

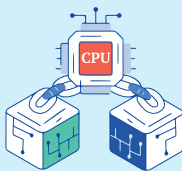
大话 计算机科学

生活中的计算思维

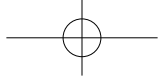
商静波 赵馨 著

×

费赛尔 绘



清华大学出版社
北京



内 容 简 介

计算机科学是当下最火热的学科之一，看似复杂、高大上，其实离每个人并不遥远。本书作为该学科的专业科普图书，通过常见的生活场景切入 38 个计算机科学知识点，涵盖计算机科学本科教育中的经典课程，包括大一、大二必修的“程序设计”“数据结构”“算法”课程，以及大三、大四选修的“数据科学”“机器学习”“智能系统”“信息安全”“计算机硬件”等课程。每个知识点通过独立故事呈现，读者可以按任意顺序阅读。

本书最大的特色是，以一个个日常生活中的寻常事例来讲计算机科学，没有公式和代码，旨在从原理和本质上讲透计算机科学的重要概念，同时让读者真切体会到计算思维在生活中随处可见。本书不仅故事丰富有趣，还配了精美插图，方便读者透彻理解内容。

本书适合对计算机科学感兴趣的中学生和非专业大众读者，亦可以作为高等院校计算机相关专业的导论课程、非计算机专业的通识课程的教材。

版权所有，侵权必究。举报：010-83470000，beiqinquan@tup.tsinghua.edu.cn。

图书在版编目（CIP）数据

大话计算机科学：生活中的计算思维 / 商静波，赵馨著；费赛尔绘. -- 北京：清华大学出版社，2024.

9. -- ISBN 978-7-302-67199-2

I . TP3-49

中国国家版本馆 CIP 数据核字第 2024NK7568 号

责任编辑：王中英

封面设计：杨玉兰

责任校对：徐俊伟

责任印制：刘 菲

出版发行：清华大学出版社

网 址：<https://www.tup.com.cn>，<https://www.wqxuetang.com>

地 址：北京清华大学学研大厦 A 座

邮 编：100084

社 总 机：010-83470000

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

印 装 者：北京博海升彩色印刷有限公司

经 销：全国新华书店

开 本：170mm×240mm

印 张：17.5

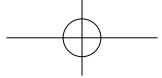
字 数：276 千字

版 次：2024 年 9 月第 1 版

印 次：2024 年 9 月第 1 次印刷

定 价：99.00 元

产品编号：108089-01



推荐语

(按收到的时间先后排序)

This is an excellent book on computer science for beginners. It takes complex concepts and breaks them down into simple, everyday language that anyone can understand. The illustrations are top-notch, and the connection between technical ideas and daily life is masterfully done. Whether you're new to computer science or just curious, this book is an absolute must-read. It's engaging, enlightening, and incredibly inspiring.

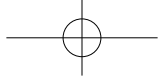
这是一本非常适合初学者的计算机科普读物，它将复杂的概念分解成简单的、日常的语言，使得任何人都能轻松理解。这本书的插图非常出色，技术思想与日常生活的联系处理得非常巧妙。无论你是计算机科学的新手，还是只对计算机科学感到好奇的读者，都非常推荐阅读本书——它既有趣又有启发性，令人非常振奋！

John Hopcroft (约翰·霍普克罗夫特)

1986 年图灵奖 (计算机科学领域最高奖) 得主，
2024 年中华人民共和国国际科学技术合作奖得主，

康奈尔大学终身教授

A truly interesting book that introduces advanced computer science concepts in such story-telling way that everyone can understand and enjoy! I love reading it!



● | 大话计算机科学：生活中的计算思维 | ●

这本书极其有趣！它通过讲故事的方式介绍了计算机科学中的复杂概念，让每个人都能轻松理解并沉浸其中。我爱不释手！

韩家炜

美国伊利诺伊大学厄巴纳香槟分校（UIUC）终身教授，

Michael Aiken 讲席教授，ACM Fellow，IEEE Fellow，

加拿大皇家科学院 Fellow

《大话计算机科学：生活中的计算思维》以生动有趣的实例和深入浅出的讲解，帮助读者掌握重要的计算机科学知识，并且在日常生活中运用计算思维来解决问题。这本书应该成为追求卓越的科技工作者和计算机爱好者的必读书目。

李航

字节跳动研究部门负责人

在数据和智能技术席卷全球的新时代，计算机科学成为国家发展和行业变革的重要基础，掌握计算机科学基础知识也将获得不同维度的思考视角和思维方式，从而激发更大的生产动能。本书用生动有趣的事例，深入浅出地演绎了计算机学科的基础知识，帮助读者轻松进入数字世界，开启一场奇妙的探索之旅。

郑宇

京东集团副总裁，IEEE Fellow，

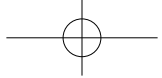
国家万人计划科技领军人才

本书通过生动有趣的例子，深入浅出地科普了计算机科学中从基础到前沿的各种知识。对于想入门计算机科学的同学、想引导孩子进入计算机领域的家长、希望获取灵感普及知识的专业人员，这本书都是非常适合的，强烈推荐给大家。

陈天奇

卡耐基梅隆大学（CMU）机器学习和计算机系助理教授，

知名开源项目 TVM、MXNet、XGBoost 作者



● | 推荐语 | ●

计算思维是现代生活中应对复杂问题的关键能力。本书以日常生活案例为切入点，为读者打开计算机科学世界的大门，是一本激发兴趣、培养思维的难得佳作，强烈推荐！

李亚洲

“机器之心”联合创始人、主编

我和业界很多前辈讨论过一个问题——能在计算机科学领域做得好的人，最重要的思维能力是什么？所有人最终的结论都一样，那就是具备“抽象思维能力”，这是一种能从复杂的事务中抽象出本质的能力，是学习计算机科学最重要的能力。可是，抽象思维能力本身就是一个抽象的东西，很难传授这种能力。在看到《大话计算机科学：生活中的计算思维》这本书之后，我意识到这本书讲的其实就是抽象思维能力，从日常生活中的现象中理解计算机科学的道理，就是对这种能力的锻炼。希望有更多年轻的朋友能够通过这本书了解到计算机科学并不是什么神秘莫测的东西，能够走进计算机科学的殿堂。

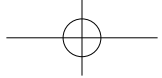
程墨

迪士尼流媒体技术总监，知乎新知答主

计算机的发明初衷是让它像人一样思考，因此它的另一个中文译名“电脑”格外贴切。经过近百年的发展，计算机已经凝聚了全球最杰出的人脑的智慧。如今，随着人工智能时代的到来，人类更应向计算机学习，将智慧用到自己的生活中。愿读者能通过阅读本书，学会商老师和赵律师的生活中的“电脑”智慧。

黎珍辉

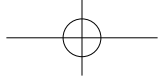
商老师在上海交大和 UIUC 的双科学姐，
也被评过宾夕法尼亚州立大学终身教授



推荐序

当我翻开《大话计算机科学：生活中的计算思维》这本书时，有一种如释重负的心情——这是我一直想做而一直没时间做的事。作为一名长期致力于计算机科学教育的学者，看到这本书的问世，我心中充满了难以抑制的喜悦和激动。这是一本真正意义上的计算机科普读物，尽管此类科普读物不少，但是在不同程度上都难以摆脱“专业味道”，很难用“大白话”解释计算机科学的概念与术语。另外，现有的科普读物往往局限于某一具体技术领域，或通过漫画等形式呈现，但未能全面涵盖计算机科学的重要知识点。还有，我们多见的中文科普读物多半是由外国作者撰写再翻译的，结合中国本土事例的原创计算机科普读物乏善可陈。《大话计算机科学：生活中的计算思维》巧妙地将计算机科学与我们的日常生活紧密结合，真正实现了知识与现实的无缝对接。

在我 38 年的教育生涯中，有幸培养了许多非常优秀的学生，而商静波无疑是其中最出色的一位。他不仅拥有扎实的计算机专业知识，更重要的是，他还具备全面发展的卓越能力。商静波不仅在学术上表现卓越，还拥有将复杂的计算机知识化繁为简的能力，他把这些知识融入日常生活，与非计算机专业的人群对话和沟通。更难能可贵的是，商静波愿意把他的知识、心得与更多的人分享。商静波从上海交通大学 ACM 班本科毕业后，赴美深造，他在博士毕业后获得了在加州大学圣迭戈分校计算机系和数据科学学院的助理教授职位；2024 年 5 月，被评为了加州大学的终身教授（Tenure），同时晋升为副教授。在商静波加入加州



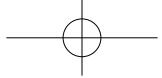
大学的短短 4 年多的时间里，他便获得这一瞩目的成就，其背后的付出、辛劳和不易可想而知。尽管如此，他不只是埋头专注于自己的学术领域和关心自身的得失，还无私地抽出了大量的个人时间来创作本书，以期鼓励下一代思考和学习，这令我深深感动和自豪——他不仅是一位学术领域的翘楚，更是一位有着教育情怀和心怀大爱的知识传播者。

《大话计算机科学：生活中的计算思维》写得非常精彩，书中巧妙地将计算机科学的复杂概念与生活中的平凡小事相结合，娓娓道来，令人耳目一新。我相信，无论你是计算机领域的初学者，还是对这个学科充满好奇的读者，都能在书中找到乐趣和启发。

我认为，好的科普读物，扎实的“科学知识”是基础，巧妙的“常理解释”是关键。《大话计算机科学：生活中的计算思维》通过实际生活中的例子开篇，有很强的趣味性，能够激发读者的兴趣和好奇心；同时，本书的概念讲解和实际应用结合紧密，非常有助于读者理解。逛超市、就医这些平常事在本书里化作一个个生动的例子，增强了本书的可读性，也拉近了读者和计算机科学的距离。这本书的创作团队不仅有专业人员，还有非专业人员，这也正是本书和其他科普读物不同的地方——商静波对计算机科学的核心理念讲解到位，赵馨的润色使其更加平易近人，费赛尔的插画细致传神，三者相得益彰，让枯燥的技术概念变得既通俗易懂又生动活泼。这种完美的融合不仅让读者从中获得知识，更为读者提供了一场视觉与心灵的盛宴。

