

Манифест Process Mining

(финальная версия)

Wil van der Aalst, Arya Adriansyah, Ana Karla Alves de Medeiros, Franco Arcieri, Thomas Baier, Tobias Blickle, Jagadeesh Chandra Bose, Peter van den Brand, Ronald Brandtjen, Joos Buijs, Andrea Burattin, Josep Carmona, Malu Castellanos, Jan Claes, Jonathan Cook, Nicola Costantini, Francisco Curbera, Ernesto Damiani, Massimiliano de Leoni, Pavlos Delias, Boudewijn van Dongen, Marlon Dumas, Schahram Dustdar, Dirk Fahland, Diogo R. Ferreira, Walid Gaaloul, Frank van Geffen, Sukriti Goel, Christian Günther, Antonella Guzzo, Paul Harmon, Arthur ter Hofstede, John Hoogland, Jon Espen Ingvaldsen, Koki Kato, Rudolf Kuhn, Akhil Kumar, Marcello La Rosa, Fabrizio Maggi, Donato Malerba, Ronny Mans, Alberto Manuel, Martin McCreesh, Paola Mello, Jan Mendling, Marco Montali, Hamid Motahari Nezhad, Michael zur Muehlen, Jorge Munoz-Gama, Luigi Pontieri, Joel Ribeiro, Anne Rozinat, Hugo Seguel Pérez, Ricardo Seguel Pérez, Marcos Sepúlveda, Jim Sinur, Pnina Soffer, Minseok Song, Alessandro Sperduti, Giovanni Stilo, Casper Stoel, Keith Swenson, Maurizio Talamo, Wei Tan, Chris Turner, Jan Vanthienen, George Varvaressos, Eric Verbeek, Marc Verdonk, Roberto Vigo, Jianmin Wang, Barbara Weber, Matthias Weidlich, Ton Weijters, Lijie Wen, Michael Westergaard, and Moe Wynn

IEEE Task Force on Process Mining*

<http://www.win.tue.nl/ieeetfpm/>

Аннотация Методы process mining (методы глубинного анализа процессов) позволяют *извлекать знания из журналов событий*, доступных в современных информационных системах. С помощью данных методов появляются новые возможности *извлечения, мониторинга и улучшения процессов* в различных областях применения. Растущий интерес к process mining обусловлен двумя основными факторами. С одной стороны, все больше событий фиксируется в цифровом виде, что позволяет получить детальную информацию об истории различных процессов. С другой стороны, существует потребность в улучшении и сопровождении бизнес-процессов в высококонкурентной и быстро изменяющейся среде. Данный манифест создан рабочей группой *IEEE Task Force on Process Mining* и нацелен на продвижение тематики process mining. Кроме того, изложенный здесь набор руководящих принципов, наряду с перечнем актуальных вызовов, позволяет надеяться, что манифест послужит *руководством для разработчиков, ученых, консультантов, менеджеров и пользователей*. Целью манифеста является развитие process mining в качестве нового инструмента совершенствования возможностей (пере)проектирования, контроля и сопровождения операционных бизнес-процессов.

1 Рабочая группа IEEE Task Force on Process Mining

Манифест – это «публичное выражение принципов и намерений» группой людей. Данный манифест написан членами и сторонниками *IEEE Task Force on Process Mining*. Цель этой группы – продвижение исследований, разработок, образовательной деятельности, внедрения, эволюции и понимания process mining.

Process mining – сравнительно молодая область исследований, которая находится между направлениями исследований вычислительного интеллекта (computational intelligence) и глубинным анализом данных (data mining), с одной стороны, и моделированием и анализом процессов – с другой. Суть process mining заключается в *извлечении, мониторинге и*

* Оригинальная версия опубликована в *BPM 2011 Workshops proceedings, Lecture Notes in Business Information Processing*, Springer-Verlag, 2011

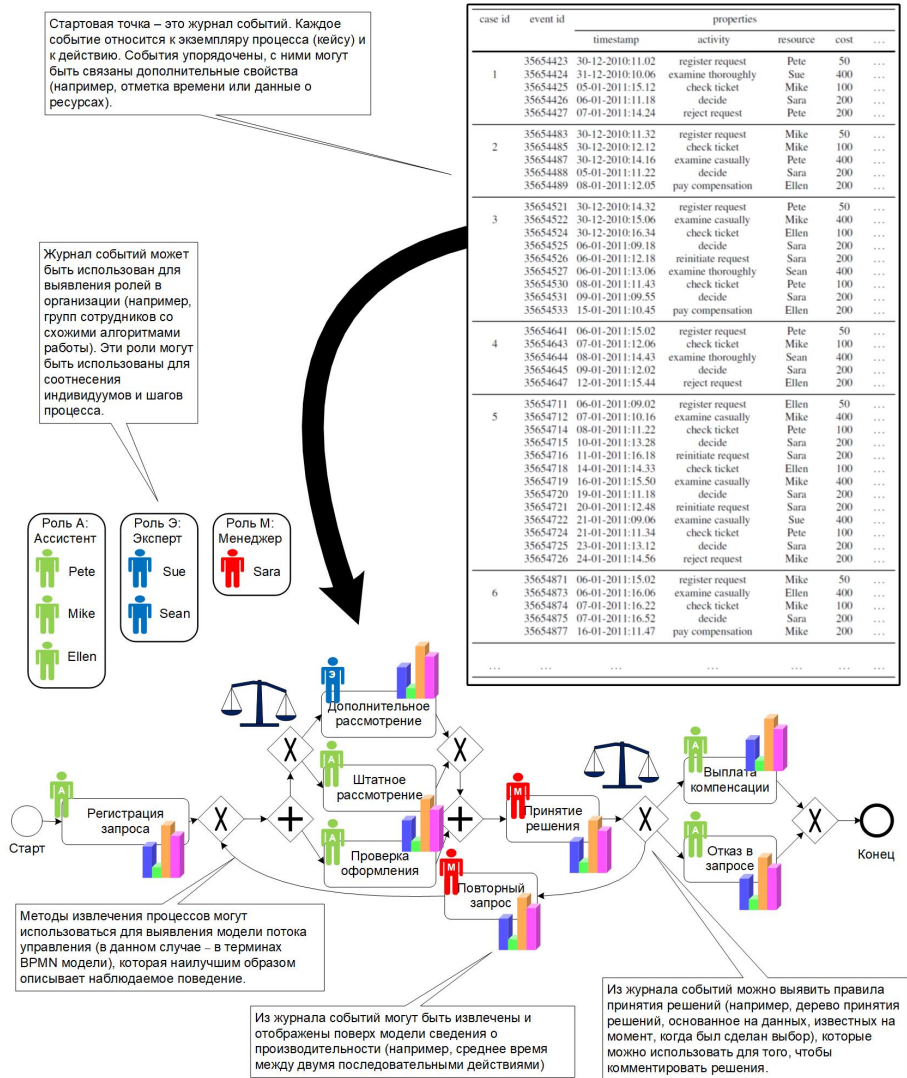


Рис. 1. Методы process mining позволяют получить новые знания из журналов событий для целей извлечения, мониторинга и улучшения процессов.

усовершенствовании реальных процессов (т.е., не предполагаемых, а реально существующих процессов) с помощью извлечения новых знаний из журналов событий, которые доступны в современных (информационных) системах (см. Рис. 1). Process mining включает в себя (автоматизированное) извлечение процессов (т.е. получение моделей процессов из данных журналов событий), проверку соответствия (т.е. мониторинг отклонений при сравнении модели и журнала событий), извлечение сведений о социальных сетях и организационном ракурсе процесса, автоматизированное построение имитационных моделей, расширение и исправление моделей, предиктивную аналитику и выработку рекомендаций на основе исторических данных процессов.

Process mining является важным связующим элементом областей data mining с одной стороны и моделирования и анализа бизнес-процессов – с другой стороны. В последнее время появилось немало громких аббревиатур в рамках общего «зонтичного» термина Business Intelligence (BI), которые, по сути, обозначают достаточно простые инструменты построения отчетов и информационных панелей. В частности, Business Activity Monitoring (BAM) – класс технологий для мониторинга бизнес-процессов в реальном времени. Complex Event

Processing (СЕР), обработка сложных событий, – технологии для обработки больших количеств событий, анализа паттернов и обнаружения сложных событий с целью мониторинга, контроля и оптимизации бизнеса в реальном времени. *Corporate Performance Management* (СРМ), система управления эффективностью организации, – еще одно громкое наименование для измерения эффективности процессов организации. Кроме того, можно отметить такие управленческие подходы, как *Continuous Process Improvement* (СРІ) – непрерывное усовершенствование процессов, *Business Process Improvement* (ВРІ) – улучшение бизнес-процессов, *Total Quality Management* (ТQМ) – концепция всеобщего управления качеством, а также *Six Sigma* – концепция «6 сигма». Общим для этих подходов является то, что процессы изучаются «под микроскопом» с целью получения информации о возможностях их улучшения. Process mining – это технология, которая открывает новые возможности для применения СРМ, ВРІ, ТQМ, Six Sigma и тому подобных.

Несмотря на то, что инструменты ВІ и управленческие подходы наподобие Six Sigma и ТQМ нацелены на улучшение операционной эффективности (например, уменьшение времени обработки или снижение количества брака), организации делают все больший упор на корпоративное управление (*corporate governance*), управление *рисками* и комплаенс (*compliance*). Важность этих требований можно проиллюстрировать, например, законодательными актами, такими, как закон Сарбейнза-Оксли (Sarbanes-Oxley Act, SOX) и соглашение «Базель II» (Basel II Accord). Методы process mining позволяют скрупулезно контролировать соответствие внутренних процессов организации требованиям внутреннего и внешнего контроля и обеспечить валидность и надежность информации об этих процессах.

