大数据：颠覆的力量

陈国青

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。

大数据已经成为了我们耳熟能详的词汇和概念。实际上它已经变成了这个时代的符号，这个符号是什么样子？又意味着什么？

我想围绕若干个对于大数据的认识，诠释什么是大数据、大数据时代，以及我们置身其中所要遇到的、见到的、面对的那些变化、冲击、挑战。同时，我们也可以从中发现和把握大数据带来的机遇以及可能的创新和发展空间。

英国脱欧、美国大选有数据公司的影子

大数据已经提出若干年了，大数据本身在概念和应用上已经与我们越来越密切了。在开始大数据这个话题前，先从两个国际事件说起。

一个是纠结了很长时间的英国首相要辞职了，另外一个是美国不断“退群”，最近又四处打贸易战。实际上这两个事件都是由三四年前两个转折性的事件引发的，这就是英国脱欧和美国大选。这三四年来，大家已经看到这些事件在他们国家带来的社会撕裂、对世界格局的冲击，以及对我们比较熟悉的国际秩序、多边关系的挑战。

非常巧的是这两个事件的背后都有一家公司的影子，就是英国剑桥分析公司，简称 CA。这家公司是一个数据公司，它用数据做选民分析，做助选服务。这家公司主要做的是心理画像，它根据收集到的大量数据，包括千万级的 Facebook 数据，采用国际心理学界比较有名的 Ocean 心理模型来刻画选民的人格特征。它能刻画一个人喜欢什么、担忧什么、对什么感兴趣、宗教的取向以及价值的态度。用公司 CEO 的话来讲：“我们可以预测每个美国成年人的人格特征”。其实这家公司并不是那么出名，还有更出名的公司也在选举中提供各种服务，但是其他公司的服务基本都基于人口统计学的信息，也就是性别、肤色、宗教、收入、年龄、教育等信息，但是这家公司从心理的视角刻画一个人的心理历程或者叫做数字脚印，这个视角是别的公司没有的。另一方面，他们可以获得更加细粒度的数据，使得心理层面、人格层面的刻画成为可能。这就是我

们所说的大数据的影响。

当然，这家公司的影响是有限的，一个选举、一个大的事件还受诸多其他因素影响，但是这些因素的叠加造成了我们现在看到的，包括这几年演化出来的世界格局的改变。

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 年的春运是世界上最大的人口迁徙，有 30 亿人次的流动。2018 年“双十一”有 2135 亿的销售额度。现在，每天会产生 450 亿的微信条目。手机网民已经达到 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？现实世界在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。数字化生活的两个要素之一：像素，其数据的粒度已经具备。像素够高的时候我们要干什么？形象地说是“成像”，就像手机、相机，像素越高成像的质量越好。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时代。

大数据时代可以分成两个阶段，我们用商务的形态来说明这个问题。

第一阶段是数据商务阶段。不断地把现实生活中的要素都进一步数据化，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，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。但是算法商务阶段材料已经足够丰富，这个时候要比的就是手艺，看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

什么是大数据？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，但是对它们含义的理解有相当不同的认识。

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，没有说一定要达到多少 G、多少 P、多少 Z 才是大数据。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该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。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。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，源源

不断地出来，不需要我们等很久。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，一定要可以通过数据表示。同时可以像地图一样，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。而且你把外部要素和内部要素联系起来了，所以你在关联。

比如管理学中，传统企业管理最常见的就是怎么把业务流程做好，优化流程，提高质量，同时改进人力资源环节、财务环节，制定企业战略。这个过程是站在企业内部看不同的部门，往左看一看是供应商，往右看一看是客户。

企业花了很多努力，突然有一天一个人在网上发难，说企业产品有问题、服务不好。还没容得企业进行辩解，成百上千万跟贴瞬间就把企业的产品、形象、品牌定格成了某一个形象、某一个状态。企业可能很委屈，觉得这些人既不是我的客户，也不是我的员工，他们只是原本跟企业没有联系的社会大众，但是他们的口碑却对产品、质量、品牌、形象产生影响。因此，当管理决策的视角不仅是考虑内部，而且要考虑外部和企业相关的因素时，这个问题就开始变成大数据问题了。你要跨界，跨出你的传统边界。

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。前面每一个点、每一个环节的数据叫做粒度缩放，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，这个画像本身又是全景式的，从内涵来讲，我们希望既关联又因果。