此外，商静波在本书中不仅仅传递了计算机的专业知识，他更鼓励大家理解书中倡导的计算机科学思维（简称计算思维）。计算思维是一种非常重要的思维模式，它不仅能全方位提高人们的学习、工作和生活效率，还能帮助人们在面对复杂问题时，找到科学、合理的解决方案。这种思维方式也正是我们在当今快速变化的社会中所需要的。这本书不仅能带领读者探索计算机科学知识的奥秘，更能引导读者去拓展思维的边界和培养创新解决问题的能力。

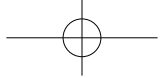
总之，这本书是难得一见的佳作，它如一座桥梁，将日常生活连接于科学的高峰，又将抽象的专业领域连接于每一个人的生活。毫无疑问，这本书值得每一



个对计算机科学感兴趣的人仔细阅读，也强烈地推荐给家有适龄孩子的家长，你们可以与孩子一起欣赏阅读。希望这本书能成为广大读者通向计算机科学世界的一扇门，能够让更多人喜爱和学习计算机科学知识！

俞勇

国家高层次人才特殊支持计划教学名师，
CCF（中国计算机学会）杰出教育奖获得者，
上海交通大学教授、ACM 班创始人



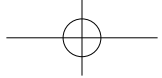
前言

计算机科学，这个看似复杂、高大上的学科，其实已经几十年如一日地、润物细无声地影响了人们的日常生活。从硬件的角度来看，个人计算机、智能手机、可穿戴设备，已经深入生活；从软件的角度来看，搜索引擎、推荐系统、人工智能等，已经无处不在。作为当下最火热的学科之一，计算机科学其实就在每个人的身边。

很多人都想学习计算机科学，但是不知道从何学起。因为计算机科学是建立在数学之上的科学，其理论本身对数学和逻辑的要求较高，所以计算机科学是一门门槛较高的学科，因而非专业读者对计算机科学相关的书籍有畏难情绪也就不难理解了。

本书作为计算机科学的科普图书，旨在通过日常生活里的寻常事例，将看似晦涩深奥的计算机科学的核心理念和知识鲜活地呈现在读者面前，为非专业但是对计算机科学感兴趣的读者提供一个独特而生动的学习途径。本书没有冗长复杂的理论、公式，取而代之的是中国本土广为大众所熟知的事例。本书适合对计算机科学感兴趣的广大读者阅读。

本书的内容可以解构为“科”和“普”两部分，由商静波和赵馨负责文字部分，由费赛尔负责配图工作。下面由商静波和赵馨分别从“科”和“普”两个角度向读者介绍本书。



从“科”的角度看本书——商静波

我有幸能很早期地接触、钻研计算机科学：小学五年级暑假初学 Pascal 编程，拿到本市小学生竞赛第一名；初高中参加信息学奥林匹克竞赛，获得亚太地区国际金牌；在上海交通大学 ACM 班完成本科学习，并代表学校参加国际大学生程序设计竞赛，获得世界亚军；在美国伊利诺伊大学厄巴纳香槟分校（UIUC）师从数据挖掘开山鼻祖韩家炜教授，获得博士学位；在美国加州大学圣迭戈分校（UC San Diego）任教，用了 4 年多的时间，成为加州大学历史上最快获得计算机系和数据科学学院终身教授职位的人之一。

根据这一路的观察和总结，我认为在计算机科学的学习过程中，最重要的并不是对于知识点的掌握，而是对一整套思维模式的训练，这种思维模式就是计算机科学思维，简称**计算思维**。

计算思维，是一种将具体问题抽象化，并在抽象层面进行逻辑推理的模式。这种思维模式和以记忆、学习知识点为主的模式有本质的区别。它是一种更高效、简洁的思维模式，非常适合应对高速发展的现代社会——新知识不断涌现，亟须用计算思维加以推理总结、活学活用。计算思维也恰好是 ChatGPT 这类大模型所欠缺的：ChatGPT 所体现出来的智能很大程度上来自对海量知识点的记忆和学习。写作这本书的初衷是让每个人都能接触到计算思维这个概念。通过具体实例，以小见大，在传递知识的同时，更注重思维模式的培养，让读者可以领略到计算机科学家思考问题的方式，从而做到一通百通。阅读本书不仅能让读者对计算机科学产生浓厚的兴趣，还能帮助读者更深刻地学习计算机专业知识、梳理体系。我期待和鼓励读者在读完本书后，通过计算思维来解读、思考日常生活中的小事，从而得到不同的理解。这样的思维方式不仅局限于计算机科学，更是一种通用的从科学角度分析问题和理解问题的精神。

为了实现这个目标，全书通过“我”（即计算机科学家商老师）的视角，把生活中的衣食住行、吃喝玩乐同计算机科学联系起来，将抽象的计算机科学知识同具体的生活事例相结合，具有很强的可读性和趣味性。通过由具体到抽象再到具体的过程，让读者对计算思维有一个直观的体验，从而在阅读后激发思考，逐步接受这种思维模式。

从知识体系的角度来看，全书涵盖了几乎所有计算机科学本科教育中所涉及的经典课程，从大一、大二必修的“程序设计”“数据结构”“算法”课程，逐步进阶到大三、大四选修的“数据科学”“机器学习”“智能系统”“信息安全”“计算机硬件”等课程。各章节也依照本科教育的顺序来编排，从底层设计讲到顶层应用。由于底层设计更偏向抽象的数学和逻辑，而顶层应用本身更贴近生活，因此阅读难度可能会随着章节的深入而逐步降低。这也是计算机系本科生通常感到前两年的学习非常痛苦，但真正到了后两年偏运用阶段反而更自如的原因。

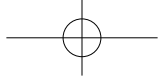
在全书的写作上，我将每个知识点通过独立的故事呈现，因此，即便读者对某一个知识点或者故事理解不够透彻，也不会影响对其他知识点和故事的理解。所以，尽管本书参考了本科计算机教育的系统顺序，读者仍然可以挑选自己感兴趣的部分独立阅读。本书的创作初衷是鼓励读者阅读和思考，尤其是在这个忙碌的时代下，我期待读者即使花了碎片时间来阅读本书，也能够很好地理解计算思维。

考虑不同读者的基础和接受程度，我在一定程度上平衡了知识点的广度和深度。一些章节里的部分知识点可能对于有些读者来说具有一定的挑战性，如“最短路与负环：套餐定价和外汇兑换的约束”（第9章），“梯度下降：驾驶汽车和登山都用到了导数”（第21章），“非对称加密：公开的密钥能加密却不能解密”（第33章），但基于我的写作设计，读者即使跳过这些章节内偏技术的部分内容，对该章节的整体理解也不存在太大影响。

当然，我更鼓励读者能够系统地阅读本书，毫不夸张地说，如果读者能将本书的内容完全融会贯通，就可以打败90%的计算机系本科生了。因此，本书亦可以作为高等院校计算机相关专业的导论课程、非计算机专业的通识课程的教材。

从“普”的角度看本书——赵馨

熟悉我的朋友们应该知晓我是一个数学和逻辑都不太行的人（当初我选择法学院有一大半原因是不用学习数学），他们应该还知道我从小就有创作一本书的心愿。不过，他们大概没想到，我真地创作了一本书，还是和计算机科学、数学和逻辑相关的，我自己也没有想到！大概是2022年的春节时分，商老师和我彼时还是一对新手父母，每天焦头烂额地学习如何更好地照顾新生儿。在这个过程



中，我们第一次接触到了很多写给小朋友甚至小婴儿的科普读物，遗憾的是，大多数这类书籍都是由外国作者结合外国事例撰写的。商老师和我不服气：咱们来写一本中国本土化的、大众化的计算机科普图书！

这是本书诞生的一个契机。

可是，光有雄心壮志还不行，真正开始着手写，发现这是一个艰难的过程，如何结合中国本土化事例来讲解计算机科学呢？因为商老师的工作就是和计算机科学打交道的，所以我们在生活中不可避免地会谈论计算机的很多话题，于是我们开始慢慢回忆生活中我们何时何地会讨论计算机相关的话题——在给小宝宝讲“老和尚给小和尚讲故事”的故事时，我们聊到过计算机科学里的递归；一起看《最强大脑》的时候，我们尝试过用启发式搜索和选手一起解题；用外卖 App 下单时，我们讨论过计算机科学中的最佳匹配和推荐系统；因为自然语言处理的复杂性，我们和朋友约着一起吃饭时闹过“粤菜”（一家餐厅的名字）和“粤菜”（菜系）的笑话。

我们意识到，计算机科学和生活是息息相关的。于是，我们从生活中的这些真实的经历出发，结合计算机科学的专业知识来构建这本书。本书中出现的每一个事例、每一个故事和每一段对话，都是我们生活中真实发生过的，只不过为了增加可读性，我们以商老师和赵律师为主体出发，增加了更多的角色。我们通过计算机科学家商老师的视角，把生活中的点滴和计算机科学联系起来，将抽象的计算机科学知识借由生活事例具象化。我们尽可能地将计算机科学中的很多概念通过日常生活里的寻常事例呈现给读者，期待以一种平易近人的方式为非专业但是对计算机科学感兴趣的读者来讲解这些概念和知识。

由此，我们厘清了本书的大致思路 and 结构，但接下来的另一个难题是如何撰写本书才能激发读者的兴趣并便于读者的理解。换句话说，就是如何平衡本书的科学性和通俗性，这对于我们来说是一个巨大的挑战。这本书里的很多概念和知识，对于商老师来说都是耳熟能详、手到擒来的，但是对于像我这样的门外汉而言，这些概念和知识是复杂且晦涩的。在创作本书的过程中，我总是期待以打比方、举例子的方式将这些概念和知识以大众容易接受的方式呈现；但是从科学性和严谨性的角度，商老师通常无法接受我打的比方或举的例子，因此商老师和我

常常争论不休。我和商老师在一起创作本书之前，几乎从来不吵架，而在创作本书的过程中，我们两个吹胡子瞪眼、拍桌子争吵了不知道多少回。我记得有一次，商老师几乎崩溃地跟我说道：“要不咱不写了，再写下去感觉婚姻都要破裂了。”（现在想想，觉得是很好笑的事。）

