Unsupervised Root-Cause Analysis for Integrated Systems

Renjian Pan[1], Zhaobo Zhang[2], Xin Li[1], Krishnendu Chakrabarty[1] and Xinli Gu[2]

[1] Department of ECE, Duke University

[2] Futurewei Technologies, Inc.

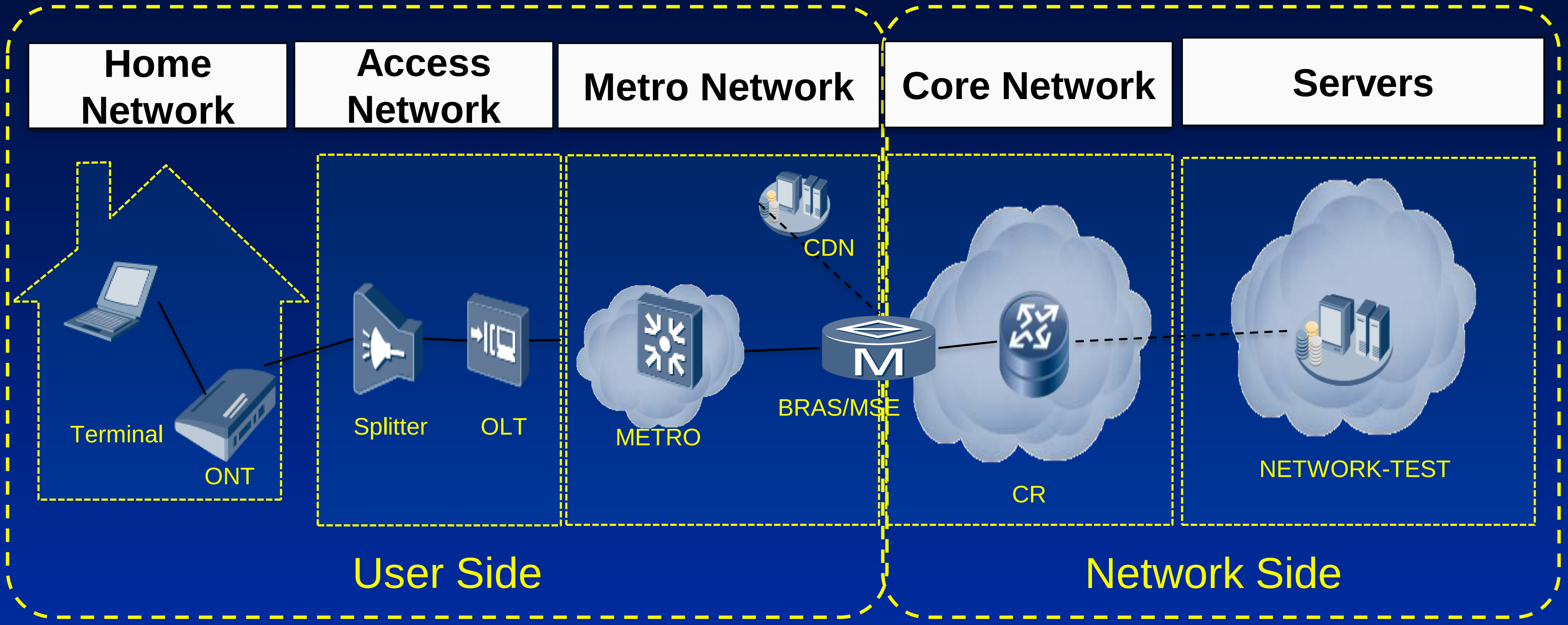
Purpose

- Analyze the root causes for integrated systems with an unsupervised machine-learning method

