arXiv:2304.02699v1 [cs.HC] 5 Apr 2023

Tracing and Visualizing Human-ML/AI Collaborative Processes through Artifacts of Data Work

Jen Rogers
University of Utah
Salt Lake City, UT, USA
jen@sci.utah.edu

Anamaria Crisan
Tableau Research
Seattle,WA, USA
acrisan@tableau.com

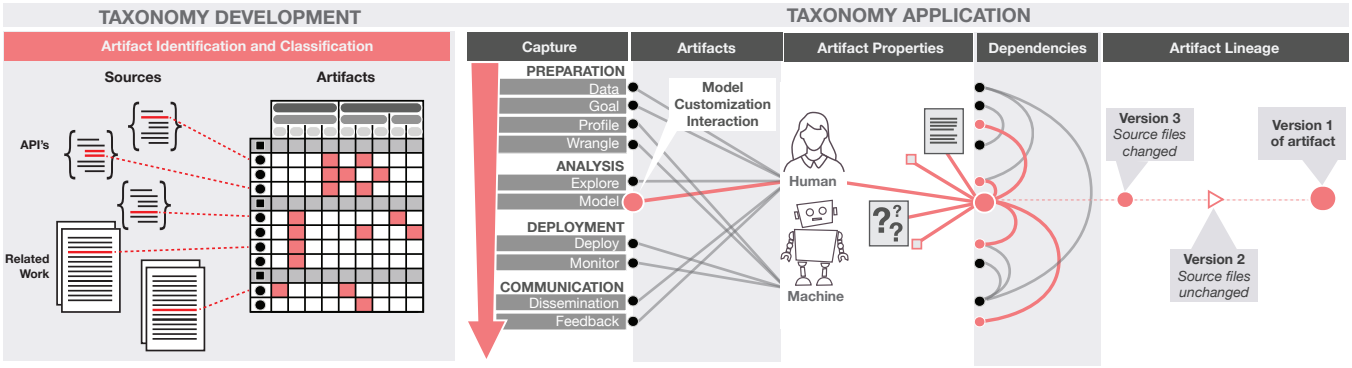

Figure 1: We developed an artifact taxonomy that captures both human and ML/AI processes in automated data work. We assess our taxonomy's utility through a collaboration with an enterprise software development team creating an AutoML system. .

ABSTRACT

自动化机器学习（AutoML）技术可以降低数据工作的门槛，但要发挥作用，仍需要人工干预。然而，人类与机器分担工作的复杂协作流程，使得追踪工作由谁（或什么）在何时完成变得困难。在本研究中，我们构建了一个数据工作产物分类体系，用于捕捉自动化机器学习和人类过程。我们提出了创建该分类体系的严谨方法，并探讨其向视觉设计过程的迁移能力。我们通过开发 AutoML Trace 将该分类体系付诸实践：这是一种可视化交互草图，展现了数据工作中人-ML/AI协作的情境与时间性。最后，我们通过一个企业软件开发团队的使用场景，展示了该方法的实用价值。总体而言，我们的研究过程与发现探讨了开发数据可视化工具所面临的挑战与可行方向，以审视自动化数据工作中的社会技术关系。

Availability of Supplemental Materials: https://osf.io/3nmyj/?view_only=19962103d58b45d289b5c83421f48b36

允许为个人或课堂使用而免费制作本作品全部或部分内容的数字或纸质副本，但不得为营利或商业利益而制作或分发副本，且副本首页必须附有本声明和完整引文。必须尊重本作品中由 ACM 以外其他方拥有版权的组件。注明出处后可以摘录。以其他方式复制、转载、发布到服务器或再分发至列表，均须事先获得明确许可和/或支付费用。请通过 permissions@acm.org 申请许可。
CHI '23，2023年4月23日至28日，德国汉堡© 2022年 Association for Computing Machinery。ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/1122445.1122456

CCS CONCEPTS

• **Human-centered computing** → *Visualization theory, concepts and paradigms*; • **Computing methodologies** → *Artificial intelligence.*

KEYWORDS

AutoML, Taxonomy, Data Visualization, Human-Machine Collaboration

ACM Reference Format:
Jen Rogers 和 Anamaria Crisan. 2022年。通过数据工作的产物追踪和可视化人类与机器学习/人工智能的协作流程。载于 汉堡 *'23*：*ACM*计算系统中的人为因素会议，*2023*年*4*月*23*日至*28*日，德国汉堡。ACM，美国纽约州纽约，22页。https: //doi.org/10.1145/1122445.1122456

1 INTRODUCTION

数据工作包含多个相互关联的阶段，这些阶段利用统计和计算技术进行数据准备、分析、部署和交流 [15]。开展数据工作所需的技能仍然相当复杂，使得许多具备相关领域背景但缺乏技术能力的专家难以企及 [44]。近期的创新开发了自动执行数据工作的技术，例如模型选择或某些数据准备步骤，从而降低了非技术专家的使用门槛 [18]。最初，这种所谓的自动化机器学习技术（自动化机器学习）主要聚焦于分析阶段。然而，近期研究正在推动自动化机器学习的边界，以涵盖更端到端的数据工作流程 [14, 36, 44, 119]。自动化机器学习的扩展范围现在可以

涉及自动化数据准备阶段的整理、分析阶段的超参数调优和算法选择、已部署模型中的漂移警报，甚至用于交流和传播的自动生成报告。然而，在实践中，自动化机器学习仍需要大量人类劳动和协调才能发挥作用 [14, 21, 109]。此外，这些先前的研究指出，当自动化机器学习仍是一个"黑箱"时，数据团队之间存在摩擦，并且难以探究人们如何与自动化机器学习技术协作完成数据工作。即使完全自动化成为可能，人类监督和干预仍然是被期望的 [14, 44, 50, 101]。不幸的是，现有的自动化机器学习工具很少考虑这一人类要素，导致即使该技术不断进步，仍有许多未满足的需求。为了弥合这一差距，我们探索了可视化分析如何帮助技术专家和非技术专家追踪由自动化机器学习辅助的数据工作。

为了认可自动化机器学习与人类共同承担的劳动，我们将自动化数据工作中的可追溯性挑战视为人-ML/AI协作的一部分。HCI 社群近期的研究 [112] 强调了人-ML/AI协作中的若干困难，其中两个问题尤其推动了我们的研究。第一个挑战是，人-ML/AI协作增加了对机器学习/AI系统能力和输出的不确定性。他们指出，现有的设计方法难以应对这种不确定性。尽管已有技术和可视化系统通过溯源追踪和审计使自动化机器学习流程更加透明（例如，[70, 84, 87, 103]），但这些方法通常聚焦于管道的分析阶段，优先考虑机器学习工程师和数据科学家，而不是技术专长较少的数据工作者。对于自动化机器学习系统的开发而言，这一差距十分重要，因为它没有考虑参与人-ML/AI协作的团队的多样性 [4, 14, 37, 40, 55, 72, 102, 109]。开发能够为完整范围的数据工作者提供透明度支持的系统至关重要，尽管这颇具挑战性 [81]。这反映了 Yang 等人在其工作中识别出的第二个挑战：面向用户的研究人员与ML/AI 工程师之间的密切协作十分重要，但由于缺乏对这种跨学科对话的调解机制，这种协作存在障碍，例如"用于支撑的共享工作流、边界对象或通用语言" [112]。

我们在与企业团队的工作中遇到了这些人类-机器学习/AI协作的挑战。我们的初始目标是开发一种解决方案，用于可视化追踪由他们的软件促进的数据工作流中的人类和AutoML流程。然而，我们意识到协作缺少一种用于共享讨论的通用语言，以支持他们正在开发的AutoML系统。通过更广泛的HCI文献的视角，特别是Yang等人，我们所缺乏的是边界对象——抽象或具体的信息，这些信息为与软件团队的协作建立了共享理解 [43]。我们确定，在开发可视化分析工具之前，我们首先需要表征AutoML流水线中可以从人类和自动化流程中捕获的内容。

人际协作与知识共享领域的研究强调，捕捉工件 [47, 49] 对于追踪 [27, 59, 95] 复杂协作流程至关重要。基于这些先前研究以及 Yang 等人 [111], 的工作，我们提出了一种通过以下方式使自动化数据工作可追踪的方法：开发 AutoML 工件分类法和 AutoML Trace 可视化工具。我们的分类法基于对现有和理论上的自动化机器学习及人类与机器学习/人工智能交互系统的考察。它界定了人类和自动化机器学习所产生工件的广泛范围。工件的确切含义取决于其所处的语境。在我们的研究中，工件指数据工作过程中由人类（即目标、任务、文档、数据集、源代码等）或自动化机器学习流程（例如特征集、模型选择、自动化洞察等）生成的有形或抽象项目。该分类法充当边界对象，为与软件团队的对话提供了支架，使我们能够围绕其现有系统和未满足的终端用户需求进行构思。我们进一步通过 AutoML Trace 将该分类法付诸实践——这是一种高保真可视化交互草图，用 Greenberg 和 Buxton [35],的话来说，它旨在"将模糊的想法具体化，反思可能出现的问题及用途，发现其他新想法，并完善现有想法。"。我们开发的是交互式草图，而非高保真、功能更全面的原型，这让我们能够更灵活地与软件团队开展协同设计过程 [82]。AutoML Trace 可识别、捕捉并情境化由我们的分类法界定的工件，并展示它们的依赖关系及其随时间的演变。虽然我们将 AutoML Trace 应用于协作者的自动化机器学习软件，但它也可以应用于其他软件。

总体而言，我们的研究提出了三项贡献。**本研究的第一项贡献是 AutoML 工件分类法**，该分类法涵盖了数据工作自动化连续谱中人-ML/AI协作的细微差异。本研究的另外两项贡献，是在分类体系的制定过程以及我们与企业团队的合作中自然发展而来的。**第二项贡献是 AutoML Trace 交互式草图，它将我们的分类体系付诸实践**，并作为与自动化机器学习系统互动的媒介。**最后，第三项贡献是对可追溯性的定义**，该定义阐明了可理解、可观察的数据工作流所包含的内容。虽然我们主要关注数据工作自动化面临的挑战，但我们也探讨如何运用分类法和可视化草图，广泛地开发用于设计人-ML/AI协作的框架和系统。

2 RELATED WORK

We review related work concerning taxonomies' utility for creating a common language within and between complex systems, existing taxonomies for AutoML and data visualization, and existing visualization systems for AutoML that can surface these artifacts.

2.1 Taxonomies, Ontologies, and Schemas

分类法为知识提供结构，并能够实现项目之间的比较和关系识别 [67]。可视化、HCI 和机器学习社区使用分类法来指导系统的开发、定义需求，并提供交流的通用语言 [20, 48, 51]。我们希望我们的分类体系也能发挥同样的作用。然而，我们力求以严谨的方式开发我们的自动化机器学习工件分类体系，以确保我们的工作建立在坚实的理论基础之上。我们的分类法开发受到 Nickerson *et al.* 在信息系统中进行严谨分类法开发工作的启发 [67], ，该工作的动机是

他们自己社区中经常采用的临时方法来构建分类法。我们回顾了自动化机器学习和数据可视化中的现有分类法，以了解它们各自的概念特征、效用和粒度，以及与我们分类体系的关系。我们将现有分类法和类似工作分为三组：机器学习流程、人在回路自动化和可视化分析。

2.1.1 机器学习流程中的溯源性、可追溯性与可复现性。我们并不是第一个将机器学习和人工智能流程形式化为分类体系的人。Tatemen 等 [93] 提出了一个关于机器学习研究可复现性的分类体系。他们的研究根据研究流程所产生的工件，识别出从低到高的可复现性示例。出于类似的可复现性目标，Publio 等 [74] 提出了 ML-Schema，这是一种用于表示和交换机器学习流程制品的本体，其中包括代码、数据和实验文档。他们旨在自动生成机器学习模型元数据描述符，以提高机器学习流程的可解释性。Souza 等 [87] 基于 ML Schema 以及PROV-DM，创建了一个用于追踪多个机器学习工作流溯源信息的特定模式。虽然这些关于机器学习溯源性的分类法和模式十分重要，但它们并未充分考虑人类过程以及各个阶段的人为干预方式，而这正是我们的研究试图实现的目标。不过，在构建我们的分类体系时，我们也考虑了现有分类体系如何与我们的分类体系相连接，从而为特定的数据科学流程增添更细致的详情。

2.1.2 人在回路与混合自动化。在对可视分析中溯源进行刻画时，Ragen 等人展示了特定工作流的异质性，以及"记录用户想法、分析推理和洞察"的重要性 [75]。近期研究 [19, 97] 提出了一些分类体系，开始明确纳入机器学习流程中的各种人工生成工件。Dellerman 等人 [19]聚焦于自动化机器学习技术中的人工干预；其研究在理念上与我们的工作最为接近，并且采用了与我们相同的方法来构建分类体系。不过，这些分类体系主要关注模型优化阶段，而我们的研究则将其视为一个端到端的数据科学流程，涵盖从准备到交流的各个环节。人机交互（HCI）和计算机支持的协同工作（CSCW）领域的分类体系 [44, 101] 提出了在端到端数据科学流程中，将不同程度的自动化与人类协作相结合的方法。Karamaker 等人 [44]根据任务成功实现自动化的程度，提出了六个自动化级别。其附录详细介绍了不同的机器学习方法、自动化范围以及人工干预的作用。Wang等人。 [101] 提出了类似的人类主导和系统主导的自动化级别，并将其置于一个更大的人在回路自动化机器学习框架中加以描述。

2.1.3 机器学习溯源、可追溯性和模型的可视化。由于我们的方法探讨了如何通过数据可视化呈现制品，我们考察了可视化领域的先前研究。Sacha 等 [77] 提出了一个用于可视化辅助机器学习的本体，该本体契合人类参与其中的机器学习/人工智能范式。它将制品表示为构成数据、模型或知识的输入和输出实体；然而，他们没有提供关于这些实体属性的更细粒度信息。Spinner 等 [90] presents a framework for explainability in visual and interactive ML whose processes align with those of automated data science processes driven by AutoML technology. They also primarily view artifacts as input/output entities but do not further define what those entities are.

2.1.4 弥合差距。 这些不同的分类法、本体和框架都旨在定义自动化数据科学工作中的一组实体和操作。然而，它们缺乏对数据工作中生成或共享的实体的一致描述。我们提出将工件作为这一实体。通过构建我们的分类体系，我们认为我们的研究可以帮助弥合这些既有工作之间的差距。

2.2 AutoML Visualization Systems

机器学习管道的交互与可视化既能促进用户参与和干预，也能增强用户对机器学习流程结果的信任 [8]。近年来，涌现出了许多用于自动化机器学习的可视化工具。ATMSeer [103] 自动搜索机器学习模型，并通过自动生成的联动视图仪表板，向最终用户展示搜索空间的汇总统计信息。ModelLineUpper [64]也使用不同视觉编码的多联动视图，比较由自动化机器学习流程生成的机器学习模型。AutoVizAI [104]同样探索模型配置这一狭窄范围，但使用条件平行坐标图来可视化模型在各种可能配置下的生成过程。最后，Visus [83] 着眼于领域专家如何具体运用自动化机器学习来构建模型。

其他系统更广泛地看待 AutoML 流程，超越了建模阶段。PipelineProfiler [70] 与 Jupyter 笔记本集成，并使用与对齐视图并置的矩阵来提供结果概览，以指示 AutoML 流水线在每一步中的不同组件和输出。AutoDS [100]使用网络图来展示配置端到端 AutoML 流水线的不同可能方式。AutoDS 可作为独立工具存在，或嵌入 Jupyter 笔记本中。Boba [56] 系统及其底层 DSL 使用与 AutoDS 类似的可视化设计，来可视化不同数据科学流程的阶段和结果。Boba 的设计灵感源于 Liu [54] 早期进行的用户研究，这些研究通过网络图可视化了数据工作者的分析模式。Swatai 等人 同样发现网络图能有效捕获用户通过交互式分析流的各种路径 [60]。Xin 等人 [108] 利用这种图结构开发了将人类插入自动化机器学习流程的技术。研究也面向捕获用户与可视分析系统的交互；Knowledge Pearls [91] 和 Trrack [16] 是两个示例，它们也使用底层图来管理和可视化分析路径。

通过我们的分类体系，我们旨在拓展可视化的工件范围，并为工件补充有关其来源、依赖关系和历史的背景信息。在实现 AutoML Trace（第 6 节）时，我们借鉴了这些既有系统在视觉编码方面的选择。

