系统与决策丛书

基于数据科学的复杂元网络方法及应用

刘　潇　杨建梅/著

科学出版社

系统与决策丛书

基于数据科学的复杂元网络方法及应用

刘　潇　杨建梅　著

国家自然科学基金部分资助

科学出版社

北京

内 容 简 介

复杂性是系统研究的前沿，复杂网络是复杂系统结构的模型，元网络则刻画了社会网络的复杂性。本书基于元网络与复杂网络，提出了复杂元网络的概念与分析方法。由于具有大规模性与多样性，复杂元网络与大数据科学有着天然的联系，所以按照知识论域体系，本书属于方法、技术与应用层面的专著，其特色具体体现在复杂网络与元网络糅合的建模方法、大数据处理的技术以及在国际贸易研究中的应用上。

本书可为系统工程、管理学、统计学、计算机科学和信息管理等相关专业的学者、研究生提供有价值的研究借鉴，也适合于对社会经济网络、复杂网络及管理复杂性研究感兴趣的人士阅读和参考。

图书在版编目（CIP）数据

基于数据科学的复杂元网络方法及应用/刘潇，杨建梅著. —北京：科学出版社，2015

（系统与决策丛书）

ISBN 978-7-03-043420-3

Ⅰ.①基… Ⅱ.①刘…②杨… Ⅲ.①数据管理-研究 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2015）第 031420 号

责任编辑：王京苏/责任校对：胡小洁

责任印制：李 利/封面设计：蓝正设计

科学出版社出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

三河市骏杰印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2015 年 3 月第 一 版 开本：720×1000 1/16

2015 年 3 月第一次印刷 印张：12 1/2

字数：250 000

定价：60.00 元

（如有印装质量问题，我社负责调换）

序

该书属于“系统与决策丛书”系列，截至目前，该丛书已出版了《软件过程改进的复杂性工作程序研究》《企业知识管理方法研究：利益协调软系统方法论的应用》以及《基于行为的公司资本投资决策方法研究》三本专著。在先前丛书序中，基于沃菲尔德的科学论域体系，作者提出了“哲学信仰—理论—方法论—方法—应用”的知识论域体系，并陆续阐述了上述专著在各自知识论域体系链条中的位置以及与丛书主题的内在联系。

不过在近几年，团队的研究重点逐步转向了复杂网络及人类行为动力学领域，所以原丛书序中所列的《企业战略管理的系统方法论》《利益协调软系统方法论》等方法论与方法层面的专著就暂时被搁置出版了，取而代之的是管理复杂网络等方面的专著。出版计划的改变既有遗憾，也为我们带来了追随学科发展探索新领域的乐趣。

系统科学揭示了1加1大于2这个系统现象，但是没有回答1加1大于2究竟是如何产生的这个令人着迷的问题。复杂性科学则以回答这个问题为宗旨，因此复杂性科学是系统科学的新发展。同时，复杂网络与人类行为动力学又是兴起不久的复杂性科学的重要分支。因此可以说，新的出版计划仍然围绕着丛书的系统与决策这个主题。

复杂网络是物理学家创建的复杂系统的结构模型，元网络是社会学家奠基的分析组织多重复杂关系的建模方法，两者虽然起源不同，但却有着共同的研究对象。该书结合复杂网络与元网络，提出了复杂元网络的概念与分析技术；同时由于复杂元网络具有大规模性与多样性，其必然与大数据科学有着天然的联系。所以，按照知识论域体系，与上述已出版的偏向方法论的专著相比，《基于数据科学的复杂元网络方法及应用》则更多地偏向于方法与应用，其特色具体体现在复杂网络与元网络相糅合的建模方法、大数据处理的技术以及国际贸易研究等方面。

杨建梅

2015年1月

前　言

数据科学是关于数据的科学，是研究探索网络空间中数据界奥秘的理论、方法和技术、是在数据爆炸背景下涌现出来的处理大数据的跨学科领域。数据科学理论基础和工具来自多个不同的学科领域，包括计算机科学、统计学、数据库理论、人工智能、机器学习及社会科学等。数据科学的重点是研究联系大数据的关系网络，因此对大数据所形成的复杂数据网络的特性与功能进行研究的复杂网络分析是数据科学的重要基石。

