

第1章

意义：人类文明的基石



没有调查，没有发言权。

——毛泽东（伟大的无产阶级革命家）

除了上帝，任何人都必须用数据来说话。

——爱德华兹·戴明（美国著名质量管理专家）

在当今社会，“数据”会伴随一个人的一生。设想一下：从你呱呱坠地的那一刻，就有一堆数据被记录下来——妊娠周数、出生时刻、体重、身长、体温，等等；在你成长的过程中，还是被大量数据所环绕——年龄、住址、学习成绩、工作经历、婚姻状况，等等；就算你离开人世之后，还是摆脱不了数据的纠缠——死亡时间、死亡原因、生前工作单位、生前声誉、生前贡献……如果没有这些数据，你都无法客观地认识自己和评价别人。

站在更高的层次上看，人类文明的产生与进步也是通过对数据进行收集、整理和提炼而达成的。在史前时代，人类的祖先在没有发明记事的媒体工具时，已经开始使用数据了——从父辈和周围人的口耳相传中，知道了哪些环境可以居住、哪些动植物可以食用、哪些情况暗藏危险；有了文字之后，人类通过记录下来的历史数据来获取更多的经验教训——“秀才不出门，全知天下事”“以史为鉴，可以知兴替”；到了近代自然科学萌芽之后，数据的重要性逐渐提升到了一个前所未有的高度：不论是在哪个领域，科学家们很重



要的一项工作就是做实验采集数据，因为科学发明需要通过这些数据来推导或证实。

那么，什么是数据呢？如图 1.1 所示，传统意义上的“数据”，是指“有根据的数字”，例如，人们常说的实验数据、统计数据，就是以数字的形式表现出来的，这些其实只是狭义上的数据。

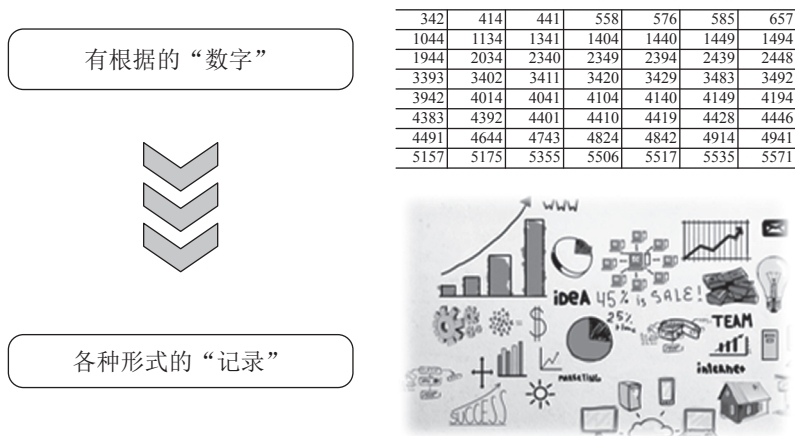


图 1.1 数据定义的变化

随着技术的进步，“数据”的范围得以扩大，可以代指许多“结构化的信息和情报”，比如，人们经常提到的一个词——“数据库”，其实就是指符合一定格式的信息的汇总。数据库里的数据，可以是某个机构所有成员的基本情况，包括姓名、年龄、通信方式、学历以及履历等（文字信息），这些已经超出数字的范畴。

进入信息时代之后，“数据”的含义更加宽泛，包括任意形式的信息，比如互联网上的全部内容，档案资料、设计图纸、病例、影像资料，等等。可以说数字、文本、音频、视频、图形各种形式的“记录”组成了广义的“数据”。

1.1 如何做出决策

如图 1.2 所示，人们在生活中经常痛苦于“彷徨”和“纠结”，这两者的区别在于：彷徨是因为无路可走，纠结是因为有太多路可走……后者的痛苦更需警惕，因为“选择的

河流越宽，淹死的人就越多”。人们的时间和精力都是有限的，面对各种可能，总是需要做出取舍，进行抉择。



图 1.2 纠结的痛苦

那做决策的依据应该是什么呢？有时候是依靠本能（直觉），例如，微生物来到酸碱度不适合的环境中就会逃走，人类感到了灼烧的疼痛就会退缩和躲避；有时候是依靠随机，例如一个人喜欢吃馒头、包子、面条、米饭，到了食堂随便买哪一样都行；但面对复杂问题时，直觉会失灵，而随机的后果难以承受——一旦选错了方向走错了路，失败就不可避免。有句话说得好，“朝相反的方向奔跑，停下来就是前进”。

➤ 1.1.1 祈求神灵的启示

清朝末年，有一位金石学家名叫王懿荣，他生病后去药房拿药。出于某种敏感的职业直觉，他发现自己服用的一味中药——“龙骨”，上面有一些很像文字的东西，如图 1.3 所示。经过仔细观察，王懿荣先生认为这些符号很可能是上古时期遗留下来的文字，极具考古价值，于是开始收集这种曾经被当作药材的“龙骨”，再加以研究，终于确认了它的来历。

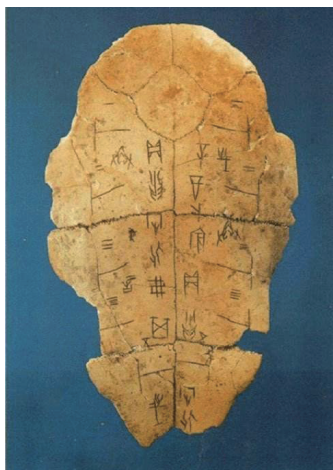


图 1.3 甲骨文示例



原来，所谓的龙骨是 3000 多年前的商朝人用来占卜的工具。他们把想要卜算的事情写到乌龟的甲壳或者其他动物的骨头上，然后再把这些甲或骨放在火上面烧，骨头受热以后，正面就会随着“卜”的一声出现裂纹。在古人看来，这些奥妙无穷的裂纹代表的就是上天赐下的旨意。最初，“卜”这个字的本意就是用火来烧（也有用尖锐物品钻）龟壳，让它产生裂纹，从而解读吉凶祸福。

在先民缺乏历史经验和科学知识的情况下，“遇事则卜”不失为一种非常实用的决策方式，毕竟面对多条路，无论走哪一条也比犹豫不前要好得多。况且这种方式还能够凝聚人心，具有极强的说服力——神灵给予的启示，还不赶快遵从！对于后人来说，这也留下了一笔宝贵的文化遗产，如果没有这种占卜活动，可能就没有人们今天称作“甲骨文”的商朝文字。所以，王懿荣发现了龙骨的秘密之后，众多文化学者立刻投入对这种文字的研究，现在已经大体上完成了对它的解读。

《礼记·曲礼》中说“龟为卜，策为筮”，这里的“龟”是龟甲，而“策”就是蓍草（如图 1.4 所示）。可见，在西周时期，用龟甲算命叫“卜”，用蓍草算命叫“筮”。按照《史记》的说法，当时的人们可以用卜筮来解决生活中的一切难题，包括卜财、卜居、卜岁、卜天、卜徙等。不管人们有什么疑问、有什么心愿，首先想到的就是通过卜筮来求得神灵的启示。就像现在的青少年，不管有什么想要的东西，都会试着在微博上转发锦鲤的图片。



(a) 鲜活的蓍草



(b) 干枯的蓍草

图 1.4 鲜活的蓍草与干枯的蓍草

根据民俗学家的研究，早期人们仅仅是点燃晒干枯的蓍草来烧灼龟甲，也就是说蓍草是占卜活动中的一个配角。但是随着时间的推移，出现了占卜的需求越来越多和乌龟数量不太够用的矛盾。当“一次性用品”龟甲极度稀缺时，蓍草就逐渐成为人们的首选。

