计算机前沿

一、 云计算：

1. 狭义定义

   一种IT基础设施的交付和使用模式，用户可以借助网络以按需使用、按量计费的方式获得各种硬件和软件资源。

2. 广义定义

   服务的交付和使用模型，用户可以借助网络获得所需的服务。

3. 维基百科定义（重要）

   云计算（英语：cloud computing），也被意译为**网络计算**，是一种基于互联网的计算方式，通过这种方式，共享的软硬件资源和信息可以按需求提供给计算机各种终端和其他设备，使用服务商提供的电脑基建作计算和资源。

4. 基本特征

   互联网上汇聚的计算资源、存储资源、数据资源和应用资源正随着互联网规模的扩大而不断增加，互联网正在从传统意义的通信平台转化为泛在、智能的计算平台。与计算机系统这样的传统计算平台比较，互联网上还没有形成类似计算机操作系统的服务环境，以支持互联网资源的有效管理和综合利用。在传统计算机中已成熟的操作系统技术，已不再能适用于互联网环境，其根本原因在于：互联网资源的自主控制、自治对等、异构多尺度等基本特性，与传统计算机系统的资源特性存在本质上的不同。为了适应互联网资源的基本特性，形成承接互联网资源和互联网应用的一体化服务环境，面向互联网计算的**虚拟计算环境**（Internet-based Virtual Computing Environment，iVCE）的研究工作，使用户能够方便、有效地共享和利用开放网络上的资源。 

   互联网上的云计算服务特征和自然界的云、水循环具有一定的相似性，因此，云是一个相当贴切的比喻。根据美国国家标准和技术研究院的定义，云计算服务应该具备以下几条特征：

   - 随需应变自助服务。
   - 随时随地用任何网络设备访问。
   - 多人共享资源池。
   - 快速重新部署灵活度。
   - 可被监控与量测的服务。

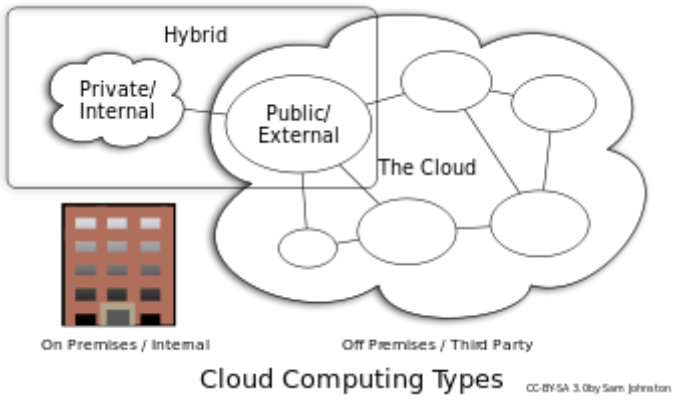
   一般认为还有如下特征：

   - 基于虚拟化技术快速部署资源或获得服务。
   - 减少用户终端的处理负担。
   - 降低了用户对于IT专业知识的依赖。

5. 美国国家标准和技术研究院的云计算定义中明确了三种服务模式：

- 软件即服务（SaaS）：消费者使用应用程序，但并不掌控操作系统、硬件或运作的网络基础架构。是一种服务观念的基础，软件服务供应商，以租赁的概念提供客户服务，而非购买，比较常见的模式是提供一组账号密码。例如：Adobe Creative Cloud，Microsoft CRM与Salesforce.com。
- 平台即服务（PaaS）：消费者使用主机操作应用程序。消费者掌控运作应用程序的环境（也拥有主机部分掌控权），但并不掌控操作系统、硬件或运作的网络基础架构。平台通常是应用程序基础架构。例如：Google App Engine。
- 基础设施即服务（IaaS）：消费者使用“基础计算资源”，如处理能力、存储空间、网络组件或中间件。消费者能掌控操作系统、存储空间、已部署的应用程序及网络组件（如防火墙、负载平衡器等），但并不掌控云基础架构。例如：Amazon AWS、Rackspace。

6. 部署模型/云种类



美国国家标准和技术研究院的云计算定义中也涉及了关于云计算的部署模型

公用云

简而言之，公用云（Public Cloud）服务可透过网络及第三方服务供应者，开放给客户使用，“公用”一词并不一定代表“免费”，但也可能代表免费或相当廉价，公用云并不表示用户资料可供任何人查看，公用云供应者通常会对用户实施使用访问控制机制，公用云作为解决方案，既有弹性，又具备成本效益。

私有云

私有云（Private Cloud）具备许多公用云环境的优点，例如弹性、适合提供服务，两者差别在于私有云服务中，资料与程序皆在组织内管理，且与公用云服务不同，不会受到网络带宽、安全疑虑、法规限制影响；此外，私有云服务让供应者及用户更能掌控云基础架构、改善安全与弹性，因为用户与网络都受到特殊限制。

社群云

社群云（Community Cloud）由众多利益相仿的组织掌控及使用，例如特定安全要求、共同宗旨等。社群成员共同使用云资料及应用程序。

混合云

混合云（Hybrid Cloud）结合公用云及私有云，这个模式中，用户通常将非企业关键信息外包，并在公用云上处理，但同时掌控企业关键服务及资料。

7. 云计算的隐私安全问题

主条目：云计算的安全性

云计算受到业界的极大推崇并推出了一系列基于云计算平台的服务。然而在用户大量参与的情况下，不可避免的出现了隐私问题。用户在云计算平台上共享信息使用服务，那么云计算平台需要收集其相关信息。实际上，云计算的核心特征之一就是数据的储存和安全完全由云计算提供商负责。对于许多用户来说，这一方面降低了组织内部和个人成本，无需搭建平台即可享受云服务。但是，一旦数据脱离内网被共享至互联网上，就无法通过物理隔离和其他手段防止隐私外泄。因此，许多的用户担心自己的隐私权会受到侵犯，其私密的信息会被泄露和使用。 云计算的隐私安全问题主要包括：

