新概念新法则

大数据处理应遵循的原则

□张 鼎

信息革命将人类带入一个革命性的“大数据时代”，人们通过电脑网络，利用数据分析、数据集成、数据设计、数据模型、数据决策和数据虚拟化等手段，针对具体的认知对象，对反映其数量、性质、结构、功能、价值、前景等方面的数据，进行全面的收集、整理、分析、综观和萃取，然后作为决策和实践的根据。由此，使人类的认识变得更精确系统、更全面广泛、更快速有效。然而面对各种数据源收集来的海量数据，究竟如何对其进行归类、计算、建模和解释，却需要高超的智慧。因数据本身始终沉默，需要依赖能动的认识主体的赋予。为此，本文想就大数据处理应遵循的几条原则给予阐释。

简约原则

这是牛顿创设的一条“极简主义”的节约规则。在牛顿看来，神奇的自然界在创设过程中选择的简单性和对繁琐的讨厌，使得人类也形成如下观念：“在用很少的东西就能够解决问题的情况下，决不劳力费神和兴师动众”；要始终体现大自然所遵循的简约性、精准性、合理性与有效性。欲达此目的，就需要在云计算或大数据处理中，遵循简约原则，选择有用数据，淘汰无用数据；识别有代表性的本质数据，去除细枝末节或无意义的非本质数据。要能够确识数据之间的巨大差距或差异；要能够鉴别和挑出那些“以一当十”的数据和信息。这种简约原则在大数据的收集、挖掘、算法和实施中的最有效途径，就是对“数据规约”的运用。所谓数据规约就是简化现有的数据集，使得一种小规模的数据就能够产生同样的分析效果。常用的数据规约策略有数据立方体聚集、维规约、数据压缩、数值压缩、离散化和概念分层等，而常用的数据规约方法则主要包含粗糙集、遗传算法、主成分分析、逐步回归分析、公共因素模型分析等。运用这些规约方法，就可以获取可靠数据，减少数据集规模，提高数据抽象程度，提升数据挖掘效率，使之在实际工作中，可以根据需要选用具体的分析数据和合适的处理方法，以达到操作上的简单、简洁、简约和高效。具体地说，当一位认知主体面对收集到的大量数据和一些非结构化的数据对象，如文档、图片、饰品等物件时，不仅需要掌握大数据管理、大数据集成的技术和方法，遵循“简约原则”和“数据集成原则”，学会数据的归档、分析、建模和元数据管理，还需要在大量数据激增的过程中，学会规约、选择、评估和发现某些潜在的本质性变化，包括对新课题、新项目的兴趣和开发。

综观原则

所谓综观，就是对认知对象进行综合性的观察、分析和探索；就是从总体上对认识对象、认识过程和认识结果进行抽象、概括或直觉，并通过具体的信息数据超越那涵盖于总体性中的局部或个别。这种综观既针对构成事物之个体的全部，也针对构成事物的诸要素组成的统一体，以及总体上显现的本质和规律。综观较整体观察更加辩证。它坚持从大处着眼，从总体上去“观其状，求其法，探其道”，以求得解决问题的策略和战略。它坚持整体的具体统一性，凸显认知对象的具体实在性。至于现实中，人们究竟如何对具体的认知对象进行综观，这里需要借助与综观紧密相关的大数据集合的理论与实践。因为大数据集成，既包括对存贮在结构化数据结构中的数据进行移动和集成，也包括对一大部分非结构化数据中的数据进行移动、调节和集成。比如面对复杂的信息和数据，人们就可以将“云架构、实时数据集成、数据虚拟化、数据集成建模”等先进技术用到具体问题的解决中，使用一种根据大数据制作的“可预测模型描述语言”（PMML），为其提供一种快速简便的程序和模型。此时，通过使用标准的XML（可扩展标记语言）解析器对PMML进行解析，应用程序就能够决定模型输入和输出的数据类型，及模型的详细格式，并会按照标准的数据挖掘术语来解释模型的结果。通过对大数据的综观、模型化和虚拟化，可以做到花最小气力，获最大效益。特别是数据虚拟化，不仅可以为数据使用者提供极具真实性、完整性和精准性的“实时集成的数据视图”，还可以将来自不同数源的数据信息整合为一，并转化成使用者所需要的图式和模型。因为有些对象，绝不是仅仅用数字就可以解释和认知的，比如人类复杂多变的面部表情，就很难用单纯的数据给予精确表达，只有通过数据集成、智能技术和虚拟技术将大数据虚拟化，使反映认知或实践对象的海量信息和数据，变成一种实时图像或视频供主体观察研究，他们才可能从中获得相关的认识、结论和决策。

