吴军/著

大数据与智能革命重新定义丝

序—

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

吴军/著

大数据与智能革命重新定义絲

序—

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 在了一起。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 同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，S夸越 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 —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AlphaGozK 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的对手 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预 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 —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 搜索算法一一

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CPUa的内核 以及100多个GPUb)。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

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

数据一一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

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 心数据一一包括我们的信件、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。信息既可以是我们人 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一一3K背 景辐射M]、物理学定律中的参数、曰 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。

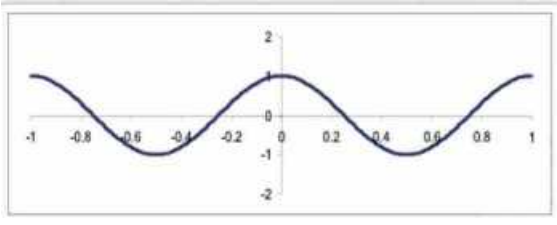


图1.2信息(上图）埋藏在噪声(下图) 中

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(2ir),

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

\/^/2倍，这个高度保证了两面墙的

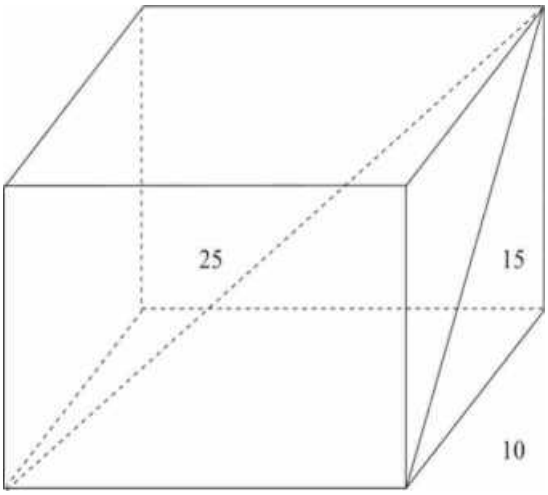
对角线长度是个整数——15埃及古尺， 因为根据勾股定理[3]:

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

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：$15^2 + 20^2 = 625$二 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 x4 +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。

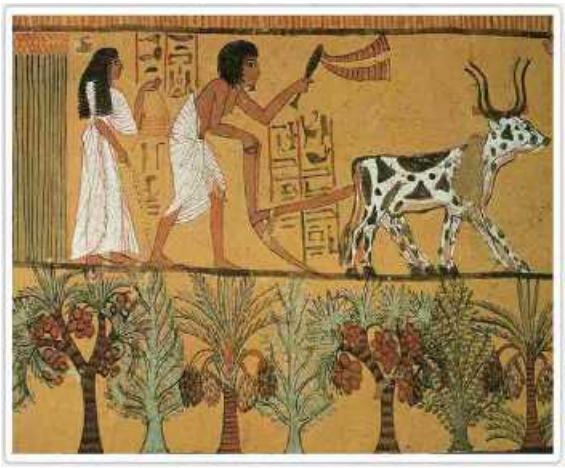


图1.4古埃及人为了农业的收成而发 展起天文学

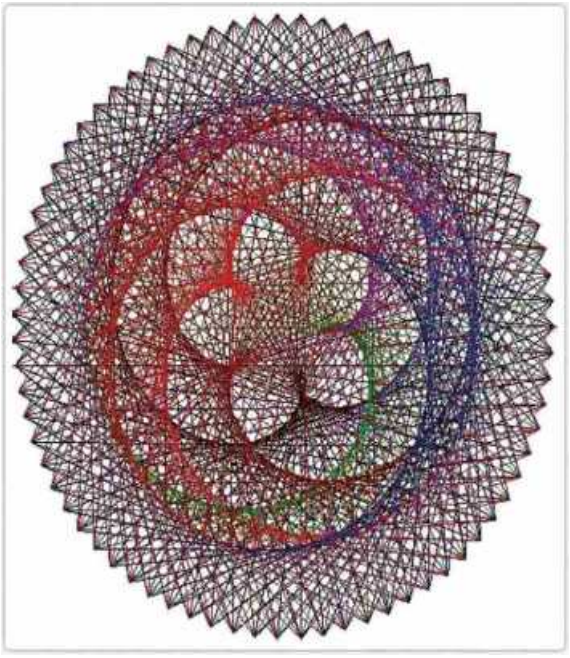


图1.5从地球上看到的金星运行轨迹

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28~2妖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_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：

获取数据	► ►	分析数据	► ►	建立模型	► ►	预测未知

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 (八_3此，前287年一前212年）贝腱 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。

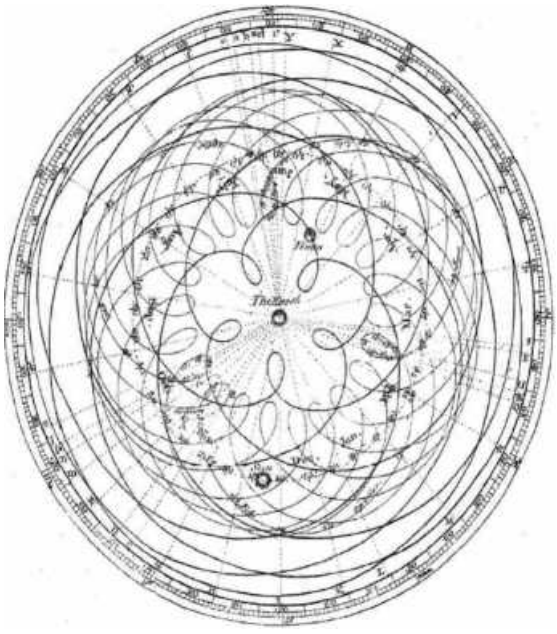


图1.7托勒密用多个圆相互嵌套的模 型来描述行星运动

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 的曰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。由于早期的曰心 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，楠圆的焦点 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_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。

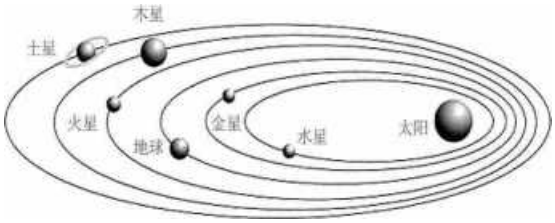


图1.9开普勒的太阳系模型

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_定是在中国极北的地区， 曰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。

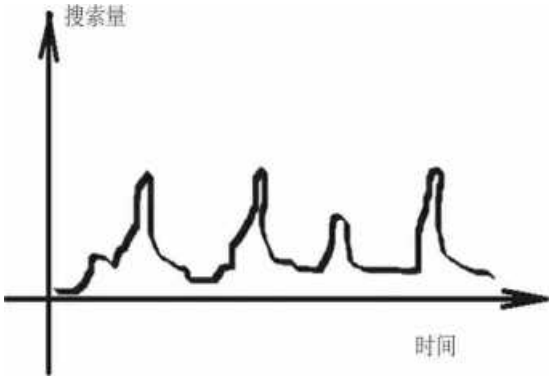


图1.11 Google用户在不同时间点对 某个电视节目的搜索量

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丨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。

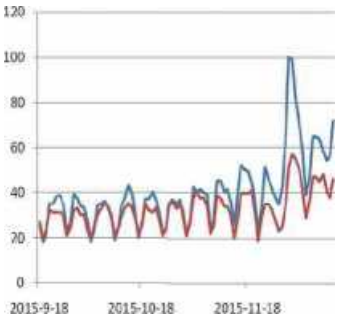
几个月后，我加入了Google，发现 帕特尔在Google确实不是很受人重视。 他加入Google很早，但是人们知道他仅 仅是因为他要求和当时新来的CEO (首 席执行官)施密特挤一间办公室，而不是 他所做的工作。好在Google总是支持每 个人干自己所喜欢的事情，因此帕特尔 就在Google内部一直研究搜索的模式。

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 毒一甲型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。



一气候变化 一全球变暖

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 ～ 4哟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 对偏高一一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，以便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 颇为准确的结论一一共和党候选人杜威 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 了质疑一一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，^(火）是 该变量的数学期望值，/7是实验次数（或 者是样本数），f是误差，a^2 是方差。 这个公式的含义是，当样本数足够多时, _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 们要想将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_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 —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干秋，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。而曰本缺乏这种水 平的技工，但是善于用机器加工，因此曰 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。 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干秋，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。而曰本缺乏这种水 平的技工，但是善于用机器加工，因此曰 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，这并非是因为我们对特定问题 的认识有了多大的提高，而是因为在很 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 件一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■倍撒）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，全世界开始了新的一轮技术革 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没有太大的差别一一它们都是能够 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。

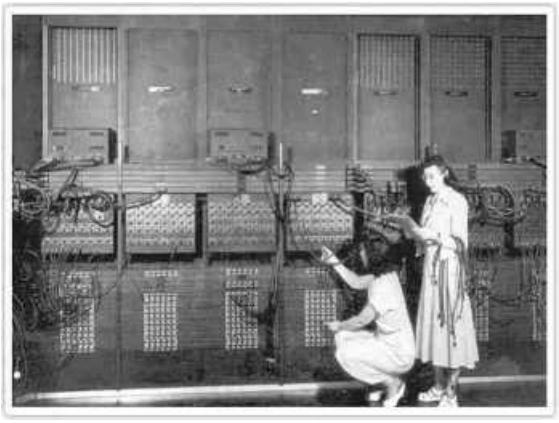


图2.2世界上第一台电子计算机

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 (AlanTuring，1912—1954)博士。 1950年，图灵在《思想》 ()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_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 如在下棋方面，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。

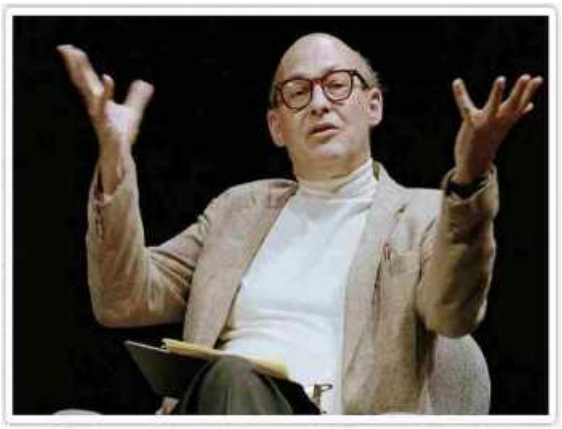


图2.4人工智能的奠基人明斯基

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，后来所有这些人都成了计算机科 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 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 (Ba「-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，旧M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，还不如旧M多，因此旧M快 被司法部进行反垄断调查了。这时，旧M 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)到旧M做学术休假[3]， 正好这时旧M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BM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_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，只有旧M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），好在旧M有大量的电传文本，这成 了BM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旧M 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！

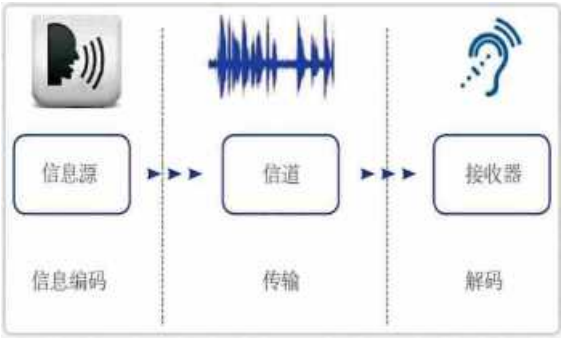


图2.7语音识别的通信模型

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 旧M。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。

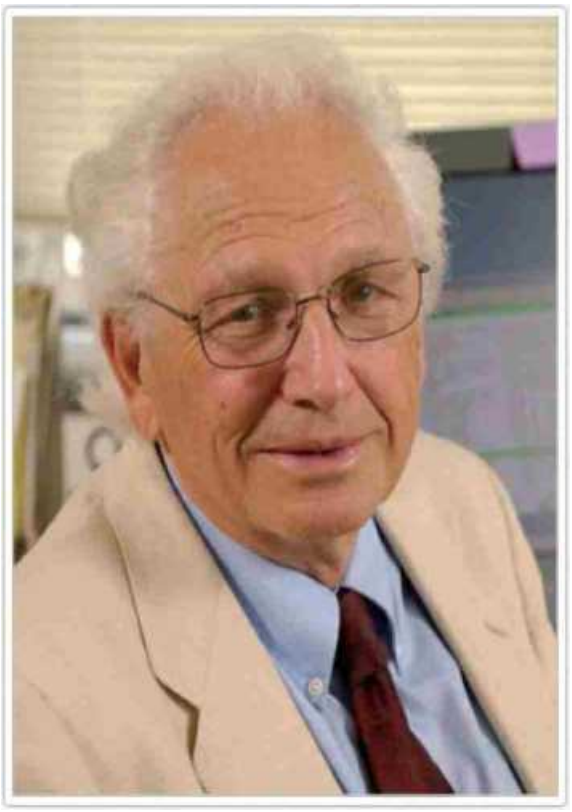


图2.8语音识别和机器学习的先驱贾 里尼克

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。

后来在旧M和Google先后担任过 主管研究的副总裁艾尔弗雷德_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 授们去旧M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讨论，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 _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“x x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“X X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 据说得好听点是具有多样性，说得不好 P斤是“相当杂乱” 0勺。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_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_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，大部分人都不想让自己表现 得“非常怪”，因此是不会在答卷上写下 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“你好一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。

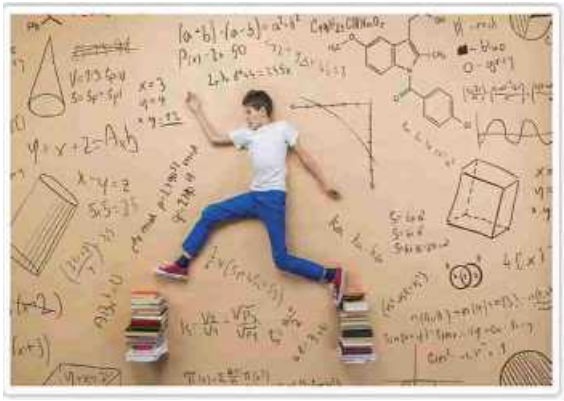


图2.10死记硬背的学习方式

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 是比较全面地覆盖中文、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_，是非常小 概率事件。

但是到了 2012年，情况发生了变化， —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Twitte「(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 到一起，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。

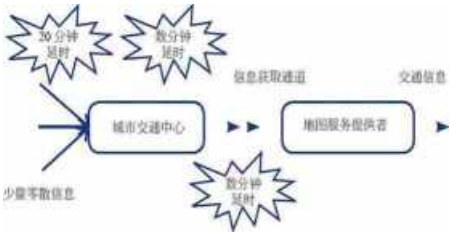


图2.12过去交通路况信息发布的流程

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		
地图和交通 信息提供者	► ► ►	地图服务提供者
、、 交通信息		7 / / / / / ✓ ✓

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,丨arge和vas倒底有什么差 别呢。丨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，全世界开始了新的一轮技 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，旧^/l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曰。

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 旧M承认，这步怪棋其实是源于程序的_ 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一卡罗-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 夫拿它没有办法一一它甚至像人一样会 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，以至于旧M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。

旧M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。旧M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 ^、又……心，计算

出相应的获胜概率

尸（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

Cfff artr of/r

^ 1、〇 2.....^ M

，再计算出每一个对应的新状态的胜率

P(S"\).P(S"'$_{2}$)……P (S”'$_{M}$)，然后挑一个产生最大胜率的走法，比如 是

尸(57"),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。

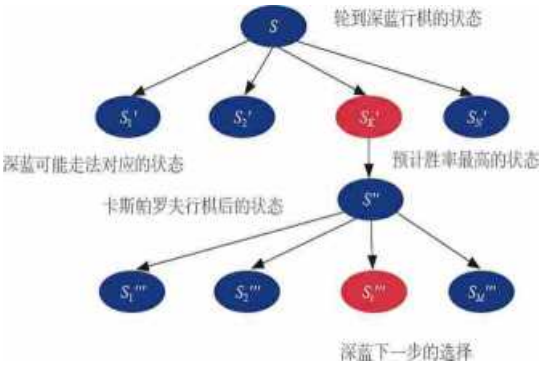


图2.14计算机下棋的博弈图

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，—卜36〇战胜了人类 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。一些人是

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 beon scattorod out and away from the lino of sight Why is the sky Blue?

math.ucf.edu/.../BlueSky/blue_sky.hlml University of California. Rhierside 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_，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 —，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）〇

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 们曰常做事行为的先前最重要的一种思 维方式一一机械思维。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_。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的天文观察数 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。

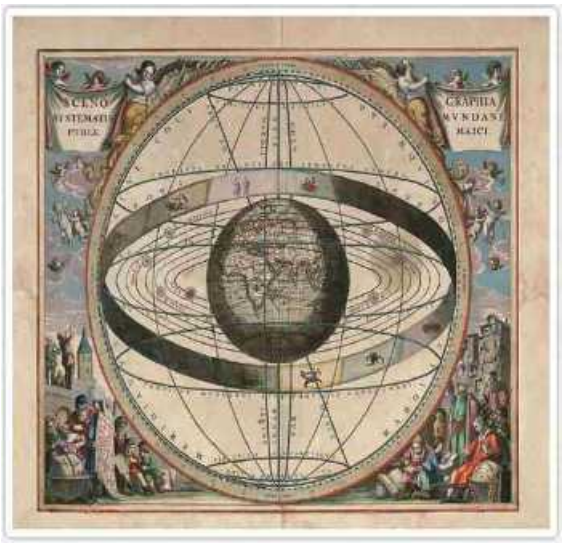


图3.2托勒密的地心说模型

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，比如像训练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_颗彗星围绕太阳运 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， 这在过去是无法想象的。这也同时让确定性这个词深深地印入了人类的思想 中。

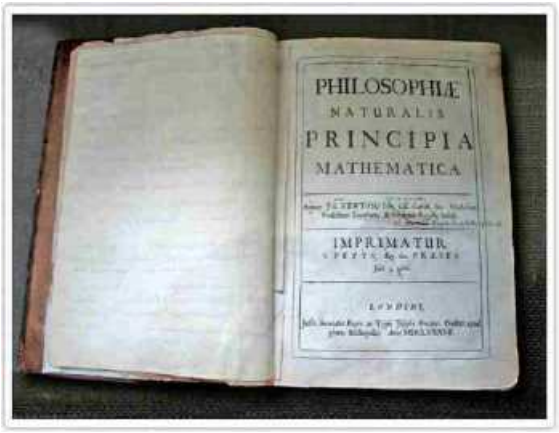


图3.3牛顿自己的那本第一版《原理》 一书，上面的笔记是在第十二版修订时 做的（现保存于剑桥大学三一学院）

牛顿作为思想家的贡献还在于他指 出了任何正确的理论从形式上讲都是简 单的，同时又有非常好的通用性，这与东 方哲学中的大道至简思想不谋而合。牛 顿在科学上的各种发明和发现，从物理 学的定律到数学微积分的定理，都可以 用非常简单的公式描述出来，而这些公 式又具有普遍意义。因此，从牛顿的时代 开始，科学家们都在致力于通过几个公 式来描述我们的世界，并且应用它们预 测未知。在牛顿之后，英国的焦耳也通过 一个简单的公式描述了能量守恒原理， 而他们的另一位同胞麦克斯韦则通过几 个简单的方程式描述了我们看不见摸不 着的电磁世界。这些科学原理简单的形 式，使得它们很容易地被应用到发明中。

从欧几里得到托勒密再到牛顿，在 思想方法上可以说是一脉相承而又不断 发展的。牛顿不仅把欧几里得通过逻辑 推理建立起一个科学体系的方法论从数 学扩展到自然科学领域，而且把托勒密 用机械运动模型描述天体的规律，扩展 到对世界任何规律的描述。后来人们将 牛顿的方法论概括为机械思维，其核心 思想可以概括成这样几句话：

第一，世界变化的规律是确定的，这 一点从托勒密到牛顿大家都认可。

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 是他们不知道该怎样改进。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，但其 实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一机械 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1821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（EliWhitney, 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，世界各地的人们的生活水平其 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岁之刚嗰，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 -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世界各地区人均寿命在当地开 始工业革命之后大幅提局

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。

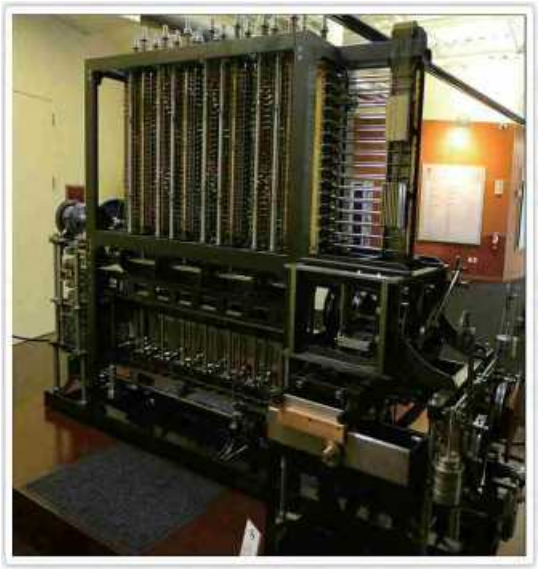


图3.7巴贝奇的差分机(硅谷计算机博 物馆的复制品）

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。

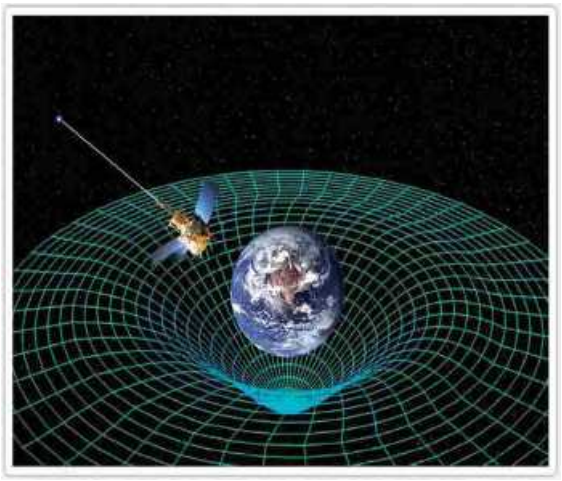


图3.8和牛顿万有引力原理不同的是，

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At ■ A p> s 熵的定义H二-2： i已log Pi

