一份关于数据开源的“少数派报告”

■本报记者 陈彬

如果要发明一辆汽车，是不同的人根据自己的喜好，开发出一堆规格尺寸各异、功能互不兼容的零件，还是在前人探索的基础上进一步跟进，形成彼此数据相关和吻合的各种部件，最终完成组装呢？

只要具备常识的人，在这道“选择题”中应该都会选择后者，但问题是——前人探索时所产生的数据，你怎么能知道呢？

这是在接受《中国科学报》采访时，南京大学人工智能学院教授俞扬打的一个比喻。对于这个比喻，远在天津的南开大学计算机科学与技术系主任程明明也深有同感。就在几个月前，为了能够实现科研数据在科研生态中的彼此共享，程明明在网上发起了一项开放共享科研记录行动倡议。

在倡议中，程明明提议国内科研同行尽量开源自己论文的代码以及相关实验数据，同时，通过将论文“英译中”，设立主页、Demo（演示）等形式，给国内学者更多学习、交流的机会。

被忽略的沟通协作

借用电脑 Word 软件的文件名后缀，程明明将这份倡议命名为“DOCX”。这其中的每个字母都代表了他的一个主张。

“D 代表 Demo，即希望国内科研人员尽可能为论文中的每个问题做一个在线 Demo，方便自己使用，也方便同行的教学实验和科普；O 代表 Open Source，即在不违反保密和商业协议的前提下，尽量开源自己论文的代码以及相关实验数据；C 代表 Chinese version，即建议科研同行能将自己发表的英文顶刊论文共享中文翻译版，方便国内读者阅读；X 则代表 eXplain，即倡议科研人员尽量为每篇论文建立项目主页，方便读者留言提问。”程明明解释道。

在倡议中，程明明表示，近年来，国内科研水平进步很快。以计算机视觉技术为例，目前全球顶级会议论文的第一作者中，华人已超过半数。“但是，我们还得用英文而非母语阅读大部分论文。我们常需验证别人的工作并纠结于为什么实现的结果不一样。我们看到了论文里面的酷炫结果，但尝试这些新技术但并不是很容易。”

究其原因，相关科研数据不公开是个大问题。

此前，国内某科研机构进行过一项持续 9 年的数据实践研究。他们发现，国内科研数据的交流范围很少超出生产数据的科研团队本身，而且对周边环境数据的共享请求也很少，其原因既包括缺乏专业技能、资源以及共享数据的激励措施，也牵扯科研道德问题。

“在我国的科研领域，开源氛围相对较弱，‘论文’意识过强，导致科研人员忙于发表论文，反而忽略了科研工作其实是科研同行之间沟通协作、共同攻克科学难题的过程。”程明明说，这还导致我们实现前人工作时大量重复。更重要的是，最终的论文不可能把所有细节说清楚，自己的版本往往不如原始作者的版本。

科研生态的有力保障

十多年前，还在求学的俞扬得到国际科研共享数据的支持。“我受益最多的是来自新西兰怀卡托大学的某开源机器学习包和美国加州大学尔湾分校的机器学习公开数据集。”从这些开源数据中，他不但学习到一个优秀项目的工程实现的过程，也贡献过错误纠正。

就在俞扬受益于数据开源几年后，2011 年，程明明养成了代码开源的习惯，在不违反保密和商业协议的前提下，尽可能最大化地开放科研成果中的代码和数据。“若干年后，我猛然意识到，自己没有及时开源的代码和数据，由于工作单位和常用电脑的多次更换，大部分都找不到了，而开源数据却可以随时在网络上找到。”程明明说，一次数据开源既方便了别人，又方便了自己。

