Prix de Thèse GDR GPL 2024

A Data Mining Perspective on Explainable AIOps with Applications to Software Maintenance

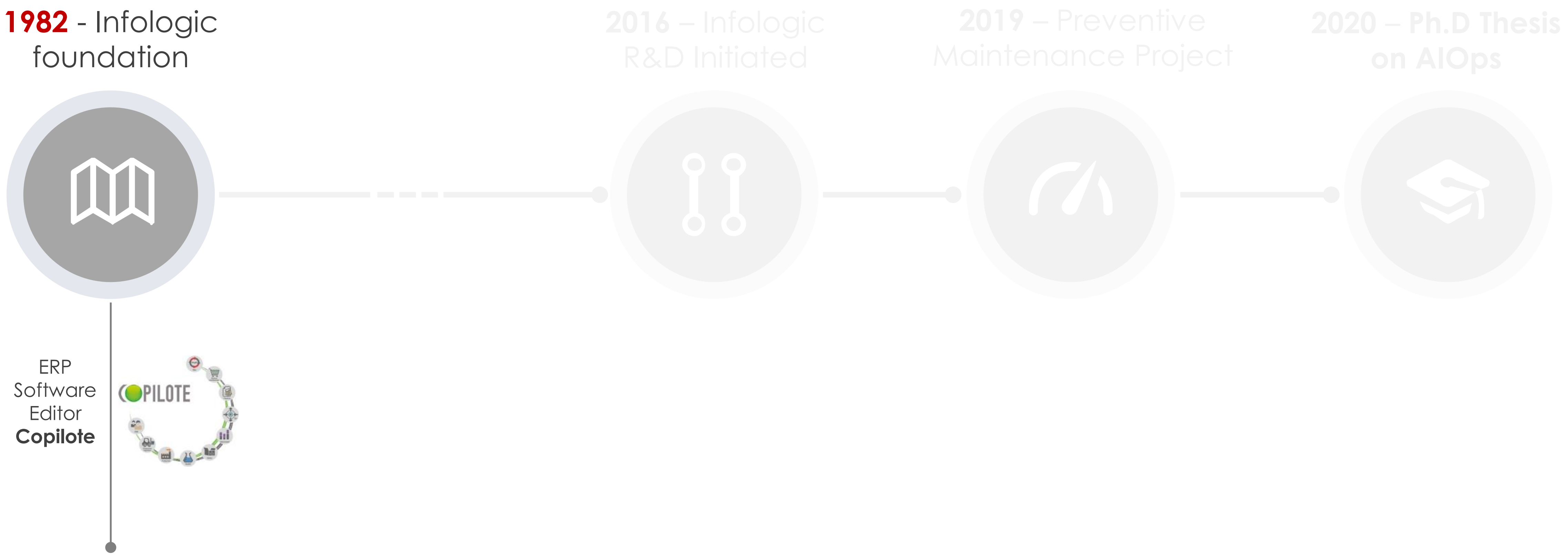
Presented by **Youcef REMIL**

Proposed by **Infologic R&D**

Advised by **Pr. Jean-François BOULICAUT**
Dr. Mehdi KAYTOUE
Dr. Anes BENDIMERAD

June 04, 2024

Introduction and Motivation – CIFRE Thesis



1982 - Infologic foundation

2016 – Infologic R&D Initiated

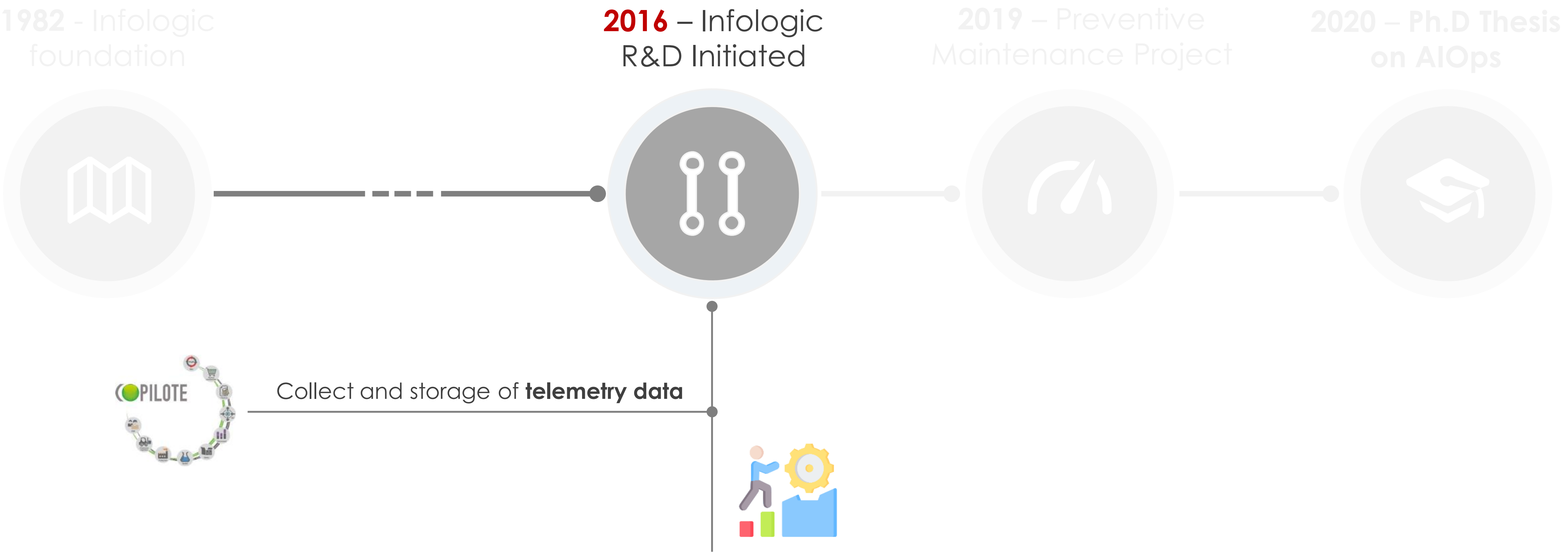
2019 – Preventive Maintenance Project

2020 – Ph.D Thesis on AIOps

ERP Software Editor **Copilote**

- Significant annual growth
- More than **600 sites**
- Over **200K workstations**

Introduction and Motivation – CIFRE Thesis



1982 - Infologic foundation

2016 – Infologic R&D Initiated

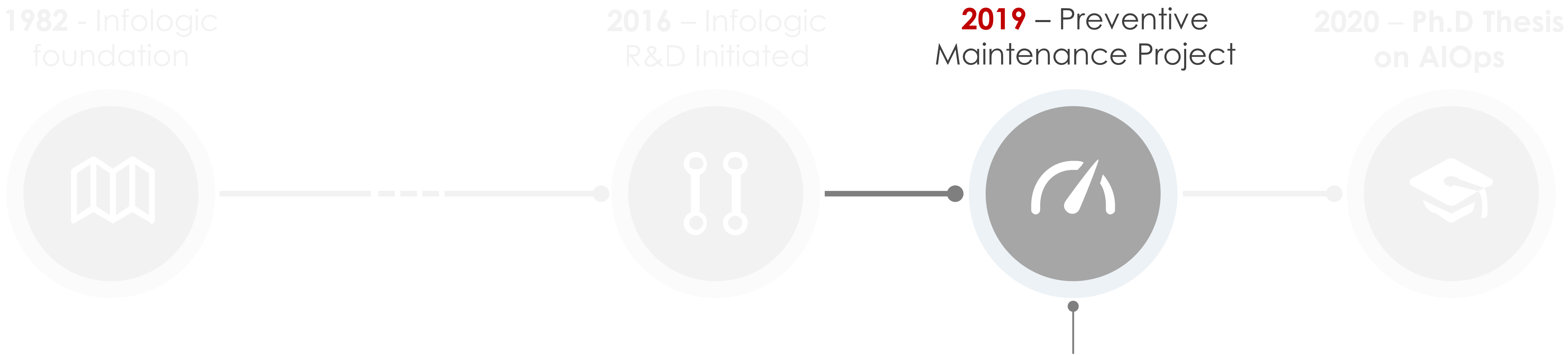
2019 – Preventive Maintenance Project

2020 – Ph.D Thesis on AIOps

Collect and storage of **telemetry data**

- Boosting **efficiency/reliability**
- Service **quality**
- Need for **automation**

Introduction and Motivation – CIFRE Thesis



1982 - Infologic foundation

2016 – Infologic R&D Initiated

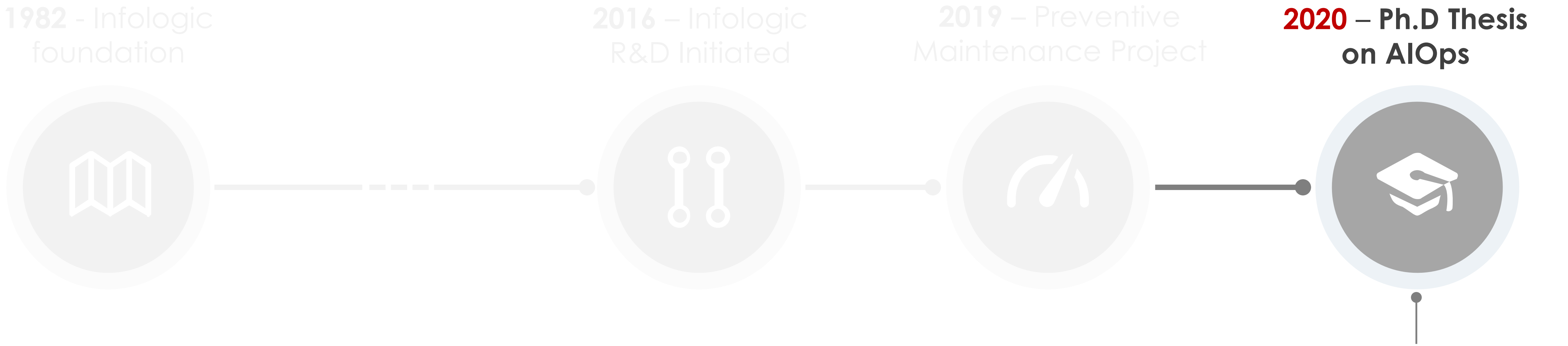
2019 – Preventive Maintenance Project

2020 – Ph.D Thesis on AIOps

- **Data-centric** approach
- **Real-time** monitoring
- **Proactive** maintenance
- **AIOps*** and Automation

*AIOps: **AI** for **Op**erating **S**ystems [Pankaj Prasad and Charley Rich. Market guide for AIOps platforms]

Introduction and Motivation – CIFRE Thesis



1982 - Infologic foundation

2016 – Infologic R&D Initiated

2019 – Preventive Maintenance Project

2020 – **Ph.D Thesis on AIOps**

- Study of **AIOps** field
- **Limitations** of AIOps
- **Development** of effective AIOps solutions
- **Applicative** and **Research Contributions**