近十几年来，复杂网络作为复杂系统建模工具在社会、经济、生物和技术网络中取得大量实证研究成果和理论进展，然而经典的社会网络或复杂网络方法主要用于研究单模至多双模的静态网络，面临着如何处理多模式和多重关系网络数据及如何分析动态网络或纵向网络数据的挑战。

美国卡内基梅隆大学的 Carley 教授提出采用元网络（meta network）方法建模多模式和多重关系网络；其领导的社会和组织系统计算分析中心将社会网络分析、链接分析和多智能体仿真等技术结合起来开发元网络分析软件 ORA，除提供传统社会网络分析功能，还具备矩阵运算、多重关系网络指标计算、纵向网络统计分析、网络结构优化、元网络可视化和多智能体仿真等功能。

本书在数据科学背景下，从网络方法的概念和理论出发，以 Carley 的元网络方法为基础，提出复杂元网络概念及跨学科方法论框架，具体包括网络数据采集和管理、复杂多重关系网络建模和分析、动态网络分析及网络数据挖掘等方法和工具，同时结合国际贸易实例介绍上述方法和工具的具体实现和应用，可为社会经济系统复杂性研究提供方法借鉴和参考。

第 1 章分析数据世界及大数据形成背景，对数据科学涌现及其学科体系形成进行探索和诠释，尝试性地定义数据科学的“理论、方法论、方法和应用”学科框架，并从技术演化视角，介绍大数据存储和处理及大数据分析的关键技术。

第 2 章从发生学视角介绍图论方法、社会网络方法和复杂网络方法的起源和演化，然后介绍元网络方法及工具；之后从知识论域体系视角比较了社会网络、复杂网络和元网络三种方法范式在“哲学、理论、方法论、方法和应用”等方面的异同；指出随着大数据所形成的复杂数据网络分析需求增长，网络方法必将与数据科学糅合，成为系统复杂性研究的重要工具。

第 3 章首先从复杂系统建模视角，将元网络拓展为复杂元网络概念。其次结合国际贸易网络研究问题和理论，将复杂元网络应用于国际贸易系统建模，提出国际

贸易复杂元网络模型及动力学分析框架。最后结合应用实例及其意义，总结了复杂元网络的知识论域体系的“哲学”为整体论和诠释学；“理论”为复杂性理论与具体学科理论的融合；“方法”是图论方法与多智能体仿真，以及数据科学相关方法的整合，可“应用”于复杂社会经济系统多种复杂多重关系问题研究。

第 4 章针对复杂元网络应用面临的挑战，基于数据科学相关方法技术提出复杂元网络方法论框架：集成数据库、数据仓库、联机分析处理（on-line analysis processing，OLAP）、元网络分析、动态网络分析、数据挖掘等多样化技术工具；对网络表达语言、网络建模技术等进行专题研究；基于商业智能平台介绍动态网络数据库、复杂元网络数据仓库和复杂元网络计算机建模的具体技术细节。

第 5 章介绍复杂元网络分析方法。指出复杂元网络分析既要重视静态分析也要重视动态分析；既要有个体层面、群组层面的分析也需要进行网络层面和元网络整体层面的分析。因此，需要多样化的分析视角和分析工具。本章重点介绍元网络分析方法、动态网络改变诊断、网络社团分析和网络数据挖掘方法等。

第 6 章介绍复杂元网络分析工具，属于技术层面内容。重点介绍 ORA 软件在元网络分析及组织系统分析中的主要工具和操作细节，以及开源软件 R 中网络分析和数据挖掘的相关工具和实现代码。

第 7 章在上述模型、方法论及具体建模和分析技术的基础上，研究国际贸易问题。内容包括出口贸易复杂动态元网络模型建模和网络结构演化分析，以及企业行为模式挖掘、出口市场网络结构模式演化分析等。基于复杂元网络方法的研究提供了出口贸易系统结构特征方面和要素关系及企业行为方面的丰富信息，是目前管理部门采用传统经济统计方法不能提供和解释的。

不同学科与领域背景的读者对本书会有不同的阅读角度。读者如果对国际贸易网络感兴趣，可以选读第 3、4 章的部分章节和第 7 章。如果对元网络方法有兴趣，可选读第 2、5、6 章的部分章节。如果对 R 语言中网络分析和数据挖掘工具感兴趣，可选读第 5～7 章的部分章节和附录。

