Overview BPM Analysis Techniques & Simulation

prof.dr.ir. Wil van der Aalst
Overview Analysis

- **Process mining**
- **Simulation**
- **Gaming**
- **Invariants**
- **Siphons/traps**
- **Reachability graph**
- **Coverability graph**
- **Markov chain analysis**
- **Queueing networks**
- **Validation**
- **Verification**
- **Performance analysis**

**Design-time analysis**

**Run-time analysis**

- e.g. process models represented in BPMN, BPEL, EPCs, Petri nets, UML AD, etc. or other types of models such as social networks, organizational networks, decision trees, etc.
- e.g., dedicated formats such as IBM's Common Event Infrastructure (CEI) and MXML or proprietary formats stored in flat files or database tables.
- e.g., systems like WebSphere, Oracle, TIBCO/Staffware, SAP, FLOWer, etc.
Let’s Play
Play-Out

process model

event log
Play-Out (Classical use of models)
Play-Out: From Model to Behavior (traces, runs, state spaces, …)

- Simulation (random walks visiting parts of the state space)
- Verification/model checking (logical statements about behavior)
- Enactment (WFM/BPM engine).
Play-In

event log → process model
Play-In

A B C D  A E D  A E D
A C B D  A B C D  A C B D
A C B D  A E D  A C B D

start A B C D
p1 B
p2 E
p3 D
p4 end

A E D  A E D
A C B D
A B C D
A C B D
Example Process Discovery
(Vestia, Dutch housing agency, 208 cases, 5987 events)
Example Process Discovery
(ASML, test process lithography systems, 154966 events)
Example Process Discovery
(AMC, 627 gynecological oncology patients, 24331 events)
Replay

- event log
- process model

- extended model showing times, frequencies, etc.
- diagnostics
- predictions
- recommendations
Replay

A B C D

start

p1

E

p2

p3

p4

dep

end
Replay

A E D

start

p1

B

p2

C

p3

p4

end
Replay can detect problems

Problem! token left behind
Problem! missing token
Replay can extract timing information

$A^5B^8C^9D^{13}$
Performance Analysis Using Replay
(WOZ objections Dutch municipality, 745 objections, 9583 event, f= 0.988)
Verification
Generic Property: Soundness

1. **Option to complete**: It is possible to reach the marking with just a token in place end.
2. **Proper completion**: If there is a token in end, the rest of the net is empty.
3. **No dead parts**: For each transition there is at least one part enabling it.
Corrected model
□(book car → ◊time-out) is true
□(select car → ◊too late) is not true
◊too late) → (◊supply car) is true
◊supply car) → (◊too late) is not true

<table>
<thead>
<tr>
<th>name</th>
<th>notation</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>eventually</td>
<td>◊F</td>
<td>$F$ has to hold eventually, e.g., $[F,A,B,C,D,E]$, $[A,B,C,F,D,E]$, $[ABFCDFEF]$, etc.</td>
</tr>
<tr>
<td>always</td>
<td>□F</td>
<td>$F$ has to always hold, e.g., $[F,F,F,F,F,F]$.</td>
</tr>
<tr>
<td>until</td>
<td>$F □ G$</td>
<td>$G$ holds at the current state or at some future state, and $F$ has to hold until $G$ holds. When $G$ holds $F$ does not have to hold any more. Examples are $[G,A,B,C,D,E]$, $[F,G,A,B,C,D,E]$, $[F,F,F,F,G,A,B,C,D,F,G]$, etc.</td>
</tr>
</tbody>
</table>
Simulation
Many Jokes About Simulation ...
...but still one of the most widely used analysis techniques.
Other simulation “models”
Discrete Event Simulation (DES)

- Basic idea: Given an executable model, repeatedly execute the model and compare the results.
- Close to workflow models but there is a need to model the environment!
- Stochastics and other abstractions are used to "model the unknown".
- Steady-state versus transient analysis.
- From a technical point of view just a "walk" in the reachability graph.
- By making many "short walks" (in case of transient behavior) or one or more "long walks" (in case of steady-state behavior), it is possible to make reliable statements about properties/ performance indicators.
Transient versus steady-state analysis

Transient analysis

Steady-state analysis (I)
In the remainder, we only consider steady-state analysis.
BPM|one and Protos provide the same simulation functionality. On the sheets I will use Protos rather than BPM|one.
Subruns in BPM|one and Protos

total length = (x+z)*y
subruns used for analysis = x
Understanding subruns

"low level" measurements

aggregation per subrun
(average, min, max, variance, etc.)

