Discovery of Resource Working Calendars from Process Event Log

Master's Degree Course in Computer Science

Candidate: Alessandro Pegoraro (ID: 1240466)
Supervisor: Prof. Massimiliano de Leoni

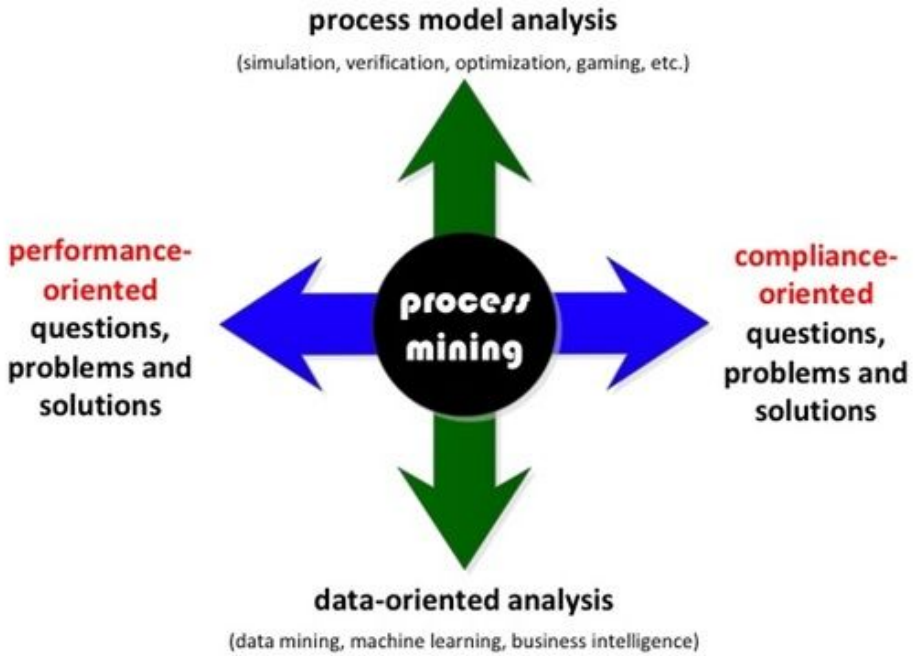
April 21, 2022

Università degli Studi di Padova

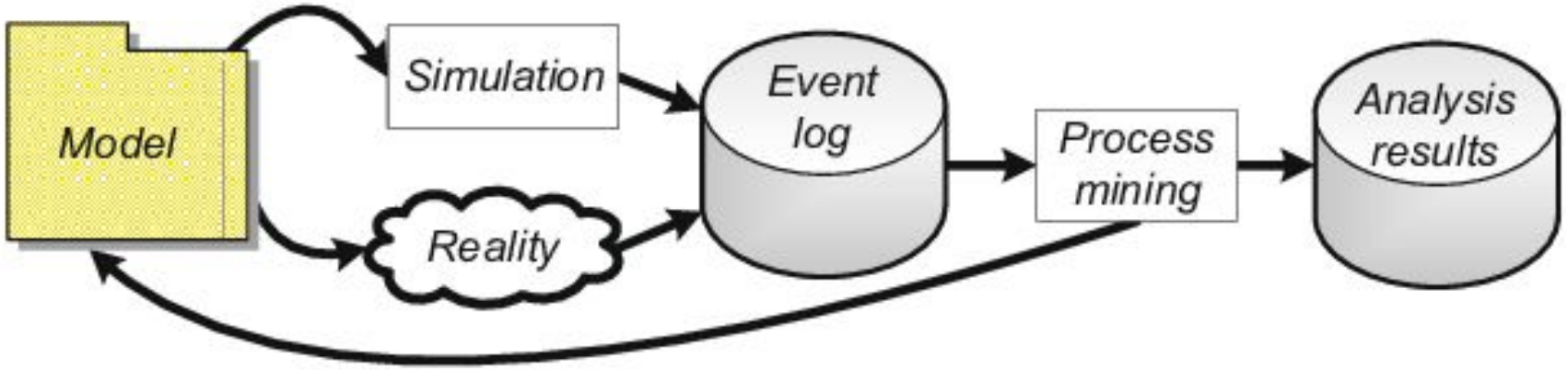
DIPARTIMENTO MATEMATICA
Dipartimento di Matematica "Tullio Levi-Civita"

