arXiv:2303.01575v1 [cs.HC] 2 Mar 2023

DataPilot：在可视化数据准备过程中利用质量和使用信息进行子集选择

Arpit Narechania 佐治亚理工学院 亚特兰大，美国
ar
pitnarechania@gatech.edu

Fan Du Adobe Research 圣何塞，美国
dufan2013
@gmail.com

Atanu R Sinha
Adobe Research
班加罗尔，印度
atr@adobe.com

Ryan A. Rossi
Adobe Research 圣何塞，美国
ryrossi@adobe.com

Jane Hoffswell
Adobe Research
西雅图，美国
jhoffs@adobe.com

Shunan Guo
Adobe Research
圣何塞，美国
sguo@adobe.com

Eunyee Koh
Adobe Research 圣何塞，美国
eunyee@adobe.com

Shamkant B. Navathe
佐治亚理工学院
亚特兰大，美国
sham@cc.gatech.edu

Alex Endert
佐治亚理工学院
亚特兰大，美国
endert@gatech.edu

摘要

从大型、陌生的数据集中选择相关的数据子集可能很困难。我们通过建模和可视化两类辅助信息来应对这一挑战：(1)质量——执行某些分析任务所需数据的有效性和适当性；以及 (2) 使用情况——数据在多个用户中的历史利用特征。通过一项有 14 名数据工作者参与的设计研究，我们将这些信息整合到可视化数据准备与分析工具 DataPilot 中。DataPilot 呈现有关数据“好、坏与丑陋”方面的视觉提示，并提供图形用户界面控件作为交互功能，引导用户进行子集选择。通过一项有 36 名参与者参与的研究，我们考察了 DataPilot 如何帮助用户浏览大型、陌生的表格数据集，准备相关子集，并构建可视化仪表板。我们发现，用户选择了更小、更有效的子集，这些子集具有更高的质量和使用情况，并且用户取得了更大的成功度和信心。

CCS 概念

• **以人为中心的计算 → 可视化中的实证研究；可视化系统与工具。**

关键词

数据质量、数据使用情况、子集选择、数据准备、可视化、可视化数据分析、设计研究

ACM 引用格式：

Arpit Narechania、Fan Du、Atanu R Sinha、Ryan A. Rossi、Jane Hoffswell、Shunan Guo、Eunyee Koh、Shamkant B. Navathe 和 Alex Endert。2023。DataPilot：在可视化数据准备过程中利用质量和使用信息进行子集选择。载于*2023 年 ACM 人机交互大会论文集*

允许免费制作本作品部分或全部内容的数字副本或纸质副本，供个人或课堂使用，但不得为营利或商业目的制作或分发副本，且副本首页必须附有本声明和完整引文。本作品中第三方组件的版权必须得到尊重。其他用途请联系所有者/作者。*CHI '23，2023年4月23日至28日，德国汉堡*© 2023 版权所有归所有者/作者所有。ACM ISBN 978-1-4503-9421-5/23/04。
https://doi.org/10.1145/3544548.3581509

*Human Factors in Computing Systems（CHI '23），2023 年 4 月 23 日至 28 日，德国汉堡。*ACM，美国纽约州纽约市，18 页。
https://doi.org/10.1145/ 3544548.3581509

1 引言

数据从来都不是真正的原始数据 [42]，在用于预期目的之前，仍需经过清洗、集成、转换和筛选等处理 [92]。现代组织通常会以原生形式摄取所有传入数据，计划稍后再进行分析 [32]。这种“先加载”理念导致的信息过载，给数据导航和知识发现带来了若干挑战 [24, 33, 44]。例如，设想这样一项用户任务：“分析一个大型电子商务数据集，并构建一个仪表板，以可视化近期地理趋势，从而预测未来销售额。”要完成这项任务，用户必须先找出与客户位置相关的数据属性（例如“邮政编码”），然后应用时间过滤器（例如按月）选择所需的数据记录。遗憾的是，不熟悉数据的新用户可能会采用“试错”式检查策略 [21]，从而选中无关或质量较差的属性，同时遗漏重要属性，进而损害后续分析的结果。即使是经验丰富的用户，也可能依赖自己过去的使用经验，而不去探索新数据集中的新属性，这同样会使分析结果受到质疑。此外，用户可能会花更多时间查找相关数据，而不是执行手头的分析任务 [33]。因此，我们提出以下问题：“如何设计用户界面，引导用户分析大型且陌生的数据集，并为下游分析和可视化任务（例如构建仪表板和客户细分）选择相关且有效的子集？”

我们采访了某大型科技公司的 14 位数据工作者，他们会选择数据子集（从较大的数据集中提取较小的属性和记录集合），用于制作仪表板（数据分析师）、训练机器学习模型（数据科学家）以及开展数字营销活动（营销人员）。所有数据工作者都强调了数据质量的重要性；其中一些人由于缺乏必要的技能而依赖他人准备这些数据子集，他们还谈到了呈现其他

数据特征（例如数据在不同用户间的使用情况）的潜力。数据工作者的这些反馈表明，我们需要一种交互式自助工具，利用两类辅助信息促进数据准备：(1) 质量和 (2) 使用情况。我们使用数据、相关元数据及相应的使用日志对这些辅助信息进行建模，并以可视化方式呈现给用户，帮助他们进行子集选择与分析；尽管目的各异，这项任务是他们都会执行的。

既有研究从多个视角界定数据质量：消费者 [39], 商业 [35, 51, 74, 87, 89], 和基于标准的 [17, 77]。难以用一个定义涵盖不同上下文 [39]。结合本研究的上下文，我们将质量 定义为 “执行特定分析任务所需数据的有效性和适用性。” 质量很重要，因为数据往往杂乱无章，而组织“先加载再说”的理念往往会导致“大数据坟场” [105]，其中充斥着大量缺失、错误和无关的信息。理想情况下，这些数据缺陷应促使组织采取纠正措施，甚至停止使用数据；然而，大多数组织未能维护数据质量标准 [83] ，因为 “ 每个人都想做 [机器学习] 模型工作，而不是数据工作” [99]。在本研究中，我们对专家认为重要的三个质量维度 [89],进行建模：（1）完整性：非.....的频率
(2) 正确性：数据中正确值的频率。
(3) 客观性：值符合目标分布的程度。
数据中的缺失值。

我们将使用情况 定义为“多个用户对数据的历史使用特征”，这一概念受到“数据效用”描述符 [101]的启发。用户在工作中经常协作 [6, 26, 64, 99, 128], ，但协作更多围绕代码而非数据展开 [65]。组织内部数据的创建和共享方式尚未得到充分研究 [65]。我们认为，利用当前和过去用户的使用日志以及元数据，可以成为引导其他用户的一种方式。受到数据工作者使用案例的启发，我们为子集选择和仪表板构建任务归纳出使用情况的三个维度，其中数据指表格数据集的属性和记录：

(1) 子集中的使用：将数据放入其子集的用户百分比。(2) 筛选器中：对数据应用筛选条件的用户百分比。(3) 可视化中：将数据可视化的用户百分比。

我们将质量和使用信息整合到可视化数据准备与分析工具 DataPilot 中。DataPilot 可帮助用户从大型表格数据集中准备一个子集，用于构建可视化仪表板。具体而言，DataPilot 会针对质量和使用情况的各个维度计算一个标准化分数，满分为 100 分。例如，“Profit”属性的 子集中的使用 得分为 94 分。DataPilot 还会通过视觉提示引导用户了解数据的“优点”和“问题”，例如将缺失值和错误数据值标红突出显示。最后，DataPilot 提供图形用户界面（GUI）控件作为交互功能，帮助用户进行子集选择，例如使用范围滑块筛选掉不太受欢迎的数据，并使用排序控件将具有相似特征的数据排序并归为一组。现代数据工具 [7, 25, 47, 79, 81, 107, 110, 112] 提供了各种功能，例如交互式 GUI，以帮助用户准备数据；然而，据我们所知，尚无工具利用使用日志及相关元数据中的使用信息，提供有助于交互式子集选择与分析的交互功能。

我们对36名参与者进行了用户研究，考察DataPilot用户界面如何引导用户（将他们推向某个方向）浏览大型且陌生的表格数据集，准备相关子集，并构建可视化仪表板。研究结果表明，质量和使用信息相结合，有助于用户更成功、更有信心地创建更小且有效的子集。我们将有效子集定义为其中质量和使用情况总体得分较高的属性和记录占比更高的子集。值得注意的是，参与者对过度依赖以往用户的使用行为表示谨慎，因为这可能减少对高质量数据的探索（降低对新颖性的追求），转而侧重于利用（重复迄今为止行之有效的做法）。与惯常观点相悖，我们的研究结果还呼吁可视化数据分析工具优先考虑将数据准备功能直接整合到分析工作流程中，以促进更有效地利用数据。

本研究的主要贡献包括：
(1) 对 14 位数据工作者开展设计研究，探讨他们在数据准备与分析过程中面临的任务和挑战。研究揭示了数据质量的重要性，以及呈现使用情况等其他特征的潜力，这有助于改进这些工作流并促进用户协作（第 3 节）。(2) 利用数据、相关元数据和用户的使用日志，对两类辅助信息进行建模：**质量** 和**使用情况**（第 4 节）。(3) 将质量和使用信息整合到可视化数据准备与分析工具 **DataPilot** 中，以在子集选择与分析过程中为用户提供指导（第 5 节）。(4) 对 36 位参与者开展用户研究，结果表明，在子集选择和仪表板构建任务中，DataPilot 能帮助用户从大型且陌生的数据集中更成功、更有信心地选出较小的有效子集（第 6 节）。

请注意，判断选定子集和创建的仪表板是否真正有效，取决于最终目标和其他情境因素，因此需要专家评估；由于参与者并非领域专家，我们没有探讨这一方面。此外，尽管 DataPilot 专注于子集选择，但仍需借助其他工具和研究，评估其在其他应用中的下游分析任务，例如排序和聚类。

2 相关工作

2.1 数据准备

数据准备（或预处理）包括通过收集、集成、转换、清洗、缩减和离散化来分析数据，以确保获得高质量的结果 [129]。

2.1.1 子集选择。子集选择（或数据缩减）涉及缩小数据集 [36, 56, 91] 的规模；可通过两种方式实现：特征集缩减（缩减表格数据集的属性或列）或样本集缩减（缩减表格数据集的记录或行）。特征集缩减常用于训练机器学习模型，用户可删除无关特征 [57] ，或通过降维技术 [36]减少特征。样本集缩减常见于市场细分 [111]，其中会筛选出特定的消费者群体，以满足细分市场的特定目标。这些技术已被用于应对选择偏差，例如通过可视化子集与原始数据集 [12, 44] 的对比情况。在这项工作中，我们通过向用户呈现数据质量与使用信息，支持子集选择任务。

2.1.2 数据质量评估与工具。真实世界数据集往往"脏乱"，包含各种数据质量问题 [58]，据估计，这些问题每年给各类组织造成数万亿美元的损失 [48, 95]。数据质量至关重要，它能确保使用这些数据的系统以高效、可扩展、准确且无偏差的方式执行预期任务 [8, 19, 22, 46, 78]。Umbrich 等人 [114] 指出，即使元数据质量较低（元数据缺失），也会影响数据集的发现与使用。然而，Kandel 等人 [61]发现，从业者认为数据整理既枯燥又耗时。Sambasivan 等人 [99] 提供了"数据级联"的实证证据——一系列相互叠加的事件，会使数据质量问题产生负面影响并逐步向下游传递。因此，越来越多的研究致力于理解和改善数据质量，以避免"垃圾进，垃圾出"问题 [49, 96]。

数据营养标签 [53]框架与食品上的营养成分标签类似，会突出显示数据集的"成分"，帮助判断该数据集是否适用于特定的统计使用场景。DataPilot 提供类似的一目了然的数据集质量信息。Tableau Prep [107], OpenRefine [47], 和 Wrangler [60]是自助式数据准备工具，提供交互功能，供用户在分析前探索、清理、构建数据结构并调整数据形态。与 DataPilot 最相关的是 Profiler [62], ，这是一款用于评估表格数据中质量问题的可视分析工具；Profiler 会应用数据挖掘方法，自动标记有问题的数据，并提供协调式摘要可视化，以便结合上下文评估数据，不过它不提供类似 DataPilot 的使用信息。

Pipino 等人 [89] 最初提出了十六个客观和主观维度，用于评估数据质量；此后这些维度又得到了扩展 [3, 69, 87, 101, 103, 115] ，因为"并不存在一套适用于所有情况的指标" [89] ，也"没有一种占据主导地位的工具" [3]。根据领域专家受访者的反馈，我们对其中三个维度（完整性、无错误性和客观性）进行建模，以引导用户关注潜在的数据质量问题。虽然数据清理（例如填补缺失值）[18, 20, 47, 60, 72, 73, 76, 88, 94,104, 125] 留待未来工作处理，但 DataPilot 目前提供了新颖的图形用户界面交互功能，以支持子集选择。

2.1.3 用户之间的协作。 先前研究探讨了人们如何协作以共享和获取信息[11, 55, 64, 65, 99, 128] [31] 。

社会透明性理论指出，通过让参与者及其活动彼此可见，可以设计数字系统来支持大型群体的协作。这些协作会产生直接成本（例如员工薪酬）和间接成本（例如因用户偏好和可用性而导致的时间延迟）[98, 108, 109], ，因此需要努力减轻低效问题。

在可视化领域，协作系统 [55, 117, 120,123] 致力于同时支持同步 [116] 和异步 [50, 122] 模式。在数据科学领域，Auto-suggest 会在计算笔记本中推荐数据准备步骤 [126], ，但这些建议基于先前编写的代码，而非使用日志。研究表明，在计算笔记本中进行数据整理时，向用户提供易读且可复用的代码（而不是要求手动编程）能够提升用户的效率、信任和信心 [27]。

在数据库领域，研究表明，专家提供丰富的入门查询能够帮助非专家使用 SQL 对临时数据库进行数据分析 [54]。在机器学习（ML）和人工智能（AI）领域，Almahmoud 等人 [6] 研究了不同团队成员如何交流机器学习模型的质量；他们发现，以用户为中心和以模型为中心的性能概念之间存在不匹配，并且人们难以理解自身角色之外的关切。Ehsan 等人 [29] 发现，社会透明度有可能校准对人工智能的信任、改善决策、促进组织集体行动，并培养整体可解释性。这些先前研究促使我们对数据以往使用情况中的使用特征进行建模和呈现，以在子集选择过程中帮助用户。除了数据质量之外，DataPilot 对使用信息的建模也是其一项关键创新。

