第3章 数据思维

开篇案例

别轻易点赞，它会泄露你的性格秘密

美国科学院院报(PNAS)最新的一篇研究表明，在社交网站上别轻易点赞，因为点赞能够泄露你一些比较私密的性格特质。该项目的研究人员邀请脸书上8.6万名志愿者参与这项性格测试，并且收集了他们的“点赞”数据(即对什么帖子或内容发生点赞行为)。同时邀请了被试的亲朋好友参与测试，给出有关该被试者性格的评价。这样就获得了被试者的三份性格数据，一份是自我的评价，一份是亲朋好友的评价，一份是基于点赞数据计算的结果。研究结果表明，算法得到的性格倾向指数比亲朋好友的判断更为准确。

想一想在点赞的时候，我们希望向脸书好友展示我们对特定内容(包括状态更新、照片、书籍、产品、音乐)的积极态度。与此同时，“点赞”行为也暴露了你很多的私密信息、敏感特质、性格偏好和行为倾向等。例如，宗教信仰、政治观点、性取向和酒量等。具体的结论是，大概只需要10个“赞”，计算机就能比同事更准确地判断你的性格；通过70个“赞”，计算机的判断就能超过你的朋友；140个“赞”便能超过你的家人(父母亲兄妹)。300个“赞”则能“击败”你的伴侣。

这个结论你想到了吗？你的性格密码被大数据“熟知”了会怎样？

学习目标

学完本章，你应该牢记以下概念。

- 统计思维——采样、总体。
- 计算思维——递归、容错(抽象、自动化)。
- 数据思维——全数据思维、相关性思维、容错思维。

学完本章，你将具有以下能力。

- 比较统计思维与数据思维的本质区别。
- 说明计算思维与数据思维是如何解决容错性问题的。

学完本章，你还可以探索以下问题。

- 探究你在某平台/App上留下了哪些数据，你自己的“画像”你了解吗？
- 探究大数据案例中数据思维的具体应用(全数据、相关、容错)。
- 关于“一切皆可量化”，对现在、对未来意味着什么？


视频讲解

3.1 统计学与统计思维

3.1.1 什么是统计

什么是统计？顾名思义，“统”是总括、概括，“计”是计算，合在一起就是概括的计算。所以，统计是指对某个事件进行概括性的计算，以得出支撑结论的统计数据。很久以前，古代人们就掌握了记数的技术，主要用于记录食物的数量。但是随着人们智慧的增长，人们不再局限于记数，对于记录下来的数据，总有人会去探索一些有趣的事情，其中最简

单的一种计算就是均值，计算一组数据的平均数来衡量这组数据的平均水平。有了均值来衡量平均水平，那么人们自然会关注个体与平均水平的差异，这时方差应运而生，基于均值来衡量整体水平之间的差异程度。假设共有 n 个样本数据 $x_i(i=1,2,3,\cdots,n)$，则该样本集的均值、方差(标准差)的计算公式如下。

$$\text{均值：}\overline{X}=\frac{\sum_{i=1}^{n}X_i}{n}$$

$$\text{标准差：}s=\sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}}$$

$$\text{方差：}s^2=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$$

随着统计学继续发展，人们很快发现，之前定义的整体只是当前收集到的全部数据，对于某个事件(偶尔发生的个体为随机变量的具体值)不可能穷尽搜集到它的所有数据，这所有的数据称为总体，之前定义为整体的那部分数据称为这个总体下的一份样本。总体与样本之间的关系如图 3.1 所示。

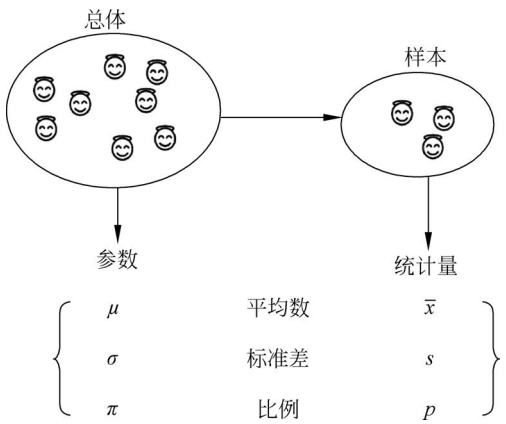


图 3.1　总体与样本

样本的数据表现并不稳定，但是在多次实验的情况下，事件的某种情况发生的频率趋于稳定，结合极限的概念，可以给总体中事件出现的频率一个定义，即“概率”。进而为了理解某个事件的规律，我们希望穷尽事件所有可能的概率，因此需要知道总体数据大概以什么样的方式呈现。为了刻画总体的模样，分布又应运而生，即事件所有可能的概率分布。有了分布的概念，人们开始研究各种不同事件的分布形式，进化出 0-1 分布(伯努利分布)、二项分布、泊松分布、指数分布、正态分布等。正态分布的发现是一个里程碑式的事件。正态分布曲线形状优美，与其对应的密度函数的数学表达及分布图如图 3.2 所示。正态分布的期望值 μ(随机变量的均值)可解释为位置参数，决定了分布的位置；

其方差 σ^2 的平方根或标准差 σ 可解释为尺度参数，决定了分布的幅度。

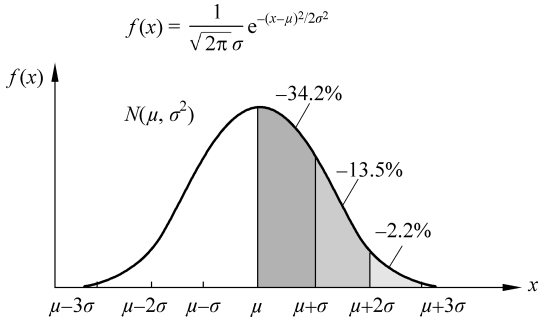


图 3.2　随机变量 x 的正态分布

总之，在统计学家眼里，世上所有发生的事件都是随机的，但所有的随机事件都可以用概率分布来描述。

应用案例 3.1：面包的故事

由于战争，德国有一个时期物资特别紧缺，对面包实行配给制：政府把面粉发给指定的面包房，面包师傅烤好了面包再发给居民。有一个统计学家，怀疑他所在区域的面包师傅私扣面粉，于是就天天称自己的面包。几个月以后，他去找面包师傅，说："政府规定配给的面包是 400g，因为模具和其他因素，你做的面包可能是 398g、399g，也可能是 401g、402g，但是按照统计学的正态分布原理，这么多天的面包质量平均应该等于 400g，可是你给我的面包平均质量是 398g。我有理由怀疑是你使用较小的模具，私吞了面粉"。面包师傅承认确实私吞了面粉，并再三道歉保证马上更换正常的模具。又过了几个月，统计学家又去找这个面包师傅，说："虽然这几个月你给我的面包都在 400g 以上，但是这可能是因为你没有私吞面粉，也可能是因为你从面包里特意挑大的给我。同样根据正态分布原理，这么多天不可能没有低于 400g 的面包，所以我认为你只是特意给了我比较大的面包，而不是更换了正常的模具。我会立刻要求政府检查你的模具"。面包师傅只好当众认错道歉，接受处罚。

这个故事用到了正态分布原理，是不是很有趣啊？其实统计学离我们的生活并没有那么遥远，很多时候可以利用统计学解决一些生活中的小问题。

3.1.2　统计学原理与统计思维

统计学源于人们对客观现象的描述及分析，标志从理论数学到统计学、从确定到不确定、从部分推整体、从演绎到归纳的思维变化，而基本统计量如均值、方差、概率分布等是这一思维实施的方法及工具。统计学将现实世界的问题抽象成某种描述（模型），借此就可能发现问题的本质及其能否求解。

