

## 进入黑盒子:机器学习能为环境健康研究提供什么?

Charles W. Schmidt

<https://doi.org/10.1289/EHP5878-zh>

在 1956 年的一次会议上,来自加利福尼亚圣塔莫尼卡市(Santa Monica)兰德公司(RAND Corporation)的科学家们公布了被称为第一个人工智能(artificial intelligence, AI)的项目。一个被叫做逻辑理论机器(Logic Theory Machine)<sup>1</sup>的智能装置能通过模仿人类解决问题的能力来证明复杂的数学定理。如同其创建者所描述的那样,这个“思考机器”当时并未引起人们的关注。如今, AI 已完全不可同日而语了: AI 软件嵌入在我们日常使用的许多电子设备中, 2019 年, 全球用于该项技术上的花费接近 360 亿美元—比前一年增加了 44%。<sup>2</sup>

什么是 AI? 这不好说, 因为这个术语缺乏统一的定义。“我们很难分析和衡量人类的智能, 因为它是我们脑子中感受的事情,” 北卡罗莱纳州研究三角园的 RTI International 的高级人工智能研究员 Sam Adams 说。“那么我们又怎么会知道这个智能是不是人工的呢?”

位于费城的宾夕法尼亚大学(University of Pennsylvania)佩雷尔曼医学院(Perelman School of Medicine)信息学教授 Jason Moore 将 AI 描述为构建像人类一样解决问题和推理的软件及计算机的一门科学。他举了自动驾驶汽车的例子, 自动驾驶汽车必须识别道路上的行人和车辆, 阅读路牌并能作出瞬间的反应避免撞车。AI 技术还可以增强人类的智能, 因为它可以让科学家在庞大的数据集中识别出他们自己无法探测到的重要关联。Moore 说, 科学家们提出了一个新的术语“增强智能”来描述这种能力。

如今, AI 正在成为环境健康领域中一个强大的研究工具。<sup>3,4</sup> “我认为它是环境健康科学创新的催化剂, 在如何最好地利用大型复杂的数据这方面, 它可以帮助我们解决许多尚未解决的挑战,” 美国国立环境健康科学研究所(NIEHS)的代理所长 Rick Woychik 说道。“理想情况下, AI 可以帮助我们提出新的假设或者为棘手的问题提供有效的解决方案。”

环境卫生的科学家们已经在使用 AI 来搜索有用信息的文献, 对细胞和组织中污染物的影响进行建模,<sup>5</sup> 并根据遥感数据对空气质量进行评估。<sup>6,7</sup> 根据美国国家毒理学计划(National Toxicology Program, NTP)替代毒理学方法评估跨部门中心的代理主任 Nicole Kleinstreuer 介绍, AI 可能最终在细胞蛋白质合成器的转录组研究和评估“暴露组”或个体一生的化学暴露中发挥重要作用。

尽管如此, 专家指出如果 AI 的使用不当也会产生误导性结果。AI 算法很难被训练, 其中很多都是“黑匣子”——

这意味着其内部计算要么是专有的信息, 要么由于过于复杂人们难以理解。<sup>8</sup> 科学家们或许有理由怀疑, 当黑匣子处理真实世界的的数据时, 它是否会如预期那样运行, 是否会因混淆信号影响其预测。

德州农工大学(Texas A&M University)的毒理学教授 Ivan Rusyn 慎重地表示, 一些科学家可能会夸大了这项技术, 吹嘘 AI 对医学和环境卫生难题的解决方案“就在眼前且触手可及”来误导公众。

Moore 对此表示赞同, 并补充道, AI 在环境健康方面的应用应该缓慢而稳步推进。“在科学家们寻求解决每个问题的正确方法时, 我们需要保持热情的同时降低期望,” 他说。

