

Day: \_\_\_\_\_

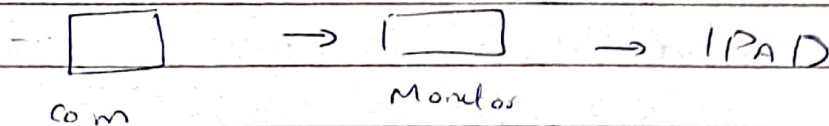
Date: \_\_\_\_\_

## GSP

Data Mining: extract useful knowledge from data

data: discrete sequence

A sequence is an ordered list of symbols



sequences

1 → go → back → home

The goal is to find all subsequences that appear frequently in a set of discrete sequences

1 Items

$I = \{a, b, c, d, e\}$

a → apple, b → bread, c → cake, d → doll

e → egg

An itemset is set of items that is a subset of

$\{a, b, c\}$

$\{d, e\}$

$I \setminus I$	$I$
	An itemset cannot have duplicate items A same item twice

Day: \_\_\_\_\_

Date: \_\_\_\_\_

A discrete sequence is an ordered list of itemsets  $S = \langle X_1, X_2, \dots, X_n \rangle$  where  $X_j \subseteq I$  for any  $j \in \{1, 2, \dots, n\}$

Example

$\langle \{a, b\}, \{c\} \rangle \quad a, b \rightarrow c$

It means customer purchased a, b at same time and then purchased c

$\langle \{a\}, \{a\}, \{c\} \rangle$

$a \rightarrow a \rightarrow c$

subsequence

$\langle \{a, c\} \rangle \subseteq \langle \{a, b, c\} \rangle$

$\langle \{a, c\} \rangle \not\subseteq \langle \{a\}, \{c\} \rangle$

$\langle \{a\}, \{c\} \rangle \subseteq \langle \{a, b\}, \{d\}, \{b, c\} \rangle$

Q

Day: \_\_\_\_\_

Date: \_\_\_\_\_

$D = \{S_1, S_2, S_3, S_4\}$

Input

Sequence Database

$S_1 = \langle \{a, b\}, \{c\}, \{a\} \rangle$

$S_2 = \langle \{a\}, \{b\}, \{c\} \rangle$

$S_3 = \langle \{b\}, \{c\}, \{d\} \rangle$

$S_4 = \langle \{b\}, \{a, b\}, \{c\} \rangle$

$\text{Support}(\langle \{a\} \rangle) = 3$

$\text{sup}(\langle \{b\} \rangle) = 4$

$\text{sup}(\langle \{a\}, \{b\} \rangle) = 1$

$\text{sup}(\langle \{a, b\} \rangle) = 2$

Output

$\text{minsup} = 3$

$\langle \{a\} \rangle$        $\text{sup} = 3$

$\langle \{b\} \rangle$        $\text{sup} = 4$

$\langle \{c\} \rangle$        $\text{sup} = 4$

$\langle \{a\}, \{c\} \rangle$        $\text{sup} = 3$

$\langle \{b\}, \{c\} \rangle$        $\text{sup} = 4$

$\langle \{a, b\}, \{c\} \rangle$        $\text{sup} = 3$



Day: \_\_\_\_\_

Di 100  
20

1 sequence  
1 1 3 2 2 3 3 4 5  
2 sequence  
Date: \_\_\_\_\_

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100

3-sequence  
< {1} {2} {3} >  
< {1} {2 5} >  
< {1} {5} {3} >  
< {2} {3} {4} >  
< {2 5} {3} >  
< {3} {4} {5} >  
< {5} {3 4} >

Merge for  
candidate sequence

< {1} {2 3} {3} >  
< {2} {3} {4 5} >  
< {1} {2 3} {3 4} >

Candidate Generation

< {1} {2} {3} {4} >  
< {1 3} {2 5} {3} >  
< {1} {5} {3 4} >  
< {2} {3} {4} {5} >  
< {2 5} {3 4} >

Candidate Generation

Candidate Pruning

## Candidate Generation

### Frequent 3-sequences

$$\begin{aligned} & \langle \{1\} \{2\} \{3\} \rangle \\ & \langle \{1\} \{2\ 5\} \rangle \\ & \langle \{1\} \{5\} \{3\} \rangle \\ & \langle \{2\} \{3\} \{4\} \rangle \\ & \langle \{2\ 5\} \{3\} \rangle \\ & \langle \{3\} \{4\} \{5\} \rangle \\ & \langle \{5\} \{3\ 4\} \rangle \end{aligned}$$

