arXiv:2302.07187v1 [cs.HC] 14 Feb 2023

使用 ISHMAP 框架开发火星毅力号探测车异常检测界面的经验

Austin P. Wright Peter Nemere Adrian Galvin 佐治亚理工学院 昆士兰科技大学 喷气推进实验室 亚特兰大，美国 布里斯班，澳大利亚 帕萨迪纳，美国

Duen Horng Chau 佐治亚理工学院 亚特兰大，美国

Scott Davidoff 喷气推进实验室 帕萨迪纳，美国

摘要

尽管异常检测是众多科学领域中最重要且最有价值的问题之一，但异常检测研究往往聚焦于人工智能方法，而这些方法可能缺乏开展科学探究所必需的细致性与可解释性。我们认为，这种对采用固定框架的算法的单一关注，最终阻碍了科学家在许多科学应用场景中采用即使准确率很高的异常检测模型。在本文中，我们介绍了一种替代方法的应用结果：将基于机器学习的异常检测的数学框架置于参与式设计框架之中。我们与美国国家航空航天局的科学家开展合作，他们使用PIXL仪器研究火星行星地球化学，并将其作为寻找地外生命的一部分；在此过程中，我们进行了为期18个月以上的情境化用户研究与协同设计，以确定美国国家航空航天局科学家在检测和解释光谱异常时所面临的关键问题。针对这些问题，我们为PIXL科学家开发了一套新颖的光谱异常检测工具包。该工具包具有很高的准确率（检测衍射异常的测试准确率达到93.4%），同时保持了对科学解释的高度透明性。我们还介绍了该算法及其配套界面为期一年的实地部署成果。如今，它已作为PIXL科学团队工作流程的核心组成部分，每日投入使用，并直接使该算法成为有关火星潜在宜居性发现的关键贡献者。最后，我们介绍了一种在此次合作过程中开发的新设计框架，用于共同创建异常检测算法：异常现象的迭代语义启发式建模（ISHMAP）。该框架为科学家和研究人员提供了一套流程，以生成原生可解释的异常检测模型。本文展示了一个在科学领域内成功衔接人工智能与人机交互方法的案例，并通过ISHMAP提供了一项资源，供其他研究人员和从业者使用

允许出于个人或课堂用途免费制作本作品部分或全部内容的数字副本或纸质副本，前提是不得出于盈利或商业目的复制或分发这些副本，且副本首页须注明本声明及完整引文。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。*IUI '23，2023*年*3*月*27*日至*31*日，澳大利亚新南威尔士州悉尼
© 2023年版权归所有者/作者所有。ACM ISBN 979-8-4007-0106-1/23/03。
https://doi.org/10.1145/3581641.3584036

希望与其他科学团队合作，通过更有效且可解释的异常检测工具推动科学取得更大进展。

CCS 概念

• **应用计算** → **物理科学与工程**；•**计算方法** → **异常检测**；• **以人为中心的计算** → 开源软件；交互系统与工具。

关键词

异常检测、以人为中心的机器学习、科学计算、X 射线光谱学

ACM 引用格式：
Austin Wright、Peter Nemere、Adrian Galvin、Duen Horng Chau 和 Scott Davidoff。2023年。使用 ISHMAP 框架开发火星毅力号探测车异常检测界面的经验教训。载于第*28*届国际智能用户界面会议（*IUI'23*），*2023*年*3*月*27*日至*31*日，澳大利亚新南威尔士州悉尼。ACM，纽约，美国，15页。https://doi.org/10.1145/3581641.3584036

1 引言

检测和解释异常是科学过程的核心任务之一[24]。然而，尽管新型机器学习模型正迅速提升异常检测领域的技术水平[34]，，科学应用仍为异常检测带来了有趣而复杂的问题。由于科学的目标往往不只是检测异常，还要理解其成因，因此，开发兼具高准确率和可解释性的异常检测模型，成为支持科学探究的关键途径之一。尽管已有研究探讨如何让算法解释其推理过程[18, 26, 35]，，本文研究如何将人机交互方法深度融入模型开发的框架之中，推动将可解释性视为用户定义的质量，并将其作为首要目标，而非事后计算的解释。

在这项工作中，我们展示了这样一种设计流程的实际应用：我们与美国国家航空航天局的一支科学家团队进行了多年合作，他们分析来自PIXL仪器的数据，以了解火星地球化学，进而探究地外生命存在的可能性 [3, 7, 9, 13]。我们的合作围绕以下研究问题展开：

（1）**研究问题1：** 在PIXL科学家的特定科学工作流程中，建模方法必须满足哪些具体要求？（2）**研究问题2：** 现有的、基于机器学习的异常检测标准方法在多大程度上满足或违背了这些要求？（3）**研究问题3：** 如何构建一种不同的异常检测建模框架，以便根据这些要求开发更有效的系统？

在探讨这些问题的过程中，我们研究了科学情境下异常检测的更广泛方法，并报告了以下四项贡献：

(1) **形成性设计研究**。我们介绍一项为期18个月的研究成果。在该研究中，我们通过一系列情境访谈，梳理了美国国家航空航天局 PIXL 科学团队的分析工作流程，发现科学家在检测和解读 X 射线荧光（XRF）数据中的光谱异常时面临的关键挑战，并全面建立了 PIXL科学家开展异常检测的模型。这项研究揭示了三个关键设计目标，用以指导开发辅助这一工作流程的工具。

(2) **新颖的光谱学异常检测算法**。我们介绍了一种新方法，能够直接在原始未量化光谱中自动检测并准确分类衍射及其他光谱异常（测试准确率为93.4%），较现有方法有显著改进。

(3) **部署的算法与可视化系统** 。我们将该算法嵌入了一款可视化工具中，该工具已投入使用，如今已成为过去一年所有基于 PIXL 开展的分析中常规且重要的组成部分。全球 97 名以上的美国国家航空航天局科学家及其合作科学家每天都在使用该工具。我们还通过考察该工具所促成的一些新颖行星科学研究实例，进一步评估了这款应用的成功之处。

(4) **我们提出了一种名为 ISHMAP 的新型框架，用于像 PIXL 这样的科学团队协作开发异常检测工具**。最后，我们介绍了异常现象迭代语义启发式建模（ISHMAP）框架。该框架源于这款应用的成功，可作为一种实用工具，支持未来在类似科学异常检测环境中的协作。该框架融合了人机交互与人工智能在模型开发方面的观点，并以一种不同的方式表述异常检测问题，专门满足科学用户的需求。ISHMAP 提出了一套流程，用于构建默认即可提供一流科学解释的异常检测模型，以确保科学用户认可这些模型，同时还能与现有建模技术（包括深度学习方法）紧密集成。

本文深入探讨这一具体应用，展示了一个将人工智能异常检测研究与人机交互领域开发的方法相结合的可行范例。我们相信，通过跨学科结合不同方法，研究人员能够与科学界合作，更好地应对异常检测等优先级较高的问题，并助力推动科学发现。

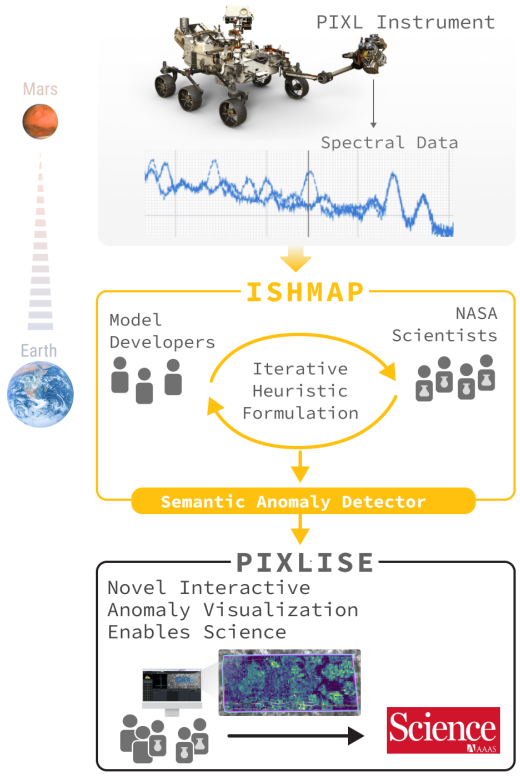

图 1：我们如何运用 ISHMAP 协助 PIXL科学任务的概览。借助这一协作流程，我们开发出一种新颖且可解释的异常检测模型，并将交互式可视化部署到广泛使用的 PIXLISE 可视分析程序中。此次部署为正在进行的重大科学发现提供了关键见解。

我们以 NASA PIXL 科学团队为例，展示这种衔接如何为我们自身的探究提供支持。在下一节中，我们将把这项工作置于人机交互、人工智能和天体生物学研究的更广阔背景下。

2 背景与相关工作

2.1 地外生命的探索

寻找地外生命是当代重大的科学事业之一 [9]，，而火星是这项探索的关键组成部分 [13]。世界各地的科学家了解火星是否可能孕育生命，一个主要途径是长期研究其行星尺度地质学和地球化学 [3, 7, 27]。美国国家航空航天局的 PIXL 仪器通过在一次实验中采集相互配准的可见光成像，以及数千组成对的空间定位*X*射线荧光 （XRF）光谱，为这项持续进行的研究提供支持 [2]。

PIXL仪器在地外XRF测量的规模和灵敏度方面实现了跨代飞跃[15, 36,

42]。虽然这能带来分析能力的飞跃，但也意味着数据对光谱异常更加敏感，可能导致科学家误读数据。

XRF光谱中的异常历来由人工识别。但每个实验地点都包含数千条需要人工检查的光谱，每条光谱又有数百个峰，因此异常越来越难以人工识别，也越来越耗时。漏检光谱异常可能导致对所研究样品的元素化学、矿物学乃至作用于这些样品的行星尺度环境产生错误解读。因此，科学界对可靠且可解释地检测PIXL传入数据中的异常抱有浓厚兴趣。