Introduction - Process Mining

- **Process Mining** is a research field that sits between machine learning and data mining
- Its aim is to **discover**, **monitor** and **improve** real business processes by extracting knowledge from event log
- It employees **information systems** to collect and record all the data required to initiate its techniques

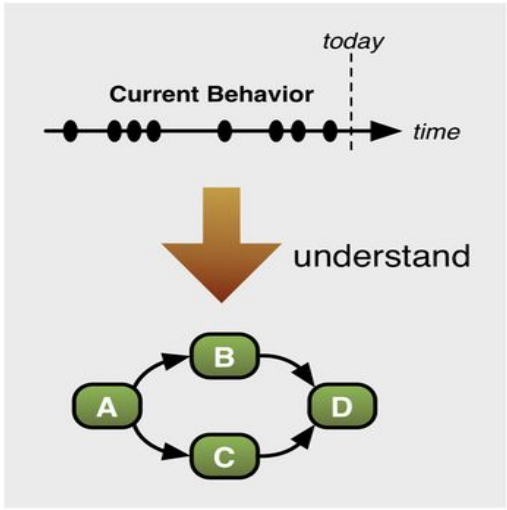


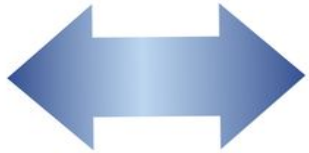
Introduction - Business Process Simulation

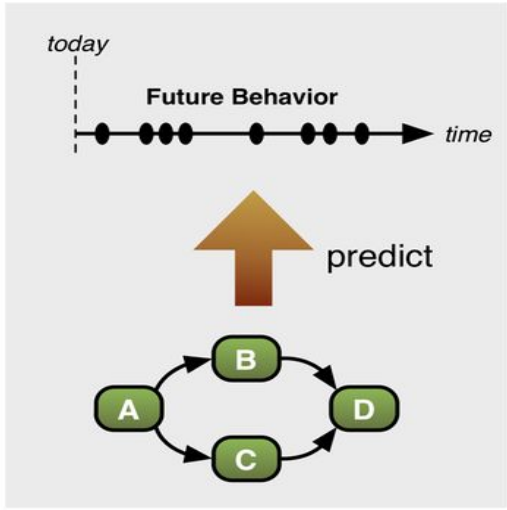


- **Business process simulation** are techniques for the simulation of business process behavior on the basis of a **simulation model**
- **Simulation** provides a flexible approach to **analyse** and **improve** business processes, it **cannot** prove correctness!
- The main idea is to **repeat many times** the simulations to pinpoint the aspects that are more critical and require improvements

Introduction







Process Mining:
Generates models to understand current behavior

Simulation:
Predicts future behavior based on models

- **Problem:**
  - Process model that are manually designed by an analyst **may not capture all dependencies and patterns** of the process
- **Solution:**
  - **Automatically discover** simulation models from business process execution logs (also known as **event logs**)

Preliminaries - Event Log

patient	activity	timestamp	doctor	age	cost
5781	make X-ray	23-1-2014@10.30	Dr. Jones	45	70.00
5541	blood test	23-1-2014@10.18	Dr. Scott	61	40.00
5833	blood test	23-1-2014@10.27	Dr. Scott	24	40.00
5781	blood test	23-1-2014@10.49	Dr. Scott	45	40.00
5781	CT scan	23-1-2014@11.10	Dr. Fox	45	1200.00
5833	surgery	23-1-2014@12.34	Dr. Scott	24	2300.00
5781	handle payment	23-1-2014@12.41	Carol Hope	45	0.00
5541	radiation therapy	23-1-2014@13.57	Dr. Jones	61	140.00
5541	radiation therapy	23-1-2014@13.08	Dr. Jones	61	140.00
…	…	…	…	…	…

