大数据棒球

Big Data Baseball: Math, Miracles, and the End of a 20-Year Losing Streak

一个终结俱乐部20年败绩的奇迹

「美」特拉维斯·索契克(Travis Sawchik)◎著　史丹丹◎译　夏毅◎审校

体育产业发展清华丛书

大数据棒球

Big Data Baseball: Math, Miracles, and the End of a 20-Year Losing Streak

一个终结俱乐部20年败绩的奇迹

「美」特拉维斯·索契克(Travis Sawchik)◎著　史丹丹◎译　夏毅◎审校

清華大學出版社
北京

Travis Sawchik
Big Data Baseball: Math, Miracles, and the End of a 20-Year Losing Streak, 1st Edition
ISBN: 978-1250094254

图书在版编目（CIP）数据

大数据棒球：一个终结俱乐部20年败绩的奇迹 /(美) 特拉维斯· 索契克 (Travis Sawchik) 著；史丹丹译 . — 北京 : 清华大学出版社，2017
（体育产业发展清华丛书）
书名原文 : Big Data Baseball:Math,Miracles,and the End of a 20-Year Losing Streak
ISBN 978-7-302-46055-8

Ⅰ . ①大 … Ⅱ . ①特 … ②史 … Ⅲ . ①棒球运动 – 俱乐部 – 介绍 – 美国 Ⅳ . ① G848.167.12

中国版本图书馆 CIP 数据核字（2017）第 001070 号

责任编辑：张 伟
封面设计：众智诚橙
责任校对：王凤芝
责任印制：杨 艳

出版发行：清华大学出版社
网 址：http://www.tup.com.cn，http://www.wqbook.com
地 址：北京清华大学学研大厦 A 座 邮 编：100084
社 总 机：010-62770175 邮 购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市中晟雅豪印务有限公司
经 销：全国新华书店
开 本：148mm × 210mm 印 张：11 字 数：210 千字
版 次：2017 年 1 月第 1 版 印 次：2017 年 1 月第1次印刷
定 价：69.00 元

产品编号：071720-01

体育产业发展清华丛书编委会

丛书序

开卷开步开创，发展体育产业

半年前，得赖于一批忠诚母校、热心体育的校友的支持，以及英迈传媒的带头出力，清华大学体育产业发展研究中心成立，希望能够充分发挥清华大学学科齐全、人才密集、体育传统深厚的优势，创造性地开展研究，发挥体育产业一流思想与行动平台的作用，为落实国家体育产业发展战略、推动体育产业升级及企业发展提供智力支持。

中心筹建之初，就发现虽然国家把体育产业作为绿色产业、朝阳产业加以培育和扶持，政府官员、专家学者和实践者也已经达成共识，认为体育产

业将会成为推动中国经济转型升级的重要力量，但遗憾的是，毕竟中国的体育产业尚在起步期，呈现为价值洼地、人才洼地和研究洼地的现状。因此，中心决定与清华大学出版社合作，策划出版“体育产业发展清华丛书”，组织专家团队选书、荐书。在出版社的大力支持和密切配合下，令人高兴的是，中心成立半年之后，丛书首批即将与读者见面。

“体育产业发展清华丛书”计划分批、分层次地出版体育产业相关的书籍，既包括引进版权的国际经典著作，也包括国内学者原创的对于体育产业发展和体育管理方面的真知灼见；既有对于具体运动项目的精准聚焦研究，也有结合某一体育管理领域的深度剖析探查。我们相信，只要开始第一步，踏实耕耘，探索创新，日积月累，坚持下去，这套丛书无论是对促进体育产业的研究，还是对指导体育产业发展的实践，都是有价值的。

清华大学的体育传统非常悠久。马约翰先生曾经说过：“体育可以带给人勇气、坚持、自信心、进取心和决心，培养人的社会品质——公正、忠实、自由。”在庆祝马约翰先生服务清华五十年的大会上，蒋南翔校长特别号召清华学生“把身体锻炼好，以便向马约翰先生看齐，同马约翰先生竞争，争取至少为祖国健康地工作五十年。”2008 年，时任清华大学党委书记的陈希同志说过：“五十年对一个人来讲，跨越了青年、中年和老年，为祖国健康地工作五十年，就是要在人生热情最高涨、精力最充沛、经验最丰富的各个阶段为党和人

民的事业做出贡献。”就在中心成立这半年来，国家先后发布《全民健身计划（2016—2020年）》和《“健康中国2030”规划纲要》，国民强身健体、共建健康中国，成为国家战略。“为祖国健康工作五十年”这种清华体育精神在当下绝非赶时髦，而是清华体育传统的强化与传承。

