上海市教育委员会组编

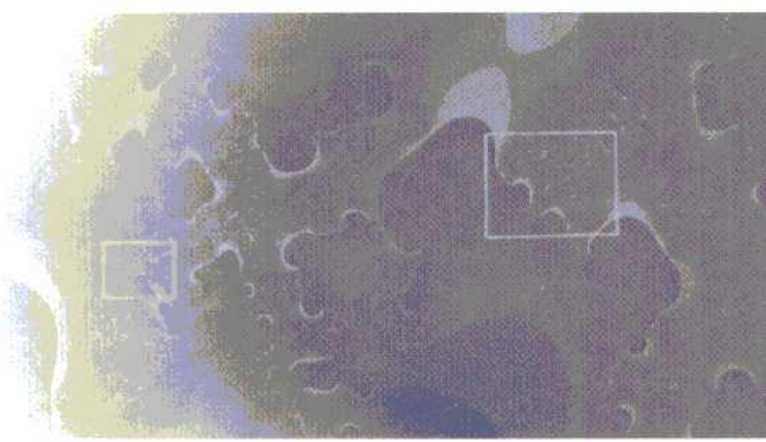
张奠宙

daxuesheng renwen suyang jiangzuo

大学生人文素养讲座

shuju yu ren sheng

数据与人生



49

上海交通大学出版社

上海市教育委员会组编

大学生人文素养讲座

数据与人生

张奠宙

上海交通大学出版社

内 容 提 要

“数字地球”的口号风靡世界。人类面临着处处充满着“数字”的环境。如何处理汹涌而来的“数字”大潮，是新世纪摆在每个中国公民面前的课题。我国王选院士实现了中文排版印刷技术上的革命，这归功于数学上的数据压缩技术。美国在数字电视上击败日本，是数据处理上的成功范例。本书还将揭露一些“数据欺诈”的事例，包括广告、中奖、伪科学等诸多社会生活中的数据问题。世界已经充满数据，人生必须处理数据。

图书在版编目(CIP)数据

数据与人生/张奠宙. —上海：上海交通大学出版社，2000
(大学生人文讲座/王铁仙主编，叶敦平、陈卫平副主编)
ISBN 7-313-02415-0

Ⅰ.数… Ⅱ.张… Ⅲ.数据-人生-青年读物 Ⅳ.G.416

中国版本图书馆 CIP 数据核字(2000)第 15837 号

大学生人文素养讲座
数据与人生
张奠宙
上海交通大学出版社出版发行
(上海市番禺路 877 号 邮政编码 200030)
电话：64071208 出版人：张天蔚
常熟市文化印刷厂印刷 全国新华书店经销
开本：787mm×960mm 1/32 总印张：46.5 总字数：912 千字
2000 年 10 月第 1 版 2000 年 10 月第 1 次印刷
印数：1—3050
ISBN 7-313-02415-0/G·342 全套定价：72.50 元
本册定价：5.50 元

前　言

自从有了人类，便有了“数字”，因为猎物是需要计数的。不过，直到20世纪前半叶，人们应付日常生活，还只需要扩大了的算术——知道数的加减乘除，以及一些几何图形知识就够了。数据，还只是数学家和科学家关注的对象。

1945年出现的电脑一步一步地在改变世界。身份证是一串数据，超市里的商品标是“条形码”的数据，人事处电脑里储存着“个人的一列数据”。20世纪90年代，美国的数字电视打败了日本的模拟电视，中国的中文排版技术占领了世界的汉字印刷业大半江山。1998年，美国副总统戈尔在“信息高速公路”的基础上，又提出“数字地球”的概念，整个地球都要数字化了。人们终于知道，社会人生离不开数据，任何人必须处理数据，不得不和数据打一辈子的交道。

在数据背后的是数学。统计学是处理数据的，数理统计学更揭示出数据后面的数学规律。与20世纪50年代相比，中学生已经减少了对平面几何的学习，数学课却增加了“统计”内容。当电视上播出“去掉一个最高分，去掉一个最低分”的时候，数据处理就在我们的身边了。

用数据说话往往是最有说服力的。可是，数据也可以造假。原苏联的李森科，就用假数据来制造

伪科学。广告上的数据欺诈不可不防，但是合理的数据表示又是时代的要求。

本书是为“非数学专业”的大学生们写的，里面没有数学公式，也没有大段的推理。它只向读者表明，世界在数字化，人生将面对数据。“预则立”，让我们更多地关注那些“枯燥的数据”，把它们处理得富有人情味，具有社会感，在新世纪中面对从未有过的“数据人生”。

