Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Study Guide Business Process Intelligence

Wil van der Aalst,
Niklas Adams,
Bianka Bakullari,
Ali Norouzifar,
Marco Pegoraro,
Mahsa Pourbafrani,
Mahnaz Qafari,
Majid Rafiei, and
Miriam Wagner

Disclaimer: Due to COVID-19 things may change and information may be incomplete. Changes and additional information will be announced via RWTH Moodle.

2021

Study Guide Business Process Intelligence (SS 2021)

Lecturers

- prof.dr.ir. Wil van der Aalst (lectures),
- Niklas Adams (instructions & assignment),
- Bianka Bakullari (instruction & assignment),
- Ali Norouzifar (instruction & assignment),
- Marco Pegoraro (instruction & assignment),
- Mahsa Pourbafrani (instructions & assignment),
- Majid Rafiei (instructions & assignment),
- Mahnaz Qafari (instructions & assignment), and
- Miriam Wagner (instructions & assignment).

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, myInvenio, 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 14th 2021. Lectures are planned on Wednesdays from 10.30 to 12.00 and Thursdays from 08.30 to 10.00 and Instructions are held on Fridays from 8.30 to 10.00 online. Please note that the lectures will be recorded and instead there will be regular Q&A sessions. These Q&A sessions are only meaningful if you actually watched the lectures. Also there are few exceptions in the planning (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 instructions on Thursdays.

The textbook "W. van der Aalst. Process Mining: Data Science in Action. Springer-Verlag, Berlin, 2016" (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 (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). Over 135,000 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.

Also, note that there is a 20%-50% discount on the textbook W.M.P. van der Aalst. Process Mining: Data Science in Action. Springer-Verlag, Berlin, 2016. http://www.springer.com/978-3-662-49850-7 if you register.

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.)

Also the lectures of previous year are available via Video AG and YouTube.

Software

The course uses the following analysis tools:

- **ProM lite 1.3:** The software can be downloaded from the ProM web site: http://www.promtools.org/ (see http://www.promtools.org/doku.php?id=promlite13). 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=promlite13.
- **Disco 2.12.2 (or later):** Download from http://fluxicon.com/academic/, and visit http://fluxicon.com/academic/material/ (use RWTH email address!)
- **RapidMiner 9.9 (or later):** Download the latest RapidMiner Studio version via 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/ carefully and use your RWTH account).
- **Celonis:** Obtain access to the "Celonis Academic Edition" via https://www.celonis.com/academic-signup. See 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 an academic program (e.g., LANA Labs).

Note that the RapidProM extension 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) consisting of several parts, counting for 40% of the final result, and the final Written Test which counts for the remaining 60% of the final result. Both the Assignment *and* the Written Test need to be passed to pass the whole course. 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.

Assignment Part 1 and 2 are done in groups of 2-3 persons. Part 3 is done individually in Dynexite with each subpart being open for one week.

- **Final Written Test** (60%):
  - First option (PT1): To be announced later (exam planning of semester is ongoing).
  - Second option (PT2): To be announced later (exam planning of semester is ongoing).
- **Schriftliche Hausarbeit Teil 1/DS Assignment Part 1** (10%): deadline Friday 04/06/2021
- **Schriftliche Hausarbeit Teil 2/DS Assignment Part 2** (20%): deadline Friday 16/07/2021
- **Schriftliche Hausarbeit Teil 3/DS Assignment Part 3** (10%)
  - **Subpart 1:** deadline Friday 30/04/2021
  - **Subpart 2:** deadline Friday 07/05/2021
  - **Subpart 3:** deadline Thursday 20/05/2021
  - **Subpart 4:** deadline Friday 11/06/2021
  - **Subpart 5:** deadline Friday 02/07/2021

Important: Successful participation in the Schriftliche Hausarbeit/Assignment is a prerequisite to passing the whole course. Since the Assignment cannot be redone, there is no point in taking the Written Test if you failed the Assignment. 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. As stated before, 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 (APM). 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 PADS group is one of the globally leading research groups in process mining and other topics combining data science and process science. The group also closely collaborates with the Fraunhofer Institute for Applied Information Technology (FIT) and is one of the key groups in the Cluster of Excellence Internet of Production (IoP) and the RWTH Artificial Intelligence center. 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/ to learn more about possible Bachelor and master theses.

