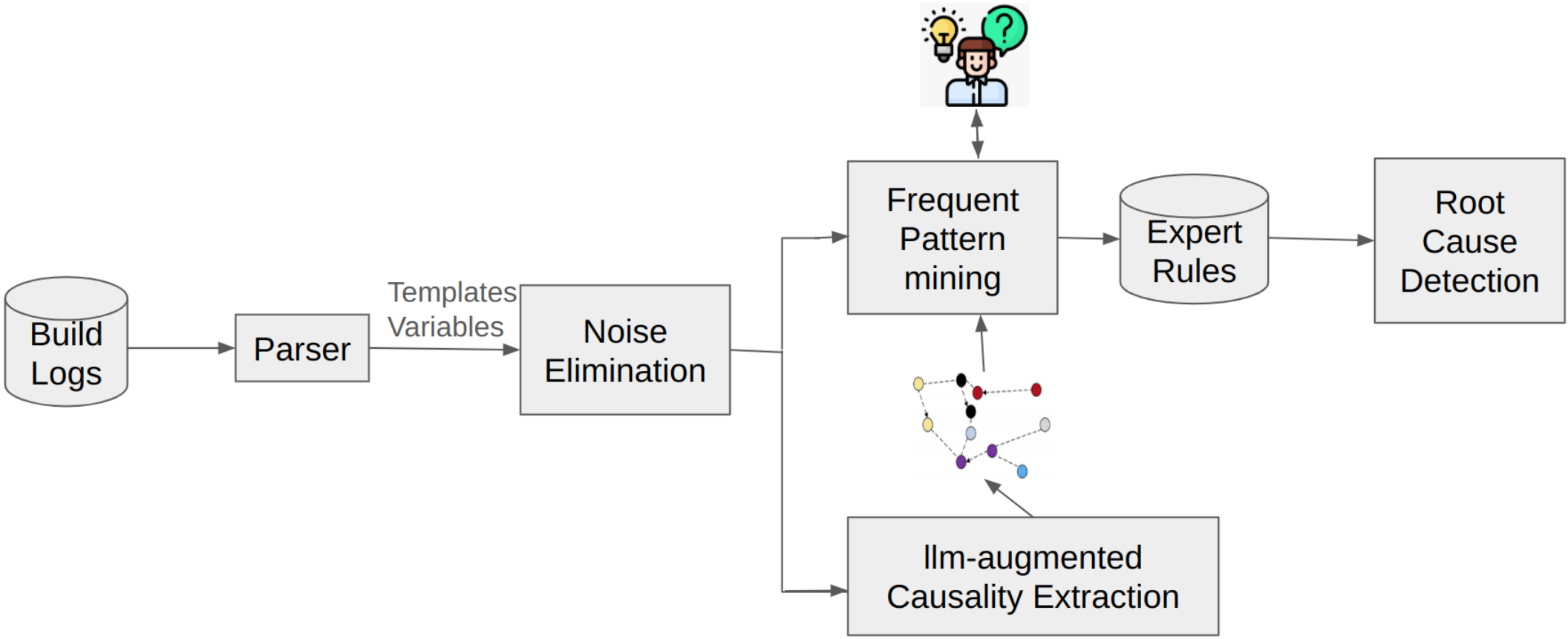
Log Grouping and Causality Analysis

Fateme Faraji Daneshgar

Polytechnique Montréal

DORSAL Laboratory

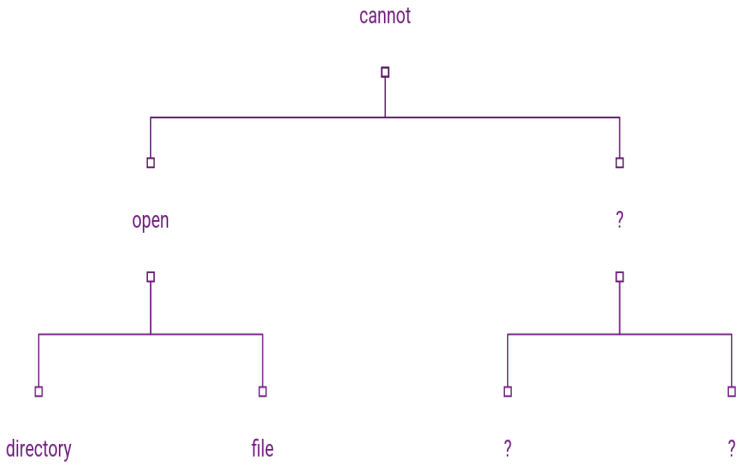
Root Cause Detection pipeline



Parser

- Regular expression
  - Cannot open directory X

- Drain approach
  - Clustering logs using a prefix tree
  - Incremental prefix tree



