# Process Mining

## Generalized Alignment-Based Trace Clustering

Mathilde Boltenhagen[1]

J.Carmona[2] T.Chatain[2]

5th, June 2019

[1]LSV, CNRS, ENS Paris-Saclay, Inria, Université Paris-Saclay
[2]Universitat Politècnica de Catalunya

# Data analysis

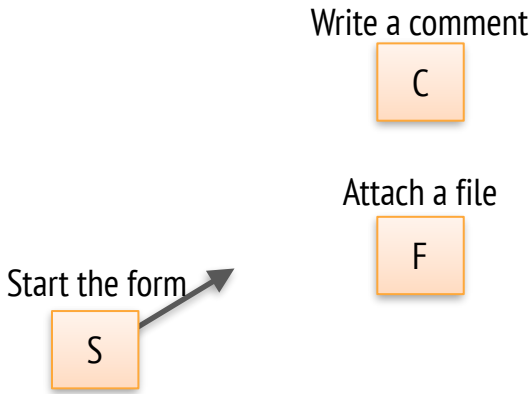
# Data analysis

# Data analysis

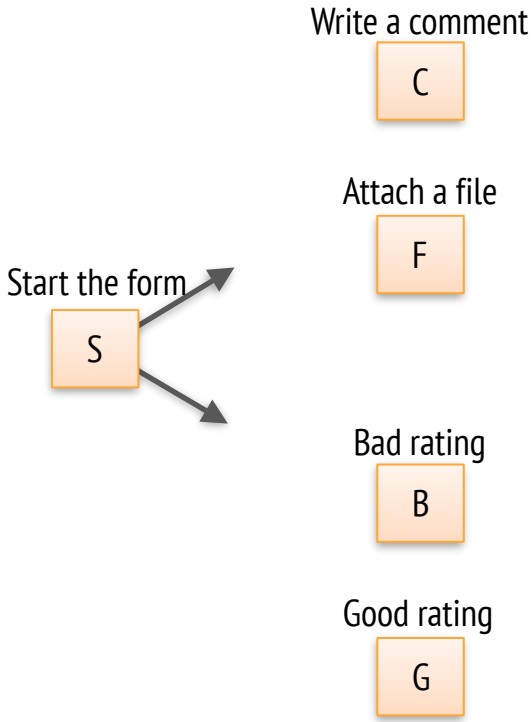
Start the form

S

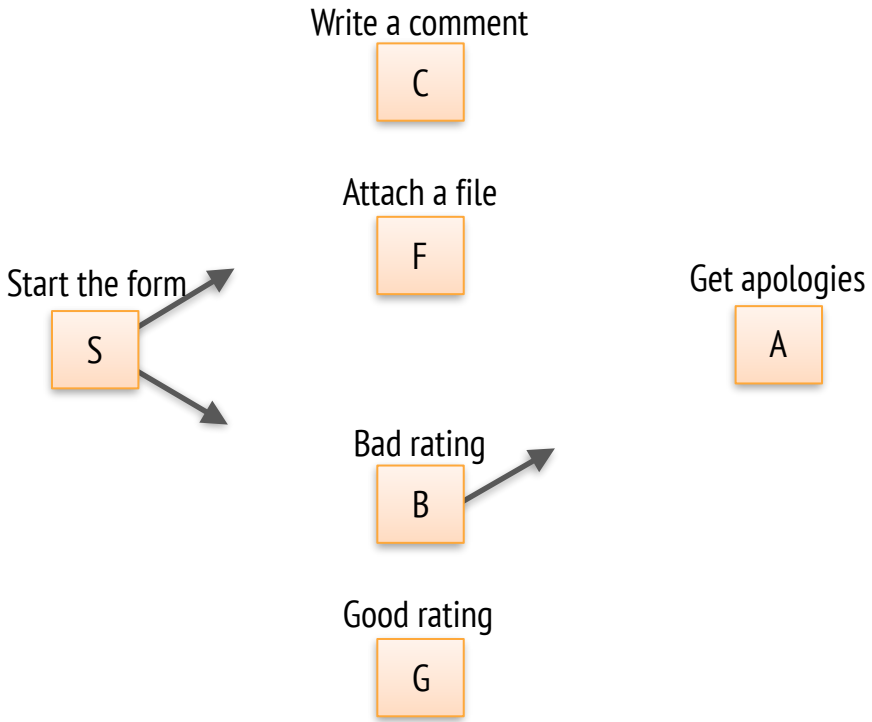
# Data analysis



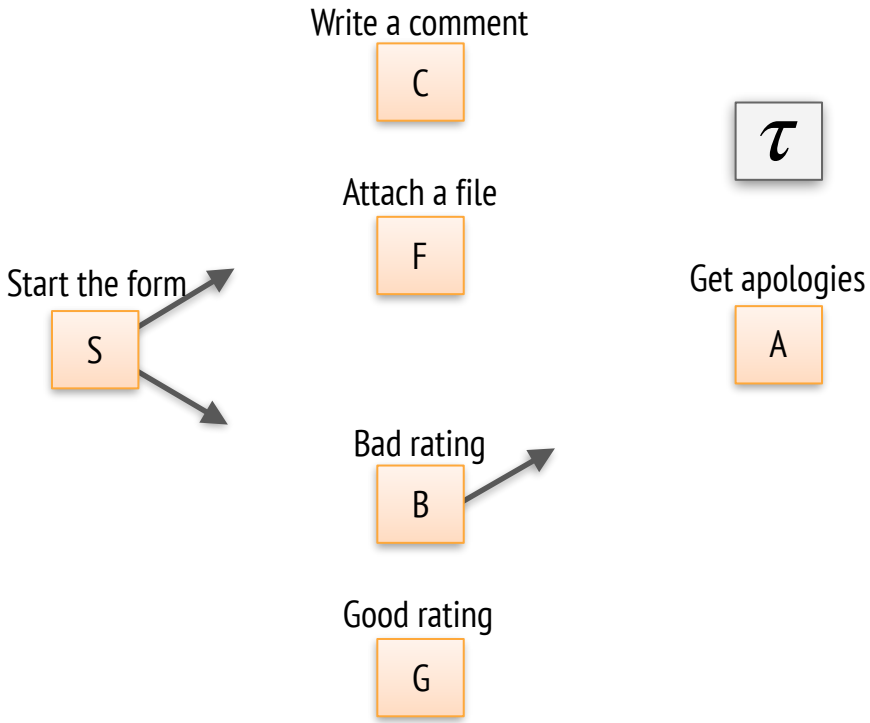
# Data analysis



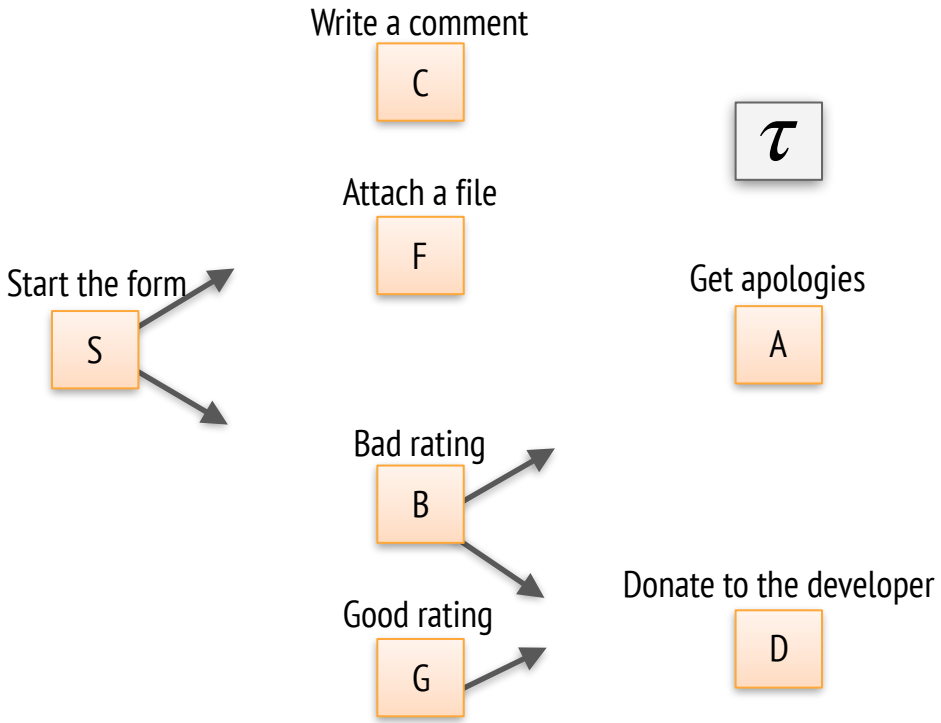
# Data analysis



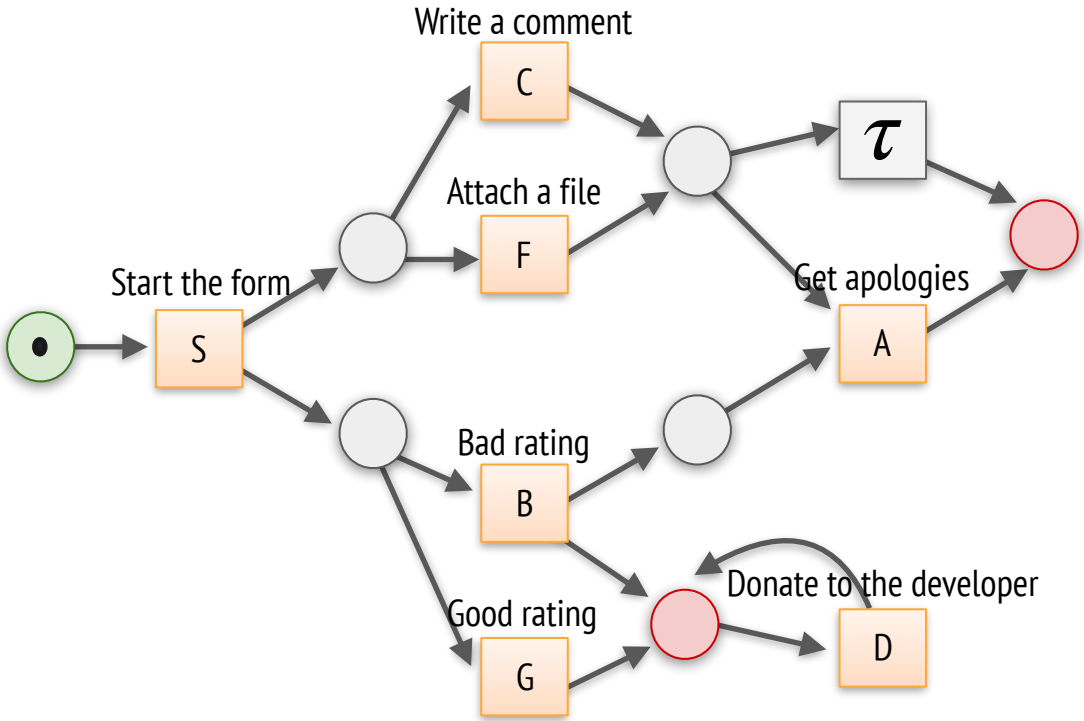
# Data analysis



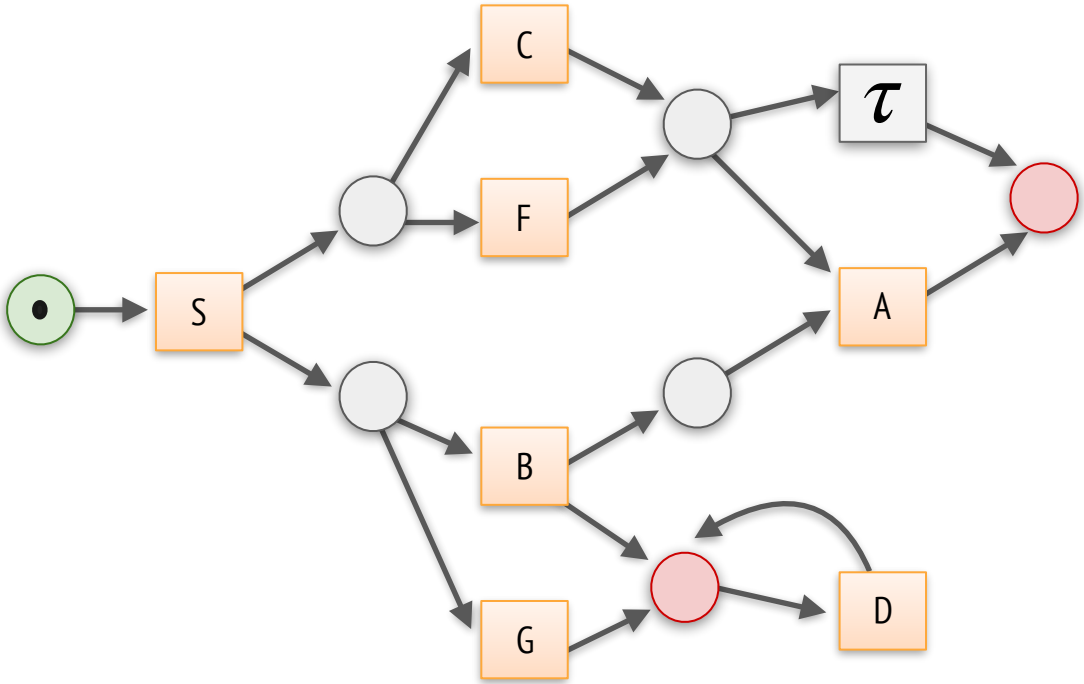
# Data analysis



# Data analysis



# Data analysis



# Data analysis

< Start the form, Give a comment, Good rating >

< Start the form, Give a comment, Good rating Donate to the developer>

< Start the form, Bad rating, Send a File, Get Apologies >

< Start the form, Send a File, Send a File, Get Apologies >

< Start the form, Send a File, Bad rating, Get Apologies >

< Start the form, Good rating, Send a File, Donate to the developer, Donate to the developer, Donate to the developer Donate to the developer >

< Good rating, Give a comment,  Start the form, Donate to the developer, Donate to the developer >

< Start the form, Donate to the developer, Donate to the developer, Donate to the developer >

# Data analysis

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >

# Data analysis

## Log Traces

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >

# Data Clustering

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >

**Data clustering** is the task of grouping objects by similarity.

# Supervised and Unsupervised methods

## Data clustering

- Unsupervised algorithms
- No classes
- Number of clusters is unknown

To mine databases

(k-means, self-organizing map … )

## Data classification

- Supervised algorithms
- Labelled data/ known classes
- Number of classes is known

To respond to defined problems

(k-NN, SVM … )