- 在未经授权的情况下，他人以不正当的方式进行数据侵入，获得用户数据。
- 政府部门或其他权利机构为达到目的对云计算平台上的信息进行检查，获取相应的资料以达到监管和控制的目的。
- 云计算提供商为获取商业利益对用户信息进行收集和处理

维基百科地址

云计算课程链接

二、 密码学：

1. 维基百科定义

密码学（英语：Cryptography）可分为古典密码学和现代密码学。在西方语文中，密码学一词源于希腊语*kryptós*“隐藏的”，和*gráphein*“书写”。

古典密码学主要关注信息的保密书写和传递，以及与其相对应的破译方法。而现代密码学不只关注信息保密问题，还同时涉及信息完整性验证（消息认证码）、信息发布的不可抵赖性（数字签名）、以及在分布式计算中产生的来源于内部和外部的攻击的所有信息安全问题。古典密码学与现代密码学的重要区别在于，古典密码学的编码和破译通常依赖于设计者和敌手的创造力与技巧，作为一种实用性艺术存在，并没有对于密码学原件的清晰定义。而现代密码学则起源于20世纪末出现的大量相关理论，这些理论使得现代密码学成为了一种可以系统而严格地学习的科学。

密码学是数学和计算机科学的分支，同时其原理大量涉及信息论。著名的密码学者罗纳德·李维斯特解释道：“密码学是关于如何在敌人存在的环境中通信”，自工程学的角度，这相当于密码学与纯数学的差异。密码学的发展促进了计算机科学，特别是在于电脑与网络安全所使用的技术，如访问控制与信息的机密性。密码学已被应用在日常生活：包括自动柜员机的芯片卡、电脑用户访问密码、电子商务等等。

2. 相关术语

直到现代以前，密码学几乎专指加密算法：将普通信息（明文）转换成难以理解的资料（密文）的过程；解密算法则是其相反的过程：由密文转换回明文；加解密包含了这两种算法，一般加密即同时指称加密与解密的技术。

加解密的具体运作由两部分决定：一个是算法，另一个是密钥。密钥是一个用于加解密算法的秘密参数，通常只有通信者拥有。历史上，密钥通常未经认证或完整性测试而被直接使用在加解密上。

密码协议是使用密码技术的通信协议。近代密码学者多认为除了传统上的加解密算法，密码协议也一样重要，两者为密码学研究的两大课题。在英文中，“cryptography”和“cryptology”都可代表密码学，前者又称密码术。但更严谨地说，前者（cryptography）指密码技术的使用，而后者（cryptology）指研究密码的学门，包含密码术与密码分析。密码分析是研究如何破解密码学的学门。但在实际使用中，通常都称密码学（即cryptography），而不具体区分其含义。

编码：它意指以码字取代特定的明文。例如，以‘苹果派’（apple pie）替换‘拂晓攻击’（attack at dawn）。编码已经不再被使用在严谨的密码学，它在信息论或通信原理上有更明确的意义。

在汉语口语中，电脑系统或网络使用的个人账户通行码也常被以密码代称，虽然通行码亦属密码学研究的范围，但学术上通行码与密码学中所称的密钥并不相同，即使两者间常有密切的关连。

3. 对称密钥加密

对称密钥加密是密码学中的一种加密法，是以转换其中一个数字、字母或仅字符串随机字母，一个秘密密钥会以特定的方式变更消息里面的文字或字母，例如更换字母相对位置（例如hello变成lohel）。只要寄件者与收件者知道秘密密钥，他们可以加密和解密并使用这个资料。

4. 公开密钥加密

公开密钥加密（也称为非对称加密）是密码学中的一种加密法，非对称密钥，是指一对加密密钥与解密密钥，某用户使用加密密钥加密后所获得的资料，只能用该用户的解密密钥才能够解密。如果知道了其中一个，并不能计算出另外一个。因此如果公开了其中一个密钥，并不会危害到另外一个。因此公开的密钥为公钥；不公开的密钥为私钥。

5. 数字签名

数字签名（又称公钥数字签名、电子签名）是一种类似写在纸上的签名，但是使用了公钥加密领域的技术实现，用于鉴别数字信息的方法。在网络上，我们可以使用“数字签名”来进行身份确认。数字签名是一个独一无二的数值，若公钥能通过验证，那我们就能确定对应的公钥的正确性，数字签名兼具这两种双重属性："可确认性"及"不可否认性（不需要笔迹专家验证）"。

6. 现代密码学

现代密码学大致可被区分为数个领域。对称密钥密码学指的是发送方与接收方都拥有相同的密钥。直到1976年这都还是唯一的公开加密法。

现代密码学重视分组密码与流密码的研究及应用。分组密码在某种意义上是阿尔伯蒂的多字符加密法的现代化。分组密码取用明文的一个区块和密钥，输出相同大小的密文区块。由于消息通常比单一区块还长，因此有了各种方式将连续的区块编织在一起。DES和AES是美国联邦政府核定的分组密码标准（AES将取代DES）。尽管将从标准上废除，DES依然很流行（三重资料加密算法变形仍然相当安全），被使用在非常多的应用上，从自动交易机、电子邮件到远程访问。也有许多其他的区块加密被发明、发布，质量与应用上各有不同，其中不乏被破解者。

