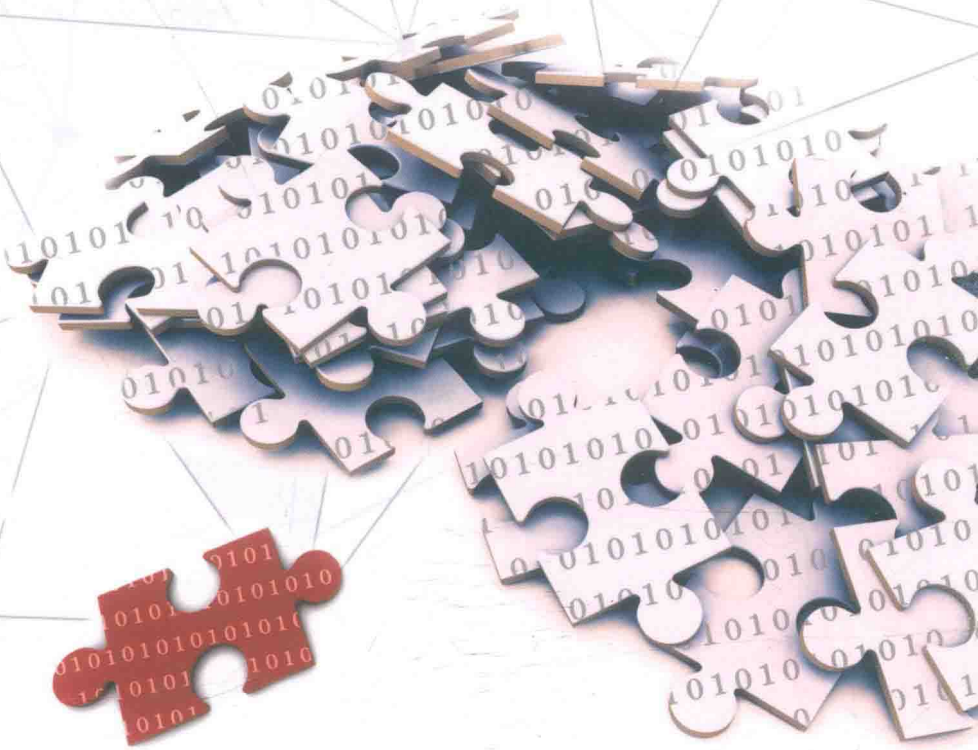
清华大学电子工程系核心课系列教材

Data and Algorithms

数据与算法

◎吴及　陈健生　白铂　编著



清華大學出版社

清华大学电子工程系核心课系列教材

Data and Algorithms

数据与算法

吴及　陈健生　白铂　编著

清华大学出版社
北京

内 容 简 介

本书从数据与算法的相互关系入手，内容涵盖了传统的数据结构和数值分析，并增加了数学模型和算法设计思想的介绍。第一部分数据、数学模型和算法的基本概念是全书的基础；在数据结构部分主要从数学模型和问题的角度介绍了线性结构、树结构、图结构，最常见的非数值问题查找和排序；在数值分析部分从问题的角度介绍了误差分析，实数的表示和运算，一元非线性方程，线性方程组，拟合与插值，最优化问题；最后一部分则从算法设计思想的角度介绍了蛮力法、分治法、贪心法、动态规划、搜索算法和随机算法，以及在求解具体问题时的应用实例。书中问题和算法两个视角构成了纵横交织的网络，希望能够帮助读者更清楚地看到数据和算法的相互关系，更透彻地理解数值和非数值问题的差异和共性，更全面地提升利用计算机作为工具解决实际问题的能力，为今后的学习和未来的发展打下扎实的基础。

