QUT | Centre for Data Science

Building Strong Data Foundations for Process Mining

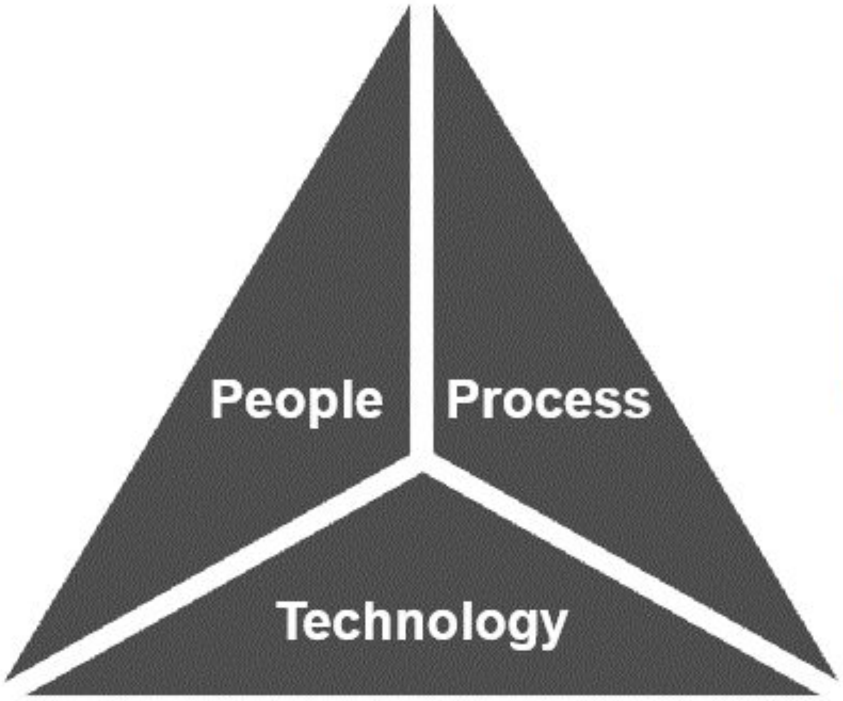
Prof Moe Thandar Wynn
Co-Director, QUT's Centre for Data Science
Leader, Process Science Discipline
Queensland University of Technology, Brisbane, Australia
m.wynn@qut.edu.au

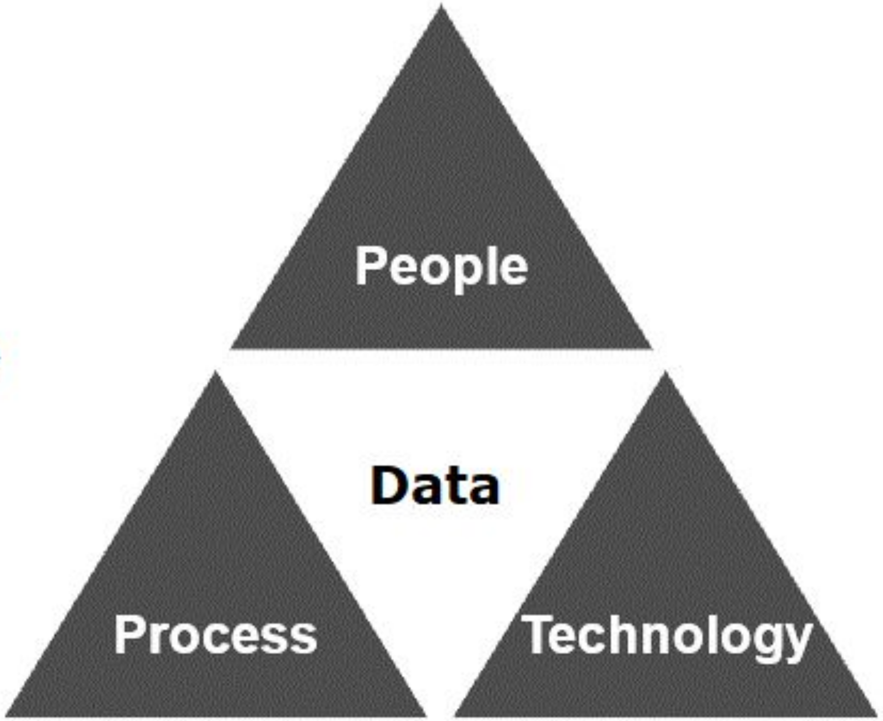
Agenda

- Data Requirements for Process Mining
  - Case-Centric Event Log
  - Object-Centric Event Log
  - Typical Data Sources
  - Event Log Formats
- Data Pre-processing
- Process-Data Quality Management

Source: Gartner, Inc., 2018

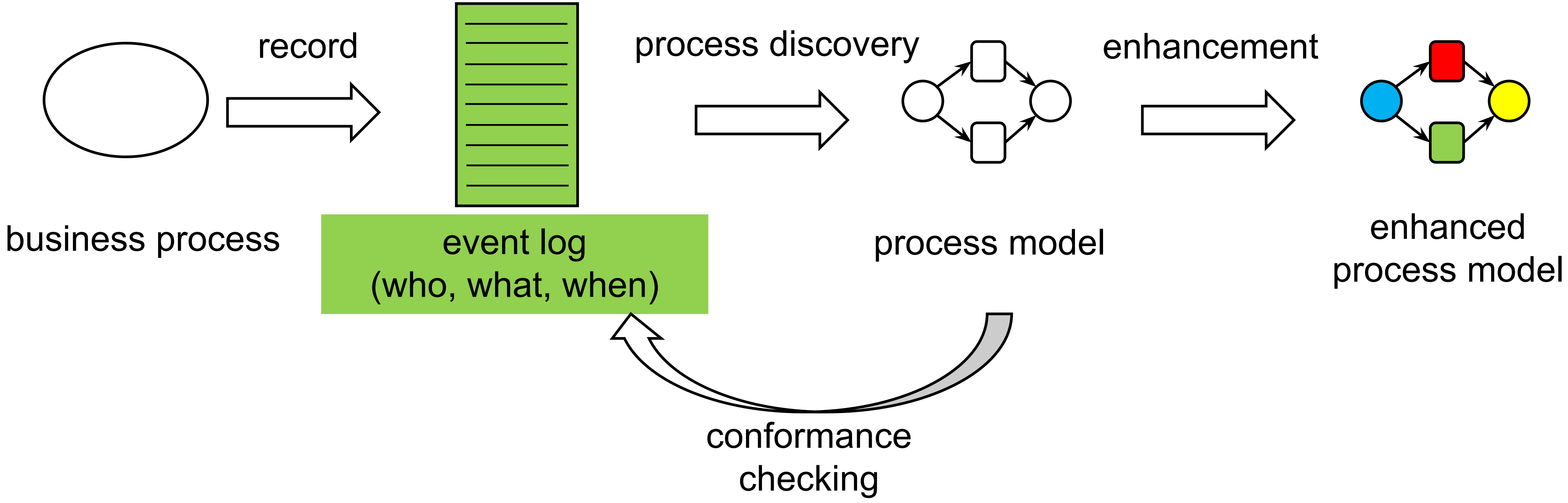
Data captures the critical interplay between people, process and techno





Source: Gartner, Inc., 2018

Data Requirements for Process Mining



QUT Centre for Data Science

Case-Centric Log: Case-id, Activity Labels, Timestamps

Case ID	Activity	Timestamp
C001	Receive emergency call	18/03/2026 14:05:00
C001	Ambulance arrive on scene	18/03/2026 14:18:00
C001	Arrive at hospital	18/03/2026 15:35:00
C001	Triage	18/03/2026 16:00:00
C002	Receive emergency call	18/03/2026 06:30:00
C002	...	...
C002	Discharge	20/03/2026 12:30:00

QUT Centre for Data Science

Object-Centric Log: Object(s) and Events

Event ID	Activity	Patient	Transport	Admission	Timestamp
e_1	Receive emergency call	P001	TR001		18/03/2026 14:05:00
e_2	Ambulance arrive on scene	P001	TR001		18/03/2026 14:18:00
e_3	Provide on-scene treatment	P001	TR001		18/03/2026 14:30:00
e_4	Initiate transport	P001	TR001		18/03/2026 14:55:00
e_5	Arrive at hospital	P001	TR001	EN001	18/03/2026 15:35:00
e_6	Triage	P001		EN001	18/03/2026 16:00:00
e_7	Admitted to ward	P001		EN001	18/03/2026 17:10:00
e_8	Discharge	P001		EN001	19/03/2026 11:20:00
e_9	Receive emergency call	P002	TR002		18/03/2026 06:30:00
e_{10}	Ambulance arrive on scene	P002	TR002		18/03/2026 07:10:00
e_{11}	Provide on-scene treatment	P002	TR002		18/03/2026 07:30:00
e_{12}	Initiate transport	P002	TR002		18/03/2026 08:15:00
e_{13}	Arrive at hospital	P002	TR002	EN002	18/03/2026 08:50:00
e_{16}	Admitted to ward	P002		EN002	18/03/2026 12:00:00
e_{17}	Discharge	P002		EN002	20/03/2026 12:30:00

XES Survey

IEEE TASK FORCE ON PROCESS MINING

Working Group: Moe Wynn, Julian Lebherz, Wil van der Aalst, Rafael Accorsi, Claudio Di Ciccio, Eric Verbeek & Lakmali Jayarathna

Data Transformation Challenges for PM:
(**12 Qs**)

Four perspectives: Academia, Professional Services, Vendors, End Users

Released: **June – July 2021**

Received: **290** responses

Q1. How much experience do you have with Process Mining?

