第5章 从数据到知识

开篇案例　“百度指数”能告诉你什么？

百度指数是以百度海量网民行为数据为基础的数据分享平台。在这里，可以研究关键词搜索趋势、洞察网民兴趣和需求、监测舆情动向、定位受众特征。以“数据科学与大数据技术”为关键字，从2016年1月1日到2023年8月7日的“百度指数”搜索趋势如图5.1所示。

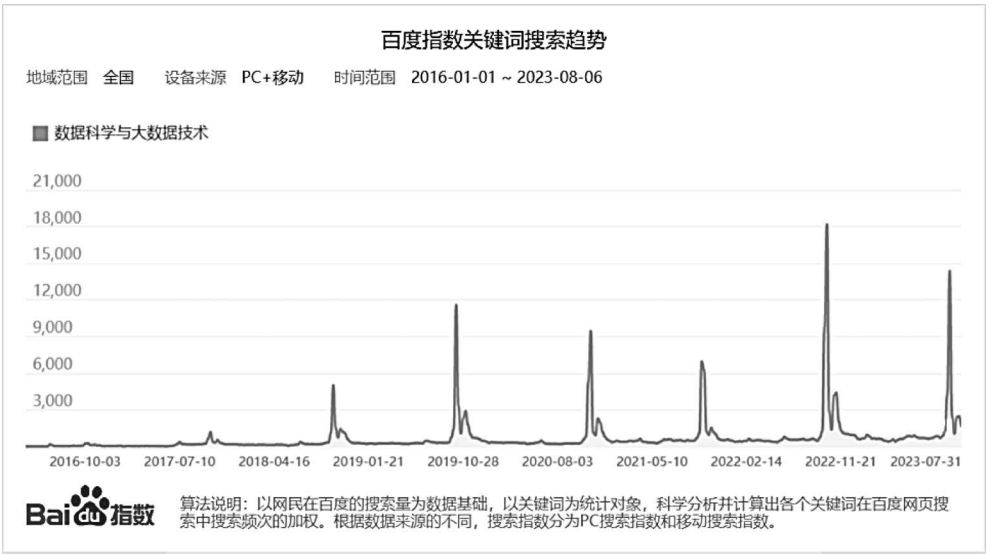


图5.1　百度指数搜索趋势

描述性分析、探索性发现、让数据讲故事……你有什么“发现”？从DIKW视角来看，“百度指数”是如何得到的呢？有什么价值？又可以“驱动”什么商业“行动”？

学习目标

学完本章，你应该牢记以下概念。

- 知识与知识表示、知识发现。
- 数据分析、数据挖掘、机器学习、模型与算法。
- 描述性分析、探索性分析、预测分析。
- A/B测试、决策支持。

学完本章，你将具有以下能力。

- 理解从商业问题到数据科学问题的重要性。
- 从时间的维度理解数据分析、数据挖掘与机器学习的区别与联系。
- 数据科学项目开发方法选择及流程。

学完本章，你还可以探索以下问题。

- 通过案例分析，比较商业需求与不同数据分析方法的关联性。
- 借助DIKW模型，理解不同数据分析的结果呈现方式及价值。

5.1 知识与知识发现

5.1.1 什么是知识

“知识”是人们熟悉的名词。但究竟什么是知识呢？维基百科给出的定义是：知识是对某个主题“认知”与“识别”的行为藉以确信的认识，并且这些认识拥有潜在的能力为特定目的而使用。意指通过经验或联想，而能够熟悉进而了解某件事情；这种事实或状态就称为知识，其包括认识或了解某种科学、艺术或技巧。此外，也指通过研究、调查、观察或经验而获得的一整套知识或一系列资讯。简言之，知识就是人们对客观事物（包括自然的和人造的）及其规律的认识，具体来说，包括对事物的现象、本质、属性、状态、关系、联系和运动等的认识，即对客观事物原理的认识。此外，知识还应包括人们利用客观规律解决实际问题的方法和策略，既包括解决问题的步骤、操作、规则、过程、技术、技巧等具体的微观方法，也包括诸如战术、战略、计谋、策略等宏观方法。如图 5.2 所示，“知识”来源于“数据”与“信息”，“知识”指导“智慧”行动。

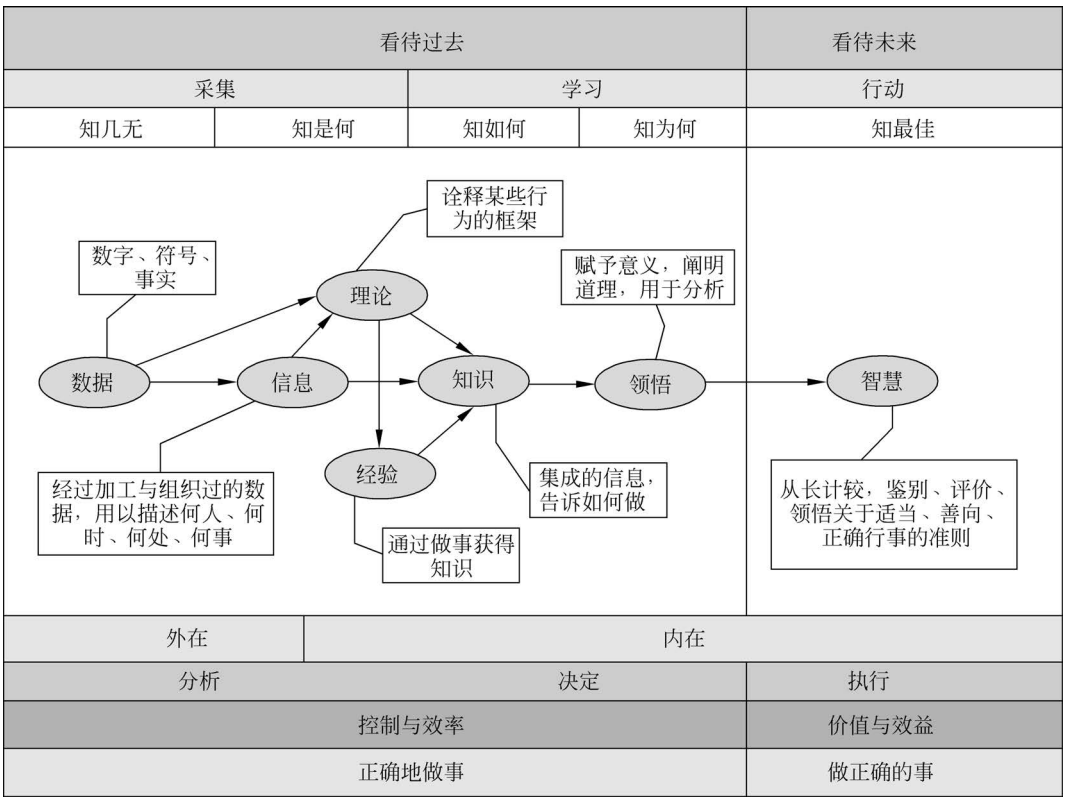


图 5.2 数据驱动的旅程

就形式而言，知识可分为显性的和隐性的。显性知识是指可用语言、文字、符号、形象、声音及他人能直接识别和处理的形式，明确地在其载体上表示出来的知识。例如，我

们学习的书本知识就是显性表示的知识。隐性知识则是不能用上述形式表达的知识，即那些“只可意会，不可言传或难以言传”的知识。

就严密性和可靠性而言，知识又分为理论知识和经验知识。理论知识是严密而可靠的，经验知识一般是不严密或不可靠的。就确定性而言，知识又可以分为确定性知识和不确定性知识。就确切性而言，知识又可以分为硬的、确切描述的知识和软的、非确切描述的知识。

