No Knowledge Without Processes
Process Mining as a Tool to Find Out What People and Organizations Really Do

prof.dr.ir. Wil van der Aalst
www.processmining.org
process mining
intro
Process discovery
Process Discovery
Conformance checking
Conformance Checking

ddfeabb
Let's play
### Play-Out

<table>
<thead>
<tr>
<th>Case</th>
<th>Activity</th>
<th>Timestamp</th>
<th>Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>432</td>
<td>register travel request (a)</td>
<td>18-3-2014:9.15</td>
<td>John</td>
</tr>
<tr>
<td>432</td>
<td>get support from local manager (b)</td>
<td>18-3-2014:9.25</td>
<td>Mary</td>
</tr>
<tr>
<td>432</td>
<td>check budget by finance (d)</td>
<td>19-3-2014:8.55</td>
<td>John</td>
</tr>
<tr>
<td>432</td>
<td>decide (e)</td>
<td>19-3-2014:9.36</td>
<td>Sue</td>
</tr>
<tr>
<td>432</td>
<td>accept request (g)</td>
<td>19-3-2014:9.48</td>
<td>Mary</td>
</tr>
</tbody>
</table>

**Diagram:**

- Start
- Register travel request (a)
- Get detailed motivation letter (c)
- Get support from local manager (b)
- Check budget by finance (d)
- Decide (e)
- Accept request (g)
- Reject request (h)
- Reinitiate request (f)
- End
Play Out: A possible scenario

**a b d e g**

<table>
<thead>
<tr>
<th>Case</th>
<th>Activity</th>
<th>Timestamp</th>
<th>Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>432</td>
<td>register travel request (a)</td>
<td>18-3-2014:9.15</td>
<td>John</td>
</tr>
<tr>
<td>432</td>
<td>get support from local manager (b)</td>
<td>18-3-2014:9.25</td>
<td>Mary</td>
</tr>
<tr>
<td>432</td>
<td>check budget by finance (d)</td>
<td>19-3-2014:8.55</td>
<td>John</td>
</tr>
<tr>
<td>432</td>
<td>decide (e)</td>
<td>19-3-2014:9.36</td>
<td>Sue</td>
</tr>
<tr>
<td>432</td>
<td>accept request (g)</td>
<td>19-3-2014:9.48</td>
<td>Mary</td>
</tr>
</tbody>
</table>
Play Out: Another scenario

register travel request (a)
get detailed motivation letter (c)
get support from local manager (b)
check budget by finance (d)
decide (e)
accept request (g)
reject request (h)
reinitiate request (f)
start
end
Play Out: Process model allows for many more scenarios

register travel request (a) ->
get detailed motivation letter (c) ->
check budget by finance (d) ->
decide (e) ->
accept request (g) ->
reject request (h) ->
reinitiate request (f) ->
start ->
end
<table>
<thead>
<tr>
<th>Case</th>
<th>Activity</th>
<th>Timestamp</th>
<th>Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>432</td>
<td>register travel request (a)</td>
<td>18-3-2014:9.20</td>
<td>John</td>
</tr>
<tr>
<td>432</td>
<td>get detailed motivation letter (c)</td>
<td>18-3-2014:9.30</td>
<td>Sam</td>
</tr>
<tr>
<td>432</td>
<td>get support from local manager (b)</td>
<td>18-3-2014:9.40</td>
<td>John</td>
</tr>
<tr>
<td>432</td>
<td>check budget by finance (d)</td>
<td>19-3-2014:8.50</td>
<td>John</td>
</tr>
<tr>
<td>432</td>
<td>decide (e)</td>
<td>19-3-2014:9.60</td>
<td>Sue</td>
</tr>
<tr>
<td>432</td>
<td>accept request (g)</td>
<td>19-3-2014:9.70</td>
<td>Mary</td>
</tr>
</tbody>
</table>

Play-In
Play In: Simple process allowing for 4 traces

abdeg  adbeg  adbeg  adbeg
abdeh  abdeh  abdeh  abdeh
abdeh  abdeh  abdeh  abdeh
abdeh  abdeh  abdeh  abdeh
Play In:  
Process allowing for more traces
No modeling needed!
Example Process Discovery
(Dutch housing agency, 208 cases, 5987 events)
Example process discovery for hospital
(627 gynecological oncology patients, 24331 events)
Process discovery algorithms
(small selection)

