arXiv:2303.04501v2 [cs.DC] 1 Jun 2024

面向数据驱动的环境政策制定的行星计算

Patrick Ferris、Michael Dales、Sadiq Jaffer、Amelia Holcomb、Eleanor Toye Scott、Thomas Swinfield、Alison Eyres、Andrew Balmford、David Coomes、Srinivasan Keshav、Anil Madhavapeddy 计算机科学与技术系、植物科学系和动物学系 剑桥保护倡议 英国剑桥

摘要

我们阐述了行星计算的必要性：构建基础设施，以摄取、转换、分析和发布全球数据产品，从而推动环境科学发展，并促进制定更明智的政策。我们结合自身经验，介绍一支计算机科学家团队如何与环境科学家合作开展森林碳和生物多样性保护研究，并根据现有解决方案在可扩展地处理地理空间数据方面的灵活性，以及通过可追溯性和可复现性帮助建立对结果的信任的能力，对这些方案进行分类。我们还指出了计算与环境科学交叉领域的研究空白，重点关注如何处理持续变化的数据集：这些数据集往往跨数十年收集，需要谨慎实施访问控制，而非完全开放访问。

1 引言

旨在应对二氧化碳排放不断上升$_2$、生物多样性迅速丧失 [99], 和荒漠化 [98], [63] 的政策，正变得日益依赖数据。例如，卫星和无人机遥感 [45, 96], 无线地面传感器测量 [36] 以及气候建模 [64]，都在现代气候研究和政策制定中发挥着重要作用。利用计算流水线处理和分析各类数据集，推动了生态学领域的快速进展 [78]。然而，若要有效地摄取、清理、汇总、处理、探索和存档原始数据，并据此制定政策决策，所需的计算机系统 目前非计算机科学专业人士无法使用；这些系统对于科学和政治决策而言还不够可靠，也未能广泛、公开地提供给全球所有相关方。

因此，我们提出行星计算：为推动环境科学而处理全球数据产品的摄取、转换、分析和发布所需的数字基础设施。行星计算系统不仅在提炼有关我们世界的洞见方面发挥着至关重要的作用，还能通过在决策过程中落实透明度、可复现性、问责制和及时性标准，增强公众对由此产生的政策行动的信任。

虽然行星计算表面上与传统的大数据处理相似，但它也带来了一些独特的挑战，而我们对此有切身体会。我们将首先介绍过去三年里，由计算机科学家和环境科学家组成的联合团队所构建的全球生态分析（§1.1），以及从这一实施过程中汲取的经验（§2）。我们归纳出行星计算的一些共同需求（§3），并发现现有解决方案只能部分解决计算机系统问题（§3.4）。接着，我们将讨论行星计算带来的尚未解决的计算研究机会，并探讨构建供环境科学家和政策制定者使用的系统的利弊（§4）。我们的最终目标是发展一个涵盖各种组织的生态系统（从非政府组织到政府，再到企业；它们都需要获取有关地球健康的情报），使其在未来几十年中不因任何单一控制实体的失效而瓦解，并在尊重底层观测数据集敏感性的同时，支持透明且可复现的计算科学。

1.1 环境情景的动因

在剑桥保护倡议，我们的计算机科学家与生态学家组成的联合团队多年来一直携手合作，应对以下环境挑战；我们将以此说明行星计算的必要性。

分析雨林保护的成效。赤道带沿线有许多保护项目，保护着数百万公顷的原始热带雨林，从而减少了森林砍伐造成的CO_2 排放（这种方法称为 REDD+ [27]）。我们面临的挑战是计算这些干预措施的“额外性”，以便量化每年避免了多少吨CO_2 排放。这些干预措施通常采取可持续生计的形式（例如混合森林可可种植园），作为皆伐森林的替代方式 [71], ；衡量CO_2 额外性对于证明资助这些干预措施的合理性至关重要。目前的衡量机制是由项目自行认证该地区的背景森林砍伐率，然后手动测量地上生物量相对于基线的净增量，而。然而，这种方法往往会高估避免的排放量，而且不同项目的基线各异，导致结果无法相互比较 [29]。相反，我们的计算解决方案利用过去二十年的全球卫星数据，处理全球森林生物量估算 [26, 45]，并运用统计反事实分析 [30], 以一致的方式识别干预措施带来的额外性。结果会根据非永久性进行调整 [94]，并转化为高质量碳信用，用于抵消国际航空旅行等无法避免的排放，同时确保其气候效益与避免的林业排放相当 [93]。

计算全球灭绝率。生态学家评估栖息地数据区域，以生成全球灭绝统计数据 [34],, 但不得披露单个观测点，否则物种

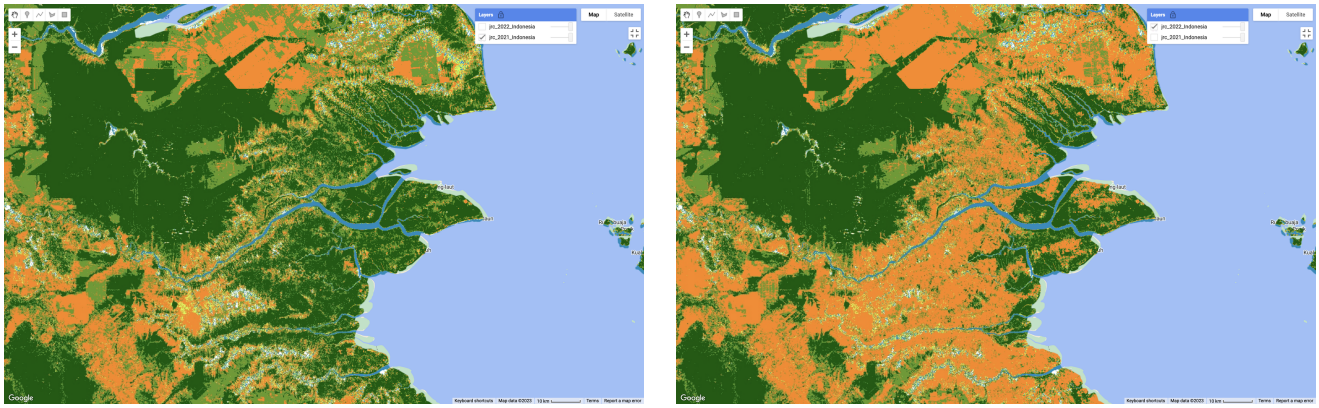


图1：JRC TMF 数据集的两个版本，展示了印度尼西亚某地区在*2008*年相同土地利用类别的数据。左图显示的是*2021*年的数据，右图显示的是*2022*年的数据。■ 未受干扰，■ 退化，■ 已毁林，■ 再生，■ 水域，■ 其他

可能会受到偷猎者的威胁 [68]。为生成这些汇总数据，他们将卫星数据（Landsat [106], MODIS [65],哥白尼计划 [35], GEDI [44]）与数十年来人工收集的读数相结合 [57]。数据的质量差异很大，因此在使用机器学习训练模型来插补缺失数据之前，必须先进行清理和标准化 [61],。根据数据得出的信息用于指导栖息地恢复和保护工作，但随着新数据不断到来，每月都必须重新生成这些信息。受到质疑时，应能够向审计人员展示结论的数据溯源，即使这些结论来自数十年前的观测。汇总数据（约 92 拍字节）涵盖 29,000+种陆生物种，经处理后用于创建名为“LIFE”的全球灭绝率指标 [48]。该指标以带版本号的栅格地图集形式发布（最终结果为 ≈500 GB），供公众使用；但生态学家也需要访问单个物种的数据。

生物多样性保护的土地利用政策。粮食和纤维生产与自然栖息地相互竞争；要了解应在哪里开展此类生产，需要进行辖区土地管理 [31]。评估土地利用变化对生物多样性影响的不同方法时，公务人员需要访问本国具有合理分辨率（<100米/像素，因此每层需要150GB存储空间）的数据集，涵盖世界自然保护联盟灭绝物种名录中的所有物种（10000+ 条目 [57]），并可追溯至30年前。同样，自然资源管理者需要能够对缩小视图或裁剪后的数据进行交互式、迭代式探索，以制定潜在的土地利用政策，然后扩展到集群计算规模，开展全国范围的运算。

