让科学数据不再沉寂

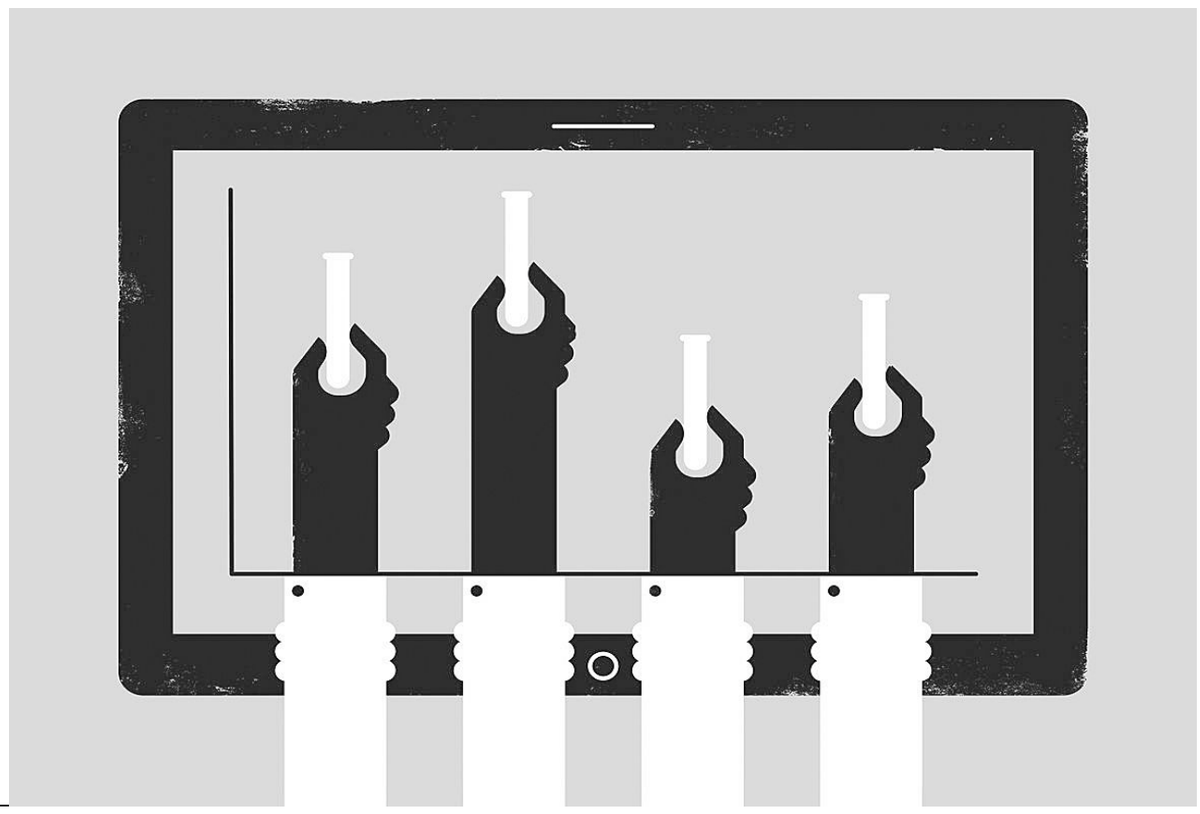
新可视化工具使在线发表更具交互性和再现性

当 Benjamin Delory 开始着手撰写关于记录一种量化植物形态新方法的论文时，他意识到其中的一批数据可能会带来问题。该论文提出一个“持久性的条形码”来描述植物根系的分支结构。其中的挑战是如何解释它。德国吕讷堡大学博士后 Delory 说，该条形码的基础算法是“连续和动态的”。而表示动态的最佳办法“是让它动起来”。

科学数据被认为是典型的静态图像。但这些静态图像却与基础数据相互分离，这会阻碍读者更详细地探索它们，例如放大一些感兴趣的特征。对于那些需要将数百万个数据点填入仅有几厘米大的密集视觉效果的基因组学家来说，这会特别棘手。

对于计算机运算领域的研究人员来说也是如此。科学家经常会把软件放到开源程序库如 GitHub 等网站，但让该代码正常运行却是“说起来容易做起来难”。评审人以及感兴趣的人经常需要另外的软件和配置才能让这些算法运行。