Q2. Which area and role best describes how you have interacted with PM?

Q3. What share of effort is typically spent on data pre-processing?

Q4. Which process mining solutions have you used?

Q5. Which technologies have you used in data pre-processing for process mining?

Q6. In Which formats is your source data available in?

Q7.Which source systems have you analyzed with process mining?

Q8. How big was the largest data set you worked with in process mining?

Q9. To what extent did you encounter the following data-related challenges while undertaking PM projects?

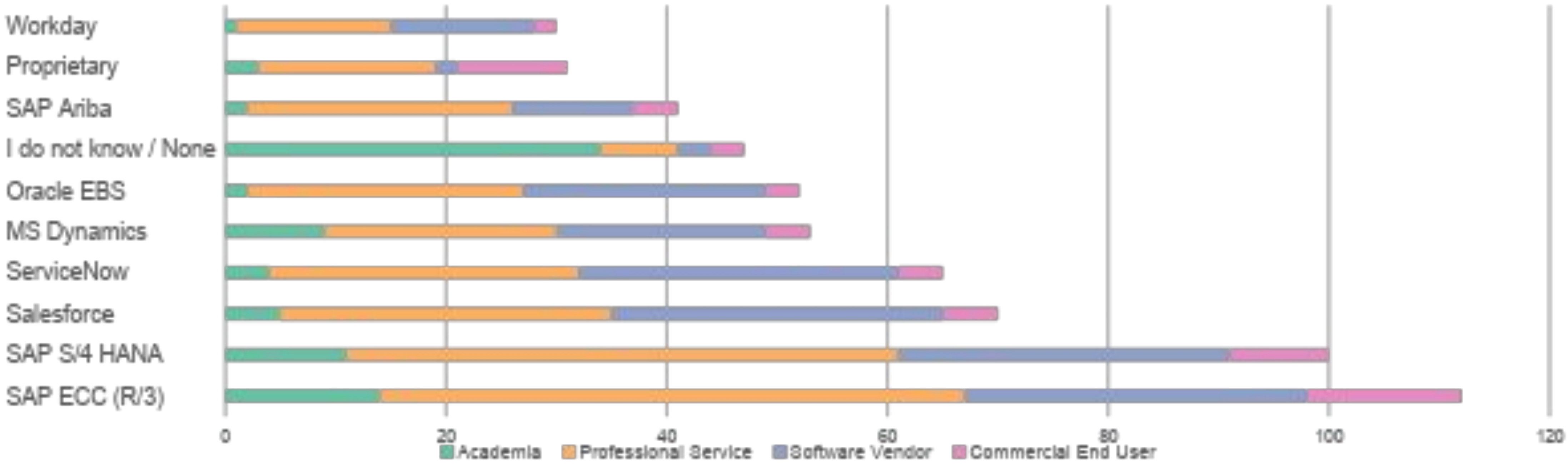
Q10. Which data-related challenges have you encountered beyond the ones listed in question #9?

Q11. There is general consensus amongst practitioners that data pre-processing tasks still consume most of the effort put into process mining initiatives. How could we speed up the data pre-processing to focus on analysis?

Q12. How could a reimagined industry-wide process mining data standard help you excel in your role?

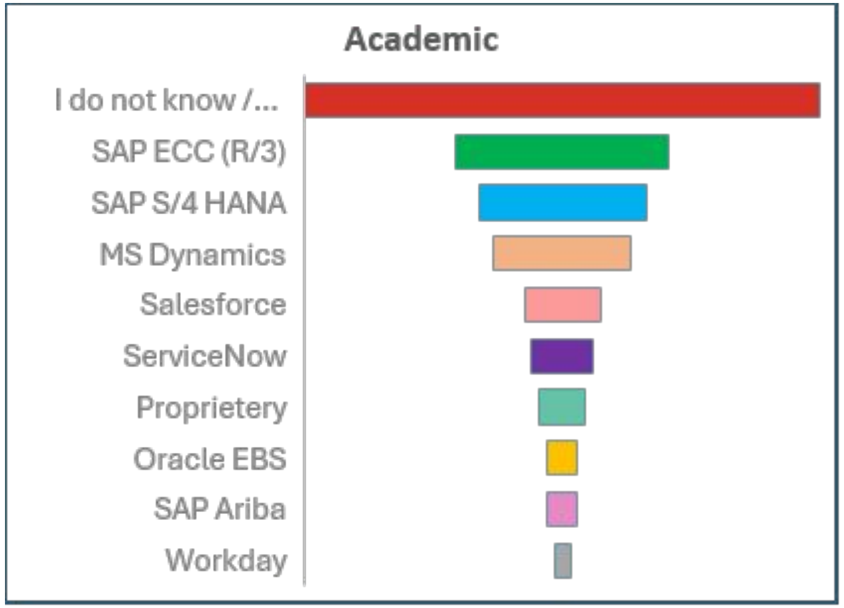
Wynn, M.T. et al. (2022). Rethinking the Input for Process Mining: Insights from the XES Survey and Workshop. In: Munoz-Gama, J., Lu, X. (eds) Process Mining Workshops. ICPM 2021. Lecture Notes in Business Information Processing, vol 433. Springer, Cham. https://doi.org/10.1007/978-3-030-98581-3_1

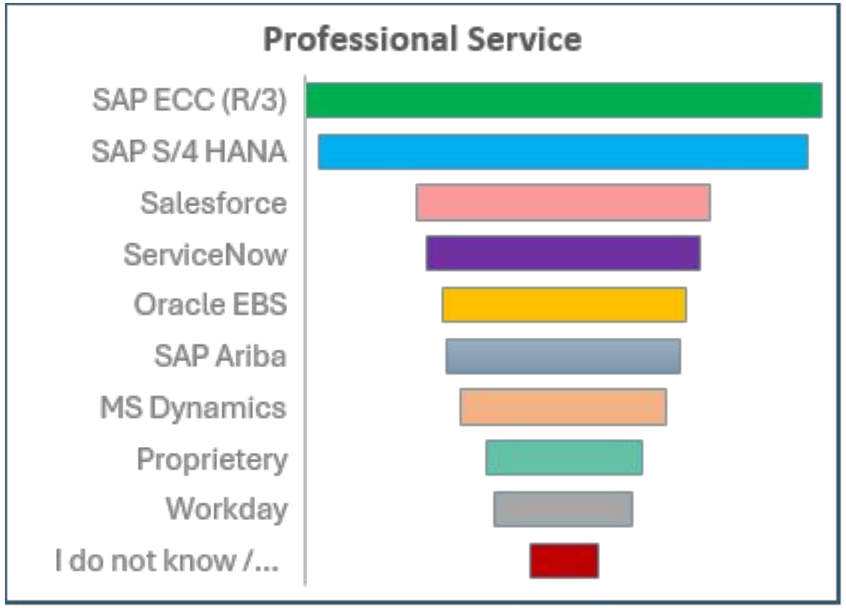
Q7 - Which source systems have you analyzed with process mining?

