

04010r01000101061000010101020r0T01nUr 均070101or0!05101020791010E000totui

01010i0fODr0r,0ut607oroT0194 (Y0rr6C01040rg10 每您70F0!0+010-01Gr0r0107Q2910

0iu100T0i91070701010181srorchfor0E0+分40104069461pnu501010p 两FGfoo 010f

iG000340601003106010106ro1040061070 01010100 力西0101010Tur91010070r0r05

1011001001

**第16章** **数据科学案例**



在不同来源、各种类型的数据之上，利用各种分析方法，挖掘其中蕴含的规律，获得 对事物的深刻理解，是数据科学的目标。数据科学技术已经应用到各个行业，包括电子商 务、政府服务、医疗、教育、科学研究等，并取得了显著的效果。本章选取近年来数据处 理/分析技术在各行各业应用的若干精彩案例加以介绍。

**16.1** **Google流感趋势预测(Google** **Flu** **Trends)**

Google Flu Trends 是Google 公司提供的一个web 服务，最早于2008年启动，它基 于Google 搜索引擎的查询日志，预测流感的爆发。

Google Flu Trends 使用如下的方法进行流感爆发预测。首先，在2003—2008年间从 美国本土提交到Google 搜索引擎的50000000条查询的基础上，构造一个时间序列数据 集。根据IP 地址可以确定某个查询是从哪一州提交的。对于每个州单独构造一个时间序 列，并且进行规范化处理。接着，建立如下的回归模型：

logit(P)=β+β×logit(Q)+ε

式中，P 为看医生的病人里流感疑似病例的比例 (Percentage of Influenza-like Illness (ILI)Physician Visit);Q 为和流感疑似病例相关的搜索引擎的查询 (ILI-Related Query Fraction Computed in Previous Steps);β 为截距 (Intercept);β1 为回归系数 (Coeffi- cient);e为误差。 Google Flu Trends的预测准确度见图16-1。

在历史数据上进行检验，看看哪些查询最能够预测实际的流感疑似病例(从CDC 获 得实际数据， CDC 即美国疾病预防与控制中心， U.S.Centers for Disease Control and Prevention) 。 保 留Top 45个查询，通过训练获得模型的参数。最后.这个模型用来预测 美国全境各个州的流感爆发。

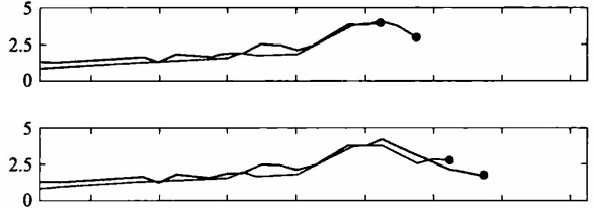
在2009年的冬季流感大流行中， Google Flu Trends运用已经训练的模型，成功地对 流感的爆发进行了预测，具体到特定的地区和州。该模型预测到了2010年2月在Mid-At- lantic 地区的病例峰值 (Case Spiking in the Mid-Atlantic Region),比 CDC 的报告早2个



**Data available as of 4 February 2008**



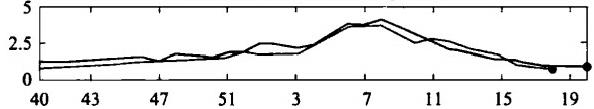
**Data available as of 3 March 2008**



ILI percentage

**Data available as of 31 March 2008**

**Data available as of 12 May 2008**



Week

**图16-** **1** **Google Flu Trends的预测准确度**

说明：以2008年数据作为测试样本.Google Flu Trends的性能表现如下：在第五周，该模型预测到在 Mid-At- lantic地区，流感疑似病例比例 (ILI Percentage) 将有一个急刷的上升：2008年3月8日，即第8周，流感疑似病例 比例到达最高点；最后，在第9周和第10周，这个比例急剧下降。这些预测后来都由CDC的流感疑似病例实际数据 (CDC ILI Data) 证实。

星期左右。尽早预测到疾病的爆发，疾控部门就可以提前采取预防和控制措施 (Preven- tion and Control Measure), 减少疾病的进一步传播，减少死亡病例和损失。

Google Flu Trends 是群体智能 (Collective Intelligence) 用于识别趋势和进行预测的 一个活生生的案例。搜索引擎收集的大量搜索日志体现了人们的需求和倾向性。这些趋势 不是这么显而易见的，只有通过适当的分析，我们才可以从中了解到事物发展的规律，获 得对事物的理解。

早期的Google Flu Trends 获得较好的预测效果，预测的准确率达到97%,比CDC 的 预测准确率要高得多。但是2011 — 2013年， Google Flu Trends的预测结果不是那么准确 了，它明显高估了流感的流行。即便这样，我们也不能否认Google 的开创性工作。这个 工作启发了后续的研究，比如奥斯纳布吕克大学 (Osnabruck) 的认知科学中心 (Institu-

te of Cognitive Science) 发 起 了 新 的 流 感 预 测 项 目 (Flu-Prediction Project)。他 们 把 Google 的思想继续向前推进，把来自社交媒体比如 Twitter 的数据和CDC 的数据结合起来， 构造更加复杂的模型，推测疾病在时间和空间上的传播 (Spatial and Temporal Spreading)。