За последние десять лет данные о событиях стали более доступными. Одновременно с этим сформировались технологии process mining. Более того, как упоминалось выше, тренды в менеджменте организации, связанные с улучшением процессов (например, Six Sigma, ТQМ, СРІ и СРМ) и комплаенсом (SOX, ВАМ и т.п.), могут быть поддержаны и усилены технологией process mining. К счастью, алгоритмы process mining реализованы в различных академических и коммерческих программных продуктах. В настоящее время активно действует сообщество ученых, работающих с управлением process mining, и это направление стало одной из горячих тем в развитии управления бизнес-процессами. Более того, интерес к process mining проявляется и коммерческий сектор. Все больше вендоров программного обеспечения добавляет функциональность process mining в свои продукты. В качестве примеров можно назвать: ARIS Process Performance Manager (Software AG), Comprehend (Open Connect), Discovery Analyst (StereoLOGIC), Flow (Fourspark), Futura Reflect (Futura Process Intelligence), Interstage Automated Process Discovery (Fujitsu), OKT Process Mining suite (Exeura), Process Discovery Focus (Iontas/Verint), ProcessAnalyzer (QPR), ProM (TU/e), Rbminer/Dbminer (UPC), и Reflect|one (Pallas Athena). Растущий интерес к анализу процессов на основе записей журналов событий послужил движущей силой для создания рабочей группы Task Force on Process Mining.

Рабочая группа была создана в 2009 году в рамках Data Mining Technical Committee (DMTC) при обществе Computational Intelligence Society (CIS) Института инженеров электротехники и электроники (Institute of Electrical and Electronic Engineers, IEEE). В текущем составе группы представлены *вендоры программного обеспечения* (например, Pallas Athena, Software AG, Futura Process Intelligence, HP, IBM, Infosys, Fluxicon, Businesscape, Iontas/Verint, Fujitsu, Fujitsu Laboratories, Business Process Mining, Stereologic), *консалтинговые компании / пользователи* (например ProcessGold, Business Process Trends, Gartner, Deloitte, Process Sphere, Siav SpA, BPM Chile, BWI Systeme GmbH, Excellentia BPM, Rabobank), и *научно-исследовательские организации* (например, TU/e, University of Padua, Universitat Politècnica de Catalunya, New Mexico State University, IST - Technical University of Lisbon, University of Calabria, Penn State University, University of Bari, Humboldt-Universität zu Berlin, Queensland University of Technology, Vienna University of Economics and Business, Stevens Institute of Technology, University of Haifa, University of Bologna, Ulsan National Institute of Science and Technology, Cranfield University, K.U. Leuven, Tsinghua University, University of Innsbruck, University of Tartu).

Целями рабочей группы являются:

- распространение среди пользователей, разработчиков, консультантов, руководителей бизнеса и ученых информации о современном состоянии дел в области process mining,
- продвижение использования методов и инструментов process mining, стимулирование инноваций в применении,
- влияние на процесс стандартизации журналов событий,
- организация обучающих семинаров, специальных сессий, воркшопов, панельных дискуссий,
- публикация статей, книг, видео и специальных выпусков журналов.

Начиная с 2009 года группой были предприняты различные действия для достижения указанных целей. В частности, при участии группы были организованы несколько воркшопов и докладов, например, Business Process Intelligence (BPI'09, BPI'10, и BPI'11) и отдельные доклады на основных конференциях IEEE (например, CIDM'11). Осуществлялось распространение знаний в виде учебных пособий (например, WCCG'10 и PMPM'09), летних школ (ESSCaSS'09, ACPN'10, CICH'10, etc.), видео (см. www.processmining.org), и публикаций, включая первую книгу о process mining, опубликованную издательством Springer.¹ Группа также принимала участие в организации первого Business Process Intelligence Challenge (BPIC'11): состязания, в котором участники извлекают новые знания из больших и сложных журналов событий. В 2010, группа также ввела стандарт *XES* (www.xes-standard.org), содержащий требования к структуре журнала событий. Этот стандарт является расширяемым и поддерживается *OpenXES library* (www.openxes.org), а также такими программными инструментами, как ProM, XESame, Nitro, etc.

Читатель может ознакомиться с ресурсом <http://www.win.tue.nl/ieeetfpm/> для получения дополнительной информации о деятельности группы.

¹ W.M.P. van der Aalst. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag, Berlin, 2011. <http://www.processmining.org/book/>

2 Process mining: современное состояние

Развитие возможностей информационных систем и всего, что напрямую связано с вычислительными мощностями, хорошо описано Законом Мура. Со-основатель компании Intel Гордон Мур в 1965 году предсказал, что количество транзисторов, размещаемых на кристалле интегральной схемы, будет удваиваться каждый год. В последующие 50 лет этот рост действительно имел экспоненциальный характер, хотя и не настолько стремительный. Это, в свою очередь, привело к бурному росту «цифровой вселенной» (т.е. количества хранимых и передаваемых цифровых данных). Более того, цифровой и реальный миры становятся все сильнее связанными друг с другом. Рост той составляющей «цифровой вселенной», которая увязана с процессами организации, позволяет записывать данные о происходящих событиях и анализировать их. Под событиями подразумевается самый широкий набор фиксируемых в электронном виде действий, таких, как снятие денег в банкомате, настройка рентгеновского аппарата специалистом, подача документов на права, уплата налогов или получение электронного авиабилета. Задачей анализа является интерпретация данных событий таким образом, чтобы получить новую информацию, определить узкие места, предупредить возникновение проблем, определить отклонения от правил, выработать рекомендации по устранению нарушений и обеспечить оптимальное исполнение процесса. Именно на это направлен process mining.

Отправной точкой process mining является *журнал событий*. Все методы process mining предполагают возможность получения *последовательно* записанных *событий*, таких что каждое событие относится к *действию* и является частью *кейса* (т.е. экземпляра процесса). Журналы событий могут содержать дополнительную информацию о событиях. По сути, при малейшей возможности методы process mining позволяют использовать дополнительные атрибуты событий для анализа, такие как *ресурс* (например, исполнитель или задействованное оборудование), *временная отметка* события или *элементы данных* записанные для события (например, сумма заказа).

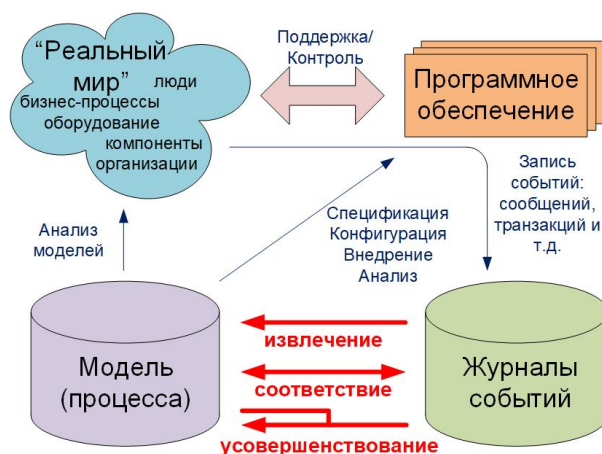


Рис. 2. Позиционирование трех основных форм process mining: (а) *извлечение*, (б) *проверка соответствия* и (в) *усовершенствование*.

Как показано на Рис. 2, журналы событий могут быть использованы для осуществления трех форм process mining. Первая форма – *извлечение* процесса. Методы извлечения процессов используют журнал событий для получения модели процесса без использования какой-либо априорной информации. Извлечение процессов является самой значимой составляющей process mining. Для многих организаций поразительным является сам факт того, что существующие алгоритмы позволяют извлечь модели реально действующих процессов

просто за счет обработки записей событий. Вторая форма process mining – *проверка соответствия*. Здесь производится сопоставление существующей модели процесса с журналом событий этого же процесса. Проверка соответствия может быть использована для оценки того, насколько реальные данные журнала соответствуют модели, и наоборот. Необходимо отметить, что можно рассматривать разные виды моделей: проверка соответствия может быть применена и к процедурным моделям, и к организационным, и к декларативным, а также к внутренним правилам и распоряжкам, законодательству и проч. Третья форма process mining – это *усовершенствование* процесса. В данном случае существующую модель процесса улучшают с использованием информации о реально осуществляемом процессе, зафиксированном в каком-либо журнале событий. Если проверка соответствия сравнивает модель и реальность, то здесь речь идет об изменении и/или расширении априорной модели процесса. В частности, с использованием сведений временных отметок журнала, можно расширить априорную модель сведениями о частотных и временных характеристиках процесса и его шагов, указать узкие места.

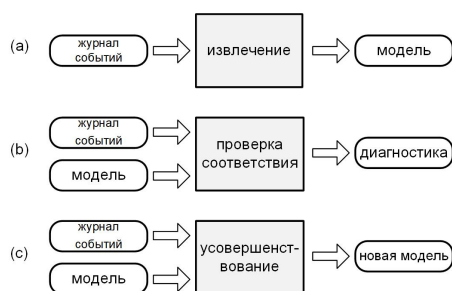


Рис. 3. Три основных формы process mining в терминах «входов» и «выходов»: (а) извлечение, (б) проверка соответствия и (в) усовершенствование.

Рисунок 3 описывает три формы process mining в терминах «входов» и «выходов». Методы извлечения процесса позволяют сформировать модель процесса на основе журнала событий. Полученная модель, как правило, является моделью именно процесса (например, сеть Петри, BPMN, EPC или UML-диаграмма деятельности), но, помимо этого, модель может также формализовать и другие ракурсы (например, социальную сеть). Проверка соответствия на входе использует журнал событий и модель. На выходе выдается диагностическая информация, указывающая на отклонения или соответствие модели и журнала событий. Усовершенствование модели (уточнение или расширение) также на входе использует журнал событий и модель. На выходе предоставляется улучшенная или расширенная модель. Process mining может показать разные ракурсы процесса. *Ракурс потока управления* показывает порядок и структуру выполнения действий. Цель извлечения такого ракурса – получение описания и характеристик всех используемых путей осуществления процесса. Результат такого извлечения процесса обычно отображается в виде сетей Петри либо в другой нотации моделирования процессов (например, EPCs, BPMN или UML-диаграмма деятельности). *Организационный ракурс* отражает информацию о ресурсах, указанных в журнале событий, т.е. какие акторы вовлечены в процесс (например, люди, системы, роли или организационные единицы) и как они взаимосвязаны. Целью построения такого ракурса является либо структурирование организации путем классификации участников (людей) в терминах ролей и организационных единиц, либо извлечение социальной сети. *Ракурс кейсов* фокусируется на свойствах самих кейсов. Очевидно, кейс может быть охарактеризован либо путем в структуре процесса либо акторами, которые с ним работают. Тем не менее, кейсы также могут быть описаны через значения связанных с ними элементов данных. Например, если кейс связан с заказом на пополнение запасов, то интересно получить информацию о поставщике или количестве заказанных единиц товара. *Временной ракурс* связан с показателями времени и частоты

событий. Если у событий есть временные отметки, то возможно обнаружить узкие места, измерить уровни сервиса, отслеживать использование ресурсов и предсказывать остающееся время обработки активных кейсов.

