arXiv:2304.05767v1 [cs.DL] 12 Apr 2023

引导科学家实现数据可检索性的决策树 *

Andrea Bianchi、Giordano d'Aloisio、Francesca Marzi 和 Antinisca Di Marco

意大利拉奎拉大学{andrea.bianchi,giordano.daloisio}
@graduate.univaq.it{francesca.marzi,antinisca.dimarco}
@univaq.it

摘要。 可复现性是科学研究的一个关键方面，指通过分析相同的数据或重复相同的实验，独立复现实验结果的能力。多年来，人们提出了许多研究，旨在使实验结果真正具备可复现性。然而，很少有研究关注 *data reproducibility*的重要性，它被定义为独立研究人员保留实验输入所用同一数据集的能力。妥善解决数据可复现性问题至关重要，因为仅仅提供数据链接往往不足以让结果具备可复现性。事实上，要使数据集完全可复现，还必须提供适当的元数据（例如预处理说明）。本研究旨在弥补这一空白，提出一棵决策树，引导研究人员实现其数据集的可复现性。具体而言，该决策树将引导研究人员判断数据集是否确实可复现，以及是否还必须提供额外的元数据（即复现数据所需的其他资源）。这棵决策树将成为未来应用程序的基础，该程序将根据具体情境（例如数据可用性、数据预处理等）自动提供必要的元数据，从而实现数据复现流程的自动化。值得注意的是，本文详细阐述了使数据集可获取的步骤，而数据集文档等其他对可复现性至关重要的方面将在未来的研究中详述。

关键词： 数据可复现性 · FAIR原则 · 决策树

1 引言

可复现性是科学研究的重要方面，指通过对相同内容进行独立分析来复现实验结果的能力。

* 本研究由欧盟—NextGenerationEU—国家复苏与韧性计划（PNRR）项目“ So-BigData.it——强化意大利社会挖掘与大数据分析研究基础设施”资助，项目编号：IR0000013，依据 2021 年 12 月 28 日第 3264 号公告；并由欧盟 NextGenerationEU 资助的“ICSC——高性能计算、大数据与量子计算国家研究中心”资助。

通过重复相同的数据分析或实验。众多研究，尤其是生物医学领域的研究，揭示了科学发现未能通过可重复性测试的普遍程度。例如，根据《自然》的一项调查，对来自不同学科的1576名研究人员开展的研究可重复性在线问卷显示，超过70%的研究人员无法复现其他科学家所做的实验，超过50%的研究人员无法复现自己所做的实验 [1]。

这种结果缺乏可复现性的现象被称为可复现性危机，对科学界产生了重大影响，使结果难以复用、比较或拓展。此外，缺乏可复现的实验不仅妨碍他人利用我们的工作，也会损害我们的声誉。实证结果的可靠性和有效性是科学的基本原则，而无法复现实验则会引发对研究可信度的担忧。

尽管许多文章介绍了如何遵循最佳实践来确保实验可复现 [9], ，通常侧重于算法的维护、发布和文档编写 [5], ，但很少有文章关注数据可复现性的重要性，即独立研究人员能够保留实验所用的同一数据集。仅提供原始数据集的链接通常不足以确保结果可复现 [11]。通常，数据集会在预处理阶段经过处理，以满足实验需求，而这一阶段可能会对结果产生重大影响。因此，除了原始数据集之外，还需要提供用于生成最终数据集（即作为算法输入的数据集）的预处理脚本。此外，数据通常受到法规保护（例如在医疗或公共领域），无法公开共享，因此必须提供其他访问方式（如果可行）。

本文旨在构建一棵决策树，引导研究人员评估其数据集的可复现性。所提出的决策树提供了一套结构化指南，不仅能够帮助研究人员核实其数据集是否具备可复现性，还明确列出了确保可复现性所需采取的步骤。通过遵循这棵决策树，研究人员可以了解影响其数据集可复现性的因素，并采取适当措施解决潜在问题。

值得注意的是，在本研究中，若一个数据集满足两个标准，我们便认为它具有可复现性：该数据集是可获取的且具备可用性。我们所说的可获取性，是指其他研究人员能够以与原始实验中使用时相同的格式获取该数据集。我们所说的可用性，是指存在适当的文档，对正确使用数据集所需的全部信息进行说明（例如，描述各列的含义、数据集的用途等）。本文重点讨论可复现性的第一个方面（即可获取性），而可用性将留待未来研究探讨。