Outline

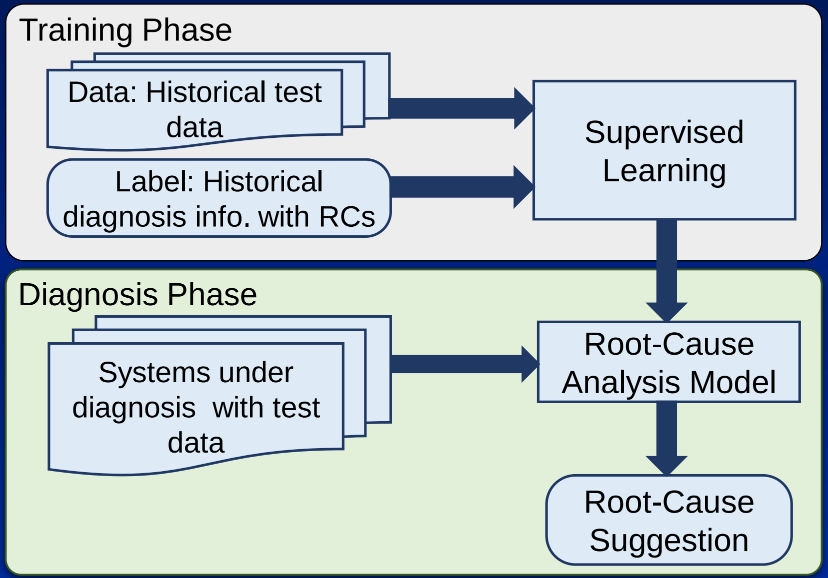
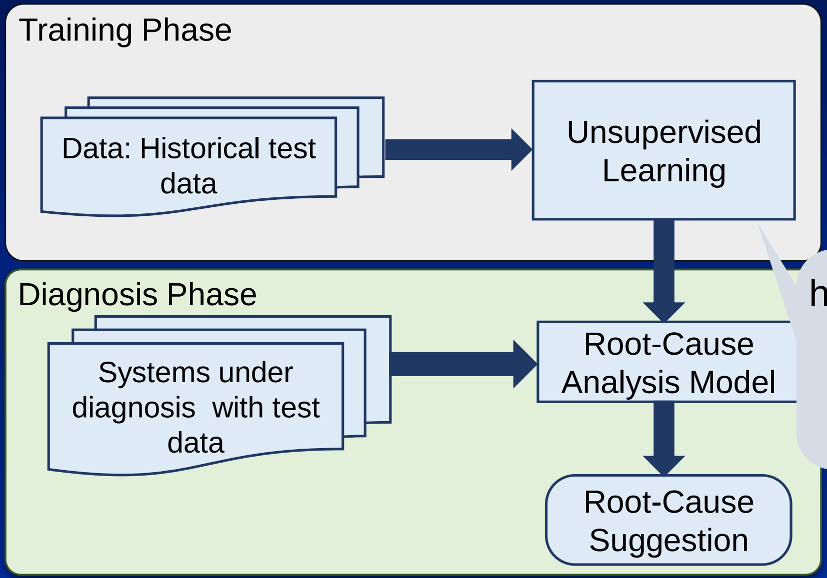
- Background
- Problem Formulation
- Proposed Method
- Experimental Results
- Conclusions

Motivation



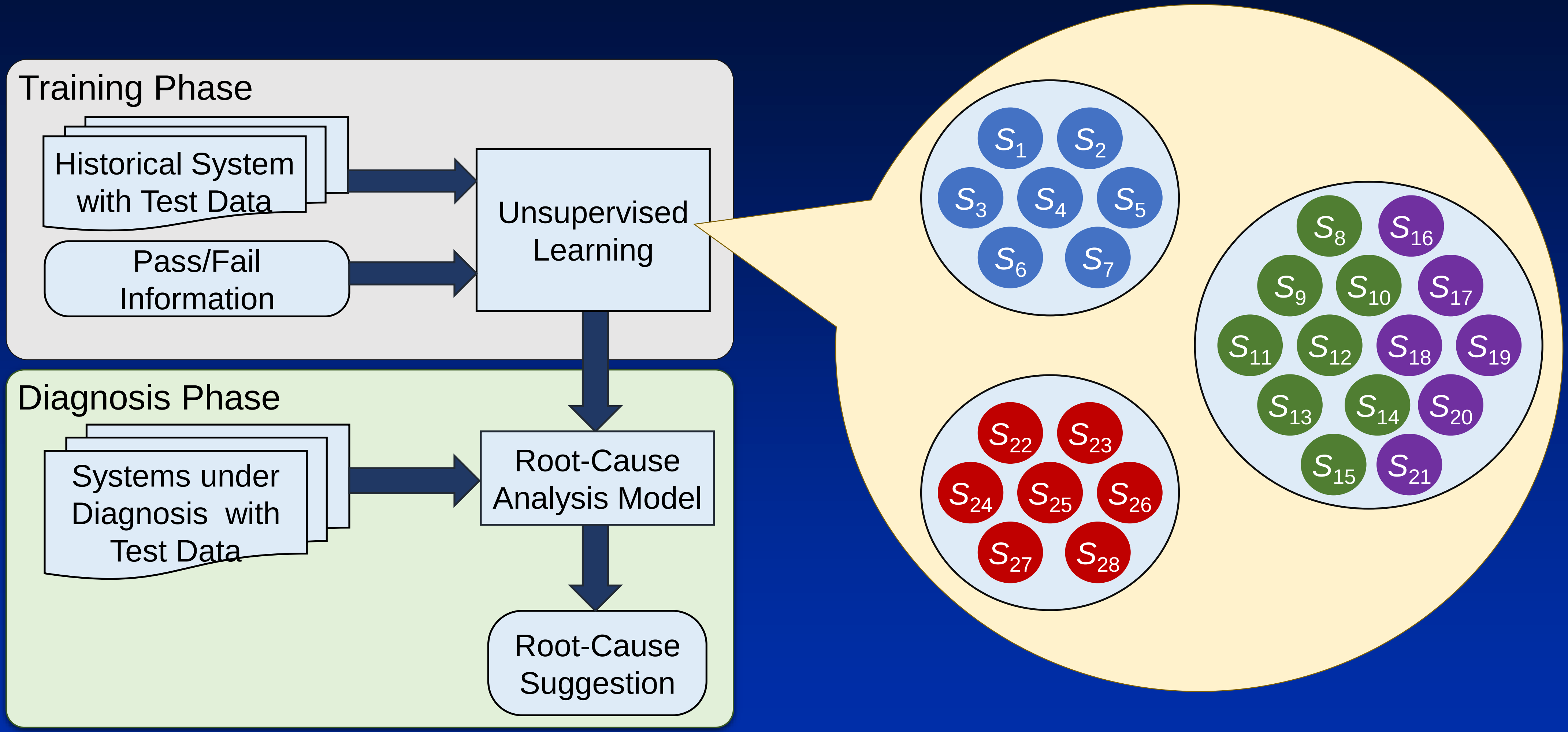
- **Intelligently locate the root cause**

