
10 光明讲坛

大数据给我们带来了哪些改变

10版:光明讲坛 上一版 下一版

大数据给我们带来了哪些改变

演讲人：陈国青 演讲地点：人文清华讲坛 演讲时间：二〇一九年六月

作者： 《光明日报》（2019年07月13日 10版）

陈国青 清华大学经济管理学院EMC讲席教授，学术委员会主任。2005年度受聘教育部长江学者特聘教授，担任教育部高等学校管理科学与工程类专业教学指导委员会主任委员，国家信息化专家咨询委员会成员，国际模糊系统学会（IFSA）副主席，中国信息经济学会副理事长，中国系统工程学会副理事长等职。同时担任国家自然科学基金委大数据重大研究计划指导专家组组长，主持国家自然科学基金委重大项目等多个国家级科研项目，以及多个国际合作、企业信息战略和管理项目。主要研究与教学领域为商务智能与电子商务、IT战略与管理、模糊逻辑与数据模型。曾获国际模糊系统协会2009年度“IFSA Fellow”；复旦管理学奖基金会2007年度“管理学杰出贡献奖”；1999年度国家杰出青年科学基金等荣誉。

精彩推荐 原创精华

精神的力量 新时代之魂

探讨“中国精神”的时代内涵

奋力谱写云南发展新篇章 最美乐章 大美彩云南

[不忘初心牢记使命]国家民委:努力做好新时代民族工作

全国公安政务服务实现全网通办 群众万千事 监督一张网

十万大军睡马路：何曾见过这样秋毫无犯的军队

做好当前经济工作:奋发有为,努力实现全年主要目标任务

关于生命教育的补习班 在情感共鸣中给出思考和答案

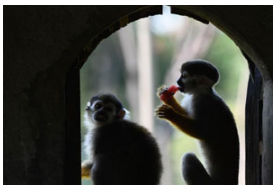
军运会火炬传递启动仪式南昌举行 一次庄严的火炬传递

光明图片

台风“韦帕”带来强降雨

在马出生的第二只大熊猫获名“谊谊”

烈日下的货运列车“体检员”



动物清凉度夏

7月9日，大学生们在安徽淮南市大数据展示中心参观。新华社发

参观者在位于济南市的山东省档案馆推出的山东省大数据科普主题展上体验基于5G网络传输的VR全息眼镜。新华社发

贵州铜仁市一家蔬菜公司的工作人员在管护蔬菜。通过大数据云平台，蔬菜公司可以根据订单来决定种植品种和规模。新华社发

大数据时代的两个阶段

我们现在处在一个数据的海洋当中。

2019年的春运被媒体戏称为“世界上最大的人口迁徙”，有30亿人次流动。2018年“双十一”网购达到了2135亿元的销售额度。现在，每天会产生450亿的微信条目。用手机的网民已经达到8.17亿。总体来说，我们国家的GDP数字经济占比已经达到了34.8%，超过了1/3，这方面实际能够体现出，我们这个社会已经开始越来越数字化了。

说起大数据、大数据时代，主要的时代背景是什么呢？我们现实世界有多大程度上可以被数据表示？用一个形象的话来讲，我们的社会像素正在急剧提升。这个“像素”来自到处可见的感测设备——探头、智能手机、可穿戴设备、车载设备，林林总总。这些使我们这个社会的数字化程度越来越高，数据的粒度因此也越来越细。也就

是说，数字化生活的两个要素之一：像素、数据的粒度已经具备。像素够高的时候我们要干什么？形象地说就是“成像”，就像手机、相机，像素越高成像的质量可能越好，因此，成像是我们数字化生活中另外一个重要的要素，像素和成像对应起来，就把数据和算法联系起来了。这就是我们所说的大数据的时代背景。

我认为，大数据时代可以分成两个阶段。

第一阶段是数据商务阶段。不断地把现实生活中的要素，人财物，都进一步数据化，同时根据这些数据化的人财物进行算法的应用。

第二阶段是算法商务阶段。当像素足够高的时候，重点就变成了成像了，也就是说，重点变成算法应用。

数据商务阶段和算法商务阶段都围绕着数据和算法进行，但是重点有所不同。数据商务阶段就像做菜一样，数据化的过程就是不断准备材料的过程，不停地增加和丰富材料，然后根据已有的材料提供不同的菜品。但是在算法商务阶段，材料已经足够丰富了，这个时候要比的就是手艺了，你是不是能够做得更好、更多。这就是我们所说的算法进阶及应用创新，如“智能+”，我们可以用更加高尖的智能技术，包括人工智能的很多技术，在现有的大规模数据下进行应用。