统计学的精髓就在于通过可以观测的样本来推测总体的情况。统计学有三大基石，即大数定律、中心极限定理及正态分布。在统计活动中，人们发现，随着实验次数的增加，一个事件发生的概率会收敛于一个稳定的值。在数学上的表达就是，n 个随机变量的均值（或者说期望）会随着 n 趋近于无穷而收敛于总体均值，也就是实际的均值。

技术洞察 3.1：大数定律与中心极限定律——统计学的基石

大数定律表明：在实验不变的条件下，重复实验多次，随机事件的频率近似于它的概率。随着样本的增大，随机变量对平均数的偏离是下降的。

大数定律解决了样本和总体的关系问题，其核心思想就是当样本量足够大的时候，样本的分布（均值）与总体的分布（真实均值）充分接近，也就是可以把两者视为相等的。大数定律告诉我们只要获取适合的数据样本量就可以把握住事物的分布规律，而不需要所谓的海量数据。关键是数据样本的代表性、数据的真实性、有效性以及适合的样本量。大数定律反映了一个自然规律：在一个包含众多个体的大群体中，偶然性而产生的个体差异，使得个体都是毫无规律、难以预测的，但由于大数定律，整个群体能呈现出稳定的形态。但要注意的是，大数定律仅在样本数量足够多的情况下才成立。

中心极限定律表明：在随机变量的个数无限多的时候，随机变量的分布会趋近于正态分布，并且这个正态分布以 μ 为均值，以 σ^2/n 为方差。

综上所述，这两个定律都是在阐述样本均值性质。随着 n 增大，大数定律表明：样本均值几乎必然等于均值。中心极限定律则进一步阐述，它越来越趋近于正态分布，并且这个正态分布的方差越来越小。

想一想：

“数据富足”时代，这两个定律给你的启发是什么？

统计学是研究如何有效收集、整理、分析数据的一门学科，它以数据为研究对象，以统计描述、统计建模和统计推断等方式分析处理数据，是数据科学最重要的理论基础与方法论。

统计描述是利用各种数学方法对数据的结构和特征进行描述的方法。常用的统计描述方法包括统计图表、分布函数、数字特征等。统计推断是用样本推断数据总体分布或分布的数值特征的统计方法，而统计建模主要是指如何选择合适的模型（分布）去描述给定的数据和数据与数据之间的关系，内容涉及变量选择（或称特征选择）、模型构造与模型选择等方面。

简单来说，统计学所做的工作就是从随机性中寻找规律性，这是统计的基本思想，也是统计的魅力所在。统计学里所表达的两个核心理念就是：允许误差下的概率保证及允许误差下的统计推断。概率和误差构成了统计思维的两大支柱，并发展出统计学里几乎所有的关键要点。

3.1.3 像统计学家一样思考

统计思维从属于一般思维，是人脑和统计学原理、方法、统计学工具交互作用并按照一般思维规律认识各类现象的内在的批判性思维活动。它是一种思维方式、行为方式、工作方式及决策方式。统计思维最终要指导人们如何和数据打交道，解决客观现实问题。思维的基本特点体现在数量性、总体性、客观性、历史性、对比性、综合性、具体性、创造性和实用性等。

统计学家思考的前提是基于数据不完全或者数据过于庞大，导致无法全面研究和分析的情况。利用事物或数据整体（总体）与部分（样本）之间的内在联系通过研究挖掘部分数据的某些特征达到推测整体特征的目的（抽样分析法），根据对采取样本进行分析而

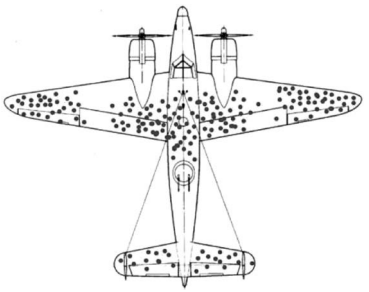
推测总体的结论。采样的绝对随机性成为随机采样成功的关键因素，但保证绝对的随机性是非常困难的事情。要想保证结果的可信性，往往对样本有严格的限制（或假设），如独立同分布等。

应用案例 3.2：幸运者偏差

1941 年第二次世界大战中，盟军的战机在多次空战中损失严重，无数次被纳粹炮火击落，盟军总部秘密邀请了一些物理学家、数学家以及统计学家组成了一个小组，专门研究“如何减少空军被击落概率”的问题。当时军方的高层统计了所有返回飞机的中弹情况，发现飞机的机翼部分中弹较为密集，而机身和机尾部分则中弹较为稀疏，于是当时的盟军高层的建议是加强机翼部分的防护。但这一建议被小组中的一位来自哥伦比亚大学的统计学教授沃德（Abraham Wald）驳回了，沃德教授提出了完全相反的观点：加强机身和机尾部分的防护。

那么这位统计学家是如何得出这一看似不够符合常识的结论的呢？沃德教授的基本出发点基于以下三个事实。

- 统计的样本只是平安返回的战机。
- 被多次击中机翼的飞机，似乎还是能够安全返航。
- 而在机身机尾的位置，很少发现弹孔的原因并非真的不会中弹，而是一旦中弹，其安全返航的概率极小，即返回的飞机是幸存者，仅依靠幸存者做出判断是不科学的，那些被忽视了的非幸存者才是关键，它们根本没有回来！



军方采用了教授的建议，加强了机尾和机身的防护，并且后来证实该决策是无比正确的，盟军战机的击落率大大降低。

可见，统计学中样本的随机性如此重要！

统计学里的思维方法和人们的思维方式有一定的对应关系。具备统计思维的第一步就是要求假设任何随机现象都是服从某一分布的，有了这个认识才能去做出后续的判断。统计思维的实践体现在以下几点。

（1）要有善于利用数据的思维。做决策要有数据，每一项数据，都可能是有用的信息。统计学家要善于运用数据，具有对数据的“敏感性”。

（2）要有善于捕捉不确定性的思维。宇宙的运转，必然性与随机性交错着进行。人们对未来，知道大致会发生哪些事，以及何时发生，但又不能完全掌握。由于不确定性的存在，人们所能做的，就是要了解它，很多时候还要设法减少这些不确定性。因此，先辈针对随机的世界，总结了一些所谓的法则来应对这样的不确定性。例如大数定理及中心极限定理。在统计里做预测和估计，本质上是在做以偏概全的事。虽偏却能概全，这是统计学家的本领。

（3）要有相信概率的思维。数学家拉普拉斯（Pierre-Simon Laplace）曾说过，“大部分生活中最重要的疑问，都只是概率的问题”。在随机世界里，通常以“相同的可能性”来解释概率。在随机的世界里，要相信概率，而不是要挑战概率。

（4）要有合理估计的思维。随着统计学的发展，各种估计方法百家争鸣。这些有道

理的估计方法，往往有各自的优点，并且适用于某些场合，不会有哪种方法永远是最佳的。例如，有时觉得给个范围能更清楚地描述，这就是著名的置信区间估计方法。

(5) 要有疑罪从无的假设检验思维。英文中的假设 Hypothesis 一词，是由古希腊文 Hypotithenai 演变而来，科学上的假说(或称假设学说)也是这个词。在数学里，常在证明一个命题是真或伪。但在随机世界中，很多现象都只能视为假设，就看更愿意接受哪一个。接受不表示就完全相信该假设为真，拒绝也不表示该假设为伪。统计里的假设，经检定后，不论接受哪一个，都无法让该假设成为定律，假设永远是假设。