本书是国内首部尝试介绍元网络方法及拓展和应用的书籍，作者希望能在数据科学背景下，对网络方法创新及分析工具实践等方面做些开创性工作，但由于网络方法领域发展很快，而数据科学领域又太新，涉及学科知识太多，也罕有系统全面的学术参考，所以尽管付出了许多努力，但毕竟学识有限，最后完成的工作难免存在疏漏之处，恳请广大读者批评指正。本书设置了网上交流园地，并提供书中实例研究数据及分析代码下载，我们热诚期待与各位读者进行交流。

本书最终能顺利出版，得到了科学出版社的大力支持，在此表示感谢！

刘　潇　杨建梅

2014 年 12 月于广州

目　录

第 1 章　数据科学概述 …… 1
1.1　数据世界演化及特征 …… 2
1.2　数据科学的概念 …… 4
1.3　数据科学的学科体系 …… 6
1.4　大数据关键技术 …… 12
1.5　本章小结 …… 16
第 2 章　网络方法概述 …… 17
2.1　网络方法起源与演化 …… 17
2.2　元网络方法 …… 24
2.3　网络方法比较 …… 28
2.4　本章小结 …… 31
第 3 章　复杂元网络模型 …… 32
3.1　复杂元网络模型概述 …… 32
3.2　复杂元网络模型实例 …… 34
3.3　复杂元网络应用面临的挑战 …… 44
3.4　复杂元网络知识论域体系 …… 45
3.5　本章小结 …… 46
第 4 章　复杂元网络方法论 …… 48
4.1　复杂元网络方法论框架 …… 49
4.2　网络表达方法研究 …… 55
4.3　网络建模技术研究 …… 61
4.4　复杂元网络建模实例 …… 63
4.5　本章小结 …… 70
第 5 章　复杂元网络分析方法 …… 72
5.1　静态元网络分析方法 …… 72
5.2　网络社团分析算法 …… 80
5.3　网络改变诊断方法 …… 83
5.4　网络数据挖掘方法 …… 87
5.5　本章小结 …… 103

第 6 章 复杂元网络分析工具 …… 105
6.1 元网络分析工具 ORA …… 105
6.2 R 语言与网络分析 …… 118
6.3 R 语言与数据挖掘 …… 122
6.4 本章小结 …… 127
第 7 章 复杂元网络实例建模与分析 …… 128
7.1 实例背景 …… 128
7.2 数据与模型 …… 129
7.3 网络结构演化分析 …… 130
7.4 仿真与情景分析 …… 145
7.5 agent 行为模式挖掘 …… 147
7.6 关联结构挖掘和社团分析 …… 161
7.7 本章小结 …… 170
参考文献 …… 171
附录 实例分析 R 代码 …… 183
后记 …… 188

第 1 章　数据科学概述

据百科辞典解释，数据是对客观事物的符号表示，是用于表示客观事物未经加工的原始素材，如图形符号、数字、字符等。或者说，数据是通过物理观察得来的事实和概念，是关于现实世界中的对象或概念的描述。

20 世纪中期计算机科学诞生，大量数据被快速生产并以二进制数位的形式存储在计算机系统中，人类社会的数据量剧增，逐步形成一个有别于真实自然界或人类社会的“数据世界”。步入 21 世纪，在互联网络和 Web 技术的推动下，电子商务、移动服务、云计算和传感器的普及，以及科学实验和计算机仿真的应用，大量从宏观到微观、从自然到社会、从科学研究到个人活动，数据不受时间和地点限制源源不断产生；数据的种类和规模以前所未有的速度增长和累积，标志着人类社会在不知不觉中步入大数据时代。

近年来以探讨如何更好地利用数据来产生良好社会效率的“大数据研究和发展倡议”[1]，以及如何在科学研究、环境、生物医学领域利用大数据进行突破的“大数据计划”[2]等，引起产业界、学术界和政府机构的密切关注。在学术界，*Nature* 杂志曾在 2008 年刊登“Big Data”专刊[3]，阐述了在数据驱动的背景下，解决大数据问题所需要的技术以及面临的挑战；*Science* 杂志则在 2011 年刊登“Dealing with Data”专刊[4]，围绕科学研究中大数据的问题展开讨论，说明大数据对于科学研究的重要性。

