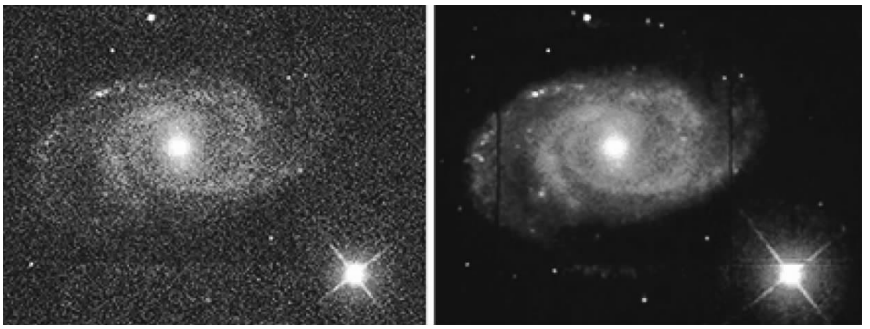
# AI在行动 :改变科学界的人工智能

胡德良 / 编译



人工智能“知道”星系应该是什么样的,它把一张模糊的图像(左)变成了一张清晰的图像(右)

● AI可以通过你的智能手机与你对话,可以在无人驾驶汽车里上路行驶,心怀疑虑的未来主义者担心AI的飞速发展将有可能导致大规模失业,但对于科学家来说,AI前景是光明的:它将加速科学发现的进程。

## 算法如何分析大众情绪

社交媒体拥有数以亿计的用户,他们每年发布数以千亿计的微博和帖子,这将社会科学推向了大数据时代。心理学家马丁·塞利格曼(Martin Seligman)认识到:社交媒体提供了一个前所未有的机会——利用人工智能从大众沟通中提取意义。在宾夕法尼亚大学积极心理学研究中心,塞利格曼和20多位心理学家、医生和计算机科学家投身于世界福祉项目,利用机器学习和自然语言处理技术大量筛选数据,以期衡量公众的情绪和身体健康状况。从传统上来讲,这些都是通过调查完成的。但塞利格曼说:“社交媒体数据并不引人瞩目,非常廉价,而且所获数据的数量级也更大。”社交媒体上的数据是凌乱的,但是人工智能可以提供一种揭示其中模式的高效方法。在最近的一项研究中,塞利格曼及同事研究了29 000名用户在脸谱网上更新的内容,他们对于是否患有抑郁症进行了自我评估。研究人员利用其中28 000名用户的数据资料,通过机器学习算法,发现用户用词和抑郁程度存在关联。这样一来,仅仅根据更新内容,这种算法就可以成功地判定出其余用户的抑郁程度。

在另一项研究中,研究小组分析了1.48亿篇微博以预测一个县城中的心脏病致死率。结果证明,跟愤怒和消极关系有关的话语成为危险指标。相较于根据吸烟、糖尿病等10个主要危险指标进行的预测,通过社交媒体进行的预测与实际死亡率更加接近。除此之外,研究人员还利用社交媒体来预测人的性格、收入和政治思想意识,并研究医院护理、神秘经历和刻板印象等情况。通过推特的数据,研究人员甚至创建了一张地图,用不同颜色描绘了美国每个县居民的幸福感、抑郁度、信任度和五种人格特质。德克萨斯大学奥斯汀分校的社会心理学家詹姆斯·潘尼贝克(James Pennebaker)说:“语言分析与心理学的联系正在发生一场革命。”潘尼贝克注重的不是内容,而是风格。例如,他发现:可以根据大学招生考试的文章中所使用的功能词来预测成绩。冠词和介词的使用体现了分析思维能力,可以预测其成绩会较高;代词和副词的使用体现了叙事思维能力,可以预测其成绩会较低。此外,潘尼贝克发现的证据表明:1728年的剧本《双重背叛》(*Double Falsehood*)中的大部分内容可能是由威廉·莎士比亚撰写的,根据认知复杂性和罕见词等因素,机器学习算法认定该剧本与莎士比亚的其他作品一致。潘尼贝克称:“现在,我们可以分析一个人曾经撰写和发布过的所有内容,并且逐渐分析你和他人的谈话方式。结果就是,我们可以越来越详细地描绘

出大家到底是什么样的人。”