Cannot open directory x
Cannot open directory y
Cannot open directory z

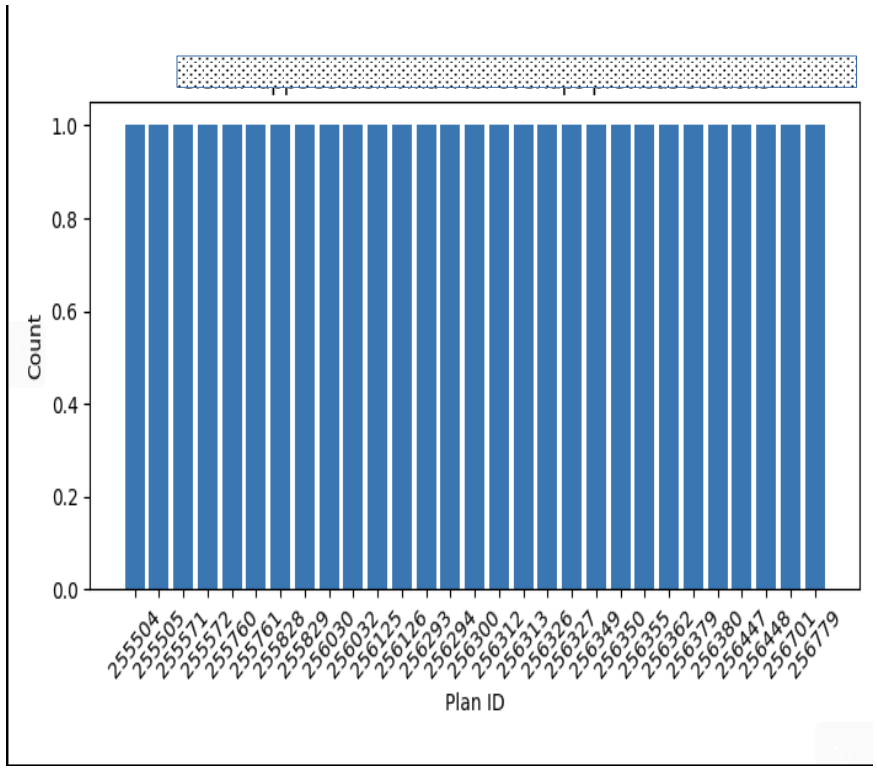
Cannot open directory x

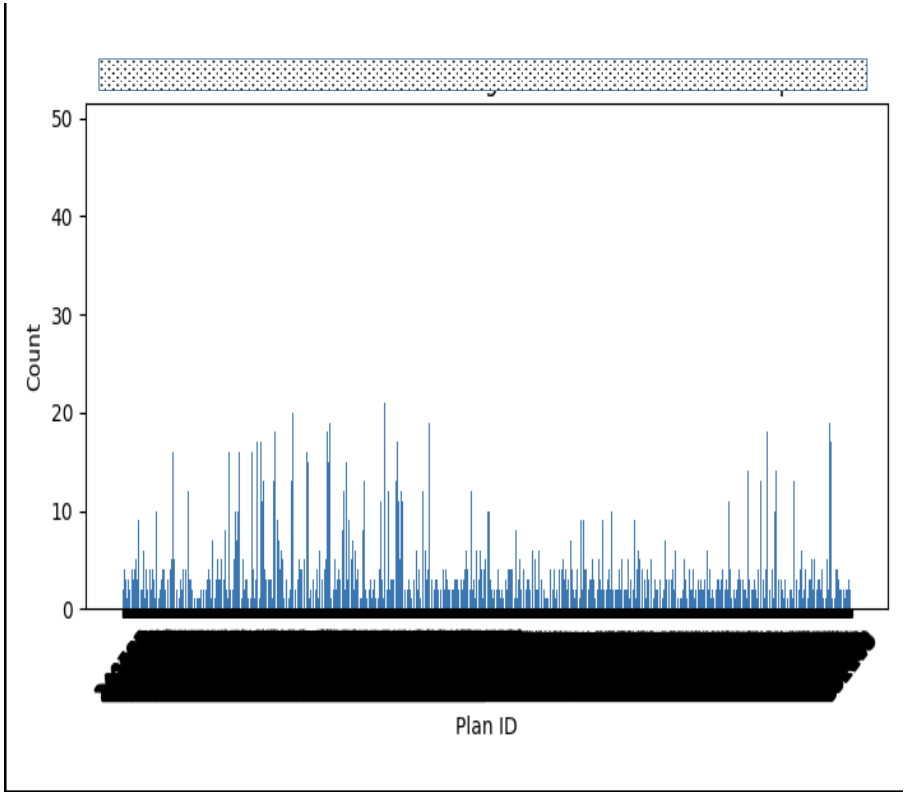
Parser Results

- Same Templates
- Different Templates
  - Improvement (61.5%)
  - Errors(2.8%)
  - Low frequency and invariant variables (34.8%)