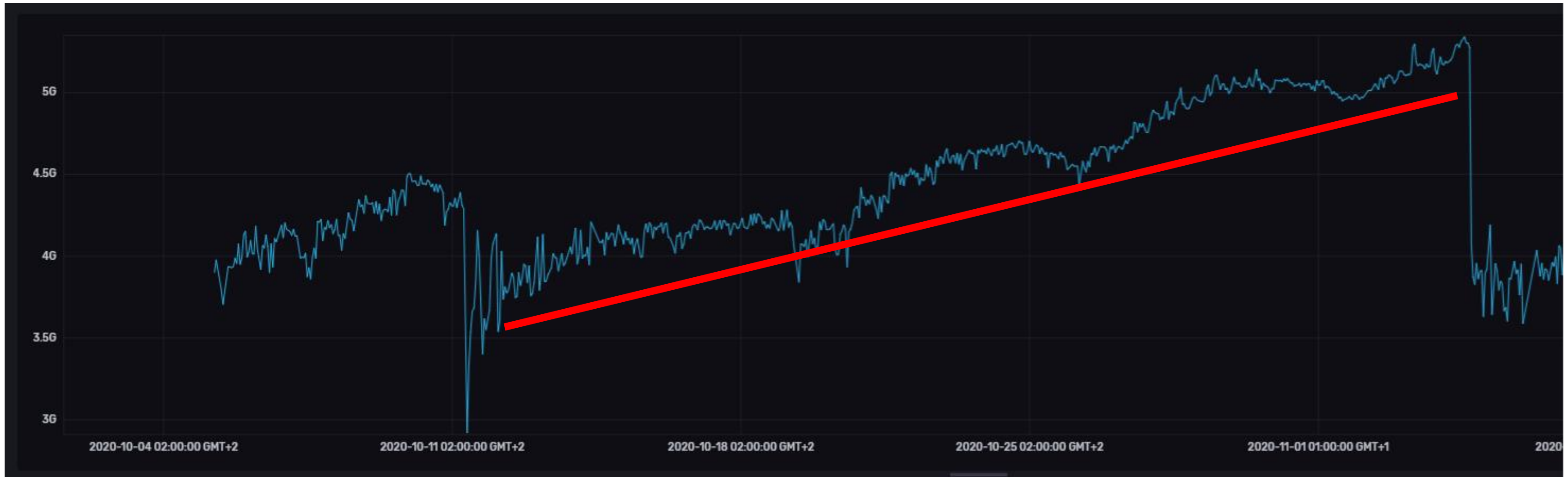
***AIOps: AI** for **Op**erating **S**ystems [Pankaj Prasad and Charley Rich. Market guide for AIOps platforms]

Introduction and Motivation

❑ Real pain points of maintenance routines at Infologic

Lack of standardized and automated maintenance routines with higher costs

- Relying mostly on corrective maintenance



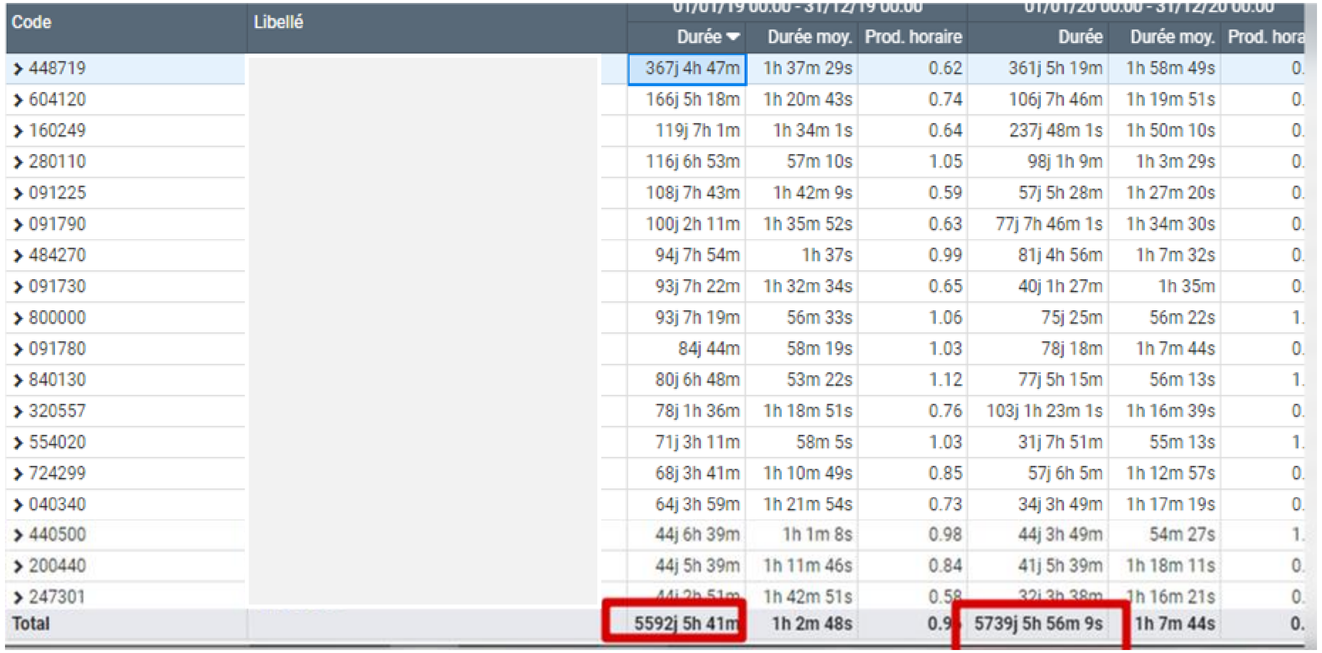
- Example: A detectable **memory leak** at a customer's premises (with **+€450m** annual revenue) **blocked** the departure of all delivery trucks from a factory for **30 minutes**.

Introduction and Motivation

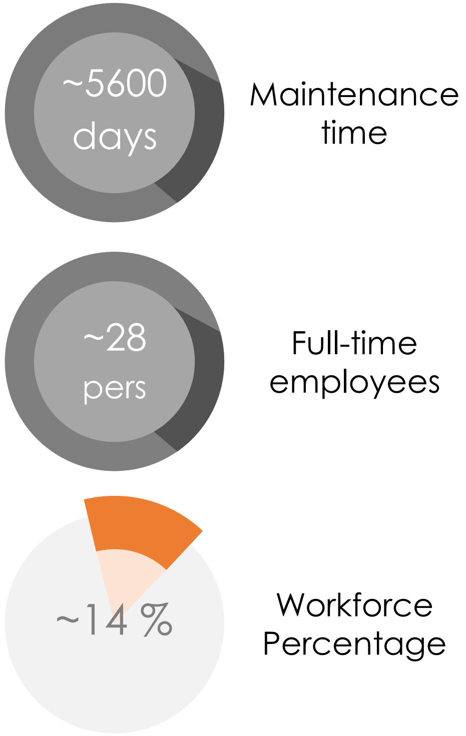
❑ Real pain points of maintenance routines at Infologic

Lack of standardized and automated maintenance routines with higher costs

- Higher human and resource costs [statistics by the end of 2019]



Code	Libellé	01/01/19 00.00 - 31/12/19 00.00			01/01/20 00.00 - 31/12/20 00.00		
		Durée ▾	Durée moy.	Prod. horaire	Durée	Durée moy.	Prod. hora
› 448719		367j 4h 47m	1h 37m 29s	0.62	361j 5h 19m	1h 58m 49s	0.
› 604120		166j 5h 18m	1h 20m 43s	0.74	106j 7h 46m	1h 19m 51s	0.
› 160249		119j 7h 1m	1h 34m 1s	0.64	237j 48m 1s	1h 50m 10s	0.
› 280110		116j 6h 53m	57m 10s	1.05	98j 1h 9m	1h 3m 29s	0.
› 091225		108j 7h 43m	1h 42m 9s	0.59	57j 5h 28m	1h 27m 20s	0.
› 091790		100j 2h 11m	1h 35m 52s	0.63	77j 7h 46m 1s	1h 34m 30s	0.
› 484270		94j 7h 54m	1h 37s	0.99	81j 4h 56m	1h 7m 32s	0.
› 091730		93j 7h 22m	1h 32m 34s	0.65	40j 1h 27m	1h 35m	0.
› 800000		93j 7h 19m	56m 33s	1.06	75j 25m	56m 22s	1.
› 091780		84j 44m	58m 19s	1.03	78j 18m	1h 7m 44s	0.
› 840130		80j 6h 48m	53m 22s	1.12	77j 5h 15m	56m 13s	1.
› 320557		78j 1h 36m	1h 18m 51s	0.76	103j 1h 23m 1s	1h 16m 39s	0.
› 554020		71j 3h 11m	58m 5s	1.03	31j 7h 51m	55m 13s	1.
› 724299		68j 3h 41m	1h 10m 49s	0.85	57j 6h 5m	1h 12m 57s	0.
› 040340		64j 3h 59m	1h 21m 54s	0.73	34j 3h 49m	1h 17m 19s	0.
› 440500		44j 6h 39m	1h 1m 8s	0.98	44j 3h 49m	54m 27s	1.
› 200440		44j 5h 39m	1h 11m 46s	0.84	41j 5h 39m	1h 18m 11s	0.
› 247301		44j 2h 51m	1h 42m 51s	0.58	32j 3h 38m	1h 16m 21s	0.
Total		5592j 5h 41m	1h 2m 48s	0.9	5739j 5h 56m 9s	1h 7m 44s	0.



Introduction and Motivation

❑ Real pain points of maintenance routines at Infologic

Inefficient incident triage and classification

- Need for automatic **assigning**, **ranking** and **classification**
- Problem of tossing sequence*
- Presence of recurring similar issues in historical maintenance calls

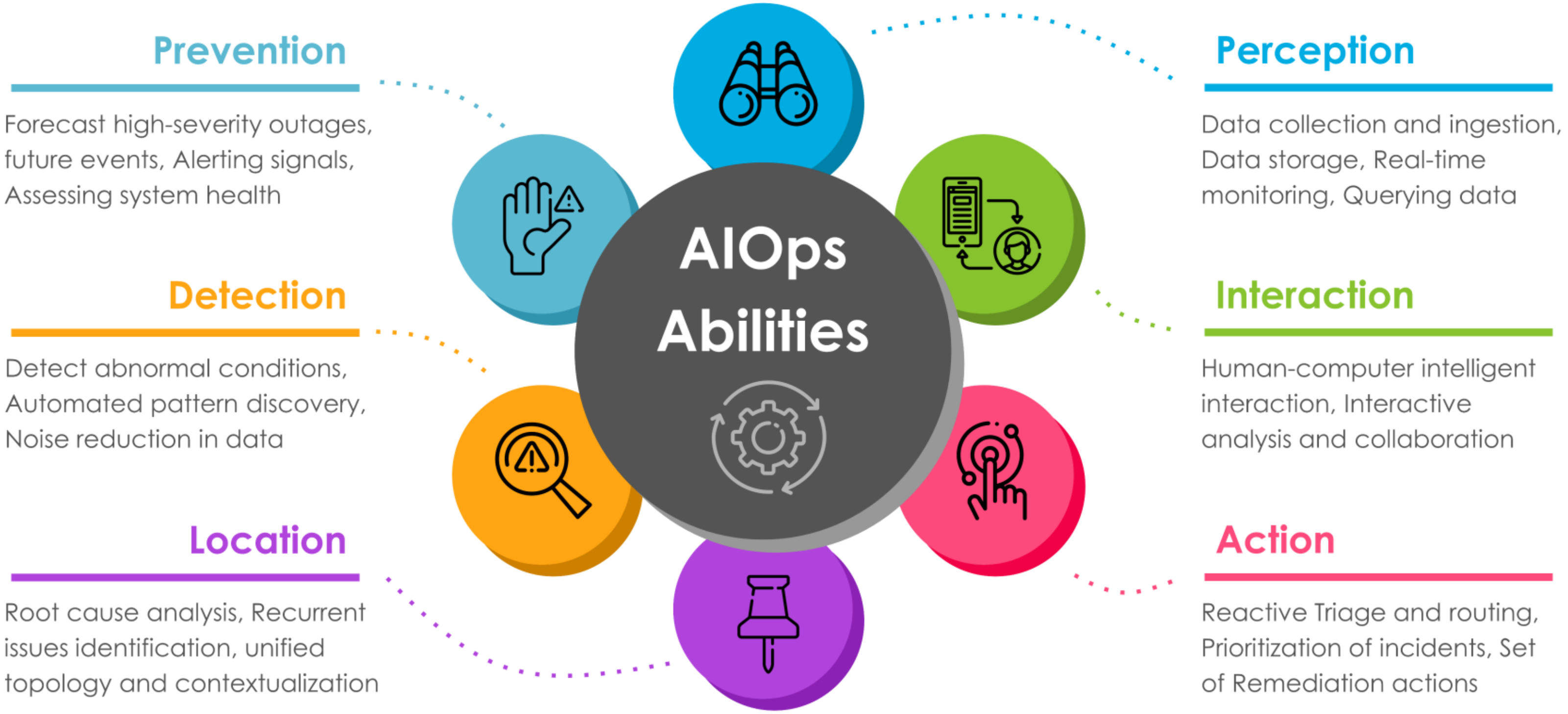
Ineffective root cause analysis and incident correlation