清华体育，在精神层面也格外强调“Sportsmanship”（运动家道德）的传统，这里回顾一下老清华时期的概括：承认对手方是我的敌手，不在他面前气馁也不小视他；尽所能尽的力量去干；绝对尊重裁判人的决定，更要求学生“运动比赛时具有同曹互助之精神并能公正自持不求徼幸”。据我所知，许多企业的核心价值观中亦有Sportsmanship的表达，甚至直接就用这一词汇作为组织成员的行为规范（如韩国SK集团）。当我在“体育产业发展清华丛书”中看到描述体育产业中的历史追溯、颠覆创新、变革历程以及行业规范时，这个词再次浮现在眼前，这其实也是商业的基本规则和伦理，也是产业成长的核心动力和引擎。

体育产业发展，需要拼搏精神，需要脚踏实地，来不得投机，也无捷径可走，因此，中国的体育产业发展，就更需要所有利益相关者多些培育心态，方能形成健康的生态共同体。同时，体育产业发展，需要尊重规则和规律，无论是运动项目的发展规律，还是商业活动的规则、规范，无论是与资本握手的契约精神，还是商业模式中利益相关者准确定位的角色意识。我很希望“体育产业发展清华丛书”能借他山

之石对中国体育产业发展的路径和模式有所启发，能用严谨、规范的研究和最佳的实践案例对中国体育产业和体育管理的具体问题有所探究。

每一步，都算数！无体育，不清华！

杨斌

清华大学副校长、教务长

2016 年 12 月

序 言

这本书从某种意义上讲，就是《点球成金》（*Money ball*）[1]的续集和升级版，不过它的棒球大数据故事，发生在奥克兰运动家队的故事十余年之后，主角换成了另一支美国职业棒球大联盟MLB的穷队——匹兹堡海盗队。

---

1 《点球成金》(*Money ball*)，美国迈克尔·刘易斯著，讲述美国职棒大联盟奥克兰运动家棒球队总经理比利·比恩如何通过棒球大数据改造球队、以弱制强的故事。此书名列《财富》杂志评选的75本商业必读书，被《福布斯》(*Forbes*)评价为“既是关于棒球，更是关于管理的最佳图书之一”，后被改编成电影，由明星布拉德·皮特饰演主角，于2011年在美国上映，被广为所知。

故事几乎用一段话就能说清楚：一位名校毕业、棒球爱好者出身的大联盟球队总经理，请来了少年成名但一直不得志的主教练，加上两位名校毕业、有数学和计算机天赋的棒球迷做数据分析师及系统架构师，收购了数名在走下坡路但别有所长的大联盟球员，组成了一个团队，通过棒球大数据的采集和应用，使得在美国职业棒球大联盟垫底 20 年的穷队——匹兹堡海盗队咸鱼翻身的故事。

这么一个故事，有什么可看的呢？会比《点球成金》更精彩？我带着几个问题，仍然饶有兴趣地用两天看完了这本书。

问题一：十年间，大数据棒球发生了怎样的演变？

说到棒球大数据，首先无法回避一个人——比尔·詹姆斯 (Bill James)，这位被称为“当今数据分析学革命之父”、创造了赛伯计量学 (Sabermetrics，又称棒球统计学) 的统计学家兼棒球爱好者，其实是《点球成金》一书中的真正幕后主角。早在 20 世纪 70 年代，毕业于堪萨斯大学、在堪萨斯劳伦斯城 (Lawrence) 一家猪肉豆类罐头厂做巡夜保安的比尔·詹姆斯就开始研究棒球数据了。1977 年，他自费出版了第一本《比尔·詹姆斯棒球摘要》[1]。自 1977 年至 1988 年，每年出一期《比

1 *Bill James Baseball Abstract*，《1977 年棒球摘要：提供 18 项你在其他地方找不到的棒球统计资料》。

尔·詹姆斯棒球摘要》，从而奠定了赛伯计量学这门统计学的基础。

要知道，到1975年，在美国才诞生了第一台个人计算机的雏形Altair 8800计算机；1978年，Intel公司的16位微处理器8086才出现；1982年，英特尔公司在8086的基础上，研制出了80286微处理器，就是我们当年所说的286电脑；1989年，有关互联网应用的分类互联网信息协议World Wide Web方确定。

在美国棒球界开始进行数据分析革命时，金融等商业领域尚未普遍使用数据分析，只是在这之后，数据分析才逐渐广泛应用到金融、证券乃至政治解析等各领域，而今美国不少金融和政治数据分析师都是棒球统计员出身。2011年，比尔·詹姆斯甚至写了一本书《热门罪案》(*Popular Crime*)，将数据分析应用到了连环杀手身上。