<table>
<thead>
<tr>
<th>subruns</th>
<th>average</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>5.7</td>
<td>0.21</td>
</tr>
</tbody>
</table>

confidence = 0.9
confidence interval = 

\[5.7 - 0.117, 5.7 + 0.117\] = [5.58, 5.82]
Activity/task parameters

- **Number of persons needed** (The number of people needed to carry out the activity one time.)
- **Costs** (The fixed costs of carrying out the Activity.)
- **Frequency** (The number of times that the activity is carried out relative to other activities sharing an input place in a deferred choice construct.)
- **Processing time** (The time needed to carry out the activity one time.)
- **Priority** (To indicate that certain activities have precedence over others when queueing for a resource.)
Note that frequencies appear at two places: (1) for resolving deferred choices and (2) for XOR-splits.
Replacing details by stochastics

- Select right level of detail: there may be thousands of customers, products, etc.
- You cannot look into the heads of people or under the hood of applications!
- Details are replaced by probability distributions for durations and routing.
- Random generators are used to "implement" these probability distributions.

Most random generators generate a series of pseudo-random numbers $\frac{X_i}{m}$ according to the formula:

$$X_n = (aX_{n-1} + b) \mod m$$

In the specific example given:

$$X_n = 16807X_{n-1} \mod (2^{31} - 1)$$

$$X_n = 3125X_{n-1} \mod (2^{35} - 31)$$
Dice

\[ P[X = a] = \begin{cases} \frac{1}{6} & \text{if } a \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{else} \end{cases} \]

\[
\mathbb{E}[X] = \frac{(1+2+3+4+5+6)}{6} = \frac{21}{6} = 3.5
\]

\[
\text{variance} \quad \text{Var}[X] = \frac{(1-3.5)^2 + (2-3.5)^2 + \ldots + (6-3.5)^2}{6} = \frac{2.916}{6} = 1.7^2
\]
### Standard discrete distributions

<table>
<thead>
<tr>
<th>distribution</th>
<th>domain</th>
<th>$\Pr[X = k]$</th>
<th>$\mathbb{E}[X]$</th>
<th>$\text{Var}[X]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernoulli</td>
<td>$k \in {0, 1}$</td>
<td>$\begin{cases} 1 - p &amp; k = 0 \ p &amp; k = 1 \end{cases}$</td>
<td>$p$</td>
<td>$p(1 - p)$</td>
</tr>
<tr>
<td>homogeneous</td>
<td>$k \in {a, \ldots, b}$</td>
<td>$\frac{1}{(b-a)+1}$</td>
<td>$\frac{a+b}{2}$</td>
<td>$\frac{(b-a)((b-a)+2)}{12}$</td>
</tr>
<tr>
<td>binomial</td>
<td>$k \in {0, 1, \ldots, n}$</td>
<td>$\binom{n}{k} p^k (1 - p)^{n-k}$</td>
<td>$np$</td>
<td>$np(1 - p)$</td>
</tr>
<tr>
<td>geometric</td>
<td>$k \in {1, 2, \ldots}$</td>
<td>$(1 - p)^{k-1} p$</td>
<td>$\frac{1}{p}$</td>
<td>$\frac{1-p}{p^2}$</td>
</tr>
<tr>
<td>Poisson</td>
<td>$k \in {0, 1, \ldots}$</td>
<td>$\frac{\lambda^k}{k!} e^{-\lambda}$</td>
<td>$\lambda$</td>
<td>$\lambda$</td>
</tr>
</tbody>
</table>
## Standard continuous distributions