- automata-based learning
- heuristic mining
- genetic mining
- stochastic task graphs
- ETM genetic algorithm
- fuzzy mining
- mining block structures
  - α algorithm
  - α# algorithm
  - α++ algorithm
- distributed genetic mining
- language-based regions
- state-based regions
- LTL mining
- Inductive Miner (infrequent)
- neural networks
- hidden Markov models
- conformal process graph
- partial-order based mining
- ILP mining
Region \( R = (X,Y,c) \) corresponding to place \( p_R \): \( X = \{a_1,a_2,c_1\} \) = transitions producing a token for \( p_R \), \( Y = \{b_1,b_2,c_1\} \) = transitions consuming a token from \( p_R \), and \( c \) is the initial marking of \( p_R \).
Basic idea: enough tokens should be present when consuming

A place is feasible if it can be added without disabling any of the traces in the event log.

for any $\sigma \in L, k \in \{1, \ldots, |\sigma|\}$, $\sigma_1 = h d^{k-1}(\sigma)$, $a = \sigma(k)$, $\sigma_2 = h d^k(\sigma) = \sigma_1 \oplus a$:

$$c + \sum_{t \in X} \partial_{\text{multiset}}(\sigma_1)(t) - \sum_{t \in Y} \partial_{\text{multiset}}(\sigma_2)(t) \geq 0.$$
Replay

<table>
<thead>
<tr>
<th>Case</th>
<th>Activity</th>
<th>Timestamp</th>
<th>Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>432</td>
<td>register travel request (a)</td>
<td>18-3-2014:9:15</td>
<td>John</td>
</tr>
<tr>
<td>432</td>
<td>get support from local manager (b)</td>
<td>18-3-2014:9:25</td>
<td>Mary</td>
</tr>
<tr>
<td>432</td>
<td>check budget by finance (d)</td>
<td>19-3-2014:8:55</td>
<td>John</td>
</tr>
<tr>
<td>432</td>
<td>decide (e)</td>
<td>19-3-2014:9:36</td>
<td>Sue</td>
</tr>
<tr>
<td>432</td>
<td>accept request (g)</td>
<td>19-3-2014:9:48</td>
<td>Mary</td>
</tr>
</tbody>
</table>
Replay

register travel request (a)
get detailed motivation letter (c)
get support from local manager (b)
check budget by finance (d)
decide (e)
accept request (g)
reject request (h)
reinitiate request (f)
start end
Replay

a c e g

check budget (d) is missing!
Alignments: Relating reality and model

Register travel request (a) 
Get detailed motivation letter (c) 
Get support from local manager (b) 
Check budget by finance (d) 
Decide (e) 
Accept request (g) 
Reject request (h) 
Reinitiate request (f)

Check budget (d) did not happen but should have according to the model.
a c h d e g

? reject request (h) is impossible
Alignments: Relating reality and model

register travel request (a)  
get detailed motivation letter (c)  
get support from local manager (b)  
check budget by finance (d)  
reinitiate request (f)  
accept request (g)  
reject request (h)

reject request (h) happened but could not happen according to the model
Any trace in reality can be related to a path in the model

register travel request (a)
get detailed motivation letter (c)
get support from local manager (b)
check budget by finance (d)
decide (e)
accept request (g)
reject request (h)
reinitiate request (f)

start

end
Any trace in reality can be related to a path in the model

start

check is missing

check is missing

c

c

c

c

c

check budget by finance (d)

get detailed motivation letter (c)

one check too many

cannot both be done

decide (e)

reinitiate request (f)

accept request (g)

reject request (h)

end

optimization problem using a cost function
process model

event log

synchronous move

move on log only

move on model only
Replay with timestamps

- Register travel request (a) at 9.15
- Get detailed motivation letter (c) at 9.20
- Check budget by finance (d) at 9.35
- Decide (e) at 10.15
- Accept request (g) at 11.30
- Reject request (h)
Replay with timestamps for many traces

- Frequencies of activities
- Durations of activities
- Waiting times and other delays between activities

1. Register travel request (a)
2. Get support from local manager (b)
3. Get detailed motivation letter (c)
4. Check budget by finance (d)
5. Decide (e)
6. Accept request (g)
7. Reject request (h)
8. Reinitiate request (f)
9. End
Alignments are essential!

