Process Mining
Generalized Alignment-Based Trace Clustering

Mathilde Boltenhagen[1]

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

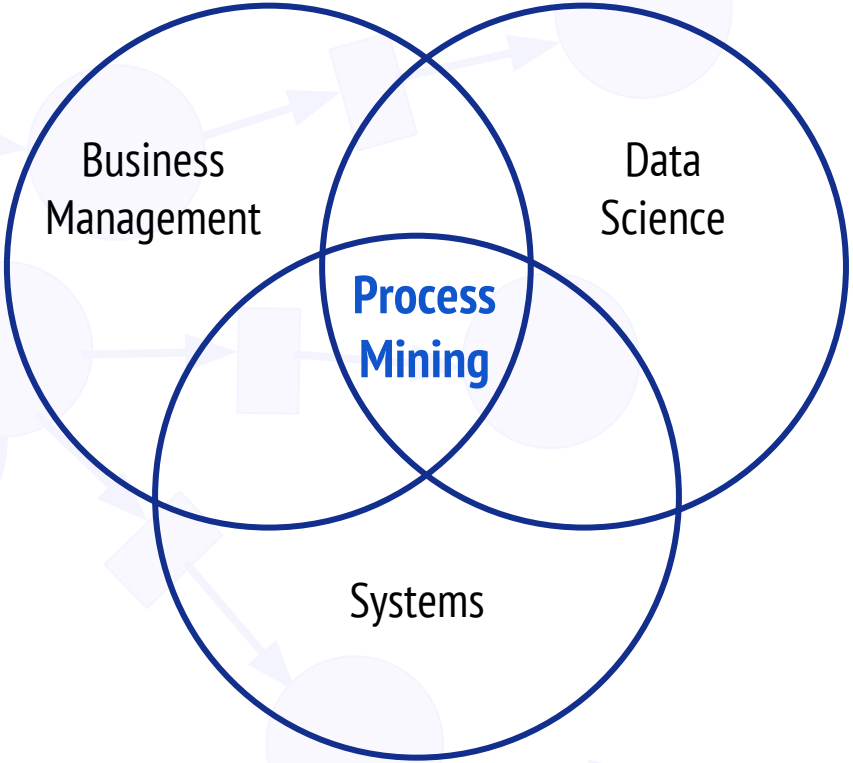
CIRM - 5 mars 2019

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

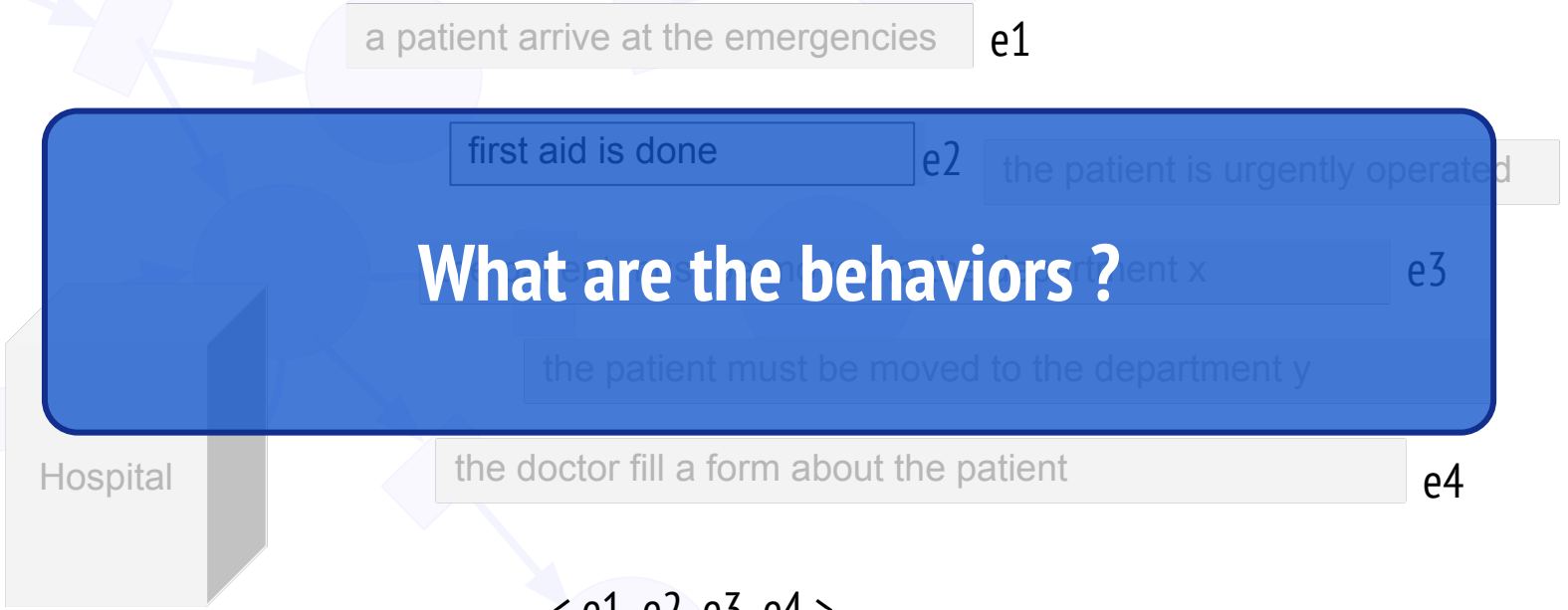
Sommaire

- Preliminaries
- State-of-the-Art
  - Main research topics
  - Quality of models
  - Trace Clustering
- Contribution
- Opening

