Lecture 7

Monday, September 23, 2024 10:02 AM

Categorical data

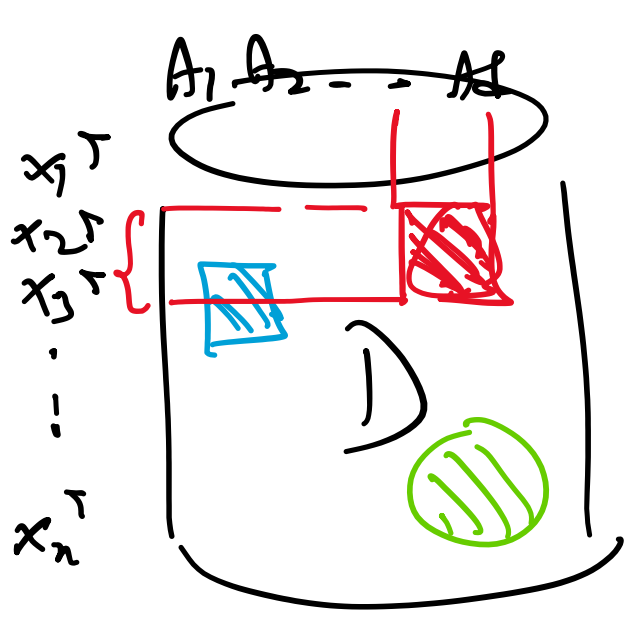
- ↳ frequent patterns
- ↳ Association Rule

$X \Rightarrow Y$

X implies Y, (Confidence, frequency)

Patterns

- ↳ itemset "set" of co occurring events (unordered)
- ↳ sequence → temporal / → positional (ordered)
- ↳ graphs (subgraphs that are 'Common')



→ model for the whole data

↳ local model for a subset of the points and a subset of the attributes

---

Itemset mining

frequently occurring sets of 'items'

$I = \{$ set of possible items $\}$

$\equiv \{A, B, C, D, E\}$

$T = \{$ set of instances / transactions $\}$

$= \{1, 2, \ldots, n\}$ — ids

$X \subseteq I$

↳ itemset is simply some subset of I

$X = AB = \{A, B\}$

$X = ACE = \{A, C, E\}$

$Y \subseteq T$ is called a tidset (transaction id set)

D — binary data matrix

	A	B	C	D	E	
1	1	1	0	1	1	→ ABDE
2	0	1	1	0	1	→ BCE
3	1	1	0	1	1	→ ABDE
4	1	1	1	0	1	→ ABCE
5	1	1	1	1	1	→ ABCDE
6	0	1	1	1	0	→ BCD
	1345	123456	2456	1356	12345	

Sparse matrix

nnz # of non-zero entries

$nnz \ll |I| \times |T|$

Task: Submatrix of all ones

$X \times Y$

such that

$|Y| \geq minsup$

then we also call X a frequent itemset

all tids in Y support X (contain all items in X)

X: (BCE) × Y: (245)

	B	C	E
2	1	1	1
4	1	1	1
5	1	1	1

BCE is therefore a frequent pattern

minsup ≡ Minimum support

eg. 3

Combinatorial enumeration task!

find all such frequent patterns X, that occur ≥ minsup times

I : $2^{|I|}$ possible subsets / itemsets

$I = \{A, B, C, D, E\}$

2^5 possible subsets ≡ powerset

- 0 — ∅ — empty pattern (always frequent)
- 1 — A, B, C, D, E
- 2 — AB, AC, AD, AE, BC, BD, BE, CD, CE, DE
- 3 — ABC, ABD, ABE - -
- 4 — ABCD, ABCE, . .
- 5 — ABCDE