图书在版编目(CIP)数据

数据与算法/吴及，陈健生，白铂编著.—北京：清华大学出版社，2017
(清华大学电子工程系核心课系列教材)
ISBN 978-7-302-46881-3

Ⅰ.①数… Ⅱ.①吴… ②陈… ③白… Ⅲ.①数据结构–高等学校–教材 ②算法分析–高等学校–教材 Ⅳ.①TP311.12

中国版本图书馆 CIP 数据核字(2017)第 064079 号

责任编辑：文　怡
封面设计：台禹微
责任校对：李建庄
责任印制：杨　艳

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　张**：22.5　　**字　数**：565 千字
版　　次：2017 年 10 月第 1 版　　**印　次**：2017 年 10 月第 1 次印刷
印　　数：1～2000
定　　价：59.00 元

产品编号：072343-01

丛 书 序

清华大学电子工程系经过整整十年的努力，正式推出新版核心课系列教材。这成果来之不易！在这个时间节点重新回顾此次课程体系改革的思路历程，对于学生，对于教师，对于工程教育研究者，无疑都有重要的意义。

一

高等电子工程教育的基本矛盾是不断增长的知识量与有限的学制之间的矛盾。这个判断是这批教材背后最基本的观点。

当今世界，科学技术突飞猛进，尤其是信息科技，在 20 世纪独领风骚数十年，至 21 世纪，势头依然强劲。伴随着科学技术的迅猛发展，知识的总量呈现爆炸性增长趋势。为了适应这种增长，高等教育系统不断进行调整，以把更多新知识纳入教学。自 18 世纪以来，高等教育响应知识增长的主要方式是分化：一方面延长学制，从本科延伸到硕士、博士；一方面细化专业，比如把电子工程细分为通信、雷达、图像、信息、微波、线路、电真空、微电子、光电子等。但过于细化的专业使得培养出的学生缺乏处理综合性问题的必要准备。为了响应社会对人才综合性的要求，综合化逐步成为高等教育主要的趋势，同时学生的终身学习能力成为关注的重点。很多大学推行宽口径、厚基础本科培养，正是这种综合化趋势使然。通识教育日益受到重视，也正是大学对综合化趋势的积极回应。

清华大学电子工程系在 20 世纪 80 年代有九个细化的专业，20 世纪 90 年代合并成两个专业，2005 年进一步合并成一个专业，即“电子信息科学类”，与上述综合化的趋势一致。

综合化的困难在于，在有限的学制内学生要学习的内容太多，实践训练和课外活动的时间被挤占，学生在动手能力和社会交往能力等方面的发展就会受到影响。解决问题的一种方案是延长学制，比如把本科定位在基础教育，硕士定位在专业教育，实行五年制或六年制本硕贯通。这个方案虽可以短暂缓解课程量大的压力，但是无法从根本上解决知识爆炸性增长带来的问题，因此不可持续。解决问题的根本途径是减少课程，但这并非易事。减少课程意味着去掉一些教学内容。关于哪些内容可以去掉，哪些内容必须保留，并不容易找到有高度共识的判据。

探索一条可持续有共识的途径，解决知识量增长与学制限制之间的矛盾，已是必需，也是课程体系改革的目的所在。

二

学科知识架构是课程体系的基础，其中核心概念是重中之重。这是这批教材背后最关键的观点。

布鲁纳特别强调学科知识架构的重要性。架构的重要性在于帮助学生利用关联性来理解和重构知识；清晰的架构也有助于学生长期记忆和快速回忆，更容易培养学生举一反三的迁移能力。抓住知识架构，知识体系的脉络就变得清晰明了，教学内容的选择就会有公认的依据。

核心概念是知识架构的汇聚点，大量的概念是从少数核心概念衍生出来的。形象地说，核心概念是干，衍生概念是枝、是叶。所谓知识量爆炸性增长，很多情况下是“枝更繁、叶更茂”，而不是产生了新的核心概念。在教学时间有限的情况下，教学内容应重点围绕核心概念来组织。教学内容中，既要有抽象的概念性的知识，也要有具体的案例性的知识。

梳理学科知识的核心概念，这是清华大学电子工程系课程改革中最为关键的一步。办法是梳理自 1600 年吉尔伯特发表《论磁》一书以来，电磁学、电子学、电子工程以及相关领域发展的历史脉络，以库恩对“范式”的定义为标准，逐步归纳出电子信息科学技术知识体系的核心概念，即那些具有“范式”地位的学科成就。

