大数据在体育科学中的应用及思考

马国全，杨建文，张虎祥，田 宇

（兰州理工大学 体育教学研究部，兰州 730050）

摘 要：近年来，大数据成为经济界、政府以及学术界热议的话题和研究热点。目前大数据在体育领域的研究处于起步阶段。采用文献研究、案例分析等研究方法，首先阐述大数据科学理念的缘起及本质特征，然后探讨大数据在体育科学中反兴奋剂、科学选材、赛绩提升等方面的具体应用实践，分析和思考大数据在体育科学应用中面临的问题与挑战。

关键词：大数据；反兴奋剂；科学选材；赛绩提升

中图分类号：G80-058 **文献标志码：**A **文章编号：**1008-3596（2015）02-0011-06

近年来，大数据引起了经济界、政府部门以及学术界人士的广泛关注[1]，大数据不仅在信息产业领域得到成功应用，也日渐渗透于教育、医疗、商务、传媒等各个领域，体育科学也不例外。笔者查阅文献发现，国内外的研究大都侧重于大数据的技术层面，针对体育科学具体应用的文献非常少。本文拟从厘清大数据的概念入手，继而重点探讨大数据在体育科学中的具体应用及作为，并对大数据技术在体育科学应用中存在的问题与挑战进行初步分析和思考。

1 认识大数据

科技的发展，使得人、机、物三元世界高度融合，带来了数据规模的爆炸式增长，世界已进入大数据时代[1-2]。现代科学研究面临的一个巨大挑战就是如何处理与日俱增的海量数据。近几年，《Science》和《Nature》等国际顶级学术刊物相继出版专刊探讨大数据的理论与实践。2008年9月，《Nature》专门就此问题以特刊、社论及评议等文章形式进行了讨论[3-5]。2011年，《Science》亦推出关于数据处理的专刊《Dealing with data》[6]，讨论了大数据对当前社会带来的挑战。美国政府认为大数据是“未来的新石油”，并于2012年3月宣布投资2亿美元启动“大数据研究和发展计划”（Big Data Research and Development Initiative），至此将对大数据的研究上升为国家意志。可以看出大数据将对未来的科技、经济、生活及文化等诸多方面的发展带来深远的影响。

在我国，大数据的研究与应用也得到高度重视。为了在大数据时代抢得先机，使科技、经济立于不败之地[7]，中国计算机学会于2012年6月成立了“大数据专家委员会”。2013年12月，中国计算机学会大数据专家委会发布了《中国大数据技术与产业发展白皮书（2013）》，在一定程度上反映了我国大数据学术界和产业界的共识。目前，大数据正处于方兴未艾、众说纷纭的时刻，学术界对“大数据”这一新兴的科学概念还没有明确的定义。中国科学院李国杰院士认为[1,8]：一般意义上，大数据是指无法在可容忍的时间内用传统信息技术和软硬件工具对其进行感知、获取、管理、处理和服务的数据集合（定义1）。

世界著名咨询机构麦肯锡公司于2011年5月发布《大数据：下一个创新、竞争和生产力的前沿》的技术报告，报告认为[9]：大数据是指其大小超出了典型数据库软件的采集、储存、管理和分析等能力的数据集（定义2）。

定义1和定义2所处的角度有所不同，定义1主要立足于信息科学，强调传统软硬件无法轻易处理的数据便为大数据；定义2的视角则更加宽泛，凡是原有“典型的”数据处理模式无法胜任的数据集合都可以称为“大数据”。

国际数据公司（IDC）站在信息科学的角度，将大数据的特征归纳为4个V，即Volume（体量浩大）、Variety（模态繁多）、Velocity（生成快速）和Value（价值巨大但密度很低）[10]。相比而言，被誉为“大数据时代预言家”的牛津大学教授舍恩伯格（Schonberger）等[11]则以不同的全局视野，归纳出大数据的三大特征，笔者认为其更符合体育科学领域的应用范围，即全体性、相关性和混杂性。这些特征也更切合麦肯锡所给出的大数据定义。

