目录

第 1 章数据科学概论

1.1 数据和大数据

1.1.1 数据

1. 数据的定义和类型

今天的人们对于“数据”二字，一定不会感到陌生。翻开书本，打开手机或计算机，甚至不必我们自己去搜寻，就已经有各种各样的数据源源不断地向我们涌来。大至政府发布的各种经济数据和税务数据，小至物价数据、气温数据和身体的健康数据，可以说，我们生活在一个完全离不开数据的世界。

在特定背景下的数据中蕴含的信息能够帮助人们做出合理的决策。政府可以通过统计数据制定合适的政策，健身教练可以根据人们的身体健康数据为我们制订合适的训练计划，我们自己也可以根据天气数据决定今天如何着装等等。数据的重要性不言而喻。

不同的学科中对数据的定义是不同的。统计学中的数据[1]，是指为了找出问题背后的规律而需要的，与问题相关的变量的观测值，是对客观现象进行计量的结果。计算机科学中的数据[2]，是指所有能输入到计算机，并被计算机程序处理的符号，是用于输入电子计算机进行处理，具有一定意义的数字、字母、符号和模拟量等的通称。

从上面两个定义中不难看出，数据只有在特定的背景下才是有意义的，对数据的研究不能脱离其产生背景。**本书从数据科学的角度，将数据定义为，在一定背景下有意义的对于现实世界中的事物定性或定量的记录。**

数据可以有多种分类方式。依据结构分类，可以分为结构化数据和非结构化数据。比如，数字、字符、日期等属于结构化数据类型，而文字、图片、视频、音频都属于非结构化数据。依据形式分类，可以分为文本数据、数字数据、声音数据、图片数据、视频数据等等；依据来源分类，可以分为观测数据和实验数据；如何对数据进行分类，取决于我们想要用数据解决什么样的问题。

2. 数据的 DIKW 模型

正如上一部分所说，我们研究数据是为了得到数据背后蕴藏的规律，以指导人们做出正确的决策，帮助人们解决在现实中遇到的问题。在这个过程中，有四个概念需要读者理解，它们分别是数据、信息、知识和智慧。其中，数据处于相对表象的位置；当我们有目的地对数据进行处理，便可以从中抽取出对问题有意义的部分，这便是信息；信息具有一定的时效性，一些信息随着时间的推移逐渐被证明是错误的，失实的，含糊的，缺乏价值的，而另一些信息经过时间的锤炼，逐渐沉淀累积，形成了知识；知识比信息更具抽象性、逻辑性和价值性，而最后的智慧便是知识积累到一定程度而产生的一种对于规律的掌握。这样的描述或

许有些抽象，为了帮助读者梳理这几个概念之间的区别与联系，这里简单介绍信息科学和知识管理领域著名的 DIKW 模型。

DIKW 模型也被称为知识金字塔。它的前身，其实源自一首名为《岩石》[3]的诗，作者是诺贝尔奖获得者 Thomas Stearns Eliot。诗中有这样一句：“我们在知识中失去的智慧在哪里？我们在信息中丢失的知识在哪里？”。后来，这一概念被 Milan Zeleny[4]、Russell Ackoff[5]等人不断地扩充和细化，最终形成了现在我们看到的 DIKW 理论，如图 1-1 所示。

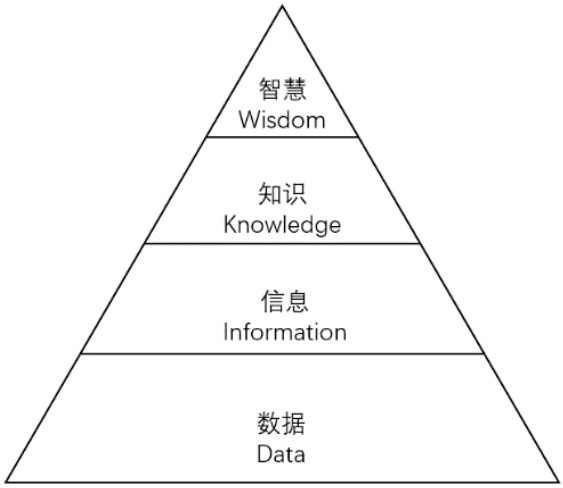

图 1-1 DIKW 模型图

（1）数据（Data）：数据位于模型的第一层，也是模型中的“原始材料”。它是对客观事物的数量、属性、位置及其相互关系等进行的表示，以便系统对其进行保存和处理。

（2）信息（Information）：信息位于数据的上一层。它具有一定的时效性，且具有一定意义，是已经过加工处理，并对决策有指导作用的数据流。

（3）知识（Knowledge）：知识位于信息的上一层。它是经过人类长期选择与积累的，具有价值的信息。

（4）智慧（Wisdom）：智慧位于模型的顶层。它是人类所具备的，基于已有知识和相关信息对问题进行分析和解决的能力。这种能力运用的结果是将有价值的信息挖掘出来，并使之成为已有知识结构的一部分，进而促进智慧的产生。

数据科学的任务就是以数据为研究对象，提炼出数据中蕴含的对决策有益的信息和知识。

1.1.2 数据化进程

数据是对真实世界的记录。人类最早数据记录的产生，可以追溯到三万年前的旧石器时代。那时的人类祖先，就开始在岩石、洞穴上，绘制描述自然生活的壁画。法国肖维岩洞壁画是人类已知最早的史前艺术，创作年代距今约 3 万 2 千至 3 万 6 千年。壁画的内容大都为动物和捕猎的人类，贴近生活，反映了该时期的部分生活风貌。这些壁画在当时具有何种用途，我们已经无从得知，但这些壁画真实的记录了当时人类的生活状况，为发现它们的后人打开了一扇通往神秘远古时期的大门。这是人类通过外部媒介记录自己精神状态的开始。

到了新石器时代，一些早期的社会形态逐渐形成，人们对于记录的需求日益增强，出现

了各种各样的记事方法，其中使用较多的有结绳记事。在我国，《易经·系词》是有关结绳记事的最早文献记载，其中提到，“上古结绳而治，后世圣人易之以书契。百官以治，万民已查。”直到近代，一些民族依旧沿用了结绳记事。比如，我国哈尼族、瑶族、独龙族、高山族等少数民族直到上 20 世纪 50 年代，依旧保留了这种记事方法。结绳记事的方法使得人类的记忆凭借着外物得以延续，在人类历史上有着十分重要的意义。

除了结绳记事，这一时期的人类还掌握了很多别的方法，比如，刻木记事、编贝记事、积石记事等等。这些通过实物记事的方法，使数据信息在人类大脑以外的地方得到保存，前人的智慧、经验、教训得到了更大范围和更长时间的传承。文明成果得以积累，文明发展速度开始加快，我们今天所熟知的数据记录形式——文字开始逐渐形成。

目前，考古发现的最早的真正意义上的文字，是公元前 3200 年左右乌鲁克古城中刻有象形符号的泥板文书，这是最早的楔形文字，也是世界上最早的文字记载。最古老的文字外观并不像楔形，只是一些平板图形。而随着人类文明的发展和交流范围的扩张，原始图形无法满足应用需求，于是苏美尔人逐渐简化符号，增加其意义，使得象形符号逐渐过渡为以音节表意的抽象楔形文字。事实上，汉字也起源于图画，之后从图画逐渐抽象为图案符号，再由图案符号逐渐抽象为具有意义的文字单元，这一过程持续了几千年之久。目前学术界公认的最早的汉字，是殷商时期，刻在龟甲和兽骨上的甲骨文和铸造在青铜器上的金文，存在的时期约为公元前 17 世纪到公元前 11 世纪。

文字的出现是人类文明史上的一个巨大进步，人类终于不再只使用大脑存储记忆，使用语言口口相传信息，或者仅仅使用简单的工具，简略记录见到的事物和发生的事情。文字出现之后，人类在实践过程中的所见所得，所思所想，宝贵的经验和知识，通过文字得以广泛传播，长久传承，属于人类文明的知识和智慧才能够开始积累，文明化的进程进入了一个新的阶段。

