# 从具有结构化缺失的数据中学习

Robin Mitra[1,2,*], Sarah F. McGough[3, *], Tapabrata Chakraborti[1,4], Chris Holmes[1,5], Ryan Copping[3], Niels Hagenbuch[6], Stefanie Biedermann[7], Jack Noonan[8], Brieuc Lehmann[2], Aditi Shenvi[9], Xuan Vinh Doan[10], David Leslie[1,11], Ginestra Bianconi[1,12], Ruben Sanchez-Garcia[1,13], Alisha Davies[1,14,15], Maxine Mackintosh[1, 16], Eleni-Rosalina Andrinopoulou[17], Anahid Basiri[1, 18], Chris Harbron[19,*], 以及 Ben D. MacArthur[1,13,20,*]

[1]艾伦·图灵研究所，伦敦，英国[2]Statistical Science，伦敦大学学院，伦敦，英国[3]基因泰克，美国加利福尼亚州南旧金山[4]伦敦大学学院医学物理与生物医学工程系及伦敦大学学院癌症研究所，伦敦，英国[5]牛津大学统计学系，牛津，英国[6]F. Hoffmann-La Roche AG，瑞士巴塞尔[7]开放大学数学与统计学院，米尔顿凯恩斯，英国[8]卡迪夫大学数学学院，加的夫，英国[9]华威大学统计学系，考文垂，英国[10]华威商学院，华威大学，考文垂，英国[11]伦敦玛丽女王大学数字环境研究所，伦敦，英国[12]伦敦玛丽女王大学数学科学学院，伦敦，英国[13]南安普顿大学数学科学，南安普顿，英国[14]斯旺西大学健康与生命科学学院，斯旺西，英国[15]威尔士公共卫生署，加的夫，英国[16]英格兰基因组学，道森楼，查特豪斯广场，伦敦，英国[17]伊拉斯谟医学中心生物统计学系与流行病学系，荷兰鹿特丹[18]格拉斯哥大学地理与地球科学学院，英国[19]罗氏制药，韦林花园市，英国[20]南安普顿大学医学院，南安普顿，英国[*]通信请联系：Robin Mitra (ucakrmi@ucl.ac.uk)、Sarah McGough (mcgough.sarah@gene.com)、Chris Harbron (chris.harbron@roche.com) 和 Ben MacArthur (bmacarthur@turing.ac.uk)

arXiv:2304.01429v1 [stat.ML] 4 Apr 2023

## 摘要

缺失数据是许多机器学习任务中不可避免的复杂问题。当数据属于“随机缺失”时，有一系列工具和技术可以应对这一问题。然而，随着机器学习研究变得更加雄心勃勃，并试图从规模不断扩大的异质数据中学习，一种日益常见的问题随之出现：缺失值会明确或隐含地呈现某种关联或结构。这种“结构化缺失”带来了诸多挑战，而这些挑战尚未得到系统性解决，也从根本上阻碍了规模化学习。本文概述了当前的研究文献，并提出一系列关于从具有结构化缺失的数据中学习的重大挑战。

## 1 引言

处理缺失数据是统计学和机器学习（ML）中长期存在的问题（1；2）。针对“随机缺失”的数据，有多种处理方法（3），其依据是成熟的理论（4；5；1；6）。然而，在许多机器学习问题中，数据可能并非随机缺失，而是呈现某种多变量结构或模式（7）。在需要广泛开展数据融合、多视图学习和/或数据关联，并从多个来源和研究中提取信息的领域，这些问题尤为突出（8；9；10）。要推动这些领域的发展，我们需要解决如何整合具有高度“结构化缺失”（SM）的数据并从中学习的问题。

结构化缺失可能由多种原因导致，因此是一个普遍存在的问题。在**框**1 中，我们概述了结构化缺失的一些常见成因，反映了在现代机器学习和建模研究中，结构化缺失自然产生的多种方式。对于大规模机器学习而言，尤为重要的是，在合并多模态数据集时，或数据描述的是由特征各异的个体组成的异质群体的属性时，结构化缺失便会自然产生。例如，在健康领域，关联纵向临床、基因组和影像数据时会出现结构化缺失（11）；在处理范围广泛的任务时也会出现结构化缺失，例如开发泛癌症预测模型，这类模型纳入了针对多种不同癌症的诊断测量，因此相关测量数据仅适用于部分患者（12）。更一般而言，合并多项研究的信息时，常会出现结构化缺失；各项研究的设计和测量集可能各不相同，因此所包含的变量也只是各种测量模态并集中的一个子集。在这些情况下，缺失值可能与用于收集数据的各种抽样方法有关，也可能反映目标人群的特征，因此能够提供有用信息。尽管存在这种可能性，结构化缺失通常会严重妨碍我们使用当前机器学习方法有效利用数据，并可能严重阻碍分析流程的各个方面，包括构建推断模型、设计预测与分类算法，以及生成有信息量的数据可视化摘要。数据缺失生命周期相关的各种挑战概览见**图**。1。

用于处理缺失数据的成熟方法，例如（多重）插补（13），通常不会考虑缺失数据的结构，因此可能无法恰当或有效地处理结构化缺失。随着我们致力于开发能够从海量数据集中规模化学习并执行众多下游任务的机器学习模型，例如基础模型（14），和/或采用联邦学习方法（15；16），这些问题将变得更加突出。事实上，说缺失问题是阻碍规模化学习效率的主要因素之一，并不为过（17）。然而，尽管这一问题普遍存在，结构化缺失尚未得到系统研究；我们既缺乏关于结构化缺失的理论，也缺少能够从存在结构化缺失的数据中高效学习的工具。

相关研究虽有涉猎，但往往零散地分布于不同学科，且通常着眼于改造现有方法，以应对具体的结构化缺失问题。例如，（18）探讨了对$k$近邻方法进行调整，以插补合并多个高通量组学实验数据时产生的块状缺失数据。在类似的生物医学背景下，（19）借助多任务学习工具，插补成块缺失的基因型。从更偏理论的角度来看，（20）探讨了调整多层模型的方法，以便在缺失值系统性地产生时进行多重插补。相关地，（21）提出了一种插补块状缺失值的方法，拓展了经典多重插补方法。更广泛而言，机器学习方法（如基于树的工具和深度学习）在插补缺失值和/或理解缺失结构方面的能力正逐渐受到重视，尽管相关研究仍处于起步阶段（22；23；24）。因此，迄今的研究虽已在特定情境下部分解决了结构化缺失问题，但所提出的方法通常是为特定情况量身定制的；尽管它们在理解结构化缺失方面取得了重要进展，其可推广性仍不明朗。

我们认为，要在不同的应用和部署领域大规模推进机器学习，就需要采用一种更通用的结构化缺失方法，运用计算科学各领域的方法，更好地理解结构化缺失、它对后续学习的影响，以及如何高效地处理结构化缺失；甚至还要探究在什么情况下，结构化缺失的模式能够传递重要信息，因而可能有用。

为此，艾伦·图灵研究所举办了一系列研讨会，汇聚多学科专家，共同识别结构化缺失领域尚待解决的问题，并勾勒出解决这些问题的路线图。本文由社区成员撰写，是这些会议的成果。

本文的组织结构如下。在第 2 节中，我们将介绍 SM，包括回顾理解 SM 所需的一些缺失数据基本概念，描述 SM 的常见来源，并提供相关的 SM 分类体系，以便概念化其中涉及的问题。我们通过实例说明这些入门概念，其中包括一个复杂的临床基因组数据集，它很好地展现了 SM 面临的一些挑战（25）。这些实例凸显了 SM 是多么容易发生，为什么标准方法无法应对 SM 带来的挑战，以及 SM 对后续学习可能产生的影响。在第 3 节中，我们概述了一系列重大的 SM 挑战；若能解决这些挑战，将大幅推动该领域的发展，并使我们更有能力开发新一代机器学习方法，从而高效地大规模学习数据。最后，我们在第 4节作出一些展望未来的总结。

**框1：结构化缺失的成因**

结构性缺失（SM）是一个广义术语，指数据集中呈现非随机多变量模式的缺失，因此可能以多种方式出现，每种方式都有其自身的特征和挑战。尽管如此，在机器学习情境中，仍有一些常见过程会导致 SM。下面我们概述其中五种。

多模态关联：首先，当来自不同来源、流程或实验的数据被关联起来时，自然会出现结构化缺失。这种数据关联正变得越来越普遍，关联后的数据被用于训练一些当今最强大的机器学习模型。在这种情况下，数据中的每个单元可能只测量了部分可用的测量模态，而其他模态则未被观测到。框2 举例说明了这种结构化缺失在大型生物医学数据集中造成的挑战（25；26）。

多尺度关联：其次，当不同空间或时间尺度上的测量结果被合并时，自然会出现结构化缺失。例如，在地理空间分析中，不同传感器可能测量气候的不同方面。其中一些测量值——例如可通过卫星热红外传感器相对容易地获取的海表温度（27）——在时空上的分辨率可能较高；而其他测量值，例如降水，则较难获取，分辨率也可能较低（28）。综合考量时，所得数据会呈现出特有的缺失模式，这些模式与不同传感器技术在精度和复杂性上的差异有关。

批次故障：第三，结构化缺失可能源于数据采集流程中的故障或差异，例如传感器故障、传感器覆盖范围差异，或由批次问题导致的系统性测试故障。这些问题可能是暂时性的，也可能是数据采集流程中不可避免的问题。例如，通过卫星上的遥感设备获取的数据，可能因卫星暂时发生故障而在某些时段缺少观测；也可能因卫星轨道或视场的限制，而在某些地理区域系统性地受到限制（例如所谓的“极区空洞”（29））。

跳题模式：第四，结构化缺失常见于调查数据中，表现为“跳题模式”（30）。只有在对较早问题的回答为“是”时，才会收集一组回答；若回答为“否”，参与者则会跳到调查问卷后面的部分。这种情况下，大块缺失是研究设计中固有且可预期的特征。