2 来自实地的经验教训

当我们启动这些项目时，构建用于应对环境挑战的计算机系统看起来轻而易举。汇总后的卫星数据约有十几太字节，因此，甚至一台多核计算机就能运行相关的数据处理流程 [72]。然而，随着计算机科学家深入研究，处理生态分析时那些微妙的难题开始浮现。本文着重介绍将计算应用于生态学后，导致结果不确定性增加的四种方式。

2.1 数据不确定性

我们的首要任务是获取一套稳定的地球观测数据，以此为基础开展分析；我们很快就发现，看似相似的观测数据竟有如此多的差异。例如，我们使用了欧盟委员会联合研究中心（JRC）的热带湿润森林（TMF）数据集 [102]，该数据集根据卫星观测计算全球土地利用类别（LUC）数据，引用次数已超过 200 次。[1] 随着新算法和分析方法的出现，TMF 历史数据会得到更新（这是良好的做法），但数据发布方式并未清楚标示这些更新涉及哪些历史数据。数据更新后，曾用于其他计算（以及已发表论文）的早期版本便不再容易获取，这严重影响了研究可重复性。

这些差异的重要程度在很大程度上取决于分析所针对的地区。在我们的研究中，我们考察了印度尼西亚的热带森林带，发现汇总全国土地利用变化数据时存在显著差异。比较同一历史年份（2008年）的两个版本数据（一份来自2021年，另一份来自2022年）时，我们发现，被划为相同 年份的数据中，被归为森林砍伐 的土地增加了3.05%，被归为未受扰动 的土地减少了0.59%，被归为退化土地 的土地减少了2.00%[2]。这些数字看似不大，但应结合印度尼西亚约190万平方公里的国土面积来看。被错误划为森林砍伐的土地面积变化，约相当于多哥的国土面积！图1直观展示了这些变化；部分差异是由于“年度变化数据集中老龄次生林森林砍伐数值序列的改进和错误修正……” [23]。由于无法获取欧盟联合研究中心2021年的原始数据发布版本，这些差异会悄无声息地传导至后续研究中。我们在其他卫星数据产品中也观察到了类似的不稳定数据集 [44]。

或许有悖直觉的是，数据不确定性有时是一种可取的属性：某些数据集高度敏感，一旦意外泄露，可能导致物种遭到屠杀。例如，传感器进行的野生动物追踪可能会暴露动物的位置，

[1]根据谷歌学术，截至2023年14 2 12- - 分析与源代码：https://github.com/quantifyearth/jrc-diff

从而帮助人们捕捉、伤害或杀死带标签的动物 [39, 92]。曾有案例中，有人请愿要求开放追踪数据集，以协助狩猎（这种狩猎往往违法） [55]。为化解这一矛盾，生态学家发布了详尽的决策树，指导人们决定是否开放数据集 [101]。处理方式包括逐步屏蔽物种身份信息、其所在位置的高分辨率地图，以及汇总结果，从而限制数据的开放程度。此外，收集数据可能对当地社区造成不利影响，这也带来了伦理难题。遥感技术可能扰乱现有的集体责任机制 [90] ，并可能对生计造成不利影响，增加人们转而采取不可持续做法（例如森林皆伐）的压力，以弥补收入损失 [87]。这些因素使我们的保护合作伙伴不愿将其数据集发布到传统托管平台，而是选择手动管理。

2.2 代码中的不确定性

因此，生态学家事实上已成为数据科学家。他们不仅需要了解自己的学科领域，还必须学会处理庞大繁杂的数据集和高性能计算（HPC）基础设施，并使用编程语言编写代码：这些语言要么主要面向数据科学（R 或 Julia），要么是通用语言，后来才通过添加相关支持来适用于数据科学（例如配备了 NumPy、Pandas 等库的 Python）。但需要特别指出的是，这些语言都不是为拍字节级计算而设计的。软件工程的主要产出是代码；这个领域花了许多年编写代码，也经历了许多软件质量方面的公开失败，才逐渐掌握版本控制、测试和代码检查等如今常见的实践。因此，其他领域迟迟未采用这些工具，也就不足为奇了；在这些领域，代码被视为手段而非目的。

我们和其他人一样，在与生态学家同事合作时亲眼观察到这一现象：脚本被传来传去、反复修改，结果一出来就被搁置。这部分是因为所开发的工具——例如无处不在的 Git——即使对专家而言，要正确、充分地使用也颇具挑战性。但这些工具的架构也没有考虑到科学家工作的特点：它们鼓励人们重视最终经过反复检验的代码，而数据科学往往是一个探索性、迭代式的过程，其中大多数版本的代码都不被视为正确，因而也不值得提交；等最终版本真正成形时，作者又会立刻带着新鲜出炉的数据投入下一个挑战。对代码和 数据进行某种形式的自动版本管理，将有助于探索式编程 [66]。

如今，科学期刊往往要求在已发表的论文之外公开提供代码 [32] ，这使得改进工作的动机日益显现。然而，这通常是在最后一步才完成的，此时已发表的数据与用于生成这些数据的特定版本代码之间的关联往往已经减弱。持续集成（CI）等软件工程实践并未被用于将输出成果与生成它们的代码关联起来，部分原因是人们需要掌握另一种规范语言，此外，从本地开发环境导出大型数据集也存在困难

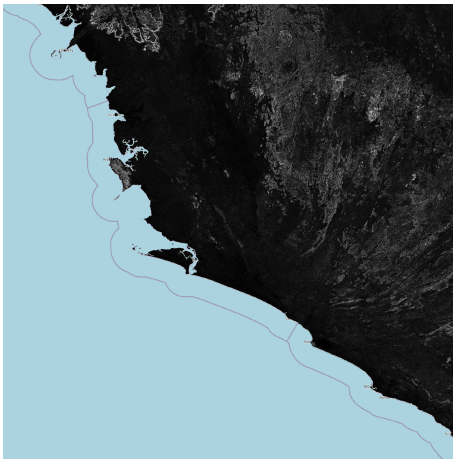


图 2：展示了使用 gdaldem tri 命令计算的地形崎岖度指数（TRI）在 GDAL 3.2 和 3.3 版本之间的差异。理想情况下，差异应为零（全部为黑色）。

团队清楚地认识到，结果与代码之间必须保持紧密关联。

在计算机科学领域，人们知道可以使用形式化方法来评估代码是否按预期运行，或者至少通过一套完善的单元测试来增强信心。然而，对于处理这些大型数据集的研究人员来说，情况并非如此，这同样是因为在发表之前，研究工作具有探索性质。由于简单的编程错误很容易导致论文需要撤回，缺乏系统化测试令整个团队感到担忧 [74]。数据集的规模也意味着，在以 <100m 的分辨率处理数万个物种并制作地图时，可视化工具无法轻松检查中间结果。

2.3 依赖项中的不确定性

即使格外谨慎，如果没有可复现的计算环境（例如使用 Nix [43] 这样的系统），在传统操作系统中，可能影响最终结果的众多因素也会变得难以管理。例如，可以看看广受欢迎的地理空间数据处理库 GDAL [8]。图 2 展示了使用相同的 命令行计算地形崎岖度指数（TRI）时，GDAL 3.2 与 GDAL 3.3 得出的数据之间的差异。[3] 这种差异源于默认算法的变更。虽然这一变更是必要的，但在本例中，它导致像素之间的差异最大达到 372m。

同样，硬件版本的细微差异也会导致结果发生变化，即使同一条处理流程在不同机器上运行也是如此。例如，在处理生物多样性地图时，我们发现栖息地移除情景中出现了意料之外的微小增幅（按理说，这种情景中只应有物种损失）。追查原因后，我们一路深入探究，最终发现，某些新版本的 AMD EPYC 处理器在数组运算与标量运算之间存在舍入差异；这些处理器新增了 512 位 SIMD 指令，其舍入特性与较窄位宽的 SIMD 指令和标量运算不同。单独来看，这些

[3]源代码：https://github.com/quantifyearth/gdal-tri-diff

舍入误差微不足道（低于 CPU 浮点数机器精度），但在数千个物种上累加后，便产生了显著影响。所有生物多样性情景都受到了该误差的影响，但在大多数预期会出现损失与增益的情景中，这一问题被掩盖了。庞大的数据量使得我们难以细致检查中间结果。幸运的是，我们运行了一个完整的损失情景，发现了这一差异；团队也具备足够的计算机科学知识，得以追溯软件和硬件中的真正根源，并与上游团队讨论这一问题。[4]