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 性一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，弗菜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，距离弗菜明首次发现青霉素已经过 去11年了，而弗菜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——种被称为青霉 烷的物质。青霉院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 系了起来，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。

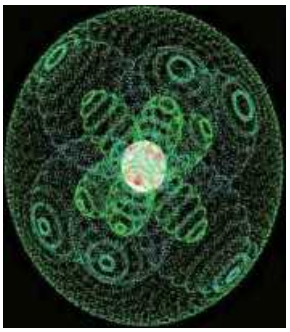


图3.10我们无法剛棚准电子的位置 和动量，只能计算出它们的分布,因此电 子就如同散布在原子核之外的云，也被 称为电子云

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。

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了。 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 是，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。如果我们有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_。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 _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－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，10000111 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，但是有的方法效率局，有的则效率 低，或者说用了很长的编码才能表示_ 个信息。香农第一定律洪的是，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，有些字用得多，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

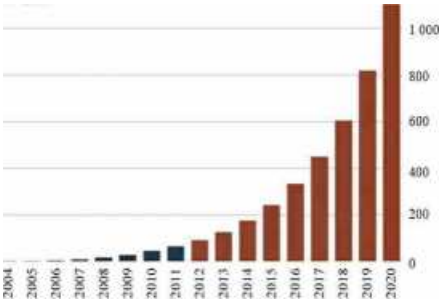


图3.12 2002-2020年全球互联网的 带宽呈指数级增长

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 弓丨入主观的假设。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 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，而是由库尔贝克等人提出 的，因此在英文里更多地被称为库尔贝 克-莱伯勒距离（Ku丨丨back-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 P尼?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_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——青霉烧——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，离赛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。这也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 是有足够多的数据支持。

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(Michae丨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 了很多疾病（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 -(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。

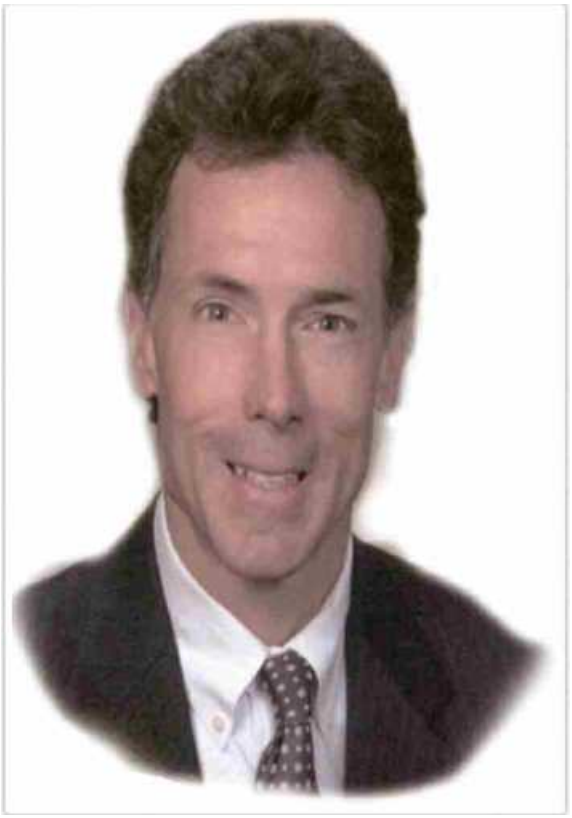


图3.13告倒烟草公司的密西西比州总

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_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上, 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 维方法一一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是一家局科 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，搜狗等小规模的搜 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 它不仅可以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1*3~2* 5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， -■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_点到另外任意一点可以

画直线。

2. —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 —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，而且还让他攒够了首付又 买了一栋房子。[1]

图4.2种植大麻的豪宅外景

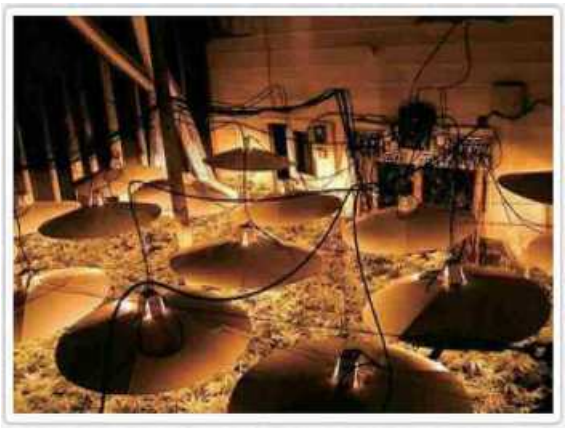


图4.3豪宅内实际上是这样的大麻种

类似的情况在美国各州和力嗥大不 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(Edisto日ectric)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 —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_ 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比 对，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一些营 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 了解我们自己的需求了。

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 商店一进门的货架，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， 过去基本上是凭经验来，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进门的 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 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 接但是需要非常大量数据的方法，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卜 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(或者遐掉部 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:丨. 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一一 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

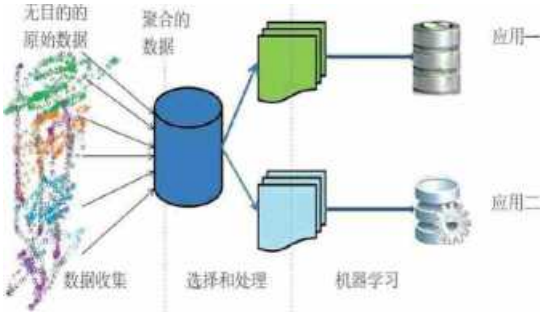


图4.7大数据收集、处理和建模的流程

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 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椁近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 决方案一一改造酒吧的酒架，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_ 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“万物联 网”（Internet of Things,简称loT)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_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 尽力地设计好每一款产品，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 和布置专卖店时非常尽心尽力，比如某 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 问题一一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)。艮「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 道问题出在设计和制作上，还是出在销 售上。

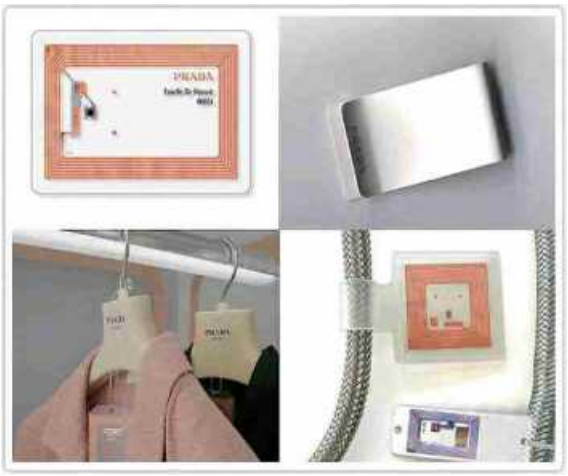


图4.10普拉达的衣服价牌里藏着一个 RFID芯片

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 _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_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；但是同时计算机会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 1丁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/ew/ 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，处理不确定性问题的能 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（Lada「）到高速摄像机，再到红外 传感器等，最根本的原因是Google米用 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

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, 计算机学习“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_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_。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。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产生、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 双方在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丨T行业做类 比就会发现一个有趣的现象：在纳斯达

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。

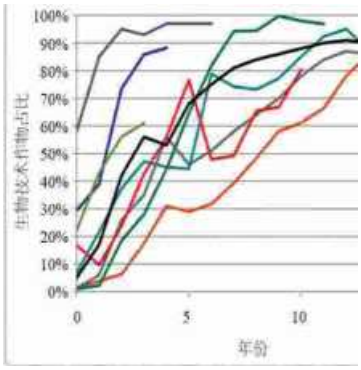
农业是另一个看似和计算机关系不 大的产业，在几千年的历史长河里，这个 产业变化非常缓慢。但是这种情况在过 去的30年里得到了根本性的改变，比如 农民们不再像过去那样自己育种，而是 从种子公司购买种子。而种子的培育，背 后用到了大量的信息技术作为支持。作 为全球最大的种子供应商之一的孟山都 公司也因此由一个化工企业变成了一家 生物公司,年收入143亿美元，利润却高 达82亿美元。相比之下，美国每年的农业 收入才不过1200亿美元。[20]如果我 们对比一下农业的这种模式和19世纪末 美国使用电力的模式就会发现，孟山都 等公司在某种程度上起到了通用电气和 西屋电气的作用，而大量的农民所扮演 的角色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



一娜Jit 一巴

—蜘莒蓿 一加$大汕菜 一美国棉花 —中闽棉花 一印度職 一 ■甜菜 一鲷大豆

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_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，S艮装等商品 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，_是产业 链从_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挎 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1\/1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 导致了商业链的出现;到了信息时代，商 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 —面。

在大数据时代，丨T软件和服务业依 然会是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。

加（+)大数据缔造新产业

2015年，“互联网+ ”是一个热门 词。不过，我觉得用“+互联网”这个词 更合适。类似地，对于大数据的应用，我 们也可以像过去“+蒸汽机”“+电气”那样，把它概括成“+大数据”。

我们在前面提到过的金风公司的故 事在2015年又有了新的进展。在和我进 行了多次关于大数据时代商业模式的探 讨后,该公司决定向BM学习，在商业模 式上做根本性的转变，主营业务从风力 发电机的制造，转变成发电设备的运营 和服务。当然，并非什么公司想做服务就 能做得好并净到钱，金风公司有底气转 型，源于其在宏观上对全球风能市场的 了解，在微观上对每一台风能发电机运 营细节的了解，加上通过大数据对发电 机可能出现的问题的分析，能够比一般 工程公司更有效地维护发电机。至于发 电机的生产，该公司只负责研制，然后将 设备制造交给其他公司去做。这样一来, 金风公司就在风力发电领域成功地复制 了旧M服务的模式。大多数亚洲制造企 业虽然在全球市场上占的份额不小，但 是通常竞争的手段就是压低利润降价，最后把整个行业变得都没有利润。金风 公司转型的做法，或许能给这些企业_ 些启发，当然如果没有大数据这样的机 遇，这种转型是非常困难的。

与金风公司面临类似情况的还有诸 多的电器生产厂商。这些电器无论是高 端的还是低端的，厂家只能赚到一次钱， 而且由于亚洲制造业同行相互压价，利 润也不可能很高。为了解决利润的问题， 一些对新技术敏感的公司想到了利用大 数据和移动互联网来改变商业模式。

GE公司是美国电器行业的龙头老 大，在过去它的冰箱和其他大电器的利 润一直不错，但是自从亚洲制造的相关 产品开始冲击美国市场后，GE家电部门 的利润率开始下降。在2008年金融危机 之前，它靠给购买家电的顾客贷款维持 利润，每年平均12.99%的利息实际上让 GE把一次性买卖变成了细水长流的生 意。但是在2008 ~ 2009年的金融危机 中，很多人还不上借款，导致GE家电部 门严重号损。提供贷款这条路也走不通 了，GE开始想别的办法来维持家电部门 的利润，它们想到了移动互联网和大数 据。

图4.16 GE的智能冰箱

GE将Wi-Fi安装到它的冰箱和其 他大型家电上，用来提示用户更换冰箱 取水器的滤芯等消耗性材料。这些滤芯 通常需要每半年更换一次，但是大部分 用户都难得更换，即使冰箱上的指示灯 亮了。GE将冰箱通过Wi-Fi连到互联网 上之后，可以通过手机APP (应用程序） 来提醒用户及时更换滤芯，这样一来用 户更换滤芯的比例提高了很多。值得一 提的是，用户订购滤芯只需要在手机 APP上点击确认即可,GE可以用快递将 滤芯直接邮寄给顾客，这样就省去了很 多中间环节。对GE来讲，两个滤芯（可 以使用_年，大约100美元左右）的利润 就抵得上一台冰箱本身的利润。

图4.17 GE智能冰箱能及时提示更换 接水器的滤芯

当然，GE通过Wi-Fi获得的信息远 不止滤芯的寿命，它可以全面了解用户 使用电器的情况，并且可以从千百万用 户那里收集到关于用户的大数据。通过 分析这些数据，GE可以牢牢地把握住这 些用户，知道他们接下来需要什么，有的 放矢地推销后续产品。有人说这是利用 移动互联网，而非大数据。诚然，在这个 过程中，移动互联网是必需的，否则GE 等公司无法收集数据，但是光有移动互 联网，没有大数据分析，GE无法了解用 户的具体情况，如果强行推广营销，会适 得其反。在有了大数据之后，如果制造业 厂商能够把思维方式变成“ +大数据”， 那么其产业就将得到全面的升级。当那 些厂商能够把控每一个用户、每一个产 品和每一次交易细节，它们就能绕过很 多经销的中间环节,直接和顾客做生意。 善用大数据之后，家电的销售可以不再 是一锤子买卖。如果我们对比拥有大数 据思维前后冰箱作用的变化，就会发现 这个原本只是家庭贮藏柜的大件电器，一下子成了连接顾客和商家的渠道。

GE的做法实际上是今天很多传统 的电器公司都可以采用的，但事实上大 多数公司并没有这么做，因为大部分企 业还没有形成大数据的思维方式。2013 年中国工业界发生了一件在媒体上被热 议的事情，即所谓的“雷军和董明珠”之 争。这一年的12月12日，中国经济年度 人物奖获得者小米手机公司的创始人雷 军先生和格力电器公司的CEO董明珠 女士在全国电视观众面前打了_个10亿 人民币的赌——前者表示当时年收入不 足百亿元的小米公司能够在5年内超过 当时年收入已经过千亿的格力电器公司。外人不论对他们的观点是赞同还是 反对，对他们这种张狂的豪赌行为其实 是贬多于褒，而且很多人是抱着看热闹 的心态静静地等待结果的产生。从表面 上看，这是两家企业负责人之间相互赌 气，前者对自己公司早期的成功信心满 满，对家电行业的老前辈颇有不敬，后者 对前者注重表面文章的做法看不上眼， 对自己所拥有的核心技术和长期以来形 成的市场经营更有信心。但是，在这场豪 赌的背后，其实突显出的是两种不同的 办企业思维之间的冲突。

小米是一家手机制造公司，其主要 收入来源就是它的手机销售，比较单一。 单从这_点上看，它与中国的两个主要 竞争对手华为和联想没有什么区别，它 甚至还不如华为，因为华为自己能生产 手机处理器，而小米主要的元器件完全 要从高通和东芝等厂家购买。至于智能 手机的核心——操作系统，小米用的是 Google的安卓（Android),尽管它修 改了一些UI的功能和接口。在很多人看 来，这样的企业就是一个没有核心技术 的亚洲制造企业，未来免不了陷入以打 价格战为主的低层次竞争中，事实上从 2013年开始，小米为了增加市场份额, 已经开始用极低的价格推广它的低端手 机了。这似乎完全没有摆脱过去亚洲制 造企业一贯的做法。因此，多年来致力于 发展自主知识产权、打造基于技术的核 心竞争力的格力电器，看不起移动互联 网暴发户小米是在情理之中的事情。

但是，就是这样一家产品单一、仍在 亏损的“电器”企业,2015年7月再融资 时，却被国际上知名的风险投资公司估 值为450亿美元。而与此同时，手机出货 量和小米相当、个人电脑全球占有率第 一、连续多年赢利的联想公司，市值只有 100亿美元左右。为什么会出现这样奇 怪的现象，只有两种可能性，要么DST等 风险投资公司的决策者是一群傻子，严 重高估了小米的价值，要么从长远来讲 小米具有联想等公司所不具有的价值。 作为成功投资了 Facebook和阿里巴巴 等公司的DST,投资人显然不是傻子，即 便高估了小米的价值，也不应该太离谱。 那么小米一定有联想等公司所不具备的 价值，这个价值就体现在大数据上。

小米从一开始就以一家互联网公司 的方式来经营它的手机业务。从本质上 讲手机只是小米获得用户的手段，在获 得用户后，它需要通过其他方式挣钱，这 一点小米和华为、联想都不同，后两者在 卖掉手机后就完成了交易。事实上，最早 利用智能手机特点开发移动社区的公司 不是腾讯，而是小米，只不过小米因为用 户的数量远不如腾讯多，它的米聊才最 终败给了腾讯的微信。在拥有一定数量 的用户后，小米也拿到了大量的用户数 据，但是怎么能不断有效地从每个用户 身上挣到钱，这是小米必须解决的问题。 目前，它通过手机成功地推销出不少配 件，包括一些可穿戴式设备，但是这些都 还不足以让它挣到足够的利润。为了进 一步绑定用户，小米还开发了其他的产 品线，比如电视、空气净化器等。因此， 从某种程度上讲，小米更像是一个以家 电为主的垂直电商，而不是家电生产厂 商。与传统电商所不同的是，小米从一开 始就注重对用户行为的分析和数据的作 用，因此它有可能在一些垂直领域做得 比传统电商更有效。至于它能否做到这 些，就要看它的技术水平和执行力了。正是基于其互联网公司的定位，风险投资 公司才给小米那么高的估值。而对于联 想这样的公司，投资人把它定位成制造 型企业，因此估值不高。至于为什么联想 这些制造型企业无法转型成为互联网公 司，那是由它们固有的基因决定的。[22

在争论小米和格力哪一家更有前途 这个话题时，董明珠问了雷军一个问题: 如果没有生产工厂,小米还能有销售吗？ 显然，董明珠按照思维定式把小米当作 制造型企业来看待。作为制造型企业，有 关它的产品的核心技术、自主知识产权 当然很重要，所以董明珠才会认为格力 经过20多年的积累，有深厚的技术沉淀， 是不可能被小米超越的。但是，正如我们 前面分析的，小米根本就没有把自己定 位为制造型企业，它卖手机并非满足于 挣硬件的利润，而在于获得用户,然后再 从每一个用户身上获得

长期的收益。因 此，雷军和董明珠之争，其实体现了大数 据时代和摩尔时代不同的思维方式的冲 突。至于小米是否能在5年内（即2013～ 2018年）超过格力，我倒认为雷军话说 得太满了，从长远看，如果小米不出现重 大失误，它一定能够超过格力，但是这个 时间点恐伯不是2018年。

说回到格力电器，它其实是传统的 家电企业的典型代表。这类企业在过去 20多年里一直在努力地发明和创新，但 是在外人看来却缺乏创造力。它们一方 面是全球的专利大户，比如索尼、东芝和 三星一直是获得美国专利前10名的大公 司，但是另_方面它们在世界经济中的 ±也位却在不断下降。在互联网大潮中，很 多这类企业已经变相落伍了，甚至开始 苦苦挣扎，比如索尼公司。当大数据时代 到来时，它们应该非常有所作为，因为它 们已经占据了家庭的客厅和卧室，但是 如果它们自己的思维方式还局限在摩尔 时代做硬件、卖产品的定式上，那么它们 将失去一次绝好的转型机会，其中很多 不免会被淘汰。就以格力电器的核心产 品空调为例，每一台能用10年左右甚至 更长的时间，而且它的购买和安装也不 像买个手机那么容易(虽然价格差不多)， 因此很少有人经常换空调。在工业化国 家，除非某年遇到了特别极端的天气，否 则空调的销售增长非常有限；在中国由 于城市化的进程还没有完成，大家只是 暂时感觉不到市场快要饱和而已。其他 家电，比如电冰箱、洗衣机，都是如此。 家电行业不仅增长缓慢，而且它在世界 各国的利润其实都非常薄，因此家电企 业的投资回报率都不高。

怎样才能让家电行业获得稳定的利 润呢？这在大数据时代之前是很难做到 的。虽然商学院很早就在教授吉列公司 送刀架卖刀片的商业模式，但是这种做 法过去在家电领域很难模仿，因为家电 的交易完成之后，用户和商家就没有关 系了，商家不知道向谁提供服务。即便知 道，后续的增值服务(如果能够进行的话) 也不是换一个刀片那么简单，用户的需 求常常千变万化，如果不了解用户的具 体情况强行提供所谓的服务，会让他们 感到反感。更重要的是，过去生产厂商和 经销商通常不是同一个，经销商刻意要 切断厂商和用户的联系，以便他们有可 能做后续的增值服务。比如在美国销售 办公用品的连锁店Staple (史泰博）或 Office Depot (欧迪办公），会预备好复 印机、激光打印机的各种OEM (代工生 产)的耗材，提供给购买办公电器的顾客， 销售电器的百思买和Fry's做法也类似。 因此，大部分家电企业只能从家电本身 赚钱，它们根本不知道自己的产品卖给 了哪一个具体的消费者，更不要说了解 消费者更多的个人情况和生活习惯了。

但是在大数据时代，家电厂商可以 通过_些产品跟踪技术（我们后面会讲 到）知道自己出产的每一个电器是如何 一步步进入顾客家的，并且知道用户是 淮。而每个大件电器本身又是收集用户 数据的采集器，因此家电公司可以完全 了解用户的很多生活细节，比如他们在 哪里，每天使用该电器的情况，使用其他

电器的情况，甚至什么时候回家，什么时 候吃饭等。从宏观的角度看，商家可以了 解到它的商品是通过什么渠道卖给了具 体的消费者，从而优化它的销售网络;从 微观的角度看，它可以了解每一位顾客 的生活，知道接下来每一个人需要什么。 这样，生产厂商其实就不再受经销商控 制了，厂商和用户的直接联系就建立起 来了。这样不仅厂商能获得更多的利润， 顾客因为消除信息的不对称性也能在价 格上获得优惠，而且厂商和顾客之间能 够建立一种细水长流的商业关系。这时,厂商之间的核心竞争力不再是商品本 身，而是更重要的服务。未来产品的服务 水平不完全取决于厂商对它的重视程度 (比如服务态度）和相关技术，而更多要 依靠智能化。未来,商家将在数据层面和 智能化方面展开竞争。

当然，像格力这样的传统企业必须 做出一个选择——是否愿意利用大数据 转型。从蒸汽机时代、电气时代到半个多 世纪前开始的信息时代，它们一直验证 着这样一个规律，即原有的产业加上新 技术就成为新产业，否则将被淘汰。在今 天的大数据和机器智能时代，这条规律 依然成立。

对于选择踏上新时代浪潮的公司， 是否都要成立大数据部门，是否都要转 型成为IT公司，这类问题没有一个简单 的“是”或者“非”的答案，不同的企业 会有不同的选择。但是，有两点是共同的: 首先，它们在人员构成上一定会有大数 据的专家加入;其次，大部分企业并不需 要自己成为大数据和机器智能开发的公 司。

在每一次技术革命的大潮中，并不 需要所有的公司都从事新技术本身的开 发和产品研制。当然，这些为全社会提供 技术的公司，比如GE、英特尔公司等, 站在了浪潮之巅，成为相应时代伟大的 公司。在今天大数据和机器智能的时代， 虽然每一个公司都会得益于数据的使用 以及机器智能带来的好处，但这并不意 味着每家公司都要聘请数据科学家或者 机器智能方面的专家。更切合实际的是， 他们付费使用第三方的服务。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。

小结

从工业革命开始，几次主要的技术 革命都遵循相似的规律。首先，是大部分 现有产业加上新技术等于新产业。或者 说原有产业需要以新的形态出现。其次， 并非每一家公司都要从事新技术产品本 身的制造，更多时候它们是利用新技术 改造原有产业。这次以大数据为核心的 智能革命也不例外，我们将看到它依然 会延续这两个特点。每次技术革命都会 诞生新的思维方式和商业模式，企业只 有在思维上跟上新的时代，才能在未来 的商业中立于不败之地。 注释

[1] http://www.seattletimes . com/seattle-news/big -time-pot-growers-use-seattlearea-homes/.