大数据的规模效应导致数据采集、管理、分析和利用的复杂性，并最终导致传统的数据管理方式、数据处理方式、数据思维的颠覆式的改变，探测数据的科学和技术变得越来越重要。目前，一门探测数据世界中数据的奥秘和规律的新兴学科——数据科学正涌现出来[5]；一群精通数据和数据处理技术、将数据价值从繁冗的数据中抽离出来的数据研究人员——数据科学家（data scientist）走进人们的视野[6]；一种基于密集型数据的知识发现的科学研究方式——“第四范式”一经发布就引发革命性轰动效应[7]；一个以开发和利用数据资源、生成和制造数据产品的产业已初具规模。

本章首先回顾数据世界形成的历史及演化过程，其次结合相关历史事件介绍数据科学的形成及学术界对数据科学体系的探索和诠释，再次借鉴知识论域框架尝试性定义了数据科学的“理论、方法论、方法和应用”，最后综述了大数据的关键技术和工具。

1.1 数据世界演化及特征

1.1.1 数据世界的形成与演化

回顾社会发展历史，人类最初是通过大脑来记忆对现实世界的感知，这是最初的数据化。但是由于人脑记忆的有限性和不完全可靠性，人类开始寻求各种辅助设施来帮助记忆。例如，在树桩或龟壳上刻录图形和符号等，这种方式不仅创造了文字，也实现了对自然界各种事物的记录和传播。

造纸和印刷术的发明带来了人类历史上的第一次“数据爆炸”[8]。纸质书的发明使得记载真实世界的符号、图形等数据能长期保存并广泛传播。这期间的作者和出版商是主要的数据生产者，书籍和图书馆成为数据存储和传播的媒介和场所。20 世纪初，音频和视频等多媒体设备的发明和使用成为数据存储和传播的新载体，如录音带和录音机，缩微胶片和投影仪等。至此，关于人类能感知的任何东西，都能以数字、字符、声音、图像和照片等形式被记录、存储和传播。

20 世纪中叶，计算机及存储设备的发明带来了第二次“数据爆炸”，人类将原来存储在不同载体上的数据数字化，以二进制数位形式存储在计算机系统中。数据处理技术从最初的文件处理系统发展到基于数据库技术的运营系统，如航空售票系统、银行交易记录系统、超级市场销售记录系统等，数据伴随着运营活动产生并记录和存储在计算机系统中。这阶段，数据生成本质上是按定制格式录入计算机的“被动式”生产方式（表 1-1）。随着各种计算机系统的运作，区别于真实世界的数据世界开始形成，且规模随着时间推移不断增长扩大。

真正的“数据爆炸”源于因特网和万维网技术的发展，特别是进入 Web2.0 时代，以博客、微博为代表的各种新型社交网络的出现，3G 网络和 Wi-Fi 等网络基础设施的普及，以及以智能手机、平板电脑为代表的新型移动设备的广泛使用，激发用户主动创建和传播数据的意愿，数据产生方式转变为以用户原创内容（user generated content，UGC）为标志的[9]“主动式”生成，最终导致数据呈爆炸式的增长。

数据密集型的科研方式和人类对感知式系统的广泛使用，进一步将人类社会带入到“大数据”时代。一方面，计算机模拟或仿真产生海量实验数据，如人类基因组数据库的建立；另一方面，真实世界各领域数据被广泛布置于各角落的设备自动采集，如地图数据、海洋数据、天气和气象数据等，各种来源和不同类型的数据自动地、源源不断地流入并存储在网络服务器、政府部门数据库、企业数据库、个人电脑和便携式设备中。

表 1-1　数据世界演化历史

阶段	特征
人工管理阶段（前计算机时代）	数据主要来源于人类脑力活动（对数据进行采集、编码或转换）；数据产生方式以手工记录、辅助设备记录为主；形成书籍、音频、视频和图片资料等不同载体形式，数据规模小
计算机管理时代（20 世纪 50 年代起）	数据主要来源于各种运营系统（金融、零售业、制造业等）；数据产生方式以被动式为主（按定制格式录入）；采用文件系统和数据库技术，数据是结构化的，数据规模 MB 或 GB
互联网与 Web 时代（20 世纪 90 年代末起）	数据主要来源于内容网站、新兴社交网络、开源社区、电子商务、Web 服务器日志；数据产生方式以主动式为主（用户自定义）；采用超文本技术 XML、Web 数据库等，数据是半结构化或非结构化的，数据规模 GB 或 TB
数据密集型科学研究及感知系统时代（21 世纪起）	数据主要来源于科学实验、计算机仿真、传感器等自动采集系统；数据产生方式为自动式（仿真/数据流）；形成数据高速产出、数据规模 TB 或 PB 甚至更大

