_6^L$ | $x_6^H$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| O1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| O2 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| O3 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| O4 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| O5 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| O6 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| O7 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| O8 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| O9 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| O10 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |

Table 1: The ten first observations of the airline safety dataset binarized considering the attribute $x_1$–$x_6$.

We again consider the airline safety data and the rule

$$\{x_2^H, x_3^H, x_4^H, x_5^H\} \rightarrow \{x_6^H\}.$$

What is the confidence of the rule?

A. 0.0 %

B. 20.0 %

C. 66.7 %

D. 100.0 %

E. Don't know.

# Solution:

| | $x_1^L$ | $x_1^H$ | $x_2^L$ | $x_2^H$ | $x_3^L$ | $x_3^H$ | $x_4^L$ | $x_4^H$ | $x_5^L$ | $x_5^H$ | $x_6^L$ | $x_6^H$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| O1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| O2 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| O3 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| O4 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| O5 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| O6 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| O7 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| O8 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| O9 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| O10 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |

Table 1: The ten first observations of the airline safety dataset binarized considering the attribute $x_1$–$x_6$.

The confidence is given as

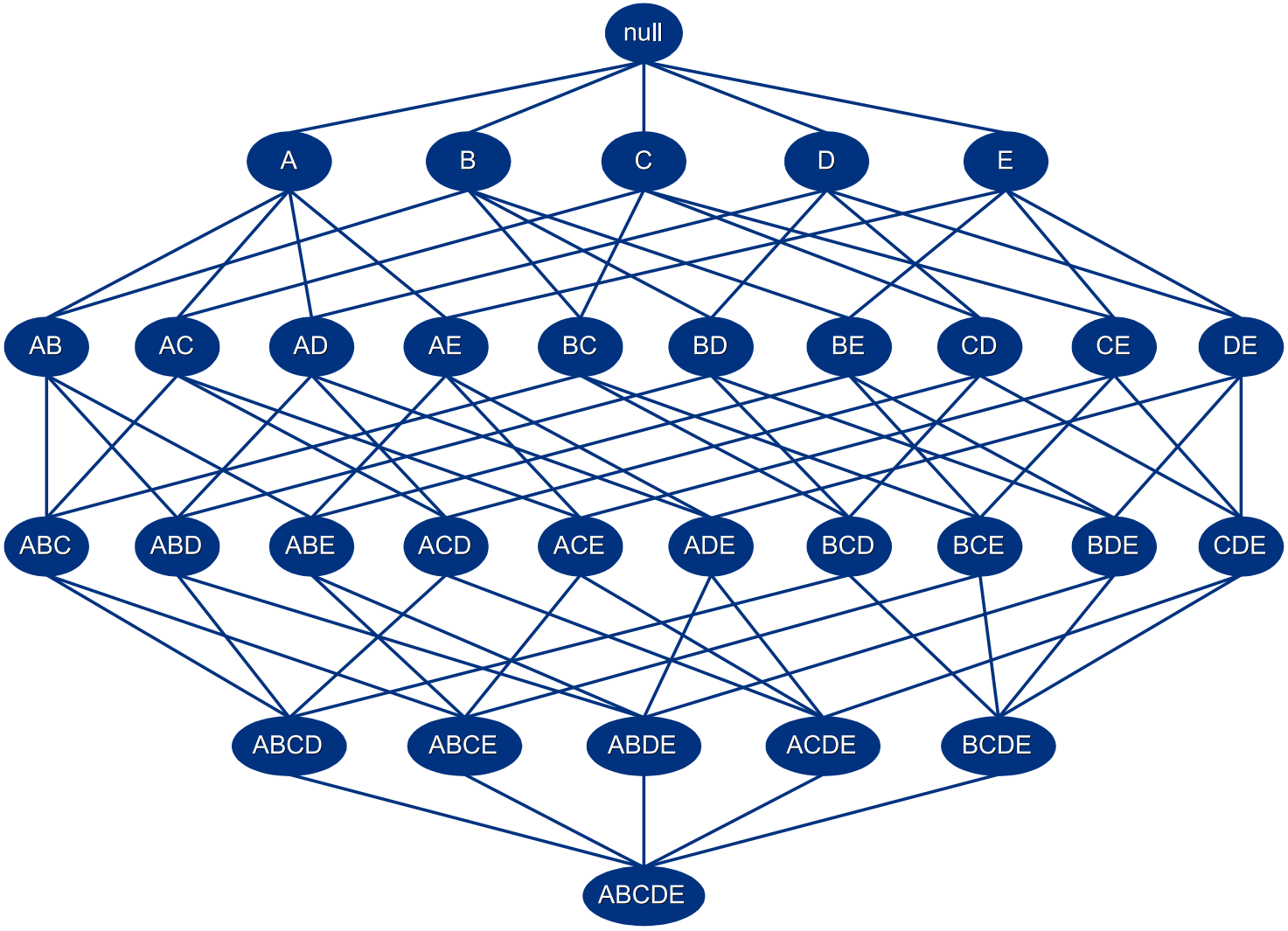
$$c(\{x_2^H, x_3^H, x_4^H, x_5^H\} \to \{x_6^H\})) = \frac{s\left(\{x_2^H, x_3^H, x_4^H, x_5^H, x_6^H\}\right)}{s\left(\{x_2^H, x_3^H, x_4^H, x_5^H\}\right)} = \frac{2/10}{3/10} = 2/3 = 66.7\%$$

