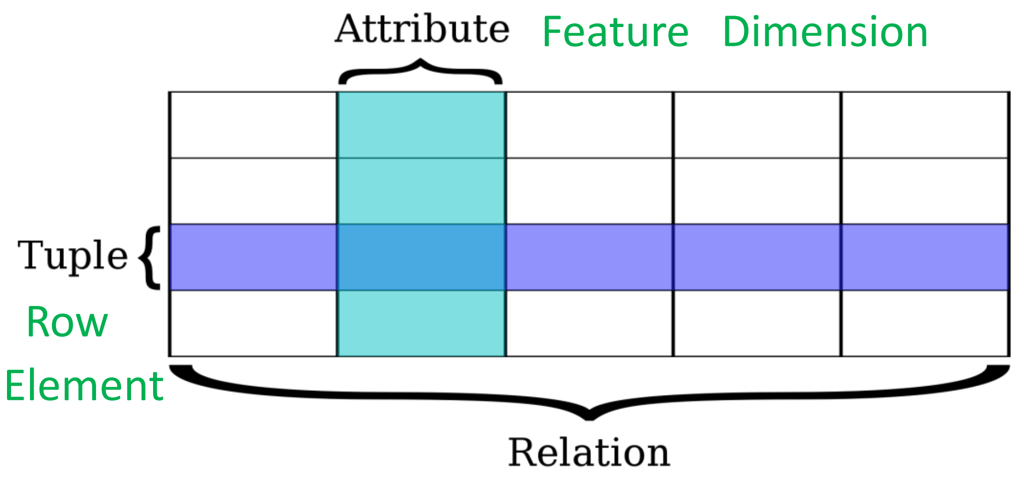
Unsupervised Learning (cont'd)

Praphul Chandra

1. James, Gareth, et al. *An introduction to statistical learning*. Vol. 6. New York: springer, 2013.
2. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning.* Vol. 1. Springer, Berlin: Springer series in statistics, 2001.
3. Kuhn, Max, and Kjell Johnson. *Applied predictive modeling*. New York: Springer, 2013.

What does data look like?



$$\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbb{R}^p$$

$$X \in \mathbb{R}^{n \times p}$$

Unsupervised Learning: Definitions

- … algorithms used to draw inferences from datasets consisting of input data without labeled responses.

- … task of inferring a function to describe hidden structure from unlabeled data.
    - Distribution / Density
    - Summary statistics
    - Clustering
    - Principal Components Analysis

Patterns in data

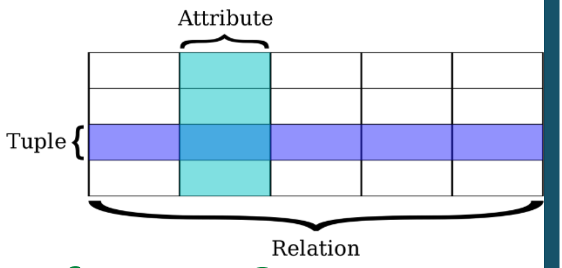
- They describe structure (patterns) in the data
  i. Which value(s) occur most frequently?
  ii. How much does the data vary?
  iii. How symmetrically does data vary around center?
  iv. Is data clustered around value(s)?
  v. Sub-space where data is “concentrated”
- Summary statistics
  i. Median
  ii. Variance, Standard Deviation
  iii. Skewness, Kurtosis
  iv. Mode
- Multiple dimensions
  i. Are two features / dimensions correlated

- Clustering
  - Find data elements which are similar.
  - Finding “areas” in space where data is concentrated
- Association Rules
  - Find features (dimensions) which occur together
  - Find features (dimensions) which are “correlated”
- Dimensionality Reduction
  - Find smaller dimensional representations of the data which preserve it’s essential structure.
  - Find subspaces where data varies the most.

