当数据如潮水般涌来，如何不被数据误导而迷失方向？

探寻深藏于数据海洋中的“因果关联”

图/视觉中国

■方陵生/编译

鲨鱼袭击事件和冰淇淋销量能联系到一起吗？它们之间只是简单的相关关系，还是有前因后果的因果关系？有时候，简单的数据统计和分析会得出令人瞠目结舌的结论。

为什么会这样？那是因为隐藏和纠缠在一大堆数据和事实中的因果关系是那么扑朔迷离，让人难辨真相。不过，数学工具也许可以为我们理清这些关系，从而让真相浮出水面。

不了解因果关系，很容易为数据偏离所左右

上世纪90年代中期，以病人住院资料作为数据基础的人工智能（AI）算法得出了一个令人惊讶的预测结果：患有哮喘症的肺炎患者生存率更高。

这是一个与所有医学知识相悖的结论。医学常识告诉我们，肺炎病人如果同时患有哮喘，死于肺炎的风险会增加。然而，多家医院收集到的数据为何会推导出截然相反的结论？这到底是怎么回事呢？

进一步调查发现，在AI预测中，算法疏漏了一个关键问题，那就是对于同时患有哮喘的肺炎患者，医生会更加重视，甚至会直接把他们送到重症监护室，那里的积极治疗和精心护理，大大降低了病人死于肺炎的风险。

这个案例说明，如果不了解因果关系，很容易为数据偏离所左右。而且仅靠数据分析，也很难可靠区分哪些属于因果关系，哪些只是偶然和巧合。所以，解开因果之间的真正联系，对于现代科学至关重要。从药物开发到基础设施设计，都离不开对因果关系的了解。

然而，一个多世纪以来，科学家们一直缺乏正确理清因果关系的工具。当人类社会进入大数据时代，面对海量涌来的数据，必须要找到可靠的工具来理清因果关系，才能在数据海洋中保持清晰的航向。

区分相关性与因果关系，亟需全新科学工具

“相关性并不一定就是因果关系”，这是大多数科学家哪怕在睡梦中也会奉为圭臬的一句至理箴言。给科学研究提供迫切需要的“因果关系语言”，数学可以成为解决这个难题的工具。

剖解因果关系的数学工具，在我们这个拥有丰富大数据的时代有着相当广泛的应用前景——了解事物的因果关系，将成为解决算法致命缺陷的必要工具。

道理人人都懂，但问题是，要理清数据中的因果关系，我们还需要数据以外的一些额外信息。鲨鱼伤人和冰淇淋销量之间的关系并不难理清，但如果涉及到一些更复杂或对其背景了解不多的数据，要区分两者之间是相关关系还是因果关系，就没那么容易。

事实上，了解因果关系对人类生活意义重大。例如，数据相关性研究可以告诉我们，哪种治疗方案可以让病人更快恢复，但却不能告诉我们这是为什么。而且，数据相关性也不能告诉我们如何更加有效地治愈病人，甚至不能成为给病人开处方的依据。

“若想治疗某种疾病，或者知道如何降低某种疾病的风险，就需要理解其中的因果关系。”丹麦哥本哈根大学的乔纳斯·彼得斯说。美国哥伦比亚大学的伊莱亚斯·巴伦布瓦姆则认为，科学研究和科学系统的运行都绕不开对因果关系的理解。

遗憾的是，可以用来理清因果关系的科学工具太少了。从伽利略时代开始，现代科学的研究工具之一是代数和等式。物理学家用等式来表达大气压和气压计读数之间的关系，但这样的等式说明不了它们之间的因果关系——是气压导致了气压计读数的变化，还是正好反过来呢？显然，代数语言不可能解决哪个是因哪个是果的问题。

创新数据“游戏规则”，为因果推理理论奠定基石

上世纪90年代初，美国加利福尼亚大学洛杉矶分校的朱迪亚·珀尔开始创建科学迫切需要的“因果关系语言”。

珀尔的解决方法是引进一种被称为“doing”（表示做、作为、动作、行动的意思）的数学语言。比如，如果通过“do”这个新引入的运算符采取某种干预“行动”，让气压计周围的大气压产生变化，那么气压计上的读数也会随之变动；但如果干预“行动”是移动气压计上的读数，显然大气压不会因此发生任何变化。所以，通过这样的数据变动，就能找出数据之间的因果关系——“因”变“果”也随之变，但无论“果”如何变，“因”却不会变。