case id | **activity name** | **timestamp** | **resource** | **other data**

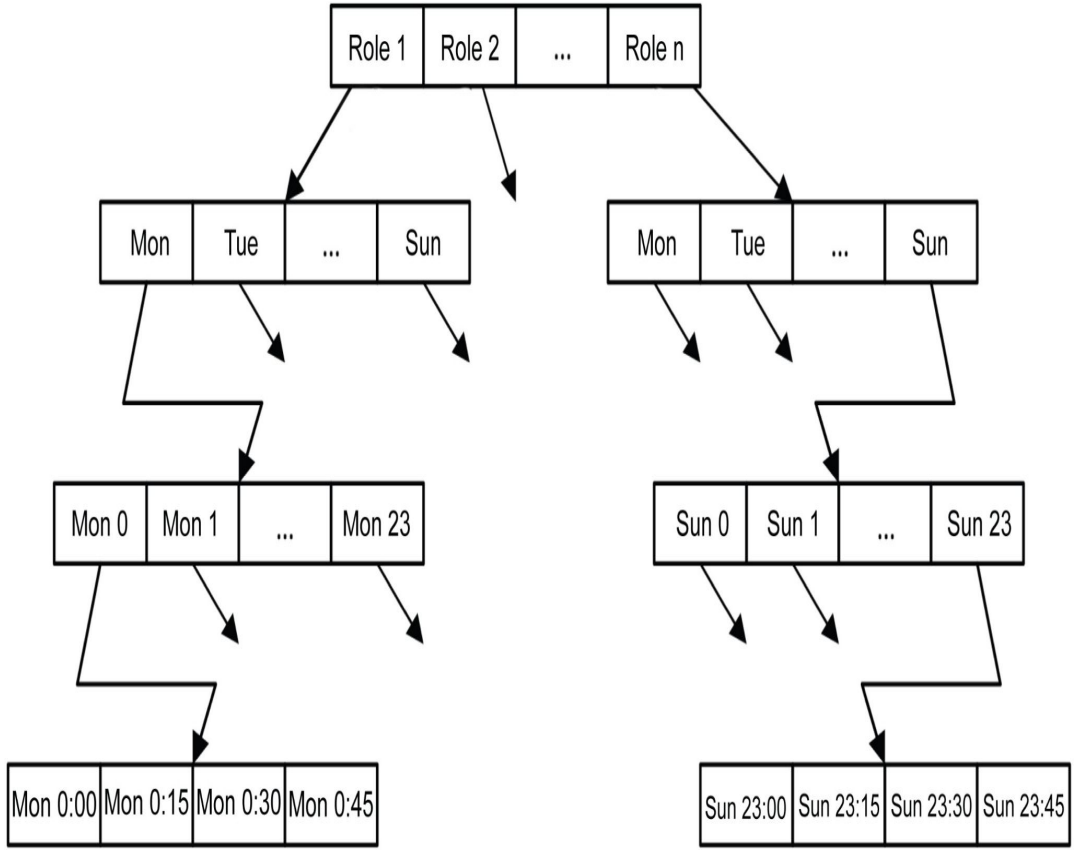
- Represents a register of **past executions of a process**
- Every operation is recorded as an **event**, which represents one specific execution of an activity

Preliminaries - Working Calendar

- Describe the **availability constraints** of resource pool
- A resource pool is simply a set of roles/resources that can **perform a given activity**
- Each calendar is associated to a specific role/resource
- **Availability constraints** specifies the periods of the day, week, month and year in which a role/resource **can perform** a task
- With **work-shifts** we refer to the possibility of a role/resource to work part-time with different **shifts**
  (on some weeks it works in the morning, while, on other weeks it works in the afternoon)