Position of the field



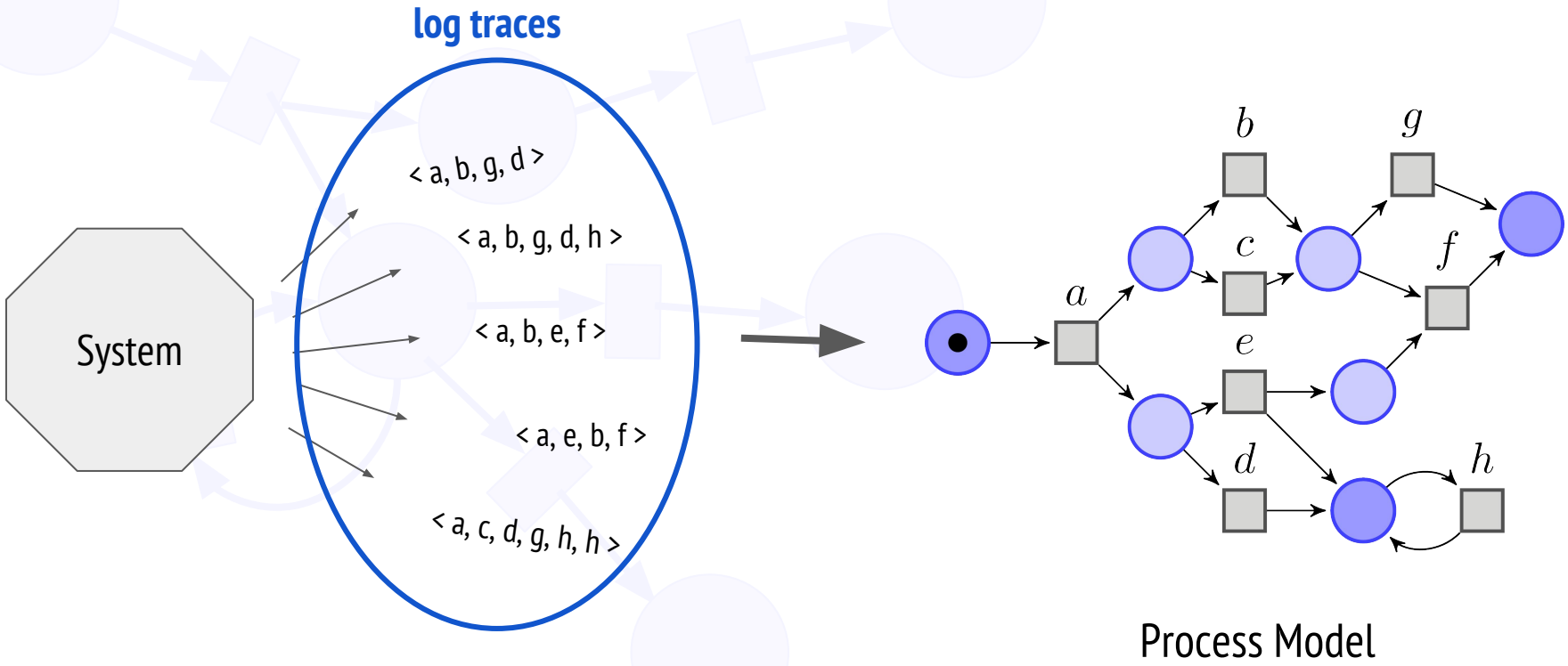
Example



