Finding Similar and Dissimilar Events for Preprocessing Logs and Improving Mining Results - *A Snapshot of Thesis and Tools*

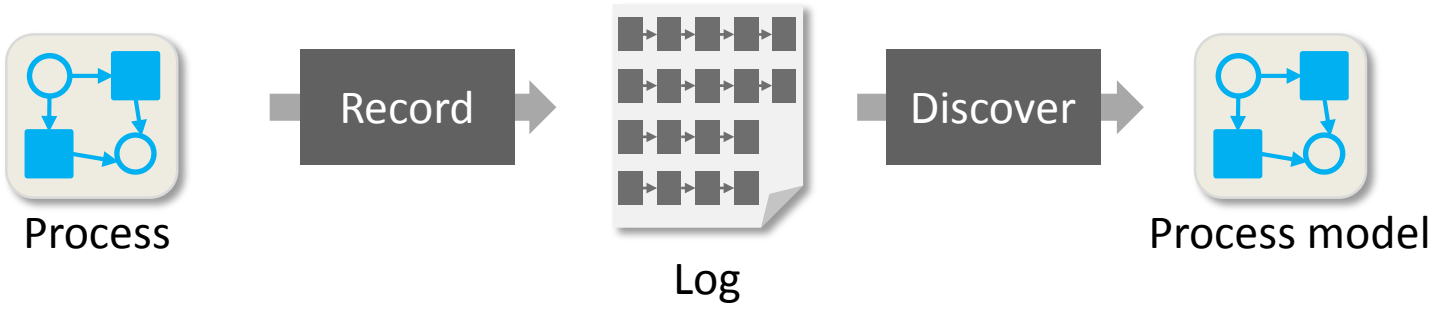
Xixi Lu, Dirk Fahland, Wil van der Aalst

Eindhoven University of Technology

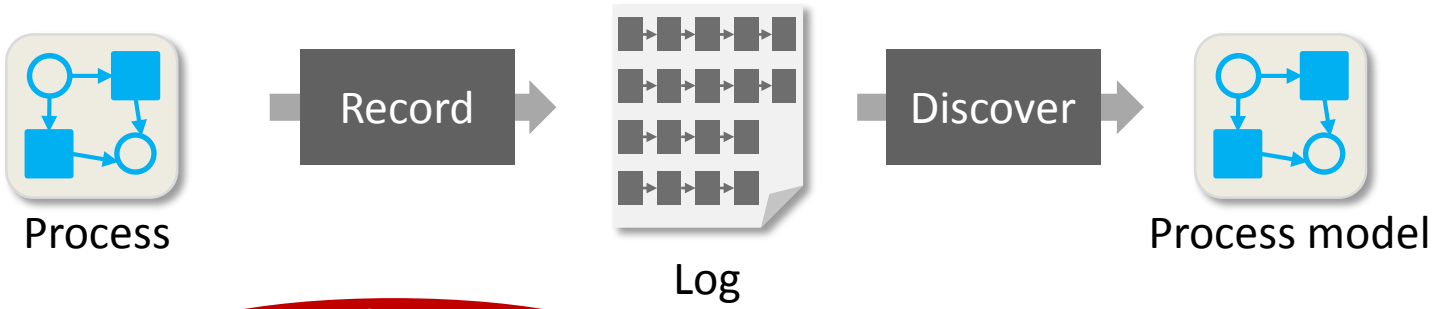
Agenda

- (5 min) I. Process discovery and challenges
- (5 min) II. Research problem
- (10 min) III. Approach and applications (Recall)
- (20 min) IV. Demo and Discussions