本文的组织结构如下。第2节基于我们的经验和相关工作，阐述我们对数据可复-现性的讨论动机。第 3 节介绍我们的决策树。第 4 节总结全文。

2 动机与相关工作

在本节中，我们首先描述促使我们定义第3 节所述决策树的动机。接下来，我们讨论数据可复现性背景下的一些相关工作。

近年来，数据可复现性问题在文献中受到越来越多的关注。在 [7] 中，作者指出，缺少原始数据可能是所谓的可复现性危机的原因之一。然而，在 [11] 中，作者还指出，即使原始数据可以访问，仅提供一个数据链接也远远不够。事实上，还必须提供元数据（即正确使用数据所需的附加信息），才能使实验可复现。例如，数据可能需要经过预处理，因此也必须提供预处理说明。尽管如此，元数据取决于具体情境，而且必须根据所用数据的类型或可访问性级别等因素提供不同的元数据。例如，在医学领域，数据通常受到法规保护，不能公开共享。这意味着必须提供不同的数据使用方式（例如，使用未连接网络的计算机在本地访问数据）。类似地，教育领域的数据通常可以访问和使用，但不能以原始格式公开共享，只能以预处理后的格式共享。最后，在更广泛的数据科学领域中使用的数据通常可以公开获取和共享。

我们之前的两项工作也证实了，为实现数据可复现性所需的元数据存在异质性。在 [4] 中，我们使用了一个开放获取数据集进行实验。具体而言，这些数据集取自开放获取的存储库，并在相关论文中进行了适当说明。此外，我们还执行了一些预处理步骤，为实验准备数据。因此，在这种情况下，由于数据可开放获取且可完整分发，为了使其可复现，我们直接提供了预处理后数据的链接，以及最初介绍这些数据的论文引用。与此不同，在 [3] 中，实验所用数据并非开放获取，因此无法提供数据的直接链接。在一项医学领域的实验研究中，数据最初因隐私问题而无法访问。为了获取这些数据，我们与收集该数据集的实体建立了正式合作。通过这项合作，该实体向我们提供了有关数据集的所有必要信息，包括各个实例的详情，以及对原始数据所执行的预处理步骤。这使我们得以生成可用于实验的预处理版数据集，从而继续开展研究。这两段经历很好地展现了实现数据可复现性的不同途径。

我们的目标是通过提供一棵决策树，引导研究人员检查其数据集的可复现性，从而迈出解决这一问题的第一步。具体而言，本文聚焦于数据可检索性，其他可复现性的基本方面（即可用性）将在未来的工作中详细探讨。

关于相关工作，很少有研究探讨数据可复现性问题。Gebru 等人在 [6] 提出了一种数据表，用于记录数据集并促进数据集创建者与使用者之间的沟通。在

具体而言，他们提供了一组有助于记录数据集并使其可复现的问题。 [8] Monk 等人提出了 STRESS 指南，这是一份旨在使实证模拟结果可复现的检查清单。该清单分为六个部分，即目标、逻辑、数据、实验、实现和代码访问。在我们的工作中，我们旨在通过提出一棵决策树，自动化这些论文中描述的数据文档编制流程的部分环节，并实际引导研究人员实现数据集的可复现性。 [10], Ning 等人提出了 CROWDAQ，这是一种工具，通过可定制的用户界面组件、自动化的标注员资格认证，以及以可重复使用的格式保存流程，来规范机器学习场景中的数据收集流程。作者展示了该工具如何帮助规范数据收集流程并使其可复现。然而，他们没有考虑若干数据可复现性场景，例如所用数据并非开放获取的情况。此外，他们也没有考虑数据可复现性所需的其他元数据（例如数据文档）。

3 决策树

为了确保数据可复现性，清楚了解数据集的可访问性以及（如有）处理方法非常重要。如上文各节所述，本研究聚焦于数据集的可获取性（即使其他研究人员能够访问数据集），而数据集的可用性将在未来的工作中详细探讨。本节中，我们使用决策树构建一个用于评估数据集可获取性的框架，重点关注其对未来研究和分析的可访问性。

决策树如图1所示，也可在 Zenodo 存储库中获取 [2]。该决策树由八个主要问题组成，以内部菱形节点表示。叶节点以方形节点表示，并针对应提供哪些元数据或材料给出具体建议。每组答案都会导向树中的一个特定叶节点。通过提出相关问题，并在每个决策点提供具体建议，树的结构有助于以清晰且有条理的方式处理共享数据集时遇到的不同情形。

该树可以分为两个宏观区域，这两个区域基于这样一个事实：数据集可以被共享（即使其公开可用），也可以不被共享。如果数据集无法被共享，那么我们就转到图1的左侧，询问是否存在其他访问数据的方法（例如，数据集在一台未连接到网络的机器内本地可用）。如果没有访问数据的方法，那么该数据集不可复现，研究人员必须说明其不可复现的原因（即法规）。反之，如果存在其他访问数据的方法，并且数据也可以被完全访问（即，数据集中的所有信息都可以查看），那么研究人员必须描述如何访问数据。如果并非数据集中的所有信息都可以访问（例如，在医疗记录中，这种情况非常常见，其中本地、国家和/或

