过程挖掘宣言

Wil van der Aalst[1,2,*], Arya Adriansyah[1], Ana Karla Alves de Medeiros[50], Franco Arcieri[26], Thomas Baier[11,53], Tobias Blickle[6], Jagadeesh Chandra Bose[1], Peter van den Brand[4], Ronald Brandtjen[7], Joos Buijs[1], Andrea Burattin[28], Josep Carmona[29], Malu Castellanos[8], Jan Claes[45], Jonathan Cook[30], Nicola Costantini[21], Francisco Curbera[9], Ernesto Damiani[27], Massimiliano de Leoni[1], Pavlos Delias[51], Boudewijn van Dongen[1], Marlon Dumas[44], Schahram Dustdar[46], Dirk Fahland[1], Diogo R. Ferreira[31], Walid Gaaloul[49], Frank van Geffen[24], Sukriti Goel[12], Christian Günther[5], Antonella Guzzo[32], Paul Harmon[17], Arthur ter Hofstede[2,1], John Hoogland[3], Jon Espen Ingvaldsen[14], Koki Kato[10], Rudolf Kuhn[7], Akhil Kumar[33], Marcello La Rosa[2], Fabrizio Maggi[1], Donato Malerba[34], Ronny Mans[1], Alberto Manuel[20], Martin McCreesh[15], Paola Mello[38], Jan Mendling[35], Marco Montali[52], Hamid Motahari Nezhad[8], Michael zur Muehlen[36], Jorge Munoz-Gama[29], Luigi Pontieri[25], Joel Ribeiro[1], Anne Rozinat[5], Hugo Seguel Pérez[23], Ricardo Seguel Pérez[22], Marcos Sepúlveda[47], Jim Sinur[18], Pnina Soffer[37], Minseok Song[39], Alessandro Sperduti[28], Giovanni Stilo[26], Casper Stoel[3], Keith Swenson[13], Maurizio Talamo[26], Wei Tan[9], Chris Turner[40], Jan Vanthienen[41], George Varvaressos[16], Eric Verbeek[1], Marc Verdonk[19], Roberto Vigo[21], Jianmin Wang[42], Barbara Weber[43], Matthias Weidlich[48], Ton Weijters[1], Lijie Wen[42], Michael Westergaard[1], and Moe Wynn[2]

[1] Eindhoven University of Technology, The Netherlands
[2] Queensland University of Technology, Australia
[3] Pallas Athena, The Netherlands
[4] Futura Process Intelligence, The Netherlands
[5] Fluxicon, The Netherlands
[6] Software AG, Germany
[7] ProcessGold AG, Germany
[8] HP Laboratories, USA
[9] IBM T.J. Watson Research Center, USA
[10] Fujitsu Laboratories Ltd., Japan
[11] BWI Systeme GmbH, Germany
[12] Infosys Technologies Ltd, India
[13] Fujitsu America Inc., USA
[14] Fourspark, Norway
[15] Iontas/Verint, USA
[16] Business Process Mining, Australia
[17] Business Process Trends, USA
[18] Gartner, USA
[19] Deloitte Innovation, The Netherlands
[20] Process Sphere, Portugal
[21] Siav SpA, Italy

[*] 通讯作者：Wil van der Aalst, e-mail: w.m.p.v.d.aalst@tue.nl

[22] BPM Chile, Chile
[23] Excellentia BPM, Chile
[24] Rabobank, The Netherlands
[25] ICAR-CNR, Italy
[26] University of Rome "Tor Vergata", Italy
[27] Università degli Studi di Milano, Italy
[28] University of Padua, Italy
[29] Universitat Politècnica de Catalunya, Spain
[30] New Mexico State University, USA
[31] IST - Technical University of Lisbon, Portugal
[32] University of Calabria, Italy
[33] Penn State University, USA
[34] University of Bari, Italy
[35] Vienna University of Economics and Business, Austria
[36] Stevens Institute of Technology, USA
[37] University of Haifa, Israel
[38] University of Bologna, Italy
[39] Ulsan National Institute of Science and Technology, Korea
[40] Cranfield University, UK
[41] K.U. Leuven, Belgium
[42] Tsinghua University, China
[43] University of Innsbruck, Austria
[44] University of Tartu, Estonia
[45] Ghent University, Belgium
[46] Technical University of Vienna, Austria
[47] Pontificia Universidad Católica de Chile, Chile
[48] Hasso Plattner Institute, Germany
[49] Telecom SudParis, France
[50] Capgemini Consulting, The Netherlands
[51] Kavala Institute of Technology, Greece
[52] Free University of Bozen-Bolzano, Italy
[53] Humboldt-Universität zu Berlin

摘要

过程挖掘技术能够从现代信息系统普遍产生的事件日志中抽取信息，该技术为各种应用领域中的过程发现、监测和改进提供了新的手段。过程挖掘日益受到关注有两个方面的原因：一方面，越来越多的事件得以记录，从而可以提供关于过程历史的详细信息；另一方面，在竞争和飞速变化的环境中改善和更好的支持业务过程这一需求日益突显。此宣言由 IEEE 过程挖掘工作组制定，其目的是推广过程挖掘技术。此外，通过定义一系列的指导原则并指明其重要的挑战，该宣言希望成为软件开发者、科学家、咨询顾问、业务经理和最终用户的指南。其最终目标是提高过程挖掘作为一种新工具来改进业务过程的（重）设计、控制和支持的成熟度。

1. IEEE 过程挖掘工作组

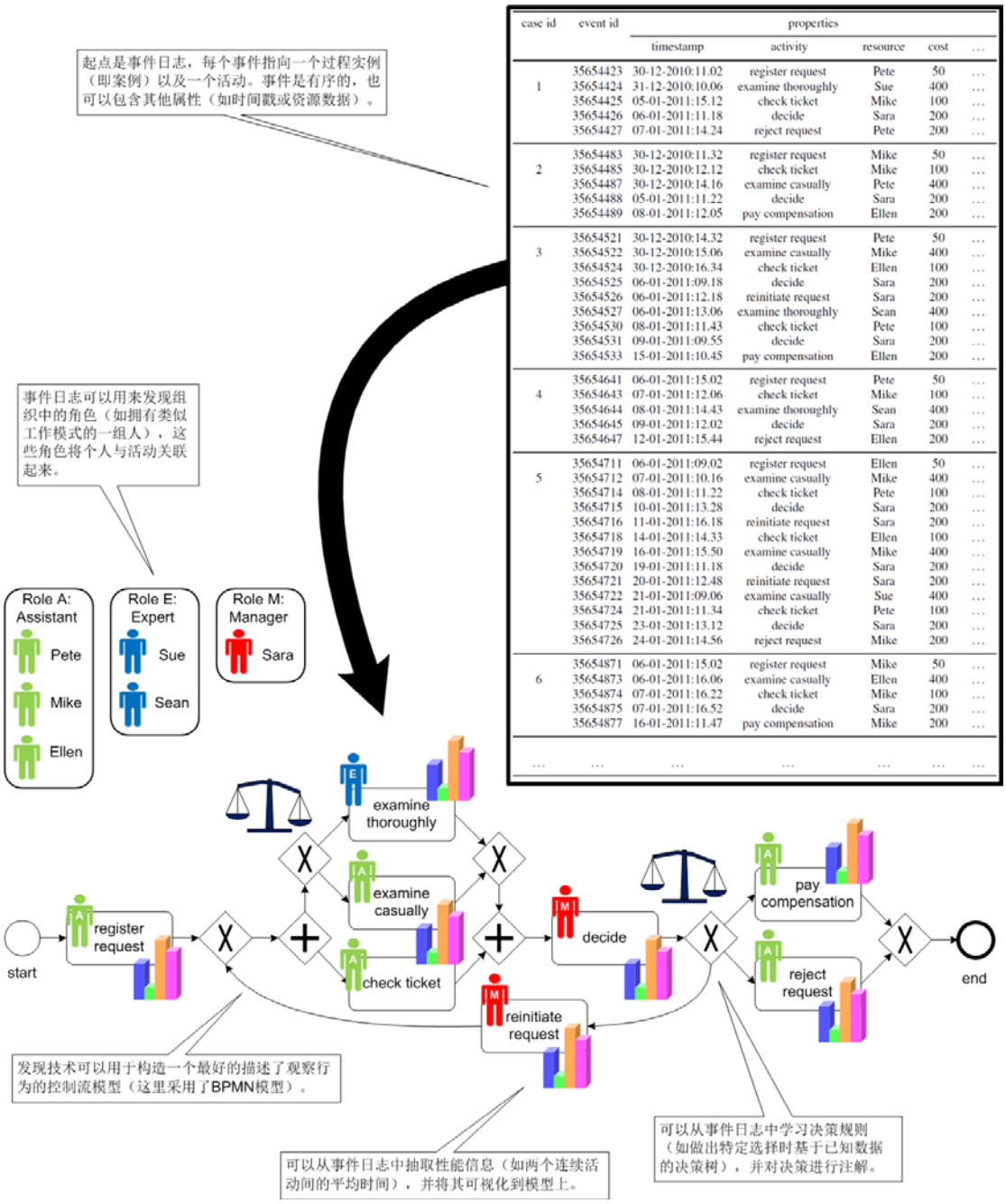
宣言是由一群人发表的“准则和内涵的公开声明”。本宣言是由 IEEE 过程挖掘工作组的成员和支持者撰写的，该工作组的目的是推动过程挖掘的研究、发展、教育、实现、演变和理解。