- Need for deep fault localization and figure out dependencies among components and services

*__Xie et al.__, Bug Triaging Based on Tossing Sequence Modeling. In **Journal of Computer Science and Technology 2019**

Introduction and Motivation

❑ Capabilities of AI for Operating Systems (AIOps)†



AIOps Abilities

Prevention

Forecast high-severity outages, future events, Alerting signals, Assessing system health

Detection

Detect abnormal conditions, Automated pattern discovery, Noise reduction in data

Location

Root cause analysis, Recurrent issues identification, unified topology and contextualization

Perception

Data collection and ingestion, Data storage, Real-time monitoring, Querying data

Interaction

Human-computer intelligent interaction, Interactive analysis and collaboration

Action

Reactive Triage and routing, Prioritization of incidents, Set of Remediation actions

†**Remil et al.** AIOps Solutions for Incident Management: Technical Guidelines and A Comprehensive Literature Review, In **TOSEM 2023** [Under Submission]
*__Dang et al.__ AIOps: real-world challenges and research innovations. In **ICSE** 2019

Introduction and Motivation

❑ Research challenges of AIOps addressed in this thesis

Introduction and Motivation

- ❑ Research challenges of AIOps addressed in this thesis

Novel and Unstructured Field

AIOps lacks unified terminology, complete taxonomy, desiderata, technical details

Introduction and Motivation

❑ Research challenges of AIOps addressed in this thesis

Novel and Unstructured Field

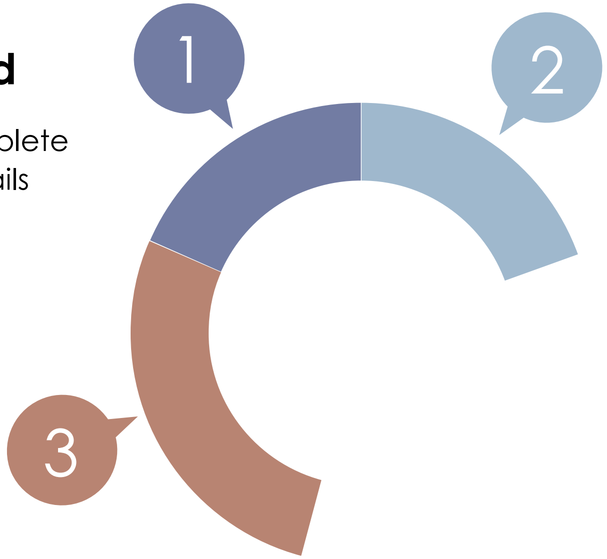
AIOps lacks unified terminology, complete taxonomy, desiderata, technical details

Data Requirements

Noisy, unstructured, missing, unlabeled, non-homogeneous and complex data

Introduction and Motivation

❑ Research challenges of AIOps addressed in this thesis



Novel and Unstructured Field

AIOps lacks unified terminology, complete taxonomy, desiderata, technical details

Data Requirements

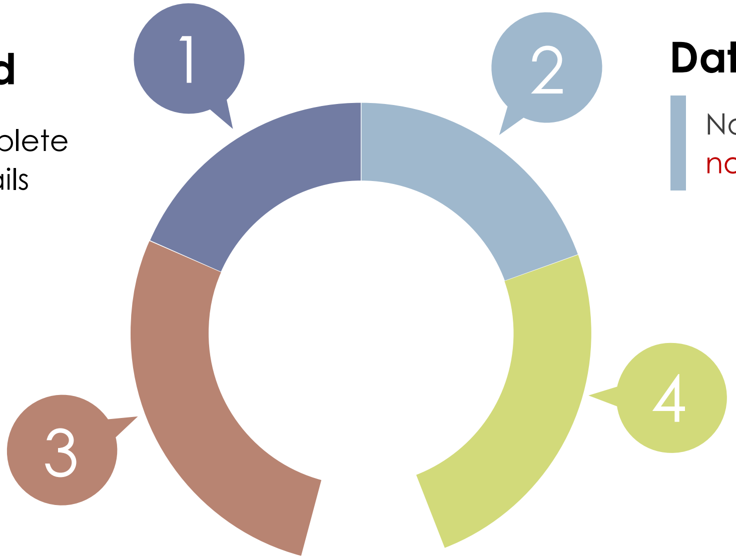
Noisy, unstructured, missing, unlabeled, non-homogeneous and complex data

Model Design

Impractical supervised methods, overlooking descriptive models

Introduction and Motivation

❑ Research challenges of AIOps addressed in this thesis



Novel and Unstructured Field

AIOps lacks unified terminology, complete taxonomy, desiderata, technical details

Data Requirements

Noisy, unstructured, missing, unlabeled, non-homogeneous and complex data

Model Design

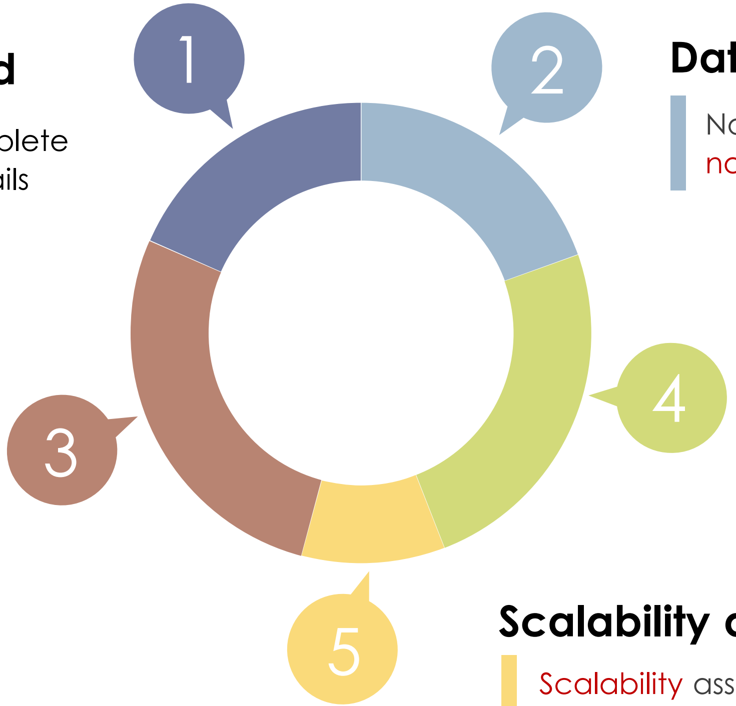
Impractical supervised methods, overlooking descriptive models

Interpretability

Best models are black box, transparency is preferred over performance

Introduction and Motivation

- ❑ Research challenges of AIOps addressed in this thesis



Novel and Unstructured Field

AIOps lacks unified terminology, complete taxonomy, desiderata, technical details

Data Requirements

Noisy, unstructured, missing, unlabeled, non-homogeneous and complex data

Model Design

Impractical supervised methods, overlooking descriptive models

Interpretability

Best models are black box, transparency is preferred over performance

Scalability and Robustness

Scalability assessment often overlooked, temporal and in-context evaluation

Introduction and Motivation

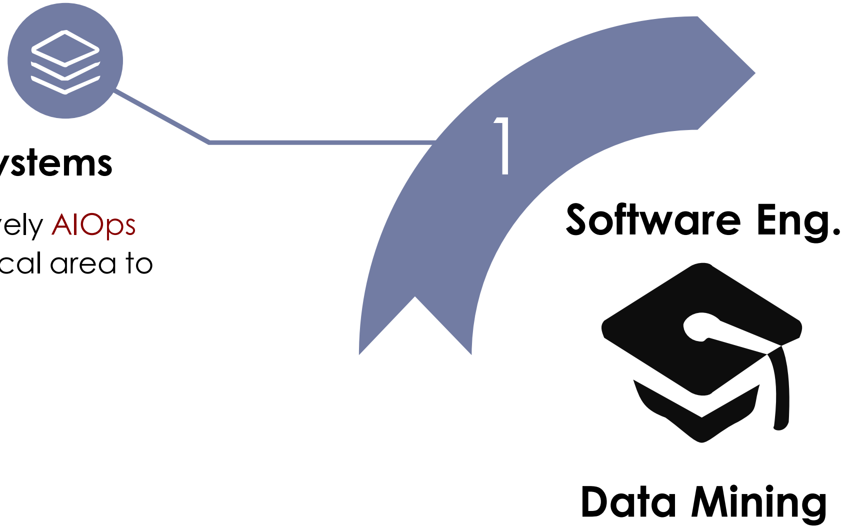
❑ Contributions and Key Research Areas

Software Eng.

Data Mining

Introduction and Motivation

❑ Contributions and Key Research Areas



AI for Operating Systems

Study Comprehensively AIOps research and technical area to address challenge 1

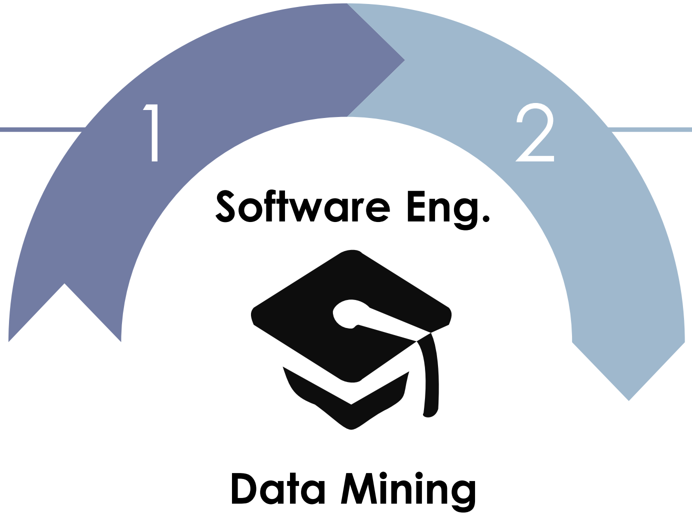
Remil et al. AIOps Solutions for Incident Management: Technical Guidelines and A Comprehensive Literature Review, In **TOSEM 2023 [Under revision, Core 2021, A*]**
Bendimerad, **Remil** et al. On-premise Infrastructure for AIOps in a Software Editor SME: An Experience Report, In **ESEC/FSE 2023 [Published, Core 2021, A*]**

Introduction and Motivation

❑ Contributions and Key Research Areas

AI for Operating Systems

Study Comprehensively AIOps research and technical area to address challenge 1



1

2

Software Eng.

