Day:________ Date:________

GSP

Data Mining : extract useful knowledge from data

data. discrete sequence

A sequence is an ordered list of symbols

□ → □ → IPAD

com Monitor

sequences

I → go → back → home

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

Items

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

a → apple, b: bread, c: cake d: doles
e = eggs

An itemset is set of items that is a subset of I

IMP

{a, b, c} {d, e}

An itemset cannot have duplicates items

a same item two

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 $j \in \{1, 2, n\}$

Example

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

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

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

$a \rightarrow a \rightarrow c$

subsequence

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

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

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

Day:______ Date:______

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

Input Sequence Database

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

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

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

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

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

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

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

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

Output

minsup = 3

$\langle \{a\} \rangle$	sup = 3
$\langle \{b\} \rangle$	sup = 4
$\langle \{c\} \rangle$	sup = 4
$\langle \{a\}, \{c\} \rangle$	sup = 3
$\langle \{b\}, \{c\} \rangle$	sup = 4
$\langle \{a,b\}, \{c\} \rangle$	sup = 3

Scanned with CamScanner

$\frac{25}{\cancel{100}}$ ~~x 5~~
20

1sequence
{1} {2} {3} {4} {5}

Day: 2sequence Date:

25 <{1} {2}> , { . . . }

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 genera

< {1} {2} {3} >

< {2} {3} {4} >

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

Candidate Generalhm

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

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

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

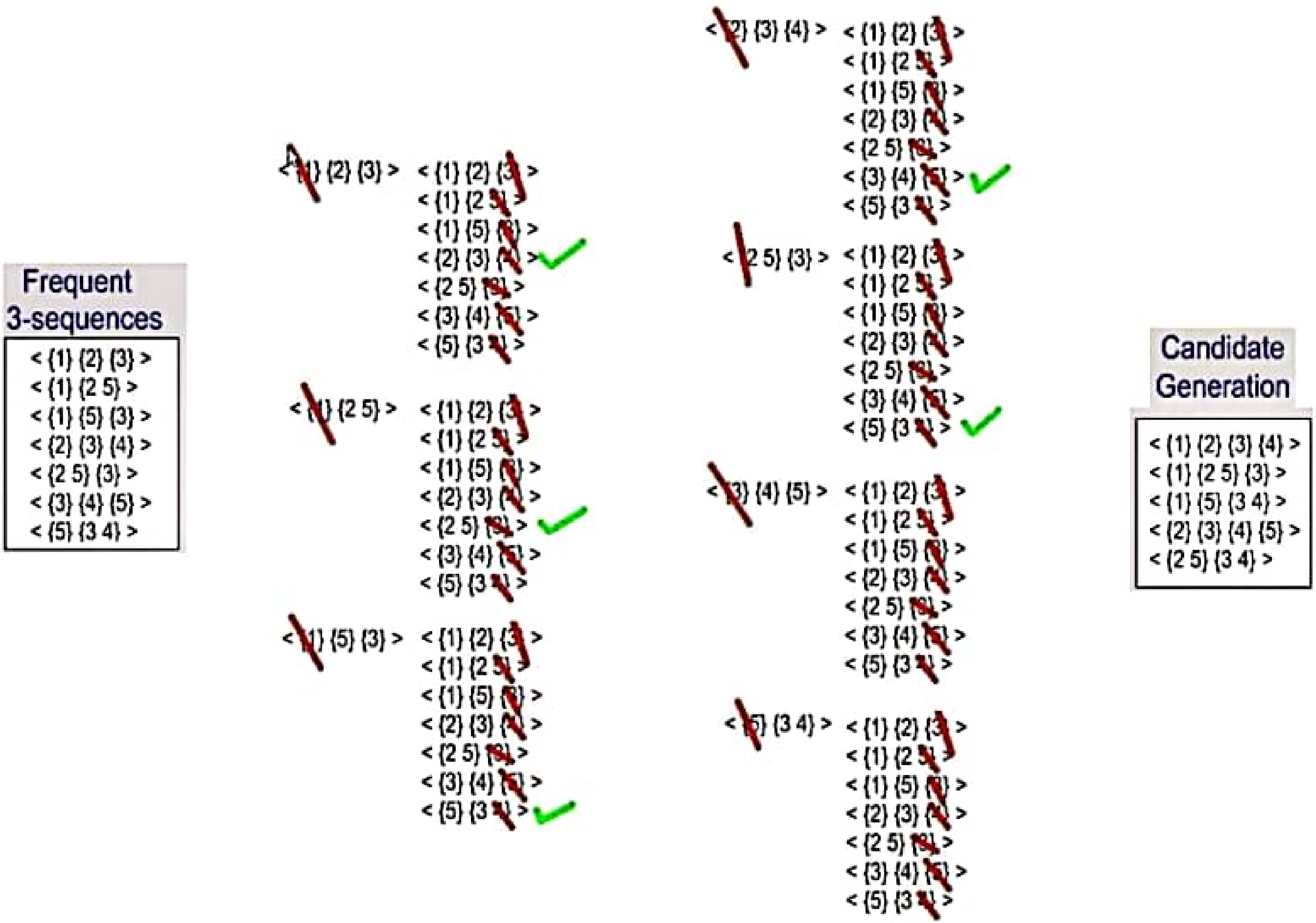
<{2} {3} {4} {5} >

< {2 5} {3 4} >

Candidate Generation

Candidate Pruning

Candidate Generation



Frequent 3-sequences

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

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

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

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

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

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

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

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

Candidate Generation

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

Day:________ Date:________

Candidate Pruning

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

Fwd support

< {1} {2} {3} {45} >

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

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

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

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

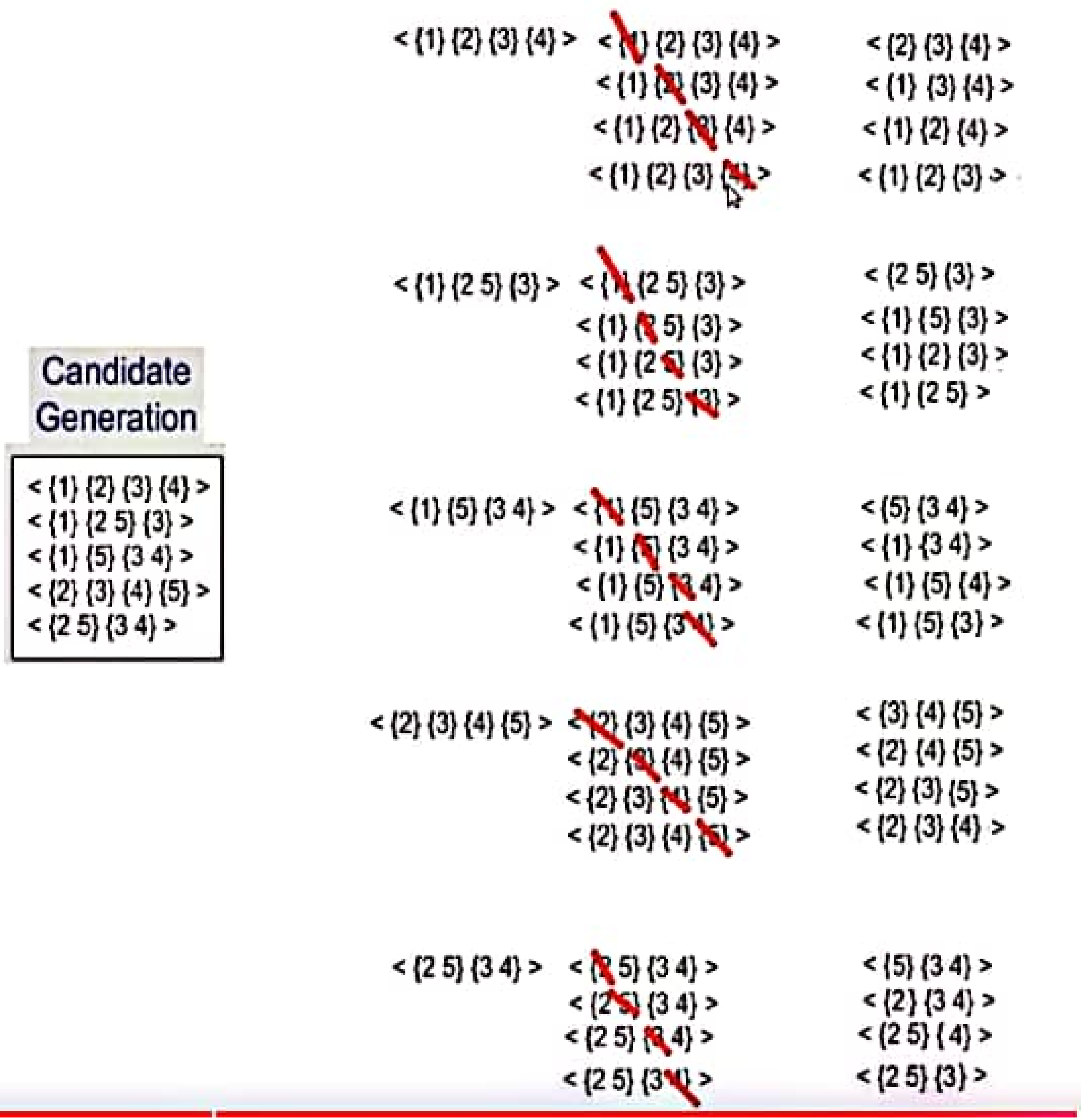
Prunning Imppoint

{~~1~~, 2} {4}

{2} {4} {~~6~~}

{1, 2} {4} {6} ✓ (X {1} {2} {4} {6})

