数据科学家如何审阅学术文献

Sheshera Mysore
smysore@cs.umass.edu
马萨诸塞大学阿默斯特分校
USA

Mahmood Jasim
mjasim@cs.umass.edu
马萨诸塞大学阿默斯特分校
USA

Haoru Song
hsong@umass.edu
马萨诸塞大学阿默斯特分校
USA

Sarah Akbar
sakbar@umass.edu
马萨诸塞大学阿默斯特分校
USA

Andre Kenneth Chase Randall
andrekenneth@umass.edu
马萨诸塞大学阿默斯特分校
USA

Narges Mahyar
nmahyar@cs.umass.edu
马萨诸塞大学阿默斯特分校
USA

arXiv:2301.03774v1 [cs.IR] 10 Jan 2023

摘要

跟进研究文献在科学家的工作流程中发挥着重要作用——帮助他们理解某一领域，确定所关注的问题，并提出相应的解决方案；这些工作反过来又塑造着学科的面貌。在本文中，我们考察了数据科学家的文献综述实践。数据科学正经历论文数量的指数级增长，并日益借鉴众多不同学科的成果，也日益被应用于这些学科。近期的工作开发了若干旨在帮助数据科学家应对大量研究成果的工具，同时还开展了协调性的工作，开发旨在揭示研究前沿的人工智能工具。尽管这些趋势表明数据科学家正面临信息过载，但此前没有研究考察这些科学家在学术规范不断演变的跨学科领域中所面对的具体实践与挑战。在本文中，我们通过对工业界和学术界数据科学家开展一系列半结构化访谈和出声思维协议，弥补了这一空白（$N = 20$）。我们的结果在印证其他知识工作者实践的同时，也揭示了若干新发现：个体（1）在学科信息茧房之外寻找论文并进行意义建构时面临挑战；（2）面对缺失的细节和数学内容时，难以理解论文；（3）通过利用代码、博客和演讲中的知识背景来应对大量研究成果；（4）依赖线上和线下的同行。此外，我们还提出了未来可能帮助数据科学家应对不断增长的研究文献的发展方向。

CCS概念

• 信息系统 → 用户与交互式检索；数字图书馆与档案；协作与社交计算系统和工具；• 以人为中心的计算 → 人机交互设计与评估方法。

ACM参考格式：
Sheshera Mysore、Mahmood Jasim、Haoru Song、Sarah Akbar、Andre Kenneth Chase Randall 和 Narges Mahyar。2023年。《数据科学家如何评阅学术文献》。载于*ACM* 信息检索研究与发展国际会议人类信息交互与检索会议（*CHIIR '23*），*2023*年*3*月*19*日至*23*日，美国德克萨斯州奥斯汀。ACM, 美国纽约州纽约市，16页。https://doi.org/10.1145/3576840.3578309

允许免费制作本作品全部或部分内容的数字或纸质副本，供个人或课堂使用，前提是这些副本不得用于营利或商业用途，且须附上本声明和首页的完整引文。对于本作品中由作者以外的其他方拥有版权的部分，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、再版、发布到服务器或分发给列表成员，须事先获得特定许可和/或支付费用。请向 permissions@acm.org 申请许可。人类信息交互与检索会议 *'23，2023*年*3*月*19*日至*23*日，美国德克萨斯州奥斯汀© 2023年版权归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 979-8-4007-0035-4/23/03...$15.00 https://doi.org/10.1145/3576840.3578309

1 引言

文献综述在各类科学家的工作流程中发挥着重要作用——研究人员每周阅读研究文献的时间超过 11 小时 [91]。他们出于各种目的查阅这些文献，包括探索问题与解决方案前沿、跟进并加深理解，以及快速查找事实等 [46, 94]。这一领域面临的一项挑战是论文数量不断增加：研究估计，过去几十年里，已发表研究的数量每 9 年就翻一番 [71, 129]。正如 Chu 和 Evans [23] 所指出的，这种信息洪流造成了一种局面：整体科学进步放缓，而个别科学家和同行评审者则难以识别和理解新颖创意。在这项研究中，我们考察了一个正在兴起的知识工作者群体——数据科学家——的文献综述实践。

与科学领域的整体趋势相似，近年来数据科学出版物激增，每 2 年便翻一番[11, 14, 66], ，同时还开发了若干专门用于帮助数据科学家的工具[1]，并开展了协调一致的工作，以开发人工智能工具来识别数据科学研究的前沿 [66]。这表明该群体正面临信息过载。此外，数据科学作为一门学科，对商业 [103], 、政府 [98], 和科学 [13], 等领域产生了重大影响。尽管存在这些趋势，据我们所知，此前尚无研究考察数据科学家在综述研究文献时的实践或所面临的挑战。我们的目标是弥补这一空白。

在此，我们将数据科学家 视为从事数据工作、应用工程和研究工作，并接受过计算机科学、统计学以及经济学和生物学等特定应用领域学科训练的个体 [24]。我们将文献综述 视为一种知识构建过程，涉及研究文献收集、建立所收集数据之间的关系，以及综合信息的产生——涵盖信息寻求、意义建构和撰写 [122, 140]等任务。不过，在这项工作中，鉴于撰写 的关注范围较窄，主要面向出版，且这些阶段难以明确划分，因此我们不对其进行考察 [126, Sec 3]。

具体而言，我们考察信息寻求与意义建构这一迭代过程中的实践与挑战，其中

[1]https://search.zeta-alpha.com/, https://papers.labml.ai/papers, https://alphasignal.ai/

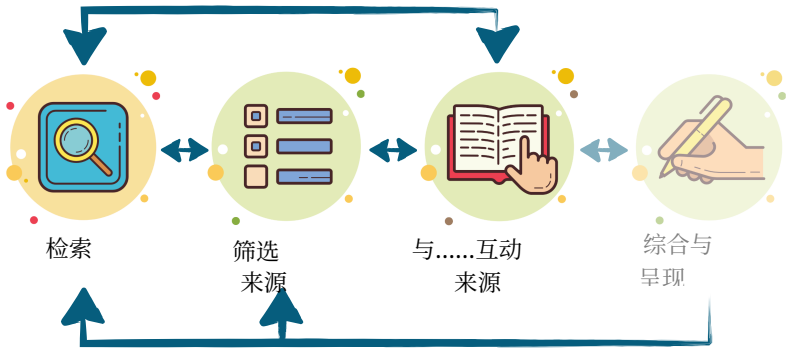


图 1：在本研究中，我们考察了数据科学家的文献综述实践。我们探讨了涵盖信息寻求与意义建构的实践和挑战：搜索与发现、文献筛选、略读与阅读，以及对一组论文形成综合理解。§4 展示了以这些阶段为基础的主题分析结果。然而，我们未研究长期综合与撰写。

研究文献会被搜索、筛选、阅读，并综合成心智模型，以实现特定任务目标 [110, 133]。在密切关联地研究信息寻求与意义建构这两项活动时，我们借鉴了搜索即学习 领域的一系列研究。这些研究将搜索视为与意义建构密切相关的学习过程，并主张支持用户收集信息以及利用所找到的结果 [69, 78, 117, 126]。我们的研究不同于以往的相关研究：那些研究分别孤立地 考察信息寻求与意义建构的实践，侧重于搜索行为 [7], 使用学术社交网络 [4], 使用信息管理系统 [92], 以及通过笔记记录捕捉创意 [52]。以往将信息寻求与意义建构结合起来考察的研究，要么聚焦于特定文献综述系统的上下文 [80, 122]，要么从抽象认知过程的视角出发[6, 125]，较少基于实际应用来考察专业知识工作者如何进行文献综述。

此外，通过考察数据科学家的文献综述实践，我们的工作也拓展了近期一系列揭示数据科学家工作实践的研究，涉及教学实践 [68], 代码与数据工作流 [79, 88, 120], 以及文档编写实践 [130], 等。我们特别关注数据科学家的实践，这也源于先前研究指出，不同学科的科学家在文献综述实践方面存在差异，这是由于研究方法、交流规范和学科价值观各不相同 [4, 12, 17, 49]。

为开展这项研究，我们从业界和学术机构招募了 20 名数据科学家，并采用两种不同的数据收集方法进行分析：半结构化访谈，以及参与者执行开放式搜索任务时的出声思维观察。我们的分析采用了多轮编码和主题分析。在一定程度上，我们的研究印证了以往的发现，同时也在信息寻求和意义建构方面揭示了若干新的、尚未得到充分研究的发现：1）参与者在寻求文献、略读文献，以及判断专业领域之外文献的可信度时遇到了困难。2）面对缺失的细节和数学内容时，他们难以理解论文。3）他们遭遇了大量看似相似的论文，并试图通过比较它们之间的差异来理解这些论文。4）在应对这些问题时，他们借助同行和一系列在线资源，形成了论文的知识背景：讨论论坛和社交媒体、代码、演讲以及博客。根据我们的研究发现，我们重点指出了一些尚未得到充分探索的领域，这些领域为数据科学家开展文献评述提供了丰富的未来研究空间和有意义的解决方案。

2 相关工作

这里我们概述四个相关的研究方向。第一个方向概述了各类知识工作者的信息寻求系统与实践。此外，我们还概述了面向学术文献的意义建构系统与实践。通过考察知识工作者的实践，我们以先前研究为基础，梳理数据科学家的实践发现。同样，我们对现有系统的考察与数据科学家面临的挑战（§4）相结合，使我们得以探讨有意义的未来研究方向（§5）。接下来，我们概述有关搜索即学习的研究；这类研究考察了在学习任务（如文献综述）中搜索系统的使用情况。与信息寻求和意义建构研究相辅相成的是，这一研究方向也揭示了搜索、阅读和记笔记等任务对以学习为导向的目标所产生的重要影响。尽管这三个研究方向在一定程度上由彼此独立的社群开展，但它们可以视为紧密相关：信息寻求、意义建构与学习相互关联 [78]。最后，我们概述近期有关数据科学家实践的研究——我们的工作与这部分文献的联系最为直接。

2.1 信息寻求与意义建构实践

2.1.1 信息寻求。 大量研究探讨了不同知识工作者的信息寻求实践，涵盖开展系统性文献综述的医学研究人员 [63], 、他们对搜索界面的使用 [76],、设计专业学生利用搜索进行构思的方式 [95], 、社会科学与教育研究人员的实践 [51], 、天体物理学家的实践 [112],，以及工程学教授对机构图书馆服务的看法 [29]。Niu 等人 [91] 和 Alhoori 等人 [4] 对该领域早期研究进行了广泛综述

。鉴于我们关注的是数据科学家，我们综述了考察相关学科人群信息寻求实践的研究 [4, 7, 53, 81, 118]。

Athukorala 等人 [7] 考察了计算机科学家进行文献检索的目标及其使用的工具。主要发现表明，受访者最常使用搜索来了解主题的最新进展，而探索未知领域则是他们认为最具挑战性的事情。他们发现，浏览论文最常见的方式是追踪某篇已知种子论文的引文，通常这篇种子论文是通过最初的关键词检索找到的。Alhoori 等人 [4] 结合社交媒体使用和参考文献管理工具，考察了科学、技术、工程和数学领域研究人员的信息寻求行为。Hoeber 等人 [46] 考察了不同资历层次的科学、技术、工程和数学领域研究人员在实践方面的差异。最后，Soufan 等人近期的研究 [118]开展了一项面向科学、技术、工程和数学领域研究人员的大规模综述，旨在揭示探索式搜索的理论模型与研究人员实际采用的做法之间的一致性。另有一些研究考察了基于特定系统的行为：Ishita 等人 [53] 调查研究人员会查看搜索结果中返回的论文的哪些部分；McCay-Peet 等人 [81] 则考察社会科学家和计算机科学家在使用数字图书馆的控制功能方面的差异。

2.1.2 意义建构。尽管已有大量研究考察信息寻求行为，但专门探讨意义建构任务实践的研究较少，例如管理收集到的文献、阅读和记笔记。Nosheen 等人 [92] 考察了工程领域研究人员在个人信息管理方面面临的挑战，发现数据分散在不同系统中，以及难以判断信息未来的价值，都是他们面临的挑战。同样，Inie 等人 [52] 探讨了研究人员管理创意时的工具使用模式，发现工具之间的互操作性至关重要。最后，Morabito 和 Chan [85] 研究了研究人员如何使用工具记录文档的上下文和元数据，以便有效地重新利用找到的信息。

与这类将信息
寻求和意义建构任务分开考察的研究不同，我们通过一系列访谈和出声思维观察，结合考察这两类任务。这种综合考察更真实地呈现了信息寻求与意义建构的迭代过程[110, 133], ，尤其是意义建构的早期阶段。这使我们得以深入了解这两类活动交汇处的实践与挑战，为开发支持这两类活动的系统提供参考。

2.2 信息寻求与意义建构系统

2.2.1 信息寻求。 以科学文献为对象构建信息寻求系统的早期研究探索了多种策略：通过种子论文和引文追踪发现文献 [21, 40, 74, 101], ；利用作者或关键词等元数据进行检索 [27, 28, 60,102], ；以及探索更专门查询方法的系统[19, 22, 30, 61, 106]。