大数据的数据特征

那么，什么是大数据？首先看它的数据特征，可以从4个维度来理解，即4V：volume（规模）、variety（多样）、value（价值）、velocity（速度）。大家对这四个维度没有什么大的争议，但是对它们的含义的理解还是有相当不同的认识的。

第一是规模，我们称之为超规模。大数据规模会很大，但是没有绝对的量纲标准，没有说一定要达到多少G多少P多少Z才是大数据，这个不一定，因为大数据的大规模和问题、领域有关。只要这个大的规模超出了这个领域和问题的传统边界，那就是大规模里的超规模。

第二是多样，即富媒体的意思。现在80%～90%的数据都是文本、语音、图像、视频，不再是特别传统的二维的整齐的结构化的数据了。

第三是价值。我们处在数据的海洋中，四周都是数据，但是跟我个人有关，跟我企业有关的那种有价值的信息相对少了，因为数据量的分母太大了，即密度在降低，这个后面直接的隐喻就是要深度挖掘才能发现我们希望的价值。

第四是速度。数据就像开着的水龙头一样，源源不断地出来，而不是我们上传下载图片要等很久。因此，大数据里的数据是一个流数据的概念。

大数据的问题特征

那么，什么样的问题才是大数据问题？这要看它的问题特征。

第一个特征，是粒度缩放。粒度缩放是指我们碰到的这个问题的要素一定是数据化的，即这个要素不管是宏观的还是微观的，一定能通过数据表示。同时，可以像地图一样，可以在特别大的范围和特别细的范围之间缩放，能够在宏观、微观之间进行映射。

第二个特征，是大数据外部性导致的特征，称之为跨界关联。考虑问题的时候要看视角，问题边界是在哪儿，如果考虑问题的时候这个边界到了传统边界之外，就是跨界了，而且你把这个外部的要素和内部要素联系起来了，所以你在关联。

第三个特征，全局视图。大数据实际是希望了解全貌的，它最后是要看画像，因为前面我的每一个点、每一个环节的数据叫作粒度缩放，同时和我相关的要素我又关联了，但是我最后要干什么，要了解全貌，要有个人画像、企业画像、政府画像、社会画像等，所以这个画像本身又是全景式的，从范围来讲是全景式的，从内涵来讲，我们希望既关联又因果。

这里，我举一个共享单车的例子，方便大家审视大数据问题的特点。有的人会认为共享单车其实就是我们的代步工具，但是这是传统的概念。现在一般每辆共享单车都有自己的感应器和定位装置，也就是说感测的数据粒度到了车和部件。这时候就不单是一个单车了，可能我走到什么地方，共享单车的App就告诉我附近有什么商圈、酒店、餐馆，我在什么地方买东西可能还可以用移动支付，当视角从单车走到了其他行业、要素时，就开始跨界关联了。可能在这个地区人特别多，共享单车不够，可能在另外的地方单车冗余了。因此，共享单车的平台应该清楚什么地方需要车，什么地方不需要车，怎样调动，这就是全局视图。当共享单车具备粒度缩放、跨界关联和全局视图时，共享单车的运营、优化，就是一个大数据问题。

这些年来，社会上比较流行一个论断，说“大数据只讲关联不讲因果”。这个论断虽然有一定道理，但是总体来讲是误导的。特别是在重要决策的时候，如果涉及的后果可能会有严重的人财物的损失，然后我告诉你“你就这么干吧，没有为什么”，谁敢作决策？所以，在大数据环境下作管理决策，既要看关联也要看因果。另外，因果是认识论的基本诉求，我们要知道原因。

大数据冲击各行各业

我们作为个人不仅是数据的接收者，也是数据的生产者。一方面我们下载、阅读、浏览，因此我们在消费数据；另一方面，我们又上传、撰写、参加各种活动，各种活动就可以留下我们的很多痕迹，因此我们也在留痕，我们实际又在产生数据。在这样一个既是消费又是生产的环境中，我们从方方面面已经和数据分不开了。

大数据已经在冲击各行各业。

比如经济金融领域。股价的预测其实一直是个难题，传统的股价预测，实际是通过一些专业的模型来估计风险、收益、评价企业，有专门的理论和方法来估计股价。但是影响股价的除了这些因素之外还有人们的“期望”，而估计“期望”是非常难的，因为“期望”既涉及外部因素，又涉及心理预期。现在一个新视角是考虑公众关注，比如搜索。若对某些企业比较关心，可能就搜索其企业状况、新闻事件，这种搜索体现了大众对具体企业的股票价格和价值走向的关心。这是一个跟过去特别不同的角度，因为这不是特别专业的角度，它是从专业外人士的行为来估计的角度。这种关注和搜索与股价的走势有相当强的关联度。但是，要特别指出，仅用这一个因素来估计股价是不够的，还有大量的因素需要专业模型。因此，一方面能够扩展或者冲击传统的定式和视角，另外应该把其他视角引入进来，大数据的股价预测应该是包括内部与外部、专业与非专业因素的模型构建。