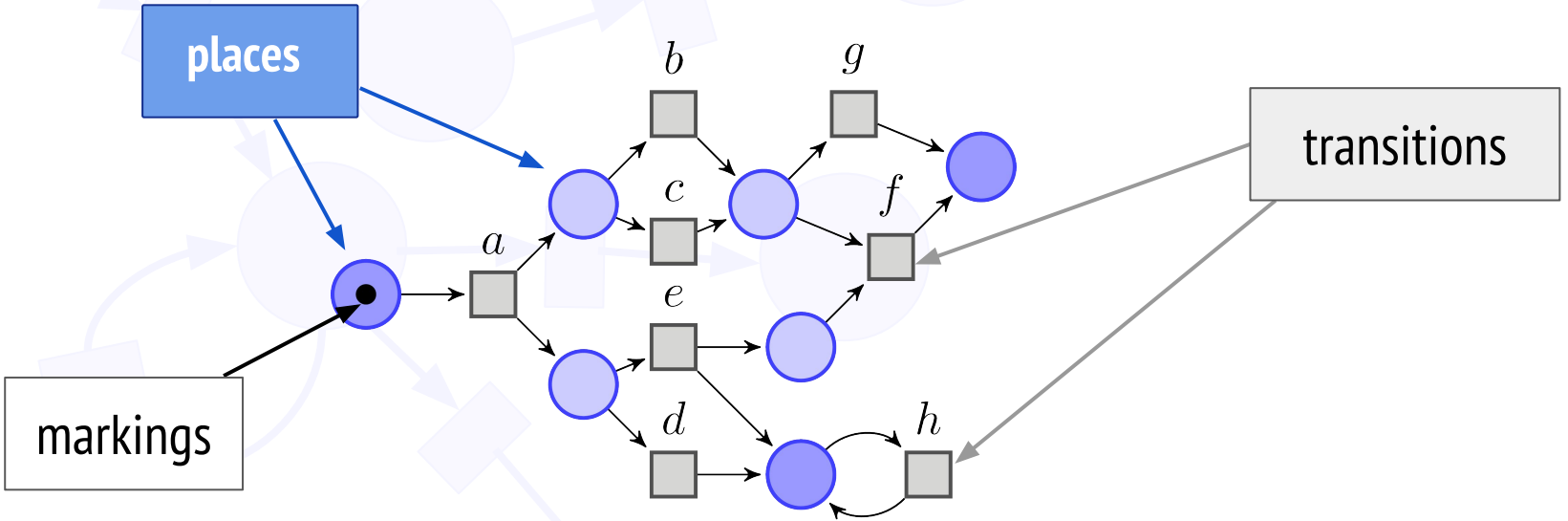
What are the behaviors ?

