Disclaimer: please ignore the dates mentioned in this studyguide. Due to Corona things need to be replanned.

Study Guide Business Process Intelligence

Wil van der Aalst,
Gyunam Park
Mahsa Pourbafrani,
Majid Rafiei, and
Miriam Wagner

2020
Study Guide Business Process Intelligence (SS 2020)

Lecturers
- prof.dr.ir. Wil van der Aalst (lectures),
- Gyunam Park (instructions & assignments),
- Mahsa Pourbafrani (instructions & assignments),
- Majid Rafiei (instructions & assignments), and
- Miriam Wagner (instructions & assignments).

Course Contents and Motivation
This course starts with an overview of approaches and technologies that use event data to support decision-making and business process (re)design. Subsequently, the course focuses on process mining as a bridge between data mining and business process modeling. Business Process Intelligence (BPI) and process mining enable engineers to understand, diagnose, improve, and streamline operational processes for a wide variety of organizations and systems (hospitals, banks, high-tech systems, governments, electronic shops, transportation systems, trading systems, etc.).

Process mining is part of the larger data science discipline. Data science aims to answer questions as:
- What really happened? (discovery)
- Why did it happen? (root cause analysis)
- What will happen? (prediction)
- What is the best that can happen? (recommendation)

Process mining is an enabling technology to answer such questions about operational processes in different domains. There is a huge demand for engineers having the skills and tools to turn event data into real value.

The course is at an introductory level with various practical assignments.

The course covers the three main types of process mining. The first type of process mining is discovery. A discovery technique takes an event log and produces a process model without using any a-priori information. An example is the alpha-algorithm that takes an event log and produces a Petri net explaining the behavior recorded in the log. The second type of process mining is conformance. Here, an existing process model is compared with an event log of the same process. Conformance checking can be used to check if reality, as recorded in the log, conforms to the model and vice versa. The third type of process mining is enhancement. Here, the idea is to extend or improve an existing process model using information about the actual process recorded in some event logs. Whereas conformance checking measures the alignment between model and reality, this third type of process mining aims at changing or extending the a-priori model. An example is the extension of a process model with performance information, e.g., showing bottlenecks. Process mining techniques can be used in an offline, but also online setting. The latter is known as operational support. An example is the detection of non-conformance at the moment the deviation actually takes place. Another example is time prediction for running cases, i.e., given a partially executed case the remaining processing time is estimated based on historic information of similar cases.

Process mining provides not only a bridge between data mining and business process management; it also helps to address the classical divide between "business" and "IT". Evidence-based business process management based on process mining helps to create a common ground for business process improvement and information systems development.

In recent years, process mining has become the primary data-driven BPM (Business Process Management) approach. Process mining is also increasingly applied in other domains (auditing, production, etc.). The attention for Big Data and the uptake of data science strengthen this development. Process mining is where “Data Science” and “Process Science”
meet! Currently, there are about 30 software vendors offering process mining tools and all consultancy firms are offering process mining services. There are many larger organizations using process mining at a global scale (e.g., within Siemens over 6000 people are using process mining). Next, to open-source tools like ProM, Pm4Py, and RapidProM there are commercial tools such as Celonis Process Mining, Fluxicon, ProcessGold/UiPath, Signavio, Minit, mylnvenio, QPR ProcessAnalyzer, Everflow, Puzzledata, PAFnow, Software AG, Stereologic, Logpickr, Mehrwerk, Lanalabs, etc. The availability and application of these tools illustrate the uptake of process mining.

The course uses many examples using real-life event logs to illustrate the concepts and algorithms. After taking this course, one is able to run process mining projects and have a good understanding of the Business Process Intelligence (BPI) field. Moreover, students will be able to directly apply process mining techniques in all kinds of practical settings, including internships and master projects.