AgroScholar [74] 和 PaperQuest [101] 支持从论文种子集开始，对引文网络进行可视化和交互操作——用户可以根据系统可能推荐的其他论文，将论文逐步添加到核心论文集中。Apolo[21] 和 PaperPoles [40], 中的工作进一步支持用户进行如下交互：对论文进行分组，并指定关键词偏好，以影响推荐论文集。虽然这一研究方向侧重于基于引文的发现，但另一些研究则更多地利用论文元数据。PivotPaths [28] 提出了一种可视化系统，使研究人员能够通过作者和关键词元数据探索论文集，特别强调探索元数据元素之间的关系，并促进意外发现。类似地，Portenoy等人 [102] 和 Kang 等人 [60] 支持从与种子作者或论文所属学科不同的学科中发现相关作者和论文。最后，一系列研究探索了超越论文或其元数据的专门查询方法，更充分地利用论文内容。Chan等人 [19] 和 Kang 等人 [61] 的工作支持根据论文所处理的“挑战”“解决方案”或“结果”搜索论文，而 DataHunter [31] 支持根据研究问题查询搜索数据集。

2.2.2 意义建构。 我们将用于意义建构的早期系统分为两类：辅助用户处理文档集合 的系统 [45, 80, 113], ，以及帮助用户阅读 [32, 42, 131, 139]和对单篇 文档进行记笔记的系统 [38, 58, 121, 142] 。

由于意义建构通常包含反复进行的信息寻求阶段，以及基于收集到的数据构建心智模型的阶段[100]，，面向文档集的意义建构系统也包含辅助信息寻求的组件——这些组件既允许用户探查并构建关系，也允许用户发现更多相关研究 [21, 28]。其他研究则致力于帮助用户识别科学文献中显著的时间对齐趋势。Shahaf等人 [113] 根据科学论文集构建“地铁图”，以描绘多条相关研究脉络的演变；类似地，Heimerl等人 [45] 借助流图隐喻，让用户能够考察引文、主题和研究社群随时间变化的模式。最后，PaperForager [80] 针对的是另一个问题——尽量减少文档筛选、略读和阅读过程中的上下文切换。它将论文集完全呈现为持续可见的文档图像。

在增强阅读体验方面，既有研究致力于帮助读者将文档与其他概念或论文联系起来，也有研究旨在丰富单篇文档的阅读体验。Zhang等人 [139] 利用气泡树图，帮助读者直观呈现并探索文档中的概念层级。类似地，Wang等人 [131] 帮助用户更好地理解单篇论文的相关工作章节。另一方面，针对单篇论文的阅读辅助工具也研究了如何通过高亮标记突出重要文本，辅助略读 [32], ；还有一些研究探索了如何增强对论文中方程和术语的阅读体验 [42]。除阅读之外，研究还探索了各种形式的记笔记支持，以便综合和重组收集到的数据。Passages [38] 和 Threddy [58] 等项目帮助用户从阅读的论文中提取并组织文本片段和笔记，使其能够在不同应用和文档间保留。Passages着重利用持久化元数据和跨应用迁移来保留信息来源，而 Threddy则侧重于根据笔记集发现更多论文。最后，也有研究探索了用于文献综述的一体化系统，涵盖搜索与发现、组织、综合和撰写 [122, 140]。与开发用于研究特定系统中实践的系统的工作相比，

我们在特定系统的上下文中开展了一项更具探索性的研究，考察科学家在其自然工作流程中的活动，旨在为未来系统的发展方向提供更广泛的参考。

2.3 搜索即学习

有关搜索即学习的研究探讨了搜索过程中学习如何发生的多个不同方面，并得出了一些与信息寻求和意义建构相关的发现。大部分研究重点关注学习成果，同时考察了以下因素：1）用户特征；2）搜索系统特征；以及 3）搜索行为[126]。在这些研究中，学习最普遍的定义是知识心理模型的形成、这些模型随时间的保持，以及它们的应用 [78]。这些研究还提供了一种超越查找和收集结果的搜索视角。相反，它们将搜索视为一个由一系列任务构成的建构过程——这一过程既需要对搜索提供支持，也需要支持对搜索结果的运用 [69, 78, 117, 126]。不过值得注意的是，搜索即学习领域的大部分研究都聚焦于课堂学生，或通过对众包工作者开展受控实验得出推论，而非研究数据科学家等知识工作者。

2.3.1 用户特征。 关于用户特征的研究常常考察用户的领域知识。Willoughby 等人 [134] 考察了在论文写作任务中，领域知识与使用搜索引擎对结果的影响。他们发现，搜索仅对领域知识较丰富的参与者撰写论文有帮助，而参与者也指出，领域知识有助于搜索。Roy 等人 [108] 在参与者探索某个主题的搜索会话期间，通过词汇评估考察学习情况；他们发现，主题知识较丰富的参与者在搜索会话后期获得了知识增益，而主题知识较少的参与者则在会话初期获得了更多知识。尽管考察用户特征的研究往往聚焦于领域知识，Vakkari 等人[128] 也发现，领域知识只有对熟悉搜索系统的用户才会影响搜索。

2.3.2 系统特征。 在考察系统特征时，先前研究探讨了搜索功能 [18, 25, 104] ，以及与阅读和记笔记相关的功能 [33, 109, 123]。在这方面，di Sciascio 等人 [25] 发现，与广泛使用的 PubMed 相比，透明且可控的搜索系统能带来更好的学习成果；Qiu 等人 [104] 发现，与对话式搜索界面相比，网页搜索界面能带来更大的知识增益，但后者有助于提升知识保持。在研究阅读体验时，Freund 等人 [33] 发现，较简单的文本环境有助于提升理解；Roy 等人[109] 发现，阅读时添加高亮标记有助于提高论文写作任务中的主题覆盖度，而记笔记有助于纳入更多事实；Syed 等人 [123] 则发现，将自动生成的问题嵌入文本中有助于提升学习成果。

2.3.3 搜索行为。关于搜索行为的研究带来了一些发现。Vakkari 等人 [128] 发现，随着词汇量的提升，心理学学生会使用更具体的查询，但其采用的搜索策略始终保持一致。Dosso 等人 [26] 发现，领域知识不会影响医学和计算机科学学生所使用查询的数量或长度，不过，他们发现医学学生会在查询中使用更多领域专用词汇。Vakkari 和 Huuskonen [127] 发现，投入更多精力查看文档与论文质量提升有关，即使搜索结果的精确率较低也是如此；他们还发现，学生会设法弥补糟糕的搜索结果，这一发现也在其他研究中得到印证 [76, 116]。Moraes 等人 [86] 的研究发现，将搜索与教师讲座结合，比单独听讲座更能提升学习成果。最后，Urgo 和 Arguello [125] 不再关注具体交互，而是将不同认知复杂度任务中通往学习目标的“路径”概括为以学习为导向的子目标。这也与一系列探讨任务复杂度和学习成果的研究相关。

在我们的工作中，我们并未聚焦于具体的学习成果，也没有实施促进学习的干预措施，而是开展了一项探索性的需求发现研究，旨在揭示数据科学家在以学习为导向的任务中，从搜索到使用搜索结果这一过程中的实践与挑战。对此，我们借鉴搜索即学习（SAL）的观点，将搜索视为一种建构性过程，认为它不仅需要关注信息寻求，也需要关注意义建构，而非仅仅收集搜索结果 [69, 126]。

2.4 数据科学家的实践

除上述研究方向外，近期还有大量研究考察了数据科学家的广泛实践——数据科学家是一个新兴的知识工作者群体 [24], 。我们在此简要讨论这方面的研究。Crisan 等人 [24]通过回顾以往有关数据科学家的研究，描绘了数据科学家的群体画像。也有一些论文考察了数据科学家与数据集相关的信息寻求需求。Kross 和 Guo [68] 采访了在工业界和学术界从事数据科学家培训的人员，并发现，讲师认为，搜索具有教学相关性的数据集是他们面临的挑战之一。Koesten 等人 [64] 探讨了数据科学家在搜索结构化数据存储库时遇到的挑战。还有一系列研究通过代码本考察数据科学工作流 [79, 120], 、其文档实践 [130],, 以及他们如何做出分析决策 [62]。其他近期研究则考察了数据科学家在开发公平机器学习系统 [48], 以及采用可解释机器学习系统 [67]时面临的挑战与需求。

尽管这一系列研究考察了数据科学的开展方式，但数据科学家如何寻求并理解不断增长的研究文献仍不为人知——正如我们所指出的，近年来面向数据科学家的工具不断涌现，学术文献也持续增长，这表明这是一项重要而具有挑战性的活动。因此，在开展需求调查时，我们的工作旨在加深对数据科学家工作实践的理解，尤其关注他们对学术文献的信息寻求与意义建构。

3 研究方法

3.1 研究设计

我们的研究主要包含两个部分：半结构化访谈，以及随后在参与者所选工作环境中进行的出声思维观察。半结构化访谈旨在了解参与者自述的实践和挑战，而出声思维观察则旨在观察更为隐性的行为、

表1：参与者概况。研究领域由参与者在半结构化访谈中自行描述；为保护参与者对其研究领域保密的要求，本文在介绍中对此予以尊重。在20名参与者中，13人就读于美国和欧盟各大学的博士项目，7人在营利性行业或非营利组织中担任数据科学家、机器学习工程师或统计学家职称。

参与者	组织	职称	研究领域
P1	工业界	数据科学家	对话式信息检索
P2	非营利组织	统计学家	统计学
P3	工业界	数据科学家	计算机视觉
P4	工业界	数据科学家	法律文本自然语言处理
P7	工业界	数据科学家	语音处理
P5	大学	博士生	用于公共卫生的机器学习
P6	大学	博士生	视频处理
P8	大学	博士生	搜索即学习
P9	大学	博士生	自然语言处理鲁棒性
P10	大学	博士生	问答与推理
P11	大学	博士生	自然语言处理鲁棒性与仇恨言论
P12	大学	博士生	新闻业自然语言处理
P13	大学	博士生	变分推断
P14	工业界	机器学习工程师	NLP
P15	大学	博士生	科学文档自然语言处理
P16	大学	博士生	差分隐私
P17	大学	博士生	计算机科学理论
P18	非营利组织	数据科学家	密码学
P19	大学	博士生	强化学习与因果推断
P20	大学	博士生	数据管理

工作流程和实践，并将其置于真实的任务场景中。在出声思维环节中，参与者会根据三个开放式任务提示之一，完成自己选择的文献综述。我们的研究设计和内容是在一项试点研究中制定的，该研究于2022年2月至5月开展，共有15名参与者。在试点研究中，我们旨在探究文献综述中涉及的信息寻求行为；然而，在初步结果中，我们观察到大量与信息寻求密切相关的意义建构活动。在2022年7月至8月开展的第二轮研究中，我们邀请了20名新参与者，并更新研究内容，以进一步探究文献综述中更广泛的意义建构活动——本文报告的正是第二轮研究。附录A列出了我们的问题和出声思维提示。

3.2 研究参与者

我们通过在 Twitter 和 LinkedIn 上发布社交媒体帖文，以及向大学邮件列表发送电子邮件邀请，招募研究参与者。招募表中展示了对“数据科学家”的宽泛定义，并询问受访者是否认为自己是数据科学家。此外，我们还请参与者提交过去一年中觉得有用或读起来有趣的 3 篇研究论文的链接或标题。在受访者中，我们按先到先得的原则选出 20 人作为研究参与者（P1-P20），同时确保他们认为自己是数据科学家，并且在过去一年中读过论文。参与者获得了价值 25 美元的礼品卡，以补偿他们投入的时间。所有研究流程和材料均经大学机构审查委员会批准。附录A包含我们的招募表。

20 位参与者中，11 位使用 he/him 性别代词，9 位使用 she/hers，2 位使用 they/them，其中一些人使用多个代词。14 位参与者正在攻读或已获得博士学位，6 位已获得硕士学位。13 位参与者在大学工作，7 位在非营利组织或营利性行业的工业实验室工作。参与者平均发表过 4 篇研究论文。表1介绍了参与者的研究领域以及他们开展工作的组织类型。

3.3 研究流程

每次研究环节均由作者主持，时长约为 1 小时，其中半结构化访谈和出声思维环节各占一半。整个环节通过 Zoom 进行。每次环节开始时，我们会说明研究流程，征得参与者同意其参与，并请他们填写一份简短的人口统计调查。随后进行半结构化访谈，询问参与者的研究重点、目标，以及开展文献综述时的做法和面临的挑战。我们鼓励参与者将文献综述视为超越单次定向搜索的过程，并讨论他们与科学文献的所有互动。接下来是出声思维环节：参与者通过 Zoom 屏幕共享进行文献综述。我们向参与者展示了 3 个宽泛的任务情境，并请他们选择一个作为出声思维的提示。其中一个提示请他们回忆并演示以往的一次文献综述；另一个请他们加深对熟悉主题的了解；最后一个则请他们针对自己感兴趣、但经验较少的领域，为未来的项目查阅相关文献。后两个情境取自 Hoeber 等人 [46] 的研究。我们的分析显示，参与者在这 3 种情境中的分布均匀。请注意，我们没有将简单的查找搜索用作任务情境，因为我们希望了解更复杂的探索性搜索。如果论文相关性足以需要深入阅读，我们会邀请参与者继续搜索，或结束研究环节。研究过程中录制了音频和视频，以供分析。

3.4 数据收集与分析

我们的所有分析均使用了访谈的转录音频，并查看了出声思维环节的视频。音频由 Zoom 自动转录，之后我们对转录内容进行了大量纠错。随后，本文的 3 位作者进行了 3 轮编码，并对访谈和出声思维转录文本进行了主题分析——这一过程于 2022 年 8 月至 10 月开展。第一轮编码中，2 位作者分别独立检查了 5 份转录文本，并通过开放编码和主轴编码生成一组代码；随后将这些代码整合为统一的代码集。整合时，两位作者先分别检查对方的代码，再会面讨论差异，并将代码整合为统一代码集。之后，两位编码者将这套统一代码应用于这 5 份转录文本。根据名义尺度的 Krippendorff α系数，我们观察到一致性为 0.92。这套统一代码集包含 167 个代码，其中 51 个记录了参与者使用的工具名称或研究的后勤安排。随后，我们讨论并解决编码分歧，再将最终代码应用于其余 15 份转录文本，同时将新代码补充到初始编码手册中。这一过程新增了 59 个代码，其中 32 个记录了工具名称或后勤安排。之后，我们利用编码后的转录文本进行主题分析，结果发现，形成的主题部分对应于信息搜索过程的各个阶段 [126, 节 3], 下面我们将介绍这些主题。我们的代码以及各主题的扩展引文可在线查看。[2]

4 结果

在呈现主题分析的结果时，我们大体上按照信息搜索过程来组织各个主题。该过程包括形成信息需求（§4.1）、构建查询式并进行搜索（§4.2）、通过略读与阅读评估搜索结果（§4.4、§4.3），以及综合一组文档（§4.4、§4.3）。此外，最后一个主题阐述了参与者如何在信息搜索过程的各个阶段依赖社会关系（§4.5）。需要特别指出的是，尽管这些阶段提供了一个有助于理解的组织框架，但它们属于边界定义模糊的迭代过程。呈现结果时，我们会先介绍各个主题，并在适当之处随后讨论相关的具体既有研究。

4.1 为什么数据科学家要访问科学文献?

