大数据与人工智能技术丛书

大数据分析

Python爬虫、数据清洗和数据可视化

◎ 黄源 蒋文豪 徐受蓉 主编

150分钟视频讲解

教学大纲　教学课件　程序源码

电子教案　习题答案

- 采用“理实一体化”教学方式，配套大量上机操作
- 涵盖最新大数据分析知识及相关开源库的使用
- 提供150分钟视频讲解及丰富的配套教学资源

清華大學出版社

大数据与人工智能技术丛书

大数据分析

Python爬虫、数据清洗和数据可视化

◎ 黄源 蒋文豪 徐受蓉 主编

清華大學出版社
北京

内 容 简 介

本书的编写目的是向读者介绍大数据分析的基本概念和相应的技术应用。全书共10章，具体内容包括大数据、爬虫与大数据、Scrapy爬虫、数据库连接与查询、数据可视化基础与应用、大数据存储与清洗、数据格式与编码技术、数据抽取与采集、pandas数据分析与清洗，以及数据分析与清洗综合实训。本书将理论与实践操作相结合，通过大量的案例帮助读者快速掌握和应用大数据分析相关技术，通过对书中重要的、核心的知识点的练习，达到熟练应用的效果。

本书可作为大数据专业、软件技术专业、信息管理专业、计算机网络专业的教材，也可作为大数据爱好者的参考书。

图书在版编目(CIP)数据

大数据分析：Python爬虫、数据清洗和数据可视化/黄源，蒋文豪，徐受蓉主编. —北京：清华大学出版社，2020.1
（大数据与人工智能技术丛书）
ISBN 978-7-302-53054-1

Ⅰ. ①大…　Ⅱ. ①黄… ②蒋… ③徐…　Ⅲ. ①数据处理－教材　Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2019)第094686号

策划编辑：魏江江
责任编辑：王冰飞
封面设计：刘　键
责任校对：李建庄
责任印制：沈　露

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 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：20.75　　**字　　数**：492千字
版　　次：2020年1月第1版　　**印　　次**：2020年1月第1次印刷
印　　数：1～2500
定　　价：59.80元

产品编号：081893-01

前言

大数据是现代社会高科技发展的产物。大数据相对于传统的数据分析，它是海量数据的集合，它以采集、整理、存储、挖掘、共享、分析、应用、清洗为核心，正广泛地应用在军事、金融、环境保护、通信等各个领域中。

当前，发展大数据已经成为国家战略，大数据在引领经济社会发展中的新引擎作用更加明显。2014 年“大数据”首次出现在我国《政府工作报告》中。报告中提到，要设立新兴产业创业创新平台，在大数据等方面赶超先进，引领未来产业发展。“大数据”一词逐渐在国内成为热议的词汇。2015 年国务院正式印发《促进大数据发展行动纲要》，《纲要》明确指出要不断地推动大数据发展和应用，在未来打造精准治理、多方协作的社会治理新模式，建立运行平稳、安全高效的经济运行新机制，构建以人为本、惠及全民的民生服务新体系，开启大众创业、万众创新的创新驱动新格局，培育高端智能、新兴繁荣的产业发展新生态。

本书以理论与实践操作相结合的方式深入地讲解了大数据分析的基本知识和实现的基本技术，在内容设计上既有上课时教师的讲述部分，包括详细的理论与典型的案例，又有大量的实训环节，双管齐下，可极大地激发学生在课堂上的学习积极性与主动创造性，让学生在课堂上跟上老师的思维，从而学到更多有用的知识和技能。

本书共 10 章，主要包括大数据、爬虫与大数据、Scrapy 爬虫、数据库连接与查询、数据可视化基础与应用、大数据存储与清洗、数据格式与编码技术、数据抽取与采集、pandas 数据分析与清洗，以及数据分析与清洗综合实训。

本书有如下特点。

(1) 采用“理实一体化”教学方式，课堂上既有教师的讲述，又有学生独立思考、上机操作等内容。

(2) 配套资源丰富，本书提供教学大纲、教学课件、电子教案、习题答案、程序源码等多种教学资源，扫描封底的课件二维码可以下载；本书还提供 150 分钟的视频讲解，扫描书中相应位置的二维码可以在线观看、学习。