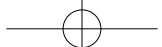
总的来说，这本书由商老师构建主体知识和逻辑，这是商老师的专业领域，也是商老师所长之在；由我来润色思路和语言。我们总说，我既是作者，也是非专业的大众读者，商老师的第一个任务是保证非计算机专业的我能够看懂，这样非计算机专业的读者才能理解。另外，为了便于读者的理解，我们还邀请了中国美术学院毕业的费赛尔为每章配插图。费赛尔有很多年的插画经验。这个过程一般是由商老师对知识点和插图进行讲解和指导，由费赛尔执笔设计和配色。在本书的创作过程中，商老师的第二个任务是保证费赛尔能够理解并以艺术的方式呈现这本书里的知识点。按照这样的节奏，我们期待本书对于非计算机专业的大众读者来说是容易阅读且容易理解的。

因为平时还要工作，照顾孩子和家庭，本书的写作很多时候都是由我和商老师在深夜进行的，我们也没有想到，这样一写便写了两年多的时间。在两年前，我们刚开始决定写这本书的时候，ChatGPT 还仅为学术界所知，而如今，ChatGPT 已经是一个人尽皆知的火热概念，我们也顺应着时代的变化，添加了一个专门讲解 ChatGPT 的章节。

诚然，计算机科学理论本身对数学和逻辑的要求高，但是计算科学并不是在庙堂之上的。我们耳熟能详的“老和尚给小和尚讲故事”的故事，我们手机里下单的外卖 App……其实都是普通人置身其中的计算机科学。数学和逻辑天分并不突出的我，都能够弄明白这本书里提到的计算机知识和概念，我想每一位读者也都可以理解本书的内容。

写在最后

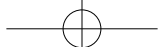
我们非常希望这本书能让更多中国人，尤其是青少年，了解、喜爱计算机科学，让它从一个复杂、高大上的学科变成一种生活中随处可见的计算思维。此外，我们想衷心地感谢王中英编辑、宋亚东编辑在创作和出版过程中给予的鼓励和帮助，还要感谢一直站在身后支持我们的家人。



当然，计算机科学知识犹如一个广袤的宇宙，我们的认知仅仅是冰山一角，又囿于有限的精力和时间，书中错谬之处在所难免，恳请广大读者批评指正，我们将不胜感激。

商静波 赵 馨

2024 年 8 月

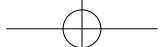


目录

第1篇

生活中的程序设计与数据结构

第1章	递归：老和尚给小和尚讲故事	002
1.1	“老和尚给小和尚讲故事”中的递归	002
1.2	《盗梦空间》中的递归	006
1.3	生活中的递归	008
第2章	二进制：靠掰手指居然能数几万个数	009
2.1	用双手表示 0~99 的任意数字	009
2.2	用双手表示 0~1023 的任意数字	011
2.3	用双手表示更大的数字	013
第3章	循环与上下文切换：怎么更有效地做重复劳动	015
3.1	按说明书组装玩具书，重复劳动多	016
3.2	优化一：同类操作集中做，减少上下文切换	019
3.3	优化二：调整组装顺序，减少额外开销	021
3.4	优化三：两人合作变身“双核 CPU”	022
第4章	二分法与二叉树：图书馆保安应该怎么找到没借过的书	023
4.1	用二分法在有序数组中定位数字	024
4.2	用二分法找没有借过的书，闹出了大乌龙	025
4.3	二叉树可以更好地对借书问题建模	027
4.4	二叉树的层数：需要的检测次数	028

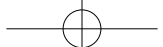


第2篇

第5章 队列与栈：明天该穿什么衣服	031
5.1 把T恤衫组织成栈结构	032
5.2 把T恤衫组织成队列结构	033
5.3 生活中的其他队列结构	035
5.4 生活中的其他栈结构	035

生活中的算法与理论

第6章 蛋糕怎么切才公平：多赢是可能的	038
6.1 切蛋糕是一个博弈论问题	039
6.2 A切B选算法	040
6.3 悬线切蛋糕算法	041
第7章 启发式搜索：《最强大脑》里的那些计算力小游戏	043
7.1 掌握正确算法，你就是最强大脑	043
7.2 启发式搜索解数独	044
7.3 启发式搜索解玲珑棋局	045
第8章 深度优先遍历：迷宫里的右手法则	050
8.1 矩形迷宫中的右手法则	050
8.2 右手法则与深度优先搜索	052
8.3 右手法则与迷宫拓扑结构	054
8.4 右手法则在其他迷宫未必奏效	055
第9章 最短路与负环：套餐定价和外汇兑换的约束	057
9.1 合理的菜单定价与三角形不等式	058
9.2 从菜单到图论	059
9.3 用最短路算法进行合理性检验	061
9.4 最短路算法与货币兑换中的负环	063
第10章 最佳匹配：外卖平台是怎样派单的	066
10.1 外卖派单与二分图	066
10.2 二分图匹配与运力判定问题	067
10.3 二分图最大匹配：匈牙利算法	069
10.4 从最大匹配到最优匹配	070

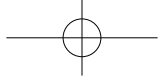


第3篇

第 11 章	旅行商问题：怎样逛超市最省时间	073
11.1	逛超市采购是一个旅行商问题	074
11.2	生活中的其他旅行商问题	075
11.3	旅行商问题极具复杂性	076
11.4	旅行商问题的近似算法	078

生活中的数据科学

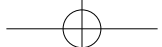
第 12 章	数据标注：验证码里的大生意	082
12.1	训练数据通常需要人工数据标注	082
12.2	先有训练数据，才有人工智能模型	084
12.3	训练数据的质量与数量都很重要	085
12.4	巧妙使用验证码来进行数据标注	086
第 13 章	数据库：抢火车票的背后发生了什么	090
13.1	关系数据库：最经典的数据库	091
13.2	索引：为了更高效的数据库查询	092
13.3	多表单数据库：提高整体查询效率	093
13.4	数据库中的并发与锁：为什么查到有票却买不到	095
13.5	为什么购物平台往往比购票平台更高效	096
第 14 章	大数据：啤酒和尿布为什么要摆放在一起售卖	098
14.1	频繁模式：看两种物品同时出现的频率	099
14.2	关联规则：用户购买 A 了，还有多大概率购买 B	100
14.3	分布式挖掘频繁模式：高效探索关联规则	100
第 15 章	最优化：为什么肯德基、麦当劳总是开在一起	102
15.1	店铺选址的“最优化”目标究竟是什么	103
15.2	生活中的其他最优化问题	105
第 16 章	特征工程：如何区分三个“一模一样”的灯泡	107
16.1	区分浣熊和猫咪的特征构建：特征工程重要且复杂	108
16.2	提取灯泡发热的特征，妙解特征工程的经典面试题	109
16.3	好的特征工程能力处处有用	111



第4篇

生活中的机器学习

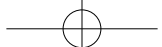
第 17 章	最近邻算法：孟母三迁背后的假设	114
17.1	最近邻算法不需要“学习”	114
17.2	最近邻算法的关键：如何定义相似度	116
17.3	从最近邻到 K -最近邻：综合考虑更稳健	117
第 18 章	支持向量：美国大选基本只需要看摇摆州	120
18.1	得摇摆州者，得美国大选	120
18.2	从二分类的角度看支持向量：寻找最优分界线	121
18.3	远离最大化边界的点对最优分界线没有影响	124
第 19 章	过拟合：高考失误真的是因为心态吗	125
19.1	人类学习与机器学习的类比	126
19.2	高考表现得不好可能是因为过拟合	126
19.3	过拟合产生的原因	129
19.4	如何避免过拟合	131
第 20 章	集成学习：疑难杂症要多看几个专家	133
20.1	集成学习与寻医问诊	133
20.2	Bagging 框架：群策群力、一人一票	135
20.3	Boosting 框架：考虑专家可信度加权平均	137
第 21 章	梯度下降：驾驶汽车和登山都用到了导数	140
21.1	一尺之捶，日取其半，万世不竭	140
21.2	位移、速度、加速度之间的导数关系	141
21.3	梯度下降：利用导数寻找最优解	143
21.4	登山、高尔夫球中的“梯度下降”	144
第 22 章	朴素贝叶斯：为什么我的邮件被识别为垃圾邮件	147
22.1	垃圾邮件中的关键词	147
22.2	关键词分类背后的朴素贝叶斯算法	148
22.3	元数据也是垃圾邮件分类的重要依据	150
22.4	如何避免正常邮件被误分为垃圾邮件	151



第5篇

生活中的智能系统

第 23 章	个体与系统的博弈：外卖骑手的困局	154
23.1	博弈论的概念	154
23.2	外卖平台中的博弈	155
23.3	博弈的初衷是“三赢”	156
23.4	算法的“智能”会打破博弈的平衡	157
23.5	个体之间的团结有利于博弈的平衡	159
第 24 章	搜索引擎：孕妇到底能不能吃螃蟹	161
24.1	早期搜索引擎：看字面相似度	162
24.2	现代搜索引擎：从字面到语义	164
24.3	搜索关键词的设计	166
第 25 章	同温层效应：为什么我的 App 内容越来越同质化	168
25.1	协同过滤：相似的用户喜欢相似的内容	169
25.2	现代个性化推荐系统：用户画像	170
25.3	用户如何与个性化推荐系统更好地磨合	172
25.4	个性化推荐系统的弊端：同温层现象	173
第 26 章	天梯匹配系统：网络游戏如何让玩家欲罢不能	175
26.1	天梯积分是一种 Elo 等级分制度	176
26.2	玩家在天梯匹配系统下的期望胜率通常是 50%	178
26.3	天梯匹配系统也不仅仅是 Elo 等级分制度	180
26.4	如何利用天梯匹配系统科学上分	181
第 27 章	大语言模型：啥是 ChatGPT	182
27.1	什么是 GPT	183
27.2	GPT 以外的其他语言模型	185
27.3	Scaling Law：模型越大，能力越强	185
27.4	ChatGPT 带来的机遇、挑战和风险	188
第 28 章	人脸识别：我的脸解锁了妈妈的手机	190
28.1	人脸识别的常见流程	190
28.2	人脸识别常用的特征	192
28.3	人脸识别的挑战：如何识别伪造的人脸	193



第6篇

第 29 章 自然语言处理：NLP is so hard 194

- 29.1 基于自然语言处理的应用随处可见 195
- 29.2 中文的分词极具挑战性 196
- 29.3 英语的语义也常有歧义 197
- 29.4 歧义性也是一种魅力 198