**16.2** **Target利用数据分析预测到一个高中女生的怀孕**

Target 是美国 一家大型连锁零售商店(见图16- 2)。2012年，围绕 Target 发生了一



件有趣的事情。明尼苏达州一家Target 门店被客户投诉， 一位中年男子指控Target 将婴 儿产品优惠券寄给了他的女儿，他的女儿只是一个高中生，实在是不可理喻。但是没过多 久他却给 Target 来电道歉，因为经他逼问，他女儿承认自己真的怀孕了。



**图16-2** **Target 连锁零售商店**

这位高中生没有告诉过他父亲她怀孕了，也没有在Target 调查问卷上留下过类似的 记录。Target 是如何了解到她怀孕了，给她发放婴儿产品优惠券呢?首先，它建立了大型 的数据仓库，围绕用户ID, 数据仓库记录了如下信息并且关联起来，包括用户的购买历 史、信用卡的使用情况、对客户支持中心的电话投诉、调查问卷反馈、 Email 的点击、网 站的访问等信息。在上述信息之上，Target 还补充了其他信息，包括人口统计数据 (De- mographic Data), 比如年龄、种族、教育背景、婚姻状况、小孩的数量、预期收入、工作 历史以及生活中的各种事件，包括上次搬家是什么时候、是否离过婚、是否宣告破产等。

Target 把注册到商店的Baby Shower Registry①的购物者和这些购物者的历史购买行 为关联起来，创建了一个模型。这个模型可以给所有的女性购物者一个可能怀孕的评分 (Pregnancy Prediction Score), 预测该女性购物者是否已经怀孕。通过该模型的评分， Target 可以找出哪些妇女有可能已经怀孕，即便她们并没有告诉其他任何人，也没有通过 问卷调查等方式告知 Target。

针对这个用户群体 (Customer Segment),Target针对性地进行婴幼儿用品促销，甚 至可以精准到根据怀孕的不同阶段进行不同的营销，从而提高产品的销量。通过数据分 析，Target 发现孕妇在怀孕的不同阶段的购物习惯是不一样的。在怀孕的前20周，孕妇 开始购买钙、镁、锌等补充营养成分的药片。从第四个月到第六个月 (Second Trimes- ter), 孕妇开始购买更宽松的牛仔裤 (Jeans), 以及大量的消毒杀菌剂 (Hand Sanitizer)、 不加香水的洗液 (Unscented Lotion)、不加香水的肥皂 (Fragrance Free Soap) 以及棉球 (Cotton Ball) 等 。Target 找出了25种孕妇经常购买的产品。

Target 通过了解用户的历史购买行为以及生活的变化如何改变购物习惯(搬家后可能 购买一些家具用品),把营业收入从2002年的440亿美元提高到2010年的670亿美元， 这是一个了不起的增长。Target 公司总裁Gregg Steinhafel在对公司的投资者做报告时提 到 ，Target 公司对母亲以及婴幼儿类产品的营销给予了特殊的关注和投入。这些产品的销 售额对整个公司销售额的提高和公司走向成功确实起了巨大的作用。

①Baby Shower,是西方的准爸妈为即将出生的孩子举办的Party。Baby Shower Registry,是由孕妇选定商家注 册自己想要的礼物，到时朋友们可以按单选购，并且由商店送货给准妈妈的一种商家促销方式。



**16.3** **互联网舆情监控与管理**

网络空间是一个开放、自由的空间，人们可以自由地发表自己的见解.其中，微博、 博客等是近几年来兴起的一种自媒体，具有传播主体分散、传播速度快、受众广泛等 特 点 。

网络舆情监控系统，是利用搜索引擎技术和网络信息挖掘技术.通过网页内容的自动 采集、处理、聚类、分类、主题检测以及统计分析，实现企事业单位、政府部门对相关网 络舆情监督管理的需要而设计的系统。该系统能够提供舆情简报以及分析报告，为决策层 全面掌握舆情动态，做出正确舆论引导，提供分析依据。

自媒体的兴起给舆情管理带来了新的挑战。主要的挑战体现在两个方面； 一个是需要 迅速准确捕抓到微博、博客等信息：另一个是需要适时对舆情进行引导。

近年来，国内外涌现了大批提供舆情监控系统的创业公司，各自发布了舆情监控拳头 产品。包括军犬网络舆情监控系统、鹰隼网络舆情监控系统、 Goonie 网络舆情监控系统、 麦知讯第三方网络舆情监控系统、中国舆情网PALAS (帕拉斯)网络舆情监控系统、红 麦软件舆情监测系统等。

为了实现对包括微博、博客，以及论坛、微信、传统媒体的电子版等各个网络信息发 布平台的舆情监控，首先，需要数据监测技术实现对上述各个数据源的数据爬取和保存. 并对爬取的数据进行适当的预处理，比如对图片、音频、视频进行自动识别等。其次，需 要大规模数据存储技术。通过建设具有海量数据存储能力的大数据平台，实现对大规模结 构化数据、非结构化数据的高效读写、存储和交换。再次，需要数据挖掘技术，从海量数 据中快速挖掘有价值的信息，发现数据背后隐藏的规律性。利用关联分析、聚类分析、话 题分析、情感分析等技术，自动分析网上言论蕴含的意见和倾向性，揭示舆情发展趋势。 最后，需要安全技术，保证数据安全，包括身份认证、授权、入侵检测、防火墙等技术。

舆情监控系统不仅能够提供信息监测、分析功能，而且应该能够提供报警、处置功 能。舆情监控系统能够提供不同的信息观察角度，比如从个人角度分析出个人言论倾向、 个人社交关系、个人与事件关联关系等，从事件角度分析出事件发展趋势、事件言论倾 向、事件与人物关联关系等。