## 通过基因组搜寻自闭症的根源

对于遗传学家来说，自闭症是一个棘手的问题，而遗传模式表明自闭症具有很强的遗传因素。但是，对自闭症有影响的数十种已知基因的变体，只能解释所有病例的大约 20%。要想找到可影响自闭症的其他变体，就需要在 25 000 个其他人类基因及其周围 DNA 的相关数据中寻找线索——这是人类研究者难以招架的任务。因此，普林斯顿大学的计算生物学家奥尔加·特洛伊安斯卡亚（Olga Troyanskaya）和纽约市西蒙斯基金会引入了人工智能工具。

纽约基因组中心创始人、洛克菲勒大学医学科学家罗伯特·达内尔（Robert Darnell）解释说：“我们所能做的，就是像生物学家一样揭示自闭症等疾病是由什么引起的。在某个方面，当一个科学家提出 10 个问题时，机器却有能力提出 1 万亿个问题，机器将改变整个游戏规则。”

特洛伊安斯卡亚整合了数百个数据集，这些数据集包含了哪些基因在特定的人类细胞中表现出活性、蛋白质之间是如何发生相互作用、转录因子结合位点以及其他关键基因组特征所在之处的描述。然后，特洛伊安斯卡亚及其研究小组利用机器学习创建了一份基因相互作用的图谱，并将少数已经熟知的自闭症风险基因与数千个其他未知基因进行对比，试图寻找其中的类似之处。2016 年，他们在《自然–神经科学》上发文说：对比表明，另外 2 500 个基因可能跟自闭症有关。

然而，就像最近遗传学家才认识到的那样——基因并不是单独起作用的，基因的行为是由附近数百万个非编码碱基决定的，而非编码碱基与 DNA 结合蛋白以及跟其他因子发生相互作用。与寻找这些基因相比，识别哪些非编码变体可能会影响到附近的自闭症基因是个更加困难的问题。在普林斯顿大学特洛伊安斯克亚的实验室里，有位名叫周健（Jian Zhou）的研究生正在利用人工智能解决这个问题。

为了训练这个深度学习系统，周健为这个系统引入了 DNA 元素百科全书和人类表观基因组学收集的数据，这两个项目记录了数万个非编码 DNA 位点是如何影响附近基因的。在评估非编码 DNA 未知片段上潜在的活动时，该系统实际上学会了如何确定应寻找的特征。2015 年 10 月，当周健和特洛伊安斯克亚在《自然–方法学》上描述他们这个被称为 DeepSEA 的计划时，加州大学尔湾分校的计算机科学家谢晓辉（Xiaohui Xie）将其称为“把深度学习应用到基因组学的里程碑”。目前，普林斯顿大学的研究团队正在利用 DeepSEA 计划对自闭症患者的基因组进行分析，希望能够对非编码碱基带来的影响进行排序。

同样，谢晓辉也正在利用人工智能处理基因组，但他关注的范围不仅仅是自闭症，他更希望根据突变的危害程度对其进行分类。但是，谢晓辉警告说：在基因组学中，深度学习系统只有在受过训练的领域才能像数据集那样有效。他说：“在我看来，当前人们会质疑这种系统在分析基因组方面的可靠性。但是将来越来越多的人会接受深度学习。”

## 能够理解太空的机器

2017 年 4 月，天体物理学家凯文·沙文斯基（Kevin Schawinski）在推特上发布了 4 个星系的模糊图片，并随图附上了一个请求，希望同行的天文学家能够帮助他进行分类。同行们意见一致：这些图像看起来像椭圆和螺旋结构，属于熟悉的星系类型。

但一些天文学家仍心存疑虑，因为沙文斯基是有头脑的，所以他们直截了当地问道：这些是真正的星系吗？它们是利用相关物理学原理在电脑上模拟出来的模拟星系吗？沙文斯基回答说：其实两者都不是。在瑞士苏黎世联邦理工学院，沙文斯基、计算机科学家张策（Ce Zhang）及其合作人员，在一个神经网络中模拟了这些星系，该神经网络对物理学原理一无所知，似乎只是帮助我们理解在深层次观察中，星系看起来应该是什么样的。

沙文斯基只不过是想利用在推特上发布的帖子来看看神经网络创造的星系在多大程度上是可信的，但是他更远大的目标是开创一项技术，能够像电影中那样将模糊的观测图像奇迹般地清晰化——神经网络能够使模糊的星系图像看起来更加清晰，就好像是用较为高级的望远镜拍摄的。这样，天文学家们就能够从大量的观察中得到更为精确的细节画面。沙文斯基说：“在巡天工程中，我们花费了数亿甚至是数十亿美元的资金。从某种程度上说，一旦有了

这项技术我们即刻就能提取更多的信息。”