虽然龟甲和蓍草都是古人心目中和鬼神沟通的神器，但在商周时期有一种说法认为“小事则筮，大事则卜”，即龟甲的神性比蓍草还要更强一些。在春秋时期，晋国的国君晋献公想要娶一名叫骊姬的美女，先用龟甲占卜，结果是不吉利；再用蓍草做卜筮，结果却是大吉大利，于是晋献公就相信了蓍草。这时，为晋献公做占卜的官员悠悠叹道：“筮短龟长，不如从长。”翻译成大白话就是：“蓍草这种新生事物，怎么能比得上我们沿用了好几百年的龟甲占卜传统文化呢！”

虽然晋献公在是否迎娶骊姬的问题上选择相信了蓍草，但并不因为他与时俱进、求新求变，也不代表他是一个狂热的“蓍草粉丝”。在后来不久，面对是否答应秦国提亲的问题上，这位国君就选择了相信龟甲占卜的结果——因为龟甲这次的结果更符合他的心意。也就是说，在他心中早已有了决断。不管是蓍草，还是龟甲，都只是用来支持自己决策的一个道具而已。

大约从战国末期开始，蓍草在日常生活中的使用频率就一路下滑。如图 1.5 所示，人们越来越不耐烦用复杂烦琐的方式来卜筮，而是倾向于使用简便的做法，如扔铜钱、掷杯筮、摇竹签，简单到了极致就是偏远农村的方法——“扔鞋”。近些年来，随着中西方文化交流的不断加强，有些人抱着“外来的和尚会念经”的观念，相信用塔罗牌占卜更加有效。



(a) 扔铜钱



(b) 掷杯筮



(c) 摇竹签

图 1.5 各种简便的卜筮方式



那么问题就来了，就这样烧灼龟甲、摆弄草棍、抛掷铜钱，真的能够获得神灵的启示吗？真的可以洞察先机、做出正确决策吗？站在唯物主义的角度，现在的人们当然是不相信的，其实古人也未必多么相信。唐代大诗人白居易写过一组政治抒情诗《放言五首》，其中第三首的内容如下。

赠君一法决狐疑，不用钻龟与祝蓍。
试玉要烧三日满，辨材须待七年期。
周公恐惧流言日，王莽谦恭未篡时。
向使当初身便死，一生真伪复谁知？

为了便于大家的理解，在这里简单翻译了一下（最后又多加两句扩展）。

我要送给你一种解决疑问的好方法，
既不用钻裂龟甲，也不用摆弄蓍草。
判断玉的真假，要耐心地烧满三天，
辨别木材好坏，还须等到七年之后。
周公辅佐成王初期，到处流言蜚语，
王莽篡夺皇位之前，天天毕恭毕敬。
假如这两人在未成事之前就死掉了，
他们本性的好坏，又有谁能知道呢？
奉劝大家不要被眼前的表象所迷惑，
尽可能多地收集数据才能做好决策！

► 1.1.2 战争背后的规律

《左传·成公十三年》里有一句话“国之大事，在祀与戎”，其中的“祀”指的是祭祀，“戎”指的就是军事或者战争。祭祀为的是凝聚大家的信心、团结大家的力量，而军事则是要捍卫领土和权力、保卫大家的劳动成果。中国自古以来就有“富国强兵”的理念——不解决挨打的问题，一切幸福指数都是空谈，所以各种战争案例和军事著作多得数

不胜数。其中，春秋时期孙武所著的《孙子兵法》^①不仅是世界上最早的军事著作，而且体系比较完备，对后世的影响非常之大。

《孙子兵法》的第一篇中就提到了以数据作为决策依据的思想：

“夫未战而庙算胜者，得算多也；未战而庙算不胜者，得算少也。多算胜少算，而况于无算乎！吾以此观之，胜负见矣。”

这段话大致的意思是：

在战争未发动之前，先进行数据分析和计算，就可以预判最终的结果了。如果充分估算好敌我态势、采用了合适策略，开战之后往往会取得胜利；如果很少了解双方的有利条件和不利条件，开战之后就会遭遇失败。多做分析计算就能取胜，少做分析计算就会失败，更何况根本不做计算呢！

而且，孙武还在书中对使用数据的具体方法做了论述：“一曰度，二曰量，三曰数，四曰称，五曰胜。”其中，“度”是指国土的大小，“量”是指粮草资源的多少，“数”是指军队的数量，“称”是指双方实力的对比。孙武的意思是，战争的胜负可以通过这四个因素进行估计，而这四个因素本质上都是数据。作战双方都不断刺探对方的实力，试图获得准确的数据，同时也不断释放数据“烟雾”，以迷惑对方，掩盖自己的实力。

通过释放数据“烟雾”，以计诈敌，这样著名的战例在中国古代有不少。战国时期，魏将庞涓率领10万大军进攻韩国，韩国不敌，向齐国求救。驰援韩国的齐军采用了军师孙臆的“减灶计”：开始的时候设10万个灶，其后设5万个灶，最后减到了3万个灶。庞涓见到齐军所留的灶迹不断减少，就判定齐军出现了大量掉队、减员的现象，因此撤下步兵率领骑兵分队加速追击，结果在马陵中了孙臆的埋伏，兵败身亡。

东汉时期，西北边陲的羌族入侵武都郡，名将虞诩率兵前往救援。因为一开始兵力不足，虞诩需要避免正面决战，等待援军集结，就在行军途中使用“增灶计”：让官兵每人各做两个锅灶，以后每日增加一倍。羌兵见此，认为郡兵已来接应，不敢纠缠，因此争取到了行军时间。到达郡府后，虞诩集合全部军队，命令他们次日先从东门出城，再从北门入

^① 《孙子兵法》，共十三篇，总结了春秋以前战争胜负的许多经验，被誉为“兵学圣典”，也是世界三大兵书之一（另外两部是克劳塞维茨的《战争论》和宫本武藏的《五轮书》）。

城，然后改换服装，往返多次。羌人不知城中有多少汉军，惊恐不安，最终被虞诡击败。

在中外战争史上，大规模、系统化运用数据的经典战役莫过于威廉·特库赛·谢尔曼将军在美国南北战争期间领导的“向大海进军”（March to the Sea）。1864年8月，谢尔曼率6万大军挺进南方的中心城市亚特兰大之后，采取了后世史学家认为整个南北战争中“最为大胆、最为关键的一次行动”：挥师东进、横穿佐治亚州，一路打到美国东海岸线。

兵马未动，粮草先行。在还没有飞机的时代，军队的行进路线安排必须充分考虑后勤补给。北宋年间的中国科学家沈括分析过，对一支10万人的军队而言，随军辎重就要占去1/3的兵力，最后真正能上阵打仗的士兵其实不足7万，如果一个士兵需要3个民夫供应，那就要征召30万民夫和额外的管理人员，但就是这样庞大的后勤规模，也只能支持行军31天^①。沈括因此得出结论：凡行军作战，应该争取从当地获取粮草和补给，这是最为紧迫的事情，否则不仅耗费大，而且走不远、跑不快，作战能力极为有限！但是如何在当地获取补给，沈括却只字未提。

谢尔曼一方面从国家的人口普查部门获得了南方的人口、资源等方面的宏观数据，另一方面在亚特兰大搜寻了一切关于佐治亚州的地图、财税明细和各种统计表格。然后，他计划主动切断后方补给，以统计数据为“航标”，根据农场、牲畜、集市、车站等重要资源的分布，通过后勤参谋的精心计算，确定最佳的行军路线和在各地停留的时间。五路大军沿着规划好的路线向东部沿海重镇萨凡纳突进，部队不仅在当地完成了补给，摧毁了敌方重要的基础设施，而且遭遇了最少的正面阻击。攻占萨凡纳之后，谢尔曼向联邦军总司令格兰特报告说，部队沿途消耗骡子15 000头、牛10 000余头，各种粮食都来自当地，和战争开始前的预测相差无几。经此一战，南方的战略资源被掠夺和破坏殆尽，几个月后南北战争结束。

谢尔曼在战后给美国普查办公室主任约瑟夫·肯尼迪发去了感谢信：“此战证明，您

① 每个民夫可以背6斗米，一个士兵可以自带5天的干粮。1个民夫供应1个士兵，两人同吃同行，可以维持18天，如果计回程的话，只能进军9天；2个民夫供应1个士兵的话，单程可以维持26天，若计回程，只能进军13天；3个民夫供应1个士兵，且每吃完1袋粮食就遣返1名民夫，单程最多进军31天，若计回程，只能进军16天。

