HZ BOOKS
华章教育

张诚 张琦◎编著

商业数据科学

数据价值与机器学习实战

DATA SCIENCE FOR BUSINESS

DATA ANALYSITCS AND MACHINE LEARNING

数据科学与大数据管理丛书

机械工业出版社
China Machine Press

数据科学与大数据管理丛书

商业数据科学

数据价值与机器学习实战

张诚 张琦◎编著

DATA SCIENCE FOR BUSINESS

DATA ANALYSITCS AND MACHINE LEARNING

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

商业数据科学：数据价值与机器学习实战 / 张诚，张琦编著．—北京：机械工业出版社，2019.1
（数据科学与大数据管理丛书）

ISBN 978-7-111-61835-5

I. 商… II. ①张… ②张… III. 机器学习－应用－商业信息－数据处理 IV. F713.51

中国版本图书馆 CIP 数据核字（2019）第 002123 号

如果用一段话来总结这本书的内容，我很愿意引用 2013 年第一次写下课程教案时对课程的描述：“它不是一门人云亦云的课程，不讲理论，以实战为主，用一套套实际数据来讲如何从数据里发掘商业问题和检验商业假设；它是一门商业素养和技术算法综合应用的课程，需要有开放思想和开放学习能力的同学来参与和体验；它是一门动手性极强的课程，以脸书、腾讯、雅虎等公司的分享数据为基础，培养学生过硬的推理和分析能力；它是一门跨学科的课程，为同学未来领导跨部门商业数据分析团队铺路。你，准备好了吗？

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：冯小妹　　责任校对：李秋荣
印　　刷：三河市宏图印务有限公司　　版　　次：2019 年 2 月第 1 版第 1 次印刷
开　　本：185mm×260mm　1/16　　印　　张：14.5
书　　号：ISBN 978-7-111-61835-5　　定　　价：59.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379210　88361066　　投稿热线：（010）88379007
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjg@hzbook.com

前言

2013年开始我为学院MBA学生讲授数据分析选修课程，至今已超过5年。开设当年正值大数据概念引起国内关注之时，为了吸引学生眼球，给课程名加上个“大”字。商学院的同学多在企业各业务职能部门从事管理工作，他们的实际需求并不在于成为技术专家，而在于直面日渐增长的数据对业务决策带来的挑战和机遇，能在实际工作中将商业挑战与企业数据结合提出正确的商业分析需求，并有效组织和管理团队完成基于数据的决策。按照这样的教学思路和近五年的实战授课，如今终于可以完成此书，供读者阅读了解企业如何结合数据和算法实现新的商业应用。

如果用一段话来总结这本书的内容，我很愿意引用2013年第一次写课程教案时的简介：“它不是一门人云亦云的课程，不讲理论，以实战为主，用一套套实际数据来讲如何从数据里发掘商业问题和检验商业假设；它是一门商业素养和技术算法综合应用的课程，需要有开放思想和开放学习能力的同学来参与和体验；它是一门动手性极强的课程，以各大互联网公司分享数据为基础，培养学生过硬的推理和分析能力；它是一门跨学科的课程，为同学未来领导跨部门商业数据分析团队铺路。你，准备好了吗？”回首过去五年多的持续教学和不断改进，在不惑之年再来面对这颇有些年少轻狂的描述，觉得自己算是做到了不忘初心。

将课程教案改编成书的重要原因在于市场中始终缺少这样一本将商业需求和数据分析结合在一起的书。为了备课，我在过去五年参阅了很多相关书籍，包括偏重介绍商业应用的《大数据时代：生活、工作与思维的大变革》（Viktor Mayer-Schönberger，Kenneth Cukier）、《决战大数据：驾驭未来商业的利器》（车品觉）等，以及偏重技术算法的《数据挖掘十大算法》（Xindong Wu，Vipin Kumar）、《数据分析：数据科学应用场景与实践精髓》（Bart Baesens）和《机器学习》（周志华）等，这些书

各有特点，我也选用作为课程推荐读物。但通过几年授课经验和学生的反馈，觉得市场已有书籍用于MBA同学或企业管理人员学习和实践始终还差两点：介绍商业应用的书籍通常缺乏技术的详细做法，当同学们被书中的精彩描述所打动，希望能深入下去实践时却不知该如何开展；而介绍技术的书籍则通常缺乏与企业从需求到问题的结合，同学可以学到工具却不太能对应到真正的商业现象和管理思维。工作实践中，只有将商业现象转变为需求定义、问题描述并最终确定具体技术方案后，工具才可以上阵解题。借《庄子·列御寇》中的“朱泙漫学屠龙于支离益，单千金之家，三年技成而无所用其巧”故事做个不恰当的比喻，算法类书籍通常能教读者屠龙之技，却不教读者如何找龙，而商业类数据帮读者找到龙，却不教屠龙之技。本书希望能弥补这一缺陷。

