

漫话大数据

作者：郎为民

漫话大数据
郎为民◎编著
人民邮电出版社

北京

前言

与《小崔说事》中的白云、黑土一样，《大话物联网》《大话云计算》《大话移动互联网》和《手机那点事儿》出版之后，我参加的活动那是相当多。活动之余，经常遭人调侃：“热点IT技术都让你大话完了，下一次看你小样儿拿啥开刀。”还好，有大数据！单位同事更是拿我开涮：“老狼，快写《漫话大数据》吧，院里厕所可没纸用了！”其实，写一本大数据科普图书的想法由来已久，只不过我一直在用大数据掐算什么时候出版、写些什么东东（东西）才能火。

网络社会新概念真是层出不穷，人们还没有完全适应现代互联网生活，“下一代互联网”的竞赛已经如火如荼；信息高速公路带来的海量信息资源还不曾被充分利用，又有人在想通过“物联网”玩转“智慧的地球”；在云尚未聚雨带来甘霖之时，大数据的概念又被热炒起来，它俘获了无数创业者和技术男的春心。

大数据是当前最火的科技词语，每个人都在谈论大数据，仿佛谁不谈论大数据就落伍了似的。2013年大数据当仁不让地坐上IT江湖的头把交椅。套用赵本山在小品《功夫》中所说的台词：反正你接受它也来了，不接受它也来了，接受不接受大数据时代它都带着诚意扑面而来。

男大当婚，女大当嫁，数据大了不中留，留来留去留成仇，需要赶紧为她找个归宿。2013年可谓是大数据的花季之年，提亲媒婆似乎踏破了每个数据中心的门槛，大家都对大数据产生了无限的好奇和遐想。天涯何处无芳草，人家大数据才最懂你的心。手牵手，跟她一起走，创造幸福的生活。昨天你来不及，明天就会可惜。

大数据开启了一次重大的时代转型，一场生活、工作与思维的大变革。就像望远镜让我们能够感受宇宙，显微镜让我们能够观测微生物一样，大数据正在改变我们的生活以及理解世界的方式，成为新发明和新服务的源泉，而更多的改变正蓄势待发……套用一句广告词：有了大数据，生活好滋味；更多选择，更多欢笑，就在大数据！

在国家自然科学基金资助项目“节能无线认知传感器网络协同频谱感知安全研究”（编号：61100240）资金的支持下，结合自己多年来在大数据领域的研究经验和体会，特撰写拙作，以期抛砖引玉，为我国大数据事业的发展尽一份微薄之力。

本书是一本关于大数据的幽默科普读物，它使用大量的漫画、故事、笑话、网络流行语、相声小品台词等生动风趣的语言，采取比喻、夸张、排比、拟人等多种表现手法，以独特的视角深入浅出地为大家解读了大数据的产生背景、前世今生、衡量标准和基本特征，描述了大数据采集、存储、分析、处理和展现等关键技术，列举了大数据在政治、经济和生活中的应用实例，分析了大数据在隐私、安全、宕机（死机）、数据公开等方面面临的困境，展望了大数据的发展趋势，最后总结了国内外知名大数据公司的发展策略与应用现状。本书通过大量实例和漫画式插图来帮助读者理解晦涩、枯燥的技术，向读者诠释了大数据的巨大魅力，为初学者打开了一扇深入学习大数据的大门。

本书由郎为民主编，武汉职业技术学院的焦巧，解放军第二炮兵指挥学院的陆雪娟，解放军国防信息学院的刘建国、苏泽友、钟京立、刘勇、陈凯、张国峰、吴帆、陈红、夏白桦、毛炳文、刘素清、邹祥福、瞿连政、徐延军、张锋军、陈于平、余亮琴、张丽红、王大鹏、王昊、陈虎、姜斌、和湘、朱元诚参与了本书部分章节的撰写，高泳洪、蔡理金、王会涛、崔遥绘制了本书的全部图表。孙少兰、刘建中、靳焰对本书的初稿进行了审校，并更正了不少错误，在此一并向他们表示衷心的感谢。