1 数字化与数学观

数字化生存，预示着数字化的未来。当世间万物都数字化的时候，没有数学的生活是不可想象的。信息时代要求人们掌握基本的数学知识，具备良好的数学能力，更为重要的是树立正确的数学观。数学不再被人们只当做“升学”的敲门砖。数学将浸透在人们的日常生活之中，成为谋生立业的基础。数学，将是为发展生产和兑现经济效益的技术。数学技术体现了新的数学文化现象。

1.1 数学技术：从几则新闻说起

1978年，徐迟的报告文学《哥德巴赫猜想》，风靡大江南北。新闻媒介使陈景润成为中国青少年的科学偶像。数学，在人们的心目中几乎等于陈景润和他的哥德巴赫猜想研究。陈景润的科学业绩已经载入中国的史册，无需赘言。20年过去了，数学正在发生悄悄的变化。除了华罗庚、陈景润等的数论研究之外，数学新闻正向数学技术倾斜。请看以下的新闻和事例。

(1) 1998年9月7日《文汇报》报道：8月16日，沙市水位从44.88米涨到晚上11时45分的

45.08 米。估计到凌晨 5 时，水位可能涨到 45.20 米。只等中央一声令下，早已准备好的 20 吨炸药将立即在荆江大堤分洪。次日凌晨 5 时 30 分，中央命令的传真文件下达，要求继续严防死守。作出这一重大决策的根据有：

> “由多方专家组成的水利专家组用‘有限元素法’对荆江大堤的体积渗透进行了测算，确定出一个安全系数。照这一系数推定，即使沙市水位涨到 45.30 米，也可以坚持对大堤严防死守，不用分洪。”

这“有限元素法”便是一种求解偏微分方程的数学方法。我国已故数学家冯康在 1965 年曾做出前驱性的贡献。

(2) 1998 年底，中央电视台、香港凤凰卫视先后播送王选院士的访问记。方正集团，中国高科技产业的“大哥大”，已成为改革开放时代的成功代表。但是，这项告别“铅与火”的印刷业革命，其核心技术之一是数学技术。方正创始人王选，1962 年毕业于北京大学数学系，从事汉字排版技术研究。20 世纪 70 年代，直接采取“数字化”方案，运用自己发明的“汉字”信息压缩技术，取得了世界领先的水平。台湾地区《中央日报》也买方正的技术。

技术的核心是数字信息的压缩。一个汉字包含许多笔画，形状各异，如何使用较少的信息加以刻划，是一个核心问题，当然也是核心商业机密。

(3) 1996 年 8 月，英国的格拉斯哥地区法庭审理一桩工业诉讼案。原告是一所剧院，起诉某建筑设计院的设计不好，影响票房收入。而证人是一台电脑。法庭诉讼中，将剧院设计的各种设计参数输入电脑，并用直观的动画模拟剧院内空气流动的情况。结果是剧院业主胜诉。电脑中存储的空气动力学软件其实是一组数学方程式。此举开创了电脑模拟作为法庭证据的先例。

(4) 美国对冲基金 1998 年损失了 100 亿美元，原因是运用的金融数学技术出现了失误，一个小概率事件导致了错误的决策，损失惨重。

(5) 周光召著文谈中国科学的发展，提到美国现在掌握着未来电视的技术，日本的电视王国地位受到动摇。原因是美国掌握了数字化技术，其核心是“小波”数学技术。

此外，密码编制和破译，CT 扫描技术，海湾战争模拟技术，喷气机和航天器的控制技术，说到底，其核心都包含着一种特殊的数学技术。

因此，人们对数学的认识发生了重大变化。如果说，陈景润在解决哥德巴赫猜想上的贡献是激励人们攻克数学难题，那么王选的贡献表明数学是可以成为直接创造经济效益的技术。数学技术可以成为企业技术的核心。

那么，请读者审视一下我们流行的世界观，回答一个基本的问题：什么是数学？前苏联的加里宁有一句名言：“数学是思想的体操”，数学通过逻辑证明、严密推理培养人们的良好思维习惯。这句名言尤其成为数学教育的指针。极端的说法是把数学等