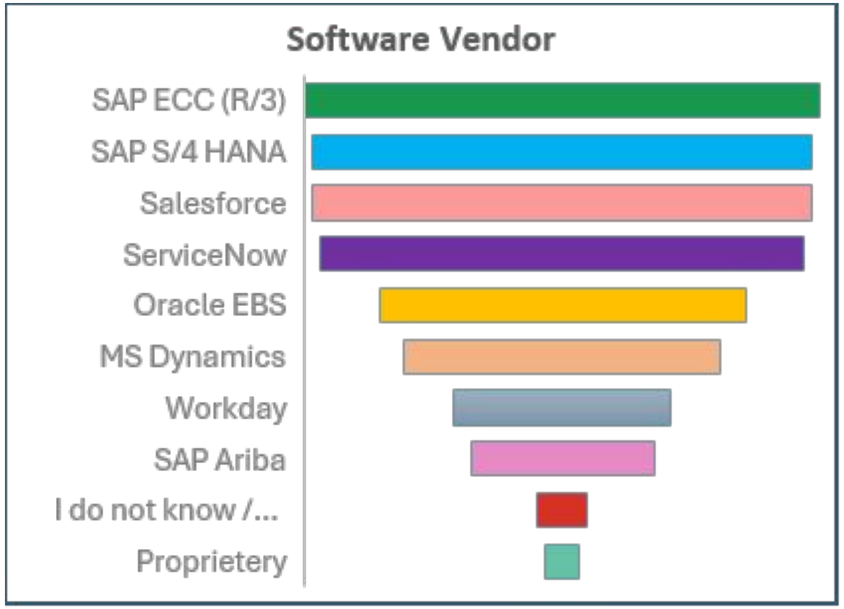


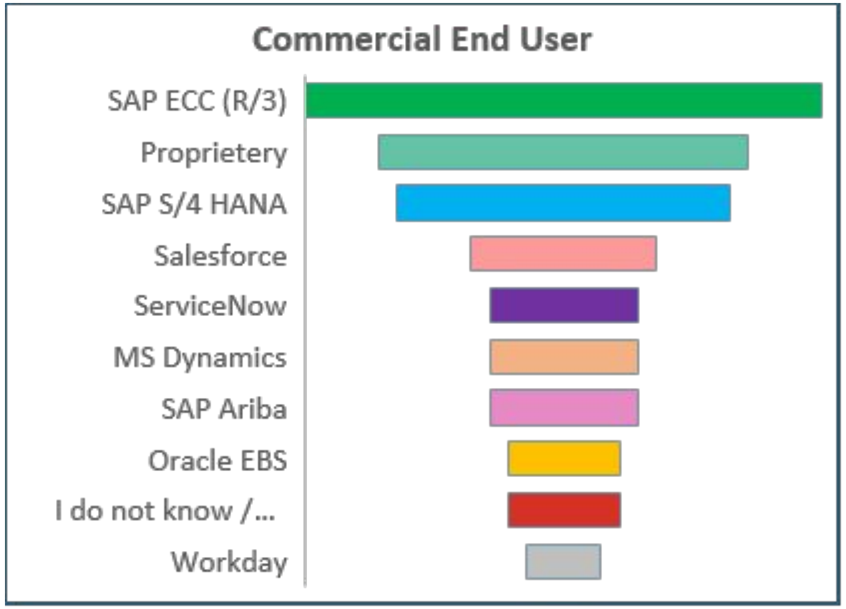
	Academia	Professional Service	Software Vendor	Commercial End User	Total
SAP ECC (R/3)	14	53	31	14	**114**
SAP S/4 HANA	11	50	30	9	**101**
Salesforce	5	30	30	5	**71**
ServiceNow	4	28	29	4	**66**
MS Dynamics	9	21	19	4	**54**
Oracle EBS	2	25	22	3	**53**
I do not know / None	34	7	3	3	**48**
SAP Ariba	2	24	11	4	**42**
Proprietary	3	16	2	10	**31**
Workday	1	14	13	2	**30**

Q7 - **Which source systems have you analyzed with process mining?** Top 10 Source systems based on Roles

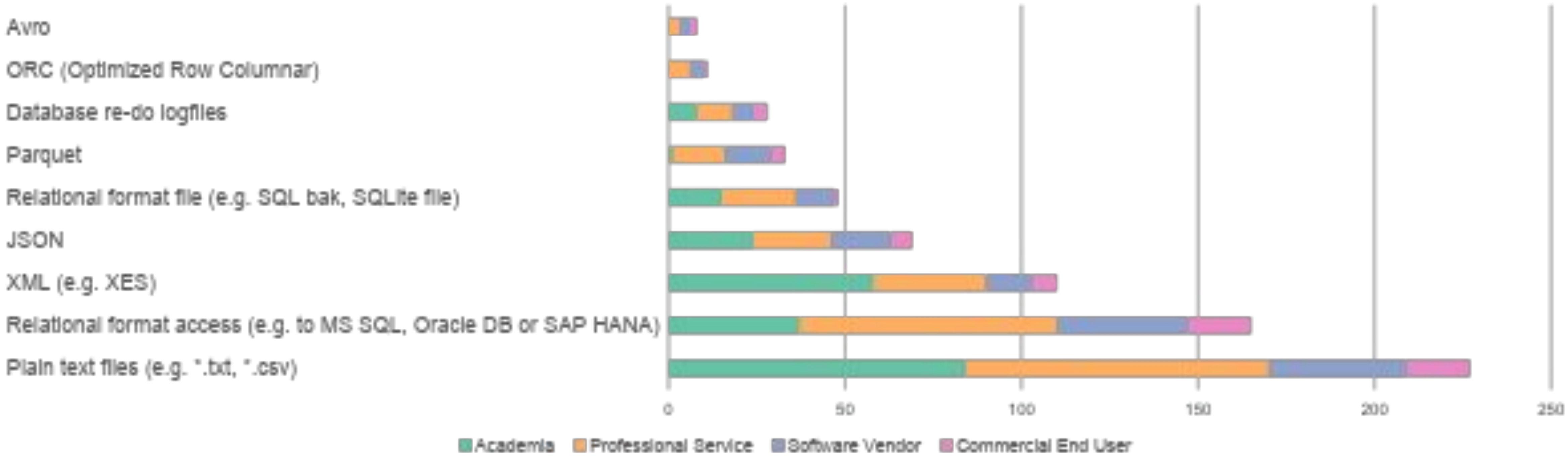






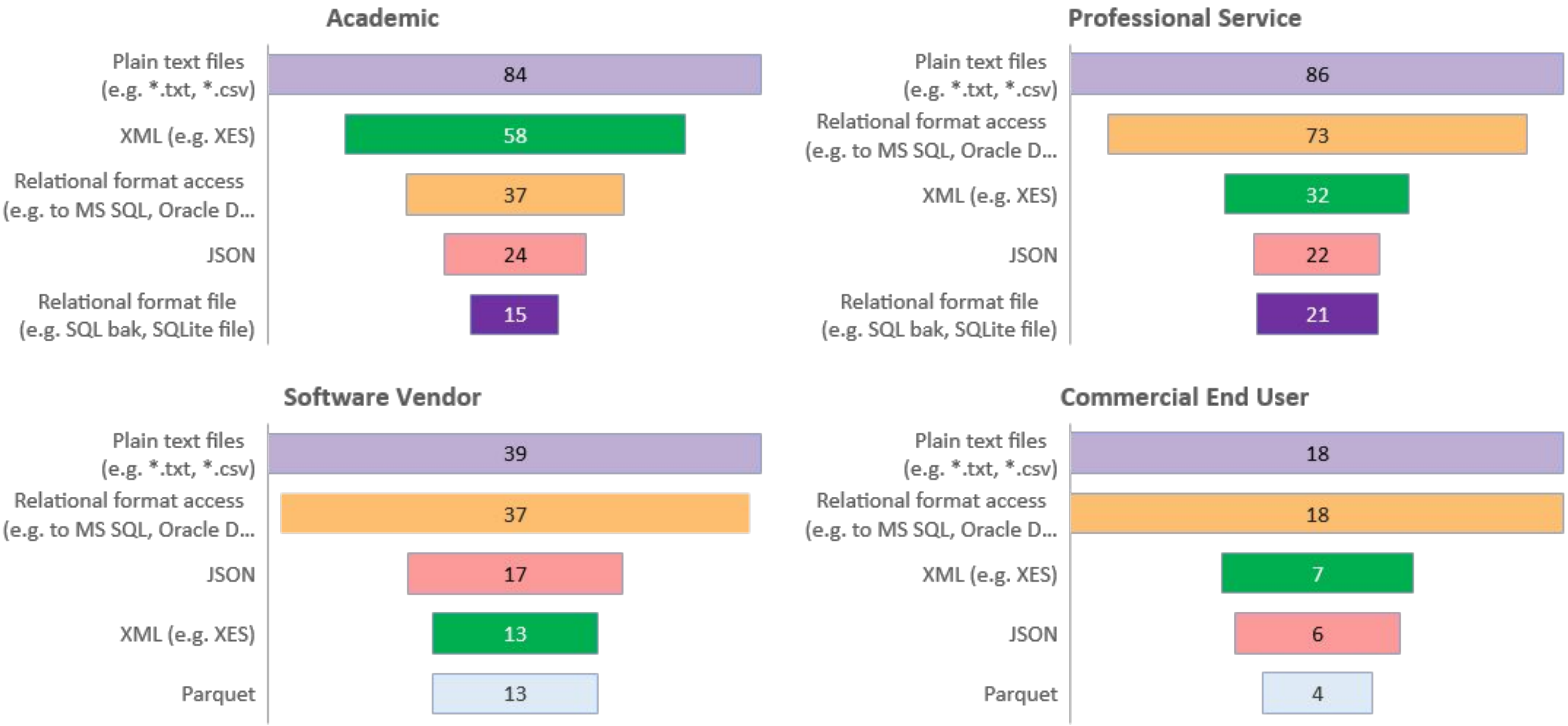


Q6 - **In which formats is your source data available**? – Top 10 responses



	Academia	Professional Service	Software Vendor	Commercial End User	Total
Plain text files (e.g. *.txt, *.csv)	84	86	39	18	**229**
Relational format access (e.g., to MS SQL, Oracle DB or SAP HANA)	37	73	37	18	**168**
XML (e.g. XES)	58	32	13	7	**112**
JSON	24	22	17	6	**70**
Relational format file (e.g. SQL bak, SQLite file)	15	21	11	1	**48**
Parquet	1	15	13	4	**34**
Database re-do logfiles	8	10	6	4	**29**
ORC (Optimized Row Columnar)	0	6	4	1	**11**
Avro	0	3	3	2	**9**

