数据世界探秘

第一章　欢迎来到数据科学时代

授课要点

- 一、引言
- 二、数据科学的内涵和发展
- 三、数据科学的学科地位
- 四、数据科学的成熟度曲线
- 五、数据科学的理论体系

一、引言

- 数据蕴含了大量的信息，人们对传统数据的加工、整理和分析，已经能够提炼出许多有价值的信息。

一、引言

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 思政案例1.1：每个数字后面都暗藏线索——来自于上财研究生新生的大数据揭秘

2019年9月5日，上海财经大学研招办微信公众号发布了一篇文章。通过对2649名财大硕士博士新生的分析，试图找出蕴含在上财研究生新生背后的数字线索。

——参考：上海财经大学研招办微信公众号

一、引言

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 上海财经大学

- 教育部直属高校
- 国家“双一流”建设高校
- 国家“211工程”
- “985工程优势学科创新平台”重点建设高校…

全国首批博士
教育部人文社会科学重点研究基地
国家级大学生创新创业训练计划
国家“111计划”
全国首批获得博士学位授予权的高校之一
硕士学位授予单位
卓越法律人才教育培养计划
国家经济学基础人才培养基地
国家建设高水平大学公派研究生项目

学校源于1917年南京高等师范学校开设的商科，是中国人自主创办的第一所研究商学的高等学府

一、引言

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 学科发展稳步发展

- 据ESI（基本科学指标数据库）2019年5月9日公布的数据，上海财经大学经济学与商学（Economics & Business）、社会科学（Social Sciences, general）双双进入ESI全球前1%
- 在教育部学位与研究生教育发展中心全国第四轮学科评估中，应用经济学、工商管理被评为A（全国2%-5%），统计学被评为A-（全国5%-10%），理论经济学、马克思主义理论被评为B+（全国10%-20%）

一、引言

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

- 国际影响力不断增强：“长三角国际论坛”

- 由上海财经大学主办的长三角国际论坛在上海开幕，来自全球各政府机构、知名企业、高校、智库和媒体等机构的嘉宾代表共300余人参加了论坛。

一、引言

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

我们曾在同在一个省份生活

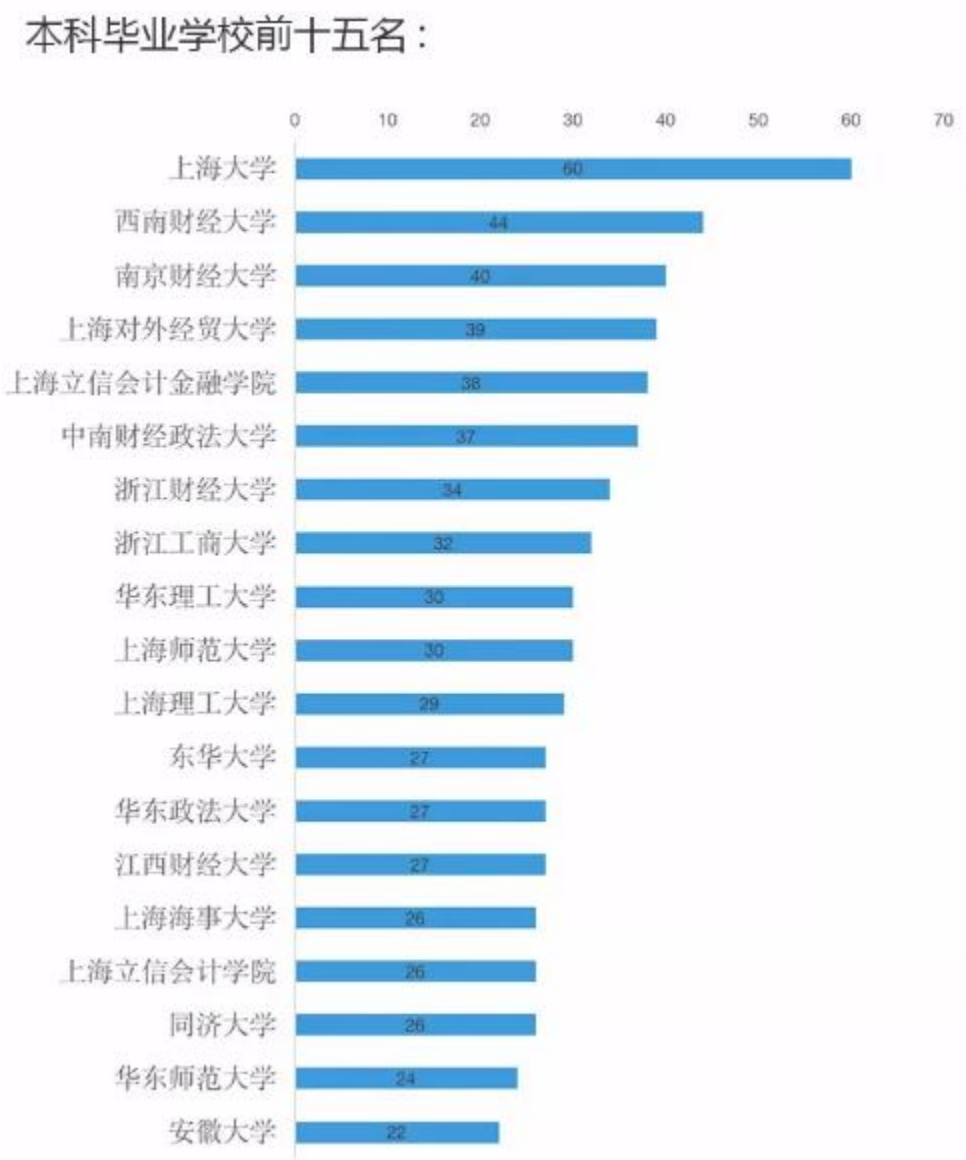
2649名硕士博士新生来自于祖国的32个省（直辖市、自治区、地区）。

一、引言

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

我们曾在同在一个学校学习

除我财以外，本科来自上海大学、西南财经大学、南京财经大学、上海对外经贸大学、上海立信会计金融学院的萌新最多。财经类兄弟院校的才子们对我财青睐有加。



一、引言

甚至曾经在同在一间教室考研

一、引言

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

也许我们来自同一个专业

所有新生都来自经管法文理？不存在的。除此之外，我们还有……

专业背景丰富到你无法想象！

一、引言

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

男生多？女生多？

院所男女比例

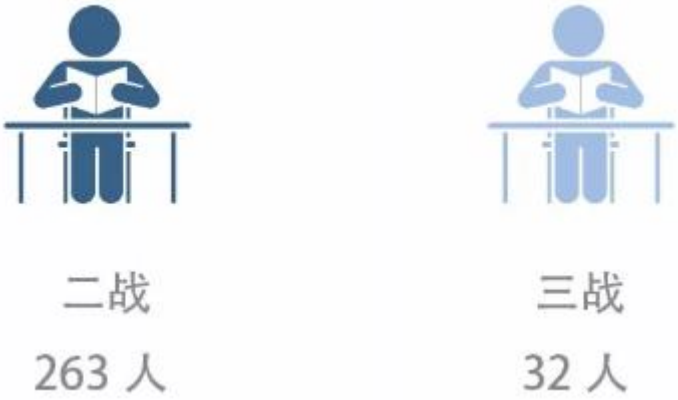
一、引言

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

只要坚持不懈

咬定青山不放松

就能迎来人生的美好春天！



一、引言

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 思政案例1.2：《哪吒之魔童降世》内地票房

8月23日，《速度与激情：特别行动》上映

——参考：猫眼app

一、引言

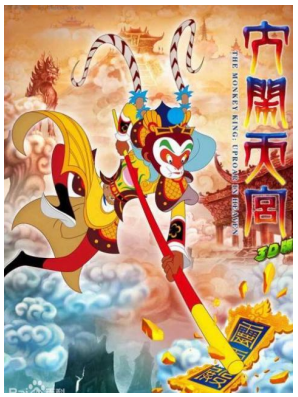
中国动画电影发展史

- 初期～1950年代：1937年，万籁鸣和万古蟾兄弟二人决定绘制动画长片《铁扇公主》，这是亚洲第一部也是当时继美国的《白雪公主》、《小人国》和《木偶奇遇记》之后的第四部动画长片，标志着当时中国的动画艺术已经接近世界先进水平。