1984年，针对各大联盟球队不愿对外公开球队数据的情况，比尔·詹姆斯在其当年出版的《比尔·詹姆斯棒球摘要》中，倡议发起一场名为“记录纸项目”(Project Scoresheet，也称为“记分表计划”)的草根活动，号召遍布大城小镇的广大球迷详细为每场球赛记录，然后把记录的信息输入一个电脑数据库，而这个记录纸，至今仍被棒球赛事计分时所广泛使用。

与比尔·詹姆斯同期，另一位棒球数据统计爱好者、制药公司研究员出身的迪克·克莱摩尔(Dick Cramer)于20世

纪80年代设立了一家名为STATS的公司[1]，从事棒球数据收集及分析。比尔·詹姆斯给STATS公司投资，并出任创意部总监。该公司的数据产品，被ESPN[2]和《今日美国》报等所采用，直到1999年该公司被默多克的新闻集团所属福克斯广播公司以4 500万美元的价格收购以前，其一直是业界领先的棒球数据供应商。

《点球成金》一书中，奥克兰运动家队的故事只记录到了2003年，其所采用的数据分析系统，也主要是以STATS公司为代表的数据分析产品；而本书中匹兹堡海盗队的故事则一直记述到了2014年，那么在这10年间，棒球数据统计又发生了什么重大变化呢？我们可以看看本书中提及和笔者所知的几个重要数据应用系统和重大事件。

一、PITCHf/x系统

2007年，位于芝加哥的运动大观公司(Sportvision)开发推出了PITCHf/x系统。PITCHf/x系统问世以前，棒球运动并没有一个真真正正的大数据工具。由于PITCHf/x系统问世，职业棒球行业随之产生了一个前所未有的工作部门——数据学部门。

1 STATS(运动队分析和跟踪系统，Sports Team Analysis and Tracking Systems)；本书中提到的约翰·迪万(John Dewan)是STATS公司的CEO。

2 ESPN(Entertainment and Sports Programming Network，娱乐与体育节目电视网)，是一个24小时专门播放体育节目的美国有线电视联播网。

PITCHf/x是在垒包内置摄像头的运动跟踪系统，研发的目的本是改进ESPN的一款叫作K区的产品(K-Zone product)，这个产品用于测定投手的投球是否落在好球区内。PITCHf/x每年自动生成将近2 000万个可用的数据点，差不多相当于20世纪记录的数据总量。

2007年，投球自动辨识系统PITCHf/x开始安装在各个大联盟球场运行，采集实时投球数据，2008年遍及每个大联盟球场。当年赛季，每个赛场都装上了60赫兹的摄像头。摄像头和物体识别软件会拍下球自脱离投手的手至穿越本垒板为止这段时间的运行情况。PITCHf/x会依据所拍照片，实时把球的速度、轨迹、三维位置计算出来，速度误差小于每小时一英里，位置误差小于一英寸。此外，PITCHf/x也会实时标记投球的类型。有史以来，投手的准确投球速度和各种投球类型所占的准确比例终于得以为人掌握。投球的速度、类型、运动和位置也终于有了一套标准单位，并可轻易在FanGraphs.com、BrooksBaseball.net等网站上查到。

伊利诺伊大学教授艾伦·M. 内森(Alan M. Nathan)在2012年一篇论物理与棒球的论文中写道："(PITCHf/x)记录投球速度、球与本垒板相对位置等物理量之精确，前所未有。然而，更为重要的是，以前未加度量的物理量，我们现在也有了度量标准。"

二、TrackMan 投球跟踪系统

丹麦 TrackMan 公司是运动大观公司一个强有力的竞争对手，以利用雷达跟踪高尔夫球飞行和滚动轨迹而名声大噪。2009 年，TrackMan 公司开始利用雷达跟踪技术，进行投出和击出的棒球研究，同时该公司在三个大联盟球场开始测试自己的技术。

TrackMan 的主要目的是整理数据，给球队提供一些基础性的信息，最引人注目的东西是有效速率。其读数与 PITCHf/x 的基本一样，但 TrackMan 所测的是球在空中运行的（整个）轨迹，而不是以五十英尺间隔距离为准，在球行轨迹上选取（二十个）不同的点来测，而且还测投手的伸展长度。

