20 Years of Process Mining Research
Accomplishments, Challenges, and Open Problems

prof.dr.ir. Wil van der Aalst
RWTH Aachen University
W: vdaalst.com T:@wvdaalst
Positioning

“bridging gaps”
“process management by modeling”

Petri nets
Formal methods
Concurrency theory
BPM, WFM, etc.
Simulation

< 1999

Process mining
Process discovery
Conformance checking

≥ 1999

Predictive analytics

“process management by mining”
Starting point: Event data

<table>
<thead>
<tr>
<th>Case ID</th>
<th>Activity</th>
<th>Resource</th>
<th>Timestamp</th>
<th>product</th>
<th>prod-price</th>
<th>quantity</th>
<th>address</th>
</tr>
</thead>
<tbody>
<tr>
<td>6350</td>
<td>place order</td>
<td>Aiden</td>
<td>2018/02/13 14:29:45.000</td>
<td>APPLE iPhone 6 16 GB</td>
<td>639,00 €</td>
<td>5</td>
<td>NL-7751DG-21</td>
</tr>
<tr>
<td>6283</td>
<td>pay</td>
<td>Lily</td>
<td>2018/02/13 14:39:25.000</td>
<td>SAMSUNG Galaxy S6 32 GB</td>
<td>543,99 €</td>
<td>3</td>
<td>NL-7828AM-11a</td>
</tr>
<tr>
<td>6253</td>
<td>prepare delivery</td>
<td>Sophia</td>
<td>2018/02/13 15:01:33.000</td>
<td>APPLE iPhone 6 16 GB</td>
<td>639,00 €</td>
<td>3</td>
<td>NL-7887AC-13</td>
</tr>
<tr>
<td>6257</td>
<td>prepare delivery</td>
<td>Aiden</td>
<td>2018/02/13 15:03:43.000</td>
<td>SAMSUNG Galaxy S6 32 GB</td>
<td>543,99 €</td>
<td>1</td>
<td>NL-9521KJ-34</td>
</tr>
<tr>
<td>6185</td>
<td>confirm payment</td>
<td>Emily</td>
<td>2018/02/13 15:05:36.000</td>
<td>SAMSUNG Galaxy S4</td>
<td>329,00 €</td>
<td>1</td>
<td>NL-9521GC-32</td>
</tr>
<tr>
<td>6218</td>
<td>confirm payment</td>
<td>Emily</td>
<td>2018/02/13 15:08:11.000</td>
<td>APPLE iPhone 6s Plus 64 GB</td>
<td>969,00 €</td>
<td>2</td>
<td>NL-7948BX-10</td>
</tr>
<tr>
<td>6245</td>
<td>make delivery</td>
<td>Michael</td>
<td>2018/02/13 15:14:04.000</td>
<td>APPLE iPhone 6 16 GB</td>
<td>639,00 €</td>
<td>3</td>
<td>NL-7905AX-38</td>
</tr>
<tr>
<td>6272</td>
<td>pay</td>
<td>Emily</td>
<td>2018/02/13 15:20:36.000</td>
<td>APPLE iPhone 6 16 GB</td>
<td>639,00 €</td>
<td>1</td>
<td>NL-7821AC-3</td>
</tr>
<tr>
<td>6269</td>
<td>pay</td>
<td>Charlotte</td>
<td>2018/02/13 15:25:21.000</td>
<td>SAMSUNG Galaxy S4</td>
<td>329,00 €</td>
<td>1</td>
<td>NL-7907EJ-42</td>
</tr>
<tr>
<td>6212</td>
<td>prepare delivery</td>
<td>Sophia</td>
<td>2018/02/13 15:43:39.000</td>
<td>HUAWEI P8 Lite</td>
<td>234,00 €</td>
<td>1</td>
<td>NL-7905AX-38</td>
</tr>
<tr>
<td>6323</td>
<td>send invoice</td>
<td>Alexander</td>
<td>2018/02/13 15:46:08.000</td>
<td>APPLE iPhone 6 16 GB</td>
<td>639,00 €</td>
<td>1</td>
<td>NL-7833HT-15</td>
</tr>
<tr>
<td>6246</td>
<td>confirm payment</td>
<td>Jack</td>
<td>2018/02/13 15:56:03.000</td>
<td>SAMSUNG Galaxy S4</td>
<td>329,00 €</td>
<td>3</td>
<td>NL-7833HT-15</td>
</tr>
<tr>
<td>6347</td>
<td>send invoice</td>
<td>Jack</td>
<td>2018/02/13 15:57:42.000</td>
<td>SAMSUNG Galaxy S4</td>
<td>329,00 €</td>
<td>3</td>
<td>NL-7905AX-38</td>
</tr>
<tr>
<td>6351</td>
<td>place order</td>
<td>Zoe</td>
<td>2018/02/13 16:17:37.000</td>
<td>APPLE iPhone 5s 16 GB</td>
<td>449,00 €</td>
<td>3</td>
<td>NL-9521GC-32</td>
</tr>
<tr>
<td>6204</td>
<td>prepare delivery</td>
<td>Sophia</td>
<td>2018/02/13 16:31:28.000</td>
<td>SAMSUNG Core Prime G361</td>
<td>135,00 €</td>
<td>1</td>
<td>NL-7828AM-11a</td>
</tr>
<tr>
<td>6204</td>
<td>make delivery</td>
<td>Kaylee</td>
<td>2018/02/13 16:51:54.000</td>
<td>SAMSUNG Core Prime G361</td>
<td>135,00 €</td>
<td>1</td>
<td>NL-7828AM-11a</td>
</tr>
<tr>
<td>6265</td>
<td>confirm payment</td>
<td>Lily</td>
<td>2018/02/13 16:55:55.000</td>
<td>SAMSUNG Galaxy S4</td>
<td>329,00 €</td>
<td>4</td>
<td>NL-9521GC-32</td>
</tr>
<tr>
<td>6250</td>
<td>confirm payment</td>
<td>Jack</td>
<td>2018/02/13 17:03:26.000</td>
<td>MOTOROLA Moto G</td>
<td>199,00 €</td>
<td>4</td>
<td>NL-7942GT-2</td>
</tr>
<tr>
<td>6328</td>
<td>send invoice</td>
<td>Lily</td>
<td>2018/02/13 17:30:16.000</td>
<td>APPLE iPhone 6s 64 GB</td>
<td>858,00 €</td>
<td>4</td>
<td>NL-9514BV-16</td>
</tr>
<tr>
<td>6352</td>
<td>place order</td>
<td>Aiden</td>
<td>2018/02/13 17:53:22.000</td>
<td>APPLE iPhone 6 16 GB</td>
<td>639,00 €</td>
<td>2</td>
<td>NL-9514BV-16</td>
</tr>
<tr>
<td>6317</td>
<td>send invoice</td>
<td>Jack</td>
<td>2018/02/13 18:45:30.000</td>
<td>APPLE iPhone 6s 64 GB</td>
<td>858,00 €</td>
<td>5</td>
<td>NL-7907EJ-42</td>
</tr>
<tr>
<td>6353</td>
<td>place order</td>
<td>Sophia</td>
<td>2018/02/13 20:16:20.000</td>
<td>APPLE iPhone 5s 16 GB</td>
<td>449,00 €</td>
<td>4</td>
<td>NL-7751AR-19</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