一些期刊和平台正在通过支撑交互性数据和代码弥补这一鸿沟。其中之一是 F1000Research（是针对生命科学研究者的开放研究发表平台），该平台去年曾与加拿大蒙特利尔计算机企业 Plotly 和美国纽约的一个机构代码海洋合作。正是因为这些功能以及 F1000Research 开放获取的思想，才让 Delory 及合作者把论文递交到那里。该成果已在 1 月发表。



Project Twins 制图

交互出版

让读者可以深入到一篇文章中的基本数据的交互式图表是很多网站上频繁出现的特征，比如《纽约时报》和 fivethirtyeight.com 等网站，但这类图表在科学出版中却不常见。

F1000Research 高级出版编辑 Thomas Ingraham 说，该期刊的“活数据”——2014 年引入的可持续用新数据升级的交互式图表不仅制作起来耗时耗力，而且不可伸缩。而 Plotly 则可让用户创建和共享从散点图和线图到等高线图和地图等可视化内容。其得到的图像可让用户放大数据、平移图像和移动鼠标查看所绘值。学生订阅费用从每年 59 美元起步。开源程序库可让研究人员创建从 R、MATLAB、Python 到 Julia 代码等免费 Plotly 图表。

代码海洋每月向学者免费开放 10 小时和 50G 字节的存储空间；付费类则从每月 19 美元起步。它把代码、数据、结果和计算环境融合在一起，该计算环境可在一个含有复制作者计算配置的“计算胶囊”中执行任务。其他用户则可从代码海洋网站或是论文中的一个部件来下载、修改和运行该代码。

F1000Research 现已经发表了 6 篇含有 Plotly“活图表”的论文以及含有代码海洋小部件的 5 篇论文。今年，该期刊计划增加对交互式“蛋白质—蛋白质相互作用”地图的支持，这些地图是利用网络制图工具 Cytoscape 生成的。

研究人员不必为感受到的复杂性困扰。据布鲁金斯南达科他州立大学计算生物学家 Xijin Ge 说，他在自己的一篇论文中就包含了交互式 Plotly 图表，创建相关数据仅需要一个额外代码行数。西澳大利亚大学海洋研究所和地球科学系珊瑚学者 Tom DeCarlo 已经为多个期刊创建了 6 个代码海洋项目，其中包括《古海洋学期刊》《古气候学期刊》和《生物地球科学杂志》。“我认为它对于科学交流和再现性非常重要。”他说。

开源方法

对于那些寻求开源计算替代方案的人来说，一个叫作 Binder 的工具可将任何包含 Jupyter 记事本（交错文档、代码和数据的文档）或 R 代码的公共 GitHub 存储库转换为一个包裹，从而可以让用户从其浏览器一端运行。用户只需在 mybinder.org 网站上把记事本存储库的地址输入到搜索栏中，该程序就能创建一个可共享的交互式工作区。圣路易斯奥比斯波加州州立理工大学 Binder 项目团队的 Carol Willing 说：“它真的适用于再现性，并且易于使用。”

瑞士苏黎世 Binder 项目团队成员 Tim Head 说，类似工具还可以简化同行评审。Head 有点沮丧，因为此前他受邀审阅一篇期刊文章时不能使用该软件。“如果他们当时给我发送了 Binder 的链接，那么我们现在已经完成了。”他说。

开源选择方案也可用于创建交互式图像，包括 bokeh、htmlwidgets、pygal 和 ipywidgets 等。这些大多数是以编程方式使用的，通常在 R 或 Python 代码中使用，这在科学中应用得很普遍。例如，程序员可以使用 ipywidgets 将交互的三维绘图、地图和分子可视化到 Jupyter 记事本中。另一个用 JavaScript 编写的选择是 Vega-Lite。由于该语言在科学上的使用度不那么广泛，加州州立理工大学的 Brian Granger 和西雅图华盛顿大学的 Jake VanderPlas 开发了一个叫作“Altair”的 Python 接口，使它变得更易访问。