(3) 紧跟时代潮流，注重技术变化，书中包含了最新的大数据分析知识及一些开源库的使用。

(4) 编写本书的教师都具有多年的教学经验，重难点突出，能够激发学生的学习热情。

本书可作为大数据专业、软件技术专业、信息管理专业、计算机网络专业的教材，也可作为大数据爱好者的参考书。

本书建议学时为80学时，具体分布如下表所示。

章　节	建议学时	章　节	建议学时
大数据	4	大数据存储与清洗	6
爬虫与大数据	12	数据格式与编码技术	6
Scrapy爬虫	8	数据抽取与采集	12
数据库连接与查询	6	pandas数据分析与清洗	12
数据可视化基础与应用	10	数据分析与清洗综合实训	4

本书由黄源、蒋文豪、徐受蓉编写。其中，黄源编写了第1章、第6～10章；蒋文豪编写了第2章，蒋文豪和黄源共同编写了第3章；徐受蓉编写了第4章和第5章。徐受蓉教授对书中内容进行了审阅工作，全书由黄源负责统稿工作。

本书是校企合作共同编写的结果，在编写过程中得到了中国电信金融行业信息化应用重庆基地总经理助理杨琛的大力支持。

在编写过程中，我们参阅了大量的相关资料，在此表示感谢！

由于编者水平有限，书中难免出现疏漏之处，恳请广大读者批评指正。

编　者

2019年10月于重庆

目录

配套资源下载

第1章

大 数 据

本章学习目标

- 了解大数据的定义。
- 了解大数据的特征及技术框架。
- 掌握不同数据分类。
- 了解大数据与云计算的关系。
- 了解大数据与人工智能的关系。
- 了解发展大数据的意义。
- 了解大数据在我国的发展现状。

本章先向读者介绍大数据的定义，再介绍大数据的特征及技术框架，接着介绍大数据与云计算的关系，最后介绍大数据在我国的发展现状。

1.1 大数据概述

视频讲解

1.1.1 大数据介绍

1. 大数据的定义

大数据(big data)是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

大数据是现代社会高科技发展的产物，它不是一种单独的技术，而是一个概念、一个

技术圈。相对于传统的数据分析，大数据是海量数据的集合，它以采集、整理、存储、挖掘、共享、分析、应用、清洗为核心，正广泛地应用在军事、金融、环境保护、通信等各个行业中。

2006 年，全球知名咨询公司麦肯锡最早提出了大数据的概念。在这 14 年间，大数据从商业新概念发展成了新经济增长和企业战略的关键引擎。麦肯锡认为："大数据的应用，重点不在于堆积数据，而在于利用数据，做出更好的、利润更高的决策。"因此，大数据的核心在于对海量数据的分析和利用。

按照麦肯锡的理念来理解，大数据并不是神秘的，不可触摸的，它是一种新兴的产业，从提出概念至今不断在推动着世界经济的转型和进一步的发展。如法国政府在 2013 年投入近 1150 万欧元，用于 7 个大数据市场研发项目。目的在于通过发展创新性解决方案，并将其用于实践，来促进法国在大数据领域的发展。法国政府在《数字化路线图》中列出了 5 项将大力支持的战略性高新技术，大数据就是其中一项。

综上所述，从各种各样的大数据中，快速获得有用信息的能力，就是大数据技术。这种技术已经对人们的生产和生活方式有了极大的影响，并且还在快速地发展中，不会停下来。

2. 大数据的发展历程

大数据的发展主要历经了 3 个阶段：出现阶段、热门阶段和应用阶段。

1）出现阶段(1980—2008 年)

在 1980 年未来学家阿尔文·托夫勒著的《第三次浪潮》中将"大数据"称为"第三次浪潮的华彩乐章"。1997 年美国宇航局研究员迈克尔·考克斯和大卫·埃尔斯沃斯首次使用"大数据"这一术语来描述 20 世纪 90 年代的挑战：模拟飞机周围的气流——是不能被处理和可视化的。数据集之大，超出了主存储器、本地磁盘，甚至远程磁盘的承载能力，因而被称之为"大数据问题"。