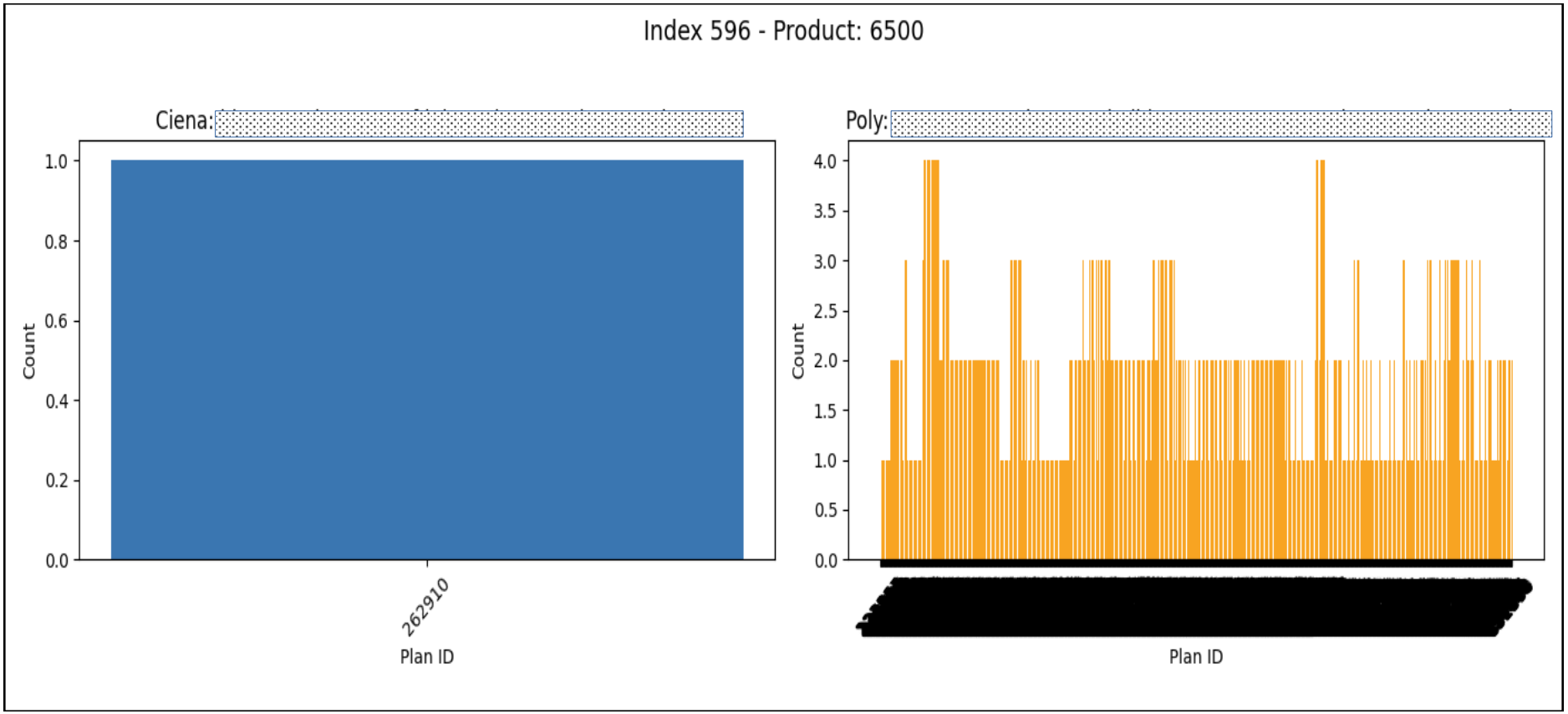
Noise Elimination

- Criterion
  - Fluctuation of the frequency of log templates

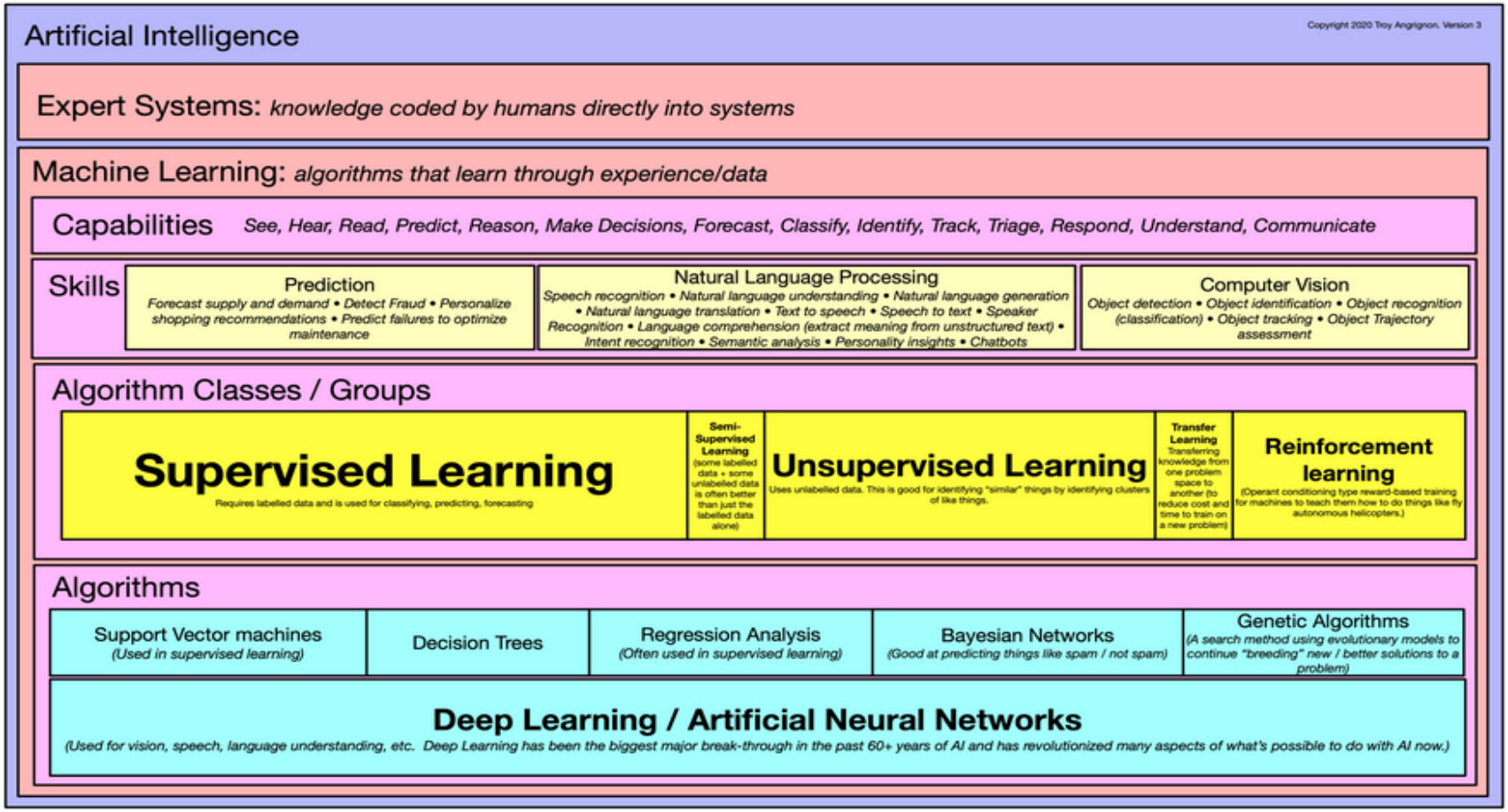