统计分析的局限性在于，采样数据更适用于宏观分析，在微观领域能发挥的作用有限。不能掌握全面的数据，不能适应算力、存储能力、传输能力高速发展的今天。因此在大数据时代背景下，统计学的知识体系需要一定程度的调整，统计学本身的理念是注重方式方法的，为数据科学进行数据价值化奠定了一定的基础。而大数据催生出的数据科学则更关注整个数据价值化的过程，数据科学不仅需要统计学知识，还需要数学知识和计算机知识。

技术洞察 3.2：统计描述与统计推断

统计描述与统计推断是统计学中常用的词汇，百度百科给出的定义是：描述统计学是研究如何取得反映客观现象的数据，并通过图表形式对所搜集的数据进行加工处理和显示，进而通过综合概括与分析得出反映客观现象的规律性数量特征的一门学科。推断统计学是研究如何根据样本数据去推断总体数量特征的方法，它是在对样本数据进行描述的基础上，对统计总体的未知数量特征做出以概率形式表述的推断。想一想以下两类问题的异同点。

统计描述问题：

- 样本中家庭年观测的收入是不是无偏差的？
- 某产品在不同区域的月销售量均值/方差是多少？
- 变量的量级差异大吗？(决定是否需要对数据标准化。)
- 使用模型中的预测变量缺失情况如何？
- 问卷调查回复者的年龄分布范围是多少？

统计推断问题：

- 参与促销活动和没有参与促销活动的消费者购买量有差异吗？
- 男性是不是比女性更倾向于购买我们的产品？
- 用户满意度在不同商业区是不是有不同？

需要注意的是，数据挖掘及机器学习算法同样能够解决统计推断问题，旨在进行精确预测，而统计学处理问题从严格的统计假设开始，主要用于推断变量之间的关系。

思考题

1. 什么是样本？统计学为什么需要采样？对采样的基本要求(假设)是什么？
2. 大数定律的现实意义是什么？结合正态分布图进行解释说明。
3. 什么是统计思维？解释说明什么是统计描述、统计推断、统计建模。


视频讲解

3.2 计算机与计算思维

3.2.1 计算与自动计算

计算是指由数据和运算符形成的运算式，按运算符的计算规则对数据进行计算并获得结果。如我们从幼儿开始就学习和训练的算术运算：

$$3+2=5,\quad 3\times 2=6,\quad 8-3=5,\quad 8-(3\times 2)=2$$

在这里不断学习和练习的内容包括两方面：一是用各种运算符组合来表达对数据的变换，即熟悉各种运算式；二是能够按照运算符的计算规则对前述运算式进行计算并得到正确的结果。这种运算式的计算是需要人来完成的，可以被称为“人”计算。

广义地讲，一个函数 $f(x)$（如正态分布函数）就是把 x 变成 $f(x)$ 的一次计算。在高中及大学阶段，我们也是不断学习各种函数及其计算规则并应用这些规则来求解各种问题，得到正确的计算结果，如对数与指数函数、微分与积分函数等。

计算规则可以学习及掌握，但应用计算规则进行计算可能超出了人的计算能力，即人知道规则但却没有办法得到计算结果。自动计算就是让机器来完成计算，即用机器来代替人类按照计算规则自动计算，这就是计算机科学家要研究的内容，即怎样实现自动计算。

技术洞察 3.3：“人”计算与“机器”计算的思维差异

人和机器是如何求解一元二次方程 $ax^2+bx+c=0$ 的整数解呢？如果是“人”计算，则可以直接利用公式 $x=\dfrac{-b\pm\sqrt{b^2-4ac}}{2a}$ 进行求解。如果是“机器”计算，则采取如下方法：从 $-n$ 到 n，产生 x 的每一个整数值，将其依次代入到方程中，如果其值使方程成立，则该值即为其解。

进一步思考可以发现，“人”进行计算，计算规则可能很复杂，如求根公式，但计算量可能很小，只需按照求根公式计算一次即可。人需要知道数据的计算规则才能完成计算（这是数学家要提供的），有时人所应用的规则只能满足特定方程的求解，如上述公式可求解一元二次方程，但却不能应用于一元三次方程或一元任意次方程。而“机器”进行计算，规则可能很简单，只需要简单加减乘除的运算，但计算量却很大，有多少个 x 值就需要按照方程重复计算多少次。机器使用的方法可以应用于一元任意次方程，并不限于一元二次方程。

自动计算是由计算机来实现的。1642 年，法国科学家帕斯卡发明了著名的帕斯卡机械计算机，它告诉人们用“纯机械装置可代替人的思维和记忆”，开辟了自动计算的道路。1854 年，布尔基于二进制创立了布尔代数，为一百年后的数字计算机的电路科技提供了重要的理论基础。

正是由于前人对机械计算机的不断探索与研究，不断追求计算的机械化、自动化、智能化，即如何能够自动存储数据？如何能够让机器识别可变化的计算规则并按照规则执行计算？这些问题促进了机械技术和电子技术的结合，最终导致了现代计算机的

出现。现代计算机基于二进制，设计了能够理解和执行任意复杂计算的程序，如数学计算、逻辑推理、图像图形变换、数理统计、人工智能与问题求解，计算机的功能在不断提高。

3.2.2 算法与程序

算法是计算机和软件的灵魂。算法是指解题方案的准确而完整的描述，是一系列解决问题的清晰指令，算法代表着用系统的方法描述解决问题的策略机制。也就是说，能够对一定规范的输入，在有限时间内获得所要求的输出。如果一个算法有缺陷，或不适合于某个问题，执行这个算法将不会解决这个问题。不同的算法可能用不同的时间、空间或效率来完成同样的任务。一个算法的优劣可以用空间复杂度与时间复杂度来衡量。

算法的5大特征如下。

(1) 有穷性：算法必须能在执行有限个步骤之后终止。

(2) 确切性：算法的每一步骤都必须有确切的定义。

(3) 输入项：一个算法有一个或多个输入，以刻画运算对象的初始情况，所谓0个输入是指算法本身定出了初始条件。

(4) 输出项：一个算法有一个或多个输出，以反映对输入数据加工后的结果。没有输出的算法是毫无意义的。

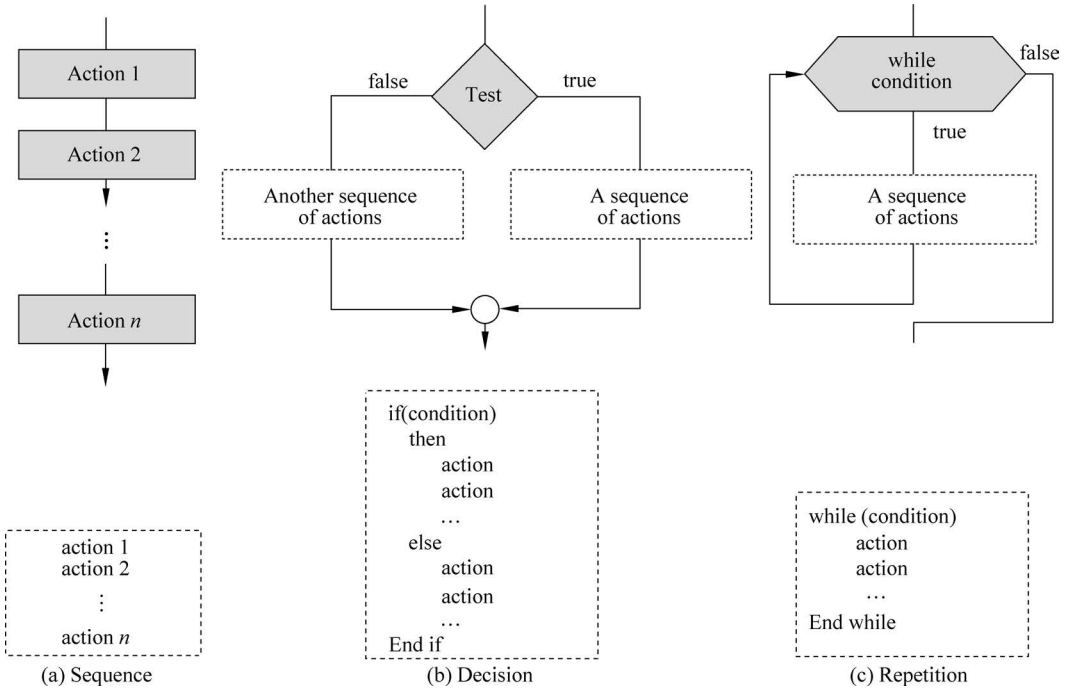
(5) 可行性：算法中执行的任何计算步骤都可以被分解为基本的可执行的操作步，即每个计算步都可以在有限时间内完成(也称为有效性)。