**16.4** **投资与信用**

**16.4.1** **大数据指数碁金**

普通老百姓为了资产的保值增值需要进行投资，他们可以选择自己进入楼市或者股 市，也可以申购基金.把钱交给基金公司进行投资。近年来，互联网公司利用自身掌握的 数据和分析能力，力图为投资增加新的决策依据。它们和基金公司合作，推出基于大数据

分析的基金产品，提供了新的投资选项。

百度百发策略100指数是中证指数与百度金融中心、广发基金合作编制的互联网大数 据指数。百发100指数的选股模型依赖百度金融的大数据分析技术，结合传统的交易数据 分析、基本面数据分析以及量化选股策略，实现了成分股的选择和调整。广发基金于2014 年10月30日成立了广发中证百度百发策略100指数基金，供投资者选择。图16-3为该 基金在2016年5月至2016年11月的收益走势图。

**广发百发100指数A[000826]基金实时行情好买基金网**

**0.9863** (+0.54%) 实时估值

**2016/11/1515:03:30(北京时间)**

**15.8%**

11.7%

7.55%

3.42%

-0.71%

05/13 **11/14**

**估值分时图**  **收益走势图** 

昨日净值 累计净值

昨日涨幅

**一周涨幅**

近三月涨幅 近三月排名

0.9810

1.3410

+0.20%

+2.08%

+5.37%

60/333

**图16** **-** **3** **广发百发100指数A [000826] 基金收益走势图**

新浪财经也在大数据指数方面积极投人，推出了i100 指数和 i300指数。通过对新浪 在财经领域的大数据以及社交媒体数据进行定性与定量分析，挖掘投资者情绪，并参考股 票基本面数据以及市场整体情况分别选取综合排名靠前的100只和300只股票组成指数样 本股。基于上述指数的南方大数据100指数基金、300指数基金也适时发售，给投资者以 多样选择。图16-4为截至2016年11月15日的南方大数据100指数基金的净值和收益情 况。

**南方大数据100指数(001113)** **用手机查看**

最新估值：0.9438个2016-11-1515:00:00

般要平 中二风的 段星译吸 暂老

**单位净值**

**0.9382**

涨跌幅

**0.30%**

近3月张幅

2.59%

**近1年涨幅**

**1.08%**

**近3年涨幅**

成立日期：2015年04月24日

累计单位净值：0.9382元

最新规模：71.01亿

累计分红：0元

管理人：南方基金

**基金经理：雷俊**

**图16** **-4** **南方大数据100指数基金净值与收益**

虽然大数据指数基金的收益好坏情况还需要更多时间的考验，但是这些指数基金给我 们提供了 一 个启示，大数据分析的结果可以帮助我们进行决策，包括投资决策。

**16.4.2** **基于大数据的信用评价**

深圳前海微众银行是2014年7月首批获得中国银监会批复成立的3家民营银行之一。 2014年12月28日，微众银行官网面世，成为第一家上线的互联网银行。它在业务模式上 定位于“个存小贷”,服务于个人消费者和小微企业用户，经营范围涉及个人及小微企业 存款、贷款等。

在用户的信用评价方面，微众银行充分利用腾讯的数据资源和数据处理技术优势，实 现基于大数据的信用评估。在官方报道中，微众银行的优势是“该银行既无营业网点，也 无营业柜台，更无须财产担保，而是通过人脸识别技术和大数据信用评级发放贷款”。

微众银行的征信体系，从不同数据来源中 · 采集并处理包括电商交易、即时通信、 SNS (社交网络)、关系链、游戏行为、虚拟消费、媒体行为等信息，分析用户的登录行 为、在线时长、虚拟财产、购物习惯、支付频率、社交行为等，对用户进行画像，并利用 统计学、机器学习等方法，得出信用主体的信用得分，授予相应金额的贷款。它运用社交 网络上的海量信息，提供了一个全新的信用评价角度，创建了一个新的体系。微众银行的 大数据征信，“以信用作担保，用数据防风险”,其最大的优势是降低征信成本。

2015年1月4日，李克强总理来到微众银行考察，他敲下电脑回车键后，卡车司机徐军 就获得了3.5万元贷款。这是微众银行的第一笔放贷业务，它通过人脸识别技术和大数据信 用评级完成贷款发放。这个客户是第一位受到微众银行特别邀请的客户(见图16-5)。



**图16-5** **李克强总理视察微众银行**

在试营业期间，微众银行通过内部分析和信息筛选，小范围地邀请目标客户群体参与 试营业，并且逐步增加受邀客户的数量。正式营业以后，它将致力于服务工薪阶层、自由 职业者、进城务工人员、普通白领等草根阶层(这些年轻人都善于使用互联网),以及符 合国家政策导向的小微企业及创业企业，实现其普惠金融的目标。

基于大数据的征信，到底可信不可信，或者说大数据能否解决征信问题，人们莫衷一 是。比如，美国征信公司FICO, 通常只收集个人的十几项基础数据，便可以做到99%的 准确度。大数据里面包含着噪声，人们的网上行为不一定反映其真实意愿。如果数据与用 户的信用状况没有直接挂钩，那么数据的效用就会打折扣。这些都是基于大数据的征信系 统需要解决的问题。