2.2 数据、分析溯源与引导

随着大数据的迅速发展，越来越多的数据 [24] 和元数据 [113]（例如应用程序日志）被存储和处理。在数据库环境中，"数据溯源"用于推理数据对象的当前状态 [127], ，例如描述其溯源特征（"数据描述符" [101]）、研究安全溯源方案及相关问题（Zafar 等人 [127]），以及记录数据和模型（"事实说明书" [9]）以及计算工作流（Wings/Pegasus [63]）的用途、性能、安全性和防护性。

溯源信息也被用于研究数据集再利用。Koesten 等人 [66] 描述了一项案例研究，该研究确定了以 GitHub 特有的参与度指标（例如复刻、关注者、星标和提交者的数量）形式存在的数据集溯源信息，如何预测数据集被再利用的可能性。为促进数据导航并推动再利用，许多开放数据门户会利用和/或向用户呈现溯源信息 [2]，包括"浏览量最高" [1], 、"高价值" [1], 和"热门" [59, 90] 数据集，以及示例项目和用户讨论区的访问入口 [58, 66]。

在可视化上下文中，尤其与本文工作最相关的是，"分析溯源"会追踪用户与可视化系统的交互，以概览其意义建构过程 [86]。随后，这些信息会用于产品和用户行为分析，例如生成个性化内容 [121], 、减轻有偏差的分析行为 [84, 119], 、增强用户信任 [9],、推荐其他设计选择 [28, 71], 以及开展可视化数据探索 [122]。在人机交互上下文中，先前交互轨迹已被用于通过滚动条历史记录重新访问页面的常用区域 [4], 、追踪用户与文档的交互 [52], 、促进群件协作 [45], ，以及在浏览网页时使用眼动 [85] 和鼠标追踪 [10]来追踪用户注意焦点。

在描述可视化与数据分析中的数据溯源时，Ragan 等人 [93]提出了一个组织框架，涵盖分析溯源的五种类型和六种用途；我们的工作最符合"数据"类型（数据变更和移动的历史，其中可以包括子集选择、数据合并、格式设置、转换，或执行模拟以摄取或生成新数据）以及"协作交流"用途（与进行相同分析的其他人交流和分享数据、信息与想法）。Ceneda 等人 [16] 将可视分析中的引导划分为三个程度（定向、指导、规定），用于说明用户需要引导的程度以及系统提供引导的程度。DataPilot 通过视觉提示指出数据质量和使用情况的优缺点，提供干预程度最低的"定向"引导。

3 设计研究与专家访谈

为了解数据分析前数据准备过程中与子集选择相关的用户任务和挑战，我们采用了设计研究方法论 [102]，通过半结构化访谈、头脑风暴和反馈会议开展研究，并围绕原型设计与开发进行迭代。

我们采访了某大型科技公司的 14 位数据工作者（E_{1-14}；10 位男性；4 位女性），以了解他们当前的任务、面临的挑战及需求。这些专家包括数据工程师（*3* 位：E_{1-3}）、数据分析师（*3* 位：E_{4-6}）、数据科学家（*4* 位：E_{7-10}）、数字营销人员（*3* 位：E_{11-13}）以及一位用户界面工程师（*1* 位：E_{14}），相关工作经验为 3 至 34 年（中位数：15.5；μ=16.86）。我们通过定向邮件联系和滚雪球抽样策略相结合的方式，从企业内部网络中招募这些专家。为期十二周的访谈均通过 Microsoft Teams 远程进行；与后期每周一次的访谈相比，前期访谈更加频繁，也更为灵活。

总计进行了八场会议（包括三场后续会议），每场约 45 分钟，通话中有 1 至 7 位领域专家和 2 至 3 位研究管理员参与（会议S_1: E_2, S_2: $E_{7,8}$, S_3: $E_{3,9}$, S_4: $E_{11,12}$, S_5: $E_{1,4-6,10,13,14}$, S_6: E_2, S_7: $E_{4,5}$, S_8 E_{14} ：）。这些专家彼此建立了良好的工作关系，也了解对方的优势和专业知识，这与一群完全陌生的人开会不同。对我们而言，与这样一个凝聚力强的团队讨论问题和解决方案很有价值。例如，项目经理常常依赖工程师对某个想法的技术可行性的看法；同样，工程师在涉及优先级排序和时间安排的问题上，也会征询项目经理的意见。在这些会议期间，一位研究管理员分享了 PowerPoint 演示文稿，另一位则在主持讨论的同时做笔记，讨论围绕以下问题展开：

(1) “你日常会完成哪些与数据与分析相关的任务？(2) “你处理哪些类型的数据？如何准备这些数据？你会使用哪些工具？”(3) “你会如何与组织内外的其他人就相关任务进行协作？”(4) “你在执行任务时会遇到哪些挑战？如何克服这些挑战？”

3.1 用户任务与挑战

我们采用归纳式主题分析对领域专家的口述引文进行了编码 [13],，并根据他们的角色、所执行的任务、面临的挑战，以及对质量和使用情况的看法进行分类。我们在补充材料中提供了这些内容，但已进行匿名化处理。

我们发现，不同用户角色所承担的任务差异很大。数据分析师会选择关键绩效指标（属性），并筛选相关记录，以设计交互式仪表板并为企业高管准备报告。营销人员会制定战略细分规则，对客户行为数据和人口统计数据进行筛选，例如设置筛选条件，筛选出客户以开展定向数字营销传播。数据科学家会从现有数据集中选择相关属性（特征）和记录（观测值）子集，以构建预测模型。数据工程师帮助其他用户（如营销人员）为各种分析或运营用途准备数据；他们还会监控组织的数据存储库，以控制其存储与成本占用。用户界面（UI）工程师帮助设计可扩展的交互式 Web 应用，以满足各种面向用户的使用场景。

我们还了解到，这些专家在执行任务时面临许多普遍性和特定领域的挑战。所有人都向我们表示，数据质量很重要，例如：“数据是否完整？是否正确？是否存在偏差？数据类型是否正确？”他们还要求通过在数据准备和分析过程中提供额外引导，让用户更加了解质量问题。特别是，E_1（数据工程师）指出，偏斜是一个重要问题，“如果所有数据都来自北加利福尼亚（美国），会怎样？”E_{11}（营销人员）从营销角度指出，按偏斜（和/或稀疏）属性对客户数据集进行细分，可能导致传播定向效果不佳；E_7（数据科学家）也从机器学习的角度确认了这一担忧。

关于稀疏性，E_3（数据工程师）根据经验分享说，缺失数据可能是“空字符串（“”）、仅含 *ASCII* 空格的字符串（“ ”）、空值字符串（“*null*”）、缺失字符串（），这与 *[显式空]* 值 *[不同]*；具体取决于数据的来源，这些情况可能会 *[扰乱]*你的细分逻辑。”E_2（数据工程师）指出，“糟糕的”筛选策略会增加计算成本（例如，如果第一次筛选将搜索集缩小了 95%，用户往往会通过运行新的查询来撤销该操作，而这会带来高昂的成本）。相反，E_2 建议预先展示（质量）洞察，因为这可能“激发用户的好奇心和关切”，并帮助发现“糟糕的”筛选条件。另一个常见挑战是，从大型数据集中选择“重要/最佳/有效/相关”的属性和记录，以便准备有效的仪表板（E_4）、训练准确且公平的机器学习模型（E_9），以及制定细分规则，以开展成功的营销传播（E_{11}）。

许多领域专家指出，数据准备过程中的协作可能会很繁琐，因此主张使用更好的工具来支持这一过程。E_{11}（营销人员）提到，向数据工程师传达与数据相关的需求往往是一个繁琐的过程，需要多次迭代、耗费时间，而且交流过程中也可能丢失关键信息，因此主张提供一种交互式自助工具。E_1（数据工程师）建议提供一种类似 LinkedIn [80] 个人资料完整度的可视化报告卡，不过是针对数据集的，因为它“还可以让负责执行这些任务的用户以外的决策者了解数据质量如何，从而促使他们落实数据政策。”有人建议，根据组织内部数据的使用情况来呈现洞察。E_3（数据工程师）起初叹道：“即使摄取的数据质量可能很低，用户有时也并不在意；他们实际上并没有使用这些数据。我们只能接受这一点。”不过，E_3 也指出：“数据的使用情况 *[usage]* 在这里可能非常重要，因为它能带来另一种认知。”并从商业角度界定了两种可能有益的使用维度：“用于客户细分的‘布尔逻辑’*[筛选条件]*（例如，‘年龄’ > *18* 是定位成年人的常用条件），以及‘投影’*[引用属性]*,，例如，在营销活动中，‘姓名’和‘电子邮件’属性经常用于定位用户。”E_3 还指出：“如果我正在定义细分群体并引用一些可能质量不佳 *[或]* 很糟糕的字段，而且空值比预期更多，我希望能知道。所以我很喜欢这里关于使用情况的这一方面。”

E_9（数据科学家）建议，数据溯源（例如，“数据集最近一次使用的时间、地点和使用者”）可用于协助数据维护，因为“有几个无人使用、质量低劣的

数据集就这么闲置着，或许可以将它们归档。”E_4（数据分析师）建议将这一想法加以推广，整理出“一个关键绩效指标（*KPI*）目录，列出不同的指标、计算方法，以及它们在仪表板、细分群体和用户旅程中的使用频率，并将这些评分放在一旁。”E_2（数据工程师）总结道，“要把数据整理成高质量的形式需要付出辛苦，因此任何形式的重复利用都是好事，无论是重复利用输出结果，还是重复利用获得该结果的工作流。”受到这些专家认可的启发，我们确认，向用户介绍数据的质量和使用特征，有望帮助他们更好地完成任务。

3.2 设计需求练习

在了解用户任务和挑战后，我们分别与 E_2（数据工程师）和 $E_{4,\ 5}$（数据分析师）进行了两次后续讨论，探讨如何构建 DataPilot，以及如何从访问（身份验证和伦理考量）、处理（可扩展计算策略）和持久化（最佳存储机制）等方面处理数据。随后，我们与 E_{14}（用户界面工程师）进行了一次后续讨论，并开展了设计练习：我们以数字形式和纸面形式绘制低保真设计草图，并展示这些草图以征求反馈。这些草图涵盖了用户界面中的可视化、小组件、布局、工作流和交互，旨在及时发现日后可能出现的错误。我们集思广益，讨论每种设计的优缺点，并据此进行了多项修改和完善。对于部分设计，我们开发了快速软件原型，旨在探索潜在技术（如软件库）并评估其可行性。这也进一步帮助我们淘汰了实用性较低的设计，完善了需要支持的用户任务，并提炼出设计目标，详见下文。

3.3 设计目标

我们从专家访谈中提炼出**六项**关键设计目标，指导 DataPilot 的设计与开发。

DG1. 支持在原位进行数据准备和可视化数据分析。数据准备是分析之前必不可少的步骤。然而，即使在分析过程中，用户也常常需要返回数据准备步骤。我们提出这一核心设计目标，旨在通过同一工具支持这两个环节（原位完成），减少不必要的学习成本，以及在多个工具和窗口之间切换的情况。具体而言，DataPilot 支持用户在浏览大型且陌生的表格数据集后，从中选取子集并据此构建可视化仪表板。

DG2. 将数据质量和使用信息建模为标准化评分。由于非技术型营销人员（E_{11}）常常不得不依赖数据工程师（E_1），我们提出这一目标，利用智能且基于启发式方法确定的规则，将质量和使用信息的各个维度建模为标准化评分（满分 100 分）。这一评分策略能够对不同维度进行比较和聚合，为用户带来“自助服务”体验，减少他们对其他用户的过度依赖及繁琐沟通。高级用户仍可自行交互式地指定约束，从而获得一定的配置控制权。

DG3. 在数据准备过程中提供有关数据质量和使用情况的视觉引导，同时平衡用户自主权与控制。这一目标意味着向用户提供有关其数据质量和使用特征的引导。为了按照领域专家的期望，在用户自主权与控制之间取得平衡，我们提供干扰最小的“方向性”引导 [16],，例如提供视觉提示（如突出显示缺失值、显示计算得分），而不打断用户的分析。

DG4. 为数据发现、子集选择和可视化仪表板创建提供交互功能和规范配置功能。这一目标的一项关键创新，是提供“自助式”交互功能（例如排序和筛选界面控件），帮助用户检查原始数据集及选定子集中的数据属性和记录的质量和使用信息；并提供规范配置功能，将数据属性分配给视觉编码（例如 X 轴、Y 轴），从而构建可视化仪表板。

DG5. 通过可配置性实现控制和上下文。E_3（数据工程师）曾指出，质量和使用信息可能并非总是可用或适用（例如，新上传的数据集尚无使用信息）。因此，这一目标意味着在 DataPilot 用户界面中提供选项，用于配置与质量和使用相关的不同组件的可见性（即显示其中一项、均不显示或两项都显示），从而使用户和应用程序能够获得多个层级的控制和上下文。

DG6. 面向可扩展性和性能进行设计。E_{14}（用户界面工程师）曾再次强调，在用户界面上呈现大量数据所面临的挑战（例如，加载时间过长、交互体验迟缓）。据此，我们确立了这一目标：设计高性能的前端应用程序，将复杂操作卸载至可扩展的后端服务器，以全面提升用户体验的流畅度 [30]。

4 数据质量与使用情况建模

基于第 3 节所述的设计研究，我们现在讨论如何对质量和使用信息进行建模。质量和使用信息的建模策略取决于数据类型、用户和应用场景。在本研究中，我们聚焦于一种仪表板应用：用户首先上传表格数据集，准备一个相关子集（通过选择相关属性并筛除无关记录），然后利用该子集创建可视化，构成仪表板。用户定义的约束以及与图形用户界面元素的交互（例如属性级选择复选框、用于记录级筛选条件的范围滑块）用于对质量和使用信息进行建模、量化，并与之交互。

4.1 质量

基于领域专家提出的关于数据偏斜（E_1，E_7，E_{11}）、稀疏性（E_1，E_2，E_3，E_{11}）和不正确性（E_2）的现有挑战，以及先前研究 [89],，我们对属性级的三个质量维度进行了建模：完整性、正确性和客观性，以及记录级的两个维度：完整性、正确性（**DG2**）。