过程挖掘是一个年轻的跨领域的交叉学科，即跨计算智能与数据挖掘领域和过程建模与分析领域。过程挖掘的想法是通过从现代（信息）系统普遍可见的事件日志中提取知识，来发现、监测和改进实际过程（而非假想过程），见图 1。过程挖掘包括过程（自动）发现（即从一个事件日志中提取过程模型）、符合性检查（即通过对比模型和日志来监测偏差的发生）、社交网络/组织挖掘、仿真模型的自动生成、模型扩展、模型修复、案例预测，以及基于历史的推荐等。

过程挖掘在数据挖掘和业务过程建模与分析之间搭建了重要的桥梁。商务智能领域引进了很多流行词来说明比较简单的报告和仪表盘工具。业务活动监测（Business Activity Monitoring，即 BAM）指的是用来实时监测业务过程的技术。复杂事件处理(Complex Event Processing，即 CEP)指的是处理大量事件，并将其应用到业务监测、引导和优化中去的技术。公司业绩管理（Corporate Performance Management, 即 CPM）是测量过程或者组织表现的又一个流行词。同样相关的管理方法还有持续过程改进(Continuous Process Improvement，即 CPI)、业务过程改进(Business Process Improvement，即 BPI)、全面质量管理(Total Quality Management，即 TQM)和 6-Sigma。这些方法有一个共同点，即过程被“放在显微镜下”来观测是否存在进一步改进的可能性。过程挖掘是 CPM、BPI、TQM、6-Sigma 等的使能技术。

尽管商务智能工具和管理方法，如 6-Sigma 和 TQM，目的在于改进运作性能，如减少流转时间和缺陷，同时组织也更注重企业的治理、风险和合规性。比如，萨班斯奥克斯利法案(Sarbanes-Oxley Act，即 SOX) 和巴赛尔资本（Basel II Accord）协定阐明了合规性问题的焦点。过程挖掘技术为更严格的合规性检查，以及查明关于组织核心过程相关信息的有效性和可靠性提供了手段。

在过去的十年里，事件数据越来越容易获得，过程挖掘技术正在趋于成熟。此外，刚刚提到的和过程改进（如 6-Sigma、TQM、CPI 和 CPM）及合规性 （SOX, BAM 等）相关的管理需求也可从过程挖掘技术中受益。幸运的是，过程挖掘算法已经在各类学术和商业系统中得以实现。当前有一群很活跃的研究者致力于过程挖掘工作，过程挖掘也成为了业务过程管理（BPM）研究中的热门课题之一。此外，工业界对过程挖掘也表现出浓厚的兴趣。越来越多的软件制造商将过程挖掘功能添加到他们的软件产品中。带有过程挖掘能力的软件产品包括：ARIS PPM（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)和 Reflect|one (Pallas Athena)。日益增长的对事件日志进行过程分析的兴趣，导致了过程挖掘工作组的设立。



case id	event id	properties				
		timestamp	activity	resource	cost	…
1	35654423	30-12-2010:11.02	register request	Pete	50	…
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	…
	35654425	05-01-2011:15.12	check ticket	Mike	100	…
	35654426	06-01-2011:11.18	decide	Sara	200	…
	35654427	07-01-2011:14.24	reject request	Pete	200	…
2	35654483	30-12-2010:11.32	register request	Mike	50	…
	35654485	30-12-2010:12.12	check ticket	Mike	100	…
	35654487	30-12-2010:14.16	examine casually	Pete	400	…
	35654488	05-01-2011:11.22	decide	Sara	200	…
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	…
3	35654521	30-12-2010:14.32	register request	Pete	50	…
	35654522	30-12-2010:15.06	examine casually	Mike	400	…
	35654524	30-12-2010:16.34	check ticket	Ellen	100	…
	35654525	06-01-2011:09.18	decide	Sara	200	…
	35654526	06-01-2011:12.18	reinitiate request	Sara	200	…
	35654527	06-01-2011:13.06	examine thoroughly	Sean	400	…
	35654530	08-01-2011:11.43	check ticket	Pete	100	…
	35654531	09-01-2011:09.55	decide	Sara	200	…
	35654533	15-01-2011:10.45	pay compensation	Ellen	200	…
4	35654641	06-01-2011:15.02	register request	Pete	50	…
	35654643	07-01-2011:12.06	check ticket	Mike	100	…
	35654644	08-01-2011:14.43	examine thoroughly	Sean	400	…
	35654645	09-01-2011:12.02	decide	Sara	200	…
	35654647	12-01-2011:15.44	reject request	Ellen	200	…
5	35654711	06-01-2011:09.02	register request	Ellen	50	…
	35654712	07-01-2011:10.16	examine casually	Mike	400	…
	35654714	08-01-2011:11.22	check ticket	Pete	100	…
	35654715	10-01-2011:13.28	decide	Sara	200	…
	35654716	11-01-2011:16.18	reinitiate request	Sara	200	…
	35654718	14-01-2011:14.33	check ticket	Ellen	100	…
	35654719	16-01-2011:15.50	examine casually	Mike	400	…
	35654720	19-01-2011:11.18	decide	Sara	200	…
	35654721	20-01-2011:12.48	reinitiate request	Sara	200	…
	35654722	21-01-2011:09.06	examine casually	Sue	400	…
	35654724	21-01-2011:11.34	check ticket	Pete	100	…
	35654725	23-01-2011:13.12	decide	Sara	200	…
	35654726	24-01-2011:14.56	reject request	Mike	200	…
6	35654871	06-01-2011:15.02	register request	Mike	50	…
	35654873	06-01-2011:16.06	examine casually	Ellen	400	…
	35654874	07-01-2011:16.22	check ticket	Mike	100	…
	35654875	07-01-2011:16.52	decide	Sara	200	…
	35654877	16-01-2011:11.47	pay compensation	Mike	200	…
…	…	…	…	…	…	…

图 1　过程挖掘技术从事件日志中抽取信息，从而发现、监测和改进过程

过程挖掘工作组于 2009 年由 IEEE 计算智能协会（Computational Intelligence Society）下的数据挖掘技术委员会（Data Mining Technical Committee）设立。当前工作组拥有的成员代表了软件制造商（如 Pallas Athena、 Software AG、Futura Process Intelligence、HP、IBM、Infosys、Fluxicon、 Businesscape、Iontas/Verint、Fujitsu、Fujitsu Laboratories、Business Process Mining、Stereologic）、咨询公司/最终用户（如 ProcessGold、Business Process Trends、Gartner、Deloitte、Process Sphere、Siav SpA、BPM Chili、BWI Systeme GmbH、Excellentia BPM、Rabobank）和研究机构（如 TU/e、University of Padua、Universitat Politècnica de Catalunya、New Mexico State University、IST - Technical University of Lisbon、University of Calabria、Penn State Uni-versity、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）。

过程挖掘工作组的具体目标包括：

- 使用户、开发者、咨询顾问、业务经理和研究人员洞悉过程挖掘的发展现状；
- 促进过程挖掘技术和工具的使用，激发新的应用软件；
- 致力于事件日志记录的标准化；
- 组织辅导、专题会议、专题讨论会、专家小组；
- 发表文章、出版书籍、发布专刊和视频。

过程挖掘工作组自 2009 年成立以来，陆续开展了与上述目标相关的各类活动。组织了多个研讨会和专题讨论，如业务过程智能研讨会（BPI'09, BPI'10 和 BPI'11）和在IEEE会议（如CIDM'11）上的专题讨论。相关知识通过各种途径得以传播，包括专题（如WCCI'10 和PMPM'09）、暑期学校（如ESSCaSS'09, ACPN'10, CICH'10 等）、视频（www.processmining.org）和一些出版物，包括Springer最近出版的第一本关于过程挖掘的书籍。工作组也（参与）组织了第一届业务过程智能挑战赛（BPIC'11）：在比赛中参与者需要在庞大并且复杂的事件日志中提取有用信息。2010 年，本工作组提出了XES日志格式(www.xes-standard.org)。XES是一个标准化可扩展的日志格式，已被OpenXES library (www.openxes.org)及ProM、XESame、Nitro等工具支持。

如果您想了解关于IEEE过程挖掘工作组的更多活动信息，请访问：http://www.win.tue.nl/ieeetfpm/。

2. 过程挖掘的发展现状

依赖于计算机的信息系统及其他系统的扩展能力很好地遵循了摩尔定律（Moore's law）。Gorden Moore，Intel 的创始人之一，于 1965 年预测集成电路上可容纳的晶体管数目约每年增加一倍。过去的 50 年中，尽管增长速度逐渐变慢，但仍是呈指数增长。这个速度导致了数字空间（如电子化数据的存储或交换）惊人地增长，使得数字空间和物理空间变得更加一致。

数字空间增长和组织中的过程协调一致，使事件的记录和分析变得随处可见。事件涉及的范围很广，包括从 ATM 机中取款、医生校准 X 射线机、公民申请驾照、纳税申报的提交、游客获取电子客票号码的收据。其挑战在于以一种有意义的方式利用事件数据，例如提供洞察细节、识别瓶颈、预测问题、记录违规、推荐对策和简化处理，过程挖掘的目的就是要解决上述难题。

过程挖掘的起点是事件日志。所有的过程挖掘技术均假设能够连续地记录事件，从而每个事件代表一个活动（即在过程中一个良好定义的步骤），也使得每个事件和一个特定的案例 (即一个过程实例)相关。事件日志可能存储了关于事件的额外信息。事实上，在很多情况下，过程挖掘技术都使用了额外信息，如资

