

足球比赛中的实时数据是如何统计出来的？人工 or 人工智能？

Part1

说到足球，恐怕即便是上了年纪，又从不怎么看球的人也能说出一两个球星的名字，比如贝克汉姆，卡卡。



左图为金嗓子一哥罗纳尔多



右图为金嗓子二哥卡卡

对足球有稍微有一点点了解的人，应该对控球率、红黄牌这些数据也不陌生。那么，作为球迷的你，有没有想过，足球比赛中的实时数据是怎么统计出来的？

2017 年 x 月 x 日 一场西班牙国家德比赛事的实时转播 转播过程中出现了诸如比分、传球成功率等实时数据：



图为直播中左上角弹出皮克和拉莫斯的传球成功率对比

图片来源：PPTV 赛事回放截图

“一定是有人拿着小本本在场边画“正”字统计！皇马射门+1+1+1，巴萨进球

+1+1+1！”老实人甲说道。

“那肯定是用热成像摄像头捕捉球员运动轨迹，用人工智能分析球员动作自动生成实时数据。”程序猿乙说道。

以上的两种回答代表了绝大多数人内心的想法，但这两种说法都不全面。实际上，当今足坛主流的实时数据收集方法是“图像识别”+“无线定位”+“人工统计”。在详细剖析实时数据如何统计之前，我们先大概预测一下各类实时数据的统计难易程度。



图为 PPTV 半场数据统计图

图片来源：PPTV 赛事回放截图

上图为西班牙国家德比转播过程中的半场统计数据，其中：

- 球员跑动距离、范围 预计收集难度：两颗星
- 球队射门数、角球 预计收集难度：一颗星
- 点球、任意球 预计收集难度：一颗星
- 红牌、黄牌次数 预计收集难度：一颗星
- 球员传球成功率 预计收集难度：四颗星
- 球员拦截成功率 预计收集难度：四颗星
- 球队控球率 预计收集难度：三颗星

上面给出的预计统计难度，完全是基于笔者对当前足球科技水平的意淫得出，而实际情况究竟如何，科技发展有没有给足球数据统计行业带来红利，这就要从

早期足球比赛的数据统计说起。

Part2 足球数据统计历史

现代足球起源于 19 世纪末的英国，紧接着便疯狂发展并席卷全球，当时主要的传播媒介是报纸。足球比赛结束后，报社将比分、照片以及比赛中的各项数据，配以简单的评述传播至世界各地。可以想象当时的足球数据仅限于“场面数据”的统计，简单的比分、扑救、定位球个数、红黄牌数等由各大报社工作人员拿小本本记录下来，赛后大家一比对，记的都差不多，那就给写稿子的人发稿吧，如此就是最初的数据统计。这种在现在看来缺乏时效性与参与度的方法，在当时却在世界各地播种下了足球文明的种子。



图为早期的英国体育报纸的足球报道

科学不断发展，球迷们显然对于纸媒的时效性非常不满，于是 1927 年 1 月



27 日，英国 BBC 电台首次通过电台广播对阿森纳 VS 谢菲尔德的比赛进行了转播。这场比赛 BBC 安排了两名评论员，一位负责对场上状况进行评述，一位指出足球在下图所示的网格中的位置，两人配合，给听众传播最正确的比赛信息。这种转播方式在现在看来是一种非常考验想象力以及反应力的转播方式，由此可以想到早期的球迷足球素养非常高。

图为电台解说分区图：用于描述足球位置

同样是在英国，同样是阿森纳，同样是 BBC，1939 年，一场阿森纳一线队与预备队的比赛录像被呈现在电视屏幕上。录播虽然是令人遗憾的，但这却开了球迷们用眼睛欣赏比赛的先河。电视机前的观众可以独立进行数据统计及核对，这就对转播公司的实时数据统计提出了更高的要求，要更准确、更及时。所以，在众多足球黑科技出现之前，**早期的足球数据统计依然是依靠人工。**

就在此时，计算机技术也发展起来，计算机成了专业的统计辅助工具，专业的统计人员将球场上的每一个动作变成一个一个事件，将每一分钟 30 次以上的高密度动作转化为事件存储下来，再由人工进行验证，并通过程序进行实时发布。这种方法对人工要求极高，因而准确度也高，多常见于第三方数据服务商，为转播方、解说员、博彩公司等提供服务。

然而，那些如雨后春笋般涌起的黑科技，究竟有没有改变这个行业呢？

Part3:

要回答这个问题，首先要清楚足球比赛中的实时数据是谁来统计的。作为一名电视机前的观众，比赛过程中得到信息的渠道有两个，一是眼睛获取的画面，二是耳朵听到的解说。转播画面带给球迷的是最原始的视觉冲击，而解说则是帮助球迷更加深入的理解比赛。球迷们经常能听到“***队本场的射门次数已经达到 15 次之多，其中射正的仅仅 3 次”诸如此类的话，难道解说员在比赛过程中还要一边拿小本本记数据，一边解说，再把数据加加减减，算出控球率？

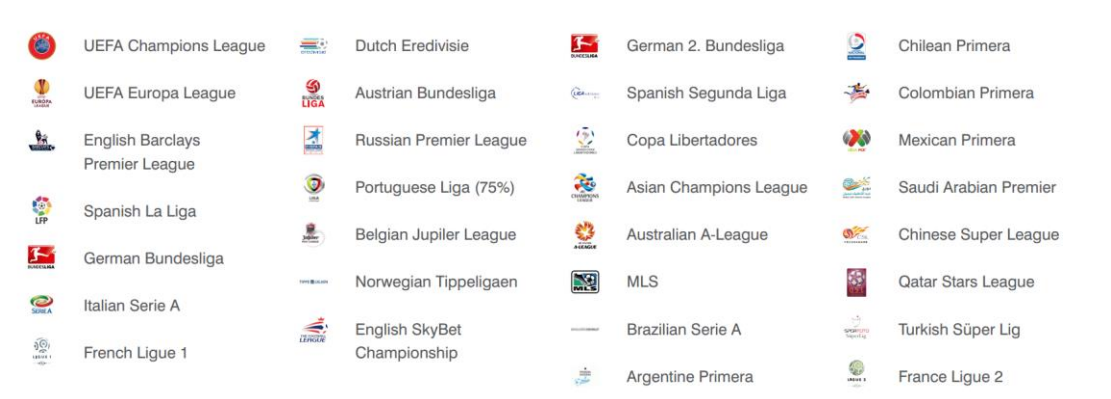
如果真是这样的话，德艺双馨的老解说员们早就累死了。

一般来说，一些大型的足球联赛，比如欧洲五大联赛，转播是由专业的信号制作公司提供。他们在比赛场馆架设 6 台、8 台甚至更多的摄像机，尽力追捕球

员从赛前采访到进球后的每一个眼神，然后在一个布满电线的大型制作场所内，将比赛信号通过卫星传输到世界各地。而赛中的实时数据则是由专门的数据公司提供的，转播公司向数据公司购买服务，数据公司使用专业的数据采集方法对实时数据进行采集。一部分数据在直播中展示，一部分更深入详细的数据在赛后进行整理分析和展示。简单的说，我们在 CCTV5 台看到的中超转播信号制作其实是一家叫体奥动力的公司提供的，而看到的实时数据，又是其名下的盈动力大数据公司提供的。

国外从事数据服务的公司非常多，而且很多公司的技术已经非常成熟，竞争相当激烈。这些公司中不乏一些巨头，比如英国的 OPTA 公司，法国的 Sport Universal 公司等等，不同的公司有着不同的数据采集系统，下面以 OPTA 为例展开分析。

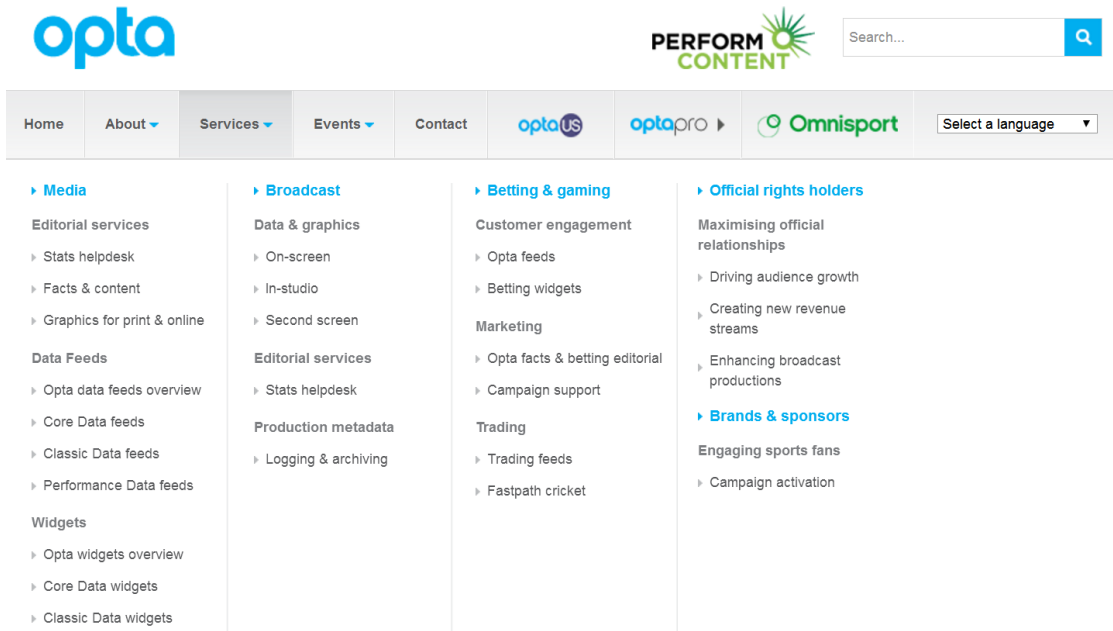
来自英国的 OPTA 实况体育数据公司是全球领先的数据提供商，该公司目前为包括英超、德甲、荷甲在内的多个大型比赛的官方数据合作伙伴，同时也为下图所示的大型联赛提供数据服务，包括中超。除了为俱乐部提供服务以外，OPTA 同时也与 Canal+，天空体育，ESPN 等电视台合作，在赛前、赛中以及赛后提供数据分析服务。