依赖关系中的不确定性也延伸到了用于地理空间分析的热门平台，例如 Google Earth Engine (GEE)——一个集代码和数据于一体、用于开展“行星尺度分析”的平台 [54]。它产生了广泛的积极影响，使许多生态学家得以开展原本难以实现的有用分析。一个例子是医疗机构出行时间的全球地图，可在生态分析中用作“偏远程度”的替代指标 [104]。然而，我们开始遇到 GEE 作为闭源平台的局限；例如，我们无法以能够确保数据不可变、因而可复现的方式，充分指定特定的历史数据集（如前述的 JRC TMF 数据集）。相反，我们依赖可能发生变化的不透明标识符；若这是一个开放平台，这种行为本可以得到验证。

这引发了人们对计算平台长期存续性的担忧，而保护研究工作正是依托该平台开展的。有些干预措施通常会规划数十年（树木寿命很长！），并涉及数千人以及分布在不同地区的数十个组织。像 GEE 这样的平台——无论短期内多么有效——都依托于一家大型企业，而该企业有关闭非核心产品的先例 [52]。GEE 中的许多数据集也经过处理，以便于使用，但处理这些数据集所采用的方法并未发布，因此一旦数据集经过改编，就难以迁移离开 GEE。因此，我们对 GEE 的使用可谓喜忧参半——它非常适合对大规模数据集进行交互式科学探索，但无法扩展到满足我们对科学可重复性（这是全面同行评审过程的关键）和长期承诺的要求

以确保几十年后仍能解释我们的研究结果。

2.4 政策不确定性

基于劣质数据制定政策并非新鲜事。例如，Herndon 等人在 Reinhart 和 Rogoff 关于高公共债务与经济增长之间联系的论文中，发现了选择性排除数据、电子表格编码错误和不恰当的统计方法；这篇论文曾被颇具影响力的政治人物广泛引用 [60]。然而，尽管环境科学长期以来一直在应对不确定性（例如生物多样性分析中的抽样误差 [49], ），一种纳入大型数据集并经过多个计算阶段的科学方法，可能会加剧该领域原本就存在的方法学不确定性。

图 3 展示了我们所应对的生态挑战（§1.1）中这一过程如何发生。计算流水线还需要将源数据的不确定性、不同算法带来的不确定性，以及对政策制定者在推行环境变化时所用数据产品的追踪纳入考量

[4]参见 https://github.com/numpy/numpy/issues/25269

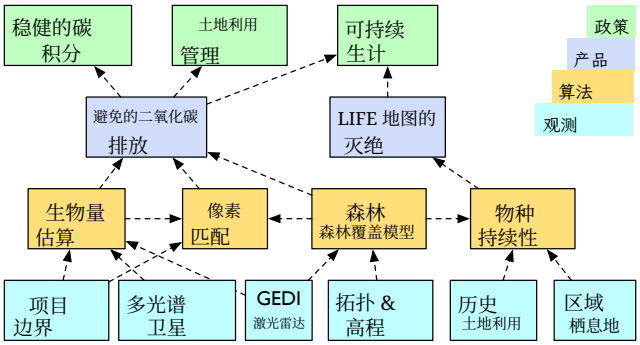


图 3：针对 §1.1 中两种动机的示例性数据流水线，以及可由结果推导出的政策情景。有关更完整的规范，请参见 [30]或 [48]。

。因此，接下来我们将探讨行星计算如何帮助人们信任基于全球数据分析结果的决策，以及如何构建一个充分顾及政策制定者需求、并顾及希望质疑拟议政策者需求的系统。

3 行星计算的范围界定

为了紧急加快应对气候危机的行动，我们倡导建立一个以环境科学家和政策制定者为主要用户的系统，从而促进围绕数据驱动的情景开展协作（§1.1）。

3.1 能力

环境科学家需要访问由以下内容组成的**大规模输入数据集**：（*i*）来自卫星的原始观测数据 [45]，其规模达到 PB 级，或直接地面测量数据 [80, 100]；（*ii*）通过算法转换或基于人工智能的推断得到的派生数据源；以及（*iii*）由第三方或先前运行产生的既有结果。随后，程序员在这些数据集上**描述计算过程**，该过程：（*i*）可以是算法式的，也可以基于机器学习，并结合使用 CPU 和 GPU；（*ii*）需要支持自动扩缩，以便先进行本地开发，再开展全球分析；以及（*iii*）可以由非计算机科学专家表达，理想情况下通过可视化界面 [33], ，使用 R 或 Python 等语言，甚至借助大语言模型 [86]。这些**派生产品**提供更高层次的洞见，必须长期存档，同时还应：（*i*）跟踪输入数据的数据溯源，并对输出结果执行隐私约束；（*ii*）在能够访问源数据时可由独立方验证；以及（*iii*）进行增量重新计算，以支持交互式探索和本地数据的纳入。这些要求对应于图 4 所示的数据流水线，该流水线包含四个阶段：

- **摄取**：即获取遥感数据集。发布者通常通过临时搭建的 HTTP/FTP 服务器提供这些数据，因此需要定期轮询并下载脚本。数据格式各不相同（例如 GeoJSON、TIFF 文件），并且由于数据量庞大，通常需要将其标准化为一种格式，例如空间索引列式存储。数据提供者很少对数据实施系统化的版本控制（§2.1）。
- **转换**：指针对超大型数据集进行规模化处理的动态计算流水线，通常使用多种语言（常见的有 Python、R、Julia 或 Fortran）编写。算法必须

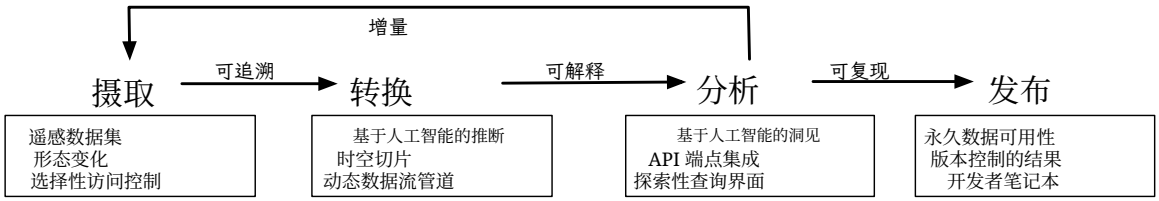


图 4：行星计算引擎的理想数据流管道

能够扩展，以便生态学家先在本地进行实验，再将其工作转移到高性能计算集群。此外，机器学习常用于插补稀疏数据集，并通过时空切片进行优化，以聚焦于目标区域。

- **分析**：是外部系统可使用的外部接口，可通过 API 驱动的端点（例如网络钩子）、查询接口或人工智能驱动的语言模型访问。

- **发布**：是对可复现结果的长期存储（例如用于科学出版物）。此外，提供在线笔记本也很有用，可为非专业用户提供一个快速开发环境，其中包含所有数据、代码和工具。

3.2 能动性

除非该工作流程不仅满足大型机构的需求，也能让所有可能受到任何政策决策影响的人使用，否则它就毫无用处。因此，任何行星计算系统都必须满足以下要求：

- **可扩展性：** 用户（而非平台维护者）能否通过导入库或工具，为系统添加新功能，以支持新技术？用户能否轻松地将该系统整合到现有工作流程中？

- **可访问性：** 用户能否无需请求许可（从而避免可能被拒绝）就访问该系统？任何人是否都能使用该系统，而无需具备专业的计算机系统知识？它是否支持科学家最常用的工作模式，例如渐进式数据探索和探索性研究？

- **保密性：** 在涉及数据敏感性的情况下，已发布的结果能否以受限方式共享，包括后续转发共享？

鉴于许多保护行动发生在全球南方，而相关计算机系统的大部分开发工作则在全球北方进行 [67]，因此，在行星计算的背景下，能动性的考量尤为重要。这要求行星计算认真对待公平性、偏差和问责制问题，并审视各种计算方法，例如：机器学习模型的数据如何收集，以及如何公平地训练和部署这些模型 [62, 73]，；自动化决策在任何政策决策中发挥什么作用 [103]；以及《遥感》的可访问性和使用所涉及的伦理考量 [108]。

3.3 可存续性

系统还必须确保其结果能够持久留存，因为这些结果将在未来很长一段时间内为科学和政策提供依据。

- **可追溯性：** 结果必须能够通过密码技术追溯至其数据和代码输入，同时尊重

并非所有输入都能直接披露，可能是由于治理限制或敏感性（例如，受威胁物种会成为偷猎目标 [107]）。必须在中间数据集中追踪这些限制，并确保输出符合隐私要求 [47]。