BAT(Baidu,Tencent,Alibaba) 之一的阿里巴巴集团推出的阿里小贷，其业务也

是给个人或者小微企业提供贷款，它依赖于阿里巴巴拥有的庞大的电商交易数据，作为信 贷管理决策的基础。在淘宝和天猫上的电商小微企业，都成为阿里小贷的潜在客户。

**16.5** **IBM** **Watson计算机与医疗**

2011年，IBM 研发的沃森 (Watson) 计算机，在智力问答节目《危险边缘》中打败人 类选手，获得胜利，给人们留下了深刻的印象(见图16-6)。赢得智力问答比赛冠军，并不 是 IBM 研发Watson 的最终目的。IBM 的愿景是利用大数据分析、认知计算技术与人工智能 技术，解决各行各业的实际问题。目前， IBM 已经把该电脑系统的应用延伸到了医疗领域。



**图16-6** **IBM Watson打败两个伟大的人类冠军**

注：图片来源IBM。

2015年，东京大学的医学科学研究所的医生，借助 Watson 计算机成功诊断了一位病 人的病情，她患有一种罕见的白血病。

这位病人于2015年1月来到东京大学的医学科学研究所附属医院入院。医生的最初 诊断是急性骨髓性白血病 (Acute Myeloid Leukemia), 是 白 血 病 ( 俗 称 血 癌 (Blood Cancer)) 的一种。

经过化疗之后，医生观察到她的恢复过程非常缓慢，于是怀疑自己的最初诊断可能有 问题。医院的研究团队使用Watson 计算机寻求解决办法。 Watson 超级计算机是基于云平 台的计算机，它使用人工智能技术，分析和互相参照了 (Cross-Referencing) 世界上各个 研究机构的上千万篇关于肿瘤学 (Oncology) 的论文里的数据。

从大量的数据中， Watson 计算机迅速把相关的信息提取出来。 Watson 对照了病人 DNA 里的上千个基因突变 (Genetic Mutation) 和 Watson 已经建立的数据库，把和病人 的病情相关的基因突变从上千个基因突变里过滤出来。有些基因突变是和家族遗传特点 (Hereditary Characteristics) 相关的，但是和疾病无关。

利用其强大的计算能力， Watson 仅仅用十分钟就完成了这个任务(见图16- 7),普 通的人类科学家则需要两周时间才能阅读完相关材料，对数据进行比对和分析，然后做出 诊断。



**图16-7** **医生使用Watson计算机辅助进行医疗诊断**

资料来源：<https://www.ibm.com/watson/health/imaging/.>

Watson 的分析结果帮助研究人员和医生得出了正确的结论，病人得了一种罕见的白 血病，这种白血病是由骨髓增生异常综合征 (Myelodysplastic Syndromes) 引起的。正确 的诊断为后续的治疗争取了时间。我们不能说 Watson 拯救了病人，但至少可以说它帮助 医生拯救了病人。

Watson 计算机的数据处理能力给人们留下了深刻的印象，但是它仍然可能犯错。随 着工程师不断改进其算法，在不远的将来(也许是10年之内)其能力将获得极大提升， 在医疗领域大展其能，成为医生的有力助手。

在医疗领域，Watson 计算机还可以完成一系列任务，包括：(1)在医生严重短缺的 国家和地区，帮助进行癌症病人的诊断和治疗。IBM 已经和印度的Manipal Hospital( 印 度第三大医疗网络)合作，使用Watson 计算机技术，在Manipal Hospital的16个医疗中 心，每年帮助诊断和治疗200000个癌症患者。(2)依托基于云平台的Watson 数据分析 服务，开发基于智能终端的App, 提供健康服务。2015年12月， IBM 发布Nutrino App。 这是和Nutrino 公司 (Nutrino Inc.) 合作推出的智能终端应用，它给孕妇提供科学的、 个性化的营养餐推荐 (Meal Recommendations) 和全天候的营养建议(7×24 Nutrition Support) 。(3) 基于病人基因轮廓的个性化治疗方案。利用Watson 的深度学习能力，把 病人的 DNA 数据转换成一个基因画像 (Genetic Profile), 结合相关的医学论文的搜索、 分析，医生可以基于分析结果.为病人提供更加个性化的治疗方案。

百度公司在人工智能和大数据处理方面进行了持续的投入。他们开始致力于利用大数 据和人工智能技术改善医疗服务。2016年底，百度公司的CEO 李彦宏提出互联网+医疗 的四个层次的愿景。(1)第一个层次是(O2O 服务。也就是怎样通过线上服务把用户引流 到线下，并分发到那些适合处理用户疾病的医院去。百度医生平台现在已经有50万名医 生参与咨询，累计有800万人通过百度医生平台来获得相关的医疗服务。(2)第二个是智 能问诊。在北大国际医院做过的一个测试显示，在80%的情况下，百度医生的诊断和北大 国际医院的医生诊断是一致的，它的准确度还将不断提升。在一些罕见病的诊断中，百度 医生的表现也很好。要达成这个目标，不仅需要对大量的医疗知识进行机器学习，而且需 要对病人表述的理解能力进行不断提升，也就是需要对自然语言理解的强大能力。(3)第 三个层次是基因分析和精准医疗 (Precision Medicine)。很多疾病是多个基因共同作用导 致的。搞清楚一个疾病是由哪些基因共同作用导致的需要大量的计算。搞清楚这些以后， 就可以使用基因编辑等先进的治疗手段，进行精准治疗。(4)第四个层次是新药研发。已 知的有可能形成药物的小分子化合物的数量大概是10的33次方(10³3),这是一个非常大

