



智能时代

吴军/著

大数据与智能革命重新定义

序一

大数据与机器智能催生智能时代

大数据是当今信息社会的热词。关于数据，狭义上，在计算机科学中，数据是指所有能输入计算机并被计算机程序处理的符号介质的总称，是用于输入电子计算机进行处理的具有一定意义的数字、字母、符号和模拟量等的通称。广义上的数据，按照维基百科的定义，则是以适于更好使用或处理的方式来表示或编码的信息或知识，它可以被测量、收集和报告及分析，能够使用图形或图像来显示。按照上述定义，数据是伴随人类社会而出现的，从狭义的计算机数据的角度来看，数据从有计算机算起到现在也有70年历史了，从摩尔定律的提出到现在也有50年了。这几十年来，全球数据量按每年平均40%的速度增长。由摩尔定律所驱动的计

还在督促自己每天进步一点吗？

还在坚持每天阅读的习惯吗？

还在为找不到自己喜欢的书籍烦恼吗？

那～

你愿意与我成为书友吗？

国内外当下流行书籍

各图书销量排行榜书籍

大量工具书籍

使我们受益终生的书籍

.....

海量电子版、纸质版书籍及音频课程

还有贴心的“学习管家”服务哦！

微信：shuyou055





吴军/著

大数据与智能革命重新定义

序一

大数据与机器智能催生智能时代

大数据是当今信息社会的热词。关于数据，狭义上，在计算机科学中，数据是指所有能输入计算机并被计算机程序处理的符号介质的总称，是用于输入电子计算机进行处理的具有一定意义的数字、字母、符号和模拟量等的通称。广义上的数据，按照维基百科的定义，则是以适于更好使用或处理的方式来表示或编码的信息或知识，它可以被测量、收集和报告及分析，能够使用图形或图像来显示。按照上述定义，数据是伴随人类社会而出现的，从狭义的计算机数据的角度来看，数据从有计算机算起到现在也有70年历史了，从摩尔定律的提出到现在也有50年了。这几十年来，全球数据量按每年平均40%的速度增长，由摩尔定律所驱动的计算机处理能力也在持续增长，现在每年新

增的数据量与计算机处理能力都是以前无法相比的，但数据量与计算机处理能力之比并没有因为年份而有数量级的大变化。问题是为什么现在才出现大数据热呢？

吴军先生的《智能时代》一书给出了答案。该书回顾了科学研究发展的四个范式，即描述自然现象的实验科学、以牛顿定律和麦克斯韦方程等为代表的理论科学、模拟复杂现象的计算科学和今天的数据密集型科学。即便在实验科学和理论科学及计算科学范式时期，数据仍然起了重要作用。作者在介绍科学发展史时用实例说明了数据在科学发现中的位置，在牛顿和麦克斯韦时代，他们所导出的简洁的公式给出的确定性的规律是由大量观察数据所验证的。现在我们面对的是更复杂的自然和社会现象，多维度和多变量导致很大的不确定性，虽然还不能用解析式来说明因果关系，但如果从足够多的数据中发现相关性也能把握事物发展的轨迹，这就是数据密集型科学产生的背景。大数据的应用缘于需求，更是得益于技术的发展：互联网的宽带化和移动互联网及物联网的技术与应用源源不断产生数据，摩尔定律所支撑的计算能力几乎是以十年千倍的速度提升，云计算的集约化运用模式降低了信息化的成本，更重要的是机器智能的发展。计算机的计算与存储能力是人远远不及的，唯一不足是智能，但人的智能也不是与生俱来，只是学习的结果。机器智能可以通过深度学习得到，从而将大数据挖掘问题转化为可计算问题来处理。大数据挖掘的需求加速了机器智能技术的成熟，可以说，大数据与机器智能相伴而生，促进物联网从感知到认知并智能决策的升华，催生了智能化时代。这是一个计算无所不在、软件定义一切、数据驱动发展的新时代。相比以蒸汽机的发明为标志以机械化为特征的第一次工业革命，以电的发明为标志以电气化为特征的第二次工业革命，现在以大数据应用为标志之一和以智能化为特征的新一轮产业革命到来了，它对人类文明和社会进步及经济发展的影响将不亚于前两次工业革命。

读吴军先生的《智能时代》和同样出自其手的《数学之美》和《文明之光》，

我感受到作者深厚的数学与物理功底。他对科学发展史研究情有独钟，见解深刻，以历史的眼光引导读者认识现代科技的发展趋势。他的书深入而浅出，既专业又通俗。《智能时代》一书与前两本书相比更关注产业变革，从工业革命谈起，顺理成章导出大数据与智能化，并积极评价了大数据与机器智能对社会与产业发展的贡献，同时根据

历史经验分析了 智能时代可能产生的负面影响，指出技 术时代的变迁总是会引起现有产业格局 的重大调整，要抓住智能时代的机遇并 认真对待挑战，力争在新_轮产业变革 浪潮中占领先机。作者过去在谷歌和腾 讯公司的工作经历和多年从事大数据与 机器智能的研究实践，反映到《智能时 代》一书中对相关技术的准确把握。但作 者并没有将笔墨的重点放在对技术的深 入解读上，而是着眼从技术的应用中体 现大数据的理念，聚焦于启迪创新思维。 综观全书，这是一部近代科技的历史书， 也是一部科普书，也可以说是一部指导 创新的教科书。由于大数据的应用必然 会渗透到所有的领域，因此本书不仅值 得IT行业科技人员一读，对关注信息化 应用的其他行业的科技人员和管理人员 来说也必定开卷有益。

中国工程院院士

邬贺铨 2016年7月26日

智能时代，未来已来

最近几年，人类在一些科技前沿领域取 得了重大的突破，这些领域包 括:人工智 能、基因技术、纳米技术等。过去一年, 我们看到了许多存在于 科幻小说中的内 容成为现实：人工智能击败了人类顶尖 棋手,自动驾 驶汽车技术日趋成熟，生产 线上大批量的机器人取代工人.....甚至 在 我们有生之年，也许可以期待看到星 际航行技术的成熟。当这些曾经 是对人 类社会“未来”描述的事情一件件成真， 或许我们可以说，已经 初露端倪的“智能 时代”就是人类想象中“未来”的样子。

《智能时代》这本书展现了吴军博士 的真知灼见和前瞻思维，这些都 来自于

他在大数据和机器智能领域的多年第一 线实践经验。全书对大数据与 智能革命 带来的思维革命、技术上的挑战，以及机 器智能如何改变人 类社会，都做了全面 的讲解。与其他一些写机器智能的书不 同，这本书与作者之前的几本书一样，维 持了作者对科学生动而易于理解的、 有 温度感的一贯的表述方式。

大数据是解决不确定性的良药

“用不确定的眼光看待世界，再用信 息来消除这种不确定性”，是大数据

解决智能问题的本质。吴军博士在书中提到了世界的不确定性来自两个方面，一是影响世界的变量太多以至于无法用数学模型来描述；二是来自客观世界本身：不确定性是我们所在宇宙的特性。因此，用

机械论已经完全无法对未来进行预测。

香农，这位不世出的天才，则通过借用热力学中“熵”的概念，引入“信息熵”，用信息论将世界的不确定性与信息联系在一起。这个建立在不确定性上的理论，正是今天人类研究大数据与机器智能的基石。

解决智能问题，就是将问题转化为消除不确定性的问题，大数据则是解决不确定性问题的良药。可以预见，在这里会诞生无数的机会。

现有产业+新技术=新产业

吴军博士在书中总结了从第一次工业革命以来历次技术革命中的一个规律，即每一次技术革命都会围绕着一个核心技术展开，第一次工业革命是蒸汽机，第二次工业革命是电，信息革命是计算机和半导体芯片，当下的智能革命则是大数据和机器智能。而在每一次技术革命中，只有率先采用新技术，才能立于不败之地。在智能革命中，现有产业采用了新技术后，将会全面升级，成为新产业，这将给我们带来无限的机会。智能革命带来前所未有的不连续性挑战

本书的一个重要观点是：机器智能革命的发生来自大数据量的积累达到质变的奇点。从这个角度来看，机器的学习与人类的学习并没有什么本质的不同。几千年以来，我们人类的知识都建立在归纳法之上，归纳法隐含的假设是“未来将继续和过去一样”，换句话说应该叫连续性假设。但即将到来的这个“智能时代”，可以说人类将遭遇前所未有的“不连续性”。如何在新的时代里生存，跨越底层认知的不连续性，是前进的第一步。

与工业革命相比，人工智能带来的革命程度将更深更广。书中也提到，一些人对变化开始有了一定程度的担心，认为机器智能将在未来危及整个人类的工作机会，大多数人在未来将不再被社会需要。不可避免，每一次大的技术革命都会带来阵痛，但同时诞生的，还有更多新的机会。而要想在智能时代取得胜利，成为“2%的人”，我们需要做的第一步，是打破现有的认知束缚。

如何在智能时代开始跨越思维的不连续性?寻找答案,此书也许是最恰当的一本。

李善友

混沌大学创始人

人类的胜利

AlphaGo在第一盘出人意料地轻松获胜。当然,大部分人在赞誉AlphaGo水平的同时,依然认为这可能是李世石在试探计算机而已,毕竟那是五盘棋的比赛,用一盘棋试探自己毫不了解对手未尝不是明智之举。但是当AlphaGo在第二盘获得连胜并且下出了很多人类意想不到的好棋后,对机器智能持怀疑态度的聂卫平等人都对它产生了敬意。在AlphaGo获得第三盘胜利之后,很多超一流的棋手都渴望和它一战,希望以此检验自己的水平,并且能够提高技艺。虽然李世石在第四盘抓住AlphaGo的一个失误打了一个漂亮的翻身仗,但是

AlphaGo在最后一盘稳稳地控制着局面,直到胜利。可以讲在那一次人机大战之后,围棋界对机器智能从怀疑变成了顶礼膜拜,大家都意识到,按照AlphaGo在过去几个月里的进步速度,只要Google愿意继续进行科研,很快人类所有的围棋高手都无法和它过招了。

计算机之所以能战胜人类,是因为机器获得智能的方式和人类不同,它不是靠逻辑推理,而是靠大数据和智能算法。在数据方面,Google使用了几十万盘围棋高手之间对弈的数据来训练AlphaGo,这是它获得所谓的“智能”的原因。在计算方面,Google采用了上万台服务器来训练AlphaGo下棋的模型,并且让不同版本的AlphaGo相互对弈了上千万盘,这才保证它能做到“算无遗策”。具体到下棋的策略,AlphaGo里面有两个关键的技术。第一个关键技术是把棋盘上当前的状态变成一个获胜概率的数学模型,这个模型里面没有任何人工的规则,而是完全靠前面所说的数据训练出来的。第二个关键技术是启发式搜索算法——蒙特卡罗树搜索算法(Monte Carlo Tree Search),它能将搜索的空间限制在非常有限的范围内,保证计算机能够快速找到好的下法。虽然AlphaGo的训练使用了上万台服务器,但是它在和李世石对弈时仅仅用了几十台服务器(1000多个CPU的内核以及100多个GPU)。相比国际象棋,围棋的搜索空间要大很多倍,AlphaGo的计算能力相比深蓝,其实并没有这么多倍的提高,它靠的是好的搜索算法,能够准确地聚

焦搜索空间，因此能够在很短的时间里算出最佳行棋步骤。由此可见，下围棋这个看似智能型的问题，从本质上讲，是一个大数据和算法的问题。

当然，Google开发AlphaGo的最终目的，并非要证明计算机下棋比人强，而是要开发一种机器学习的工具，让计算机能够解决智能型问题。AlphaGo和李世石对弈，实际上是对当今机器智能水平的一个测试。从樊麾到李世石，他们实际上是用自己的专才在帮助Google测试当今机器智能的发展水平。在人机对弈的第四盘李世石反败为胜的过程中，他无意中发现了AlphaGo的一个缺陷。因此，Google的成功里面也有李世石等棋手的功劳。从这个角度来讲，AlphaGo的胜利标志着人类在机器智能方面达到了一个崭新的水平，因此它是人类的胜利。

AlphaGo无论是在训练模型时，还是在下棋时所采用的算法都是几十年前大家就已经知道的机器学习和博弈树搜索算法，Google所做的工作是让这些算法能够在上万台甚至上百万台服务器上并行运行，这就使得计算机解决智能问题的能力有了本质的提高。这些算法并非专门针对下棋而设计，其中很多已经在其他智能应用的领域(比如语音识别、机器翻译、图像识别和大数据医疗)获得了成功。AlphaGo成功的意义不仅在于它标志着机器智能的水平达到了一个新的台阶，还在于计算机可以解决更多的智能问题。今天，计算机已经开始完成很多过去必须用人的智力才能够完成的任务，比如：医疗诊断，阅读和处理文件，自动回答问题，撰写新闻稿，驾驶汽车，等等。可以讲，AlphaGo的获胜，宣告了机器智能时代的到来。

AlphaGo的获胜让一些不了解机器智能的人开始杞人忧天，担心机器在未来能够控制人类。这种担心是不必要的，因为AlphaGo的灵魂是计算机科学家为它编写的程序。机器不会控制人类，但是制造智能机器的人可以。而科技在人类进步中总是扮演着最活跃最革命的角色，它的发展是无法阻止的，我们能做的就是面对现实，抓住智能革命的机遇，而不是回避它、否定它和阻止它。未来的社会，属于那些具有创意的人，包括计算机科学家，而不属于掌握某种技能做重复性工作的人。

在AlphaGo取得人机大战胜利之际，我们出版这本书，希望能让大家更多地了解大数据的本质、它的作用、它和机器智能的关系、机器智能的原理和发展历程，以及它们对未来产业和社会的影响。本书一共分为七章，分别介绍了数据的作用，大数据和机器智能，机器智能的原

理及其发展历程，大数据思维的核心 及其重要性，大数据和机器智能与商业 的关系，它们对社会正反两个方面的巨 大影响。书中的核心内容来自我在研习 社和一些大学商学院讲课的讲义，但是 考虑到大家读书和听课毕竟有很大的区 别，因此在将讲义改写成书的时候，我在 书中增加了大量的案例和历史背景介 绍，以方便大家能够系统地了解大数据 和机器智能的来龙去脉，以及我们对未 来进行分析的依据。

本书的出版，在很大程度上是研习 社负责人曾兴晔女士、空无边处出版团 队的张娴和郑淳女士，以及中信出版社 经管分社的朱虹社长和赵辉编辑等相关 人员积极推动的结果。著名的信息领域 专家、中国互联网协会理事长邬贺铨院 士，以及混沌学院创始人李善友教授，在 百忙中为本书写了序言。上海交通大学 电子信息与电气工程学院副院长王延峰 副教授对本书的内容提供了宝贵的参考 意见。在此我对他们表示衷心的感谢。由 于本人水平有限，书中不免有这样或者 那样的错误，希望广大读者朋友不吝赐 教指正。

吴军

2016年4月25日于硅谷

第一章

数据——人类建造文明的基石

如果我们把资本和机械动能作为大 航海时代以来全球近代化的推动力 的话，那么数据将成为下一次技术革 命和社会变革的核心动力。

- [现象、数据、信息和知识](#)
- [数据的作用：文明的基石](#)
- [相关性：使用数据的钥匙](#)
- [统计学：点石成金的魔棒](#)
- [数学模型：数据驱动方法 的基础](#)
- [什么是机器智能](#)
- [鸟飞派：人工智能1.0](#)
- [另辟蹊径：统计+数据](#)
- [数据创造奇迹：量变到质 变](#)

- [大数据的特征](#)
- [变智能问题为数据问题](#)
- [思维方式决定科学成就：从欧几里得、托勒密到牛顿](#)
- [工业革命，机械思维的结果](#)
- [大数据的本质](#)
- [从因果关系到强相关关系](#)
- [从大数据中找规律](#)
- [巨大的商业利好:相关性、时效性和个性化的重要性](#)
- [把控每一个细节](#)
- [重新认识穷举法 完备](#)
- [从历史经验看大数据的作用](#)
- [技术改变商业模式](#)
- [加 \(+\)大数据缔造新产业](#)
- [技术的拐点](#)
- [数据收集：看似简单的难题](#)
- [数据存储的压力和数据表示的难题](#)
- [并行计算和实时处理：并非增加机器那么简单](#)
- [数据挖掘：机器智能的关键](#)
- [数据安全的技术](#)
- [保护隐私：靠大数据长期挣钱的必要条件](#)
- [未来的农业](#)
- [未来的体育](#)
- [未来的制造业](#)
- [未来的医疗](#)
- [未来的律师业](#)
- [未来的记者和编辑](#)
- [智能化社会](#)
- [精细化社会](#)
- [无隐私的社会](#)
- [机器抢掉人的饭碗](#)
- [争当2%的人](#)

现象、数据、信息和知识

在很多人的印象中，数据就是数字，或者必须是由数字构成的，其实不然，数据的范畴比数字要大得多。互联网上的任何内容，比如文字、图片和视频都是数据；医院里包括医学影像在内的所有档案也是数据；公司和工厂里的各种设计图纸也是数据；出土文物上的文字、图示，甚至它们的尺寸、材料，也都是数据；甚至宇宙在形成过程中也留下了许多数据，比如宇宙中的基本粒子数量。

虽然数据本身是客观存在的，但是它的范畴是随着文明的进程不断变化和扩大的。在计算机出现之前，一般书籍上的文字内容并不被看成是数据，而今天，这种以语言和文字形式存在的内容是全世界各种信息处理中最重要的数据，也是全世界通信领域和信息科技产业的核心数据——包括我们的信件、电话和电子邮件内容、电视和广播节目、互联网网页，以及各种社交产品中由用户产生的内容（User Generated Content,简称 UGC）。这些数据的共同特点是以语音和文字为载体。因此，研究人员为了更好地研究和处理它们，还建立了专门针对语音和文字的数据库，即所谓的语料库 (Corpus)。在语料库中，数据主要是语音和文字的内容，反而没有多少数字的内容。



图1.1互联网上的数据

将数据的外延再扩大，那些医学影像资料、工业中的各种设计图纸都可以被划分为数据。事实上它们已经是今天大数据处理的对象了。我们人类的活动本身，也可以被看成是一种特殊的数据，比如我们玩游戏的行为、我们的社会关系、我们每天的活动等。可以想象，我们的下一代所谈论的数据，一定比今天的范围更广泛。可以说，数据是文明的基石，人类对它的认识也反映了文明的程度。

在今天，谈论数据时，人们常常把它和信息的概念混同起来，比如人们在今天谈论数据处理和信息处理时，其实想要表达的意思相差不大。然而严格地讲，数据和信息虽然有相通之处，但还是不同的。

信息是关于世界、人和事的描述，它比数据来得抽象。信息既可以是人类创造的，比如两个人的语音通话记录，也可以是天然存在的客观事实，比如地球的面积和质量。不过信息有时藏在事物的背后，需要挖掘和测量才能得到，比如宇宙大爆炸时留下的证据——3K背景辐射 M]、物理学定律中的参数、日月星辰运行的周期等。在西方很多物理学家看来，上帝在创造这个宇宙时，将很多信息埋藏在了黑暗之中，他们的工作就是找到这些信息，并且用数据把它们描述清楚。因此，在这种前提下，将信息和数据混为一谈倒也无害。

不过，数据和信息还是稍有不同，虽然它最大的作用在于承载信息，但是并非所有的数据都承载了有意义的信息。数据本身是人造物，因此它们可以被随意制造，甚至可以被伪造。没有信息的数据通常没有太大意义，人们也不太关心，因此这些数据不是本书想要讨论的重点。伪造出的数据则有副作用，比如我在《数学之美》中不断提到的为了优化网页搜索排名而人为制造出来的各种作弊数据。另外，我们还需要强调，那些有用的数据、毫无意义的数据和伪造的数据常常是混在一起的，后面两种数据无疑会干扰我们从数据中获取有用的信息，因此如何处理数据，过滤掉没有用的噪声和删除有害的数据，从而获取数据背后的信息，就成为技术甚至是一种艺术。只有善用数据，我们才能够得到意想不到的惊喜，即数据背后的信息。我们不妨看一个如何通过数据得到信息的例子。

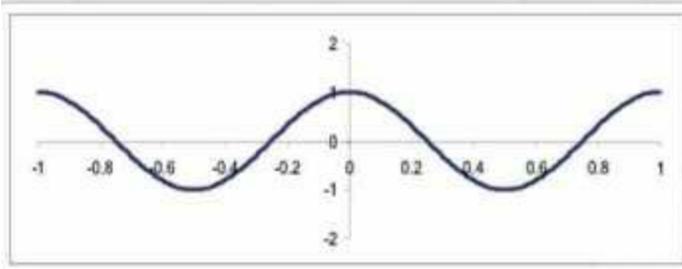


图1.2信息(上图) 埋藏在噪声(下图) 中

在距今4500多年前的公元前26世纪，古埃及人已经掌握了很多数学知识，他们在建造胡夫大金字塔时，将这样的信息通过数据告诉了我们。比如，大金字塔的周长和高度的比值大约为6.29。这大致是圆的周长和半径的比例，即两倍的圆周率($2\pi r$)，误差在千分之_左右。当然，大金字塔留下的最有意义的数字可能是法老墓室的尺寸。它有20埃及古尺[2]长，10埃及古尺宽，比例正好是2:1，但是高度为11.18埃及古尺，这并不是个整数。为什么法老要选用这样一个奇怪的数字呢？因为11.18正好是

也就是墓室宽度的

$\sqrt{2}$ 倍，这个高度保证了两面墙的

对角线长度是个整数——15埃及古尺，因为根据勾股定理[3]:

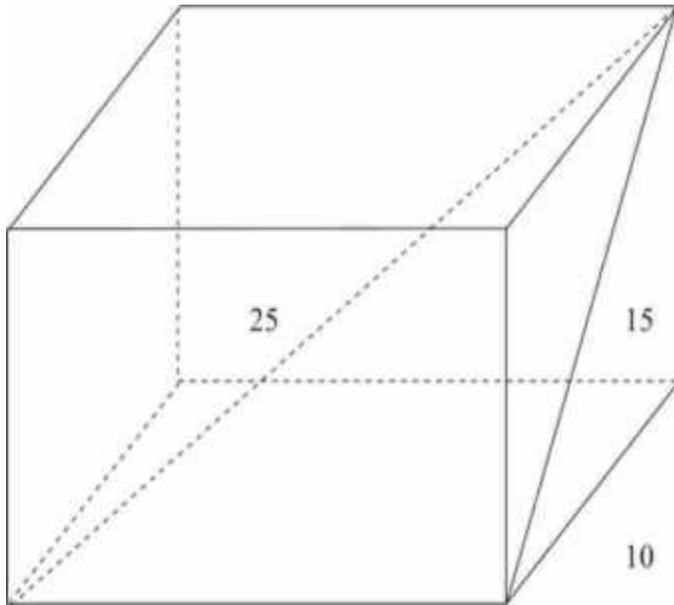
$$10^2 + (5\sqrt{2})^2 = 225 = 15^2$$

不仅如此，墓室的两个最远的顶点之间的距离也是整数，即25埃及古尺，因为同样根据勾股定理： $15^2 + 20^2 = 625 = 25^2$ 。

从这个墓室的尺寸，我们分析出4600年前的古埃及人已经知道了勾股定理，进而可以知道那个时期古埃及文明大致发展到了什么水平，这就是从数据得到信息很好的例子。

数据中隐藏的信息和知识是客观存在的，但是只有具有相关领域专业知识的人才能将它们挖掘出来。比如大金字塔的这组数据，如果让一个盗墓者看到，他可能联想不到任何事情，但是在数学家或者考古学家眼里却意义重大，因为前者不具备后者所拥有的数据处理能力。处理信息和数据可以说是人类所特有的本事，而这个本事的大小和现代

智 人的社会发展有关。今天我们还能找到 这样的原始部落，他们对数字的认识只 有1、2、少量和很多一共四种衡量方式， 但是随着人类的进步以及处理数据和信 息的能力不断增强，人类从数据中获取 有用信息 的本事就越来越大，这就是今 天所说的大数据应用的基础。



20

图1.3大金字塔墓室的尺寸示意图

对数据和信息进行处理后，人类就 可以获得知识。知识比信息更高一个层 次，也更加抽象，它具有系统性的特征。 比如通过测量星球的位置和对应的时 间，就得到数据;通过这些数据得到星球 运动的轨迹，就 是信息;通过信息总结出

开普勒三定律，就是知识。人类的进步就 是靠使用知识不断地改变我们的生活和 周围的世界，而数据是知识的基础。在下 一节里我们不妨 看看人类是如何利用数 据改变世界的。

数据的作用：文明的基石

早期人类得到的数据是从哪里来的？其中一个重要的来源是对现象的观察。从观察中总结出数据，是人类和动物的重要区别，后者虽具有观察能力，却无法总结出数据，但是人类有这个能力。而得到数据和使用数据的能力，是衡量文明发展水平的标准之一。

我们的文明从一开始就伴随着对数据的使用，可以说数据是文明的基石。人类最初希望了解到的是周围的世界，这样可以更好地生活。早在埃及法老们开始修建金字塔的几千年之前，闪米特人[4]和当地的土著就在尼罗河畔辛勤耕耘了。为什么他们会选择在那个地方定居呢？除了气候温暖之外，最重要的原因是每年尼罗河都会发洪水，洪水退去之后留下大片肥沃的土地供他们耕耘收获。为了准确预测洪水到来和退去的时间，以及洪水的大小[5]，当时的埃及人开始观察天象，并且在观察数据的基础上开创了天文学。他们根据天狼星和太阳同时出现的位置来判断一年中农耕的时间和节气，然后准确地判断洪水可能到达的边界和时间。古埃及人观察到一年的时间不是正好365天，而是多了一点，但在古埃及的历法中又没有闰年，于是他们用了—一个非常长的“季度”长达 $365 \times 4 + 1 = 1461$ 天，因为每隔这么多天，太阳和天狼星就一起升起。事实证明，以天狼星和太阳同时出现作为参照系比以太阳作为参照系更准确些。这实际上也说明了好的模型要和数据相吻合的道理，因此古埃及人已经有了从数据中总

结数学模型的基本能力。

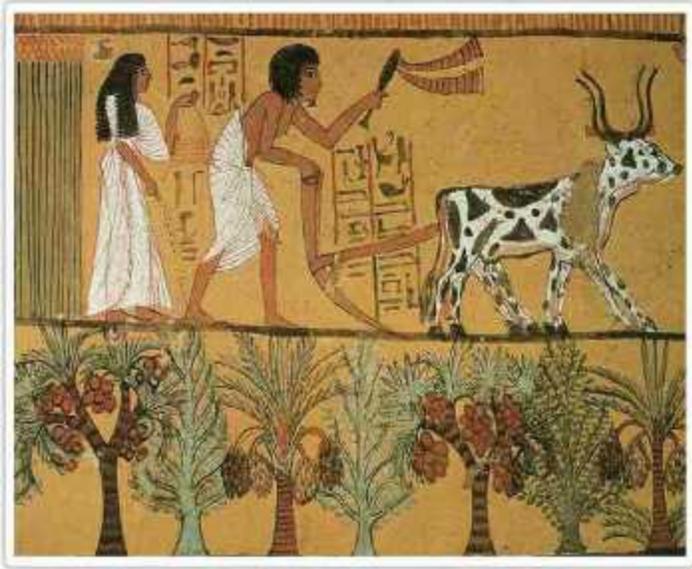


图1.4古埃及人为了农业的收成而发展起天文学

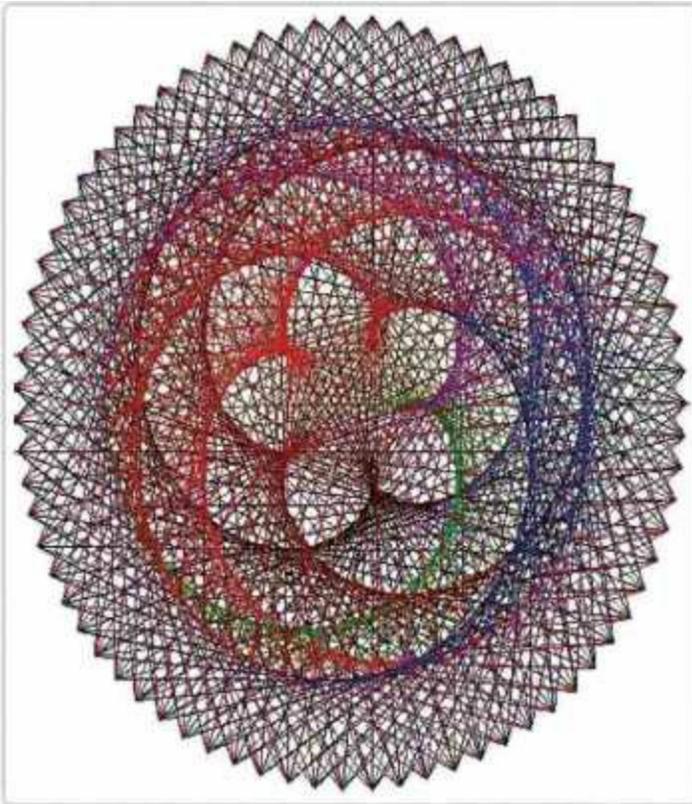


图1.5从地球上看到的金星运行轨迹

透过上述天文学的起源和发展历程，我们可以清晰地了解到数据在人

类 发展过程中所产生的巨大作用。人类另 一个古老的文明中心是美索不达米亚[6]平原，那里的苏美尔人对天文学有了进 一步的发展，他们根据观察发现月亮每 隔28~29天就完成从新月到满月再回到 新月的周期。他们同时观察到每年有四季之分，每过12~13个月亮的周期，太阳就回到原来的位置，这样他们就发明 了太阴历，历法实际上就是对天文现象 的一个数据化描述。苏美尔人还观测到 了五大行星（金、木、水、火、土，因为 肉眼看不到天王星和海王星）运行的轨 迹不是简单地围绕地球转，而是波浪形 的。西方语言中行星（planet）—词的意思就是漂移的星球。他们还观测到行星 在近日点运动比远日点快，以及金星大 约每4年在天上画一个五角星，他们记录 了这些信息。在美索不达米亚文明中，当地的数学家一直试图利用他们所获得的天文观测数据建立起我们今天所说的数 学模型，来完成从数据到知识的过程。利用这些模型，美索不达米亚人能够计算 出月亮和五大行星的运行周期，并且能 够预测日食和月食。

从这些例子可以看出，人类的文明 过程其实伴随着如图1.6所示的这样一个过程：

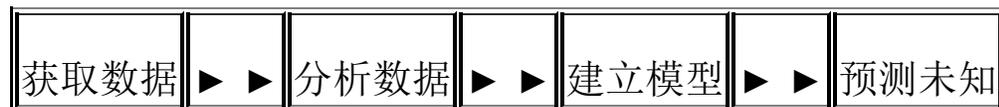


图1.6使用数据的标准流程

由此可见，数据在人类文明中起到了基石的作用。

到了古希腊文明时期，地中海沿岸 的学者们学习继承了美索不达米亚文明 的成果。公元前551年，古希腊科学和哲 学的集大成者毕达哥拉斯来到米利都 (Miletus)[7]、得洛斯 (Delos)[8]等地，拜访了当时著名的数学家和天 文学家泰勒斯 (Thales of Miletus, 前 624—前546)[9]、阿那克西曼德 (Anaximander,前611—前546) [10]和菲尔库德斯 (Pherecydes of Syros,生平不详) [11]等人，并成了 他们的学生，把美索不达米亚的数学和 天文学成就带回了古希腊地区。在这之 后，古希腊成了全世界数学和天文学研 究的中心。后来柏拉图的学生欧克多索 (Eudoxus of Cnidus,前408—前347) 建立了地心说的早期模型，阿基米德 (阿_3此，前287年—前212年) 贝隼 立了日心说模型的原型。而最终利用数 据建立起描述天体运动模型的是著名天 文学家托勒密。

托勒密的伟大之处在于用40~60个 小圆套大圆的方法，精确地计算出了

所有行星运动的轨迹，如图1.7所示。托勒密继承了毕达哥拉斯的一些思想，他也认为圆是最完美的几何图形，因此，所有天体均以匀速度按完全圆形的轨道旋转。事实上，后来日心说的提出者哥白尼也坚持认为天体运动的模型必须符合毕达哥拉斯的思想。但是实际上天体以变速度按椭圆轨道绕地球以外的中心——太阳——运动。为了维护原来的基本假设，就必须用小圆套在大圆之上的方法解释了。托勒密使用了3种尺寸的圆相互嵌套的模型，即本轮、偏心圆和均轮，这样，他就能对五大行星的轨道给出合理的描述。不过这五大行星的轨道无法用一组圆来统一描述，因此，托勒密用了很多个圆分别描述，互相嵌套的大小

圆多达40~60个。

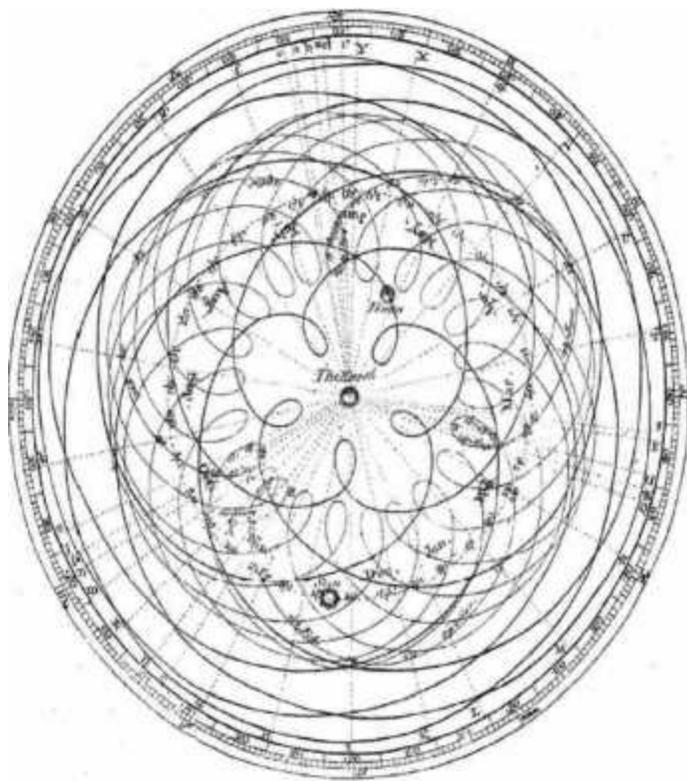


图1.7托勒密用多个圆相互嵌套的模型来描述行星运动

托勒密认为模型必须与观测数据相吻合（这种想法从古埃及开始就有了），

要感谢喜帕恰斯为托勒密留下了很多观测数据，使得他的模型能够建立得很准确。托勒密的追随者宣称托勒密地心说的模型和前面800多年

的观测数据相吻合，这种说法可能有点夸大，今天有证据表明在托勒密的时代，人类可能只记载了一百多年的观测数据。不过即使只能和一百多年的数据相契合，这个模型也很了不起了。托勒密根据自己的模型绘制了一张表，预测了今后某个时候某个星球所在的位置。托勒密模型的精度之高，让后来所有的科学家都惊叹不已。即使今天，在计算机的帮助下，我们也很难解出40个套在一起的圆的方程。每每想到这里，我都由衷地佩服托勒密。托勒密根据计算，制定了关于日月星辰位置的 [(实箱天SiMiHandy Tables)，和当时的儒略历[12]相吻合，即每年365天，每4年增加一个闰年，多一天。其后1500年，人们根据儒略历和《实用天文表》决定农时。但是，经过了1500年后，托勒密对太阳运动的累积误差还是多出了10天。由于这10天的差别，欧洲的农民从事农业生产的日期几乎差了一个节气，很影响农业生产。1582年，教皇格里高利十三世在日历上取消掉10天，然后将每一个世纪最后一年的闰年改成平年，每400年再插回一个闰年，这就是我们今天用的日历，这个日历几乎没有误差。为了纪念格里高利十三世，我们今天 的日历也叫作格里高利日历。



图1.8格里高利的墓碑，左下角那本书代表他的历法

格里高利十三世之所以能“凑出”准确的历法，即每400年比儒略历减去3个闰年，其实也是根据上千年的历史数据。当然，格里高利十三世没有本事修正托勒密的模型，而波兰天文学家哥白尼则从另一个角度看问题，提出了日心说的模型，它的好处是只需要8~10个圆，就能计算出一个行星的运动轨迹。但遗憾的是，哥白尼正确的假设并没有得到比托勒密更好的结果，他的模型误差比托勒密模型的误差要大不少，很

重要的原因是哥白尼缺乏数据。由于早期的日心说模型并不比托勒密的地心说模型更准确，因此不能让人心服口服地接受，日心说要发展，就得更准确地描述行星运动。

完成这一使命的是约翰内斯·开普勒(Johannes Kepler, 1571—1630)。开普勒在所有一流的天文学家中，资质较差，一生中犯了无数低级的错误。但是他有两样别人没有的东西，第一是从他的老师第谷(Tycho Brahe, 1546—1601)手中继承的大量的、在当时最精确的观测数据，第二是运气。开普勒很幸运地发现了行星围绕太阳运转的轨道实际上是椭圆形的，这样不需要用多个小圆套在大圆之上，而只要用一个椭圆就能将星体运动规律描述清楚。开普勒为此提出了三个定律，形式都非常简单，而且非常准确。至于为什么是椭圆的，开普勒也说不清楚，他其实只是碰巧找到了一个模型能够比较好地拟合全部观测数据罢了。在开普勒之后，牛顿提出了万有引力定律，这才彻底解释了为什么天体运动的轨迹是椭圆形的。牛顿还修正了开普勒的椭圆模型，椭圆的焦点从太阳移到了太阳系的重心（两者有微小的差别）。

数据的重要性不仅表现在科学研究中，而且渗透到我们社会生活的方方面面。虽然中国古代不像古希腊和古罗马那样重视自然科学，但是在使用数据上一点也不比西方少。中国的历史从某种意义上讲是通过对数据进行收集、处理和总结而写成的。在中国的远古传说中，有伏羲演八卦的故事。伏羲是中国上古的三皇之一，比我们说的炎、黄二帝还要早得多。也有人说他其实不是一个人，而是代表一个部落，当然这个并不重要。据说他发明了八卦，并且可以通过它推演未来的吉凶。伏羲演八卦准不准，我们这里不做评论，但是这件事说明在远古人们已经懂得把未来的吉凶根据不同的条件（实际上是输入数据）归纳成8种或者64种可能的结果(输出数据)。之所以能够对未来这样分类并且有很多人相信它（虽然我不太相信），是因为很多人认为过去所听到的、看到的事情（也是数据）证明了这么归纳分类的正确性。到了农耕文明的时代，先前的很多生活经验，比如什么时候要开始播种，什么时候可以收获，常常就是从“数据”中总结出来的，只是那时还没有文字或者很多人不识字，大家只能一代代地口口相传。

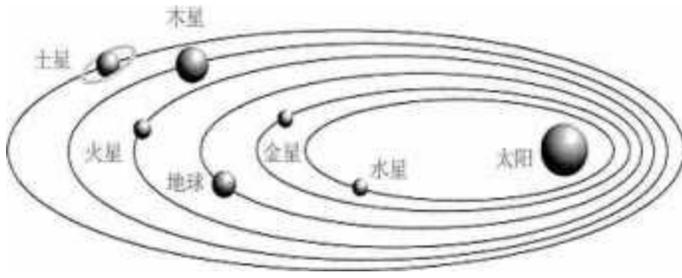


图1.9开普勒的太阳系模型

我们从天文学的发展历程中可以看出，数据的作用自古有之，并非到了今天 大数据时代大家才意识到。但是在过去 数据的作用常常被人们忽视。这里面有两个原因，首先是由于过去数据量不足，积累大量的数据所需要的时间太长，以至于在较短的时间里它的作用不明显。其次，数据和所想获得的信息之间的联系通常是间接的，它要通过不同数据之间的相关性才能体现出来。可以说，相关性是让数据发挥出作用的魔棒。

相关性：使用数据的钥匙

我们不妨通过下面的例子来说明数据相关性的重要性。

20世纪70年代，中国的国际交往开始恢复正常，为了加快中国的建设，中国政府决定向其他国家就一些重大建设项目进行招标，其中一项是大庆油田石油设备。当时大庆油田的情况中国政府对外保密，西方国家了解甚少，甚至连它的具体地点都不知道。但是来自日本的投标却非常有针对性并且一举中标。其背后的原因是，日本人通过1964年中国的《人民画报》上刊登的铁人王进喜的照片，分析出了关于大庆油田的许多细节

在照片中，王进喜穿着厚棉袄，戴着大皮帽，握着钻井机的扳手眺望远方，背景是高高的井架。在一般人看来，这张照片除了体现出石油工人的豪迈之气，并没有什么特别的地方，但是在日本情报人员看来却披露出许多信息。

首先它泄露了大庆油田的位置。根据王进喜穿的厚棉袄和戴的大皮帽，可以断定油田一定是在中国极北的地区，日本人估计油田应该在哈尔滨和齐齐哈尔之间。其次从背景中井架的密度，大致可以估算出油田的产量。最后从王进喜握手柄的方式，大致能推算出油井的直径。由于日本人获得了关于大庆油田相对准确的信息，因此他们提供的设备非常有针对性，中标也就没有悬念了。



图1.10 1964年《人民画报》上刊登的

王进喜的照片

从这个事例中我们可以看出，数据之间常常有我们想象不到的关联性，利用这种关联性，不仅可以获得想要的信息，而且还可能得到意想不到的惊喜。在大数据时代即将到来的时候，一些人敏锐地觉察出了这一点。

2002年年初我到Google面试的时候，面试我的其中一位工程师是阿米

特·帕特尔 (Am it Patel), 他是一位数 学博士, 考了我一些数学问题, 由于我回 答得很快, 所以剩下很多时间聊一些别 的事情。我就问他在 Google里面做些什 么, 通常Google人喜欢故弄玄虚不告诉 你他们工作的细节, 但是帕特尔倒是挺 坦诚。他给我随手画了下面这样一张图。

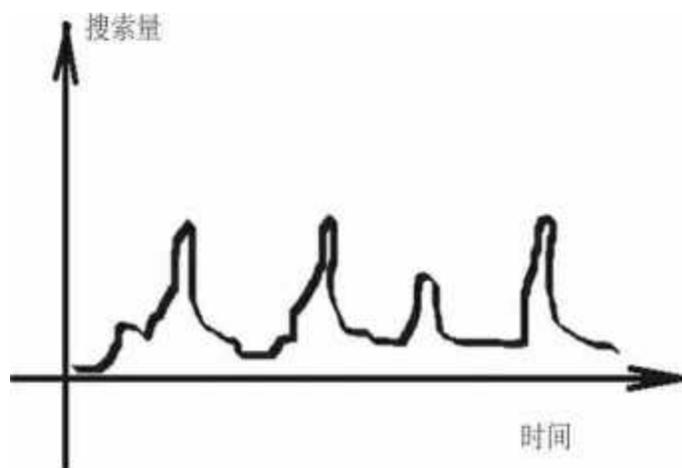


图1.11 Google用户在不同时间点对 某个电视节目的搜索量

他在图中画出的是从Google内部 看到的用户在不同时间点对某个电视节 目的搜索量。帕特尔问我为什么会出现4 个高峰, 我说可能是大家在看 节目的前 后回到Goog | e上搜索这个节目, 至于4 个高峰, 是因为美国 跨了 4个时区, 节目 播出的时间各差一个小时。帕特尔同意 我这个说 法, 他又补充道, 其实通过它以 及各个时区的人口, 可以了解到不同 电 视节目在不同地区 (各个时区) 的收 视 率。这样, 帕特尔就将搜索 量和收 视率联 系起来了。我称赞他这个发现很有意思, 帕特尔感慨 道, 因为这个工作没有太多 经济利益, 因此在公司里无法获得多少 资 源。

几个月后, 我加入了Google, 发现 帕特尔在Google确实不是很受人重 视。 他加入Google很早, 但是人们知道他仅 仅是因为他要求和当时新 来的CEO (首 席执行官)施密特挤一间办公室, 而不是 他所做的工作。 好在Google总是支持每 一个人干自己所喜欢的事情, 因此帕特尔 就在 Google内部一直研究搜索的模式。

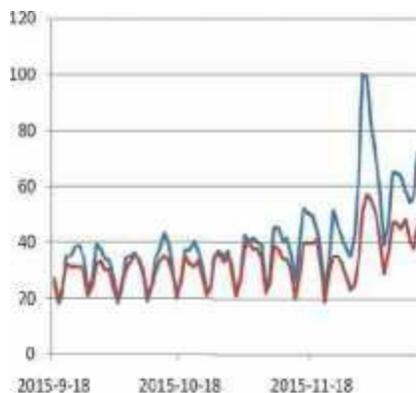
到了2007年, 帕特尔突然在全世界 声名薛起, 因为他的研究成果被几 个工 程师开发成了 Google的一款产 品——Google趋势(Google Trends)。 利用这款产品, 任何人都可以看到全世 界用户在Google上搜 索的关键词随着 时间和地点变化的趋势, 从而知道大家 关注什么事

情。比如在2015年年底的巴黎气候大会期间，全球范围内“气候变化”（climatechange)的搜索量暴增。

当然，如果仅仅是看看搜索趋势的变化，这可能不过是一个小玩具而已。但是，如果把搜索和其他事情关联起来，就能发现非常重要的信息。

2009年，人类发现一种新的流感病毒—甲型H1N1禽流感病毒，短短的一个月内由该病毒导致的疾病就在全球迅

速蔓延开来。这让大家想起了1918年欧洲的大流感，当时有5亿人口受到威胁，并且有5000万~1亿人死亡[13]，因此甲型H1N1禽流感引起了全世界的恐慌。当时还没有研制出对抗这种流感的疫苗，因此公共卫生专家只能先设法知道这种禽流感流行到了哪里，以便防止它的进一步传播。



—气候变化 —全球变暖

图1.12 “气候变化”和“全球变暖”在

Google上搜索量的变化

数据来源：Google Trends导出数据

过去预报疫情传统的方法是由各地医院、诊所和医务人员向美国疾病控制和预防中心（Centers for Disease Control and Prevention,简称CDC)上报。但是这种方法的延时大约有10天至两周，而两周内疫情早已迅速扩散，因此公共卫生专家需要找到新的办法预测和监控疫情。值得庆幸的是，疾病控制和预防中心的科学家和Google的工程师从2007年

到2008年一起合作研究了流 行病传播和各地区搜索量变化的关系， 并且于2009年2月在著名的《自然》杂志 上发表了他们的研究成果[14]——通过 各地区用户在Google上搜索和流感有 关的关键词的趋势变化， 预测流感流行 到什么地方了。Google的工程师们从4.