另外从内容而言，知识可分为（客观）原理性知识和（主观）方法性知识两大类。就性质而言，原理性知识具有抽象性、概括性，因为它是特殊事务的概括和升华；而方法性知识具有一般性、通用性，因为只有通用才有指导意义，才配称为知识。这两个条件是知识与数据、信息的分水岭，也是对数据的不断深入理解（领悟、洞见）的升华。当然，所有原理性知识都是方法性知识的基础。

培根说过：“知识就是力量，但更重要的是运用知识的技能”。很显然，后面这句话才是培根要重点强调的，这也和他的哲学思想相吻合。从这一点来看，无论是原理性知识还是方法性知识都是有价值的。

想一想 5.1：知识的不确定性及不确切性的表示

尽管在人类的知识和思维行为中，精确性只是相对的，不精确性才是绝对的，但就知识的不确定性和不确切性而言，知识的表示通常用概率或程度来表示及度量。不确定性就是一个命题（亦即所表示的事件）的真实性不能完全肯定，而只能对其为真的可能性给出某种估计，它们描述的是人们的经验性知识。例如，“如果乌云密布并且电闪雷鸣，则很可能要下暴雨”“如果头痛发烧，则大概是患了感冒”。不确切性就是一个命题中所出现的某些言词其含义不够确切（模糊），从概念角度讲，也就是其代表的概念的内涵没有硬性的标准或条件，其外延没有硬性的边界，即边界是软的或者说是不明确的。例如，“小王是个高个子”“张三和李四是好朋友”“如果向左转，则身体就向左稍倾”。

狭义上的不确定性知识和不确切性知识的表示一般采用概率或信度来刻画。例如，{这场球赛甲队取胜，0.9}，这里的 0.9 就是命题“这场球赛甲队取胜”的信度。它表示“这场球赛甲队取胜”这个命题为真（即该命题所描述的事件发生）的可能性程度是 0.9，而{如果乌云密布并且电闪雷鸣，则天要下暴雨，0.95}{如果头痛发烧，则患了感冒，0.8}中的 0.95 和 0.8 就是对应规则结论的信度。它们代替了原命题中的“很可能”和“大概”，可视为规则前提与结论之间的一种关系强度。信度一般是基于概率的一种度量，或者就直接以概率作为信度。概率论研究和处理的是随机现象，事件本身有明确的含义，只是由于条件不充分，使得在条件和事件之间不能出现决定性的因果关系。无论采用什么数学工具和模型，都需要对规则和证据的不确定性给出度量。

“一切皆可量化”你体会到了吗？

你还能举出一些其他的例子吗？

5.1.2 知识发现的任务

从数据科学的角度来看，虽然历史上由于数据的匮乏及技术的局限，只能对有限的

数据进行汇总统计及简单的定量及定性分析，也在一定程度上对决策起到辅助的作用。传统的知识发现任务都是围绕结构化数据展开的，具体包括以下几点。

(1) 数据汇总及描述。其目的是对数据进行浓缩，给出它的紧凑描述。传统的也是最简单的数据总结方法是计算各种变量的求和值、平均值、方差值等统计值，或者用直方图、饼状图等图形方式表示。

(2) 分类与聚类。分类的目的是提出一个分类函数或分类模型(也常称为分类器)，该模型能把数据库中的数据项映射到给定类别中的某一类中。聚类则是根据数据的不同特征，将其聚集在一起，它的目的使得属于同一类别的个体之间的差异尽可能小，而不同类别上的个体间的差异尽可能大。分类及聚类往往通过计算机编程实现，也称为面向数据库的方法(算法)。

(3) 相关性分析及偏差分析。相关性分析的目的是发现特征之间或数据之间的相互依赖关系。偏差分析的基本思想是寻找观察结果与参照量之间有意义的差别，发现异常。

(4) 建模。建模就是构造出能描述一种活动、状态或现象的数学模型，常用于预测分析。

以上这些任务通常都是用存储在数据库中的数据、面向某个特定的商业需求展开的，习惯上称之为“数据挖掘”。随着信息化技术的不断推进，从数据到知识的研究方法及工具也发生了翻天覆地的变化，但其本质还是知识发现(获取)，同时这类知识还应该是指面向计算机的知识描述或表达形式和方法。

知识表示与知识本身的性质、类型有关。面向人的知识表示可以是语言、文字、数字、符号、公式、图标、图形和图像等多种形式，这些表示形式是人所能接受、理解和处理的形式。但面向人的这些知识表示形式，目前还不能完全直接用于计算机，因此就需要研究适于计算机的知识表示模式。具体来讲，就是要用某种约定的(外部)形式结构来描述知识，而且这种形式结构还要能够转换为机器的内部形式，使得计算机能方便地存储、处理和应用，这类知识就是“可执行的知识”。当适用于计算机描述的知识由于具有“可执行”的特点，也就构成了实现基于数据驱动的基础。

面向非结构化数据的知识发现任务往往更强调“理解”，即像人一样“感知”周围世界并理解，如自然语言理解、图像理解等。数据科学研究的目的就是发现复杂数据的关系(知识)，并以“隐性知识显性化、显性知识结构化”为目标，实现真正的数据驱动。

应用案例 5.1：什么是“可执行的知识”

由于互联网的发展，产生的数据中绝大部分(超过 80%)都是以文本、图像等非结构或半结构的方式存储。所以，挖掘数据价值首先就是要系统地研究如何挖掘无结构数据的价值，也就是说，要实现从“大数据”到“可执行的知识”的转变。

一个“可执行的知识”的例子如图所示，即“驾驶行为识别”的可量化结果，通常是以概率(可能性)的形式出现的，这种计算机可存储且表示的“知识”就可以方便地作为下一步决策的客观依据。

5.1.3 决策与决策支持

决策是所有组织(企业)经营活动中最重要的环节之一,决策决定着组织的成败。做出正确决策的回报可能非常高,而做出不正确决策的损失也可能非常严重。由于内部与外部的因素,做决策变得越来越困难。多年来,管理者认为做决策纯粹是一种艺术,一种需要长时间的经历(即在反复尝试中吸取经验)和依靠直觉的才能。这种自顶向下的决策过程,具体表现在:决策是在"业务驱动"情况下的行为;使用还原论把复杂问题简单化,找到关键点改善决策;数据分析目标是定性的,依据定性分析做出判断和决策。

数据时代为自底向上的决策流程提供了可能性,即业务数据化后以"数据驱动"作为判断和决策的依据。具体体现在:围绕数据以定量分析发现趋势,其核心是高深的计算机技术、模型和算法。即从系统论总体考虑问题,发现、解释、可视化和讲述数据中的模式以推动业务战略,如图 5.3 所示。