大数据也开始在改变会计学。传统的会计学衡量企业的状况是通过三张报表：资产负债表、现金流量表、利润表，这三张报表反映了一个企业的运营能力、偿债能力和盈利能力。虽然这三张报表是非常基础和非常重要的，但是大家突然发现，有一大类企业是高风险的，特别是一些IT企业、创业企业、新行业企业，长期负债，但同时又有非常高的市值，人们又有非常强的忠诚度，如果用这三张报表衡量，似乎不能完全体现它的价值，也就是说，传统会计学的三张报表现在可能就不够用了。因此，人们正在呼唤“第四张报表”的出现，业界和学界都在做这方面的研究。长周期、高负债、高不确定性企业的价值可能受到的是口碑、忠诚度、品牌、公允价值，包括无形资产的影响。这些东西，我们可以称之为数据资产。

大数据也在为体育界带来变革。篮球项目像美职篮NBA就做得非常好，他们通过收集肌肉、血液、心脏、动作、战术、团队等全景式的数据来帮助训练和比赛，因为这些因素，都有可能影响整个比赛的结果。科技体育这几年有巨大的空间，传统的师傅带徒弟，师傅的传帮带确实非常重要，但是应该有更细粒度，更加多角度、更加全景式的手段，采用大数据技术来提升整体的竞赛水平。

大数据在艺术上也有很多影响。传统绘画，不管是古典的还是现代的画，都有自己的素材和表现形式。现在出现了一种新的素材——数据素材，也就有了新的表现形

式。比如飞机航班的数据轨迹就可以构成一幅新颖的画。

大数据在其他领域也有非常多的应用，比如农业方面就有蔬菜革命、精准扶贫。在医疗健康领域，医院内医院外，得病和未得病之间的关联，也是大数据问题。文学上通过大数据技术对一些词语、作者、关系、背景等进行分析。这些都是利用大数据的例子。

哲学里一个重要的方向是认识论和方法论，这里包括我们近些年提炼出来的新的研究成果。传统的哲学认识论追求探索因果关系，因此基本叫作模型驱动范式，也就是说通过刻画变量之间的联系，比如自变量和因变量，通过构建这两个之间的函数关系，比如线性、非线性等，可以知道一个自变量一个单位的变化会导致因变量有几个单位的变化，这里试图反映变量之间的逻辑的因果上的机理。但是，这个模型驱动的范式，在大数据时代会受到一些挑战，或者说它碰到一些问题时会捉襟见肘。比如，当数据变量的组合数特别多时，当很多变量是潜变量和隐变量时，当很多变量虽然重要，但是不可测不可获时，还有当数据的样本规模特别大时，这些问题用传统的模型驱动做法就会比较困难。因此，就出现了一个新的范式转变，催生了大数据驱动范式。这个范式想表达的是，对于管理决策，我们希望能够实现既有关联又有因果的诉求，这个新范式简单地说由外部嵌入、技术增强和使能创新三方面构成。外部嵌入是指引入视角之外的变量，有些变量我们知道重要，但是没有办法放进模型里，比如我知道股价，我预测股价有个计量模型，但是如果今天这个公司出了一件事情，或者是有个关联新闻，或者行业里有个新的政策，我们觉得可能会影响股价，但是这些变化很可能是视频、语音或者文本，没有办法融入传统的模型中去。所以，需要引入外部视角。这些图像、视频、新闻文本要引入进来，就是要使得我们引入的变量可测、可获，这就需要技术上的增强。当这些变量引入进来的时候，我的变量空间就发生了变化，这时候我们可能会研究新的X到Y的转换，也就是变量关系和映射要重新定义和审视，这就是使能创新。

历史学其实也和大数据密不可分。传统的历史记录内容都是帝王将相、英雄豪杰、国家、政治、重大的军事事件等，很难在历史中看到平民和我们自己。一个是过去的粒度不够，第二手段也不行，存不下来。大数据环境下就可能自下而上反映历史。比如国家图书馆互联网信息战略保存项目，就是和新浪网合作，把新浪公开的相关博客文章作为历史资料记录下来，通过自上而下与自下而上的史学观的融合，能够让我们在更细粒度上反映历史和社会，同时也可以获得更加全面的历史画面。

