Preface

This book encompasses a revised version of the Ph.D. dissertation, written by the author, at the Mathematics Department of the University of Padua (Italy), and at the Computer Science Department of the University of Bologna (Italy).

In 2014, the dissertation won the "Best Process Mining Dissertation Award", assigned by the IEEE Task Force on Process Mining to the most outstanding Ph.D. thesis, discussed between 2012 and 2013, focused on the area of business process intelligence.

The increasing availability of storage and computing capability, combined with the advent of new "smart" devices, represents the fundamental basis of the so-called "Internet of Things" (IoT). Business companies are focusing their attention to IoT as well, since it could be exploited in a valuable manner. One of the results of such IoT diffusion, but more generally, a common trend of these years, is that the data collection is monumentally increasing.

It is important to remind that the value of data is intimately connected to the *knowledge* that it is possible to synthesize from them. Moreover, in order to strengthen their business, the focus of companies should be on the consolidation and improvement of their business processes, rather than on their data. This is the scenario where process mining sits: in between data mining, and business process modeling.

After a brief presentation of the state of the art of process mining techniques, this book proposes different scenarios for the deployment of process mining projects. In particular, a characterization of companies, in terms of their "process awareness" (and process awareness of their information systems), is detailed.

The work continues identifying and reporting the possible circumstances where problems, both "practical" and "conceptual", can emerge. We identified these three areas as possible sources of problems: *(i)* data preparation (e.g., syntactic translation of data, missing data); *(ii)* the actual mining phase (e.g., mining algorithm exploiting all data available); and *(iii)* results interpretation. Several problems are not limited to a single phase, but orthogonal to all the mentioned sources: for example, the configuration of parameters by non-expert users or the computational complexity of some techniques. In this book we will analyze at least one solution for each of the presented problems. The descriptions of these solutions are kept general, in order to easily allow their tailoring into specific application domains.

The solutions proposed in this book belong to two different computational paradigms: the first considers the classical "batch process mining" (also known as "off-line"); the second introduces the "on-line process mining".

Concerning batch process mining, we are going to investigate first the data preparation problem and we will analyze and present a solution for the problem of hidden data (i.e., when a required field is not explicitly indicated). In our example we are going to consider the "case-id". In particular, our approach tries to identify this missing information by looking at *metadata* recorded for each event.

After that, we will concentrate on the second step (the mining phase) and, in particular, on the problem of exploiting all the available information. As example, we propose the generalization of a well-known control-flow discovery algorithm (i.e., Heuristics Miner) in order to exploit non-instantaneous events. The usage of interval-based recording leads to an important improvement of the algorithm performance. As another example of data exploitation, we present an automatic approach for the extension of a control-flow model with social information (i.e., roles), in order to simplify the analysis of these two perspectives (the control-flow and resources) combined.

Later on, we will focus our attention on another important and, for non-expert users, impacting problem: the parameters configuration. As example, we considered the configuration of a control-flow discovery algorithm. Our approach consists of two steps: first, we introduce a method to automatically discretize the space of parameter values. Then, we present two approaches to select the "best" parameters configuration. The first, completely autonomous, uses the Minimum Description Length principle to balance the model complexity and the data explanation; the second requires human interaction to navigate a hierarchy of models and find the most suitable result.

The data interpretation and results evaluation phase is not problem free, as well. Also in this case, we will analyze the problems and propose two new metrics: a *model-to-model* and a *model-to-log* (the latter considers models expressed in declarative language).

The final part of this book deals with the adaptation of process mining to on-line settings. We will consider, as example, the problem of on-line control-flow discovery. Specifically, we are going to propose a formal definition of the problem and then present two baseline approaches. These two basic approaches are used only for validation purposes. The actual mining algorithms proposed will be two: the first is the adaptation, to the control-flow discovery problem, of a well-known frequency counting algorithm (i.e., Lossy Counting); the second constitutes a framework of models which can be used for different kinds of streams (for example, stationary streams or streams with concept drifts)

Innsbruck, Austria — Andrea Burattin
February 2015

http://www.springer.com/978-3-319-17481-5

Process Mining Techniques in Business Environments
Theoretical Aspects, Algorithms, Techniques and Open Challenges in Process Mining
Burattin, A.
2015, XII, 220 p. 101 illus., Softcover
ISBN: 978-3-319-17481-5