4.1.1 主动理解学科规范。参与者最常通过主动查阅文献来理解自己尚不熟悉的学科规范（18/20）。他们希望了解学科词汇，形成对该学科的心智模型，并理解问题、解决方案及评估实践所处的上下文和社群偏好：

> “对于任何研究者来说，首先想到的问题是：已经有哪些工作完成了？是否存在已经解决的类似问题，或是相关问题，其对应的方法或许可以用于我的问题。”——P16

> “当我开始着手解决一个问题时……我对相关工作还不够熟悉，不知道通常采用哪些方法、如何进行评估，也不了解哪些方法正逐渐失宠，哪些方法正越来越受到社区认可。”——P15

参与者关注问题和解决方案，这反映了他们在科学和创造性问题解决学科中的既有理解；他们往往会从“问题”和“解决方案”的角度思考 [44, 95]。对评估和指标的关注，也体现了他们认识到数据科学高度重视定量表现 [12]。

4.1.2 被动地关注某个学科。除了主动寻求了解某个学科，参与者还提到，他们也会以更被动的方式查阅文献（10/20）——关注社群中的趋势，或了解同行的最新工作：“了解社群的发展方向，或我之前关注过的那些人的最新工作” - P10；也会留意未来的项目：“与我自己的研究相关的阅读，以及我曾考虑过……在某个时候开展未来研究的内容。*[但]* 现在我觉得这件事已经搁置了” - P17。这种保持关注的动机也反映在相关研究中 [7]。

4.1.3 头脑风暴解决方案。大多数参与者也利用现有科学文献作为辅助头脑风暴的资源（18/20）。在这些参与者中，有人借助文献寻找尚待解决的问题（12/20）：“确保我没有重复做别人已经做过的事，这是一个主要方面——弄清楚有哪些尚未解答的问题，以及有哪些现有工作可以作为我的研究基础”——P10。另一些人则考察了以往研究中的解决方案（13/20）：“我会寻找能够实现我正在做的事情的现有方法。其他人是如何解决这个问题的？这个问题已经解决了吗？是否已有方法能够解决这个问题？——P16。还有一些人通过寻找类似的解决方案，来确立自己创意的新颖性（10/20）：“在你弄清楚 *[问题],*之后，就会想到可以如何做得更好，然后就要看看其他人以前是否做过类似的事情”——P5。

这一寻找问题、提出解决方案并确立创意新颖性的过程，形成了一个迭代式头脑风暴循环。其他研究也指出，搜索可用于创造性头脑风暴 [95], ；而在科学文献中寻找问题和解决方案的需求，推动了旨在促进这一过程的方法和资源的发展 [19, 61, 89]——我们的研究结果进一步支持了这一努力。对新颖性的重视以及在前人工作的基础上开展研究，也与人们对其在数据科学研究 [12] 及更广泛的创造性学科 [52]中的重要性的认识相符。

4.1.4 寻求应用解决方案。除了为开发解决方案而查阅文献外，若干参与者也会针对问题寻求解决方案（17/20）。在这里，参与者通常是为了直接应用而寻找解决方案：“我还会看看其他人是如何应用 *[这种方法],* 的，看看其他人是否将其应用于与我的数据或用例类似的场景” ——P2；或者将其作为有助于研究的基线系统：“人们通常会采用哪些方法来解决某项任务……这在我们尝试发表论文时会有所帮助……这样我们就能了解现有的基线方法，以便在定量研究中进行比较” ——P6。

将解决方案作为比较基线来寻求，或将其作为应用的解决方案来寻求，这与引文分析的发现相吻合，后者表明

[2]代码和扩展引文：https://github.com/MSheshera/dslitreview-study

解决方案要么是经过验证且广泛采用的方法，要么处于研究知识的前沿并展现当前的发展，因此有必要进行比较 [57,第 8] 节。此外，“创建”和“应用”这两种行为也对应于常用于明确界定学习目标的认知过程 [125]。

4.2 数据科学家如何获取科学文献?

4.2.1 寻找文献的数据科学家。不出所料，大多数参与者通过基于关键词的网络搜索（17/20）、追踪源文章的引文及其所引用的文献（15/20）（这些源文章通过搜索或推荐获得），或查阅相关出版物作者的论文（13/20）来寻找文献。这与以往的认识大体一致 [7] ，并促使人们开发了若干旨在帮助进行引文追踪的工具 [21, 101, 102]。尽管这些方法早已广为人知，参与者仍常常感叹，难以找到恰当的关键词来描述自己的信息需求（13/20）；有时他们需要花上数周才能偶然找到合适的关键词。在不熟悉的学科中，这一问题尤为严重：

> “我脑子里有个想法，但不确定该如何用社群使用的规范术语来表达，甚至不确定社群究竟对什么感兴趣……最后只能不断试错，比如找一些论文，再过几周回来继续找，最后我才逐渐发现一些真正符合需求的内容。”——P15

为应对复杂检索中的这一挑战，先前研究探索了查询推荐策略 [22, 83, 96], ，大型系统也采用了这些策略 [115]。更普遍地说，先前研究指出，文献检索往往具有探索性，搜索过程不断演进、界定不明且开放式；用户通常会在搜索过程中逐渐了解相关术语 [118]，这也印证了构造查询时面临的挑战。尽管已有许多系统专为探索式搜索而开发（见 §2.2），大型文献检索系统却缺乏对此类搜索的支持 [90]。

最后，面对不知从何开始检索的情况，参与者还提到会向专家同行寻求推荐（4/20）：“如果你不知道从哪里开始，只要联系某个人，尤其是在像 *[已隐去]* 这样的公司里，这就容易多了；那里有很多主题专家，他们会帮助我们。” - P7。正如我们接下来将指出的，如何利用社交关系仍未得到充分探索。

4.2.2 数据科学家如何发现文献。除了通过搜索查找文献外，参与者也会以更被动的方式发现科学文献（20/20）。他们获取自动推荐的主要方式包括在社交媒体上关注他人（13/20）、订阅来自 arXiv 或 Google Scholar 的电子邮件提醒（7/20）、关注特定作者的研究成果（13/20），或订阅旨在精选文献的新闻简报（4/20）。不过，参与者指出，那些本该为他们提供信息的提醒反而让他们应接不暇（5/20）：“能跟上大家的研究进展当然很好……至少你会打开其中 *10%* 的内容，而不是觉得这些全是垃圾邮件。” - P1。此外，参与者也指出，发现工具往往会把他们困在学科泡泡中（4/20）：“我大概深陷于自己的论文圈子里……如果我研究的是仇恨言论，我收到的大多数推荐都会非常偏向计算机科学，但社会科学领域或许也有相关内容，而我可能永远都不会接触到这些内容。” - P11.

除了通过自动化方法发现文献外，大多数参与者还表示会收到同行的推荐（14/20），不过频率低于自动化方法。这种方式有几个优点：同行熟悉参与者的兴趣、审阅过论文，并且能进行更深入的交流：

> “我有一些来自工业界和学术界的同行，如果他们发现了有趣的东西，而且知道 [name] 正在研究这些问题，或者她觉得这些内容有意思，就会直接发给她。有时发来的甚至不是论文，而只是一篇博客文章，里面附有论文链接。”——P9
>
> “我想，[同行推荐的好处是]，它已经经过了一双眼睛的审视。因此，我们会更愿意去读；有人认为它很有意思，而且其中的论点也说得通。”——P10

Kang 等人的近期研究 [59] 发现，在学术邮件提醒中加入自动生成的基于社交网络的解释，甚至可以提升科学家的参与度和留存率，同时让他们感受到自己是社区的一员。

4.3 数据科学家如何选择论文?

4.3.1 论文如潮水般涌来。通过搜索与发现获得文献后，参与者接下来要面对的是筛选和整理文献。他们经常提到，面对大量相似的论文（16/20），并将此归因于数据科学这一学科领域竞争激烈。这使得他们难以在众多相似的论文或渐进式改进与奠基性论文之间做出区分：

> “如果一个领域竞争非常激烈——有时我觉得强化学习以及我关注的问题都很拥挤，那么就会让人感到沮丧，因为你总会发现一些[做着同样事情]的论文。”——P19
>
> “有很多论文只是渐进式改进，所以要找到那篇真正开创一切的奠基性论文会很困难；除非有人帮你，否则确实非常难。”——P7

为应对这股论文洪流，参与者会借助综述或优质的文献评述来寻找重要论文（7/20）。一些人选择专注于少数几篇论文，尽量减少所查阅论文之间的重叠，并将深入阅读少数论文与引文追踪相结合，以平衡探索结果的广度和深度（9/20）。还有一些人会将某些概念或论文被反复引用视为发现值得研读的论文的信号（6/20）：“通常会有一组论文，在所有 *[论文]* 中都会被引用；对我来说，这很好地表明了哪些是我应该关注的核心论文。”——P16。面对许多相似的变体，信息觅食领域的研究指出，用户会根据变体之间的差异来理解它们，或将其视作按时间顺序展开的故事[119]。虽然这一点是在寻找代码的情境下进行探讨的，但我们注意到，在更具挑战性的文本文件场景中，人们也有类似的需求；第4.4.3节将进一步阐述这一点。

4.3.2 确立论文可信度。应对这场信息洪流的一项重要挑战，是确立论文的可信度

(13/20)。在此过程中，参与者依赖一些预期指标：以熟悉的作者、所属机构、发表渠道和引用次数作为可信度的标志。正如我们在§4.2.2 中所指出的，有些人依赖同行推荐所提供的审查，另一些人则会寻找论坛讨论："一件难事是判断一篇论文的可信度，所以我会尝试根据 *Twitter*、*Reddit* 或 *Openreview* 等在线论坛上的讨论来判断。即使这篇论文经过了高度评价，其他在类似领域工作的人又是怎么看的呢？"- P14。

然而，可信度方面的挑战并未随着论文的筛选而结束。参与者还指出，论文内容有时会违背其信息气味（12/20）。这在一定程度上或许可以归因于学科写作规范，但 [50],参与者也提到了论文贡献不明确或被夸大、思想重新包装，以及论文结构未围绕读者的信息需求展开等问题：

> "人们常常会给思想重新包装，有时很多创意并不新颖，但动机部分写得像诗一样；读到细节时，你会觉得[他们] 所做的并非他们声称的那样。..... [或者] 夸大自己的贡献，而实验又达不到预期。因此，从论文中识别这些趋势非常重要。"- P19

出版元数据在热门搜索引擎中得到了充分应用，其在判断可信度方面的作用也经过了严格审视 [10, 36]。然而，利用论文相关的社交话语来辅助略读和筛选的方法与系统，尚未得到充分探索。此外，尽管科学传播领域已研究过贡献夸大等问题 [132]，但专家进行文献检索时遇到的此类问题仍研究不足——我们在此发现了这一现象的证据。这些由夸大或重新包装的主张，以及需要从大量相似论文中筛选信息（§4.3.1）所引发的挑战，可能是数据科学这一竞争激烈且激励机制高度集中的学科所导致的。尽我们所知，我们首次报告了这些挑战；但它们值得在数据科学及其他快速发展的学科中得到更深入的研究。

4.3.3 每个人都会略读论文。最后，在筛选结果时，所有参与者都会略读单篇论文（20/20），通常是为了快速判断其正确性或了解论文的贡献。他们往往依靠对学科的熟悉程度来判断应去哪里查找特定信息；反之则会遇到困难：

> "我会直接跳到他们说'在本文中'的地方，因为我对这个领域相当熟悉，不需要真的去看引言。"——P7 "一个挑战是， [这个学科] 在 90 年代非常活跃.....相关研究发表在不同的发表渠道上，而且他们使用的方法我并不熟悉。这让略读论文、把握其贡献精髓变得困难得多。"——P15

略读通常与信息寻求交织在一起，二者之间频繁切换上下文；然而，先前很少有研究考察信息寻求与略读的紧密整合 [80], ，或用于辅助论文略读的工具 [32]。

4.4 数据科学家在阅读论文时面临哪些挑战?