5亿种关键词的组合中， 最终挑出45个重 要的检索词条和55个次重要词条（归并 成12类）作为特征， 训练了一个线性回归 模型[15]预测2007年和2008年冬季流 感传播的趋势和地点， 并且将机器预测 的结果和疾病控制与预防中心公布的数据进行比对， 发现准确率高达97%以上。

受到这篇论文的启发， 疾病控制与 预防中心在2009年了解禽流感疫情时 采用了同样的方法， 获得了更有效、更及 时的数据。这个案例后来被各种媒体报 道， 成为利用大数据解决医疗问题的经 典案例。在这个例子中， 最关键的是建立 起了数据之间的相关性， 即疾病传播和 该地区搜索关键词变化的关系。

很多时候， 我们无法直接获得信息（比如疫情传播情况）， 但是我们可以将 相关联的信息(比如各地搜索情况)量化， 然后通过数学模型， 间接地得到所要的 信息。而各种数学模型的基础都离不开 概率论和统计学。

统计学：点石成金的魔棒

最初研究概率论的并非数学家，而是一群赌徒和投机者。直到今天，很多研究纯数学的数学家都不把概率论当作数学，而将它看成是一门独立的学科。统计学，有时又被称为数理统计，是建立在概率论基础之上，收集、处理和分析数据，找到数据内在的关联性和规律性的学科。在这里，我们就不详细介绍概率论和统计学了，关于它们在信息技术中的应用，可以参见拙著《数学之美》。不过我们这里要强调统计学中数据采集上的两个要点-量和质。

先讲讲数据量的问题。要想取得准确的统计结果，统计首先要求数据量充足。比如我们想了解电影院的观众年龄分布，以便做市场推广，假定我们把观众群分为15岁及以下、16~25岁、26~40岁和41岁及以上4个人群。要了解每个人群的比例，一个简单的办法就是到电影院门口去问一问那些看电影的人的年龄。比如我们通过调查了解到大约有343人在15岁及以下，459人在16~25岁，386人在26~40岁，而490人在41岁及以上，我们大致可以得出这样的结论：

15岁及以下的观众占20%左右，16~25岁的观众超过四分之一，但不到三成，26~40岁的观众略少于四分之一，41岁及以上的观众最多，大约占到了三

成。

但是，如果我们只在周末的晚上抽样调查了10个人，发现有3个15岁及以下的观众，5个16~25岁的观众，2个26~40岁的观众，我们显然不能说25岁及以下的观众占了八成，而41岁及以上的中年人从来不来电影院这样的结论。我想大部分读者都会同意这样一个观点，在统计样本数量不充分的情况下，统计数字毫无意义。至于需要多少数据统计结果(在我们这个问题里是概率的估计)才是准确的，这就需要进行定量分析了。

越想要得到准确的统计结果，需要的统计数据量就越大。在上面的例子中，统计的样本总数是1678人，要得出大致结论是足够了，但是如果我们一定要说“41岁及以上的观众就是29.2%”，或者“15岁及以下的观众一定超过20%”那样非常确定的话，大家就可能会挑战

这个结论了，因为统计是有随机性的，也是有误差的，仅仅上千人的数据得不到这样准确的结论。

统计除了要求数据量必须充分以外，还要求采样的数据具有代表性。有些时候不是数据量足够大，统计结果就一定准确。统计所使用的数据必须和我们想统计的目标相一致。为了说明这一点，让我们来看一个大量统计却没有得到准确估计的案例。

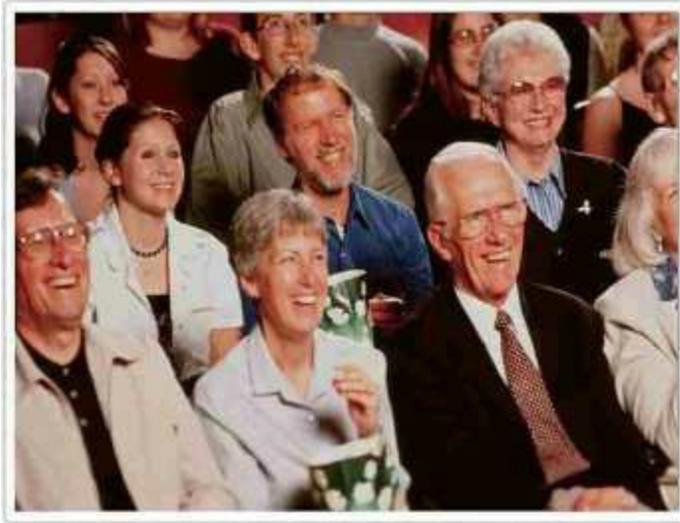


图1.13这场电影显然中老年观众偏多，如果统计量不够，得到的结论未必反映真实情况

在1936年的美国总统大选前夕，当时著名的民意调查机构《文学文摘》

(The Literary Digest)预测共和党候选人兰登会赢。此前，《文学文摘》已经连续4次成功地预测了总统大选的结果，这一次它收回来240万份问卷，比前几次多得多，统计量应该是足够了，因此民众们相信其预测。不过，当时一位名不见经传的新闻学教授（也是统计学家）乔治■盖洛普（George Gallup, 1901—1984）却对大选结果提出了相反的看法，他通过对5万人意见的统计，得出了民主党候选人罗斯福会连任的结论。后来的大选结果证实是采用少量样本的盖洛普对了。面对迷惑的民众，盖洛普解释了其中的原因：《文学文摘》统计的样本数虽然多，但是不具有代表性，它的调查员们是根据电话本上的地址发送问卷的，而当年美国只有一半的家庭安装了电话，这些家庭的收入相对偏高——他们大多支持共和党。而盖洛普在设计统计样本时，考虑到

了美国 选民种族、性别、年龄和收入等各种因素， 因此虽然只有5万个样本， 却更有代表性。这个例子说明统计样本代表性的重要性。

* Weekly poll ^•ucorTsm 、 一一

Iatliluti' Foivw^h llw H(H*Icflion of Franklin D. Kooscvrll, (iivc; * Him 51% of Popubr Vote, Mininuun of 315 EKrlorw



图1.14 1936年盖洛普正确地进行了 总统大选结果的预测

在盖洛普之后， 各种民意调查和统计公司都试图设计出具有代表性的样本， 以使用相对少的数据精确地统计出 所想知道的结论， 然而是否做好了， 没有人知道。有时人们甚至根据结论倒推当初的样本设计， 结论准确了， 就说当初的 样本假设是没有问题的， 否则就说样本 没有设计好。这其实是马后炮， 但是在大数据出现之前， 这个问题难以解决。

我们不妨依然用盖洛普的例子来说 明样本设计之难。在1936年成功地预测 了大选结果之后， 盖洛普不仅个人一夜 成名， 而且还催生出一个直到今天仍具 权威性的民调公司——盖洛普公司。在 这之后， 该公司又成功地预测了 1940年 和1944年两次大选。在1948年年底美 国大选前夕， 盖洛普公布了一个自认为 颇为准确的结论——共和党候选人杜威将在大选中以比较大的优势击败当时的 总统、 民主党候选人杜鲁门。由于盖洛普 公司前三次的成功， 在大选前很多人， 包 括蒋介石， 都相信这个结论。但是， 大选 的结果大家都清楚， 杜鲁门以比较大的 优势获胜。这不仅让很多人大跌眼镜， 而 且让大家对盖洛普公司的民调方法产生 了质疑——虽然盖洛普公司考虑了选民的 收入、 性别、 种族和

年龄的因素，但是 还有非常多的其他因素，以及这些因素 的组合他们没有考虑。

迷信了 1948年盖洛普预测的第一 大输家无疑是远在大洋彼岸的蒋介石先生。他本来就和杜鲁门关系不是很好，在 得知杜威将战胜杜鲁门这个预测后，非 常兴奋，公开支持杜威，并且期待着在杜 鲁门下台后能从美国得到更多的援助。 草根出身的杜鲁门本来就非常不喜欢蒋介石的独裁和腐败，对这次蒋介石公开 支持他的竞争对手的行为更是大为不 满，因此他在连任总统后，对蒋更加不待 见了。当然这是题外话，不过这至少说明， 使用不具有代表性的数据得到的结论可 能要“坑死人”。

在互联网出现之前，获得大量的具 有代表性的数据其实并非一件容易事， 在误差允许的范围内做一些统计当然没 有问题，但是只有在很少的情况下能够 单纯依靠数据来解决复杂的问题。因此 在20世纪90年代之前，整个社会对数据 并不是很看重。

数学模型：数据驱动方法的基础

在上面统计电影观众分布的例子中，我们大致可以估计出4个年龄组观众的人数分布情况。现在的问题是这个估计是否可信，因为毕竟抽样有很大的随机性。从概率论一诞生人们就有这种担忧，人们希望能够从理论上证明当观察到的数据量足够多了以后，随机性和噪声的影响可以忽略不计。19世纪的俄国数学家切比雪夫（Chebyshev，1821—1894）对这个问题给出了肯定的回答。他给出了这样一个不等式，也称作切比雪夫不等式：

$$\sigma^2$$

$$ne^2$$

其中 x 是一个随机变量， μ 是该变量的数学期望值， n 是实验次数（或者是样本数）， f 是误差， σ^2 是方差。这个公式的含义是，当样本数足够多时， n 个随机变量（比如观察到的各个年龄段观众的比例）和它的数学期望值（比如真实情况下所有看电影的观众中不同年龄段的比例）之间的误差可以任意小（小于不等式右边的数值）。

将切比雪夫不等式应用到我们这个例子中，我们大致可以计算出4个年龄组的人分别占到观众人数的20%、27%、24%和29%左右，误差小于5%（在统计中也称为置信度大于95%）。但是如果我们要将4个年龄段观众的准确率提高到小数点后一位数，那么我们大约需要10倍的数据，即两万个左右样本。

用抽样数据来估计一个概率分布是一类非常简单的问题，用统计数据做一做加减乘除即可。但是在大多数复杂的应用中，需要通过数据建立起一个数学模型，以便在实际应用中使用。要建立数学模型就要解决两个问题，首先是采用什么样的模型，其次是模型的参数是多少。

模型的选择不是一件容易的事情，通常简单的模型未必和真实情况相匹配，一个典型的例子就是，无论支持地心说的托勒密，还是提出日心说的哥白尼，都假定行星运动轨迹的基本模型是最简单的圆，而不是更准确的椭圆。由此可见，如果一开始模型选得不好，那么以后修修补补就很困难。因此，在过去，无论在理论上还是工程上，大家都把

主要的精力放在寻找模型上。

有了模型之后，第二步就是要找到模型的参数，以便让模型至少和以前观察到的数据相吻合。这一点在过去的被重视程度远不如找模型。但是今天它又有了一个比较时髦而高深的词——机器学习。

鉴于完美的模型未必存在，即使存在，找到它也非常不容易，而且费时间，因此就有人考虑是否能够通过用很多简单不完美的模型凑在一起，起到完美模型的效果呢？比如说，是否可以通过很多很多圆互相嵌套在一起，建立一个地心说模型，和牛顿推演出的日心说模型[16]一样准确呢？如今这个答案是肯定的，从理论上讲，只要找到足够多的具有代表性的样本（数据），就可以运用数学找到一个模型或者一组模型的组合，使得它和真实情况非常接近。

这种思路在现实生活中已经被用到。比如美国和苏联在设计飞机、航天器和其他武器上的理念和方法就不同。苏联拥有大量数学功底非常深厚的设计人员，但是缺乏高性能的计算机和大量的数据，因此其科学家喜欢寻找比较准确但是复杂的数学模型；而美国的设计人员相比之下数学功底平平，但是美国的计算机拥有强大的计算能力和更多的数据，因此其科学家喜欢用很多简单的模型来替代一个复杂的模型。这两个国家做出的东西可谓各有千秋，但从结果来看，似乎美国的更胜一筹。

在工程上，采用多而简单的模型常常比一个精确的模型成本更低，也被使用得更普遍。比如在光学仪器的设计上，一个完美的镜头里面的透镜其实不应该是球面镜，因为那样边缘的图像会变形，只有采用抛物面或者其他复杂曲面，才能使得整个画面都清晰。但是这些非球面透镜的加工需要技艺高超的技工。德国因为拥有最好的技工，因此敢于在镜头设计上采用非球面透镜，这样整个光学仪器就非常小巧。而日本缺乏这种水平的技工，但是善于用机器加工，因此日本人在设计光学仪器时，就用好几个球面透镜来取代一个非球面透镜，这样的光学仪器虽然显得笨重，但是容易大规模生产，而且成本非常低。“二战”后，日本超过德国成为全球光学仪器（包括相机）第一大制造国。

回到数学模型上，其实只要数据量足够，就可以用若干个简单的模型取代一个复杂的模型。这种方法被称为数据驱动方法，因为它是先有大量的数据，而不是预设的模型，然后用很多简单的模型去契合数据（Fit Data）。虽然这种数据驱动方法在数据量不足时找到的一组模型可

能和真实的模型存在一定的偏差，但是在误差允许的范围内，单从结果上看和精确的模型是等效的[17]，这在数学上是有根据的。从原理上讲，这类似于前面提到的切比雪夫大数定律。模型来替代一个复杂的模型。这两个国家做出的东西可谓各有千秋，但从结果来看，似乎美国的更胜一筹。

在工程上，采用多而简单的模型常常比一个精确的模型成本更低，也被使用得更普遍。比如在光学仪器的设计上，一个完美的镜头里面的透镜其实不应该是球面镜，因为那样边缘的图像会变形，只有采用抛物面或者其他复杂曲面，才能使得整个画面都清晰。但是这些非球面透镜的加工需要技艺高超的技工。德国因为拥有最好的技工，因此敢于在镜头设计上采用非球面透镜，这样整个光学仪器就非常小巧。而日本缺乏这种水平的技工，但是善于用机器加工，因此日本人在设计光学仪器时，就用好几个球面透镜来取代一个非球面透镜，这样的光学仪器虽然显得笨重，但是容易大规模生产，而且成本非常低。“二战”后，日本超过德国成为全球光学仪器（包括相机）第一大制造国。

回到数学模型上，其实只要数据量足够，就可以用若干个简单的模型取代一个复杂的模型。这种方法被称为数据驱动方法，因为它是先有大量的数据，而不是预设的模型，然后用很多简单的模型去契合数据（Fit Data）。虽然这种数据驱动方法在数据量不足时找到的一组模型可能和真实的模型存在一定的偏差，但是在误差允许的范围内，单从结果上看和精确的模型是等效的[17]，这在数学上是有根据的。从原理上讲，这类似于前面提到的切比雪夫大数定律。

当然，数据驱动方法要想成功，除了数据量大之外，还要有一个前提，那就是样本必须非常具有代表性，这在任何统计学教科书里就是一句话，但是在现实生活中要做到是非常难的。我们在后面的章节中将会看到，这在大数据出现之前，其实都没有做得很好。

在今天的IT领域中，越来越多的问题可以用数据驱动方法来解决。具体讲，就是当我们对一个问题暂时不能用简单而准确的方法解决时，我们可以根据以往的历史数据，构造很多近似的模型来逼近真实情况，这实际上是用计算量和数据量来换取研究的时间。这种方法不仅仅是经验论，它在数学上是有严格保障的。

数据驱动方法最大的优势在于，它可以在最大程度上得益于计算机技术的进步。尽管数据驱动方法在一开始数据量不足、计算能力不够

时，可能显得有些粗糙，但是随着时间的推移，摩尔定律保证了计算能力和数据量以一个指数级增长的速度递增，数据驱动方法可以非常准确。相比之下，很多其他方法的改进需要靠理论的突破，因此改进起来周期非常长。在过去的30年里，计算机变得越来越聪明，这并不是因为我们对特定问题的认识有了多大的提高，而是因为很大程度上我们靠的是数据量的增加。

可以用来说明数据驱动方法对机器智能产生作用的最佳案例，恐怕要数2016年在计算机行业最热门的事件—Google的AlphaGo计算机战胜天才围棋选手李世石了。AlphaGo在围棋方面有很高的智能，来源于它对能找到的全部几十万盘人类高手对弈的分析总结。这么多的对弈是任何人类高手一辈子也学习不完的。在总结了几十万盘的数据后，AlphaGo得到了一个统计模型，对于在不同的局势下该如何行棋有一个比人类更为准确的估计。这就是AlphaGo显得很聪明的原因。

关于数据驱动方法，我们在后面的章节里还会详细介绍，它是大数据的基础，也是智能革命的核心，更重要的是，它带来一种新的思维方式。

小结

数据的范畴远比我们通常想象的要广得多。人类认识自然的过程，科学实践的过程，以及在经济、社会领域的行为，总是伴随着数据的使用。从某种程度上讲，获得和利用数据的水平反映出文明的水平。在电子计算机诞生、人类进入信息时代之后，数据的作用越来越明显，数据驱动方法开始被普遍采用。如果我们把资本和机械动能作为大航海时代以来全球近代化的推动力，那么数据将成为下一次技术革命和社会变革的核心动力。接下来，我们将在这样一个高度上来理解大数据，以及由它带来的全球智能革命。附录Google预测流感传播论文所用的搜索词条种类

Influenza Complication (流感并发症)

Cold/Flu Remedy (感冒/流感治疗法)

General Influenza Symptoms (常见流感症状)

Term for Influenza (流感术语)

Specific Influenza Symptom (特殊流感症状)

Symptoms of an Influenza Complication (流感并发症的症状) Antibiotic Medication (抗生素药物)

General Influenza Remedies(常见流感疗法)

Symptoms of a Related Disease (相关疾病症状)

Antiviral Medication(抗病毒药物)

Related Disease (相关疾病)

Unrelated to | nfluenza (与流感无关)

注释

[1] 关于3K背景辐射的更多描述, 读者朋友可以参阅拙著《文明之光》。

[2] 埃及古尺: 英文为Royal cubits,又名皇家肘, 估计和英尺类似, 是某个法老的肘长。1埃及古尺约为0.524米。

[3] 勾股定理的严格证明直到古 埃及两千年后的毕达哥拉斯 (Pythagoras of Samos)才完成。

[4] 闪米特人: 亚非大陆上一个 古老的民族, 今天的阿拉伯人和犹太人都 是闪米特人的分支。

[5] 预测洪水的大小是为了准确 测量可耕种土地的边界。

[6] 美索不达米亚的原意是指两 条河之间的土地。

[7] 米利都: 位于安那托利亚西 海岸线上的一座古希腊城邦, 靠近米安 得尔河口, 今属土耳其, 以米利都学派而 闻名。

[8] 得洛斯:古希腊的宗教圣地, 相传是太阳神和月神的出生地。

[9] 泰勒斯: 希腊七贤之一, 古 希腊及西方第一个自然科学家和哲学 家, 他开创了米利都学派, 该学派用理性 思维和观测到的事实而不是

用古希腊神话来解释世界。在几何学上，泰勒斯懂得了相似三角形的原理，并利用影子长度计算出大金字塔的高度。

[10] 阿那克西曼德:泰勒斯的学生，米利都学派重要学者。

[11] 菲尔库德斯:得洛斯著名学者，提出了物质不灭和生物进化的理论。

[12] 儒略历：由罗马共和国独裁官儒略·恺撒（即盖乌斯·尤里乌斯·恺撒）采纳数学家兼天文学家索西琴尼的计算后，于公元前45年1月1日起执行的取代旧罗马历法的一种历法。儒略历中，一年被划分为12个月，大小月交替；四年一闰，平年365日，闰年366日。为在当年二月底增加一闰日，年平均长度为365.25日。

[13] Knobler, S.; Mack, A.; Mahmoud, A.; *et al.* (eds.). *The Story of Influenza . The Threat of Pandemic Influenza : Are We Ready ? Workshop Summary (2005)*. Washington, D. C.: The National Academies Press, pp. 60-61.

[14] Jeremy Ginsberg , Matthew H. Mohebbi, Rajan S . Patel, Lynnette Brammer, Mark S. Smolinski and Larry Brilliant, *Detecting influenza epidemics using*

search engine query data , *Nature* Vol 457 , 19 February 2009.

[15] 关于线性回归模型的更多细节，请参见拙著《数学之美》。

[16] 哥白尼的日心说模型非常不准确。

[17]当然，运气好的话从数据出发也有可能得到和真实模型完全一致的结果，但是这并非数据驱动方法的目标。

Ajt - 3±E

第一早 大数据和机器智能

在有大数据之前，计算机并不擅长于解决需要人类智能的问题，但是今天 这些问题换个思路就可以解决了，其核心就是变智能问题为数据问题。由此，全世界开始了新一轮技术革命 智能革命。

当我们有可能获得大量的、具有代表性的数据之后，能够获得什么好处呢？大家很快就想到把一些模型描述得更准确，或者对一些规律认识得更深刻。比如当开普勒从他的老师手上接过大量的天文数据之后，他终于找到了准确描述行星围绕太阳运动轨迹的模型——椭圆模型。类似的情况在今天不断地发生。但是，这还远远不足以让我们兴奋，因为那还只是一个量的改变，不足以产生颠覆这个世界的创新。

大量数据的使用，最大的意义在于它能让计算机完成一些过去只有人类才能做到的事情，这最终将带来一场智能革命。我们不妨用一些具体的例子来说明这种趋势。

在过去，只有人类才有用语音交流的能力，尽管人类从1946年开始就努力让计算机有听得懂人的语音的智能，但是一直不成功。20世纪70年代，科学家们采用数据驱动方法，找到了解决这个问题的途径，并且不断地改进方法。但是语音识别准确率的提高，主要是靠20世纪90年代以后数据的大量积累。从这个研究领域，大家开始看到了数据的重要性。类似地，图像识别也取得了根本性的突破。

在2000年以后，由于互联网特别是后来移动互联网的出现，数据量不仅剧增，而且开始相互关联，出现了大数据的概念。科学家和工程师们发现，采用大数据的方法能够使计算机的智能水平产生飞跃，这样在很多领域计算机将获得比人类智能更高的智能。可以说我们正在经历一场由大数据带来的技术革命，其最典型的特征就是计算机智能水平的提高，因此我们不妨把这场革命称为智能革命。当计算机的智能水平赶上甚至超过人类时，我们的社会就要发生天翻地覆的变化，这才是大数据的可怕之处。

那么为什么大数据会最终导致这样的结果，大数据和机器智能是什么关系呢？要说清楚这一点，首先要说明什么是机器智能。

什么是机器智能

能够辅助计算的机械很早就有了，它的历史可以上溯至美索不达米亚人时代、希腊人时代，以及中国人发明算盘的时代，并且后来经过帕斯卡（Blaise Pascal, 1623—1662）、莱布尼茨（Gottfried Wilhelm Leibniz, 1646—1716）、巴贝奇（Charles Babbage, 1791—1871）和楚泽（Konrad Zuse, 1910—1995）等人的努力，人类制造出了可以编程计算的机器。但是很少有人将它们和具有类似人类智能的思维机器联系起来，后者只存在于科幻小说中。



图2.1帕斯卡发明的机械计算器复原模型（收藏于硅谷计算机博物馆）

1946年，第一台电子计算机ENIAC 诞生，这使得人类重新开始考虑机器能否有智能的问题。从功能上讲，ENIAC 与德国工程师楚泽研制的继电器计算机 Z3没有太大的差别——它们都是能够实现编程功能的图灵机[1]。Z3是一台 继电器计算机，每秒的运算速度只有5 ~ 10次；ENIAC则是一台基于电子管开关 电路的计算机，按照今天的标准来衡量, 它还远远不够完善，因为它每改变一次

程序就要在计算机里面重新连接线路， 因此使用并不方便。但是ENIAC比起Z3 有一个非常突出的优点，就是计算速度 能够达到每秒5000次。虽然这个速度连 今天手机里面处理器速度的十万分之一 都不到，但是比最聪明的人脑运算起来 不知道要快几千倍，因此量变带来了质变。

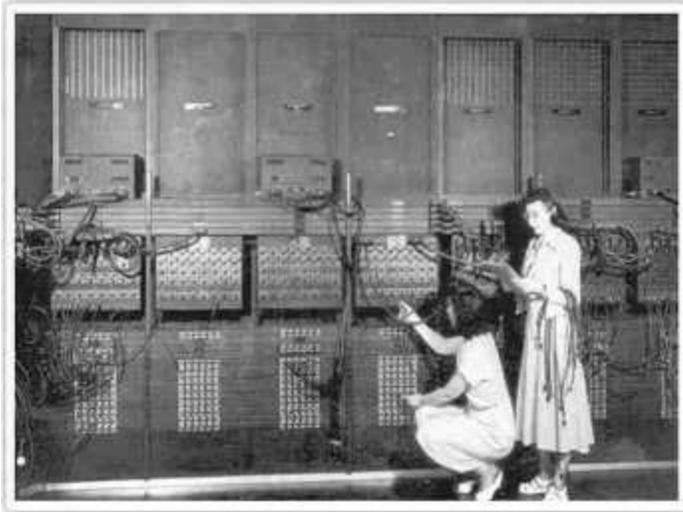


图2.2世界上第一台电子计算机

ENIAC

实际上发明“电脑”一词的不是任何科学家，而是一位英国的元帅——蒙巴顿伯爵。作为英美联军的英军统帅，蒙巴顿参观了 ENIAC 的演示。由于这台计算机最初设计的目的是研制远程火炮的弹道，因此在它被制造出来后，虽然“二战”已经结束，那个远程火炮项目被停止了，但是科学家们依然用计算火炮弹道来展示计算机的计算速度。在过去，计算火炮弹道需要工程师们用计算尺算上好几天，但是 ENIAC 每秒 5000 次的计算速度可以在炮弹打出去后还没有落地之前，就准确地计算出弹道的轨迹。这让蒙巴顿元帅无限感慨，不禁赞道：“这真是电脑啊！”当然，有同样感慨的不止他一人，在 ENIAC 诞生后，各行各业的人，当然也包括科学家们都在问自己，机器能否产生智能？

真正科学地定义什么是机器智能的还是电子计算机的奠基人阿兰·图灵 (Alan Turing, 1912—1954) 博士。1950 年，图灵在《思想》(Mind) 杂志上发表了一篇题为《计算的机器和智能》的论文。在论文中，图灵既没有讲计算机怎样才能获得智能，也没有提出什么解决复杂问题的智能方法，而只是提出了一种验证机器有无智能的判别方法。

图2.3图灵测试

让一台机器和一个人坐在幕后，让一个裁判同时与幕后的人和机器进行交

流，如果这个裁判无法判断自己交流的对象是人还是机器，就说明这台机器有了和人同等的智能。这种方法被后人称为图灵测试 (Turing Test)。计算机科学家们认为，如果计算机实现了下面几件事情中的几件，就可以认为它有图灵所说的那种智能：

1. 语音识别
2. 机器翻译
3. 文本的自动摘要或者写作
4. 战胜人类的国际象棋冠军
5. 自动回答问题

今天，计算机已经做到了上述这几

件事情，有些时候还超额完成了任务，比如下棋方面，不仅战胜了国际象棋的世界冠军，而且还战胜了围棋的世界冠军，后者的难度比前者高出6~8个数量级 ($10^6 \sim 10^8$)。当然，人类走到这一步并非一帆风顺，而是先走了十几年的弯路。

鸟飞派：人工智能1.0

据记载，1956年的夏天，香农和一群年轻的学者在达特茅斯学院召开了一次头脑风暴式的研讨会。会议的倡议者是当时在该学院任教的约翰·麦卡锡 (John McCarthy, 1927—2011), 以及同岁的马文·明斯基 (Marvin Minsky, 1927—2016), 他们当时都只有29岁, 另外两个倡议者是纳撒尼尔·罗切斯特 (Nathaniel Rochester, 1919—2001)和克劳德·香农 (Claude Shannon, 1916—2001), 年龄也不大, 还有6位年轻的科学家参加, 其中包括后来得了图灵奖的赫伯特·西蒙 (Herbert Simon, 1916—2001)和艾伦·纽维尔 (Allen Newell, 1927-1992) ○ 会议虽然叫作“达特茅斯夏季人工智能研究会议”, 其实它不同于今天一般召开几天的学术会议, 因为一来大家并没有可以报告的科研成果, 二来这个会议持续了一个暑期。事实上, 这是一次头脑风暴式的讨论会。这10位年轻的学者讨论的是当时计算机科学尚未解决, 甚至尚未开展研究的问题, 包括人工智能、自然语言处理和神经网络等。人工智能这个说法便是在这次会议上提出的。

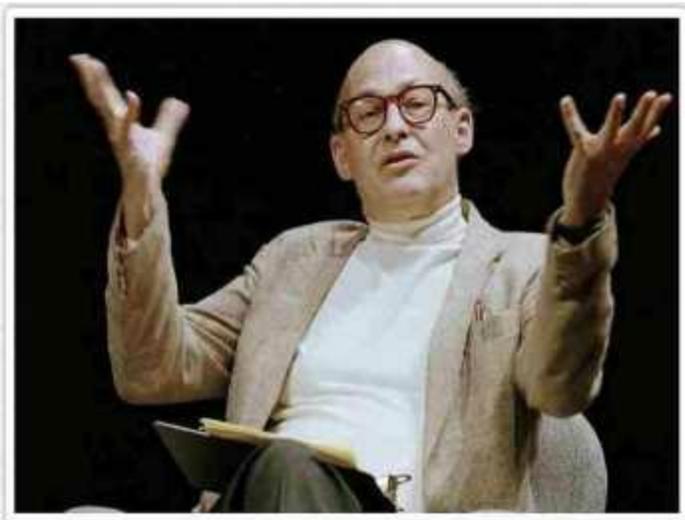


图2.4人工智能的奠基人明斯基

参加达特茅斯会议的10个人, 除了香农, 当时大多都没有什么名气。但是没关系, 这些年轻人籍籍无名的时间不会太久, 后来所有这些都成了计算机科学领域或者认知科学领域的泰斗, 包括4位图灵奖获得者 (麦卡锡、明斯基、西蒙和纽维尔), 而香农作为信息论的发明人, 他的名字被用来冠名通信领域的最高奖——香农奖。

虽然达特茅斯会议本身没有产生什么了不起的思想，10个最聪明的大脑一个暑假的思考甚至比不上今天一位一流的博士毕业生，但是它的意义超过10个图灵奖，因为它提出了问题。好几个未来非常热门的研究领域的研究工作，其中包括人工智能和机器学习，就始于那次会议之后。

人工智能这个名词严格地讲在今天有两个定义，第一个是泛指机器智能，也就是任何可以让计算机通过图灵测试的方法，包括我们在本书中要经常讲的数据驱动方法。第二个是狭义上的概念，即20世纪五六十年代特定的研究机器智能的方法。今天，几乎所有书名含有“人工智能”字样的教科书(包括全球销量最大的由斯图亚特·罗素和诺威格编写的《人工智能：一种现代的方法》一书)依然用主要的篇幅介绍那些“好的老式的人工智能”(Good Old Fashioned AI) [2]。后来那些利用其他方法产生机器智能的学者为了划清自己和传统方法的界限，特地强调自己不是用人工智能的方法。因此，学术界将机器智能分为传统人工智能的方法和现代其他的方法(比如数据驱动、知识发现或者机器学习)。当然，计算机领域之外的人在谈到人工智能时，常常是泛指任何机器智能，而并不局限于传统的方法。因此为了便于区分，我们在本书中尽可能地使用机器智能表示广义上的概念，而在使用人工智能表达时，通常是指传统的人工智能方法，甚至我们有时会强调为人工智能1。

0。

那么传统的人工智能方法是什么呢?简单地讲，就是首先了解人类是如何产生智能的，然后让计算机按照人的思路去做。今天几乎所有的科学家都不坚持“机器要像人一样思考才能获得智能”，但是很多的门外汉在谈到人工智能时依然想象着“机器在像我们那样思考”，这让他们既兴奋又担心。事实上，当我们回到图灵博士描述机器智能的原点时就能发现，机器智能最重要的是能够解决人脑所能解决的问题，而不在于是否需要采用和人一样的方法。

为什么早期科学家们的想法会和今天的门外汉一样天真呢?这个道理很简单，因为这是根据我们的直觉最容易想到的方法，在人类发明的历史上,很多领域早期的尝试都是模仿人或者动物的行为。比如人类在几千年之前就梦想着飞行，于是就开始模仿鸟，在东方和西方都有类似的记录，将鸟的羽毛做成翅膀绑在人的胳膊上往下跳，当然实验的结果都可想而知。后来人们把这样的方法论称作“鸟飞派”，也就是看

看鸟是怎样飞的，就能模仿鸟造出飞机，而不需要了解空气动力学。事实上我们知道，怀特兄弟发明飞机靠的是空气动力学而不是仿生学。在这里，我们不要笑话前辈来自直觉的天真想法，这是人类认识的普遍规律。

在人工智能刚被提出来的时候，这一个研究课题在全世界都非常热门，大家仿佛觉得用不了多长时间就可以让计算机变得比人聪明了。遗憾的是，经过十几年的研究，科学家们发现人工智能远不是那么回事，除了做出了几个简单的“玩具”，比如让机器人像猴子一样摘香蕉，解决不了什么实际问题。到了20世纪60年代末，计算机科学的其他分支都发展得非常迅速，但是人工智能研究却开展不下去了。因此，美国计算机学界开始反思人工智能的发展。虽然一些人认为机器之所以智能水平有限，是因为它还不够快、容量不够大，但是，也有一些有识之士认为，科学家们走错了路，照着那条路走下去，计算机再快也解决不了智能问题。1968年，明斯基在《语义信息处理》（Semantic Information

Processing）_书中分析了所谓

人工智能的局限性，他引用了巴希勒（Bar-Hillel）使用过的一个非常简单的例子：

The pen was in the box（钢笔在盒子里），这句话很好理解，如果让计算机理解它，做一个简单的语法分析即可。但是另一句语法相同的话：

The box was in the pen.



图2.5钢笔在盒子里，这句话很好理解

就让人颇为费解了。原来，在英语中，pen（钢笔）还有另外一个不太常用的意思——小孩玩耍的围栏。在这里，理解成这个意思整个句子就通顺了。但是，如果用同样的语法分析，这两句话会得到相同的语法分析树，而仅仅根据这两句话本身，甚至通篇文章，是无法判定pen在哪一句话中应该作为围栏，在哪一句话中应该是钢笔的意思。事实上人对这两句话的理解并非来自语法分析和语意本身，而是来自他们的常识或者说关于世界的知识(world knowledge)，这个问题是传统的人工智能方法解决不了的。因此，明斯基给出了他的结论：“目前”（指1968年）的方法无法让计算机真正有类似于人的智能。由于明斯基在计算机科学界具有崇高的声望，他的这篇论文导致美国政府削减了几乎全部人工智能研究的经费，在之后大约20年左右的时间里，全世界人工智能在学术界的研究是处于低谷的。

图2.6 pen（钢笔）的另一个含义-围栏

另辟蹊径：统计+数据

到了20世纪70年代，人类开始尝试 机器智能的另一条发展道路，即采用数据驱动和超级计算的方法，而这个尝试始于工业界而非大学。

在那个年代，BM在全世界计算机乃至整个IT产业可以说是处于独孤求败的地位。20世纪60年代末，IBM的市值达到500亿美元，这在当时是个很大的数目，占到了美国GDP（国内生产总值）的3%以上。当时，全世界制造大型计算机的只有8家公司，它们被比喻成白雪公主和7个矮人。白雪公主是BM，7个矮人是其他7家公司。如果将这7家公司的营业额加在一起，再加上当时生产小型机动数字设备公司DEC（美国数字设备公司）和惠普，还不如IBM多，因此IBM快被司法部进行反垄断调查了。这时，IBM考虑的不能再是如何占有更大的市场份额，而是如何让计算机变得更聪明。

1972年，康奈尔大学的教授弗雷德·贾里尼克（Fred Jelinek，1932-2010）到IBM做学术休假[3]，正好这时IBM想开发“聪明的计算机”，贾里尼克就“临时”负责起这个项目。至于什么是聪明的计算机，当时大家的共识是它要么能够听懂人的话，要么能将一种语言翻译成另一种语言，要么能够赢得了国际象棋的世界冠军。贾里尼克根据自己的特长和IBM的条件，选择了第一个任务，即计算机自动识别人的语音。

在贾里尼克之前，各个大学和研究 所的专家们在这个问题上已经花了20多年的时间，主流的研究方法有两个特点，一个是让计算机尽可能地模拟人的发音特点和听觉特征，二是利用人工智能的方法理解人所讲的完整的语句。对于前一项研究，有时又被称为特征提取，各个研究单位都有自己的见解，采用各自不同的方法，很难比较哪个更好，而且这些和人的发音或者听力相关的特征也很难统一到一个系统中。对于后一项研究，大家采用的方法倒是差不多，具体讲就是传统人工智能的方法，它基于语法规则和语义规则，打一个比方，有点像教大家学外语。在20世纪70年代初，语音识别这个智能问题解决到了什么水平呢？当时最好的语言识别系统大约能够识别百十来个单词，识别率只有70%左右，而且讲话时要口齿清晰，没有噪声。

贾里尼克从来不是一位人工智能专家，他是一位通信专家，因此他看待语音识别问题的角度和先前的计算机科学家们都不相同——在他看来，语音识别不 是一个人工智能的问题，而是一个通信问题。

贾里尼克认为，人的大脑是一个信息源，从思考到找到合适的语句,再通过发音说出来，是一个编码的过程,经过媒介（声道、空气或者电话线、扬声器等）传播到听众耳朵里，是经过了一个长长的信道的信息传播问题，最后听话人把它听懂，是一个解码的过程。既然是一个典型的通信问题，就可以用解决通信问题的方法来解决，为此贾里尼克用两个数学模型(马尔可夫模型)分别描述信源和信道。至于计算机识别时 需要从语音中提取什么特征，贾里尼克的想法很简单，数字通信采用什么特征，语音识别就采用什么特征。这样，贾里尼克就用当时已经颇为成熟的数字通信的各种技术来实现语音识别，而彻底抛开了人工智能的那一套做法。

正如我们在前面介绍的，找到了数学模型之后，下一步就是要用统计的方法“训练出”模型的参数，这在今天来讲就是机器学习。在这个过程中，需要使用大量的数据，同时要有足够的计算能力。在当时，只有IBM具备这些条件。那时不仅没有互联网上大量的内容，甚至没有很多存在计算机里的文本（又称机读文本），好在IBM有大量的电传文本，这成了IBM语音识别系统使用的最早期的数据。此外，在当时没有第二家公司有IBM那样的计算能力，当然，那时贾里尼克整个团队所拥有的计算能力还不及今天一部iPhone(苹果)手机呢！

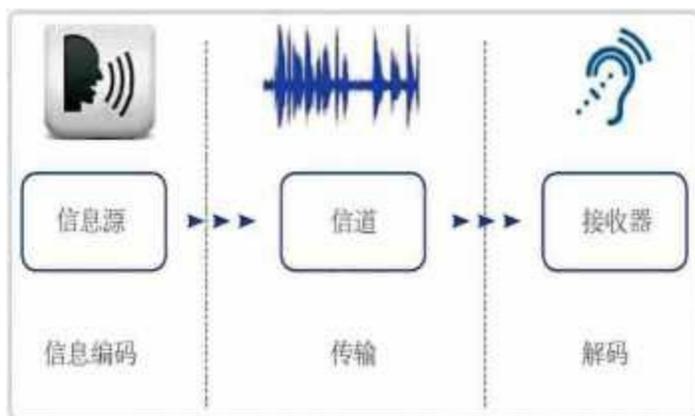


图2.7语音识别的通信模型

贾里尼克除了找到一条不同于传统人工智能的语音识别方法，另一个特点就是喜欢招收数学基础好的，特别是学习过理论物理的员工。出于某种原因，他不喜欢语言学家并且把他们都请出了IBM。贾里尼克的团队花了4年的时间，就开发了一个基于统计方法的语音识别系统，它的语音识别率从过去的70%左右提高到90%以上，同时语音识别的规

模从几百词上升到两万多词。这样语音识别就有了本质的飞跃。我们不妨想想，如果一个语音识别系统每10个汉字就错3个，我们是无法读懂这句话的，但是如果每10个汉字只错1个，我们就能准确还原原来语句的意思。更何况，几百个英文单词只能进行类似幼儿园小孩之间的交流，而两万多个英文单词足够母语是英语的人进行各种交流了。从此，语音识别就能够从实验室走向实际应用了。

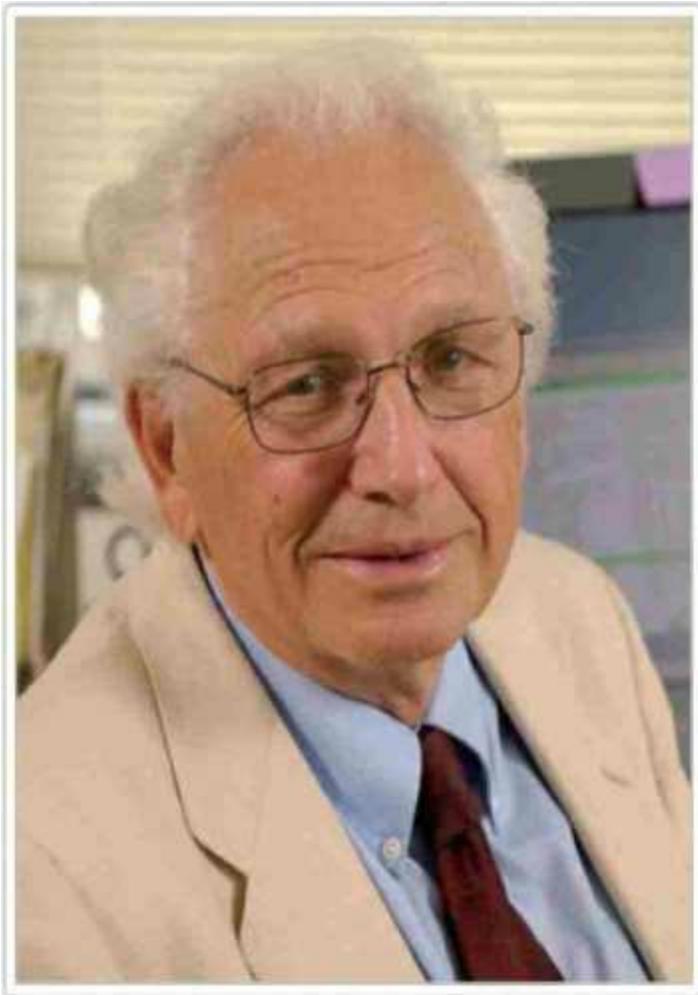


图2.8语音识别和机器学习的先驱贾里尼克

贾里尼克和他的同事在研究语音识

别时，无意中开创了一种采用统计的方法解决智能问题的途径，因为这种方法需要使用大量的数据，因此又被称为数据驱动方法。这种方法最大的好处是，随着数据量的积累，系统会变得越来越好，相比之下过去人工智能的方法很难受益于数据量的提升。