人口异质性：第五，考虑人类行为或特征的模式时，也会出现结构化缺失。例如，人们的兴趣和愿望各不相同，活动模式各异，身体属性也有所不同。汇集大量人口（31）的这些行为或属性的数据，自然会包含大量结构化缺失，反映出底层人口在内在特征和行为方面固有的多样性。

虽然这些例子并不全面，但它们凸显了处理结构化缺失时面临的一些困难：在某些情况下，结构化缺失是可以预料的，也可能是恰当的；而在另一些情况下，它可能与可避免或不可避免的数据采集缺陷有关，并会严重阻碍后续学习。因此，理解和处理结构化缺失并非易事：我们需要一套适用于不同情境的工具，也需要相应的流程，以便随着社群对结构化缺失的认识不断发展，能够在这些工具的基础上继续完善。

## 2 结构化缺失的概念化

本节回顾缺失数据领域的标准文献，并讨论现有分类体系为何未能充分涵盖结构化缺失的复杂性。我们还将回顾一些导致结构化缺失的常见途径，并以一个真实世界的健康数据库为例，结合具体情境探讨这一问题。

假设我们有一个$n \times p$数据矩阵$X$，它对应$n$个单位在$p$个变量上的测量值。现在考虑一个相应的$n \times p$指示矩阵$R$，其元素为0或1；若$R_{ij} = 0$，则表示$X^{ij}$缺失；若$R_{ij} = 1$，则表示$X^{ij}$已观测到。我们可以将$X$分解为已观测到的部分（$X_{\text{obs}}$）和缺失数据部分（$X_{\text{mis}}$），如下所示：

$$X_{\text{obs}} = \{X_{ij} : R_{ij} = 1\} \quad \text{and} \quad X_{\text{mis}} = \{X_{ij} : R_{ij} = 0\}.$$

借助这些术语，我们可以定义缺失数据中的两个重要基本概念：（1）缺失数据机制；（2）缺失数据模式。

缺失数据机制指的是缺失值在数据中产生的过程。1976年，Rubin提出了一种基于条件分布 p(R |X $\phi$) 的缺失机制分类体系，此后该体系被采纳为缺失数据机制分类的标准 (3)。这一分类体系将值缺失的概率建模为数据中变量的函数，其中 p(·) 表示关系的一般函数形式，而 $\phi$ 表示模型中的参数，用于确定 R 与 X 之间的具体关系。具体而言，Rubin提出了以下三类缺失机制：（1）完全随机缺失 (MCAR)。在这种情况下，一个变量中的缺失读数与已观测和未观测的数据均相互独立；存在缺失数据的单位与数据完整的单位之间不存在系统性差异。对于完全随机缺失的数据，完整案例分析（即简单地将数据缺失的单位排除在分析之外）可能会降低统计功效，但不会引入偏倚。数学上，对于完全随机缺失的数据，p(R|X $\phi$) =p(R|$\phi$)。（2）随机缺失 (MAR)。在这种情况下，一个变量中的缺失读数取决于另一个变量的观测值。例如，在填写有关心理健康问题的问卷时，男性可能不如女性愿意透露信息。因此，假设性别数据完整，回答数据中的缺失就会与一个已观测到的特征相关。对于随机缺失的数据，完整案例分析可能会引入偏倚，因为它可能会选出并不代表整体数据的单位子集。数学上，对于随机缺失的数据，p(R|X $\phi$) = p(R|$X_{obs}$,$\phi$) (3)。（3）非随机缺失 (MNAR)。在这种情况下，一个变量中的缺失读数取决于其他变量的缺失情况，或取决于实验或数据收集过程范围之外的系统性因素。对于非随机缺失的数据，完整案例分析可能会引入偏倚；此外，由于偏倚来源并不明显，后续分析难以处理。数学上，对于非随机缺失的数据，p(R|X $\phi$) =p(R|X $\phi$)（即在此框架下不存在可用于简化的关系）。

缺失数据模式指缺失值在数据中的分布方式。尽管量化模式的方法有很多种（见下文挑战 2），但实践中通常采用两种分类：单调模式和非单调模式（见**图**。2 和（1））。单调模式是指一种模式，在这种模式下，可以对$p_{ij} = 0$，$R_{ik} = 0$进行排序，使得对所有$k > j$均成立。非单调模式是指无法以这种方式对变量进行排序的模式。单调缺失数据模式通常能更清晰地将数据分解为缺失部分（$X_{\text{mis}}$）和观测部分（$X_{\text{obs}}$），进而有助于采用可通过解析方式获得结果的方法

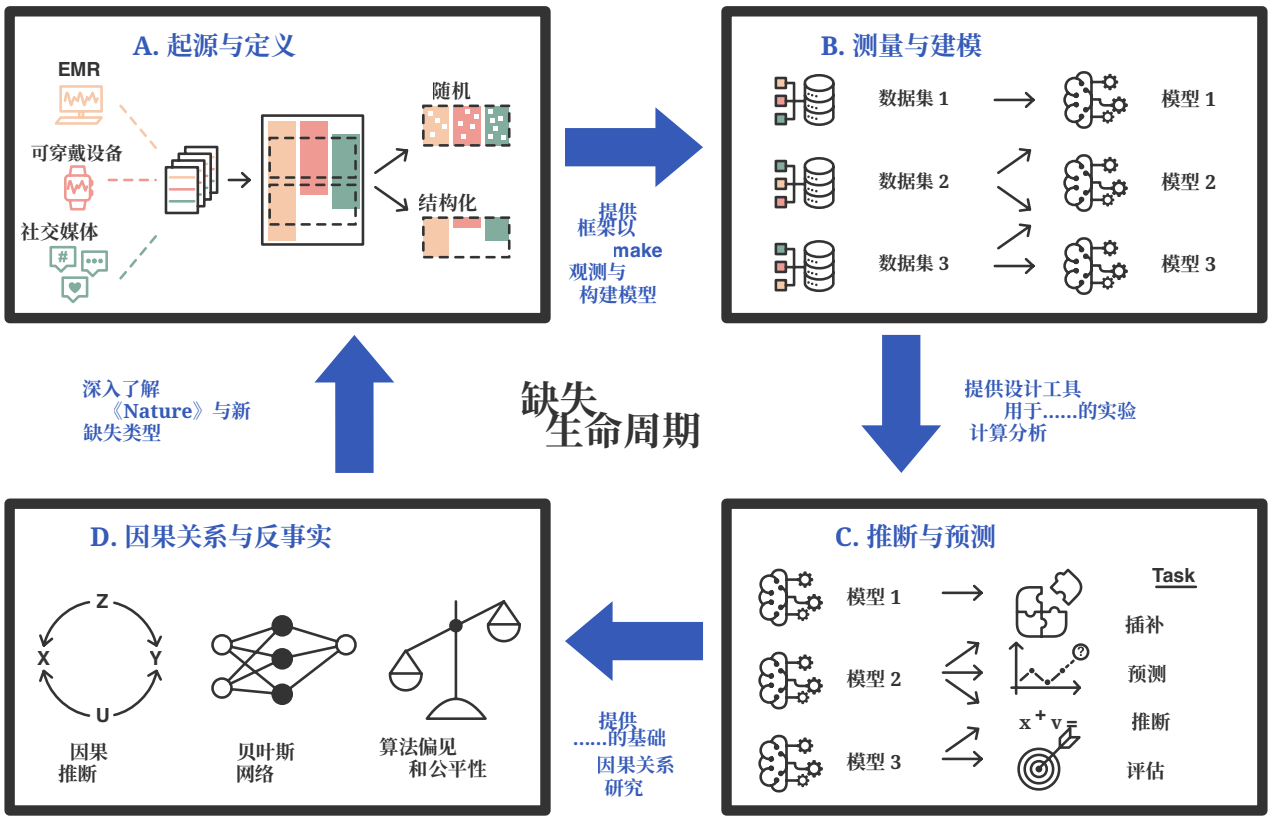


**图 1。数据缺失生命周期**。结构化缺失（SM）的存在会影响数据收集与分析生命周期的各个方面。（**A**）数据可能来自众多不同来源——本图示例包括电子病历、可穿戴设备和社交媒体。将这些数据关联起来后，可能会产生随机和结构化的缺失模式。要理解随机缺失数据与结构化缺失之间的相互作用，需要新的理论和工具，包括新的设计工具，以尽量减少结构化缺失对后续分析的影响。挑战 1–3 涉及这些问题。（**B**）结构化缺失会影响我们构建模型的能力，使模型能够恰当地、无偏倚地从数据中学习。我们需要新的工具来妥善处理结构化缺失，也需要能够在预测或推断场景下适应结构化缺失的新模型。挑战 4–5 涉及这些问题。（**C**）模型的效能往往（但并非总是）取决于我们准确插补缺失读数的能力。因此，我们需要新的结构化缺失插补工具，以及用于基准测试和评估各种方法表现的工具，这些方法旨在对结构化缺失进行建模和处理，包括预测和推断任务。挑战 4–7 涉及这些问题。（**D**）在许多科学领域，学习的最终目标是更好地理解因果关系。结构化缺失的存在可能严重妨碍这一努力。我们需要更好的工具，以便从存在结构化缺失的数据中量化并推断因果结构；还需要方法评估缺失数据与数据采集流程中的系统性偏差之间的关联程度，而这反过来可能妨碍因果机制的推断，并导致模型结果产生偏倚或不公平。挑战 8–9 涉及这些问题。总而言之，解决缺失生命周期各阶段的挑战，对于确保对存在结构化缺失的数据进行恰当分析并从中生成洞见至关重要。