源（即人或设备）的执行或初始化活动、事件的时间戳、或者在事件中记录的数据元素（如订单规模）。

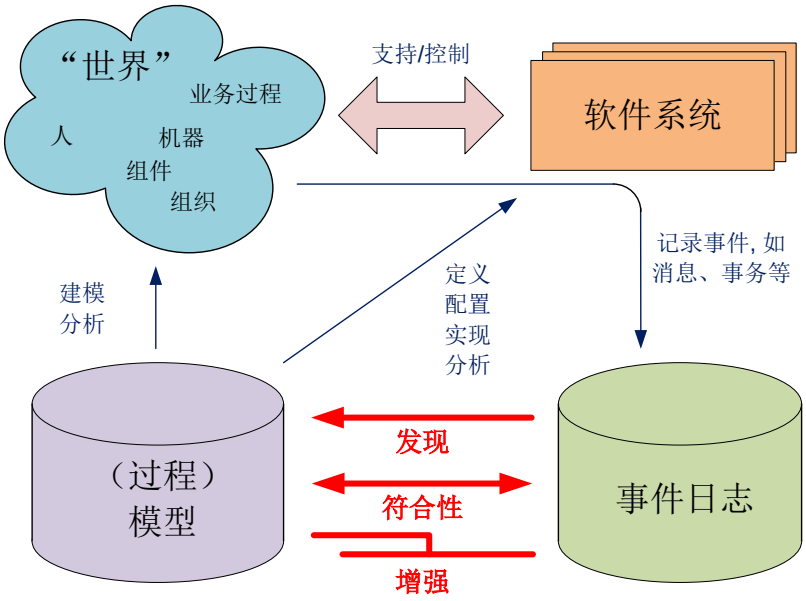


图 2　三种类型的过程挖掘场景：发现、符合性检查和增强

如图 2 所示，事件日志可以被用于三种类型的过程挖掘场景。第一类过程挖掘场景是“发现”，即根据一个事件日志生成一个模型，并不使用任何先验信息。过程发现是最典型的过程挖掘技术。许多组织已惊奇地发现，基于事件日志中的执行样例就可以发现实际的业务过程。第二类过程挖掘场景是“符合性检查”，将一个已知的过程模型与这个模型的事件日志进行对比。符合性检查可以用来检查日志中记载的实际情况是否符合这个模型，反之亦然。第三种类型的过程挖掘场景是“增强”，指的是使用一些事件日志中记录的实际过程来扩展或者改进现存的过程模型。与符合性检查不同，第三种过程挖掘场景的目的在于改进或扩展已有模型。例如，使用事件日志中的时间戳，我们可以扩展模型以显示瓶颈、服务水平、生产周期和频率。

图 3 描述了这三种过程挖掘场景的输入和输出。“发现”场景的输入是事件日志，输出是过程模型。发现的结果模型通常是一个过程模型（如 Petri 网、BPMN、EPC 或 UML 活动图），不过，该模型也可以表达其他视角（如社交网络）。“符合性检查”场景的输入需要一个事件日志和相应的过程模型，输出是反映模型和日志间差异与共性的诊断信息。“增强”（修复或者扩展）场景的输入也是一个过程日志和相应的过程模型，输出是一个改进或者扩展后的过程模型。

过程挖掘可以覆盖多个不同的视角。控制流视角关注活动间的次序，其目的在于找到反映所有可能路径的最佳特征，结果通常用 Petri 网或者其他过程描述语言（如 EPC、BPMN 或 UML 活动图）进行表达。组织视角关注隐藏在日志中的资源信息，例如，涉及的角色（如人、系统、角色或部门）以及他们如何关联，其目标是发现角色和组织单元，或者展示社交网络。案例视角关注案例的性质。很明显，一个案例可以由它在过程中的路径或者为其工作的参与者来刻画。不过案例也可以通过相应数据元素的值来刻画，例如，如果一个案例代表一个进货订单，那么就要关注供应商和订单的产品数量。时间视角注重事件的周期和频率，

如果事件带有时间戳，就有可能发现瓶颈、度量服务水平、监视资源的使用、预测运行案例的剩余处理时间。

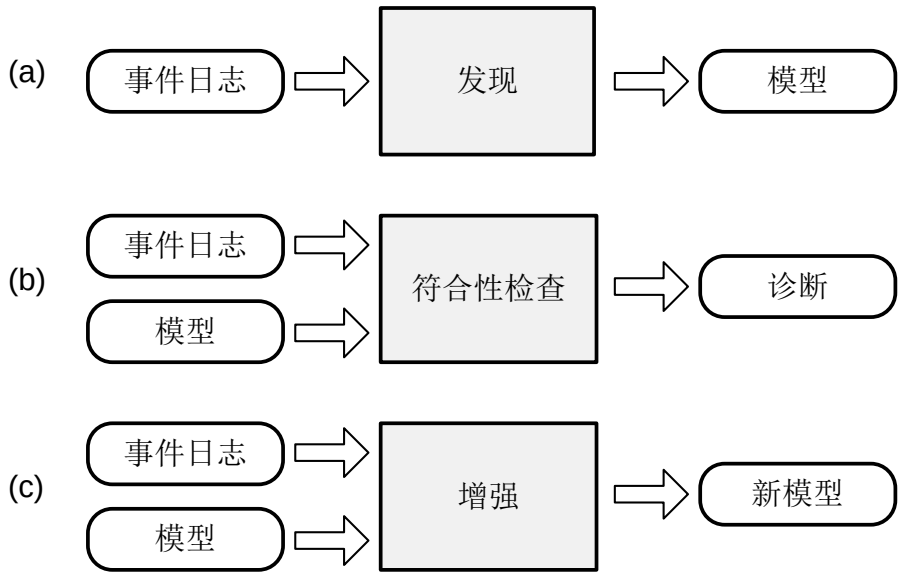


图 3：三种基本过程挖掘类型的输入和输出：发现、符合性检查和增强

关于过程挖掘有一些普遍存在的错误认识。比如，一些制造商、分析师和研究人员把过程挖掘的范围局限为一类特定的数据挖掘技术，仅能够用于离线分析的过程发现。事实上并非如此，因此我们重点强调如下三个方面。

- 过程挖掘不仅仅局限于控制流发现。从事件日志中发现过程模型激发了实干家和学者们的想象力。所以，控制流发现经常被认为是过程挖掘中最令人兴奋的部分。然而，过程挖掘不仅仅局限于控制流发现。一方面，发现仅仅是三种基本过程挖掘场景（发现、符合性检查和增强）中的一种。另一方面，其范围不仅局限于控制流方面，组织、案例和时间方面都扮演着重要角色。
- 过程挖掘不仅仅是数据挖掘的一种特定类型。过程挖掘被看做是数据挖掘和传统的模型驱动 BPM 中缺少的一环。绝大多数数据挖掘技术根本不是以过程为中心的。从根本上支持并发语义的过程模型，同简单的数据挖掘结构（如决策树和关联规则）不可同日而语，因此 需要全然不同的表示和算法。
- 过程挖掘不仅仅局限于离线分析。过程挖掘技术从历史事件数据中提取知识。尽管使用历史数据，但是结果可以应用于运行案例中。例如，一个部分完成的用户订单的完成时间可以借助发现的过程模型进行预测。

为定位过程挖掘，我们使用如图 4 所示的 BPM 生命周期。BPM 生命周期说明了一个业务过程的 7 个阶段和对应的信息系统。（再）设计阶段创建一个新的过程模型或者修改一个已经存在的过程模型。分析阶段对候选模型和其它可选模型进行分析。（再）设计阶段后，模型在实现阶段得以实现，或者一个已存在的系统在（重）配置阶段得以重新配置。在执行阶段，设计好的模型得以执行和监测。此外，一些较小的调整可以不需要重新设计过程，而是直接在调整阶段进行改动。在诊断阶段，被执行的过程得到分析，此阶段的输出可能会触发一次新的

（再）设计阶段。对于图 4 中所示的大多数阶段，过程挖掘都是一个很有价值的工具。很显然，过程挖掘对诊断阶段最有帮助。然而，过程挖掘不仅仅局限在诊断阶段。例如在执行阶段，过程挖掘技术可以被用于运作支持。依据历史信息得到的模型可以做出预测和推荐，来影响运行中的案例。类似形式的决策支持可以用来调整过程以及指导过程的（重）配置。