~~$\langle 1 \rangle \langle 2 \rangle \langle 3 \rangle$~~      ~~$\langle \{1\} \{2\} \{3\} \rangle$~~   
 ~~$\langle \{1\} \{2\} 3 \rangle$~~   
 ~~$\langle \{1\} \{5\} \{3\} \rangle$~~   
 ~~$\langle \{2\} \{3\} \{4\} \rangle$~~  ✓  
 ~~$\langle \{2\} 5 \{3\} \rangle$~~   
 ~~$\langle \{3\} \{4\} \{5\} \rangle$~~   
 ~~$\langle \{5\} \{3\} \{1\} \rangle$~~

~~$\langle \{1\} \{2\} 5 \rangle$~~      ~~$\langle \{1\} \{2\} \{3\} \rangle$~~   
 ~~$\langle \{1\} \{2\} 3 \rangle$~~   
 ~~$\langle \{1\} \{5\} \{3\} \rangle$~~   
 ~~$\langle \{2\} \{3\} \{1\} \rangle$~~   
 ~~$\langle \{2\} 5 \{3\} \rangle$~~  ✓  
 ~~$\langle \{3\} \{4\} \{5\} \rangle$~~   
 ~~$\langle \{5\} \{3\} \{1\} \rangle$~~

~~$\langle 1 \rangle \{5\} \{3\}$~~      ~~$\langle \{1\} \{2\} \{3\}$~~   
 ~~$\langle \{1\} \{2\} 5 \rangle$~~   
 ~~$\langle \{1\} \{5\} \{3\}$~~   
 ~~$\langle \{2\} \{3\} \{1\}$~~   
 ~~$\langle \{2\} 5 \{3\}$~~   
 ~~$\langle \{3\} \{4\} \{5\}$~~   
 ~~$\langle \{5\} \{3\} \{1\}$~~  ✓

~~$\langle 2 \rangle \{3\} \{4\}$~~      ~~$\langle 1 \rangle \{2\} \{3\}$~~   
 ~~$\langle 1 \rangle \{2\} \{5\}$~~   
 ~~$\langle 1 \rangle \{5\} \{7\}$~~   
 ~~$\langle 2 \rangle \{3\} \{4\}$~~   
 ~~$\langle 2 \rangle \{5\} \{7\}$~~   
 ~~$\langle 3 \rangle \{4\} \{5\}$~~  ✓  
 ~~$\langle 5 \rangle \{3\} \{4\}$~~

~~$\langle 2\ 5\ \{3\} \rangle$~~      ~~$\langle \{1\}\ \{2\}\ \{3\} \rangle$~~   
 ~~$\langle \{1\}\ \{2\}\ 3 \rangle$~~   
 ~~$\langle \{1\}\ \{5\}\ \{3\} \rangle$~~   
 ~~$\langle \{2\}\ \{3\}\ \{1\} \rangle$~~   
 ~~$\langle \{2\ 5\}\ \{3\} \rangle$~~   
 ~~$\langle \{3\}\ \{4\}\ \{5\} \rangle$~~   
 ~~$\langle \{5\}\ \{3\}\ 2 \rangle$~~  ✓

~~$\langle 3 \rangle \langle 4 \rangle \langle 5 \rangle$~~      $\langle 1 \rangle \langle 2 \rangle \langle 3 \rangle$   
 $\langle 1 \rangle \langle 2 \rangle \langle 5 \rangle$   
 $\langle 1 \rangle \langle 5 \rangle \langle 3 \rangle$   
 $\langle 2 \rangle \langle 3 \rangle \langle 4 \rangle$   
 $\langle 2 \rangle \langle 5 \rangle \langle 4 \rangle$   
 $\langle 3 \rangle \langle 4 \rangle \langle 5 \rangle$   
 $\langle 5 \rangle \langle 3 \rangle \langle 4 \rangle$

$$\begin{aligned} & \langle \cancel{5} \rangle \langle 3 \ 4 \rangle & \langle \{1\} \{2\} \{3\} \rangle \\ & & \langle \{1\} \{2 \ 3\} \rangle \\ & & \langle \{1\} \{5\} \{4\} \rangle \\ & & \langle \{2\} \{3\} \{4\} \rangle \\ & & \langle \{2 \ 5\} \{3\} \rangle \\ & & \langle \{3\} \{4\} \{5\} \rangle \\ & & \langle \{5\} \{3 \ 4\} \rangle \end{aligned}$$