2.2 融合人机交互与人工智能方法，实现可解释建模

尽管起源更加独立，HCI-AI[22], HCML [17], 和人类引导的机器学习[16] 等混合学科的近期发展，体现了人们希望借鉴不同领域的知识，共同指导涉及各学科的系统开发。比如，人机交互研究者致力于理解设计符合用户需求的人工智能系统所面临的挑战[47],，并探索如何将这些系统呈现的新特性用作设计资源[5]。处理某种形式的嵌入式智能所带来的复杂性，促使其他研究者引入特定方法，以规范普适计算[10, 33] 或对话系统[46] 中人工智能和机器学习系统的构思与迭代。

另外，人工智能和机器学习领域的研究人员借助专家知识来指导机器学习模型 [8],，或将交互性引入学习系统 [12],，并尝试利用交互来定义机器学习模型的规则 [19]。这些方法已应用于网络安全、时空分析和行为建模情境中的异常检测[23]。

在科学领域，这种交叉研究探讨了交互性和玻璃盒模型，旨在支持可解释且可配置的已部署机器学习模型 [43]。然而，海洋学等学科的实地研究表明，尽管解读和理解代码与模型至关重要，但这不足以促成对更广泛科学工作流程的贡献，因为推动变革的主要动力往往来自异常的“变化时刻”，这些时刻自然而然地促使人们重新构想，而这种重新构想并不适用于任何传统数据科学或机器学习模型所隐含的固定数据视角；正如文中所言：“一味专注于解决问题，可能会错失创新机会，而这些机会本可以推动社群参与，进而带来发展势头并促进采纳”[25]。因此，本研究旨在拓展参与式设计实践在科学人工智能方法中的应用传统，通过采用更灵活的建模方法，同时保留可解释性的既有关键要素。

3 形成性研究

本节介绍我们对科学家分析 PIXL 数据工作流程的探究结果，重点关注异常。我们通过一系列情境探究访谈[44] 开展了这项研究，访谈大约每两周进行一次，持续了 18 个月，以了解不同的用户

PIXL 数据。虽然我们与数十位使用 PIXL 数据开展工作的科学家进行了交流与合作，但主要关注五位用户：三位光谱学家（下文称为 R1、R2 和 R3）、一位沉积学家 R4，以及一位地球化学家 R5。

本研究开始时提出的研究问题是：在这一情境下，任何建模干预要想对科学家的分析工作流程有用，必须满足哪些主要约束（研究问题1）？通过这些访谈，我们确定了三项主要设计约束，阐明了任何异常检测方法要对PIXL科学家有用所需满足的要求。虽然每项设计目标都牢牢立足于PIXL分析的情境，但其背后的依据以及为各项目标提供参考的PIXL数据特征并非PIXL独有，很可能适用于广泛的科学应用。

3.1 PIXL 科学工作流程背景

为了理解PIXL异常检测设计目标所处的背景，我们必须先介绍科学家所处理的数据格式和处理步骤的一些基本信息。在实验期间，PIXL会对一种专门指定的样品进行操作，这种样品称为目标。PIXL会将X 射线束照射到岩石表面的某个位置。在每个位置，PIXL的X 射线束会使目标中的化学元素产生荧光，并将其记录为一种名为光谱 [4], 的数据类型，即一个包含4096个通道的数组。每个通道记录在特定能级上检测到的电子计数，该能级以千电子伏特为单位。PIXL的A 和B探测器分别从两个不同的相位角记录目标上各点发出的独特荧光模式。这些独特响应呈现为荧光峰，即宽度固定、横跨若干通道的高斯峰，叠加在背景测量值之上；背景测量值取决于 X 射线束照射位置的化学组成。实验期间，PIXL的相机还会拍摄目标的可见光图像。数据传回地球后，研究人员会重建 X 射线束几何结构，并将每个光谱点定位到返回图像中的相应位置。总体而言，对一个目标进行实验会返回一系列可见光图像以及约4,000个光谱点，每个光谱点均包含A和B光谱，以及在每张图像中的x、y坐标 [2] （见图2）。

在火星上使用PIXL仪器进行实验后，该实验的光谱和图像数据会传回地球，并主要通过PIXLISE分析用户界面 [39]进行分析。领域专家会通过多个步骤对这些数据进行分析和转换。在分析的第一步，光谱学家会将光谱中的峰转换为他们认为存在于目标中的元素 [20]。

随后，光谱学家使用元素列表来定量分析光谱，将各个峰的强度转换为经验估算值，以估计每种元素在目标 [20]中占总元素质量的百分比。PIXL使用PIQUANT定量算法 [11, 21], ，并通过PIXLISE用户界面提供这些功能。PIQUANT会独立定量分析每条光谱，而光谱学家则使用数据集中所有光谱的总和来确定要定量分析的元素集合。

将光谱量化为元素重量百分比后，更广泛的科学团队便开始分析数据集。由于

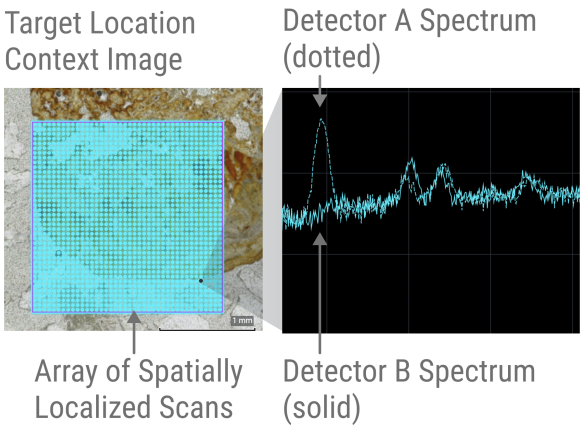

图 2：PIXL仪器所提供数据概览

纯元素在自然界中并不常见，科学团队接下来的任务是确定已检测并量化的元素如何组合，形成目前已知的某种或某些5700+矿物 [37]。这项分析主要依据数据集的元素重量百分比图，并运用许多标准的地质学与地球化学数据可视化技术进行呈现，例如三元图 和热图；同时也结合 PIXLISE 独有的可视化方式，例如弦图。此外，还会结合图像中岩石的颜色、形状和纹理等信号，也就是其形态，以辅助判断候选矿物。

随着候选矿物逐渐浮现，科学家会将其指定为感兴趣区域（ROI），并逐步构建理论，解释使岩石及其各个ROI得以形成并随时间发生变化的历史地质过程 [3]。随着讨论范围扩大、纳入能够汇总地球化学、地球物理学和气候细节的天体生物学家，这些理论也越来越依赖背景信息和视觉信息，以及矿物的分布情况，以此构建关于目标、地点和区域整体背景的长期理论，并探讨其对生物宜居性的意义。

3.2 设计目标 1：关注原始数据，而非处理后和定量分析后的数据（*G1*）

在考虑这一工作流程中的异常检测问题时，我们首先试图弄清楚应分析哪种数据结构：原始光谱还是定量分析结果。我们观察到，对许多科学家而言，PIXL仪器光谱所提供的信息几乎完全是通过荧光现象的元素定量分析，以及在PIXLISE中对这些定量结果的可视化来呈现的。这意味着，在大多数情况下，异常是通过发现定量分析中的意外结果来识别的，随后再回溯查找导致定量结果出错的非荧光光谱现象。正如R2所述：

> “……我在氧化钡图中看到相当多的活动，这么多是正常的吗……？ ……所以我可能会……查看那些单独的光谱。”

这种方法有明显的缺点：任何导致定量分析出错、但又不够异常而不值得深入审查的异常，都可能传播误导性信息。此外，后续纠正所花费的时间和精力总量可能远远超过尽早检测，尤其考虑到 PIXL 生成的数据规模。因此，我们提出了第一个设计约束：希望能够**对原始未定量光谱进行分析**，以便发现可能被批量求和定量分析所掩盖的现象。

3.3 设计目标 2：在真实标签有限的情况下保持稳健性（*G2*）

我们发现的下一个限制是，目前PIXL数据集中实际存在的可靠真实标签数量非常有限。这是因为，要得出可靠结论，需要少数专家用户进行人工处理，而PIXL生成的数据规模又前所未有。R1表达了他们希望在深入分析之前，先筛除潜在异常的愿望：

> “在自动化阶段，我们希望……把每天一个包含*5000*条光谱的数据集……缩减到每周只需查看几条光谱。因为……调查一个数据集需要多日 *[for]*，而且……还有所有其他科学产出……所以……之后你会希望把它标记出来，以便稍后查看。”

因此，我们将第二个限制确定为：我们的方法必须**能够在真实标签数量较少的情况下保持稳健**，从而提供数量适当的标记，供专家进行精确分析。

3.4 设计目标 3：允许按科学因果过程区分异常（*G3*）

在最初的访谈中，我们反复听到另一项主要限制，受访者明确强调了理解为何某些异常会出现在科学模型中至关重要。例如，R1 解释了为何目前需要人工查看光谱：光谱的建模方式种类繁多，因此需要具备背景知识：

> “荧光数据需要人工拟合是有原因的……光谱中可能会有……*15* 条稀土元素谱线。总会有某位专家用户了解样品，并据此拟合数据……由于组合方式无穷无尽，这件事……无法自动完成。”

此外，异常现象本身可能包含有关其成因所涉及的物理过程的有用信息，因此值得单独分析。我们发现，手动调查定量异常时最常发现的非荧光现象是衍射。衍射通常会在专门设计的X射线衍射仪器中作为独立现象进行研究 [6]，，但其信号响应有时会与荧光峰重叠。因此，我们假设，如果能充分区分这类光谱异常，就可以将其作为另一种辅助信息来源。

这与矿物鉴定过程中用于观察背景信息的背景成像类似。目前缺少的是一种能够在流程早期发现并表征异常，进而利用这些检测结果改进下游分析的方法。R1也以类似的方式阐述了这一目标：

> “最终的检验是：如果荧光产生了一种含义模糊的矿物，而衍射能够使其变得明确。”