Review of Root-Cause Analysis

	Supervised Methods	Unsupervised Methods
Example	Classification models (SVM, random forest, etc.)	Clustering models (K-means, DBSCAN, etc.)
Framework		
PROS	High accuracy	No need of historical diagnosis information
CONS	Need historical diagnosis information with root cause from experts	Relatively low accuracy

have not used the pass/fail info. which contains great value

Problem Formulation



Evaluation Metric

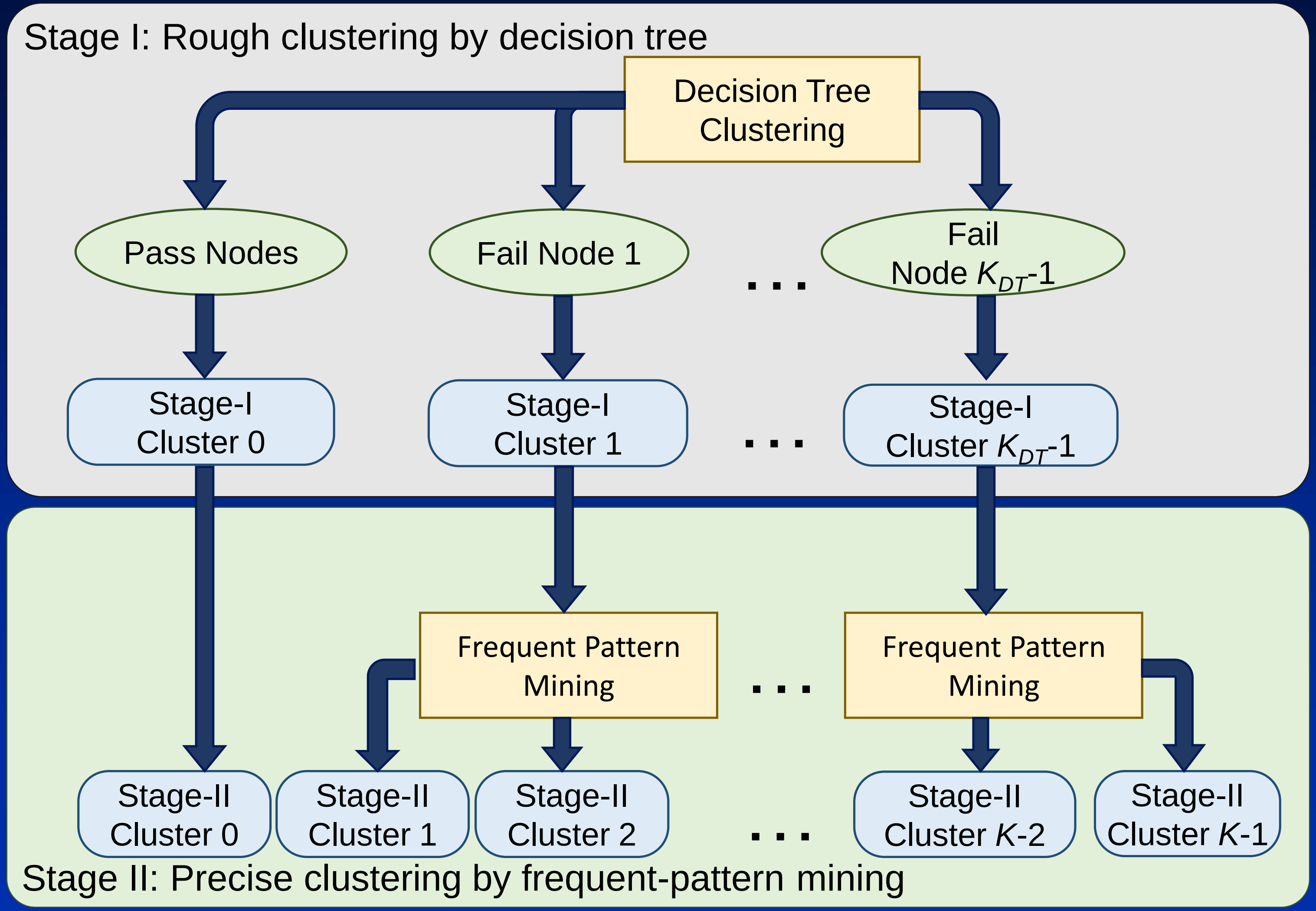
- Normalized Mutual Information (NMI)

$$\mathrm{I}_{NMI}(r,c) = \frac{\mathrm{I}_{MI}(r,c)}{(\mathrm{H}(r)+\mathrm{H}(c))/2}$$