4.4.1 理解隐藏的细节。在阅读和理解论文时，参与者还指出，论文中缺少细节会带来挑战（13/20）。他们提到，论文的写作目标是被接收，而作者几乎没有动力去纳入那些使解决方案有效的细节："论文中经常会包含那些能够构成一个好故事的内容，但论文中真正重要的东西，例如如何设置超参数，却被遗漏了。"- P5。一些参与者还指出，纳入大量细节与读者希望了解论文大意之间存在矛盾：当然，我总会遇到这样的情况：我有时间阅读一篇论文，于是开始深入其中，却面对大量细节。有时是因为篇幅，有时是因为某些细节必须被省略，而有时候，恰恰是这些细节会让读者感到困惑。- P9。为应对细节缺失，参与者指出，由代码提供的扩充内容具有价值（7/20）。

> [我会询问作者，你们所做的工作是否有公开可用的代码。因为这些论文中有很多看起来写得不错，但实际如何实现却不清楚；或者他们具体选择了哪些超参数也不明确。——P16

虽然 paperswithcode.com 等平台将论文与代码配对，但利用代码提升科学论文阅读体验，仍是未来可改进的方向。更广泛地说，模型和数据的文档不清晰所带来的挑战，促使数据科学界推动相关举措，以记录这些产物 [34, 84], 有时还会在发表时为其提供激励。[3] 然而，截至 2022年9月，论文附带代码的比例仍为 25%；与此同时，针对文档编写的激励措施也正在逐步形成共识 [20]。

4.4.2 理解展示出来的数学。 参与者指出，缺少细节是一方面；另一方面，大量数学内容也带来了挑战（10/20）。参与者提到，学科亚文化往往会奖励数学内容丰富的论文，并将难以理解等同于高质量：

> 面向小众受众写作时，需要证明 [一个想法] 很重要或有用，而这往往意味着他们会为 [一个想法] 添加方程或定理，尽管它们实际上并没有那么复杂.....如果数学内容很多，或者很难理解，那它一定很了不起。- P5

在应对这一问题时，参与者指出了代码、博客和演讲等替代信息来源的重要性，或是在论文中加入示例以帮助理解（9/20）。他们还指出，演讲和博客能更好地激励受众确保自己理解了内容：博客很有帮助，因为有时候文字 *[在论文中]* 不像非正式写作那样容易理解 *[I*：所以，论文可能数学性更强，而博客会介绍高层次的创意？ *]* 没错 - P6。

阅读论文时辅助理解数学的工具，在近期研究中已有一些先例，形式是旨在

[3] https://naacl2022-reproducibility-track.github.io/ [4] https://paperswithcode.com/trends

更好地解释论文中的数学符号 [42, 43]。然而，这仍是一个尚待解决的问题，目前还没有研究探索如何利用代码、博客或演讲来辅助数学理解。还值得注意的是，论文细节缺失和数学理解方面的挑战，以及借助代码和论文周边知识背景的应对策略，可能是数据科学等跨学科和计算学科特有的挑战——我们的工作对此进行了初步报告。

4.4.3 难以理解进展增量。 在阅读论文时，参与者常常提到，理解论文相较于其他有影响力或已知论文所提出的具体"增量"很重要（9/20）。这对于理解该学科并将其贡献置于上下文中至关重要：

> "读过 10、20、30 篇论文之后，核心思想是什么，以及相对于核心思想的增量是什么，就会清晰得多。……对于那些你不太熟悉的领域，如果你不了解论文撰写时的上下文，往往会觉得一团糟。要弄清文献中已有和缺失的内容很难，接着要弄清贡献是什么、为何重要也很难。" - P5

为了理解进展增量，参与者指出，熟悉子学科的规范很重要，而缺乏这种熟悉则会带来挑战： 在研究生院，教授会综合这些内容并告诉你：嘿，这些论文的核心主题是什么。当这些信息已经为你梳理好，你开始阅读论文时，就知道该期待什么；否则，你会花很多时间，却不明白它与以往论文有何不同。" - P7。

然而，要判断一篇论文是写得不好，还是参与者缺少必要的上下文，是个难题："贡献表述不清——可能是因为论文的写法，也可能是因为我完全不了解他们可能在做什么。" - P19。

正如我们在 §4.3.1 中所指出的，当某个项目存在许多变体时，已有研究发现，用户会试图了解这些项目之间的差异 [119] ——我们在这里也观察到了这一点。用于探索文档中的自由文本差异、以帮助筛选、略读和理解文集的方法与系统仍未得到充分研究。然而，近期自然语言处理领域的研究开始探讨这一问题——Luu 等人[77] 探索了解释论文之间关系的方法，而 Rajagopal 等人 [107] 和 Kuznetsov 等人 [70] 则研究了其他应用上下文中的文本编辑。

4.5 数据科学家如何借助社交关系？

除了在 §4.2.2 中借助社会关系发现论文外，参与者也会通过其他方式依靠同行。

4.5.1 协作进行头脑风暴并理解论文。所有参与者都指出，他们会借助社会关系来理解文献（20/20）。他们提到，以论文为中心的小组讨论很有价值，可以帮助他们跟进文献、构思创意，或激发新的研究方向（12/20）：

> "至少在独立研究中，我会和别人聊聊自己的想法，看看方向是否正确。另一方面，我们每两周会进行一次头脑风暴，讨论至少一篇与我们的工作密切相关的论文，因此能获得很多意见……这可能有用，也可能是个局限，也可能行不通。"——P7

其他参与者提到，与合作者讨论有助于理解特定重要论文的细节（11/20）："如果我和别人进行一对一讨论，我们就会深入探讨人们为何在论文中做出某些决定，以及我们是否也应该采取相同做法。" - P10。最后，少数参与者（5/20）提到，与合作者分享笔记和文献，有助于确立思想来源：我通常会与合作者分享这些笔记，让他们了解我的这个想法或这个假设是从哪里来的- P19；也有助于核查信息的正确性：这也有助于合作者再次检查我的写作。- P6。

除了亲近同伴，参与者也会寻求较弱的社会关系和线上讨论（14/20）。在这里，他们提到了与同行互动带来的许多相同益处——在论坛上向专家征求建议：你会在 *Quora* 或 *Yahoo* 上找到由杰出数学家主导的回答。有人提出过与你类似的问题 *[*，而且*]* 有人为他们指出了一些相关研究 ——P17；如§4.3.2所述，这有助于确立论文的可信度；他们也会参与讨论，以了解论文的具体内容： *[ML Collective]* 的 *Discord* 频道里有很多志愿者和贡献者。如果你能找到一位对同类工作感兴趣的人……你们就可以讨论算法如何从这一步转到下一步，或者某个变量是什么 —— P14。

已有大量研究探讨协作式信息搜寻[87], , 其中一些研究考察了社交阅读 [97, 138], 以及通过协作式注释进行意义建构 [114, 143]。然而，在具体任务情境中对协作式阅读和意义建构的考察仍是研究不足的领域，而且可能会带来特定挑战，正如软件团队中的信息再利用 [75] 或协作验证主张的可信度 [41]等案例所示。

4.5.2 借助作者的力量。 虽然同行在线上和线下都很有帮助，参与者也发现与作者交流颇有价值（11/20）。有些人会主动联系作者（5/20），另一些人则提到，他们会较为被动地通过录制演讲以及 Twitter 或 Reddit 等论坛寻找作者（6/20）。直接交流有助于更全面地理解研究成果、与其他研究人员建立联系，有时还会促成富有成效的合作（2/20）：不久前，一篇新论文发表了，内容与我的工作非常相似，于是我给作者发了邮件。在这种情况下，我会这样做，以便展开交流，也能与从事相似工作的研究人员建立联系。我和他们通过邮件往来讨论了好几天，我想向他们确认我的方法 *[*与*]* 他们的方法之间的差异，这非常有帮助 - P18。其他人也发现，这种做法有助于探讨较少被讨论的方面：我会给他们发消息，祝贺他们发表了这项成果，并了解他们工作背后的一些秘诀 - P4。

在这方面，先前的研究探索了专家同行推荐 [102] 以及更广义的"人员推荐" [37]。然而，如何设置激励机制以促进这些互动，仍然存在重要挑战。例如，Breitinger 等人 [15] 指出，在发现正在进行的研究与确保其保密之间存在权衡。

积极参与对于深入理解很重要，但通过演讲或论坛帖子被动地与作者互动，比阅读他们的论文更容易理解其研究。参与者指出了视觉传播的价值，以及激励作者讲解其想法有助于理解： 有时候，了解他们的动机能让人形成很好的直觉，从而帮助理解。论文里可能讲得不够清楚，但他们在视频里向你解释时，就会有趣得多。我认为视觉传播 *[很有用]*, 他们的演讲采用了许多出色的设计和幻灯片动画.....这些内容无法通过论文传达。- P19。不过，有些人指出，只有"知名人士"的演讲才会被录制： 但是， *[演讲]* 并不总是适用，因为只有更知名人士的项目才会得到如此多的报道。不过，我想象自己进入一个全新的领域，并从身边最知名的人士开始 - P9。

对作者通过社交媒体进行互动的考察直到最近才开始受到研究 [35, 65]，将这方面的信息纳入意义建构辅助工具仍未得到充分探索。作者在论文之外的互动也体现了替代性出版形式的价值 [1, 47], ，然而，如何为出版商和作者建立激励机制仍然面临挑战 [124]。

5 讨论

在结果部分，我们考察了数据科学家审阅科学文献时的实践与挑战，并围绕信息需求的形成、查询的制定与搜索、通过略读与阅读评估搜索结果、综合一组文档，以及借助社会关系完成若干任务等环节展开讨论。接下来，我们指出贯穿这些主题的挑战，并推测未来可能提出解决方案的研究方向。

5.1 支持跨学科获取

数据科学领域的论文数量呈指数级增长，每两年翻一番 [66]。随着各领域的发展，围绕特定问题和方法形成了多个子学科，每个子学科都有各自的学科词汇和规范 [57]。如今，科学家个人需要在日益碎片化的学科中开展工作，其中只有一部分学科为他们所熟悉；但他们也知道，通过跨越多个学科开展工作，可以实现重大创新与进步 [111]。在所考察的信息搜索过程的每个阶段，我们的研究发现都反映出碎片化知识带来的挑战。未来，为跨学科研究的信息寻求和意义建构提供支持，可以采取多种形式。

为应对在陌生学科中进行搜索的挑战，人们已探索了查询推荐等辅助工具（§4.2.1）；然而，旨在让用户更充分地说明自身上下文的查询策略仍未得到充分探索。这类策略可以采用冗长查询 [2, 89], 、对话式和交互式搜索[5, 39] ，并结合跨领域检索机制的形式——Kang 等人 [60] 提出了一个初步示例。此外，在陌生学科中探索论文，可能会受益于展示"解释"，以帮助理解其可信度和相关性。尽管推荐系统中的解释已得到广泛研究 [141], ，但针对解释的考察跨学科探索中的解释仍是一个悬而未决的问题。也可以开发略读辅助工具。由于略读依赖于对某一学科写作规范的了解，这类辅助工具可以采用自适应文档布局 [54] ，或自动生成针对读者所属学科个性化定制的论文"常见问题解答" [8], 。同样，阅读辅助工具可以通过面向不同学科读者释义改写文本，或呈现理解某篇论文所必需的"先备"概念或论文 [73] ，从而帮助读者不受自身知识缺口的影响来判断论文质量。然而，由于科学家在认为具有挑战性的任务中会取得更好的学习成果，简化这一过程所涉及的权衡仍有待观察 [76, 127]。

5.2 促进对亲近同伴的依赖

在文献评阅的多个阶段，参与者都依赖亲近同伴——队友、实验室成员、合作者、受指导者和导师。他们从同行那里获得推荐、开展头脑风暴、判断论文的可信度，并理解论文的细节。数据科学家的协作实践此前已在代码和数据工作 [137] 的语境下得到考察——我们将这一研究拓展至创意的提出与理解。认真理解并支持这些实践很可能会带来成效，尤其是在远程工作日益普及的背景下 [135]。

信息寻求领域的主流研究采取了自我中心视角，而对利用社群的方法和系统的探索，大多见于早期关于协作式搜索 [87] 和群体推荐 [55]的研究。在当前情境下，重新探讨这一路径或许大有裨益——Piao 等人[99] 近期的研究为此提供了证据：他们发现，将"朋友纳入其中"能够让推荐系统给出更准确、更多样的推荐。重新探索早期将意义建构、阅读和发现融入协作式信息流阅读器 [3] 的研究，也有望利用科学家对同行的信任，帮助他们筛选和获取研究成果。Morris[87] 提出了另一条研究路径：用户更倾向于交替进行自我中心搜索和轻量级通信。这类研究的范例包括用于总结群聊 [136] 或协作式对话代理 [9] 的辅助工具，可用于头脑风暴。然而，开发这些工具需要精心设计——Brucks 和 Levav [16] 指出，虚拟交流（如视频会议）会抑制创意构思。同样，Inie 等人 [52] 指出，用户倾向于使用自己熟悉的学术工具链，因此新工具能否实现互操作性，对其普及至关重要。