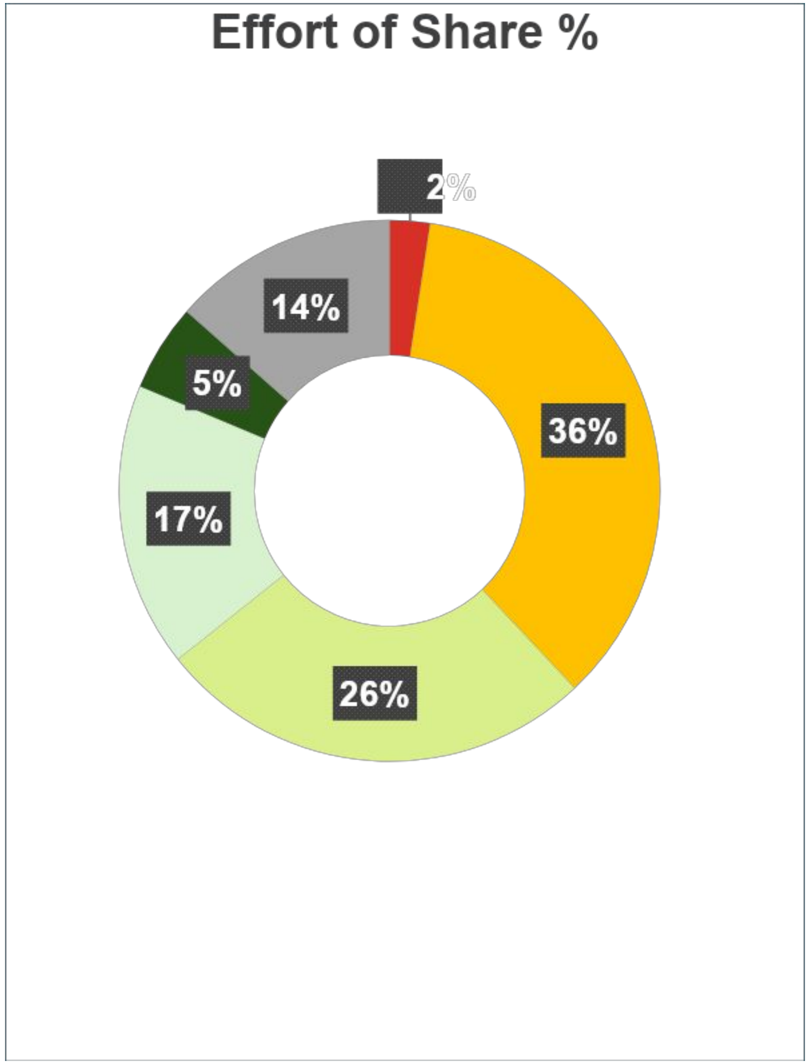
Q6 - In which formats is your source data available? (check all that applies).

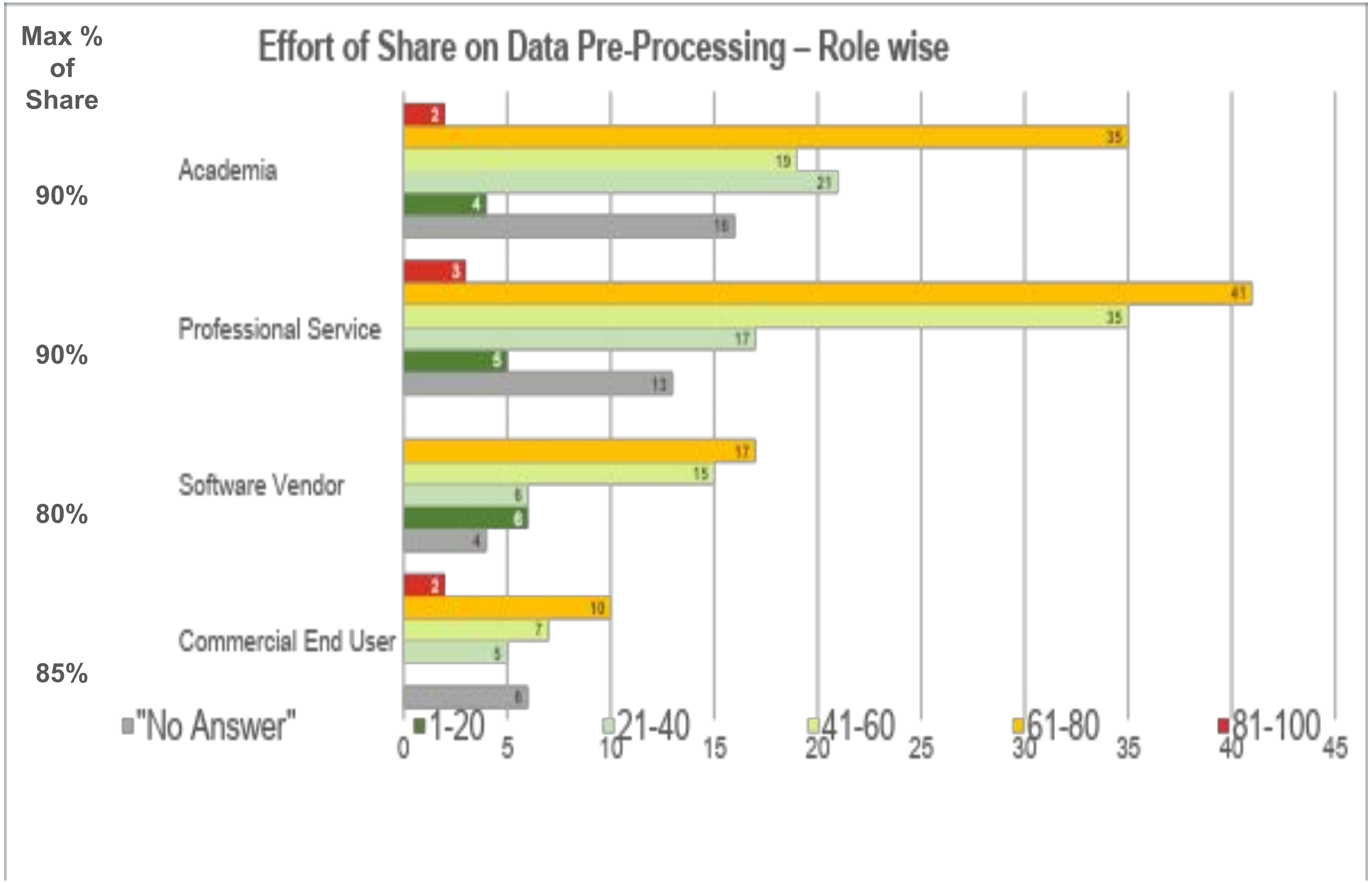


Centre for Data Science

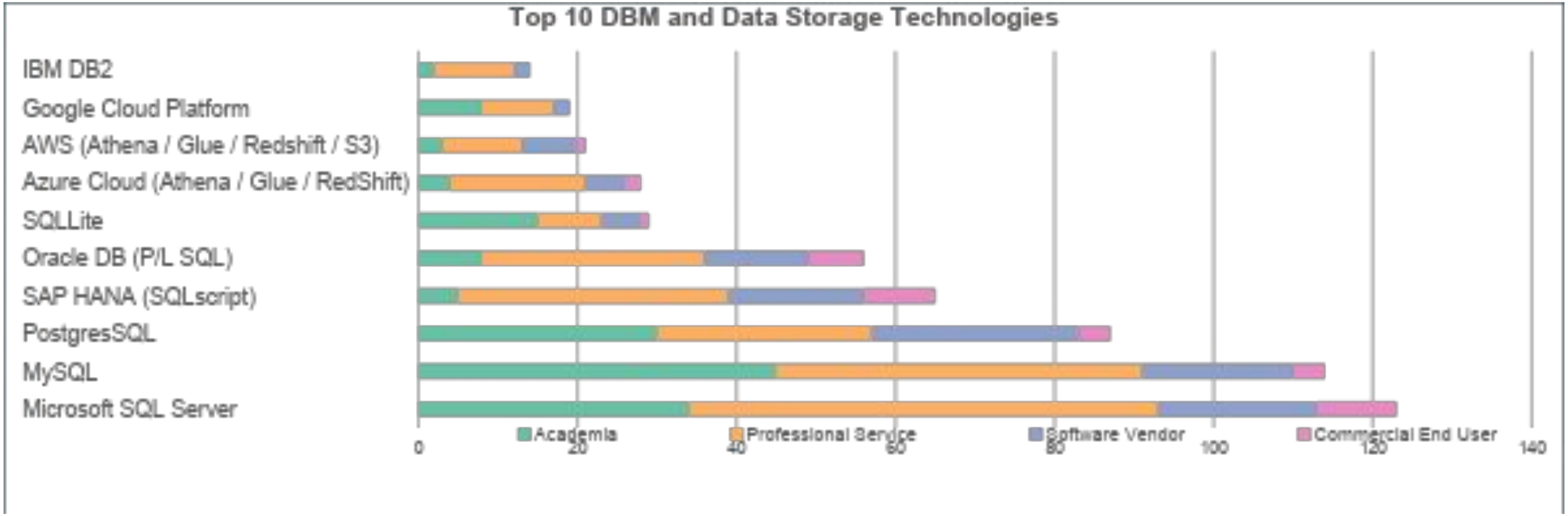
Data Pre-processing

Q3 – What share of effort is typically spent on data pre-processing (all tasks between extraction and analysis)?