谷歌(Google)在 2006 年首先提出云计算的概念。2007—2008 年随着社交网络的激增，技术博客和专业人士为"大数据"概念注入新的生机。"当前世界范围内已有的一些其他工具将被大量数据和应用算法所取代"。《连线》的克里斯·安德森认为当时处于一个"理论终结时代"。一些国家的政府机构和美国的顶尖计算机科学家声称："应该深入参与大数据计算的开发和部署工作，因为它将直接有利于许多任务的实现"。2008 年 9 月，《自然》杂志推出了名为"大数据"的封面专栏，同年"大数据"概念得到了美国政府的重视；计算社区联盟(Computing Community Consortium)发表了第一个关于大数据的白皮书《大数据计算：在商务、科学和社会领域创建革命性突破》，其中提出了当年大数据的核心作用：大数据真正重要的是寻找新用途和散发新见解，而非数据本身。

2）热门阶段(2009—2012 年)

从 2009—2010 年"大数据"成为互联网技术行业中的热门词汇。2009 年印度建立了用于身份识别管理的生物识别数据库；2009 年联合国全球脉冲项目研究了如何利用手机和社交网站的数据源来分析预测从螺旋价格到疾病暴发之类的问题；2009 年美国政府通过启动 Data.gov 网站的方式进一步开放了数据的大门，该网站的超过 4.45 万个数据集被用于保证一些网站和智能手机应用程序来跟踪信息，这一行动促使肯尼亚及英国

政府相继推出类似举措；2009 年，欧洲一些领先的研究型图书馆和科技信息研究机构建立了伙伴关系致力于改善在互联网上获取科学数据的简易性。2010 年肯尼斯库克尔发表大数据专题报告《数据，无所不在的数据》；2011 年 2 月扫描 2 亿兆的页面信息或 4 亿兆字节磁盘存储，只需几秒即可完成。IBM 的沃森计算机系统在智力竞赛节目《危险边缘》中打败了两名人类挑战者。后来《纽约时报》评论这一刻为一个大数据计算的胜利。“大数据时代已经到来”出现在 2011 年 6 月麦肯锡发布了关于“大数据”的报告，正式定义了大数据的概念，后来逐渐受到各行各业的关注。

2012 年，“大数据”一词越来越多地被提及，人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。数据正在迅速膨胀并变大，它决定着未来发展，随着时间的推移，人们将越来越多地意识到数据的重要性。

2012 年，美国奥巴马政府在白宫网站发布了《大数据研究和发展倡议》，这一倡议标志着大数据已经成为重要的时代特征；2012 年 3 月 22 日，奥巴马政府宣布花费 2 亿美元投资大数据领域，是大数据技术从商业行为上升到国家科技战略的分水岭；2012 年美国颁布了《大数据的研究和发展计划》；英国发布了《英国数据能力发展战略规划》；日本发布了《创建最尖端 IT 国家宣言》；韩国提出了“大数据中心战略”；世界其他一些国家也制定了相应的战略和规划。

3）应用阶段（2013—2016 年）

2014 年“大数据”首次出现在我国《政府工作报告》中。报告中提到要设立新兴产业创业创新平台，在大数据等方面赶超先进，引领未来产业发展。“大数据”一词逐渐在国内成为热门词。

2015 年国务院正式印发的《促进大数据发展行动纲要》明确指出，要不断地推动大数据发展和应用，在未来打造精准治理、多方协作的社会治理新模式；建立运行平稳、安全高效的经济运行新机制；构建以人为本、惠及全民的民生服务新体系；开启大众创业、万众创新的创新驱动新格局；培育高端智能、新兴繁荣的产业发展新生态。

2015 年，“十三五”规划出台，规划通过定量和定性相结合的方式提出了 2020 年大数据产业的发展目标。在总体目标方面，提出到 2020 年，技术先进、应用繁荣、保障有力的大数据产业体系基本形成，大数据相关产品和服务业务收入突破 1 万亿元，年均复合增长率保持 30%左右。

此外，随着我国大数据产业规模的迅速扩张，2016 年我国大数据市场规模约为 168 亿人民币，预计到 2020 年大数据产业市场规模将达到 578 亿元，年均增长率在 30%以上。

3. 大数据的影响

大数据的影响主要有以下 4 点。

1）大数据对科学活动的影响

人类在科学研究上先后历经了实验、理论和计算 3 种范式。当数据量不断增长并积累到今天，传统的 3 种范式在科学研究，特别是一些新的研究领域已经无法很好地发挥作用，需要有一种全新的第四种范式来指导新形势下的科学研究。这种新的范式就是从以计算为中心，转变到以数据处理为中心，确切地说也就是数据思维。