Существует несколько распространенных заблуждений, связанных с process mining. Ряд вендоров, аналитиков и ученых узко трактуют process mining как один из методов глубинного анализа данных (data mining) для извлечения процесса, используемый исключительно для анализа в режиме офлайн. Это *неверно*, и поэтому ниже мы подчеркнем следующие три характерные особенности.

- *Process mining не ограничивается извлечением моделей потока управления.* Извлечение моделей процесса из журналов событий разжигает воображение как специалистов-практиков, так и представителей академического сообщества. Поэтому извлечение моделей потока управления зачастую рассматривается как наиболее примечательная составляющая process mining. Однако, process mining не ограничен извлечением потока управления. С одной стороны, извлечение модели – это лишь одна из трех основных форм process mining (извлечение, подтверждение и усовершенствование). С другой стороны, область применения не сводится только к потоку управления: важную роль также играют организационный, временной ракурсы и ракурс кейсов.
- *Process mining не является просто еще одним видом data mining.* Process mining может рассматриваться как «недостающее звено» между data mining и традиционным управлением процессами (BPM). Большинство методов data mining вовсе не процессно-ориентированы. Модели процессов, потенциально демонстрирующие параллелизм, не сравнимы с простыми структурами data mining, такими как дерево принятия решений и ассоциативные правила. Поэтому необходимы совершенно новые алгоритмы и формы представления данных.
- *Process mining не ограничен офлайн-анализом.* Методы process mining извлекают знания из накопленных данных о событиях. Несмотря на то, что используются данные постфактум, результаты применимы и к тем кейсам, которые еще находятся в обработке. Например, используя выявленную модель процесса, можно спрогнозировать время завершения обработки заказа клиента.

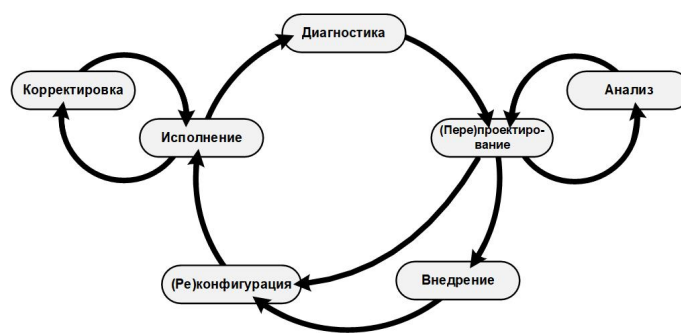


Рис. 4. Жизненный цикл BPM, описывающий различные фазы бизнес-процессов и соответствующих информационных систем; process mining (потенциально) играет роль на всех фазах (не считая фазы внедрения).

Чтобы позиционировать process mining, мы используем жизненный цикл управления бизнес-процессами (BPM), приведенный на Рис. 4. Жизненный цикл BPM содержит семь фаз бизнес-процессов и связанных с ними информационных систем. На *фазе (пере)проектирования* либо создается новая модель процесса, либо изменяется существующая. Следующая за ней – *фаза анализа* предлагаемой новой модели и ее альтернатив. После фазы (пере)проектирования следует *фаза внедрения* модели, в ином случае существующая система (ре)конфигурируется

(фаза (ре)конфигурации). На фазе исполнения производится запуск созданной модели. Во время выполнения этой фазы производится *мониторинг* модели. Кроме того, могут быть произведены небольшие корректировки, не требующие изменения процесса (фаза *корректировки*). На фазе *диагностики* запущенный процесс анализируется, и результаты анализа могут инициировать повторный запуск фазы (ре)инжиниринга. Process mining является полезным инструментом для реализации большинства приведенных на Рис. 4 фаз. Очевидно, что прежде всего это фаза диагностики; тем не менее, применение process mining не ограничено лишь диагностикой процессов. Например, методы process mining можно использовать на фазе исполнения для *операционного сопровождения*. Прогнозы и рекомендации, основанные на извлеченных моделях, можно использовать для воздействия на обрабатываемые кейсы. Похожие формы поддержки принятия решений можно использовать также для корректировки процессов и для управления (ре)конфигурацией процессов.

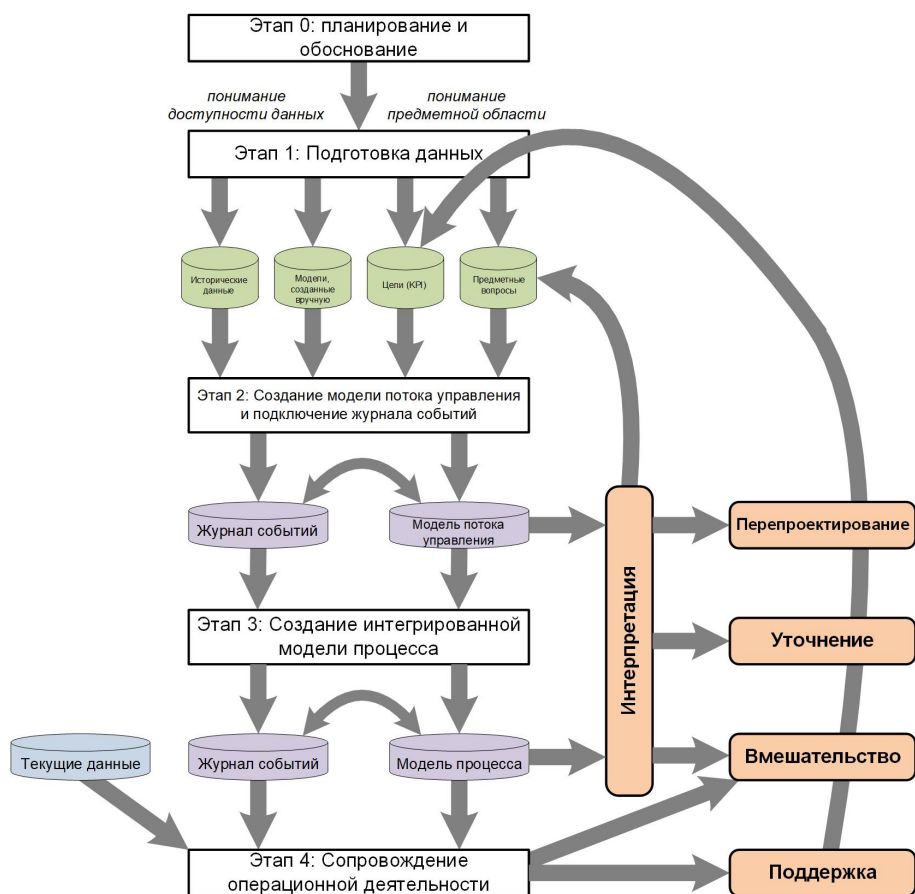


Рис. 5. Модель жизненного цикла L^* формализует проект process mining через пять этапов: планирование и обоснование (Этап 0), подготовка данных (Этап 1), создание модели потока управления и подключение журнала событий (Этап 2), создание интегрированной модели процесса (Этап 3), сопровождение операционной деятельности (Этап 4).

Тогда как Рис. 4 демонстрирует общий жизненный цикл BPM, Рис. 5 фокусируется на конкретных действиях и артефактах process mining. Рисунок 5 описывает возможные этапы проекта process mining. Любой такой проект начинается с этапа планирования и обоснования для формирования плана (Этап 0). После инициирования проекта, данные событий, модели, цели и вопросы необходимо получить из систем, от экспертов в предметных областях и у

менеджмента (Этап 1). Это, в свою очередь, требует осведомленности о доступных данных («Какие данные могут быть использованы для анализа?») и понимания предметной области («Какие вопросы важно задать?»). Результатом этапа являются артефакты, приведенные на Рис. 5 (т.е. накопленные исторические данные, созданные вручную модели, цели и вопросы). На Этапе 2 формируется модель потока управления, которая связывается с журналом событий. На данном этапе можно использовать автоматизированные методы извлечения моделей процессов, и полученная модель процесса может либо позволить сразу ответить на некоторые вопросы, либо инициировать реинжиниринг или корректирующие действия. Более того, возможно отфильтровать или адаптировать журнал событий с использованием полученной модели (например, убирая из него редкие действия, кейсы или дополняя его отсутствующими событиями). Иногда требуются значительные усилия для того, чтобы связать между собой события, относящиеся к одному и тому же экземпляру процесса. Оставшиеся события относятся к сущностям модели процесса. Когда процесс относительно структурирован, модель потока управления может быть расширена при помощи других ракурсов (например, данные, время, ресурсы) во время Этапа 3. Связь между журналом событий и моделью, установленная на Этапе 2, используется для расширения модели (например, временные отметки или связанные события можно использовать для расчета времени ожидания для действий). Такое расширение было использовано для получения новых ответов на вопросы, и в результате также могут быть сразу инициированы какие-либо дополнительные действия. Наконец, созданные в рамках Этапа 3 модели могут быть использованы для операционного сопровождения (Этап 4). Знания, извлеченные из накопленных исторических данных событий, сопоставляются с информацией обрабатываемых кейсов. Результаты можно использовать для корректировки, прогнозирования и подготовки рекомендаций. Этапы 3 и 4 можно реализовать только в том случае, если процесс достаточно стабилен и структурирован.

