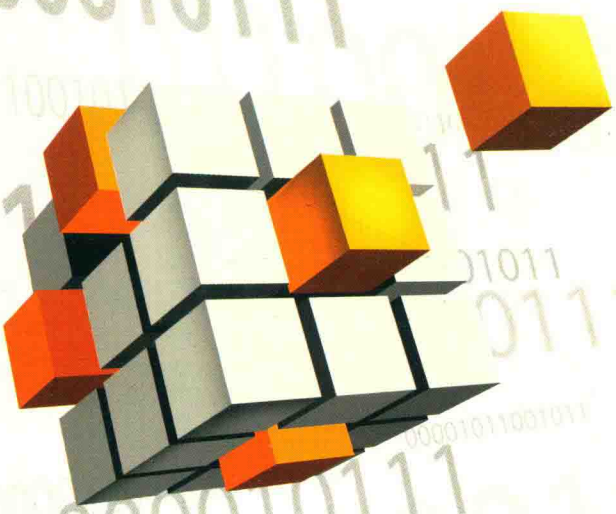
新世纪高等学校规划教材 · 大数据系列



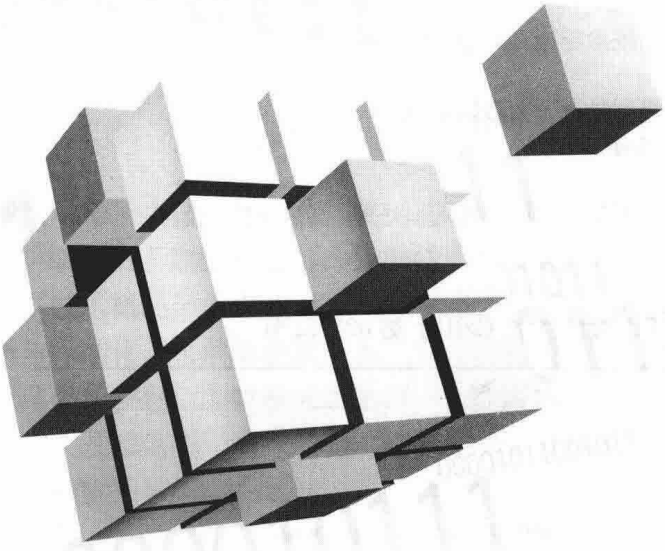
陈 明◎编著

数据科学与大数据技术导论

- 注重基本内容与基本方法的介绍
- 走进数据科学与大数据技术大门

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

新世纪高等学校规划教材 · 大数据系列



陈 明◎编著

数据科学与大数据技术导论

SHUJU KEXUE YU DASHUJU JISHU DAOLUN

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

图书在版编目（CIP）数据

数据科学与大数据技术导论/陈明编著. —北京：北京师范大学出版社，2018.6
新世纪高等学校规划教材·大数据系列
ISBN 978-7-303-23452-3

Ⅰ. ①数… Ⅱ. ①陈… Ⅲ. ①数据处理-高等学校-教材
Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 025112 号

营 销 中 心 电 话 010-62978190 62979006
北师大出版社科技与经管分社网 http://jsws.bnupg.com
电 子 信 箱 kjjg@bnupg.com

出版发行：北京师范大学出版社 www.bnupg.com
北京新街口外大街 19 号
邮政编码：100875
印 刷：北京京师印务有限公司
经 销：全国新华书店
开 本：787mm×1092mm 1/16
印 张：22.75
字 数：511 千字
版 次：2018 年 6 月第 1 版
印 次：2018 年 6 月第 1 次印刷
定 价：49.80 元

策划编辑：赵洛育 责任编辑：赵洛育
美术编辑：刘 超 装帧设计：刘 超
责任校对：王 云 责任印制：赵非非

内 容 简 介

大数据技术是一个面向实际应用的技术。从大数据中获取有价值信息是大数据技术的精髓。本书详细介绍了数据科学与大数据技术的主要内容，全书分为13章，主要包括数据科学与大数据技术概述、大数据获取与存储管理技术、大数据抽取技术、数据清洗技术、大数据去噪与标准化、大数据约简技术、大数据集成技术、大数据挖掘技术、大数据分析、大数据分析结果的解释、大数据机器学习、大数据离线计算技术、大数据流式计算技术等。本书在内容上，注重概念、方法介绍，实例丰富、语言精练、逻辑层次清晰，可作为大学数据科学与大数据技术专业和相近专业的教材，也可以作为科技人员的参考书。