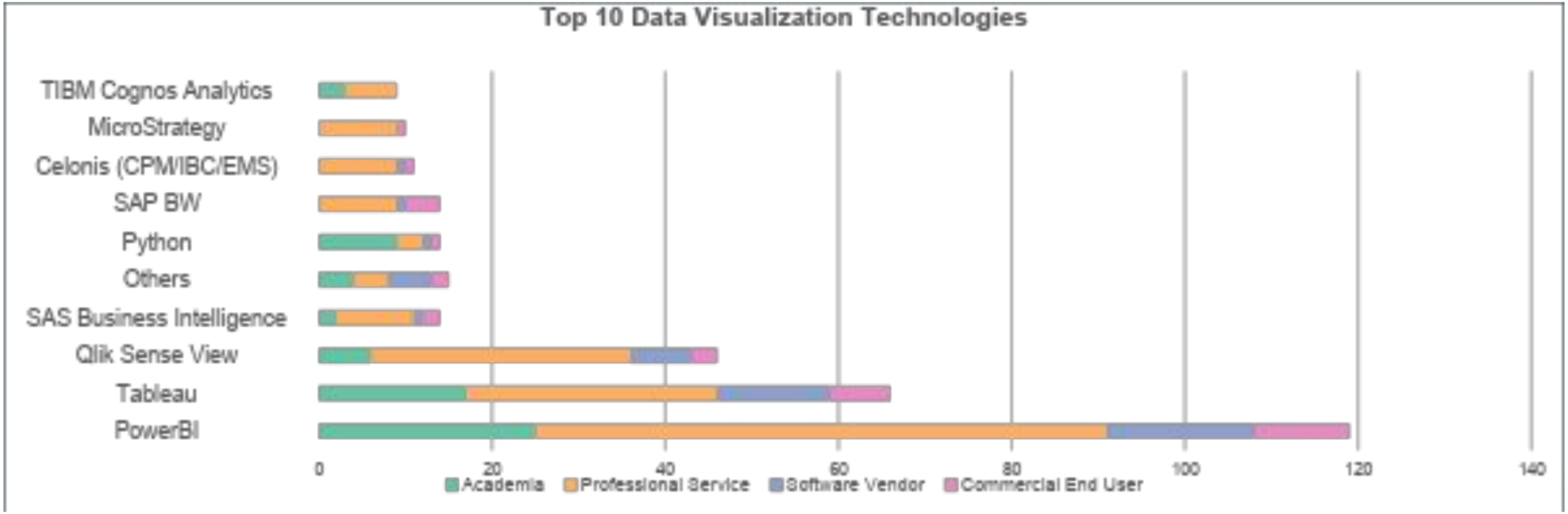


Q5 - Which technologies have you used in data preprocessing for process mining?



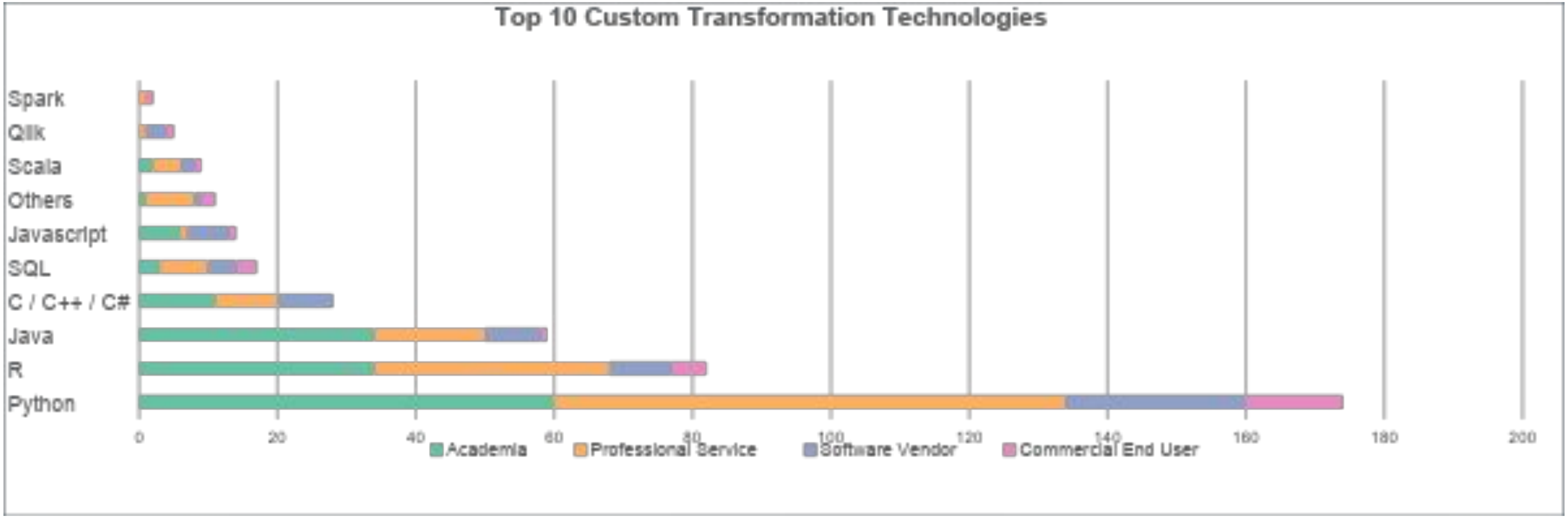
	Academia	Professional Service	Software Vendor	Commercial End User	Total
Microsoft SQL Server	34	59	20	10	**125**
MySQL	45	46	19	4	**116**
PostgresSQL	30	27	26	4	**89**
SAP HANA (SQLscript)	5	34	17	9	**68**
Oracle DB (P/L SQL)	8	28	13	7	**57**
SQLLite	15	8	5	1	**30**
Azure Cloud (Athena / Glue / RedShift)	4	17	5	2	**29**
AWS (Athena / Glue / Redshift / S3)	3	10	7	1	**21**
Google Cloud Platform	8	9	2	0	**19**
IBM DB2	2	10	2	0	**14**

Q5 - Which technologies have you used in data preprocessing for process mining?



	Academia	Professional Service	Software Vendor	Commercial End User	Total
PowerBI	25	66	17	11	**122**
Tableau	17	29	13	7	**68**
Qlik Sense View	6	30	7	3	**47**
SAS Business Intelligence	2	9	1	2	**15**
Others	4	4	5	2	**15**
Python	9	3	1	1	**14**
SAP BW	0	9	1	4	**14**
Celonis (CPM/IBC/EMS)	0	9	1	1	**11**
MicroStrategy	0	9	0	1	**10**
TIBM Cognos Analytics	3	6	0	0	**9**

Q5 - Which technologies have you used in data preprocessing for process mining?



	Academia	Professional Service	Software Vendor	Commercial End User	Total
Python	60	74	26	14	**177**
R	34	34	9	5	**83**
Java	34	16	8	1	**60**
C / C++ / C#	11	9	8	0	**28**
SQL	3	7	4	3	**17**
Javascript	6	1	6	1	**14**
Others	1	7	1	2	**11**
Scala	2	4	2	1	**9**
Qlik	0	1	3	1	**5**
Spark	0	1	0	1	**2**

Process Data Quality Management

Process Mining relies on event data as its source of truth …

Challenge C1 in the PM Manifesto: "C1: Finding, Merging, and Cleaning Event Data"

Garbage In □ Garbage Out

Maturity Levels of Event Logs

Level	Characterization	Examples
★★★★★	Highest level: the event log is of excellent quality (i.e., trustworthy and complete) and events are well-defined. Events are recorded in an automatic, systematic, reliable, and safe manner. Privacy and security considerations are addressed adequately. Moreover, the events recorded (and all of their attributes) have clear semantics. This implies the existence of one or more ontologies. Events and their attributes point to this ontology.	Semantically annotated logs of BPM systems.
★★★★	Events are recorded automatically and in a systematic and reliable manner, i.e., logs are trustworthy and complete. Unlike the systems operating at level ★★★, notions such as process instance (case) and activity are supported in an explicit manner.	Events logs of traditional BPM/ workflow systems.
★★★	Events are recorded automatically, but no systematic approach is followed to record events. However, unlike logs at level ★★, there is some level of guarantee that the events recorded match reality (i.e., the event log is trustworthy but not necessarily complete). Consider, for example, the events recorded by an ERP system. Although events need to be extracted from a variety of tables, the information can be assumed to be correct (e.g., it is safe to assume that a payment recorded by the ERP actually exists and vice versa).	Tables in ERP systems, event logs of CRM systems, transaction logs of messaging systems, event logs of high-tech systems, etc.
★★	Events are recorded automatically, i.e., as a by-product of some information system. Coverage varies, i.e., no systematic approach is followed to decide which events are recorded. Moreover, it is possible to bypass the information system. Hence, events may be missing or not recorded properly.	Event logs of document and product management systems, error logs of embedded systems, worksheets of service engineers, etc.
★	Lowest level: event logs are of poor quality. Recorded events may not correspond to reality and events may be missing. Event logs for which events are recorded by hand typically have such characteristics.	Trails left in paper documents routed through the organization ("yellow notes"), paper-based medical records, etc.