Process Discovery

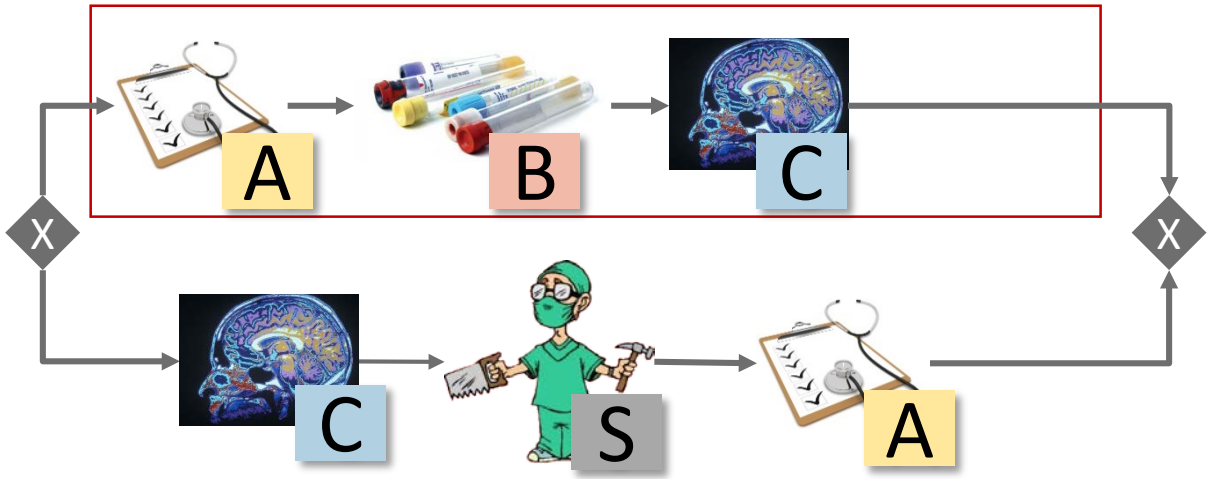


Challenges in Flexible Environments



1. High variety of behavior
2. Deviating behavior
3. Duplicated tasks

1. High Variety of Behavior

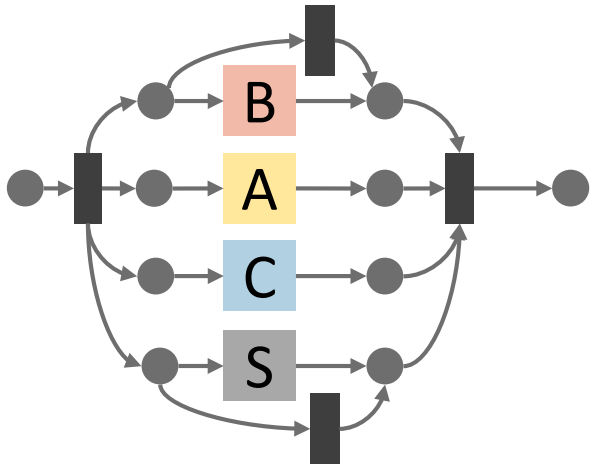


Record

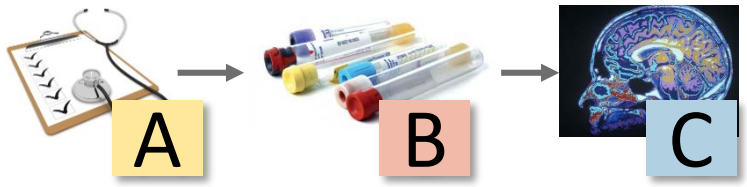
A → B → C

C → S → A

Discover



2. Deviating behavior



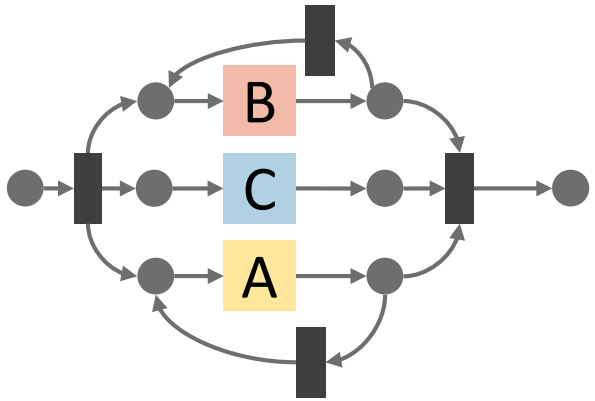
Record

A → B → C

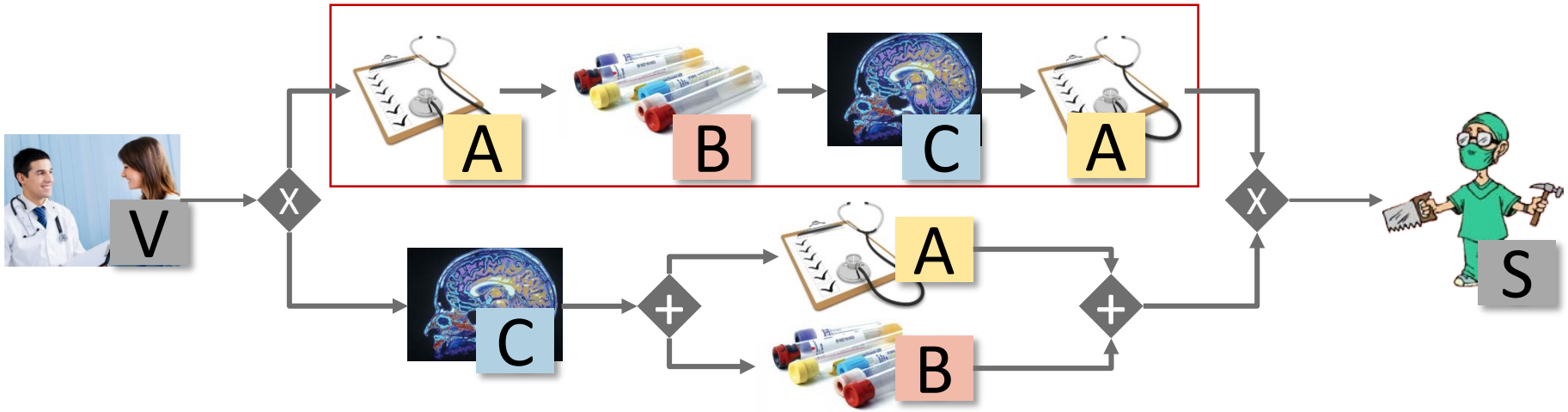
A → B → C → A

A → B → C → B

Discover



3. Duplicated Tasks



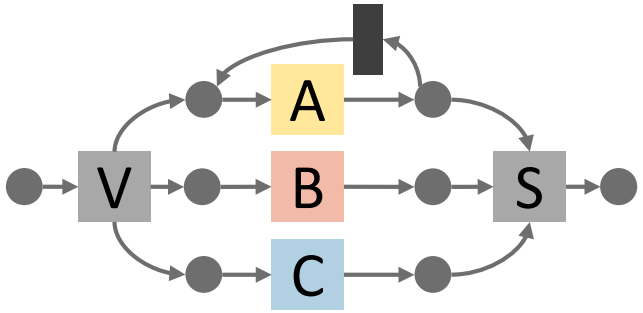
Record

V → A → B → C → A → S

V → C → B → A → S

V → C → A → B → S

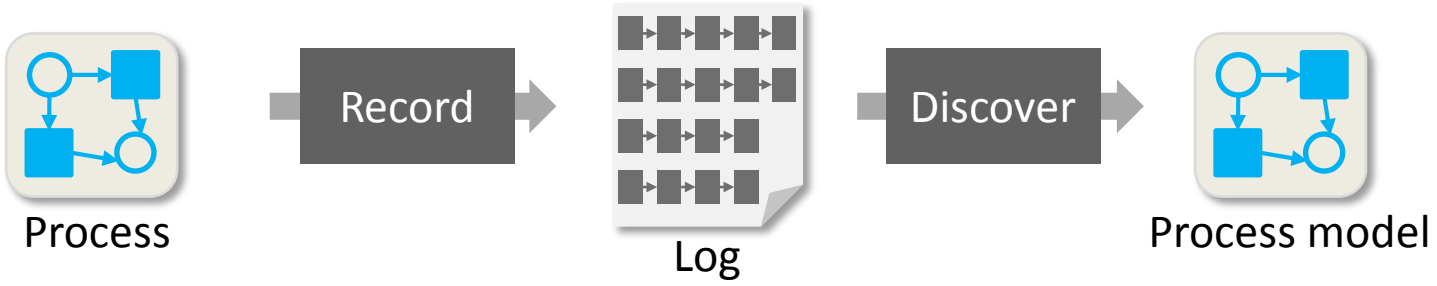
Discover



Challenges

System unknown!

We don’t know what is optimal!



1. High variety of behavior
2. Deviating behavior
3. Duplicated tasks

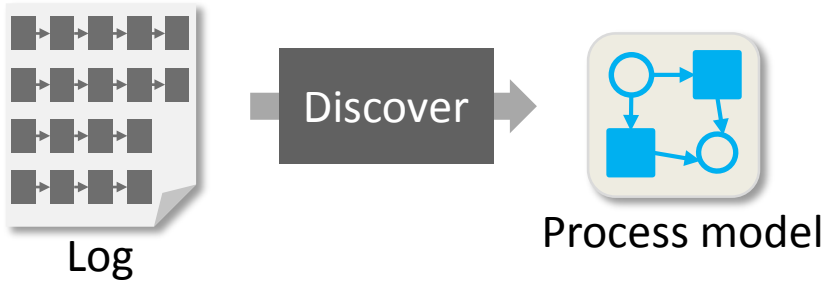
Cluster?

Filter?

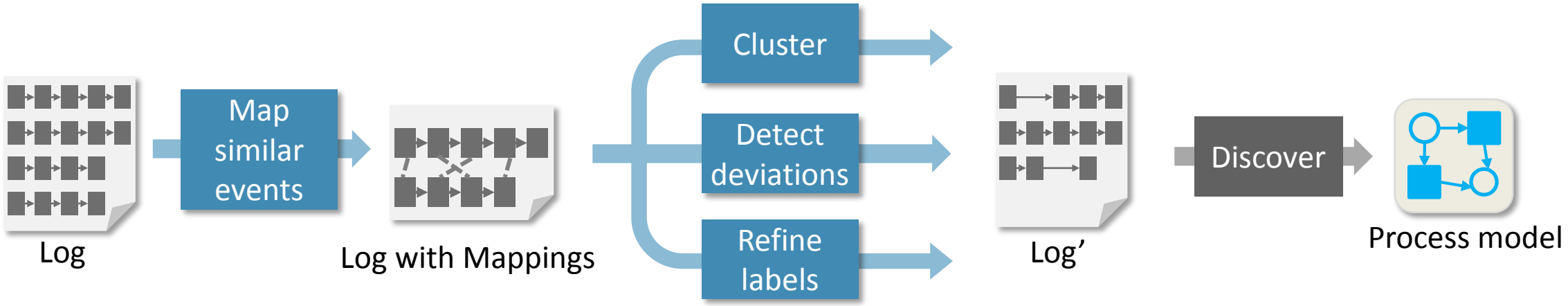
Yet Another Discovery?

Having dedicated algorithms?