生活中的信息安全

第 30 章 浏览器：隐私浏览模式到底有多隐私 200

- 30.1 上网冲浪的过程到底能被谁看到 201
- 30.2 隐私浏览模式到底保护了什么隐私 203

第 31 章 区块链：比特币的共识 205

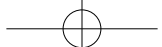
- 31.1 区块链本质上是一个分布式账本 206
- 31.2 区块链的独特之处：去中心化的账本 207
- 31.3 比特币的价值在哪里 209

第 32 章 哈希算法：好网站都不保存用户的明文密码 210

- 32.1 好的哈希函数是一个特级名厨 211
- 32.2 用户识别：比较登录密码的哈希值就够了 212
- 32.3 黑客如何根据哈希值反向破解密码 213
- 32.4 “适量加盐”可以让密码更难破解 214
- 32.5 靠浏览器记住密码靠谱吗 216

第 33 章 非对称加密：公开的密钥能加密却不能解密 217

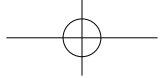
- 33.1 加密已经是上网冲浪的标配了 218
- 33.2 对称加密的风险问题 218
- 33.3 非对称加密的优势 219
- 33.4 图灵奖级别的工作：RSA 算法 220
- 33.5 中间人攻击：非对称加密也没法避免的问题 222



第7篇

生活中的硬件系统

第 34 章	文件系统：我的聊天记录没有了还能找回来吗	226
34.1	文件和文件系统	227
34.2	快捷方式与懒删除	229
第 35 章	内存与存储：为什么手机“内存”比电脑内存还大	231
35.1	计算机存储的分类	232
35.2	外存是一种非易失性存储器	233
35.3	内存是一种易失性存储器	233
35.4	手机广告中提到的“内存”通常不是计算机科学中的“内存”	235
第 36 章	网络：为什么“1000M”的宽带网络没有“1000M”的运行速度	237
36.1	bit 和 Byte 的区别	237
36.2	网速的瓶颈可能让人意想不到	238
36.3	为什么上传速度通常比下载速度慢	239
36.4	网速为什么会随着时间波动	240
36.5	Wi-Fi 里的 2.4GHz 和 5GHz 是什么意思	241
第 37 章	物联网：体感游戏的秘诀	242
37.1	体感游戏如何评分	243
37.2	体感游戏评分难免有误差	244
37.3	记步数到底有多可靠	245
37.4	传感器之间如何联动提高准确度	247
第 38 章	分布式计算：一小时做完年夜饭	249
38.1	锅碗瓢盆都是烹饪过程中的“计算节点”	250
38.2	分布式计算和烹饪共同面对的问题	251
38.3	高效烹饪与高效分布式计算的核心都在于调度协调	252
38.4	计算资源决定了分布式计算的上限	254
38.5	生活中的其他分布式计算	255

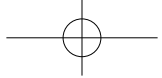


第 19 章 过拟合： 高考失误真的是因为心态吗

又是一年高考季，商老师学校实验室里的小姚同学的家对面就有一个高考考点，考点前聚集着密密麻麻的家长们，在焦急地等待着自己的孩子（如图 19-1 所示）。



图 19-1 高考考点示意图



小姚这天在实验室里跟自己的妈妈视频，妈妈给他看了看家对面的高考考点实况。商老师路过实验室，正好看到挂掉视频的小姚同学心事重重的样子，上前问小姚怎么了，小姚感叹道：“今天是今年高考第一天，刚才我妈妈给我视频看我们家对面的高考考点，我不由自主地想起了我的高考，居然是好几年以前的事儿了，有时候都不想回忆自己的高考，自己心态太差了，导致那年高考发挥失误……”

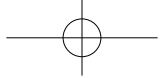
商老师安慰他：“其实未必是你的心态原因，可能是你平时的学习和高考之间产生了‘过拟合’。”

► 19.1 人类学习与机器学习的类比

在机器学习中，**拟合 (Fitting)** 指的是机器学习模型基于给定的数据（即**训练数据**）学习规律的过程。拟合效果是用来评估**机器学习模型**功能的一项重要指标。这个过程还要用到**验证数据集**，主要是检查机器学习模型的训练是否正常，是否需要调整方向。如图 19-2 所示，机器学习和学生学习有极大的共性。例如在典型的人类学习场景中，学生学习的最终目的是建立自身的知识体系和逻辑推理模型，这里我们简称为学生**自身的学习模型**。学生在学校学习的过程，可以被看作自身学习模型拟合的过程，老师通过课堂上的讲解、课后的作业等形式，为学生的学习模型提供**训练数据**，学生需要在这些训练数据的基础上学习和了解这些数据背后的知识和规律。在这个过程中，老师还会提供一些单元小考、周考或者月考来帮助学生查缺补漏，这些阶段性的考试就可以看作**验证数据集**。此外，验证数据集通常和训练数据呈独立同分布 (independent identical distributed, i.i.d.)。代入学习的背景下，用通俗的语言解释，独立同分布就是指平日的讲解和考试的出题是由同一个学校的老师甚至是同一位老师完成的。

► 19.2 高考表现得不好可能是因为过拟合

如何评估一个机器学习模型的拟合效果呢？机器学习中常见的一个合理性检验 (Sanity Check)，是将该机器学习模型运用到训练数据上来检验结果。这就好比老师上完课后，给学生们提供随堂测验，考查课堂上教授的知识点，以此来检验学生们的掌握程度。如果一位学生在随堂测试中表现不佳，老师通常可以认为该学生对该堂课上的知识点掌握得不到位。与之对应，当一个机器学习模型在



训练数据上表现差强人意时，这个模型便存在**欠拟合（underfitting）**。

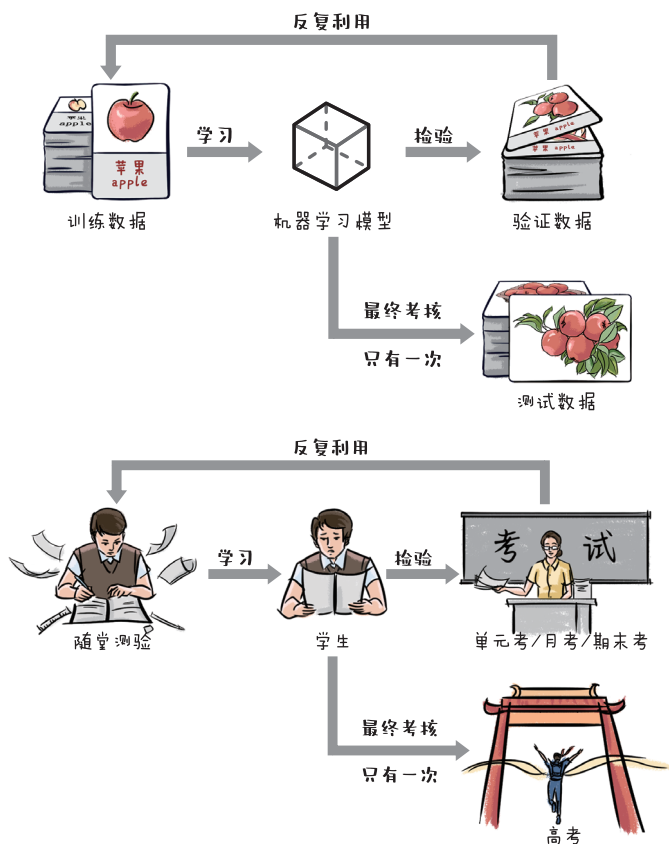
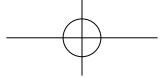


图 19-2 训练数据、验证数据、测试数据和课堂讲解、各种小考、高考的类比关系

在机器学习中，当机器学习模型见过足够多的训练数据并且自身模型足够复杂时，几乎能在所有的训练数据、验证数据上实现近似 100% 的准确率。这就好比如果老师就某堂课的知识点反复讲解并提供多样的针对性习题，学生们反复积累相同题型的解题方法，那么学生们在之后的且由同一位老师出题的随堂测验、单元小考、周考或者月考中的表现会越来越好，甚至可以多次获得满分。

此时，训练数据和验证数据不足以区分机器学习模型的拟合效果，科研人员会引入一个更具挑战的测试——使用新的数据（即**测试数据**，通常和训练数据没有交集）来评估模型的拟合效果。

当机器学习模型见到新的数据时，可以通过之前学习训练数据得到的规律进行运算，并就测试数据给出对应的结果。一个成熟的机器学习模型应当具备一定



的**泛化（Generalization）**能力——在面对新的数据时，应当有良好的表现。机器学习模型的泛化能力主要是通过比较其在训练数据和测试数据上的表现进行评估的。当一个模型在训练数据上表现得很好，但在测试数据或新的评估指标上表现得差强人意时，我们就说这个模型**过拟合（overfitting）**；当一个模型在训练数据和测试数据或新的评估指标上表现得都不错时，我们就说这个模型**拟合得恰到好处（proper fitting）**（如图 19-3 所示）。机器学习中的测试数据类似于学生们面临的高考——高考是根据考试大纲制定的全新的考试，高考考题就是用来检验每个学生的学习模型的测试数据。本质上，高考就是在检查学生对过去十几年学习和积累的知识的调用，是在测试每个学生建立起的学习模型的泛化能力。

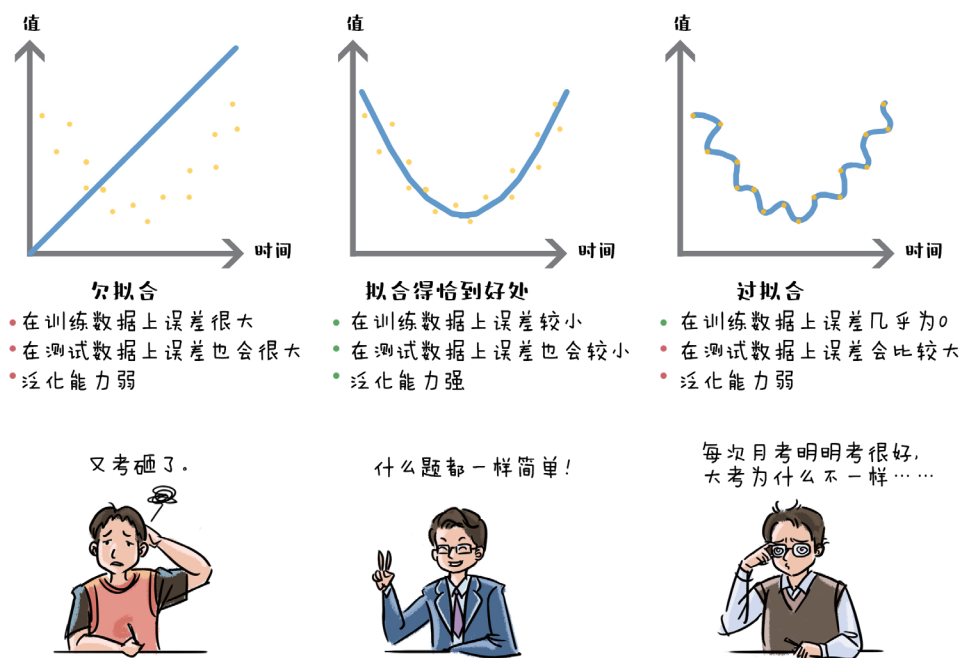
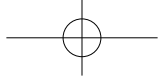