[2] http://www.dispatch.com *content* stories/local/20111021281 police - suspectinghome - pot -growing - get - power - use - data • html.

[3] https://www.sba.gov/sites *default* files/FAQ_Sept_2012.pdf.

[4]

IRSReleasesNewTaxGapEstimates ,2008, www.irs.gov .

[5] Jeff Butler. Discusses the IRS Research Division ' s Big Data Techniques , Meritalk, 2016.

[6] http://www.governing.com *columns* tech-talk/gov-states-big

-data-tax-fraud.html 这家网站给出 了一些利用大数据查处偷漏税的案例。

[7] 第一大连锁百货店是沃尔玛。

[8] 美国一些商店提供将发票发 到顾客邮箱中的服务，一些顾客为了和 信用卡对账方便，愿意提供邮箱或者手 机号。

[9] http://www.nytimes.com/ 2012/02/19/magazine/shopping -habits.html?pagewanted = 1 &r=1& hp.

[10]在美国，很多大公司的IT业 务是外包给专门的IT服务公司的。

[11] http://gadgets.ndtv.com *internet* news / netflix - now -accou nts-for-34-percentof-us-internet - traffic - at - peak - times -524323.

[12]关于 RFID (Radio Frequency Identification,射频识别， 简称RFID)的原理，我们在第五章中介

绍。

[13] John McCormick, Prada : The Science of Desire , http://www.baselinemag . com *c* a *Projects* - *Customer* -*Service* Prada-The-Science-Of-Desire，2002年12月 16日.

[14]在拉丁语系的语言中，比如 西班牙语中，不同性别使用的名词、定冠 词甚至动词都是不同的，这种情况比在 英语中的还要复杂得多，在中文里用词 基本没有性别的区分。

[15] 1 英里=1 .609344 公 里。——编者注

[16] 有一次交通意外是自动驾驶 汽车被其他车辆撞了。

[17] 关于瓷器的诞生时间，专家 们说法不一，从汉末三国到后唐五代的 说法都有。不过瓷器真正成为中国重要 的产业是从北宋时期开始的。关于瓷器 历史更多的内容，读者朋友可以参阅拙 著《文明之光》。

[18]冶金业虽然是人类最古老的 行业之一，但是在没有电之前，人类只能 生产很少几种金属（金、银、铜、铁、锡 和铅等）和合金（青铜），而且一般都很 难做到精纯。法国皇帝拿破仑三世是一 个喜欢奢华的人，他常常大摆宴席。宴会 上，客人的餐具是用银制成的，而他自己 却用铝制品，因为当时冶炼铝十分困难， 招的价格比黄金局昂得多。有了电之后， 人们发明了电解铝的制造方法，铝的价 格就跌到了我们今天说的白菜价，也正 因为如此，铝才能够被广泛地应用于各 行各业。即使是人类最早使用的金属铜， 在过去的几千年里，人类使用的都是粗 铜，如果用来做导线，不仅电阻比较大， 而且容易折断。而真正的精铜，也需要靠 电解才能获得。至于其他各种金属和合 金的制造，则更离不开电了。有了这些合 金，才有了后来的航天和航空工业。

[19]今天投资者依然可以要求上 市公司提供纸质的股票，但是没有人这 么做。

[20] http://www . bloomberg. com/news/articles/2013-08-27/u-s -farm-income-for-2013-seen-at -record-120-6-billion.

[21] AMD:超微半导体公 司。——编者注

[22] 关于这一点，有兴趣的读者 可以参考拙著《浪潮之巅》。

第五章 大数据和智能革命的技术挑战

大数据的数据量大、维度多、数据完 备等特点，使得它从收集开始，到存 储和处理，再到应用，都与过去的数 据方法有很大的不同。因此，使用好 大数据也需要在技术和工程上采用 与过去不同的方法。

每一次技术革命除了有生产力发展 需要，还要有很多技术准备，只有当所有 这些必要的技术都成熟时，技术革命才 变为可能。历史上虽然不乏“穿越时空” 的人，比如达■芬奇和尼古拉•特斯拉， 他们能设计出很多后世才用得到的东 西，但是由于市场没有准备好，配套的技 术不成熟，他们的想法在当时只能算是

空想而已。

以大数据为核心的智能革命也是如 此，它之所以在今天这个时间点爆发，除 了在商业上有了应用的可能性之外，也 是因为很多相关技术已经成熟。在未来 它要想进一步发展和普及，还需要解决 很多技术上的瓶颈。在这一章，我们首先 分析产生大数据的技术基础，然后再探 讨它所面临的技术挑战。

技术的拐点

科学技术的发展并非是匀速的。重 大的科技突破常常需要酝酿很长的时 间，在这段时间里，我们发现技术进步是 一个缓慢的量的积累，有人把它称为相 对停顿的状态，因为这个阶段一切发展 都是平衡的。但是当这些量的积累到一 定程度后，科技在短时间内获得单点突 破，然后新科技全面迸发，这便是拐点。 在历史上有很多关键性的拐点，比如 1666年，牛顿发明了微积分，发现了力 学三定律和万有引力定律，完成了光学 分析，从此世界进入科学近代社会，因此 这一年这被看成是科学史上的一个拐 点。到了 1905年，爱因斯坦完成了分子 说，发现了光电效应，提出了狭义相对论， 从此开启科学的现代社会，随后物理学 的各个领域全面繁荣。1965年，摩尔博 士提出了摩尔定律，同时在工业界大规 模集成电路出现，从此开始了待续半个 世纪的信息产业高速发展。在这些拐点 上，原有的平衡被迅速打破，人类从此进 入一个新的时代。

机器智能进步速度

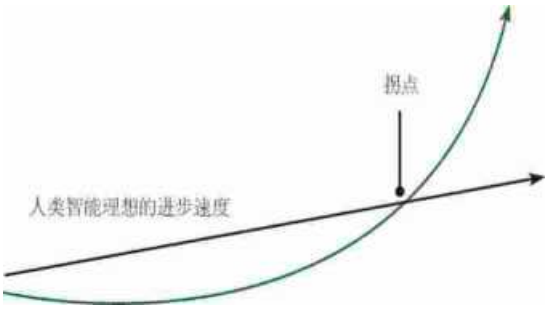


图5.1我们今天可能正处在机器智能 将要超过人类的拐点上

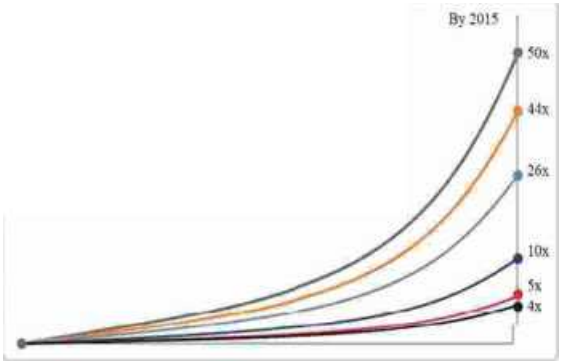
机器智能的概念已经被提出来60多 年了，但是真正的突破却在具有了大数 据的今天。大数据本身真正引起科技行 业的注意，也仅仅是10年前的事情，然而 在短短的几年里，它就井喷式地爆发了， 并且让机器智能水平有了本质的提高。 因此智能技术的拐点可能就发生在从10 年前开始到接下来的一二十年这一段时 间。再过一两个世纪，回顾我们今天所处 的这个时代，后人会感叹这是人类文明 史上的一个大时代，就如同我们今天谈 论大航海时代和工业革命那样。

为什么大数据的拐点会发生在今 天?从过去的10年开始，最容易看到的特 征就是全球数据量呈爆炸式增长。图5.2 是2011年思科公司根据它自己、 Gartne「（高德纳咨询公司）和IDC (国

P示数据公司）的数据对全球互联网、硬件 发展和企业数据量增长的估计。图中的 起点是2009年,终点是2015年，因此从 2009-2011年是对历史数据的总结，再 往后则是预测。从图中可以看出，企业级 数据的增长幅度比6年前涨了 50倍。相 比之下，计算机硬件(存储设备和服务器) 和互联网本身(传统互联网、移动互联网 和视频)的增长虽然也是指数级的，却都 显得缓慢很多。

在全球企业的商业每年不到10%的 增长率情况下，数据量却能够这么快地 增长，这看上去是一个违反常理的现象。 其中的原因概括来讲就是所有和数据相 关的准备条件在这个时间点开始成熟。 我们可以从数据的产生、存储、传输和处 理四个角度来分析一下大数据形成的技

术条件



5〇s数据摄以企ik数据中心的垧长为代表 44x存fgfiff预ii将从2009年的0.SZB W长到2020年的35ZB 数据通的垧丨沃自子移动设备数世的增长 K)x 1务器教!81[1界_内的坳板 5x褚于IP的视柳(8丨自$及实时应用数fit的拊fe 4x IPilfite他长

图5.2数据量的增长在所有的增长中 是最快的

数据来源：Cisco VNL，2011.6; Gartner, 2009 & 2001

数据的产生

大数据的第一个来源是电脑本身。 全球数字化让几乎每_个使用电的设备

都有了一个“电脑”，这些电脑或者设备 中内置的处理器、传感器和控制器一直 在产生数据，比如记录设备状态的曰志 (Log) 〇在过去，很多数据并不会被记 录下来，比如电话交换机除了记录少量 的设备运行状态之外，并不记录来往通 话的控制信息，

包括打电话的时间、双方 的电话号码、通话时长等，但是当人们发 现这些数据有价值之后，由计算机控制 的程控交换机很容易把这些细节都记录 下来，这就产生了很多和电信相关的数 据。

另外，由于企业级的IT系统和软件 越来越复杂，它们的设计者不得不记录 更多的细节，以便在发生异常时能够跟 踪找到问题所在。在Google,工程师们 在编写程序时，每隔几行代码就要插入 一句记录状态的日志语句，以便今后查 找错误、完善程序和进行数据分析。

大数据的第二个来源是传感器。传 感器技术的进步使得收集数据变得非常 容易。我们在前一章中提到无源的射频 视频芯片（RFID)就是一种帮助收集数 据的工具。今天无所不在的摄像头，其作 用与收集数据的传感器也有着相似之 处。

我们先看看射频视频芯片是怎样工 作的。这种芯片里面可以存储一些信息, 芯片外有一个回形的天线（线圈），用于 接收阅读器发出的无线电波。当天线线 圈接收到无线电信号后，根据电磁感应 原理，它会产生微小的电流让芯片工作, 并将里面的信息发出，再由阅读器读取。 这种射频视频芯片非常便宜，零售价也 不过4美分一片。将它装到各种物品上, 就可以自动识别各种物品，进而跟踪物 品。它的体积可以做得非常小，甚至可以 被植入生物体内，用以跟踪它们的活动。

RFID的用途非常广泛，将它贴到商 品上，当该商品通过一个RFID阅读器时， 阅读器就知道该商品经过。那么在未来 的超市中，其实不需要在收银台用人工 扫描每一件购买的商品，记账付款，而只 要将装满货物的推车推出安装了 RFID 阅读器的通道，所购买的商品就会被_ 一计价结算，然后再通过移动互联网将 购货金额发送到购买者的手机上，经过 购买者确认后，直接手机付款即可。这样 整个商场只需要几个保安确认购买者守 秩序即可。

图5.3 RFID芯片

除了用于零售业结算，RFID还可以 用于商品的防伪和跟踪货物的移动等很 多方面。由于有了RFID，物品从生产到 消费，整个流程都可以被跟踪，这样就产 生了大量的数据。

类似于RFID这样的传感器很多，比 如可穿戴式设备中，一个核心的传感器 是感知加速度的芯片，它根据加速度的 积分算出速度，这样就可以追踪人的身 体的各种活动了。另外，在万物联网中, 需要大量使用各种传感器，它们在不断 地提供各种各样的数据。

RF1D

阅读器

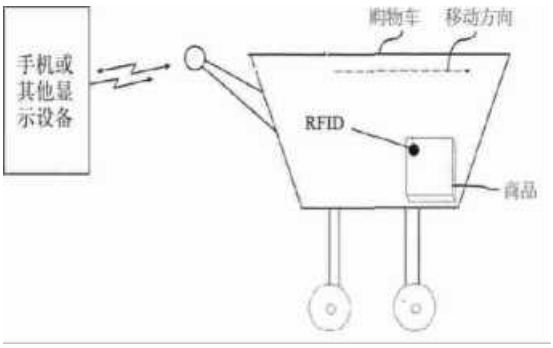


图5.4采用RFID自动计价付款系统的

图5.5万物联网离不开传感器

数据的第三个来源是将那些过去已 经存在的、以非数字化形式存储的信息 数字化，这个过程开始于2000年左右。 非数字化的数据包括语音、图片、设计图 纸、视频、档案、古稀图书和医学影像等, 这些信息过去都是以各种各样的形式存 储的，由于积累的时间很长，因此数量巨 大。据约翰■霍普金斯大学生物工程系 系主任麦克维（曰liot McVeigh)教授介 绍，在2010年时，全美国病例档案的文 件规模比互联网上(非重复)的网页数量 高出一个数量级，当然，在过去的几年里 互联网上的内容增加很快，很难说今天 病例的数据量是否依然超过互联网，但 至少说明它的规模很大。

产生上述数据的主要是企业，而非 个人。在互联网时代，网络用户产生的数 据（UGC)以更快的速度在增长。对于 用户产生的数据，大家可能并不陌生，因 为我们每一个人都是这些数据的制造 者。我在拙著《浪潮之巅》中讲到过互联

网2.0的特点，它的本质是一个互联网的 平台，而上面的文字、图片、视频和各种 其他信息都是由用户提供的。在图片共 享网站Pinterest中，每天有7000万张图 片[1]被上传，累计上传了300亿张。 在Google旗下的YouTube视频网站， 数据量更是大得惊人，每分钟有300小时 的视频被上传到YouTube。至于互联网 用户每天在社交网络上的聊天和互动所 产生的内容就更多了。

EB丨百亿£1=节1 9 000 8 000 7 000 6 000

图5.6不同类型数据的增长

图5.6是思科公司对过去的5年里各 类数据增长的估计(和预估），其中增长 最快的是传感器带来的数据和用户产生 的数据。总的来讲,数据量的增长是惊人 的，甚至很多人在怀疑这是大数据的鼓 吹者夸大其词，他们说：“怎么过去没觉 得有这么多数据，一夜之间全冒了出 来?”事实上很多数据是大家不在意时被 收集的，比如各种传感器产生的数据，包 括摄像头、可穿戴式设备、手机的GPS (全球定位系统），以及各种采集声、光、 热和运动的传感器等。我们每天携带手 机，苹果公司就可以把每一个苹果手机 用户的出行路径记录得一清二楚。这类 数据总量之大远远超出常人想象，比如 像北京和上海这样千万人口的大都市, 摄像头的数量超过10万个，如果每个都 是每周7天、每天24小时监控，每个城市 产生的录像时长高达每分钟1700小时 以上，是YouTube的6倍左右。在过去， 因为没有条件存储这么多视频记录，常 常不存储或者只存一两天就删除，但是 今天人们已经发现它在城市管理中有着 重要的用途，比如通过录像识别违章的 车牌号,因此存储了大量的监控数据。这 样一来，我们存储的数据总量就陡然增 长了，仿佛是一夜之间从地下冒出来的。 从这里也可以看出，大数据兴起的第二 个必要条件就是存储技术的发展。

信息的存储

由于摩尔定律导致各种存储器的容 量成本增加，同时价格迅速下降,使得原

本不得不丢弃的一些数据现在有条件存 起来以供使用。比如在Google提供 Gmai暇务之前，企业级用户的电子邮箱 容量有限，公司雇员不得不经常删除邮 件，但是在Google宣布提供无容腳艮制 的邮箱服务后，各个电子邮件提供商不 得不采用同样的策略以维持用户，公司 里每个人的邮箱容量在不到10年的时间 里涨了 3个数量级，从几百兆(～100MB) 上升到几千万兆（~100GB)。邮箱里 的数据实际上只是企业级数据中很小的 一部分，在企业级的软件和服务中，大量 的中间数据被保存下来。各种互联网企 业所提供的服务，_般都会记录下详细 的曰志数据，比如任何一个合格的电子 商务公司都会记录下交易的详情，而搜 索引擎会记录每一次搜索非常详尽的信 息，比如搜索从哪里来，发生在哪一天的 几点几分，搜索的关键词是什么,用户点 击了哪些结果（或者广告），每一条结果 看了几分钟，等等。这些信息对改进产品 非常有用。

只是存储的容量上去还不够，因为 随着数据量剧增，查找和使用数据的时 间会变得相当长，因此存储设备的读写 速度也必须随着容量的增加而大幅度提 高。早期海量存储设备采用的是顺序访 问数据的磁带，因此大数据的使用显然 是不可能的，人们连存储数据的兴趣都 不大。20年前硬磁盘取代了磁带成为海 量存储设备，数据访问的时间缩短到原 来的大约1/1000,这时批处理数据不再 是个问题，人们开始重视收集和存储数 据。但是随机存储和访问数据依然很缓 慢，而且由于硬盘的速度取决于机械运 动，不可能大幅度提高，因此数据的使用 受到限制。直到大约七八年前，半导体的 固态存储器（Solid State Drives,简称SSD)的容量增加成本下降，才使得人们 能够很方便地使用数据，这时从存储技 术上讲，使用大数据的时机才成熟。

在能够产生大量的数据，也能够存 储这些数据之后，还有一个问题必须解 决，那就是这些数据怎样才能从采集端 传到存储设备上，这就要求数据传输技 术有所突破了。 传输的技术

由于数据的来源和采集点分布在不 同的地点，可能是许多不同的设备，也可 能在每个人身上、各个物件上面,在互联 网发展的早期阶段，人们还考虑不到把 这些东西通过互联网来连接，那时互联 网首先要解决的是把当时已有的计算机 连接到一起。在那样的通信环境里，即便 产生了大量的数据，也收集不到一起，因 此人们也不会去考虑大数据的问题。

到了移动互联网时代，这个情况发 生了根本性的改变，相比10年前第二代 移动通信系统GSM (全球移动通信系 统)只有不超过10OKB/s的数据传输率， 今天的第四代LTE (通用移动通信技术 的长期演进）的有效数据传输率达到2 MB/s~10MB/s，增长了几十到上百倍。 同时，Wi-Fi在主要城市的覆盖率已经非 常高，蓝牙也成为很多设备的标准配置， 这才使得数据在产生后可以迅速传到服 务器上。

信息的处理

当海量的数据被传到服务器上之 后，能否用得好就要看是否有足够强大 的数据处理能力了，因此信息处理的速 度也是大数据的一个先决技术条件。虽 然计算机处理器的速度可以按照摩尔定 律规定的速度每18个月翻一番，但是仅 仅靠单一处理器性能的提升依然无法应 对增长更快的数据量，这不仅是因为数 据量太大单机处理不了，而且因为当数 据量提高一万倍时，计算量通常不是成 线性增加的，大部分情况下，它会增加几 十万倍乃至上亿倍。[2]虽然有少量的 超级计算机有能力处理这样海量的数 据，但是这些计算机价格动辄上亿美元, 远不是一般公司和机构可以用得起的。

因此，应用大数据的一个前提就是 能够将一个大的计算任务分到很多台便 宜的服务器上去做并行计算。单一维度 数据的处理

不是一件难事，但是大数据 有多维度的特点，有时并行化是非常困 难的。没有相应的软件支持，很难将一个 复杂的大问题拆成很多小问题分配到多 台服务器上去做并行计算。并行计算的 另一个必要的技术条件是交换机和网络 速度得非常快，否则网络就成为计算的 瓶颈，服务器的处理器使用效率会非常 低下。事实上，市面上能够买到的最快的 交换机可能也达不到无传输障碍的海量 并行计算的要求。为了提高服务器之间 通信的速度，Google需要自己设计最快 的交换机。

上述计算问题直到2002年之后才 被Google等公司陆续解决，也就是在那 个时期，云计算开始兴起。通过互联网、 廉价服务器，以及比较成熟的并行计算 工具，实现了大规模并行计算，大数据的 处理才成为可能。

由于这些技术条件在10年前逐渐成 熟，才使得大数据出现井喷式的爆发。但 是今天大数据的应用水平依然处在初级 阶段，在机器智能方面人类其实才刚刚 起步。未来大数据和机器智能的发展，需 要在技术上有进一步的突破。