3 TRACEABILITY FOR HUMAN-MACHINE COLLABORATION

追踪人类与机器学习/人工智能过程之间的协作关系，对于确保整个数据工作流程透明且可供审查，而不仅仅是最终产品（即模型或结果）[105]至关重要。长期以来，软件与设计工程领域 [76, 92], 、社会科学 [47, 49],以及知识管理社群 [27, 59, 95] 一直在探索工件的可追溯性，近来这一问题也开始在机器学习领域受到关注 [11, 63, 84]。然而，可追溯性的定义差异很大。在此，**我们将机器学习/人工智能的可追溯性定义为涵盖来源、透明度和背景**。来源是记录单个工件及其来源的过程，包括生成该工件的内容以及依赖于它的其他工件。透明度关乎理解模型如何得出结论的能力。最后，背景指明工件在分析中所处的位置。在此，我们提出追踪数据工作中的工件，涵盖从准备到交流的各个阶段，这些工件是在各阶段随时间推移而形成的人-ML/AI协作的产物。若工件有明确的定义，能够说明它是什么、如何以及何时生成，并且存在记录其变化过程的沿袭关系，我们就认为该工件具有可追溯性。

4 MOTIVATION AND METHODOLOGY FOR AN AUTOML ARTIFACT TAXONOMY

分类体系是一种广泛使用的知识组织系统，它根据概念共有的属性，将其分层归入逻辑关联之中 [67, 73]。分类体系提供了一种通用语言，使人们能够推测并进一步构建概念，从而促进由不同领域专家组成的团队内部的交流 [79]。以往的数据可视化研究采用过任务分类体系（例如， [9, 52, 96]）、数据分类体系（例如，[7]）以及可视化技术分类体系，以推动工具开发。自动化机器学习和人-ML/AI协作领域也开发了类似的分类体系（见第 2 节），但其对工具开发的影响尚不明确，且缺乏一致的开发机制。因此，文献中分类体系的稳健性在质量和范围方面可能存在显著差异。我们的分类体系整合并协调了现有的分类体系、框架、本体，以及现有系统和理论系统的产物，从而提供一套全面的自动化机器学习产物。我们采用了信息系统研究中的一种稳健方法，对简洁性、稳健性、全面性、广度和可解释性进行评估 [67, 73]。作为分类体系的贡献之一，我们介绍了开发方法（概述见图 2），以阐明稳健性在分类法创建中的重要性。

4.1 方法概述

Nickerson [67] 提出了一种多阶段、等集成的方法，用于定义和评估分类体系。他们的方法植根于其对分类体系的定义：一组根据分类描述符进行分类的对象；分类描述符是由维度、类别和特征构成的层级集合。对象可以指各种事物，例如生物、商店出售的产品类型，或工件（如本文所述）。

他们将分类法创建定义为三个阶段：预开发、开发和评估。预开发阶段 定义分类体系对象的元特征以及用于结束分类法开发的结束条件集合。随后，开发阶段 采用从实证到概念或从概念到实证的方法来定义对象及其属性。最后，在评估阶段 ，通过结合客观和主观标准，以迭代流程对分类体系进行评估。

回顾其方法论，Nickerson等 [67] 强调，分类体系是一种难以求解的“设计搜索流程”。然而，他们认为——我们也认同这一点——其方法论提高了最终分类体系的透明度、稳健性和可扩展性。下面，我们将详细介绍在这些分类法开发阶段所作的选择。我们研究流程中产生的制品，包括笔记、文档和材料，均来自分类法开发的 8 次迭代，可在线查阅[1]。受篇幅限制，有关分类体系及其开发的更多详情请参见补充材料，本文中以［SM1］标注（这些也是可点击链接）。本文末尾提供了这些补充材料的完整说明。

4.2 开发前期阶段

4.2.1 定义元特征。 分类法开发过程始于界定元特征的具体定义，用以描述所研究的对象（图 2.1）。**在本研究中，我们将分类体系中的对象定义为由人工或自动化机器学习驱动的任务生成并交换的自动化机器学习工件。这类任务贯穿端到端的数据科学工作流，其中涵盖数据准备、分析、模型部署和交流等流程。**

4.2.2 定义结束条件。 我们预先设定了一组客观和主观的结束标准，用于在每个开发阶段评估分类体系（图 2.1）。若在评估阶段满足这些标准，我们便结束分类法开发。分类体系在不同迭代中的结构稳定性也是客观结束标准的一部分。为满足这一结束条件，我们的分类体系应符合以下标准：

(1) 与上一次迭代相比，不新增或修改维度、特征或对象（工件）；(2) 不对维度、特征或对象（工件）进行合并或拆分；(3) 每个维度的每项特征下至少归类一个对象（工件）

The subjective ending conditions are defined by Nickerson *et als* [67] as the minimum criteria for the utility of a taxonomy. These subjective conditions include conciseness, robustness, comprehensiveness, extensibility, and explanatory. As these are subjective criteria, they serve as a function to reflect on the taxonomy's internal validity.

[1] https://osf.io/3nmyj/?view_only=19962103d58b45d289b5c83421f48b36. This is an OSF view-only link for the review process, meaning it does not collect any data that could identify reviewers

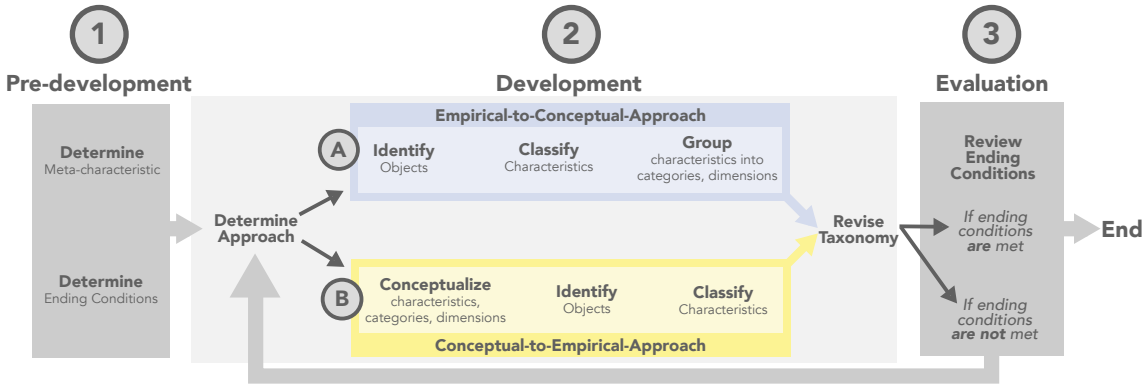

图 2：我们的分类法开发方法概览。我们遵循了 Nickerson等。 [67] 提出的方法。分类法开发过程由三个阶段组成。(1) 流程的预开发阶段包括定义元特征和开发的结束标准。(2) 开发阶段会在流程中重复进行，直到满足结束条件。该阶段采用两种方法之一：(A)从实证到概念，或 (B) 从概念到实证。(3) 我们在流程的评估阶段确定是否满足结束标准。如果未满足结束条件，我们会重复开发阶段，直到满足结束条件。

4.3 Development Stage

开发阶段始于从实证到概念的方法（图 2.2.A）或从概念到实证的方法（图2.2.B）。在前一种方法中，从可用的数据源中识别对象，通过定量（即统计聚类）或定性（即主题分析）方法进行分类，并根据一组不断涌现的属性（特征、类别、维度）进行分组。而在后一种方法中，则先对一组属性进行概念化，并用其识别数据源和对象，随后再对这些对象进行分类。每个开发阶段开始时所采用的方法可能有所不同。我们主要采用从实证到概念的方法来识别对象以进行分析。

4.3.1 文献来源。我们根据涵盖机器学习、人机交互、计算机支持的协同工作、信息可视化和可视分析领域的研究文献，定义了自动化数据工作中的人类生成和机器生成工件。我们采用两种方法抽样研究文献。首先，收集了由 13 篇论文组成的初始便利抽样，熟悉相关方法，并创建了初始分类体系。便利抽样中的论文是作者已知的论文，以及通过快速搜索“artifacts AutoML”“taxonomy AutoML”“capturing AutoML”和“visualizing AutoML”找到的论文，再由我们主观选取并讨论。接着，我们确定了一组关于“AutoML”的系统性已发表研究和预印本。搜索截至 2021 年 6 月 14 日，共检索到 153 篇文章，来源包括 KDD、AAAI、NeurIPS、CHI 等会议。大多数出版物来自 arXiv（153 篇中的 100 篇，占 65%），且发表于过去两年内。我们分析中使用的所有来源的完整清单，以及关于其使用方式的说明，见在线材料。我们对全部 153 篇论文进行了初步扫描。根据扫描结果，我们制定了纳入和排除标准。我们排除了范围过于狭窄、专注于某一特定技术亮点的论文。可用的文献来源清单见 [SM2]。

4.3.2 对象分类。我们从文献来源中识别并提取了约 400 个可能代表人类或机器生成工件的条目。首先，我们将这些条目粗略归类到数据工作流的各个阶段（准备、分析、部署和交流）[15]。在这些阶段内，我们进一步将条目划分为工件组。最后，我们利用这种分组构思了一组工件属性。我们采用开放式编码和轴心编码技术，推导出一组用于描述工件属性的特征、类别和维度。在编码过程中，我们使用了对象来源文献中的描述和定义。随着每轮编码逐渐明确工件及其属性的定义，我们将各个条目合并（例如，T-SNE 和 PCA 被合并为映射变换工件，因为二者都会将数据从较高维度映射到较低维度）。我们将最初的 400 个条目精简为 52 个工件。工件及其分类的完整清单见 [SM1] 和 [SM1-F]。

4.4 Evaluation Stage

在每个开发阶段之后，我们评估是否达到了我们的结束标准（图 2.3）。根据我们来自 [67]和 [73],的定义，如果在一个开发阶段结束时，能够使用一组充分且不过度的维度、类别和特征对对象进行全面分类，那么该分类体系就是 简洁、稳健 和 全面的。如果能够在迭代过程中轻松添加新的维度、类别和特征，那么它就是 可扩展 的。最后，如果它可用于描述对象的性质，那么它就是 解释性 的。

我们的分类法开发经过八轮迭代，才满足结束条件。两位作者阅读了文献来源，提取符合元特征定义的制品，将这些项目分类，最后根据不断演变的一组制品属性对其进行分组。两位作者连续一个月每天会面，讨论各自的分类结果。虽然最终达成了共识，但我们并未试图穷尽所有冲突、歧义或不同解读并加以解决。

5 自动化机器学习产物分类体系

我们的分类体系包含52种制品，根据其属性分为11组。我们依据4个维度、17个类别和41项特征来定义这些制品的属性。重要的是，没有任何一个自动化机器学习系统包含所有这些工件 [44]。相反，我们综合了各个自动化机器学习工具包、系统和理论研究论文中的设计决策。我们认为，广泛考察现有系统、了解它们是什么以及它们的目标是什么，能让我们的分类体系延伸至尚未开发的系统。图3总结了这些制品、其分组情况，以及它们所属的数据科学流程（以及交互流程）。

5.1 自动化机器学习产物

5.1.1 数据科学工作流的产物与流程。随着自动化机器学习系统中的创新不断发展，任务自动化的范围也随之扩大。截至本文撰写时，许多提出的系统尚无法投入实际使用 [44]。借鉴先前的框架，我们将端到端的数据科学工作流定义为由准备、分析、部署和交流流程构成。这些阶段也与Karmaker等人 [44] 为自动化机器学习系统提出的任务和自动化级别相对应。同样，由数据科学基本构件 [38, 70] 组成的自动化机器学习系统，也按照这些流程划分。虽然我们将这些流程应用于产物分类（第4.3节），但也发现，大多数产物通常只属于一个流程。例如，初始数据集是一种产物，通常由人类提供，属于数据准备阶段——未来的自动化机器学习系统或许能够为数据工作者寻找这些数据集。该产物属于准备阶段，但其属性会体现出其机器来源。相反，模型结果仪表板是一种存在于交流流程中的产物，既可以由人类精心整理，也可以自动生成 [41]。

自动化机器学习工件不仅仅是这些数据流程中各项任务的输入和输出。工件也可以是由数据科学流程创建的，或为数据科学流程创建的元数据或其他文档。先前的研究考察了机器学习或软件系统中的元数据，以及它们与溯源的关系（第 2.1节）。例如，组织流程会创建人类需求文档，这是一种由人类生成的制品，可以直接规定数据分析目标，并影响数据集或模型的选择。

*5.1.2*工件组和单个工件。我们现在根据数据科学流程描述工件组和单个工件的示例。有关制品属性层级的示例，请参见图 4 中对 " 需求文档 " 制品的分解。虽然这里以线性方式呈现这些流程，但实际上，它们可以按任意顺序发生。

准备流程包含两类工件组：目标和数据（图 4.2）。数据工作始于某个目标，它可以表现为分析目标、需求规范或任务 [30, 39, 48]。目标也可以转化为任务 [9, 44, 106] 和意图 [28, 85]，进一步明确具体的分析目标。这些目标对于确定用于分析的数据集，以及对初始数据及其模式表示进行的任何转换或扩充都是必要的 [19]。这些转换可能源于数据清理或数据整理操作 [45]，数据划分 [116], 或映射转换。我们还观察到，在准备阶段会引入其他数据集，以进一步对模型性能进行基准测试 [30, 119] 或评估其稳健性。数据扩充可以包括由人类提供的语义标注 [25]。我们观察到，准备阶段目前仍主要由单个人类或多个协同工作的人类开展。这些活动目前是数据工作中最耗时的部分 [14],，同时也是影响最深远的部分 [81]。作为数据准备的一部分，我们将探索性数据分析纳入其中，该分析会生成自动或由人类整理的摘要，包括描述性统计和可视化摘要 [107]。

分析流程在以往文献中得到最广泛的探讨，并涵盖了许多人认为是自动化机器学习核心功能的内容。我们将分析制品分为四组：与单个模型、单个AutoML 流水线配置、所有可能的管道配置所构成的搜索空间，以及计算相关的制品。第一组制品涉及模型，包括其任务（即分类、回归、聚类，或神经网络中各种更细致的任务）、特征编码 [11, 44, 111, 113], 生成 [101, 118],与选择，以及模型优化 [69, 94, 110] （其中包括模型架构，例如深度神经网络的架构 [42, 118]）和性能评估 [101]。

然而，模型只是AutoML流水线中的一个组件 [99, 111] 或原语 [38, 70] 。流水线本身由更广泛的可选配置搜索空间 [2, 26, 38, 80, 99, 103, 113, 119]确定。用于可视化自动化机器学习系统的工具越来越关注搜索空间和流水线配置 [70, 103]。这两个工件来源共同影响最终模型的选择，因为它们确定了模型可能采取的形式范围。这三个工件组——模型、流水线和搜索空间——共享相似的工件，包括初步配置、性能评估、优化摘要，以及对拟合（或搜索）计算的描述性摘要。

More recent AutoML systems place computation more prominently in the analysis stage. While these can include source code [11, 101] (including analysis notebooks), they also include system configurations and environments [11]. Recently, computational budgets [103] are used to calibrate model performance against computation time.