数据思维是指在大数据环境下，一切资源都将以数据为核心，人们从数据中去发现问题，解决问题，在数据背后挖掘真正的价值，科学大数据已经成为科技创新的新引擎。在维克托·迈尔-舍恩伯格撰写的《大数据时代》(中文版译名)中明确指出，大数据时代最大的转变，就是放弃对因果关系的渴求，取而代之关注相关关系。也就是说，只要知道“是什么”，而不需要知道“为什么”。这就颠覆了千百年来人类的思维惯例，据称是对人类的认知和与世界交流的方式提出了全新的挑战。虽然第三范式和第四范式都是利用计算机来计算，但它们在本质上是不同的。第四范式彻底颠覆了人类对已知世界的理解，明确了一点：如果能够获取更全面的数据，也许才能真正做出更科学的预测，这就是第四范式的出发点，也许是最迅速和实用的解决问题的途径。

因此，大数据将成为科学研究者的宝库，从海量数据中挖掘有用的信息会是一件极其有趣而复杂的事情。它要求人们既要依赖于数据，又要有独立的思考，能够从不同数据中找出隐藏的关系，从而提取出有价值的信息。如图 1-1 所示是科学范式的发展过程。

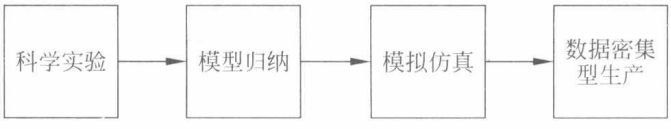

图 1-1　科学范式的发展过程

2) 大数据对思维方式的影响

(1) 人们处理的数据从样本数据变成全部数据：面对大数据，传统的“样本数据”可能不再适用，由于大数据分析处理技术的出现，使得人们对“全量数据”的处理变得简易可行。大数据时代，带来了从“样本数据”到“全量数据”的转变。在大数据可视化时代，数据的收集问题不再成为人们的困扰，采集全量的数据成为现实。全量数据带给人们视角上的宏观与高远，这将使人们可以站在更高的层级全貌看待问题，看见曾经被淹没的数据价值，发现藏匿在整体中有趣的细节。因为拥有全部或几乎全部的数据，就能使人们获得从不同的角度更细致、更全面地观察研究数据的可能性，从而使得大数据平台的分析过程成为惊喜的发现过程和问题域的拓展过程。

(2) 由于是全样本数据，人们不得不接受数据的混杂性，而放弃对精确性的追求：传统的数据分析为了保证精确性和准确性，往往采取抽样分析来实现。而在大数据时代，往往采取全样分析而不再采用以往的抽样分析。因此追求极高精确率的做法已经不再是人们的首要目标，速度和效率取而代之，如在几秒内就迅速给出针对海量数据的实时分析结果等。同时人们也应该容许一些不精确的存在，数据不可能是完全正确或完全错误的，当数据的规模以数量级增加时，对大数据进行深挖和分析，能够把握真正有用的数据，才能避免做出盲目和错误的决策。

(3) 人类通过对大数据的处理，放弃对因果关系的渴求，转而关注相关关系：在以往的数据分析中，人们往往执着于现象背后的因果关系，总是试图通过有限的样本来剖析其中的内在机制；而在大数据的背景下，相关关系大放异彩。通过应用相关关系，人们可以比以前更容易、更便捷、更清楚地分析事物。例如，美国一家零售商在对海量的销售数据处理中发现每到星期五下午，啤酒和婴儿尿布的销量同时上升。通过观察发现星期五下

班后很多青年男子要买啤酒度周末，而这时妻子又常打电话提醒丈夫在回家路上为孩子买尿布。发现这个相关性后，这家零售商就把啤酒和尿布摆在一起，方便年轻的爸爸购物，大大提高了销售额。再如，谷歌开发了一个名为“谷歌流感趋势”的工具，它通过跟踪搜索词相关数据来判断全美地区的流感情况。这个工具会发出预警，告诉全美地区的人们流感已经进入“紧张”的级别。这样的预警对于美国的卫生防疫机构和流行病健康服务机构来说非常有用，因为它及时，而且具有说服力。此工具的工作原理为通过关键词（如温度计、流感症状、肌肉疼痛、胸闷等）设置，对搜索引擎的使用者展开跟踪分析，创建地区流感图表和流感地图（以大数据的形式呈现出来）。然后再把结果与美国疾病控制和预防中心的报告做对比，进行相关性预测。

