Association Rule Mining

Arend Hintze

example: Market Basket

Market-Basket transactions

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

Example of Association Rules

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

informal definition

- Given a set of transactions where each transaction is a set of items

- extract a set of rules that summarize the relationship among subsets of those items

constraints

- attributes must be asymmetric binary: words in documents, items in basket, medical conditions in patients

DocID	*Words*
1	big, data, analytics
2	big, fish, small, pond
3	data, driven, discovery
4	computing, cloud, big, data
5	fish, lake, nutrients

VisitID	*Medical Conditions*
1	Cough, Fever
2	Seizure, Chest Pain
3	Vomit, Diarrhea, Stomach Pain
4	Dizziness, Headache
5	Cough, Running Nose, Sneezing

considerations

- number of possible rules is large
- how to decide what is a “good” rule and what isn’t?
- the set X can not contain elements from Y (tautology)

Rule Evaluation Measures

- given rule {X}->{Y}

support (s): how often can the rule be applied?

$$s = \frac{\sum_{t=1}^{T} I(X \cup Y \in t)}{\mathrm{T}}$$

I is an indicator function

T is number of transactions

$$I(z) = \begin{cases} 1 & \text{if } z \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

confidence (c): how strong is the rule?

$$c = \frac{\sum_{t=1}^{T} I(X \cup Y \in t)}{\sum_{t=1}^{T} I(X \in t)}$$

→ **Number of transactions that contain X and Y**

→ **Number of transactions that contain X**

example

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

$\{\text{Milk, Diaper}\} \Rightarrow \{\text{Beer}\} \quad s = \frac{2}{5} = 0.4,\ c = \frac{2}{3} = 0.67$

$\{\text{Bread}\} \Rightarrow \{\text{Milk}\} \quad s = \frac{3}{5} = 0.4,\ c = \frac{3}{4} = 0.75$

$\{\text{Milk}\} \Rightarrow \{\text{Bread}\} \quad s = \frac{3}{5} = 0.4,\ c = \frac{3}{4} = 0.75$

s ~ fraction of entire rule true
of that
c ~ fraction of time the result is true

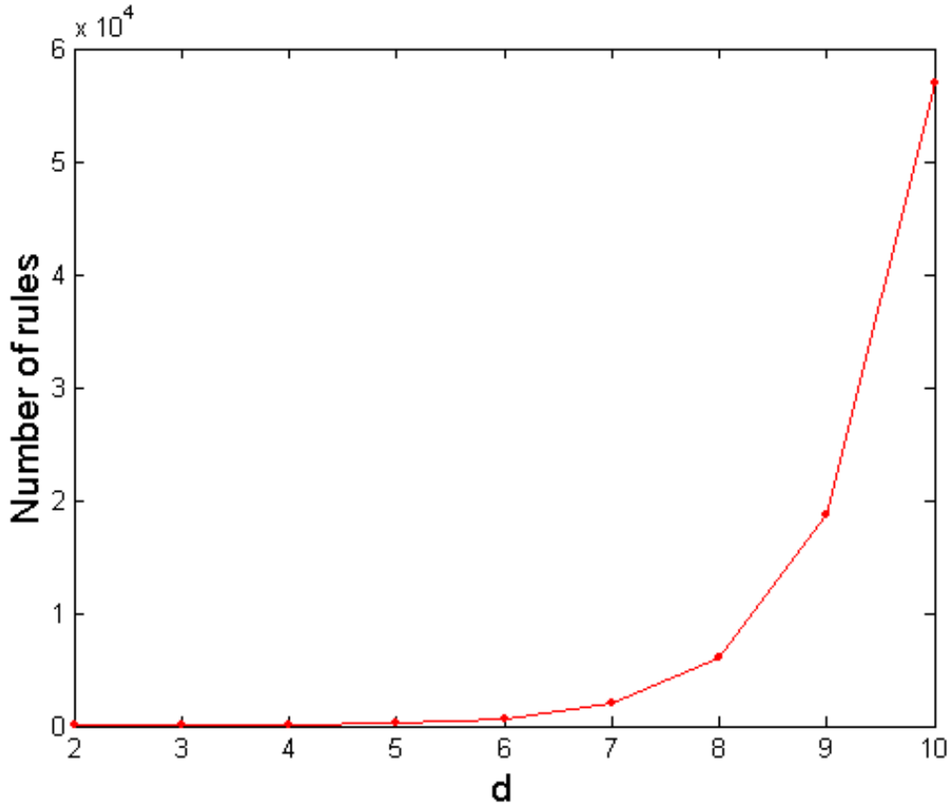
turning this into an algorithm

- given a set of transactions
- find a set of rules that:
  - support >= *minsup* threshold
  - confidence >= *minconf* threshold
    - *minsup* and *minconf* are USER defined!

but how to choose?