71,043 events
12,666 cases
7 activities
Starting point: Event data

<table>
<thead>
<tr>
<th>Case ID</th>
<th>Activity</th>
<th>Resource</th>
<th>Timestamp</th>
<th>product</th>
<th>prod-price</th>
<th>quantity</th>
<th>address</th>
</tr>
</thead>
<tbody>
<tr>
<td>6350</td>
<td>place order</td>
<td>Aiden</td>
<td>2018/02/13 14:29:45.000</td>
<td>APPLE iPhone 6 16 GB</td>
<td>639,00 €</td>
<td>5</td>
<td>NL-7751DG-21</td>
</tr>
<tr>
<td>6283</td>
<td>pay</td>
<td>Lily</td>
<td>2018/02/13 14:39:25.000</td>
<td>SAMSUNG Galaxy S6 32 GB</td>
<td>543,99 €</td>
<td>3</td>
<td>NL-7828AM-11a</td>
</tr>
<tr>
<td>6253</td>
<td>prepare delivery</td>
<td>Sophia</td>
<td>2018/02/13 15:01:33.000</td>
<td>SAMSUNG Galaxy S6 32 GB</td>
<td>543,99 €</td>
<td>3</td>
<td>NL-7887AC-13</td>
</tr>
<tr>
<td>6257</td>
<td>prepare delivery</td>
<td>Aiden</td>
<td>2018/02/13 15:03:43.000</td>
<td>SAMSUNG Galaxy S6 32 GB</td>
<td>543,99 €</td>
<td>1</td>
<td>NL-9521KJ-34</td>
</tr>
<tr>
<td>6185</td>
<td>confirm payment</td>
<td>Emily</td>
<td>2018/02/13 15:05:36.000</td>
<td>SAMSUNG Galaxy S4</td>
<td>329,00 €</td>
<td>1</td>
<td>NL-9521GC-32</td>
</tr>
<tr>
<td>6218</td>
<td>confirm payment</td>
<td>Emily</td>
<td>2018/02/13 15:08:11.000</td>
<td>APPLE iPhone 6s Plus 64 GB</td>
<td>999,00 €</td>
<td>2</td>
<td>NL-7948BX-10</td>
</tr>
<tr>
<td>6245</td>
<td>make delivery</td>
<td>Michael</td>
<td>2018/02/13 15:14:04.000</td>
<td>APPLE iPhone 6s Plus 64 GB</td>
<td>999,00 €</td>
<td>3</td>
<td>NL-7905AX-38</td>
</tr>
<tr>
<td>6272</td>
<td>pay</td>
<td>Emily</td>
<td>2018/02/13 15:20:36.000</td>
<td>APPLE iPhone 6s Plus 64 GB</td>
<td>999,00 €</td>
<td>1</td>
<td>NL-7905AX-38</td>
</tr>
<tr>
<td>6269</td>
<td>pay</td>
<td>Charlotte</td>
<td>2018/02/13 15:25:21.000</td>
<td>SAMSUNG Galaxy S4</td>
<td>329,00 €</td>
<td>1</td>
<td>NL-7907JE-42</td>
</tr>
<tr>
<td>6212</td>
<td>prepare delivery</td>
<td>Sophia</td>
<td>2018/02/13 15:43:39.000</td>
<td>HUAWEI P8 Lite</td>
<td>349,00 €</td>
<td>1</td>
<td>NL-7905AX-38</td>
</tr>
<tr>
<td>6323</td>
<td>send invoice</td>
<td>Alexander</td>
<td>2018/02/13 15:46:08.000</td>
<td>APPLE iPhone 6s Plus 64 GB</td>
<td>999,00 €</td>
<td>3</td>
<td>NL-7833HT-15</td>
</tr>
<tr>
<td>6246</td>
<td>confirm payment</td>
<td>Jack</td>
<td>2018/02/13 15:56:03.000</td>
<td>SAMSUNG Galaxy S4</td>
<td>329,00 €</td>
<td>3</td>
<td>NL-7833HT-15</td>
</tr>
<tr>
<td>6347</td>
<td>send invoice</td>
<td>Jack</td>
<td>2018/02/13 15:57:42.000</td>
<td>SAMSUNG Galaxy S4</td>
<td>329,00 €</td>
<td>3</td>
<td>NL-7905AX-38</td>
</tr>
<tr>
<td>6351</td>
<td>place order</td>
<td>Zoe</td>
<td>2018/02/13 16:17:37.000</td>
<td>APPLE iPhone 5s 16 GB</td>
<td>449,00 €</td>
<td>3</td>
<td>NL-9521GC-32</td>
</tr>
<tr>
<td>6204</td>
<td>prepare delivery</td>
<td>Sophia</td>
