

专业技术人员继续教育培训用书

我们的

大数据时代

The Era of
Big Data

李广建 ◎ 主 编

化柏林 ◎ 副主编



中国人事出版社

专业技术人员继续教育培训用书

我们的

大数据时代

The Era of
Big D

李广建 ○ 主 编

化柏林 ○ 副主编



中国人民大学出版社

图书在版编目(CIP)数据

我们的大数据时代/李广建主编. —北京: 中国人事出版社, 2015
专业技术人员继续教育培训用书
ISBN 978-7-5129-0893-2

I. ①我… II. ①李… III. ①互联网络-数据采集-继续教育-教学参
考资料 IV. ①TP274

中国版本图书馆 CIP 数据核字(2015)第 085578 号

中国人事出版社出版发行

(北京市惠新东街 1 号 邮政编码: 100029)

*

保定市中国画美凯印刷有限公司印刷装订 新华书店经销

787 毫米×960 毫米 16 开本 12 印张 200 千字

2015 年 5 月第 1 版 2015 年 5 月第 3 次印刷

定价: 26.00 元

读者服务部电话: (010) 64929211/64921644/84643933

发行部电话: (010) 64961894

出版社网址: <http://www.class.com.cn>

版权专有 侵权必究

如有印装差错, 请与本社联系调换: (010) 80497374

我社将与版权执法机关配合, 大力打击盗印、销售和使用盗版
图书活动, 敬请广大读者协助举报, 经查实将给予举报者奖励。

举报电话: (010) 64954652

编委会名单



主 编：李广建

副主编：化柏林

编 委（按姓氏拼音为序）：

江信昱 王昊贤 王晓笛 祝振媛

人类社会已经进入了大数据时代。大数据是继云计算、物联网之后 IT 产业的又一次颠覆性的技术革命。其对数据价值的分析和应用，正在逐步地改变人类世界。无论是科研学术界，还是工商企业界，无论是政府组织管理，还是个人日常生活，随处可见大数据的影子。大数据给时代带来的改变也无处不在，从精准营销到战略定位，从课程开设到人才就业，从桌面办公到移动互联网，从产业升级到社会变革，从网络安全到国家安全，从社会治理到国家战略，大数据不仅改变了人们工作、生活和学习的方式，也改变了人们思维与决策的方式。可以说，大数据不仅仅是一项技术，更是衍生成为一种社会现象。

大数据时代，谁掌握了数据并实现了数据的价值，谁就将在竞争中胜出。为应对当前激烈的竞争环境，并在国际竞争与合作中占据有利地位，我国政府高度重视大数据的应用价值与产业发展。《中共中央关于全面深化改革若干重大问题的决定》指出，“建立全社会房产、信用等基础数据统一平台，推进部门信息共享”，“充分利用信息化手段，促进优质医疗资源纵向流动”。《“十二五”国家战略性新兴产业发展规划》将“智能海量数据处理相关软件研发和产业化”（大数据技术）列为重点发展技术方向之一。除了中央政府的高度重视和大力推进，北京、上海、贵州、广东、武汉、济南等地均推出了针对大数据产业发展的相关规划或举措，大数据产业快速发展的政策环境已经成型。

专业技术人员作为我国人才队伍的主体和骨干力量，是我国科技、教育、文化、卫生、体育等专业化公共服务的主要承担

者，是推动科技创新、管理创新、文化创新的主力军。在以颠覆与创新为主导的大数据时代，专业技术人员不只是数据的见证者，还是数据的生产者、加工者和消费者。如何在大数据时代培养自身的数据基因和数据思想并整合各种分析方法，对复杂现象及其关系做出审慎判断，将是现代社会中专业技术人员必备的个人修养和生存技能。

为帮助广大专业技术人员全面系统地掌握大数据相关知识和技能，提升综合能力和素质，加快推进专业技术人才知识更新工程实施，我们编写了《我们的大数据时代》这样一部教材。本书以基础性与系统性为导向，对当前大数据的概念进行了辨析，讨论了大数据的机遇与挑战，介绍了大数据的技术，论述了大数据的管理，探讨了大数据的安全与应对，归纳了大数据的应用价值，全面展现了大数据的应用现状和未来发展方向。考虑到专业技术人才队伍特征和工作实际，本书在编写过程中力求通俗，尽量避免艰深晦涩，并在每章章后设计了思考题，内含情景分析题，以供读者参考。希望读者能够借助本书初步掌握大数据的基本理论、方法、技术与应用等方面的基础知识。

