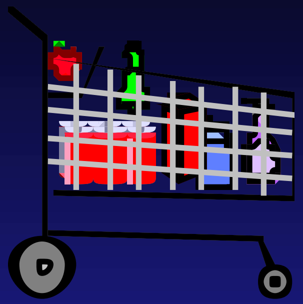
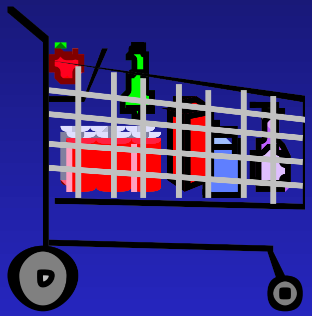
KDD Laboratory
College of IT, UNC-Charlotte





Association Rules

presented by

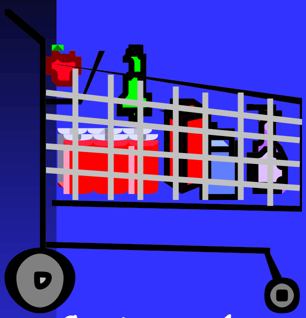
Zbigniew W. Ras[*,#)]

[*)] *University of North Carolina – Charlotte*
[#)] *Polish-Japanese Academy of Information Technology*

Market Basket Analysis (MBA)

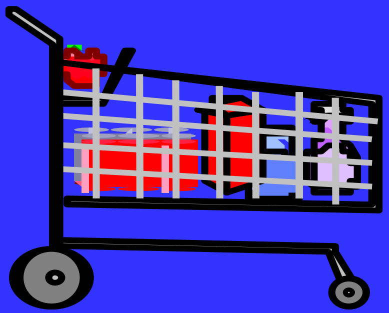
Customer buying habits by finding associations and correlations between the different items that customers place in their “shopping basket”