- **可解释性：** 为了依据计算结果制定切实可行的政策，这些结果以及产生这些结果的算法必须能够让一群非专业政策制定者理解。这要求清晰、简洁地表达算法，并对机器学习模型进行自省。

- **可复现性：** 系统产生的结果必须能够由他人独立重新运行；对于包含 CPU/GPU 运算的异构工作负载，以及内部算法会随版本发布而变化的库而言，这一点出奇地难以实现。 [105]

3.4 当前技术水平

我们通过两种策略调查现有系统：使用能够尽可能覆盖我们所确定生命周期各个环节的端到端平台，或者利用现成组件组建定制系统。我们的调查总结见表1。

现有的端到端解决方案。谷歌的 Earth Engine（GEE）[53]和微软的 Planetary Computer（MPC） [14] 都提供面向生态学家的云端端到端解决方案。这两个平台已经摄取并提供了热门卫星影像，可供用户交互式探索现有数据集，并能使用 Python 和 JavaScript 等熟悉的语言轻松处理数据，同时隐藏扩展算法所涉及的计算复杂性。两者的做法有所不同：GEE 侧重于通过自有的专有用户界面和库提升易用性，而 MPC 则侧重于提供一个开放平台，以便大规模运行现有工具。两个平台的端到端能力都表现出色，但在非技术性要求方面稍逊一筹。两者均需通过申请才能访问，且不保证一定获准，因此也无法保证任何工作都能复现。为 GEE 编写的代码无法直接在其他平台上运行，与 MPC 更开放的做法相比，这也降低了可复现性。相较之下，MPC 向用户暴露系统方面的问题，程度远高于 GEE；GEE 则使用自有框架对内存和并行性进行抽象。两个系统都无法在保护私有数据的同时支持可追溯性和可解释性。

平台。多年来，开源社区开发了许多组件，用于构建定制解决方案（此处仅介绍其中一小部分）。这些方案通常会让用户面对更多系统层面的事项，也需要更多管理工作。解决全球环境问题所需数据集的规模和复杂度已达拍字节级，而且还在不断增长，这意味着

		端到端能力				能动性			可存续性		
	描述	**摄取**	**转换**	**分析**	**发布**	**可扩展**	**可访问**	**保密性**	**可追溯**	**可解释**	**可复现**
Cloud	Earth Engine [53]	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗
	MPC [14]	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓
	ArcGIS [4]	✓	✗	✓	✗	✗	✓	-	✗	✗	✗
	AWS [15]	✓	✓	✗	✗	✓	✗	-	-	✓	-
	Snowflake [18]	✗	✓	✗	✗	✓	✗	-	-	✓	-
	Hugging Face [11]	✗	✓	✓	✗	✗	✓	-	-	✗	✗
	Zenodo [2]	-	-	-	✓	✗	✓	✗	✓	✓	-
	DataDryad [7]	-	-	-	✓	✗	✓	✗	✓	✓	-
	GitHub [10]	-	✗	-	✓	✗	✗	✓	✓	✓	-
	OneDrive [17]	-	-	-	✓	✗	✗	✓	✗	-	-
	Google 云端硬盘 [16]	-	-	-	✓	✗	✗	✓	✗	-	-
Framework	Apache Sedona [76]	✗	✓	✓	✗	✗	✓	-	✗	✓	✓
	Apache Hadoop [3]	✗	✓	✓	✗	✗	✗	-	✗	✓	✓
	Pachyderm [79]	✗	✓	✓	✗	✗	✗	-	✗	✓	✓
	PostGIS [1]	-	✗	✓	✗	✓	✗	-	✗	✓	✓
	TileDB [21]	-	✓	✓	✗	✗	✗	-	✗	✓	✓
	Jenkins [12]	-	✗	-	✓	✓	✗	✗	✓	✓	✓
Libraries	GDAL [8]	-	✗	✗	-	✗	✓	-	-	✓	✓
	CuPy [5]	-	✗	✗	-	✗	✗	-	-	✓	✓
	Dask [6]	-	✓	✗	-	✓	✓	-	-	-	✓
	TerraLib [20]	-	✓	✓	-	✗	✗	-	-	✓	✓
	PyTorch [82]	-	✓	✓	-	✗	✓	-	-	✗	✓
Tools	R [25]	✓	✗	✓	-	✓	✓	-	✗	✓	✓
	QGIS [24]	-	✗	✓	-	✗	✓	-	-	-	-
	STAC [19]	✓	-	-	✓	✓	✓	-	✗	✓	✗

表1：环境科学中现有工具的比较（破折号表示不在讨论范围内的领域）

存储和处理成本高昂。许多人转而使用通用云平台（GPCP），将其作为经济且可扩展的替代方案，例如亚马逊云服务（AWS）[15] 和 Snowflake [18]。有些平台（如 AWS）已提供可直接使用的开放科学数据集，但这种情况并不常见。对于没有额外培训的非技术人员而言，通用云平台更难上手：配置基础设施不可避免地会让人接触到系统的复杂性。用户需要确保其架构与基础设施配置 CPU/GPU/内存的机制相匹配，而 GEE 能很好地隐藏这些复杂性。

另一种选择是使用本地部署解决方案，在本地硬件上搭建可扩展的计算平台，并采用 Hadoop [3] 或 Pachyderm [79], 等通用框架，以及 Sedona [76]等面向地理空间的框架。这些框架负责调度存储和计算资源，但确实需要进行技术配置和维护。这些系统会偏向某一类算法：例如，Hadoop 的 MapReduce 会要求用户按照其模式组织代码，而这对深度学习 [77]等算法来说效率不高。还有各种支持 GIS 的数据库（如 PostGIS [1] 和 TileDB [21]），可简化特定类别地理空间数据的处理；这些工具配置完成后功能强大，但要掌握如何运行能够处理海量数据的高性能数据库，本身就是一项专业工作。

为了实现增量处理，可以设置持续集成（CI）系统（如 Jenkins [12]）来帮助运行端到端流水线，但要让它们在不同系统间具备可复现性和可移植性却相当困难；此外，由于其脆弱性和对最终用户不够友好，这类系统通常由专门的 DevOps 专业人员管理。GitHub [10] 确实提供了一些易于管理、类似 CI 的功能，设置起来更简单，但并非用于行星尺度的计算。

在发布方面，代码和数据的托管服务（例如 GitHub、Zenodo [2] 和 DataDryad [7]）既是第一手来源，又支持版本管理，因此能够提供一定程度的可追溯公开引用。不过，由于尚未采用去中心化信息流控制（DIFC）[110]，要确保私有数据不会通过这些服务泄露十分困难。值得注意的是，尽管学术界人士已采用这类数据门户，但美国国家航空航天局和欧盟联合研究中心等主要机构仍自行发布数据，因而出现了 §2.1 所述的问题。通常，这些平台仅用于发布最终成果；但实际上，生态学家经常需要与其他机构的同事分享尚在进行中的成果，而地理空间数据集的大小很容易超过大多数人云服务商文件存储的容量限制（例如谷歌云端硬盘或 Microsoft OneDrive [16, 17]），因此人们会采用临时解决方案，导致可追溯性受限。

数据转换。 要在地理空间数据集所涉及的规模下高效编写代码，通常需要具备计算机科学专业知识，才能使用 GPU 等硬件。在这一领域中常用的传统工具（如 Python 或 R）无法充分抽象内存管理和任务并行，使用户仅凭自身就能高效地表达转换与分析算法；不过，也有一些分层框架可以提供帮助。例如，PyTorch [82] 提供了一种“务实”的编程接口，通过抽象并行性和 GPU 的使用来简化深度学习。Dask [6] 对基于 NumPy 的数值工作负载也能实现类似功能，可将其分布到多个核心或集群上。将计算表达与数据资源管理分离的能力，是帮助用户处理这些数据集的关键。

框架只能为数据生命周期提供部分解决方案，而其边界则暴露出系统层面的关注事项。例如，CuPy [5], 这款基于 CUDA 的图形处理器数值计算框架 NumPy 的即插即用替代品 [58], 并未考虑针对特定数据集，使用图形处理器是否比使用 CPU 更高效：它把这一判断留给用户，并假设用户熟悉相关细节，例如需要在图形处理器上执行足够多的工作，才能抵消内存传输成本。框架之间的交互也可能