Data Mining

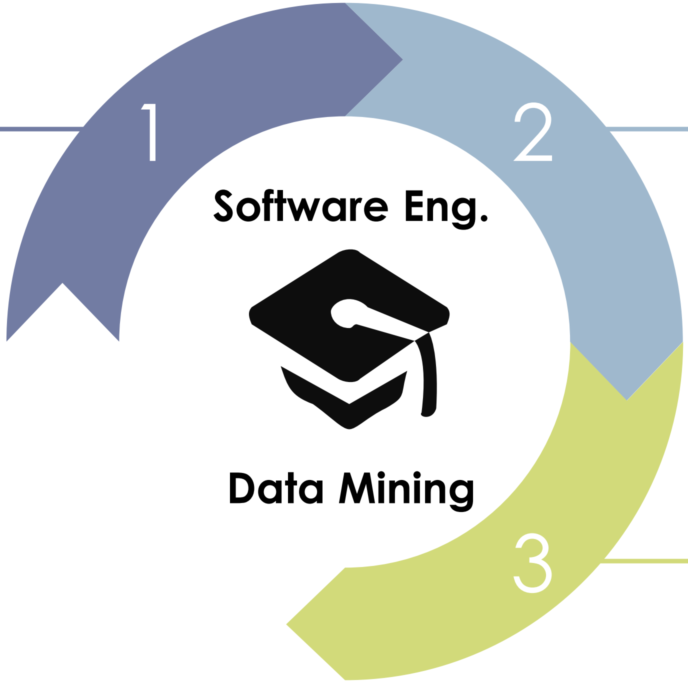
Supervised Rule Discovery

Study SD and EMM to address challenges 2 and 3 related to data quality and models

Remil et al. What makes my queries slow: Subgroup Discovery for SQL Workload Analysis, In **ASE 2021 [Published, Core 2021, A*]**
Remil et al. Interpretable Summaries of Black Box Incident Triaging with Subgroup Discovery, In **DSAA 2021 [Published, Core 2021, A]**
Remil et al. Mining Java Memory Errors using Subjective Interesting Subgroups with Hierarchical Targets, In **ICDMW 2023 [Published, Workshop]**

Introduction and Motivation

❑ Contributions and Key Research Areas



AI for Operating Systems

Study Comprehensively AIOps research and technical area to address challenge 1

Supervised Rule Discovery

Study SD and EMM to address challenges 2 and 3 related to data quality and models

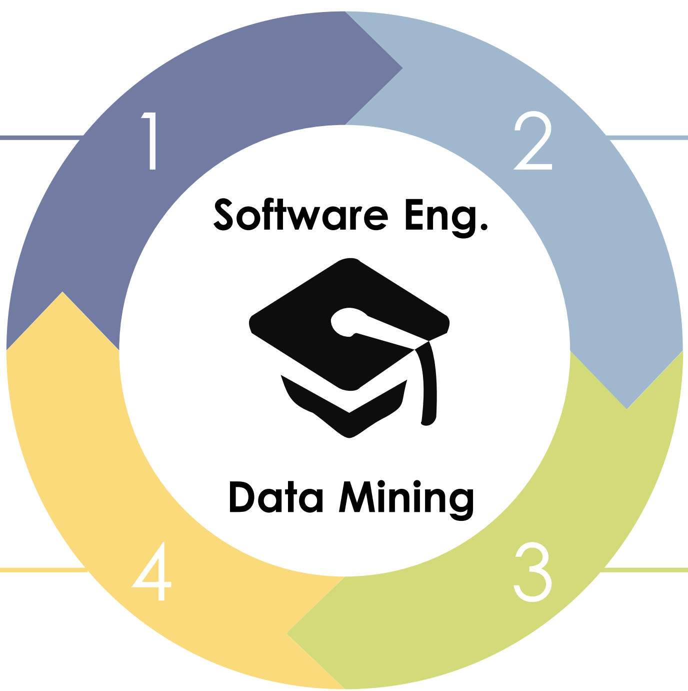
Explainable AI

Study Explainable AI to cope with challenge 4 of interpreting black box models

Remil et al. Interpretable Summaries of Black Box Incident Triaging with Subgroup Discovery, In **DSAA 2021 [Published, Core 2021, A]**
Remil et al. Découverte de Sous-groupes Interprétables pour le Triage d'incidents, In **EGC 2022 [Published, National Conf]**

Introduction and Motivation

❑ Contributions and Key Research Areas



AI for Operating Systems

Study Comprehensively AIOps research and technical area to address challenge 1

Supervised Rule Discovery

Study SD and EMM to address challenges 2 and 3 related to data quality and models

Locality Sensitive Hashing

Study LSH for fast and efficient similarity search to tackle challenge 5

Explainable AI

Study Explainable AI to cope with challenge 4 of interpreting black box models

Remil et al. DeepLSH: Deep Locality-Sensitive Hash Learning for Fast and Efficient Near-Duplicate Crash Report Detection, In **ICSE 2024 [Published, Core 2021 A*]**

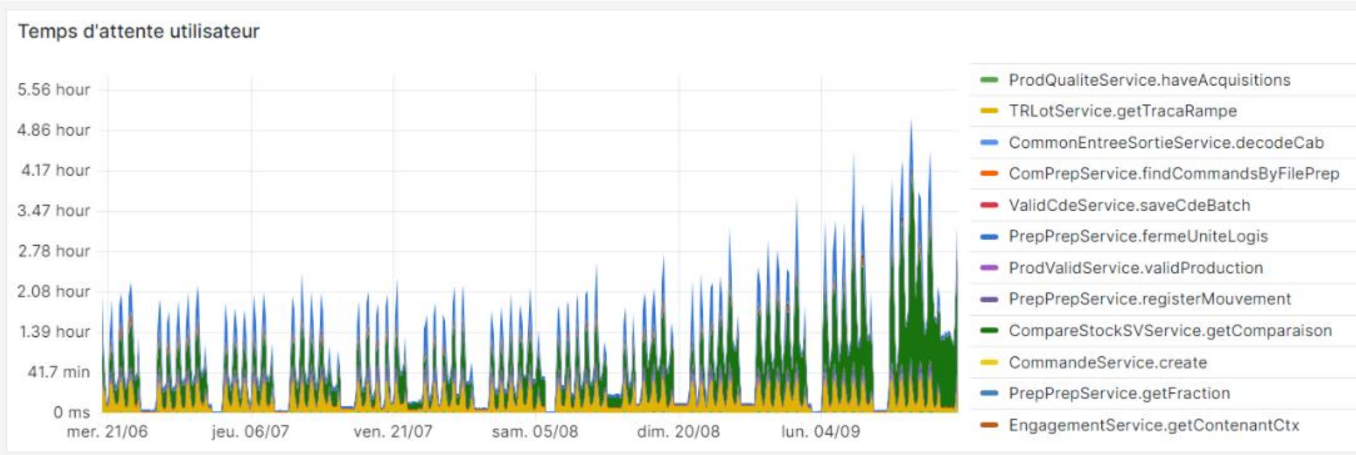
Subgroup Discovery for SQL Workloads

Database server

Huge SQL Workload

- SQL queries
- execution time
- # rows
- History sessions …

Performance analysis



- Query execution plan
- …

Database schema

DBAs

- Slow queries
- Index recommendation
- Concurrency issues
- …

Subgroup Discovery for SQL Workloads

Need for a **generic** framework to analyse **batches** of SQL queries and bring answers to the question:

How to characterize SQL queries that foster some properties of interest?

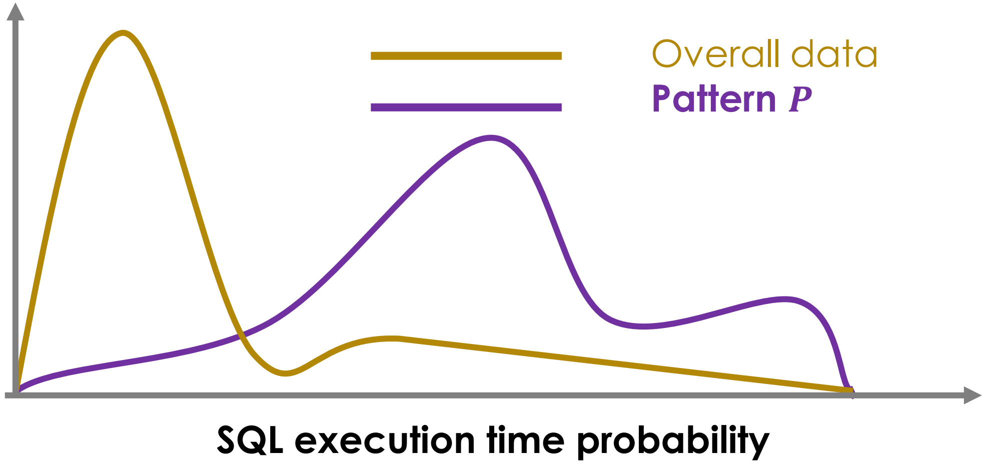
Subgroup Discovery for SQL Workloads

Need for a **generic** framework to analyse **batches** of SQL queries and bring answers to the question:
How to characterize SQL queries that foster some properties of interest?

Illustrative example of SQL queries

Predicates		Topology		…	Targets	
ik	date	Db. version	…	…	time	slow
1	**1**	**V2**	…	…	**14**	**1**
0	1	V1	…	…	2	0
…	…	…	…	…	…	…
1	0	V2	…	…	3	0
0	**1**	**V2**			**25**	**1**

