CHAPTER – 3

PROCESS MINING

This chapter deals with the Descriptive, Prescriptive and explanatory goals of process models and its uses. The Business Process Model life-cycle and the three activities of process mining i.e. Discovery, Conformance & Enhancement are presented. Various stages of Lasagna and Spaghetti process also discussed.

3.1 Data Explosion

Information systems are becoming more and more intertwined with the operational processes they support. As a result, multitudes of events are recorded by today’s information systems. Nevertheless, organizations have problems extracting value from these data. The goal of process mining is to use event data to extract process related information, e.g., to automatically discover a process model by observing events recorded by some enterprise system (Wil M.P. van der Aalst, 2010). Process mining can play an important role in realizing the promises made by contemporary management trends such as SOX and Six Sigma.

Most of the data stored in the digital universe is unstructured and organizations have problems dealing with such large quantities of data. One of the main challenges of today’s organizations is to extract information and value from data stored in their information systems. The importance of information systems is not only reflected by the spectacular growth of data, but also by the role that these systems play in today’s business processes as the digital universe and the physical universe are becoming more and more aligned (Wil M.P. van der Aalst, 2010). Technologies such as RFID (Radio Frequency Identification), GPS (Global Positioning System), and sensor networks will stimulate a further alignment of the digital universe and the physical universe.

The growth of a digital universe that is well-aligned with processes in organizations makes it possible to record and analyze events. Events may range from the withdrawal of cash from an ATM, a doctor setting the dosage of an X-ray machine, a citizen applying for a driver license, the submission of a tax declaration, and the receipt of an e-ticket number by a traveler.
The challenge is to exploit event data in a meaningful way, for example, to provide insights, identify bottlenecks, anticipate problems, record policy violations, recommend countermeasures, and streamline processes. This is called process mining.

### 3.2 Process Models

The goals of a process model (Wil M.P. van der Aalst, 2010) are to be:

- **Descriptive**
  - Track what actually happens during a process.
  - Take the point of view of an external observer who looks at the way a process has been performed and determines the improvements that must be made to make it perform more effectively or efficiently.

- **Prescriptive**
  - Define the desired processes and how they should/could/might be performed.
  - Establish rules, guidelines, and behavior patterns which, if followed, would lead to the desired process performance. They can range from strict enforcement to flexible guidance.

- **Explanatory**
  - Provide explanations about the rationale of processes.
  - Explore and evaluate the several possible courses of action based on rational arguments.
  - Establish an explicit link between processes and the requirements that the model needs to fulfill.
  - Pre-defines points at which data can be extracted for reporting purposes.

**Process Models are used for**

- **Insight**: while making a model, the modeler is triggered to view the process from various angles.
- **Discussion**: the stakeholders use models to structure discussions.
- **Documentation**: processes are documented for instructing people or certification purposes (cf. ISO 9000 quality management).
• **Verification**: process models are analyzed to find errors in systems or procedures (e.g., potential deadlocks).

• **Performance analysis**: techniques like simulation can be used to understand the factors influencing response times, service levels, etc.

• **Animation**: models enable end users to “play out” different scenarios and thus provide feedback to the designer.

• **Specification**: models can be used to describe a PAIS before it is implemented and can hence serve as a “contract” between the developer and the end user/management.

• **Configuration**: models can be used to configure a system.

Clearly, process models play an important role in larger organizations. When redesigning processes and introducing new information systems, process models are used for a variety of reasons. Typically, two types of models are used: (a) informal models and (b) formal models (also referred to as “executable” models). Informal models are used for discussion and documentation whereas formal models are used for analysis or enactment (i.e., the actual execution of process). On the one end of the spectrum there are “PowerPoint diagrams” showing high-level processes whereas on the other end of the spectrum there are process models captured in executable code. Whereas informal models are typically ambiguous and vague, formal models tend to have a rather narrow focus or are too detailed to be understandable by the stakeholders. Independent of the kind of model—informal or formal—one can reflect on the alignment between model and reality. A process model used to configure a workflow management system is probably well-aligned with reality as the model is used to force people to work in a particular way (Wil M.P. van der Aalst, 2010). Unfortunately, most hand-made models are disconnected from reality and provide only an idealized view on the processes at hand. Moreover, also formal models that allow for rigorous analysis techniques may have little to do with the actual process.