then count; find all the tids that support each itemset

Apriori Algorithm! level-wise / BFS

D

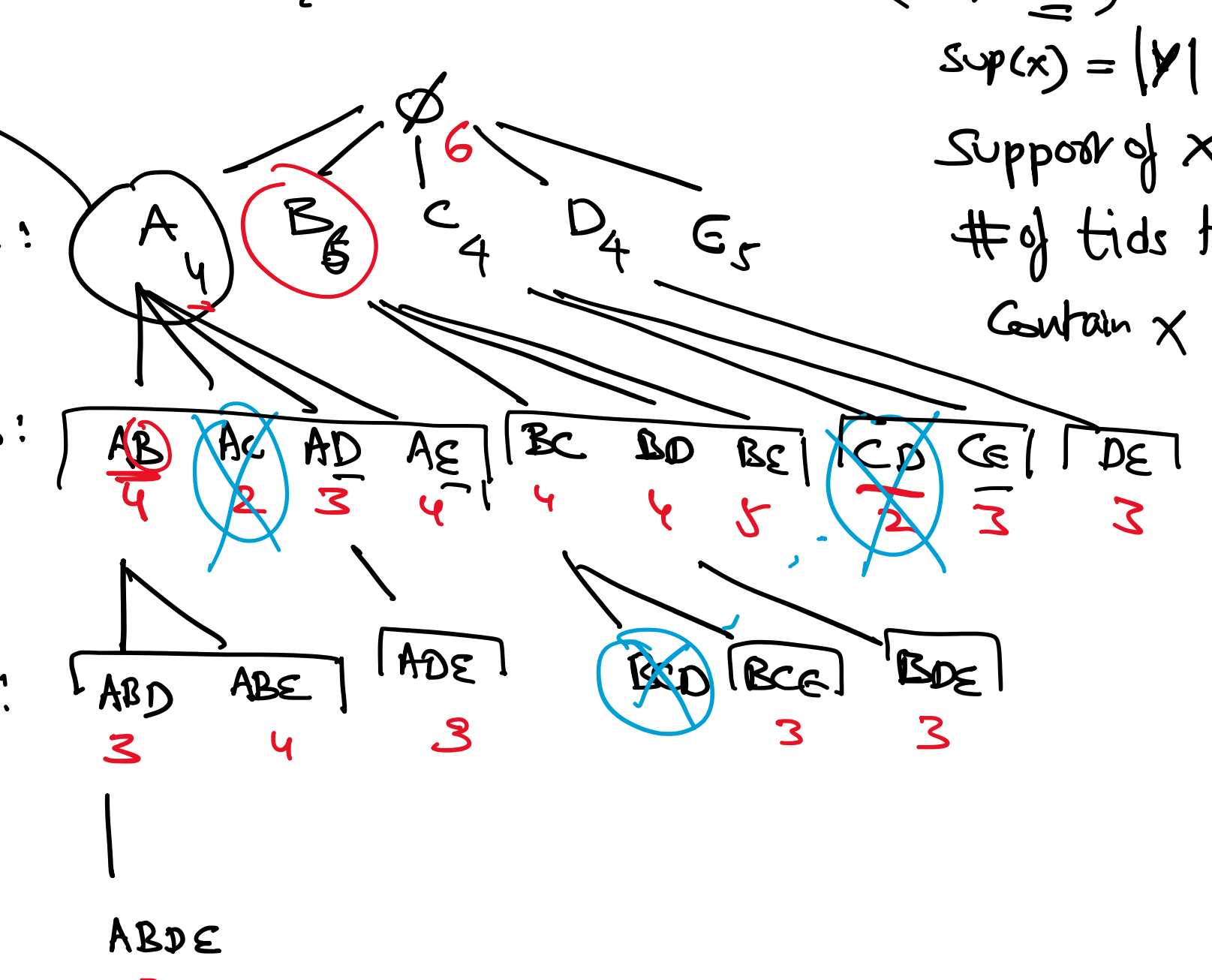
1.	ABDE
2.	BCE
3.	ABDE
4.	ABCE
5.	ABCDE
6.	BCD

minsup = 3

$D = 10^6 \times 10^3$

$|I| \times |T|$

Huge dataset



$(X \times Y)$

$sup(X) = |Y|$

Support of X ≡ # of tids that contain X

$F = \{A, B, C, D, E, AB, AD, \ldots - ABDE\}$

Collection of frequent patterns.