这些工具中大多数都倾向于为特定的图表类型提供函数，Vega-Lite 和 Altair 都类似于灵活的“语法”，它们可用于描述变量如何映射不同的视觉特性，如颜色或形状等。它们还让图表产生关联，如此一来当用户选择一个绘图区域时，其附近的显示就会相应地更新。华盛顿大学计算机学家 Jeffrey Heer（其所在团队开发出 Vega-Lite）说：“实际上，它可以让我们以多维方式探索相关性。”

另外两款产品则可以让研究人员创建可利用小部件的互动应用程序，如可用于混合数据、图表和代码的下拉菜单和滑块控件，包括马萨诸塞州波士顿的 RStudio 制作的用于 R 编程的 Shiny 以及 Plotly 的用于 Python 编程的 Dash。它们通过把用户的小部件的动作传递给一台远程服务器起作用，远程服务器可运行基础代码并更新页面。

由此产生的应用程序可让那些不喜欢编程的研究人员获得相关数据和工具。例如，以色列特拉维夫大学研究生 Tal Galili 与同事合作，开发了一个基于 Plotly 的工具箱，并据此从上传的数据集中构建交互式热地图，Shiny 的一个界面可在幕后运行该代码。北卡罗来纳州杜克大学统计学家 Mine Cetinkaya-Rundel 为本科统计学课程建立了 Shiny 资源，以帮助其在课堂上解释一些有难度的概念。“这种感觉非常好，把它停下来然后说，‘好，现在我们已经介绍完了，当我们移动小部件时会发生什么呢？’”她说。

在期刊网页上发表这样的集成需要对编辑工具、编辑流程和基础设施做改变。它还涉及到把科学数据交付给不能永远保证其表现的第三方。

为了解决这一问题，开放获取出版商 eLife 产品开发负责人 Giuliano Maciocci 说，eLife 的“可再现文档堆栈”项目旨在创建一个端到端工具包，用于编辑、提交和发表在计算上可再现的文档。他说，该计划旨在把一篇论文的多个核心科学“产品”——其文本、数据、代码、图表和计算环境等——压缩到一个可下载的对象中。为了鼓励使用，该期刊已将堆栈设置为开放资源。

大步向前

其他若干家杂志和出版商也在支持代码海洋的集成，包括 GigaScience、IEEE、SPIE、剑桥大学出版社和 Taylor&Francis 等。《细胞生物学期刊》的 JCB DataViewer 基于开源性 OMERO 软件，可让读者探索原始的显微镜图像，而非通常看到的经过处理的压缩文件。一个相关的工具——图像数据资源，可为发表在任何期刊的论文提供类似功能。《自然》杂志也发表了交互性的数据，例如一篇描述“DNA 元素百科全书”项目的论文。一位发言人称，该杂志正在研究若干其他交互代码和数字的选择方案。与此同时，研究人员经常从其文章链接到外部的可视化效果。

得克萨斯州休斯敦贝勒医学院的 Erez Lieberman Aiden 说，随着越来越多的期刊拥抱交互性，科学信息的在线呈现方式很可能会从根本上发生变化，它代表着可再现性的胜利。Aiden 近日在《细胞》杂志的一项成果中发表了交互性的核染色质互动地图，他表示静态图标只是数据的一个方面。“有洞察力的读者需要具备能力得出自己的结论。”他说，“1974 年阅读一篇论文的行为不应该与 2017 年阅读一篇论文的行为相同。”　（晋楠编译）

首个牛胚胎干细胞诞生

有望带来更加健康和高产的家畜

经过几十年的努力，科学家最终成功地从牛身上获得胚胎干（ES）细胞，并在培养皿中使其保持原始状态。获得这些可变成从皮肤到肌肉、骨头等各种组织的多功能细胞，将使调整和保存肉牛以及乳牛品种的遗传性状变得更加容易。这反过来又促成了产生更多牛奶或者更嫩牛肉、产仔时面临更少并发症以及拥有更强的疾病抵抗力的动物。此项发现或许还为研究牛的基本发育和使其成为人类疾病模型开辟了道路。

“我原以为这辈子都不会看到此事发生。”美国密歇根州立大学发育生物学家 Jose Cibelli 表示。他是上世纪 90 年代末试图获得牛 ES 细胞的团队成员之一。自从以后的诸多努力中，当来自牛胚胎的干细胞在实验室中生长出来时，它们会发育成其他细胞类型。这意味着它们很快丧失了多能性。