The value of models is limited if too little attention is paid to the alignment of model and reality. Process models become “paper tigers” when the people involved cannot trust them. For example, it makes no sense to conduct simulation experiments while using a model that assumes an idealized version of the real process. It is likely that-based on such an idealized model-incorrect redesign decisions are made. It is also precarious to start a new
implementation project guided by process models that hide reality. A system implemented on the basis of idealized models is likely to be disruptive and unacceptable for end users. A nice illustration is the limited quality of most reference models (Wil M.P. van der Aalst, 2010). The idea is that “best practices” are shared among different organizations. Unfortunately, the quality of such models leaves much to be desired. For example, the SAP reference model has very little to do with the processes actually supported by SAP. In fact, more than 20 percent of the SAP models contain serious flaws (deadlocks, livelocks, etc.). Such models are not aligned with reality and, thus, have little value for end users. Given (a) the interest in process models, (b) the abundance of event data, and (c) the limited quality of hand-made models, it seems worthwhile to relate event data to process models. This way the actual processes can be discovered and existing process models can be evaluated and enhanced. This is precisely what process mining aims to achieve.

3.2.1 Business Process Model (BPM)

To position process mining, we first describe the so-called BPM life-cycle using Fig. 3.1

![BPM Life Cycle Process Model](image)

**Fig 3.1 - BPM Life Cycle Process Model**

The life-cycle describes the different phases of managing a particular business process. In the design phase, a process is designed. This model is transformed into a running system in
the configuration/implementation phase. If the model is already in executable form and a WFM or BPM system is already running, this phase may be very short. However, if the model is informal and needs to be hardcoded in conventional software, this phase may take substantial time. After the system supports the designed processes, the enactment/monitoring phase starts. In this phase, the processes are running while being monitored by management to see if any changes are needed. Some of these changes are handled in the adjustment phase shown in Fig. 3.1. In this phase, the process is not redesigned and no new software is created; only predefined controls are used to adapt or reconfigure the process. The diagnosis/requirements phase evaluates the process and monitors emerging requirements due to changes in the environment of the process (e.g., changing policies, laws, competition). Poor performance (e.g., inability to meet service levels) or new demands imposed by the environment may trigger a new iteration of the BPM lifecycle starting with the redesign phase (Wil M.P. van der Aalst, 2010).

Process models play a dominant role in the (re)design and configuration/implementation phases, whereas data plays a dominant role in the enactment/monitoring and diagnosis/requirements phases. Until recently, there were few connections between the data produced while executing the process and the actual process design. In fact, in most organizations the diagnosis/requirements phase is not supported in a systematic and continuous manner. Only severe problems or major external changes will trigger another iteration of the life-cycle, and factual information about the current process is not actively used in redesign decisions. Process mining offers the possibility to truly “close” the BPM life-cycle. Data recorded by information systems can be used to provide a better view on the actual processes, i.e., deviations can be analyzed and the quality of models can be improved.

### 3.2.2 Process Mining

Process mining is a relative young research discipline that sits between machine learning and data mining on the one hand and process modeling and analysis on the other hand.
The idea of process mining is to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs readily available in today’s systems. Process mining, i.e., extracting valuable, process-related information from event logs, complements existing approaches to Business Process Management (BPM). BPM is the discipline that combines knowledge from information technology and knowledge from management sciences and applies this to operational business processes. It has received considerable attention in recent years due to its potential for significantly increasing productivity and saving cost. BPM can be seen as an extension of Workflow Management (WFM) (Wil M.P. van der Aalst, 2010). WFM primarily focuses on the automation of business processes, whereas BPM has a broader scope: from process automation and process analysis to process management and the organization of work. On the one hand, BPM aims to improve operational business processes, possibly without the use of new technologies. For example, by modeling a business process and analyzing it using simulation, management may get ideas on how to reduce costs while improving service levels. On the other hand, BPM is often associated with software to manage, control, and support operational processes. This was the initial focus of WFM. Traditional WFM technology aims at the automation of business processes in a rather mechanistic manner without much attention for human factors and management support.

Process-Aware Information Systems (PAISs) include the traditional WFM systems, but also include systems that provide more flexibility or support specific tasks. For example, larger ERP (Enterprise Resource Planning) systems (SAP, Oracle), CRM (Customer
Relationship Management) systems, rule-based systems, call center software, high-end middleware (Web Sphere), etc. can be seen as process aware, although they do not necessarily control processes through some generic workflow engine. Instead, these systems have in common that there is an explicit process notion and that the information system is aware of the processes it supports. Also a database system or e-mail program may be used to execute steps in some business process. However, such software tools are not “aware” of the processes they are used in. Therefore, they are not actively involved in the management and orchestration of the processes they are used for. Some authors use the term BPMS (BPM system), or simply PMS (Process Management System), to refer systems that are “aware” of the processes they support. We use the term PAIS to stress that the scope is much broader than conventional workflow technology (Wil M.P. van der Aalst, 2010).

Fig 3.3 - Three types of process mining Discovery, Conformance and Enhancement
Process mining is a blend of computational intelligence and data mining. It is applicable to a wide range of systems. These systems may be a pure Information systems or systems where the hardware plays a more prominent role. The only requirement is that the system produces Event Logs, thus recording the actual behavior. Process mining is practically relevant and the logically next step in Business Process Management. Process mining provides many interesting challenges for scientists, customers, users, managers, consultants, and tool developers (Wil M.P. van der Aalst, 2010).