<td>2018/02/13 16:31:28.000</td>
<td>SAMSUNG Galaxy S4</td>
<td>329,00 €</td>
<td>3</td>
<td>NL-7828AM-11a</td>
</tr>
<tr>
<td>6204</td>
<td>make delivery</td>
<td>Kaylee</td>
<td>2018/02/13 16:51:54.000</td>
<td>SAMSUNG Galaxy S4</td>
<td>329,00 €</td>
<td>3</td>
<td>NL-7828AM-11a</td>
</tr>
<tr>
<td>6265</td>
<td>confirm payment</td>
<td>Lily</td>
<td>2018/02/13 16:55:50.000</td>
<td>SAMSUNG Galaxy S4</td>
<td>329,00 €</td>
<td>3</td>
<td>NL-9521GC-32</td>
</tr>
<tr>
<td>6250</td>
<td>confirm payment</td>
<td>Jack</td>
<td>2018/02/13 17:03:26.000</td>
<td>MOTOROLA Moto G</td>
<td>199,00 €</td>
<td>4</td>
<td>NL-7942GT-2</td>
</tr>
<tr>
<td>6328</td>
<td>send invoice</td>
<td>Lily</td>
<td>2018/02/13 17:30:16.000</td>
<td>APPLE iPhone 6s 64 GB</td>
<td>858,00 €</td>
<td>4</td>
<td>NL-9514BV-16</td>
</tr>
<tr>
<td>6352</td>
<td>place order</td>
<td>Aiden</td>
<td>2018/02/13 17:53:22.000</td>
<td>APPLE iPhone 6 16 GB</td>
<td>639,00 €</td>
<td>2</td>
<td>NL-9514BV-16</td>
</tr>
<tr>
<td>6317</td>
<td>send invoice</td>
<td>Jack</td>
<td>2018/02/13 18:45:30.000</td>
<td>APPLE iPhone 6s 64 GB</td>
<td>858,00 €</td>
<td>5</td>
<td>NL-7907JE-42</td>
</tr>
<tr>
<td>6353</td>
<td>place order</td>
<td>Sophia</td>
<td>2018/02/13 20:16:20.000</td>
<td>APPLE iPhone 5s 16 GB</td>
<td>449,00 €</td>
<td>4</td>
<td>NL-7751AR-19</td>
</tr>
</tbody>
</table>

event = case + activity + timestamp + ...
Let’s look at orders 6350, 6351, and 6352

<table>
<thead>
<tr>
<th>Case ID</th>
<th>Activity</th>
<th>Timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>6350</td>
<td>place order</td>
<td>2018/02/13 14:29:45.000</td>
</tr>
<tr>
<td>6351</td>
<td>place order</td>
<td>2018/02/13 16:17:37.000</td>
</tr>
<tr>
<td>6352</td>
<td>place order</td>
<td>2018/02/13 17:53:22.000</td>
</tr>
<tr>
<td>6352</td>
<td>send invoice</td>
<td>2018/02/19 09:20:28.000</td>
</tr>
<tr>
<td>6351</td>
<td>send invoice</td>
<td>2018/02/19 16:08:07.000</td>
</tr>
<tr>
<td>6350</td>
<td>send invoice</td>
<td>2018/02/21 09:38:16.000</td>
</tr>
<tr>
<td>6350</td>
<td>pay</td>
<td>2018/03/02 12:39:37.000</td>
</tr>
<tr>
<td>6352</td>
<td>pay</td>
<td>2018/03/05 15:46:47.000</td>
</tr>
<tr>
<td>6351</td>
<td>cancel order</td>
<td>2018/03/06 10:17:01.000</td>
</tr>
<tr>
<td>6350</td>
<td>prepare delivery</td>
<td>2018/03/07 13:50:35.000</td>
</tr>
<tr>
<td>6350</td>
<td>make delivery</td>
<td>2018/03/07 16:41:01.000</td>
</tr>
<tr>
<td>6350</td>
<td>confirm payment</td>
<td>2018/03/07 16:53:00.000</td>
</tr>
<tr>
<td>6352</td>
<td>prepare delivery</td>
<td>2018/03/07 17:05:59.000</td>
</tr>
<tr>
<td>6352</td>
<td>confirm payment</td>
<td>2018/03/07 17:59:55.000</td>
</tr>
<tr>
<td>6352</td>
<td>make delivery</td>
<td>2018/03/08 09:54:36.000</td>
</tr>
</tbody>
</table>
Let’s look at orders 6350, 6351, and 6352