的数目。对未知分子式进行大量的筛选，找到有效的新药，需要极其强大的计算能力和更 、 加先进的算法，数据科学将发挥重要的作用。

**16.6** **奥巴马竞选团队利用大数据**

**分析技术，帮助奥巴马竞选连任成功**

2012年，美国总统奥巴马竞选连任，他的竞争对手是罗姆尼。2012年11月，奥巴马 大选连任成功。胜利果实部分归功于大数据分析技术，他的竞选团队收集了大量数据，并 且进行了深入分析，从而制定了有效的竞选策略。《时代》杂志断言，在政治领域，仅仅 依靠直觉与经验进行决策的优势在急剧下降，大数据的时代已经到来。

奥巴马的竞选团队是怎么做到的呢?首先，功劳归于奥巴马竞选活动负责人Jim Mes- sina, 是他决定用数据和数据分析来驱动整个竞选活动。竞选团队分成若干小组，现场小 组负责组织志愿者、处理注册事宜、鼓励选民投票给奥巴马等。数字小组负责在线展示、 Email 营销、在线募集资金、社交媒体管理等事宜。通信与媒体小组负责媒体访谈、广告 购买投放等事宜。财务小组负责资金募集的总体策略。这些小组掌握的数据都交给数据分 析小组，数据分析小组的分析结果指导这些小组的后续行动。

民主党全国委员会 (Democratic National Committee) 的数据架构负责人(Director of Data Architecture)Chris Wegrzyn描述了建立一个数据分析驱动的竞选团队 (Analytics- Driven Campaign) 的挑战、机遇和路线图。基于数据分析驱动的竞选，围绕数据本身， 进行建模和实验模拟，用分析结果指导竞选策略。这些数据包含了关于选民和竞选活动的 事实性信息，模型则对选民进行精确的画像和理解，精细到单个选民的粒度，实验结果则 告诉竞选团队，具体的竞选活动将怎样影响到选民的倾向。

奥巴马的竞选团队建立了由100个分析师组成的强大的数据分析团队， Dan Wagner 是首席数据科学家(见图16-8)。为了实现数据的实时分析，他们使用了HP Vertica MPP 数据库，并且使用R 和 Stata 统计软件包建立了预测模型。他们收集的数据达到几十 TB 的规模。



**图16-82012年奥巴马竞选团队的首席数据科学家** **Dan Wagner**

资料来源：<https://www.technologyreview.com/s/509026/how-obamas-team-used-big-data-to-rally-voters/.>

竞选团队利用数据分析结果，帮助奥巴马较快筹集到了10亿美元资金，改变电视广 告投放策略、提高投放效果，制定出拉拢摇摆选民的最有效的拉票方法，比如邮寄信件、 电话或者利用社交媒体等。通过分析他们发现，参加“快速捐赠计划”①的人所捐献的资 金是其他捐献者的4倍，这一计划在后期得到大力推广。在大选最后阶段，竞选团队决定 让奥巴马在知名社交新闻网站Reddit 上回答问题，因为数据分析结果显示，他们的很大 一部分目标选民就在Reddit 上。

在整个竞选活动中，竞选团队发起的两个计划展示了数据分析对竞选活动的辅助决策 的作用，这两个计划分别是AirWolf 和 MediaOptimizer 。(1)AirWolf 的目的是打通现场 小组和数字小组，实现数据共享。在2008年的竞选活动中， 一个普遍存在的问题是，现 场小组获得的关于选民的信息，数字小组难以跟进。通过AirWolf, 当现场小组通过门对 门的方式接触了某个选民以后，该选民的一些喜好被记录下来。之后，数字小组基于选民 的喜好，针对性地进行 Email 营销，Email 的内容是针对选民关心的问题进行裁剪的。 Email 营销的精准度提高了，有利于争取摇摆的选民。(2) Media Optimizer的目的是更加 有效的广告购买和投放。在此之前，电视广告购买和投放，仅仅依赖于一般的人口统计学 方面的数据，花销大，效率不高。使用Media Optimizer之后，竞选团队使用统计分析模 型，从数据库里确定目标选民，了解他们喜欢什么节目。然后，结合人口统计学方面的数 据以及广告价格数据，确定最优的广告投放策略，既减少了成本，又提高了广告的效果。 比如，奥巴马竞选团队在Cable TV 上的广告购买量是罗姆尼竞选团队购买量的两倍，最 后证明是行之有效的。

奥巴马竞选团队基于数据分析驱动的竞选获得了成功，启示有两个：(1)通过整合来自不 同渠道(竞选团队的不同小组)的数据，竞选团队可以对选民和竞选活动进行360度的观察， 获得全面的了解；(2)利用数据分析系统的实时分析能力，竞选团队可以尽快做出反应。

**16.7** **数据科学与科学研究**

科学研究的观测结果和模拟结果的数据规模往往非常大。下面举两个例子，分别是 CERN(European Organization for Nuclear Research, 欧洲原子能研究中心)的大型强子 对撞机实验结果分析和 LIGO(Laser Interferometer Gravitational-Wave Observatory) 观 测结果分析。

CERN 的大型强子对撞机是世界上最大的粒子加速器，通过把质子加速到接近光速， 然后进行对撞，在对撞的残骸中，寻找亚原子粒子的蛛丝马迹，从亚原子层面了解宇宙的 秘密(见图16-9)。