1981 年，科学家成功分离出小鼠 ES 细胞，从而得以研究早期胚胎发育并且测试遗传缺陷的影响。此后不久，研究人员将目光转向了牛。不过，直到 1998 年，研究人员才找到合适的营养基培养人类 ES 细胞。

作为目前正在分离牛 ES 细胞的团队成员之一，Jun Wu 在其职业生涯中极少关注牛。这位来自得克萨斯州西南医学中心的干细胞生物学家一直对从小鼠和人类身上获得新的干细胞种类更感兴趣。他还对开发将来自两个物种的细胞混合的嵌合体感兴趣，比如拥有人类细胞并且可能有一天生长出移植用器官的猪。2015 年，他同合作者报告称，发现了很容易在实验室和小鼠体内生长出来的新型人类多能干细胞的正确培养条件。

不过，该项目的一位合作者盯上了另一项研究。此前曾在 Cibelli 实验室工作的加州大学戴维斯分校生殖生物学家 Pablo Juan Ross 希望，这些相同的培养条件最终维持来自家畜的 ES 细胞的生长。这会使改善动物遗传性状变得更加容易。为此，这个还将发育生物学家 Juan Carlos Izpis ú a Belmonte 及其在圣地亚哥索尔克研究所的团队包括在内的研究小组，将从牛胚胎中分离出来的干细胞暴露在新的培养基中。这种混合物拥有两个关键组分：促使细胞生长和增殖的蛋白以及另一种抑制它们分化成更多成熟细胞类型的分子。

家畜育种者如今可利用胚胎干细胞改善畜群的遗传性状。　图片来源：HeshPhoto

“他们同时利用了‘加速器’和‘刹车片’。”牧场主、科罗拉多州立大学生殖生理学家 George Seidel 表示。结果是：在实验室中生长了一年多以后，细胞仍保留了多能状态。“多年来，我的很多同事和学生一直试图做到这一点。”Seidel 说。当被注射进拥有较弱免疫系统的小鼠体内时，这些细胞生长成由多种细胞构成的畸胎瘤——真正的多能干细胞的关键特征。研究人员在日前出版的美国《国家科学院院刊》上报告了这一成果。

不过，Seidel 表示，对牛 ES 细胞的兴趣随着克隆技术的发展而有所消退。利用产生多莉羊的相同技术，即来自成年细胞的 DNA 被植入去掉了 DNA 的卵子中，家畜育种者能复制出想要的动物遗传性状，比如快速生长或者丰富的牛奶产量。拥有这些性状的公牛为育种者带来了丰厚的利润，因为他们可以把精子出售给牛肉和牛奶生产商，而后者利用其为母牛实施人工授精从而将更好的性状带给下一代。

不过，Cibelli 介绍说，通常被用于产生这些克隆品的细胞——被称为纤维母细胞的结缔组织细胞很短命，并且仅能分裂 20 或者 30 次。而有了能生存很长时间的 ES 细胞，育种者可以更加轻松地抓住优势细胞系，并且通过诸如 CRISPR 等技术对牛基因组进行多轮编辑。

即便未经过基因改造——一种消费者可能不愿看到被用于牛排和奶昔的技术，ES 细胞也能让育种者更加轻松地选择优势动物。他们可以测试来自不同胚胎的 ES 细胞，以寻找诸如同更多牛奶产量相关的基因等遗传优势。Cibelli 表示，一旦辨别出想要的一系列性状，研究人员便能利用这些细胞创造出无限的克隆品。

对于 Ross 来说，最令人激动的应用依赖于其团队目前的工作——阐明如何将这些 ES 细胞发展成牛的精子和卵细胞。如果他们成功了，家畜基因公司便可以将这些精子和卵子结合起来，以创建拥有新的遗传组合的胚胎，然后从最好的胚胎中分离出更多的干细胞。他们可以利用这种循环——干细胞、精子和卵细胞、胚胎、干细胞——在无须任何动物出生的情况下加速改善后代的性状。这意味着等待牛妊娠的时间变短，同时被浪费的动物变少。“它能使遗传进度加速几个数量级。”Ross 表示。　（宗华编译）