- minsup:
    - start with high threshold
    - gradually reduce until you see interesting results
- minconf:
    - start with high threshold
    - gradually reduce it
        - the lower the threshold the more rules you get
        - for binary data mincing should exceed 0.5

there is but one problem



$$R = \sum_{k=1}^{d-1}\left[\binom{d}{k} \times \sum_{j=1}^{d-k}\binom{d-k}{j}\right]$$

$$= 3^{d} - 2^{d+1} + 1$$

If d=6, R = 602 rules

rule and it's item set

these rules:
(milk,beer) -> (bread)
(milk,bread) -> (beer)
(beer,bread) -> (milk)

as do these rules:
(milk) -> (beer,bread)
(bread) -> (milk,beer)
(beer) -> (bread,milk)

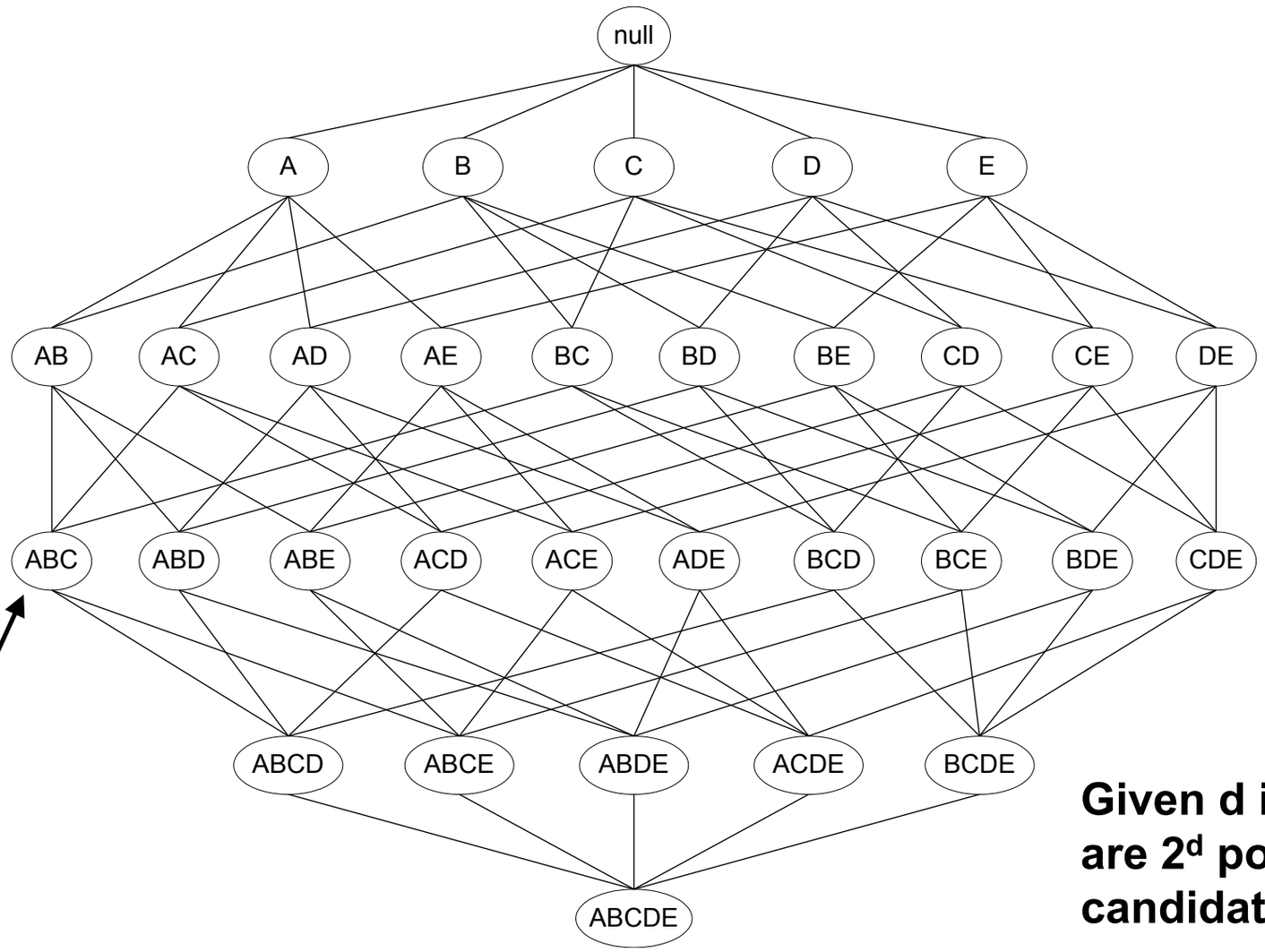
all belong to one item set:
(milk,beer,bread)