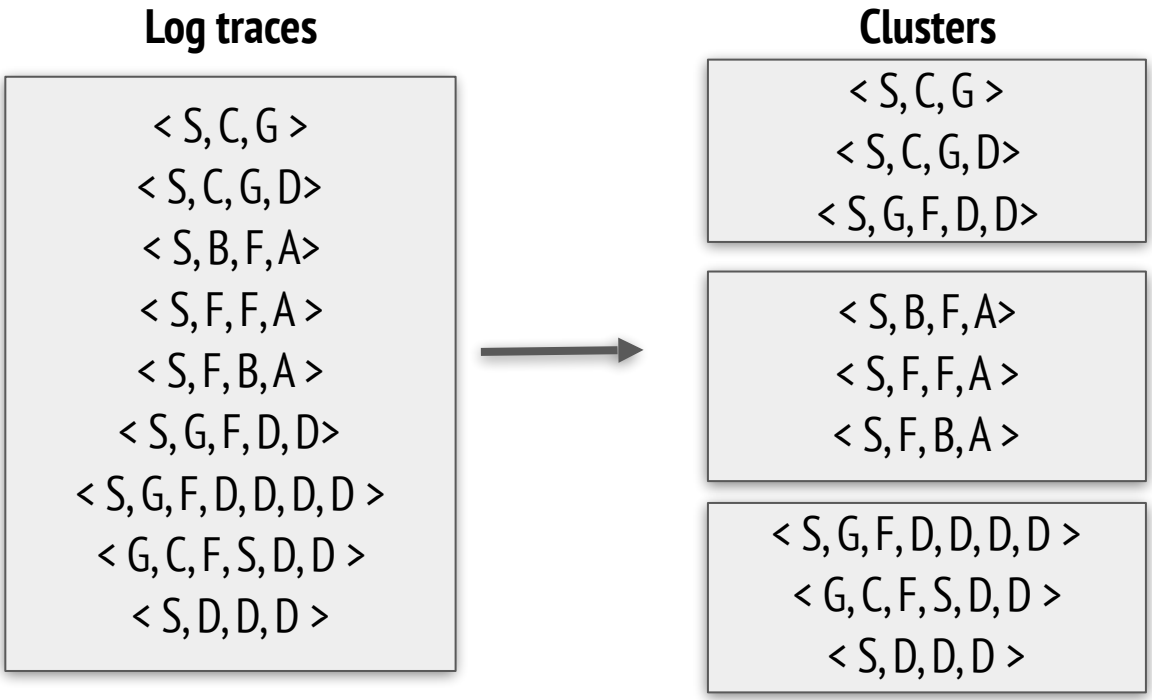
# Trace clustering

**Log traces**

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >

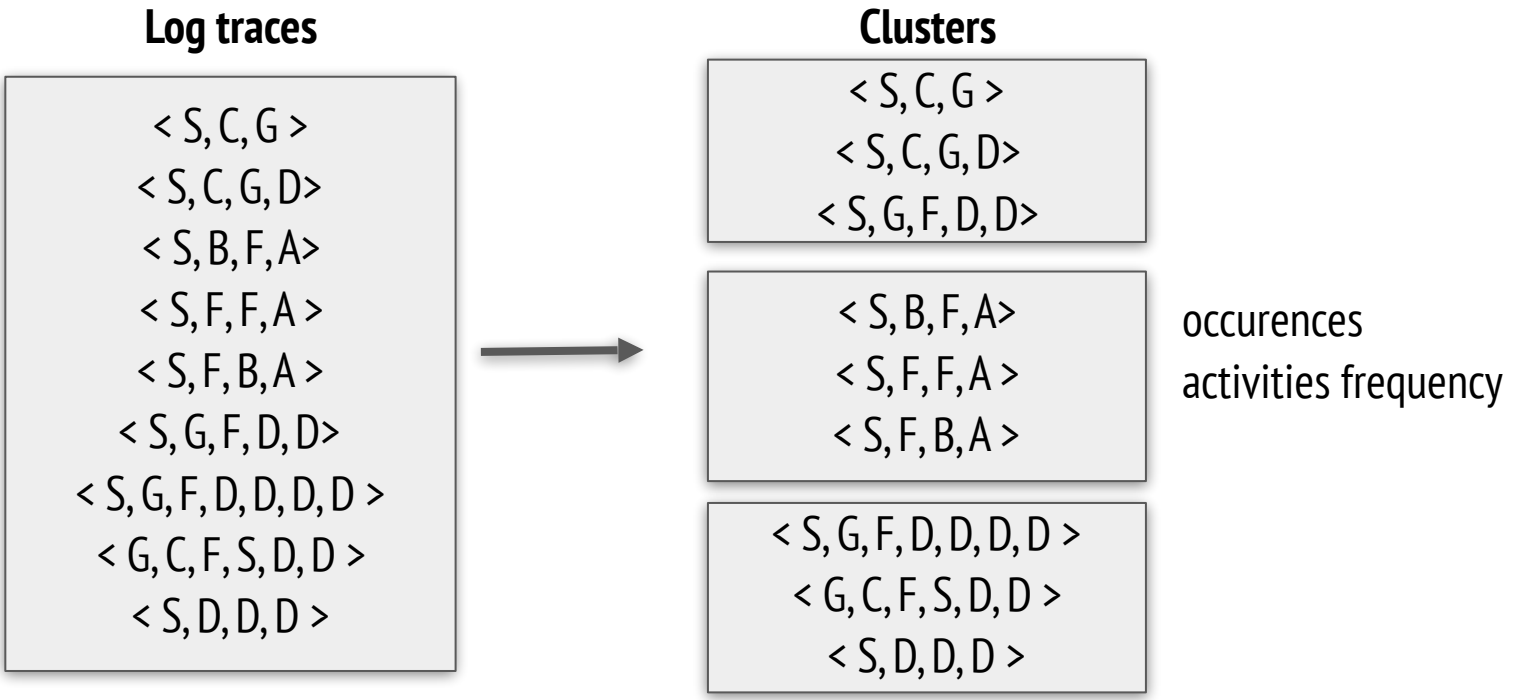
[ Greco 2006 ] ; [ Ferreira 2007 ]

# Trace clustering




[ Greco 2006 ] ; [ Ferreira 2007 ]

# Trace clustering




[ Greco 2006 ] ; [ Ferreira 2007 ]

## Model-based Trace Clustering

**New idea** : to cluster data based on an existing process model

> highlight parts of models that are executed

> show deviating traces

> model repair

[ Chatain 2017 ]

# Alignment-based Trace Clustering

**Log traces**

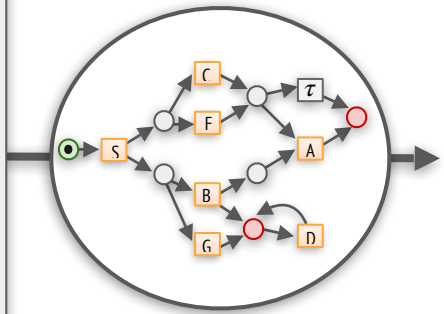
< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >

[ Chatain 2017 ]

# Alignment-based Trace Clustering

**Log traces**

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >

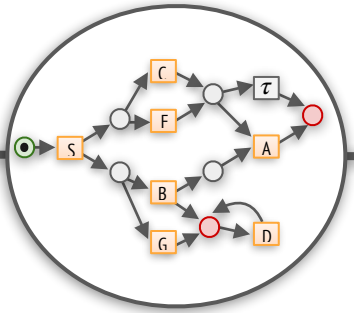



[ Chatain 2017 ]