图 19-3 欠拟合、拟合得恰到好处、过拟合

商老师向小姚举例解释：“你的高考经历看起来和过拟合的概念很相似。平时的作业还有校内的周考、月考，其实都是自己学习模型的训练数据和验证数据。所谓高考失误，主要指你在这些训练数据、验证数据上的表现不错，但是在最后的高考，也就是最重要的测试数据上的表现欠佳。从计算机科学的角度而言，这很有可能是你建立的学习模型在老师提供的训练数据上**过拟合**了。你肯定也认识



一些贪玩的学生，平时不认真对待作业以及校内的周考、月考，直接导致了平时的作业和考试表现都不行，最后的高考也获得不了满意的结果，这种情况其实就对应了**欠拟合**的概念。相反，还有些同学平时学习成绩优异，高考发挥也很稳定，这其实就是他们的学习模型**拟合得恰到好处**。”

► 19.3 过拟合产生的原因

那么机器学习中为什么会产生过拟合呢？

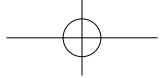
机器学习模型过拟合现象的产生原因主要有以下几种，这些原因也可以很好地与学生的学习和高考的过拟合产生的原因对应起来：

(1) 训练数据不够多。在机器学习中，当训练数据不够多时，机器学习模型会错误地学习到过于简单的规律。这种过于简单的规律对于寥寥无几的训练数据来说可能绰绰有余，但是面对测试数据时就不一定有效了。这就好比一位学生如果平日学习的知识点有限，那么在高考中便会有极大的概率遇到平时没有掌握的知识点，从而影响发挥。

(2) 训练数据不够准确。机器学习最忌讳的一点是训练数据噪声过大。所谓的噪声，是指训练数据本身不够准确，存在一定的误差。机器学习模型会被噪声误导，学习到不准确或者错误的规律。这就好比一个学生平日学习的解题思路方法总是错误的，那么该学生在高考中很难有好的表现。

(3) 机器学习模型过于复杂。在机器学习中，过于复杂的模型甚至能将所有训练数据都记忆下来，因此总能在训练数据上轻易地取得 100% 的正确率。对应到学生的学习和高考情景中，一个极端的例子就是学生在平时的学习中，并没有真正学到任何知识，仅仅是把平时练过的题和解题步骤都死记硬背下来，那么这个学生在高考中就只会平时死记硬背的题型，遇到任何新的题型都无法招架。

(4) 训练数据和测试数据的组成有一定的偏差。机器学习的模型要在测试数据上有效，就需要遵循一个基本的假设，即训练数据和测试数据也是独立同分布的。但在现实生活中，这个假设有时候未必成立。当这个假设被严重违反时，在训练数据上训练得再完美、在验证数据上验证得再有效的机器学习模型，在新的数据面前也会毫无用武之地。就高考而言，高考的命题、阅卷和平时学校考试

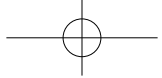


的命题、阅卷的风格未必相同，甚至有时还会迥异。这样的差别在文科考试中尤为明显。如果一位学生平日的学习不懂得举一反三，只拘泥于本校老师的风格，那么在高考中若遇到不同风格的命题，就会容易发生发挥失常的现象。

进一步来看，众所皆知，高考的试题主要分为客观题和主观题两大类。客观题的评分受阅卷人主观因素的影响很小，而主观题的评分虽然是在统一的评分细则下进行的，阅卷人的主观因素仍然会在一定程度上且不可避免地影响到打分。客观题一般情况下会有唯一的标准答案，从计算机科学的角度而言，即评估指标是固定的；而主观题的评分有较大的人为因素的参与，从计算机科学的角度而言，评估指标是因人而异的。

在学生平时建立学习能力模型的过程中，很重要的一部分训练数据来源于老师平日里的讲解、课后作业及考试，因为是由同一位老师或者同一所学校的老师来负责的，练习和考试覆盖的范围有很大的一致性，也就是上文中提到的训练数据和验证数据服从一个分布。学生在这个过程中的学习目标可以简单地描述为对训练数据（课堂的讲解内容）进行学习和理解，从而使得自己在验证数据集上的错误最少（在作业和考试中取得不错的成绩）。

但当训练数据和测试数据不服从同一个分布时，模型是极易发生过拟合的。平时老师上课讲解的重点，会自然而然地成为该老师出题的重点，所以学生只要把该老师强调的重点理解透彻，在完成作业和考试时，便会顺理成章地取得不错的成绩。然而，高考客观题的一个挑战是，该题的题型或者知识点可能是该老师平时教学里没有重点强调过的。这便会导致部分平时成绩不错的同学在高考考试中发挥失常。对于这种情况，用机器学习的语言来讲，其实就是测试数据和训练数据的分布相差过大，导致在训练数据和验证数据上效果很好的机器学习模型，在测试数据中呈现出非常差的效果。这些平时成绩好的同学，很有可能只是对该老师或者该学校强调的知识点掌握得不错，因为平时的考试是由该老师或者该学校的教研组进行命题的，所以这些同学能够呈现出一个不错的考试结果——用机器学习的语言来说，其实这些同学建立的学习模型在训练数据和验证数据上产生了**过拟合**。当这些同学在高考中遇到没有重点学习过的知识点时，无法调用和运用所学的知识，从而导致了高考的“发挥失常”——用机器学习的语言来说，所



谓的“发挥失常”其实是这些同学建立起的学习模型的泛化能力不够。

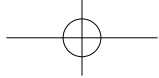
主观题则更容易产生过拟合，因为有阅卷人的主观因素参与评分中。举个小例子：老师们会经常强调卷面整洁的重要性，这是因为一个整洁、美观的卷面是会潜移默化地影响评分老师的印象的。学生平时解答主观题的思路和行文是在本班老师的指导下形成的。通过平日里的作业、随堂考试和月考，学生其实是在不停地探索老师喜爱的答题方式和评分偏好，这便会导致学生和本班老师的评估指标之间产生过拟合。在高考这类大型考试中，阅卷人来自不同的学校，各自拥有各自偏好的思路和行文方式，这些阅卷人的评分偏好其实是一个新的评估标准。在这种新的评估标准下，学生能够获得怎样的分数，取决于学生建立的知识模型的泛化能力，即学生的思路和行文方式等能否在不同的老师的评分偏好下仍然获得好评。这也是为什么在主观性最强的作文中，会存在某位老师非常喜欢的一篇作文却不被另一位老师青睐的现象，也是文科性质的考试比理科性质的考试存在更大的不确定性的原因。

商老师总结道：“因此，人们有时候惋惜的‘高考失误’，很有可能并不是因为某位同学的抗压能力太弱或心态不够稳定，而是该同学平时的学习和高考之间产生了过拟合。该同学平时可能非常好地学习和掌握了本班老师强调的知识点，也能够在校内的校内考试中获得不错的成绩，但是和最终的高考拟合效果欠佳。”

► 19.4 如何避免过拟合

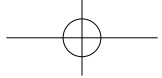
小姚追问：“那有没有什么好方法可以避免这样的过拟合呢？”

商老师回答道：“结合机器学习模型过拟合产生的原因，我们其实不难理解如何降低高考中‘过拟合’的产生风险。从学生的角度来说，学生们要尽可能多地接触不同的、高质量的训练数据，即学生们要尽可能地熟悉每一个知识点，尽可能多地做一些练习题，并且要把平时的学习从本校的范围里扩展出去——多参与和了解一些不同学校、不同区域的考题。这也是一些多校联考、多市联考甚至全省联考的意义所在。从机器学习的角度来说，这些联考就是在让学生们了解和接触不同的数据，通过**数据增强（Data Augmentation）**的手段来提高模型的泛化能力。从任课老师的角度来说，要尽可能地为学生提供准确的训练数据。如果任课老师能够参与高考的命题和改卷，并能够深入研究每一年高考的命题和改



卷，那么将会有第一手的经验来了解高考的规律，从而能够在平日的教学中有的放矢。当然，高考的命题设计和评卷标准也需要充分考虑测量评估上的统一性和科学性。”

“难怪说高考不是一个人的战斗啊！”小姚笑着发出了如此的感叹。



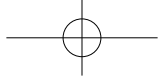
第23章 个体与系统的博弈： 外卖骑手的困局

赵律师有个学妹去做公益律师了，致力于农民工法律援助。学妹及其同事带着一个外卖骑手维权的故事去《一席》做了一场演讲。学妹邀请赵律师和商老师去看演讲，看完演讲后，赵律师和商老师内心五味杂陈。学妹演讲完和赵律师、商老师一起吃夜宵，三人不由自主地又聊起了学妹的演讲，学妹感叹了一句：“还记得《人物》里那篇《外卖骑手，困在系统里》吗？人工智能发展到今天，毫无疑问给人们提供了巨大的便利，可是又束缚了多少像外卖骑手（以下简称‘骑手’）这样的人的生活啊！”

商老师想了一会儿，说道：“其实困住骑手的并不只是平台这个系统，骑手们也被困在了社会这个系统里，在平台、用户和自己博弈的夹缝中生存。”