大数据实际上是对计算机科学、电 机工程、通信、应用数学和认知科学发展 的一个综合考量。在这里，我们选择一些 具有挑战性的关键技术进行一一介绍。 这些技术难题目前并不一定有最佳的解 决方案，甚至不存在什么绝对好的解决 办法，但是这些问题必须得到解决才能 保证大数据的普及应用。

数据收集：看似简单的难 题

按照信息论的观点，要消除不确定 性就需要信息，因此信息的收集非常关 键。大数据与传统的数据统计方法相比， 在收集数据方面有了很大的不同。

首先，传统的数据方法常常是先有 _个目的，然后开始收集数据。比如，人 们在发现天王星之后，发现它的运动轨 迹和牛顿力学预测出来的不一样，于是 预测在天王星之外应该有一个质量较大 的行星干扰它的轨迹。根据这个设想，天 文学家拍了很多星空的照片，想看看有 没有一颗位置在移动的、未知的星星，后 来找到了,这就是海王星。在大数据时代， 在收集数据时常常没有这样预先设定的 目标，而是先把所有能够收集到的数据 收集起来，经过分析后，能够得到什么结 论就是什么结论。正是因为在收集数据 时没有前提和假设，大数据分析才能给 我们带来很多预想不到的惊喜，也才使 得大家觉得计算机变得很聪明了。

在获取数据方面，大数据和传统的 统计方法另一个不同点在于，过去我们 是通过少量的采样获得所谓具有代表性 的数据，这些数据被称为样本。根据统计 学的原理，只要样本具有代表性,通过分 析这些少量的样本数据，就可以总结出 规律性。在过去的几个世纪里，科学家们 就是这么做的，不过他们在宣布自己从 有限数据中获得的规律性具有普遍意义 时，很快便有其他科学家会找到反例，在 局部范围内推翻原来的理论。这里面固 然有人类认知局限性的原因，也有样本 数据太少难以具有代表性的因素。亚里 士多德曾经给出一个看似很荒谬的结论 “男人的牙齿比女人多”，一些人认为他 可能是拍脑子想象出来的，不过作为格 物致知的先行者，亚里士多德并非一个 说话没有根据的人，他或许是数了几个 男人和几个女人的牙齿，恰巧那几个男人的智齿长了出来，而那几个女人的智 齿埋在牙龈中，于是他得出了“男人的牙 齿比女人多”的结论。这说明，我们常常 认为具有代表性的数据，可能并不那么 具有代表性。

当然，可能会有读者朋友质疑我对 亚里士多德的分析本身就是一种猜测， 不能说明人类无法获得具有代表性的数 据。但真实世界的情况是，获得足够量的 具有代表性的数据远比我们想象的要难 得多。回到电视收视率统计，对于收视率 较高的几个电视节目，统计结果一般是 比较准确的，但是对于那些收视率较低 的节目，统计的结果和真实情况相差一 两倍是很正常的事情。在Google内，我 们也发现类似的现象。我们过去一直采 用1 %的流量预测用户在搜索某个关键 词时所点击的搜索结果，对于常见搜索, 这个准确率非常高。但是对于不常见的 关键词组合，或者说长尾搜索，搜索结果 的概率分布比真实情况相差一两倍是很 常见的，甚至有时会相差一个数量级。

当然有人可能会问，你为什么要那 么较真，对于那些一天搜索不了几次的 关键词，点击数据的准确性差个一两倍 又有何妨，而在大部分情况下，传统统计 方法得到的结果也不过3%～5%的误差, 是可以接受的。但是，在商业上对这些细 节进行准确了解真的很重要，Google和 必应(Bing)在搜索质量上的一点点差异 就体现在这些细节上。如果统计永远有3 %～5%的误差，我们就无法在多维度上 得到可信的统计结果。

大数据则避免了采样之苦，因为大 数据常常以全集作为样本集。但是怎样 收集到全集就是一件很有挑战的事情 了，因为不能再采用过去抽样调查的方 式了。比如我们想要了解电视的收视率， 我们采用大数据思维，去了解每一个人 收视的情况，显然我们不能再采用过去 发调查问卷的方式了。事实上，收集到这 个数据(全集)最好的方法是通过电视机 的机顶盒记录用户的收视情况，当然那 些智能电视机也能记录这些信息。如果 能够获得这些数据，那么不仅能知道各 个电视节目的收视率，还能得知所插播 的广告的效果，如果进一步分析,还能够 知道每一个观众的特点。因此，这种没有 目的性的、全面的数据收集看上去优点 非常多。

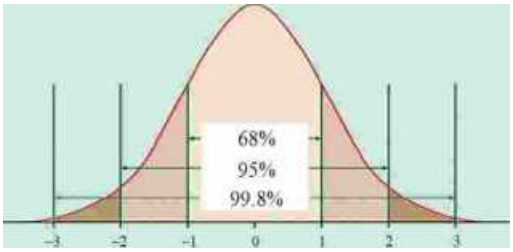


图5.7如果一个随机变量呈高斯分布， 总有5%左右的样本会落在两倍方差之 外

上面的想法固然不错，但是能够通 过机顶盒设备和电视机掌握用户数据的 只有它们的生产厂商和有线电视运营 商，而二者都不会轻易把这个数据分享 出来。这便是很多想利用大数据做事的 人和公司所面临的困境。因此，数据的收 集可以说是一个看似简单的难题。

那么，聪明的公司会怎样解决收集 数据的难题呢？最常见的方法就是绕一 个弯路，间接地收集数据，然后利用数据 的相关性，导出自己所要知道的信息。但 是这条路并不好走。

Google是一个重视数据的公司，它 很想了解每一个家庭的具体情况。为此， 它做了很多尝试，但大部分都失败了。 2010年，Google推出了自己的电视机 顶盒Google TV，为获取数据进入电视 广告市场做准备，但是，Google TV的销 售如此之差，以至于后来每个季度退回 来的机顶盒比卖出去的还多。最终 Google彻底放弃了这个产品，但是没有 放弃收集数据的想法。2014年,Google 斥巨资32亿美元收购了只有130名员 工、用户数量200万左右、还处在亏损状 态的nest公司。该公司的产品是具有自主学习功能和Wi-Fi的智能空调控制器， 可以根据人在家里活动的习惯控制空调 节省20%左右的电。如果单算经济账，这 笔投资不知道猴年马月才能赚回来，或 许永远挣不回来。Google之所以花如此 高的价格购买nest,最主要的目的是获 取每一个家庭的数据。nest智能空调控 制器的工作原理是靠跟踪家里人在每一 个房间里的活动，比如几点回家,几点看 电视，几点吃饭，晚上都待在哪里，什么 时候睡觉。在Google收购nest不久，它 又花了 5.55亿美元的巨资收购了家庭录 像监控公司Dropcam,这样就能获得更 多的居家数据。

Go gleTV

rilHtra WE!. KEB MEETS N,

图5.8 Google颇为失败的Google TV 产品，当初它的广告是电视机与互联网 的结合

在现实的世界里有一个匪夷所思的

现象。一方面，微软、苹果和Google这 些IT公司，为了挖掘每一个家庭的消费 潜力，想尽办法干方百计地要掌握每一 个家庭客厅的数据。它们有的通过游戏 机，有的通过类似机顶盒的设备 (Google过去的机顶盒、后来的 Chromecast,苹果的AppleTV),在为 用户提供服务的同时，在不经意之间收 集用户数据。另一方面，拥有这些数据的 公司除了统计一下收视率，计算一下可 能的广告观众,并没有什么大的作为。从 这个现象可以看出，一些公司已经敏锐 地看到了数据的价值，而另外一些公司 却拿着金饭碗在要饭，这其实反映出两 种类型的公司在方法论上的差异。

图5.9 Google旗下nes增能空调控制 器，其实是一个数据收集器

在收集数据时，我们还需要再一次 强调它是在无意之间完成的。我们前面 提到的关于央视收视率调查的例子就是 一个很好的反例。在数据的收集过程中， 非常忌讳那种“大胆假设，小心求证”的 思维方式，因为在很多时候，如果事先有 了定论，再找数据来证实它，总能找到有 利的证据，而这些看似被数据证实的结 论，很可能与真实情况相差十万八千里。 经济学家马光远先生曾经讲过一个故 事，说明带有偏见的“大胆假设，小心求 证”的危险性。

在2008年夏天，中国经济领域自认 为潜在的风险是热钱的涌入，只要翻翻 当时的报纸就能看到媒体天天在谈防止 热钱涌入这件事。但是，经济学家马光远 无意中从银行里接电话的业务员那里了 解到，在电话里，客户们都是要换外汇把 钱转移走，这和媒体上的说法完全相反。 事实证明，电话一端的基层业务员的信 息是正确的，媒体反而错了。媒体上的说 法来自官方，官方的说法也是有数据支 持的，并非完全拍脑袋想的。但是，由于 中国的经济结构并非完全市场化的，很 多政策是官方顶层设计的结果，那么想 要找到支持官方观点的数据很容易，这 样一来，看似有数据支持，但这样的数据 已经不具有任何代表性了。

在大数据这个概念不断地被炒作， 数据变得越来越值钱时，一些公司和个 人开始赤裸裸地收集用户的数据，然后 想办法卖钱[3]。事实上这样刻意收集 来的数据意义不大，因为收集数据的过 程会引起用户的警觉、恐慌和反感，一部 分对信息安全敏感的人可能会关闭收集 数据的传感设备，导致收集的数据不全 面;而另一部分人的行为会变得不自然。 这种变了形的数据，既不具有统计意义， 也失去了大数据的完备性。因此,真正高 明的公司都会像微软、苹果和Google那 样采用曲线救国的方法。有些时候，为了 收集数据，这个弯需要绕得特别大。

Google为了推出它的基于手机的 语音识别系统Google Voice,需要大量 的语音数据。在过去，各家语音识别公司 和实验室都是找人来录入数据，比如美 国标准的电话语音库Switchboard就是 这么构造的。这种类似于采样方法导致 的缺陷我们在前面已经介绍了，不再赘 述。Google的方法则不同，它为了收集 数据，先推出了一个类似玩具的电话语 音识别系统Google-411 [4](识别率相比后来真正的产品Google Voice是 非常低的），很多人出于试验和玩的目的 打这个电话，这样就在无意中为Google

提供了大量的电话录音。

数据的收集是一个开放性的话题， 不存在唯一的、最佳的方法。但是好的方 法一定能够保证数据的全面性(完备性) 和不变性。

数据存储的压力和数据表 示的难题

摩尔定律固然使得存储的成本大幅 下降，但是当大数据出现后，数据量增长 的速度可能超过摩尔定律增长的速度。 首先，很多原来不打算存储的数据被存 了下来，比如我们前面提到的很多情况。 而有些产品产生的数据多得惊人，比如 有了Google眼镜，就有可能将人一辈子 看到的事情全部记录下来，如果这件事 做成了，会彻底改变我们对世界甚至对 自己人生的了解。但是，将这些视频（包 括音频）数据都存下来可不是件容易的 事情。Google为街景地图服务的汽车每 天产生的原始数据量更是大得惊人，每 辆车每天产生的数据就是1TB,假如一 份数据存三个拷贝，一年下来就是1PB [5]。即使用当今最大容量的10TB硬盘来 存，也需要用100个（当然，数据在处理 之后并没有那么大）。

从图5.2中可以看出，数据量增长的 速度是高过存储设备发展速度的，越往 后，它们之间的差距越大。因此，不能简 单地依靠更多地生产和购买设备来解决 数据存储的问题，而是需要技术解决方 案来提高存储的效率，保证不断产生出 来的数据都能够存得下。

目前节约存储设备的技术体现在两 方面，第一类技术是存储同样的信息占 用的空间小。当然，这并不是简简单单的 数据压缩。从信息论的角度讲，就是要去 除数据的冗余，但是在去除冗余之时，相 应的数据读写处理要做改变。比如在邮 件中，同一份附件在所有的邮件中只存 一份,就可以大大节省空间，当然这会导 致邮件中文件管理系统的改变。再比如， 图像的存储由点阵变成向量，也可以大 大节省空间，但是这样就要改变图像的 读和写的方式。

图5.10 Google拍摄街景的汽车的摄 像设备为15个500万像素的照相机

第二类技术涉及数据安全，在这里 所讲的数据安全是指数据不丢失、不损 坏，而不是指防止数据被盗。在过去，防 止数据不丢失的最简单的办法就是多存 几个拷贝，放到不同的地点，比如过去AT &T (美国电话电报公司）关于业务的数 据就在美国三个不同地区要存三个完整 的备份。但是大数据都是存在云端，而且 数据量大，也不可能在世界各地存非常 多个拷贝，因此需要有特殊的方式保证 数据的安全性。Google的文件系统GFS [6]的设计从一开始就是为了方便存储 大数据而进行的。早期的GFS每个文件 在一个数据中心需要存3个拷贝,然后同 时存放在地理上相距较远的3个数据中 心，这样就是9份拷贝。虽然数据是安全 了，但是显然并不经济。后来改进成存3 +1份，前3份内容相同，最后一份是为了 方便校验和恢复信息，内容不同，这样只 需要存4份即可，大大节省了存储的空 间。

信息存储相关技术并不局限在研究 如何节省存储量上，还需要研究怎样存 储信息才能便于使用。在大数据之前，人 们在设计文件系统和数据存储格式时， 主要考虑的是规模较小、维度较少的结 构化的数据。到了大数据时代，不仅数据 量和维度都剧增，而且因为大数据在形 式上并不遵循什么固定的格式，过去需 要重新优化数据的格式，按照过去数据 特点优化设计的文件系统对大数据的使 用未必是高效率的，因此需要重新设计 通用、有效和便捷的数据表示方式和存 储方式。

我们不妨通过一个例子来说明大数 据的存储和过去数据的不同。大数据由 于量大，随机的访问就成为_个难题，为 了做到这一点，需要对数据建立索引，而 过去数据量不大时，索引实际上并非必 需的。建立索引对于有些数据并非难事， 比如机器系统产生的日志和互联网的网 页数据。前者虽然量大，但是每一条记录 中字段是清晰的，它们的表示（描述）、 检索和随机访问并不是什么大问题。网 页的数据虽然显得杂乱一些，但是它们 都是通过超链接文本组织起来的，从一 个网页就可以找到下一个，而且网页文 本的颗粒度都很小（是单词），因此我们 很容易通过关键词把它们索引起来。但 是到了富媒体数据大量出现时，要想随 机访问它们就不那么容易了，比如要想 从视频中找出一个画面就非常复杂，因 为我们即使找到了视频每一个主帧 (mainframe),也很难根据那些画面 对所有的视频建立索引。当数据量更大， 尤其颗粒度更大之后，这就是一个非常 难的技术问题了。比如对很多与医疗相 关的数据的随机访问就不是那么容易， 它们的基本单元动不动就是几百兆、上 千兆，用现有的技术来检索它们是不可 能的。如果不检索，就无法随机访问，那 么使用时在这么大量的数据中找到所需 要的，耗时特别长，很不实用。除了医疗， 还有很多行业，比如半导体设计、飞机设 计制造，它们的数据量都很大，而且很复

is :v

•r:l

••#•• •

••• •

•••••••

图5.11围棋的棋谱怎么表示能够便于 查询和搜索，就不是一个简单的问题(图 为AlphaGo和李世石对弈的第一盘）

大数据面临的另一个技术难题就是 如何标准化数据格式，以便共享。在过去， 各个公司都有自己的数据格式，它们只 在自己的领域使用自己的数据。但是，到 了大数据时代，我们希望通过数据之间 的相关性，尤其是大数据多维度的特性， 找到各种事物之间的关联。回到前面百 度知道的那个例子，如果我们能够往前 再走几步，将每一个用户的饮食习惯收 集起来，通过可穿戴式设备了解他们的 生活习惯，然后再和他们的医疗数据甚 至是基因数据联系起来，就能研究出不 同人、不同生活习惯下各种疾病的发病 可能性，并且可以建议他们改进饮食习 惯，预防疾病。这个前景看起来很美好, 但是要实现它就必须先解决数据的表 示、检索和随机访问等问题。显然，对于 世界上各种各样的大数据，无法用一个 统一的格式来描述，但是大家需要一些 标准的格式，以便相互交换数据和使用 数据。最早进入大数据领域的Google公 司设计了一种被称为Protocol Buffer的 数据格式。在Google内部，Protocol Buffe「是数据存储的主要格式，也是它所 开发的各种软件在进行数据通信时标准 的接口。今天，Google已经将Protocol Buffer开源出来供大家使用，旨在便于全 世界能够共享数据。

大数据的应用方法和场景与过去使 用数据完全不同，这不仅带来了上面所 说的在数据存储和表示方面的挑战，也 带来了数据处理的挑战，而这些挑战并 非简单增加处理器就能够解决的。

并行计算和实时处理：并 非增加机器那么简单

大数据由于体量大、维度多，处理起 来计算量巨大，它的使用效率取决于并 行计算的水平。我们在前面提到了 Google的MapReduce (编程模型）和 雅虎的Hadoop (海杜普）等工具，它们 能够把相当_部分大型计算任务拆成若 干小任务在很多并行的服务器上运算。 这确实给大数据处理带来了福音，但是 并没有完全解决计算瓶颈的问题。在一 般人想象中，增加10倍的处理器并行计 算，可以同样成倍地节省时间，但是在工 程上这是做不到的。

首先，任何一个问题总有一部分计 算是无法并行的，这类计算占比越大，并 行处理的效率越低。在计算机科学中，通 常用可并行比例（Parallel Portion)来 度量在一个任务中有多少是可以并行计 算的，有多少不能。图5.12给出了在不同 的可并行比例下，并行计算处理器的数 量和实际加速(speed up)之间的关系。 从图中可以看出，如果在一个任务中能 够并行处理的比例越高，实际的加速越 多，但是即便只有5%的计算不能并行, 那么无论使用多少台服务器，实际的加 速也不会超过20倍。

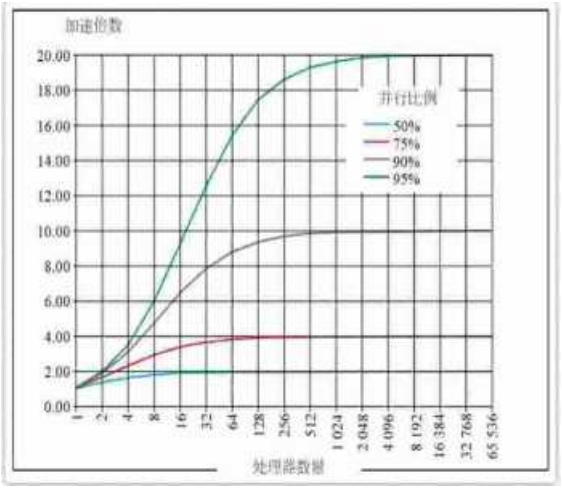


图5.12在不同的可并行比例下，增加 处理器数量和实际加速的关系曲线

另一个影响并行计算效率的因素在 于无法保证每个小任务的计算量是相同 的。例如，我们要进行两个1000000 x 1000000的大矩阵相乘，一台服务器显 然难以完成这样的任务，因此我们使用 MapReduce或者Hadoop在1万台服 务器上进行并行计算。我们通常会把这 个大矩阵按照行或者列分成1万份，每份 100行（列），每个服务器上分一份。但 是，虽然一个服务器上的任务看上去都 是计算100行（列），但是这些小任务的 计算量未必均衡，其中_个是另外一个 的两三倍是一件很常见的事情。这样一 来，并行计算的效率就大打折扣——完 成了自己计算任务的服务器，在等待个 别尚未完成计算的服务器。最终的计算 速度取决于最后完成的子任务。如果考 虑到一些子任务会因为系统不稳定出现 计算错误需要重新计算，并行计算的效 率还会进一步降低。因此，并行计算的时 间是远远做不到和服务器数量成反比。 事实上，使用的处理器越多，并行计算的 效率越低。

大数据处理的另一个挑战是对实时 性的要求。一些看似简单的操作一到大 数据头上就特别费时间。比如过去用 Excel (数据处理软件）在几万行数据中 找到最大值只要一两秒钟的时间，排个 序所需要的时间也不过十几秒钟。但是 在一个几千万行的电商销售日志中要找 到销量最好的商品，或者将商品按照销 量排一个序，即使采用上千倍的处理器， 也不可能在几秒或者几十秒内完成。这 里面的原因除了我们前面提到的并非所 有的计算都可以并行化之外，还因为早 期的大数据都是存储在硬盘上的，而且 并行计算工具，比如MapReduce或者 Hadoop,都是批处理形式的。通常上述 操作的处理时间至少要几十分钟，这对离线的数据分析可能不是一个大问题， 但是如果公司主管想实时了解经营情 况，这个等待时间就无法忍受了。要解决 实时处理大数据的问题，就需要从根本 上改变系统设计和算法，而不是增加机 器那么简单。事实上对任何大数据问题 都做到实时处理是不可能的，但是对于 很多特定问题，比如对于日志等结构化 或者半结构化数据，还是有可能的。比如， Google为了解决上述问题，专门设计了 —个被称为Dremel[7]的工具，专门针 对曰志、数据库等大数据，解决实时访问 和简单的数据处理问题。与传统的文件 系统或者数据库不同的是，它的文件是 基于内存的而不是硬盘的，而且在数据 的存放上和传统数据库系统不同， Dremel采用以数据列为优先的方式存 储，而传统的数据库系统是以行为优先 方式存储的。Dremel这样的特殊设计是 为了方便多维度数据按照某个特定维度 进行处理和数据挖掘。当然，类似 Dremel的工具还有很多，通过它我们只 是想说明针对大数据的实时处理需要开 发很多新的工具，而不是简单地把过去 的工具并行化就可以。

数据挖掘：机器智能的关 键

使用大数据，相当于在一堆沙子中 淘金，不经过处理的原始数据是给不出 什么新知识的，大数据能产生的效益在 很大程度上取决于使用(和挖掘)数据的 水平。在Google,至少有四成的工程师 天天在处理数据，然后通过数据得到知 识，通过知识使得计算机变得更智能。