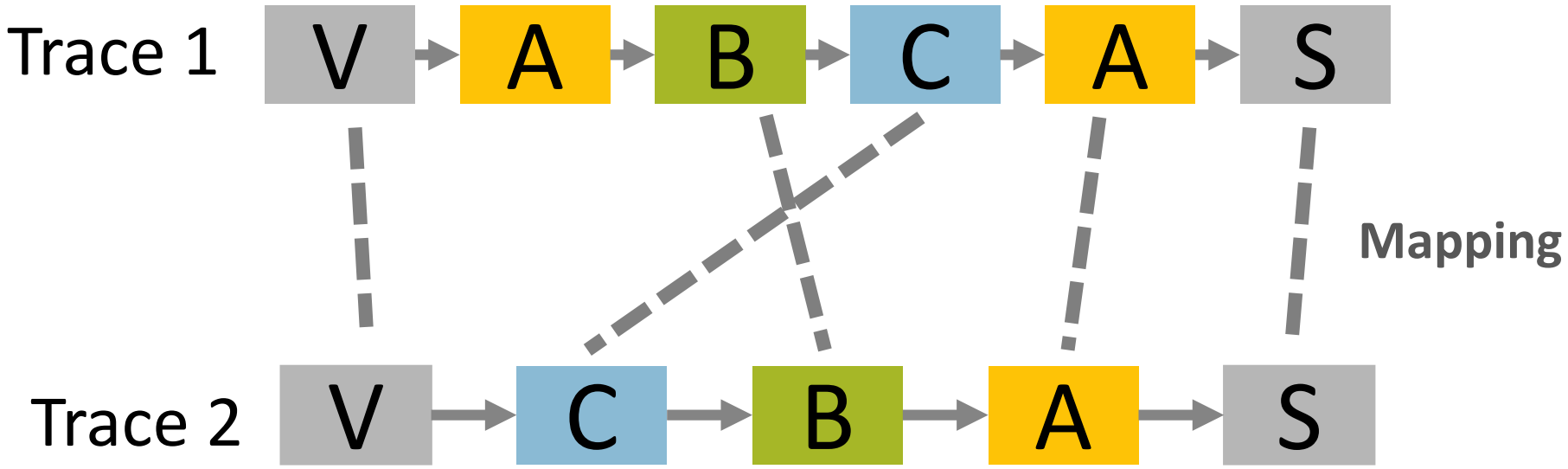
Challenges



Research Problem and Approach

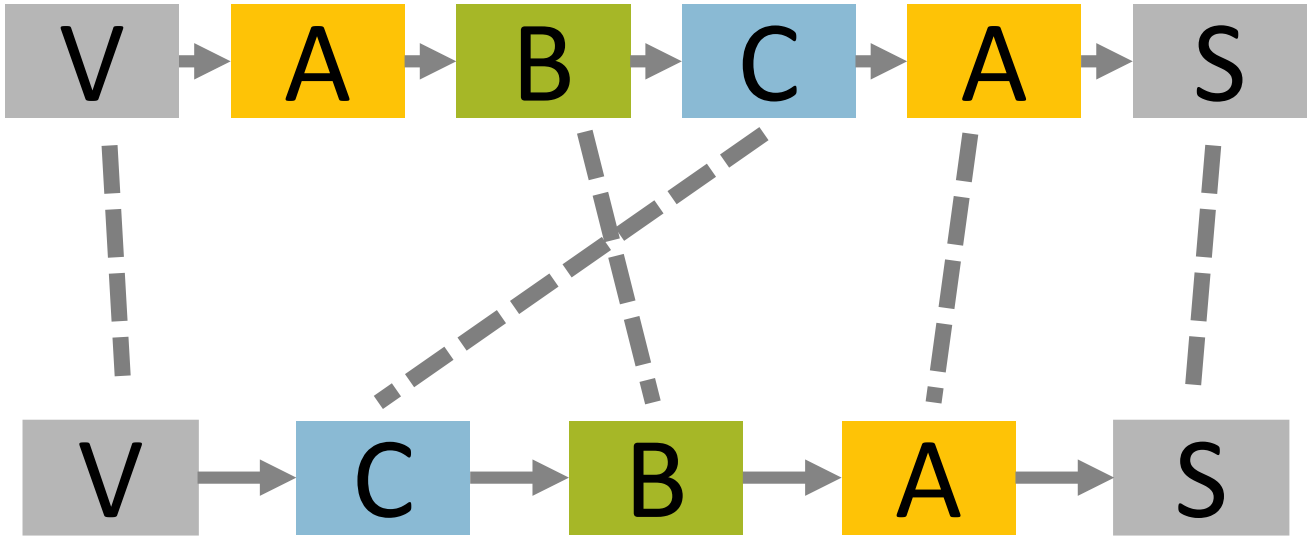


Mapping Between Events



Quantify Similarity of Mapped Events

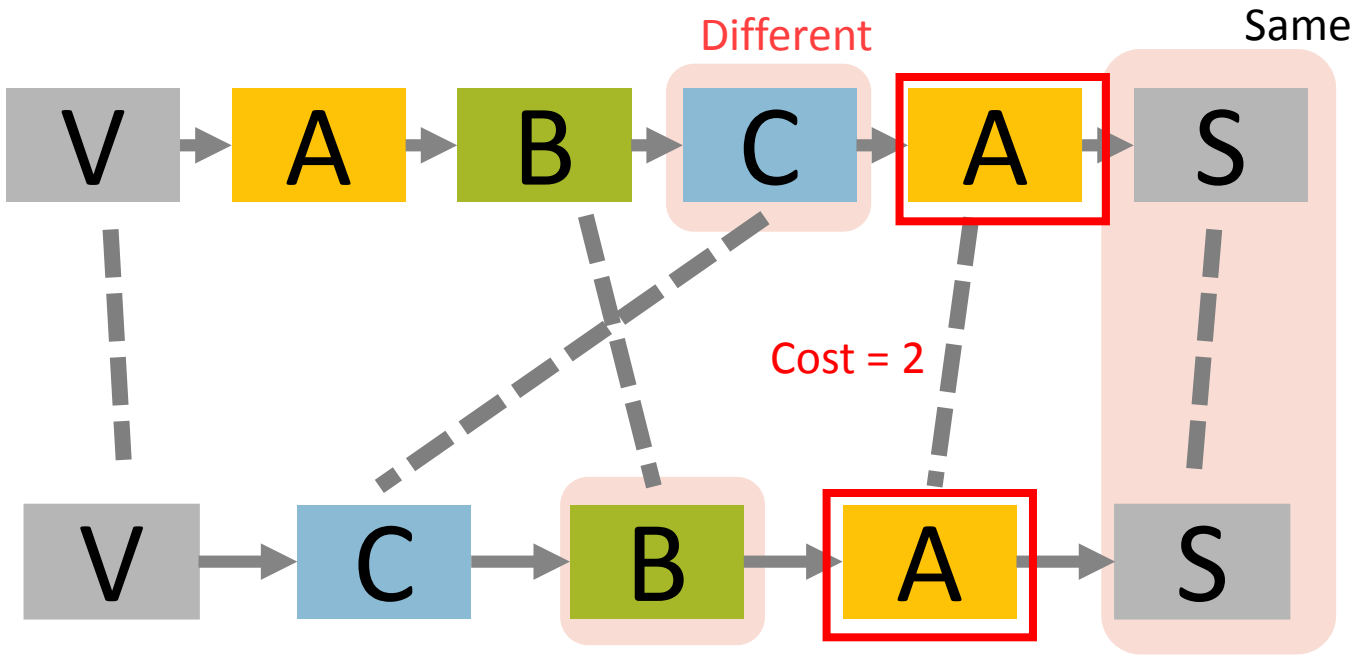
… based on “structural context of events”



Quantify Similarity of Mapped Events

… based on “structural context of events”

(1) Differences in neighbors

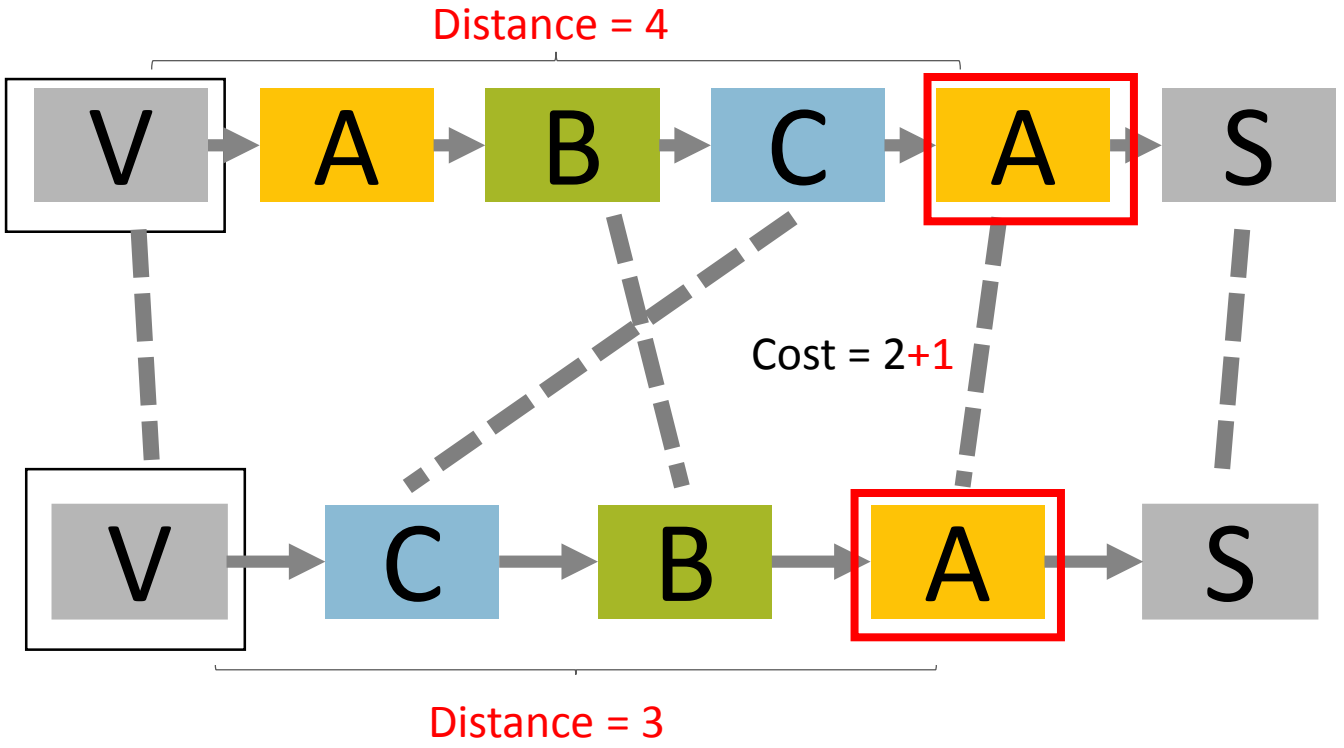


Quantify Similarity of Mapped Events

… based on “structural context of events”

(1) Differences in neighbors

(2) Differences in structure

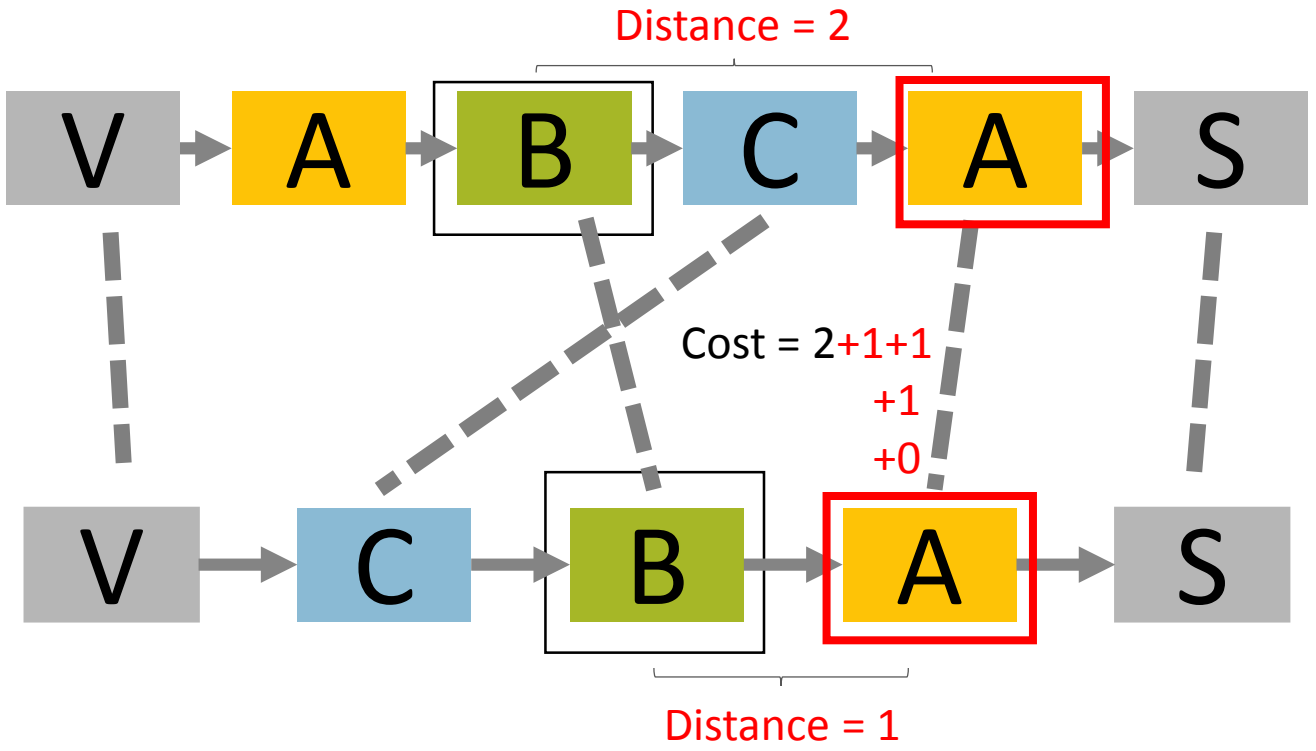


Quantify Similarity of Mapped Events

… based on “structural context of events”