此外，很多大数据相关书籍不太适合自学。如果读者有买过大数据方面偏技术的书，想必或多或少都有被书中大段的数学公式和推导“震撼”，最终将书“供奉”起来的经验吧。我猜其震撼程度和供奉速度随读者的年龄、职位增加而增大、变快。由于大数据这一概念已经达到举国皆知、全民皆知的程度，绝大多数希望了解和掌握数据分析精髓的读者并不具备计算机、统计等相关专业知识，这就要求一本好的教材做到深入浅出，并能教会“文科生”掌握这方面的知识。“文科生”并不是我的某种偏见，而是有太多我的学生在课程伊始会来问我“老师，我是文科生，能学好这门课吗”之类的问题，有其代表意义。另外一类常碰到的问题是：“老师，我工作很忙，但很希望了解数据分析，能学好这门课吗？”借此机会，我隆重宣布：这本书适合广大文科生，而且适合“三心二意”的同学（即本职不是数据分析相关工作，但希望了解、参与和领导这方面工作的企业人员）。此书的目的就在于满足广大日常忙于本职工作，但有兴趣掌握数据分析精髓的企业文科生们的学习需求。对于这个承诺，此书毫无压力，因为它本就来自MBA实践授课，经过最近五年的改进，获得了学生的良好反馈。

那么它是如何做到的呢？首先，它保证能用小学和中学的知识将大数据分析中主要需要掌握的计算机算法（也叫机器学习算法）思路讲清楚！要知道，文科生、理科生在中小学期间学到的数学知识可没有什么差异。每次我在课程第一节课时都会给学生一个承诺：保证用中小学（主要是小学和初中，偶尔用到点中学奥数）知识教会大家机器学习，超纲的话可以投诉我。承诺一直维持到2017年，由于那只狗（AlphaGo）大热，在授课时加入神经网络知识才没法保持。所以这本书里，我承诺主要用中小学知识教会读者大数据分析的算法精髓，偶尔会涉及大学一年级的微积分课程知识（幸运的是，这也是文科生必修的课程，而且已经成为部分高中的教

学内容)。如果读者能将书中关于神经网络的知识内容用中小学知识讲清楚，或有别的改进建议，请通过网络告诉我（我在网上建了反馈系统 @ https://www.wjx.top/jq/20812517.aspx  ）。一旦采用，我将给您邮寄此书的签名新版（如果此书还能再版，哈哈），并奉上 100 元（或捐赠给指定慈善机构或个人）作为感谢。

其次，根据 2012 年一篇发表在美国科学院院刊上的学术研究[⊖]，论文正文里每页多 1 个数学公式会导致该论文的引用次数（相对公式少的）减少 28%（引用是指其他学术论文提及它，是学术领域评价论文影响的一个重要指标，可以类比为认同程度)。因此，本书杜绝各类高等数学公式（简单来说，就是那些使用超过 x，y，z 和 a，b 的变量，或者超过加减乘除符号的方程、不等式和其他数学公式)，但请读者允许我保留中小学时学到的四则运算式、二元一次方程组和解析 / 立体几何，毕竟初等数学公式可以比文字更简洁易懂地说清楚一些道理。从这个角度来说，本书也适合中小学生学习，或者作为家庭指导儿女了解大数据的教材。希望本书可以为大数据分析领域的书籍带来一点小清新。

当然，采用形象易懂的描述并不意味着降低对大数据分析的理解程度或本书的质量。事实上，把前人推导出来的公式和算法整理列出的工作量，通常远少于在理解算法本质后将其和日常现象结合描述出来的付出。而这，也是我在任教十余年中感受到的教学的最大挑战和责任：把大部分学生教懂的难度远远大于把大部分学生教不懂。想必不少读者在大学期间（也可能从中学开始）就有“这些公式好难”或者“学这些数学公式有什么用”的困惑，以及崩溃于老师在罗列公式和推导步骤时说的“显而易见，我们可以得到……”或者“很简单，我们可以看出……”的经验。如果回忆一下当年某些名为“数理统计和概率论”“线性代数”“随机过程”“(非）线性优化”等课程的学习感受，想必留给大多数同学的记忆不仅是很多难以理解的数学推导，还有不知道为什么要学这些的无助。尽管我在科研道路探索了十几年后越来越能感受到数学之美，但也不无遗憾：如果当年老师能让我明白这些道理所对应的社会事物和现象，那我会更早热爱科学。

所以，在课堂上和这本书里，我会努力用粗浅的语言去描述高深知识的本质，让读者明白：原来不过如此。教学不是吓倒学生，让他们畏惧，而是让学生感受到自己也能掌握和做好的自信。实现了这步，就可以更好地帮助学生融会贯通，创造

⊖ Fawcett T W, Higginson A D. Heavy Use of Equations Impedes Communication among Biologists[J]. Proceedings of the National Academy of Sciences of the United States of America, 2012, 109(29).