流密码，相对于区块加密，制造一段任意长的密钥原料，与明文依比特或字符结合，有点类似一次一密密码本（one-time pad）。输出的流根据加密时的内部状态而定。在一些流密码上由密钥控制状态的变化。RC4是相当有名的流密码。

密码散列函数（有时称作**消息摘要函数**，杂凑函数又称**散列函数**或**哈希函数（Hash）**）不一定使用到密钥，但和许多重要的密码算法相关。它将输入资料（通常是一整份文件）输出成较短的固定长度散列值，这个过程是单向的，逆向操作难以完成，而且碰撞（两个不同的输入产生相同的散列值）发生的几率非常小。

消息认证码或押码（Message authentication codes, MACs）很类似密码散列函数，除了接收方额外使用秘密密钥来认证散列值。

公钥密码学

公开密钥密码学，简称**公钥密码学**，又称**非对称密钥密码学**，相对于**对称密钥密码学**，最大的特点在于加密和解密使用不同的密钥。

在对称密钥密码学中，加密和解密使用相同的密钥，也许对不同的消息使用不同的密钥，但都面临密钥管理的难题。由于每对通信方都必须使用异于他组的密钥，当网络成员的数量增加时，密钥数量成二次方增加。更尴尬的难题是：当安全的通道不存在于双方时，如何创建一个共有的密钥以利安全的通信？如果有通道可以安全地创建密钥，何不使用现有的通道。这个“鸡生蛋、蛋生鸡”的矛盾是长年以来密码学无法在真实世界应用的阻碍。

1976年，惠特菲尔德·迪菲与马丁·赫尔曼发表开创性的论文，提出公开密钥密码学的概念：一对不同值但数学相关的密钥，公开密钥（公钥, public key）与私密密钥（私钥, private key or secret key）。在公钥系统中，由公开密钥推算出配对的私密密钥于计算上是不可行的。历史学者David Kahn这样描述公开密钥密码学：“从文艺复兴的多字符取代法后最革命性的概念。”

在公钥系统中，公钥可以随意流传，但私钥只有该人拥有。典型的用法是，其他人用公钥来加密给该接受者，接受者使用自己的私钥解密。Diffie与Hellman也展示了如何利用公开密钥密码学来达成迪菲-赫尔曼密钥交换协议。

1978年，麻省理工学院的罗纳德·李维斯特、阿迪·萨莫尔和伦纳德·阿德曼发明另一个公开密钥系统，RSA。

直到1997年的公开文件中大众才知道，早在1970年代早期，英国情报机构政府通信总部的数学家James H. Ellis便已发明非对称密钥密码学，而且Diffie-Hellman与RSA都曾被Malcolm J. Williamson与Clifford Cocks分别发明于前。这两个最早的公钥系统提供优良的加密法基础，因而被大量使用。其他公钥系统还有Cramer-Shoup、El Gamal、以及椭圆曲线密码学等等。

除了加密外，公开密钥密码学最显著的成就是实现了数字签名。数字签名名副其实是普通签名的数字化，他们的特性都是某人可以轻易制造签名，但他人却难以仿冒。数字签名可以永久地与被签署消息结合，无法自消息上移除。数字签名大致包含两个算法：一个是签署，使用私密密钥处理消息或消息的散列值而产生签名；另一个是验证，使用公开密钥验证签名的真实性。RSA和DSA是两种最流行的数字签名机制。数字签名是公开密钥基础建设（public key infranstructures, PKI）以及许多网络安全机制（SSL/TLS, 虚拟专用网等）的基础。

公开密钥算法大多基于计算复杂度上的难题，通常来自于数论。例如，RSA源于整数因数分解问题；DSA源于离散对数问题。近年发展快速的椭圆曲线密码学则基于和椭圆曲线相关的数学难题，与离散对数相当。由于这些底层的问题多涉及模数乘法或指数运算，相对于分组密码需要更多计算资源。因此，公开密钥系统通常是复合式的，内含一个高效率的对称密钥算法，用以加密消息，再以公开密钥加密对称密钥系统所使用的密钥，以增进效率。

密码分析

密码分析又称破密术。密码分析的目的是发现密码机制的弱点，从事者可能是意图颠覆系统恶意的攻击者或评估系统弱点的设计人。在现代，密码算法与协议必须被仔细检查和测试，确定其保证的安全性。

大众普遍误解认为所有加密法都可以被破解。香农在二战时的工作就已证明只要密钥是完全随机，不重复使用，对外绝对保密，与消息等长或比消息更长的一次一密是不可能破解的。除了一次一密以外的多数加密法都可以以暴力攻击法破解，但是破解所需的努力可能是密钥长度的指数成长。

密码分析的方式有很多，因此有数个分类。一个常见的分别法则是攻击者知晓多少信息。在唯密文攻击中，密码分析者只能访问密文，好的现代密码系统对这种情况通常是免疫的。在已知明文攻击中，密码分析者可以访问多个明文、密文对。在选择明文攻击中，密码分析者可以自选任意明文，并被赋予相对应的密文，例如二战时布列颠所使用的园艺法。最后，选择密文攻击中，密码分析者可以自选任意密文，并被赋予相对应的明文

