process mining

hybrid process discovery
Positioning PM
Launched in 2013

DSC/e
Data Science Center Eindhoven
Turning Data into Value

Wil van der Aalst
Scientific Director DSC/e

THE PERFECT DATA SCIENTIST

OPEN MIND
CREATIVE
HUMAN INTEREST
ANALYTICAL
KNOW HOW TO DO BUSINESS

©Marlon van de Wiel 2014
Data Science as “the enabler”

infrastructure
“volume and velocity”
- instrumentation
- big data infrastructures and distributed systems
- databases and data management
- programming
- security
- ...

analysis
“extracting knowledge”
- statistics
- data/process mining
- machine learning/artificial intelligence
- operations research
- algorithms
- visualization
- ...

effect
“people, organizations, society”
- ethics & privacy
- IT law
- human technology interaction
- operations management
- business models
- entrepreneurship
- ...

the data science pipeline
Uptake of process mining

Evolved from 29 plug-ins in 2004 to 274 plug-ins 2009 (ProM 5.2) to over 1500 plug-ins today.

Start of process mining at TU/e (1999)

Alpha miner & Heuristic miner (2000-2002)

First version of ProM (2004)

Conformance checking, other perspectives, prediction, etc. (2005 - )

Process mining book, courses, uptake commercial tools, etc. (2011 - )

publications on process mining (2016/17 incomplete) based on Scopus data
process mining as the missing link
12,666 cases (orders)
80,609 events

discovered using ILP miner
Taxonomy: Not just CF discovery!

Task
- Process discovery
- Conformance checking
- Extending process models
- Decision making

Perspective
- Control Flow (CF) only
- CF + time
- CF + data
- CF + resources

Type
- Offline
- Online
Taxonomy: Not just CF discovery!

- **Task**
  - Process discovery
  - Conformance checking
  - Extending process models
  - Decision making

- **Perspective**
  - Control Flow (CF) only
  - CF + time
  - CF + data
  - CF + resources

- **Type**
  - Offline
  - Online
process model discovered using the inductive miner (showing only the most frequent paths)
zooming in

prepare delivery 11015
11015
11015
11015
+
11015
make delivery 11015
11015
confirm payment 11015
1651
cancel order 1651
1651
11015
using conformance checking to see all deviations
bottleneck analysis: enriching the model with performance information
animating the event log
showing real cases
seamless abstraction: one log many views
Data-driven & process-centric

event data

Data-driven & process-centric

process models
Answering two types of questions

**Performance questions**
- Why are these cases late?
- Where are the real bottlenecks?
- Which resources are overloaded?

**Conformance questions**
- How often is the four-eyes principle violated?
- Which activities are often skipped?
- Which resources cause deviations?

Event data → Process-centric → Data-driven

Process models → Data-driven → Process-centric

©Wil van der Aalst (use only with permission & acknowledgements)
Process mining results may hurt and trigger resistance, but this only supports the need for it.
Process Mining Software
1500+ plug-ins available covering the whole process mining spectrum

>130k downloads

©Wil van der Aalst (use only with permission & acknowledgements)
Interaction with industry

- ideas
- challenges
- new techniques and approaches
- data

©Wil van der Aalst (use only with permission & acknowledgements)
Job done?

Not really ...
Concurrency and Semantics
Boxes and arrows: What do they mean?
An example log and “proper models”

1000 cases and 5000 events

Inductive miner, ILP miner, Alpha miner, etc.
Model with 3 concurrent activities

Too difficult 😊?
Concurrency & Semantics

Much easier?

Semantics?
Concurrency & Semantics

Do frequencies make sense?
Do frequencies make sense?

- do not add up
- were performed in any order
Do times make sense?
Do times make sense?
Correct numbers
Concurrency & Semantics

Correct times

51.14 hours → pay ticket → 194.39 hours
81.48 hours → confirm receipt → 164.05 hours
117.71 hours → get ticket → 127.81 hours

Visit concert
But ...
faking confidence
(even with fitness = 1.0)
Need to have a PhD in Petri nets?
How to combine the best of both worlds?

- informal when needed & fast and scalable
- precise (having formal semantics) whenever possible & useful

- commercial tools
- heuristic miner
- fuzzy miner
- etc.