解释原则

尽管数据集成、数据建模、云计算和数据虚拟化是大数据处理的一些主要形式，能够给认知主体以质、量和度等多方面的总体性的形象和认识，但面对“不能言语”的具体的数字、信息、图像和虚拟视频，还是需要认知主体进行能动的和创造性的解读与阐释。这不只因为包括数字、数码、文字和一切符号在内的语言“是存在的故乡”，即要认识客体或对象，总是需要利用语言给予建构、包装、说明和解释，还因为一切语言自身所拥有的价值和意义，也需要使用它和阅读它的主体所“赠予”。换句话说，这些摆在人们面前的大数据，究竟表达什么或意味什么，很大程度上，并不取决于由数据信息自身所标明的“客观实在性”，而是主要取决于认知主体对其解读时所拥有的整体上的解释力、构建力和知解力，取决于由数据构架起来的理论形态和实践目的。因为只有通过人的感悟、觉识、分析、推理、判断和阐释才能够赋予数据和信息以多重的或异乎寻常的结构和意义，才能够由表及里，揭示出深藏于内的隐蔽之物；进而通过各种解释之间的矛盾和冲突，获悉被解释的存在和本质。

智慧原则

尽管信息革命将人类带进大数据的春天，而且使越来越多的人确信“数据多多益善，即数据越多，分析越深入，所得的结论就越全面”，但面对“僵死的数据”，要想点石成金，还需依赖于人的智慧和学识。为此，在大数据处理过程中，真正的智者既要兼具数据分析、机器学习、数据挖掘以及数据统计的能力，也要具备应用算法和编写代码的经验。尤其是面对琳琅满目的大数据，不仅要关注海量数据的多样性、差异性、精确性和实效性，否则缺少其中任何一个性能，都可能使所获数据达不到预期的效果和目标；还要全面深入地挖掘各种类型的数据，并在此基础上运用数据建模和数据算法在不同的数据集成中分析不同的假设情境，建构不同的可视化图像，进而揭示数据集成的变化及其产生的效用。特别是今天，面对激烈的社会竞争，必须不断寻找新的数据处理方法，不断加快数据处理速度。要意识到各种数据都并非生而就有价值，只有通过主体智慧的挖掘，才能将其变为现实。

此外，还要善于从数据集成、数据建模和数据虚拟化中发现和解决问题；提升自己观察、思考、批判和扬弃的能力；锤炼自己的理性思维和逻辑思维；培养自己统筹决策、高瞻远瞩、见微知著的预见力和洞察力。当然，在大数据时代更需要掌握对已有的数据模型进行精练，以及利用新的训练数据对原有内容和规则集进行修改、操作和运行的技艺。在此过程中，要尽可能做到思想活跃，思维清晰，头脑开放，认识深远，能够不失时机地打破陈规旧套，抓住新机遇，尝试新途径，开辟新天地，以多元智能的理念来认知和实践，以便在大数据处理中，既不忽略任何一个未经深度分析的数据，也不丢弃任何一个异常数据。在许多情况下，异常数据往往比常规数据更有价值。这样，也就自然地要求认知主体工作上缜密细心，时时关注事件的每一个细节与数据，真正做到明察秋毫、细心研制，直至收获完美的认识和成功的实践。

科技与社会

人工智能与人类智慧的交锋

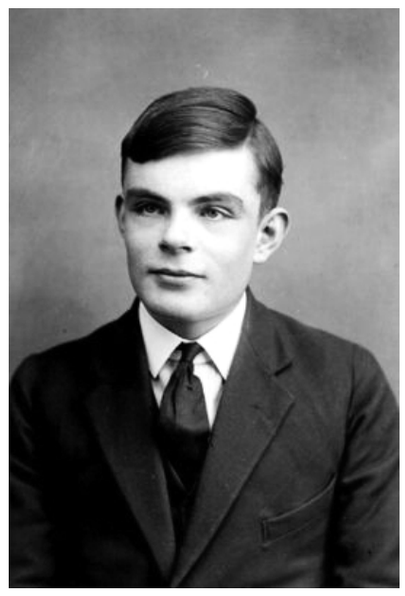
□吴保来

阿兰·图灵被称为计算机科学之父、人工智能之父。1912年6月23日，图灵生于英国的帕丁顿，1931年进入剑桥大学国王学院，1938年在美国普林斯顿大学取得博士学位。