（1）全体性，即大数据旨在收集和分析与某事物相关的“所有”数据，而非分析少量的抽样数据。传统的小数据时代的随机采样，其核心是以最少的数据（抽样）推断出最多的信息。小数据时代的成功主要依赖于采样的绝对随机性，其精确性也随着采样随机性的增加而大幅提升。然而实现采

收稿日期：2014-10-21

基金项目：甘肃省体育社会科学研究项目（GST201450）；兰州理工大学科研发展基金项目

作者简介：马国全（1978—），男，河南罗山人，讲师，硕士，研究方向为竞技体育理论与方法。

样分析的绝对随机性是非常困难的，一旦采样过程中存在有偏性（biased sampling），或者说存有偏见，那么分析得出的结果就可能大相径庭。

（2）相关性，指两个或两个以上变量的取值之间存在某种规律性。在大数据时代，我们的研究思维要发生重要转变，即大数据鼓励我们更多地关注事物间的相关性（correlation），而非紧盯事物之间的因果关系（causal relation）。也就是说，在大数据时代，很多情况下，弄清楚“是什么（what）”比致力于找寻“为什么（why）”更为重要。

（3）混杂性，即大数据可接受数据的纷繁复杂，而不再单一地追求精确性。在小数据时代，人们总试图收集一些非常干净的、高质量的数据。在数据采集时，由于噪音数据（noisy data）的存在，导致数据存在混杂性通常是不可避免的。如果这些噪音数据是偶然的，那么它会被更多的正确的大数据淹没掉，达到“瑕不掩瑜”的效果；如果噪音数据存在规律性，足够的大数据分析则可协助我们发现这个规律，从而进一步地把系统性的噪音数据过滤掉。因此，在大数据时代，容许人们不必追求严格的精确性，可倾向于满足某种大方向的结论，而不是迷失于现象的点滴细节。这样，适当忽略微观层面上的精确度，可让我们在宏观层面拥有更好的洞察力。

2 大数据在体育科学中的应用

目前，大数据应用面临着许多挑战，其研究尚处于初级阶段，仍需挑战更多的研究领域以解决数据储存、数据挖掘及数据分析效率等方面的问题[12]。尽管大数据已成为信息产业炙手可热的流行词汇，在中国也已经被上升至国家信息产业战略层面，但大数据的实际应用才刚刚开始，还没有大量的实际应用成果出现在现实生活中[13]，在体育科学中亦如此。下面针对大数据的三个特性（全体性、相关性和混杂性），笔者将分别对大数据在反兴奋剂、科学选材和赛绩提升三个方面的应用实践实施案例进行分析，以拓展大数据在体育科学中的创新应用。

2.1 大数据在反兴奋剂中的应用

2012年7月28日，我国游泳运动员叶诗文以4 min 28 s43的成绩夺得伦敦奥运会混合泳400 m的金牌，并打破了该项目的世界记录。然而西方媒体甚至学术期刊却纷纷发文质疑叶诗文成绩的有效性[14-16]。例如，2012年8月1日，世界顶级学术刊物《Nature》对此议题亦发表自己的看法。Ewen Callaway在《Nature》官方网站撰文《超凡奥运成绩为何会引发质疑》[15]，文中配发叶诗文奥运泳池比赛的照片，暗示叶诗文成绩是“异常的（anomalous）”。该文刊出后，在世界范围内引起广泛争议。

该文[15]对叶诗文成绩的有效性提出两点质疑：①叶诗文在混合泳400 m的最后50 m的成绩（28.93 s）甚至比美国运动员罗彻特（Ryan Lochte）的男子400 m混合泳的最后50 m成绩（29.10 s）还要快很多，这不符合常理。②叶诗文前后两次大赛（即2011年游泳世锦赛和2012奥运会）的成绩在短时间内提升太快，居然获得近7 s的成绩提升，这也是不正常的。

如前文描述，在大数据时代，我们需要分析更多的数据，有时甚至需要与某个现象相关的全部数据，而不是依赖于数据的随机采样。事实上，该文所谓的“正常”数据仅为除叶诗文之外的2012年伦敦奥运会男、女子400 m混合泳决赛的运动员，样本总数也仅为15个[17]，如此小的数据集合，难免是有偏的，而有偏的数据推出的结论也势必存在误导性。