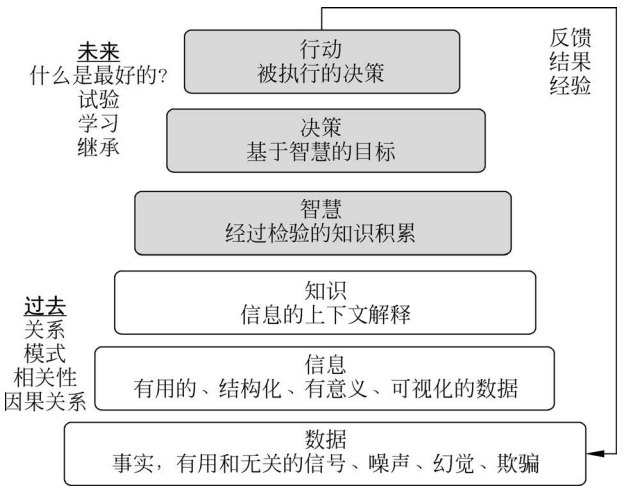


图 5.3 基于数据驱动的决策

决策支持系统(Decision Support System,DSS)是管理信息系统(Management Information System,MIS)向更高一级发展而产生的先进信息管理系统。它为决策者提供分析问题、建立模型、模拟决策过程和方案的环境,可调用各种信息资源和分析工具,帮助决策者提高决策水平和质量。从如图 5.4 所示的时间轴来看,早期的决策基于结构化的定期报告,后期发展为以管理系统为基础的不同维度、不同级别的各种报告,以便更好地理解和应对业务不断变化的需求与挑战。20 世纪 80 年代出现的 CRM 与 ERP 系统为后续出现的数据挖掘与商务智能(Business Intelligence,BI)提供了数据基础。2010 年以来,由于数据获取和使用方式又进行了一次范式转变,大数据与人工智能的出现正在改变 BI 的现状,使得机器学习在图像、视频及语音识别领域取得的成果融入决策过程中。

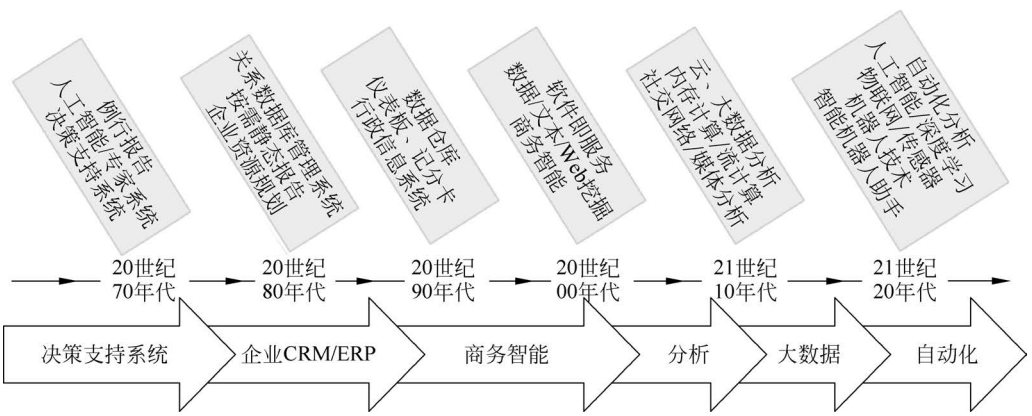


图 5.4 决策支持、数据分析、商务智能与人工智能的发展

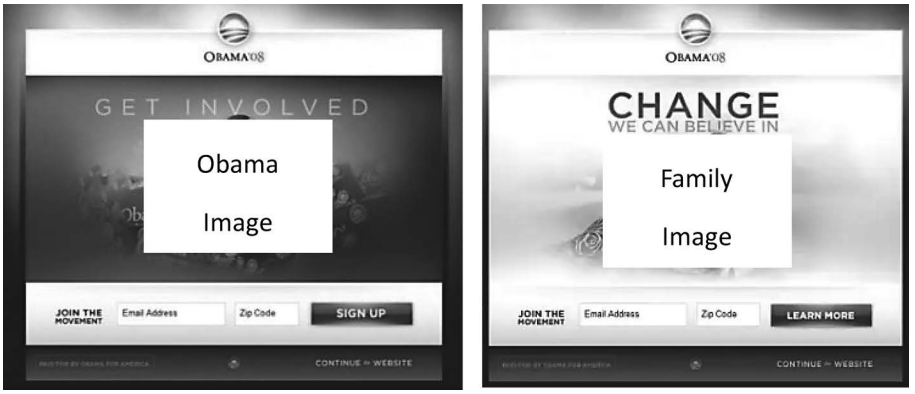
(图片来源:《商业分析:基于数据科学及人工智能技术的决策支持系统》)

从上述描述可以看出,知识发现的进程随着数据的不断充裕、计算技术的不断强大,越来越趋于向全自动化(全智能)的方向发展,使得基于数据驱动的自动化智慧决策(DIKW)逐渐成为可能。从这一视角来看,目前较为流行的机器学习与数据挖掘、人工智能、统计这些领域是相通的,其中,机器学习是最有力的工具,应用范围也更广泛。

技术洞察 5.1:什么是 A/B 测试——奥巴马当选美国总统背后的故事

A/B 测试是一种新兴的网页优化方法,可以用于增加转化率、注册率等网页指标。具体做法是为 Web 或 App 界面或流程制作两个(A/B)或多个(A/B/n)测试版本,在同一时间维度,分别让相似的访客群组(目标人群)随机地访问这些版本,收集各群组的用户体验数据和业务数据,最后分析、评估出最好版本,正式采用。

奥巴马成功当选总统的背后也有 A/B 测试的功劳,称为政治竞选中的经典案例。时任奥巴马竞选团队开发了简单易用的 A/B 测试系统如图所示,左图是最初奥巴马的竞选网站设计,右图是测试版本。可以看到的是,左图是典型的个人英雄角色,而右图改成了更美式价值观的家庭图片。右图的"Change, We Can Believe In"更突出了奥巴马竞选中的口号。另外,左图有注册(Sign Up)栏,对访问者可能有一定的过度要求,而右图则只是"Learn more",显得更轻松友好。



经过 A/B 测试改善的版本得到了惊人的结果，如通过网站登记的访问者提升了 40.6%，新增了 280 万的联系 E-mail，增加了 28.8 万名志愿者，获得 5700 万美元的捐助等。

想一想：

你理解“基于数据驱动的决策”的实质了吗？

思考题

1. 什么是隐性知识？你能举例说明吗？
2. 为什么说“一切皆可量化”？这里的“量化”的含义是什么？举例说明。
3. 什么是“可执行的知识”？决策支持的自动化程度与该类知识的依存关系如何？
4. 为什么要用 A/B 测试？举例说明 A/B 测试的具体应用。