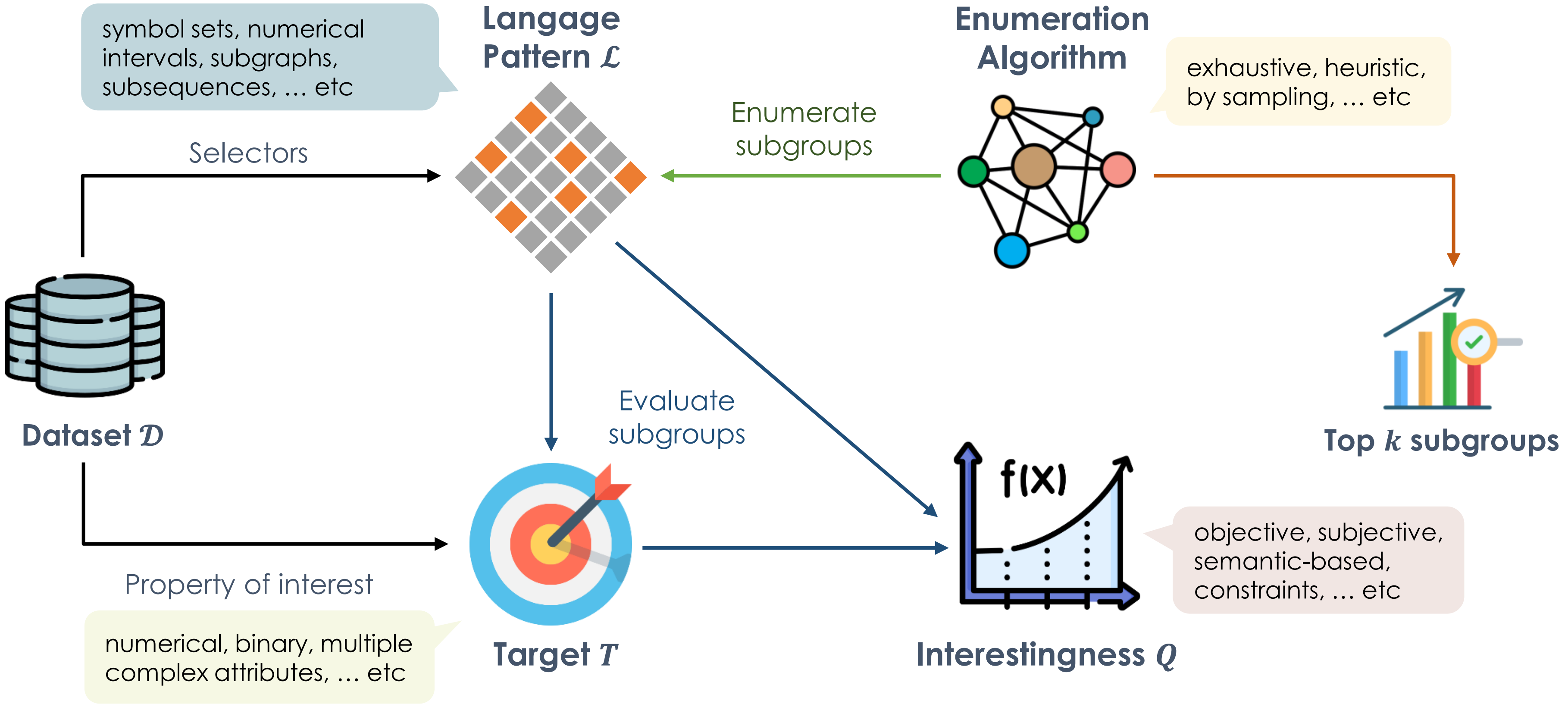
Pattern *P*: Predicate = verrou.date ∧ Db. Version = V2 ⟶ slow queries



Atzmueller. Subgroup Discovery, In **DAMI 2015**
Wrobel. An algorithm for multi-relational discovery of subgroups. In **PKDD 1997**

Subgroup Discovery for SQL Workloads

Subgroup Discovery building blocks



symbol sets, numerical intervals, subgraphs, subsequences, … etc

Langage Pattern $\mathcal{L}$

Enumeration Algorithm

exhaustive, heuristic, by sampling, … etc

Enumerate subgroups

Selectors

Dataset $\mathcal{D}$

Evaluate subgroups

Top k subgroups

Property of interest

numerical, binary, multiple complex attributes, … etc

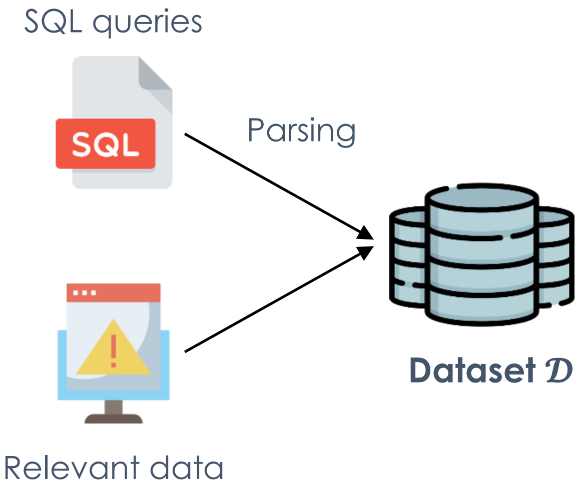
Target T

Interestingness Q

objective, subjective, semantic-based, constraints, … etc

Subgroup Discovery for SQL Workloads

SD building blocks for SQL Workload Analysis



Example of an SQL query

```
SELECT m.ik
FROM model AS m
JOIN prod AS p
WHERE m.ik = p.ik
  AND m.uex = p1
  AND (m.uex in collection0
       OR m.ik in collection1)
  AND (m.dossier = p3
GROUP BY m.ik
HAVING (COUNT(DISTINCT p.ik) = p2)
AND (SUM(m.nbembal) = MAX (p.nbembal))
```

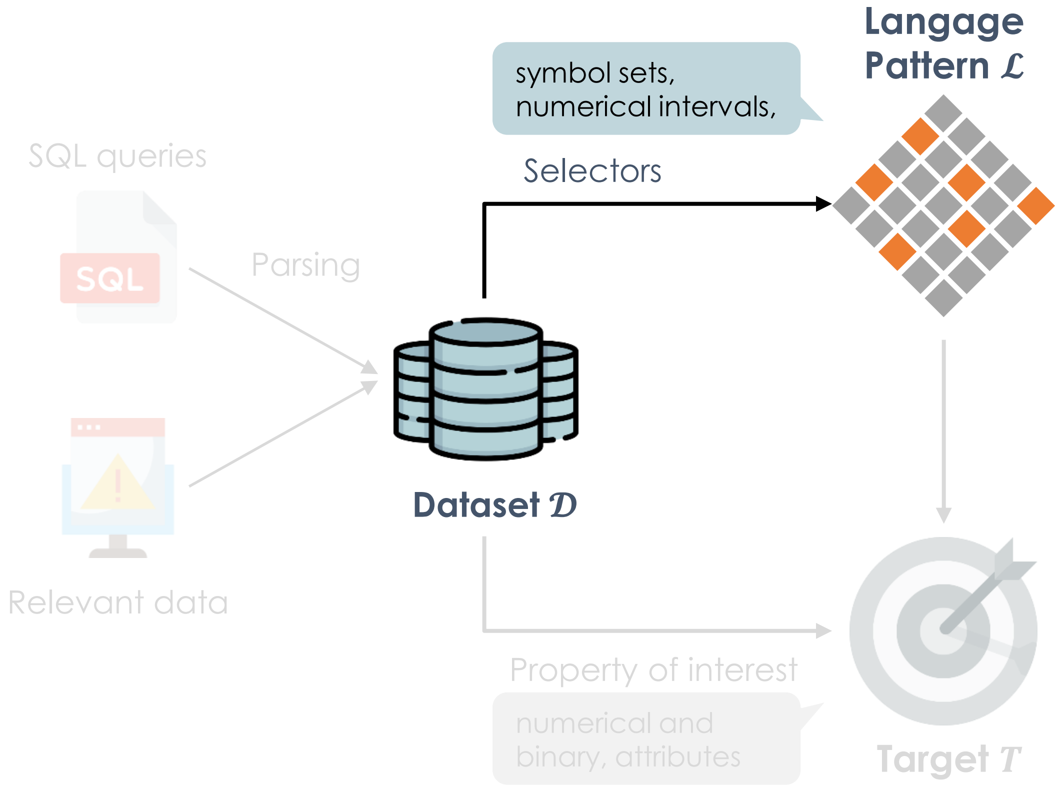
Our parser*

SELECT_model_ik	1
FROM_model	1
JOIN_prod	1
WHERE_model_ik	2
WHERE_model.uex	1
WHERE_model.dossier	1
WHERE_prod.ik	1
GROUPBY_ik	1
HAVING_prod_ik	1
HAVING_model.nbembal	1
HAVING_prod.nbembal	1
COUNT_prod.ik	1
SUM_model.nbembal	1
MAX_prod.nbembal	1

*https://github.com/klahnakoski/mo-sql-parsing/pull/26

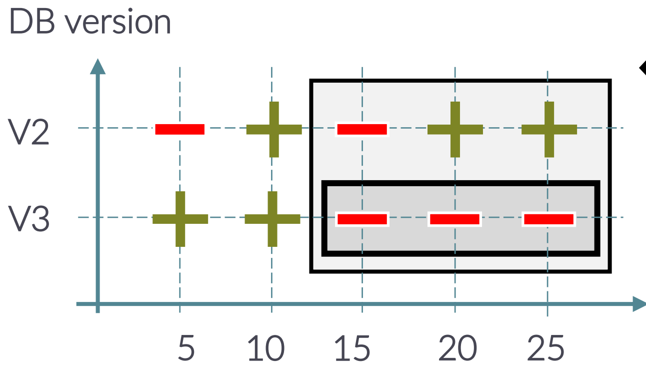
Subgroup Discovery for SQL Workloads

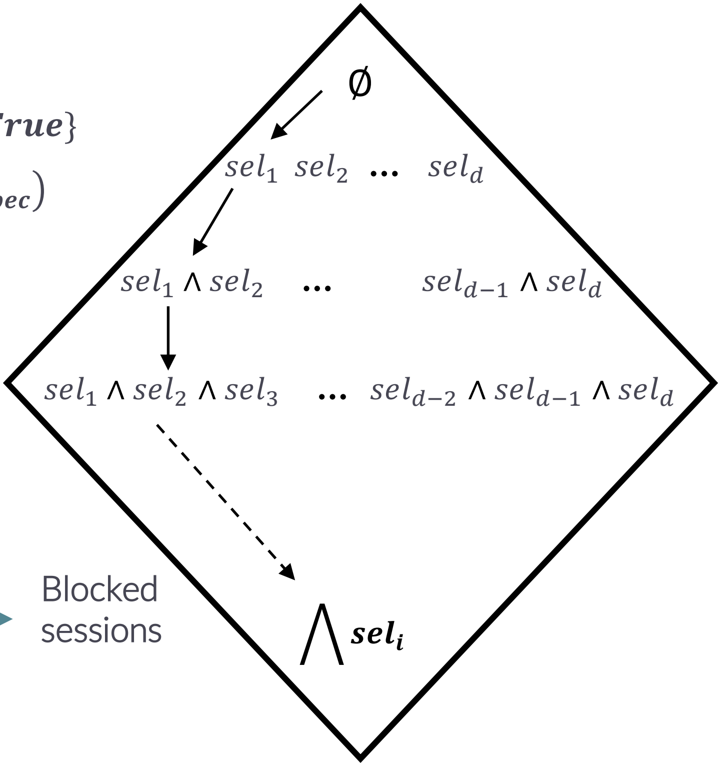
SD building blocks for SQL Workload Analysis


Langage Pattern $\mathcal{L}$
symbol sets, numerical intervals,
Selectors
SQL queries
SQL
Parsing
Relevant data
Dataset $\mathcal{D}$
Property of interest
numerical and binary, attributes
Target T

$$\text{Search Space } \mathcal{L} \xrightarrow{\text{Made of}} \text{Selectors} \xrightarrow{\text{Conjunctive combinations}} \text{Patterns}$$

- $sg(P) = ext(P) = \{c \in \mathcal{O} \mid P(o) = True\}$
- $P_{gen} \subset P_{spec} \Rightarrow sg(P_{gen}) \supseteq sg(P_{spec})$