给我提供的各种统计表格和数据价值巨大，没有它们，我不可能完成任务……”在回忆录中，他总结说：“历史上没有任何一次行军远征，曾经建立在像这次一样完善和肯定的数据之上。”

在南北战争正式爆发的前一年，也就是1860年，是美国的大选年。这一年，一位伟大的美国平民——亚伯拉罕·林肯，经历种种逆袭，最终入主白宫，登上了美国政治的中心舞台。而这一年也是美国第八次人口普查年，所以林肯总统可以直接从普查办公室主任肯尼迪那里获取到大量的相关数据，来预测这次内战的最终胜负。

围绕着《孙子兵法》所论述的“度、量、数、称、胜”五个角度，我们看一看当时美国南北方的潜在军事力量对比：国土面积上，美国当时有三十三个省，其中二十二个省站在联邦一方（北方）；粮草资源上，虽然南方最大的优势就是其棉花出口占据全国出口额的60%，堪称国民经济的命脉，但是棉花毕竟不是粮食，北方的农业生产足以保证战时粮草的供应；军队数量上，全国18～45岁的青壮年劳力约有69%集中在自由州，即使中间的摇摆州全部倒戈，南方的力量也不过31%；经济实力上，联邦更是占有绝对优势，北方集中了全国三分之二的铁路和百分之九十的工业产量。这些数据起到了“定心丸”的作用，林肯深信，如果打持久战，胜利最终将属于北方。

可以看出，无论是林肯的信心还是谢尔曼的传奇，都源于他们手头的数据是大量的、系统的、成片的，背后有专业人员给予支持和维护的。这种“有数可用”，得益于美国建国之后就开始的、长期的、周期性的努力和强大的制度保障。这种制度化的数据收集体系，才是近代战争中美国和其他国家在数据使用方面拉开差距的根本原因。

► 1.1.3 数据治国的理念

美国是个年轻的国家，其开国至今不过200多年，但数据在其政治活动和社会生活中的历史，却几乎和其建国史相生相伴。美国的建国者一开始就把人口普查写进了宪法。他们认为，国家权力应该在人口之间平均分配，而这个政策的落实，必须用数据来说话。

当时的政论家、教育家诺亚·韦伯斯特继而指出，在所有的事实当中，用数据描述的事实是最准确、最锐利、最有说服力的。因此，描述一事实，增强客观性、减少主观性

的最好方法，就是尽可能地使用数据。

美国人对数据依赖到何种地步，从1787年在费城召开的制宪会议中的一个片段——讨论“一个黑奴应该拥有多少权利和义务”——就可以看出来：

既然人口的多少在一定程度上决定了权力的大小，那南方拥有庞大的黑奴群体，是否也应该计入总数？一开始，大家都认为，奴隶本来就不拥有政治权力，因此不应该计入总数，但在后续的讨论中，人口的多少不仅成为分权的依据，还和纳税的义务挂上了钩，即人口多的州，国会占的席位多，也要缴纳更多的税收。南方则主张，黑奴既然不享受政治权利，也不应该承担义务，但北方又认为这样南方占了便宜。争论又起，最后的结论是，每个黑奴按3/5个白人（自由人）的标准纳入南方人口的总数，这个总数才是南方权力分配和纳税的依据。

每个黑奴等同于“3/5”个白人，这一规则被写进了宪法，成为种族不平等的历史明证。宪法颁布之后，曾引起很多追问，例如，为什么是“3/5”，而不是“1/2”或者“2/3”？当时主导辩论的亚历山大·汉密尔顿也说不清楚。他后来坦承：这是一个瑕疵，但当时必须找出一个数字，这个数字可能不完美，但比没有强。这就是美国人对数据的执着，哪怕是歧视，也要用数据来衡量。

人口普查的作用从政治领域不断扩张，首先蔓延到了政策制定领域，然后是社会生活领域。人口普查也转化为向社会寻找“真正事实”的统计活动，通过收集充足的数据，国家可以掌握整个社会“出生率、性别、年龄、婚姻状况、健康、职业、寿命”等方方面面的情况。这就逐渐形成了一种数据文化：一方面提高识字率，减少文盲；另一方面要推广数学教育，减少“数盲”，以提高公民的思辨能力，使其学会独立思考。

美国的国父们都推崇数据文化，乔治·华盛顿、托马斯·杰斐逊和本杰明·富兰克林就是其中的突出代表^①。1788年，华盛顿曾经这样描述数学教育：“从某种程度上说，文明生活的方方面面都不可缺少数字的科学，对数学真理的追踪可以训练推理的方法和正确

^① 华盛顿的第一份工作是弗吉尼亚州的土地测量员，他深知数据对于认识客观世界的重要性，在第一次人口普查期间，他甚至亲力亲为，组织了美国的第一次农业调查；杰斐逊也曾做过土地测量员，除了是一名政治家，他还研究密码学、测量学和考古学；富兰克林则是一名政治家、外交家和科学家，年轻时曾沉迷于捕捉雷电，后来发明了避雷针。

性，这是一项有益的活动，尤其适合理性的人类。”杰斐逊则建议，所有的小学除了教授阅读、写作外，还应该开设数学课。他认为：“就像身体的其他组织一样，大脑的功能也可以通过练习而改善、加强。因此，基于数学的推理和演绎，是人类了解深奥法则的有益准备。”到了1802年，数学已经正式成为哈佛大学入学考试内容。

在这样一批建国者的推动下，数学教育很快在这个新生国家普及，并影响到了美国的货币体系改革和测量单位的统一。这些工作对后人数据意识的形成、数据文化的建立，也产生了深远的影响。

美国的货币体系改革

在英国殖民期间，北美大陆一直沿用英国的货币系统及测量单位。当时，英国货币单位分为英镑、先令和便士，其中，1英镑=12先令，1先令=20便士，换算过程比较麻烦^①。杰斐逊认为，美国应该简化自己的货币体系，以方便大众、推动商业发展。于是，在他的主导下，美国以十进制为基础，推出了以“元、角、分”为单位的新货币体系（1美元=10角，1角=10美分）。

为了推动新的货币体系在民间尽快流通，杰斐逊还在全美教育系统鼓励“数学和换算”方面的教学，并在随后出版的教材序言中写道：“我亲爱的同胞，我请求你——别再使用英国的货币计算方法，让他们用他们的，我们用我们的！他们的方法确实适用于他们的政府——专制的暴君把会计系统尽可能搞复杂、把人搞糊涂，以操纵税收和财务工作，但一个共和国的货币系统应该简单，简单到最普通的人也能方便地使用。”换句话说，美国货币体系改革的目标就是让一切关于数据的计算变得简单，让每个人在商业活动中能够方便地利用数据进行思考和决策。

十进制在亚里士多德时代就被发明了，但美国是全世界第一个在货币体系中普及十进制的国家。几年后，法国也跟进，制定了以十进制为标准的货币、测量和重量单位。随后这套标准逐渐推广到整个欧洲，乃至全世界。

^① 从1971年起，英国也将货币体系改为十进制，即1英镑=100便士，并取消先令。



19 世纪 30 年代，英国哲学家托马斯·汉密尔顿来到美国游历，并把他的亲身见闻写成了一本书《美国人及其作风》。他发现，美国人已经习惯于通过数据来做决策。例如，美国人会根据他人的财务状况对其进行分门别类，“我已经被清楚地告知，我的熟人当中谁有良好的名声和信誉以及他们每年的开支。”他最后在书中得出结论说：“我认为，在这群不断猜测、估算、预期和计算的美国人当中，算术就像是一种与生俱来的本能。”

类似的观察还有很多。1825 年，费城的一名医生统计了 7077 名新生儿的体重，并制作了一张重量分布表，发放给新生儿的母亲，以方便她们对比掌握自己孩子的情况。他还监测了孕妇在 280 天孕期中每天增长的体重，并发放给孕妇作为其每天饮食标准以及体重增长的参考。