对称密钥加密的密码分析通常旨在查找比已知最佳破解法更有效率的方式。例如，以最简单的暴力法破解DES需要一个已知明文与255解密运算，尝试近半数可能的密钥。线性分析攻击法对DES需要243已知明文与243 DES运算，显然比暴力法有效。

公开密钥算法则基于多种数学难题，其中最有名的是整数分解和离散对数问题。许多公开密钥密码分析在研究如何有效率地解出这些计算问题的数值算法。例如，已知解出基于椭圆曲线的离散对数问题比相同密钥大小的整数因数分解问题更困难。因此，为了达到相等的安全强度，基于因数分解的技术必须使用更长的密钥。由于这个因素，基于椭圆曲线的公开密钥密码系统从1990年代中期后逐渐流行。

当纯粹的密码分析着眼于算法本身时，某些攻击则专注于密码设备执行的弱点，称为副通道攻击。如果密码分析者可以访问到设备执行加密或回报通行码错误的时间，它就可能使用时序攻击法破解密码。攻击者也可能研究消息的模式与长度，得出有用的信息，称为流量分析，对机敏的敌人这相当有效。当然，社会工程与其它针对人事、社交的攻击与破密术一并使用时可能是最有力的攻击法。

密码学原型

多数的密码学理论研究在探讨密码学原型：具备基本密码学特质的算法以及和其他问题的关连。例如，容易正向运算却难以逆向运算的单向函数。通常而言，密码应用如果要安全，就必须保证单向函数存在。然而，如果单向函数存在，就表示P ≠ NP。既然目前P与NP问题仍是未解，我们就无从得知单向函数是否存在。如果单向函数存在，那安全的准随机数产生器与准随机数函数就存在。 目前已知的密码学原型仅提供基本的机能。通常是机密、消息完整、认证、和不可否认。任何其他机能都是基本算法的组合与延伸，这类组合称为密码系统。例如PGP、SSH、SSL/TLS、公开密钥基础建设和数字签名等。其他密码原型还有加密算法本身、单向排列、暗门排列等。

密码协议

议题，像是交互证明、秘密分享与零知识，更复杂的有电子钞票和安全多方计算。

当一个好的密码系统的安全失效时，很少是密码学原型出现漏洞。大部分的弱点都发生于协议设计、系统实现、或是某些人为错误。许多密码学协议都在非系统化的过程中发展出来，很少有安全上的证明。一些正规分析协议安全的方式都本于数学逻辑（例如BAN逻辑）或近期的具体安全原则，这些都是数十年来研究人员的主题。很不幸的，这些工具都相当的笨重也无法用于复杂的设计。如何实现与集成密码学的应用本身是截然不同的领域，参见密码学工程与安全工程。

密码学维基百科

密码学课程链接

三、 区块链：

1. 特点

   去中心化、防篡改、公开透明、共识

2. 定义

   区块链本质就是一个去中心化的分布式账本，原来需要通过银行或第三方支付才能完成的交易，现在可以通过分布式节点组成的网络来完成。

   广义来讲，区块链是利用块链式数据结构来验证与存储数据、利用共识算法来生成和更新数据、利用密码学的方式保证数据安全、利用智能合约来编程和操作数据的一种全新的分布式基础架构与计算范式。

3. 维基百科定义

区块链（英语：blockchain或block chain）是借由密码学与共识机制等技术创建与存储庞大交易资料区块链的点对点网络系统。

每一个区块包含了前一个区块的加密散列、相应时间戳记以及交易资料（通常用默克尔树（Merkle tree）算法计算的散列值表示），这样的设计使得区块内容具有难以篡改的特性。用区块链技术所串接的分布式账本能让两方有效记录交易，且可永久查验此交易。

目前区块链技术最大的应用是数字货币，例如比特币的发明。因为支付的本质是“将账户A中减少的金额增加到账户B中”。如果人们有一本公共账簿，记录了所有的账户至今为止的所有交易，那么对于任何一个账户，人们都可以计算出它当前拥有的金额数量。而区块链恰恰是用于实现这个目的的公共账簿，其保存了全部交易记录。在比特币体系中，比特币地址相当于账户，比特币数量相当于金额。

4. 概念

以比特币的区块链账本为例。每个区块基本由上一个区块的散列值，若干条交易，一个调节数等元素构成，矿工通过工作量证明实现对交易整理为账本区块和区块安全性的维持。一个矿工通过交易广播渠道收集交易项目并打包，协议约定了区块速度生成速度而产生的难度目标值，通过不断将调节数和打包的交易数据进行散列运算而算出对应散列值使其满足当时相应的难度目标值，最先计算出调节数的矿工可以将之前获得上一个区块的散列值、交易数据、当前算出对应区块的调节数集成为一个账本区块并广播到账本发布渠道，其他矿工则可以知道新区块已生成并知道该区块的散列值（作为下一个区块的“上一个区块的散列值”），从而放弃当前待处理的区块数据生成并投入到新一轮的区块生成。

对于其他基于区块链的应用，主要是针对所负载的数据，区块安全性的维持方式等进行调整。

5. 应用项目

国家货币

a. e-Dinar是突尼斯共和国政府用区块链技术发行的数字货币。也是第一个国家数位货币。
b. eCFA是塞内加尔共和国政府用区块链技术发行的数位货币。
c. 数字人民币（E-CNY）是中华人民共和国政府用区块链技术发行的数字货币。