- conformance checking to diagnose deviations
- squeezing reality into the model to do model-based analysis

\[
\begin{array}{cccccccc}
  a & c & e & f & d & d & b & c & e & h \\
  a & c & d & e & f & d & d & » & b & » & e & h \\
\end{array}
\]
Many moves on log of "O_CANCELED", "O_CREATED", "O_SELECTED", "O_SENT" occurred with the same frequency value (i.e. 60) before parallel branch.

Loops of "W_Completeren aanvraag" and "W_Nabellen offertes" are often performed.

"O_DECLINED" and "W_Wijzigen contractgegevens" are often skipped.

"W_Afhandelen leads" ( > 2200 times) occurred in the end of traces.

Work of Arya Adriansyah (Replay project)
Loops of “W_Completeren aanvraag” and “W_Nabellen offertes” are often performed in the end of traces. Many moves on log of “O_DECLINED” and “W_Wijzigen contractgegevens” are often skipped.

Synchronous moves of “Completeren aanvraag” and “O_DECLINED” and “W_Wijzigen contractgegevens” are often performed.

Many moves on log of “W_Afhandelen leads” occurred in the end of traces. Loops of “W_Completeren aanvraag” and “W_Nabellen offertes” are often performed.

Synchronous moves of “Completeren aanvraag” and “O_DECLINED” and “W_Wijzigen contractgegevens” are often skipped.

Moves on log of “O_CANCELLLED” and “A_CANCELLLED” occurred in the end of traces. Moves on model towards end of traces.
"O_ACCEPTED" has average sojourn time of 27.07 minutes, while "A_REGISTERED", "A_ACTIVATED", and "A_APPROVED" have average sojourn time of 29.56 minutes. Activity "W_Wijzigen contractgegevens" is the bottleneck, but it occurred rarely (only 4 times). The average waiting time for the input place of "W_Nabellen offertes+START" is very long (2.83 days) compared to the average waiting time of other places.
600+ plug-ins available covering the whole process mining spectrum
Overview: Role of process models

- **models** analyzes
- **analyzes**
- **configures**
- **implements**
- **analyzes**
- **supports/controls**

- **Play-Out**
- **discovery**
- **conformance**
- **enhancement**

- **software system**
  - records events, e.g., messages, transactions, etc.

- **event logs**

- **(process) model**

- **business processes**
  - people
  - machines
  - components
  - organizations

- "world"
decomposed/distributed process mining
“DATA IS THE NEW OIL.”

From the beginning of recorded time until 2003, we created 5 exabytes (5 billion gigabytes) of data. In 2011 the same amount was created every two days. By 2013, it’s expected that the time will shrink to 10 minutes.

Every hour create enough photos to fill all of North America. Every second we create 247 billion emails, sent every day. (Up to 60% are spam.)

There are 133 million blogs on the web. As of August 2012, there were just over 4 million articles in the English Wikipedia.

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, high-frequency traders, with the help of computer algorithms, use Big Data to follow trends and act quickly.

80% of all humans own a mobile phone of some sort. Out of 3 billion mobiles, 1 billion are smartphones. In Singapore, 56% of citizens are smartphone users.

60% of all humans (8.4 billion people) are active Facebook users. In 2010, 103,000 text messages were sent every second.

10% of all photos ever taken were taken in 2011.

These specialized algorithms make split-second decisions to buy or sell at up to 0.1 second from the time a trade is being placed (5 milliseconds).

With new fiber-optic cable, the round-trip time between New York and London will be 59.8 milliseconds. This 5-millisecond saving is worth many millions of dollars to the firms who use the cable (and pay millions to do so).

How they save 5 m:
The depth of the Atlantic:
The new cable will lie on the seafloor that area up to 1.4 times the current for a different route. It will be shorter, making it take for more information along its way.