### 机器学习教程

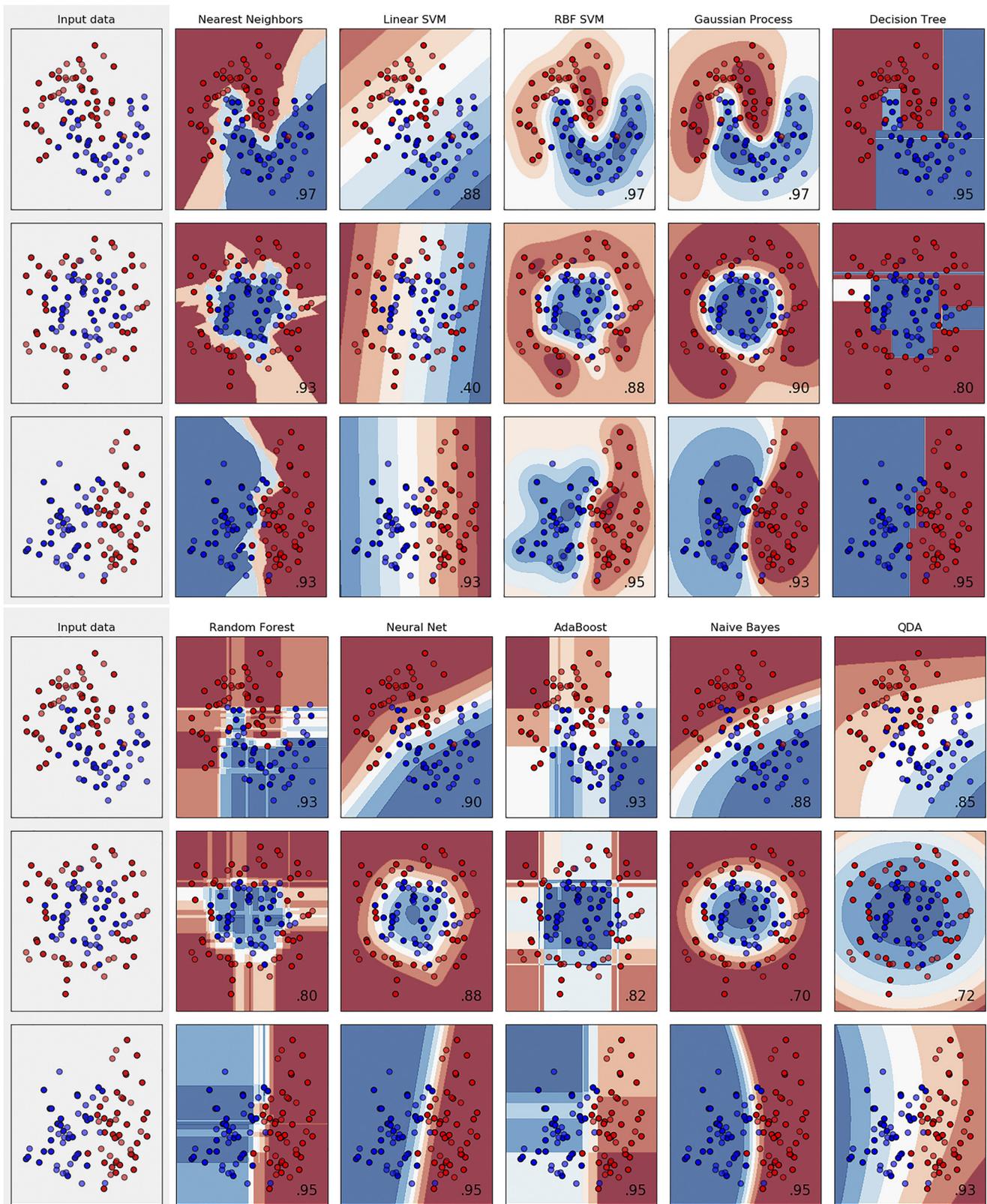
AI 背后的驱动力是机器学习, 这是指计算机算法如何随着经验的增加改进对执行指定任务的能力。<sup>9</sup> 其中一种方法是学习识别数据中的模式。模式识别的训练可以是有监督的(由人指导), 也可以是无监督的, 这就意味着算法随意根据数据自行去识别模式。

监督算法首先必须训练它们如何识别标签数据集, 例如, 一张数码照片中一只猫、一个 DNA 序列中的一个基因, 或者某个社区中可能的房价。根据数据的基本性质可将算法的预测分为两类: 一类是离散类(如“猫”或者“基因”), 另一类是回归类(如“价格”), 预测描述的是连续变量的测量值。

无监督算法是在没有任何指导的情况下自行组织数据。例如, 使用一种称为聚类分析的常用技术, 这些算法会自动将具有相似特征的数据分组。由于科学家们可能事先不知道要寻找这些数据组群, 所以聚类分析可能会牵出新的和预料之外的发现。