后来在IBM和Google先后担任过 主管研究的副总裁阿尔弗雷德_斯伯格特 (Alfred Spector)博士, 20世纪80 年代时是卡内基-梅隆大学的教授, 据 他介绍, 当年卡内基-梅隆大学已经在 传统的人工智能领域走得非常远了, 大 家遇到了很多跨不过去的障碍。后来教 授们去IBM沃森实验室参观, 看到那里采 用数据驱动方法取得的巨大成绩, 回来 以后很多教授接受了这种新的方法论。 李开复就是在这样的背景下, 在传统的人工智能实验室里, 采用基于统计的方 法开展他的博士论文的工作, 并且最终 和洪小文一起构建了世界上第一个大词 汇量、非特定人、连续语音识别系统[4]。按照斯伯格特的说法, 如果没有李开 复等人的工作, 他们的论文导师瑞迪 (Raj Reddy)不可能获得图灵奖。

在语音识别之后, 欧洲和美国的科 学家开始考虑能否用数据驱动方法解决 其他智能问题。贾里尼克的同事彼得-布朗 (Peter Brown)在20世纪80 年代, 将这种数据驱动方法用于机器翻 译[5]。但是由于缺乏数据, 最初的翻 译结果并不令人满意, 虽然一些学者认 可这种方法, 但是其他学者, 尤其是早期 从事这项工作的学者认为, 解决机器翻 译这样智能的问题, 光靠基于数据的统 计是不够的。从20世纪80年代初到90年代中期大约十多年的时间里, 在计算机 界大家一直有个争议, 那就是数据驱动 方法是否适用于各种领域, 语音识别是 否只是一个特例。简单地讲, 当时无论是 做语音识别、机器翻译、图像识别, 还是 自然语言理解的学者, 分成了界限很明 确的两派, 一派坚持采用传统的人工智 能方法解决问题, 简单来讲就是模仿人, 另一派在倡导数据驱动方法。这两派在 不同的领域力量不一样, 在语音识别和 自然语言理解领域, 提倡数据驱动的一 派比较快地占了上风; 而在图像识别和 机器翻译方面, 在较长时间里, 数据驱动 这一派处于下风。这里面主要的原因是, 在图像识别和机器翻译领域, 过去的数 据量非常少, 而这种数据的积累非常困 难。图像识别就不用讲了,在互联网出现 之前, 没有一个实验室有上百万张图片。在机器翻译领域, 所需要的数据除了_般的 文本数据, 还需要大量的双语(甚至 是多语种)对照的数据, 而在互联网出现 之前, 除了《圣经》和少量联合国文件, 再也找不到类似的数据了。[6]—直到 21世纪初, SYSTRAN (系统翻译) [7]等研究机器翻译的公司, 依然在组织 大量的人力编写机器翻译使用的语法规 则。针对 一对语言, 比如英语和汉语, 他 们要编写几万条规则。在这几万条规则 的帮助下, 直到2002年, SYSTRAN公 司的中英翻译系统依然是全世界做得最 好的, 但是进入21世纪之后, SYSTRAN 很快便落伍了, 因为数据驱动方法在数 据量不断增加之后, 它的优势便渐渐显 现出来。

在20世纪90年代互联网兴起之后，数据的获取变得非常容易。从1994年到2004年的10年里，语音识别的错误率减少了一半，而机器翻译的准确性[8]提高了一倍，其中20%左右的贡献来自方法的改进，80%则来自数据量的提升。虽然在每一年，计算机在解决各种智能问题上的进步幅度并不大，但是十几年量的积累，最终促成了质变。

数据创造奇迹：量变到质变

从某种意义上讲，2005年是大数据元年，虽然大部分人感受不到数据带来的变化，但是一项科研成果却让全世界从事机器翻译的人感到震惊，那就是之前在机器翻译领域从来没有技术积累、不为人所知的Google，以巨大的优势打败了全世界所有机器翻译研究团队，一跃成为这个领域的领头羊。

故事要从这一年的2月说起。全世界拿了美国政府机器翻译科研经费的研究机构，不论是大学还是公司，照例都要参加由美国国家标准与技术研究所 (National Institute of Standards and

Technologies,简称NIST)主持的测评和交流，而且需要介绍自己研究方法的细节。当然，没有拿美国政府机构的研究团队也可以参加，但是没有义务披露太多的细节。这一年的测评从2月开始，Google的机器翻译团队是第一次参加这个测评，其他的团队要么过去曾经取得过很好的成绩，比如德国亚琛工学院，要么研究的历史非常长，比如旧"和SYSTRAN,因此在测试之前谁也没有关注Google团队的表现。

当年4月，测评的结果出来了，让除了Google以外的所有人大吃一惊，在所有4项测评中，之前从来没有做过机器翻译的Google均比其他研究团队同类的系统领先了一大截。表2.1是2005年NIST评比的结果，表中所给的数据是机

器翻译结果和人工翻译结果之间的BLEU分数。关于BLEU分数，简单地讲，它反映了两种翻译结果的一致性，因此这个分数越高越好。当然，并非BLEU分数要达到100%才算翻译完全正确，因为人和人之间的BLEU分数大约只有50%左右。明确了评分标准，我们不妨看两项评比的结果：从阿拉伯语到英语的翻译，Google系统的得分为51.31%，领先第二名将近5%，而提高这5个百分点在过去需要研究5-10年的时间[9]；而在中文到英语的翻译中，Google51.37%的得分比第二名领先了17%，这个差距已经超出了一代人的水平。表2.1 2005年美国国家标准与技术研究所 (NIST)对全世界多种机器翻译系统的评比结果

从阿拉伯语到英语的翻译

Google	51.31%
南加州大学	46.57%
IBM沃森实验室	46.46%
马里兰大学	44.97%
约翰·霜龄触学	43.48%
SYSTRAN 公司	10.79%
从中文到英语的翻译	
Google	51.37%
SAKHR公司	34.03%
美军ARL研究所	22.57%

至于为什么Google能够做到这一点，其中一个原因行业里的人都知道，就是Google花重金请到了当时世界上水平最高的机器翻译专家弗朗兹·奥科 (Franz Och)博士。事实上参加测评的系统中，有两个可以说是Google系统的姊妹系统，那就是亚琛工学院的系统和南加州大学的系统，前者是奥科读博士时写的，后者是奥科做研究教授时写的。但是，奥科是2004年7月才正式到Google上班的，[10]这一年的7月到第二年的2月，奥科也只能赶时间把他过去的工作在Google重新演练一下，根本不可能做实质性的改进。那么为什么Google的系统要比它的姊妹系统好很多呢？

根据NIST的要求，大家在测评结果出来后，一般是在5月到7月之间，要开一次研讨会，交流各自的研究方法。以前，大家的兴趣在于相互i寸论，当然在学术界，老朋友们也通过这种方式见见面，联络一下感情。但是这一次大家的目的非常明确，就是看看Google的秘密武器到底是什么。



图2.9 Google翻译的发明人奥科博士

这一年的7月，大家来到NIST所在的弗吉尼亚州北部开会交流经验，奥科则是这次会议的焦点人物。大家都想听他的秘诀，但是这个秘诀一讲出来就不值钱了，他用的还是两年前的方法，但是用了比其他研究所多几千倍甚至上万倍的数据。其实，在和自然语言处理有关的领域，科学家们都清楚数据的重要性，但是在过去，不同研究组之间能使用的数据通常只相差两三倍，对结果即使有些影响，也差不了很多。但是，当奥科用了上万倍的数据时，量变的积累就导致了质变的发生。奥科能训练出一个六元模型，而当时大部分研究团队的数据量只够训练三元模型[11]。简单地讲，一个好的三元模型可以准确地构造英语句子中的短语和简单的句子成分之间的搭配，而六元模型则可以构造整个从句和复杂的句子成分之间的搭配，相当于将这些片段从一种语言到另一种语言直接对译过去了。不难想象，如果一个系统对大部分句子在很长的片段上直译，那么其准确性相比那些在词组单元做翻译的系统要准确得多。在Google之则，不是没有人想到五元或者六元模型，但是如果没有充足的数据，那么训练出来的五元或六元模型准确性非常差，对翻译没有任何帮助。

从表2.1中可以看到，采用传统人工智能方法的SYSTRAN公司和那些采用数据驱动的系统相比，差距之大已经不在一个时代了，因此从2005年NIST测评之后，它就逐渐退出了历史舞台，如今已经没有多少人知道它了。在那次测评之后，其他大学和研究所则把大部分精力都用到收集数据上了。在第二年的测评中，所有研究组都使用了比前一年至少多100倍的数据，它们和Google的差距迅速缩小。

如今在很多与“智能”有关的研究领域，比如图像识别和自然语言理解，如果所采用的方法无法利用数据量的优势，

会被认为是落伍的。

数据驱动方法从20世纪70年代开始起步，在八九十年代得到缓慢但稳步的发展。进入21世纪后，由于互联网的出现，使得可用的数据量剧增，数据驱动方法的优势越来越明显，最终完成了从量变到质变的飞跃。如今很多需要类似人类智能才能做的事情，计算机已经可以胜任了，这得益于数据量的增加。

全世界各个领域数据不断向外扩展，渐渐形成了另外一个特点，那就是很多数据开始出现交叉，各个维度的数据从点和线渐渐连成了网，或者说，数据之间的关联性极大地增强，在这样的背景下，就出现了大数据。

大数据的特征

大数据一词经常出现在媒体上是 2007年以后的事情，但是大家对它的理解并不统一，有些甚至是误解，比如将大数据和大规模数据混为一谈。要谈大数据的问题，我们先要讲清楚什么是大数据，它都有哪些特征。

大数据最明显的特征是体量大，这一点无论是内行还是外行都认可，没有什么异议。但是仅仅有大量的数据并不一定是大数据，比如一个人基因全图谱的数据，是在上百GB (吉字节)到TB (太字节)数量级[12]，这个数据量不可谓不大，但是它没有太大的统计意义。再比如，如果记录下全世界70亿人的出生日期，这个数据量也不小，但是如果仅仅有这一项数据，它除了能够非常准确地给出全世界人口的年龄分布外，也得不到太多其他统计信息。事实上，要了解全世界人口的年龄分布，用传统的抽样统计方法就可以得到，因此这个大量的数据意义也不大。

大数据之所以有用，是因为它除了数据量大以外，还具有其他的特征。一些数据专家将大数据的特征概括成三个V，即大量 (Vast)、多样性 (Variety)和及时性 (Velocity)，这种说法虽然方便记忆，但并不全面准确。首先，尽管一些大数据具有及时性的特点，我们也会在后面详细介绍及时性的好处，但它并非所有大数据所必需的特征，一些数据没有及时性，一样可以被称为大数据。其次，多样性虽然是大数据的一个特征，但是含义上有歧义性，其中最重要的含义是多维度。实际上，多维度的讲法更加简明而准确。因此，在不引起混淆的情况下，我们今后把Variety解释成多维度。至于多维度的重要性和它的威力，我们不妨通过下面一个简单的例子来看一看。

2013年9月，百度发布了一个颇有意思的统计结果——《中国十大“吃货”省市排行榜》。百度没有做任何民意调查和各地饮食习惯的研究，它只是从“百度知道”的7700万条与吃有关的问题里“挖掘”出来一些结论，而这些结论看上去比任何学术研究的结论更能反映中国不同地区的饮食习惯。我们不妨看看百度给出的一些结论：

在关于“xx能吃吗”的问题中，福建、浙江、广东、四川等地的网友最经常问的是“XX虫能吃吗”，江苏、上海、北京等地的网友最经常问的是“XX的皮能不能吃”，内蒙古、新疆、西藏的网友则最关心“蘑菇能

吃吗”，而宁夏网友最关心的竟然是“螃蟹能吃吗”。宁夏网友关心的事情一定让福建网友大跌眼镜，反过来也是一样，宁夏网友会惊讶于有人居然要吃虫子。

百度做的这件小事，其实反映出大数据多维度特征的重要性。百度知道的数据维度很多，它们不仅涉及食物的做法、吃法、成分、营养价值、价格、问题来源的地域和时间等显性的维度，而且还藏着很多外人不注意的隐含信息，比如提问者或回答者使用的计算机（或手机）以及浏览器。这些维度并不是明确地给出的（这一点和传统的数据库不一样），因此在外行人看来，百度知道的原始数据说得好看点是具有多样性，说得不好听是“相当杂乱”的一堆。但恰恰是这些看上去杂乱无章的数据将原来看似无关的维度（时间、地域、食品、做法和成分等）联系了起来。经过对这些信息的挖掘、加工和整理，就得到了有意义的统计规律，比如百度公布出来的关于不同地域的人的饮食习惯。

当然，百度只公布了一些大家感兴趣的结果，只要它愿意，它可以从这些数据中得到更多有价值的统计结果。比如，它很容易得到不同年龄、性别和文化背景的人的饮食习惯（假定百度知道用户的注册信息是可靠的，即使不可靠，也可以通过其他方式获取可靠的年龄信息），不同生活习惯的人（比如正常作息的人、夜猫子们、在计算机前一坐就是几个小时的游戏玩家、经常出差的人或者不爱运动的人等）的饮食习惯等。如果再结合每个人使用的计算机（或者手机等智能设备）的品牌和型号，大抵可以了解提问者和回答者的收入情况，这样就可以知道不同收入阶层的人的饮食习惯。当然，为了不引起大家对隐私问题的担忧，百度是不会公布这些结果的。由于百度的数据收集的时间跨度比较长，通过这些数据还可以看出不同地区人饮食习惯的变化，尤其是在不同经济发展阶段饮食习惯的改变。而这些看似很简单的问题，比如饮食习惯的变化，没有百度知道的大数据，尤其是它的多维度特征，还真难得到答案。

说到这里，大家可能会有个疑问，上面这些统计似乎并不复杂，按照传统的统计方法应该也可以获得。在这里，我不是说传统的统计方法行不通，而是其成本非常高，难度相当大，比一般人想象的要大很多。我们不妨看看如果是用过去传统的统计方法得到同样准确的结果必须做哪些事情。首先，需要先设计一个非常好的问卷（并不容易），然后要从不同地区寻找具有代表性的人群进行调查（这就是盖洛普一直在做的事情），最后要半人工地处理和整理数据[13]。这样不仅

成本高，而且如同盖洛普民调一样，很难在采样时对各种因素考虑周全。如果后来统计时发现调查问卷中还应该再加一项，对不起，补上这一项要让整个成本几乎翻一番，因为大部分人工的工作要重新来。

传统方法难度大的第二个原因是填写的问卷未必反映被调查人真实的想法。要知道大家在百度知道上提问和回答是没有压力，也没有功利目的的，有什么问题就提什么问题，知道什么答案就回答什么答案。但是在填写调查问卷时就不同了，大部分人都不想要自己表现得“非常怪”，因此是绝不会在答卷上写下自己有“爱吃臭豆腐”的习惯，或者有“喜欢吃虫子”的嗜好。中央电视台过去在调查收视率时就遇到这样的情况，他们发现通过用户填写的收视卡片调查出的收视率，和自动收视统计盒子得到的结果完全不同。在从收视卡得到的统计结果中，那些大牌主持人和所谓高品位的节目收视率明显地被夸大了，因为用户本能地要填一些让自己显得有面子的节目。我本人也做过类似的实验，从社交网络的数据得到的对奥巴马医疗改革的支持率(大约只有24%)比盖洛普民调的结果(41%)要低得多。

现在有了百度知道这样多维度的大数据，这些在过去看来很难处理的问题便可以迎刃而解了。

大数据的第三个重要特征，也是人们常常忽视的，就是它的全面性，或者说完备性。我们不妨再用中英文翻译的例子来说明大数据的完备性。

小明是在中国出生长大的小学生，在学校里学习了一句“早上好——Good morning”，他将这个句子的中英对应关系背了下来。因此，如果你让他翻译“早上好”这句话，他是会的，但是这并不说明他对英语有多少了解，而仅仅是因为他的脑子里有了这种对应关系。当然，如果他又学会了“你”(you)这个词，他按照自己理解中文的方式去翻译，会翻译出一种洋泾浜式的句子“Good you”，这显然翻译错了。早期计算机自动翻译的很多错误也是这样来的。那么如果我们再教小明一句英语“你好——How are you”，他又背了下来，现在小明就可以翻译两句话了。当然，小明不可能将所有中文句子到英文的翻译背下来，死记硬背学习英语的方法是所有老师都反对的。因此，小明为了将汉语的文章翻译成英语，需要先学会这两种语言，然后读中文写的文章，逐句理解它的含义之后，根据语法和语义，翻译成英语。

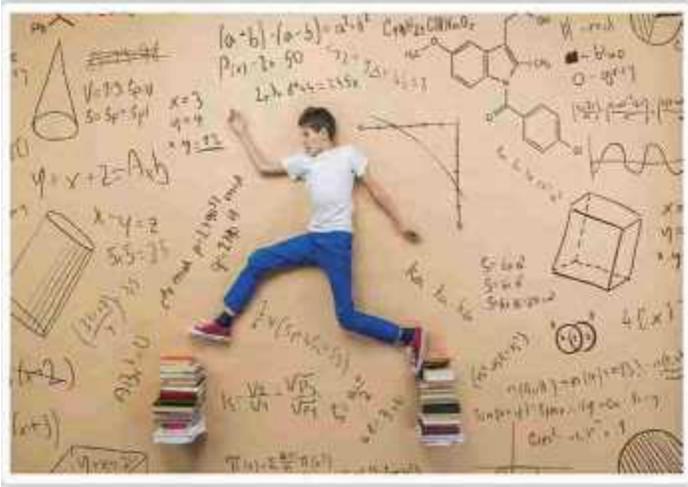


图2.10死记硬背的学习方式

过去科学家们研制机器翻译系统就是这个思路。而奥科在Google做的翻译系统没有采用这种思路，而是采用类似于死记硬背的笨办法，也就是说通过数据学到了不同语言之间很长的句子成分的对应，然后直接把一种语言翻译成另一种，当然前提是，奥科使用的数据必须是全面地覆盖中文、英语和阿拉伯语所有的句子，然后通过机器学习，获得两种语言之间各种说法的翻译方法，也就是说具备两种语言之间翻译的完备性。奥科幸运的是，他当时是在Google工作，有条件获得完备的主要语言常见说法的数据和两种语言对应的译法，而其他研究单位没有这么完备的数据，因此奥科才能够做得比别人好。

美国媒体还报道过另一个大数据完备性的例子——预测2012年美国总统大选结果。我们在上一章提到，盖洛普博士靠成功地预测了1936年美国总统大选的结果而出名，并且从此他的公司在每次美国大选时都做预测。总的来讲，盖洛普公司的预测虽然结果正确的时候占大多数，但是也错了不少次，而且即使在它预测正确的时候，也没有一次能够正确预测美国全部50个州再加上华盛顿特区的选举结果。为什么准确预测美国各州的选举结果那么重要呢？因为美国总统的选举不像法国那样是简单的一人一票制，而是现有各州选举出该州的获胜者，这个获胜者通吃全州被分配的选票数额（比如加州是55票）[14]，因此准确预测各州的选票很重要。在过去，盖洛普公司做了这么多年的预测也做不到准确预测全部50+1个州的结果，因此统计学家们认为这不是盖洛普公司本事不大，而是这件事本身就办不到。美国每次大选时选举结果事先不明朗的州大约有10个左右，在那些州里，各候选人支持率民意调查的差距比标准差要小很多，因此可以讲各种民调给出的结论

基本上是随机的。要随机猜对10个州的大选结果，这个概率其实不到千分之₁，是非常小概率事件。

但是到了2012年，情况发生了变化，一个名叫内德■斯维尔（Nade Silver）的年轻人，利用大数据，成功地预测了全部50+1个州的选举结果。这让包括盖洛普公司在内的所有人都大吃一惊。斯维尔是怎样解决这个难题的呢？其实他的思路很简单，如果有办法在投票前了解到每一个人会投哪个候选人的票，那么准确预测每一个州的选举结果就变得可能了。于是，他在互联网上，尤其是互联网的各种社交网络上，尽可能地收集所有和美国2012年大选有关的数据，其中包括各地新闻媒体上的数据，留言簿和地方新闻中的数据,Facebook(脸谱网)和Twitter(推特)上大家的发言及其朋友的评论，以及候选人选战的数据等，然后按照州进行整理。

虽然斯维尔还做不到在大选前得到每一个投票人的想法，但是他统计的数据已经非常全面了，远不是民意调查公司所能比拟的。另一个重要的因素是，斯维尔的数据反映了选民在没有压力的情况下真实的想法，准确性很高。两点结合起来，斯维尔获得了对选民想法的全面了解，或者说在某种程度上具有了数据的完备性，因此他能够准确预测2012年美国大选结果也就不奇怪了。



图2.11 2012年,斯维尔预测的美国大选结果（左）和实际的结果（右），红色代表共和党获胜，蓝色代表民主党获胜

当然，并非在所有时候，数据的完备性都可以获得，但是局部数据的完备性还是可能获得的，因此利用局部完备性，我们可以解决部分问题。在下一节计算机自动回答问题的例子中，我们可以看到局部的完备性也能够帮助我们。

大数据的时效性其实不是必需的，但是有了时效性可以做到很多过去做不到的事情，城市的智能交通管理便是一个例子。在智能手机和智

能汽车(特斯拉 等)出现之前, 世界上的很多大城市虽然 都有交通管理 (或者控制) 中心, 但是它们能够得到的交通路况信息最快也有20 分钟滞后, 这是Google在2007年最初 推出Google地图交通路况信息服务时所面临的情况。这些信息虽然以较快的 方式加入Google的服务中, 但用户看到 时, 却已经有了半小时的延时。如果没有 能够跟踪足够多的人 出行情况的实时信 息的工具, 一个城市即使部署再多的采 样观察点, 再频繁地报告各种交通事故 和拥堵的情况, 整体交通路况信息的实 时性也不会比2007年有多大改进。



图2.12过去交通路况信息发布的流程

但是, 在能够定位的智能手机出现 后, 这种情况得到了根本的改变。由于智 能手机足够普及并且大部分用户开放了 他们的实时位置信息 (符合大数据的完 备性), 使得做地图服务的公司, 比如 Google或者 百度, 有可能实时地得到任 何一个人口密度较大的城市的人员流动 信息, 并且根据其流动的速度和所在的 位置, 很容易区分步行的人群和 行进的 汽车。

由于收集信息的公司和提供地图服 务的公司是一家, 因此从数据采 集、数据 处理, 到信息发布中间的延时微乎其微, 所提供的交通路况 信息要及时得多。使 用过Google地图服务或者百度地图服 务的人, 对 比六七年前, 都很明显地感到 了其中的差别。当然, 更及时的信息可 以 通过分析历史数据来预测。一些科研小 组和公司的研发部门, 已经 开始利用一 个城市交通状况的历史数据, 结合实时 数据, 预测一段时 间以内(比如一个小时) 该城市各条道路可能出现的交通状况, 并且帮助 出行者规划最好的出行路线。 我们在后面的章节里还会介绍大数据帮 助改进城市交通的案例, 并且分析大数 据时效性对社会带来的影响。

大量全面的信息

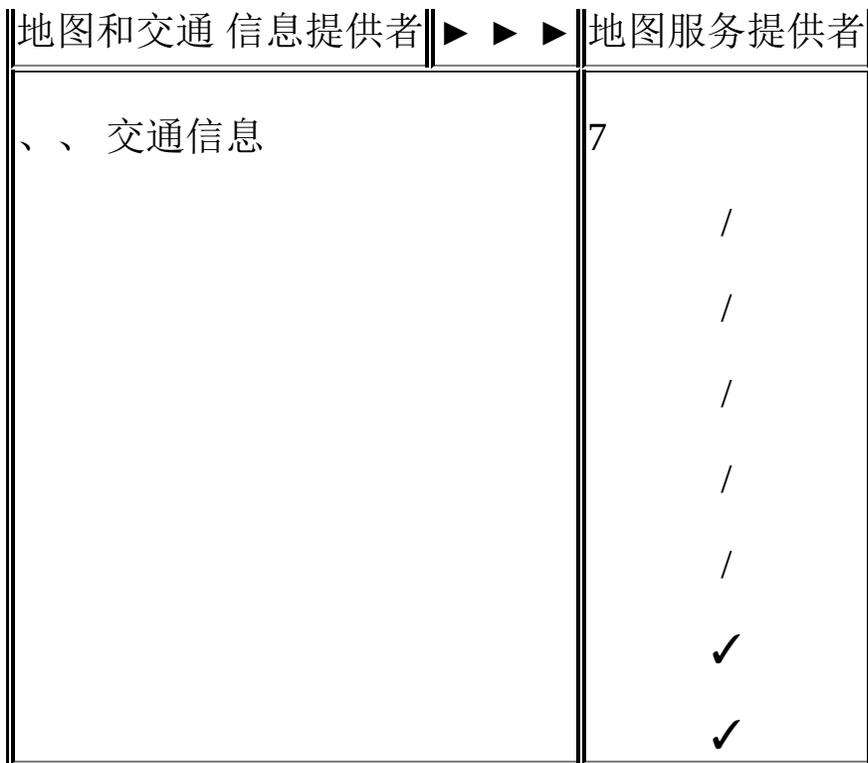


图2.13智能手机出现之后，交通路况 信息发布的流程

大数据的最后一个，或许也是最重要的一个特点，通过分析它名称的英文 写法就能够知道。英语里的large和big 翻译成中文都是大的意思，因此很少有人关心为什么大数据使用“big data”这 个英语词组，而不是“large data”。但是，在大数据被提出之前，很多通过收集 和处理大量数据进行科学研究的论文， 都采用large或者vast (海量) 这两个英 文单词，而不是big。比如我们常常可以 看到论文的标题包含“large Scaled...” “Vast Data...” "Large Amount..." 等词组，但是很少用Big。

SP么big, l arge和vas倒底有什么差 别呢。l arge和vast在程度上略有差别， 后者可以看成是very large的意思。而 big和它们的差别在于， big更强调的是相 对小的大，是抽象意义上的大，而large 和vast常常用于形容 体量的大小。比如 “large table”常常表示一张桌子尺寸很 大，而如果 说“big table”其实是要表示 这不是一张小桌子，真实尺寸是否很大 倒不 一定，但是这样的说法是要强调已 经称得上大了，比较抽象。

仔细推敲英语中big data这种说法， 我们不得不承认这个提法非常准 确，它 最重要的是传递了一种信息——大数据 是一种思维方式的改 变。现在的数据量 相比过去大了很多，量变带来了质变，思 维方式、

做事情的方法就应该和以往有所不同。这其实是帮助我们理解大数据概念的一把钥匙。在有大数据之前，计算机并不擅长解决需要人类智能来解决的问题，但是今天这些问题换个思路就可以解决了，其核心就是变智能问题为数据问题。由此，全世界开始了新一轮技术革命——智能革命。

变智能问题为数据问题

尽管在过去的半个世纪，计算机的运算速度一直呈指数级提升，可以做的事情越来越多，可是给人的印象依然是“快却不够聪明”，比如它不能回答人的提问，不会下棋，不认识人，不能开车，不善于主动做出判断.....然而当数据量足够大之后，很多智能问题都可以转化成数据处理的问题，这时，计算机开始变得聪明起来。第一次让全世界感到计算机智能水平有了质的飞跃是在1996年，那一年计算机第一次战胜人类的国际象棋世界冠军。不过相比2016年AlphaGo战胜李世石，那一次的比赛更加一波三折、惊心动魄。

1996年，IBM的超级计算机深蓝和当时的国际象棋世界冠军卡斯帕罗夫进行了一场六番棋的比赛。卡斯帕罗夫是世界上最富传奇色彩的国际象棋世界冠军，他的等级分之高在当时创造了纪录，这个纪录直到十多年后，才由今天的世界冠军卡尔松打破，即便如此卡斯帕罗夫依然保持着有史以来第二高的等级分。在那次对局的第一盘，学习了卡斯帕罗夫过去棋谱的深蓝执白先行，并且先声夺人赢下了这一盘。这让全世界感到震惊，虽然大家觉得计算机最终可能在国际象棋上战胜人类的冠军，但是这一天来得比绝大部分人预料得要早。不过，善于应变的卡斯帕罗夫在随后的五盘棋中没有再输，最后以3.5：1.5的比分战胜了深蓝。对于这次比赛，媒体认为一方面深蓝的表现足够好了，虽然在总比分上它输了，但这毕竟是计算机第一次在国际象棋上战胜人类的冠军；然而，另一方面，计算机还不够聪明，它不仅缺乏应变能力，而且还会出现低级的错误[15]。因此，大家的结论是计算机在下国际象棋方面全面超过人类还有待时日。

但是，时隔一年，1997年5月，经过改进后的深蓝卷土重来。

这一次比赛还是六盘决胜负，不过第一盘是由卡斯帕罗夫执白先行。卡斯帕罗夫以自己熟悉的王翼印度进攻开局[16]，然后牢牢把握住先行的优势。到了第44步时，深蓝走出了一步非常怪异的棋，这让卡斯帕罗夫误以为计算机有了超级智能。当然，他还是平稳地走出了第45步，并且让深蓝放弃认输了。事后IBM承认，这步怪棋其实是源于程序的一个bug，使得深蓝找不到合适的走法，而采用了预先设定的保守走法。深蓝虽然输了第一盘，但是给卡斯帕罗夫在心理上造成了压力，因为他不知道计算机到底有多么聪明。后来斯维尔评论道，能够不按

常规行事其实是超级智能的表现。

第二盘由深蓝先行，它走了常见的洛普兹开局[17],双方行棋平稳，但是进行到残局时，深蓝又走出了一步非常规的走法，在45步后，卡斯帕罗夫想不出破解的方法，推盘认输了。那时候，他观棋的朋友告诉他实际上这盘棋还有救，能够走成和局。不过，今天一些国际象棋下得最好的计算机，比如Stockfish [18]，能够在深蓝那局棋的基础上，在各种应变的情况下都获胜。因此，那一盘棋卡斯帕罗夫输得并不冤。

在接下来的三盘里，双方下成和棋，其中在第四盘卡斯帕罗夫因为用时过多，被迫弈和;第五盘卡斯帕罗夫在盘面占优的情况下被深蓝逼和。在这两盘棋中，深蓝显示出了超强的计算能力。应该讲在前五盘中，双方发挥正常。

到了第六盘，卡斯帕罗夫在开局时采用了他第四盘的下法—卡罗-康防御[19]，这是执黑的棋手为了抵消后手劣势采用的一种迅速简化棋盘、拼比残局实力的走法。但是，深蓝没有重复第四盘的走法，通过大胆弃马攻破了卡斯帕罗夫的防线。这一盘只下了20多手，卡斯帕罗夫还没等到进入残局就认输了，这比通常国际象棋的进度短了一半。

从我描述的这个过程来看，似乎计算机已经足够聪明了，以至于卡斯帕罗夫拿它没有办法——它甚至像人一样会做出一些想象不到的反应。当时的媒体对深蓝的评论也是这样的，以至于IBM的股票都因此而飙升。但在这看似聪明的表象背后，其实是大量的数据、并不算复杂的算法和超强计算能力的结合——深蓝从来没有，也不需要像人一样思考。

IBM其实在1996年那次对弈之前，就收集了所有能够找到的卡斯帕罗夫的对弈记录。IBM深蓝小组所做的事情，就是利用这些数据建立了一些模型。具体的做法如下：

计算机利用数学模型，能够在棋盘的任何一个状态下，比如说某个状态叫作S，评估出自己和对方获胜的概率为 $P(S)$ 。当它要考虑接下来可能的走法，比如说有N种[20]走法时，先要考察这些走法分别对应状态，假设是A、B、C.....心，计算

出相应的获胜概率

$P(A)$ 、..... $P(S, N)$ 。根据这些概率，深蓝找出一个让自己获胜概率最大的状态，我们不妨假设是

它就往这个方向走。接下来，该

对方走棋了，对方走出一步棋后棋盘进入一个新的状态 S'' 。这时深蓝再根据自己能够选择的有限种走法，假如这回是 M 种，分别对应状态

S_1'' 、 S_2'' S_M''

，再计算出每一个对应的新状态的胜率

$P(S_1'')$ 、 $P(S_2'')$ $P(S_M'')$ ，然后挑一个产生最大胜率的走法，比如是

$P(S_2'')$ ，如图2.14所示。

当然，深蓝在评估自己和对方的胜率时，会根据历史的数据考虑卡斯帕罗夫可能采用的走法，对不同的状态给出可能性的估计，然后根据对方下一步走法对盘面的影响，核实这些可能性的估计，找到一个最有利于自己的状态，并走出这步棋。因此，深蓝的团队其实把一个机器智能的问题变成了一个大数据的问题和大量计算的问题。顺便提一句，AlphaGo在具体的算法上和深蓝略有差异，但是它们博弈的原理是相同的。

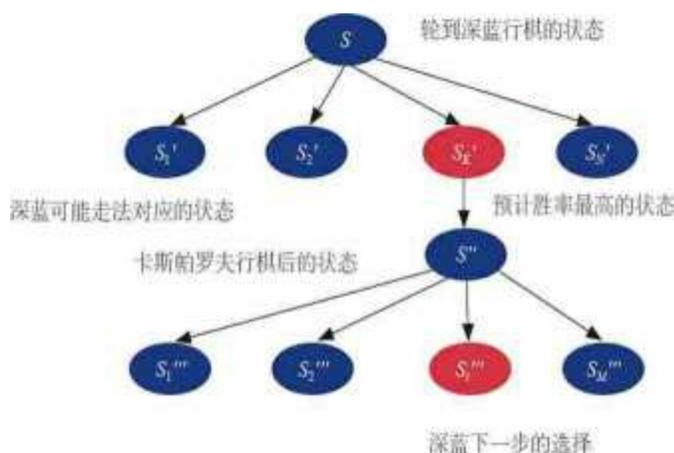


图2.14计算机下棋的博弈图

在1996年的那次对弈中，深蓝的团队研究了卡斯帕罗夫的历史数据，对他的棋风还是颇有了解的，如果卡斯帕罗夫按照通常的习惯走，深蓝应该是能够应付的。这或许是深蓝能够出奇制胜第一盘的原因。但是，深蓝使用的数据量显然不够，因此卡斯帕罗夫稍微变着数，深蓝就处于被动状态。到了1997年，深蓝团队不仅把计算机的速度提升了两个数量级，而且召集了全世界上百位国际大师[21]，收集和整理全世界各位大师的对弈棋谱，供计算机学习。这样一来，深蓝其实看到了名家们在各种局面下的走法，或者说人类能够想到的各种好棋，它都见识过了，这就具备了大数据的完备性。在第二次六局对弈中，除了第一盘深蓝因为bug最后负于卡斯帕罗夫，最后五盘非胜即平，一些走法甚至出乎卡斯帕罗夫的意料，也就是说，深蓝看过的棋局其实已经超过了后者。此外，作为机器，深蓝还具有卡斯帕罗夫所不具备的另一个优势，那就是不受情绪的影响，发挥可以相对稳定。这个性质在很多智能应用中至关重要。

自1997年之后，计算机下棋的本领越来越高，而且进步速度超出人们的想

象。今天在国际象棋上，任何人都无法与好的计算机抗衡了。按照早期对机器智能的定义，如果计算机能够在国际象棋上超过人，就说明它有了智能。然而尽管如此，大部分人，包括围棋界和科技界的权威人士，在2015年年底仍然认为AlphaGo还达不到顶级围棋手的水平。但是2016年1月，一卜360战胜了人类的欧洲围棋冠军樊麾二段。2016年3月，AlphaGo再次用事实证明了它的水平已超过人类的顶级高手——它与韩国著名棋手李世石九段进行了五番棋比赛，结果以4：1大胜，震惊世界围棋界和科技界。关于AlphaGo的具体算法，我们在后面介绍深度学习时再详细讨论。

机器智能真的在很多方面已经超过了人类?大家对此的看法不一。一些人

持否定看法的，他们习惯于把计算机已经完成的问题归结到非智能问题中。在过去，当计算机能够识别语音并理解其含义时，这个问题也从智能问题中被删除出去了。当计算机战胜人类的象棋冠军后，他们会说计算机还不会下围棋;当计算机在围棋上也表现卓越时，他们就把下棋这件事由过去的智能问题改成了计算问题。当然，虽然机器的智能在不断地提高，但总是有几件事情一直做得不好，因此人类还可以很自豪地说自己的智能水平比机器高。

在计算机尚未做好的事情中，回答那些需要进行推理的复杂问题或许可以算是一种。比如计算机是否能够回答“夏天为什么比冬天热”这样的问题。在计算机自动问答研究领域，科学家们已经研究了多年。通常我们把问题归结为7类：“是什么”(What)、“什么时候”(When)、“什么地点”(Where)、“哪一个”(Which)、“是谁”(Who)、“为什么”(Why)和“怎么做”(How)。由于它们都是以W或者H开头的，这7个疑问词又被称为WH单词(WH words)，各种问题也被称为WH语句。在这7类问题中，容易回答的是询问事实，包括“是什么”(What)、“什么时候”，(When)、“什么地点”(Where)、“哪一个”(Which)和“是谁”(Who)，比如：“中国的国家主席是谁？”难回答的是询问原因的“为什么”(Why)问题，以及询问过程的“怎么做”(How)问题。全世界的自然语言处理专家和机器智能专家对这两类问题的机器自动问答研究了很多年，直到2012年，都没有找到好的

方法。

2012年，我离开腾讯回到Google，我的上级领导辛格博士和尤斯塔斯对我讲，不指望我做什么马上见成效的产品，希望我解决一些和机器智能有关的根本性问题，前提是这些问题解决之后，微软要花5年时间才能追赶上。我花了一个多月的时间在公司里寻找要解决的问题。当时Google的云计算平台和大数据平台已经搭建得非常完善了，自然语言处理的基础工作（比如所有网页中主要语言每一句话都做了句法分析）都已经完成，对前五类简单问题的回答在林德康博士的领导下已经做得非常完善了。但是，还没有人触及对复杂问题的回答，因为大家都觉得这件事情太难，以前学术界几十个研究所、上百名一流的科学家

都没能解决这个问题。

不过，根据我对Google基础条件和数据准备情况的考察，发现如果换一个思路来解决计算机回答复杂问题的难题，就有可能另辟蹊径解决或者至少部分解决这个难题。当我把这个想法告诉辛格博士时，他的第一反应是“如果其他公司和研究所做不到，我们是否有一些别人没有的条件，使得我们能做到”，我回答他说，是数据。接下来我向他介绍说，可以将这个智能问题变成一个大数据的问题。

我们解决问题的方法是这样的：

第一步，根据网页确定哪些用户在Google问过的复杂问题可以回答，而

哪些回答不了。根据我们的研究发现，大约 70%~80%的问题，在 Google 第一页搜索结果中都有答案。大家如果想要验证这一点，不妨做一个简单的实验：在 Google、必应 (Bing) 或者百度问一个“为什么”的问题，比如问“天为什么是蓝色的”或者“为什么夏天比冬天热”，然后打开上述搜索引擎给出的前10条搜索对应的网页，通常都能找到想要的答案。但是，如果只看这些搜索引擎的摘要，只有 20%~30%的问题的答案正好在摘要中。这实际上反映出在2012年的时候，计算机与人在理解问题和回答问题上的差异。那么如果我们把目标设定在只回答那些在网页中存在答案的问题，我们其实就具备了大数据的完备性。

第二步，就是把问题和网页中的每一句话一一匹配，挑出那些可能是答案

的片段，至于怎么挑，就要依靠机器学习了。

第三步，就是利用自然语言处理技术，把答案的片段合成为一个完整的段

落。

听了我的介绍，辛格博士觉得这个道路似乎走得通，于是我们在山景城很快就成立了一个团队来开发计算机回答复杂问题的原型系统。

出于保密的考虑，我在这里不便透露我们做法的细节。简单地讲，我们建立起了一个由世界各地科学家和工程师组成的联合团队，按照大数据处理的思路，经过两年的努力，使得计算机能够回答 30% 的复杂问题，包括“天为什么是蓝色的”，“为什么夏天比冬天热”，或者“怎样烤蛋糕”之类的问题，我们将计算机产生的答案和人回答的答案拿给测评人评估，对于大部分问题的答案，测评人无法判断机器产生的答案与人回答的哪个更准确、更好。按照当年图灵博士的定义，我们实际上已经让计算机具有了某种等同于人类的智能。

GO^ why is the sky blue

Books More» Search tools

About 309,000,000 results (0.42 seconds) A clear cloudless day-time sky is blue because molecules in the air scatter blue light from the sun more than

they scatter red light. When we look towards the sun at sunset, we see red and orange colours because the blue light has been scattered out and away from the line of sight Why is the sky Blue?

math.ucf.edu/.../BlueSky/blue_sky.html University of California, Riverside
图2.15 Google自动问答（问题为“天为什么是蓝色的”，问题下面是计算机产生的答案）

计算机下棋和回答问题，体现出大数据对机器智能的决定作用。我们在后面会看到很多各种各样的机器人，比如 Google自动驾驶汽车、能够诊断癌症或者为报纸写文章的计算机，它们不需要像科幻电影里的机器人那样长着人形，但是它们都在某个方面具有超过人类的智能。在这些机器人的背后，是数据中心强大的服务器集群，而从方法上讲，它们获得智能的方法不是和我们人一样靠推理，而更多的是利用大数据，从数据中学习获得信息和知识。如今，这一场由大数据引发的改变世界的革命已经悄然发生，我们在后面的几章会更深入地介绍它。这次技术革命的特点是机器的智能化，因此我们称之为智能革命也毫不为过。

我们对大数据重要性的认识不应该停留在统计、改进产品和销售，或者提供决策的支持上，而应该看到它(和摩尔定律、数学模型一起)导致了机器智能的产生。而机器一旦产生和人类类似的智能，就将对人类社会产生重大的影响。毫不夸张地讲，决定今后20年经济发展的是大数据和由之而来的智能革命。

注释

[1] 关于图灵机，请参阅拙著《文明之光》第三册第十八章“计算的时代”。

[2] 诺威格本人也是数据驱动方法的倡导者之一，但是他和罗素所编写的教科书依然花了大量的篇幅介绍传统的人工智能。

[3] 在美国的大学里，教授每7~10年左右的时间可以带全薪休假半年，或者带半薪休假一年，这被称为学术休假。在此期间，大部分教授会选择到合作单位做一些科研，以拓宽自己的视野，另一些教授则选择找一个地方去写书。

[4] 旧M的早期系统只能识别孤立语音，在连续语音识别上，李开复的斯芬克斯(Sphinx)系统领先于旧M的同类系统。

[5] Peter. Brown at el,/I

Statistical approach to

Machine Translation ,

Computational

Linguistics , vol 16, no 2,

[9] 在机器翻译、语音识别和图像识别等领域，依靠技术进步大约每年可以改进0.5%左右。

[10] 奥科于2004年4月28日 Google宣布上市的当天加盟Google，但是随后请假回南加州大学完成教学任务，直到放暑假才正式开始在Google上班。

[11]简单地讲，N元模型是考虑 N个单词前后的关联，六元模型就是考虑6个单词，而大家当时普遍使用的三元模型只考虑3个单词。

[12] Reid J. Robison, How big is the human genome ?