В настоящее время существуют методы и инструменты для поддержки осуществления всех этапов, представленных на Рис. 5. Тем не менее, process mining – сравнительно новая парадигма, и, как следствие, большинство доступных инструментов находятся на ранних стадиях развития. Более того, предполагаемые пользователи этих инструментов зачастую не осведомлены о реальном потенциале и ограничениях process mining. С учетом этого, в настоящем манифесте перечислены руководящие принципы (см. Раздел 3) и вызовы (см. Раздел 4) для пользователей методов process mining, а также для ученых и разработчиков, заинтересованных в развитии этой области.

3 Руководящие принципы

Существуют очевидные ошибки, которые можно совершить, используя process mining на практике – как и со всякой новой технологией. Принимая это во внимание, мы приводим здесь шесть *руководящих принципов*, которые позволят пользователям и аналитикам избежать подобных ошибок.

3.1 РП1: Данные о событиях должны считаться объектами первого класса

Стартовой точкой для какого-либо действия в process mining являются зарегистрированные события. Такие «сборники событий» обозначаются термином *журналы событий*, хотя это и не означает, что события должны храниться в специализированных файлах-логах. События могут храниться в таблицах баз данных, логах сообщений, почтовых архивах, логах транзакций и других источниках данных. Более важным, чем формат хранения, является *качество* логов этих событий. Качество результатов process mining напрямую зависит от данных на входе. Следовательно, к журналам событий следует относиться как к *объектам первого класса* информационных систем, в которых реализованы процессы, которые предполагается анализировать. К сожалению, зачастую журналы событий являются лишь «побочными продуктами», используемыми при отладке или профилировании. Например, медицинское оборудование Philips Healthcare регистрирует события только потому, что разработчики программного обеспечения вставили в код команду «print statements». Несмотря

на то, что существуют некоторые неформальные методические рекомендации добавлять подобные выражения в код, нужен более системный подход к улучшению качества журналов событий. Данные о событиях должны рассматриваться как объекты первого класса (вместо того, чтобы считаться объектами второго класса).

Существует несколько критериев для оценки качества данных о событиях. События должны быть *достоверными*, т.е. надо быть уверенным в том, что зарегистрированные события действительно произошли и что характеристики событий корректны. Журналы событий должны быть *полными*, т.е., с учетом конкретного контекста, ни одно событие не должно быть пропущено. Каждое зарегистрированное событие должно иметь четко определенную *семантику*. Кроме того, данные о событиях должны быть в *безопасности* в том смысле, что при регистрации событий должны быть учтены вопросы приватности и защиты данных. Например, исполнители должны быть осведомлены о том, какого рода события регистрируются и как они используются.

Таблица 1 описывает пять уровней зрелости журналов событий, варьирующихся от великолепного качества (★★★★) до низкого качества (*). Например, журналы событий компании Philips Healthcare находятся на уровне ★★★, т.е. события регистрируются автоматически и регистрируемое поведение соответствует реальности, но нет системного подхода к присвоению событиям семантики и к обеспечению охвата на требуемом уровне. Методы process mining применимы с журналами уровня ★★★★★, ★★★ и ★★★. Принципиально методы process mining применимы и с журналами уровня ** или *. Однако, анализ таких логов обычно проблематичен, а результаты не достоверны. На самом деле, применять process mining с журналами * не имеет особого смысла.

Чтобы получать выгоды от process mining, организациям следует стремиться к максимально высокому качеству журналов событий.

3.2 РП2: Журналы должны извлекаться в соответствии с предметными вопросами

Как показано на Рис. 5, действия process mining должны быть обусловлены вопросами. Крайне затруднительно извлечь осмысленные данные о событиях без предметных вопросов. Рассмотрим, например, тысячи таблиц в базе данных ERP системы вроде SAP: без конкретных вопросов невозможно отобрать релевантные таблицы для извлечения данных.

Модель процесса, как показано на Рис. 1, описывает жизненный цикл кейсов (т.е., экземпляров процессов) определенного типа. Следовательно, прежде чем применять какой-либо метод process mining, необходимо определиться с типом кейсов для анализа. Этот выбор должен быть продиктован вопросами, на которые надо будет ответить, и это может оказаться нетривиальной задачей. Рассмотрим, например, обработку заказов клиентов. Каждый заказ клиента может состоять из нескольких строк (позиций), так как одним заказом клиент может заказать несколько продуктов. Один заказ может быть доставлен за несколько доставок. Одна доставка может соотноситься со строками (позициями) нескольких заказов. Таким образом, между заказами и доставками существует отношение «многие ко многим», а между заказами и строками заказов – отношение «один ко многим». При наличии базы данных о событиях, относящихся к заказам, строкам (позициям) и доставкам, могут быть извлечены разные модели процесса. Можно извлечь данные с целью описания жизненного цикла отдельно взятых заказов. Однако, также возможно извлечение данных с целью описания жизненного цикла отдельно взятой строки (позиции) заказа или жизненного цикла отдельно взятых доставок.

3.3 РП3: Необходимость поддержки операторов параллелизма, выбора и других базовых элементов моделей потока управления

Существует множество нотаций описания процессов (например, BPMN, EPCs, сети Петри, PREL и UML-диаграммы деятельности). Некоторые из этих нотаций предоставляют большое

Таблица 1. Уровни зрелости журналов событий.

Уровень	Характеристика
*****	Наивысший уровень: журнал событий великолепного качества (т.е., достоверный и полный), и события четко определены. События регистрируются в автоматическом, систематическом, надежном и безопасном режиме. Надлежащим образом обеспечены конфиденциальность и безопасность. Регистрируемые события (и все их атрибуты) обладают понятной семантикой, что подразумевает наличие одной или нескольких онтологий, а события и атрибуты указывают на эту онтологию. <i>Пример:</i> семантически аннотированные журналы систем BPM.
****	События регистрируются автоматически и в систематическом и надежном режиме, т.е. логи достоверны и полны. В отличие от систем, оперирующих на уровне ***, явным образом поддерживаются такие сущности, как экземпляр процесса (кейс) и действие. <i>Пример:</i> журналы событий традиционных BPM систем/систем документооборота.
***	События регистрируются автоматически, но нет системного подхода к регистрации событий. Однако, в отличие от журналов уровня **, есть некоторый уровень гарантии того, что регистрируемые события соотносятся с реальностью (т.е., журнал событий достоверен, но не обязательно полон). Рассмотрим, например, события, регистрируемые в системе ERP. Несмотря на то, что приходится извлекать события из многочисленных таблиц, информация в них, предположительно, корректна (например, наверняка можно предположить, что платеж, зарегистрированный в ERP, действительно существует, и наоборот). <i>Примеры:</i> таблицы в системах ERP, журналы событий в системах CRM, логи транзакций систем обмена сообщениями, журналы событий высокотехнологичного оборудования и т.д.
**	События регистрируются автоматически в режиме «побочного продукта» какой-либо информационной системы. Степень покрытия различна, нет системного подхода для принятия решений о том, какие события следует регистрировать. Кроме того, есть возможность обходить информационную систему. Следовательно, события могут отсутствовать, либо они могут быть неправильно зарегистрированы. <i>Примеры:</i> журналы событий системы управления документами или продуктами, журналы ошибок встроенных систем, рабочие таблицы сервисных инженеров и т.д.
*	Низший уровень: журналы событий низкого качества. Зарегистрированные события могут не соответствовать реальности, часть событий может отсутствовать. Обычно такими характеристиками обладают журналы событий, в которых события регистрируются вручную. <i>Примеры:</i> закладки («желтые стикеры»), заложенные в бумажных документах, перемещающиеся по организации, бумажные медицинские карты и т.д.

количество элементов для моделирования (например, в BPMN более 50 различных графических элементов), тогда как другие могут быть очень простыми (сети Петри конструируют только из трех элементов: позиции, переходы и дуги). Описание потока управления – основа любой модели процесса. Базовые конструкции потока работ (также известные как *паттерны*) поддерживаются всеми распространенными нотациями: это последовательное выполнение, параллельное выполнение («И»), выбор (XOR-операторы исключающего «ИЛИ») и циклы. Очевидно, что эти паттерны должны поддерживаться методами process mining. Тем

не менее, некоторые методы не применимы к ситуациями, где существует параллелизм, и поддерживают только цепи Маркова/системы переходов.

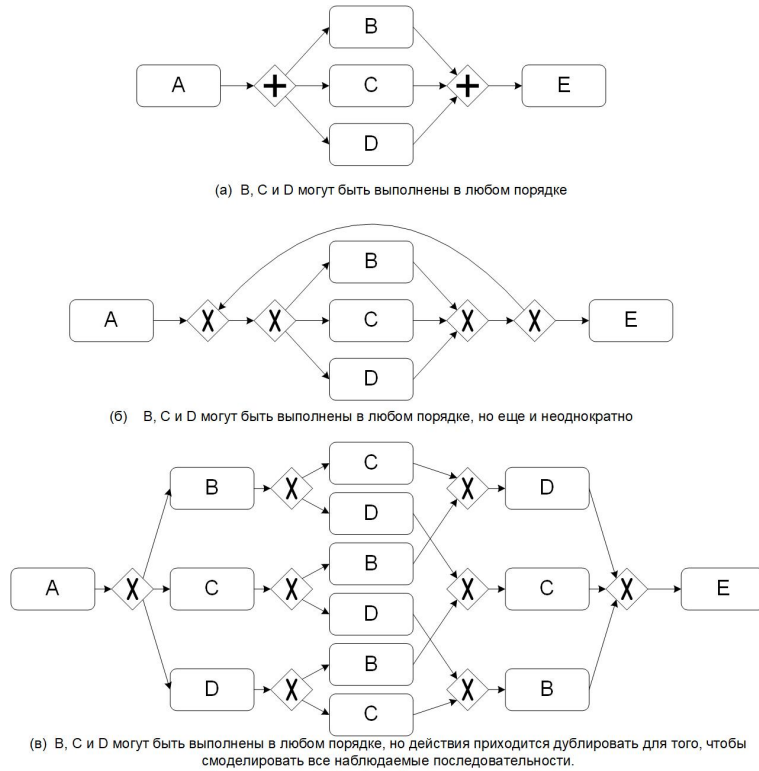


Рис. 6. Пример, иллюстрирующий проблемы, когда параллельное выполнение (операторы И) нельзя указать явным образом. В данном примере только действия (B, C и D) являются параллельными. Представим себе результирующую модель процесса, в которой 10 параллельно выполняемых действий ($2^{10} = 1024$ состояний и $10! = 3,628,800$ возможных последовательностей их выполнения).