今天，现代化的医院一般都秉承了这种数据传统，从体检、诊断到治疗，大部分的医疗环节都以数据为支撑。例如，孩子一出生就要开始接受体检，身高、体重、头围是三个基本的检查指标，美国医院除了提供各项指标的大小，还会提供该项指标的百分位^①。

在美国做手术，术前病人或家属会被告知手术的风险，例如 0.03% 的死亡率、0.1% 的感染率以及各种并发症的可能性。这些百分比的得出，都建立在长期收集数据的基础上。2013 年，美国外科医师协会（ACS）利用信息技术推陈出新，收集了 2009—2012 年全国 393 所医院、140 多万病人的数据，在这个基础上开发了一个手术风险计算器（ACS/NSQIP Surgical Risk Calculator）。该计算器能针对病人的情况，计算 1557 种手术的风险及各种并发症的可能性，为医生和病人提供手术前的决策参考和准备。

反观古代中国，数据意识淡薄由来已久，甚至可以称为当时国民性的一部分。从古至今的大量典籍里面，人们都能发现各种非常模糊、夸张的描述，例如，各种正史文字中的“千余轻骑”“几十万大军”“向北百余里”“身高丈余”，仔细想想这些已经不是“差之毫厘”了，怎么都能算得上“谬以千里”。

著名信息管理专家涂子沛先生在《大数据：正在到来的数据革命》一书中提及，在

① 百分位（percentile）就是个体指标相对于大众平均水平的参照。例如，体重 38% 的百分位，意味着这个个体的体重超过了 38% 的同类个体。

国外留学期间，通过工作和生活中的对比，他感觉到了中国人缺乏“用数据说话”的素养。中国的语言表达方式中“重定性、轻定量”的特点非常明显，口语中常常使用“大概”“差不多”“少许”“若干”“一些”等高度模糊的词语。例如，中国菜的烹调方法就会令美国教授抓耳挠腮、不知所措，其中关于“盐少许”“酒若干”“醋一勺”的提法，完全是跟着感觉走，让初学者无从下手。

中国近现代著名的思想家胡适就对一些人“凡事差不多、凡事只讲大致如此”的习惯和作风深感忧虑。1919年，他写下了著名的《差不多先生传》^①，活灵活现地描画了当时国人不肯认真、缺乏逻辑、甘于糊涂的庸碌形象。

差不多先生传

你知道中国最有名的人是谁？

提起此人，人人皆晓，处处闻名。他姓差，名不多，是各省各县各村人氏。你一定见过他，一定听过别人谈起他。差不多先生的名字天天挂在大家的口头，因为他是中国全国人的代表。

差不多先生的相貌和你我都差不多。他有一双眼睛，但看得不很清楚；有两只耳朵，但听得不很分明；有鼻子和嘴，但他对于气味和口味都不很讲究。他的脑子也不小，但他的记性却不很精明，他的思想也不很细密。

他常说：“凡事只要差不多，就好了。何必太精明呢？”

他小的时候，他妈叫他去买红糖，他买了白糖回来。他妈骂他，他摇摇头说：“红糖白糖不是差不多吗？”

他在学堂的时候，先生问他：“直隶省的西边是哪一省？”他说是陕西。先生说：“错了。是山西，不是陕西。”他说：“陕西同山西，不是差不多吗？”

后来他在一个钱铺里做伙计；他也会写，也会算，只是总不会精细。十字常常写成千字，千字常常写成十字。掌柜的生气了，常常骂他。他只是笑嘻嘻地赔礼道：“千字比十

^① 胡适先生创作的一篇传记题材寓言，对过去中国人数据意识淡薄、不肯认真、拒绝精准的庸碌形象进行了讽刺。全文原载于1919年出版的《新生活》杂志第二期。



字只多一小撇，不是差不多吗？”

有一天，他为了一件要紧的事，要搭火车到上海去。他从从容容地走到火车站，迟了两分钟，火车已开走了。他白瞪着眼，望着远远的火车上的煤烟，摇摇头道：“只好明天再走了，今天走同明天走，也还差不多。可是火车公司未免太认真了。八点三十分开，同八点三十二分开，不是差不多吗？”他一面说，一面慢慢地走回家，心里总不明白为什么火车不肯等他两分钟。

有一天，他忽然得了急病，赶快叫家人去请东街的汪医生。那家人急急忙忙地跑去，一时寻不着东街的汪大夫，却把西街牛医王大夫请来了。差不多先生病在床上，知道寻错了人；但病急了，身上痛苦，心里焦急，等不得了，心里想道：“好在王大夫同汪大夫也差不多，让他试试看罢。”于是这位牛医王大夫走近床前，用医牛的法子给差不多先生治病。不上一点钟，差不多先生就一命呜呼了。差不多先生差不多要死的时候，一口气断断续续地说道：“活人同死人也差……差……差不多，凡事只要……差……差……不多……就……好了，何……何……必……太……太认真呢？”他说完了这句话，方才绝气了。

他死后，大家都称赞差不多先生样样事情看得破、想得通；大家都说他一生不肯认真，不肯算账，不肯计较，真是一位有德行的人。于是大家给他取个死后的法号，叫他作圆通大师。

他的名誉越传越远，越久越大。无数无数的人都学他的榜样。于是人人都成了一个差不多先生——然而中国从此就成为一个懒人国了。

1.2 科学方法的产生

著名的华人历史学家黄仁宇，曾经在《中国大历史》等著作中对古代中国文化的一些缺陷做了剖析。他认为：在中国传统的理学和道学当中，一直都分不清伦理之“理”与物理之“理”的区别。这两个“理”混沌不分的结果，是中国人倾向于粗略的主观性、排斥精确的客观定量，从而养成了重形象、重概括、轻逻辑、轻数据的文化习惯。这种文化习

惯，使中国人长期沉浸在含蓄、模糊的审美意识中，凡事只能在美术化的角度来印证，满足于基于相似的“模糊联想”，止步于用逻辑来分析、用数据来证明，最终将表象上的相似当作本质上的相同。

归根结底，古代很多中国人对数据的漠视，缘于一种文化上的欠缺：随意、盲目、不求甚解、理性不足。从某种意义上讲，正是因为这种文化上的问题，科学技术最终在西方国家产生，近代中国的坎坷命运也就此铸成。

► 1.2.1 站在前人的肩上

中国的瓷器^①是一个伟大的发明，它对世界的政治文化和人类的日常生活都产生了巨大的影响。尤其是在宋代和明代，中国瓷器在世界上每到一处，就会掀起一股奢侈品购置的热潮，并改变了当地人的生活方式、当地的文化，甚至改变了当地的制造业。世界上还没有第二种商品能在几百年的时间里长期做到这一点。

葡萄牙国王曾经用 260 件中国瓷器装饰了桑托斯宫的天顶，这表明在当时欧洲最富有的皇室眼里，瓷器是美和财富的象征。大航海时代，西班牙人从美洲带走了一万六千吨（约五亿两）白银，这些白银的三分之一都用来购买了中国的货物，主要是瓷器和茶叶。这让中国赚足了欧洲人发现新大陆后 150 年的红利。在欧洲，还有后来的美国，中产家庭大都有一个带玻璃门的瓷器柜（这种瓷器柜就叫 china），里面展示着各种瓷质的餐具。家里没有瓷器柜，会被认为没有品位。

人们也对历史上的中国名瓷耳熟能详，例如代表性的唐宋青瓷、元明青花瓷，还有宋代著名的五大名窑——汝^②、官、哥、钧、定。如图 1.6 所示，这些都是人造的奇迹、祖先智慧的结晶，也是中国人的骄傲。但你可能不知道，当今欧洲人占据着世界高端瓷器市场 90% 的份额，其余份额由美国和日本瓜分，Made in China（中国制造）的瓷器只是在中低端市场。究竟是什么因素导致的呢？而且欧洲人喜欢讲“中国人发明了瓷器，后来欧洲人