然而，数据开源对于科研的好处并不局限于“保存数据”。在程明明看来，开源本身就是对科研生态的有力保障。

“近年来，国内科研界对论文造假的质疑时有发生，但这类事件在计算机科学领域却相对较少。一个重要原因在于，该领域开源风气比较浓厚。”程明明说。

对此，俞扬表示，至少在机器学习领域，开源几乎已是业界共识。“算法开源十分高效地完成了成果的有效性检验，也加速了有效成果的传播普及，使得后人的工作可以建立在前人的基础上，推进领域前行。目前我所在的 LAMDA 研究组已经开源了 150 多个项目。”

然而，如果将视野放置于整个科研领域，目前国内从事数据开源相关工作的依然属于“少数派”。

在“DOCX”倡议后的留言中，武汉某高校一位遥感专业博士生感叹道：“计算机现在为什么能‘吊打’生化环材，从某种程度上与共享政策有很大关系。科研人员可以在网上找到共享资源，再继续共享开源，进而把整个蛋糕做大。而不是像‘生化环材’那样搞封闭实验。”

那么，是什么在阻碍“生化环材”领域的数据开源？

不必着急的过程

自 2020 年 10 月程明明将 DOCX 倡议发布于网络，至今已有几个月的时间。对于大家的反馈，程明明总结为三句话——几乎全部赞同、没有反对、很多质疑。

在质疑声中，主要的质疑点在于此倡议不存在约束力，同时开源者本身不会有任何获利，因此担心难以执行下去。对此，程明明却看得很开。“这只是我个人倡议，本身就不可能有约束力。如果有人认可并这样做了，那对于建立健康的学术生态就会产生益处，但即使没有人做也不会带来坏处。”

至于“获利”问题，程明明表示，开源科研记录在短期内不会带来收益，但一方面，如果科研人员都能这样做，任何人都可以利用其他人的数据，这在无形中会大大节省时间和精力；另一方面，即使是从纯“获利”的角度来看，开源也未必不会带来好处。

“比如，此前我曾公开过一些研究项目的代码。多年后，企业主动找到我洽谈合作项目，原来他们的负责人在求学时就曾使用过我们开源的代码。开源无形中扩大了合作对象，对此很多人并没有注意。我们这些年很多项目都是这样产生的。”程明明说。

在俞扬看来，目前数据开源面临的重要争议之一，还在于需不需要中文版共享内容。在这个问题上，他的态度十分明确——民族、文化与语言不可分割，中文版的内容十分必要。

“要得到其他语言文化的认同，其根源还是工作的引领性。能带领人们翻越巅峰，才能聚集更多协作者。”俞扬表示，“跑在第二的位置，可以简单地用距榜首多远来衡量成绩。然而，轮到跑在榜首的国产引领，评价就变得十分困难，方向、速度等都会失去参照，质疑声往往比掌声多。”

虽然发布了 DOCX 倡议，但对于数据开源工作未来的发展，程明明并不着急。“必须承认，数据开源工作在我国还处于发展的初期，并不是政府出台政策就可以‘立竿见影’的。”在他看来，数据开源的推进需要“优胜劣汰”的自然过程。

“作为数据开源的‘先行者’，我们已感受到这项工作带来的益处，而周围的人也会在我们的带动下，体验到其中的好处。至于不开源者，则会在过程中慢慢被边缘化，最终被淘汰。这是一个缓慢的过程，我们需要保持耐心、坚持到底。”程明明说。

简讯

海南大学全健康研究院揭牌

本报讯 2 月 6 日，海南大学召开全健康研究院成立大会暨新时代传染病防控研讨会。海南大学全健康研究院在会上揭牌成立。

据悉，海南大学全健康研究院将重点开展全健康教育、科学研究、国际合作和产业发展等四大工作内容，开展跨学科、跨部门协作和交流。海南大学将与海南省疾病预防控制中心、海口市人民医院、海南省动物疫病预防控制中心共建一流的全健康研究与创新平台，汇集一批全健康研究方向的高水平人才，抢占国内外全健康研究高地，建成具有海南特色的、国际领先的全健康学科群。