3.3 Event Logs to Process mining

3.3.1 Data Sources

The goal of process mining is to analyze event data from a process-oriented perspective. Fig 3.4 shows the overall process mining workflow.

Starting point is the “raw” data hidden in all kinds of data sources. A data source may be a simple flat file, an Excel spreadsheet, a transaction log, or a database table. However, one should not expect all the data to be in a single well-structured data source. The reality is that event data is typically scattered over different data sources and quite often some efforts are needed to collect the relevant data. Events can also be captured by tapping of message exchanges and recording read and write actions. Data sources may be structured and well-described by meta data. Unfortunately, in many situations, the data is unstructured or important meta data is missing. Data may originate from web pages, emails, PDF documents, scanned text, screen scraping, etc. Even if data is structured and described by meta data, the sheer complexity of enterprise information systems may be overwhelming. There is no point in trying to exhaustively extract events logs from thousands of tables and other data sources. Data extraction should be driven by questions rather than the availability of lots of data.

In the context of BI and data mining, the phrase “Extract, Transform, and Load” (ETL) is used to describe the process that involves: (a) extracting data from outside sources, (b) transforming it to fit operational needs and (c) loading it into the target system, e.g., a data warehouse or relational database. A data warehouse is a single logical repository of an organization’s transactional and operational data. The data warehouse does not produce data but simply taps off data from operational systems. The goal is to unify information such that it can be used for reporting, analysis, forecasting so that ETL activities can be used to populate a
It may require quite some efforts to create the common view required for a data warehouse. Different data sources may use different keys, formatting conventions, etc.
If a data warehouse already exists, most likely it holds valuable input for process mining. However, many organizations do not have a good data warehouse. The warehouse may contain only a subset of the information needed for end-to-end process mining, e.g., only data related to customers is stored. Moreover, if a data warehouse is present, it does not need to be process oriented. For example, the typical warehouse data used for Online Analytical Processing (OLAP) does not provide much process-related information. OLAP tools are excellent for viewing multidimensional data from different angles, drilling down, and for creating all kinds of reports. However, OLAP tools do not require the storage of business events and their ordering. The data sets used by the mainstream data mining approaches also do not store such information. Process mining requires information on relevant events and their order. Whether there is a data warehouse or not, data needs to be extracted and converted into event logs. Often the problem is not the syntactical conversion but the selection of suitable data. Typical formats to store event logs are XES (eXtensible Event Stream) and MXML (Mining eXtensible Markup Language). Once an event log is created, it is typically filtered. Filtering is an iterative process. Based on the filtered log, the different types of process mining can be applied: discovery, conformance, and enhancement.

### 3.3.2 Event logs

All process mining techniques assume that it is possible to sequentially record events such that each event refers to an activity (i.e., a well-defined step in some process) and is related to a particular case (i.e., a process instance). Event logs may store additional information about events. In fact, whenever possible, process mining techniques use extra information such as the resource (i.e., person or device) executing or initiating the activity, the timestamp of the event, or data elements recorded with the event (e.g., the size of an order). A few assumptions about an event log are:

- A process consists of cases.
- A case consists of events such that each event relates to precisely one case.
- Events within a case are ordered.
- Events can have attributes. Examples of typical attribute names are activity, time, etc.
- Costs and resource.
Not all events need to have the same set of attributes; however, typically, events referring to the same activity have the same set of attributes.

3.3.3 Play-In, Play-Out and Replay

Play-out refers to the classical use of process models. Given a Petri net, it is possible to generate behavior. The traces could have been obtained by repeatedly “playing the token game” using the Petri net. Play-out can be used both for the analysis and the enactment of business processes. A workflow engine can be seen as a “Play-out engine” that controls cases by only allowing the “moves” allowed according to the model. Hence, Play-out can be used to enact operational processes using some executable model. Simulation tools also use a Play-out engine to conduct experiments. The main idea of simulation is to repeatedly run a model and thus collect statistics and confidence intervals. Note that a simulation engine is similar to a workflow engine. The main difference is that the simulation engine interacts with a modeled environment whereas the workflow engine interacts with the real environment (workers, customers, etc.). Also classical verification approaches using exhaustive state-space analysis—often referred to as model checking—can be seen as Play-out methods (Wil M.P. van der Aalst, 2010).