<table>
<thead>
<tr>
<th>distribution</th>
<th>domain</th>
<th>$f_X(x)$</th>
<th>$\mathbb{E}[X]$</th>
<th>$\text{Var}[X]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>uniform $a &lt; b$</td>
<td>$a \leq x \leq b$</td>
<td>$\frac{1}{b-a}$</td>
<td>$\frac{a+b}{2}$</td>
<td>$\frac{(b-a)^2}{12}$</td>
</tr>
<tr>
<td>exponential $\lambda &gt; 0$</td>
<td>$x \geq 0$</td>
<td>$\lambda e^{-\lambda x}$</td>
<td>$\frac{1}{\lambda}$</td>
<td>$\frac{1}{\lambda^2}$</td>
</tr>
<tr>
<td>normal $\mu \in \mathbb{R}$, $\sigma &gt; 0$</td>
<td>$x \in \mathbb{R}$</td>
<td>$\frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$</td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>gamma $r, \lambda &gt; 0$</td>
<td>$x &gt; 0$</td>
<td>$\frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)}$</td>
<td>$\frac{r}{\lambda}$</td>
<td>$\frac{r}{\lambda^2}$</td>
</tr>
<tr>
<td>Erlang $\lambda &gt; 0$, $r \in {1, 2, \ldots}$</td>
<td>$x &gt; 0$</td>
<td>$\frac{\lambda^r (x-1)! x^{r-1} e^{-\lambda x}}{(r-1)!}$</td>
<td>$\frac{r}{\lambda}$</td>
<td>$\frac{r}{\lambda^2}$</td>
</tr>
<tr>
<td>$\chi^2$ $v \in {1, 2, \ldots}$</td>
<td>$x &gt; 0$</td>
<td>see gamma $r = \frac{v}{2}$ and $\lambda = \frac{1}{2}$</td>
<td>$v$</td>
<td>$2v$</td>
</tr>
<tr>
<td>beta $a &lt; b$ $r, s &gt; 0$</td>
<td>$a \leq x \leq b$</td>
<td>$\frac{1}{b-a} \frac{\Gamma(r+s)}{r \Gamma(s)} \left( \frac{x-a}{b-a} \right)^{r-1} \left( \frac{b-x}{b-a} \right)^{s-1}$</td>
<td>$a + (b-a) \frac{r}{r+s}$</td>
<td>$\frac{rs(b-a)^2}{(r+s)^2(r+s+1)}$</td>
</tr>
</tbody>
</table>
Negative exponential distribution \((\lambda = 0.5, 1.0, 2.0)\)

probability density function (pdf)
Normal distribution
Gamma distribution
Beta distribution
Example

24 arrivals per hour

2 resources, average service time of 4 minutes

2 resources, average service time of 4 minutes

initial design: sequence

Both interarrival times and service times have neg. exponential distributions
The utilization rate is approximately 80%. The service times are approximately 4+4 minutes. The flow time is approximately 22 minutes.
24 arrivals per hour

2 resources, average service time of 4 minutes

redesign: parallel
utilization of approx. 80%

service times of approx. 4+4min.

flow time of approx. 15 min.

+/-22 => +/- 15 min. !!!!
24 arrivals per hour

4 resources, average service time of 7 minutes

redesign: compose
<table>
<thead>
<tr>
<th>Roles</th>
<th>Utilization Rate</th>
<th>Activities</th>
<th>Queue Time</th>
<th>Work Time</th>
<th>Wait Time</th>
<th>Wait+Queue Time</th>
<th>Lead Time</th>
<th>Work Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Lower 90%</td>
<td>Mean Lower 90%</td>
<td>Mean Lower 90%</td>
<td>Mean Lower 90%</td>
<td>Mean Lower 90%</td>
<td>Mean Lower 90%</td>
<td>Mean Lower 90%</td>
<td>Mean Lower 90%</td>
</tr>
<tr>
<td>&lt;&lt;No role&gt;&gt;</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>role12</td>
<td>0.711539562</td>
<td>0.703584111</td>
<td>0.719520712</td>
<td>0.698894727</td>
<td>0.724184396</td>
<td>0.7095813564</td>
<td>0.7035127572</td>
<td>0.7156499756</td>
</tr>
</tbody>
</table>

- Utilization of approx. 70%.
- Service times of approx. 7 min.
- Flow time of approx. 9.5 min.

+-22 => +/- 9.5 min. !!!!
24 arrivals per hour

4 resources, average service time of 4 minutes

redesign: pool
The utilization rate is approximately 80%.