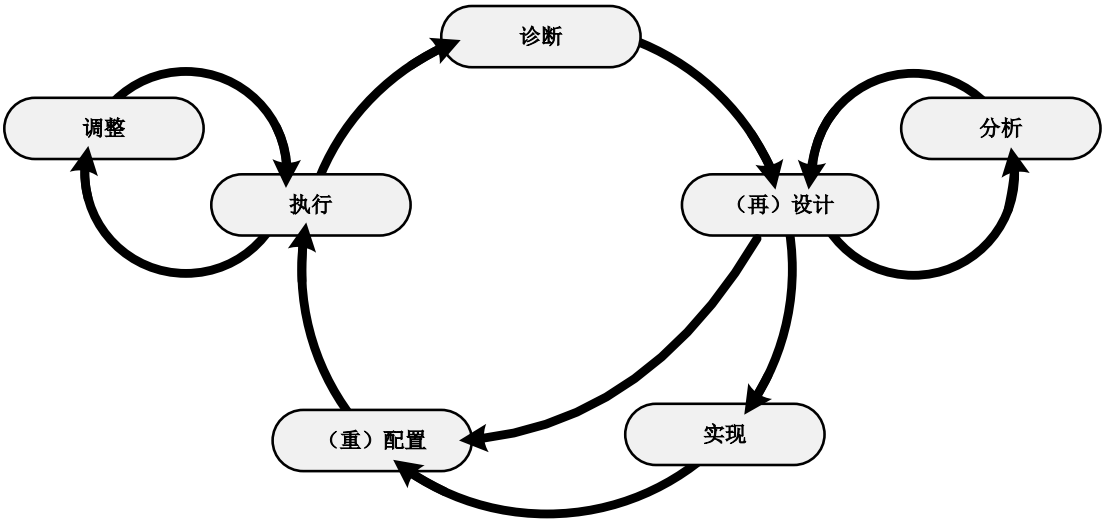


图 4：确定业务过程的各个阶段，以及对应信息系统的 BPM 生命周期；过程挖掘在各个阶段都能发挥作用（实现阶段除外）。

图 4 给出了完整的 BPM 生命周期，图 5 则侧重具体的过程挖掘活动和制品。图 5 描述了一个过程挖掘项目中包含的典型阶段。任何过程挖掘项目都开始于一个计划和该计划的一个需求（阶段 0）。项目初始化后，事件数据、模型、目标和问题需要从系统、领域专家和管理中获取（阶段 1）。这就要求理解可用数据（“什么数据可以用于分析？”），也要求理解特定领域（“什么是重要的问题？”），并得到图 5 中所示的制品 (即历史数据、手工模型、目标和问题)。阶段 2 创建控制流模型并将其与事件日志关联，在这里可以采用过程自动发现技术。发现的过程模型可能已经为一些问题和触发重设计或调整动作提供了解决方法。此外，可能会使用模型来过滤或者修改事件日志（如删除不常见的活动或异常案例，以及插入丢失的事件）。有时需要投入很大精力来关联属于同一个过程实例的事件，即事件与过程模型实例相关联。当过程已相对结构化时，控制流模型可能在阶段 3 中被扩展以集成其他视角（如数据、时间和资源）的信息。基于阶段 2 中建立的事件日志与模型之间的关系，来扩展模型（例如，相关事件的时间戳可以用来估计活动的等待时间）。这可能用来回答更多的问题和触发更多的行为。最终，阶段 3 中构造的模型可能用于阶段 4 中的运作支持。从历史事件数据中提取的信息与运行案例的信息相结合，可用来干预、预测和推荐。只有在过程足够稳定和结构化时才能到达阶段 3 和阶段 4。

尽管支持图 5 中所示所有阶段的技术和工具目前都可以找到，但是过程挖掘毕竟是一个相对年轻的领域，大多数现存的工具还不够成熟。此外，大多数用户还不了解过程挖掘的潜能和限制。所以，本宣言为过程挖掘技术的用户以及有兴趣发展过程挖掘技术的研究者和开发者，分门别类地说明了一些重要的指导原则

（见第 3 节）和本领域当前面临的挑战（见第 4 节 。

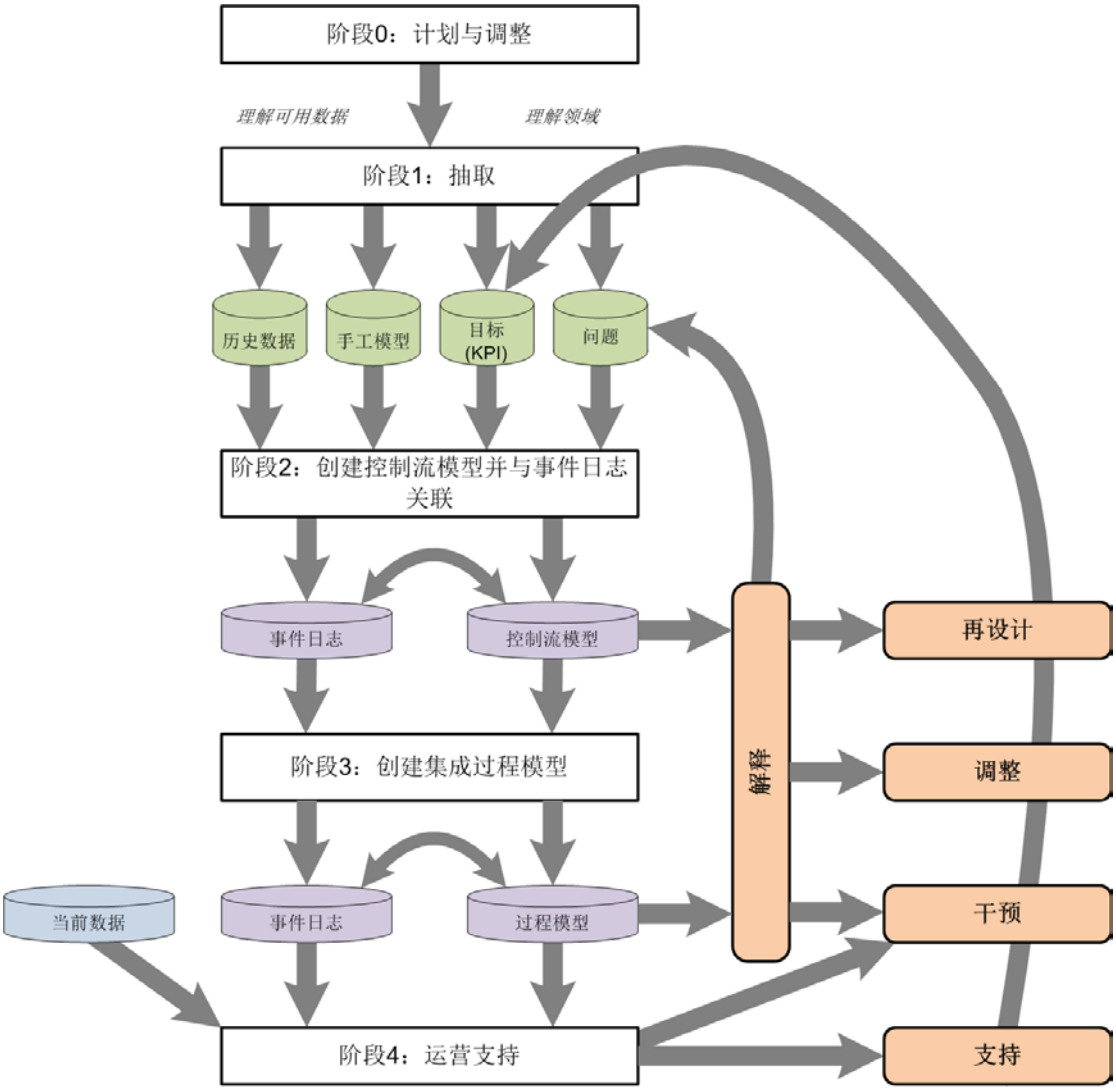


图 5：模型生命周期描述了过程挖掘项目的 5 个阶段：计划与调整阶段（阶段 0），提取阶段（阶段 1），创建控制流模型并与事件日志关联（阶段 2），创建集成过程模型（阶段 3），提供运作支持（阶段 4）

3. 指导原则

与所有的新技术一样，过程挖掘在实际应用时会遇到一些常见的错误。为了用户和分析师避免类似错误，我们列出以下六条指导原则。

3.1 指导原则 1：事件数据应该被当做“头等公民”

任何过程挖掘活动的起点都是记录的事件，我们将事件的集合称作事件日志，不过这不代表事件就一定要存储在专门的日志文件中。事件可能存储在数据库表、消息日志、邮件档案、事务日志和其他数据源中。比存储格式更重要的是存储日志的质量，过程挖掘结果的质量很大程度上取决于输入。所以事件日志在支持过程分析的信息系统中应该被当作“头等公民”对待。不幸的是，事件日志

经常仅仅被作为“副产品”用于调试或记录。例如，飞利浦卫生保健的医疗设备会记录事件，仅仅是因为软件开发者在代码中插入了“打印语句”。尽管针对类似语句有很多非正式的规范，但是提高事件日志的质量需要更加系统化的方法。事件日志应该被当作信息系统中的头等公民对待，而不是二等公民。

评判事件数据的质量有几个准则。事件应该是值得信任的，即可以认为记录的事件是真正发生过的并且事件属性是正确的。事件日志应该是完备的，即在给定的范围内不应该有丢失的事件。任何记录的事件应该有定义良好的语义。此外，在记录事件时出于隐私和安全的考虑，事件数据应该是安全的。例如，参与者应该明晰记录事件的类型和其被使用的方式。

表 1 事件日志的成熟度

级别	特征
★★★★★	最高级别：事件日志质量非常好（即可信的和完备的），并且事件定义良好。用自动、系统、可靠和安全的方式来记录事件。充分考虑隐私和安全性。此外，记录的事件（及其属性）有清晰的语义。这就意味着存在一个或多个本体，事件及其属性指向本体。 例如：BPM 系统的语义标注日志。
★★★★	用系统、可靠的方式自动记录事件，即事件是可信和完备的。与运行在级别★★★的系统不同，显式支持过程实例（案例）和活动等概念。 例如：传统 BPM/工作流系统中的事件日志。
★★★	自动记录事件，但不遵循系统化的方法。与级别★★的日志不同，此级别有一定的保障使记录的事件和事实相符（即事件日志是可信的，却不一定完备），例如一个 ERP 系统记录的事件。尽管事件需要从多张表中提取，信息仍然可以被认为是正确的（例如，可以认为 ERP 记录的付款记录是真实存在的，反之亦然）。 例如：ERP 系统中的表、CRM 系统的事件日志、通信系统中的交易日志、高科技系统中的事件日志等。
★★	自动记录事件，即一些信息系统的“副产品”。覆盖性是变化的，即没有系统化的方法来判断哪些事件得以记录。此外，某些业务活动可能会不通过信息系统从而不记录事件，所以事件可能会丢失或者没有得到正确记录。 例如：文档和产品管理系统的事件日志、嵌入式系统的错误日志、服务工程师的工作单等。
★	最低级别：事件日志质量很差。记录的事件可能与事实不符，并且可能会丢失事件。一般手工记录事件的事件日志会有这样的特点。 例如：组织内部路由的纸质文档流转记录、纸质病历卡等。