4.1.1 属性级质量维度。

完整性是某个属性的值中非缺失值所占的百分比。例如，如果 50 个属性值中有 10 个是空值或空字符串，则其完整性为 100*(50-10)/50 = 80%。完整性有助于用户检测稀疏属性；例如，稀疏属性可能会影响机器学习算法做出准确预测的能力。

正确性 是某个属性的值中正确值 所占的百分比。例如，如果 50 个属性值中有 5 个不正确，那么其正确性为 100*(50-5)/50 = 90%。要计算正确性，

企业可以在 DataPilot 源代码中，通过关系运算符（>，<，=）、范围运算符（BETWEEN）、模式匹配（LIKE）和成员运算符（IN）预先配置类 SQL 约束；例如，“`SELECTCount(*) WHERE` 电子邮件 `NOT LIKE '%_@__%.__%'`”会计算电子邮件地址不正确的记录数。借助正确性，用户可以评估单个属性的准确性。

客观性是指值符合目标分布的程度。例如，若性别属性中有 120 名男性和 45 名女性，则显然男性占比偏高，因此从性别平等的角度来看，该数据并不客观。我们采用 Wall 等人提出的属性分布（AD）指标 [118] 来衡量观察到的分布与预期客观分布（基线）之间的偏差；AD 分数的范围为 [0,1]，因此我们将其乘以 100 进行标准化。借助这一维度，用户可以发现异常现象，例如，当大多数申请人的性别与预期不符时。与正确性一样，企业可以在 DataPilot 源代码中预先配置客观性约束。

4.1.2 记录级质量维度。

完整性是每条数据集记录中非缺失值所占的百分比。例如，一条记录有 50 个值（每个属性对应一个值），其中 20 个为空值或空字符串，则其完整性为 100*(50-20)/50 = 60%。借助这一维度，用户可以在营销活动中剔除稀疏的客户档案（记录）；在这类活动中，成功度取决于档案的丰富程度。

正确性是每条记录中正确值所占的百分比。例如，如果一条记录有 50 个属性值，其中 15 个不正确（根据集合约束判断），则其正确性为 100*(50-15)/50 = 70%。通过这一维度，营销人员可以舍弃电子邮件地址无效的客户档案（记录），以及无法用于开展营销活动的社交媒体账号。

客观性不适用于记录级维度，因为每条记录都包含来自不同且无法比较的属性的值。

4.1.3 总体得分：聚合与自定义。

我们会为每个属性和每条记录计算可配置的、基于启发式方法的总体得分；默认情况下，该得分取相应维度的算术平均值。根据 Vaziri 等 [115], 的研究，用户可以为不同维度指定不同权重（例如，用户可能更倾向于将总体维度设置为完整性占 75%、正确性占 25%，并忽略客观性），也可以为不同属性和记录指定不同权重（例如，数字营销人员可能希望在正确性方面赋予“电话”属性的权重高于“电子邮件地址”属性）。

4.2 使用情况

根据领域专家的积极反馈，我们在属性级的三个维度上对使用信息 **（DG2）**进行了建模：子集中的使用、筛选器中和可视化中；并在记录级的一个维度上进行了建模：子集中的使用。

4.2.1 属性级使用维度。

子集中的使用 分数是指选择某个属性并将其纳入子集以供日后使用的用户所占百分比。例如，如果 20 位用户中有 15 位选择某个特征来训练机器学习模型，那么该属性的子集中的使用 分数为 100*15/20 = 75%。借助这一维度，新用户例如，可以根据先前用户的子集，选择使用频率较高（或较重要？）的属性，从而快速高效地进行分析。

筛选器中得分表示对该属性应用筛选器的用户百分比，例如选择多选下拉菜单中的选项（性别=“女性”）或拖动范围滑块的滑块（年龄 ∈ *[40,50]*）。借助这一维度，数字营销人员可以根据以往的营销活动确定细分规则（用于筛选特定客户档案的筛选条件），从而开展营销活动。需要注意的是，筛选器中并不是子集中的使用的子集；用户可以按某个属性进行筛选（或不筛选），并在其子集中选择（或不选择）该属性，反之亦然。

可视化中得分表示将该属性分配给一个或多个视觉编码（例如 X 轴），并在仪表板中使用所得可视化的用户百分比。借助这一维度，用户可以参考过去业务报告中常用的（重要的？）属性，以协助设计当前的业务报告。

4.2.2 记录级使用维度。

子集中的使用得分表示选择将某条记录纳入其子集（由筛选条件产生）的用户所占的比例。借助此维度，用户可以选择一个由热门（重要？）记录构成的子集，并根据以往成功的营销活动中的客户档案（记录）重新开展营销活动。该维度本质上与记录级筛选器中和可视化中的使用维度相同，因为 DataPilot 会将筛选后的数据集视为选定子集，并用于可视化。

4.2.3 总体得分：聚合与自定义。

与总体质量类似，我们基于启发式方法计算每个属性和记录的总体得分，取各组成维度中的最大值。由于属性很少会同时用于子集、筛选器和可视化中，采用平均值会导致得分偏低，无法有效激励用户；因此，我们选择了最大值。如果某个或多个使用维度与用户的用例无关，用户可以将其忽略，例如筛选器中的使用情况。

5 DATAPILOT 用户界面

为了在同一工具（**DG1**）中支持子集选择与分析，我们将 DataPilot 用户界面设计为三步工作流，各步骤都可通过左上角的导航入口相互切换（图 1 和图 2）。我们根据四位用户参与的试点研究敲定了这一设计；第 5.5 节讨论了一些备选但最终弃用的设计。

5.1 步骤 1：查看原始数据

此步骤也是 DataPilot 的首页，可让用户分析数据集并选择相关子集（图 1）。它由以下视图组成：

（A）属性视图以扁平列表（ ）或嵌套列表（ ）的形式显示所有属性，后者适用于层级数据集。为了高效显示大量属性，我们采用虚拟滚动原理，避免渲染视口中不可见的对象（**DG6**）。搜索字段支持通过基于关键词的查询快速查找属性。用户还可以按属性级的质量和使用维度进行排序。每个列表项都会显示属性的名称（例如“sales.product.name”）、数据类型（例如：类别型、：数值型）以及一个双色圆形字形（**DG3**），

A #

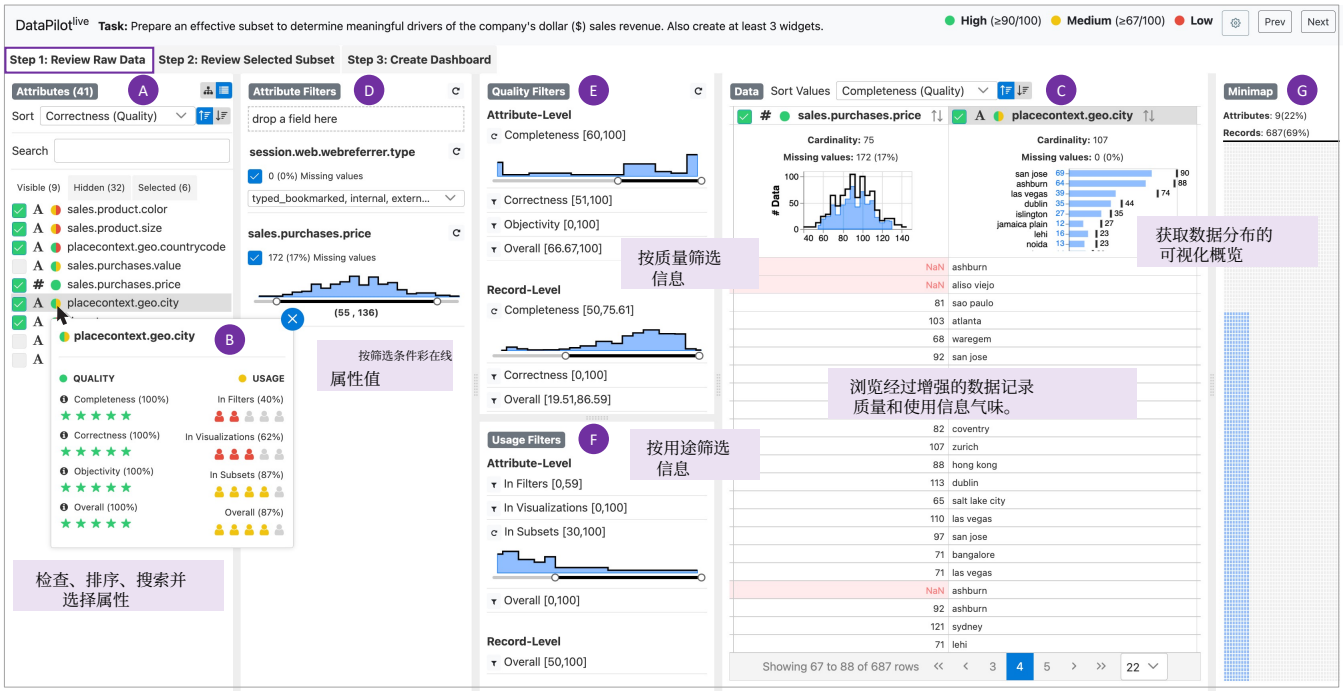

图 1：DataPilot用户界面展示了三步工作流中的第 1 步（查看原始数据）。用户可以检查数据集属性列表（A. 属性视图），查看某个属性的质量和用途维度评分（B. 属性详情视图），可视化属性分布并浏览数据集记录（C. 数据视图），根据属性值逐步筛选记录（D. 属性筛选视图），根据质量（E. 质量筛选器视图）和使用维度（F. 使用情况筛选器视图）逐步筛选属性和记录，以缩小搜索空间；查看筛选后数据集的可视化摘要（G. 缩略图视图）；并为所需子集明确选择属性（A. 属性视图）和记录（根据筛选条件自动选择）。

例如(由绿色、黄色、红色组合而成)，其中左半部分显示总体质量分数，右半部分显示总体使用情况分数。请注意，如果上传的数据集仅包含质量或使用信息中的一种，这些双色字形会自动变为单色字形；用户也可以在右上角的设置中手动配置（**DG5**）。高（≥90）、中（≥67，但 <90）、低阈值（用于确定这三个类别）以及对应的颜色（以满足色觉无障碍需求），均可在右上角的图例中配置。每个复选框都允许用户在子集中选中或取消选中属性（**DG4**）。将鼠标悬停在属性名称上时，工具提示会显示其描述。点击双色字形可打开**属性详情视图**。

(B) 属性详情视图 是一个叠加层，展示属性质量和使用情况的详细信息，例如 LinkedIn 的 [80] 个人资料完整度（**DG3**）。与双色字形类似，左列显示数据质量维度，右列显示使用维度，并通过 5 分制图标阵列评分量表可视化分数。例如，“place- context.geo.city”的完整度评分为 100%（），总体使用情况评分为 87%（）。将鼠标悬停在信息图标上，可查看该维度的定义（例如，“完整性是属性中非缺失值所占的百分比（%）”）以及计算所依据的任何预配置规则（例如，“*sales.purchases.price* 被视为正确，当且仅当其为 ≥ 0”），以帮助用户了解相关信息（**DG2**）。

(C) 数据视图 通过交互式表格展示整个数据集。第一行汇总显示属性特征，例如基数（唯一值的数量）、缺失值和分布图（数值型属性使用面积图，分类属性使用条形图；图中以黑色显示底层数据分布，以蓝色显示筛选后的数据分布）（**DG3**）。包含缺失值或错误值的表格单元格（例如，“*sales.purchases.price*”=“NaN”）会以红色突出显示，悬停时可查看详情（**DG3**）。搜索、分页和排序等常规操作均集成在表格控件中。用户还可以在记录级别按质量和使用维度排序（**DG4**）。在图 1 中，记录按完整性 排序（在**数据视图**的“排序值”下拉菜单中设置），列按正确性 排序（在**属性视图**的“排序”下拉菜单中设置），且二者均按升序排列。

(D) 属性筛选视图 使用户能够通过将属性从**属性视图** 或**数据视图**拖入此视图的放置区（**DG4**），为每个属性应用筛选条件，从而筛选数据集。针对类别型属性的多选下拉菜单、针对数值型属性的范围滑块，以及用于展示原始数据集（黑色）和应用筛选条件后（蓝色）属性值分布的视觉气味（嵌入式可视化，为用户在信息空间中导航提供信息气味线索 [122]），有助于用户确定合适的筛选条件（**DG3**）。与选择属性不同，选择属性时必须显式勾选复选框



图 2：DataPilot 第 2 步（审查选定子集）和第 3 步（创建仪表板）。用户审查选定的属性（H. 属性视图）和记录（I. 数据视图），将属性（J. 属性视图）分配给编码（K. 编码视图），检查生成的可视化（L. 可视化画布），并将其保存到仪表板（M. 已保存的可视化）。用户可以在这三个步骤之间自由切换。

以将其添加到子集，而 DataPilot 会在筛选后自动将所有剩余记录选入子集。

（E）质量筛选器视图使用户能够按属性级和记录级别的质量维度筛选数据集（**DG4**）。例如，应用属性级完整性筛选器 $\in [60, 100]$会移除完整度评分超出范围的数据属性（列）。同样，记录级完整性筛选器 $\in [50, 75.61]$ 会筛除超出该范围的所有记录（行）。

（F）使用情况筛选器视图与**质量筛选器视图**类似，允许用户根据使用维度（**DG4**）筛选数据集。例如，应用属性级子集中的使用筛选条件 $\in [30, 100]$ 会移除所有被少于30%的用户选中的属性。

（G）缩略图视图提供了一种新颖的可视化概览，展示数据集中原有属性和记录所占的比例（灰色），应用筛选器后当前可见的部分（蓝色），以及在数据集子集中选中的部分（绿色）（**DG4**）。默认情况下，我们禁用了绿色（选中）状态，因为试点用户觉得它会造成视觉干扰（第 5.5 节）。矩形区域的宽度和高度分别表示属性和记录的数量。此视图被划分为许多小矩形，其大小与数据集大小成比例。

5.2 步骤2：查看选定的子集

此检查步骤由**（H）属性视图**和**（I）数据视图**组成，仅显示所选的属性和记录（图 2✅。将所有选定属性堆叠在一起进行查看，用户便可以检查高、中、低质量分数和使用情况得分的相对分布；此视图还便于检查所选表格单元格中红色高亮（缺失或错误值）的分布；如果存在未选中的属性，这两项任务将很难完成