Candidate Pruning



Candidate Generation

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

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

Candidate Pruning

Candidate Generation

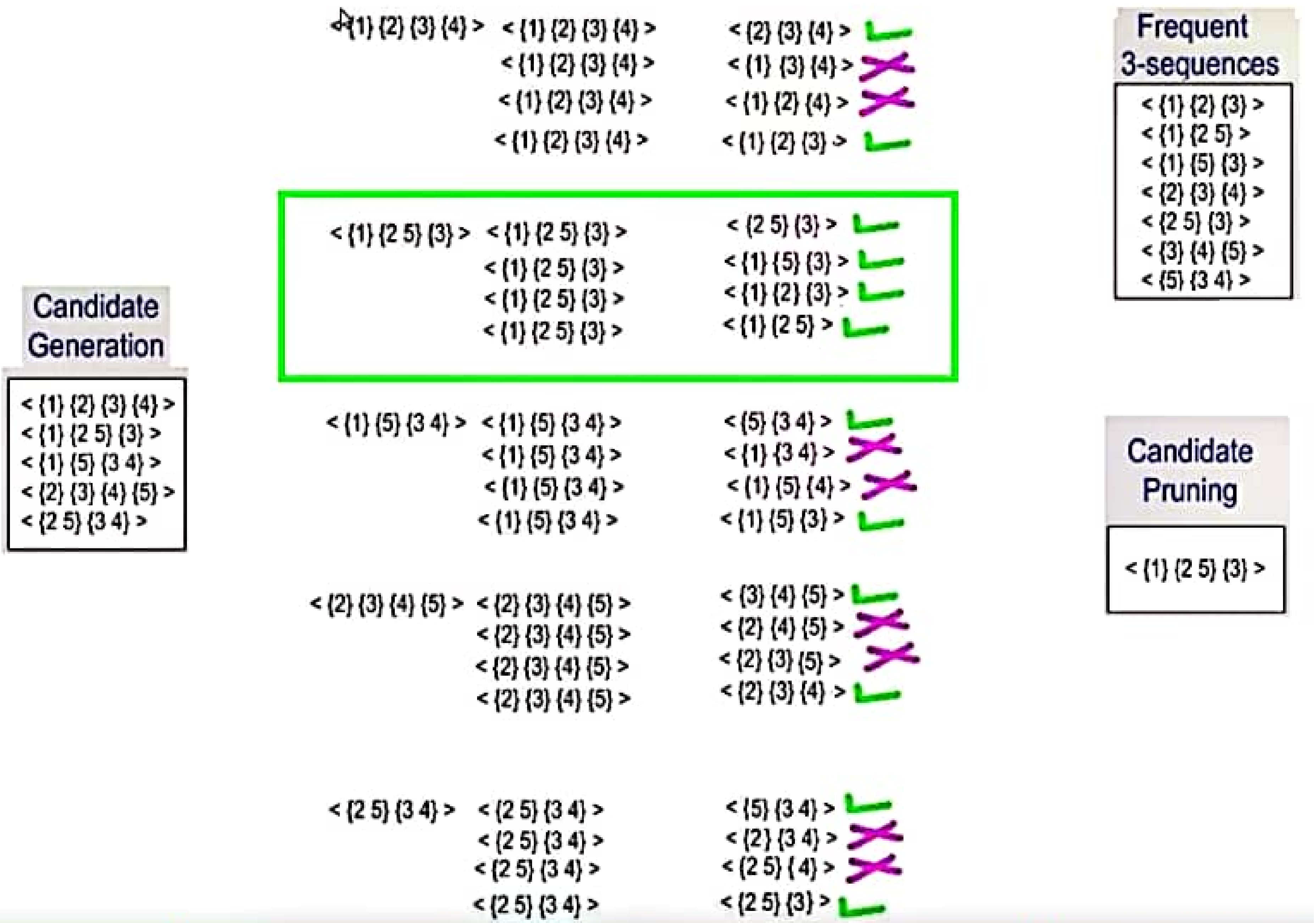
< {1} {2} {3} {4} >
< {1} {2 5} {3} >
< {1} {5} {3 4} >
< {2} {3} {4} {5} >
< {2 5} {3 4} >

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

Frequent 3-sequences

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

Candidate Pruning



Frequent 3-sequences

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

Candidate Generation

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

< {1} {2} {3} {4} >
< {1} {2} {3} {4} > < {2} {3} {4} > ✓
< {1} {2} {3} {4} > < {1} {3} {4} > ✗
< {1} {2} {3} {4} > < {1} {2} {4} > ✗
< {1} {2} {3} {4} > < {1} {2} {3} > ✓

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

< {1} {5} {3 4} >
< {1} {5} {3 4} > < {5} {3 4} > ✓
< {1} {5} {3 4} > < {1} {3 4} > ✗
< {1} {5} {3 4} > < {1} {5} {4} > ✗
< {1} {5} {3 4} > < {1} {5} {3} > ✓

< {2} {3} {4} {5} >
< {2} {3} {4} {5} > < {3} {4} {5} > ✓
< {2} {3} {4} {5} > < {2} {4} {5} > ✗
< {2} {3} {4} {5} > < {2} {3} {5} > ✗
< {2} {3} {4} {5} > < {2} {3} {4} > ✓

< {2 5} {3 4} >
< {2 5} {3 4} > < {5} {3 4} > ✓
< {2 5} {3 4} > < {2} {3 4} > ✗
< {2 5} {3 4} > < {2 5} {4} > ✗
< {2 5} {3 4} > < {2 5} {3} > ✓

Candidate Pruning

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