前　言

大数据技术与应用展现出锐不可当的强大生命力，科学界与企业界寄予无比的厚望。大数据成为继20世纪末、21世纪初互联网蓬勃发展以来的又一轮IT工业革命。

大数据技术是指从数据采集、清洗、集成、挖掘、分析与结果解释，进而从各种各样的巨量数据中快速获得有价值信息的全部技术。大数据技术的精髓是从数据挖掘和分析中获取具有重要价值的信息、产生新见解的能力、识别复杂关系和做出更加精准的预测。

大数据技术是现代科学与技术发展，尤其是计算机科学技术发展的重要成果和结晶，是科学发展史一个新的里程碑。大数据的出现对计算机许多领域提出了挑战与冲击，推动了计算机科学技术的发展。

大数据技术的出现凝集了多学科的研究成果，是一门多学科的交叉融合技术，随着科学技术的进步，大数据技术发展更为迅速，应用更为深入与广泛，并凸显其巨大潜力和应用价值。

本书在体例的设计方面，以大数据技术为核心展开。大数据技术是指在大数据处理周期中所使用的全部技术的集合，也就是说，从数据获取之后，经过存储、抽取、清洗、约简、分析、挖掘等步骤后，获得有价值信息的全过程所需要的技术集合。本书的第1章～第11章主要介绍了上述过程。除此之外，还在第12章～第13章中介绍了大数据离线计算技术和大数据流式计算技术等内容。其中，在大数据离线计算技术中，主要包括大数据离线处理架构、MapReduce的体系结构、基于Hadoop框架的分布计算和MapReduce程序设计实例分析。在大数据流式计算技术中，主要包括流式数据的计算模式、流式计算的应用、流式计算的系统架构、高可用技术、Storm处理过程和大数据流式计算的应用案例等。

本书在内容的描述方面，注重大数据技术的主要概念、结构和方法的清晰描述。对主要的算法，如分类算法、聚类算法等给出了形式化描述。

本书在结构上为积木状，各章内容独立地进行概念性与方法性论述。出于篇幅考虑，书中所提及定理没有给出证明，如需要可以查阅相关文献。由于作者水平有限，书中不足之处在所难免，敬请读者批评指正。