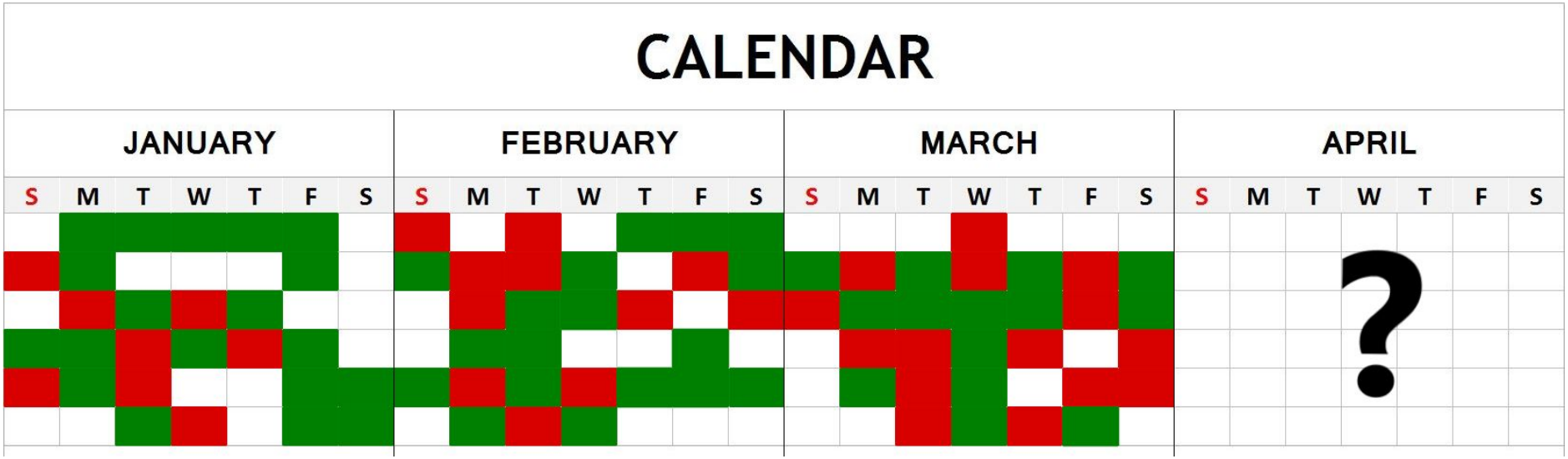
Discovering business process simulation models in the presence of multitasking and availability constraints

- Chooses the tuple of more frequent **time granules** to create **calendar expression**
- Does not consider the possibility of **work-shifts**
- For each possible **calendar expression** it has to check **every** related case in the event log



Related Works

Retrieving Resource Availability Insights from Event Log



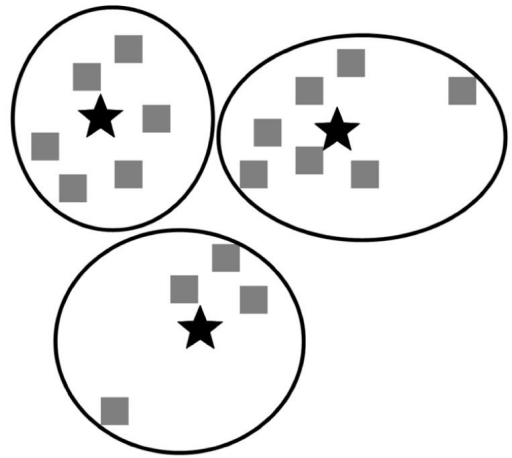
- Defines the **availability** and **unavailability** of resources **only** for the dates in the event log
- No **general calendar** for resources