Рисунок 6 показывает результат использование методов Process mining, не поддерживающих параллельное выполнение действий (без операторов «И»). Рассмотрим журнал событий $L = \{\langle A, B, C, D, E \rangle, \langle A, B, D, C, E \rangle, \langle A, C, B, D, E \rangle, \langle A, C, D, B, E \rangle, \langle A, D, B, C, E \rangle, \langle A, D, C, B, E \rangle\}$. L содержит кейсы, которые начинаются с A и заканчиваются E. Действия B, C и D появляются в произвольном порядке между A и E. BPMN-модель на Рис. 6 (а) демонстрирует компактную визуализацию процесса, используя два оператора «И». Предположим, что метод process mining не поддерживает оператор «И». В таком случае, очевидными кандидатами будут две BPMN-модели на Рис. 6. Модель на Рис. 6 (б) сравнительно компактна, но позволяет избыточные варианты исполнения (избыточное поведение), в частности, согласно модели, возможны такие кейсы как $\langle A, B, B, B, E \rangle$, что маловероятно, согласно журналу событий. Модель на Рис. 6 (в) соответствует всем кейсам L , но избыточна и не является компактным представлением журнала событий. Этот пример демонстрирует, что для процессов, которые в реальности содержат множество потенциально параллельных действий, результирующие модели, созданные в нотации, которая не поддерживает параллельное выполнение, будут обладать минусами недоподгонки (допуская избыточное поведение), и/или будут чрезмерно усложненными.

Как показано на Рис. 6, важно наличие поддержки хотя бы базовых паттернов потока работ. Кроме указанных базовых паттернов, желательно поддерживать и операторы включа-

ющего «ИЛИ», поскольку они позволяют создать компактную визуализацию множественного выбора и частичных синхронизаций.

3.4 РП4: События должны соотноситься с элементами модели

Как указано в Разделе 2, утверждение, что process mining сводится к извлечению потока управления, является заблуждением. Как показано на Рис. 1, извлеченная модель может представлять различные ракурсы (организационный ракурс, временной ракурс, ракурс данных и т.д.). К тому же, извлечение – это лишь одна из трех форм process mining, показанных на Рис. 3. Две других формы process mining (подтверждение соответствия и усовершенствование) во многом основываются на взаимосвязи между *элементами модели и событиями в журнале*. Это отношение может быть использовано для «воспроизведения» (replay) журнала событий в модели. Воспроизведение используется для того, чтобы выявить расхождения между журналом событий и моделью, например, если некоторые события в журнале невозможны согласно модели. Методы подтверждения соответствия количественно оценивают и диагностируют такие расхождения. При воспроизведении можно использовать временные отметки в журнале событий для анализа поведения во времени. Разница во времени между обычно связанными действиями может быть использована для того, чтобы добавить в модель время ожидания. Эти примеры показывают, что для различных типов анализа установление соотношения между событиями в журнале и элементами в модели является стартовой точкой.

В некоторых случаях установить такое соотношение – нетривиальная задача. Например, событие может относиться к двум разным действиям или может быть неизвестно, к какому действию оно относится. Для правильной интерпретации результатов process mining необходимо устранить подобные двусмысленности. Помимо задачи соотнесения событий и действий, есть еще задача соотнесения событий и экземпляров процесса, общеизвестная как *корреляция событий*.

3.5 РП5: Модели следует рассматривать как целенаправленные абстракции реальности

Модель, созданная на данных о событиях, представляет собой *точку зрения на реальность*. Эта точка зрения должна представлять собой целенаправленную абстракцию поведения, выявленного в журнале событий. При заданном журнале, полезными могут оказаться различные точки зрения. Кроме того, различным стейкхолдерам могут понадобиться различные точки зрения. На самом деле, модели, извлеченные из журналов событий, должны рассматриваться как «карты» (аналогично географическим картам). Данный руководящий принцип раскрывает важные идеи, две из которых описаны ниже.

Прежде всего, важно отметить, что не существует такого понятия, как «просто карта» для какого-то географического района. В зависимости от предполагаемого способа использования существуют различные карты: карты автомобильных дорог, карты для пешего туризма, велосипедные карты и т.д. Все эти карты показывают точку зрения на одну и ту же реальность, и нелепо предполагать, что существует некая «идеальная карта». То же самое относится и к моделям процессов: модель должна выделять то, что релевантно для конкретной категории пользователей. Извлеченные модели могут фокусироваться на различных аспектах (поток управления, поток данных, время, ресурсы, издержки и т.д.) и отражать их на различных уровнях детализации и точности. Например, руководитель может захотеть увидеть общую модель процесса с фокусировкой на издержках, в то время как процессный аналитик может захотеть увидеть детальную модель процесса с фокусировкой на отклонениях от стандартного течения. Также отметим, что различные стейкхолдеры могут захотеть увидеть процесс на различных уровнях: *стратегическом уровне* (решения на этом уровне имеют долгосрочный эффект, и они базируются на агрегированных данных о событиях за продолжительный период), *тактическом уровне* (решения на этом уровне имеют среднесрочный эффект, и они в

основном базируются на недавних данных), и *операционном уровне* (решения на этом уровне имеют незамедлительный эффект, и они базируются на данных о событиях текущих кейсов).

Во-вторых, когда речь заходит о создании понятных карт, полезно перенимать идеи из картографии. Например, дорожные карты абстрагируются от менее значимых дорог и населенных пунктов. Менее значимые вещи либо убираются с карты, либо динамически группируются в агрегированные фигуры (например, улицы и пригороды сливаются в города). Картографы не только устраняют несущественные детали, но также используют цвета для выделения важных особенностей. Кроме того, для обозначения значимости объектов используются графические элементы определенного размера (например, толщина линий и размеры точек могут различаться). Географические карты также имеют четкую привязку к осям x и y , т.е. размещение объектов на карте не является произвольным, поскольку координаты элементов имеют значение. Все это резко контрастирует с общепринятыми моделями процессов, которые обычно не используют цвет, размер и особенности расположения, чтобы сделать модели более понятными. Однако, идеи из картографии могут быть легко внедрены в конструирование карт извлеченных процессов. Например, размер действия может показывать его частоту или какое-либо другое свойство, обозначая его значимость (например, издержки или использование ресурса). Толщина дуги может отражать важность соответствующей причинно-следственной связи, а обозначение цветом может использоваться для выделения узких мест.

Наблюдения выше показывают: правильное представление и тонкая его настройка для целевой аудитории важны. Это важно при визуализации результатов для конечных пользователей и для подводки алгоритмов извлечения к подходящим моделям (см. также Вызов В5).

3.6 РП6: Process mining должен быть непрерывным процессом

Process mining позволяет строить содержательные «карты», напрямую связанные с данными о событиях. На такие модели можно проецировать как исторические данные о событиях, так и текущие данные. Вместе с тем, процессы меняются в процессе их анализа. Учитывая динамическую природу процессов, нецелесообразно рассматривать process mining как однократное действие. Цель должна состоять не в том, чтобы создать стационарную модель, но в том, чтобы вдохнуть жизнь в модели процессов – так, чтобы пользователи и аналитики стремились рассматривать их на ежедневной основе.

Сравним это с применением приложений, использующих геотеги. Существуют тысячи приложений, использующих Гугл-карты (например, приложения, выдающие информацию о дорожной обстановке, недвижимости, ресторанах быстрого питания или времени показа фильмов на выбранной карте). Люди могут легко увеличивать и уменьшать масштаб этих карт и взаимодействовать с ними (например, на карте отображаются пробки, а пользователь может выбрать конкретную проблему, чтобы уточнить детали). Так и в process mining должна быть возможность работы на основе данных, поступающих в реальном времени. Продолжая сравнение с картой, мы можем представить, что у событий есть GPS-координаты, которые проецируются на карту в режиме реального времени. По аналогии с автомобильными навигационными системами, инструменты process mining могут помогать пользователям в (а) навигации по процессам, (б) проецировании динамической информации на карты процесса (например, показывая «пробки» в бизнес-процессах) и (в) создании прогнозов для обрабатываемых кейсов (например, оценивая «время прибытия» кейса, который запаздывает). Эти примеры наглядно показывают, что было бы упущением не использовать модели процессов активнее. Process mining следует рассматривать как непрерывный процесс, дающий практическую информацию в различных временных масштабах (минутах, часах, днях, неделях и месяцах).

4 Вызовы

Для современных организаций, которые должны управлять нетривиальными операционными процессами, process mining является важным инструментом. С одной стороны, наблюдается невероятный рост объема данных о событиях. С другой стороны, информация должна быть идеально согласована с процессами, чтобы они удовлетворяли требованиям, связанным с комплаенсом, эффективностью и качеством обслуживания клиентов. Нет необходимости оспаривать применимость process mining в целом, тем не менее есть ряд важных вызовов, которые необходимо принимать во внимание. Наличие этих проблемных вопросов показывает, что process mining является формирующей дисциплиной. В следующей части мы рассмотрим некоторые из этих проблемных вопросов. Данный перечень не претендует на полноту: с течением времени могут возникнуть новые вызовы и задачи, которые потребуют значительных усилий, а существующие могут исчезнуть по мере развития методов process mining.