人民邮电出版社的李强老师作为本书的责任编辑，为本书的面世付出了辛勤的劳动。人民邮电出版社对本书的出版给予了大力支持。在此一并表示感谢。

由于大数据仍在发展之中，新的技术和应用不断涌现，加之作者水平有限，编写时间仓促，因而本书难免存在错漏之处，恳请各位专家和读者不吝指出。我的E-mail是：wenlang@163.com。

谨以此书献给我聪明漂亮、温柔贤惠的老婆焦巧，活泼可爱、机灵过人的宝贝郎子程！

郎为民

2014年1月于武汉

第1章 大数据的春天

2013年度最心疼奖应当颁发给歌手汪峰，理由是有一种创伤叫上不了头条。2013年9月13日在微博宣布离婚碰上“菲鹏”婚变；11月9日公开告白章子怡又撞上球队夺冠；11月13日发表新作品，却遇上吴奇隆与刘诗诗公开恋情以及杨幂与刘恺威的婚讯；11月27日登苍山，好不容易攻顶，但同一天，王力宏与李云迪先后认爱，头条再次被抢走；12月2日汪峰发布新专辑，想不到遇上了嫦娥三号的发射，让他第5次冲击头条失败。

与汪峰狂抢头条未果不同，2013年大数据当仁不让地坐上IT江湖的头把交椅（如图1-1所示）。我们在经历了以PC、网络、服务器和存储为中心的时代后，大数据如约而至，它将深刻地改变我们每个人的生活。大数据的江湖很热闹，热闹到许多与大数据没有半毛钱关系的人，都会义无反顾地投身到大数据的江湖里来，包括贩菜的、修车的、卖书的、开店的、做饭的……都纷纷声称要向大数据转型。



图1-1 大数据上头条

从2012年起，大数据一词越来越多地被提及，人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。它光顾过《纽约时报》《华尔街日报》的专栏封面，进入了美国白宫官网的头条新闻，现身在国内一些互联网主题的讲座沙龙，甚至被嗅觉灵敏的证券投资商等写进了投资推荐报告。

一桩桩围绕争夺大数据制高点而展开的并购案，一家家发布大数据战略的IT厂商，一个个关于大数据的传奇故事，一场场以大数据为主题的研讨会议，一名名轮流登场布道的专家大佬（老大，指有权威的人），无一不在宣告：大数据时代来了。

1.1 四部影视剧帮你看懂大数据

盘点近年来的西方影视剧，最靓丽的风景莫过于其中的高科技奇观。这些影视剧使用先进的高科技进行制作，拥有无可比拟的观赏性和征服力，总能给人以极大的震撼和幻想，充分体现了科技和艺术结合的魅力。同时，它们又总能紧跟潮流，与时俱进，将最时尚、最前沿的新东东（东西）融入到影视剧当中，因而一些年轻人对西方影视剧情有独钟也就不足为奇了。

大数据，无疑是2013年IT业内最热的词。厂商热推，媒体热炒。仿佛不谈大数据，不推大数据，不用大数据，就没有了未来。当IT业界正在爆炒大数据的时候，影视圈的导演和主角们自然也不甘寂寞，他们雄赳赳、气昂昂地投身到这场轰轰烈烈的造星运动之中。娱乐没有圈，没有谁规定电影人不能玩高科技！

1.1.1 英剧《黑镜》：“复活”爱人

当深爱之人不幸殒命，你会用科技的力量复活他么？2011年热播的英国电视剧《黑镜》给出了答案：Yes！（是的！）

《黑镜》（如图1-2所示）是英国电视4台于2011年12月播出的迷你电视剧，由英国制片人查理·布鲁克制作。“黑镜”的灵感来源于人人都有的“黑镜子”——每个家庭、每张桌子、每个手掌之间都有一个屏幕、一个监视器、一部智能手机，这是一面反映时下现实的黑镜子。