本书共分为六章，具体分工如下：

第一章 大数据时代 执笔人：化柏林 李广建

第二章 大数据的机遇与挑战 执笔人：李广建 化柏林

第三章 大数据的技术 执笔人：王昊贤 王晓笛

第四章 大数据的管理 执笔人：祝振媛 江信昱

第五章 大数据的安全 执笔人：王晓笛 王昊贤

第六章 大数据的应用 执笔人：江信昱 祝振媛

全书由李广建统筹策划，确定全书主体内容、逻辑框架与写作风格等，化柏林、江信昱在编写过程中对各章的内容提出了重

要的修改意见并提供了大量的资料，李广建对全书进行了修改审定。

中国人事出版社的刘明波同志负责本书的编辑工作，他对本书的定位、内容布局等一系列关键问题给出了非常中肯的意见与建议，在此表达我们的谢意。

在写作过程中，我们参考和借鉴了大量的中外文资料。由于篇幅所限或工作疏忽，本书未能一一列出全部的参考文献，在此，对本书具名和未具名的参考文献的作者表示衷心的感谢。

大数据是最近几年出现的新生事物，我们对她的了解也是初步的。本书的各位执笔人虽然努力，但限于能力和水平，加之时间有限，缺点不足与错误疏漏之处在所难免，恳请各位专家和读者批评指正。

李广建

2015年4月于北京大学

目 录

Contents

第一章 大数据时代	1
第一节 大数据概述	1
第二节 大数据的价值	10
第三节 大数据各国行动	19
本章思考题	25
第二章 大数据的机遇与挑战	26
第一节 大数据的公众分歧	26
第二节 大数据的商业模式	32
第三节 大数据的企业应对	36
第四节 大数据的专业技术人员应对	40
第五节 大数据的发展趋势	42
本章思考题	49
第三章 大数据的技术	51
第一节 大数据的采集技术	51
第二节 大数据的存储技术	59
第三节 大数据的预处理技术	66
第四节 大数据的挖掘技术	71
本章思考题	84
第四章 大数据的管理	86
第一节 大数据的生命周期管理	86
第二节 数据的再利用	89

第三节 数据的重组与扩展	93
第四节 数据的估值与折旧	103
第五节 数据废气	110
本章思考题	112

第五章 大数据的安全 114

第一节 大数据安全概述	114
第二节 本地安全	115
第三节 网络安全	121
第四节 隐私安全	128
第五节 技术安全	140
本章思考题	144

第六章 大数据的应用 145

第一节 医疗卫生领域	145
第二节 科研教育领域	152
第三节 经济管理领域	163
第四节 社会服务领域	169
第五节 其他领域的大数据应用现状	173
本章思考题	175

参考文献	177
------------	-----

第一章 大数据时代

本章导读

本章主要介绍大数据的基础知识。通过分析大数据时代来临的历史背景，引入大数据的基本概念、特点、分析理念和价值，梳理典型国家的大数据政策与计划，并结合其在当前商业运营和管理中的成功案例，帮助广大专业技术人员理解。

第一节 大数据概述

当今世界已经进入了大数据时代。大数据无处不在，它正在深刻影响人们的工作、生活和学习，并将继续产生更大的影响。专业技术人员身处大数据时代，需要更好地认识大数据、掌握大数据、利用好大数据。

一、大数据产生的背景

（一）大数据产生的技术背景

大数据不是某一天突然产生的。随着互联网信息的急速增长，机器设备信息的实时采集，产生了大量数据，这些数据如何存储、挖掘以及利用成为一个人们必须解决的问题，文本、图片、音频、视频等多媒体信息对