我们在强调收集大数据是无目的性 的同时，也给处理大数据增加了难度。由 于大数据的原始数据常常是没有固定格 式、显得杂乱无章的，因此使用大数据的 第一步是对数据的过滤和整理，去除与 要解决的问题无关的维度，将与问题有 关的数据内容进行格式化的整理，以便 进一步使用。数据的过滤和整理有时很 容易，比如我们希望通过日志分析一款 游戏玩家的行为，这只要把相应的维度 保留下来，把无关的信息过滤掉即可。但 是，在很多应用中,即使这一步也不容易 做到。比如在前面提到的机器自动问答 的例子中，虽然问题的答案存在于网页 之中，但是答案的内容通常是七零八落 地分布在不同网页里的，对网页的结构、 内容进行分析就成了使用大数据的先决 条件。当然,如果没有很好的自然语言理 解技术，这第一步都无法完成。

虽然香农告诉我们，信息越多，我们 就越能消除系统的不确定性，但是数据 中常常不仅仅是信息，还不可避免地夹 杂着噪声，这个问题在大数据中特别明 显。使用数据的人常常会发现某些数据 的质量高，而另外一些数据的质量不是 SP么高。当然，简单地用质量高和质量低 是不足以准确定量地衡量数据质量的。 在信息处理领域，大家使用一个被称为 信号与噪声之比（Signal Noise Ratio, 简称信噪比SNR)的度量来描述信号的 质量。如果数据中的信噪比很高，数据就 可靠；相反，如果信噪比太低，可能有限 的信息会被淹没在噪声中，这样的数据 使用后可能产生不了什么好的结果。对 于那些夹杂着大量噪声的数据，为了提 高数据的信噪比，在使用数据之前，我们 常常需要进行降噪处理，损失一部分数 据，以提高信噪比。 图5.13信号中常常夹杂着噪声，当信 噪比较高的时候，依然能够恢复出原有 的信号

现在，我们可以认为这样处理过的 数据能直接使用了，接下来关键的一步 就是机器学习。机器学习并不是什么新 鲜事，今天广泛使用的机器学习算法，比 如人工神经网络算法、最大熵模型、逻辑 自回归等，早在40年前就已经成熟了。但 是由于数据量不够，导致机器学习的应 用范围比较窄，再加上它是介于应用数 学、统计学和计算机科学之间的交叉领 域，因此一直没有受到太大的重视。2000 年以后，随着计算机速度的增加和数据 量的暴增，机器学习在很多领域发挥了 重大作用。2016年Google创造奇迹的 AlphaGo,其训练算法就是人工神经网 络。

但是如果认为机器学习就是把几十 年前的论文拿过来用计算机的程序实现 一遍，那也未免太天真了，因为机器学习 —旦上了规模，实现起来可不是一件容 易的事情。不幸的是，大数据的机器学习 还真是一个上规模的难题。要理解为什 么数据量一大机器学习就变得非常困 难，我们不妨简单介绍一下机器学习的 原理。

机器学习的过程无一例外是一个不 断迭代、不断进步的过程，用机器学习的 专业术语来说就是“期望值最大化” (Expectation Maximization)的过程: 只要事先定出一个学习的目标，这些算 法就会不断地优化模型，让它越来越接 近真实的情况。可以说，机器学习训练算 法迭代的次数越多，或者通俗地说学习 得越深入，得到的数学模型效果越好。因 此，同样的数据，同样的算法，采用不同 深度的机器学习方法，得到的结果会有 所不同。

但是机器学习的算法通常都比较 “慢”，用比较专业的术语讲，就是计算 复杂度太高[8]，因此随着数据量的增 加，计算时间会剧增。在过去，由于计算 能力的限制，以及并行计算工具不够有 效，人们在机器学习时，通常要在下面两 种情况下二选一：

1. 数据量大，但是采用比较简单的 模型，而且比较少的迭代次数，也就是说 用大量的数据做一个浅层的机器学习。

2. 数据量较小，但是采用比较复杂 的模型，而且经过很多次迭代训练出准 确的模型参数。

通常，由大量的数据、较少迭代训练 出的“较粗糙”的模型，要比用少量的数 据、深度的学习精耕细作得到的模型效 果更好。

是否有可能用大量的数据，进行深 度的学习，然后得到更好的模型呢？从理 论上讲，有这个可能性，而且结果一定会 更好。但是，在实际应用中非常难做到， 原因是这样的计算量很大，不仅计算时 间长，而且需要计算机系统有非常大的 内存空间，通常不是几台计算机能够完 成的。Google的AlphaGo虽然在和李世 石下棋时只用到几十台服务器，但是训 练时可是需要上万台服务器的。如果采 用成千上万甚至几十万台计算机并行处 理，那么过去老的机器学习算法是无法 搬到成千上万台计算机构建的并行处理 系统的，需要将过去的机器学习算法重 新工程化才可行。

2010年，Google宣布开发出名为 Google大脑(Google Brain)的深度学 习工具。从机器学习理论上来讲，它没有 任何突破，只是把过去的人工神经网络 并行地实现了。但是从工程的角度上来 讲，它有非常大的意义。首先，过去的人 工神经网络无法训练很大的模型，即使 计算的时间再长也做不到，因为内存中 根本放不下和模型参数相关的数据。 Google的突破在于找到了一种方法，可 以将_个很大的模型上百万参数同时训 练的问题，简化为能够分布到上万台(甚 至更多)0g务器上的小问题，这样使得大型的人工神经网络训练成为可能。当然， Google还找到了（不是发明了）一些对 大模型并行训练收效比较快的训练算 法，可以在能够接受的时间内，深度训练 出一个大型的数学模型。Google在几个 带有智能特色的问题上，用这个深度学 习的工具对语音识别的参数进行重新训 练，就将识别的错误率降低了 15%(相对 值）[9]，这对于机器翻译效果同样显 著。

Google大脑的成功不仅向业界展 示出机器学习在大数据应用中的重要 性，而且通过实现一种机器学习并行算 法（人工神经网络），向大家证明了深度 学习所带来的奇迹。至于Google选择人 工神经网络作为机器学习的算法的原 因，听上去匪夷所思，细想起来却很有道 理——人工神经网络的核心算法几十年 来基本上没有变过。人们从直觉上一般 会认为不断改进的方法才是好的、应该 采用的,但是在工程上却不然，像Google 大脑这样试图解决各种问题（而不是一 个特定问题)的大数据机器学习工具，实 现起来工作量巨大，一旦实现，就希望能 够使用很长时间，因此算法需要稳定，不 能三天两头地改进。说到这里，读者朋友 可能要问，采用一个几十年前的算法，是 否会让机器学习的效果受影响。对于某 些特定的问题，确实会有一个机器学习 算法比其他的好这种情况，但是总体来 讲，大部分机器学习算法是等效的，只有 量的差别，没有质的差别，而量的差别可以通过规模和数据量来弥补，因此， Google的做法不失为一种好的折中。事 实上，Google的AlphaGo采用的是同样 的训练算法，这也是Google强调它的算 法是通用的原因。[10]

到上一层的输入 8个通道

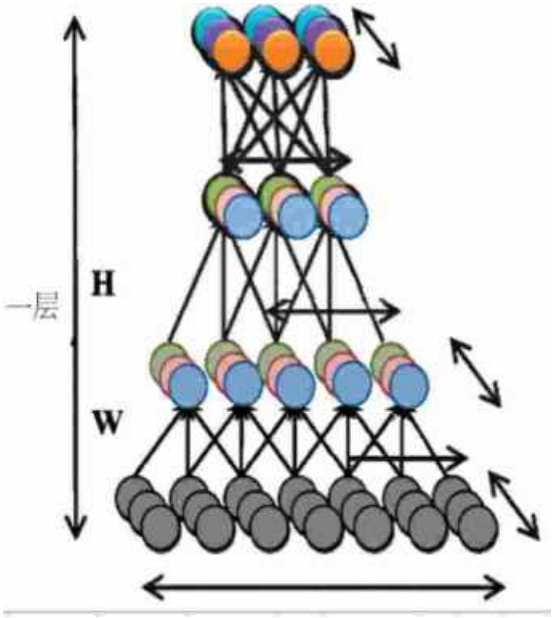


图5.14 Google大脑的核人工神 经网络

当然，机器学习的算法很多,Google 或者某个大公司也不可能把每一种算法

都实现得最有效，而一般的公司也不可 能有技术力量去开发工程难度很大的机 器学习软件，因此最好的解决方式就是 出现一些专门做机器学习的公司，来为 需要使用大数据和机器智能的公司提供 服务。2012年Google在安迪•鲁宾的 主导下，以5亿美元的巨资收购了只有 100人左右的小公司DeepMind。这家公 司对外宣传其所做的事情是让计算机有 思维，其核心技术就是研究通用的机器 学习算法，而AlphaGo就是DeepMind 团队为了证明他们机器学习算法有效性 而开发的一款智能程序。在此之后， Google还收购了多家小型机器学习和 人工智能公司。Facebook、雅虎等公司 也做了相应的并购或者战略人才引进， 由此可以看出目前很多IT公司都在大数 据的挖掘和处理上进行战略性布局。

机器学习的方法不可能由每家公司 自己去研究，最终会由专业的公司为大 众提供机器学习的服务。但是这样又会 引发大家的一个忧虑，那就是数据安全 和隐私保护的问题。

数据安全的技术

大数据应用的一个挑战来自对数据 安全性的担忧和对隐私的诉求。这一节 我们重点讨论数据安全，在下一节中我 们将讨论隐私保护的问题。

数据安全有两层含义，首先是要保 证用户的数据不损坏、不丢失。10年前, 云计算刚开始普及，大家所担心的是数 据存储在云端会丢失。经过了 10年，互联 网用户或多或少都有了使用云计算的经 历，已经体会到数据放在云端上的方便 性。在这10年里，也没有发生什么数据存 在云端取不回来的情况，实际上比放在 自己的电脑上或者手机上安全多了，因 此大家不再担心这方面的问题。

但是数据安全还有第二层的含义， 即要保证数据不会被偷走或者盗用。在 过去的10年里，由于不断传出有犯罪分 子或者恶意的黑客进入计算机系统中偷 盗数据的事件，而且确实给公司和个人 带来了很大的麻烦，因此今天大家担心 的是自己的数据是否会被别人偷盗，以 至于让自己蒙受很大的损失。

令人遗憾的是，过去各种安全防范 的方法，虽然防住了绝大多数黑客和数 据偷盗者的入侵，但总是有一些漏网的。 好在过去数据通常不是集中存放，因此 即使系统被黑客攻破，丢失的数据常常 有限，损失也有限。更重要的是，由于数 据大多是单一维度，或者是低维度的，所 以损失比较直接，可以估量。直到2002 年，由于计算机的存储量不足，像美国最 大银行之一的美洲银行，各个州储户的 账号还是单独存放。那时在美国西海岸 加州的美洲银行开户后，到东海岸的马 里兰州办事处想要调出个人信息，是一 件近乎办不到的事情。对银行来讲，这当 然操作不方便，但是也避免了用户信息 被一锅端。

但是在大数据时代，由于数据量巨 大，数据一旦丢失，损失也是巨大的。比 如2013年美国百货连锁商塔吉特数据 丢失造成的损失高达1.6亿美元。[11] 2014年曝出的索尼丢失数据事件，造成 的损失高达1亿美元。[12]更早的时候， 美国折扣连锁店TJ Maxx的数据丢失造 成了 2亿美元的损失。在这几起信息被偷 盗的案件中，用户信息都是被一锅端的。

比商业数据丢失后损失更大的是医 疗记录被盗，据加州几家信息安全公司 给出的参考数据，在美国黑市上,一个医 疗记录的卖价是个人商业数据的50倍左 右。美国每年有不少医疗健康信息被盗, 甚至个别的医疗仪器被黑客劫持，整个 医疗系统被黑客勒索的赎金高达数十亿 美元，只是绝大部分患者不知道而已。

Recent Hacking Incidents around@ti'© world

CRED[T CARDS & CUSTOMS INFO

图5.15塔吉特的计算机系统被黑客攻

击，半个多月后才被发现，4000万顾客 信用卡信息被盗

当然，比数据集中存放更让业内人 士不踏实的是一旦黑客得到多维度的数 据，从理论上讲，黑客也像数据科学家一 样对大数据进行分析，那么机密泄漏的 损失就大得难以估量。我们在前面提到， 既然合法使用数据的用户利用大数据分 析可以得到意想不到的惊喜，那么非法 入侵的黑客同样可以得到。

有经验的IT系统主管和架构设计师 都知道要尽量将敏感信息放到不同的地 方，以免多种敏感数据同时丢失。但是这 件事情执行起来并不容易，因为如果一 项安全措施导致操作麻烦，很多人就会 不遵守，比如在很多公司里，操作人员为 了方便，通常习惯把分开存放的数据又 拷贝到同一个地方一起处理，原先出于 信息安全所做的设计就形同虚设。通常 人们在方便性和安全性方面会优先考虑 方便性，这是人的天性使然。

在大数据时代，虽然计算机系统在 设计时对安全性的考虑比过去周全了许 多，但肯定无法百分之百地防止黑客入 侵行为，因为很多时候这是人为失误造 成的，而不是防火墙不够先进。比如前面 提到的塔吉特遭受黑客入侵的事件，其 实防火墙已经报警了，但是由于报警频 率太高，操作人员嫌烦而关闭了报警系 统，这才惹出大祸。有些时候，人的安全 防范意识要比想象的差得多，我本人就 遇到过一件事情令人匪夷所思的事情。

几年前，我陪同母亲到加拿大游玩， 然后回到美国。根据美加两国的协议，美 国把海关设在加拿大机场内，这样在登 机前实际上要先经过美国海关。或许是 因为海关的计算机设备放在了加拿大， 或许是其他原因，总之在海关窗口的那 台计算机里找不到我母亲的信息。于是 海关工作人员就不得不让她到旁边办公 室办理入关手续，因为在那里有一台计 算机直接连到美国外交部，能够访问完 整的数据库。由于母亲的英语不流利，我 就被允许和她一起进入那个房间。或许 是因为那台计算机连着外交部的数据 库，颇为敏感，如果操作人员两分钟不操 作，就自动退出了，这时海关的官员必须 重新输入密码登录。那位海关官员对我 母亲问话的时间显然不止两分钟，因此 他不得不一遍遍输入密码。当然，这样敏 感的账号自然要求密码设置非常复杂， 比如要各种类型的字符组合在一起，这 样的密码其实很难记得住，于是他把密 码写在了一张纸上，那张纸就放在办公 桌

上，密码被旁边的我看得一清二楚。我 当然对进入外交部的数据库没有兴趣, 但是如果是一个黑客而不是我看到那张 纸，麻烦可能就大了。

当然，可能有读者会觉得我遇到的 是个案，事实上，我们在Google接到用 户账号被盗的报案中，一大部分情况是 用户自己把账号写到什么地方，被人给 偷盗了。既然不能够完全把偷盗者挡在 外面，就需要有更好的方式来保障信息 安全。

^tmazon： C7Lt〇nff&f enL ui57Tt0J §.mail： lJ〇33<KsSpO 图5.16那些很复杂的密码反正记不 住，只好抄到纸上，结果更加不安全

科学家和工程师们首先想到的是在 文件系统和操作系统设计上加以改进。 直到今天，文件系统和操作系统的设计 和40年前没有本质的差别，而在那个年 代，信息安全的矛盾并不突出，因此在数 据安全性上的考虑并不多。早在2001年， 一位计算机科学家在IEEE (电气和电子 工程师协会）的_个信息安全的研讨会 上就指出，计算机系统的设计和高楼设 计很大的不同是，前者事先并不考虑安 全的隐患，而后者在每一个环节都要考 虑安全的问题，这就是我们面临的现实。 因此，从系统上根本地解决信息安全的 问题，其必要性是显而易见的。不过这并 非一朝_夕能够办到的事情。

另一种行之有效的方法恰恰是利用 大数据本身的特点，来保护大数据的信 息安全。通常_家机构里的业务流程是 固定的，被授权操作员的使用习惯也是 可以学习的。比如我前面提到的海关官 员操作外交部信息档案库的流程，通常 可能会是从A点到B点，再到C点、D 点……但是，假如外来的闯入者真拿到 了密码进入外交部的计算机系统，由于 他对外交部内部的业务流程并不了解， 他的操作可能直接从A点绕到C点，然后 跳到E点，因此可以通过大数据发现并制 止异常的操作。麻省理工学院计算机和 人工智能实验室(MIT CSAIL)的研究表 明，利用大数据（2000万用户产生的36 亿行的系统日志）分析来防范黑客攻击， 要比传统的在防火墙设置各种规则的做 法有效5倍。[13]而在工业界，一些信 息安全公司已经开始按照这种思路来设 计和研制产品了。

硅谷的Trustlook公司和中国一家 电信运营商的信息安全服务公司就是这 么做的。它们都利用大数据分析和机器 学习了解公司正常的业务流程，发现并 防止异常操作。不仅_家公司正常的业 务流程可以学习，当数据量足够大时，每 个被授权的使用者的操作习惯也可以学 习，那么不符合这些习惯的操作就可能 来自非法的闯入者，这些操作就会被禁 止。类似地，日本一个发明家将这种思路 用于汽车的防盗。他发明了_套检测驾 驶员身材信息和操作信息的监控系统， 能够根据平曰里经常驾驶某辆车的人的 身材信息、坐姿和动作，判别是原来的司 机还是新来的人。如果某个偷车贼偷到 了钥匙试图把车开走，那么该系统一旦 发现这个人平时没见过，就会要求他输 入密码，如果密码输入错误，汽车会完全 关闭，不能启动。这种防盗方式和保护信 息安全的方式异曲同工。

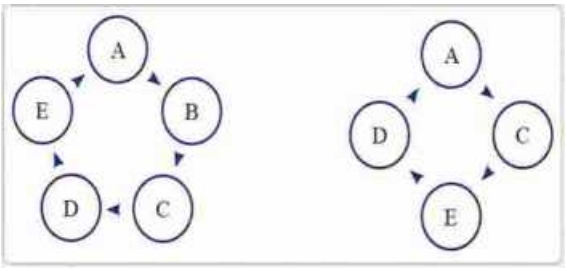


图5.17用大数据分析流程是否正常,

左边为常规的流程，右边为异常的，可能 是由黑客在操作

保护隐私：靠大数据长期 挣钱的必要条件

由于大数据具有多维度和全面的特 点，它可以从很多看似支离破碎的信息 中完全复原一个人或者一个组织的全 貌，并且了解到这个人生活的细节或者 组织内部的各种信息。这样就会引发大 家对隐私权的担忧。比如在塔吉特公司 预测孕期那个案例中，实际上那个怀孕 的未成年少女的隐私已经被泄露。好在 塔吉特作为年销售额超过700亿美元的 百年老店，没有必要冒着毁坏百年信用 的危险泄露那个人的隐私。

为什么要保护隐私，对这个问题的 回答恐怕是仁者见仁，智者见智,但通常 大家有一点看法是一致的，那就是赤裸 裸地生活在众人的目光下不舒服。我们 每一个人都不是完人，都或多或少有些 并非十分光彩的一面，那一面如果被熟 人知道了，对生活会有很坏的影响，比如 艳照门的那几位主角，受到的负面影响 是一生的。

在过去的历次技术革命中，都没有 过多涉及个人隐私的问题，因为那时技 术的发展和个人隐私关系不大。但遗憾 的是，在大数据时代，技术的发展和保护 隐私开始产生矛盾。比如我们前面在介 绍各种智能交通管理工具时展示出了它 的好处，但是另_方面，如果某家提供这 种服务的公司无限制、无节制地收集用 户数据，实际上每个人的行踪都可能暴 露在大众面前，这是非常危险的。很多公

司现在已经具有了这样的能力，只是大 家不知道或者不注意而已。

2013年，一位特斯拉汽车的主人抱 怨他的电动车在充满电之后，跑的距离 没有像特斯拉公司声称的那样远。特斯 拉马上回应道，车主人所走的路线并不 是他向媒体讲的那条路，而是一条绕远 的路。这件事情曝光之后大家的注意力 马上从电池的续航能力转移到了个人隐 私方面。根据位于硅谷的圣塔克拉拉大 学的法律学教授葛兰西（Dorothy Glancy)介绍[14]，不只是特斯拉一 家公司在获取汽车车主的数据。96%的 新车都有类似于特斯拉这种追踪车主行 踪的功能，而且获取的数据比我们想象 的多很多，比如是否系了安全带。更关键 的是，绝大部分车主并不知情，而且无法

关闭这些监控的功能。

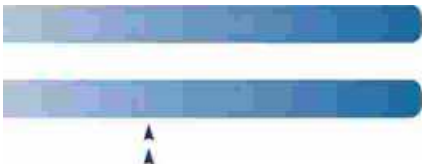
或许我们已经习惯了出门在外被这 种摄像头监视，对于私家汽车里面安装 上述数据采集装置也只好听之任之。我 们天真地以为，至少在家里关起门来外 面是不可能知道家里发生的事情的，但 是情况并非如此。像nest这样的智能家 居可以知道家里每个人的活动，甚至知 道什么人来访。如果一个家庭妻子出差 了，丈夫带另一位女士来家里过夜，这件 事在今后恐怕是瞒不住的。当然，从好的 方面想，通过获取这些细节的数据，至少 有助于反腐败，能够发现私底下的官商 勾结、权钱交易，帮助查处贪腐获取证据。 但是这样一来，我们也毫无隐私可言了。

大家通常会夸大大数据带来的便利 之处而忽视它对个人隐私带来的危害， 因为大众对于隐私的重要性普遍不够重 视。当然可能会有一些读者挑战我的这 个说法，认为大众，特别是欧美国家的人， 还是很重视保护个人隐私的。但实际上, 欧美国家的人常常也只是把保护隐私挂 在嘴边而已。