因此，我们认为，对于科学家而言，能够发现异常还不够，**还必须在帮助他们理解潜在科学原理的背景下，解释并区分不同类型的异常**。

3.5 与现有方法的比较：使用设计目标评估基于标准机器学习的异常检测方法

在理解了该领域及相关设计目标后，我们尝试评估异常检测问题的标准方法，并考虑其中可能存在的问题（研究问题2），同时为解决方案中需要处理的具体问题提供指导（研究问题3）。这些用于异常检测和机器学习的标准方法大致可分为传统方法和基于深度学习的方法。传统方法通常针对表格数据设计，因此一般不适合分析设计目标 G1 所要求的原始数据；而深度学习模型则以适应复杂的非表格数据格式而闻名。基于深度学习的异常检测方法通常遵循这样一种通用结构：训练模型将数据编码为紧凑表示，以发现结构模式和离群点 [34]。这些无需大量标注（设计目标 G2 所要求）的主要方法类别是特征提取方法和正常性表示方法[34]，二者都包含与我们的设计目标相冲突的明确假设。特征提取方法假设：“深度学习模型提取的特征表示保留了有助于将异常与正常实例区分开的判别信息。”这与我们在目标 G2 中的发现相悖：测量成本高昂，因此没有足够的数据来构建一个全面的特征集，其中涵盖罕见而微妙的异常类别，且这些数据没有明确标签。正常性表示模型采用潜在空间编码器的形式，例如自编码器或生成对抗网络，并假设与异常实例相比，这些编码模型能够更好地表示和重构非异常实例。我们发现，在实践中情况并非如此。考虑到目标 G2 的限制，我们无法预先区分正常实例和异常实例以进行半监督学习，这意味着异常实例必须包含在训练数据中。这可能造成问题，因为模型随后会学会像表示正常实例一样良好地表示这些实例。考虑到重构损失函数的选择“是为降维或数据压缩而设计，而非为异常检测而设计。因此，所得表示只是对潜在规律的通用概括，并未针对检测不规则性进行优化”，这一点也就更容易理解。结合 G2 缺乏监督标签这一情况，该特性从根本上违背了目标 G3，因为通用模式明确不会优先关注真正罕见或类别迥异的异常，而总是优先关注单个点异常，或正常数据类别中碰巧更罕见的简单子样本。

我们通过科学解释实现异常区分的目标 G3，清楚地表明科学领域的最终用户最关注的是潜在的因果过程，而不是最清晰的表层经验模式。在缺乏稳健标注的情况下，无监督方法最终只能以各自不同的方式优化表层经验相似性，而无法考虑不同的因果过程。因此，即使这类方法能够发现部分现有异常，它们仍从根本上将筛选重要异常类别与不重要异常类别的任务保留为人工流程。我们希望能够颠倒这一操作顺序：先让科学家区分他们所关心的异常类型，然后直接检测这些异常，从而在处理新数据时获得预先分类的解释。

4 方法

考量深度学习和传统数据科学式异常检测的弱点时，问题的一个根本原因在于，问题框定要么完全没有科学家的直接参与（这不可避免地违反设计目标G3），要么只能通过标注来获取科学家的意见（为了传达所需的细微差别，这就需要采用违反设计目标G2的量表）。因此，我们着手开发一种替代的模型开发框架，以便更有效、更高效地将科学家的先验知识融入异常检测模型（研究问题3）。

开发这一框架的一个关键洞见，是区分现象 与数据。大多数现有方法纯粹聚焦于数据空间，因此必须将异常定义为单个数据点。然而，科学家所采用的本体比数据空间更抽象；物理世界中存在许多潜在过程，而这些过程可以通过许多不同或不完整的方式进行测量。重要的并非某个单一数据项，而是任何数据子集或超集能够揭示潜在现象的什么信息。因此，我们所说的现象可以出现在多个尺度层级上。一个高维或高度复杂的数据点本身可能包含多个不同现象的实例。

这种分析形式体现了权衡的一端。数据空间分析追求完整性，通过对数据集中样品所包含的所有可测量信息进行编码，直至达到数据集规模的限制，而不考虑正确性，也不关注任何给定数据项为何具有特定的一组特征。对现象建模必然会限制完整性，因为只有明确纳入考虑的现象才能被建模，而这必然少于在掌握完备知识的情况下可能建模的无数潜在因素。然而，我们由此获得了正确性：感兴趣的现象可以按照更贴合其自然尺度的方式进行建模，并且默认彼此分离以供解读，而不是事后分析数据空间的潜在编码。

本文提出了一种设计框架，异常现象的迭代语义启发式建模（ISHMAP，见图 3），该框架

可为开发人员与科学家协作开展基于现象的异常分析提供范本。该框架以先前提出的设计目标为指导，可根据科学确定的、有意义的异常现象，生成用于提取原始数据特征的启发式模型，并根据科学家有限的可用时间进行自适应迭代。我们在 PIXL 科学研究中运用了这一框架，本文将讨论其在 PIXL 科学研究中的具体应用方式，说明它如何助力取得新的科学发现；同时也将阐述其一般原则，为其他遵循类似设计目标的协作项目提供参考。

4.1 科学家对异常现象的描述

ISHMAP 的入口，以及现象分析与以数据为中心的分析之间的关键区别，在于首先由科学家明确识别出某个特定的异常现象（图 3A）。这一过程始于在给定科学领域的本体中概述异常现象的一个语义类别。科学家选择这一特定类别，是因为它传达了某些重要信息。对于 PIXL，我们从衍射现象入手。科学家选择从这一现象开始，是因为衍射只会在一组特定且理论上已有充分理解的条件下发生，而这些条件涉及晶体结构。这意味着，衍射的存在能够向科学家提供有关样品物理结构的重要信息，而仅凭元素组成无法区分这些信息。

确定某一现象后，必须确定的第一个特征，是现有数据测量该现象所处的尺度。该现象是否表现为某种特定类型的数据点？是否表现为相邻数据点之间的一种模式？或者，它是否可能在单个数据点内作为一个子集出现多次？分析尺度取决于对物理现象的既有理解，以及生成可用数据的测量方法的特性。确定这一尺度对于启用可正确解释的启发式方法至关重要，因为它决定了启发式函数的输入，从而决定了该函数如何将某种现象在更大数据集中的表现形式表达为最自然的解释。确定尺度时，必须制定一个**采样过程，以便从原始数据集中提取目标现象的所有潜在实例**，并建立一个映射回原始数据集的对应关系，以确定具体数据点中哪些部分被纳入样品。在我们的示例中，科学家知道衍射会以光谱中的离散峰形式出现；根据对 PIXL 仪器分辨率特性的了解，这些峰响应被假定为探测器在大小为 0.2 千电子伏的窗口内产生的离散信号，而 0.2 千电子伏正是可检测高斯峰的半峰全宽尺寸 [1]。这意味着，启发式方法的采样输入将是所有宽度为 0.2 千电子伏的连续窗口，并可将每个窗口单独评估为潜在的衍射峰。

确定正确的分析尺度后，科学家应描述该现象相对于数据测量的差异性因果过程。具体而言，就是描述给定现象如何与测量过程相互作用，以及这种相互作用如何基于已知或假设的因果关系，使所描述的异常数据与非异常数据产生差异。这种描述不必极其详尽，因为之后还会经过各种有损处理；只需说明该现象在哪些主要方面偏离模型的默认假设，以及这些差异如何体现在数据中。与必须完全基于原始数据集进行描述相比，恰当选择分析尺度通常能大幅降低描述的复杂性。这种描述将构成设计和反复改进启发式方法的基本起点。对于衍射，其因果过程已得到充分理解：当 PIXL 仪器与样品中的晶体结构恰好对准，且 PIXL 发出的 X 射线频率与晶格共振时，就会发生衍射效应；响应会以相长干涉的方式散射，并在该共振频率处形成响应峰。这些响应峰的形状类似于荧光产生的高斯峰，其宽度由探测器分辨率决定。然而，两者的因果过程不同：衍射峰可出现在任意频率，而荧光峰只能出现在元素的荧光频率处；此外，该效应的空间依赖性足够敏感，因此 PIXL 的两个探测器之一中出现的衍射响应，极少会在另一探测器的相同频率处出现。科学家完成对某一现象的尺度及其产生差异性数据测量的因果前向过程的描述后，开发人员即可进入 ISHMAP 的下一步（图 3B）。

4.2 转换为启发式模型

科学家给出现象的定义后，开发人员的任务便是将该定义转化为可计算的启发式模型（图3B）。根据所提供描述的性质，框架的这一阶段可以有多种不同形式。唯一的要求是，每轮开发最终都要得到一个程序：该程序以符合现象尺度表征所规定形式的数据样品作为输入，并输出一个标量值，用以表示给定样品与异常表征的正向过程的符合程度。该程序的形式取决于多种因素，包括正向过程描述的格式、现象尺度上可用数据的数量和格式，以及可用的计算资源。在为衍射设计初始启发式方法时，我们选择手动实现对上一步所提出假设的统计检验。我们定义了一个函数：给定宽度为 0.2 千电子伏的光谱计数窗口，该函数检验以下假设：PIXL 的两个探测器中，一个探测器是否存在高于光谱噪声阈值且具有统计显著性的响应峰，而另一个探测器中则不存在。具体做法是对窗口中每个通道的两个探测器进行配对差值 t 检验 [40]。之所以采用这种方法，是因为各通道中的计数并非相互独立，每个通道的底层计数取决于该通道的 X 射线频率。我们计算并返回该窗口的 t 统计量的绝对值（因为衍射由哪个探测器检测到并不重要），以此衡量潜在衍射的统计效应。