注：MB、GB、TB、PB 等为计算机存储单位。1 比特（bit，计算机二进制位，取值为 0 或 1）；1 字节（B）=8 比特；$1KB=2^{10}B$；$1MB=2^{20}B$；$1GB=2^{30}B$；$1TB=2^{40}B$；$1PB=2^{50}B$；$1EB=2^{60}B$；$1ZB=2^{70}B$

1.1.2　数据世界的特征

数据世界在技术演化的推动下不断发生质的飞跃和改变，并逐渐形成并表现出数据量巨大、数据类型多样、数据复杂但富含价值等典型特征。国际数据公司（International Data Corporation，IDC）给出了大数据的 4V 定义[10]，即海量的数据规模（volume）、快速的数据流转和动态的数据体系（velocity）、多样的数据类型（variety）和巨大的数据价值（value）。

数据规模性是指数据存储或处理的规模从原来的 MB 发展到 GB，甚至 TB 或 PB；数据管理从数据库（database，DB）到大数据库（very large database，VLDB），再到超大规模数据库（extremely large database，XLDB）的管理。但是多大规模的数据才能称为“大数据”呢？根据麦肯锡公司的观点，“大数据”不是按具体的 TB 值来衡量的，而是其大小超出了常规数据库的获取、存储、管理和分析能力[11]；亚马逊公司的大数据科学家 John Rauser 也指出，大数据是任何超过了一台计算机处理能力的数据量[12]。因此，数据的规模效应不仅带来数据管理和处理的困难，而且在数据的真实性、一致性、安全性等方面也面临难于控制等问题。

数据多样性是指数据来源的多样性。例如，数据来源包括个人数据、企业数据、政府数据、公共数据、地理数据（如 GPS 数据）、生命数据（如 DNA 序

列）、经济数据（如股票数据）、文化和社会数据（如新闻出版）、空间数据、海洋数据、科学数据等各领域。这些数据记录了人的行为、企业的业务活动、科学实验和社会经济的发展等。不同来源的数据在产生方式和产生频率、数据存储和表达语言、数据类型和格式等方面也呈现多样性特征。

数据价值性是指在海量数据中隐含的大量的丰富信息，探索和挖掘数据背后的规律或模式成为理解自然、生物、社会和技术系统复杂性的重要依据。大数据给科学研究和产业带来挑战的同时，也带来无穷机遇。

数据复杂性也是海量数据表现出来的重要特征之一。在文件管理和数据库管理阶段，先有模式才产生数据，数据世界的数据多是结构化的数据；但进入Web2.0和大数据时代，数据从结构转变为无结构，如网页（超链接文本）、Web日志和电子邮件等，以及音频、视频和图片等多媒体数据。数据复杂性对创新的数据存储技术、数据管理技术和数据分析工具产生需求。

1.2 数据科学的概念

在科学领域，当数据不再是科学研究的成果，而是变成了科学研究的基础，该如何对其采集、管理和分析呢？在商业领域，当数据从简单的处理对象演变为重要资源，该如何对其管理、开发和利用呢？对这些问题的思考，是否会导致学术界产生以数据为研究对象的数据科学或数据学（datalogy）的探讨呢？

尽管目前对数据科学是否存在或者是否是一门独立学科的问题尚无统一界定，但越来越多的学者认识到：探索数据是人类认识和理解真实世界的有效方法，以数据为中心的科学既不同于以自然界为研究对象的自然科学，也不同于以人类为研究对象的社会科学，但它正成为推动自然、社会和人文科学发展的动力。

事实上，自20世纪60年代开始，学者们开始对与数据相关的科学产生兴趣，尝试着对“数据科学”的研究对象、研究内容、学科体系等方面进行探索和诠释。文献［13］研究了相关历史事件，作者对其进行了整理和扩充，见表1-2。

表1-2 数据科学相关的事件

年份	人物/组织	主要贡献
1966	Peter Naur	提出并定义数据学
1974	Peter Naur	提出并定义了数据科学
1996	IFCS东京会议	数据科学首次作为会议主题之一
2001	William S. Cleveland	提出数据科学：拓展统计技术领域的行动计划
2002	CODATA	*Data Science Journal* 发行