**Objectives**

After taking this course, students should:

- have a good understanding of Business Process Intelligence techniques (in particular process mining),
- understand the role of Big Data and Data Science in today's society,
- be able to relate process mining techniques to other analysis techniques such as simulation, business intelligence, data mining, machine learning, and verification,
- understand the relation between process mining and data mining techniques like classification, clustering, and association rules,
- be able to apply basic process discovery techniques such as the alpha algorithm to learn a process model from an event log (both manually and using tools),
- understand how more advanced process discovery techniques like region-based mining, genetic mining, and heuristic mining work
- be able to apply basic conformance checking techniques (such as token-based replay) to compare event logs and process models (both manually and using tools),
- be able to extend a process model with information extracted from the event log (e.g., show bottlenecks),
- have a good understanding of the data needed to start a process mining project,
- be able to characterize the questions that can be answered based on such event data,
- explain how process mining can also be used for operational support (prediction and recommendation), and
- be able to execute process mining projects in a structured manner using the L* life-cycle model.

**Organization & Lecture Material**

The course starts on Wednesday, April 8th 2020. Lectures are on Wednesdays from 10.30 to 12.00 in AH IV (2354|030) and Thursdays from 08.30 to 10.00 and Instructions are on Fridays from 10.30 to 12.00 also in AH IV (2354|030). Please note that due to planning problems there are few exceptions (the goal is to have on average 3 hours of lectures per week and allow time for working on the assignments). Especially towards the end, there will be several lectures on Fridays and instructions on Wednesdays and Thursdays.

The textbook "W. van der Aalst. Process Mining: Data Science in Action. Springer-Verlag, Berlin, 2016" ([http://springer.com/9783662498507](http://springer.com/9783662498507)) is the primary source of information and the lectures will be linked to chapters in the book. It can be ordered from [http://springer.com/9783662498507](http://springer.com/9783662498507) (also note the 50% discount when registering for the Coursera Process Mining MOOC, see later).

Next to the book, the following material will be distributed via the RWTH-moodle E-learning platform:

- Slides,
- Exercises,
Event logs, and
Assignments.

**Coursera MOOC in Process Mining**

The material of the BPI course was used for the successful MOOC (Massive Open Online Course) *Process Mining: Data Science in Action* (see [https://www.coursera.org/learn/process-mining](https://www.coursera.org/learn/process-mining)). Over 120,000 participants joined the course over the last couple of years, illustrating the global interest in the topic. This way you can watch lectures and do additional assignments at your own pace (only as background or in case things are not clear). When things are not clear you can watch the corresponding lecture.


The online lectures are listed below (cover about 60-70% of the course):

- Lecture 1.1: Data Science and Big Data (17 min.)
- Lecture 1.2: Different Types of Process Mining (21 min.)
- Lecture 1.3: How Process Mining Relates to Data Mining (20 min.)
- Lecture 1.4: Learning Decision Trees (27 min.)
- Lecture 1.5: Applying Decision Trees (21 min.)
- Lecture 1.6: Association Rule Learning (18 min.)
- Lecture 1.7: Cluster Analysis (13 min.)
- Lecture 1.8: Evaluating Mining Results (15 min.)

- Lecture 2.1: Event Logs and Process Models (14 min.)
- Lecture 2.2: Petri Nets (1/2) (16 min.)
- Lecture 2.3: Petri Nets (2/2) (18 min.)
- Lecture 2.4: Transition Systems and Petri Net Properties (21 min.)
- Lecture 2.5: Workflow Nets and Soundness (17 min.)
- Lecture 2.6: Alpha Algorithm: A Process Discovery Algorithm (25 min.)
- Lecture 2.7: Alpha Algorithm: Limitations (23 min.)
- Lecture 2.8: Introducing ProM and Disco (25 min.)

- Lecture 3.1: Four Quality Criteria for Process Discovery (19 min.)
- Lecture 3.2: On The Representational Bias of Process Mining (17 min.)
- Lecture 3.3: Business Process Model and Notation (BPMN) (15 min.)
- Lecture 3.4: Dependency Graphs and Causal Nets (21 min.)
- Lecture 3.5: Learning Dependency Graphs (21 min.)
- Lecture 3.6: Learning Causal nets and Annotating Them (18 min.)
- Lecture 3.7: Learning Transition Systems (15 min.)
- Lecture 3.8: Using Regions to Discover Concurrency (18 min.)