Association Rule Mining

Conceptual Overview

Association Rules



- What does the value of one feature tell us about the value of another feature?
  - People who buy diapers are likely to buy baby powder
  - If (people buy diaper), then (they buy baby powder)
  - Caution : Watch the directionality! (A➔B does not mean B ➔A)
- Association rules
  - Are statements about relations among features (attributes) : across elements (tuples)
  - Use a transaction-itemset data model

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

	Beer	Bread	Milk	Diaper	Eggs	Coke
T_1	0	1	1	0	0	0
T_2	1	1	0	1	1	0
T_3	1	0	1	1	0	1
T_4	1	1	1	1	0	0
T_5	0	1	1	1	0	1

Association Rules = Market Basket Analysis?

T_1	0	1	1	0	0	0
T_2	1	1	0	1	1	0
T_3	1	0	1	1	0	1
T_4	1	1	1	1	0	0
T_5	0	1	1	1	0	1

- Most common use
  - Each basket (purchase) is a row and each item is a column
- Not the only use
  - Can work in any dataset where features take only two use values : 0/1
  - Can work in any dataset where features can be *represented as* taking only two use values : 0/1
    - Preprocessing: Discretization, Feature selection
- Association Rules beyond Market Basket Analysis
  - People who visit webpage X are likely visit webpage Y.
  - Nodes which run a web server are likely to run linux.
  - People who have age-group [30,40] & income [>$100k] are likely to own home

Measures of effectiveness

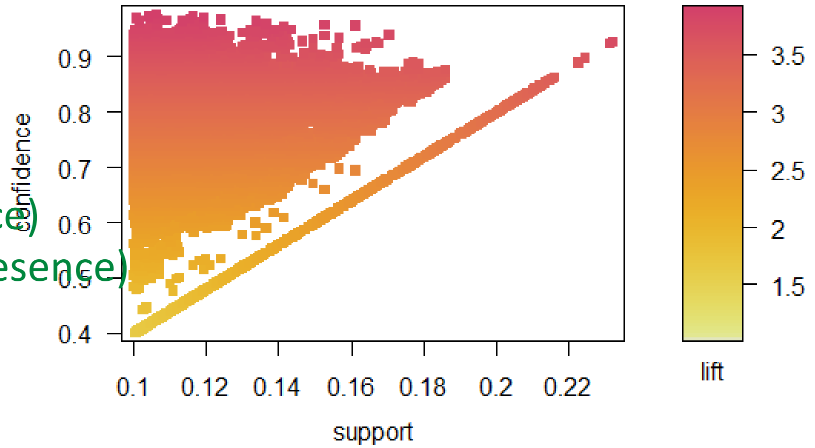
- What do association rules look like?
  - {diapers} ➔ {baby powder}
  - {bread, butter} ➔ {milk}
  - {bat, ball, pads} ➔ {helmet}
  - X ➔ Y :: If {X}, Then {Y}
  - If Precondition, Then Conclusion
  - If Antecedent, Then Consequent
- How good / significant is a rule?
  - An association rule is a probabilistic statement
  - How much historical data **supports** your rule?
  - How **confident** are we that the rule holds?
- Support (a.k.a. Coverage) of X➔Y
  - Fraction of rows containing both X & Y
  - P(X and Y): Joint Probability
  - Support (X ➔ Y) = Support (Y ➔ X)
- Confidence of X➔Y
  - Among rows containing X, fraction of rows containing Y
  - P(Y|X) : Conditional Probability
  - Confidence (X ➔ Y) ≠ Confidence (Y ➔ X)
- What do association rules really look like?
  - $X \rightarrow_{\text{support, confidence}} Y$

Measures of effectiveness (cont'd)

- {Diaper, Beer} ➔ Milk
  - Support = 2/5, Confidence = 2/3
- {Milk} ➔ {Diaper, Beer}
  - Support = 2/5, Confidence = 2/4
- {Milk, Diaper} ➔ Bread
  - Support = 2/5, Confidence = 2/3
- {Milk, Beer} ➔ Diaper?