# Association rule mining

- Find all association rules that have
  - **Support** $\geq$ *minsup*
  - **Confidence** $\geq$ *minconf*

- Approach
  - **Frequent itemset generation**
    - Generate a list of all **itemsets** with **Support** $\geq$ *minsup*
  - **Association rule generation**
    - Generate all **association rules** with **Confidence** $\geq$ *minconf*

# Frequent itemset generation

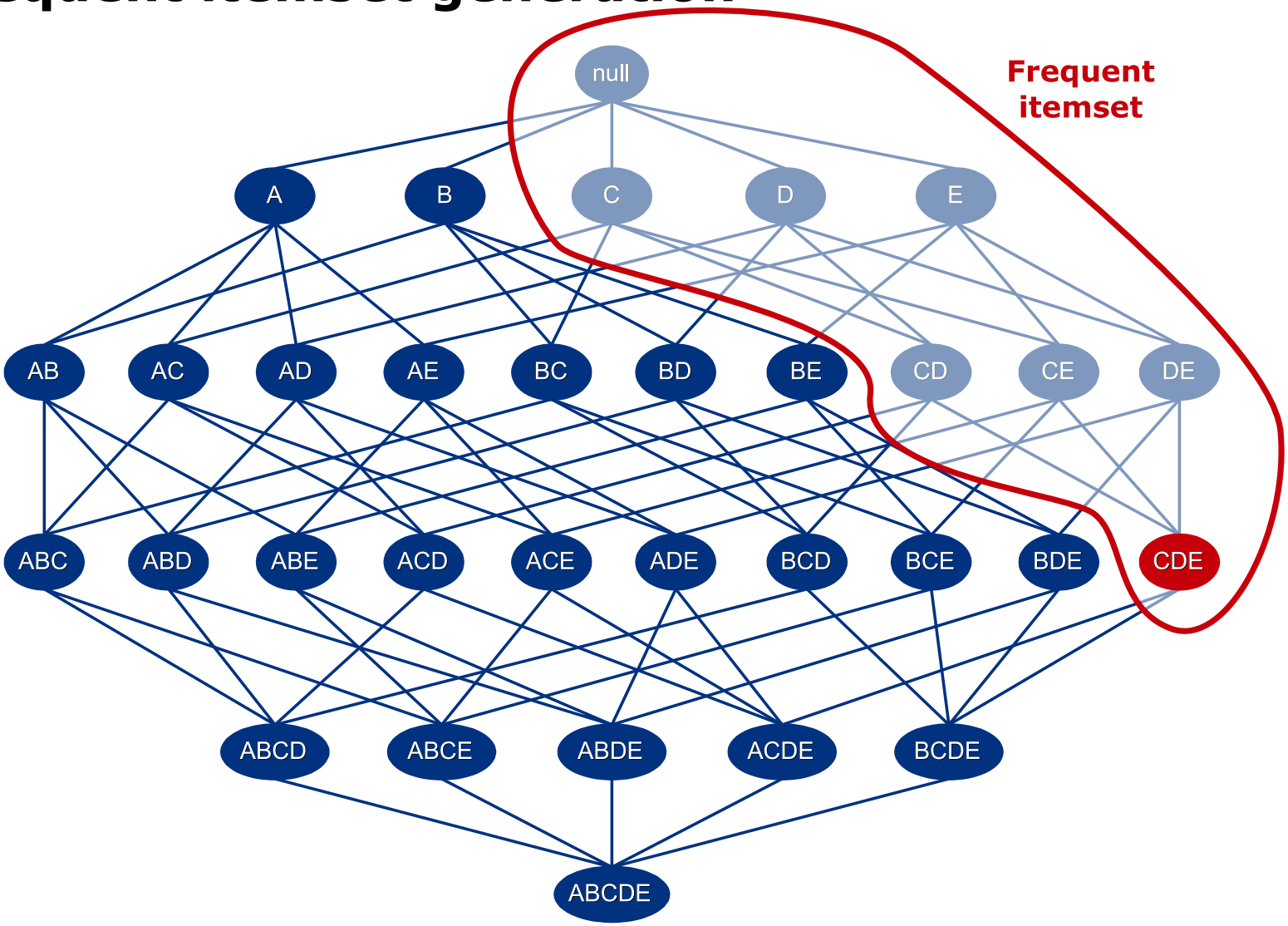



How many different itemsets can be created for a problem with a total of D items?



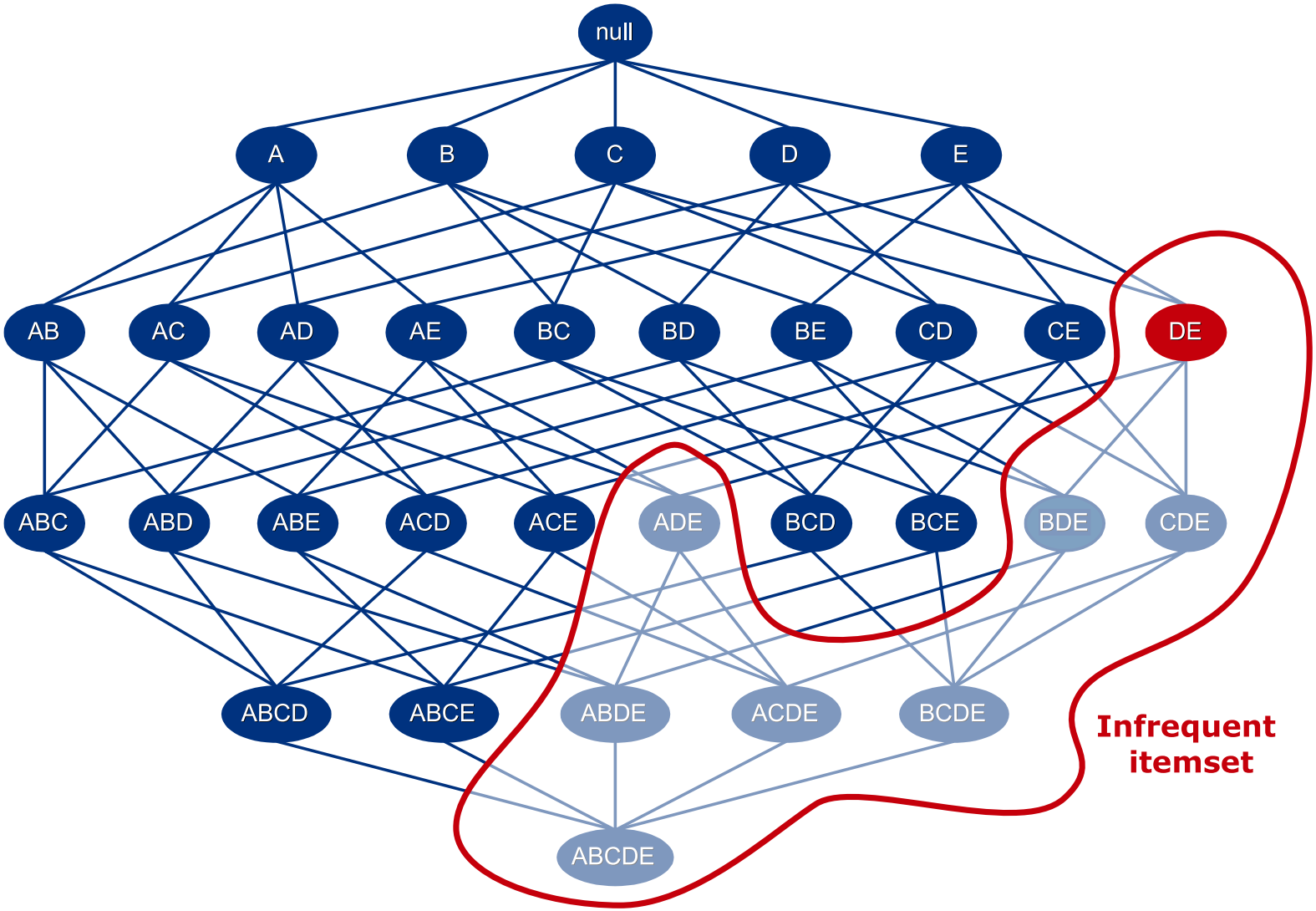
This itemset latice representation is adapted from:
"Introduction to Data Mining" by Pang-Ning Tan, Michael Steinbach, Vipin Kumar