同于逻辑。另一种说法是，“数学是科学的语言”。数学是物理学、工程学、天文学等等学科的工具。科学规律都是用数学公式表示的。因此，不懂数学也就无法掌握其他科学知识。这当然是完全正确的论断。但是，20 世纪下半叶数学的进展，特别是电脑技术的推动，数学已经从科学的后台走到幕前，成为一门能够直接产生经济效益的技术，前面的例子说明了这一点。

大百科全书的“数学”条目是这样写的：“数学是反映现实世界空间形式和数量关系的科学”。它仍然是数学的最好概括。不过，我们的理解应当加深。数学从反映现实到服务于现实，从“思想体操”和“科学工具”，进一步提出“数学技术”，乃是数学观的一项革命性进步。

1.2 数字化：电视大战的历史经验

《纽约时报》畅销书，尼葛洛庞帝的《数字化生存》，于 1997 年由海南出版社出版。一年内三次印刷，一时供不应求。封面上写着：“计算不再只和计算机有关，它决定着我们的生存。”

数字化将是未来社会的主要特征。世间万物都会数字化。地球、气象、资源、海洋甚至交通指挥，都会数字化。未来卫星的分辨率将是一米。当人们用卫星导航驾驶汽车的时候，数字信号的传送是我们生存空间的灵魂。一个明显的事实是：一切信息都将以数字的方式而存在，数字的计算和信息交换将密不可分，那么，人们离开数学怎么行？

现在让我们回顾在20世纪90年代发生的电视大战。这场媒介革命的历史经验证明了数字意识和数学技术是何等重要。

电视是20世纪最重要的发明，它对社会发展的影响怎样估计都不会过高。那么未来的电视技术将是什么样子呢？

日本是世界电视工业的领头羊。索尼、松下等公司一向领导电视技术的新潮流。早在1972年，几位富于前瞻的日本人对自己提出问题，下一步的电视往何处走？他们的答案是："用高分辨率提高清晰度。"确实，高清晰度电视是每个人都会提出的要求。

大家知道，现在的电视使用的是模拟技术。先用摄像机将画面的光学特性转变为电的特性（光电效应），即将图像上明亮和色彩不同的光点，逐点、逐行、逐帧地转变为一串电信号。这一转变是模拟性质的，例如光的亮度和电信号的强度一致。然后电视台将此电信号发射出去，电视机再将它们逐点、逐行、逐帧地同步恢复原来各个光点，重构原来的画面。中国现行标准是每秒传送25幅画面，每个画面上有400条水平线，每条线上有400个点，屏幕宽高比是4:3。显然，每幅画面分解的行数和点数越多，图像就会越清晰。用通俗的语言来说，"电视屏幕上的点子越细越好。"计算机屏幕上有1150条线，当然就比电视画面要清晰得多。由于光信号转变为电信号的方法多种多样，于是有美国的NTSC制，法国的SECAM制，德国和亚洲的PAL制，互不统一。

日本人的超前意识，意味着更多的帧频（每秒60幅），更高的行数（每帧1125条或1250条），每行

有800个点，以及更合理的屏幕宽高比16:9，更科学的电子技术设置等等。他们花了12年的时间，提出了各种制式，希望能够得到国际的公认，而且最好是全世界统一的制式。然后按新制式生产各种新的电视摄像机、发射机、接收机，从而获得新一轮电视贸易的主动权。但是，这是一个落后的技术意识，数字化电视把日本的“高清晰度”模拟技术彻底打垮了。

大约在1981年，欧洲警觉到日本可能独霸电视市场，也研究起“高清晰度”电视。从贸易保护主义的立场出发，1986年欧洲共同体立法通过了自己的“高清晰度”模拟技术。20世纪80年代以前的美国，几乎没有电视工业，所有的电视设备全由国外进口，特别是日本。在日本和欧洲电视战的初期，美国的一些研究机构支持日本的方案。后来，美国又反对日本和欧洲方案。可笑的是，一些厂商拾人牙慧，所提出的方案仍然是模拟技术。1987年，美国通讯委员会(FCC)决定开发新一代电视，争夺市场。当时美国有四家公司，另有日本的方案作为第五家共同竞争。

1991年，美国通用仪器公司(General Instrument Corporation)和麻省理工学院联合宣布实行电视的全数字化技术方案。几乎在一夜之间，美国所有研究电视技术的计划全部改弦易辙，抛弃模拟思想的陈旧温床，站到“数字化”的大旗之下。有充分的证据显示，数字信号的处理有更好的成本效益。美国的四家公司都提出全数字化方案，日本的模拟方案明显落后，不得不于1993年退出竞争。