< e1, e2, e3, e4 >

Process Model Discovery



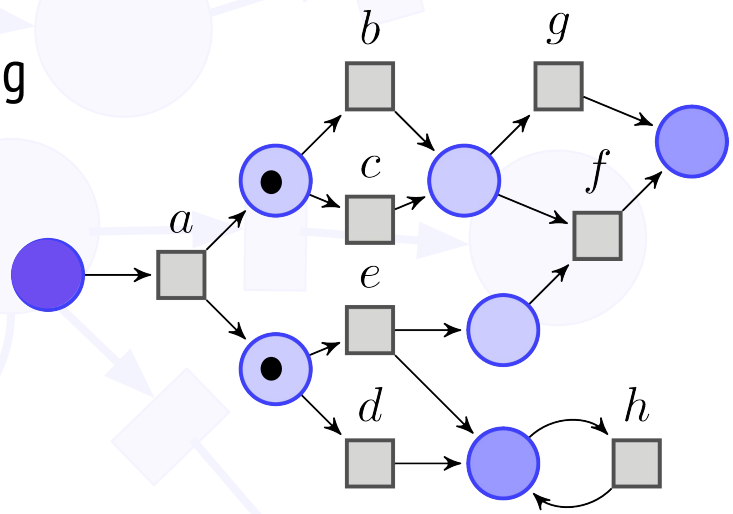
Petri Nets



Token game

> Firing transition

> Reachable marking



Main research problems in Process Mining

> Process model discovery

> Conformance checking

> Model repair

Quality criteria

How good is the discovered model ?

> Fitness

*Has the recorded behavior been modelled ?**

> Precision

*Has the modelled behavior been recorded ?**

> Generalization

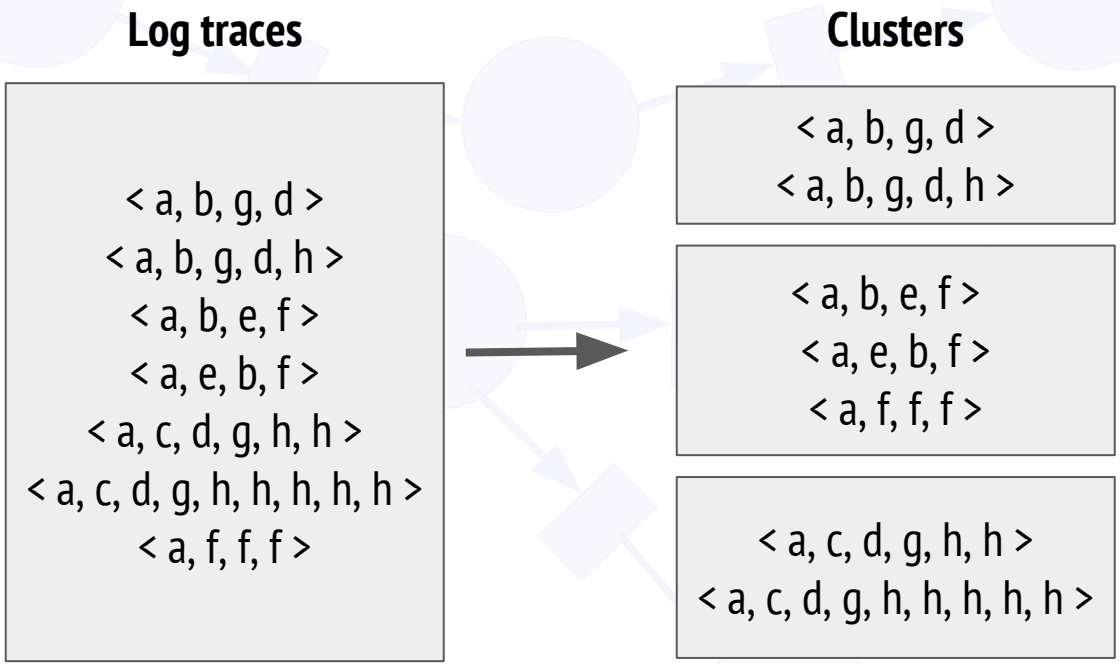
Will the model fit future behaviors ?