---

1) itemset search space combinatorial } BFS vs. DFS

2) counting ← data organization (indexing)

Eclat (Equivalence Class transformation)

new db format

A	B	C	D	E
1	1	2	1	1
3	2	4	3	2
4	3	5	5	3
5	4	6	6	4
	5			5
	6			

Inverted index

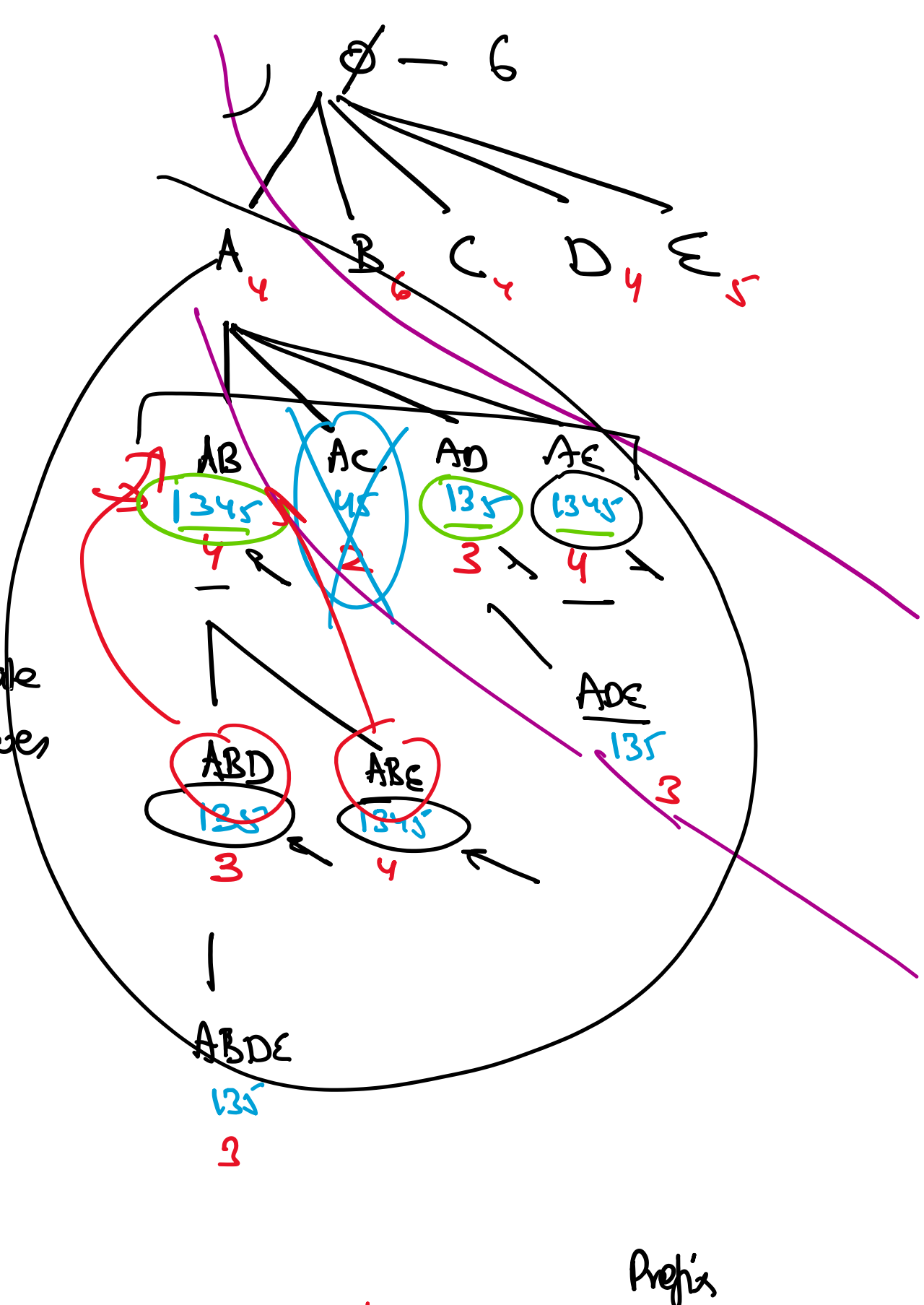
$t(A)$ = tidset for A

$t(X)$ = tidset for itemset X

$sup(X) = |t(X)|$

$t(AB) = t(A) \cap t(B)$

Con: intermediate storage goes up



X = ABE

$t(X)$ = tidset of X

$= \{1, 3, 4, 5\} = 1345$

$diffset(X) = t(\text{prefix of } X) - t(X)$

difference of tidsets

= tids that contain prefix but not the last item in X

Prefix

$d(ABD) = t(AB) - t(ABD)$ (X = ABD)

$= \{1345\} - \{135\}$

$= \{4\}$

---

dEclat: Eclat on diffsets only

1) level 1: Use tidsets to create diffsets

2) for other levels use only diffsets! (propagate only differences)