表 1 定义了从最高（★★★★★）到最低（★）五个级别的事件日志成熟度。例如，飞利浦卫生保健的事件日志处于级别★★★，即事件被自动记录并且记录的行为符合事实，不过没有用系统化的方法来为事件分配语义，并确保覆盖性处在一个特定级别。过程挖掘技术可以应用于级别★★★★★，级别★★★★和级别★★★的日志。原则上，过程挖掘技术可以应用于级别★★或级别★的过程日志。不过类似这样日志的分析存在着明显问题并且结果不可信。事实上，将过程挖掘应用于级别★的日志没有太大意义。

为了更好的从过程挖掘中受益，组织应该着眼于尽可能高质量级别的事件日志。

3.2 指导原则 2：日志提取应该由问题驱动

如图 5 所示，过程挖掘活动应该由问题驱动。如果问题不清晰，提取有意义的事件数据是非常困难的。如 SAP 这样的 ERP 系统，其数据库拥有数千张表，没有具体问题就不知道哪些表与数据提取相关。

如图 1 所示的过程模型描述了一种特定类型案例（即过程实例）的生命周期。所以，在应用任何过程挖掘技术前，应该选择要分析的案例类型。这个选择应该由需要回答的问题驱动，且并且绝非轻而易举的事情。例如 考虑客户订单的处理，每个客户订单可能包括多条订单项，因为客户可能在一个订单中订购了多个产品。一个客户订单可能产生多次送货，一次送货可能指向多个订单的订单项。因此，在订单和送货之间存在多对多的关系，一个订单和订单项之间存在一对多的关系。给定一个包含与订单、订单项和送货相关事件数据的数据库，就可能发现不同的过程模型。这样就可以以描述每个订单生命周期这一目的来提取数据，也可能以发现每个订单项或每次送货的生命周期为目的来提取数据。

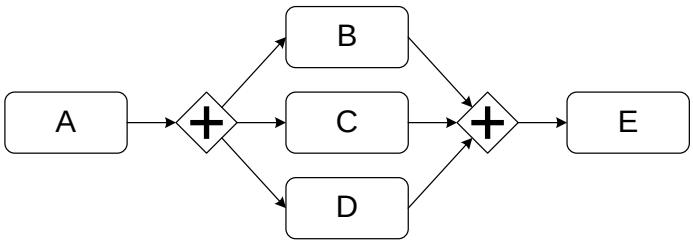
3.3 指导原则 3：应该支持并发、选择和其他基本控制流结构

目前存在大量的过程建模语言（如 BPMN、EPC、Petri 网、BPEL 和 UML 活动图），一些语言提供了很多建模元素（如 BPMN 提供 50 多种不同的图形元素），相比之下其他过程建模语言提供的建模元素则非常有限（如 Petri 网仅由三种不同元素组成：库所、变迁和弧）。控制流描述是任何过程模型的基础，所有主流语言都支持的基本控制流结构（也叫模式）是：顺序、并行路由（*与*分裂/合并）、 选择路由（*异或*分裂/合并）和循环。显然，过程挖掘技术应该完整支持上述模式。然而，一些挖掘技术不能处理并发，而仅仅支持马尔可夫链/变迁系统。

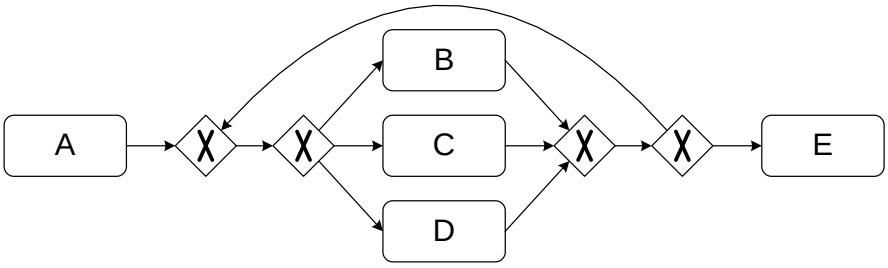
图 6 展示了使用支持并发（没有‘*与*分裂/合并’）的过程挖掘技术的应用效果。给定事件日志 L= {<A,B,C,D,E>, <A,B,D,C,E>, <A,C,B,D,E>, <A,C,D,B,E>, <A,D,B,C,E>, <A,D,C,B,E>}。L 包括从 A 起始，结束于 E 的案例。活动 B、C、D 在 A 和 E 之间以任意顺序发生。图 6(a)中的 BPMN 模型通过简洁的表现形式表达了使用两个与门的基本过程。假设过程挖掘技术不支持‘*与*分裂/合并’，这种情况下，图 6 中的另外两个 BPMN 模型成为候选。图 6(b)中的 BPMN 模型很简洁，不过却允许过多的行为（例如案例<A, B, B, B, E>按照这个模型是可能出现的，不过根据事件日志不可能出现）。图 6(c)中的 BPMN 模型符合 L 中的案例，却显式硬编码了所有序列，因此不是该日志的一种简洁表现形式。这个例子说明了对于拥有很多潜在并发活动的实际模型，结果模型非常欠拟合（即允许过多的

行为），或者不支持并发时结果模型就会极端复杂。

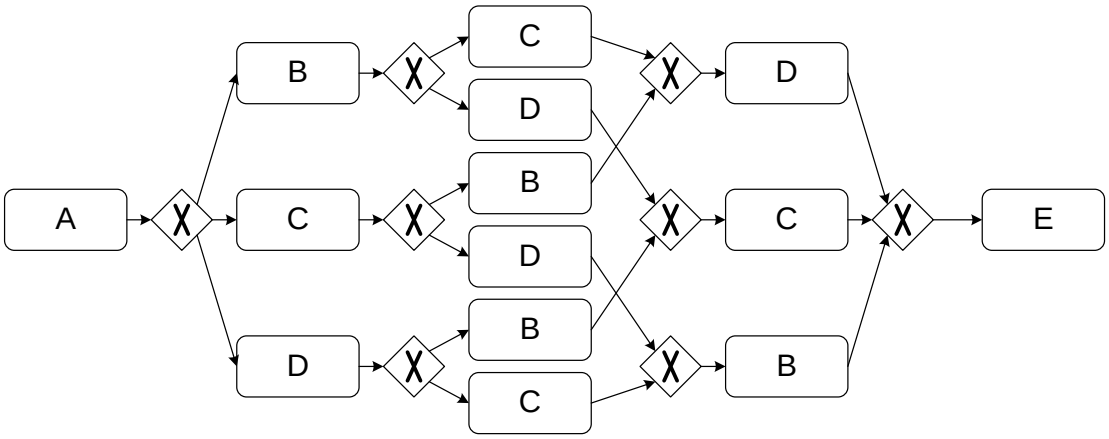
正如图 6 表明，支持最基本的控制流模式是非常重要的。除了提到的基本模型，支持*或*分裂/合并也非常重要，因为其提供了非排他性选择和部分同步的简洁表示形式。



(a) B、C、D可以任意次序执行



(b) B、C、D可以任意次序执行任意多次



(c) B、C、D可以任意次序执行，但为建模所有观察到的序列，活动需要被重复

图 6：举例说明当*并发*（即*与*分裂/合并）不能直接表达时的问题。在这个例子中，只有三个活动 B、C、D 是并发的，想象一下当有 10 个并发活动时的结果过程模型（$2^{10} = 1024$种状态和 10！=3,628,800 种可能的执行序列）

3.4 指导原则 4：事件应该与模型元素相关

第 2 节说明了过程挖掘被局限于控制流发现是一种错误认识。如图 1 所示，发现的过程模型可能覆盖多种视角（组织视角、时间视角、数据视角等），而且发现只是图 3 中三种过程挖掘场景中的一种。另外两种过程挖掘场景（符合性检查和增强）非常依赖于模型中元素和日志中事件之间的关系，这种关系可以被用来在模型上“重播”事件日志。重播可以揭示事件日志和模型之间的差异，例如日志中的一些事件相对模型而言不应存在。符合性检查技术量化这类差异并给予

诊断。事件日志中的时间戳可以用于在重播时分析时序行为，因果行为间的时间差可以用来增加模型的预期等待时间。这些例子都表明了日志中事件和模型中元素之间的关系可以作为不同类型分析的起点。

在一些情况下，建立这一关系可能并不容易。例如，一个事件可以指向两个不同的活动，或者指向的活动并不清楚。为了合理解释过程挖掘的结果，需要消除类似的模糊性。除了关联事件到活动的问题，还存在关联事件到过程实例的问题，这通常被称作事件关联。

3.5 指导原则 5：模型应该被看作对现实的有针对性的抽象

从事件数据中得到的模型提供了现实的抽象视图，这样一个视图应该是为了某一目标而对事件日志中获取的行为所进行的抽象，即这种抽象是为了某一目标而存在的。给定一个事件日志，可能存在多个有用的视图。此外，不同参与者可能需要不同的视图。事实上，从事件日志中发现的模型应该被看做地图（类似地理的地图）。这个指导原则提供了更为深刻的见解，下面描述其中的两个。