新的想法和做法！为此，我采用了几点小创新：首先，除最开始的介绍部分和结束的管理部分，其余各章内容先从商业现象入手，展开为商业需求和问题描述，然后提出解决思路，最终介绍与此思路相关的机器学习算法，希望通过这个方式打通业务和技术的隔阂，建立从商业到数据的分析和决策逻辑，真正帮助大家读有所得，并能很快转化为解决工作中其他实际问题的能力。其次，本书将主要基于通俗易懂的语言（而非专业术语），让读者感受到我娓（不）娓（够）道（专）来（业）的知识梳理。再次，本书正文里采用文字和初等数学的方式来介绍相关机器学习算法，将更严谨专业的数学表达放在对应插文里，供有兴趣深入了解的读者参阅。最后，本书附带教课课件，也即将推出网络教学视频选集，可供大家借鉴。

这本书的完成离不开五年来选修本课程的我院数百位 MBA 同学的反馈和建议，也离不开三任助教（王西蒙、田婧和张琦）的付出。西蒙作为首任助教，协助我应对了很多早期由于准备不足带来的慌乱；田婧持续帮助我改进授课内容和分析案例，做了很多教学基础工作；张琦完善授课视频，并协助我整理教案文稿和丰富书中内容。现在西蒙已远赴北美亚马逊工作；田婧现在北美佐治亚大学交流，即将博士毕业走上工作岗位；张琦即将通过我院博士班资格考，成为新一届的女博士。他们都有各自人生的精彩，也通过本书留下各自经历的痕迹，特此记之。

在本书出版之际，需要感谢在这大半年时间里一直督导本书出版的编辑张有利老师和责编冯小妹老师，没有你们的鞭策和鼓励，患有严重拖延症加上语言表达困难的我很难按时完成本书。此外，要感谢 2018 年毕业前往苏州大学工作的师妹沈怡，你第一个帮忙通读初稿，并给出了很多中肯的意见。还要感谢浙江大学陈熹老师、南京大学宋培建老师、清华大学卫强老师和清华大学出版社刘向威老师，你们或交流课程心得，或推荐相关读物，让我受益匪浅。

此外，感谢 2015 ～ 2017 年上过本课的全体 MBA 同学，你们众志成城帮忙打 Call 取名字，并通过民主投票最终确定本书书名。特别感谢徐建林同学，甘心做我的小白鼠，周末熬夜阅读帮忙检查错漏字，并测试作业数据。很抱歉之前上课的同学没有建微信群，无法再联系到。但希望你们看到本书后，能感受到我想表达的心意，也祝愿所有同学在工作岗位上取得更大的进步，将新一代商业数据科学的思维和机器学习方法真正用到企业实践中。

