Voorbereiden van de Data

Ngi

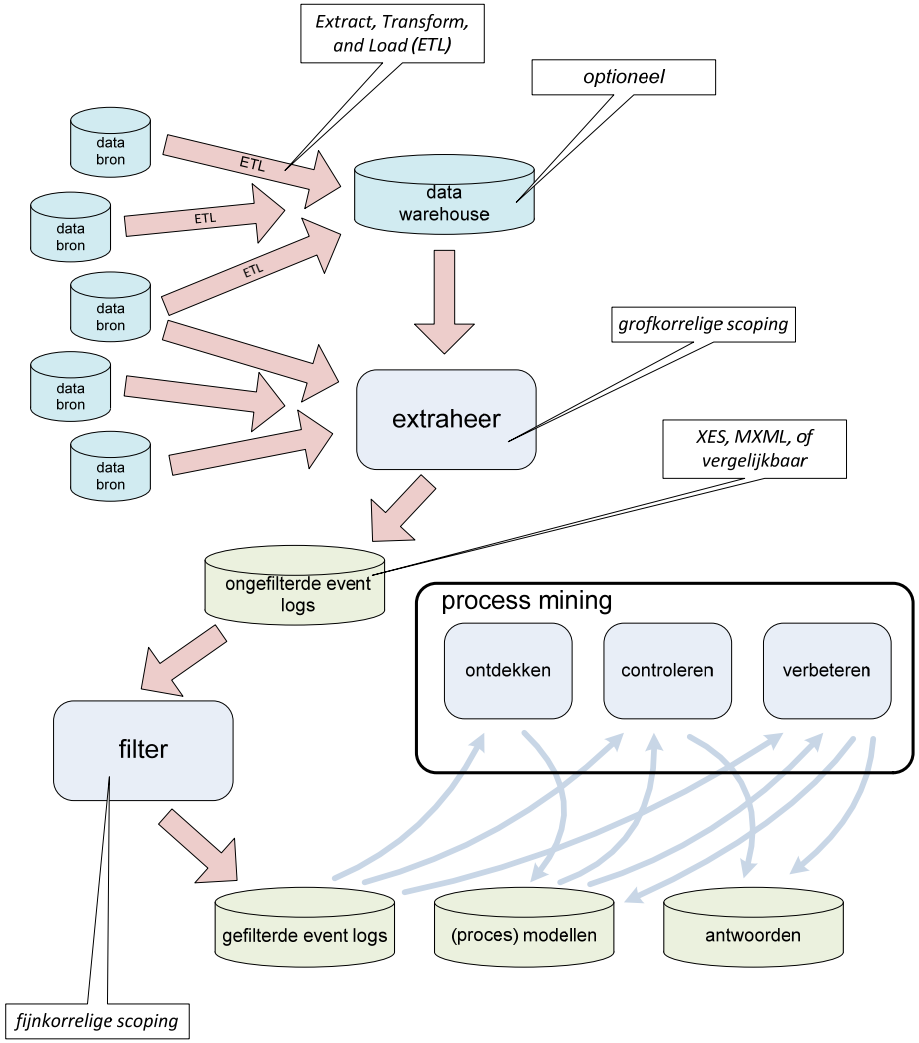
12-11-2014

Dr.ir. Ronny Mans

TU/e Technische Universiteit Eindhoven University of Technology

Where innovation starts

Van heterogene data bronnen naar process mining resultaten

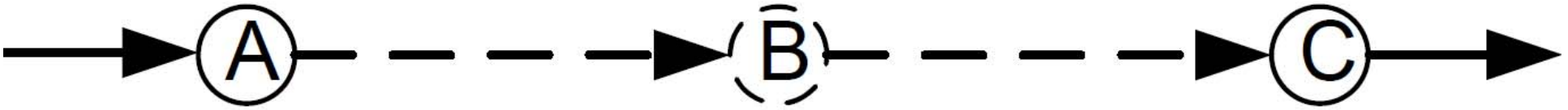


Datakwaliteits issues

- Ontbrekende case ID
- Onnauwkeurige tijdstempels
- Granulariteit van events
- Ontbrekende events
- ….