Service times are approximately 4+4 minutes.

The flow time is approximately 13 minutes.

The variation is +/- 22, which translates to approximately +/- 13 minutes.
5.6 5.6 6.2 5.5 5.7 5.9 5.4 5.8 5.6 5.7 5.7

is not the same as

4.6 6.6 3.2 8.5 1.7 9.9 4.4 6.8 4.6 6.7 5.7

although the average over the subrun results is the same (5.7)
Central limit theorem

For a set $X_1, X_2, \ldots, X_n$ of independent uniformly distributed random variables with expectation $\mu$ and variance $\sigma^2$, the random variable

$$\frac{(X_1 + X_2 + \ldots + X_n) - n\mu}{\sigma \sqrt{n}}$$

converges for $n \to \infty$ to a standard normal distribution.

Hence, the result of one subrun is normally distributed if it is based on averages and the average over many subrums is also normally distributed.
The situation with many subruns (>30)

- Subruns need to be independent!
- Result per subrun does not need to be normally distributed.
- However, the average over the subruns can be considered to be normally distributed by Central Limit Theorem.
- \((1-\alpha)\) confidence interval for \(n\) subruns:

\[
\left[ \bar{x} - \frac{s}{\sqrt{n}} z\left(\frac{\alpha}{2}\right), \bar{x} + \frac{s}{\sqrt{n}} z\left(\frac{\alpha}{2}\right) \right]
\]

\[
\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}
\]

\[
s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}
\]

<table>
<thead>
<tr>
<th>(x)</th>
<th>(z(x))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>3.09</td>
</tr>
<tr>
<td>0.005</td>
<td>2.58</td>
</tr>
<tr>
<td>0.010</td>
<td>2.33</td>
</tr>
<tr>
<td>0.025</td>
<td>1.96</td>
</tr>
<tr>
<td>0.050</td>
<td>1.64</td>
</tr>
<tr>
<td>0.100</td>
<td>1.28</td>
</tr>
</tbody>
</table>
Sample mean and variance

- **Sample mean** is best estimator for expected value.
- **Sample variance** is best estimated by the unbiased estimator of the population variance (divide by n-1 rather than n since not the whole population is seen).

\[
\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}
\]

\[
s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}
\]
The sample mean is 0.9408 and the sample variance is 0.000617. So, all the data needed to set up a \((1 - \alpha)\)-confidence interval are known: \(n = 30\), \(\bar{x} = 0.9408\), \(s^2 = 0.000617\) and therefore \(s = 0.02485\). If we take \(\alpha\) equal to 0.010 we will find the following confidence interval:

\[
\left[ 0.9408 - \frac{0.02485}{\sqrt{30}} \cdot z\left(\frac{0.010}{2}\right), 0.9408 + \frac{0.02485}{\sqrt{30}} \cdot z\left(\frac{0.010}{2}\right) \right]
\]

\[
[0.9291, 0.9525]
\]
The situation with fewer subruns ($<30$)

- Subruns need to be independent!
- Result per subrun need to be normally distributed!
- $(1-\alpha)$ confidence interval for $n$ subruns:

\[
\left[ \bar{x} - \frac{s}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right), \bar{x} + \frac{s}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right) \right]
\]

$t_v(x)$ is the critical value of a Student’s $t$-distribution.

\[
\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}
\]

\[
s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}
\]
Critical values for a Student's t-distribution with $v$ degrees of freedom