续表

年份	人物/组织	主要贡献
2003	www. jdsruc. org	*Journal of Data Science* 的发行
2005	美国国家科学委员会	给出了数据科学家的定义
2009	Nathan Yau	发表《数据科学家的崛起》论文
2009	Troy Sadkowsky	在 Linkedin 成立数据科学家组（datascientists. net）
2010	Mike Loukides	指出数据科学是生产制造数据产品
2010	Mason and Wiggins	数据科学分为数据获取、清洗、探索、建模和解释
2010	Drew Conway	提出数据科学文氏图，是多学科或技术的交叉
2011	Pete Warden	数据科学与传统科学家所做的事不同
2011	Harlan Harris	指出数据科学是数据科学家所做的事情
2011	DJ Patil	建立数据科学团队
2013	美国国家科学委员会	《2025 年的数学科学》报告

基于时间上的观察，大致可以划分为三个阶段。第一个阶段是“数据科学”概念的形成。早在 1966 年，丹麦计算机科学家、图灵奖的获得者 Peter Naur 就曾撰写 *The Science of Datalogy* 一文[14]，提出并定义“datalogy”是使用数据的科学（datalogy as the science of the nature and use of data），并建议用该术语替代计算机科学。之后 Peter Naur 在 1974 年出版的 *Concise Survey of Computer Methods*[15] 一书中，定义“数据科学”是处理数据的科学（the science of dealing with data）[15]。后来“数据科学”在 20 世纪 90 年代中期被分类社团联盟采用，在 IFCS 东京会议首次作为会议主题，即“data science，classification，and related methods”[16]。2001 年贝尔实验室的 Cleveland 发表论文[17]，提出将数据科学设立为一个新的学科，吸收“计算在数据方面取得的进展”作为统计学的延伸，并提出具体实施的六个行动计划。2002 年国际科技数据委员会（Committee on Data for Science and Technology，CODATA）采用了术语“数据科学”[18]，并发行官方杂志 *Data Science Journal*。2003 年 *The Journal of Data Science* 创刊（www. jds-online. com），标志着“数据科学”概念在学术界得到认可。

第二个阶段是“数据科学家”概念的形成。2005 年美国国家科学委员会发表了《数字数据收集万岁：促进 21 世纪的研究与教育》报告[19]，提出并将数据科学家定义为“信息与计算机科学家，数据库、软件工程师及程序员，学科专家”，是数字数据收集的关键人物。2009 年 Natahn Yau 在“数据科学家崛起”一文中提及数据科学家的职位头衔[20]；同年拥有“科学性程序员”头衔的 Sadkowsky 在 Linkedin 建立了数据科学家小组，对其 datasceintists. com 网站进行

辅佐[21]。这些举动被视为是数据科学走向职业化的标志。2011 年 Patil 建立数据科学团队，指出数据科学家就是使用数据和科学创造新的东西的群体[22]。

第三个阶段是关于“什么是数据科学”的讨论。Loukides 在其发表的 *What is Data Science* 一文中指出，数据科学是创造数据的科学[23]。Mason 和 Wiggins 指出数据科学粗略的步骤是数据获取、清洗、探索、建模和解释，是融合统计学、机器学、数学等学科的专业领域[24]。Conway 提出数据科学的文氏图，指出数据科学是数学、统计学、专业知识、黑客技能等的交叉领域[25]。Warden 指出数据科学“是有缺陷的但却有用的术语”，尚无被广泛接受的范围边界，也没有完整的定义，但肯定是与传统科学家所做的事情不同[26]。Harris 指出数据科学是数据科学家所做的事情[27]。

在这个阶段，除最早成立于 2002 年的权威机构 CODATA（www. codata. org）外，英国、澳大利亚等国家相继成立了各种数据科学研究中心或实验室[5]。我国学者也开始加入数据科学研究领域，2007 年中国科学院成立虚拟经济和数据科学研究中心。2008 年复旦大学成立了数据学和数据科学研究中心，该中心的朱杨勇教授等比较系统地对数据学或数据科学进行了探索性研究[28,29]，在对相关文献回顾和总结的基础上，尝试性地定义了数据科学的概念。

定义 1-1 数据科学，也称为数据学或数据的科学，是探测数据世界奥秘和规律的理论、方法和技术。

从定义可知，数据科学的研究对象是数据，而不是信息，也不是知识。信息是在自然界、人类社会及人类思维活动中存在和发生的现象；知识是人们在实践中所获得的认识和经验。尽管数据可以作为信息和知识的符号表示或载体，但数据本身并不是信息或知识。数据科学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，因此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和信息科学有着本质的不同；也不同于自然科学和社会科学领域的学科（如数学、物理学、统计学等），是人类社会发展至今，伴随数据世界诞生的新科学。