社区货币

a. 空卢（英文：Colu）公司在英国伦敦用区块链发行了东伦敦社区英镑（Local Pound, East London），主要为中小企业提供支付平台。2017年6月止，空卢在全球发行了多款社区货币，共有50,000用户。

私有链、公有链和联盟链的区别

	公有链	联盟链	私有链
参与者	不限	联盟成员	链的所有者

	公有链	联盟链	私有链
共识机制	PoW/PoS	分布式一致性算法	solo/pbft等
验证者	自愿提供算力或质押加密货币者	联盟成员协商确定	链的所有者
激励机制	需要	可选	无
去中心化程度	较高	偏低	极低
如初特点	解决双重支付	效率和成本优化	安全性高、效率高
吞吐量	7笔/秒至数千笔/秒（TPS）	<10万笔/秒（TPS）	视配置决定
应用领域	区块链游戏、非同质化代币、去中心化金融等	供应链管理、金融服务、医疗保健等	大型组织或私人企业之业务等
代表项目	比特币、以太坊	R3、Hyperledger	

公有链

公有链可称为公共区块链，指所有人都可以参与的区块链。换言之它是公平公开，所有人可自由访问，发送、接收、认证交易。另外公有链亦被认为是“完全去中心化”的区块链。公有链的代表有BTC区块链，ETH、EOS等，它们之间有存在不同架构。举个例子说，以太坊（ETH）是一条公有链，在以太坊链上运作的每一项应用都会消耗这条链的总体资源；EOS只是一个区块链的基础架构，开发人员可以自由地在EOS上创建公链，每条链与链之间都不会影响彼此拥有的资源，换言之不会出现因个别应用资源消耗过多而造成网络拥挤。

私有链

商业组织正在为各种应用开发分布式分类账和其他区块链启发的软件。由于这些软件被中心化机构控制，不具有区块链去中心化的属性，被称为私有链 （private blockchains）、区域链、或者联盟链。因为其应用范围与用分布式数据库处理信息的云端运算相似，2017年6月止，不但没有任何私有链得到认可和使用，而且国际银行界纷纷退出所参与的项目；加拿大中央银行也在2017年5月放弃了国家私有链的开发，主要原因是与中央银行体系格格不入。相反，用现有区块链进行ICO众筹，来开发新型去中心化社区的项目，如雨后春笋般地涌现，得到不同凡响的结果。以下是部分私有链及联盟链开发项目：

- 德勤和ConsenSys2016年宣布计划创建一个数位银行ConsenSys计划
- R3计划连接42家银行分布式分类帐，主要由以太坊，Chain.com，英特尔和IBM牵头
- Microsoft Visual Studio正在使Ethereum Solidity语言可供应用程序开发人员使用。
- SafeShare保险提供一项区域链为基础的主打共享经济的保险，由英国保险巨头劳合社承保。

- 一家瑞士工业联盟，包括瑞士电信，苏黎世州银行和瑞士股票交易所，以柜台买卖为原型的资产交易，基于以太坊科技的区域链。
- Context Labs a 2013 company developing blockchain enabled platforms
- R3区域链联盟
- Digital Asset Holdings
- Satoshi Citadel Industries
- 方舟私有链Arkblockchain一个比特币并行区域链项目，特别面向供应链、电子商务、物联网、医疗服务、政务等应用开发的高可信任私有链。
- 美国期货和期权交易所CME集团于2017年4月11日宣布，正在测试基于区域链的黄金期货平台的正处于最后测试阶段，该区块链是为比特币提供认证的对等网络。
- 台湾林产品生产追溯系统

侧链

区块链中的侧链（sidechains）实质上不是特指某个区块链，而是指遵守侧链协议的所有区块链，该名词是相对与比特币主链来说的。侧链协议是指：可以让比特币安全地从比特币主链转移到其他区块链，又可以从其他区块链安全地返回比特币主链的一种协议。

ICO代币

首次代币发行（英语：Initial Coin Offering，简称ICO），也称为ICO众筹，是用区块链筹集资金，以便开发新型区块链社区的项目。

非营利组织

- 比尔及梅琳达·盖茨基金会《基层项目 / Level One Project》旨在利用区块链技术帮助世界各地20亿缺乏银行账户的民众。
- 联合国世界粮食计划署的《区块建设 / Building Blocks》旨在使粮食计划署越来越多的现金扶贫业务更快，更便宜，更安全。“区块建设”于2017年1月在巴基斯坦开展了现场试点工作，将在整个春季继续进行。2017年6月，该项目已经扩大到叙利亚等国，计划在2030年前在全球实现零饥饿。

去中心化的社会网络

a. 回馈项目（Backfeed project）正在基于区块链分布式自治系统，开发共识主动性创建和分配价值的社会网络。
b. 亚历山大项目（The Alexandria project）是一个基于区块链开发的去中心化图书馆网络。
c. 它自主（Tezos）是一个根据它代币（token）持有者们的投票结果，让电脑程序自我演变，来实现区块链自主的开发项目。比特币区块链是一个去中心化的加密货币和支付的金融自主体系。以太坊区块链在前者的基础上增加了去中心化的智能合约的法律自主体系。它自主将在前两者的基础上增加去中心化的电脑程序开发功能，以便创建社会管理自主权体系。