程序是由若干指令构造的一个指令组合或一个指令序列，使外界使用者用于表达其期望计算系统实现的千变万化功能的一种手段。计算系统应该是能够执行程序的系统，程序用计算机语言实现求解某些问题的算法。“是否会编程序”本质上讲，首先是能否想出求解问题的算法，其次才是将算法用计算机可以识别的形式(程序)书写出来。

技术洞察 3.4：三种基本算法的结构及流程

任何简单或复杂的算法都可以由顺序结构、选择结构和循环结构这三种基本结构组合而成，三种基本结构的流程如下图所示。

顺序结构是最简单的程序结构，程序中的各个操作是按照它们在源代码中的排列顺序，自上而下，依次执行，流程如图(a)所示。选择结构用于判断给定的条件，进而控制程序的流程。它会根据某个特定的条件进行判断后，选择其中一支执行，流程如图(b)所示。循环结构是指在程序中需要反复执行某个或某些操作，直到条件为假或为真时才停止循环一种程序结构。它由循环体中的条件判断继续执行某个功能还是退出循环，流程如图(c)所示。



构造与设计算法需要从问题本身来挖掘求解的思想。各学科利用计算系统进行问题求解的关键是发现构造与设计求解问题的算法，包括以下两点：构造与设计在有限的时间内可以执行的算法；构造与设计尽可能快速的算法。不同环境可能产生不同的算法，不同的审视问题的视角也可能产生非常简单但却很重要的算法。社会自然中的问题求解同样有助于产生计算问题的求解算法。将具体问题抽象出其数学模型，更是有利于算法的发现与构造。

3.2.3 什么是计算思维

正如数学家在证明数学定理时有独特的数学思维、工程师在设计制造产品时有独特的工程思维、艺术家在创作诗歌音乐绘画时有独特的艺术思维一样，计算机科学家在用计算机解决问题时也有自己独特的思维方式和解决方法，统称为计算思维(Computational Thinking)。从问题的计算机表示、算法设计直到编程实现，计算思维贯穿于计算的全过程。计算思维是运用计算机科学的基础概念去求解问题、设计系统和理解人类行为的涵盖了计算机科学广度的一系列思维活动。

基于上述定义，可以挖掘出如下三个层次的内涵。

(1) 求解问题中的计算思维。利用计算手段求解问题的过程是：首先要把实际的应用问题转换为数学问题，可能是一组偏微分方程(Partial Differential Equations，PDE)；其次将 PDE 离散为一组代数方程组；然后建立模型、设计算法和编程实现；最后在实际

的计算机中运行并求解。前两步是计算思维中的抽象，后两步是计算思维中的自动化。

(2) 设计系统中的计算思维。R. Karp 认为：任何自然系统和社会系统都可视为一个动态演化系统，演化伴随着物质、能量和信息的交换，这种交换可以映射为符号变换，使之能用计算机实现离散的符号处理。当动态演化系统抽象为离散符号系统后，就可以采用形式化的规范来描述，通过建立模型、设计算法和开发软件来揭示演化的规律，实时控制系统的演化并自动执行。

(3) 理解人类行为中的计算思维。计算思维是基于可计算的手段，以定量化的方式进行的思维过程。计算思维就是能满足信息时代新的社会动力学和人类动力学要求的思维。在人类的物理世界、精神世界和人工世界三个世界中，计算思维是建设人工世界所需要的主要思维方式。

试一试 3.1：排序算法——计算思维的实践

排序是现实世界中常见的问题，其本质是对一组对象按照某种规则进行有序排列的过程。通常是把一组对象整理成按关键字递增(或递减)的排列，关键字是对象的一个用于排序的特性。

一个升序排序算法的简单描述是：对给定的一个数据表，算法从第一个元素开始扫描整个列表，找到最小的元素，并将其与第一个位置的元素交换。然后算法从第二个位置的元素开始扫描剩下的列表，找到次小的元素，并将其与第二个位置的元素交换，如此循环，直到排完所有的元素。

排序算法的伪代码如下。

```
BubbleSort(A[1..n]) {
  for i = 1 to n do
    for j = i + 1 to n do
      if (A[i] > A[j]) then
        swap(A[i], A[j]);   //数据交换
}
```

初始排序数据：[49 78 65 97 36 13]
第一轮排序后：13 [78 65 97 36 49]
第二轮排序后：13 36 [65 97 78 49]
第三轮排序后
第四轮排序后
第五轮排序后
最后排序结果

假设数据表 A[i]如右侧所示，$n=6$，试着写出右侧几轮排序的结果。

你能理解什么是算法吗？你体会到循环结构的算法在计算机中是如何运行的吗？

你能体会到计算思维中提到的“抽象”和“自动化”吗？

想一想

还有哪些地方的数据需要排序？

学习计算思维，就是学会像计算机科学家一样思考和解决问题。计算思维的本质是抽象(Abstract)和自动化(Automation)。它反映了计算的根本问题，即什么能被有效地自动进行。

计算是抽象的自动执行，自动化需要某种计算机去解释抽象。从操作层面上讲，计算就是如何寻找一台计算机去求解问题，隐含地说就是要确定合适的抽象，选择合适的计算机去解释执行该抽象，后者就是自动化。

与数学相比，计算思维中的抽象显得更为丰富，也更为复杂。数学抽象的特点是抛开现实事物的物理、化学和生物等特性，仅保留其量的关系和空间的形式。而计算思维

中的抽象却不仅如此。例如，算法也是一种抽象，也不能将两个算法简单地放在一起构建一种并行算法。

抽象层次是计算思维中的一个重要概念，它使人们可以根据不同的抽象层次，进而有选择地忽视某些细节，最终控制系统的复杂性。在分析问题时，计算思维要求将注意力集中在感兴趣的抽象层次或其上下层，还应当了解各抽象层次之间的关系。

计算思维中的抽象最终是要能够机械地一步一步自动执行的。为了确保机械地自动化，就需要在抽象过程中进行精确、严格的符号标记和建模，同时也要求计算机系统或软件系统生产厂家能够向公众提供各种不同抽象层次之间的翻译工具。因此，从思维的角度看，计算科学主要研究计算思维的概念、方法和内容，并发展成为解决问题的一种思维方式，极大地推动了计算思维的发展。