为了证实这一点，凯文•凯利和我 分别在挂谷地区对社交网络的用户做过 这样一个调查，看看他们到底是在乎自 己的隐私还是更多地希望获得便利性。 我们的调查方法基本上相同，都是通过 一份精心设计的调查问卷对各种文化背 景的男女老少进行抽查。调查问卷包括 三部分内容。首先我们列举了很多社交 网络产品和移动互联网产品，让被调查 者说明他们都用了哪一些。对这些产品， 我们当然清楚使用者会暴露多少隐私， 但是被调查者未必清楚。其次，我们列举 了一些需要牺牲隐私换取便利的服务， 看看被调查者有多少原因使用那些服 务。最后，我们给被调查者一个可以拉动 的游标，让他们在保护隐私和方便性之 间做一个选择，比如最左边是彻底保护 P 急私，但是也彻底失去移动互联网和社 交网络带来的方便性，最右边是彻底放 弃隐私，但是得到当下各种技术和产品 带来的全部便利性。

绝大部分被调查者在最后一项调查 中会选择把游标拉到50%的位置（图5. 18)，或许他们认为这样便保持了在隐私 方面和便利性之间的平衡。但是，根据前 两项的统计结果我们发现，用户在行动 上的选择是放弃隐私以换取便利性，或 者说把游标拉到了非常右边的位置，如 图5.18所示。

0%隐私保护 100%便利性



100%隐私保护 0%便利性

用户心理上的选择 A 社交网结的实际谕® T

图5.18用户在心理上和行动上对保护 隐私认识的差异

当然，在这里我要说明的是我们调 查的人群仅限于挂谷地区，加上调查的 人数有限，因此我们的结果难免有所偏 差，但是被调查者如此一致地选择便利 性而不是保护自己的隐私，应该足以说 明大家对隐私问题的忽视。

我们也对各种APP和社交网络产品 在保护隐私和提供方便两方面做了评 估，有意思的是，它们和用户在行动上的 选择非常一致，即将便利性推到极点而 不顾及大家的隐私。这或许是巧合，但也 可能是社交网络和移动互联网APP的提 供者在不断地测试用户对暴露隐私的承 受底线，只要用户不抱怨，它们就做得越 来越过分。比如，一款手电APP要求访问 移动互联网的所有联系人的信息时，它 要访问的信息和它提供的功能是完全无 关的。当然，这些刺探个人隐私的公司能 够得逞的原因，是用户自己将自己的隐 私交给了那些毫不相干的公司（或者个 人）。

大众在大数据时代对自己的隐私如 此不在意，可能有三个原因。首先是不清 楚大数据按照目前这个方式发展，最终 会严重侵犯个人隐私，因为在过去的技 术革命中这不是问题。其次是抱着侥幸 的心理，认为那么多用户数据，怎么可能 数据的拥有者和操作者正好能挖掘到我 的隐私，这是因为他们对大数据所带来 的机器智能不了解，事实上这不需要人 工去做人肉搜索，计算机可以自动完成 挖掘任务，而且做得非常智能。最后，很 多人会觉得，我既不做什么坏事，也不担 心行踪被暴露，也不是什么名人怕大家 知道什么秘密，那些拥有我的数据的公 司即便知道我的隐私，也损害不了我的 利益。这种想法实际上是大错特错，因为 用户的利益在隐私暴露之后很容易就被 损害。我们不妨看看如下几个例子。

人们在中国某大型电子商务网站上 发现，某些人总是买到假货，而另外一些 人以同样价格却买到真货。这并不是因 为前者比后者的运气差，而是商家掌握 了太多的个人数据,或者说我们的隐私。 当商家知道前者是买了假货也不会吭声 的软柿子，后者是睚眦必报的刺头的时 候，欺软怕硬的行为一定能够给他们带 来最多的利益。在利用大数据方面，个人 用户相比商家永远是弱势群体，一旦他 们的秘密被商家知道，他们的利益就难 免受到损害。

美国很多航空公司在利用个人隐私 大发其财。当航空公司发现某个机票的 询票者最近必须旅行，而且在过去对票 价不是很敏感时，它给出的报价就会比 给其他人的高很多。尤其当两个城市间 仅此一家航空公司有直飞的航班时，价 格上的差异就更明显。这些航空公司甚 至出钱聘请了美国一些著名的大学帮助 研究这样利用用户隐私赚钱的方法。据 一所世界名校里承接这些项目的团队介 绍，利用对用户行为的分析，可以让航空 公司提高10%左右的销售额。虽然10% 的提高听起来不算太多，但是对净利润 率只有0.2%的航空业来说，这是几十倍 的利润的提高。而对于乘客来说,由于只 是部分乘客受到伤害，整体上10%票价 的提高意味着他们的额外支出要远比10 %多，实际上多付出的票价可以高达50 %。

我们必须清楚地认识到，保护好隐 私对大数据长远的发展非常重要。人们 不可能看到自己的隐私最终完全受到侵 犯，而依然任由大数据继续发展下去，因 此，隐私的问题不处理好,将对大数据的 长期发展不利。今天，在医疗卫生这样敏 感的行业，担忧隐私受到侵犯已经成为 这个行业大数据和机器智能发展的障 碍。关于大数据和隐私在未来社会中可 能产生的矛盾，我们在下_章还会详细 论述，在这里我们探讨是否有可能在技 术上做到既能够利用大数据带来的便 利，又能保护好个人隐私。

为什么必须在技术上保护隐私，而 不仅仅是在法律层面靠处罚来解决侵犯 隐私的行为呢？虽然在法律层面保护隐 私是必需的，但是光靠法律是解决不了 问题的。首先，很多侵犯隐私的行为是个 人行为，比如偷窥，很难发现和查处。其 次，法律的制定永远落后于案件的发生， 尤其是在大陆法系的国家。因此,除了法 律手段外，我们还必须有相应的技术手 段维护个人的隐私，但是这又谈何容易。

需要指出的是，保护隐私并非简简 单单地屏蔽掉一些个人信息那么简单。 在过去这种方法是有效的，因为各种维 度的数据联系不到一起。但是在大数据 时代，由于大数据多维度和全面性的特 点，简单屏蔽掉的很多信息是可以从其 他维度利用相关性恢复的。因此，保护隐 私需要新的技术。

一类保护隐私的技术是从收集信息 的一开始就对数据进行一些预处理，预 处理后的数据保留了原来的特性，使得 数据科学家和数据工程师能够处理数 据，却“读不懂”数据的内容。这样至少 能防止个人窃取和泄露隐私，但是并不 能限制那些拥有非常多数据的大公司了 解每一个人的隐私。

另一类保护隐私的技术是所谓的双 向监视。这是一个很新颖的保护隐私的 想法，简单地讲就是当使用者看计算机 时，计算机也在盯着使用者看。大部分人 喜欢偷窥别人隐私的一个原因是，这种 行为是没有任何成本的。但是，如果有人 在刺探别人隐私时，他的行为本身暴露 了，那么他就会多少约束自己的行为。这 就好比一个偷窥者悄悄推开门缝往里面 窥视，发现里面有双眼睛正在看着他，那 么他的反应可能是马上把门关上。凯 文•凯利对各种保护隐私的技术做了评 估，他和研究人员发现，如果给窥视者_ 个选择，输入自己的真实信息然后才可 以窥视他人，那么绝大多数人会选择直 接离开。正如制约权力最好的办法是使 用权利，解决一种技术带来的漏洞最好 的办法是采用另一种技术，那么保护隐 私最好的办法或许是让侵犯隐私的人必 须以自己的隐私来做交换。

总结上述两种技术的特点，我们可 以看出，为了在使用大数据的同时尽可 能地保护隐私，数据从采集到使用都需 要是双向知情的，也就是说不再是数据 的所有者暴露在大庭广众之下，数据的 采集者和使用者（偷窥者也是_种特殊 的数据使用者)也是同样被监督的，或许 这样是最有效地保护隐私的方式。

图5.19双向监控，当偷窥者通过计算 机和网络刺探别人隐私时，被窥视者也 在看着他

保护隐私对个人的好处不言而喻， 对商家其实也有好处。这不仅在于它们 能够“合法地挣钱”，而且还能让好的商 家长期挣钱。为了理解这一点，我们不妨 看看美国银行发展的历史。在很长的时 间里，美国的银行业可以用“胡作非为” 来形容，挪用储户存款进行非法经营的 情况时有发生。当偷钱的银行工作人员 在非法经营中净到钱后，他们把利润装 进自己的口袋，相反，当他们亏了钱之后， 他们是不会从自己口袋里掏出钱还给储 户的，而是让银行破产，因此在1933年 罗斯福新政之前，美国银行破产是家常 便饭。1933年之后，美国一方面杜绝银 行进入股市等高风险的资本市场（法律 监管），并且提供了 一个技术手段来保护 储户利益，即FDIC(联邦存款保险公司) 的再保险；另一方面通过竞争让那些胡 作非为的银行纷纷倒闭，这才让储户能 放心地把钱放在银行里。类似地，如果两 家公司同样挣钱，一家有能力保护用户 隐私，另一家总是侵犯用户隐私,可以想 象，后者会逐渐丧失用户。 小结

大数据在今天这个时间点爆发，是 各种技术条件具备的结果。但是，要让大 数据真正发挥巨大作用，让计算机变得 更聪明’还有很多技术挑战需要应对。

大数据的数据量大、维度多、数据完 备等特点，使得它从收集开始，到存储和 处理，再到应用，都与过去的数据方法有 很大的不同。因此，使用好大数据也需要 在技术和工程上采用与过去不同的方 法，尤其是要改变我们过去的很多思维 定式。大数据和机器智能的发展和应用 过程，还会带来很多新的技术挑战，需要 解决很多技术上的难题，比如对数据安 全的考虑，对隐私保护的考虑等。有些问 题虽然在大数据之前并不重要，但是今 天在大数据时代它们变得非常突出而且 敏感，使得我们不得不认真考虑。

我们已经向大家展示了大数据能给 我们带来的诸多好处，但是这些好处的 获得需要有扎实的技术和工程基础做保 障。在今后,任何一个能够提供某些大数 据关键技术的公司和个人，在未来的智 能革命中，都将有大展宏图的机会。 注释

[1] 这是2015年的上传速度。

[2] 计算量的增加取决于算法的 复杂度。对于排序这样的计算，数据量增 加N倍，计算时间会增加NLogN倍;对于

矩阵运算，则可能增加N^2倍。

[3] 比如做一个公司或者APP, 直接卖数据或者把公司卖掉。

[4] 在美国，帮助接线的电话是 411。

[5] 1PB = 1024TB。一-编者

注

[6] GFS是一个可扩展的分布式 文件系统，用于大型的、分布式的、对大 量数据进行访问的应用。

[7] 在Google内部，Dreme丨原 先的项目代号是BigTable。

[8]很多机器学习的算法都不是 多项式复杂度的，因此算法专家致力于 将这些算法在特定应用中做一些近似和 简化。即便如此，这些简化后的算法也是 高阶多项式的，数据量增加一点点，复杂 度会增加很多。

[9]见参考文献（Quoc Le, 2012)。

[10]关于Google大脑的技术细 节，有兴趣的读者可以参看拙著《数学之 美》。

[11] http://techcrunch.com/ 2015/02/25 /target - says - credit -card-data-b「each-cost-it-162m-in-2013-14.

[12] http : / / www .

businessinsider . com / sonys -hacking - scandal - could - cost -thecompany -100 - million - 2014 -12.

[13] https://it.slashdot.org/ story/16/04/18/148252/mit-reveals -ai - platform - whichdetects - 85 -percent-of-cyberattacks.

[14] http://www.usatoday.com *story* money/cars/2013/03/24/car-spying-edr-dataprivacy/1991751 /.

第六章 未来智能化产业

现有产业+机器智能=新产业，未来 的农业、制造业、体育、医疗、律师，甚至编辑记者行业都将迎来崭新形 态，新产业将取代旧产业满足人类的 个性化需求，大数据将导致我们整个 社会的升级和变迁。

在过去的300多年里，人类所经历的 历次重大技术革命都沿袭这样的规律: “现有产业+新技术=新产业”，这也是 贯穿本书的主题。有些新产业是旧的产 业以新的形态出现，比如在互联网出现 后，广告业从过去的印刷广告和电视广 告逐渐转变为互联网广告；有些则是全 新的产业，比如电报和电话的产生催生 出电信业。在接下来的智能革命中，依然 将是现有产业的转变和新产业的诞生并 行。但是，无论是哪一种，它们都有共同 的特点，即智能化和精细化，因此我们不 妨将它们统称为“智能产业”。在这些产 业中，具有智能的计算机可以帮助我们 完成相当多的工作，甚至是绝大部分工 作。

在接下来的篇幅里，我们就通过一 些未来产业的形态，进一步理解智能革 命对产业和社会的影响。在这些产业中， 有些今天已经存在，虽然看似和机器智 能没有多大的关联，但是它们会受到智 能革命的影响而彻底改变。这些改变并 非我们的预测，而是已经发生和正在发 生的事实。让我们从最古老的行 业——农业一开始审视智能革命的影

未来的农业

农业是人类所从事的最古老的行 业，也是支撑人类文明的基础。根据斯坦 福大学教授兰■莫里斯(Lan Morris)的 观点，人类文明的水平可以用人均产生 的能量来衡量，比如像原始社会，人类产 生的能量是所消耗能量的2～3倍;那么到 发达的农业社会时，这个比值可能高达 10倍;到了工业革命之后，由于机械在农 业上的应用，每一个人能够耕种的土地 和收获的粮食大大增加，使得大量的人 口能够被释放出来从事工业和服务业的 劳动。但是，自然环境，比如土地的面积 和降雨量，依然是制约农业发展的瓶颈。

在过去，解决土地短缺问题的方法 就是施用化肥和农药增加单产，解决水

资源短缺问题的方法就是挖更多的井抽 水，挖更多的渠引水，但这实际上是将短 期矛盾转变为长期的危机。如果跳出定 式思维来考虑农业用水的问题，我们首 先要问：“种田是否需要那么多水，那么 多土地？”

2005年，Google一些好事者学着 以色列人的做法，在总部门前开辟了很 小的一片蔬菜种植园，试图重现以色列 人在过去几十年里在农业上取得的成 就。几年试验下来，证明以色列人的做法 是可以复制的。那么以色列人是怎么做 的呢？我们还得先看看以色列人的生存 环境。

图6.1 Google员工学习以色列的滴灌 技术种植的蔬菜瓜果

1990年我去中国西部出差，参观一 些治理沙漠的项目。当地人告诉我这样 一件事，他们听说以色列人能在干旱的 土地上实现农业高产，就请了一些以色 列的专家来指导农业。这些以色列人到 中国的大西北考察了自然条件之后说，

你们这里哪儿叫缺水，水比我们以色列 多多了。以色列的自然环境实在是太差， 绝大部分土地为沙漠，可耕种面积不到 国土面积的五分之一，而且土层是世所 罕见的贫瘠，更要命的是水资源严重匮 乏。在以色列境内只有一条约旦河(还要 和阿拉伯人共享水源)，以及一个小得微 不足道的淡水湖。以色列降雨极少，年降 水量约200毫米，占土地面积一大半的南 部内盖夫沙漠，每年平均降雨量仅有25～ 50毫米。这么少的降雨量是什么概念呢? 对比一下我们常说的缺水的大西北就知 道了。兰州年降雨量达325毫米，西宁 380毫米，乌鲁木齐200~800毫米不等， 都比以色列多很多。

然而，就是在这样一片生存条件恶 劣之地，以色列人创造了令人咂舌的奇

迹，许多农产品的单产量领先于世界先 进水平。他们的奶牛单产奶量居世界第 一，平均每头年产奶10500公斤，每只鸡 年均产蛋280个，棉花单产居世界之首， 亩产近1000斤（中国为228斤）[1]， 柑橘年均亩产多达3吨（中国为0.5吨）， 西红柿年均亩产20吨[2]。由于单产高， 以色列居然成为农产品出口大国，每年 向欧洲出口大量的蔬菜和水果，有“欧洲 的厨房”之称。不仅如此，这个干旱的沙 漠国家还成为仅次于荷兰的世界第二大 花卉供应国。2007年，以色列农业总产 值为55亿美元，其中，农业出口占40%， 达21.72亿美元，也就是说，以色列平均 一个国民贡献了世界上1.7个人的食物。 以色列取得这样的成就，其根本原因是 靠科技兴农，而不是靠破坏生态环境，竭 泽而渔。至于以色列人如何通过科技手

段提高单产，我们暂不讨论，这里我们不 妨看看以色列人如何在农业中节省水资

源。

图6.2以色列将荒漠改造成良田和牧 场

作为严重缺水的国度，以色列人发 明了滴灌技术——装有滴头的管线直接 将水和肥料送达植物的根系，大大节约 了水和肥料。所有灌溉方式都采用计算

机进行自动化控制，灌溉系统中有传感 器，能通过检测植物茎果的直径变化和 地下湿度，来决定对植物的灌溉量，这样 可以节省人力和水资源。由于有大量的 传感器在采集数据，这种自动滴灌系统 可以对用水量和产量的关系进行学习, 改进灌溉量。自“二战”后立国以来，以 色列的农业生产增长了十多倍，而每亩 地的用水量仍保持不变。靠着农业高科 技，以色列给传统的农业带来了质的革 命，“二战”前是一片荒漠的内盖夫地区 (以色列所在地），现在已经出现大片绿 洲了。

如果农田可以靠精确的灌溉，那么 草地、花园、院落等凡是需要用水灌溉的 地方是否也能够采用类似的方法大幅度 节省用水呢?答案是肯定的。2013年7月

的《时代》周刊报道了硅谷一家小公司发 明的Drople像庭院落自动喷水机器人。 这种机器人从外观上和行走的方式上看 与目前很多家庭使用的扫地机器人有点 像，但是智能水平要高很多。Droplet的 喷水机器人首先会对每个家庭的院落扫 描一遍，看看院子里有多少植物和草坪 需要浇灌，同时它还测试各处土地的湿 度和植物的高度，以决定喷水量。在浇水 时，Droplet会根据事先计算出的喷水量 拖着水管子走到相应的位置，调整好喷 水的角度、流量和时间开始浇水，并且根 据事先规划好的路径完成整个院落的浇 灌，而不会漏掉任何一处植被。在使用的 过程中，Droplet可以根据湿度调整水量, 并且和天气预报相连，如果明后天会下 雨，那么Droplet会停止浇灌。根据《时 代》周刊的报道，一些家庭使用Droplet

之后，可以节省95%以上的浇水量。在 2015年加州最干旱的季节，很多小区为 了节水，由物业补贴钱让住户购买这种 喷水机器人。

图6.3自动浇水的机器人Droplet

在引入机器智能之后，农业这个人 类最古老的产业将会以崭新的形态出 现，它将验证“现有产业+机器智能=新 产业”这样一个已被证明的技术革命进 步的规律。

未来的体育

在2015~2016年的NBA(美国职业 篮球联赛)赛季，位于硅谷地区的金州勇 士队（Golden State Warriors)创造了 NBA历史上常规赛获胜率最高的纪录， 在全部82场比赛中获胜73场[3]，同时 它还创下主场54连胜的纪录。在一年前， 该队获得了 NBA总冠军。读到这里,一般 人会觉得金州勇士队应该是一个老牌强 队，同时拥有很多大牌球星加上一个金 牌教练，否则难以创下这样的纪录。但事 实并非如此，勇士队长期以来一直是 NBAM的一支“鱼腩球队”。在2009年， 金州勇士队还是NBA里最烂的球队之 _，SP—年它的成绩排名倒数第二，当然 勇士队也不可能有什么球星和大牌教

练。因此该队能取得这样的成绩,实在是 一个奇迹，而它创造奇迹的方式在体育 史上恐怕是独一无二的。

一般来讲，一个弱队的崛起常常是 因为有一个大老板喜欢这个体育项目， 买下全部或者部分球队，然后砸钱买球 星和请大牌教练，再做各种广告招揽球 迷。中国恒大足球队走的就是这条路，恒 大集团在里面投资最高，使它的估值居 然高达33.5亿美元，甚至超过了皇家马 德里队。当然，砸钱容易，取得成绩却并 非花钱就能做到，因此弱队崛起通常并 非易事。金州勇士队的成功并非砸钱的 结果，而是因为它处在一个特别的地 区-硅谷。

硅谷地区有两种人最不缺，即风险

投资人和工程师，勇士队的奇迹从很大 程度上讲是靠他们创造的，前者善于看 到其他人还没有发现的投资潜力，然后 把它经营成值钱的实业；后者善于利用 技术创造奇迹。勇士队的成功就是他们 合作的结果。6年前勇士队的比赛成绩跌 到了谷底，因此价值较低，一些风险投资 人决定将这支不值钱的球队买下来好好 经营，让它成为美国体育界最耀眼的明 星。这个计划看上去有点疯狂，不过投资 人有自己的考虑，他们有秘密武器，那就 是能够应用大数据的工程师。最终，投资 人花了 4.5亿美元这个相对较低的价格 完成了对勇士队的收购。

在收购完成后，投资人为球队委派 了新的管理层，在管理层的背后,有一些 工程师在利用大数据制定球队的发展战

略和比赛战术。新的管理层在上任后所 做的第一件事，不是购买大牌球星，反倒 是把队伍中的明星给卖掉了，然后他们 围绕一位当时毫无名气的球员重新制定 球队的风格和战术，当然管理层的决策 依据是从大数据中得到的结论。

根据数据分析的结果，管理层认为 现在NBA以及很多职业联赛所追求的打 法是低效率甚至是错误的。几十年来， NBA的发展一直在追求制空权，球队寻 找个人身体条件突出的球员们，他们要 么伸手就能将篮球装进球筐(比如姚明), 要么能高高跃起从上往下把篮球扣进球 筐（比如乔丹）。这样的打法虽然看起来 漂亮，但是效率很低，因为需要全队费很 大力气攻到篮下，把球传给那个大高个 儿，即便不出现传球失误，也就是得2分,

扣篮也是如此，在耗费巨大的体力之后， 也是得2分。勇士队的管理层设计的新打 法却是尽可能地从24英尺（大约7.3米） 外的三分线投篮，这样可以得3分。正是 因为不再按照篮球传统的战术作战，勇 士队才卖掉了那些价钱高却效率低的明 星，而着重培养自己看中的新人。