而随着文字产生的另一个事物，就是数字。人类早期的结绳记事等实物记事方法中，其实就蕴含着计数的思想。比如，部落中需要记录人数，那么有几个人就在绳子上系几个结，从这一点可以看出，计数源于人类的生活需要。当文字产生之后，随之产生了各式各样的数学符号。发展到后期，产生的较为成熟且一直沿用至今的，便是由印度人发明，由阿拉伯人改造并传播到西方的阿拉伯数系。印度数字在公元前 3 世纪就已经出现，在经过阿拉伯人的使用流通之后，随着阿拉伯鼎盛时期的远征，传入了欧洲。1202 年，数学家 Fibonacci 发布著作《计算之书》，标志着印度数字在欧洲获得认可。后来，人们就将其称为阿拉伯数字系统，也是今日最为常见的全球通用的一种数据形式。

数字的出现使得人类对事物的描述开始变得精准量化，为一系列高级计算方式的诞生提供了可能，这也是为什么有人会说，数学每往前前进一小步，人类文明就往前前进一大步。

回顾历史，算盘是人类历史上最早的用来计算的专门工具。关于其准确的产生时间，学者们说法不一，各个地区的算盘也有不同的外在结构和使用方法。但在中国，算盘的产生，大约可以追溯到汉朝时期的一种更为简单的工具——算筹。算盘由算筹在实际应用的过程中长期改进而来，并于宋元时期广泛流行。使用算盘的计算称为珠算，珠算有对应于四则运算的相应法则。这说明在这个时期，人类就已经具有了通过工具，计算数据量较大且较为复杂

难解的问题的能力。

后来，欧洲逐渐产生了一些机械计算器。17 世纪中叶，法国数学家 Blaise Pascal 发明滚轮式加法器，可以透过转盘进行加法运算。几十年后，德国数学家 Gottfried Wilhelm Leibniz 将其进行改造，制作出可以进行四则运算的步进计算器 1820 年之后，机械式计算器得到了广泛使用，也随之产生了一系列其他类型的机械式计算器。除此以外，十九世纪还诞生了基于穿孔纸带的计算器。1801 年，法国人 Joseph Marie Jacquard 在前人创造的基础上发明了提花织布机，利用打孔卡控制织花的纹样，这是可编程化机器的里程碑。1822 年，英国科学家 Charles Babbage 制造出了第一台差分机，可以处理 3 个不同的五位数，并且精度达到了 6 位小数。1834 年，他提出了分析机的概念，并将机器分为三个部分：堆栈、运算器、控制器。而他的助手，Ada Lovelace，著名诗人 George Gordon Byron 之女，为分析机编制了人类历史上第一批计算机程序，她也成为了世界上第一位程序员。他们的工作相较真正计算机的出现，超前了一个世纪以上，为后来计算机的出现奠定了坚实的基础。

除了计算水平因各式计算器的出现获得了突飞猛进的发展，人类信息社会在这一时期所产生的数据类型也日渐多样化。除了传统的社会生活方方面面所产生的文字，数字以及绘画类型的数据，其他类型的数据记录，比如照片数据、音频数据、视频数据伴随着人类的发明创造逐渐产生。

20 世纪初期，在英国数学家 George Boole 创立了布尔代数这一数字计算机的基础理论，英国工程师 John Ambrose Fleming 利用爱迪生效应发明了电子管之后，1913 年，麻省理工的教授 Vannevar Bush 制造出了第一台模拟式计算机微分分析仪。而第一台电子计算机的发明人是美国人 John Vincent Atanasoff。他和他的学生 Clifford E. Berry 于 1939 年 10 月，研制了人类第一台电子计算机。Atanasoff 把这台机器命名为 ABC（Atanasoff-Berry-Computer）。此后，科学家和工程师们对早期体型巨大并且价格昂贵的计算机进行了一次次的改进和优化，为人类迎接信息时代提供了高效的工具。

20 世纪 50 年代，通信领域的学者们开始意识到，不同计算机用户之间也有着通信的需求，于是他们开始对分散网络、排队论、分组交换等展开研究。1960 年，美国国防部高等研究计划署，出于冷战考虑创建了 ARPA 网。此后，网络技术日益进步，ARPA 网络逐渐成为了互联网发展的中心。1973 年，ARPA 网络被扩展为互联网，接入了来自英国和挪威的计算机。在互联网几十年的发展过程中，ARPA 的 Robert Elliot Kahn 和斯坦福的 Vint Cerf 提出了 TCP/IP 协议，Timothy John Berners-Lee 在瑞士欧洲核子研究组织构建了万维网项目，如今，互联网已经达到了高度普及的程度。根据中国互联网信息中心在 2020 年 4 月发表的报告，截止 2020 年 3 月，中国的网民规模达 9.04 亿[6]，为世界首位。世界各地的人们都在互联网上分享、下载、上传各种类型的数据，庞大的互联网数据正成为一种全新的数据的表现形式，相应的并行计算、分布式计算、集群计算和云计算技术等的出现，也为数据科学的研究指明了未来的方向。

今日，我们被生活中方方面面的数据包围。面对大数据，我们依旧在努力创造更有力的计算工具，设计更科学的计算流程。希望身处于由我们自身创造出的数据迷雾中的我们，能够找到这些数据之后蕴藏的关于人类本身，关于我们生存的这个世界的本质规律。

1.1.3 大数据

1. 大数据的定义

跟随着信息化的浪潮，海量数据的产生和流转已经成为了常态，我们已经真正的进入了大数据时代。梅宏院士曾在《中国信息化周报》中发表文章说道[7]“所谓大数据，是信息化到一定阶段之后的必然产物。”那么大数据究竟是什么呢？

在 1998 年的 USENIX 大会上，美国硅图公司的首席科学家 John Mashey 首次提出了“大数据”这一概念，发表了名为《大数据与下一次基础设施压力的浪潮》[8]的报告。接着，在 2000 年，宾夕法尼亚大学的经济学家 Francis Diebold 发布了报告《宏观经济评估与预测的大数据动能因素模型》，再次提及了这一概念。2011 年，牛津大学客座教授 Viktor Mayer-Schönberger 开始在经济学人杂志上发布一系列大数据专栏文章，并出版了被视为大数据研究的先河之作的《大数据时代：生活、工作与思维的大变革》[9]。至此，这一概念才逐渐走进人们的视野，并迅速占据了人们的注意力。

对于这样一个产生时间不久的概念，业内还没有一个统一的说法，从不同的角度会对大数据这一概念产生不同的理解，其中最有名的是以下几种定义：

数据科学家 John Rauser 认为，大数据是指任何超过了一台计算机处理能力的数据。

咨询公司麦肯锡将其定义为[10]，大数据是无法在一定时间内用传统数据库软件工具对其进行抓取、管理和处理的数据集合。

咨询公司高德纳将其定义为，大数据是大量、高速、或多变的信息资产，它需要新型的处理方式去促成更强的决策能力、洞察力与最优化处理。

以上的定义都指出，首先，大数据依旧是数据，或数据相关的过程，其次，大数据的规模并非一定要达到某一确切的数值，关键在于，是否超过了实际情况下的数据存储能力和数据计算能力。可见，大数据这一信息化的自然产物，对现存的技术手段来说是巨大的挑战，但如果着眼于挖掘海量数据背后蕴藏的丰富规律，它无疑也是我们这个时代的机遇和财富。

2. 大数据相关定理与模型

1）5V 模型

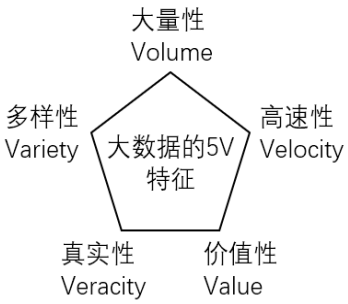

图 1-2 大数据特征的 5V

目前对大数据本质特征的总结中，有最初的 3V 模型，由高德纳公司的高级分析师 Doug Laney 提出。之后人们对这一概念进行了扩充，最具代表性的是 IBM 公司在随后的提出的 4V 模型和现在演变出的 5V 模型[11]，即，多样性(Variety)、大量性(Volume)、高速性(Velocity)、价值性(Value)、真实性(Veracity)。

（1）多样性(Variety)：大数据的来源与类型多样。比如，从生成类型上分为交易数据、交互数据、传感数据；从数据来源上分为社交媒体、传感器数据、系统数据；从数据格式上分为文本、图片、音频、视频、光谱等；从数据关系上分为结构化、半结构化、非结构化数据；从数据所有者分为公司数据、政府数据、社会数据等。