- Lecture 4.1: Two-Phase Process Discovery and Its Limitations (15 min.)
- Lecture 4.2: Alternative Process Discovery Techniques (23 min.)
- Lecture 4.3: Introduction to Conformance Checking (12 min.)
- Lecture 4.4: Conformance Checking Using Causal Footprints (10 min.)
- Lecture 4.5: Conformance Checking Using Token-Based Replay (15 min.)
- Lecture 4.6: Token-Based Replay: Some Examples (15 min.)
- Lecture 4.7: Aligning Observed and Modeled Behavior (18 min.)
- Lecture 4.8: Exploring Event Data (21 min.)

- Lecture 5.1: About the Last Two Weeks of This Course (10 min.)
- Lecture 5.2: Mining Decision Points (17 min.)
• Lecture 5.3: Discovering Data Aware Petri Nets (12 min.)
• Lecture 5.4: Mining Bottlenecks (11 min.)
• Lecture 5.5: Mining Social Networks (17 min.)
• Lecture 5.6: Organizational Mining (9 min.)
• Lecture 5.7: Combining Different Perspectives (13 min.)
• Lecture 5.8: Comparative Process Mining Using Process Cubes (13 min.)
• Lecture 5.9: Refined Process Mining Framework (11 min.)

• Lecture 6.1: Operational Support: Detect, Predict and Recommend (17 min.)
• Lecture 6.2: Getting the Right Event Data (17 min.)
• Lecture 6.3: Guidelines for Logging (10 min.)
• Lecture 6.4: Process Mining Software (16 min.)
• Lecture 6.5: How to Conduct a Process Mining Project (11 min.)
• Lecture 6.6: Mining Lasagna Processes (6 min.)
• Lecture 6.7: Mining Spaghetti Processes (8 min.)
• Lecture 6.8: Process Models as Maps (12 min.)
• Lecture 6.9: Data Science in Action (9 min.)

Software
The course uses the following analysis tools:

• **ProM lite 1.2:** The software can be downloaded from the ProM web site: [http://www.promtools.org/](http://www.promtools.org/) (see [http://www.promtools.org/doku.php?id=promlite12](http://www.promtools.org/doku.php?id=promlite12)). It requires Java 7.0 or higher. For more information about the requirement of this tool, please see [http://www.promtools.org/doku.php?id=promlite12](http://www.promtools.org/doku.php?id=promlite12).

• **Disco 2.2.1:** Download from [http://fluxicon.com/academic/](http://fluxicon.com/academic/), and visit [http://fluxicon.com/academic/material/](http://fluxicon.com/academic/material/) (use RWTH email address!)

• **RapidMiner 9.6:** Download the latest RapidMiner Studio version via [http://rapidminer.com/educational-program/](http://rapidminer.com/educational-program/). Note that as a RWTH student, you can apply for a license and get an unlimited version of RapidMiner (please follow the workflow on [https://rapidminer.com/educational-program/](https://rapidminer.com/educational-program/) carefully and use your RWTH account).

• **Celonis:** Obtain access to the “Celonis Intelligent Business Cloud - Academic Edition” via [https://academiccloud.celonis.com](https://academiccloud.celonis.com). See [https://www.celonis.com/de/academic-alliance/](https://www.celonis.com/de/academic-alliance/) for information (you need to apply for an academic license).

• Optionally, you are encouraged to use additional process mining tools. Several vendors (next to Fluxicon and Celonis) offer and academic program (e.g., LANA Labs).

Note that the RapidProM extension [www.rapidprom.org](http://www.rapidprom.org) is not needed for this course, but will be used in later specialized courses and may already be interesting.