5.3 利用论文的知识背景

尽管专家同行组成的紧密社区很重要，但参与者并不总能获得这样的支持。数据科学是一门不断发展、不断有新人加入的学科，这种情况在数据科学中也很可能出现 [56]。在这种情况下，人们会寻求各种其他资源来补充论文，例如论坛和社交媒体上的讨论、录制演讲和视频、博客以及代码。这些资源为探索提供了切入点，有助于确立论文的可信度，并帮助理解论文中缺失的细节和数学内容。传统上，这类信息（例如实体卡片、视频）被称为知识背景 ，已应用于网络搜索引擎结果页（SERP）中，并且

有助于用户做出具备信息素养的决策。Smith 和 Rieh[117] 主张更多地使用这种知识背景，而不是对其加以简化，也不应让搜索引擎“退到一旁”，以免干扰用户。在此，我们认同这一主张，并指出它对搜索文献的数据科学家有何价值。此外，我们还注意到，它对于增强已找到文档的阅读和略读辅助功能也很有价值。

关于如何将知识背景与学术搜索一同呈现，仍有许多问题有待探讨。TREC博客赛道的先前研究考察了博客检索，重点关注其主观内容——学术出版物的知识背景检索也需要类似的研究 [93]。这种信息的呈现方式也很可能需要进一步研究，例如，Levi等人 [72] 报告称，只有部分查询会受益于呈现来自社区问答网站的一组结果。此外，使用知识背景丰富学术搜索结果所带来的公平性影响仍有待探讨 [82]。除了用于信息寻求，这类知识背景也可能有助于略读与阅读辅助工具。Rachatasumrit等人 [105] 近期的研究提供了一个早期范例：将后续研究中的讨论内容置于论文页边空白处。不过，更复杂的增强方式仍有探索空间——例如，利用博客和视频帮助理解数学，利用代码推断缺失的细节，或利用社交媒体和博客辅助略读。这一领域仍未得到充分探索，信息的每种类型都会给检索、呈现及后续使用影响带来挑战。

6 未来工作与局限

我们指出了本研究的三项局限，每一项都指向未来的研究工作。这些局限源于我们的研究设计、参与者的选择，以及所探讨挑战的性质。首先，鉴于本研究仅持续一个时段，我们没有考察长期开展的活动，例如长期综合与撰写；考察这一方面也可能有助于进一步了解科学家的组织和笔记策略。这需要对参与者进行更长期的观察，并让他们进行自我反思。其次，在招募参与者时，我们注意到所有参与者都来自美国或欧盟，且能够充分接入基础设施；参与者群体更加多元，可能会带来不同的研究结果。此外，开展规模更大的研究也可能有所裨益。第三，在若干案例中，我们注意到与传播研究相关的激励机制所引发的挑战。虽然事后看来这些挑战是合理的，但本研究对其探讨不足，值得在未来进一步研究 [20]。最后，尽管这不太可能构成一项局限，我们也注意到，本研究是在COVID-19疫情结束后（2022年2月至10月）开展的，而疫情对工作实践的影响 [135] ——这很可能影响了我们的研究及其结果。

7 结论

学术出版物数量的指数级增长、影响力的不断扩大、大量新进入者的涌现，以及它可能对数据科学产生的影响，促使我们考察数据科学家在审阅文献时的信息寻求与意义建构实践。在这项研究中，我们对从工业界和学术机构招募的 20 位数据科学家开展了探索性访谈研究。我们首先了解他们查阅文献的目标，随后考察他们在搜索与发现、筛选搜索结果、略读与阅读，以及在多项任务中依赖同行方面的实践与挑战。我们的许多结果与先前研究的发现相印证，同时也综合了以往分散的研究成果。此外，我们的研究还揭示了一些特定发现，这些发现源于我们对数据科学家的关注，以及对信息寻求和意义建构的联合考察。研究结果凸显了碎片化的科学学科给信息搜索过程各阶段带来的挑战；我们认为，鉴于数据科学的跨学科性质，这对数据科学而言是一个特殊挑战。此外，我们还指出，阅读论文时缺少细节和数学内容也是一项挑战，这很可能是数据科学等计算学科的一个特点。通过联合考察信息寻求与意义建构，我们揭示了这两类活动交汇处的信息需求：由于论文的主张被夸大或重新包装，或是因为对某一学科不熟悉，人们希望确认论文的可信度；同时，他们也希望理解许多看似相似的论文之间渐进式差异。这很可能是数据科学中子学科拥挤且受到激励机制驱动所带来的结果。为应对这些挑战，我们发现参与者会借助科学论文周边的知识背景，包括代码、博客、演讲和论坛，也会依靠同行；我们的研究也考察了这些实践。最后，通过考察参与者面临的挑战及其应对机制，我们谨慎地推测了未来研究方向：既有望为这些挑战提供有意义的解决方案，也能为信息检索、自然语言处理、人机交互和计算机支持的协同工作社群提出有意义的科学问题。

致谢

感谢参与研究的参与者分享宝贵见解并付出宝贵时间。我们还要感谢 HCI-VISLab 的 Alyx Burns，在本项目的初期阶段，他担任高级人机交互课程的助教，为项目提供了大量指导。我们感谢美国国家科学基金会（资助编号：IIS-1922090）、陈-扎克伯格倡议（科学知识库构建项目）、美国国家GEM联盟和英特尔学者计划提供的部分资金支持。

参考文献

[1] ICLR 2022. 2022年。*ICLR 2022* 博客赛道。https://iclr-blog-track.github.io/ 访问日期：2022年10月15日。[2] Jafar Afzali、Aleksander Mark Drzewiecki 和 Krisztian Balog。2021年。POINTREC：叙事驱动的兴趣点推荐测试集。载于第*44*届国际*ACM SIGIR*信息检索研究与发展会议论文集 （虚拟会议，加拿大）（*SIGIR '21*）。计算机协会，美国纽约州纽约市，2478–2484。https: //doi.org/ 10 1145 3404835 3463243 。/。[3] Netta Aizenbud-Reshef、Ido Guy 和 Michal Jacovi。2009年。社区中的协作式信息流阅读。载于*2009*年*ACM*国际群体工作支持会议论文集 （美国佛罗里达州萨尼贝尔岛）（*GROUP '09*）。计算机协会，美国纽约州纽约市，277–280。https://doi.org/ 10 1145 1531674 1531716 。/。[4] Hamed Alhoori、Mohammed Samaka、Richard Furuta 和 Edward A Fox。2019年。学术社交媒体平台兴起后的学术信息行为模式剖析。《国际数字图书馆期刊》20，4（2019），369–389。https://doi.org/10.1007/s00799-018-0255-9[5] Mohammad Aliannejadi、Leif Azzopardi、Hamed Zamani、Evangelos Kanoulas、Paul Thomas 和 Nick Craswell。2021年。分析对话式搜索中的混合式主动交互与搜索

策略。载于第30届ACM国际信息与知识管理会议论文集（虚拟会议，澳大利亚昆士兰州）(CIKM '21)。计算机协会，美国纽约州纽约市，
16 26 10 1145 3459637 3482231 –。https://doi.org/。/。
[6] Lorin W Anderson 和 David R Krathwohl。2001。学习、教学与评估分类法：布鲁姆教育目标分类法的修订版。朗文。[7] Kumaripaba Athukorala、Eve Hoggan、Anu Lehtiö、Tuukka Ruotsalo 和 Giulio Jacucci。2013年。计算机科学家的信息寻求行为：电子文献检索工具面临的挑战。《美国信息科学与技术学会论文集》50，1（2013），1–11。https://doi.org/10.1002/meet.14505001041 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/meet.14505001041[8] Tal August、Lucy Lu Wang、Jonathan Bragg、Marti A Hearst、Andrew Head 和 Kyle Lo。2022年。Paper Plain：借助自然语言处理让医疗研究论文更易于医疗保健消费者理解。arXiv预印本 arXiv: 2203 00130 2022 10 48550 2203 00130 。
[9] （）。https://doi.org/。/arXiv。。Sandeep Avula、Gordon Chadwick、Jaime Arguello 和 Robert Capra。2018年。SearchBots：协作式搜索期间用户与聊天机器人的互动。载于2018年人类信息交互与检索会议论文集（美国新泽西州新不伦瑞克）（CHIIR '18）。计算机协会，美国纽约州纽约市，
52 61 10 1145 3176349 3176380 –。https://doi.org/。/。
[10] Leif Azzopardi。2021。搜索中的认知偏差：信息检索中认知偏差的综述与反思。载于2021年人类信息交互与检索会议论文集（澳大利亚首都领地堪培拉）（CHIIR '21）。计算机协会，美国纽约州纽约市，27–37。https://doi.org/ 10 1145 3406522 3446023 。/。
[11] Alina Beygelzimer、Emily Fox、Florence d'Alché Buc 和 Hugo Larochelle。2019年。我们从NeurIPS 2019数据中学到了什么。
h
t
t
p
s://
neuripsconf.medium.com/what-we-learned-from-neurips-2019-data-111ab996462c
[12] Abeba Birhane、Pratyusha Kalluri、Dallas Card、William Agnew、Ravit Dotan 和 Michelle Bao。2022年。机器学习研究中编码的价值观。载于2022年ACM公平性、问责制与透明度会议（韩国首尔）（FAccT '22）。计算机协会，美国纽约州纽约市，
173 184 10 1145 3531146 3533083 –。https://doi.org/。/。
[13] David M. Blei 和 Padhraic Smyth。2017。科学与数据科学。《美国国家科学院院刊》114，33（2017），8689–8692。10 1073 1702076114（），–。https://doi.org/。/pnas。arXiv:https://www.pnas.org/doi/pdf/ 10 1073 1702076114 。/pnas。
[14] Marcel Bollmann 和 Desmond Elliott。2020。忘记引用较早论文：对ACL Anthology的分析。载于第58届计算语言学协会年会论文集。计算语言学协会，线上，
7819 7827 10 18653 1 2020 699 –。
[15] https://doi.org/。/v /。acl-main。Corinna Breitinger、Patrick Wortner、Bela Gipp 和 Harald Reiterer。2019。'协作已为时过晚'：发现进行中研究的挑战。载于2019年ACM/IEEE数字图书馆联合会议（JCDL）。134–137。https: //doi.org/ 10 1109 2019 00028 。/
[16] JCDL。。Melanie S Brucks 和 Jonathan Levav。2022年。虚拟交流抑制创意生成。《自然》605，7908（2022），108–112。https://doi.org/10. 1038/s41586-022-04643-y [17] Hilary Bussell、Jennifer Schnabel 和 Amanda K. Rinehart。2020年。满足研究生需求：对学科差异的探索。《公共服务季刊》16，4（2020），213–233。https://doi.org/10.1080/15228959.2020.1818663 arXiv:https://doi.org/ 10 1080 15228959 2020 1818663 。/。。Arthur Câmara、Nirmal
[18] Roy、David Maxwell 和 Claudia Hauff。2021年。借助教学支架通过搜索学习。载于2021年人类信息交互与检索会议论文集（澳大利亚首都领地堪培拉）（CHIIR '21）。计算机协会，美国纽约州纽约市，209–218。https: //doi.org/ 10 1145 3406522 3446012 。/。
[19] Joel Chan、Joseph Chee Chang、Tom Hope、Dafna Shahaf 和 Aniket Kittur。2018年。SOLVENT：用于发现研究论文之间类比的混合主动系统。《ACM人与计算机交互论文集》2，CSCW，文章31（2018年11月），页数。https://doi.org。/。Jiyoo Chang 和 Christine Custis。理解机器学习文档的实施挑战。载于算法、机制与优化中的公平与可及性（美国弗吉尼亚州阿灵顿）（EAAMO '22）。计算机协会，美国纽约州纽约市，文章16，8页。https://doi.org/ 10 1145 3551624 3555301 。/。Duen Horng Chau、
[21] Aniket Kittur、Jason I. Hong 和 Christos Faloutsos。2011年。Apolo：通过结合丰富的用户交互与机器学习理解大型网络数据。载于SIGCHI计算机系统中的人为因素会议论文集（加拿大不列颠哥伦比亚省温哥华）（CHI '11）。计算机协会，美国纽约州纽约市，167–176。https://doi.org/10. 1145/1978942.1978967[22] Kiroong Choe、Seokweon Jung、Seokhyeon Park、Hwajung Hong 和 Jinwook Seo。2021年。Papers101：支持新手研究者在文献综述工作流程中的发现过程。载于2021年IEEE第14届PacificVis研讨会（PacificVis）。176–180。https://doi.org/10.1109/PacificVis52677.2021.00037
[23] 科学的交汇领域。美国国家科学院院刊 118, 41 (2021)。https://doi.org/ 10 1073 2021636118 。/pnas。Anamaria Crisan、Brittany
[24] Fiore-Gartland 和 Melanie Tory。2020。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* 可视化与计算机图形学汇刊 27, 2 (2020), 1860–1870。https://doi.org/ 10 1109 2020 3030340 。/TVCG。. Cecilia di
[25] Sciascio、Eduardo Veas、Jordan Barria-Pineda 和 Colleen Culley。2020。理解控制与透明度对作为学习的搜索的影响。载于第*25*届国际智能用户界面会议论文集 （意大利卡利亚里）（*IUI '20*）。计算机协会，美国纽约州纽约市，
498 509 10 1145 3377325 3377524 –。
[26] https://doi.org/。/。Cheyenne Dosso、Lynda Tamine、Pierre-Vincent Paubel 和 Aline Chevalier。2022。专业知识对复杂学习任务求解过程中查询构建策略的影响：一项针对医学与计算机科学专业学生的研究。载于国际人类工效学协会第*21*届大会论文集（*IEA 2021*），Nancy L. Black、W. Patrick Neumann 和 Ian Noy（编）。施普林格国际出版公司，Cham，621–627。https://doi.org/10.1007/978-3-030-74614-8 77[27] Marcel Dunaiski、Gillian J Greene 和 Bernd Fischer。2017。使用交互式标签云可视化探索学术出版物和引文数据。《科学计量学》110, 3 (2017), 1539–1571。https://doi.org/10.1007/s11192-016- 2236-3[28] Marian Dörk、Nathalie Henry Riche、Gonzalo Ramos 和 Susan Dumais。2012。PivotPaths：漫游于分面信息空间。*IEEE* 可视化与计算机图形学汇刊 18, 12 (2012), 2709–2718。https://doi.org/ 10.1109/TVCG.2012.252[29] Debra Engel、Sarah Robbins 和 Christina Kulp。2011。工程学教师的信息寻求习惯。《学院与研究图书馆》72, 6 (2011), 548–567。https://doi.org/ 10 5860 155 。/crl-
[30] Michael Färber 和 Ann-Kathrin Leisinger。2021。*DataHunter*：一个基于科学问题描述查找数据集的系统。计算机协会，美国纽约州纽约市，749–752。https://doi.org/10.1145/3460231. 3478882[31] Michael Färber 和 Ann-Kathrin Leisinger。2021。DataHunter：一个基于科学问题描述查找数据集的系统。载于第*15*届*ACM*推荐系统会议论文集 （荷兰阿姆斯特丹）（*RecSys '21*）。计算机协会，美国纽约州纽约市，749–752。https://doi.org/10.1145/3460231.3478882[32] Raymond Fok、Andrew Head、Jonathan Bragg、Kyle Lo、Marti A Hearst 和 Daniel S Weld。2022。Scim：用于交互式、多轮略读科学论文的智能分面式重点内容。（2022）。https://doi.org/10.48550/arXiv.2205. 04561[33] Luanne Freund、Rick Kopak 和 Heather O'Brien。2016。文本环境对阅读理解的影响：对作为学习的搜索的启示。《信息科学杂志》42, 1 (2016), 79–93。https://doi.org/10.1177/ 0165551515614472[34] Timnit Gebru、Jamie Morgenstern、Briana Vecchione、Jennifer Wortman Vaughan、Hanna Wallach、Hal Daumé Iii 和 Kate Crawford。2021。数据集数据说明表。*Commun. ACM* 64, 12 (2021), 86–92。https://doi.org/10.1145/ 3458723[35] Katy Ilonka Gero、Vivian Liu、Sarah Huang、Jennifer Lee 和 Lydia B. Chilton。2021。推文教程为何引人入胜：专家如何在Twitter上讲解复杂主题。*ACM*人与计算机交互论文集 5, CSCW2, 文章422（
2021 26 10 1145 3479566 年10月），页数。
[36] https://doi.org/。/ Charles J Gomez、Andrew C Herman 和 Paolo Parigi。2022。在全球科学领域，领先国家获得的引文数量日益超过开展类似研究的其他国家。《自然·人类行为》（2022），1–11。https://doi.org/10. 1038/s41562-022-01351-5[37] Ido Guy 和 Luiz Pizzato。2016。人员推荐教程。载于第*10*届*ACM*推荐系统会议论文集（美国马萨诸塞州波士顿）（*RecSys '16*）。计算机协会，美国纽约州纽约市，
431 432 10 1145 2959100 2959196 –。https://doi.org/。/。Han L.
[38] Han、Junhang Yu、Raphael Bournet、Alexandre Ciorascu、Wendy E. Mackay 和 Michel Beaudouin-Lafon。2022。Passages：跨文档文本交互。载于*2022*年计算机系统中的人为因素*CHI*会议论文集 （美国路易斯安那州新奥尔良）（*CHI '22*）。计算机协会，美国纽约州纽约市，文章338，共17页。https://doi.org/ 10 1145 3491102 3502052 。/。Abram Handler、Narges Mahyar 和
[39] Brendan O'Connor。2022。ClioQuery：用于全面调查历史新闻档案的交互式面向查询的文本分析。*ACM Trans. Interact. Intell. Syst.* 12, 3, 文章22（
2022 49 10 1145 3524025 年7月），页数。https://doi.org/。/
[40] Jiangen He 、 Qing Ping 、 Wen Lou 和 Chaomei Chen。2019。PaperPoles：借助引文链接促进对科学出版物的自适应可视化探索。《信息科学与技术协会期刊》70, 8 (2019), 843–857。https://doi.org/10.1002/asi.24171