(1）。然而，在高度多变量设置中，非单调模式更有可能出现。在这种情况下，通常只能借助马尔可夫链蒙特卡罗等计算方法近似求得结果，不过这些近似结果往往仍然相当准确。缺失数据模式看似与缺失数据机制存在相似之处，但二者是对缺失数据的不同且相互独立的刻画。例如，数据中的非单调模式可能由完全随机缺失、随机缺失或非随机缺失机制引起，也可能由这三种机制共同作用而产生。尽管通常很难明确说明模式与机制之间的关系，但确实存在一些简单的经验法则：例如，由于单调模式通常具有复杂的结构形式，因此不太可能由完全随机缺失机制产生。

Rubin缺失机制与模式之所以重要，是因为它们为识别不同的缺失类型提供了有力指引；有助于判断何时可能出现偏倚等问题，以及何时不会出现；也有助于判断插补等工具何时有效，何时无效。例如，目前已有许多强大的方法，可在假设数据为完全随机缺失的情况下处理缺失数据，其中包括各种插补工具（1）。对于随机缺失的数据，处理起来较为困难，但有时可以使用更复杂的插补工具来应对，例如多重链式方程插补（MICE）（32）。非随机缺失的数据定义最不精确、情况也最复杂，因此处理起来更加困难，通常需要非常谨慎地对待，并采用更务实、量身定制的方法处理：例如，手动检查并修正缺失模式，或通过敏感性分析，确定在不同的数据剔除或插补选择下分析结果会如何变化。

然而，尽管Rubin分类体系具有毋庸置疑的实用价值，但它们并不能完全解释现代机器学习应用中日益常见的结构化缺失的高维模式。例如，虽然显然结构化缺失与Rubin分类体系中的随机缺失和非随机缺失类别相关，但结构化缺失并不专属于其中任何一类。一个说明性示例见**图**。2。在此，医生决定哪些患者接受一系列检查，其依据要么完全是患者的年龄（见**图**。2A、B），要么是疑似疾病状态（见**图**。2C）。当决策基于年龄时，数据属于随机缺失，因为年龄是完全观测到的；而当决策基于未知的疾病状态时，数据则属于非随机缺失。更重要的是，缺失数据结构是不可观测的，因为它取决于未知变量，这是非随机缺失所导致的一种隐蔽且有害的后果。此外，根据检查是以确定性方式还是概率性方式进行（即临床医生是否总是将特定年龄的患者转诊接受检查，如**图**。2A所示，或是根据其他可能未被记录的特征运用其判断，如**图**。2B所示），最终形成的模式可能是单调的，也可能是非单调的。

在高维情境中——多种不同的生成机制可能会叠加在这些数据的不同缺失模式上——此类细微差异会使检测和描述数据集中存在的不同类型的结构化缺失变得困难。例如，单调模式（如块状缺失数据模式）通常表现得清晰明显，可能很容易可视化和描述。另一方面，非单调模式若源于底层结构，使缺失值稀疏分布于高度多变量空间，或出现在非随机缺失情境中，则可能带来更大危害，也更难描述，甚至难以检测。

随着人们对从复杂高维数据中学习的兴趣日益增长，这类缺失模式将越来越常见。若不加区分地使用此类数据训练机器学习模型，会带来风险，因为标准缺失数据方法无法充分应对结构化缺失带来的挑战。例如，对 **图**. 2A 中的数据进行完整案例分析（许多机器学习库的默认做法），会剔除所有年龄超过 85 岁的患者，从而因舍弃人口中的重要部分而引入偏倚。另一方面，常用的插补方法（如多重插补或基于树的插补）可能因数据是有意缺失的而不适用，也可能因缺失变量和观测的数量而引入很大的不确定性。因此，在此类情境中，开发能够解析结构化缺失（往往复杂）的几何结构的方法至关重要（见下文挑战 2）。

在**框**1 中，我们概述了结构化缺失（SM）的一些常见成因；在**框**2 中，我们通过一个详细的示例说明，在处理复杂的真实世界数据集中的结构化缺失时会出现哪些问题。这些示例表明，结构化缺失涵盖了范围极广的现象，无法整齐地归入当前的缺失数据模式或机制类别，而且每种现象都各自带来

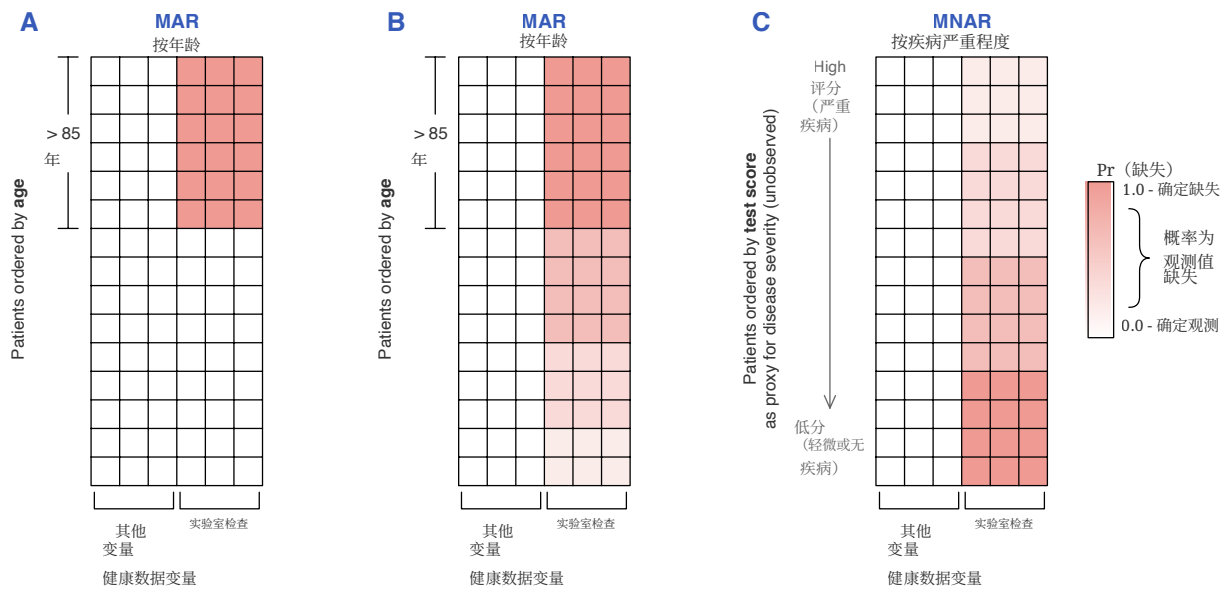


**图 2. 结构化缺失的示例**。缺失可能以微妙的方式产生，而经典理论往往无法充分解释这些情况。例如，医生可能会根据某位患者的健康状况或年龄、性别等其他相关特征，决定是否建议进行一系列检查，导致数据中的一组变量同时缺失或同时被观测到。由于临床医生的决策很少是非此即彼的，由此产生的缺失关系可能是概率性的，而非确定性的，因此一组变量更有可能同时缺失。（**A**）描述了一种假设情境：对所有低于某个临界年龄（此处为85岁）的患者进行实验室检查，而不对高于该年龄的患者进行检查。这是单调缺失数据模式的一个例子。（**B**）描述了一种更现实的情境：随着患者年龄增长，进行检查的概率会降低。虽然这是类似随机缺失（MAR）的缺失机制，但它会产生非单调模式。（**C**）描述了一种情境：医生更可能根据疑似疾病状态开具诊断性实验室检查，而该状态正是由检查本身揭示的（因此数据属于非随机缺失（MNAR）），由此产生结构化缺失（SM）的非单调模式。此处展示结构化缺失模式仅为说明之用：在现实中，由于该模式及其结构取决于一个未观测变量——作为潜在疾病严重程度代理指标的诊断检查评分——因此无法观测到缺失模式及其结构。

挑战。在复杂的多变量场景中，这些问题会进一步加剧：同一数据集中可能存在多种不同形式的结构化缺失，每种形式都由不同机制导致，呈现出各自的模式，并且这些模式可能与数据其他部分的缺失模式相关，也可能无关。这种极大的多样性阻碍了将结构化缺失确立为一个独立研究领域，因此我们目前仍缺乏对结构化缺失的各种常见形式进行分类、理解其对下游学习的影响或有效处理这些缺失的理论工具。

### 迈向结构化缺失分类体系

上文强调的这些示例表明，要从大型异质关联数据集中学习，我们需要以新的方式来理解结构化缺失，并将处理结构化缺失的工具作为机器学习流程的组成部分（见**图**。1）。

要做到这一点，会面临若干不同的数学和计算挑战。首先，现有分类体系无法充分涵盖结构化缺失作为一种现象的丰富内涵，而且我们缺乏用来定义和量化结构化缺失的机制与模式的术语：因此，我们需要新的数学表述，既扩展 Rubin 的经典分类法，又与之互为补充，并且

将结构化缺失视为一种本质上多变量的现象。事实上，仅仅检测并归类大型多变量数据集中存在的各种结构化缺失类型，就可能极具挑战性；据我们所知，目前还没有工具能够做到这一点。

其次，由于缺乏量化结构化缺失的正式框架，我们也缺少评估结构化缺失在多大程度上损害或造成机器学习模型偏倚的工具。在某些情况下，缺失数据即便数量显著且高度结构化，也不会影响模型性能。例如，如果缺失数据与感兴趣的结局关系不大，或者能够根据观测准确推断缺失读数，就可能出现这种情况。然而，在其他情况下，即便结构化缺失程度不高，也可能严重损害模型性能。例如，如果缺失数据涉及决定感兴趣结局的因果决定因素，或涉及对该结局具有很强预测力的变量，就可能如此。尽管在一定程度上可以针对具体情况处理这些问题，但目前我们缺少一个能够以有原则的方式普遍应对这些问题的正式框架。

第三，不同类型的结构化缺失对后续推断和学习的影响尚未得到充分探讨。例如，假设变量之间存在某种（也许未知的）关系，且其中一部分变量呈现结构化缺失，那么这会对涉及其他变量的后续分析产生什么影响?