# Alignment-based Trace Clustering

**Log traces**

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >




**Clusters**

< S, C, G >
< S, C, G, D>

< S, B, F, A>
< S. F. F. A >

< S, F, B, A >

< S, G, F, D, D>

< S, G, F, D, D, D, D >

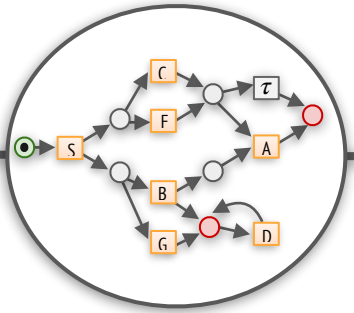
< G, C, F, S, D, D >
< S. D. D. D >

[ Chatain 2017 ]

# Alignment-based Trace Clustering

**Log traces**

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >




**Clusters**

< S, C, G >
< S, C, G, D>

< S, B, F, A>
< S. F. F. A >

< S, F, B, A >

< S, G, F, D, D>

< S, G, F, D, D, D, D >

< G, C, F, S, D, D >
< S. D. D. D >

**Centroids : runs**

< S, C, τ, G >

< S, B, F, A >

< S, F, B, A >

< S, G, F, D, D>

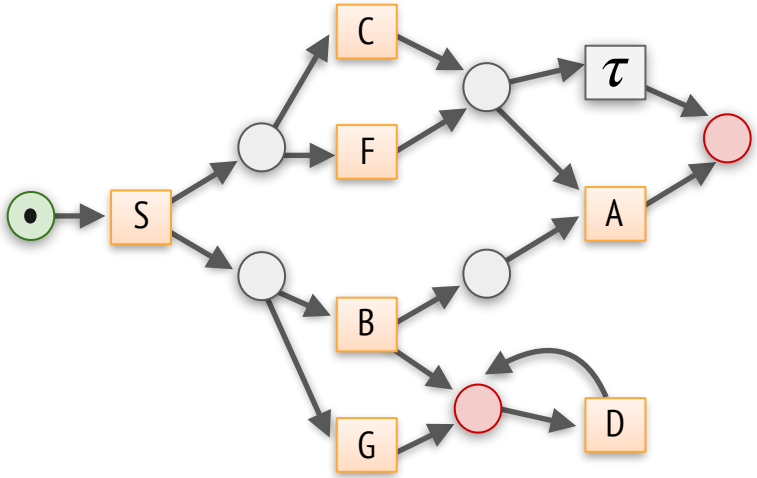
< S, G, F, D, D, D, D >

Non-clustered

[ Chatain 2017 ]

# Alignment-based Trace Clustering

## Full runs as centroids




Example of full run : < S, C, $\tau$ ,G>

[ Chatain 2017 ]

# Alignment-based Trace Clustering

## Full runs as centroids

**N** a process model, **L** a log :

For every trace $\sigma \in \mathbf{L}$,

Find $u \in \text{Runs}(\mathbf{N})$, $dist(\sigma, u)$ is small

u is a centroid

[ Chatain 2017 ]

# Alignment-based Trace Clustering

## Full runs as centroids

**N** a process model, **L** a log :

For every trace $\sigma \in \mathbf{L}$,

Find $u \in$ Runs(**N**), $dist(\sigma,u)$ is small

u is a centroid

Example :

$\sigma 1 = < S, C, G >$

$\sigma 2 = < S, C, G, D >$

[ Chatain 2017 ]

# Alignment-based Trace Clustering

## Full runs as centroids

**N** a process model, **L** a log :

For every trace $\sigma \in \mathbf{L}$,

Find $u \in$ Runs(**N**), $dist(\sigma,u)$ is small

u is a centroid

Example :

$\sigma 1 = < S, C, G >$

$\sigma 2 = < S, C, G, D >$

$\mathbf{u = < S, C, \tau, G >}$

$dist(\sigma 1, u) = 0$

$dist(\sigma 2, u) = 1$

[ Chatain 2017 ]

# Alignment-based Trace Clustering

## Full runs as centroids

**N** a process model, **L** a log :

For every trace $\sigma \in \mathbf{L}$,

Find $u \in$ Runs(**N**), $dist(\sigma,u)$ is small

u is a centroid

Example :

$\sigma 1 = < S, C, G >$

$\sigma 2 = < S, C, G, D >$

$\mathbf{u = < S, C, \tau, G >}$

$dist(\sigma 1, u) = 0$

$dist(\sigma 2, u) = 1$

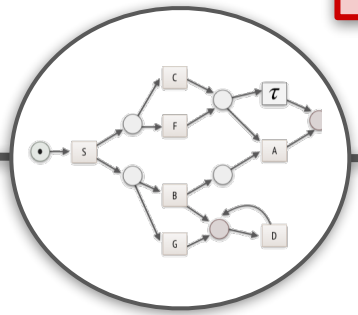
*dist* is a distance between words (Hamming distance, Edit distance..)

[ Chatain 2017 ]

# Alignment-based Trace Clustering

**Log traces**

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >



**Clusters**

The use of full runs as centroids does not allow concurrency

< S, B, F, A>
< S. F. F. A >

< S, F, B, A >

< S, G, F, D, D>

< S, G, F, D, D, D, D >

< G, C, F, S, D, D >
< S. D. D. D >

**troids : runs**

τ, G >

< S, B, F, A >

< S, F, B, A >

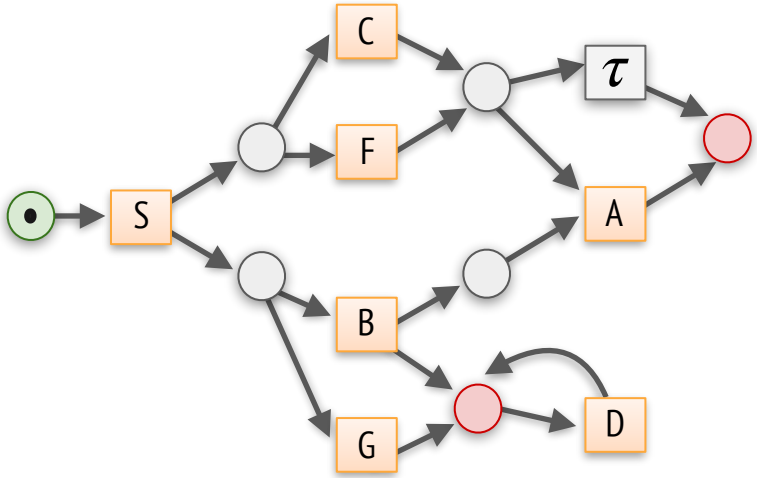
< S, G, F, D, D>

< S, G, F, D, D, D, D >

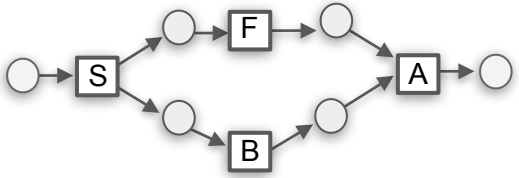
Non-clustered

[ Chatain 2017 ]

## Processes as centroids




Example of process :




[ Boltenhagen 2019 ]

Linearizations of the process :