50% of 6-year-old kids in the U.S. are given access to a smartphone.
What if?

there are more than 100.000.000 events?

there are more than 1000 different activities?

there are more than 1.000.000 cases?
Decompose event log!
vertical or horizontal

sets of cases

sets of activities
Vertical distribution: Split cases

sets of cases
Horizontal distribution

sets of activities

\{a,b,e,f,g\}

\{b,c,d,e\}

\begin{itemize}
  \item \text{abceg}
  \item \text{abdceg}
  \item \text{abcdegc}
  \item \text{abedefbcdeg}
  \item \text{abdcefgd}
  \item \text{abcdeg}
  \item \text{abdceg}
  \item \text{abcdeg}
  \item \text{abedefbcde}
  \item \text{abedefbcdeg}
\end{itemize}

\begin{itemize}
  \item \text{abeg}
  \item \text{abefbeg}
  \item \text{abeg}
  \item \text{abefbeg}
  \item \text{abeg}
  \item \text{abefbeg}
  \item \text{abeg}
  \item \text{abeg}
  \item \text{abefbeg}
  \item \text{abeg}
\end{itemize}

\begin{itemize}
  \item \text{bcde}
  \item \text{bdcebcde}
  \item \text{bdce}
  \item \text{bcdebcde}
  \item \text{bdcebdce}
  \item \text{bcdebdce}
  \item \text{bcde}
  \item \text{bdce}
  \item \text{bcdebdce}
  \item \text{bcdebdce}
\end{itemize}
Horizontal distribution: The key idea

projected on \{a,b,e,f,g\}

projected on \{b,c,d,e\}
Two foundational ways of splitting event data: horizontal or vertical
Decomposing Conformance Checking

- Decomposition technique
  - SN process model
  - Decompose model
  - L event log
  - Decompose event log

- Conformance checking technique
  - M\(^1\) submodel
  - M\(^2\) submodel
  - M\(^n\) submodel

- Example decomposition techniques:
  - Maximal decomposition, passage-based decomposition, or SESE/RPST-based decomposition
  - A* based alignments, token-based replay, or simple replay until first deviation

- Yields a (valid) activity partitioning

See "divide and conquer" framework by Eric Verbeek.
Example of a valid decomposition

Log can be split in the same way!
Example of an alignment for observed trace a,b,c,d,e,c,d,g,f

\[ \gamma_3 = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ a & b & c & d & e & c \tau & d \tau & \gamma & d \gamma & \gamma & \gamma \\ \end{bmatrix} \]

Etc.

\[ \gamma_3^1 = \begin{bmatrix} 1 \\ a \\ \end{bmatrix}, \quad \gamma_3^2 = \begin{bmatrix} 1 & 2 & 4 & 5 & 7 & 8 \\ a & b & d & e \tau & \gamma & d \\ t1 & t3 & t5 & t6 & t2 & t5 \\ \end{bmatrix}, \quad \gamma_3^3 = \begin{bmatrix} 1 & 3 & 5 & 6 \\ \gamma & \gamma & \gamma & \gamma \\ a & c & c & c \\ t1 & t4 & t6 & t4 \\ \end{bmatrix} \]

\[ \gamma_3^4 = \begin{bmatrix} 3 & 4 & 6 & 8 \\ c & d & c & d \\ c & d & c & d \\ t4 & t5 & t4 & t5 \\ \end{bmatrix}, \quad \gamma_3^5 = \begin{bmatrix} 4 & 5 & 8 & 9 & 10 & 11 \\ d & e & d \tau & \gamma & \gamma & d \\ t5 & t6 & t5 & t7 & t9 & t8 \\ \end{bmatrix}, \quad \gamma_3^6 = \begin{bmatrix} 10 & 11 & 12 \\ g & f \tau & \gamma & \gamma & \gamma \\ g & f \tau & \gamma & \gamma & \gamma \\ t9 & t8 & t11 & t11 & t11 & t11 \\ \end{bmatrix} \]
Conformance checking can be decomposed!!!