步骤 1要求用户在未选中的属性存在的情况下暂停并反思其子集选择性能，然后再继续构建仪表板（**DG1**）。

5.3 步骤 3：创建仪表板

查看所选子集后，此步骤可帮助用户创建并保存单变量和双变量可视化，最终组成一个仪表板（图 2）（**DG1**）。此步骤包括：

(J) 属性视图与**步骤 2**中的**属性视图**相同。

(K) 编码视图允许用户通过指定图表类型（条形图、散点图、折线图）、将属性拖至视觉编码（X、Y），以及在适用时确定聚合方式（总和、平均值、最大值、最小值）来创建可视化（**DG4**）。

(L) 可视化画布 根据**编码视图**中配置的规范渲染可视化。用户可以为可视化添加标题并点击保存图标来保存它。

(M) 已保存可视化视图显示从**可视化画布**保存的所有可视化列表。此视图还允许用户根据需要删除单个或全部已保存的可视化（**DG4**）。

5.4 实现

我们使用 Angular [43], 开发了 DataPilot 前端，该前端通过 HTTP REST [97] 和 WebSocket [34] 协议与 Python [37] 服务器实时交互。数据集、用户交互日志（从前端收集）和辅助信息均存储在 PostgreSQL 中，之后使用 SQL 查询（**DG6**）。

5.5 设计备选方案

在敲定 DataPilot 的设计之前，我们向四位试点用户展示了界面的初始版本，以评估某些设计的可行性，以及为用户研究（第 6 节）规划的评估任务的保真度。接下来将介绍一些未纳入当前版本的设计考量。

在确定采用属性名称旁的双色字形之前，我们尝试过其他视觉变量，例如大小（如更大的圆圈表示更高的分数）和形状（如质量用方形表示，使用情况用圆形表示）。为了满足 **DG5**（配置 DataPilot，以支持质量和使用信息中的一项、两项或均不支持），我们没有选择这些方案；双色字形更美观，因为它们在使用不同颜色描述不同维度的同时，保留了统一的圆形形状，适用于不同配置。接下来，我们选择离散的三档（高、中、低）量表，而非连续量表，以使用户通过颜色色相，而不是效果较差的饱和度，更直观地区分属性并将其归类 [67]。对于属性详情视图中的五档评级量表，我们考虑过类似进度条的连续控件，用长度编码数值大小，但最终选择了离散图标阵列，因为它们易于阅读 [41]。要将记录选入子集，我们考虑过是否应像属性一样，通过复选框手动选择；然而，这种逐条选择被认为十分繁琐，因此没有采用。最后，为了兼顾数据准备和分析，我们考虑了若干与工作流相关的问题，例如：应该设置多少个步骤？这些步骤是否应该按线性顺序进行？审核步骤是否必要？试点用户帮助我们最终确定了灵活的线性三步工作流。

5.6 DataPilot 示例场景

为了展示 DataPilot 如何帮助用户从大型且陌生的数据集中准备出相关子集，我们围绕两位假想用户——Sunny（数据工程师）和 Kiran（数据分析师）——设计了两个使用场景；这些场景是在与领域专家协作的过程中开发的，以确保其与领域相关（第 3 节）。

5.7 案例1：专家用户，性能提升

Sunny 是一位经验丰富的数据工程师，经常为分析师准备数据子集，供他们随后制作业务报告。他们打开 DataPilot，上传一批近期的电子商务应用客户交易数据，并开始分析。凭借其领域专长，他们通过搜索字段快速查找已知属性，并为子集选择了五个 属性：“sales.product.name”、“sales.purchase.price”（以美元计）、“timestamp”（购买时间）、“placecontext.geo.countrycode”（例如，印度的“IN”）以及“environment.operatingsystem”（例如，“iOS”）。

他们切换到**第2步：审查选定子集**，发现数据表中的几个单元格（此时表中仅显示五个选定属性）背景为红色。具体来说，“placecontext.geo.countrycode”列中值为“*AA*”的单元格被标出（），而“environment.operatingsystem”列中为空（缺失值）的单元格被标出（）。他们意识到，没有任何国家使用“*AA*”作为国家代码（根据 DataPilot 的正确性 约束以及他们自己的知识），并且操作系统值中有相当一部分（1000 个值中有 706 个）缺失，于是返回**第1步：查看原始数据** 进行修改。

他们将“placecontext.geo.countrycode”属性从**属性视图**拖入筛选面板，以移除所有值为“*AA*”的记录（），并单独向数据收集团队提醒该问题。为绝对确保其数据在所有属性上均正确，他们应用了记录级“正确性”筛选器（），仅保留 100% 正确的记录。最后，他们从子集中取消选择“environment.operatingsystem”，转而选择了另一个包含类似信息的属性“environment.browserdetails.useragent”，例如‘*Mozilla/5.0 (iPhone; CPU OS 12_0 like Mac OS X; en_US)*’。尽管该属性此前并不常用（右半部分为红色），但其总体质量很高（左半部分为绿色）。通过这种方式，在整个工作会话中，DataPilot 帮助 Sunny 识别其数据中的问题，并引导他们准备出更完整、更正确的子集。

5.8 案例2：新用户，有效入职

Kiran 最近加入了一家数据分析公司，负责熟悉客户的数据，以便设计未来的仪表板。他们将一个包含电子商务交易的客户数据集上传到 DataPilot，并开始分析。该数据集规模庞大且十分陌生。他们开始从**属性视图**中检查属性名称和描述，以及**数据视图** 中的相应值和分布图（）。由于数据量过于庞大，为了加快入职熟悉过程，他们采用了策略，只针对重要属性。

他们尝试在**质量筛选器视图**和**使用情况筛选器视图**中应用属性级筛选器，以此作为重要性的代理指标，从而缩小属性搜索空间。具体而言，他们检查相应范围滑块上的分布，并筛除具有总体质量分数 < 75 （）和总体使用情况分数 < 25 （）的属性，将属性数量减少到易于管理的 17 个。最后，他们按照总体质量分数降序排列这些属性，并开始在**属性详情视图**中检查它们的姓名、描述和分数（通过双色圆形字形）。通过这种方式，DataPilot 帮助 Kiran 快速而有效地熟悉一个全新且陌生的数据集。

6 评估

我们开展了一项用户研究，以评估并了解 DataPilot 中的质量和使用信息如何指导用户为后续分析准备有效的数据子集。

任务： 我们设计了一项包含子集选择和可视化分析的任务，要求参与者：

> “探索一个关于电子商务网站在线客户行为的数据集，准备一个有效子集[1]，以确定对公司销售收入（美元）有意义的驱动因素，并创建一个包含至少三种可视化的仪表板来展示他们的发现。”

参与者： 我们招募了36名参与者，他们由来自业界和学术界的专业人士与研究人员组成：学生 （23名）、商业顾问 （2名）、资深数据分析师 （2名）、助理教授、副产品经理、数据科学经理、博士后学者、项目经理、质量保证工程师、科学家（临床试验）、软件开发人员和用户体验设计师。参与者正在攻读或已经获得以下学位：学士（3名）、硕士（14名）或博士（19名），专业领域包括计算机科学（21名）、以人为中心的计算（4名）、人机交互（2名）、工商管理（3名）、制药科学、经济学、电子工程、系统工程、数据科学或信息学。在人口统计方面，他们的年龄（岁）分布为*18—24*岁 （13名）、*25—34*岁 （19名）、*35—44*岁 （3名）或不愿透露 （1名）；性别为女性（16名）、男性（19名）、其他（0名）或不愿透露（1名）。参与者自我报告称，他们使用可视分析工具（例如 Excel、Tableau）或编程进行任何形式数据分析的经验为：每天或工作的一部分（10名）、经常（13名）、偶尔（13名）、很少（0名）或从不（0名）。

数据集： 为全面评估 DataPilot 的所有功能，并确保在规定的研究时长内完成任务，我们从一个开源数字营销数据集中随机抽取了 1000 条记录和 42 个属性（列） [68]，并通过设置适当的约束，引入了与正确性 和客观性 相关的若干质量问题。我们对质量、使用情况和总体得分进行了设定，使得 ≥ 90 被标记为高， ≥ 67 而 < 90 被标记为中，其余则被标记为低。我们固定了这些阈值，以便在三个类别（高、中、低）之间合理分配属性和记录，从而避免参与者士气低落（所有得分都很低）或过度自信（所有得分都很高）。

[1]请注意，数据子集所包含的属性和记录数量均不超过原始数据集中的相应数量。

作为用户研究条件的系统配置： 为实现**DG5**，我们设计 DataPilot 支持四种配置：（1）既不使用质量信息，也不使用使用信息；（2）仅使用质量信息；（3）仅使用使用信息；（4）同时使用质量信息和使用信息。在这四种配置中，我们没有明确评估（3）仅使用使用信息的配置，因为专家访谈强调，解决数据质量问题最为重要，而且仅凭使用信息绝不能驱动“数据驱动”的分析和决策，至少不能在缺少质量等更重要因素的情况下这样做。因此，我们将 DataPilot 的另外三种配置作为独立的研究条件，用于接下来所述的被试间评估。

[B] 基线：通过此配置，我们旨在了解用户在没有数据质量与使用信息的情况下所采用的策略，同时模拟许多当前系统（例如 Tableau [106]）的做法。具体而言，属性名称旁的双色字形、筛选和排序选项，以及表格中用于表示使用情况和质量的视觉气味均被隐藏。

[Q] 质量：通过此配置，我们旨在了解用户如何仅利用质量信息完成研究任务，同时模拟许多当前系统（例如 Profiler [62], Trifacta [110]）的做法。此条件还使我们能够与以下*D*配置（同时包含数据质量与使用信息）进行比较。具体而言，属性名称旁仅显示并启用与质量相关的单色圆形字形、排序和筛选选项，以及数据表中的视觉气味。

[D] DataPilot：此综合配置在界面中同时显示数据质量与使用信息。具体而言，第 5 节中所述的所有功能均已启用。D 条件的使用信息是通过处理参与者在 *B* 和 *Q* 条件下的交互日志（24 名参与者）计算得出的。我们将每个属性的子集中的使用得分计算为选择该属性纳入其子集的参与者所占百分比，筛选器中的使用得分计算为按该属性进行筛选的参与者所占百分比，可视化中的使用得分计算为将该属性分配给某种视觉编码的参与者所占百分比，并将总体得分设为上述三个得分中的最大值。同样，对于每条记录，我们通过计算将该记录（因应用筛选条件而自动）纳入其子集的参与者所占百分比，得出其子集中的使用得分（此处也即总体得分）。为排除分析过程中临时、非计划和意外的选择，我们仅依据任务结束时界面的最终状态（选定子集、应用的筛选条件、已保存的可视化）计算这些信息。

研究阶段： 我们将参与者分配到三种研究条件（*B*、*Q*、*D*）之一，并尽量平衡他们的背景、人口统计特征以及视觉数据分析素养。每个研究阶段持续 60 至 90 分钟；由于参与者的训练和练习时间不同，*D*所需时间比*Q*长，而*Q*又比 *B* 长。我们为每位参与者提供了一张价值 15 美元的礼品卡，以补偿他们投入的时间。我们通过 Microsoft Teams [82] 远程开展研究；实验人员共享自己的电脑屏幕，并授予参与者输入控制权，从而让参与者访问研究环境。参与者同意参与后，观看了一段视频教程（*B*：5 分钟，*Q*：7 分钟，*D*：10 分钟），视频演示了

DataPilot[2] 的特征。随后，参与者在房屋数据集 上完成了一项练习任务（改编自 [23]），以便在开始实际任务前熟悉用户界面。

实际任务最长持续30分钟。为模拟真实的工作环境，参与者无需在任务期间进行出声思考（尽管部分参与者对此感到自在）。任务期间，参与者与系统的交互（例如，他们应用的筛选器、选择的数据子集）均被记录。研究以参与者填写问卷结束，问卷用于评价 DataPilot 特征的实用性；随后进行了为期10分钟的半结构化总结访谈，参与者在访谈中反思了整体经验、提供了反馈并回答了其他问题。在总结访谈结束时，实验人员还向 D 配置的参与者和B及Q 配置的参与者进行了演示，以了解他们的初步反应，并征求反馈，了解这组新辅助工具在假设情况下将如何帮助他们以不同方式完成任务。每次总结访谈均进行了屏幕和音频录制，以便后续开展定性分析。

6.1 假设

我们根据以下假设来组织研究分析。这些假设是在研究开始前预先确定的，依据包括我们对工具预期用途的设想、以往的感知研究、试点研究的反馈以及我们自己的直觉。 $>$ 表示多于或大于； $<$ 表示少于或小于。

H1 B（基线） $>$ Q（质量） $>$ D（DataPilot），就选定子集中属性和记录的数量而言。H2 B $>$ Q $>$ D，就选定子集中低质量和低使用情况的属性和记录所占比例而言。H3 B $<$ Q $<$ D，就选定子集中高质量和高使用情况的属性和记录所占比例而言。H4 B $<$ Q $<$ D，就任务完成后的成功度和信心而言。H5 B $<$ Q $<$ D，就执行任务时所需的努力程度、时间需求、心理需求和挫败感而言。H6 参与者在执行任务时会认为质量信息比使用信息具有更高的效用。

6.2 结果

下面我们将介绍用户研究的发现，并结合参与者的定性反馈进行讨论。$B_{1,...,12}$，$Q_{1,...,12}$，$D_{1,...,12}$ 分别指 基线（B）、质量（Q）和 DataPilot（D）条件下的 36 名参与者。事后访谈中说出的参与者引语以及问卷中的书面回答，均采用亲和图法进行编码和分类 [40], 归纳式主题分析 [13] 技术。一名实验人员提出了初步的类别集合，随后与另外三名实验人员经过多轮迭代对其进行完善，直至达成共识；最终的编码手册由 6 个高层类别和 43 个详细的低层代码组成。相关研究材料包括用户的交互日志、问卷、访谈记录、分配的定性代码、数据分析脚本，以及使用 RainCloudPlots 绘制的相关图表 [5] （而非箱线图），均已收录于补充材料中。