表1 以一年（12个月）为时间窗口400 m混合泳成绩提升前32名

提升名次	运动员姓名	提升成绩/s	提升名次	运动员姓名	提升成绩/s
1	Grainne Murphy	19.97	17	Vasilii Danilov	6.61
2	Dmitriy Gordiyenko	14.13	18	Roberto Pavoni	5.91
3	Stephanie Rice	12.14	19	Alexa Komarnycky	5.45
4	Anja Klinar	11.55	20	Travis Nederpelt	5.18
5	Mireia Garcia	11.04	21	Thomas Fraser-Holmes	5.00
6	Katinka Hosszu	10.73	22	Stina Gardell	4.85
7	Ting Wen Quah	10.57	23	THOMAS Haffield	4.75
8	Camille Muffat	10.19	24	Jing Liu	4.61
9	Dinko Jukic	9.60	25	Joerdis Steinegger	4.59
10	Alessio Boggiatto	8.75	26	Luca Marin	4.59
11	Laszlo Cseh	8.74	27	Tanya Hunks	4.22
12	Brian JOHNS	8.35	28	Yana Martynova	4.12
13	Mateusz Matczak	7.67	29	Andrey Krylov	4.10
14	Barbora Zavadova	7.59	30	Federico Turrini	3.56
15	Julie Hjorth-Hansen	6.88	31	Hannah Miley	3.34
16	Shiwen Ye	6.72	32	Samantha Hamill	3.25

其实，即使仅仅看混合泳最后 50 m 的冲刺表现，也早有女子超过男子的案例。例如，2011 年在上海举行的上海游泳世锦赛 800 m 自由泳的比赛中，来自英国的女子运动员瑞贝卡·阿德灵顿（Rebecca Adlington）最后 50 m 的 28 s91 的成绩，不仅超过叶诗文最后 50 m 的表现（28 s93），而且也超过罗彻特的表现（29 s10）。

相比而言，美国堪萨斯大学信息与通信技术中心 Huan 等[18]收集了 2007—2012 年游泳运动员的所有数据。在他们的大数据集里，包括超过 2 600 名运动员、500 场不同的赛事、40 000 个运动员不同赛段的成绩数据。他们的研究表明，叶诗文伦敦奥运成绩的提升在大数据视野下属于正常，Callaway 等人对叶诗文的评判有偏见之嫌。表 1 所示是的以 12 个月为时间窗口，400 m 混合泳成绩提升前 32 名的运动员。由表 1 可见，在一年内，成绩提升排名第一的爱尔兰女子运动员 Grainne Murphy 在成绩上提升了 19.97 s，哈萨克斯坦男子运动员 Dmitriy Gordiyenko 成绩提升了 14.13 s，澳大利亚女子运动员 Stephanie Rice 成绩提升 12.14 s，而叶诗文的成绩提升（6.72 s）仅位列第 16 位。倘若将时间窗口拉长至 24 个月，Grainne Murphy 成绩提升了 21.37 s，Dmitriy Gordiyenko 成绩提升了 16.31 s，匈牙利女子运动员 Katinka Hosszu 成绩提升 15.88 s，叶诗文的成绩提升（6.72 s）位列第 19 位。由此可见，在类似时间段内成绩提升较快的案例在年轻游泳运动员身上存在普遍性，叶诗文的成绩提升不存在所谓的“异常”。

由以上案例分析可知，大数据会弱化抽样的有偏性，能带来更为正确的大视野，从而避免“小数据”的有偏性带来的误导性。因此，有理由相信，随着大数据技术的日臻成熟，在未来反兴奋剂斗争中，大数据必将扮演重要角色。