(1) Differences in neighbors

(2) Differences in structure

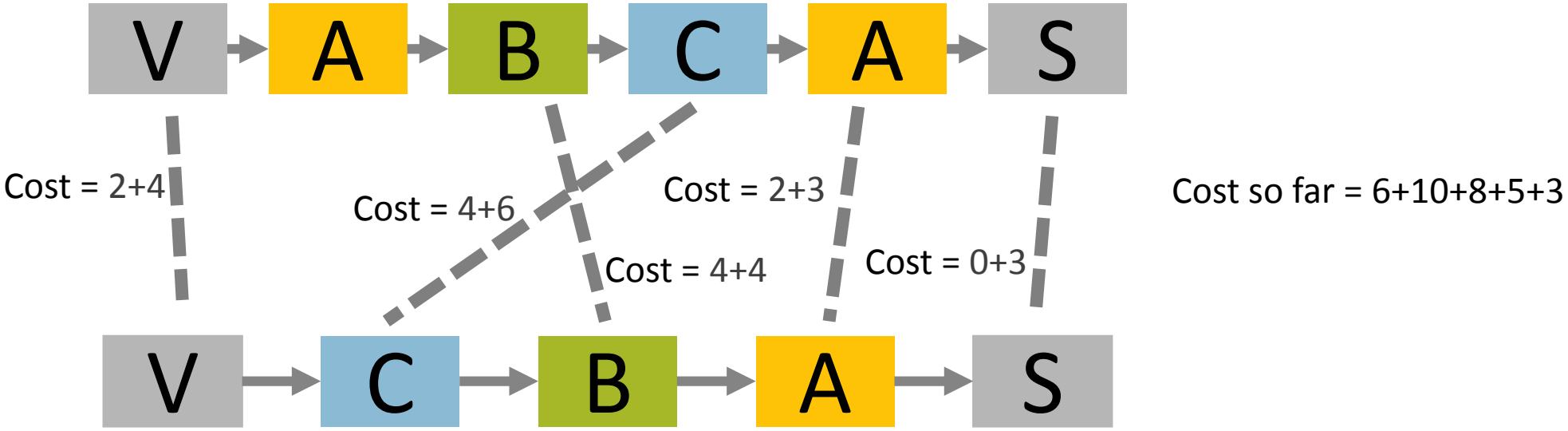


Quantify Similarity of Mapped Events

… based on “structural context of events”

(1) Differences in neighbors

(2) Differences in structure



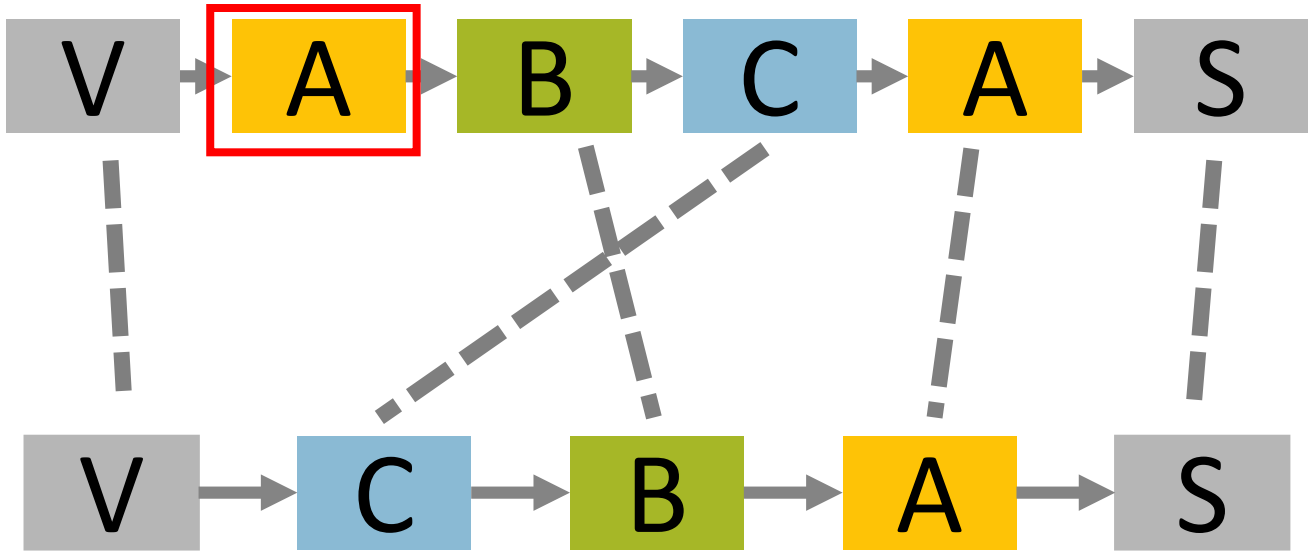
Quantify Similarity of Mapped Events

… based on “structural context of events”

(1) Differences in neighbors

(2) Differences in structure

(3) #Dissimilar events

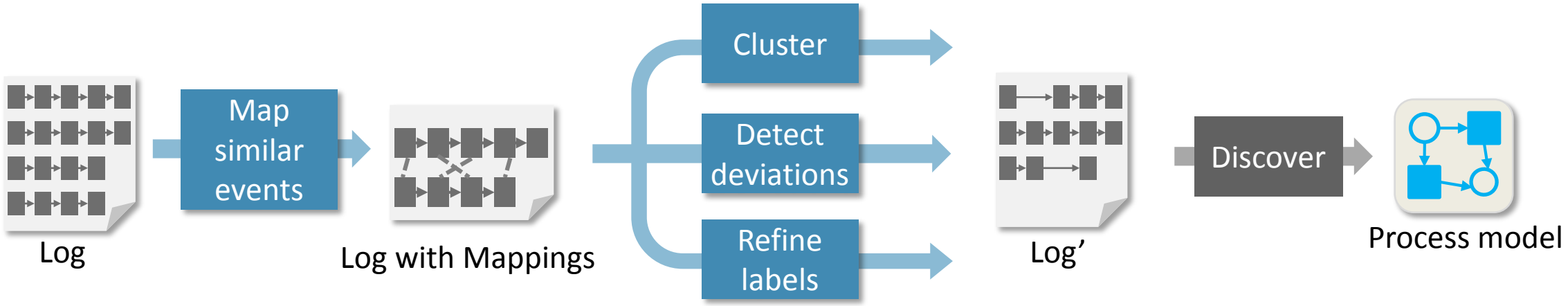


Total Cost = 6+10+8+5+3 +1

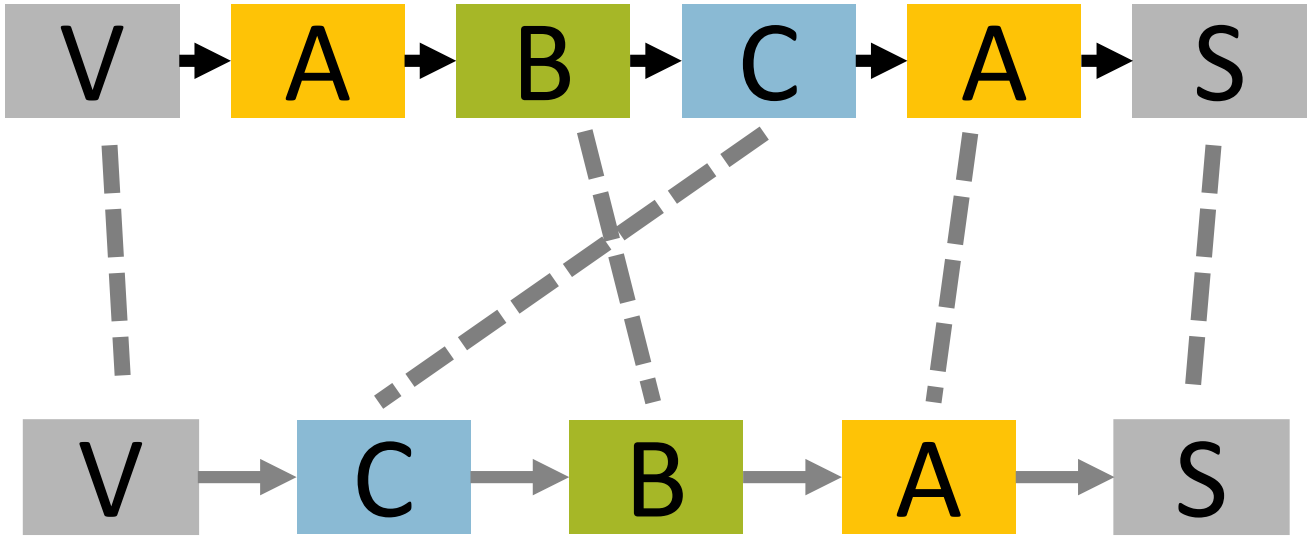
There is an algorithm to compute an optimal mapping :

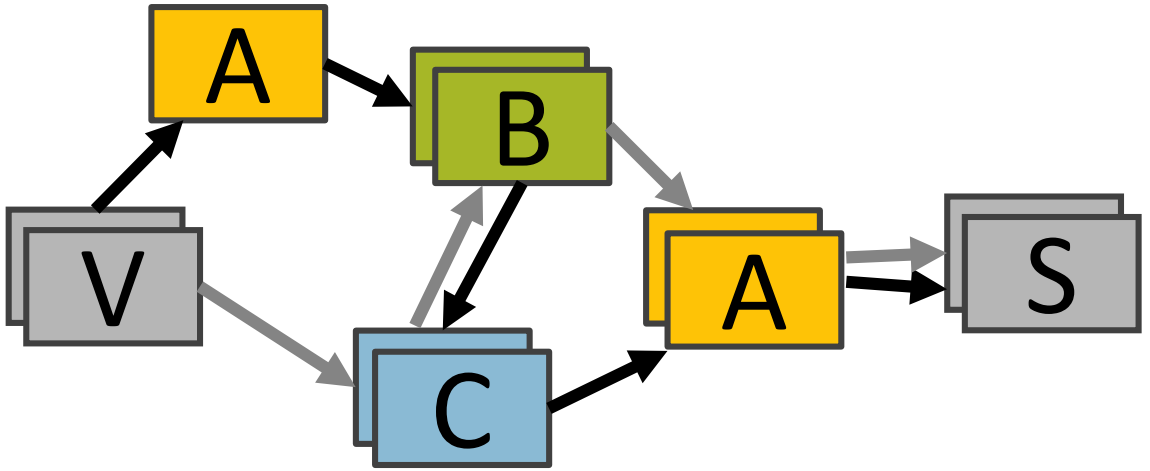
Xixi Lu, Dirk Fahland, Frank J.H.M. van den Biggelaar, and Wil M.P. van der Aalst. **Detecting Deviating Behavior without Models**. In *BPM Workshops* 2015, volume 256, pp.126–139, Springer International Publishing, 2015.