<table>
<thead>
<tr>
<th>Case ID</th>
<th>Activity</th>
<th>Timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>6350</td>
<td>place order</td>
<td>2018/02/13 14:29:45.000</td>
</tr>
<tr>
<td>6351</td>
<td>place order</td>
<td>2018/02/13 16:17:37.000</td>
</tr>
<tr>
<td>6352</td>
<td>place order</td>
<td>2018/02/13 17:53:22.000</td>
</tr>
<tr>
<td>6352</td>
<td>send invoice</td>
<td>2018/02/19 09:20:28.000</td>
</tr>
<tr>
<td>6351</td>
<td>send invoice</td>
<td>2018/02/19 16:08:07.000</td>
</tr>
<tr>
<td>6350</td>
<td>send invoice</td>
<td>2018/02/21 09:38:16.000</td>
</tr>
<tr>
<td>6350</td>
<td>pay</td>
<td>2018/03/02 12:39:37.000</td>
</tr>
<tr>
<td>6352</td>
<td>pay</td>
<td>2018/03/05 15:46:47.000</td>
</tr>
<tr>
<td>6351</td>
<td>cancel order</td>
<td>2018/03/06 10:17:01.000</td>
</tr>
<tr>
<td>6350</td>
<td>prepare delivery</td>
<td>2018/03/07 13:50:35.000</td>
</tr>
<tr>
<td>6350</td>
<td>make delivery</td>
<td>2018/03/07 16:41:01.000</td>
</tr>
<tr>
<td>6350</td>
<td>confirm payment</td>
<td>2018/03/07 16:53:00.000</td>
</tr>
<tr>
<td>6352</td>
<td>prepare delivery</td>
<td>2018/03/07 17:05:59.000</td>
</tr>
<tr>
<td>6352</td>
<td>confirm payment</td>
<td>2018/03/07 17:59:55.000</td>
</tr>
<tr>
<td>6352</td>
<td>make delivery</td>
<td>2018/03/08 09:54:36.000</td>
</tr>
</tbody>
</table>
Using the whole event log

<table>
<thead>
<tr>
<th>Case ID</th>
<th>Activity</th>
<th>Timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>6350</td>
<td>place order</td>
<td>2018/02/13 14:29:45.000</td>
</tr>
<tr>
<td>6351</td>
<td>place order</td>
<td>2018/02/13 16:17:37.000</td>
</tr>
<tr>
<td>6352</td>
<td>place order</td>
<td>2018/02/13 17:53:22.000</td>
</tr>
<tr>
<td>6352</td>
<td>send invoice</td>
<td>2018/02/19 09:20:28.000</td>
</tr>
<tr>
<td>6351</td>
<td>send invoice</td>
<td>2018/02/19 16:08:07.000</td>
</tr>
<tr>
<td>6350</td>
<td>send invoice</td>
<td>2018/02/21 09:38:16.000</td>
</tr>
<tr>
<td>6350</td>
<td>pay</td>
<td>2018/03/02 12:39:37.000</td>
</tr>
<tr>
<td>6352</td>
<td>pay</td>
<td>2018/03/05 15:46:47.000</td>
</tr>
<tr>
<td>6351</td>
<td>cancel order</td>
<td>2018/03/06 10:17:01.000</td>
</tr>
<tr>
<td>6350</td>
<td>prepare delivery</td>
<td>2018/03/07 13:50:35.000</td>
</tr>
<tr>
<td>6350</td>
<td>make delivery</td>
<td>2018/03/07 16:41:01.000</td>
</tr>
<tr>
<td>6350</td>
<td>confirm payment</td>
<td>2018/03/07 16:53:00.000</td>
</tr>
<tr>
<td>6352</td>
<td>prepare delivery</td>
<td>2018/03/07 17:05:59.000</td>
</tr>
<tr>
<td>6352</td>
<td>confirm payment</td>
<td>2018/03/07 17:59:55.000</td>
</tr>
<tr>
<td>6352</td>
<td>make delivery</td>
<td>2018/03/08 09:54:36.000</td>
</tr>
</tbody>
</table>

- place order: 8016 x
- send invoice: 1651 x
- pay: 2962 x
- prepare delivery: 30 x
- make delivery: 7 x
- confirm payment: 16 x
- cancel order: 16 x
Using the whole event log

place order
send invoice
pay
prepare delivery
make delivery
confirm payment

place order
send invoice
pay
prepare delivery
make delivery
confirm payment

place order
send invoice
pay
prepare delivery
cancel order

place order
send invoice
pay
prepare delivery
confirm payment
make delivery

8016 x

1651 x

2962 x

30 x

7 x
Performance and Compliance

What happens?

Where are the bottlenecks?