① 瓷器是彻底的人造物，它和金属、玻璃（包括水晶）不同，在自然界是找不到的。它完全是人类活动的结果和文明的标志。

② 汝窑青瓷流传到至今的真品，已知的仅 67 件，“纵有家财万贯，不如汝瓷一片”。香港苏富比 2012 年“中国瓷器及工艺品”拍卖，“北宋汝窑天青釉葵花洗”经 34 口叫价，以 2 亿多港元成交。



又发明了它”，这又有什么鲜为人知的故事呢？



图 1.6 中国名瓷的代表：汝窑杯盏（左）与元代青花瓷器（右）

其实欧洲人制造瓷器的历史很富有戏剧性。由于和瑞典开战，萨克森公国的国王奥古斯都二世的财力几乎枯竭，于是他在 1706 年抓住了两个炼金术士来为自己炼制黄金，当然很快他就发现这件事是不可能的。由于在欧洲的瓷器售价堪比黄金，他就命令两个炼金术士研制瓷器，其中一个叫约翰·弗里德里希·伯特格尔的人因此而名垂青史。

从被奥古斯都二世软禁在阿尔布莱希茨堡到制造出欧洲的第一件瓷器，伯特格尔花了 4 年时间，做了 3 万次实验。他不仅记录了全部的实验过程和结果，而且把每一次实验之间的细小差异全都记录了下来。与熟练掌握瓷器制造工艺却不明白其中化学原理的亚洲工匠不同，这种科学实验和材料分析的方法，让欧洲人对瓷器烧制的原理有了理性认识和定量的了解，他们可以通过细微调节瓷土中元素的配比和调整烧制过程，来制造各种精致的瓷器。

伯特格尔的成功给萨克森公国带来了巨大的财富和荣誉，到了 18 世纪，德国麦森瓷器的售价已经是中国瓷器的两倍。今天麦森仍然是世界瓷都之一，并且在国际高端瓷器市场占有很大的份额。随后，奥地利和法国都在麦森瓷器的基础上不断研发新的工艺，例如，西洋珐琅彩瓷器^①被欧洲人带到中国，康熙皇帝非常喜欢，下令在大内仿制，这实际上标志着中国在瓷器制造技术上已落后于欧洲了。

18 世纪中后期，“英国陶瓷之父”乔赛亚·韦奇伍德先是在工厂中搞出了一种叫作

^① 一种将玻璃液化后烧制在瓷器表面的技术，不仅在瓷器表面营造出一种晶莹剔透的效果，也使得瓷器更加经久耐用。

“流水线”的生产管理方式^①，后来又把当时最先进的科技产品——蒸汽机——引入瓷器制造。这些措施不仅极大提高了瓷器的制造效率，而且不同批次的瓷器品质都能得到保障。他的后人在 1812 年还发明了骨质瓷器，这种加入牛骨粉的制瓷工艺让瓷器更加结实，因此可以做得更薄，甚至薄到半透明的状态（如图 1.7 所示）。正是从韦奇伍德的时代开始，瓷器首次在世界范围内供大于求。



图 1.7 欧洲制瓷工艺的代表：西洋珐琅彩瓷器（左）与韦奇伍德骨质瓷（右）

从这段历史看来，欧洲人之所以在瓷器制造上超越中国，正是重视科学方法和数据记录的结果。欧洲人在研制瓷器的过程中，保留了全部的原始数据和实验报告，这样，前人每取得一点进步，后人都可以直接受益。例如，前面提到的伯特格尔把 3 万多次尝试的点点滴滴都保留了下来，同样，韦奇伍德在研制碧玉细炆器时，进行了 5000 多次实验，也把所有的细节都记录了下来。

相比之下，中国工匠更多的是具有对制瓷工艺的悟性，他们靠“师傅带徒弟”的方法将经验代代相传，而徒弟是否能超越师傅，则完全靠悟性。中间即使有一些发明和改进，却因为没有详细的过程记载，或出于保密故意略去，很多精湛的工艺都无法传世，例如，宋代五大名窑的制作工艺大多失传了。这样，后世常常不得不重复前人的失败，而无法直接“站在巨人的肩上”进行攀登，久而久之，造成了瓷器制造技术“起点很高，进步缓慢”的窘境。这种对数据记录的不重视，不是中国瓷器制造业特有的问题，而是中国古代

^① 1769 年，韦奇伍德在自己开办的埃特鲁里亚陶瓷工厂里实行精细的劳动分工，把原来由一个人从头到尾完成的制陶流程分成几十道专门工序，分别由专人完成。这样一来，原来意义上的“制陶工”就不复存在了，分成了专门的挖泥工、运泥工、扮土工、制坯工等，他们必须按固定的工作节奏劳动，服从统一的劳动管理。

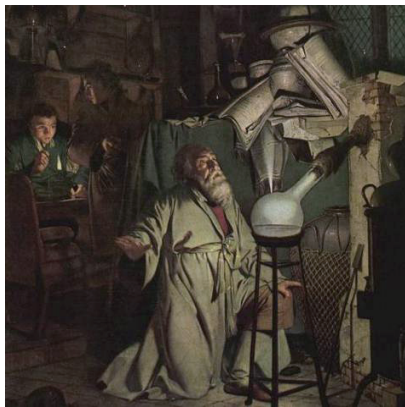
很多手工业普遍存在的现象。其实，中国古代的文献记录里面一直有这么一种现象：注重帝王，不注重平民；注重人文，不注重科学；注重定性，不注重定量。这可能也是中华文明在近代逐渐落后于西方文明的一大诱因。

➤ 1.2.2 事实胜过雄辩

1.2.1 节提到了“欧洲人再发明瓷器”的一个里程碑式的人物——约翰·弗里德里希·伯特格尔，他最初的职业是一个炼金术士。如图 1.8 所示，炼金术历史悠久，横跨了多个文明^①：在西方和伊斯兰世界，人们企图将廉价的金属变成贵重的黄金；在古代中国，则主要是为了制造万灵丹药和长生不老药，因此也叫“炼丹术”。人们学过化学之后，知道这些“炼金术”是行不通的，但正是这些术士们一代一代地前赴后继，催生了火药的发明，找到了各种矿物质，积累了实验的方法，制造了很多设备，进而产生了化学这门学科。



(a) 古代中国的炼丹道士



(b) 中世纪欧洲炼金术士

图 1.8 炼金术士的形象

为什么化学这门学科诞生于近代欧洲而不是中国，有很大一部分原因归于欧洲的炼金术士有意无意地采用了科学的方法。首先他们对自己做过的实验都有详细的实验记录，这

^① 炼金术有 2500 ~ 3000 年的历史，存在的地域包括美索不达米亚、古埃及、波斯、印度、中国、古希腊和罗马，以及中世纪的欧洲等。

些实验记录至今还保留在很多国家的档案馆里。还是拿 1.2.1 节提到的伯特格尔和韦奇伍德发明瓷器的过程举例，由于有了他们这些人的完整数据记录，人们现在才能轻而易举地复制欧洲历史上任何一件名瓷，但是中国的很多工艺却免不了“发明、失传、再发明、再失传”的命运，以至于现在，人们还无法完全仿制出宋代的汝瓷。

今天大多数中学生可能对物理和化学实验都颇有兴趣，但是写实验报告恐怕就没那么认真了，一般记录实验结果时常常随便找张纸潦草地写几个数据了事，更有甚者可能过分相信自己的大脑，记在脑子里回去再整理成实验报告。不仅丢失了实验细节，还会为了应付老师，篡改实验数据来迎合教材上的结论。笔者非常赞同吴军博士的说法：“一旦养成不做记录的习惯，就很难改，这么做实验无法很好地积累经验，后人只好重复前人的错误。”例如，人们今天不是很了解中国的道士们在炼丹技术上都做了哪些改进，明清道士炼丹的水平恐怕并不比隋唐时期的道士高多少，因为没有实验的数据积累，或者记录过于粗略。