张诚

2018 年 11 月 2 日立冬

目录

| 第 1 章 |

大数据及其应用

很高兴能够和大家分享大数据的相关知识。首先，我们用一串数字作为开篇，如图 1-1 所示。

3.14159265358979323846264338327950288419716939937510582097494459230781640628620899862803482534211706798214808651328230664709384460955058223172535940812848111745028410270193852110555964462294895493038196442881097566593344612847564823378678316527120190914564856692346034861045432664821339360726024914127372458700660631558817488152092096282925409171536436789259036001133053054882046652138414695194151160943305727036575959195309218611738193261179310511854807446237996274956735188575272489122793818301194912983367336244065664308602139494639522473719070217986094370277053921717629317675238467481846766940513200056812714526356082778577134275778960917363717872146844090122495343014654958537105079227968925892354201995611212902196086403441815981362977477130996051870721134999999837297804995105973173281609631859502445945534690830264252230825334468503526193118817101000313783875288658753320838142061717766914730359825349042875546873115956286388235378759375195778185778053217122680661300192787661119590921642019893809525720106545858632788659361533818279682303019520353018529689957736225994138912497217752834791315155748572424545150695950829533116861727855889075098381754637464939319255060400927701671139009848824012858361603563707660104710181942955596198946767837449448255379774726847104047534646208046684259069491293313677028989152104752162056966024058038150193511253382430035587640247496473263914199272604269922796782354781636009341721641219924586315030286182974555706749838505494588586926995690927210797509302955321165344987202755960236480665499119881834797753566369807426542527862551818417574672890977772793800081647060016145249192173217214772350141441973568548161361157352552133475741849468438523323907394143334547762416862518983569485562099219222184272550254256887671790494601653466804988627232791786085784383827967976681454100953883786360950680064225125205117392984896084128488626945604241965285022210661186306744278622039194945047123713786960956364371917287467764657573962413890865832645995813390478027590099465764078951269468398352595709825822620522489407726719478268482601476990902640136944374553050682034962524517493996514314298091906592509372216964615157098583874105978859597729754989301617539284681382686838689427741559918559252459539594310499725246808459872736446958486538367362226260991246080512438843904512441365497627807977156914359977001296160894416948685558484063534220722258284886481584560285060168427394522674676788952521385225499546667278239864565961163548862305774564980355936345681743241125150760694794510965960940252288797108931456691368672287489405601015033081679286809208747609178249385890097149096759852613655497818931297848216829989488
图 1-1　π

1.1　大数据的特性

读者应该对这串数字比较熟悉——3.141 592 6，通常我们会在小学三年级学到圆周率 π，老师会要求我们背到小数点后 2 至 4 位数，用于计算圆的周长或面积。那么，在不到 10 岁接触到这个数时，你是否想过 20 年后会和它产生无

数交集呢？比如，你为人生第一张银行卡设立的 6 位数密码其实早就存在于 π 里面了。

为什么这么说呢？首先我们来回顾一下什么是 π。π 具有无限不循环的特性，这意味着所有数字组合都可能在里面出现。我用大学同学的故事来举个例子，他的生日是 1 月 23 日，于是他取 π 中第 123 位开始的 6 个数作为密码。其实无论大家如何选择密码，在无限不循环的 π 中，总会有 6 位数和你的密码重复，只要 π 里面的数字列得足够长。如果大家有兴趣，不妨看看你的密码是否在其中。这里只展示了几千位，如果不够用，可以去网上查更长的 π 值。

接下来进一步思考，假设我们用任何 4 个数字代表 1 个汉字（中文常用字不超过 1 万字），那么如果这个数字足够长，里面会不会出现你人生写过的第一篇日记或第一封情书呢？小说《三体》的作者刘慈欣的另一部短篇小说《诗云》中提到这样一个问题：任何 4 个数字组成 1 个汉字，能不能通过汉字的排列组合创造出一篇最好的、超越历史上所有诗词大家（如李白、白居易等）作品的七言绝句？后来外星人用整个宇宙的物质来创建数字组合，但最终也没能找出来，因为排列组合的可能太多，即便知道那首诗就在宇宙中，但是可能永远也不知道它在哪儿。就像 π，它的无限不循环的数字可以长到容纳所有事实，而我们却可能永远也找不到自己想要的事实。

小学学到的圆周率 π 体现出了当代**大数据的两个特点**。第一，**数据爆炸式增长**，已经多到能逐渐逼近现实的所有情况。就像周星驰电影《国产凌凌漆》里面的一句台词：一张草纸也有它的价值。尽管一个用户的信息不能影响企业决策，但当企业掌握的消费者数据由原来的 1 千条、1 万条，变成百万条、千万条、亿条的时候——这个世界的人口也不过 50 多亿，情况就完全不同了。企业现在可以拥有某个区域内大部分用户的表现和信息，也就可以从这些数据中挖掘出更多的价值用于决策。比如做市场调研时通常会遇到一个问题：随机抽几百人能否代表市场需求？大家做企业用户需求或产品设计调研时应该都遇到过类似的困惑。如果现在我们可以掌握 1 亿用户的信息，这个困惑就被大大削弱，甚至不用调研也可以了解所有用户的偏好和感受，那么企业产品部门就有更多机会做出好的产品，市场部门也会有更多机会提供好的服务。这就是数据增长带来的价值。

第二，**数据处理和发现价值的成本不低**。企业知道有价值的数据在公司里面，但是更需要知道它存储在哪儿。因此，挖掘出价值需要更高的技能，需要新的投入。好比在圆周率 π 中，即便你相信你的密码一定存在其中，但是却不知道存在于哪个位置。相对应地，这本书主要包含两个主题：一个是数据的价