[41] Lu He 和 Changyang He。2022。Help Me #DebunkThis：剖析个人与社区在信息可信度评估中的协作工作。ACM人与计算机交互论文集 6, CSCW2, 文章413（2022年11月），31页。https://doi.org/10.1145/3555138[42] Andrew Head、Kyle Lo、Dongyeop Kang、Raymond Fok、Sam Skjonsberg、Daniel S. Weld 和 Marti A. Hearst。2021。通过即时、位置敏感的术语与符号定义增强科学论文。载于2021年CHI计算系统中的人为因素会议论文集（日本横滨）（CHI '21）。计算机协会，美国纽约州纽约市，文章 413, 18 10 1145 3411764 3445648 页。[43] https://doi.org/。/。Andrew Head、Amber Xie 和 Marti A. Hearst。. 数学增强：作者如何运用新颖的视觉设计实践提高公式的可读性。载于2022年计算机系统中的人为因素CHI会议论文集（美国路易斯安那州新奥尔良）（CHI '22）。计算机协会，美国纽约州纽约市，文章491，共18页。https://doi.org/ 10 1145 3491102 3501932 。/。Kevin Heffernan 和 Simone Teufel。2018。识别科学文本中的问题与解决方案。《科学计量学》116, 2 (2018), 1367–1382。https://doi.org/10. 1007/s11192-018-2718-6[45] Florian Heimerl、Qi Han 和 Steffen Koch。2016。CiteRivers：引文模式的视觉分析。IEEE 可视化与计算机图形学汇刊 22, 1 2016 190 199 10 1109 2015 2467621 （1月），–。[46] https://doi.org/。/TVCG。. Orland Hoeber、Dolinkumar Patel 和 Dale Storie。. 学术搜索场景与信息寻求行为研究。载于2019年人类信息交互与检索会议论文集（英国苏格兰格拉斯哥）（CHIIR '19）。计算机协会，美国纽约州纽约市，231 235 10 1145 3295750 3298943 –。https://doi.org/。/。Fred Hohman、Matthew Conlen、Jeffrey Heer 和 Duen Horng (Polo) Chau。2020。通过交互式文章进行交流。Distill（2020）。https://doi.org/ 10.23915/distill.00028 https://distill.pub/2020/communicating-with-interactive-articles.[48] Kenneth Holstein、Jennifer Wortman Vaughan、Hal Daumé、Miro Dudik 和 Hanna Wallach。2019。提升机器学习系统的公平性：行业从业者需要什么？载于2019年CHI计算系统中的人为因素会议论文集（英国苏格兰格拉斯哥）（CHI '19）。计算机协会，美国纽约州纽约市，1–16。https://doi. org/ 10 1145 3290605 3300830 。/。Chien-yu Huang、Arlene Casey、Dorota Głowacka 和 Alan Medlar。2019。大纲中的空白：摘要质量因学科而异及其对科学文献检索的影响（CHIIR '19）。计算机协会，美国纽约州纽约市，289 293 10 1145 3295750 3298953 –。[50] https://doi.org/。/。Chien-yu Huang、Arlene Casey、Dorota Głowacka 和 Alan Medlar。2019。大纲中的空白：摘要质量因学科而异及其对科学文献检索的影响。载于2019年人类信息交互与检索会议论文集（英国苏格兰格拉斯哥）（CHIIR '19）。计算机协会，美国纽约州纽约市，289–293。https://doi.org/ 10 1145 3295750 3298953 。/。Sharon Favaro Ince、Christopher Hoadley 和 Paul A. Kirschner。2018。博士生学术工作流程中的搜索实践研究。载于2018年人类信息交互与检索会议论文集（美国新泽西州新不伦瑞克）（CHIIR '18）。计算机协会，美国纽约州纽约市，245 248 10 1145 3176349 3176877 –。[52] https://doi.org/。/。Nanna Inie、Jonas Frich 和 Peter Dalsgaard。. 研究人员如何管理创意。载于创造力与认知（意大利威尼斯）。计算机协会，美国纽约州纽约市，83 96 10 1145 3527927 3532813 –。[53] https://doi.org/。/。Emi Ishita、Yasuko Hagiwara、Yukiko Watanabe 和 Yoichi Tomiura。. 研究人员在选择学术文献时会查看搜索结果的哪些部分？载于第18届ACM/IEEE数字图书馆联合会议论文集（美国得克萨斯州沃思堡）（JCDL '18）。计算机协会，美国纽约州纽约市，345 346 10 1145 3197026 3203867 –。[54] https://doi.org/。/。Charles Jacobs、Wil Li、Evan Schrier、David Bargeron 和 David Salesin。. 自适应文档布局。Commun. ACM 47, 8（2004年8月），60–66。https: // doi.org/ 10 1145 1012037 1012063 。/。Anthony Jameson 和 Barry Smyth。2007。群体推荐。载于《自适应网络》。施普林格，596 627 10 1007 978 3 540 72079 9_20 –。[56] https://doi.org/。/ - - - - Nicole Janssen。. 数据科学人才缺口：为何存在，以及企业可以采取哪些措施。https://www.forbes.com/sites/forbestechcouncil/2022/10/11/the-data-science-talent-gap-why-it-exists-and-what-businesses-can-do-about-it/[57] David Jurgens、Srijan Kumar、Raine Hoover、Dan McFarland 和 Dan Jurafsky。2018。通过引文框架衡量科学领域的演变。《计算语言学学会汇刊》6 (2018), 391–406。https://doi.org/ 10 1162 _ _00028 。/tacl a Hyeonsu Kang、Joseph Chee Chang、Yongsung Kim 和 Aniket Kittur。2022。Threddy：用于科学文献个性化线程式探索与组织的交互系统。载于第35届ACM用户界面软件与技术年度研讨会论文集（美国俄勒冈州本德）（UIST '22）。计算机协会，美国纽约州纽约市，文章94，15页。https://doi.org/10.1145/3526113.3545660[59] Hyeonsu B Kang、Rafal Kocielnik、Andrew Head、Jiangjiang Yang、Matt Latzke、Aniket Kittur、Daniel S Weld、Doug Downey 和 Jonathan Bragg。2022年。从认识的人到阅读的内容：利用隐式社交网络增强科学推荐。载于《*2022*年计算机系统中的人为因素*CHI*会议论文集》（美国路易斯安那州新奥尔良）（*CHI '22*）。计算机协会，纽约，美国，文章302，23页。https://doi.org/ 10 1145 3491102 3517470 。/。Hyeonsu B Kang、Sheshera Mysore、Kevin J Huang、Haw-Shiuan Chang、Thorben Prein、Andrew McCallum、Aniket Kittur 和 Elsa Olivetti。2022年。通过跨知识领域检索增强科学创造力。载于*NAACL*人机交互与自然语言处理衔接研讨会第二届 *2022* 10 48550 2206 01328 。https://doi.org/。/arXiv。/。[61] Hyeonsu B. Kang、Xin Qian、Tom Hope、Dafna Shahaf、Joel Chan 和 Aniket Kittur。2022年。利用类比搜索引擎增强科学创造力。《*ACM*计算机与人类交互汇刊》（2022年3月）。https://doi.org/10.1145/3530013 已接收。[62] Mary Beth Kery、Bonnie E. John、Patrick O'Flaherty、Amber Horvath 和 Brad A. Myers。2019年。迈向有效的数据科学家探索：寻找以往的分析选择。载于《*2019*年*CHI*计算系统中的人为因素会议论文集》（英国苏格兰格拉斯哥）（*CHI '19*）。计算机协会，纽约，美国，1 13 10 1145 3290605 3300322 –。[63] https://doi.org/。/。Ian A. Knight、Max L. Wilson、David F. Brailsford 和 Natasa Milic-Frayling。2019年。受困于数据：医学系统综述的认知工作分析。载于《*2019*年人类信息交互与检索会议论文集》（英国苏格兰格拉斯哥）（*CHIIR '19*）。计算机协会，纽约，美国，203–212。https://doi.org/10.1145/3295750.3298937[64] Laura M. Koesten、Emilia Kacprzak、Jenifer F. A. Tennison 和 Elena Simperl。2017年。结构化数据工作的艰辛历程：信息寻求行为研究。载于《*2017*年*CHI*计算系统中的人为因素会议论文集》（美国科罗拉多州丹佛）（*CHI '17*）。计算机协会，纽约，美国，1277–1289。https://doi.org/ 10 1145 3025453 3025838 。/。Kaisu Koivumäki、Timo Koivumäki 和 Erkki Karvonen。2020年。“在社交媒体上，科学似乎更具人情味”：探索作为数字科学传播者的研究人员。《媒体与传播》8，2（2020），425–439。https://doi.org/10.17645/mac.v8i2.2812[66] Mario Krenn、Lorenzo Buffoni、Bruno Coutinho、Sagi Eppel、Jacob Gates Foster、Andrew Gritsevskiy、Harlin Lee、Yichao Lu、Joao P Moutinho、Nima Sanjabi 等。2022年。用人工智能预测人工智能的未来：在指数级增长的知识网络中进行高质量链接预测。（2022年）。https://doi.org/10.48550/arXiv.2210.00881[67] Satyapriya Krishna、Tessa Han、Alex Gu、Javin Pombra、Shahin Jabbari、Steven Wu 和 Himabindu Lakkaraju。2022年。可解释机器学习中的分歧问题：从业者视角。*arXiv*预印本 *arXiv:2202.01602*（2022 10 48550 2202 01602 ）。https://doi.org/。/arXiv。/。Sean Kross 和 Philip J. Guo。数据科学从业者在行业和学术界的教学：期望、工作流程与挑战。载于《*2019*年*CHI*计算系统中的人为因素会议论文集》（英国苏格兰格拉斯哥）（*CHI '19*）。计算机协会，纽约，美国，1 14 10 1145 3290605 3300493 –。[69] https://doi.org/。/。Carol Kuhlthau。《寻求意义：图书馆与信息服务的过程方法》Ablex Publishing。（1993年1月）。[70] Ilia Kuznetsov、Jan Buchmann、Max Eichler 和 Iryna Gurevych。2022年。修改后重投：同行评审中基于文本协作的互文模型。*arXiv*预印本 *arXiv:2204.10805*（2022年）。https://doi.org/10.48550/arXiv.2204.10805 [71] Esther Landhuis。2016年。535，7612，2016，457，458，10，1038，7612，457。科学文献：信息过载。《自然》（），–。https://doi.org/。/nj-a[72] Or Levi、Ido Guy、Fiana Raiber 和 Oren Kurland。2018年。搜索结果页面上的选择性聚类呈现。《*ACM*信息系统汇刊》36，3，文章28（2月 2018 42 10 1145 3158672 ），页。https://doi.org/。/Irene Li、Alexander R Fabbri、Robert R Tung 和 Dragomir R Radev。2019年。我应该先学什么：为自然语言处理教育和先修关系链学习推出lecturebank。载于*AAAI*人工智能会议论文集，第 33 01 33016674 卷。–。https://doi.org/。/ aaai.v i。Kevin Li、Haoyang Yang、Anish Upadhayay、Zhiyan Zhou、Jon Saad-Falcon 和 Duen Horng Chau。2021年。Argo Scholar：在浏览器中交互式视觉探索文献。（ 2021 10 48550 2110 14060 ）。[75] https://doi.org/。/arXiv。/。Michael Xieyang Liu、Aniket Kittur 和 Brad A. Myers。是重复使用还是不重复使用？评估摘要知识的框架与系统。《*ACM*人与计算机交互论文集》5，CSCW1，文章166（2021年4月），35页。https://doi.org/ 10 1145 3449240 。/Ying-Hsang Liu、Paul Thomas、Tom Gedeon 和 Nicolay Rusnachenko。2022年。生物医学搜索的搜索界面：注视、用户感知、搜索行为和搜索性能之间有何关系？载于*ACM SIGIR*人类信息交互与检索会议（德国雷根斯堡）（CHIIR