希格斯玻色子是物理学“标准模型”预言的一个基本粒子。这个模型是统一描述宇宙 强力、弱力和电磁力这三种基本力以及组成所有物质的基本粒子的理论。希格斯玻色子一 直没有在实验中观测到。2012年， CERN 根据其最新实验结果的分析，宣告发现了希格 斯玻色子。这个发现使得人类到达了理解自然的新的里程碑，成为人类理解物质的更多细 节的基础。



①Quick Donate Program, 即可以通过在线或者短信的方式进行捐赠而无须重复输人信用卡信息。

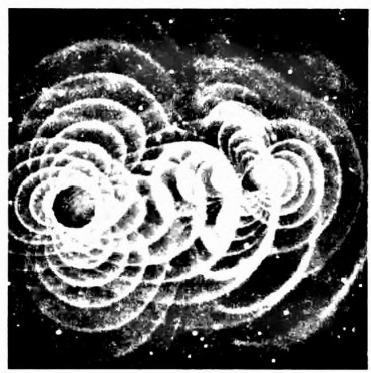


**图16-9** **计算机生成的质子(Proton)对撞图像**

注：该对撞成为支持希格斯玻色子存在的证据(图片来源CERN)。

大型强子对撞机使用了大量的探测器，以纳秒的时间精度进行数据收集，产生了极大 规模的数据，达到PB 级别。如何快速准确地分析这些数据是一个严峻的挑战。数据分析 工具能够有效处理这些巨大规模的数据集，从大量数据中寻找支持希格斯玻色子存在的证 据，即数据中所表现出来的特性和模式。

激 光 干 涉 引 力 波 观 测 站 (Laser Interferometer Gravitational-Wave Observatory, LIGO) 是加州理工学院 (Caltech) 和麻省理工学院 (MIT) 联合发起的实验室，实验资金 来源于美国国家科学基金会，现在已经有来自世界各地的其他大学参与其中。建设LIGO 的目的是寻找宇宙中的引力波，从而验证黑洞的存在和检验爱因斯坦的广义相对论(见图 16-10) 。LIGO 有两个观测点，分别是位于路易斯安那Livingston Parish的 Livingston 观 测点和华盛顿Hanford 的 Hanford 观测点。



**图16-10** **两个黑洞所产生的引力波的3维模拟图**

注：图片来源NASA。

在物理学上，引力波是爱因斯坦广义相对论所预言的一种以光速传播的时空波动。宇 宙中大质量天体的加速、碰撞和合并等事件，才可以形成强大的引力波，但能产生这种较 强引力波的波源，距离地球都十分遥远，传播到地球时变得非常微弱。

美国科学家于2016年2月11日宣布，他们在前一年(2015年)9月首次探测到引力

波，印证了爱因斯坦100年前的预言。 LIGO 实验装置的观测数据，其数据量是非常庞大 的 。LIGO 实验室的科学家使用Python 语言，专门设计和编写了分析软件，对这些数据 进行及时的分析，从中寻找引力波的证据。

**16.8** **奥斯卡大奖预测、总统选举预测、世界杯预测**

2013年，微软研究院(纽约)的经济学家大卫 · 罗斯柴尔德 (David Rothschild) 利 用大数据，成功预测第85届奥斯卡金像奖24个奖项中的19个。2014年.他改进了已有 的模型，成功预测第86届奥斯卡金像奖24个奖项中的21个。这件事情向人们展示了数 据科学的魅力，成为人们津津乐道的话题。①大卫 ·罗斯柴尔德的模型使用了交易市场数 据 (Market Data)、调查数据 (Polling Data) 和历史数据 (Historical Data)。交易市场 里，人们对将要到来的事情有一个预期，他们据此进行买入和卖出。调查数据则反映了特 定人群的情感和期待。罗斯柴尔德对这些数据进行预处理和规范化，在此基础上创建预测 模型进行预测.取得了令人振奋的效果。

2014年巴西世界杯期间.百度推出了世界杯预测栏目。②为了进行预测.百度预先设 计好了赛事预测模型。为了让这个模型更加精准，百度大数据研究院指派资深数据科学家 团队，全面搜索过去5年内全世界987支球队的3.7万场比赛数据。同时，百度公司与国 内著名彩票网站乐彩网、欧洲 SPDEX 指数等公司建立战略合作伙伴关系，将博彩市场数 据融入预测模型中。该模型在淘汰赛阶段取得了非常高的预测准确率。

Nate Silver 是一位非常年轻的统计学家(2012年34岁),他成功地预测了2008年美 国总统大选和2012年美国总统大选的结果。2008年他成功预测了50个州中49个州的选 举结果，2012年他成功预测了50个州的选举结果。他的模型依赖于大量的变量，包括历 史上的选举结果、各种民调数据等，在此之上进行精细的计算和分析。③

在2012年大选的最后阶段，对纽约时报网站的访问 (NY Times Website) 有 2 0 % 的 流量访问了Nate Silver的博客，用户就是想上去看一眼，Nate Silver的预测结果是怎么 样 的 。

**16.9** **用户画像与精准广告**

广告投放得当，商家可以借此提高商品的销量，广告投放不当，则商家在付出广告成 本以后却没有带来应有的收益。精准的广告投放依赖于对用户的理解(用户画像),然后 针对用户的不同特点，为其定制不同的广告内容，推送到他们面前，以期提高转化率。