4.1 В1: Поиск, сведение и очистка данных о событиях

По-прежнему требуются значительные усилия для извлечения данных о событиях, пригодных для process mining. Обычно, необходимо преодолеть несколько препятствий:

- Данные могут быть *рассредоточены* по множеству источников, и эту информацию требуется свести. Решение такой задачи становится проблематичным в случае, когда в различных источниках данных используются разные идентификаторы. Например, для идентификации человека в одной системе используется имя и дата рождения, в то время как в другой системе используется номер его полиса социального страхования.
- Данные о событиях часто являются «объектно-ориентированными», а не «процессно-ориентированными». Например, отдельные изделия, паллеты и контейнеры могут иметь RFID-метки, и зарегистрированные события будут относиться к этим меткам. Однако для отслеживания отдельно взятого заказа клиента такие объектно-ориентированные события должны быть сведены и предварительно обработаны.
- Данные о событиях могут быть *неполными*. Распространенной проблемой является то, что события не связаны явным образом с соответствующими экземплярами процесса. Обычно такую информацию вполне возможно связать, но это может потребовать значительных усилий. Кроме того, для некоторых событий может отсутствовать информация о времени. Возможно, потребуется интерполировать временные отметки для того, чтобы все-таки использовать доступную информацию о времени.
- Журнал событий может содержать *выбросы*, т.е. аномальное поведение, также называемое *шумом*. Что следует считать выбросами? Как выявить эти выбросы? На эти вопросы необходимо ответить для очистки данных.
- Журналы могут содержать события *различной степени детализации*. В журнале событий информационной системы больницы событием могут быть как простой анализ крови, так и сложная хирургическая процедура. Временные отметки также могут различаться по степени детализации в диапазоне от миллисекундной точности (28-9-2011:h11m28s32ms342) до просто информации о дате события (28-9-2011).
- События происходят в определенном *контексте* (погода, загруженность, день недели и т.д.). Этот контекст может объяснить некоторые явления, например, большее чем обычно время отклика может произойти по причине незавершенной ранее работы или отпуска сотрудника. Желательно, чтобы этот контекст учитывался при анализе, что влечет за собой необходимость сведения данных о событиях с данными о контексте. Здесь начинается «проклятие размерности», так как при добавлении слишком большого количества переменных анализ становится слишком громоздким.

Для решения вышеуказанных проблем необходимы более совершенные инструменты и методологии. К тому же, как указывалось выше, организации должны воспринимать журналы событий как объекты первого класса, а не как побочный продукт. Целью является добиться

для журналов событий уровня ★★★★★ (см. Таблицу 1). В этом смысле, опыт, накопленный при использовании хранилищ данных, полезен и для целей обеспечения высококачественных журналов событий. Например, простая проверка на этапе ввода данных может существенно сократить долю некорректных данных.

4.2 В2: Работа со сложными журналами событий, имеющими различные характеристики

Журналы событий могут иметь очень разные характеристики. Некоторые журналы событий могут быть чрезвычайно большими, что затрудняет их обработку, в то время как другие журналы событий настолько малы, что данных недостаточно для того, чтобы сделать достоверные выводы. В некоторых сферах события регистрируются в умопомрачительных количествах, поэтому приходится прилагать дополнительные усилия по повышению производительности и масштабируемости. Например, компания ASML на постоянной основе осуществляет мониторинг всех своих фотолитических установок, используемых различными компаниями (например, компаниями Samsung и Texas Instruments) для производства чипов интегральных микросхем (около 70% чипов в мире произведено на фотолитических установках ASML). Существующие инструменты испытывают трудности при работе с петабайтами данных, собранных в таких сферах. Помимо количества зарегистрированных событий, существуют и другие характеристики, такие как среднее число событий в кейсе, сходство кейсов, количество уникальных событий и количество уникальных путей процесса. Рассмотрим журнал событий *L1* со следующими характеристиками: 1000 кейсов, в кейсе в среднем содержится 10 событий, вариативность невелика (например, ряд кейсов проходят по одному и тому же или похожему пути). Журнал событий *L2* содержит всего 100 кейсов, но в кейсе в среднем содержится 100 событий и все кейсы проходят по уникальному пути. Очевидно, что анализировать журнал *L2* гораздо сложнее, чем журнал *L1*, несмотря на то, что по размеру журналы схожи (около 10 000 событий).

Поскольку журналы событий содержат лишь примеры поведения, их не следует считать полными. Методы process mining вынуждены иметь дело с неполнотой, используя «предположение об открытом мире»: тот факт, что что-то не произошло, не означает, что это не может произойти. Это затрудняет работу с небольшими журналами событий с большой вариативностью.

Как уже упоминалось ранее, некоторые журналы содержат события с очень низким уровнем абстракции. Эти журналы, как правило, чрезвычайно велики, но отдельно взятые низкоуровневые события мало интересны стейкхолдерам. Поэтому можно было бы объединить низкоуровневые события в высокоуровневые события. Например, при анализе процессов диагностики и лечения конкретной группы пациентов отдельные тесты, зарегистрированные в информационной системе больницы лабораторий, сами по себе могут оказаться неинтересными.

На данный момент времени организациям приходится использовать метод проб и ошибок для того, чтобы понять, подходит ли журнал событий для process mining. Поэтому программное обеспечение должно позволять проводить быстрый тест пригодности отдельно взятого набора данных. Такой тест должен указать на потенциальные проблемы с получением результатов и предупредить о том, что журналы неполны или избыточно детальны.

4.3 В3: Разработка показательных бенчмарков

Process mining является развивающейся технологией, и это объясняет, почему хорошие бенчмарки до сих пор отсутствуют. Например, на данный момент доступны десятки методов process mining, различные вендоры предлагают различные продукты, но консенсуса насчет качества всех этих методов не наблюдается. Несмотря на то, что различия в функциональности и производительности огромны, сравнивать разные методы и различные программные продукты затруднительно. Поэтому необходимо разработать хорошие бенчмарки, состоящие из примеров наборов данных и показательных критериев качества.

Для классических методов data mining существуют хорошие бенчмарки. Эти бенчмарки подтолкнули поставщиков программного обеспечения и ученых к повышению производительности разрабатываемых ими методов. В случае с process mining это более сложная задача. Например, реляционная модель данных, сформулированная Э.Ф.Коддом в 1969 году, является очень простой и повсеместно поддерживается. Как следствие, для преобразования данных из одной базы в другую не требуется больших усилий и не возникает сложностей интерпретации. Для процессов настолько простой модели не существует. Стандарты, предлагаемые для моделирования процессов, гораздо сложнее, и лишь небольшое число вендоров поддерживает один и тот же набор понятий. Процессы просто-напросто гораздо более сложны, чем табличные данные.

И все же, создание показательных бенчмарков для process mining является важным, и некоторая предварительная работа уже проделана. Например, существуют различные метрики для измерения качества результатов process mining (подгонка, простота, точность и общность). Кроме того, ряд журналов событий находится в открытом доступе (см. www.processmining.org). См., например, журнал событий, использованный на первом Business Process Intelligence Challenge (BPIC'11), организованном рабочей группой (см. [doi:10.4121/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54](https://doi.org/10.4121/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54)).

С одной стороны, нужны бенчмарки, основанные на наборах данных из реальной жизни. С другой стороны, есть необходимость создать синтетические наборы данных с определенными заданными характеристиками. Такие синтетические наборы данных будут способствовать разработке техник process mining, направленных на обработку неполных журналов событий, журналов событий с шумами или специфических категорий процессов.

Помимо разработок показательных бенчмарков, также следует прийти к большому согласию в части критериев, используемых для оценки результатов process mining (см. также Вызов В6). Причем, для оценки результата могут быть адаптированы методы *перекрестной проверки*, используемые в data mining. Рассмотрим, например, k -блочную проверку. Разделим журнал событий на k частей. $k - 1$ частей могут быть использованы для извлечения модели процесса, а оставшуюся часть можно использовать для проверки соответствия полученного результата. Процедуру можно повторить k раз, и таким образом получить некоторые вводные для понимания качества модели.

4.4 В4: Работа с концептуальным дрейфом

Термин *концептуальный дрейф* описывает ситуацию, при которой процесс видоизменяется по ходу анализа. Например, в начале журнала событий два действия могут осуществляться параллельно, в то время как ближе к концу журнала эти действия становятся последовательными. Процессы могут видоизменяться в силу периодических/сезонных колебаний (например, «в декабре спрос возрастает» или «по пятницам после обеда на рабочих местах меньше сотрудников») или по причине изменившихся условий (например, «рынок становится более конкурентным»). Подобные изменения влияют на процессы, и их выявление и анализ принципиально важны. Концептуальный дрейф процесса может быть обнаружен посредством разделения журнала событий на журналы меньшего размера и анализа «следов» в этих меньшего размера журналах. Такой анализ «второго порядка» требует существенно большего объема данных о событиях. Однако, стабильных во времени процессов совсем немного, и понимание концептуального дрейфа является важнейшей задачей в управлении процессами. Поэтому для адекватного анализа концептуального дрейфа нужны дополнительные исследования и разработка дополнительного функционала программного обеспечения.

4.5 В5: Снижение предвзятости представления при извлечении процесса

Методы извлечения процесса создают модель в конкретной нотации (например, BPMN или сети Петри). Однако, важно разделять визуализацию результата и представление, получаемое во время извлечения самого процесса. Выбор целевой нотации зачастую включает в себя несколько неявных допущений, что ограничивает пространство исследований: не могут

быть извлечены те процессы, которые не смогут быть отображены средствами выбранной нотации. Эта так называемая «предвзятость представления», получаемая при извлечении процесса, должна быть предметом осознанного выбора и не должна определяться только предпочитаемым графическим представлением.

