第1章
引　论

1.1　序言

谷歌的首席经济学家哈尔·瓦里安说过一番话："与数据相关的能力——包括获取数据、理解数据、处理数据、从数据中提取价值、用可视化方式展现数据、交流数据——将成为未来数十年间至关重要的一项能力。不仅专业技术人员应该掌握它，即使是在小学、中学和大学，也都应该传授相关的技巧。因为，我们已经进入了大数据时代，数据无处不在，无孔不入。"

这也是本书编著及修订的目的所在。"数据科学"是一门包罗万象的学科，涉及很多方面的内容，涵盖数学、统计学、计算机科学、人工智能、模式识别、分布式计算、图形学等多个领域的技术和理论。研究数据科学，正是为了让人们拥有和数据打交道的能力，运用这些技术和工具，帮助人们应对学习、工作和生活，在大数据时代如鱼得水。

1.2　数据科学简述

1.2.1　数据科学的定义

"数据科学"这个词从最早提出到现在已经有超过 50 年的历史了。要理解什么是数据科学，首先要明白什么是科学。

图 1－1 对现有的科学体系做了一个总结，可以看到，科学研究的范围小到粒子，大到整个可见宇宙，囊括了客观世界的方方面面，还涉及主观世界中人的逻辑思维、社会行为等。那究竟什么是科学？

按照达尔文的说法，"科学就是整理事实，从中发现规律，得出结论"。而在我国的《辞海》中对科学有如下定义，"科学是关于自然界、社会和思维的知识体系，它是为适应人们生产斗争和阶级斗争的需要而产生和发展的，它是人们实践经验的结晶"。维基百科如是说，"科学是一项系统性工程，它以各种可验证、可测试的关于宇宙万物的解释和预测的形式，来创造、构建和组织知识体系"。

可见，科学就是一项系统性的，通过不断的探索和尝试，去获取知识、了解世界的工程。

由此，可以引出数据科学的定义：

数据科学就是一门通过系统性研究来获取与数据相关的知识体系的科学。这里有两个层面的含义：

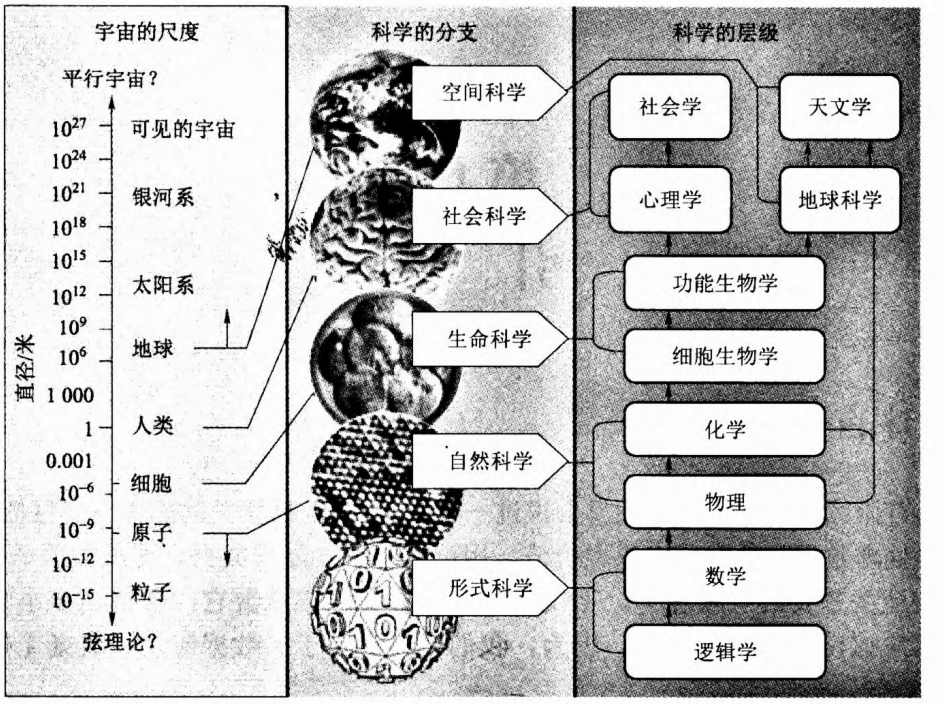

图 1-1　科学体系

1）研究数据本身，研究数据的各种类型、结构、状态、属性及变化形式和变化规律；

2）通过对数据的研究，为自然科学和社会科学的研究提供一种新的方法——称为科学研究的数据方法，其目的在于揭示自然界和人类行为的现象和规律。

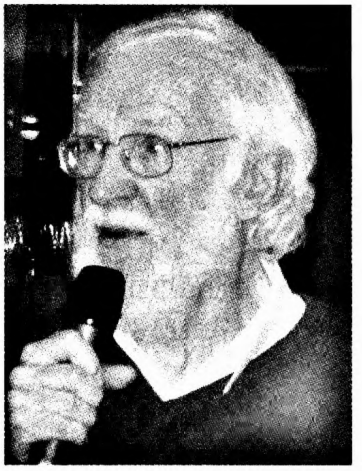

图 1-2　彼得·诺尔（前图灵奖得主，丹麦人）

1.2.2　数据科学的由来

“数据科学”这个词最早出现在 1960 年，是由丹麦人，前图灵奖得主，计算机科学领域的先驱彼得·诺尔（图 1-2）所提出的。最初，彼得·诺尔打算用它来代称计算机科学。