Related Works

Retrieving the resource availability calendars of a process from an event log

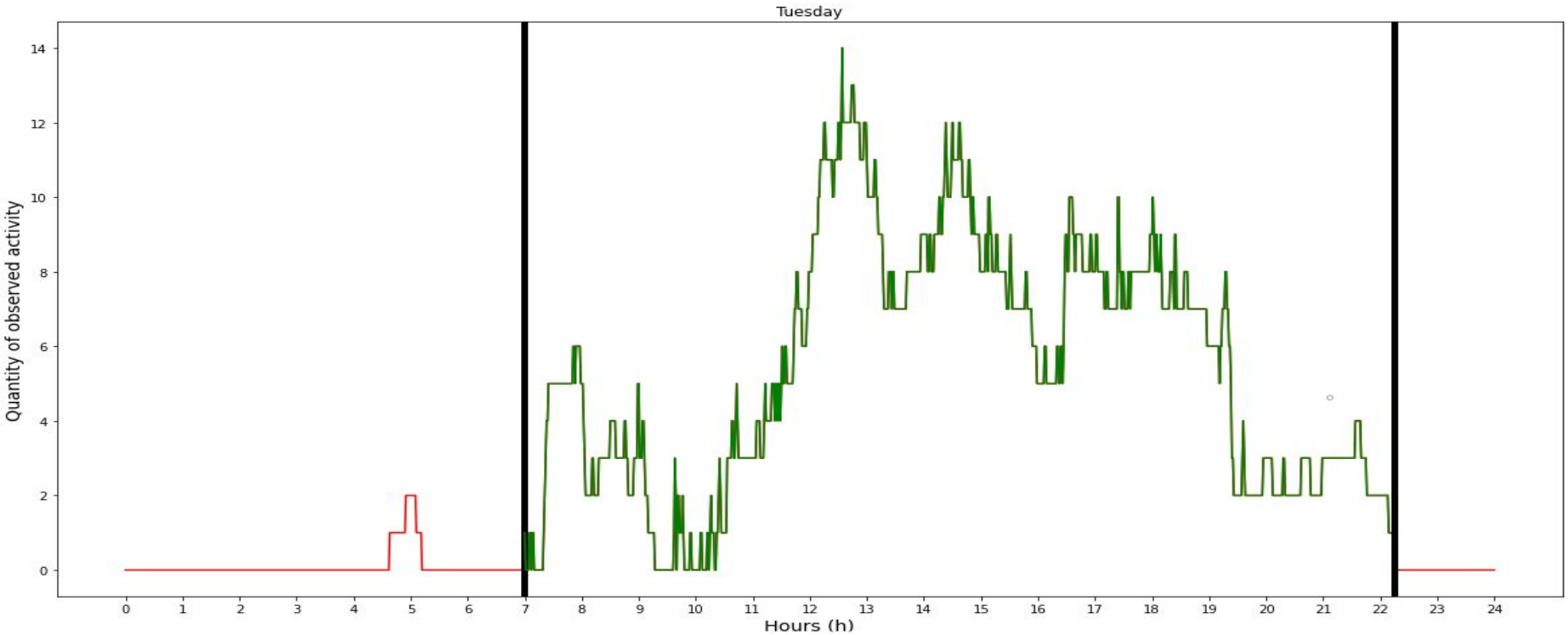
- Propose two methods to obtain **general calendar** for resources from the calendars of only the dates in the event log
- Both approaches rely on a **random draw**
- In the presence of work-shifts a random draw could lead to **skewed calendars**

(a) Direct sampling



(b) Cluster-based sampling

Algorithms - Calendar Discovery for Roles



- Tries to find and remove activities **occurring outside the usual work time** (noise), dependent on two parameters:
- **Threshold:** minimum quantity of activity not considered noise
- **Tolerance:** maximum distance between activities

Algorithms - Calendar Discovery for Roles

- Parameter optimization of **Threshold** and **Tolerance** aiming at the maximization of γ

$$\gamma = 2 * \frac{precision * recall}{precision + recall} - numerosity + calendar\ size$$

  - **Precision:** percentage of activities that are covered by the calendar and are registered in the event log
  - **Recall:** percentage of activities in the event log that completely belongs to the intervals composing the calendar
  - **Numerosity:** number of intervals composing the calendar
  - **Calendar Size:** size of the intervals composing the calendar

Algorithms - Work-Shift Discovery

Role 1							Resource 4						
Monday	:	13:50	-	20:42			Monday	:	14:00	-	20:31		
Tuesday	:	13:47	-	20:34			Tuesday	:	14:01	-	20:34		
Wednesday	:	13:45	-	20:37			Wednesday	:					
Thursday	:	08:04	-	13:46	13:49	- 23:51	Tuesday	:	08:08	-	13:40		
Friday	:	08:14	-	12:29			Friday	:	08:14	-	12:29		
Saturday	:						Saturday	:					
Sunday	:	13:50	-	20:43			Sunday	:	13:58	-	20:43		

- It merges the availability intervals that it deems **similar** to obtain the final work-shifts
- Applies a **similarity function** to decide which intervals to merge