2016年，12306官方网站全面上线了商业广告， 一时间引起了广告商的疯抢，展示了



①<http://predictwise.com/.>

②<http://trends.baidu.com/worldcup!.>

③<https://www.theguardian.com/science/grrlscientist/2012/nov/08/nate-sliver-predict-us-election.>

铁路系统作为广告发布平台的吸引力。铁路总公司筹建的铁路旅客用户画像系统，使得新 的12306 App 可以根据用户的历史购票、出行数据，推送符合用户口味的广告内容。比 如，商家如果想要定点开拓某个区域市场(某些城市),12306 App 可以根据用户的常住 城市，推送合适的广告。此外，旅客消费能力不同，购票座席的档次就不同，12306 App 就可以根据旅客的消费能力，推送分类广告。对用户进行画像，然后针对性地进行广告推 送，既节省了商家的广告费，又由于用户接收到的是他所需的广告内容，有可能提高广告 的购买转化率，铁路总公司也能从中获得更多的收益。

成立于2009年的传漾科技公司是中国互联网广告领域的技术流代表。传漾科技的主 要业务是依托其网络广告技术/营销平台，为客户量身定制智能数字化营销方案。

网络广告连接着媒体、广告主(商家)、广告公司和用户，传递广告信息，承载塑造 品牌形象，影响日标人群的功能。要准确传达广告信息.就需要了解用户的需求、行为和 习惯，需要对用户进行理解和精细画像。传漾科技公司积累了大量数据，通过数据分析对 网民兴趣进行刻画。他们把网民兴趣划分为33个兴趣大类、168个兴趣中类和857个兴趣 小类，对用户兴趣进行全方位追踪，贴近目标受众。利用大数据分析技术.传漾科技公司 为广告主迅速定位目标客户群体，把广告内容递送到目标客户面前，提高转化率。

**16.10** **自动短文评分**

自动短文评分 (Automated Essay Scoring.AES) 是使用特殊的计算机程序，对学生 撰写的短文自动进行评价，最后给出一个评分。AES 软件系统对短文数据集中的每篇短 文给出一个等级分，比如从1～5的分数。由此可见， AES 可以看作一个分类问题。

近年来，人们对 AES 的兴趣与日俱增，其原因在于：(1)成本。利用人力(专家) 进行短文评分，效率低、成本高。(2)标准化。利用计算机程序进行评分，有利于评分过 程的标准化，能够给短文一个客观的评分，利用人力(专家)进行评分，则容易受到主观 因素的影响。(3)成熟度。机器学习、数据挖掘技术的进步，使得人们可以设计出更加成 熟的AES 系 统 ，AES 系统的评分结果和专家的评分结果差别已经变得很小，于是给教育 机构(包括中小学、大学)更大的信心来使用AES 系统。

AES 的使用也遭到了一些专家的反对，他们认为计算机程序到目前为止仍然不能对 短文给出精确的评分。在一些关键的考试里，比如高考的作文评分，使用 AES 系统的条 件仍然不成熟，些许的不精确给考生的命运造成的影响是巨大的。此外，大量使用AES 系统进行短文评分，将催生应试教育，老师将教授给学生一些技巧，如何在AES 评分的 考试里拿到高分，忽视了创造力、想象力的培养。

即便如此，在人力短缺的场合，AES 仍然不失为一种可行的选择。AES 是一个教育 评估工具 (Educational Assessment), 它是自然语言处理技术的实际应用。AES 需要从短 文中分析出关键的特征和指标，然后馈入评分模型，实现短文的评分。主要的特征包括：

(1)词态错误、语法错误情况；(2)短文的组织和展开；(3)使用的词汇的复杂度；(4)句 子的长短；(5)文章的长短等。针对中文短文的评分，有研究者使用更多的特征，比如使用 修辞格 (Figures-of-Speech) 来衡量文章的质量.包括：(1)排比句的使用情况；(2)比喻句 的使用情况等。

在打分模型方面，可以使用分类和回归技术进行打分，包括决策树 (Decision Tree)、 多元线性回归等技术。

**16.11** **数据产品**

根据我们的理解，数据产品 (Data Product) 是对数据进行加工得到的一种产品形式， 这种产品的最终目的是帮助用户做出更优的决策和行动。为了设计和实现数据产品，首先 要分析和理解用户需求；然后采集和整理必要的数据；接着，建立分析模型，对数据进行 深度分析，生成分析结果，这个结果就是数据产品；利用分析结果，用户可以解决实际业 务问题，对数据产品的不足给出反馈；最后，数据产品生产者对数据产品进行改进。

从以上的论述似乎看不出数据产品和传统的数据分析系统有什么区别。在这里，我们 要指出一个重要的区别，那就是数据产品的生产者并不为单一的用户开发一套系统，而是 把数据和分析整合起来，为大量的用户(他们提出个性化的需求)提供定制化的分析结 果，为此，他们往往需要建设必要的数据中心和大数据管理和分析的基础设施。用户看不 到数据，也不掌握模型，他们付费以后，拿到的是分析的结果。

淘宝的“生意参谋”①就是一款数据产品，它为淘宝店的店主提供大量决策信息。用 户通过付费和定制，获得数据分析服务，他们无须关心数据的收集、存储以及用何种方法 进行分析，他们也无须关心基础设施如何建设，这些都由淘宝提供。