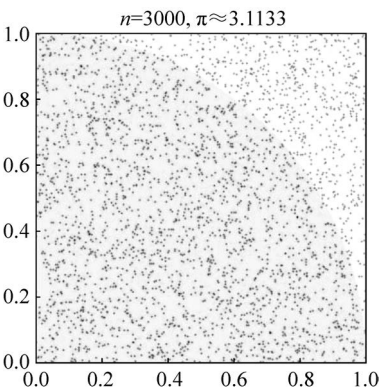
技术洞察 3.5：蒙特卡罗方法——统计模拟法

蒙特卡罗方法(Monte Carlo Method)是一种“统计模拟方法”。20 世纪 40 年代，为建造核武器，冯·诺依曼等人发明了该算法。因赌城蒙特卡罗而得名，暗示其以概率作为算法的基础。

假设要计算一个不规则形状的面积，只需在包含这个不规则形状的矩形内，随机地掷出一个点，每掷出一个点则 $N+1$，如果这个点在不规则图形内则 $W+1$，落入不规则图形的概率即为 W/N。当掷出足够多的点之后，可以认为：不规则图形面积＝矩形面积$\times W/N$。

要应用蒙特卡罗算法的问题，首先要将问题转换为概率问题，然后通过统计方法将其问题的解估计出来。蒙特卡罗方法基本思想就是：当所求解问题是某种随机事件出现的概率，或者是某个随机变量的期望值时，通过某种“实验”的方法，以这种事件出现的频率估计这一随机事件的概率，或者得到这个随机变量的某些数字特征，并将其作为问题的解。该方法的理论基础是中心极限定理，样本数量越多，其平均就越趋近于真实值。不断抽样，逐渐逼近。

π 是一个无理数，没有任何一个精确公式能够计算出来，只能采用近似计算。通过蒙特卡罗算法求 π 的示例如图所示，图中构造了一个正方形和一个 1/4 单位圆，往整个区域随机投入点，根据点到原点的距离判断是落在圆内还是圆外，从而根据落在不同区域点的数目，求出两个区域的比值，进而可以求出 1/4 单位圆的面积，再进一步可以求出圆周率 π。图中给出了模拟 3000 次 π 的结果，如果模拟 100 000 次，得到 π 的值是 3.140 76(注意，这个值每次模拟是不确定的)。



蒙特卡罗方法是统计思维与计算思维的完美结合，你体会到了吗？用统计思维再想一想，为什么每次模拟计算的结果会不一样呢？这样还有意义吗？

3.2.4 像计算机专家一样思考

计算机专家的思考方式是将数学、工程学和自然科学中一些最好的特征结合在一起。像数学家一样，计算机专家使用形式化语言来表达思想（即语义符号化、逻辑的抽象）。像工程师一样，他们设计事物，把部件装配为系统，在候选方案中寻求一种平衡。像科学家一样，他们观察复杂系统的行为，形成假设，并对其进行检验。

计算思维强调可行与实践，追求“可交付、可使用”。和哲学式的探究与纯逻辑的符号推导不同，计算思维强调实践性——解决方案不是理论正确就好了，要在实际中可行才可以。这一特点是由这一思维的诞生背景所决定的，当计算机科学家处理问题时，除了要知道如何将一个问题抽象为计算机能够理解的可计算模型，还要能够将计算收敛到有限空间中得到结果。如果算法的时空复杂度过大，以当前的算力在有效求解的时间内无法得出结果，那么再完美的理论算法也无法在现实中奏效。计算思维能够让我们明白正确性和可行性的关系，明白“实验室结果”和“日常使用效果”的必然差距。

归纳起来，计算机思维的特点体现在以下几个方面。

（1）通过简约、嵌入、转换和仿真的方法，把一个看起来困难的问题变成一个可计算的解决方案。

（2）一种递归思维、一种并行处理，采用抽象和分解来控制庞杂的任务。

（3）按照预防、保护的原则，通过冗余、容错、纠错的方法，并从最坏情况进行系统恢复及维护。

（4）在时间和空间之间、在处理能力和存储容量之间进行折中的思维方法。

技术洞察 3.6：计算中的递归与迭代

递归（Recursion）：常被用来描述以自相似方法重复事物的过程，在数学和计算机科学中指的是在函数定义中使用函数自身的方法（A 调用 A）。递归是一个树结构，从字面可以理解为重复“递推”和“回归”的过程，当“递推”到达底部时就会开始“回归”，其过程相当于树的深度优先遍历。

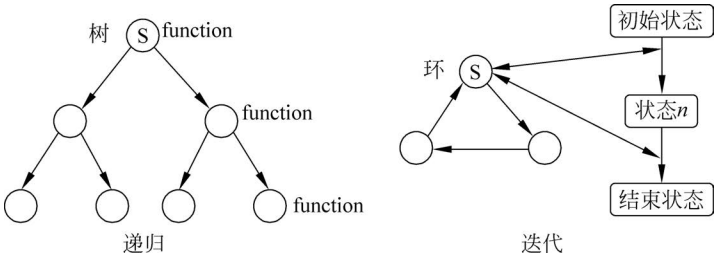
迭代（Iteration）是一种重复反馈过程的活动，每一次迭代的结果会作为下一次迭代的初始值（A 重复调用 B）。迭代是一个环结构，从初始状态开始，每次迭代都遍历这个环，并更新状态，多次迭代直到到达结束状态。

关于阶乘的计算有两种表示方式：递归与迭代。你发现了吗？递归中一定有迭代，迭代中不一定有递归。多数情况下上二者可以相互转换。计算机所实现的编程算法也是这样实现的。

$n!=n\times(n-1)!$

$n!=n\times(n-1)\times(n-2)\times(n-3)\times\cdots\times3\times2\times1$

从直观上讲，递归是将大问题转换为相同结构的小问题，从待求解的问题出发，一直分解到已经已知答案的最小问题为止，然后再逐级返回，从而得到大问题的解（自上而下）。而迭代则是从已知值出发，通过递推式，不断更新变量新值，一直到能够解决要求的问题为止（自下而上）。



想一想：

“思维决定行动!”计算思维决定计算机的编程实现。

思考题

1. 什么是算法？简述算法的 5 大特征是什么。
2. 三个典型算法的基本流程是什么？排序算法中使用了哪种流程？
3. 什么是计算思维？关于“抽象”与“自动化”你能举例说明吗？

3.3 大数据与数据思维

3.3.1 数据思维的特点

数据科学需要全新的数据思维，有其鲜明的“以数据为中心”的特点。“数据是 21 世纪的石油”这句话已经充分说明了数据的价值。大数据时代信息的不断整合及分析，已然使得信息、数据量化及互联转变为多维度的发展状态。换言之，数据思维渗透至各个领域及行业的不同维度，是大数据发展的初始动机和直接目的。


视频讲解

1. 整体性——整体反映全貌

从基本特征层面分析，数据思维的主要特征之一就是整体性(全数据)。整体性是相对于系统的部分或者元素来讲的，整体性是事物系统的本质特性，没有整体性就无法维持系统自身的存在及其发展。系统的自身性质及其功能由自身系统的整体性赋予。每种数据来源均有一定的局限性和片面性，事物的本质和规律隐藏在各种原始数据的相互关联之中，只有融合、集成各方面的原始数据，才能反映事物全貌。因此，以整体性的思维把握事物本身，才能真正客观而全面地把握对象的真实本质及其变化发展的趋势。

2. 动态性——动态多维多层

世界事物的本原是以多维状态和层次形态呈现，传统的静态思维只是一维结构，无形中制约了人类对数据价值的判断和更高层次的认知。

采用动态观点在同一时间从多个角度看问题，则可以正确看待各类数据存在的价值。这种模糊、非确定、灵活且立体型的思维决定了在多个维度上。事物亦此亦彼，亦黑亦白，即没有绝对的对错判断，必须结合具体问题和背景环境才能做出对错判断。数据思维摆脱了静态思维的束缚，从动态视角多维且多层次认知数据的价值，从而进一步接