support (s):

$$s = \frac{\sum_{t=1}^{T} I(X \cup Y \in t)}{\mathrm{T}}$$

all will have the same support!

itemset instead of rules



a-bc
ab-c
ac-b
b-ac
bc-b
c-ab

one computation instead of 6

Given d items, there are 2^d possible candidate itemsets

anti-monotone property of support

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

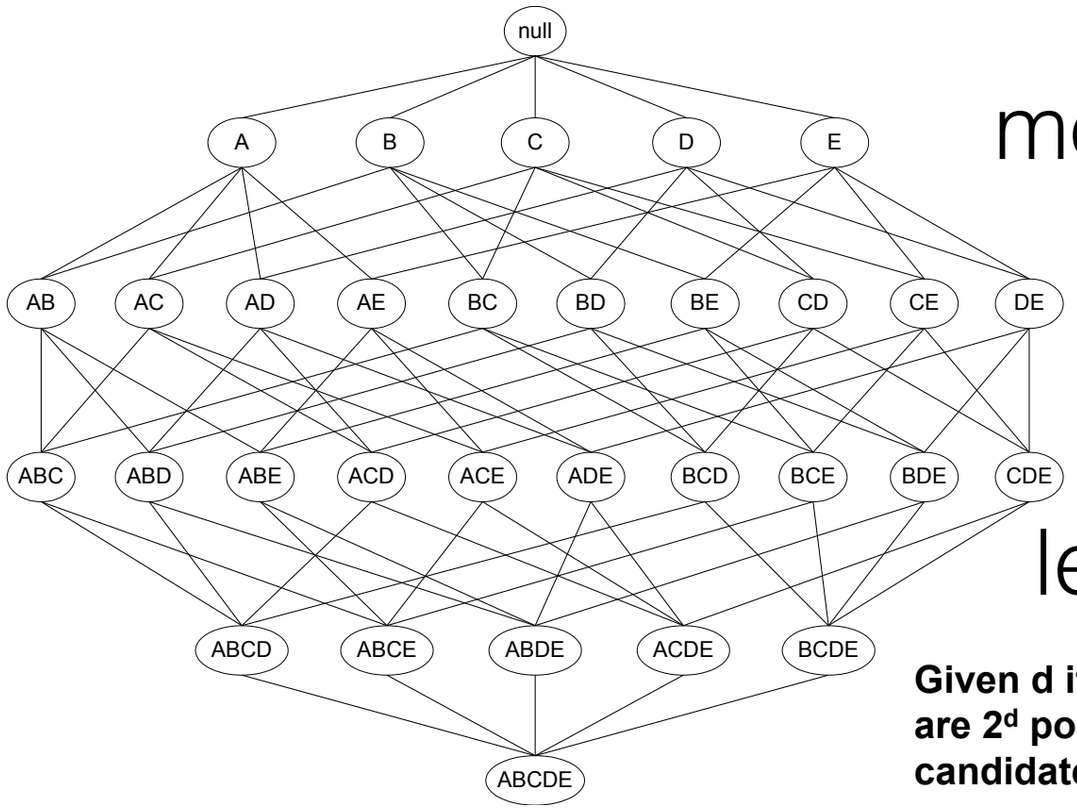
Itemset	Support
{Bread}	4/5
{Bread, Milk}	3/5
{Bread, Milk, Diaper}	2/5