第四，我们对结构化缺失对算法偏见与公平性的影响知之甚少。在医疗保健和刑事司法等高度敏感和/或受到监管的领域，机器学习工具如今已开始投入使用（33；11），公平性并非无关紧要的问题：为公共利益开发和部署的任何机器学习工具，都必须以严谨、透明的方式处理公平性问题（34）。为此，我们需要更深入地了解结构化缺失如何影响算法偏见，以及机器学习决策的公平性。

基于这些观察，艾伦·图灵研究所召开了一系列研讨会，以梳理这些问题，并确定结构化缺失未来研究的方向。在下一节中，我们将概述结构化缺失领域的九项重大挑战，这些挑战源自上述讨论。这些挑战可归为以下四类（见**图** 1）：（1）结构化缺失的起源与定义；（2）结构化缺失的测量与建模；（3）结构化缺失的推断与预测；（4）结构化缺失的因果关系。这些挑战共同界定了与结构化缺失相关的问题，为将结构化缺失发展为一个研究领域提供了路线图，并概述了一系列有助于推动机器学习大规模发展的未来研究方向。

**框 2：实践中的结构化缺失：一个启发性示例**

Flatiron Health-Foundation Medicine 临床基因组数据库（FH-FMI CGDB，以下简称 CGDB）是一个覆盖美国全国的纵向去标识化肿瘤学数据库，整合了约 280 家美国癌症诊所（约 800 个诊疗机构）收治患者的真实世界临床数据及结果，并结合了对 ~600 超过 400,000 份样本进行测序所得的独特癌症相关基因的全面基因组分析。纵向回顾性临床数据来源于电子健康记录（EHR）数据，包括经技术辅助信息提取整理的患者级结构化与非结构化数据；这些数据通过去标识化确定性匹配，与 CGDB 中源自 FMI 全面基因组分析（CGP）检测的基因组数据相关联（25；26）。CGDB 总计包含来自 ~100,000 名患者的数据，涵盖多种疾病特异性和非疾病特异性数据集。

由于涵盖范围广且复杂，CGDB 生动展现了处理结构化缺失所面临的挑战。此外，它还说明了，由于数据收集和关联实践存在差异，同一数据集中可能以多种不同方式出现结构化缺失。举例来说，我们讨论两种情况：首先，CGDB 中的基因组样本来自检测（称为检测项目），每项检测都会测量一组特定基因的基因改变状态。然而，诊断检测并非都相同，而且检测项目往往会随时间演变。根据开具检测的医生的治疗或预后目标，具有某种疾病状态的患者A可能接受包含400个基因的检测项目，而疾病状态相同的患者B可能接受包含150个基因的检测项目。任何结合

患者A和患者B及类似患者的记录，对患有该疾病的患者进行分析时，都必须处理两项检测中未测量基因所对应的数据缺失块。

将其与第二类结构化缺失数据问题进行对比：对于某些患者群体而言，某些数据在临床上并不相关（或无法）测量。例如，前列腺特异性抗原（PSA）是一项用于前列腺癌筛查、具有高度预后意义的血液检测；同样，血清CA125也广泛用作卵巢癌的肿瘤标志物。在CGDB中合并特定的癌症队列，例如转移性前列腺癌患者和卵巢癌患者，将揭示这些癌症特异性变量中的结构化缺失模式（女性患者不会测量PSA，男性患者不会测量CA125）。其他癌症特异性变量——例如格里森评分，即针对前列腺癌患者的预后分级评分——也会呈现类似的缺失结构。重要的是，在相关数据收集群体之外（例如男性/女性，或前列腺癌/卵巢癌患者），对这些变量进行估计并不恰当。然而，对数据进行更广泛的分析用途——例如使用包含男性和女性的队列构建预后模型，或跨多种癌症类型构建泛癌症模型——可能有理由纳入这些变量。在这些情况下，结构化缺失的出现并非源于数据采集问题，而是与部分特征对部分患者缺乏临床相关性有关。在考虑大型复杂数据集时，理解并处理此类结构化缺失模式会带来相当大的方法学挑战；这类数据集包含从众多不同来源获取的信息，CGDB就是其中一个例子。

## 3 结构化缺失的重大挑战

本节介绍应对结构化缺失并减轻其影响的九项重大挑战。这些挑战源于2021年11月至12月在艾伦·图灵研究所举办的一系列研讨会及后续讨论。

**挑战 1：定义结构化缺失**

尽管在鲁宾的奠基性论文发表之前，缺失数据已是一个众所周知的问题，但当时对它的处理缺乏科学严谨性，其成因也大多被视为偶然现象而遭到忽视（3）。直到鲁宾提出缺失数据机制，对缺失现象进行了正式的数学刻画，围绕如何处理缺失以及如何更有效地利用存在缺失读数的数据，才形成了活跃的研究领域。我们认为，结构化缺失也是如此。

如今显而易见，结构化缺失这一概念涵盖了各种无法由鲁宾的原始分类体系充分描述的缺失类型。毋庸置疑，难以应对一个定义不清的问题。因此，要有意义地处理结构化缺失，我们必须首先对其进行严格定义。一项关键挑战是识别结构化缺失的重要类型、它们与随机缺失和非随机缺失等现有表述有何不同，以及彼此之间有何关联。

有三个关键考量。首先，Rubin 分类体系的一个决定性特征是其数学严谨性：对 SM 的任何系统性刻画，都必须同样以严谨的数学理论为基础。一项核心挑战是制定一种能够有意义地对应现有机制和模式的刻画，并确保在适当情况下，过去的研究进展能够纳入这一新的框架。其次，任何新定义都必须克服现有刻画的不足，例如消除对 Rubin 最初提出的 MAR 的某些解释歧义（ 35 ； 36 ），并将 MAR 和 MNAR 假设扩展到多变量场景；这既不简单，也不直观（37）。第三，也可能是最重要的一点，任何针对 SM 的正式刻画都必须以便于实际应用的形式呈现：例如，用于判断何时适合进行插补、何时不适合，并为构建相应工具奠定基础（38）。

**挑战 2：探索结构化缺失的几何结构**

结构化缺失的根本特征是缺失值之间存在非随机的多变量关联模式。一种简单但可能十分有效的刻画这些模式的方法，是借助网络等数学结构以及

它们的推广形式，捕捉其几何和/或拓扑学特征，从而能够学习缺失值之间潜在的多变量关系。通过这种方式，所得到的网络有助于解决与数据缺口相关的问题，例如为设计数据库丰富策略提供指导。然而，这类用于理解缺失的几何方法尚未得到广泛探索。

网络是表示复杂数据的自然方式，也是关联表的有力替代方案。网络科学可提供便捷的工具来概括数据的结构属性，例如变量之间的关联模式（39）（例如，指示矩阵 $R^{ij}$ 可视为网络的关联矩阵）。然而，按照定义，网络模型只能捕捉变量之间的成对相互作用。在多变量结构性缺失情境下，高阶相互作用（即涉及两个以上变量的协同缺失模式）至关重要。此类高阶关联可通过单纯复形和超图等网络推广形式加以捕捉（40），并借助信息论工具进行分解（41）。近年来，这类高阶网络理论与模型及相关几何和拓扑工具在数据分析中的理论发展和实际应用取得了长足进展（42；43；44）。重要的是，这些高阶模型可同时用于统计、拓扑学和几何分析，因此有望以统计学和几何学上的严谨方法研究结构性缺失。同样，网络也可以通过元数据加以丰富，例如依据不同类别或层级对相互作用进行分类，从而形成所谓的多层网络（45；46）。总的来说，这些进展为研究结构性缺失开辟了新工具。挑战在于提出新颖的计算与数学表示，并将其与信息论和统计科学的工具结合起来，以统计上稳健的方式表示结构性缺失的结构，并对其几何与拓扑组织进行定量表征和可视化。

**挑战3：结构化缺失下的实验设计**

防患于未然，因此应尽可能在数据收集（即设计）阶段考虑结构化缺失。在研究的设计阶段考虑缺失已被证明具有优势（47；48；49），而且这在更广泛的结构化缺失情境中尤其重要，例如，简化因数据中的结构性失衡而受到不利影响的下游分析。要应对这些挑战，尤其需要以下三方面：（1）开发新的实验设计工具，以尽量减少结构化缺失的程度及其影响。尽管某些结构化缺失可能无法避免，但能否以一种尽量减轻其影响并预见未来缺失来源的方式收集数据？例如，未来某项实验的数据可能会与其他原本并非为合并而设计的实验数据相结合。设计能否预见并应对这类情况？为此，应探索采用通用标准。例如，建立统一的数据收集与记录框架，并将未来可能涉及的变量纳入考虑，可减少缺失块的产生和累积。（2）开发新颖的设计方法，以便表征并考虑结构化缺失的不同形式（见**框**1和挑战1）。此类方法可借鉴并拓展空间填充设计理论（50；51），以定位（并减轻其影响）特别重要的缺失区域。（3）开发新工具，结合先验知识与领域专长，预见结构化缺失的形式，并将其无缝纳入设计方法。如果没有可用的结构化缺失形式先验知识，则可考虑借鉴自适应设计理论的方法（52）。两阶段自适应设计可能尤其有用：第一阶段设计识别数据中存在的结构化缺失形式，随后优化第二阶段设计，以减轻这些形式对下游分析的影响。

**挑战4：结构化缺失下的预测**

利用多模态异质数据生成洞见，是预测建模的一项常见目标。然而，在许多预测性机器学习研究中，缺失数据的报告或处理方式并不完善，导致模型性能受损（53）。从存在结构化缺失的数据中进行预测是一项特殊挑战，因为许多能够处理复杂多变量关系的机器学习算法都要求数据完整，不含（或仅含少量）缺失值。例如，尽管在开发无监督深度生成模型方面已取得很大进展——这类模型能够生成具有复杂统计关联的合成数据，其关联模式与训练数据中的相匹配——