Proposing Approach

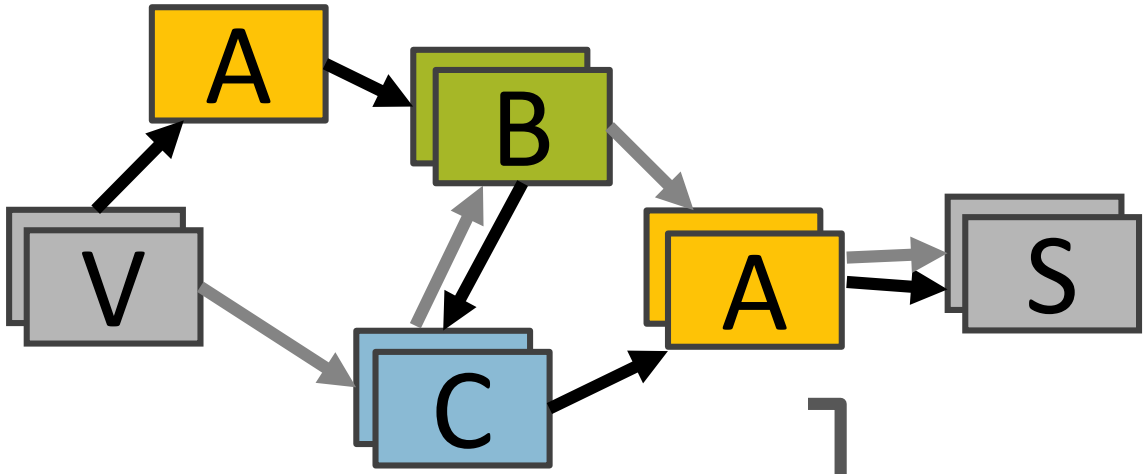


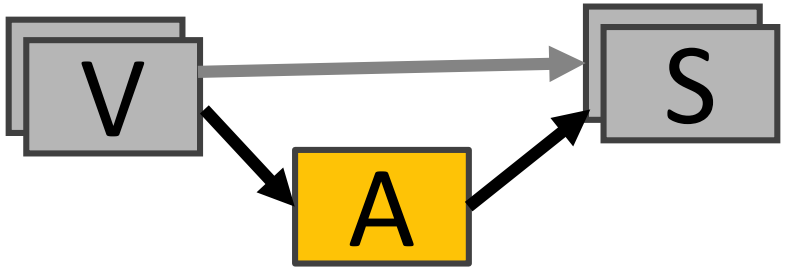
Cluster Traces Using Mappings and Fusion