一、引言

中国动画电影发展史

- 1950年代～1980年代初：当时中国发展出自有的风格，包括水墨动画、皮影动画、木偶动画等，极具民族特色，此时期的作品一般称为美术片。最具代表性的有《大闹天宫》、《哪吒闹海》、《天书奇谭》，在世界亦获得崇高赞誉和众多奖项。此时期的动画画风独特，想象奇瑰，不乏精品，是中国动画最为辉煌的时期。



一、引言

中国动画电影发展史

- 1980年代～1990年代：改革开放后，中国虽然也产出《葫芦兄弟》、《黑猫警长》、《阿凡提的故事》等给人留下深刻印象的作品
- 1990年代～2000年代：其间推出的全年龄动画如《宝莲灯》、《风云决》等不受关注，相比而言幼儿动画《喜羊羊与灰太狼》则很吃香

一、引言

- 从2015年至2019年，基于优秀传统文化题材的动画作品显著增多，优秀传统文化成为连接影片与消费者的有力纽带；国产动画电影作品产量不断攀升，成为动画生产大国。与此同时，国际合作愈加频繁；动画基地布局总体完成；创意、研发、衍生品授权等产业链日趋完善；人才培养体系初步建立，全国多所高校开办动漫专业，就业人数屡创新高；动画公司和企业大量涌现，产业辐射人群日趋广泛，产业产值和社会影响不断增大，目前已经达到世界领先水平。

一、引言

- 随着信息技术的快速发展，数据搜集和数据存储技术亦快速进步，使得各组织机构可以积累海量数据。特别是大数据的兴起，使得数据量越来越大、数据越来越多样化。大数据对各学科领域的传统知识提出了挑战，传统知识已经无法解释和有效利用新兴的大数据，进而促使传统理论与方法的革命性变化，从而迈入了数据科学时代。

一、引言

- 数据科学时代改变了人们对数据价值的认识，不再认为数据是无价值的，也不再简单认为数据是死的、被动的东西，而是更加重视数据的积极作用。大数据在各个领域得到了充分的应用。

一、引言

- “数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。”——麦肯锡

一、引言

- 案例1.3：Facebook的数据泄露与政治斗争

一、引言

2018年3月17日当地时间，美国《纽约时报》和英国《观察者报》（英国《卫报》的周日版）共同发布了深度报道。声称英国一家基于数据分析的政治咨询公司剑桥分析（Cambridge Analytica）被控利用Facebook的信息管理不力，窃取了高达5000万名Facebook用户的个人资料，在2016年美国大选期间帮助共和党候选人、现任总统特朗普投放针对性的政治广告，可能影响到大选结果。这篇报道在世界范围内引发了轩然大波。

一、引言

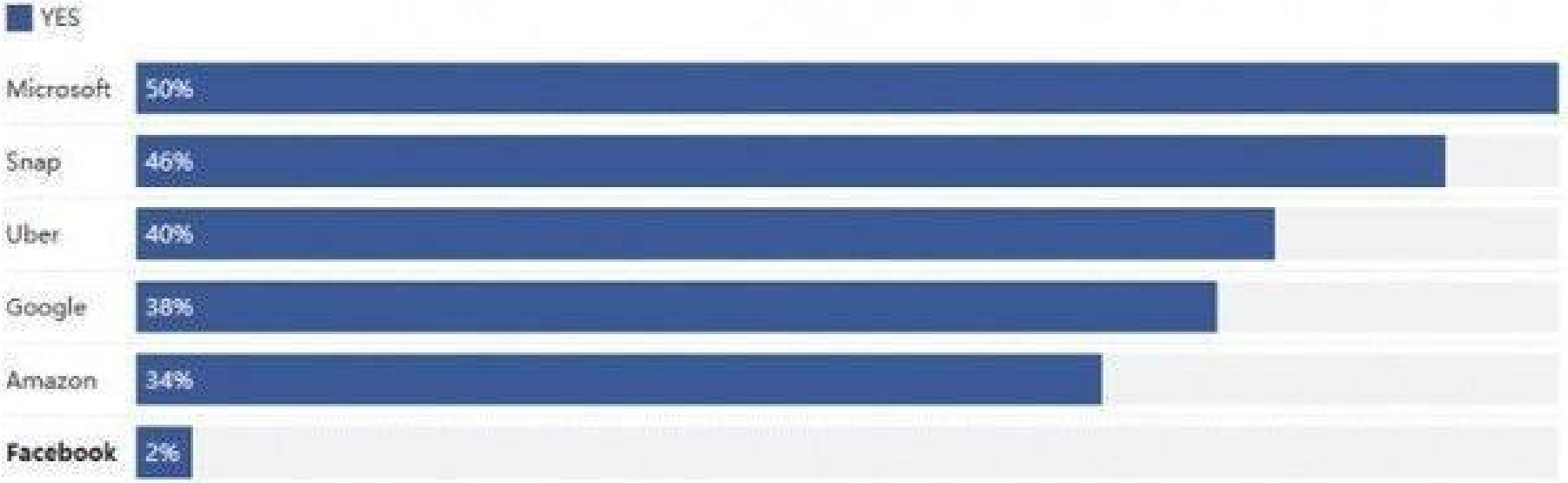
事件曝光之后，Facebook迅速封杀了剑桥分析。但这一事件在美国主流媒体产生了巨大反响，各大媒体出现了大量关于Facebook的谴责负面报道。#删除Facebook账号#迅速成为了Twitter的热门话题，甚至成为了一些科技媒体的大标题。连WhatsApp联合创始人布莱恩·阿克顿(Brian Acton)也不失时机地呼吁删除Facebook账号。(在把WhatsApp作价190亿美元出售给Facebook三年后，他已经离开Facebook。)

一、引言

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

#DeleteFacebook? Here's How Tech Workers Answered

We surveyed over 2,600 tech workers, asking if they will delete Facebook after the data mishap with Cambridge Analytica. The survey ran from March 20, 2018 through March 24, 2018. Overall, 31% answered 'YES' and 69% answered 'NO.' Below, are the results from employees from the top 5 tech companies (as calculated by highest volume of responses) as well as the responses from Facebook employees.



Blind is an anonymous work talk app for tech employees. Blind users were surveyed and asked if they planned to delete Facebook by answering 'YES' or 'NO.' The chart above includes data from employees at tech companies with the most responses to our survey. To be included in the "Top 5", each company had to have a minimum of 50 responses.

Source: Blind

一、引言

受这一负面事件影响，Facebook股价从此前的185美元高点下滑到168美元左右，市值直接蒸发了近10%，相当于500多亿美元。这是Facebook自2012年以来的股价最低迷表现，甚至引发了投资者的诉讼。

