Event Log Data Quality in Process Mining

Modelling, Assessing, Improving

Marco Comuzzi

Department of Industrial Engineering
Ulsan National Institute of Science and Technology (UNIST)
mcomuzzi@unist.ac.kr

About me

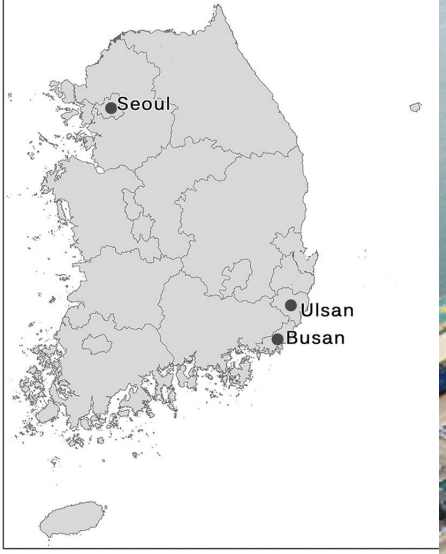
CS Undergraduate, master degree, Politecnico di Milano (밀라노 공대)
PhD Information Technology, Politecnico di Milano (밀라노 공대), 2007

2009 – 2013 Assistant Professor, Eindhoven University of Technology (네덜란드)
2013 – 2016 Assistant Professor, City, University of London (영국)
2016 – Associate Professor, Department of Industrial Engineering, UNIST

Research interests

Process mining
Blockchain

Marco Comuzzi mcomuzzi@unistac.kr

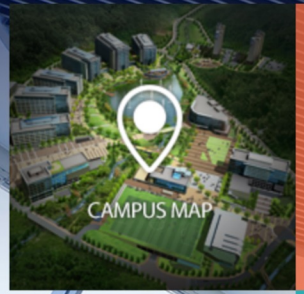

Seoul
Ulsan
Busan


EUKOR

https://www.unist.ac.kr

UNIST

About UNIST | Admissions | Academics | Research | Campus Life | UNIST News Center | Students | Parents | Faculty & Staff | K | E



UNIST

Community Contribution

Community Contribution

UNIST

Campus Life Guide Book

UNIST

Admission

Undergraduate | Graduate

UNIST

UNIST at Glance

more

UNIST NEWS

UNIST Recognized for "Excellence" in Customer Satisfaction!

[2020 Public-service Customer Satisfaction Survey]
Honored for "Excellence" in Customer Satisfaction!

UNIST

UNIST Identity

기획재정부 Ministry of Economy and Finance

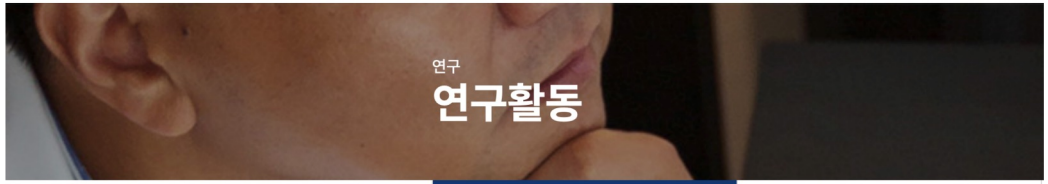
국민권익위원회 Anti-Corruption & Civil Rights Commission

kr/category/all-news/community-contribution/

https://ie.unist.ac.kr/research-activities/

UNIST | 산업공학과

소개 구



연구활동 | 연구실소개

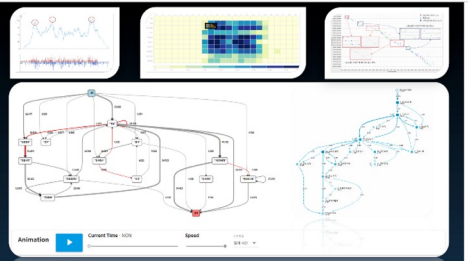
교수	Comuzzi Marco
연구그룹	스마트서비스, 스마트 비즈니스 & 파이낸스
전문분야	데이터마이닝, 기계학습 알고리즘(연합학습, 표현학습), 산업인공지능

https://iel.unist.ac.kr

UNIST

PROCESS-AWARE ARTIFICIAL INTELLIGENCE LAB

HOME
ABOUT
MEMBERS
PUBLICATIONS
PROJECTS
TEACHING
SOFTWARES & TOOLS
GALLERY



WELCOME TO
PROCESS-AWARE ARTIFICIAL INTELLIGENCE LAB

PROFESSOR MARCO COMUZZI

ABOUT

Process-Aware Artificial Intelligence Lab
PAAI Lab.

Office: UNIST Main Campus - 112 - Building 302-2 room

INTRODUCTION OF LAB (SINCE 2016 - .)

This lab was established in 2016. It focuses on the application of machine learning techniques (mainly classification) to the analysis of business process event logs. As a lab, we are always trying to contribute to the main international academic conferences in the field: Int. Conf. on Business Process Management, Int. Conf. on Process Mining and Int. Conf. on Advanced Information Systems Engineering. Our research has always practical relevance. For instance we have applied the results of our research on event logs obtained from the Ulsan Port Authority and from a large university hospital in Korea.

MEMBERS

Head of Lab

MARCO COMUZZI
Associate professor in School of Management Engineering (UNIST)

EMAIL
mcomuzzi@unist.ac.kr

OFFICE

Plan for today

Marco Comuzzi mcomuzzi@unistac.kr

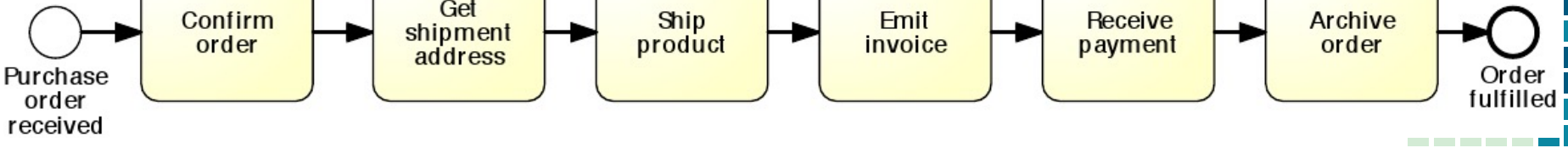
Business Processes and Process Mining 101

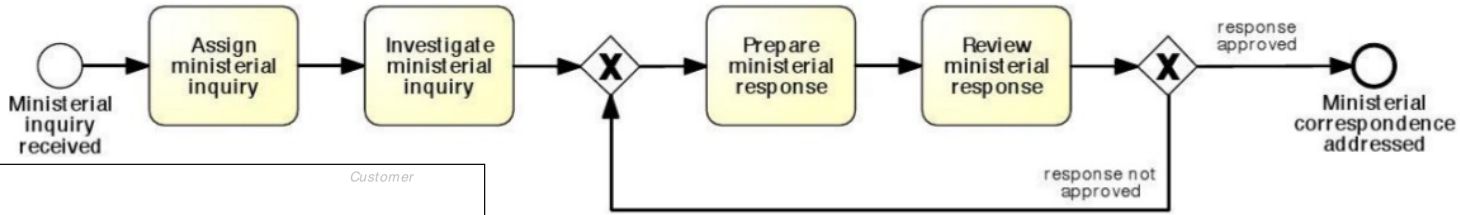
Marco Comuzzi mcomuzzi@unistac.kr

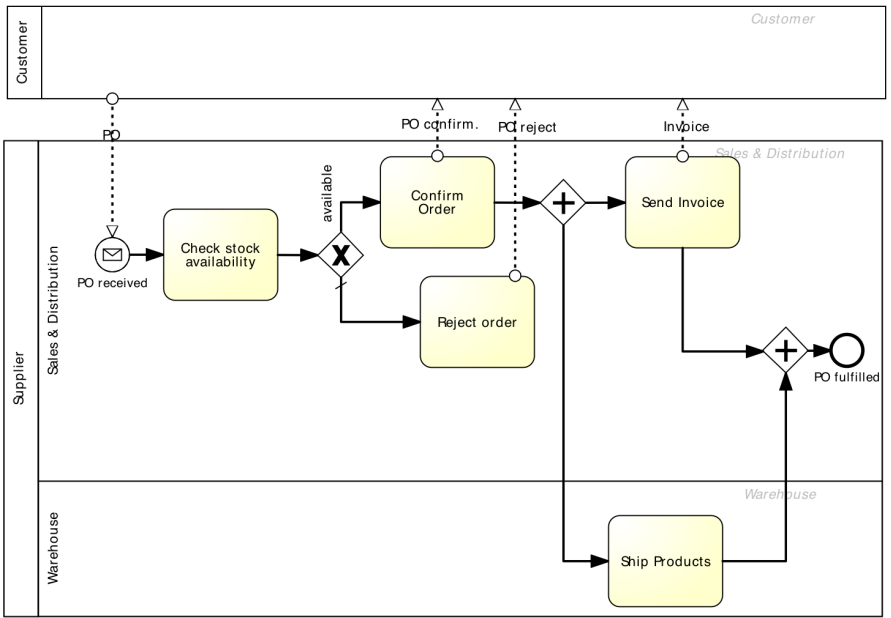
What is a business process?

"A collection of related events, activities and decisions that involve a number of actors and resources, and that collectively lead to an outcome that is of value to an organisation and/or its customers"

Marco Comuzzi mcomuzzi@unistac.kr



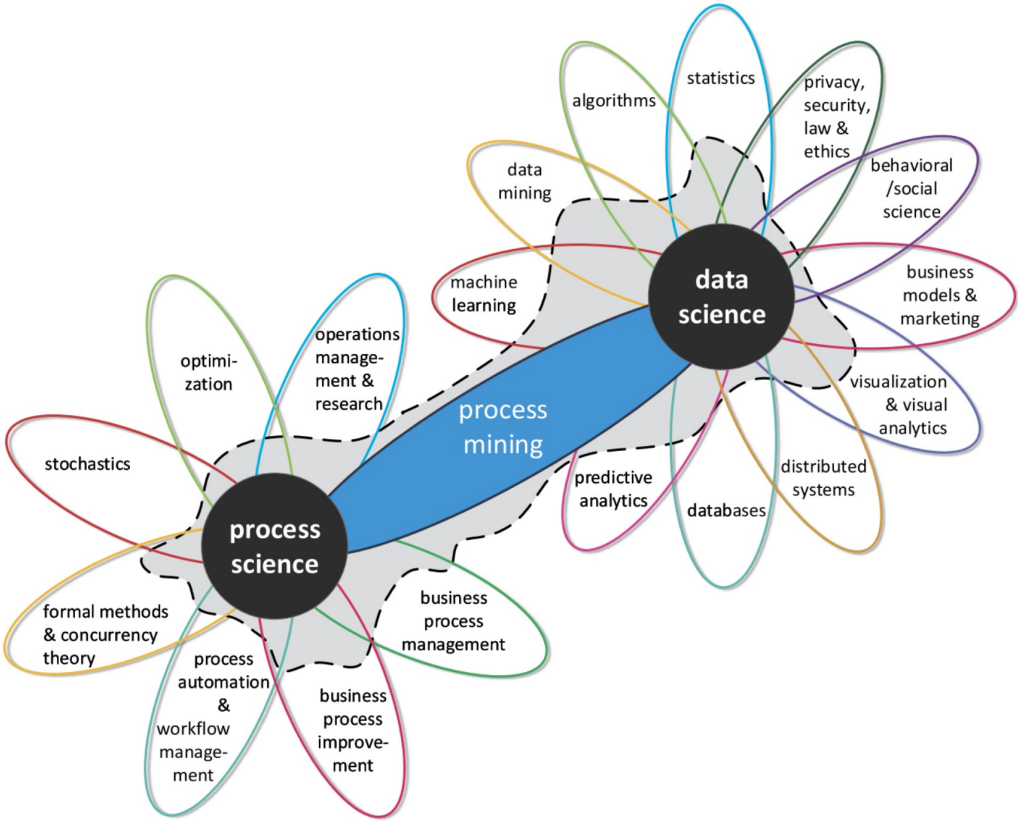




mcomuzzi@unistac.kr

Processes are everywhere

Marco Comuzzi mcomuzzi@



A bridge between data and process science

Marco Comuzzi mcomuzzi@unistac.kr

Nokia transforms business processes with Celonis

By Derek du Preez July 25, 2022 Audio mode

SUMMARY: Multinational communications company Nokia has also creat
Excellence for Celonis' Execution Management System, whicl
the business to scale the platform.

Process mining become the terrain

Enterprise technology companies like SAP and Micros technology, but the current rate of acquisition is rapid

Home › Cloud › Why Process Mining Is the top skill to learn in 2022

Cloud News

Why Process Mining Is the top skill to learn in 2022

By **CRN Team** - July 22, 2022 2

Like 0

What are “process data”?