（2）大量性(Volume)：聚合在一起供分析的数据规模非常庞大。谷歌的前执行董事长艾瑞特·施密特曾说，现在全球每两天创造的数据规模等同于从人类文明开始至 2003 年间产生的数据量总和。

（3）高速性(Velocity)：数据的增长速度快，同时要求数据访问、处理、交付等的速度快，实时性要求高。比如搜索引擎要求几分钟前的新闻能够被用户查询到，推荐算法尽可能要求完成实时推荐。

（4）价值性(Value)：尽管我们拥有大量数据，但是发挥价值的仅是其中非常小的部分。大数据背后潜藏的价值巨大。比如通过对于社交网站的用户信息进行分析，广告商可根据结果精准投放广告。

（5）真实性(Veracity)：一方面，对于大量的数据需要采取措施确保其真实性、客观性；另一方面，通过大数据分析，真实地还原和预测事物的本来面目也是大数据应用的内在要求。

2）5R 模型

除了 5V 模型，如果从数据管理的角度认识从大数据中获取有用信息的过程，还可以得到由 Stidston 提出的 5R 模型[12]。该模型包括大数据的相关特性(Relevant)、实时特性(Real-time)、真实特性(Realistic)、可靠特性(Reliable)，以及投资回报特征(Return on investment, ROI)。

关于 5R 模型与 5V 模型的区别与联系，计算机科学家吴信东曾在文章《从大数据到大知识: HACE + BigKE》中提到[13]：

从 5R 模型的内容来看，它和 5V 模型具有类似的地方。它们都着眼于大数据的本质特征。相比较而言，5R 是基于商业用途而提出，它对于大数据的五大特征的描述是基于数据管理在商业上的应用进行阐释。从数据管理的角度来看待大数据，其关键在于数据的组织形式。大数据的海量多源异构特征已经得到了普遍的认可。针对这些特征，采取一种怎样的数据组织形式以提升数据收集、存储、处理和应用的效率，获取对商业发展与决策具有价值的“知识”，是 5R 模型中提出的需要解决的问题。

3）4P 模型

信息化医疗系统的推广使得医疗数据也具有了大数据的特点，针对医疗数据体量的庞大，

以及在医疗诊断中病因与病状之间多样化的复杂对应关系，在医疗大数据的环境中产生了医学 4P 模型[13]，包含预测性(Predictive)、预防性(Preventive)、个体化(Personalized)、参与性(Participatory)。该医疗模型基于大数据，对疾病做出预测，并基于个人数据对病人做出个性化的服务。同时，诊疗过程中的数据将再次被记录到数据库中，从而为病人提供基于大数据的健康建议。5V 和 5R 模型主要阐述了大数据的本质特征，而 4P 模型概括了大数据与医疗模型的结合。

4）HACE 定理

除此之外，吴信东基于大数据的本质特征，提出了 HACE 定理[13]和与其关联的大数据处理框架，从大数据的来源，大数据复杂的数据结构，以及数据之间的关系这三个方面，对大数据的特征进行了阐述。

HACE 定理将大数据描述为，始于异构（Heterogeneous）、自治（Autonomous）的多源海量数据，旨在寻求探索复杂的（Complex）和演化的（Evolving）数据关联的方法和途径。其中异构和自治主要是针对大数据的数据源而言的，比如，物联网中的每一个独立的传感器和万维网中不同的作者和读者都可以作为数据源，他们产生的数据也具有不同的媒体形式和表现形式。大数据分析就是从这些异构自治的多源数据中，探索出随时间和空间演化的数据关联。

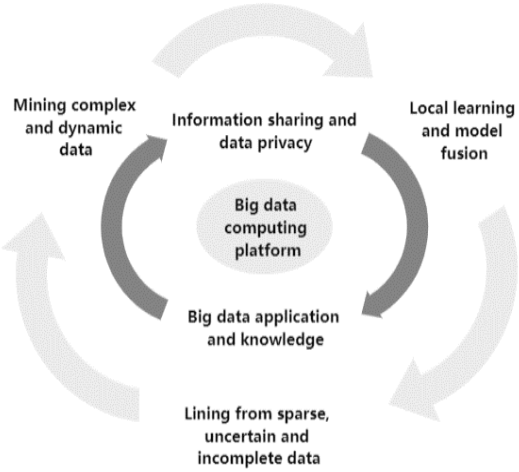

图 1-3 基于 HACE 定理的大数据三层架构[13]

基于 HACE 定理，文章中还提出了大数据处理的三层框架[13]。

框架的第一层是大数据的计算平台。这一层提出使用带有高计算性能的集群计算机。它的特点在于，每一个计算节点都可以并行处理计算任务，使得单个计算机上的计算量有所降低，从而减小对每个计算节点的硬件依赖性。

框架的第二层是大数据的语义和应用知识，包含数据共享与隐私、领域和应用知识的问题。基于第一层的大数据计算平台，我们需要分析大数据中的隐含知识。在对大数据下的隐含知识分析的过程中需要数据的共享，而这也会带来信息隐私的问题。因此，第二层从存储角度，对访问数据的权限进行了限制，从信息共享的渠道，对数据的一部分特征进行匿名化，

使得数据中的敏感信息得到模糊处理，保证了数据的安全性。

框架的第三层是大数据分析算法。针对不同类型的数据和问题，提出大数据挖掘的具体算法；例如，局部学习，多信息源的模型融合，稀疏不确定和不完整的数据挖掘，动态的复杂数据的挖掘等方面。

1.2 数据科学理论基础

1.2.1 数据科学发展历程

1974 年，图灵奖获得者 Peter Naur 在美国和瑞典出版了《计算机方法简明调查》[14]。这本书调查了当时被广泛运用的数据处理方法，首次提到了“数据科学”的概念。作者也对“数据科学”做出了简单的定义：它是研究处理数据的科学，一旦被建立起来，数据和数据所代表的意义之间的关系需要通过其他领域和其他科学去解释。

1996 年，IFCS（International Federation of Classification Societies）的成员召开名为“数据科学，分类和相关方法”的会议，“数据科学”首次在会议的主题中出现。

1997 年，密歇根大学的 H. C. Carver 学院举行了统计学教授的就职演讲。吴建福（CF Jeff Wu）教授在演讲中提议，将统计学改名为数据科学，将统计学家改名为数据科学家。

2001 年 William S. Cleveland 发表文章《数据科学：一个用来扩大统计学领域技术范畴的行动计划》[15]。作者指出，这种对统计学领域中技术工作的扩展是实质性的，能够给该领域带来的变化，因此，新的领域应该有新的名字，叫作数据科学。

2002 年 4 月，国际数据委员会（CODATA）创立了学术期刊 *DataScience Journal*，这是首个关于数据科学的学术期刊。*DataScience Journal* 的创建被认为是 CODATA 成立以来迈出的最重要的一步。

2003 年 1 月，*Journal of datascience* 创刊，不同于上面提到的学术期刊 *DataScience Journal*，该杂志提供了一个人人参与的交流平台，数据工作者们可以发表自己的见解，促进该领域的学术交流。

2007 年，复旦大学成立数据学和数据科学研究中心。中心成立后，每年都会举办关于数据学和数据科学的研讨会。两年后，中心研究员朱扬勇和熊赟出版了《数据学》[16]。作者提出了信息化、CYBER 空间、数据爆炸、数据界等概念；并指出，数据学和数据科学是关于数据的科学或研究数据的科学，可以将其定义为，“研究探索 Cyberspace 中数据界奥秘的理论、方法和技术，研究的对象是数据界中的数据”。作者还指出，“与自然科学和社会科学不同，数据学和数据科学的研究对象是 Cyberspace 中的数据，是新的科学”。

2009 年 1 月，谷歌首席经济学家 Hal Ronald Varian 曾说，“未来十年最受欢迎的工作将是统计学家”。2009 年 6 月，统计学家 Nathan Yau 在 FlowingData 上发表名为《数据科学家的崛起》的文章[17]，文章对 Hal Ronald Varian 提出的“统计学家”进行探究，将其解释为，“能够从大型数据集中提取信息，并具备计算机科学，数学及统计学，数据挖掘等能力的人，而数据科学家正是能够做到这一切的人”。