- basic inductive miner
- ILP miner
- other region-based approaches
- etc.
Idea: Hybrid process models

Joint work with Riccardo De Masellis, Chiara Di Francescomarino, Chiara Ghidini
People do not hate Petri nets (BPMN, etc.): they hate to be precise (when ...!)

Vagueness can be a feature!

Semi-structuredness can be deliberate!
Hybrid Petri nets have three types of arcs:

1. **Sure and precise**
   - Strong causality determined based on thresholds
   - Logic can be captured*

2. **Sure but not precise**
   - Weak causality determined based on thresholds
   - Logic cannot be captured*

3. **Unsure**
   - Unsure

* = easily / of a certain quality / within the representational bias

© Wil van der Aalst (use only with permission & acknowledgements)
Hybrid Petri nets have three types of arcs

- **Type I**: strong causality (sure)
- **Type II**: weak causality (unsure)
- **Type III**: informal

**Informal** (annotations that are deliberately vague)

**Formal** (firm statements about the inclusion or exclusion of traces)

- **Strong causality** ("sure")
- **Weak causality** ("unsure")
Phase 0: Get data
Phase 1: Learn a Causal Graph

Different approaches possible (business as usual)

- Start
- Place order
- Send invoice
- Send reminder
- Cancel order
- Pay
- Make delivery
- Send delivery
- Pay
- Prepare delivery
- Confirm payment
- Make delivery
- End

Strong causal relation $R_s$
Weak causal relation $R_w$
Phase 2: Learn a Hybrid Petri Net

1. Start
2. Place order
3. Send invoice
4. Pay
5. Make delivery
6. Send reminder
7. Cancel order
8. Prepare delivery
9. Confirm payment
10. Make delivery
11. End
Phase 2: Learn a Hybrid Petri Net
Phase 2: Learn a Hybrid Petri Net
Phase 2: Learn a Hybrid Petri Net
Phase 2: Learn a Hybrid Petri Net

- **How?** Generate **candidate places**
- **A candidate place** is characterized by two sets of activities $A_{\text{input}}$ and $A_{\text{output}}$ such that sure arcs are connecting any activity in $A_{\text{input}}$ to any activity in $A_{\text{output}}$


![Diagram of a Hybrid Petri Net]

©Wil van der Aalst (use only with permission & acknowledgements)
Phase 2: Learn a Hybrid Petri Net

- **How?** Generate **candidate places**
- A candidate place is characterized by two sets of activities $A_{input}$ and $A_{output}$ such that sure arcs are connecting any activity in $A_{input}$ to any activity in $A_{output}$
Phase 2: Learn a Hybrid Petri Net

- Determine the quality of each candidate place $p=(A_{input}, A_{output})$

- Ideally: start empty, finish empty, not negative (compare ILP miner)
Phase 2: Learn a Hybrid Petri Net

A_{input}  A_{output}

- Three possible scoring functions:
  a) Fraction of cases perfectly fitting
  b) Fraction of relevant cases perfectly fitting
  c) Global score (extremely efficient)
Generate candidate places and evaluate

\[
\begin{align*}
&\text{co} \quad \text{cancel order} \\
&\text{cp} \quad \text{confirm payment} \\
&\text{md} \quad \text{make delivery} \\
&\{\text{co}\},\{e\} \\
&\{\text{cp}\},\{e\} \\
&\{\text{md}\},\{e\} \\
&\{\text{co,cp}\},\{e\} \\
&\{\text{co,md}\},\{e\} \\
&\{\text{cp,md}\},\{e\} \\
&\{\text{co,cp,md}\},\{e\} \\
\end{align*}
\]
Generate candidate places and evaluate

- cancel order: \(\{\{co\},\{e\}\}\)
- confirm payment: \(\{\{cp\},\{e\}\}\)
- make delivery: \(\{\{md\},\{e\}\}\)
- cancel order, confirm payment: \(\{\{co, cp\},\{e\}\}\)
- cancel order, make delivery: \(\{\{co, md\},\{e\}\}\)
- confirm payment, make delivery: \(\{\{cp, md\},\{e\}\}\)
- cancel order, confirm payment, make delivery: \(\{\{co, cp, md\},\{e\}\}\)
ProM Implementation