导致显著的性能问题：GDAL [8],这个用于读取和写入地理空间数据的标准库，通常会建议采用高效的分块大小来读取 GeoTIFF 文件，但这种分块方式与 CuPy 的预期不兼容；除非用户知道要先重新调整数据维度，否则图形处理器上只会有一个算术逻辑单元（ALU）投入使用。就不确定性而言，尽管在 CPU 和图形处理器之间迁移数据很方便，但根据实验的敏感程度，结果也会有所不同，因为图形处理器的浮点行为与 CPU 的有所差异 [105]。

分析与可视化。 使用该领域通用的数据格式标准（例如 GeoJSON、GeoTIFF 等），便能更轻松地与现有分析工具互操作，例如使用 Leaflet [22] 和 QGIS [24] 进行探索性可视化，以及使用 R [25] 进行统计分析。这些工具对于生态学和地理空间分析至关重要，有助于进行可视化调试和统计调试。例如，在地图上绘制点位，能让人更容易发现纬度和经度是否存储颠倒。将这些工具更深入地集成到计算流水线中，有助于改进对计算进度和正确性的检查过程。数据集可视化极其重要，我们曾观察到这样的情况：在必须权衡选择最优地理空间数据格式（例如，考虑到地图投影的局限性而不使用栅格数据，转而使用等面积六边形网格）时，生态学家会退而使用自己能够可视化的格式。

发布与可追溯性。 尽管科学领域长期以来一直关注数据沿袭 [89],，常见的地理空间数据存储格式却不支持版本管理，更不用说可追溯性了。GeoTIFF 是卫星影像数据的常用格式，但其标准中没有任何与版本管理相关的字段 [41],，即将推出的 geozarr 等格式也同样如此 [9]。Arrow 和 Parquet 格式允许使用嵌入式元数据记录数据沿袭，但没有任何强制字段或通用约定。时空资产目录（STAC） [19] 是一种基于 JSON、用于发布地理空间数据集的格式，许多服务都在使用它，包括 MPC。STAC 扩展可以记录数据的处理方式（有助于提高可复现性），但没有要求提供所引用数据文件的版本信息；而且，可选的版本信息也不包含任何内容（例如数据哈希值）来确保版本

与特定数据实例相关联。

当前的端到端解决方案因缺乏可复现性、可追溯性和可解释性而令人失望——这些都是应对气候危机、确保做出正确决策的关键。尽管可以将现成组件组装成具备这些特性的端到端系统，但这样做会使用户面临需要技术领域专家才能处理的系统相关问题，从而阻碍这些系统原本能够推动的环境干预措施得到更广泛的应用。

4 开放研究挑战

我们已经指出，实施数据驱动的环境政策制定（§1.1）所面临的难题，有赖于弥补若干缺失环节（§3）；而单一的计算框架或云平台无法解决所有问题。问题的核心在于，全球观测数据、算法、派生产品以及最终的政策制定（§2）层层叠加，带来了大量不确定性。我们

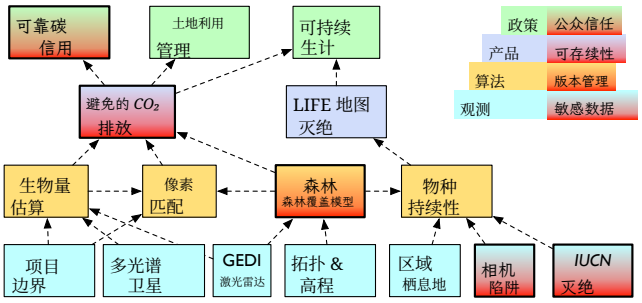


图 5：图 3 所示的流程管线，重点标示了围绕隐私、版本管理和长期存续性的计算研究挑战。要实现可信的政策结果，计算挑战与环境挑战彼此依存。

现在来探讨计算机科学领域的研究挑战：若能将这些挑战联系起来，便可为构建可持续行星计算模式铺平道路。图5总结了我们在构建自己的数字化流水线时发现的总体机遇。我们必须处理敏感数据观测（§4.1），继而为这些大型环境建立版本管理机制（§4.2），确保它们能够持续数十年（§4.3），最终实现我们的目标：建立对数据驱动政策制定的信任（§4.4）。

4.1 协调隐私与透明度

从数据保密性的角度来看，气候变化监测和环境保护都面临着巨大挑战，而向公众完全开放数据可能弊大于利。任何数据收集过程都涉及多个利益相关方；有许多案例表明，出于经济动机的恶意行为者会利用环境数据造成危害（§2.1）。

因此，从计算机科学的角度来看，我们需要从一开始就在任何行星计算基础设施中构建强有力的安全和隐私控制措施。虽然大多数现有系统都以牺牲隐私为代价，优先考虑开放性或透明度，但我们已经指出，过度开放会让不良行为者摸清如何钻系统的空子 [84]。为缓解这一问题，我们提出一种可能的做法：遵循“最终开放原则”，先对数据实施禁运，之后再将其公开 [69]。此外，差分隐私 [47] 和去中心化信息流控制 [110] 可以在保护数据隐私的同时，允许一定程度的透明度，即使是在敏感的初期阶段。能够跨多个用户遵守隐私约束的完整查询引擎也是一个新兴领域，例如多元宇宙数据库架构 [70]。这些能力可以为科学家和政策制定者提供所需工具，帮助他们平衡隐私与透明度。此外，部分披露源数据或许也有帮助，既可避免完全披露，又允许获准访问源信息的第三方随后进行审计并独立验证数据。

4.2 代码和数据中的版本化不确定性

我们观察到（§2.1），即使是看似静态的过往观测数据集，也可能随着时间推移而发生变化，因为处理这些数据集的算法会不断改进。图 5 展示了这一点如何影响下游政策制定，例如碳信用，从而对预期的保护成效造成不利影响。因此，我们

需要在代码、数据集及其处理环境中建立实用的版本控制方案，使政策制定者能够预测即将发生的变化所带来的影响。

目前已有针对气候建模所涉及数据格式进行结构化的倡议（例如 CMIP6 [28]）。他们所确定的提高效率的关键，是建立一种通用数据格式，避免在后续计算流水线中反复进行数据转换。尽管现有许多工作流引擎，但由于动态数据流图的表达能力各不相同，没有任何引擎能够支持可移植且高效的流水线组合；其范围从图灵完备引擎 [77] 到不支持数据依赖控制流的静态图 [3]。协议驱动方法可以规定库接口，以便使用不同语言表达多个代码阶段、编排容器构建、GPU 执行，并从外部系统（例如 GEE）检索结果，而这些系统可能没有可用的本地副本。与摄取一样，系统必须在分布式节点之间传播数据集的保密信息 [109]。因此，行星计算流水线需要在加入版本管理的同时，兼顾可追溯性、隐私和共享；值得研究的方向包括：*(i)* 使用按预期访问模式排序的压缩列式格式 [85]，并采用可选的有损压缩 [13, 42]，以优化向 CPU/GPU 的流式处理 [97]；*(ii)* 使用去中心化信息流控制（DIFC）标签标记输入源 [110]，确保可以在分布式流水线或查询引擎中的任意位置执行访问控制检查 [70]；*(iii)* 将数据哈希为空间数据块，以便随后对数据子集进行版本控制 [50]。

4.3 构建可持续存续的数据产品

乍看之下，像 GEE 这样的云平台解决了短期的行星计算问题（表1），并且已得到广泛应用 [78]。挑战在于如何让它在未来几十年持续存续。很难相信一个封闭平台——无论其工程设计多么出色——能够汇集全球数据，并成为政策制定情报的中央来源。过去，计算机领域曾齐心协力，为新兴技术构建联邦式测试平台，例如 PlanetLab [83], CloudLab [46] 和 Emulab [59]。自互联网诞生以来，我们积累了共同的知识，包括如何培育和壮大开源社区 [95], 、整理开放数据集 [81], ，以及推动互联网工程任务组（IETF）等共享治理机制 [56]。

行星计算需要同样雄心勃勃的努力，以确保 GEE 等平台能够超越任何单一组织或服务提供商的存续期限，持续发展。目前尚不清楚，解决之道是否在于构建一个完全联邦式系统，但朝这个方向迈出的第一步，是定义协议，使独立实现能够互操作。类似 IETF 的流程将涉及为图4中的每个流水线模块制定协议。然而，在大规模去中心化访问控制和加密方面，研究文献仍存在显著空白。