但在监督式深度学习场景中，如何处理缺失数据却鲜少受到关注（54）。要在这一领域取得进展，需要推进以下五个方面：（1）开发新工具，更好地理解数据中存在的结构与缺失模式之间的关系，并将这些模式纳入预测模型（另见挑战 2）。（2）开发适用于预测任务的新型插补技术，以处理缺失结构，并在适当情况下加以利用（见挑战 6）。（3）探索组合模型的新方法，使模型能够从数据的子结构中学习，并整合为同时涵盖完全观测和部分观测特征的整体模型。（4）开发新工具，以便稳健地预测结构化缺失本身，包括准确预测训练数据之外的缺失模式。（5）探索评估模型性能和不确定性的新方法，并将缺失结构纳入考量（见挑战7）。例如，在临床情境中，与未接受基因检测的患者相比，模型对拥有遗传数据的患者所作的预测可能更准确。因此，需要开发工具，评估模型对这两类患者的差异化性能，从而间接评估遗传数据缺失对模型性能的影响。

**挑战5：结构化缺失下的推断与估计**

统计推断是根据数据估计目标人群特征的过程。结构化缺失的存在可能会严重损害这一过程。例如，如果缺失值主要集中在数据的某一部分（例如与某个特定人群亚组相关），那么基于这些数据的估计程序，其不确定性可能高于基于数据中更完整部分的估计程序。在机器学习情境中，这可能导致有偏模型：模型在部分数据子集上表现良好，在其他子集上却表现不佳。例如，研究已发现，多种自动化人脸识别算法在识别浅肤色男性面孔时准确率很高，但在识别深肤色女性面孔时则存在不确定性（55）。这种不确定性部分源于底层训练数据集中深肤色受试者代表性不足，导致模型无法准确学习与这一群体相关的特征（56）。在这种情况下，模型的不确定性提供了有关缺失模式的有用信息，而这些问题能够也应当得到解决。因此，这些问题对相关工具的公平性有着深远影响，并非仅具有技术层面的重要性（参见挑战9）。

此外，在某些类型的结构化缺失下，某些推断可能无法进行。例如，如果对于两个分类变量中的每一个变量，某个类别水平的数据都同时缺失，那么在基于标准似然的推断中，将无法估计这些类别水平交互作用的参数。贝叶斯推断程序可通过施加信息性先验分布（例如借助领域专家知识）来缓解其中一些问题，从而能够估计模型参数（57）。贝叶斯方法还可以自然地将不确定性（包括模型不确定性）纳入推断，并通过马尔可夫链蒙特卡罗等计算方法对数据拟合复杂模型（58）。然而，目前我们仍缺乏将结构化缺失纳入这类贝叶斯框架的严谨方法，也缺乏用于评估何时能够可靠地进行推断、何时不能的工具。此外，我们也缺乏能够在不同类型的缺失并存时进行推断的工具。例如，如果数据的一部分存在非随机缺失，另一部分存在随机缺失，但缺失值还呈现出一般性的全局结构（见挑战 2），那么在局部层面（例如仅限于包含随机缺失的数据时），推断会受到怎样的影响？此外，结构化缺失的全局模式会如何影响此类推断的可解释性？关键挑战在于开发能够明确考虑结构化缺失且不损害可解释性的推断工具，区分不同缺失机制对后续推断的影响，并生成无偏且可解释的总体特征估计，从而支持有充分把握地得出结论。

**挑战 6：填补结构化缺失**

插补是根据观测数据，用合理且具有代表性的数值替换缺失值（或可疑值）的过程（59）。通常会进行多重插补，以估计不确定性或变异性（32）。随后，这些集成结果既可用于估计隐含生成模型的参数，也可将插补后的数值直接作为后续分析的输入，例如用于构建预测模型。

虽然有许多方法可以对完全随机缺失的数据进行插补，但在存在结构化缺失的情况下，插补既带来了挑战，也带来了机遇。当缺失数据以块结构出现时，对数据中缺失比例很高的一大块独立部分进行插补是否明智，值得商榷，因为这很可能导致插补值不可靠。此外，对于某些类型的结构化缺失，插补可能并不适用（示例见 **方框** 2 ）。重要的是，对于存在大量结构化缺失的复杂多变量数据集，要判断何时适合插补、何时不适合，可能并不简单，而且可能取决于许多因素。这些因素可能是数据本身固有的，例如其他缺失数据的插补值；也可能是外在的，例如领域专门知识。亟需制定严谨的方法来解决这一问题，包括纳入先验知识。

在某些情况下，结构化缺失也提供了构建信息更充分的插补模型的机会，这些模型将缺失值本身作为信息加以利用。例如，在一项关于性健康的（假设）研究中，男性可能不愿回答有关性功能丧失的问题。如果随后我们在研究中发现一条性别和性功能的回答均缺失的记录，那么我们可能有理由相信，缺失的性别很可能是男性，并且可以将这一信息纳入我们的插补模型和/或任何下游学习中。

**挑战7：结构化缺失工具的基准测试与评估**

评估一种处理缺失数据的新技术时，通常会遵循两个步骤：首先，在原本完整的数据集（可以是模拟数据集，也可以是真实数据集）中，按照预先设定的结构人为引入缺失读数。这个过程称为“人工删失”（60；61；62）。其次，应用这项新技术（例如插补）及任何相关的后续分析（例如分类或推断），并根据适当的成功度量指标（例如分类中的平衡准确率或相关指标），将结果与未进行人工删失时的结果进行比较。

原则上，评估SM方法可以遵循同样的流程。然而，在复杂的多变量场景中，以能准确反映所关注结构化缺失模式的方式对数据进行人工删失，本身就可能是一项重大挑战。通常，逐个变量移除观测的标准方法并不奏效，因为它无法捕捉预期的多变量缺失模式（60）。模拟复杂的多变量缺失结构，需要新的工具来学习并量化缺失模式本身（见挑战2和挑战4）。因此，基准测试问题与理解结构化缺失的问题密不可分：也就是说，要对SM工具进行基准测试，就必须先理解结构化缺失，而这恰恰是这些工具旨在帮助我们理解的内容。我们需要创新的缺失基准测试方法，以避免陷入这一僵局。

**挑战8：结构化缺失下的因果关系**

众所周知，缺失数据通常会损害因果推断。结构化缺失的影响尤其严重。例如，如果一组特定变量的数据系统性缺失，就无法确定它们与感兴趣的结局之间的因果关系。由于固有的困难，关于缺失数据下因果推断的文献非常少；当输入变量和潜在结果都存在系统性缺失时，研究尤其具有挑战性（63）。因此，即使能够克服刻画数据缺失结构这一难题（见挑战1-2），衡量不同缺失结构对因果推断过程的影响，并开发用于解决这些问题的基准工具，也仍是一项挑战（见挑战7）。

因果推断与结构化缺失（SM）密不可分，还有两个原因。首先，因果推断在确保算法公平性或验证公平性方面发挥着重要作用（64），因此，因果关系问题本质上与

偏倚、推断和预测问题相关（见挑战4、5和9）。例如，生物医学数据集在性别、种族等受保护属性方面往往存在偏倚。在某些情况下，这是由真实的生理原因造成的：心脏病数据集中男性患者的样本可能多于女性患者；皮肤癌数据集中较浅肤色患者可能多于较深肤色患者。虽然这些偏倚可能准确反映患者人口统计特征，因此从统计意义上说并无偏倚，但它们自然会导致类别不平衡，因此从机器学习的角度看仍然存在偏倚，并可能导致算法在针对某些患者群体时系统性地失效。重要的是，因果工具有助于识别这些偏倚，并指出减轻偏倚的方法（65）。开发能够应对结构化缺失（SM）的新型因果推断方法，或识别出必须先加以改进才能回答因果问题的数据子集，对于以公平、透明的方式大规模推进机器学习至关重要。

其次，因果推断为整体理解结构化缺失（SM）提供了一个框架。事实上，因果推断问题可以重新表述为缺失数据问题（66），即对于所有可能的结果，至多只有一个潜在结果可以被观测到，其余结果则无法观测，或属于“缺失”。因此，因果推断目标是使用适当的方法估计这些本质上缺失的结果（即反事实）。将这一思路拓展到结构化缺失，我们可以通过逆概率加权和插补等因果推断方法，进一步了解如何估计缺失结构（67；68；69）。一项关键挑战是理解并量化已测量值和未测量值对目标估计量的影响。

**挑战9：结构化缺失的伦理影响**

结构化缺失可能源于数据采集与分析生命周期各个阶段的社会文化偏倚（见**图**。1），从采集哪些数据及如何采集，到数据处理、分析以及将数据用于决策（70）。因此，在人口的特定社会人口学亚组中，某些变量（或变量组）的缺失程度可能更高（71；72）。例如，受保护群体中的个体可能不愿提供可能带来不利影响的信息或敏感信息。医疗从业者也可能对边缘化群体中的患者抱有有意或无意的偏倚，并基于这些偏倚作出不同的诊断和治疗决策（73）。在某些情况下，历来服务不足的群体的数据可能完全缺失，或几乎全部缺失。例如，生物医学数据集中的样本明显更多来自西方人群——以基因组学为例，86%的样本来自欧洲血统个体（74）；而在人类微生物组研究中，超过70%的样本来自欧洲、美国或加拿大（75）。

结构化缺失在伦理层面是否存在问题，取决于研究的目的。当偏倚导致研究结果无法推广至更广泛的人群，或由于某些变量缺失或统计效能不足而无法理解解释因素的重要性时，问题便会出现（另见挑战8）。首先，在数据集中引入记录缺失情况的标准化流程，有助于确保结论恰当地承认这些偏倚（76；77）。