如何用数学语言来表达这个概念呢？珀尔创建了一套包括加减和其他运算法则在内的运算方法。就像其他运算符一样，他的“do”运算符可以作为一种特殊变量加入到运算中。

再让我们回到海边场景。通过数学模型模拟，珀尔的“do”运算符改变了冰淇淋的消耗量，而不考虑其他任何对吃冰淇淋或被鲨鱼攻击产生影响的混杂因素。在实验中，如果只改变冰淇淋的消耗量，那么鲨鱼袭击频率如果有任何相应变化就应该是由吃冰淇淋引起的。

珀尔的实验表明，使用可观察到的数据，“do”运算符的变化可有效模拟随机控制实验，从而提取其中的因果关系。珀尔因这项研究获得了2011年图灵奖，他也由此奠定了因果推理理论的基石。

因果推理“工具包”，破解“结论不可重复”窘境

除了赋予科学以更坚实的因果推理基础之外，珀尔的数学框架还有助于解决许多学科问题，包括困扰医学和社会科学领域的“研究结论不可重复危机”。

过去十年，因为相关的实验结果无法复制，人们对一些领域中的研究产生了怀疑。比如，有研究认为，学生用模糊字体更容易解答出数学问题；还有研究提出，意志力是一种有限的、可耗尽的资源。事实上，心理学领域于2015年进行的一项关于实验结果复现性的大规模研究发现，该领域60%的研究成果无法复制，这给整个学科蒙上了巨大阴影。

巴伦布瓦姆认为，因果推理可以帮助解决这些问题。他说，在许多情况下，最初的测试结果容易受到多种混杂因素的影响，而这些因素可能是实验者没有意识到或被忽略的，而随后的复现性尝试可能会在混杂因素中发现新的因果关系。

一个典型例子是关于幸福感对经济决策的影响。最初，实验通过向参与者展示美国喜剧演员罗宾·威廉姆斯的镜头来衡量幸福感。可到了进行复现性实验时，威廉姆斯已经去世，同样的实验可能会对参与者的反应产生不同影响。另一个因素是，原始研究实验中的受试者为美国人，而复现研究中的受试者是英国人。由于这些混杂效应的影响，后来的复现实验显然无法对最初的研究发现作出合理评价。

因果推理理论的应用远远超越了科学的范畴。“如果你想要做出更好的决策，就要了解因果关系。也就是说，在做决策前要考虑一下，如果我这么做，会有什么后果，世界会发生什么变化。”美国约翰·霍普金斯大学的苏奇·萨里亚说。

相关性与因果关系有何不同

一些海边城市的数据告诉我们，哪天冰淇淋销售量多，海滨游泳者遭遇鲨鱼袭击的概率就高。那么，这是否意味着，出于公众安全考虑，应该取缔海边卖冰淇淋的小摊呢？人们大概率不会这么做。

因为常识和理智告诉我们，酷热天气会使海滨的人流量激增，这是一个明显的事实。人多，意外事件发生的概率也会更高。所以，冰淇淋销售量的增加与鲨鱼袭击频率增高的原因，很可能都是海滨游客增多，而冰淇淋销售量和鲨鱼袭击之间只存在相关性，并不存在因果关系。

图/视觉中国

因果关系工具包“求解”生活关切

区分相关性和因果关系是长期以来困扰人们的一个难题。而今，研究人员发现，解决这个难题也许并非不可能。

从论文数据中挖掘可靠结论

一些经济学家很早就认识到，对于他们想要解决的许多复杂的经济社会问题，都需要某种因果关系工具包，用以确定具体政策的效果。这些工具包更加高级、更加复杂。于是，科学家不断开发新的工具，以适应各种社会需求，尤其是来自医疗领域的需求。

例如，增加香烟税是否会减少吸烟对健康的影响？吸烟与健康之间的关系会受到一系列混杂因素的影响，包括年龄、性别、饮食、家族史、职业和受教育年限等。为弄清我们所关心的因果关系，我们只需要其他因素不变的部分数据。但每去除一种混杂变量，相应的数据集都会变小，最终只会剩下很少的数据，根本无法让我们据此得出可靠结论。

为克服这类困难，美国斯坦福大学的苏珊·埃塞和同事开发了一种技术。该技术与朱迪亚·珀尔的“do”运算符方法非常相近，同时又保留了尽可能多的数据。这样的工具在医疗保健领域将产生重大影响，甚至可能挽救生命。