> Simplicity

*How complex is the model ?**

[Buijs 2012] ; [Carmona 2018]

Trace clustering



> Data analysis

> Model discovery

[Greco 2006] ; [Ferreira 2007]

Model-based Trace Clustering

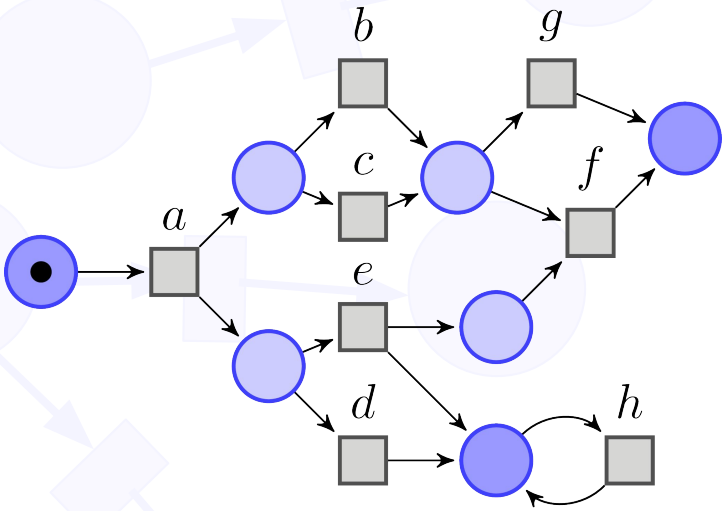
Idea : to cluster data with the existing model

> highlight parts of models that are executed

> show deviated traces

> model repairs

Full Runs



Example of full run : < a, b, d, g, h >

Full Runs as Centroids

Centroids (full runs)	Traces	Distances
< a, b, g, d >	< a, b, g, d >	0
	< a, b, g, d, h >	1
< a, b, e, f >	< a, b, e, f >	0
< a, e, b, f >	< a, e, b, f >	0
< a, c, d, g, h, h >	< a, c, d, g, h, h >	0
< a, c, d, g, h, h, h, h >	< a, c, d, g, h, h, h, h >	0

Non-clustered	< a, f, f, f >

[Chatain 2017]

Alignment-based Trace Clustering

Limits

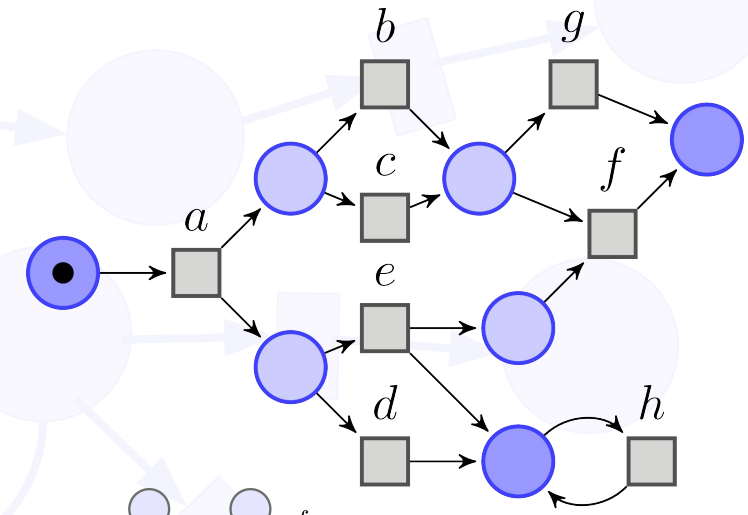
Centroids (full runs)	Traces	Distances
< a, b, g, d >	< a, b, g, d >	0
	< a, b, g, d, h >	1
< a, b, e, f >	< a, b, e, f >	0
< a, e, b, f >	< a, e, b, f >	0
< a, c, d, g, h, h >	< a, c, d, g, h, h >	0
< a, c, d, g, h, h, h, h >	< a, c, d, g, h, h, h, h >	0
Non-clustered	< a, f, f, f >	

concurrency

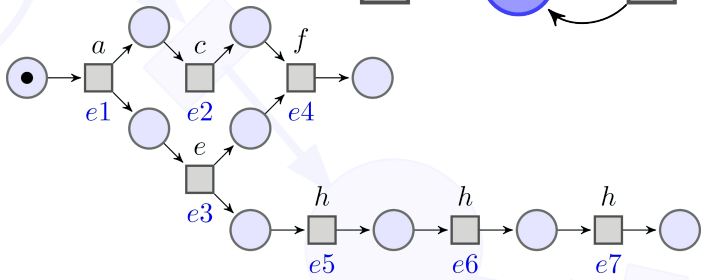
loops

[Chatain 2017]