3）大数据对社会发展的影响

在大数据时代，不管是物理学、生物学、环境生态学等领域，还是军事、金融、通信等行业，数据正在迅速膨胀，没有一个领域可以不被波及。“大数据”正在改变甚至颠覆着人们所处的整个时代，对社会发展产生了方方面面的影响。

在大数据时代，用户会越来越多地依赖于网络和各种“云端”工具提供的信息作出行为选择。从社会这个大方面上看，这有利于提升人们的生活质量、和谐程度，从而降低个人在群体中所面临的风险。例如，美国的网络公司 Farecast 通过对 2000 亿条飞行数据记录的搜索和运算，可以预测美国各大航空公司每一张机票的平均价格的走势，如果一张机票的平均价格呈下降趋势，系统就会帮助用户作出稍后再购票的明智选择。反过来，如果一张机票的平均价格呈上涨趋势，系统就会提醒用户立刻购买该机票。通过预测机票价格的走势及增降幅度，Farecast 的票价预测工具能帮助消费者抓住最佳购买时机，节约出行成本。

现在，Google 的无人驾驶汽车已经在加州行驶了几千千米，未来人们可以通过人工智能与汽车产生互动，从而使自动驾驶得以实现，这些都是基于大量数据解析的结果，背后都有大数据的功劳。

4）大数据对就业市场的影响

大数据激发内需的剧增，引发产业的巨变。生产者具有自身的价值，而消费者则是价值的意义所在。有意义的东西才会有价值，消费者如果不认同，就卖不出去，价值就实现不了；消费者如果认同，就卖得出去，价值就得以体现。大数据可以帮助人们从消费者这里分析意义所在，从而帮助生产者实现更多的价值。

此外，随着大数据的不断应用，随之带来各行各业数据业务转型升级。例如，在金融业，原来的主业是做金融业务，靠佣金赚钱；而如今清算结算可能免费，利用支付信息的衍生信息增值业务赚钱。

1.1.2 大数据的特征

随着对大数据认识的不断加深，人们认为大数据一般具有 4 个特征：数据量大、数据类型繁多、数据产生速度快及数据价值密度低。

1. 数据量大

大数据中的数据量大，就是指的海量数据。由于大数据往往是采取全样分析，因此大数据的“大”首先体现在其规模和容量远远超出传统数据的测量尺度，一般的软件工具难以捕捉、存储、管理和分析的数据，通过大数据的云存储技术都能保存下来，形成浩瀚的数据海洋，目前的数据规模已经从太字节（TB）级升级至拍字节（PB）级。大数据之“大”还表现在其采集范围和内容的丰富多变，能存入数据库的不仅包含各种具有规律性的数据符号，还囊括了各种如图片、视频、声音等非规则的数据。

2011 年，马丁·希尔伯特和普里西利亚·洛佩兹在《科学》上发表了一篇文章，对 1986—2007 年人类所创造、存储和传播的一切信息数据量进行了追踪计算。其研究范围大约涵盖了 60 种模拟和数字技术：书籍、图画、信件、电子邮件、照片、音乐、视频（模拟和数字）、电子游戏、电话、汽车导航等。

据他们估算：2007 年，人类存储了超过 300EB 的数据；1986—2007 年，全球数据存储能力每年提高 23%，双向通信能力每年提高 28%，通用计算能力每年提高 58%；在 2013 年，世界上存储的数据能达到约 1.2ZB，预计到 2020 年，世界上存储的数据总量将达到不可思议的 35ZB（1ZB=1024EB，1EB=1024PB，1PB=1024TB，1TB=1024GB）。