（余梦月 潘锭钰）

16 门食品营养与健康专业教材将出版

本报讯 近日，高等学校食品营养与健康专业教材建设启动会在西北农林科技大学召开。此次将启动建设 16 门课程新教材，新教材由西北农林科技大学等国内 50 余所高校和科研单位的 150 多名专家编写，在 2021 年至 2022 年间出版并投入使用。

这 16 门课程新教材编写具有三大亮点：一是拓展强化了医学与生物学基础，二是强化了科学研究与功能食品开发训练，三是体现了食品科学和信息科学、传统医学的结合。（靳军 张行勇）

山西艺术职业学院成立

本报讯 日前，山西艺术职业学院、山西戏剧职业学院和山西省晋剧院、山西省京剧院合并组建为新的山西艺术职业学院。

在该院“院团合一、人才共享”聘书颁发会议上，相关人员共为首批 10 位挂职干部，58 位业务骨干、教师和演员，10 位名家工作室专家和非遗大师颁发了聘书。

据了解，建立“院团合一、人才共享”机制是山西省委、省政府关于高等教育“三个调整优化”重大决策部署。该省通过有效整合校、院、团，形成教学、科研、演出一条龙，打通人才培养、艺术生产、文化创新各环节，逐步形成“产教融合、院团合一”的办学新模式。

（程春生）

北京建筑大学明湖交叉学科创新论坛开讲

本报讯 近日，北京建筑大学明湖交叉学科创新论坛（简称明湖论坛）首场开讲。首场明湖论坛聚焦智慧城市与智能建造前沿交叉技术，邀请了中国科学院大学教授陆锋、清华大学教授贾庆山作特邀报告。

据介绍，论坛取名自该校大兴校区的明湖，有明德、明诚、明辨、明盛等寓意。明湖论坛旨在打造一个学术交叉平台，加强合作、共享技术，促进学科融合，孵化跨界创新团队，实现科技创新跨越式发展。首场明湖论坛采取“专家演讲 + 交流研讨”的形式，侧重交叉、突出互动。（温才妃）

受新冠肺炎疫情的影响，上海交通大学学生决定留校做实验，过一个“学术年”。

上海交大学生处副处长龚强表示，除少数家在中高风险地区暂时无法返乡的学生，留校学生中大部分都希望利用这段时间精进学业、推进科研进度，因此学校在寒假开放实验、学习场馆，公共科研平台分析测试中心，让他们的科研没有后顾之忧。

黄辛摄影报道

多所大学延迟开学报到时间，专家指出——

警惕疫情成为高校懒政借口

本报讯（记者袁一雪）日前，国务院联防联控机制举行新闻发布会。教育部应对新冠肺炎疫情工作领导小组办公室主任王登峰在会上表示，教育部要求春季学期全面开学、正常开学和安全开学。在特殊情况下，可能还需要分批错峰开学。如果春节过后还有中高风险地区，这些地区的学生可能要暂缓返校。同时，各地各校要做好线上线下教学衔接准备，需要时随时开启线上教学。

随后，有高校发布通知，推迟到校报到时间，并明确了线上教学的时间。相较去年新冠肺炎疫情来临时，各大高校被动响应，今年确实积极得多。不过，去年临时从辅助教学转变为主要教学方式的在线教育，虽然保证了学生的学习进度，也暴露出多方面问题。如任课教师准备不充分，线上教学质量参差不齐；线上课堂对于教学平台稳定性的过度依赖；教师对线上课堂监管不便；师生在线互动不强等。

对此，中国教育科学研究院研究员储朝晖在接受《中国科学报》采访时表示，在线教学今年保持了发展的势头，基础越来越成熟，但在线教育规模与数量的扩展并不意味着其本身有了“质”的飞跃。“在线教育除了涉及技术问题，还涉及行为心理学等，所以使用人数多、规模大，不代表在线教育已经成熟到可以替代线下教育。”

对于一些高校已经发布通知延迟到校报到时间，储朝晖并不认同。在他看来，当下疫情的发展与去年不同。去年对疫情走向更多的是未知，今年已知可控的因素较多，例如更好的防控意识和疫苗的普及。