Examination
The exam consists of two parts: one assignment (Schriftliche Hausarbeit) delivered in two parts, counting for 40% of the final result, and the final written test which counts for the remaining 60% of the final result. Successful participation in the assignment is required for participation in the final test. Only the final test can be retaken in this semester (there will be one re-exam). The assignment can only be redone in the next academic year. The assignment is done in groups of 2-3 persons (the same group for both parts of the assignment).

• **Final written test (60%):** <<check later for possible changes>>
  o First option (PT1): Friday 14/08/2020 (11:30 - 13:30)
  o Second option (PT2): Wednesday 09/09/2020 (8:30 - 10:30)
• **Schriftliche Hausarbeit Teil 1/DS Assignment Part 1** (20%): deadline Friday 29/05/2020
• **Schriftliche Hausarbeit Teil 2/DS Assignment Part 2** (20%): deadline Friday 10/07/2020
**Important:** Successful participation in the Schriftliche Hausarbeit/Assignment is a prerequisite for taking the written exam. The two parts form a whole and it is not possible to retake parts of the course, i.e., the results of the assignment expire after the end of the semester. Moreover, to pass the course it is required to pass both the Schriftliche Hausarbeit/Assignment and the written test. It means that you should obtain a minimum score of 50% in the Schriftliche Hausarbeit/Assignment and a minimum score of 50% in the final written test.

**Plagiarism:** We will systematically check for plagiarism between groups making the Schriftliche Hausarbeit/Assignment. Both groups and all members of both groups are responsible in case of plagiarism, and it will result in failure, will be reported, and may lead to removal from your studies.

Detailed descriptions of the assignments will be handed out separately.

**Who can take the course?**
The course can be taken at the master or bachelor level. It is a Wahlpflichtfach for several programs and an elective for many other programs. Students from other programs are welcome to participate, but it is up to the management and rules of the corresponding programs to decide whether the course "counts" (we cannot help you there). The course is related to the PADS courses Introduction to Data Science (IDS) and Advanced Process Mining (API). There is a small intentional overlap allowing you to take these courses in any order.

**Questions**
All the questions related to the lectures and instructions should be asked via Moodle. In case of urgent personal questions regarding the course, contact bpi@pads.rwth-aachen.de. Avoid sending e-mails to an individual or even multiple lecturers. If you have problems with RWTHonline, RWTHmoodle, etc., that are not specific for this course, please contact the persons responsible for these systems and not the lecturer.

**About the Process and Data Science (PADS) group @ RWTH**
The Process and Data Science (PADS) group, headed by prof.dr.ir. Wil van der Aalst, is one of the research units in the Department of Computer Science. The scope of PADS includes all activities where discrete processes are analyzed, reengineered, and/or supported in a data-driven manner. Process-centricity is combined with an array of Data Science techniques. The group’s research and teaching activities can be characterized by the keywords: Data Science, Process Science, Process Mining, Business Process Management, Data Mining, Process Discovery, Conformance Checking, and Simulation.

The group has been established in the context of the Alexander von Humboldt Professorship awarded to prof.dr.ir Wil van der Aalst in 2018. The award is Germany’s most prestigious and valuable prize for international researchers. The PADS group supports RWTH’s strategy to further strengthen its Data Science capabilities. The group also closely collaborates with the Fraunhofer Institute for Applied Information Technology (FIT).

Currently, the main research focus is on Process Mining (including process discovery, conformance checking, performance analysis, predictive analytics, operational support, and process improvement). This is combined with neighboring disciplines such as operations research, algorithms, discrete event simulation, business process management, and workflow automation.

Visit [http://www.pads.rwth-aachen.de/](http://www.pads.rwth-aachen.de/) to learn more about possible Bachelor and master theses.

