第5章

数字技术通识

5.1 大 数 据

早在20世纪，科学家们已经开始意识到，随着计算机存储数据量的快速增长，传统的数据分析面临着数据获取、理解、处理和组织等方面的困难。因此，计算机领域开始广泛使用“大数据”(big data)这一术语来描述这些挑战(图5-1)。如今，大数据已经引起了人们的广泛关注，并推动了一系列涵盖广泛主题的新型研究。简言之，大数据的数据规模已经超出了传统计算机软件和技术的处理能力，因此，人们迫切需要更大的存储容量来承载和管理数据，也需要更有效且高效的技术来处理这些数据。

图5-1 大数据

5.1.1 大数据基本概念

应知应会

大数据通常由规模庞大的数据集构成，这些数据集往往超出了人类在可接受的时间范围内收集、管理和处理的能力。随着时间的推移，人们对于数据规模的认识也在不断发展。例如，2000年，1TB的数据通常被视为大数据，而到了2020年，大数据的门槛已经提高到PB(TB的数千倍)级别。除了数据规模，大数据的挑战还体现在数据类型的多样性、对数据处理时间的严格要求等方面，这些挑战通常被学术界概括为“nV”特性。下面将通过相对广为人知的“3V”特性来描述大数据的三大特征：Volume(规模)、Variety(多样性)和Velocity(速度)。

1. 数据规模大

数据规模大(big volume)是大数据当中最基础的特性，意味着数据集合的存

储需要消耗 TB、PB 甚至 EB 级别的存储介质。以某款高清摄像头为例，每天能产生超过 500G 的图像数据，相当于每年的数据量累计高达 180TB。因此，大数据无法使用一般的存储介质进行存储，必须首先对数据进行切分，然后使用数十台、数百台甚至数千台计算机，利用云存储技术来分摊存储压力（图 5-2）。

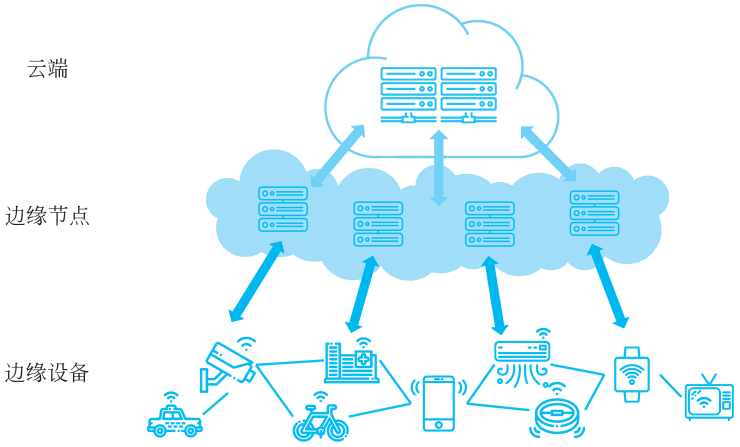

图 5-2　大数据和云存储

除了历史数据的积累所产生的存储压力，针对大数据的处理压力已经显现出来。例如，沃尔玛公司需要在 1 小时内处理掉高达几百万条的消费信息。因此，大数据几乎无法使用传统的数据库管理系统来管理、维护，而必须使用在数十、数百甚至数千台服务器上同时并行运行的软件。

大数据这一特征同样考验着数据持有机构的数字技术能力。对某些组织来说，当突然面对 1GB 的数据集时，它们的数字系统就会崩溃；而对另一些组织来说，数据集的大小可能需要高达 1TB 才会对它们造成困扰。

2. 数据种类多（big variety）

大数据的获取来源将影响其应用的效益与质量，依照获取的直接程度一般可分为以下三种。

第一方数据（first party data），为己方单位自己和消费者、用户、目标客群交互产生的数据，具有高质量、高价值的特性，但易局限于既有顾客数据，如企业搜集的顾客交易数据、追踪用户在 App 上的浏览行为等，拥有者可弹性地将这类数据用于分析研究、营销推广等用途。

第二方数据（second party data），取自第一方数据，通常与第一方具有合作、联盟或契约关系，因此可共享或采购第一方数据。第二方数据的例子有订房品牌与飞机品牌共享数据，当客人购买某一方的商品后，另一单位即可向客户推荐相关的旅游产品；或是已知某单位具有己方想要的数据，通过协议采购，直接从第一方获取数据。