► 23.1 博弈论的概念

博弈论，英文叫 **Game Theory**。从名字来看，好像是一个打游戏的理论，其实博弈论和游戏确实有一定的相关性。在博弈论中，人们需要定义一个“游戏”和所有参与它的“玩家”，并将各个“玩家”在这个“游戏”中的目标量化。在这个量化的过程中，人们会用到一系列公式，其中最重要的就是**效用函数 (Utility Function)**。效用函数源自微观经济学的概念，是指消费者对商品或者服务的满意度。在博弈论中，效用函数主要用来描述在“游戏”的某一个局面下，“玩家”



达成了多少“游戏”中的目标。在通常情况下，当研究博弈论时，我们往往假设这些“玩家”都足够聪明，且能够基于他们能获取的信息，做出一个“最优”的决策。所以，博弈论也被称为决策论。关于博弈论，也可以参考第6章的介绍。

博弈论的奠基人之一正是现代计算机之父约翰·冯·诺依曼。从计算机科学的角度看，博弈论可以被应用到很多不同的场景，比如说用来指导互联网平台上的搜索广告的拍卖、定价等。

在博弈论中，有一类博弈是发生在多个个体“玩家”和一个系统“玩家”之间的。每个个体和系统之间的目标通常不一样，也就是说，其各自的效用函数不一致。当博弈发生时，个体之间是否有协同作战、系统是否考虑个体之间的差异，将对博弈的结果产生巨大的影响。

► 23.2 外卖平台中的博弈

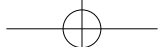
在骑手的例子中，“游戏”的核心其实就是外卖订购和外卖配送，大致可以分为以下几个环节：

- (1) 用户在外卖平台上选定餐厅和具体的菜肴，下单。
- (2) 餐厅接单，估计备餐所需要的时间。
- (3) 平台指派骑手，规划路线，并估计送达时间。
- (4) 骑手按照平台的派单，尽快将外卖送到用户手中。
- (5) 用户在骑手送达订单后，点击“已送达”按钮通知平台。

从博弈论的角度来说，这个“游戏”中其实存在四类主要“玩家”：骑手、用户、餐厅和平台。为了简化问题，此处假设餐厅永远可以准时出餐，无须将餐厅纳入此处的讨论。每一个个体（骑手、用户）都直接参与“游戏”，并与外卖平台这个系统进行交互来优化各自的目标。系统和个体之间的效用函数不同，因此其行为出发点也会各不相同；同时，系统和个体的行为又会进一步相互影响。为了便于理解，如图23-1所示，我们将简化各方的效用函数，将这个框架下各方的目标定义为：

- 用户：希望订单可以按照预计送达时间准时送达，甚至能够提前送达。
- 平台：希望在单位时间内完成更多的订单，保证用户满意度和黏性的同时降低成本、增加平台的收入。





- 骑手：希望在工作时间内配送更多的订单，以增加自己的收入。

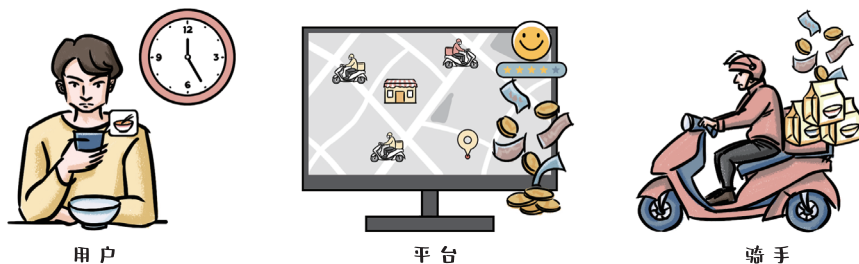


图 23-1 三方的效用函数

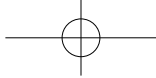
► 23.3 博弈的初衷是 “三赢”

用户、平台和骑手在各自效用函数的驱动下，在这个“外卖游戏”中进行博弈。整个博弈的过程，可以理解为在很长的一段时间内，各方为了优化自己的目标，做出对应的决策，并相互影响各自决策、行为以及收益的过程。

在整个流程中，有非常重要的一环连接着用户、平台和骑手的博弈——**平台是否能准确地估计配送的时间**，包括从骑手当前位置到餐厅所需的时间，以及从餐厅到用户的地址所需的时间。在理想的状态下，配送时间的预估会在平台不断积累数据的过程中被算法不断优化。当这个预估时间能够无比精确时，平台就可以通过一些优化算法来调配骑手，从而提高整体的配送效率，达到三方共赢的局面。

平台在建立之初会怎么制定“外卖游戏”的规则呢？受限于数据量不足的情况，配送时间一般可以简单地通过地图导航来预计。结合地图导航和骑手配送的路线、速度、当前的路况等信息，算法可以预估一个配送时间。平台可能会给这个估计的时间加 10% 左右的缓冲，以保证大部分订单可以准时送达。这样一来，用户不会因为过高的期望而对平台失望。同时，为了鼓励骑手尽快送达，平台还会为配送超时设定惩罚，为提前配送设定奖励。

对于一个这样的“外卖游戏”，骑手又会如何为自己争取利益呢？为了获得更多的奖励、尽量避免惩罚，骑手们自然会发挥主观能动性，开动脑筋去优化自己的配送时间。例如，如图 23-2 所示，在配送的过程中，骑手可能会发现某个小区有一个侧门，从侧门进入小区比从正门进入小区要节省时间，那么骑手接下



来的每次配送都会选择从这个侧门进入，相应地，骑手的配送时间就会比平台预计的配送时间提前。



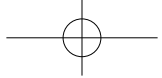
图 23-2 骑手的主观能动性

骑手的行为和平台算法之间是紧密连接的。平台算法会激发骑手的主观能动性，同时，骑手们的行为又为平台算法优化提供了大量的数据支持。这种紧密连接看似构成了一个和谐的生态系统，但其实背后也有巨大的隐患。

► 23.4 算法的“智能”会打破博弈的平衡

平台其实是很聪明的。当不断地有骑手优化自己的配送路线、节省配送时间、获得平台奖励时，平台就会意识到自己现有算法的不足。在逐利的驱动下，平台势必会通过积累的训练数据进一步优化算法。在上文的例子里，当平台发现大多数骑手向某小区配送的时间都比平台预计的配送时间提前时，算法就会缩短预计的配送时间。从数据科学的角度来看，骑手在无形之中给平台提供了海量的配送时间的数据，这些数据将会进一步优化算法模型，并进一步优化平台的决策。反过来，在新的算法模型下，配送该小区的骑手就会渐渐发现，自己原来向该小区配送可以比平台预计的配送时间提前，但是现在只能勉强卡点送达；当遇到特殊情况时，比如小区侧门关闭，配送就会超时。配送时间的压缩，意味着骑手获得奖励的机会越来越少，与此同时，这意味着超时的可能性越来越大，被惩罚的风险越来越高。

在《外卖骑手，困在系统里》的文章里，我们可以了解到，有的算法模型甚至会激进地只考虑取餐点和送餐点之间的直线距离。随着算法模型越来越“智能”，预计配送时间会被进一步压缩，骑手们为了生存，往往会陷入一个无从选



择的困境——他们需要跟算法抢时间。这也是为什么人们看到很多骑手不得已选择逆行、闯红灯来缩短自己的配送时间。为了抢赢算法，骑手们只能被迫降低自身的安全性。

而在骑手和算法抢时间的同时，这些数据又会被反馈到平台。算法模型无从知晓骑手经历过什么，当大多数骑手可以在被压缩的配送时间内完成配送时，算法模型就会误认为模型是准确、可靠的，并且存在可以进一步压缩的空间。

此时，博弈各方之间的平衡就已被打破：①平台通过收集的数据不断地优化预计配送时间，预计配送时间会一次又一次地被压缩。平台期望骑手能尽快完成配送；期待预计的配送时间与骑手实际的配送时间最大化匹配，从而降低平台的成本，减少平台支付的奖励，将收益最大化；②骑手们会逐渐发现自己获得奖励的机会越来越少，甚至如果不逆行、不闯红灯都无法在规定的时间内完成配送。

当然，在这种博弈中，还有一个重要的“玩家”，即平台的用户。有些时候，骑手们的确会在做决策时把用户考虑进来。

商老师问学妹：“你有没有接到过骑手的电话，说自己还有5分钟就能送餐，请您先提前点一个‘已送达’，这样自己就不用被平台惩罚了？”

学妹点点头。

此时，用户是作为一个重要因素参与骑手的这个决策的。大多数的用户此时都能理解骑手并配合骑手的要求。在这个新的决策中，骑手在当前的订单里会获得额外的收益——避免了超时送达的惩罚，还有可能会获得提前送达的奖励，如图 23-3 所示。

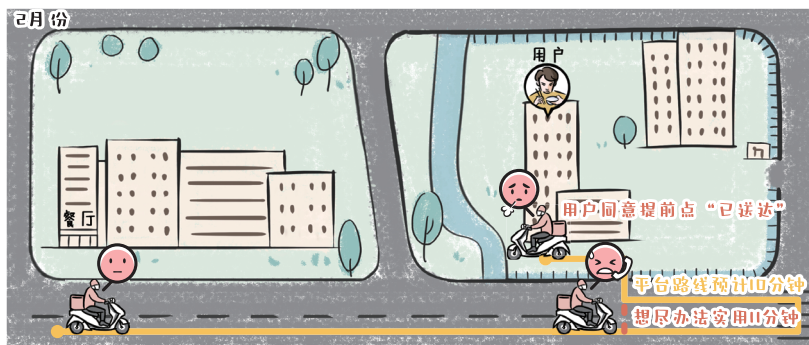
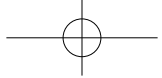


图 23-3 平台算法从骑手的行为中学习新的路线，优化了预计到达时间。骑手想尽办法、用户“帮助”骑手，最终“按时”送达



博弈中的每个“玩家”都会对博弈的结果产生影响。用户的这一行为，从眼前的角度来看是帮助了骑手，但从长远的角度而言，未必如此。在用户的帮助下，骑手原本无法按时送达的订单在预估的时间内送达了，这种“超前”的送达会让平台和算法产生一种错觉。如图 23-4 所示，随着平台慢慢积累和收集这类过于乐观的配送时间数据，平台毫无疑问地会进一步优化算法模型，进一步压缩预计的配送时间，导致按时配送会成为一个越来越不可能完成的任务。这样发展下去，博弈将会越来越不平衡，最终形成一个三输的局面：①平台为用户呈现了过于乐观的预计配送时间，用户满意率下降；②骑手疲于奔命，危险倍增，却无法准时配送，也无法保证自己的收益；③用户永远无法准时收到外卖。