在某些情境下，结构化缺失可能显而易见；但如果未能察觉，也可能出现在更隐蔽的情境中。根据某些社会人口学特征量化结构化缺失的方法，或可揭示数据中存在的社会偏见（78）。要有效缓解偏倚，必须充分关注数据生成背后的社会过程，并结合具体情境，认识到歧视模式可能受到社会、文化和历史决定因素的影响。因此，邀请具备领域知识和社会科学训练的专家参与至关重要。

如果未得到妥善处理，SM 可能会延续或加剧现有的不平等现象（79；80）。尽管应对上述各项挑战的进展都有望减轻此类危害，但要彻底解决这一问题，还需要开展明确的跨学科努力，着力处理偏倚与公平性问题，并将代表性或同质性方面的统计问题与相关伦理原则和价值观联系起来。因此，在尝试应对我们所述的所有 SM 挑战时，伦理考量应处于核心位置。例如，SM工具的基准测试和评估应纳入公平性考量；预测和插补方法应考虑 SM 可能导致的偏倚；还应采取措施，确保可能出现 SM 的数据收集尽可能具有代表性。

## 4 结论

近年来，廉价数字存储和计算能力的可用性迅速提升，机器学习（ML）方法也取得了颠覆性进展。这些方法能够从海量数据中学习——包括为研究目的而收集、整理并日益广泛开放的数据——并执行众多下游任务（14）。然而，这些进展也带来了自身的挑战。尽管自然语言处理和计算机视觉等特定机器学习领域取得了非凡进步，但值得注意的是，这些领域往往依赖大量免费开放的训练数据资源（81）。而在医疗保健和社会经济政策等其他重要领域，训练数据并不容易获取（82），且往往由众多彼此独立的数据资源汇编而成，这些资源源自不同的现实世界实验，可能在生成时并未考虑后续整合。在这种情况下，模型训练所需的合并数据往往充斥着缺失或不完整的数据，以及其他广为人知的问题（83）。处理随时间不断积累新信息的“动态”数据集时，这些问题尤为突出（84）：此类数据集会将新信息与旧信息合并，并在新的测量模态出现时纳入相关数据，例如基因测序领域不断发展的新进展（85）。要使机器学习方法能够从这类动态异质数据中学习并实现稳健的泛化，就必须设计为能够应对不可避免的“结构化缺失”（SM）。这些问题有别于标准数据清洗流程相关的模型退化和偏倚问题（86；87）。因此，我们认为，当前亟需将 SM 作为一个独立议题加以解决，因为它对机器学习的未来至关重要。为此，我们提出了九项 SM 挑战；若能解决这些挑战，将有助于我们大规模地从现实世界数据中有效学习。应对这些挑战时，有两个贯穿始终的关键考量。

首先，充分了解方法所应用的学科背景或领域背景至关重要，同时也要认识并清晰阐明这些方法的局限性。例如，如果计算资源或数据资源不足以全面考察分析流程中结构化缺失（SM）的程度及其影响，就应明确说明这一点，并将可能由此产生的偏倚及其实际影响作为报告流程的一部分，清楚、毫不含糊地告知利益相关方。为此，需要在更广泛的背景下，以整体方法应对结构化缺失，涉及机器学习、数学、网络科学、伦理学和统计设计等多个学科。因此，有必要举办由社区主导的跨学科研讨会，并开发开源协作平台等平台，充分利用各领域的专业知识，推动这一新兴领域的发展。

其次，不应仅将结构化缺失视为事后 问题，并留待分析阶段处理。要尽可能有效地应对结构化缺失，关键是从设计阶段就开始考虑这一问题，并在适当范围内尽可能减少结构化缺失的数量及其影响（见挑战3）。尽管尽了最大努力，某些情况下结构化缺失仍不可避免，甚至可能是适当的（见框1）；此外，其存在也可能未被发现。为更好地检测和表征结构化缺失，需要采用创新方法，例如使用网络模型或拓扑模型（见挑战 2），以及其他数据驱动方法，如无监督聚类。重要的是，这类事后分析工具并非独立于设计考量：通过良好的设计与分析相结合，并将二者作为一体化分析流程中相互关联的环节，才能最有效地应对结构化缺失。

最后，我们提出一些可能的思路，以应对前文所述的挑战。尽管要全面解决这些挑战，无疑需要开发新的数学和计算方法，但现有的一些工具和技术或许有助于为未来研究开辟有前景的方向。

应对结构化缺失（88）的一种可能方法，是借鉴稀疏与协同表示的概念。协同表示模型借鉴了协同过滤器的原理，而协同过滤器常用于在线推荐系统（89）。协同过滤器会长期跟踪许多用户的活动——每位用户都可能因其特定兴趣和档案特征而呈现复杂且独特的缺失数据模式（见框1）——并通过将个体活动与整个社群所学到的模式进行比较，为其提供个性化推荐（90）。协同

表示分类器（91）还利用了这样一个事实：个体档案通常可以用稀疏基集简洁地表示，而稀疏基集可通过汇集更广泛人群的数据来学习。这些方法利用了两个可能有助于克服结构化缺失所带来挑战的关键事实：（1）对于某些分析，通过以分布式方式在大型社群中收集和汇总部分数据，可以减轻缺失数据的影响。（2）在许多情况下，表面上具有高维特征的现象可以在较低维度上得到有效表示，而不会显著损失信息；这一压缩过程还可以起到“去噪”作用（例如，人类初级视觉皮层利用这种稀疏编码，有效表示复杂的自然场景（92））。

另一种可能是借鉴合成数据领域的近期概念（93；94）。直观而言，生成合成数据的问题可视为插补的一种极端情形，即所有数据均缺失。因此，在适当情况下，合成数据生成工具或可用于插补结构化缺失。深度生成模型等方法，尤其是自注意力变换器（95）和对抗网络（96），可能特别有用。例如，用于生成深度伪造内容（97）和迁移风格（98）的技术，正开始应用于科学应用，例如对组织病理学图像进行标准化（99）。同样，基于这些技术开发的强大工具，已能够在具有理论保证的情况下插补完全随机缺失的数据；对于稀疏的随机缺失和非随机缺失数据，这些工具也具有实证有效性（100）。尽管这些方法尚不能处理结构化缺失的所有形式，但它们展现出非常promising 的发展前景。此外，它们也暗示了机器学习与结构化缺失研究之间可能形成良性循环：机器学习的进步有助于我们更好地理解结构化缺失，而结构化缺失研究的进展则有助于我们充分利用现有数据资源，并开发更强大的机器学习方法。

## 致谢

本研究由Turing-Roche战略合作伙伴关系资助。作者感谢ChloeMatus博士为图示创作与设计贡献才华，也感谢Vicky Hellon在社区管理方面提供的专业支持。

## 利益冲突

作者声明不存在利益冲突。

## 参考文献

**1.** Roderick JA Little 和 Donald B Rubin。 缺失数据的统计分析，第793卷。John Wiley & Sons，2019年。**2.** Bojan Karlaˇs、Peng Li、Renzhi Wu、Nezihe Merve G¨urel、Xu Chu、Wentao Wu 和 Ce Zhang。针对不完整信息的最近邻分类器：从确定答案到确定预测。 *arXiv*预印本 *arXiv:2005.05117*，2020。**3.** Donald B Rubin。推断与缺失数据。 *Biometrika*，63(3):581–592，1976年。**4.** Therese D Pigott。缺失数据方法综述。教育研究与评估，7(4):353–383，2001 年。**5.** Joseph L Schafer 和 John W Graham。缺失数据：我们对现状的看法。 心理学方法，7(2):147，2002年。**6.** Daniel F Heitjan 和 Donald B Rubin。可忽略性与粗粒度数据。 *The Annals of Statistics*，第 2244–2253 页，1991 年。**7.** Tlamelo Emmanuel 、 Thabiso Maupong 、Dimane Mpoeleng、Thabo Semong、Banyatsang Mphago 和 Oteng Tabona。机器学习中的缺失数据综述。*Journal of Big Data*，8(1):1–37，2021 年。**8.** Jing Gao 、 Peng Li 、 Zhikui Chen 和 Jianing Zhang。多模态数据融合深度学习综述。 *NeuralComputation*，32(5):829–864，2020。