[13] 大量人工统计的数据的处理量是非常大的，耗时也很长。在美国历史上，常常出现人口普查结果10年还统计不完的情况，为了解决这个难题,才催生出旧M公司。

[14] 缅因州和内布拉斯加州除外，这两个州是按照州内选区分配选举人票数。

[15] 后来发现这个低级的错误是程序的bug (漏洞) 导致的。

[16] 国际象棋中最常见的开局之一，先行的一方先将王前面的兵跳两步，然后用后兵上前一步保护王兵，这种开局进攻性很强。

[17]西班牙的一种开局法，虽然也是先将王前的兵跳两步，但是接下来

以王翼的马跳上去保护，然后出象，这种 开局能够以最短时间实现王车易位，相 对攻守平衡。

[18] 如今这些计算机在国际象棋 上能够轻松战胜任何人。

[19] 卡罗-康 (Caro-Kann) 防御是由两位德国棋手卡罗和康共同创 立而 得名。它的开局的思路是，黑方避开 各种复杂的变化，经过兑子快速 过渡到 中残局，然后拼比后半盘的棋力。

[20] 对于国际象棋，这些可能性 并不多。

[21] 国际象棋的最高等级是国际 特级大师 (Grandmaster,等级分为 2500 以上)，其次是国际大师 (Master, 等级分2400以上) ○

思维的革命 在无法确定因果关系时，数据为我们 提供了解决问题的新 方法，数据中所 包含的信息可以帮助我们消除不确 定性，而数据之间 的相关性在某种程 度上可以取代原来的因果关系，帮助 我们得到我们 想知道的答案，这便是 大数据思维的核心。

在上一章我们从技术的层面分析了 大数据为什么如此重要，尤其是在 机器 智能方面的应用，机器智能的革命将导 致计算机在越来越多的领 域超过人类， 并最终让我们的社会发生天翻地覆的变 化。在后面的章 节中我们会进一步描述 大数据对社会的影响。在这一章，我们着

重分析大数据重要性的另一个方面，即 在方法论的层面，大数据是一 种全新的 思维方式。按照大数据的思维方式，我们 做事情的方式与方 法需要从根本上改

变。

要说清楚大数据思维的重要性，需 要先回顾一下自17世纪以来一直指 导我 们日常做事行为的先前最重要的一种思 维方式——机械思维。虽 然有些希望速 成的读者认为我们没有必要把篇幅花在 描述那些历史性的 知识和结论上，但是 如果我们要想在“道”的层面了解大数 据，了解 一种新的思维方式的重要性，而 不仅仅是将自己的追求停留在“术”的层 面，那么我们就需要了解人类认识世界 方法的演变和发展过程。

今天说起机械思维，很多人马上想 到的是死板、僵化，觉得非常落

伍，甚至“机械”本身都算不上什么好词。但是在两个世纪之前，这可是一个时髦的词，就如同今天我们说互联网思维、大数据思维很时髦一样。可以毫不夸张地讲，在过去的三个多世纪里，机械思维可以算得上是人类总结出的最重要的思维方式，也是现代文明的基础。今天，很多人的行为方式和思维方式其实依然没有摆脱机械思维，尽管他们嘴上谈论的是更时髦的概念。那么，机械思维是如何产生的？为什么它的影响力能够延伸至今，它和我们将要讨论的大数据思维又有什么关联和本质区别呢？我们不妨把目光投回2000年之前。

思维方式决定科学成就：从欧几里得、托勒密到牛顿

机械思维的形成可以追溯至古希腊。欧洲之所以能够在科学上领先于世界其他地方，在很大程度上是依靠从古希腊建立起来的思辨的思想和逻辑推理的能力，依靠它们可以从实践中总结出最基本的公理，然后通过因果逻辑构建起整个科学的大厦。其中最具有代表性的是欧几里得的几何学和托勒密的地心说。

欧几里得最大的成就不是发现了那个几何定理，而是在人类所积累起来的几何学和数学知识的基础上，创立了基于公理化体系的几何学。人类对几何学的知识，在欧几里得之前就已经积累了几千年，比如在古埃及、美索不达米亚和古代中国的文明中，人们就已经知道勾股定律。但是当时世界上其他任何文明都没有建立起公理化体系的知识结构，因此对世界的了解免不了支离破碎。在欧几里得公理化的几何学中，他首先总结出5条简单得不能再简单而且相互独立的公设（Five Axioms）[1]，也就是说任何一条公理都无法从另外4条中推导出来，而且这5条公理本身是不证自明的。接下来几何学的一切定理都由定义和简单得无法证明的5条公理直接（仅以公理和定义为前提）或者间接地（除了公理和定义，还可以使用已经证明的定理）演绎得出。

欧几里得将他的公理化体系几何学写成了一本书，名为《几何原本》，这也是对世界影响力最大的一本书。欧几里得的这种基于逻辑推理的公理化系统不仅为几何学、数学和自然科学后来的发展奠定了基础，而且对西方人的整个思维方式方法都有极大的影响。甚至在法学界，整个罗马法都是建立在类似于欧几里得公理系统这样的基础上的，当然罗马法里面的公理不是几何学的，而是自然法[2]——所有的法律都可以从自然法中演绎出来。

在欧几里得之后大约5个世纪，古希腊罗马时代最伟大的天文学家托勒密将欧几里得的这种方法论应用到天文学上，建立起一套完整、严格而且相当精确的描述天体运动规律的理论体系，即地心说。讲到托勒密要顺便提一句，有些时候，一些好心人建议我将书中“最伟大”之类的词改成“最伟大的之一”，以免犯错误，或者他人有异议。其实，写书表达思想是一件颇为主观的事情，最重要的不是避免犯错误，而是不可缺乏思想。在我看来，托勒密在近代之前是当之无愧的最伟大的天文学家，没有之一。除了地心说，托勒密的贡献还包括：发明了球坐标

（我们今天还在用），定义了包括赤道 和零度经线在内的经纬线（今天的地图 就是这么划的），提出了黄道，发明了弧 度制，等等。这些贡献随便拎出_条，都足以让托勒密名垂青史。

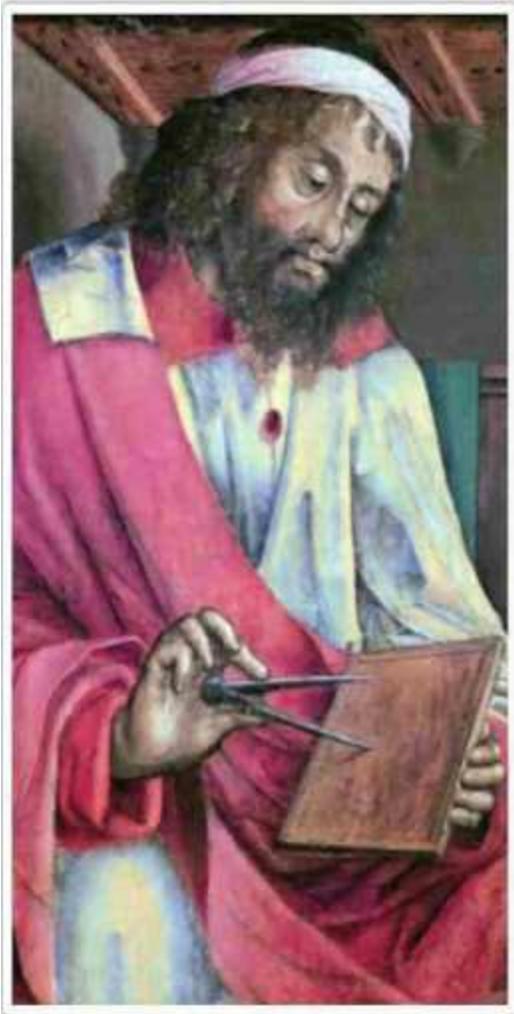


图3.1创立公理化系统的欧几里得

和欧几里得一样，托勒密不仅是一个构建大系统的人，也是一个善于总结

方法论的人。托勒密的方法论可以被概括为“通过观察获得数学模型的雏形，然后利用数据来细化模型”。托勒密的成就首先得益于过去上百年来天文观察数据，其次受益于欧几里得和毕达哥拉斯的学说。托勒密将各种天文现象的共性，用最基本的、无法再简化的原型（Meta Model）来描述。至于原型应该是什么，托勒密认为是圆，因为毕达哥拉

斯说圆 是最完美的图形。托勒密仅仅通过圆这 种曲线，以及不同大小的圆相互嵌套，把 当时人们所知的天体运动的规律描述得 清清楚楚。至于他提出的为什么是地心 说而不是日心说，原因很简单，因为这最符合人们看到的现象——日月星辰都是 从东边升起，西边落下。

托勒密的思想影响了西方世界一千

多年，这倒不完全是因为他的地心说，而 是他这种思维方式和方法论。事实上后 来的哥白尼和伽利略依然没有摆脱托勒 密的思维方式，尽管他们相信日心说。哥 白尼只是发现如果把托勒密坐标系的中心从地球移到太阳，就可以让天体运动 的模型简单一些，但是他依然需要采用 托勒密多个圆相互嵌套的模型。伽利略 在科学上比哥白尼进步了很多，事实上 真正让人们相信日心说的是伽利略，而 不是哥白尼（或者布鲁诺）[3]。但是， 即便是伽利略，其研究方法和托勒密也 如出一辙。

应该讲，托勒密等人的方法虽然很 朴素，但是很管用，直到今天，我们在做 事情的时候还是会首先想到这种方法， 比如几乎所有经济学家的理论，都是按

照这种方法提出来的。如果我们把他们 的方法论做一个简单的概括，其核心思想有如下两点:首先，需要有一个简单的 元模型，这个模型可能是假设出来的，然 后再用这个元模型构建复杂的模型；其 次，整个模型要和历史数据相吻合。这在 今天动态规划管理学上还被广泛地使用，其核心思想和托勒密的方法论是一 致的。

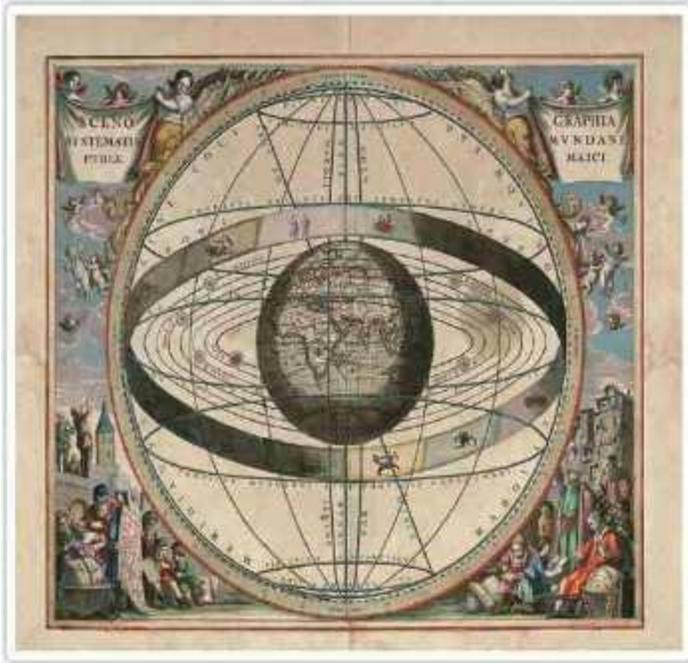


图3.2托勒密的地心说模型

思维方式和方法远不如方法论对科学的发展至关重要，东方的文明长期以来在技术上领先于西方，但是在科学体系的建立上远远落后于西方，关键是输在方法论上。

不过，托勒密的方法论有两大缺陷。首先整体模型很复杂，原因是元模型用了再简单不过的圆，这么复杂的模型依靠手工计算就难以准确。不过托勒密的这种方法论在今天机器学习领域倒是很常见，比如训练AlphaGo所用的Google大脑，就是简单的人工神经网络在几万台服务器上复杂的实现。托勒密方法论的第二缺陷是致命的，那就是确定性假设。它假定模型一旦产生，就是确定的和不会改变的。机械论延续了这种先验假设。托勒密的地心说模型和过去的观测数据吻合得天衣无缝，但是对未来的预测还是有微小的误差的，而这个误差无法被修正。这个无法被修正的细微的误差积累上千年后，一年就要差出10天时间，以至于后来预测农时极不准确，于是教皇格里高利十三世不得不让日期一次性地跳过10天。当然这些瑕疵无损托

勒密的伟大。

在古希腊罗马以后，人类对自然界的认识进步非常缓慢，西方进入了中世纪的黑暗时代。东方的中国和阿拉伯帝国虽然在工程和技术上不

断进步，但是既没有形成科学体系，也没有在方法论方面做出太多的贡献。最终，发展科学方法的任务留给了笛卡儿和牛顿。笛卡儿的贡献在于提出了科学的方法论，即大胆假设，小心求证，这个方法论在我们今天的工作中还在使用。不过对近代社会思想贡献最大的还是著名科学家和思想家牛顿。

西方人对牛顿评价之高是强调官本位的中国人难以想象的。牛顿去世后，被葬在威斯敏斯特教堂(又称为西敏寺)里最显眼的地方，其墓碑建筑远远超过包括伊丽莎白一世在内的英国任何一位君主，每天到那里拜谒的人不计其数。在大部分中国人看来，牛顿不过是一个科学家，而且他的理论今天看起来也颇为简单，为什么会如此受敬重呢？因为在欧美人看来，牛顿不仅是一位杰出的科学家，而且是人类历史上最重要的思想家之一。牛顿甚至被一些历史学家认为是人类历史上第二具有影响力的人物，不仅排在爱因斯坦等所有的科学家之前，而且超过了耶稣和孔子。牛顿通过他在数学、物理学、天文学和光学等诸多领域开创性的成绩，总结出一种全新的方法论，不仅开创了科学的时代、理性的时代，而且开启了西方的近代社会。

牛顿最直接的贡献，在于他用简单

而优美的数学公式破解了自然之谜。牛顿在他的巨著《自然哲学之数学原理》(简称《原理》)一书中，用几个简明的公式(力学三定律和万有引力定律)破解了宇宙中万物运动的规律，用微积分的概念把数学从静止的变量拓展为连续变化函数。在他的《光学》一书中，他把看上去虚幻的光分解为单个原色。

牛顿通过自己的伟大成就宣告了科学时代的来临，作为思想家，他让人们相信世界万物的运动变化规律是可以被认识的。他告诉人们：世界万物是运动的，而且这些运动遵循着确定性的规律，这些规律又是可以被认识的。牛顿的这些发现，给人类带来了从未有过的自信。在牛顿之前，人类对自己能否认识自然是缺乏信心的，那些我们今天看似不需要解释的自然现象，比如苹果为什么会落地，日月星辰为什么升起又落下，在当时却是无法被人们认识的，因此人类对自然恐惧而迷信。直到牛顿出现，人们才开始摆脱这种在大自然面前被动的状态，能够主动地应用科学来把握未来。与牛顿同时代的大科学家哈雷利用牛顿提出的原理，计算出了彗星围绕太阳运转的周期，以及彗星每一次造访地球的时间，这颗彗星后来就用他的名字命名了。后人利用牛顿的理论，能够精确地预测出1000年后出现日食和月食的时间，

第二，因为有确定性做保障，因此规律不仅是可以被认识的，而且可以用简单的公式或者语言描述清楚。这一点在牛顿之前，大部分人并不认可，而是简单地把规律归结为神的作用。

第三，这些规律应该是放之四海而皆准的，可以应用到各种未知领域指导实践，这种认识是在牛顿之后才有的。

这些其实是机械思维中积极的本

质。

工业革命，机械思维的结果

机械思维直接带来工业大发明的时代。

虽然牛顿本人就利用光学原理发明了牛顿天文望远镜，并且因此当选为英国皇家学会会员，但是第一个自觉应用牛顿力学原理做出重大发明的是伟大的发明家瓦特。我们常说瓦特发明了蒸汽机，其实蒸汽机在瓦特之前就有了，更准确的说法应该是瓦特改进了蒸汽机，或者说瓦特发明了一种万用蒸汽机。在18世纪时，英国的一些矿井使用的是非常笨拙、适用性差、效率低下的纽卡门蒸汽机。虽然纽卡门蒸汽机有诸多缺点，但是半个世纪的时间里都没有人能够改进它——这不是因为工匠们不想改进，而是他们不知道怎样改进。在牛顿和瓦特之前，一项技术的进步需要非常长的时间来积累经验，或者用今天的话讲就是获得数据、信息和知识，这个过程常常要持续经过很多代人。

瓦特和他之前的工匠都不同，他是通过科学原理直接改进蒸汽机，而不是靠长期经验的积累。虽然各种励志的读物把他描写成没有上过大学的人，但实际上他系统地学习过大学物理的课程和高等数学的很多内容。瓦特从20岁出头就在格拉斯哥大学工作，利用工作之便，他在那里听了力学、数学和物理学的课程，并与教授们讨论理论和技术问题。瓦特改进蒸汽机的大部分理论工作都是在这所大学里完成的。后来瓦特离开了大学，和工厂主博尔顿一起专心发明新的、适合各种场合的蒸汽机，因此瓦特蒸汽机也被称为万用蒸汽机。

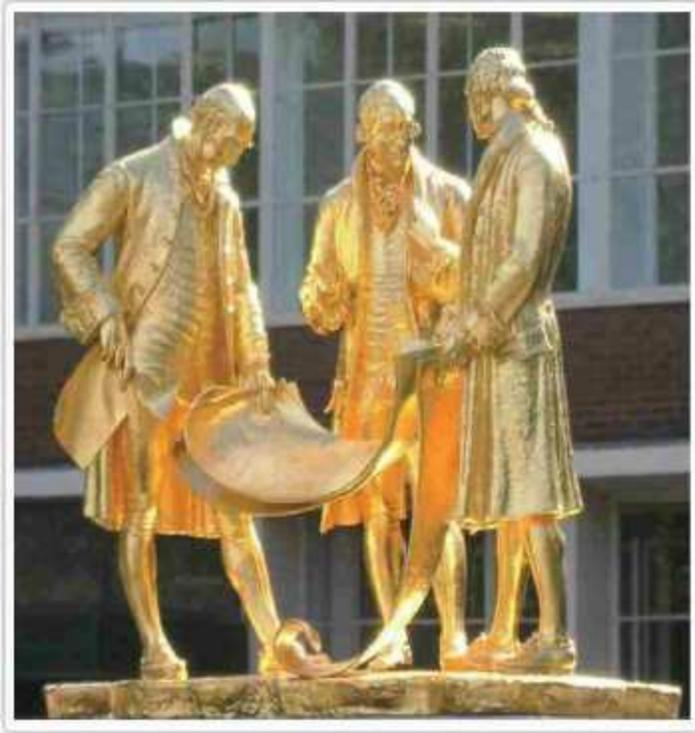


图3.4 (从左到右) 博尔顿、瓦特和他们的助手默多克（位于英国伯明翰市中）

瓦特还发明了一种通用的机器用以解决所有的问题。在瓦特之前的蒸汽机是为特定目的设计和制造的，很难从一个厂矿拆下来用于其他地方。瓦特的蒸汽机的通用性则要好很多，同一种蒸汽机可以卖到不同的工厂。这也是机械思维的重要特征——所有问题有一个通用的解决方法。瓦特的合伙人博尔顿对通用性的重要性有着先见之明，他明确地指出，他和瓦特所做的事情是为工业提供动力，而不简简单单是一种机器。

正是因为瓦特蒸汽机的这个特性，才使得工业革命后有了“蒸汽机+现有产业=新产业”的模式。博尔顿和瓦特在月光社[4]的朋友、后来的瓷器大王韦奇伍德，将瓦特蒸汽机用于瓷器的制造，这是世界上第一个采用蒸汽机动力的行业。蒸汽机的使用，使得在全世界一千多年里供不应求的瓷器，从此出现了供大于求的情况。在此之后，工业革命导致全世界财富量迅速增长。后人这样评价牛顿和瓦特这两位英国的杰出人物：牛顿找到了开启工业革命大门的钥匙，而瓦特拿着这把钥匙开启了工业革命的大门。

瓦特的成功不仅是技术的胜利，更重要的是他掌握了新的方法论——机械思维。在瓦特之后，机械思维在欧洲开始普及，工匠们发明了解决各种问题的机械。19世纪初，英国技师史蒂芬森利用机械发明了火车，并且在1825年实现了英国斯托克顿和达灵顿之间的铁路连接，从此人类之间的距离开始大大地缩短。1843年，英国发明家查尔斯·瑟伯 (Charles Thurber, 1803—1886) 第一次用机械的方式实现了替代手写字的转轮打字机，从此几千年来人类通过书写来记录文明的方式，被一种机械运动取代了。在工业革命前夕，机械思维从英国传到了大西洋彼岸的美国，一位毫无工作经验的耶鲁机械学毕业生伊莱·惠特尼 (Eli Whitney, 1765-1825)，利用自己所学习到的物理学知识和机械原理发明了轧棉机，把过去要用手工技巧摘除棉花里的棉籽的工作交给了机器来完成。轧棉机使得摘棉籽的效率提高了50倍以上，并因此彻底改变了美国南方种植园经济，间接地导致了后来的美国南北战争。和惠特尼同年出生的美国发明家罗伯特·富尔顿 (Robert

Fulton, 1765—1815) 则发明了使用机械动力取代风力的蒸汽船，为全球自由贸易时代的到来做好了准备。

机械的广泛使用和机械的思维方式直接导致了人类迄今为止最为伟大的事件——工业革命。在工业革命之前的两千年里，世界各地人们的生活水平其实没有太大的提高。已故著名历史学家安格斯·麦迪森 (Angus Maddison, 1926—2010) 对全球各个文明在不同历史时期所做的经济学研究发现，世界人均财富从公元元年左右到18世纪工业革命前是没有提高的[5]。但是，到了工业革命之后，情况就大不相同了。马克思曾经讲过：“资产阶级在其不到100年的阶级统治中所创造的生产力，比过去一切时代创造的全部生产力还要多，还要大。”[6]相比工业革命，任何王侯将相所谓的丰功伟绩都显得微不足道。

工业革命带来的不仅是财富，也大大延长了人类的寿命。在工业革命之前，无论是欧洲、东亚还是印度，人均寿命都在30-40岁之刚，因此古人才会有“人生七十古来稀”之叹。而在1800年之后，世界各国的人均寿命都先后翻了一番（见图3.5）。由此可见，一种新的思维方式对人类文明进步的重要性。

图3.5 世界各地人均寿命在当地开始工业革命之后大幅提高

机械思维对世界的影响力并没有随着工业革命的结束而结束，从牛顿时代开始接下来的3个世纪里，人类越来越习惯于用机械的方式描述一

切，这就如同在托勒密的时代人们习惯于把一切运动归结为圆周运动一样。机械思维从此渗透到社会生活的方方面面，人们相信能够用机械解决一切问题，包括很多过去无法解决的问题。

瑞士的能工巧匠们将机械的威力发挥到了极致，他们制造的那些精致而昂贵的机械表不仅可以指示时间，而且可以准确地预测上百年的太阳历、阴历和主要星辰的运动，甚至可以通过机械振

动演奏音乐。



图3.6能够奏出音乐的雅典表 (Ulysse Nardin)

不仅时间、音乐与机械挂上了钩，计算也可以用机械来实现。在19世纪中叶，发明家巴贝奇用机械实现了复杂的差分计算，70年后的20世纪30年代，德国计算机科学家和机械师楚泽则用机械实现了制造人类第一台可编程的计算机Z1。

在当时人们的眼里，世界上任何事情都是可以用机械来实现的，只是时间早晚

而已。

机械思维更广泛的影响力是作为一种准则指导人们的行为，其核心思想可以概括成确定性(或者可预测性)和因果关系。牛顿可以把所有天体运动的规律用几个定律讲清楚，并且应用到任何场合都是正确的，这就是确定性。类似地，当我们给物体施加一个外力时，它就获得一个加速度，而加速度的大小取决于外力和物体本身的质量，这是一种因

果关系。没有这些确定性和因果关系，我们就无法认识世界。

如同我们今天在谈论大数据思维和 互联网思维时无意中会带有一种优越感

一样，在19世纪时，机械思维是一个非常 时髦的词汇，人们喜欢用这个词汇表示 自己对近代科技的了解和所具有的理性 精神。在客观上，机械思维也确实促进了 世界近代化，乃至现代化的过程——它 导致了 很多重大的发明和发现，比如爱 因斯坦的相对论的提出，也促进了一些 现代科学的诞生，比如现代医药学。

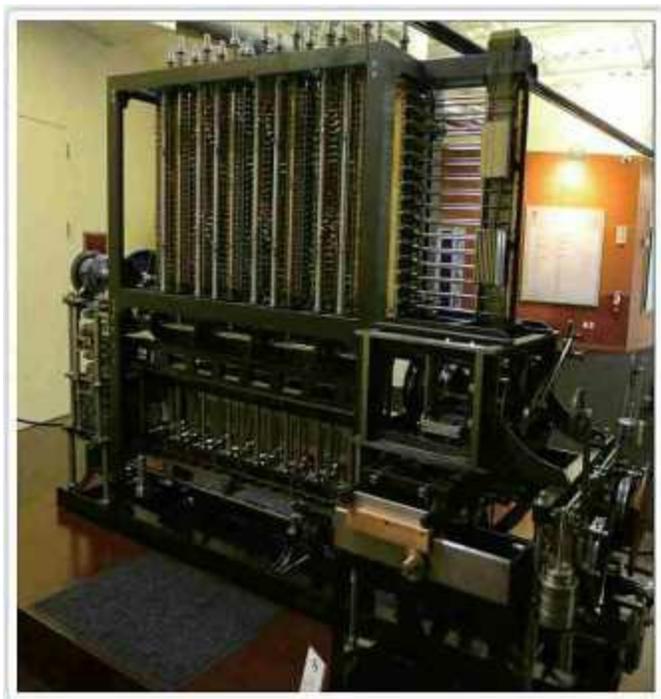


图3.7巴贝奇的差分机(硅谷计算机博物馆的复制品)

要理解机械思维深远的影响力，就 必须谈谈爱因斯坦。大家都知道，爱因斯 坦是现代物理学的集大成者，他不仅在 物理学上突破了牛顿理论，而且在物理 学几乎每个领域都有所建树，但是他的 思维方式其实和牛顿是一致的。牛顿的 物理学理论是建立在确定性基础，即所 谓的绝对时空[7]之上的，他发现万有 引力定律则是寻找因果关系的结果。牛 顿发现行星围绕太阳运动这个结果，然 后找到了万有引力这个原因。爱因斯坦 的研究方式是类似的，他的理论也是建 立在一种确定性 一一光速恒定的基础之 上的，基于这种假设，利用逻辑推理，就 可以

推导出整个狭义相对论。就连爱因斯坦自己也说，如果不是他，也会有人在很短的时间内发现狭义相对论，因为狭义相对论就是光速恒定的必然结果。类似地，如果将重力和加速度等价起来，利用因果逻辑，就能推导出广义相对论。爱因斯坦的相对论在形式上和牛顿力学也有相似之处，简单而美妙，几个公式就把整个理论描述清楚了。

至于牛顿和爱因斯坦能找到这些因果关系的原因，除了拥有过人的智慧之外，他们的运气还特别好，或者说都曾有过灵光一闪的灵感。如果说牛顿被苹果砸了一下的说法是伏尔泰杜撰出来的，并不靠谱，那么爱因斯坦从白日梦中获得另类想法搞清楚了广义相对论却是一件真实的事情。当年，爱因斯坦在瑞士专利局无所事事时，坐在窗前看着外面明媚的阳光，想着有人在窗外坐着椅子从天上加速而下的怪事，从此想清楚了重力和加速度的联系，发现了广义相对论。这个例子说明，人类找到真正的因果关系是一件很难的事情，里面运气的成分很大，因此机械思维在认识世界时还是有很多的局限性的。

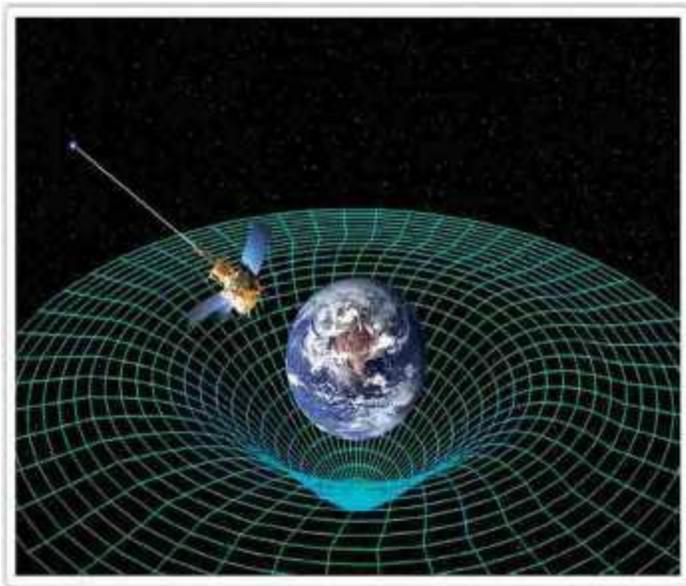


图3.8和牛顿万有引力原理不同的是，

爱因斯坦在广义相对论中用引力场解释引力现象

当然，机械思维的局限性更多来源于它否认不确定性和不可知性。爱因斯坦有句名言——“上帝不掷色子”，这是他在和量子力学的发明人波尔等人争论时讲的话。今天我们知道，在这场争论中，波尔等人是正确的，爱因斯坦错了，上帝也掷色子。著名物理学家张首晟教授喜

欢用三个公式概括人类最高的文明成就：

爱因斯坦的质能转换公式 $E = mc^2$ 量子力学中的测不准原理 $\Delta t \Delta p > s$
熵的定义 $H = -\sum p_i \log p_i$

张教授把波尔和爱因斯坦的公式同时放上去了，反映出机械思维的两面性——善于把握确定性而难以解决不确定性问题。

张首晟教授也让我给出三个公式。有两个公式我们是不约而同想到的，即质能转换和熵的定义。但是和张首晟教授略有不同的是，我用一个更简单基本的公式 $1 + 1 = 2$ 取代了测不准原理。张首晟教授是著名的物理学家，他喜欢物理学的原理，而我对数学更感兴趣，给出了这个最为朴素的公式，因为它不仅是整个数学的基础，而且概括了因果逻辑，从大前提和小前提，一定能够得到确定的结论。反过来看，要想让结果被人们接受，就必须知道原因。这是从笛卡儿开始总结出科学方法以来全世界科学家们都必须遵守的原则。利用这个方法论，在“二战”之前，人类可谓是无往不利，世界上许多发明就是在这样的方法论下产生的。在这些发明中，青霉素的发明不仅非常重要，而且极具代表性。

青霉素对人类的重要性无须多言，它不仅仅是一种抗生素，能够杀菌治病，而且在很大程度上消除了人类对疾病的恐惧，从此不必生活在对疾病的恐惧中。在青霉素被发明和使用之前，不论是东方人还是西方人，一旦得了病，能否治好很大程度上只有听天由命。我们今天无法想象天天生活在对疾病和死亡的恐惧中是怎样的感觉，但是半个多世纪前人类就是生活在对未来不确定的阴影中。青霉素改变了这一切，因此它的发明过程被无数文学作品过度地渲染也就不足

为奇了。

对青霉素最戏剧化的渲染就是将青霉素的发明过程归结为：英国医生亚历山大·弗莱明（Alexander Fleming, 1881—1955）在1928年很幸运地发现霉菌可以杀死细菌，从而发明了这种万灵药。但真实情况要复杂得多，弗莱明的偶然发现仅仅是发明青霉素漫长过程的一个开始而已，其重要性也被远远地夸大了。事实上，弗莱明并不清楚霉菌杀菌的原理，也没有能力浓缩和提炼其中的有效成分，如果仅仅靠他偶然的发现，青霉素的普及不知道要晚多少年。

青霉素真正得以从偶然的发现变成一种万灵药，在很大程度上是科学家们自觉应用因果逻辑的结果。在制药这个行业，直到今天其核心的方法都遵循“研究病理找到真正致病的原因，然后针对这个原因找到解决方案”。世界上最早真正采用科学方法研究青霉素杀菌原理和提炼青霉素的，是霍华德■弗洛里 (Howard Florey, 1898—1968)和厄恩斯特■钱恩 (Ernst Chain, 1906—1979)等人，当时已经是1939年，距离弗莱明首次发现青霉素已经过去11年了，而弗莱明本人也已经不再研究青霉素。钱恩和他的同事爱德华■彭利■亚伯拉罕 (Edward Penley Abraham, 1913—1999)等人找到了青霉素的有效成分——一种被称为青霉烷的物质。青霉烷能够破坏细菌的细胞壁，而人和动物的细胞没有细胞壁，青霉素可以杀死细菌却不会伤害人和动物，这样才算搞清楚了青霉素杀菌的原理。后来根据这个原理，美国麻省理工学院的科学家约翰■希恩 (John Sheehan, 1915-1992)成功地合成出青霉素，而不再像过去那样需要通过培养霉菌的方法提炼这种药物了。[8]同时，了解了青霉素的杀菌原理，也有助于科学家们搞清楚为什么某些细菌会产生抗药性[9]，亚伯拉罕等人再应用青霉烷的杀菌原理，发明了头孢类的抗生素等多种新型抗生素，解决抗药性问题。青霉素和其他抗生素的发明，实际上遵循了“分析找到原因，根据原因得到结果”的思维方式，或者说知其然也知其所以然。这种方法带来的好处是有目共睹的，工业革命后人类寿命的提高都是依靠这种方法。相反，传统医学常常不遵循因果关系，是“不知其所以然”，因此治病的效果也是时好时坏，然后医生们用一些似是而非的语

言解释他们其实并没有搞清楚的原因。

从牛顿开始，人类社会的进步在很大程度上得益于机械思维，但是到了信息时代，它的局限性也越来越明显。首先，并非所有的规律都可以用简单的原理描述；其次，像过去那样找到因果关系已经变得非常困难，因为简单的因果关系规律性都被发现了。另外，随着人类对世界认识得越来越清楚，人们发现世界本身存在着很大的不确定性，并非如过去想象的那样一切都是可以确定的。因此，在现代社会里，人们开始考虑在承认不确定性的情况下如何取得科学上的突破，或者把事情做得更好。这也就导致一种新的方法论诞生。



图3.9因发明青霉素而获得诺贝尔奖的三名科学家:弗莱明、弗洛里和钱恩(从左至右)

世界的不确定性

不确定性在我们的世界里无处不在。我们经常可以看到这样一种怪现象，很多时候专家们对未来各种趋势的预测是错的，这在金融领域尤其常见。如果读者有心统计一些经济学家们对未来的看法，就会发现它们基本上是对错各一半。这并不是因为他们缺乏专业知识，而是由于不确定性是这个世界的重要特征，以至于我们按照传统的方法——机械论的方法难以做出准确的预测。

世界的不确定性来自两方面，首先是当我们对这个世界的方方面面了解得越来越细致之后，会发现影响世界的变量其实非常多，已经无法通过简单的办法或者公式算出结果，因此我们宁愿采用一些针对随机事件的方法来处理它们，人为地把它们归为不确定的一类。

我们可以通过下面的例子来理解这种不确定性。如果我们在平整的桌子上掷一次色子，在色子落到桌子上停稳以前，我们一般都认为无法知道到底哪一面朝上，哪一面朝下。但是其实在色子离开手的一瞬间，如果能够知道色子准确的形状和密度分布、出手的力量和旋转的角速度、空气流动的速度，同时我们的计算足够精确，其实我们是能够算出色子的哪个点或者哪个面接触到桌面的。如果我们还知道桌面的弹性系数和色子的弹性系数，以及这两种材质的物理性质等因素，我们就能够算出这个色子弹起来多高、运动的方向等，最终可以算出它停下来时哪一面朝上。但是，由于这里面很多细节难以准确测量，比如出手的速度和力量，因此考虑了所有的因素后计算出来的结果也未必正确。在这种情况下，一般人干脆假定色子每一面朝上的概率都是 $1/6$ 。

回到前面提到的预测股市，各种专家预测的准确性大抵在一半左右，这和掷色子的道理很相似。美国政府和一些研究所公布的各种经济数据多达两万个，最好的经济学家一辈子能够研究透的经济指标不到它们的1%（当然他们认为很多数据并不重要），有太多的不确定因素是他们考虑不到的，因此他们无法准确预测市场也就不奇怪了。美国各大投资机构出于对利润的考虑，利用计算机尽可能地考虑了各种经济数据的影响，但是最终预测的准确性依然在50%左右，这是因为人们对这些因素的测量也未必准确。事实上，美国大部分基金的投资回报率并没有市场的平均值高。

不确定性的第二个因素来自客观世界本身，它是宇宙的一个特性。在宏观世界里，行星围绕恒星运动的速度和位置是可以计算得很准确的，从而可以画出它的运动轨迹。但是在微观世界里，电子在围绕原子核做高速运动时，我们不可能同时准确地测定出它在某时刻的位置和运动速度，当然也就不能描绘出它的运动轨迹了。这并非我们的仪器不够准确，而是因为这是原子本身的特性。在量子力学中有一个测不准原理，也就是说，像电子这样的基本粒子的位置的测量误差和动量的测量误差的乘积不可能无限小。这与机械思维所认定的世界的确定性是相违背的。为什么会有这样的现象存在呢？因为我们测量活动本身影响了被测量的结果。对于股市上的操作也类似，当有人按照某个理论买或者卖股票时，其实给股市带来了一个相反的推动力，这导致股市在微观上的走向和理论预测的方向相反。

如果世界充满了不确定性，我们对

未来世界的认识是否又回到了牛顿之前的不可知状态？答案是否定的。就拿微观世界的电子运动来说，虽然我们无法确定电子的准确位置和速度，但是能够知道它在一定时间内在核外空间各处出现的概率，因此科学家们用一种密度模型来描述电子的运动。在这个模型里，密度大的地方，表明电子在那里出现的机会多，反之，则表明电子出现的机会少。这个模型很像在原子核外有一层密度不等的“云”，因此也被形象地称为“电子云”。在现实生活中情况也是类似的，不论是因为数据量太大导致的不确定性，还是因为世界本身带有的不确定性，总之，世界上很多事情是难以用确定的公式或者规则来表示的。但是，它们并非没有规律可循，通常可以用概率模型来描述。在概率论的基础上，香农博士建立起一套完整的理论，将世界的不确定性和信息联系起来，这就是信息论。信息论不仅仅是通信的理论，也给了人们

一种看待世界和处理问题的新思路。

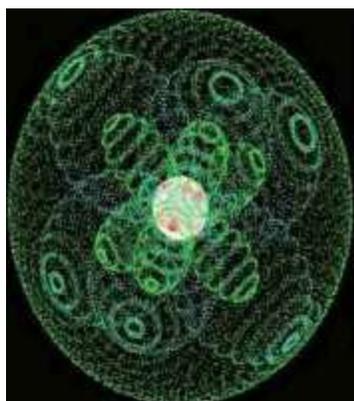


图3.10我们无法刚棚准电子的位置 和动量，只能计算出它们的分布,因此电子就如同散布在原子核之外的云，也被 称为电子云

信息论最初是通信的理论。信息这个词如今我们每天都能够听到，有时我们会用信息量大、信息量小这类说法，但是到底有多少信息算是信息量大，其实 很多人并没有仔细地想过。我们进一步 刨根问底，信息是否能够被量化地度量？如果能，又应该怎么度量，大部分人对这个问题并不清楚。当然，脑筋快的人会马上想到，既然信息和数据有直接的联系， 能否以数据量来表示信息量，因为数据 量很容易度量。应该讲数据量有些时候 可能和信息量有点关系，但是两者不能 画等号。比如，一本50多万中文字的《史 纪》和两本80万英文单词的《圣经■旧约》和《圣经■新约》，谁的信息量更大？

这似乎不是由篇幅和字数来决定的。再 比如，大家都明白,看似大量却不断重复 的数据，其实里面的信息量是很少的。

那么如何度量信息呢？这个问题其 实是几千年来很多人想知道却无法回答 的问题。直到1948年，克劳迪·香农在 他著名的论文《通信的数学原理》（A Mathematic Theory of Communication)中提出了“信息熵”的概念，才解决了对信息的度量问题，并 且量化地给出了信息的作用。同时，香农 还把信息和世界的不确定性，或者说无 序状态联系到了一起。

首先意识到无序状态这个问题的是 奥地利物理学家路德维希·玻尔兹曼(Ludwig Boltzmann, 1844—1906) ○ 他发现一个封闭容器内的微观状态的有序程度，即每个原子的位置和动量，与这 个容器内气体的热力学

性质有关。在玻尔兹曼之前，制作蒸汽机的工程师们已经发现了热力学第二定律[10]，其中鲁道夫·克劳修斯（Rudolf Clausius）提出了一种叫作“熵”的概念，来描述一个系统中趋向于恒温的程度。当这个系统完全达到恒温时，就无法做功了，这时熵最大。但是在玻尔兹曼之前的工程师和科学家们都没能解释其中的原因。玻尔兹曼则把熵（宏观特性Entropy）和封闭系统的无序状态（每一个分子的微观特性Q）联系起来，即：

$$E = k \log(Q)$$

其中k被称为玻尔兹曼常数。玻尔兹

曼等人还发现，在一个封闭的系统中，熵永远是朝着不断增加的方向发展的，也就是说从微观上讲，这个系统越来越无序，从宏观上看它趋于恒温。

••

•参 | ••

•••

参

图3.11两个容器中，左边的气体温度低，右边的温度高，处于一种有序状态，熵的值较低，混合之后，变成无序状态，熵增加

香农在信息论中借用了热力学里熵

的概念，他用熵来描述一个信息系统的 不确定性。接下来香农指出，信息量与不确定性有关：假如我们需要搞清楚一件 非常不确定的事，或是我们_无所知的事情，就需要了解大量的信息。相反，如果我们对某件事已经有了较多的了解，那么不需要太多的信息就能把它搞清楚。所以，从这个角度来看，可以认为，信息量的度量就等于不确定性的多少，这样香农就把熵和信息量联系起来。他还指出要想消除系统内的不确定性，就要引入信息。

信息论最初是关于通信的理论。人类进入文明社会之后，除了吃饭睡觉之外，大部分时间其实都在做和通信有关的事情，我们在工作中讨

论问题、开会、写邮件，平时和家人聊天，闲暇之余看书、读报、看电视、看电影，都是某种形式的通信，而通信所传输的是某种信息。在科学上，香农的贡献在于第一次量化地度量信息，并且用数学的方法将通信的原理解释得一清二楚。

虽然香农提出信息论最初的目的只是建立通信的科学理论，但是，信息论的作用远不止在科学和工程上——它也是一种全新的方法论。与机械思维是建立在一种确定性的基础上所截然不同的，信息论完全是建立在不确定性基础上，而要想消除这种不确定性，就要引入信息。至于要引入多少信息，则要看系统中的不确定性有多大。这种思路成为信息时代做事情的根本方法。我们不妨用互联网广告的例子来说明上述原理的作用。

在我们对用户一无所知的情况下，在网页上投放展示广告，点击率非常低，每1000次展示也只能赚不到0.5美元的广告费，因为这等于随机猜测用户的需求，很不准确。如果有10万种广告，只有10种与用户相关，那么猜中的可能性就是万分之一。如果用信息论的方法来度量，它的不确定性为14比特左右[11]。搜索广告因为有用户输入的关键词，准确率会大幅提高，至于提高了多少，取决于关键词所提供的信息量。以汉字词为例，如果一个搜索输入了两个词，每个词平均两个汉字，那么大约能提供10-12比特的信息量，这样大部分不确定性就消除了。假定还是从10万种广告中猜10个，这时猜中的可能性就是十几分之一到几分之一，因此读者点击广告的可能性大增。在实际情况中，Google搜索广告每1000次展示所带来的收入大约是50美元，比展示广告高出两个数量级。这就说明了信息的作用。类似地，我们大致计算出，像Facebook或者Google通过挖掘注册用户的使用习惯，大约能够获得1-2比特的信息量，这样就将广告匹配的难度下降了大约一半，事实上，那些与用户相关的展示广告比完全随机的正好产生高一倍左右的广告收入。

上面虽然是一个特定的例子，但是反映出在信息时代的方法论：谁掌握了信息，谁就能够获取财富，这就如同在工业时代，谁掌握了资本谁就能获取财富一样。

当然，用不确定性这种眼光看待世界，再用信息消除不确定性，不仅能够赚钱，而且能够把很多智能型的问题转化成信息处理的问题，具体说，就是利用信息来消除不确定性的问题。比如下象棋，每一种情况都有几种可能，却难以决定最终的选择，这就是不确定性的表现。再

比如要识别一个人脸的图像，实际上可以看成是从有限种可能性中挑出一种，因为全世界的人数是有限的，这也就把识别问题变成了消除不确定性的问题。我们在前面一章里讲到了贾里尼克等人的工作，从那时开始，人类在机器智能领域的成就，其实就是不断地把各种智能问题转化成消除不确定性的问题，然后再找到能够消除相应不确定性的信息，如此而已。

我们在利用信息时使用的很多原理和方法，在信息论中都能找到根据。比如

用信息论中的一个重要概念——互信息 (Mutual Information),可以解释为什么信息的相关性可以帮助我们解决很多问题。在很多时候，我们能够获取的信息和要研究的事物并非一回事，它们之间必须“有关联”，所获得的信息才能帮助我们消除不确定性，搞清楚我们想要研究的问题。比如前面提到的王进喜的照片和大庆油田的位置、产量等情报就属于有关联。当然“有关联”这种说法太模糊，不科学，最好能够量化地度量两件事之间的“相关性”。为此，在信息论里用互信息这个概念，实现了对相关性的量化度量。比如通过对大数据文本进行统计就会发现，“央行调整利率”和“股市短期浮动”的互信息很大，这证实了它们之间有非常强的相关性。而“央行调整利率”和“北京机场大量航班晚点”的互信息则接近于零，说明二者没有什么相关性，甚至无关。

香农除了给出对信息和互信息的量化度量之外，还给出了两个相关信息处理和通信的最基本的定律，即香农第一定律和香农第二定律。这两个定律对于信息时代的作用堪比牛顿力学定律对机械时代的作用。