储朝晖认为，每一次人类社会发生大灾难，都是人类发展的分水岭，会让勇敢的人更勇敢，智慧的人更智慧，也会让畏缩的人更畏缩，愚昧的人更愚昧。“不论是高校还是其他部门，在防控的同时，更应该让疫情成为一次教育学生的机会，让他们变得更加智慧和勇敢。”今年是否推迟开学，现在就作出判断为时尚早，至少与目前疫情在可控范围的大趋势不相符。“应警惕疫情成为高校懒政的借口。”

广东打造职业教育“深圳方案”

本报讯（记者温才妃）近日，教育部、广东省人民政府联合出台《关于推进深圳职业教育高端发展 争创世界一流的实施意见》（以下简称《意见》），计划到 2022 年深圳职业教育累计投入 100 亿元，打造世界一流职业教育，进一步服务国家战略和建设粤港澳大湾区、支持深圳建设中国特色社会主义先行示范区。

《意见》提出，在专业建设方面，将重点围绕集成电路、新一代信息通信技术等万亿级和千亿级产业集群建设 10—15 个一流专业群。瞄准芯片、新材料等关键技术，加大相关学科专业建设和高素质技术技能人才培养力度，支持深圳建设职业教育微电子人才培养示范基地等平台。

推进中职学校每个专业群联合行业领军企业或行业组织、高职院校共建 1 个产教联合体；高职院校每个专业群联合世界一流企业或行业领军企业建设 1 所特色产业学院，共同建设高水平专业、开发课程标准、打造教学创新团队、设立研发中心、开发高端认证证书，共同“走出去”，到 2022 年建成 10 个示范性特色产业学院。

支持深圳职业技术学院、深圳信息职业技术学院开设部分本科专业，创建中国特色世界一流职业学校。支持广东新安职业技术学院打造民办品牌高职院校。依托深圳广播电视大学建设开放大学。支持广东省按程序将深圳技师学院纳入高等职业学校序列。支持深圳技术大学招收职业学校毕业生。支持有条件的高职院校建设科技园区，培育孵化育成体系，助推师生创新创业。

中国人民大学评价研究中心执行主任周光礼指出，现代职业教育具有三大特点，一是职业性，二是高科技性，三是区域性。世界一流的职业教育体系需要依托拥有世界级产业的区域，进行高水平的产教融合，而纵观国内，唯有深圳或粤港澳大湾区，才拥有世界级的产业群。深圳要打造世界一流的职业教育、贡献高端职业教育的中国方案，不仅在职业教育的积淀上有基础，而且也契合了国家“十四五”规划建立高质量教育体系的目标，因此，创造新模式是很有可能的。

他指出，此过程中有诸多新看点，如推进纵向贯通，将职业教育与中小学对接，帮助中小学校进行劳动教育和职业启蒙教育；支持职业院校创办本科，加上之前主打专业学位研究生培养的名校研究院和产业学院，将实现职业教育体系从基础教育贯穿至博士教育。加强横向融通，如支持深圳技术大学招收职业院校毕业生，从而解决职校毕业生的升学出口问题。

热点微评

栏目主持：温才妃

首轮“双一流”建设成效评价将在今年完成

2 月 4 日，教育部公布了 2021 年工作要点。该要点中指出，2021 年，将完成首轮“双一流”建设成效评价，实施一流学科培优行动。

点评：

“双一流”建设成效评价要经历自评、第三方评价、教育部组织的专家评价等环节。目前，一流建设高校、一流建设学科均已完成自评，所有动态监测数据材料已提交教育部。教育部正在积极开展一流学科的专家评议。

“双一流”建设成效评价的结果最终估计会分档呈现。表现优异的高校或学科，可能会获得加大建设的支持；表现不佳的高校或学科，可能通过动态调整被剔除出“双一流”建设范围。但也有部分专家建议，不要着急剔除暂时表现不佳的高校或学科，毕竟“双一流”建设为时尚短，要给予它们一次再观察的机会。