图片来源：OPTA 官网

比赛过程中，OPTA 会对场上超过 200 项的技术统计进行统计分析，实时更新，并为客户提供 XML 数据接口，方便来自全世界的客户调取。OPTA 兼有世界上最完备的体育历史数据库，从 1996 年起搜集各种比赛数据，渐渐形成了世界上数量最大、权威性最高的比赛数据库。另外，OPTA 会按照不同需求提供不

同的数据服务，其中就包括解说支持。球迷对比赛过程中历史数据的对比一定不陌生，解说员总是清楚的知道下一个进球是球员职业生涯的第几个进球，甚至是欧冠历史上第几个进球等等，其实这很大程度上要依靠 OPTA 这样的数据服务商。总而言之，这个公司真的是在做与足球数据相关的几乎所有生意。



数据来源：OPTA 官网

而赛中的数据又是如何去统计呢？大家普遍认为职业顶级联赛都用了很多了不起的高科技，只需要把机器一架，数据就像水一样哗啦啦自己流出来了，这种想法显然是不现实的。其实无论是哪家公司，采用什么样的高科技，其实核心还是解决两个问题，一是球在哪儿，人在哪儿；二是人对球做了什么，人对人做了什么。

对于人和球在哪儿，解决方案有很多，其中一种就是安装多台高速摄像机，保证每个区域都有两个以上的摄像头能覆盖。图像识别能做到判断出是谁拿了球，球速是多少，但是具体这个人想做什么，这个动作是射门还是传球，就需要经验丰富的记录员来进行判断。优秀的记录员能够在极短的时间内判断出球员的意图，再通过辅助记录的软件，选择球员，并选择此刻与该球员对应的事件，这些事件包括传球、传中、拦截、射门、扑救、犯规等等，如此便将每场比赛 2000 次以上的事件记录了下来。

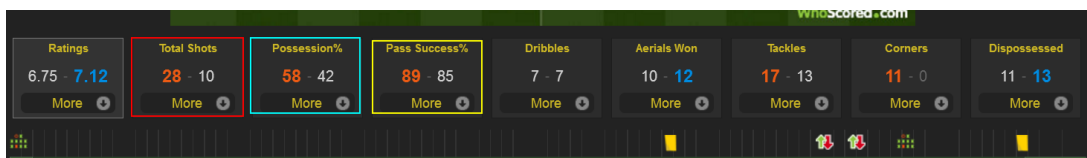


图为 OPTA 的数据专家在进行实时数据采集

图片来源：OPTA 官网

OPTA 的实时统计系统就是这种人工+智能的方式,每个记录小组有三个人,两个人进行实时数据采集,一个人进行实时的视频素材回放校验。可能有人会质疑这种掺杂人工的方式的准确度。实际上,不同公司给出的比赛数据的确是存在差异的,每一家公司的数据收集背后都有一套独特的哲学。

2018 年 1 月 13 日的一场西甲常规赛中,皇家马德里主场迎战黄色潜水艇比利亚雷亚尔,赛后的数据统计中,Whoscored 的数据显示出的控球率为 58% VS 42%,传球成功率为 89% VS 85%。而国内足球数据公司创冰科技给出的控球率则是 54.9%VS 45.1%,传球成功率为 84%VS 84%,差异明显。甚至连射门数这样看似简单的统计指标,两个网站都分别给出了 28 VS 10 和 27 VS 9 这样不同的结果,说明两个公司的统计人员对于射门这个事件的判断有差异,而这便是足球比赛数据统计中的“人为误差”。