能测出投手的伸展长度是球队对 TrackMan 产品感兴趣的一大主要原因。另外两大原因是该产品能跟踪球被打击时的初速度和场内球的末速度。PITCHf/x 能够告诉球队投手的垂直释球点，但不能告诉球队投手的水平释球点，因而也就不能显示投手球出手时球离本垒板的距离。这一点很重要，因为如果球出手时球离本垒板更近，则投手的有效速率会更大。举个例子，甲、乙两个快球初始时速同为 93 英里，甲行进 53 英尺，乙行进 55 英尺，则甲的整体速度比乙的大。

三、MLBAM 公司设立

在 2000 年互联网泡沫破裂时，美国职业棒球大联盟的 30 个俱乐部，各家每年出资 100 万美元，逆势成立了子公司 MLBAM(MLB Advanced Media，美国职业棒球大联盟高级媒体公司)，利用互联网改善棒球比赛播放体验。MLBAM 成立后，先是为职棒大联盟和下属的 30 个俱乐部建立官方网站。

2002 年，MLBAM 开始将业务延伸到流媒体，在网站上提供比赛视频播放。

2003 年后，MLBAM 便已经开始盈利，并在 2006 年将早期投资还给了俱乐部。

2005 年，MLBAM 以 6 600 万美元的价格买下了售票网站 Tickets.com。

2005 年，花旗银行、高盛、瑞士信贷、摩根大通等投行曾试图劝说 MLBAM 上市，当时只有在线视频业务的 MLBAM 估值已经达到 20 亿 ~25 亿美元。

2014 年，MLBAM 已经有 5% 的棒球赛门票通过手机应用销售。MLBAM 的 At The Ballpark 应用允许已经购票入场的用户在场内直接在手机上付费升级看台座位。

四、Statcast 球员跟踪系统

2014 年 3 月，MLBAM 开发的 Statcast 球员跟踪系统面世。

Statcast从不同的系统接收数据，然后将数据整合在一起：

Statcast利用TrackMan的SABR40棒球雷达将军队追踪飞机和导弹的3D多普勒雷达用于追踪棒球的飞行轨迹。雷达每秒扫描2 000次场地，根据返回电波的变化判断棒球的运动。它的高精度扫描不光能获得棒球的飞行速度和轨迹，还能知晓对于比赛有影响的旋转角度。

同时利用两组美国蔡润合古公司(ChyronHego)的双套摄像机列阵追踪赛场上的球员，由于摄像机为立体布局，每个列阵由两个倒挂着的方块构成，每个方块相隔15米，因而具备三维追踪能力，它们就像人的双眼一样通过不同角度“看到”的画面获得立体的影像，据此判断运动员的运动速度。这些摄像机追踪赛场上每位球员的运动，并与TrackMan的多普勒雷达读数同步，然后，球员和球的运动由Statcast系统的软件转化成具体数据。PITCHf/x能够追踪所投之球的运动、位置、速度，Statcast则能追踪赛场上的一切运动。原先只能凭肉眼主观判断的东西终于可以量化了。

Statcast系统可跟踪量化球员最高时速、加速度、起点至拦截点距离、路程、路径效率等实时数据，同时还可实时追踪球速、角度、行进距离、滞空时间等，是PITCHf/x的超强版。其问世后，防守范围、准确度、路径效率、手臂力量等得以准确量化。

仅这两个系统，每场比赛就会生成7TB的数据，一个赛季2 430场比赛就是17PB。球场获得数据后会实时将数据上传到亚马逊AWS平台上，由MLBAM开发的软件进行加工，绘制成普通人能看得懂的图形表现。

1TB=1 024GB，1PB=1 024TB。17PB数据形象地表述一下，就是1TB(1 024GB)容量的电脑硬盘，能够装满17 000块！

Statcast系统目前的问题是，可以实时追踪并分享数据，却没法实时生成图表。原因是每场比赛产生的数据极多，别说实时处理，即使传输也很费时。但Statcast系统确实为未来的大数据棒球留下了巨大的想象空间。

作为一名学历史文物专业、律师出身的文科生，我如此费劲地整理出上述几个数据应用系统和企业的介绍，主要是实在很难用语言描述本书中匹兹堡海盗队与《点球成金》一书中奥克兰运动家队所处时代的差别，只好用数据说话。

上述的系统和事件说明，在这十年间，随着科技迅猛的发展，彼时的奥克兰运动家队尚处在**大量数据**时代，而现时的匹兹堡海盗队则已进入真正的**大数据**时代，在数据的Volume(体量)、Velocity(速度)、Variety(多样性)、Veracity(真实性)等方面，已经不可同日而语，因此，两书中有关数据分析师及系统架构师在棒球数据分析、应用上的细节描述，也迥然有所不同。