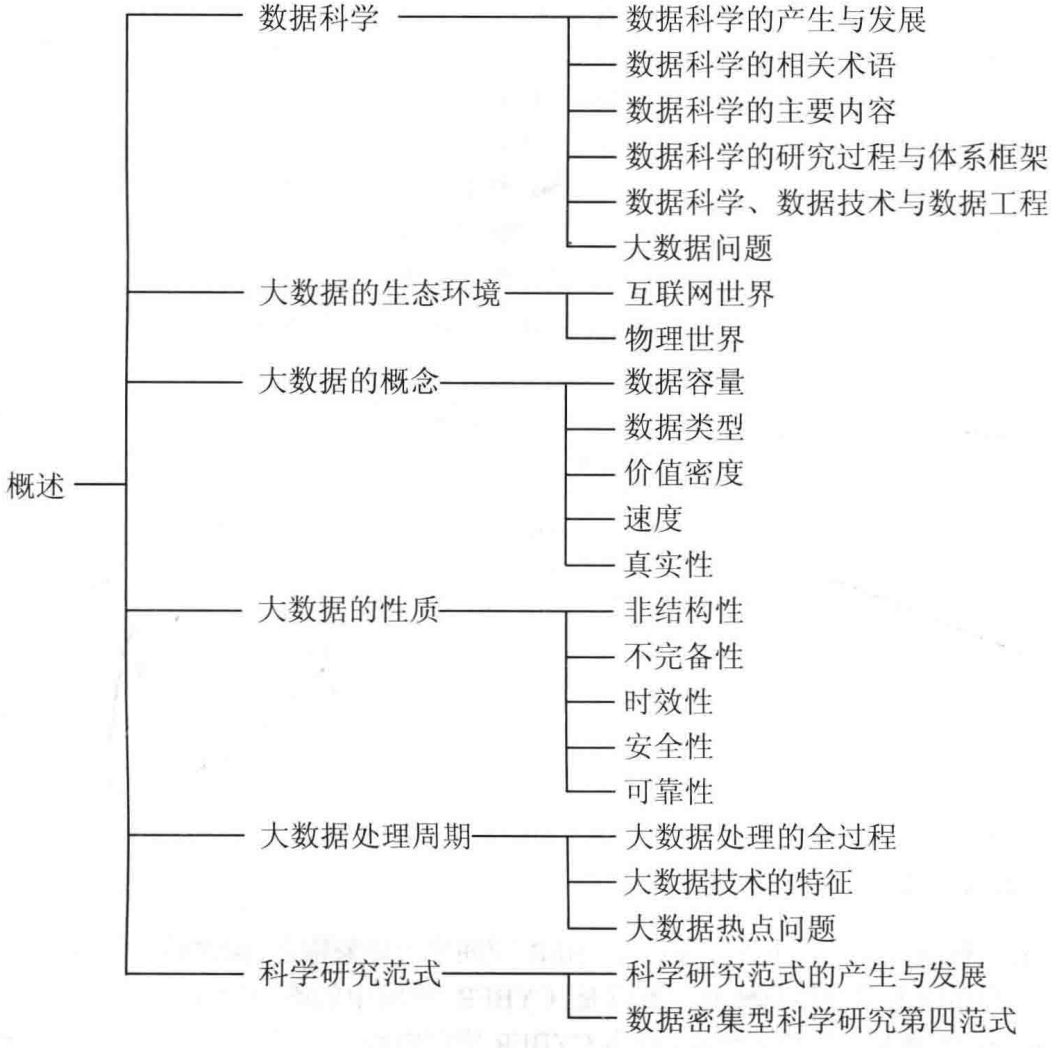
陈明

2017.8

目　录

第1章 概 述

本章主要内容



世界已经进入了一个模型和假设逐渐清晰的大数据时代，计算机科学是算法与算法变换的科学，数据科学研究范围更为广泛。数据科学研究与进展不仅可以推动数学、计算机科学、统计学、天体信息学、生物信息学、计算社会学等学科的发展，而且还能够大力助推产业发展与技术进步。

1.1 数据科学

数据科学是关于数据的科学，数据科学是为研究探索 CYBER 空间中数据界的理论、方法和技术。

1.1.1 数据科学的产生与发展

数据科学出现在 20 世纪 60 年代，1974 年彼得·诺尔出版的《计算机方法的简明调查》中将数据科学定义为“处理数据的科学，一旦数据与其代表事物的关系被建立起来，将为其他领域与科学提供借鉴”。1996 年在日本召开的数据科学、分类和相关方法会议上，将数据科学作为会议的主题词。2001 年美国统计学教授威廉·S.克利夫兰发表了《数据科学：拓展统计学的技术领域的行动计划》，首次将数据科学作为一个单独的学科，并把数据科学定义为统计学领域扩展到以数据作为现金计算对象相结合的部分，奠定了数据科学的理论基础。

1.1.2 数据科学的相关术语

1. CYBER 空间

CYBER 空间意译为异次元空间、多维信息空间、计算机空间、网络空间等。其本意是指以计算机技术、现代通信网络技术、虚拟现实技术等信息技术的综合运用为基础，以知识和信息为内容的新型空间，是人类运用知识创造的人工世界并用于知识交流的虚拟空间。信息化是将现实世界中的事物和现象以数据的形式存储到 CYBER 空间中，是一个数据生产的过程。数据是自然和生命的一种表示形式，记录了人类的行为，包括工作、生活和社会的发展。

2. 数据爆炸

将数据快速大量地产生并存储在 CYBER 空间中的现象称之为数据爆炸，数据爆炸在 CYBER 空间中形成数据自然界。数据是 CYBER 空间中的唯一存在，需要研究和探索 CYBER 空间中数据的规律和现象。探索 CYBER 空间中数据的规律和现象就是探索宇宙的规律、探索生命的规律、寻找人类行为的规律、寻找社会发展的规律的一种重要手段。