Data logged by information systems supporting the process execution

Marco Comuzzi mcomuzzi@unistac.kr

Event logs for process mining

A process consists of **cases** (or "instances")

A case consists of a "trace" of **events**

Each event relates precisely to one case

Events in a case are ordered in time

Events can have attributes:

Activity label

Timestamp

Cost

Resource

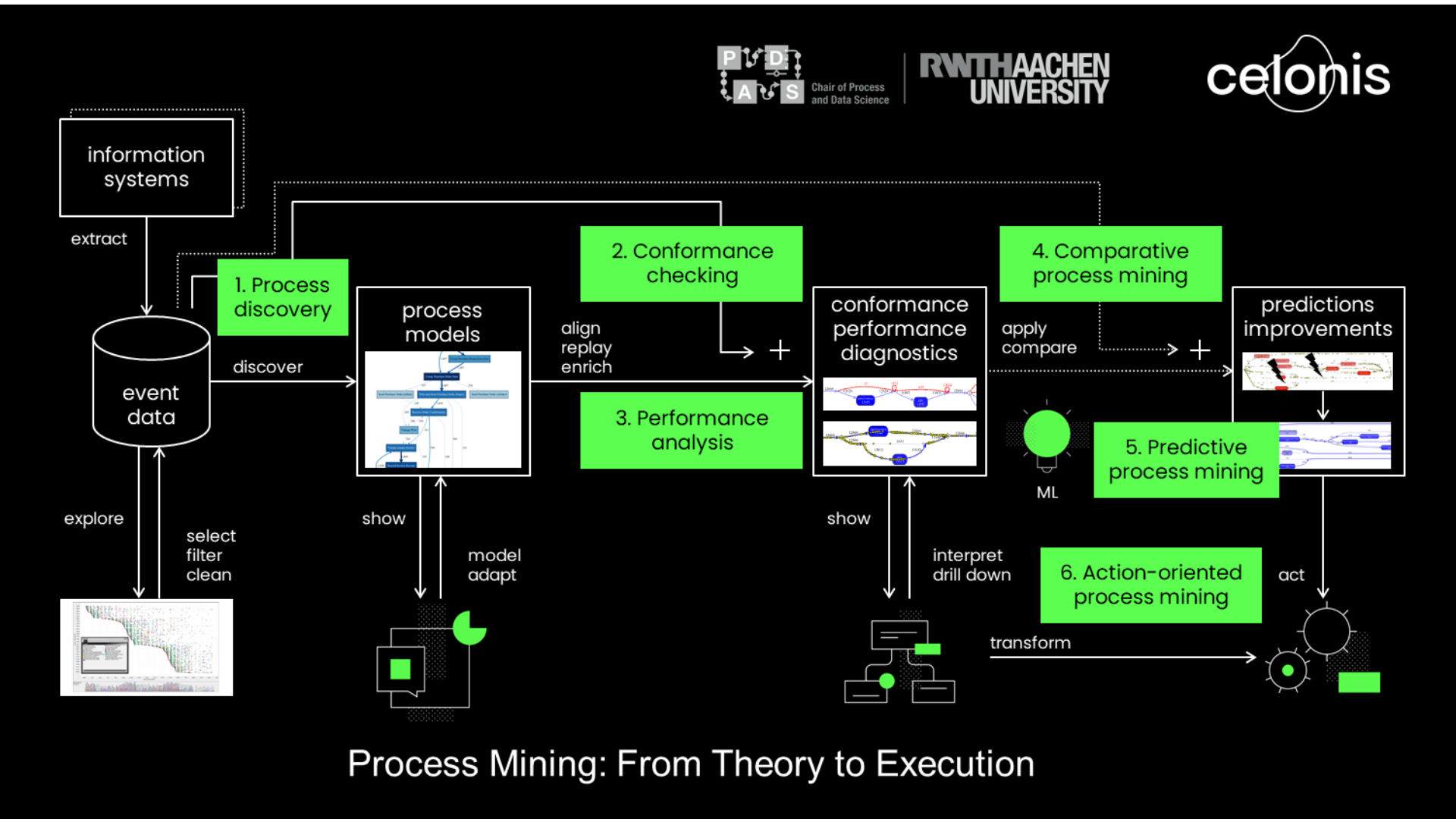
...

case id	event id	properties				
		timestamp	activity	resource	cost	...
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	35654426	06-01-2011:11.18	decide	Sara	200	...
	35654427	07-01-2011:14.24	reject request	Pete	200	...
2	35654483	30-12-2010:11.32	register request	Mike	50	...
	35654485	30-12-2010:12.12	check ticket	Mike	100	...
	35654487	30-12-2010:14.16	examine casually	Pete	400	...
	35654488	05-01-2011:11.22	decide	Sara	200	...
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	...
3	35654521	30-12-2010:14.32	register request	Pete	50	...
	35654522	30-12-2010:15.06	examine casually	Mike	400	...
	35654524	30-12-2010:16.34	check ticket	Ellen	100	...
	35654525	06-01-2011:09.18	decide	Sara	200	...
	35654526	06-01-2011:12.18	reinitiate request	Sara	200	...
	35654527	06-01-2011:13.06	examine thoroughly	Sean	400	...
	35654530	08-01-2011:11.43	check ticket	Pete	100	...
	35654531	09-01-2011:09.55	decide	Sara	200	...
	35654533	15-01-2011:10.45	pay compensation	Ellen	200	...
4	35654641	06-01-2011:15.02	register request	Pete	50	...
	35654643	07-01-2011:12.06	check ticket	Mike	100	...
	35654644	08-01-2011:14.43	examine thoroughly	Sean	400	...
	35654645	09-01-2011:12.02	decide	Sara	200	...
	35654647	12-01-2011:15.44	reject request	Ellen	200	...
...	...	...	...	...	...	...

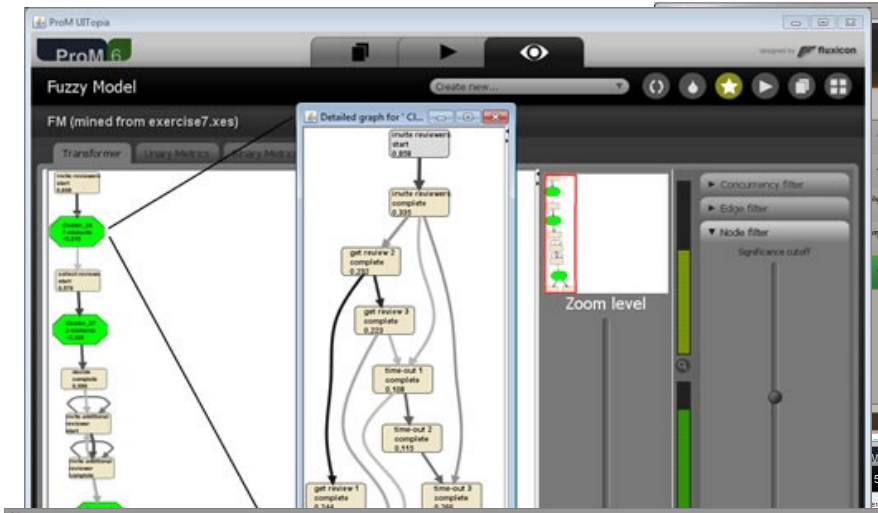
Marco Comuzzi mcomuzzi@unistac.kr

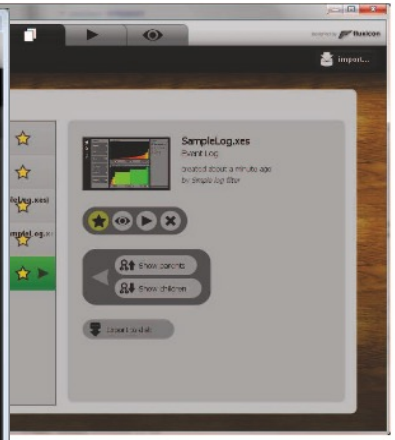
Case_ID	Event_ID	Activity_Name	CompleteTimeStamp	Resource ID	Department	Reserved Timestamp	Doctor ID	System
P001	e_1042	Reserve a consultation	2020/10/27 10:25:21	R0002	Administration	2020/10/29 12:30:00		Sys_admin
	e_1049	Receipt treatment	2020/10/29 12:31:48	R0104	Neurology		D1023	Sys_neuro
	e_1050	Start treatment	2020/10/29 13:01:13	D1023	Neurology		D1023	Sys_neuro
	e_1051	Finish treatment	2020/10/29 13:14:09	D1023	Neurology		D1023	Sys_neuro
	e_1053	Start an examination	2020/10/29 13:29:44	R3211	Radiology			Sys_radi
	e_1054	Finish an examination	2020/10/29 14:02:12	R3211	Radiology			Sys_radi
	e_1057	Payment	2020/10/29 14:11:02	R2062	Administration			Sys_admin
	e_1062	Print a prescription	2020/10/29 14:31:52	R0912	Pharmacy			Sys_phar
P002	e_1045	Reserve a consultation	2020/10/27 14:11:32	R0004	Administration	2020/10/29 13:15:00		Sys_admin
	e_1052	Receipt treatment	2020/10/29 13:18:21	R0422	Cardiology		D1030	Sys_cardi
	e_1055	Start treatment	2020/10/29 14:02:42	D1030	Cardiology		D1030	Sys_cardi
	e_1056	Finish treatment	2020/10/29 14:11:01	D1030	Cardiology		D1030	Sys_cardi
	e_1061	Payment	2020/10/29 14:27:23	R0426	Administration			Sys_admin
	e_1071	Print a prescription	2020/10/29 14:47:07	R0912	Pharmacy			Sys_phar

Marco Comuzzi mcomuzzi@unistac.kr

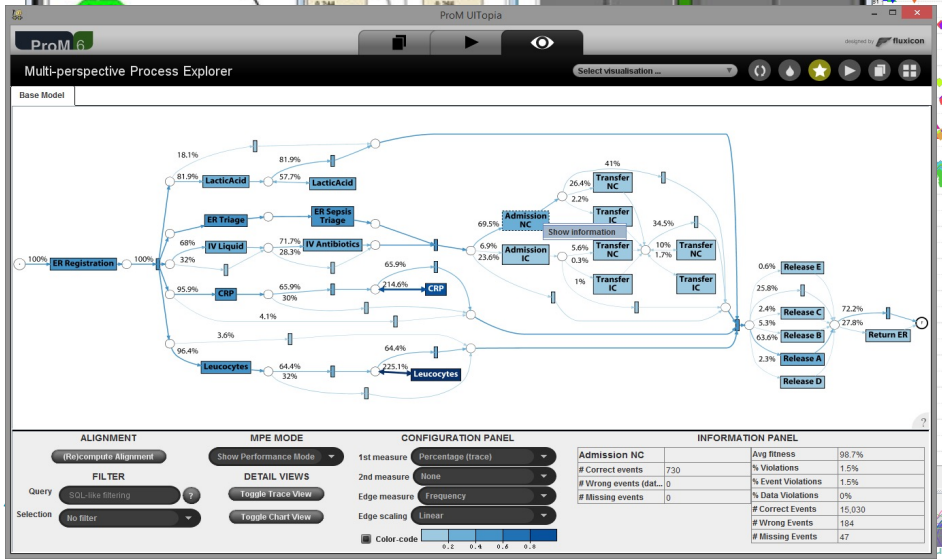


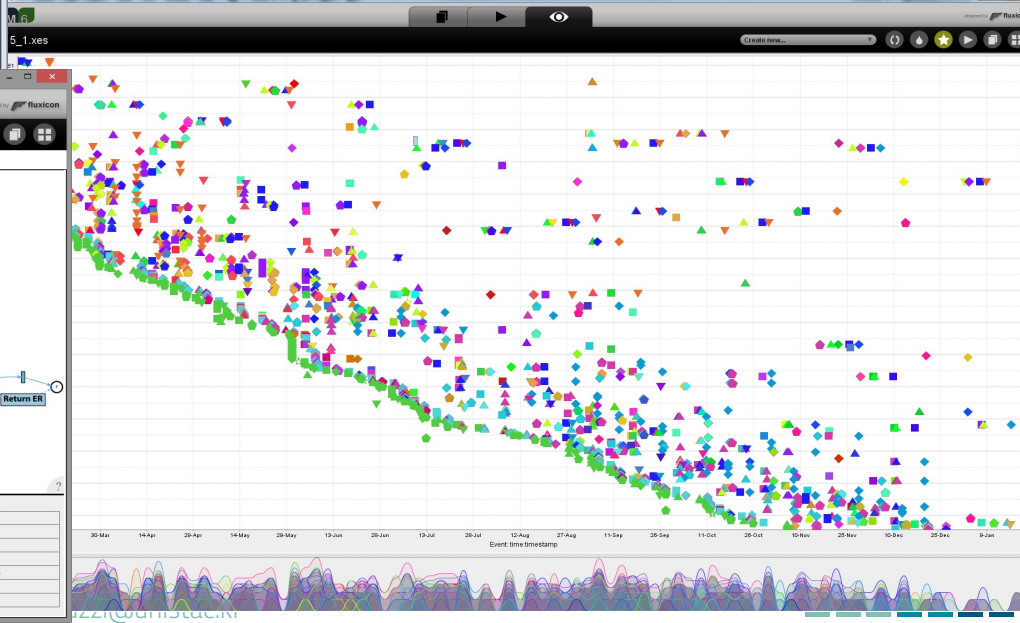
Process Mining: From Theory to Execution