第三方数据（third party data），提供数据的来源单位并非产出该数据的原始作者，该数据即为第三方数据。提供第三方数据的单位通常为数据供应商，其广泛搜集各式数据，并将数据贩售给数据需求者，其数据可来自第一方、第二方与其他第三方数据，如爬取网络公开数据、市调公司所发布的研究调查、经去识别化的交易信息等。

大数据的应用广泛，科学研究、企业应用和 Web 应用等都在源源不断地产生新类型

的大数据。大数据的应用示例已经涵盖了各行各业，典型的大数据类型包括生物大数据、交通大数据、医疗大数据、工业大数据、城市大数据、金融大数据、银行大数据、消费大数据等。

综上所述，大数据的类型非常丰富，而它们通常可以被抽象成三类：结构化数据、半结构化数据和非结构化数据。


数据的分类

结构化数据，指的是具有固定的结构、类型和属性划分的数据，通常可以采用二维表结构来表述。例如学生信息表(表 5-1)，它包含了学号、姓名、性别、出生日期等属性。

表 5-1　结构化数据示例

学　　号	姓　　名	性　　别	出 生 日 期
1001	Tom	male	19870604
1002	Jerry	female	19890707

半结构化数据，这种数据保留了结构化数据相同的表达能力，但兼顾了灵活性。例如，XML(可扩展标记语言)允许将数据属性的自描述、数据结构和数据内容集中在一起，现在已经成为国际上进行数据交换的一种公共语言。使用 XML 格式来表示表 5-1 中的数据记录，代码如下所示。

```
<students>
    <stu>
        <id>1001</id>
        <name>Tom</name>
        <sex>male</sex>
        <birth>19870604</birth>
    </stu>
    <stu>
        <id>1002</id>
        <name>Jerry</name>
        <fname>Li</fname>
        <sex>female</sex>
        <birth>19890707</birth>
    </stu>
</students>
```

注：表 5-1 中的第二条数据记录包含一项特殊的属性 fname，在代码中用浅紫色表示。

非结构化数据(图 5-3)，指的是无法采用固定的结构来表示的数据，如图片、视频、音频等。非结构化数据的格式丰富多样，且无法采用结构化或半结构化表示。在技术层面，非结构化数据包含的信息量远超结构化数据，也需要占用更多的存储空间。根据国际数据组织的调查显示，非结构化数据约占全部数据总量的 80%。

3. 数据处理速度快

数据处理速度快(big velocity)这一特性指的是数据从生成到处理进而产生结果的速度。传统的数字技术往往将准确性摆在第一位，通常不会对处理数据所需的时间有严格的

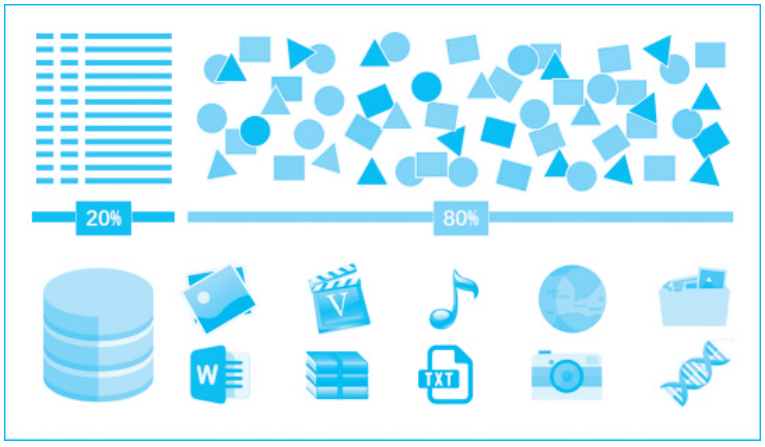

图 5-3　非结构化数据

限制。但对于需要借助大数据做出实时性决策的公司来说，该特性就显得非常重要。所谓“秒级响应”即要求公司所采用的数字技术能够在 1 秒的时间内针对海量规模数据做出实时分析并反馈结果，否则这些数据的价值就会丧失，甚至造成不可挽回的损失。数据分析过程的效率正在成为人们关注的核心问题。

5.1.2　大数据的典型应用　　应知应会

大数据正在对社会发展、人们生活产生深远的影响，具体体现在以下三方面。

1. 大数据决策

所谓大数据决策，即指组织或机构的领导者采用手中掌握的数据做出决策，而不是根据传统的经验积累进行决策。目前，基于数据实时采集、分析得出的宏观决策、实时营销方案、个性化推荐服务等已经成为一种备受追捧的全新决策方式。例如，我国政府部门将大数据技术融入舆情分析，通过对微博、微信、论坛等多种来源数据进行综合分析，揭示信息中隐含的情报内容，协助实现政府决策，可以有效应对各类突发事件。