## Candidate Generation

$$\begin{aligned} & \langle \{1\} \{2\} \{3\} \{4\} \rangle \\ & \langle \{1\} \{2\ 5\} \{3\} \rangle \\ & \langle \{1\} \{5\} \{3\ 4\} \rangle \\ & \langle \{2\} \{3\} \{4\} \{5\} \rangle \\ & \langle \{2\ 5\} \{3\ 4\} \rangle \end{aligned}$$

Day: \_\_\_\_\_

Date: \_\_\_\_\_

## Candidate Pruning

$\langle \{1\} \{2\} \{3\} \{4\} \rangle$

find support

$\langle \{1\} \{2\} \{3\} \{4\} \rangle$

$\langle \{1\} \{2\} \{3\} \{4\} \rangle$

$\langle \{2\} \{3\} \{4\} \rangle$

$\langle \{1\} \{2\} \{3\} \{4\} \rangle$

$\langle \{1\} \{3\} \{4\} \rangle$

$\langle \{1\} \{2\} \{3\} \{4\} \rangle$

$\langle \{1\} \{2\} \{4\} \rangle$

$\langle \{1\} \{2\} \{3\} \{4\} \rangle$

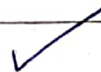
$\langle \{1\} \{3\} \{4\} \rangle$

## Pruning Important

$\{1, 2\} \{4\}$

$\{2\} \{4\} \{6\}$

$\{1, 2\} \{4\} \{6\}$



$(\{1\} \{2\} \{4\} \{6\})$



# Candidate Pruning

## Candidate Generation

$\langle \{1\} \{2\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{2\} \{5\} \{3\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \{5\} \rangle$   
 $\langle \{2\} \{5\} \{3\} \{4\} \rangle$

$\langle \{1\} \{2\} \{3\} \{4\} \rangle$   ~~$\langle \{1\} \{2\} \{3\} \{4\} \rangle$~~   
 ~~$\langle \{1\} \{3\} \{3\} \{4\} \rangle$~~   
 ~~$\langle \{1\} \{2\} \{3\} \{4\} \rangle$~~   
 ~~$\langle \{1\} \{2\} \{3\} \{4\} \rangle$~~

$\langle \{2\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{2\} \{4\} \rangle$   
 $\langle \{1\} \{2\} \{3\} \rangle$

$\langle \{1\} \{2\} \{5\} \{3\} \rangle$   ~~$\langle \{1\} \{2\} \{5\} \{3\} \rangle$~~   
 ~~$\langle \{1\} \{5\} \{3\} \rangle$~~   
 ~~$\langle \{1\} \{2\} \{3\} \rangle$~~   
 ~~$\langle \{1\} \{2\} \{5\} \{3\} \rangle$~~

$\langle \{2\} \{5\} \{3\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \rangle$   
 $\langle \{1\} \{2\} \{3\} \rangle$   
 $\langle \{1\} \{2\} \{5\} \rangle$

$\langle \{1\} \{5\} \{3\} \{4\} \rangle$   ~~$\langle \{1\} \{5\} \{3\} \{4\} \rangle$~~   
 ~~$\langle \{1\} \{5\} \{3\} \{4\} \rangle$~~   
 ~~$\langle \{1\} \{5\} \{3\} \{4\} \rangle$~~   
 ~~$\langle \{1\} \{5\} \{3\} \{4\} \rangle$~~

$\langle \{5\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{5\} \{4\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \rangle$

$\langle \{2\} \{3\} \{4\} \{5\} \rangle$   ~~$\langle \{2\} \{3\} \{4\} \{5\} \rangle$~~   
 ~~$\langle \{2\} \{3\} \{4\} \{5\} \rangle$~~   
 ~~$\langle \{2\} \{3\} \{4\} \{5\} \rangle$~~   
 ~~$\langle \{2\} \{3\} \{4\} \{5\} \rangle$~~

$\langle \{3\} \{4\} \{5\} \rangle$   
 $\langle \{2\} \{4\} \{5\} \rangle$   
 $\langle \{2\} \{3\} \{5\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \rangle$

$\langle \{2\} \{5\} \{3\} \{4\} \rangle$   ~~$\langle \{2\} \{5\} \{3\} \{4\} \rangle$~~   
 ~~$\langle \{2\} \{5\} \{3\} \{4\} \rangle$~~   
 ~~$\langle \{2\} \{5\} \{3\} \{4\} \rangle$~~   
 ~~$\langle \{2\} \{5\} \{3\} \{4\} \rangle$~~