二战爆发后，图灵的祖国危在旦夕，他怀着满腔报国热情，返回英国。他接受了英国外交部的“超级机密”研究工作，即主持对德军通讯密码的破译工作，协助军方破解了德国的著名密码系统ENIGMA，对盟军在大西洋战场上的胜利发挥了关键作用。1945年获得英国政府的最高奖——大英帝国荣誉勋章。

1950年，他发表了题为《机器能够思考吗?》的论文，在计算机科学界引起巨大震撼，为人工智能学的创立奠定了基础。1951年，当选英国皇家学会会员。1954年，图灵因食用浸染过氰化物溶液的苹果而死亡，神秘的死亡给人留下难解之谜。为了纪念他对计算机科学的巨大贡献，美国计算机协会于1966年设立一年一度的图灵奖，以表彰在计算机科学中做出突出贡献的人。

如何才能判断一台机器是否像人类那样思考呢？或者说一台机器是否能像人类那样具有思考和解决问题的能力呢？1950年，图灵提出了一个测试标准，用来判断电脑能否被认为是“能思考”的机器。在图灵看来，如果一台机器的行为，让人类无法辨别它是机器还是人类，那么就可以认为，这台机器具有了人类智能。这个测试被称为图灵测试，它已被多数人所承认。



图灵测试中测试的一方是人，被测试的另一方是一个人和一台声称有人类智力的机器。测试时，双方在隔开的情况下，测试人通过一些装置（如键盘）向被测试人或机器随意提问。问过一些问题后，如果测试人不能确认被测试者30%的答复哪个是人、哪个是机器的回答，那么这台机器就通过了测试，并被认为是具有人类智能。到目前为止，还没有一台机器能通过图灵测试。其中，最接近人类智能的机器是一个由俄罗斯专家设计的“叶甫根尼”的电脑程序，它在2012年英国举行的一场国际人工智能机器测试竞赛中，以29.2%的回答都成功地“骗过”了测试人，取得了仅差0.8%便通过图灵测试的最好成绩。

当时间回溯到1997年5月时，美国IBM公司的“深蓝”超级计算机首次击败了国际象棋男子世界冠军卡斯帕罗夫，人类真的败给机器了吗？很多人开始忧虑人类将被自己所创造的东西取代，人类的智慧和尊严受到了挑战。图灵曾预言，到2000年将会出现足够好的电脑，能够让很多人误认为它们是“人”。目前，也的确有一些电脑在测试中“骗”过了部分测试者。那么，是否电脑就已经超过人脑了呢？现在看还言之过早，虽然人工智能取得了一个又一个成功，但却远未达到像人类那样思考，光靠存储器和程序永远也不能解决思维和智能领域的所有问题，甚至人类自己都无法确认自己的思维运行模式，更谈不上让机器来模仿人类的思维。

如果一台机器通过了图灵测试，那么它就具备思维能力了吗？或者说，一个能和人类正常交流的机器，能不能算是有思想的机器呢？1980年，约翰·西尔勒提出了著名的“中文屋子”思想实验，揭示了一个深刻的道理：有言语交际能力并不表明有思想。

西尔勒的“中文屋子”假设：完全不懂中文的西尔勒博士（假定是计算机中的CPU）想象自己被锁在一个封闭的屋子里，可以通过输入和输出的微小缝隙与外部连通。输入的是中文问题，房子内有一本英语的指令手册（相当于程序），从中可以找到相应的规则。他按照规则办事，把作为答案的中文符号写在纸（相当于存储器）上，输出到屋子外面。这里，表面上看他能够处理中文问题并给出正确答案（如同一台计算机通过了图灵测试），但是，他对那些问题一窍不通，他根本不理解那些问题，甚至不认识其中的任何一个词汇。这个思想实验说明：即使机器能通过图灵测试，我们还是不能断定机器具有理解能力，图灵测试中定义的智慧是不能充分描述人类智慧的。