6. 黑客事故

区块链目前多用于民间自定义的各种虚拟货币领域，众多黑客事件也发生在这些场景，区块链本身可以确保记账内容万无一失但目前几乎都是不记名设计，所以谁能夺取账号文本就能声称为所有者，而民间公司保存账号的服务器防骇条件不一使此类“抢劫”行为提供可能性。

2018年

- 1月，日本数字货币交易所Coincheck遭黑客攻击，约价值超过5.34亿美元的NEM于平台上被非法转移。
- 2月11日，意大利加密货币交易所BitGrail遭黑客攻击，约价值1.7亿美元的NANO被盗。
- 4月22日，BeautyChain智能合约出现重大漏洞，黑客通过此漏洞无限生成代币，导致BEC的价值接近归零。
- 4月25日，SmartMesh出现疑似重大安全漏洞，宣布暂停所有SMT交易和转账直至另行通知，导致损失约1.4亿美金。
- 7月10日，以色列数字货币交易所Bancor遭黑客攻击，约价值超过23.5亿美元的ETH，NPXS，和BNT于平台上被非法转移。
- 7月25日，EOS Fomo 3D狼人游戏的游戏合约遭受溢出攻击，60686个EOS从奖励池中被盗取，导致部分奖励没有按照游戏规则奖励用户。EOS核心仲裁论坛（ECAF）对黑客进行仲裁后，冻结黑客EOS账户：eosfomoplay1。
- 9月20日，日本数字货币交易所Zaif遭黑客攻击，导致损失67亿日元（约6000万美元加密货币），其中包括5,966比特币。根据CNN报道指出，被盗金额约4000万美元属客户资金，另外2000万则属于交易所。
- 12月3日，EOS Dice3D黑客攻击，损失10569个EOS。黑客将被盗的EOS转至火币，Dice3D官方决定自费拿出部分EOS给予玩家作补偿。

2019年

- 2月22日，EOS42被黑客攻击，黑客利用EOS节点没有更新黑名单的漏洞去攻击系统，使EOS42损失二百万个EOS。这个安全事件发生后，EOS社群开始作防备措施，避免类似情况再出现。
- 3月30日，韩国加密货币交易所Bithumb遭到黑客入侵，超过300万EOS（约1270万美元）和2000万XRP（约620万美元）的资产被盗。
- 5月8日，全球最大加密货币交易所Binance发布公告表示，遭到黑客攻击，共计7000枚比特币遭窃，损失估计超过4000万美元。
- 7月12日，日本金融厅认证的合法加密货币交易所币宝（BitPoint），遭窃取上千颗比特币，各类加密货币合计损失高达35亿日元。而币宝台湾分公司从7月23日开始也全面暂停服务，所有用户不能交易加密货币外，连台币账户都无法提领。

四、 人工智能

1. 维基百科定义

人工智能（英语：artificial intelligence，缩写为**AI**）亦称**智械、机器智能**，指由人制造出来的机器所表现出来的智能。通常人工智能是指通过普通计算机程序来呈现人类智能的技术。该词也指出研究这样的智能系统是否能够实现，以及如何实现。同时，通过医学、神经科学、机器人学及统计学等的进步，常态预测则认为人类的很多职业也逐渐被其取代。

人工智能的定义可以分为两部分，即“人工”和“智能”。“人工”即由人设计，为人创造、制造。

关于什么是“智能”，较有争议性。这涉及到其它诸如意识、自我、心灵，包括无意识的精神等等问题。人唯一了解的智能是人本身的智能，这是普遍认同的观点。但是我们对我们自身智能的理解都非常有限，对构成人的智能必要元素的了解也很有限，所以就很难定义什么是“人工”制造的“智能”。因此人工智能的研究往往涉及对人智能本身的研究。其它关于动物或其它人造系统的智能也普遍被认为是人工智能相关的研究课题。

人工智能目前在电脑领域内，得到了愈加广泛的发挥。并在机器人、经济政治决策、控制系统、仿真系统中得到应用。

2. 人工智能历史

年代	20世纪40年代	20世纪50年代	20世纪60年代	20世纪70年代	20世纪80年代	20世纪90年代	2世0年
计算机	1945 电脑（ENIAC）	1957 FORTRAN 语言					
人工智能研究		1953 博弈论 1956 达特矛斯会议		1977 知识工程宣言	1982 第五代电脑计划开始	1991 人工神经网络	20 深学
人工智能语言			1960 LISP 语言	1973 PROLOG 语言			

年代	20世纪40年代	20世纪50年代	20世纪60年代	20世纪70年代	20世纪80年代	20世纪90年代	2世
知识表达				1973 生产系统 1976 框架理论			
专家系统			1965 DENDRAL	1975 MYCIN	1980 Xcon		

3. 人工智能三大流派

符号主义(symbolicism),又称为逻辑主义(logicism)、心理学派(psychologism)或计算机学派(computerism),其原理主要为物理符号系统(即符号操作系统)假设和有限合理性原理。

连接主义(connectionism),又称为仿生学派(bionicsism)或生理学派(physiologism),其主要原理为神经网络及神经网络间的连接机制与学习算法。