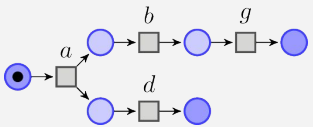
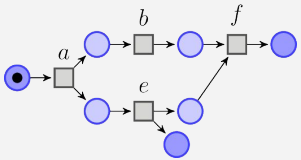
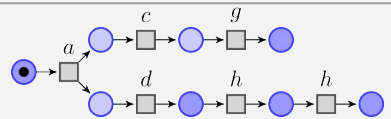
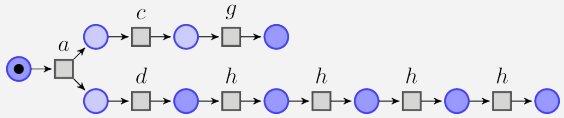
Processes



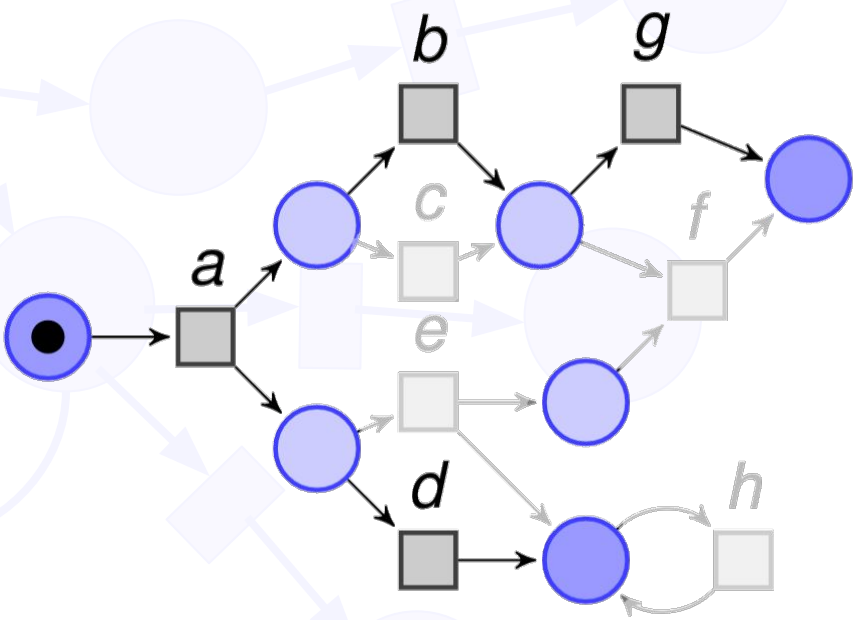
Example of process :



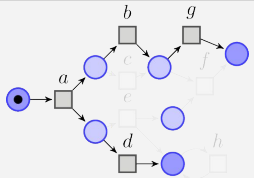
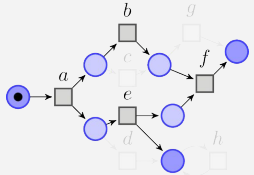
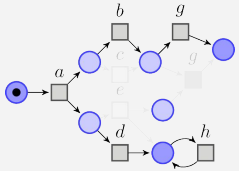
Processes as Centroids

Centroids (process)	Traces	Distances
	< a, b, g, d >	0
	< a, b, g, d, h >	1
	< a, b, e, f >	0
	< a, e, b, f >	0
	< a, c, d, g, h, h >	0
	< a, c, d, g, h, h, h, h >	0
Non-clustered	< a, f, f, f >	

Subnets



Subnets as Centroids

Centroids (subnets)	Traces	Distances
	< a, b, g, d >	0
	< a, b, g, d, h >	1
	< a, b, e, f >	0
	< a, e, b, f >	0
	< a, c, d, g, h, h >	0
	< a, c, d, g, h, h, h, h >	0
Non-clustered	< a, f, f, f >	