图1-2 英剧《黑镜》

在《黑镜》第二季第1集《去去就来》中，艾什生前是一个社交网络控（控来自Complex，即情结、极度喜欢的意思），几乎天天泡在Facebook、Twitter这类网站上。每天玩社交网络的艾什伤不起啊！除了做梦的时候不上，其余时间每一分钟都在上网，没事儿就想着点刷新，手贱得根本无法自控。在玛莎的劝说下，艾什同意和她搬到一处偏僻的小屋去住。糟糕的是，艾什在一次车祸中不幸遇难。在艾什的葬礼上，玛莎的朋友萨拉称某公司可以利用艾什生前在社会媒体上留下的“生活轨迹”，再造一个“真实”的、具有“人工智能”的艾什。一开始玛莎认为这个想法太不可思议，但最终还是同意了。此后，玛莎很快发现自己怀孕了。人工智能艾什给玛莎发来很多邮件，玛莎决定回复其中的一封。这个“死后重生”的艾什会做何反应呢？

寂寞无助的玛莎经不起亲情和爱情的双重诱惑，她太爱艾什了！于是，在供应商为玛莎量身定制从文字聊天到语音通话的套餐吸引下，她无法自拔，自愿加入了一个尚在测试阶段的项目，该项目利用艾什在Facebook、Twitter等社交网络上留下的大量数据，重建了一个模拟艾什人格的人工智能（Artificial Intelligence, AI）机器人。

在社交网络上分享信息实际上是一个量化自我的过程。量化自我（Quantified Self, QS）是指通过科技方式将一个人日常生活的各方面，包括物质摄入、身体状况以及体能情况记录下来的一项活动。量化自身的追随者相信收集、分析数据可以改善一个人的生活。

千金难买爱人一笑。剧中替身机器人的生产商采用了软件免费+硬件收费这一营利模式。先提供免费的文字和语音沟通服务令用户产生依赖，然后借机推销昂贵的人形终端设备。当厂商成功地将用户对逝去亲人的感情转移至人工智能，有谁会拒绝一掷千金来挽回爱人的音容笑貌呢？

供应商提供了3种服务：初级服务、中级服务和高级服务。初级服务提供在线聊天服务，但这里的在线聊天是文字类型的，用户通过IM（Instant Message，即时通信）的形式与机器人进行交流；中级服务在文字沟通交流的基础上，拓展至可语音通话，从而使机器人更加立体化和情感化；高级服务则直接给你一个活生生的人。他可以陪你说话，可以陪你睡觉，可以给你做饭，而且他还遵循机器人三大定律（机器人不得伤害人；机器人应服从人的一切命令，但不得违反第一定律；机器人应保护自身的安全）。对于数据提供商来说，初级服务是可以免费提供的，而中、高级服务则是需要收费的。

通过对艾什在社交网络上留下的信息进行分析，获得模型、发现规律、统计比较，最终实现“预测”这一终极目标，即预测在特定的情景下，“如果艾什活着，他会怎么做？”

当然，需要提醒的是，社交网络上不可能包含所有的生活情景，这使得重建的人格中存在数据空白。如艾什身上有颗痣，而替身机器人（如图1-3所示）却没有，因为艾什可能不会在社交网络上公布身体的这些细节信息；替身机器人并不知道艾什生前触碰女友胸部时的性反应，因为这部分属于艾什在网上不能说的秘密；当玛莎将替身机器人赶出卧室时，他不会和玛莎吵架，因为艾什不会把与老婆打架，被老婆抓了个满脸花的照片上传到社交网站；当玛莎让替身机器人跳海时，他不会向玛莎哭着求饶，因为他不知道艾什性格中还有怕老婆的一面；当玛莎驱车播放她和艾什喜欢的音乐时，替身机器人评价“好俗气”，因为他不知道艾什夫妇其实就是一对俗人。



图1-3 人工智能艾什

这些数据的空白环节需要替身机器人自行学习。人工智能学习的知识有两类：一类是数据库；一类是使用过程中用户的实时反馈。例如，剧中的替身机器人从AV（Adult Video，成人影片）中学习各类性爱姿势，还会经常询问女主角某个词的含义。重建人格并填补上数据的缺项后，人工智能就拥有了一个情景库，其中包含可能遇到的各种情况，并有与之对应的处理方式。人工智能接下来要做的，就是将实际遭遇的场景与这个库中的范例进行匹配。