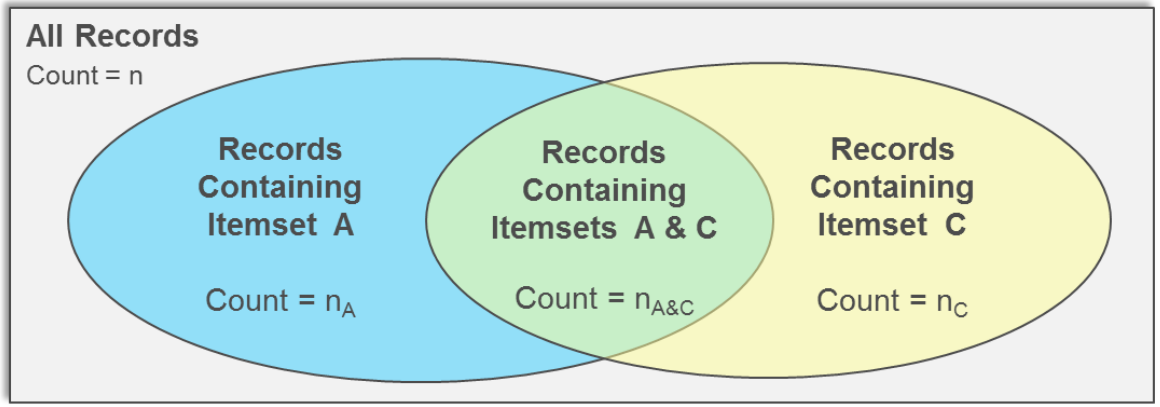
	Beer	Bread	Milk	Diaper	Eggs	Coke
T_1	0	1	1	0	0	0
T_2	1	1	0	1	1	0
T_3	1	0	1	1	0	1
T_4	1	1	1	1	0	0
T_5	0	1	1	1	0	1

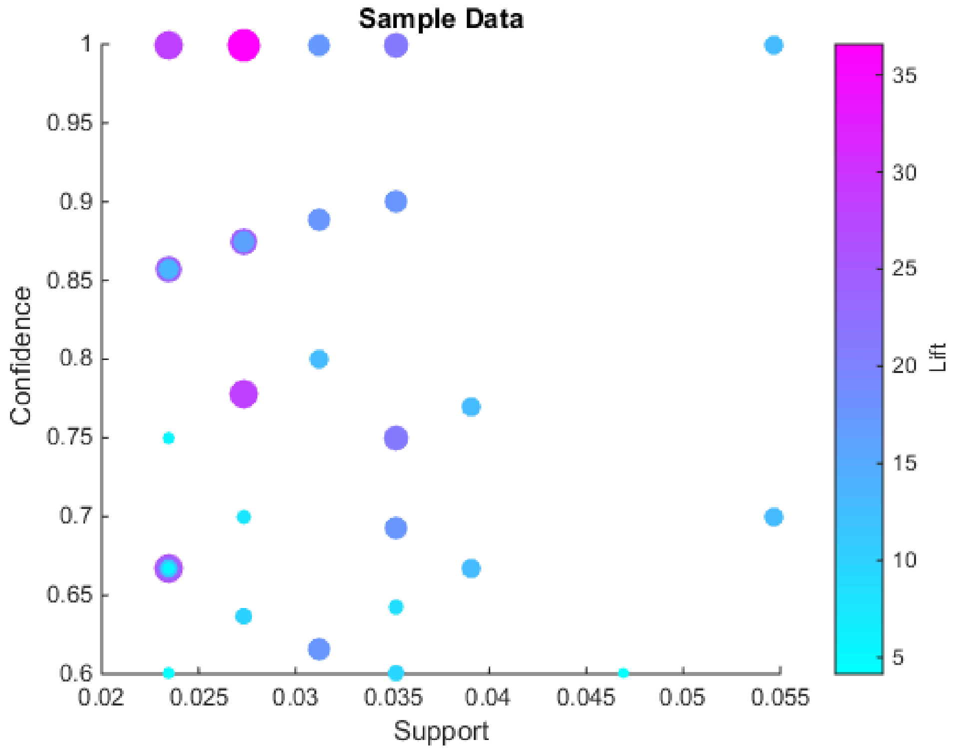
- Confidence = 1?
  - Caution : Diaper is very popular!
  - Does the inclusion of {Milk, Beer} increase the probability of Diaper?
- Lift
  - Confidence (X➔ Y)/Support(Y) or equivalently P(Y|X) / P(Y)
  - > 1 : X & Y positively correlated (Presence of X lifts probability of Y's presence)
  - < 1 : X & Y negatively correlated (Presence of X reduces probability of Y's presence)
  - = 1 X & Y not correlated