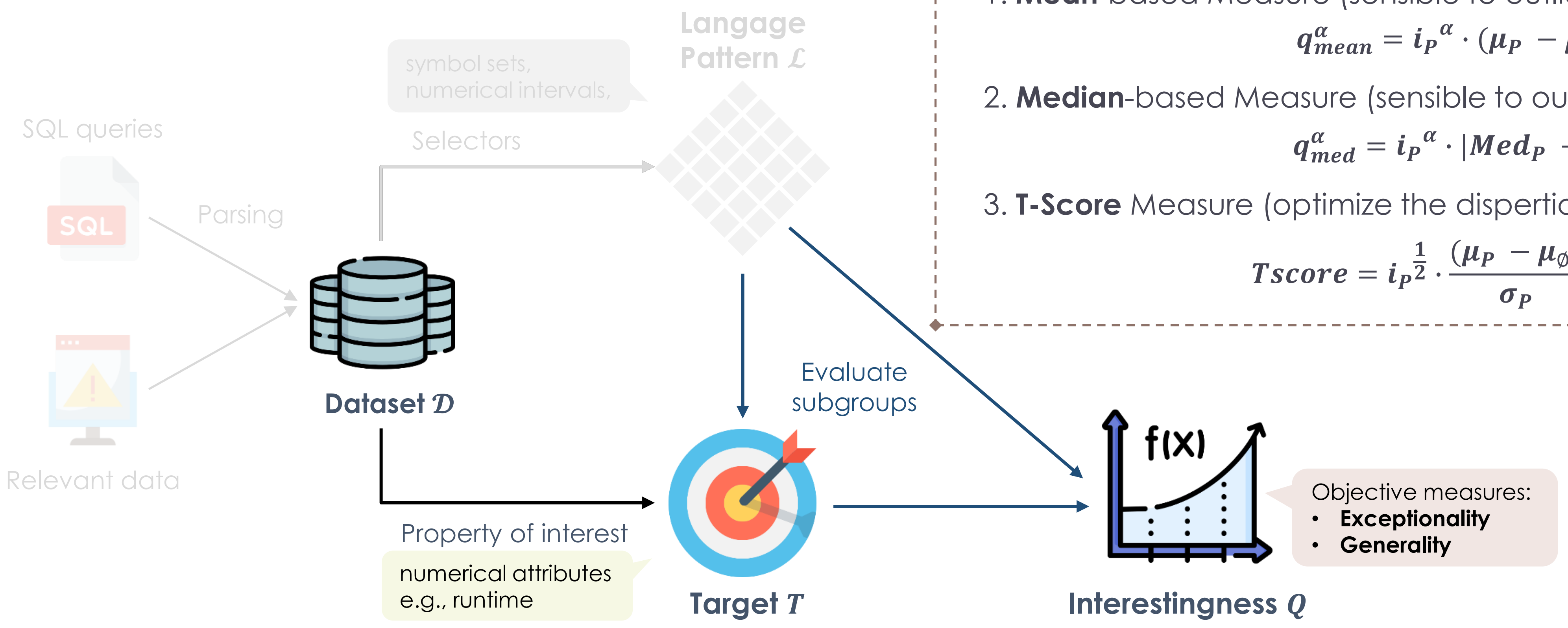

$\emptyset$
sel_1 sel_2 … sel_d
$sel_1 \wedge sel_2$ … $sel_{d-1} \wedge sel_d$
$sel_1 \wedge sel_2 \wedge sel_3$ … $sel_{d-2} \wedge sel_{d-1} \wedge sel_d$
$\bigwedge sel_i$

P_{gen} : $blockedSessions \in [15, 25]$

P_{spec} : $blockedSessions \in [15, 25] \wedge dbVersion = V3$

Subgroup Discovery for SQL Workloads

SD building blocks for SQL Workload Analysis



SQL queries

Parsing

Relevant data

symbol sets, numerical intervals,

Selectors

Langage Pattern $\mathcal{L}$

Dataset $\mathcal{D}$

Evaluate subgroups

Property of interest

numerical attributes e.g., runtime

Target T

Interestingness Q

Objective measures:

- **Exceptionality**
- **Generality**

Numerical measures used to evaluate the subgroup patterns*

1. **Mean**-based Measure (sensible to outliers)

$$q_{mean}^{\alpha} = i_P{}^{\alpha} \cdot (\mu_P - \mu_\emptyset)$$

2. **Median**-based Measure (sensible to outliers)

$$q_{med}^{\alpha} = i_P{}^{\alpha} \cdot |Med_P - Med_\emptyset|$$

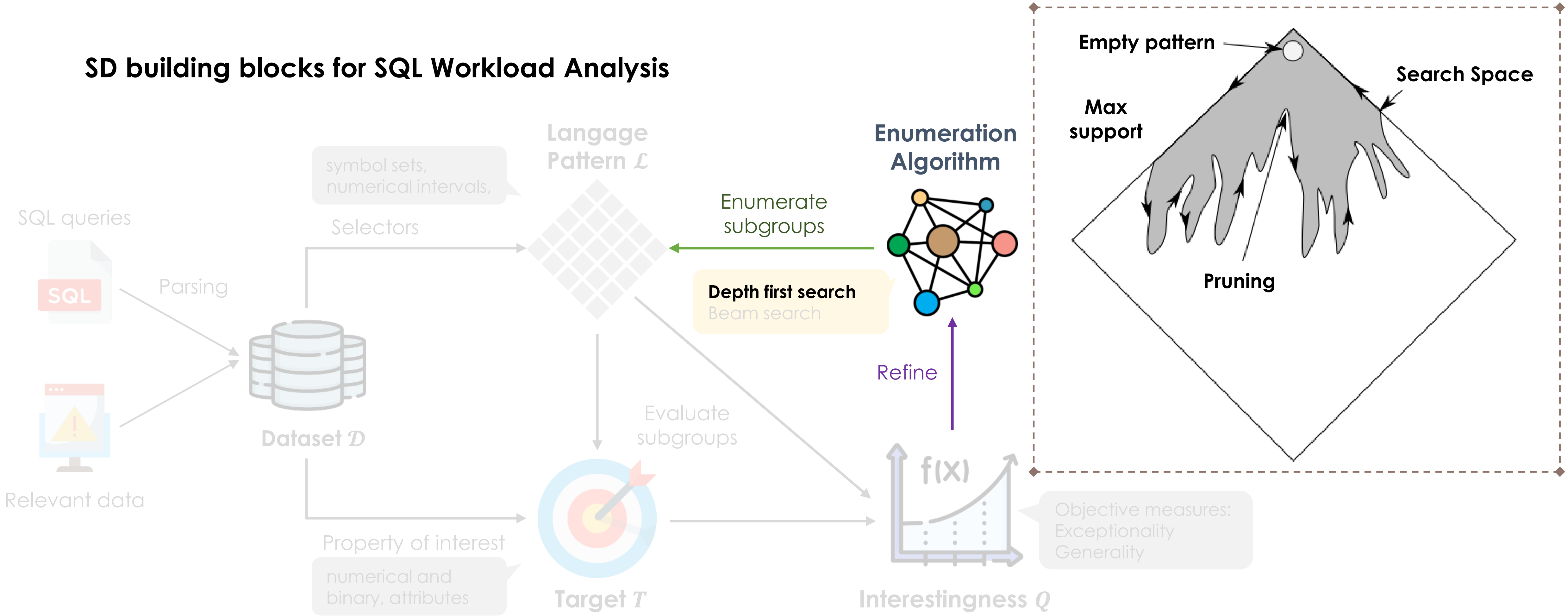
3. **T-Score** Measure (optimize the dispertion)

$$Tscore = i_P^{\frac{1}{2}} \cdot \frac{(\mu_P - \mu_\emptyset)}{\sigma_P}$$

*Lemmerich. Novel Techniques for Efficient and Effective Subgroup Discovery. **PhD thesis**, **2014**

Subgroup Discovery for SQL Workloads

SD building blocks for SQL Workload Analysis



SQL queries

SQL

Parsing

Relevant data

Dataset $\mathcal{D}$

Selectors

symbol sets, numerical intervals,

Langage Pattern $\mathcal{L}$

Property of interest

numerical and binary, attributes

Target T

Evaluate subgroups

Interestingness Q

Objective measures: Exceptionality Generality

Enumerate subgroups

Enumeration Algorithm

Depth first search

Beam search

Refine

Empty pattern

Search Space

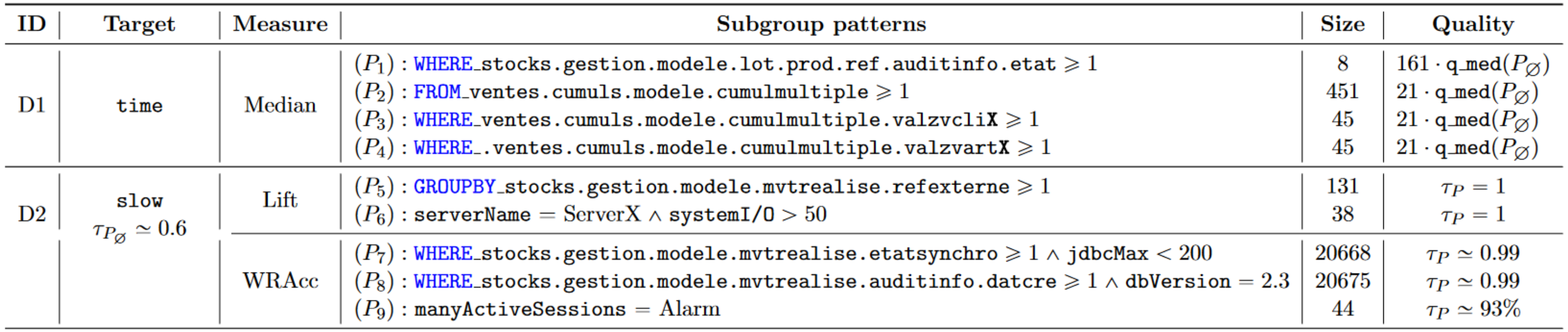
Max support

Pruning

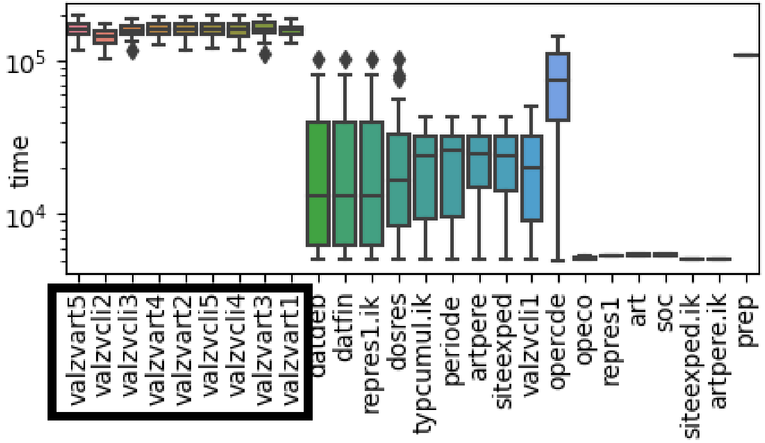
Lemmerich et al., Fast exhaustive subgroup discovery with numerical target concepts. In **DAMI 2016**