2.2　大数据在科学选材中的应用

运动员科学选材是体育强国获得优异成绩的重要战略保障，特别是在当前经济、科技高度发展，体育强国之间的运动水平日益接近，训练手段和方法、训练条件的差异逐渐缩小的背景下[19]。此外，竞技体育竞争日益激烈的发展趋势和成才率相对低下的客观事实，使得科学地实施运动员选材的重要性更加突显[20]。“选材的成功意味着训练成功的一半”。《运动员科学选材》将科学选材定义为：科学选材是根据不同运动项目的特点和要求，用现代科学的手段和方法，通过客观指标的测试，全面综合评价和预测，把先天条件优越、适合从事某项运动的人才从小选拔出来，进行系统培养，并且不断地监测其发展趋势的一个过程[21]。

传统意义上的运动员选材，多是依据运动员的静态数据来选取“未来可用”之才，但是对于已成年且成熟的球员，传统的选材方式就显得“鞭长莫及”。譬如，目前一名优秀运动员的转会身价可能动辄几百万乃至几千万美元。在此背景下，为确保自己的球队（运动队）获胜，一个教练或一个运动队的管理层必须考虑的问题是，如何在有限的预算下选取最有（潜在）价值的球员，这对俱乐部来说至关重要。随着信息技术的快速发展，基于大数据分析的运动员选材，就可成为传统选材方式的有益补充。下面以棒球为例加以说明。

一直以来，棒球教练们选择球员的惯例是依据球员的“击球率”（Batting Average，AVG），其值等于安打数/打数。击球率代表一位打者能击出安打的机率，高者判定为该球员有潜力。迈克尔·刘易斯（Michael Lewis）在其著作《Moneyball：The Art of Winning an Unfair Game》中描述了一个真实的案例：美国职业棒球队大联盟奥克兰“运动家球队”的总经理比利·比恩（Billy Beane）依据其独特的运动员选材方式，以最经济的成本带领自己的球队赢得多次比赛[22]。

比利·比恩另辟新径，采用“上垒率”（On-Base Percentage，OBP）来挑选球员，上垒率＝（安打＋四坏保送＋触身球）/（打数＋四坏保送＋触身球＋高飞牺牲打），它代表一个球员能够上垒而不是出局的能力。这样挑选球员的策略并非比恩凭空而来，而是十余年来他对数千场球赛的大数据分析的结果。他用统计学方法把人的因素及运气成分剥离，将棒球场上每一个区块用坐标表示，把每一球击出去的力道、角度与落点加以分类，考虑每一球形成“安打”的概率后，换算成实际得分的预期值（expectation），进而套入每个选手实际比赛的历史数据，去换算成每位选手实际上所贡献的得分值[23]，该比分值最终可用赢球概率 P（win）来表示，如表 2 所示。表 2 统计了 1999—2004 年美国棒球比赛数据，对赛场上能对赛事结果产生重要影响的几种典型的行为（如保送球、触身球、一垒安打等）进行了一系列条件概率估算，得出各种赛场行为的赢球概率 P（win）值。

由表 2 可见，四坏保送＋触身球的获胜概率之和达到 5.65%，其值已大于一垒安打（4.18%）的贡献率了，因此教练值得重视。但是它们对整个比赛的全局影响看起来远没有打击率直观。通过精细的数学模型分析，比恩发现高“上垒率”与比赛的胜负有某种关联（correlation），据此他提出了自己的独到见解，即一个球员怎样上垒并不重要，不管他是地滚球还是三跑垒，只要结果是上垒就够了。虽然偷垒会让棒球比赛看起来更精彩，但对比赛的结果却没有太大影响，教练不应太多关注这类华而不实的技能，赢得比赛最为重要。在广泛的批评和质疑声中，比恩通过自己的大数据分析，创立了“赛伯计量学”（Sabermerrics）[11]。据此理论，比恩依据“高上垒率”选取了自己所需的球员，带领自己的球队在 2002 年的美国联盟西部赛事中夺得冠军，并取得了 20 场连胜的战绩。

从比恩对棒球运动员的选材可以看出，基于证据（Evidence-Based）的决策比基于经验（Experience-based）的惯性思维来得更加理性且更加有效[24]。体育教练或上层管理者应从依靠自身经验做判断过渡到依靠数据做决策，这一重要转变是体育大数据做出的最大贡献之一。如前文所述，大数据更看重事物间的相关性，而其背后的因果关系则容许在后期进行研究。通过大数据分析，比恩发现了“上垒率”与赛事的胜负存在关联，因此在运动员选材上，他并没有拘泥于其背后的因果关系。也就是说，大数据强调的是“是什