Quality criteria

> Distance between traces and centroids

> Number of non-clustered traces

> Number of clusters

> Intra-cluster distance

> Inter-cluster distance

Pseudo-SAT

> Tool *DARK SIDER**

> Optimal clusterings

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

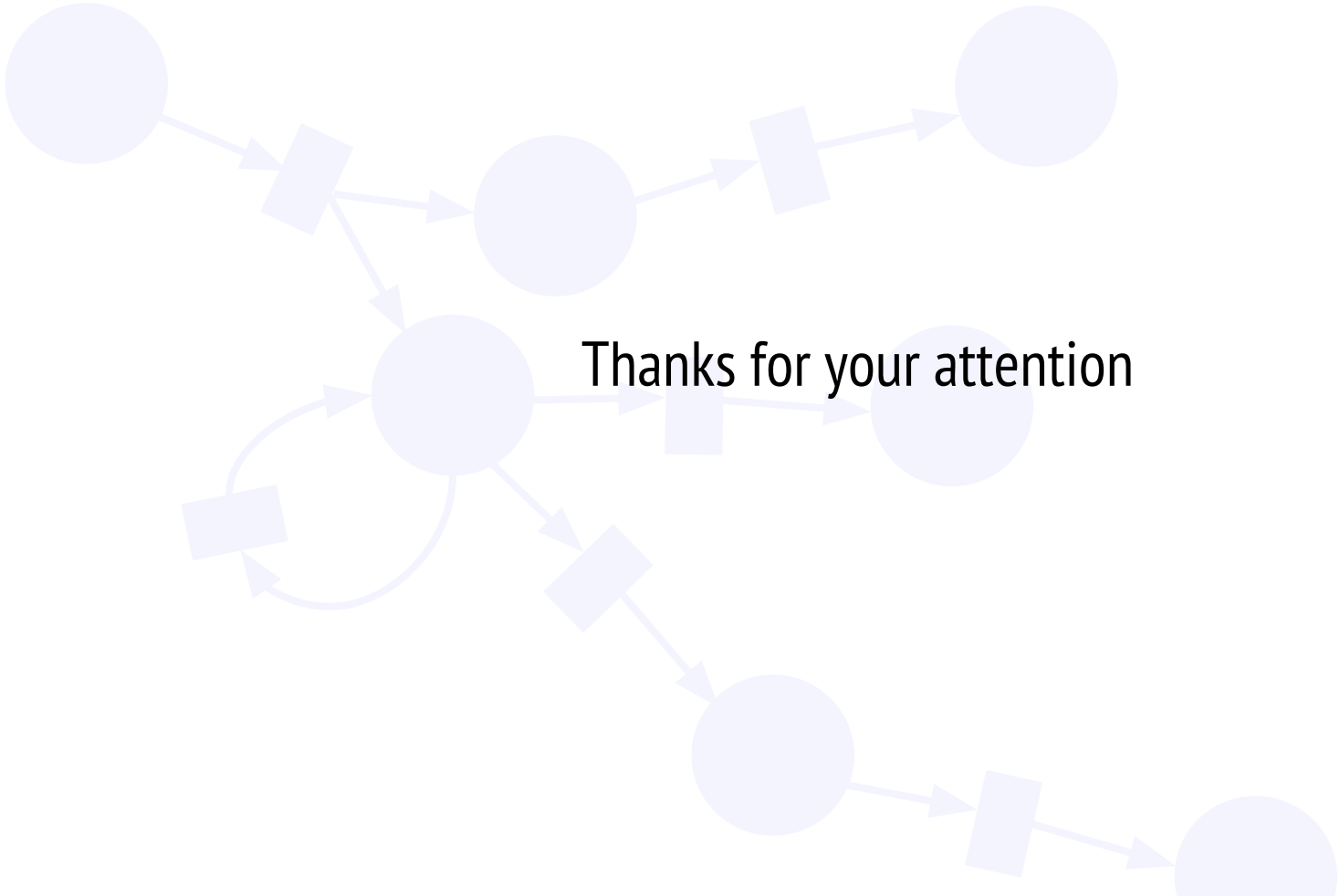
Model Repair

At least a better fitness

Not too far from the original model

How can one repair an existing process model ?

....



Thanks for your attention