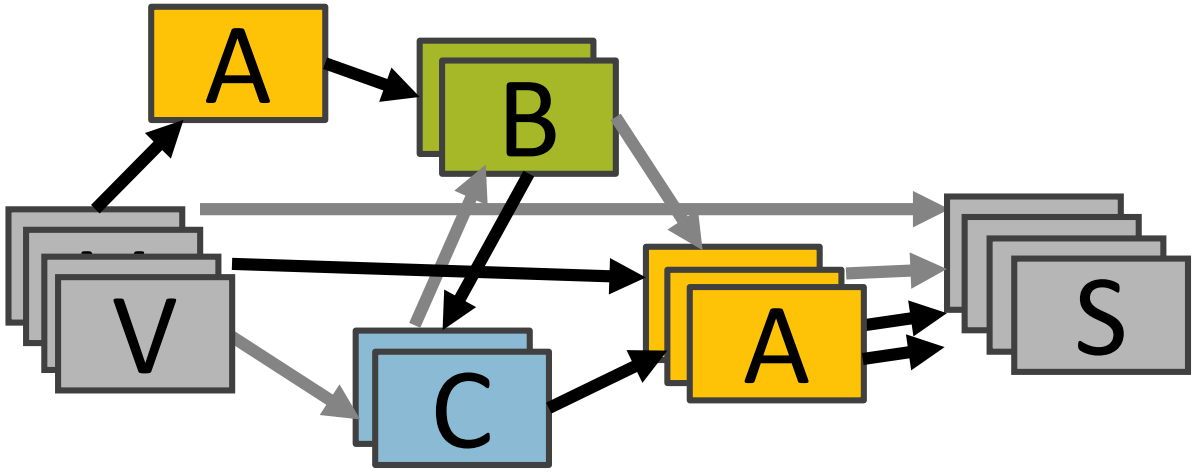


Cluster Traces Using Mappings and Fusion

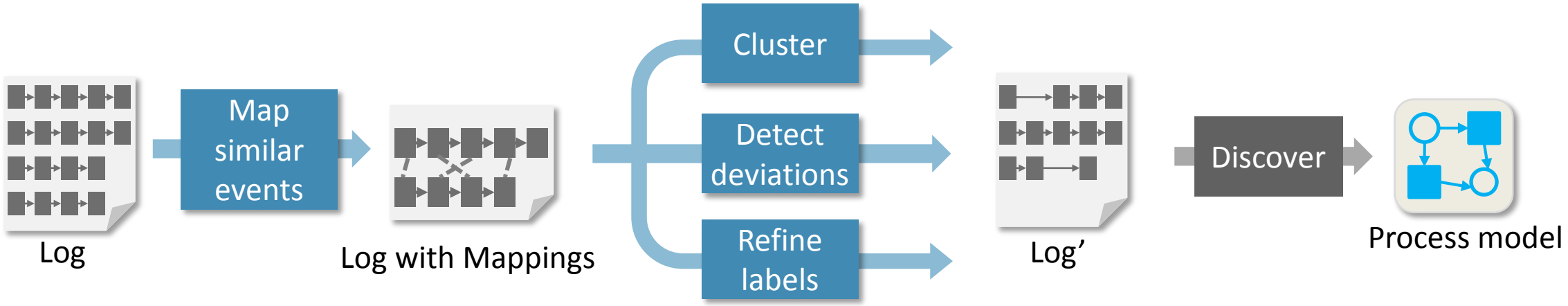




Hierarchical Cluster of traces



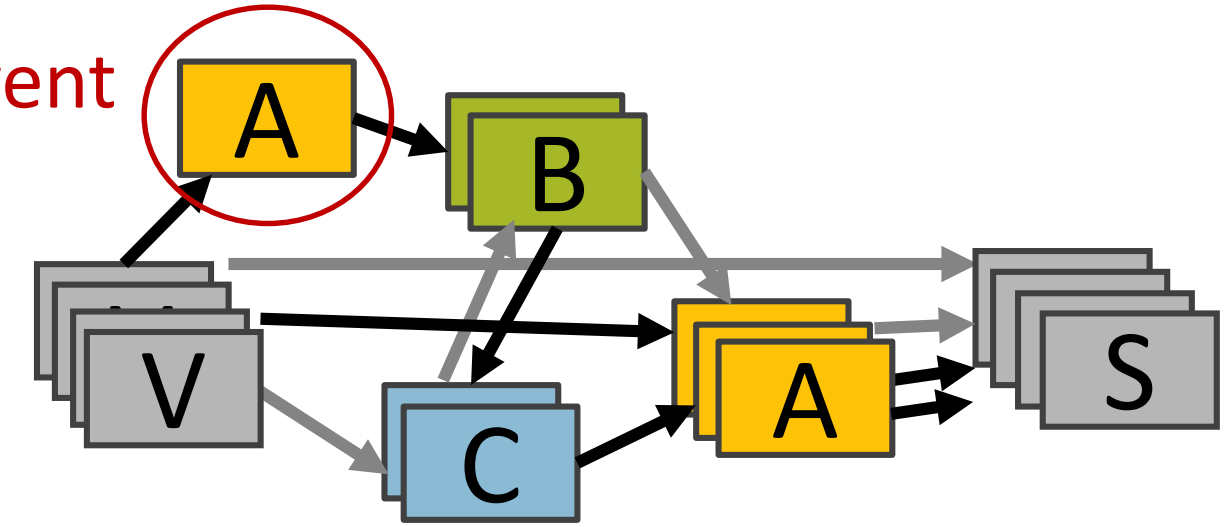
Proposing Approach



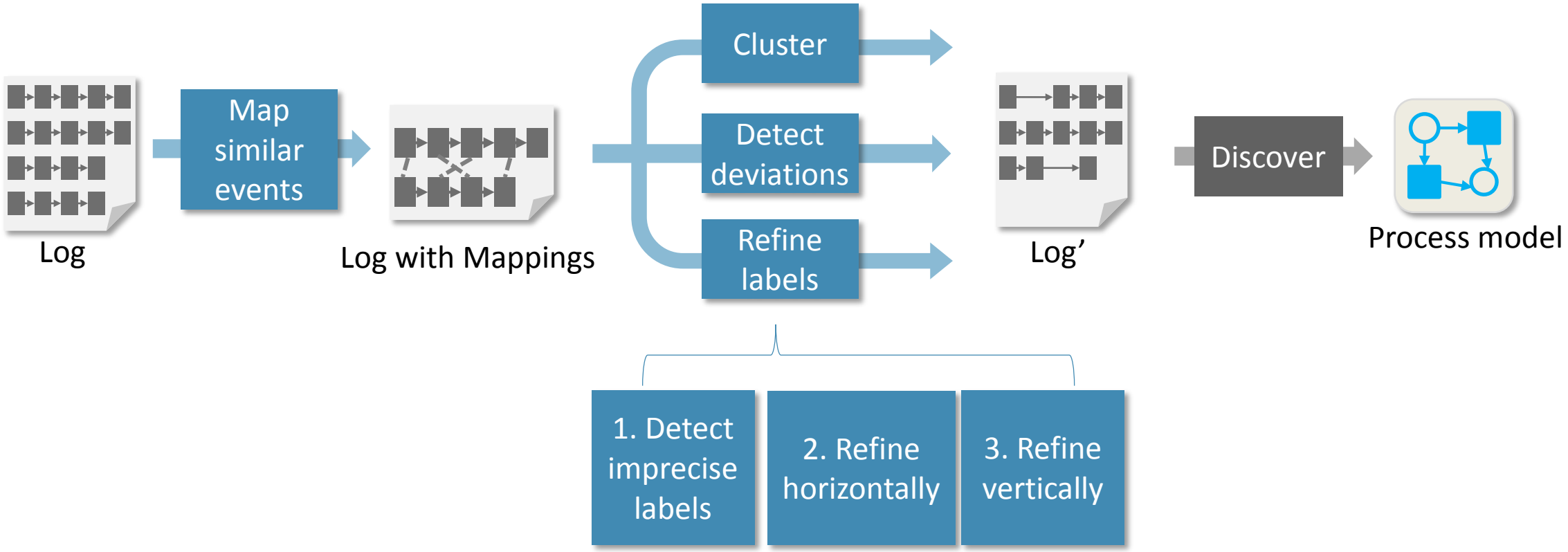
Detect Deviation Using Mappings

Threshold: “Deviating event” if < T% of cases

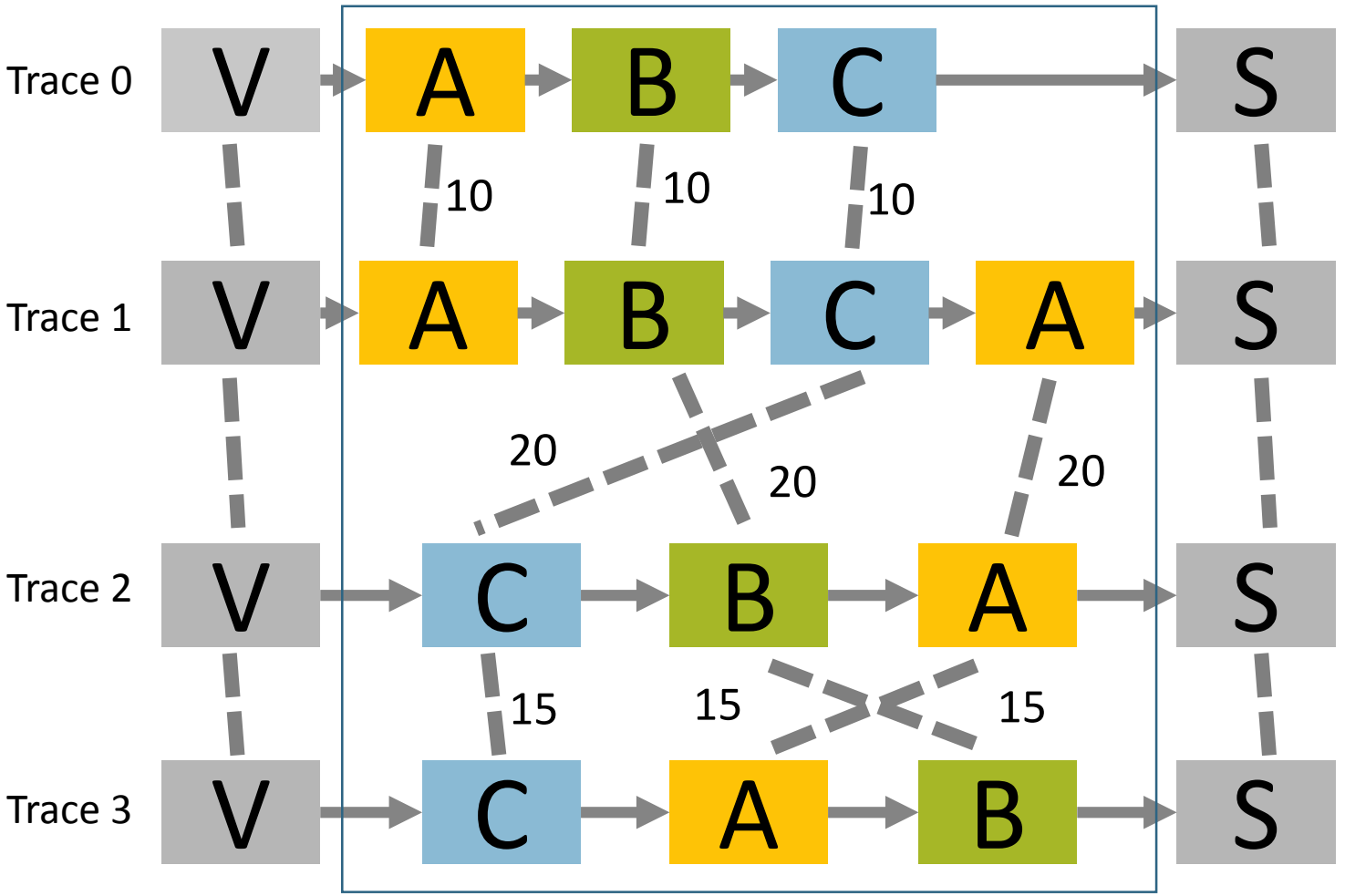
Deviating event
for T = 30%



Proposing Approach



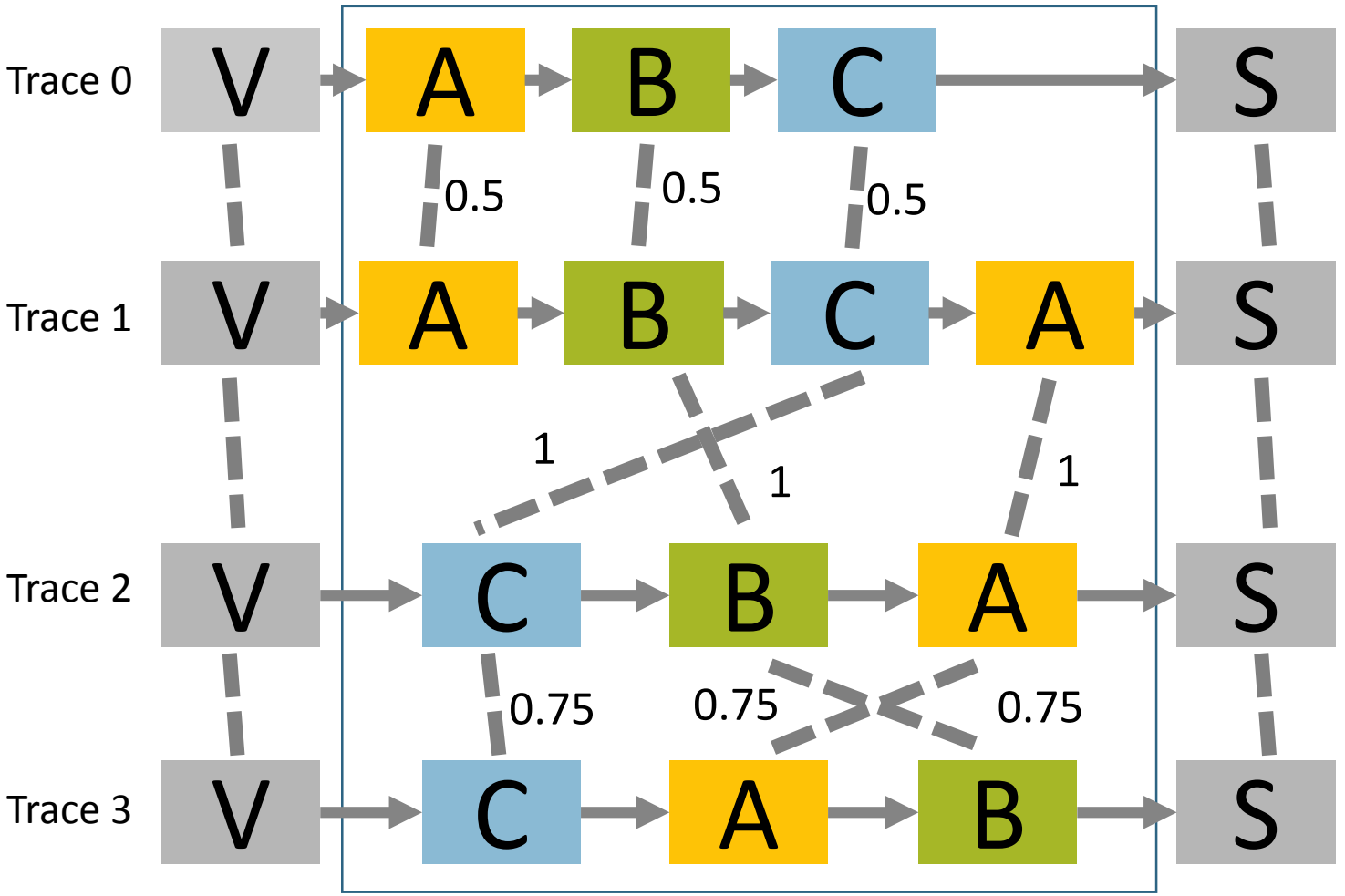
Refine Labels Horizontally



Imprecise label candidates {A, B, C}

a) Normalize costs w.r.t. maximal cost seen in the log

Refine Labels Horizontally