么”，而非“为什么”，前者表明客观事实才是我们生活、思维的基础，这一思维的转变，是大数据的重要精髓之一。

显然，依据大数据技术实施运动员科学选材，大大拓展了体育选材的视野，提升了选材的可靠性，为教练员评估运动员的当前性能和未来潜能提供了另外一种强有力的策略。

表2 部分赛场行为的棒球取胜概率

赛场行为(Even)	发生次数(Frequency)	平均获胜概率P(win)	标准方差(Std. Error)
四坏保送(Walk)	17 208	0.028 1	0.000 2
触身球(Hit by pitch)	1 752	0.028 4	0.000 6
一垒安打(Single)	29 866	0.041 8	0.000 3
二垒安打(Double)	8 902	0.064 6	0.000 7
三垒安打(Triple)	952	0.094 8	0.002 6
本垒打(Home run)	5 963	0.121 7	0.001 3
三振出局(Strike out)	31 254	−0.027 6	0.000 1
地滚球(Ground out)	35 911	−0.022 0	0.000 1
飞球出局(Fly out)	25 279	−0.024 8	0.000 1
双杀(Ground into double play)	3 833	−0.075 3	0.001 0

注：表2中出现的负值概率，是指出现了某项赛场行为导致比赛失败的概率，即P（win）的相反面。数据来源：Stats Inc。

2.3 大数据在赛绩提升中的应用

对于精英运动员来说，他们的赛场性能基本达到了其收益递减（diminishing return）的临界点，即提高运动员的赛场表现已经非常困难，再进一步地增大训练负荷可能适得其反。因此，对于运动员或教练员来说，安排合理的比赛战术显得非常重要。但是，这种情况下，不能单凭运动员或教练员自己的感觉和经验来说话，来自体育大数据的深度挖掘与分析，并以分析结果作为竞技赛场上的战术指导依据，已逐渐成为未来体育赛事竞争的趋势所在。下面以网球为例加以论述。

判断运动员竞技水平高低的一个重要标准就是看他/她能否成功地赢得比赛。但是一场比赛的影响因素很多，譬如在网球比赛中有双误、ACE球、一发成功率、挽救破发点、接发球得分率及成功破发率等12项技术指标。有关网球致胜的技术因素，很多研究人员从不同的角度出发进行了研讨，呈现出不同的结论。有的研究者认为是失误率低，有的研究者认为发球成功率高，还有研究者认为是接发球能力强，甚至还有研究人员认为一个回合中ACE个数是取胜的关键。在众多“混杂”因素中，如何抓住主要获胜因素至关重要。

自2005年以来，IBM通过Slam Tracker应用软件追踪了网球四大满贯赛事的8 000多场比赛，每场比赛收集了4 100万个数据点，包括5 500多个分析模型。在IBM的Slam Tracker中，大数据分析的精华主要体现在“制胜关键指标”（Keys to the Match）里，它在每场比赛中为对阵双方的选手找到了三个获胜的关键指标，且为每一个指标设定一个量化达标线。

澳大利亚网球教练、体育大数据专家Craig O’Shannessy认为[25]，在竞技赛场上，一个教练不仅要有专业的技能，还要有赛场数据的统计分析能力，让大数据发出自己的声音，才能安排更合理的比赛战术，最终赢得比赛。

譬如，莎拉波娃（Maria Sharapova）和小威廉姆斯（Serena Williams）在网球赛场上是宿敌。莎拉波娃对阵小威廉姆斯常是输多赢少，如何击败小威廉姆斯，对于莎拉波娃和她的教练来说是个难题。对此，O’Shannessy使用IBM的数据分析软件，对数以千计的网球比赛的各项数据实施了深度分析，发现了影响比赛成败的关键因素所在，而且找到了真正有效并且非常重要的比赛模式和指标。基于对小威廉姆斯赛场的大数据分析，O’Shannessy对莎拉波娃提出了自己独到的建议[26]：