5.2 数据分析、数据挖掘与人工智能

5.2.1 知识发现的方法

知识发现的方法可简单归为两类：统计方法和机器学习方法。事物的规律性一般从其数量上会表现出来，而统计方法就是从事物的外在数量上的表现去推断事物可能的规律性。因此，统计方法就是知识发现的一个重要方法。常见的统计方法有回归分析、判别分析、聚类分析以及探索分析等。机器学习方法包括符号学习、连接学习以及统计学习等。可视化就是把数据、信息和知识转换为图形的表现形式的过程。可视化可使抽象的数据信息形象化。于是，人们便可以直观地对大量数据进行考察、分析，发现其中蕴藏的特征、关系、模式和趋势等。因此，信息可视化也是知识发现的一种有用的手段。

就像“一千个人眼里有一千个哈姆雷特”一样，对于什么是数据科学也有很多种不同的解读，并由此衍生出很多相关概念，如数据驱动、大数据、分布式计算等。这些概念虽然各有侧重点，但它们都毫无争议地围绕同一个主题：如何从实际的生活中提取出数据，然后利用计算机的运算能力和模型算法从这些数据中找出一些有价值的内容，而“知识

发现”为商业决策提供支持。这正是数据科学的核心内涵。

在科学的历史上，任何词汇的出现与流行都深深印刻着时代的烙印。“数据分析”“数据挖掘”与“机器学习”等词汇是经常与数据科学同时出现的热门话题，但其本质都是对“从数据到知识”过程的描述，也体现了“让数据变得有用”的不同程度及发展脉络。但随着时代的变迁，从数据到知识的方法与技术发生了巨大的变化，这些词汇的本质内涵也不尽相同，但理解这些词汇的内涵是非常必要的。

想一想 5.2：你能从下面对“知识”的描述中得到什么

- 知识是我们已知的，也是我们未知的。基于已有知识，去发现未知，由此，知识得到扩充。我们获得的知识越多，未知的知识就会更多。因此，知识的扩充永无止境。
- 在终极的分析中，一切知识都是历史；在抽象的意义下，一切科学都是数学；在理性的基础上，所有判断都是统计学。
- 不确定性的知识加上所含不确定性度量的知识最终称为“可用的知识”。

5.2.2 数据分析与业务分析

百度百科给出的数据分析的定义是：“数据分析是一种统计学常用方法，指用适当的统计分析方法对收集来的大量数据进行分析，将它们加以汇总和理解并消化，以求最大化地开发数据的功能，发挥数据的作用。数据分析是为了提取有用信息和形成结论而对数据加以详细研究和概括总结的过程”。从这个定义可以看出，数据分析强调的是统计分析与统计推断、数据可视化、实验设计、领域知识与沟通。

企业的数据分析（业务分析）往往围绕关键绩效指标（Key Performance Indicator，KPI）展开，这些 KPI 是衡量流程绩效的一种目标式量化管理指标，是把企业的战略目标分解为可操作的工作目标的工具，是企业绩效管理的基础。商业盈利的三要素是“增加收入、减少支出、防范风险”，而 KPI 是商业问题转换为数据科学问题的具体体现。数据指标必须适配业务目标，是企业走向成功的关键一环。在互联网与移动通信时代，企业盈利的关注点（KPI）也有所不同，相关部分行业的 KPI 示例如表 5.1 所示。

表 5.1 不同领域常用的 KPI 指标

业　务	价　值	典 型 指 标
网站	流量	页面流量（Page View，PV）、独立访客（Unique Visitor，UV）、平均停留时间、跳出率、退出率、人均浏览次数、新独立访客等
电商	交易	购买转化率、客单价、成交金额、成交人数、搜索点击次数、关注人数、收藏人数、活跃商品数
游戏	留存付费	注册用户数，日/月活跃用户数（Daily/Monthly Active Users，DAU）、最高同时在线玩家数（Peak Concurrent Users，PCU）、每用户平均收入（Average Revenue Per User，ARPU）、付费率、任务停滞率
社交网站	互动（内容/好友）	病毒增长率、活跃用户、发送消息数、关注人数、回复率、转发率

续表

业　务	价　值	典型指标
视频	观看付费	观看次数、观看时长、评论数、付费率
移动应用	推广，交互	自然用户数、渠道用户数、渠道增长率、使用时长、使用路径、留存率、活跃用户

狭义上的数据分析中一般数据规模都不会太大，也相对简单，具体细分可以包括描述性分析、探索性分析等，当然也可能包括简单的因果分析，如回归分析等。所以在这里把“数据分析”理解为早期的研究数据的范畴，其结果往往是以数据分析报告呈现与沟通的。所以数据分析往往与业务分析有着紧密的联系，其结果为特定操作提供决策或建议。

从 DIKW 的视角来看，数据统计与分析的过程也是追求实现“数据—信息—知识—智慧”持续变化的过程。即从数据开始，以形成智慧为最终目的。具体过程是：借助相关操作对数据进行处理、加工，明确数据之间的关系，提取出有意义的信息，进而将信息组织成知识，在明确“如何去使用”及“应该何时使用”及“为什么要使用”时，便形成了智慧。显然，数据统计与分析中的几个关键词，即数据、统计、分析为智慧决策奠定了基础，而智慧决策又必须在前面环节的基础上展开。具体来说，数据需要为统计服务，统计是建立在数据提供的基础上；统计的结果是为了进行分析，分析必须依赖于统计结果；分析的目的是提供决策的依据。

5.2.3　数据挖掘与知识发现

数据挖掘(Data Mining，DM)可以视为信息技术自然进化的结果。自 20 世纪 60 年代以来，数据存储及管理从原始的文件处理演变成复杂的、功能强大的数据库系统，完成了对大量数据的存储、检索及事务性处理，并逐渐出现了对数据进行高级分析的需求，即在统计分析的基础上进行深层次规律、模式的探究。因此，数据挖掘一般特指从结构化数据库(如 CRM 或 ERP)中挖掘知识。知识发现(Knowledge Discovery in Database，KDD)是与数据挖掘同时出现的概念，其目的就是从数据集中抽取和精化一般规律或模式。从其他数据挖掘的定义中也可以看出两者之间的关系，如“数据挖掘就是对数据库中蕴涵的、未知的、非平凡的、有潜在应用价值的模式(规则)的提取”“数据挖掘就是从大型数据库的数据中提取人们感兴趣的知识。这些知识是隐含的、事先未知的潜在有用信息”。因此可以看出，数据挖掘的本源是大量、完整的数据；而数据挖掘的结果是知识(规则)。

5.2.4　机器学习与人工智能

如前所述，经验积累、规律发现和知识学习等能力都是智能的表现。那么，要实现人工智能就应该赋予计算机这些能力。简单来讲，就是要让计算机或者说使其具有自学习

能力。试想，如果机器能自己总结经验、发现规律、获取知识，然后再运用知识解决问题，那么，其智能水平将会大幅度提升，这也是数据科学追求的终极目标。

机器学习（Machine Learning，ML）是一类从数据中（特别是非结构化数据中）自动分析获得规律，并利用规律对未知数据进行预测的算法。机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法。机器学习是人工智能（Artificial Intelligence，AI）的一个分支。人工智能的研究历史有着一条从以“推理”为重点，到以“知识”为重点，再到以“学习”为重点的自然、清晰的脉络。显然，机器学习是实现人工智能的一个途径，即以机器学习为手段解决人工智能中的问题。