这个“复活”了的男朋友最初通过文字，继而通过语音与女主角聊天，最后甚至通过一具机器人身体“实实在在地”陪伴在女主角身边。在剧中，替身机器人会用艾什的口吻安慰玛莎，会恰到好处地说些艾什风格的俏皮话，但这并不意味着替身机器人能够完全理解人类的行动和情感。女主角将机器人当作复活的男友，但恐怕存在于服务器里的“男友”只是把与她的每一次沟通在情景库中进行匹配，并用糅合了艾什人格和一般化数据的“应对方案”进行回应。

替身机器人的本体是保存在云端服务器的人工智能。眼下方兴未艾的一“云”多“端”思路在剧中的时代已经生根发芽，观众看到的替身机器人同女主角最初用来与人工智能沟通的计算机、智能手机一样，只是一台终端。

科幻从来不是无根之水，剧中的替身机器人涉及的技术已初现端倪，“复活”爱人并不遥远。如果这事儿发生在古代，活生生的一出《人鬼情未了》。还好，剧集发生在未来，所以一切都经得起推断和解释。

此类替身机器人面临着诸多伦理问题：一是中国有个不成文的传统，即人死为大，逝者为尊，入土为安。在亲人不幸离世之后，使用逝者生前留下的数据重建他的人格甚至外貌，并将重建的数据用于订制替身机器人是否合适？是让爱人“复活”还是假以怀念？二是发表在社交网络的大部分数据都是公开的，如何避免有未经授权的机构或个人擅自盗用这些数据，复制出死者甚至活人的替身用于非法行为？三是不同于普通的家政机器人，替身机器人作为爱人的替代品，会在非常私密的场合下使用，如何防止个人隐私的泄露？剧中，女主角在短暂的蜜月期之后，逐渐认识到“复活”的艾什并非原本的那个“TA”。无法狠心摧毁替身机器人的女主角，最终将“TA”关在阁楼，每周上去一次怀念曾经深爱的男友。

社交网络的大数据分析，除了复活逝者，还能在人海中寻找拥有相同思考逻辑与兴趣爱好的“灵魂伴侣”。当曲终人散，科技不该让人沉湎过去，而该助人迈步前行。我们不想去纠结剧集本身的压抑和黑暗以及主创所想要表达的思想，只想通过这么一个剧集，来看一下大数据未来的应用。无疑，在这部剧集中大数据最为直观的应用就是能够生产人。

1.1.2 电影《点球成金》：用数据拿冠军

一名棒球好手走在路上，忽然看到一只小猫在树上摇摇欲坠。他赶忙奔去将小猫接个正着，然后朝一垒方向扔去。在国内，棒球仍属于非常小众的运动，但在美国却红得发紫，本人在美国田纳西大学留学时，每当体育馆有棒球比赛时，诺克斯维尔市的道路堵得跟北京似的，两旁店铺中球迷们边看电视边狂饮酒，搞得跟圣诞节一样。但是，打好棒球其实并不是一件容易的事情，就连棒球巨星米奇·曼陀都

说：“难以置信，打了一辈子球，却依然对它知之甚微！”一个0.4亿元家底的棒球队如何能够与一个1.4亿元家底的球队相抗衡，而且还能赢得1亿元？电影《点球成金》（如图1-4所示）给出了答案。



图1-4 电影《点球成金》

影片《点球成金》改编自迈克尔·刘易斯的《魔球：逆境中制胜的智慧》。讲述的是一个真实的故事，介绍奥克兰运动家棒球队总经理比利·比恩的经营哲学，描写了他抛弃几百年一直依赖的选择球员的传统惯例，采用了一种依靠计算机程序和数学模型分析比赛数据来选择球员的方法。他并没有采用那些像“棒球击球率”这样传统的标准，而是采用了看上去很奇怪的、类似“上垒率”这样的标准。这个方法发现了这项体育赛事的另一面，始终存在却一直被忽略的一面。一个球员怎样上垒并不要紧，不管是地滚球还是三垒跑，只要他上垒了就够了。当数据表明盗垒不实用的时候，即使这会让比赛更有看头，比利·比恩也不会再关注这种华而不实的技能。