Planning

See spreadsheet.

#	Lecture	week	date	day	description	book chapters	mooc lectures
1	Introduction to Process Mining	1	14/04/2021	Wednesday	Introduction to Data Science, Process Mining, and the organization of the course.	1 & 2	1.1-1.3
2	Decision Trees	1	15/04/2021	Thursday	Basic introduction to classification and decision trees (intended for those not having a data mining background).	4	1.4-1.5
Instruction 1	Tool introduction	4	06/05/2021	Thursday	ProM and Disco		
3	Association Rules & Clustering	2	21/04/2021	Wednesday	Basic introduction to unsupervised learning, frequent item sets, pattern mining, and clustering (for those not having a data mining background).	4	1.6-1.8
4	Introduction to Process Discovery	2	22/04/2021	Thursday	Basic introduction to process discovery. What is the problem and what approaches are possible?	2 & 3	1.2-2.1
Q&A 1	Lecture 1-3	2	22/04/2021	Thursday			
Instruction 2	Data Mining + RapidMiner	2	23/04/2021	Friday	Some exercises about Association Rules, Decision Tree, and Clustering, Introduction to RapidMiner		
5	Petri Nets & Alpha Algorithm	3	28/04/2021	Wednesday	More on Petri nets. Introduction to the Alpha algorithm.	3, 5 & 6	2.2-2.7
6	Alpha Algorithm Continued	3	29/04/2021	Thursday	Limitation and properties of the Alpha algorithm.	6	2.5-2.7
Q&A 2	Lecture 4-6	3	29/04/2021	Thursday			
Instruction 3	Petri Nets	3	30/04/2021	Friday	Petri net exercise and initial discovery questions		
	Deadline for Assignment Subpart 1	3	30/04/2021	Friday			
7	Quality of Discovered Models and Representations	4	05/05/2021	Wednesday	How to evaluate discovered models? Notions like fitness, precision, generalization, simplicity. Discussion on representations.	6	3.1-3.2
8	Heuristic Mining	4	06/05/2021	Thursday	Introduction to an algorithm that can handle noise and incompleteness.	7	
Instruction 4	Alpha Algorithm and Model Evaluation	4	07/05/2021	Friday	Alpha miner exercises, Quality measures		
	Deadline for Assignment Subpart 2	4	07/05	Friday			
9	Region-Based Mining	5	12/05/2021	Wednesday	Introduction to algorithms that provide guarantees but cannot handle noise.	7	3.7-4.2
Instruction 5	Heuristic Mining and Region-Based Mining	5	14/05/2021	Friday	Heuristic Mining and Region-Based Mining		
	Group forming deadline Part 1	5	14/05/2021	Friday	Students need to have chosen a group by now!		
10	Inductive Mining	6	19/05/2021	Wednesday	Introduction to a state-of-the-art approach (Inductive mining).	7	not in MOOC
Q&A 3	Lecture 7-10	6	19/05/2021	Wednesday			
Instruction 6	Advanced Process Discovery Techniques	6	20/05/2021	Thursday	Inductive miner exercise		
	Deadline for Assignment Subpart 3	6	20/05/2021	Thursday			
Instruction 7	First part of the Assignment Q&A	6	21/05/2021	Friday			
11	Event Data and Exploration	8	02/06/2021	Wednesday	What types of event data exist and what are the problems when data is "not flat"?	5	4.8 6.2
12	Conformance Checking (1/2)	8	04/06/2021	Friday	Conformance checking using footprint matrices and token based replay	8	4.4-4.6
Deadline 1	**Deadline for Assignment Part 1**	8	04/06/2021	Friday	Deadline at the end of the day 23.59 (CET)	until HM	
13	Conformance Checking (2/2)	9	09/06/2021	Wednesday	Conformance checking using token based replay and alignments	8	4.5-4.7
Instruction 8	Conformance Checking	9	10/06/2021	Thursday	Token-based replay and footprint exercises		
Instruction 9	Conformance Checking	9	11/06/2021	Friday	Alignment exercise and using ProM for computing the fitness		
	Deadline for Assignment Subpart 4	9	11/06/2021	Friday			
14	Decision Mining	10	16/06/2021	Wednesday	How to learn factors influencing decisions in processes? Application of classification techniques in processes.	9	5.2-5.3
Q&A 4	Lecture 11-14	10	16/06/2021	Wednesday			
Instruction 10	Decision Mining	10	18/06/2021	Friday	Decision Mining Exercises		