4.4 非专业用户的可及性

科学家和政策制定者是行星计算存在的根本原因 和主要用户，不应被视为事后才考虑的对象；计算必须用于实现积极的最终成果，惠及 [88] 受气候变化影响最大的人群。数据管道的目的

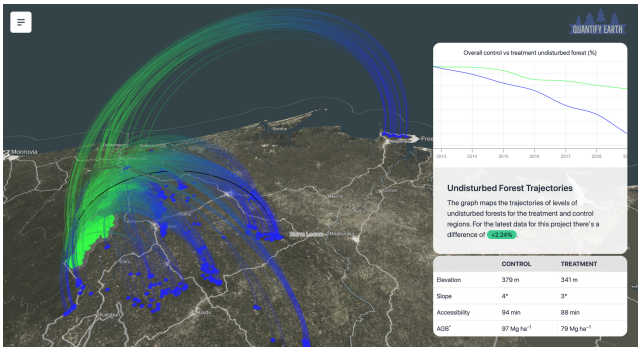


图 6：通过反事实像素匹配，评估避免毁林项目的额外性（§1.1）的可视化。绿色像素来自该项目，并与蓝色区域中的相似地点相匹配。

（图 5）是赋能我们分析复杂的反事实情景，从而制定艰难的政策决策。因此，我们需要更简单的查询接口来访问由专家驱动的底层数据产品，同时保留这些产品的数据溯源和可解释性，以维持人们对最终决策的信任。在我们研究热带森林和碳信用的过程中 [93],，最具影响力的接口之一是可视化探索器，它逐步展示了其中涉及的反事实分析（图 6）。这个动态接口最初是用于调试算法的原型，却很快成为向生态学家和金融从业者解释复杂算法的工具 [30] 。值得探索的研究方向包括用于组合算法和数据构件的“无代码”接口 [38], ，以及基于大语言模型的自然语言协作助手，通过对话式探索生成可重复使用的输出 [37]。

另一个重要考量是能动性和独立性——政策制定者必须能够将公共数据集与自己的私有数据（通常针对特定辖区）混合使用，而不将其透露给第三方 [91]。安全地混合此类数据，需要仔细分析，以确定可能对其在特定区域的适用性产生不利影响的偏差。数据集数据表 [51] 等方法或许有助于建立相应标准，使非专业用户能够确保其决策所依据的数据来源适合本地情境。

5 结论

我们结合计算机科学家和生态科学家的跨学科团队经验，阐述了我们对行星计算的愿景：将其作为加速协作式数据驱动环境政策制定的工具。本文概述的研究方向绝非面面俱到，主要聚焦于计算机系统研究。不过，我们希望本文对行星计算的阐释能成为一个有益的着力点，帮助计算机科学界弥补数据驱动环境科学应用中的关键知识空白，并最终形成对行星健康状况准确且具代表性的认知，为制定保护和改善自然生态系统的政策提供依据，惠及未来数十年。