近事实真相，更全面地认识世界。

3. 相关性——泛在的相关性

数据思维的互联性源于事物泛在的相关关系，任何一个事物都有其内部结构，且与同一系统内的其他事物存在广泛的联系，这种泛在的相关关系要求在面对问题时具备相关思维。大数据作为由各种数据构成相互联系的整体，在数据相互作用的状态中生存和发展。

相关性思维将事物与其周边事物联系起来进行考察，既注重内部各部分数据之间的相互作用关系，又重视大数据与其外部环境的相互作用关系。通过数据的重组、扩展和再利用，突破原有的框架，开拓新领域，发掘数据蕴含的价值。

4. 多样性——多维多角度

数据思维的多样性特征是通过数据种类的不同体现的。关系数据库中存储的基本是结构化数据，而非关系数据库中存储的多源异构数据成为数据思维多样性的主要来源。多样性并不仅存在于大数据领域，人类生活的方方面面均存在多样性。因此，应尽可能全方位把握多样性的存在，搞清楚多样性在数据思维中的具体表现，为利用数据思维奠定基础。

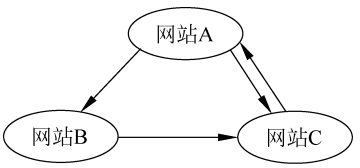
5. 量化互联性——要么数字化，要么死亡

“不论是有形之物还是无形之物，一切皆可量化”，这是道格拉斯·W.哈伯德在《数据化决策》中的名句。知名投资人孙正义也认为：“要么数字化，要么死亡”。数字化成为时代发展的必然趋势，而量化思维是数字化的必然思维结果。

量化可以解释为使用共性语言描述和解释世界的一种方式，体现在充分运用最新的技术手段，对于各个领域进行信息全面定量采集以及信息互通，打通信息间的隔阂，并进行全新的信息整合，实现分析实用性及数据科学性，创造更具价值的数据应用和信息资产。

试一试 3.2：网站重要性度量

PageRank 算法的核心是“网站重要性”度量，以一个“权重”来表示，该算法的“精妙”之处在哪里呢？以一个小型网络为例，首先给每个网站设定相同的权重。然后，让我们把网站想象成一个桶，给每个桶里放 8 个球，表示网站的初始权重相同（第 1 列）。现在，每个网站必须将球交给它链接的其他网站，如果链接多个网站，那么就将球均分给那些网站。如图所示，由于网站 A 链接了网站 B 和网站 C，它将为每个网站提供 4 个球；而网站 B 只链接了网站 C，它就需要将拥有的 8 个球全部放入网站 C 的桶中。第 1 轮分配后（第 2 列），网站 C 得到的小球数最多。



但是我们需要继续重复这个分配过程，因为现在位于最高排名的网站 C 链接了网站 A，所以又会产生新的分配结果。9 轮重复分配过程中各网站小球数量的变化情况如下表，表明网站 A 及网站 C“重要性”都很高，而网站 B 则最低。你也不妨试着算一下，把空白列补齐。

	第 1 轮	第 2 轮	第 3 轮	第 4 轮	第 5 轮	第 6 轮	第 7 轮	第 8 轮	第 9 轮
网站 A	8	8	12	8		10	9		9.5
网站 B	8	4	4	6		5	5		5
网站 C	8	12	8	10		9	10		9.5

想一想：

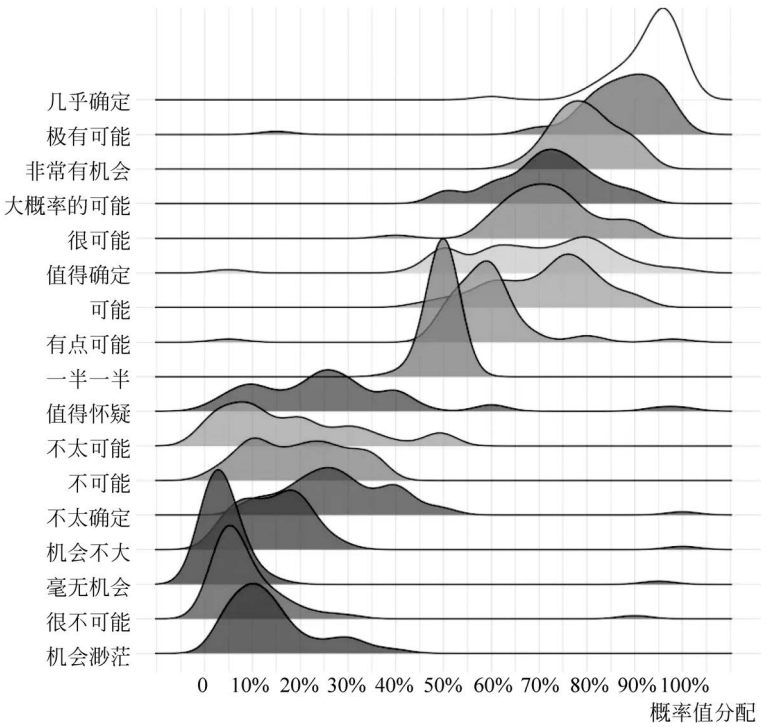
PageRank 算法还可以用到哪里？

3.3.2 一切皆可量化

大数据发展的核心动力来源于人类测量、记录和分析世界的渴望。数据无处不在，它们躲在暗处嘲笑不会善加利用的人们，真相往往隐藏在数字的排列组合里。数据看似枯燥而烦琐，但它们是指向真相的最佳路径。如果一切皆可通过合理方法予以量化，那么就可以说“认识”了这个世界。不论是有形之物还是无形之物，一切皆可量化，量化是一切决策的有益助手，甚至包括婚姻、感情、幸福。量化一切，是数据化的核心，也是大数据时代的基石。

想一想 3.1：文字“可能”“差不多”等词可以量化吗

想一想，我们通常说的程度用词可以定量描述吗？也就是说，可以将定性的描述变换成定量的描述吗？如果能，通过什么方法？这样的定量描述准确吗？假如有大量的样本数据后会更准确吗？这样的定量描述（概率统计分布描述）有应用价值吗？用“迭代思维”再想一想。



当文字变成数据时，数字图书馆孕育而生；当方位变成数据时，GPS 系统横空降世；当沟通变成数据时，Twitter 家喻户晓。我们所有的行为、兴趣爱好甚至是情绪都在不知不觉中被记录，成为数据并组成信息。不同的商业目的会截取不同的信息，再进行交叉组合，事实上，我们也越来越发现，搜索引擎的推送更加贴心精准。这一切都是大数据的功劳。

谈到量化的概念，其核心就是“减少不确定性”，但没有必要完全消除不确定性，“不求精确，但求有效”是概率思维的核心体现。一般来说，量化方法就隐藏在量化目标中，一旦管理者弄清楚要量化什么以及被量化的事物为什么重要，就会发现事物显现出更多可量化的方面。确定真正要量化什么，是几乎所有科学研究的起点。商业领域的管理者需要认识到，某些事物看起来完全无形无影，只是因为你还没给所谈论的事物下定义。