2003年，此书出版后在美国掀起了一股热潮，从波士顿、纽约到旧金山、洛杉矶的球迷、新闻媒体乃至金融精英都津津乐道于书中的只言片语。数年后，该书的影响力甚至跨越太平洋传到了欧洲足球界。这本颠覆了美国体育管理层思路的书，讲述了精明的奥克兰运动家棒球队总经理比利·比恩如何采用统计学和数学建模的方式分析数字，从而取得最终胜利的经营哲学。他是逆向投资的表率，用极少的资金经营着这家俱乐部，并使用复杂的计算机程序分析比赛数据，用“数据”的方式将一个小球队打造成超级劲旅，使得这支球队取得了一场又一场的胜利，甚至有能力和大名鼎鼎的纽约扬基队竞争。2011年由原著改编的同名电影正式上映，布拉德·皮特扮演了书中主角奥克兰运动家队的总经理比利·比恩。

《华尔街日报》评价说：“从来没有一部电影将枯燥的数据转化为如此令人愉悦的娱乐体验。”《芝加哥太阳报》则认为：“这是一部聪明、紧张且感人的电影……虽然入场前我已经知道电影的故事，但影片的智慧与深度是我没有预料到的。”该片获得第84届奥斯卡最佳影片、最佳男主角、最佳男配角、最佳改编剧本、最佳音响效应、最佳电影剪辑6项提名，以及第69届金球奖最佳剧情片提名和最佳剧本提名。

“一鼓作气，再而衰，三而竭”。这句话的意思是，做事情必须要一鼓作气，才能成功。可是，影片《点球成金》的拍摄却是一波三折，在经历了大卫·弗兰科尔和斯蒂文·索德伯格的参与和退出之后，这部电影最终被交到了《卡波特》的导演贝尼特·米勒的手中，而影片的剧本也是一改再改，到了米勒这里，已经不知道是几易其稿了，而且影片的制作成本也被一再削减。不过，好在原著小说的作者迈克尔·刘易斯没有在这些事件中公开表态支持哪个导演，否则这部电影还要继续“命运多舛”下去。

在美国职业棒球大联盟（Major League Baseball, MLB）比赛中，比利所属的奥克兰运动家队败给财大气粗的纽约扬基队，三名主力被重金挖走。总经理比利很早就发现自己的天赋不足以成为大联盟的球员，因而他下定决心成为一个棒球界的高层管理人员。他在奥克兰运动家队做球探，8年后就成了总经理。他暗下决心改造球队。

比利是一个“特立独行”、“思维怪异”的家伙，就是在这样一个经理人的掌控下，他的一切行事和工作几乎皆不按常理出牌，处理一切皆采用逆向思维的方式。就是这样一个比利，按照他自己所谓的对事物真谛的顿悟，打破一切惯例常规之后，成功组建和塑造了一支具有强大战斗力的棒球队。

在竞争激烈的美国职业棒球大联盟，比利的奥克兰运动家棒球队无论在人员构成、物质配备，还是在资金实力上都仅仅位于“下三流”之列，不可能像扬基队那样一掷千金来购买高身价的球星。一次偶然的会，他认识了耶鲁大学经济学硕士、大胖子彼得，两人对于球队运营的理念不谋而合。

比利聘请彼得作为自己的顾问，查询所有球员的历史数据，利用数学建模定量分析不同球员的特点，合理搭配，重新组队，颠覆棒球界靠重金挖明星球员的传统理念。在新的赛季中，奥克兰运动家队创造20场连胜的战绩，刷新了大联盟纪录，如图1-5所示。



图1-5 比利和彼得用数据拿冠军

大块头有大智慧。在彼得的辅佐下，比利召集和物色了一批表面看上去都各有瑕疵、性格怪僻，但骨子里却都在棒球运动某方面拥有超强能力的队员，以打破常规、突破传统的经营模式，在一片批评与质疑声中取得了骄人的比赛成绩，甚至达到了比肩实力雄厚的纽约扬基队的程度！