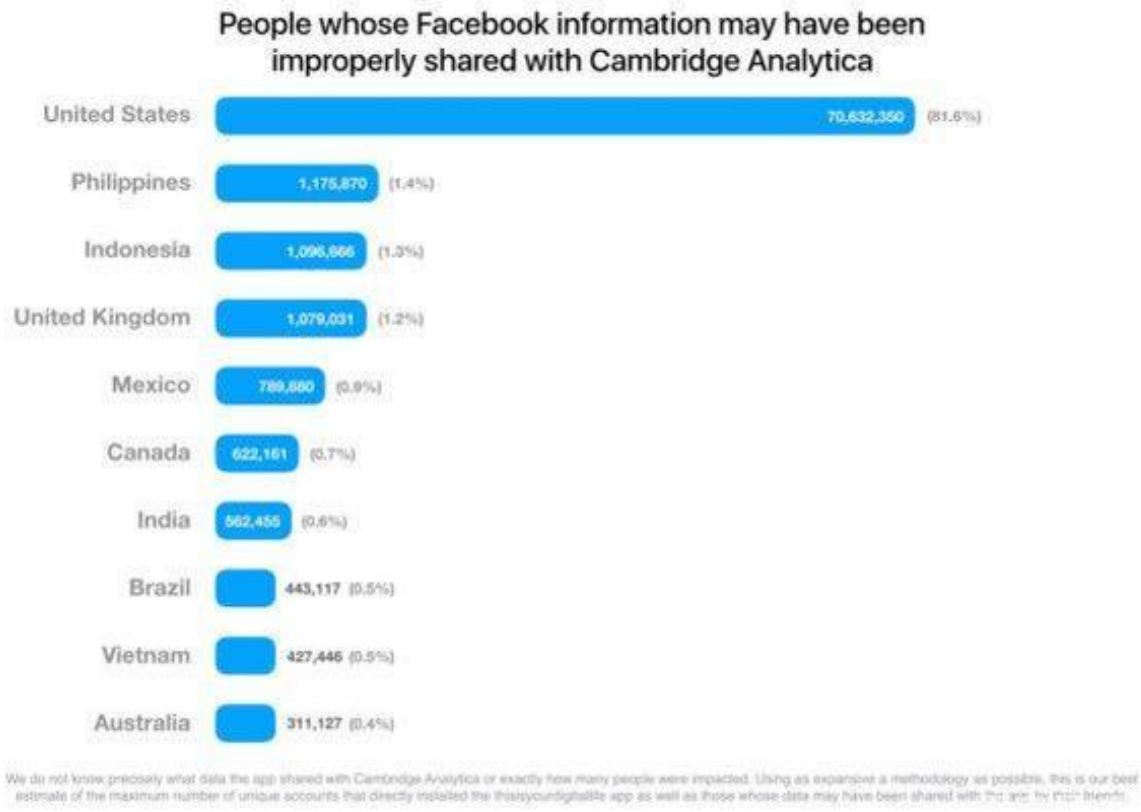
围绕数据的罗生门

此次Facebook数据泄露事件的来龙去脉：2013年，英国剑桥大学心理学研究员阿列克桑德·考根(Aleksandr Kogan)在Facebook上创建了一个应用“这是你的数据生活”(This is your digital life)，通过问答来预测用户的性格喜好，用于他的学术研究。这类似于国内长久不衰的性格测试，无论是东西方网民对这类趣味测试总是缺乏抵抗力。

一、引言

共有近30万名Facebook用户下载了这一应用，通过自己的Facebook账号登录并进行了测试。考根也因此通过这一应用获得了30万名用户的居住地和测试内容等信息，并接入了他们Facebook好友的信息流。高达5000万的涉及用户数目就是《纽约时报》根据这30万人的好友信息流网络所估算得出的。考根通过应用获取用户数据并没有违反Facebook的数据政策，用户在点击测试的时候也同意交出自己的个人资料。

一、引言



2018年4月5日，Facebook表示涉及人数可能多达 8700万人，其中美国人的比例大概占到81.6%，也就是7060万人左右。

一、引言

2015年，Facebook从媒体报道中得知，考根违反了信息管理政策，私下把这些数据交给了一家叫做战略传播实验室(SCL)的公司。SCL旗下还有一家政治数据分析公司，就是此次事件的核心——剑桥分析。Facebook明确规定，第三方应用开发者不得出售或转让他们从Facebook获得的用户数据。Facebook当时采取了一些应对措施，撤下了考根那款性格测试应用，并要求考根销毁他所收集的用户数据。

2016年8月，Facebook又派律师要求剑桥分析立即删除未经授权获取的用户数据，并得到了后者肯定的答复。但Facebook不是执法部门，无法核实后者是否真的删除了数据。

一、引言

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

2016年7月，共和党总统候选人特朗普的竞选团队聘请了剑桥分析和美国数据营销专家帕斯卡尔(Parscale)进行数字广告投放。而Facebook正是特朗普团队进行政治营销的重中之重。

具体来说，剑桥分析负责确定潜在的受众目标，而帕斯卡尔设计相对应的广告。根据这些数据中用户的个人喜好，判断出哪些人可能会投票给特朗普，再向他们投放广告，促使这些选民在大选期间把票投给特朗普。此外，特朗普团队还会向希拉里的潜在选民投放广告，劝说他们不要投票。

特朗普在2016年美国总统大选中获胜。而Facebook两年后成为了美国民主党和主流媒体声讨“全民公敌”特朗普的最新靶子，也成为了希拉里阵营大选失败的又一个替罪羊。

——参考：《Facebook的罪与罚：隐私痼疾与政治斗争》，网址链接：

https://baijiahao.baidu.com/s?id=1595594155886422440&wfr=spider&for=pc

一、引言

➢ 我国在个人信息保护、数据安全方面的相关措施

- 中央网信办、工信部、公安部、市场监管总局四部门联合发布《关于开展APP违法违规收集使用个人信息专项治理的公告》，开展APP违法违规收集使用个人信息专项治理
- 十三届全国人大二次会议将《个人信息保护法》列入本届立法规划，相关部门正在抓紧研究和起草，争取早日出台
- 安部网络安全保卫局、北京网络行业协会、公安部第三研究所等单位共同研究制定的《互联网个人信息安全保护指南》正式发布，旨在全面贯彻落实《网络安全法》，有效指导个人信息持有者健全公民个人信息安全

一、引言

➢ 我国在个人信息保护、数据安全方面的相关措施

- 工信部印发《电信和互联网行业提升网络数据安全保护能力专项行动方案》，提出2019年10月底前，完成全部基础电信企业、50家重点互联网企业以及200款主流APP数据安全检查；2020年7月底前，进一步完善网络数据安全制度标准体系，形成行业网络数据保护目录，制定15项以上行业网络数据安全标准规范，贯标试点企业不少于20家等要求
- 《个人金融信息（数据）保护试行办法》（初稿）出炉，央行已下发到各家银行征求意见…

一、引言

- 案例1.4：2012年美国总统大选

一、引言

数据科学的理念在2012年奥巴马成功竞选美国总统中得到直接应用并发挥了重要作用。

在竞选活动的初期，奥巴马竞选团队的主管吉姆·梅斯纳(Jim Messina)就曾宣布，即将打造一个以数据为驱动力的竞选活动——“政治是最终目标，但政治嗅觉已不再是总统候选人取胜的唯一方法。我们会在此次竞选活动中对每个事件进行数据分析。”

以数据为中心的决策方式成就了奥巴马在第二任期连任美国总统。“以往依赖于预感和经验的华盛顿特区竞选专家们的受欢迎程度正在迅速下降，而他们的地位则将由善于收集数据并加以分析的程序员所取代。”

一、引言

竞选团队构建

奥巴马竞选团队由团队主管、数据科学家、数据分析团队和团队发言人等构成。

聘请著名政治顾问吉姆·梅斯纳为竞选团队主管。

在竞选团队中设置了首席科学家。吉姆·梅斯纳邀请数据挖掘领域拥有丰富经验的雷伊德·加尼（Rayid Ghani）出任芝加哥竞选团队总部的首席科学家。加尼具有丰富的大数据处理经验，曾经基于数据分析成功地提出过超市销售效率达到最大化的方法。

聘请了一大批数据分析员，人员规模甚至达到了2008年竞选时数据分析部门的5倍。

聘请本·拉-波尔特（Ben LaBolt）等作为团队发言人，负责对外数据发布与交流。

一、引言

数据分析——明确关键环节

竞选活动的关键绩效指标是准备投票给奥巴马的选民数除以准备投票的总人数得到的比值。有三个关键环节可以最大限度地提高这个数值：登记、说服和投票。竞选团队必须鼓励他们的目标人群进行选民登记；说服犹豫不决的选民投票给奥巴马；然后尽最大的努力让奥巴马的选民在选举日投票。

为了应对挑战，数据分析团队被分成了几个不同的小组。现场工作组负责组织志愿者、办理登记、鼓励选民投票等。数字组负责在线宣传、邮件广告、网上募捐以及社交媒体等。通信和媒体组负责发布奥巴马的个人信息、安排采访、进行广告投放等。财务组则负责整体的竞选筹款。而且，各个小组采用了统一建设和集中管理，而不是分别进行各自的数据分析。

一、引言

奥巴马的竞选团队广泛采用在Facebook收集的用户数据来投放广告。竞选团队的信息收集手段公开直接。支持奥巴马的选民在竞选网站上用Facebook账号登录之后，奥巴马的竞选团队就自动获得了他们的用户数据。然后奥巴马竞选团队在这些用户(本就是他的支持者)的同意下，获得他们Facebook好友的数据。再接下来，数据分析团队用这些用户数据和他们所在地的实际选民数据进行综合对比梳理，则生成了政治广告所需的详尽用户资料——不仅知道用户的身份，还能知道如何引导用户。