沙文斯基在推特网上发布的星系图像是生成对抗性网络的作品，该网络是一种机器学习模型，包括两个互相对抗的神经网络，其中一个网络是可以创造图像的生成器，另一个是可以挑出瑕疵、去除伪造成分的鉴别器，它可以形成优化的生成器。沙文斯基的研究小组拍摄了数千张真实的星系图像，然后人为分解它们。接着，研究人员训练生成器拼接图像，使它们具有鉴别器的功能。最终，神经网络会胜过其他技术，成为消除星系图像杂乱成分最好的技术。

伊利诺伊州巴达维亚地区费米国家加速器实验室的天体物理学家布莱恩·诺德（Brian Nord）说：沙文斯基的方法是机器学习在天文学领域一个特别前卫的例子，但并非绝无仅有。2017 年 1 月，在美国天文学会的一次会议上，诺德提出了一种机器学习策略，用来搜寻强效引力透镜——遥远星系的图像在通往地球的过程中经过扭曲的时空时，会在太空中形成罕见的光弧。这些引力透镜可用于测量宇宙间天体的距离，并发现肉眼看不见的质量密度。

强效引力透镜的视觉效果十分独特，但难以利用简单的数学规则描述——利用传统的计算机很难分辨出来，但是对于人类来讲却相对容易分辨。诺德和其他科学家意识到，利用数千个透镜对神经网络进行培训之后，神经网络就可以获得类似人类的直觉。诺德称：“实际上，在接下来的几个月里，我们撰写了十几篇论文，都是研究应用机器学习搜寻强效引力透镜的，这是一件激动人心的事情。”

在整个天文学领域，这只是冰山一角。天文学家越来越认识到，人工智能提供了一个强有力的工具，可以利用 PB 级数据发现有趣的天体并对其进行分类。沙文斯基称：“人人都在惊呼：‘天哪，我们拥有的数据太多啦！’我认为，在大数据时代利用人工智能，终将会有真正的发现。”

## 神经网络学习化学合成的艺术

有机化学家善于反向思维。类似于烹饪大师先从看到成品菜肴开始，然后再研究如何烹制……许多化学家都是从他们想要制造的分子的最终结构开始思考如何组装它。德国明斯特大学的研究生马尔文·塞格勒（Marwin Segler）说：“要想知道如何合成分子，你所需要的只是合适的成分和方法。”目前，塞格勒和其他研究人员正在将人工智能引进他们的分子实验室。

研究人员希望人工智能能够帮助他们应对分子制造过程中的一个关键挑战：从数百个潜在的构建材料和数千个相关的化学规则中做出最合适的选择。几十年来，一些化学家利用已知的化学反应煞费苦心地为计算机编程，希望创建一个能够快速计算出最灵敏的分子合成法的系统。然而，塞格勒说：“化学可能是非常微妙的，很难以二进制的方式写下所有的规则。”

因此，塞格勒、明斯特大学计算机科学家迈克·普罗伊斯（Mike Preuss）和塞格勒的顾问马克·沃勒（Mark Waller）将目光转向了人工智能。他们没有利用化学反应的严格规则进行编程，而是设计了一个深度学习的神经网络程序。通过数以百万计的化学反应实例，该程序能够自行学习反应是如何进行的。塞格勒称：“你提供的数据越多，效果就越好。”随着时间的推移，这个神经网络学会了如何预测化学合成过程中目标分子的最佳反应。它从零开始，最终拿出了自己的分子制造方案。

这 3 位研究人员测试了 40 种不同的目标分子，并与传统的分子设计程序进行对比。根据 2017 年研究人员在一次会议上的报告，在两个小时的计算时间内，传统程序完成了 22.5%目标分子的合成方案；而人工智能程序则完成了 95%的合成方案。塞格勒不久将要前往伦敦的一家制药公司工作，他希望通过这个方法改进医药的生产过程。

加州帕洛阿尔托市斯坦福大学的有机化学家保罗·温德（Paul Wender）认为，现在判断塞格勒的方法是否有效，还为时尚早。然而，温德也正在将人工智能应用到有机化学合成，他认为：不仅在合成已知分子方面，而且在寻找制造新分子的方法方面，人工智能都可能会产生深远影响。塞格勒接着说，人工智能不会很快取代有机化学家，因为化学家们所做的远远不只是预测反应将会如何进行。就化学来说，人工智能就像 GPS 定位系统，它可能适于寻找合成的路线途径，但它本身却不能自行设计和执行整个合成过程。

当然，人工智能开发人员也已经着眼于完成其他任务了。

[资料来源：Science][责任编辑：松 石]