在很多科幻文学和电影作品中，机器人与人类之间的斗争是一个永恒的主题，这恰恰反映了人类对自己未来命运的担忧。人工智能只是对人类智慧的模拟，它可以解决模拟人类的逻辑思维能力，但是却永远无法实现模拟人类的情绪和情感；可以解决客观知识的问题，但却无法克服社会和文化因素的差距；可以实现推理和设计，但却无法实现人工直觉和人工想象力。人工智能是科技进步的重要成果，能够为人类的发展作出重要贡献，但是，仅仅依靠科技是无法解决人类所面临的所有问题。一方面，人类自身要跟上科技发展的步伐，抛弃愚昧、无知的陈腐观念，不断变革、创新；另一方面，人类要寻求内在超越之路，从精神、思想、道德、伦理等方面迎接时代的挑战，不断调整人际交往、人类社会和人类文明的进化方向。

国外组织实施专项科技任务的经验

编者按：按照近期深化中央财政科技计划（专项、基金）管理改革提出的要求，政府部门主要负责科技计划的宏观管理，不再直接管理具体项目，科技布局上要形成聚焦重点任务配置资源、集成攻关的新体制。国外相关经验可以提供哪些启示，本报特编发此文，供读者参考。

□李哲 苏楠 郝君超

专项科技任务是指与常规科技计划（基金）渠道设立的单项项目相对应，面向特定目标集成研发、中试、试点示范等创新链各环节，以项目群的方式组织实施的科技创新活动。采取专项方式快速提高特定领域的科技和产业竞争力是许多国家的做法。

以综合性的长期论证为基础 确保技术经济上的可行性

专项任务在启动实施之前都要进行充分严密的综合性论证，论证周期较长。如美国国家纳米技术计划（NNI）在出台前经过了4年多的调查研究，伽利略计划经过长达3年的论证，韩国高铁计划经过8年的调查，国际热核聚变实验堆计划（ITER）经过近20年的论证设计才启动。

从技术、经济效益、运营管理等角度进行全方位、多层次的评估。ITER计划对项目的选址、费用、采购责任、管理和运行等方面也进行了多轮谈判，尤其是在费用预算方面，通过多次的论证将经费由原来的预算造价100亿美元降至约46亿美元。

设立专业的协调机构和咨询机构 协同推进实施

针对专项涉及部门多的特点，大多数国外科技项目都建立了跨部门乃至国际性的高层协调机构。其职能除了负责项目运行管理、保障各参与方的有效合作外，也包括协调政府各部门在技术研发、配套基础设施、配套政策等方面的工作，建立政府公共科技投入成果与企业需求的联系，解决突出的技术路线争议，保证研发活动对所有利益相关者公开透明等。

采取全过程的专家参与，发挥专家的决策咨询作用。为解决科技专项实施过程中遇到的技术、管理等难题，许多项目都设立了专家咨询机构，在项目的立项、过程管理、成果评估等各个阶段都充分发挥专家的作用。ITER计划成立了两个咨询委员会：一是科学技术咨询委员会，负责项目建设和运行期间发生的科学技术问题；二是管理咨询委员会，负责项目的管理问题，如预算分配、权利的运用、提出推进项目实施的管理建议。欧洲新药创新计划（IMI）科学委员会由会员国推选，最多由15人构成，对计划项目完成目标的情况进行指导。

依据任务目标导向 因地制宜选取管理模式

为有效管理科技专项，除了在总体上加强组织协调和领导之外，还需要有相应的组织管理模式。国外专项管理模式一般有两类，一类是分散式组织方式，另一类是成立新专业化组织机构。

分散式的管理方式往往用于涉及多个领域且研发成果具有较高独立性的科技专项。美国国家纳米技术计划（NII）包括七个领域，涉及26个部门，这些部门自身已经拥有一套比较成熟的研究开发项目管理体系，因此它们对于负责的领域，会根据自身的项目管理体系确定项目承担单位，进行监督、管理和评估。纳米技术计划的开展只是增加了项目的数量，并没有新增任何管理体系。