一、引言

奥巴马的竞选团队还首次利用Facebook等社交网络进行大规模的游说，就像以前挨家挨户敲门拉票的方式一样。在竞选的最后几周，下载特定应用软件的用户收到了包括他们在“摇摆州”好友照片在内的多条消息，应用软件鼓励用户通过点击按钮来呼吁“摇摆州”选民采取行动：鼓励选民投票注册、更早地进行投票并积极参与到民意调查工作。竞选团队发现，大约有五分之一收到此信息的选民做出了回应，主要原因在于请求来自于自己熟悉的朋友。

一、引言

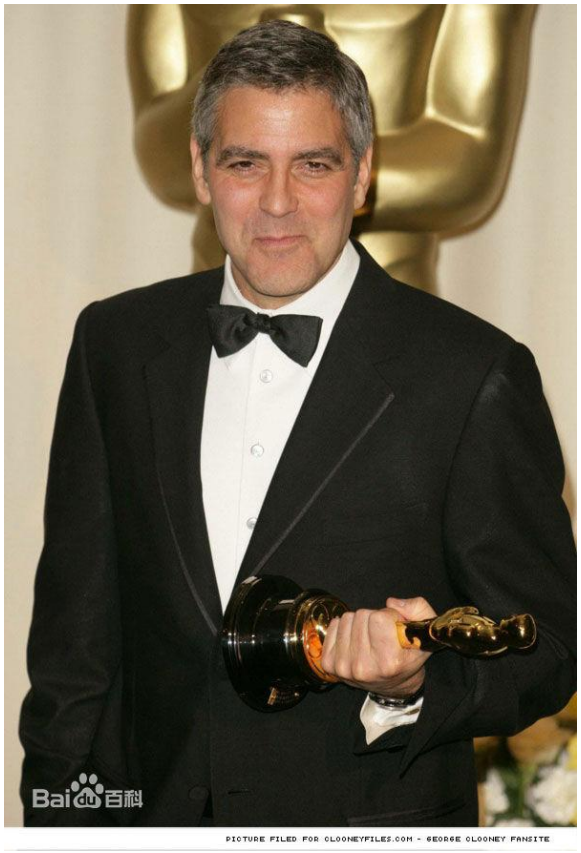
数据还帮助奥巴马竞选团队更好地做出了广告购买的决策。例如，在电视剧《混乱之子》、《行尸走肉》、《23号公寓的坏女孩》中就出现了奥巴马的竞选广告。而此前竞选广告通常只会出现在本地新闻节目中。奥巴马的芝加哥竞选总部曾表示：“我们在电视上的广告购买效率提升了14%，因此，我们能够确保与摇摆州选民产生交流。”

此外，奥巴马竞选团队在大选最后阶段还采取了不同于以往的战略。2012年8月，奥巴马决定在知名社交新闻网站Reddit上回答问题。原因在于“根据数据显示，我们的很大一部分目标选民就在使用Reddit。”

一、引言

数据洞见——乔治·克鲁尼效应

2012年年初，奥巴马竞选团队幕后的数据分析团队注意到了一个现象——著名演员兼导演乔治·克鲁尼(George Clooney)对美国西海岸40～49岁女性具有非常大的吸引力，她们甚至愿意不远万里为克鲁尼和奥巴马共进晚餐而慷慨解囊。



一、引言

因此，竞选团队创造性地提出了一个新的想法——在美国东海岸找到一位具备相同号召力的名人，从而复制“克鲁尼效应”，为奥巴马筹集竞选资金。他们最终选择了主演电视剧《欲望都市》的著名演员莎拉·杰西卡·帕克（Sarah Jessica Parker）。数据分析团队深入研究了帕克粉丝群体，并洞见了他们的主要特征——喜欢竞赛、小型聚会和名人。因此，竞选团队组织了一场与奥巴马在帕克位于纽约的豪宅共进晚餐的“竞争”。

一、引言

媒体评价

当奥巴马通过社交媒体收集用户数据进行精准投放的时候，美国主流媒体一片交口称赞，欢呼技术手段帮助奥巴马获胜。知名杂志《麻省理工科技评论》(MIT Tech Review) 在2012年专门撰文《奥巴马是如何通过大数据来团结选民的》。文中谈到“奥巴马团队甚至知道给他投票的6945万名选民中每一个人的名字……奥巴马竞选团队的分析师可以看到每一个选区的民主党人投票总数，确认最有可能投票给他的民众。”

Intelligent Machines

How Obama's Team Used Big Data to Rally Voters

How President Obama's campaign used big data to rally individual voters.

by Sasha Issenberg December 19, 2012

The Obama 2012 campaign used data analytics and the experimental method to assemble a winning coalition vote by vote. In doing so, it overturned the long dominance of TV advertising in U.S. politics and created something new in the world: a national campaign run like a local ward election, where the interests of individual voters were known and addressed.
This story was originally posted in three installments.

美国知名左翼媒体《纽约时报》当时也同样撰文，褒扬奥巴马的数字营销策略非常成功。

一、引言

——参考：《Facebook的罪与罚：隐私痼疾与政治斗争》，网址链接：
https://baijiahao.baidu.com/s?id=1595594155886422440&wfr=spider&for=pc

——参考：朝乐门，《数据科学理论与实践》，清华大学出版社

一、引言

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

- 案例1.5：今日头条的巨大成功

一、引言

随着资讯市场的成熟和发展，人们需要一个在碎片时间消费有趣资讯的产品，来解决用户的需求。这里的有趣因人而异，就需要用个性化的推荐技术去满足。而依靠个性化推荐技术获得成功的典型案例就是——今日头条。

今日头条是北京字节跳动科技有限公司开发的一款基于数据挖掘的推荐引擎产品，提供连接人与信息的服务。

今日头条由张一鸣于2012年3月创建，2012年8月发布第一个版本。短短几年，今日头条就获得了巨大的成功。目前，每天活跃的用户总数超过2亿。

一、引言

产品特色

今日头条基于个性化推荐引擎技术，根据用户的个人特征（兴趣）、环境特征（位置）、文章特征三者的匹配程度进行个性化推荐。目前，推荐内容不仅包括狭义上的新闻，还包括音乐、电影、游戏、购物等资讯。

根据用户社交行为、阅读行为、地理位置、职业、年龄等挖掘出兴趣。可实现实时推荐。0.1秒内计算推荐结果。3秒完成文章提取、挖掘、分类。通过社交行为分析，5秒计算出新用户兴趣分配。通过用户行为分析，用户每次动作后，10秒内更新用户模型。

根据用户所在城市，自动识别本地新闻，精准推荐给当地居民。可根据用户年龄、性别、职业等特征，自动计算并推荐其感兴趣的资讯。

一、引言

个性化推荐方案

为用户找到有趣的资讯有两条路可以走：人工运营和算法推荐。在类头条产品出现之前，请新闻方面专业人才来运营是最稳妥的方式。但人工运营成本越来越高，局限性越来越明显。走算法推荐的路是一条必由之路。下表简要对比两者的差别。

对比项	人工运营	算法推荐
风险把控能力	强	弱
投入产出比	有瓶颈	有空间
覆盖度（用户、内容）	部分人群、部分内容	无限大
个性化程度	一般	精细

一、引言

推荐算法应用在资讯类产品存在一些挑战，这也是资讯推荐能否做好的关键所在。

- 可扩展性

推荐本质是建立用户（user）和项目（item）的关联，一般问题要么是user侧量级大，要么是item侧量级大，而资讯推荐是典型的“双大”场景。又由于是高度依赖个性化的场景，还不能简单地将某一侧大幅降维，所以可扩展性显得尤为重要。

- 稀疏性

资讯的高度个性化自然而然得带来一个很棘手的问题就是稀疏性。举个最简单的例子，如果将user和item的点击行为用矩阵形式表示出来，会发现比一般问题更多的0项存在。而稀疏问题是一直困扰机器学习高效建模的一大难题。

一、引言

- 冷启动

每天都有大量的新闻产生，如何将如此多的新闻快速、合理地冷启动，尽快将高质量的新闻推给合适的用户是个大问题。

- 时效性

不同于商品、书籍、电影、视频等的推荐，新闻的鲜明特点是生命周期非常短，有的甚至只有几个小时。如何在最短的时间里把新闻推送给感兴趣的人，在新闻进入“暮年”之前发挥它的最大价值是个非常重要的问题。