我们观察到，自动化机器学习系统会在合理范围内尽可能自动化分析，同时也提供人工干预的途径。自动化机器学习系统的复杂性日益增加，若不考察完整的工件谱系，就越来越难以追溯系统是如何选定某个模型的。例如，搜索可能的 AutoML 流水线配置空间的系统，既依赖初始配置，也依赖可用的原语集合。施加计算预算也会限制所探索的搜索空间范围。

部署流程 将最终模型应用到生产环境。我们识别出部署的两组工件：涉及 验证 和 监督。验证工件源于监控模型性能（部署前后） [100]。它们包括生成摘要统计、与现有基准测试的显式比较 [116, 118, 119], [13, 22, 90] 以及检测模型漂移或异常。这些工件对于捕获模型随

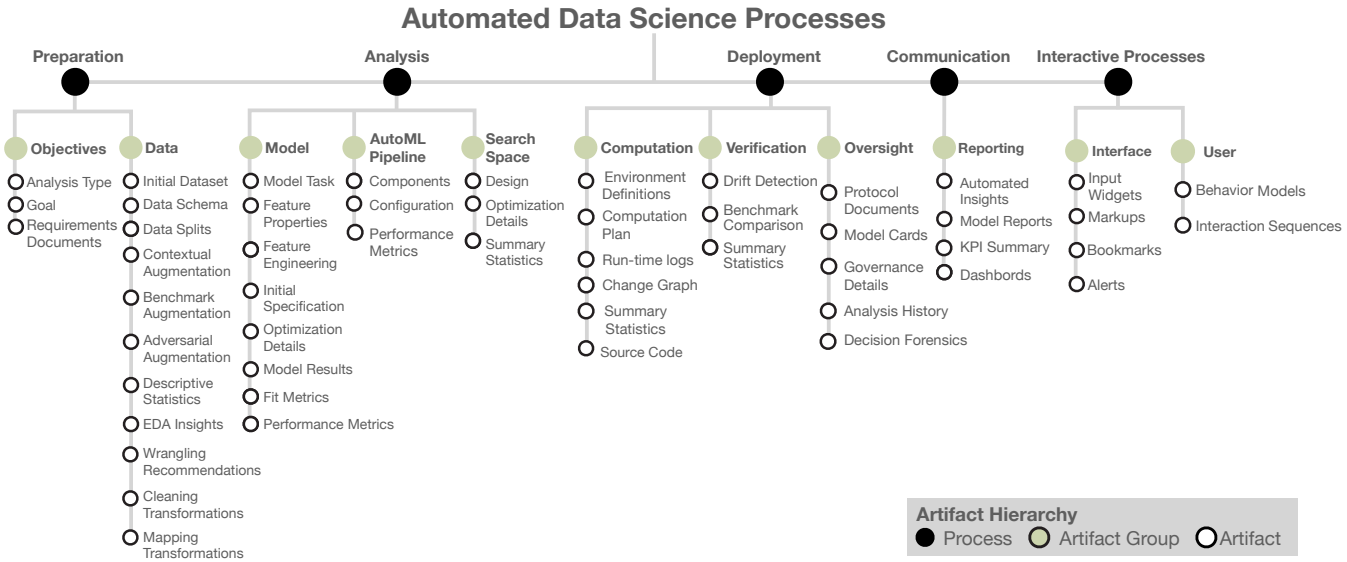

Figure 3: Artifacts elicited from AutoML toolkits, libraries, systems, and user studies. We summarized approximately 400 artifacts from these sources into 11 Artifact Groups and 52 artifacts. The properties of these artifacts are further delineated according to a taxonomy and a hierarchical set of dimensions, categories, and *characteristics*.

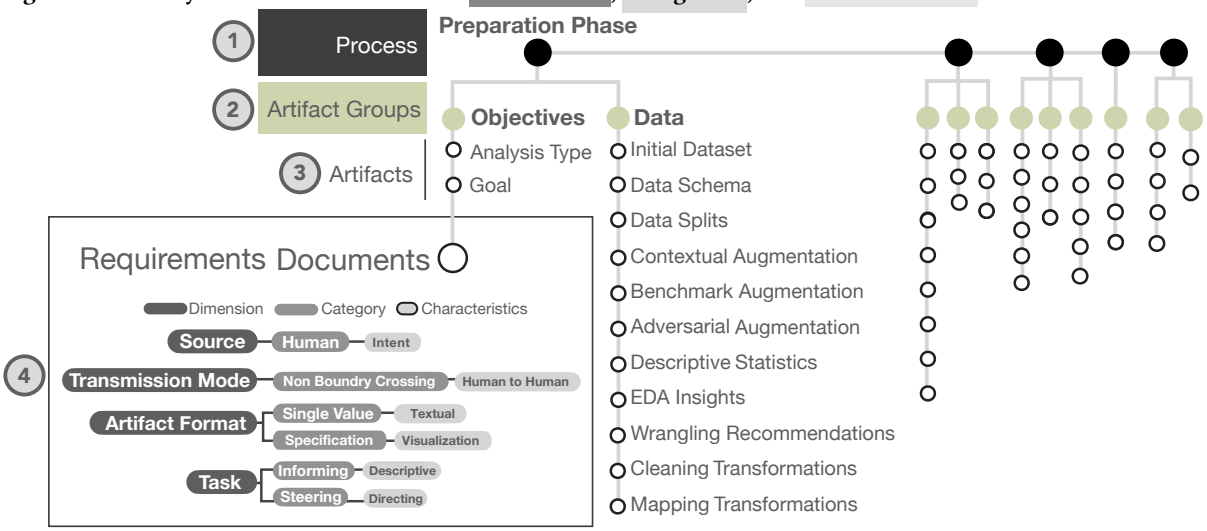

图 4：(1) 准备阶段 流程的信息层次分解。这里我们看到 (2) 工件组，以及 (3) 准备阶段的工件。每个由白色圆圈表示的制品都有其自己的维度、类别和特征。此示例显示了需求文档的制品属性 (4)。每个制品至少有一个来自四个维度的示例；来源、传输模式和任务。一些工件对于每个来源都有多个类别和特征。例如，需求文档对于格式和任务都有两个类别。

时间的变化很重要，并经常输入到监督工件中。这些监督工件包括描述模型特征的文档，例如模型卡 [61],决策取证报告 [100], 使用来源工件 [90], 以及治理模型使用的文档 [14, 100]。监督工件提供了知识共享的关键点，人类在此监控模型以确保其负责任地应用 [100]。此外，这些由分析师自动生成的工件提供了重要途径

for humans to intervene in automated work. For example, suppose a deployed model in production begins to exhibit poor performance on benchmark datasets. In that case, oversight artifacts can initiate a process where a human returns to the analysis and manually re-initiates the model fitting processes.

我们分类体系中的通信流程制品 主要是文档，既可以是静态的（例如报告），也可以是交互式的（例如



Figure 5: Artifact properties are a set of hierarchical taxonomic descriptors. The top level of this hierarchy is a dimension, followed by category, and finally characteristics.

仪表板），用于 报告 信息。交流不仅包括人与人之间的交流，自动化机器学习系统也必须与人类交流。我们再次有机会借鉴人与人通信，从而提升人机通信的有效性。通信产物包括洞察的自动生成摘要，或对模型决策的解释。建模解释由系统自动生成，对于实现透明度也日益重要 [62, 88, 90, 101]。

交互流程 相较于其他流程是一个例外。我们认为应将其单独处理，因为它们代表了无法自动化、但试图影响自动化流程的独特人类行为。其他阶段中的许多工件可以由人类或机器行为的某种组合生成。我们将交互流程分为图形用户界面和用户自身的工件。用户界面的元素包括已加入书签或已保存的洞见 [16, 107], 注释 [16, 25, 87]。人类还可以触发或修改自动化流程 [13]，贯穿数据科学流程。这些用户操作日益被捕获为行为图、交互日志或交互序列 [6, 12, 39, 91]，并可被可视化 [16, 91]，通过语义交互影响机器学习组件 [23, 29]。

5.2 Artifact Properties

The AutoML artifacts described in the previous section were determined by their properties. We used the initial set of 400 artifacts collected from the literature to derive a set of properties that allowed us to further group them into a smaller set.

完整的制品属性集如图 5 所示，但为避免过多重复，工件及其特征的详细分解见附录 A。虽然分类体系的初始目标是描述 工件，但我们也发现属性能够 比较 它们同样有用。例如，两条自动机器学习流水线可能都包含特征生成阶段，这将产生一个共同的制品：特征集。然而，在一条流水线中，特征生成可以自动完成，而在另一条中，这是人类的工作。在两条流水线中，随后的超参数调优可能都是自动完成的。我们努力使我们的分类体系能够描述自动化机器学习系统的广泛设计空间；包括已实现的和理论的。

在顶层，我们的分类体系包含四个维度，用于回答以下四个问题：**"**是什么生成了该制品？**"**（来源），**"**它是否跨越了人类与自动化机器学习流程之间的边界**"**（传输模式），**"**该制品呈现什么形态？**"**（工件格式），最后，**"**它的预期用途是什么？**"**（任务）。

来源制品表明它由谁或什么产生。我们确定了五种来源：人类、人员组织、数据、AutoML 流程和计算系统。前两种来源区分了单独行动和协作行动的人类，以及可能影响这些人的一组一般组织实践（即业务实践、法律或监管要求）。对初始数据集进行的计算、转换和其他推导也会产生新的工件。最后，AutoML 流程和支持该自动化的计算基础设施会产生互补但彼此独立的工件。例如，前者可能会生成模型损失的实时摘要，而后者则记录并返回代码故障，或记录计算预算已达到的情况。

传输模式属性描述工件是否跨越了人类与自动化机器学习来源之间的边界，以及跨越的方向。我们优先关注可能在人类与机器之间传递的工件（$h \rightarrow m$），以及反向传递的工件（$m \rightarrow h$）；方向性是通过阅读文献来源确定的。有些工件虽然在特定的自动化机器学习系统中不会跨越边界，但仍至关重要，因为它们为跨越边界的工件补充了背景信息。

工件格式 属性有助于比较不同的自动化机器学习管道。在构建分类体系的过程中，我们发现工件格式与自动化机器学习系统的设计选择密切相关。例如，面向机器学习专家终端用户的自动化机器学习系统，在展示信息时，其工件仅限于单一值、文本或张量。面向领域专家的系统则以可视化或交互式的方式呈现相同的数据。我们总结出四种格式：单一值、多个值、规范说明和报告。第 2 节讨论的可视化系统和仪表板被视为报告，具有静态或交互式特征。

该**任务** 描述了工件的可供性。我们提出了四类任务：告知、治理、共享和引导。告知类工件用于描述数据科学流水线的先前或当前状态，可以包括报告、汇总统计或仪表板（以及其他可能形式）。治理类工件专用于规范、审计和监控自动化与人类驱动的工作。共享类工件旨在在人类之间分发，而不仅仅是在分析与 AutoML系统之间共享。最后，引导类工件会介入数据科学流水线中的任意环节，以进行改变。这些工件源于人类或自动化流程的操作，例如对数据质量问题发出警报。

5.3 进一步扩展

The taxonomy itself can be further expanded over time, accommodating new artifacts that emerge as the capabilities of AutoML

系统扩展或包含特定系统实现的高度定制化品质。在我们开发分类体系时，我们不断反思其可扩展性，将其作为我们评估标准的一部分。具体而言，当我们合并了许多不同的先前分类法，这些分类法专门针对自动化机器学习和机器学习 [19, 93, 97],可视化分析的类型学 [9, 48], 以及其他分类系统 [44, 77, 87,88], 我们仔细审查了我们的分类体系的稳定性，以纳入这些变化。此外，我们的停止标准基于分类法结构的稳定性。我们依赖未来的工作来持续反思其可扩展性，因为当前的分类体系纳入了目前可用且相关的先前研究。

6 AUTOML TRACE

我们通过创建AutoMLTrace（一种交互式可视化草图）来实现我们的产物分类体系， [35]。与更复杂的交互式原型相比，可视化草图的保真度较低，但在促进研究人员与其合作者之间的共创活动方面发挥着重要作用， [82]。相比之下，Sanders等人。 [82] 将原型定义为在构思和执行方面更为成熟的形式；这与Buxton和Greenberg [35], 的观点一致，即原型对于共创和构思可能适得其反。本着这一理念，我们开发了 ˜AutoMLTrace，以探究将我们的分类体系应用于现有 AutoML系统的可视化分析的效用。尽管我们的目标最终是开发 AutoML Trace，以促进与合作者之间的对话（见第7节），但 AutoML Trace 与我们的分类体系相结合，也可以重新用于更广泛地分析人工智能/机器学习系统。我们尤其旨在通过捕捉工件并 delineate 其属性来突出人类要素，从而阐明自动化机器学习系统中的人类与机器学习/人工智能协作过程。本节介绍：（1）我们的分类体系如何使我们能够识别、分类、提取并可视化人类和机器派生的工件；（2）我们的交互式草图 AutoML Trace 的整体设计，包括其支持的数据和任务。

6.1 将我们的分类体系操作化

我们的AutoML 工件分类法捕获了数据工作端到端流程中从人类和机器派生的工件，该流程涵盖从准备到交流的各个环节。单个工件及其属性使我们能够适应不同程度的自动化，从人类驱动到完全自动化，以及介于两者之间的混合自动化模式 [14, 44, 72]。在混合自动化模式下，我们捕获工作的方向性——从人类到机器的流程（$h \rightarrow m$），反之亦然（$m \rightarrow h$）。通过加入时间信息，我们使用分类体系来派生工件创建的上下文和时间。通过持续捕获自动化数据工作流程中的工件，我们可以展示数据工作和人类与机器学习/人工智能协作过程随时间的演变。

6.1.1 AutoML Trace 中使用的工件AutoML Trace。我们在交互式草图中使用了从企业团队的 AutoML 流水线捕获的工件。直接使用他们的工件，不仅展示了如何借助实际运行的系统将我们的分类体系付诸实践，也促进了团队对其 AutoML 工具开发相关数据的有意义参与。将分类体系付诸实践的第一步，是利用它来识别和描述来自 AutoML 系统的工件。一些工件可以通过编程方式捕获，例如作为自动化机器学习系统的输入或来自不同 API 的输出。例如，人类可以通过交互式界面指定目标。或者，自动化机器学习流程可以初始化并遍历搜索空间，以寻找最优的模型参数组合。用户输入和搜索空间探索都可以从系统日志中捕获。其他工件则需要手动捕获。例如，说明系统需求的文档或传达结果的演示文稿，需要从现有的文档管理系统或其他整理工作中获取。在自动捕获或通过整理工作捕获这些内容时，分类体系的结构和工件的属性会提供其创建背景（例如，准备、分析、部署或交流阶段）。

我们的分类体系使我们能够识别这些工件的生成方式，并通过手动注释为其分配属性。例如，设计师和机器学习/数据科学工程师可以讨论工作流中的各种输入和输出，识别其可能属于的工件类型，并使用分类体系的受控词汇对其进行一致描述。AutoML Trace可以通过定义工件的默认模板，并以可视化方式指示哪些内容已被捕获或缺失，来支持这一流程。无论这些工件以何种方式被捕获，最终结果都是在人类与数据工作中的自动化流程之间交换的一组工件。尽管我们在 AutoMLTrace 中使用了团队自动化机器学习系统中的特定管道，但所捕获的工件具有我们为一系列自动化机器学习系统开发的分类体系中的属性。考虑到分类体系为工件提供的脚手架，AutoMLTrace 仍适用于其他管道。未来的工作将评估交互式草图与其他管道的交互效果，并进一步实现工件注释流程的自动化。

6.1.2 追踪制品的时间顺序、依赖关系和可变性。除了创建上下文之外，我们还可以收集制品的创建时间戳，以便检查其创建顺序和依赖关系。例如，特征生成制品可作为模型拟合的输入。我们还可以检查制品如何随时间变化。例如，假设初始特征集由自动化机器学习算法自动生成，一位检查制品的人类决定手动挑选并更新这些特征。此时，便存在两个版本的制品。通过制品的属性，可以识别出第一个版本是自动创建的，而后续版本则是人工干预的结果。

6.1.3 描述和比较人类-*ML/AI*协作分析。 人类与ML/AI系统之间的协作使得审计和比较分析变得困难。我们提出，通过我们的产物分类体系对分析进行注释，我们可以直接描述和比较不同的分析选择，并推断出一定程度的自动化，从完全自动化到无自动化以及介于两者之间的不同程度 [44, 50, 72]。

6.2 Data and Tasks

我们将单个工件及其集体元数据都用作 AutoML Trace 的输入数据集，以供其可视化。单个工件有不同的格式，这些格式会影响它们的捕获方式以及

它们向最终用户呈现的方式；我们在分类体系中将这些不同格式定义为工件属性的一部分（图 5.2）。分类体系以及时间戳和流水线结构等补充信息，共同定义了一组工件的元数据。为了围绕这些工件促进富有参与感的协作对话，我们定义了一组交互式可视化草图应支持的任务：

- **T1 Present a Contextual Overview of Artifacts:** The contextual overview ties the artifact creation with its specific data science phase (see Section 3). Whether an artifact was generated automatically or by a human was important – this consideration would become a key component of the AutoML Tracedesign. The dependencies of artifacts on each other were also an important contextual component.
- **T2 Locate an Artifact:** Enable end-users to filter out artifacts they are not interested in and to focus on a specific artifact, or group of artifacts, that are of interest to them.
- **T3 Summarize the Details of the Artifact:** Artifact details, like its properties and dependencies, should be progressively revealed to the end-user. Similarly, an artifact's taxonomic descriptors should reveal artifacts that share the same properties, not just those that a selected artifact depends on.
- **T4 Compare an Artifact over its History:** The end-user should be able to compare the states of an artifact over time and relative to its upstream and downstream dependencies.

These tasks align with those for information seeking that were defined by Shneiderman [86] (Overview, Zoom, Filter, Details on Demand, Relate, Histories, and Extracts), but described using a terminology of more recent task typology defined by Brehmer and Munzner [9].