——同济大学发展规划部部长 蔡三发

高校新添一名女校长

2 月 4 日，中国石油大学（北京）召开干部教师大会，宣布吴小林同志任中国石油大学（北京）校长，与华南理工大学前校长王迎军、西安电子科技大学前校长郑晓静、中国石油大学（华东）前校长山红红、山东大学校长樊丽明等耳熟能详的名字一起，成为了备受关注的中国高校女掌门人中的一员。

点评：

翻开人类的教育史，很早就有这样一群“智慧女神”的身影，她们以女性特有的视角和智慧，倾情奉献于人类文化教育的发展与进步。她们带给教育的感觉，让我想起了意大利的博洛尼亚市——一座被称为最母性的城市：古城的人行道上覆盖着拱廊，为行人遮阳挡雨，使行人避开来往车辆，夜晚为行人照明……整个城市充满了一种介于世界和自己的身体之间的母性调节感。

有统计数据显示，近几年从普通高中到本科、研究生，各个学段女生比例都整体呈现出持续上升趋势，接受本科教育的群体中女生的比例超过男生，社会地位逐渐提高使女性对接受教育的重视程度也在逐步加强。在这样的高等教育发展背景下，女校长队伍的蓬勃发展对大学中的女生具有更好的榜样作用。

哈佛大学第 28 任校长德鲁·吉尔平·福斯特是该校历史上的第一位女校长。这位始终站在提倡种族平等战线上的女权主义者，当选校长后强调“我不是哈佛女校长，我是哈佛校长”。

——对外经济贸易大学副研究员 李芳

名校生竞争高校辅导员

近日，武汉大学对外公示了 2021 年拟招聘的辅导员人员名单。在这份名单里，应聘者均来自国内名校，当中不乏清华、北大博士，还有海归名校博士。

点评：

这一现象暴露出一个已被聚焦多次的问题：为什么许多人明明不想做学术，却还是要攻读博士学位？面对越来越严峻的就业挑战，他们中很多人在明知没有专业前景的领域苦苦读完博士，然后去做一份完全与专业无关的工作。

近年来，研究生报考人数持续大涨，其中大多来自人文社科专业。相较理科生，人文社科专业的毕业生要想获得一份体面或者有竞争力的工作，似乎难度更大。当然，这可能无关学科价值高低的问题，主要还是由社会需要决定的。从这个角度来说，未来博士生的扩招，应与实际需求相匹配。当这种结构性的调整与社会需求结构匹配度日趋接近，自然可以减少讨论中所提到的优质教育资源的浪费。

——中国教育在线总编辑 陈志文

清华、北大开放互选课程

2021 年春季学期，北京大学和清华大学两校继续开放互选课程，学分互认。北大将向清华本科生开放 56 门本科生通识核心课、含 396 个名额；清华开放 27 门次优质课程、含 299 个名额。有网友发问，何时其他学校的学生也能选北大、清华的课程？

点评：

课程互选和学分互认，需要各校课程质量、学分含金量基本一致。否则，如果某校的课程质量不高，那选某校的课程就会被质疑是“混学分”，认可该校的学分，就可能拉低培养标准。

如北京一所非“双一流”建设高校的学生，选读“双一流”建设学校，甚至北大、清华的课程，那么既可实现“双一流”建设学校优质资源的辐射、共享，也能提高学生选课的积极性。但要实现互选而非单选，包括在获得本专业学位证书的同时，获得辅修学位，就必须提高非“双一流”建设高校的本科教学质量，达到和“双一流”建设高校本科教育一样的水准。

推进校际课程互选和学分互认，与整体提高所有本科院校的教育质量是相辅相成的，对破除我国的“唯学历论”也有重要的推动作用。

——21 世纪教育研究院院长 熊丙奇