3. 数据科学的定义

数据科学是关于数据的科学或者研究数据的科学、探索 CYBER 空间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。与自然科学和社会科学不同，数据科学的研究对象是 CYBER 空间的数据，是新的科学。数据科学主要包括两个方面：一个是研究数据本身，研究数据的各种类型、状态、属性及变化形式和变化规律；另一个是为自

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现象和规律。

4. 数据科学的方法和技术

数据科学采用收集数据的形式，进行开放式分析，不做预先假定。在许多数据科学研究项目中，首先需要浏览原始数据，形成一个假定，然后基于假定进行调查确认。数据科学的关键概念是，数据科学是一个经验科学，直接基于数据进行科学处理。数据科学已经有一些方法和技术，如数据获取、数据存储与管理、数据安全、数据分析、可视化等。其基础理论和新技术主要有数据存在性、数据测度、时间、数据代数、数据相似性与簇论、数据分类与数据百科全书、数据伪装与识别、数据实验、数据感知等。数据学的理论和方法将改进现有的科学研究方法，形成新型的科学研究方法，并且针对各个研究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，如行为数据学、生命数据学、脑数据学、气象数据学、金融数据学、地理数据学和过程数据学等。

数据科学不仅完成分析，还要涉及整个端到端的生命周期，数据系统本质上是用于研发真实世界理解模型的科学设备。这就表明必须深刻理解数据的来源、数据转换的适用性和准确性、转换算法和过程之间的相互作用以及数据存储机制。进而能够确保所有事物都正确执行，从而探索数据、创建并验证各项科学假设。

1.1.3　数据科学的主要内容

数据科学的内容主要包括基础理论研究、数据技术、应用研究及数据科学的学科体系研究，如图 1-1 所示。数据科学学科建立，需要完成知识结构、课程设置、专业设置等学科体系建设，探讨数据科学与自然科学和社会科学之间的关系，数据科学与计算机科学和信息科学之间的关系等。

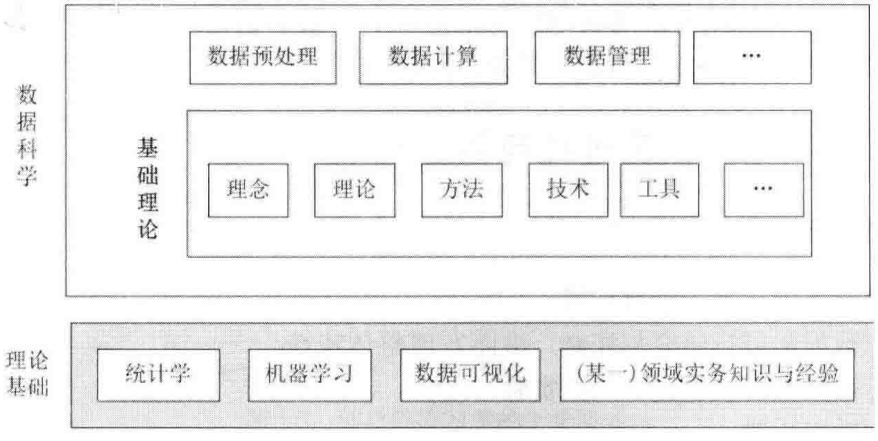


图 1-1　数据科学的内容

数据科学包括用数据的方法研究科学技术和用科学的方法研究数据的两方面。前者包括生物信息学、天体信息学和数字地球等领域，后者包括统计学、机器学习、数据挖掘等领域，这些学科都是数据科学的重要组成部分。

1. 用科学的方法研究数据

比较常用的数据有表格、点集、时间序列、图像、视频、网页和网络数据等。基本的数据结构包括度量结构、网络结构和代数结构。

用科学的方法研究数据，主要包括数据处理周期的所有技术，其核心问题是数据分析。数据分析的基本问题是找出模型，由于数据中含有噪声导致了随机模型的出现。为了减少求解的难度，可以找到模型一部分，或将随机模型简化为确定性模型。观察和逻辑推理是科学的基础，数据自然界中主要采用了观察方法与数据推理的理论和方法，包括数据的存在性、数据测度、时间、数据代数、数据分类、数据相似性与簇论等。