存储技术提出了新的要求，而位置信息、关系信息使得数据种类越来越丰富，其价值挖掘也日益受到人们的重视。大数据的理念和方法正是在这种环境中产生的。

互联网是一种最为突出的大数据环境。在 2000 年前后，互联网网页呈现爆发式增长，到 2000 年年底，全球网页数达到 40 亿个，而且每天以大约 700 万个网页的速度飞速增长。在这种情况下，用户查找信息越来越不方便。为了帮助互联网用户从亿万数据中快速找到所需信息，谷歌（Google）等公司率先建立了覆盖数十亿网页的索引库，开始提供较为精确的搜索服务，大大提升了人们使用互联网的效率，这是大数据应用的起点。当时搜索引擎要存储和处理的数据，不仅数量之大前所未有，而且形式以非结构化数据为主，传统技术已经无法应对。为此，谷歌提出了一套以分布式为特征的全新技术体系，即后来陆续公开的分布式文件系统（GFS, Google File System）、分布式并行计算（Map Reduce）和分布式数据库（Big Table）等技术。通过这些技术，利用较低的成本实现了之前技术无法达到的数据处理规模。这些技术奠定了当前大数据技术的基础，可以认为是大数据技术的起源。

随着互联网的进一步发展，特别是 Web 2.0 发展，万维网之父蒂姆·伯纳斯-李（Tim Berners-Lee）等人在 2007 年发起开放数据运动，将关联数据串联起来形成一个巨大的数据网，从而构建更多的应用与服务，开放数据运动同时也带来了更多的数据。云计算、物联网、分布式并行计算及数据库、社交网络及智能终端等新兴信息技术的发展，也正在不断丰富数据的采集方式。此外，为了解决数据遗失问题，数据存储设备及其功能也不断完善，使得数据的保存更加便捷，这也让数据量变得越来越大，据 IDC 报告估计，2005—2020 年间，全球数据量将增长 300 倍，达到 40 万亿 GB^①。数据的快速增长，引出了更多的数据管理、硬件环境与分析服务等需求。政府、电子商务、互联网、金融、医疗保健等行业的相关组织开始使用多种新兴信息技术不断搜集不同来源的各类数据，以便从中挖掘出更多有价值的信息或知识。对组织来说，数据采集已不是主要障碍，关

① GB (Gigabyte) 表示计算机存储容量。计算机存储容量基本单位是字节 (Byte)，一个字节能够容纳一个英文字符，一个汉字需要两个字节的存储空间。表示计算机存储容量的单位还有 KB、MB、GB、TB、PB、EB、ZB、YB、BB。各单位表示的容量换算公式如下：1 KB = 1 024 Byte；1 MB = 1 024 KB；1 GB = 1 024 MB；1 TB = 1 024 GB；1 PB = 1 024 TB；1 EB = 1 024 PB；1 ZB = 1 024 EB；1 YB = 1 024 ZB；1 BB = 1 024 YB。一张 CD 光盘的容量约为 600 MB。

键在于更完善、更有效的数据分析与挖掘方法，让信息变得更容易理解且利于决策行动。

（二）大数据产生的社会背景

伴随着互联网产业的崛起，新的海量数据处理技术在电子商务、精准广告、智能推荐、社交网络等方面得到了广泛应用，并取得巨大的商业成功。这启发了全社会开始重新审视数据的巨大价值，于是，金融、电信等数据密集型行业开始尝试使用新的理念和技术并取得初步成效。与此同时，业界也在不断对谷歌公司提出的数据处理技术体系进行扩展，使之能在更多的场景下使用。2011年，麦肯锡、世界经济论坛等知名机构对这种数据驱动的创新进行了研究总结，随即在全世界兴起了一股“大数据”研究和应用的热潮。

大数据概念的提出，对人们的生活、思维及工作方式产生了巨大的影响，并将当前的信息化社会推进到了一个新的发展阶段。当前数据急速膨胀，使得传统的系统平台已无法支持大数据处理，现有分析方法也难以从纷繁复杂的大数据中凝练出更多有价值的信息以及提供新的深刻洞察，这就要求对大数据存储技术、处理技术、分析方法、应用服务等方方面面做全新的思考，进而也引出了大数据存储、分析、管理与服务等一系列基于数据链的“大数据产业”。

大数据被多个领域视为下一代信息技术与数据分析管理的热点，影响人们思维与生活、企业运营与管理、国家治理与政府决策等各个方面。对国家决策及政府管理来说，数据资源已成为新时代中的一种战略优势。对科研人员来说，大数据并不是突然出现的新概念或新技术，而是由过去的分布式、数据挖掘等专业术语演变成的广为人知的流行词，这种演变的重要意义在于启发了人们重新审视数据的重要意义及潜在价值。