Where do we deviate from the happy path?
Control-flow: Backbone of any process

- time/frequencies
- data/decisions
- resources/organization
- costs
- etc.
Initial focus academia

Classical Petri nets
Initial focus industry

Directly Follows Graphs

time/frequencies
Towards a more complete picture

discovery

conformance checking

descriptive process models
predictive process models
prescriptive process models

time/frequencies
data/decisions
resources/organization
costs
etc.
Towards a more complete picture

from offline to online

time/frequencies
data/decisions
resources/organization
costs
etc.

from insights to action
Why should I care?
One year to get “Kindergeld” and “Kinderzuschläge” having 0, 1, 3, 4 children and dealing with friction within and between four organizations.

>25 letters, >35 phone calls
Good news: 80% of the cases are described by 20% of the variants.

Bad news: the remaining 20% of the cases account for 80% of the variants and these may take 80% of the time (operational friction).
Example: Process Mining @ Siemens
(thanks to Lars Reinkemeyer, head of process mining Siemens)

• > 6000 active Celonis users (P2P, O2C, etc.)
• Millions of savings by reducing rework, process unification, etc.
• Improved reliability and responsiveness.
• At an amazing scale, e.g., Order to Cash (O2C) process with >30M cases, >300M events, and >900K variants.
data process

process mining
BPM reality

process mining
IT business process mining
• 1999 start of process mining research at TU/e
• 2000-2002 Alpha and Heuristic miner
• 2004 first version of ProM
• 2004-2006 token-based conformance checking, organization mining, decision mining, etc.
• 2007 first process mining company (Futura PI)
• 2009 IEEE Task Force on Process Mining created
• 2010 alignment-based conformance checking
• 2011 first process mining book
• 2011 Process Mining Manifesto published
• 2014 Coursera process mining MOOC
• 2016 “Process mining: data science in action” book
• 2018 Market Guide for Process Mining by Gartner
• 2018 25+ process mining companies
• 2018 Celonis becomes a Unicorn
• 2019 First ICPM conference

20 years of process mining
Accomplishments

“from an academic puzzle to something that works in the real world”
Scientific publications on process mining
(Scopus, June 2019)
Scientific publications on process mining
(Scopus, June 2019)

Examples of progress in:
- process discovery
- conformance checking
- tooling

- IEEE Task Force
- Celonis
- ProcessGold
- Fluxicon
- many new process mining companies created
- second process mining book
- PM manifesto
- first process mining book
- first version of ProM
- first process mining company (Futura PI)

start process mining at TU/e

1999  2001  2003  2005  2007  2009  2011  2013  2015  2017

© Wil van der Aalst (use only with permission & acknowledgements)
Progress in Process Discovery

(Disclaimer: not intended to be complete, but to illustrate the steps towards maturity.)
\[
\left[ \langle a, b, e, f \rangle^2, \langle a, b, e, c, d, b, f \rangle^3, \langle a, b, c, e, d, b, f \rangle^2, \langle a, b, c, d, e, b, f \rangle^4, \langle a, e, b, c, d, b, f \rangle^3 \right]
\]

<table>
<thead>
<tr>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
<th>(f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>#</td>
<td>(\rightarrow)</td>
<td>#</td>
<td>#</td>
<td>(\rightarrow)</td>
</tr>
<tr>
<td>(b)</td>
<td>(\leftarrow)</td>
<td>#</td>
<td>(\rightarrow)</td>
<td>(\leftarrow)</td>
<td>(|)</td>
</tr>
<tr>
<td>(c)</td>
<td>#</td>
<td>(\leftarrow)</td>
<td>#</td>
<td>(\rightarrow)</td>
<td>(|)</td>
</tr>
<tr>
<td>(d)</td>
<td>#</td>
<td>(\rightarrow)</td>
<td>(\leftarrow)</td>
<td>#</td>
<td>(|)</td>
</tr>
<tr>
<td>(e)</td>
<td>(\leftarrow)</td>
<td>(|)</td>
<td>(|)</td>
<td>(|)</td>
<td>#</td>
</tr>
<tr>
<td>(f)</td>
<td>#</td>
<td>(\leftarrow)</td>
<td>#</td>
<td>#</td>
<td>(\leftarrow)</td>
</tr>
</tbody>
</table>
Alpha Algorithm in Action (2001)

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>#</td>
<td>→</td>
<td>#</td>
<td>→</td>
<td>#</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>←</td>
<td>#</td>
<td>→</td>
<td>←</td>
<td>→</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>#</td>
<td>←</td>
<td>#</td>
<td>→</td>
<td>#</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>#</td>
<td>→</td>
<td>←</td>
<td>#</td>
<td></td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>←</td>
<td>#</td>
<td>#</td>
<td>→</td>
<td></td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>#</td>
<td>←</td>
<td>#</td>
<td>←</td>
<td>#</td>
<td></td>
</tr>
</tbody>
</table>