机器学习的一个更强大的亚类叫做深度学习, 它依赖于分层排列的算法去模仿人类大脑的结构。<sup>10</sup> 例如, 卷积神经网络(convolutional neural networks, CNNs)是一种灵感来自于人类视觉系统排布和功能的深度学习模型。CNNs 是当今大多数计算机视觉应用的核心, 例如 Facebook 的照片自动标签系统或遥感数据的判读。

还有许多其他类型的深度学习模型。循环神经网络(recurrent neural network)是一种特别擅长于在时间序列数据中发现模式的模型, 这意味着数据组会随着时间的变化而变化(比如股票市场价格或一天中臭氧浓度的波动)。另一种深度学习模型被叫做自动编码器(autoencoder), 用于无监督机器学习, 并且能应用于从极为有限的键信息组中重建完整



这个图展示了三组数据(行), 其中包含两个变量 ( $x$ -和  $y$ -轴)以及两个结果(蓝色或红色)。在 10 种机器学习方法(列)中, 每一个都试图通过构建  $x$  和  $y$  变量的数学函数来将圆点分类为蓝色或红色。颜色的深浅反映了模型把每个点分类为蓝色或红色的置信度。这些数字反映了分类的精确性, 或由模型正确分配的蓝色或红色结果的比例。每种方法检测不同的模式并对每个数据集执行不同的操作。机器学习的挑战之一就是弄清楚哪种方法才是对一组特定数据的最佳选择。Image: 3-Clause BSD License. Figure created using the Scikit-learn library. <sup>18</sup>

的数字图像和其他数据展示技术。在某些情况下，它们被用来过滤掉无关的“噪点”，这对锐化数字图像很有用。

为研究课题选择合适的模型至关重要，尽管要选择哪一个并不总是显而易见。“我遇到的一个最常见的问题是，‘我应该用什么样的模型来处理我的数据？’”哥伦比亚大学梅尔曼公共卫生学院 (Mailman School of Public Health) 的助理教授 Marianthi-Anna Kioumourtzoglou 说，她在化学混合物的健康研究中使用了AI。她说，答案是研究人员应该从清晰地构建他们想要回答的问题开始。

杜克大学土木与环境工程专业的助理教授 David Carlson 表示，需要避免的风险是“过度拟合”，即一种选择不当的模型倾向于捕捉数据中的噪点而不是真实信息。在这些情况下，模型将产生不可靠的预测，他解释说，精心选择的模型将是可推广的。换句话说，一个精心挑选的模型将能够很好地适应它以前从未见过的新数据。科学家可以应用一些统计测试来验证他们的模型，这样他们就可以对模型的普遍适用性更有信心。

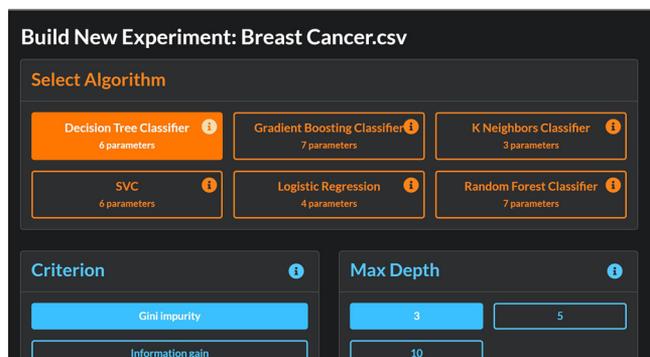
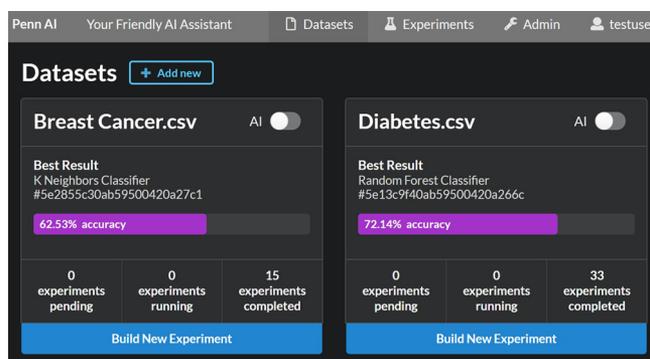
尽管模型选择涉及到统计和计算机科学方面的专业技能，但一些研究人员也正在转向日益增多的，能自动将模型与他们的数据进行匹配的开源软件包。在 2019 年的一场主题为 AI 促进环境健康的会议上，<sup>11</sup> 宾夕法尼亚大学的 Moore 介绍了一种他自己研究团队开发的，名为 PennAI 的软件包。“你只需加载你的数据集，按一下按钮，人工智能就会接管并启动它认为最佳的运行模式，”他说。据 Moore 介绍，PennAI 之所以能够做到这一点是因为它创建的这个知识库能够判断哪个模型能处理哪种类型的数据，类似于亚马逊等商业实体根据你的购物历史推荐你可能想买商品的系统。