Milk, eggs, sugar, bread



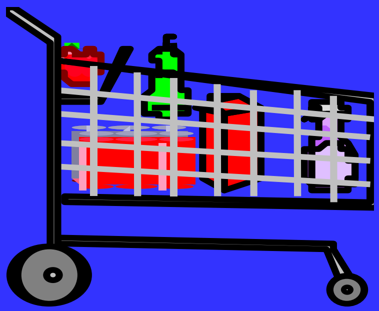
Customer1

Milk, eggs, cereal, bread



Customer2

Eggs, sugar

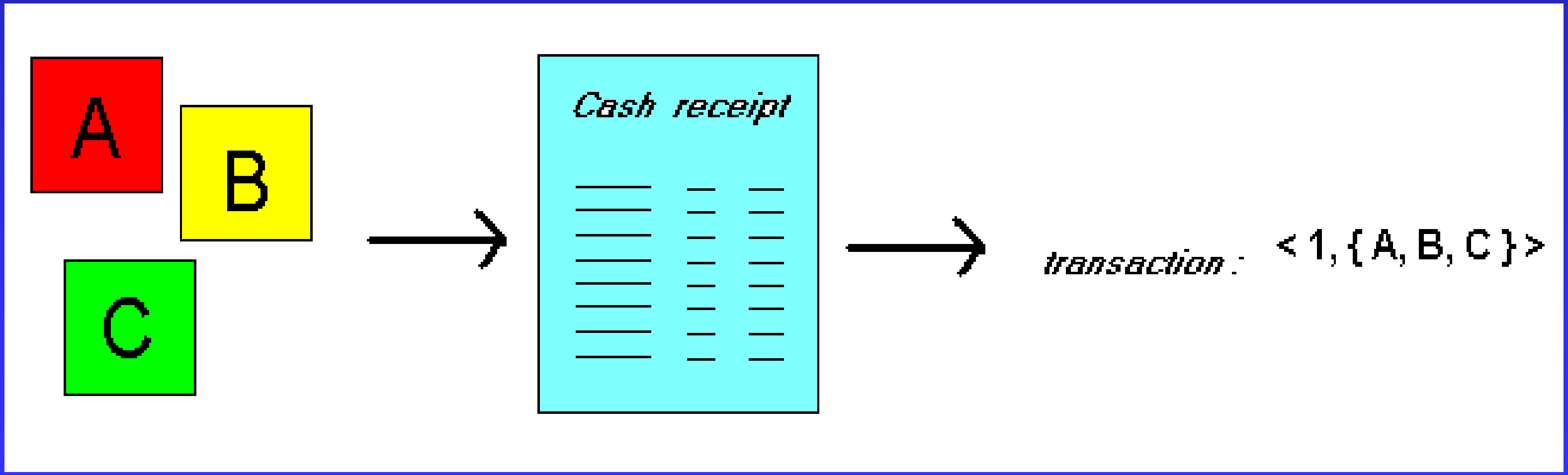


Customer3

Market Basket Analysis

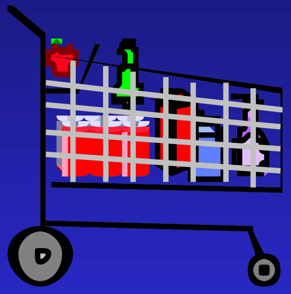
Given: a database of customer transactions, where each transaction is a set of items

Find groups of items which are frequently purchased together



Goal of MBA

- Extract information on purchasing behavior
- Actionable information: can suggest
  - new store layouts
  - new product assortments
  - which products to put on promotion



MBA applicable whenever a customer purchases multiple things in proximity

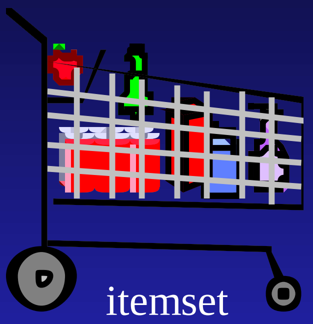
Association Rules

- Express how product/services relate to each other, and tend to group together
- "*if* a customer purchases three-way calling, *then* will also purchase call-waiting"
- Simple to understand
- *Actionable information*: bundle three-way calling and call-waiting in a single package

Basic Concepts

Transactions:

Relational format	Compact format
<Tid,item>	<Tid,itemset>
<1, item1>	<1, {item1,item2}>
<1, item2>	<2, {item3}>
<2, item3>	



Item: single element, Itemset: set of items

Support of an itemset I [denoted by sup(I)]: card(I)

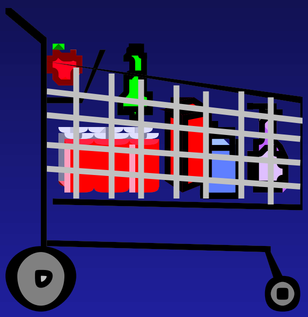
Threshold for minimum support: σ

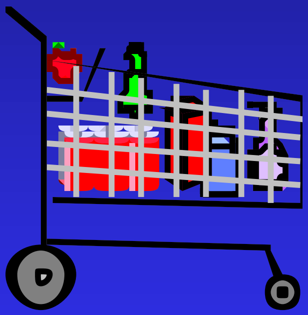
Itemset I is Frequent **if**: $sup(I) \geq \sigma$.

Frequent Itemset represents set of items which are positively correlated

Frequent Itemsets

Transaction ID	Items Bought
1	dairy,fruit
2	dairy,fruit, vegetable
3	dairy
4	fruit, cereals





sup({dairy}) = 3
sup({fruit}) = 3
sup({dairy, fruit}) = 2

If $\sigma = 3$, then

{dairy} and {fruit} are frequent while {dairy,fruit} is not.