Measures of effectiveness (cont'd)

- Support
- Confidence
- Lift
- Others: Affinity, Leverage





Rule = A → C

$$\text{Support (A)} = \frac{n_A}{n} \qquad \text{Support (C)} = \frac{n_C}{n} \qquad \text{Support (A\&C)} = \frac{n_{A\&C}}{n}$$

$$\text{Confidence (A} \rightarrow \text{C)} = \frac{\text{Support(A\&C)}}{\text{Support(A)}} = \frac{n_{A\&C}}{n_A}$$

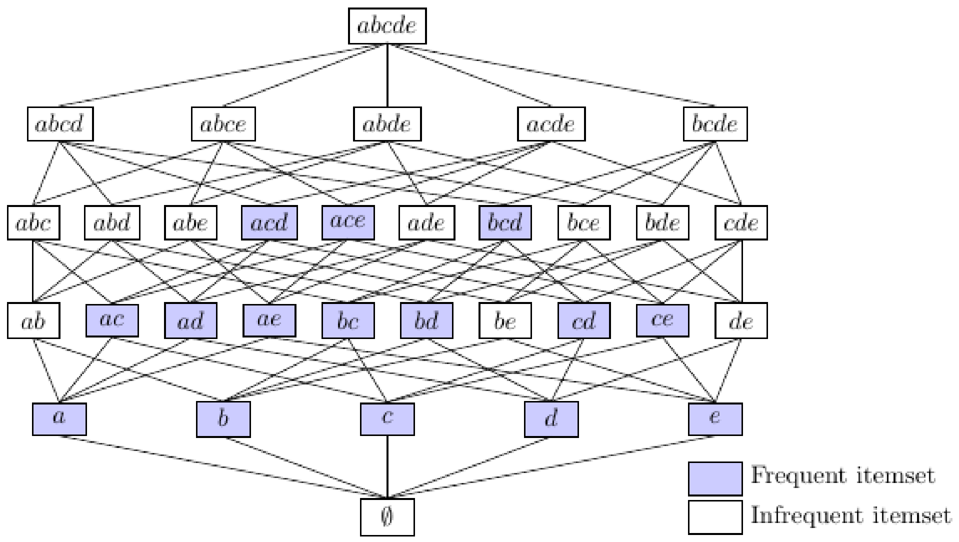
$$\text{Lift(A\&C)} = \frac{\text{Confidence(A} \rightarrow \text{C)}}{\text{Support(C)}} = \frac{\text{Support(A\&C)}}{\text{Support(A)} * \text{Support(C)}} = \frac{n * n_{A\&C}}{n_A * n_C}$$

$$\text{Affinity(A\&C)} = \frac{\text{Support(A\&C)}}{\text{Support(A)} + \text{Support(C)} - \text{Support(A\&C)}} = \frac{n_{A\&C}}{n_A + n_C - n_{A\&C}}$$

$$\text{Leverage(A\&C)} = \text{Support(A\&C)} - [\text{Support(A)} * \text{Support(C)}] = \frac{n_{A\&C}}{n} - \frac{n_A * n_C}{n^2}$$

Apriori

- Key Idea
  - If {a,c,f} is frequent, {a,c} must be frequent
  - Downward closure a.k.a. anti-monotonicity
- Algorithm
  - Find all frequent 1-itemsets (frequent ➔ > support)
  - Find all frequent 2-itemsets for filtered 1-itemsets
  - Find all frequent 3-itemsets for filtered 2-itemsets
  - …
- Salient Features
  - Exploits downward closure to optimize search
  - Lower Support ➔ Higher computational complexity
  - Confidence, Lift as post-processing filters



Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

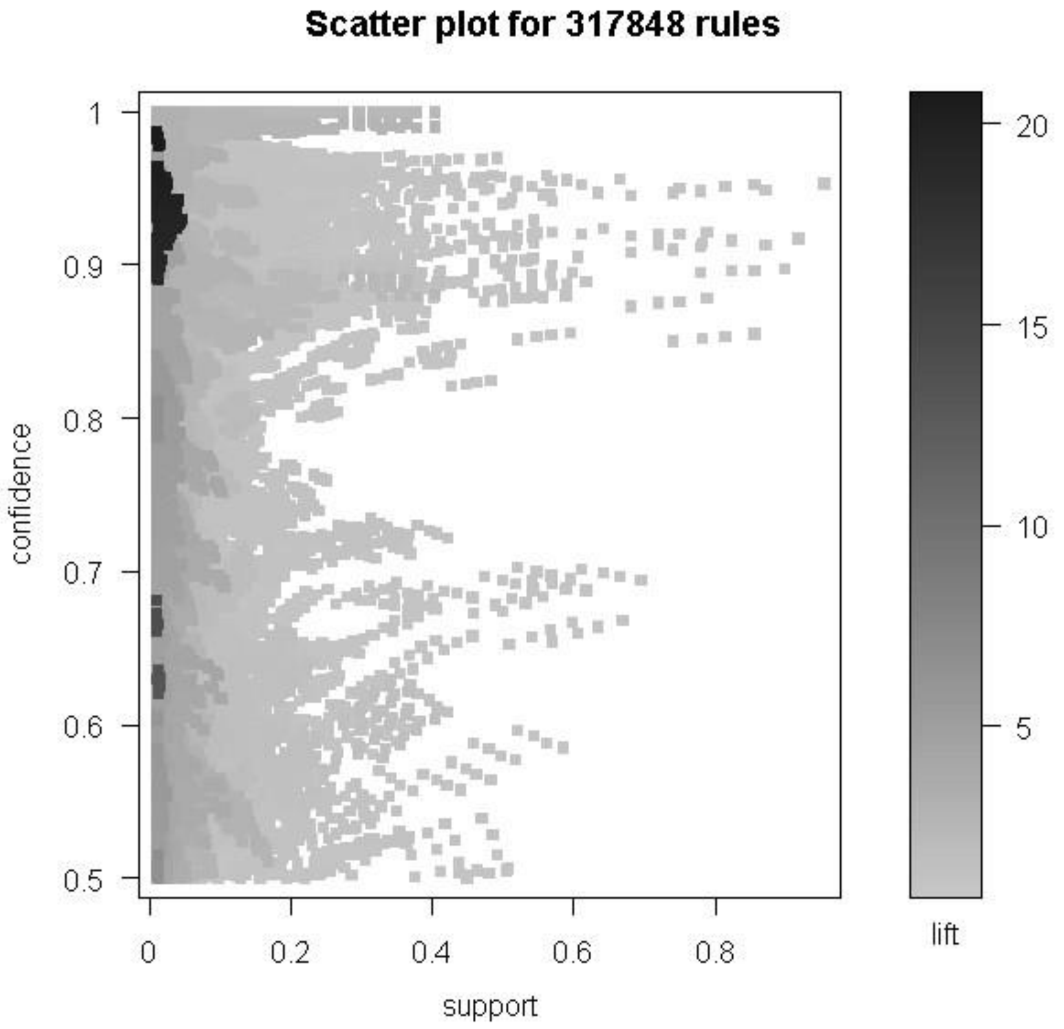
Itemset	Support
{1 3}*	2
{1 4}	1
{3 4}	1
{2 3}*	2
{2 5}*	3
{3 5}*	2
{1 2}	1
{1 5}	1