6.3 AutoML Trace Interface

AutoML Trace 以工件集合及其元数据作为可视化的输入。它包含三个互补视图：来源视图（图 6）、依赖关系视图（图 6）和历史视图（图 7）。所有视图对工件采用相同的编码方式，以保持视觉语言的一致性。工件以圆圈表示，根据其来源（人类或机器）进行颜色编码，并按照数据科学阶段（准备、分析、部署和交流）排列。这些视图的灵感来自先前自动化机器学习系统和研究中的图与网络可视化方法（见第 2 节），不过我们确实考虑了其他设计方案（见补充材料 - [SM4]）。由于这只是一个交互式草图，我们没有对其与其他设计备选方案进行详尽比较。

Origin View: ***What artifacts are human versus machine-***生成？制品来源视图展示了从自动化机器学习系统分析中收集的制品，并说明这些制品是由人类生成的，还是由系统自动生成的 6。我们使用冲积图来展示制品来源之间的流动与权衡（**T1** （当前））。我们将人类和机器生成的制品作为此视图的焦点，以此展示交织在一起的协作流程。

Hovering triggers additional taxonomic details to be revealed on demand via an information card (**T3**). End-users can further hover 基于分类描述符和上下文数据，例如依赖关系和数据科学阶段**(T2)**。一旦选择了制品，最终用户还可以查看该制品的原始来源文件输出。**依赖视图:** 哪些工件相互依赖？ 依赖视图显示工件之间的关系(Fig. 6)。此视图的设计受数据级联插图的启发 [81]；实际上，此视图是通过工件呈现这些级联的直接响应。类似于来源视图，最终用户会看到概览 **(T1)**，信息通过悬停操作显示 **(T3)**。然而，在此视图中，选择制品会突出显示其依赖关系 **(T2)**。

版本历史视图： 一个制品中的更改如何影响其他制品中的更改？此视图用于深入查看制品历史，并了解一个制品中的更改如何影响依赖制品中的更改（图7）。用户可以在来源视图或依赖视图中选择某个制品，以查看该制品的历史。此视图使最终用户能够**T4** （比较）该制品本身随时间的变化以及其他制品的变化。在图7中，有四条水平线，分别对应分析的四个修订版本或迭代。新的制品或因更新而被修改的制品以圆形表示。未发生变化的制品以向下三角形表示。所选制品的依赖关系也会显示出来。与前两个视图一样，将鼠标悬停在制品上会显示其其他分类描述符。

7 USAGE SCENARIO

我们展示了一个使用场景：与一支企业软件开发团队合作，使用我们的分类体系和 AutoML Trace来探索和分析他们现有的自动化机器学习系统。我们的合作目标是分析其现有的自动化机器学习系统，以了解该系统的最终用户何时、对什么以及如何进行干预，或覆写自动化机器学习系统的决策。当时，他们尚无用于开展此类探究的机制。我们描述了AutoML Trace如何支持与该团队讨论期间的对话，以反思系统当前的能力，并围绕尚未满足的终端用户需求进行构思。这些讨论每次持续一小时，约三个月期间每两周举行一次。整个团队都参加了会议和讨论。在这些对话中，我们向合作方展示了对其自动化机器学习系统的分析结果，这些结果以制品的形式呈现，并且是我们使用分类体系收集和标注的。我们根据合作方的反馈和讨论，改进了制品的收集工作以及 AutoML Trace 的可视化设计。

7.1 Collaboration Context

现有系统概述。其自动化机器学习系统可以将数据工作从准备到部署的各个方面实现自动化（第 3节），包括呈现自动标记的洞察，以便探索数据、生成特征和进行自动化模型选择。图形用户界面（GUI）引导最终用户完成分析以及对结果的修订。最终用户可以通过GUI中的输入小部件和交互来修改分析，例如更改模型类型，从而介入其中。系统的某些方面还要求在启动自动化流程之前明确输入人类信息。例如，系统会以非技术性的方式呈现多重共线性，并要求

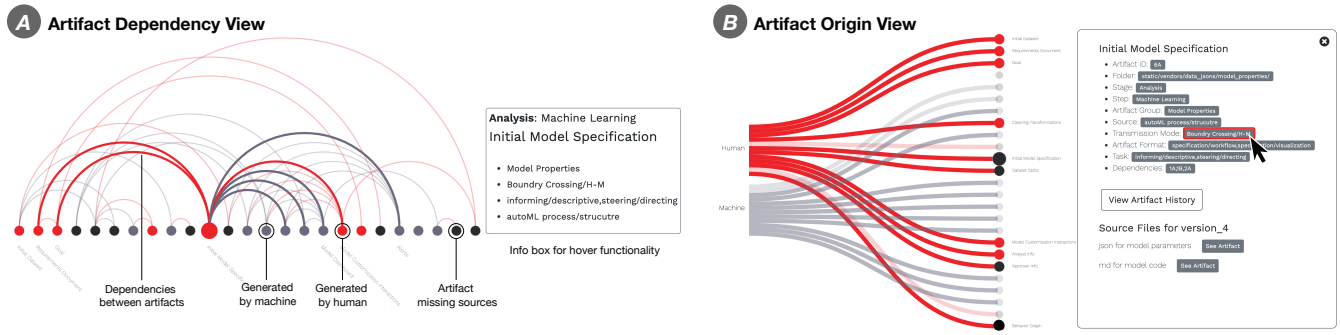

图 6：（A）构件依赖视图。此视图展示了在之前定义的数据工作阶段背景下相互依赖的工件级联。圆圈的颜色表示该工件是由机器生成还是由人类生成。弧线表示一个或多个工件之间的显式依赖关系。当选定某个工件后，该工件的依赖关系会在可视化中突出显示，从而更容易追踪某个工件发生变化时会影响哪些内容。在该示例中，选定了初始模型规范，同时显示其他工件依赖关系以及分类体系中的该工件特征。（B）工件来源视图展示哪些工件由人类生成，哪些由机器生成。此外，此视图还允许用户查找分类体系所定义的工件属性中的共性。再次选定初始模型规范后，用户可以将鼠标悬停在所选工件的参数上，以查看其他具有相同特征的工件。在该示例中，传输模式表明该工件跨越了从人类到机器的边界。将鼠标悬停其上会突出显示所有其他跨越从人类到机器这一边界的工件。

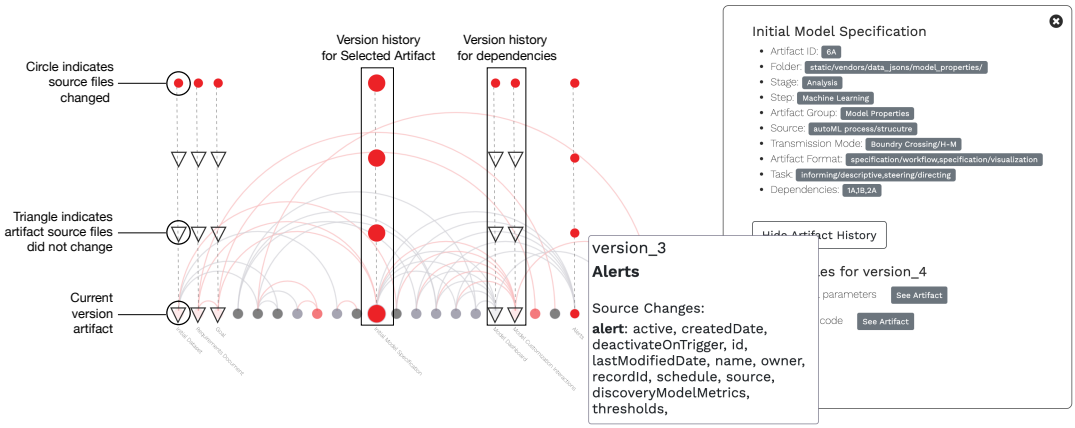

Figure 7: Artifact History View. This view shows the history of the selected artifact, differentiated by analysis versions. From the dependency view, the histories of the selected artifact dependencies are also shown. This example shows the history of the selected artifact "Initial Model Specification". The tool-tip shows details for the third version of the "Alerts" artifact, which is a dependency of "Initial Model Specification".

end-user confirm which features to remove from the analysis. The end-users could deploy a model to be used by others; automated processes would also monitor for concept drift and, if necessary, alert the end-user to trigger updates.

Team composition and collaboration goals. Our collaboration had two primary goals. First, for the team to reflect on their existing system and better understand what AutoML systems are capable of. Our taxonomy created an avenue for this reflection by providing a structured vocabulary to describe their system and compare it to others. The second goal was to examine what the additional traceability would add to their system. The project team consisted of software engineers, designers, user researchers, and a project manager. We also recruited one customer of their system for additional feedback. The team worked together to implement different components of AutoML work and implement the system.

7.2 Artifact Identification, Classification, and Extraction

We briefly describe how we analyzed our collaborator's existing system to develop a collection of artifacts visualized with `AutoML Trace`.

Generating Artifacts. Within the GUI environment of the AutoML system, we created an end-to-end data analysis. We began

以准备开始并以交流结束。在此流程中，我们返回到早期步骤并进行了修改。例如，我们初始没有应用自动数据清洗建议，但在后续迭代中这样做了。我们还让系统在第一次迭代中为模型选择特征，随后更改了它们。进行此分析有三个目标。首先，产生各种可能的工件，其次，记录工件之间的依赖关系，最后，观察工件如何响应用户交互而变化。结果是一组从同一分析派生的工件，这些工件随时间而变化。

收集制品。 我们使用团队开发的 API 来收集一组JSON 文件，以供分析。相比其他方法（即使用日志），我们采用了 API的输出，因为这些输出包含了完整的制品，使我们更容易根据分类体系对制品进行分类。我们还存储了对象的创建顺序，并能够确定工件之间的依赖关系（第 6.1.2节）。与许多现有的自动化机器学习系统一样，它们没有明确说明任何人类参与。我们不得不手动记录制品何时由人工干预生成或修改。在少数无法捕获分析全部方面的情况下，如果我们认为某个制品很重要，就会对该制品截图。例如，数据探索过程中自动生成的一些洞察包含无法从 API 中提取的可视化内容，因此我们改为截图。

工件分类我们使用产物分类体系对 API 调用或屏幕截图中的文件进行了注释。然而，在大多数情况下，单个文件包含多个工件。例如，针对初始数据集信息的 API 调用返回了这些信息，同时还返回了有关推荐的数据整理转换的信息。作者首先通过在简单的开发环境中手动检查这些 API，划分并标记其中的工件实例，从而识别工件。接下来，作者逐一检查每个工件，并根据我们的分类体系对其进行分类，同时根据分析需要修改其属性（即判断其是由人类生成的，还是自动生成的）。最后，我们检查并记录了工件之间的依赖关系。作者重复这两个步骤，直到就工件类型及其属性达成共识；我们还与合作者进行了沟通，以验证我们的工件是否准确。工件及其分类注释的最终列表见补充材料（[SM3]）；该列表构成了 AutoML Trace 实现的基础。

7.3 Collaboration and Question Elicitation

作为最后一步，我们通过视频会议平台向合作者进行了由专人引导的演示 [57] 。我们展示了自动化机器学习Trace 的功能和可供性，合作者也获得了提供反馈的机会。我们在讨论使用他们现有平台开展的分析，以及通过自动化机器学习 Trace 收集并可视化的工件之间反复切换。这是我们评估过程中的重要一步，因为真实分析中的所有工件都通过他们的API被捕获，这让合作者更加确信，可以将可追溯性添加到他们现有的系统中。团队当时

excited to view their artifacts and their system's capabilities in this way.

我们想收集我们的原型在这些反馈会议期间会引出的问题类型。互动是动态的，作者们都提出并回答了关于工件、它们的来源、依赖关系和随时间变化的问题。我们的合作者最欣赏的是能够根据我们的分类体系看到他们的系统布局。这种对他们系统的新视图使他们从以前未考虑过的角度审视他们工作的各个方面。我们将讨论总结为三个常见主题：查看和描述依赖关系、比较随时间的序列分析，以及比较他们的系统与其他系统的差异。

查看并描述依赖关系。他们此前无法可视化单个工件及不同类型工件之间的依赖关系。通过分析，他们尤其关注并兴奋地看到人类生成的流程与机器生成的流程是如何交织在一起的。这种来源视图与依赖视图的结合，使他们能够推断某个工件当前状态与其他操作之间潜在的因果关系。正如我们之前所指出的，许多自动化机器学习系统并未明确说明人类所发挥的作用；但借助 AutoML Trace，人类作用的影响和效果是不可否认的。团队看到了以这种方式将分析过程可视化的益处，并将其视为反思系统设计的一种方式。他们还看到了呈现此类关系对于支持治理分析管道的益处。例如，如果部署模型需要授权，他们认为 AutoML Trace 是审计现有分析的有用方式，以便提出部署建议或拒绝部署。

比较分析序列。我们的合作者还对使用 AutoML Trace 比较多位分析师随时间开展的分析感兴趣。他们特别希望获得由不同参与者生成的多个分析序列，并对其进行比较。这种异步人类协作场景与个体人机协作相结合，表明我们的分类体系有望用于解决更复杂的问题。虽然我们可以对制品进行版本管理，从而支持详细比较，但当前的设计并未针对多人协作进行充分优化——不过，这同样为未来的研究指明了富有成效的方向。

一位合作者尤其关注了解人类何时采纳并应用机器建议，以及何时忽略这些建议。一个特定的、令人感兴趣的制品序列始于初始数据集，随后是来自机器来源的数据整理转换建议。接着，对用户操作的记录将表明结束用户是否应用了任何数据整理转换建议。最后，序列以对初始数据集的任何潜在更新告终。这是另一个有趣的使用场景，它不仅使我们能够了解系统的自动化程度，还可能定义每位用户的自动化特征。此外，还可以评估某些工件是否比其他工件更常被修改。总体而言，这些特征可用于识别存在问题的特征（例如，如果机器的结果不断被覆盖）或分析行为模式。

将他们的系统与其他系统进行比较。我们开发的分类体系是对涵盖人类流程和机器流程的各种系统的汇总。我们的分类体系为比较这些系统，以及反思相对于其他系统可能缺少哪些工件，提供了标准词汇。例如，自动化机器学习技术的近期发展加入了计算预算，使这些自动化流程能够在合理的时间范围和预算约束内完成。然而，并非所有AutoML系统都具备此类特征。我们的分类体系促使我们讨论了合作方系统的设计影响。他们首先感到欣慰，因为他们看到现有系统中包含了与其他系统重叠的要素；但同时也看到了当前实现中缺失的其他有趣方面。

7.4 Summary

在研究的第二阶段，我们通过与一个正在开发复杂自动化机器学习系统的团队开展合作，探究了分类体系的实用性和生态效度。AutoML Trace草图表明，分类体系是一个有用的边界对象，可用于与负责设计人类-机器学习/人工智能协作系统的软件专家和ML/AI专家团队开展互动。它还表明，可追溯性在人机协作和人类-人类协作中都具有有价值的应用。尽管我们的方法并未解决演化型和自适应系统的所有设计挑战 [112], 但它确实朝着这一目标迈出了初步步伐。

8 DISCUSSION

人类与ML/AI系统的协作将随着自动化机器学习技术日益集成到数据工作中而变得更加普遍。这些系统降低了数据工作的门槛，并帮助数据科学家更有效地分诊他们的工作 [101]。然而，由于现有系统仍需要人类监督和干预，这些半自动系统需要可观察和可理解。人机交互社区的近期工作识别了审查ML/AI系统的挑战，这些挑战源于人类-ML/AI协作工作的复杂性，并强调了在此空间中进行讨论需要一种通用语言 [112]。

我们的工作从多个方面解决了人-ML/AI协作中的局限性。首先，我们通过创建 AutoML 工件分类法，形式化了一种通用语言，该语言兼顾了这些系统的人类和机器方面。其次，我们在交互式草图 AutoML Trace 中将此分类体系付诸实践，刻画这些工件以促进可追溯的工作流。第三，我们刻画了可追溯性，以审视这一高度复杂且异构的流程。虽然先前研究通过溯源工具（即 [56, 70, 103]）捕获了可追溯性的某些方面，但它们未能区分人类过程和自动化过程，并且经常完全忽略人类过程。人类-人类协作和知识共享方面的研究强调了工件在捕获 [47, 49] 和追踪 [27, 59, 95] 复杂协作流程中的重要性。通过考虑可追溯性，我们为工件提供了不同的视角。我们认为，可追溯性鼓励更广泛地考虑工件的谱系及其生成和使用的上下文因素。此外，通过工件，我们的研究承认并提升了人类与 ML/AI系统之间的社会技术关系。