$T_L = \{a, b, c, d, e, f\}$
$T_I = \{a\}$
$T_f = \{f\}$
$X_L = \{(\{a\}, \{b\}), (\{a\}, \{e\}), (\{b\}, \{c\}), (\{b\}, \{f\}), (\{c\}, \{d\}), (\{d\}, \{b\}), (\{e\}, \{f\}), (\{a, d\}, \{b\}), (\{b\}, \{c, f\})\}$
$Y_L = \{(\{a\}, \{e\}), (\{c\}, \{d\}), (\{e\}, \{f\}), (\{a, d\}, \{b\}), (\{b\}, \{c, f\})\}$
$P_L = \{P_{(i, a)}, P_{(b, c)}, P_{(d, e)}, P_{(b, c, f)}\}$
$F_L = \{(a, P_{(i, a)}), (P_{(b, c)}, e), (c, P_{(b, c, f)}), (P_{(i, d)}, d), (e, P_{(i, c, f)}), (P_{(i, c, f)}), (a, P_{(i, a)}), (d, P_{(i, c, f)}), (P_{(i, a)}), (b, P_{(i, a)}), (P_{(i, a, c, f)}), (P_{(i, a, c, f)})\}$
$\alpha(I) = (P_L, T_L, F_L)$

$i_L \bullet a \xrightarrow{P_{(a, d), (b)}} b \xrightarrow{P_{(a, e), (b)}} e \xrightarrow{P_{(a, f), (d)}} c \xrightarrow{P_{(c, d), (e)}} d \xrightarrow{P_{(c, e), (f)}} f \xrightarrow{P_{(c, f), (d)}} o_L$
Problems Alpha Algorithm and the Like

- Inability to distinguish frequent from infrequent behavior 😞
- May yield unsound process models 😞
- Representational bias that is too narrow 😞

\[ \langle a, b, c \rangle^{20}, \langle a, c \rangle^{30} \]
Problems Alpha Algorithm and the Like

- Inability to distinguish frequent from infrequent behavior 😞
- May yield unsound process models 😞
- Representational bias that is too narrow 😞

\[
[\langle a, b, c\rangle^{20}, \langle a, c\rangle^{30}]
\]

- Able to filter out infrequent behavior 😊
- Able to discover inclusive splits/joins and skips 😊
- Models may be unsound 😞
Inductive Miner  
(2013-2019 Sander Leemans)

- Able to filter out infrequent behavior 😊
- Able to discover inclusive splits/joins and skips 😊
- Models are guaranteed to be sound 😊

- All paths and all activities
- Only the most frequent paths and activities
Not a solved problem!

• Many other approaches (state-based regions, language-based regions, split miner, etc.).
• Human is still able to outperform algorithms.
• Representational bias is still problematic (e.g., multiple instances, label splitting, and non-local dependencies).
• Performance issues for more advanced approaches.
• Most commercial tools still use simple DFGs 😞
Progress in Conformance Checking

(Disclaimer: not intended to be complete, but to illustrate the steps towards maturity.)
Naïve approaches based on comparing process graphs / footprint matrices (2003)

- Very sensitive to the abstraction used 😐
- Very different behaviors can have similar graphs/matrices 😐
- Similar behaviors can lead to different graphs/matrices 😐
- No reliable diagnostics and not linked directly to events/cases 😐

Log represented by a DFG or footprint matrix

Process model is a graph and can be used to create a footprint matrix
Token-based replay (Anne Rozinat 2005-2010, Alessandro Berti 2019)

Fitness example = 0.9102382
Token-based replay  (Anne Rozinat 2005-2010, Alessandro Berti 2019)

Fitness = 0.9102382

220 times the activity happened when it was impossible (according to the model)
492 times the activity did not happen when it was supposed to happen

Addresses all the problems mentioned for the naïve approaches 😊

Does not provide valid paths through the model 😞

Diagnostics tend to be misleading (token flooding) 😞
Alignments (2010-2013 Arya Adriansyah)

- 1849 synchronous moves
- 805 move in model only
- 2651 synchronous moves
- 3 moves in model only
- Many moves on log
- Fitness = 0.7964

Addresses the problems of token-based replay and yields paths through the model 😊

Diagnostics tend to be rather arbitrary 😞

Still time consuming 😞
Not a solved problem!

- Fitness is well understood, but there are multiple measures possible.
- There is no consensus on the other three quality dimensions (precision, simplicity, and generalization).
- Alignments are very powerful but also still time-consuming.
- Less than half of the commercial tools support conformance checking (and rarely alignments)
- Still a gap between BPMN modeling tools and process mining
Two examples covering 20 years of research

- Alpha algorithm
- Heuristic miner
- Inductive miner

process discovery

- Matrix/graph comparison
- Token-based replay
- Alignments

conformance checking

1999 2019
Process Mining Books
Open-Source Process Mining Software (far from complete)

ProM 1.1-5.2 (>286 plugins) (2004-2010)

ProM 6.X (>1500 plugins) (2010-now)

MiMo (2001-2002)
EMiT (2002-2004)
Thumb (2002-2004)
InWolvE (2000-2004)

convergence

RapidProM (2014-now)
Apromore (2015*-now)
bupaR (2016-now)

PM4Py (2018-now)

divergence
Futura Process Intelligence
(Peter van den Brand, started 2007, Gartner Cool Vendor 2009)
Disco - Lowering the threshold to do process mining

(Anne Rozinat & Christian Günther, Fluxicon started in 2009)
Over 30 process mining vendors today

Celonis
ProcessGOLD
Disco by Fluxicon
QPR
Timeline Pi
my invenio
Minit
Paf Now
EverFlow
Puzzle Data
Mehrwerk
Stereo Logic
Logpicker
Signavio
Lana Labs
OnBase by Hyland