<table>
<thead>
<tr>
<th>$t_v(x)$</th>
<th>$x =$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.100</td>
</tr>
<tr>
<td>$v = 1$</td>
<td>3.08</td>
</tr>
<tr>
<td>2</td>
<td>1.89</td>
</tr>
<tr>
<td>3</td>
<td>1.64</td>
</tr>
<tr>
<td>4</td>
<td>1.53</td>
</tr>
<tr>
<td>5</td>
<td>1.48</td>
</tr>
<tr>
<td>6</td>
<td>1.44</td>
</tr>
<tr>
<td>7</td>
<td>1.41</td>
</tr>
<tr>
<td>8</td>
<td>1.40</td>
</tr>
<tr>
<td>9</td>
<td>1.38</td>
</tr>
<tr>
<td>10</td>
<td>1.37</td>
</tr>
<tr>
<td>15</td>
<td>1.34</td>
</tr>
<tr>
<td>20</td>
<td>1.33</td>
</tr>
<tr>
<td>25</td>
<td>1.32</td>
</tr>
<tr>
<td>50</td>
<td>1.30</td>
</tr>
<tr>
<td>100</td>
<td>1.29</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$x$</th>
<th>$z(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>3.09</td>
</tr>
<tr>
<td>0.005</td>
<td>2.58</td>
</tr>
<tr>
<td>0.010</td>
<td>2.33</td>
</tr>
<tr>
<td>0.025</td>
<td>1.96</td>
</tr>
<tr>
<td>0.050</td>
<td>1.64</td>
</tr>
<tr>
<td>0.100</td>
<td>1.28</td>
</tr>
<tr>
<td>subrun number</td>
<td>average load factor</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>1</td>
<td>0.914</td>
</tr>
<tr>
<td>2</td>
<td>0.964</td>
</tr>
<tr>
<td>3</td>
<td>0.934</td>
</tr>
<tr>
<td>4</td>
<td>0.978</td>
</tr>
<tr>
<td>5</td>
<td>0.912</td>
</tr>
<tr>
<td>6</td>
<td>0.956</td>
</tr>
<tr>
<td>7</td>
<td>0.958</td>
</tr>
<tr>
<td>8</td>
<td>0.934</td>
</tr>
<tr>
<td>9</td>
<td>0.978</td>
</tr>
<tr>
<td>10</td>
<td>0.976</td>
</tr>
</tbody>
</table>

$$
\left\lfloor 0.9408 - \frac{0.02485}{\sqrt{30}} \cdot t_{29} \left( \frac{0.100}{2} \right),
0.9408 + \frac{0.02485}{\sqrt{30}} \cdot t_{29} \left( \frac{0.100}{2} \right) \right\rfloor
$$

As $t_{29}(0.050) = 1.699$ this yields the interval $[0.9331, 0.9485]$. 
So make sure ...

- Subruns should be independent!
- If the utilization of some resources is high, longer subruns are needed to satisfy this requirement.
- If only a few runs, only calculate confidence intervals over normally distributed subrun results, i.e., averages are OK but utilizations not!
You are expected to ...

- understand the concept of simulation and the need for abstraction/stochastics,
- understand the way confidence intervals are calculated and the requirements that need to be satisfied, and
- be able to build simulation models with BPM|one and interpret the results.
Recommended reading :-)
Another example
(taken from http://avoinelama.fi/hingo/kirjoituksesta/misleadingvisualizations.html)

Chart 18: 3-dimensional moneybags

220   251   370   1506  409

Chart 17: True 2-dimensional moneybags

220   251   370   1506  409

Chart 16: Moneybags that grow in both directions

220   251   370   1506  409

Chart 15: Soneras result drawn with moneybags as bars

220   251   370   1506  409
• **M/M/1 queue:** arrival rate $\lambda$, service rate $\mu$, utilization $\rho = \frac{\lambda}{\mu}$.
• **Flow time** = $\frac{1}{(\mu - \lambda)}$, # in system = $\frac{\rho}{1 - \rho}$

![Graph showing flow time and utilization relationship]
Problems when modeling human resources

- People do not work at a constant speed, cf. Yerkes-Dodson Law of Arousal, coffee breaks, weather, etc.
- People are involved in multiple processes. Hence, different processes/tasks compete for attention and availability is “fluid”.
- People tend to work part-time and in batches. Different working patterns: every Friday, when the pile is too large, …
- Priorities are difficult to model. Competing processes/resources have undefined precedence rules.
- Processes may change depending on context. Things are skipped or done in a sloppy manner when …
Yerkes-Dodson Law of Arousal

Stress Performance Connection

- High Stress: Disorganization
- Medium Stress: Anxiety
- Low Stress: Sleep, Alertness
- Optimal Performance
5*0.2 ≠ 1
Classical simulation assumptions

• A resource is:
  • eager to start working,
  • dedicated to a single process,
  • works at a constant speed,
  • does not work in batches,
  • does not have coffee breaks,
  • etc.