从 DIKW 模型的角度来看，数据挖掘与机器学习都是完成从数据到知识的转换过程，只是数据挖掘更强调与业务领域结合，因此往往与数据库关系密切，而机器学习强调“学习”，特别是从非结构化数据中“感知”，所以更习惯将机器学习看成是人工智能的必要且关键环节。

技术洞察 5.2：自动驾驶中的数据科学、机器学习与人工智能

关于数据科学及相关领域，经常会有这样的问题出现，如“数据科学和机器学习有什么区别”或者“如何体现某人正在从事人工智能研究”。这些领域确实有很多重叠之处，再加上三个领域都充斥着媒体的营销炒作，导致人们很容易对它们产生混淆。

从 DIKW 模型的角度来看，机器学习实现从“信息到知识”的过程，而人工智能则强调的是“从知识到智慧”的行动，数据科学研究则是贯穿始终的理论、方法、技术及实践。这三个领域之间的差异可以简单理解为：机器学习产生预测、人工智能产生行为、数据科学产生见解（洞见）。

以自动驾驶研究领域为例，如需要研究车可自动停靠在有停车标识位置这个特定的问题，就需要从这三个领域分别进行思考与实践。

机器学习：汽车必须使用摄像头识别停车标志。在构建了包含数百万个街景标识图像数据集的基础上，训练一个算法来预测哪里会有停车标识。

人工智能：一旦车可以识别停车标志，就需要决定何时采取制动这个行为。过早或过晚制动都是很危险的，并且应该可以处理不同的道路状况（例如，如果识别是一条光滑道路，它并不能很快减速），这是一个控制理论问题。

数据科学：在自动驾驶街道测试中，如果发现汽车的表现不足够好，停车标识出现了不少误判（如“停车”标识被识别为“禁止入内”）或漏判（错过该停车标识）。这需要进一步分析这些测试数据，结果得到的结论是漏判率与时间有关：在日出之前或日落之后，更有可能错过停车标志。当发现现存的大部分训练数据仅包含白天时段，就必须构建包含夜间图像的更合适的数据集，并在此返回到机器学习步骤对识别算法及模型进行优化。

可见数据科学、机器学习与人工智能三者关注的核心问题不同，你理解了吗？为什么说发现问题比解决问题更重要？

5.2.5 从数据到知识

如前所述，数据科学及数据思维可以用 DIKW 模型来诠释，大数据的核心是挖掘（提

取)数据的价值,数据科学将解决这一过程中出现的问题,也就是解决在“使数据变得有用(获取知识)”过程中出现的各类问题。数据科学的方法论概括为“建模、分析、计算和学习的杂糅”。目标决定路径,任务选择方法。从数据到知识,就是从复杂的数据中提取有用的信息,并进而转换为指导行动的知识和决策,“实现对现实世界的认识与操控”。这一过程可以简单理解为“建模”,建模就是把问题形式化,特别是数学化的过程,如图 5.5 所示。

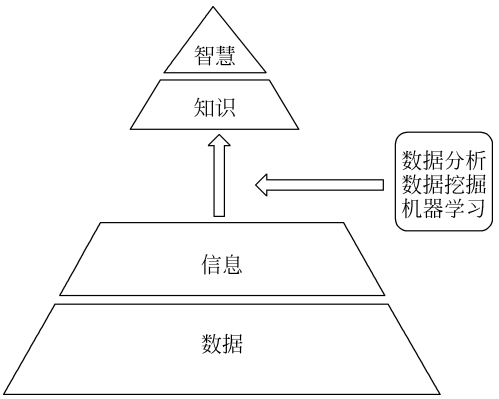


图 5.5　从数据(信息)到知识

对数据科学来说,机器学习算法是核心,而学习是数据赋能的工具。学习是数据科学处理数据的独有方法。数学与统计学解决数据科学基础建模问题;计算机解决“算得出、算得准、算得快”等问题;而人工智能将帮助数据科学解决应用问题,即完成“操控现实世界”的目标。

人工智能与大数据技术常常也难以区分,如果一定要区分,前者更强调与领域知识技术结合的“行动”,如与自然语言处理、计算机视觉、机器人、自动驾驶、竞技游戏等技术结合,更聚焦数据价值链的后端(从 K 到 W)。而后者更关注如何从现实世界中获取汇聚数据,更强调数据价值链的前端(从 D 到 K)。当然,对于完成数据价值链中段的分析和处理(从 I 到 K)角度来看,无论是前者还是后者都是非常重要的,这也说明将实现三个转换(D→I→K→W)作为数据科学的学科任务是适宜的,而基于大数据的问题发现是数据科学任务的起点(“数据密集型科学发现”范式)。

总体来说,采用任何方法(数据分析、数据挖掘、机器学习)都是完成从信息(即加工过的数据)到知识过程,只是所获得的知识的深浅、呈现方式、可应用程度有所不同,致使人们在决策时参与的程度不同。简单来说,数据分析强调统计描述与推断,数据挖掘强调“挖掘”业务场景及模式,而机器学习强调“学习”过程的自动化。作为导论性课程,这里不严格区分这几个词。从 DIKW 模型的角度来理解,数据分析、数据挖掘及机器学习统一理解为将“信息”转换为“知识”的过程即可。因此,将面向结构化数据的数据挖掘及面向非结构化数据的机器学习方法均称为算法,不严格区分“数据挖掘”和“机器学习”这两个词也是合理的。因为从获取知识(认知现实)的角度来看,二者所处的地位和所完成的数据科学任务都是一致的。

想一想 5.3：到底是“算法”还是“模型”

从定义上说，两者是完全不同的。严格来说，“算法”是完成某项任务时需要遵循的一组规则或步骤，而“模型”是对世界一种附有假设的数学描述，模型是算法实现后的结果。这两个概念看起来虽然是不同的，它们的区别也应该是显而易见的。然而，由于不同学科发展的历史原因，要精确区分两者之间的差别实在浪费时间，也毫无必要。

从某种程度上说，这是一个历史遗留问题。统计学和计算机科学一直在并行发展，它们常常使用不同的词汇描述同样的东西。这也就导致很难确定某个概念到底是机器学习算法还是统计模型。统计模型出自统计学家之手，机器学习算法则是计算机科学家所开发的，但某些技术和方法在统计模型和机器学习算法中都会用到。所以这两个词有时可以换着使用。当人们谈起这些时，既可以说是算法，也可以说是模型，尽量不要受到这些的干扰。

以结构化数据为例，还有几个类似容易混淆的术语，如数据集由数据对象组成，一个数据对象代表一个实体，关系型数据库的行对应于数据对象，而列对应于属性。属性是一个数据字段，表示数据对象的一个特征。在文献中，属性、维、特征和变量也常常互换使用。术语“维”一般用在数据仓库中，机器学习相关文献更趋于使用“特征”，而统计学家则更愿意使用“变量”，数据挖掘和数据库的专业人士一般使用“属性”。