Support is non-increasing when size of an itemset increases

apriori Algorithm part A

- support of an itemset Y will never exceed the support of its subsets

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

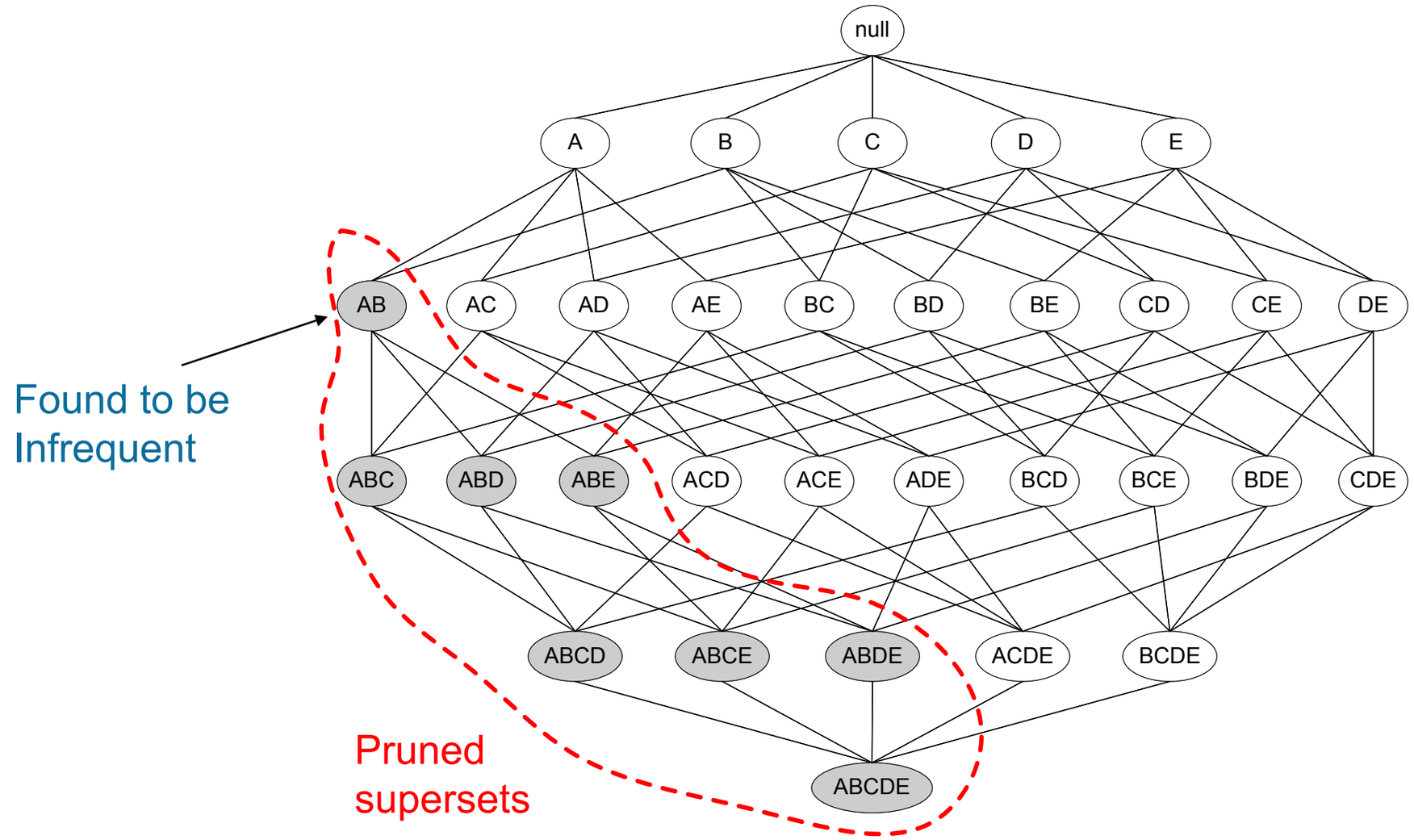


more support

less support

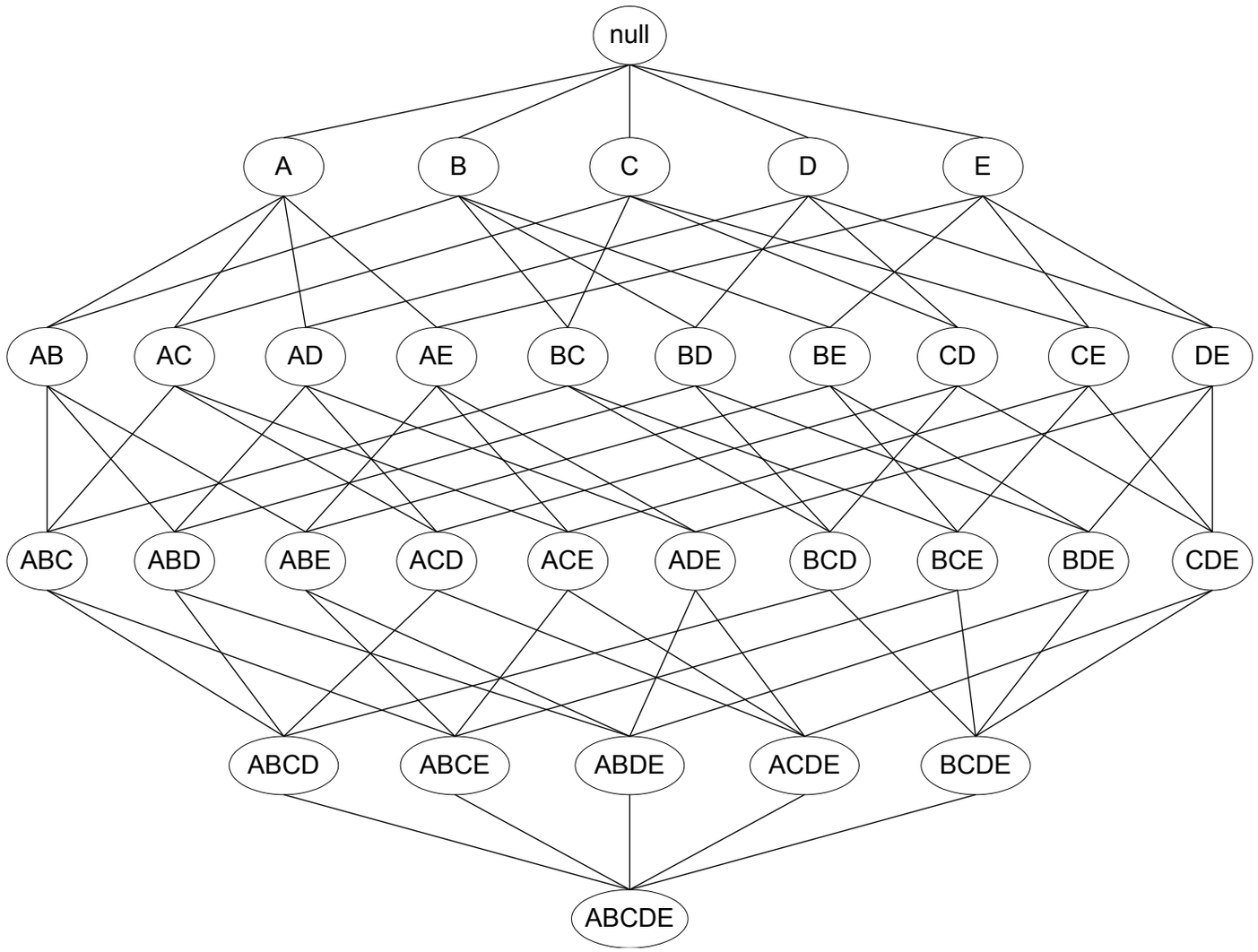
Given d items, there are 2^d possible candidate itemsets

apriori Algorithm part B



If an item set is infrequent then all of its supersets must be also infrequent

what have we just learned?



- use item sets instead of rules
- support will decrease
- frequency will drop from top to bottom

minsup minconf

Illustration

Item	Support
Bread	**0.8**
Coke	**0.4**
Milk	**0.8**
Beer	**0.6**
Diaper	**0.8**
Eggs	**0.2**

Items (1-itemsets)

Itemset	Support
{Bread,Milk}	**0.6**
{Bread,Beer}	**0.4**
{Bread,Diaper}	**0.6**
{Milk,Beer}	**0.4**
{Milk,Diaper}	**0.6**
{Beer,Diaper}	**0.6**

Pairs (2-itemsets)

(No need to generate candidates involving Coke or Eggs)

Minimum Support = 0.6

If every subset is considered,
$^6C_1 + {}^6C_2 + {}^6C_3 = 41$
With support-based pruning,
$6 + 6 + 1 = 13$