Play-in is the opposite of Play-out, i.e., example behavior is taken as input and the goal is to construct a model. Play-in is often referred to as inference. The α-algorithm and other process discovery approaches are examples of Play-in techniques. Note that the Petri net can be derived automatically given an event log. Most data mining techniques use Play-in, i.e., a model is learned on the basis of examples. However, traditionally, data mining has not been concerned with process models. Typical examples of models are decision trees (“people that drink more than five glasses of alcohol and smoke more than 56 cigarettes tend to die young”) and association rules (“people that buy diapers also buy beer”). Unfortunately, it is not possible to use conventional data mining techniques to Play-in process models. Only recently, process mining techniques have become readily available to discover process models based on event logs.

Replay uses an event log and a process model as input. The event log is “replayed” on top of the process model. Simply “play the token game” by forcing the transitions to fire (if possible) in the order indicated.
Process mining is impossible without proper event logs. Necessary information should be present in the event logs. Depending on the process mining technique used, these requirements may vary. The challenge is to extract such data from a variety of data sources, e.g., databases, flat files, message logs, transaction logs, ERP systems, and document management systems.

When merging and extracting data, both syntax and semantics play an important role. Moreover, depending on the questions one seeks to answer, different views on the available data are needed.
3.3.4 Phases of Process Mining

The goal of process mining is to extract information from event logs i.e., process mining describes a family of posterior analysis techniques exploiting the information recorded in the event logs. Typically these approaches assume that it is possible to record sequential events such that each event refers an activity (well defined step in the process) and is related to a particular process instance. Furthermore, some mining techniques use additional information such as the performer or the originator of the event, the timestamp of the event or data elements recorded with the event. Process mining (Fig.3.6) also aims to discover, monitor and improve real process by extracting knowledge from event logs. The three basic types of process mining are (A) Discovery (B) Conformance (C) Enhancement.

![Fig 3.6 - Three Phases of Process Mining](image)

(A) Discovery: Traditionally, process mining has been focusing on discovery i.e., deriving information about the organization context, and execution properties from enactment logs. An example of a technique addressing the control flow perspective is the alpha algorithm, which constructs Petri net model describing the behavior observed in the event
logs. Process mining is not limited to process models and recent process mining techniques are more and more focusing on other perspectives e.g., the organizational perspective, performance perspective or data perspective. For example, there are approaches to extract social networks from event logs and analyze them using social network analyzer. This allows organizations to monitor how people, groups or software/system components are working together. Also there are approaches to visualize performance related information e.g., there are approaches which graphically show the bottlenecks and all kinds of performance indicators e.g., total flow time or the time spent between two activities.

(B) Conformance: Conformance checking relates events in the event log to activities in the process model and compares both. The goal is to find commonalities and discrepancies between the modeled behavior and the observed behavior. Conformance checking is relevant for business alignment and auditing.

For example, the event log can be replayed on top of the process model to find undesirable deviations suggesting fraud or inefficiencies. Moreover, conformance checking techniques can also be used for measuring the performance of process discovery algorithms and to repair models that are not aligned well with reality. There is an a-priori model. This model used to check if reality conforms to the model. For example there may be a process model indicating that purchase orders for very huge amount require two checks. Another example is the checking of the so-called “four-eye” principle. Conformation checking may be used to detect deviation, to locate and explain these deviations, and to measure the severity of these deviations.

Such conformation analysis results in global conformance measures and local diagnosis. The interpretation of non-conformance depends on the purpose of the model. If the model is intended to be descriptive, then discrepancies between model and log indicate that the model needs to be improved to capture reality better.

If the model is normative, then such discrepancies may be interpreted in two ways. Some of the discrepancies found may expose undesirable deviations, i.e., conformance checking signals the need for a better control of the process. Other discrepancies may reveal desirable deviations.
The goal of business alignment is to make sure that the information systems and the real business processes are well aligned. People should be supported by the information system rather than work behind its back to get things done. Unfortunately, there is often a mismatch between the information system on the one hand and the actual processes and needs of workers and management on the other hand. There are various reasons for this. First of all, most organization uses product software, i.e., generic software that was not developed for a specific organization.

A typical example is the SAP system which is based on so-called “best practices”, i.e., typical processes and scenarios are implemented. Although such systems are configurable, the particular needs of an organization may be different from what was envisioned by the product software developer. Second, processes may change faster than the information system, because of external influences. Finally, there may be different stakeholders in the organization having conflicting requirements, e.g., a manager may want to enforce a fixed working procedure whereas an experienced worker prefers to have more flexibility to serve customers better.

Process mining can assist in improving the alignment of information systems, business processes, and the organization. By analyzing the real processes and diagnosing discrepancies, new insights can be gathered showing how to improve the support by information systems. The term auditing refers to the evaluation of organizations and their processes. Audits are performed to ascertain the validity and reliability of information about these organizations and
associated processes. This is done to check whether business processes are executed within certain boundaries set by managers, governments, and other stakeholders. For instance, specific rules may be enforced by law or company policies and the auditor should check whether these rules are followed or not. Violations of these rules may indicate fraud, malpractice, risks, and inefficiencies. Traditionally, auditors can only provide reasonable assurance that business processes are executed within the given set of boundaries. They check the operating effectiveness of controls that are designed to ensure reliable processing. When these controls are not in place, or otherwise not functioning as expected, they typically only check samples of factual data, often in the “paper world”. However, today detailed information about processes is being recorded in the form of event logs, audit trails, transaction logs, databases, data warehouses, etc.