2. 数据类型繁多

大数据包括结构化数据、非结构化数据和半结构化数据。

（1）结构化数据常指存储关系在数据库中的数据，该数据遵循某种标准，如企业财务报表、医疗数据库信息、行政审批数据、学生档案数据等。

（2）非结构化数据常指不规则或不完整的数据，包括所有格式的办公文档、XML、HTML、各类报表、图片及音频、视频信息等。企业中 80%的数据都是非结构化数据，这些数据每年都按指数增长 60%。相对于以往便于存储的以文本为主的结构化数据，越来越多的非结构化数据的产生给所有厂商都提出了挑战。在网络中非结构化数据越来越成为数据的主要部分。值得注意的是，非结构化数据具有内部结构，但不通过预定义的数据模型或模式进行结构化。它可能是文本的或非文本的，也可能是人为的或机器生成的。它也可以存储在像 NoSQL 这样的非关系数据库中。

（3）半结构化数据常指有一定的结构与一致性约束，但在本质上不存在关系的数据，如常用于跨平台传输的 XML 数据及 JSON 数据等。

据 IDC 的调查报告显示：随着互联网和通信技术迅猛发展，如今的数据类型早已不是单一的文本形式，如网络日志、音频、视频、图片、地理位置信息等多类型的数据对数据的处理能力提出了更高的要求，并且数据来源也越来越多样，不仅产生于组织内部运作的各个环节，也来自于组织外部的开放数据。其中内部数据主要包含：政府数据，如征信、户籍、犯罪记录等；企业数据，如阿里巴巴的消费数据，腾讯的社交数据，“滴滴出行”的数据等；机构数据，如第三方咨询机构的调查数据。而开放数据主要包含网站数据和各种 APP 终端数据及大众媒介数据等。

例如，苹果公司在 iPhone 手机上应用的一项语音控制功能 Siri 就是多样化数据处理

的代表。用户可以通过语音、文字输入等方式与 Siri 对话交流，并调用手机自带的各项应用，读短信、询问天气、设置闹钟、安排日程，乃至搜寻餐厅、电影院等生活信息，收看相关评论，甚至直接订位、订票，Siri 则会依据用户默认的家庭地址或是所在位置判断、过滤搜寻的结果。企业与企业间的电子商务又称为 B2B，B2B 是 business to business 的缩写。它是指企业与企业之间通过专用网络或 Internet，进行数据信息的交换、传递，开展交易活动的商业模式，它将企业内部网和企业的产品及服务，通过 B2B 网站或移动客户端与客户紧密结合起来，通过网络的快速反应，为客户提供更好的服务，从而促进企业的业务发展。

3. 数据产生速度快

在数据处理速度方面，有一个著名的“1 秒定律”，即要在秒级时间范围内给出分析结果，超出这个时间，数据就失去价值了。大数据是一种以实时数据处理、实时结果导向为特征的解决方案，它的“快”有两个层面。

(1) 数据产生得快。有的数据是爆发式产生，例如，欧洲核子研究中心的大型强子对撞机在工作状态下每秒产生拍字节(PB)级的数据；有的数据是涓涓细流式产生，但是由于用户众多，短时间内产生的数据量依然非常庞大，例如，点击流、日志、论坛、博客、发邮件、射频识别数据、GPS(全球定位系统)位置信息。

(2) 数据处理得快。正如水处理系统可以从水库调出水进行处理，也可以处理直接对涌进来的新水流。大数据也有批处理(“静止数据”转变为“正使用数据”)和流处理(“动态数据”转变为“正使用数据”)两种范式，以实现快速的数据处理。

例如，电子商务网站从点击流、浏览历史和行为(如放入购物车)中实时发现顾客的即时购买意图和兴趣，并据此推送商品，这就是数据“快”的价值，也是大数据的应用之一。

4. 数据价值密度低

随着互联网及物联网的广泛应用，信息感知无处不在，信息海量，但价值密度较低，如何结合业务逻辑并通过强大的机器算法来挖掘数据价值，是大数据时代最需要解决的问题。以视频为例，一段 1 小时的视频，在连续不间断监控过程中，可能有用的数据仅仅只有一两秒。但是为了能够得到人们想要的视频，不得不投入大量资金用于购买网络设备、监控设备等。

因此，由于数据采集的不及时，数据样本不全面，数据可能不连续等，数据可能会失真，但当数据量达到一定规模，可以通过更多的数据达到更真实全面的反馈。

1.1.3 大数据技术应用与基础

1. 大数据应用

大数据的应用无处不在，从金融业到娱乐业，从制造业到互联网行业，从物流业到运输业等各行各业都有大数据的身影。

金融业：通过大数据预测企业的金融风险，并通过描绘用户画像，清楚用户的消费行