< S, B, F, A >

< S, F, B, A >

**Log traces**

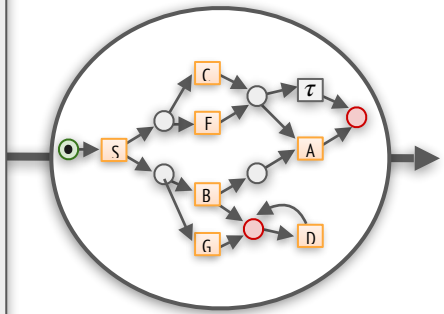
< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >

[ Boltenhagen 2019 ]

# Alignment and Model Partial-Order-based Trace Clustering

**Log traces**

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >

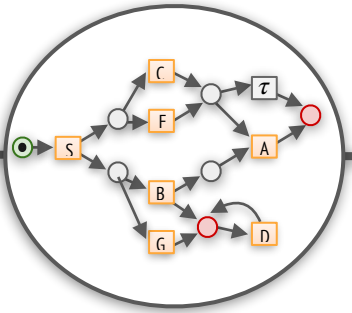



[ Boltenhagen 2019 ]

# Alignment and Model Partial-Order-based Trace Clustering

**Log traces**

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >




**Clusters**

< S, C, G >
< S, C, G, D>

< S, B, F, A>
< S, F, F, A >
< S, F, B, A >

< S, G, F, D, D>

< S, G, F, D, D, D, D >

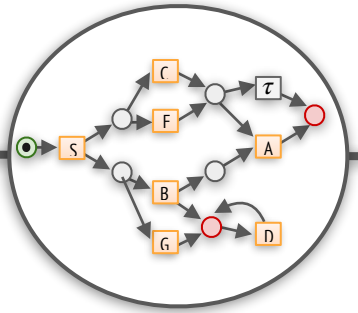
< G, C, F, S, D, D >
< S, D, D, D >

[ Boltenhagen 2019 ]

# Alignment and Model Partial-Order-based Trace Clustering

**Log traces**

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >




**Clusters**

< S, C, G >
< S, C, G, D>

< S, B, F, A>
< S, F, F, A >
< S, F, B, A >

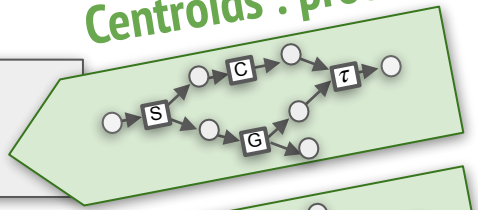
< S, G, F, D, D>

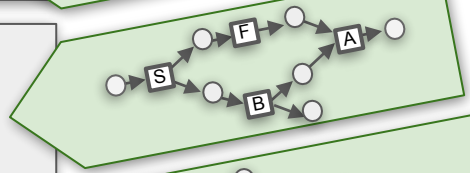
< S, G, F, D, D, D, D >

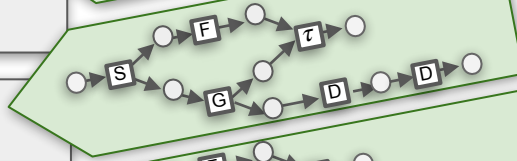
< G, C, F, S, D, D >
< S, D, D, D >

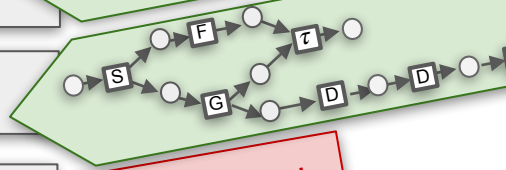
Non-clustered

**Centroids : processes**




[ Boltenhagen 2019 ]

# Alignment and Model Partial-Order-based Trace Clustering

**N** a process model, **L** a log :

For every trace $\sigma \in \mathbf{L}$,

Find $P \in$ Processes(**N**), $dist(\sigma, P)$ is small

P is a centroid

[ Boltenhagen 2019 ]

# Alignment and Model Partial-Order-based Trace Clustering

**N** a process model, **L** a log :

For every trace $\sigma \in \mathbf{L}$,

Find $P \in$ Processes(**N**), $dist(\sigma, P)$ is small

P is a centroid

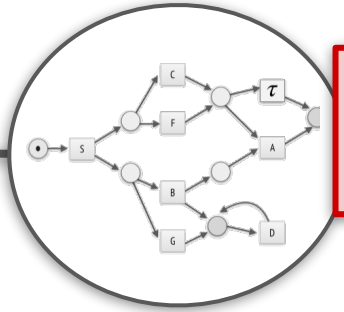
*dist* is the minimal distance between a linearization of P and the trace (computed as distance between words : Hamming distance, Edit distance..)

[ Boltenhagen 2019 ]

# Alignment and Model Partial-Order-based Trace Clustering

**Log traces**

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >



**Clusters**

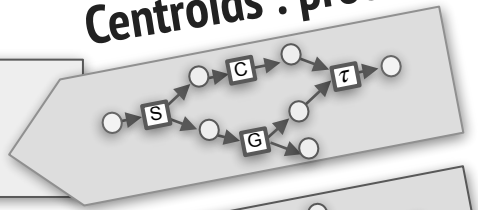
< S, C, G >
< S, C, G, D>

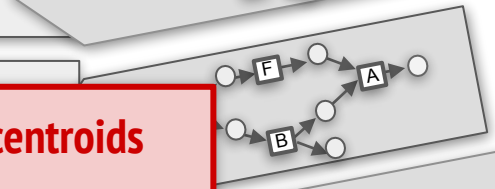
< S, G, F, D, D>

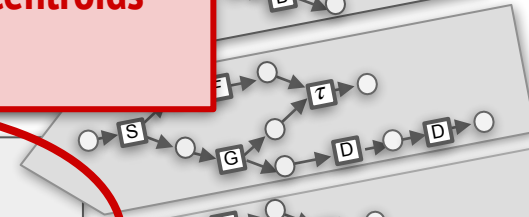
< S, G, F, D, D, D, D >

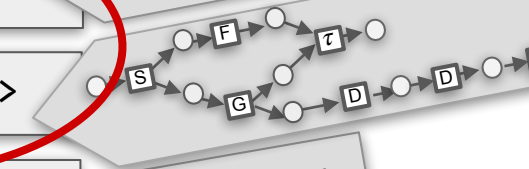
< G, C, F, S, D, D >
< S. D. D. D >

**Centroids : processes**





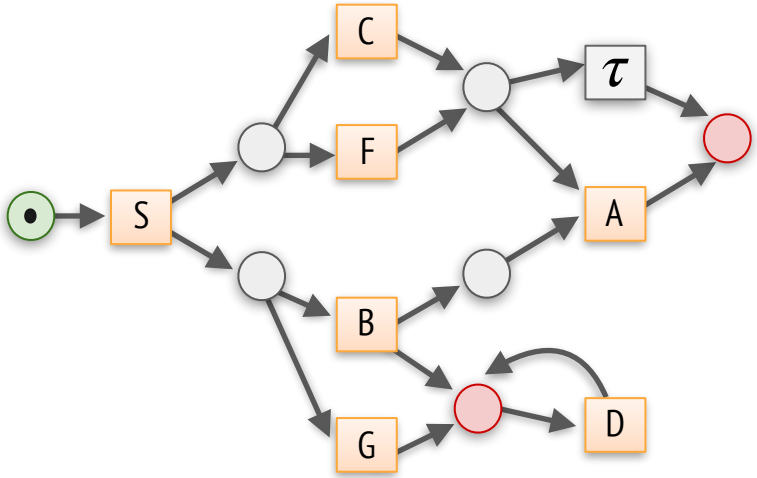




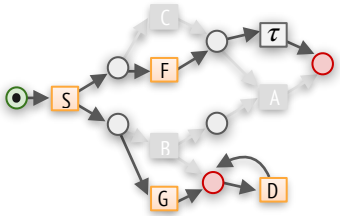
Non-clustered

**The use of processes as centroids does not allow loops**

[ Boltenhagen 2019 ]