Association Rules: AR(s,c)

- $\{A,B\}$ - partition of a set of items
- $r = [A \Rightarrow B]$

  Support of r: $sup(r) = sup(A \cup B)$

  Confidence of r: $conf(r) = sup(A \cup B)/sup(A)$

- Thresholds:
  - minimum support - s
  - minimum confidence - c

$r \in AS(s, c)$, if $sup(r) \geq s$ and $conf(r) \geq c$

Association Rules - Example

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Min. support - 2 [50%]
Min. confidence - 50%

Frequent Itemset	Support
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

- For rule $A \Rightarrow C$:
  - $\sup(A \Rightarrow C) = 2$
  - $\mathrm{conf}(A \Rightarrow C) = \sup(\{A,C\})/\sup(\{A\}) = 2/3$
- The Apriori principle:
  - Any subset of a frequent itemset must be frequent

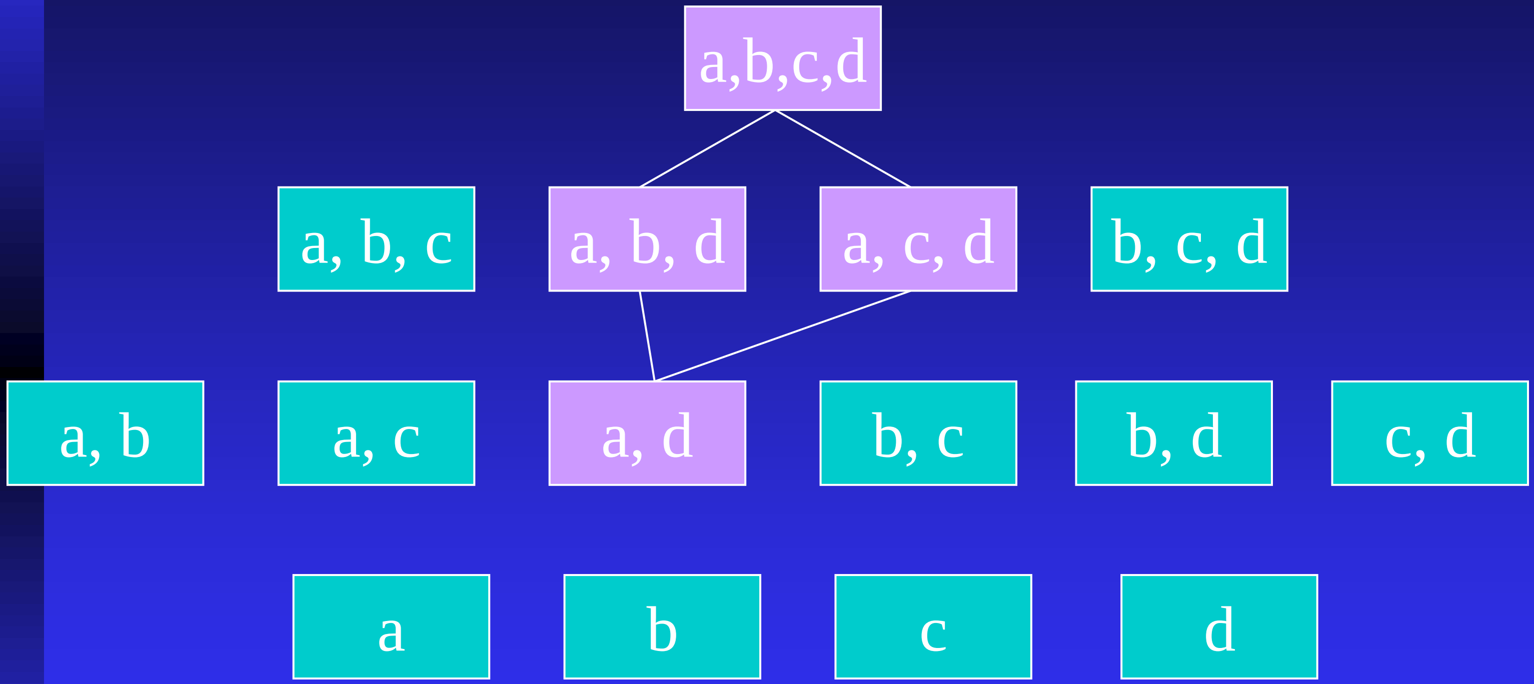
The Apriori algorithm [Agrawal]

