大数据时代的解决方案

块数据 4.0

人工智能时代的激活数据学

大数据战略重点实验室◎著

中信出版集团

块数据

4.0

人工智能时代的激活数据学

大数据战略重点实验室◎著

中信出版集团 · 北京

图书在版编目（CIP）数据

块数据 4.0：人工智能时代的激活数据学 / 大数据战略重点实验室著. -- 北京：中信出版社，2018.5

ISBN 978-7-5086-8886-2

Ⅰ. ①块… Ⅱ. ①大… Ⅲ. ①经济管理－数据管理－通俗读物 Ⅳ. ①F2-39

中国版本图书馆CIP数据核字（2018）第 073607 号

块数据 4.0——人工智能时代的激活数据学

著　　者：大数据战略重点实验室
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲 4 号富盛大厦 2 座　邮编　100029）
承 印 者：北京通州皇家印刷厂

开　　本：880mm×1230mm　1/32　　印　　张：11.5　　字　　数：230 千字
版　　次：2018 年 5 月第 1 版　　印　　次：2018 年 5 月第 1 次印刷
广告经营许可证：京朝工商广字第 8087 号
书　　号：ISBN 978-7-5086-8886-2
定　　价：59.00 元

编撰委员会

眼看乾坤一局棋，满枰黑白子离离。

铿然一子成何劫，唯有苍苍妙手知。

这是被称为“波斯李白”的诗人奥马尔·海亚姆（1048—1122）的《鲁拜集》中的诗句。作为一名精通天文和数学的大学者，奥马尔认为，宇宙的规律是可以探知的，并可以用严密而美妙的数学方式表示出来。前定与随机，必然与偶然，向来是人文科学中长期争论不休的命题。自然科学理论始终受实验和观测的检验，而它的每一个重大发现又都会反馈到文化和社会的层面，对人的哲学和历史观有所启示。

决定性和概率性一直被当作数学、物理等学科对自然界的描述方式。在牛顿创立古典力学之后的250年间，直至20世纪20年代，决定论长期处于主导地位，基于概率论的统计描述或者说数据的描述，则一直属于不得已情况下所采用的辅助手段。决

定性的牛顿力学从计算和预测的观点来看，实际上也具有内秉随机性，这就是微观层次上的混沌运动。大量隐藏在暗数据背后的某些看似简单原因所导致的复杂后果，则渐渐成为混沌研究的重要信息。混沌不是无序和混乱。与人们习以为常的周期排列或对称形状的数据相比，大自然和人类社会中的很多数据其实就是一种没有周期性次序的混沌。在理想模型中，它可能包含着无穷的内在层次，层次之间存在自相似性或不尽相似。在观察手段和技术的分辨率不高时，只能看到每一个层次或某一种类型的结构。但技术条件改变或提高后，在远离不能识别之处就会出现更小尺度上的结构。零维的点、一维的线、二维的面、三维的体和四维的时空，是人们现在所能认知的数据空间。如果在不远的将来，我们真的进入一个超数据时代，现有的技术和描述手段也许就无法对这种高度无序数据的混沌运动进行分形，而关于相变和临界现象理论的框架也需要一个新的重构。

这时我们不由得想到那个著名的洛伦兹“蝴蝶效应”理论，其实和这个理论相联系的还有一个被称为“湍流发生机制”的观点，认为向湍流的转变由少数自由度决定，经过两三次突变，运动就到了维数不高的奇怪吸引子上。这里所谓的吸引子是指运动轨迹长时间之后的终极形态，它可能是稳定的平衡点或周期性的轨道，也可能是继续不断变化、没有明显规则或次序的许多回转曲线。无论是蝴蝶效应还是湍流发生机制，其实都是对我们现在正在研究的激活数据学的一种理论上的关照和呼应。事实上，大数据乃至超数据时代的数据运动，就是这样一种处于混沌和分形

之下的对数据运动轨迹及其规律的研究。许多看起来杂乱无章、随机起伏的数据变化或时空穿越，可能造成的就是类似亚马逊级别的数据风暴。如果说上述蝴蝶效应粉碎的是本就无法实现的长期天气预报的幻梦，那么紧接着的奇怪吸引子告诉我们的是，人类对于天气的实际预报能力并没有因那只蝴蝶的翅膀而受到任何影响，相反，却因对于更加混沌的数据的研究而提高了。激活数据学就是一种基于复杂理论及混沌研究的关于未来大数据乃至超数据时代的理论假说，就像上面讲的天气预报，但它所关心的并不是下个星期的晴雨冷热，而可能是未来 10 年耕种季节的平均降水量和平均气温。激活数据学研究使以往根据统计原则所做的预报上升为数据动力学的预报，也就是应用了似是随机现象的内在规律，从而提高了预测单个轨道近期行为的精确度，并丰富了长期预报的办法。