1991 年 9 月,《数字化生存》作者尼葛洛庞帝在一次午餐会上向法国总统密特朗和内阁成员建议:放弃他们自称的"领先地位"——电视模拟技术。当时法国没有接受。英国的撒切尔夫人听从了尼葛洛庞帝的建议。1992 年末,英国的梅杰首相否决了给"高清晰度"电视计划补贴 8 亿美元的提案。到了 1993 年,欧洲共同体终于承认了"数字化"方案的优越性,宣布放弃模拟技术的发展计划。

1992 年,尼葛洛庞帝向日本首相宫泽喜一说明"高清晰度电视没有前途",这使宫泽喜一大吃一惊。1994 年,日本邮政省放送行政局长讲山晃正提议日本跨入"数字世界"时,遭到日本产业领袖的围攻。日本人在"高清晰度"电视上投入的钱实在太多了,不肯轻言放弃是可以理解的。但是,当美国和欧洲相继放弃模拟技术之后,世界市场的大部分已经在"数字化"方案的控制之下,日本在"高清晰度"电视模拟技术上的近 20 年努力,终于宣告失败。

1.3 烽火台:"比特"的胜利

数字信息的单位是"比特",上节所述的电视大战以数字化方案获胜而告终,因此被称为"比特的胜利"。

那么,什么是"数字化"呢?它何以有如此大的威力?

以电话为例,过去的老式电话是将声波模拟成电波传送,现在的数字式电话是把声波变换成一组数字。接受者将这串数字还原为声音,传送即完毕。

再如，如果要将你的基本情况告诉别人，只要把你的身份证号码传过去就行，对方按这串数字就可以读出你的住址、性别和出生年月。传一串数字比传一组汉字要方便多了。

数字有0,1,2,3,4,5,6,7,8,9共十个。从技术上说，只传送两个数字0、1就够了。这是因为有以下的对应关系：

0 ⟶ 000
1 ⟶ 001
2 ⟶ 010
3 ⟶ 011
4 ⟶ 100
5 ⟶ 101
6 ⟶ 110
7 ⟶ 111
8 ⟶ 1000
9 ⟶ 1001

因此，所谓数字化，实际上是把信息化为一串由0、1组成的数串。身份证不过15个数字，转换成0、1数串就要几十个数字。如果要传声音、图像等信息量很大的信息，那就需要传送几万几亿以至更高数量级的0、1数串了。

那么，用什么方法来衡量信息量的大小呢？这便是“比特”(Bit)。这要从中国古老的烽火台说起。

我国古代传递军事信息，烽火台十分有用。平日边关太平无事，烽火台便没有动作。一旦敌人来犯，烽火台上便燃起烽烟，瞭望哨发现烽烟，立即通报大本营决策迎敌。这是最原始的通讯，只有“无

烟”、“有烟”两种情况(分别用0、1表示)。因为这种0、1通讯最原始、最基本、最简单,我们很自然地将它定为信息量的单位,即定义为一个信息量。信息论的奠基人申农(Shannon)将这样的单位信息量称为1比特。

定义:在通讯过程中,如果传送的信息只有0、1两种情况,则称该信息的信息量为1比特。$\log_2 2 = 1$。一般地,如果信息有N种情况,则定义它的信息量为$\log_2 N$。以2为底的对数,在这里很自然地用上了。

例如,在烽火台通讯中,如果某前沿阵地有甲、乙两座烽火台。甲台表示有无敌人来犯,乙台表示是否需要补给。于是有以下四种情况:

甲台	乙台	
0	0	(没有敌人,不必补给)
0	1	(没有敌人,需要补给)
1	0	(敌人进犯,不必补给)
1	1	(敌人进犯,需要补给)

这时的信息量显然应该定为2比特:

$$\log_2 4 = 2\log_2 2 = 2(\text{比特})。$$

比特的定义,使原来无法确切衡量的信息得到数量表示。如果说烽火台所传送的早期信息不过是几个比特,那么传送一封信,或一篇文章,那就需要几千甚至几万比特。过去使用电报传送信息,用滴滴嗒嗒的短声长声(分别是0和1),可以表示数字、英文字母以及汉字(四个数字表示一个汉字)。到了无线电通讯和广播时代,信息量又成倍增加。后来,发明了电视,一个画面的数字化信息量相当于100