C_3

Itemset	Support
{1 3 4}	1
{2 3 5}*	2
{1 3 5}	1

Example : Apriori in R

```
data("AdultUCI");
Adult = as(AdultUCI, "transactions");
rules = apriori(Adult, parameter=list(support=0.01, confidence=0.5));
```



https://www.r-bloggers.com/association-rule-learning-and-the-apriori-algorithm/

Apriori : Limitations

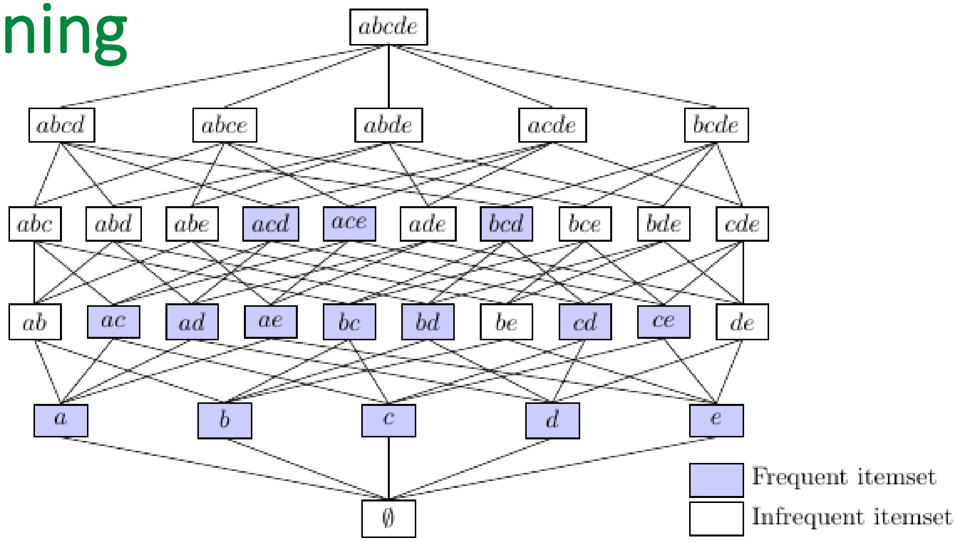
- Computational Complexity
  - How long does it take to run?
  - How much memory does it need?

- Approaches
  - Throw more compute / RAM at it
  - Parallelize
  - Increase support
  - Leverage item hierarchy
  - Another algorithm?

- Rare patterns
  - Rules with low support but maybe very valuable
  - People who buy ______ likely to buy luxury cars

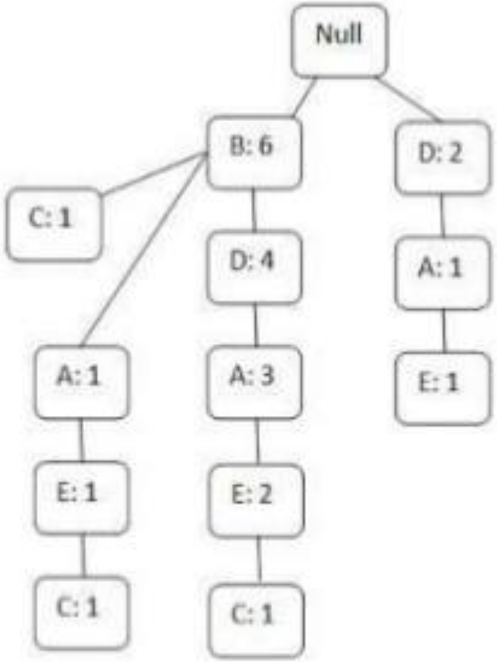
- When sequence of transactions matters
  - Define a sequence as an item
  - Combinatorial Explosion : Computational Complexity
  - Read-Up!