Example of subnet :




[ Boltenhagen 2019 ]

# Alignment and Model Subnet-based Trace Clustering

**Log traces**

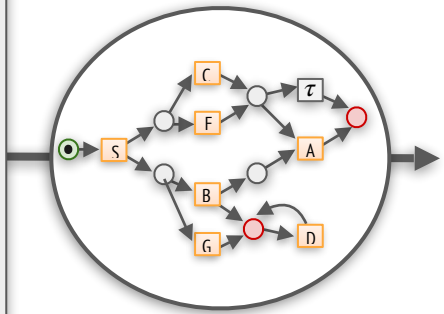
< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >

[ Boltenhagen 2019 ]

# Alignment and Model Subnet-based Trace Clustering

**Log traces**

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >

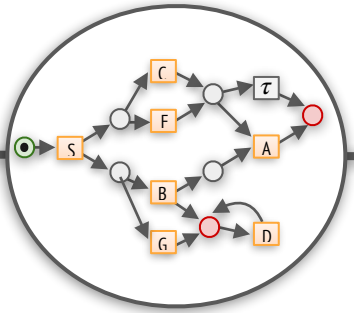



[ Boltenhagen 2019 ]

# Alignment and Model Subnet-based Trace Clustering

**Log traces**

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >




**Clusters**

< S, C, G >
< S, C, G, D>

< S, B, F, A>
< S, F, F, A >
< S, F, B, A >

< S, G, F, D, D>
< S, G, F, D, D, D, D >

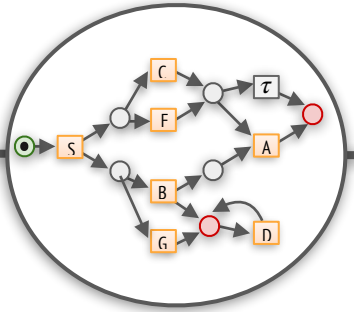
< G, C, F, S, D, D >
< S. D. D. D >

[ Boltenhagen 2019 ]

# Alignment and Model Subnet-based Trace Clustering

**Log traces**

< S, C, G >
< S, C, G, D>
< S, B, F, A>
< S, F, F, A >
< S, F, B, A >
< S, G, F, D, D>
< S, G, F, D, D, D, D >
< G, C, F, S, D, D >
< S, D, D, D >




**Clusters**

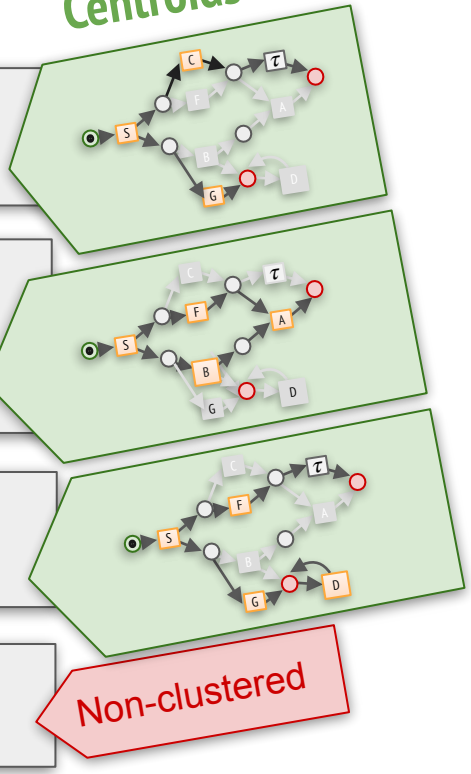
< S, C, G >
< S, C, G, D>

< S, B, F, A>
< S, F, F, A >
< S, F, B, A >

< S, G, F, D, D>
< S, G, F, D, D, D, D >

< G, C, F, S, D, D >
< S. D. D. D >

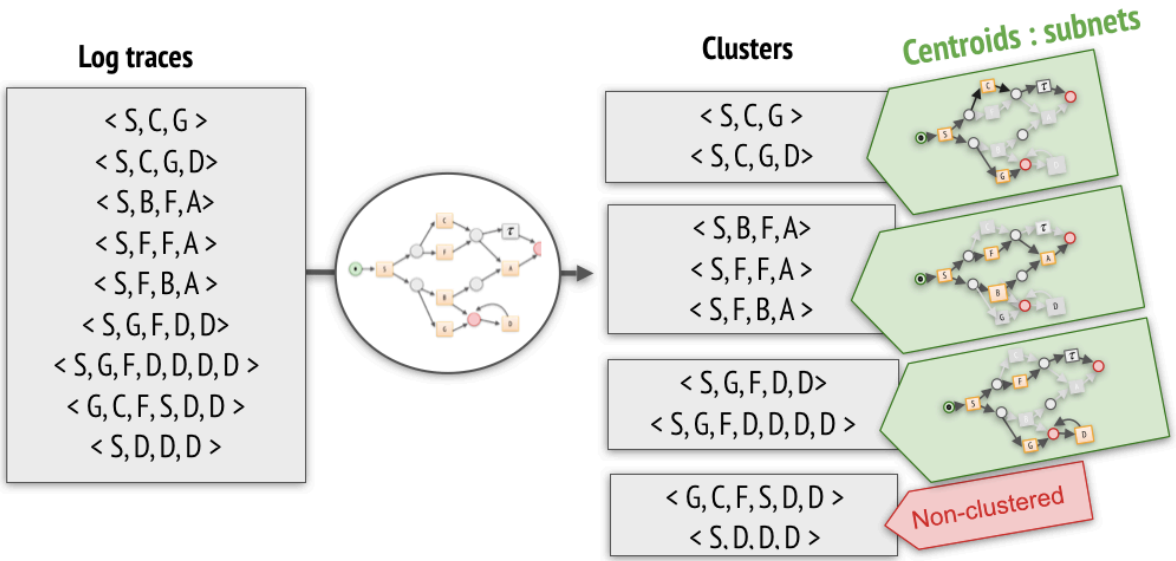
**Centroids : subnets**




Non-clustered

[ Boltenhagen 2019 ]