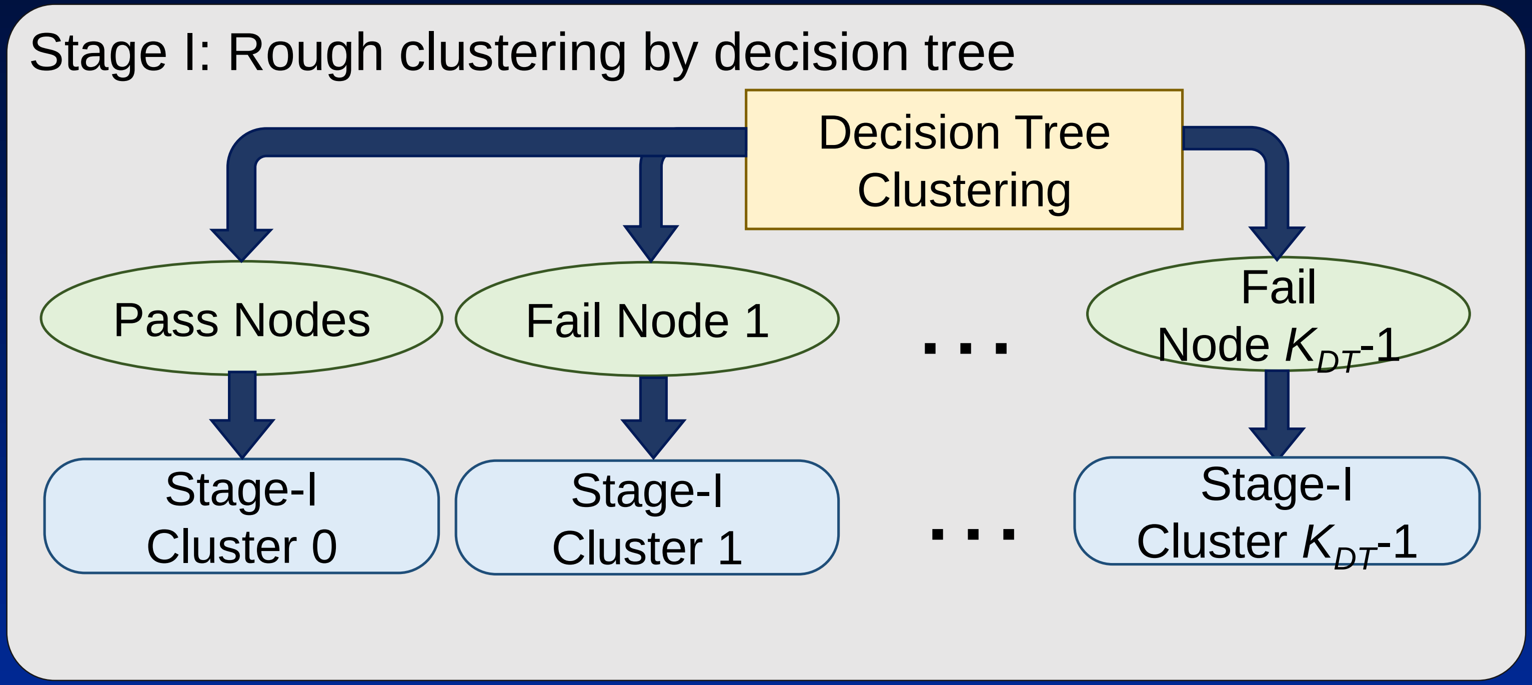
  - r: correct root causes c: clustering result
  - $\mathrm{I}_{MI}(\cdot)$: mutual information $\mathrm{H}(\cdot)$: entropy

Clusters	**NMI**
[0,0],[1,1],[2,2],[3,3]	1.0
[0,1,2,3], [0,1,2,3]	0.0
[0,0,1,1], [2,2,3,3]	0.67
[0,0,1,3], [2,2,1,3]	0.33