A	B	C	D	E
1345	123456	2456	1356	12345

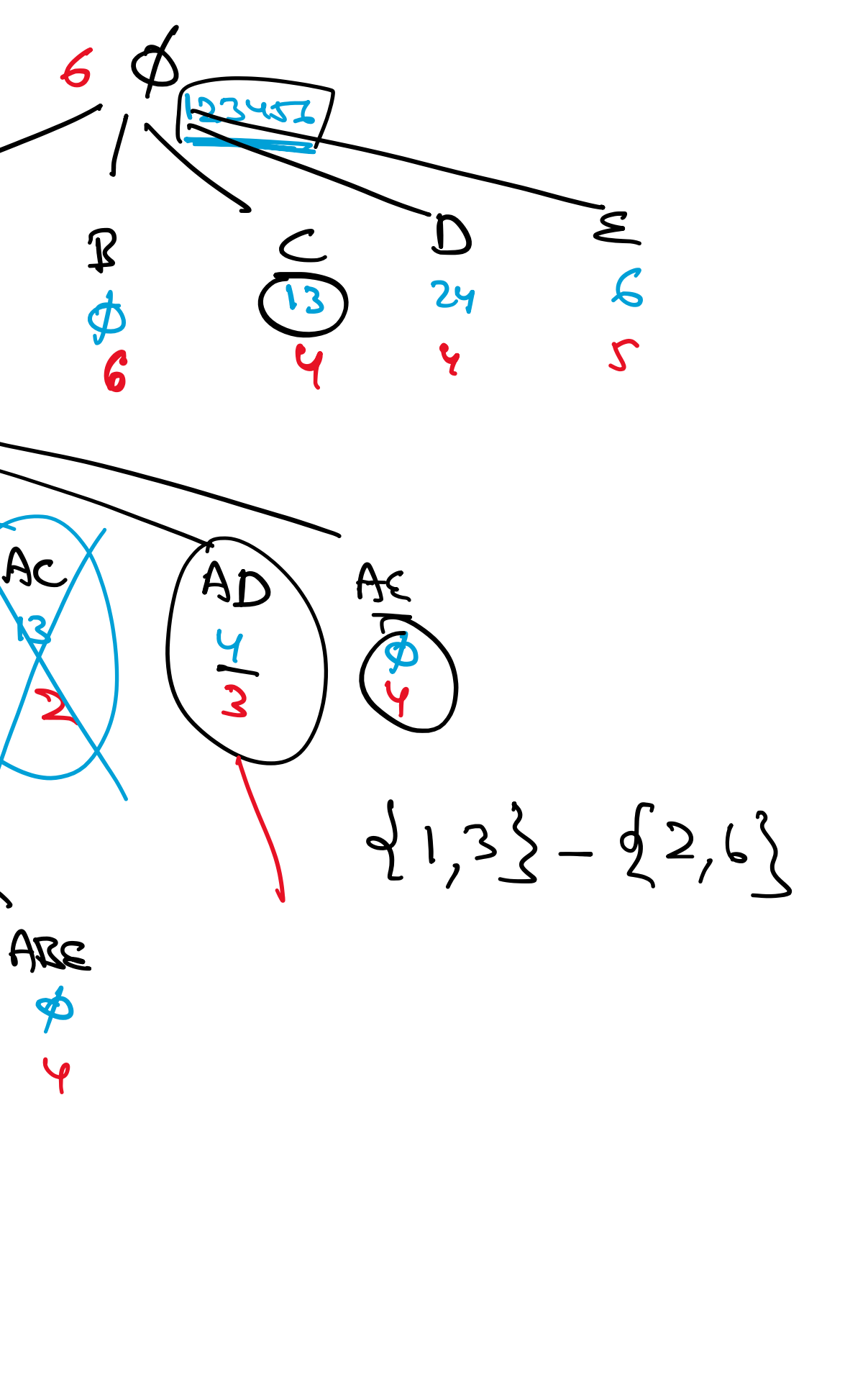
dataset

diffsets →

$sup(X) = sup(\text{prefix}) - len(\text{diffset})$

$d(AB) = d(B) - d(A)$

$d(AD) = d(D) - d(A)$



$\{1, 3\} - \{2, 6\}$