Frequent Pattern Growth : Association Rule Mining



- Apriori
  - Use **frequent** k-itemsets to generate k+1-itemsets candidates
  - Scan DB to determine frequent k+1-itemsets
  - Iterate
  - ➔ Multiple scans of DB;
  - + Multiple itemsets (Computational Complexity; Does not scale)

- FP Growth: Key !dea
  - Scan the DB only twice;
  - Summarize itemsets in an efficient data structure (FP-Tree)
  - Extract frequent itemsets from the FP-Tree



FP-Growth : Growing the Tree

TID	Items
1	E, A, D, B
2	D, A, C, E, B
3	C, A, B. E
4	B, A, D
5	D
6	D,B
7	A,D,E
8	B,C

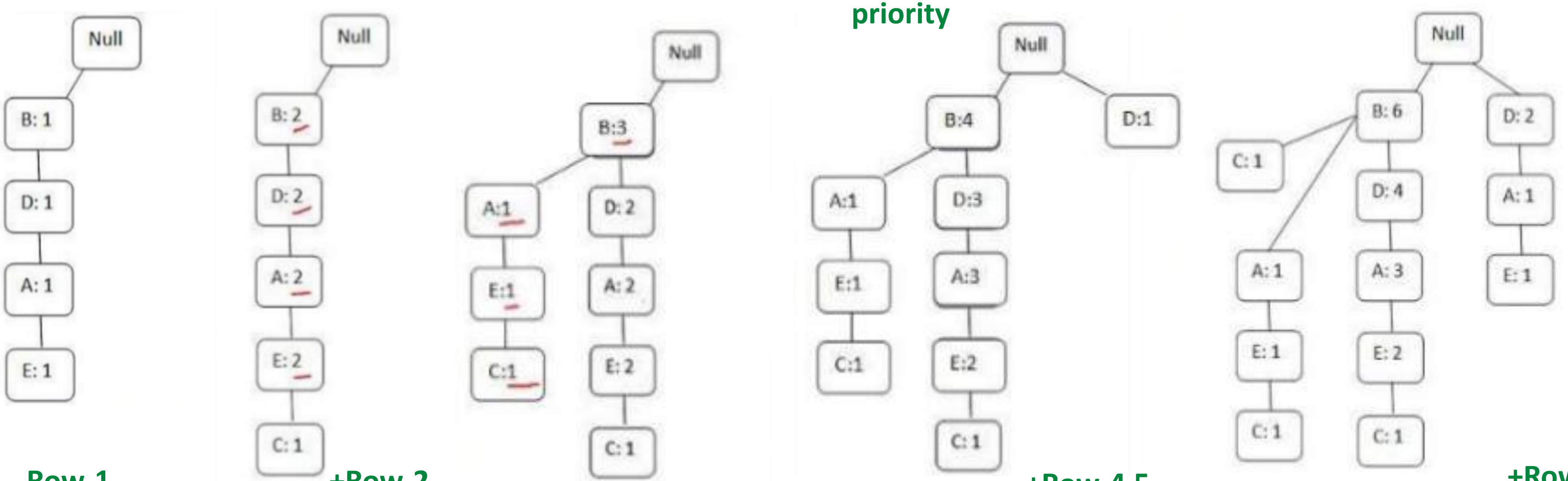
Transaction data in DB

TID	frequency	priority
A	5	3
B	6	1
C	3	5
D	6	2
E	4	4

1-Itemset Support / **1-Itemset priority**

TID	Items	Ordered Items
1	E, A, D, B	B,D,A,E
2	D, A, C, E, B	B,D,A,E,C
3	C, A, B. E	B,A,E,C
4	B, A, D	B,D,A
5	D	D
6	D,B	B,D
7	A,D,E	D,A,E
8	B,C	B,C