Overall Flow of the Two-Stage Clustering



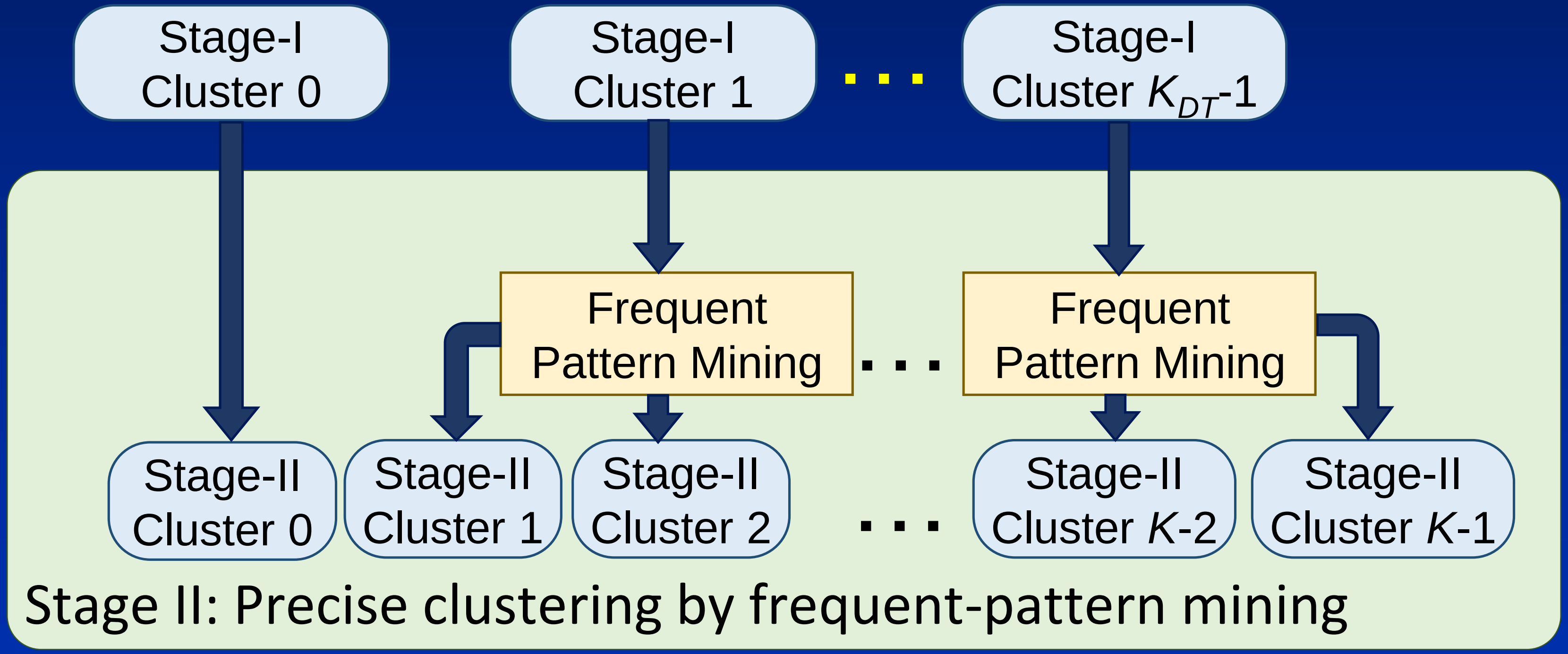
Stage-I: Rough Clustering by Decision Tree



- **Learning:** Build a decision tree with pass/fail info. as the label
  - **Fail node**: fail-labeled data as majority
  - **Pass node**: pass-labeled data as majority
- **Clustering:** Data in pass nodes form a pass cluster (Stage-I Cluster **0**)
  - Data in each fail node forms a stage-I cluster

Stage-II: Precise Clustering with Frequent-Pattern Mining

- **Initialization**: Discretize numerical data
- **Learning**: Mine frequent patterns to extract important info. on each stage-I cluster
- **Clustering**: Split each stage-I cluster based on the frequent patterns for data samples