- F_k : Set of frequent itemsets of size k
- C_k : Set of candidate itemsets of size k

F_1 := {frequent items}; k:=1;
while card(F_k) $\geq$ 1 do begin
C_{k+1} := new candidates generated from F_k ;
for each transaction t in the database do
increment the count of all candidates in C_{k+1} that
are contained in t ;
F_{k+1} := candidates in C_{k+1} with minimum support
k:= k+1
end

Answer := $\bigcup\{ F_k$: $k \geq 1$ & card(F_k) $\geq 1\}$

Apriori - Example



{a,d} is not frequent, so the 3-itemsets {a,b,d}, {a,c,d} and the 4-itemset {a,b,c,d}, are not generated.

Representative Association Rules

- Definition 1.
  Cover C of a rule $X \Rightarrow Y$ is denoted by $C(X \Rightarrow Y)$
  and defined as follows:
  $C(X \Rightarrow Y) = \{ [X \cup Z] \Rightarrow V : Z, V \text{ are disjoint subsets of } Y\}$.

- Definition 2.
  Set $RR(s, c)$ of Representative Association Rules
  is defined as follows:
  $RR(s, c) =$
  $\{r \in AR(s, c): \sim(\exists r_l \in AR(s, c)) [r_l \neq r \;\&\; r \in C(r_l)]\}$
  s – threshold for minimum support
  c – threshold for minimum confidence

- Representative Rules (informal description):
  [as short as possible] $\Rightarrow$ [as long as possible]

Representative Association Rules

Transactions:
{A,B,C,D,E}
{A,B,C,D,E,F}
{A,B,C,D,E,H,I}
{A,B,E}
{B,C,D,E,H,I}

Find RR(2,80%)

Representative Rules

From (BCDEHI):
{H} ⇒ {B,C,D,E,I}
{I} ⇒ {B,C,D,E,H}

From (ABCDE):
{A,C} ⇒ {B,D,E}
{A,D} ⇒ {B,C,E}

1-element sets	2-element sets	3-element sets	4-element sets	5-element sets
(A, 4)	(AB, 4)	(ABC, 3)	(ABCD, 3)	(ABCDE, 3)
(B, 5)	(AC, 3)	(ABD, 3)	(ABCE, 3)	
(C, 4)	(AD, 3)	(ABE, 4)	(ABDE, 3)	
(D, 4)	(AE, 4)	(ACD, 3)	(ACDE, 3)	
(E, 5)	(BC, 4)	(ACE, 3)		
(H, 2)	(BD, 4)	(ADE, 3)		
(I, 2)	(BE, 5)	(BCD, 4)	(BCDE, 4)	(BCDEH, 2)
	(BH, 2)	(BCE, 4)	(BCDH, 2)	(BCDEI, 2)
	(BI, 2)	(BCH, 2)	(BCDI, 2)	(BCDHI, 2)
	(CD, 4)	(BCI, 2)	(BCEH, 2)	(BCEHI, 2)
	(CE, 4)	(BDE, 4)	(BCEI, 2)	(BDEHI, 2)
	(CH, 2)	(BDH, 2)	(BCHI, 2)	
	(CI, 2)	(BDI, 2)	(BDEH, 2)	
	(DE, 4)	(BEH, 2)	(BDEI, 2)	
	(DH, 2)	(BEI, 2)	(BDHI, 2)	
	(DI, 2)	(BHI, 2)	(BEHI, 2)	
	(EH, 2)	(CDE, 4)	(CDEH, 2)	(CDEHI, 2)
	(EI, 2)	(CDH, 2)	(CDEI, 2)	
	(HI, 2)	(CDI, 2)	(CDHI, 2)	
		(CEH, 2)	(CEHI, 2)	
		(CEI, 2)	(DEHI, 2)	
		(CHI, 2)		
		(DEH, 2)		
		(DEI, 2)		
		(DHI, 2)		
		(EHI, 2)		