Table 1: Maturity levels for event logs.

Challenges when extracting event logs

Correlation: Events in an event log are grouped per case. This simple requirement can be quite challenging as it requires event correlation, i.e., events need to be related to each other.

Timestamps: Events need to be ordered per case. Typical problems: only dates, different clocks, delayed logging.

Snapshots: Cases may have a lifetime extending beyond the recorded period, e.g., a case was started before the beginning of the event log.

Scoping: How to decide which "tables" to incorporate?

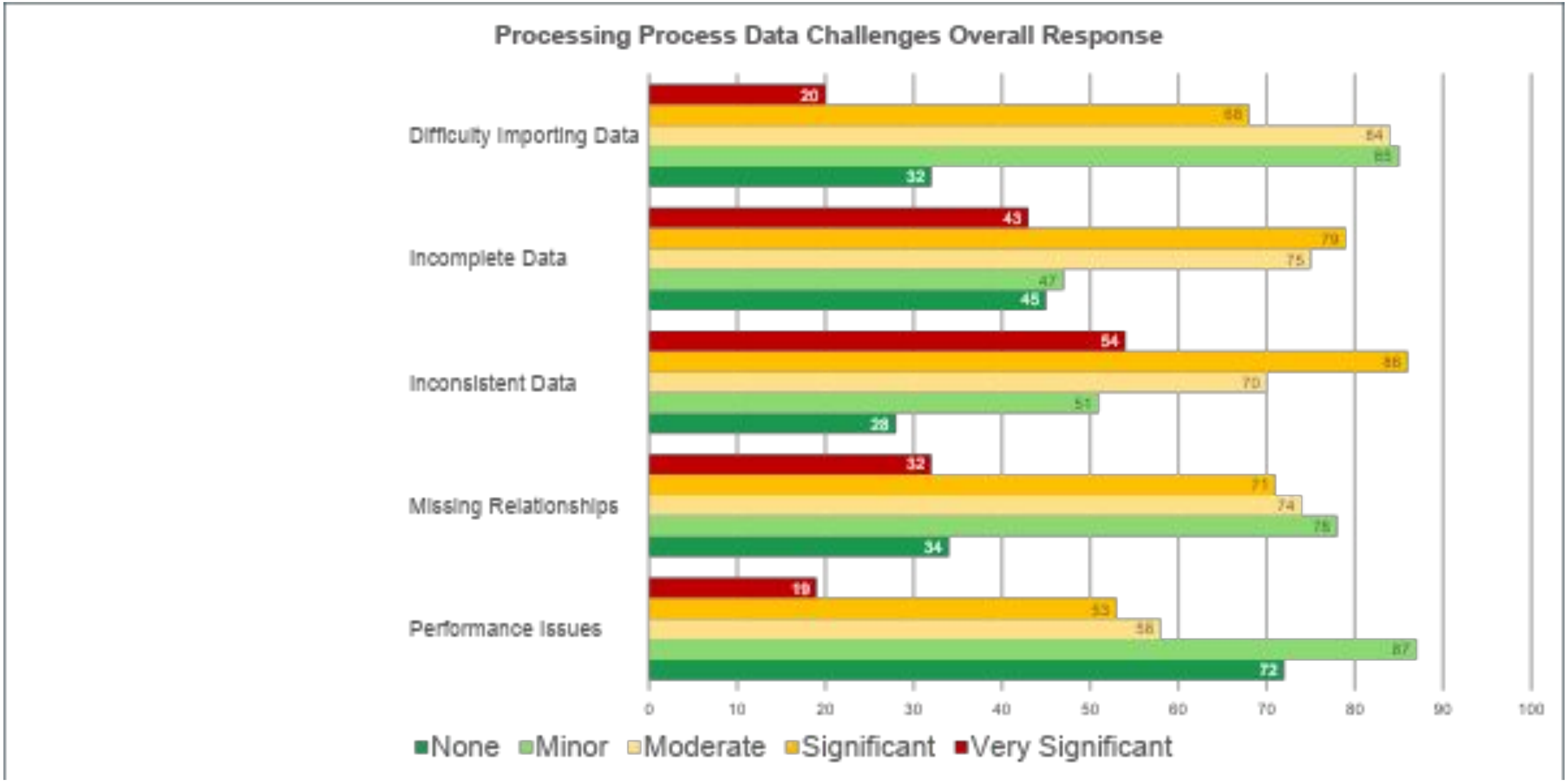
Granularity: the events in the event log are at a different level of granularity than the activities relevant for end users.

Data quality matrix

	case	event	belongs to	c attribute	position	activity name	timestamp	resource	e attribute
missing data	In reality a case has been executed but it has not been recorded in the log	Events are missing within the trace although they occurred in reality.	Association between events and cases is lost (correlation problem)	Case attribute was not recorded.	Ordering of events in the trace is lost.	Activity names of events are missing.	Timestamps of events are missing.	Resources that executed an activity have not been recorded.	Event attribute was not recorded.
incorrect data	Some cases in the log belong to a different process.	Events that were not actually executed for some cases are logged	Association between events and cases are logged incorrectly.	Values correspondin g to case attributes are logged incorrectly.	Order is mixed up.	Wrong activity names are recorded.	Incorrect timestamps.	Incorrect resource assigned to event.	Attributes of events are recorded incorrectly.
imprecise data			Difficult to correlate events to specific cases (too coarse).	Provided value is too coarse, e.g., city but no address.	For example concurrent events may have become been totally ordered.	Activity names are too coarse.	Days rather than minutes or seconds. Hence, precise order cannot be derived.	Just role or department is recorded.	Provided value is too coarse.
irrelevant data	Irrelevant cases are included and cannot be removed easily.	Events may be irrelevant and difficult to remove							

Bose, R.P.J.C.; Mans, R.S.; van der Aalst, W.M.P., "Wanna improve process mining results?," *Computational Intelligence and Data Mining (CIDM 2013)* , doi: 10.1109/CIDM.2013.6597227

Q9 -To what extent did you encounter the following data related challenges while undertaking PM projects?



Process Data Quality Management

- How do we **recognise** quality issues in event data?
- How do we **assess** the quality?
- How do we **improve** the quality?
  - **Patterns-based DQ detection and repair**
- How can we **quantify** its impact on process mining analyses?
- How do we keep **track** of these improvements?
  - **Quality-informed Process Mining**
- How do we **prevent** these issues from recurring?
  - **Root cause detection**

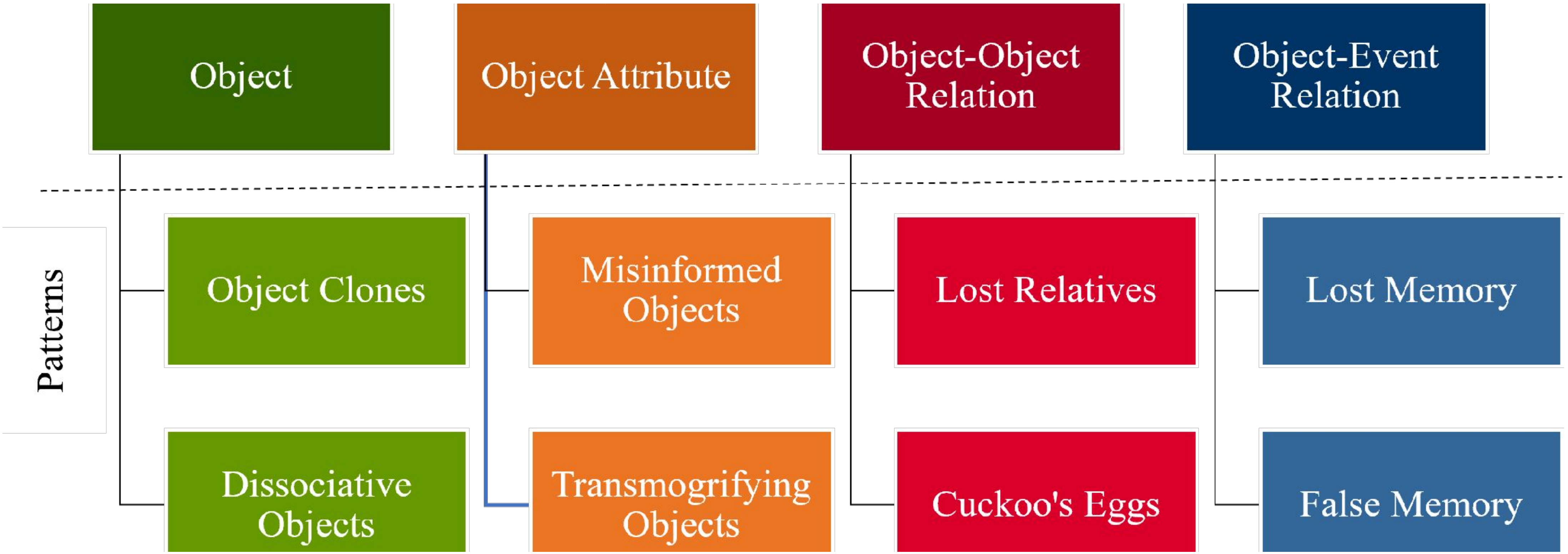
https://www.processdataquality.com/

Event Log Imperfection Patterns (case-centric)