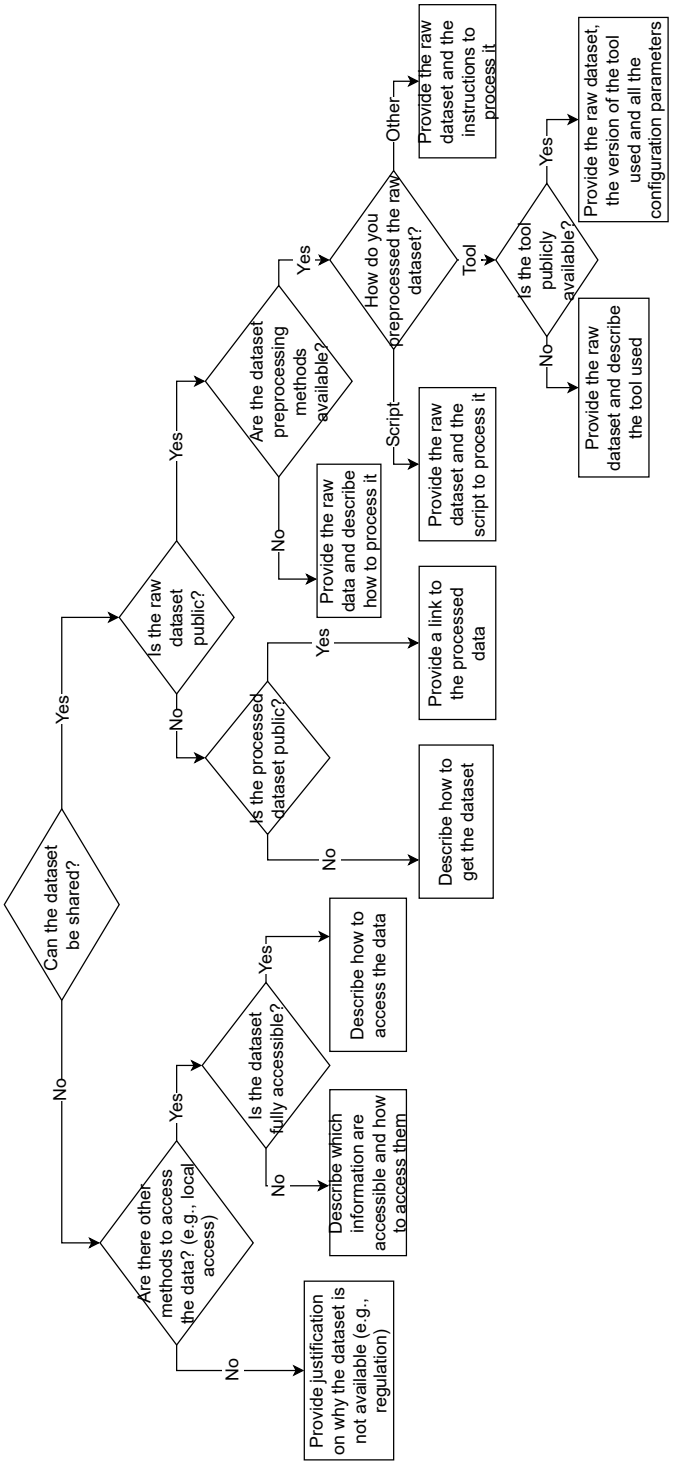

Can the dataset be shared?
Yes
No
Is the raw dataset public?
Yes
No
Are the dataset preprocessing methods available?
Yes
No
How do you preprocessed the raw dataset?
Other
Tool
Script
Provide the raw dataset and the instructions to process it
Is the tool publicly available?
Yes
No
Provide the raw dataset, the version of the tool used and all the configuration parameters
Provide the raw dataset and describe the tool used
Provide the raw dataset and the script to process it
Provide the raw data and describe how to process it
Is the processed dataset public?
Yes
No
Provide a link to the processed data
Describe how to get the dataset
Are there other methods to access the data? (e.g., local access)
Yes
No
Is the dataset fully accessible?
Yes
No
Describe how to access the data
Describe which information are accessible and how to access them
Provide justification on why the dataset is not available (e.g., regulation)

图 1. 数据可检索性树

国际法规可能限制某些数据 [3]），那么研究人员还必须说明哪些信息可以访问，以及如何访问这些信息。

相反，当数据集可以共享时，就会到达图1中树的右侧。具体而言，数据集可以以两种形式共享：原始（未处理）数据集和最终的预处理后数据集。首先要根据原始数据集是否公开这一问题作出分支决策。如果原始数据集不公开，但预处理后数据集公开，则研究人员必须提供预处理后数据集的来源链接。反之，如果原始数据或预处理后的数据不公开，则研究人员必须说明如何获取原始数据集和/或预处理后数据集（例如，向平台提交正式请求）。