2010 年 2 月，数据编辑 Kenneth Cukier 在《经济学人》中发表特别报告提出，“数据科学家作为一种新的职业出现，他们具备了软件程序员、统计学家和讲故事者的技能，用来提取大量数据背后隐藏的规律”。

2010 年 9 月，Drew Conway 在文章中指出[18]，“能够胜任工作的数据科学家需要学习很多方面东西”，并将其以韦恩图的形式总结为黑客技能、数学和统计知识、以及专业领域知识。

2012 年 9 月，Tom Davenport 和 Dhanurjay Patil 在《哈佛商业评论》上发表名为《数据科学家：21 世纪最有魅力的工作》的文章[19]。

2015 年 2 月 18 日，美国白宫宣布 Dhanurjay Patil 成为白宫首位数据政策副首席技术官兼首席数据科学家。在向公众发表讲话时，Dhanurjay Patil 表示，“美国首席数据科学家的使命，就是负责任地释放数据的力量，使所有美国人受益”。

近几年来，数据科学开始广泛的进入人们的视野，相关研究的文献数量迅速增加。这一新兴学科的发展历史虽然并不长，但如今发展迅速，对数据科学的研究，也能使人们更好地面对大数据时代的机遇和挑战。

1.2.2 数据科学的概念

什么是数据科学？从字面意思理解，数据科学就是以数据为中心的科学。对于这样一门起源时间并不长的“年轻”学科，业界在目前还没有一个统一的定义，但几十年来，也有很多学者和机构发表过自己的见解。

最早提出数据科学概念的 Peter Naur 将数据科学简单定义为[14]，研究处理数据的科学，它一旦被建立起来，数据和数据所代表的意义之间的关系需要通过其他领域和其他科学去解释。

美国计算机科学家 William S. Cleveland 认为[15]，随着计算机科学的发展而扩展的，统计学中数据分析的技术领域，叫作数据科学。

复旦大学的数据科学研究中心把数据科学定义为[16]，关于数据的科学，或者研究数据的科学，是用来研究探索 Cyberspace 中数据奥秘的理论、方法和技术。

李国杰院士在《对大数据的再认识》一文中说道[20]，“数据科学是数学（统计、代数、拓扑等）、计算机科学、基础科学和各种应用科学融合的科学，类似钱学森先生提出的‘大成智慧学’”。

我们可以从上一小节的数据科学发展史，和这一小节中对数据科学概念的梳理中看出，数据科学并非一个全新的领域，而是起源于统计学的，为了探索当前学术界和工业界所产生的大量数据中蕴含的信息与知识，结合了诸如大数据、计算机科学、机器学习、数据挖掘等新的技术和理论的一门新兴交叉学科。

1.2.3 数据科学的主要内容

1. 研究内容和研究对象

数据科学通过一系列科学的流程，研究现实世界方方面面产生的数据，从而完成从数据中抽取出信息和知识的任务，发现事物背后隐藏的规律，最终使数据的集成度更高，价值密度更大。因此，数据科学依赖数据，也依赖研究数据的方法。那么数据科学的研究内容是什么呢？针对这一问题，鄂维南院士曾在《数据科学导引》的绪论中有着这样的概括[21]：

数据科学主要包括两个方面：用数据的方法研究科学和用科学的方法研究数据。前者包括生物信息学、天体信息学、数字地球等领域；后者包括统计学、机器学习、数据挖掘、数据库等领域。这些学科都是数据科学的重要组成部分，只有把它们有机地整合在一起，才能形成整个数据科学的全貌。

数据科学对什么对象进行研究？如果粗略的看，我们可以认为数据科学的研究对象就是现实世界中来源不同、类型不同的数据。但由于数据科学具备科学的性质，其科学意义上的研究对象也同样有待进一步的探讨。比如，李国杰院士就曾在《大数据研究的科学价值》中，针对数据科学的研究对象进行了深刻的讨论，他在文章中提到[22]：

计算机科学是关于算法的科学，数据科学是关于数据的科学。但任何研究领域的研究，若要成为一门科学，研究内容一定是研究共性的问题。数据研究能成为一门科学的前提是，在一个领域发现的数据相互关系和规律具有可推广到其他领域的普适性。事实上，过去的研究已发现，不同领域的数据分析方法和结果存在一定程度的普适性。但抽象出一个领域的共性科学问题往往需要较长的时间，提炼“数据界”的共性科学问题还需要一段时间的实践积累。计算机界的学者至少在未来 5 至 10 年内，还需要多花一些精力协助其他领域的学者解决大数据带来的技术挑战问题。通过分层次的不断抽象，大数据的共性科学问题才会逐渐清晰明朗。技术上解决不了的问题积累到相当的程度，科学问题就会浮现出来。（有删减）

2. 理论体系

数据科学作为支撑大数据研究与应用的新兴交叉学科，其理论基础来自多个不同的学科领域。2010 年 9 月，美国数据科学家 Drew Conway 第一次使用韦恩图定义了数据科学的理论体系。

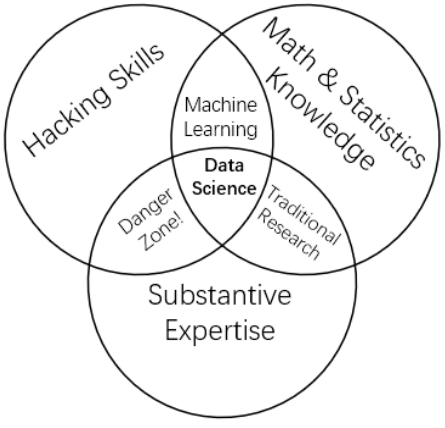

图 1-4 Drew Conway 提出的数据科学韦恩图[23]

在图 1-4 中，黑客技能（Hacking skills）指，在收集数据、清理数据、处理数据、分析

数据等一系列流程中，需要用到的计算机科学、人工智能等方面的方法与技术。

数学和统计学知识（Math & Statistics Knowledge）指，在对数据进行分析处理的过程中，需要用到的数学和统计学方法理论。

实质性专业知识（Substantive Expertise）指，数据科学工作中涉及到的实质性领域知识。对于数据科学家来说，发现问题的能力离不开实质性专业知识的掌握，正如俗语所言“外行看热闹，内行看门道”。而强大的发现问题的能力正是数据科学家与一般数据分析师的不同之处。

黑客技能与数学和统计学知识的交叉区域是机器学习（Machine Learning）领域。数学和统计学知识与实质性专业知识的交叉区域是传统研究（Traditional Research）领域。而三者重叠的区域便是数据科学（Data Science）领域。这说明数据科学的理论基础应该是这三种理论知识的结合。

可以注意到，黑客技能与实质性专业知识的交叉区域是危险区域。对此，Drew Conway表示，这并不代表同时具备两方面的理论基础就会带来危险，而只是因为缺少统计学知识可能会对数据科学的工作带来损害。这也体现了数学与统计学知识在数据科学领域中的重要性。

该韦恩图以其形式的简洁明了，一经发布便受到了业界的广泛好评。如今我们见到的文献中，也大多引用此图对数据科学的概念进行介绍。受到该图的启发，在此之后，众多业内人士也纷纷提出自己对数据科学理论体系的见解。其中较为著名的是KDD会议联合创始人、数据科学家 Gregory Piatetsky-Shapiro 提出，Matthew Mayo 于 2016 年在 KDnuggets 上发布的数据科学韦恩图。

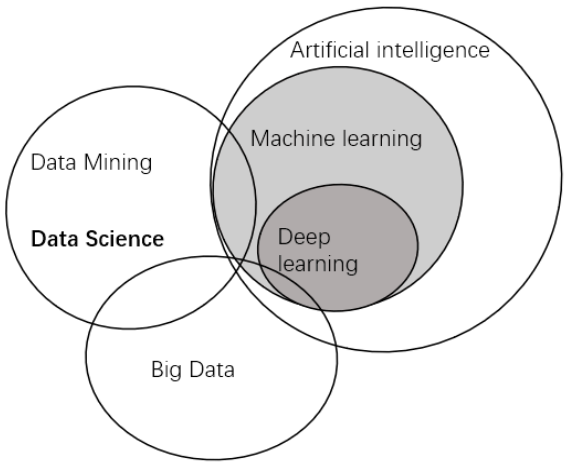

图 1-5Gregory Piatetsky-Shapiro 数据科学韦恩图[24]