Noise Elimination

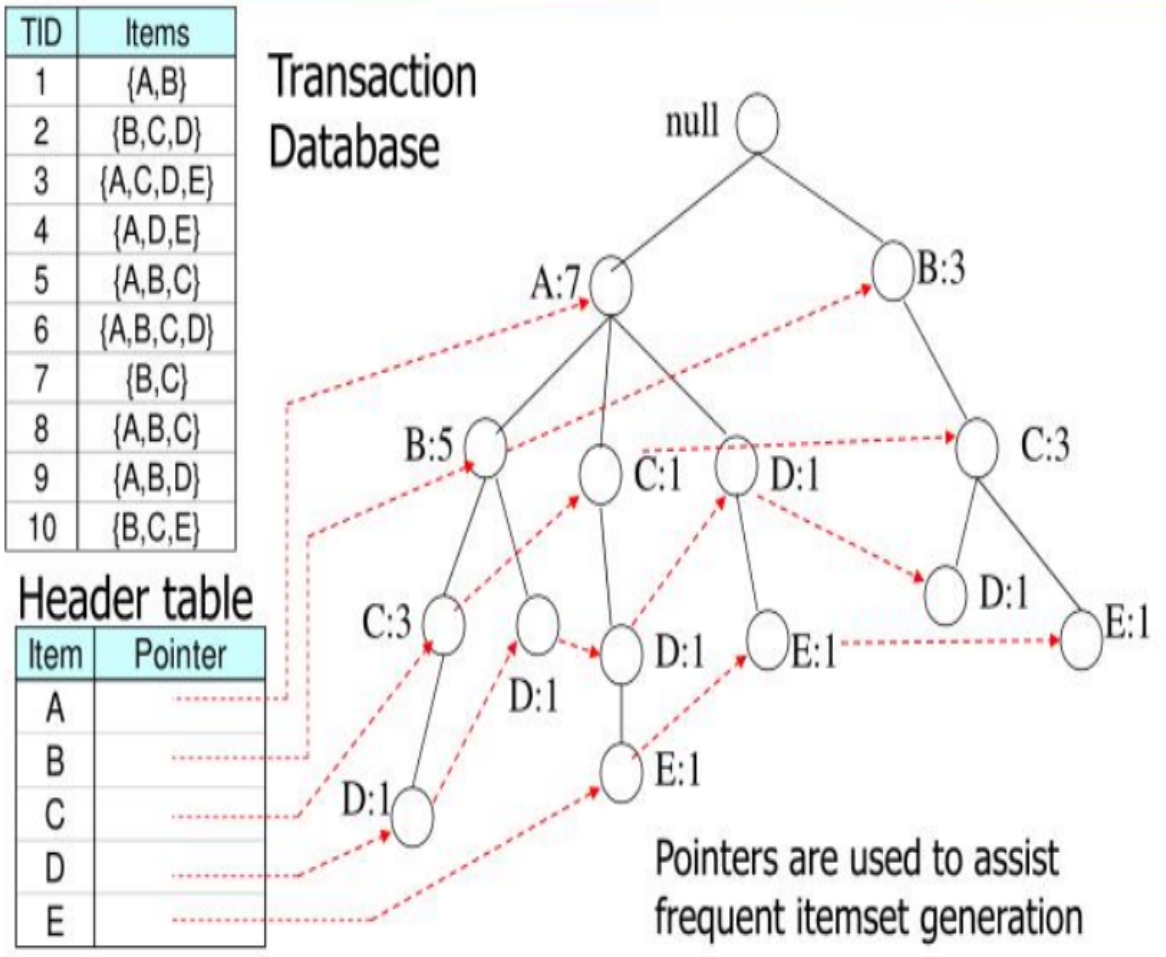


Expert Knowledge Extraction



Frequent Pattern mining

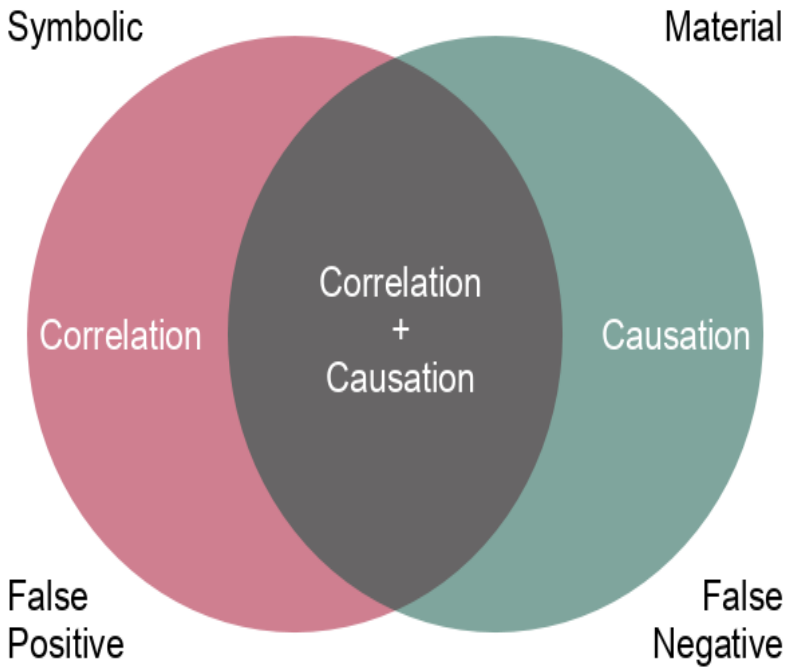