同样，我们还可以考察一个似静实动的模型。让沙子从一个漏斗孔中缓缓落到桌面上，形成渐渐变大的沙堆，总有最后新添加的某一粒沙子会在整个沙堆勉强维持平衡的锥面上导致一次“雪崩”，使一撮沙子滑到堆底，雪崩留下的小洼地会被后续的沙流填平，直到下一次更大的雪崩。在我们收集所有这些雪崩的数据后，可以发现它们的大小和间隔遵循某些数据动力规律，而沙堆模型无疑也启发了我们对于数据激活状态中的相变和突变的研究。无论数据的平衡态的相变或非平衡的临界多么不确定，可以确定的是，在搜索、融合、激活和碰撞等一系列状态下，数据在某一个临界点附近的扰动必然会导致某种全局性后果。当然，是

否存在可以被“激活”的“数据蝴蝶”或“数据吸引子”，还需要我们进一步探寻，但这并不否定我们的所愿，而仅仅需要我们从实际数据的研究和挖掘中进一步加以发现。

作为一种理论假说，激活数据学就像一座朝向深邃的大数据宇宙的“天眼”。它是未来人类进入云脑时代的预报，是关于混沌的数据世界的跳出决定论和概率论的非此即彼、亦此亦彼的复杂理论的大数据思维范式的革命。从一定意义上来说，大数据就是面向未来社会人类需要破译的“基因”。正如因发明一种DNA（脱氧核糖核酸）快速测序方法而获得1980年诺贝尔化学奖的吉尔伯特针对生物学研究范式的变化指出的，“正在兴起的新的范式在于，所有的基因将被知晓，今后生物学研究项目的起点将是理论的。一位科学家将从理论的假设开始，然后才转向实验室去检验该假设”。是的，借助日渐深入的人工智能的发展，大数据的理论研究正在激发人类的新的假想和猜测。正是这种假想和猜测，让我们以某种“对称破缺”的方式去探知深邃未知的数据海洋，发现诸多社会发展法则背后产生影响甚至支配的物质和数字的力量。

人生是一种快变量，语言是一种慢变量，而数据将是一种突变量。虽然“未知”依然是现实的一部分，但是身处海量数据大爆发时代，人们坚信，未来已来！从“块数据1.0”到现在的“块数据4.0”，我们一直在持续探讨这个已来的未来，尤其是基于对“以人为原点的数据社会学的范式革命”的认知。事实上，从一开始，我们就没有把大数据仅仅看作所谓的“大”的数据，而是把大数据看作一种“活”的数据，因为只有激活，大数据才有生命，

才有社会属性，才能成为未来世界人们赖以生存与发展的土壤和空气。最后，套用《爆裂》一书中关于现代世界生存的九大原则中“系统优于个体”的表述：真正具有竞争性的是一个系统，而非一个特别强大的个体；是一套能够保证不断成功的制度，而不是一个天才个人的行为。同样，激活数据学就是这样一个思想的系统，就是要为我们身处的这个大数据时代找到一个解决方案，这个方案可以构建一个融合数据、计算和场景的系统，让我们在大数据的时空中真正“思考和行动”起来。世界正处于根本结构性变革中，我们必须具备这样一种能力，即下意识地适应和发现因不适应我们的旧习惯而被忽视的事情。

连玉明

大数据战略重点实验室主任

2018 年 4 月 3 日于北京

目 录

大数据时代的解决方案

本书探讨的主题是大数据时代激活数据学的提出、运行机理及场景应用。激活数据学是以充分发挥人机群体智能为核心，综合运用数据科学、生命科学和社会科学提出的海量数据存储、处理的解决方案。激活数据学将确立一个新的观察人类智能和机器智能的视角，引导人们重新审视数据无限膨胀可能造成的人类认知障碍，重新思考维持一个健康、安全和有效的数字社会的根本办法，建立与人类智能复杂性同步的人工智能系统，开启用复杂性系统思维认识未来世界和改造未来世界之旅。

大数据时代面临的问题与挑战

在人类文明的伊始，人与人的第一声交流即意味着“连接”的开始。语言使人与人连接，并促使用于记载事物的文字、数字

符号产生，这样的“连接”便产生了“数据”，并演化为人类文明最初的信息与知识。在漫长的农耕文明时代，“连接”主要以语言沟通和书面文字沟通的形式存在。进入工业文明时代，“连接”开始通过无线电台、电报、电视的形式存在，但这样的“连接”产生的信息往往是单向性且缺乏互动的。互联网时代，人和人开始通过网络进行复杂交错的互动连接。社交网站、电子邮件、搜索引擎、聊天工具……人类建立连接的方式趋于多样化、多维化，人类社会产生的数据也因而大量积累。与此同时，数据的价值越来越受到人们的重视。数据深刻作用于政治、经济、文化等领域，带来更多的创新机会，从生产、生活到科研，一个大数据时代正在开启。

在美丽的贵州省黔南州布依族苗族自治州平塘县，被称为“中国天眼”的世界上最大的单口径射电望远镜——FAST（500米口径球面射电望远镜）已于2016年9月25日落成启用。FAST的计算速度需达到每秒200万亿次以上，存储容量需达到10PB[①]以上。这一世界级的工程将帮助人们捕捉到更多来自宇宙的信息，它的背后是“天文级”的海量数据存储和复杂的计算。

随着时间的推移、科学任务的深入，以及数据的大量采集，未来对计算速度和存储容量的需求将爆炸式增长，数据量和计算量都将“大得惊人”。

数据是没有边际的，而计算力、存储力始终存在物理极限。

---

① $1PB=2^{50}B$。