值；另一个是数据的使用，即怎样把价值从数据中挖掘出来。大家在学习的过程中也会慢慢感受到数据的成本其实并不低。有了这样全面的认识，未来就能在企业里做出很好的应用。

1.1.1　数据价值实现的挑战：智能投顾

2017 年我参加了一个银行家沙龙，当时的讨论涉及金融科技里面的智能投顾，简单理解就是让机器人或算法帮客户理财。现在的投顾大多由人工完成，一般银行会设置 VIP 室，客户经理会请有理财意愿的客户进去喝喝茶、聊聊天，提供一对一的投资理财服务。当客户资产达到一定程度时是值得使用这项服务的，但是一位客户经理能够服务的人数有限，那么是否可以让算法来帮客户理财呢？首先要了解相应背景知识，应法律规范要求，金融产品的所有条款都写得非常明确，因此这些规则可以被量化计算。

业界在 AlphaGo 出来后预测过多次哪些行业的工作可能会被取代，其中理财服务永远名列前茅。这里的金融不是指交易，而是指理财咨询、顾问，以及保险。因为这些领域内的服务条款都必须量化，不能存在任何含糊不清的表述，这与下围棋的明确规则相同，所以计算机非常容易掌握这些规律。接下来的问题是怎样匹配用户的需求。当时有位银行分行行长说，银行掌握了很多客户的需求信息，比如客户的每一笔消费记录，但是他们不清楚应该如何利用这些信息向客户推荐产品。一般来说，客户经理都经历过严格的训练，不仅要学理财、心理学，还要在和客户的交流中揣摩对方的想法和风险偏好。这些风险偏好原来是人工识别的，如果用交易系统，怎样识别客户的风险偏好呢？这是一个很难回答好的问题，因为即便掌握了用户的全部消费数据，也不一定能掌握用户的风险偏好。消费购买产品是为了获得产品的使用价值，这不是风险投资，风险投资往往会面临损失。因此，根据消费数据推断出用户的风险偏好是有难度的。

1.1.2　大数据时代的数据

图 1-2 简单罗列了近年报告中总结的数据。现在大数据之所以变得热起来，原因之一就是大家意识到当企业拥有全市场的数据（或者大部分市场的数据）后，决策方式将发生变化。五年里这些数字不断增长，而每次我都会问同一个问题：看到这些数字，你们觉得数据有什么价值？

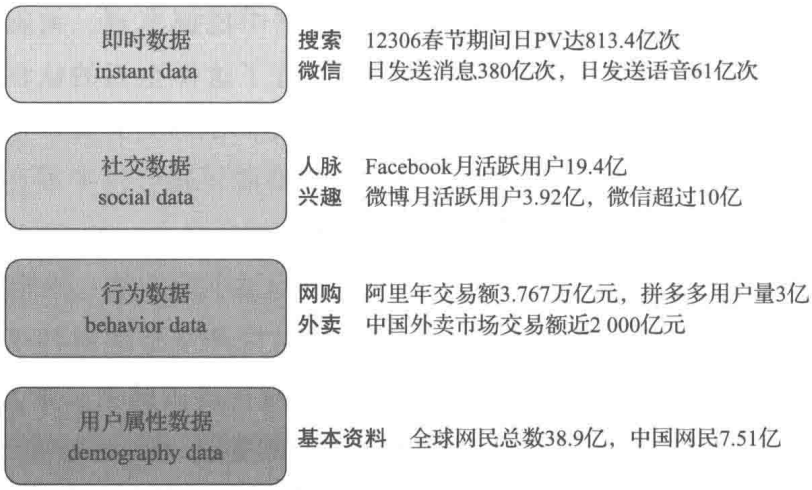

图 1-2 常见的用户数据规模（截至 2017 年）

众所周知，股票定价的基础在于其市场价值，但其价格的波动往往来自市场对公司的不同理解。因此，一个行之有效的交易策略是利用市场的反应（如恐慌或乐观）来预测股价。以前市场舆情方法效果不好，原因在于只有媒体和股评家的反应，尽管他们可以影响很多人，但是我们永远不清楚每次评论会影响多少人，因而无从得知全市场的反应。而现在我们有机会通过微博收集全市场的反馈来进行分析，这样的预测就会精确得多。前几年有论文（参见本章扩展部分）用 Twitter 数据作为市场舆情来预测股价波动，因为 Twitter 用户包含美国大多数用户，只要这些用户通过 Twitter 表达情感，就可以通过文本分析获得全市场恐慌或者乐观的倾向。巴菲特说过，在别人恐慌的时候我贪婪，在别人贪婪的时候我恐慌。当有办法知道整个市场对特定事件、特定公司的看法时，我们自然而然就能设计出新的交易策略。