举两个简单的例子看一看大数据问题的一些特点。

一个例子是旅游。比如某个景点，经常在一些时间人满为患，有时候服务跟不上，可能出现游客不满意投诉的现象。作为景点管理方，如果想改进，可以增加员工、提高运力，但是增加了车、人，有可能下一个时间段没有那么多人来。所以如果我们仅从景点这个边界出发来提供优化的方案，解决质量的问题就很困难。实际上如果要解决景点的问题，一定要走到景点之外，比如旁边的餐饮、酒店、交通、气象如何，以及附近有没有其他景点、其他活动，当我们考虑了景点内外要素时，就有了跨界关联的属性，作为整个的旅游我们来看景点内外时，我们有这样一个全局的视图，我们面对的就是一个大数据问题了。

另外一个例子是共享单车。有的人认为共享单车是我们的代步工具，这是传统的概念。现在每辆共享单车都有自己的感应器和定位装置，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。可能这个地区人特别多，共享单车不够，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，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，如果涉及到的后果可能会有严重的人财物的损失，我告诉你“就这么干吧，没有为什么”，谁敢做决策？所以在大数据环境下做管理决策，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

作为个人，我们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”。估计“期望”是非常难的，因为“期望”既涉及外部因素环境，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于具体企业的股票价格和价值走向的关心。这跟过去特别不同，因为这不是专业的角度，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。

这和几年前谷歌通过搜索来估计流感是一样的，它不是采用所谓流行病学的专业模型，而是某段时间很多人有症状，头疼、咳嗽、吃了什么药，大家有很多信息的交换，这种交换的强度、交流的走向，恰恰可能跟流行病的流行模式非常相关。所以，我们也可以从搜索的角度来估计它和股价之间的影响和关联度。但要特别指出，虽然搜索和股价的走势有联系，但是只凭借这一个因素来估计股价是不够的，还有大量因素需要专业模型。因此，一方面能够扩展或者冲击传统的定式和视

角，另外应该把其他专业视角引入进来，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，同时有非常高的市值，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可能不够用了。因此，人们在呼唤“第四张报表”的出现，业界和学界都在做研究。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。所以，这是从会计学的角度来看我们碰到的一个冲击，很多新的现象导致呼唤新的模型、新的理论框架出来。

大数据也在为体育界带来变革。现在我们都在积极筹备冬奥会，国家有少数冰雪项目水平比较高，但总体水平不是特别高。主要的问题是长期的传统做法比较粗犷、比较经验型。冰雪项目中有一大类是姿态类项目，运动员的关节、角度、力量和跳跃的高度、旋转的速度与动作的完成质量密切相关。现在大家已经意识到这个问题，有些队会用手机拍照片、视频，但是数据粒度没有到关节这项，也没有到姿势、力量和角度上，所以数据粒度不够。第二，视角也不够，可能需要更加专业的采集设备，更加专业的还原设备来完成。比如现在简单的二维的图像应该变成动态三维的还原，并且可以分解，这样就能帮助总体的竞赛水平得到提高。

别的项目像篮球，NBA 就做得非常好，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。垒球、网球的角度、落点、战术都有不同的大数据分析。可见，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。由数据作为素材，有新的视角进来，作为一个整体的新型创作出现。

大数据已经影响到经济、管理、体育、艺术

等领域，在其他的领域也有非常多的应用，比如农业就有蔬菜革命、精准扶贫，这些都是利用大数据的例子。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。

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，因此叫做模型驱动范式。也就是说通过刻划变量之间的联系，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，当很多的变量虽然重要，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的做法就会比较困难。

因此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个事情，或者行业里有新的政策，我们觉得可能会影响股价，这些变化很可能是视频、语音或者文本，没有办法融入到传统的模型中去。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，这就是第二条，技术上要增强。当这些变量引入进来的时候，变量空间就发生了变化，这时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。这是大数据驱动范式框架的三个方面。

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，和新浪合作，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，使得我们可以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。比如下载一个 App，凭什么问我要这么多权限？我不给权限行不行？没有办法，不给就不能用。我在网上购物、浏览，我的痕迹、数据脚印，都被公司采集了，用户有没有权利要求公司把这些痕迹抹掉、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，国家 2018 年高考 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。这也是由大数据激发出来的新问题，对传统的法学研究产生了新的挑战，或者说带来了新的发展空间。

人工智能的难点是黑盒子问题

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，现在井喷式的发展才变成了可能。人工智能是现在这个时代中很多技术的一类，它本身已经发展好几十年了，但为什么在近些年才得到了快速发展？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。过去为什么不行呢？比如本来想学一个圆，但是过去的数据只有一个半圆，它怎么能学出圆呢？人工智能的主流技术首先是要基于大规模数据进行学习，所以进入大数据时代，当我们的数据有足够的粒度和像素时它才成为可能。