围绕核心概念选择具体案例是每一位教材编者和教学教师的任务，原则是具有典型性和时代性，且与学生的先期知识有较高关联度，以帮助学生从已有知识出发去理解新的概念。

三

电子信息科学与技术知识体系的核心概念是：信息载体与系统的相互作用。这是这批教材公共的基础。

1955 年前后，斯坦福大学工学院院长特曼和麻省理工学院电机系主任布朗都认识到信息比电力发展得更快，他们分别领导两所学校的电机工程系进行了课程改革。特曼认为，电子学正在快速成为电机工程教育的主体。他主张彻底修改课程体系，牺牲掉一些传统的工科课程以包含更多的数学和物理，包括固体物理、量子电子学等。布朗认为，电机工程的课程体系有两个分支，即能量转换和信息处理与传输。他强调这两个分支不应是非此即彼的两个选项，因为它们都基于共同的原理，即场与材料之间相互作用的统一原理。

场与材料之间的相互作用，这是电机工程第一个明确的核心概念，其最初的成果形式是麦克斯韦方程组，后又发展出量子电动力学。自彼时以来，经过大半个世纪的飞速发展，场与材料的相互关系不断发展演变，推动系统层次不断增加。新材料、新结构形成各种元器件，元器件连接成各种电路，在电路中，场转化为电势（电流电压），“电势与电路”取代“场和材料”构成新的相互作用关系。电路演变成开关，发展出数字逻辑电路，电势二值化为比特，“比特与逻辑”取代“电势与电路”构成新的相互作用关系。数字逻辑电路与计算机体系结构相结合发展出处理器（CPU），比特扩展为指令和数据，进而组织成程序，“程

序与处理器”取代“比特与逻辑”构成新的相互作用关系。在处理器基础上发展出计算机，计算机执行各种算法，而算法处理的是数据，“数据与算法”取代“程序与处理器”构成新的相互作用关系。计算机互联出现互联网，网络处理的是数据包，“数据包与网络”取代“数据与算法”构成新的相互作用关系。网络服务于人，为人的认知系统提供各种媒体（包括文本、图片、音视频等），“媒体与认知”取代“数据包与网络”构成新的相互作用关系。

以上每一对相互作用关系的出现，既有所变，也有所不变。变，是指新的系统层次的出现和范式的转变；不变，是指“信息处理与传输”这个方向一以贯之，未曾改变。从电子信息的角度看，场、电势、比特、程序、数据、数据包、媒体都是信息的载体；而材料、电路、逻辑（电路）、处理器、算法、网络、认知（系统）都是系统。虽然信息的载体变了，处理特定的信息载体的系统变了，描述它们之间相互作用关系的范式也变了，但是诸相互作用关系的本质是统一的，可归纳为“信息载体与系统的相互作用”。

上述七层相互作用关系，层层递进，统一于“信息载体与系统的相互作用”这一核心概念，构成了电子信息科学与技术知识体系的核心架构。

四

在核心知识架构基础上，清华大学电子工程系规划出十门核心课：电动力学（或电磁场与波）、固体物理、电子电路与系统基础、数字逻辑与 CPU 基础、数据与算法、通信与网络、媒体与认知、信号与系统、概率论与随机过程、计算机程序设计基础。其中，电动力学和固体物理涉及场和材料的相互作用关系，电子电路与系统基础重点在电势与电路的相互作用关系，数字逻辑与 CPU 基础覆盖了比特与逻辑及程序与处理器两对相互作用关系，数据与算法重点在数据与算法的相互作用关系，通信与网络重点在数据包与网络的相互作用关系，媒体与认知重点在媒体和人的认知系统的相互作用关系。这些课覆盖了核心知识架构的七个层次，并且有清楚的对应关系。另外三门课是公共的基础，计算机程序设计基础自不必说，信号与系统重点在确定性信号与系统的建模和分析，概率论与随机过程重点在不确定性信号的建模和分析。