科学方法的另一个要素也是炼金术士的贡献，即对每次实验的结果进行定量分析。量杯、天平、比重计和各种简单的测量工具都被用于炼金试验中，有了这些定量的记录和分析，后人就可以重复前人的实验结果，并在此基础上进行自己的改进和创新。这一点也成为今天在高级别学术杂志和学术会议上发表论文的前提条件。例如，在信息科学领域，要证明一种新的算法比以往的算法都好，就必须先重复近期发表的同类算法的实验结果。如果你只是给出自己算法的效果，而没有对比前人的算法在同等条件下取得的结果，任何权威的学术机构都不会承认你的工作。

定量分析带来的另一个结果就是，在科学上从尊重权威变成尊重事实。没有定量的衡量，很多观点和结论是不可比的，人们只好相信权威。在古代，人们喜欢这么论证，例如“亚里士多德是这么说的”“孔子是这么说的”，等等。到了近代，人们立论的证据不再是经卷上的教条，而是根据自己的观察或做实验的结果，因为定量的结果很容易比出好坏对错。笛卡儿就非常强调：“是事实而不是权威，才是验证一个结论正确与否的前提。”



拉瓦锡的实证精神

安托万·洛朗·德·拉瓦锡是法国化学家、生物学家，被后世尊称为“近代化学之父”。他提出规范的化学命名法，撰写了第一部真正现代化学教科书《化学基本论述》(Traité Élémentaire de Chimie)；提出了“元素”的定义并于1789年发表第一个现代化学元素列表，列出33种元素；他还统一了法国的度量衡，并且最终形成了当今现行的公制。

发现氧气和证实质量守恒定律是拉瓦锡的两个重大成果。在此过程中，他坚持采用了科学的方法：首先对命题进行怀疑；然后通过实验寻找证据，并对实验进行详细记录和定量分析；有了这些证据之后，再通过逻辑推理得出正确结论。可以说，拉瓦锡在研究过程中，再次确认了科学方法的重要性，对整个学科进行了综合，提出了新的学术思想，并建立了近代化学的学科体系。

法国大革命爆发后，拉瓦锡被雅各宾派领导人送上了断头台，据说这是他进行的最后一次“科学实验”——验证人的脑袋砍下来之后是否还有感觉。行刑前，他和刽子手约定自己被砍头后尽可能多地眨眼睛，据说拉瓦锡的眼睛一共眨了十一次（另一种说法是十五次）。虽然这个故事不见于正史，但是人们还是愿意相信它，因为拉瓦锡一生都在强调实验是认识的基础，这个传奇桥段的确是太符合他的做事风格了。

► 1.2.3 提高质量的法宝

近些年，每逢节假日都会有新闻报道，大量中国游客去日本游玩，回国之前抢购了大量日本产品，其中还不乏在中国制造的日本品牌，等等。这种现象说明在国人心目中，“日本制造”已经成为品质的象征。但大家可能不知道：第二次世界大战前后日本商品在国际上恰恰以“山寨”“低劣”而闻名；“日货”的崛起是在第二次世界大战之后短短十几年间完成的；而为此做出巨大贡献的竟是一个美国物理学博士——爱德华兹·戴明。

让我们把视线挪到1950年7月13日，虽说戴明早已多次搭乘军用飞机来日本了

（帮助指导人口普查和战后重建），但这一天的意义极为特殊。在日工盟^①主席石川一郎的安排下，戴明在晚餐会上见到了日本的21位行业巨头，和他们一起坐榻榻米、喝清酒、看艺妓表演。面对着掌管日本80%财富的行业巨头们，戴明向他们承诺说：“如果按照我倡导的原则去做，你们就可以生产出高质量的产品。5年内，日本的产品将占领整个国际市场”。5年！当时晚餐会上的所有人都认为这匪夷所思，但事实证明了戴明博士预言的准确性。日本的产品质量总体水平在四年后（大约1955年）就超过了美国，到20世纪70—80年代，不仅在产品质量上，而且在经济总量上，对美国工业造成了巨大的挑战。

跨界造就的管理大师

爱德华兹·戴明是耶鲁大学的物理学博士，由于在物理试验中产生的大量的数据，处理这些数据使他深刻体会到了“实际偏差是如何产生的，又该如何控制”。与数学博士乔治·盖洛普^②长时间的合作讨论，加上参与美国人口普查的经历，使得他逐渐偏离了原来的研究方向，进入了统计领域，成为美国首屈一指的抽样专家。接下来，他开始研究如何用统计方法进行质量控制；再后来，他又进入管理领域，成为名扬世界的质量管理大师。

戴明先物理、后统计、再管理，用现代的话来说，就是“跨界”。跨界是指跨越不同的领域、行业甚至文化，对其中的相关因素进行融合和嫁接，进而开创一片新领域、一种新风格或者一个新模式。戴明的跨界，开创了一个应用统计科学进行质量管理的新领域，其中的过程曲折起伏，令人感叹。感兴趣的读者可以翻阅涂子沛先生的《数据之巅》第五章——抽象时代：统计革命的福祉。

戴明的质量管理立足于一个基本信念，即高质量可以降低成本。控制质量，需要在生产过程中尽可能收集数据，利用偏差控制图和鱼骨图等可视化工具来进行分析。戴明认

^① 日本科学与工程联盟，JUSE，简称日工盟。

^② 乔治·盖洛普，美国数学家，抽样调查方法的创始人、民意调查的组织者，他几乎是民意调查活动的代名词。他于1935年创立盖洛普公司，该公司是全球知名的民意测验和商业调查/咨询公司。

为，无论企业的管理者还是生产者，都要学会制作这两类图表。

如图 1.9 所示，偏差控制图为每个偏差定义了一个变化的上限和下限，一旦波动超出了这个限度，就说明可能发生了特殊原因，应该首先消除。但这还不够，真正的质量控制，不仅要使偏差落在规定的范围之内，还要让偏差波动的范围越小越好，即在生产过程中也要全力消减共同原因，达到“稳定的一致性”。他认为，是否追求这种一致性，正是后来日本成功、美国失败的原因。

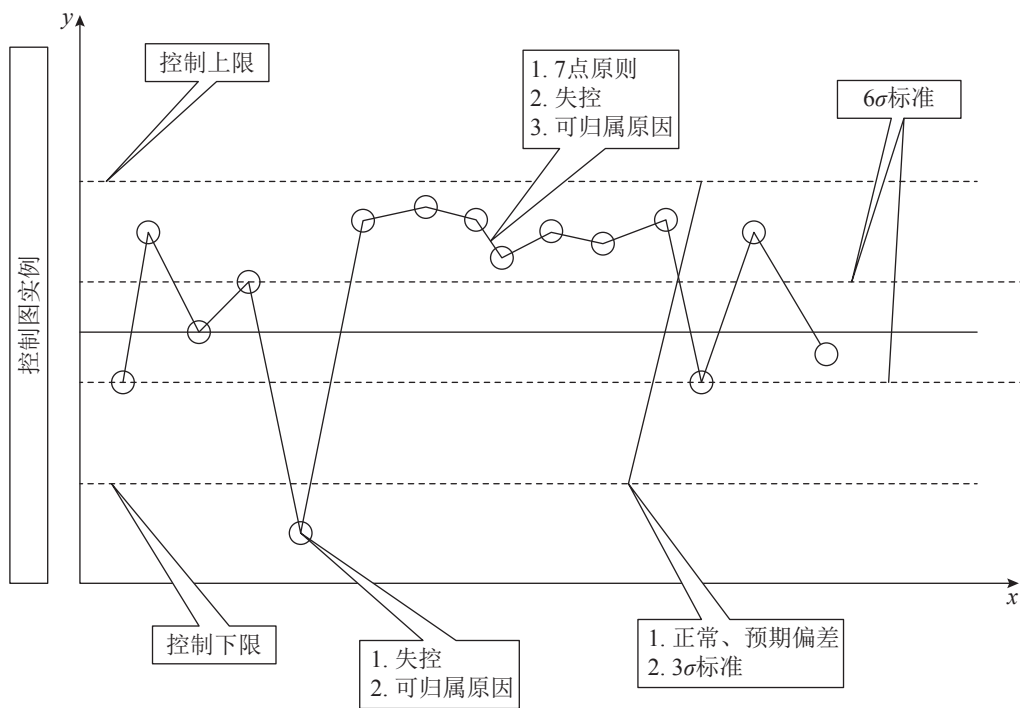


图 1.9 偏差控制图示例

发现了偏差，确定了偏差发生的类型，接下来就要针对偏差产生的原因进行因果关系分析，分析工具就是鱼骨图（因为全图像鱼的骨头，故称鱼骨图）。鱼骨图由日本学者石川馨^①提出，得到了戴明的充分肯定，从 20 世纪 60 年代开始在全世界企业管理领