参考文献

[1] https://qgis.net/. [2] 关于 Zenodo。https://about.zenodo.org/. [3] Apache Hadoop。https://hadoop.apache.org. [4] ArcGIS Online。https://www.esri.com/en-us/arcgis/products/arcgis-online/overview.[5] CuPy：面向图形处理器的 NumPy 和 SciPy。https://cupy.dev/. [6] Dask | 扩展你喜爱的 Python 工具的规模。https://dask.org/. [7] Dryad——发布并保存你的数据。https://datadryad.org/. [8] GDAL。https://gdal.org/. [9] Geozarr-spec。https://github.com/zarr-developers/geozarr-spec. [10] GitHub。https://github.com/. [11] Hugging Face。https://huggingface.co/. [12] Jenkins。https://www.jenkins.io/. [13] 有限误差栅格压缩（LERC）。https://github.com/Esri/lerc. [14] Microsoft Planetary Computer。https://planetarycomputer.microsoft.com/. [15] 亚马逊云服务概览。https://docs.aws.amazon.com/whitepapers/latest/aws-overview/introduction.html.[16] 个人云存储和文件共享平台——谷歌。https://www.google.com/intl/en-US/drive/. [17] 个人云存储——Microsoft OneDrive。https://www.microsoft.com/en-gb/microsoft-365/onedrive/online-cloud-storage. [18] Snowflake。https://www.snowflake.com/. [19] STAC：时空资产目录。https://stacspec.org/. [20] TerraLib5。http://www.terralib.org/. [21] TileDB——通用数据库。https://tiledb.com/. [22] Leaflet：一款开源 JavaScript 库，可创建适用于移动设备的交互式地图。https://leafletjs.com，2023年。[23] TMF 2022产品中整合的更新列表。https://forobs.jrc.ec.europa.eu/TMF/data#update，2023年。[24] QGIS——免费开源地理信息系统。https://www.qgis.org/，2023年。[25] R 统计计算项目。https://www.r-project.org，2023年。[26] Achard, F., Beuchle, R., Bodart, C., Brink, A., Carboni, S., Eva, H., Mayaux, P., Raši, R., Simonetti, D., and Stibig, H.-J. 全球尺度森林覆盖监测：欧盟联合研究中心的方法。载于第 33 届 ISRSE 会议论文集（2009年），第 4 卷。[27] Angelsen, A. 实现 REDD+：国家战略与政策选项。Cifor，2009年。[28] Balaji, V., Taylor, K. E., Juckes, M., Lawrence, B. N., Durack, P. J., Lautenschlager, M., Blanton, C., Cinqini, L., Denvil, S., Elkington, M., et al. 支持 CMIP6 的全球数据基础设施需求。《地球科学模型发展》11, 9（2018年），3659–3680。[29] Balmford, A., Brancalion, P. H. S., Coomes, D., Filewod, B., Groom, B., Guizar-Coutiño, A., Jones, J. P. G., Keshav, S., Kontoleon, A., Madhavapeddy, A., Malhi, Y., Sills, E. O., Strassburg, B. B. N., Venmans, F., West, T. A. P., Wheeler, C., and Swinfield, T. 碳信用可信度威胁森林。《科学》380, 6644（2023年），466–467。[30] Balmford, A., Coomes, D., Hartup, J., Jaffer, S., Keshav, S., Lam, M., and Wheeler, C. Pact 热带湿润森林认证方法。技术报告，2023年。[31] Balmford, B., Green, R. E., Onial, M., Phalan, B., and Balmford, A. 土地隔离的效果可以有多不理想，才会使土地共享对野生物种更有利？《应用生态学杂志》56, 1（2019年），73–84。_预印本：https://onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2664.13282. [32] Barnes, N. 发布你的计算机代码：它已经足够好了。《自然》467, 7317（2010年10月），753–753。[33] Blinn, A., Moon, D., Griffis, E., and Omar, C. 面向交互式编程助手的综合性以人为中心架构。载于《2022 IEEE 可视化语言与以人为中心计算研讨会（VL/HCC）》（2022年），第 1–5 页。[34] Brooks, T. M., Pimm, S. L., Akçakaya, H. R., Buchanan, G. M., Butchart, S. H., Foden, W., Hilton-Taylor, C., Hoffmann, M., Jenkins, C. N., J oppa, L., et al. 陆地区域栖息地面积（AOH）的测量及其对 IUCN 濒危物种红色名录的用途。《生态学与进化趋势》34, 11（2019年），977–986。[35] Buchwitz, M., Reuter, M., Schneising, O., Bovensmann, H., Burrows, J. P., Boesch, H., Anand, J., Parker, R., Detmers, R. G., Aben, I., et al. 哥白尼气候变化服务（C3S）全球大气二氧化碳和甲烷卫星观测。《航天科学与技术进展》1（2018年），57–60。[36] Burgess, S., Kranz, M., Turner, N., Cardell-Oliver, R., and Dawson, T. 利用无线传感器技术推进森林生态学和农业研究。《农业与森林气象学》150, 1（2010年1月），30–37。[37] Busch, D., Nolte, G., Bainczyk, A., and Steffen, B. 将 ChatGPT 纳入流程：领域特定建模语言的自然语言扩展。载于《弥合人工智能与现实之间鸿沟国际会议》（2023年），Springer，第 375–390 页。[38] Cai, Y., Mao, S., Wu, W., Wang, Z., Liang, Y., Ge, T., Wu, C., You, W., Song, T., Xia, Y., Tien, J., and Duan, N. 低代码大语言模型：基于大语言模型的可视化编程，2023年。[39] Cooke, S. J., Nguyen, V. M., Kessel, S. T., Hussey, N. E., Young, N. 和 Ford, A. T. 保护与管理领域动物追踪前沿的棘手问题。《保护生物学》31，5（2017年），1205–1207。[40] Culina, A., van den Berg, I., Evans, S. 和 Sánchez-Tójar, A. 生态学中代码的可用性低：呼吁立即采取行动。《PLOS Biology》18，7（2020年7月），1–9。[41] Devys, E., Habermann, T., Heazel, C., Lott, R. 和 Rouault, E. OGC GeoTIFF 标准。[42] Di, S. 和 Cappello, F. 使用 SZ 实现快速、有误差界限的有损高性能计算数据压缩。载于2016年IEEE国际并行与分布式处理研讨会（IPDPS）（2016年），第730–739页。[43] Dolstra, E. 和 de Jonge, M. Nix：用于软件部署的安全、无政策系统。载于第18届大型安装系统管理会议（LISA 04）（亚特兰大，佐治亚州，2004年11月），USENIX协会。[44] Dubayah, R., Armston, J., Healey, S. P., Bruening, J. M., Patterson, P. L., Kellner, J. R., Duncanson, L., Saarela, S., Ståhl, G., Yang, Z. 等。GEDI开启太空生物量推断的新时代。《环境研究快报》17，9（2022年），095001。[45] Dubayah, R., Blair, J. B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S. 等。全球生态系统动态调查：对地球森林和地形进行高分辨率激光测距。遥感科学1（2020年），100002。[46] Duplyakin, D., Ricci, R., Maricq, A., Wong, G., Duerig, J., Eide, E., Stoller, L., Hibler, M., Johnson, D., Webb, K. 等。CloudLab的设计与运行。载于2019年USENIX年度技术会议（USENIX ATC 19）（2019年），第1–14页。[47] Dwork, C. 和 Roth, A. 差分隐私的算法基础。Now Foundations and Trends，2014年。[48] Eyres, A., Ball, T., Dales, M., Swinfield, T., Arnell, A., Baisero, D., Durán, A. P., Green, J., Madhavapeddy, A. 和 Balmford, A. LIFE：定量绘制土地覆盖变化对全球灭绝影响的指标，2023年11月。[49] Freckleton, R. P., Watkinson, A. R., Green, R. E. 和 Sutherland, W. J. 普查误差与密度制约的检测。《动物生态学杂志》75，4（2006年7月），837–851。[50] Gazagnaire, T., Chaudhry, A., Madhavapeddy, A., Mortier, R., Scott, D., Sheets, D., Tsipenyuk, G. 和 Crowcroft, J. Irminsule：分支一致的分布式库数据库。载于第4届ACM OCaml 用户与开发者研讨会（瑞典哥德堡，2014年9月），ACM。[51] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D. 和 Crawford, K. 数据集数据表。《ACM 通讯》64，12（2021年），86–92。[52] Gerken, T. 玩家告别谷歌Stadia，平台关闭，2023年。访问日期：2023年10月23日。[53] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. 和 Moore, R. Google Earth Engine：面向所有人的行星尺度地理空间分析。《环境遥感》202（2017年12月），18–27。[54] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. 和 Moore, R. Google Earth Engine：面向所有人的行星尺度地理空间分析。《环境遥感》202，18–27，2017年。[55] Grover, J. 一次超越——公众的知情权——无线电遥测。《Fisherman》26，5（2001年），18–22。[56] Harcourt, A., Christou, G. 和 Simpson, S. IETF、W3C和IEEE的内部治理：结构、决策与国际化。载于《互联网治理中的全球标准制定》。牛津大学出版社，2020年。[57] Harfoot, M. B. J., Johnston, A., Balmford, A., Burgess, N. D., Butchart, S. H. M., Dias, M. P., Hazin, C., Hilton-Taylor, C., Hoffmann, M., Isaac, N. J. B., Iversen, L. L., Outhwaite, C. L., Visconti, P. 和 Geldmann, J. 利用IUCN濒危物种红色名录在全球尺度绘制陆生脊椎动物所面临的威胁。《自然·生态与进化》5，11（2021年11月），1510–1519。编号：11；出版商：Nature Publishing Group。[58] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. 和 Oliphant, T. E. 使用NumPy进行数组编程。《自然》585，7825（2020年9月），357–362。编号：7825；出版商：Nature Publishing Group。[59] Hermenier, F. 和 Ricci, R. 如何构建更好的测试平台：Emulab十年网络实验的经验教训。载于《测试平台与研究基础设施：网络与社区的发展》（柏林、海德堡，2012年），T. Korakis、M. Zink 和 M. Ott 编，《计算机科学、社会信息学与电信工程讲义》，Springer，第287–304页。[60] Herndon, T., Ash, M. 和 Pollin, R. 高公共债务是否始终会抑制经济增长？对Reinhart和Rogoff的批评。《剑桥经济学杂志》38，2（2013年12月），257–279。[61] Holcomb, A., Mathis, S. V., Coomes, D. A. 和 Keshav, S. 使用GEDI激光点云和森林变化图评估森林恢复的计算工具。遥感科学8（2023年），100106。