'22）。计算机协会，纽约，美国，78–89。https://doi.org/10.1145/3498366.3505769。/。Kelvin Luu、Xinyi Wu、Rik Koncel-Kedziorski、Kyle Lo、Isabel Cachola 和 Noah A. Smith。2021年。解释科学文献之间的关系。载于第59届计算语言学协会年会和第11届自然语言处理国际联合会议论文集（第1卷：长篇论文）。计算语言学协会，线上，2130–2144。https://doi.org/10.18653/v1/2021.acl-long.166。/。Gary Marchionini。搜索、意义建构与学习：弥合差距。《信息与学习科学》（2018年）。https://doi.org/10.1108/ILS-06-2018-0049 Moshe Mash、Stephanie Rosenthal 和 Reid Simmons。2021年。DSWorkFlow：记录数据科学家工作流程的框架。计算机协会，纽约，美国。https://doi.org/10.1145/3411763.3451683。/。Justin Matejka、Tovi Grossman 和 George Fitzmaurice。Paper Forager：支持快速探索研究文献集合。载于Graphics Interface 2021。https://openreview.net/forum?id=QhN4tUZd8r Lori McCay-Peet、Anabel Quan-Haase 和 Dagmar Kern。2015年。数字图书馆中的探索式搜索：对界面功能用途及作用的初步考察。信息科学与技术协会论文集52, 1（2015），1–4。https://doi.org/10.1002/2.2015.14505201070。/pra。。Graham McDonald、Craig Macdonald 和 Iadh Ounis。学术搜索中有效公平排序的搜索结果多样化。《信息检索期刊》25，1（2022年），1–26。https://doi.org/10.1007/s10791-021-09399-z Alan Medlar、Jing Li 和 Dorota Głowacka。2021年。将查询建议作为摘要用于探索式搜索。载于2021年人类信息交互与检索会议论文集（澳大利亚首都领地堪培拉）（CHIIR '21）。计算机协会，纽约，美国，119–128。https://doi.org/10.1145/3406522.3446020。/。Margaret Mitchell、Simone Wu、Andrew Zaldivar、Parker Barnes、Lucy Vasserman、Ben Hutchinson、Elena Spitzer、Inioluwa Deborah Raji 和 Timnit Gebru。2019年。用于模型报告的模型卡。载于公平性、问责制与透明度会议论文集（美国佐治亚州亚特兰大）（FAT* '19）。计算机协会，纽约，美国，220–229。https://doi.org/10.1145/3287560.3287596。/。John S Morabito 和 Joel Chan。管理学术知识综合过程中的上下文：流程模式与系统机制。载于创造力与认知（意大利线上活动）。计算机协会，纽约，美国，文章39, 5页。https://doi.org/10.1145/3450741.3465244。/。Felipe Moraes、Sindunuraga Rikarno Putra 和 Claudia Hauff。将作为学习活动的搜索与教师设计的学习进行对比。载于第27届ACM信息与知识管理国际会议论文集（意大利都灵）（CIKM '18）。计算机协会，纽约，美国，167–176。https://doi.org/10.1145/3269206.3271676。/。Meredith Ringel Morris。重新审视协作式搜索。载于2013年计算机支持的协同工作会议论文集（美国得克萨斯州圣安东尼奥）（CSCW '13）。计算机协会，纽约，美国，1181–1192。https://doi.org/10.1145/2441776.2441910 Michael Muller、Ingrid Lange、Dakuo Wang、David Piorkowski、Jason Tsay、Q. Vera Liao、Casey Dugan 和 Thomas Erickson。2019年。数据科学工作者如何处理数据：发现、采集、整理、设计与创建。载于2019年CHI计算系统中的人为因素会议论文集（英国苏格兰格拉斯哥）（CHI '19）。计算机协会，纽约，美国，1–15。https://doi.org/10.1145/3290605.3300356。/。Sheshera Mysore、Tim O'Gorman、Andrew McCallum 和 Hamed Zamani。2021年。CSFCube——一个用于基于示例的分面查询的计算机科学研究论文测试集。载于第三十五届神经信息处理系统会议数据集与基准赛道（第二轮）。https://doi.org/10.48550/arXiv.2103.12906 Ya R Nedumov 和 Sergei D Kuznetsov。2019年。科学文章的探索式搜索。《程序设计与计算机软件》45，7（2019年），405–416。https://doi.org/10.1134/S0361768819070089 Xi Niu、Bradley M Hemminger、Cory Lown、Stephanie Adams、Cecelia Brown、Allison Level、Merinda McLure、Audrey Powers、Michele R Tennant 和 Tara Cataldo。2010年。美国学术研究人员信息寻求行为的全国性研究。《美国信息科学与技术学会会刊》61, 5（2010），869–890。https://doi.org/10.1002/asi.21307。/。Fatima W. Nosheen、Irfan Ali 和 Shazia Yasmeen。寻回已找到的内容。《信息与学习科学》119，12（2018年），712–720。https://doi.org/10.1108/07-2018-0064。/ILS---Iadh Ounis、Craig MacDonald 和 Ian Soboroff。2021年。关于TREC博客赛道。《AAAI国际网络与社交媒体会议论文集》2，1（9月2021），93–101。https://ojs.aaai.org/index.php/ICWSM/article/view/18622）， –。Elisabeth Pain。如何跟进科学文献。《科学职业》30（2016年）。https://www.science.org/content/article/how-keep-scientific-literature-rev2 Srishti Palani、Zijian Ding、Stephen MacNeil 和 Steven P. Dow。2021年。"主动搜索"假设：搜索策略与创造性学习的关系。计算机协会，美国纽约州纽约市，325–329。https://doi.org/10.1145/3406522.3446046。/。Srishti Palani、Zijian Ding、Austin Nguyen、Andrew Chuang、Stephen Mac-Neil 和 Steven P. Dow。2021年。CoNotate：基于笔记推荐查询，促进知识发现。载于《2021年计算机系统中的人为因素CHI会议论文集》（日本横滨）（CHI '21）。计算机协会，美国纽约州纽约市，文章编号726，共14页。https://doi.org/10.1145/3411764.3445618。/。Jennifer Pearson、Tom Owen、Harold Thimbleby 和 George R. Buchanan。2012年。Co-Reading：协作小组阅读研究。载于《第12届ACM/IEEE-CS数字图书馆联合会议论文集》（美国华盛顿特区）（JCDL '12）。计算机协会，美国纽约州纽约市，325–334。https://doi.org/10.1145/2232817.2232876。/。Timothy Persons。数据与分析创新：新兴机遇与挑战。（2016）。[99] Jinghua Piao、Guozhen Zhang、Fengli Xu、Zhilong Chen、Yu Zheng、Chen Gao 和 Yong Li。2021年。将好友纳入推荐系统：一项探索性研究。《ACM人与计算机交互论文集》5，CSCW2，文章编号439（2021年26月），10.1145/3479583（10月），页码。https://doi.org/。/ Peter Pirolli 和 Stuart Card。通过认知任务分析识别出的意义建构过程及分析师技术的杠杆点。载于《国际情报分析会议论文集》，第5卷。美国弗吉尼亚州麦克莱恩，2–4。[101] Antoine Ponsard、Francisco Escalona 和 Tamara Munzner。2016年。PaperQuest：支持文献综述的可视化工具。载于《2016年CHI计算机系统中的人为因素会议扩展摘要集》（美国加利福尼亚州圣何塞）（CHI EA '16）。计算机协会，美国纽约州纽约市，2264–2271。https://doi.org/10.1145/2851581.2892334 Jason Portenoy、Marissa Radensky、Jevin D West、Eric Horvitz、Daniel S Weld 和 Tom Hope。2022年。打破科学过滤气泡：通过发现新作者促进创新。载于《2022年计算机系统中的人为因素CHI会议论文集》（美国路易斯安那州新奥尔良）（CHI '22）。计算机协会，美国纽约州纽约市，文章编号309，共13页。https://doi.org/10.1145/3491102.3501905。/。Gil Press。数据科学简史。https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/?sh=310bcc2a55cf Sihang Qiu、Ujwal Gadiraju 和 Alessandro Bozzon。2020年。迈向令人难忘的信息检索。载于《2020年ACM SIGIR信息检索理论国际会议论文集》（线上活动，挪威）（ICTIR '20）。计算机协会，美国纽约州纽约市，69–76。https://doi.org/10.1145/3409256.3409830。/。Napol Rachatasumrit、Jonathan Bragg、Amy X. Zhang 和 Daniel S Weld。2022年。CiteRead：将局部引文语境融入科学论文阅读。载于《第27届国际智能用户界面会议》（芬兰赫尔辛基）（IUI '22）。计算机协会，美国纽约州纽约市，707–719。https://doi.org/10.1145/3490099.3511162。/。Behnam Rahdari 和 Peter Brusilovsky。2021年。PaperExplorer：面向会议论文集的个性化探索式搜索。。载于IUI研讨会。[107] Dheeraj Rajagopal、Xuchao Zhang、Michael Gamon、Sujay Kumar Jauhar、Diyi Yang 和 Eduard Hovy。2022年。一份文档，多次修订：用于编辑意图分类与描述的数据集。载于《第十三届语言资源与评测会议论文集》。欧洲语言资源协会，法国马赛，5517–5524。https://aclanthology.org/2022.lrec-1.591 Nirmal Roy、Felipe Moraes 和 Claudia Hauff。2020年。探索用户在搜索会话中的学习收益。载于《2020年人类信息交互与检索会议论文集》（加拿大温哥华）（CHIIR '20）。计算机协会，美国纽约州纽约市，432–436。https://doi.org/10.1145/3343413.3378012 Nirmal Roy、Manuel Valle Torre、Ujwal Gadiraju、David Maxwell 和 Claudia Hauff。2021年。记录高亮内容：在搜索即学习环境中融入主动阅读工具。载于《2021年人类信息交互与检索会议论文集》（澳大利亚首都领地堪培拉）（CHIIR '21）。计算机协会，美国纽约州纽约市，229–238。https://doi.org/10.1145/3406522.3446025。/。Daniel M. Russell、Mark J. Stefik、Peter Pirolli 和 Stuart K. Card。1993年。意义建构的成本结构。载于《INTERACT '93与CHI '93计算机系统中的人为因素会议论文集》（荷兰阿姆斯特丹）（CHI '93）。计算机协会，美国纽约州纽约市，269–276。https://doi.org/10.1145/169059.169209。/。Andrey Rzhetsky、Jacob G. Foster、Ian T. Foster 和 James A. Evans。2015年。选择实验以加速集体发现。《美国国家科学院院刊》112，47（2015），14569–14574。https://doi.org/10.1073/pnas.1509757112。/pnas. Hemant Kumar Sahu 和 Surya Nath Singh。天文学／天体物理学科学家的信息寻求行为。载于《Aslib会议录》。Emerald Group Publishing Limited。https://doi.org/10.1108/00012531311313961 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1509757112。/。