^① 石川馨，日本式质量管理的集大成者，20 世纪 60 年代初期日本“质量圈”运动最著名的倡导者。石川馨是日工盟主席石川一郎的儿子，曾担任过戴明的翻译。

域风行。

图 1.10 就是针对某产品出现“尺寸超差”问题而绘制的鱼骨图，问题的起因可能有“材料、人员、环境、方法和设备”五大来源，每一个来源又分为若干个小因素，每个箭头都表示一个因素。戴明主张通过一线生产小组的集体讨论，共同绘制出这种分析图，并通过这个过程，让生产者、管理者一起积极地确定问题产生的原因，增强大家对于问题的理解并竭力避免。

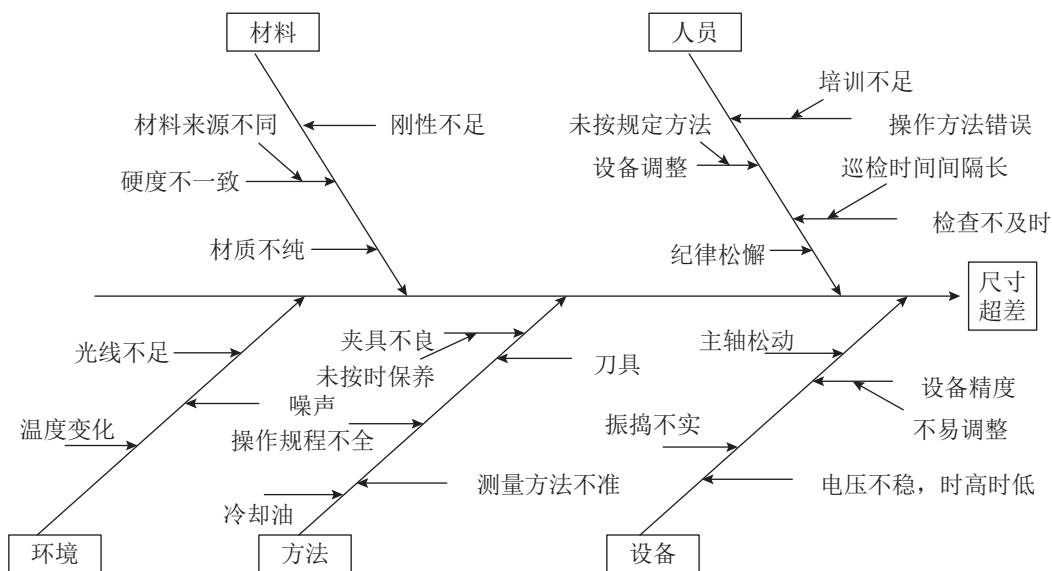


图 1.10 鱼骨图示例

丰田公司可以说是戴明质量控制理论最早、最大的受益者。到 1961 年，丰田公司已经在戴明和石川馨等人的指导下开创了一套全面质量控制体系（TQC），不仅在生产过程中全力缩小偏差范围，还完全吸纳了消费者调查方法。例如，在进入一个新市场时，公司甚至会派人去测量当地人的身高、腿长，以调整变速杆的高度和乘客腿部空间的大小。

丰田公司以及日本工业的战绩卓越：1975 年，丰田超过德国大众，成为美国最大的汽车进口商；1981 年，日本主导了整个国际汽车市场，成为全球最大的汽车生产国和出口国，其出口量是美、德、法三国轿车出口量之和；1983 年，丰田推出的佳美车型独步



天下，之后 10 年中有 9 年都是美国市场最畅销的车型（唯一一年屈居第二，输给的是一个日本品牌——本田雅阁）。而同时期的美国汽车巨头如通用、福特、克莱斯勒经营业绩不断下滑，每年都有高达十几亿的亏损。更要命的是，除了汽车，电视机、摩托车、录音机、复印机等日本商品在美国大行其道，“美国制造”黯然失色。

1980 年，丰田总裁丰田章一男在采访中说：“我没有一天不在思考，戴明博士于我们的意义何在——戴明是我们整个管理思想的核心！”



图 1.11 戴明质量奖章

日本人为了表达感激与敬意，用戴明捐赠的课程讲义稿费和募集到的资金设立了著名的“戴明奖”^①——一个刻着戴明侧像的银质奖章，用以奖励在质量管理方面取得重大成就的企业。如图 1.11 所示，在其肖像下面镌刻着戴明的一句话：“良好的质量和稳定性是商业繁荣与和平的基础。”

回顾戴明的故事，可以看到，戴明对日本的贡献不仅在于质量，戴明更大的遗泽在于推进了日本社会对数据统计的普及和重视。因为产品质量的崛起，日本的企业、政府甚至全社会都认识到了统计和数据的重要性。1973 年 7 月 3 日，日本内阁经会议讨论决定，将每年的 10 月 18 日定为“统计日”，帮助国民理解统计的重要性，鼓励他们形成对统计的兴趣，并在国家进行各项普查时予以最大限度的配合。日本政府内务部负责每年统计日的宣传、组织和实施，包括印制海报、组织知识竞赛、成果展览等。

除了国家统计日，日本每年还在中小学教师中组织“统计讲习会”，在中小学之间开展统计图表大赛，入选作品在东京的统计资料博览会上展出，最佳作品将获得总务大臣特别奖。此外，日本政府还在全国各地建设统计广场、统计资料馆、统计图书馆，以生动活泼的形式向大众介绍、展示统计的历史及最新的图书资料，在全民中推广数据的概念和知识。

① 1951 年以来，日本每年都评选戴明质量奖，国家电视台会现场直播每次颁奖典礼，视其为年度盛事。

1.3 智慧从哪里来

在日常使用中，人们总是混淆“数据”“信息”“知识”“智慧”这四个词语，其实从专业角度来看，它们是完全不同的概念。如图 1.12 所示：数据是信息的载体，但并非所有的数据都承载了有意义的信息；信息是有背景的数据，需要对相关领域有所了解的人才能够将其提取出来；知识要更高一个层次，也更加系统，是经过人类的归纳和整理，最终呈现规律的信息；而智慧则是根据运用已有知识，对获取的信息进行分析，并找出解决问题的方案的能力。

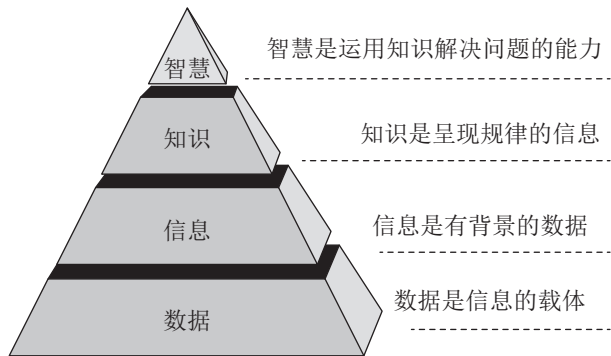


图 1.12 数据—信息—知识—智慧体系（DIKW 体系）

更加严谨一点描述如下：

- 数据是对现实世界的测量或抽象。
- 信息是经过处理、结构化、附加上下文解释的数据。
- 知识是人类已经理解和整理好的信息，具有规律性。
- 智慧是根据已有知识适时采取行动。