# Frequent itemset generation




If an itemset is frequent, then all of its subsets must also be frequent

# Frequent itemset generation




If an itemset is infrequent, then all of its supersets must also be infrequent

# The Apriori Algorithm

Find all 1-itemsets

Generate k-itemsets by merging single items to the k-1-itemsets

Remove all the generated itemsets for which subsets are not part of the k-1-itemsets

Keep remaining k-itemsets with enough support.

Output all frequent itemsets

**Algorithm 8:** Apriori algorithm

1: Given $N$ transactions and let $\epsilon > 0$ be the minimum support count
2: $L_1 = \{\{j\} | \text{supp}(\{j\}) \geq \epsilon\}$
3: **for** $k = 2, \ldots, M$ and $L_k \neq \emptyset$ **do**
4: $\quad C'_k = \{s \cup \{j\} | s \in L_{k-1}, j \notin s\}$
5: $\quad$ Set $C_k = C'_k$
6: $\quad$ **for** each $c \in C'_k$ **do**
7: $\quad\quad$ **for** each $s \subset c$ such that $|s| = k - 1$ **do**
8: $\quad\quad\quad$ **if** $s$ is not frequent, i.e. $s \notin L_{k-1}$ **then**
9: $\quad\quad\quad\quad C_k = C_k \setminus \{c\}$ (Remove $c$ from $C_k$)
10: $\quad\quad\quad$ **end if**
11: $\quad\quad$ **end for**
12: $\quad$ **end for**
13: $\quad L_k = \{c | c \in C_k, \text{supp}(c) \geq \epsilon\}$ (compute support)
14: **end for**
15: $L_1 \cup L_2 \cup \cdots \cup L_k$ are then all frequent itemsets

# Piazza quiz 3: A-priori (Fall 2018)

We will consider a binary dataset consisting of the $M = 6$ features $f_1$, $f_2$, $f_3$, $f_4$, $f_5$, $f_6$. We wish to apply the Apriori algorithm to find all itemsets with support greater than $\varepsilon = 0.15$. Suppose at iteration $k = 3$ we know that:

$$L_2 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Recall the key step in the Apriori algorithm is to construct $L_3$ by first considering a large number of candidate itemsets $C_3'$, and then rule out some of them using the downwards-closure principle thereby saving many (potentially costly) evaluations of support. Suppose $L_2$ is given as above, which of the following itemsets does the Apriori algorithm *not* have to evaluate the support of?

A. $\{f_2, f_3, f_4\}$

B. $\{f_1, f_2, f_6\}$

C. $\{f_2, f_3, f_6\}$

D. $\{f_1, f_3, f_4\}$

E. Don't know.

# Solution:

Recall the Apriori algorithm obtain $L_3$ from $L_2$ in three steps. First, the Apriori algorithm construct $C_3'$ by, for each itemset $I$ in $L_2$, loop over all items not already in $I$ and consider all such combinations where $I$ is enlarged by a single item as a candidate itemset in $C_3'$. Specifically we get:

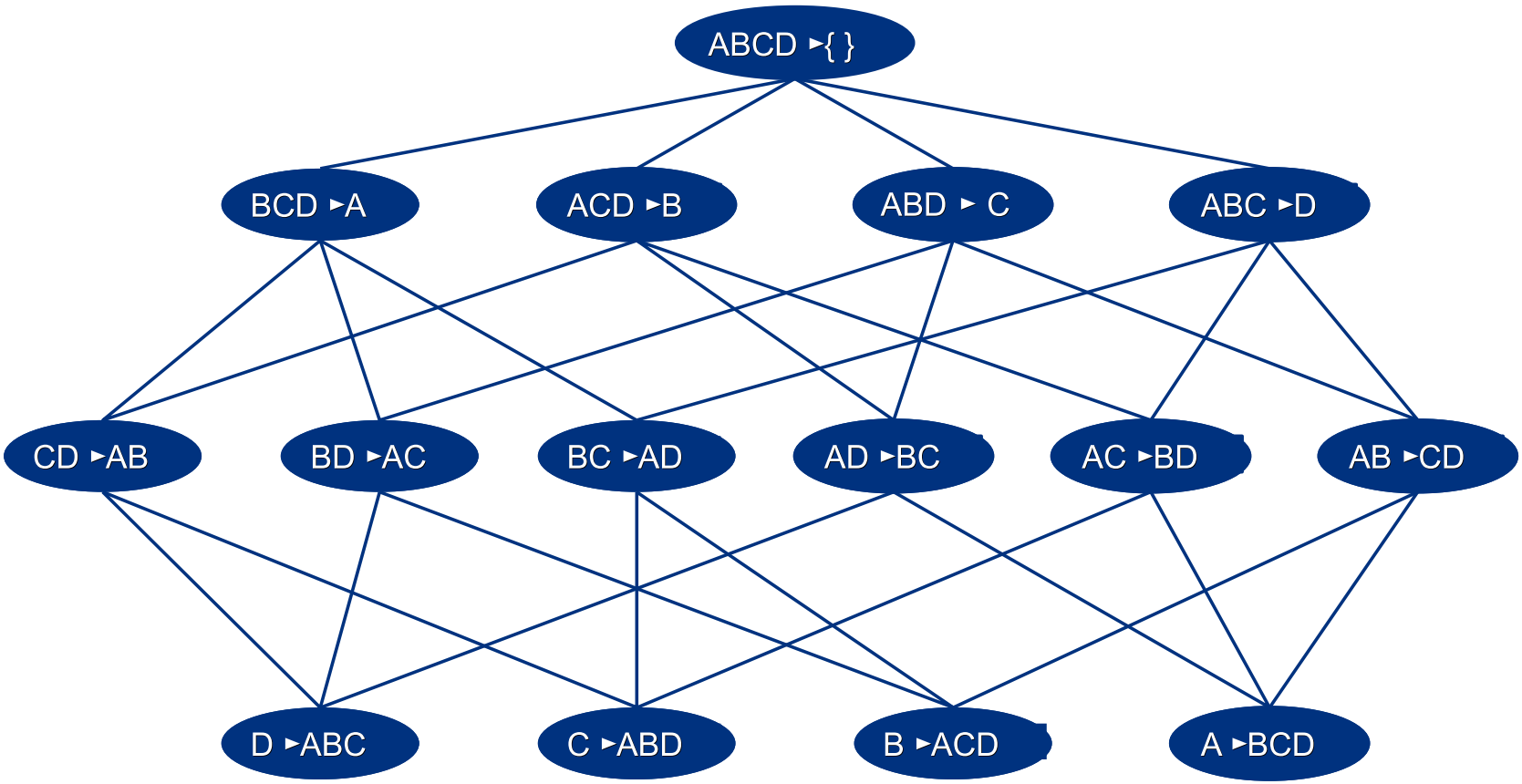
$$C_3' = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}.$$

The downwards closure principle is then applied by removing and itemset $I$ in $C_3'$ if $I$ contains a subset of 2 items not found in $L_2$. We thereby get:

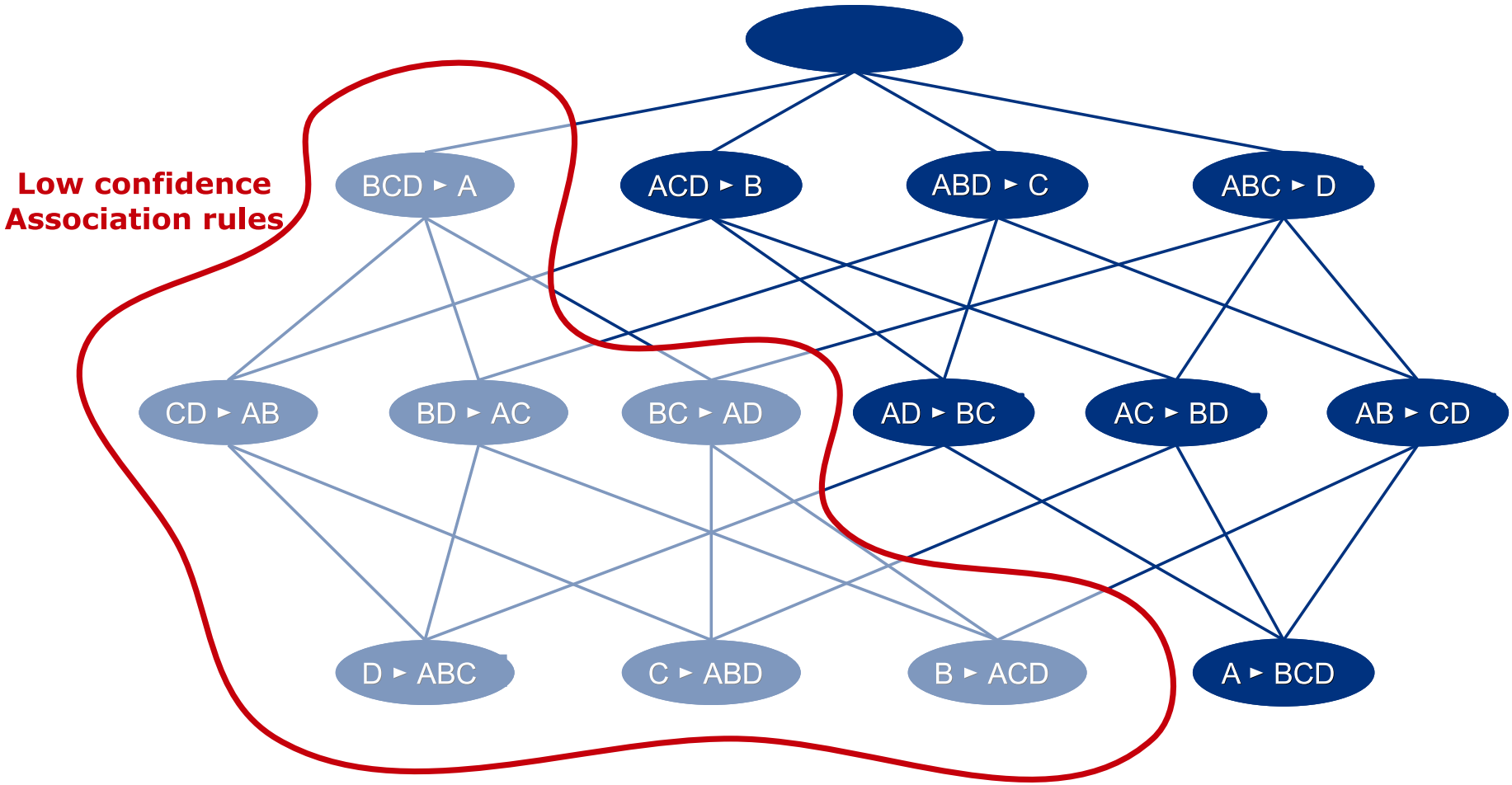
$$C_3 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

Finally, $L_3$ is constructed from $C_3$ by removing those itemsets with a support lower than $\varepsilon$. Thus, the itemsets we don't have to compute support from are those itemsets found in $C_3'$ but not in $C_3$, or as an even simpler criteria, those which have a subset of size 2 not found in $L_2$. This rules out all options except D.

# Association rule generation



# Association rule generation



# Results for market basket example

| Itemset | Support |
|---|---|
| Milk | 80% |
| Bread | 60% |
| Soda | 60% |
| Beer | 60% |
| Diaper | 60% |
| Diaper Milk | 60% |
| Soda Milk | 60% |
| Bread Beer | 40% |
| Bread Milk | 40% |
| Soda Diaper | 40% |
| Beer Diaper | 40% |
| Beer Milk | 40% |
| Soda Diaper Milk | 40% |
| Beer Diaper Milk | 40% |

| Association rule | Support | Confidence |
|---|---|---|
| {} ▸ Milk | 80% | 80% |
| Soda ▸ Milk | 60% | 100% |
| Diaper ▸ Milk | 60% | 100% |
| Soda, Diaper ▸ Milk | 40% | 100% |
| Beer, Diaper ▸ Milk | 40% | 100% |
| Beer, Milk ▸ Diaper | 40% | 100% |

| ID | Bread | Soda | Milk | Beer | Diaper |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 | 1 | 0 |
| 3 | 0 | 1 | 1 | 1 | 1 |
| 4 | 1 | 0 | 1 | 1 | 1 |
| 5 | 0 | 1 | 1 | 0 | 1 |

- How can we do association mining for continuous data?