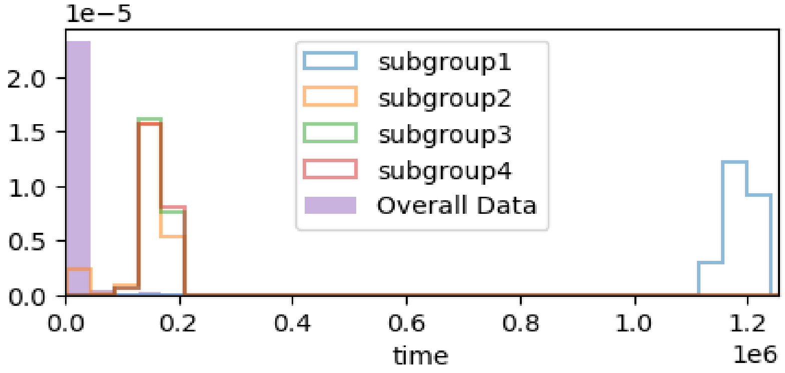
Subgroup Discovery for SQL Workloads

Results on a large workload of Hibernate queries made available by Infologic



ID	Target	Measure	Subgroup patterns	Size	Quality
D1	`time`	Median	(P_1) : WHERE_stocks.gestion.modele.lot.prod.ref.auditinfo.etat $\geqslant 1$	8	$161 \cdot$ q_med($P_\varnothing$)
			(P_2) : FROM_ventes.cumuls.modele.cumulmultiple $\geqslant 1$	451	$21 \cdot$ q_med($P_\varnothing$)
			(P_3) : WHERE_ventes.cumuls.modele.cumulmultiple.valzvcliX $\geqslant 1$	45	$21 \cdot$ q_med($P_\varnothing$)
			(P_4) : WHERE_.ventes.cumuls.modele.cumulmultiple.valzvartX $\geqslant 1$	45	$21 \cdot$ q_med($P_\varnothing$)
D2	`slow` $\tau_{P_\varnothing} \simeq 0.6$	Lift	(P_5) : GROUPBY_stocks.gestion.modele.mvtrealise.refexterne $\geqslant 1$	131	$\tau_P = 1$
			(P_6) : serverName = ServerX $\wedge$ systemI/O > 50	38	$\tau_P = 1$
		WRAcc	(P_7) : WHERE_stocks.gestion.modele.mvtrealise.etatsynchro $\geqslant 1 \wedge$ jdbcMax < 200	20668	$\tau_P \simeq 0.99$
			(P_8) : WHERE_stocks.gestion.modele.mvtrealise.auditinfo.datcre $\geqslant 1 \wedge$ dbVersion = 2.3	20675	$\tau_P \simeq 0.99$
			(P_9) : manyActiveSessions = Alarm	44	$\tau_P \simeq 93\%$

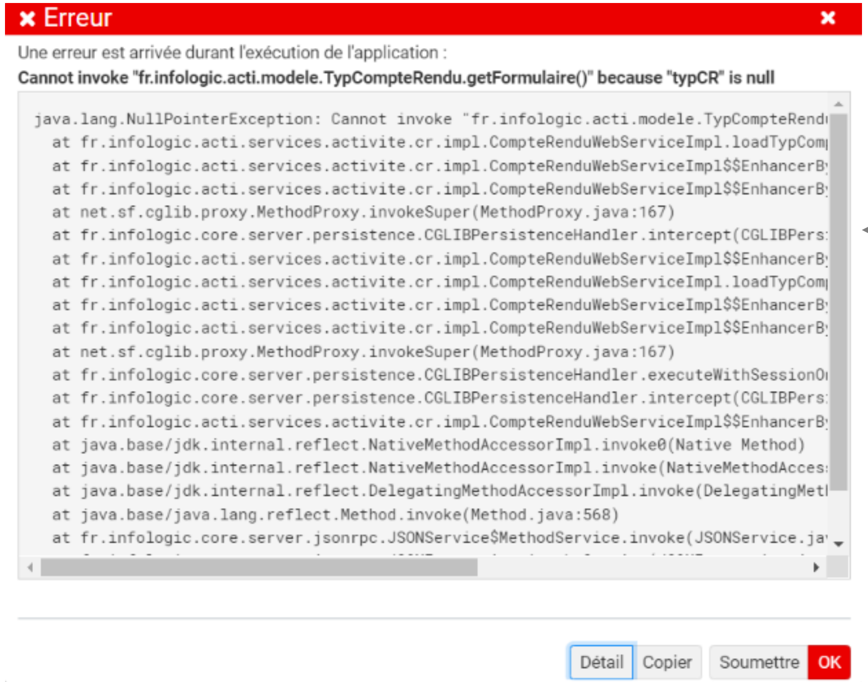




Enhancing Duplicate Crash Report Retrieval

Problem of deduplication*

A bug reported for the **ACTI** service



Une erreur est arrivée durant l'exécution de l'application :

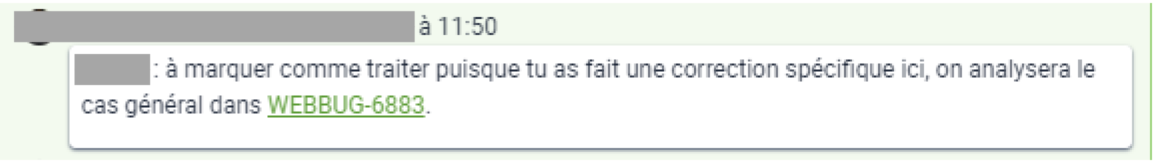
Cannot invoke "fr.infologic.acti.modele.TypCompteRendu.getFormulaire()" because "typCR" is null

```
java.lang.NullPointerException: Cannot invoke "fr.infologic.acti.modele.TypCompteRend
  at fr.infologic.acti.services.activite.cr.impl.CompteRenduWebServiceImpl.loadTypCom
  at fr.infologic.acti.services.activite.cr.impl.CompteRenduWebServiceImpl$$EnhancerB
  at fr.infologic.acti.services.activite.cr.impl.CompteRenduWebServiceImpl$$EnhancerB
  at net.sf.cglib.proxy.MethodProxy.invokeSuper(MethodProxy.java:167)
  at fr.infologic.core.server.persistence.CGLIBPersistenceHandler.intercept(CGLIBPers
  at fr.infologic.acti.services.activite.cr.impl.CompteRenduWebServiceImpl$$EnhancerB
  at fr.infologic.acti.services.activite.cr.impl.CompteRenduWebServiceImpl.loadTypCom
  at fr.infologic.acti.services.activite.cr.impl.CompteRenduWebServiceImpl$$EnhancerB
  at fr.infologic.acti.services.activite.cr.impl.CompteRenduWebServiceImpl$$EnhancerB
  at net.sf.cglib.proxy.MethodProxy.invokeSuper(MethodProxy.java:167)
  at fr.infologic.core.server.persistence.CGLIBPersistenceHandler.executeWithSessionO
  at fr.infologic.core.server.persistence.CGLIBPersistenceHandler.intercept(CGLIBPers
  at fr.infologic.acti.services.activite.cr.impl.CompteRenduWebServiceImpl$$EnhancerB
  at java.base/jdk.internal.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
  at java.base/jdk.internal.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAcces
  at java.base/jdk.internal.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMet
  at java.base/java.lang.reflect.Method.invoke(Method.java:568)
  at fr.infologic.core.server.jsonrpc.JSONService$MethodService.invoke(JSONService.ja
```

Near-duplicates

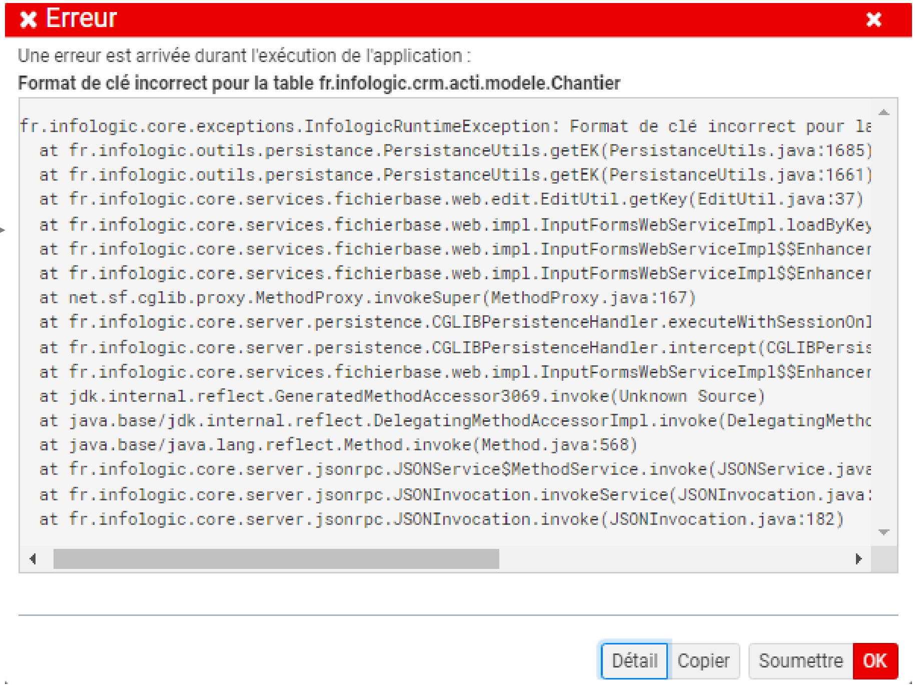
≉

But the bug is generic and related to a **web** feature



[redacted] à 11:50

[redacted] : à marquer comme traiter puisque tu as fait une correction spécifique ici, on analysera le cas général dans WEBBUG-6883.



Une erreur est arrivée durant l'exécution de l'application :

Format de clé incorrect pour la table fr.infologic.crm.acti.modele.Chantier

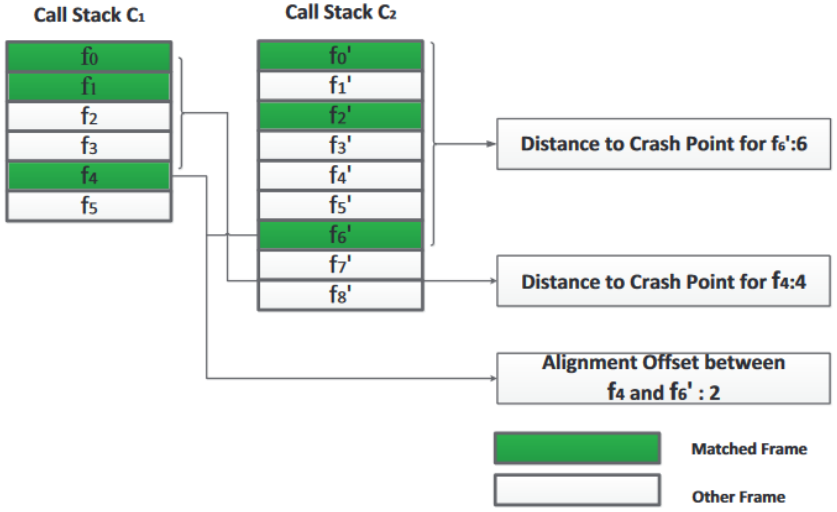
```
fr.infologic.core.exceptions.InfologicRuntimeException: Format de clé incorrect pour la
  at fr.infologic.outils.persistance.PersistanceUtils.getEK(PersistanceUtils.java:1685)
  at fr.infologic.outils.persistance.PersistanceUtils.getEK(PersistanceUtils.java:1661)
  at fr.infologic.core.services.fichierbase.web.edit.EditUtil.getKey(EditUtil.java:37)
  at fr.infologic.core.services.fichierbase.web.impl.InputFormsWebServiceImpl.loadByKey
  at fr.infologic.core.services.fichierbase.web.impl.InputFormsWebServiceImpl$$Enhancer
  at fr.infologic.core.services.fichierbase.web.impl.InputFormsWebServiceImpl$$Enhancer
  at net.sf.cglib.proxy.MethodProxy.invokeSuper(MethodProxy.java:167)
  at fr.infologic.core.server.persistence.CGLIBPersistenceHandler.executeWithSessionOnl
  at fr.infologic.core.server.persistence.CGLIBPersistenceHandler.intercept(CGLIBPersis
  at fr.infologic.core.services.fichierbase.web.impl.InputFormsWebServiceImpl$$Enhancer
  at jdk.internal.reflect.GeneratedMethodAccessor3069.invoke(Unknown Source)
  at java.base/jdk.internal.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMetho
  at java.base/java.lang.reflect.Method.invoke(Method.java:568)
  at fr.infologic.core.server.jsonrpc.JSONService$MethodService.invoke(JSONService.java
  at fr.infologic.core.server.jsonrpc.JSONInvocation.invokeService(JSONInvocation.java:
  at fr.infologic.core.server.jsonrpc.JSONInvocation.invoke(JSONInvocation.java:182)
```