除了溯源之外，当代研究越来越关注透明度、可解释性和可说明性对 ML/AI系统的重要性 [3, 5, 46, 62, 90]。然而，这项先前工作聚焦于模型本身，而错过了数据级联 [81] 中的影响因素。我们的研究扩大了范围，通过工件和分类法捕获数据科学工作端到端流程中的工件。我们证明分类法可以稳健地创建，并可作为设计人-ML/AI协作系统的边界对象。我们的方法表明，可以同时拥有 "流程的透明度和产品的透明度；前者指研究和创新的人类过程的透明度，后者指如此开发的 *[...] AI* 系统的透明度。" [105]。

最后，我们的研究承认并描述了在数据工作中开发用于人类与机器学习/人工智能协作的可视化交互系统所面临的困难。设计研究和其他应用型研究主要关注最终用户，但复杂系统仍然需要 ML/AI专家的参与。研究人员与并非最终用户的专家之间的协作仍然复杂，并且可能需要借助可视化工具作为中介来促进对话 [112]。由于缺乏可靠的对话支架，我们承担了创建这些支架这一雄心勃勃的任务。开发 ***AutoML*** 工件分类法和 `AutoML Trace`创建了边界对象，我们利用这些对象来应对这些挑战。我们描述自身流程的意图，是为面临类似挑战的其他研究人员提供可能的途径。

8.1 Implications of Our Findings and Future Work

*8.1.1*关于以人为中心的自动化机器学习系统的设计与评估。我们的产物分类体系可用于反思现有系统并构思新系统。现有的人-ML/AI交互指南的局限之一在于，它们针对的是系统的初始构思，而当系统已经存在时，效果就不那么理想了。在我们的案例研究中，我们观察到，产物分类体系具有回顾性和前瞻性地反思设计的潜力。这种潜力对于识别和修改无效方法至关重要。我们的分类体系有助于研究人员和实践者构思新系统，并推测机器学习/人工智能系统能够做什么 [112]，同时促进对人类作用的反思。

8.1.2 关于数据科学协作。不同类型的数据工作者参与数据工作 [14, 100, 102]。我们还需要进一步研究不同的数据科学角色 [15]，（从 ML 工程师到技术分析师）会如何使用这一分类体系。先前研究表明，人们会彼此分担数据工作的不同方面 [15, 114]。捕获并追踪工件有助于数据工作团队了解完成了哪些工作，以及由谁（或什么）完成。此外，围绕工件展开讨论，并借助 AutoML Trace 等工具将其可视化，可以帮助数据工作团队理解并批判分析及其结果 [65]。最后，尽管已有一些研究探索数据工作者与自动化程度之间的关系（即 Wang 等人 [101]），但尚未有人探讨人际协作与人机协作之间的复杂关系。我们的分类体系或许有助于将这些先前研究拓展至更混合式的数据工作流程。

8.1.3 关于数据可视化与视觉分析工具。视觉分析工具利用机器学习的进展，创新视觉分析（VA）系统的可供性。通过捕获更详尽的工件目录来加深我们对 VA 系统的理解，将使“系统用户能够更投入地进行视觉数据探索，而不是主动训练模型和系统，”[24]。受 Yang 等人 [112], 的启发，我们希望拓展对 ML/AI 管道中可捕获并呈现内容的认识，使其不局限于模型或分析阶段。我们的分类体系强调将更广泛的人类要素纳入捕获和呈现的内容，是迈向更普遍视角的一步：在不依赖单一系统或管道的情况下，捕获整个工作流程中人工生成的工件。我们希望其他人继续在 VA 系统领域应用并拓展这一成果。可视化和人机交互领域的研究者可以从两个方面在我们的研究基础上继续推进。第一，拓展 VA 系统针对 ML/AI 可视化内容的范围。我们的分类体系提出了对 AutoML 流水线更全面的视角，而当前研究（第 2.2 节）尚未考虑这一点。研究者可以运用我们的分类体系分析并可视化其他 ML/AI 系统，包括但不限于 AutoML 系统，还可以进一步扩展分类体系，为扩充 ML/A 工件目录作出贡献。尽管我们的 AutoML Trace 交互式草图提出了一种可能的可视化方案，但我们认为，在可视化设计领域仍有许多值得探索的机会。第二，拓展人-ML/AI交互范式。数据可视化工具是人类与机器学习系统协同工作的媒介 [88, 90]。虽然与这些系统的交互可用于干预 ML 模型 [10, 23, 29], ，未来的工作可以将这一潜力拓展至 AutoML 管道中的其他类型的原语和要素 [38, 70]。

8.2 Limitations

数据工作涉及各种各样的数据工作者。我们还需要进一步了解不同的数据科学角色 [15], 从 ML 工程师到技术分析师会如何使用这一分类体系。借用统计模型领域的一句格言，“所有分类体系都有缺陷，但有些很有用。”与此前的分类体系一样，我们力求使这一分类体系对 HCI、可视化领域以及机器学习研究人员和从业者有所助益。为实现这一目标，我们遵循了 Nickerson 等人 [67] 提出的严谨分类法开发流程，并结合了 Prat 等人 [73]的扩展。我们认真记录了分类法开发过程，并将研究流程中的工件作为补充材料公开，以便他人评议或扩展我们的工作。虽然这种方法涉及的工作更多，但它也为分类法开发中难以审视和复现的临时拼凑方法提供了一种重要的替代方案。更广泛地说，我们认为，研究流程使人们更加关注数据科学工作中由自动化和人类劳动所产生的工件。我们工作的另一项局限是，使用场景未涵盖最终用户，也就是实际开展分析的人。这样做有两个原因。首先，我们需要一个基准来支撑类似 AutoML Trace 这样的系统的开发。缺少这一基准时，我们就需要创建一个，因此本分类体系是我们的主要贡献。第二个原因是，就目前的贡献而言，开发团队是更适合进行初步评估的群体。

在第7节中，我们指出了若干拓展工作的有效方向，以推动终端用户评估的开展，其中包括更广泛地探索视觉设计空间，以及研究异步多人人类协作。

最后，我们反思了我们的分类体系开发方法。尽管我们的分类体系是通过广泛的文献综述开发的，但我们主要通过与单个团队的合作来评估其效用。可视化研究聚焦于一个核心协作者群体并不罕见，但仍需进一步工作来评估其普适性。我们乐观地认为，利用该分类体系比较从自动化机器学习到可视化分析的不同系统和管道，具有巨大潜力。我们鼓励社区根据自身数据工作的需要，采用合适的媒介来使用这一分类体系。在此之外，我们也期待看到，随着机器学习/人工智能技术的发展，这一分类体系将如何不断扩展。

9 CONCLUSION

自动化机器学习技术日益无处不在，使更广泛的利益相关者能够开展数据工作，但也可能使追踪完成了哪些工作以及由谁（或什么）完成变得更加困难。应对这些挑战的尝试往往受到系统本身复杂性的阻碍，也受到缺乏可用于与创建这些系统的机器学习与软件开发者开展互动的支撑机制的限制。我们提出了一种两阶段方法，用于阐明人类与自动化机器学习系统之间为开展数据工作而形成的协作关系。第一阶段构建了一个产物分类体系，可用于识别、分类和描述来自社会与技术来源的工件。第二阶段是与一个企业开发团队共同开展的使用场景，展示了我们的分类体系作为边界对象的实用性。该使用场景还进一步具体体现了在设计和开发协作性人类—机器学习/人工智能系统的流程中追踪工件的价值。我们的研究发现与贡献对于用于促进数据工作自动化的自动化机器学习系统的设计与评估具有重要意义。

Availability of Supplementary Materials. Notes taken during the development of our taxonomy and the `AutoML Trace` prototype are available online[2].

- `[SM1]` **AutoML Artifact Taxonomy Development :** A spreadsheet of each artifact and their classification across each iteration of the taxonomy
- `[SM1-F]` **AutoML Artifact Taxonomy Final :** A spreadsheet of the final taxonomy, with artifacts, properties, and their classification. A summary is also shown in
- `[SM2]` **Taxonomy Iterations Details :** A summary of our decision process for adjusting the taxonomic dimensions, categories, and objects. It has a detailed annotation of each iteration, what changed, and how we assess our end conditions for taxonomy development.
- `[SM3]` **Design Specification :** Design specifications for AutoML Trace prototype that provide additional context for the content in Section 6.
- `[SM4]` **设计备选方案：** 与合作者共同考虑并讨论过的设计备选方案。

[2]https://osf.io/3nmyj/?view_only=19962103d58b45d289b5c83421f48b36

参考文献

[1] Oshin Agarwal、Heming Ge、Siamak Shakeri 和 Rami Al-Rfou。2021。基于知识图谱的合成语料生成，用于知识增强型语言模型预训练。https://arxiv.org/abs/ 2010 12688 。Stefano Alletto、Shenyang [2] Huang、Vincent Francois-Lavet、Yohei Nakata 和 Guillaume Rabusseau。2020。RandomNet：迈向多模态学习的全自动神经网络架构设计。arXiv:2003.01181 https://arxiv.org/abs/2003.01181[3] Saleema Amershi、Dan Weld、Mihaela Vorvoreanu、Adam Fourney、Besmira Nushi、Penny Collisson、Jina Suh、Shamsi Iqbal、Paul N. Bennett、Kori Inkpen、Jaime Teevan、Ruth Kikin-Gil 和 Eric Horvitz。2019。人机交互指南。发表于*CHI'* *19* 1 13 10 1145 3290605 3300233 会议论文集。–。https://doi.org/。/。Gagan Bansal、Besmira Nushi、Ece Kamar、Daniel S. [4] Weld、Walter S. Lasecki 和 Eric Horvitz。2019。人机团队的最新进展：理解并应对性能与兼容性权衡。*AAAI'19* 会议论文集33, 01（2019年7月），2429– 2437 10 1609 33 01 33012429 。https://doi.org/。/aaai.v i。Alejandro [5] Barredo Arrieta、Natalia Díaz-Rodríguez、Javier Del Ser、Adrien Bennetot、Siham Tabik、Alberto Barbado、Salvador Garcia、Sergio Gil-Lopez、Daniel Molina、Richard Benjamins、Raja Chatila 和 Francisco Herrera。2020。可解释人工智能（XAI）：迈向负责任人工智能的概念、分类体系、机遇与挑战。*Information Fusion* 58（2020），82–115。 https://doi.org// 10 1016 2019 12 012 。/j.inffus.。。Leilani Battle 和 Jeffrey Heer。 [6] 2019 。探索性可视分析的特征：文献综述及 Tableau 中分析溯源的评估。*Co mputer Graphics Forum* 38, 3 2019 145 159 10 1111 13678 ( )，–。 [7] https://doi.org/。/cgf。Fabian Beck、Michael Burch、Stephan Diehl 和 Daniel Weiskopf。 2017 。动态图可视化的分类体系与综述。*Computer Graphics Forum* 36, 1（2017），133–159。 https://doi.org/10.1111/cgf.12791[8] Nadia Boukhelifa、Anastasia Bezerianos、Remco Chang、Christopher Collins、Steven Drucker、Alexander Endert、Jessica Hullman、Chris North 和 Michael Sedlmair。2020。交互式可视机器学习系统评估中的挑战。*IEEE Computer Graphics and Applications* 40, 6（2020），88–96。[9] Matthew Brehmer 和 Tamara Munzner。2013。抽象可视化任务的多层级分类体系。*IEEE Transactions on Visualization and Computer Graphics* 19, 12 2013 2376 2385 10 1109 2013 124 ( )，–。 [10] https://doi.org/。/TVCG.。Eli T. Brown、Jingjing Lie、Carla E. Brodely 和 Remco Chang。 2012 。Dis-function：交互式学习距离函数。发表于*2012 IEEE* 可视分析科学与技术会议（*VAST*）。 83–92。https://doi.org/10.1109/VAST.2012.6400486[11] José Pablo Cambronero。2021。挖掘软件工件以用于自动化机器学习。博士论文。MIT-CSAIL。 https://www.josecambronero.com/pdf/JCambronero-PhD-EECS-June2021.pdf[12] Dylan Cashman、Shah Rukh Humayoun、Florian Heimerl、Kendall Park、Subhajit Das、John Thompson、Bahador Saket、Abigail Mosca、John Stasko、Alex Endert、Michael Gleicher 和 Remco Chang。2019。用于探索性模型分析的基于用户的可视分析工作流。 *Computer Graphics Forum* 38, 3（ 2019 185 199 10 1111 13681 ），–。 [13] 2021 。https://doi.org/。/cgf。Bilge Celik 和 Joaquin Vanschoren。 。演化数据的自动化机器学习适应策略。*IEEE Transactions on Pattern Analysis and Machine Intelligence*（ 2021 1 1 10 1109 2021 3062900 ），–。 [14] https://doi.org/。/TPAMI.。Anamaria Crisan 和 Brittany Fiore-Gartland。 2021 。一波三折：企业对 AutoML 的使用及人在回路中的作用。发表于*CHI'21* 会议论文集601, 15。文章页码。[15] Anamaria Crisan、Brittany Fiore-Gartland 和 Melanie Tory。2021。传递数据接力棒：对数据科学工作及从业者的回顾性分析。*IEEE Transactions on Visualization and Computer Graphics* 27, 2（2021），1860–1870。https://doi.org/ 10 1109 2020 3030340 。/TVC G.。Zach Cutler、Kiran Gadhave 和 Alexander Lex。2020。Trrack：用于基于 Web 的可视化溯源追踪的库。发表于*2020 IEEE* 可视化会议（*VIS*）。 [16] 2020 。116 120 10 1109 47514 2020 00030 –。https://doi.org/。/ [17] VIS。 2013 。。Tim Davies 和 Mark Frank。 。“不存在所谓的原始数据”：探索政府数据集的社会技术生命历程。发表于*WebSci '13* 会议论文集。 75 78 10 1145 2464464 2464472 –。 [18] https://doi.org/。/。Tijl De Bie、Luc De Raedt、José Hernández-Orallo、Holger H. Hoos、Padhraic Smyth 和 Christopher K. I. Williams。2022年。数据科学自动化。*Commun. ACM* 65, 3（2022年2月），76–87。https://doi.org/10.1145/3495256[19] Dominik Dellermann、Adrian Calma、Nikolaus Lipusch、Thorsten Weber、Sascha Weigel 和 Philipp Ebel。2021。人机协作的未来：混合智能系统设计知识的分类体系。arXiv:2105.03354 https://arxiv.org/abs/ 2105 03354 。Veronika Domova 和 Katerina [20] Vrotsou。2022。可视分析中自动化类型与级别模型：调查、分类体系与示例。*IEEE Transactions on Visualization and Computer Graphics*（2022）。[21] Jaimie Drozdal、Justin Weisz、Dakuo Wang、Gaurav Dass、Bingsheng Yao、Changruo Zhao、Michael Muller、Lin Ju 和 Hui Su。2020。对 AutoML 的信任：探索建立对自动化机器学习系统信任所需的信息。发表于*IUI' 20* 会议论文集。297–307。https://doi.org/10.1145/3377325.3377501