Conformance checking can be used for improving the alignment of business processes, organizations, and information systems. Replay techniques and footprint analysis help to identify differences between a process model and the real process as recorded in the event log. The differences identified may lead to changes of the model or process. For example, exposing deviations between the model and process may lead to better work instructions or changes in management. Conformance checking is also a useful tool for auditors that need to make sure that processes are executed within the boundaries set by various stakeholders. Conformance checking can also be used for other purposes such as repairing models and evaluating process discovery algorithms. Moreover, through conformance checking event logs get connected to process models and thus provide a basis for all kinds of analysis.

(C) Enhancement: Enhancement aims to extend or improve an existing process model using information about the actual process recorded in some event log. There is an a-priori model. This model is extended with a new aspect or perspective, i.e., the goal is not to check conformance but to enrich the model with the data in the event log. An example is the extension of a process model with performance data, i.e., some a-priori process model is used on which bottlenecks are projected.

As indicated, process models show only the control-flow. However, when extending process models additional perspectives are added. Moreover, discovery and conformance techniques are not limited to control-flow. For example, one can discover a social network and check the validity of some organizational model using an event log.
3.3.5 Mining Perspective

Orthogonal to the three types of mining (discovery, conformance, and enhancement), different perspectives can be identified.

The control-flow perspective focuses on the control-flow, i.e., the ordering of activities. The goal of mining this perspective is to find a good characterization of all possible paths, e.g., expressed in terms of a Petri net or some other notation (e.g., EPCs, BPMN, and UML ADs) (Wil M.P. van der Aalst, 2010).

- The **organizational perspective** focuses on information about resources hidden in the log, i.e., which actors (e.g., people, systems, roles, and departments) are involved and how are they related. The goal is to either structure the organization by classifying people in terms of roles and organizational units or to show the social network.

- The **case perspective** focuses on properties of cases. Obviously, a case can be characterized by its path in the process or by the originators working on it. However, cases can also be characterized by the values of the corresponding data elements. For example, if a case represents a replenishment order, it may be interesting to know the supplier or the number of products ordered.

- The **time perspective** is concerned with the timing and frequency of events. When events bear timestamps it is possible to discover bottlenecks, measure service levels, monitor the utilization of resources, and predict the remaining processing time of running cases.

3.4 Process Mining: Discovering and Improving the Lasagna and Spaghetti processes

3.4.1 Lasagna Process

Lasagna processes are relatively structured and the cases flowing through such processes are handled in a controlled manner. Lasagna processes have a clear structure and most cases are handled in a prearranged manner. There are relatively few exceptions and stakeholders have a reasonable understanding of the flow of work. It is impossible to define a formal requirement characterizing Lasagna processes (Wil M.P. van der Aalst, 2010). As a rule of thumb we use the following informal criterion: a process is a Lasagna process if with limited efforts it is possible to create an agreed-upon process model that has a fitness of at
least 0.8, i.e., more than 80% of the events happen as planned and stakeholders confirm the validity of the model.

In a structured process (i.e., Lasagna process) all activities are repeatable and have a well-defined input and output. In highly structured processes most activities can, in principle, be automated.

The goal of process mining is to improve operational processes. In order to judge whether process mining efforts are successful, we need to define Key Performance Indicators (KPIs). KPIs related to time (e.g., lead time, service time, waiting time, and synchronization time), KPIs related to costs, and KPIs related to quality. Note that quality may refer to compliance, customer satisfaction, number of defects, etc. To evaluate suggested improvements, the effectiveness and efficiency of the as-is and to-be processes need to be quantified in terms of KPIs.

For Lasagna processes, process mining can result in one or more of the following improvement actions:

- **Redesign.** Insights obtained using process mining can trigger changes to the process, e.g., sequential activities no longer need to be executed in a fixed order, checks may be skipped for easy cases, decisions can be delegated if more than 50 cases are queuing, etc. Fraud detected using process mining may result in additional compliance regulations, e.g., introducing the 4-eyes principle for critical activities.

- **Adjust.** Similarly, process mining can result in (temporary) adjustments. For example, insights obtained using process mining can be used to temporarily allocate more resources to the process and to lower the threshold for delegation.

- **Intervene.** Process mining may also reveal problems related to particular cases or resources. This may trigger interventions such as aborting a case that has been queuing for more than 3 months or disciplinary measures for a worker that repeatedly violated compliance regulations.