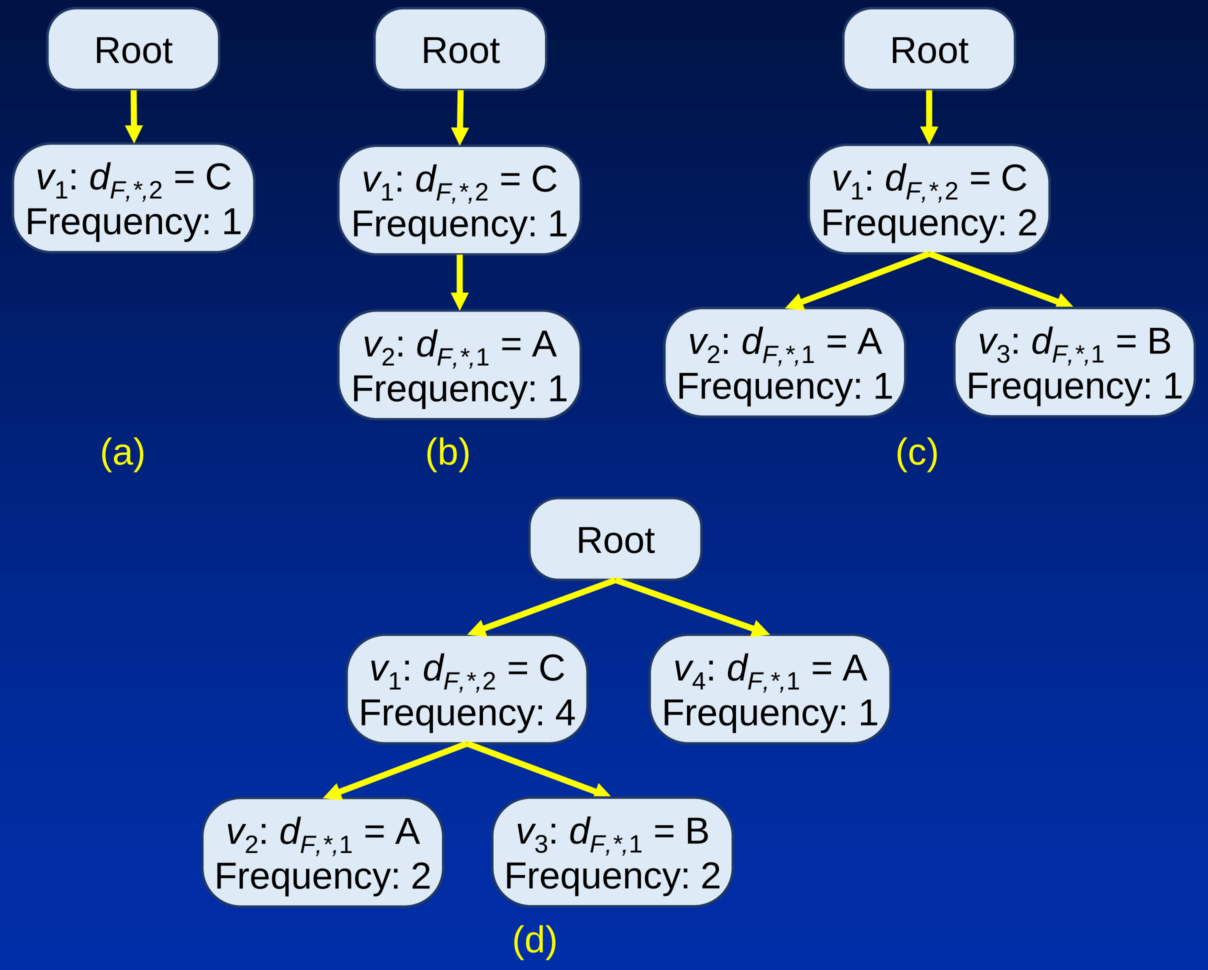


Frequent-Pattern Mining

Index	$d_{F,*,1}$	$d_{F,*,2}$	**Frequent Items**
1	A	C	[$d_{F,*,2}$ = C, $d_{F,*,1}$ = A]
2	B	C	[$d_{F,*,2}$ = C, $d_{F,*,1}$ = B]
3	B	C	[$d_{F,*,2}$ = C, $d_{F,*,1}$ = B]
4	A	C	[$d_{F,*,2}$ = C, $d_{F,*,1}$ = A]
5	A	D	[$d_{F,*,1}$ = A]

Order: [$d_{F,*,2}$ = C, $d_{F,*,1}$ = A , $d_{F,*,1}$ = B]

Frequent Patterns	**Frequency**
{$d_{F,*,2}$ = C}	4
{$d_{F,*,1}$ = A}	3
{$d_{F,*,1}$ = B}	2
{$d_{F,*,2}$ = C, $d_{F,*,1}$ = A}	2
{$d_{F,*,2}$ = C, $d_{F,*,1}$ = B}	2



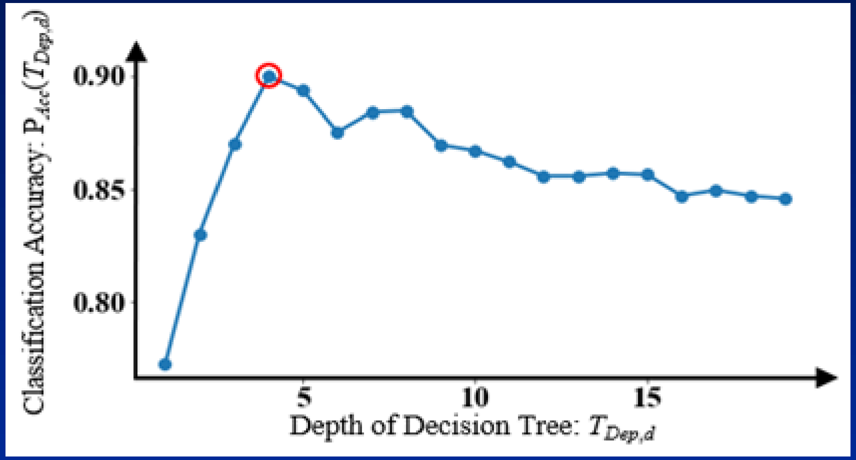
Hyper-Parameter Tuning

- **Depth of decision tree: T_{Dep}**
  - Select from a set of candidates $\{T_{Dep,1}, T_{Dep,2}, \ldots\}$
  - Cross-validation (CV)

    J-fold of data $\{D_{L,CV,1}, D_{L,CV,2}, \ldots, D_{L,CV,J}\}$
    J-1 folds for training and 1-fold for validation