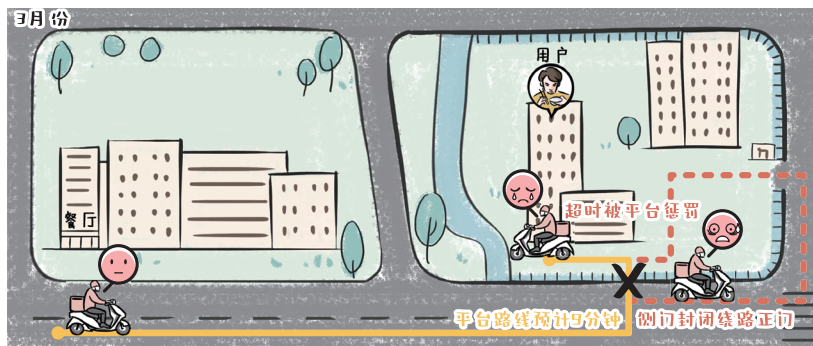


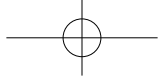
图 23-4 随着“按时”送达的数据越来越多，平台的估计越来越激进，骑手却越来越疲惫。
按时配送变成了一个不可能完成的任务

► 23.5 个体之间的团结有利于博弈的平衡

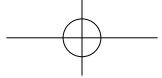
博弈时个体之间是否有协同作战也会在很大程度上影响博弈的结果。在理想状态下，如果所有的骑手都能按照力所能及的速度配送，那么平台能积累到更可靠的预计配送时间数据，博弈的结果自然也会不同。

学妹听完思考片刻后说：“可是在这场博弈中，各方的初始地位就是不平衡的。骑手迫于生计，几乎没有任何议价能力；平台为了逐利，可以居高临下地把风险和成本都最大化地转嫁给骑手。骑手的困局让人非常失落，也让我忍不住去怀疑算法和人工智能究竟是在帮助人类还是在束缚人类。”

商老师安慰道：“没错，尤其在这个博弈中，规则在很大程度上是由平台这个‘玩家’设定的。这也是为什么社会不仅仅需要像你这样的公益律师来发声，



也需要计算机科学家来研究博弈论。我们研究、学习博弈论的终极目标应该是辅助人类，而不是制造对立和不平等。当我们深入地了解了博弈论的原理后，我们不仅可以运用博弈论来模拟和预见各方博弈后可能的结果，还可以为‘系统’设计一套更合理的效用函数，使得博弈中的‘系统’和‘个体’能够处于一种相对‘势均力敌’的地位，达到一种相对平衡的博弈状态。”



第 32 章 哈希算法： 好网站都不保存用户的明文密码

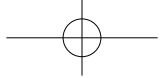
“英伟达的员工和客户密码泄露了！”赵律师下班路上跟商老师讲大新闻的时候，邻居家的奶奶正好听到了。奶奶感叹道：“真是可怕呀，这就是我不能上网就不上网的原因。要不我在一个网站上的账号密码被泄露了，用相同密码的其他网站就都跟着遭殃了。”

商老师笑了笑说道：“奶奶，其实也没有那么可怕，我和您聊聊密码安全的事儿吧。”

密码安全至关重要，尤其在高度信息化的今天。人们的日常生活几乎离不开各种网站和手机应用。在使用这些网站和手机应用时，用户都需要输入用户名和密码。这些用户名和密码的背后，涉及人们的邮件、银行账户和社交平台等。毫无疑问，密码安全关乎人们的信息安全、财产安全和隐私安全等。

在大众的认知里可能存在一个误区——人们普遍误认为当发生黑客攻击、密码被泄露时，如果用户的密码是 123456，则密码 123456 就直接被泄漏了。事实上，这样危险的情况只有在密码被明文储存的时候才会发生。

明文密码存储是指在密码的存储和传输时，密码本身没有被加密而以纯文本的方式被存储：如用户的密码是 123456，系统存储的密码为 123456。显而易见，明文密码存储是具有极高风险的。在发生密码泄露的时候，明文密码存储没有任



何额外的安全保护。

正是因为这种高风险性，越来越多的网站和应用程序从用户安全的角度出发，都对用户的密码进行了**加密存储**。大多数网站和应用程序是通过**哈希算法**来实现这种加密的。

► 32.1 好的哈希函数是一个特级名厨

哈希算法里的“哈希”是英语 Hash 的音译。如图 32-1 所示，Hash 这个词本来的意思是在烹饪的过程中，人们把一些原料切碎，混合起来做成一道菜。在享用菜品的过程中，人们不再关心原材料被切碎前的样子，只关心烹饪后它们呈现出来的菜品和味道。这就好比，当提到青椒肉丝时，人们不会关心肉被切成肉丝之前的样子；当提起土豆泥时，人们也不会关心土豆成为土豆泥之前的样子。经过复杂的烹饪过程后，呈现出来的菜品也很难再可逆到最初原材料的状态。

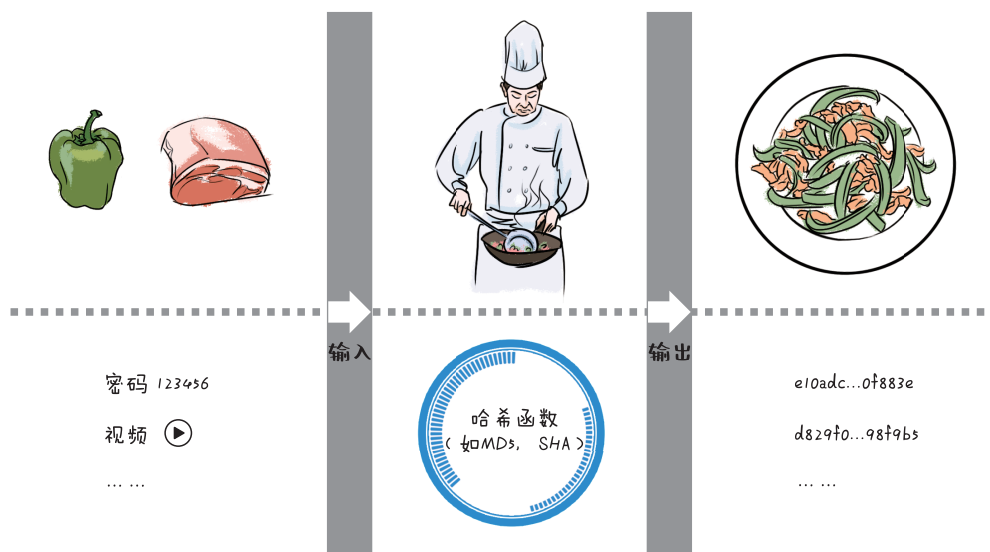
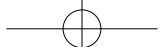


图 32-1 烹饪与哈希的类比

哈希一词非常形象地解释了哈希算法。在计算机科学中，哈希算法的核心思想就是将一个**输入数据**（类似于烹饪中原材料的品种、数量、形状等非常具体的内容）通过一个相对复杂、不可逆的**哈希函数**（类似于烹饪的过程）映射到一个新的值。在计算机中，哈希函数的输入数据既可以是由一串数字、字母、标点符号组成的密码，也可以是一首歌的 MP3 文件，还可以是一部电影的 MP4 文件等。



哈希函数对输入数据进行映射后的值，被称为**哈希值**（类似于烹饪后呈现的菜品）。哈希值是一个固定位数的数字，通常是通过 16 进制的形式呈现的。在计算机科学中，为了方便显示十六进制数字，通常用数字 0 ~ 9 和字母 a ~ f，一起表示十进制中的 0 ~ 15。

在通常情况下，可以作为哈希函数输入的数据远比可能的哈希值多。因此，理论上一定存在两个及以上不同的输入数据被哈希函数映射到同一个哈希值。当这种情况发生时，我们就说哈希发生了“冲突”。这就好比在烹饪中，不同形状或类别的土豆最后都能做出口味几乎相同的土豆泥。

常见的哈希函数有 MD5、SHA1、SHA2 等。这些函数的背后有一系列数论的理论支撑，以确保“冲突”发生的概率很低。这些常见的哈希函数就像特级名厨一样，总能通过一些巧妙的烹饪技巧，把食物原材料上的细节尽可能地在菜品中呈现出来。

► 32.2 用户识别：比较登录密码的哈希值就够了

有了这些常见哈希函数的强力支持，如图 32-2 所示，网站和应用程序便可以通过比较用户本次登录输入的密码的哈希值和用户创建账户时设置的密码的哈希值是否相等，来进行用户识别。从不太严谨的角度来说，这个过程和人们在数学考试中使用的特殊值代入“蒙”题法是类似的：在验证两个公式 $f(x)$ 和 $g(x)$ 是否等价（这里的等价指的是可以通过化简、等价变换等操作，将两个公式相互转换）时，我们可以通过代入几个特殊的便于计算的 x 值，分别根据 $f(x)$ 和 $g(x)$ 两个公式计算出具体值；如果多次代入的值均相等，那么就可以大胆地猜测这两

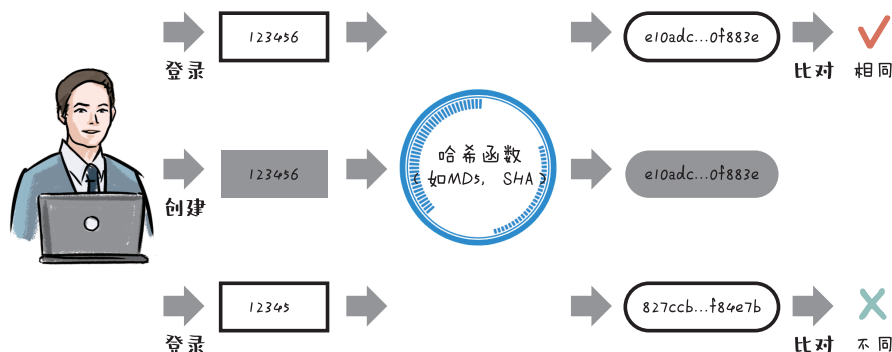
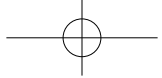