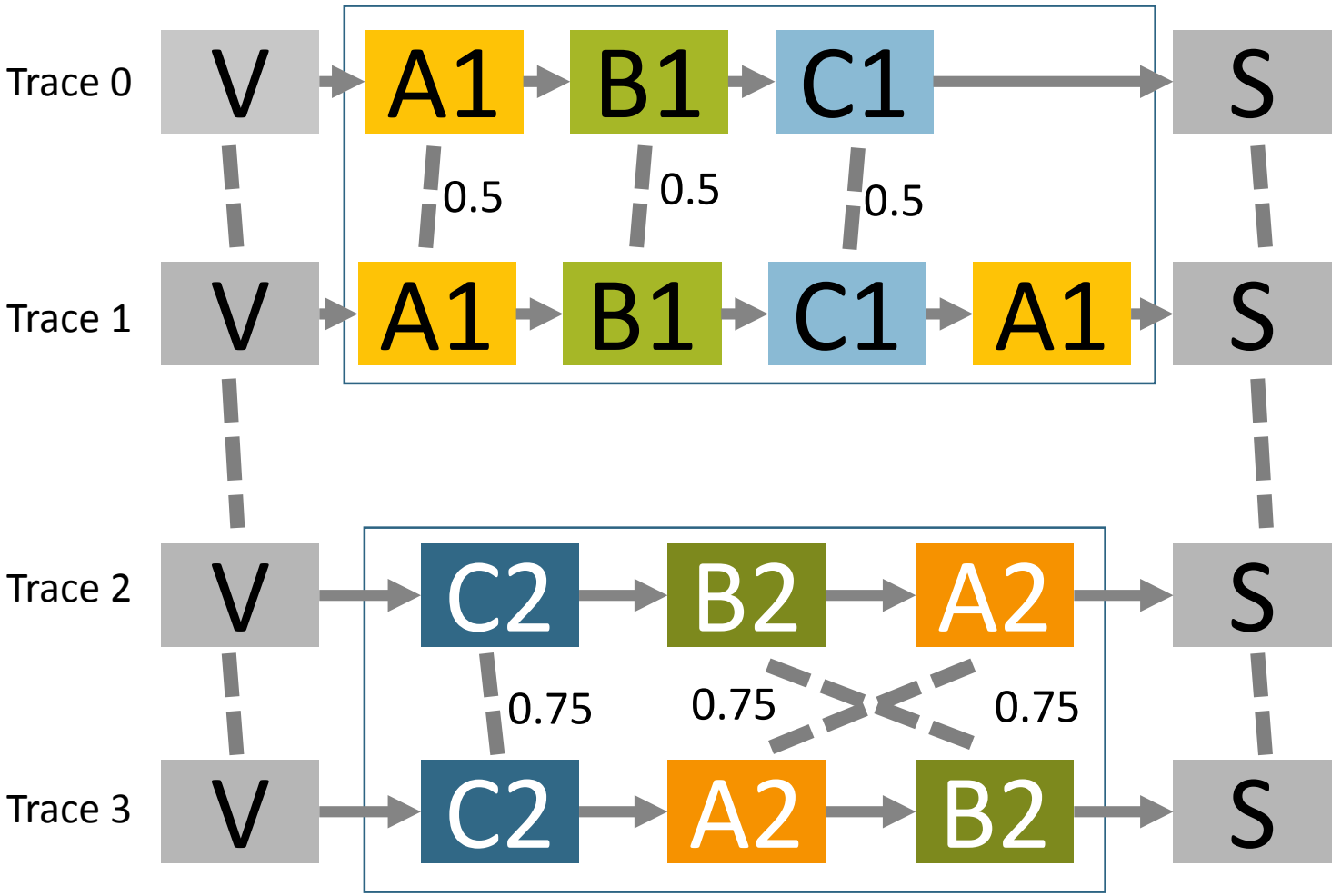
Imprecise label candidates {A, B, C}

a) Normalize costs w.r.t. maximal cost seen in the log

b) Set variant threshold, say, 0.8

c) Remove edges if cost > variant threshold

Refine Labels Horizontally

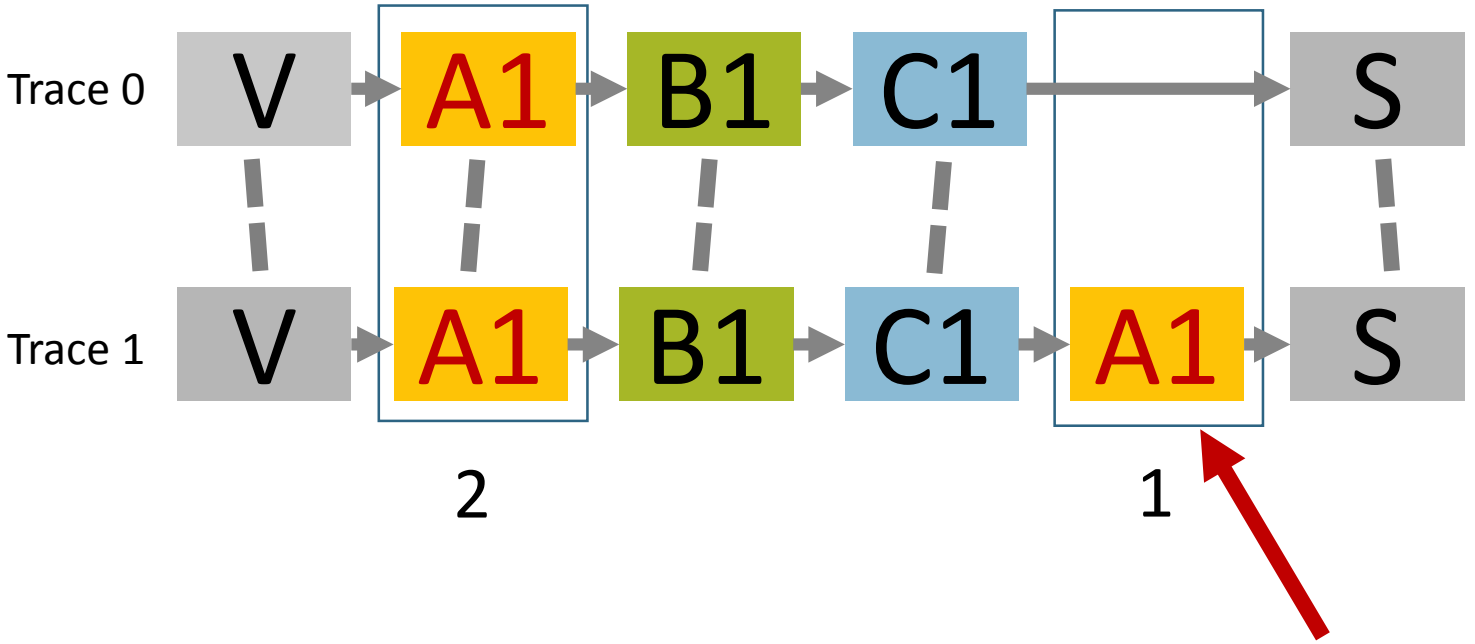


a) Normalize costs w.r.t. maximal cost seen in the log

b) Set variant threshold
say, 0.8

c) Remove edges
if cost > variant threshold

Refine Labels Vertically

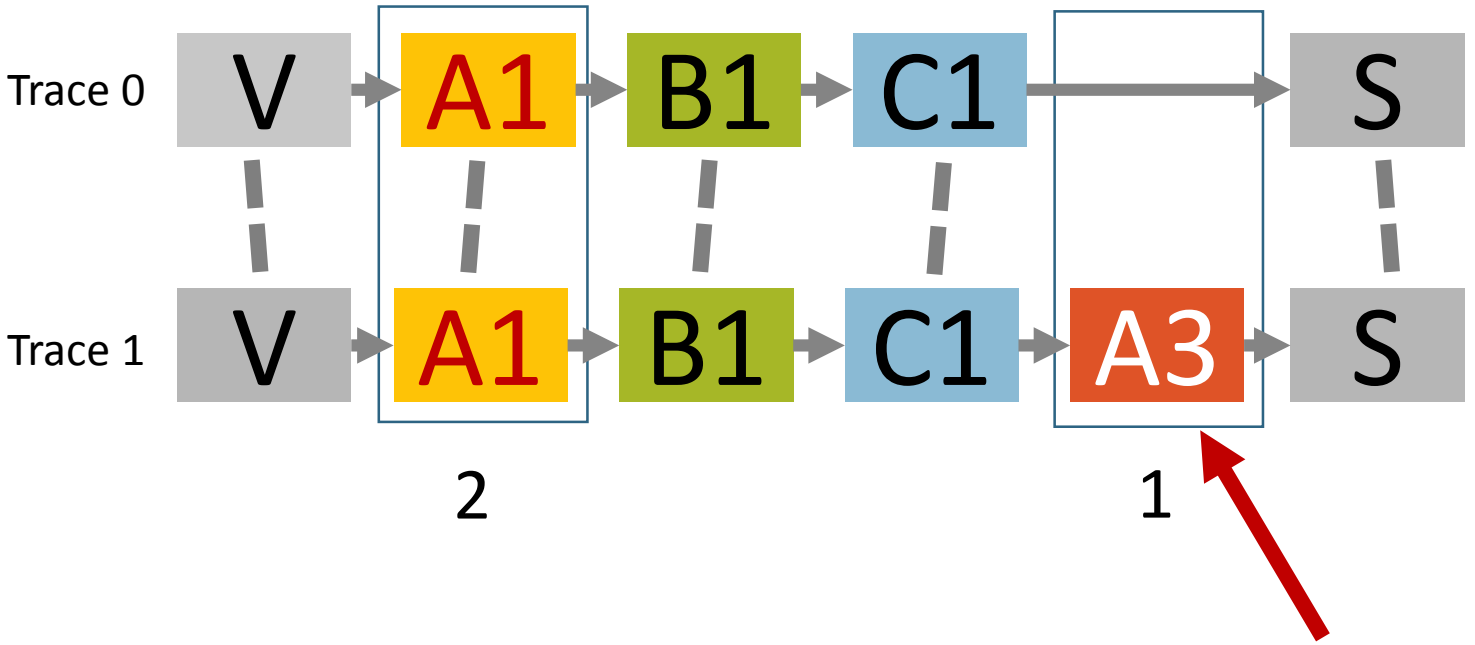


I) Set unfolding threshold,
say, 60%

II) Refine labels
if freq > unfolding threshold

Is this A1 a 2nd iteration of a loop?