$$P_{Acc}(T_{Dep,d}) = \frac{1}{J}\sum_{j=1}^{J} P_{Acc}(j, T_{Dep,d})$$

    Select T_{Dep} with maximum $P_{Acc}(T_{Dep})$



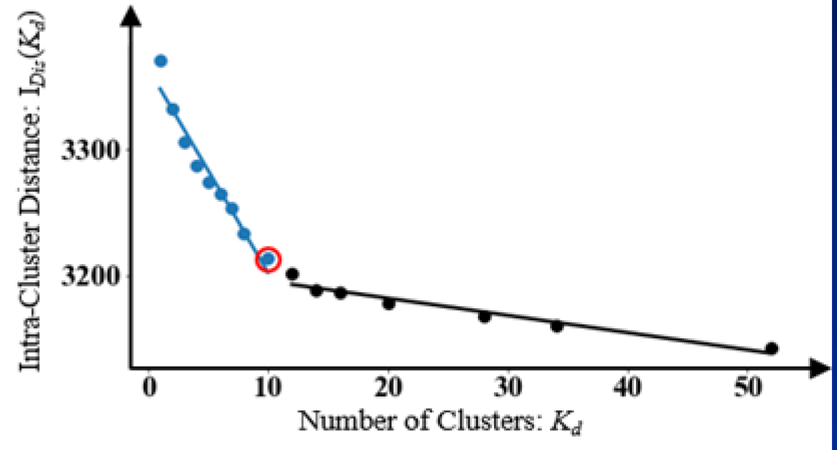
Hyper-Parameter Tuning

- **Frequency threshold for frequent-pattern mining: N_{Th}**
  - Select from a set of candidates $\{N_{Th,1}, N_{Th,2}, \ldots\}$
  - Calculate a set of triplets $\{(N_{Th,1}, K_1, I_{Dis,1}), (N_{Th,2}, K_2, I_{Dis,2}), \ldots\}$

$$I_{Dis,d} = \sum_{k=1}^{K_d} \sum_{n=1}^{N_{d,k}} \left\| d_{L,Fail,d,k,n} - \bar{d}_{L,Fail,d,k} \right\|_2^2$$

K: # of clusters

L-method to determine the "knee point" and the effective N_{Th}



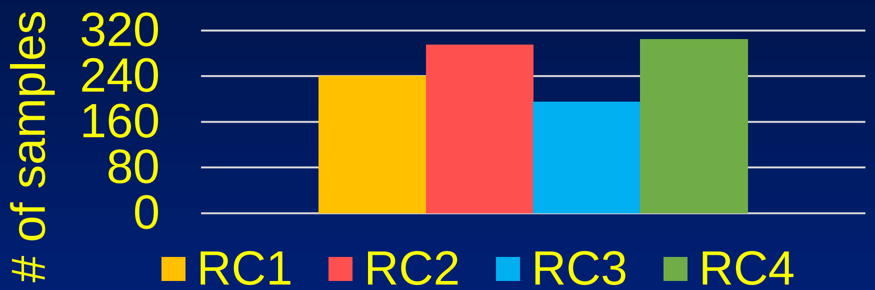
Experimental Results on Network Product #1

- **Industrial Network test data:**
  - **Training: 1597 samples with pass/fail label**
  - **Diagnosis: 1012 samples with RCs labeled by human experts**
  - **19 numerical features and 4 RCs**

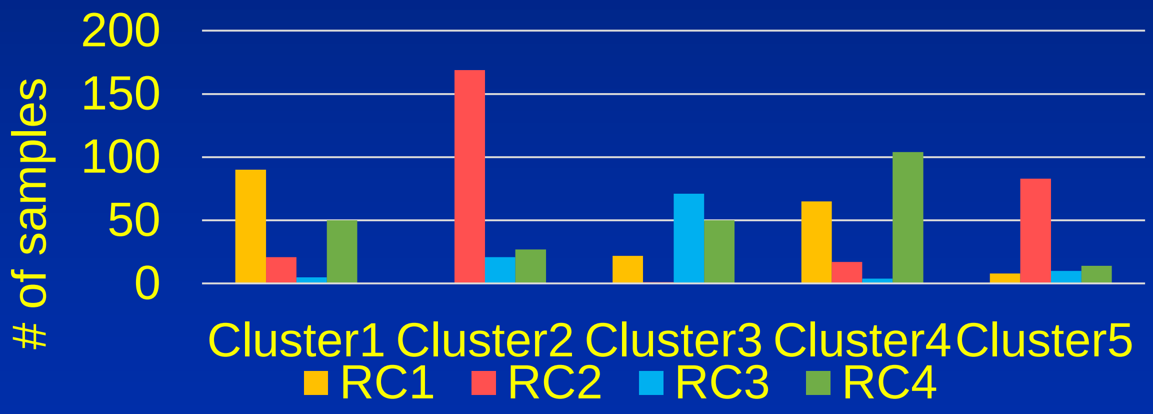
	NMI	**Optimum *K***	**Runtime (s)**
Hierarchical clustering	0.062	50	10.7
Proposed algorithm	**0.293**	**8**	**19.0**

Experimental Results on Network Product #1

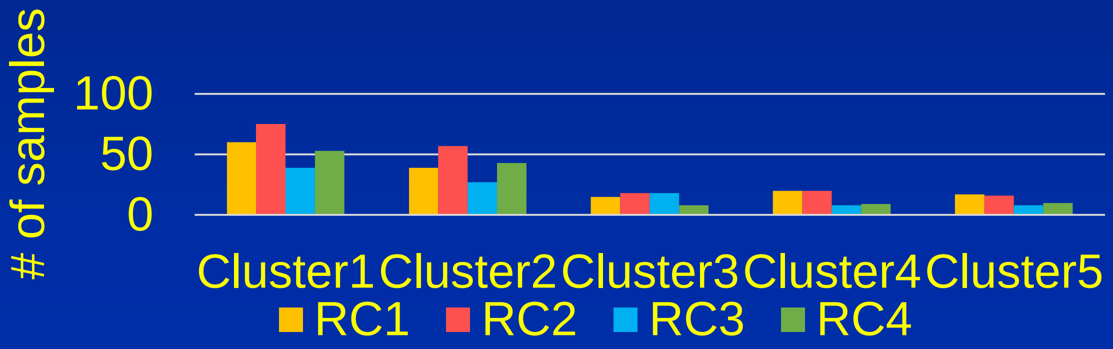
- **Root-cause distribution on test data**