数据科学维恩图所描述的“交叉”你感受到了吗？

综上所述，数据科学的经典定义是统计学、计算机学科和领域专业知识结合的交叉学科（数据科学维恩图），数据科学的本质是发现数据的价值，其数据价值的呈现形式可能是多样的，如图 5.6 所示，包括发现规律、现象、模式、模型等。数据科学与机器学习和人工智能主要的区别在于，在数据科学中，人是循环中不可缺少的一部分：算法得出数据结果，人们通过数据得到见解（洞见）或从结论中受益。而机器学习与人工智能的结合强调决策行动的“全自动化”。

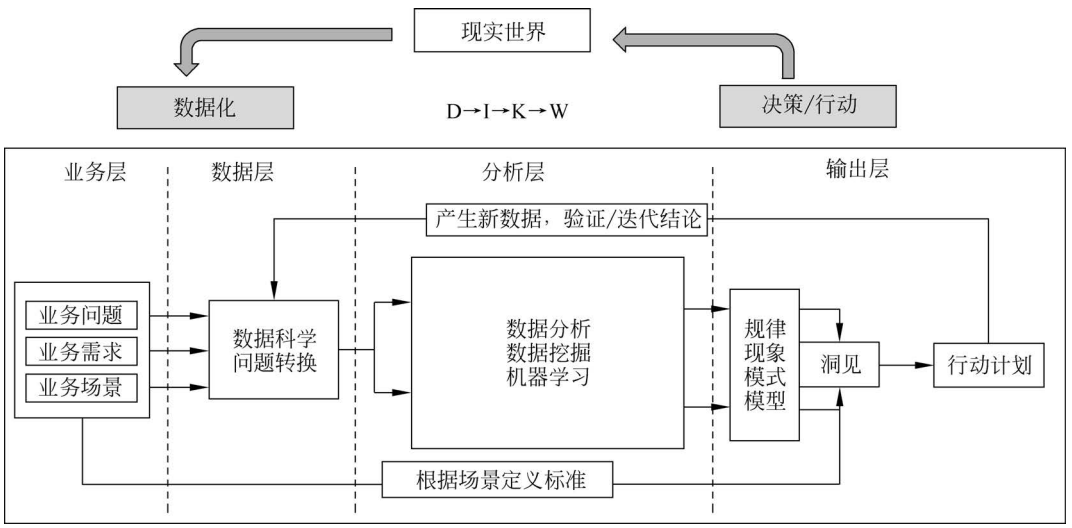


图 5.6　数据分析、数据挖掘与机器学习

思考题

1. 模型与算法的区别是什么？为什么常常不去区分它们？
2. 举例说明容易混淆的术语还有哪些。为什么会产生这些混淆？
3. 什么是 KPI？举例说明。
4. 为什么说从 DIKW 视角来看，数据分析、数据挖掘和机器学习所起的作用都是一样的？

5.3 数据科学项目的选择

5.3.1 数据科学的认知误区

数据科学并不玄虚。做数据科学，首先要梳理行业的商业逻辑，抽象定位这个业务的本质是什么；抓住本质后要用数学工具去量化它，处理庞大的数据问题。知其然，然后知其所以然。所谓数据科学的本质，只有放到环境的“上下文”中，才能发挥正确的价值。数据科学是一门综合性学科，既有科学问题也有工程问题，它是科学和艺术的结合。从图 5.7 可以看出，数据科学是由客观存在与主观意识结合的一门“艺术”。

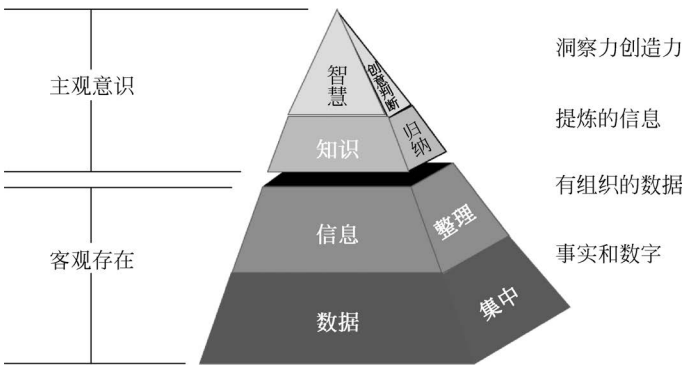


图 5.7 数据科学的艺术

数据科学的认知误区常常包括以下几点。

1. 误区一：让数据自动去寻找问题的答案

数据科学的各个处理阶段都需要数据科学家的介入。问题分解、解决方案设计、数据准备、选择最合适的机器学习算法、精准解释分析结果、根据分析结果采取必要的干预措施，这些环节都需要数据科学家的参与，特别是掌握不同技能的数据科学家团队。

2. 误区二：每个项目都需要大数据和深度学习

一般来说，拥有“更多”的数据是很有帮助的，但是拥有“正确”的数据更重要。数据科学项目经常在多个组织中进行，在数据量和计算能力方面，一般组织的资源明显少于谷歌、百度或微软等巨头。数据量根本达不到百万级，就没有必要考虑太字节（TB）级数

据下的数据架构。

3. 误区三：数据科学很容易实施

目前，市场上有很多相关的软件可以使用，这就导致很多人觉得数据科学借用这些软件就很容易实现。正确地进行数据科学实践既需要适当的领域知识，也需要关于数据属性的专门知识，以及各种机器学习算法底层假设的支持。数据科学需要投资开发数据的硬件设施，还需要具有数据科学专业背景的研发人员。

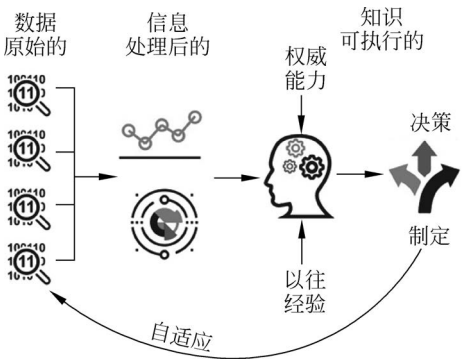
4. 误区四：利用数据科学一定能成功

数据科学并不能给每个项目都带来积极的结果，有时数据中没有金矿只有砂砾。数据科学往往是一个加分项，适当的数据和专业的团队可以为组织提供成功所需的竞争优势，但无法保证一定能成功。

想一想 5.4：数据科学还是什么

关于数据科学，还有很多说法，你觉得他们说得有道理吗？

- 数据科学是一个过程，而非事件。在这个过程中使用数据来了解事物，了解世界。例如，当你有一个问题的模型或假设，你会试着通过数据来验证这个假设或模型。
- 数据科学是一门艺术，揭开那些隐藏在数据背后的观点和趋势，将数据编译成一个故事，以说故事的方式激发新的视角，再利用这些视角、观点、想法为企业或机构做出战略选择。
- 数据科学是一个领域，是关于从各种形式中进行数据提取的过程和系统，无论数据是非结构化的还是结构化的。
- 数据科学是对数据的研究，正如生物科学是研究生物、物理科学是研究物理反应一样。数据是真实的，具有实际属性，是需要我们对其进行研究的。