- **Support.** Process mining can be used for operational support, e.g., based on historic information a process mining tool can predict the remaining flow time or recommend the action with the lowest expected costs.
Some typical use cases for process mining are:

- Identification of bottlenecks to trigger a process redesign that reduces the overall flow time with 30%.
- Identification of compliance problems using conformance checking. Some of the compliance problems result in ad-hoc interventions whereas others lead to adjustments of the parameters used for work distribution.
- Harmonization of two processes after a merger based on a comparison of the actual processes. The goal of such a harmonization is to reduce costs.
- Predicting the remaining flow time of delayed cases to improve customer service.
- Providing recommendations for resource allocation aiming at a more balanced utilization of workers.
- Identification of exceptional cases that generate too much additional work. By learning the profile of such cases, they can be handled separately to reduce the overall flow time.
- Visualization of the 10 most complicated or time consuming cases to identify potential risks.
The ten activities of process mining are grouped into three categories: cartography (activities discover, enhance, and diagnose), auditing (activities detect, check, compare, and promote), and navigation (activities explore, predict, and recommend). Although the framework helps to understand the relations between the various process mining activities, it does not guide the user in conducting a process mining project. Therefore, we introduce the L* life-cycle model for mining Lasagna processes.

### 3.4.2 L* Lifecycle model for Lasagna process

The L* life-cycle model refers to the ten process mining related activities (explore, discover, check, etc.) and the four improvement actions (redesign, adjust, intervene, and support). There are five stages in the lifecycle model (Wil M.P. van der Aalst, 2010).

#### Stage 0: Plan and Justify

Any process mining project starts with a planning and a justification of the planned activities. Before spending efforts on process mining activities, one should anticipate benefits that may result from the project. There are basically three types of process mining projects:

- A **data-driven** (also referred to as “curiosity driven”) process mining project is powered by the availability of event data. There is no concrete question or goal, however, some of the stakeholders expect that valuable insights will emerge by analyzing event data. Such a project has an explorative character.

- A **question-driven** process mining project aims to answer specific questions, e.g., “Why do cases handled by team X take longer than cases handled by team Y?” or “Why are there more deviations in weekends?”.

- A **goal-driven** process mining project aspires to improve a process with respect to particular KPIs, e.g., cost reduction or improved response times.

For an organization without much process mining experience it is best to start with a question-driven project. Concrete questions help to scope the project and guide data extraction efforts. Like any project, a process mining project needs to be planned carefully. For instance, activities need to be scheduled before starting the project, resources need to be allocated, milestones need to be defined, and progress needs to be monitored continuously.
Stage 1: Extract

After initiating the project, event data, models, objectives, and questions need to be extracted from systems, domain experts, and management. In a goal-driven process mining project, the objectives are also formulated in Stage 1 of the L* life-cycle. These objectives are expressed in terms of KPIs. In a question-driven process mining project, questions need to be generated in Stage 1. Both questions and objectives are gathered through interviews with stakeholders (e.g., domain experts, end users, customers, and management).

Stage 2: Create Control-Flow Model and Connect Event Log

Control-flow forms the backbone of any process model. Therefore, Stage 2 of the L* life-cycle aims to determine the de facto control-flow model of the process that is analyzed. The process model may be discovered using the process discovery techniques. However, if there is a good process model present, it may be verified using conformance checking (activity check) or judged against the discovered model (activity compare). It is even possible to merge the handmade model and the discovered model (activity promote). After completing Stage 2 there is a control-flow model tightly connected to the event log, i.e., events in the event log refer to activities in the model. The output of Stage 2 may be used to answer questions, take actions, or to move to Stage 3. The output (control-flow model connected to an event log) needs to be interpreted before it can be used to answer questions or trigger a redesign, an adjustment, or an intervention.

Stage 3: Create Integrated Process Model

In Stage 3, the model is enhanced by adding additional perspectives to the control flow model (e.g., the organizational perspective, the case perspective, and the time perspective). The result is an integrated process model that can be used for various purposes. The model can be inspected directly to better understand the as-is process or to identify bottlenecks.

Moreover, a complete process model can also be simulated. The output of Stage 3 can also be used to answer selected questions and take appropriate actions (redesign, adjust, or intervene). Moreover, the integrated process model is also the input for Stage 4.
Fig 3.9 – Stages of Lasagna Process
Stage 4: Operational Support

Stage 4 of the L* life-cycle is concerned with the three operational support activities: detect, predict, and recommend. For instance, using short-term simulation or annotated transition systems it is possible to predict the remaining flow time for running cases. Stage 4 requires current data (“pre mortem” data on running cases) as input. Moreover, the output does not need to be interpreted by the process mining analyst and can be directly offered to end users. For example, a deviation may result in an automatically generated e-mail sent to the responsible manager. Recommendations and predictions are presented to the persons working on the corresponding cases. Note that operational support is the most ambitious form of process mining. This is only possible for Lasagna processes. Moreover, there needs to be an advanced IT infrastructure that provides high-quality event logs and allows for the embedding of an operational support system.