[22] Radwa Elshawi、Mohamed Maher 和 Sherif Sakr。2019。自动化机器学习：现状与开放性挑战。arXiv:1906.02287 https: //arxiv.org/abs/ 1906 02287 。 [23] Alex Endert、Patrick Fiaux和 Chris North。2012。用于可视文本分析的语义交互。发表于*CHI'12*论文集。473–482。https://doi.org/10.1145/2207676。2207741[24] Alex Endert、William Ribarsky、Cagatay Turkay、BL William Wong、Ian Nabney、IDíaz Blanco 和 Fabrice Rossi。2017。将机器学习融入可视分析的最新进展。发表于 *ComputerGraphics Forum*，第36卷。Wiley Online Library，458–486。[25] Suilan Estevez-Velarde、Yoan Gutiérrez、Andrés Montoyo和Yudivián Almeida- Cruz。2019。基于语法演化的自动化机器学习策略：从文本中进行知识发现的案例研究。发表于 *ACL'19*论文集。4356–4365。https://doi.org/10.18653/v1/P19-1428[26] Matthias Feurer、Katharina Eggensperger、Stefan Falkner、Marius Lindauer 和 FrankHutter。2020。Auto-Sklearn 2.0：新一代系统。arXiv:2007.04074 https://arxiv.org/abs/ 2007 04074 。Gerhard .Fischer 和 Jonathan. Otswald。 [27] 2001。知识管理：问题、承诺、现实与挑战。*IEEE*智能系统 16，1（2001），60–72。https://doi.org/10.1109/5254.912386[28] Kiran Gadhave、Jochen Görtler、Zach Cutler、Carolina Nobre、Oliver Deussen、Miriah Meyer、 JeffPhillips和 Alexander Lex。2020。在散点图中进行刷选时捕捉用户意图。 https://doi.org/10.31219/osf.io/mq2rk[29] Sebastian Gehrmann、Hendrik Strobelt、Robert Krüger、HanspeterPfister和AlexanderM.Rush。2020。通过协作式语义推理与深度学习模型进行可视交互。*IEEE Transactions on Visualizationand Computer Graphics* 26, 1 2020 884 894 10 1109 2019 2934595 ( )，– 。https://doi.org/ 。/TVCG. 。Pieter Gijsbers 、 Erin LeDell 、 Janek Thomas 、 [30] Sébastien Poirier、Bernd Bis- chl 和 JoaquinVanschoren。2019。一个开放源码的自动化机器学习基准测试。arXiv: 1907 00909 1907 00909 。 https://arxiv.org/abs/ 。LisaGitelman。 [31] 2013 。“原始数据”是一个矛盾修辞。MIT Press，剑桥，美国。[32] Michael Gleicher。2018。可视化比较的考量。*IEEE Transactions on Visualization and Computer Graphics* 24，1（2018），413–423。https://doi.org/ 10 1109 2017 2744199 。/TVCG. 。Michael Gleicher、Danielle [33] Albers、Rick Walker、Ilir Jusufi、CharlesD. Hansen和Jonathan C. Roberts。2011。面向信息可视化的视觉比较。信息可视化 10，4（2011年10月），289–309。 https://doi.org/10.1177/ 1473871611416549[34] Daniel Golovin、Benjamin Solnik、Subhodeep Moitra、Greg Kochanski、John Karro和 D. Sculley。2017。Google Vizier：一种黑盒优化服务。发表于*KDD'* *17* 1487 1495 10 1145 3097983 3098043 。– 。 [35] https://doi.org/ 。/ 。SaulGreenberg和William Buxton。 2008 。可用性评估被认为是有害的（有时）。*CHI'08*论文集，111–120。https://doi.org/10.1145/ 1357054.1357074[36] Xin He、KaiyongZhao和 Xiaowen Chu。2021。自动化机器学习：现状综述。基于知识的系统 212（2021年1月），106622。https://doi. org/ 10 1016 2020 106622 。/ j.knosys.。Jeffrey Heer。 [37] 2019 。。代理加自动化：将人工智能设计进交互式系统。美国国家科学院院刊 116，6（ 2019 1844 1850 10 1073 1807184115 ），– 。 https://doi.org/ 。/pnas。Yuval Heffetz、Roman Vainshtein、Gilad Katz和 [38] LiorRokach。2020。DeepLine：使用深度强化学习和分层动作筛选生成管道的自动化机器学习工具。发表于 *KDD '20*论文集。2103–2113。https://doi.org/10. 1145/3394486.3403261[39] Sungsoo Ray Hong、Sonia Castelo、Vito D'Orazio、Christopher Benthune、Aecio Santos、Scott Langevin、David Jonker、Enrico Bertini 和 Juliana Freire。2020。探索使用自动化机器学习系统的模型构建流程评估。 arXiv: 2009 00449 2009 00449 。https://arxiv.org/abs/ 。 [40] Sungsoo Ray Hong、Jessica Hullman 和 Enrico Bertini。2020。模型可解释性中的人为因素：行业实践、挑战与需求。*CSCW'20*论文集。 4，CSCW1，文章068（2020年5月），26页。https://doi.org/10.1145/ 3392878[41] KevinHu、Michiel A. Bakker、Stephen Li、Tim Kraska和César Hidalgo。2019。VizML：一种用于可视化推荐的机器学习方法。发表于*CHI'19*论文集。Association forComputing Machinery，纽约，纽约州，美国，1– 12。https://doi.org/ 10 1145 3290605 3300358 。/ 。HaifengJin、 [42] Qingquan Song和Xia Hu。2019。Auto-Keras：一种高效的神经架构搜索系统。发表于 *KDD '19*论文集 （美国阿拉斯加州安克雷奇）。Association for Computing Machinery，纽约，纽约州，美国，1946–1956。https://doi.org/10. 1145/3292500.3330648[43] Bonnie E John、Len Bass、RickKazman 和 Eugene Chen。2004。识别HCI、软件工程与设计之间的差距，以及用于弥合这些差距的边界对象。发表于 *CHI'04*计算系统中的人因扩展摘要。1723–1724。[44] Shubhra Kanti Karmaker、Md. Mahadi Hassan、Micah J. Smith、Lei Xu、ChengXi- ang Zhai和 Kalyan Veeramachaneni。2021。自动化机器学习至今及未来：挑- l

[45] Stephen Kasica、Charles Berret 和 Tamara Munzner。2021。《表格残片：基于计算新闻学制品研究的多表数据整理实用框架》。IEEE Transactions on Visualization and Computer Graphics 27, 2 (2021), 957–966。https://doi.org/10.1109/TVCG.2020.3030462[46] Harmanpreet Kaur、Harsha Nori、Samuel Jenkins、Rich Caruana、Hanna Wallach 和 Jennifer Wortman Vaughan。2020。《解读可解释性：了解数据科学家如何使用机器学习可解释性工具》。1–14。https://doi.org/h 10 1145 3313831 3376219 . / . Kristian Kreiner。. 默会知识管理：制品的作用。[47] 《知识管理杂志》6, 2 (2002年1月1日), 112–123。https://doi.org/ 10.1108/13673270210424648[48] Heidi Lam、Melanie Tory 和 Tamara Munzner。2018。《通过设计研究分析报告连接目标与任务》。IEEE Transactions on Visualization and Computer Graphics 24, 1 (2018), 435–445。https://doi.org/10.1109/TVCG. 2017.2744319[49] Charlotte P. Lee。2007。《边界协商制品：打破边界对象的惯例，拥抱协作工作中的混乱》。Proc. CSCW'07 16, 3 01 2007 307 339 10 1007 10606 007 9044 (Jun), – 。https://doi.org/ 。 /s - - - [50] D. Lee、Stephen Macke、Doris Xin、Angela Lee、Silu Huang 和 Aditya G. Parameswaran。2019。《以人为中心的自动化机器学习视角：里程碑与未来发展方向》。IEEE Data Eng. Bull. 42, 2 (2019), 59–70。http://sites. computer.org/debull/A19june/p59.pdf [51] Doris Jung-Lin Lee、Vidya Setlur、Melanie Tory、Karrie G Karahalios 和 Aditya Parameswaran。2021。《解构可视化推荐中的分类：分类体系与比较研究》。IEEE Transactions on Visualization and Computer Graphics (2021)。[52] Doris Jung-Lin Lee、Vidya Setlur、Melanie Tory、Karrie G. Karahalios 和 Aditya Parameswaran。2021。《解构可视化推荐中的分类：分类体系与比较研究》。IEEE Transactions on Visualization and Computer Graphics (2021), 1–1。https://doi.org/10.1109/TVCG.2021. 3085751[53] Doris Jung-Lin Lee、Dixin Tang、Kunal Agarwal、Thyne Boonmark、Caitlyn Chen、Jake Kang、Ujjaini Mukhopadhyay、Jerry Song、Micah Yong、Marti A. Hearst 和 Aditya G. Parameswaran。2021。《Lux：面向探索式数据科学的常驻可视化推荐》。https://arxiv.org/abs/ 2105 00121 . Shixia Liu、Xiting Wang、[54] Mengchen Liu 和 Jun Zhu。. 面向更优机器学习模型分析：可视分析视角。《可视信息学》 1, 1 2017 48 56 10 1016 2017 01 006 ( ), – 。[55] https://doi.org/ . /j.visinf. . . Yang Liu、Tim Althoff 和 Jeffrey Heer。. 探索的路径、遗漏的路径、被遮蔽的路径：端到端数据分析中的决策点与选择性报告。载于 Proc. CHI' *20* 1 14 10 1145 3313831 3376533 . – 。https://doi.org/ . / . [56] Yang Liu、Alex Kale、Tim Althoff 和 Jeffrey Heer。.《Boba：创作并可视化多元分析》。IEEE Transactions on Visualization and Computer Graphics 27, 2 (2021年2月), 1753–1763。https://doi.org/10.1109/tvcg. 2020.3028985[57] David Lloyd 和 Jason Dykes。2011。《地理可视化设计中的以人为中心方法：通过长期案例研究考察多种方法》。IEEE Transactions on Visualization and Computer Graphics 17, 12 (2011), 2498 2507 10 1109 2011 209 – 。https://doi.org/ . /TVCG. . Octavio [58] Loyola-González。. 黑盒与白盒：从实践角度理解其优势与弱点。IEEE Access 7 (2019 154096 154113 10 1109 2019 2949286), – 。https://doi.org/ . / [59] ACCESS. . Stefania Mariano 和 Yukika Awazu。. 知识管理研究中的制品：系统性文献综述与未来研究方向。《知识管理杂志》20, 6 (2016年1月1日), 1333–1352。https://doi.org/10. 1108 05 2016 0199 /JKM- - - Swati Mishra 和 Jeffrey M Rzeszotarski。2021。《为机器[60] 学习非专业人士设计交互式迁移学习工具》。载于 Proc. CHI'21。文章 364，共 15 页。https://doi.org/ 10 1145 3411764 3445096 . / . Margaret Mitchell、Simone Wu、Andrew Zaldivar、Parker Barnes、Lucy Vasserman、Ben Hutchinson、Elena Spitzer、Inioluwa Deborah Raji 和 Timnit Gebru。2019。《用于模型报告的模型卡》。载于 Proc. FAccT'19。https: //doi.org/ 10 1145 3287560 3287596 . / . Brent Mittelstadt、[61] Chris Russell 和 Sandra Wachter。2019。《解释人工智能中的解释》。载于 Proc. FAccT' *19* 279 288 10 1145 3287560 3287574 . – 。[62] https://doi.org/ . / . Marçal Mora-Cantallops、Salvador Sánchez-Alonso、Elena García-Barriocanal 和 Miguel-Angel Sicilia。2021。《可信人工智能的可追溯性：模型与工具综述》。Big Data and Cognitive Computing 5, 2 (2021年5月), 20。https://doi.org/ 10 3390 5020020 . /bdcc Shweta Narkar、Yunfeng Zhang、Q. Vera [63] Liao、Dakuo Wang 和 Justin D. Weisz。2021。《Model LineUpper：支持自动化机器学习中多层级交互式模型比较》。载于 Proc. IUI'21。170–174。https://doi.org/10.1145/3397481. 3450658[64] Gina Neff、Anissa Tanweer、Brittany Fiore-Gartland 和 Laura Osburn。2017。《批评与贡献：改进批判性数据研究与数据科学的实践框架》。Big data 5, 2 (2017), 85–97。[65] Habib Asseiss Neto、Ronnie C. O. Alves 和 Sergio V. A. Campos。2020。《NASirt：基于自动化机器学习、利用实例级复杂度信息进行学习》。arXiv: 2008 11846 2008 11846 。https://arxiv.org/abs/。[66] Robert Nickerson、Upkar Varshney 和 Jan Muntermann。2013。分类体系开发方法及其在信息系统中的应用。*European Journal ofInformationSystems* 22 05 2013 10 1057 2012 26 ( )。[67] https://doi.org/。/ejis.。Nikolay O. Nikitin、Pavel Vychuzhanin、Mikhail Sarafanov、Iana S. Polonskaia、Ilia Revin、Irina V. Barabanova、Gleb Maximov、Anna V. Kalyuzhnaya 和 Alexander Boukhanovsky。2021。用于设计复合机器学习流水线的自动化进化方法。https://arxiv.org/abs/2106. 15397[69] Randal S. Olson、Nathan Bartley、Ryan J. Urbanowicz 和 Jason H. Moore。2016。用于数据科学自动化的基于树的流水线优化工具评估。载于*Proc GECCO'16*。485–492。https://doi.org/10.1145/2908812.2908918 [70] Jorge Piazentin Ono、Sonia Castelo、Roque Lopez、Enrico Bertini、Juliana Freire 和 Claudio Silva。2021。PipelineProfiler：用于探索自动化机器学习流水线的可视分析工具。*IEEE Transactions on Visualization and Computer Graphics* 27, 2 2021 390 400 10 1109 2020 3030361 ( ), –。[71] https://doi.org/。/TVCG.。Sinno Jialin Pan 和 Qiang Yang。关于迁移学习的综述。*IEEE Transactionson Knowledge and Data Engineering* 22, 10 (2010), 1345–1359。https://doi.org/ 10 1109 2009 191 。/TKDE.。Raja Parasuraman、Thomas B. Sheridan 和 Christopher D. Wickens。2000。人类与自动化交互类型和层级模型。*IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30, 3 2000 286 297 10 1109 3468 ( ), –。https://doi.org/。/ 844354 [72] Nicolas Prat、Isabelle Comyn-Wattiau 和 Jacky Akoka。2015。信息系统工件评估方法分类体系。*Journal of Management Information Systems* 32, 3 (2015), 229–267。https://doi.org/10.1080/07421222. 2015.1099390[74] G. Publio、Diego Esteves、Agnieszka Lawrynowicz、P. Panov、L. Soldatova、Tommaso Soru、J. Vanschoren 和 Hamid Zafar。2018。ML-Schema：通过模式和本体揭示机器学习语义。https://openreview.net/forum?id=B1e8MrXVxQ [75] Eric D Ragan、Alex Endert、Jibonananda Sanyal 和 Jian Chen。2015。可视化和数据分析中的溯源特征：溯源类型和目的的组织框架。*IEEE transactionson visualization and computer graphics* 22, 1 (2015), 31–40。[76] Jen Rogers、Austin H. Patton、Luke Harmon、Alexander Lex 和 Miriah Meyer。2021。严谨性在进化生物学设计研究中的实验所得洞察。*IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 1106–1116。https://doi.org/ 10 1109 2020 3030405 。/TVCG.。Dominik Sacha、Matthias Kraus、Daniel A. Keim 和 Min Chen。2019。VIS4ML：用于可视分析辅助机器学习的本体。*IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 385–395。https://doi.org/ 10.1109/TVCG.2018.2864838[78] Dominik Sacha、Michael Sedlmair、Leishi Zhang、John A. Lee、Jaakko Peltonen、Daniel Weiskopf、Stephen C. North 和 Daniel A. Keim。2017。所见即所得：通过交互式可视化实现以人为中心的机器学习。*Neurocomputing* 268 (2017), 164–175。https://doi.org/10.1016/j. neucom.2017.01.105[79] Janet Salmons。2008。期待原创性！利用分类体系构建支持原创工作的作业。载于 *Student plagiarism in an online world: Problems and solutions*。IGI Global，208–227。[80] Manuel Martin Salvador、Marcin Budka 和 Bogdan Gabrys。2016。利用 Auto-WEKA 在流程工业中采用混合适应策略调整多组件预测系统。载于*Proceedings of the Workshop on Automatic Machine Learning (Proceedings of Machine Learning Research, Vol. 64)*，Frank Hutter、Lars Kotthoff 和 Joaquin Vanschoren（编）。PMLR，纽约，美国纽约州，48–57。http://proceedings.mlr.press/v64/salvador_adapting_2016. html[81] Nithya Sambasivan、Shivani Kapania、Hannah Highfill、Diana Akrong、Praveen Paritosh 和 Lora M Aroyo。2021。"每个人都想做模型工作，而不是数据工作"：高风险人工智能中的数据级联。载于*Proc. CHI'21*。文章 39, 15 10 1145 3411764 3445518 页。https://doi.org/。/。[82] Elizabeth B.-N. Sanders 和 Pieter Jan Stappers。2014。探针、工具包和原型：协同设计中的三种创作方法。*CoDesign* 10, 1 (2014), 5–14。https://doi.org/10.1080/15710882.2014.888183[83] Aécio Santos、Sonia Castelo、Cristian Felix、Jorge Piazentin Ono、Bowen Yu、Sungsoo Ray Hong、Cláudio T. Silva、Enrico Bertini 和 Juliana Freire。2019。Visus：用于自动构建和整理机器学习模型的交互式系统。载于*Proc. HILDA'19*。文章 6，7 页。https://doi.org/10.1145/ 3328519.3329134[84] Sebastian Schelter、Joos-Hendrik Boese、Johannes Kirschnick、Thoralf Klein 和 Stephan Seufert。2017。自动跟踪机器学习实验的元数据和溯源信息。载于*Machine Learning Systems Workshop at NIPS*。27–29。[85] Vidya Setlur、Melanie Tory 和 Alex Djalali。2019。在可视化分析中推断描述不充分的自然语言话语。载于*Proc IUI '19*。40–51。https: //doi.org/10.1145/3301275.3302270