万比特。现在市面上通用的5英寸光盘可存储100亿比特。以美国为代表的工业化国家正在建立“信息高速公路”，就是要在全国畅通无阻地传送这类极大容量的信息。

1.4 现在的计算机运算速度仍然是太慢了

电子计算机的运算速度，以令人眩晕的方式在发展。从每秒运算几千次，到几万次，乃至上亿次，都曾经令人激动万分，把它看做是人类智慧的伟大胜利。

计算速度是障碍人类科学前进步伐的主要原因之一。许多科学问题，有了可靠的理论，也有实行的社会需要，就是因为计算速度的制约，理论只能束之高阁。

气象预报是一个典型的例子。大气环流的动力学理论，包括它的数学方程，在19世纪已经完成了。无线电通讯技术使得气象数据的收集和传输成为可能。但是，用这些数据，求解这些方程，至少需要几天甚至数月的时间。天有不测风云，气象预报必须在几小时之内发出才有实际价值。明日的天气要几天之后才能算出来，成了事后诸葛亮，还有何用？1950年4月，电子计算机设计方案的创始人冯·诺依曼，和气象学家合作，运用早期的计算机成功地进行了世界上第一次“数值天气预报”，成为计算机解决科学问题的一个里程碑。

现在的计算机每秒运算1亿次，已经不是新闻。但是，从数学计算的要求来说，这个速度还是太慢

了。现用著名的货郎担问题为例加以说明。

若有一货郎想走遍 N 个村庄，试问怎样走可使路线最短？这是一个世界性的难题，至今未获得满意解决。所谓满意解决，是指找到一种算法，使得能用计算机在可以承受的时间内得到结果。遗憾的是，现有算法都需要太长的计算机时间——几年，因而事实上做不到。

为便于说明起见，我们用最笨的算法来计算。第一步，找出所有的路线。第二步，将所有的路线长度两两比较，长的弃去，短的留下。这样，把所有的路线都比较过，最短路线自然就出来了。我们取 $N=31$。31 个村庄，对货郎担来说，不算多。

先计算路线数。设货郎在某村庄出发。他的第一站有 30 种选法。第二站有 29 种选法，第三站有 28 种选法。依此类推，全部可以走的货郎路线是 30！条，其数量级是 10 的 32 次方：

$$30!\approx 2.6\times 10^{32}$$

然而，1 天 = 86400 秒。一年 = 3.2×10^7 秒。倘若我们用每秒一亿次（即 10 的 8 次方）电子计算机，不停地运转一年，可以运算

$$10^8\times 3.2\times 10^7\approx 3.2\times 10^{15}\text{次。}$$

因此，要完成 30！次，需要大约 0.8×10^{17}年，即 8 亿亿年！即使用每秒 1 亿亿次的电子计算机（现在还没有），用这一算法来计算，还得 8 亿年，也是办不到的事。

当然，大家会说，这个办法太笨了，有些一眼就看出是兜圈子的路线，根本不需要拿来比，及早把它们剔除就是了。确实，数学家正在努力这样做。他

们定的标准是:找一种算法,能在 N 的若干次方的运算时间内完成货郎担最短路线问题。这种算法,称为"多项式算法"。大家知道,$N!$ 的数量级和 10 的 N 次方差不多,我们把这种计算次数等于某数的 N 次方的算法,称为指数式算法。上述的货郎担问题笨算法就是一种指数式算法。至于货郎担问题是否有多项式算法,至今尚未找到答案。记得 1979 年,《纽约时报》刊登消息,说苏联的哈奇扬找到了货郎担问题多项式算法。其他各国的报纸也多方报道,包括中国的《参考消息》。后来经过验证,原来是哈奇扬提出了一种求解线性规划问题"椭球算法",这是多项式算法。但椭球算法不能用于货郎担问题。即便如此,哈奇扬也获得了很高的国际声誉。可以想像,如果有朝一日找到了货郎担的多项式算法,一定会轰动世界,为世界各大报纸争相报道。

1.5 数据压缩与方正集团

数学是思维的体操,数学是科学的工具。如果说数学是智慧的源泉,那么人们把数学奉为"科学的女王"。如果说数学推动社会的发展,那么数学是一门服务性的科学,一直是幕后英雄,"科学的侍女"。

但是,到了 20 世纪的下半叶,数学发生了重大变化:从幕后走向前台,成为能够直接创造财富的数学技术。且不说 CT 扫描、密码破译、军事模拟、最优控制等广为人知的数学技术,只就最火红的计算机软件来说,数学也在其中发挥重要作用。反过来,软件事业也为数学的发展带来了生机。这里我们介