[113] Dafna Shahaf、Carlos Guestrin 和 Eric Horvitz。2012年。科学地铁图。载于《第18届ACM SIGKDD知识发现与数据挖掘国际会议论文集》（中国北京）（KDD '12）。计算机协会，美国纽约州纽约市，1122–1130。https: //doi.org/ 10 1145 2339530 2339706 . / . Rina Shaikh-Lesko。网页注释工具Hypothesis达成一个里程碑。《自然》569，7756（2019），295–296。[115] Namit Shetty。2017年。详细查询的查询建议。 https://scholar.googleblog.com/2017/08/query-suggestions-for-detailed-queries.html [116] Catherine L. Smith 和 Paul B. Kantor。2008年。用户适应：从糟糕的系统中获得良好结果。载于《第31届年度国际ACM SIGIR信息检索研究与发展会议论文集》（新加坡）（SIGIR '08）。计算机协会，美国纽约州纽约市， 147 154 10 1145 1390334 1390362 –。https://doi.org/。/。 Catherine L. Smith 和 Soo Young Rieh。搜索系统中的知识情境：迈向信息素养行为。载于《2019年人类信息交互与检索会议论文集》（英国苏格兰格拉斯哥）（CHIIR '19）。计算机协会，美国纽约州纽约市，55–62。https://doi.org/ 10 1145 3295750 3298940 . / . Ayah Soufan、Ian Ruthven 和 Leif Azzopardi。2022年。搜索文献：对探索式搜索任务的分析。载于《ACM SIGIR人类信息交互与检索会议》（德国雷根斯堡）（CHIIR '22）。计算机协会，美国纽约州纽约市，146–157。https://doi.org/10.1145/3498366.3505818[119] Sruti Srinivasa Ragavan、Sandeep Kaur Kuttal、Charles Hill、Anita Sarma、David Piorkowski 和 Margaret Burnett。2016年。在过多相似变体中搜寻。载于《2016年CHI计算机系统中的人为因素会议论文集》（美国加利福尼亚州圣何塞）（CHI '16）。计算机协会，美国纽约州纽约市，3509–3521。https://doi.org/10.1145/2858036.2858469[120] Krishna Subramanian、Johannes Maas 和 Jan Borchers。2020年。TRACTUS：理解并支持假设驱动的数据科学中的源代码实验。计算机协会，美国纽约州纽约市，1–12。https://doi.org/ 10 1145 3313831 3376764 . / . Hariharan Subramonyam、Colleen Seifert、Priti Shah 和 Eytan Adar。2020年。TexSketch：通过笔墨注释进行主动绘图。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，美国纽约州纽约市， 1 13 10 1145 3313831 3376155 –。https://doi.org/。/。Nicole Sultanum、Christine Murad 和 Daniel Wigdor。理解并支持借助LitSense开展的学术文献综述工作流程。载于《国际高级可视化界面会议论文集》（意大利萨莱诺）（AVI '20 67, 5 10 1145 3399715）。计算机协会，美国纽约州纽约市，文章页码。https://doi.org/。/。 3399850 Rohail Syed、Kevyn Collins-Thompson、Paul N. Bennett、Mengqiu Teng、Shane Williams、Dr. Wendy W. Tay 和 Shamsi Iqbal。2020年。通过视线追踪和自动问题生成改善学习成果。载于《2020年万维网大会论文集》（中国台北）（WWW '20）。计算机协会，美国纽约州纽约市，1693–1703。https: //doi.org/ 10 1145 3366423 3380240 . / . 编辑团队。Distill停刊。Distill（2021）。https://doi.org/10.23915/distill.00031 https://distill.pub/2021/distill-hiatus。[125] Kelsey Urgo 和 Jaime Arguello。2022年。理解迈向搜索者学习目标的"路径"。《ACM信息系统汇刊》40，4，文章编号77（2022年1月），42 10 1145 3495222 页。https://doi.org/。/ Pertti Vakkari。搜索即学习：基于文献的系统化。《信息科学杂志》42，1（2016），7–18。 https://doi.org/10.1177/0165551515615833[127] Pertti Vakkari 和 Saila Huuskonen。2012年。搜索投入会降低搜索输出，但能改善任务结果。《美国信息科学与技术学会会刊》63，4（2012），657–670。https://doi.org/10.1002/asi.21683 arXiv:https://onlinelibrary.wiley.com/doi/pdf/ 10 1002 21683 . /asi. Pertti Vakkari、Mikko Pennanen 和 Sami Serola。撰写研究计划书时搜索词和策略的变化：一项纵向案例研究。《信息处理与管理》39，3（2003），445–463。https://doi.org/10. 1016 0306 4573 02 00031 6 /S - ( ) - Richard Van Noorden。全球科学产出每九年翻一番。《自然》新闻博客（2014）。[130] April Yi Wang、Dakuo Wang、Jaimie Drozdal、Xuye Liu、Soya Park、Steve Oney 和 Christopher Brooks。2021年。怎样的笔记本才算文档齐全？Kaggle上的数据科学家文档实践案例研究。计算机协会，美国纽约州纽约市。 https://doi.org/10.1145/3411763. 3451617[131] Yun Wang、Dongyu Liu、Huamin Qu、Qiong Luo 和 Xiaojuan Ma。2016年。文献综述导览：通过叙事可视化促进学术论文阅读。载于《第9届国际视觉信息交流与交互研讨会论文集》（美国德克萨斯州达拉斯）（VINCI '16）。计算机协会，美国纽约州纽约市，17–24。https: //doi.org/ 10 1145 2968220 2968242 。/。Jevin D. West 和 Carl T. Bergstrom。2021年。科学之中及关于科学的错误信息。美国国家科学院院刊118，15（2021），e1912444117。https://doi.org/ 10 1073 1912444117 。/ pnas。Ryen W White 和 Resa A Roth。。探索式搜索：超越查询—响应范式。信息概念、检索与服务综合讲座 1, 1 2009 1 98 10 2200 00174 1 01 200901 003 （），–。 https://doi.org/。/S ED V Y ICR Teena Willoughby、S. Alexandria Anderson、Eileen Wood、Julie Mueller 和 Craig Ross。2009年。在学习情境中快速搜索互联网信息：领域知识的影响。计算机与教育52，3 2009 640 648 10 1016 2008 11 009 （），–。 https://doi.org/。/j.compedu。。[135] Longqi Yang、David Holtz、Sonia Jaffe、Siddharth Suri、Shilpi Sinha、Jeffrey Weston、Connor Joyce、Neha Shah、Kevin Sherman、Brent Hecht 等。2022年。远程工作对信息工作者之间协作的影响。《自然·人类行为》 6, 1 2022 43 54 10 1038 41562 021 01196 4 （），–。 https://doi.org/。/s - - - Amy X. Zhang 和 Justin Cranshaw。。通过协作式标注和摘要生成理解群聊。*ACM*人与计算机交互论文集 2, 196 2018 27 10 1145 3274465 CSCW，文章（11月），页码。https://doi.org/。/ Amy X. Zhang、Michael Muller 和 Dakuo Wang。。数据科学工作者如何协作？角色、工作流程与工具。*ACM*人与计算机交互论文集4，CSCW1，文章22（2020年5月），23页。https://doi.org/10.1145/ 3392826[138] Huiwen Zhang、Dana McKay 和 George Buchanan。2021年。我所有的读者都与我同在：一种将阅读视为社交活动的模型。载于*2021*年人类信息交互与检索会议论文集（澳大利亚首都领地堪培拉）（*CHIIR '21*）。计算机协会，美国纽约州纽约市， 185 195 10 1145 3406522 3446022 –。https://doi.org/。/。Xiaoyu Zhang、Senthil Chandrasegaran 和 Kwan-Liu Ma。2021年。ConceptScope：基于领域本体组织和可视化文档中的知识。载于*2021*年*CHI*计算机系统中的人为因素会议论文集（日本横滨）（*CHI '21*）。计算机协会，美国纽约州纽约市，文章19，13页。 https://doi.org/10.1145/3411764。3445396[140] Xiaolong Zhang、Yan Qu、C. Lee Giles 和 Piyou Song。2008年。CiteSense：支持研究文献的意义建构。载于*SIGCHI*计算机系统中的人为因素会议论文集（意大利佛罗伦萨）（*CHI '08*）。计算机协会，美国纽约州纽约市，677–680。https://doi.org/ 10 1145 1357054 1357161 。/。Yongfeng Zhang、Xu Chen 等。。可解释推荐：综述与新视角。*Foundations and Trends®* 信息检索14，1（ 2020 1 101 10 1561 1500000066 ），–。https://doi.org/。/ Sacha Zyto、David Karger、Mark Ackerman 和 Sanjoy Mahajan。2012年。社交文档注释系统的成功课堂部署。载于*SIGCHI*计算机系统中的人为因素会议论文集（美国得克萨斯州奥斯汀）（*CHI '12*）。计算机协会，美国纽约州纽约市， 1883 1892 10 1145 2207676 2208326 –。https://doi.org/。/。 Sacha Zyto、David Karger、Mark Ackerman 和 Sanjoy Mahajan。2012年。社交文档注释系统的成功课堂部署。载于*SIGCHI*计算机系统中的人为因素会议论文集（美国得克萨斯州奥斯汀）（*CHI '12*）。计算机协会，美国纽约州纽约市，1883–1892。 https://doi.org/10.1145/2207676.2208326

A 研究材料

这里展示了本研究使用的招募表、半结构化访谈中使用的问题，以及出声思维中使用的提示。此外，我们在表1中报告了参与者的人口统计信息。

A.1 参与者招募

以下内容显示在网络表单中，用于收集参与者招募所需的信息：

(1) 您的姓名。回答：自由文本(2) 您的电子邮件地址。这将有助于我们联系您，以安排研究访谈。回答：自由文本 (3) 您是否认同自己是从事"数据科学"的实践者？数据科学实践者的范围可能非常广泛，涵盖数据工作、应用工程工作，以及由接受过计算机科学和统计学训练的人员开展的研究，也包括经济学、生物学等具体应用领域学科的相关研究。回答：是，否

(4) 请列出您在过去一年中阅读过且觉得有趣或在其他方面有用的 3 篇研究论文。请对“阅读”作广义理解——这可以包括完整阅读论文、略读，或通过博客或推文阅读摘要。回答：自由文本

A.2 阶段 1：半结构化访谈问题

在半结构化访谈中，我们提出了以下问题，并在适当时提出后续问题（如下所列子问题），并在提出下一个问题前向参与者总结其回答。

（1）请简单介绍一下你作为数据科学家的工作角色。

- 例如：您从事哪些类型的问题研究？您日常开展的工作是什么性质的？(2) 您进行文献综述的目标是什么？(3) 您在进行文献综述时会使用哪些辅助工具？例如搜索引擎、文献管理器、笔记应用等。(4) 在进行文献综述的过程中，您是否会与他人交流（线上或当面）？(5) 请谈谈您曾经觉得进行文献综述令人沮丧或枯燥乏味的一次经历。

(6) 您能否说说自己通过哪些不同途径接触到研究文献？时间允许时询问

- 你认为这些不同方法各自有哪些优点和取舍？（7）在结束这一阶段之前，对于与文献的互动，你还有什么想补充分享的吗？

A.3 阶段2：任务场景下的出声思维提示

在这一阶段，参与者进行了文献综述，并共享屏幕以便接受观察。

参与者阅读了以下提示，并被要求开始进行文献综述。

> “回想一次你过去做过的文献综述。想象你要重新开始这一过程，请向我们展示你是如何进行文献综述的。”“想象你想要查找并记录某个自己感兴趣主题的最新研究成果，请向我们展示你如何完成这一过程。”“想象你正计划围绕一个自己感兴趣的问题开展未来研究，请进行文献综述，以帮助你规划未来的工作。”