1974 年时，彼得·诺尔出版了 *Concise Survey of Computer Methods* 一书，对当时的数据处理方法进行了广泛的调研，在书中他多次提到了“数据科学”的字眼。

1997 年，国际知名的统计学家吴建福（图 1-3）在美国密西根大学做了名为“统计学是否等同于数据科学”的讲座。他把统计学归结为由数据收集、数据建模和分析、数据决策所组成的三部曲，并认为应将“统计学”重命名为“数据科学”。

2002 年，*Data Science Journal* 创稿。2003 年，*The Journal of Data Science* 创稿。

随着大数据时代的来临，“数据科学”这门学科在近些年来受到了越来越多的关注。

1.2.3 数据科学的研究范畴

虽然有着近 60 年的历史，但数据科学仍可算是一门新兴的学科。它涉及的范围非常广泛（图 1－4），主要涵盖以下几个方面：

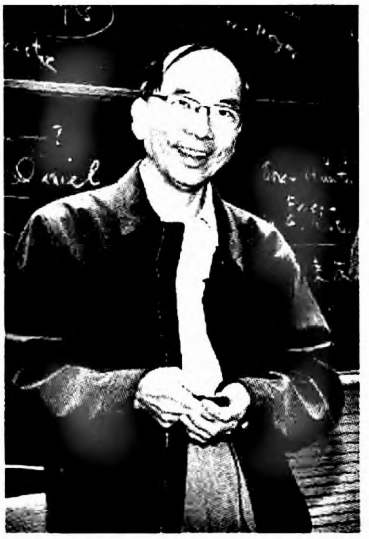

图 1－3　吴建福（国际知名统计学家，美国国家工程院院士）

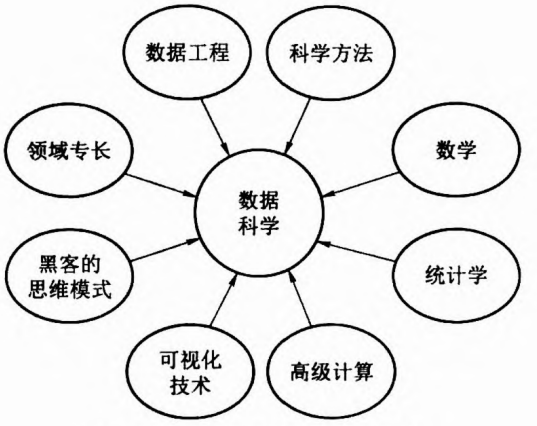

图 1－4　数据科学的研究范畴

1）数据与统计学相关知识，包括数据模型、数据过滤、数据统计和分析、数据结构优化等；

2）计算机科学的相关知识，包括数据的获取技术、数据的处理方法、数据的存储和安全性保障等；

3）图形学的相关知识，包括数据的可视化、数据的协同仿真、虚拟环境的实现等；

4）人工智能的相关知识，包括机器学习算法的应用、神经网络的运用等；

5）领域相关知识，包括处理特定领域的数据分析和解读时需要用到的理论和方法等。

除了上面这些已知的领域外，数据科学在未来还会深入许多目前未知的领域。虽然已有近 60 年的历史，但对数据科学的探索只能算是刚开始，还有许多未知的领域亟待进行探索。

从图 1－5 中可以看到，数据科学的研究对象是数据本身，通过研究数据来获取对自然、生命和行为的认识，进而获得信息和知识。数据科学的研究对象、研究目的和研究方法等，都与已有的计算机科学、信息科学和知识科学有着本质的不同。自然科学研究自然现象和规律，认识的对象是整个自然界，即自然界物质的各种类型、状态、属性及运动形式。行为科学是研究自然和社会环境中人的行为以及低级动物行为的科学，包括心理学、社会学、社会人类学等。数据科学支持了自然科学和行为科学的研究工作。而随着数据科学的发展，越来越多的科学研究工作将会直接针对数据进行，这将使人类更好地认识数据，从而更加深刻地认识自然和社会。