同理，我们可以想象一下，投资者如果关注某只股票或公司，很可能会先通过搜索引擎（如百度）去搜索相关信息。当你掌握了全中国大部分网民每人每天网络搜索的信息，并从中间选取与股票和上市公司有关的那部分，你就能够知道市场对它们的关注度，从而有可能建立一个对股市股价波动做出解释的系统。

那么，淘宝和京东的数据是否可以应用到金融市场呢？我认为可以。当然，这些数据不一定能分析全市场，使用数据的方法也会和百度有区别。淘宝和京东有中国最丰富的消费数据，通过这些数据可以了解老百姓每天究竟需要什么日常用品。我又要提到巴菲特了，他说他最喜欢投资生产社会日常用品的公司。淘宝或京东利用中国几亿用户的日常消费购买数据，可以更好地判断消费品的走势，以及市场对各类品牌的消费产品的需求，这样是否可以更准确地预测相应公司的中长期价值呢？

腾讯是中国最大的互联网综合服务提供商之一，提供最广泛的社交和新闻服

务，在某种程度上它能够很好地刻画人之间的相互影响。后面我们将给大家分享一个思路——如何找出市场中影响力大的个体，这样更有可能预测股价后面的波动。其实股民中存在显著的羊群效应，如果我们能找出其中的舆论领袖或有影响的媒体、股评师报道，并能准确地刻画他们的影响程度，那么追踪他们的反应就会有很大收获。我们经常说，在中国股市里谁最有发言权，你跟着他们行动就可以了，其实说的就是这个道理。

1.2　数据发展历程

1.2.1　用户成为数据产生的源头

在了解数据价值之前，我们首先要明了为什么大数据会兴起，这样可以更好地理解后续的算法思路。我们用图 1-3 来总结一下过去 50 年信息技术和数据的发展历程。20 世纪 60 年代芯片的小型化加上存储的出现，形成了第一台电脑，这是计算机的启蒙时代。然后出现了像 Excel 电子表格那样的能够分门别类存储数据的数据库，再加上网络的发展，使得机器之间可以通信。企业随之发现，部门之间可以采用这种方式来促进信息沟通，每个部门产生的信息（如财务、运营）都可以汇总到中心数据库，同时其他部门也可以使用，这样就可以促进企业部门间的协作，如财务和运营共同参与物料采购管理。读者在本科阶段可能已经通过《信息导论》或《管理信息系统》了解了相应知识，在这里我们称之为“联结的部门和流程”。公司里每天上班打卡都会用到这样的系统。

当企业发现部门之间的信息可以连通之后，自然而然会想到可否将企业和企业直接连通在一起。20 世纪 90 年代末到千禧年初，出现了相应的电子商务模式，一个典型的案例是亚马逊和戴尔。这种模式通过联结企业更有效率地实现了企业间协作，亚马逊和戴尔自身并不制造产品，却分别成为美国著名的零售商和计算机制造商，这是一种全新的商业模式。在这一阶段，数据都来自业务的驱动，也就是说，只有发生了事才会产生数据，数据用于记录和追踪事件的发展。我们可以称这个阶段为“联结的企业”。

那时之所以没有出现大数据，与数据产生源的差别及后续的改变有关。当时主要的信息源是企业业务部门收集整理的数据，其成本与商业技术的投入有关。随着企业间以及企业和市场间的联结不断增强，技术的成本逐渐降低，而且被更多的人接受，突变由此产生。举个例子，我出生于 20 世纪 70 年代，当时电视里

明星演唱歌曲，观众们非常兴奋时都是挥舞双手，而现在都是挥舞数码照相机、平板和手机，随意拍个照，立即就分享给了别人。由此大家在不经意间跨入了一个大数据时代——每天每时每刻，你输入和发送信息或照片时，就为数据时代贡献了一份数据，你跟你的朋友发送的每一条信息都同时告诉了公司几件事情：第一，你什么时间什么地点在干什么事情；第二，你喜欢什么内容，喜欢跟谁说；第三，那个人跟你什么关系。之前的时代是由企业业务驱动产生数据，而当前这个时代是由人驱动产生数据，每个用户随时随地随性地就可以和其他人互动，表达自己的情感和完成购买等。因此，这个时代把人、物和社会联结起来了，用户（而非企业）成为数据的最重要来源。