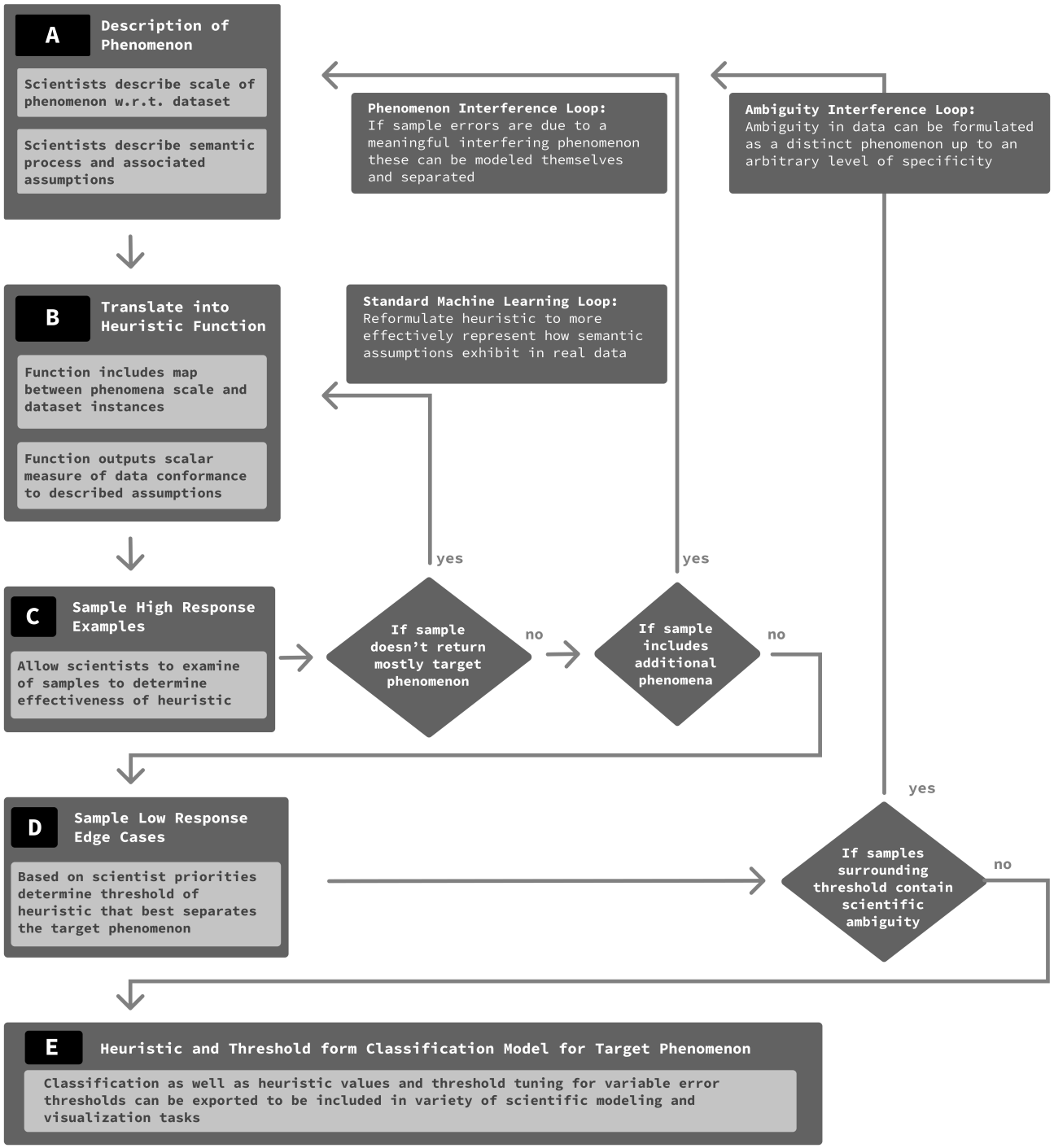

图3：ISHMAP设计框架概览。从科学家主导的现象描述（A）开始，由开发人员将其转化为可计算的启发式函数（B），进而对典型数据实例进行采样（C）；反复迭代，直至启发式方法能够为高响应样本提供清晰信号，随后对低响应样本进行采样以确定分类阈值（D）。确定阈值后，即可得到目标现象的最终模型（E）。图4展示了这一流程所生成的模型示例。

在 PIXL 案例中，启发式模型采用了基于科学家所提供假设的假设检验形式，但构建启发式方法还有许多可能的途径。如果给定的科学描述更容易用贝叶斯先验来表达，那么采用此类概率表述的模型（包括基于深度学习的模型）可能会有所帮助。或者，如果针对特定异常提供了直接模拟技术，那么基于正向过程模拟进行直接相似性比较可能更合适。ISHMAP 框架中的启发式模型并非要为所有现象强制采用单一建模范式，而是要为不同模型提供接口，使其能够在更大的框架中集成协同工作。重要的是，所选模型必须具有最合适的现象尺度输入、可比较的标量启发式输出，并能基于科学家先验进行参数化，同时支持微调。满足所有这些要求后，即可针对数据集中的每个现象尺度数据项计算启发式值，并在下一步中使用这些数据与启发式值组成的数据对（图 3C）。

4.3 基于高响应原型采样的启发式模型评估

在获得启发式模型的一个版本后，ISHMAP 的下一步是评估该启发式，并据此确定模型改进所需的迭代类型（图 3C）。首先，对启发式响应分布高端的数据进行采样。如果启发式模型表现良好，这些高响应样本应构成所要建模现象的鲜明原型。在此阶段，应让科学家有机会检查模型原型类别，并判断该类别的连贯性。如果高响应样本始终被认定为所需现象的良好示例，那么就可以确认该启发式模型，并进入下一个阈值调优阶段。否则，高响应样本集合中就会包含目标之外的其他现象实例。

在这种情况下，科学家必须确定还包含了哪些内容。如果该集合中包含数量可观的不同异常现象实例，那么我们便可以递归地重复 ISHMAP 流程，对这一其他类别进行建模，从而实现区分。评估最初的衍射启发式时，恰恰发生了这种情况。在此阶段，R2 指出：

> “嗯，好吧，是的，我认为这其实是需要指出的一点，因为算法将其识别为衍射。但这并不是衍射。实际上，我会说，我们看到的是强度失配，这种情况不只出现在那个峰上，也出现在光谱中的其他位置、其他峰上。我认为，这种强度失配有一部分与测量粗糙表面有关，比如测量这些较大晶粒时就会出现。”

我们发现了另一类异常现象，这类现象由表面粗糙度导致。随后，我们使用 ISHMAP 对这一现象进行建模。科学家的背景知识告诉我们，表面粗糙度效应与频率无关，因此粗糙度现象空间涵盖了整个光谱，并且这种效应预计会使单个探测器中的信号受到恒定衰减。基于这一背景知识，启发式粗糙度检测器通过计算某个光谱在整个频率范围内的探测器差值均值来进行检测，这本质上是对假定恒定衰减因子的最大似然估计。这一启发式方法在第一轮迭代中有效地区分了粗糙度效应。完成更深层次的 ISHMAP 迭代，并获得有效的粗糙度检测启发式方法后，我们便能在高响应区域中将衍射与粗糙度效应区分开来，从而利用原有的衍射启发式方法，筛选出一类连贯的衍射实例。请注意，粗糙度的现象尺度与衍射的现象尺度并不相同。这说明，两种不同现象的效应可能在一个尺度上相互重叠，但在另一个尺度上却很容易区分。ISHMAP 的递归迭代和以现象为中心的结构，确保所有启发式方法都在现象的原生尺度上对其进行建模，同时仍能提供跨尺度的信息。

在区分高响应样本集中的所有不同异常类别后，可以直接迭代主要启发式模型，以利用非异常正常数据及其相关背景噪声优化区分效果。该迭代循环可根据科学家可提供的标注量灵活调整：基线建模假设能够确保初始启发式模型具有一定的最低语义连贯性，因此额外优化是可选的；同时，启发式建模接口的灵活性也支持在获得额外反馈时采用能够从中受益的模型。因此，预期可获得的下游科学家标注情况，是选择启发式模型类别时需要考虑的重要约束。在经过充分迭代、确保启发式模型中的高响应样本始终代表目标现象的典型实例后，ISHMAP 的这一阶段即告完成，我们便可以进入下一阶段（图 3D ）。

4.4 从采样中等响应边界案例进行启发式阈值调优

确定启发式模型能够对正确特征作出响应后，还需要确定启发式响应值的阈值，以便进行分类。这个过程与前一阶段非常相似：生成启发式值接近潜在阈值的样品，然后由科学家进行评估。如果这些样品仍然高度一致，就可以降低阈值；如果一组样品中正确异常的数量不足，则可以提高阈值。这个过程可以反复进行并加以调整，具体取决于科学家对假阳性和假阴性的相对重视程度。此阶段（图 3D）仍保留了高响应阶段（图 3C）中的迭代机会：如果在边界区域发现更多现象，就可以递归地对其建模，从而使边缘案例区域更加清晰。此外，在这一阶段，我们还会考虑一类特殊现象，称为歧义样本。在现象检测的决策边界区域，科学家经常会发现某些方面存在模糊性的实例，使得难以或无法为其指定底层标签。这些模糊实例不应与启发式值恰好接近决策边界的定义明确的样品相混淆；后者的

启发式值恰好接近决策边界。相反，歧义样本指的是科学真实情况本身也不十分明确的示例。这些现象给任何分类方案都带来了挑战。在 ISHMAP 中，我们通过将模糊性建模为一种独立现象来处理；即使科学先验按定义而言不够稳健，科学家用来解释不同解读的特征仍可作为启发式建模的依据。这些特征必然存在，否则科学家便没有任何经验依据来产生不确定性。这个额外的迭代阶段会持续进行，直到不再发现歧义实例，此时便可根据误差容限精确确定阈值；或者，由于歧义类别可能包含数量无限、各不相同且具有个体特征的可建模特征，其迭代过程会受到科学家可用时间的限制。

在对衍射边缘案例进行采样时，科学家发现了两种不同的模糊现象。第一种是单个探测器响应的案例：可能发生衍射的窗口中，探测器之间存在显著差异，但这种差异并未像先前预期的衍射响应那样呈现明显的峰，而是相对平坦。随后通过测量窗口中心处探测器差异峰的相对高度，对峰值明显和不明显的情况进行区分，并据此建立模型。第二种模糊现象是：非衍射探测器的背景并不恒定，基线探测器（窗口内平均计数较低的探测器）也出现了一个小峰，但其幅度低于另一个探测器。这导致不同科学家作出了不同解读，因此通过考察窗口内基线探测器的变异系数，将其作为模糊案例建模。区分这些主要的模糊现象后，衍射启发式阈值的边界区域变得足够清晰，可以据此确定用于判定衍射的阈值；在本例中，阈值的设置倾向于减少假阳性判定。

一旦所有干扰现象均已建模并加以区分，且确定了分类阈值，ISHMAP 的最终输出便是一个可部署的分类模型（图3 E）。该模型具有与图4所示决策树相同的基本分类结构，因为上述现象区分由 ISHMAP 返回的分类器通过类似的递归迭代实现。