一、引言

- 质量保证

新闻本身量大，且时效性强，如何在短时间里快速评估每篇稿子的质量和合法性，做到最高效、最精准的内容审核是个大课题。

- 动态性

这里的动态性主要体现为用户兴趣随时间改变、当前热点随时间改变。用户在一天里的不同时刻、不同地点、不同上下文里的阅读兴趣都有所差别，动态变化。

一、引言

个性化推荐算法

围绕上面这几个挑战，业界各大资讯类产品在做推荐时想出了各种招儿来解决，下面介绍Google News、今日头条等产品的推荐算法的基本思想。

- Google News

2007年，Google News首次发表论文《Google News Personalization: Scalable Online Collaborative Filtering》公开资讯推荐技术，采用CF（协同过滤）技术。无论是基于用户(user-based)还是基于项目(item-based)的CF技术，其计算量直接取决于特征维数和用户项目对的数目，而资讯类产品这两个数目都非常大，导致计算量巨大。

一、引言

Google这篇论文的核心就是将CF改造为支持大规模计算的方法。

其原理也很简单：根据用户对页面的点击历史将用户事先分成群，再做user-based CF时实际变成了user cluster-based CF。这样大大简化了计算。同时，线上只需要记录每群用户喜欢什么。一个用户来了之后，先找到其对应的群，再推荐这个群喜欢的资讯就好。而线下则借助MapReduce（面向大数据并行处理的计算模型、框架和平台）实现了聚类分群算法，定时把最新分群结果推到线上。

一、引言

user cluster-based CF的算法也有一些明显的缺点：

（1）它不能解决新用户、新资讯的冷启动，因为没有行为数据来支撑CF运转。

（2）推荐精度不够高，没有做到真正的个性化。这是cluster-based CF算法本身的特点决定的。

（3）实时性不够。用户聚类不能做到快速更新，这导致了对用户最新兴趣把握有不及时的风险。

一、引言

Google News在2010年推出了另一篇文章《Personalized News Recommendation Based on Click Behavior》。这篇文章探讨了基于贝叶斯理论进行建模，重点解决推荐精准性和新资讯的冷启动问题。

它针对CF遗留的问题进行了很好的解决：（1）引入新闻类别解决了新新闻的冷启动；（2）引入用户兴趣解决了个性化和推荐精确度的问题。但新用户冷启动还有优化的空间，因为按照这个方法，同一地区不同新用户推荐的都是该地区最热门的内容。

一、引言

- 今日头条

作为国内个性化推荐产品的代表之一，今日头条技术经历了三个阶段：

（1）早期以非个性化推荐为主，重点解决热文推荐和新文推荐，这个阶段对于用户和新闻的刻画粒度也比较粗，并没有大规模运用推荐算法。

一、引言

（2）中期以个性化推荐算法为主，主要基于协同过滤和内容推荐两种方式。协同过滤技术和前面介绍的大同小异，不再赘述。基于内容推荐的方式，则借助传统的NLP（自然语言处理）、word2vec（用来产生词向量的相关模型）和LDA（文档主题生成模型）对新闻有了更多的刻画，然后利用用户的正反馈（如点击，阅读时长、分享、收藏、评论等）和负反馈（如不感兴趣等）建立用户和新闻标签之间的联系，从而进行统计建模。

（3）当前以大规模实时机器学习算法为主，用到的特征达千亿级别，能做到分钟级乃至秒级更新模型。

——参考：《今日头条成功的核心技术秘诀是什么？深度解密个性化资讯推荐技术》，网址链接：https://www.leiphone.com/news/201707/YYoc8r9MWsBy2QAC.html

——参考：《今日头条公布算法原理，承认机器理解有限需引入更多人工》，网址链接：https://baijiahao.baidu.com/s?id=1589753420651219200&wfr=spider&for=pc

一、引言

➢ 案例1.6：对于谷歌流感趋势分析的困惑和思考

未卜先知

2009年，Google公司的谷歌流感趋势（Google Flu Trends，GFT）团队在《自然》(Nature)杂志发表标题为《基于搜索引擎查询数据的流感疫情监测》(Detecting in Fluenza Epidemics Using Search Engine Query Data）的文章，介绍了GFT项目。

当时，传统的流感监测体系，包括美国疾病控制与预防中心(Centers for Disease Control and Prevention，CDC)依赖于包含疑似流感的病例占比、住院人数、死亡人数等数据，对流感进行监测，这些数据通常滞后一两周发布，无法实时预测美国流感的爆发情况。

一、引言

GFT团队借鉴了前人利用与流感相关的间接信息对流感进行快速监控的思想。前人研究中利用的间接信息包括：电话分诊咨询热线的流感相关电话量；治疗流感药物的销售量；与流感相关的检索词的在线搜索量；健康医疗网站的访问量等。

GFT团队则利用了与流感相关的检索词的搜索量。 GFT团队认为，人们输入的检索词代表了他们的即时需要，反映了用户情况。团队利用2003年至2008年的数据进行建模，其中，使用2003至2007年的数据作为训练集构造模型，该模型在2007年至2008年的测试数据集中表现良好。GFT团队从5000万个关键词中选择了“一揽子”共45个检索词，包括流感并发症、感冒/流感治疗、流感症状、抗生素用药等作为预测变量，模型表示流感关键词的搜索量能较好地预测CDC公布的流感发病率（疑似流感病例占比），两者具有显著的相关性。GFT团队的研究成果发布之后引起了广泛的重视。

一、引言

在训练集中，模型结果与CDC报告的数据相关系数为0.85。

在测试集中，模型结果与CDC报告的数据相关系数为0.96。

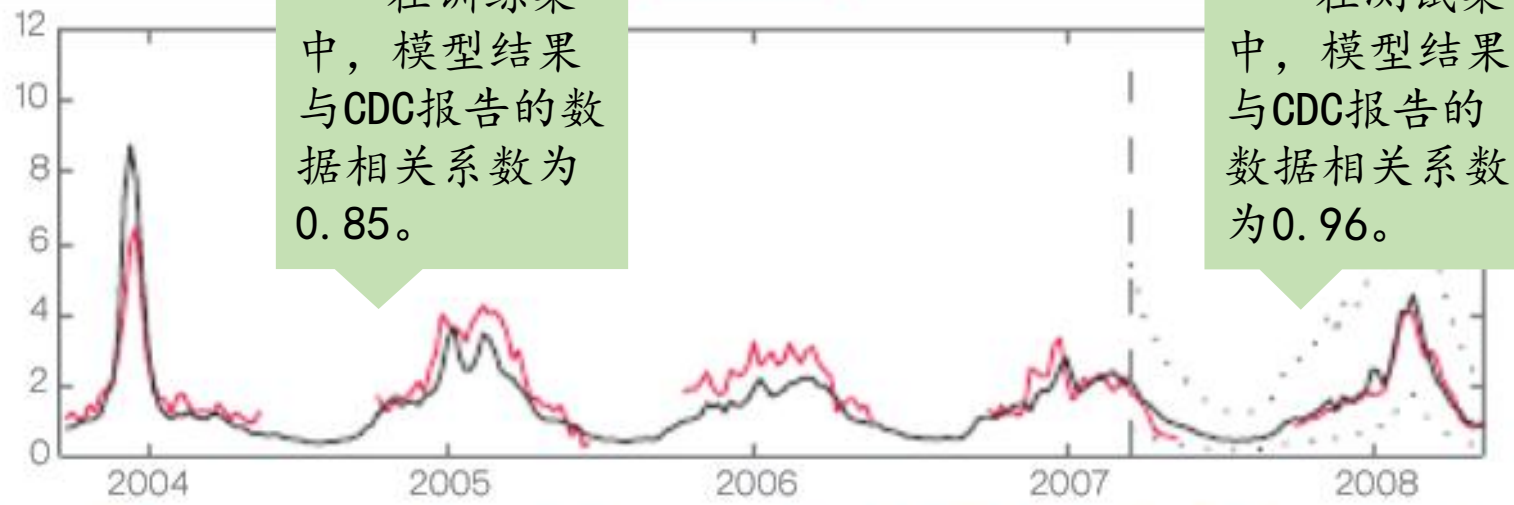
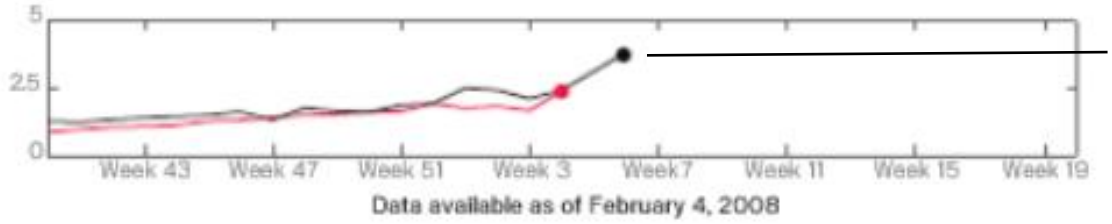
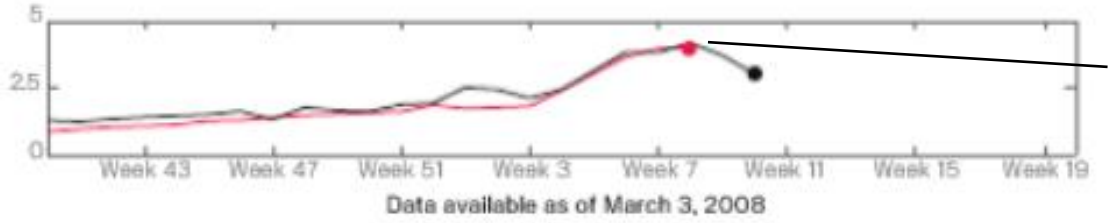