香农第一定律，也称为香农信源编码定律，它大致的含义是这样的:假定有一个信息源，里面有N种信息，现在我们需要对这N种信息一一进行编码，比如我们用0011表示第一种信息，1000111表示第二种.....这些编码当然不能重复，否则我们就无法根据编码来断定是哪一种信息了。虽然编码可以有很多种方法，但是有的方法效率低，有的则效率低，或者说用了很长的编码才能表示一个信息。香农第一定律讲的是，对于信源发出的所有信息设计一种编码，那么编码的平均长度一定大于该信源的信息熵，但同时香农还指出，一定存在一种编码方式，使得编码的平均长度无限接近于它的信息熵。

对于没有学过信息论的读者而言，上面这段话可能有点费解，让我们看一个具体的例子就好理解了。比如要对汉字编码，有些字用得

有些字用得少，因此可以把常用字的编码做得短些，生僻字的编码做得长些，但是不论怎么做，编码的平均长度一定会超过汉字的不确定性，即它们的信息熵，这是香农第一定律的第一层意思。同时，香农第一定律还有第二层意思，也就是说一定存在一种（最优的）编码方法，使得每个汉字的平均编码长度可以非常接近它的不确定性（信息熵）。至于怎么才能做到，霍夫曼 (Huffman) 给了一个非常简单的方法——只要把最短的编码分配给最常见的汉字即可。这种编码方法具有通用性，又称为霍夫曼编码，它可以被认为是对香农第一定律的补充。

香农第一定律不仅是现代通信的基础，也代表了一种新的方法论。经济学上的吉尔德定律 (Gilder's Law)，即尽量多地采用便宜的资源，尽可能节省贵的资源，与信息论中的霍夫曼编码从本质上讲是相同的。在信息时代，由于摩尔定律的作用，计算机是便宜的资源，而且越来越便宜，人力成本则会越来越局，因此聪明的公司懂得利用计算机来取代人的工作，像Google或者Facebook这样的公司，都是尽可能地将越来越多的事情交给机器去做，而不是雇用很多人。在过去的半个世纪里，生产力的提高实际上就是靠用便宜的机器取代人工，这种做法有意无意地和信息论的原理相符合。当然，也有的企业主不愿意在IT方面进行投入而坚持使用人工，因为这种投入在初期看上去显得比人工昂贵，这些企业后来就逐渐地被淘汰了。

在信息论中，还有香农第二定律，通俗地讲就是信息的传播速率不可能超过信道的容量，这和我们的现实生活也是契合的。我们经历了互联网发展全过程的这一代人都有这样一种体会，互联网发展的各个阶段实际上是建立在不断拓

宽带宽的基础之上的。早期，我们使用电话调制解调器，然后开始使用DSL(数字用户线路)，再到后来使用宽带电缆，最后到光纤，都是围绕着不断增加信道容量而进行的，只有信道的容量增加了，传输率才能上去，我们才能从阅读文字，到看图片，到看视频，再到看高清视频，整个互联网才能得到发展。在香农提出他的第二定律之后，人类就开始有意识地不断扩展带宽。

•历史数据 .预細数据

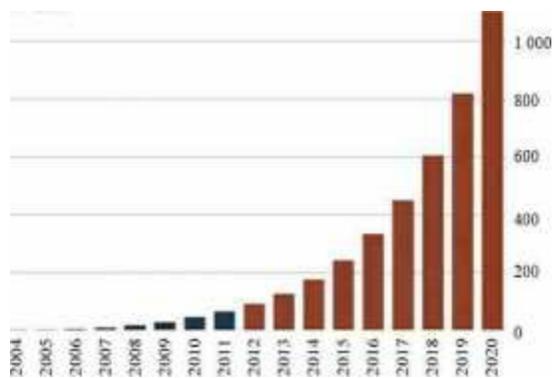


图3.12 2002-2020年全球互联网的 带宽呈指数级增长

香农第二定律不仅描述了通信领域 最根本的规律，而且它是自然界本身所 固有的规律，能够解释很多商业行为。比 如我们常说做生意要靠人脉，其实这个 人脉就是人与人交往的带宽。如果人脉 不够，发出的信息和获得的信息都有限， 生意一定做不大。现代通信手段的本质，就是以相对低廉的成本让人们获得人 脉，而媒体行业的不断进步，本质上是不 断地在为企业拓宽对外连接的带宽，使 得它们做生意越来越方便。

关于信息论，还有一个原理必须了 解，那就是最大熵原理。这个原理的大意 是说，当我们要对未知的事件寻找一个 概率模型时，这个模型应当满足我们所 有已经看到的数据，但是对未知的情况 不要做任何主观假设。在很多领域，尤其 是金融领域，采用最大熵原理要比任何 人为假定的理论更有效，因此它被广泛 地用于机器学习。最大熵原理实际上已 经不同于我们使用了几百年的“大胆假 设、小心求证”的方法论，因为它要求不 引入主观的假设。当然，不做主观假设的 前提是取得了足够多的数据，否则最大 熵模型只能给出一些平均值而已，而 不能对任何细节进行描述和预测。[12]

今天，信息论已经被广泛地用于管 理，因为它为我们提供了信息时代的方 方法论。而熵这个词，也成了信息论和不确 定性的代名词。也正是 因为如此，张首晟 教授和我都认为它代表了人类对我们的

世界认知度的最高境界。

大数据的本质

有了信息论这样一个工具和方法 论，我们便很容易认清大数据的本质了。首先我们必须承认世界的不确定性，这样我们就不会采用确定性的思维方式去 面对一个不确定性的世界。当我们了解 到信息或者说数据能够消除不确定性之 后，便能理解为什么大数据的出现能够 解决那些智能的问题，因为很多智能问 题从根本上来讲无非是消除不确定性的 问题。对于前面提到的大数据的三个特 征，即数据量大、多维度和完备性，我们 可以从信息论出发，对它们的重要性和 必要性一一做出解释。在这个基础之上， 我们就能够讲清楚大数据的本质。

先谈谈数据量的问题。在过去，由于 数据量不够，即使使用了数据，依然不足 以消除不确定性，因此数据的作用其实 很有限，很多人忽视它的重要性是必然 的。在那种情况下，哪个领域先积攒下足 够多的数据，它的研究进展就显得快一 些。具体到机器智能方面，语音识别是 最早获得比较多数据的领域，因此数据驱 动的方法从这个领域产生也就不足为奇 了。

关于大数据多维度的重要性问题， 可以从两个角度来看待它。第一个视角 是前面提及的“互信息”，为了获得相关 性通常需要多个维度的信息。比如我们 要统计“央行调整利息”和“股市波动” 的相关性，只有历史上央行调整利息一 个维度的信息显然是不够的，需要上述 两个维度的信息同时出现。第二个视角 是所谓的“交叉验证”，我们不妨看这样 一个例子:夏天的时候，如果我们感觉很 闷热，就知道可能要下雨了。也就是说，“空气湿度较高”和“24小时内要下雨” 之间的互信息较大。但是，这件事并非很 确定，因为有些时候湿度大却没有下雨。 不过，如果结合气压信息、云图信息等其 他维度的信息，也能验证“24小时内要 下雨”这件事，那么预测的准确性就要大很 多。因此，大数据多维度的重要性，也 是有信息论做理论基础的。

最后，我们从信息论的角度来看看 数据完备性的重要性。在说明这件事情 之前，我们还需要介绍信息论里一个重 要的概念——交叉熵，这个概念并非由 香农提出的，而是由库尔贝 克-莱伯勒距离 (Kullback-Leibler Divergence),它可以反映两个信息源 之间的一致性，或者两种概率模型 之间 的一致性。当两个数据源完全一致时，它 们的交叉熵等于零，当

它们相差很大时，交叉熵也很大。所有采用数据驱动的方法，建立模型所使用的数据和使用模型的数据之间需要有一致性，也就是盖洛普所讲的代表性，否则这种方法就会失效，而交叉熵就是对这种代表性或者一致性的一种精确的量化度量。

回过头来讲大数据的完备性。在过去，使用任何基于概率统计的模型都会有很多小概率事件覆盖不到，这在过去被认为是数据驱动方法的死穴。很多学科把这种现象称为“黑天鹅效应” [13]。在大数据出来之前，这件事是无法避免的，就连提出数据驱动方法的鼻祖贾里尼克也认为，不论统计数据量多大，都会有漏网的情况。这些漏网的情况反映到交叉熵时，它的值会达到无穷大，也就是说数据驱动方法在这个时候就失效了。

怎样才能防止出现很多漏网的情况呢？这就要求大数据的完备性了。在大数据时代，在某个领域里获得数据的完备性还是可能的。比如在过去把全国所有人的面孔收集全是一件不可想象的事情，但是今天这件事情完全能做到。当数据的完备性具备了之后，就相当于训练模型的数据集合和使用这个模型的测试集合是同一个集合，或者是高度重复的，这样，它们的交叉熵近乎零。在这种情况下，就不会出现覆盖不了很多小概率事件的灾难。这样数据驱动才具有普遍性，

而不再是时灵时不灵的方法论。

由此可见，大数据的科学基础是信息论，它的本质就是利用信息消除不确定性。虽然人类使用信息由来已久，但是到了大数据时代，量变带来质变，以至于人们忽然发现，采用信息论的思维方式可以让过去很多难题迎刃而解。

从因果关系到强相关关系

逻辑推理能力是人类特有的本领，给出原因，我们能够通过逻辑推理得到结果。在过去，我们一直非常强调因果关系，一方面是因为我们常常是先有原因，再有结果，另一方面是因为如果我们找不出原因，常常会觉得结果不是非常可信。比如在过去，现代医学里新药的研制，就是典型的利用因果关系解决问题的例子。

我们在前面讲到的青霉素的发明过程就非常具有代表性。首先，在19世纪中期，奥匈帝国的塞麦尔维斯（Ignaz Philipp Semmelweis, 1818—1865）[14]、法国的巴斯德等人发现微生物细菌会导致很多疾病，因此人们很容易想

到杀死细菌就能治好疾病，这就是因果关系。不过，后来弗莱明等人发现，把消毒剂涂抹在伤员伤口上并不管用，因此就要寻找能够从人体内杀菌的物质。最终在1928年弗莱明发现了青霉素，但是他不知道青霉素杀菌的原理。而牛津大学的科学家钱恩和亚伯拉罕搞清楚了青霉素中的一种物质——青霉烷——能够破坏细菌的细胞壁，才算搞清楚青霉素有效性的原因，到这时青霉素治疗疾病的因果关系才算完全找到，这时已经是1943年，离塞麦尔维斯发现细菌致病已经过去近一个世纪。两年之后，女科学家多萝西·霍奇金(Dorothy Hodgkin)搞清楚了青霉烷的分子结构，并因此获得了诺贝尔奖，这样到了1957年终于可以人工合成青霉素[15]。当然，搞清楚青霉烷的分子结构，有利于人类通过改进它来发明新的抗生素，亚伯拉罕就因此而发明了头孢类抗生素。在整个青霉素和其他抗生素的发明过程中，人类就是不断地分析原因，然后寻找答案(结果)。当然，通过这种因果关系找到的答案非常让人信服。

其他新药的研制过程和青霉素很类似，科学家们通常需要分析疾病产生的原因，寻找能够消除这些原因的物质，然后合成新药。这是一个非常漫长的过程，而且费用非常高。在七八年前，5开制一种处方药已经需要花费10年以上的时间，投入10亿美元的科研经费，如今，时间和费用成本都进一步提高；一些专家，比如斯坦福医学院院长米纳（Lloyd Minor）教授则估计需要20年的时间，20亿美元的投入。这就不奇怪为什么有效的新药价格都非常昂贵，因为如果不能在专利有效期内[16]赚回20亿美元的成本，就不可能有公司愿意投钱研制新药了。

按照因果关系，研制一种新药就需要如此长的时间、如此高的成本。这显然不是患者可以等待和负担的，也不是医生、科学家、制药公司想要的，但是过去没有办法，大家只能这么做。如今，有了大数据，寻找特效药的方法就和过去有所不同了。美国一共只有5000多种处方药，人类会得的疾病大约有一万种。如果将每一种药和每一种疾病进行配对，就会发现一些意外的惊喜。比如斯坦福大学医学院发现，原来用于治疗心脏病的某种药物对治疗某种胃病特别有效。当然，为了证实这一点需要做相应的临床试验，但是这样找到治疗胃病的药只需要花费3年时间，成本也只有1亿美元。这种方法，实际上依靠的并非因果关系，而是一种强关联关系，即A药对B病有效。至于为什么有效，接下来3年的研究工作实际上就是在反过来寻找原因。这种先有结果再反推原因的做法，和过去通过因果关系推导出结果的做法截然相反。无疑，这样的做法会比较快，当然，前提是足够多的数据支持。

但是在过去，由于数据量有限，而且常常不是多维度的，这样的相关性很难找得到，即使偶尔找到了，人们也未必接受，因为这和传统的观念不一样。20世纪90年代中期，在美国和加拿大围绕香烟是否对人体有害这件事情的一系列诉讼上，如何判定吸烟是否有害是这些案子的关键，是采用因果关系判定，还是采用

相关性判定，决定了那些诉讼案判决结果。

在今天一般的人看来，吸烟对人体有害，这是板上钉钉的事实。比如美国外科协会的份研究报告显示，吸烟男性肺癌的发病率是不吸烟男性的23倍，女性则是相应的13倍[17]，这从统计学上讲早已经不是随机事件的偶然性了，而是存在必然的联系。但是，就是这样看似如山的铁证，依然“不足够”以此判定烟草公司就是有罪，因为它们认为吸烟和肺癌没有因果关系。烟草公司可以找出很多理由来辩解，比如说_些人之所以要吸烟，是因为身体里有某部分基因缺陷或者身体缺乏某种物质；而导致肺癌的，是这种基因缺陷或者某种物质的缺乏，而非烟草中的某些物质。从法律上讲，烟草公司的解释很站得住脚，美国的法律又是采用无罪推定原则[18]，因此，单纯靠发病率高这一件事是无法判定烟草公司有罪的。这就导致了在历史上很长的时间里，美国各个州政府的检察官在对烟草公司提起诉讼后，经过很长时间的法庭调查和双方的交锋，最后结果都是不了了之。其根本原因是提起诉讼的原告一方(州检察官和受害人)拿不出足够充分的证据，而烟草公司又有

足够的钱请到很好的律师为它们进行辩护。

这种情况直到20世纪90年代中期 美国历史上的那次世纪大诉讼才得到改变。1994年，密西西比州的总检察长麦克-摩尔(Michael Moore)又一次提起了对菲利普·莫里斯等烟草公司的集体诉讼，随后，美国40多个州加入了这场有史以来最大的诉讼行动。在诉讼开始以前，双方都清楚官司的胜负其实取决于各州的检察官们能否收集到让人信服的证据来证明是吸烟而不是其他原因导致了更多疾病（比如肺癌）更高的发病率。

我们在前面讲了，单纯讲吸烟者比不吸烟者肺癌的发病率高是没有用的，因为得肺癌可能是由其他更直接的因素引起的。要说明吸烟的危害，最好能找到吸烟和得病的因果关系，但是这件事情短时间内又做不到。因此，诉讼方只能退而求其次，他们必须能够提供在(烟草公司所说的)其他因素都被排除的情况下，吸烟者发病的比例依然比不吸烟者要高很多的证据，这件事做起来远比想象的困难。虽然当时全世界的人口多达60亿，吸烟者的人数也很多，得各种与吸烟有关疾病的人也不少，但是在以移民为主的美国，尤其是大城市里，人们彼此之间基因的差异相对较大，生活习惯和收入状况也千差万别，即使调查了大量吸烟和不吸烟的样本，能够进行比对的、各方面条件都很相似的样本并不多。不过在20世纪90年代的那次世纪大诉讼中，各州的检察长下定决心要打赢官司，而不再是不不了了之，为此他们聘请了包括约翰·霍普金斯大学在内的很多大学的顶级专家作为诉讼方的顾问，其中既包括医学家，也包括公共卫生专家。这些专家们为了收集证据，派下面的工作人员到世界各地，尤其是第三世界国家的农村地区（包括中国的西南地区），去收集对比数据。在这样的地区，由于族群相对单一(可以排除基因等先天的因素) [19]，收入和生活习惯相差较小（可以排除后天的因素），有可能找到足够多的可对比的样本，来说明吸烟的危害。



图3.13告倒烟草公司的密西西比州总

检察长摩尔

各州检察官们和专家们经过三年多 的努力，最终让烟草公司低头了。1997 年，烟草公司和各州达成和解，同意赔偿 3655亿美元。在这场历史性胜利的背后， 靠的并非是检察官们找到了吸烟对人体 有害的因果关系的证据，而依然是采用了统计上强相关性的证据，只是这一次的 证据能够让陪审团和法官信服。在这场马拉松式的诉讼过程中，其实 人们的思维方式已经从接受因果关系，转到接 受强相关性上来了。

如果在法律上都能够被作为证据接 受，那么把相关性的结果应用到其他领 域更是顺理成章的事情。

2003年Google推出了根据网页内 容安插广告的AdSense服务[20],以 与那

些在网页中随机投放广告的产品竞争。根据我们的直觉，如果在一个和照相机有关的网站(或者)网页中放上照相机的广告，效果应该最好。这其实就是用到了相关性的特点，但是大部分时候，相关性并不是那么直接，不能一眼就看出来。根据大量数据的统计结果，我们发现这样一些广告和内容的搭配效果非常好，很多和我们的想象不大相同，比如：

在电影租赁和收看视频的网站，放上零食的广告；

在女装网站上，放男装的广告；

在咖啡评论和销售网站上，放信用卡和房贷的广告；

在工具 (Hardware)评论网站上，放上快餐的广告；

rfr-rfr

寺芬0

这些搭配，如果没有大量的数据统计作为基础，一般人是想不到的。当然，如果仔细分析有些看似不太相关的搭配，还是能够找到合理的解释，比如电影租赁和视频播放网站与零食广告的搭配，符合人在看视频时喜欢吃零食的习惯。但是，有些搭配会让人完全摸不到头脑，比如把咖啡和信用卡或者房贷联系起来。不管是能够找到原因的，还是想不出原因的（可能背后存在着我们一时想不到的原因），只要使用了这些相关性，广告的效果就好。当然，在利用相关性时，我们希望是那种可信度比较高的，即数学上所谓的强相关性，而不是随便把一些看似相关的东西扯到一起。

我们在前面提到，能通过因果关系找到答案，根据因果关系知道原因固然好，但是对于复杂的问题，其难度非常大，除了靠物质条件、人们的努力，还要靠运气。牛顿和爱因斯坦都是运气很好的人。遗憾的是，大部分时候我们并没有灵感和运气，因此很多问题得不到解决。在大数据时代，我们能够得益于一种新的思维方法——从大量的数据中直接找到答案，即使不知道原因。这一方面给了我们一个找捷径的方法，同时我们不会因为缺乏运气而被问题难倒；另一方面，这种找不出原因的答案我们是否敢接受呢？如果我们愿意接受，那么我们的思维方式已经跳出了机械时代单纯追求因果关

系的做法，开始具有大数据思维了。

当然，这种思维方式的改变有一个过程，我们不妨以最受益于大数据的 Google 公司为例，来说明转变思维方式的重要性。

数据公司 Google

在一般人眼里，Google 是一家高科技公司，不断地研发新的技术，并且成功地将一部分技术转化成了产品。但是，它从根本上讲其实是一家数据公司。著名的机器智能专家，前 Google 研究院院长 诺威格博士对 Google 的这个本质有深刻的认识。他在接受母校(加州大学伯克利分校)授予他的荣誉证书时，曾经这样讲述他为什么要加入 Google:

2001年，当全球互联网泡沫破碎后，大家都在逃离这个领域，很多人从互联网行业回到了学术界。人们问我为什么在这样一个时候离开 NASA (美国国家航空航天局)，加入 Google 这家不大的互联网公司。我和他们讲了大萧条时期 (1929-1933年) 的一个故事。在大萧条时，有些人买了银行的股票，后来都发了财。事后人们问那些买了银行股票的人为什么在银行如此糟糕时敢买它们的股票，那些投资人讲，“因为全世界的钱都在它们那里。”所以，加入 Google 的决定并不难做，因为全世界的数据都在 Google 那里。

诺威格在 Google 负责搜索质量部门（也是我所在的部门）。在 2005 年之前，虽然我们不断地使用数据来提高搜索质量，但是主要的工作方法还是遵循因果关系。比如我们发现有些搜索结果相关性不好，那么我们需要先分析原因，再寻找答案。在那个时候，网页搜索质量可以提升的空间还比较大，靠这种方法我们每年可以将搜索质量提高 3~5 个百分点。不过随着搜索质量接近完美，再按照这样一种方式工作，每年的进步连一个百分点都到不了。但与此同时，依靠数据的积累，大家发现搜索质量和很多数据特征有很强的相关性，利用这些特性可以迅速提升搜索结果的质量。

在所有的数据中，与搜索质量相关性最高的是大量的点击数据，即对于不同的搜索关键词，用户们都点击了哪些搜索结果（网页）。比如对于“虚拟现实”这个查询，用户有 31000 次点击了网页 A，15000 次点击了网页 B，11000 次点

击了网页 C 在这种情况下，网页 A 应

该被排在第一位，但是如果搜索排序算法不好，有可能出现它没有被排在第一位的情况。这时搜索引擎的设计者就面临一个选择，是采用通过研究改进原有的排序算法，还是干脆相信用户的点击结果，或者是将它们结合在一起。如果单纯改进排序算法，这个周期特别长。如果相信用户点击的结果，其实就是用相关性取代因果关系，当然这里面有两个风险：首先是用户点击容易形成马太效应，排在前面的结果即使不是很相关，也容易获得更多的点击；其次是单纯依靠点击，搜索结果的排名容易被一些使用者操纵。因此，比较稳妥的办法是对用户的点击数据建立一个简单的模型，作为搜索排序算法的一部分。

今天，各个搜索引擎都有一个度量用户点击数据和搜索结果相关性的模型，通常被称为“点击模型”。随着数据量的积累，点击模型对搜索结果排名的预测越来越准确，它的重要性也越来越大。今天，它在搜索排序中至少占70%~80%的权重[21]，也就是说搜索算法中其他所有的因素加起来都不如它重要。换句话说，在今天的搜索引擎中，因果关系已经没有数据的相关性重要了。

当然，点击模型的准确性取决于数据量的大小。对于常见的搜索，比如“虚拟现实”，积累足够多的用户点击数据并不需要太长的时间。但是，对于那些不太常见的搜索（通常也被称为长尾搜索），比如“毕加索早期作品介绍”，需要很长的时间才能收集到“足够多的数据”来训练模型。一个搜索引擎使用的时间越长，数据的积累就越充分，对于这些长尾搜索就做得越准确。微软的搜索引擎在很长的时间里做不过Google的主要原因并不在于算法本身，而是因为缺乏数据。同样的道理，在中国，搜狗等小规模搜索引擎相对百度最大的劣势也在于数据量上。

当整个搜索行业都意识到点击数据的重要性后，这个市场上的竞争就从技术竞争变成了数据竞争。这时，各公司的商业策略和产品策略就都围绕着获取数据、建立相关性而开展了。后进入搜索市场的公司要想不坐以待毙，唯一的办法就是快速获得数据。比如微软通过接手雅虎的搜索业务，将必应的搜索量从原来Google的10%左右陡然提升到Google的20%~30%，点击模型估计得准确了许多，搜索质量迅速提高。但是即使做到这一点还是不够的，因此一些公司想出了更激进的办法，通过搜索条(Toolbar)、浏览器甚至输入法来收集用户的点击行为。这种办法的好处在于它不仅收集到用户使用该公司搜索引擎本身的点击数据，而且还能收集用户使用其他搜索引擎的数据，比如

微软通过旧浏览器收集用户使用Google搜索时的点击情况。这样来，如果一家公司能够在浏览器市场占很大的份额，即使它的搜索量很小，也能收集大量的数据。有了这些数据，尤其是用户在更好的搜索引擎上的点击数据，一家搜索引擎公司可以快速改进长尾搜索的质量。当然，有人诟病必应的这种做法是“抄”Google的搜索结果，其实它并没有直接抄，而是用Google的数据改进自己的点击模型。这种事情在中国市场上也是一样，因此，搜索质量的竞争就成了浏览器或者其他客户端软件市场占有率的竞争。虽然在外人看来这些互联网公司竞争的是技术，但更准确地讲，它们是在数据层面竞争。

在Google内，点击模型的使用标志着工作方法从传统的“遵循因果关系”，逐步变成了“寻找相关性”。今天，Google至少有13~25的工程师每天的工作就是处理数据。Google的关键词广告系统AdWords不仅是互联网世界最赚钱的产品，对广告商来讲也是广告效果最好的平台。Google是如何做到兼顾自己的利益和广告商的利益的呢？Google的销售人员对外宣传是技术好，这种说法当然没有错，但是更准确的说法是它从一开始就积累了大量的各种数据，并且善于利用数据。Google在搜索结果页投放广告时，不仅要考虑广告主的出价，还要考虑它与搜索的结果是否相关，该广告本身的质量，以及在历史上用户点击这个广告的比例。这样一来，那些不太可能产生点击的广告，或者质量不高的广告，Google就展示得很少。对广告主来讲省了钱，对Google来讲，把资源(有限而宝贵的搜索流量)留给了可能被点击的广告，收入也有所增加。更重要的是，给用户的体验要比到处放广告的网站要好很多。值得一提的是，Google的广告系统每次播放什么广告，不是由任何规则决定的，而完全是利用数据、挖掘相关性的结果。



图3.14 Google其实是一家数据公司,这是它的超级数据中心

Google和很多互联网公司之所以能够取得成功,不仅仅是靠技术,靠数据,更是靠采用了大数据时代的方法论,或者说大数据思维。作为数据公司,它们在做事情的方法上有着和传统工业公司不同的思维方式。相对来讲这些公司很少花大量的时间和资源来寻找确定的因果关系,而是通过从大量数据中挖掘相关性,直接用于产品,因此它们给外界的感觉是产品更新非常快。大数据思维对Google等公司的帮助,我们会在后面的章节里进一步介绍。

小结

很多时候,落后与先进的差距,不是购买一些机器或者引进一些技术就能够弥补的,落后最可怕的地方是思维方式的落后。西方在近代走在了世界前列,很大程度上靠的是思维方式全面领先。

机械思维曾经是改变了人类工作方式的革命性的方法论,并且在工业革命和后来全球工业化的过程中起到了决定性的作用,今天它在很多地方依然能指导我们的行动。如果我们能够找到确定性(或者可预测性)和因果关系,这依然是最好的结果。但是,今天我们面临的复杂情况,已经不是机械时代用几个定律就能讲清楚的了,不确定性,或者说难以找到确定性,是今天社会的常态。在无法确定因果关系时,数据为我们提供了解决问题的新方法,数据中所包含的信息可以帮助我们消除不确定性,而数据之间的相关性在某种程度上可以取代原来的因果关系,帮助我们得到我们想知道的答案,这便是大数据思维的核心。大数据思维和原有机械思维并非完全对立,它更多的是对后者的补充。在新的时代,一定需要新的方法论,也一定会产生新的方法论。

欧几里得几何学的五条公设 (Five Axioms):

1. 由任意一点到另外任意一点可以画直线。
2. 一条有限直线可以继续延长。
3. 以任意点为心及任意的距离[22]可以画圆。

4. 凡直角都彼此相等。

5. 平面内一条直线和另外两条直线 相交，若在某一侧的两个内角的和小于二直角的和，则这二直线经无限延长后 在这一侧相交。[23]

欧几里得几何学的五条公理 (Five Notions):

1. 等于同量[24]的量彼此相等。
2. 等量加等量，其和仍相等。
3. 等量减等量，其差仍相等。
4. 彼此能重合的物体是全等[25] 的。
5. 整体大于部分。

注释

[1] Axiom应该翻译成公理，但 是早期《几何原本》就译成了公设，因此我们沿用这种习惯。具体内容参见附录

[2] 关于对古希腊科学和罗马法 的更详细的内容，读者朋友可以参阅拙著《文明之光》第一册。

[3] 伽利略发现木星的4颗卫星 后，他告诉人们在地球以外的天体也可以成为一个中心，这才否认了地球的独特性，进而让人们相信日心说。

[4] 月光社是当时在英国伯明翰 的一个小的学术圈，成员包括博尔顿、老 达尔文(查尔斯■达尔文的爷爷)、瓦特、韦奇伍德、约瑟夫·普里斯特里 (Joseph Priestley,发现了氧气助燃原 理) 等，以及通信会员法国的拉瓦锡、美 国的富兰克林和杰弗逊。月光社对整个 欧美的工业革命产生了巨大的影响，18 世纪英国的名人传记中或多或少都会提

到月光社。

[5] 详见本书第七章。

[6] 摘自《共产党宣言》。

[7] 时间和空间本身不随运动变

[8] 培养霉菌的方法不仅成本高，而且产量很低。

[9] 某些细菌会产生一种酶，溶解掉青霉素的有效成分。

[10] 不可能把热量从低温物体传递到高温物体而不产生其他影响。

[11] 关于信息论的基础知识，请读者参阅拙著《数学之美》。

[12] 读者朋友如果想了解最大熵原理的更多细节，可以阅读拙著《数学之美》。

[13] 在18世纪欧洲人发现澳大利亚之前，由于他们所见过的天鹅都是白色的，所以当时的欧洲人认为所有天鹅都是白色的。后来欧洲人在澳大利亚看到了黑天鹅，原来通过对白天鹅无数次观察得到的结论就失效了。因此，从以往数据得到的结论未必能反映未来的小概率事件。在科学方法上，或者经济学和社会学的研究中，“黑天鹅”隐喻那些极为罕见、在通常的预期之外的事件，它们在发生之前没有前例可以证明，但一旦发生，就会产生极端的影响。

[14] 奥匈帝国医生，在1847年发现了细菌是导致很多疾病的原因。

[15] 在此之前要靠培养霉菌提炼青霉素。

[16] 虽然美国的专利有效期长达17年，并且可以延长3年，但是因为大部分核心专利在药品进行实验时已经申请，中间有非常长的各种实验过程，等到药品上市，剩下的专利有效期通常不超过10年。

[17] The Health Consequences of

Smoking , A Report of

The US Surgeon General , 2004.

[18]意指被告的一方在法庭上先

被假定为无罪，除非有足够的证据证明 其有罪。

[19] 在人口流动性较大的地区，比如城市里或者经济发达地区，很难找到一群基因非常接近的人，而这一点在 经济不太发达、人们世代代住在一起 很少流动的地区才能做到。

[20] 今天这项服务被称为 AdSense for Content (谷歌内容广告) 。

[21] 各家搜索引擎对点击模型的 依赖权重虽然有大有小，但是都在60%以上。

[22] 原文中无“半径”二字出现， 此处“距离”即圆的半径。

[23] 这就是大家提到的欧几里得 第5公设，即现行平面几何中的平行公理 的原始等价命题。

[24] 这里的“量”与第4条公理 中的“物体”在原文中是同一个字thing。

[25] 为了区别面积相等与图形相 等，《几何原本》译者将图形“相等”译为“全等”。

第四章 大数据与商业 在未来我们可以看到,大数据和机器 智能的工具就如同水和电这样的资源，由专门的公司提供给全社会使用。

大数据思维不是抽象的，而是有一 整套方法让人们能够通过数据寻找相关性，最后解决各种各样的难题。每一个人、 每一个企业在接受大数据思维，改变做 事情的方式之后，就有可能实现一些在 过去想都不敢想的梦想。在这些梦想的 基础上，我们能够构建一个完美的商业 环境和一个更加现代化的社会。大数据 对社会的影响，涉及社会的方方面面，描 述清楚需要很长的篇幅。这一章，我们集 中讨论大数据对商业的影响，并通过一 些具体的案例，看看大数据思维是怎样 解决商业活动中所遇到的各种问题，进 而构建出一个全新的商业社会。

从大数据中找规律

当人们改变思维方式后，很多过去难以解决的问题在大数据时代可以迎刃而解。

在美国，毒品问题是一大社会毒瘤。按照一般人的想法，切断毒源就可以从根子上解决这个问题，因此过去美国把缉毒的重点放在切断来自南美洲的毒品供应上。尽管美国在这方面做得不错，但是仍然无法禁止毒品的泛滥，其中一个重要的原因就是很多提炼毒品所需的植物，比如大麻，种起来非常容易，甚至可以在自己家里种。

在马里兰州的巴尔的摩市东部，有一些废弃的房屋（见图4.1），当地一些穷人就进去把四周的门窗钉死，然后在里面偷偷用LED（发光二极管）灯种植大麻，由于周围的社区比较乱，很少有外人去那里，因此那儿就成了毒品种植者的天堂。



图4.1巴尔的摩东部贫民窟有大量废弃的住房，毒品生产者在里面偷偷种植和提炼毒品

对图4.1中这一类街区进行重点排

查是否就能解决问题呢？答案并不是那么简单。在环境优美生活水准高的西雅图地区，比如在图4.2那样的社区里，把门窗钉起来种毒品自然是行不通的，但是毒品种植者也有办法。有一家人花了50万美元买下了一栋豪宅，周围是种满了玫瑰的花园，平时很少有人来。这栋四卧两厅的大宅子其实没有人住，占据它的是里面658株盆栽的大麻。房

主每年 卖大麻的收入，不仅足够付房子的分期 付款和电费，而且还让他擴够了首付又 买了一栋房子。[1]



图4.2种植大麻的豪宅外景

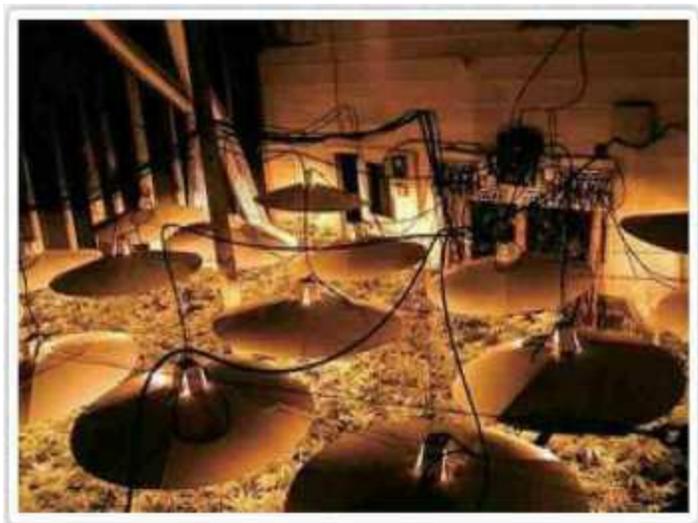


图4.3豪宅内实际上是这样的大麻种

类似的情况在美国各州和力喙大不 少地区都有发生。据估计，仅加拿大的不 列颠哥伦比亚省，每年这种盆栽大麻的 收入就高达65亿美元，在当地是仅次于 石油的第二大生意。

由于种植毒品的人分布的地域非常 广，而且做事隐秘,定位这样种植毒品的 房屋的成本非常高。再加上美国宪法的 第四修正案规定“人人具有保障人身、住 所、文件及财物的安全，不受无理之搜查 和扣押的权

利”，警察在没有证据时不得 随便进入这些房屋进行搜查。因此，过去警察虽然知道一些嫌犯可能在种植毒品，也只能望洋兴叹，这使得美国的毒品 屡禁不止。

但是到了大数据时代，私自种植毒品者的好日子就快到头了。2010年，美国各大媒体报道了这样一则新闻：

在南卡罗来纳州的多切斯特县 (Dorchester County),警察通过智能电表收集上来的各户用电情况分析，抓住了一个在家里种大麻的人。

这件事引起了美国社会的广泛讨论，当然话题除了围绕当地的供电公司爱迪斯托(Edison Electric)是否有权利 将用户的数据提供给警察之外，更多的是在探讨大数据能够帮助我们解决过去的难题，以及这项技术对社会产生的影响。不过，不论社会怎么看，我觉得倒是 该给警察们一些赞誉，他们能够在新的技术环境下改变思维方式，把过去难以

解决的问题解决好。

无独有偶，这则消息出来以后不久，媒体陆续报道出在美国其他州，警察也 用类似的方法抓到在房间里种大麻的人，[2]截至2011年，仅俄亥俄一个州，警察就抓到了 60个这样的犯罪嫌疑人。为什么最近这些年警察抓嫌疑犯的效率 一下子变得如此之高呢？因为以前供电公司使用的是老式的电表，只能记录每家每月的用电量，而从十几年前开始，美国逐渐采用智能电表取代传统的电表，这样不仅能够记录用电量，还能记录用电模式。种植大麻的房子用电模式和一般居家是不同的，只要把每家每户的用电模式和典型的居家用电模式进行对比，就能圈定一些犯罪嫌疑人。

对于查处毒品种植的案例，我们看到了大数据思维的三个亮点：第一是用 统计规律和个案对比，做到精准定位。第二是社会其实已经默认了在取证时利用 相关性代替直接证据，即我们在前面所说的强相关性代替因果关系。第三是执法的成本，或者更广泛地讲，运营的成本，在大数据时代会大幅下降。

类似的使用大数据的不仅有警察局，还有税务局。

在美国99.7%的企业是500人以下 的小企业，它们雇用的员工占了私有企业员工的一半左右，而每个小企业平均人数只有5人左右。[3]这些小

企业，尤其是涉及可以进行现金交易的零售企业（比如餐馆、商店、服务行业等），时常有偷漏税现象发生。据估计，美国每年仅偷漏的联邦税就高达3000多亿美元[4]，在最多的年份2006年是近4000亿美元。如果没有偷漏税，美国是可以避免财政赤字的。而在美国偷漏税比例最高的是小企业，因为查这些企业偷漏税的成本太局0

不过从2006年开始，美国偷漏税的金额开始下降了，这主要是因为国税局和各州州税局采用了大数据的技术，比较准确地圈定了可能偷漏税的小企业以及个人骗退税的情况。[5]后一种情况需要一些美国个人所得税的背景知识，我们略过不讲，重点看看前一种情况，即小企业偷漏税的情况。联邦和州两级税务局防止小企业偷漏税的做法其实很简单。首先，税务局将企业按照规模（场地大小）、类型和地址做一个简单的分类，比如旧金山拿骚大街上的餐馆分为一类，圣荷西第十大街上的某个理发店分为另一类等。然后，税务局根据历史的数据对每一类大致的收入和纳税情况进行分析。比如前一类餐馆每平方米的营业面积每年产生1万美元左右的营业额，整个餐馆的年收入大约是200万~280万美元左右，纳税20万美元；后一类年收入是8万~12万美元左右，纳税5000美元。如果前一类中有一家餐馆的营业面积和其他各家差不多，自称收入只有50万美元，那么就会被调查；后一类如果有一家理发店每年有10万美元的收入，只纳税1000美元，也会被调查。

在有大数据之前，我们寻找一个规律常常是很困难的，经常要经历“假

设——求证——再假设——再求证”这样一个漫长的过程，而在找到规律后，应用到个案上的成本可能也是很高的。但是，有了大数据之后，这一类问题就变得简单了。比如通过对大量数据的统计直接找到正常用电模式和纳税模式，然后圈定那些用电模式异常的大麻种植者，或者有嫌疑的偷漏税者。由于这种方法采用的是机器学习，依靠的是机器智能，大大降低了人工成本，因此执行的成本非常低。在美国有大量类似的报道[6]，在各种媒体上都可以看到。



图4.4税务部门利用大数据查处偷税

漏税

既然行政监管机构通过大数据分析可以得到它们想要得到的信息，那么商家也可以通过类似的方法做更多的生意。《纽约时报》的记者查尔斯·杜西格在2012年详细地报道了美国第二大连锁百货店塔吉特[7]用大数据做生意的事情。

2002年，塔吉特连锁百货店聘请统计学硕士安德鲁■波尔(Andrew Pole)来分析数据。在此之前，塔吉特通过信用卡号、接收发票的邮箱[8]能把某些顾客与其所购买的商品联系起来（回顾大数据的多维度特征）。但是这些数据有什么用、怎么用，塔吉特并没有考虑。波尔来了以后，就用这些数据分析用户行为。有一天市场部的同事来找他，问他能否判断一位女性顾客是否怀孕了，因为如果一个家庭有了孩子，他们的购物习惯将改变，甚至会疯狂购物，这时，百货店就可以给这些顾客推送相应商品的优惠券，牢牢把握住这些有刚需的用户。

波尔的数据分析团队经过对怀孕顾客行为的分析发现，这些女性在怀孕的不同阶段购买的东西有很大的相似性。在最初阶段，她们会购买无味的大瓶润肤油，这是因为她们会出现皮肤干燥的症状，接下来就是购买维生素和营养品，然后就是购买大包无味的香皂和棉球。等到购买婴儿用的毛巾等用品时，一般就到了快分娩的时间了。虽然每位孕妇购买的东西不完全相同，塔吉特所拥有的数据也并非完整，但是这个大趋势还是能够被系统自动归纳出来的。波尔说，如果一位女性买过大瓶椰子油润肤露、一个能装两大包尿不湿的大挎包、维生

素和鲜亮的孩子玩耍的地毯，那么 根据这看似不多的信息，就能估计出她 怀孕的可能性是87%，而且如果确实怀 孕了，那么预产期可以预测得非常准确。

依靠大量的数据，波尔团队给出的 预测还是相当准确的。塔吉特根据波尔

统计出的结论，找出25类商品，一旦确定 一个家庭有人怀孕了，就在孕妇怀孕的 不同时期向她们推送这25类商品的优惠券。利用大数据精确地做生意的做法，让 塔吉特能够在美国零售市场趋于饱和且 被电商瓜分的情况下，保持稳定的增长。 2002年，也就是波尔受聘于塔吉特的那 一年，该连锁店的营业额是440亿美元。到了 2010年，营业额则上升到670亿美 元。至于波尔的工作对此有多少贡献，塔 吉特的老板认为是非常大的，因为塔吉 特从那以后专注于给像母婴这样的特定 顾客有针对性地推荐产品。

塔吉特利用大数据的故事非常具有 代表性，它反映出大数据和未来商业的 关系。但是塔吉特的故事并没有到此结 束，接下来的事情就非常戏剧化了。接下 来的这一段内容被《福布斯》等多家媒体 不断报道和转载，因此读者可能已经读 到过了，在这里我就不赘述细节了，只是 为了便于讨论，介绍下故事的梗概[9

有一天，一位中年男子闯进明尼阿 波利斯的一家塔吉特商店，要求找他们 的经理。在见到经理后，这位男子说：“我 那个才上高中的女儿收到了这些优惠券——婴儿的衣服、婴儿的摇车等，你们 这是鼓励她过早怀孕么？”经理开始时一 头雾水，看了男子手里拿的信件的地址 和里面的优惠券，确实是他们寄出去的。 于是经理就向这位男子道歉。

几天后，这位经理又专门打电话给 这位男子，再次道歉，并且了解一下后者 对他们的处理是否满意。这回让这位经 理吃惊的是，在电话的另一端，那位男子 说：“我和女儿谈了，家里有些事情我确 实不知道，她真的怀孕了，预产期是8月。 我应该向你道歉。”

记者杜西格在他的长文中这样评论 道：“塔吉特比一个十几岁女孩的父亲先 知道他的孩子怀孕了。事实上它很清楚 顾客家庭的情况，却装作不知道。这件事 就如同跑去相亲的男女，虽然事先已经 把对方了解得一清二楚，还装作什么都 不知道。”当然，塔吉特挖掘大数据并非 为了刺探隐私，而是为了做生意,但是这 也从另一方面说明商家掌握了大数

据之后，对顾客的需求可以说是了如指掌。

相比电子商务公司，塔吉特的IT技术力量并不强，而且作为传统的连锁店，它所收集到的与用户行为相关的数据并不算多，即便如此，在使用大数据之后，它比客户的家庭更了解自家的情况。那些手握更多数据的电子商务公司，诸如亚马逊和阿里巴巴，就更可能比我们更了解自己的需求了。

巨大的商业利好:相关性、时效性和个性化的重要性

在大数据出现之前，并非我们得不到信息直接的关联性，而是需要花费很长的时间才能收集到足够多的数据，然后再花费更长的时间来验证它，这也是过去大部分传统的企业对于细节数据的收集和处理不是很重视的原因，相比之下他们更看重经验和宏观数据。但是到了大数据时代，这些企业的观念也在慢慢转变。