Last set: (BCDEHI, 2)

Frequent Pattern (FP) Growth Strategy

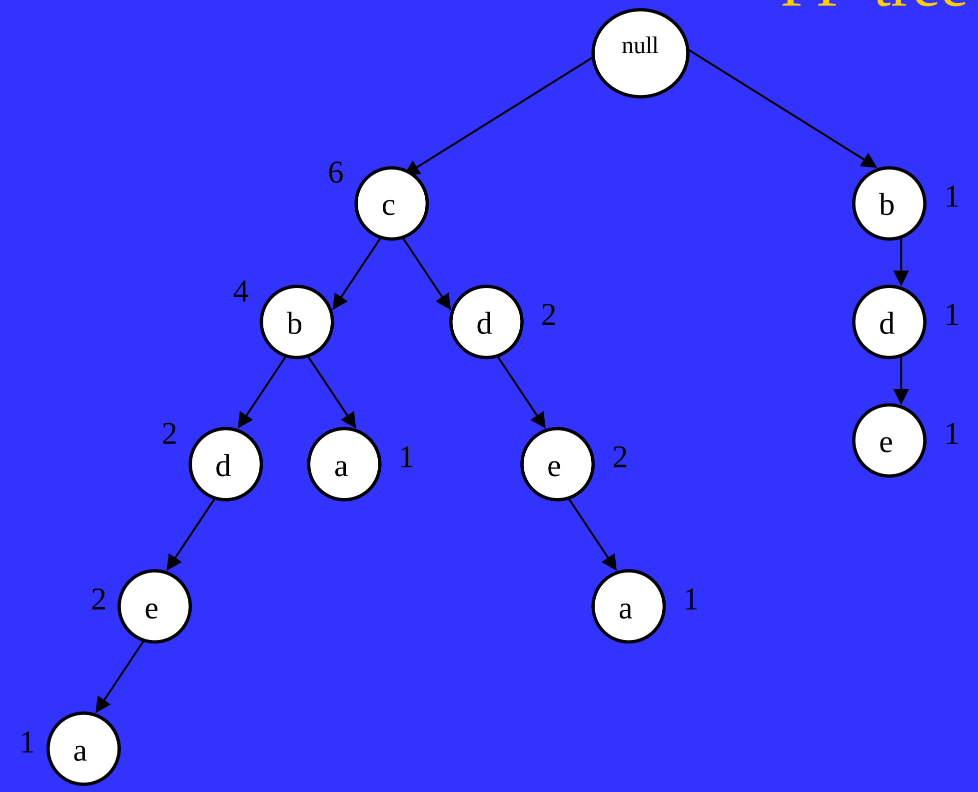
Minimum Support = 2

Transactions:
abcde
abc
acde
bcde
bc
bde
cde

Transactions ordered:
cbdea
cba
cdea
cbde
cb
bde
cde

Frequent Items:
c – 6
b – 5
d – 5
e – 5
a – 3

FP-tree

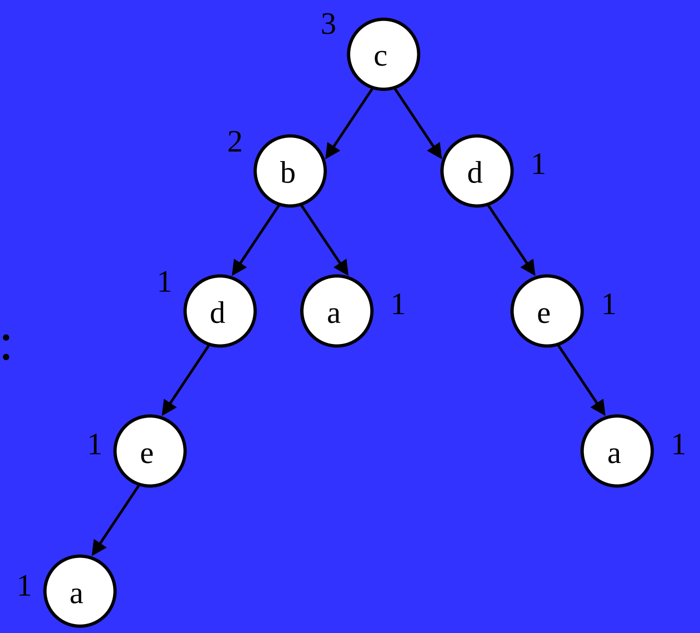


Frequent Pattern (FP) Growth Strategy

Mining the FP-tree for frequent itemsets:
Start from each item and construct a subdatabase of transactions (prefix paths) with that item listed at the end.
Reorder the prefix paths in support descending order. Build a conditional FP-tree.

a – 3
Prefix path:
(c b d e a, 1)
(c b a, 1)
(c d e a, 1)

Correct order:
c – 3
b – 2
d – 2
e – 2



Frequent Pattern (FP) Growth Strategy

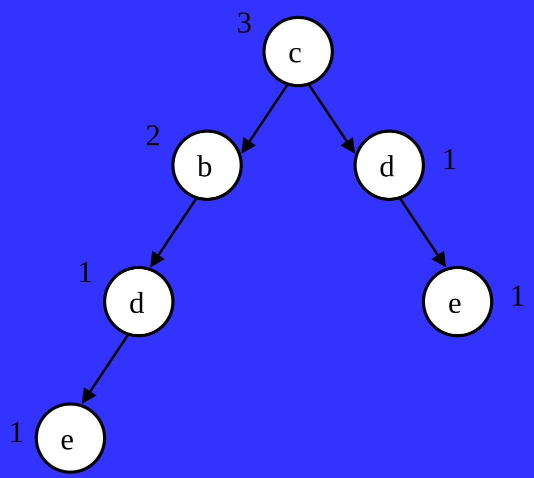
a – 3

Prefix path:

(c b d e a, 1)

(c b a, 1)

(c d e a, 1)



Frequent Itemsets:

(c a, 3)

(c b a, 2)

(c d a, 2)

(c d e a, 2)

(c e a, 2)

Multidimensional AR

Associations between values of different attributes :

CID	nationality	age	income
1	Italian	50	low
2	French	40	high
3	French	30	high
4	Italian	50	medium
5	Italian	45	high
6	French	35	high

RULES:

[**nationality** = French] $\Rightarrow$ [**income** = high] [50%, 100%]

[**income** = high] $\Rightarrow$ [**nationality** = French] [50%, 75%]

[**age** = 50] $\Rightarrow$ [**nationality** = Italian] [33%, 100%]

Single-dimensional AR vs Multi-dimensional

Multi-dimensional

<1, Italian, 50, low>

<2, French, 45, high>

Single-dimensional

<1, {nat/Ita, age/50, inc/low}>

<2, {nat/Fre, age/45, inc/high}>

Schema: <ID, a?, b?, c?, d?>

<1, yes, yes, no, no>

<2, yes, no, yes, no>

<1, {a, b}>

<2, {a, c}>

Quantitative Attributes

- Quantitative attributes (e.g. age, income)
- Categorical attributes (e.g. color of car)

CID	height	weight	income
1	168	75,4	30,5
2	175	80,0	20,3
3	174	70,3	25,8
4	170	65,2	27,0

Problem: too many distinct values

Solution: transform quantitative attributes into categorical ones via discretization.

Discretization of quantitative attributes

- Quantitative attributes are statically discretized by using predefined concept hierarchies:
  - elementary use of background knowledge

  Loose interaction between Apriori and Discretizer

- Quantitative attributes are dynamically discretized
  - into "bins" based on the distribution of the data.
  - considering the distance between data points.