5.3.2 成功的数据科学项目

数据科学项目的成败取决于人类的参与和关注，这些应该从理解“可能的”业务操作开始，它是每个项目中最应该问的第一个问题，这是一切的起点。不能仅通过知道一个商业问题的答案(K)来赚钱，而是当你采取行动时(W)才能赚钱，而商业行动是受到现实世界中能做到的能力范围的限制。一个业务行动可能依据企业大的宏观策略，需要创建

外部伙伴关系，需要让整个团队都参与决策，这才是真实世界的样子。但是可以采取的影响现实世界的好的、有效的行动的数量通常相对较小。一旦知道了可以采取的业务操作，就应该使用这些操作驱动数据分析，而不是相反。图 5.8 说明了“可能的分析”和“可能的行动”之间的关系。

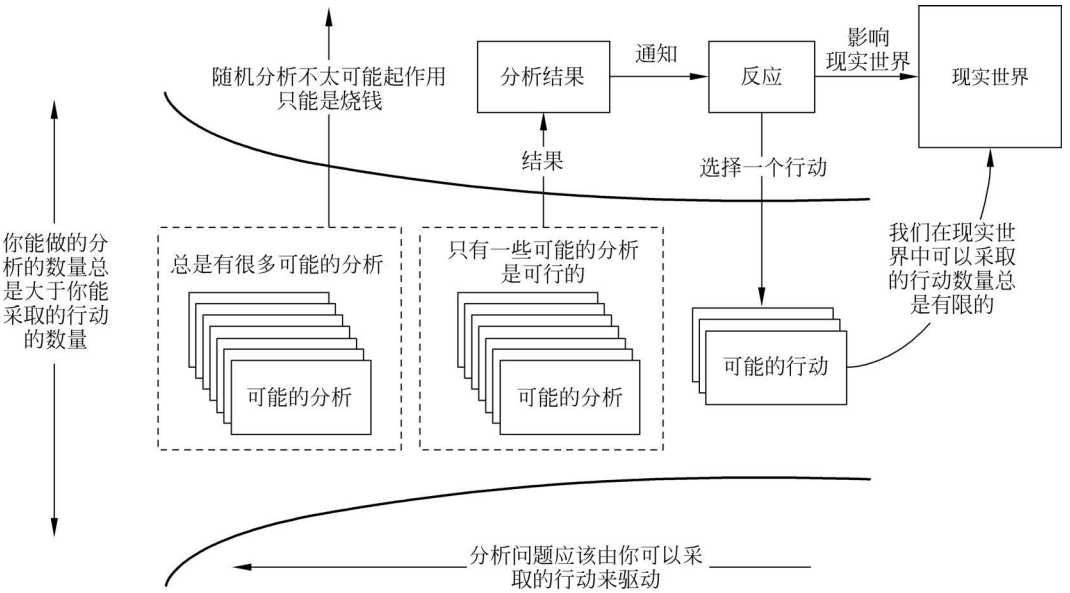


图 5.8 “可能的分析”与“可能的行动”

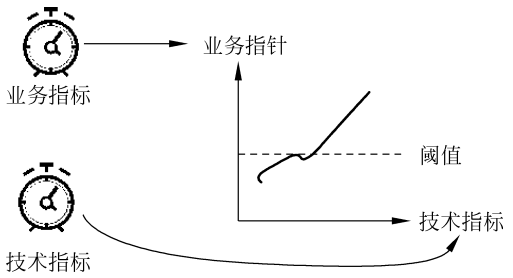
（图片来源：《Succeeding with AI: How to make AI work for your business》）

在每一个数据科学项目中，需要记住两点：①唯有行动使你获利，没有行动的数据分析只是成本，只有当企业执行适当的业务操作时才能赚钱，而只完成某些数据分析时则不能赚钱。分析可以成为营利的推手，但要从会计角度来看分析是一种成本。只有当它能帮助你采取良好的商业行动时，分析才不再是一种成本，而是一种投资。②要想成功要关注整个系统，而不是其中的个别部分。数据科学项目的最终结果取决于整个系统的运作情况。

一个数据科学项目要想成功，需要能够衡量科学项目数据结果对商业的影响，而且这种衡量是必须可量化的。机器学习算法不能使用直觉指标作为正在进行这个项目的反馈，所以需要有人为定义的一个量化指标。在衡量数据科学家将如何影响业务之前，必须先考虑需要衡量业务的指标，即应该考虑的是有没有办法根据一些数值指标来衡量相关业务做得有多好，这类指标与业务收益直接相关，业务度量可能是现成的，也可能是你自己开发的。

技术洞察 5.3：什么是利润曲线

利润曲线是建立在业务和技术指标之间的关系曲线，即建立以机器学习算法使用的技术指标与业务指标的阈值（业务指标项目必须达到的最小值才能可行）的对应关系。尽管在指标之间建立数学关系的一般概念大家都清楚，但构建利润曲线则更加突出其重要的相互关系。



利润曲线指定了一个技术指标和一个商业指标之间的关系。它允许你理解技术结果(以技术指标的形式)对商业条款的意义。在定义利润曲线时,你会通过一种数学关系将商业和技术指标结合起来,这样可以将研究问题与你要解决的商业问题联系起来,将技术和商业结合起来。价值阈值是指你的项目必须达到的商业指标的最小值,以保证项目的可行性。

均方误差(Mean Square Error,MSE)是回归模型的误差度量方法,它就是一个技术指标。

5.3.3 数据科学项目的选择之旅

对应不同的业务场景需求,数据科学问题也有所不同。但最终目的都是在一定程度上助力企业的决策,企业在进行数据分析时常涉及描述性分析、预测性分析与规范性分析等分析类别。从基于数据的决策驱动及自动化程度来看,它们之间的区别与联系如图 5.9 所示。结合 DIKW 模型,可以更深刻理解各层次之间的关系及给企业带来的不同价值。