3.4.3 Process Mining Opportunities For Lasagna Process

The main functional areas that can be found in most organizations:

- **Product development** is concerned with all the preparations and engineering work needed to start producing a particular product. Products do not need to be physical objects (e.g., a car or copier); the product may also be a piece of information or a service (e.g., a new kind of insurance). Product development processes are typically Spaghetti-like because they have a lower frequency and depend on problem solving, expertise, and creativity rather than repetition, routine, and efficiency.

- **Production** is the functional area where the products are actually produced. Processes may range from classical manufacturing (assembling a car) to information creation (opening a back account). Most production processes are Lasagna processes because they need to be reproducible and efficient.

- **Procurement** entails all activities to get the materials needed for production. Note that the input for the production process may also be information from other parties. The input materials need to be purchased, stocks need to be monitored, deliveries need to be checked, etc. Processes in this functional area are typically Lasagna processes.
• **The functional area Sales/CRM** is concerned with all activities related to “lead-to-order” and “order-to-cash”. Besides the actual sales function, most organizations need to market their products and manage long-term relationships with their customers (CRM). Both Lasagna processes and Spaghetti processes can be found in this functional area. The handling of sales activities can be very structured whereas marketing-related activities may be rather unstructured.

• **Logistics** is concerned with the movements of products and materials, e.g., shipping the product to the customer and managing the storage space. Most processes in logistics are Lasagna processes.

• **The functional area Finance/accounting** deals with all financial aspects of an organization, e.g., billing customers, checking invoices, financial reporting, and auditing. Processes in this functional area are also typically Lasagna processes.

• **Resource management** is the functional area that makes sure there are sufficient resources to perform all other functions. HRM (Human Resource Management) is concerned with human resources and similar functions exist for machines, buildings, etc. Both Lasagna processes and Spaghetti processes can be found in this functional area, e.g., the handling of job applications may be very structured whereas the handling of a problematic employee may be rather ad-hoc.

• **The functional area Service** deals with all activities after the product has been shipped and paid for, e.g., activities related to product support, maintenance, repairing defective products, and help-desk operations. Service related processes are typically Spaghetti-like. Customers will use products in many different ways and repair processes are rather unpredictable for most products, e.g., no faults are found in the product returned by the customer or the wrong component is replaced and the product still malfunctions intermittently (Wil M.P. van der Aalst, 2010).

The process mining opportunities in different sectors and industries are described below:

The **primary sector** of the economy is concerned with transforming natural resources into primary products (e.g., agriculture, agribusiness, fishing, forestry and all mining and quarrying industries). Information technology tends to play a minor role in these industries. Hence, the application potential of process mining is limited.
The secondary sector of the economy refers to the manufacturing of tangible products and includes the automotive industry, chemical industry, aerospace manufacturing, consumer electronics, etc. Organizations in the secondary sector typically have an organizational structure covering all functional areas. Hence, both Lasagna processes and Spaghetti processes can be encountered.

The tertiary sector of the economy consists of all organizations that produce “intangible goods” such as services, regulations, and information. The term “services” should be interpreted in the broadest sense including transportation, insurance, wholesaling, retailing, entertainment, etc. Note that goods may be transformed in the process of providing the service (cf. preparing food in a restaurant). However, the focus is on serving the customer rather than transforming physical goods. In many industries in the tertiary sector, information plays a dominant role and many events are being recorded.

Process mining can be used to improve a variety of Lasagna and Spaghetti processes encountered in the tertiary sector. We sketch below some of the most interesting industries.

- The healthcare industry includes hospitals and other care organizations. Most events are being recorded (blood tests, MRI scans, appointments, etc.) and correlation is easy because each event refers to a particular patient. The closer processes get to the medical profession, the less structured they become. For instance, most diagnosis and treatment processes tend to be rather Spaghetti-like. Medical guidelines typically have little to do with the actual processes. On the one hand, this suggests that these processes can be improved by structuring them. On the other hand, the variability of medical processes is caused by the different characteristics of patients, their problems, and unanticipated complications.

- Governments range from small municipalities to large organizations operating at the national level, e.g., institutions managing processes related to unemployment, customs, taxes, and traffic offences. Both local and national government agencies can be seen as “administrative factories” as they execute regulations and the “products” are mainly informational or financial. Processes in larger government agencies are characterized by a high degree of automation. Consider, for example, tax departments that need to deal with millions of tax declarations. Processes in smaller government agencies (e.g., small municipalities) are typically not automated and managed by office workers.
rather than BPM systems. However, due to the legal requirements, all main events are recorded in a systematic manner.