换句话说，Carlson 解释道，PennAI 和许多其他软件包的目的是使 AI 工具更容易被广大用户所接受。但是，他补充说，尽管这样的工具比以往任何时候都更容易获得，“我个人认为这种[一劳永逸]的系统现在还不存在，因为你需要大量的专业知识和理解来使用和正确地解释系统所输出的信息。”

## 当今环境卫生领域的人工智能景象

随着 AI 进入环境健康研究领域，该技术近期的机遇在好几个方面得到体现。文本分析(也称为文本挖掘)使用机器学习算法从论文和报告中提取有用信息。这是“我们感兴趣的一个很大的领域，”美国环境保护署 (EPA) 科学与信息管理办公室主任 Jerry Blancato 说。Blancato 介绍道，理想情况下文本分析将提供更好的方法来管理、查询和分类不同来源的数据。

据英国兰卡斯特大学 (Lancaster University) 和美国循证毒理学协作 (Evidence-Based Toxicology Collaboration) 的研究人员 Paul Whaley 介绍，文本分析的进步将在很大程度



开源软件 PennAI 旨在简化用户的机器学习。在运行界面上(上图)，用户选择可用的数据组进行分析。“最佳结果”框内显示了在每个数据组上哪个算法执行得最准确。用户还可以通过点击“实验完成”框来浏览每个数据集的所有结果。从这里，用户可以切换到“AI”选项，让软件自动选择合适的机器学习算法和参数。或者在“新建实验”界面(下图)，用户可以手动选择算法和参数设置。Image: Courtesy Jason Moore.

上提高了系统综述的效率，这是一个高度系统化过程，科学家能从多个来源收集有助于回答某些问题的证据。就目前的情况，系统综述很大程度上依赖于研究助理，他们必须阅读数百甚至数千份文献。Whaley 说 EPA 和 NIEHS 都已经开始使用机器学习算法自动化完成初步筛选，根据标题或摘要中的关键词对文献进行分类。

更复杂的文本分析可能最终会允许算法去阅读和理解整个句子，尽管这些程序对语言还没有丰富和细腻的理解力。“这正是我们真正在寻找的能力，”Whaley 说。“分类是很有用的，但更重要的是我们需要能够通读报告并为我们提取相关信息的机器学习系统。这样，相比从 25 份报告中手工提取数据，你可以从成千上万个可能有用的文档中自动提取比手工收集容量更大、更丰富的数据集。”

Whaley 补充道，朝这个方向迈出的重要一步将是建立一个注释研究的“全文语料库”，以用来训练算法更有效地阅读技术语言。全文语料库是一组重要信息被手工强调或标识的文档。据 Whaley 所说，在这样一个知识库上训练过的算法，日后应用于其他文档中时，它将学习去识别和提取类似的信息。

NTP 的研究人员正在使用类似的方法着眼于开发能预测化学毒性的计算机化系统。为此，Kleinstreuer 的团队和

橡树岭国家实验室 (Oak Ridge National Laboratory) 的研究人员正在联合开发一种算法, 开发的第一步是系统将识别出毒理学文献中的高质量论文。在这个初始过程中, 审阅人必须阅读论文, 然后提取诸如, 方案、测试的化学物质类型和观察到的效果等信息。其目的是把这些论文中的信息作为数据库的原始资料, 这些数据库将化学结构与毒性终点如: 死亡率、内分泌紊乱和蛋白质反应活性等关联起来。反过来, 这些数据库可以被其他研究化学安全的团队用来训练机器学习模型。