（1）在小威廉姆斯的发球局中，在右侧平分区发球时，她的一发几乎全部发向外角，即莎拉波娃的正手；而二发则更多地发向场地中间的T点，也就是莎拉波娃的反手。据此，O’Shannessy给出建议，莎拉波娃可调整自己在接发球的站位，来获得更佳的接发球效果。

（2）在双方的多拍回合中，莎拉波娃认为自己的大角度进攻容易得分；而数据显示的恰恰相反，莎拉波娃因此失分更多。其原因是小威廉姆斯不仅能及时赶到，而且在击球角度上有更多的选择。据此，O’Shannessy建议，莎拉波娃应该放弃“自以为是”的战术，如果能够让小威廉姆斯在跑动中停下脚步、再次启动并重新组织击球线路，这样莎拉波娃就会有更多的机会。

一些有意义的结论并不是常规方法分析能够显现的，而通过赛场大数据的分析，才可发现包括网球在内的很多竞技运动，不是简单的移动和对策。大数据分析的介入，让运动员和教练们可以从另一个维度解析他们所从事的运动，从而获得有见地的洞察，而这些洞察有助于运动员和教练制定有针对性的训练计划，调整比赛策略。由此可预见，大数据技术在未来日趋激烈的比赛中，将发挥越来越大的作用。

大数据在体育科学中的应用与发展越来越广泛，譬如，大数据在体育产业、体育传播等相关领域均有广泛的应用前景[12,27]。正如Google的首席经济学家Hal Varian所指出的[28]，数据是广泛可用的，而我们所缺乏的就是从中提取出有用知识的能力。正是由于大数据的广泛存在，才使得大

数据问题的解决很具挑战性。而它的广泛应用则促使越来越多的研究人员开始关注和研究大数据问题。

3　大数据应用面临的主要问题与挑战

随着大数据时代的到来，包括体育科学在内的各个领域，都需要同步“革新”我们的基本生活、工作和思维方式[11]。目前，越来越多的行业、领域已开始在数据爆炸性增长的时代寻找机遇。为了能让体育科学在大数据时代焕发新的生机、争得一席之地，如下几个方面值得思考：

（1）体育科学研究的思维范式需要发生根本性变革。在大数据时代，大数据所代表的不仅是一种技术手段的创新，同时也意味着所有人的思维方式都将发生巨大变革。而在体育科学中，长期以来，基于“小数据”的随机样本、精确性、因果关系等已成为一种极为普遍的思维范式。基本上，几乎所有的探索与研究都是为了解答一个问题——“为什么?”，而这恰恰与“大数据思维”截然相反。正如舍恩伯格等[11]所强调的，目前最重要的是人们可在很大程度上从对因果关系的追求中解脱出来，转而将关注力放在相关关系的发现和应用上。

（2）目前体育科学领域中数据的“流动性”和“可获取性”亟需改进。当前大数据发展的最大障碍在于数据的“流动性”和“可获取性”[29]。早在 2009 年，美国政府就创建了专门的数据获取网站（http：//www.data.gov），公众能够通过这个网站获得各种包括体育在内的政府数据。IBM 很早亦确立了其体育大数据的发展方向。早在 2000 年悉尼奥运会上，IBM 已帮助搭建了奥运会所需的赛事 IT 系统，并从此开始了体育赛事的大数据挖掘[30]。开放的、流通的数据是时代趋势的要求，我国要赶上这样一场大数据变革，首先体育管理机构应公开各项数据，其次是相关的体育俱乐部等体育企业，最后是运动员和教练员等以自愿的原则公开个人性能数据。

（3）与体育相关的大数据收集和提取的合法性需要得到保障。体育运动的主体是人，而与人相关的数据势必涉及个人隐私及各种行为细节的记录。因此，我们需要在数据隐私保护和数据隐私应用之间进行权衡。任何体育俱乐部或体育管理部门从运动员群体中提取私人数据，运动员都应有必要的知情权，将运动员的隐私数据用于商业开发时，都需要得到运动员的认可。数据源头的采集受限可能会大大限制大数据的商业应用与开发[31]。因此，如何做到既深入挖掘大数据给体育科学带来利益的智慧部分，又充分保护运动员隐私不被滥用，在大数据的利用中找到运动员信息开放和保护的平衡点，是体育大数据提出的又一巨大难题。