1.3 数据科学的学科体系

尽管数据科学的术语由来已久，但是作为新学科正处于萌芽阶段。数据科学学科体系如何构成，数据科学的研究范围是什么，包含哪些研究内容，与其他传统学科的关系等问题，引发了各领域学者的思考。

朱杨勇等认为，数据科学的学科体系是由数据学基础和数据学应用两部分构成[28]。其中，数据学基础包括了数据学基本假设和基础理论、数据获取、数据分析和数据感知等，具体可细分为数据管理（文件系统、数据库管理系统、数据

仓库等)、数据勘探、数据伪装、数据辨伪、数据整合、数据实验、数据挖掘、数据分类、数据可视化、数据可嗅化、数据可听化、数据可触化等技术门类。数据学应用指在数据学基础上针对某个研究领域开发出的专门领域的理论、技术和方法，所形成的专门领域的数据学。

此外有学者建议将数据科学粗略地划分为理论数据学、工程数据学和应用数据学三大部分。其中，理论数据学包括数学、哲学、管理学、社会学等自然科学与社会人文学科的内容；工程数据学包括资源、技术及其实践；应用数据学则体现了数据科学在不同行业领域及不同学科中的应用。

科学论域体系是管理复杂性学派的创始人 Warfield 提出的，他认为一门学科包括“基础—理论—方法论—应用”四个环节。其中“基础”是指科学的普遍前提在本学科具体化后形成的基本概念等要素，其作用是指导“理论”；而“理论”提供概念之间关系的定律等以指导“方法论”；“方法论”提供程序指导“应用”；“应用”反过来影响“基础”。杨建梅对 Warfield 的科学论域体系进行扩展[30]，她将“基础”合并进“理论”，认为哲学不属于科学，但是哲学信仰会决定知识工作者构建何种类型的理论，应该将其作为一层引入“理论”；而方法论为做事情的原则和程序，与具体的方法不同，所以应将方法从方法论中剥离出来，就此提出了“知识的论域体系”由“哲学信仰—理论—方法论—方法—应用”五个环节有机构成。

作者参照两种论域体系观点，采用“理论—方法论—方法和技术—应用”框架，尝试性地建构数据科学的体系架构（图 1-1）。

1.3.1　数据科学理论

数据科学“理论”包括基础理论和专业理论两大类别。基础理论粗略地划分为数学理论和非数学理论两个分支。数学理论包括计算数学和应用数学（如图论）；非数学理论包括计算机科学、统计学、机器学习、信息科学等，以及与数据全生命周期相关的理论，如数据测度、数据代数、数据分类、数据百科全书、数据实验学等。其中计算机科学是 20 世纪中叶发展起来的学科，其基础理论追根溯源包括离散数据、密码学、计算理论、量子信息、数字逻辑、编译原理和信息论等。

2013 年美国国家科学基金会数学科学部发表《2025 年的数学科学》报告[31,32]。该报告指出，由于越来越多的领域依赖于复杂计算机模拟和海量数据分析，数学科学为计算机模拟与海量数据分析提供基本的语言，正成为社会科学的基础和许多新兴领域不可或缺的重要组成部分。数学科学涉及最广泛意义上的数学、统计学和计算综合，以及这些领域与潜在应用领域的相互作用。该基金会委员发现，数学正在不断扩大，数学科学和其他研究领域之间的边界变得模糊并