- **Banking and insurance** are two industries where BPM technology has been most effective. Processes are often automated and all events are recorded in a systematic and secure manner. Examples are the processing of loans, claims management, handling insurance applications, credit card payments, and mortgage payments. Most processes in banking and insurance are Lasagna processes, i.e., highly structured. Hence, all of the techniques discussed can be applied. Process discovery is less relevant for these organizations as most processes are known and documented. Typical use cases in these industries involve conformance checking, performance analysis, and operational support.

- **Organizations involved in education** (e.g., high-schools and universities) are recording more and more information related to the study behavior of individuals. For instance, at TU/e (Technical University of Eindhoven) we are applying process mining to analyze study behavior using a database containing detailed information about exam results of all students who studied computer science. Moreover, this database also contains information about high-school exam grades, etc. Some of these educational processes are structured, others are much unstructured.

  For example, it is very difficult to predict the remaining study time of students at a university because the curriculum often changes and students tend to have very different study patterns. Nevertheless, valuable insights can be obtained. By visualizing that few students follow the courses in the order intended, one can show that the design of a curriculum should not only focus on the “ideal student”, but also anticipate problems encountered by other students.

- **The products manufactured by organizations** in the secondary sector are distributed through various retail organizations. Here it is interesting to see that more and more information about products and customers is being recorded. Customers are tracked using loyalty cards or through online profiles. Products are tagged and the shop has real-time information about the number of items still available. A product that has an
RFID tag has a unique identifier, i.e., two identical products can still be distinguished. This allows for the correlation of events and thus facilitates process mining.

- **The transportation** industry is also recording more and more information about the movement of people and products. Through tracking and tracing functionality the whereabouts of a particular parcel can be monitored by both sender and receiver. Although controversial, smartcards providing access to buildings and transportation systems can be used to monitor the movement of people.

- **The capital goods** industry is also transforming from the situation in which customers purchase expensive machines to the situation in which customers only pay for the actual use of the machine.

### 3.4.4 Spaghetti Process

Spaghetti processes are the counterpart of Lasagna processes. Because Spaghetti processes are less structured, only a subset of the process mining techniques described are applicable. For instance, it makes no sense to aim at operational support activities if there is too much variability. Nevertheless, process mining can help to realize dramatic process improvements by uncovering key problems.

### 3.4.5 L*Lifecycle model for Spaghetti process

Only the initial stages of the Lasagna process are applicable for Spaghetti processes. Note that Stage 4 has been removed because operational support is impossible for the processes just described. To enable history-based predictions and recommendations, it is essential to first make the “Spaghetti-like” process more “Lasagna-like” In fact, Stage 3 will also be too ambitious for most Spaghetti processes. Moreover, it is also possible to view dotted charts, create social networks, etc. However, it is very unlikely that all of these can be folded into a meaningful comprehensive process model as the basis (the control-flow discovered) is too weak.
3.4.6 Process Mining Opportunities for Spaghetti Processes

- Processes in Product development tend to be rather unstructured because they are low frequent (compared to production processes) and rely on creativity and problem-solving capabilities.

- An interesting development is that more and more products are monitored while being used in their natural habitat, e.g., modern high-end copiers, expensive medical devices, and critical production facilities collect event logs and can be observed remotely.

- Resource management and Sales/CRM are two functional areas where a mixture of Spaghetti and Lasagna processes can be encountered.

Process mining can be seen as the “missing link” between data mining and traditional model-driven BPM. Process mining is an important tool for modern organizations that need to manage nontrivial operational processes. On the one hand, there is an incredible growth of event data. On the other hand, processes and information need to be aligned perfectly in order to meet requirements related to compliance, efficiency, and customer service. The digital universe and the physical universe are amalgamating into one universe where events are recorded as they happen and processes are guided and controlled, based on event data (Wil M.P. van der Aalst, 2010).

Process mining can be brought into play for many different purposes. Process mining can be used to diagnose the actual processes. This is valuable because in many organizations most stakeholders lack a correct, objective, and accurate view on important operational processes. Process mining can subsequently be used to improve such processes. Conformance checking can be used for auditing and compliance. By replaying the event log on a process model, it is possible to quantify and visualize deviations. Similar techniques can be used to detect bottlenecks and build predictive models.
Fig 3.10 - Stages of Spaghetti Process
3.5 Summary

This chapter begins with the explanation on data explosion and then discussed the various process models. It further listed the uses of process models. The life cycle process model of BPM has been discussed. The three types process mining – Discovery, Conformance and Enhancement are described followed by a detailed note on event logs. There are two types of processes – Lasagna and Spaghetti. The lifecycle models of these processes are described. The insurance claim process considered in this thesis is a Lasagna process. The various perspectives of process mining – control flow perspective, organizational perspective, case perspective and time perspective are also dealt with in this chapter.