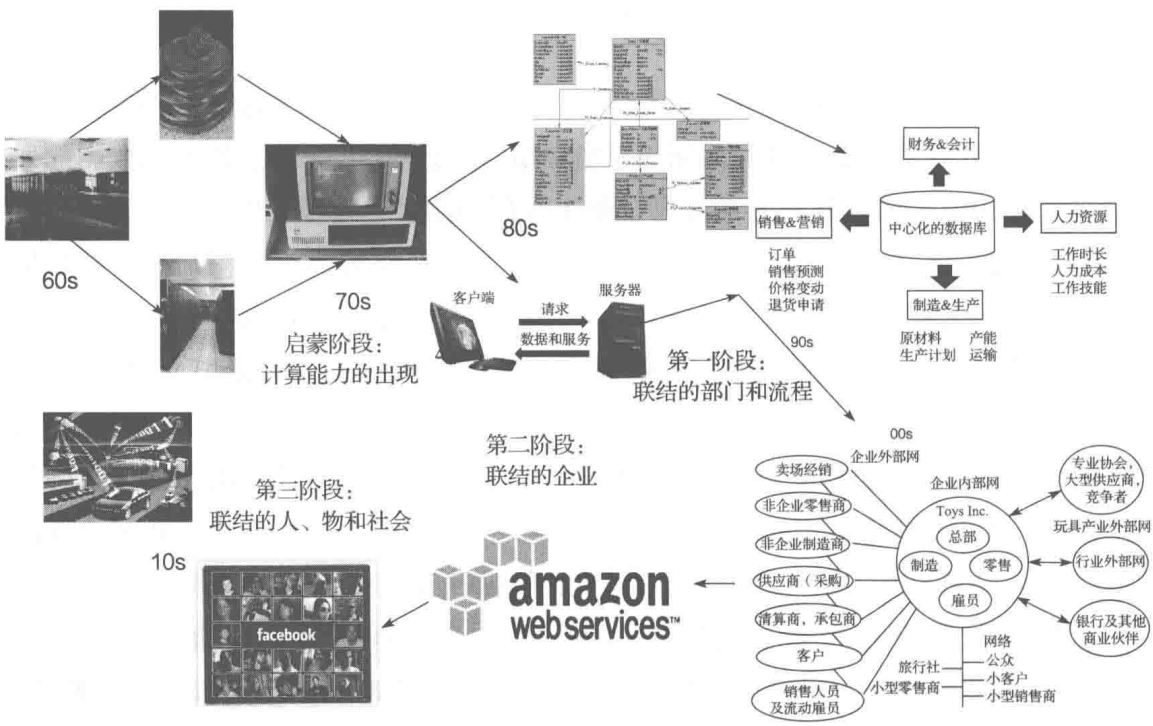

图 1-3　信息技术和数据的发展历程

这个转变直接带来了数据量的爆发和数据成本的降低。本科时我有位做语音识别的师兄在微软研究院实习，他当时负责采集普通话的中文标准发音，大概一个字的费用是两元钱。大家可能觉得成本有点高，但按汉语常用字 1 万来计算，这个费用对微软来说并不算贵。但考虑到中国十几亿人发音习惯和方言的差异，如果要做一个能识别各地区中文语音的程序，公司的投入将远超过标准发音的采集。

现在我们有类似 Siri 和小娜这样的语音识别程序，安装在手机上可以根据用户的语音提示完成操作。它们同样需要采集和处理语音，不过与过去相比投入和

成本大大降低。除了语音识别技术自身的发展之外，另一个重大改变是语音数据的增加和采集成本的降低。人们随时随地都会自发地通过网络传播语音，这些语音保留在网络和公司中，从而为语音程序提供了大量低成本语料。文本分析和舆情监控也是这个道理。比如你花 10 万元请一位计算机工程师，让他写一段网络爬虫代码，只要运行得时间够长，就可以在几个月之内获得上千万人的上亿甚至几十亿条网络信息，而每条信息的获得成本仅有万分之一元，每个人信息的成本只有千分之一元。由此可见，单位数据的成本在急剧降低。

现在大数据应用层出不穷，背后有一个必然的原因：数据量级增加的同时，单位采集成本在降低。人们自由表达的机会越多，留在网上的数字化痕迹也就越多，每条信息的采集成本也就越低。在数字化过程中跻身前列的是 20 世纪 90 年代后期开始兴起的数字照片，实现了从胶卷照片到电子文件的转换。接着发展下去是视频数字化，之后是语音数字化，近几年比较流行的是人体数字化。我们有幸处在了这样一个时代，体验着这样的变革。我们的孩子也许会对此习以为常，而我们正好处在了这个交界点上。我在 20 世纪 90 年代读大学时还在讨论信息技术是否有用，以及如何产生价值，而现在讨论的却是如何利用信息技术把公司（包括亚马逊、脸书、谷歌、BAT 等）联结起来，把社会联结起来。这个时代已经过去了 10 年，取得了激动人心的成就。