试一试 3.3：余弦定理与文本相似度

一段文本的相似度可以通过统计词频然后通过比较词频向量的余弦距离计算其相似程度。其计算公式如下，其中，A 与 B 分别是词频统计的结果。

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

以如下两个用户购物后的评价信息（文字组）为例，你认为二者的相似度多少？应该如何计算呢？

- 句子 A：这只皮靴号码大了。那只号码合适。
- 句子 B：这只皮靴号码不小，那只更合适。

文本相似度算法实现的步骤如下。

第一步，分词：句子 A：这只 / 皮靴 / 号码 / 大了。/ 那只 / 号码 / 合适

句子 B：这只 / 皮靴 / 号码 / 不 / 小，/ 那只 / 更 / 合适

第二步，列出所有的词：这只、皮靴、号码、大了、那只、号码、合适、不、小、更

第三步，计算词频：句子 A：这只 1，皮靴 1，号码 2，大了 1，那只 1，合适 1，不 0，小 0，更 0

句子 B：这只 1，皮靴 1，号码 1，大了 0，那只 1，合适 1，不 1，小 1，更 1

第四步，写出词频向量：句子 A：(1,1,2,1,1,1,0,0,0)

句子 B：(1,1,1,0,1,1,1,1,1)

第五步，按照相似度公式，计算相似度为 0.70 即 70%的相似度。

想一想：

如果评价的内容（文字）稍作变化，试计算一下它们的相似度会有什么变化。

用户画像是“一切皆可量化”的最好范例。用户画像的概念，最早由交互设计之父 Alan Cooper 提出，是对产品或服务的目标人群做出的特征刻画。在早期，也就是用户数据的来源渠道比较少，数据量也相对比较小的时期，用户画像的研究主要基于统计分析层面，通过用户调研来构建用户画像标签。近年来，随着互联网海量数据的爆炸式增长，众多企业的用户画像研究有了新的机遇。目前，用户画像泛指根据用户的属性、用户偏好、生活习惯、用户行为等信息而抽象出来的标签化用户模型。通俗地说，就是给用户打标签。而标签是通过对用户信息分析而来的高度精炼的特征标识。通过打标签可以利用一些高度概括、容易理解的特征来描述用户，可以让人更容易理解用户，并且可以方便计算机处理。用户画像的构建及应用全过程较好地体现了数据思维的 6 大特点。

技术洞察 3.7：用户偏好计算——TF-IDF

TF-IDF(Term Frequency-Inverse Document Frequency)是一种用于文本分析的常用加权技术。TF是词频，IDF是逆文本频率指数。TF表述的核心思想是，在一条文本中反复出现的词更重要。而IDF的核心思想是，在所有文本中都出现的词是不重要的，IDF用于修正TF所表示的计算结果。

$$\mathrm{TF}=\frac{\text{该词语在文本中出现的次数}}{\text{文本的总词数}}$$

$$\mathrm{IDF}=\log\left(\frac{\text{文本总数}}{\text{出现该词语的文本数}+1}\right)$$

$$\mathrm{TF-IDF}=\mathrm{TF}\times\mathrm{IDF}$$

某电商平台用户A、用户B的浏览记录如下所示。

用户A	
浏览记录1	白色 短袖 女 XXL 可爱
浏览记录2	黑色 长袖 女 XL 皮卡丘 宠物小精灵
浏览记录3	黑色 短袖 男 XL 哪吒 国潮

用户B	
浏览记录1	白色 短袖 女 XL 职业
浏览记录2	黑色 长袖 男 XXXL 商务

假设将一名用户浏览记录类比为一篇文章，用户浏览的商品标题在分词汇总后作为其中的词库，平台的用户总数即为文本总数。TF-IDF计算及用于用户的偏好标签的简单示例如下。

用户A拥有三条浏览记录，分词后总计17个词；用户B拥有两条浏览记录，分词后总计10个词。假设平台的用户总数为10 000人，用户浏览过的商品标题带有“黑色”一词的用户有500人，那么以底数为2，可计算用户A和用户B对标签“黑色”的TF-IDF。

用户	词频(TF)	逆文本频率指数(IDF)	TF-IDF
用户A	$\frac{2}{17}=0.12$	$\log_2\left(\frac{10000}{500+1}\right)=4.32$	$0.12\times4.32=0.52$
用户B	$\frac{1}{10}=0.1$	$\log_2\left(\frac{10000}{500+1}\right)=4.32$	$0.1\times4.32=0.432$

得到“黑色”这个标签对用户A和用户B的权重分别为0.52和0.432，说明用户A对“黑色”的偏好高于用户B。因此有了权重，就能够将其运用于寻找相似用户。

想一想：

这个方法可以用于更大范围的文本分类吗？如新闻稿件的分类、论文的分类、用户评论的分类。

3.3.3 像数据科学家一样思考

如前所述，数据科学的使命是完成“三个转换和一个实现”，数据分析生命周期定义了从项目开始到项目结束整个分析流程的最佳实践，像数据科学家一样去思考，体现在数据分析生命周期的全过程中，用数据思维去思考并提出问题，再通过反复迭代循环，最

终解决商业问题，为企业带来价值。图 3.3 概述了数据分析生命周期的 6 个阶段，它是一个循环，箭头代表了项目在相邻阶段之间可能的反复迭代，而最大的环形箭头则代表了项目最终的前进方向。

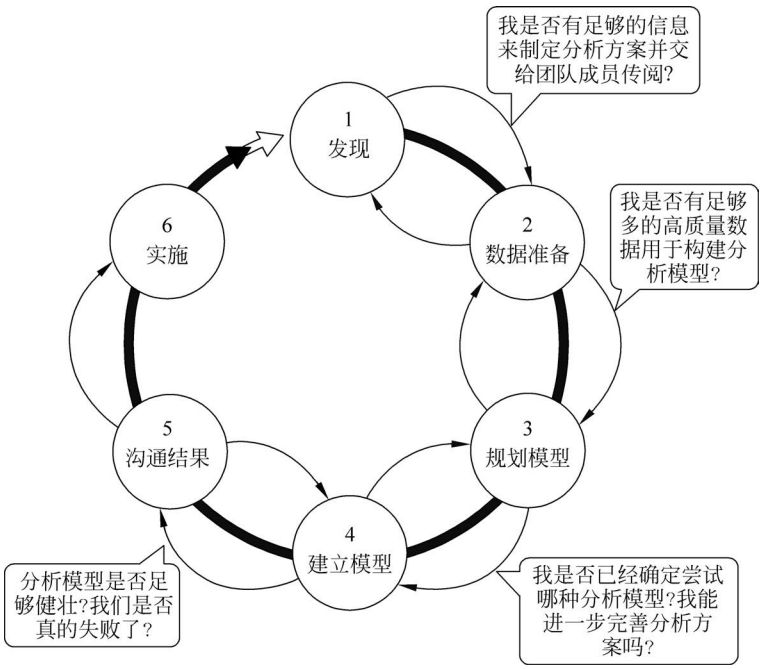


图 3.3　数据分析的生命周期

数据分析生命周期（数据科学项目）几个主要阶段应该完成的具体任务如下。

（1）发现阶段：在这个阶段理解业务领域的相关知识是关键，其中包括项目的相关历史。例如，可以了解该组织或者企业以前是否进行过类似项目，能否借鉴相关经验。这时还需要评估可以用于项目实施的人员、技术、时间和数据。在这个阶段，重点要把业务问题转换为分析挑战以待在后续阶段解决，并且制定初始假设用于测试和开始了解数据。

（2）数据准备：该阶段需要执行提取、加载和转换（ELT）。数据应在这一过程中被转换成可以被使用和分析的格式。在这个阶段，需要彻底熟悉数据，并且逐步治理数据。

（3）规划模型：在该阶段需要确定在后续模型构建阶段所采用的方法、技术和工作流程。探索性分析用于探索数据以了解变量之间的关系，然后挑选关键变量和最合适的模型。

（4）建立模型：该阶段首先创建用于测试、训练的数据集。此外，在这个阶段构建并运行由上一阶段确定的模型，同时还需要考虑现有的工具是否能够满足模型的运行需求，或者需要一个更强大的模型和工作流的运行环境（例如，更快的硬件和并行处理）。