Triplets (3-itemsets)

Itemset	Support
{Bread,Milk,Diaper}	**0.6**

Item sets not rules yet!

association rule generation

- only compute all the bipartition rules from item sets that have high support

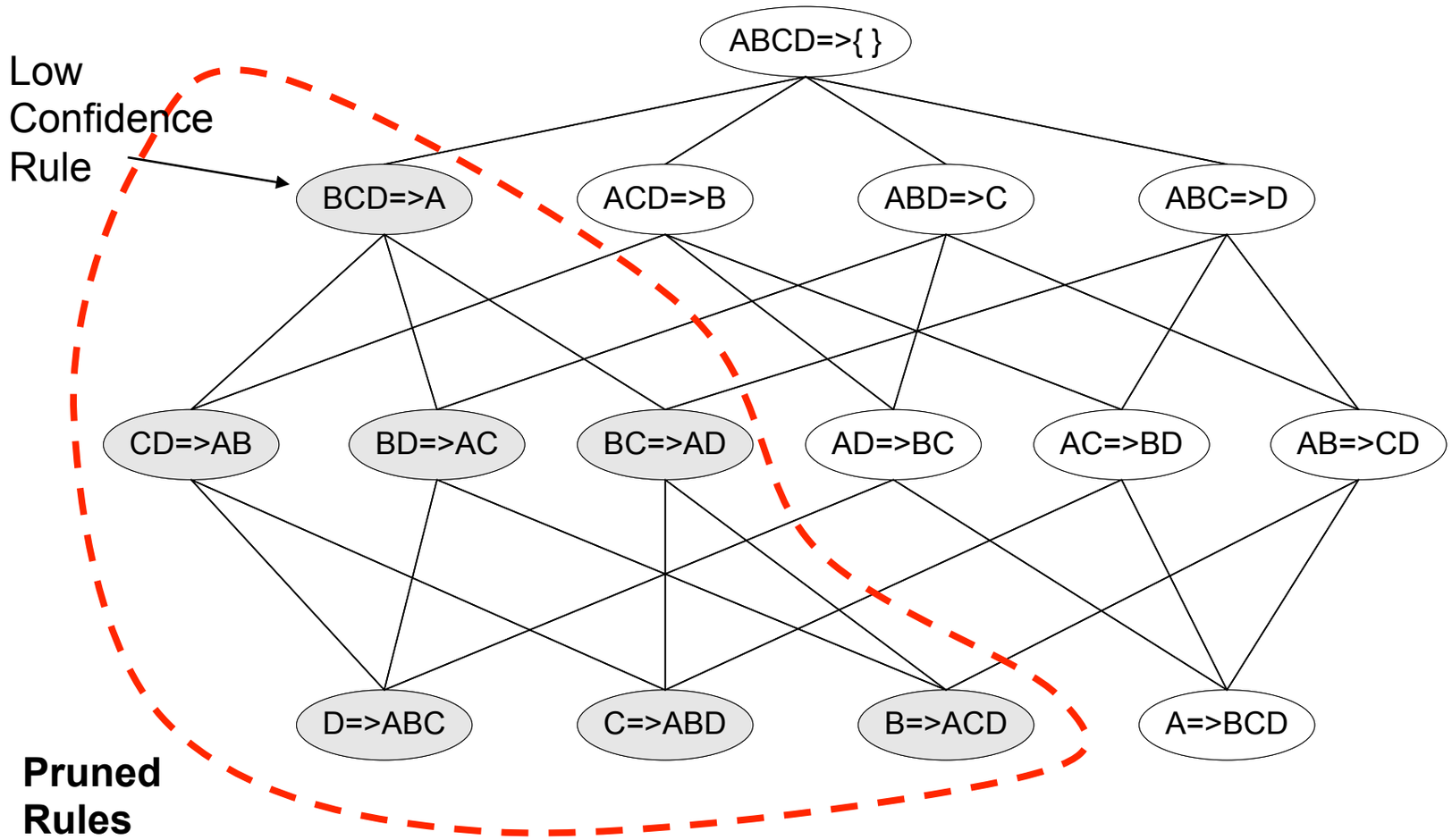
- those rules have the same support but different confidence: (ABC-D) (AB-CD) (A-BCD)

confidence (c):