“生意参谋”的主要功能包括：(1)店铺快照。“生意参谋”的首页，通过实时指标、 商品排行、行业排名、店铺经营概况、流量分析、商品分析、交易分析、服务分析、营销 分析、市场行情等信息，给店主提供店铺经营的全方位的核心分析数据，相当于店铺的一 个快照。(2)实时直播。提供店铺实时交易流量数据、实时地域分布、流量来源分布、实 时热门宝贝排行榜、实时催付榜单、实时客户访问等功能，让店主了解当前状况。(3)经 营分析。经营分析包括四个方面的内容。①流量分析，提供全店流量的概况，流量的来源 和去向，访客时段地域等特征分析，店铺装修的趋势和页面点击分布分析结果，帮助店主 对客户进行引流，转化成实际购买。②商品分析，提供店铺所有商品的详细效果数据， 包括商品概况、商品效果、异常商品、分类分析、采购进货，让店主尽快识别问题商品 和有潜力的商品。③交易分析，包括交易概况和交易构成两大功能，通过交易漏斗.从 店铺整体到不同粒度，细分店铺交易情况，让店主及时了解店铺交易情况，了解盈亏状 况。④营销推广，包括营销工具、营销效果两大功能，帮助店主实现精准营销，提升产 品销量。(4)市场行情。是一款行业数据分析工具，提供行业大盘、品牌分析、产品分 析、属性分析、商品店铺榜、买卖家画像等功能，帮助店主把握市场动态，发掘市场 机 会 。

此外，“生意参谋”还通过专题栏目，整合了竞争情报、热词选取、行业排行、单品 分析、销量预测等功能。其中，热词选取功能帮助店主了解来自PC 端和无线移动端对店 铺形成引流的店外搜索关键字，反映买家需求的店内搜索关键字，相关的行业内搜索关键 字，以及这些关键字的搜索热度、引流效果(即通过搜索把用户引导到店铺)。



①<https://sycm.taobao.com/.>



**路透社和彭博社的数据产品①**

以前投资者一般通过报纸、电视等媒体获得各种金融信息，如今他们逐渐向数字化的 金融信息过渡。提供金融信息服务的公司很多，路透社和彭博社凭借其多年在新闻媒体领 域的优势，成为最领先的两家，占据绝大部分市场份额。

彭博社通过 Bloomberg Terminal 提供金融信息服务。Bloomberg Terminal 最初是一 个硬件终端，现在已经演化成为一款软件，方便用户远程访问。用户通过Bloomberg Ter- minal 存取彭博社的各种金融信息服务。Bloomberg Tradebook 是 Bloomberg Terminal的 一个扩展插件，用户可以通过彭博社的消息服务执行各种证券交易。路透社也提供类似于. Bloomberg Terminal 的软件产品，称为 Thomson Reuters Eikon 。Eikon 可以用来接收、 监控、分析各种金融信息。Eikon 还可以根据用户的要求，对某个主题的Tweet 进行持续 监控，得出随时间变化的正面/负面情感指标。利用社交媒体的非结构化数据，分析市场 情绪，是近10年来(自2000年以来)投资者和投资机构借以了解市场趋势的辅助手段。

2013年， Bloomberg Terminal 拥有315000个用户，占据57%的市场份额， Eikon 拥

有190000个用户，占据34%的市场份额。

**16.12** **其他数据科学案例**

数据科学在各行各业都可以找到用武之地。其他数据科学的成功案例还有很多，比 如：(1)迈凯伦一级方程式车队 (McLaren's F1 Racing Team),利用数据分析技术降低 事故，为车队保驾护航。迈凯伦一级方程式车队通过汽车传感器，在赛前的场地测试中采 集实时数据，结合历史数据，通过预测模型的分析，发现赛车潜在的问题，预先对赛车进 行调校. 降低正式比赛中事故的发生率，提高比赛获胜的概率。(2) UPS 快递，利用数 据分析技术，确定最佳行车路线。 UPS 快递公司对其货车安装必要的传感器、 GPS 系统 和无线网络适配器。利用这些传感器，UPS 快递公司采集上千种数据，包括油耗、胎压、 引擎运行状况、车辆的GPS 信息等。利用这些信息优化车队(员工)管理、提高生产力、 降低油耗，每年节省的运营成本达数百万美元。 UPS 为货车定制的最佳行车路线是根据 过去的行车经验总结而来的。2011年，UPS 的驾驶员总共少跑了近4828万公里的路 程。(3)智慧城市。 一个城市的管理和运行离不开管理部门的科学决策。只有数据支撑 才能实现智慧城市的高效运行。通过大数据的采集、整合、共享、分析/挖掘和应用，

**精确了解城市运行情况和发展趋势，为领导决策提供支撑，最终服务于普通市民，提高**

城**市生活的质量。智慧城市包含很多方面，有可能在交通、健康和公共安全等领域率**先 获得突破。

更多的数据科学案例，请参考 Kaggle 网站。②

① <http://www.investopedia.com/articles/investing/052815/financial-news-comparison-bloomberg-vs-reu>- ters.asp.

②<https://www.kaggle.com/wiki/DataScienceUseCases.>

文

**16.13** **思考题**

(1)请举出数据科学在政府服务方面的应用。

(2)请举出数据科学在卫生和医疗方面的应用。

(3)请举出数据科学在教育和培训方面的应用。

(4)请举出数据科学在零售业方面的应用。

(5)请举出数据科学在广告业方面的应用。

(6)请举出数据科学在投资和信用评价方面的应用。

(7)请举出数据科学在互联网舆情监控方面的应用。

(8)请举出数据科学在科学研究方面的应用。