5 评估

为了评估我们的方法，我们既针对研究中的模型组件进行了标准的定量误差分析（第 5.1 节），也对最终部署系统的成功情况进行了定性评估（第 5.2 节）。这样做并不是要证明，作为一种设计框架，ISHMAP 确实优于其他设计方法（因为这需要掌握许多不同设计框架在同一问题上的成功情况等反事实信息，而这显然远超本研究的范围）。我们只是希望展示 ISHMAP 能够在**这**一应用中产生出色的科学成果，因此可能有助于其他应用。

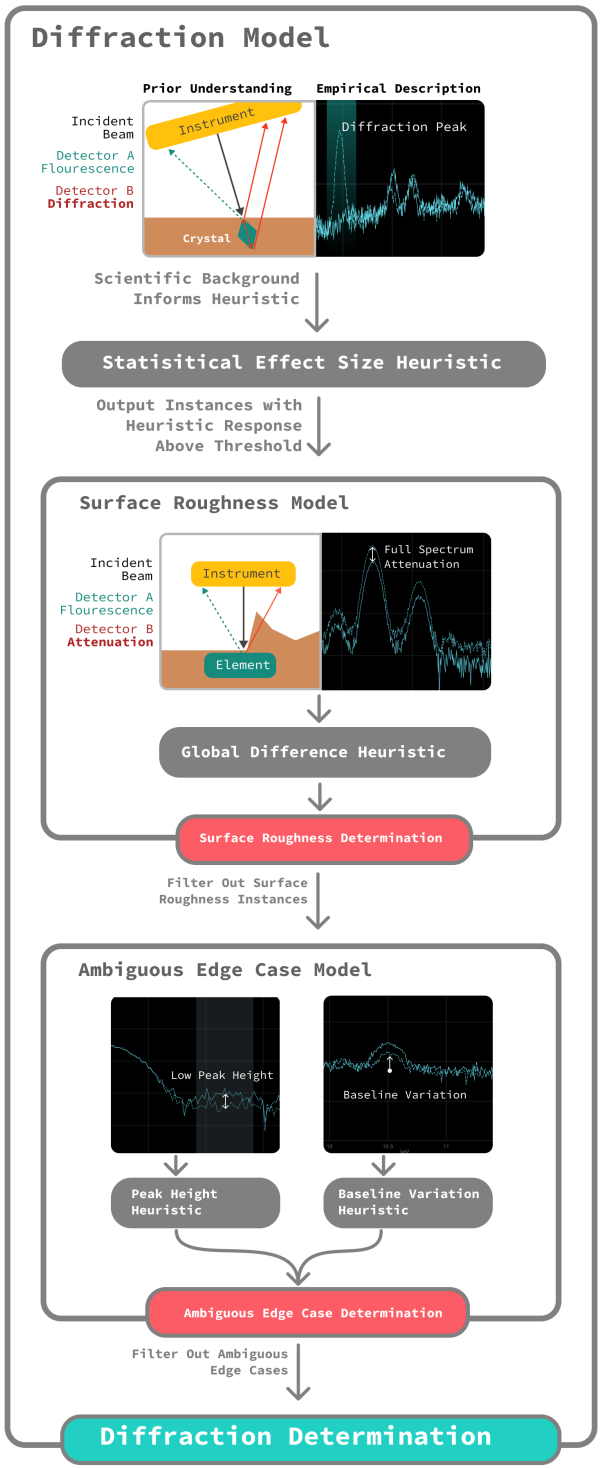

图4：利用 ISHMAP 检测衍射峰的输出模型（图3E）架构。

5.1 异常检测模型评估

完成针对衍射这一目标现象的 ISHMAP 流程后，我们得到一个分类模型（图3E / 图4）。为了在矿物鉴定的实际应用背景下评估该模型的准确率，同时避免信息泄漏并检验模型的泛化能力，我们使用 R2 和 R3 提供的输入开发衍射模型后，又使用来自其他数据集、由其他科学家（R1、R4 和 R5）提供的标签测试模型。我们邀请了多位科学家，以确保标签可靠。我们向三位科学家提供了 213 条具有代表性的随机抽样光谱，这些光谱未用于训练过程。样品经过均衡处理，其中 107 条光谱是在模型判定含有衍射峰的光谱中均匀随机抽取的，另有 106 条则是在模型判定不含衍射的光谱中均匀抽取的（这些光谱表现为表面粗糙度、模糊性或无异常）。随后，科学家需要判断所展示的光谱中是否有任何一条含有衍射。尽管三位科学家都看到了全部 213 条光谱，但由于时间有限，有些人花在判断上的时间明显更少，只对所展示的光谱标注了少数几条。此外，我们还发现，各位科学家在不确定情况下判定存在衍射的敏感度各不相同，因此他们的判断经常不一致。因此，对于 16 条光谱，我们无法形成可靠的真实标签：这些光谱仅有两位意见不一致的科学家给出标签；另有 45 条光谱仅有一位科学家给出标签（我们发现，缺少第二意见时，这种标签并不可靠）。最终，我们得到了 152 条可用于评估的光谱（其中 144 条的判断完全一致，8 条的判断以多数意见为准）。在这组可靠的真实标签中，模型对是否存在衍射的预测准确率为 **93.4%**。

这些结果与科学家在检查模型输出时表达的定性体验相符。在一次对 R2 的访谈中，他们表示：

> “你的工具运行得很好，在几乎所有情况下都能找到 *[衍射峰]*……因此，实际上并没有发现任何错误的结果。”

从科学家领域专家最终用户的角度来看，这种定性可靠性和满意度构成了对该工具效能最切题的评估。与那些没有提供任何系统性替代基线的先前方法相比，这也充分展示了 ISHMAP 框架在为科学用户设计实用模型方面的成效。

5.2 已部署界面的影响

ISHMAP 协作式设计流程的一项重要优势是，一旦完成了用于检测特定异常现象的模型，就可以确保该模型解决的是科学家实际关注的问题，而参与协作的科学家也会对这一技术产生一定的主人翁意识和认同感[38]。这意味着，采用此类模型的工具在实际部署后，更有可能被纳入科学工作流程。为展示这种协作式设计和优先考虑科学解释的建模方法，不仅能够在通用分类基准方面表现出色，还能为科学领域的最终用户带来有意义的新能力，我们可以考察将我们的 ISHMAP 衍射模型部署到 PIXLscience 工作流中的情况。模型定稿后，我们将模型输出集成到了 PIXL 科学所使用的现有主要视觉分析工具包 PIXLISE 中 [14, 31, 48]。

模型及相关可视化的首个完整版本于2021年11月部署，科学家早在2021年6月就已能使用初步版本。截至撰写本文时，全球有97位以上的美国国家航空航天局科学家及其合作科学家每天使用该工具，共同参与PIXL科学任务。在PIXL科学小组的讨论板上，该工具的功能在工作组讨论中被提及80多次，其中39次指出，衍射支持的地球科学解释揭示了原本不易察觉的信息，另有28次提及对光谱学结果的纠正和定量误差检测。如今，每当有新数据从火星传回，我们的模型便能帮助科学家即时发现可用于后续矿物鉴定的衍射峰。

5.3 衍射面板界面

数据集一旦载入PIXLISE，科学家就可以立即使用衍射面板（图5A）识别特定的衍射峰。衍射面板包含一幅直方图，显示检测到衍射峰的所有能级。科学家可以在直方图中选择一组能量范围，进而筛选出所有包含这些范围内衍射峰的位置（图5A.1）。这使科学家能够快速发现不同类型的衍射峰以及相应矿物的位置（因为衍射能量直接取决于对应矿物的晶体结构），如图6所示。这一功能对于科学解释尤为重要，因为我们的衍射模型采用了ISHMAP规定的正确衍射尺度框架，而非基于机器学习的模型所使用的异常光谱默认框架。随后，科学家可以通过可排序的检测到的衍射峰列表进一步核验各个峰的识别结果（图5A.2）。科学家可以选择一个峰，然后在PIXLISE光谱视图图表上查看其对应的光谱和能量位置（图5A.3），并在界面中确认该分类是否正确判定为衍射，还是假阳性。这一步不可或缺，因为即使非常准确的模型也会出错；允许用户进一步完善可用数据，不仅能让模型持续更新和改进，也有助于建立科学家对模型的信任，确保他们始终能够推翻模型的解释。

5.4 衍射图可视化

除了衍射面板支持的针对特定峰的工作流程外，我们还通过衍射图 界面实现了对异常结构更高层次的可视化（图 5B）。随着 PIXL 不断传回新数据，衍射图已成为 PIXL科学团队进行矿物鉴定时易于获取、便于共享且不可或缺的信息。

衍射图可视化会在视觉背景图像上叠加一张热图，显示每个

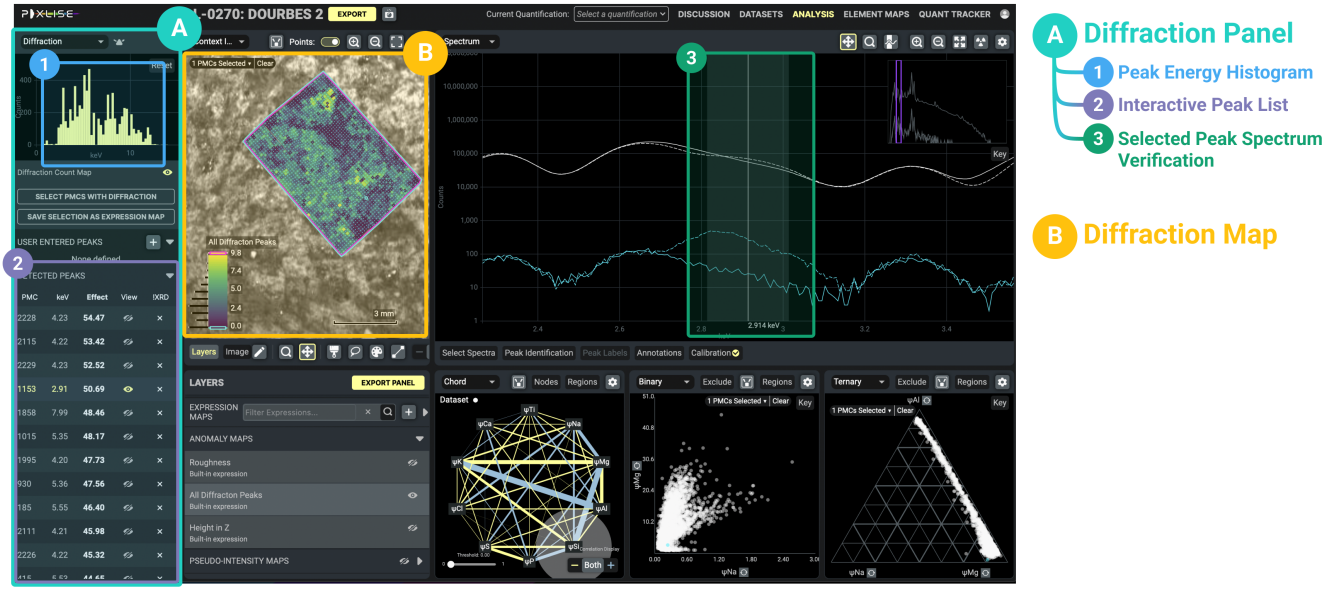

图 5：PIXLISE 应用程序界面组件概览，展示了 Guillaumes 火星数据集。（A）衍射面板可快速识别并验证单个衍射峰或归为一组的相似峰。（B）衍射图显示衍射峰在空间中的分布，可覆盖整个光谱范围、特定能量子集，也可结合其他 PIXLISE 分析工具中用户自定义的表达式进行显示。