像沃尔玛连锁店或者梅西百货店这样传统的商店，货物的摆放是很有讲究的。这些店的货架基本上可以分为两种。第一种摆放的商品基本上是固定的，比如1~10排是药品和洗漱用品，11~15排是文具，16-20排是生活用品，等等。这类固定货架是为了方便老顾客每次能够顺利找到他们想要的东西。第二种是商店一入门的货架，摆放的是促销的、当下热门的或者与季节相关的商品，这类货架虽然数量不多，却产生了可观的营业额。但是，第二类货架该摆什么商品，过去基本上是靠经验来，而积累经验时虽然也用到数据的相关性，但是过程非常缓慢。比如沃尔玛发现在下雨天或者天气恶劣时，手电筒等应急物品卖得很好，这听起来很合理，因此沃尔玛就在坏天气来临之前把这些商品放在一入门的货架上。当然，沃尔玛也发现坏天气时一些方便早餐，比如甜甜圈和蛋糕的销量特别好，因此这些方便早餐和手电筒等应急物品可以放在一起卖。

一些人把这种相关性也看成是大数据的应用，其实它更多的是传统意义上数据的应用，因为它的规律性是慢慢被观察到的。事实上，沃尔玛在20世纪80年代就遍布美国和世界上很多国家了，但是它通过销售数据改进货物摆放搭配是到了21世纪之后的事情。

新一代的百货店做法就不同了，它们从一开始就直接利用数据提升销售。沃尔玛在20多年前每次向美国证监会提交财报时，列举的主要竞争对手是塔吉特连锁店或者Costco（好市多）仓储店，但是如今它最大的竞争对手成了网上的百货店亚马逊。亚马逊的优势倒不在于价格便宜，事实上美国实体店和网上的价格差不太多，这和中国的电商有很大区别，它的优势是能够有针对性地给用户推荐商品，这占到亚马逊销售额的1/3。为什么亚马逊能够做到这一点而沃尔玛做不到呢？这就涉及大数据的时效性等特点了。



图4.5亚马逊会把男性护肤用品和古典音乐一同推荐

亚马逊在推荐商品方面做得最成功，今天它的销售额中有1/3是靠给用户推荐而产生的。相比沃尔玛，亚马逊有三个优势，首先它的交易数据是即时而完整地记录下来的，而且是随时可以用，可以分析的，因此亚马逊挖掘到类似廉价早餐点心和应急用品的搭配只需要几个小时，而不是多少年。沃尔玛等传统的公司，虽然交易数据都是保留的，但都是支离破碎地存放在各处，有些还是存放在第三方[10]，用起来并不方便。亚马逊的第二个优势在于它拥有顾客全面的信息，比如张三上周买了一台数码相机，之前他还购买了几个玩具，同一个地址的李四前两天买了婴儿用的浴液。那么可以联想到张三和李四是一家人，他们有一个出生不久的婴儿，张三买数码相机或许是为了给孩子照相。他们或许会对在线冲印照片（并做成贺年卡），或者电子相框有兴趣。如果将他们的地址和美国个人住宅信息网站zillow.com联系起来，很容易了解到他们的住房价值，进而估计出他们的收入。这些条件是沃尔玛不具备的。亚马逊的第三个优势在于它的任何市场策略都能马上实现，比如它能够随时捆绑商品，并且随时调整价格进行促销；而美国所有的实体店，调整价格都需要在晚上关门之后进行，因此即使它们数据挖掘的速度和亚马逊一样快（当然这是不可能的），在市场上的反应也跟不上亚马逊这样的电商公司。

对比亚马逊和沃尔玛，我们能够看到大数据时效性和个性化特征带来的好处。今天，在各大电商网站上，商品数量多得已经无法靠浏览来选择。对于购买目标很明确的顾客，可以靠搜索来完成选择，但大部分人逛网店其实并没有太明确的目标。这时候，有针对性的推荐就变得特别重要了。今天，亚马逊的个性化推荐不仅能针对个人的喜好，

而且有较强的时效性。当然，亚马逊能做到今天这一步，也是靠较长时间大数据的积累。在亚马逊开始做商品推荐的初期，由于数据量不足，不得不采用不需要太大数据量的同类顾客归类的推荐方式。事实证明将顾客聚类的方式效果非常不好，最终亚马逊不得不放弃这种方式。好在随着亚马逊数据量的积累，它可以采用直接但是需要大量数据的方法，即它所谓的“由商品直接推荐商品”（Item to Item），这才使得亚马逊的推荐系统变得准确而有时效性。像沃尔玛这样的百货店，今天能做到把两类商品准确地关联起来已经很不错了，而且比过去大大地提高了营业额，但是，亚马逊却能做到两件具体的商品直接的关联。这样一来，两家商店在吸引顾客方面的差异就显而易见了。2015年7月，亚马逊的市值超过了沃尔玛，这标志着一个新时代的到来——以大数据为基础的电子商务将超越传统的零售商业。后者并非不能利用大数据，只是在个性化和时效性等方面，很难做得像电子商务公司那么有效而已。

美国在线电影、电视租赁公司 Netflix（网飞）在业务上比亚马逊更依赖于数据。这家公司是在第一次互联网泡沫期间（1997年）诞生的，在今天算得上是资格最老、上市最早的公司之一了，但是其业务真正发展起来却是几年前大数据时代到来以后的事情。Netflix原本指望通过互联网的优势与原有的电影租赁公司百视通（Blockbuster）和好莱坞录像（Hollywood Video）竞争：用户可以在互联网上选定自己想看的电影，Netflix将电影的DVD（数字多功能光盘）用快递送给用户，用户看完后再将DVD 放到一个已付邮资的信封中寄回给 Netflix。不过用户手上只能同时保留1-4张DVD，只有Netflix收到寄回的DVD，才会给用户寄出他想看的下一张。Netflix的收费从每月8美元到18美元不等，取决于用户手上能同时保留几张 DVD。考虑到邮寄的周期通常是一周，因此算下来大约相当于花2~3美元在家看一场电影。

Netflix在它早期的10年间发展并不快，这不仅因为用户增长不快，而且活跃度也不高。Netflix早期的用户（包括我本人和周围的人）都有一个共同的特点，就是在头几个月把过去想看的电影都看了，接下来就不知道该看什么了。虽然 Netflix也会推荐一些好片子给用户，但是由于它并不了解每个人的需求，因此推荐的常常是最热门的或者评分最高的电影，但是个人的口味相差很大，这种缺乏个性化的推荐效果并不好，因此原本订18美元一次保留4部电影的用户，就改成每月花8美元，而原本看得不多每月花8美元的用户就干脆退订了。Netflix后来将邮寄改为通过宽带在线观看，这有点像我在拙著《浪潮之巅》里描述

的“根据需求收看”（on demand）。虽然从理论上讲观众省了来回邮寄的时间，应该能看更多的电影，事实却是大部分观众并非如此，因为一开始在线观看并没有解决如何有针对性地推荐电影的问题，大部分用户的活跃度并不高，因此在很

长时间里，大家都不看好这家公司。

但是，随着数据量的积累，尤其是和每一个用户相关的各种维度数据的积累，Netflix给每一个用户的推荐越来越靠谱，越来越准确。Netflix不仅知道每个用户看什么风格的电影（风格、题材、导演、演员等）最多，而且知道它给用户推荐的效果是否好（是否点击观看，是否看到一半就转去看别的节目了，等等），这些数据是过去其他传媒公司无法获得的。今天，它的用户所观看的节目有3/4是Netflix推荐的。靠着精准的推荐，Netflix用户的活跃度在不断提升，而一些原先有线电视和卫星电视的付费用户，也开始终止原来的服务（或者退掉部分套餐），改用Netflix。从2008年开始，Netflix的业务量剧增，到了2014年，

Netflix的流量已经占到美国峰值流量的1/3以上，[11]并且为全世界除中国以外的主要国家提供在线电影服务。2016年年初，Netflix的市值已经超过传统的电视网、默多克的Direct TV。

和亚马逊类似，Netflix的数据具有较强的时效性，它可以根据用户的反应很快调整它的市场策略，这种灵活性也是过去那些事先安排好一周节目的有线电视网所不具备的。

时效性很强的个性化的推荐不仅体现在商品上，还可以用于任何意义上的信息搜寻。在Google内部，直到2005年，反对为用户提供相关搜索的声音依然占上风，因为很多人认为应该由用户自己输入他们所要查找的关键词，而不是由搜索引擎引导用户去搜索。事实上，早在2005年，我们就开发出了利用搜索关键词之间的相关性提供相关搜索的技术，我们甚至能够在搜索条中自动地根据用户搜索习惯和输入的一两个字提示出完整的关键词组合。但是这个服务迟迟未上线，因为佩奇和布林并不喜欢这种服务。最终，我们不得不先利用中、日、韩文字打字慢的特殊性说服了两位创始人允许我在这三种语言中试一试，结果这种相关搜索一下子让这三种语言的搜索量增加了10%。不到一年后，佩奇同意把这项技术应用到英语和其他语言中，与中、日、韩语言类似，它对提升英语等其他语言的流量有同样明显的帮

助。到了 2008年，佩奇在这方面变得激进起来，不仅同意我们在搜索结果页的下方提供 相关搜索，而且希望能够在搜索栏内根 据用户当前部分输入和历史数据，自动 提示搜索的关键词，这使得搜索关键词 输入的速度大大提高，Google搜索在用户中的黏性进一步提升。再到后来，由于 数据量的增加，特别是能做到针对每一 个用户都积累了足够量的历史数据，以 至于关键词的提升能够做到完全个性 化，也就是说，两个不同用户，在输入一 半关键词后，Google给他们的提示常常 是不同的。到了2011年，Google不仅积 累了大量的用户数据，而且了解了用户 使用互联网的行为，甚至是生活的习惯 (比如住在哪里，每天工作做些什么事情 等)，因此进一步提出“无关键词的搜索”， 也就是说，对特定用户，根据他某个时间 过去的行为，以及当前使用Google产品 的场景，自动产生搜索关键词（在用户看 来自己没有输入任何关键词），从互联网 上查找信息，然后提供给用户。Google 基于这项技术最重要的产品就是安卓手 机上的Google Now——它可以提示用 户接下来该做什么，而这种提示靠的是 当时的时间、地点、应用场景和不同用户 本身的习惯特点。

Google李世石: | . Q

李世石对弈 Ranov*

李设 S alphago Renore

李逛石

雜: S柯洁 |

图4.6 Google的搜索关键词提示功 能，输入部分搜索关键词，Google可 以 根据该用户的搜索历史和其他用户的常 见搜索，提示全部关键词，并且自动填充 到搜索框中

从Google的这个案例中我们可以 看到，技术的进步可以改变人们的思维 方式，从而让产品呈现出新的形态。

大数据商业的共同点—— 尽在数据流中

在上述大数据应用的案例中，存在 着一些普遍的规律，这些规律可以

通过 数据流(Data Flow)的一致性体现出来。 我们不妨分析一下在上面的几个案例中 数据是如何流动的。

首先，大量看似杂乱无章的数据点， 从很多不同的地方(可以是不同的人、不 同的公司， 甚至是不同的采样点)收集上 来， 这些数据在生成时常常是彼此独立 的， 而且在收集上来之前是原始的、未加 工的、无目的的。 无论是亚马逊上顾客的 购买行为， Netflix上用户收看电影的行为， 还是Google用户上网搜索或者做其 他事情的行为， 事先与这些服务的提供 商都是没有沟通和商量的， 而且彼此是 独立的。 这些大量独立的数据聚合在一 起， 才能得到客观而准确的统计结论， 比 如网页搜索和结果之间的相关性， 不同 商品之间的相关性， 或者不同电影之间 的联系等。 在这个过程中,各种数据如同 百川入海一般汇聚到一起。

筛选、处理后 数学模型

的数据

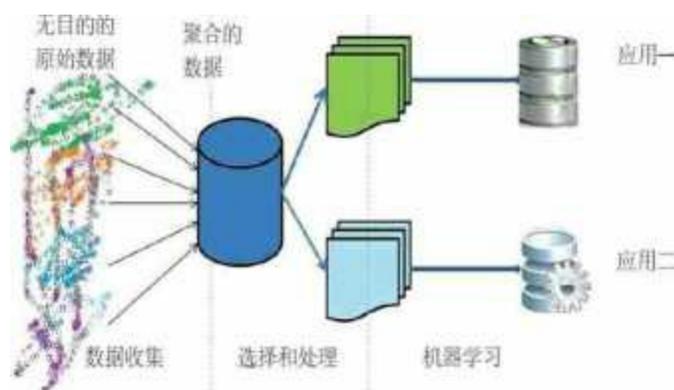


图4.7大数据收集、处理和建模的流程

其次， 由于这些数据在产生和收集 时是没有特定目的的， 因此怎样使用它 们需要视特定的应用而定， 比如Google 在网页搜索排序中使用的 数据， 与它在 给用户搜索提示时使用的数据不同， 虽 然它们都是从同 一个来源收集到的。 由 于大数据的多维度特征， 使用者可以根 据自己的 需求进行筛选、过滤和处理。 同 时， 由于在收集数据时事先没有太 多的 目的性， 从这些数据中能够得到什么结 果事先也无从知晓， 最终 从数据中得出 什么结论就是什么结论。

在上述过程中， 数据的流向是从枝 末的局部到整体。 而当我们利用从 大数 据得到的规律指导商业行为和其他行为 时， 数据的流向则是从整

体到局部，如图 4.8所示。

从每一个细节到整体 ▶▶

从整体到每一个细节

图4.8在大数据的商业应用中，数据通常要完成两个方向的流动

前面的几个例子无一不是先从大数据找到普遍规律，然后再应用于每一个具体的用户，并且影响到每一个具体的操作。以抓毒品种植和偷漏税为例，警察局或者税务局首先需要根据大数据了解用电或者纳税普遍的模式，然后要准确地估算出每一个地址正常的模式，这样就能够发现每一个异常的情况。对于互联网公司的那些应用也如此，那些公司可以对每一个用户提供不同的服务，甚至做到每一次的服务都不相同。比如电商公司在用户浏览打印机或者电动牙刷时，如果发现他们在阅读产品介绍和评价，那么可能用户尚未完成购买，推荐相应的产品给用户是合理的；而当用户完成购买后，再搜索或浏览这些产品，推荐给用户打印机墨盒或电动牙刷头等耗材，就比推荐那些耐用产品本身更合理了。经常在亚马逊上购物的人对这一点会有体会，不仅不同的人看到的网页内容是不一样的，而且同一个人今天和昨天看到的内容也是不一样的，尤其是在完成一些购买行为之后。这种精细到每一次交易，甚至每一次内容展示的服务，在过去是想都不敢想的，但是靠大数据今天这已经变成了可能，而且它还代表着未来商业的趋势。

把控每一个细节

大数据在商业活动中从细节到整体 再从整体到细节双向的流动，使得我们 不仅能够利用大数据对商业进行整体提 升，更能够精确到每一个细节。这在互联 网公司已经不是什么稀奇事，不过即使 在所谓的传统行业里，大数据也能帮助 我们做到这一点。我们不妨看看下面这 几个例子。

戴维是硅谷地区一位创业者，他喜 欢根据技术发展的大趋势寻找特定领域 里的商机。我在见到他之前他已经创办 过两家公司，一家公司表现平平，于是他 在经营到第四个年头时不得不将它关 闭。但是戴维的第二家公司经营得不错， 并且在5年后被一家大公司收购了，这样 戴维获得了财务上的自由。戴维在接下 来的一年里走访了美国100多家酒 吧，然 后考虑如何利用大数据和移动互联网来 帮助它们提升业务。在美国，一半小型企 业（包括餐馆等）的寿命不超过5年，酒 吧也是如此。戴维发现它们之所以经营 不下去，除了 一般所说的经营不善，更重 重要的是大约23%的酒都被酒保们偷喝了。

那么酒保们是如何偷喝掉将近1/4 的酒的呢？戴维说，这其实很简单， 主要 是酒保们趁老板不在的时候偷喝酒，或 者给熟人朋友免费的和超量的酒饮。比 如小王是酒保，小李是他的朋友，这天小 李来到酒吧 时，小王看老板不在，就给小 李倒上一杯没有算钱。甚至即使老板 在， 小王本来该给小李倒4两酒，结果倒了 6 两。由于每一次交易的损 失都非常小，不 易察觉，因此在过去酒吧的老板平时必 须盯得紧一点， 如果有事离开一会儿，只

好认倒霉。图4.9传统酒吧的酒很难管理

开过小餐馆的人都会有这样的经 验，自己是否在店里看着，对营业额 的影响特别大，因此做这种餐饮买卖的人特 别辛苦，稍微不注意就开始亏损。针对酒 吧老板的这些麻烦，戴维设计了一套解 决方案——改造酒吧的酒架，装上可以 测量重量的传感器，以及无源的射频识 别芯片（RFID）[12]的读写器，然后 再在每个酒瓶上贴上一个RFID的芯片。 这样，哪一瓶酒在什么时候被动过，倾 倒 了多少酒都会被记录下来，并且和每 一笔交易匹配上。酒吧的老板可以用平板 电脑查询每一笔交 易，因此即使出门办 事也可以了解酒吧经营的每一个细节。

当然，戴维提供的服务如果只是停留在这个层面，那么更像是一个“物联网”（Internet of Things,简称IoT)的应用，与我们所说的大数据其实关系并不大。戴维对酒吧的改造带来了一个额外的好处，就是积累了不同酒吧比较长时间的经营数据。在这些数据的基础上，他为酒吧的主人提供了一些简单的数据分析。我把他提供的服务概括为以下三个方面：

首先，分析每一家酒吧过去经营情况的统计数据，这有助于酒吧的主人全面了解经营情况。在过去，像酒吧这样传统的行业，业主除了知道每月收入多少钱，主要几项开销是多少，其实对经营是缺乏全面了解的。至于哪种酒卖得好，哪种卖得不好，什么时候卖得好，全凭经验和自己是否上心，没有什么分析。戴维提供的数据分析让这些酒吧老板首先对自己的酒吧有了准确的了解。

其次，为每一家酒吧的异常情况提供预警。比如戴维可以提示酒吧老板某天该酒吧的经营情况和平时相比很反常，这样就可以引起酒吧老板的注意，找到原因。在过去，发生这种异常情况时老板很难注意到，比如某个周五晚上的收入比前后几个周五晚上少了20%，老板们一般会认为是正常浮动，也无法去一一检查库存是否和销售对得上。有了戴维提供的数据服务，这些问题都能及时被发现。

最后，综合各家酒吧数据的收集和分析，戴维会为酒吧老板们提供这个行业宏观的数据作为参考。比如从春天到夏天，旧金山市酒吧营业额整体在上升，如果某个特定酒吧的销售额没有增长，那么说明它可能有问题。再比如，戴维还可以提供不同酒的销售变化趋势，比如从春天到夏天，啤酒的销量上升比葡萄酒快，而烈酒的销售平缓等。这些有助于酒吧老板们改善经营。

2013年，戴维从硅谷几家风险投资基金获得了融资，专注于利用大数据改进传统的酒吧行业。在这个例子中，我们会发现大数据可以让商业行为在准确把控宏观规律的同时，精确到每一个细节，从而提高利润。在未来，即使是那些标榜时尚和艺术的传统企业，也需要利用大数据重塑它的商业竞争力，而很多知名企业已经开始这么做了。

普拉达(Prada)是意大利著名的奢侈品品牌，有着100多年的历史，它的产品主要包括服装、皮具和皮鞋等。通常购买奢侈品的过程和一般商品不同，购买者不仅需要购得一件奢侈品，而且希望享受购物的过程。这些体验常常只有在顾客密度不高的专卖店才能享受到，因此随

着业务的增长，普拉达在全球开了 250家专卖店。和很多奢侈品一样，普拉达的销售有一半来自它的专卖店，而不是高端百货店或者网站直销。奢侈品销售还有一个特点，就是它的销量要看是否赢得了消费群体的喜爱，而与价格关系不是很大，因此很难通过降价促销来提高业绩。至于能否赢得人数并不多的消费群体的喜爱，在过去主要是看设计师的经验和专卖店营销的水平。

不过，经验和营销水平在过去常常靠不住，或者说不可能靠得住。据《奢侈舰》(Deluxe : How Luxury

Lost Its Luster) 书的作者、专门研究奢侈品的获奖作家戴娜·托马斯 (Dana Thomas) 女士介绍，这些奢侈品时装的销售好坏常常看运气。虽然在外界看来大牌时装设计师有很高的艺术水平和经验，而且他们也是非常尽心尽力的设计好每一款产品，但是市场反应如何他们完全不知道。至于销售水平也是如此，虽然这些奢侈品品牌在设计 and 布置专卖店时非常尽心尽力，比如某一家大牌公司在北京新开一家专卖店之前，1:1的模型就做了3个，但是其实没有人事先确定专卖店的设计应该是什么样的，里面的时装应该如何摆放。更糟糕的是，公司和设计师在过去甚至无法根据销售的结果了解成功或者失败的原因。比如一款时装卖得不好，是设计的问题或制作的问题，还是在专卖店销售的问题——比如没有把它放到明显的位置，这些都无从得知，当然就谈不上总结经验教训了，因此一切都是靠运气。

但是，这些问题在大数据时代开始有了答案。早在2001年，普拉达就开始利用最新的IT技术来提升它的销售。首先，它在商品的标签里嵌入一个很小的RFID芯片(图4.10)。良「旧是_种不需要电源的芯片，里面存储的信息可以被专门的阅读器发出的无线电波探测出来。我们在下一章会介绍RFID的一些技术细节。根据《普拉达：欲望的科学》[13]—文的描述，销售人员挥动一下商品，RFID的阅读器就可以识别这件商品并且给出它的详细信息。更重要的是，这个芯片可以把客户正感兴趣的这一件商品和他们可能感兴趣的其他商品联系起来，这有点像亚马逊的商品推荐。据普拉达的销售副总裁丹·斯坦尼克 (Dan Stanek) 讲，通常顾客和店员的交互越多，购买的可能性越大，因此相关的推荐非常有用，没有这种智能芯片之前，其实店员不知道该推荐什么给顾客。当然，普拉达所做的远不止嵌入一个小芯片做商品推荐，它还改造了专卖店的试衣间，这样每一次顾客把时装拿到试衣间试穿，店里都能记录下来。普拉达的数据分析师根据这些数据就能知

道如果_件时装 卖得不好，是因为放在店里没有人注意到（根本没有拿去试穿），还是因为试穿 后顾客不喜欢。根据这些信息，公司就知道问题出在设计 and 制作上，还是出在销售上。



图4.10普拉达的衣服价牌里藏着一个 RFID芯片

普拉达的智能试衣间能够做的事情 远不止收集试衣的次数和时间这些简单的信息,它有一个屏幕，能够让顾客从各个方位“看”到自己试穿上一件衣服或者 戴上围巾、皮具的效果。它还可以让顾客 看到自己试穿不同尺码、不同颜色类似 服装的效果，这样顾客不仅不需要拿一大堆衣服到试衣间，而且有欲望“试”不同的搭配。在过去，如果这家专卖店没有一些颜色和尺寸的搭配，顾客常常转身就走了。现在，顾客可以通过试衣间的屏幕，大致了解自己试穿那些自己并没有试的服装的效果，如果他们喜欢，普拉达 的专卖店可以从其他商店为顾客调来他们所喜欢的服装。

利用大数据，普拉达的销售额从 2001年的15亿美元左右，提高到2013年的40多亿美元，这个增长速度要远远 高于全球的经济增长速度，也高于服装 行业总体水平。

在一般人看来，针对终端用户的商业，也就是俗称To C (2C)的生意，很容易积累数据，同时能够利用大数据精细到每一笔交易的特点以提升商业水平。而对于那些针对企业级客户的生意，即所谓的To B (2B)的生意，在过去一笔是一笔，而且通常要靠比较长时间的 市场铺垫

(Marketing)才能拿到合同，因此可能大数据带来的帮助并不大，但是真实的情况并非如此，大数据对商业的帮助是全方位的，即使是对企业级的用户。

中国的金风公司是一家生产风能发电设备的公司，2015年时它的风能发电机在全世界的占有率已经排到第二位，这是一个相当好的业绩。但是，金风公司在海外面临着中国制造业企业通常都会遇到的困境，就是虽然市场占有率不低，营业额也不少，却没有多少利润，其根本原因在于中国的企业常常只能控制从设计到销售诸多环节中的制造环节，其他六七个环节的收益则被外国公司赚走了。当然，像金风这样具有不少自主知识产权和技术的公司，有能力赚取设计环节的利润，却无法掌控市场。这并不是因为金风的市场能力不够强，而是由企业级设备销售的特点决定的。

在过去，企业级的设备采购常常是购买者的主动行为，也就是说购买者有了需求后，向销售者购买。在世界贸易中，销售者和制造者常常不是同一家公司。比如巴西的某个连锁的零售百货店要更新1亿美元的计算机和通信设备，它通常会找一个工程的合同商(中间商)来承包整个工程，这样方便设备的运行和维护。这些中间商一方面搭建了制造商和顾客之间的桥梁，另一方面在主观或者客观上也阻断了买卖双方的联系。在过去一旦买卖双方货款两清，它们的关系就基本中断了。接下来买方的设备使用得怎么样，是否有新的需求，卖方是一无所知的，直到买方有了再次购买的意愿而通知设备生产厂商来竞标。当然，比较主动的卖家会做一些市场分析，当然这些市场分析很难做到准确。即使像波音和空中客车这种两家就完全垄断了全球市场的公司，对市场的预测也常常是错的。只要读一读波音公司每年向美国证监会提供的年度财报就会发现，它对未来一到两年市场预测的准确率只有60%左右。

具体讲金风公司，以前它虽然卖了不少风力发电机，但是那些发电机用在哪里?使用得怎么样，哪些地区还有潜力，哪些地区已经饱和，它所知甚少，对国外的用户更是一无所知，在过去这些售后的服务也不是它们工作的重点。到了大数据时代，该公司的管理层逐步意识到数据的重要性，开始转换经营理念，而且还专门到硅谷来取经。在此之后，该公司利用互联网，将发电机的各种数据(地点、发电量、运行情况)全部收集到公司，进行大数据分析。这样他们一方面可以全面地了解全球的风能分布情况、各地的风力利用情况等宏观信息，有利于公司

有针对性地做市场推广;另一方面，他们 可以了解每一台发电机日常运行的每一个细节，不仅发电机有了问题可以及时发现并解决，而且如何进一步改进也有了数据依据，这样一来该公司的经营策略就从依赖市场预测、打价格战等传统 的营销手段，提升到成为高质量的服务商，业绩也得到明显的提升。再到后来， 它的商业模式也开始发生变化，这一点 我们后面还会讲到。

像金风这样的中国企业非常多，我 在给一些传统行业的企业家讲课时了解到，在中央空调、工业制冷等很多行业，中国的企业在完成制造和安装后，就和 海外顾客鲜有联系了，更不用说通过对 那些顾客的服务了解全球的市场状况。 金风公司所做的尝试，或许对很多行业 都有借鉴意义。

重新认识穷举法--完备

性带来的结果

在商业上，大数据不仅便于掌握大局和每一个具体细节，而且改变了人们开发产品和解决问题的思路，这些做事方法的变化在很大程度上是大数据的完备性带来的。

在我们的认识里，穷举法在工作中并不是一个好的方法。首先，在大多数情况下无法穷举所有的情况；其次，即使在一些场合能够穷举出各种情况，这种方法也被称为笨办法，用穷举法会被人瞧不起。以笛卡儿和牛顿为代表的方法论都是在强调寻找一种普遍规律，然后用数据来验证。一旦这种普遍规律被找到，它就一劳永逸地解决问题。当然，当过去认为是普遍适用的规律遇到意外时，人们会找到相应补救的规律。但是，不论我们找到多少新的规律来处理那些不常见的意外情况，可能还会有意外发生，这种工作方式到后来效率就变得非常低了。当我们所找到的规律只能覆盖不常见的个案时，这种方法其实就和穷举法差不多了。既然如此，我们可能需要重新认识穷举法这种笨办法，或许在大数据时代它并不像想象中的那么笨。

下面_个例子是我在Google遇到的实际案例，从这个例子中，大家可以看出来我们在研究和开发工作中方法的变化，当然这个变化是基于我们有非常多的数据和非常强大的计算能力。

网页搜索最早是用关键词索引查找的，这很容易被想到。但是在欧洲语言中，用词受限于时态、语态、性别（阳性和阴性）[14]，同一个意思在不同上下文可能用了不同的拼写，因此严格按照关键词匹配，例如查找时使用单数名词，就可能找不到有复数名词的内容。当然，这也难不倒工程师们，大家很容易想到把意思相同的词归为一类，按照类别来查找，而归类最简单的方法就是采用词干（stem，有时也叫作词根），比如计算一词的动词形式在英语中是compute，变化形式是computed、computing或者computes等，名词形式是computation，形容词是computed，计算机是computer、computers.....这些都可以对应_个词干comput，如果用comput查找，似乎比用每一个单独的衍生词（compute、computer、computation等）更合理。这个想法几乎在一有文献搜索时就有人想到了，可以追溯到40多年前，但是奇怪的是直到2003年，在真

正的产品中都没有使用这种方法。自从有了互联网和网页搜索，不断有人尝试用这种方法改进搜索质量，但是发现它带来的问题和好处同样多，比如在搜索计算机产品时，单数名词 computer 和复数名词 computers 是等价的，但是如果我们说计算机科学 computer science 时，就不能用复数 computers 取代 computer 了，在前一种情况下，使用词根是合情合理的，而后一种情况就会找到一些不相干的结果。因此，无论是学术界还是工业界，在进行了多年尝试后，都先后放弃了上述想法。

是否有办法确定在什么情况下应该使用词干搜索，在什么情况下必须严格按照关键词的原型搜索呢？对于具有较高语言水平的人，实际上是能够做到这一点的，但是要让计算机做到这一点就很困难，因为在什么情况下可以让一些近义词相互替换，什么情况下不可以，这并非几条规则就能够写清楚的，也不是简单地使用一个概率模型就能估摸出来的，在很多时候它们都需要按照个例来处理（case by base），或者说随时按照具体情况做具体的分析。至于这些具体情况有多少种，基本上讲，亿万用户能够想到的每一种搜索关键词组合，都是一种情况。在大数据时代之前，没有人奢望有一种方法能够把这么多情况一一考虑到，但是在大数据背景下，列举每一种（常见）情况，并且有针对性地做出不同

的处理，则成为可能。

2003年，在Google内部，辛格博士和我等4个人，再一次尝试使用词干进行搜索，尽管我们知道前面有很多人尝试失败。与之前其他人不同的是，我们找到了一种方法，能够对每一种关键词的组合做专门的处理，比如我们知道在什么情况下动词compute和computes、computed、computing,甚至和名词 computer 或者近义词 calculate、estimate是同义词，可以混为一谈，什么时候必须严格分开，也就是说，对于每一次搜索我们都能找到最好的匹配方式。Google在2003年一整年中，搜索质量的改进一半是靠这个方法。至于我们是怎么做到的，说起来可能会显得很没有技术含量。我们事先把多年来用户搜索过的关键词搭配都整理出来，然后在2003年美国独立日的长周末期间（有4天的假期），我们停掉了公司当时5个最大的数据中心中的一个，利用4天时间，对每一个关键词的搭配做了特殊处理。这实际上就是一种穷举法，Google的优势在于它有足够的数据和计算能力用“笨办法”把每一种搜索事先试一遍，而这一点大部分公司做不到。当然有人会问，如果将来遇到过

去没有见过的新的 关键词怎么办，办法也很简单，第一次遇到它时，用户只能认倒霉，搜索引擎只能按照旧的搜索方法给出结果；但是同时 计算机离线地把这个关键词处理一遍，这样以后别人再搜索这个关键词时，就可以使用针对它的特定搜索方法进行 搜索了。

在这个例子里，我们看到大数据思维改变了我们的做事方式，因为过去被 看作笨办法的穷举法变成了可行的方法。更为颠覆我们思维方式的是，穷举法 可以方便我们对特殊情况做特殊处理，这反而是过去那些放之四海而皆准的机械思维做不到的。

通过这件事我们也能进一步体会大数据完备性的特点。在过去，统计学家们 一直试图寻找好的采样方法，以便在有限的样本中找到覆盖尽可能全的规律，但是在大数据时代，这些努力都不需要了，因此样本集可以等于全集。另外，我们还可以从这个案例中看到大数据时效性的特点。对于新的、过去没有见过的情况，Google的服务器反应是非常及时的，即在第二次就能把新鲜的数据提供给用户使用，这在大数据时代之前也是做不到的。

如果Google搜索的例子对很多非 IT行业的读者来说还不够直接，Google 自动驾驶汽车则是一个利用大数据思维解决问题的极佳案例。

Google的自动驾驶汽车可以算是一个非常聪明的机器人，因为它可以像人一样控制汽车，识别道路，并且对各种 随机突发性事件快速地做出判断。如果 单从驾驶的安全性来看，它的表现甚至超过了人。从有做无人驾驶汽车的想法 开始，到研制出让人眼前一亮的原型车， Google只花了 4年多的时间，这让全世界大吃一惊，其震惊程度不亚于当年深蓝战胜卡斯帕罗夫。其原因是，在所有专家们看来，自动驾驶汽车这件事太难了，而Google在这个领域进步的程度超出了最乐观的专家们最大胆的想象。

在Google之前，全世界的学术界已经花了几十年来研制自动驾驶汽车。20 世纪90年代初在清华大学上班和上学的人或许还能记得，在学校的主楼前一条 几十米长的弧形马路上时常有人在试验 自动驾驶汽车。在我和我同学的印象中，那辆车的时速只有每小时一两公里，在 无人干涉的情况下自动行驶的距离从来没有超过100米，这显然和实用性相差太远，当然后来清华大学也放弃了这个尝试。

世界上其他大学和研究所在这个领域的进展也快不了多少。在2004

年，美国国防部高级研究计划局（Defense Advanced Research Projects Agency,简称DARPA)组织了世界上 第一届自动驾驶汽车拉力赛。由于当时 各个研究团队水平都不高，因此比赛不 敢在真正的道路上进行，而是选择了 150 英里[15]长的废弃道路。不过后来的结果表明根本不需要准备这么长的赛道，因为最终取得第一名的汽车花了几个小时才开出8英里，然后就抛锚了。至于其 他参赛的汽车，不是提前抛锚了，就是撞坏了。

恰巧也是在这一年，经济学家弗兰 克■李文（Frank Levy)和理查德■默南（Richard Murnane)出版了《劳工 新种类》（A/new/ Division of Labor)_书，在书中他们列出了一些 在近期内不会受到技术进步威胁的工作，其中货车司机的工作赫然在列。李文 和默南在写书时并不知道DARPA拉力 赛的结果，他们的判断是根据他们自己 当时对科技进步的了解而做出的。在作 者给出的很多理由中，很重要的一条是 这样说的：计算机善于执行事先制定好 的规则，解决确定性问题，而驾驶汽车会 遇到很多的不确定性，并非规则能够解 决的，需要实时做出聪明的判断。这两位 经济学家认为，处理不确定性的能 力是人所特有的，机器暂时不会具有这个能力。

但是，就在DARPA拉力赛过去仅仅 6年之后，2010年Google就研制出了自 动驾驶汽车，并且已经在各种道路上，从 闹市区到高速路，行驶了14万英里，没有 出一次事故[16]。为什么Google能在 如此短的时间里做到这一点呢？除了它 聘用了在这个领域世界上最好的专家， 即几年前获得自动驾驶汽车拉力赛第一 名的卡内基-梅隆大学的团队，以及采 用了当时最好的信息采集技术，从激光 雷达（Ladar）到高速摄像机，再到红外 传感器等，最根本的原因是Google采 用了和其他研究单位不同的研究方 法——它把自动驾驶汽车这个看似是机 器人的问题变成了一个大数据的问题。



图4.11 Google自动驾驶汽车，注意，里面没有方向盘

首先，Google自动驾驶汽车项目其实是它已经成熟的街景项目的延伸。对Google自动驾驶汽车的各种报道通常都会忽视一个事实，那就是它只能去Google“扫过街”的地方。对于这些已经去过的地方，Google都收集到了非常完备的信息，比如周围的各种目标的形状大小、颜色，每条街道的宽窄、限速，不同时间的交通情况、人流密度等，Google都事先处理好以备未来使用。因此，自动驾驶汽车每到一处，对周围的环境是非常了解的，它可以迅速把这些数据调出来作为参考。而过去那些研究所里研制的自动驾驶汽车使用的是人的思维方式，每到一处都要临时识别目标，这样即使所搭载的计算机再快，也来不及进行太深入的计算，因此无法做出准确判断。

其次，自动驾驶汽车上装有十多个传感器，每秒钟进行几十次的各种扫描，这一方面超过了人所能做到的“眼观六路、耳听八方”，同时大量的数据要在短时间内处理完，计算的压力是非常大的。Google的自动驾驶汽车是通过移动互联网与Google的超级数据中心相连的，虽然它本身携带的电脑不过是一台简单的服务器，但是整体的数据量和计算能力要远远超出过去其他公司和大学那些自动驾驶汽车上面所携带的计算机。

再次，我们人开车，常常是根据周围情况临时做出判断，遇到死胡同，转弯掉头再找其他的道路。Google拥有一个最好的全球地图数据，它的自动驾驶汽车不仅行驶的路线大部分是事先规划好的，而且对各地的路况以及不同交通状况下车辆行驶的模式有准确的了解，因

此它可以规避很多不必要的麻烦。当然,如果开到了事先(扫街汽车)没有去过的地方,自动驾驶汽车常常会无计可施。

在2016年年初,Google的无人驾驶汽车在道路上安全行驶了200多万英里之后,终于出了第一起负主动责任的交通事故。出事的原因与其说是它的判断出了问题,不如说是数据的缺失。出事的那辆汽车在道路上检测到一个5公斤大小的小沙袋,那种沙袋一般是家庭用在院落的水沟旁防止洪水的。一般司机遇到这种情况就直接压过去了,但是Google自动驾驶汽车没见过这个东西,因此试图换道绕过去,而那辆车并没有方向盘,乘客也无法人为控制方向,结果出了一次小事故。



图4.12让Google自动驾驶汽车出事故的就是这样一个小沙袋

我们讲这件事情,并非想要讨论自

自动驾驶汽车的产品设计是否应该允许人 能够控制它，也不是讨论它是否安全，事实上它比人开车安全得多，而是从反面 证明这是一个利用数据获得智能的典型 案例。在今天的很多智能产品和服务上，可以说没有数据就没有智能。

Google在数据上的优势，是大学和 各个研究所并不具备的。即使是全球著名的汽车公司，包括丰田、大众和美国通用，也不具备如此多的数据。因此，它们 虽然在自动驾驶汽车研制方面早起步几十年,但是很快就被Google超越。另外，机器学习“经验”的速度远远比人快得多，这也是大数据多维度的优势，因此 Google自动驾驶汽车的进步才能如此快。这并非说明Google的科研能力超过了过去那么多大学、研究所和公司的总和，反而是体现出大数据的威力，以及采用大数据思维的重要性。

从历史经验看大数据的作用

在历史上，一项技术带动整个社会变革的事情也曾经发生过。它们通常遵循一个模式，即：

新技术+原有产业=新产业

那些有意或者无意接受了这个规律的企业家，常常在新的时代又站到了浪潮之巅。

近代第一次带来全社会变化的技术是以蒸汽机为核心的动力革命。在瓦特发明万用蒸汽机之后，很多有上千年历史的古老行业，使用蒸汽机之后摇身一变成为新产业。



图4.13韦奇伍德瓷器博物馆中的蒸汽机

瓷器在蒸汽机诞生之前已经有近千年的历史了，[17]而且一直供不应求。但是自从瓦特和博尔顿在月光社的朋友韦奇伍德开始采用蒸汽机生产瓷器后，这种一度被誉为白色的黄金的商品就在全球范围内变得供大于求了，而且瓷器的用途也从盛器和装饰品扩展到各行各业。英国巴拉斯顿（Barlaston）的韦奇伍德博物馆依然保留着它早期使用蒸汽机以及使用蒸汽机制造瓷器的各种设计文档。韦奇伍德公司在它的历史回顾中写道，它不断将新技术应用于制造。

纺织业的历史比瓷器还要长得多，几千年来这个行业一直是_家_户式的小手工业。英国的纺织业在蒸汽机出现之前已经有了很大的发展，靠水能驱动的各种纺织机在19世纪之前是高科技产品，它们的生产效率比东方纯释手工的纺织机要高很多。但是，在那个年代，英国的纺织品并没有多到要向全世界倾销。等到蒸汽机用于纺织业，情况就不同了，英国需要打开东方市场才能消化全部的产能。当最终那些洋布卖到中国和印度之后，当地几千年来传统的家庭纺织业在短短的100年里就消失了。从此全世界的纺织业被重新定义，各个迈向工业化的国家开始建纱厂、织布厂，一时间纺织业成了工业化进程中的全新产业。

运输业的历史几乎和人类的文明史一样长，可以追溯到美索不达米亚的苏美尔文明时期。相比陆路运输，水路运输的能力要大得多，因此航运占了运输总量的大部分，为此中国还修建了大运河。从苏美尔文明开始，帆船就是运输的主要工具，到了18世纪，西班牙、荷兰等积极参与航海的国家，把大帆船技术推向了一个顶峰。在那时，大帆船是最可靠、最便捷的长途航运工具，当然也是高科技产品。但是当蒸汽机被应用在轮船上之后，大帆船就退出了历史舞台。类似地，在陆路运输方面，火车取代了马车，成为客运和货运的主要工具。一个崭新的运输业就此诞生了。

至于蒸汽机在工程方面的作用就更大了，世界上大规模建设城市和港口就始于那个时期。港口的建设后来帮助英国把工业品卖到全世界。

就在英国人开始采用蒸汽机改造这些产业时，它的GDP还远比不上传统的经济大国中国。但是在广泛使用蒸汽机的同时，英国实际上按照以下思路重新定义了很多产业：

现有产业+蒸汽机二新产业

这一思路使得英国把各个古老的文明都甩在了后面。中国虽然在洋务运动

之后开始使用蒸汽机，并且开始学习使用新技术，但是在思维方式上一直没有推广当时最先进的机械思维，还是坚持“中学为体、西学为用”的落后思想。

需要指出的是，英国当时并不是每一个工厂都在制造蒸汽机。制造蒸

汽机的是非常少的几个工厂，大部分是使用蒸汽机改造原有的产业。

到了19世纪末，电的应用改变了世界。其发挥作用的方式和蒸汽机有相似之处，也有不同的地方。相似之处在于，它也是靠单点突破，带动社会的全面变革。但不同之处在于，电的使用所带来的不仅是一种取代蒸汽能量的动力源，还是一种新的生产和生活方式，因此它催生了很多看似新的产业。从宏观的角度看，电的使用导致了人口高密度的大都市的出现，因为电梯的出现，人们可以把楼盖得高，公共交通（有轨和无轨电车、地铁等）的出现可以把城市拓宽。西方各国的大都市都是在19世纪末20世纪初形成的。

电对世界的巨大影响还在于各种电器的发明，它们导致了新产业的出现。比如，以电报和电话为核心的通信产业就是在那个时期奠定的基础，今天它是全球最大的产业之一。留声机、电影和后来的收音机的发明，导致了大众娱乐产业的出现。至于电灯、电动机、电炉等依靠电能工作的电器，作用就不消说了。因此电改变的不仅是经济，还改变了国家的政治形态、生活方式和社会结构。电本身还有一些特殊的性质，比如正负极性，对这些性质的利用可以让物质发生化学变化，比如将化合物变成另一种化合物或者单质。这样电的使用就伴随着很多新产业的出现和革命，比如电彻底改变了冶金工业的状况[18]。

此外，电也是化学工业的催化剂。在19世纪，化学有了突飞猛进的发展，但是几乎所有的成就都是在实验室里，人类还无法大规模地生产化工产品。电的使用，让化学从实验室走向产业化。从化肥到农药，从人造纤维到各种生活用品，从建筑和装修材料到油漆涂料，没有电，今天我们使用的大部分化工产品就制造不出来。电的使用创造出今天产值高达3万亿美元的化工产业。

不过，如果我们深究一下上述新产业的历史渊源就会发现，其实很多所谓的新行业在电出现之前就已经有了，比如建筑业、交通运输业、娱乐业、冶金业，在使用电之后，这些行业发生了质的变化。但是，如同工业革命并不需要所有使用蒸汽机的工厂都制造蒸汽机一样，在整个19世纪，美国主要供电的公司只有两家，即通用电气和西屋电气，而使用电、得益于电的公司却有千千万万。类似地，在当时第二大工业国德国，发电的也只有西门子和德国电气总公司两家。电带来了“第二次工业革命”，因此我们不妨把这个时代总结成：

现有产业+电=新产业

“二战”后信息技术带来了新的产业革命。信息革命其实有两方面的革命，首

先是创造了一批与信息的生产、传输和处理有关的产业，比如电视和传媒、通信、卫星，以及与信号处理相关的产业，比如军事上的雷达、地质上的遥感等，这些都是很大的产业。另一方面，原有的很多产业在使用计算机之后产生了本质的变化，形成了全新的产业。在过去的半个世纪里，很难找到哪些产业没有受到计算机的影响。我们不妨看两个看似与信息技术的关系不是那么密切的行业——金融业和农业，来体会信息革命对全球经济和社会的影响。