首先，对某一特定的地域，统一的地图并不存在。根据不同的用法存在不同的地图：道路地图、登山地图、骑车地图等。所有的这些地图都是在描述同一个现实世界，只是视角不同，认为存在完美的、普适的地图是很荒谬的。这同样适用于过程模型：模型应该着眼于和特定类型用户相关的事情。发现的模型可以把重点放在不用的方面（控制流、数据流、时间、资源、成本等），用不同的粒度和精度表现它们。例如，一位经理可能想要查看关于成本的粗略非正式的过程模型，然而一位过程分析师可能想要查看详细的关于流程偏差的过程模型。同样要注意不同参与者可能想要查看一个过程的不同层次：决策层次（有长期效应，并且基于长期积累的聚类事件数据的决策）、战略层次（有中期效应且基于近期数据的决策）和运作层次（有即时效果，基于与运行案例相关的事件数据的决策）。

其次，采用制图学的思想，有助于制作易于理解的地图。例如，道路地图抽象掉了不是很重要的道路和城市，不重要的事物要么是被忽略要么就被动态并入到聚集形状中（例如街道和郊区被并入城市中）。制图者不仅省略不相关的细节，还要使用颜色来突出重要特征。此外，图形元素有特定的尺寸来表明它们的重要性（例如线条粗细和点大小的变化）。地理地图对 x 轴和 y 轴也有清楚的解释（比如上北下南、左西右东），即地图不是任意布局的，它的元素的坐标是有意义的。所有这些都和主流的过程模型形成了鲜明的对比，过程模型并不使用颜色、尺寸和位置特征使模型更易理解。然而，制图学的想法可以很容易地被结合到发现过程地图的构造中。例如，活动的尺寸可以用来反映活动的频率，或者其他表明重要性（例如成本和资源使用）的性质，一条弧的宽度可以反映对应因果依赖的重要性，弧的颜色可以用来突出瓶颈。

上述观察表明，选择正确的表达形式并且根据不同的参与者做出相应调整是非常重要的。这对向最终用户可视化的呈现结果，以及指导发现算法构造合适的模型也是非常重要的（见挑战 5）。

3.6 指导原则 6：过程挖掘应该是一个持续的过程

过程挖掘有助于提供直接与事件数据相连的有意义的地图，历史事件数据和当前数据都可以投影到这样的模型上。此外，过程在被分析的同时也会发生改变。

考虑到过程的动态本质，不建议将过程挖掘看作是事后的一次性的活动。我们的目标不应该是创建一个一成不变的模型，而应为过程模型注入活力，从而激励用户和分析者在日常工作中经常查看它们。

可以把这一方面与使用地理标记的应用程序进行对比。有数千个应用程序使用 Google 地图（例如一些应用程序将有关交通状况、房地产、快餐店或者电影放映的信息投影到一幅选中地图上）。人们可以流畅的放大和缩小这些地图，并且和地图进行交互（例如，堵车的情况投影到了地图上，用户可以选择某一特定问题查看具体信息）。可以基于实时事件数据构造过程挖掘。借用地图中的概念，我们可以认为一个事件拥有 GPS 坐标，这些坐标可以实时投影到地图上。类似于汽车导航系统，过程挖掘工具可以通过以下几种方式帮助最终用户：（a）对过程的导航，（b）将动态信息投影到过程地图中（例如，在业务流程中显示“堵车”情况），（c）对运行案例进行预测（例如，估计延迟案例的“到达时间”）。这些例子表明不充分的使用过程模型将是非常令人遗憾的事情。所以，过程挖掘应该被看作是一个持续的过程，根据不同的时间尺度（分钟、小时、天、周和月）提供可实施的信息。

4. 挑战

过程挖掘是现代组织用来管理运作过程的一种重要工具。一方面，事件数据规模以惊人的速度增长；另一方面，过程和信息需要进行对齐，以满足与符合性、效率和用户服务相关的需求。以上这些都表明过程挖掘仍然是一门新兴学科。除了适用性（见上面第 3 节），过程挖掘还面临一些需要解决的重大挑战。下面，我们将列出其中一部分挑战。显然，这个列表并不能列举全部的挑战，因为随着时间的推移，过程挖掘的发展将导致新挑战的产生，以及现存挑战的消失。

4.1 挑战 1：发现、合并和清洗事件数据

目前，抽取适合过程挖掘的事件数据仍然需要花费相当大的精力。一般来讲，有几个需要克服的障碍：

- 数据可能分布在多个数据源上。这类信息需要合并，当不同标识符被应用在不同数据源中时将出现问题。例如，一个系统用名字和生日来标识一个人，而另一个系统却使用这个人的社会保险号来标识。
- 事件数据往往“以对象为中心”，而不是“以过程为中心”。例如，独立产品、集装箱和容器可能使用 RFID 标签，并且被记录的事件会指向这些标签。但是，为了监控一个特定客户订单，这类以对象为中心的事件需要被合并和预处理。
- 事件数据可能是不完备的。一个普遍的问题是事件不会显式指向过程实例。通常，有可能推导出这个信息，但是这将花费相当多的精力。此外，有些事件的时间信息有可能会丢失。为了能继续使用时间信息，只好人为添加时间戳。
- 一个事件日志可能包含异常值，即异常行为，也被称为噪音。怎样定义异常值？如何检测这些异常值？这些问题需要通过事件数据的清洗来解答。

- 日志会包括不同粒度级别的事件。在一个医疗信息系统的事件日志中，事件可能是简单的验血，也可能是复杂的外科手术过程。此外，时间戳也有不同粒度级别，范围从精细的毫秒级别（28-9-2011：h11m28s32ms342）到粗糙的日期信息级别（28-9-2011）。
- 事件发生在特定环境中（如天气、负载、一周的某一天等等）。环境可能用来解释特定现象，比如，因为工作中或者假期会导致响应时间比平时长。为了便于分析，将环境因素考虑在内是合适的。这意味着事件数据与环境数据的合并。在分析中，当加入过多变量后产生的“维数灾难”会使得分析变得非常棘手。

通常，需要比较好的工具和方法来处理上面的问题。此外，在前面提过，组织要将事件日志视为头等公民，而不是某些副产品。目标是为了获得★★★★★质量级别的事件日志（见表 1）。利用数据仓库环境下学习到的经验，可以提高事件日志的质量。例如，在数据载入的时候进行简单的检查，可以在很大程度上减少错误事件数据的比例。

4.2 挑战 2：处理包含多种特征的复杂事件日志

各类事件日志的特征有着非常大的差异。有些事件日志特别大，以至于难以处理，而有的事件日志特别小，以至于没有足够的可用数据来得出可信的结论。

在一些领域，记录的事件多到令人难以置信。因此，为了提高性能和可扩展性需要做出额外的努力。例如，ASML 持续监控它的所有晶圆扫描仪，这些晶圆扫描仪被许多组织使用（例如，三星电子和德州仪器）来生产芯片（大约 70% 的芯片是利用 ASML 的晶圆扫描仪生产的）。现存工具处理在这个领域收集的 PB 级数据时困难重重。除了记录的事件数量，还有其他特征，比如每个案例中的平均事件个数、案例的相似性、不同事件的个数、不同路径的个数。考虑一个事件日志 L1，它包含以下特征：1000 个案例，每个案例平均包含 10 个事件和很少的变化（例如，多个案例走的路径相同或者相似）。事件日志 L2 仅仅包含 100 个案例，但是平均每个案例有 100 个事件，并且所有案例的路径都是不同的。显然，L2 比 L1 更难分析，尽管两个日志有相同的大小（约 10,000 个事件）。

由于事件日志只包含样例行为，它们不应该被假定为完备的。过程挖掘技术需要通过“开放世界假设”来处理不完备性，所谓“开放世界假设”指的是：某些事情没有发生并不意味着它不能发生。这给处理包含许多变化的小规模事件日志带来了挑战。

正如前面提到的，一些日志包含抽象级别很低的事件。这些日志的规模倾向于极大，并且利益相关者对单个的低级别事件几乎没有兴趣。所以，应该将低级别事件聚集成高级别事件。例如，在分析一组特定病人的诊断和治疗过程时，可能对医院实验室信息系统中单个病人的化验记录并不感兴趣。

在此情况下，组织需要用反复试验法来看一个事件日志是否适合过程挖掘。因此，工具应该允许对给定的特定数据集进行快速的可行性测试。这样的测试会指出潜在的性能问题，并且对非常不完备的日志或过于细节的日志给予警告。

4.3 挑战 3：构建广泛接受的基准

过程挖掘是一门新兴技术，这解释了为什么它仍然缺乏一个好的基准。尽管有许多可用的过程发现技术，并且不同供应商提供了多种不同的产品，但是在这些技术的质量方面还没有形成统一的标准。由于在功能和性能上存在着巨大的差异，将不同的技术和工具进行比较仍然很困难。因此，由样本数据集和典型质量指标组成的好基准仍然需要不断探索。

对于经典的数据挖掘技术，有许多好的基准可用，这些基准激励工具提供商和研究人员不断改进各自技术的表现。在过程挖掘环境下，这就更富有挑战性。例如，1969 年由 E. F. Codd 提出的关系模型很简单并且得到了普遍支持，只需要花费很少的精力就可以将数据从一个数据库迁移到另一个数据库，并且不需要进行额外解释。对于过程来说，目前尚不存在这样一个简单模型。被提出的用于过程建模的标准要复杂的多，只有几个供应商准确地提供与这些概念集相对应的支持工具。过程的复杂性也远远超过了表格数据。