（5）沟通阶段：该阶段需要与主要利益相关人进行合作，以第 1 阶段制定的标准来判断项目结果是成功还是失败。鉴别关键的发现、量化其商业价值，并以适当的方式总结发现并传达给利益相关人。

（6）实施：提交最终报告、简报、代码和技术文档是一种数据分析的成果物。此外，

也可以在生产环境中实施一个试点项目来应用模型。运行模型并产生结果后，根据受众采取相应的方式阐述成果非常关键。此外，阐述成果时展示其清晰价值也非常关键。如果经过精确的技术分析，但是没有将成果转换成可以与受众产生共鸣的表达，那么人们将看不到成果的真实价值，也将浪费许多项目中投入的时间和精力。

应用案例 3.3：淘宝的“淘气值”

用户画像可理解成“用户标签”，用户标签是用来概括用户特征的，如姓名、性别、职业、收入、养猫、喜欢美剧等。需要强调的是，组成用户画像的标签要跟业务或产品结合。为了实现大数据“杀熟”，电商企业建立用户画像标签可以说是煞费苦心。例如，商家要判断用户喜欢什么类型的活动，就要监测用户促销敏感度、满减促销敏感度、满赠敏感度、打折促销敏感度、换购促销敏感度、团购促销敏感度等。2017 年 6 月，阿里巴巴宣布“淘气值”将作为阿里巴巴会员等级的统一衡量标准，成为阿里巴巴最核心的用户画像标签，标志着淘宝会员评级从之前“买买买”的评分维度，到现在以“购买、互动分享、购物信誉”三个属性为重要用户特征，进而对用户进行更加多维度的精准分类。

针对不同“淘气值”的会员阿里巴巴提供更具特色的个性化服务。例如，2017 年“双十一”期间，淘气值超过 1000 的超级会员只需要 88 元就能买到一年能省 2000 元的 88VIP，而淘气值在 1000 分以下的普通会员需要花 888 元才能买到同样的“吃/玩/听/看/买”一卡通。

想一想：

你的“淘气值”是多少？对于“双十一”淘宝歧视“穷人”的说法你怎么看？

结合数据分析的生命周期来看，你又有哪些感想？

思考题

1. 什么是全数据思维、相关性思维、容错思维？结合统计学的“大数定律”又如何理解呢？

2. 什么是语言模型？为什么建立语言模型非常重要？

3. 如何理解数据思维的 5 个特点？请举例说明。

3.4 探究与实践

1. 百度“情感倾向分析”。

百度大脑是百度 AI 核心技术引擎，包括视觉、语音、自然语言处理、知识图谱、深度学习等 AI 核心技术和 AI 开放平台。进入“情感倾向分析”主题页，在功能演示区测试一下，你觉得“量化”结果靠谱吗？可以应用到哪些场景？

https://ai.baidu.com/tech/nlp_apply/sentiment_classify

2. 从生活中的用户画像你想到了什么？

秋高气爽，爸爸带着 Coco 出去玩，在某个湖边看到好多人在放风筝。突发奇想：我们也去放吧！Coco 表示：嗯！于是两人一起去走鬼（广东话，指无证流窜小摊贩）大叔那

买风筝(场景还原1)。看到Coco喜欢,爸爸就准备掏钱了,然而峰回路转,没想到又有下边一段(场景还原2)。

场景还原1

爸爸:买风筝。

大叔:大人放?小孩放?

爸爸:小孩。

大叔:男孩?女孩?

爸爸:男孩。

大叔:这个小号海豚风筝看一下……

大叔从背包里,抽出一个卷起来的风筝。

大叔摊开风筝给爸爸看。

爸爸:Coco喜欢吗?

Coco:喜欢!(#^.^#)

场景还原2

爸爸:多少钱?

大叔:20块。

爸爸:我扫哪里?

大叔:给小孩玩的话,可以换这个安全绕线轮,只要30块,线不会割着孩子的手哦。

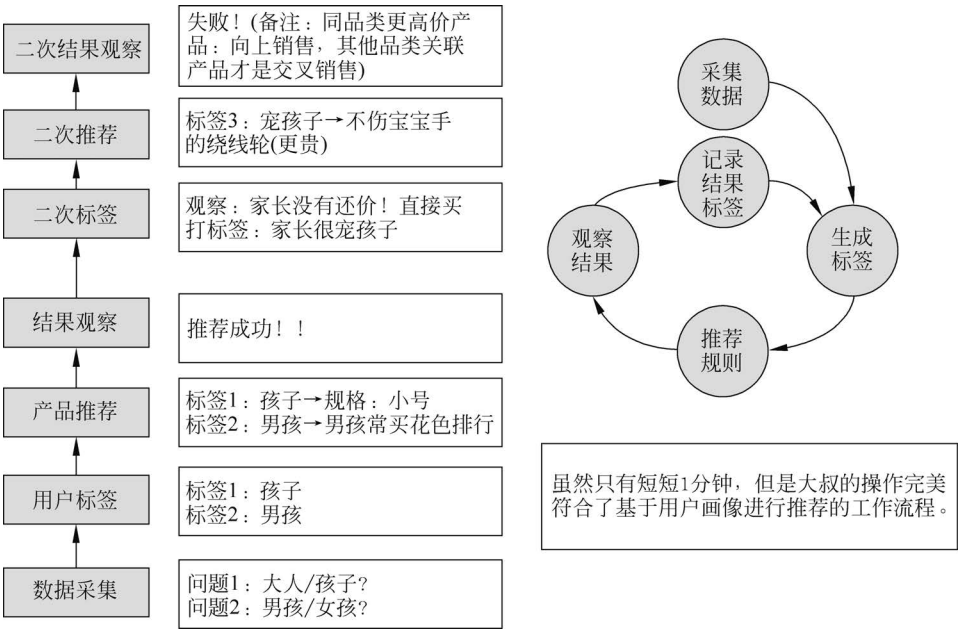
爸爸:(拿起20块的普通线轮,放在手上割了下试试)没事,就这个了,我扫哪里?

大叔:扫这里,微信支付宝都行。

爸爸:好了,走咯!

全程不到1分钟搞掂!!!

然后爸爸就和Coco愉快地放风筝去了。仔细想一想,是不是基于用户画像的推荐系统提升了交易的完整流程呢?数据采集→打标签→产品推荐→向上销售一气呵成,还做了二次推荐,把成交率和客单价分开提升,真是巧妙!



进一步探究你在×××平台/App上留下了哪些数据,你自己的"画像"你了解吗?

3. 相关思维与推荐算法。

推荐算法中较为传统的算法是协同过滤算法。假设已知小明、小张、小李、小王分别买了以下几本书。用"相关思维"思考可以得到的假设是:如果买书习惯跟小明类似的人购买了小明没有买的书,那么就认为,小明很有可能买这本书。于是,这类问题就变成了"找买书习惯跟小明类似的人"的相似度量化问题。

假设几位同学的购书记录数据如下，“0”和“1”分别表示未购买和已购买。

	《Hadoop权威指南》	《Java核心技术》	《新东周列国传》	《论语别裁》	《男装手册》	《世界是平的》
小明	1	0	1	0	1	0
小张	0	1	1	0	0	0
小李	1	1	0	1	1	0
小王	0	0	0	0	1	1

利用余弦定理提供的计算公式，你能得出如下结论吗？小李与小明“更”相似，应该推荐小明购买《Java 核心技术》这本书。