如图 1-5 所示， Piatetsky-Shapiro 整理了数据科学与其他几个相关领域之间的关系。Matthew Mayo 在发布的文章中对其进行了解释。他认为[24]：

大数据作为“超出常用软件工具捕获，管理和处理能力”的数据集，是所有数据相关科学邻域的基础。

机器学习是数据科学的核心。数据科学是为了从数据中获取知识和智慧，那么机器学习就是让这个过程自动化的引擎。它使用样本进行推断和预测，这一点与传统的统计学有很多

共同点。机器学习与数据挖掘也有着关联，数据挖掘是一个过程，而机器学习被用作工具来提取数据集中的潜在价值。

数据挖掘对数据科学也至关重要。Fayyad，Piatetsky-Shapiro 和 Smyth 将数据挖掘定义为“从数据中提取模式的特定算法的过程”，这一概念在数据科学的概念推广之前就已经大受欢迎。但数据挖掘更多的被视为一个过程，数据科学是一门科学，它既是数据挖掘的同义词，也是包含数据挖掘的概念的超集。

深度学习是一个相对较新的概念，它是应用深度神经网络来解决问题的过程。它不会取代所有其他的机器学习算法和数据科学技术，但可以以额外的过程和工具的形式为数据科学提供大量帮助以解决问题，它是数据科学领域的一个有价值的补充。

数据科学包含机器学习和其他分析过程，统计学和相关的数学分支，而且越来越多的借鉴了高性能的计算。所有这些都是为了最终从数据中提取知识和智慧，并使用这些新发现的信息来讲故事。这些故事需要以可视化的形式展现，主要应用于某些具体的行业和研究。数据科学使用来自各种相关领域的各种不同工具。（有删减）

3. 数据科学与第四范式

2007 年 1 月，图灵奖得主 Jim Gray，在演讲中提出了“指数级增长的科学数据”背景下的数据密集型科学研究的第四范式。他认为，科学范式在历史上经历了几次变革，几千年前，科学的星星之火刚刚点燃时的实验科学范式；几百年前以牛顿的经典力学，麦克斯韦理论解释的电磁学，所代表的理论科学范式；到几十年前的计算机科学范式，再到信息爆炸的今天，我们需要“从计算机科学中把数据密集型科学区分出来，作为一个新的、科学探索的第四种范式”，这就是第四范式的由来。

2009 年 10 月，微软出版了《第四范式：数据密集型科学发现》一书，在 Jim Gray 提出的第四范式概念的基础上进行了展开。译者在文中提到[25]：这是“第一本、也是迄今为数不多的从研究模式变化角度来分析‘大数据’及其革命影响的著作”。书中从地球与环境科学、生命与健康科学、数字信息基础设施和数字化学术信息交流等方面出发，介绍了基于海量数据的科研活动、过程、方法和基础设施，从不同角度介绍了数据密集型科学研究的内容。

2011 年 12 月，CODATA 中国全国委员会召开“数据密集型科研与数据科学研讨暨 CODATA 中委会人才团队建设启动会”，会议对数据科学的基本科学问题，数据密集型科研的特点、面临的问题和挑战、未来的发展方向，如何推动数据科学发展等问题进行了探讨。

数据密集型科学由三个基本活动组成：数据采集、数据管理和数据分析。这里的数据是指大型国际实验室、跨实验室、单一实验室，甚至发展到以后还包括个人生活之中所产生的数据。实验之中涉及到不同种类的学科，并且数据规模巨大，这都是数据密集型科学活动中面临的挑战。

数据科学与第四范式的联系在于，这二者是大数据研究的两大理论基础，前者是更广泛意义上的数据科学，后者是针对科学研究范式而言的。正是由于两位图灵奖获得者对这两个概念的提出和后来的科学家们对它们的扩展，大数据科学研究的理论体系才更加完善，数据科学的发展才越来越快。

1.3 数据科学应用实践

1.3.1 数据科学家

1. 数据科学家的定义

马云曾说，“数据是新一轮技术革命最重要的生产资料”。在数据已经渐渐成为“生产资料”，深刻得影响着社会生活方方面面的当下，为了探索数据中蕴含的巨大价值，业界对数据科学家的需求也在不断增大。我们首先关心的问题是，什么是数据科学家？

2005 年 9 月，美国国家科学委员会发布名为 *Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century* 的报告[26]，这一报告将数据科学家定义为计算机科学家，数据库和软件工程程序员，专业领域专家，专家评论员，图书管理员，以及其他人的组合。这一定义说明在数据科学这样一个交叉学科当中，数据科学家应当具有多领域的技能，或者至少在一个数据科学家团队中，需要有不同学科背景的人。

2010 年 Drew Conway 发布了数据科学韦恩图[23]后，人们逐渐认可了这一简洁的表达，并在此基础上把数据科学家定义为具有计算机科学技术，数学和统计学知识基础和实质性专业理论知识的人。因此也出现了“数据科学家是计算机科学家中的统计学家，统计学家中的计算机科学家”这样有趣的描述。

从数据科学家所做的工作方面考虑，我们可以把数据科学家定义为能够发现现实世界的问题，收集问题相关的数据，抽取数据中的信息，并解释数据背后的规律意义的人。

2. 数据科学家的技能

正如 Rachel Schutt 在《数据科学实战》[28]中所说，数据科学家首先需要对可能存在问题的领域有着深入的了解，发现数据科学需要解决的问题；然后需要用到统计学和软件工程的知识和技巧，对问题相关的数据进行采集、清理和处理；当数据被整理成型之后，数据科学家需要构建模型，设计算法，设计实验进行数据分析；最后还需要对分析结果进行可视化的展示，使用明白无误的语言和图形同别人交流，使他们明白数据背后的规律和含义。

2015 年 SAP 公司的数据科学家 Stephan Kolassa 发布了一个数据科学家韦恩图[27]，以此来说明数据科学家需要具备的技能。值得注意的是，沟通技能在其中也占据了一席之地。

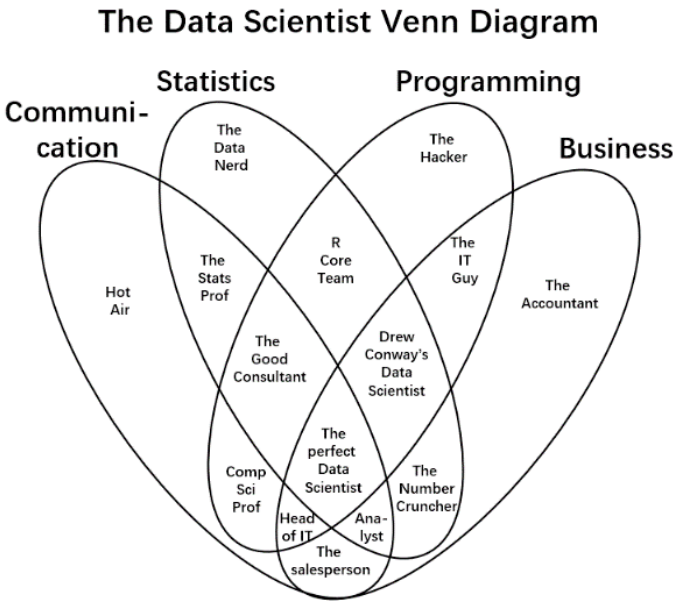

图 1-6 数据科学家韦恩图

从通用技能来说，数据科学家需要具备的技能应该涉及：数据分析、数学、统计学、人工智能、机器学习、深度学习、自然语言处理、工程管理、软件工程、计算机科学、沟通能力等。从技术技能来说，数据科学家的工作需要使用到 Python、R、SQL、Hadoop、Spark、Java、SAS、C++、TensorFlow 等语言、库或是工具。

1.3.2 数据科学工作流程

数据科学工作有一套完整的工作流程，但这套工作流程并非一成不变，根据实际情况的不同，可以不必按照完整的流程进行工作，而只完成工作的一部分或者重复进行某些工作。数据科学工作流程图如图 1-7 所示。