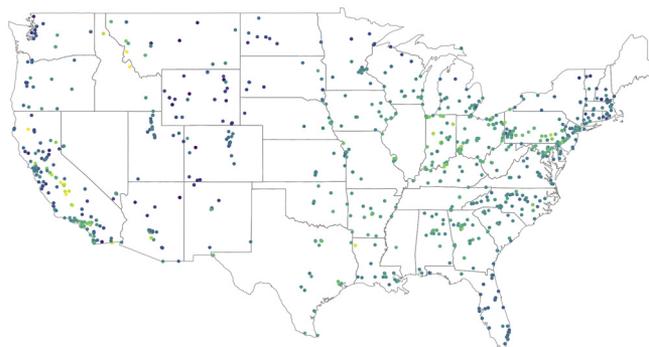
组建数据库需要 NTP 的研究人员将公布的信息转换成计算机可读的格式, 以便计算机算法进行处理。“我们现在所做的很多事情对数字化研究来说都是计算机无法识别的蛮力管理,” Kleinstreuer 说。她补充道, NTP 的研究人员最近建立了一个与大约 1.5 万个化学结构相关的啮齿动物 LD<sub>50</sub> 值(即杀死一组 50% 化学暴露动物的剂量)的数据库。Kleinstreuer 说, 随着模型开发的发展, 整个过程——从选择论文, 到管理数据库, 到最终开发出预测未经测试的化学品的毒性的算法——都可以在 AI 的帮助下及时完成。

将机器学习应用于基于现场和卫星的遥感数据是另一个新兴的发展。在 EPA, 科学家们正在利用这项技术来绘制洪泛区和蚊子栖息地的地图并开发预测模型, 以便预警有毒藻类爆发。在别的地方, 其他研究人员用它来评估空气污染水平。麦吉尔大学 (McGill University) 的流行病学专家 Scott Weichenthal 就是其中的一位科学家。在最近的一个研究项目中, Weichenthal 的团队发现, 当应用于卫星图像时, CNNs 预测细颗粒物浓度 (PM<sub>2.5</sub>) 的准确性几乎与世界卫生组织 (WHO) 全球疾病负担研究中所用的空气质量评估模型相同。<sup>12</sup>

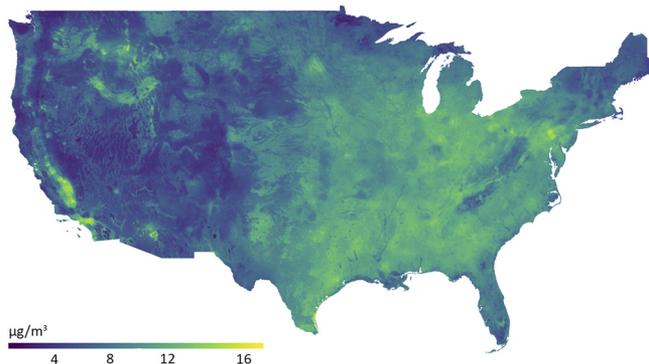
WHO 的模型称为空气质量数据集成模型 (Data Integration Model for Air Quality), 它依赖于许多不同的输入, 比如从地面传感器收集到的化学传输特征和污染测量数据。Weichenthal 和他的同事通过将来自 98 个国家约 6000 个地点的地面传感器数据与每个传感器位置对应的卫星数据进行配对来训练他们的模型。一旦训练完成, 这个模型就可以仅根据地面特征来预测 PM<sub>2.5</sub> 的变化, “你需要的只是卫星图片,” Weichenthal 说。

基于这种方法, 哈佛大学生物统计学家、数据科学计划 (Data Science Initiative) 联合主任 Francesca Dominici 将机器学习估算出空气中 PM<sub>2.5</sub> 的浓度值与美国老年人死亡率的变化相关联。<sup>13</sup> 为了实现这一目标, 她和同事们采用了一个模型<sup>14</sup>, 该模型结合了基于地面和卫星的测量方法并将机器学习算法应用到数据中, 以估算全美国平方公里范围内的污染水平。他们将预测值与 2000 年到 2012 年间每个地区所收集的数百万医疗保险支付的数据进行配对。他们的分析表明, 当空气中 PM<sub>2.5</sub> 增加 10 μg/m<sup>3</sup>, 臭氧增加 10 ppb, 分别与总死亡率增加 7.3% 和 1.1% 相关。<sup>13</sup>

Average annual PM<sub>2.5</sub> measured by Air Quality Service monitors (2012)



Estimated annual average PM<sub>2.5</sub> (2012)



包括 Frederica Dominici 在内的一些研究人员开发了一个机器学习模型来预测美国各地的 PM<sub>2.5</sub> 浓度。该模型整合了遥感数据、地面 PM<sub>2.5</sub> 估算值和大气中气溶胶总量、气象数据、土地利用数据等。训练组(上图)是基于美国环境保护署空气质量系统的监测数据。该模型生成的图像(下图)与地面真实数据非常接近, 而且提供了更精细的空间尺度。Image: Courtesy Benjamin M. Sabath.