2. 大数据推动新技术发展

大数据的应用需求已经成为新技术开发的源泉，目前已经涌现出了多种形式的大数据技术，并且得到了广泛的应用。数据所蕴含的能力正在得到释放。在不远的将来，原本依靠人类自身判断力的应用正在被各种基于大数据的应用所取代。例如，DeepMind 公司的围棋软件 AlphaGo 战胜了李世石、柯洁在内的人类职业棋手；特斯拉汽车公司推出能在特定环境下工作的自动驾驶汽车。

3. 大数据促进数字技术与各行业的深度融合

当前有普遍的观点认为，一方面，大数据将在未来十年抹除几乎每一项行业的业务功能，包括互联网、银行、金融、交通、能源、服务等；另一方面，不断积累、发展的大数据将加速推进这些行业与数字技术的深度融合，开拓行业发展的新方向。例如，国内银行裁减了业务部门中很大比例的员工，而为大数据的维护、应用新增了大量的就业岗位。大数据的影响无处不在，表 5-2 展示了大数据在各个领域的应用情况。

表 5-2　大数据在各个领域的应用

领　域	大数据应用
制造	利用工业大数据提升制造业水平，优化生产计划，优化供应链
汽车	无人驾驶汽车，自动导航
互联网	个性化推荐，针对性广告投送
能源	优化电网运行，优化电力需求响应，保障电网运行安全
物流	优化物流网络，提高物流效率，降低物流成本
交通	智能交通，优化交通线规划
体育	预测比赛结果，选拔选手
银行	信贷风险预测，出台客户个性化挽留策略

5.2　云　计　算

5.2.1　云计算基本概念

背景知识

云计算实现了通过互联网提供可伸缩的、廉价的分布式计算能力，用户只需要处于具备互联网接入条件的地方就可以随时随地获得各类所需的数字技术资源。云计算代表了以虚拟化技术为核心，以低成本为目标，动态可伸缩的网络应用基础设施，是近十年来最有代表性的网络计算技术与模式。

云计算主要包括三种典型的服务模式：基础设施即服务（Infrastructure as a Service，IaaS）、平台即服务（Platform as a Service，PaaS）和软件即服务（Software as a Service，SaaS），如图 5-4 所示。IaaS 将计算、存储资源等基础设施作为服务出租，PaaS 将平台作为服务出租，而 SaaS 则把软件作为服务出租。

图 5-4　云计算的服务模式

所谓的云即提供的服务，根据服务的对象可以分为以下三类。

（1）公有云，面向所有的用户。只要是注册付费的用户都可以使用，典型的公有云服务商有 Amazon 和阿里巴巴。

（2）私有云，只为特定用户提供服务，例如企业自建的云环境，只能为内部网络用户提供服务。相比公有云，私有云可以更好地保证数据的安全性。

(3) 混合云,可以理解为公有云和私有云的混合搭配应用。

云计算的关键技术包括虚拟化、分布式存储和分布式计算。

虚拟化是云计算基础架构的基石。通过虚拟化,一台计算机可以被虚拟为多台逻辑上的计算机,逻辑上的计算机彼此独立。这意味着每台逻辑计算机可以安装不同的操作系统,应用程序在独立的空间内运行,不会相互影响,安全性大大提高。典型的虚拟化软件有VMware、Virtualbox、QEMU、Docker 等。

面对海量数据的存储压力,将数据集中到一台机器上进行存储不太现实。分布式存储可以很好解决这一难题。谷歌文件系统(Google File System,GFS)是一款最典型的分布式文件系统(图 5-5),可以自动将 TB 甚至 PB 级别的超大文件拆分成大小相等的块,分散存储在由数百台,甚至数千台服务器组成的巨大服务集群当中。此外,GFS 还提供了很好的硬件容错性,即使集群中的小部分服务器突然损坏也不会造成数据丢失。