如何在不公平的竞争中以弱胜强？这是比利·比恩面临的难题。所谓“不公平的竞争”便是指美国棒球大联盟不同球队间巨大的薪资差距。2001年，纽约扬基队的总年薪高达114 457 768美元，这个天文数字不仅在当时美国体育界鹤立鸡群，即便是在世界范围内也毫不逊色于皇家马德里、巴塞罗那、曼联之类的足球豪门。然而，奥克兰运动家队同时期的总年薪约为纽约扬基的三分之一，只有39 722 689美元，在全联盟位列倒数第3位。薪资的巨大差距意味着优秀球员的流失，更难以寻觅联盟中那些当红的超级巨星。除了捉襟见肘的工资预算，运动家队还是一支典型的“小市场”球队。一穷二白的状况不仅使其无法招揽球员，甚至连一座像样的专业棒球场都供养不起。自1968年至今，运动家队都不得不与同城另一支美式橄榄球队共用体育场。设施陈旧和全美倒数的观众上座率几乎成为运动家队主场的“特色”。

然而，运动家队却是近年来“投入产出比”最高的职业棒球队。2000——2003年间，他们每赢一场球的成本约50万美元，而扬基队每赢一场的成本则几乎是其3倍（近150万美元）。至于联盟中其他那些“富队”（如巴尔的摩金莺队、德州游骑兵），则需要花费近300万美元才能赢一场“天价”般昂贵的胜利。金融圈出身的作者刘易斯非常敏锐地捕捉到了这个诡异的现象，于是便有了这本书。他希望借此来勾画与解释比利·比恩是如何扮演这么一位近乎“无米”的“巧妇”角色。

棒球是一项强调数据的运动。所谓“数据”不仅包括球队的各项胜败指数，还有每个职业球员的各类成绩——防御率、胜投数、打击率、长打率、全垒打数、打点数等多达几十种类型。在谈论某个球员时，资深棒球迷都会如数家珍般报出一连串数字。若被不谙棒球者听到，或许以为是两位证券分析员在交流工作心得。长久以来，美国棒球界也将这些数据的记录工作看作重中之重。因此，即便是20世纪初某场比赛的交战数据都能毫不费力地找到。然而，在比利·比恩看来，棒球界却没有将这些数据转换成真正提升球队战绩的“不二法门”。基于这个想法，他开始摸索一套全新的方法来解读棒球数据背后的“真谛”，甚至不惜向所谓“百年传统”宣战。

比利的成功之道是运用一整套的数据分析法（Sabermetrics）来代替传统的球队运作，这是一种美国棒球研究协会（Society for Advanced Baseball Research, SABR）所倡导的统计方法。直到现在，美国高级棒球研究协会一直是一个奇特亚文化的中心。比利最重视的是上垒率，而这种统计方法帮助他成为出色的管理人员，也使得运动家队最终成功。

这套全新方法被称为“棒球统计学”，其创始人并不是比利·比恩，而是一位统计学家比尔·詹姆斯。他是刘易斯《魔球：逆境中制胜的智慧》书中的另一位主角。比利认为，棒球界传统的统计数据无法准确反映出球队或球员的价值，也无法准确预测其未来的表现，而对这些数据的解读也缺乏一种科学的方式。于是，他雄心勃勃地设计了一套统计学公式来计算各类既有的棒球数据。在一片批评与质疑声中，比利的“棒球统计学”在奥克兰运动家棒球队的办公室里被铭记了下来。

早在1977年，比尔就自费出版了自己的“研究成果”，但却几乎毫无反响。棒球界的元老们根本瞧不上这样一位从没真正打过棒球的“门外汉”。在他们看来，这类书呆子式的纸上谈兵可能连“票友”水平都不如。问世20年后，穷则思变的比利·比恩才成为第一个真正吃螃蟹的实践者。

比利破天荒地将“棒球统计学”作为球队的经营方针。他尽可能地将球员能力数据化，并以此作为衡量球员能力的唯一标准，而非某些基于主观经验的判断。通过这套统计学公式，比利以有限预算去寻找那些价值被低估的球员。同时，他还强迫整个球队摒弃传统的成绩评估标准：既然让棒球比赛结束的因素是27个出局数，而不是时间；那么就忘记“打击率”、“盗垒”等华而不实的成绩，“上垒率”才最重要。因为只有上垒才能减少出局的概率，并提高得分的概率。比利打破一切常规惯例，就如同伽利略用“太阳中心论”来挑战天主教的权威一样。最终，在全新理念指引下，运动家队在2000年后曾5次打入季后赛，4次获得分区冠军，共赢了1045场比赛。期间，甚至还创下了美国职业棒球大联盟百年历史上连胜20场的空前纪录。从那以后，统计学家取代球探成为棒球专家，其他很多球队也开始争相采用“棒球统计学”来指导球队运作。