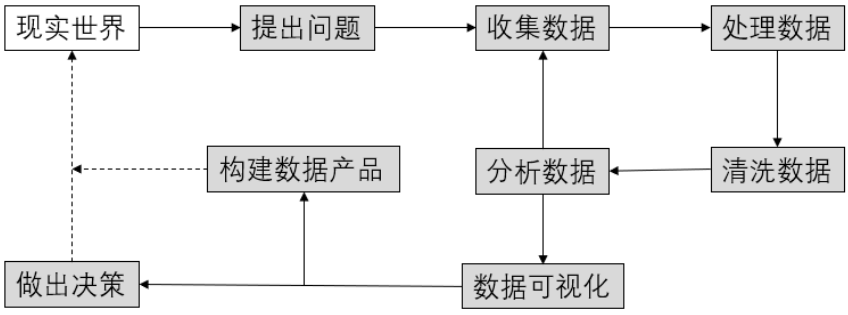

图 1-7 数据科学工作流程图

（1）提出问题：数据科学工作的第一步是从现实世界发现问题并提出问题，这是数据科学家与数据分析师的一个重要区别，数据科学家需要对某一领域的专业知识更加了解，能够提出可以通过数据科学解决的问题。

（2）收集数据：当问题被提出后，数据科学家需要从现实世界中收集大量与问题相关

的原始数据。

（3）处理数据和清洗数据：这一阶段通常是对数据的再加工，通常使用 Python 语言、R 语言、Shell 脚本、SQL 等完成，最终得到格式化的数据，便于分析。

（4）分析数据：分析数据包括两步：第一步是探索性分析；第二步是通过机器学习算法和统计学模型对数据进行分析。探索性分析是指对已有的数据在尽量少的先验假定下进行探索，通过制图、制表、方程拟合等手段探索数据结构和规律的一种数据分析方法，适用于面对大量数据不知如何下手从何处分析的情况。在探索性分析的过程中，可能会发现数据中有重复值，缺失值，或者异常值，一些数据未被记录或错误记录。此时，为了保证数据的正确性和完整性，需要重新对数据进行采集和处理清洗。在分析数据的第二步，通常会使用机器学习的一些算法或统计学的一些模型，比如 K 近邻和朴素贝叶斯等对数据分析模型进行设计，在设计模型时需要对这是何种类型的问题进行考虑，以选取最适合的算法。

（5）数据可视化：在得到分析结果后，数据科学家需要将分析结果可视化，以便于他人理解数据分析的结果，明白数据背后蕴含的规律。

（6）构架数据产品：最后，可以通过数据分析的结果构建数据产品。Rachel Schutt 认为[28]，在数据科学中，构建数据产品的要点在于，在数据分析中将用户与产品的交互形成的反馈数据考虑在内，以此对模型产生的偏差进行调整，这是数据科学与统计学以及单纯的数据分析的一大区别。模型不仅预测未来，它还在影响未来。

1.3.3 数据科学实践案例

数据科学与我们并不遥远。今日，数据科学已经渗透进了我们生活的方方面面，帮助我们解决生活中的大小难题。为了帮助大家更好地理解数据科学工作，引发同学们的学习热情，下面将给大家介绍几个数据科学的实践案例。

案例 1 医疗健康大数据

2017 年中国大数据技术大会公布了十项年度大数据应用最佳实践。其中颇受瞩目的一个是传统医疗行业中的医疗大数据应用系统。

传统医疗行业中累积了类型各异的海量数据，但这些数据在过去并未得到有效的利用。同时，我国的医疗现状存在着医疗资源分配不均，重复诊疗等问题。近年来，国家大力推行分级诊疗制度，移动医疗、远程医疗的发展也随着互联网技术的进步有了很大的突破。而临床一线，医疗事故仍然不时发生。资历尚浅的医生受制于医疗资源分布不均衡，医疗水平提高缓慢，这一情况直接导致了优秀医生的数量严重不足。

医疗大数据的治理，可以在海量医疗数据和医疗行业中的现存问题之间架起一座桥梁。建立临床医疗诊断辅助决策平台，主要目的之一就是提高医疗的安全性和诊疗质量，减少医疗差错，提升病人的就诊体验。

具体来说，医疗大数据是指，在人们健康管理及医疗行为的过程中产生的，与健康医疗相关的数据。它具有医疗大数据的特性，包括体量大、多态性、不完整性、冗余性、时效性

和隐私性。体量大是指，医疗大数据体量巨大。比如，一张 CT 图像中含有的数据量约为100MB，而一个标准的病理图接近 5GB。多态性是指，数据来源多样，形式丰富，包括文本、医学影像等。不完整性指医疗数据的收集和处理过程经常相互脱节，这使得医疗数据库难以全面地反映任何疾病信息。大数据来源于人工记录，这也导致了数据记录的偏差和残缺，许多数据的表达、记录本身，如“大约”“不确定”等也具有不确定性。冗余性是指，同一人在不同医疗机构会产生相同的信息，而在诊疗过程中由于对病情推理的不确定也会产生大量与真实病理的不相关诊疗记录，因此整个医疗数据库包含着大量重复和无关紧要的信息。时效性是指，医疗数据的创建速度快，更新频率高。隐私性是医疗大数据的重要特点。个体的患病情况、诊断结果、基因数据等的泄露将会导致严重后果，侵犯公民权，威胁公众安全。

医疗大数据系统将医疗大数据与机器学习、深度学习等技术，和循证医学、影像组学等学科进行结合，最终达到优化诊疗流程、提升医疗行为效率的效果。

以对医疗文本数据的挖掘为例。电子化的医疗数据方便存储和传输，但是并未达到能够直接进行数据分析的要求，大约 80%的医疗数据是非结构化的文本数据。通过使用自然语言处理技术对文本数据进行结构化，包括数据清洗、短句切分、主干提取、短句聚类、统计筛选、模板整合、模板应用等步骤，使医疗文本达到数据分析的要求。而通过对文本数据进行分析处理，也可以构建医疗知识图谱，为智能系统提供可用的学习材料。医疗知识图谱是一种从海量医疗文本中抽取结构化知识的手段，也可以应用于医疗影像数据。医疗知识图谱通过将图形学、应用数学、信息可视化技术、信息科学等学科的理论与共现分析等方法结合，利用可视化的图谱形象地展示实体之间的关系。

再以医疗影像数据的挖掘为例。医疗影像数据是 X 光、CT、核磁共振等医学影像设备所产生的影像数据的集合，具有数量巨大、维度高和复杂度高的特点，是典型的非结构化数据。作为疾病诊疗的最大信息来源，医学影像数据占全部临床医疗数据的 80%以上。通过深度学习如卷积神经网络，可以学习到医疗影像数据的特征表示。卷积网络通过一系列的方法，能够将数量庞大的图像识别问题不断降维，最终使其能够被训练，读懂医学影像，进行疾病的风险评估。

案例 2 沃尔玛与社交大数据

谈起大数据，不得不谈及沃尔玛“啤酒与尿布”的经典大数据应用案例。早在 20 世纪 90 年代，沃尔玛超市管理人员分析销售数据时发现：在某些特定的情况下，“啤酒”与“尿布”两件看上去毫无关系的商品会经常出现在同一个购物篮中。经过后续调查，美国有婴儿的家庭，一般是年轻的母亲在家照看婴儿，年轻的父亲负责购买尿布。他们在市场一般会顺路购买啤酒犒劳自己。于是，沃尔玛将尿布与啤酒摆在同一区域，两个商品的销售迅即获得增长。

在“啤酒与尿布”中饱尝甜头后，沃尔玛开启了大数据挖掘的大幕，陆续并购大数据企业，增强数据分析与运营实力。特别是在 2011 年，专门成立大数据公司 Walmart labs，旨在通过深度挖掘消费者在社交网站上产生的峰值数据预测商品和消费需求，将这些数据转为有助于决策的信息，通过移动终端向用户进行精准推送。

Walmart labs 成立之初，主要实现两大功能。首先是数据挖掘，分析消费者在社交网络上展现的兴趣，从而预测他们可能在沃尔玛电商平台 walmart.com 下一个购买的产品。其次是发展地理位置科技，实验室的工程师们希望能够开发出一个地理位置应用，引导用户寻找自己感兴趣的商品。