[2]D 参与者同时看到了质量信息和使用信息；Q 只看到了质量信息；而C 两者都没有看到；因此，他们观看相应视频教程所花的时间有所不同。

6.2.1 对 *DataPilot* 质量信息和使用信息的反馈。

DataPilot，即该系统。总体而言，参与者认为 DataPilot 很有用；在三个条件下，系统可用性（SUS [15]）得分均高于平均水平：{B：80.21，Q：74.17，D：71.67}。 D_4评论道：“提供详细的辅助信息，例如每个数据属性的质量和使用情况，这非常重要，而 *Tableau* 和 *PowerBI* 等现有工具却缺少这些信息。” Q_8 还解释了质量和使用信息为何重要，并指出：“真正的数据分析、数据科学和机器学习中，*80%* 至 *90%* 的工作都在 *[*数据准备*]*这一步。这些 *[*质量和使用情况*]*衡量指标能让用户开始 *[*处理任务*]*，并促使他们探索一些意料之外的后果，这非常有价值。它为未来的持续探索提供了充足机会，让产品不断改进，因此这确实非常、非常令人着迷。”

质量信息。 参与者对质量信息总体给予了积极反馈。 Q_{10} 评论道：“数据中存在一些看不见的问题，只有开始尝试使用可视化时，你才不一定能发现它们。 *[*此外，*]* 在聚合可视化中，你识别质量问题的能力有限，甚至完全无法识别，因此我很欣赏 *DataPilot* 对这些质量问题的明确提示。” Q_7 指出：“系统提供这类开箱即用的洞见很重要，这样像我这样不会编写代码的用户就不会完全忽略这些方面，而是可以依靠绿色属性，直接开始分析。” Q_8认为“这些质量测量信号很有价值，能帮助用户更快地筛选 *[*属性和记录*]*，而不必自己去学习 *[*这些信息*]*。”然而， Q_8 也对“混杂因素表示谨慎，尤其是缺失数据，因为很多时候数据并非随机缺失，而是确实缺失并传递着某种信息”，这表明质量信息可以提供良好的起点，但用户可能仍需开展进一步分析。

使用信息。对于使用信息，参与者的反馈不一。给予积极反馈的参与者建议利用使用信息进行快速、高效的分析（Q_6），通过“进行一些小调查”（Q_1）来寻求验证，“检查自己是否与其他人持有相似的观点”（D_4），“发现其他人没有关注的新事物”（D_3），向前人寻求引导（例如，$Q_{2,\ 11}$），避免重蹈覆辙（D_3），以及在相互冲突的选择之间做出决定（例如，“有些属性不容易决定……但使用情况可以帮助选择”- D_8）。给予混合或负面反馈的参与者表示，他们不会在意（D_3），也不会依赖其他人的做法，因为他们不了解其他用户，不得不假设这些用户的分析做得很好（Q_1, D_{10}）。参与者还表达了对偏差和盲目从众的担忧，因为“人们可能会错过某个不常见但同样有用的属性”（B_7）。

6.2.2 比较准备好的子集。 表1和图3展示了参与者所选择的子集规模（从42个属性和1000条记录中选出的总属性数和记录数）（图3a），以及 高、中、低 属性级和记录级质量及使用信息的分布（图3b）。验证 **H1**时，D 选择的属性和记录最少，其次是Q，再其次是B。

此外，D选择了更高比例的总体质量较高的属性，相较于Q和B。由于数据集较为稀疏（每条记录中的大多数值均为空），没有任何记录具有较高的总体

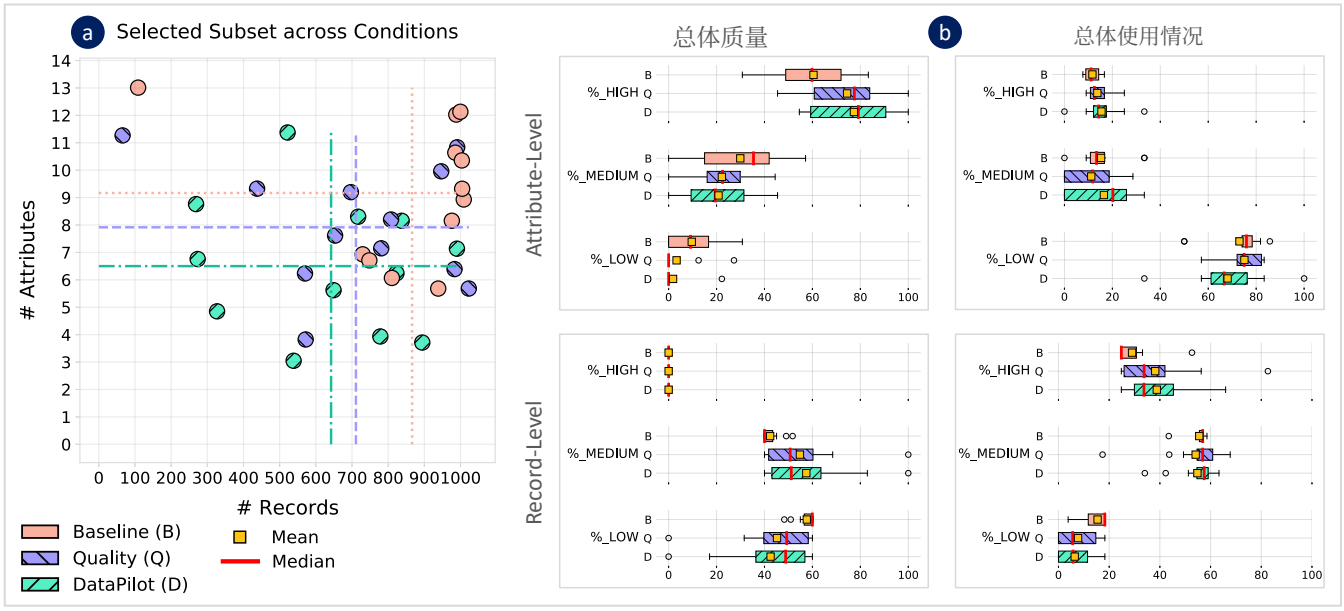

图3：（a）参与者选定子集中的属性和记录数量；（b）在三种研究条件下，质量和使用情况的高、中、低总体得分的属性级和记录级分布（B、Q、D）。

质量分数，因此相应的 μ_R， σ_R 值对于B、Q、D 均为 0。D 还选择了更高比例的总体使用情况较高的属性和记录，相较于Q 和B。这些结果验证了**H3**。

同样，D 选择的总体质量较低的属性和记录比例低于 Q，也低于 B。此外，D 选择的总体使用情况较低的属性和记录比例低于 Q 和 B，验证了 **H2**。这些发现表明，质量和使用情况信息促使用户准备更小、更有效的子集。

6.2.3 任务保真度得分。图4展示了参与者在七点制"不同意"（1）至"同意"（7）量表上对任务保真度的反馈。与Q和B相比，D报告了更高或相当的心理需求（M_D=5；M_Q=5；M_B=4.5；M=中位数）、辛苦程度（M_D=5；M_Q=4；M_B=4）和挫败感（M_D=2.5；M_Q=2.5；M_B=2），为支持H5提供了一些证据。我们将这一结果归因于D中额外用户界面元素所带来的复杂性增加，这可能影响了用户的认知负荷。然而，D最终报告了更高的成功度（M_D=6；M_Q=5.5；M_B=5）和信心（M_D=5.5；M_Q=5；M_B=4.5），验证了H4，并表明辅助信息帮助参与者更有效地完成了任务。

6.2.4 一般信息、质量信息和使用情况信息的重要性。我们询问了参与者，界面中不同类型的一般信息、质量信息和使用情况信息的重要性，采用从完全不重要 （1）到非常重要 （7）的量表。除属性数据类型外，其他一般信息，如属性名称、值、分布、基数和描述，大多被认为是有用的（图5a）。

图 5b、5c 显示，总体而言，Q和 D参与者都认为质量信息很有用（M_D=5；M_Q=5；M=中位数）。在属性级，完整性 (M_D=6；M_Q=6)比正确性 (M_D=5；M_Q=6)和总体 (M_D=5；M_Q=5)更为重要，而客观性 (M_D=3.5；M_Q=4.5)的评分则褒贬不一。许多参与者认为完整性最为重要（Q_5,$D_{3、6、9、10}$）因为*"[他们]并非制定正确性和客观性规则的人"*(D_6)。记录级维度的评分褒贬不一：总体 (M_D=4；M_Q=4)、正确性 (M_D=4；M_Q=3.5)和完整性 (M_D=4；M_Q=4)。Q_4试图让自己的子集尽可能真实，所选记录大多完整，但 Q_7 没有这样做，因为他们觉得应用属性级筛选器后再这么做会适得其反。B参与者在总结环节看到质量信息后表示，他们要么以为不存在缺失值（$B_{2、10}$），要么忘了检查缺失值并承诺下次会更加留意（B_9），要么注意到了缺失值但选择忽略（$B_{4、7}$）。

图5b、图5d显示，总体而言，D参与者对使用信息的反馈不一（M_D=5；M=中位数）。$D_{2、7、8}$认为这些信息有用，D_1 则不太认同，而$D_{3、4、5、10}$对偏差和原创性丧失表示担忧，认为应在特定情境下谨慎提供使用信息。在属性级层面，总体（M_D=5）比子集中的使用（M_D=4）、可视化中的使用（M_D=3.5）和筛选器中的使用（M_D=3）更重要。大多数参与者也表示，总体维度最为重要，除了 D_6，他"选择了筛选器中使用情况得分最高的属性"。参与者认为记录级维度的用处较小（子集中的使用：M_D=3）。Q组和B组参与者在事后访谈中看到模拟使用信息后，表示使用信息可以"让 *[他们]* 更有信心地选择属性"（Q_4）、帮助核验他们的工作（Q_1），并让他们参考他人的工作（$B_{8、}$ Q_2）。总体而言，参与者认为质量比使用情况更重要，正如 D_4所说："在我们的日常生活中，数据质量重要得多，只有在多人使用同一个数据集或工具时，数据使用情况才可能有帮助"；Q_{12}也表示："如果某个属性质量高但使用率低，我仍然会选择该属性。"综合来看，这些结果验证了假设H6。

表1：与已准备的数据集子集相关的统计数据，包括其“大小”以及在三种研究条件（*B*、*Q*、*D*）下，属性级（“A”）和记录级（“R”）质量分数与使用情况得分中 高 （“% H”）、中（“% M”）、低 （“% L”）值的分布。每行中加粗并突出显示的值支持我们的假设，尤其是H1、H2和H3。例如，6.5(D) 的属性（“A”）数量（“大小”）最少， μ 这些属性被选入子集，支持了H1。由于所选数据集较为稀疏，没有任何记录（“R”）具有高（“% H”）总体质量分数。此外，中（“% M”）值不属于我们的假设，因此，与这些值对应的表格单元格既未突出显示，也未设置格式。

		基线（B）		质量（Q）		DataPilot（D）	
		μ	σ	μ	σ	μ	σ
准备好的（选定子集）大小							
Size	A	9.17	2.44	7.92	2.19	**6.5**	2.32
	R	866.17	253.08	710.83	282.85	**642.17**	249.18
总体质量评分分布							
% H	A	60.45	16.63	74.47	18	**77.32**	17.73
	R	0	0	0	0	0	0
% M	A	29.90	20.43	22.22	13.94	20.83	16.36
	R	42.36	4.12	54.78	17.09	57.45	18.7
% L	A	9.65	10.32	3.31	8.36	**1.85**	6.42
	R	57.64	4.12	45.22	17.09	**42.55**	18.70
总体使用得分分布							
% H	A	11.67	33.20	13.72	4.54	**15.45**	8.27
	R	29.03	8.14	38.16	16.89	**38.78**	12.97
% M	A	15.29	9.53	11.18	9.81	16.46	12.95
	R	55.54	3.90	54.11	13.16	54.82	8.59
% L	A	73.04	11.37	75.09	7.87	**68.08**	16.45
	R	57.64	4.12	45.22	17.09	**42.55**	18.70

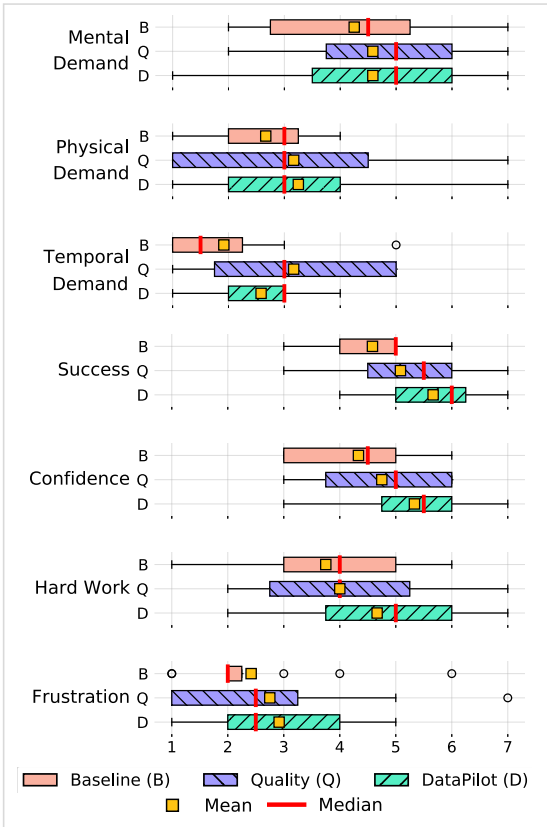

图4：参与者报告的研究任务保真度评估，采用七点不同意（1）至同意（7）量表。*D* 参与者报告称，与*Q*和*B*相比，他们在任务结束时感受到更高或相当的心理需求、辛苦程度和挫败感，但成功度和信心更高。

7 讨论

7.1 参与者选择子集的策略。

仅关注质量。十位*Q* 参与者和两位*D* 参与者仅依据质量进行选择：Q_4 通过应用完整性筛选条件，舍弃不完整的记录，$D_{1,\ 5}$ 根据完整性筛选掉一些属性，并 Q_3 通过彩色字形寻找总体质量较高的属性。

仅关注使用情况。没有*D* 参与者仅依据使用情况进行选择，这印证了领域专家的判断：在数据驱动的准备和分析过程中，质量仍是最关键的信息。

质量与使用情况并用。12名*D* 参与者中有7名同时使用了质量和使用情况。例如， D_9 应用质量筛选器，随后关注双色字形，以避开使用情况较低的属性。 D_8 先按总体使用情况得分对属性排序，再应用质量筛选器； D_{11} 应用质量筛选器后，检查子集中的使用情况维度；而$D_{4,\ 6}$ 则先依据质量进行初步选择，再利用使用情况进行核验和验证。