$\mathbf{X}=$

| Attribute 1 | Attribute 2 | Attribute 3 |
|---|---|---|
| 0.3689 | 0.9827 | 0.6999 |
| 0.4607 | 0.7302 | 0.6385 |
| 0.9816 | 0.3439 | 0.0336 |
| 0.1564 | 0.5841 | 0.0688 |
| 0.8555 | 0.1078 | 0.3196 |
| 0.6448 | 0.9063 | 0.5309 |
| 0.3763 | 0.8797 | 0.6544 |
| 0.1909 | 0.8178 | 0.4076 |
| 0.4283 | 0.2607 | 0.8200 |
| 0.4820 | 0.5944 | 0.7184 |
| 0.1206 | 0.0225 | 0.9686 |
| 0.5895 | 0.4253 | 0.5313 |
| 0.2262 | 0.3127 | 0.3251 |
| 0.3846 | 0.1615 | 0.1056 |
| 0.5830 | 0.1788 | 0.6110 |
| 0.2518 | 0.4229 | 0.7788 |
| 0.2904 | 0.0942 | 0.4235 |
| 0.6171 | 0.5985 | 0.0908 |
| 0.2653 | 0.4709 | 0.2665 |
| 0.8244 | 0.6959 | 0.1537 |

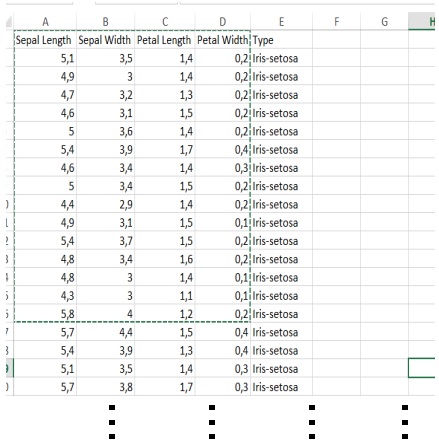
# Binarize data according to percentiles

| AttributeNames= | Attribute 1 | Attribute 2 | Attribute 3 |
|---|---|---|---|
| | 0.3689 | 0.9827 | 0.6999 |
| | 0.4607 | 0.7302 | 0.6385 |
| | 0.9816 | 0.3439 | 0.0336 |
| | 0.1564 | 0.5841 | 0.0688 |
| | 0.8555 | 0.1078 | 0.3196 |
| | 0.6448 | 0.9063 | 0.5309 |
| | 0.3763 | 0.8797 | 0.6544 |
| | 0.1909 | 0.8178 | 0.4076 |
| | 0.4283 | 0.2607 | 0.8200 |
| **X**= | 0.4820 | 0.5944 | 0.7184 |
| | 0.1206 | 0.0225 | 0.9686 |
| | 0.5895 | 0.4253 | 0.5313 |
| | 0.2262 | 0.3127 | 0.3251 |
| | 0.3846 | 0.1615 | 0.1056 |
| | 0.5830 | 0.1788 | 0.6110 |
| | 0.2518 | 0.4229 | 0.7788 |
| | 0.2904 | 0.0942 | 0.4235 |
| | 0.6171 | 0.5985 | 0.0908 |
| | 0.2653 | 0.4709 | 0.2665 |
| | 0.8244 | 0.6959 | 0.1537 |

| AttributeNamesBin= | Attribute 1 0-50 % | Attribute 1 50-100 % | Attribute 2 0-33.3 % | Attribute 2 33.3-66.7% | Attribute 2 66.7-100 % | Attribute 3 0-50% | Attribute 3 50-100% |
|---|---|---|---|---|---|---|---|
| | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| **Xbinary**= | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| | 0 | 1 | 0 | 0 | 1 | 1 | 0 |

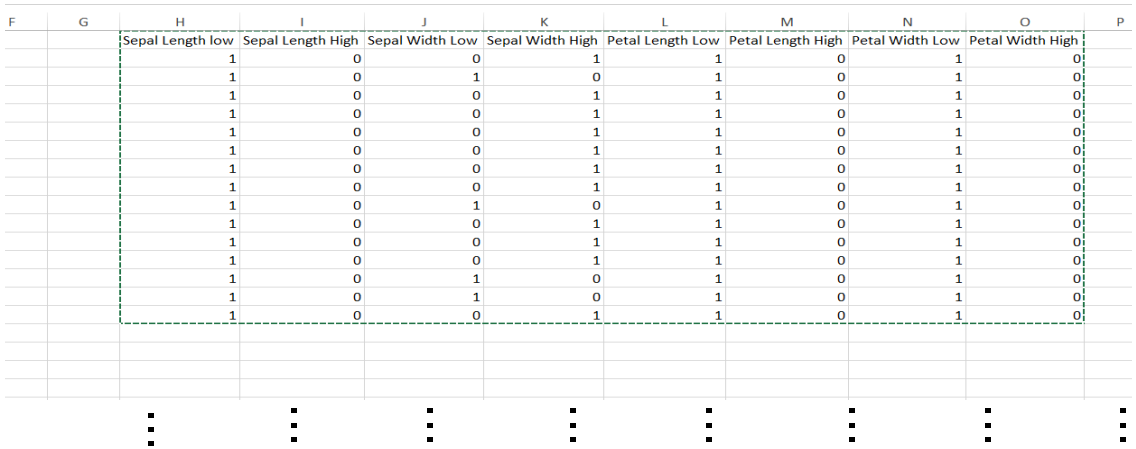
# Recap of association rule discovery on Iris data