为更好地进行大数据深度挖掘，沃尔玛并购了多个社交网站和移动技术企业，推出了语义搜索服务——加入社交媒体内容，扩大搜索引擎的知识储备，搜索引擎可以更好决定用户所寻找的上下文；利用社交网络上的峰值数据，预测商品需求，将这些数据转为有助于决策的信息，并推送给沃尔玛的客户。

如今，通过自身数据积累整合及并购研发，沃尔玛已然拥有一个涵盖消费者线下交易数据、沃尔玛网络商城电子数据与社交媒体应用数据为一体的实时更新积累的大数据库。沃尔玛大数据可以细化到全球 27 个国家 11457 家门店任一时段的销售数据和销售细节。通过 Walmart labs 工程师们的努力，这些数据会通过计算机系统，从扩散到集中，详尽地呈现顾客消费习惯的变化。通过数据挖掘和分析，得出不同地域、不同购物偏好，为采购、开店决策提供依据，将执行成本降到最低，并且创造新的消费机会。这被沃尔玛称为大脑中枢神经的终端。

在大数据的强力驱动下，沃尔玛业务飞速增长。2014 年，沃尔玛营利 270 亿美元，同比上年增长 1%；全球电子商务销售额约为 120 亿美元，相对于 2011 年翻了 3 倍。2015 财政年度（2014 年 2 月 1 日至 2015 年 1 月 31 日）的净销售金额达到近 4 857 亿美元。2015 年以来，Walmart .Com 线上商品由 4 年前的 100 万种品类增至 700 万多种。

大数据会随着数据的结构化和规模化滚动雪球，越来越“大”，越来越“快”，沃尔玛这个世界上最大的零售商，已经利用大数据技术在电子商务的发展浪潮中抢得先机。

案例 3 大数据与智慧城市

智慧城市(smart city)这一概念发端于 20 世纪 80 年代的信息城市(information city)，经历了 20 世纪 90 年代的智能城市(intelligent city)与数字城市(digital city)，在 2000 年后逐步演化为智慧城市。2009 年 IBM 公司首次提出了智慧城市愿景，使得智慧城市理念与实践在全球范围内迅速传播。目前，在欧洲和北美已有数百座城市宣布建设智慧城市，IBM 公司参与的智慧城市项目多达 2 500 余个，微软、思科、西门子、日立、松下等科技公司以及埃森哲、奥雅纳等商业或工程咨询公司也在积极涉足智慧城市建设，预计至 2020 年智慧城市相关产业市场规模将达到 4 000 亿美元。

数字城市技术把基础地理数据、正射影像、街景景象数据、全景影像数据、三维模型数据结合在一起。在政务网上，通过注册可以进行服务共享；在公共平台、互联网、公网上，通过二次开发可以提供各种交通、导航、旅游、文物、购物等服务系统。

高速发展的物联网能够实现人与人、人与机器、机器与机器的互联互通，实现智慧城市的各种应用。在经济发展方面，可以推动智慧制造、工业互联网、物联网。在文化交流方面，可以考虑智慧户外流媒体、智慧教育、智慧旅游等等。在社会交往方面，有智慧交通、购物、社会综合管理。

智慧城市涉及多个方面的概念，智慧管理、智慧出行、智慧环境、智慧生活等等。

在智慧管理方面。由城市运行所产生的交通、环境、市政、商业等各领域数据量是巨大的，这些数据经过合理的分析挖掘可产生大量传统数据所不能反映的城市运行信息。目前与智慧管理相关的大数据来源主要包括由遍布全市的摄像头收集的视频影像，由各类传感器收集的环境等方面信息，由各类终端收集的刷卡信息，由市民通过手机应用或社交网站贡献的相关信息等。其应用方式主要体现在三个领域。

一是实时监控与突发事件处理。如巴塞罗那和格拉斯哥都计划在全市大规模布置摄像头或传感器以及时识别火灾、犯罪等异常情况；巴西里约热内卢还开设了一座建设有 80 m 宽监视屏的城市运行控制中心，显示来自全市 900 多个摄像头的监控影像，由来自 30 个不同部门的 50 名工作人员对洪水威胁、交通事故、管道泄漏等突发事件做出应急控制。

二是市政服务。如维也纳、波士顿、格拉斯哥都推出(或计划推出)用于报告市政故障的手机应用；而瑞典斯德哥尔摩自 2007 年至今已投资 7000 万欧元开发 50 多项电子服务，并藉此降低了城市的管理成本。

三是公众参与。大数据使人们得以构建反映城市建成环境实时变化的三维可视化系统，这类系统可作为公众参与的平台。如 Autodesk 公司在德国班贝格市(Bamberg)开发的城市三维可视化系统被用于讨论新铁路线建设，市民使用 iPad 即可了解铁路线对周边环境的影响，节省了公众参与的时间。

在智慧出行方面。交通流的合理规划与疏导是几乎所有城市长期面临的问题，而大数据的广泛性与实时性则为解决这类问题提供了新的可能。目前大数据在智慧出行领域的应用主要体现在两方面。一是交通流量实时监控，如利用遍布全市的摄像头监控实时交通流量。二是交通信息实时提供，如通过安装在停车场的传感器为市民提供实时停车位信息，以引导居民合理出行。

在智慧环境方面。在智慧城市概念出现之前，生态城市、低碳城市等概念就已被广泛接受，也是新千年后全球城市发展的关注重点。目前大数据在智慧环境领域的应用主要体现在两方面。一是能源使用管理。安装在电网系统中的传感器可实时收集用户的能耗信息，并按时段调配能源供给或在电力峰值不同的建筑物之间进行电力融通，提高能源使用效率。如伦敦、阿姆斯特丹、西雅图、斯德哥尔摩等许多城市都计划推行智慧电网（Smart Grid），日本千叶与日立公司合作建立了地区能源管理系统（AEMS）。二是环境质量监控。如哥本哈根利用安装在自行车轮上的传感器收集空气质量信息，巴塞罗那利用安装在路灯上的传感器收集噪声、污染信息等。

在智慧生活方面。虽然智慧城市涉及大量技术内容，但其核心价值仍在于为市民提供更高质量的生活，这也是几乎所有国外智慧城市建设项目所不断强调的。目前大数据在此领域的应用主要体现在生活服务方面。如维也纳、巴塞罗那、纽约等城市在开放数据的基础上众包开发了几十种至上百种生活服务类手机应用；多伦多、格拉斯哥等城市则通过云计算等技术对实时信息进行分析并据此为市民提供更多生活服务实时信息。此外，思科公司提出了智慧连接社区概念（Smart + Connected Communities）。通过智能网络系统将社区的服务、信息和人群等各类资源相结合，将物理空间的社区转化为一个更加紧密联系的社区。但也可以看

到，在医疗、教育这两个智慧生活的重要方面，大数据尚未获得较多实质性的应用。

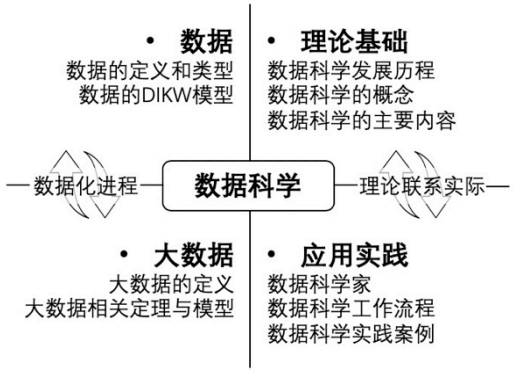

图 1-8 数据科学知识体系图

1.4 小结

本章主要介绍了数据科学的产生背景、基础知识、基本理论、数据科学家和数据科学的实践案例。数据科学知识体系包括数据、大数据、理论基础和应用实践四大部分。其中，数据部分，应该掌握数据的定义和类型以及数据的 DIKW 模型。大数据部分，应该掌握大数据的定义以及大数据相关的重要定理与模型。从数据到大数据，应当从宏观的角度了解到人类社会所发生的数据化进程。在理论基础部分，应该了解数据科学的发展历程，掌握数据科学的重要概念以及数据科学的主要内容。应用实践部分，应该了解数据科学家的概念和技能，熟悉数据科学的工作流程，了解真实世界的数据科学实践案例。在理论基础和应用实践的学习过程中，应学会相互联系，实践永无止境，理论创新也永无止境。