ProM UITopia
ProM 6
fluxicon
Fuzzy Model
FM (mined from exercise7.xes)
Transformer
Detailed graph for "Cl...
Concurrency filter
Edge filter
Node filter
Significance cutoff
Zoom level


fluxicon
SampleLog.xes
Event Log
Show parents
Show children


ProM 6
fluxicon
Actions
Input
Output
Heuristics Miner
Log Footprint
PomPom
Replay Log in Simple Precedence Diagram
Week Abstraction Log Editor
Reset
Start


ProM UITopia
ProM 6
fluxicon
Multi-perspective Process Explorer
Select visualisation ...
Base Model
ER Registration
LacticAcid
ER Triage
ER Sepsis Triage
IV Liquid
IV Antibiotics
CRP
Leucocytes
Admission NC
Admission IC
Show information
Transfer NC
Transfer IC
Release A
Release B
Release C
Release D
Release E
Return ER
ALIGNMENT
(Re)compute Alignment
FILTER
Query
SQL-like filtering
Selection
No filter
MPE MODE
Show Performance Mode
DETAIL VIEWS
Toggle Trace View
Toggle Chart View
CONFIGURATION PANEL
1st measure
Percentage (trace)
2nd measure
None
Edge measure
Frequency
Edge scaling
Linear
Color-code
INFORMATION PANEL
Admission NC
Correct events 730
Wrong events (dat... 0
Missing events 0
Avg fitness 98.7%
% Violations 1.5%
% Event Violations 1.5%
% Data Violations 0%
# Correct Events 15,030
# Wrong Events 184
# Missing Events 47


5_1.xes
Create new...
Event: time:timestamp

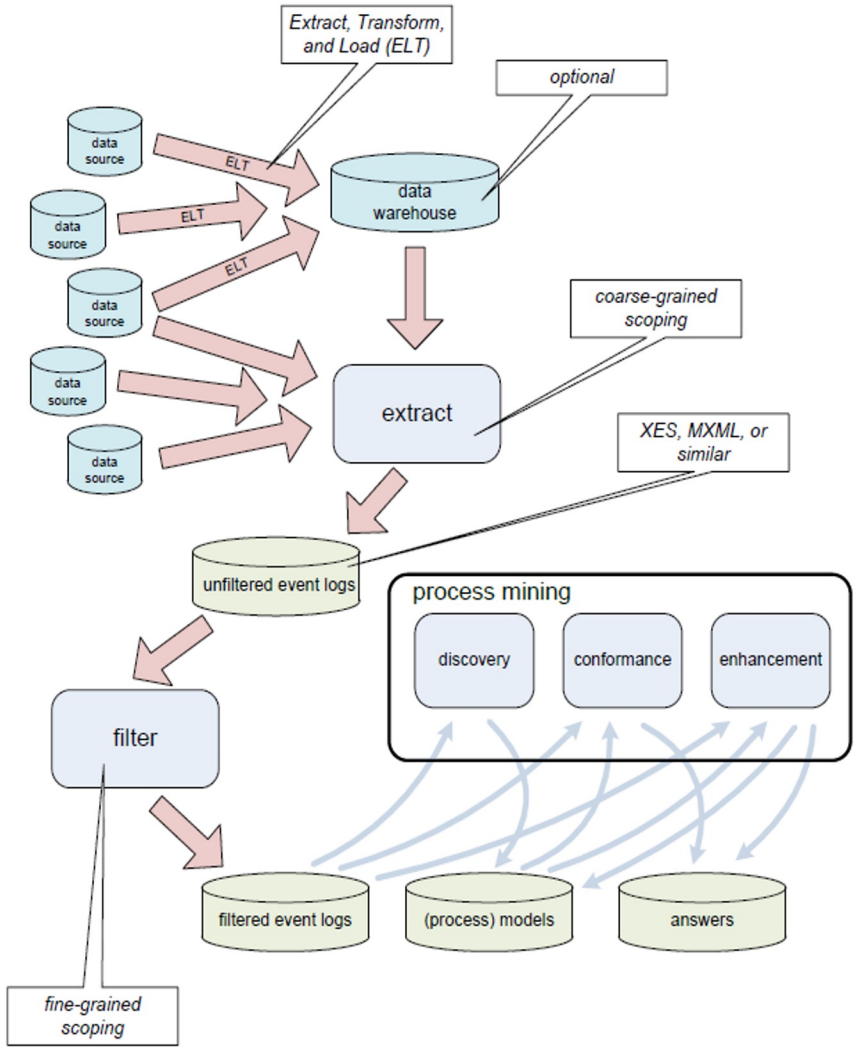
Ter Hofstede, A.H., Koschmider, A., Marrella, A., Andrews, R., Fischer, D.A., Sadeghianasl, S., Wynn, M.T., Comuzzi, M., De Weerdt, J., Goel, K. and Martin, N., 2023. Process-data quality: the true frontier of process mining. *ACM Journal of Data and Information Quality*, *15*(3), pp.1-21.

Event Log Data Quality

Marco Comuzzi mcomuzzi@unistac.kr

Any system used in a process can be a source of process data:

ERP, CRM, Sharepoint, spreadsheets, accounting/management software, IoT data streams, …



Event log preparation is critical, indeed

Ensuring event log has sufficient quality is crucial

Event log extraction and preparation can take up to 80% of the time in a process mining project

Lack of structured approaches to deal with this phase

Marco Comuzzi mcomuzzi@unistac.kr

Data Quality

Accuracy

Degree of adherence of the data to reality

Completeness

Capability of the data of representing all relevant aspects of reality

Reliability

Capability of the data to comply with the properties of reality (integrity constraints and business rules)

Bose, R. J. C., Mans, R. S., & Van Der Aalst, W. M. (2013, April). Wanna improve process mining results?. In *2013 IEEE symposium on computational intelligence and data mining (CIDM)* (pp. 127-134). IEEE.

DQ of event logs

Bose et al. 2012

Top-down approach

Missing Data

Incorrect Data

Imprecise Data

Irrelevant Data

	case	event	belongs to	c_attribute	position	activity name	timestamp	resource	e_attribute
Missing Data	I1	I2	I3	I4	I5	I6	I7	I8	I9
Incorrect Data	I10	I11	I12	I13	I14	I15	I16	I17	I18
Imprecise Data			I19	I20	I21	I22	I23	I24	I25
Irrelevant Data	I26	I27							

Marco Comuzzi mcomuzzi@unistac.kr

Event Log Imperfection Patterns

A different, bottom-up approach

Derive patterns from existing process mining case studies

11 patterns identified

1. Form-based Event Capture
2. Inadvertent Time Travel
3. Unanchored Event
4. Scattered Event
5. Elusive Case
6. Scattered Case
7. Collateral Events
8. Polluted Label
9. Distorted Label
10. Synonymous Labels
11. Homonymous Label

http://www.workflowpatterns.com/patterns/logimperfection/

Marco Comuzzi mcomuzzi@unistac.kr

Form-based event capture

Episode ID	Event	Timestamp	Description	...
ID1	Primary Survey	2012-11-23 15:42:38	.....	....
ID1	Airway Clear	2012-11-23 15:42:38	.....	....
ID1	...	2012-11-23 15:42:38	.....	....
	Primary Survey	2012-11-24 09:58:33	.....	....
	Airway Clear	2012-11-24 09:58:33	.....	....
	...	2012-11-24 09:58:33	.....	....
	Procedure 1	2012-11-24 09:58:33	Completed on 2012-11-24 06:58:34	....

These events are recorded on a form …

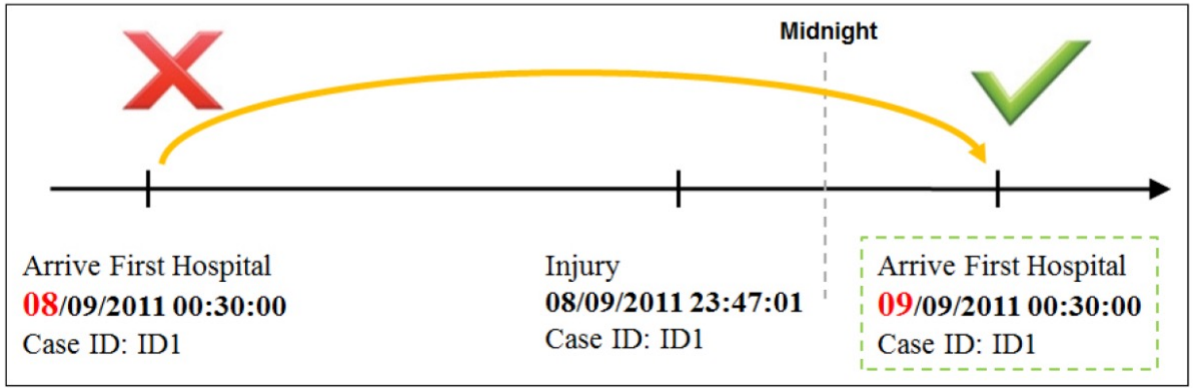
… and all have the same timestamp.

Marco Comuzzi mcomuzzi@unistac.kr

Inadvertent time travel

Episode ID	Activity	Timestamp	...
ID1	Arrival first hospital	2011-09-08 00:30:00	.....
ID1	Injury	2011-09-08 23:47:01	.....
...	.....	.....	.....
ID1	Operation	2011-09-09 16:30:00	.....

'Midnight' problem. Time portion correct but date part in error.



Collateral events

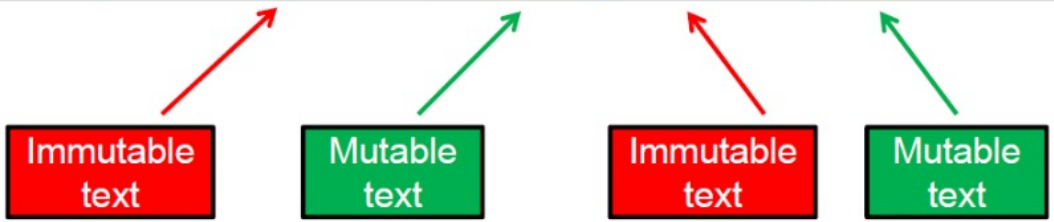
caseID	Activity	Timestamp
1234567	Adjust recovery cost	19/06/2014 12:15:18
1234567	Adjust recovery cost	19/06/2014 12:16:53
1234567	Email	19/06/2014 12:19:25
....	....	.....
1234567	Pay assessor fee	19/06/2
1234567	Adjust admin cost	19/06/2014 12:22:48

All events refer to single process step 'Pay Insurance Claim Assessor'.

Marco Comuzzi mcomuzzi@unistac.kr

Polluted labels

caseID	activity	timestamp	.....
xxxx	Notification of Loss - AAAA Incident No. aaaa	xxxx-xx-xx xx:xx:xx	.....
xxxx	Notification of Loss - BBBB Incident No. bbbb	yyyy-yy-yy yy:yy:yy	.....
xxxx	Notification of Loss - CCCC Incident No. cccc	zzzz-zz-zz zz:zz:zz	.....
.....	Notification of Loss - DDDD Incident No. dddd	.....	.....



Marco Comuzzi mcomuzzi@unistac.kr

Distorted labels

caseID	activity	timestamp	Description
1234567	a/w inv to cls.	06/09/2013 12:33:17	.....
8912345	a/w inv to cls	06/09/2013 13:10:23	.....
1234567	XX – Further Information Required	06/09/2013 13:15:00	.....
8912345	XX – Further Infomation Required	13/09/2013 07:24:36	.....

Marco Comuzzi mcomuzzi@unistac.kr

Homonymous labels

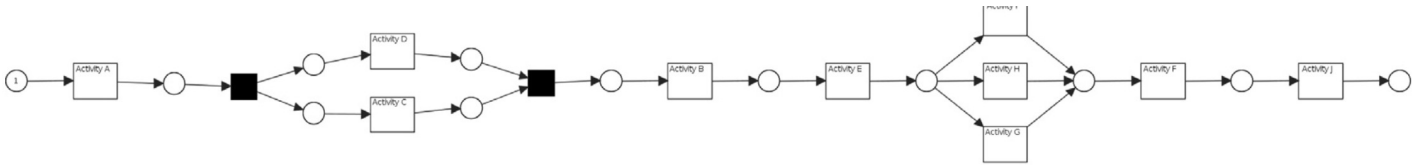
caseID	activity	timestamp	Description
1234567	Triage Assessment	06/09/2013 12:33:17	.....
1234567	Progress Note	06/09/2013 13:10:23	.....
1234567	Discharged	06/09/2013 13:15:00	.....
1234567	Triage Assessment	13/09/2013 07:24:36	.....
1234567	Triage Assessment	13/09/2013 07:28:51	.....