为了便于理解，这里举一个生活中的例子：“30”是一个传统意义上的数据；给它赋予背景之后可以成为“今年北京 7 月 16 日，气温 30℃”，这就是一个有逻辑含义的信息；结合每年 7 月北京的温度信息，就可以进一步提炼出来气候规律——“北京 7 月的平均气温全年最高，天气炎热”，这就形成了知识；如果能够利用这个气候方面的知识，7 月在北京策划一次防暑产品或避暑旅行的推介会，进而解决了某公司的经营业绩问题，这就可

以称得上有智慧了。

再举一个科学史上的例子：人们通过测量星球的位置和对应的时间，得到了大量的天文数据；在这些数据的基础上可以计算星球运动的轨迹，提取更为抽象的信息；基于这些信息进一步总结出来的开普勒定律^①，就是更有意义的知识；如果利用这些知识能够预测天文现象、确定时间节气，从而改变人们的生活和周围的世界，这就是智慧的体现了。

数据从哪里来？

数据主要的来源之一是“测量”，其狭义的定义为“有根据的数字”，强调的就是对客观世界的测量结果。数字之所以出现，是因为人类在实践中发现，仅仅用语言、文字和图形来描述这个世界是不精确的，也是远远不够的。例如，有人问“天安门广场有多大？”，如果回答说“很大”“非常大”“最大”，别人听了只能得到一个抽象的印象，因为每个人对“很”“非常”有着不同的理解，即使“最”也是相对的，但如果回答说“44万平方米”，就一清二楚了。

除了测量，新数据还可以由“原始数据”经计算衍生而来。这里说的“原始数据”，并不是“原始森林”这个意义上的“原始”，原始森林是指天然存在的，而原始数据仅仅是指第一手的，没有经过人为篡改的。毕竟，无论测量和计算都是人为的，没有“纯天然”。有了计算这个手段，人们就可以得到一些衍生的、间接的数据。在很多生产实践中，这些衍生数据甚至比原始数据更能起到直接的作用。例如，人们无法直接测量地球的质量，但还是可以通过测量地球上的物体质量和自然现象，来计算出重力加速度、万有引力恒量、地球半径等数据，然后再通过这些数据进一步计算地球的质量。

进入信息时代之后，“数据”二字的外延开始不断扩大：不仅指代“有根据的数字”，还统指一切保存在计算机中的信息，包括档案资料、设计图纸、病例、影像资料等。“文本、音频、视频”的来源往往不是对世界的测量，而是对世界的一种记录，所以信息时代的数据又多了一个来源——记录。

^① 开普勒定律是德国天文学家开普勒提出的关于行星运动的三大定律，分别称为椭圆定律、面积定律和调和定律。

从前面介绍的历史事件和科学实例中，可以看出数据的作用自古有之，并非到了今天的信息社会才突然显现。但是在过去，数据的作用常常被人们忽视，这是什么原因呢？可以比较一下人类文明的三个基本要素——物质、能量和信息，会发现，这三者之中物质资源相对直观，信息资源比较抽象，而能量资源则介于两者之间。由于人类的认识过程也是从简单到复杂、从直观到抽象的，所以材料科学与技术往往发展在前，接着是能源科学与技术的发展，最后才是信息科学与技术的发展。

况且在生产力和生产社会化程度不高时，要想积累足够的数据，可能需要几代甚至几十代人的努力。在如此漫长的时间里，用原始的工具把先人留下的数据保存完好，本身就是非常困难的事情。如果再想从中提取出信息、总结出规律，更是需要过人的天赋和非凡的运气！所以，祖辈传下的经验（真假掺杂的知识）就显得弥足珍贵，“不听老人言，吃亏在眼前”，靠着这些口耳相传的经验就基本上满足了当时认识世界和改造世界的需要。

20 世纪中期以后，随着信息时代的到来和信息技术的普及，各行各业的数据数量和种类激增，产生了一大堆问题。例如，信息过量，难以消化；鱼龙混杂，真假难辨；形式不一，不好处理……各种信息系统的建立和运行，虽然可以高效地实现数据的录入、查询、统计等功能，但难以发现数据中隐含的关系和规律，无法根据现有的数据预测未来的发展趋势，这就导致了“数据爆炸但知识贫乏”的现象。20 世纪 90 年代，管理大师彼得·德鲁克^①就曾经发出感叹：迄今为止，我们的系统产生的仅仅是数据，而不是信息，更不是知识！

数据挖掘（data mining）就是通过特定的计算机算法来取代人工，对大量的数据进行自动的分析，从而揭示数据之间隐藏的关系、模式和趋势，为决策者提供新的知识。由于早期各行各业的主要数据大都按照固定的格式存储在数据库中，这样也有利于提高计算机处理的效率。所以在某些场合下，数据挖掘也被人们称为数据库中的知识发现（Knowledge Discovery in Database, KDD）。

如图 1.13 所示，可以简单地把数据挖掘理解为“对数据进行挖山凿矿式的开采”，它的主要目的有两个：一是要发现潜藏在数据表面之下的历史规律，二是通过现有数据对未

^① 彼得·德鲁克，现代管理学之父，其著作影响了数代追求创新以及最佳管理实践的学者和企业家们，各类商业管理课程也都深受其思想的影响。



来进行预测。前者称为描述性分析，后者称为预测性分析。在商业应用上，很多超市会从购物记录中挖掘“哪些商品常常会被顾客同时购买？”这就是一种典型的描述性分析；如果通过考察现有的历史数据，以特定的算法估计某种商品下个月的销售量（以确定进货量），则是一种预测性分析了。



图 1.13 从数据中挖山凿矿

数据挖掘把数据分析的范围从“已知”扩大到了“未知”，从“过去”推向了“将来”，这也是商务智能（business intelligence）^①真正的生命力和“灵魂”所在。它的发展和成熟，最终推动了商务智能在各行各业的广泛应用。

利用数据挖掘进行营销策划

零售帝国沃尔玛^②拥有世界上数一数二的数据库系统，也是最早应用数据挖掘技术的企业之一。在一次例行的数据分析之后，研究人员突然发现：跟尿布一起搭配购买最多的商品竟然是啤酒！尿布和啤酒，听起来风马牛不相及，但这是对历史数据进行挖掘的结果，反映的是潜在的规律。于是，沃尔玛随后对啤酒和尿布进行了捆绑销售，并尝试着将两者摆在一起，结果使得两者销量双双激增，为公司带来了大量的利润。后来的跟踪调

^① 1989 年，高德纳咨询公司的德斯纳在商业界给出了“商务智能”的一个正式定义：商务智能指的是一系列以事实为支持、辅助商业决策的技术和方法。

^② 沃尔玛（Wal-Mart），是世界最大的零售商，拥有 8400 多家分店、200 多万雇员（和美国联邦政府的雇员等量齐观），它的收入在 2010 年突破了 4000 亿美元，超过了许多国家的 GDP 总值。

查发现，在美国有孩子的家庭中，太太经常嘱咐丈夫下班后要去超市为孩子买尿布，而30%～40%的丈夫们会在买完尿布以后又顺手买点啤酒犒劳自己……

天睿公司^①与沃尔玛进行合作，从2004年开始对沃尔玛所有的历史交易记录进行整合与分析。发现每次飓风来临，不仅手电筒、电池、水这些商品热销，而且一种袋装小食品“Pop-Tarts”的销量也会明显增加。于是，飓风来袭之前，沃尔玛就提高Pop-Tarts的仓储量，以防脱销，并且把它和水捆绑销售。研究人员后来发现，这个规律的背后原因是：一方面美国人喜欢此类甜食，另一方面Pop-Tarts在停电的时候吃起来非常方便……如果没有数据挖掘，Pop-Tarts和飓风的微妙关系就难以被发现。

^① 天睿公司（Teradata），是美国十大上市软件公司之一，已经成为全球最大的专注于大数据分析、数据仓库和整合营销管理解决方案的供应商之一。