（4）体育大数据结论的解读和应用有待加强。体育大数据可从某些赛事数据分析上揭示各个变量（如棒球运动员的上垒率、长打率、获胜率等）之间可能的关联。但是，数据层面上的关联如何具体应用到体育科学实践中？如何制定可执行方案从而合理应用体育大数据的结论？这些问题要求执行者（如教练、体育管理部门等）不但能够解读大数据，同时还需深谙体育发展各个要素之间的关联。这一环节基于大数据技术的发展又涉及到管理和执行等各方面因素。

（5）体育大数据人才缺乏的现状亟待改变。人的因素是体育大数据战略的制胜关键。从技术角度，执行人需要理解大数据技术，能够解读大数据分析的结论；从管理的角度，执行人需要制定出可执行的解决问题的方案，并且确保在利用体育大数据解决问题的同时，没有制造出新的问题。这些需求，要求执行人掌握体育科学的大数据技术，有系统论的思维，能够从复杂系统的角度关联地看待大数据与体育行业的关系。此类人才的稀缺性将制约体育大数据的发展，故此迫切需要培养善用大数据的人才。

4　结语

大数据时代已经到来，世界各国将在这一新的科学领域展开新一轮的竞争。“大数据”虽然已成为包括体育科学在内的众多研究领域的热点议题，但目前对大数据的研究仍处于起步阶段，还有很多基础性的问题有待解决，如大数据的科学定义、大数据的形式化表述、大数据的结构模型等[12]。

在目前条件下，中国开展体育大数据研究与应用是有困难的，最需迫切解决的就是赛务、纪录等信息系统均未完全建立，导致历史数据要么丢弃，要么难以公开获取。数据量不足、数据难以获取，致使从这些数据中获取有意义的信息难度很大。譬如，运动员选材，即使教练员有再好的经验和分析策略，但倘若没有数据作为支撑，也难免会陷于“巧妇难为无米之炊”的困境。一个新生事物的出现必将导致传统观念和技术的革命，要解决体育科学中各类大数据问题仍有很长的路要走。

参考文献：

[1] 李国杰，程学旗. 大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 中国科学院院刊，2012，27(6)：647-657.

[2] LOHR S. The Age of Big Data[N]. New York Times，2012-02-12(Sunday Review).

[3] LYNCH C. Big data：How do your data grow[J]. Nature，2008，455(7209)：28-29.

[4] HOWE D，COSTANZO M，FEY P，et al. Big data：The future of biocuration[J]. Nature，2008，455(7209)：47-50.

[5] WALDROP M. Big data：wikiomics[J]. Nature，2008，455(7209)：22-25.

[6] LENNARD P R. Dealing with data[J]. Nature，2011，347(6288)：104-106.

[7] 李健，王丽萍，刘瑞. 美国的大数据研发计划及对我国的启示[J]. 中国科技资源导刊，2013，45(1)：17-23.

[8] 李国杰. 大数据研究的科学价值[J]. 中国计算机学会通讯，2012，8(9)：8-15.

[9] McKinsey Global Institute. Big data：The next frontier for innovation，competition，and productivity[R]. 2011.

[10] Gantz J，Reinsel D. Extracting value from chaos[R]. IDC iview，2011：1-12.