这位新人叫斯蒂芬•库里 (Stephen Curry),今天他在美国已经 家喻户晓，中国的篮球迷对他也非常熟 悉，但当年他可是一个没有人要的球员。 库里身高只有1.91米，在篮球场上和那 些明星大腕相比可谓相形见绌，高中毕 业时，那些篮球强校的教练都看不上他。 2009年他被勇士队以很便宜的价格签 约（4年只有1270万美元，而姚明登陆 NBA第一年的薪酬就高达1250万美

元），尽管他在大学篮球队表现不错，但 那并不是一支顶级的大学篮球队，因此 他的对手们也没有把他放在眼里。

勇士队的管理层之所以要重用库 里，是因为他有一个特长，S隙t是投篮准 确，勇士队最终把他培养成了一位三分 球的神投手。在2014 *2015年赛季中, 库里的神投让勇士队夺得了 40多年来的 第一个总冠军，他自己也成为当年的最 有价值球员（MVP)。到了 2015* 2016 赛季，库里投进了403个三分球，创造了 NBA历史上的纪录，打破了由雷•阿伦 所保持的个人单赛季269记三分命中数 的纪录。库里投篮的准确率高达50%，三 分球的命中率也高达45%，这意味着他 的三分球比那些大牌球星的篮下投球更 准。到后来，很多球迷跑去看勇士队训练，

主要就是为了欣赏库里投三分球。当然， 库里成名之后，在赛场上各个球队都要 派人盯紧他，但这也给勇士队其他选手 在篮下创造了机会。

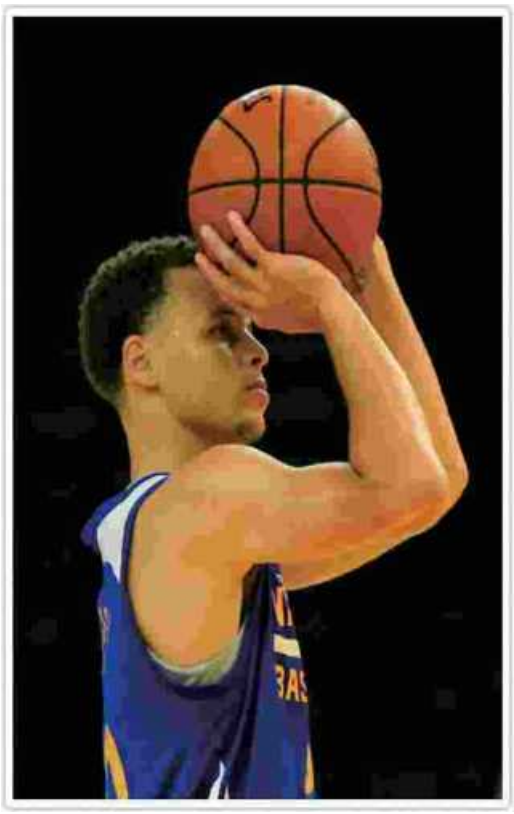


图6.4勇士队神投手库里

按照一般教练的想法，勇士队应该 趁机加强内线进攻才对，但教练史蒂 夫■科尔却不这么看。2014年科尔执教 勇士队时，没有任何执教NBA的经验，但 勇士队的老板乔•拉格布(Joe Lacob) 坚持使用这位新教练。拉格布是个篮球 迷，却并不是篮球界的人士，他是著名风 险投资公司凯鹏华盈的合伙人，套用俗 话讲，他是一位“技术控”。他的合伙人 很多甚至就是工程师出身，比如 YouTube的联合创始人查德•赫利 (Chad Hurley)。因此他们更相信自己 根据数据得到的结论，而不是来自NBA 的经验。拉格布看中科尔的是他在NBA 生涯中准确的投篮，后者作为和乔丹同 时代的公牛队队员，夺得过5次总冠军,

个人的投篮命中率高达45.4%，位列当 时NBA球员之首。科尔在执掌勇士队之 后，坚持用数据说话，而不是凭经验，他 根据背后团队对历年来NBA比赛的统 计，发现最有效的进攻是眼花缭乱的传 球和准确的投篮，而不是彰显个人能力 的突破和扣篮。在这个思想的指导下，勇 士队队员苦练神投技，全队在一个赛季 中投进1000个三分球，又创造了一项 NBA纪录。同时，就在对手防守库里时, 勇士队的第二投手汤普森大展神威，在 一个赛季投进了270个以上三分球，成为 第二个跨越之前历史纪录的篮球选手。

除了利用数据制定战略，勇士队还 利用实时数据及时调整比赛中的战术。 早在2012年，勇士队的总裁兼COO (首 席运营官）里克.威尔茨（Rick Weltz)

就在一次大数据会议（TUCON 2012) 上介绍了该球队应用大数据的成果。根 据威尔茨的介绍，大数据可以帮助球队 改进精细到两个人配合的细节。正是靠 高科技，勇士队才得以在短短6年里从倒 数第二名登顶NBA的总冠军。

鉴于勇士队的战术和成绩给NBA带 来的巨大冲击，奥巴马在白宫专门接见 了勇士队，并且讲道：“（这）看起来正 在打破这项运动的格局，这似乎是不公 平的比赛。”篮球界的人士则认为，勇士 队是 NBAM 的 Google。

利用数据提高球队整体运动成绩的 想法并非今天才有，20世纪80年代，与 中国女排同时崛起的还有美国女排。与 中国队主要靠技战术水平和女排精神所

不同的是，美国女排的秘密武器是高速 摄像机和统计。也是美国女排的运气不 好，遇上了在巅峰状态的中国队，接连几 次世界大赛都和冠军失之交臂。过去由 于数据量有限，统计作用不是非常明显, 因此在体育中利用数据指导训练的情况 并不普遍，但是在未来，大数据一定会改 变体育界的格局。

大数据对体育训练的帮助还在于分 析和总结优秀选手的动作姿势，纠正其 他运动员的动作。今天高尔夫球运动员 和网球运动员会在身上安装各种传感 器，测定动作，然后和优秀选手的动作比 对，纠正自己的动作。

机器智能对体育的帮助，还体现在 计算机可以训练棋牌选手。今天，很多国

际象棋学校在训练小棋手时，使用的是 计算机而不是真人教练。近年来，计算机 也开始训练围棋选手了。

可以预见，未来的竞技体育是离不 开大数据和机器智能的。体育依然会是 人类最喜爱的娱乐活动，但是仅靠天赋 和苦练将不足以取得最好的成绩。

未来的制造业

2011年德国提出工业4.0的概念， 即通过数字化和智能化来提升制造业的 水平。相应地，中国也提出了中国制造 2025的概念，其核心是通过智能机器、 大数据分析来帮助工人甚至取代工人， 实现制造业的全面智能化。在美国，特斯 拉汽车公司已经尝试全部使用机器人来 装配汽车，这不仅使得工厂雇用工人的 数量大幅度减少，而且还让出厂的汽车 性能和质量更稳定。

曾几何时，产业工人的数量被看成 是制造业竞争力的重要标志，大量低工 资的生产线上的工人造就了全球制造业 的繁荣。被称为“世界工厂”的中国在改 革开放以后正是靠这一项核心竞争力跻

身世界制造业大国行列。在中国，全球最 大的OEM制造商富士康雇用了 130万 名廉价的工人，使得全球的电子产品制 造商无法在成本上和它竞争。当然，富士 康也得到了“血汗工厂”的恶名。由于雇 用的工人太多，像富士康这样的公司即 便有心将自己办成高福利的企业，也是 做不到的。另一方面，它也不可能通过进 一步压榨工人来降低制造成本。为了解 决这些矛盾，富士康一直在研制取代生 产线工人的工业机器人。富士康预计未 来将装备上百万台机器人，逐渐取代装 配工人。这使得工人们不再需要从事繁 重而重复性的工作，但由于工厂所需要 的工人数量大幅度减少，很多低技能的 工人将失去工作。

美国“二战”后的汽车行业有上百万

装配工人，但是现在只剩下当年的一个 零头。而新的汽车公司比如特斯拉，已开 始尽可能地使用机器人取代装配工人。 硅谷东部的弗利芒特市(Fremont)有特 斯拉最大的汽车装配厂。在该厂的门口 每天都有几个人举着骷髅抗议，停下来 一问，才知道特斯拉根本不从汽车工会 招装配工人，甚至很少招生产线上的工 人，因此汽车工会天天跑去抗议。

图6.5汽车工会每天在特斯拉公司门

口示威，抗议该公司不招汽车工会的工 人

抗议归抗议，特斯拉就是不雇生产 线上的工人，外界也拿它没有办法。事实 上，在过去的5年里，特斯拉员工数量增 长非常快，不过它所雇的都是IT人员，以 至于它更像是一个IT公司而非汽车公 司。那么大家可能有一个疑问，特斯拉的 汽车是怎么制造出来的呢？答案很简单， 尽可能地使用机器人。

图6.6特斯拉的汽车装配厂，全部由机 器人操作

机器人取代人类从事制造业的另一 个巨大优势在于，产品很容易按照个性 化定制。在大工业时代，机器所解决的是 确定性问题，因此，_旦_个产品设计出 来，它就是确定的,按照事先确定的设计 复制，成本是很低的。但是，如果哪个顾

客想要根据自己的需求订购一款特定的 产品，那么成本是很高的。而在机器人取 代生产线上的装配工人的智能制造时 代，只要通过设定产品参数，机器人就可 以根据用户需求制造出个性化的产品， 其成本不会比大规模生产高多少。

特斯拉很少雇用原来汽车行业的人 员，除了降低成本外，还有一个更深层次 的原因一_它一直把自己定位成一个IT 公司，而不是汽车公司。汽车其实就是承 载着特斯拉IT技术的平台，特斯拉内部 将汽车看成是一个巨大的智能终端，通 过这个智能终

端，特斯拉把它的各种技 术服务提供给大家，同时也参与到消费 者的曰常生活中，这和我们在前面提到 的小米手机有不少相似之处。

特斯拉颠覆现有汽车行业所做的另 一件事，就是取消存在了一个世纪的汽 车代理商制度。为什么特斯拉能够做到 这一点，而比它更大的、更有话语权的那 些大牌汽车公司却不得不分利给各地的 代理商呢？这就要从产品生产和流通的 产业链说起。

产品生产本身只是商品经济中几个 主要环节中的一个。除了生产，商品的设 计和研发、仓储和物资管理、物流和运 输、fft发和零售，在过去都是不可或缺的 环节。我在《浪潮之巅》中介绍过戴尔的 商业模式，它的成功在于一方面出让了 最需要人力的生产环节，以降低成本，另 一方面依然牢牢把控着其他重要的环 节，以保证利润。过去，在生产以外的环 节，要么需要所谓知识型的员工来完成，

要么需要本地的员工。比如汽车的销售 在过去依靠的就是本地员工，如果由汽 车厂直接在销售地雇人，成本会比交给 代理商更高。但是到了大数据时代，除了 商品的设计和研发，剩下的环节要么高 度智能化（比如仓储和物资管理），要么 干脆被砍掉（比如批发行业），因此在制 造业中那些所谓高端的工作也面临着被 机器智能所取代。比如阿里巴巴的崛起， 就让很多批发行业的工作从此消失了, 当然，同时也带来了全社会效率的提升。

戴尔公司从早期直到2004年的成 功原因在于，它率先采用智能化的管理 降低了各个环节的成本。但是，当联想等 很多企业也采用类似的管理方式时，靠 低成本竞争的戴尔就不再具有优势了。 特斯拉则比戴尔更进了一步，它除了大

量雇人研发汽车的各种新功能外，还从 设计开始，直到汽车送到顾客手上，加上 售后服务，这中间各个环节里尽可能地 采用计算机而不是人来工作。因此，特斯 拉才能够做到所有事情都由自己来做， 因为计算机帮了它的忙。

特斯拉其实在悄无声息地重新定义 汽车行业，它对汽车的理解已经和当年 的福特或者奔驰完全不同了。汽车这个 老行业，在引入大数据和机器智能之后 就脱胎换骨，变成了一个新的行业。

特斯拉只是未来制造业一个典型的 案例，其他商品的制造和流通也可以得 益于大数据和机器智能。当机器智能逐 步渗入到商品制造和销售的各个环节 时，不仅工人的数量将逐渐减少,而且整

个制造业都将被重新洗牌。仅仅靠降低 工人工资的低水平竞争将不再具有制造 业方面的优势，因为它在未来的竞争要 靠从设计到销售全过程的智能化水平。 当然，在我们欢呼整个制造业效率提升、 产品质量提升的同时，有一个问题值得 关注，那就是被机器智能取代的劳动力 如何安排，这个问题我们放到下一章讨 论。

未来的医疗

医疗保健在任何发达国家都是一个 大产业，甚至是最大的产业，因为人类发 展经济和科技最重要的目的就是增进健 康、延年益寿。在历史上，历次重大科技 进步都伴随着人类医疗保健水平的飞 跃。在工业革命之后，人类搞清楚了细菌 致病的原理，并且通过科学的方法完成 了传统医学到现代医学的转变。在第二 次工业革命之后，人类发明了抗生素，我 们在前面讲到，抗生素的发明过程是自 觉应用机械思维的结果。“二战”后，随 着信息革命的开展，各种诊断仪器和治 疗仪器被发明出来，包括今天常用的CT (计算机体层摄影)扫描仪、核磁共振机、 心脏起搏器和进行各种微创手术的仪

器。毫无疑问，大数据和机器智能也将对 未来的医疗产生全面而重大的影响。

今天，人类在医疗保健上遇到了一 些瓶颈，主要体现在以下几方面:首先是 医疗的成本越来越高。以美国为例，今天 医疗保健的开销已经占到GDP的17%~ 18%左右[4],而且按照目前的发展趋 势，到2020年，这个比例将上升到20%。 在中国，虽然这个比例很难准确估计，因 为很多与医疗保健有关的花销是隐形 的，但是“看不起病”是社会的共识。其 次，医疗资源不平衡，这一点几乎每一个 中国老百姓都认同。在医疗发达的美国， 这个问题同样存在，拥有约翰•霍普金 斯医院[5]、海军总医院[6]、协和医 院（Union Memorial Hospital)和国家 医学院的马里兰州，人均医疗资源是全

美国平均水平的3倍。由于医院集中，像 协和医院这样历史悠久的大医院居然会 为病人的数量发愁。在全世界范围内，医 疗资源不平衡的问题更加严重。最后，也 是最关键的，很多疾病治不好，比如癌症、 帕金森综合征和阿尔茨海默症（即人们 常说的“老年痴呆”）。尽管全世界医生 和科学家们已经努力了许多年，世界各 国也投入了大量的资金来寻找上述疾病 的治疗方法，但是在过去的20多年里，医 学在这些领域的进展十分缓慢。我们不 妨从这三个方面来看看大数据和机器智 能将如何改变全世界医疗保健以及制药 行业的现状。

降低医疗成本

先看看医疗成本的问题。美国医疗

系统有一个制度上的缺陷，就是医疗事 故赔偿过高，律师拿钱太多。普华永道估 计这笔开销占了全部医疗保健的10%，[7]即上干亿美元，最保守的估计也有550 亿美元(2011年的估计，以2008年美元 不变价计算）[8]，平均每个美国人每 年要负担170美元。另外，医疗保险系统 和医院的管理费偏高，加上相当一部分 没有保险的患者赖账[9],医疗系统就 把这部分钱加到了有能力支付的患者头 上。当然，这些都不是技术问题，不在我 们讨论之列，我们重点谈谈利用技术手 段降低医疗成本的问题。

从医疗本身讲，医疗成本高的前两 个重要原因是药品的研制周期太长、费 用太高，以及医务人员培养的成本太高。

让我们先来看看美国新药研制的费 用问题。斯坦福大学医学院院长米纳教 授说：“今天一款新药从关于它的第一批 最重要的论文发表，到药品上市,大约需 要20年的时间，在这个过程中全部的科 研投入至少为20亿美元。”根据美国专利 法，专利保护的期限从申请之日算起只 有20年(如果申请后三年还没有被批准， 则按照批准之曰算起17年)。但是专利的 申请并非是药品上市那一天才开始的， 通常要早于药品上市十几年，也就是说 药品上市后，受到专利保护的年限只有 几年。据强生公司介绍，主要的处方药在 上市后，能够享受专利保护的时间只有7 年。也就是说，新药即使能够顺利研究出 来，也只有7年时间的独家销售以挣回成 本。因此，每一种特效的新药都卖得非常

虫

造成美国医疗成本非常高的第二个 原因是医务人员的收费很高。在欧美等 发达国家，医生可以说是“三高”的职 业——高学历、高收入和高地位，而在医 生中间，专科医生，比如诊断癌症的放射 科医生或者做手术的胸外科医生、脑外 科医生，又是医生群体中收入最高的群 体，他们的平均收入远远高于上市公司 高管的平均水平。那么这些人的收入具 体有多高呢？ 2014年我与斯坦福大学医 学院的几位教授和医学博士们聊到专科 医师收入的问题，他们以放射科医师为 例来形容专科医师的生活和收入。“当你 获得放射科医生行医执照并且得到第一 份工作时，你的高中同学的孩子都上小 学了，而且他们也都事业有成了。但是， 你可以很自豪地告诉他们：‘我才拿到第 一份工作，不过年薪是50万美元！’”虽

然50万美元的年薪并非所有的专科医师 都能拿到，但是相当一部分专科医生的 年收入就是这个水平，甚至更高。年薪50 万美元是什么概念，这大约相当于美国 中位数工资的10倍[10]，比美国总统 高 1/4 [11]。

如果专科医师们挣的那么多，他们 的收费一定更高。比如与放射科有关的 医学影像分析这个行业，2014年的花费 就高达330亿美元左右[12]，摊到每一 个美国人头上居然高达每年110美元，不 论你这一年是否做过任何透视、CT或者 核磁共振的检查。更可怕的是，这个花费 年增长率为7%，远远高于GDP的增长 水平。

为什么在美国专科医师收费要那么

高，主要原因是成为专科医师太难，这个 群体人数太少。要成为_名合格的专科 医师，除了要智力水平高，还要经过长时 间系统的训练，并且花费很多的学费和 培养费。具体讲，在美国培养一名合格的 专科医生的过程大体如下：

首先，他们要完成4年大学本科学 习，因为在美国只有获得本科学位之后 才能够学医。在本科毕业后，S 卩些所谓的 医学预科生(p $_{re}$ Med)要经过激烈的竞 争才能进入医学院，好的医学院的录取 率要远比哈佛大学低[13]。在医学院里， 这些幸运的未来的医学博士要接受4～5 年的医科学习，医学院的学习负担要比 一般的研究生专业重得多。在完成医学 院学习之后，如果运气好的话，经过2年 左右的医院实习（实习医生)和2～3年的

专科实习（Fellow)，才能获得专科的行 医执照。整个过程平均要花费13年之久, 中间还会有很多次被淘汰的可能。实际 上，高中毕业时想成为专科医生的人并 不算少，但是真正获得行医执照的少之 又少。

其次，成为专科医师的学习费用也 是相当高的，因为读医学博士的人是没 有奖学金的，如果再考虑到读本科时也 要自己掏钱，那么一个成绩优秀的学生 从本科算起，到医学院毕业，大概需要花 费50-70万美元。在欧美国家，大多数 人又不愿意啃老，因此，每一个专科医师 在能够开始挣钱时都已经负债累累。从 投资回报的角度讲，既然时间和金钱的 投入都如此巨大，他们必须有高收入才 合算。

美国医疗系统的这些症结不是简单 要求医生、医院和药厂少收费就能做到 的，这也是奥巴马医保计划在美国难以 推行的原因。

在过去，像放射科医生这一类工作， 被认为需要太多的专业技能，工作性质 太复杂，不可能被机器所取代。但是，今 天智能的模式识别软件通过医学影像的 识别和分析，可以比有经验的放射科医 生更好地诊断病情，这将从根本上改变 医疗行业的现状。

科学家和医生们通过模式识别和图 像理解进行医学影像分析的想法其实不 是在有了大数据之后才开始的。早在20 世纪70年代图像处理开始起步时，人们 就想到了它在医学上的应用。但是真正

取得突破性进展，并且能够做到比人做 得更好，则是近几年的事情，因为计算机 有了大量的数据可以进行学习。

在中国很多患者的心目中，看病要 找“老大夫”，因为他们有经验。实际上， 老大夫经验的积累就是一个通过病例 (数据）学习的过程，而人学习再快，也 学不过计算机，这一点我们在前面分析 Google的AlphaGo和李世石下棋的案 例中已经指出了。一个放射科大夫一生 阅读研究的病例很难超过10万个，而计 算机则很容易从上百万病例中学习。 2012年Google科学比赛的第一名授予 了一位来自威斯康星的高中生，她通过 对760万个乳腺癌患者的样本数据的机 器学习，设计了一种确定乳腺癌癌细胞 位置的算法，来帮助医生对病人进行活

检，其位置预测的准确率高达96%,超过 目前专科医生的水平。这位年轻学生采 用的图像处理和机器学习算法都不复 杂，她的成功完全得益于大数据,没有哪 个大夫一生能够见识760万个病例。

在医学影像分析方面，很多软件已 经开始商用化，只是由于目前在临床诊 断上需要有真人签署检验报告，因此这 些软件给出的结果还需要由人来核实后 签字。即便如此，由于放射科医生的工作 效率能大大提高，诊断的费用可以逐步 降低。

图6.7手术机器人达■芬奇的手术台

具有了智能的计算机不仅能帮助诊 断，承担放射科医生的工作，还可以进行 手术。今天,世界上最有代表性的做手术 的机器人就是达•芬奇手术系统。 达•芬奇手术系统分为两部分：手术室 的手术台和医生可以在远程控制的终 端。手术台是一个有三个机械手臂的机 器人，它负责对病人进行手术，每一个机

械手臂的灵活性都远远超过人，而且带 有摄像机可以进入人体内手术，因此不 仅手术的创口非常小，而且能够实施一 些人类医生很难完成的手术。在控制终 端上，计算机可以通过几台摄像机拍摄 的二维图像还原出人体内的高清晰度的 三维图像，以便监控整个手术过程。医生 也可以在远程对手术的过程进行人工干 预。达•芬奇手术系统的主要发明人之 一，约翰•霍普金斯大学的拉塞尔•泰 勒（Russell Taylor)教授是我的朋友和 师长，因此我有幸亲身体验操作该机器 人。他为我在手术台上设置的是一个仿 制的人脑，我在远程用手术刀虚拟切割 时，手的感觉和切割真实的组织是一样 的。目前全世界共装配了3000多台 达•芬奇机器人，完成了300万例手术。