按照“宽口径、厚基础”的要求，上述十门课均被确定为电子信息科学类学生必修专业课。专业必修课之前有若干数学物理基础课，之后有若干专业限选课和任选课。这套课程体系的专业覆盖面拓宽了，核心概念深化了，而且教学计划安排也更紧凑了。近十年来清华大学电子工程系的教学实践证明，这套课程体系是可行的。

五

知识体系是不断发展变化的，课程体系也不会一成不变。就目前的知识体系而言，关于算法性质、网络性质、认知系统性质的基本概念体系尚未完全成型，处于范式前阶段，相应的课程也会在学科发展中不断完善和调整。这也意味着学生和教师有很大的创新空间。电动力学和固体物理虽然已经相对成熟，但是从知识体系角度说，它们应该覆盖场与材料（电荷载体）的相互作用，如何进一步突出“相互作用关系”还可以进一步探讨。随着集

成电路发展，传统上区分场与电势的条件，即电路尺寸远小于波长，也变得模糊了。电子电路与系统或许需要把场和电势的理论相结合。随着量子计算和量子通信的发展，未来在逻辑与处理器和通信与网络层次或许会出现新的范式也未可知。

工程科学的核心概念往往建立在技术发明的基础之上，比如目前主流的处理器和网络分别是面向冯·诺依曼结构和 TCP/IP 协议的，如果体系结构发生变化或者网络协议发生变化，那么相应地，程序的概念和数据包的概念也会发生变化。

六

这套课程体系是以清华大学电子工程系的教师和学生的基本情况为前提的。兄弟院校可以参考，但是在实践中要结合自身教师和学生的情况做适当取舍和调整。

清华大学电子工程系的很多老师深度参与了课程体系的建设工作，付出了辛勤的劳动。在这一过程中，他们表现出对教育事业的忠诚，对真理的执着追求，令人钦佩！自课程改革以来，特别是 2009 年以来，数届清华大学电子工程系的本科同学也深度参与了课程体系的改革工作。他们在没有教材和讲义的情况下，积极支持和参与课程体系的建设工作，做出了重要的贡献。向这些同学表示衷心感谢！清华大学出版社多年来一直关注和支持课程体系建设工作，一并表示衷心感谢！

王希勤

2017 年 7 月

前　言

近年来，信息科学技术呈现快速发展的态势，云计算、移动互联网、大数据、人工智能，给我们所处的时代和社会带来了一波又一波的冲击。人们日常生活中的信息获取方式、社会交往方式、生产工作方式都已经发生了很大的变化，而且更为巨大的变化似乎就在并不遥远的未来。我们的教学和人才培养模式如何能够适应这样的变化，对于高等教育的从业者来说是一个严峻的挑战。

为了解决膨胀的知识量与有限的学制之间的矛盾，提高教学效率和质量，培养拔尖型创新人才，清华大学电子工程系进行了全面的教学改革。在梳理出电子信息科学知识构架的基础上，构建起了全新的课程体系，数据与算法就是其中的一门核心课程。数据是客观世界的描述，是信息的载体，也是算法的处理对象；算法是解决问题的方法和步骤，是处理数据的系统。因此数据与算法的关系，本质上是信息载体与系统的相互作用。同时，数据的特性是算法设计中不可忽视的关键性因素，对数据特性利用得越充分，算法的性能和效率就越高，但与此同时，算法的针对性越强，适用面也就越窄。