**Planning**
See spreadsheet.
<table>
<thead>
<tr>
<th>#</th>
<th>Lecture</th>
<th>date</th>
<th>day</th>
<th>description</th>
<th>book chapters</th>
<th>mooc lectures</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction to Process Mining</td>
<td>4/8/2020</td>
<td>Wednesday</td>
<td>Introduction to Data Science, Process Mining, and the organization of the course.</td>
<td>1 &amp; 2</td>
<td>1.1-1.3</td>
</tr>
<tr>
<td>2</td>
<td>Decision Trees</td>
<td>4/9/2020</td>
<td>Thursday</td>
<td>Basic introduction to classification and decision trees (intended for those not having a data mining background).</td>
<td>4</td>
<td>1.4-1.5</td>
</tr>
<tr>
<td>3</td>
<td>Association Rules &amp; Clustering</td>
<td>4/15/2020</td>
<td>Wednesday</td>
<td>Basic introduction to unsupervised learning, frequent item sets, pattern mining, and clustering (for those not having a data mining background)</td>
<td>4</td>
<td>1.6-1.8</td>
</tr>
<tr>
<td>4</td>
<td>Introduction to Process Discovery</td>
<td>4/16/2020</td>
<td>Thursday</td>
<td>Basic introduction to process discovery. What is the problem and what approaches are possible?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Data Mining</td>
<td>4/17/2020</td>
<td>Friday</td>
<td>Some exercises about Association Rules, Decision Tree, and Clustering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Petri Nets &amp; Alpha Algorithm</td>
<td>4/22/2020</td>
<td>Wednesday</td>
<td>More on Petri nets. Introduction to the Alpha algorithm.</td>
<td>3, 5 &amp; 6</td>
<td>2.2-2.7</td>
</tr>
<tr>
<td>7</td>
<td>Alpha Algorithm Continued</td>
<td>4/23/2020</td>
<td>Thursday</td>
<td>Limitation and properties of the Alpha algorithm.</td>
<td>6</td>
<td>2.5-2.7</td>
</tr>
<tr>
<td>8</td>
<td>Instruction 1 Data Mining</td>
<td>4/24/2020</td>
<td>Friday</td>
<td>Petri net exercise and practical experiments with RapidMiner.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Instruction 2 Data Mining and Petri Nets</td>
<td>4/30/2020</td>
<td>Thursday</td>
<td>More challenging questions on data mining and Petri nets and initial discovery questions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Instruction 3 Data Mining and Petri Nets</td>
<td>5/6/2020</td>
<td>Wednesday</td>
<td>How to evaluate discovered models? Notions like fitness, precision, generalization, simplicity. Discussion on representations.</td>
<td>6</td>
<td>3.1-3.2</td>
</tr>
<tr>
<td>11</td>
<td>Instruction 4 Alpha Algorithm and Model Evaluation</td>
<td>5/7/2020</td>
<td>Thursday</td>
<td>Introduction to an algorithm that can handle noise and incompleteness.</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Instruction 5 Heuristic Mining and Region-Based Mining</td>
<td>5/15/2020</td>
<td>Friday</td>
<td>Heuristic Mining and Region-Based Mining</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Event Data and Exploration</td>
<td>5/20/2020</td>
<td>Wednesday</td>
<td>What types of event data exist and what are the problems when data is &quot;not flat&quot;?</td>
<td>5</td>
<td>4.8-6.2</td>
</tr>
<tr>
<td>14</td>
<td>Advanced Process Discovery Techniques</td>
<td>5/22/2020</td>
<td>Friday</td>
<td>Inductive miner exercise and practical session for using ProM for discovery techniques</td>
<td>4</td>
<td>4.4-4.6</td>
</tr>
<tr>
<td>15</td>
<td>Conformance Checking (1/2)</td>
<td>5/27/2020</td>
<td>Wednesday</td>
<td>Conformance checking using footprint matrices and token based replay</td>
<td>8</td>
<td>4.4-4.6</td>
</tr>
<tr>
<td>16</td>
<td>Conformance Checking (2/2) / invited lecture?</td>
<td>5/28/2020</td>
<td>Thursday</td>
<td>Conformance checking using token based replay and alignments</td>