与之相对应，商业模式也发生了重大变化。首先是出现了免费模式，原来我们只能在注册网上会员时享受免费的注册服务，很少遇到其他免费的产品或服务，但近年来越来越多的商业开始尝试免费的模式。我们见证了淘宝、百度、腾讯这些“免费”商业模式的成功，未来可能还会见证它们如何被新的模式取代。阿里巴巴这家不收费的公司市值已经达到上千亿元，然后它又把支付系统以及金融服务拆分出来，又分别达到千亿元级的水平。它是怎么赚钱的？市值那么大是靠烧钱烧出来的吗？这个细节我们后面会讨论。现在，亚马逊、阿里巴巴提供的云服务用事实证明信息或者数据的确会赚钱。

1.2.2　大数据应用：新零售下数据联结的价值、挑战和机遇

2017 年最火的一个词是新零售，或者说是线上线下、O2O。阿里巴巴三年前就开始深耕支付，想办法打通线下，大家也慢慢能感觉到线下买东西用支付宝比较方便，而就在用户支付的同时它也记录下了用户的消费内容，从而把店铺交易剖视成了数据。当然，一些连锁超市并不愿意把信息提供给阿里巴巴。

如果知道了用户每天在超市买什么，吃什么，而线下支付的账号是用户的支

付宝账号，而支付宝又恰好是淘宝账号，这样线上和线下就串联在了一起。既然连锁商业超市不给线上公司提供数据，那就入股、并购，比如腾讯入股永辉、阿里巴巴入股银泰。这些现象前几年就已有征兆，只不过到现在才集体爆发。

BAT 除了出于财务投资的考虑，为什么要进军零售业呢？互联网之前的零售毛利本就很低。第一种可能是它们找到了一个原来不盈利的行业，但现在有办法帮这个行业赚钱。比如精准营销，线下大型连锁零售商用入户海报做营销，一期印 24 页，一年成本可能就达 3 000 万元。如果把 24 页减少到 12 页，一年就可以节约 1 500 万元以上，而这 1 500 万元就是净利润。可见入户海报是一种低效高成本的营销方式。而线上公司可以通过精准营销更便宜有效地影响用户，线下原来做亏的事，线上可以做到盈利。

第二种可能是和模式运作有关。比如京东，2018 年财报变正，市场普遍认为不仅仅是因为销售增加了，更是金融和物流（包括物流资产）带来的。之前淘宝盈利就不错，大家会好奇背后的原因，因为淘宝本身不卖货，没有传统意义上的通过采购销售价差带来的销售收入。但是它使得阿里巴巴成为全球前五的广告平台，最近几年阿里巴巴的盈利大约 80% 都来自电子广告，所以淘宝之前的盈利模式不是交易卖货，而是卖广告。

我们从模式的角度仔细想想淘宝过去主要的收入可能来自哪里。它提供的注册、发布信息和交易服务是免费的，也没有收厂商的入场费，过去也没有所谓的云服务。除了少量的增值服务费，主要就是淘宝收取的站内产品推广费，或者叫广告费，那么这一部分收入占了总营收的多少呢？

首先来看一下阿里巴巴的营收情况。表 1-1 展示了阿里巴巴从 2012 年至 2016 年的财务报表情况。以 2016 年为例，该年阿里巴巴集团销售收入为 159 亿美元，折旧和摊销费用为 10.4 亿美元，不包含折旧和摊销成本 48.2 亿美元。总收入达到了 100.4 亿美元。

表 1-1　阿里巴巴集团财务年报（2012 ～ 2016 年）（财务年度为 4 月到下年 3 月）

	2012	2013	2014	2015	2016
销售 / 收入	3.14B	5.5B	8.58B	12.3B	15.9B
包含折旧 & 摊销的销货成本	1.05B	1.57B	2.24B	4.18B	5.86B
不包含折旧 & 摊销的销货成本	915.26M	1.42B	1.96B	3.47B	4.82B
折旧 & 摊销费用	136.37M	148.86M	270.16M	712.63M	1.04B
折旧	112.08M	128.16M	218.71M	368.34M	581.63M
无形资产摊销	24.3M	20.7M	51.45M	344.29M	460.87M
总收入	2.09B	3.93B	6.34B	8.12B	10.04B

注：数据来自企业公开数据。M 代表百万美元，B 代表 10 亿美元。