*__Jiang et al.__ Igor: Crash Deduplication Through Root-Cause Clustering, In **CCS 2021**

Enhancing Duplicate Crash Report Retrieval

Similarity measures for stack trace comparison embedded in Clustering algorithms

- **Complex** similarity measures based
- **Computational** Complexity is very **costly**
- Measures embedded in **clustering** with several **issues**
- It should be handle as **Nearest Neighbours Search** problem



Call Stack C1
f0
f1
f2
f3
f4
f5
Call Stack C2
f0'
f1'
f2'
f3'
f4'
f5'
f6'
f7'
f8'
Distance to Crash Point for f6':6
Distance to Crash Point for f4:4
Alignment Offset between f4 and f6' : 2
Matched Frame
Other Frame

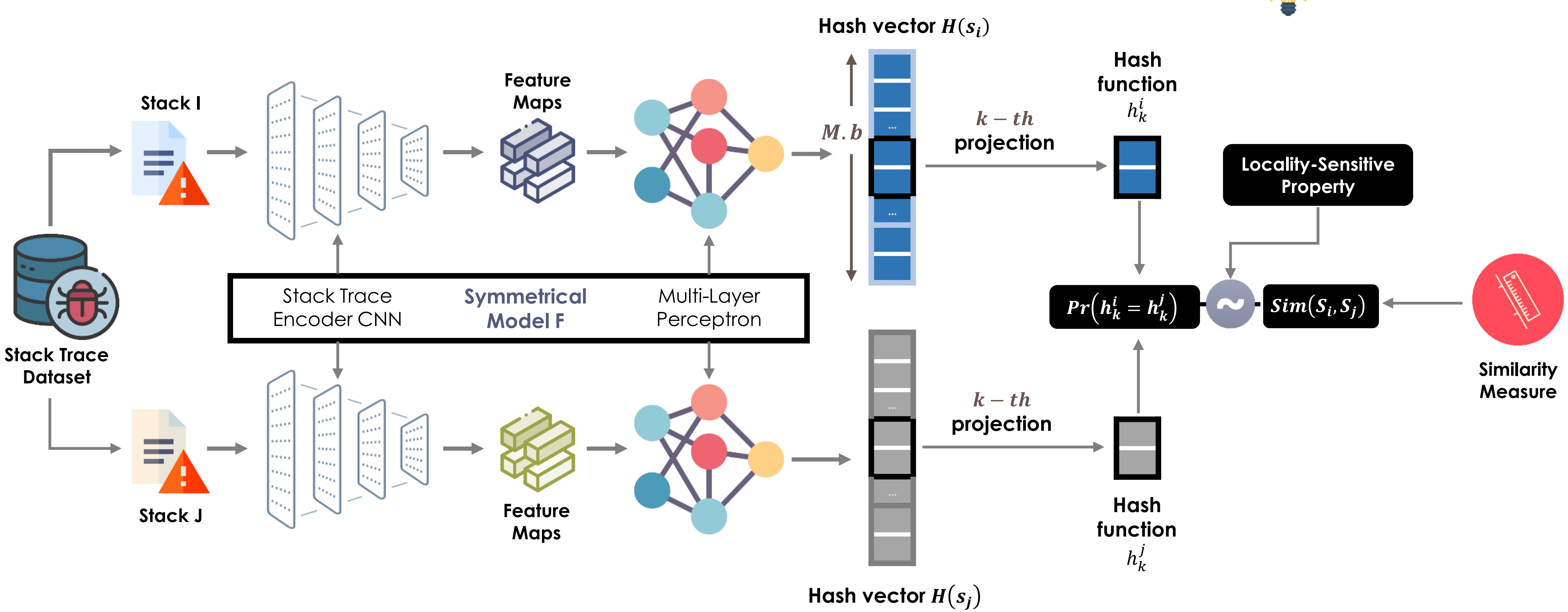
[**Dang et al.**, in **ICSE 2012**]*

*__Dang et al.__ ReBucket: A Method for Clustering Duplicate Crash Reports Based on Call Stack Similarity. In **ICSE 2012**
†**Wu et al.** CrashLocator: Locating Crashing Faults Based on Crash Stacks. In **ISSTA 2013**
‡**Moroo et al**. Reranking-based Crash Report Deduplication. **SEKE 2017**

Enhancing Duplicate Crash Report Retrieval

Contribution

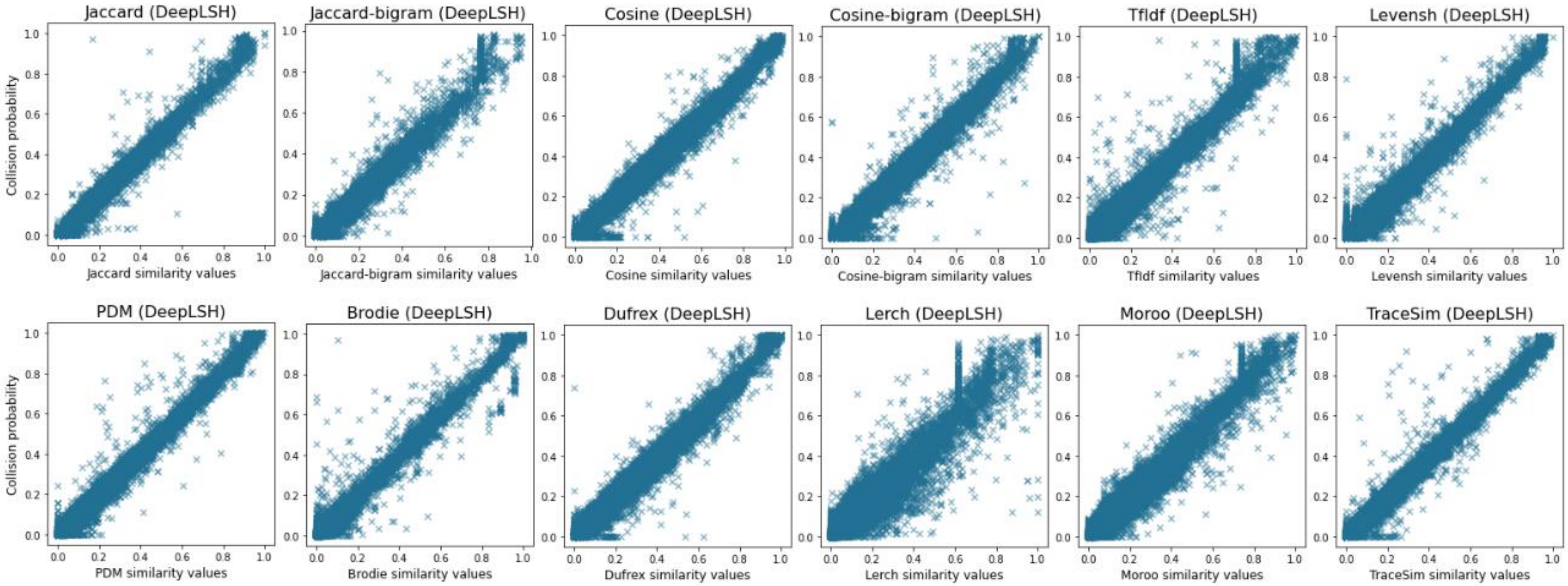
Learn a family of hash functions with a constrained hashing Siamese neural network



Enhancing Duplicate Crash Report Retrieval

Experiments

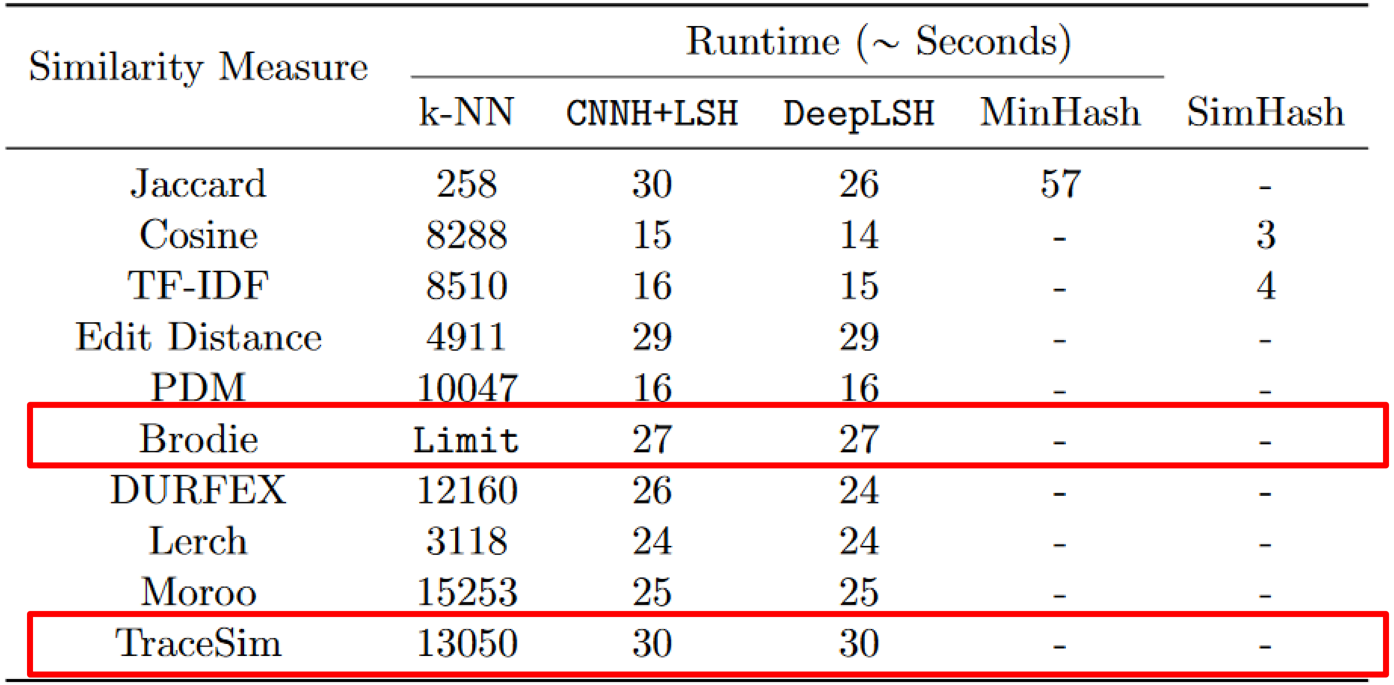
Does the model manage to converge to the LSH property?



Enhancing Duplicate Crash Report Retrieval

Experiments

Is the model fast enough compared to linear scans?



Similarity Measure	Runtime (~ Seconds)				
	k-NN	CNNH+LSH	DeepLSH	MinHash	SimHash
Jaccard	258	30	26	57	-
Cosine	8288	15	14	-	3
TF-IDF	8510	16	15	-	4
Edit Distance	4911	29	29	-	-
PDM	10047	16	16	-	-
Brodie	Limit	27	27	-	-
DURFEX	12160	26	24	-	-
Lerch	3118	24	24	-	-
Moroo	15253	25	25	-	-
TraceSim	13050	30	30	-	-

End

Thanks for your attention

End

Q/A?