Рассмотрим пример из Рис. 6: от того, допускает выбранная нотация параллельное выполнение действий или нет, зависят как визуализация извлеченной модели, так и класс моделей, изучаемых алгоритмом. В случае, если предвзятость представления не допускает параллельного выполнения (на Рис. 6 (а) это не допускается) и не допускает различным действиям иметь одну и ту же метку (на Рис. 6 (в) это не допускается), то тогда возможны только проблематичные модели, как на Рис. 6 (б). Этот пример показывает, что выбор представления должен быть как можно более тщательным и продуманным.

4.6 В6: Обеспечение баланса критериев качества: подгонки, простоты, точности и обобщенности

Достаточно частой является ситуация неполноты журналов событий, т.е. даны только некоторые примеры поведения. Модели процессов обычно допускают экспоненциальное или даже бесконечное количество различных трасс (traces) (при наличии «петель»). Кроме того, одни трассы встречаются с меньшей вероятностью, чем другие. Следовательно, нереалистичным будет предположить, что каждая возможная трасса представлена в журнале событий. Для того, чтобы проиллюстрировать непрактичность принятия полных журналов как должного, рассмотрим процесс, состоящий из 10 действий, которые могут выполняться параллельно, и соответствующий журнал, который содержит информацию о 10 000 кейсов. Общее количество возможных переходов между такими действиями (interleavings), переплетенных между собой, в модели с 10 параллельными действиями составляет $10! = 3,628,800$. Соответственно, в логе не представлено каждое переплетение, поскольку количество кейсов (10 000) меньше, чем количество потенциальных трасс (3,628,800). Даже если в журнале миллионы кейсов, крайне маловероятно то, представлены все возможные вариации. Дополнительная сложность состоит в том, некоторые альтернативы встречаются реже, чем другие, и могут быть признаны «шумом». Невозможно построить разумную модель для слишком «шумного» поведения. Извлеченная модель должна абстрагироваться от этого, а низкочастотные варианты поведения предпочтительнее исследовать с использованием подтверждения соответствия.

Шум и неполнота делают извлечение процесса сложной задачей. По сути, есть четыре конкурирующих измерения качества: (а) подгонка, (б) простота, (в) точность и (г) обобщенность.

Модель с хорошей *подгонкой* учитывает большинство вариантов поведения, наблюдаемого в журнале. У модели идеальная подгонка в случае, если все трассы из лога могут быть воспроизведены моделью от начала до конца. *Наипростейшая* модель, способная объяснить поведение, отраженное в журнале событий, является наилучшей. Этот принцип известен как «брита Оккама». Однако, только подгонки и простоты недостаточно, чтобы оценить качество извлеченной модели процесса. Например, очень легко построить крайне простую сеть Петри («лепестковую модель»), которая будет способна воспроизвести все трассы журнала событий (а также любого другого журнала событий, содержащего тот же набор действий). Аналогично, нежелательна модель, которая воспроизводит исключительно содержащиеся в журнале событий варианты поведения. Необходимо помнить, что журнал содержит лишь примеры поведения, и возможно появление различных, пока не наблюдаемых, трасс. Модель *точна*, если она не допускает избыточного поведения. Очевидно, что «лепестковой модели» не хватает точности. Неточная модель является «недоподогнанной». Недоподгонка – это такая проблема, когда модель чрезмерно обобщает представленное в журнале поведение (т.е., модель допускает варианты поведения, сильно отличающиеся от вариантов в журнале событий). Модель должна обобщать и не сводить варианты поведения к только тем примерам, которые наблюдаются в журнале. Модель, которая не *обобщает*, является «переподогнанной». Переподгонка – это такая проблема, когда модель чрезмерно специфицирована, между

тем как очевидно, что журнал содержит только примеры поведения (т.е., модель объясняет выборку из конкретного журнала, но следующая выборка событий из журнала того же процесса может дать совершенно иную модель процесса).

Обеспечение баланса подгонки, простоты, точности и обобщенности является сложной задачей. По этой причине большинство наиболее значимых методов извлечения процесса предусматривают разносторонние параметры. Необходимо совершенствовать алгоритмы, чтобы лучше балансировать эти четыре конкурирующих критерия. Причем любой используемый параметр должен быть понятен конечным пользователям.

4.7 В7: Кросс-организационный process mining

Традиционно, process mining применяется в рамках одной организации. Однако, по мере того, как сервисные технологии, интеграция цепей поставок и облачные вычисления распространяются все шире, возникают ситуации, когда журналы событий нескольких организаций становятся доступными для анализа. В принципе, возможны два режима *кросс-организационного process mining*.

Во-первых, мы можем рассмотреть коллаборативный режим, когда различные организации совместно обрабатывают экземпляры процесса. Такой кросс-организационный процесс можно сравнить с пазлом, т.е. общий процесс разделен на части и распределен по организациям, которые взаимодействуют для успешного исполнения кейсов. Анализа журнала событий только одной из вовлеченных организаций недостаточно. Для извлечения сквозных процессов необходимо объединить журналы событий различных организаций. Это нетривиальная задача, так как необходимо соотнести события, разделенные организационными границами.

Во-вторых, мы можем рассмотреть режим, когда различные организации в основном выполняют один и тот же процесс, делясь опытом и знаниями или используя одну инфраструктуру. Рассмотрим пример платформы Salesforce.com, на базе которой у многих организаций реализованы процессы продаж. С одной стороны, у этих организаций общая инфраструктура (процессы, базы данных и т.д.). С другой стороны, их никто не принуждает следовать строго определенной модели процесса, поскольку система может быть сконфигурирована так, чтобы поддерживать различные варианты одного и того же процесса. В качестве другого примера рассмотрим основные процессы, исполняемые каким-либо муниципалитетом (например, выдача разрешения на строительство). Несмотря на то, что все муниципалитеты обязаны применять один и тот же набор процессов, могут иметь место и вариации. Очевидно, анализ этих вариаций в различных организациях представляет интерес. Данные организации могут учиться друг у друга, а поставщики услуг имеют возможность улучшать качество обслуживания и предлагать услуги с добавленной ценностью, основанные на результатах кросс-организационного process mining.

Необходимо разрабатывать новые методы для обоих типов кросс-организационного process mining. Эти методы должны принимать во внимание вопросы конфиденциальности и безопасности. Организации могут не захотеть делиться информацией по причинам конкуренции или из-за отсутствия доверия. Поэтому важно разрабатывать методы process mining, обеспечивающие конфиденциальность.

4.8 В8: Сопровождение операционной деятельности

Изначально process mining был сфокусирован на анализе накопленных данных. Однако на сегодняшний день многие источники данных обновляются в режиме реального времени или близком к нему, а вычислительные мощности позволяют анализировать события в момент их возникновения. Соответственно, не стоит сводить process mining к офлайн анализу, он может использоваться и для поддержки текущей операционной деятельности. Определены три задачи поддержки операционной деятельности: *обнаружение*, *предсказание* и *рекомендации*. Момент отклонения кейса от предписанного процесса может быть обнаружен, и тогда система

выдаст предупреждение. Зачастую желательно генерировать такие уведомления незамедлительно (и иметь возможность повлиять на события), а не в режиме офлайн. На накопленных данных могут быть построены предсказательные модели, которые можно использовать для управления выполняющимися экземплярами процессов. Например, можно предсказать время, оставшееся до окончания обработки кейса. На основе таких предсказаний становится возможным построение систем рекомендаций, предлагающих какие-либо действия для снижения издержек или сокращения времени цикла. Применение методов process mining в режиме онлайн ставит дополнительные задачи, предъявляющие требования к вычислительным мощностям и качеству данных.

4.9 В9: Сочетание методов process mining с другими типами анализа

Операционный менеджмент (Operations management) и в частности исследование операций является разделом наук о менеджменте, в значительной степени опирающимся на моделирование. Используются разнообразие математические модели: от линейного программирования и планирования проектов до моделей массового обслуживания, цепей Маркова и имитационного моделирования. Глубинный анализ данных (data mining) может быть определен как «анализ наборов данных (часто больших) для обнаружения неожиданных взаимосвязей и обобщения данных нестандартными способами, которые понятны и полезны владельцу данных». Разработано широкое многообразие методов анализа: классификация (например, обучение дерева решений), регрессия, кластеризация (например, метод k-средних) и выявление паттернов (например, обучение ассоциативным правилам).

Обе области знаний (операционный менеджмент и глубинный анализ данных) используют самодостаточные методы анализа. Задачей-вызовом является сочетание методов этих областей с process mining. Рассмотрим на примере. Методы process mining позволяют разработать на основе накопленных данных имитационную модель. Соответственно, имитационная модель может использоваться для обеспечения операционной поддержки. В силу тесной связи модели и журнала событий, модель может использоваться для исторической реконструкции: можно приступить к моделированию и нажать «кнопку быстрой перемотки вперед» из текущего состояния в будущее на основе текущего данных.

Схожим образом желательно совместить process mining с *визуальной аналитикой*. Визуальная аналитика сочетает автоматизированный анализ с интерактивными визуализациями для лучшего понимания больших и сложных наборов данных. Визуальная аналитика использует удивительные способности человека находить закономерности в неструктурированных данных. При сочетании методов process mining с интерактивной визуальной аналитикой становится возможным извлечение большего количества идей из данных о событиях.

4.10 В10: Повышение удобства использования неспециалистами

Одной из целей process mining является создание «живых моделей процессов», т. е. таких моделей процессов, которые используются ежедневно, а не статичных моделей, которые в конечном итоге оказываются где-то в архиве. Новые данные о событиях могут использоваться для извлечения нового поведения. Связка между данными о событиях и моделями процессов дает возможность спроецировать текущее состояние и недавние действия на актуальные модели. Таким образом, конечные пользователи могут взаимодействовать с результатами process mining на ежедневной основе. Такое взаимодействие очень ценно, но оно требует интуитивных пользовательских интерфейсов. Задача состоит в том, чтобы скрыть сложные алгоритмы process mining за удобными интерфейсами, которые автоматически задают параметры и предлагают подходящие типы анализа.