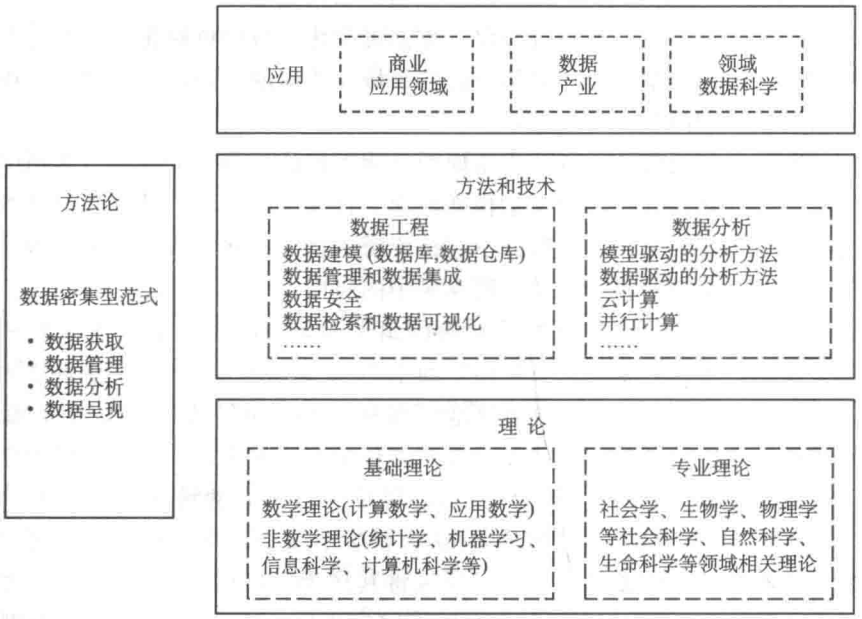

图 1-1 数据科学的学科体系

逐渐消失。自然科学、社会科学、生命科学、计算机科学和工程领域的研究与数学之间的联系越来越密切。数学科学的用途在不断扩展，21 世纪的大部分科学与工程都将建立在数学科学的基础上。

数据科学有两个内涵：一是研究数据本身，研究数据的各种类型、状态、属性及变化形式和变化规律；二是为自然科学和社会科学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。具体来说，数据科学的理论和方法可以应用于许多领域，开发出专门的理论、技术和方法，从而形成专门领域的数据学，如金融数据学、地理数据学、生物数据学、气象数据学等。因此，从这个视角来看，数据科学“理论”还应是包括在自然科学和社会科学中的各学科门类的基础理论。

1.3.2 数据科学方法论

方法论为做事情的原则和程序，与具体的方法不同，是人们用什么样的方式、方法来观察事物和处理问题。因此方法论主要解决“怎么办”的问题。

文献［28］认为数据学研究的一般工作流程是：从数据世界获得一个数据集合；对该数据集的整体性进行探测性分析；然后进行数据研究分析或进行数据实验；发现数据规律；将发现的结果和规律可视化等。文献［7］在对人类社会曾

采用的科学研究范式进行总结的基础上（表 1-3）指出，在数据爆炸时代，一种新的科学研究模式是：通过仪器收集数据或通过模拟方法产生数据，然后利用计算机软件处理和分析数据，再将形成的信息和知识存储在计算机中。这种模式与先前的实验研究、理论研究和计算仿真迥然不同，称为科学探索的“第四范式”或 eScience 科研模式。

表 1-3　科学发现的数据四种范式

时期	科学研究范式
几千年前	实验范式：以探索自然规律和描述自然现象为目标
最近几百年	理论范式：采用模型和归纳的方法
最近几十年	计算范式：对复杂现象进行仿真或模拟
当今 eScience	理论-实验-计算机仿真范式：数据密集型的知识发现

文献［9］从商业领域典型的大数据应用中，总结和归纳出大数据处理的基本流程：在合适的工具辅助下，对广泛异构的数据源（来自数据库、网页、文档等）进行抽取和集成，然后按照一定的标准统一存储，之后再利用合适的数据技术对存储的数据进行分析（包括机器学习、数据挖掘、统计分析等），从中提取有益的知识并利用恰当的方式将分析结果展现给终端用户（包括研究者、企业和政府等）。

综合上述观点，无论是在科学研究领域还是在商业应用领域，数据科学的“方法论”应涵盖采用恰当的工具，支持从数据获取、数据管理、数据分析、数据呈现到可视化整个周期。

1.3.3　数据科学方法

数据科学的“方法”是围绕数据获取、数据管理、数据分析、数据呈现或可视化等主题的一系列方法或技术。文献［28］将数据科学的方法划分为数据获取、数据分析和数据感知三大类。本书将其划分为数据工程和数据分析两大类别。

数据工程重点研究数据建模、数据标准化、数据管理、数据集成、数据和信息检索、数据库安全等方面的相关方法和技术。例如，数据管理包括数据库技术、数据仓库、数据存储和数据备份等技术，以及数据质量管理的方法。

数据分析方法包括：传统的模型驱动的方法，如统计方法；数据驱动的方法，如机器学习和数据挖掘等；处理大数据的云计算和并行计算等。伴随各学科理论和信息技术的发展，数据科学“技术”不断创新和演化，见表 1-4。