相比医生，计算机在诊断和做手术 等方面有三大优势:首先，它们漏判（或 者失误)的可能性非常低，也就是说它们 能够发现一些医生们忽略的情况；其次， 它们的准确率很高，而且随着数据量(病 例）的增加提高得非常快;最后，也是人 所不具备的，这些智能程序的稳定性非 常好,它们不会像人那样受情绪的影响。 而这些智能程序的成本，通常不到人工 的百分之一。

解决医疗资源短缺问题

如果说机器智能通过帮助放射科医 师和外科医生可以降低医疗成本的话, 那么它在解决医疗资源不足的问题上同 样有效。自然语言处理专家和医生们让 计算机理解人的语言，然后让它能够根

据化验结果和病人的描述来诊断简单的 疾病。IBM公司从20世纪70年开始就致 力于机器智能的研究，并且在工业界_ 直处于领先地位。旧M开发的沃特森 (Watson)智能系统可以理解自然语言， 分析各种数据和医学影像，帮助疾病诊 断和医疗信息的

管理。在一些医学领域， 比如肿瘤科，它能够非常准确地给医生 提供诊断的建议和帮助。目前，如果不引 入医师的干预，仅仅靠计算机通过阅读 病例、倾听病人的描述和分析化验结果 进行疾病诊断，它也能达到中等医生水 平。虽然这样的水平远没有达到取代医 生的程度，而且在医疗资源较多的大城 市里必要性不大，但是在缺少医生的非 洲和印度，有这种“机器医生，’总比没有 强。何况，考虑到医疗数据增长很快[14]，计算机学习能力又很强，这一类系统

会进步非常快，可以预见在不久的将来， 计算机在一些疾病的诊断方面会超过 人。

图6.8能够帮助看病的旧M沃特森计 算机

制药业的革命

2013年Google宣布成立独资的IT 医疗公司Calico,并且聘请了世界知名

的生物系统专家阿瑟■李文森博士担任 CEO。李文森博士曾经是世界上最大的 生物制药公司基因泰克[15]的CEO, 在接受Google 任命时，他依然担任着基 因泰克的董事会主席以及当时全球市值 最高的公司一一苹果公司的董事会主 席，可谓整个工业界最有权势的人物之 _。在一些人看来，李文森接受Google 的邀请担任其一个子公司的CEO有点 屈尊了，但是他自己认为他有可能开创 —个改变人类命运的事业，因为他将利 用大数据和其他IT技术设法延长人类的 寿命。

李文森用了一个大家熟知的例 子一医治癌症，来说明大数据在未来 医疗卫生中将扮演什么角色。

治愈癌症是人类半个多世纪以来的 梦想。在20世纪50年代，著名的工程师、 被誉为晶体管之父的皮尔斯（John Robinson Pierce, 1910—2002)把治 愈癌症和登月、识别语音、水变油、海水 里提炼黄金并列为人类难以解决的5个 难题。1969年，人类实现了登月，从20 世纪70年代开始，计算机语音识别也取 得了长足的进步，今天这个问题被认为 已经解决了，但是攻克癌症还显得遥遥 无期，尽管对特定的人来说，_些癌症是 可以控制的。

人类在抗癌研究方面投入的资金比 阿波罗登月或者语音识别要多得多，但 为什么至今依然难以根治癌症呢？李文 森博士讲，世界上并不存在一种一劳永 逸的万灵药，能够像青霉素杀死细菌那

样杀死所有的癌细胞，这是今天医学界 普遍的认知，与半个世纪前大不相同。我 们知道，癌细胞是动物和人自身细胞在 复制的过程中基因出了错，而非来自体 外，因此它们与人和动物正常的细胞非 常相似。今天最有效的方法是，使用基因 技术研制出的抗癌药来治疗，从机理上 讲是找到病变的基因并且把相应的癌细 胞杀死。不过，由于不同人即使得了同一 种癌，其癌细胞病变的基因未必相同，因 此一种抗癌药可能对某些病人管用，但 是对其他病人并不管用。我们通常听到 的发生在身边的故事就是这样。实际上, 大部分医生在给癌症患者用药时，需要 对患者进行基因比对，以确定是否能用 某种抗癌药。

医治癌症第二个难点，也是最根本

的难点在于癌细胞本身的复制也会出 错。这一点其实并不难理解，因为基因在 复制的过程中出了一次错误就可能出第 二次。这样一来，原本管用的抗癌药就变 得不管用了。抗癌药在杀死癌细胞时，未 必能够把所有的都杀死[16]，剩下哪怕 只有一个癌细胞未被杀死，它依然可以 迅速繁殖，并且可能出现新的基因突变。 我们通常会听到这一类故事：某个患有 癌症的亲友已经将病情控制了很长时 间，突然一夜之间复发，而且药物不起作 用，很快便离世了。这里面的原因就是癌 细胞基因的变化使得原有的抗癌药不灵 了。

由于癌细胞基因的突变和人有关， 而且可能一再突变，因此要想彻底解决 问题，就需要针对不同的患者设计特定

的抗癌药，而且要根据患者癌细胞每一 次新的变化研制新药。李文森博士认为， 只要这个研制速度能够赶得上癌细胞的 变化，那么，即使不能彻底杀死所有的癌 细胞，患者仍可以长期和癌症共存。从理 论上讲，这种方法是可行的。但是这样做 的成本太高：首先要有一个专门的研发 团队围绕着每一个患者进行药品的研 制，而且研发的速度还要足够快；其次， 它的耗费至少在每人10亿美元以上。因 此，全世界除了个别的亿万富翁，都不可 能用这种方法来治疗癌症。这就是目前 人类在抗癌方面遇到的困境，这个困境 是无法通过传统的医学进步走出来的。 事实上，在过去的20多年甚至更长的时 间里，全世界医学界对癌症机理的理解 和治疗方式的改进都是非常有限的。

那么出路在哪里呢？李文森博士认 为这要依靠最新的IT技术，尤其是大数 据。根据基因泰克的科学家解释,我们已 知的各种可能导致肿瘤的基因错误不过 在万这个数量级，而已知的癌症不过在 百这个数量级。也就是说,即使考虑到所 有可能的恶性基因复制错误和各种癌症 的组合，不过是几百万到上千万种，这个 数量级在IT领域是非常小的，但是在医 学领域则近乎无穷大。如果能利用大数 据技术，在这不超过几千万种组合中找 到各种真正导致癌变的组合，并且对这 样每一种组合都找到相应的药

物，那么 对于所有人可能的病变都能够治疗。针 对不同人的不同病变，只要从药品库中 选一种药即可，比如对患者约翰，他原本 是使用第1203号药品，如果发生新的病 变，经过检查确认后，改用256号药品即

可，这样并不需要每一次重新研制药品。 如此一来，便可以控制癌症了。虽然这样 成千上万种药总的研发成本不低，但是 如果摊到全世界每一个癌症患者身上， 李文森博士估计只需要人均5000美元 左右。

李文森博士所倡导的为每一个患者 设计个性化特效药的思路，如今已被制 药行业和医学界普遍认可。在美国著名 的加州大学旧金山分校医学院里，阿图 尔•巴特（Atul Butte)教授建立起医学 大数据中心，专门从事利用大数据寻找 个性化药品的研究工作。根据该中心的 陈斌副教授介绍，美国只有1/7左右的临 床证明有效的药品最终能够走完FDA (食品药品监督管理局)全部审批流程并 最终上市。剩下来的6/7的药品，虽然在

小范围内使用时对一些病人确实有很好 的疗效，但是在使用到大量患者身上时， 平均的效果并不显著，因此最终被FDA 否决。该中心通过研究发现，其中不少药 其实对特定的人群有效，现在的关键是 找到那些特定的人群，让那些研制过程 中被淘汰的所谓“废药”经过改造后能够 重新被利用。在未来，可能一种疾病会有 不同的药品医治，而不同的人会有不同 的特效药。

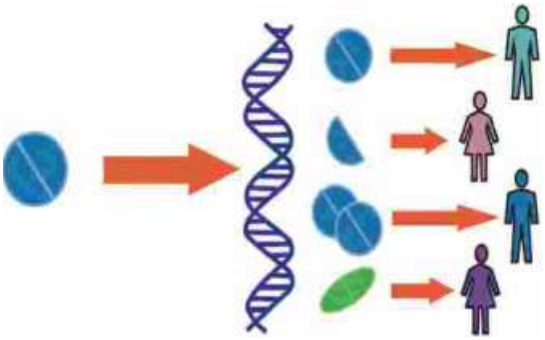


图6.9个性化药品

人类是否可以长生不老

除了个性化制药，李文森博士认为 大数据可以帮助治疗那些用传统医学方 法难以医治的疾病，而这个意义甚至比 治愈癌症更大。根据李文森博士的数据， 人类即使能够解决治疗癌症的难题，也

不过将平均寿命延长3.5年左右[17]。 在他和Google创始人佩奇看来,治疗癌 症的意义远没有大众想象的大，而人类 长寿面临的最大挑战是衰老问题——只 要人们活得足够长（而且不患癌症），最 后的结局都会是阿尔茨海默症，无一例 外。可以想象，当人类的平均寿命延长到 90岁以上后，所见之处是成群的阿尔茨 海默症患者。据麻省理工学院理学院院 长迈克尔•斯普瑟(Michael Sipser)博 士介绍，在过去的10年里，美国癌症、艾 滋病、心脏病和中风的死亡率都在下降, 下降的幅度在20%-40%左右，但是阿 尔茨海默症导致的死亡率却上升了 40 %。在李文森博士看来，延长人的寿命关 键是要找到衰老基因。至于怎么找，则需 要使用大数据，而Google的特长是善于 处理大数据，因此这便促成了李文森博

士和Google共同创建大数据医疗保健 公司Calico_事。

The Iran Opportunity SVE-Cigarettes/$20K Homes

TIME

CAN I

Goum

SOLVE ○

DEATH?

ftie search giant Is launching a venture to ©(tend the human life span.

That would be crazy-lf it weren't Google

色 t%ry.McCraal^i Bnrt L» fin班nar

图6.10《时代》周刊的封面文章 《Google能否战胜死神》

媒体对Calico给予了厚望，《时代》 周刊登载了题为《Google能否战胜死 神》的封面文章，它与其说是揭示 Google的野心，不如说寄托了大众对新 的医学研究的期望。从意识到死亡以来， 人类一直想找出终止走向死亡过程的方 法。从哲学的层面看，有生就必定有死， 长生不老是妄想。但是找到导致衰老的 基因，同时修复我们身体细胞在复制时 出错的基因，或许是一条人类延年益寿 的有效途径。

当然，Google也明白，光靠自己一 家的力量是无法解决如何防止衰老这一 难题的，为了便于全球科学家们一同努 力来解决这

个难题，Google和斯坦福大 学医学院以及杜克大学医学院一起，将 建立一个标准的人类医疗数据库，这个 数据库中包括5000人全部的生理和医 疗信息。三家参与方希望该数据库能成 为全球科学家们做研究和发表科研成果 的基准(Baseline)数据库。除了 Google 之外，更多的IT公司和IT人士开始涉足医 疗领域。事实上，由加州大学圣迭戈分校 教授约翰■克雷格■温特（John Craig Venter)等人创办的人类长寿（Home Longevity)公司在这方面甚至走到了 Google的前面，该公司于2013年成立， 今天已经开始为一些大的制药厂提供与 基因技术有关的服务了。人类长寿公司 的方法完全基于大数据，它为此聘请了 Google著名科学家、Goog丨e翻译的负 责人奥科博士担任首席科学家，而奥科 每天所做的事情依然是机器学习，这和 他在Google没有太多的不同，只是数据 从语言数据变成了生物数据。与Calico 所不同的是，人类长寿公司拥有临床的 数据，因此在将基因和疾病联系起来并 且找到治疗疾病的方法方面与应用更接 近。

如果Calico、人类长寿公司或者其 他什么公司能够利用庞大的数据找到很 多疾病的基因根源，那么接下来的问题 就是如何修复基因了。2014年，麻省理 工学院评选出的当年10项重大科技突破 中有一项技术恰恰就是基因编辑技术， 其主要发明单位和发明者是中国云南省 灵长类生物医学重点实验室，以及加州 大学伯克利分校的珍妮弗•多德娜 (Jennifer Doudna)博士、麻省理工学 院的张峰博士和哈佛大学的乔治•丘奇 (George Church)博士。其中多德娜 博士和从事这项技术应用的瑞士科学家 伊曼纽尔勒•卡彭特尔（Emmanuelle Charpentier)获得了 2015年的突破奖[18]。如果我们能够发现那些致病的基 因，并且使用这项技术修复基因,那么人 类的寿命有希望大大延长。当然，这项技 术也带来一个巨大的伦理上的挑战，不 过这不在我们的讨论范围内。

至于Calico和人类长寿公司的成果 何时能够商品化，这两家公司都不愿意 透露细节，但是至少它们给了我们长寿 的希望。

机器智能能够改变的所谓高级工 作，不仅在医学领域，还在法律、金融和

新闻等诸多领域。

图6.11因发明基因编辑技术而获得突 破奖的多德娜（右）和卡彭特尔（左）

未来的律师业

我们在前面讲到的大数据思维其实 已经在改变司法领域的工作方式，诉讼 的一方会通过数据之间的强相关性寻找 证据，而司法领域也认可这一类证据。大 数据对司法领域的另一个重大影响在于 机器智能会逐渐取代律师做一些案例分 析工作，这使得诉讼的成本有可能大幅 度下降。

与医生类似，律师过去在发达国家 也被认为是最“高大上”的职业。由于打 官司的过程长、费用高，而且法庭的判罚 常常带有惩罚性质（而不是简单的赔偿 性质），因此律师的工作显得特别重要， 而诉讼双方付出的律师费用也高得惊 人。2010年Viacom国际公司（ABC电 视网的母公司）诉Google旗下的 YouTube侵犯其视频的版权，并且要求 10亿美元左右的赔偿。后来Viacom被 发现是自己一边在YouTube上传视频， 一边告YouTube,因此而败诉。但是, Google为了打赢官司依然付出了 1亿美 元左右的律师费。在苹果与三星一场更 大的诉讼中，双方付出的律师费更高。虽 然小一些的公司之间的诉讼未必像 Google、三星和苹果这样的公司那么花 钱，但是费用绝对不低。根据美国知识产 权法律协会的调查结果，对于专利赔偿 诉求在100万美元之下的小官司,双方的 律师费花销居然高达65万美元（中位 数），对于赔偿诉求在100万到2500万 美元之间的专利官司，律师费花销为250 万美元（中位数），而对于那些赔偿诉求 在2500万美元以上的大官司，律师费用

更是高达500万美元（中位数）[19]

高昂的律师费不仅对大公司来讲是 个负担，而且使得小公司几乎难以赢得 官司，因为它们常常在打赢官司之前就 已经拿不出律师费将官司继续打下去。 在美国打官司，律师费用高昂的原因有 很多，其中最重要的一个是英美法系是 判例型法律体系（又称海洋法系），打一 场大官司，需要将历史上相关的官司法 律文件都拿出来分析，这个工作量巨大。 像Google和Viacom之间的官司，需要 分析上百万份历史文挡。

到了大数据时代，这个情况会慢慢 得到改变。今天，一些公司利用自然语言 处理和信息检索技术，发明了让计算机 阅读和分析法律文献的软件，可以取代 很多人工。位于硅谷帕罗奥图市的 Blackstone Discovery (黑石发现）公 司发明了一种处理法律文件的自然语言 处理软件，使得律师的效率可以提高500 倍，而打官司的成本可以下降99%,这意 味着未来将有相当多的律师（尤其是初 级水平的律师)可能失去工作。事实上这 件事情在美国已经发生，新毕业的法学 院学生找到正式工作的时间比以前长了 很多。

2015年，统计调查公司Altman Weil Flash对美国律师事务所的老板们 进行了一项民意调查，了解他们对自然 语言处理软件是否能够取代律师的看 法。总的来讲，大多数人相信计算机最终 能够取代人类当律师。只有20%的受调 查者认为计算机无法取代人类，另外有 38%的受调查者认为在5～10年内还不 能取代人类。47%的受调查者认为，在5 -10年内，律师助理将失去工作，这对律 师事务所的合伙人以及客户来讲，或许 是个好消息，因为诉讼的成本可以下降。 不过，尽管智能计算机取代有经验的律 师要稍微难一些，依然有13.5%的受调 查者认为，律师事务所里面处在金字塔 顶端的合伙人也会被计算机取代掉。因 此对这个行业来说，机器智能其实是一 把双刃剑。

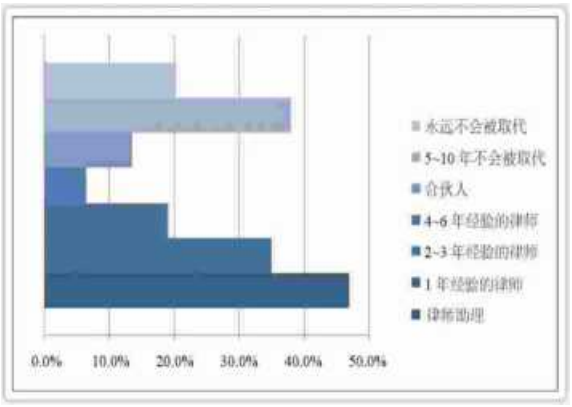


图6.12 Altman Weil Flash对计算机 是否能取代律师的调查结果

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，计算机会先“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旧M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 人元;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 *PMC 3048809* 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。

[1〇] 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]旧M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_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_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 多地占据了我们人类的工作机会，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_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_ 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时 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 于背后的一项技术一一区块链（Block Chain) O

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_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 得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_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_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 医生好的原因。

到了智能时代，机器的智能水平足 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 万物联网技术)、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 的可能性变成肝癌。其他很多疾病也有 这样的关联关系。这家医学院研究的目 的当然是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_般对投保者_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-F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(Goodwi丨丨）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—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，FB丨（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FB丨交出数据。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曰。 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 —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。机器智能如此天翻地覆的革命， 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 的西汉末年，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倍一从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体，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 Gradwell) 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，而比 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扯 也是先从少数精英开始，经过长达半个 世纪的时间，才

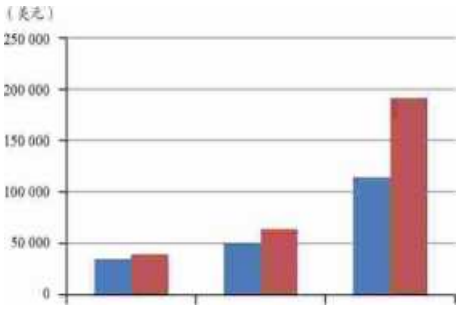
开始造福技术革命的中 心地区。而世界上大部分地区享受到第 二次工业革命的成果，是第二次世界大 战之后的事情。

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 的提升。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

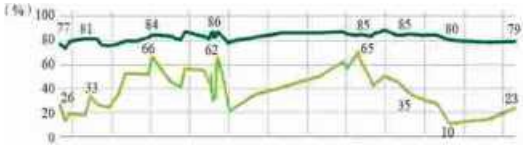
数据来源：美国国家统计局

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; 下方的青线是大家对物质生活的满意程 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。

对个人舌的_度与对国家11|]满濟度Wfif总体来说.你对段时间的!I35状况培沟滿S?

■对整体生活的确肩度 ■对拘质牛活的盹8:度



1980 1983 1986 1989 199； 1995 1998 2001 2004 2007 2010 20131^)

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，

P余去这两个国家，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。

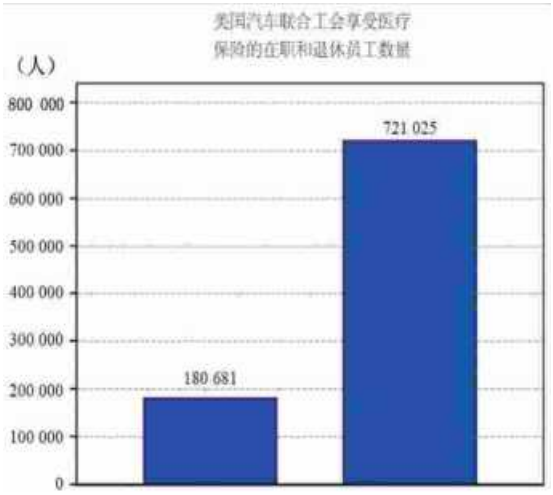
解决问题只有靠时间

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工会中在 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。



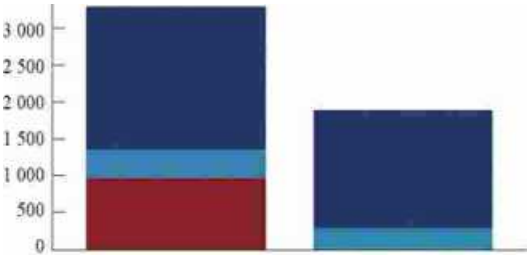
在职员工 退休员工及工

会成员的M

图7_ 10 2008年金融危机之前美国汽 车联合工会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曰本 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 是心非一一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 _。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 感一机器不能思考。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 非常需要脑力的工作。

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 人口所需的食品，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、丨T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

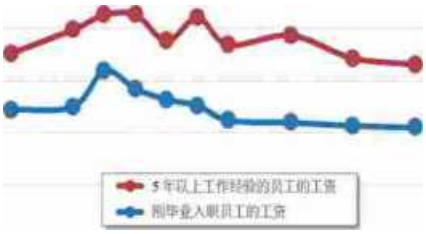
100 000

80 000

60000

40 000

20 000



1997 1999 2001 2003 2005 2007 2009 20丨丨.（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， 实际上等于鼓励懒情，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，剩下来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_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 们的希望一体面的工作。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，我问了他几个问题：

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 会变得更好;不过，和以往的技术革命_ 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%〇

[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

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

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