Fuzzy Causal Graph

Start
Send
Invoice
Confirm
Payment
Pay
Make
Delivery
Send
Reminder
Cancel
Order
Place
Order
Prepare
Delivery

End

ProM Implementation

- Fuzzy causal graph
- Fuzzy Petri net
- Sure arc
- Unsure arc
- Sliders

Workflow:

- Start
- Send invoice
- Confirm payment
- Pay
- Make delivery
- Send reminder
- Cancel order
- Place order
- Prepare delivery
- ?
- End

Nodes:

- p1
- p2
- p3
- p4
- p5
- p6
- p7
- p8
- p9
Parameters

- Threshold for activity frequency ($t_{freq}$)
- Parameters used to compute strength of relations taking into account concurrency and loops ($c$ and $w$)
- Thresholds for strong and weak causalities ($t_{RS}$ and $t_{RW}$ with $t_{RS} > t_{RW}$)
- Threshold for place quality ($t_{replay}$)
Evaluation (see paper and technical report for more details)

<table>
<thead>
<tr>
<th>Log</th>
<th>$t_{freq}$</th>
<th>$t_{RS}$</th>
<th>$t_{RW}$</th>
<th>$w$</th>
<th>$t_{replay}$</th>
<th>$T$</th>
<th>$P$</th>
<th>$\hat{F}_1$</th>
<th>$F_2$</th>
<th>$F_3$</th>
<th>Fitness</th>
<th>Precision</th>
<th>Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPI-2011</td>
<td>343</td>
<td>0.81</td>
<td>0.80</td>
<td>0.10</td>
<td>0.80</td>
<td>38</td>
<td>6</td>
<td>4</td>
<td>200</td>
<td>6</td>
<td>0.84</td>
<td>0.04</td>
<td>11772</td>
</tr>
<tr>
<td>BPI-2012</td>
<td>3926</td>
<td>0.90</td>
<td>0.89</td>
<td>0.10</td>
<td>0.80</td>
<td>14</td>
<td>8</td>
<td>7</td>
<td>20</td>
<td>1</td>
<td>0.90</td>
<td>0.26</td>
<td>12414</td>
</tr>
<tr>
<td>BPI-2014</td>
<td>13985</td>
<td>0.90</td>
<td>0.90</td>
<td>0.10</td>
<td>0.80</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>13</td>
<td>0</td>
<td>0.93</td>
<td>0.54</td>
<td>21233</td>
</tr>
<tr>
<td>BPI-2015</td>
<td>360</td>
<td>0.45</td>
<td>0.40</td>
<td>0.50</td>
<td>0.80</td>
<td>59</td>
<td>26</td>
<td>24</td>
<td>145</td>
<td>75</td>
<td>0.74</td>
<td>0.05</td>
<td>7055</td>
</tr>
<tr>
<td>BPI-2016</td>
<td>445</td>
<td>0.50</td>
<td>0.50</td>
<td>0.10</td>
<td>0.80</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>31</td>
<td>0</td>
<td>0.83</td>
<td>0.10</td>
<td>31428</td>
</tr>
<tr>
<td>BPI-2017</td>
<td>9453</td>
<td>0.51</td>
<td>0.50</td>
<td>0.50</td>
<td>0.80</td>
<td>22</td>
<td>8</td>
<td>7</td>
<td>36</td>
<td>12</td>
<td>0.95</td>
<td>0.12</td>
<td>24772</td>
</tr>
</tbody>
</table>

- Behaves as expected, e.g., when $t_{replay}$ goes up, fitness goes up and precision goes down
Performance

• Good, but room for improvement
• Smartly pruning the set of candidate places (avoid conflicting or less informative places)
• Lazy place evaluation
• Distribution/decomposition using e.g. Spark (see joint work with Long Cheng and Boudewijn van Dongen in a slightly different setting)
Evaluation is not easy

We interviewed hundreds of users and turned all of their suggestions into features.

As it turns out, every user we talked to was an idiot, and their dumb suggestions ruined our product.

In hindsight, we probably should have talked to people who work outside this building.
But, the need is obvious
But, the need is obvious
Conclusion
process mining

hybrid process discovery
Learning Hybrid Process Models from Events

Process Discovery Without Faking Confidence
Thanks!

Looking for talented PhDs and Postdocs!!