银行业是一个非常古老的行业，但是在过去的几百年里，它并没有本质的变化，存取钱和借贷都必须去银行，因此银行的大小取决于其营业网点的多少，从欧洲文艺复兴时期银行业的先驱美第奇家族，到后来犹太银行家的代表罗斯柴尔德家族，再到后来美国银行业的代表、洛克菲勒支持的花旗银行都是如此。它们需要花几代人的时间走到（它们所能触及的）世界各地，但是即便如此，在它们已知的世界中，仍有99%的人无法使用它们的金融服务。跨行的交易成本非常高，而且非常麻烦，因此人们旅行时不得不携带现金或者旅行支票。

与银行业相关的其他金融领域也是如此。比如在1971年纳斯达克诞生之前，股票的交易需要去交易所，或者打电话给中间商(broker)才能进行，更重要的是，他们常常交易的是真正纸质的股票。在交易中，报价过程是类似几百年前拍卖式的讨价还价过程。直到2000年，美国纽约证券交易所(简称纽交所)的交易价格还遗留着拍卖报价的痕迹，即买卖双方讨价还价时以一美元、半美元、四分之一美元，直到十六分之一美元为基数进行。由于这种出价方式买卖价差巨大，因此在那个年代，高盛和摩根士丹利等券商的主要收入来自交易费，每一笔交易的手续费都在100美元以上。

但是，计算机的使用彻底改变了这个行业。计算机网络的发展和自动取款机(ATM)的使用使得银行营业网点很容易部署到全世界。从20世纪70年代开始，工业化国家陆续实现了不同地区之间的跨行存取，甚至跨国存取。储户只要在一个稍微有点规模的银行开户，就可以在世界(除非洲之外)大部分地区使用存款。因此银行很容易把业务拓展到全世界。中国的招商银行成立于1987年，仅仅过了10年，它就成为全国性的银行，又过了10年，它在全世界除非洲之外的各大洲开办了分行或者办事处，相比花旗等老一代银行，这样的发展速度是惊人的，这一

切要托信息革命的福。今天，人们已经无法想象全世界的银行如果彼此不联网是多么不方便，可以说有了计算机的银行业和过去几百年的银行业已经完全不同了。

类似地，证券交易也发生了根本性的变化。1971年美国的全国证券交易商协会推出了自动报价系统，这套系统的英文全称为National Association of Securities Dealers Automated Quotations,简称NASDAQ,即我们常说的纳斯达克。纳斯达克和纽交所不同，交易者不需要再到交易所，而是通过网络和电话进行交易，交易的报价方式也是我们今天熟知的精确到一美分的方式。由于在纳斯达克上的交易完全是电子化的，纸质的股票便被淘汰了。[19] 纳斯达克的报价方式显然比纽交所的方便，于是在经历两种报价方式共存之后，纽交所放弃了上百年的传统开始向纳斯达克靠拢。纳斯达克的诞生使得一般的股民很容易通过折扣代理商(富达、先锋等证券商)自己交易股票，单笔交易的手续费只要5~10美元。这进一步改变了美国券商市场的格局，一方面让嘉信理财(Charles Schwab)这样的折扣代理商崛起，另一方面让高盛和摩根士丹利等高端代理商从股票交易转向理财业务。

如果我们把证券行业和IT行业做类比就会发现一个有趣的现象：在纳斯达克

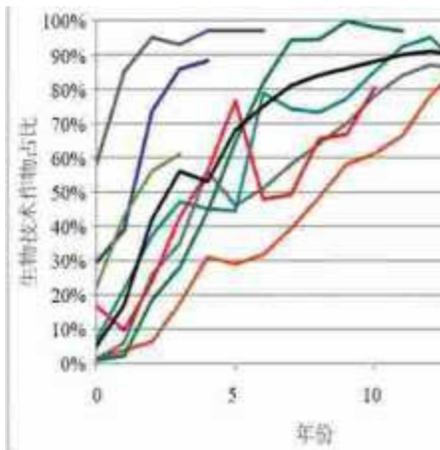
克出现之前，券商好比是生产大型设备的IT公司，它们每一笔交易都可以获得丰厚的利润，这就如同IT公司当时每卖出去一件产品，就都可以挣很多钱一样。但是，后来低端的券商靠价格优势抢了高端券商的生意，逼着高端券商从事理财这样的金融服务，就如同亚洲制造的电脑公司靠价格优势逼着和惠普从事IT服务一样。从这个趋势可以看出，各种服务在信息革命之后变得越来越重要。

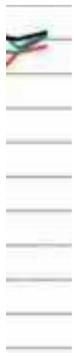
农业是另一个看似和计算机关系不大的产业，在几千年的历史长河里，这个产业变化非常缓慢。但是这种情况在过去的30年里得到了根本性的改变，比如农民们不再像过去那样自己育种，而是从种子公司购买种子。而种子的培育，背后用到了大量的信息技术作为支持。作为全球最大的种子供应商之一的孟山都公司也因此由一个化工企业变成了一家生物公司,年收入143亿美元，利润却高达82亿美元。相比之下，美国每年的农业收入才不过1200亿美元。[20]如果我们对比一下农业的这种模式和19世纪末美国使用电力的模式就会发现，孟山都等公司在某种程度上起到了通用电气和西屋电气的作用，而大量的农民所扮演的角色其实相当于19世纪末每一个使用电力的公司。这些农民和

农场主在有了孟山都之后，不再为种子发愁，就如同一个世纪前的工厂主在有了电力之后，不再为动力发愁一样。而在农产品市场上，采用传统农产品种子的农民很难和采用孟山都种子的农民竞争。就这样，农业这个最古老的产业在计算机时代被彻底改变了。

自从1965年摩尔博士提出了摩尔定律，计算机处理器和存储器的性能分别提高了2000万倍和10亿倍，价格却不断地下降，以至于它可以被应用到各行各业，以及生活的方方面面。40年前，没有计算机，中国的国民经济几乎不会受任何外在影响。今天就不同了，没有计算机是无法想象的事情，哪怕它们只有一天不工作，中国的城市也将全面瘫痪。人们会无法出行，因为无论是私家车还是公共交通，都要靠计算机才能工作。即使我们可以步行或者骑自行车出门，也进不了办公大楼，因为大楼的门禁也是用计算机控制的。如果砸开了门进入大楼，那么也不得不爬十几层到几十层楼才能进办公室，因为电梯也是由计算机控制

的。至于上班办公，更是离不开计算机。没有计算机，整个通信系统也会瘫痪，我们无法和外界取得联系。因此，可以毫不夸张地说，我们今天的生活已经完全依赖于计算机了。





—娜Jit —巴

—蜘蛛蓆 —加\$大汕菜 —美国棉花 —中闽棉花 —印度職 —■甜菜 —鯛大豆

图4.14通常，一种新的农作物种子被孟山都公司研发出来，在15年内便会占据全球80%以上的农产品市场

在过去的半个世纪里，世界的进步

背后最根本的动力可以概括为摩尔定律的应用，或者说是数字化。今天的大部分产业在“二战”后就已经存在了，但是无论什么行业，加上摩尔定律，就形成了一个新产业，我们可以把这个时代经济的特点概括为：

现有产业+摩尔定律=新产业

与前两次工业革命类似，虽然信息革命的代表产品是计算机处理器，但是并不需要每一家公司都生产处理器，甚至不需要每一家公司自己开发软件。今天大部分公司使用的处理器只有两个系列，即英特尔x86系列（加上AMD [21]兼容产品）和英国ARM公司设计的RISC（精简指令集）处理器，因此计算机实际上可以被看成是一种资源，而大部分公司需要做的只是使用好这些资源而已。

我们回顾过去是为了展望未来。今后，由大数据引发的智能革命也将是以一种与前面几次技术革命类似的方式展开，如果我们用两个简单的公式来概括的话，那就是：

现有产业+大数据=新产业

现有产业+机器智能=新产业

技术改变商业模式

历次技术革命除了缔造新产业之外，还不可避免地会带来商业模式的变化，进而导致社会生活方式的变化。

在工业革命之后，全世界从过去的物质生产供不应求，逐渐变成了供大于求。瓷器商人持续了几个世纪的好日子在蒸汽机用于瓷器制造后便一去不复返了。为了方便瓷器的销售，英国瓷器商人韦奇伍德在伦敦开办了瓷器展示店，这成为后来高端产品专卖店的前身。与此同时，由于纺织品价格下降，西装等商品采用机器生产，价格跟着大幅下降，人们由自己在家手工制作衣裳，变成直接购买制成品。其他行业和瓷器行业、服装行业情况类似。1851年，第一届世界博览会在英国伦敦郊区召开，这实际上是英国在向全世界展示它丰富的工业品。从那以后，世博会逐渐成为一种商品时代的传统，延续至今。

在第二次工业革命中，电的使用又一次改变了商业模式。现代传媒和通信业的兴起是电普及的直接产物。有了这些通信和传媒的手段，厂家和顾客之间有了顺畅的信息交流渠道，产品的营销则从过去口碑相传、实体店展示这种被动的形式，变成了采用广告主动宣传。全球性品牌在这个时期开始诞生，它们开始逐渐垄断全球市场。由于任何产品都可以很容易地买到，工厂不需要从零件开始做自己的产品，产业链开始形成，工业标准化成为必然。当然也就是在这个时期，大量本土的、地方性的品牌和产品消失了。同时，由于商品进一步供大于求，工业化国家必须依靠消费拉动经济增长，整个社会的消费价值观也开始发生变化。

在信息时代，商业模式的变化更加明显，它突出地表现在两方面，一是产业链从一种产品扩展到整个IT行业，二是服务业的重要性突显出来。

我们先来看看IT产业链的形成。在上一节我们盛赞摩尔定律给我们带来的好处，但是它也带来了一个问题，那就是让很多电子产品，尤其是与计算机相关的产品（比如个人电脑、DVD机、电视机和手机等）的价格持续下降。这对消费者看起来是福音，但是对生产厂家来讲是灾难性的，因为一旦出现这样明显的通货收缩，就不会再有消费者急于购买新产品了，这和消费拉动经济增长的格局是相违背的。为了解决这个根本性矛盾，就需要将整个IT行业整合成一条大的产业链，这条产业链可以被概括为“安迪-比尔定律”。

安迪-比尔定律的原话是：“比尔要 拿走安迪所给的。”（What Andy gives, Bill takes away.）这里面的安迪是个人 电脑巅峰时代英特尔当时的CEO安 迪·格罗夫，比尔则是大名鼎鼎的比 尔■盖茨，他当时是微软公司的CEO。这句话的含义是，在计算机领域,软件功 能的增加和改进要不断地吃掉硬件性能 的提升。这一点经历过个人电脑发展或 者智能手机历程的人都会有亲身体会。 虽然今天我们的个人电脑比1981年 旧M推出的PC (个人电脑) 快了两万倍 左右，但是我们并没有觉得它有那么 快， 因为微软操作系统使用的计算和存储资 源比30年前要多得多，给 人的感觉是它 吃掉了所有硬件性能的提升。



图4.15比尔·盖茨和安迪·格罗夫

安迪-比尔定律看起来是以盖茨为

代表的软件公司在和用户做对，但是，如 果没有这些软件公司提供新 的功能或者 不断改进现有的功能，整个计算机产业 就会缺乏发展的动 力。安迪-比尔定律 反映出计算机工业的整个生态链：以微 软为代表的 软件开发商吃掉硬件提升带 来的全部好处，迫使用户更新机器，让惠 普、戴尔和联想等公司受益，而这些PC 整机厂商再向英特尔这样的半 导体公司 订购新的芯片，同时向希捷（Seagate）等外设厂商购买新的外 设。在这个过程中，各家的利润先后得到相应的提升，股 票也随着增 长。各个硬件半导体和外设 公司再将利润投入研发，按照摩尔定律 预

定的速度，提升硬件性能，为微软下一步更新软件、吃掉硬件性能做准备。

从上述产业链中我们可以看出，主

动的一方不是各种看得见摸得着的工业品生产商，而是提供软件和服务的一方。正是出于这个原因，微软成为个人电脑时代最成功的公司，而曾经以生产计算机为主的旧M则坚决地进行了转型，将主营业务从制造计算机转向提供软件和技术服务。20世纪90年代，IBM传奇般的CEO郭士纳敏锐地觉察到摩尔定律导致IT行业的格局发生巨变，为旧M找到了一个至今依然有钱赚的商业模式——IT服务。人类对服务的需求总是有的，而且随着科技进步，人们对服务的要求越来越高，因此它的利润就有保障。事实证明20年前郭士纳主导的旧M转型是走对了路，而和旧1V1同时代的其他计算机公司，绝大多数要么关门了，要么被并购了。

摩尔定律和安迪-比尔定律到了智能手机时代照样适用，我们就不赘述了。

通过上述对历次技术革命中商业模式变迁的分析，我们可以得到这样三个结论：

首先，技术革命导致商业模式的变化，尤其是新的商业模式的诞生。

其次，生产越来越过剩，需求拉动经济增长的模式变得不可逆转。同时，单纯制造业的利润越来越低，那些行业越来越没有出路。相反，人们对服务的需求越来越强烈。在IT时代，唱主角的公司逐渐从制造设备的旧M、DEC、爱立信、诺基亚和惠普等公司，变成了提供软件和服务的微软、甲骨文和Google等公司。

最后，商业模式的变化既有继承性，又有创新性。工业革命导致了产品需要靠推销才能卖出去，第二次工业革命导致了广告业的兴起，推销的方式从展示变成了做广告，而这两者之间是有联系的。作为创新的一方面，第二次工业革命导致了产业链的出现；到了信息时代，产业链得到了发展，这是继承性的一面；而服务业的重要性突显，这是其创新性的一面。

在大数据时代，IT软件和服务业依然是IT领域最好的行业，而且这

个趋势将更加明显。提供服务虽然不像销售产品一次能挣比较多的钱，但是细水长流的技术服务最终会给这些服务的提供者带来更长久的生意、更多的利润。

样章到此结束

需要完整版

扫下面二维码



或加微信：shuyou055

免费领取

未来的记者和编辑

如果我们把计算机分析案卷和病例看成是一种阅读行为，那么今天的计算机已经发展到不仅能读，而且还能写作了。其实计算机在自动回答问题时，就已经具有了简单的写作本领，因为计算机回答问题的最后一个步骤就是将知识的片段写成优美的文字段落。当然，在回答问题时，所需要写的只是简单的段落而非完整的文章。

今天计算机写作的本领到底有多大？我们可以把写作从简单到复杂分为下面5个层次：

1. 书写完整的句子。
2. 组织几个句子构成符合逻辑的段落。
3. 给予特定格式，或者写作模板，能够清晰传递信息，表达意思。
4. 能够不限定格式地写作内容，达到一般人平均水平。
5. 能够达到专业记者、作家和学者水平。

在组织构造问题答案时，计算机已经达到了第二层次。实际上目前计算机的写作水平比这个层次还高一点，它能够完成结构比较清晰、格式固定的新闻稿，因此基本上达到了第三个层次的要求。

今天美国很多媒体的财经新闻，尤其是对公司财报的评述，其实已经是计算机产生的了。比如BM公司发布了去年四季度的财报，计算机先“读”一遍该公司财报的内容，然后提取出主要的信息，比如该季度的收入、利润，与华尔街预期的对比，人员情况，市场份额，等等；然后计算机可以写一篇关于IBM业绩的新闻稿，当然最后在发表前多少还是经过了人工的一些润色处理。至于有些新闻报道说计算机能够写诗，那只不过是媒体用机器智能做一些吸引眼球的事情而已，计算机还远不能达到自己抒发感情的地步，而且它写作的方式其实和人完全不同。

计算机是如何写作的？实际上它的写作方式和我们人在学习外语时造句的方法完全不同。它不是根据语法和所要表达的意思编句子，而是从大量文本语料中学习写作。我们常用“熟读唐诗三百首，不会作诗也会吟”说明背诵过去的范文对写作的帮助，而计算机的长处恰恰在于它能够背，而且能够快速读非常多的样本并背下来。计算机写财经评论其实是根据以前很多报纸上多年积累的财经类的文章，训练出各类财经文章的模板，然后每次根据从财报中读出的信息，合成一篇文章。当然，这样合成的文章读起来未免生硬，因此计算机还要用一种被称为语言模型[20]的概率模型，将文字构造成优美的句子，再用另一个语言模型将句子组合成段落。这些模型也是从以往的数据中训练出来的。当然，像《华尔街日报》或者《纽约时报》这样的大报在发稿前还会让编辑润色文字，而一些网络媒体常常将计算机写的财经文章直接就登了出去。计算机写作大大提高了新闻行业的效率，但是同时也让记者和编辑这类工作正在萎缩。或许再过若干年，我们在编辑部里看到的景象不再是一批伏案工作的编辑，而是一台台计算机，这个行业也就被重新定义了。小结

大数据将导致我们社会的产业升级和变迁。不过，如果对比每一次产业革命前后产业的变化，你就会发现其实人类很多基本的需求并没有变，只是采用了新技术后，新产业会取代旧产业满足人类的需求。在技术革命时，固守旧产业是没有出路的。

机器智能会给人类带来一个终极问题：既然什么事情都可以让机器来做，而且还比人做得好，那么人类怎么办？我们将在下一章中重点讨论这个问题。注释

[1] http://www.cnagri.com/mucaixw_aigeshidian_20130308/220677.html.

[2] <http://www.ishitech.co.il/0112ar8.htm>.

[3] 此前的记录是由乔丹时代的芝加哥公牛队保持，一个赛季获胜72场。

[4] 数据来源：世界银行。

[5] 绝大部分时间里，该医院被评为全美最好的医院。

[6] 美国总统看病的指定医院，类似于中国的301医院。

[7] PriceWaterhouse Coopers . The factors fueling rising healthcare cost s 2006

[Internet] New York (NY): Price Waterhouse Coopers; 2006. Jan.

[8] <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3048809/table/TI/>.

[9]按照美国的法律，急救的病人送到医院后必须救治，即使没有保险，这笔钱医院一般是拿不回来的。另外，很多患者临终前最后一笔医疗费医院是拿不到的。医院实际上将这些亏损变相地加到了有医疗保险的病人身上。

[10] 2014年美国中位数年薪是

5.2万美元。

[11]目前美国总统的年薪是40 万美元。

[12] <http://arxiv.org/ftp/arxiv/papers/1401/1401.0166.pdf>.

[13]在美国排名前三的医学院 (哈佛医学院、约翰■霍普金斯医学院和 斯坦福医学院) 录取率一般在2%左右， 而哈佛大学本科录取率为5% ~ 6%。

[14]IBM预测医疗数据每73天 翻一番，直到2020年。 <http://www.ibm.com/smarterplanet/us/en/ibmwatson/health/>.

[15] 基因泰克公司的主营业务是 利用基因技术研制抗癌药。

[16] 如果刻意用很大剂量的药物 试图杀死所有的癌细胞，可能导致人的免疫系统先被破坏，对患者反而有害无 益。在救治的过程中，因免疫系统被破坏 而死亡的病人非常多。

[17] 大部分人终身并不会得癌 症，因此将癌症患者寿命延长的时间平摊到所有人头上，远没有想象的那么多。

[18] 突破奖 (Breakthrough Prize)由布林夫妇、马云夫妇、扎克伯 格夫妇和俄罗斯著名投资人米尔纳夫妇 设立，每年授予在生命科学、数学和理论 物理学领域做出杰出贡献的科学家，由 于每个奖项的奖金数额

高达300万美元，远远超过目前诺贝尔奖的170多万美元，又被称为超级诺贝尔奖。和诺贝尔奖所不同的是，该奖获奖的项目并不需要验证其影响力，因此可以被授予最新的科技突破，而不是几十年前的重大贡献。

[19] <http://www.cnet.com/news/how-much-is-that-patent-lawsuit-going-to-costyou/>.

[20] 简单地讲统计语言模型是一个判定单词串是否像一个合理的句子的概率模型。要想了解语言模型更多的细节内容，请参见拙著《数学之美》。

第七章 智能革命和未来社会

在历次技术革命中，个人、一家企业，甚至一个国家，可以选择的道路只有两条：要么加入浪潮，成为前2 %的人，要么观望徘徊，被淘汰。

“这是最好的时代，也是最坏的时代”，这是英国文豪狄更斯在他的名著《双城记》中开篇的一句话，一百多年来不断地被人引用。在这里我们再次引用来形容智能革命给我们带来的未来社会。一方面，智能革命无疑将给我们带来一个更美好的社会，它是智能的、精细化的和人性化的，从这个角度看，智能社会无疑是迄今为止人类文明史上最好的社会。但是另一方面，智能革命也将给我们带来空前的挑战。随着大数据和机器智能的不断普及，我们会发现机器越来越多地占据了人类的工作机会，这个过程在一开始是悄无声息的，但是当发展到一个拐点，我们就会发现这个趋势将不可逆转。大数据和机器智能造福人类的同时，也会造成非常多的社会问题，以至于让我们不知所措。因此，或许有人会觉得这是最坏的时代。我们无意评论智能社会的好坏，只是希望大家对它所带来的冲击有所准备。

智能化社会

智能化社会表现在整个社会从宏观到微观的各个层面，在这一小节，我们先来关注宏观层面的变化。大数据和机器智能将把我们社会的管理水平提升到一个前所未有的高度，使我们生活的环境更加安全。

2014年跨年夜上海外滩陈毅广场踩踏事件是于2014年12月31日23时35分左右，在中国上海市黄浦区外滩陈毅广场台阶处发生的一起踩踏事故，截至2015年1月2日，事故共造成36人死亡，49人受伤。踩踏悲剧发生的根本原因是那个地区的人流量太高。据报道，事故发生前外滩地区人流量超过100万人，已超出该地区30万人的人流容量上限。如果能够 在事情发生之前，或者在事件开始时，准确地预测人流量，并且在第一时间通知给周围的行人，就能在很大程度上预防悲剧的发生。那么这件事是否能做到呢？

事实上，在上海踩踏事件发生之后，百度就开发了预测热门城市和景点的拥挤情况等相关信息的服务。而为什么百度能做到呢？其实说起来并不复杂，因为百度能够从安装了它的APP的大量用户手里得到人流的信息，这些数据汇总后，可以训练出一个根据人流和时间变化的模型，在未来的时间里，可以根据当前人流分布使用这个模型预测在未来的几个小时里人流的流动情况。如果发现过多的人流涌向某一个地点，那么就可以预警。

如果推广利用大数据预防踩踏事件的方法，就会发现它可以适用于很多类似的情况。交通拥堵是今天住在大都市里的人每天的烦心事，那么是否有可能通过城市整体上的智能交通或多或少地改进交通路况呢？从目前一些城市的实验结果来看这是能够做到的。Google自动驾驶汽车的研发团队曾经做过粗略的估算，如果道路上所有的汽车都是能够相互协调配合的自动驾驶汽车，即使不减少车的数量，只是对行车路线实现规划和协调的话，每个人平均通勤的时间至少可以缩短20%以上。对于这个结论，几乎没有人会有异议，因为对交通做整体的规划一定能够更好地利用道路，减少拥堵的发生，并且在拥堵发生后，让附近行驶的车辆能够及时地规避拥堵。虽然自动驾驶汽车的普及还显得有些遥远，但是利用智能手机在很大程度上可以取得类似的效果。

美国自然科学基金会(NSF)和国防部下属的DARPA资助了不少大学的研

究团队研究利用大数据从整个都市的层面优化交通的项目。其中一个由大学主导的项目团队已经开发了一整套基于智能手机和其他移动设备规划城市交通和优化每一个人出行的智能交通系统，并且在美国4个大都市开始试运行。由于该团队正在进行商业融资，不便于披露团队的细节情况，我们暂且称该团队为X团队，对应的项目为X项目。X项目的核心是利用实时的大数据更合理地在空间和时间上分配和利用交通资源（比如道路和停车场）。

通过手机APP有效地利用空间资源比较容易理解，未来的智能交通管理系统可以从每一个安装了这一类APP的出行的人那里，全面了解并且预测城市每一条道路的交通情况，比如哪些道路拥堵，哪些相对顺畅；同时也能够了解每一位出行者的情况，比如是自己开车、乘坐公交还是骑自行车或者步行，以预测各个道路未来可能出现的交通状况。这种智能交通管理系统的一个突出的优势是，它运行的时间越长，历史的数据收集得越多，对未来路况的预测就越准确。

在时间上优化一个城市的交通资源，就必须做到统筹每一个人每天的出行状况甚至是活动安排了。在信息时代，不少人上班的时间比较灵活，早上班半小时或者晚回家半小时其实不影响工作和生活。X项目对这一类人通常会建议一个每天最佳的上下班时间。X团队研究发现，很多时候早出发5分钟可以早到半小时，或者晚出发半小时，仅仅晚到5分钟而已。因此他们会根据每个人的工作安排，比如上午第一个会议的时间，给出这些通勤的人最佳的出门时间和路径。当然，城市里还有很多人每一天出行是很规律的（比如早上送孩子上学，然后去上班，下班后去买菜，然后回家，等等），强行要求他们改变生活习惯是行不通的，不过X项目的智能城市管理系统会给这些人提供详细的交通分析数据，帮助他们选择更好的出行时间和次序。

安装了智能交通软件的用户可能会有一个担忧，就是自己的行踪会完全暴露。为了保护个人隐私，X团队从来不保存使用者在起点和终点0.5英里范围内的活动路径。他们解释说，这样虽然损失一些信息，但是对掌控一个城市交通的整体情况已经足够了。更主要的是，如果监管部门要求他们提供使用者的相关信息，他们可以不提供，因为他们确实没有。



图7.1最左边的快速通道为拼车车道

智能交通不仅对通勤有好处，也方便市政当局优化和调整全市整体的交通

状况。首先,可以通过每天的交通情况制定拼车车道[1]的使用时间,引导大家尽可能地分散出行的时间和使用的道路。在硅谷地区,个别车道在交通高峰期是自动收费的,这个措施实行以后,不少通勤的人开始调整自己的出行时间和办事的次序。当然,目前挂谷地区这些车道的控制还没有利用大数据,如果使用,效果会更加明显。

其次,利用大数据管理交通可以根据实时流量和对未来流量的预测,调整交通信号灯的时间。目前世界上大部分城市的交通信号灯互相并不联通,而时间控制的策略总体上是固定的。我们经常看到在十字路口,另一个方向的道路已经没有了汽车而信号灯还是绿的,而自己的方向堵了一条长龙。

今天,世界上主要的大都市都已经没有了大规模扩建街道的可能性,但是其中大部分大都市的人口还在增加,流动人口也越来越多,因此除了更聪明地在时间和空间上利用好现有的道路,别无他法。

X团队目前和美国4个大型城市合作,试用了该系统,结果表明使用者每天可以节省20分钟左右的通勤时间。不要小看这20分钟时间,如果像北京这样的大都市每人每天能在通勤上节省20分钟时间,社会效益是非常可观的。

大数据对于交通状况的改进，其实只是它在帮助城市管理方面所做的一件具体的事情。相比交通拥堵，生活在大城市里的人或许更关心人身安全，而对人身安全最大的威胁就是恐怖袭击。从“9 ■ 11”事件开始，全球都面临反恐问题，尽管美国、欧洲和中国都加强了反恐的力度，但是总的来讲全球恐怖袭击事件越来越频繁，这说明按照过去的方式反恐越来越难。

大数据的出现给反恐带来了曙光。根据俄罗斯官方的报道，1996年4月21日深夜，俄罗斯在车臣叛军首领杜达耶夫用手机通话时，用A-50空中预警机根据无线电波锁定了他的位置，然后发射导弹将其炸死。这件事给大家一个提示大部分恐怖分子今天也是使用手机通信联络的。基于这个认识，斯坦福大学一位不愿透露姓名的学者开发了一个系统，可以全面跟踪一个地区所有手机和电子设备（包括各种移动设备和可穿戴式设备）使用者的行踪。据这位学者介绍，上述每种设备其实都有一个特殊的识别码，可以用一个阅读器来识别（其原理有点像RFID）。在公共场合安装这样的识别装置采集历史数据，一旦有外来的可疑分子（带有不认识的设备，或者已被怀疑的设备），就可以开始预警，并且配合视频监控跟踪那些人。这位学者在斯坦福周围的一些公共场所进行试验，能够准确地将各种外来之客从偌大的校园里识别出来。目前这位学者正在和某个国家合作，建立覆盖整个地区的反恐防范系统。在未来，有效的反恐需要更多地采用技术的手段，而不仅仅是增加人力。

智能社会体现在方方面面，但概括起来，就是让我们的生活变得更加方便，

同时社会资源的利用率极大地提高。要做到这一点，重要的是让整个社会精细化。

精细化社会

我们在第四章中介绍了大数据在商业应用中的两个方向。从每一个局部汇集到整体时，我们能够掌握全局，实现社会的智能化。而当数据再从整体流向每一个细节时，我们可以让未来的社会变成一个精细化的社会。为了说明这一点，我们不妨先看看通过区块链（Block Chain），在未来如何跟踪每一件商品从制造出来直到被消费的完整行踪。

追踪每一次交易

从2013年开始，比特币这种既没有政府信用背书，也没有实体价值支撑的虚拟货币忽然被中国的炒家从每个三十几美元，炒到了上千美元，一下子成为IT行业的一个热门话题。比特币本身到底是否有价值不是我们讨论的话题，它之所以能够在一定程度上起到货币的作用，并且成为全球很安全的洗钱工具，源于背后的一项技术——区块链（Block Chain）○

区块链由两个英文单词Block和Chain组成，顾名思义，它应该包含两个方面的意思：Block即模块、单元的意思，它像一个账户存储信息；Chain是链条的意思，即表示一连串的交易；交易的细节就存在Block中。比特币实际上是一个由随机数算法产生的随机数，这个随机数在整个互联网上是唯一的，而且是可以验证其真伪的。比特币在被挖矿者挖出来时，就产生一个带有这样特殊随机数的Block，当这个比特币通过交易转到第二个人手里时，在该Block中就记录下了交易的信息，这个过程本质上是一个加密的信息传输过程。一旦交易完成，它就被通知到整个互联网上，大家就知道相应的比特币的拥有者改变了。所有比特币散布在整个互联网中，通过公开密钥来发送和传播，拥有者和交易的过程都是匿名的，而且没有一个中心能够集中控制，因此特别适合洗钱。

既然比特币的这种区块链可以记录钱的交易，那么也应该能记录其他的交易和传输。如果在每一件商品制造出来时（或者出厂时）产生这样一个区块链，并且在它被运输和交易时利用区块链记录全过程，那么这个商品整个的流通过程就是可追踪的。当最终的消费者（顾客）购买这个商品后，他可以看到这个产品是如何从出厂开始一步步卖到自己手上的。这样，从理论上讲可以杜绝假货，因为区块链和商品是一一对应的，既然不可能产生两个相同的区块链，也就不可能复制同一个

商品。类似地，厂家也可以了解到它每一件商品是怎样流通到最终消费者手里的。

要完成上述的过程还需要使用本书前文介绍过的RFID技术,来自动记录每一件商品的流通和交易的全过程。最终，区块链和RFID等技术的应用不仅会使我们未来的社会完全是一个精细化的、智能的社会，而且会让今天很多我们不敢想象的事情变成现实。图7.2以区块链为基础的比特币交易示意图。每个比特币被挖矿者挖出时，就产生一个区块（表格）记录各种信息，每一次交易(链接)的情况就被记录在区块中

从标准化到个性化的服务

我们在上一章讲到通过个性化制药为每一个人定制特效药品，这样能治愈癌症并延长人的寿命。其实，在医疗领域，不仅用药可以个性化，整个行业的服务都应该是有针对性的，这样可以最高效地利用医疗资源并且最切合地为每一个人服务。

今天医疗领域存在两个怪现象。首先，一方面一些没有什么大病的人要费很大劲找一个专家看病，他们为了做到这一点要么一大早排队挂号，要么托关系走后门；另一方面确实需要有经验的专家看病的那些病人却得不到相应的医疗资源。其次，另一个怪现象是，那些千方百计找专家看病的人，实际上也不知道找哪个专家，只能根据他们的头衔找所谓最好的，而专家们也常常发现患者其实应该找其他专家而不是他们自己看病。虽然这里面有专家号太便宜的原因，有患者无知的因素，但是没有足够的信息，以及缺乏一个很好的医疗顾问进行就诊指导也是其中一个重要的原因。

在一个个性化和智能化的社会，上面的问题可以得到很好的解决。一方面由于每一个人都积累了非常完整的与自己健康状况有关的数据，医院、医生甚至患者本人对自己的病情都会有比较清晰的了解。另一方面由于有了比较完备的医疗从业者的数据，智能的就诊指导系统会根据患者的情况和医生的情况帮助他们选择合适的医生。这样患者在小病时不需要折腾自己，真遇到大病时能更容易地找到合适的医生。

其实，今天大家用药和就诊这件事透露出工业时代的一个特征，就是一切标准化。在工业革命开始以前，人类使用的产品、享受的服务都有细微的差别，当然这样效率很低。在近代医学开始之前，每一个人的用药都是不同的，尽管那种差异未必有科学根据。工业化的一个结

果，就是靠批量生产的效率让个性化从大众市场消失了，不仅产品是标准化的，服务也是如此。比如在医疗方面，美国医生协会要求每一个从业者遵守流程。对医院来讲，医生宁可治不好病，也不能违背流程，因为如果违背流程引起官司，医院的损失可能是巨大的。虽然不能说标准化的产品和服务不好，但在很多情况下对顾客肯定不是最优化的。然而，在工业社会里，要获得个性化的产品和服务成本太高，除了个别富人愿意花非常高的代价去享受这样的产品和服务，一般人是享受不到的。在大多数产品和服务都被标准化的时代，大家很难找到最适合自己的，只能默认最权威的或者最贵的就是最好的。这也是大家在就诊时普遍认定教授比副教授好，副教授比主治医师好的原因。

到了智能时代，机器的智能水平足以以为我们提供各种个性化的服务，同时能够做到成本和过去的标准化服务相当。这使得我们在今后可以享受到个性化为我们带来的生活的巨大改善，那是今天所谓富有的上层人士才能享受到的生活。因此，大数据和机器智能可以让我们整体的社会环境乃至文明程度都有质的飞跃。但是，在另一方面，大数据也会给未来社会带来巨大的冲击，这就是我们接下来要讨论的内容。

无隐私的社会

到目前为止，我们一直在讲的是大数据和智能革命对社会、对我们的生活所带来的正面影响。但是任何事情一定都有两面性，大数据和智能革命对未来社会的冲击也是不能小视的，我们或许会生活在一个没有隐私的环境里，或许会被一些超级权力在无形中控制，甚至很多人因为没有掌握未来生存的技能而找不到工作，财富可能会更加集中在少数人手里。根据历史的经验，这些问题是无法回避的，而且也不存在快速的解决方法。让我们先来看看大数据和机器智能对于个人隐私的影响。

虽然我们在前面的章节里也提到了隐私的问题，但只是讨论技术问题，而对

隐私的重要性以及全社会所面临的挑战一笔带过。实际上，大数据和机器智能引发的隐私问题会非常严重，在今天和未来，当移动互联网（以及正在快速发展的物联网技术）、大数据和机器智能三者叠加到一起之后，我们不再有隐私可言。

虽然媒体在谈到隐私时常常讲的都是是一些个人不愿为大众所知的信息，比如私密的图片或者银行账号的密码，但是这里我们所讲的是一个非常广泛的概念，涉及我们生活的方方面面。比如在前面提到的最近是否必须到某城市出差这样的信息，个人性格是软弱还是强悍，学历的高低和收入的多少，这些都是我们的隐私。私密的图片流出去会让我们难堪，银行账户密码丢了会让我们蒙受经济上的损失，这些损失是看得见的，我们因此非常在意，也就或多或少地有所防范。但是，关于我们生活方方面面的细节隐私，我们常常很不在意，也不加以保护。

这些生活细节的隐私泄露出去会发生什么事？简单地讲，将来可能很麻烦。这不仅仅是在淘宝上总被送来一些假货，或者买机票总是比别人贵20%那么简单，它可能涉及我们的健康和医疗，可能没有医院会接收我们住院。在美国加州大学的一所医学院里，科学家们正在做这样一项研究：每一个人从小到老生病的规律性。比如我们知道得了丙型肝炎，即使暂时治愈，还是有很大的可能性在若干年后转成肝硬化，然后又有很大可能性变成肝癌。其他很多疾病也有这样的关联关系。这家医学院研究的当然善意的一一为了提前防治疾病。

但是，研究成果如果让医疗保险公司使用，那么它们就有权利拒绝接受一位未来可能得重病的投保者。美国各大保险公司实际上掌握着投保人过去多年的身体状况信息，因为医生每一次向保险公司索要医疗费时，都会提供这些信息。在过去，由于机器智能的水平不高，这些事没法做，保险公司一般对投保者一律接受，但是在法律上，它们有拒绝投保人的权利。

在过去，我们泄露隐私有时是不得已的，比如不能不去看病，而医生也不能不去问保险公司要钱。但是在移动互联网时代，尤其是今后万物联网的时代，我们本身就是主动的隐私泄密者。绝大部分智能手机的使用者安装了太多的、很少使用甚至并不必要的APP，参加了太多的优惠促销活动。同时，在自认为安全的社交网络说了很多在公众场合不适合说的话，或者发了太多的照片。这些都可能造成人为的隐私泄露。我们还在使用的各种电子产品，从可穿戴式设备到带有GPS的照相机，再到与Wi-Fi相连的各种智能电器，不自觉地记录下了我们详细的行踪和生活信息，并且提供给了服务商。很多时候，第三方再通过服务商获得这些信息，也并非难事，究其源头，是我们自己在不设防的情况下把信息泄露出去的。

拥有数据的公司保护个人隐私的意愿远不如大家想象的那么强。除了Google、苹果、亚马逊等大型跨国互联网公司迫于欧盟和美国政府的要求（当然也是为了让它们大量的客户安心），在服务条款中特别明确地写明了从用户获得的数据属于用户本人，而它们只是保存和“借用”而已，其余的公司都没有明确声明这一点。在医疗行业，美国绝大部分医院会认为病人的病例数据属于医院，这也是该行业的传统。在中国，互联网公司并没有就数据的所有权做明确的说明，而大部分用户也默认互联网公司拥有数据。更有一些制药厂在没有得到病人同意的情况下，直接通过医生获得病人的数据用于药品研究，而这件事被认为是有助于医学研究，因此社会并没有追究。

今天很多人忽视大数据对个人隐私潜在的威胁，原因至少有以下三个：首先是对这个问题缺乏认识，他们并不知道大数据的威力，不知道多维度的信息凑到一起能够得到一个人完整的画像。其次是低估了机器智能的力量。很多人认为，虽然某个公司即使有了关于我的很多数据，但是那些数据都是杂乱无章的，该公司哪有工夫专门和我这个小人物过不去。岂不知在机器智能时代，挖掘个人隐私并不需要人来做，而是由机器完成的。最后，也是最重要的原因，就是很多人一厢

情愿地把个人隐私寄托在数据拥有者的善意(Goodwill)上。虽然到目前为止，Facebook、腾讯和阿里巴巴这些实际上已经掌握了用户隐私的公司似乎还靠得住，但是掌握了大量用户数据的公司远不止这几家。当掌握大量用户数据的公司和用户利益发生冲突时，前者会有意无意地最大化自己的利益，而牺牲掉用户的利益。我们从前面的一些案例中看到，把我们的隐私权建立在别人的善意上，是根本靠不住的。像航空公司、保险公司在获得个人隐私后，并不是把用户隐私暴露给大众让用户出丑，而是谋取利益，用户拿这些公司也没有办法。约翰·霍普金斯大学工学院院长施乐辛格(Edward Schlesinger)教授本身是大数据的倡导者，但是他也担心，如果保险公司能了解到每一个人今后会得什么病，将拒绝给那些可能得致命性疾病的人提供保险，那么那些最需要医疗保险的人反而无法买到医疗保险，或者必须支付天价保费。

既然我们不能指望我们的隐私靠一些公司的善意来保护，那么是否有希望通过立法的手段来解决保护隐私的问题，答案基本上是否定的。首先在大陆法系的国家[2]，立法永远是远远滞后于案件发生的。当科技和产业变化比较慢时，这不是什么大问题。假如产业变化的周期是几十年一变，就算立法落后了5年，产业还没有太成熟，依然可以利用法律的手段把后几十年管好。但是今天产业发展太快，如果立法的速度真落后了5年，当法律被制定出来后，已经过时了。更何况今天，大部分法律制定的时间远不止5年。比如中国的电子商务在过去的几年里迅速发展，与此同时卖假货的问题也已经发展到不容忽视的地步，但是中国至今没有相应的集体诉讼赔偿法规[3]和有效的执法手段。因此，目前在中国是无法靠法律手段杜绝假货横行的。

~'|

图7.3无处不在的监控

今天，世界各国虽然都对偷盗行为进行惩处，但是对于偷盗数据和利用大数据侵犯个人隐私的行为，并没有相应的立法。在美国，虽然有一些具有法律意义的判例，但是处罚也是相当轻的，对于偷盗数据的处罚和对于抢银行的处罚是无法相比的。我们可以毫不夸张地讲，今天的法律对保护隐私几乎是无效的。当人们开始逐渐意识到隐私的重要性时，可能会对大数据和机器智能产生恐惧，这对技术的发展并非好事。



图7.4大家出于对信息时代没有隐私的恐惧，引用奥威尔《1984》一书中“老大哥在盯着你”那句话画成漫画

大数据对隐私带来的另一个威胁在于，它会在无形中造就出一个老大哥 (Big Brother)。Big Brother一词来源于英国小说家乔治·奥威尔 (George Orwell, 1903-1950)的政治幻想小说《1984》，那里面有一句话是“Big Brother is watching you”。Big Brother是指专制政权里的老大，那句话放在小说语境中的含义是指，总有一双眼睛在盯着你。在冷战时期，只有1000万人口的东德倒有10万监视老百姓的安全部工作人员和20万线人，他们用很传统的笨办法监视每一个人，比如拆私人信件，这使得每一个人都生活在恐惧中。当然，这种做法的效率不会很高。

到了大数据时代，如果真有一个老大哥想监控每一个人，其实是可以做得到的，他也不需要采用东德安全部的笨办法，因为大家的隐私都保存在互联网的某处。假如出现一个强权，要求拥有大数据的服务提供商交出数据，建立在善意基础上的隐私保护就显得非常脆弱了。民众即便不懂得什么是大数据，不懂得大数据容易泄露隐私，对强权部门索要数据的事情也是非常担心的。

2016年，FBI（美国联邦调查局）要求苹果公司交出某些用户数据，以配合反恐调查。苹果公司如果迫于压力交出这些所谓嫌疑人的数据，这个先例一开，今后权力机构再以其他借口随意索取用户数据，那么大家就不再有隐私可言。正是出于对这个原因的考虑，苹果公司才拒绝向FBI交出数据。好在美国公司的正常商业运行不受FBI的干扰，苹果公司也

大到足以抗衡FBI，最终FBI只好放弃。

但是，是否所有拥有大数据的公司都会拒绝将用户数据交给美国政府，谁也不能保证，至少作为微软董事会主席的比尔■盖茨公开表示应该交出数据。也就是说，如果他还在负责微软的经营，使用微软产品的用户可能就没有隐私可言。我们把命运寄托在一些公司的善意上其实并不可靠。如果一家公司或者政府部门有能力获得和随意使用每一个人的隐私，那么它就拥有了某种超级权力。更进一步讲，如果拥有用户大量私密数据的公司同时具有了超级机器智能水平，那么它不仅拥有权力，而且还拥有超级执行力。历史证明，任何不受约束的超级权力最后都会带来灾难。如果真到了那一步，大数据和机器智能的负面效应

就会变得非常大。



图7.5苹果公司对决FBI,成为美国社会关注的焦点

尽管我们还可以试图通过技术手段争取做到在使用大数据时无法看到与用户隐私有关的数据内容，同时能够尽可能地防止大家在互联网上“偷窥”他人隐私的行为，但这些技术的开发有待时日。今天的大数

据是完全裸露的。以人们对 隐私问题最担心的医疗大数据为例，使用数据时对于隐私的保护现在依然是靠 君子协定，也就是说处理和使用数据的人签了一纸协议就被允许访问隐私数据。虽然那一纸协议可能具有法律效力，但是患者其实很难判定掌握自己隐私的人是否违反了事先的协议，因此如果有人违反协议使用隐私数据，患者很难状告那些破坏隐私的人。