其次，人工智能算法本身需要非常强的计算能力，也就是算力，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，才能为进一步的大数据应用创造条件，为人工智能的发展提供非常好的环境和支撑。现在可以看到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、作业机器人、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，目前的大数据技术特别是深度神经网络这样的技术，基本上属于“黑盒子”的技术，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”就不敢用这个方法做重要决策，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。

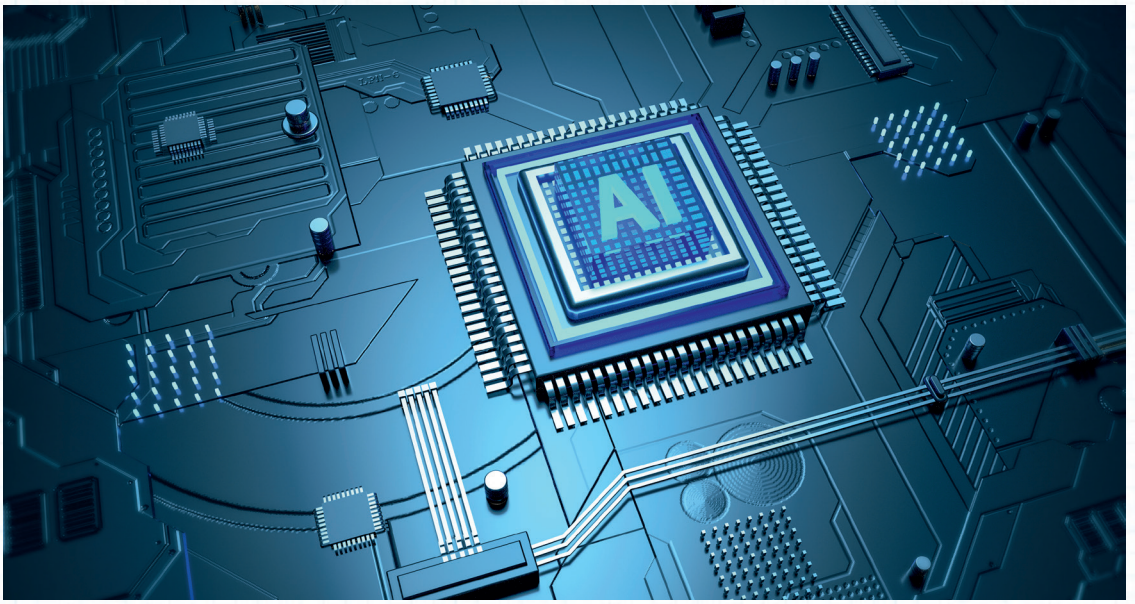
现在这么热闹的人工智能，很多都是过去成果的工程化和产品化。它本身的理论突破，包括提到的可解释性，也是大家在未来关注的重点。

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”叫做机器人。机器人的行为是不是都在我们人的设想之中呢？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。当人知道和我们打交道的是机器人时，人到底会有什么不同？机器如果只是模拟人的行为，那么我们用不用担心它会做一些其他的事情？当人和机器人一起互动时，会不会有其他的一些问题出现？这些问题实际是很革命性的。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出现在我们身边，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，这个担忧或者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。最新的《自然》杂志上有一篇文章也是呼唤学界、业界关注机器的行为以及机器和人的行为。

运用大数据要重视商业伦理

实际大数据的使用本身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、管理里面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，这种处理有个度的问题。第一客户是否知道他的信息被收集，第二他是否愿意真的出高价买。作为企业来讲，又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。实际上大数据杀熟是在商业伦理层面的问题。

还有刚才提到的 App 权限的滥用，以及数据的泄露，回到我们最开始提到的剑桥分析公司的例子，当时的 CEO 说“在美国所有的数据都可以买到”。因此这家公司 2018 年就陷入了数据泄露和商业伦理的丑闻，最终关门了。不管怎么说，在大数据时代我们跟数据打交道就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。技术发展特别快，这些问题的出现也变得越来越重要，我们应该有紧迫感，来更好地面对这些问题。

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%，网站一千多家。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，每天人均上网时间 4 小时。

在这样的时代中，简单地总结一下，我觉得就是两个词，“感测”和“响应”。时代的变化太快，我们应该敏锐地主动感测和了解这个变化。同时不管是企业还是个人都要做出自己的准备和响应，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、新的内涵、一类新的技术，可能会变成一个新时代的符号，所以当下我们要面对大数据，未来我们要融入新时代。

（本文为作者 6 月 10 日在人文清华讲坛发表的主题演讲）