4.11 В11: Повышение понятности для неспециалистов

Даже если сгенерировать результаты process mining несложно, это на самом деле не означает, что эти результаты практичны. У пользователя могут возникнуть проблемы с пониманием

выходных данных или появиться соблазн сделать неверные выводы. Чтобы избежать подобных проблем, результаты должны быть изложены с использованием подходящего представления (см. также РП5). Кроме того, всегда следует четко указывать на степень достоверности полученных результатов. Возможно, что для обоснования конкретных выводов данных слишком мало. По факту, существующие методы извлечения процессов обычно не предупреждают о недостаточной либо избыточной подгонке. Они в любом случае отображают модель, даже когда очевидно, что данных слишком мало для обоснования каких-либо выводов.

5 Послесловие

Целями рабочей группы IEEE Task Force on Process Mining являются (а) продвижение практического применения process mining, (б) консультирование разработчиков программного обеспечения, руководителей бизнеса и конечных пользователей по вопросам использования самых современных методов process mining и (в) стимулирование научных исследований в данной области. В данном манифесте излагаются основные принципы и намерения рабочей группы. После введения в тематику process mining в данном манифесте перечисляются некоторые руководящие принципы (Раздел 3) и вызовы (Раздел 4). Руководящими принципами следует пользоваться во избежание очевидных ошибок. Перечень вызовов определяет направления научных исследований и разработок. Оба раздела ставят целью повышение уровня зрелости process mining.

В заключение, несколько слов о терминологии. В пространстве process mining используются следующие термины: workflow mining (извлечение потока работ), (business) process mining, автоматизированное извлечение процессов и процессная аналитика. Похоже, что различные организации используют различные термины в пересекающихся областях знаний. Например, Gartner продвигает «автоматизированное извлечение процессов» (Automated Business Process Discovery, ABPD), а компания Software AG при описании своей контроллинговой платформы использует термин «процессная аналитика» (Process Intelligence). Термин workflow mining представляется не очень подходящим, поскольку создание моделей потоков работ является лишь одним из многих возможных вариантов применения process mining. Аналогичным образом, добавление термина «бизнес-» ограничивает диапазон применения process mining конкретной областью применения. Существуют многочисленные варианты применения process mining (например, анализ использования высокотехнологичных систем или анализ веб-сайтов), где подобное добавление представляется неуместным. Несмотря на то, что извлечение процесса является важной составной спектра задач process mining, это лишь один из способов использования. Подтверждение соответствия, предсказание, извлечение организационного ракурса, анализ социальных связей и т.д. являются вариантами использования, выходящими за рамки извлечения процесса.

Рисунок 7 показывает взаимосвязь некоторых только что упомянутых терминов. Все технологии и методы, ставящие целью предоставление ценной информации, которая может быть использована для поддержки принятия решений, могут быть позиционированы под зонтичным понятием бизнес-аналитика (Business Intelligence, BI). (Бизнес)-процессная аналитика ((Business) process intelligence) может рассматриваться как комбинация методов BI и BPM, т.е. методы BI используются для анализа и улучшения процессов и управления ими. Process mining может рассматриваться как конкретизация процессной аналитики, берущая за точку отсчёта журналы событий. (Автоматизированное) извлечение (бизнес-) процессов является только одной из трех базовых форм process mining. Рисунок 7 может немного дезориентировать, в том смысле что большинство инструментов BI не обладают той функциональностью, которая описана в данном документе. Сам термин BI часто вольно смещается в сторону описания конкретного программного обеспечения или метода, покрывающего только малую часть из широкого спектра BI.

Использование альтернативных терминов может объясняться коммерческими соображениями. Некоторые вендоры могут захотеть выделить отдельно взятый аспект (например, извлечение или аналитику). Однако, во избежание путаницы, для дисциплины, описанной в данном манифесте предпочтительнее использовать термин process mining.

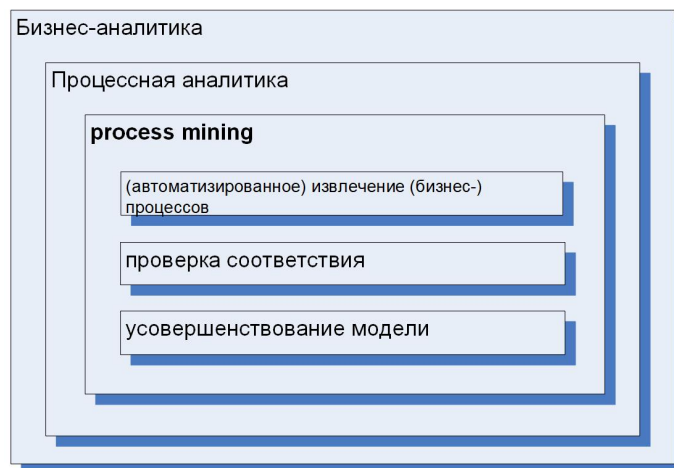


Рис. 7. Взаимосвязь различных терминов.

Глоссарий

- **Автоматизированное извлечение бизнес-процессов (Automated Business Process Discovery)**: см. **Извлечение процессов**.
- **Аналитика бизнес-процессов (Business Process Intelligence)**: см. **Процессная аналитика**.
- **Бизнес-аналитика (Business Intelligence, BI)**: широкий набор методов и инструментов работы с данными для поддержки принятия решений.
- **Глубинный анализ данных (Data Mining)**: анализ (обычно больших) выборок данных в целях поиска скрытых связей и формирования таких представлений данных, которые позволяют получить новые знания.
- **Глубинный анализ процессов (Process Mining)**: практики, методы и инструменты извлечения, мониторинга и улучшения существующих (не предполагаемых) бизнес-процессов. В основе process mining лежит извлечение новых знаний из журналов событий, которые массово доступны в современных информационных системах.
- **Действие (Activity)**: четко определенный шаг процесса. События (Events) происходят с действиями в каждом экземпляре процесса, например, начало действия, завершение, отмена и т.д.
- **Журнал событий (Event Log)**: совокупность событий, используемая в качестве входной информации для process mining. События не обязательно должны храниться в отдельном специализированном журнале (например, события могут быть распределены по различным таблицам баз данных).
- **Извлечение процесса (Process Discovery)**: один из трех основных видов Process mining. На основе журнала событий формируется модель процесса. Например, с помощью α -алгоритма можно извлечь сеть Петри, определяя паттерны процесса в совокупности событий.
- **Кейс (Case)**: см. **Экземпляр процесса**.
- **Концептуальный дрейф (Concept Drift)**: изменение процессов во времени. Наблюдаемый процесс может постепенно или скачкообразно измениться (например, в результате сезонных изменений, роста конкуренции), усложняя таким образом анализ.
- **Кросс-организационный process mining (Cross-Organizational Process Mining)**: применение методов process mining к журналам событий, сформированным на базе разных организаций.
- **Обобщенность (Generalization)**: метрика, которая определяет, насколько полученная модель допускает поведение, которое ранее не наблюдалось. Модель, которая получена в результате переподгонки, не будет достаточно обобщенной.

- **Подгонка (Fitness)**: метрика, которая определяет, насколько полученная модель допускает поведение, отраженное в журнале событий. Точная подгонка модели означает, что все трассы журнала возможно воспроизвести на модели от начала до конца.
- **Предвзятость представления (Representational Bias)**: выбор нотации для визуализации результатов извлечения процесса.
- **Проверка соответствия (Conformance Checking)**: анализ соответствия записанных в журнале событий реальных данных сформированной модели, и наоборот. Целью проверки соответствия является поиск отклонений и их оценка. Проверка соответствия – одна из трех основных форм process mining.
- **Простота (Simplicity)**: метрика, описывающая соответствие принципу «бритвы Оккама»: наиболее простая модель, с помощью которой можно описать поведение, наблюдаемое в журнале событий, является наилучшей моделью. Метрика простоты модели количественно может быть измерена разными способами, например, количеством узлов и дуг в модели.
- **Процессная аналитика (Process Intelligence)**: направление в бизнес-аналитике, фокусирующееся на управлении бизнес-процессами.
- **Событие (Event)**: записанное в журнале событий действие (например, начало шага, его завершение, отмена и т.д.) для конкретного экземпляра процесса.
- **Сопровождение операционной деятельности (Operational Support)**: онлайн-анализ поступающих данных событий с целью мониторинга и воздействия на выполняющиеся экземпляры процессов. Определены три задачи сопровождения операционной деятельности: *обнаружение* (предупреждение в случае, если наблюдаемое поведение отклоняется от модельного), *предсказание* (предсказание будущего поведения на основании прошлого, например, можно предсказать оставшееся время обработки) и *рекомендации* (предложение соответствующих действий для достижения конкретных целей, например, снижения затрат).
- **Точность (Precision)**: метрика, которая определяет, насколько полученная модель допускает поведение, значимо отличающееся от поведения, наблюдаемого в журнале событий. Модель с малой точностью – это модель с «недоподгонкой».
- **Управление бизнес-процессами (Business Process Management, BPM)**: подход к управлению организацией, сочетающий применение информационных технологий и дисциплин менеджмента к операционным бизнес-процессам.
- **Усовершенствование модели (Model Enhancement)**: одна из трех основных форм process mining. Модель процесса расширяется и улучшается на основе информации, извлеченной из определенного журнала событий. Например, узкие места можно определить при помощи воспроизведения журнала событий на модели процесса, анализируя при этом временные отметки.
- **Экземпляр процесса (Process Instance)**: сущность, обрабатываемая анализируемым процессом. События относятся к экземплярам процессов. Примерами экземпляров процессов являются заказы клиента, заявления о выплате страхового возмещения, заявки на кредит и т.д.
- **MXML**: основанный на XML формат журналов событий. MXML заменен форматом XES, новым, не зависящим от платформы форматом данных process mining.
- **XES**: основанный на XML стандарт разметки журналов событий. Этот стандарт был принят IEEE Task Force on Process Mining в качестве основного формата данных журналов событий. (ср. www.xes-standard.org).