- Time consuming
- Candidate generation
  - Frequent Pattern
  - Conditional Pattern Growth



Causality Analysis

- Conditional Independence
  - Correlation does not imply Causality
    - PC, FCI, RFCI, …

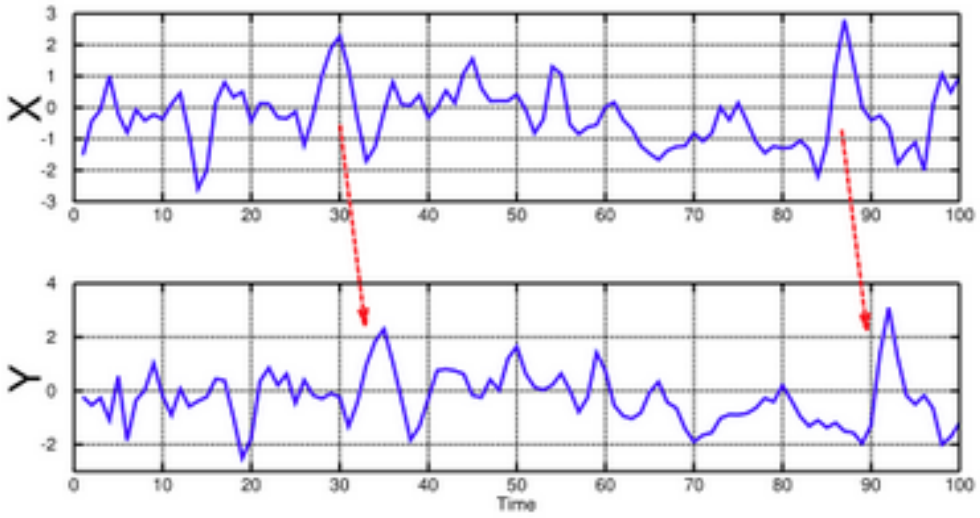
- $A \perp B|C \iff P(A|B,C)=P(A|B)$
- Variable A can predict Variable B