**X=**



| Sepal Length | Sepal Width | Petal Length | Petal Width | Type |
|---|---|---|---|---|
| 5,1 | 3,5 | 1,4 | 0,2 | Iris-setosa |
| 4,9 | 3 | 1,4 | 0,2 | Iris-setosa |
| 4,7 | 3,2 | 1,3 | 0,2 | Iris-setosa |
| 4,6 | 3,1 | 1,5 | 0,2 | Iris-setosa |
| 5 | 3,6 | 1,4 | 0,2 | Iris-setosa |
| 5,4 | 3,9 | 1,7 | 0,4 | Iris-setosa |
| 4,6 | 3,4 | 1,4 | 0,3 | Iris-setosa |
| 5 | 3,4 | 1,5 | 0,2 | Iris-setosa |
| 4,4 | 2,9 | 1,4 | 0,2 | Iris-setosa |
| 4,9 | 3,1 | 1,5 | 0,1 | Iris-setosa |
| 5,4 | 3,7 | 1,5 | 0,2 | Iris-setosa |
| 4,8 | 3,4 | 1,6 | 0,2 | Iris-setosa |
| 4,8 | 3 | 1,4 | 0,1 | Iris-setosa |
| 4,3 | 3 | 1,1 | 0,1 | Iris-setosa |
| 5,8 | 4 | 1,2 | 0,2 | Iris-setosa |
| 5,7 | 4,4 | 1,5 | 0,4 | Iris-setosa |
| 5,4 | 3,9 | 1,3 | 0,4 | Iris-setosa |
| 5,1 | 3,5 | 1,4 | 0,3 | Iris-setosa |
| 5,7 | 3,8 | 1,7 | 0,3 | Iris-setosa |

**Xbinary=**



| Sepal Length low | Sepal Length High | Sepal Width Low | Sepal Width High | Petal Length Low | Petal Length High | Petal Width Low | Petal Width High |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |

**Example of association rules where y is also appended to Xbinary, i.e. [Xbinary Y], i.e. Y has three columns indicating which of the three flowers the observation belongs to.**

{Petal length Low, Petal Width low, Iris Setosa} -> {Sepal Length Low}
(Conf 100, sup 33)

{Sepal Length Low, Petal Width low, Iris Setosa} -> {Petal length Low}
(Conf 100, sup 33)

{Sepal length Low, Petal length Low, Iris Setosa} -> {Petal Width Low}
(Conf 100, sup 33)

{Sepal length Low, Sepal Width High, Petal length Low, Petal Width Low} -> {Iris Setosa}
(Conf 100, sup 28)

# Piazza quiz 4: A-priori (Bonus)

Consider the following dataset consisting of 10 transactions

| | Juice | Milk | Beer | Cheese | Chocolate | Yoghurt | Sugar | Flour | Egg | Wine |
|---|---|---|---|---|---|---|---|---|---|---|
| Customer 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| Customer 2 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| Customer 3 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| Customer 4 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| Customer 5 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Customer 6 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Customer 7 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| Customer 8 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 |
| Customer 9 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| Customer 10 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |

Find all itemsets with support greater than 35% (i.e. found in four or more transactions). How many are there?

A: 7 itemsets
B: 9 itemsets
C: 11 itemsets
D: 13 itemsets
E: Don't know

# Solution:

| | Juice | Milk | Beer | Cheese | Chocolate | Yoghurt | Sugar | Flour | Egg | Wine |
|---|---|---|---|---|---|---|---|---|---|---|
| Customer 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| Customer 2 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| Customer 3 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| Customer 4 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| Customer 5 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Customer 6 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Customer 7 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| Customer 8 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 |
| Customer 9 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| Customer 10 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |

There are 11 itemsets with support of at least 35%. They are:

{Juice}, {Milk}, {Yoghurt}, {Egg}, {Wine}
{Juice, Milk}, {Juice, Yoghurt}, {Milk, Yoghurt}, {Wine, Milk}, {Wine, Yoghurt}
{Juice, Milk, Yoghurt}

# Resources