[86] B. Shneiderman。1996。视觉即优势：面向信息可视化的任务—数据类型分类体系。载于*Proc. VLHCC'96*。336–343。https://doi.org/10.1109/ VL.1996.545307[87] Renan Souza、Patrick Valduriez、Marta Mattoso、Renato Cerqueira、Marco Netto、Leonardo Azevedo、Vítor Lourenço、Elton F. de S. Soares、Raphael Melo、Rafael Brandão、Daniel Salles Civitarese、Emilio Vital Brazil 和 Marcio Ferreira Moreno。2019。计算科学与工程中机器学习生命周期的溯源数据。载于*Proc. WORKS'19*。1–10。https://doi.org/ 10 1109 49585 2019 00006 。/WORKS .。Fabian Sperrle、[88] Mennatallah El-Assady、Grace Guo、Rita Borgo、Duen Horng Chau、Alex Endert 和 Daniel Keim。2021。以人为中心的机器学习中以人为中心的评估综述。*Computer Graphics Forum* 40, 3 2021 543 567 10 1111 14329 ( )，–。https://doi.org/。/cgf。Fabian [89] Sperrle、Astrik Jeitler、Jürgen Bernard、Daniel A. Keim 和 Mennatallah El-Assady。2020。混合主动式可视分析中的协同自适应引导：学习与教学。载于*EuroVis* 可视分析研讨会 *(EuroVA)*。The Eurographics Association。https://doi.org/ 10 2312 20201088 。/eurova。Thilo [90] Spinner、Udo Schlegel、Hanna Schäfer 和 Mennatallah El-Assady。2020。explAIner：用于交互式、可解释机器学习的可视分析框架。*IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020 1064 1074 10 1109 2019 2934629)，–。https://doi.org/。/ [91] TVCG.。Holger Stitz、Samuel Gratzl、Harald Piringer、Thomas Zichner 和 Marc Streit。2019。KnowledgePearls：基于溯源的可视化检索。*IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 120–130。https://doi.org/10.1109/TVCG.2018.2865024[92] Senthil Karthikeyan Sundaram、Jane Huffman Hayes、Alex Dekhtyar 和 E. Ashlee Holbrook。2010。评估软件工程制品的可追溯性。*Requirements Engineering* 15, 3 (2010年9月1日), 313–335。https: //doi.org/ 10 1007 00766 009 0096 6 。/s - - - Rachel Taman 、 [93] Jake VanderPlas 和 Sohier Dane。2018。机器学习研究可复现性的实用分类体系。https://openreview.net/ forum?id=B1eYYK5QgX [94] Chris Thornton、Frank Hutter、Holger H. Hoos 和 Kevin Leyton-Brown。2013。Auto-WEKA：分类算法的联合选择与超参数优化。载于*Proc.KDD'13*。847–855。https://doi.org/10.1145/2487575. 2487629[95] Amrit Tiwana 和 Balasubramaniam Ramesh。2001。支持协作式信息产品演进的设计知识管理系统。*Decision Support Systems* 31, 2 (2001), 241–262。https://doi.org/10.1016/S0167- 9236 00 00134 2 [96] ( ) - Eliane R. A. Valiati、Marcelo S. Pimenta 和 Carla M. D. S. Freitas。2006。用于指导多维可视化评估的任务分类体系。载于*Proceedings of the 2006 AVI Workshop on BEyond Time andErrors: Novel Evaluation Methods for Information Visualization* (意大利威尼斯) *(BELIV '06)*。Association for Computing Machinery，纽约，纽约州，美国，1–6。https://doi.org/10.1145/1168149.1168169[97] Laura von Rueden、Sebastian Mayer、Katharina Beckh、Bogdan Georgiev、Sven Giesselbach、Raoul Heese、Birgit Kirsch、Michal Walczak、Julius Pfrommer、Annika Pick 等。2021。知识引导的机器学习——将先验知识融入学习系统的分类体系与综述。*IEEE Transactions on Knowledge and Data Engineering* (2021), 1–1。https://doi.org/10.1109/tkde.2021. 3079836[98] April Yi Wang、Anant Mittal、Christopher Brooks 和 Steve Oney。2019。数据科学家如何利用计算笔记本进行实时协作。*Proc CSCW'19* 3 (2019年11月), 1–30。https://doi.org/10.1145/3359141[99] Bochao Wang、Hang Xu、Jiajin Zhang、Chen Chen、Xiaozhi Fang、Yixing Xu、Ning Kang、Lanqing Hong、Chenhan Jiang、Xinyue Cai、Jiawei Li、Fengwei Zhou、Yong Li、Zhicheng Liu、Xinghao Chen、Kai Han、Han Shu、Dehua Song、Yunhe Wang、Wei Zhang、Chunjing Xu、Zhenguo Li、Wenzhi Liu 和 Tong Zhang。2020。VEGA：迈向端到端可配置的自动化机器学习流水线。arXiv: 2011 01507 2011 01507 。https://arxiv.org/abs/。 [100] Dakuo Wang、Josh Andres、Justin D. Weisz、Erick Oduor 和 Casey Dugan。2021。AutoDS：迈向以人为中心的数据科学自动化。载于*Proc.CHI' 21* 79, 12 10 1145 3411764 3445526 。文章页。 [101] https://doi.org/。/。Dakuo Wang、Q. Vera Liao、Yunfeng Zhang、Udayan Khurana、Horst Samulowitz、Soya Park、Michael Muller 和 Lisa Amini。2021。数据科学家希望实现多大程度的自动化？ arXiv: 2101 03970 2101 03970 。 [102] https://arxiv.org/abs/。Dakuo Wang、Justin D. Weisz、Michael Muller、Parikshit Ram、Werner Geyer、Casey Dugan、Yla Tausczik、Horst Samulowitz 和 Alexander Gray。2019。数据科学中的人机协作：探索数据科学家对自动化人工智能的看法。*Proc. CSCW' 19*, 24 10 1145 3359313 页。https://doi.org/。/ [103] Qianwen Wang 、 Yao Ming 、 Zhihua Jin 、 Qiaomu Shen 、 Dongyu Liu 、 Micah J. Smith、Kalyan Veeramachaneni 和 Huamin Qu。2019。ATMSeer：提升自动化机器学习的透明度与可控性。载于*Proc CHI' 19* 1 12 10 1145 3290605 3300911 。–。 [104] https://doi.org/。/。Daniel Karl I. Weidele、Justin D. Weisz、Erick Oduor、Michael Muller、Josh Andres、Alexander Gray 和 Dakuo Wang。2020。AutoAIViz：通过条件平行坐标揭开自动化人工智能的黑箱。载于*Proc. IUI' 20* 308 312 10 1145 3377325 3377538 . –. https://doi.org/ . / . [105] AlanF.T. Winfield和 Marina Jirotka。 . 伦理治理对于在机器人学和人工智能系统中建立信任至关重要。 皇家学会哲学汇刊*A*辑：数学、物理与工程科学 376, 2133 2018 20180085 10 1098 2018 0085 ( )， . [106] https://doi.org/ . /rsta. . Catherine Wong、Neil Houlsby、Yifeng Lu 和Andrea Gesmundo。 . 借助神经AutoML进行迁移学习。在 *Proc NeurIPS'18*中。Curran Associates Inc.，美国纽约州红钩，8366–8375。[107] K. Wongsuphasawat、D. Moritz、A. Anand、J.Mackinlay、D.Howe和 J. Heer。2016。Voyager：通过可视化推荐的分面浏览进行探索性分析。*IEEE Transactions on Visualization and Computer Graphics* 22，1（2016年1月），649–658。https://doi.org/10.1109/TVCG.2015.2467191[108] Doris Xin、Litian Ma、Jialin Liu、Stephen Macke、ShuchenSong和 AdityaParameswaran。2018。加速人在回路机器学习：挑战与机遇。在 *Proc. DEEM'18* 中。https://doi.org/10.1145/3209889。3209897 [109] Doris Xin、Eva Yiwei Wu、Doris Jung-Lin Lee、Niloufar Salehi 和 Aditya Parameswaran。2021。*AutoML*何去何从？理解自动化在机器学习工作流中的作用。https://doi.org/ 10 1145 3411764 3445306 。/ 。Chengrun Yang [110]、 Yuji Akimoto 、 DaeWon Kim 和 Madeleine Udell。 。OBOE：用于AutoML模型选择的协同过滤。在 *Proc KDD'19*（美国阿拉斯加州安克雷奇）中。 2019 1173 1183 10 1145 3292500 3330909 – 。https://doi.org/ 。/ 。 [111] Chengrun Yang、JicongFan、Ziyang Wu和 Madeleine Udell。2020。AutoML流水线选择：高效遍历组合空间。在 *Proc KDD '20* 1446 1456 10 1145 3394486 3403197 中。–。https://doi.org/ 。/ 。 [112] Qian Yang、Aaron Steinfeld、Carolyn Rosé 和 John Zimmerman。2020。重新审视人机交互为何以及如何在设计上具有独特的困难。 1 13 10 1145 3313831 3376301 – 。https://doi.org/ 。/ 。 [113] QuanmingYao、Mengshuo Wang、YuqiangChen、Wenyuan Dai、Yu-Feng Li、Wei-WeiTu、Qiang Yang 和 Yang Yu。2019。将人类排除在学习应用之外：自动化机器学习综述。arXiv:1810.13306 https://arxiv.org/abs/ 1810 13306 。 [114] Amy X. Zhang、MichaelMuller 和 DakuoWang。2020。数据科学从业者如何协作？角色、工作流与工具。*Proc CSCW'2020*，文章 022 2020 23 10 1145 3392826 （5月），页码。 [115] https://doi.org/ 。/ Baobao Zhang、Markus Anderljung、Lauren Kahn、Noemi Dreksler、Michael C. Horowitz 和 Allan Dafoe。2021。人工智能伦理与治理：来自机器学习研究人员调查的证据。arXiv: 2105 02117 [ ] 。cs.CY Yao Zhang、 [116] William Zame 和 Mihaela vanderSchaar。2020。AutoCP：用于准确预测区间的自动化管道。arXiv:2006.14099 https://arxiv.org/abs/2006.14099[117] Fuzhen Zhuang、Zhiyuan Qi、Keyu Duan、Dongbo Xi、Yongchun Zhu、Hengshu Zhu、Hui Xiong 和 Qing He。2021。迁移学习综合综述。*Proc. IEEE* 109，1（2021），43–76。https://doi.org/10.1109/JPROC.2020。3004555[118] Lucas Zimmer、Marius Lindauer 和 FrankHutter。2021。Auto-Pytorch：用于高效且稳健AutoDL的多保真元学习。*IEEE*模式分析与机器智能汇刊 01（2021），1–1。https://doi.org/10.1109/ TPAMI.2021.3067763 [119] Marc-AndréZöller和M. Huber。2021。自动机器学习框架的基准测试与综述。人工智能研究杂志 70（2021），409–472。[120] BarretZoph和 Quoc V. Le。2017。结合强化学习的神经架构搜索。https://arxiv.org/abs/1611.01578

A APPENDIX: AUTOML ARTIFACT TAXONOMY ADDITIONAL DETAILS

We describe the artifact properties according to our taxonomy. We use color highlighting through this subsection to to emphasize the **dimensions**, **categories**, and *characteristics* of our taxonomy (see Section 4). The exposition of our taxonomy proceeds in a hierarchical order, beginning with a dimension down to its respective characteristics.

维度 1：来源（“是什么生成了该工件？”）识别工件的来源，有助于了解背景，并追溯自动化机器学习流程中各项决策是如何作出的。在完全自动化的数据科学流程中，这些工件由计算流程生成，我们将其称为



图 8：制品的来源维度及其涵盖的类别和特征

“机器”，且没有人工干预。然而，完全自动化不仅难以实现，也并非总是可取，因此在现实中，工件可能有多种来源。例如，运行于 AutoML 流水线之上的可视分析混合主动系统。在此类系统中，分析师可以结合后端模型作出的自动化决策，以及在过程中通过界面提供的人类输入，获得一系列洞察 [37, 88, 89]。从宏观层面看，工件的来源可分为人类或机器。不过，在构建分类体系的过程中，我们还进一步细化了工件来源。人工工件可源自个人流程或组织流程。机器工件可源自自动化机器学习流程，以及负责协调自动化数据科学流程的整体软件基础设施（或系统）。最后，我们将数据单独列为一种独特来源，因为它同时涉及人类和机器来源。这些更细致的来源划分是分类体系中的类别，各自还具有其他特征。我们发现，许多工件通常都有明确的来源，但有些工件可能有多个来源。例如，许多与数据增强有关的工件，既可能源自人类意图，也可能源自对初始数据集的衍生处理。人类输入也可能由系统提示产生，即系统明确征求用户反馈。

- 人类，人类主要通过以目标和需求、规范 [30, 39], 以及与系统的交互 [12, 39, 78]的形式提供输入，来充当自动机器学习流水线的来源。虽然人类可以指提供输入的一个或多个个体，但我们更倾向于较狭义的解释，即单个人类向 AutoML 流水线提供输入或与之交互。正如后文将变得清晰的那样，“组织流程”是描述多个人类协同工作的更好来源指定。在个体人类来源中，我们发现了两个增加重要上下文的特征：角色和意图。我们发现工件类型可根据角色 *(c1.1.1)*而有所不同 [14, 44, 103]开展分析的个人。自动化机器学习系统可供未接受过数据科学或机器学习训练的个人使用。我们认为，这些个人所提供的输入的性质，以及他们用来提供这些输入的可供性，会不同于那些拥有更多领域专业知识的个人。例如，接受过数据科学或机器学习训练的个人可能会通过使用笔记本生成更多代码库工件 [98], ，而其他角色可能更多地依赖无代码解决方案，并且他们的输入更有可能通过界面控件或其他类型的语义交互被捕获 [23, 29]。

人类来源工件的另一个重要特征是个体的 意图 *(c 1.1.2)* 以表现为用户偏好模型、分析类型，甚至模型任务（分析师选择针对特定任务优化的模型）。HCI、可视化和机器学习社区使用了不同的术语来定义一个人在分析流程中希望执行的操作。任务是这三个社区都使用的常见术语（即 [9, 44, 106] ），这些任务可以与目标 [48] 或偏好相关联。最近，可视化研究人员开始使用意图作为一种概括性方式，来捕获从个体的任务到其目标 [28, 85]这一范围。我们选择使用这一术语，是因为它与我们的分析所捕获的工件多样性非常契合。

- 或许是自动化机器学习流程中最显而易见的制品，也是最不需要解释的制品。在我们的分类体系中，数据的主要特征区分它是初始输入还是从自动化机器学习流程中派生而来。

  初始 *(c1.2.1)* 数据集有时也被称为原始数据。我们避免使用“原始”一词，主要是因为没有任何数据集真正处于这种状态 [17, 31]。相反，我们使用“初始”数据集这一术语，以代替“原始”术语。此外，“初始”这一术语承认数据集可能在应用机器学习模型之前由人类或 AutoML 流程进一步转换或增强。

  相比之下， 数据集是在对初始数据应用转换时产生的。这些转换可能源于数据清理或整理操作 [45]（包括特征编码 [44, 111, 113], 新特征的派生 [101, 118], 或通过数据或特征嵌入创建新表示 [111]）。由此产生的衍生数据集由 AutoML 流程生成，其组成的变化有助于理解流程如何得出最终结果集 [13]。

- 不同程度的自动化会直接影响自动化机器学习流程生成的工件数量及类型。鉴于自动化机器学习理论上可以涵盖从超参数调优到完整端到端数据科学流水线 [38, 44, 119], 的范围，因此自动化机器学习流程产生的可能工件谱系可以非常广泛。然而，我们确定了贯穿这一谱系的工件的三个特征：结构、指标和结果。

  结构（*c1.3.1*）制品特征描述 AutoML 流水线中的一个组件，例如机器学习模型，或涵盖数据准备、特征工程和报告等环节的端到端流程 [26, 38, 44, 99, 111, 119]。此外，我们扩展了结构特征的定义，将用于训练或调优特定组件的算法制品纳入其中 [34], ，例如架构或更复杂的模型，如神经网络 [42], ，或流水线拓扑 [111], 、配置空间 [26, 119], 或搜索空间 [68, 70, 103]。最后，

我们将模型的任务作为其结构特征的一部分，因为这些任务在理解模型预期执行的功能方面发挥着重要作用，同时为架构增添了上下文。结构特征通常采取最终用户提供的规范形式，或由自动化机器学习流程自动生成。例如，我们将模型的最终架构或拟合结果视为一种由算法流程产生的、具有结构特征的自动生成制品。