[62] Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M. 和 Wallach, H. 改善机器学习系统中的公平性：行业从业者需要什么？载于2019年CHI人机交互因素会议论文集（美国纽约州纽约市，2019年），CHI '19，计算机协会，第1–16页。[63] Hu, Q. 和 Han, Z. 近几十年中亚沙漠气候的向北扩张。《地球物理研究快报》49，11（2022年），e2022GL098895。 _电子版：https://onlinelibrary.wiley.com/doi/pdf/10.1029/2022GL098895。[64] Jakob, C., Gettelman, A. 和 Pitman, A. 气候建模实现运营化的必要性。《自然·气候变化》13，11（2023年11月），1158–1160。编号：11；出版商：Nature Publishing Group。[65] Justice, C., Townshend, J., Vermote, E., Masuoka, E., Wolfe, R., Saleous, N., Roy, D. 和 Morisette, J. MODIS陆地数据处理与产品状态概述。《环境遥感》83，1-2（2002年），3–15。[66] Kuttal, S. K., Sarma, A. 和 Rothermel, G. 当你记录下来时，历史更容易重演：混搭应用的版本管理。载于2011年IEEE可视化语言与以人为中心的计算研讨会（VL/HCC）（2011年），第69–72页。[67] Kwet, M. 数字殖民主义：美国帝国与全球南方的新帝国主义。《种族与阶级》60，4（2019年），3–26。[68] Lindenmayer, D. 和 Scheele, B. 不要发布。《科学》356，6340（2017年5月），800–801。出版商：美国科学促进会。[69] Lowe, A. J., Smyth, A. K., Atkins, K., Avery, R., Belbin, L., Brown, N., Budden, A. E., Gioia, P., Guru, S., Hardie, M., Hirsch, T., Hobern, D., La Salle, J., Loarie, S. R., Miles, M., Milne, D., Nicholls, M., Rossetto, M., Smits, J., Sparrow, B., Terrill, G., Turner, D. 和 Wardle, G. M. 开放但负责任地发布。《科学》357，6347（2017年7月），141–141。出版商：美国科学促进会。[70] Marzoev, A., Araújo, L. T., Schwarzkopf, M., Yagati, S., Kohler, E., Morris, R., Kaashoek, M. F. 和 Madden, S. 迈向多元数据库。载于操作系统热点问题研讨会论文集（美国纽约州纽约市，2019年5月），HotOS '19，计算机协会，第88–95页。[71] Mboma, J. C. A. 马莱马酋长领地凯拉洪区戈拉雨林国家公园周边森林边缘社区的替代性生计。《国际创新科学、工程与技术杂志》8（2021年）。[72] McSherry, F., Isard, M. 和 Murray, D. G. 可扩展性！但代价是什么 {COST}? 载于第15届操作系统热点问题研讨会（HotOS XV）（2015年）。[73] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. 和 Galstyan, A. 机器学习中的偏差与公平性综述。《ACM计算评论》54，6（2021年7月）。[74] Merali, Z. 计算科学：……错误。《自然》467，7317（2010年10月），775–777。[75] Mislan, K., Heer, J. M. 和 White, E. P. 提升代码在生态学中的地位。《生态学与进化趋势》31，1（2016年），4–7。[76] Moussa, R. 使用Apache Spark和Apache Sedona对空气质量批次进行可扩展分析。载于第15届ACM分布式与事件驱动系统国际会议论文集（美国纽约州纽约市，2021年6月），DEBS '21，计算机协会，第154–159页。[77] Murray, D. G., Schwarzkopf, M., Smowton, C., Smith, S., Madhavapeddy, A. 和 Hand, S. CIEL：用于分布式数据流计算的通用执行引擎。载于第8届USENIX网络系统设计与实现会议论文集（美国，2011年3月），NSDI'11，USENIX协会，第113–126页。[78] Mutanga, O. 和 Kumar, L. Google Earth Engine应用。《遥感》11，5（2019年），591。[79] Novella, J. A., Emami Khoonsari, P., Herman, S., Whitenack, D., Capuccini, M., Burman, J., Kultima, K. 和 Spjuth, O. 使用Pachyderm进行基于容器的生物信息学。《生物信息学》35，5（2019年3月），839–846。[80] Nugent, J. iNaturalist：面向21世纪博物学家的公民科学。《科学视野》41，7（2018年），12–15。[81] Parsons, A. 和 Powell-Smith, A. 释放碎片化公共数据的价值，2023年。https://research.mysociety.org/publications/unlocking-fragmented-data。[82] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. 和 Chintala, S. PyTorch：命令式风格的高性能深度学习库。载于第33届神经信息处理系统国际会议论文集，第721号。Curran Associates Inc.，美国纽约州雷德胡克，2019年12月，第8026–8037页。[83] Peterson, L. 和 Roscoe, T. PlanetLab的设计原则。《ACM SIGOPS操作系统评论》40，1（2006年1月），11–16。[84] Richards, P., Arima, E., VanWey, L., Cohn, A. 和 Bhattarai, N. 巴西的森林砍伐者是否在逃避侦测？《保护快报》10，4（2017年），470–476。 _电子版：https://onlinelibrary.wiley.com/doi/pdf/10.1111/conl.12310。[85] Saeedan, M. 和 Eldawy, A. Spatial Parquet：用于地理空间数据湖的列式文件格式。载于第30届地理信息系统进展国际会议论文集（美国纽约州纽约市，2022年），SIGSPATIAL '22，计算机协会。[86] Sarkar, A., Gordon, A. D., Negreanu, C., Poelitz, C., Ragavan, S. S., 和 Zorn, B. 使用人工智能编程是什么样的？arXiv预印本 arXiv:2208.06213（2022年）。[87] Sarkar, D., 和 Chapman, C. A. 智慧森林困境：阐释热带森林保护科学中传感器和人工智能的陷阱。《热带保护科学》14（2021年），19400829211014740。[88] Seth, A. 技术与赋权（或失权）：致技术人员的呼吁。Emerald Publishing Limited，2022年。[89] Simmhan, Y. L., Plale, B., 和 Gannon, D. 电子科学中数据溯源综述。ACM SIGMOD Record 34, 3（2005年9月），31–36。[90] Slough, T., Kopas, J., 和 Urpelainen, J. 通过培训和激励巡护，基于卫星的森林砍伐警报有助于秘鲁亚马逊地区的社区监测。美国国家科学院院刊 118, 29（2021年7月），e2015171118。[91] Society, T. R. 从隐私到合作伙伴关系。参见 royalsociety.org/privacy-enhancing-technologies。[92] Storm, D. 网络盗猎：黑入GPS项圈数据以追踪并猎杀濒危虎类。技术报告，2013年10月。https://www.computerworld.com/article/2475200/cyber-poaching–hacking-gps-collar-data-to-track-and-kill-endangered-tigers.html。[93] Swinfield, T., 和 Balmford, A. 剑桥碳影响：评估碳信用额度声明及协同效益。技术报告，2023年3月。[94] Swinfield, T., Balmford, A., Coomes, D., Madhavapeddy, A., 和 Keshav, S. 实现非永久性碳信用的社会价值。《自然·气候变化》13（2023年），1172–1178。[95] Tan, X., Zhou, M., 和 Fitzgerald, B. 扩大开源社区规模：Linux内核的实证研究。载于《ACM/IEEE第42届软件工程国际会议论文集》（2020年），第1222–1234页。[96] Tang, L., 和 Shao, G. 用于林业研究与实践的无人机遥感。《林业研究》26, 4（2015年12月），791–797。[97] Tian, J., Di, S., Zhao, K., Rivera, C., Fulp, M. H., Underwood, R., Jin, S., Liang, X., Calhoun, J., Tao, D., 和 Cappello, F. Cusz：一种高效的、基于GPU的科学数据误差有界有损压缩框架。载于《ACM并行体系结构与编译技术国际会议论文集》（2020年），PACT '20，计算机协会，第3–15页。[98] Tollefson, J. 人类正导致一百万个物种走向灭绝。《自然》569, 7755（2019年5月），171–171。[99] Tollefson, J. 碳排放创下新高：COP27发出警告。《自然》（2022年11月）。Bandiera_abtest: a Cg_type: 新闻 出版者：Nature Publishing Group 主题_term: 气候变化、政策、气候科学、数据库。[100] Trolliet, F., Vermeulen, C., Huynen, M.-C., 和 Hambuckers, A. 相机陷阱在野生动物研究中的应用：综述。Biotechnologie, Agronomie, Société et Environnement 18, 3（2014年）。[101] Tulloch, A. I., Auerbach, N., Avery-Gomm, S., Bayraktarov, E., Butt, N., Dickman, C. R., Ehmke, G., Fisher, D. O., Grantham, H., Holden, M. H., 等。用于评估发布生物多样性数据风险与收益的决策树。《自然·生态与进化》2, 8（2018年），1209–1217。[102] Vancutsem, C., Achard, F., Pekel, J.-F., Vieilledent, G., Carboni, S., Simon-etti, D., Gallego, J., Aragão, L. E. O. C., 和 Nasi, R. 湿热带地区森林覆盖变化的长期（1990–2019年）监测。《科学进展》7, 10（2021年），eabe1603。[103] Waldman, A. E. 权力、过程与自动化决策。Fordham L. Rev. 88（2019年），613。[104] Weiss, D. J., Nelson, A., Vargas-Ruiz, C. A., Gligorić, K., Bavadekar, S., Gabrilovich, E., Bertozzi-Villa, A., Rozier, J., Gibson, H. S., Shekel, T., Kamath, C., Lieber, A., Schulman, K., Shao, Y., Qarkaxhija, V., Nandi, A. K., Keddie, S. H., Rumisha, S., Amratia, P., Arambepola, R., Chestnutt, E. G., Millar, J. J., Symons, T. L., Cameron, E., Battle, K. E., Bhatt, S., 和 Gething, P. W. 医疗机构出行时间全球地图。《自然·医学》26, 12（2020年12月），1835–1838。[105] Whitehead, N., 和 Fit-Florea, A. 精度与性能：NVIDIA GPU的浮点数与IEEE 754兼容性。[106] Wulder, M. A., Roy, D. P., Radeloff, V. C., Loveland, T. R., Anderson, M. C., Johnson, D. M., Healey, S., Zhu, Z., Scambos, T. A., Pahlevan, N., 等。Landsat科学及其影响的五十年。《环境遥感》280（2022年），113195。[107] Yang, J.-H., 和 Chan, B. P.-L. 来自中国南部的睑虎属（有鳞目：蜥蜴亚目：睑虎科）两个新物种。《动物分类学》3980, 1（2015年6月），67–80。编号：1。[108] York, N. D., Pritchard, R., Sauls, L. A., Enns, C., 和 Foster, T. 保护遥感中的正义与伦理：当前论述与研究需求。《生物保护》287（2023年），110319。[109] Zeldovich, N., Boyd-Wickizer, S., 和 Mazieres, D. 通过信息流控制保障分布式系统安全。载于NSDI（2008年），第8卷，第293–308页。[110] Zeldovich, N., Boyd-Wickizer, S., 和 Mazières, D. 通过信息流控制保障分布式系统安全。载于《第5届USENIX网络系统设计与实现研讨会论文集》（美国，2008年4月），NSDI'08，USENIX协会，第293–308页。