美国约翰·霍普金斯大学的苏奇·萨里亚则利用因果推理理论来创建工具，通过比较不同医疗行为的效果来帮助医生做决定。然而，医学数据的处理是很复杂的。例如，获得医疗机会的不平等，就可能导致算法得出非常荒谬的结论，明明是有些低收入人群没钱看病，而算法可能得出“低收入人群更健康”这样错误的推论。

因此，此类推理要解决的问题关键，还是在于理清因果关系。

在医学上，为计算用某种药物治疗某种症状的有效用量，医生需要知道这种药物和那些症状之间是否存在因果关系，通常解决这一问题的方法是向有关专家了解。但以色列理工学院的基拉·拉丁斯基说，从专家那里获得关于因果关系的知识可能并不那么便捷，而且需要时间。为简化这个过程，她与合作者们采取了一种新方法，即在医学论文中发掘因果关系。这些已发表的医学论文，已经通过实验和同行评议证实了它们是有效的，因此只需将论文中所述的因果关系知识再利用，就可为治疗高血压、糖尿病等疾病找到新方法。

从数据观测中了解因果关系

发掘现有论文中的相关知识是一个富有成效和强大的方法，但不是每个领域都有大量证明因果关系的在线论文等待人们去发掘利用。因此，其他学科的一些研究人员在考虑，是否可以从单纯的数据观测中发现因果关系呢？

一种方法是寻找在任何情况下都适用的模式。比如，大气压力增加总会导致气压计读数发生变化，不管是在伦敦还是在纽约，在地球上还是在火星上。同样，不同医院或国家的医生治病救人的方式可能有所不同，但疾病和症状之间的潜在因果关系却是不变的。丹麦哥本哈根大学的乔纳斯·彼得斯等人认为，这些恒定的关系可以作为潜在因果关系的标记。

为检验这一原则，彼得斯和他的同事们开始探讨一个复杂的社会学问题：一个国家总生育率变化的真正原因。

生育率在世界各地差异悬殊，了解决定生育率升降的因果关系对想要增加本国人口的政府来说是一个福音。通过在多个国家的数据中寻找一致性模式，彼得斯和他的同事发现，幼儿死亡率是对生育率产生影响的一个重要因素，这一发现与世界各地之前的研究结果相符。

美国华盛顿大学社会学家兼统计学家阿德里安·拉弗瑞指出，当儿童死亡率较高时，很多家庭会倾向于生更多孩子，即使他们自己并没有孩子夭折。这是人们的一种前瞻性反应，以确保自身的血脉延续。

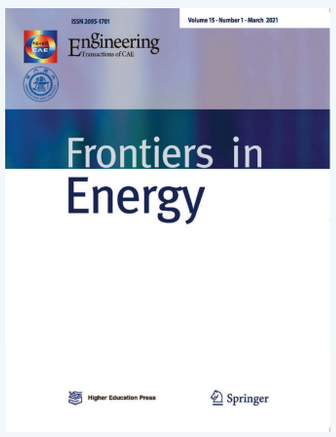
目前，彼得斯和他的合作者正利用这种不变原理来绘制生物圈和大气相互作用的因果图，这对我们理解气候变化可能将产生较大影响。

科学新知 上海期刊

“缺陷”光催化剂分解水更到位

《能源前沿》近期发表了一项来自德国汉堡大学刘军营博士、上海交通大学韦之栋博士和上官文峰教授的关于光催化分解水制氢的合作研究。光催化分解水制氢是解决不可再生能源短缺问题的有效途径，这一技术主要是通过光催化剂分解水，将太阳能转换为氢能。然而，目前光催化剂的效率受到了诸多因素（如光吸收、光生载流子的转移能力等）影响。因此，如何拓展光催化剂的光吸收范围、增强光生载流子的转移能力，进而提高其催化效率，就成为了该领域亟待解决的问题。

该研究通过在氮气氛围下，利用硼氢化钠热处理钛酸锶纳米晶，从而在钛酸锶纳米晶中引入了表面氧空位或Ti位点，使得钛酸锶纳米晶的光吸收范围获得拓展，电荷转移能力得以提升。与此同时，这些缺陷也改变了钛酸锶的氧化还原电势。在三者协同作用下，钛酸锶分解纯水时，氢气与氧气的产生比例也得到了调节。（刘瑞芹/整理）