然而，为过程挖掘构建广泛接受的基准是很重要的。一些早期工作已经开展，例如，有多种用来度量过程挖掘结果质量的指标（拟合度、简洁度、精确度和泛化度）。此外，一些事件日志也被广泛应用（参考www.processmining.org），比如，由过程挖掘工作组组织的第一届业务过程智能挑战赛（BPIC'11）所使用的事件日志(http://dx.doi.org/10.4121/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54)。

一方面，应该有基于实际真实数据集的基准。另一方面，也有必要创建合成数据集，用来捕捉特定特征。这样的合成数据集有助于开发特定的过程挖掘技术，如针对不完备事件日志、有噪音的事件日志，或者特定的过程群体。

除了构建广泛接受的基准，也需要在某些判别过程挖掘结果的质量标准上达成一致意见（也可见挑战 6）。此外，源于数据挖掘的交叉验证技术可被用来判断结果的质量。考虑 k 折校验，可以通过将事件日志分成 k 个部分，k-1 个部分用于学习得到一个过程模型，符合性检查技术可以用剩余部分来判断结果的质量。这个过程可以重复 k 次，这样从某种程度上提供了对模型质量的深层次认识。

4.4 挑战 4：处理概念漂移

概念漂移这个术语是指：过程在被分析的同时发生改变的情况。例如，在事件日志的开始，两个活动可能是并发关系，但是不久以后，日志中这些活动变成了顺序关系。过程可能周期性或者季节性的发生变化（例如，“在十二月份有更多的需求”或者“在星期五下午有较少的可用员工”），或者随着环境改变（例如，“市场竞争越来越激烈”）。这些变化影响了过程，这对于检测和分析它们极其重要。一个过程中的概念漂移可以通过将事件日志分割成更小的日志，分析这些小日志的“足迹”来发现。这种“细化”分析需要更多的事件数据。然而，只有少数过程是处于稳定状态的，理解概念漂移对于过程管理来说才更加重要。当然，为了充分分析概念漂移，需要开展新的研究和工具支持。

4.5 挑战 5：改进用于过程发现的表示方式

过程发现技术用一种特定语言（如 BPNM 或 Petri 网）描述结果模型。但是，将结果的可视化和在实际发现过程中使用的表示形式分开是很重要的。选择一个目标语言经常包含若干隐含的假设，它限制了探究空间：不能被目标语言表示的

过程不能被发现。这个在发现过程中使用的所谓的“表示方式”应该是一个独立选择，并且不应该(只)被偏好的图形表示方式驱使。

考虑图 6 中的例子：目标语言是否允许并发，对发现模型的可视化和算法考虑的模型类别都会产生影响。如果表示方式不允许并发（图 6(a)是不可能存在的），也不允许多个活动有相同的标签（图 6(c)也是不可能存在的），那么只有如图 6(b)所示的问题模型可能存在。这个例子说明选择一种更严谨和精细的表示方式非常必要。

4.6 挑战 6：权衡拟合度、简洁度、精确度和泛化度等质量标准

事件日志通常是不完备的，即只记录了一些样本行为。过程模型往往允许产生指数级别甚至无穷数量的不同轨迹（循环存在时）。此外，一些轨迹可能比其它的轨迹出现的频率低很多。因此，假设在事件日志中每个可能的轨迹都出现是不现实的。将日志理所当然的认为是完备的，这种想法并不现实。考虑一个包括 10 个活动并行执行的过程和一个对应的包括 10,000 个案例信息的日志。包括 10 个并发活动的模型可能产生的所有交叉路径数量为 10！=3,628,800。因此，每个交叉路径不可能都在日志中出现，因为案例（10,000）比潜在的轨迹（3,628,800）要少得多。就算日志中有数以百万计的案例，所有可能的变化都存在也不太可能。一个额外的复杂性在于一些可选路径的发生概率比其它路径小得多，这些发生概率小的路径可能被认为是“噪声”。对于这种类型的噪声行为，不太可能建立一个合理的模型。被发现模型需要对噪声进行抽象，最好用符合性检查检测出低概率行为。

噪声和不完备性使得过程发现面临严峻的挑战。实际上，有四个互相竞争的质量维度：(a)拟合度，(b)简洁度，(c)精确度，(d)泛化度。一个有好拟合度的模型允许出现事件日志中记录的大部分行为。当日志中的所有轨迹都能够在模型中被从头到尾的重放时，这个模型就有最好的拟合度。能解释日志中可见行为的最简单模型就是最好的模型，这个原则被称为“奥卡姆剃刀”。单独使用拟合度和简洁度不足以判断一个被发现过程模型的质量。例如，很容易构建一个可以重现日志中所有轨迹的极端简单的 Petri 网（“花儿模型”），其它 事件日志也会指向同一活动集合。类似的，只能产生事件日志中可见的那些行为的模型并不受欢迎。因为日志中只包含了样本行为，有很多可能存在的轨迹在日志中并未记录。如果一个模型不允许存在“更多的”行为，则它是精确的。很明显，“花儿模型”不够精确，一个不精确的模型是“欠拟合”的。欠拟合是指模型对于日志中的样本行为过于泛化（即模型允许产生与日志中可见行为非常不同的行为）。一个模型应该允许泛化，而不应严格局限于日志中出现的行为。一个没有泛化的模型是“过拟合”的。过拟合是指产生的模型只包含日志中可见的行为（即模型详细解释了样本日志，但是同一过程产生的下一个样本日志可能产生完全不同的过程模型）。

权衡拟合度、简洁度、精确度和泛化度是一项很有挑战性的工作，这是大多数较为强大的过程发现技术提供多种参数的原因。需要开发改良的算法，以更好地平衡四个相互制约的质量维度。此外，使用的任何参数都应该容易被最终用户理解。

4.7 挑战 7： 跨组织挖掘

传统上，过程挖掘应用于单一的组织内部。但是，随着服务技术、供应链集

成和云计算的普及，会遇到很多包含多组织的事件日志分析场景。原则上，对于跨组织的过程挖掘来说，存在以下两种情况。

第一，我们可考虑不同组织共同工作，处理过程实例的协作环境。可以将这样的跨组织过程视为“七巧板游戏”，即整个过程被分割为多个部分，并且分布于多个组织中，它们需要合作以成功完成整个案例，只分析其中一个参与组织的事件日志是不够的。为了发现端到端的过程，不同组织的事件日志需要进行合并。这并不是一项轻而易举的任务，因为事件需要跨越组织边界进行关联。

第二，我们也可考虑不同组织环境中执行相同的过程，同时分享经验、知识，或者一个共同的架构。以 Salesforce.com 为例，许多组织的销售过程被 Salesforce 管理和支持。一方面，这些组织共享一个底层架构（过程、数据库等）；另一方面，它们没有被强制遵守一个严格的过程模型，因为系统可以被配置成支持同一过程的多种变体。考虑另外一个例子，在任何城市中执行的基本过程（例如，建筑许可证发放）。尽管一个国家的所有城市都需要支持相同的一系列基本过程，但仍有可能存在不同。很明显，在不同组织中分析这种变化是很有意义的。这些组织可以互相学习，服务提供商也可以提升他们的服务，并且基于跨组织过程挖掘的成果提供增值服务。

需要为两种类型的跨组织过程挖掘开发新的分析技术。这些技术也应该考虑隐私和安全问题，组织间也许因为竞争或缺乏信任的原因而不想分享信息。因此，开发隐私保持的过程挖掘技术是很有必要。

4.8 挑战 8：提供在线运作支持

最开始，过程挖掘的焦点是历史数据的分析。但是今天，许多数据源都是（接近）实时更新的，当它们出现的时候，应该有充足的计算能力完成事件的分析。因此，过程挖掘不应该限制为离线分析，也可以用来提供在线运作支持。可以确定三种运作支持活动：检测、预测和推荐。当一个案例偏离预定义过程的时候，可以被系统检测到并且由系统产生一个警告。通常，我们希望立刻产生这样的通知（仍可影响到系统运行），而不是以离线的方式产生。历史数据可以用来构建预测模型，这些模型被用来指导正在执行的过程实例。例如，可能预测出一个案例的剩余处理时间。基于这样的预测，我们也可以构建一个能够建议特定行动的推荐系统，以减少开销、缩短流程时间。以这样的在线执行方式应用过程挖掘技术，对计算能力和数据质量方面提出了新的挑战。

4.9 挑战 9：融合过程挖掘与其它类型的分析技术

运作管理，特别是运筹学，是严重依赖于建模的管理科学的一个分支。这里使用了各种数学模型，从线性规划和项目规划到排队模型、马尔科夫链和仿真。数据挖掘被定义为“通过对数据集（通常是大数据集）所做的分析，来寻找未知关系，以及用一种容易被数据所有者理解且有用的新型方式总结概括数据”。已经开发了很多技术：分类（如决策树学习）、回归分析、聚类（例如，k 均值聚类）和模式发现（如关联规则学习）。

运作管理和数据挖掘两个领域都提供了有价值的分析技术，其中的挑战是将这两个领域的分析技术与过程挖掘融合起来。以仿真为例，过程挖掘技术可基于历史数据学习，得到一个仿真模型。随后，仿真模型可被用于提供运作支持。因

为事件日志和模型间的紧密联系，模型可以被用于重放历史，也可以从当前状态开始仿真，于是提供一个基于实时数据到达未来的“快进钮”。

类似的，将可视分析和过程挖掘结合起来也很有意义。可视分析将自动分析与交互可视化结合起来，以更好的理解大型复杂数据集。可视分析开创了人类从非结构化数据看到模式的惊人能力。通过结合自动化过程挖掘技术和交互式可视分析，有可能从事件数据中得到更多深刻的见解。