Figure A comparison of model estimates for the Mid-Atlantic Region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, while a correlation of 0.96 was obtained over 42 validation points. 95% prediction intervals are indicated.

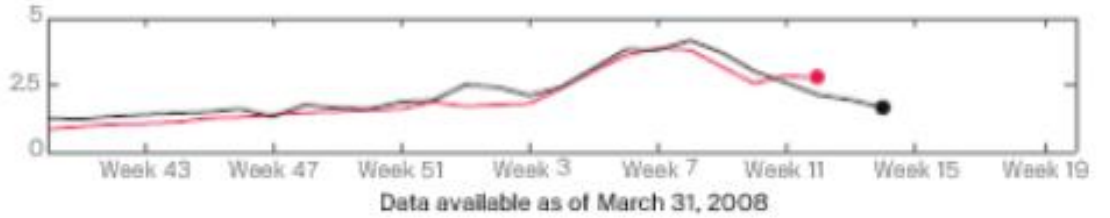
一、引言



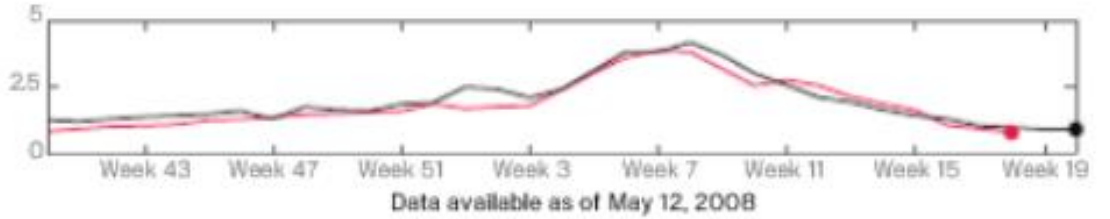
在第5周，模型结果显示疑似流感病例占比将快速上升



在第8周，模型结果显示疑似流感病例占比将达到阶段性高点，并于第9周和第10周下降



这些结论都被后续CDC公布的数据证实

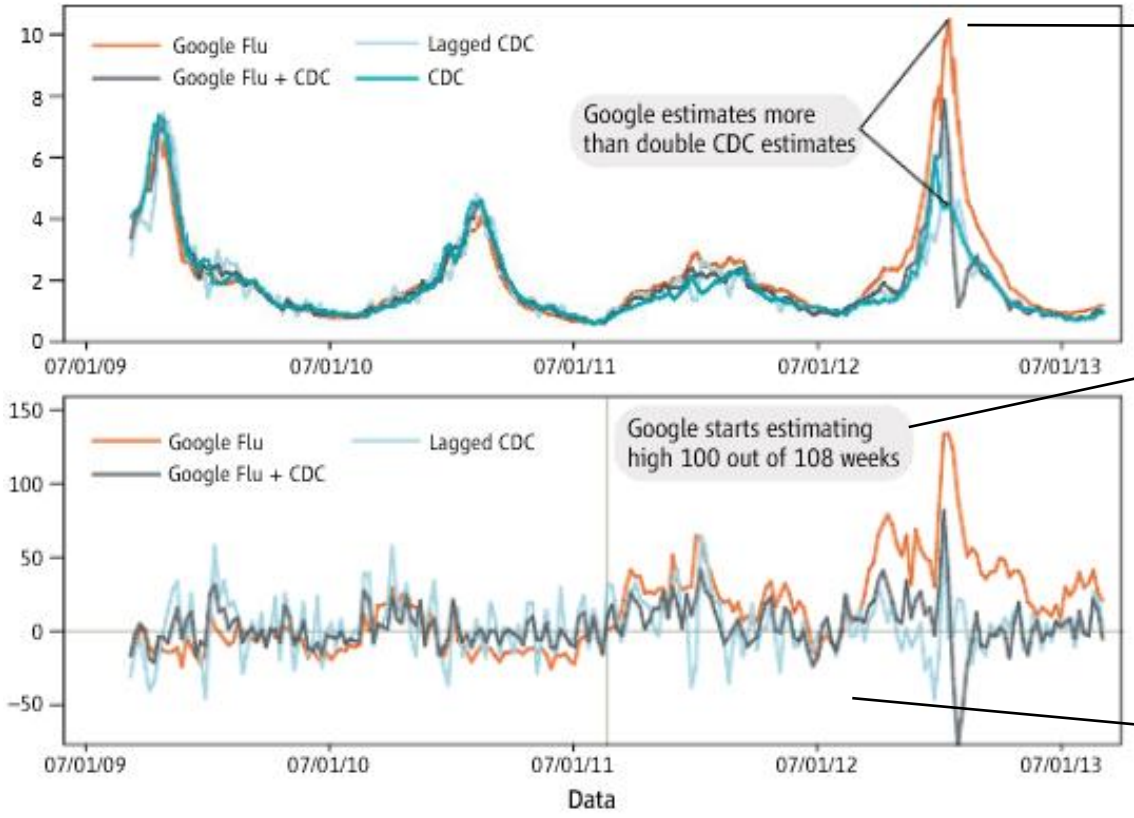


一、引言

神话破灭

然而，GFT的预测也并不总是正确。2009年，GFT没能准确预测到非季节性流感A-H1N1。从2011年8月到2013年8月的108周里，GFT有100周高估了CDC报告的流感发病率。高估达到怎样的程度呢？其中在2011年8月至2012年，GFT预测发病率与CDC报告值的误差最高达50%；而从2012年至2013年8月，GFT流感发病率与CDC报告值的误差最高达100%。2014年，David Lazer等学者在《自然》上发文《谷歌流感趋势的启示：大数据分析中的陷阱》（The Parable of Google Flu: Traps in Big Data Analysis），阐述了这一现象。同时发现，只用滞后2期的CDC的数据进行建模，预测效果比GFT都好得多。目前，GFT已经不再更新了。

一、引言



2013年2月，GFT预测结果是CDC流感发病率的2倍。

2011年8月至2013年8月的108周中，GFT预测结果超过CDC流感发病率（误差大于0）的有100周。

滞后CDC建模的预测效果好于GFT预测效果。

一、引言

对大数据的思考

David Lazer等学者认为GFT预测不准确的主要原因有两点：

- 大数据浮夸（Big Data Hubris）

“大数据浮夸”通常基于以下假定：大数据是传统数据收集和分析方法的替代品，而不是传统方法的补充。但是，数据量并不意味着可以忽略数据的基本问题，大部分大数据并不是那些可生成科学分析的有效可靠数据设备产生的。

在GFT案例中，GFT团队从5000万个搜索词中寻找可以拟合训练集数据的最佳匹配，但是，与流感趋势匹配的检索词与流感趋势的结构并不一致，很难用来预测流感发展趋势。