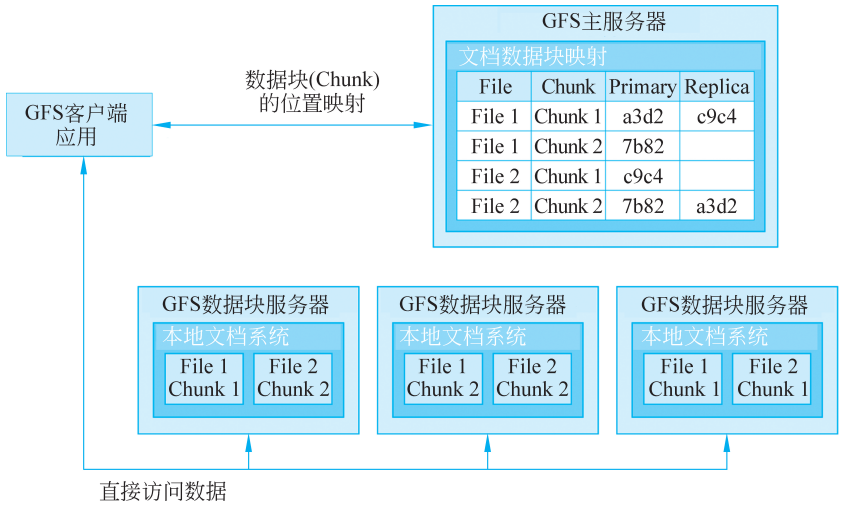

图 5-5 谷歌文件系统体系结构

分布式计算与分布式存储相辅相成。谷歌提出的并行编程框架 MapReduce 将针对海量数据执行的拆分、统计操作自动分发到不同的服务器上,执行并行处理过程,因此能够极大地提升海量规模数据的处理速度,缩短处理时间,有效满足许多应用对海量规模数据的批量处理需求。现阶段所说的分布式计算已经不仅是一种并行计算,而是融合了分布式计算、效用计算、负载均衡、并行计算、网络存储、热备份冗杂和虚拟化等计算机技术混合演进并跃升的结果(图 5-6)。

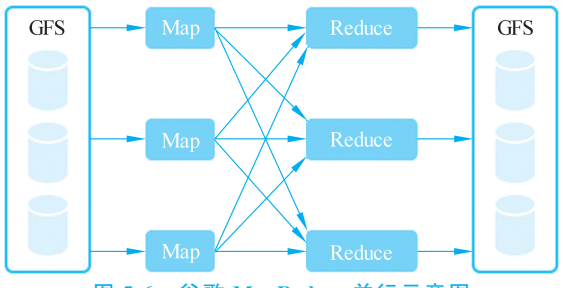

图 5-6 谷歌 MapReduce 并行示意图

云计算包含如下的五大特点。

(1) 高可伸缩性。

高可伸缩性指云计算系统能够根据用户需求快速扩展或缩减计算资源的能力。这种灵活性使得云计算能够适应不同规模和变化的工作负载,确保资源的有效利用和性能的稳定。高可伸缩性是云计算的重要特性之一,它使得企业能够根据需求实时调整计算资源,降低成本并提高效率。在云计算环境中,通常通过自动化的方式来实现高可伸缩性,包括自动扩展和收缩实例、负载均衡以及弹性存储等技术手段。

(2) 按需部署。

计算机包含了许多应用、程序软件等,不同的应用对应的数据资源库不同,所以用户运行不同的应用需要较强的计算能力对资源进行部署,而云计算平台能够根据用户的需求快速配备计算能力及资源。

(3) 高灵活性。

云计算可以灵活地扩展和缩减计算资源,满足企业不断变化的需求。

(4) 高可靠性。

高可靠性指云计算系统在面对各种挑战和故障时能够保持稳定运行的能力。即使云服务器发生故障也不影响计算与应用的正常运行,这是因为单点服务器出现故障可以通过虚拟化技术将分布在不同物理服务器上的应用进行恢复或利用动态扩展功能部署新的服务器进行计算。

(5) 高性价比。

将资源放在虚拟资源池中统一管理在一定程度上优化了物理资源,用户不再需要昂贵的、存储空间大的主机,可以选择相对廉价的 PC 组成云,一方面减少费用,另一方面计算性能不逊于大型主机。

云计算的基本载体被称为云计算数据中心。数据中心是提供云计算服务的机房,包含一整套复杂组件构成的基础设施(图 5-7)。这些组件包括刀片服务器、宽带网络连接、环境控制、监控设备以及各种安全装置等。云计算提供的计算、存储、带宽等硬件资源都集中在数据中心。它为各个平台和应用提供运行支撑环境。