- General result for any valid decomposition: Any event log or trace is perfectly fitting the overall model if and only if it is also fitting all the individual fragments

Example
(work with Jorge Munoz-Gama and Josep Carmona)

Diagnose Subprocess
Detect an unfitting subprocess, analyze it in isolation, and diagnose the cause of the problems.

Diagnose Non Fitting Net
Detect all unfitting subprocesses, compose the net that contains all them, and diagnose the cause of the problems.
Decomposing Process Discovery

- **Decomposition Technique**
  - Decomposing event log $L$ into sublogs $L_1, L_2, \ldots, L_n$
  - Yields a (valid) activity partitioning

- **Process Discovery Technique**
  - E.g., causal graph based on frequencies decomposed using passages or SESE/RPST
  - Language/state-based region discovery, variants of alpha algorithm, genetic process mining

See "divide and conquer" framework by Eric Verbeek.
conclusion
Process mining: mediating between modeled and observed behavior

Decomposition as a way to deal with "Big" process mining tasks

Many challenges, e.g., process discovery

Learn more?
More and more information about business processes is recorded by information systems in the form of so-called "event logs". Despite the omnipresence of such data, most organizations diagnose problems based on fiction rather than facts. Process mining is an emerging discipline based on process model-driven approaches and data mining. It not only allows organizations to fully benefit from the information stored in their systems, but it can also be used to check the conformance of processes, detect bottlenecks, and predict execution problems.

Wil van der Aalst delivers the first book on process mining. It aims to be self-contained while covering the entire process mining spectrum from process discovery to operational support. In Part I, the author provides the basics of business process modeling and data mining necessary to understand the remainder of the book. Part II focuses on process discovery as the most important process mining task. Part III moves beyond discovering the control flow of processes and highlights conformance checking, and organizational and time perspectives. Part IV guides the reader in successfully applying process mining in practice, including an introduction to the widely used open-source tool ProM. Finally, Part V takes a step back, reflecting on the material presented and the key open challenges.

Overall, this book provides a comprehensive overview of the state of the art in process mining. It is intended for business process analysts, business consultants, process managers, graduate students, and BPM researchers.

Features and Benefits:
- First book on process mining, bridging the gap between business process modeling and business intelligence.
- Written by one of the most influential and most-cited computer scientists and the best-known BPM researcher.
- Self-contained and comprehensive overview for a broad audience in academia and industry.
- The reader can put process mining into practice immediately due to the applicability of the techniques and the availability of the open-source process mining software ProM.
First Massive Open Online Course (MOOC) on Process Mining

About the Course

Data science is the profession of the future, because organizations that are unable to use (big) data in a smart way will not survive. It is not sufficient to focus on data storage and data analysis. The data scientist also needs to relate data to process analysis. Process mining bridges the gap between traditional model-based process analysis (e.g., simulation and other business process management techniques) and data-centric analysis techniques such as machine learning and data mining. Process mining seeks the confrontation between event data (i.e., observed behavior) and process models (hand-made or discovered automatically). This technology has become available only recently, but it can be applied to any type of operational processes (organizations and systems). Example applications include analyzing treatment processes in hospitals, improving customer service processes in a multinational, understanding the browsing behavior of customers using a booking site, analyzing failures of a baggage handling system, and improving the user interface of an X-ray machine. All of these applications have in common that dynamic behavior needs to be related to process models. Hence, we refer to this as "data science in action".

The course explains the key analysis techniques in process mining. Participants will learn various process discovery algorithms. These can be used to automatically learn process models from raw event data. Various other process analysis techniques that use event data will be presented. Moreover, the course will provide easy-to-use software, real-life data sets, and practical skills to directly apply the theory in practice.

Sessions

- Nov 12th 2014 - Dec 24th 2014
- Starts in 3 months

Eligible for

Statement of Accomplishment

Course at a Glance

- 5 weeks of study
- 4-6 hours of work / week
- English
- English subtitles

Instructors

Wim van der Aalst
Eindhoven University of Technology

Categories

- data science
- process mining
- business process intelligence
- large scale distributed computing
- visualization visual analytics
- machine learning
- data mining
- statistics stochastics
- algorithms databases
- industrial engineering
- behavioral/ social sciences
- domain knowledge
- visual analytics