无论是科研学术界，还是工商企业界；无论是政府组织管理，还是个人日常生活，大数据已渗透到各个学科领域、各行各业、多个层面。可以说，大数据不仅仅是一项技术，更是一种社会现象。如何搜集与构建大数据、存储与管理大数据、分析与挖掘大数据价值，变成为一个新热点、新领域。大数据带来的新机遇与新挑战是前所未有的，值得人们广泛关注。

二、大数据的基本概念

大数据是当前信息化社会发展的热点话题。关于什么是大数据，目前也有多种观点。

较直观的认识是按照数据规模来界定大数据，例如，麦肯锡的咨询报告将 TB 级以上的数据集视为大数据；2014 年国际数据公司预测报告指出，在大数据环境中超过 100TB 的数据集已是常态现象，包括谷歌每天处理的搜索查询、NASA（National Aeronautics and Space Administration，美国国家航空航天局）的天气观测数据存储量、电商每天处理的交易数据等；舍恩伯格在《大数据时代》一书中认为 PB 级以上的数据才称为大数据。

另一种典型观点认为，传统数据库技术与方法无法处理的海量或非结构化的数据集，称之为大数据。维基百科将大数据定义为：无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合。高德纳（Gartner）咨询公司也认为，数据的极端规模、多样性和复杂性已普遍存在于当前环境，同时也是一种颠覆现象，致使当前的数据管理、技术工具或分析方法需要进一步扩展、改造、集成甚至创新。显然，这些观点已经不把数据量作为判别是否是大数据的唯一标准了，而是引入了技术和方法标准，即大数据是无法用传统数据库技术与方法处理的数据集。

以上两种观点都是从大数据的自身条件来定义大数据的，还有一种观点，从功能或流程角度来诠释大数据。这种观点认为大数据是基于多源异构、跨域关联的海量数据分析所产生的决策流程、商业模式、科学范式、生活方式和观念形态上的颠覆性变化的总和。大数据是数据化趋势下的必然产物。数据量的爆炸性增长，不仅带来了各种计算处理数据的新要求，而且带来了互联网时代信息过载以及如何抓住消费者注意力的新问题。数据来源极大丰富，形成了大量非结构化的数据形态，并且数据之间的跨领域关联现象十分普遍。同一个数据，其表现形式可能会不同，表现方式可以是数据库、数据表格、文本、传感数据、音频、视频等多种形式；同一个事实或规律可以同时隐藏在不同的数据形式中，也可能是每一种数据形式分别支持了同一个事实或规律的某一个或几个侧面，这种跨域关联是数据量增大后从量变到质变的飞跃，是大数据巨大价值的基础。服务对象上也从针对全体的服务，变成了针对不同群体，甚至个体的服务。大数据不是数据量的简单刻画，也不是特定算法、技术或商业模式上的发展，而是从数据量、数据形态和数据分析处理方式，到理念和形态上重大变革的总和。

总结以上内容，要全面了解大数据，应该从“数据资源、技术工具、分析应用”三个认识视角来界定大数据。大数据是具有体量大、结构多

样、时效性强等特征的数据；处理大数据需采用新型计算架构和智能算法等新技术；大数据的应用注重相关分析而不是因果分析等新理念，目的在于发现新的知识与洞察并进行科学决策。

三、大数据的特点

（一）大数据的数据特点

IBM (International Business Machine Corporation, 国际商用机器公司) 用 3 个 V 来描述大数据的三个基本特性, 3V 分别是体量 (Volume)、速度 (Velocity) 以及多样性 (Variety)。也有人认为大数据包括三个要素, 即大分析 (Big Analytic)、大带宽 (Big Bandwidth) 以及大内容 (Big Content)。舍恩伯格把大数据总结为 4 个 V, 即数据规模大 (Volume)、数据类型多样 (Variety)、数据处理速度快 (Velocity)、数据价值密度低 (Value)。尽管大数据的定义与说法不尽相同, 但归结起来, 大数据具有如下几个典型特征:

1. 数据规模大

数量巨大是大数据最显著的特征, 且大数据的数据量仍以前所未有的速度持续增加。淘宝网每天的交易达数千万次, 数据产生量超过 50 TB。百度公司每天大约要处理 60 亿次搜索请求, 数据量达几十 PB。一个 8 MB 的摄像头一小时能产生 3.6 GB 数据, 一个城市若安装几十万个交通和安防摄像头, 每月产生的数据量将达几十 PB。根据麦肯锡全球研究院 (MGI, Mckinsey Global Institute) 预测, 到 2020 年, 全球数据使用量预计达到 35 ZB。医疗卫生、地理信息、电子商务、影视娱乐、科学研究等行业, 每天也都产生大量数据。如何处理超大规模的网络数据、过程行为数据、移动数据、射频采集数据、社会计算数据、语音通话数据、多媒体视频数据已经成为科研界和产业界亟待解决的关键问题, 也是大数据要解决的核心问题。

2. 数据类型多样

现在业内有一种观点认为, 数据如果类型单一, 那么即使数据量再大, 也难以称之为大数据。数据来源广泛、类型多样、结构各异是大数据的重要特点。大数据的数据类型不仅包括传统的结构化数据, 而且包括过去用常规软件无法进行深入分析处理的非结构化数据。随着互联网的飞速发展, 各种新型应用不断涌现, 如社交网络、电子商务、位置服务等, 以文本、图形、语音、视频等为代表的非结构化数据的增长速度越来越快,

远远超过了结构化数据的增长速度。未来大数据主要通过对非结构化数据的分析处理来获得有价值的结论。可以说,数据类型多样是大数据非常突出的特点,针对多源数据的分析与处理也必将成为大数据分析与管理的重要趋势之一。

3. 数据处理速度快

大数据环境下,数据的创建、分析和处理的速度在不断加快。目前,数据以传统系统不可企及的速度在传播,在短短的 60 秒内,视频分享网站 YouTube 上的用户会上传长达 48 小时的视频;谷歌会收到 200 万次搜索请求并能迅速反馈结果;社交网站推特(Twitter)可以处理 100 万条微博信息;应用商店 AppStore 有 4.7 万次的应用下载;全球新增网站达到 571 个。搜索引擎、电子商务等公司都要求实时地处理数以万计的海量数据,而且数据量增长迅速,传统的离线加批处理的方式已经不能满足要求,这对数据的处理也提出了更高的要求,数据处理的速度需越来越快,甚至要做到数据随时产生,随时处理。

4. 数据价值密度低

大数据目前仍处于数据价值密度很低的阶段。尽管数据价值不可估量,但受传统思维和技术的限制,人们很难充分发掘数据所蕴含的巨大价值,大数据的价值利用密度仍然较低。另外,有些数据本身价值密度就很低,例如,美国棱镜计划中,一般上网用户作为单个个体,其数据的利用价值并不高,只有一些恐怖分子、各国政要等特殊人士的数据才有分析价值。再如,交通视频监控或小区安保监控录像中的数据价值密度也很低。对交通事故责任认定而言,只有发生事故的那段视频,才是有价值的,其中大量其他数据对此并没有实际用途,换句话说,虽然视频监控器会记录下数量庞大的视频,但其中只有很少一部分对交通事故责任认定有用,因此说这些连续视频数据的利用价值密度很低。

(二) 大数据的技术特点

从技术视角来看,大数据对传统数据存储及管理平台发起了挑战,为了满足大数据的低耗能存储及高效率计算的要求,需要多种技术的协同合作。这些技术包括分布式云存储技术、高性能并行计算技术、多源数据清洗及数据整合技术、提供大数据存储—索引—查询等活动的云计算平台、解决海量数据结构复杂问题的分布式文件系统及分布式并行数据库、呈现复杂的数据分析结果的可视化高维展示技术等。下面主要介绍一下分布式云存储技术、高性能并行计算技术和可视化高维展示技术。

1. 分布式云存储技术

目前, 基于 Hadoop^① 技术的数据存储和处理工具已被广泛应用。为了应对大数据的复杂数据结构和提高系统存储管理效能, 非关系数据库 (NoSQL) 以及融合了可扩展性与高性能的新型数据库 (NewSQL) 也应运而生。典型的产品有谷歌公司的 BigTable 和 Spanner、阿帕奇 (Apache) 的 Hadoop 项目的子项目 HBase、甲骨文公司的 NoSQL Database、VoltDB 公司的 VoltDB、美国布朗大学 (Brown) 等单位开发的 H-Store 等。从分布管理的角度来看, 云平台已经成为大数据存储必不可少的技术支撑, 而在存储方式上, 由适合结构化数据的行存储改为适合大数据的列存储, 也是大数据对传统存储方式的挑战。