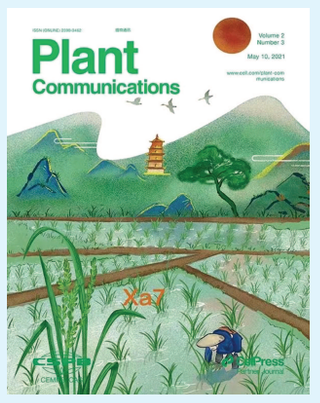
《能源前沿》由中国工程院、高等教育出版社和上海交通大学联合主办

最持久广谱水稻抗白叶枯病基因被克隆

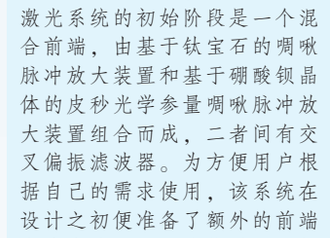
日前，浙江师范大学马伯军教授课题组与中国水稻所钱前院士团队联合在《植物通讯》上发表封面论文，宣布在国际上首次揭示了Xa7编码一个新型“执行者”抗病蛋白，解析了该基因兼具高抗、广谱、持久、耐热特性的新抗病分子机制，对于培育广谱、持久抗性的水稻新品种有重要指导意义。

由黄单胞菌水稻致病变种引起的白叶枯病是一种毁灭性细菌病害。因此，从水稻中克隆具有广谱且持久抗病特性的新基因，解析其抗病分子机制，具有重要的育种意义。

Xa7基因发现于孟加拉稻种DV85，是世界水稻抗白叶枯病育种研究中一个最具持久广谱抗性并被广泛应用的重要基因。该研究团队长期致力于水稻抗白叶枯病基因的研究，选用我国一个含Xa7基因的优质、高抗籼型强优恢复系品种镇恢084，历时5年从2万多个诱变株系中筛选到9个高感白叶枯病的株系，并通过大量的水稻转基因功能验证，最终成功克隆了Xa7基因。与此同时，该研究还揭示了一个重要的白叶枯病抗性Xa7基因及其独特的白叶枯病菌互作抗病分子机理，展示了Xa7重要的育种应用价值。（杨箫/整理）



《植物通讯》由中国科学院分子植物科学卓越创新中心与中国植物生理与植物分子生物学会主办



《高功率激光科学与工程》由中国科学院上海光学精密机械研究所主办、中国激光杂志社出版

双路10拍瓦飞秒激光系统创纪录

据《高功率激光科学与工程》近期报道，位于罗马尼亚的“极端光学装置－核物理”（ELI-NP）装置采用双路混合高能激光系统可产生2×10拍瓦的飞秒脉冲。拍瓦（PW）是功率单位，1拍瓦等于1000万亿瓦，相当于全球电网平均功率的500倍。如此高的功率，原本只在恒星内部或黑洞边缘等极端自然条件下才存在。而随着拍瓦激光的出现，研究人员可在实验室中复现这样的极端条件。

“极端光学装置－核物理”激光系统的初始阶段是一个混合前端，由基于钛宝石的啁啾脉冲放大装置和基于硼酸钡晶体的皮秒光学参量啁啾脉冲放大装置组合而成，二者间有交叉偏振滤波器。为方便用户根据自己的需求使用，该系统在设计之初便准备了额外的前端和多余的可用泵浦能量，可根据用户需求进行调整，从而显著提高了该系统的适用性。同时，每分钟一次的重复率是目前所知的10拍瓦激光系统中最高的，重复率的提升显著提高了系统的使用效率。研究人员认为，该系统高功率飞秒激光脉冲产生的极端场和压力条件，将推动基础和应用核物理领域的相关实验与研究。（都玮/整理）

《细胞研究》由中国科学院分子细胞科学卓越创新中心主办

新技术推进RNAi疗法临床应用

迄今为止，RNAi治疗方法的发展主要经历了两个阶段，即直接注射化学合成的siRNA，和通过人工载体体外包装后再进行递送。然而，这两代技术仍未完全解决siRNA体内递送，尤其是肝外递送的核心技术难题，因而现有的RNAi疗法无法完全发挥该技术的治疗价值，严重影响了小核酸药物的开发和临床应用。

日前，南京大学张辰宇团队在《细胞研究》杂志发表论文，报道了一种利用合成生物学理念设计的基因环路实现siRNA体内高效递送的新技术。利用该技术，科学家能将肝脏重编程为生物发生器，从而实现在体内合成siRNA，使其自组装进入外泌体，再分泌至循环系统，最后将siRNA递送至其他组织器官中，达到抑制靶基因表达的目的。

经过对基因环路进一步改造，科学家还能够实现siRNA对特定组织的靶向性递送。论文中提到，研究人员利用此方法对肺癌、胶质母细胞瘤、肥胖症等小鼠疾病模型进行了治疗，均产生了显著的治疗效果。（程磊/整理）