行为主义(actionism),又称为进化主义(evolutionism)或控制论学派(cyberneticsism),其原理为控制论及感知-动作型控制系统。

4. AI对人类的威胁

悲观学派

此学派的代表是天文物理学家史蒂芬·霍金(Stephen Hawking)，以及特斯拉首席执行官伊隆·马斯克(Elon Musk)。霍金认为AI对人类将来有很大的威胁，主要有以下理由：

a. AI会遵循科技发展的加速度理论。
b. AI可能会有自我改造创新的能力。
c. AI进步的速度远远超过人类。
d. 人类会有被灭绝的危机存在。

乐观学派

主要是Google、Facebook等AI的主要技术发展者，他们对AI持乐观看法的理由：

a. 人类只要关掉电源就能除掉AI机器人。
b. 任何的科技都会有瓶颈，“摩尔定律”到目前也遇到相当的瓶颈，AI科技也不会无限成长，依然存在许多难以克服的瓶颈。
c. 依目前的研究方向，电脑无法突变、苏醒、产生自我意志，AI也不可能具有创意与智能、同情心与审美等这方面的能力。

5. 实际应用

机器视觉、指纹识别、人脸识别、视网膜识别、虹膜识别、掌纹识别、专家系统、自动规划、无人载具等。

6. 应用领域

- 智能控制
- 机器人学
- 自动化技术
- 语言和图像理解
- 遗传编程
- 法学信息系统
- 下棋
- 医学领域

7. 相关

- 自动驾驶汽车
- 微软小冰
- 小米小爱同学
- 天猫精灵
- Siri

8.

人工智能维基百科

五、 机器学习

1. 维基百科定义

机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。因为学习算法中涉及了大量的统计学理论，机器学习与推断统计学联系尤为密切，也被称为**统计学习理论**。

机器学习有下面几种定义：

- 机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能。
- 机器学习是对能通过经验自动改进的计算机算法的研究。
- 机器学习是用数据或以往的经验，以此优化计算机程序的性能标准

2. 分类

机器学习可以分成下面几种类别：

- 监督学习从给定的训练数据集中学习出一个函数，当新的数据到来时，可以根据这个函数预测结果。监督学习的训练集要求是包括输入和输出，也可以说是特征和目标。训练集中的目标是由人标注的。常见的监督学习算法包括回归分析和统计分类。

监督学习和非监督学习的差别就是训练集目标是否人标注。他们都有训练集且都有输入和输出

- 无监督学习与监督学习相比，训练集没有人为标注的结果。常见的无监督学习算法有生成对抗网络（GAN）、聚类。
- 半监督学习介于监督学习与无监督学习之间。
- 增强学习机器为了达成目标，随着环境的变动，而逐步调整其行为，并评估每一个行动之后所到的回馈是正向的或负向的。

3. 算法

具体的机器学习算法有：

- 构造间隔理论分布：聚类分析和模式识别
    - 人工神经网络
    - 决策树
    - 感知器
    - 支持向量机
    - 集成学习AdaBoost
    - 降维与度量学习
    - 聚类
    - 贝叶斯分类器
- 构造条件概率：回归分析和统计分类
    - 高斯过程回归
    - 线性判别分析
    - 最近邻居法
    - 径向基函数核
- 通过再生模型构造概率密度函数：
    - 最大期望算法
    - 概率图模型：包括贝叶斯网络和Markov随机场
    - Generative Topographic Mapping
- 近似推断技术：
    - 马尔可夫链
    - 蒙特卡罗方法
    - 变分法
- 最优化：大多数以上方法，直接或者间接使用最优化算法。
- 量子机器学习

六、 人工神经网络

1. 神经网络维基百科定义

人工神经网络（英语：Artificial Neural Network，ANN），简称**神经网络**（Neural Network，NN）或**类神经网络**，在机器学习和认知科学领域，是一种模仿生物神经网络（动物的中枢神经系统，特别是大脑）的结构和功能的数学模型或计算模型，用于对函数进行估计或近似。神经网络由大量的人工神经元联结进行计算。大多数情况下人工神经网络能在外界信息的基础上改变内部结构，是一种自适应系统，通俗地讲就是具备学习功能。现代神经网络是一种非线性统计性数据建模工具，神经网络通常是通过一个基于数学统计学类型的学习方法（Learning Method）得以优化，所以也是数学统计学方法的一种实际应用，通过统计学的标准数学方法我们能够得到大量的可以用函数来表达的局部结构空间，另一方面在人工智能学的人工感知领域，我们通过数学统计学的应用可以来做人工感知方面的决定问题（也就是说通过统计学的方法，人工神经网络能够类似人一样具有简单的决定能力和简单的判断能力），这种方法比起正式的逻辑学推理演算更具有优势。