## 可靠性问题

尽管如此, Dominici 将建模的 PM<sub>2.5</sub> 预测描述为猜测, 并补充说, “我们还不能量化来自机器学习的猜测有多准。”正如她所说, 当黑匣子的预测产生了“我们在评估健康影响时不能忽视的不确定性时”, 这一点至关重要。

Weichenthal 同意这项技术并非没有缺点。他承认, 他工作中的估算在模型预先训练之外的区域越来越不可靠。此外, 鉴于模型的内部计算有些不透明, 驱动其预测所构建环境的具体特征并不为人所知。

一个预测工作特别糟糕的情况发生在 2018 年加利福尼亚森林大火期间, 谷歌使用了另一家公司专有的黑匣子机器学习算法来支持其搜索页面天气微件 (widget)。该微件称空气污染水平是安全的,<sup>15</sup> 尽管当地人们眼睁睁地看着灰尘正在往他们的汽车上不断堆积。<sup>8</sup> 据 Carlson 说, 计算机科学家目前正在试验各种方法去打开深层神经网络和其他黑匣子, 以揭示它们内部计算或生成具有同等精度的可解释模型。

同时, 任何模型的准确性在很大程度上取决于它所接触数据的数量和质量, 以及训练数据和真实数据之间的差

异。Carlson 在 2019 年写道, 这些改变准确性的差异“可能会给机器学习方法带来重大问题。”<sup>8</sup> Carlson 声称“修改单个像素可以完全改变算法对一幅图像的理解”贴在停车标志上的小贴花纸“甚至可以骗过自动驾驶汽车上的现代工业计算机视觉系统。”<sup>8</sup>

现在环境卫生领域中使用机器学习算法的一个优先事项是确保其能够充分取得高质量的数据。“如果不注重数据质量, AI 将无从下手,” Woychik 补充说, NIEHS 高度重视开发可持续性系统以生成可与世界各地的研究人员轻松共享的数据。他表示, 实现这一目标的基础是数据生产必须遵守 2016 年首次发布的 FAIR 指导原则 (FAIR Guiding Principles)。<sup>16</sup> 这些原则规定, 数据和相关的数据对象如代码应该是可查找、可获得、可互操作, 并且可以被人和机器重复使用。

为此, NIEHS 目前正在对其网络基础设施进行全面翻新, 以更好地为 AI 的使用做好准备。研究所招聘了新的工作人员, 负责制定网络基础架构管理计划, 包括更好地收集、注释和存档数据以供现在和将来使用。<sup>17</sup>一旦那些系统就位, “我们可以考虑用 AI 做更复杂的实验,” Woychik 说, “但目前很多事情还只是在设想阶段, 我们不能过分乐观。”

同样地, EPA 的官员最近成立了一个正式的指导委员会, 已成为那些对 AI 感兴趣、希望提供培训、建议或咨询的人的聚集地。“我们有许多具有深厚专业知识的人, 我们希望分享资源并建立合作关系,” EPA 的 Blancato 说。

RTI 的 Adams 也认为, 目前环境健康的重点仍然是为机器学习算法准备数据。“Facebook 和其他公司在这方面做得很成功是因为他们处理的是 TB 级别的数据,” 他说。“我们这些从事科学研究的人仍在投入资源, 给数据贴上标签让人们可以使用。我们能用这项技术做什么(取决于)我们如何整合收集到的数据。”

**Charles W. Schmidt**, 理学硕士, 居住在缅因州波特兰市的获奖记者。他的作品曾发表在《科学美国人》*Scientific American*、《自然》*Nature*、《科学》*Science*、《发现杂志》*Discover Magazine*、*Undark*、《华盛顿邮报》*Washington Post* 以及许多其他出版物上。