一、引言

- 算法演化(Algorithm Dynamics)

算法演化对大数据在实证运用中产生的影响更为深远。现实中大数据往往是公司或者企业进行主要经营活动之后被动出现的产物。以谷歌公司为例，其商业模式的主要目标是更快速地为使用者提供准确信息。为了实现这一目标，数据科学家与工程师不断更新谷歌搜索的算法，让使用者可以通过后续谷歌推荐的相关词快捷地获得有用信息。例如，2011年，谷歌对搜索结果进行了修改，能够为用户提供系统建议之外的其他搜索词。2012年2月，谷歌宣布，当用户搜索发烧、咳嗽等关键词时，谷歌会返回一些可能的诊断方法。对搜索算法的修改是为了满足谷歌的商业模式，但是从数据生成机制方面来看，却会出现使用者搜索的关键词并非出于使用者本意的现象。

一、引言

这就产生了两个问题：第一，由于算法规则在不断变化而研究人员对此并不知情，今天的数据和明天的数据容易不具备可比性。第二，数据收集过程的性质发生了变化。大数据不再只是被动记录使用者的决策，而是通过算法演化，积极参与到使用者的行为决策中。

在GFT案例中，2009年以后，算法演化导致搜索数据前后不可比，特别是“搜索者键入的关键词完全都是自发决定”这一假定在后期不再成立。这样，用2009年以后的模型就不可避免产生了预测效果不佳的过拟合问题。

一、引言

——参考：《从谷歌流感趋势谈大数据分析的光荣与陷阱》，网址链接：https://blog.csdn.net/u010999396/article/details/62070968

——参考：《谷歌流感趋势的启示：大数据分析中的陷阱》，网址链接：https://wenku.baidu.com/view/58451a86700abb68a982fbc4.html

——参考：《Detecting in Fluenza Epidemics Using Search Engine Query Data》，网址链接：https://wenku.baidu.com/view/5736e33531126edb6f1a1040.html

二、数据科学的内涵和发展

1. 数据科学的内涵：

数据科学（Data Science）可以理解为大数据时代的一门新科学，即以揭示数据时代，尤其是大数据时代新的挑战、机会、思维和模式为研究目的，由大数据时代新出现的理论、方法、模型、技术、平台、工具、应用和最佳实践组成的一整套知识体系。

二、数据科学的内涵和发展

2. 数据科学的发展：

（1）萌芽期（1974年—2009年）

- 1974年：著名计算机科学家、图灵奖获得者Peter Naur的专著《计算方法的简要综述》(Concise Survey of Computer Methods)中首次明确提出“数据科学 (Data Science)”的概念，术语“数据科学”首次出现在学术专著中。
- 2001年：当时在贝尔实验室工作的William S. Cleverland在期刊《国际统计评论》(International Statistical Review)发表了题为《数据科学——拓展统计技术的行动计划》(Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics)的论文，术语“数据科学”首次出现在学术论文的标题中，并被权威论文专题讨论。

二、数据科学的内涵和发展

- 2003年：国际科学理事会（the International Council for Science, ICSU）的CODATA（the Committee on Data for Science and Technology）发行了第一本以“数据科学”命名的学术期刊——《数据科学学报》(The Data Science Journal)。
- 2009年：Troy Sadkowsky 等在LinkedIn 上组建了第一个数据科学家群——The Data Scientists Group。

二、数据科学的内涵和发展

（2）快速发展期（2010年—2013年）

- 2010年：Drew Conway 提出了第一个解释数据科学理论基础的维恩图——数据科学维恩图（The Data Science Venn Diagram）。
- 2011年：D J Patil 出版了专著《如何组建数据科学团队》（Building Data Science Teams），系统讨论了数据科学家的能力及如何组建数据科学家团队问题。
- 2012年：数据科学中的相关思想成功地应用于奥巴马竞选团队的总统竞选工作，受到社会各界的广泛关注。Davenport T H 和 D J Patil 在《哈佛商业评论》（Harvard Business Review）上发表了题为《数据科学家——21世纪最性感的职业》（Data Scientist: The Sexiest Job of the 21st Century）的论文。Schutt R. 在哥伦比亚大学开设第一门数据科学课程《数据科学导论》（Introduction to Data Science）。

二、数据科学的内涵和发展

- 2013年：涌现了一批关于数据科学的代表性论文和专著。

✓ 代表性论文有：Mattmann C A在《自然》(Nature) 杂志上发表题为《计算——数据科学的一种愿景》(Computing: A Vision for Data Science) 的论文。Dhar V在《美国计算机学会通讯》(Communications of the ACM) 上发表题为《数据科学与预测》(Data Science and Prediction)的学术论文。

✓ 代表性专著有：Provost F 和Fawcett T出版了专著《面向商务的数据科学》(Data Science for Business) 。Mayer-Schönberger V 和Cukier K出版了专著《大数据——一场即将改变我们生活、工作和思维的革命》(Big Data: A Revolution That Will Transform How We Live, Wok, and Think)。Schutt R和O’Neil C 出版专著《数据科学实践》(Doing Data Science)。

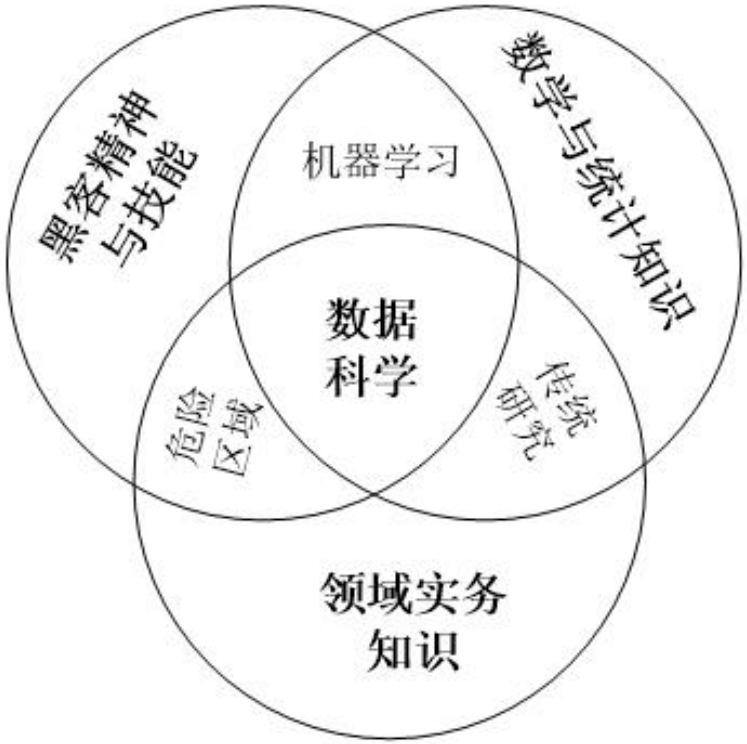
二、数据科学的内涵和发展

（3）逐渐成熟期（2014年至今）

- 2014年： Zumel N, Mount J, Porzak J等出版了专著《基于R的使用数据科学》(Practical Data Science with R)，较系统地介绍了如何运用R开展数据科学工作。
- 2015年：美国白宫任命D J Patil 为首席数据科学家。Lillian Pierson 出版专著《数据科学的傻瓜用书》 (Data Science for Dummies)。Monya Baker在《自然》杂志上发表论文《数据科学——产业诱惑》(Data Science: Industry Allure)。

三、数据科学的学科地位

2010年，Drew Conway提出了“数据科学维恩图”。根据图形可知，数据科学位于统计学、计算机科学和某一领域实务知识的交叉之处，具备较为显著的交叉型学科的特点。即数据科学是一门以统计学、计算机科学和领域知识为理论基础的新兴学科。



四、数据科学的成熟度曲线

1. 新技术成长曲线

2014年，Gartner提出了新技术成长曲线(Gartner’s 2014 Hype Cycle for Emerging Technologies)，认为数据科学的发展于2014年7月已经接近创新与膨胀期的末端，将在2～5年之内开始应用于生产高地期（plateau of Productivity）。