Marco Comuzzi mcomuzzi@unistac.kr

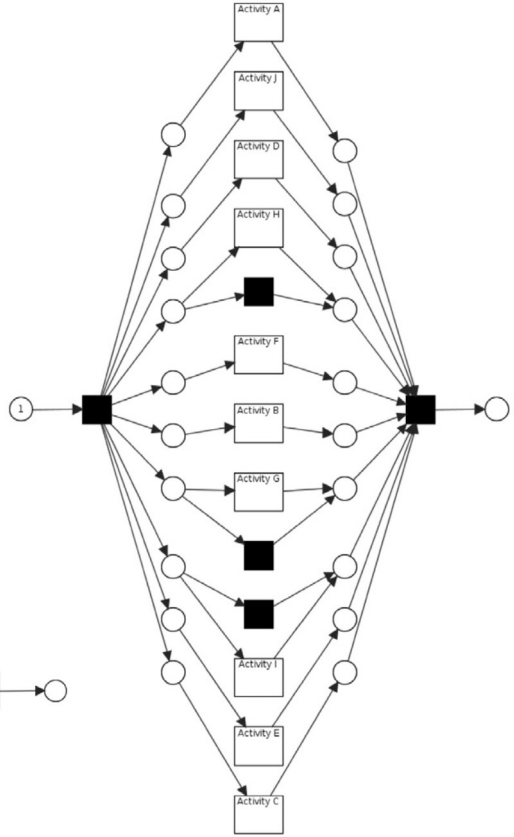
Nguyen, H. T. C., Lee, S., Kim, J., Ko, J., & Comuzzi, M. (2019). Autoencoders for improving quality of process event logs. Expert Systems with Applications, 131, 132-147.

Effects of poor DQ of event logs

Process model discovered from a synthetic log perturbed with 10% "wrong" activity labels and timestamps



(a) Process model mined from original (clean) event log



(b) Process model mined from log with injected anomalous values

Marco Comuzzi mcomuzzi@unistac.kr

Effects of poor DQ of event logs

Evident to everybody in practice…

…but rarely researched in academia, why?

Marco Comuzzi mcomuzzi@unistac.kr

No (labeled) data!

Some imperfection patterns can be easily simulated, e.g. synonym labels, distorted labels

A general, extensible approach to generate data for testing event log cleaning tools is missing

FLAWD: a language for simulating event log errors

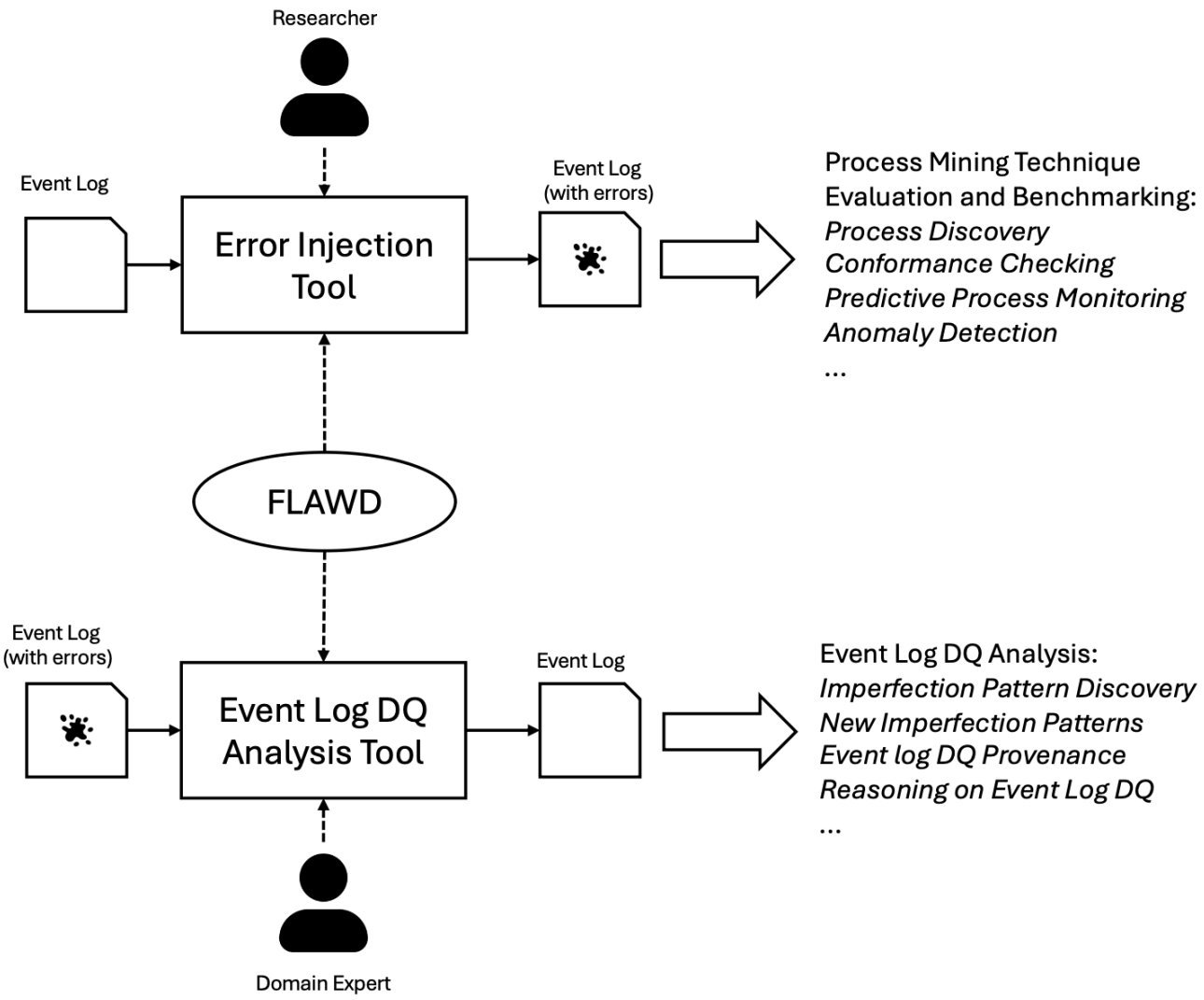
Marco Comuzzi mcomuzzi@unistac.kr

Introducing FLAWD

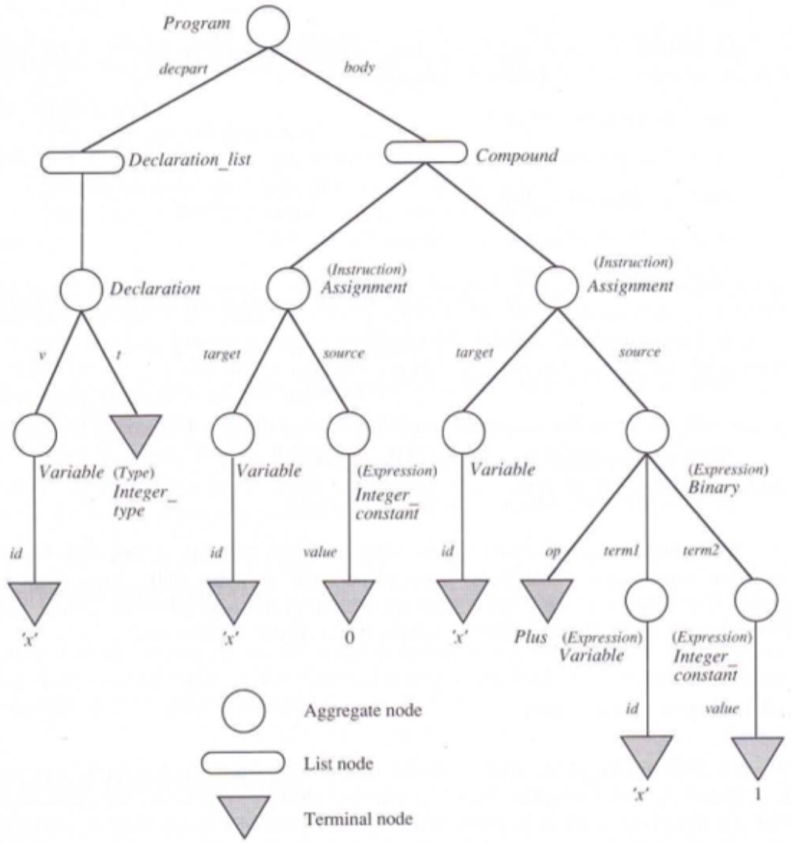
A language to model event log imperfection patterns

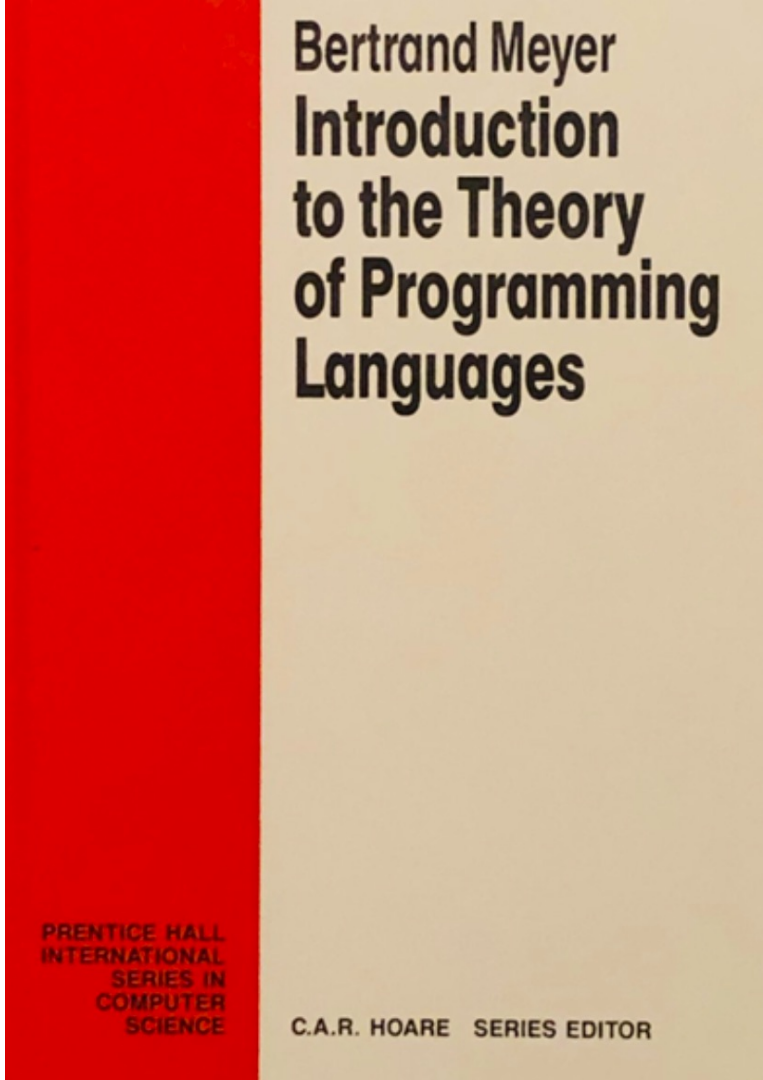
An implementation of the language as an event log error injection tool

Marco Comuzzi mcomuzzi@unistac.kr


Researcher
Event Log
Error Injection Tool
Event Log (with errors)
Process Mining Technique Evaluation and Benchmarking:
Process Discovery
Conformance Checking
Predictive Process Monitoring
Anomaly Detection
...
FLAWD
Event Log (with errors)
Event Log DQ Analysis Tool
Event Log
Event Log DQ Analysis:
Imperfection Pattern Discovery
New Imperfection Patterns
Event log DQ Provenance
Reasoning on Event Log DQ
...
Domain Expert

How to describe FLAWD?





luzzi@unistac.kr

Why an abstract syntax grammar?