<td>8</td>
<td>4.5-4.7</td>
</tr>
<tr>
<td>17</td>
<td>Instruction 7 Conformance Checking</td>
<td>5/29/2020</td>
<td>Friday</td>
<td>Token-based replay and footprint exercises</td>
<td>8</td>
<td>4.4-4.6</td>
</tr>
<tr>
<td>18</td>
<td>Deadline 1 Conformance Checking</td>
<td>5/29/2020</td>
<td>Friday</td>
<td>Deadline at the end of the day 23.59 (CET) until HM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Instruction 8 Conformance Checking</td>
<td>6/10/2020</td>
<td>Wednesday</td>
<td>Alignment exercise and using ProM for computing the fitness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Decision Mining</td>
<td>6/12/2020</td>
<td>Friday</td>
<td>How to learn factors influencing decisions in processes? Application of classification techniques in processes.</td>
<td>9</td>
<td>5.2-5.3</td>
</tr>
<tr>
<td>21</td>
<td>Decision Mining Exercise</td>
<td>6/17/2020</td>
<td>Wednesday</td>
<td>Decision Mining Exercises</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Organizational Mining &amp; Bottleneck Analysis</td>
<td>6/18/2020</td>
<td>Thursday</td>
<td>How to discover social networks from event data? How to uncover bottlenecks?</td>
<td>11.12</td>
<td>not in MOOC</td>
</tr>
<tr>
<td>23</td>
<td>Refined Process Mining Framework and Operational Support</td>
<td>6/19/2020</td>
<td>Friday</td>
<td>Putting different things together in one framework, including prescriptive and predictive analytics. Doing a PM project. Two types of processes.</td>
<td>10</td>
<td>5.9-6.1</td>
</tr>
<tr>
<td>24</td>
<td>Social Network discovery and Organizational Mining</td>
<td>6/24/2020</td>
<td>Wednesday</td>
<td>Different types of social networks and the corresponding exercises</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Dealing with Big Event Data</td>
<td>6/26/2020</td>
<td>Friday</td>
<td>Big data, event logs and event streams, streaming process mining, decomposed process mining, process mining tools</td>
<td>11.12</td>
<td>not in MOOC</td>
</tr>
<tr>
<td>26</td>
<td>Discussion of an old/possible exam</td>
<td>7/1/2020</td>
<td>Wednesday</td>
<td>Information will be provided before.</td>
<td>13, 14</td>
<td>6.4-6.7</td>
</tr>
<tr>
<td>27</td>
<td>Instruction 11 Second Assignment Q&amp;A</td>
<td>7/2/2020</td>
<td>Thursday</td>
<td>Q&amp;A for the second assignment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Instruction 12 Performance and bottleneck analysis</td>
<td>7/3/2020</td>
<td>Friday</td>
<td>Performance metrics, waiting time, executing time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Summary of the Course and Next Steps</td>
<td>7/8/2020</td>
<td>Wednesday</td>
<td>Summary of the course. What is important for the exam? What to do next? Last possibility.</td>
<td></td>
<td>6.8-6.9</td>
</tr>
<tr>
<td>30</td>
<td>Deadline 2 Deadline for Assignment Part 2</td>
<td>7/10/2020</td>
<td>Friday</td>
<td>Deadline at the end of the day 23.59 (CET) until HM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Instruction 13 Refined Process Mining, Prediction</td>
<td>7/15/2020</td>
<td>Thursday</td>
<td>Prediction with Refined Process Mining</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Instruction 14 Q&amp;A</td>
<td>7/16/2020</td>
<td>Thursday</td>
<td>Q&amp;A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Exam 1</td>
<td>8/14/2020</td>
<td></td>
<td>11:30-13:30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>Exam 2</td>
<td>9/9/2020</td>
<td></td>
<td>08:30-10:30</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Disclaimer: please ignore the dates mentioned in this studyguide. Due to Corona things need to be replanned.