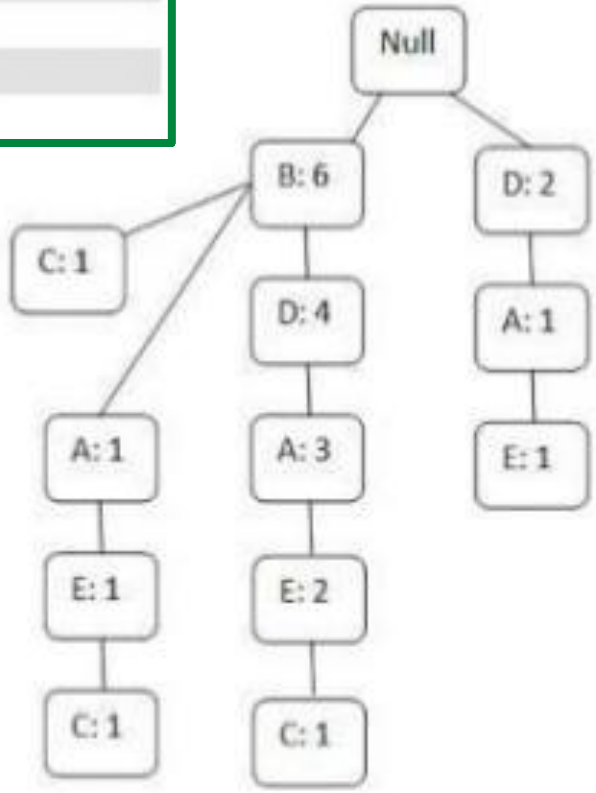
Sorted transaction data



Row-1 **+Row-2** **+Row-3** **+Row-4,5** **+Row-6,7,8**

FP-Growth : Building and Rules Extraction

TID	Items	Ordered Items
1	E, A, D, B	B,D,A,E
2	D, A, C, E, B	B,D,A,E,C
3	C, A, B. E	B,A,E,C
4	B, A, D	B,D,A
5	D	D
6	D,B	B,D
7	A,D,E	D,A,E
8	B,C	B,C



- Scan-1
  - Find support for each 1-itemset; Discard in-frequent 1-itemsets
  - Sort frequent 1-itemsets in decreasing order of support
- Scan-2
  - Read 1 transaction at a time & map it to a path in the tree
  - Fixed sorted order ensures paths overlap when transactions share itemsets (counters incremented)
    - More paths overlap ➔ More compression ➔ Tree fits in memory
    - If all transactions contain the same itemset ➔ 1 path in the tree
    - If no transactions share itemsets ➔ Tree as big as DB
- Association Rules Extraction
  - Pick an 1-itemset (Say e)
  - Check if it is a frequent itemset (Yes; support =4)
  - Check 2-itemsets ending in e: de, ce, be, ae
    - Supports : de (0), ce(0), be(0), ae(4)
    - Check 3-itemsets ending in ae: bae, cae, dae
    - …
  - Note: This is the conditional FP-tree for e.

Association Rules : Summary

- Association Rules
  - Are probabilistic statements
  - About relations among features - across elements
  - Use a transaction-itemset data model
  - The strength (statistical significance) of an association rule is measured using support, confidence, lift etc.
- Applications
  - Market Basket Analysis
  - Any dataset where features take values : 0/1
  - Can work in any dataset where features can be *represented as* taking only two use values : 0/1
    - Preprocessing: Discretization, Feature selection
- Apriori
  - Input : Dataset, minsupport
  - Output: association rules
  - Exploits downward closure to optimize search
  - Lower Support ➔ Higher computational complexity
  - Confidence, Lift as post-processing filters
- FP Growth
  - Scan the DB only twice;
  - Summarize itemsets in an efficient data structure (FP-Tree)
  - Extract frequent itemsets from the FP-Tree

Unsupervised Learning: Summary

- … algorithms used to draw inferences from datasets consisting of input data without labeled responses.
- … the task of inferring a function to describe hidden structure from unlabeled data.
  - Distribution / Density
  - Summary statistics
  - Clustering: Find data elements (rows) which are similar.
  - **Association Rules:** Find features (dimensions) which are correlated
  - Dimensionality Reduction: Find smaller dimensional representations which preserve data's essential structure.
- Unsupervised
  - Association Rules: Find patterns when we don't know what we are looking for.
    - {Diaper, Beer} ➔ **Milk**
    - {Milk} ➔ {Diaper, Beer}
    - {Milk, Diaper} ➔ **Beer**
- Supervised
  - What if we are only interested in identifying customers who bought Milk?
  - Split the customer base into two classes: Customers who bought Milk and who did not.
  - Binary classification problem : Given purchases of other customers

Q?

Praphul Chandra

Insofe