Allows to focus on the elements of a language, instead of the external appearance of programs written in that language

Facilitates understanding of properties, expressiveness, and concrete syntax generation

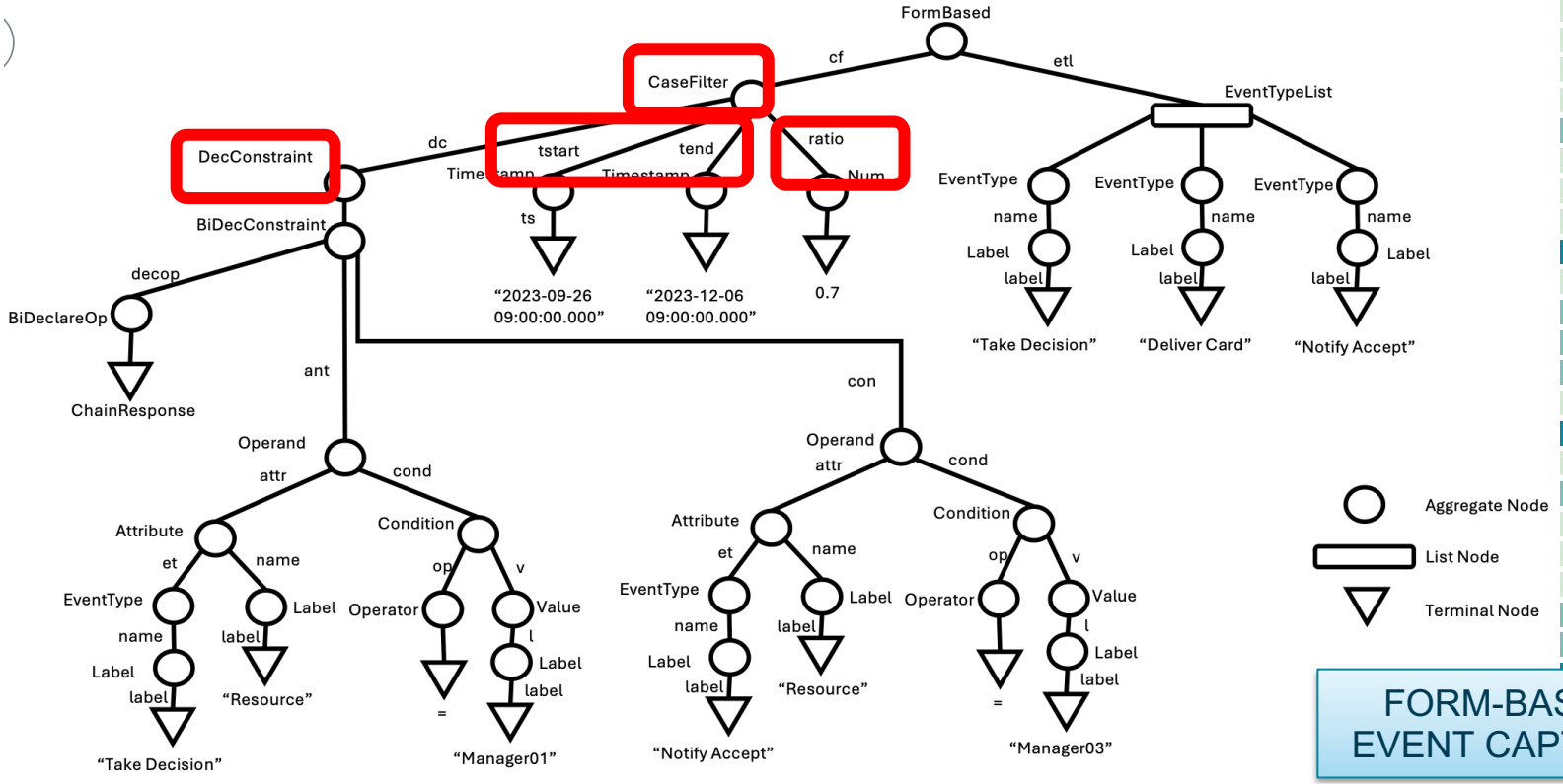
Marco Comuzzi mcomuzzi@unistac.kr

Event log

Case ID	Activity	Resource	Timestamp
1	...	...	...
1	Take decision	Manager01	2023-09-27 11:56:18
1	Deliver card	Manager01	2023-09-28 15:56:20
1	Notify accept	Manager03	2023-09-29 09:56:20
1	...	...	...

Event log with injected errors

Case ID	Activity	Resource	Timestamp
1	...	...	...
1	Take decision	Manager01	2023-09-27 11:56:18
1	Deliver card	Manager01	2023-09-27 11:56:18
1	Notify accept	Manager03	2023-09-27 11:56:18
1	...	...	...



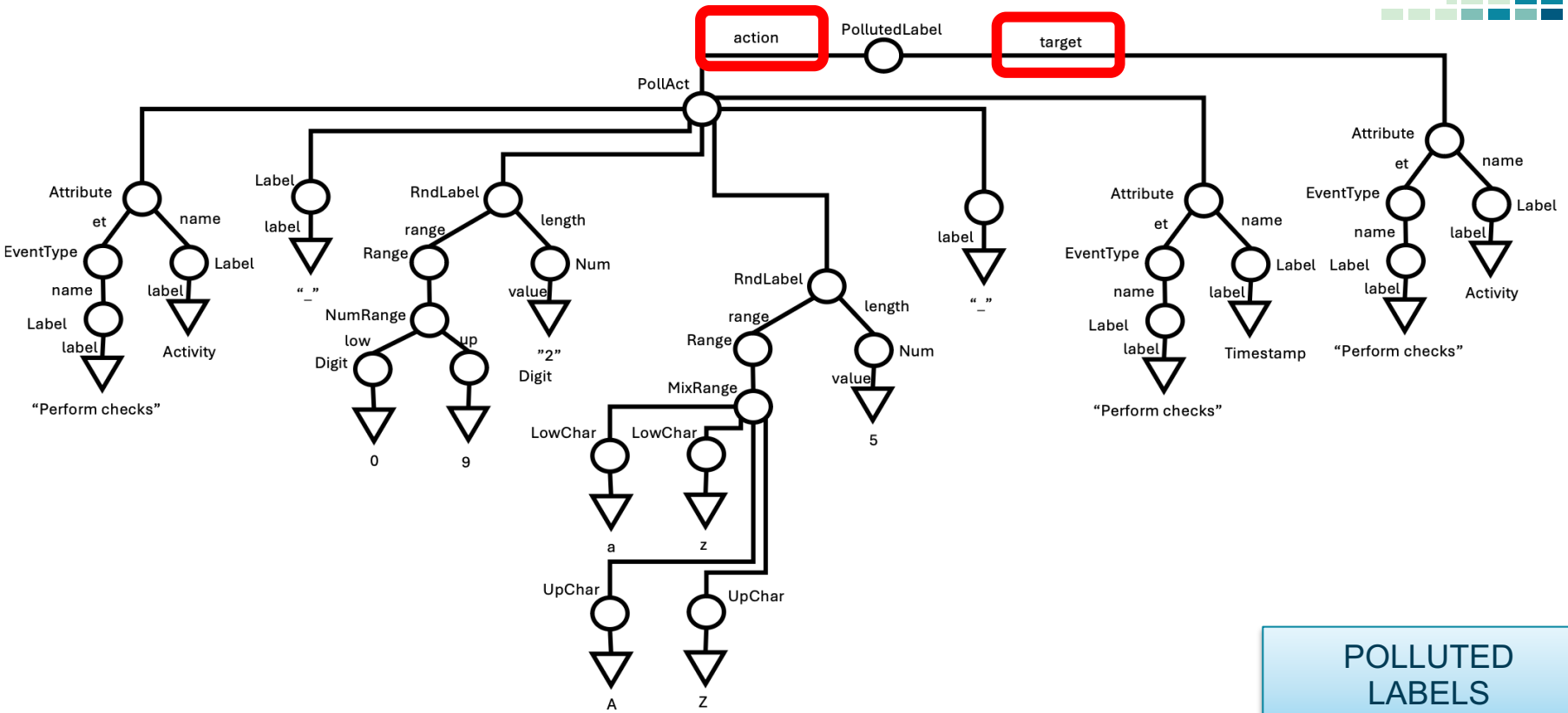
FORM-BASED EVENT CAPTURE

Event log

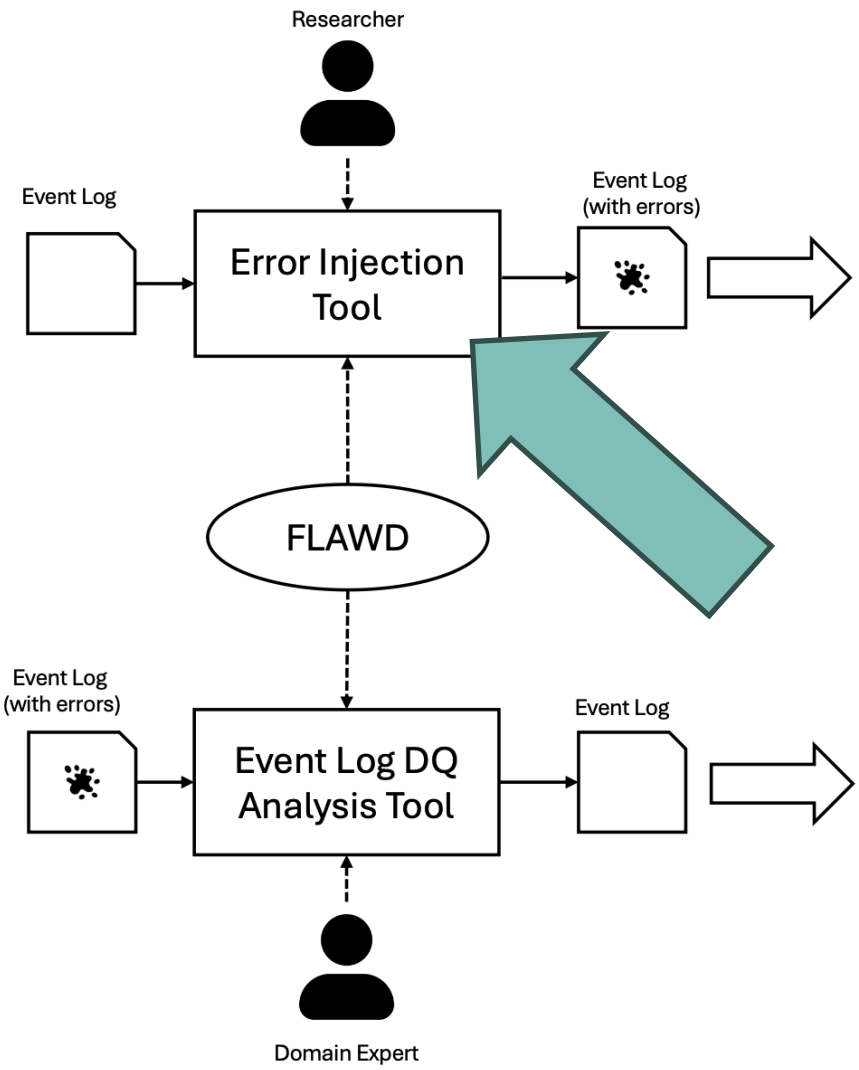
Case ID	Activity	Resource	Timestamp
1	New application received	Clerk01	2011-10-03 14:56:18
1	Perform checks	Clerk01	2011-10-04 11:56:18
1	Take decision	Manager01	2011-10-05 15:56:20
1	Notify accept	Manager02	2011-10-06 09:56:20
1	Deliver Card	Clerk03	2011-10-06 09:58:20

Event log with injected errors

Case ID	Activity	Resource	Timestamp
1	New application received	Clerk01	2011-10-03 14:56:18
1	Perform checks_91HTfsG_20111004 115618	Clerk01	2011-10-04 11:56:18
1	Take decision	Manager01	2011-10-05 15:56:20
1	Notify accept	Manager02	2011-10-06 09:56:20
1	Deliver Card	Clerk03	2011-10-06 09:58:20



POLLUTED LABELS



Process Mining Technique Evaluation and Benchmarking:
Process Discovery
Conformance Checking
Predictive Process Monitoring
Anomaly Detection
...

Event Log DQ Analysis:
Imperfection Pattern Discovery
New Imperfection Patterns
Event log DQ Provenance
Reasoning on Event Log DQ
...