既不是质量，也不是使用情况。所有*B* （因为他们没有看到任何辅助信息）、两名*Q* ($Q_{1,\ 2}$)和三名*D* 参与者($D_{2,\ 3,\ 10}$)主要依靠一般属性信息（例如，属性名称和描述）以及相关性分析和趋势分析（例如，通过创建可视化）来选择他们的子集。

其他非数据驱动策略。参与者还将先入之见（Q_3、 D_4）、常识（D_1）、直觉（$D_{2,\ 3,\ 5,\ 7}$）以及试错做法（$D_{3,\ 6}$）作为辅助策略，这凸显了人类智慧在数据驱动分析中的作用。对质量、使用情况等辅助信息进行建模，可以减少这类策略带来的不确定性和不一致性。

7.2 对三步工作流的反思。

我们设计了 DataPilot，以支持三步工作流：**(1)审查原始数据、(2) 审查选定子集**，以及 **(3) 创建仪表板**。该工作流要求用户先选择感兴趣的属性和记录，再创建可视化。这种做法不同于许多可视化数据分析工作流：有些工作流没有步骤，也没有选择或取消选择属性的方式（例如 Voyager [124],Lumos [84]）；另一些则默认选中所有属性，用户只能隐藏无关属性（例如 Tableau [106]）。在 Power BI [81],中，用户首先会看到一个独立的“查询编辑器”，可在分析前转换数据；然而，由于数据准备是一个迭代过程，用户可以在分析期间随时使用“转换数据”功能

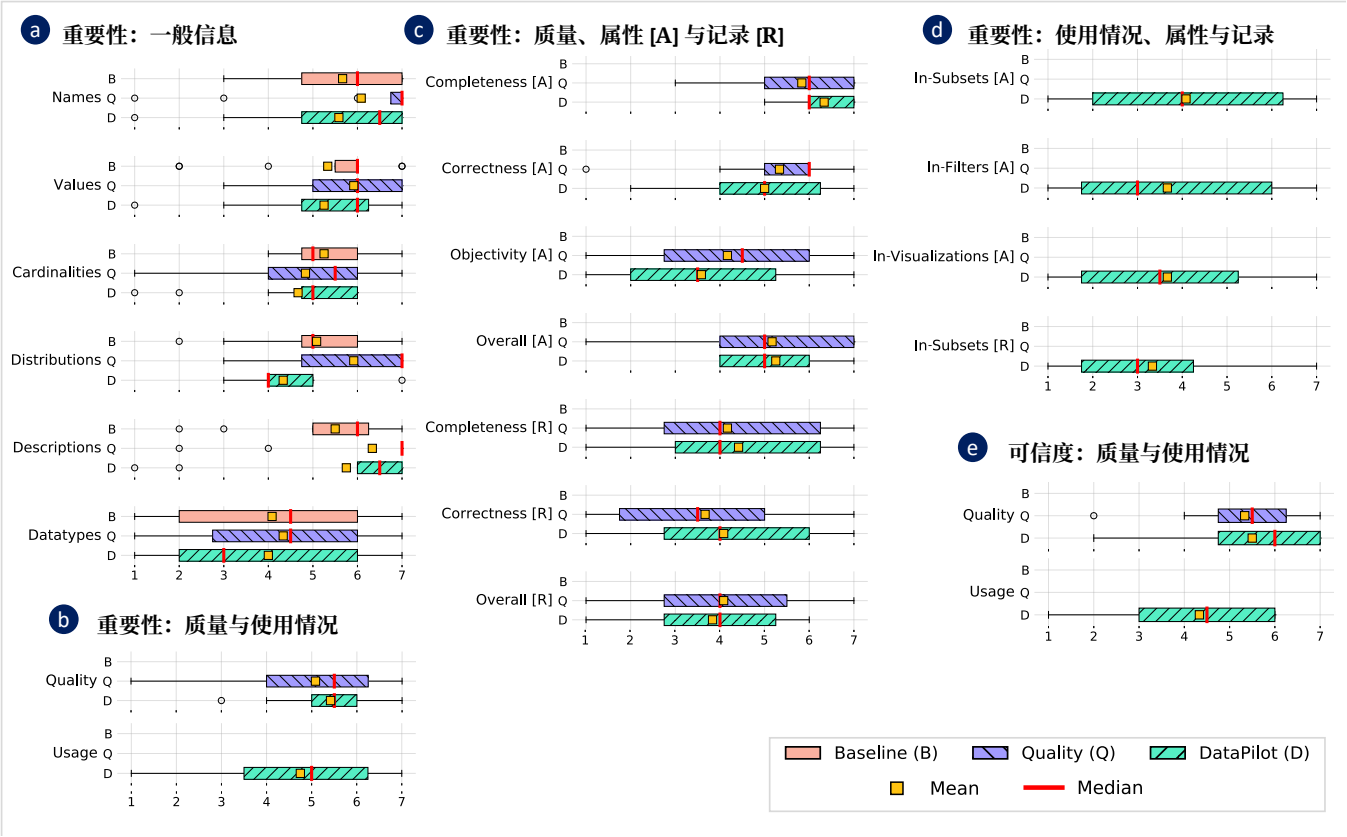

图 5：三种研究条件下，属性和记录的一般信息、质量信息和使用信息的重要性与可信度评分。某些研究条件没有箱线图，例如基线（B）在 (b)-(e) 中，因为这些条件不适用。

打开“查询编辑器”窗口。另一方面，Tableau Prep [107] 是一个独立的工具，在用于 Tableau [106]之前提供数据准备功能；不过，Tableau Prep 也提供“在 Tableau Desktop 中打开示例”功能，供用户测试当前准备中的数据样本在 Tableau 的后续分析中会如何呈现。

谈及 DataPilot 灵活的三步工作流时，参与者认为它很有用，因为“它让 *[他们]* 思考什么才是重要的，而在 *Tableau* 中，导入数据集后就会立刻进入图表制作环节，拖放属性，希望能找到有趣的内容”（B_1）。 D_{11} 评论道：“我认为第 ***1*** 步对我创建数据集来说是最重要的一步。我知道图表非常重要，但将它们恰当地放在第三步，否则面对所有属性时会让人不知所措。”对于一些参与者来说，这一工作流帮助他们专注于数据的各个方面（D_{10}），也节省了时间，因为他们可以快速判断那些听起来重要且有趣的属性是否值得查看（Q_{10}），并防止无用数据进入图表创建环节（D_7）。 B_3 使用查看原始数据这一步进行更多探索，并觉得在步骤之间来回切换很方便，“这样就能移除他们 *[不需要的]*某些属性”，他们也喜欢审查选定子集这一步，因为“他们可以看到更小、更干净的数据子集。”不过，另一些参与者要求支持“使用所有属性创建图表”（Q_{10}），这样“*[他们]* 就不必来回切换选项卡”（D_{10}）。

7.3 信任、偏差、趋同、伦理问题。

我们请 Q 和 D 参与者对他们接触到的辅助信息的可信度进行评分。如图 5e 所示，Q 和 D 参与者都认为质量信息（M_Q=5.5； M_D=6）比使用信息（M_D=4.5）更可信。有些人只是“信任整体 *[质量]* 分数” D_{12}，而另一些人在信任质量评分时则有所犹豫，并提到了正确性和客观性维度的预设约束，以及这些维度的定义不够清晰；有些参与者表示，“*[他们]*不信任 *[他们的]*管理者或其所做的设置”（D_3），因为“他们可能没有以可靠或无偏差的方式进行设置”（Q_5）。

参与者同样表示不太信任使用情况，尤其是对他人的行为或决策，因为他们不了解这些人（D_{10}）、他们的经验（Q_9）、专业知识（$Q_{1,9}$，D_1 Q_5 $D_{4\,10\,12}$ D_{10} *[* ）或任务（，，，）。有人指出，“如果我认识和我一起工作的十个人，而且大家都在做这个 *]* 相同 *[任务]*，我就会信任使用情况信息。” D_1 有人评论道，“我不想依赖之前的人对系统的理解。” D_{12} 有人猜想，创建这些数据的人可能有不同的目标，因此 *“[他们]* 可能

使用数据是为了获得与另一团队截然不同的洞察，因此 *[他们并不]* 愿意信任使用情况。”

除了讨论信任问题外，参与者还指出，使用洞察可能成为偏差的来源。 Q_7 评论说：*“[使用信息]* 肯定会使自行进行数据探索的人产生偏差……这不一定是专家式探索，而是更倾向于利用。” Q_5 解释说：“如果我必须了解其他人如何使用这个数据集，那就会使我产生偏差。” Q_7 提到了趋同理论 [75],“如果我选择一个属性，并注意到过去选择它的人不多，那么我就会忍不住取消选择。另一方面，如果先前的用户没有选择某个原本相关且重要的属性，那么由于趋同效应，它很可能会继续鲜少被使用。”

隐私是参与者和我们的领域专家都提出的另一个顾虑。除了适用性和可用性之外，参与者还质疑，提取使用数据并公开分享是否合乎伦理，即使这些数据是匿名且经过汇总的。他们还担心用户不愿分享自己的使用历史。事实上，Thom-Santelli 等人 [108] 发现，当用户（专家和新手贡献者）感受到威胁时（例如，工作场所的奖金或晋升岌岌可危时），他们之间可能会产生矛盾。

考虑到以上因素，我们认为，在子集选择与分析过程中提供质量和使用情况信息，是缓解领域专家诸多顾虑的有效方式；不过，提供更多上下文和高级配置功能是接下来重要的一步，尤其是对于高级用户。

8 经验教训

重视数据准备，而不仅仅是可视化数据分析。 受“垃圾进，垃圾出”原则的启发，应用需要投入更多数据工作 [99]„ 这也得到了用户的认同。要在数据分析工具中实现这种平衡，就需要提供检查和交互辅助信息的功能，例如 DataPilot 的质量和使用情况信息。此外，还应考虑调整工作流，将数据准备（例如，先选择相关属性和记录）置于优先位置，并整合到当前的可视化数据分析工具工作流中。

提供质量信息，以实现准确、客观的分析。 用户对质量的重要性和可信度均给予了积极反馈，这表明质量信息能够促使用户暂停并反思数据的状况，在进行分析之前采取适当的纠正措施（例如清理数据），或者干脆不使用这些数据。

呈现使用信息，但务必谨慎。用户对使用信息的重要性和信任度看法不一。使用信息有助于引导用户从以往的判断中汲取灵感，但也可能适得其反，导致不良做法（例如，有偏差的分析行为）蔓延，或抑制组织内部的创造力和原创性（例如，阻碍新想法的涌现）。实现良好平衡的一种方式是按需呈现使用信息，例如，“当我开始了解数据集时，我想隐藏双色字形的后半部分 *[使用情况]* 的 *[双色字形]*”（Q_5）。

不同任务需要灵活的信息。数据准备和后续分析都受到具体任务要求和用户偏好的影响。工具必须提供所需的灵活性，例如，为质量和使用情况的各个维度分配不同权重，或确定不同的聚合函数（最大值、平均值）来计算总体评分；覆盖预设约束以评估正确性和客观性；并调整阈值以确定高、中、低评分阈值。

可以进一步探索针对质量和使用特征的不同引导程度。参与者认为，呈现质量和使用信息的可视化交互功能，有助于“引导” [16] 他们了解该数据集。未来的工具也可以探索更高程度的引导（例如“指示”或“规定” [16]），更主动地引导用户，而不只是被动地提高他们的意识、期望他们凭良知采取行动。

组织应开始构建并利用集体智慧。组织应记录其数据库和应用程序中的使用日志，以此建模使用信息，从而提高团队内部及团队之间普通用户的认知。此外，随着新数据定期导入（旧数据归档），持续保存并随后监控辅助信息，有助于发现变化趋势、标记异常事件，并全面追踪数据溯源，从而确保数据管理准确高效。

9 局限性与未来工作

我们指出了与研究和工具相关的五项主要局限性。

研究局限。第一，在设计研究期间，我们以小组环境开展访谈；虽然受访者之间有良好的工作关系，但这种讨论方式可能会带来与性别和组织层级相关的复杂问题，因此必须加以考虑（例如，采用一对一访谈）。第二，在用户研究期间，我们合理地假设参与者不熟悉该数据集，因此具有相似的专业知识，这有助于确保内部效度；然而，从外部效度的角度来看，这一假设在现实场景中可能并不成立 [100]。未来的工作可以引入加权机制，根据使用新近性（例如，提高近期数据的重要性）、用户专业知识（倾向于专家），或使用该数据的应用的关键性，更准确地估算使用情况。第三，由于参与者并非领域专家，我们没有请专家评估选定子集或最终仪表板；未来应通过领域专家参与的用户研究，进一步评估这些结果的质量。第四，我们聚焦于为仪表板导出可视化这一特定任务，这可能影响了属性和记录的选择；未来的工作应考虑开发其他工具，以研究子集选择以外的下游分析任务，例如排序和聚类。第五，尽管数据质量维度不易跨领域迁移 [70], ，研究参与者建议将 DataPilot 用于搜索 Kaggle 上的数据集 [59] (Q_{10})、生成软件使用教程（Q_2, D_5），以及为机器学习准备公平、准确的数据集（D_4）；这些方向有待未来探索。

工具局限。第一，DataPilot 目前支持表格数据集的质量信息；未来的工作可以探索其他结构化数据集（例如关系数据库）和非结构化数据集（例如文本、文档）。第二，质量还涉及其他依赖数据的方面（例如一致性表示、易操作性和及时性 [89]）以及依赖流程的方面（例如数据收集 [38]）；同样，使用情况也不止涉及子集选择和仪表板构建任务，还包括其他方面（例如可视化中多个属性的共同使用频率，以及缩放和平移等可视化交互的频率 [14]），这些方面未来都可以进行操作化。第三，DataPilot 的仪表板视图目前支持创建非联动可视化；未来的工作可以探索刷选与联动等交互功能的影响。第四，为确保可扩展性，DataPilot 使用 SQL 查询计算质量和使用情况得分（客观性则通过 SQL 和 Python 共同计算）；然而，这些新增维度中有些可能难以通过 SQL 操作化，因此难以扩展。第五，完整性、正确性和客观性的质量约束目前以类 SQL 语法硬编码在 DataPilot 源代码中。未来的工作可以为用户提供交互功能，使其能够配置这些约束，并直接通过用户界面清理数据（例如处理缺失值）。

10 结论

DataPilot 是一种可视化数据准备与分析工具，通过对两类辅助信息——质量 和使用情况——进行建模，帮助用户分析大型且陌生的表格数据集、选择相关子集并构建可视化仪表板。DataPilot 是一项为期两个月的设计研究成果，研究涉及 14 位数据工作者；他们强调了数据质量的重要性，并建议呈现数据使用特征，以指导用户进行数据准备。一项涉及 36 位参与者的用户研究表明，结合质量和使用信息有助于用户更成功、更有信心地选择更小且有效的数据子集。我们认为，通过质量和使用信息，组织能够构建集体智慧，提升透明度和准确性，从而促进团队间更紧密的协作与合作。

致谢

我们感谢接受访谈的数据工作者、研究参与者、佐治亚理工学院可视化实验室成员，以及匿名评审者在本研究的不同阶段提供反馈。

参考文献

[1] 2012. 印度开放政府数据（OGD）平台。检索日期：2022年11月21日，网址：https://data.gov.in/ [2] 2021. CERN 开放数据门户。检索日期：2022年11月21日，网址：https://opendata. cern.ch[3] Ziawasch Abedjan、Xu Chu、Dong Deng、Raul Castro Fernandez、Ihab F. Ilyas、Mourad Ouzzani、Paolo Papotti、Michael Stonebraker 和 Nan Tang。2016年。检测数据错误：现状如何以及需要做些什么？*VLDB*基金会会刊 9，12（2016年8月），993–1004。https://doi.org/10.14778/2994509.2994518[4] Jason Alexander、Andy Cockburn、Stephen Fitchett、Carl Gutwin 和 Saul Greenberg。2009年。重新审视阅读痕迹：足迹滚动条的分析、设计与评估。载于*SIGCHI* 人机交互系统中的人为因素会议论文集。1665–1674。[5] Micah Allen、Davide Poggiali、Kirstie Whitaker、Tom Rhys Marshall 和 Rogier A Kievit。2019年。雨云图：用于稳健数据可视化的多平台工具。*Wellcome Open Research* 4（2019年）。[6] Jumana Almahmoud、Robert DeLine 和 Steven M Drucker。2021年。团队如何交流机器学习模型的质量：一项国际.....