光束位置处衍射峰或表面粗糙度异常的密度。这使科学家能够快速找到表明存在晶粒的衍射簇，将该簇中的位置分组，并创建可用于进一步分析的感兴趣区域。衍射图已成为科学家分享晶体结构发现的首选方式，正如 R3 查看 Beaujeu [28] 数据集的衍射图时所说：

> “我发现，标记有大量衍射峰的区域与我们通过地球化学分析识别为斜长石的区域之间存在有趣的相关性……看到衍射峰检测算法在实际中的用途，真是太棒了。”

这些地图可以通过多种方式进行自定义。默认视图会显示所有能级上的衍射密度；当我们对目标的具体晶体学特征缺乏特别有力的先验认知时，这种视图最为通用。不过，如果科学家对某种特定晶体构型有假设，且该构型会在可预测的频率上产生衍射，则可以通过两种方式更具体地可视化衍射。PIXLISE 内置一种自定义领域专用语言，可用于创建表达式地图。衍射模型的输出是该语言支持的查询，可让科学家将异常信息与其他现有分析结合起来。此外，衍射面板的直方图选择功能支持直接创建地图。通过查看衍射峰能量的分布，科学家无需先验知识即可分析该分布并找出簇群，从而创建地图；这类地图不是显示样品的整体晶体学结构，而是能够分离出颗粒晶粒（图 6）。

5.5 推动火星地质学领域持续取得科学发现

2021年11月，PIXL 在火星耶泽罗陨石坑底部的 Séítah 组地层 [30]，对代号为 **Dourbes** [29] 的样品进行了一系列 XRF 扫描。该地点给矿物鉴定带来了一个棘手的问题。由于风化作用，无法通过背景成像清晰识别晶粒，而 XRF 信息也不足以区分所有相关的物理属性。要推断该地点的地质历史，这些信息至关重要；在类似情形下，这也一直是以往火星任务面临的一项重大挑战。

所幸，得益于 PIXL 仪器采集的额外数据以及我们开发的一套工具，科学家能够迅速通过衍射图直观查看样品的晶体结构（图 6）。衍射图功能支持将衍射与元素分析进行稳健的空间比较，使科学家能够有力地论证特定元素颗粒及其晶体性质，从而识别矿物学特征。通过超越标准荧光和元素定量分析所提供的信息，这种比较为 Séítah 地层岩石的矿物学特征提供了决定性证据。一位 PIXL 科学家在一篇近期摘要中对此作了阐述，该摘要将发表于顶级科学期刊[41]:

> “共位的晶体大小和矿物种类对于解读岩石中的纹理关系及检验地质假设至关重要，但以往一直无法明确确定这些性质……本文展示了

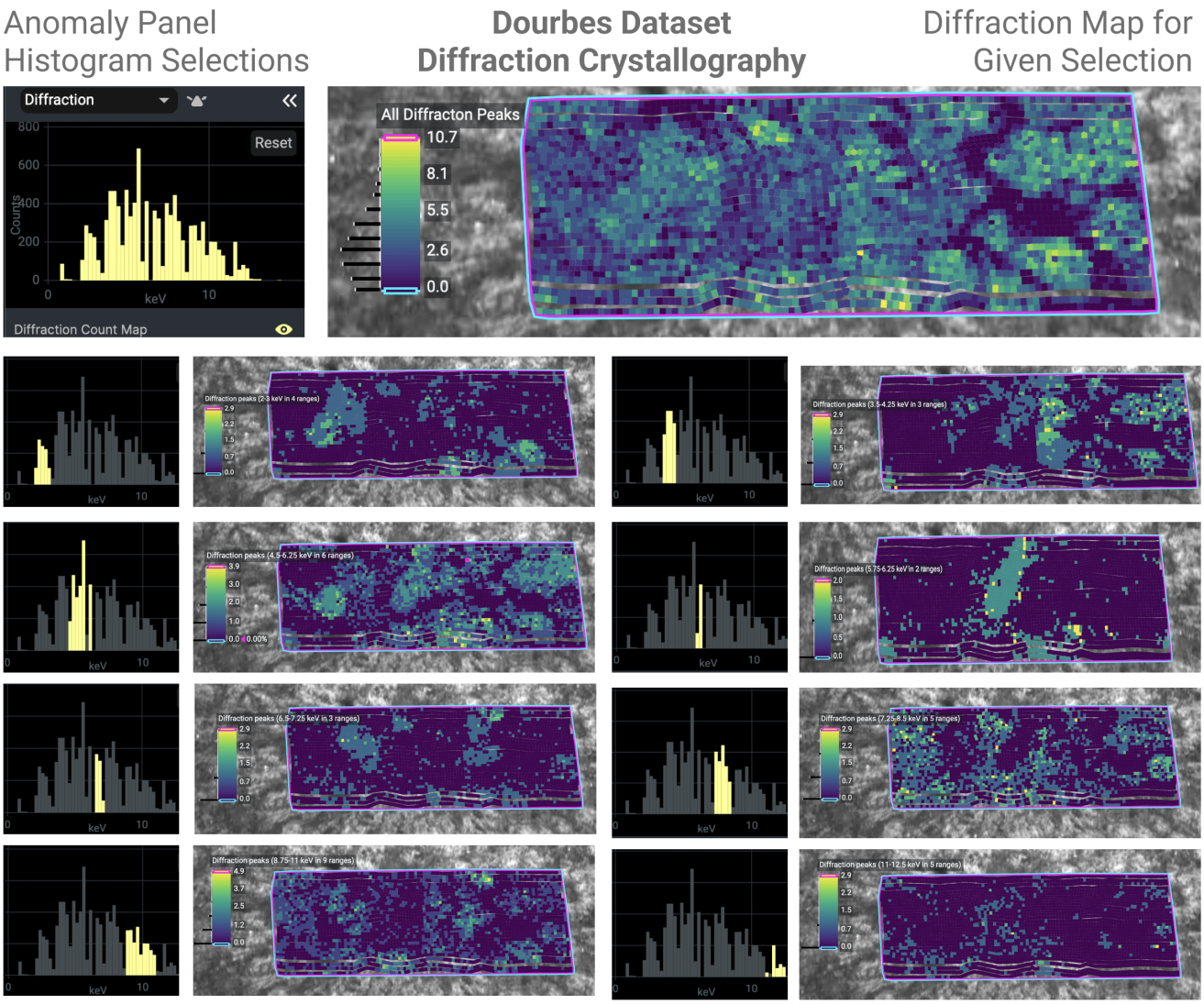

图 6：从衍射峰能量衍射面板中选择不同选项后，Dourbes[29] 数据集生成的衍射图截图。可以看到，其中包含有关不同能级下衍射峰空间分布的丰富信息；能量簇接近的衍射峰在空间上也会聚集。这表明存在独特的晶粒，利用这些衍射图便能轻松观察到它们。

PIXL 仪器检测到的衍射和荧光 *X* 射线……能够提供有关各种矿物中是否存在相干晶体域的信息。”

这一发现已成为PIXL科学团队持续研究的核心内容，而我们的模型所支持的衍射检测与可视化功能则使其得以实现。该模型能够与现有科学工作流程有效集成，并在部署后立即辅助开展高影响力分析，这进一步有力地证明了ISHMAP设计框架的有效性。

6 讨论

6.1 设计目标的可推广性

虽然这项工作展示了一个具体且成功的应用案例，但我们希望，我们的解决方案中有许多有益的见解也能够应用于其他应用场景。为了评估我们的框架对其他使用案例的适用性，关键的决定因素应当是我们所概述的设计目标之间的一致性。虽然我们是在特定的 PIXL 科学家领域背景下，严格通过嵌入式用户研究制定这些目标的，但我们根据所发现的科学家工作流程中的相关方面，对每个目标的形成进行了论证，而这些方面

在相当多元化的 PIXL 团队中不同个体及专业领域之间具有共性，同时，尽管这些方面体现在这一特定工作流程中，却不一定为该工作流程所独有。例如，我们的设计目标 G1 基于这样一个观察：数据集距离异常现象的“原生”尺度越远，异常就越容易被遗漏。由于根据定义，异常本质上是不适用默认假设的现象类别，因此，任何处理步骤都可能在异常检测方面引入错误。由此我们可以说，关注原始数据不仅是 PIXL 的一项重要设计考量，也是任何包含基于可违背假设、以不可逆方式处理数据的步骤的分析工作流程的重要考量。我们的设计目标 G2 实质上将数据和标签可用性有限这一极为常见的设计约束形式化。无监督机器学习的整个分支都在研究这种限制对模型构建的各种影响。因此，尽管发现这一约束适用于我们的特定科学领域极其重要，但显然还有许多其他领域也有类似需求。最后，我们的设计目标 G3 同样表达了基于深度学习的异常检测中一个已得到充分确立且广为人知的弱点[34]，并说明了它为何在我们的特定使用案例中十分重要。所有这些意味着，尽管我们制定的目标从一个层面上与 PIXL 科学领域紧密相关，但我们预计，很可能还有大量其他领域也具有这些目标，而在这些领域中，ISHMAP 可能会成为构建模型开发框架的有用工具。