• Do you know this person?
Avoid modeling the world in a detailed manner

Goal: Characterize resource availability with just a few parameters

Data explosion

From Bits to Zettabytes

A “bit” is the smallest unit of information possible. One bit has two possible values: 1 (on) and 0 (off). A “byte” is composed of 8 bits and can represent $2^8 = 256$ values. To talk about larger amounts of data, multiples of 1000 are used: 1 Kilobyte (KB) equals 1000 bytes, 1 Megabyte (MB) equals 1000 KB, 1 Gigabyte (GB) equals 1000 MB, 1 Terabyte (TB) equals 1000 GB, 1 Petabyte (PB) equals 1000 TB, 1 Exabyte (EB) equals 1000 PB, and 1 Zettabyte (ZB) equals 1000 EB. Hence, 1 Zettabyte is $10^{21} = 1,000,000,000,000,000,000$ bytes. Note that here we used the International System of Units (SI) set of unit prefixes, also known as SI prefixes, rather than binary prefixes. If we assume binary prefixes, then 1 Kilobyte is $2^{10} = 1024$ bytes, 1 Megabyte is $2^{20} = 1048576$ bytes, and 1 Zettabyte is $2^{70} \approx 1.18 \times 10^{21}$ bytes.
The World's Technological Capacity to Store, Communicate, and Compute Information by Martin Hilbert and Priscila López (DOI 10.1126/science.1200970)

**THE WORLD’S CAPACITY TO STORE INFORMATION**

This chart shows the world’s growth in storage capacity for both analog data (books, newspapers, videotapes, etc.) and digital (CDs, DVDs, computer hard drives, smartphone drives, etc.)

**In gigabytes or estimated equivalent**

<table>
<thead>
<tr>
<th>Year</th>
<th>Analog Storage</th>
<th>Digital Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1986</td>
<td>2.62 billion</td>
<td>0.02 billion</td>
</tr>
<tr>
<td>1993</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**COMPUTING POWER**

In 1986, pocket calculators accounted for much of the world’s data-processing power.

**Percentage of available processing power by device:**

<table>
<thead>
<tr>
<th>Year</th>
<th>Pocket calculators</th>
<th>Personal computers</th>
<th>Video game consoles</th>
<th>Servers, mainframes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1986</td>
<td>41%</td>
<td>33%</td>
<td>9%</td>
<td>17%</td>
</tr>
<tr>
<td>2007</td>
<td>66%</td>
<td>25%</td>
<td>3%</td>
<td>6%</td>
</tr>
</tbody>
</table>

**2007**

**18.86 billion gigabytes**

- Paper, film, audiotape and vinyl: 6.2%
- Analog videotapes: 93.8%
- Other digital media: 0.8%*
- Portable media players, flash drives: 2%
- Portable hard disks: 2.4%
- CDs and minidisks: 6.8%
- Computer servers and mainframe hard disks: 8.9%
- Digital tape: 11.8%
- DVD/Blu-ray: 22.8%

**276.12 billion gigabytes**

*Other includes chip cards, memory cards, floppy disks, mobile phones/PDAs, cameras/camcorders, video games
<table>
<thead>
<tr>
<th>Smoker</th>
<th>Drinker</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short ($91/10$)</td>
<td>Yes</td>
<td>$&lt;81.5$</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>$≥81.5$</td>
</tr>
<tr>
<td>Long ($30/1$)</td>
<td>Yes</td>
<td>$&lt;81.5$</td>
</tr>
<tr>
<td>Long ($150/20$)</td>
<td>Yes</td>
<td>$≥81.5$</td>
</tr>
<tr>
<td>Short ($321/25$)</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