传统上数据结构和数值分析是两门课程，前者主要研究非数值问题，后者主要研究数值问题。但是，当我们上升到更宏观的视角，也就是数据与算法相互关系的视角，我们就能够更清楚地认识到两者之间的共性和差异。从共性上来讲，它们都把现实世界的问题简化成为数学模型上的问题，并利用计算机作为工具加以求解，因此有很多算法思想不仅能用于处理非数值问题，也能有效地处理数值问题。例如，二分法既可以用于实现有序线性表的高效查找，又可以用于求解非线性方程；寻找图的最小生成树的 Prim 算法、Kruskal 算法和求解多维函数极值的最速下降法都是基于贪心算法的思想。数值和非数值问题的差异也很显著，数值问题中变量取值是连续的，符合一定精度要求的近似解可能有无穷多个，只需要得到符合精度要求的近似解就足够了，因此误差分析在数值分析中处于基础性的地位；而非数值问题的解空间是离散的，不需要考虑误差。在实际应用中，数值问题和非数值问题还经常交织在一起，例如搜索引擎已经成为人们获取信息的主要方式，在其实现过程中就既有非数值问题，也有数值问题。

数据和算法的覆盖范围包括了传统上的数据结构、数值分析两门课程，同时还特别加入了数学模型和算法设计思想的部分，并从总体上对内容上进行了取舍。我们希望这门新设计的课程，能够让同学在学习过程中更清楚地看到数据和算法的相互关系，更透彻地理解数值和非数值问题的差异和共性，更全面地提升利用计算机作为工具解决实际问题的能力，为今后的学习和未来的发展打下扎实的基础。

全书共有 9 章，分为四个部分。第一部分是第 1 章，介绍了数据，算法和数学模型的基

本概念，是全书的基础。第二部分是第 2~5 章，包括线性结构、树结构和图结构，以及查找、排序两种最常用的非数值问题及其求解，是传统数据结构的内容。第三部分是第 6、7 章，包括数值问题和最优化的初步介绍，讨论的是数值问题。第四部分是第 8、9 章，介绍了随机算法和算法设计思想。

本书第 1、2、4、5 章由吴及负责撰写；第 6、7、8 章由陈健生负责撰写；白铂撰写了第 3 和第 9 章，并参与了第 4 和第 7 章的部分工作。

由于编者水平有限，疏误之处在所难免，敬请同行及各界读者批评指正。作为突破传统教学模式和内容组织方式的一次尝试，我们也希望这样的努力能够成为电子信息学科教学改革的有益探索。

编者

2017 年 8 月

目　录

第 1 章　数据、数学模型和算法

1.1　数据时代

大数据 (big data) 是当今学术界和产业界最炙手可热的名词之一，其重要性和价值已经得到广泛的认同。数据科学，也继实验科学、理论科学、计算机仿真之后，被称为科学研究的第四范式。为什么数据处理技术会得到如此普遍的重视？为什么人类会对数据中的价值寄予如此巨大的期望？又为什么人类社会发展到今天，数据的重要性会特别地凸显出来呢？我们从什么是数据谈起。

1.1.1　什么是数据

“数”是人们用来表示事物的量的基本数学概念。在人类发展的历史上，这种抽象的“数”的概念是从具体事物中逐步获得和建立起来的。例如“一个苹果”“二个橘子”“三个香蕉”描述的是具体的事物，而“一”“二”“三”则是与具体事物无关的抽象的“数”。另一个相关的概念是“数字”，数字是人们用来计数的符号，如现在人们常用的阿拉伯数字“1”“2”“3”，又如中文的数字“一”“二”“三”和罗马数字“Ⅰ”“Ⅱ”“Ⅲ”。而我们在这里要讨论的“数据”，则是一个范围大得多的概念。

数据是客观事物的符号表示，往往是通过对客观事物的观察得到的未经加工的原始素材，是包含知识和信息的原始材料。在今天的信息社会中，数据可以说无处不在，其表现形式也是多种多样，例如：

文字和符号: 不仅普遍存在于书籍、报纸等传统的纸质媒介上，也广泛存在于计算机、手机、平板电脑等电子设备上；既包括今天人们使用的各种文字符号，也包括从远古时代遗存下来的象形文字和甲骨文等。

多媒体数据：计算机的图形界面、广播电视电影、数码相机 (DC) 和数码摄像机 (DV)，使得我们身处于丰富多彩的多媒体时代。多媒体数据的采集、保存和播放已经非常方便；图像、音频、视频等各种媒体数据在我们的日常生活中随处可见。