- **Root-cause distribution on largest 5 clusters**



Proposed algorithm



Hierarchical clustering

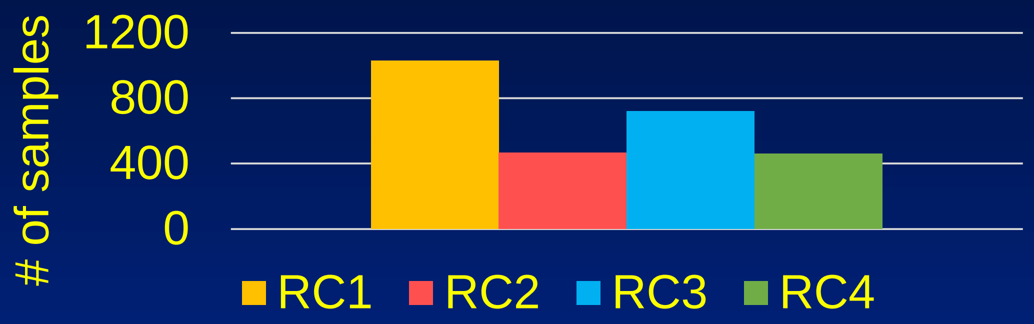
Experimental Results on Network Product #2

- **Industrial Network test data:**
  - **Training: 6299 samples with pass/fail label**
  - **Diagnosis: 2678 samples with RCs labeled by human experts**
  - **17 numerical features and 4 RCs**

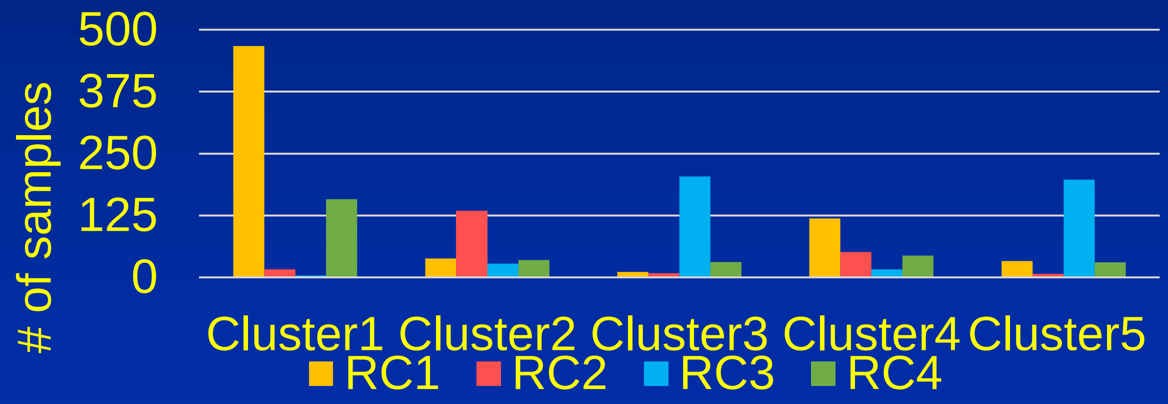
	NMI	Optimum *K*	Runtime (s)
Hierarchical clustering	0.093	45	13.6
Proposed algorithm	**0.283**	**10**	**54.4**

Experimental Results on Network Product #2

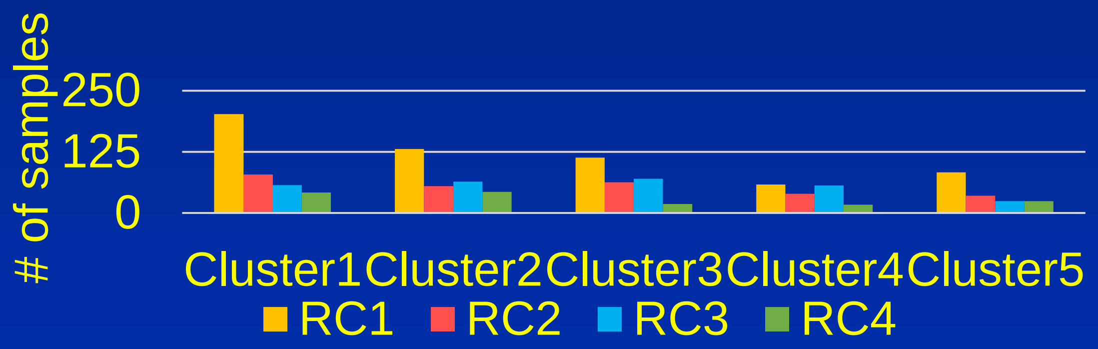
- **Root-cause distribution on test data**



- **Root-cause distribution on largest 5 clusters**



Proposed algorithm



Hierarchical clustering

Conclusions

- Formulate the root-cause analysis as an unsupervised clustering problem
- Propose a two-stage clustering method leveraging the pass/fail information
- Outperform the state-of-the-art hierarchical clustering method

Thanks!