Etc.
Trend: From Ad-Hoc to Continuous PM (e.g., Celonis and ProcessGold)

From data scientists to process managers and from insights to actions.
Terminology: Avoid the temptation ...
(process mining is broad, but also specific: event data & process models)

- **Process mining**
- Workflow mining (my mistake)
- Automated business process discovery (Gartner)
- Business process intelligence
- Process intelligence
- Business intelligence
- Process analytics
- Business process analytics
- …
Challenges

“towards better processes”
Process Mining Manifesto (2011)

Six guiding principles and 11 challenges:

1. Finding, Merging, and Cleaning Event Data
2. Dealing with Complex Event Logs Having Diverse Characteristics
3. Creating Representative Benchmarks
4. Dealing with Concept Drift
5. Improving the Representational Bias Used for Process Discovery
6. Balancing Between Quality Criteria
7. Cross-Organizational Mining
8. Providing Operational Support
9. Combining Process Mining With Other Types of Analysis
10. Improving Usability for Non-Experts
11. Improving Understandability for Non-Experts
Six guiding principles and 11 challenges:

1. Finding, Merging, and Cleaning Event Data
2. Dealing with Complex Event Logs Having Diverse Characteristics
3. Creating Representative Benchmarks
4. Dealing with Concept Drift
5. Improving the Representational Bias Used for Process Discovery
6. Balancing Between Quality Criteria
7. Cross-Organizational Mining
8. Providing Operational Support
9. Combining Process Mining With Other Types of Analysis
10. Improving Usability for Non-Experts
11. Improving Understandability for Non-Experts

* green = addressed to some degree, red=still open, orange = less imported as expected
“Old” Challenges
From **backward** looking to **forward** looking
From backward to forward looking

historic
current
tomorrow
From backward to forward looking

quality of predictions needs to improve

predictions need to be turned into actions
Context matters!
Better integration of process discovery, conformance checking, and process modeling
Hybrid process models are needed showing both the “sure/formal” and the “unsure/informal” parts.
Process modeling with “haptic feedback” based on data

Interactive process modeling still outperforms fully automated discovery techniques.
Better support for comparative process mining
What are the differences?
(different periods, departments, customer groups)

Recall comment on just comparing graphs.
What are the differences?
(different periods, departments, customer groups)

More difficult are differences in frequencies, times, etc.
Considering causality and fairness to suggest more meaningful improvements
Process mining can be used to identify compliance and performance problems

- mandatory activity is skipped
- activity is performed too late
- wrong order
- unauthorized resource
- bottlenecks and delays
- unnecessary rework
- waste
- overproduction

If Wil works on a case, check activities are more likely to be skipped.

In this department, checks are performed after the legal deadline.

Cases for this supplier tend to have many price changes.

There are often delays in the back office on Fridays.
Root case analysis?

- Flow time $\rightarrow$ deviations $\rightarrow$ crime rate $\rightarrow$ ice cream sales
Correlation ≠ Causality
Fairness

• Don’t blame overloaded resources for causing bottlenecks.
• Don’t blame the most experienced resources taking the most difficult cases for deviating.
• Discrimination-aware data/process mining aims to avoid such errors.
• Trade-off between fairness and accuracy.
Making process mining results actionable (e.g., the RPA connection)
diagnosis -> operational support (alerting, prediction, recommendations) -> action

process reengineering
Process-mining-informed automation (RPA)

- Traditional process automation
- Robotic Process Automation (RPA) candidates
- Work that can only be done by humans

Case frequency (number of similar cases in a given period)

Different types of cases (sorted by frequency)
Ensuring confidentiality in process mining
<table>
<thead>
<tr>
<th>Name</th>
<th>Activity</th>
<th>Date</th>
<th>Time</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>WvdA</td>
<td>check-in</td>
<td>30-8-19</td>
<td>8:20</td>
<td>CGN</td>
<td>VIE</td>
</tr>
<tr>
<td>WvdA</td>
<td>drop-bag</td>
<td>1-9-19</td>
<td>9:20</td>
<td>CGN</td>
<td>VIE</td>
</tr>
<tr>
<td>WvdA</td>
<td>board</td>
<td>1-9-19</td>
<td>9:40</td>
<td>CGN</td>
<td>VIE</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>WvdA</td>
<td>check-in</td>
<td>15-9-19</td>
<td>22:58</td>
<td>DUS</td>
<td>OSL</td>
</tr>
<tr>
<td>WvdA</td>
<td>drop-bag</td>
<td>17-9-19</td>
<td>18:38</td>
<td>DUS</td>
<td>OSL</td>
</tr>
<tr>
<td>WvdA</td>
<td>board</td>
<td>17-9-19</td>
<td>19:10</td>
<td>DUS</td>
<td>OSL</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>WvdA</td>
<td>check-in</td>
<td>23-9-19</td>
<td>22:46</td>
<td>DUS</td>
<td>NRT</td>
</tr>
<tr>
<td>WvdA</td>
<td>drop-bag</td>
<td>24-9-19</td>
<td>16:18</td>
<td>DUS</td>
<td>NRT</td>
</tr>
<tr>
<td>WvdA</td>
<td>board</td>
<td>24-9-19</td>
<td>17:12</td>
<td>DUS</td>
<td>NRT</td>
</tr>
</tbody>
</table>