c(ABC-D) >= c(AB-CD) >= c(A-BCD)

$$c = \frac{\sum_{t=1}^{T} I(X \cup Y \in t)}{\sum_{t=1}^{T} I(X \in t)}$$

effect on the graph

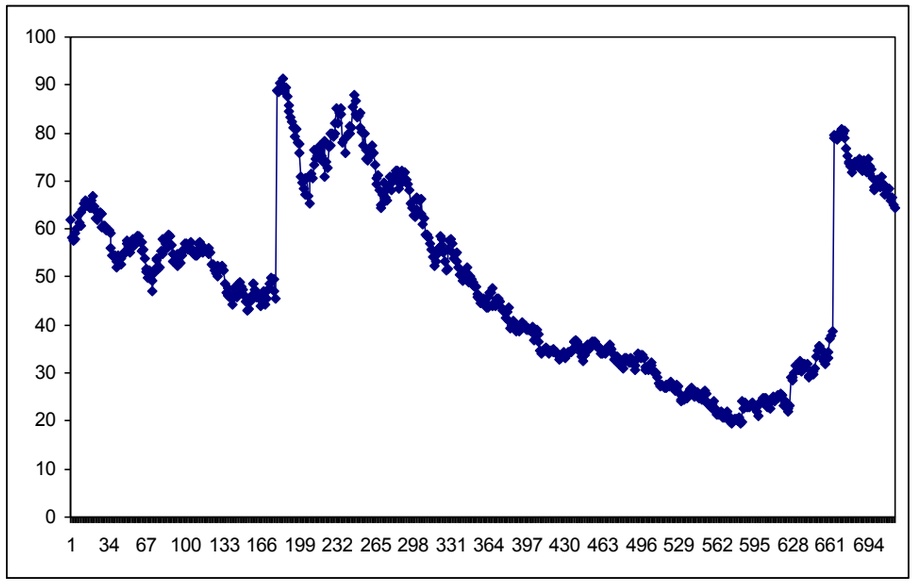


http://aimotion.blogspot.com/2013/01/machine-learning-and-data-mining.html

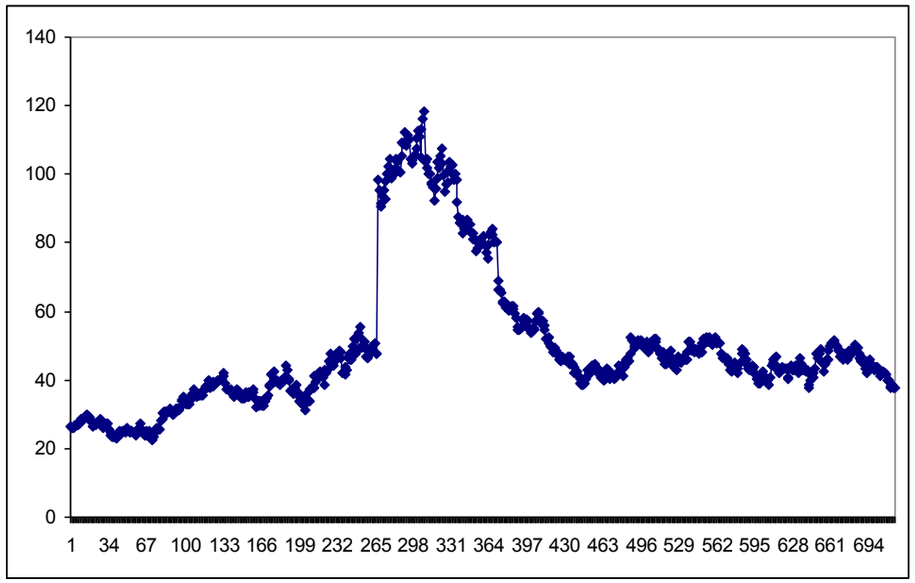
can we use this on stock data?