9. Xiaoqiang Yan、Shizhe Hu、Yiqiao Mao、Yangdong Ye 和 Hui Yu。深度多视图学习方法：综述。Neurocomputing，448:106–129，2021年。10. Chang Xu、Dacheng Tao 和 Chao Xu。多视图学习综述。arXiv预印本 arXiv:1304.5634，2013年。11. Eric J Topol。高性能医学：人类与人工智能的融合。Nature Medicine，25(1):44–56，2019年。12. Lu´ıs A Vale Silva 和 Karl Rohr。采用多模态深度学习进行泛癌症预后预测。载于2020年第17届IEEE生物医学成像国际研讨会（ISBI），第568–571页。IEEE，2020。13. Donald B Rubin。多重插补在 18+年之后。Journal of the American Statistical Association，91(434):473–489，1996年。14. Rishi Bommasani、Drew A Hudson、Ehsan Adeli、Russ Altman、Simran Arora、Sydney von Arx、Michael S Bernstein、Jeannette Bohg、Antoine Bosselut、Emma Brunskill 等。基础模型的机遇与风险。arXiv预印本 arXiv:2108.07258，2021年。15. Georgios A Kaissis、Marcus R Makowski、Daniel R¨uckert 和 Rickmer F Braren。医学成像中的安全、保护隐私的联邦机器学习。Nature Machine Intelligence，2(6):305–311，2020。16. Tian Li、Anit Kumar Sahu、Ameet Talwalkar 和 Virginia Smith。联邦学习：挑战、方法与未来方向。IEEE Signal Processing Magazine，37(3):50–60，2020。17. C Holmes。人工智能与健康：2019年1月16日举行的圆桌会议摘要报告。英国 Academy of Medical Sciences，2019年。网址：https://acmedsci.ac.uk/policy/policy-projects/ artificial--intelligence-and-health。18. Xuesi Dong、Lijuan Lin、Ruyang Zhang、Yang Zhao、David C Christiani、Yongyue Wei 和 Feng Chen。Tobmi：采用加权k近邻方法进行泛组学块状缺失数据插补。Bioinformatics，35(8):1278–1283，2019年。19. Tatsuhiko Naito、Ken Suzuki、Jun Hirata、Yoichiro Kamatani、Koichi Matsuda、Tatsushi Toda 和 Yukinori Okada。一种用于HLA插补及1型糖尿病跨族群MHC精细定位的深度学习方法。Nature Communications，12(1):1–14，2021年。20. Vincent Audigier、Ian R White、Shahab Jolani、Thomas PA Debray、Matteo Quartagno、James Carpenter、Stef Van Buuren 和 Matthieu Resche-Rigon。针对含连续变量和二元变量的多层级数据进行多重插补。Statistical Science，33(2):160–183，2018年。21. Roline Kamphuis、Shahab Jolani 和 Peter Lugtig。缺失数据的分块插补方法。2018年7月。doi：10.13140/RG.2.2.12467.32803。网址：10.13140/RG.2.2.12467.32803。22. Zhengping Che、Sanjay Purushotham、Kyunghyun Cho、David Sontag 和 Yan Liu。针对含缺失值的多变量时间序列的循环神经网络。Scientific Reports，8(1):1–12，2018年。23. Zhenhua Wang、Olanrewaju Akande、Jason Poulos 和 Fan Li。深度学习模型在大型调查缺失数据插补方面是否更优？基于实证比较的证据。arXiv预印本 arXiv:2103.09316，2021年。24. Nicholas J Tierney、Fiona A Harden、Maurice J Harden 和 Kerrie L Mengersen。利用决策树理解缺失数据中的结构。BMJ Open，5(6):e007450，2015年。

25. Gaurav Singal, Peter G Miller, Vineeta Agarwala, Jie He, Anala Gossai, Shannon Frank, David Bourque, Bryan Bowser, Thomas Caron, Ezra Baydur, Kathi Seidl-Rathkopf, Ivan Ivanov, Alex Parker, Ameet Guria, Garrett Michael Frampton, Ann Jaskiw, Dana Feuchtbaum, Nathan Coleman Nussbaum, Amy Pickar Abernethy, and Vincent A Miller. 真实世界临床基因组学数据库的开发与验证。《Journal of Clinical Oncology》，35(15):2514，2017年。26. Benjamin Birnbaum, Nathan Nussbaum, Katharina Seidl-Rathkopf, Monica Agrawal, Melissa Estevez, Evan Estola, Joshua Haimson, Lucy He, Peter Larson, and Paul Richardson. 借助模型和偏倚分析，从电子健康记录中生成大规模肿瘤学研究队列。2020。27. Emy Alerskans, Jacob L Høyer, Chelle L Gentemann, Leif Toudal Pedersen, Pia Nielsen-Englyst, and Craig Donlon. 基于被动微波测量构建海表温度气候数据记录。《Remote Sensing of Environment》，236:111485，2020。28. Pari-Sima Katiraie-Boroujerdy, Nasrin Nasrollahi, Kuo-lin Hsu, and Soroosh Sorooshian. 伊朗卫星降水估算评估。《Journal of arid environments》，97:205–219，2013年。29. Tom R Andersson, J Scott Hosking, Mar´ıa P´erez-Ortiz, Brooks Paige, Andrew Elliott, Chris Russell, Stephen Law, Daniel C Jones, Jeremy Wilkinson, Tony Phillips, et al. 利用概率深度学习预测北极海冰的季节变化。《Nature communications》，12(1):1–12，2021年。30. Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 调查方法学。John Wiley & Sons，2011。31. Heidi Ledford. Facebook、Twitter及其他数据宝库如何推动社会科学变革。《Nature》，582(7812):328–331，2020。32. Stef Van Buuren and Karin Groothuis-Oudshoorn. mice：R语言中的多重链式方程插补。《Journal of statistical software》，45:1–67，2011。33. David Leslie, Christopher Burr, Mhairi Aitken, Josh Cowls, Michael Katell, and Morgan Briggs. 人工智能、人权、民主与法治：入门指南。arXiv预印本 arXiv:2104.04147，2021年。34. Ben D MacArthur, Cosmina Dorobantu, and Helen Margetts. 有韧性的政府需要数据科学改革。《Nature human behaviour》，2022年。doi：10.1038/s41562-022-01423-6。35. Shaun Seaman, John Galati, Dan Jackson, and John Carlin. “随机缺失”是什么意思？《Statistical Science》，28(2):257–268，2013年。36. Marco Doretti, Sara Geneletti, and Elena Stanghellini. 缺失数据：以条件独立为指导的统一分类体系。《International Statistical Review》，86(2):189–204，2018年。37. Jin Tian. 图模型中的随机缺失。载于《Artificial Intelligence and Statistics》，第977–985页。PMLR，2015年。38. Luigi Antelmi, Nicholas Ayache, Philippe Robert, Federica Ribaldi, Valentina Garibotto, Giovanni Frisoni, and Marco Lorenzi. 结合多任务学习和多通道变分自编码器，利用含有缺失观测的数据集——应用于痴呆症多模态神经影像研究。HAL Open Science，HAL id: hal-03114888，2021年。

**39.** Mark Newman。 网络。 牛津大学出版社，2018年。

40. Ginestra Bianconi。《高阶网络》。剑桥大学出版社，2021年11月。doi: 10.1017/9781108770996。URL https://doi.org/10.1017/9781108770996。41. Aaron J Gutknecht、Michael Wibral 和 Abdullah Makkeh。《点滴之中：通过部分—整体关系和形式逻辑理解信息分解》。《英国皇家学会会刊 A 辑》，477(2251):20210110，2021年。42. Christian Bick、Elizabeth Gross、Heather A Harrington 和 Michael T Schaub。什么是高阶网络？arXiv 预印本 arXiv:2104.11329，2021年。43. Gunnar Carlsson。拓扑学与数据。《美国数学会通报》，46(2):255–308，2009年。44. Parvaneh Joharinad 和 J¨urgen Jost。数据几何。arXiv 预印本 arXiv:2203.07208，2022年。45. Ginestra Bianconi。《多层网络》。牛津大学出版社，2018年7月。doi: 10.1093/oso/9780198753919.001.0001。URL https://doi.org/10.1093/oso/9780198753919.001.0001。46. Narsis A. Kiani、David Gomez-Cabrero 和 Ginestra Bianconi（编）。《生物学中的网络之网络》。剑桥大学出版社，2021年2月。doi: 10.1017/9781108553711。URL https://doi.org/10.1017/9781108553711。47. Kim May Lee、Stefanie Biedermann 和 Robin Mitra。含脱落的多组试验的 D-最优设计。《医学统计学》，38(15):2749–2766，2019年3月。doi: 10.1002/sim.8148。URL https://doi.org/10.1002/sim.8148。48. Kim May Lee、Robin Mitra 和 Stefanie Biedermann。当结局值并非随机缺失时的最优设计。《中华统计学志》，2018年。doi: 10.5705/ss.202016.0526。URL https://doi.org/10.5705/ss.202016.0526。49. Kim May Lee、Stefanie Biedermann 和 Robin Mitra。针对可能存在不完整观测的实验的最优设计。《中华统计学志》，2018年。50. Jack Noonan 和 Anatoly Zhigljavsky。高维集合的非格覆盖与量化。载于《黑箱优化、机器学习与无免费午餐定理》，第273–318页。施普林格国际出版公司，2021年。doi: 10.1007/978-3-030-66515-9\ 10。URL https://doi.org/10.1007/978-3-030-66515-9_10。51. Anatoly Zhigljavsky 和 Jack Noonan。高维立方体的覆盖与量化。《SN运筹学论坛》，1(3)，2020年8月。doi: 10.1007/s43069-020-0015-8。URL https://doi.org/10.1007/s43069-020-0015-8。52. Thomas Burnett 和 Christopher Jennison。自适应富集试验：有何益处？《医学统计学》，40(3):690–711，2020年11月。doi: 10.1002/sim.8797。URL https://doi.org/10.1002/sim.8797。53. SWJ Nijman、AM Leeuwenberg、I Beekers、I Verkouter、JJL Jacobs、ML Bots、FW Asselbergs、KGM Moons 和 TPA Debray。机器学习预测模型研究中的缺失数据处理和报告不佳：文献综述。《临床流行病学杂志》，142:218–229，2022年。54. Niels Ipsen、Pierre-Alexandre Mattei 和 Jes Frellsen。如何处理监督式深度学习中的缺失数据？载于 Artemiss-ICML缺失值学习艺术研讨会，2020年。55. Joy Buolamwini 和 Timnit Gebru。性别阴影：商业性别分类中的交叉性准确率差异。载于公平、问责与透明度会议，第77–91页，2018年。

[{"id": 0," output": "18/21"}]