1.4.1 本章总结

第一节介绍数据和大数据。数据是在一定背景下有意义的对现实世界事物定性或定量的记录。其形式多样，使用不同的分类方式可以得到如结构化数据，非结构化数据；观测数据，实验数据等不同的分类结果。随着人类数据化进程的发展，从刚开始的图画，符号，到后来产生文字，照片，视频，再到后来人类发明了互联网和电子计算机，数据的形式日益多样，数据的体量与日俱增。在当下这个大数据时代，海量数据中蕴藏着我们希望找到和获得的规律和价值，与此同时也对我们今日的数据收集、数据管理、数据处理、数据分析等一系列活动带来了挑战。大数据具有独特性质，从本质特征概括有 5V 模型，从数据管理的角度有 5R 模型，针对大数据在医疗领域的应用有 4P 模型，基于本质特征有 HACE 定理和与之关联的大数据处理框架。

第二节介绍数据科学的理论基础。数据科学的概念从第一次提出到现在只有四十多年。

作为一门新兴的学科，它在早期发展缓慢，近年来发展速度显著加快，逐渐成为了热门领域。这一现象是因为数据科学的发展与大数据的发展和第四范式概念的提出有着紧密的联系。数据科学通过一系列科学的流程，从现实世界的数据中抽取出信息和知识，最后使得数据的集成度更高，价值密度更大。它以现实世界中各种各样的数据为研究对象，以数据科学韦恩图为经典的理论体系，与第四范式共同构成了大数据科学研究的理论基础。

第三节介绍数据科学的应用实践。数据科学家是能够发现现实世界的问题，收集问题相关的数据，抽取数据中的信息，并解释数据背后规律意义的人。数据科学家需要具备实质性的专业知识，需要计算机技术等处理数据的基本技能；同时也需要统计学和数据方面的理论知识以便对数据做出正确的分析处理；最后，数据科学家需要与人合作，有良好的沟通技能。数据科学家所做的数据科学工作基本分为：提出问题，收集数据，处理数据，清洗数据，分析数据，数据可视化，构建数据产品这几个关键步骤。通过数据科学流程得出的数据分析结果能够使人们理解数据背后的信息和知识，从而对现实世界的一些重要决策起到科学指导的作用。最后一节通过介绍医疗大数据治理，沃尔玛对社交大数据的应用，以及智慧城市的概念，使同学们对数据科学的应用实践有一个初步的了解。

1.4.2 扩展阅读材料

1. 维克托·迈尔·舍恩伯格. 大数据时代：生活、工作与思维的大变革[M]. 周涛, 译. 杭州：浙江人民出版社, 2012.
2. 赵国栋，易欢欢，糜万军等.大数据时代的历史机遇——产业变革与数据科学[M].北京：清华大学出版社,2013.
3. 朱扬勇，熊赟著.数据学[M].上海：复旦大学出版社,2009.
4. Rachel Schutt, Cathy O'Neil. 数据科学实战[M]. 冯凌秉，王群锋,译.北京：人民邮电出版社,2015.
5. Joel Grus.数据科学入门[M].高蓉，韩波，译.北京：人民邮电出版社,2016.
6. 朝乐门.数据科学[M].北京：清华大学出版社,2016.
7. 阿尔贝托·博斯凯蒂，卢卡·马萨罗. 数据科学导论 : Python 语言实现[M].于俊伟,靳小波,译.北京：机械工业出版社,2016.
8. 吴喜之.统计学：从数据到结论[M].北京：中国统计出版社,2009.
9. Zacharias Voulgaris. 数据科学家修炼之道.吴文磊，田原,译.北京：人民邮电出版社,2016.
10. Tony Hey, Stewart Tansley, Kristin Tolle.第四范式：数据密集型科学发现[M].潘教峰，张晓林,译.北京：科学出版社，2012.

1.5 习题

1. 查阅文献并思考，大数据的价值可以体现在哪些方面？
2. 查阅文献并思考，数据科学与统计学有何不同？

3. 查阅文献并思考，数据科学家和数据分析师有什么不同？
4. 查阅文献并思考，数据科学有哪些基本原则？
5. 查阅文献并思考，数据科学与数据密集型科学有什么不同？
6. 查找资料并思考，数据科学家需要具备哪些技能？
7. 查找资料，举出一个数据科学的实践案例。

1.6 参考资料

[1] 吴喜之.统计学：从数据到结论[M].4 版.北京：中国统计出版社，2013.

[2] 王珊，萨师煊．数据库系统概论[M].5 版．北京：高等教育出版社，2014.

[3] Thomas Stearns Eliot. The Rock[M]. London: Faber & Faber. 1934.

[4] Milan Zeleny. Management Support Systems: Towards Integrated Knowledge Management[J]. Human Systems Management, 1987, 7(1): 59-70.

[5] Russell Ackoff. From Data to Wisdom. Journal of Applied Systems Analysis[J]. 1989, 16: 3-9.

[6] 中国互联网络信息中心.第 45 次中国互联网络发展状况统计报告. http://www.cac.gov.cn/rootimages/uploadimg/1589535470391296/1589535470391296.pdf?filepath=ZBWvETi1XzcBKtOIkqelkI6LjXtsGgSA5nY19tMgqpxXM3yR3AbrYaldQRZRixKUS2HIKQQ6RgOUhhY5QbgarDEjTcORl73kmCTdYOvdSjs=&fText=全文%20第45次《中国互联网络发展状况统计报告》2020425

[7] 梅宏. 推进大数据应用 繁荣数字经济发展[N]. 中国信息化周报. 2018，7.

[8] John Mashey. Big Data and the Next Wave of InfraStress Problems, Solutions, Opportunities. https://www.usenix.org/legacy/publications/library/proceedings/usenix99/invited_talks/mashey.pdf.

[9] 维克托·迈尔·舍恩伯格. 大数据时代：生活、工作与思维的大变革[M]. 周涛, 译. 杭州：浙江人民出版社. 2012.

[10] McKinsey Digital. Big data: The next frontier for innovation, competition, and productivity. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

[11] Yuri Demchenko, Paola Grosso, Cees de Laat, et al. Addressing big data issues in Scientific Data Infrastructure. 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, 2013: 48-55.

[12] Merritte Stidston. Business leaders need R's not V's: the 5 R's of big data. https://mapr.com/blog/business-leaders-need-rs-not-vs-5-rs-big-data/.

[13] 吴信东，何进，陆汝钤等. 从大数据到大知识：HACE + BigKE[J].自动化学报. 2016，42(7)：965-982.

[14] Peter Naur. Concise Survey of Computer Methods. http://www.naur.com/Conc.Surv.html.

[15] William S. Cleveland. Data Science: An Action Plan for Expanding the Technical Areas of

the Field of Statistics[J]. International Statistical Review, 2001, 69(1): 21-26.
[16] 朱扬勇,熊赟.数据学[M].上海：复旦大学出版社，2009.
[17] Nathan Yau. Rise of the Data Scientist. http://flowingdata.com/2009/06/04/rise-of-the-data-scientist/.
[18] Drew Conway. THE DATA SCIENCE VENN DIAGRAM. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.
[19] Tom Davenport, D.J. Patil. Data Scientist: The Sexiest Job of the 21st Century. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.
[20] 李国杰.对大数据的再认识[J].大数据,2015,1(01):8-16.
[21] 欧高炎,朱占星,董彬等.数据科学导引[M].北京：高等教育出版社. 2017.
[22] 李国杰.大数据研究的科学价值[J].中国计算机学会通讯.2012,8(9): 8-15.
[23] THE DATA SCIENCE VENN DIAGRAM. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.
[24] The Data Science Puzzle. https://www.kdnuggets.com/2016/03/data-science-puzzle-explained.html.
[25] Tony Hey, Stewart Tansley, Kristin Tolle. 第四范式：数据密集型科学发现[M].潘教峰,张晓林,译.北京：科学出版社,2012.
[26] National Science Board. Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century. https://www.nsf.gov/geo/geo-data-policies/nsb-0540-1.pdf.
[27] The (Not So) New Data Scientist Venn Diagram. https://www.kdnuggets.com/2016/09/new-data-science-venn-diagram.html.
[28] Rachel Schutt, Cathy O’Neil.数据科学实战[M].冯凌秉,王群锋,译.北京：人民邮电出版社. 2015.