2. 高性能并行计算技术

数据的海量增长和复杂结构需要在提高系统性能与计算能力的基础上进行大数据的各种分析。这方面典型的技术与平台有基于 MapReduce 的各种并行计算技术 (如雅虎的 S4、Twitter 的 Storm、领英的 Kafka、美国加利福尼亚大学伯克利分校的 Spark Streaming 等) 以及微软的 Dryad 分布式并行计算平台等。传统的串行计算不仅处理能力有限, 而且时耗较长。大数据的实时快速处理需要高性能并行计算技术的强力支撑。

3. 可视化高维展示技术

面对庞大复杂的数据, 如何展示数据之间的关系以及数据分析结果, 是大数据处理的关键技术之一。可视化高维展示技术有两个方面的作用: 一是能够直观反映多维数据之间的空间关系, 包括单模数据关系 (如用户与用户的关系、商品与商品的关系) 与多模数据关系 (如用户与商品的关系); 二是能够动态演化事物的变化及变化的规律。除了传统的柱状图、饼状图、曲线图等统计图外, 大数据更多地采取网络图、地形图等来展示数据、数据与数据以及数据与时间之间的关联, 而且图形不再是静态的图片, 其中的每个要素都可以进一步点击展开, 并能随着时间变化而不断变化。

四、大数据的分析理念

大数据在分析理念上有三个重要的转变。这三个理念转变主要包括:

^① Hadoop 是一个由阿帕奇 (Apache) 基金会所开发的分布式系统基础架构, 是参考谷歌相关技术发展起来的开源系统, 其核心部分是分布式文件系统 HDFS (Hadoop Distributed File System) 和编程模型 M/R (Map/Reduce, 映射/归约)。

在数据基础上倾向于全体数据而不是抽样数据，在分析方法上更注重相关分析而不是因果分析，在分析效果上更追究效率而不是绝对精确。

（一）倾向于全体而不是抽样数据

在过去，由于缺乏获取全体样本的手段，人们发明了“随机调研数据”的方法。理论上，抽取样本越随机，就越能代表整体样本。但问题是获取一个随机样本的代价很高，而且费时费力。人口调查就是典型的例子，稍大一点的国家很难做到每年都发布一次人口调查数据，其中一个重要原因是随机调研过于耗时耗力。

有了云计算和数据库以后，获取足够大的样本数据乃至全体数据，就变得非常容易了。在电子商务应用中，电商平台不需要对用户的兴趣、收入、上网时间、每月网购花费等信息进行抽样调查，因为电商平台拥有几千万甚至上亿名用户的网络行为日志，包括用户上网渠道、浏览次数、访问路径、历史购物信息、页面停留时间等比较全面的信息与数据，这些信息足以反映用户的特征、兴趣等。而且，由于这些数据是电商平台上的所有用户留下的数据，反映的是全体用户的真实上网和购物情况，所以，针对这些真实的用户数据进行分析，比抽样调查数据更全面，更能反映整个群体的特征与规律，这些特征与规律足以为发现新的商业机会提供决策支持。

（二）注重相关分析而不是因果分析

相关性是指两个或者两个以上变量的取值之间存在某种规律性，当一个或几个相互联系的变量取一定的数值时，与之相对应的另一变量的值按照某种规律在一定的范围内变化，则认为前者与后者之间具有相关性，或者说二者是相关关系。相关性表明变量 A 和变量 B 有关，或者说 A 变量的变化和 B 变量的变化之间存在一定的正比（或反比）关系。但相关性并不一定是因果关系（A 未必是 B 的原因）。相关性（相关关系）与因果性（因果关系）是完全不同但常被混淆的两个概念。例如，根据统计结果，可以说“吸烟的人群肺癌发病率会比不吸烟的人群高几倍”，但不能得出“吸烟致癌”的逻辑结论。我国概率统计领域的奠基人之一陈希孺院士生前常用这个例子来说明相关性与因果性的区别。他说，假如有这样一种基因，它同时导致两件事情，一是使人喜欢抽烟，二是使这个人更容易得肺癌。这种假设也能解释上述统计结果，而在这种假设中，这个基因和肺癌就是因果关系，而吸烟和肺癌则是相关关系。

通过利用相关关系，可以比以前更容易、更快捷、更清楚地分析事