15	Organizational Mining & Bottleneck Analysis	11	23/06/2021	Wednesday	How to discover social networks from event data? How to uncover bottlenecks?	11,12	not in MOOC
16	Refined Process Mining Framework and Operational Support	11	24/06/2021	Thursday	Putting different things together in one framework, including prescriptive and predictive analytics. Doing a PM project. Two types of processes.	10	5.9-6.1
Instruction 11	Social Network discovery and Organizational Mining	11	25/06/2021	Friday	Different types of social networks and the corresponding exercises		
	Group forming deadline Part 2	11	25/06/2021	Friday	Students need to have chosen a group by now!		
Instruction 12	Performance and bottleneck analysis	12	01/07/2021	Thursday	Performance metrics, waiting time, executing time		
Instruction 13	Refined Process Mining, Prediction	12	02/07/2021	Friday	Prediction and theoretical questions on Big data topics		
	Deadline for Assignment Subpart 5	12	02/07/2021	Friday			
17	Dealing with Big Event Data	13	07/07/2021	Wednesday	Big data, event logs and event streams, streaming process mining, decomposed process mining, process mining tools	11,12	not in MOOC
Q&A 5	Lecture 15-17	13	07/07/2021	Wednesday			
Instruction 14	Second part of the Assignment Q&A	13	09/07/2021	Friday			
Deadline	**Deadline for Assignment Part 2**	14	16/07/2021	Friday	Deadline at the end of the day 23.59 (CET)	All	
18	Discussion of an old/possible exam	15	21/07/2021	Wednesday	Information will be provided before.	13,14	6.4-6.7
19	Summary of the Course and Next Steps	15	22/07/2021	Thursday	Summary of the course. What is important for the exam? What to do next?	15,16	6.8-6.9
Q&A 6	Lecture 18-19	15	22/07/2021	Thursday			
Instruction 15	Q&A	15	23/07/2021	Friday	Q&A		

Question & Answer Lecture Sessions (QALSs)

The lectures are pre-recorded and uploaded via RWTH Moodle, YouTube, and Video AG before the scheduled times. Actually, also the recordings of the lectures of the previous year (BPI 2020) are online, see:

- https://youtube.com/playlist?list=PLG_1ZxIPXO0uA-LolJSQH2jzJ8oUiq9hu and
- https://video.fsmpi.rwth-aachen.de/20ss-bpi.

Since the content of the course did not change much, you can also progress faster if you want. To allow for questions about the lectures **Question & Answer Lecture Sessions (QALSs)** are scheduled via Zoom. These sessions are on selected slots reserved for the lectures.

- Thursday 22-4-2021 8.30: **Lectures 1-3** -Introduction to Process Mining, Decision Trees, Association Rules & Clustering
- Thursday 29-04-2021 8.30: **Lectures 4-6** - Introduction to Process Discovery, Petri Nets & Alpha Algorithm, Alpha Algorithm Continued
- Wednesday 19-5-2021 10.30: **Lectures 7-10** - Quality of Discovered Models and Representations, Heuristic Mining, Region-Based Mining, Inductive Mining
- Wednesday 16-6-2021 10.30: **Lectures 11-14** - Event Data and Exploration, Conformance Checking (1/2+2/2), Decision Mining
- Wednesday 7-7-2021 10.30: **Lectures 15-17** - Organizational Mining & Bottleneck Analysis, Refined Process Mining Framework and Operational Support, Dealing with Big Event Data
- Thursday 22-07-2021 8.30: **Lectures 18-19** - Discussion of an old/possible exam, Summary of the Course and Next Steps

Important: These sessions are only effective if you have actually watched the lectures before and studied the slides. You can use the Zoom Group Chat to ask questions and see the questions of others. Please prepare the questions in such a way that you can copy and paste in the Zoom Chat window during the QALS. Also, pose the questions in such a way that the problem is also clear for your fellow students (e.g., self-contained and clearly linking to lectures slides). I will handle questions lecture-by-lecture (again to focus the discussion). The QALSs will not be recorded and will take as long as there are questions.