2. 用数据的方法研究科学

用数据的方法研究科学问题并不是不需要模型，只是模型的出发点不同，不是从基本原理的角度去寻找模型，通常是基于更为简单的数学模型。

需要建立数据科学的实验方法，需要提出科学假说和建立理论体系，并通过这些实验方法和理论体系进行数据自然界的探索研究，从而掌握数据的各种类型、状态、属性、变化形式和变化规律，揭示自然界和人类行为现象和规律。

3. 领域数据学

将数据科学的理论和方法广泛应用，开发出专门的理论、技术和方法，从而形成专门领域的数据科学，例如，脑数据学、行为数据学、生物数据学、气象数据学、金融数据学和地理数据学等。

4. 数据资源的开发方法和技术

数据资源是重要的现代战略资源，具有巨大的价值，越来越凸显重要性，是继石油、煤炭、矿产等传统资源之后的最重要的资源之一。人类的社会、政治和经济都将依赖于数据资源，而石油、煤炭、矿产等传统资源的勘探、开采、运输、加工、产品销售等都依赖数据资源，离开了数据资源，将无法开展与完成这些工作。

1.1.4 数据科学的研究过程与体系框架

1. 数据科学的研究过程

① 从自然界中获得一个数据集。
② 对该数据集进行研究与探索，进而发现整体特性。
③ 进行数据分析或者进行数据实验。
④ 发现数据规律。
⑤ 将数据进行感知化与可视化等。

2. 数据科学的构成

数据科学的组成要素可以从图 1-2 所示的维恩图中得到线索，主要包括计算机科学、

数学、统计学知识，以及领域的专业知识。

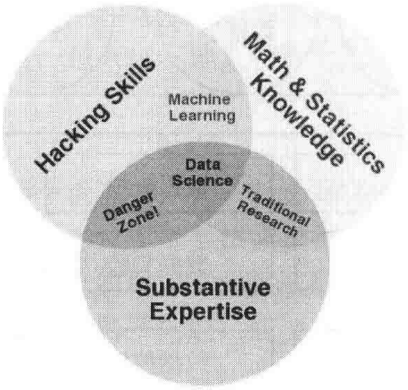


图 1-2　数据科学的组成要素

- Hacking Skills：编程技巧。
- Danger Zone：危险地带。
- Machine Learning：机器学习。
- Data Science：数据科学。
- Traditional Research：传统研究。
- Substantive Expertise：行业知识。
- Math & Statistics Knowledge：数学与统计学知识。