**Process Mining =**

**Data Mining**

**Process Analysis**
Desire lines in process models
BPM Professional
<table>
<thead>
<tr>
<th>How can process mining help?</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Uncover bottlenecks</td>
</tr>
<tr>
<td>• Detect deviations</td>
</tr>
<tr>
<td>• Performance measurement</td>
</tr>
<tr>
<td>• Auditing/compliance</td>
</tr>
<tr>
<td>• Business Process Redesign (BPR)</td>
</tr>
<tr>
<td>• Continuous improvement (Six Sigma)</td>
</tr>
<tr>
<td>• Operational support (e.g., recommendation and prediction)</td>
</tr>
<tr>
<td>• Provide new insights</td>
</tr>
<tr>
<td>• Highlight important problems</td>
</tr>
<tr>
<td>• An organization’s mirror (in two ways)</td>
</tr>
<tr>
<td>• Helps to avoid ICT failures</td>
</tr>
<tr>
<td>• Avoid “management by PowerPoint”</td>
</tr>
<tr>
<td>• From “politics” to “analytics”</td>
</tr>
</tbody>
</table>
Each line corresponds to one of the 528 requests that were handled in the period from 4-1-2009 until 28-2-2010. In total there are 5498 events represented as dots. The mean time needed to handled a case is approximately 25 days.
WMO process  
(Wet Maatschappelijke Ondersteuning)

- WMO refers to the social support act that came into force in The Netherlands on January 1st, 2007.
- The aim of this act is to assist people with disabilities and impairments. Under the act, local authorities are required to give support to those who need it, e.g., household help, providing wheelchairs and scootmobiles, and adaptations to homes.
- There are different processes for the different kinds of help. We focus on the process for handling requests for household help.
- In a period of about one year, 528 requests for household WMO support were received.
- These 528 requests generated 5498 events.
C-net discovered using heuristic miner (1/3)
C-net discovered using heuristic miner (2/3)
C-net discovered using heuristic miner (3/3)
Conformance check WMO process (1/3)
Conformance check WMO process (2/3)

40 Toetsen en beslissen start

27 Retour start

27 Retour complete

40 Toetsen en beslissen complete

50 Verzendingdoeisvervanging start

2 remaining tokens

4 missing and 23 remaining tokens

was executed while not enabled

23 missed tokens

4 missing tokens

was executed while not enabled

40 tokens

10 remaining tokens
The fitness of the discovered process is 0.99521667. Of the 528 cases, 496 cases fit perfectly whereas for 32 cases there are missing or remaining tokens.
Bottleneck analysis WMO process (1/3)
Bottleneck analysis WMO process (2/3)
flow time of approx. 25 days with a standard deviation of approx. 28
Two additional Lasagna processes

RWS ("Rijkswaterstaat") process

WOZ ("Waardering Onroerende Zaken") process
• The Dutch national public works department, called “Rijkswaterstaat” (RWS), has twelve provincial offices. We analyzed the handling of invoices in one of these offices.

• The office employs about 1,000 civil servants and is primarily responsible for the construction and maintenance of the road and water infrastructure in its province.

• To perform its functions, the RWS office subcontracts various parties such as road construction companies, cleaning companies, and environmental bureaus. Also, it purchases services and products to support its construction, maintenance, and administrative activities.
C-net discovered using heuristic miner
Social network constructed based on handovers of work

Each of the 271 nodes corresponds to a civil servant. Two civil servants are connected if one executed an activity causally following an activity executed by the other civil servant.
Social network consisting of civil servants that executed more than 2000 activities in a 9 month period.

The darker arcs indicate the strongest relationships in the social network. Nodes having the same color belong to the same clique.
WOZ process

• Event log containing information about 745 objections against the so-called WOZ ("Waardering Onroerende Zaken") valuation.

• Dutch municipalities need to estimate the value of houses and apartments. The WOZ value is used as a basis for determining the real-estate property tax.

• The higher the WOZ value, the more tax the owner needs to pay. Therefore, there are many objections (i.e., appeals) of citizens that assert that the WOZ value is too high.

• "WOZ process" discovered for another municipality (i.e., different from the one for which we analyzed the WMO process).
The log contains events related to 745 objections against the so-called WOZ valuation. These 745 objections generated 9583 events. There are 13 activities. For 12 of these activities both start and complete events are recorded. Hence, the WF-net has 25 transitions.
Conformance checker:
(fitness is 0.98876214)
Performance analysis

Bottleneck detection: places are colored based on average durations.

Time required to move from one activity to another.