指标 *(c1.3.2)*和[redacted]两个互补的特征，或许也是自动化机器学习流程制品中最受广泛审视的方面。指标指的是描述模型训练、验证和测试性能的度量。这些度量可以根据所使用的模型类型及其旨在解决的任务而呈现各种形式。然而，诸如总体准确率或平均准确率之类的基本度量往往是最常报告的。指标与应用于数据集的组件或管道的结果密切相关。同样，这一结果的确切性质取决于模型任务。两种常用类型是分类任务和聚类任务；不过，更先进的模型支持一组更复杂的任务，例如文档摘要、文本或图像生成等。

- [redacted]L 流程处于一个更大的软件生态系统中，该生态系统负责编排并执行其不同组件的计算指令（即数据清理、特征工程或机器学习步骤）。工件往往由系统生成，我们确定了此类工件的三个特征：输入、提示和流程。

这些工件的特征涉及它们由系统提供或生成的方式。一些工件作为 输入 *(c1.4.1)* 运作。它们可以来自人类流程，或来自在自动化机[redacted]系统计算层之间传输的数据或其他类型的工件。这些可以包括计算环境的配置文件 [11], 计算预算 [103], 或源代码 [11, 101]。工件也会因系统向个人呈现 提示 *(c1.4.2)* 以获取某些输入而生成，或通过自动流程 *(c1.4.3)* 生成。警报机制可以是提示个人采取某些行动的常见方式；此行动会产[redacted]发 AutoML 流水线更改的制品。例如，向个人发出其输[redacted]个变量之间高度相关的警报，可能导致他们从模型中移除一个特征。警报由执行相关性检查的自动化流程生成，其本身就是一个制品，但用户做出的选择（是否从模型中移除该特征）源于提示本身；该制品是用户的选择，并具有由提示生成的特征。

- [redacted]与现有的业务和组织实践结合使用 [14]。这些流程生成可以作为输入并直接集成到自动化机器学习流程中的工件，而其他则施加外在影响但不提供任何直接输入。

当组织制品直接影响 AutoML 流[redacted]以及最终用于决策制定的方式时，这些制品便具有集成（*c1.5.1*）特征。例如，定义数据结构的数据模式会受到业务实践的影响。然而，数据模式会影响所收集数据的类型、数据的存储与访问方式，并由此影响自动化机器学习流程 [19]中能够实现的目标或对其加以限制。构成集成组织流程的其他制品还包括数据增强，例如通过上下文增强（即人类提供的语义标注 [25] 或本体 [12]）以及基准数据集 [42, 69, 116, 119]。我们发现，这些制品紧密反映了组织开展实践的方式；相比之下，机器学习模型的底层数学规范在很大程度上与组织实践无关。

尽管这并非我们研究的重点（第 ??节），我们还考虑了外在（*c1.5.2*）[redacted]中包括组织内部规定自动化机器学习技术使用方式及其限制的法律程序或实践。这些工件并未像数据增强那样，直接集成到指定、开发或训练 AutoML 流水线各方面的流程中。尽管它们添加了相关的上下文信息，但它们与 AutoML 流程仍相隔一步；因此，我们将这些工件分类为外在。

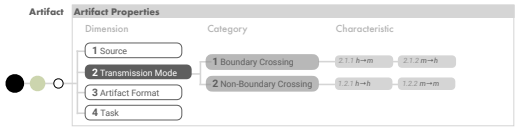

Figure 9: The Transmission dimension of an artifact and the categories and characteristics it encompasses

D2: Transmission Mode (***"Does it cross the boundaries between human and AutoML processes?"***) Our work emphasizes artifacts that cross boundaries between machine and human processes. However, there still exist artifacts that do not cross this boundary but that are important for transparently describing an AutoML processes. We include space for both in our taxonomy.

- **Category 2.1: Boundary Crossing** artifacts are passed between the human and AutoML processes, and vice-versa, which we interpret as characteristics of an artifact. An artifact that goes from a human to AutoML processes, *h→m (c2.1.1); human-to-* serves as input to the AutoML. This input includes artifact sources stemming from the data they provide, specifications and configurations, or even the imposition of integrated

或外在流程，例如需求文档 [2, 11, 101] 其可以规定自动化机器学习流程应执行的操作。我们还认为，管理自动化机器学习流程并确定其推进方式的操作也具有h→m特征，例如确定某个特定模型是否可以部署，尽管我们在构建分类体系时审阅的研究论文中没有任何一篇包含这样明确的制品。自动化机器学习流程可以自动生成报告或警报形式的输出，供人类使用。

这些制品具有*m→h*（*c*
加的 AutoML流程会生成或预期生成以报告 [101] 或自动生成的模型卡片 [61]形式呈现的输出解释。用于检测数据或模型中异常的方法 [22], , 尤其是概念漂移的存在 [13], , 由 AutoML系统启动。这些自动化方法会生成制品，以警报、自动生成的报告或仪表板的形式呈现给用户。

-
体系的非边界跨越项目数量；对这类工件进行全面调查，很可能单独就能写成一篇或几篇研究论文。此外，其他研究文献也考察了从知识管理 [47] 到 API 及计算基础设施的其他考量等一系列此类工件。为了能够将我们的分类体系与此类工件联系起来，我们纳入了

*h→h (c2.2.1;*人类对人类*)* 和 *m→*
ine)
非边界跨越工件的特征。



图 10：制品的工件格式维度及其涵盖的类别和特征

D3：(“工件呈现什么形式？”) 在我们的分类体系中，一个实际考量因素是工件所采用的不同形式。我们观察到，不同系统会以不同格式生成工件，尽管当前系统主要侧重于数值型工件。可视化自动化机器学习流程的系统，例如 PipelineProfiler [70], ATMSeer [103], 或 AutoVizAI [104], , 在考虑可视化什么以及如何可视化时，会受到这些工件格式的影响。随着我们更全面地考察自动化机器学习流程以及人类与机器流程之间的交接，并且随着自动化机器学习系统的自动化程度不断提高，我们认为工件格式将更加多样。创建未来的自动化机器学习系统或工具时，若旨在呈现来自各种数据科学及非数据科学角色的工件，就必须认识到工件格式的多样性。此外，这些工件既可以构成自动化机器学习流程的输入，也可以构成其输出。在我们的分析中，我们确定了这类工件格式的四个类别：单值、多个值、规范和报告。

-
制品通常是一个单一值，例如单个摘要统计数据，或告知个人某项结果的文本警报。我们发现，这类制品通常具有数值特征 *(c3.1.1)* 或文本特征 *(c3.1.2)*。对数据 [2
L 管
以及来自个人的反馈，也往往以评论的形式出现，因此同样是具有文本特征的制品。

-
然扩展是包含多个值的制品。我们针对这些数据的形状，分别阐述了表格 *(c3.2.1)*和张
特征。
献中，张量通常明确指代N维数值数组，其范围可以从单一标量值（N=0）到向量（N=1），最终到多维数组（N>1）。由于“张量”一词具有很强的数值含义，我们加入“表格”这一术语，以涵盖包含混合类型数据（数值、有序、类别）的制品；只有一列的表格就是一个列表。我们确实考虑过是否应将单值简单地视为表格和张量的特殊情况，但由于我们还希望强调具有相似或混合变量（列、属性）类型的数据，因此选择将单值与多个值分开，以便更容易识别这一特征。

我们还将一种图（*c3*
多种数据类型构成。“图”一词具有多重含义，既可以指可视化，也可以指一种数据结构。此处，我们使用该术语来表示由节点和边构成的图数据结构。此外，我们将“图”作为一个总称，涵盖网络和树结构；网络和树同样是图的特殊情况；树是一种具有层次化、有向且无环结构的图。我们发现，行为图、交互日志或交互序列 [6, 12, 39, 91] 是具有图特征的制品的常见示例。初始数据集也是可能具有图特征的制品，例如本体或知识图谱 [1, 97]。

-
随着 AutoML 流程从专注于模型训练扩展到数据科学流程的多个阶段，其复杂性不断增加。因此，AutoML 流程可以描述为一组可互换组件 [99, 111]（也称为‘机器学习原语’ [38, 70]），这些组件被组合成最终的工作流配置。我们在此使用扩展的 AutoML 组件定义，不仅包括机器学习模型及其依赖关系（特征工程、数据准备），还包括报告、任务制定以及其他此类阶段 [44]。根据自动化水平，规范可以表示单个组件或整个工作流。这些规范通常采用配置文件或

源代码，然而，也有人探索了更广泛的规范空间，其中包括方程和逻辑规则 [97][3]。根据自动化程度，单个组件 *(c3.3.1)* 可以完全由人类指定，也可以通过学习流程自动生成规范。规范的类型会根据所指定的组件而有所不同。例如，机器学习模型组件的规范将不同于特征生成组件的规范，而后者也将不同于数据可视化组件的规范。虽然这些流程中的许多过去是数据科学家的手动工作，但 [101, 119] 这些流程正变得越来越自动化 [44]。

目前，许多自动化机器学习系统和工具包可以说主要关注机器学习模型组件及其结构规范的优化 [69, 94, 110],；而在深度学习的情况下，则关注其架构规范 [42, 118]。然而，随着更复杂的端到端自动化机器学习系统不断涌现，个人无需指定许多组件，因为这些组件可以通过搜索自动从自动化机器学习工作流 *(c3.3.2)* 设计中派生出来。在这些情况下，指定搜索空间的初步配置 [2, 26, 38, 80, 99, 103, 113, 119]。

- 许多报告任务由人类完成，通常不被视为 AutoML 流程的一部分，但我们看到未来这种情况将会改变。随着时间的推移，AutoML 流程将自动生成解释器 [90], 报告 [40, 101], 和数据可视化 [53]。这些将与围绕 AutoML 流程的主要分析目标、决策过程溯源以及最终报告关键绩效指标的仪表板的人类报告产物并存 [101]。这些报告产物的最终输出取决于它们是由个人还是由 AutoML 流程指定的。我们发现它们通常具有两个特征，静态 *(c3.4.1)* ，例如 PDF 文件或演示文稿，或交互式 *(c3.4.2)* 仪表板。

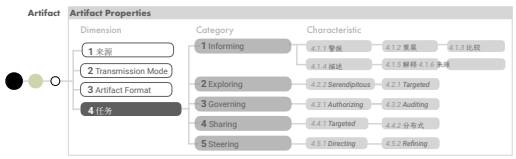

图 11：制品的任务维度及其涵盖的类别和特征

D4: (“它的预期目的是什么？”)。人工分析任务和机器学习模型任务具有互补（即便不是重叠）的目标 [88]。人类或机器希望完成的任务会影响它们之间交接的工件的特征，尤其是如果这些任务产生跨越人类与机器流程边界的制品。在我们的分析中，我们运用自己的判断来确定不同自动化机器学习产物旨在支持的任务类型。然而，在制定工件可支持的任务类型分类体系时，我们也参考了先前特定于自动化机器学习和机器学习的任务分类体系 [19, 93, 97], 、视觉分析类型学 [9, 48], 以及其他分类系统 [44, 77, 87, 88]，这些分类体系来自各个学科的报告。同样，端到端自动化机器学习系统是一项不断发展的技术，这些系统的当前实现在所支持的任务方面各不相同。因此，我们提醒读者，这些任务是我们在实际和理论自动化机器学习系统中识别出的任务的综合。

- 目前工件最常见的任务是告知。此类信息的目标受众可能有所不同，具体取决于工件是否跨越边界。我们归纳出这类工件的六项特征，以下按字母顺序介绍。警报（*c4.1.1*）这一特征与由数据质量或模型质量警报产生的工件相关 [13]，当它们以这种形式出现时，也会具有 $m \rightarrow n$ 相应特征。我们推测，未来的系统甚至可能自动纳入警报，例如以文本反馈的形式 $h \rightarrow m$；比如，人类在评估自动化机器学习流程时发现错误或遗漏，并希望向算法发出警报。此外，我们还推测存在非正式的 $h \rightarrow h$ 工件，例如评论、注释或其他形式，它们会在审查结果或决策后启动模型的纠正机制。策展（*c4.1.2*）是共享或保存的见解常见的一项特征；我们发现，这也是分享探索性可视化分析关键发现的一种常见方式 [107]。自动化机器学习系统日益能够生成一系列经过策展的见解，并按优先顺序呈现，供个人参考。具有比较（*c4.1.3*）特征的工件，在许多情况下侧重于比较单个工作流组件的不同状态；它们往往也是数据可视化。例如，个人可能希望对分类模型的某个阈值进行敏感性分析。或者，他们也可能希望比较某个组件随时间的变化，以了解其演变情况。可视化研究已对比较任务进行了广泛研究，相关分类法 [32, 33] 也能为自动化机器学习流程中的比较工件提供更多背景信息。具有描述（*c4.1.4*）特征的工件描述了自动化机器学习组件或流程的状态。这些工件包括规范、需求或监管文件，用以说明各个组件的作用，以及自动化机器学习组件的顺序或配置 [11, 101]。此外，我们认为，描述性工件还可以是统计分析，用于概述自动化机器学习流程中的数据、结果或其他属性 [12, 44]。我们将这些描述性工件与主要目的在

[3]尽管作者认为这些是知识表示，但它们符合我们对规范的定义。

旨在解释（*c4.*来。关注建模组件的解释对于透明度日益重要，而且这些解释正由这些组件自动生成 [62, 88, 90, 101]。此类工件可以采用报告的形式，也可以是 SHAP、LIME 等技术的输出 [46]。虽然这类解释往往聚焦于黑箱模型，我们也将所谓"白箱"模型的自解释特征纳入定义 [5, 58]。具有告知意图的工件的最后一项特征是溯源（*c.4.1.5*）。溯源工件关注自动化机器学习流程的沿袭关系和分析历史。即使实现了完全自动化，人类监督也能触发自动化机器学习流程配置的优化或重新训练，例如通过与其他工件交互来实现 [22, 55, 89]。在每次迭代中，无论由人类驱动还是由自动化流程触发 [13]，溯源流程都可以捕获个人与不同组件的交互，例如数据可视化 [16, 91]，可能会通过语义交互影响机器学习组件 [10, 23, 29]。虽然溯源特征也可以视为描述性特征，但我们认为，和解释一样，它们具有更具体且更积极的作用，并不只是描述系统或某个组件的状态，因此应当单独看待。

- 个人甚至自动化流程探索数据、模型或管道配置空间时生成。我们区分探索性工件，这些工件是定向 *(4.4.1)*于某些特定目标（例如，假设验证）的.以及那些通过纯粹偶然发现的 *(4.4.2)*发现产为许多来自探索性可视分析流程的工件 [6, 16, 91]。其中可以包括书签或已保存的洞见，这些被策划 [107]，成为其中偶然发现的特征的探索性工件。

- 程在较新的研究中有所出现，但在自动化机器学习文献中并未受到广泛关注 [14, 101]。随着法律要求和组织实践不断变化，以更广泛地规范机器学习/人工智能和自动化机器学习技术，这些流程的重要性只会与日俱增 [105, 115]。根据现有研究，我们提出治理制品的两个特征：授权和审计。授权（*c4.3.1*）是指允许自动化机器学习系统甚至单个分析师执行某个组件；仍取决于系统固有的自动化程度。我们在 [14] 和 [101] 中发现了一些关于此类潜在制品的证据，这些文献描述了对监督的需求，例如由数据科学家授权部署组织内其他人创建的模型，或通过自动化规则执行反偏见等规则。与这些制品相辅相成的是那些能够……的制品。

审计（*c4 .3.2)*, for example, decision optimization forensics reports [101]. Artifacts with auditing characteristics rely on a variety of artifacts, especially those that are intended to inform.

- 类完全无意干预自动化机器学习流程，这些流程的结果也会被集成到其他人类和组织流程中，以共享知识。旨在面向（*c4.4.1*）特定角色的工件具有特定用途。我们认为计算笔记本 [114] 具备此类特征，因为它们旨在与其他技术数据工作者，或仅与数据科学家本人共享。具有分发（*c4.4.2*）特征的工件旨在面向广泛的受众。这些工件可以包括演示文稿或报告，其中许多依赖于对其他自动化机器学习产物的汇编和分析。

- 发挥定向或干预作用——介入自动化机器学习流程。具有定向（*4.5.1*）于初始化自动化机器学习流程，或将其引导至特定轨迹。这些工件可以包括人类设定初始目标、提供训练数据，或为组件提供初始配置 [113]。它们还可以包括定义自动化机器学习组件最终配置的算法流程。相比之下，具有细化（*c4.5.2*）特征的工件比定向更加细微，因为它们主要建立在现有模型结构之上。微调操作（例如迁移学习）[3, 71, 117]，会产生用于细化现有神经架构的工件 [106]，，而不是进行成本高昂的神经架构搜索 [2, 42, 66, 99, 106, 120]。

我们将前者视为细化，而后者则利用初始设计空间配置来定向整体神经架构搜索流程。