图片来源：whoscored



图片来源：创冰 DATA

人工+智能的数据采集方式可以说是八仙过海各显神通，而统计指标的计算方法也会影响最终的统计结果。结合上面对采集方法的介绍，大致可以知道对于射门、角球、任意球、点球、犯规的统计方法，就是一个字：数。这一类事件，记录员有很长的时间去反应球场上究竟发生了什么，出错的可能性也比较小。控球率也是一个很容易记录的指标， $\text{控球率} = \text{控球时间} / \text{两队总控球时间}$ ，而控球时间则是指传出的球未被对方接触之前的时间。机器能够记录此时是哪一队拿球，拿球时间是多久，只需要人工进行一些简单的修正，例如剔除死球时间（庆祝时间、边角球、任意球时间），即可得出控球率。同理， $\text{传球成功率} = \text{传球成功次数} / \text{本队传球总次数}$ ，其实就是没有被抢断的次数，同样是机器和人工共同采集，人工校验的方式进行统计。

实际上，每个公司对指标的统计方法都是不一样的，这也就是为什么我们看到的结果不一样的原因。同时也说明，足球统计学，是一个值得统计学家们深入研究的领域，而懂足球的人，也是这个行业目前最核心的生产力。

人工+智能的方式当然不等同与人工智能，但这一定只是足球统计史上的一个时期。实际上，现在有很多可穿戴设备可以用于球员数据采集，能够对球员各项身体数据，甚至脑电波进行监测。但是由于竞技体育对于球员的衣着和穿戴有着严格的要求，所以可穿戴设备在赛场上并没有得到很广泛的应用，反而一些俱乐部在训练的时候用的较多。希望在不久的将来，我们能够看到梅西踢球时的脑电波，而解说员在他拿球突破的时候便吼道“梅西拿球了，他准备传给前场的阿尔巴，不，这一刻他改变主意了，他准备在前场打一脚远射，应该是在球门的左

下角，此刻门将的脑电波显示他准备扑球门的右边，梅西这一脚打门的成功率在80%左右，果然，球进了!!!”

Part4:

大数据悄然改变着每一个行业，足球也不例外。对于足球数据的应用，大家首先想到的，一定是博彩。对于博彩公司而言，一份真实完整的历史数据报告，能为精算师们计算初始赔率提供数据支持，而源源不断及时输入比赛实时数据，则会帮助博彩公司及时调整实时赔率，以获取更大的收益。值得注意的是，博彩公司初始赔率虽然是基于球队数据，但起着决定性因素的，其实是市场预期。所以，未来市场预期也是足球数据行业值得关注的领域。

在前不久进行的2017世俱杯半决赛之中，皇家马德里迎战阿布扎比半岛，这场比赛最引人注目的并非两队之间的对决，而是阿里云团队用人工智能技术对比赛进行了解说，AI不仅正确识别了球队、球员，还识别出了球员的传球、射门等动作。

解说员在紧张的比赛之中难免会出现差错，但人工智能背后强大的数据则有效地避免了这个问题。阿里工程师团队设计了多路神经网络卷积模型，用于比赛球员和球的精准定位，结合高效的跟踪算法，对场上的球员和球的运动轨迹进行实时跟踪。同时，人工智能会提前学习球衣的号码，完成对球员的人脸特征提取，结合赛前录入的球员信息，则能完成球员身份的识别。此外，对球门、边界线等球场基础信息的识别，也可以让该机器对球场形成更加全面的感知。由此可见，未来人工智能在足球领域的一大应用可能是呈现更多形式、更精准的赛事转播。

回归到足球本身，如何帮助球队取得胜利，如何用最低的溢价签到最合适的球员，或许是足球数据最具意义的应用。众所周知，每支球队都有各自赖以成名的球队风格，比如巴萨的Tiki-Taka战术以传控为主，而死敌皇马强调速度致胜。每种球风各有特色，相生相克，比如西甲中下游球队塞尔塔和英超中下游球队水晶宫，虽然在联赛中战绩不佳，但却是名副其实的强队收割机，这些隐藏在表面数据之下的特殊规律则可以交由足球大数据去探索。

可能有人会说强大的数据让足球失去了原有的魅力，甚至担心有一天人工智能机器人会打败人类足球，但其实数据只是让各方更理性的参与足球活动，甚至从中受益。足球永远是那个充满不确定的运动，因为，足球是圆的。