6.2 ISHMAP 的局限性

虽然我们展示了 ISHMAP 框架具有显著的潜在效用，但与所有框架一样，它并非普遍适用，在使用范围上也存在局限。当然，ISHMAP 是否适用，主要取决于所列出的设计目标是否相关。在优先考虑这些目标时，其他潜在优先事项便会被弱化。尤其是，ISHMAP 流程只能帮助发现已知的异常类别，以及在 ISHMAP 流程中发现的异常类别。在许多领域（包括科学领域），数据源足够新颖，可能包含许多潜在的“未知的未知”异常类别，而用户事先并未预料到这些类别。由于 ISHMAP 依赖于对已知现象的明确描述，因此无法可靠地捕捉此类未知异常。研究表明，在这类应用中，更纯粹的基于深度学习的方法在发现此类独特异常或点异常方面可能非常有效 [34]。

此外，由于要求开发人员与科学家持续迭代并相互反馈，协作流程可能会产生可观的组织开销。根据合作的性质，这种开销有时可能比直接提供更多真实标签所需的人工投入更大，从而削弱设计目标 G2。因此，在决定是否开展 ISHMAP 合作时，双方都必须考虑问题在技术和组织层面的特点，以确定这种合作是否合适。

6.3 未来工作的机遇

在介绍 ISHMAP 框架时，我们只是迈出了利用以人为中心的人工智能方法改进科学异常检测的第一步。尽管我们与 PIXL 团队共同实施的检测工具取得了明确的成功，但它仅仅是该方法的概念验证。我们鼓励研究人员在更多领域复现、评估并完善这一方法。此外，该框架本身也显然有进一步发展的空间。框架的灵活性使其适用于多种不同形式的应用和部署，因此，未来若能进一步优化科学先验编码、启发式生成和样本评估在技术与流程层面最有效的做法，或许将有助于更高效、更有效地使用 ISHMAP。事实上，ISHMAP 的开发初衷是弥补纯深度学习方法的不足；而研究如何在这个更注重可解释性的框架中整合深度学习模型及其强大的表达能力，或许能兼得两种方法的优势。最后，就 ISHMAP 当前的设计而言，科学家的专业知识固然至关重要，但模型设计者／开发人员的职责相对偏重流程执行。这为开发自动化工具或交互式界面提供了潜在机会，使科学家能够完全独立地参与 ISHMAP 流程中由其负责的部分，从而大幅提升各科学团队自行开发稳健、可解释异常检测模型的潜力。

6.4 可复现性

本文所讨论工作的代码分布在更广泛的 PIXLISE 项目中的多个代码仓库中；该项目正在根据开源许可发布其所有组成仓库，地址为 https://github.com/pixlise[32, 45]。此外，已部署的 PIXLISE 工具本身也可公开访问。任何人都可以在 https://www.pixlise.org/ 申请账户，并使用 PIXLISE 的全部功能，包括本文讨论的异常检测功能，对任何公开数据集进行分析。本文用作示例的大多数数据集都可以公开获取：要么位于 PIXLISE 上，要么以原始数据形式直接来自 NASA行星数据系统（PDS），网址为 https://pds-geosciences.wustl.edu/ m2020/urn-nasa-pds-mars 2020_pixl/。

7 结论

历史上许多最重要的科学发现，都得益于对异常的深入分析 [24]。如今，随着科学数据集变得越来越庞大、越来越复杂，用于发现其中异常的方法也日益复杂，有时甚至以牺牲解释和说明异常的能力为代价[34]，而这种能力是科学分析的根本组成部分。在这项工作中，我们尝试将人机交互方法与最先进的人工智能技术相结合，开发一种既有效又可解释的异常检测方法。我们与美国国家航空航天局一支世界领先的科学团队合作，开展了广泛的用户研究，以了解他们特定的分析工作流程，并制定能够体现科学家对可解释异常检测需求的设计目标。基于这些设计目标，我们提出了 ISHMAP，这是一种用于开发科学异常检测模型的新型设计框架。通过利用

ISHMAP 开发了一款异常检测工具包，美国国家航空航天局的科学家遍布全球，每天都在使用它，并借此推动持续进行的科学发现。我们展示了这一概念验证：通过以人为本的方式处理异常检测的技术问题与科学问题，从而实现更好的科学研究。我们希望这种方法能帮助科学家和研究人员不仅检测数据中的异常，也能理解这些异常。

致谢

这项研究的部分工作在美国国家航空航天局合同（80NM0018D0004）项下，于加利福尼亚理工学院喷气推进实验室开展。

参考文献

[1] Abigail Allwood、Ben Clark、David Flannery、Joel Hurowitz、Lawrence Wade、Tim Elam、Marc Foote 和 Emily Knowles。2015。利用 PIXL 对岩石和土壤进行纹理特定元素分析：火星2020上的行星X射线岩石化学仪器。载于 *2015*年 *IEEE* 航空航天会议。1–13。[2] Abigail C Allwood、Lawrence A Wade、Marc C Foote、William Timothy Elam、Joel A Hurowitz、Steven Battel、Douglas E Dawson、Robert W Denise、Eric M Ek、Martin S Gilbert、M.E. King、C.C. Liebe、T. Parker、D.A.K. Pedersen、D.P. Randall、R.F. Sharrow、M.E. Sondheim、G. Allen、K. Arnett、M.H. Au、C. Basset、M. Benn、J.C. Bousman、R.J. Calvet、L. Cinquini、B. Clark、S. Conaby、H.A. Conley、S. Davidoff、J. Delaney、T. Denver、E. Diaz、G.B. Doran、J. Ervin、M. Evans、D.O. Flannery、N. Gao、J. Gross、J. Grotzinger、B. Hannah、J.T. Harris、C.M. Harris、C.M. Heirwegh、C. Hernandez、E. Hertzberg、R.P. Hodyss、J.R. Holden、C. Hummel、M.A. Jadusingh、J.L. Jørgensen、J.H. Kawamura、A. Kitiyakara、K. Kozaczek、J.L. Lambert、P.R. Lawson、Y. Liu、K.M. Macneal、McLennan. S.、P. McNally、P.L. Meras、J. Napoli、B.J. Naylor、P. Nemere、N. Pootrakul、R.A. Romero、R. Rosas、J. Sachs、M.E. Schein、T.P. Setterfield、V. Singh、E. Song、M.M. Soria、N.R. Tallarida、D.R. Thompson、M.M. Tice、L. Timmermann、V. Torossian、A. Treiman、S. Tsai、K. Uckert、J. Villalvazo、M. Wang、D.W. Wilson、S.C. Worel、P. Zamani、M. Zappe 和 R. Zimmerman。2020。PIXL：行星X射线岩石化学仪器。《空间科学评论》

216, 8 2020 1 132 10 1007 11214 020 00767 7 ( ), –。[3] https://doi.org/。/s - - - Jillian F. Banfield、John W. Moreau、Clara S. Chan、Susan A. Welch 和 Brenda Little。2001。矿物学生物标志与火星生命搜索。《天体生物学》1，4（2001），447–465。https://doi.org/10.1089/153110701753593856[4] Burkhard Beckhoff、Birgit Kanngießer、Norbert Langhoff、Reiner Wedell 和 Helmut Wolff。2007。*X*射线荧光分析实用手册。施普林格科学与商业媒体。[5] Jesse Josua Benjamin、Arne Berger、Nick Merrill 和 James Pierce。2021。机器学习不确定性作为设计材料：一项后现象学探究。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，论文171，14页。https://doi.org/10.1145/3411764.3445481[6] David L Bish、DF Blake、DT Vaniman、SJ Chipera、RV Morris、DW Ming、AH Treiman、P Sarrazin、SM Morrison、RT Downs 等。2013。火星科学实验室的X射线衍射结果：盖尔陨石坑Rocknest的矿物学。《科学》341，6153（2013），1238932。[7] Janice L Bishop、Enver Murad、Melissa D Lane 和 Rocco L Mancinelli。2004。火星矿物鉴定的多种技术：研究热液岩石作为火星天体生物学地点的潜在类似物。《伊卡洛斯》169，2（2004），311–323。[8] Mu-Huan Chung、Mark Chignell、Lu Wang、Alexandra Jovicic 和 Abhay Raman。2020。用于数据外泄检测的交互式机器学习：结合人类专业知识的主动学习。载于*2020*年 *IEEE*系统、人类与控制论国际会议（*SMC*）。280–287。https://doi.org/10.1109/SMC42975.2020. 9282831[9] Christopher F. Chyba 和 Kevin P. Hand。2005。天体生物学：生命宇宙研究。《天文学与天体物理学年评》43，1（2005），31–74。[10] Scott Davidoff、Min Kyung Lee、Anind K. Dey 和 John Zimmerman。2007。通过快速约会快速探索应用程序设计。载于第九届普适计算国际会议论文集（奥地利因斯布鲁克）*(UbiComp '07)*。施普林格出版社，柏林、海德堡，429–446。[11] W. T. Elam、B. D. Ravel 和 J.R. Sieber。2002。用于X射线光谱计算的新型原子数据库。《辐射物理与化学》63，2（2002），121–128。[12] Jerry Alan Fails 和 Dan R. Olsen。2003。交互式机器学习。载于第八届国际智能用户界面会议论文集（美国佛罗里达州迈阿密）*(IUI '03)*。计算机协会，美国纽约州纽约市，39–45。https://doi.org/10.1145/604045.604056