Datakwaliteits issues

Missing events

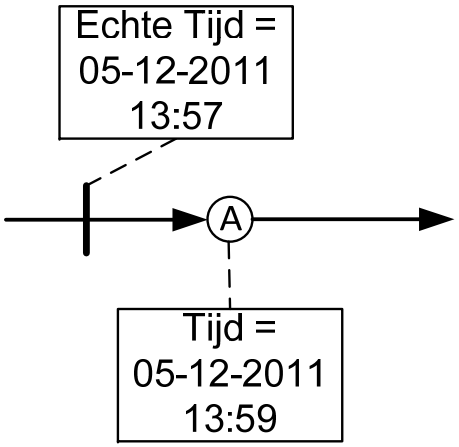


- **Process Mining analyse: ontdekking van verkeerde relaties**

- **vb: radiologie verrichtingen ontbreken**

Datakwaliteits issues

- **Foutief tijdstempel**

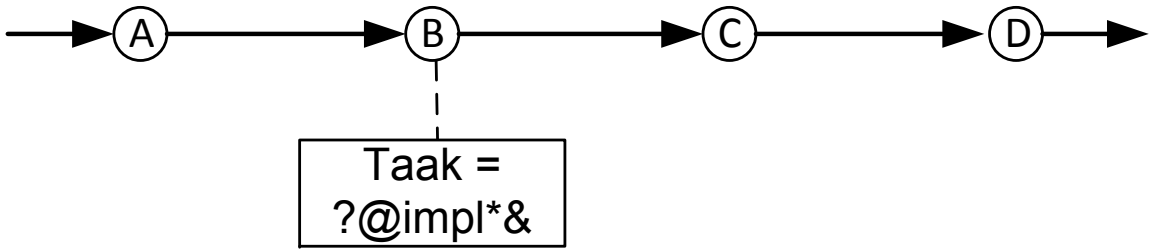


- **Process Mining analysis: Ontdekte control-flow relaties zijn onbetrouwbaar/foutief**

- **Vb: IC database: events met dezelfde tijd of 1ms verschil**

Datakwaliteits issues

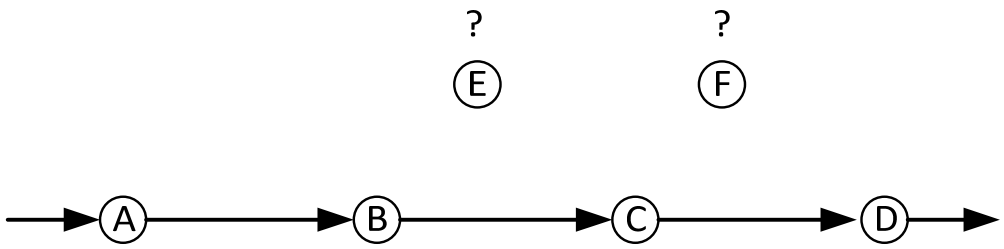
- **Onnauwkeurige activiteitsnaam**



- **vb: taaknamen:**
  - **imp. cons**
  - **impl cons: 15 min eerder!!**
  - **kaart! impl cons: 15 min eerder!!**
  - **kaart !! Impl cons: 15 min eerder!!**

Datakwaliteits issues

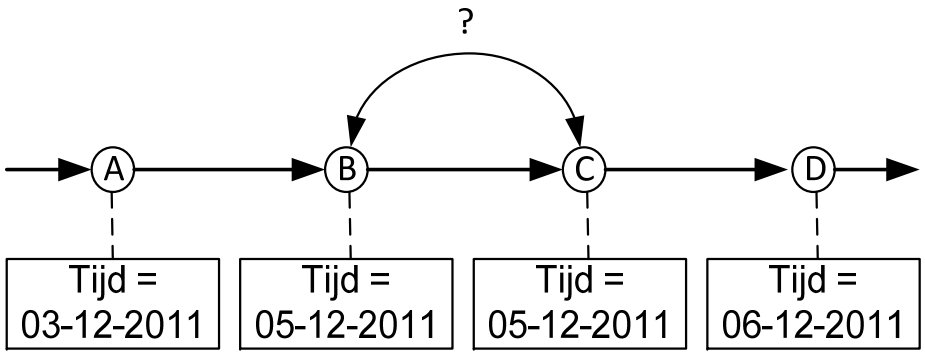
- **Onnauwkeurige relatie tussen events en case**



- **vb: tandheelkunde**
  - **Implantoloog: patiënt: J. Jansen**
  - **Tandheelkundig lab: patiënt: Jansen, J.**
  - **Tandarts: patiënt: John Jansen**

Datakwaliteits issues

- **Onnauwkeurig tijdstempel**



- **Process Mining analyse: Ontdekte control-flow relaties zijn onbetrouwbaar/foutief (veel activiteiten parallel)**

- **vb: DBC/DOT data met alleen dagtijdstempel**

Datakwaliteit 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

Datakwaliteits issues

Evaluatie van ZIS van Nederlands ziekenhuis

	case	event	relationship	c_attribute	position	Activity name	timestamp	respirce	e_attribute
Missing data	N	H	L	L	N	L	N	N	L
Incorrect data	N	L	L	L	N	L	L	N	L
Imprecise data			N	N	N	N	H	H	N
Irrelevant data									

Uitdagingen

- **Zijn de tijdstempels correct?**
- **Zijn de tijdstempels precies?**
- **Heb ik alle events?**
- **Heb ik de juiste events?**

Samenvatting

- **Data kwaliteit is belangrijk!**
- **Zoek voor problemen en beslis hoe er mee om te gaan.**
- **Regels over vastleggen van data.**

Vragen?