Information on total flow time.
## Resource-activity matrix (four groups discovered)

<table>
<thead>
<tr>
<th>user</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_4$</th>
<th>$a_5$</th>
<th>$a_6$</th>
<th>$a_7$</th>
<th>$a_8$</th>
<th>$a_9$</th>
<th>$a_{10}$</th>
<th>$a_{11}$</th>
<th>$a_{12}$</th>
<th>$a_{13}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>user 1</td>
<td>0</td>
<td>0</td>
<td>51</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user 2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>38</td>
<td>0</td>
<td>69</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user 3</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user 4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user 5</td>
<td>117</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>user 6</td>
<td>172</td>
<td>6</td>
<td>14</td>
<td>0</td>
<td>7</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>48</td>
<td>53</td>
</tr>
<tr>
<td>user 7</td>
<td>1</td>
<td>41</td>
<td>8</td>
<td>14</td>
<td>275</td>
<td>8</td>
<td>8</td>
<td>865</td>
<td>55</td>
<td>180</td>
<td>0</td>
<td>128</td>
<td>5</td>
</tr>
<tr>
<td>user 8</td>
<td>2</td>
<td>868</td>
<td>7</td>
<td>6</td>
<td>105</td>
<td>0</td>
<td>0</td>
<td>79</td>
<td>266</td>
<td>441</td>
<td>0</td>
<td>844</td>
<td>3</td>
</tr>
<tr>
<td>user 9</td>
<td>90</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td>user 10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>899</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1019</td>
</tr>
<tr>
<td>user 11</td>
<td>336</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>18</td>
<td>23</td>
</tr>
<tr>
<td>user 12</td>
<td>1</td>
<td>645</td>
<td>13</td>
<td>21</td>
<td>419</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>217</td>
<td>281</td>
<td>1</td>
<td>334</td>
<td>9</td>
</tr>
<tr>
<td>user 13</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user 14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user 15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>user 16</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user 17</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user 18</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user 19</td>
<td>13</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>user 20</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>258</td>
</tr>
</tbody>
</table>
Example of a Spaghetti process

Spaghetti process describing the diagnosis and treatment of 2765 patients in a Dutch hospital. The process model was constructed based on an event log containing 114,592 events. There are 619 different activities (taking event types into account) executed by 266 different individuals (doctors, nurses, etc.).
Fragment
18 activities of the 619 activities (2.9%)
Another example
(event log of Dutch housing agency)

The event log contains 208 cases that generated 5987 events. There are 74 different activities.
Process models should be treated as electronic maps.
Business process movies
Business information systems should offer “TomTom” functionality

**Recommend:** How to get home ASAP? Take a left turn!

**Detect:** You drive too fast!

**Predict:** When will I be home? At 11:26!
Hundreds of plug-ins available covering the whole process mining spectrum

open-source (L-GPL)

Download from: www.processmining.org
We applied ProM in >100 organizations

- **Municipalities** (e.g., Alkmaar, Heusden, Harderwijk, etc.)
- **Government agencies** (e.g., Rijkswaterstaat, Centraal Justitieel Incasso Bureau, Justice department)
- **Insurance related agencies** (e.g., UWV)
- **Banks** (e.g., ING Bank)
- **Hospitals** (e.g., AMC hospital, Catharina hospital)
- **Multinationals** (e.g., DSM, Deloitte)
- **High-tech system manufacturers and their customers** (e.g., Philips Healthcare, ASML, Ricoh, Thales)
- **Media companies** (e.g. Winkwaves)
- ...
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.
Overview Analysis

- **design-time analysis**
  - validation
  - verification
  - performance analysis

- **run-time analysis**
  - discovery
  - conformance
  - extension

- **(process) model**
  - e.g., process models represented in BPMN, BPEL, EPCs, Petri nets, UML AD, etc. or other types of models such as social networks, organizational networks, decision trees, etc.

- **event logs**
  - e.g., dedicated formats such as IBM’s Common Event Infrastructure (CEI) and MXML or proprietary formats stored in flat files or database tables.

- **(software) system**
  - records events, e.g., messages, transactions, etc.
  - specifies, configures, implements, analyzes
  - supports/controls

- **“world”**
  - business processes
  - people
  - services
  - components
  - organizations

- **event logs**
  - models analyzes
events, e.g., messages, etc.

- e.g., systems like WebSphere, Oracle, TIBCO/Staffware, SAP, FLOWer, etc.
Exercises: BPMS-instruction-7-Simulation.pdf/ppt