如果原始数据集是公开的，则需要进一步判断数据集预处理方法是否可用。如果预处理方法不可用，则研究人员必须提供原始数据集的来源，并说明如何对其进行处理（如果应用了预处理技术）。如果预处理方法可用，则需要进一步判断数据集是如何预处理的。如果数据集是通过脚本预处理的，则必须连同原始数据集一起提供该脚本。如果使用了工具进行预处理，且该工具可公开获取，则研究人员必须提供原始数据集和该工具，并提供所用工具的版本，以及（如有）其配置参数。如果使用了工具，但该工具无法公开获取，则研究人员必须提供原始数据集，并说明所用工具及其版本，以及（如有）其配置参数。最后，如果使用了其他方法（既不是脚本也不是工具），则研究人员必须提供原始数据集以及处理数据集的说明。

本节提出的决策树可作为一种实用工具，引导研究人员评估其数据集的可复现性，从而提升研究的透明度和可重复性，并推动科学知识的发展。

4 结论与未来工作

本文提出了一个初步版本的决策树，用于引导研究人员评估和验证数据集的可复现性。具体而言，我们重点讨论了使数据集可检索所需的步骤。我们首先介绍了这一方法的动机，随后详细阐述了所展示的决策树。

未来的工作中，我们计划首先将决策树扩展到可复现性的其他基本方面（例如文档）。接下来，最终的决策树将成为一项网络服务的基础，该服务旨在生成改进数据集可复现性所需的元数据。该服务将通过开展问卷调查来运行，帮助识别影响可复现性的关键因素。该服务将根据问卷答复生成所需的元数据，从而使数据集能够被高效复现。

参考文献

1. Baker, M.: 1,500 名科学家揭示可复现性的真相。《自然》**533**(7604) (2016) 2. Bianchi, A., d'Aloisio,G.,Di Marco, A., Marzi,F.:数据可复现性树（2023 年 2 月 7685205 3 10 5281 ）。https://doi.org/ ./zenodo. . Claudiani, S., Mason, C.C., Milojkovic, D., Bianchi, A., Pellegrini, C., DiMarco, A., Fiol, C.R., Robinson, M., Ponnusamy, K.,Mokretar, K., Chowdhury, A., Al-bert, M., Reid, A.G., Deininger, M.W., Naresh, K., Apperley,J.F.,Khorashad,J.S.: 卡非佐米增强鲁索替尼对骨髓纤维化的抑制作用。《癌症》**13**13194863 4 19 2021 10 3390 ( ) ( ). https://doi.org/ . /cancers . d'Aloisio, G., D' Angelo, A., Di Marco,A., Stilo,G.: 用于增强分类任务公平性的多变量去偏器。《信息处理与管理》**60**103226 5 2 103226 2023 10 1016 2022 (), (Mar). https://doi.org/. /j.ipm. . . Fehr, J., Heiland, J., Himpe,C.,Saak,J.: 以模型降阶软件为例，介绍计算机实验的可重复性、可复现性和可重用性的最佳实践。arXiv 预印本 arXiv:1607.01191 (2016) 6. Gebru, T., Morgenstern, J., Vecchione, B.,Vaughan, J.W., Wallach,H., Iii, H.D., Crawford, K.: 数据集数据表。《ACM通讯》**64**(12), 86–92（2021 年 12 月 3458723 7 10 1145 ）。https://doi.org/ ./.Miyakawa, T.:没有原始数据，就没有科学：可复现性危机的另一个可能来源。《分子脑》**13**(1), 24（2020 年 2 月）。https://doi.org/10.1186/s13041-020-0552-2 8. Monks, T., Currie, C.S.M.,Onggo, B.S., Robinson, S.,Kunc, M.,Tay- lor, S.J.E.:加强实证模拟研究报告规范：引入 STRESS 指南。《模拟学报》**13**(1), 55–67（2019 年 1 月）。https://doi.org/10.1080/17477778.2018.1442155，出版商：Taylor & Fran- cis 电子印本：https://doi.org/ 10 1080 17477778 2018 1442155 。 / 。 。 。 9 Munafo, M.R., Nosek, B.A., Bishop, D.V., Button, K.S., Chambers, C.D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.J., Ware, J.J.,Ioannidis, J.: 可复现科学宣言。《自然·人类行为》**1**(1), 1–9 (2017) 10. Ning, Q., Wu, H., Dasigi, P., Dua, D., Gardner, M., Logan IV, R.L., Marasovic,A.,Nie,Z.: 使用 Crowdaq 进行简便、可复现且质量受控的数据收集（10月 2020 10 48550 2010 06694, 2010 06694 [ ]）。https://doi.org/ . /arXiv. . arXiv: . cs . Pawlik, M., Hutter, T., Kocher, D., Mann,W., Augsten, N.: 链接并不足够——数据的可复现性。《数据库视野》**19**，107–115（2019）。