## References

1. Newell A, Simon HA. 1956. *The Logic Theory Machine: A Complex Information Processing System. P-868*. Santa Monica, CA: The RAND Corporation.
2. Shirer M, D'Aquila M. 2019. Worldwide spending on artificial intelligence systems will grow to nearly \$35.8 billion in 2019, according to new IDC spending guide. [Press release.] International Data Corporation, 11 March 2019. <https://www.idc.com/getdoc.jsp?containerId=prUS44911419> [accessed 3 February 2020].
3. Research Triangle Environmental Health Collaborative. 2019. *11th Environmental Health Summit, Artificial Intelligence in Environmental Health Science and Decision-Making*. 18–19 October 2018. Research Triangle Park, NC: North Carolina Biotechnology Center.
4. Miller TH, Gallidabino MD, MacRae JI, Hogstrand C, Bury NR, Barron LP, et al. 2018. Machine learning for environmental toxicology: a call for integration and innovation. *Environ Sci Technol* 52(22):12953–12955, PMID: 30338686, <https://doi.org/10.1021/acs.est.8b05382>.
5. Luechtefeld T, Marsh D, Rowlands C, Hartung T. 2018. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. *Toxicol Sci* 165(1):198–212, PMID: 30007363, <https://doi.org/10.1093/toxsci/kfy152>.
6. Hong KY, Pinheiro PO, Minet L, Hatzopoulou M, Weichenthal S. 2019. Extending the spatial scale of land use regression models for ambient ultra-fine particles using satellite images and deep convolutional neural networks. *Environ Res* 176:108513, PMID: 31185385, <https://doi.org/10.1016/j.envres.2019.05.044>.
7. Weichenthal S, Hatzopoulou M, Brauer M. 2019. A picture tells a thousand...exposures: opportunities and challenges of deep learning image analyses in exposure science and environmental epidemiology. *Environ Int* 122:3–10, PMID: 30473381, <https://doi.org/10.1016/j.envint.2018.11.042>.
8. Rudin C, Carlson D. 2019. The secrets of machine learning: ten things you wish you had known earlier to be more effective at data analysis. In: *Operations Research & Management Science in the Age of Analytics*. Netessine S, ed. Catonsville, MD: The Institute for Operations Research and the Management Sciences, 44–72, <https://doi.org/10.1287/educ.2019.0200>.
9. Meserole C. 2014. What Is Machine Learning? Brookings Institute. <https://www.brookings.edu/research/what-is-machine-learning/> [accessed 3 February 2020].
10. LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521(7553):436–444, PMID: 26017442, <https://doi.org/10.1038/nature14539>.
11. National Academies of Sciences, Engineering, and Medicine. 2019. *Leveraging Artificial Intelligence and Machine Learning to Advance Environmental Health Research and Decisions: Proceedings of a Workshop—in Brief*. Washington, DC: National Academies Press. <https://www.nap.edu/catalog/25520/leveraging-artificial-intelligence-and-machine-learning-to-advance-environmental-health-research-and-decisions> [accessed 3 February 2020].
12. Hong KY, Pinheiro PO, Weichenthal S. 2019. Predicting Global Variations in Outdoor PM<sub>2.5</sub> Concentrations Using Satellite Images and Deep Convolutional Neural Networks. <https://arxiv.org/abs/1906.03975> [accessed 3 February 2020].
13. Di Q, Wang Y, Zanobetti A, Wang Y, Koutrakis P, Choirat C, et al. 2017. Air pollution and mortality in the Medicare population. *N Engl J Med* 376(26):2513–2522, PMID: 28657878, <https://doi.org/10.1056/NEJMoa1702747>.
14. Di Q, Kloog I, Koutrakis P, Lyapustin A, Wang Y, Schwartz J. 2016. Assessing PM<sub>2.5</sub> exposures with high spatiotemporal resolution across the continental United States. *Environ Sci Technol* 50(9):4712–4721, PMID: 27023334, <https://doi.org/10.1021/acs.est.5b06121>.
15. McGough M. 2018. How bad is Sacramento's air, exactly? Google results appear at odds with reality, some say. *Sacramento Bee*, California section, online edition. 7 August 2018. <https://www.sacbee.com/news/california/fires/article216227775.html> [accessed 10 January 2020].
16. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A. 2019. Addendum: the FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 6(1):6, PMID: 30890711, <https://doi.org/10.1038/s41597-019-0009-6>.
17. National Institute of Environmental Health Sciences. 2019. Informatics and Information Technology Strategic Roadmap, Fiscal Years 2019–2021. <https://www.niehs.nih.gov/about/informatics-it/index.cfm> [accessed 18 November 2019].
18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. 2011. Scikit-learn: machine learning in Python. *J Machine Learning Res* 12:2825–2830.