进一步细化数据科学的 12 个主要领域，如图 1-3 所示。

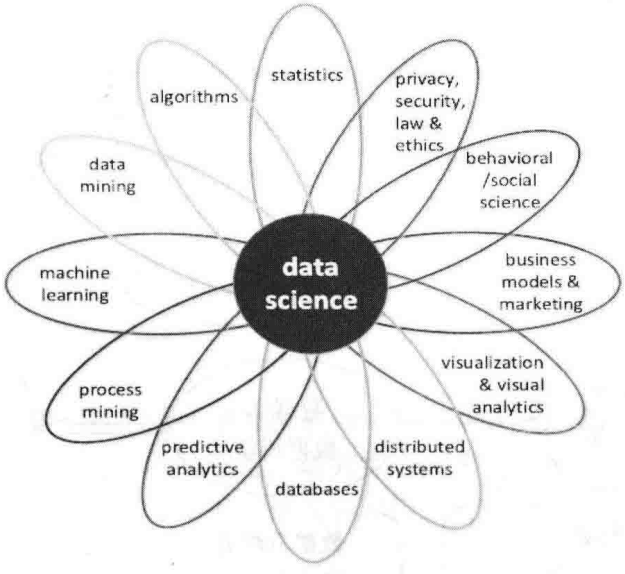


图 1-3　细化数据科学的主要领域

3. 数据科学的体系框架

数据科学的体系框架如图 1-4 所示。在图的上部描述了数据的内容，下部分是数据科学的基础描述。

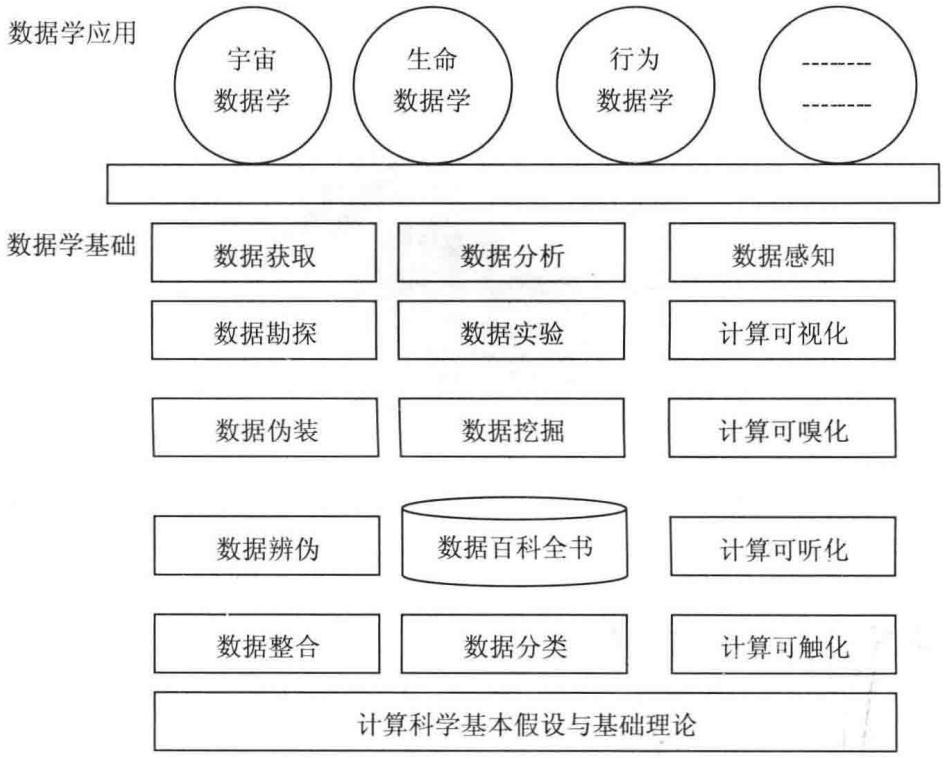


图 1-4　数据科学体系框架

（1）研究从数据中获取信息与知识

数据科学的研究对象、研究目的与研究方法等与计算科学、信息科学不同。数据存于CYBER 空间中，信息是自然界、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。数据可以作为信息和知识符号的表示或载体，但数据本身并不是信息或知识。数据科学的研究对象是数据，而不是信息，也不是知识。通过研究数据来获取对自然、生命和行为的认识，进而获得重要信息与知识。数据科学与其他学科的关系如图 1-5 所示。

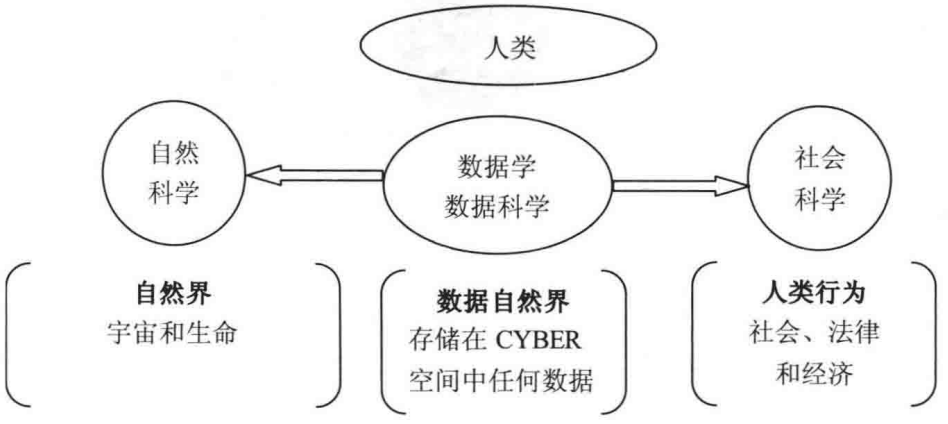


图 1-5　数据科学与其他学科的关系

（2）通过分析数据来认识自然和行为

自然科学研究自然现象和规律，认识的对象是整个自然界物质的各种类型、状态、属