[11] 维托克·迈尔-舍恩伯格，肯尼斯·库克耶. 大数据时

代:生活、工作和思维的变革[M]. 盛杨燕,周涛,译. 杭州:浙江人民出版社,2013.
[12] 张引,陈敏,廖小飞. 大数据应用的现状与展望[J]. 计算机研究与发展,2013(S2):216-233.
[13] 江和平,田洪. 试论大数据在体育频道中的应用[J]. 电视研究,2014(4):10-12.
[14] BULL A. Ye Shiwe's world record Olympic swim 'disturbing', says top US coach[EB/OL]. (2012-07-31)[2014-12-10]. http://english.caixin.com/2012-07-31/100417674.html.
[15] CALLAWAY E. Why great Olympic feats raise suspicions—'Performance profiling' could help to dispel doubts[EB/OL]. (2012-08-01)[2014-08-09]. http://www.nature.com/news/why-great-olympic-feats-raise-suspicions-1.11109.
[16] LONGMAN J. China pool prodigy churns wave of speculation[N]. The New York Times, 2012-07-31(Sports).
[17] 马国全,张虎祥. 运动员性能剖析法研究——从"叶诗文事件"谈起[J]. 中国体育科技,2013,49(1):110-116.
[18] HUAN J, LUO B. Big Data Analysis of Swimming Athletes' Performance Records[EB/OL]. (2012-08-26)[2014-08-09]. http://www.ittc.ku.edu/huanlab/swimData.
[19] 张春甫. 对新世纪运动员科学选材发展趋势的探讨[J]. 首都体育学院学报,2003,15(3):23-25.
[20] 郑晓鸿,吴铁桥. 对运动员科学选材若干问题的思考[J]. 首都体育学院学报,2003,15(3):21-22.
[21] 余竹生,沈勋章,朱学雷. 运动员科学选材[M]. 上海:中医药大学出版社,2006.
[22] MACLENNAN T. Moneyball: The Art of Winning an Unfair Game[J]. The Journal of Popular Culture, 2005,38(4): 780-781.
[23] HAKES J K, SAUER R D. An economic evaluation of the Moneyball hypothesis[J]. The Journal of Economic Perspectives, 2006, 20(3):173-185.
[24] CULLEN F T, MYER A J, LATESSA E J. Eight lessons from Moneyball: The high cost of ignoring evidence-based corrections[J]. Victims and Offenders, 2009, 4(2): 197-213.
[25] TSCHORN A. Why the tennis world's Yoda is also a data Jedi[EB/OL]. (2013-10-13)[2014-08-10]. http://techpageone.dell.com/technology/tennis-worlds-yoda-also-data-jedi.
[26] WERTHEIM J. How can Sharapova beat Williams Analytics man may have solution[EB/OL]. (2013-05-29)[2014-8-10]. http://sportsillustrated.cnn.com/tennis/news/20130529/french-open-serena-williams-maria-sharapova-analytics.
[27] 张江南. 大数据时代对体育传播的影响[J]. 武汉体育学院学报,2014,48(7):16-20.
[28] VARIAN H R. Big Data: New Tricks for Econometrics[EB/OL]. (2013-08-31)[2014-12-10]. http://people.ischool.berkeley.edu/~hal/Papers/2013/ml.pdf.
[29] 田溯宁. 中国更应推进"数据公开"[EB/OL]. (2012-10-10)[2014-12-10]. http://www.china-cloud.com/yunhudong/yunrenwu/renwuxinwen/2012/1010/15427.html.
[30] 赵楠. IBM:体育赛事背后的大数据机遇[N]. 第一财经日报,2013-09-26(C02).
[31] 周锦昌,孟昭莉. 谁来引领中国大数据的发展?[EB/OL]. (2013-09-12)[2014-08-12]. http://www.1000plan.org/qrjh/article/41038.

Applications and Thoughts of Big Data in Sport Science

MA Guo-quan, YANG Jian-wen, ZHANG Hu-xiang, TIAN Yu

(Department of P.E. Teaching and Research, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract: In recent years, big data has become a hot issue and research topic in business, government and the academic field. However, its research in the field of sports is at an initial stage. Based on the methods of literature research and case analysis, at first the origin and substantive characteristics of big data is introduced in this paper. Then innovative applications of big data in sports science, including anti-doping, scientific selection of athletes and improvement of athletes' performance are explored. Finally, the potential problems and challenges of big data, which may occur in the development of sports in the future are deeply analyzed and discussed.

Key words: big data; anti-doping; scientific selection of athletes; improvement of athletes' performance