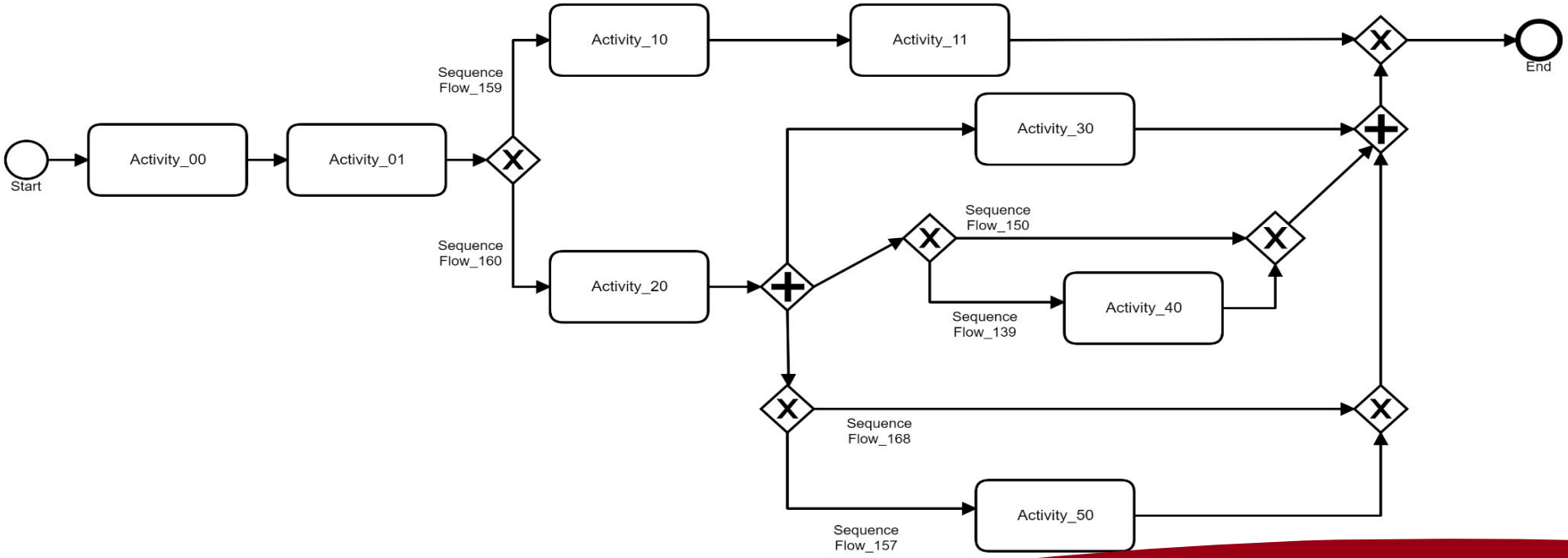
$$SIM(interval_1, interval_2) = \frac{interval_1 \cap interval_2}{interval_1 \cup interval_2}$$

Algorithms - Work-Shift Discovery

- Similarities with Cluster-Based Sampling:
  - Identifies general availability timeframes by merging similar activity intervals
- Differences with Cluster-Based Sampling:
  - All final intervals are considered as work-shifts
  - Applies noise detection and removal
  - Cluster-Based Sampling aims to remove noise by randomly choosing a cluster hoping that it does not contain noise

Experiments - Synthetic Event Logs

- We experimented on **synthetic** event logs because:
  - Start and complete timestamps **missing** from real-life event log
  - Resource information **missing** from real-life event log
  - We do not know the **"Ground Truth":** resource/role availability, the existence of work-shifts or the presence of noise



Experiments - Result Analysis

Average Accuracy	Calendar Expression	Direct Sampling	Cluster-Based Sampling	Work-shifts Discovery
Without Noise	0.9368	0.9648	0.9703	0.9770
With Noise	0.8958	0.8029	0.9137	0.9576

- We tested our methods against the state of the art from the literature
- In the case of event log **without** noise all methods performs similarly and we obtain **similar** working calendars
- In the case of event log **with** noise and more convolute work-shifts our methods performed noticeable better

Conclusions

- The aim of this thesis was to study and develop a new approach for **automatic calendar discovery**
- While the literature has developed resource-oriented methods, our approach focuses both on the role and on the resource
- Our methods introduce the possibility of **noise detection** and **removal**
- They are also able to define resource and role constraint at a **finer granularity**, especially in the case of **shift work**
- And they do not depends on any **random sampling**

Conclusions - Future Works

- Study on real-life event logs
  - We studied our methods on synthetic event logs

- Study solutions for the cases in which noisy activities expands near real work activities