4.10 挑战 10：方便普通用户使用

过程挖掘的目标之一是创造“实时过程模型”，即每天使用的过程模型，而不是终止于某个归档的静态模型。新产生的事件数据可被用于发现涌现的行为，事件数据和过程模型之间的联接允许将当前的状态和最近的活动投影到最新的模型上。这样，最终用户可以每天都和过程挖掘的结果进行交互，这种交互非常有价值，但是也需要直观的用户界面。这个挑战是：将复杂的过程挖掘算法隐藏到可以自动设置参数并提供适当分析类型的、友好的用户界面之后。

4.11 挑战 11：使普通用户易于理解

即使很容易产生过程挖掘结果，并不意味着结果是真正有用的。用户可能不能理解输出的内容，也可能被误导到不正确的结论。为了避免这样的问题，结果应该用一种适合的表示方式展示出来（也可见指导原则 5）。此外，结果的可信性也应该总是被清楚的表达出来。事实上，现存的过程发现技术一般不会对拟合度过低或者过高的结果提出警告，它们总是显示一个模型，即使有时很明显由于数据量过小而无法得到任何可信的结论。

5. 结语

IEEE 过程挖掘工作小组的目标是：(a)推动过程挖掘的应用；(b) 指导软件开发者、咨询顾问、业务经理和最终用户恰当使用最新技术；(c)激励过程挖掘方面的研究。这个宣言阐述了工作小组的主要原则和目的。在引入了过程挖掘的主题后，宣言列出了一些指导原则（第 3 节）和挑战（第 4 节）。指导原则可被用来避免明显的错误，列出一系列挑战的目的是指导研究和开发工作，这两者都是为了提高过程挖掘的成熟度。

最后，对其中的一些词汇进行说明。下面的术语被应用于过程挖掘空间：工作流挖掘、（业务）过程挖掘、自动（业务）过程发现和（业务）过程智能。不同组织往往会对这些概念使用不同的术语。例如，Gartner 公司正在推销术语“自动业务过程发现（ABPD）”，Software AG 公司使用术语“过程智能”指代他们的控制平台。术语“工作流挖掘”看起来似乎不是特别合适，因为工作流模型的建立仅仅是过程挖掘的许多可能应用之一。类似的，短语“业务”的增加缩小了过程挖掘应用的范围。有许多过程挖掘应用（例如，高技术系统的使用分析或者网站分析）在增加了这个短语的时候变得不是很适合。尽管过程发现是过程挖掘领域的一个重要部分，它仅仅是许多用例中的一个。符合性检查、预测、组织挖掘、社会网分析等，是过程发现以外的其它扩展用例。

图 7 将刚才提到的几个术语关联起来。为支持决策制定提供可执行信息当做

目标的所有技术和方法，都可以放置于商务智能（BI）这把大伞之下。(业务)过程智能可被看作是商务智能（BI）和业务过程管理（BPM）结合的产物，即商务智能技术被用于分析和改进过程及其管理。过程挖掘可以被看成是以事件日志为起点的过程智能的具体化。（自动业务）过程发现仅仅是过程挖掘的三个基本场景之一。图 7 可能存在某种程度上的误导，让人们认为商务智能工具不提供本文档所描述的过程挖掘功能。BI 这个术语经常被轻易的用到一个特定的工具或方法上，仅覆盖了广义 BI 频谱的一小部分。

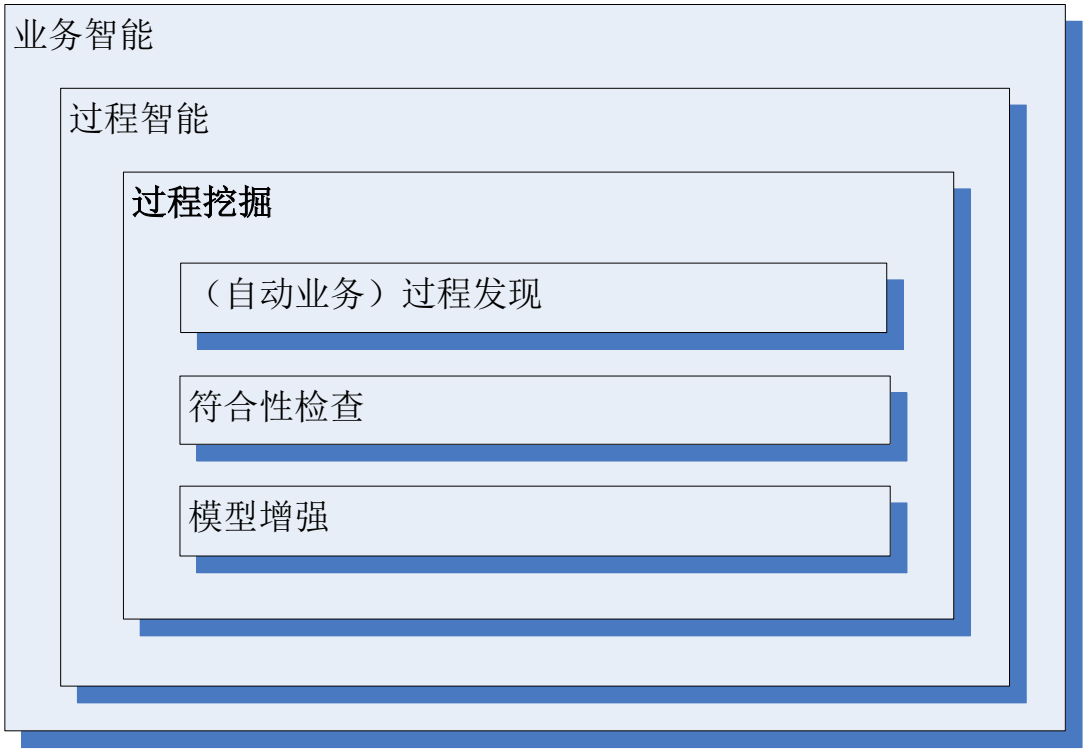


图 7：关联不同术语

使用这些可选的术语是有商业原因的，一些厂商可能想要强调一个特定的方面（例如，发现或智能）。但是，为了避免混淆，最好使用本宣言中包含的术语“过程挖掘”作为规范。

参考文献

1. W.M.P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer-Verlag, Berlin, 2011.

词汇表

活动：过程中一个良好定义的步骤。

自动业务流程发现：见**过程发现**。

商务智能 (BI)：使用数据来支持决策制定的工具和方法的广义集合。

业务过程智能：见**过程智能**。

业务过程管理（BPM）：将信息技术与管理科学中的知识进行结合，应用到运作业务过程中的学科。

案例：见**过程实例**。

概念漂移：过程随着时间的变化而变化的现象。被观察的过程由于季节变化或竞争增加，逐渐（或突然）产生变化，从而使分析复杂化。

符合性检查：分析日志中记录的现实，是否符合模型，反之亦然。目的是为了检测差异性，并衡量其精确性。符合性检查是过程挖掘三个基本场景中的一种。

跨组织过程挖掘：过程挖掘技术应用到来自不同组织的事件日志中。

数据挖掘：分析（通常是大规模的）数据集来发现非预期的关系，并且从新的视角总结数据。

事件：在日志中记录的动作，例如特定过程实例中的一个活动的开始、完成或取消。

事件日志：被用作过程挖掘的输入的一系列事件。事件不一定存储在一个单独的日志文件中（例如，事件可能分散在不同的数据库表中）。

拟合度：一种衡量给出模型是否能允许事件日志中记录行为的方法。当日志中所有的执行序列可以在一个模型从开始到结束重演时，这个模型有最高的拟合度。

泛化：一种衡量模型是否能允许未见行为的方法。一个过拟合模型是不够泛化的。

模型增强：三种过程挖掘场景中的一种。使用一些日志中提取出的信息来获得扩展和改进一个过程模型。例如，检测时间戳时，通过重演过程模型中的事件日志来发现瓶颈。

MXML：为了方便交换事件日志提出的基于 XML 的格式。目前，XES 替换了 MXML 成为不依赖于工具的、新的事件日志格式。

运作支持：为了监控和影响运行过程实例，对事件数据进行在线分析。三种运作支持活动：检测（如果观察的行为偏离模型行为，产生一个报警）、预测（根据过去的行为预测未来的行为，如预测剩余处理时间）和推荐（建议合适的行为来实现特定目标，如最小化成本）。

精确度：一种衡量模型是否允许特别偏离日志中可见行为的方法。一个低精确度的模型是欠拟合的。

过程发现：过程挖掘三个基本场景中的一种。基于事件日志获得过程模型。例如，α 算法可以通过在一系列事件日志中识别过程模式，从而发现 Petri 网。

过程实例：被分析的过程所处理的实体。事件指向过程实例。过程实例的例子包括用户订单、保险索赔、贷款申请等。

过程智能：商务智能的一个分支，关注业务过程管理。

过程挖掘：通过从现代（信息）系统中的事件日志中提取信息，发现、检测和改进实际过程的技术、工具和方法。

表示方式：为了表达和构造过程挖掘结果，选定的目标语言。

简洁度：衡量业务过程模型复杂程度的“奥卡姆剃刀”的方法。能够解释日志中行为的、最简洁的模型是最好的模型。简洁度可以用各种方式量化，如模型中节点和弧的数量。

XES：基于XML的事件日志标准。这个标准已经被IEEE过程挖掘工作组采用，作为事件日志的默认交换格式（查阅 www.xes-standard.org）。