四、数据科学的成熟度曲线

2. 数据科学成长曲线

2016年，Gartner提出了数据科学本身的成长曲线(Hype Cycle for Data Science)。

从下图可以看出，数据科学的各组成部分的成熟度不同。

R的成熟度最高，已广泛应用于生产活动。

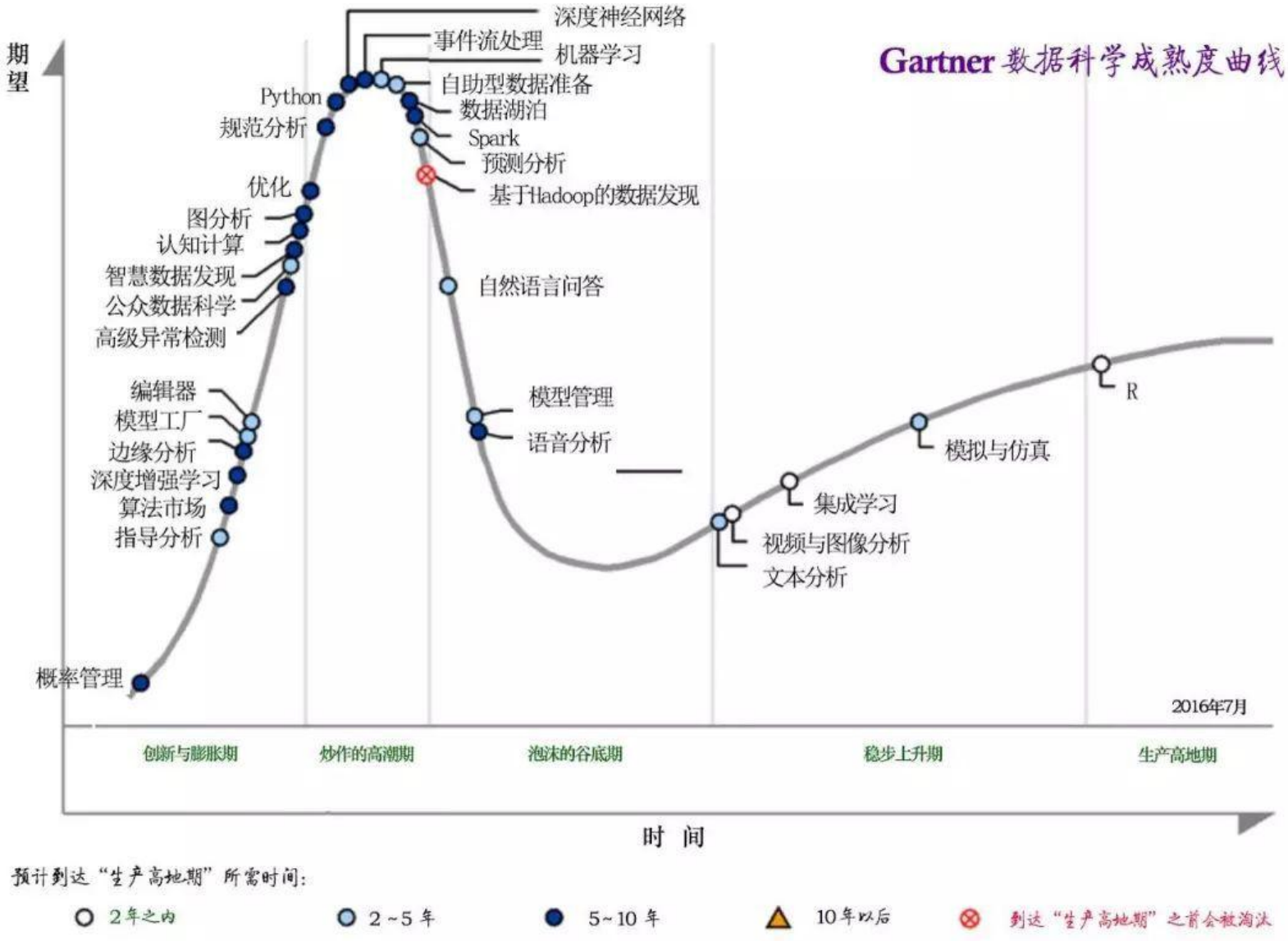
其次是模拟与仿真、集成学习、视频与图像分析、文本分析等，正在趋于成熟，即将投入实际应用。

基于Hadoop的数据发现可能要消失。

语音分析、模型管理、自然语言问答等已经渡过了炒作期，正在走向实际应用。

公众数据科学、模型工厂、算法市场（经济）、规范分析等正处于高速发展之中。

四、数据科学的成熟度曲线



五、数据科学的理论体系


数据科学
数据加工
数据计算
数据管理
数据分析
数据产品开发
…
基础理论
理念
理论
方法
技术
工具
…
理论基础
统计学
机器学习
数据可视化
（某一）领域实务知识与经验

五、数据科学的理论体系

从理论体系来看，数据科学主要以统计学、机器学习、数据可视化以及领域知识为理论基础。其研究的主要内容包括：基础理论、数据加工、数据计算、数据管理、数据分析、数据产品开发等。

1. 基础理论

主要包括数据科学中的新理念、理论、方法、技术及工具以及数据科学的研究目的、理论基础、研究内容、基本流程、主要原则、典型应用、人才培养、项目管理等。需要特别提醒的是，“基础理论”与“理论基础”是两个不同的概念。数据科学的“基础理论”在数据科学的研究边界之内，而其“理论基础”在数据科学的研究边界之外，是数据科学的理论依据和来源。

五、数据科学的理论体系

2. 数据加工（Data Wrangling 或Data Munging）

为了提升数据质量、降低数据计算的复杂度、减少数据计算量以及提升数据处理的精准度，数据科学项目需要对原始数据进行一定的加工处理工作——数据审计、数据清洗、数据变换、数据集成、数据脱敏和数据标注等。值得一提的是，与传统数据处理不同的是，数据科学中的数据加工更加强调的是数据处理中的增值过程，即如何将数据科学家的创造性设计、批判性思考和好奇性提问融入数据的加工活动之中。

五、数据科学的理论体系

3. 数据计算

在数据科学中，计算模式发生了根本性的变化——从集中式计算、分布式计算、网格计算等传统计算过渡至云计算。比较有代表性的是Google三大云计算技术（GFS、BigTable和MapReduce）、Hadoop MapReduce、Spark和YARN。计算模式的变化意味着数据科学中所关注的数据计算的主要瓶颈、主要矛盾和思维模式发生了根本性变化。

五、数据科学的理论体系

4. 数据管理

在完成“数据加工”和“数据计算”之后，还需要对数据进行管理与维护，以便进行（再次进行）“数据分析”以及数据的再利用和长久存储。在数据科学中，数据管理方法与技术也发生了重要变革——不仅包括传统关系型数据库，而且还出现了一些新兴数据管理技术，如NoSQL、NewSQL技术和关系云等。

五、数据科学的理论体系

5. 数据分析

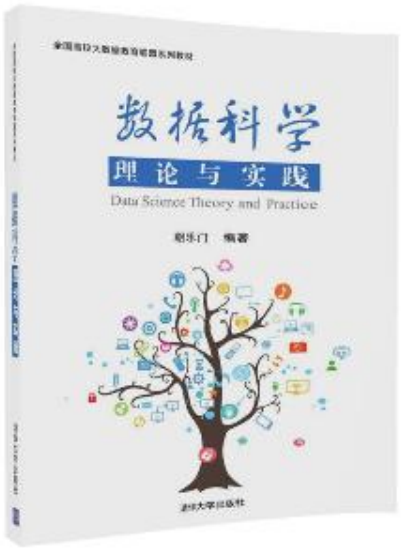
数据科学中采用的数据分析方法具有较为明显的专业性，通常以开源工具为主，与传统数据分析有着较为显著的差异。目前，R语言和Python语言已成为数据科学家较为普遍应用的数据分析工具。

6. 数据产品开发

“数据产品”在数据科学中具有特殊的含义——基于数据开发的产品的统称。数据产品开发是数据科学的主要研究使命之一，也是数据科学区别于其他科学的重要区别。与传统产品开发不同的是，数据产品开发具有以数据为中心、多样性、层次性和增值性等特征。数据产品开发能力也是数据科学家的主要竞争力之源。因此，数据科学的学习目的之一是提升自己的数据产品开发能力。

六、推荐阅读

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS



谢 谢!

Thank You