1.1.3 美剧《疑犯追踪》：有贼心也会摊上事儿

幼儿园的时候我不谈恋爱，因为不知道什么是贼；小学的时候我不谈恋爱，因为没有贼心也没有贼胆；初中的时候我不谈恋爱，因为有贼胆没贼心；高中的时候我不谈恋爱，因为有贼心没贼胆；大学的时候我不谈恋爱，因为有了贼心，也有了贼胆，可贼却没了。人们经常用“有贼心没贼胆”来调侃自己，未来有贼心可能也会犯事儿。

它几乎无所不能，全天候监视所有人的行踪，聪明地预测出谁是危险分子，谁会遭遇不测……美国政府用它攻击恐怖分子，开发者则用它拯救普通人，这是美剧《疑犯追踪》（如图1-6所示）里的“神器”。这真的只是一部科幻剧吗？



图1-6 美剧《疑犯追踪》

《疑犯追踪》讲述了一位推定死亡的前CIA（Central Intelligence Agency，中央情报局）特工里瑟与一位神秘的亿万富翁芬奇联合起来，运用一套独特办法制止犯罪的故事。该剧每集一个案子，将主角和情报局互相间的猫鼠游戏作为主线，同时配合时代错综的来回倒叙手法，可谓精彩迭连。此剧的大胆之处是将CIA和FBI（Federal Bureau of Investigation，联邦调查局）与罪犯交错复杂的人物关系以及理不清的内部白吃黑、黑吃黑、黑吃白场景一一呈现，特别是莱纳尔·弗斯科，长着黑警察的脸，怀着好警察的心，走着双面卧底的路线，还卖着为黑白两道争相奔走和相互欺瞒的命。

相比帅哥主角单一的角色设定，我更喜欢芬奇的傻劲和执着。在大数据时代，人们感到越来越不安全，生活在监视器下，没有任何隐私，不知不觉地开放着个人信息，通过社交网络、联网信息、条条光缆，随时随地向罪犯或政府输送着情报和证据，这些数据可能会在不久的将来，被拿来用作威胁的资本。

哈洛·芬奇是一位深居简出、极度重视个人隐私，且拥有亿万身家的天才软件工程师，担任着一家软件公司的幕后老板。他为政府开发了一套称作“神器”的、可侦测恐怖攻击的计算机系统（如图1-7所示）。该系统通过观测已有模式来识别有可能进行暴力犯罪的罪犯，可以监视所有人，并预测“有计划性或经谋略策划的犯罪”，诸如“9·11”事件之类的大型恐怖攻击灾难，并提供情报让有关当局防患于未然。



图1-7 “神器”

他雇用了一位被推定死亡的前美国特种部队“绿色贝雷帽”队员、前CIA探员里瑟，二人使用国家级监测技术，加以里瑟的专业技能和芬奇的无限财富，开始“法外执法”，力图在犯罪发生前就对其加以阻止。里瑟的行动引起了纽约市警察局的注意，包括警探卡特和弗斯科。面对无数的犯罪调查，里瑟和芬奇发现能够改变一切的关键就是找到正确的人、正确的信息和正确的时间。

强大的计算机集群将整个纽约市的摄像头整合在一起，结合每个人的信用卡记录、医疗及社会保险记录、行车罚单等各种数据，推断并锁定危害别人或即将被害的自然人。

1.2 大数据汹涌来袭

2013年3月1日，工业和信息化部电信研究院在北京召开了2013年ICT深度观察大型报告会暨移动互联网白皮书、中国通信产业十大关键词发布会。电信研究院在会议上发布了《移动互联网白皮书（2013）》以及“2012年中国通信产业十大关键词”的评选结果，云计算、智能终端、TD-LTE、宽带中国、移动互联网、物联网、网络与信息安全、微博、大数据、微信入选十大关键词。