图 5-7　谷歌数据中心一角

5.2.2 云计算的典型应用

应知应会

较为简单的云计算技术现已普遍服务于如今的互联网服务中，其中网络搜索引擎和网络邮箱是最为常见的云计算应用。在任何时刻，人们只要通过移动终端就可以在搜索引擎上搜索任何自己想要的资源，通过云端共享数据资源。网络邮箱也是如此。在过去，寄写一封邮件是一件比较麻烦的事情，同时也是很慢的过程，而在云计算技术和网络技术的推动下，电子邮箱成为社会生活中的一部分，只要在网络环境下，就可以轻松实现邮件的收发。

存储云，又称云存储，是在云计算技术上发展起来的新型存储技术。云存储是一个以数据存储和管理为核心的云计算系统。用户可以将本地的资源上传至云上，可以在任何地方接入互联网来获取云上的资源。读者所熟知的谷歌、微软等大型网络公司均有云存储的服务，在国内，百度云和微云则是市场占有量最大的存储云。存储云向用户提供了存储容器服务、备份服务、归档服务和记录管理服务等，大大方便了使用者对资源的管理。

医疗云，是在云计算、移动技术、多媒体、通信、大数据以及物联网等新技术的基础上，结合医疗技术，使用“云计算”来创建医疗健康服务云平台，实现了医疗资源的共享和医疗范围的扩大。因为云计算技术的运用与结合，医疗云提高医疗机构的效率，方便居民就医。像现在医院的预约挂号、电子病历、医保等都是云计算与医疗领域结合的产物，医疗云还具有数据安全、信息共享、动态扩展、布局全国等优势。

金融云，是指利用云计算的模型，将信息、金融和服务等功能分散到庞大分支机构构成的互联网“云”中，旨在为银行、保险和基金等金融机构提供互联网处理和运行服务，同时共享互联网资源，从而解决现有问题并且达到高效、低成本的目标。2013 年 11 月 27 日，阿里云整合阿里巴巴旗下资源并推出阿里金融云服务。这就是现在基本普及了的快捷支付，因为金融与云计算的结合，现在只需要在手机上简单操作，就可以完成银行存款、购买保险和基金买卖。现在，不仅阿里巴巴推出了金融云服务，像苏宁金融、腾讯等企业均推出了自己的金融云服务。

教育云，实质上是教育信息化的一种发展。具体来说，教育云可以将所需要的任何教育硬件资源虚拟化，然后将其传入互联网中，以向教育机构和学生老师提供一个方便快捷的平台。现在流行的慕课（MOOC）就是教育云的一种应用。慕课指的是大规模开放的在线课程。现阶段慕课的三大优秀平台为 Coursera、edX 以及 Udacity，在国内，中国大学 MOOC 也是非常好的平台。在 2013 年 10 月 10 日，清华大学推出 MOOC 平台——学堂在线，许多大学现已使用学堂在线开设了一些 MOOC。

5.3 物 联 网

物联网是新一代数字技术的重要组成部分，具有非常广泛的用途，同时和云计算、大数据有着密切的联系。

5.3.1 物联网基本概念

背景知识

物联网（Internet of Things，IoT）是物物相连的“互联网”，是互联网的延伸。它利用局部网络或互联网等通信技术把传感器、控制器、计算机、操作员通过新的方式连在一起，形成了人与物、物与物的相连，目的是实现信息化和远程控制。

从技术架构上看，物联网可以分为 4 层：感知层、网络层、运算层、应用层。各层的具体功能如表 5-3 所示。

表 5-3　物联网各个层次的具体功能

层次	功　能
感知层	感知层是物联网的最底层，感知层设备被广泛用来感知物理世界。感知层设备包括各类的传感器，如温度传感器、湿度传感器、加速度传感器，还有摄像头、GPS 设备等
网络层	网络层在感知层之上，起到信息传递的作用。网络层包含各种类型的网络，如互联网、移动通信网络、卫星通信网络等
运算层	运算层提供运算和存储的服务，包括数据存储、管理和分析平台等
应用层	应用层是物联网的最顶层，应用层直接面对用户，满足用户直接的应用需求，包括智能交通、智慧农业、智慧医疗、智慧工业等

物联网是物和物相连的网络（图 5-8），通过为物体加装二维码、传感器等，就可以实现物体身份唯一标识及其各种信息的采集，再结合网络连接，就可以实现人和物、物和物之间的信息交换。因此，物联网所包含的数字技术非常广泛，其中包括模式识别和感知技术、网络和通信技术、数据挖掘与融合技术等。