通信信号：电信号和电磁波已经成为人类社会信息最方便快捷的传输方式，这些用于通信和控制的电话信号、导航信号、手机信号、广播信号，无论是在发送端还是在接收端都是数据。

传感器采集的数据：通过各种各样的温度传感器、压力传感器，以及 CT、B 超、声呐等，人们可以采集到各种各样能够描述客观事物的数据。

社会性数据：人类社会生活的方方面面同样需要大量数据来描述，如社会普查数据、

人口统计数据和民意调查数据等，著名的如美国总统大选期间盖洛普所做的候选人支持率的民意测验；也包括紧密联系我们日常生活的经济运行数据，如物价、收入等。随着社交网络的发展和普及，人们之间通过互联网和移动互联网的交互行为也成为重要的海量数据来源。

可以很清楚地看到对数据的掌握和处理是当今社会的一个基本问题，在科研活动、经济活动、文化活动和政治活动中，我们随时都会面对各种各样的数据。数据和对数据的处理与我们每个人都息息相关。

我们在这里讨论的数据，进一步被特指为能够输入到计算机并被计算机处理的。

1.1.2 大数据时代

数据处理技术包括了数据的获取、数据的存储、数据的传输，以及针对数据的计算等。

数据是客观事物的表示和描述，人具有很强的获取数据的能力，如人对客观事物的观察，社会普查等；数据获取也可以通过多种多样的设备，如温度和压力等各种传感器，万用表和光谱仪等各种测量仪器，照相机和摄像机等图像视频采集设备，麦克风和录音机等声音采集设备，雷达接收机和卫星接收机等信号接收设备等。

传统的数据存储主要依靠纸质媒介，如书籍、报表和纸质文件等，典型的模拟存储介质有胶片和磁带。随着数字技术的发展，数字存储介质已经成为主流。从大型的磁盘存储系统，到容量越来越大的计算机硬盘，再到便携的移动硬盘、U 盘、光盘和闪存卡，存储容量不断增大，而且价格越来越便宜。

语言交流和书信曾是人类历史上数据传输和信息交互的主要手段。电磁波和电信号的发现和利用，造就了电话、电报等快捷的数据传输方式。互联网、移动通信，以及 USB 和 IEEE 1394 等高速率数据传输技术的发展，使数据传输的快速、高效和方便达到了前所未有的程度。

面向数据的计算涵盖了对数据的分析、管理和利用。其中既包括了以处理器性能为代表的计算能力，又包括了对数据进行处理以实现信息抽取和知识发现的技术方法。

随着信息技术的飞速发展，人类在数据采集、数据存储和数据传输方面的能力得到了长足的发展。我们都知道，二进制是数字计算机的基础，计算机存储容量的基本单位是字节 (Byte)，每个字节包含 8 个二进制位。为了描述不同规模的数据，人们定义了一系列的数据计量单位：

Bytes → Kilobyte(2^{10} Bytes) → Megabyte(2^{20} Bytes) → Gigabyte(2^{30} Bytes) → Terabyte(2^{40} Bytes) →Petabyte(2^{50} Bytes)→Exabyte(2^{60} Bytes)→Zettabyte(2^{70} Bytes)→ Yottabyte(2^{80} Bytes)

其中我们比较熟悉的有千字节 (KB)、兆字节 (MB) 和吉字节 (GB)。我们甚至难以想象更大的数据量单位意味着什么？美国国会图书馆所有藏书的数据约为 10TB。按照 2001 年的数据估算，美国国家航空航天局地球观测系统 (Earth Observing System) 三年的数据总和约为 1PB[1]。据称 1 个 ZB 大概相当于全世界所有海滩上的沙子总和，而 1 个 YB 大概相当于 7000 人体内的原子数总和 [2]。如果以每分钟 1MB 的速度不间断播放 MP3 格式的歌曲，1ZB 存储的歌曲可以让人听上 19 亿年。