科技公司。《ACM人机交互会议论文集》5，GROUP（2021），1–24。[7] 亚马逊。2022年。Deequ。检索于2022年7月29日，网址：https://aws.amazon.com/blogs/big-data/test-data-quality-at-scale-with-deequ/[8] Clemens Arbesser、Florian Spechtenhauser、Thomas Mühlbacher和Harald Piringer。2017年。Visplause：使用合理性检查对大量时间序列进行可视化数据质量评估。《IEEE可视化与计算机图形学汇刊》23，1（2017年），641–650。https://doi.org/10.1109/TVCG.2016.2598592[9] Matthew Arnold、Rachel KE Bellamy、Michael Hind、Stephanie Houde、Sameep Mehta、Aleksandra Mojsilović、Ravi Nair、K Natesan Ramamurthy、Alexandra Olteanu、David Piorkowski等。2019年。事实说明书：通过供应商符合性声明提升对人工智能服务的信任。《IBM研究与开发杂志》63，4/5（2019年），6–1。[10] Ernesto Arroyo、Ted Selker和Willy Wei。2006年。利用鼠标轨迹分析网页设计的可用性工具。载于《CHI'06计算机系统中的人为因素扩展摘要》。484–489。[11] Anant Bhardwaj、Souvik Bhattacherjee、Amit Chavan、Amol Deshpande、Aaron J Elmore、Samuel Madden和Aditya G Parameswaran。2014年。Datahub：大规模协作数据科学与数据集版本管理。arXiv预印本 arXiv:1409.0798（2014年）。[12] David Borland、Wenyuan Wang、Jonathan Zhang、Joshua Shrestha和David Gotz。2019年。高维数据的选择偏差追踪与详细子集比较。《IEEE可视化与计算机图形学汇刊》26，1（2019年），429–439。[13] Richard E Boyatzis。1998年。转化定性信息：主题分析与编码开发。Sage出版社。[14] Matthew Brehmer和Tamara Munzner。2013年。抽象可视化任务的多层级类型学。《IEEE可视化与计算机图形学汇刊》19，12（2013年），2376–2385。[15] John Brooke等。1996年。SUS——一种简便粗略的可用性量表。《工业领域的可用性评估》189，194（1996年），4–7。[16] Davide Ceneda、Theresia Gschwandtner、Thorsten May、Silvia Miksch、Hans-Jörg Schulz、Marc Streit和Christian Tominski。2017年。表征可视分析中的引导。《IEEE可视化与计算机图形学汇刊》(2017 10 1109 2016 2598468)。https://doi.org/ . /TVCG. . [17] Wo L Chang、Arnab Roy、Mark Underwood等。2019年。NIST大数据互操作性框架：第4卷，安全与隐私。（2019年）。[18] Xu Chu和Ihab F. Ilyas。2016年。定性数据清洗。《VLDB基金会会刊》9，13（2016年），1605–1608。https://doi.org/ . / . 3007320 [19] 2016 Xu Chu、Ihab F. Ilyas、Sanjay Krishnan和Jiannan Wang。数据清洗：概述与新兴挑战。载于《2016年国际数据管理会议论文集》（美国加利福尼亚州旧金山）（SIGMOD '16）。计算机协会，纽约，美国，2201–2206。https://doi.org/ 10 1145 2882903 2912574 . / . Xu Chu、Ihab F Ilyas和Paolo Papotti。[20] 2013年。整体式数据清洗：将违规项置于上下文中。载于《2013年第29届IEEE国际数据工程会议》（ICDE）。IEEE，458–469。[21] 美国国家研究委员会等。2000年。《人们如何学习：大脑、心智、经验与学校：扩展版》。美国国家科学院出版社。[22] Jeffrey Dastin。2018年。亚马逊弃用一款曾显露出对女性存在偏差的秘密人工智能招聘工具。https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [23] Dean De Cock。2011年。艾姆斯，爱荷华州：作为学期末回归项目的波士顿住房数据替代数据集。《统计学教育杂志》19，3（2011年）。https://doi.org/10.1080/10691898.2011.11889627[24] Dong Deng、Raul Castro Fernandez、Ziawasch Abedjan、Sibo Wang、Michael Stonebraker、Ahmed K Elmagarmid、Ihab F Ilyas、Samuel Madden、Mourad Ouzzani和Nan Tang。2017年。Data Civilizer系统。载于CIDR。[25] Dremio。2022年。Dremio。检索于2022年7月29日，网址：https://www.dremio.com/ [26] Ian Drosos、Titus Barik、Philip J. Guo、Robert DeLine和Sumit Gulwani。2020年。Wrex：用于合成可读数据科学家代码的统一示例驱动编程交互。载于《2020年CHI计算机系统中的人为因素会议论文集》（美国夏威夷州檀香山）（CHI '20）。计算机协会，纽约，美国，1–12。https://doi.org/10.1145/ 3313831.3376442[27] Ian Drosos、Titus Barik、Philip J. Guo、Robert DeLine和Sumit Gulwani。2020年。Wrex：用于合成可读数据科学家代码的统一示例驱动编程交互（CHI '20）。计算机协会，纽约，美国， 1 12 10 1145 3313831 3376442 –。[28] https://doi.org/ . / . Susan Dumais、Robin Jeffries、Daniel M Russell、Diane Tang和Jaime Teevan。2014年。通过日志数据和分析了解用户行为。载于《人机交互中的认知方式》。施普林格，349–372。[29] Upol Ehsan、Q Vera Liao、Michael Muller、Mark O Riedl和Justin D Weisz。2021年。拓展可解释性：迈向人工智能系统中的社会透明度。载于《2021年CHI计算机系统中的人为因素会议论文集》。

[30] Niklas Elmqvist、Andrew Vande Moere、Hans-Christian Jetter、Daniel Cernea、Harald Reiterer和TJ Jankun-Kelly。2011年。面向信息可视化的流体交互。《信息可视化》10，4（2011年），327–340。[31] Thomas Erickson和Wendy A. Kellogg。2000年。社会半透明性：一种设计支持社会过程的系统的方法。《ACM计算机-人机交互汇刊》。 7, 1 2000 10 1145 344949 345004 （3月）。[32] https://doi.org/ . / . Mina Farid、Alexandra Roatis、Ihab F. Ilyas、Hella-Franziska Hoffmann和Xu Chu。2016年。CLAMS：为数据湖带来质量（SIGMOD '16）。计算机协会，纽约，美国，2089–2092。https://doi.org/10.1145/2882903.2899391[33] Raul Castro Fernandez、Ziawasch Abedjan、Famien Koko、Gina Yuan、Samuel Madden和Michael Stonebraker。2018年。Aurum：数据发现系统。载于《2018年第34届IEEE国际数据工程会议》（ICDE）。IEEE。[34] Ian Fette和Alexey Melnikov。2011年。WebSocket协议。技术报告。[35] Mike Fleckenstein、Lorraine Fellows和Krista Ferrante。2018年。数据质量。载于《现代数据战略》。施普林格。[36] Imola K Fodor。2002年。降维技术综述。技术报告。劳伦斯利弗莫尔国家实验室，加利福尼亚州（美国）。[37] Python软件基金会。2022年。Python。检索于2022年7月29日，网址：https://www.python.org/[38] Lawrence M Friedman、Curt D Furberg和David L DeMets。2010年。数据收集与质量控制。载于《临床试验基础》。施普林格。[39] Christian Fürber。2016年。语义技术。载于《利用语义技术进行数据质量管理》。施普林格，56–68。[40] Gerry Gaffney。1999年。亲和图法。检索于1月3日（1999年），2013年。[41] Mirta Galesic、Rocio Garcia-Retamero和Gerd Gigerenzer。2009年。使用图标阵列传达医疗风险：克服低数理能力。《健康心理学》 28, 2 2009 210 10 1037 0014474 （），。https://doi.org/ . / 2013 [42] a Lisa Gitelman。《原始数据》本身自相矛盾。麻省理工学院出版社。[43] 谷歌。2022年。Angular。检索于2022年7月29日，网址：https://angular.io/[44] David Gotz、Shun Sun和Nan Cao。2016年。自适应情境化：在高维可视化和数据选择过程中应对偏差。载于《第21届国际智能用户界面会议论文集》。85–95。[45] Carl Gutwin。2002年。轨迹：可视化近期活动以支持群体交互。载于《图形界面》。Citeseer，43–50。[46] Alon Halevy、Peter Norvig和Fernando Pereira。2009年。数据的非凡效能。《IEEE智能系统》24，2（2009年），8–12。[47] Kelli Ham。2013年。OpenRefine（2.5版）。http://openrefine. org。用于清洗和转换数据的免费开源工具。《医学图书馆协会杂志：JMLA》 101, 3 2013 233 10 3163 1536 5050 101 3 020 （），。2011 [48] https://doi.org/ . / - . . . Anders Haug、Frederik Zachariassen和Dennis Van Liempd。数据质量不佳的成本。《工业工程与管理杂志》（JIEM）4，2（2011年），168–193。[49] Kim Hazelwood、Sarah Bird、David Brooks、Soumith Chintala、Utku Diril、Dmytro Dzhulgakov、Mohamed Fawzy、Bill Jia、Yangqing Jia、Aditya Kalro等。2018年。Facebook上的应用机器学习：数据中心基础设施视角。载于《2018年IEEE高性能计算机体系结构国际研讨会》（HPCA）。IEEE，620–629。[50] Jeffrey Heer、Fernanda B Viégas和Martin Wattenberg。2007年。旅行者与窥视者：支持异步协作式信息可视化。载于《SIGCHI计算机系统中的人为因素会议论文集》。1029–1038。[51] Thomas N Herzog、Fritz J Scheuren和William E Winkler。2007年。什么是数据质量？我们为何应该关注？载于《数据质量与记录链接技术》。施普林格，7–15。[52] William C Hill、James D Hollan、Dave Wroblewski和Tim McCandless。1992年。编辑磨损与阅读痕迹。载于《SIGCHI计算机系统中的人为因素会议论文集》。3–9。[53] Sarah Holland、Ahmed Hosny、Sarah Newman、Joshua Joseph和Kasia Chmielinski。2020年。数据集营养标签。《数据保护与隐私》第12卷：数据保护与民主 12（2020年），1。[54] Bill Howe、Garret Cole、Nodira Khoussainova和Leilani Battle。2011年。针对临时数据库的自动示例查询。载于《2011年ACM SIGMOD数据管理国际会议论文集》（希腊雅典）（SIGMOD '11）。计算机协会，纽约，美国。https://doi.org/ 10 1145 1989323 1989487 . / . Petra Isenberg、Niklas Elmqvist、[55] Jean Scholtz、Daniel Cernea、Kwan-Liu Ma和Hans Hagen。2011年。协作式可视化：定义、挑战与研究议程。《信息可视化》10，4（2011年），310–326。[56] George H. John、Ron Kohavi和Karl Pfleger。1994年。无关特征与子集选择问题。载于《机器学习论文集1994》，William W. Cohen和Haym Hirsh（编）。Morgan Kaufmann，旧金山（加利福尼亚州），121–129。https://doi.org/ 10 1016 978 1 55860 335 6 50023 4 . /B - - - - . - A. Jović、K. Brkić和 2015 [57] N. Bogunović。特征选择方法综述及其应用。载于《2015年第38届信息与通信技术、电子与微电子国际大会》（MIPRO）。1200–1205。https://doi.org/ 10 1109 2015 7160458 . /MIPRO. . Kaggle。Kaggle机器学习与 2019 [58] 数据科学调查。检索于2022年2月25日，网址：https: // www.kaggle.com/c/kaggle-survey-2019。[59] Kaggle。2022年。Kaggle。检索于2022年7月29日，网址：https://www.kaggle.com/[60] Sean Kandel、Andreas Paepcke、Joseph Hellerstein 和 Jeffrey Heer。2011年。Wrangler：数据转换脚本的交互式可视化规范。载于 SIGCHI计算机系统中的人为因素会议论文集。3363–3372。[61] Sean Kandel、Andreas Paepcke、Joseph M Hellerstein 和 Jeffrey Heer。2012年。企业数据分析与可视化：一项访谈研究。《IEEE可视化与计算机图形学汇刊》18，12（2012），2917–2926。[62] Sean Kandel、Ravi Parikh、Andreas Paepcke、Joseph M. Hellerstein 和 Jeffrey Heer。2012年。Profiler：用于数据质量评估的集成统计分析与可视化。载于高级视觉界面国际工作会议论文集（意大利卡普里岛）（AVI '12）。计算机协会，美国纽约州纽约市，547–554。https://doi.org/10.1145/ 2254556.2254659[63] Jihie Kim、Ewa Deelman、Yolanda Gil、Gaurang Mehta 和 Varun Ratnakar。2008年。wings/pegasus系统中的溯源轨迹。《并发与计算：实践与经验》20，5（2008），587–597。[64] Miryung Kim、Thomas Zimmermann、Robert DeLine 和 Andrew Begel。2017年。软件团队中的数据科学家：现状与挑战。《IEEE软件工程汇刊》44，11（2017），1024–1038。[65] Laura Koesten、Emilia Kacprzak、Jeni Tennison 和 Elena Simperl。2019年。结构化数据的协作实践：工具是否满足用户需求？载于2019年CHI计算机系统中的人为因素会议论文集。1–14。[66] Laura Koesten、Pavlos Vougiouklis、Elena Simperl 和 Paul Groth。2020年。数据集再利用：推动原则转化为实践。《Patterns》1，8（2020），100136。[67] Kurt Koffka。2013年。《格式塔心理学原理》。Routledge。[68] Jonathan Lancar。2022年。Luma。 h t t p s:// g i t h ub.co m/adobe/experience-platform-dsw-reference/blob/master/datasets/luma/luma_post_extended.csv。[69] Nuno Laranjeiro、Seyma Nur Soydemir 和 Jorge Bernardino。2015年。数据质量调查：劣质数据分类。载于2015年IEEE第21届太平洋地区可靠计算国际研讨会（PRDC）。IEEE，179–188。[70] Sabina Leonelli。2019年。数据治理是阐释的关键：重新构想数据科学中的数据。《哈佛数据科学评论》1，1（2019），10–1162。[71] Zipeng Liu、Zhicheng Liu 和 Tamara Munzner。2020年。图像编辑日志的数据驱动多级分段。载于2020年CHI计算机系统中的人为因素会议论文集。1–12。[72] Yuyu Luo、Chengliang Chai、Xuedi Qin、Nan Tang 和 Guoliang Li。2020年。通过复合问题实现面向渐进式可视化的交互式清理。载于2020年IEEE第36届国际数据工程会议（ICDE）。IEEE。[73] Yuyu Luo、Chengliang Chai、Xuedi Qin、Nan Tang 和 Guoliang Li。2020年。Vis-Clean：面向渐进式可视化的交互式清理。VLDB基金会会刊（aug 2020 10 14778 3415478 3415484 ）。https://doi.org/ . / . Rupa 2019 [74] Mahanti。数据、数据质量与低数据质量的成本。载于《数据质量：维度、测量、策略、管理与治理》。Quality Press。[75] Antony SR Manstead、Miles Ewe Hewstone、Susan T Fiske、Michael A Hogg、Harry T Reis 和 Gün R Semin。1995年。《布莱克威尔社会心理学百科全书》。Blackwell Reference/Blackwell Publishers。[76] Chris Mayfield、Jennifer Neville 和 Sunil Prabhakar。2010年。ERACER：用于统计推断和数据清洗的数据库方法。载于2010年ACM SIGMOD数据管理国际会议论文集。75–86。[77] Srđan Medić、Biljana Karlović 和 Zrinko Cindrić。2016年。新标准ISO 9001：2015及其对组织的影响。《复杂系统跨学科描述：INDECS》14，2（2016），188–193。[78] Ninareh Mehrabi、Fred Morstatter、Nripsuta Saxena、Kristina Lerman 和 Aram Galstyan。2021年。机器学习中的偏差与公平性调查。《ACM计算调查》（CSUR）54，6（2021），1–35。[79] Metaplane。2022年。Metaplane。检索于2022年7月29日，网址：https://www.metaplane.dev/[80] 微软。2022年。LinkedIn。检索于2022年7月29日，网址：https://linkedin.com/[81] 微软。2022年。Power BI Desktop。检索于2022年5月25日，网址：https://powerbi.microsoft.com/en-us/[82] 微软。2022年。Teams。检索于2022年7月29日，网址：https://www.microsoft.com/en-us/microsoft-teams/group-chat-software [83] Tadhg Nagle、Thomas C Redman 和 David Sammon。2017年。只有3%的公司数据达到基本质量标准。《哈佛商业评论》（2017）。[84] Arpit Narechania、Adam Coscia、Emily Wall 和 Alex Endert。2021年。Lumos：提升可视化数据分析过程中对分析行为的认知。《IEEE可视化与计算机图形学汇刊》（2021）。[85] Jakob Nielsen 和 Kara Pernice。2010年。《眼动追踪与网站可用性》。New Riders。[86] Chris North、Remco Chang、Alex Endert、Wenwen Dou、Richard May、Bill Pike 和 Glenn Fink。2011年。分析溯源：过程+ 交互+ 洞察。载于CHI'11计算机系统中的人为因素扩展摘要。33–36。[87] Boris Otto、Kai M Hüner 和 Hubert Österle。2009年。面向业务的数据质量指标识别。载于ICIQ。122–134。[88] Jinglin Peng、Weiyuan Wu、Brandon Lockhart、Song Bian、Jing Nathan Yan、Linghao Xu、Zhixuan Chi、Jeffrey M Rzeszotarski 和 Jiannan Wang。2021年。