[13] Alberto G Fairén, Alfonso F Davila, Darlene Lim, Nathan Bramall, Rosalba Bonaccorsi, Jhony Zavaleta, Esther R Uceda, Carol Stoker, Jacek Wierzchos, James M Dohm 等。2010年。《火星各个时代的天体生物学：通过研究地球类比环境来理解火星宜居性》。《天体生物学》10卷，8期（2010年），821–843页。[14] David Flannery, Scott Davidoff, Michael M. Tice, Abigail C. Allwood, William Timothy Elam, Christopher M. Heirwegh, Joel A. Hurowitz, Yang Liu 和 Peter Nemere。2021年。借助行星X射线岩石化学仪器（PIXL）的新型数据分析软件，提高火星2020探测车任务的运行效率。《2021年空间研究委员会（COSPAR）科学大会论文集》43（2021年）。B0.2期。[15] R Gellert, R Rieder, J Brückner, BC Clark, G Dreibus, G Klingelhöfer, G Lugmair, DW Ming, H Wänke, A Yen 等。2006年。α粒子X射线光谱仪（APXS）：古谢夫陨石坑的探测结果与校准报告。《地球物理研究杂志：行星》111卷，E2期（2006年）。[16] Yolanda Gil, James Honaker, Shikhar Gupta, Yibo Ma, Vito D'Orazio, Daniel Garijo, Shruti Gadewar, Qifan Yang 和 Neda Jahanshad。2019年。迈向由人类引导的机器学习。载于《第24届国际智能用户界面会议论文集》，614–624页。[17] Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, Nicolas d'Alessandro, Joëlle Tilmanne, Todd Kulesza 和 Baptiste Caramiaux。2016年。以人为中心的机器学习。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，3558–3565页。https://doi.org/10.1145/2851581.2856492[18] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf 和 Guang-Zhong Yang。2019年。XAI——可解释人工智能。《科学·机器人学》4卷，37期（2019年），eaay7120。[19] Lijie Guo, Elizabeth M Daly, Oznur Alkan, Massimiliano Mattetti, Owen Cornec 和 Bart Knijnenburg。2022年。通过用户贡献的可解释规则，在交互式机器学习中建立信任。载于《第27届国际智能用户界面会议》，537–548页。[20] Michael Haschke。2014年。实验室微区X射线荧光光谱法。查姆：施普林格国际出版公司10（2014年），978–983页。[21] C. M. Heirwegh, Elam W. T. 和 L. P. O'Neil。2022年。聚焦光束X射线荧光元素定量软件包PIQUANT。《光谱化学学报B辑：原子光谱学》196（2022年），106520。https://doi.org/10.1016/j.sab.2022.106520[22] Kori Inkpen, Stevie Chancellor, Munmun De Choudhury, Michael Veale 和 Eric P. S. Baumer。2019年。人在哪里？弥合人工智能与人机交互之间的鸿沟。载于《2019年CHI计算机系统中的人为因素会议扩展摘要集》（苏格兰格拉斯哥）（CHI EA '19）。计算机协会，美国纽约，1–9页。https://doi.org/10.1145/3290607.3299002[23] Liu Jiang, Shixia Liu 和 Changjian Chen。2019年。交互式机器学习的最新研究进展。《可视化杂志》22卷，2期（2019年），401–417页。[24] Thomas S Kuhn。1970年。《科学革命的结构》。第111卷。芝加哥：芝加哥大学出版社。[25] Kateryna Kuksenok。2016年。采用之外的影响：编程与科学实践之间的互动如何塑造四个海洋学团队的探究模式。博士学位论文。[26] Scott M Lundberg 和 Su-In Lee。2017年。一种解释模型预测的统一方法。《神经信息处理系统进展》30（2017年）。[27] S. McMahon, T. Bosak, J. P. Grotzinger, R. E. Milliken, R. E. Summons, M. Daye, S. A. Newman, A. Fraeman, K. H. Williford 和 D. E. G. Briggs。2018年。《在火星上寻找化石的野外指南》。《地球物理研究杂志：行星》123卷，5期（2018年），1012–1040页。https://doi.org/10.1029/2017JE005478 arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/

10 1029 2017 005478 . / JE 美国[28] 2021 *140* 34 国家航空航天局。PIXL Sol。https://bit.ly/Mzpl 6

[29] 2021 美国国家航空航天局。PIXL拍摄的Dourbes景象。https://www.jpl.nasa.gov/images/pia25041-pixls-view-of-dourbes [30] 美国国家航空航天局。2021年。塞塔岩石的两个视角。https://www.jpl.nasa.gov/images/pia25023-two-perspectives-of-seitah-rocks [31] Peter Nemere, Tom Barber, Adrian Galvin, Ryan Stonebraker, S. Michael Fedell, Austin P. Wright, David O. Flannery, Michael M. Tice, Yang Liu, W. Timothy Elam, Christopher M. Heirwegh, Abigail A. Allwood, Joel A. Hurowitz, Morgan Cable 和 Scott Davidoff。2021年。PIXLISE应用程序。https://www.pixlise.org/ [32] Peter Nemere, Ryan Stonebraker, Adrian Galvin, Tom Barber, S. Michael Fedell 和 Scott Davidoff。2023年。pixlise/pixlise-ui：2.0.16版。https://doi.org/10.5281/zenodo.7539750[33] William Odom, John Zimmerman, Scott Davidoff, Jodi Forlizzi, Anind K. Dey 和 Min Kyung Lee。2012年。通过用户演绎开展面向未来的实地工作。载于《交互系统设计会议论文集》（英国泰恩河畔纽卡斯尔）（DIS '12）。计算机协会，美国纽约，

338 347 10 1145 2317956 2318008 –。https://doi.org/。/。[34] Guansong Pang, Chunhua Shen, Longbing Cao 和 Anton Van Den Hengel。2021年。用于异常检测的深度学习：综述。《ACM计算综述》54卷，

2, 38 2021 38 10 1145 3439950 文章（火星），2016 [35] 页码。https://doi.org/。/。Marco Tulio Ribeiro, Sameer Singh 和 Carlos Guestrin。“为什么我应该相信你？”解释任意分类器的预测。载于第*22*届*ACM SIGKDD*国际知识发现与数据挖掘会议论文集，1135–1144页。[36] R. Rieder, T. Economou, H. Wänke, A. Turkevich, J. Crisp, J. Brückner, G. Dreibus 和 H. Y. McSween。1997年。移动式α粒子质子X射线光谱仪对火星土壤和岩石化学组成的测定：X射线模式的初步结果。《科学》278卷，5344期（1997年），1771–1774页。https://doi.org/10.1126/science.278.5344.1771[37] Steven W. Ruff 和 Jack D. Farmer。2016年。火星上的二氧化硅沉积物，其特征类似于智利El Tatio的温泉生物标志物。《自然通讯》7卷，1期（2016年11月17日），13554。[38] Elizabeth B-N Sanders 和 Pieter Jan Stappers。2008年。共创与设计的新景观。协同设计4卷，1期（2008年），5–18页。https://doi.org/10.1080/15710880701875068[39] David Schurman, Pooja Nair, Scott Davidoff, Adrian Galvin, Abigail Allwood, Yang Liu, David Flannery, Robert P. Hodyss, Santiago V. Lombeyda, Maggie Hendrie, Hillary Mushkin 和 Christopher P. Heirwegh。2019年。PIXELATE：用于分析天体生物学光谱数据的新型可视化与计算方法。《*2019*年天体生物学科学会议论文集》（*AbSciCon*）。[40] Student。1908年。均值的或然误差。*Biometrika*（1908年），1–25页。[41] Michael M Tice, Joel A Hurowitz, Abigail C Allwood, Michael WM Jones, Brendan J Orenstein, Scott Davidoff, Austin P Wright, David AK Pederse n, Jesper Henneke, Nicholas J Tosca 等。2022年。根据火星上的PIXLISE X射线荧光、X射线衍射和多光谱成像结果推断塞塔地层岩石的蚀变历史。*Science Advances* 8卷，47期（2022年），eabp9084。[42] Scott J. VanBommel, Ralf Gellert, Jeff A. Berger, John L. Campbell, Lucy M. Thompson, Kenneth S. Edgett, Marie J. McBride, Michelle E. Minitti, Irina Pradler 和 Nicholas I. Boyd。2016年。通过对火星科学实验室α粒子X射线光谱仪进行过采样，分离不同的岩性化学组成。*X*射线光谱学45卷，3期（2016年），155–161页。[43] Zijie J Wang, Alex Kale, Harsha Nori, Peter Stella, Mark E Nunnally, Duen Horng Chau, Mihaela Vorvoreanu, Jennifer Wortman Vaughan 和 Rich Caruana。2022年。可解释性，然后呢？编辑机器学习模型以体现人类知识与价值观。载于第*28*届*ACM SIGKDD*知识发现与数据挖掘会议论文集，4132–4142页。[44] Dennis Wixon, Karen Holtzblatt 和 Stephen Knox。1990年。情境设计：一种系统设计的新兴观点。载于*CHI*，329–336页。[45] Austin P. Wright, Peter Nemere, Ryan Stonebraker, Adrian Galvin 和 Scott Davidoff。2022年。*pixlise/diffraction-peak-detection*：*2.0*开源迁移版本。https://doi.org/10 5281 6959138 。/zenodo。Qian Yang, Justin Cranshaw, [46] Saleema Amershi, Shamsi T. Iqbal 和 Jaime Teevan。2019年。绘制自然语言处理草图：一项探索如何借助语言智能设计恰当内容的案例研究。载于《*2019*年*CHI*计算机系统中的人为因素会议论文集》（苏格兰格拉斯哥）（*CHI '19*）。计算机协会，美国纽约，1–12页。https://doi.org/10.1145/3290605.3300415[47] Qian Yang, Aaron Steinfeld, Carolyn Rosé 和 John Zimmerman。2020年。重新审视人类与人工智能交互为何、是否以及如何在设计上独具挑战。载于《*2020*年*CHI*计算机系统中的人为因素会议论文集》（美国檀香山）（*CHI '20*）。计算机协会，美国纽约，1–13页。https://doi.org/10.1145/3313831.3376301[48] Constance Ye, Lukas Hermann, Shravya Bhat, Nur Yildirim, Dominik Moritz 和 Scott Davidoff。2021年。PIXLISE-C：探索美国国家航空航天局科学家对矿物鉴定的数据分析需求。载于*ACM*计算机系统中的人为因素会议太空探索人机交互研讨会（美国檀香山）（*SpaceCHI 2021*）。计算机协会，美国纽约，5页。