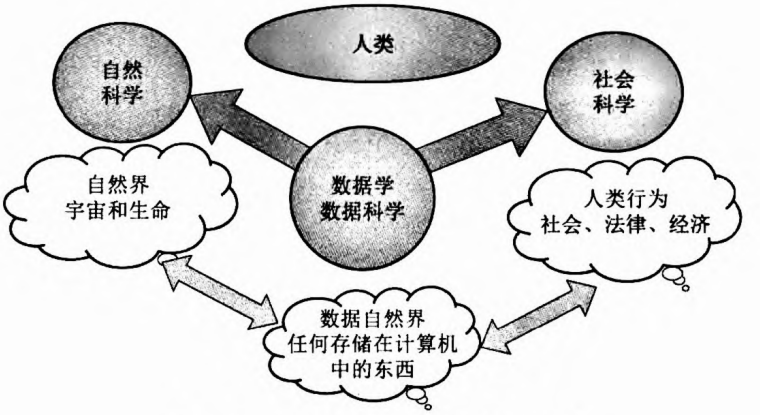

图 1-5　数据科学与其他学科的关系

归结起来看，数据科学的具体研究内容可以分为以下四个方面：

1）基础理论研究（科学）。

基础理论研究的对象是：数据的观察方法和数据推理的理论，包括数据的存在性、数据测度、数据代数、数据相似性与簇论、数据分类与数据百科全书等。

2）实验和逻辑推理方法研究（工程）。

要想做好实验和逻辑推理方法研究（工程），需要建立数据科学的实验方法，建立许多科学假说和理论体系，并通过这些实验方法和理论体系来开展对数据的探索研究，从而认识数据的各种类型、状态、属性及其变化形式和变化规律，揭示自然界和人类行为的现象和规律。

3）数据资源的开发利用方法和技术研究（技术）。

数据资源的开发利用方法和技术研究（技术）主要是指研究数据挖掘、清洗、存储、处理、分析、建模、可视化、展现等一系列过程中所遇到的各种技术难题和挑战。

4）领域数据科学研究（应用）

领域数据科学研究（应用）主要是指将数据科学的理论和方法应用于各种领域，从而形成针对专门领域的数据科学，例如，脑数据科学、行为数据科学、生物数据科学、气象数据科学、金融数据科学、地理数据科学等。

1.2.4　数据科学的学习意义

讲授数据科学这门课程的意义，就在于培养数据科学家或者数据工程师。科学的从业者被称为科学家，顾名思义，数据科学家就是数据科学的从业者。而数据工程师，就是可以熟练运用数据科学的工程人员。

根据权威咨询结构麦肯锡的报告，仅是美国，未来就需要超过 19 万个深度解析数据科学家和 150 万个初级数据管理人员，而且目前这些职位的空缺值非常大。

谷歌的首席经济学家哈尔·瓦里安曾说过：“数据科学家将会是 21 世纪最性感的工作。”当然，他本人就是目前世界上最顶尖的数据科学家，如图 1-6 所示。

1.3　本书结构

随着大数据时代的来临，数据在人们的生活中开始占据越来越大的比重。无论是在工作、

学习还是生活的过程中，人们在做决定时，都会越来越多地依赖于对数据的分析。

数据科学，作为一门正在蓬勃发展的新学科，所关注的正是如何在大数据时代背景下，运用各门与数据相关的技术和理论，服务于社会，让人们可以更好地利用身边的数据，将生活变得更加美好。数据科学已经渗入了人们生活和工作的方方面面，无论是政府还是企业，未来都需要大量的懂得数据科学相关知识的人才。

图 1-6　数据科学家——未来最性感的职业

本书将会系统地讲述与数据科学相关的各个方面的知识，着重培养数据科学所需要的各项技能与思维方法。本书作为数据类课程的基础教材，将会结合大量翔实的案例来讲解相关知识要点。一方面，大量案例使得本书更为生动具体，从各个方面阐述了数据科学知识的运用方式，让学生更容易接受。另一方面，从这些案例中，编者也提取出来数据科学各个领域的特色，寓技术于案例之中，在讲解案例的同时，把技术细节也传授给了学生。并且将数据科学学科中的各种新兴的思维模式和技术方法的介绍穿插到案例分析的过程中。结合案例来讲述相关知识和技术，可以让学生更直观、更深刻地了解各个知识点。

本书主要包含五个部分，依次回答了与数据科学相关的五个基本问题：

第一部分，即引论部分，主要是讲解数据科学的具体概念和研究范畴。

第二部分，主要是介绍大数据及其产生根源。将从数据的定义谈起，结合数据的演化，阐述大数据的概念和特征，并探究大数据的产生根源。

第三部分，主要是讲解大数据研究的重要性。将会从世界范围内大数据的研究现状谈起，结合丰富翔实的案例分析，穿插一些技术知识点的讲解，形象生动地引导学生初步了解大数据的研究方法、内容，以及理解数据科学研究的重要性。

第四部分，主要是讲解数据科学的研究方法。结合数据获取、存储、管理、处理、分析、建模及可视化的全过程，系统地、有序地讲解数据科学的研究方法，注重与工程实践相结合。

第五部分，主要是探索数据科学与生活的关系。从智慧城市的建设、智慧医疗的发展、未来就业、生活等各个方面讲述数据科学对人类社会和日常生活的深远影响。

第二部分　大数据及其产生根源