Implementation

Python scripts

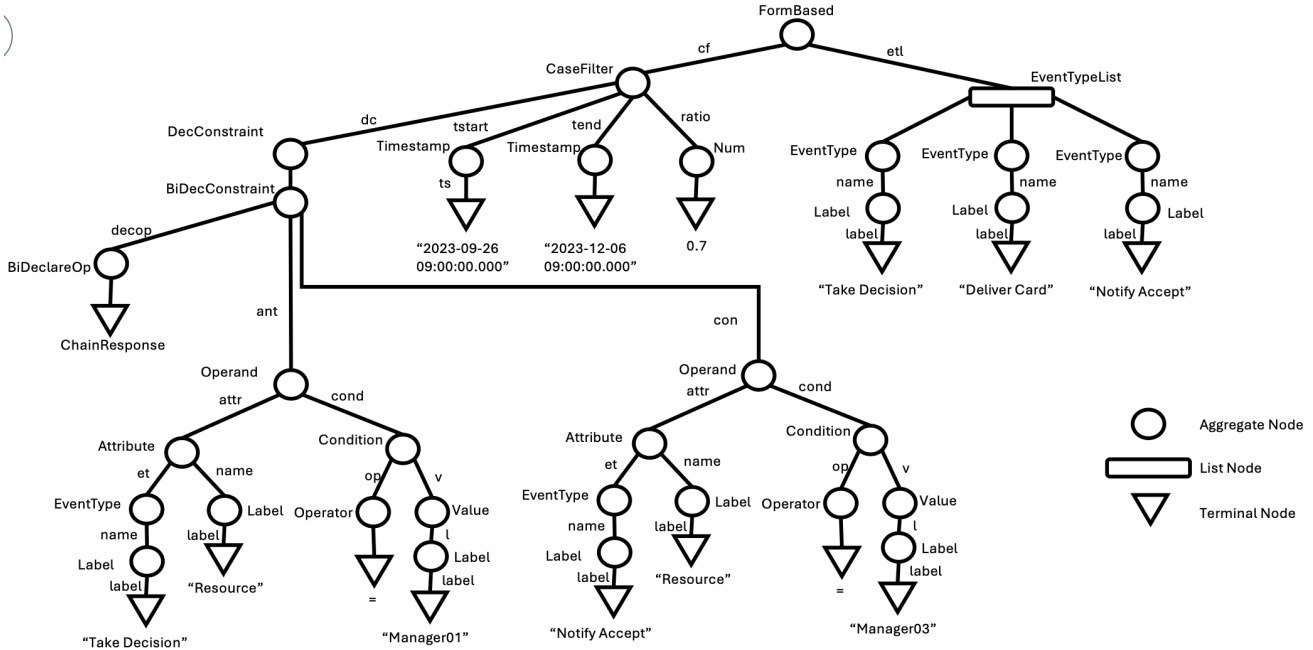
PM4Py for event log import/export

Declare4Py for case filtering based on Declare constraints

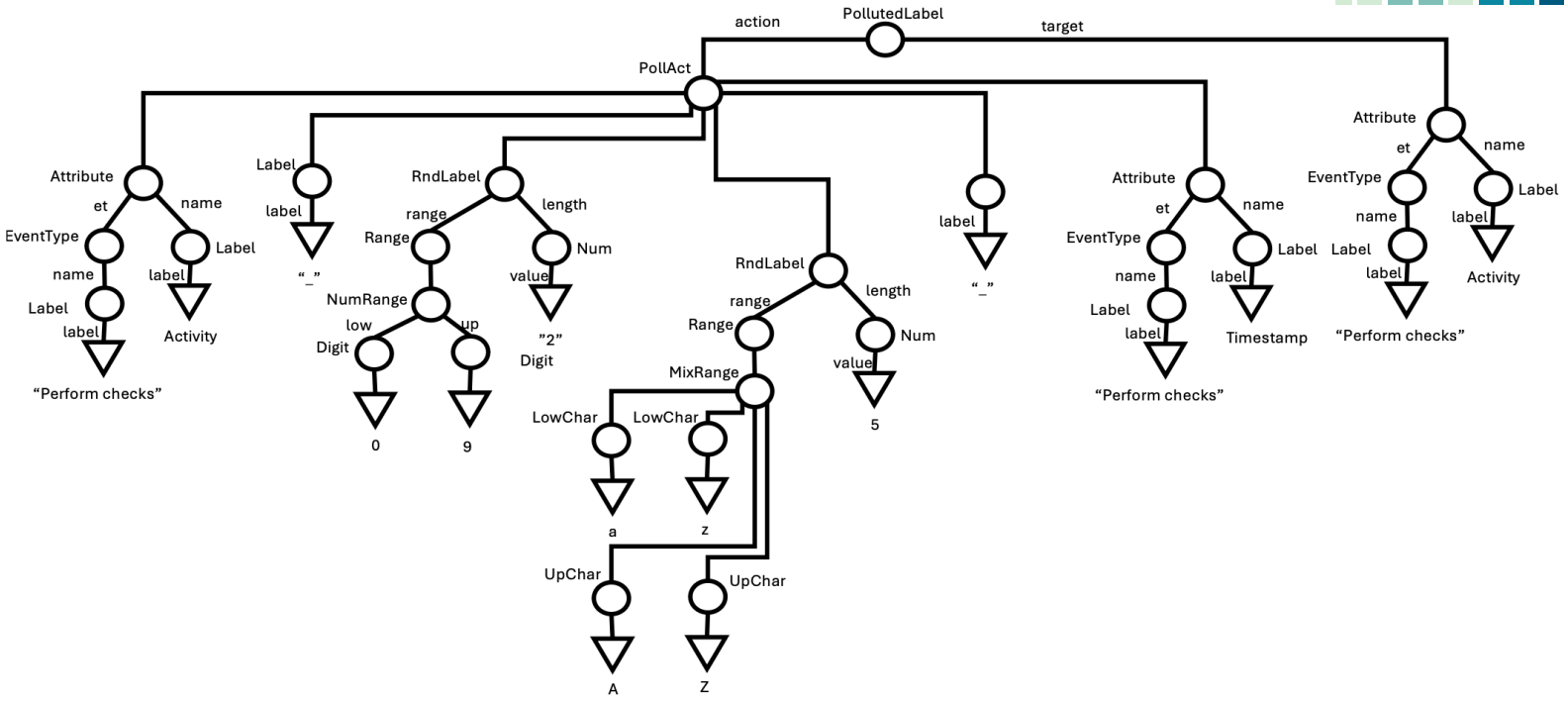
Generation of meaningful labels

Marco Comuzzi mcomuzzi@unistac.kr

```
el_polluted = form_based(el,
    etl = "['Take␣Decision','Deliver␣Card','Notify␣Accept']",
    ratio = 0.7,
    time_start = "2023-09-26␣09:00:00.000",
    time_end = "2023-12-06␣09:00:00.000",
    decConstraint = "ChainResponse[Take␣Decision,␣Notify␣Accept]␣|␣A.
        ↪ Resource␣is␣Manager01␣|␣T.␣Resource␣is␣Manager03']"
```



```
el_polluted = polluted_label(el,
    target = "[Activity:(Perform␣checks',’Take␣decision’)]"
    action = "[Activity]_[0-9:{2}][a-zA-Z:{5}]_[Timestamp*(%Y%m%d␣%H%M%
        ↪ S%f)]",
    # ratio = ...
    # filters = ...
```



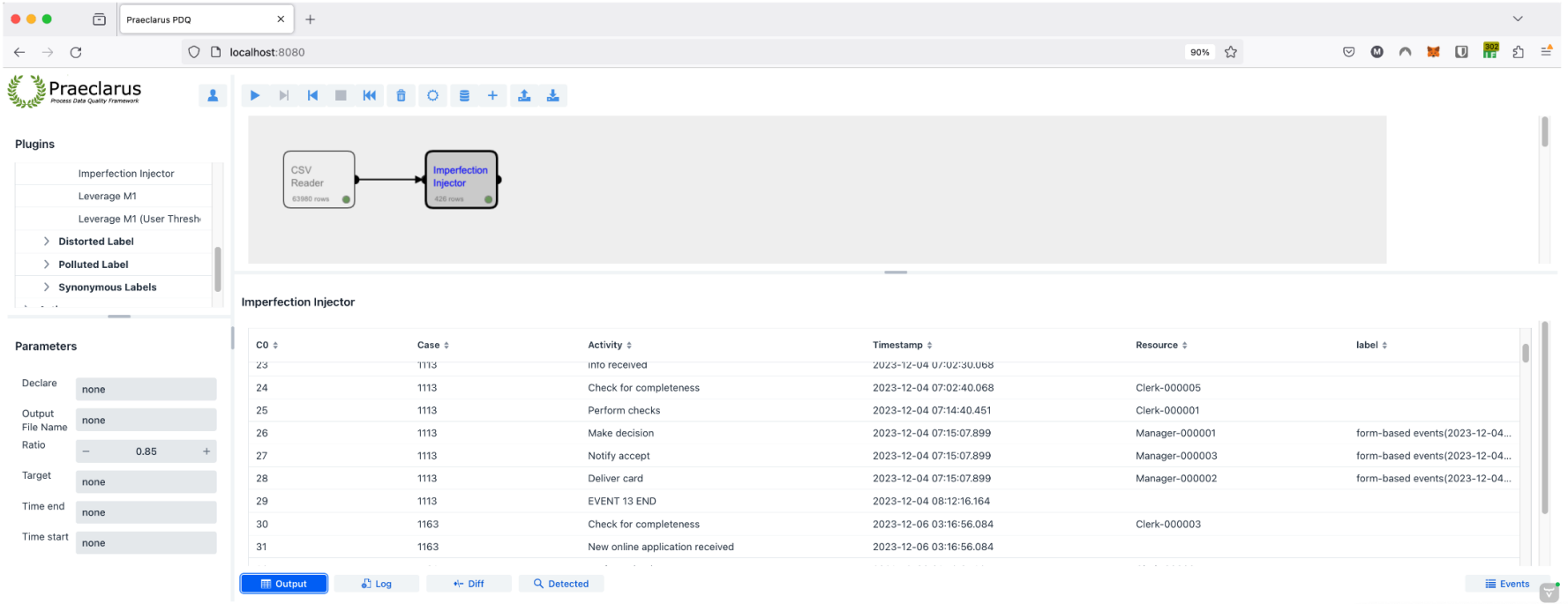
Generated labels

Trace the “provenance” of error injection

Case	Activity	Timestamp	Resource	label
1113	Check for completeness	2023-12-04 07:02:40.068	Clerk-000005	
1113	Perform checks	2023-12-04 07:14:40.451	Clerk-000001	
1113	Make decision	2023-12-04 07:15:07.899	Manager-000001	form-based events(2023-12-04 07:15:07.899000)
1113	Notify accept	2023-12-04 07:15:07.899	Manager-000003	form-based events(2023-12-04 08:02:31.947000)
1113	Deliver card	2023-12-04 07:15:07.899	Manager-000002	form-based events(2023-12-04 08:08:22.361000)