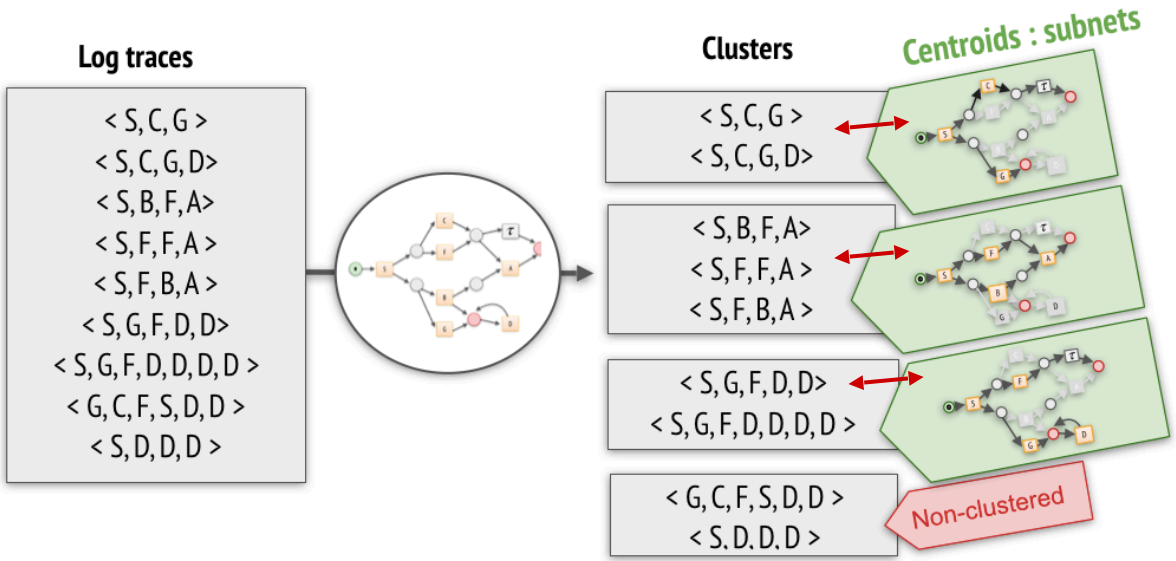
# Quality criteria




[ Boltenhagen 2019 ]

# Quality criteria

> Distance between traces and their centroid

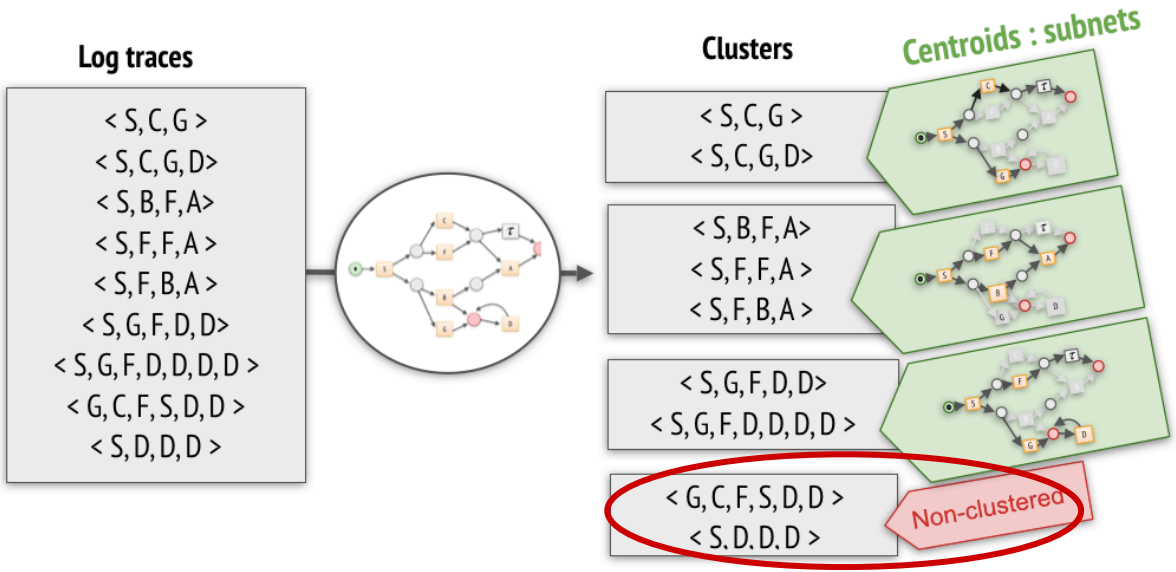



[ Boltenhagen 2019 ]

# Quality criteria

> Distance between traces and their centroid

> Number of non-clustered traces



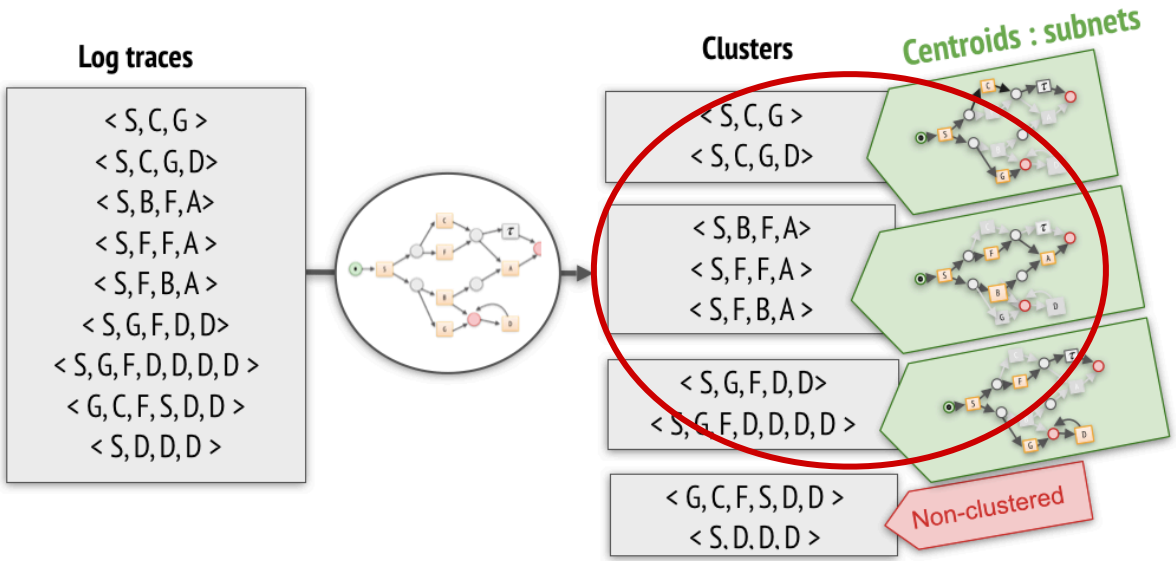

[ Boltenhagen 2019 ]

# Quality criteria

> Distance between traces and their centroid

> Number of non-clustered traces

> Number of clusters




[ Boltenhagen 2019 ]

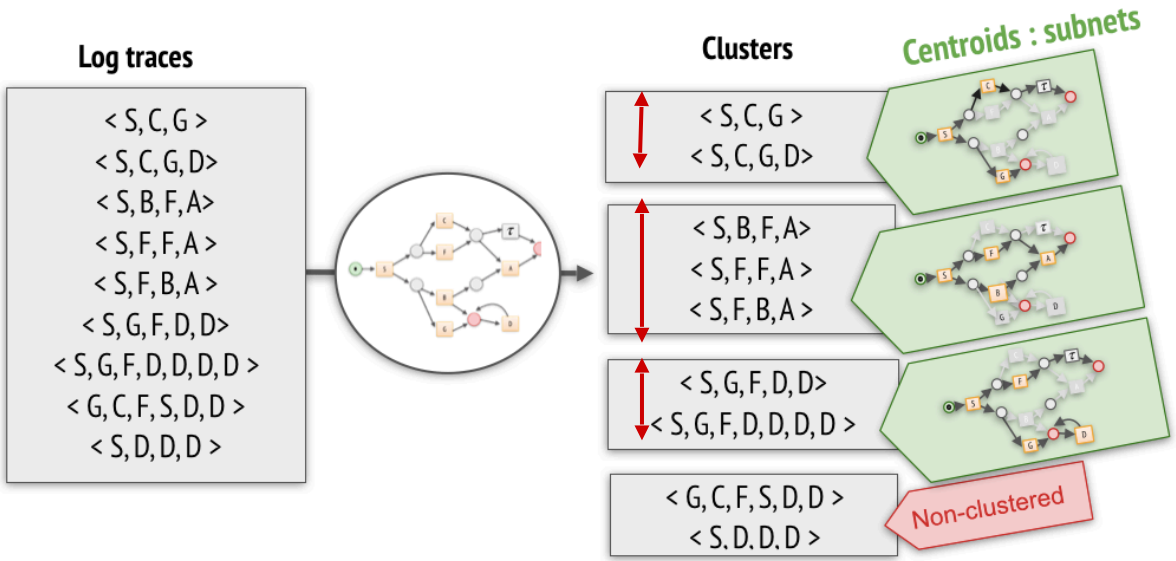
# Quality criteria

> Distance between traces and their centroid

> Number of non-clustered traces

> Number of clusters

> Intra-cluster distance




[ Boltenhagen 2019 ]