Refine Labels Vertically

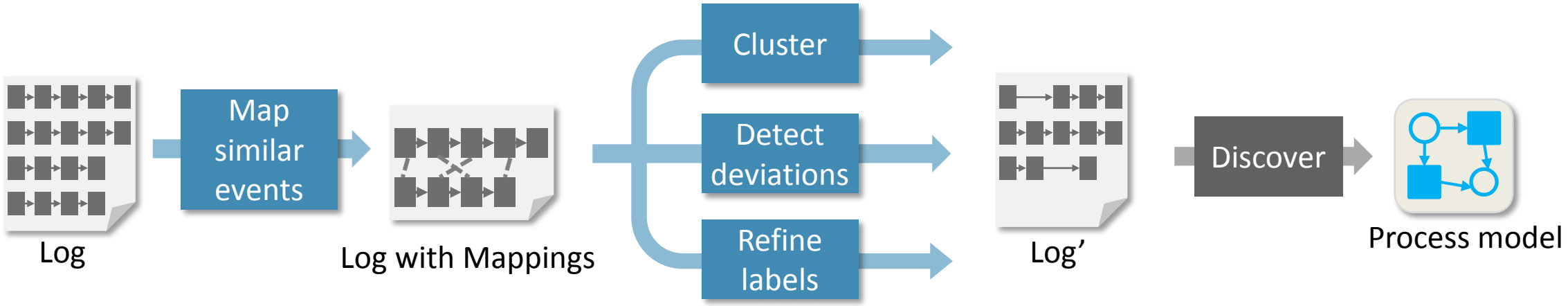


I) Set unfolding threshold,
say, 40%

II) Refine labels
if freq > unfolding threshold

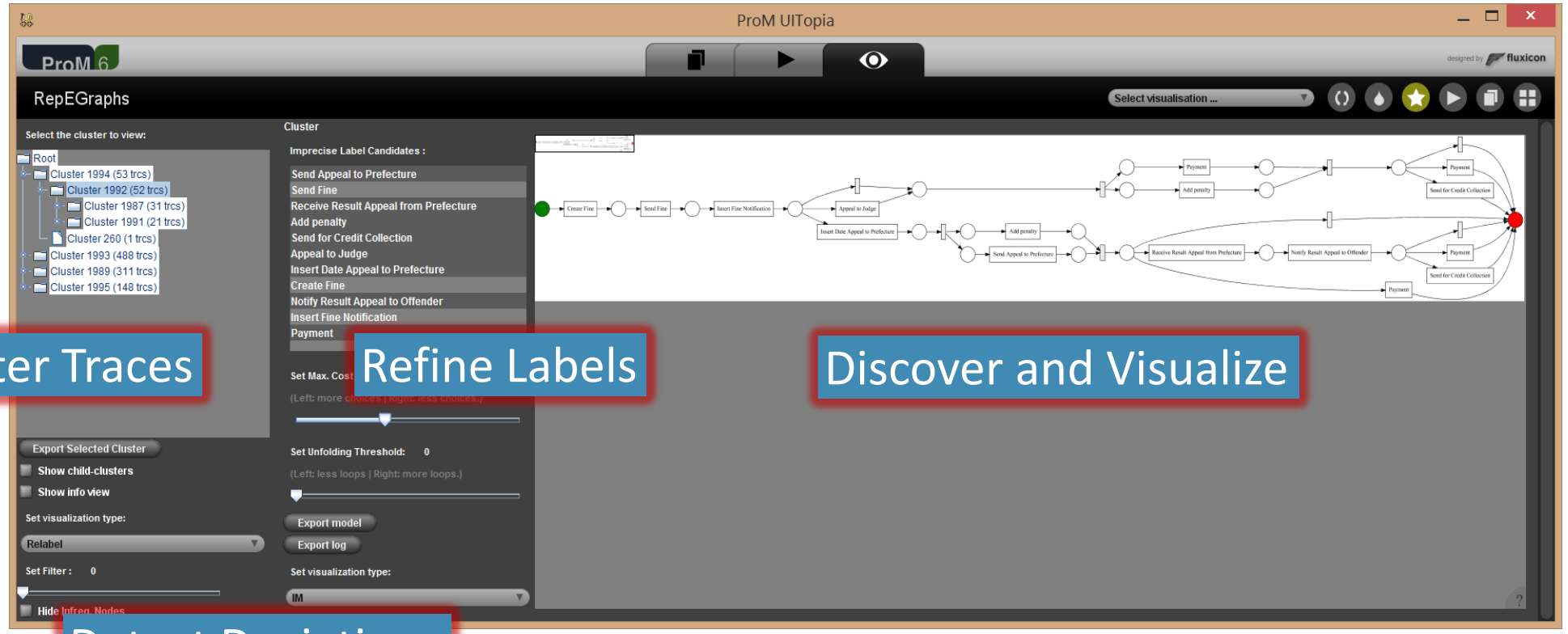
Or is this a different task?

Proposing Approach



IV. Demo and Discussion

- Plugin “Log to Model Explorer”



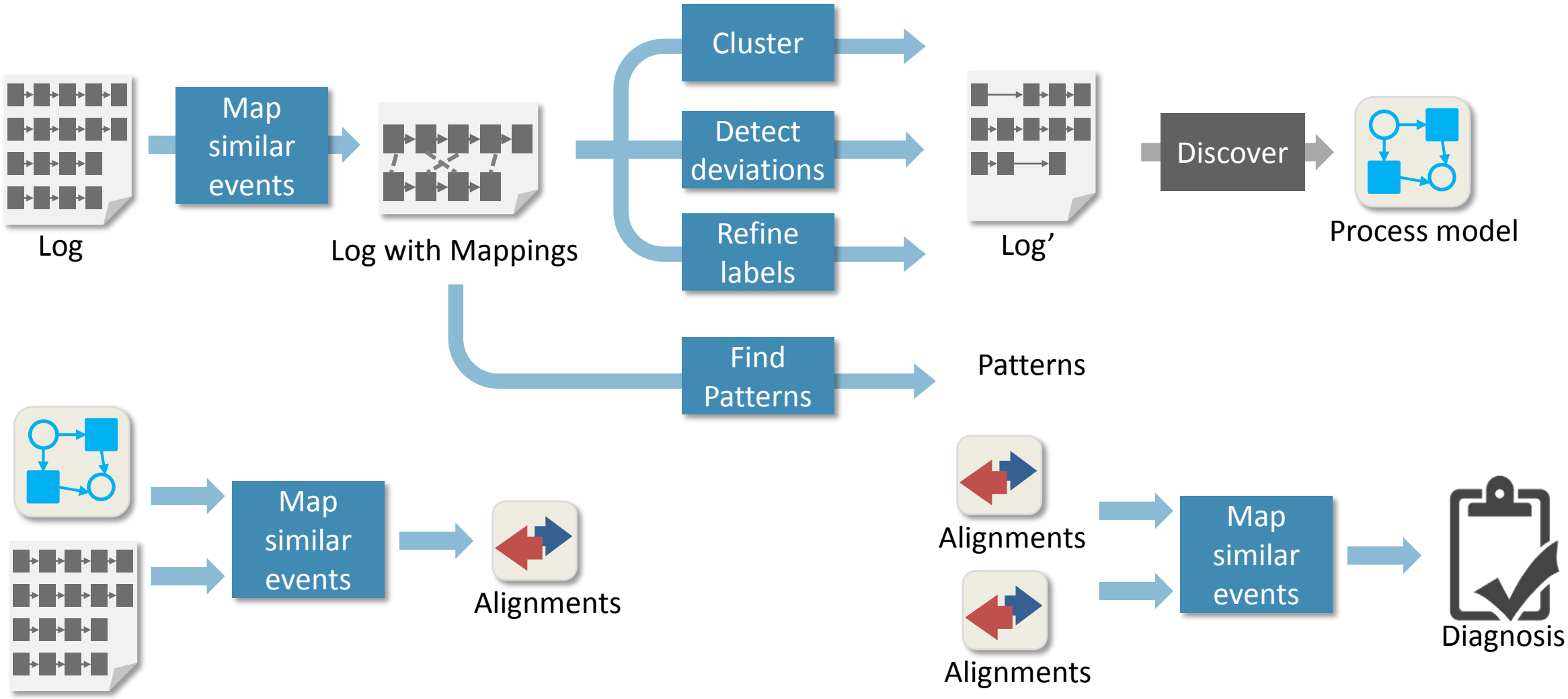
Xixi Lu, Dirk Fahland, and Wil M.P. van der Aalst. **Interactively Exploring Logs and Mining Models with Clustering, Filtering, and Relabeling**. In BPM Demo *2016*.

Limitations?

- Implementation inefficient
- Do not know what is happening in the background
- Can not handle large loops and large parallel branches

- Clustering
  - Do not know why these clusters
- Filtering
  - Do not support finding missing events (yet)
- Refining labels
  - …

The Bigger Picture?



Questions and Feedback?