按照目前大数据的发展趋势，大家会越来越没有隐私，而当我们体会到丧失隐私后的重大损失时，为时已晚。隐私就像自由，只有当人们失去它的时候，才知道它的可贵。

机器抢掉人的饭碗

技术对社会带来的影响有时候非常诡异。一方面它可以改善人们的生活，延长人类的寿命，让一些处在新的行业、掌握了新的技能的人发挥更大的作用；另一方面则可能让更多的人无事可做。智能革命也必然如此，当计算机变得足够聪明之后，一定会取代人类完成很多需要高智力的工作。

人类总体来讲是过分自信的，趋利而忽视危害，这一点研究幸福学和心理学的学者早就有了定论，我们不做过多的讨论。机器智能如此翻天覆地的革命，不可能不对社会产生巨大的负面影响。我们在给大家展示大数据和机器智能带来的美好前景时，也必须强调它们可能会给很多人的生活带来负面影响。不过遗憾的是，很多人对此不以为然，就如同历史上工业化国家的民众曾经的不以为然一样。当社会面对重大技术革命所产生的冲击不知所措，要两代人才能消除它的负面影响时，大家才开始感叹历史再一次重复。智能革命将比过去历次技术革命来得更深刻，对社会带来的冲击可能是空前的。为了说清楚这一点，我们首先来回顾一下历史。

历史上影响力可以和正在进行的智能革命相比的，只有19世纪末始于英国的工业革命、20世纪末始于美国和德国的第二次工业革命、“二战”后以摩尔定律为标准的信息革命，一共是三次。这三次技术革命都有一个共同的特点，那就是它们对当时的社会产生了巨大的冲击，都需要经过大约半个世纪甚至更长的时间才能消化掉。

从工业革命到黄金时代

首先让我们看看19世纪末的工业革命。这是人类历史上空前的伟大事件，任何其他历史事件在人类文明史上的重要性都不能和它相比。工业革命带来了三个结果：人类过得好了，人类活得长了，人类有自信和尊严了。

在工业革命开始前的两千年里，世界各地人们的生活水平其实没有太大的提高。根据已故著名历史学家安格斯·麦迪森（Angus Maddison, 1926—2010）对全球各个文明在不同历史时期所做的经济学研究可知，欧洲在古罗马时代的人均GDP就达到了600美元左右[4]，到了18世纪英国工业革命之前，人均GDP还是这么多。在中国西汉末年，人均

GDP大约为450美元，在历史上的几个太平盛世，比如两宋时期、明朝中叶和康乾盛世，中国的人均GDP达到了600美元，但是到了20世纪50年代初又退回为450美元左右，就在1979年改革开放前，中国的人均GDP也 不过800多美元[5]。虽说人均GDP未必能够完全体现人类的文明进步，但是在这么长的时间里变化不大，说明在农耕文明时期人类的进步是非常缓慢的。在工业革命之前几千年的时间里，劳动力的数量和能够提供给生产所使用的动力整体上是不足的，商品是供不应求的。

但是，到了工业革命之后，情况就大不相同了。马克思说：“资产阶级在它不

到100年的阶级统治中所创造的生产力，比过去一切时代创造的全部生产力还要多，还要大。”[6]如果用人均GDP量化地衡量一下就能发现，在南欧、西欧和北欧地区，工业革命开始以后，从1800年到2000年这200年间，人均GDP水平增长了将近20倍—从1000美元左右增加到20000美元。而中国在改革开放后的35年里（1979～2014年），人均GDP在考虑购买力以后也上涨了不止10倍，如果不考虑购买力，则上涨多达40倍，其根本原因是中国在1979年之后才真正完成工业革命，并且用35年的时间走完了欧洲花200多年走完的路，从农耕时代一直走过了早期工业时代、大工业时代和后工业时代（信息时代），并与世界同步进入后信息时代。在财富持续增长和收入不断增加的同时，工业革命也导致了人类寿命的大幅提高。可以说，如果没有工业革命，任何伟大的人物都无法做到让人类活得更好。

工业革命的影响力不仅体现在物质层面上，更体现在思想层面上，让人类有了自信和尊严。我们在前面讲到的机械论的出现，使得人类有了把握自己命运的自信。

几个世纪后再回过头来看这样一场伟大的变革，它带来的好处自然要远远大于它的负面影响，但是在当时，它的负面影响，尤其是它给社会带来的动荡是巨大的，以至于当时诅咒它的人可能比欢呼拥抱它的人更多。新技术在出现的初期，受益者是非常少的，他们通常只是那些掌握新技术或者使用新技术、从事新行业的人。具体到工业革命，最初的受益者只有博尔顿那样的工厂主、瓦特那样的发明家，或者使用蒸汽机开拓瓷器制造新行业的韦奇伍德等人。其他人在短期内是很难受益的，甚至可能因为新技术的出现变得更加贫穷，因为机器抢了他们的生计。

在工业革命后的半个世纪里，原有的经济结构被摧毁，靠有一技之长的工匠运作的小作坊纷纷破产，工匠的特长敌不过年轻劳工结实的身躯，他们从中产阶级沦为赤贫。因此从18世纪末到19世纪上半叶，是英国贫富分化严重、社会矛盾重重的半个多世纪。著名作家狄更斯用他生动的笔，记录了当时下层民众悲惨的生活，这与飞速发展的经济和暴涨的社会财富并不相称。为了节省成本便于竞争，工厂主们大量雇用低工资的童工，或者随意延长劳动时间。也正是在那个年代，英国出现了空前也是绝后的工人运动，催生出马克思主义。

英国人花了大约两代人的时间消化工业革命带来的负面影响。到了1851年，英国在伦敦郊外的水晶宫举行了第一次世博会，展示工业革命的成功，当时的维多利亚女王看完展览后，嘴里不住兴奋地念叨着“荣光啊，荣光，无尽的荣光”。后世称那个时代是英国的黄金时代，那个时代的英国人过上了全民富裕的生活。大部分人都有体面而收入不错的工作，工作时间减少到了每周48小时，童工被禁止。当时，一半的人口搬进城市，剩下的人很多在郊区买到洋楼，然后坐火车到城市和工矿区上班。周末大家可以穿着漂漂亮亮的礼服去教堂或者去逛商店。

那么工业革命的副作用是怎样被解决的呢？简单讲就是资本输出，开拓全球殖民地，推行自由贸易。英国的工业生产在工业革命之后让世界各国都无法望其项背，这使得它有能力、财力、武力按照自己的意志建立全球化市场。英国工业革命产生的产业工人只有几百万，但其巨大的生产能力却使得很多商品供大于求。由于在当时世界上没有第二个国家在国力上可以和英国匹敌，因此它的全球战略得以实施。



图7.6维多利亚时代，英国的教育已经非常普及

我们可以把工业革命对社会的影响分成三个阶段：第一个阶段只有发明家和工厂主们受益，普通英国民众并没有受益；第二阶段是全体英国民众普遍受益，但是在世界范围内大家未必受益，这两个阶段之间相差半个多世纪；第三个阶段才是整个世界受益，这和第二个阶段又相差很长时间。是否其他重大技术革命也有类似的特点呢？让我们来看看19世纪末的第二次工业革命，有趣的是，上述的模式重复出现了。

从第二次工业革命到镀金时代

第二次工业革命的核心是电的使用。这不仅让生产的效率进一步提高，而且催生了很多新产业，当然这也带来了社会财富的剧增。著名作家马尔科姆·格拉德威尔 (Malcolm Gladwell) 在《异类》_书中介绍了这样一个事实：在人类历史上最富有的75人中，有1/5出生在1830~1840年的美国，其中包括大家熟知的钢铁大王卡内基和石油大王洛克菲勒等。这一不符合统计规律的现象的背后有其必然性，卡内基等人都在自己年富力强 (30 ~ 40岁) 时，赶上了美国工业革命的浪潮，这是人类历史上产生实业巨子的高峰年代。其中洛克菲勒被认为是人类历史上最富有的人，而他年长一些的范德比尔特则一度通过建立托拉斯 (Trust, 信托) 控制了美国上市公司10%的财富。类似地，欧洲的很多工业巨子，比如克虏伯和西门子，也是那个时代的人物。

但是，和工业革命时期的英国一样，美国工人们的生活在第二次工业革命开始的一段时间里并不美好，当时美国的贫富分化程度达到了北美殖民以来的最高点，而且比今天严重得多。一方面，美国下层社会的生活非常悲惨，他们的生活和范德比尔特等人形成非常鲜明的对比，马克·吐温和西奥多·德雷塞 (Theodore Dreiser)[7]等现实主义作家对那个时代劳工的生活都有真实的描述。因此，美国历史上不多见的激进的工人运动也发生在那一段时期。另一方面，美国南方的传统经济被北方的大工业彻底碾碎了，并没有因为第二次工业革命而受益。直到今天，美国南部的经济 (除得克萨斯州外) 依然远远落后于北方。



图7.7 19世纪末美国的工人运动，后面

是警察在镇压

当美国和德国崛起时，它们已经没 有英国那么好的运气，有那么多未开发 的殖民地在等着它们。好在美国有它天 然的地理优势，它有广袤的中西部处女 地等待开发，从某种程度上解决了产能 的问题，但是贫富差距非常严重。运输业 大王范德比尔特通过建立信托控制了 10 % 的上市公司财富，而洛克菲勒聚集的 财富占全美国的1 %。为了实现社会的公 平化,美国开展了坚决的反托拉斯行动。 经过老罗斯福、塔夫脱和威尔逊三任总 统近20年的努力，美国政府强行肢解了 洛克菲勒的标准石油公司和JP摩根控 制的北方钢铁公司，并且在制度上限制 大家族过多地控制社会财富，比如征收 高额的遗产税。从1870年美国第二次工 业革命开始，到19世纪20年代，经过半 个世纪的努力，美国才基本实现了全面 繁荣。19世纪20年代被称为美国的镀金 时代，或者“柯立芝繁荣”。由于生产效 率的极大提高，美国实现了9小时（后来 是8小时）工作制[8]。到1929年大萧 条之前，美国一半的家庭有了电话和汽 车。但是，德国就没有美国那么幸运了， 为了输出产能，它最后不得不发动第 一次世界大战。在“一战”战败之后，德国 的问题并没有得到解决， 于是导致了民粹主义泛滥，最终劳工阶层把纳粹推上 了台。

今天我们站在历史的角度审视第二 次工业革命，对它都是赞誉之词,它 的代表人物爱迪生、贝尔、福特、西门子和本 茨等人，直到今天依然是创业者和企业 家们的偶像。但是它给人类带来的福祉 也是先从少数精英开始，经过长达半个 世纪的时间，才开始造福技术革命的中 心地

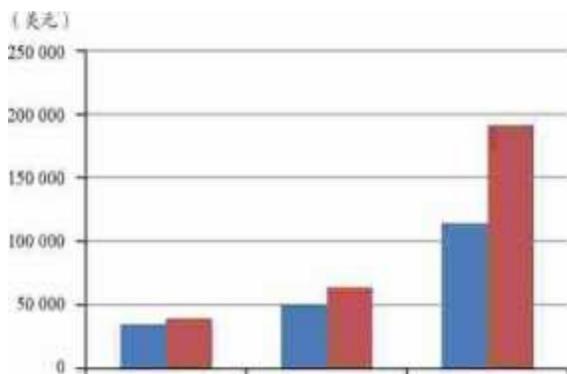
区。而世界上大部分地区享受到第二次工业革命的成果，是第二次世界大战之后的事情。

依然没有消化完的信息革命

到了“二战”后的信息时代，上述模式再次得到应验。我们都有幸亲历信息时代的繁荣，我们有了个人电脑、手机、互联网，我们的生活变得比父辈要方便得多，几乎每一个中国人都在为信息革命欢呼。中国在1979年改革开放后短短的30多年里，人均GDP从1978年的200美元剧增到2014年的7000美元，几乎每一个人的收入都有所增加。但是在过去

的30多年里，中国只是全世界的一个特例而已。中国的成功有多重原因，最根本的是它的起点比较低，生产力和创造力在被压制了几百年后被释放了出来，在短时间里爆发出巨大的能量，再加上同时完成了工业化和信息化，所有这些有利的条件叠加在一起，才导致中国无论从总体国力还是人均收入，都有大幅度提升。但是，在世界范围内，虽然每个人都看到了信息革命的结果，并且很多人使用上了最新的科技产品，然而并非每个人在经济上和社会生活方面都受益于此。即便在信息革命中心的美国，大部分人的生活质量并没有什么提高。

信息时代是人类历史上第二个创造财富的高峰年代。在美国，从20世纪50年代末到70年代初的20年间，出生了苹果公司创始人史蒂夫·乔布斯、微软公司创始人比尔·盖茨和保罗·艾伦、太阳公司创始人安迪·贝托谢姆和比尔·乔伊、戴尔公司创始人迈克尔·戴尔、Google创始人拉里·佩奇和谢尔盖·布林等人，他们在自己年富力强时幸运地赶上了信息革命的大潮。但是，美国大众的生活质量并没有很大的改变。图7.8展现了美国最富有的5%的家庭、财富值中值的家庭，以及贫困家庭从1967年到2012年(扣除通货膨胀后)财富增长的情况。我们可以看出，除了最富有的5%的家庭财富有明显增长之外，其他人的财富变化很小。



50%~75% 251^50% TOP5%

图7.8 1967~2012年美国家庭收入变化

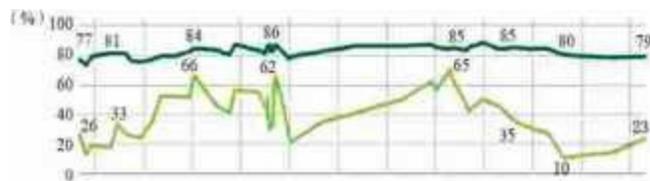
数据来源：美国国家统计局

当然，财富是社会发展和个人生活质量的一个客观标准，但绝非唯一的衡量标准。幸福指数常常被认为是生活质量的一个主观衡量标准。如果以它来衡量，美国民众的生活在过去的30多年里几乎没有什么改善。图7.9是盖洛普公司对美国民众幸福指数的调查结果。上方

的绿线是被调查者对整体生活的满意程度，从1980年到2013年它基本上持平；下方的青线是大家物质生活的满意程度，2013年比1980年还有所下降。我们可以认为，这些数据表明以摩尔定律为核心的上一次技术革命带来的社会效益，即便是作为全球信息革命中心的美国仍然没有来得及消化完。而中国作为全球信息革命的另一个中心，由于我们前面所讲的特殊情况，不太感受到它的负面影响。

对个人生活的满意度与对国家经济总体状况的满意度

■对整体生活的满意度 ■对物质生活的满意度



1980 1983 1986 1989 1999; 1995 1998 2001 2004 2007 2010 2013(19)

图7.9在过去的30多年美国民众幸福

指数的变化 数据来源：盖洛普

在过去的30年里，美国和中国两个国家贡献了全球超过一半的GDP增长，

除去这两个国家，世界上大部分地区的情况可不大美妙。我们从新闻里时常会看到，包括俄罗斯在内的很多国家似乎置身在时代之外。虽

然这一点有很多政治上的解释，但是从经

济和科技发展的角度看，以苏联为核心的东欧集团、超过10亿人的穆斯林地区、大部分欧洲国家、整个南美洲，对于信息革命的贡献微乎其微。它们自有的旧的经济结构已经落伍，甚至被摧毁，而在新的经济结构中，它们虽然能够享受到信息革命的产品，却没有享受到信息革命带来的经济增长。从全世界的范围看，消化掉信息革命的冲击波，或许还需要更长的时间。然而现在大数据和机器智能革命已经来敲门了。

解决问题只有靠时间

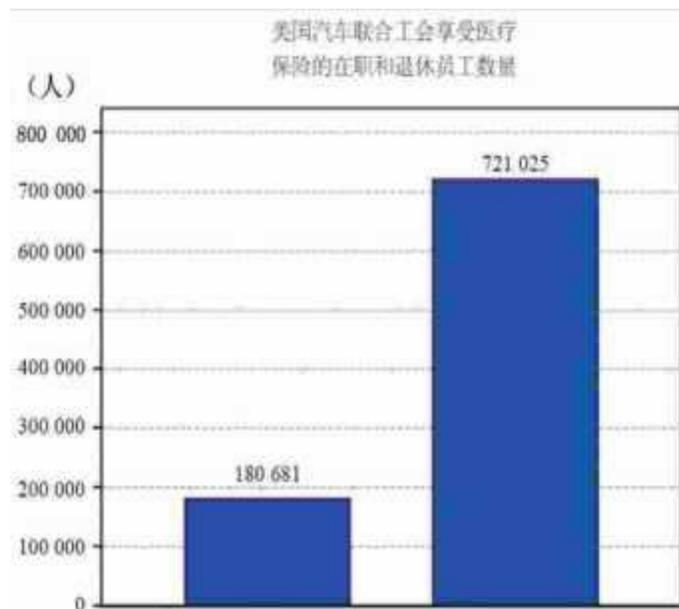
为什么每一次重大的技术革命都需要很长的时间来消除它所带来的负面影响呢？因为技术革命会使得很多产业消失，或者产业从业人口大量减少，释放出来的劳动力需要寻找出路。这个时间有多长呢？事实证明至少要一代人以上，因为我们必须承认一个并不愿意承认的事实，那就是被淘汰的产业的从业人员能够进入新行业中的其实非常少。

虽然各国政府都试图通过各种手段帮助那些从业人员掌握新的技能，但是收效甚微，因为上一代人很难适应下一代的技术发展。事实上，消化这些劳动力主要靠的是等待他们逐渐退出劳务市场，而并非他们真正有了新的出路，能够和以前一样称心如意地工作。这就是每次技术革命都需要花半个世纪来消除它带来的动荡的原因。唯一不同的是，在一百年前，各国政府认识不到关心这些被产业淘汰的从业人员的重要性，因此让社会很动荡。如今，各国意识到社会稳定很重要，因此即使很多人并不创造价值，也只好“养着”。为此，有些国家将无所事事的人强制塞到公司里（比如日本和欧盟），有些国家不肯淘汰过剩产能（比如中国），但解决问题的途径都是一个“耗”字。耗上两代，社会问题就解决了。

要更好地了解产业转型以后，消化

原有产业的从业人员有多么难，我们不妨再看看美国“二战”后发展的历程。整个20世纪50~60年代，全球规模和市值最大的公司是通用汽车公司，它和另外两家美国汽车公司一道，生产了全球90%以上的汽车，仅在美国就有70万名雇员。由于通用汽车公司的福利很好，它的每一位员工都过着幸福的生活，都能实现所谓的美国梦。今天，通用汽车公司虽然生产同样多的汽车[9]，从业人数却减少到10万以下，这是劳

动生产率提高的结果。当然，很多人以为劳动力可以转移到其他行业，但事实上并没有。我们在前一章介绍过，即便是一直不断扩大人数规模的特斯拉这样新的汽车公司，也不愿意聘用汽车行业淘汰下来的人。因此，那些汽车行业的老人只能靠工会养着。在2008年的金融危机之中，通用汽车公司宣布破产保护，其中主要的问题就是公司要养的人太多。图7.10是2008年金融危机之前美国汽车联合会中在职员工和退休员工[10]的比例，我们可以看出，1个在职工人需要养活4个不干活的人。



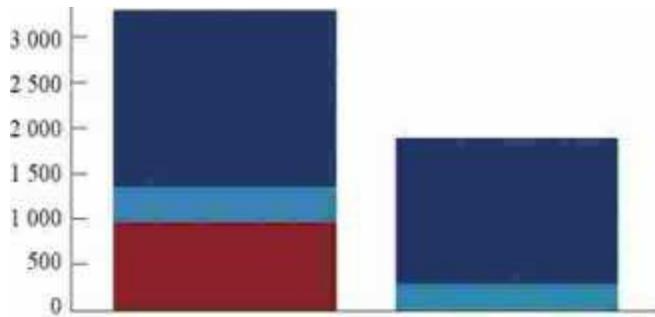
在职员工 退休员工及工

会成员的M

图7_ 10 2008年金融危机之前美国汽车联合会在在职员工和非在职员工的人数

通用的这种做法导致其汽车的成本上升，从而失去了全球竞争力。图7.11对比了在北美销售的通用汽车公司和日本丰田汽车公司每制造一辆车的福利成本，可以看出通用汽车公司大约高出了1500美元，这对平均售价只有2万美元的汽车来说是非常明显的差异。其中，福利上最大的差异来自通用公司每辆车要支付1000多美元的退休员工福利。这就是企业因为全社会技术进步和产业转型而不得不支付的成本。

3 500 厂



通用 丰田

■退休相关人员福利 ■在职人员福利 ■工资

图7.11通用和丰田在北美销售的汽车成本中的福利对比 数据来源：汽车研究中心

虽然汽车行业因为电动汽车、自动驾驶汽车等新产品不断涌现，依然有很大的发展空间，但是美国三大汽车公司[11]背负着历史的包袱，举步维艰，在未来难以有所作为。对那些曾经为人类的文明做出过贡献，但已经被技术革命所

淘汰的员工，唯一的希望就是他们的后代能够进入一个新的行业。这实际上是靠时间慢慢地消化技术革命带来的负面

影响。

智能革命的冲击

智能革命将要走的路和历史上历次技术革命的路会有很多相似之处。大数据和机器智能的趋势一旦形成，就不是人力可以阻挡的。虽然一些有识之士，包括霍金、盖茨和马斯克等人担忧机器智能将会对人类社会造成方方面面的冲击，并且呼吁有节制地发展机器智能，但是智能革命的速度不会因此而放慢。甚至就连他们自己在利益面前可能也是口是心非——马斯克的特斯拉本身就是大量使用机器人的公司，而盖茨的微软也是在机器智能领域投入最多的公司之一。与之前的三次重大技术革命一样，智能革命对社会的冲击将是巨大的，它会影响到上至国家、中到企业、下至个人的命运。从目前的发展来看，智能革命对社会的冲击甚至有可能超过过去几次技术革命。我们可以从三个角度来分析其中的原因。

首先，信息革命本身带来的影响还没有消化完。全球信息化带来的效率已经使得很多人无事可做，很少人制造出来的东西就足够全球人口消费。在美国将近一半的人是不上税甚至从政府拿补贴的，从单纯经济的角度看，他们每天所提供的劳动仅仅是让自己生存下去而已，甚至还不够，他们对社会继续发展的贡献可以说是微乎其微的。在一个民主国家，这些人最大的用途就是手中的那一张选票，以至于政客们为了选票可以轻易许诺，然后把国家的债务和赤字越堆越高。第一次和第二次工业革命带来的负面影响都花了半个世纪以上的时间来消除，而摩尔定律从1965年提出距今已经半个多世纪了，它带来的影响至今还没有消化掉。这时，智能革命又开始了，因此这次的冲击力度将是双重叠加的结

其次，今天的世界和200年前已经不同了，消化掉技术革命的影响要比工业革命时难得多。由于全球化，全世界已经没有空白的市场可以开拓了。英国人在19世纪中期能够过上相对富裕而从容的生活，是因为他们只需要解决几百万产业工人的生活和工作问题就可以了。整个19世纪，是用全球的市场，解决当时只占世界人口很小一部分的产业工人的生活问题，相对要比今天容易得多。

最后，也是最重要的一点，智能革命所要替代的是人类最值得自豪的部分——大脑。以前，当各种各样的机器可以越来越多地从事人类才能做的工作时，人类还保留了最后的尊严和自豪感——机器不能思考。过去机器只是替代人的手，因此在农机和化肥出现后，农村从事体力劳动的人可以变成需要动脑筋的工匠；在流水线出现之后，工匠们没有了市场，但是蓝领工人可以从事白领的差事。由于机械毕竟不能完成智能的工作，因此人们最终还是找到了谋生的手段。不过智能革命的结果是让计算机代替人去思考，或者说靠计算能够得到比人类思考更好的结果，能够更好地解决各种智能问题，这时，人类会突然发现自己还能做得比计算机更好的事情已经所剩不多了。我们在上_章介绍过，智能革命中，计算机所取代的不仅仅是那些简单重复性的劳动，还包括医生、律师、新闻记者和金融分析师等过去被认为是需要脑力的工作。

概括来讲，智能革命对社会的冲击可以用强度更大、影响面更广、更深刻来概括。我们必须回答一个问题：当全社会各行各业的从业人数都因为机器智能而减少时，全世界几十亿劳动力怎么办？

当然，很多人会天真地认为，船到桥头自然直，劳动力会被自然而然

地分配 到其他行业中去。但是，这种劳动力的再 分配，一来需要非常长的时间，二来依赖于产生新产业。关于时间的问题我们在 前面已经讨论过了,这里不再赘述。接下来我们来看看产生新产业的必要性及其难度。

在工业革命开始之后，机械化、电气 化和化肥农药的使用，使得发达国家只 需要2% ~ 5% [12]的人就能提供全部 人口所需的食物，因此农民就变成了工 人。虽然这个转化的时间很长，但是很多 国家基本上实现了_比_的转化，也就 是说在减少一个农民的同时，社会能够 创造出_个新的就业机会给他。但是，随 着机器革命的发展和全社会自动化程 度的提高，只需要少数的劳动力就能提供 人类所需的所有工业品和大部分依靠体 力的服务业工作。因此，全球开始了第二 次劳动力大转移，在过去的几十年里，人 类就业的希望从在工厂做工人变成从事 服务业。

服务业其实是一个非常宽泛的说 法，它既包括律师、医生、IT工程 师、股 票交易员和基金经理这一类收入和地位 较高的职业，也包括超 市、餐饮、旅游等 工作性质简单，收入水平一般的行业。其 中，第一类只占很小一部分，而且需要高 智力和长期职业培训才能胜任工作。大 部分所谓的服务业，收入可不如过去生 产线上的工人。

在1900年前后，美国东北部的波士 顿地区的人只要有一份工作，就能 在波士顿市内或者查尔斯河对岸的坎布里奇 [13]买一栋连排别墅(Town House) 。今天，那里的人需要在Google或者辉瑞 制药公司里有一份非 常好的工作，才能 买得起同样水平的住房。在20世纪60年 代，通用汽车公司一家就造就了近百万 个中产阶级家庭。今天，全球市值最大的 公司是苹果公司，它的市值（2016年） 超过6000亿美元，创造出来的 财富超出了当年通用汽车公司一个数量级，仅账 面上的现金就超过 1000多亿美元。但是， 苹果公司在全球只雇用了 8万名员工而 已。市 值和苹果类似的Google公司，雇 的人更少。今天，进入Google公司要比 被哈佛录取难得多，哈佛的录取率超过5 %，而Google的还不到千分之 二。也就 是说，受益于苹果或者Google这类公司 的人，远比20世纪50 年代普通汽车厂装 配工人的数量少很多。

那么大量淘汰下来的劳动力怎么 办?新毕业的学生如何就业?答案是要么 去从事一份工资足够低的服务性工作， 要么没有工作靠领取救济过 活。因此在 过去半个世纪里引领了信息革命大潮的 美国，国民的中位

数收入并没有提高。图 7.12是互联网时代美国有大学学历的在 职人员中位数工资变化的趋势图。上方 的红线是有5年以上工作经验的员工的工资变化情况，下方的蓝线是刚毕业入 职员工的工资变化。可以看出总体趋势 是不升反降，这验证了前面介绍的盖洛 普调查的结果。

在智能时代，一定会有一小部分人 参与智能机器的研发和制造，这是所谓 的新行业，但是这只会占到劳动力的很 小一部分。虽然很多乐观主义者认为，将

来一定会有新的行业适合人们工作，但 是这需要时间一半个世纪的时间。然 而智能革命并不打算给人类等待的时 间，它已经到来了,接下来大家不得不考 虑社会问题怎么解决。

(美元) 120 000

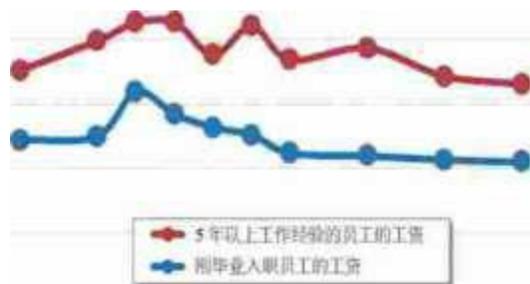
100 000

80 000

60000

40 000

20 000



1997 1999 2001 2003 2005 2007 2009 20 | | . (年).|

图7.12从互联网时代开始，美国有大学学历的中位数工资的变化

一种简单而粗暴的想法是对富人征 税。但历史证明这种劫富济贫的做法从

长远来看是阻碍经济发展的。很多经济学著作也详细分析了其中的原因，这里我们只是列举几个事实，帮助大家理解高税收的危害。首先，当税率达到100%时，一分钱的税也收不上来，因为不会有人去创造财富了。类似地，当税率过高时，实际上等于鼓励懒惰，当全社会都不去创造财富而只考虑再分配时，经济就开始衰退了。其实，只要理性地思考这个问题，而不是感情冲动地仇富，就不难理解这个道理。事实上，富人的钱财除了少部分用于个人消费并购买了一些不动产[14]外，剩下的钱并没有放在保险柜里，而是又投入了再生产。过高的税收意味着投入再生产的钱减少了。

当然，有人可能会说，宁可经济发展慢一点，也要保证社会公平。其实社会公

平只能反映在机会平等上，而不是结果的公平。实际上，只要个人的智力有差异，努力的程度不同，以及每个人的运气不同，即使劫富济贫，也无法保证社会公平。

2010年，美国爆发了所谓的占领华尔街运动，一大群无业游民、个别的低收入者和左派人士聚集到纽约街头，打着反对2%的人的旗号，表演了一场滑稽戏，并且持续了好几个月。之所以说它是滑稽戏，是因为这场运动不仅没有明确的目标，参与者不知道反对谁，反对什么，要求什么，也不知道自己的诉求是什么。而纽约的市民们照样工作生活，只当他们并不存在，因此这场闹剧最后自行收场了。从图7.13的照片里可以看出，这群人中没有一个营养不良的，因为他们实际上是在被他们所反对的2%的人养活着。2014年，美国收入在前1%的人贡献了美国45%的联邦税收，这要感谢奥巴马总统对富人的各种征税手段，在2013年这个比例是43%，2012年是40%。[15]可以说，如果没有占领华尔街的人所反对的这2%的人，美国早就成了三流国家，甚至混得比希腊还要惨。



图7.13占领华尔街的闹剧

占领华尔街运动没有得到主流民众 的同情，因为大家认为他们是不劳而获 的寄生虫。事实上，恰恰是那些他们所反 对的人为他们提供了福利，让他们能够 去搞运动。更具有讽刺意味的是，就在占 领华尔街期间，名义上美国代表中下层的左翼的民主党输掉了中期选举[16]。事实上，并没有什么主流的政客关心他 们，大部分政客只是忽悠他们以换取选票，这部分人的问题一直得不到解决。

不过，占领华尔街运动还是引起了 美国社会的反思。这些低收入或者无收 入的人出路在哪里？通过福利和救济将 他们养起来，显然是不够的，因为那些人 的人生前景依然是灰暗的。2016年，美 国总统候选人特朗普替这些人说出了他 们的希望—体面的工作。特朗普讲了一个巴尔的摩下层人的故事，那个人从 小到大生活在社会底层，在监狱进进出出很多次，有一次，实在活不下去了，又 想去抢_家药品杂货店。但是，经过一番 思想斗争后，他干脆跑到警察局把他可 怕的想法告诉警察。让他吃惊的是，那位 警官掏出了自己并不多的钱给了他，还 帮他租了一间房子住，这让他非常感动， 决定做一个对社会有用的人。但是接下 来，这位年轻人还是找不到工作,因此他 的困难虽然暂时得到了缓解，但是问题 依然没有解决。特朗普是想通过这个故事说明工作对现代人的重要性。这位年 轻人显然还有良知，也愿意通过工作养 活自己，如果有了工作，他可能完全可以 走向新生。但是在信息时代，适合这位年 轻人的工作越来越少了。到了智能革命 之后，任何简单动脑的工作可能都要消 失，甚至那些现在从事所谓高大上职业 的人，也会失去工作。

这一次由机器智能带来的革命，对社会的冲击将是全方位的，我们所依赖的那些所谓需要智力的工作也在消失。即使有新的行业出现，由于机器智能的影响，它们所需要的就业人数相比过去的老行业也会少很多。在智能革命全面到来的时候，不可能像过去那样，把农业人口变成城市人口，把第一、第二产业变成第三产业这么简单。

针对2010年的占领华尔街运动以及2015年年底以来法国、德国和比利时外来移民不断滋事的状况，大家在思考一个根本性的问题：这些不满情绪的根源在哪里？这不能简单地归结为贫富悬殊，或者宗教纷争。其根源在于，很多人被社会进步所抛弃了。随着技术革命的发展，并非每一个人的发展机会都是越来越多的，反而可能是越来越少。

是否能有良好的解决方法？坦率地讲，谁也没有。但是，即便没有好的解决方法，我们也要在观念上接受这样一个事实，即越来越多的事情人类将做不过机器。我们今后的决定，应该根据这个前提来做，只有面对现实，才能最终建设一个让所有积极向上的人都具有成就感和幸福感的社会。

虽然我们不知道如何在短期内创造出能消化几十亿劳动力的产业，但是我们很清楚如何让自己在智能革命中受益，而不是被抛弃。这个答案很简单，就是争当2%的人，而不是自豪地宣称自己是98%的人。

争当2%的人

在历次技术革命中，一个人、一家企业，甚至一个国家，可以选择的道路只有两条：要么进入前2%的行列，要么被淘汰。抱怨是没有用的。至于当下怎么才能成为这2%，其实很简单，就是踏上智能革命的浪潮。

每当我谈到机器智能对人类社会的冲击时，听众们总是要问：未来的时代是人的时代，还是机器的时代？我们是否会被机器控制？我的回答是：未来依然是人的时代，我们不会被机器控制，机器在完成任 务时甚至不知道自己在做什么。比如Google的AlphaGo，其实并不知道自己是在下棋。但是，制造智能机器的人就不同了，他们可能只占人口的不到2%甚

至更少，却在某种程度上控制着世界。

这个说法不是危言耸听，实际上今天已经发生了。大家不妨想想自己每天有多少时间挂在微信上，有多少商品是从淘宝或者京东购买的，有多少次出行是靠滴滴打车。这些公司没改变一点产品的形态，亿万用户的生活就被它们所左右了。更重要的是，这些公司完全掌握了 我们衣食住行的生活细节，它们可能比我们更了解我们自己。既然做到了对我们如此精确的把控，他们挣我们的钱便是不言而喻的事情。在销售商品的时 代，我们认为越便宜越合算；到了提供服务的时代，我们发现忽然有了很多免费的服务，我们为此欢呼，但是不久我们会发现，看似免费的东西才是最贵的，因为我们在获得这些服务的同时交出了自己的自由。而只有当我们在失去自由，利益受到损失时，才会体会到自由的可贵。

我并非要表达这些控制着我们、占不到人口2%的人要做坏事，事实上，到目前为止他们对我们的帮助比带来的危害要大得多。我想说的是他们的成功其实给予我们一个启示，那就是，如果我们不可避免地要被那2%的人通过大数据和机器智能控制，与其抱怨，不如干脆加入他们的行列。如果你现在已经在其中了，那么恭喜你，如果还不在，那么应该加入进去。讲到这里，我想大家可能会有一个疑问，那就是“我们怎样才能加入他们的行列”。我想说的不是每个人都要到上述公司去找工作，而是希望大家接受一个新的思维方式，利用好大数据和机器智能。回顾从工业革命开始的前三次

重大技术革命，首先受益的是和那些产业相关的人、善于利用新技术的人。虽然并非每一个人都能够去开发大数据和机器智能产品，但是应用这些技术远不像想象中的那么难。

我有一位在生意上还算成功的学员，在全国各地开了几百家茶叶店。这个行业有个特点，就是利润高，但是每天的交易量小，平均每家店每天只有几单生意。这位老板多少有点苦恼，因为如果要想把生意做得更大，就需要多建店面，但是店面太多他也管不过来。在我们讨论他如何转型时，我问他几个问题：

- 1.每家店每天都有多少人进门来转一转？又有多少人完成了茶叶购买？
- 2.这些客人是谁？他们什么时候来
到店里？什么时候更可能达成交易？
3. 如果有些客人是回头客，他们是谁？如果客人们买了一次不再回来，又是为什么？
4. 常客们每年消费掉多少茶叶？每个人经常消费的是哪种茶叶？价位在哪个档次？
5. 店面外每天的人流情况如何？

这些问题，除了每天有多少人达成交易他已经知道外，剩下的一无所知。如果这位老板能够在茶叶店门口装一个传感器，请人做一个手机APP,并且通过给予一些优惠券的形式鼓励到访的顾客安装，就能准确地了解上述信息，包括其中每一个细节。接下来，他就可以找人分析一下如何改进他的生意，如何做推广，等等。当然，更彻底的改变是利用所获得的大数据信息找到那些经常买茶叶的人，和他们建立起长期的供货关系，这样不仅能有比较稳定的收入，而且还能因为流通渠道成本的降低而提高利润率。其实美国的一些葡萄酒厂已经尝试这种做法好几年了，一些品质较好的葡萄酒庄已经不再依赖批发商和零售店这样的销售渠道，而主要是通过互联网向订户直销。

如果大家觉得茶叶店的生意太小，不具有代表性，我们不妨再看看现在冰箱公司在考虑什么事情。除了可以像GE那样通过消耗性材料挣

钱，一些冰箱公司开始考虑将冰箱看成商场里货架的扩展，通过摄像头和传感器，可以收集到顾客购买食物的习惯，以及顾客对食品消耗的程度，通过移动互联网提示用户补充食物。这种冰箱装有可以上网的触摸屏，顾客可以通过冰箱上的触摸屏直接从电子商务公司购买食品。这样，耐用电子产品又具有了商场货柜和电商入口的功能。虽然上述功能还没有完全实现，但是三星等公司已经在销售可以直接购物的智能冰箱的雏形了。

上述那些从事所谓传统行业的人，距离大数据和机器智能其实远比他们想象的要近得多。如果说有距离，可能心理上和观念上的距离比技术上和商业上的要远得多。

在每一个重大的技术革命开始的时候，真正勇敢地投身到技术革命大潮中的人毕竟是少数，受益者更少，大部分人则会犹豫和观望。在智能革命到来之际，每一个人也有两个选择，要么加入到这一次浪潮中，要么观望徘徊，最后被淘汰。当然，大多数人的观望、犹豫和徘徊，给了2%的人以机会，使得愿意吃螃蟹的人在奋斗的道路上少了很多竞争对手。正是因为知道自己不加入进来就会被淘汰，马斯克和盖茨一方面对机器智能的发展非常担心，另一方面却选择加入到机器智能的大潮中。



图7.14三星公司的智能冰箱，可以直接订购某些商品 小结

大数据导致机器革命的到来，这对未来社会的影响不仅仅存在于经济领域，而是全方位的。尽管总体上这些影响是正面的，从长远看会使

我们未来的社会会变得更好;不过,和以往的技术革命一样,智能革命也会带来很多负面的影响,特别是在它发展的初期,而这些影响可能会持续很久。

任何一次技术革命,最初受益的都是发展它、使用它的人,而远离它、拒绝接受它的人,在很长的时间里都将是迷茫的一代。在智能革命到来之际,作为人 和企业无疑应该拥抱它,让自己成为那2 %的受益者;而作为国家,则需要未雨绸缪,争取不要像过去那样每一次重大的技术革命都伴随半个多世纪的动荡。

我们还没有经历过机器在智能上全面超越人类的时代,我们需要在这样的环境里学会生存。这将是一个让我们振奋的时代,也是一个给我们带来空前挑战的时代。 注释

[1] 在美国,很多道路在交通高峰期要求车上必须坐有两个或两个以上的人才能使用快速车道,这些车道被称为拼车车道。

[2] 包括除了英美之外的几乎所有国家。

[3] 在商品经济比较发达的国家,法律对假货的处罚不是简单的一赔三或者一赔十这么简单,而是把赔偿的对象扩展到所有可能的受害者,通常在销售假货商家中,从销售类似产品一开始算起,把所有在那个商家购买过商品的顾客都算进去,因此我们经常看到因为产品质量而动辄赔偿上亿美元的新闻。对于大公司,这会大伤元气,对于小商家,一次假货的销售可能会导致其破产。对于其他欺诈行为,也可以通过集体诉讼的方式进行严厉的处罚。

[4] 折算成1990年的购买力。

[5] 这是按照购买力计算的,如果不考虑物价水平,1979年中国实际的人均GDP不到200美元。

[6] 摘自《共产党宣言》。

[7] 《嘉莉妹妹》《珍妮姑娘》和《美国的悲剧》等小说的作者。

[8] 美国于1916年通过了亚当森法案(Adamson Act),规定8小时工作制,一些企业比如福特公司也率先实行了8小时工作制,但是在全美国

全面实现 8小时工作制是到20世纪30年代的事情了。

[9]如果看市场份额，通用汽车 公司在全球的份额远没有20世纪五六十年代高。

[10] 包括变相下岗的所谓提前退 休的员工，以及一些老员工的遗孀。

[11] 美国三大汽车公司之一的克 莱斯勒实际上已经是欧洲菲亚特公司的子公司。

[12] 根据美国劳工部的统计，美 国农业工人早已经占不到劳动力人口的2%○

[13]哈佛大学和麻省理工学院所 在地。

[14]美国高净值家庭放在不动产 上的财富一般不超过5%。

[15] <http://www.cnbc.com/> 2015/04/13/top-1 -pay-nearly-half -of-federal-incometaxes.html.

[16]美国在4年总统任期之中偶 数的年份，需要重新选举全部的众议员、 三分之一左右的参议员和部分州的州 长，这个选举被称为中期选 举。

Table of Contents

[现象、数据、信息和知识](#)

[数据的作用：文明的基石](#)

[相关性：使用数据的钥匙](#)

[统计学：点石成金的魔棒](#)

[数学模型：数据驱动方法的基础](#)

[什么是机器智能](#)

[鸟飞派：人工智能1.0](#)

[另辟蹊径：统计+数据](#)

[数据创造奇迹：量变到质变](#)

[大数据的特征](#)

[变智能问题为数据问题](#)

[思维方式决定科学成就：从欧几里得、托勒密到牛顿](#)

[工业革命，机械思维的结果](#)

[大数据的本质](#)

[从因果关系到强相关关系](#)

[从大数据中找规律](#)

[巨大的商业利好：相关性、时效性和个性化的重要性](#)

[把控每一个细节](#)

[重新认识穷举法 完备](#)

[从历史经验看大数据的作用](#)

[技术改变商业模式](#)

[加 \(+\)大数据缔造新产业](#)

[技术的拐点](#)

[数据收集：看似简单的难题](#)

[数据存储的压力和数据表示的难题](#)

[并行计算和实时处理：并非增加机器那么简单](#)

[数据挖掘：机器智能的关键](#)

[数据安全的技术](#)

[保护隐私：靠大数据长期挣钱的必要条件](#)

[未来的农业](#)

[未来的体育](#)

[未来的制造业](#)

[未来的医疗](#)

[未来的律师业](#)

未来的记者和编辑
智能化社会
精细化社会
无隐私的社会
机器抢掉人的饭碗
争当2%的人

Table of Contents

[现象、数据、信息和知识](#)

[数据的作用：文明的基石](#)

[相关性：使用数据的钥匙](#)

[统计学：点石成金的魔棒](#)

[数学模型：数据驱动方法的基础](#)

[什么是机器智能](#)

[鸟飞派：人工智能1.0](#)

[另辟蹊径：统计+数据](#)

[数据创造奇迹：量变到质变](#)

[大数据的特征](#)

[变智能问题为数据问题](#)

[思维方式决定科学成就：从欧几里得、托勒密到牛顿](#)

[工业革命，机械思维的结果](#)

[大数据的本质](#)

[从因果关系到强相关关系](#)

[从大数据中找规律](#)

[巨大的商业利好：相关性、时效性和个性化的重要性](#)

[把控每一个细节](#)

[重新认识穷举法 完备](#)

[从历史经验看大数据的作用](#)

[技术改变商业模式](#)

[加 \(+\) 大数据缔造新产业](#)

[技术的拐点](#)

[数据收集：看似简单的难题](#)

[数据存储的压力和数据表示的难题](#)

[并行计算和实时处理：并非增加机器那么简单](#)

[数据挖掘：机器智能的关键](#)

[数据安全的技术](#)

[保护隐私：靠大数据长期挣钱的必要条件](#)

[未来的农业](#)

[未来的体育](#)

[未来的制造业](#)

[未来的医疗](#)

[未来的律师业](#)

未来的记者和编辑
智能化社会
精细化社会
无隐私的社会
机器抢掉人的饭碗
争当2%的人