Event data are very sensitive!
<table>
<thead>
<tr>
<th>Name</th>
<th>Activity</th>
<th>Date</th>
<th>Time</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>#6@7</td>
<td>check-in</td>
<td>30-8-19</td>
<td>8:20</td>
<td>CGN</td>
<td>VIE</td>
</tr>
<tr>
<td>#6@7</td>
<td>drop-bag</td>
<td>1-9-19</td>
<td>9:20</td>
<td>CGN</td>
<td>VIE</td>
</tr>
<tr>
<td>#6@7</td>
<td>board</td>
<td>1-9-19</td>
<td>9:40</td>
<td>CGN</td>
<td>VIE</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>#6@7</td>
<td>check-in</td>
<td>15-9-19</td>
<td>22:58</td>
<td>DUS</td>
<td>OSL</td>
</tr>
<tr>
<td>#6@7</td>
<td>drop-bag</td>
<td>17-9-19</td>
<td>18:38</td>
<td>DUS</td>
<td>OSL</td>
</tr>
<tr>
<td>#6@7</td>
<td>board</td>
<td>17-9-19</td>
<td>19:10</td>
<td>DUS</td>
<td>OSL</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>#6@7</td>
<td>check-in</td>
<td>23-9-19</td>
<td>22:46</td>
<td>DUS</td>
<td>NRT</td>
</tr>
<tr>
<td>#6@7</td>
<td>drop-bag</td>
<td>24-9-19</td>
<td>16:18</td>
<td>DUS</td>
<td>NRT</td>
</tr>
<tr>
<td>#6@7</td>
<td>board</td>
<td>24-9-19</td>
<td>17:12</td>
<td>DUS</td>
<td>NRT</td>
</tr>
<tr>
<td>Name</td>
<td>Activity</td>
<td>Date</td>
<td>Time</td>
<td>From</td>
<td>To</td>
</tr>
<tr>
<td>------</td>
<td>-----------</td>
<td>--------</td>
<td>------</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td>#6@7</td>
<td>check-in</td>
<td>30-8-19</td>
<td></td>
<td>VIE</td>
<td></td>
</tr>
<tr>
<td>#6@7</td>
<td>drop-bag</td>
<td>1-9-19</td>
<td></td>
<td>VIE</td>
<td></td>
</tr>
<tr>
<td>#6@7</td>
<td>board</td>
<td>1-9-19</td>
<td></td>
<td>VIE</td>
<td></td>
</tr>
<tr>
<td>#6@7</td>
<td>check-in</td>
<td>15-9-19</td>
<td></td>
<td>OSL</td>
<td></td>
</tr>
<tr>
<td>#6@7</td>
<td>drop-bag</td>
<td>17-9-19</td>
<td></td>
<td>OSL</td>
<td></td>
</tr>
<tr>
<td>#6@7</td>
<td>board</td>
<td>17-9-19</td>
<td></td>
<td>OSL</td>
<td></td>
</tr>
<tr>
<td>#6@7</td>
<td>check-in</td>
<td>23-9-19</td>
<td></td>
<td>NRT</td>
<td></td>
</tr>
<tr>
<td>#6@7</td>
<td>drop-bag</td>
<td>24-9-19</td>
<td></td>
<td>NRT</td>
<td></td>
</tr>
<tr>
<td>#6@7</td>
<td>board</td>
<td>24-9-19</td>
<td></td>
<td>NRT</td>
<td></td>
</tr>
</tbody>
</table>
Minimal information needed to get the result, e.g. a DFG.
Many open challenges and opportunities …

Therefore, we are looking for PhDs and Postdocs that would like make a difference!
Conclusion

“recommendations”
This is something completely NEW, something people said wasn’t possible, that the data wasn’t there to allow systems that really could map out a process; they were wrong. Data is now everywhere; it is accessible, there is an abundance of data and it can provide you with insights you could never find just in interviews. The goal is to get away from workflow systems that are divorced from reality and from how people really work. Today’s tools oversimplify reality when what you need is a view as close to the real world as possible.

Adoption is slow … , but the speed is accelerating!!
Free Advice
For process mining users …

• You get the tools **you deserve**. If you do not ask for it, you will not get it.

• You need: Domain/business expertise, knowledge of the information systems, and data science skills (i.e., different people). If one element is **missing**, you will **fail**.

• You can only make a business case if you are able to apply process mining **continuously** (when 80% is spent on data extraction, you want to use this multiple times).

• But, the business case is typically much **better** than most BPM, process modeling, Six Sigma, etc. initiatives.
For process mining vendors …

• Try to better adopt ideas from academia and provide feedback (even when you still need to do the engineering).
• Going beyond DFGs is unavoidable if you want to stay in business.
• Advanced features will only be used if they are usable and made scalable (adding conformance half-heartedly will not help your customers).
• Riding the AI/ML wave may backfire.
• Initiatives like XES will not help you sell more licenses, but help to establish process mining as a discipline.
For academics …

- **Good news:** Many open problems!
- **But:**
  - Get your **own case studies** ("eat your own dog food").
  - Avoid working on techniques that do slightly better on **selected data sets** using well selected criteria (even when these papers still get accepted).
  - Work on **original problems** or use truly original approaches (even when your paper gets rejected at first).
  - Understand **why** vendors are still using simple DFGs.
  - In the end, we would like to **improve processes**, right?