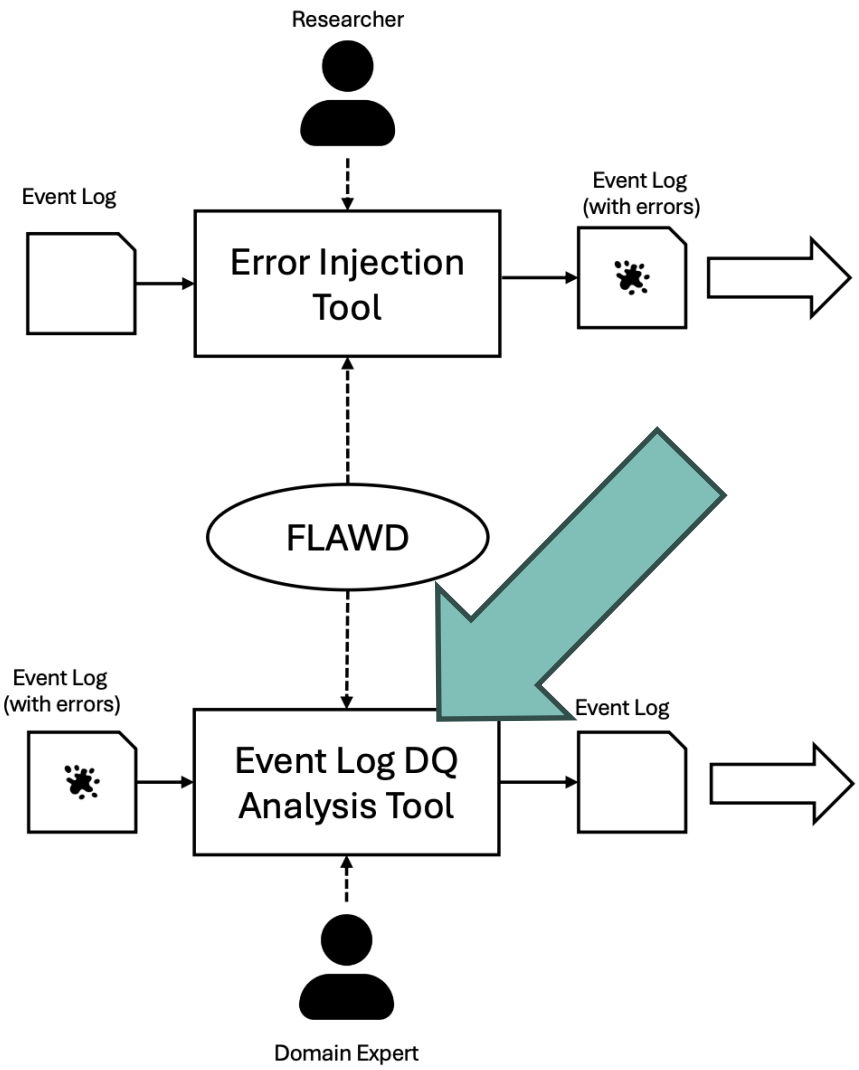
Marco Comuzzi mcomuzzi@unistac.kr

Integration into the Praeclarus Toolkit



Improving Event Log Data Quality

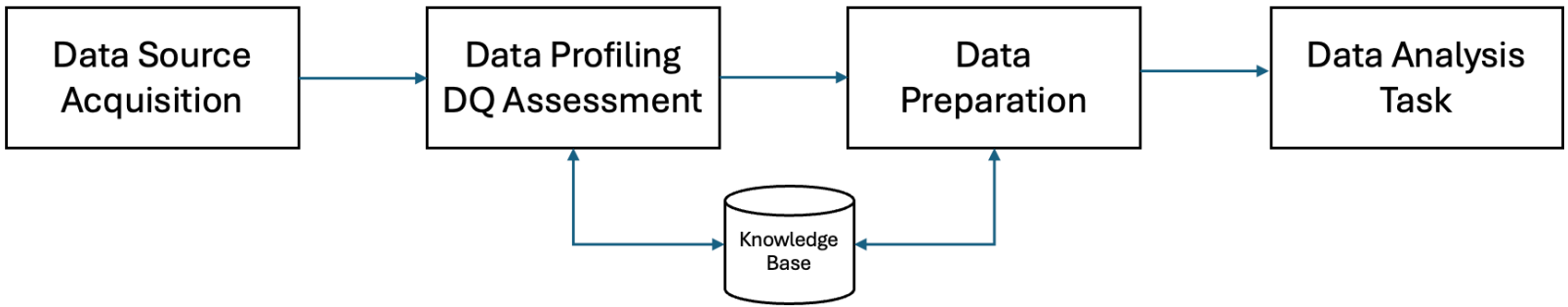
Marco Comuzzi mcomuzzi@unistac.kr



Process Mining Technique Evaluation and Benchmarking:
Process Discovery
Conformance Checking
Predictive Process Monitoring
Anomaly Detection
...

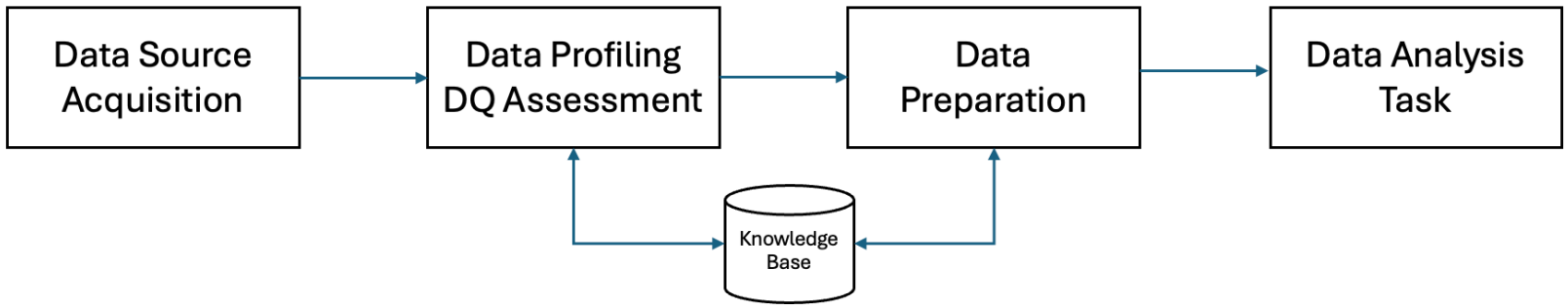
Event Log DQ Analysis:
Imperfection Pattern Discovery
New Imperfection Patterns
Event log DQ Provenance
Reasoning on Event Log DQ
...

Data preparation pipelines in data science



Marco Comuzzi mcomuzzi@unistac.kr

Data preparation pipelines in process mining



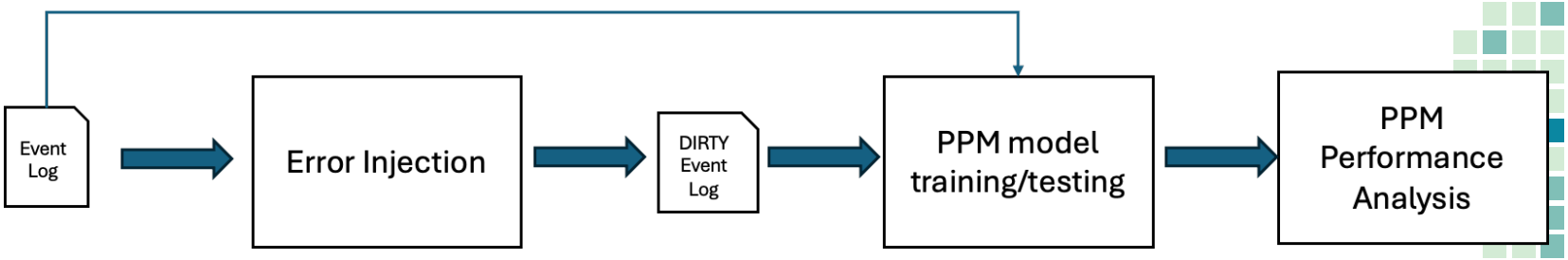
Data Source: Event Log

Data Preparation: Methods to Improve Event Log Quality

Data Analysis Task: Process Discovery, Conformance Checking, Predictive Process Monitoring (PPM)

Marco Comuzzi mcomuzzi@unistac.kr

Step 1: Empirical analysis for building knowledge base



Marco Comuzzi mcomuzzi@unistac.kr

Event log is a dataset

Can be used to train machine learning predictive models…

...which can then be used to predict "the future" of (running) process cases

case id	event id	properties				
		timestamp	activity	resource	cost	…
1	35654423	30-12-2010:11.02	register request	Pete	50	…
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	…
	35654425	05-01-2011:15.12	check ticket	Mike	100	…
	35654426	06-01-2011:11.18	decide	Sara	200	…
	35654427	07-01-2011:14.24	reject request	Pete	200	…
2	35654483	30-12-2010:11.32	register request	Mike	50	…
	35654485	30-12-2010:12.12	check ticket	Mike	100	…
	35654487	30-12-2010:14.16	examine casually	Pete	400	…
	35654488	05-01-2011:11.22	decide	Sara	200	…
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	…
3	35654521	30-12-2010:14.32	register request	Pete	50	…
	35654522	30-12-2010:15.06	examine casually	Mike	400	…
	35654524	30-12-2010:16.34	check ticket	Ellen	100	…
	35654525	06-01-2011:09.18	decide	Sara	200	…
	35654526	06-01-2011:12.18	reinitiate request	Sara	200	…
	35654527	06-01-2011:13.06	examine thoroughly	Sean	400	…
	35654530	08-01-2011:11.43	check ticket	Pete	100	…
	35654531	09-01-2011:09.55	decide	Sara	200	…
	35654533	15-01-2011:10.45	pay compensation	Ellen	200	…
4	35654641	06-01-2011:15.02	register request	Pete	50	…
	35654643	07-01-2011:12.06	check ticket	Mike	100	…
	35654644	08-01-2011:14.43	examine thoroughly	Sean	400	…
	35654645	09-01-2011:12.02	decide	Sara	200	…
	35654647	12-01-2011:15.44	reject request	Ellen	200	…
…	…	…	…	…	…	…

What can we predict about a process case?

“Which will be the next activity that will be executed?”

Predicting activities that will be executed

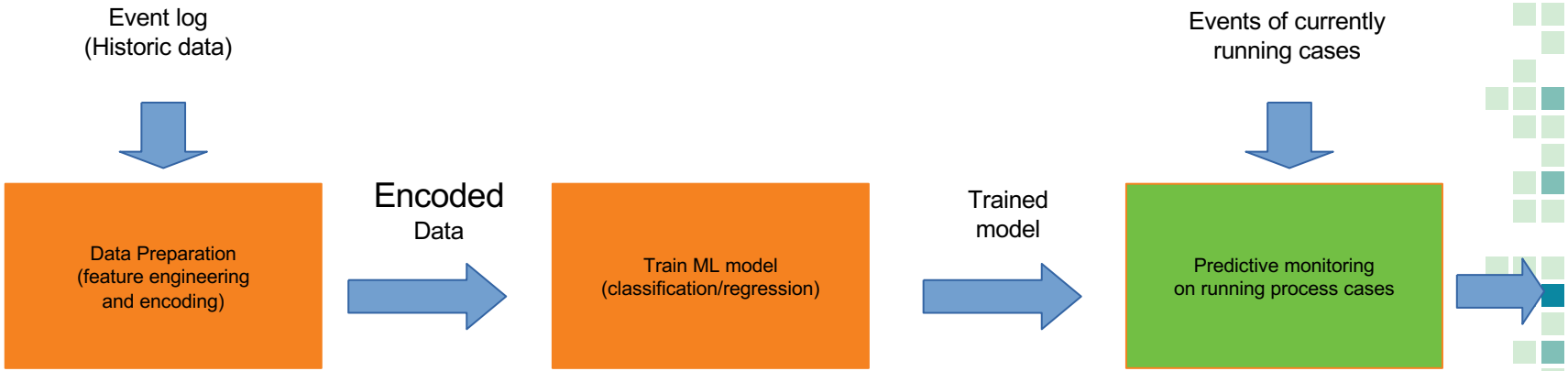
“When will the next activity be executed?
What will be the timestamp of the last activity of this case?”

Predicting timestamps

“What would be the outcome of this case, e.g., would it be possible to discharge this patient by tomorrow?”

Predict process outcomes

Predictive monitoring – General Framework



Index encoding: example

	Activity, timestamp, resource, cost				Outcome
Trace 1	A, t=1, Alice, 100	B, t=4, Alice, 50	C, t=6, Bob, 60	B, t=7, Chris, 70	1
Trace 2	A, t=3, Alice, 90	C, t=8, Chris, 100	-	-	0

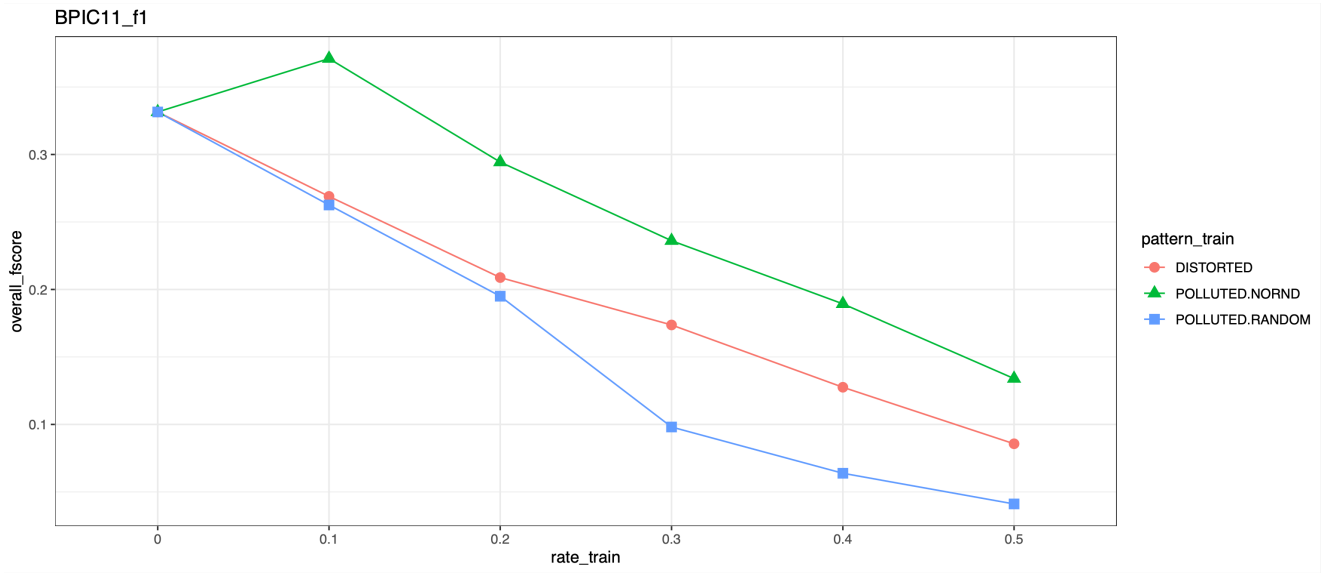
	Act_1*	Res_1*	t_1	c_1	Act_2*	Res_2*	t_2	c_2	Act_3*	Res_3*	t_3	c_3	Act_4*	Res_4*	t_4	c_4	Outcome
Trace 1	A	Alice	1	100	B	Alice	4	50	C	Bob	6	60	B	Chris	7	70	1
Trace 2	A	Alice	3	90	C	Chris	8	100	0	0	0	0	0	0	0	0	0