- Form-based event capture
- Inadvertent time travel
- Scattered event
- Unanchored event
- Collateral events
- Scattered case
- Polluted label
- Distorted label
- Synonymous labels
- Homonymous labels
- Elusive case

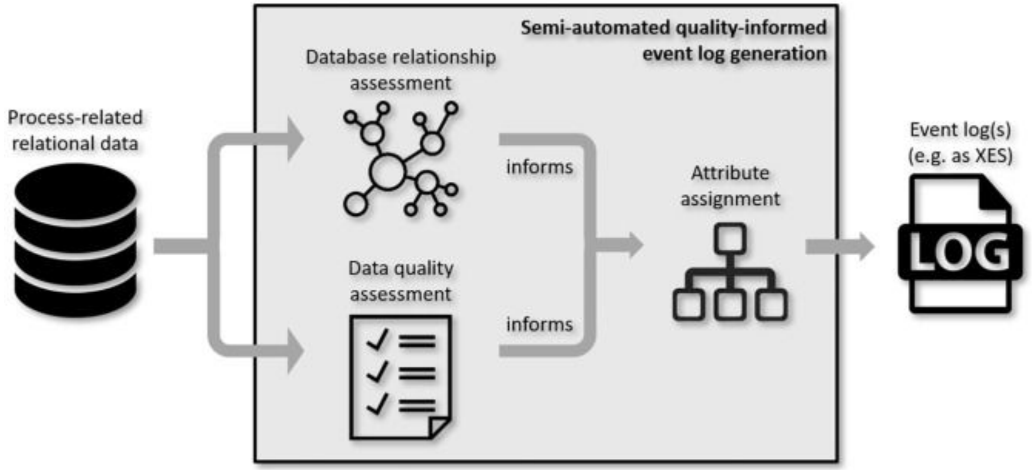
Suriadi, S., Andrews, R., ter Hofstede, A.H. and Wynn, M.T., 2017. Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. Information Systems, 64, pp.132-150.

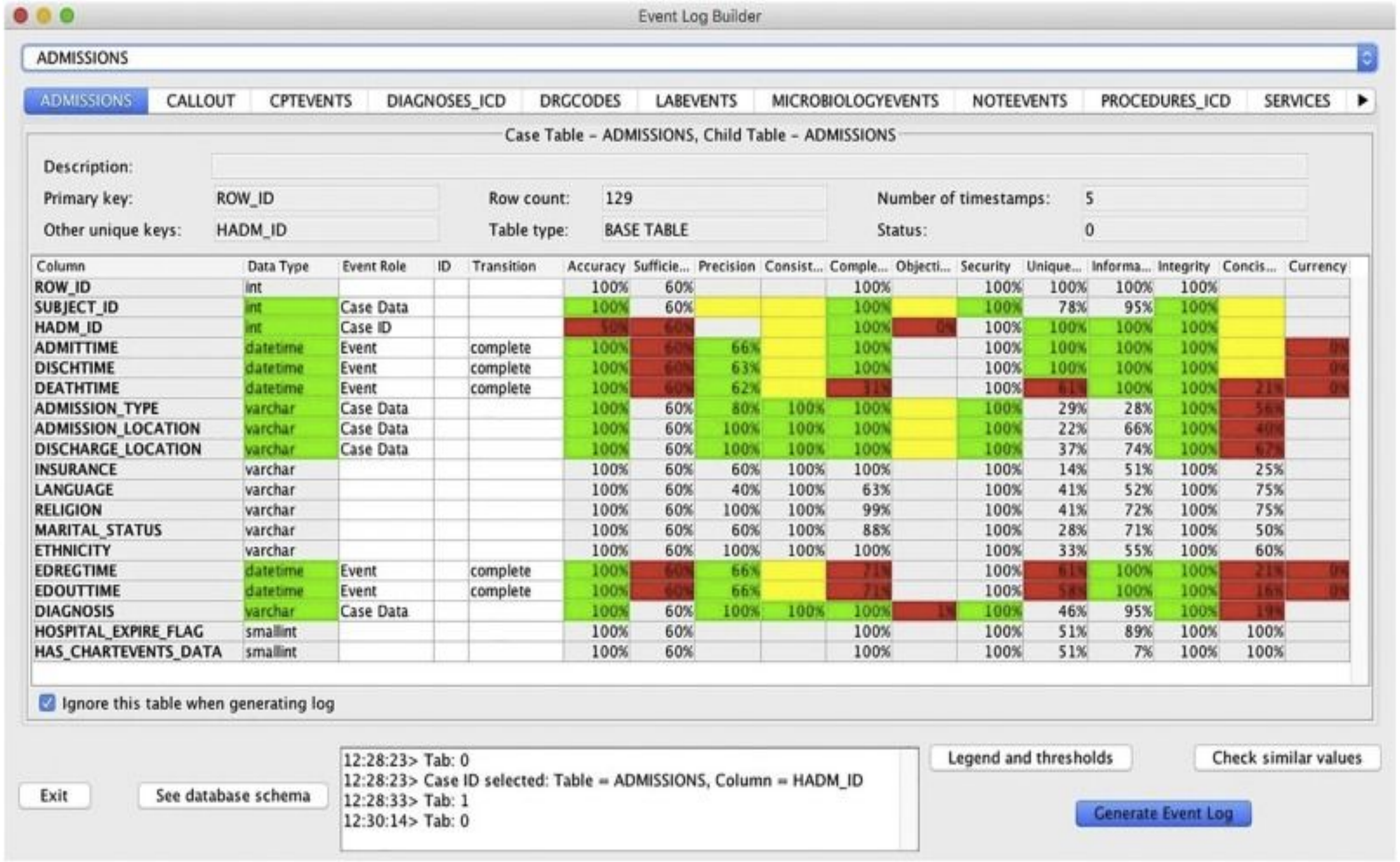
Object-Centric Event Data Imperfection Patterns



Sadeghianasl, S., Wynn, M., Andrews, R, van der Aalst, W.M.P., Ko, J. Object-Centric Event-Data Imperfection Patterns, Information Systems, Volume 141, 2026, 102766, ISSN 0306-4379, https://doi.org/10.1016/j.is.2026.102766.

Quality-Informed Generation of an Event Log

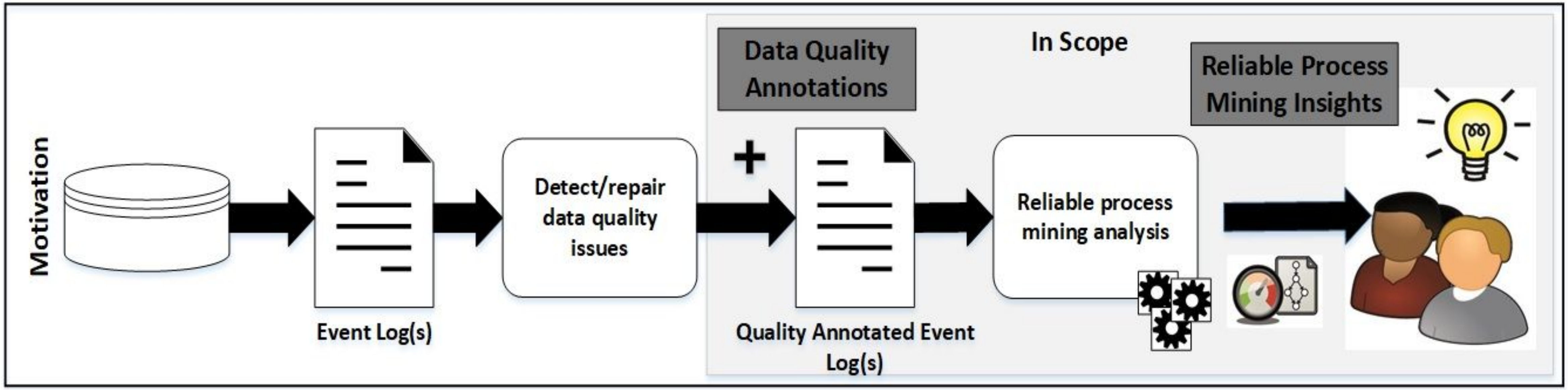




Andrews, R., van Dun, C. Wynn, M, Wratsch, W., Röglinger, M., ter Hofstede, A.H.M., Quality-informed semi-automated event log generation for process mining, Decision Support Systems, Volume 132, 2020.