problem: continuous data and not item sets…

Cisco Systems
Ticker: CSCO



Applied Materials
Ticker: AMAT



preprocessing

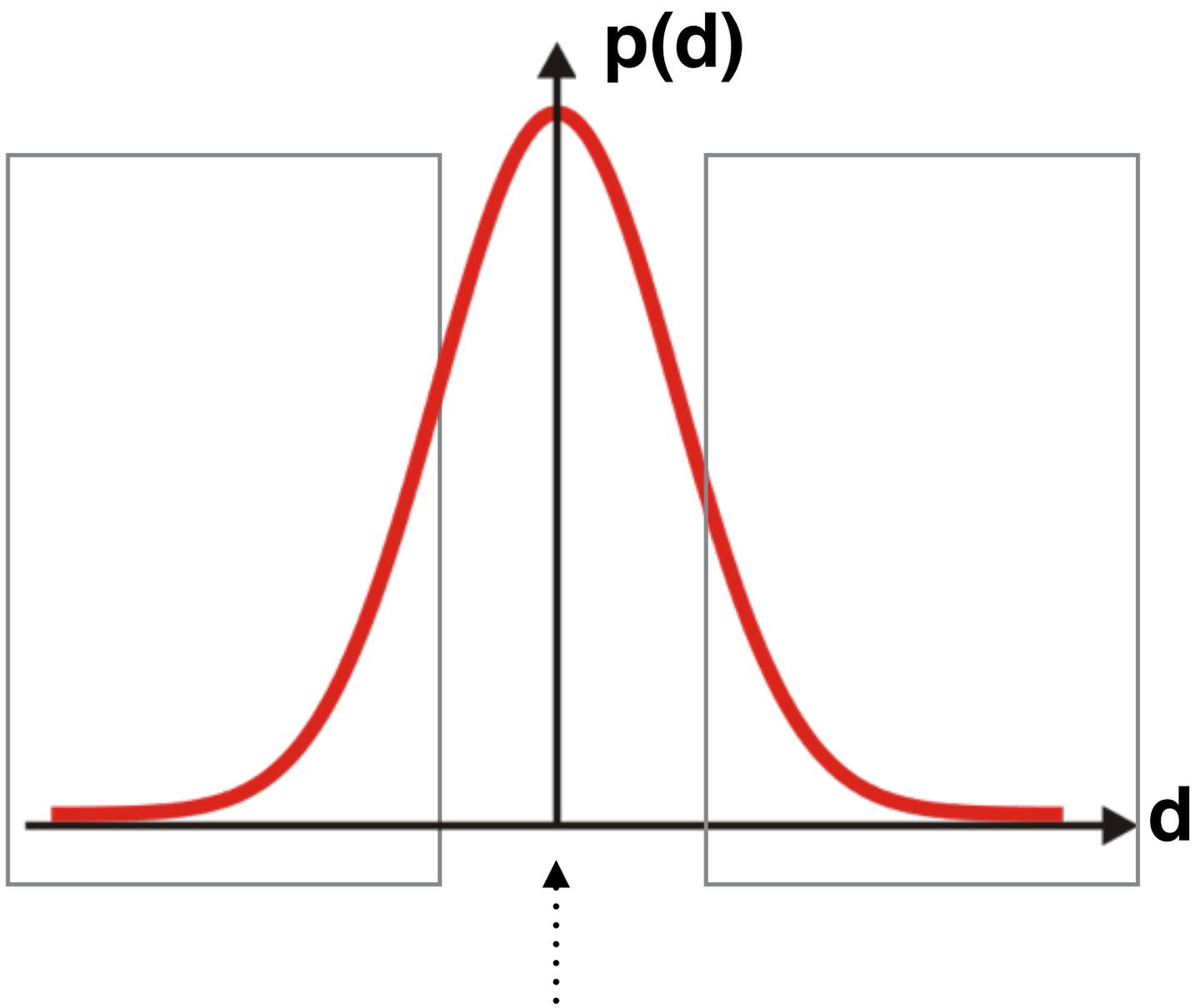
- compute the percentage change with respect to the closing price:

$$d_t = \frac{x_t - x_{t-1}}{x_{t-1}}$$

x_t : closing price on day t

- discretize the attribute d
  - stock up if d > 2
  - stock down if d < -2

sparseness is key!



these destroy confidence

$$c = \frac{\sum_{t=1}^{T} I(X \cup Y \in t)}{\sum_{t=1}^{T} I(X \in t)}$$

issues

count		
5		
75		coffee
15	tea	coffee
5	tea	

issues

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea → Coffee

Confidence= P(Coffee|Tea) = 15/20 = 0.75

but P(Coffee) = 0.9, which means knowing that a person drinks tea reduces the probability that the person drinks coffee!

⇒ Note that P(Coffee|$\overline{\text{Tea}}$) = 75/80 = 0.9375