成立专门机构的方式往往用于市场化程度较高，产品导向突出的技术领域。强子对撞机（LHC）科技专项则由欧洲11个国家共同创立的欧洲核子研究中心专门负责实施。欧洲的新药创新计划（IMI）由欧盟下属的欧洲委员会和欧洲制药工业协会共同实施。“空中客车”研制计划成立空客公司，以保证管理的独立性。

注重评估监督 强化对项目的过程管理

建立项目评估机制，强化项目管理的监督。以纳米技术计划为例，美国国会于2003年通过法案对国家纳米技术的评估做出了专门规定，要求对纳米技术趋势、对联邦纳米技术研发计划改进措施，以及对该计划在社会、伦理、法律、环境和人力方面所产生的影响进行评估。为了保证评估的有效性和科学性，该计划采取了第三方调查、电子化商务报告系统、同行评议、案例分析、统计分析、各种模型等多种手段，对各类型、各学科领域的研究进行多角度、多层次、多方位的评价。

建立项目动态管理机制，保证资源配置的有效性。美国国家信息高速公路计划（NII）采用集中上报预算的方式，保证了在各部门之间根据计划目标、部门使命而进行的资源分配。同时，每年度的计划预算都会根据实际情况而进行调整，形成了计划所涉及内容及相关管理的动态性。

以产业界的参与为先导 带动科技成果的转化

政府采取各种措施鼓励产业界参与科技专项，促进科技成果的产业化。德国的电动汽车发展规划通过“视窗”及“灯塔”两大工程，让更多的企业、组织及民众参与到德国电动汽车的发展过程中来，以加快电动汽车的市场普及进程。

欧洲新药创新计划（IMI）充分发挥企业作为科技创新及市场拓展主体的作用。第一，在研发主题提出的过程中，参与者全部为制药行业协会的企业，政府机构和研究机构并不参与这一过程。第二，在组织结构上，项目的实施通常由一个大型企业主导，其他大型企业和中小型企业共同参与解决技术和工具开发等方面的瓶颈。第三，在组织管理上，主导企业在承担的项目组织方面具有很大的管理权限，它能够合理分配和使用资金，能够与更多企业和研究机构进行合作，大大提升了管理效率，降低了管理成本。

几点启示

加强任务宏观管理。专项科技任务涉及领域广泛，参与研究单位众多，组织实施的管理部门也多，统筹协调难度大，容易出现项目平行重叠、交错运行、责权利不清晰等问题。为了保障专项顺利实施，需要建立组织高效、责权明晰、监督有力、运行通畅的专项组织管理模式和运行机制。对于各部分相对独立，系统集成难度不高的任务，可采用分散式的管理模式。对于各部分联系紧密、系统集成难度大的任务，可以委托一到两个专业化的组织机构负责统筹实施，并设立技术和管理方面的咨询委员会，从组织管理、技术把关、实施监督等各个层面明确各方职责。

明确具体的产品（工程）目标与技术指标。专项任务在实施之初，需要提出明确具体的产品（工程）目标与技术指标。只有这样才能将目标、指标与监测相衔接，将专项目标和各科技计划的项目目标相衔接。专项任务在论证过程中，需要选择适于以重大项目和重大工程的方式组织实施的重大战略任务，注重凝练形成明确具体的重大产品（工程），或关键共性技术的突破及集成应用示范。在实施过程中，基于明确的成果产出，从国内外技术发展趋势与竞争态势，专利、技术标准的布局，不同技术路线的优劣，技术经济性等方面，按进度分阶段与国际主要竞争对手进行对标。

以专项任务带动产学研结合。鼓励企业、大学和科研院所以产学研合作形式参与专项任务，建立产业技术创新合作关系，充分发挥企业在把握市场需求以及大学和科研院所在研发能力上的优势。对于有产业和产品需求的专项任务，应明确企业的实施主体地位，赋予牵头单位在项目、经费安排等方面更灵活的资源配置权限，鼓励企业自由参与和竞争。

加强项目实施的全过程监测。专项任务要加强项目实施的全过程管理、监测与评价，建立明确有效的评价指标体系和项目实施信息管理通报系统，对项目实施各项情况进行监测，随时掌握和了解项目群的实施状态，为进度管理提供动态信息。

创新政策