近年来，大数据的概念受到了学术界、商界甚至政府的热捧，一时间大数据无处不在，这让同时代其他的IT技术相形见绌，无地自容。数据正在迅速膨胀并变大，它决定着企业的未来发展，虽然目前企业可能还没有意识到数据爆炸性增长带来的问题隐患，但是随着时间的推移，人们将越来越重视数据的作用。正如《纽约时报》2012年2月的一篇专栏中所称：“大数据”时代已经降临，在商业、经济及其他领域中，决策将日益基于数据和分析作出，而不是基于经验和直觉。

大数据是2013年最火的科技词语，每个人都在谈论大数据，仿佛谁不谈论大数据就落伍了似的。2013年的大数据已经从之前的虚无缥缈逐渐开始落地，各种类型的大数据应用开始落实到实践中去，为企业和个人都带来了便利。套用赵本山在小品《功夫》中的台词：反正你接受它也来了，不接受它也来了，接受不接受大数据时代它都带着诚意扑面而来。

1.2.1 大数据有多大

互联网所涵盖的范围非常广阔，人们产生、分享和消耗的数据量很难以实体形式来衡量，那么到底有多少数据产生呢？

最近，一张以“一天之间，互联网上要发生多少事”为主题的统计图，在各大科学网站上流传，如图1-8所示。其结果是非常惊人的，真是“不看不知道，世界真奇妙”！互联网每天产生的流量信息可以装满1.68亿张碟；每天发出2940亿封邮件，如果这些是实体信件，则美国需要花费2年时间来处理；每天的社区论坛上发出200万个帖子，相当于美国《时代》杂志770年的文字量；每天世界各地有1.72亿人访问Facebook，4000万人访问Twitter，2200万人访问LinkedIn，2000万人访问Google+，还有1700万人访问Pinterest；人们每天在Facebook上耗费的时间总计47亿分钟；“状态”的更新达5.32亿人；每天有2.5亿张图片上传到Facebook，如果把它们都打印出来，堆起来相当于80座埃菲尔铁塔的高度；人们每天在Netflix观看2200万小时的电视电影节目，相当于全美剧院3天播放电影的总时长；每天人们将86.4万小时视频上传到YouTube，即使不间断地播放，全部播完这些视频也需要98年；每天有1870万小时音乐在Pandora（流媒体音乐网站）上播放，如果一台计算机从公元元年开始播放Pandora的音乐，那么它现在仍然处于播放状态；每天新增1288款应用，下载次数超过3500万次；每天卖出37.8万台手机，高于全球每天出生的婴儿数量（37.1万）。是数字时代的战利品还是一堆数字垃圾，一切由你来决定。庆幸的是，我们不用像以前那样弄个仓库把它们搬进去存储起来。



图1-8 互联网上的一天

在现实生活中，一分钟也许微不足道，连抽袋旱烟的工夫都不够，但是当你用一分钟，看完这行文字的时候：新浪已经发送了2万条微博；淘宝已经卖出了6万件商品；人人网发生了30万次访问；百度产生了90万次搜索查询；YouTube用户上传48h的新视频；电子邮件用户发送2.04亿条信息；Google收到超过200万次搜索查询；Facebook用户分享了68.4万条内容；品牌和企业收到3.5万个“赞”；消费者在网上购物花费了27.2万美元；Twitter用户发送了超过10万条微博；苹果公司收到大约4.7万个应用下载；Tumblr博客用户发布2.8万个新帖子；WordPress用户发布347篇新博文；Instagram用户分享36000张新照片；Flickr用户添加3125张新照片；Foursquare用户执行2083次签到；全球有571个新网站诞生；移动互联网新增217个用户。图1-9将帮助读者从一分钟互联网数据读懂大数据。



图1-9 从一分钟互联网数据读懂大数据

欢迎访问：电子书学习和下载网站（<https://www.shgis.cn>）

文档名称：《漫话大数据》郎为民 著.epub

请登录 <https://shgis.cn/post/1060.html> 下载完整文档。

手机端请扫码查看：