法律也和大数据相关。比如说，我作为一个消费者，在网上购物、浏览，我的网络痕迹、数据脚印都被相关公司采集了，那么，我有没有权利要求你把我的这些痕迹抹掉、遗忘掉？这就是“被遗忘权”。所谓被遗忘权是指数据主体有权要求数据控制者永久删除有关数据主体的个人数据，有权被互联网遗忘，除非数据的保留有合法的理由。2018年欧盟出台了《通用数据保护条例》，强调了被遗忘权，我们国家2018年的高考II卷一篇阅读文章的题目，也是要考生来思考、评论这个被遗忘权的问题。

大数据与人工智能的交会

大数据的冲击力量现在看来还在加剧，其中有一个力量非常值得关注，那就是人工智能。

当人工智能遇到大数据的时候，现在井喷式的发展才变成了可能。其实人工智能是现在这个时代中很多技术的一类，它本身已经发展了好几十年，但是为什么在近些年才得到快速发展？其实人工智能技术和这几个关键词有关，那就是“学习、训练、推理、演化、智能、智慧”，也就是说，它是关于这些关键词的一类技术。特别重要的一点，它要根据大量的数据来进行学习和预测，就是从数据中学习，建立模型，并用于预测未来。过去数据的粒度不够，进入大数据时代，当数据有足够的粒度和像素时它才成为可能，因为人工智能的主流技术首先是要基于大规模数据进行学习。其次，人工智能算法本身需要非常强的计算能力，只有在大数据时代，有了云计算平

台、数据传输、数据的流通、数据的管理，诸如5G技术等，才能为人工智能的发展提供非常好的支持。我们身边其实已经有很多人工智能产品了，比如工业机器人、财务机器人、下棋机器人、能做诗作画作曲的机器人等，这些机器人可以做很多我们过去认为不可能的事情。

人工智能在未来会波涛汹涌，一浪高过一浪地发展。但是它本身也有局限，目前的人工智能技术特别是深度神经网络这样的技术，基本上属于“黑盒子”技术，可以算得非常准，但是“为什么”还说不大清楚。在这种情况下，在一些重要的应用领域就受到局限，因为如果不知道“为什么”就不敢用这个方法作重要决策，如果不能通过非常清楚的机理来说明，实际它未来的应用也是有局限的。现在，业界和学界都在攻关“可解释人工智能”，实际就是人工智能在输入和输出之间，在数据和预测的结果之间，从数学上来讲需要一点定理，一些形式化的机理。从认识论上来讲需要一些因果关系。

不管怎么说，人工智能的应用已经深刻地影响到我们了。作为人类，我们自己创造了一个“亚种”叫作机器人。机器人的行为是不是都在我们人类的设想之中呢？会不会干一些我们想不到的事情呢？似乎这个担忧是必要的。所以机器行为学应运而生。传统社会学、管理学、经济学、心理学等都是研究人、由人构成的组织的行为，由人形成的网络的行为。随着各式各样的机器人越来越多地替代人的工作，越来越多地挑战人们在智力、计算上的能力，这样的研究是非常必要的。所以，我们要研究机器如何塑造人类的行为，人类如何塑造机器的行为，以及人机协作的行为。

运用大数据要重视商业伦理

在实际中，大数据的使用本身仍有很多令人担忧之处。虽然科技发展飞速，但是人们使用科技是带有价值取向的。

比如“大数据杀熟”。在传统的营销、管理中，我们都希望了解客户的行为，更好地为他们服务。在市场的环境下我们也说，既然有人愿意用高价买，那就可能要给他提供更好的服务。但是在大数据环境下，这种处理就有一个度的问题。第一是客户是否知道他的信息被收集，第二是他是否愿意真的出高价买。作为企业来讲，又需要有经营哲学上的思考：企业是以盈利为中心，还是以客户为中心？当以客户为中心时，客户满意与否就变成了主要的KPI，就是主要的决策考量，如果光考虑企业的盈利，而不考虑客户，可能就不太会考虑用户的感受。实际上，“大数据杀熟”涉及的是商业伦理层面的问题。

在大数据时代我们跟数据打交道会碰到一系列社会问题、法律问题、道德问题，需要在企业层面、商业层面，在社会和政府层面立法立规，在个人层面、在道德的层面大家来共同努力解决这些问题。

感测和响应大数据时代

过去的20年我们经历了特别大的技术变化。20年前，中国网民是62万，互联网普及率只有0.03%，网站1000多家。现在中国网民有8.29亿，互联网普及率达到59.6%，网站523万个，上网时间每天人均4小时。

时代的变化太快，我们应该敏锐地主动地感测和了解这个变化，同时不管是企业还是个人，要作出自己的准备和自己的响应，因为大数据作为一个时代会伴随我们相当长的时间。在未来的某一天，可能由大数据衍生出一个新的概念、一个新的内涵、一类新的技术，可能会变成一个新时代的符号。

返回目录 放大 缩小 全文复制