* Features from one-hot encoding of activity and resource omitted for readability

Marco Comuzzi mcomuzzi@unistac.kr

ERROR CATEGORY	ERROR TYPE	IMPERFECTION PATTERN?	HOW TO INJECT ERROR DETAILS
Errors on categorical labels (e.g. activity, resources)	Missing labels	NO	Randomly delete activity label of X% events in a log Randomly delete resource label of X% events in a log (if available) X=5, 10, 20, 30, 40, 50% File Names: <log>-MISSING-<label-type>-<X> Examples: “BPIC12-MISSING-activity-20”, “Pub-MISSING-resource-50”
	Syntactic accuracy: values outside of a pre-defined domain (e.g. typos)	#5 Polluted Label #10 Distorted Label	A) POLLUTED LABELS Randomly pollute activity labels of X% of events in a log X=5, 10, 20, 30, 40, 50% *A1) With random string:* How=“[Activity]_[a-zA-Z:{8}]_[Resource]” *A2) Without random string:* How=“[Activity]_[Resource]” B) DISTORTED LABELS Randomly distort activity labels of X% events in a log, X=5, 10, 20, 30, 40, 50% Type of distortion is chosen randomly: How=“[Activity:random(Skip,Insert,proximity,interchange,uplow)]” File Names: <log>-<type>-<label-type>-<X> Example: BPIC12-DISTORT-activity-30, Pub-POLLUTED-RANDOM-activity-40, BPIC11-POLLUTED-NORND-activity-5
	Semantic accuracy: homonym labels	“#7 Homonymous labels”	Define homonym labels (using ontology), ~X% of labels are substituted by a homonym For now: fixed homonym set, only for credit-card and pub log (see file) Do not use the same homonym: if X = 30%, then 5% of errors concern one homonym, 5% another one etc. File names: Credit-HOMONYM-30
	Semantic accuracy: synonym labels	“#6 Synonym labels”	Define synonym labels (using ontology), ~X% of labels are substituted by a synonym For now: fixed synonym set, only for credit-card and pub log (see file) Do not use the same activity synonym: if X = 30%, then 5% of errors concern one activity, 5% another one etc. File names: Credit-SYNONYM-30

Next activity prediction (SoTA LSTM architecture)
Up to 50% of events in a log affected by pattern
Performance on a “clean” test set

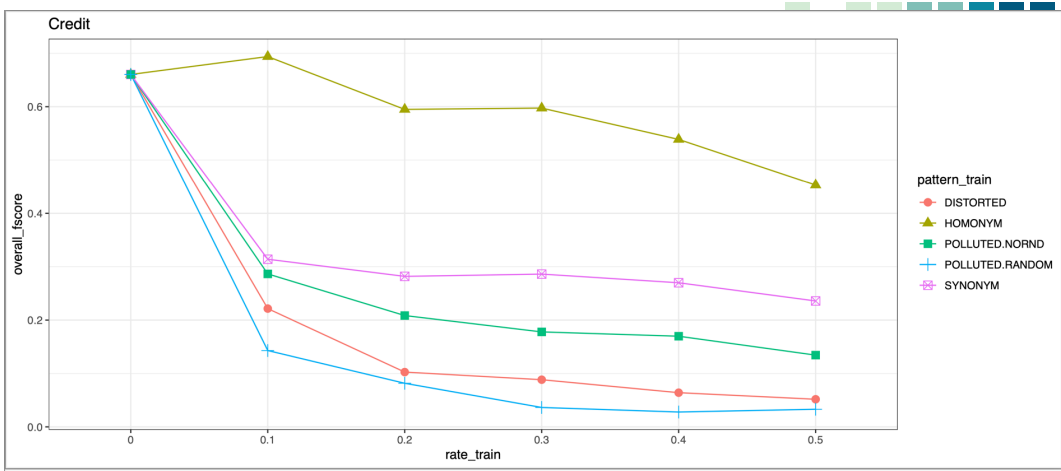


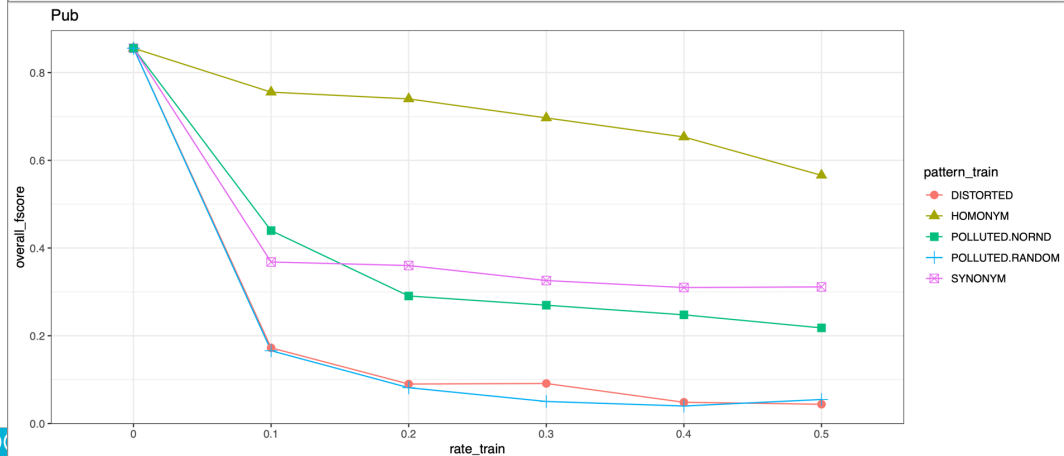
The higher the number of new activity labels introduced by errors, the lower the performance

Marco Comuzzi mcomuzzi@unistac.kr

Next activity prediction, synthetic logs

Impact of homonym labels on PPM performance is limited (reduced activity labels variability)





Some (very preliminary!) insights for a knowledge base

Focus first on errors that increase label variability (typos, random errors)

Consider synonym labels only later

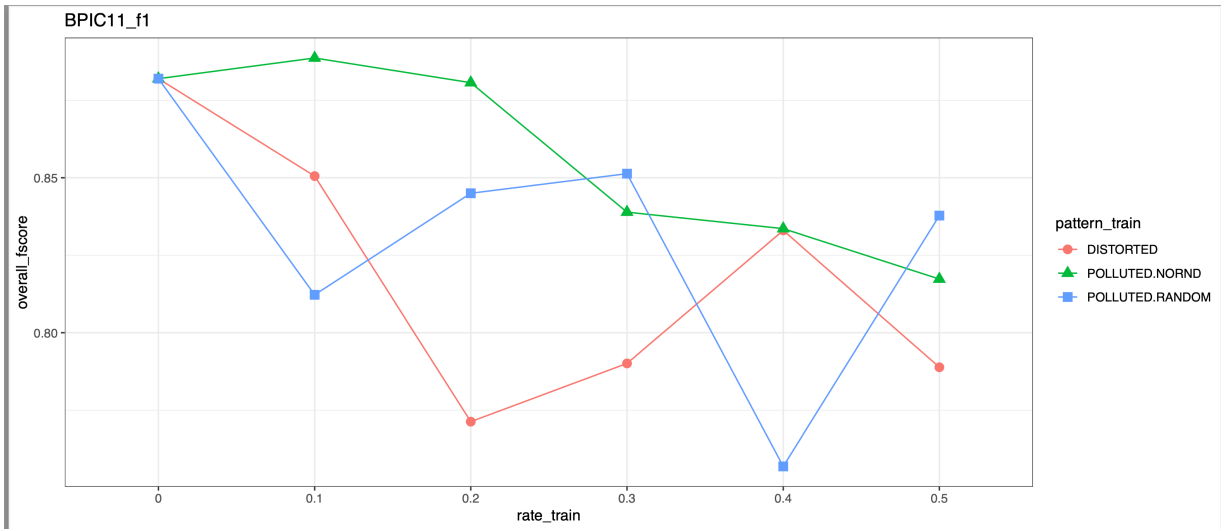
Homonym labels less critical

Marco Comuzzi mcomuzzi@unistac.kr

Outcome prediction?

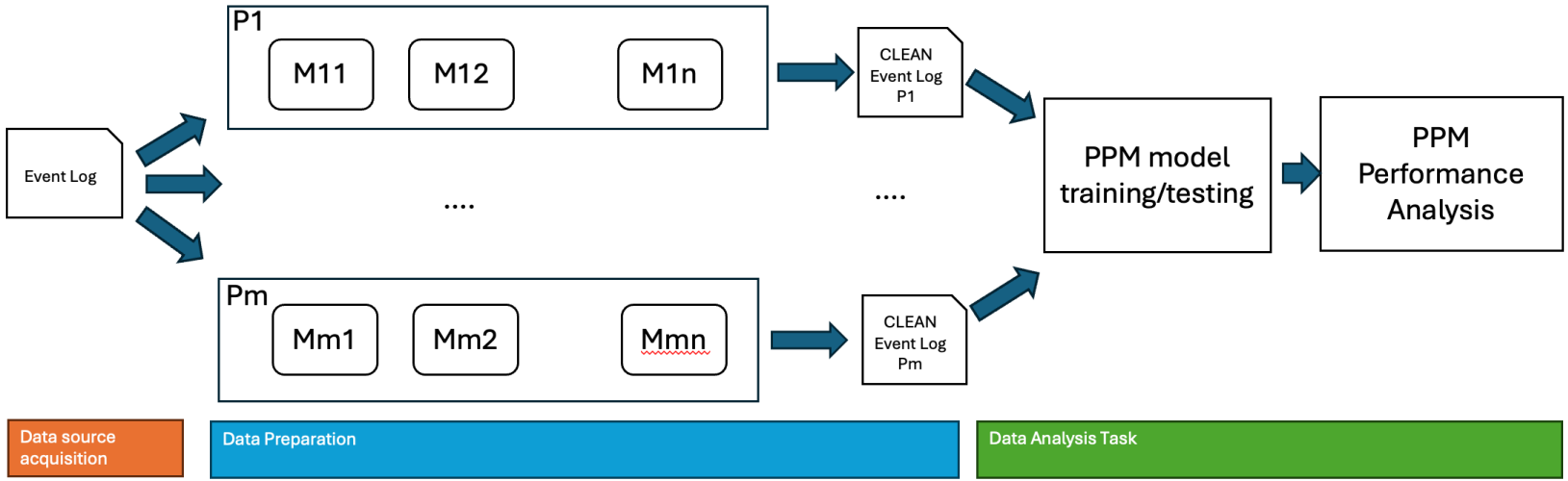
Hard to see patterns

Outcome label does not depend on which activities are executed?



Marco Comuzzi mcomuzzi@unistac.kr

Step 2: Building an (optimized) pipeline



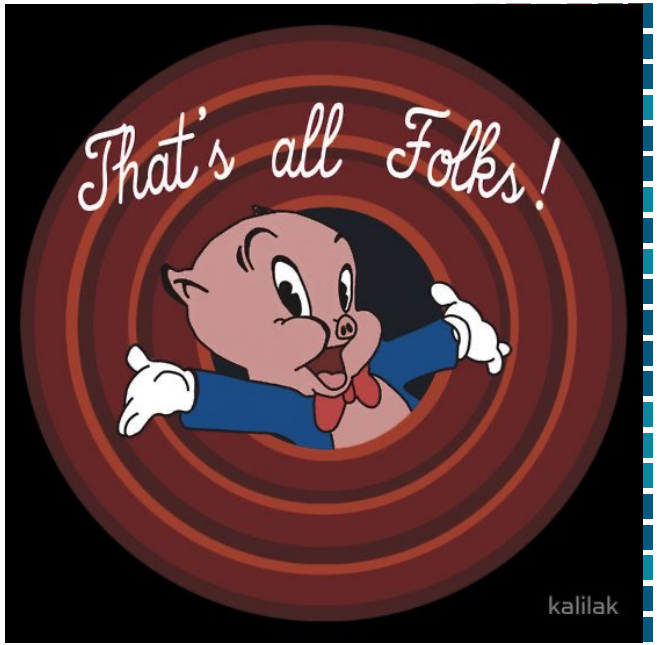
Marco Comuzzi mcomuzzi@unistac.kr

To conclude...

Process mining is now a multi-million dollar industry worldwide

Data quality can be a major hurdle in process mining projects

We need tools to create labelled logs with real-world data quality problems to approach systematically the event log preparation phase



Marco Comuzzi mcomuzzi@unistac.kr

THANKS!

https://marco-comuzzi.github.io

http://iel.unist.ac.kr/

@dr_bsad

mcomuzzi@unist.ac.kr