Process Mining
Olifantenpaden in Assurance

prof. Wil van der Aalst
Technische Universiteit Eindhoven

www.processmining.org
What Happens in an Internet Minute?

- 20 New victims of identity theft
- 47,000 App downloads
- 61,141 Hours of music
- 204 million Emails sent
- $83,000 In sales
- 20 million Photo views
- 320+ New Twitter accounts
- 3,000 Photo uploads
- 100,000 New tweets
- 639,800 GB of global IP data transferred
- 1,300 New mobile users
- 6 New Wikipedia articles published
- 100+ New LinkedIn accounts
- 277,000 Logins
- 6 million Facebook views
- 2+ million Search queries
- 1.3 million Video views
- 30 Hours of video uploaded

And Future Growth is Staggering

Today, the number of networked devices = the global population
By 2015, the number of networked devices = 2x the global population
In 2015, it would take you 5 years to view all video crossing IP networks each second
Moore's Law

Microprocessor Transistor Counts 1971-2011 & Moore's Law

D=2.03

D=1.56

D=1.92

STORAGE: FROM HIGHWAY ROBBERY TO RUNAWAY BARGAIN

$ per megabyte

1981: Apple goes magnetic with its 5MB hard drive, $1,500.
1985: Iomega appears on the scene with its Bernoulli Box, the box itself costing over $2,000 and each 10MB removable disk priced at $89.
1995: Seagate introduces its 1GB drive for $849.
2000: Hitachi introduces the first terabyte drive, the Deskstar 7K1000, $300.

D=1.92
A simple calculation

- Starting point 2010:
  - Harddisk 1 Terabyte = $10^{12}$ bytes
  - Digital Universe 1.2 Zettabyte = $1.2 \times 10^{21}$ bytes

- Disk needs to grow $2^{30.16} = 1.2 \times 10^9 = 1.2 \times 10^{21}/10^{12}$ times its current size.

- Assuming D=1.56 this takes $30.16 \times 1.56 = 47.05$ years.

- Hence, in 2060 your laptop can contain all of today's digital universe (internet, computer files, transaction logs, movies, photos, music, books, databases, etc.)!
Big Data: Even Dilbert and the "pointy-haired boss" know about it ...

http://dilbert.com/strips/comic/2012-07-29/
Process Mining

- (process) model
- event logs
- software system

models analyzes

process model event logs models analyzes discovery records events, e.g., messages, transactions, etc.
specifies configures implements analyzes supports/controls

- discovery
- conformance
- enhancement
Positioning Process Mining

process model analysis
(simulation, verification, etc.)

process mining

data-oriented analysis
(data mining, machine learning, business intelligence)

performance-oriented questions, problems and solutions

compliance-oriented questions, problems and solutions
Assurance: Part of corporate governance in which a management provides accurate and current information to the stakeholders about the efficiency and effectiveness of its policies and operations, and the status of its compliance with statutory obligations (www.businessdictionary.com).
Let's play!
Play-Out

process model

event log
Play-Out (Classical use of models)

A B C D A E D A E D A E D
A C B D A C B D A C B D
A C B D A E D A C B D
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 E 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
Replay
Replay can detect problems

A C D

Problem! token left behind
Problem! missing token

start

p1

B

E

p3

p4

end
Conformance Checking
(WOZ objections Dutch municipality, 745 objections, 9583 event, f=0.988)
Replay can extract timing information
Performance Analysis Using Replay
(WOZ objections Dutch municipality, 745 objections, 9583 event, $f = 0.988$)
Process Discovery (small selection)

automata-based learning

heuristic mining

genetic mining

stochastic task graphs

fuzzy mining

mining block structures

α algorithm

α# algorithm

α++ algorithm

distributed genetic mining

language-based regions

state-based regions

LTL mining

neural networks

hidden Markov models

multi-phase mining

conformal process graph

partial-order based mining

ILP mining
Conformance Checking
Replaying trace “abeg”

\[
\text{fitness}(\sigma, N) = \frac{1}{2} \left( 1 - \frac{1}{6} \right) + \frac{1}{2} \left( 1 - \frac{1}{6} \right) = 0.83333
\]
From “playing the token game” to optimal alignments ...

observed trace: “abeg”
Another alignment

observed trace: “abcdeg”

move in log only
Moves have costs

- Standard cost function:
  - $c(x,\top) = 1$
  - $c(\top, y) = 1$
  - $c(x, y) = 0$, if $x=y$
  - $c(x, y) = \infty$, if $x\neq y$
Any cost structure is possible

<table>
<thead>
<tr>
<th>...</th>
<th>send-letter(John, 2 weeks, $400)</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>send-email(Sue, 3 weeks, $500)</td>
<td>...</td>
</tr>
</tbody>
</table>

- **Similar activities** (more similarity implies lower costs).
- **Resource conformance** (done by someone that does not have the specified role).
- **Data conformance** (path is not possible for this customer).
- **Time conformance** (missed the legal deadline)
Alignments

- conformance checking to diagnose deviations
- squeezing reality into the model to do model-based analysis
How to get started?
Hundreds of plug-ins available covering the whole process mining spectrum

Download from: www.processmining.org
Commercial Alternatives

• Disco (Fluxicon)
• Perceptive Process Mining (before Futura Reflect and BPM|one)
• ARIS Process Performance Manager
• QPR ProcessAnalyzer
• Interstage Process Discovery (Fujitsu)
• Discovery Analyst (StereoLOGIC)
• XMAnalyzer (XMPro)
• …
How to Get Started?

Collect event data ↔ Collect questions

- Minimal requirement: events referring to an activity name and a process instance.
- Good to have: timestamps, resource information, additional data elements.
- Challenges: scoping and sometimes correlation.

- What kind problems would you like to address (cost, time, risk, compliance, service, etc.)?
- Related to discovery, conformance, enhancement?
- Iterative process: can be “curiosity driven” initially.
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.