图 32-2 创建账户时的密码和登录时的密码的哈希值比较



个函数是等价的。哈希值就可以理解为 $f(x)$ 和 $g(x)$ 经特殊值 x 代入后的具体值。如果本次登录输入的密码的哈希值和创建账户时设置的密码的哈希值完全一致，则网站和应用程序会允许用户登录；反之，则拒绝登录。这样一来，网站和应用程序只需要存储用户密码对应的哈希值，而非用户的明文密码。

虽然常见的哈希函数的计算公式是公开的，但这些函数是非常复杂且不可逆的。因此，即使黑客知晓了这些哈希函数的计算公式，也无法通过对哈希值进行“逆运算”推算密码原文。但为什么用户的密码被泄露和破解的情况仍然屡见不鲜呢？要回答这一问题，就不得不讨论“彩虹表”。

► 32.3 黑客如何根据哈希值反向破解密码

通俗而言，**彩虹表**可以理解为一个常见密码（类似于常见的原材料）与其常见哈希值（类似于家常菜）之间的对应关系表。彩虹表可以被看作一个两列的 Excel 表格：其中的一列是常见密码，另一列是这个密码可能对应的哈希值。具体来说，如图 32-3 所示，黑客会维护一个密码池（对应彩虹表中的第一列），对密码池里的明文密码用常用的哈希函数进行计算得到相应的哈希值（对应彩虹表中的第二列）。这些明文密码和哈希值存储在一起就成了一个彩虹表。黑客的密码池通常包含一些常见的密码（比如 123456）、一些较短的密码（比如 8 位以内的纯数字和小写字母的密码）、一些常见拼音和单词的组合（比如 password）。当用户使用了这些密码时，黑客便可以通过暴力枚举和查寻彩虹表得到用户的明文密码。当然，不同黑客维护的彩虹表会不尽相同。

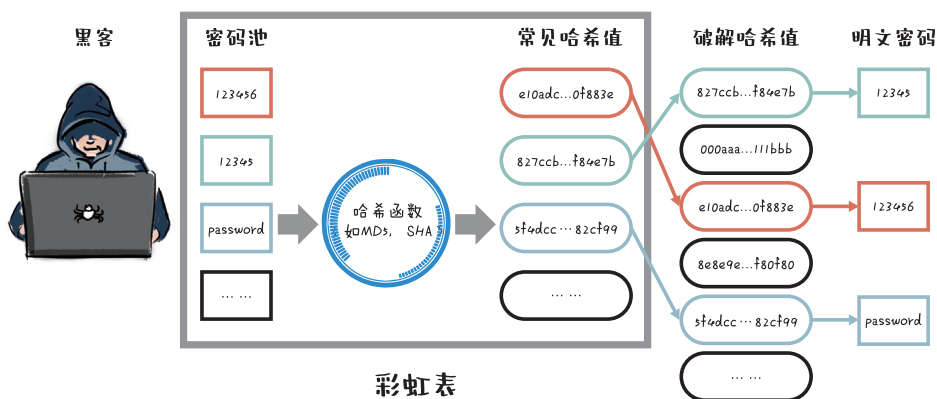
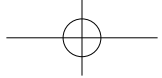


图 32-3 彩虹表和破解示意图



当黑客破解了某家网站或者应用程序存储的登录信息时，如果该网站和应用程序加密存储了用户的密码，则破解后，黑客能够获取的仅是哈希值，而非用户的明文密码。此时，黑客要想破解用户的明文密码，就需要运用到自己维护的彩虹表。如果黑客获取的哈希值已经在自己维护的彩虹表中，那么按图索骥，便可获取该哈希值下的明文密码。

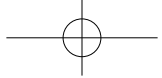
所以，大部分的黑客攻击并没有那么神奇，他们能破解的明文密码往往局限于自身维护的彩虹表的范围。由于计算、存储能力的限制，任何黑客的密码池及其衍生出的彩虹表都不可能太大。这也是为什么大家都应该使用强力的密码，即 8 位以上、至少包含各一个大小写字母、数字和符号等。越复杂的密码存在于黑客的彩虹表中的可能性越小，被破解的可能性越小，则使用该密码的账户安全性越高。

那么，是否用了强力密码后，用户就可以在不同的网站都使用相同的密码而高枕无忧了呢？并不是的，如果 A 网站明文储存了用户的强力密码，且被黑客入侵，黑客会将该明文密码加入其密码池并计算生成对应的哈希值。如果用户碰巧在 B 网站上使用了和 A 网站相同的强力密码，或者用户在 B 网站的强力密码碰巧和 A 网站的某个其他用户的密码相同（当然这个可能性很小），那么当黑客破解 B 网站得到 B 网站上的密码哈希值后，就会发现这个强力密码的哈希值和黑客曾经破解获取的 A 网站的某个明文密码的哈希值符合，因此也就无法“独善其身”。

当然，这个潜在的风险可以通过用户针对不同的网站和应用程序使用不同的强力密码来改善。但是强力密码的记忆难度较大，用户很难同时记住很多个没有强联系的强力密码。

► 32.4 “适量加盐”可以让密码更难破解

那么，网站和应用程序的开发者有什么可以帮助用户的吗？一个现有且很成熟的方法是通过开发者给用户的密码“加盐”，然后再计算哈希值来实现。“加盐”其实是从其英文名“salting”直译过来的。不得不说，这个直译非常有味道，跟人们烹饪中的加盐有异曲同工之妙。如图 32-4 所示，在烹饪过程中，厨师可以根据自己的喜好给菜肴加不同分量、不同种类的盐（比如海盐、粉盐、椒盐等），



从而做出自己“独门”的定制菜。在对明文密码计算哈希值时，网站和应用程序的开发者可以根据自己的喜好，在常见的哈希函数的基础上为哈希函数的计算过程做一些“独门”的处理，再计算哈希值。这里的“独门”，是指只有该网站和应用程序的开发者才知道的一些小众的计算公式。这些小众的计算公式不一定具有数论的理论支撑，但是在该网站和应用程序的开发者群体里能够被理解、被交流。这就好似中式烹饪里加盐的“适量”，除了烹饪者自己，其他人很难掌控“适量”的尺度。因此，“适量”的程度也成了某种菜品美味的秘诀。

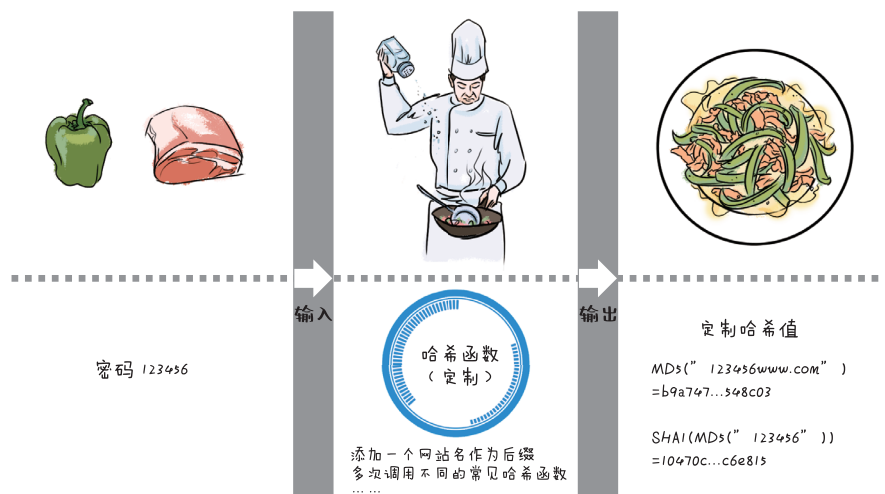
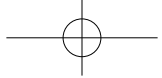


图 32-4 加盐的示意图

如此一来，哈希值的计算将跳出常用哈希函数的范畴，使得黑客维护的“彩虹表”的效力大大减弱，黑客们无法轻易地根据先验的哈希值获取明文密码。开发者“加盐”的过程可谓八仙过海各显神通，在这里举两个例子便于大家理解，如图 32-4 所示。例如，开发者可以给用户的明文密码添加一个类似于网站名称的字符串作为后缀，如明文密码是“123456”，A 网站的开发者可以先把密码转换成“123456A 网站”，再对其进行常见哈希函数的计算。这样得到的哈希值便会和直接计算“123456”得到的哈希值不同。再例如，开发者也可以多次调用不同的常见哈希函数来计算哈希值，如首先计算明文密码的 MD5 值，再对这个 MD5 值进行 SHA1 哈希值计算。

整个“加盐”处理的过程会将哈希值“定制化”——同样的密码在不同的



网站和应用程序中的哈希值会变得不同。即使用户在 A 网站和 B 网站使用同样的密码，密码映射出来的哈希值也不尽相同。这样一来，即便黑客通过 A 网站知晓该用户的明文密码，且获取了 B 网站存储的该用户密码的哈希值，也无法轻易发现 A 网站的明文密码被同样使用在了 B 网站上。

所以说，好的网站和应用程序的开发者一定会合理地运用哈希算法来保护用户的账户安全。即使在网站和应用程序被黑客攻击、数据库被泄露的情况下，用户的明文密码也无法被简单地获取。

奶奶听完之后，开心地感叹道：“那我真是放心多了！”

► 32.5 靠浏览器记住密码靠谱吗

赵律师问道：“我还有个问题，我平时上网的时候，浏览器会问我要不要记住密码，这个会不会有风险呢？有时浏览器还会提醒我密码有被泄漏的风险，那是不是意味着浏览器明文存储了我的密码呢？”

商老师解释道：“我们以 Google Chrome 为例，来简单解释浏览器存储密码背后的原理吧。” Chrome 浏览器主要是依托于计算机本地的操作系统中的加密文件存储功能，以加密的形式存储密码的。所以，当我们想调取明文密码的时候，Chrome 会要求我们输入电脑操作系统的密码。同时，Chrome 浏览器会在计算机本地使用常见哈希函数计算并保存密码的哈希值，然后通过和已知泄漏的密码的哈希值进行比较，判断是否相同，进而发现监控潜在的泄漏风险。总的来说，在浏览器中保存密码是相对安全的，但与此同时，我们需要设置一个强力的密码来保护计算机操作系统。

赵律师听后搂着邻居奶奶说：“奶奶，你之前不是一直担心记不住密码嘛，这下不仅可以放心地上网，还可以放心地让浏览器帮你记住密码呢！”