DataPrep.EDA：用于Python统计建模的任务中心型探索性数据分析。载于*2021*年国际数据管理会议论文集。2271–2280。[89] Leo L. Pipino、Yang W. Lee 和 Richard Y. Wang。2002年。数据质量评估。《*ACM*通讯》 45, 4 2002 211 218 10 1145 505248 506010（apr），–。https://doi.org/ . / .[90] 2016 。欧盟开放数据门户。欧盟开放数据。（）。2016 。[91] Pavel Pudil 和 Jana Novovičová。1998年。结合问题知识进行特征子集选择的新方法。载于特征提取、构造与选择。施普林格，101–116。[92] Dorian Pyle。1999年。面向数据挖掘的数据准备。Morgan Kaufmann。[93] Eric D Ragan、Alex Endert、Jibonananda Sanyal 和 Jian Chen。2015年。可视化与数据分析中的溯源特征：溯源类型与用途的组织框架。《*IEEE*可视化与计算机图形学汇刊》22，1（2015），31–40。[94] Vijayshankar Raman 和 Joseph M Hellerstein。2001年。Potter's Wheel：交互式数据清洗系统。载于*VLDB*，第1卷。381–390。[95] Thomas C. Redman。2016年。劣质数据每年给美国造成3万亿美元损失。https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year [96] Thomas C Redman。2018年。如果你的数据质量差，机器学习工具就毫无用处。《哈佛商业评论》2（2018）。[97] Robert Richards。2006年。表述性状态转移（REST）。载于《*Pro PHP XML*与*Web*服务》。施普林格，633–672。[98] Steven G Rogelberg、Linda Rhoades Shanock 和 Cliff W Scott。2012年。会议中的时间与金钱浪费：提高投资回报率。《小组研究》43，2（2012），236–245。[99] Nithya Sambasivan、Shivani Kapania、Hannah Highfill、Diana Akrong、Praveen Paritosh 和 Lora M Aroyo。2021年。“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*2021*年*CHI*计算机系统中的人为因素会议论文集。1–15。[100] Mark A Schmuckler。2001年。什么是生态效度？维度分析。《婴幼儿期》2，4（2001），419–436。[101] Hans-Jörg Schulz、Thomas Nocke、Magnus Heitzler 和 Heidrun Schumann。2017年。面向表格数据可视化分析的数据描述符系统化综述。《信息可视化》16，3（2017），232–256。[102] Michael Sedlmair、Miriah Meyer 和 Tamara Munzner。2012年。设计研究方法论：来自实践现场与文献资料的思考。《*IEEE*可视化与计算机图形学汇刊》18，12（2012），2431–2440。https://doi. org/10.1109/TVCG.2012.213[103] Fatimah Sidi、Payam Hassany Shariat Panahy、Lilly Suriani Affendey、Marzanah A. Jabar、Hamidah Ibrahim 和 Aida Mustapha。2012年。数据质量：数据质量维度调查。载于*2012*年信息检索与知识管理国际会议。300–304。https://doi.org/10.1109/ InfRKM.2012.6204995[104] Qinbao Song 和 Martin Shepperd。2007年。缺失数据填补技术。《国际商业智能与数据挖掘杂志》2，3（2007年10月），261–291。https://doi.org/10.1504/ IJBIDM. 2007 015485 。Brian Stein 和 Alan Morrison。2014年。数据湖与打破数据孤岛的愿景。普华永道，《技术展望：重新思考整合》（2014）。[105]
h
t
t
p
s://
w
w
w.
p
w
c.
c
o
m/
u
s/
en/technology-forecast/2014/cloud-co
mputing/assets/pdf/pwc-technology-forecast-data-lakes.pdf [106] Tableau。2022年。Tableau。检索于2022年7月29日，网址：https://www.tableau.com/ [107] Tableau。2022年。Tableau Prep。检索于2022年5月25日，网址：https://www.tableau.com/products/prep [108] Jennifer Thom-Santelli、Dan Cosley 和 Geri Gay。2010年。你知道些什么？协作系统中的专家、新手与领地性。载于*SIGCHI*计算机系统中的人为因素会议论文集。1685–1694。[109] Jennifer Thom-Santelli、Dan R Cosley 和 Geri Gay。2009年。我的东西就是我的：协作创作中的领地性。载于*SIGCHI*计算机系统中的人为因素会议论文集。1481–1484。[110] Trifacta。2022年。Wrangler。检索于2022年7月29日，网址：https://www.trifacta.com/

[111] A Caroline Tynan 和 Jennifer Drayton。1987年。市场细分。《营销管理杂志》2，3（1987），301–335。[112] Uber。2022年。Uber 数据手册。检索日期：2022年7月29日，来源：https://eng.uber.com/ databook/[113] Hannes Ulrich、Ann-Kristin Kock-Schoppenhauer、Noemi Deppenwiese、Robert Gött、Jori Kern、Martin Lablans、Raphael W Majeed、Mark R Stöhr、Jürgen Stausberg、Julian Varghese、Martin Dugas 和 Josef Ingenerf。2022年。理解元数据的本质：系统综述。《医学互联网研究杂志》24，1（2022年1月11日），e 。https://doi.org/ 。/ Jürgen Umbrich、Sebastian Neumaier 和 Axel Polleres。2015年。开放数据门户的质量评估与演化。载于*2015*年第*3*届未来互联网、物联网与云国际会议。IEEE，404–411。[115] Reza Vaziri、Mehran Mohsenzadeh 和 Jafar Habibi。2019年。使用加权指标衡量数据质量。《全面质量管理与卓越商业》30，5-6（2019），708–720。[116] Fernanda B Viegas 和 Martin Wattenberg。2006年。面向交流的可视化：行动呼吁。《*IBM*系统杂志》45，4（2006），801。[117] Fernanda B. Viegas、Martin Wattenberg、Frank van Ham、Jesse Kriss 和 Matt McKeon。2007年。ManyEyes：互联网规模的可视化网站。13，6（2007年11月），–。https://doi.org/ 。/TVCG. 。Emily Wall、Leslie M Blaha、Lyndsey Franklin 和 Alex Endert。2017年。警告，可能存在偏差：一种检测交互式视觉分析中认知偏差的方法。载于*2017*年*IEEE*视觉分析科学与技术会议（*VAST*）。IEEE，104–115。[119] Emily Wall、Arpit Narechania、Adam Coscia、Jamal Paden 和 Alex Endert。2021年。左、右与性别：探索交互轨迹以减轻人类偏见。《*IEEE*可视化与计算机图形学汇刊》（2021）。[120] Dakuo Wang、Judith S Olson、Jingwen Zhang、Trung Nguyen 和 Gary M Olson。2015年。DocuViz：可视化协作写作。载于第*33*届*ACM*计算机系统中的人为因素年度会议论文集。1865–1874。[121] Steve Wedig 和 Omid Madani。2006年。用于评估个性化机会的大规模查询日志分析。载于第*12*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集。742–747。[122] Wesley Willett、Jeffrey Heer 和 Maneesh Agrawala。2007年。气味提示控件：利用嵌入式可视化改进导航提示。《*IEEE*可视化与计算机图形学汇刊》13，6（2007），1129–1136。[123] Wesley Willett、Jeffrey Heer、Joseph Hellerstein 和 Maneesh Agrawala。2011年。CommentSpace：为协作式视觉分析提供结构化支持。载于*SIGCHI*计算机系统中的人为因素会议论文集。[124] Kanit Wongsuphasawat、Dominik Moritz、Anushka Anand、Jock Mackinlay、Bill Howe 和 Jeffrey Heer。2015年。Voyager：通过可视化推荐的分面浏览进行探索性分析。《*IEEE*可视化与计算机图形学汇刊》22，1（2015），649–658。[125] Mohamed Yakout、Laure Berti-Équille 和 Ahmed K Elmagarmid。2013年。不要害怕：使用最大似然和有界变更进行可扩展的自动修复。载于*2013*年*ACM SIGMOD*数据管理国际会议论文集。553–564。[126] Cong Yan 和 Yeye He。2020年。Auto-Suggest：利用数据科学笔记本学习推荐数据准备步骤。载于*2020*年*ACM SIGMOD*数据管理国际会议论文集（美国俄勒冈州波特兰）（*SIGMOD '20*）。计算机协会，美国纽约州纽约市， 1539 1554 10 1145 3318464 3389738 –。https://doi.org/ 。/ 。[127] Faheem Zafar、Abid Khan、Saba Suhail、Idrees Ahmed、Khizar Hameed、Hayat Mo- hammad Khan、Farhana Jabeen 和 Adeel Anjum。2017年。可信数据：安全溯源方案的综述、分类与未来趋势。《网络与计算机应用杂志》94（2017），50–68。[128] Amy X Zhang、Michael Muller 和 Dakuo Wang。2020年。数据科学工作者如何协作？角色、工作流与工具。《*ACM*人机交互会议论文集》4，CSCW1（2020），1–23。[129] Shichao Zhang、Chengqi Zhang 和 Qiang Yang。2003年。面向数据挖掘的数据准备。《应用人工智能》17，5-6（2003），375–381。https: // doi.org/10.1080/713827180 arXiv:https://doi.org/10.1080/713827180