56. D. Leslie。《理解面部识别技术中的偏倚：解读》，2020。57. Andrew Gelman、John B Carlin、Hal S Stern、David B Dunson、Aki Vehtari 和 Donald B Rubin。《贝叶斯数据分析》。2013。58. Alan E Gelfand 和 Adrian FM Smith。计算边缘密度的基于抽样的方法。《Journal of the American Statistical Association》，85(410):398–409，1990年。59. Stef Van Buuren。《缺失数据的灵活插补》。CRC出版社，2018年。60. Rianne Margaretha Schouten、Peter Lugtig 和 Gerko Vink。为模拟目的生成缺失值：一种多变量截除程序。《统计计算与模拟杂志》，88(15):2909–2930，2018年7月。doi: 10.1080/00949655.2018.1491577。URL https://doi.org/10.1080/00949655.2018.1491577。61. J. P. L. Brand。《不完整数据集统计分析的多重插补策略：开发、实施与评估》。Print Partners Ispkamp，恩斯赫德，1999年。62. Jaap P. L. Brand、Stef Van Buuren、Karin Groothuis-oudshoorn 和 Edzard S. Gelsema。用于评估多重插补方法的 SAS 工具包。《荷兰统计学》，第36–45页，2003年。63. Imke Mayer。《基于异质数据和缺失数据的因果推断：在重症监护管理中的应用》。博士论文，巴黎，法国高等社会科学研究院，2021年。64. Matt J. Kusner、Joshua Loftus、Chris Russell 和 Ricardo Silva。反事实公平性。NeurIPS，2017。65. Aili Shen、Xudong Han、Trevor Cohn、Timothy Baldwin 和 Lea Frermann。用于公平表征的对比学习。arXiv:2109.10645，2021。66. Peng Ding 和 Fan Li。因果推断：缺失数据视角。Statistical Science，33，2017年12月。doi: 10.1214/ 18-STS645。67. Shaun R Seaman 和 Ian R White。处理缺失数据的逆概率加权综述。《医学研究统计方法》，22(3):278–295，2013年。68. BaoLuo Sun、Neil J Perkins、Stephen R Cole、Ofer Harel、Emily M Mitchell、Enrique F Schisterman 和 Eric J Tchetgen Tchetgen。单调和非单调缺失数据的逆概率加权估计。《美国流行病学杂志》，187(3):585–591，2017年11月。ISSN 0002-9262。doi: 10.1093/aje/kwx350。URL https: //doi.org/10.1093/aje/kwx350。69. Daniel Westreich、Jessie K Edwards、Stephen R Cole、Robert W Platt、Sunni L Mumford 和 Enrique F Schisterman。因果推断中潜在结果的插补方法。《国际流行病学杂志》，44(5):1731–1737，2015年。70. Robert A Verheij、Vasa Curcin、Brendan C Delaney 和 Mark M McGilchrist。初级保健电子健康记录数据使用与再利用中可能的偏倚来源。《医学互联网研究杂志》，20(5):e185，2018年5月。ISSN 1438-8871。doi: 10.2196/jmir.9134。URL http://www.jmir.org/2018/5/e185/。71. Mathew V. Kiang、Jarvis T. Chen、Nancy Krieger、Caroline O. Buckee、Monica J. Alexander、Justin T. Baker、Randy L. Buckner、Garth Coombs、Janet W. Rich-Edwards、Kenzie W. Carlson 和 Jukka-Pekka Onnela。数字表型分析中缺失数据的社会人口特征。Scientific Reports，11(1)，2021年7月。doi: 10.1038/s41598-021-94516-7。URL https://doi.org/10.1038/s41598-021-94516-7。

**72.** Thomas Tsiampalis 和 Demosthenes B. Panagiotakos。缺失数据分析：ATTICA流行病学研究背景下，自我报告心理与营养相关多项目量表低响应率的社会人口学、临床及生活方式决定因素。*BMC Medical Research Methodology*，20(1)，2020年6月。doi: 10.1186/s12874-020-01038-3。URLhttps://doi.org/10.1186/s12874-020-01038-3。

**73.** David Leslie、Anjali Mazumder、Aidan Peppin、Maria K Wolters 和 Alexa Hagerty。COVID-19医疗保健时代的"人工智能"是否意味着加剧不平等？*BMJ*，第 n304 页，2021年3月。doi：10.1136/bmj.n304。URL https://doi.org/10.1136/bmj.n304。

**74.** Segun Fatumo、Tinashe Chikowore、Ananyo Choudhury、Muhammad Ayub、Alicia R. Martin 和 Karoline Kuchenbaecker。增加基因组研究多样性的路线图。*Nature Medicine*，28(2):243–250，2022年2月。doi：10.1038/ s41591-021-01672-4。URLhttps://doi.org/10.1038/s41591-021-01672-4。

**75.** Richard J. Abdill、Elizabeth M. Adamowicz 和 Ran Blekhman。公开的人类微生物组数据主要来自高度发达国家。*PLOS Biology*，20(2):e3001536，2022年2月。doi：10.1371/journal.pbio.3001536。URLhttps://doi.org/10.1371/journal.pbio.3001536。

**76.** TimnitGebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daum´e III，以及 Kate Crawford。《数据集说明书》。*Communications of theACM*，64(12):86–92，2021年12月。doi: 10.1145/3458723。网址https://doi.org/10.1145/3458723。

**77.** Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi，以及 Katherine Heller。《Healthsheet：为健康数据集开发透明度工具》。2022年。doi: 10.48550/ARXIV.2202.13028。网址https://arxiv.org/abs/2202.13028。

**78.** Nicholas J Tierney, Fiona A Harden, Maurice J Harden，以及 Kerrie L Mengersen。使用决策树理解缺失数据中的结构。*BMJ Open*，5(6)，2015年。ISSN 2044-6055。doi: 10.1136/bmjopen-2014-007450。网址 https://bmjopen.bmj.com/content/5/6/e007450。

**79.** Mart´ınez-Plumed Fernando, Ferri C`esar, Nieves David, and Hern´andez-Orallo Jos´e. 缺失缺失值：机器学习公平性中的丑小鸭。*InternationalJournalof Intelligent Systems*，36(7):3217–3258，2021年3月。doi: 10.1002/int.22415。URLhttps://doi.org/10.1002/int.22415。

**80.** Alicia R. Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. 当前多基因风险评分的临床应用可能会加剧健康差异。*Nature Genetics*，51(4):584–591，2019年3月。doi: 10.1038/s41588-019-0379-x。URLhttps://doi.org/10.1038/s41588-019-0379-x。

**81.** Aayushi Bansal、Rewa Sharma 和 Mamta Kathuria。关于深度学习中数据稀缺问题及其解决方案与应用的系统综述。*ACM Computing Surveys*，2021年。

**82.** Travers Ching 等。深度学习在生物学和医学领域的机遇与障碍。*Journal of Royal SocietyInterface*，15(141)，2018年。

**83.** Weixin Liang、Girmaw Abebe Tadesse、Daniel Ho、L Fei-Fei、Matei Zaharia、Ce Zhang 和 James Zou。创建可信人工智能数据的进展、挑战与机遇。*Nature Machine Intelligence*，4(8):669–677，2022年。

**84.** Bernard Koch、Emily Denton、Alex Hanna 和 Jacob G. Foster。精简、再利用与回收：机器学习研究中数据集的生命周期。 *arXiv:2112.01716*，2021。

85. James M Heather 和 Benjamin Chain。测序仪的演进历程：DNA测序史。《基因组学》，107(1): 1–8，2016。86. Peng Li、Xi Rao、Jennifer Blase、Yue Zhang、Xu Chu 和 Ce Zhang。CleanML：一项评估数据清洗对机器学习分类任务影响的研究。载于《2021年IEEE第37届数据工程国际会议（ICDE）》，第13–24页。IEEE，2021。87. Sanjay Krishnan、Jiannan Wang、Eugene Wu、Michael J Franklin 和 Ken Goldberg。ActiveClean：用于统计建模的交互式数据清洗。《VLDB会刊》，9(12):948–959，2016。88. L. Zhang、M. Yang 和 X. Feng。稀疏表示还是协同表示：哪种方法有助于人脸识别？IEEE国际计算机视觉会议（ICCV），第291–324页，2011。89. T. Chakraborti、B. McCane、S. Mills 和 U. Pal。图像块协同表示的广义公式（GP-CRC）。英国机器视觉会议（BMVC）论文集，2017。90. J. Ben Schafer、Dan Frankowski、Jon Herlocker 和 Shilad Sen。协同过滤推荐系统。《自适应网络》，《计算机科学讲义》，第291–324页，2007。91. T. Chakraborti、B. McCane、S. Mills 和 U. Pal。基于协同表示的细粒度物种识别。IEEE国际新西兰图像与视觉计算会议（IVCNZ）论文集，2016。92. William E Vinje 和 Jack L Gallant。自然视觉过程中初级视觉皮层的稀疏编码与去相关。《科学》，287(5456):1273–1276，2000。93. Trivellore E Raghunathan。合成数据。《统计学及其应用年度评论》，8:129–140，2021。94. James Jordon、Lukasz Szpruch、Florimond Houssiau、Mirko Bottarelli、Giovanni Cherubin、Carsten Maple、Samuel N Cohen 和 Adrian Weller。合成数据——是什么、为什么以及如何生成？arXiv预印本 arXiv:2205.03257，2022。95. Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力即一切。《神经信息处理系统进展》，30，2017。96. Han Zhang、Ian Goodfellow、Dimitris Metaxas 和 Augustus Odena。自注意力生成对抗网络。arXiv:1805.08318，2018。97. Neal Mangaokar、Jiameng Pu、Parantapa Bhattacharya、Chandan K. Reddy 和 Bimal Viswanath。Jekyll：利用深度生成模型攻击医学图像诊断。arXiv:2104.02107，2021。98. Yin Xu、Yan Li 和 Byeong-Seok Shin。利用上下文风格迁移进行医学图像处理。《以人为中心的计算与信息科学》，10(46)，2020。99. M. Runz、D. Rusche、S. Schmidt、M.R. Weihrauch、J. Hesser 和 C-A Weis。使用循环一致生成对抗网络对苏木精-伊红染色的组织学图像进行归一化。《诊断病理学》，16(71)，2021。00. Jinsung Yoon、James Jordon 和 Mihaela Schaar。GAIN：使用生成对抗网络进行缺失数据插补。载于《国际机器学习会议》，第5689–5698页。PMLR，2018。

**1**