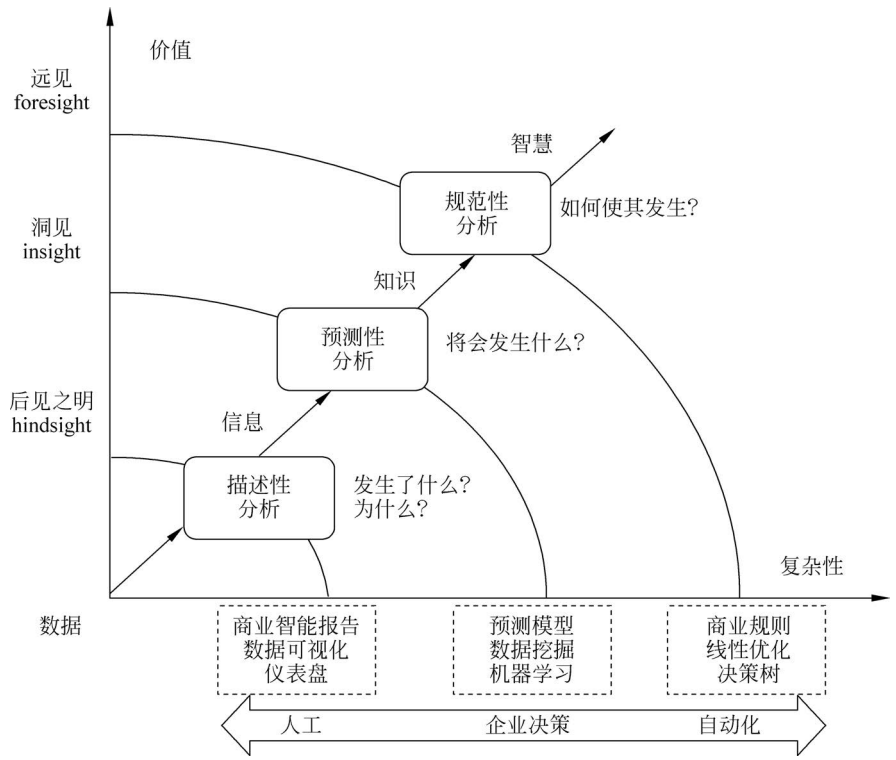


图 5.9 数据分析与企业决策管理

描述性分析获得整个企业(组织)正在发生什么的“描述”,通过这些描述获得一些事件的基本趋势和原因,并以适当的分析报告的形式为决策提供有价值的建议。这些就是目前通常狭义下的“数据分析”的内容。

预测性分析旨在确定未来可能发生的事情。传统的预测分析通常都是比较宏观的或定性的分析,随着技术的不断发展,这种分析基于统计技术或属于数据挖掘范畴。数据挖掘及机器学习提供的预测分析结果,回答“将会发生什么?”等问题,包括预报、回归、分类等。

规范性分析的目的基于可能的预测做出决策的依据,以实现最佳性能。历史上,这些属于管理学科下的优化系统的性能,是指为特定操作提供决策或建议。规范性分析结果(包括优化、决策树、启发式数学编程)等,为智慧决策的实现奠定基础,从根本上解决企业决策管理高效优化的问题(如电商的推荐系统)。

从单纯的数据科学项目的范围来看,从“数据到知识”的选择之旅示意图如图 5.10 所示,这是一个科学冒险之旅,目标不同、数据积累不同,选择之旅不同,结果也不同。这里需要数据思维,思维就是我们对客观世界的一种主观抽象描述,通过思维来分析问题,从而更为准确地找到解决问题的方法。

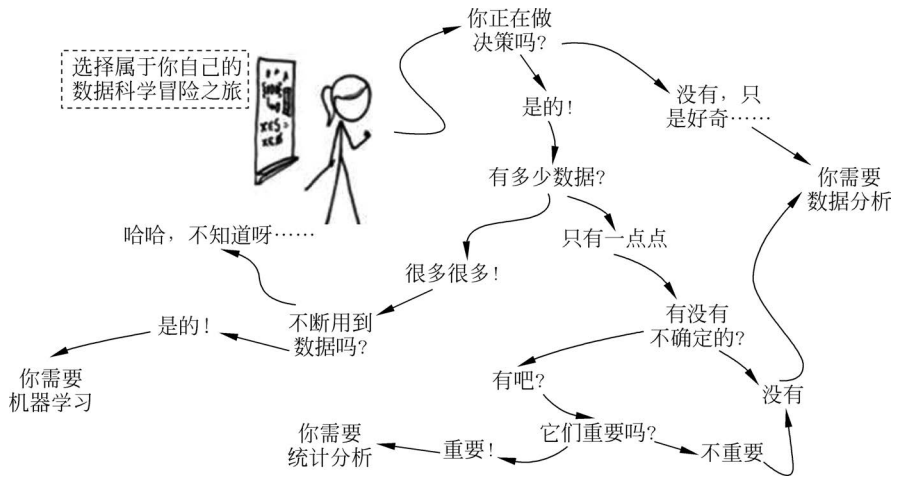


图 5.10 “从数据到知识”的选择之旅

想一想 5.5:数据收集要考虑什么

数据收集有许多缺陷,必须小心翼翼,至少需要问以下一些问题。

- 你选择的算法需要在这些数据上训练什么?它要求什么样的数据格式吗?这个算法需要多少训练数据量呢?对数据质量有什么要求?
- 数据来自哪里?谁拥有该数据集?
- 获取这个数据集的成本是多少?得到它需要多长时间?是否有必要通过谈判(甚至签订法律合同)来获得这些数据的访问权限?
- 你将获得的数据集的数据格式与在生产系统中的数据的一致性如何?是否需要对训练数据进行预处理?数据是否需要标注?

- 你需要多大的数据基础设施来存储这些数据集？
- 在构造初始数据集之后，如何收集新数据？
- 是否有这种可能性，你的组织有一些数据，但你的团队没有权利访问它？你不能访问一些你的组织已经拥有的数据，这种情形会经常发生吗？由于某些原因，数据可能是机密的吗？有道德规范、规章制度或公司隐私等政策约束吗？

思考题

1. 数据科学认知误区有哪些？你能举例描述吗？

2. 什么是利润曲线？它与数据科学项目实施成功有什么关系？

3. 什么是描述性分析、预测性分析与规范性分析？这三类分析与 DIKW 的对应关系怎样？如何理解？

5.4 探究与实践

1. 体验“让数据讲故事”。

请登录“百度指数”网站（http://index.baidu.com/v2/index.html#/），完成以下操作。

（1）确定一个你所关注的主题（一个或多个关键词），选择合适的时间段（如近 30 天）。

（2）浏览三个不同方面的所有相关信息（趋势研究、需求图谱、人群画像）。

（3）从中选择一个你最感兴趣的显示结果，截图保留（尽量完整）。

（4）分享你此刻的所思所想，你对“数据密集型科学发现范式”有何理解（可结合 DIKW 模型）？

2. 理解数据科学流程——基于数据驱动的决策。

以具体爬虫操作及结果为例，理解 DIKW，理解数据科学流程，理解数据驱动的魅力。参考步骤如下。

（1）准备工作。

① 下载安装免费的“八爪鱼爬虫”软件（https://www.bazhuayu.com/）。

② 给出爬取网页的中文名称（如京东、豆瓣等）。

③ 给出爬虫的“配置参数”（链接网址或关键字等）。

（2）数据采集（D）——爬虫及保存。

① 给出体现正在采集过程的截图。

② 给出爬虫结果的简单描述（总条数、采样时间等）。

（3）数据集成（I）——数据集成及存储。

① 用 Excel 打开你的爬虫数据，并截图。

② 给出数据集的所有特征（变量）的名称及类型描述（如文字性、数值型等）。

（4）数据分析（K）——获取数据价值。

① 选择你感兴趣的一个(或几个)变量进行分析以获取价值信息(均值、最大最小值、统计图等)。

② 需要进行哪些数据清洗工作？给出清洗过程及结果的描述。

③ 给出你获得最终的有价值结果的文字描述并至少上传一个截图。

(5) 智慧决策(W)——商业价值。

如果你是某行业的决策者,基于上面的分析结果(K),你可能采取哪些行动？给出简要的理由。

(6) 简述你的收获、体会、疑惑及畅想。