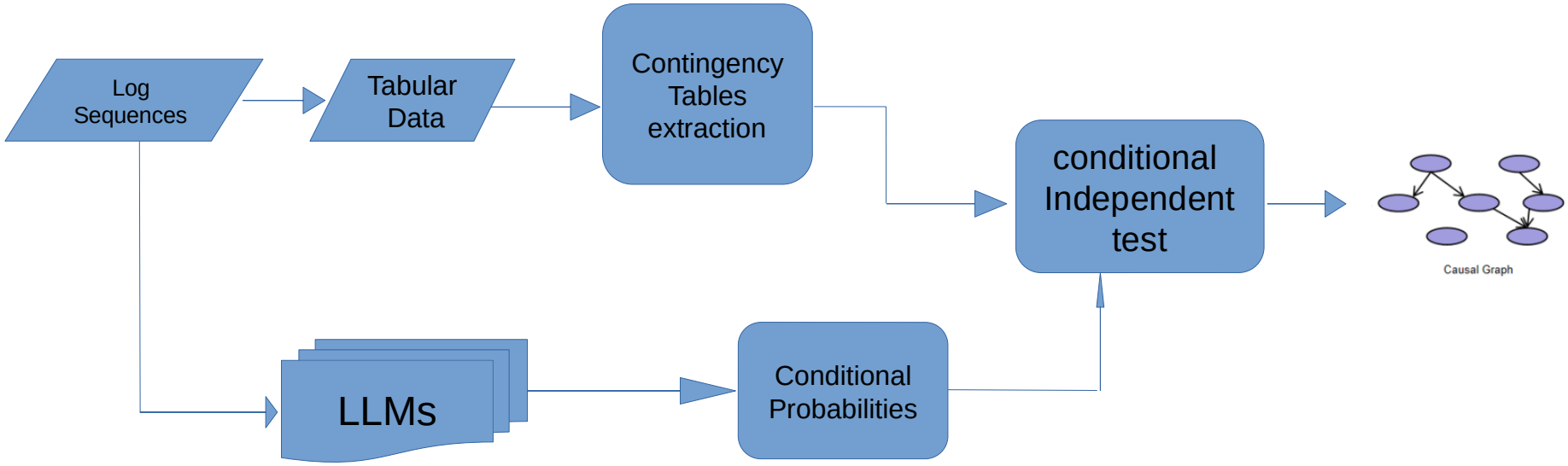
Constraint of traditiona CE for Log Analysis

- Tabular data
  - Missing the order
- Number of variables
- Time-series based
  - Sparse data
- Large language models

X	Y	C_0	C_1
0	0	0	0*
0	0	0	0*
0	0	0	0*
0	0	0	0*
1	1	1*	1
1	1	1*	1
1	1	1*	1
1	1	1*	1



LLM-Augmented PC



Contingency Table

- Log Sequences :
  - {AB,CA,BC,ACB,A}
- Conditional Independence test
  - A and B given C
- $$E_{i,j} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$
- Chi-square test

A	B	C
1	1	0
1	0	1
0	1	1
0	0	0
1	1	1
1	0	0

Contingency Table for C = 0:

A \ B	0	1	Total
0	1	1	2
1	1	1	2
Total	2	2	4

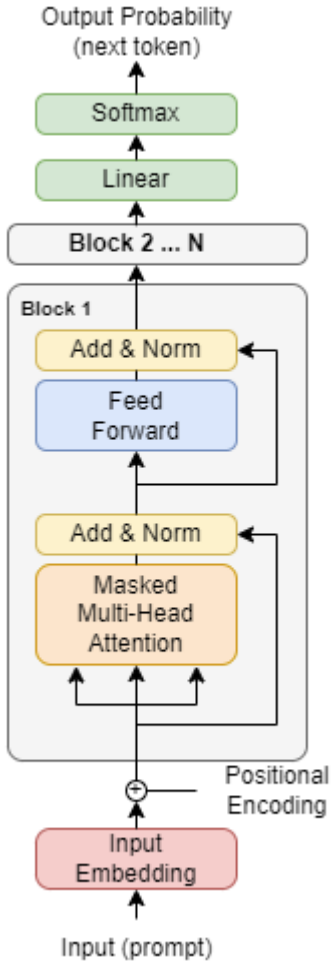
Contingency Table for C = 1:

A \ B	0	1	Total
0	1	0	1
1	1	1	2
Total	2	1	3

LLM-Augmented PC

- Decoder-only transformer
- Each log template is considered as a token
- Continual pre-train
- P(X|Y) and P(X|Y,C):
- For each sequence having Y (Y and C):
  - probability of X for all positions after Y

    Suppose si=Y S1 S2 S3 … Si Si+1 … Sn



LLM-Augmented PC

- Traditional PC
  - log template seq: S1 S2 S3 … Si Si+1 … Sn
  - 1 if we have X (before or after) and 0 otherwise

- LLM_Augmented approach:
  - log template seq: S1 S2 S3 … Si Si+1 … Sn
  - How probable is having or not having X after Y

Thanks!

fateme-faraji.daneshgar@polymtl.ca