  Tighter interaction between Apriori and Discretizer

Constraint-based AR

- Preprocessing: use constraints to focus on a subset of transactions
  - Example: find association rules where the prices of all items are at most 200 Euro
- Optimizations: use constraints to optimize Apriori algorithm
  - Anti-monotonicity: when a set violates the constraint, so does any of its supersets.
  - Apriori algorithm uses this property for pruning
- Push constraints as deep as possible inside the frequent set computation

Apriori property revisited

- Anti-monotonicity: If a set S violates the constraint, any superset of S violates the constraint.
- Examples:
  - [Price(S) $\leq$ v] is anti-monotone
  - [Price(S) $\geq$ v] is not anti-monotone
  - [Price(S) = v] is partly anti-monotone
- Application:
  - Push [Price(S) $\leq$ 1000] deeply into iterative frequent set computation.

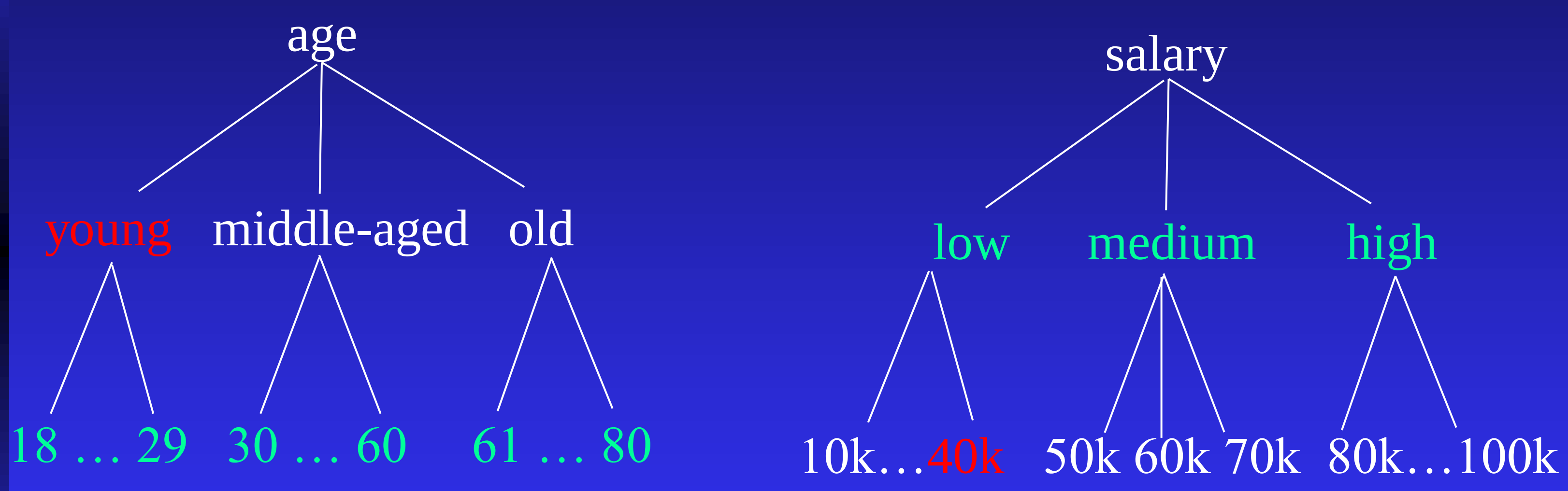
Mining Association Rules with Constraints

- Post processing
  - A naive solution: apply Apriori for finding all frequent sets, and then to test them for constraint satisfaction one by one.
- Optimization
  - Han’s approach: comprehensive analysis of the properties of constraints and try to push them as deeply as possible inside the frequent set computation.

Mining Multilevel AR

Hierarchical attributes: age, salary

Association Rule: (age, young) → (salary, 40k)



Candidate Association Rules:

(age, 18) → (salary, 40k),

(age, young) → (salary, low),

(age, 18) → (salary, low)

Questions?

Thank You