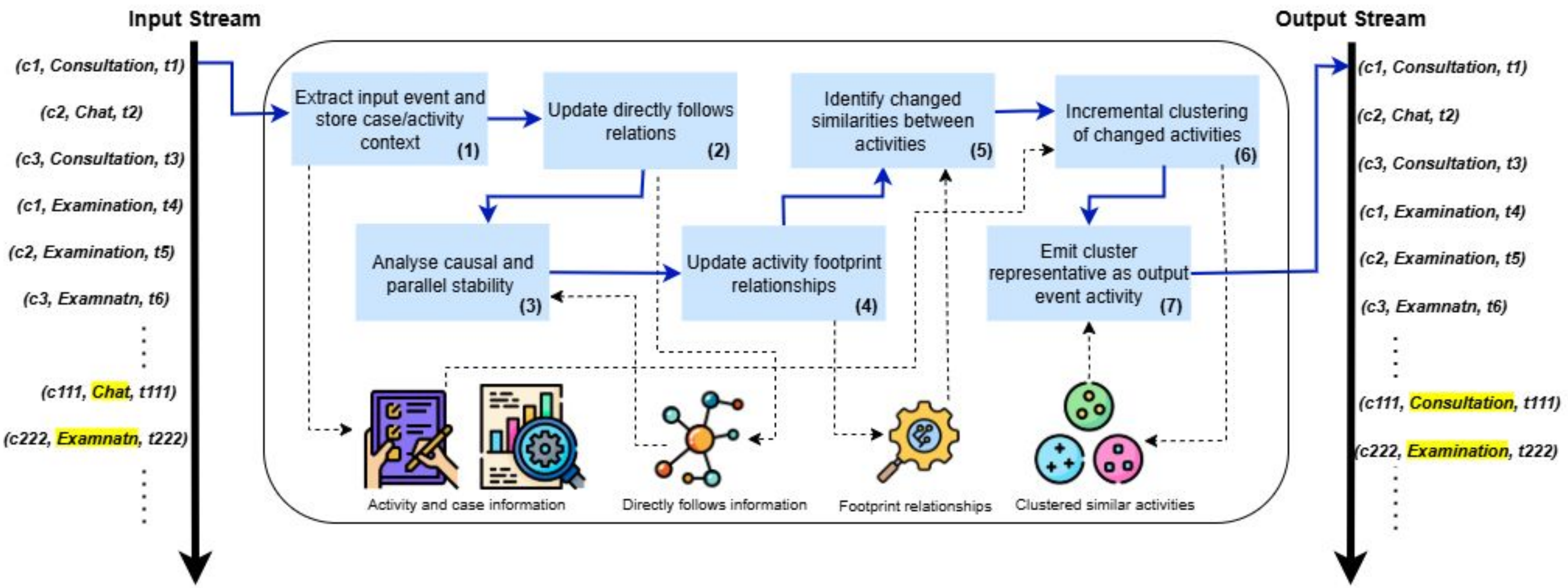
QUT Centre for Data Science

Quality-informed Process Mining



Goel, K., Leemans, S.J.J., Martin, N., and Wynn, M. . 2022. Quality-Informed Process Mining: A Case for Standardised Data Quality Annotations. ACM Trans. Knowl. Discov. Data 16, 5, Article 97 (October 2022), 47 pages. https://doi.org/10.1145/3511707

SwiftMend: Detect & Repair Labels in Process Event Streams



Kalukapuge, S., ter Hofstede, A. H., & Wynn, M. T. : SwiftMend: An Approach to Detect and Repair Activity Label Quality Issues in Process Event Streams, 2024 30th International Conference on Cooperative Information Systems (CoopIS), Porto, Portugal, 2024, pp. 131-149.

QUT Centre for Data Science

TERGEO: A Framework for Data Quality Root Cause Analysis in Healthcare

STEP 1 – Analyse Healthcare Organization's Needs

ACTIVITIES

1. Articulate business question
2. Identify measures of successful delivery

STEP 2 – Determine Scope of the Digital Health Analysis

ACTIVITIES

1. Identify the nature of the analysis
2. Identify how to best communicate analysis output
3. Determine which technological systems are relevant for the analysis
4. Determine the data required for the analysis
5. Determine the stakeholders implicated by the analysis
6. Determine the regulations surrounding the use of the data

STEP 3 – Extract and Assess Digital Health Data

ACTIVITIES

PHASE 1: Extraction of digital health data

1. Develop a conceptual model
2. Ensure relevant data is extracted

PHASE 2: Assessing digital health data quality

1. Identify and quantify digital health data quality
2. Identify data quality pattern

STEP 4 – Analyse Root Causes of Digital Health Data Quality Issues

ACTIVITIES

1. Identify issues emanating from the social, material, and personal worlds
2. Identify the severity of the root causes of the data quality issues

STEP 5 – Perform Feasibility Assessment

ACTIVITIES

1. Identify appropriate mitigation strategy through evaluating cost/complexity of change
2. Formal closure on mitigation strategy

STEP 6 – Implement prevention or mitigation strategies

ACTIVITIES

1. Execute mitigation strategy
2. Logging the mitigation strategy

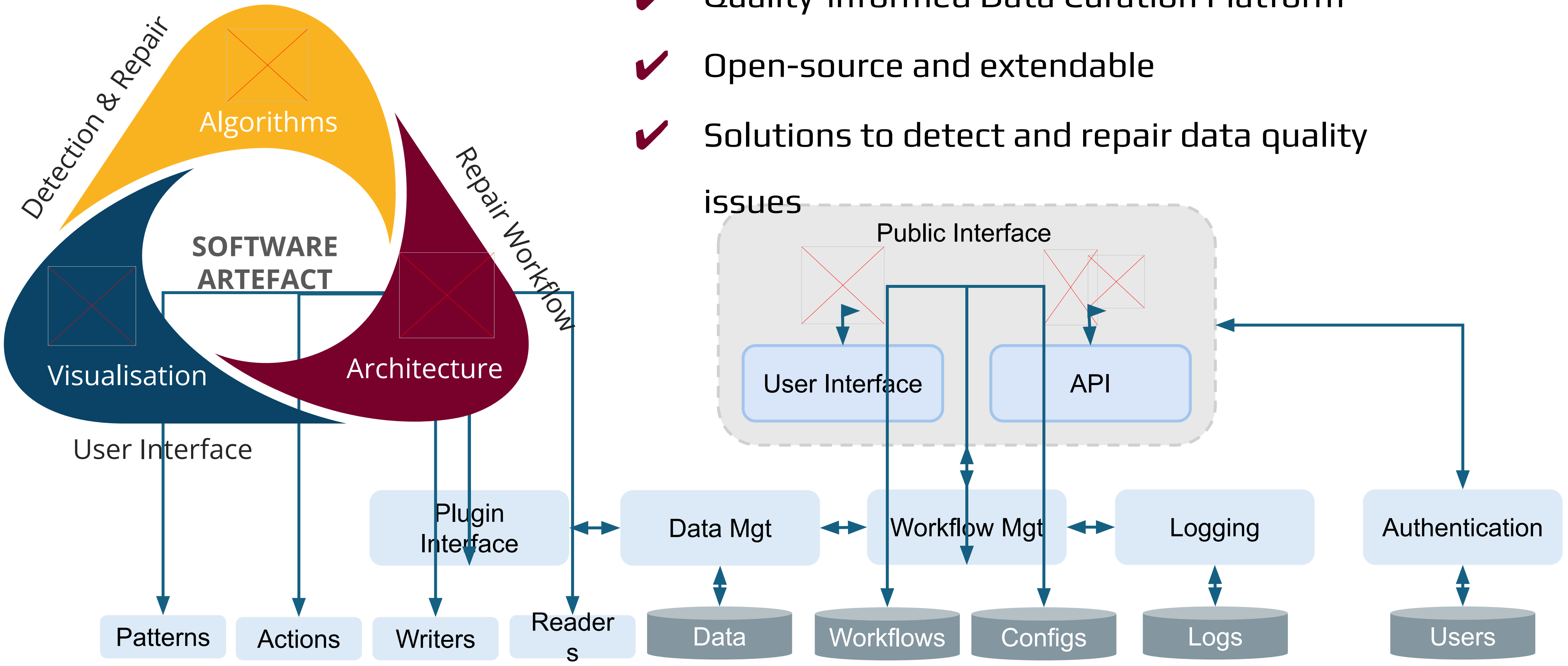
STEP 7 – Perform ongoing monitoring

ACTIVITIES

1. Determine the evaluation metrics
2. Identify how the evaluation is performed?

Praeclarus: Open-source Process-Data Quality Framework

- Quality-informed Data Curation Platform
- Open-source and extendable
- Solutions to detect and repair data quality issues



Centre for Data Science

Thank YOU! Questions?

Prof Moe Thandar Wynn | m.wynn@qut.edu.au

Acknowledgement: Many thanks to my colleagues from QUT's Process Science Group and the international BPM community for their contributions to this work being presented today.