图 5-8　物联网

5.3.2　物联网的典型应用

应知应会

物联网的应用领域涉及方方面面，其在工业、农业、环境、交通、物流、安保等基础设施领域的应用有效地推动了这些方面的智能化发展，使有限的资源被更加合理地分配、使用，从而提高了行业效率和效益。在家居、医疗健康、教育、金融、服务业、旅游业等与生活息息相关的领域的应用，使服务范围、服务方式到服务质量等方面都有了极大的改进，大大地提高了人们的生活质量；在国防军事领域，虽然还处在研究探索阶段，但物联网应用带来的影响也不可小觑，大到卫星、导弹、飞机、潜艇等装备系统，小到单兵作战装备，物联网技术的嵌入有效促进了军事智能化、信息化、精准化，极大提升了军事战斗力，是未来军事变革的关键。

1. 智能交通

物联网技术在道路交通方面的应用比较成熟。随着社会车辆越来越普及，交通拥堵甚至瘫痪已成为城市的一大问题。对道路交通状况实时监控并将信息及时传递给驾驶人，让

驾驶人及时做出出行调整，可以有效缓解公共交通的压力；高速路口设置道路自动收费系统（简称 ETC），免去进出口取卡、还卡的时间，提升了车辆的通行效率；公交车上安装定位系统，乘客可以及时了解公交车行驶路线与到站时间，根据搭乘路线确定出行计划，免去不必要的时间浪费。社会车辆增多，除了会带来交通压力外，停车难也日益成为一个突出问题，不少城市推出了智慧路边停车管理系统，该系统基于云计算平台，结合物联网技术与移动支付技术，共享车位资源，提高车位利用率和用户的方便程度。该系统可以兼容手机模式和射频识别模式，通过手机端 App 软件可以实现及时了解车位信息，提前做好预订并实现交费等操作，很大程度上解决了停车难的问题。

2. 智能家居

智能家居是物联网在家庭中的基础应用，随着宽带业务的普及，智能家居产品涉及方方面面。家中无人时，可利用手机等客户端远程操作智能空调，调节室温，甚至还可以学习用户的使用习惯，实现全自动的温控操作，使用户在炎炎夏季回家就能享受到冰爽的惬意。另外通过客户端还可以实现智能电灯的开关、调控电灯的亮度和颜色等；插座内置 WiFi，可实现遥控插座定时通断电流，监测设备用电情况，生成用电图表让用户对用电情况一目了然，安排资源使用及开支预算。智能体重秤可以监测运动效果；内置可以监测血压、脂肪量的先进传感器，内置程序可以根据用户身体状态提出健康建议；智能牙刷与客户端相连，提供刷牙时间、刷牙位置提醒，可根据刷牙的数据生成图表，监控口腔的健康状况。智能摄像头、窗户传感器、智能门禁、烟雾探测器、智能报警器等都是家庭不可少的安全监控设备，即使出门在外，用户也可以在任意时间地点查看家中任何一角的实时状况，防范安全隐患。看似烦琐的种种家居生活会因为物联网变得更加轻松、美好。

3. 公共安全

近年来全球气候异常情况频发，灾害的突发性和危害性进一步加大，互联网可以实时监测环境的不安全情况，提前预防、实时预警、及时采取应对措施，降低灾害对人类生命财产的威胁。美国布法罗大学早在 2013 年就提出研究深海互联网项目，通过将特殊处理的感应装置置于深海处，分析水下相关情况，包括海洋污染的防治、海底资源的探测，甚至对海啸也可以提供更加可靠的预警。该项目在当地湖水中进行试验并获得成功，为进一步扩大使用范围提供了基础。利用物联网技术可以智能感知大气、土壤、森林、水资源等方面的指标数据，对于改善人类生活环境发挥巨大作用。

人工智能

5.4 人工智能

5.4.1 人工智能基本概念

应知应会

人工智能（Artificial Intelligence，AI）是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。人工智能是计算机科学的一个分支，其研究包括机器人、语音识别、图像识别、自然语言处理和专家系统等。人工智能是一门极富挑战性的交叉学科，涉及计算机知识、数学、心理学和哲学等。总而言之，人工智能研究的一个主要目标是使机器能够胜任一些通常需要人类智能才能完成的复杂工作。但不同的时代、不同的人对“复杂工作”的理解是不同的。人工智能从诞生以来，理论和技术日益成熟，应用