$\langle \{5\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{5\} \{4\} \rangle$   
 $\langle \{2\} \{5\} \{3\} \rangle$

# Candidate Pruning

## Candidate Generation

$\langle \{1\} \{2\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{2\} \{5\} \{3\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \{5\} \rangle$   
 $\langle \{2\} \{5\} \{3\} \{4\} \rangle$

$\langle \{1\} \{2\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{2\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{2\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{2\} \{3\} \{4\} \rangle$

$\langle \{1\} \{2\} \{5\} \{3\} \rangle$   
 $\langle \{1\} \{2\} \{5\} \{3\} \rangle$   
 $\langle \{1\} \{2\} \{5\} \{3\} \rangle$   
 $\langle \{1\} \{2\} \{5\} \{3\} \rangle$

$\langle \{1\} \{5\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \{4\} \rangle$

$\langle \{2\} \{3\} \{4\} \{5\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \{5\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \{5\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \{5\} \rangle$

$\langle \{2\} \{5\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{5\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{5\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{5\} \{3\} \{4\} \rangle$

$\langle \{2\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{2\} \{4\} \rangle$   
 $\langle \{1\} \{2\} \{3\} \rangle$

$\langle \{2\} \{5\} \{3\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \rangle$   
 $\langle \{1\} \{2\} \{3\} \rangle$   
 $\langle \{1\} \{2\} \{5\} \rangle$

$\langle \{5\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{5\} \{4\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \rangle$

$\langle \{3\} \{4\} \{5\} \rangle$   
 $\langle \{2\} \{4\} \{5\} \rangle$   
 $\langle \{2\} \{3\} \{5\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \rangle$

$\langle \{5\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{5\} \{4\} \rangle$   
 $\langle \{2\} \{5\} \{3\} \rangle$

## Frequent 3-sequences

$\langle \{1\} \{2\} \{3\} \rangle$   
 $\langle \{1\} \{2\} \{5\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{5\} \{3\} \rangle$   
 $\langle \{3\} \{4\} \{5\} \rangle$   
 $\langle \{5\} \{3\} \{4\} \rangle$



# Candidate Pruning

## Candidate Generation

$\langle \{1\} \{2\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{2\} \{5\} \{3\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \{5\} \rangle$   
 $\langle \{2\} \{5\} \{3\} \{4\} \rangle$

$\langle \{1\} \{2\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{2\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{2\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{2\} \{3\} \{4\} \rangle$

$\langle \{2\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{2\} \{4\} \rangle$   
 $\langle \{1\} \{2\} \{3\} \rangle$

$\langle \{1\} \{2\} \{5\} \{3\} \rangle$   
 $\langle \{1\} \{2\} \{5\} \{3\} \rangle$   
 $\langle \{1\} \{2\} \{5\} \{3\} \rangle$   
 $\langle \{1\} \{2\} \{5\} \{3\} \rangle$

$\langle \{2\} \{5\} \{3\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \rangle$   
 $\langle \{1\} \{2\} \{3\} \rangle$   
 $\langle \{1\} \{2\} \{5\} \rangle$

$\langle \{1\} \{5\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \{4\} \rangle$

$\langle \{5\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{3\} \{4\} \rangle$   
 $\langle \{1\} \{5\} \{4\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \rangle$

$\langle \{2\} \{3\} \{4\} \{5\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \{5\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \{5\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \{5\} \rangle$

$\langle \{3\} \{4\} \{5\} \rangle$   
 $\langle \{2\} \{4\} \{5\} \rangle$   
 $\langle \{2\} \{3\} \{5\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \rangle$

$\langle \{2\} \{5\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{5\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{5\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{5\} \{3\} \{4\} \rangle$

$\langle \{5\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{5\} \{4\} \rangle$   
 $\langle \{2\} \{5\} \{3\} \rangle$

## Frequent 3-sequences

$\langle \{1\} \{2\} \{3\} \rangle$   
 $\langle \{1\} \{2\} \{5\} \rangle$   
 $\langle \{1\} \{5\} \{3\} \rangle$   
 $\langle \{2\} \{3\} \{4\} \rangle$   
 $\langle \{2\} \{5\} \{3\} \rangle$   
 $\langle \{3\} \{4\} \{5\} \rangle$   
 $\langle \{5\} \{3\} \{4\} \rangle$

## Candidate Pruning

$\langle \{1\} \{2\} \{5\} \{3\} \rangle$