# Quality criteria

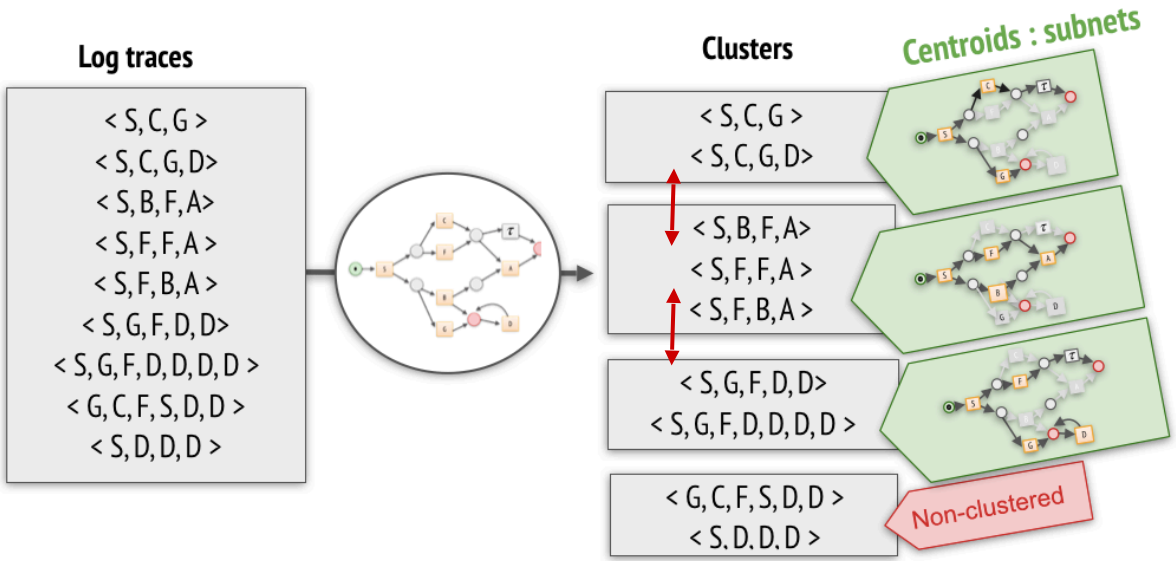
> Distance between traces and their centroid

> Number of non-clustered traces

> Number of clusters

> Intra-cluster distance

> Inter-cluster distance




[ Boltenhagen 2019 ]

> Tool DARK SIDER*

> SAT formulas

> Optimal clusterings

*https://github.com/BoltMaud/darksider

# Model Repair

How can one repair an existing process model ?

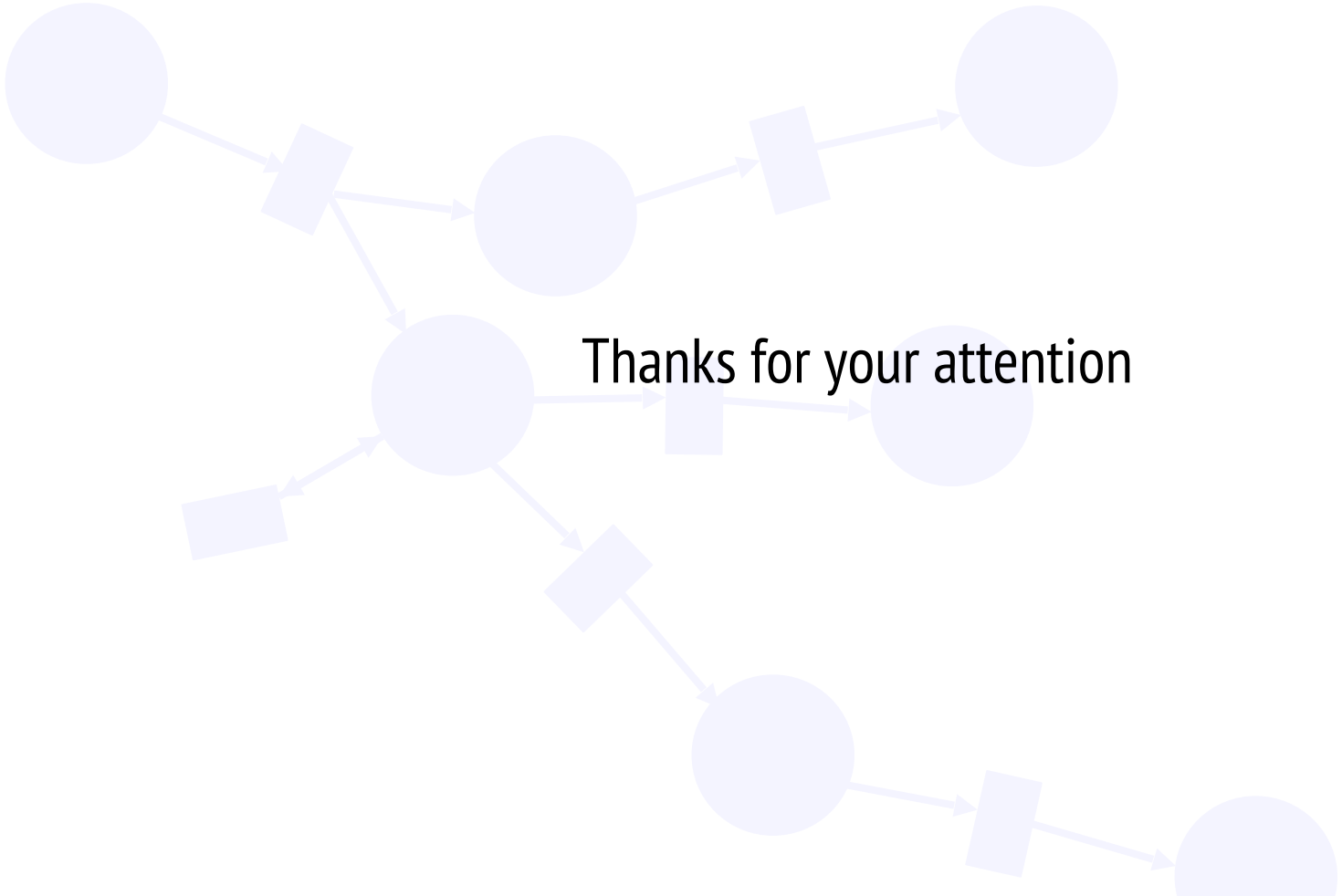
## Model Repair

At least a better fitness

Not too far from the original model

How can one repair an existing process model ?

....



Thanks for your attention