科学线人

全球科技政策新闻与解析

美国会拟提高 2020 年人口普查经费

美国商务部部长 Wilbur Ross

图片来源：U.S. Embassy Bangkok/Flickr

美国最新的短期预算协议使政府能再维持 6 周，这也为该国 2020 年人口普查计划提供了急需的推动力。

近日，美国国会通过的持续决议案还包含划拨 1.82 亿美元用于人口统计局将在 2020 年 4 月开始的 10 年人口普查。该项目预算估计是 156 亿美元，但因为国会未能在过去几年提高该普查相关工作所需的资金，人口普查官员已不得不减少这项大规模普查活动的内容。

去年 5 月，总统唐纳德·特朗普只要求为 10 年人口普查再增加 5100 万美元。2017 年 10 月，商务部部长 Wilbur Ross 对国会表示，该机构将在 2018 年额外拨款 1.87 亿美元，以保证 2020 年人口普查按时进行。

新的持续决议案将把多数政府机构的支出冻结时间延长至 3 月 23 日，但人口统计局是个例外。这笔额外经费相当于给了该机构 2018 年所需要经费的 3/4。国会还允许人口普查官员在今年剩下的时间里以更快的速度完成支出。

人口普查项目的支持者对额外经费表示欢迎，尽管他们认为这还不够。“这是令人鼓舞的一步，将有助于 2020 年的人口普查计划回到正轨和应对挑战。”美国非营利联盟“人口普查项目”的 Terri Ann Lowenthal 说。

此外，该持续决议案将 2020 年人口普查经费提高到 13.82 亿美元，其中 9.82 亿美元将用于为 2020 年的人口普查做准备。但有专家表示，10 年人口普查项目还需要在 2018 年增加 1.4 亿美元，以维持正常运转。

新经费将使人口统计局得以恢复和扩大对延伸活动的范围——例如，将与政府和民间团体合作的人员数从 40 人增加到 200 人，推进其媒体战略，并增加实地办事处的数量。

人们希望国会将在新决议案期满之前完成 2018 年预算。人口普查支持者也希望立法者能继续支持他们。“2020 年人口普查时钟正在滴答作响。”“人口普查项目”联席主席 Phil Sparks 说，“人口统计局不能暂停准备工作，因此，最终的 2018 年预算法案必须给予该项目足够的经费。”（张章）

南非将获大量天文数据

南非 MeerKAT 阵列规模将很快加倍。

图片来源：Mujahid Safodien/AFP/Getty

南非数据科学家正在为迎接海量信息做准备。这些信息将在今年 3 月该国最大射电望远镜规模扩大后到来。

每小时太字节的数据洪流——每分钟将填满超过 3 张 DVD，将从一个名为 MeerKAT 阵列的射电碟形天线网络中流出。目前，该阵列由 32 个天线组成，下个月将扩展到 64 个。

不过，与 2020 年以后的数据相比，即将到来的数据洪流也只是涓涓细流，那时国际天文学家将把 MeerKAT 阵列扩展为平方公里阵列（SKA）的一部分。此外，南非的数据科学家也在寻求将他们的专业知识转移到地球观测和生物信息学等领域。

“我们正在建立一个能让科学家获得授权的系统，以便他们能处理数据。这个系统可以让研究人员更便捷地使用数据和分析工具。”南非开普敦大学天文学家 Russ Taylor 说。

MeerKAT 阵列的目的是收集来自太空的相对较弱的无线电信号，并将它们组合起来，提取更多信息。为了将其转化为 SKA 的第一阶段，工程师们将在北开普省的 MeerKAT 站点上添加另外的 136 个碟形天线，并将它们与散布在西澳大利亚的 13 万个天线相连。

来自 10 个合作伙伴国家的科学家将分享 SKA 的数据。但 Taylor 表示，目前，南非希望保留对其 MeerKAT 阵列数据的控制，而不是将其出口到其他已经拥有数据处理设施的国家。

这在一定程度上是因为分发数据非常昂贵。“在市区连接两个点的光纤的成本是每英里几千美元。”澳大利亚帕西超级计算中心执行主任 Ugo Varetto 说，“在极端环境或水下，就成为每英里数十万美元。”

加拿大天文数据中心组长 J.J.Kavelaars 指出，天文学家也很重视望远镜所产生的数据，不想把它发送到其他地方，“把这些数据从你手中发送出去，就像把钻石运到海外切割”。　（张章）