和其他机器学习方法一样，神经网络已经被用于解决各种各样的问题，例如机器视觉和语音识别。这些问题都是很难被传统基于规则的编程所解决的。

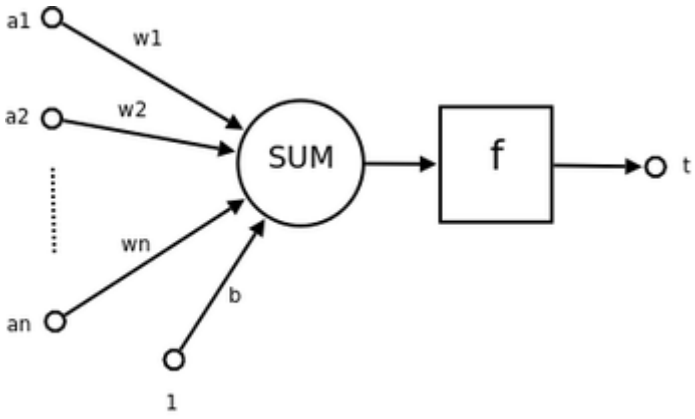
2. 构成

典型的人工神经网络具有以下三个部分：

- **结构**（**Architecture**）结构指定了网络中的变量和它们的拓扑关系。例如，神经网络中的变量可以是神经元连接的权重（weights）和神经元的激励值（activities of the neurons）。
- **激励函数**（**Activation Rule**）大部分神经网络模型具有一个短时间尺度的动力学规则，来定义神经元如何根据其他神经元的活动来改变自己的激励值。一般激励函数依赖于网络中的权重（即该网络的参数）。
- **学习规则**（**Learning Rule**）学习规则指定了网络中的权重如何随着时间推进而调整。这一般被看做是一种长时间尺度的动力学规则。一般情况下，学习规则依赖于神经元的激励值。它也可能依赖于监督者提供的目标值和当前权重的值。例如，用于手写识别的一个神经网络，有一组输入神经元。输入神经元会被输入图像的数据所激发。在激励值被加权并通过一个函数（由网络的设计者确定）后，这些神经元的激励值被传递到其他神经元。这个过程不断重复，直到输出神经元被激发。最后，输出神经元的激励值决定了识别出来的是哪个字母。

3. 神经元/感知器

神经元示意图：



- a1~an为输入向量的各个分量
- w1~wn为神经元各个突触的权值
- b为偏置
- f为传递函数，通常为非线性函数。一般有traingd(),tansig(),hardlim()。以下默认为hardlim()
- t为神经元输出

数学表示

$$t = f\left(\overrightarrow{W'}\vec{A} + b\right)$$

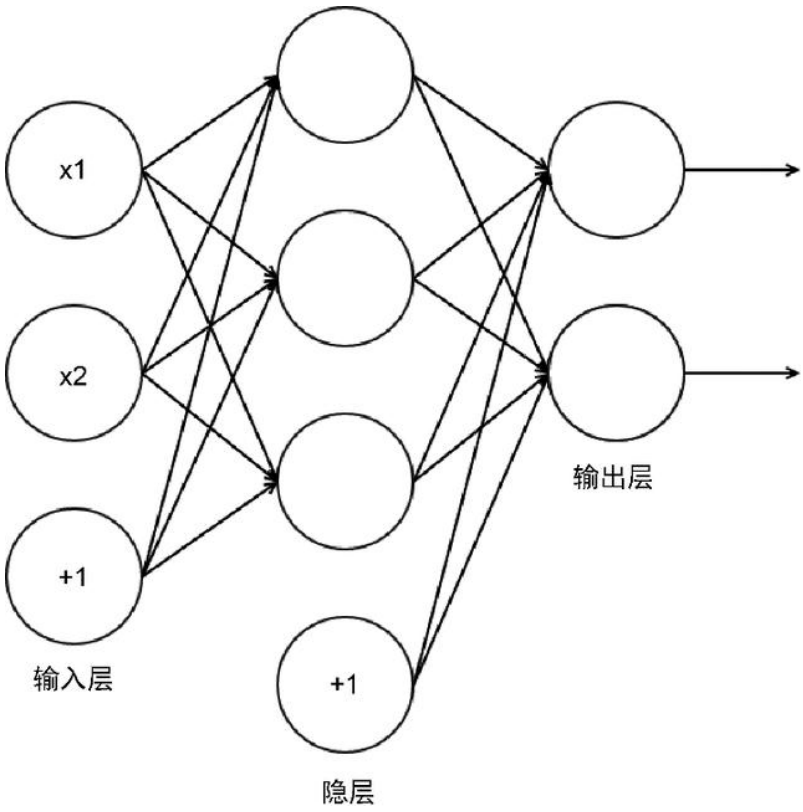
可见，一个神经元的功能是求得输入向量与权向量的内积后，经一个非线性传递函数得到一个标量结果。

单个神经元的作用：把一个n维向量空间用一个超平面分割成两部分（称之为判断边界），给定一个输入向量，神经元可以判断出这个向量位于超平面的哪一边。

该超平面的方程：

$$\overrightarrow{W'}\vec{p} + b = 0$$

4. 基本结构



一种常见的多层结构的前馈网络（Multilayer Feedforward Network）由三部分组成，

- 输入层（Input layer），众多神经元（Neuron）接受大量非线形输入消息。输入的消息称为输入向量。
- 输出层（Output layer），消息在神经元链接中传输、分析、权衡，形成输出结果。输出的消息称为输出向量。
- 隐藏层（Hidden layer），简称“隐层”，是输入层和输出层之间众多神经元和链接组成的各个层面。隐层可以有一层或多层。隐层的节点（神经元）数目不定，但数目越多神经网络的非线性越显著，从而神经网络的强健性（控制系统在一定结构、大小等的参数摄动下，维持某些性能的特性）更显著。习惯上会选输入节点1.2至1.5倍的节点。

  这种网络一般称为感知器（对单隐藏层）或多层感知器（对多隐藏层），神经网络的类型已经演变出很多种，这种分层的结构也并不是对所有的神经网络都适用。

5. 学习过程

通过训练样本的校正，对各个层的权重进行校正（learning）而创建模型的过程，称为自动学习过程（training algorithm）。具体的学习方法则因网络结构和模型不同而不同，常用反向传播算法（Backpropagation/倒传递/逆传播，以output利用一次微分Delta rule来修正weight）来验证。

6. 种类

人工神經网络分类为以下两种：

1.依学习策略（Algorithm）分类主要有：

- 监督式学习网络（Supervised Learning Network）为主
- 无监督式学习网络（Unsupervised Learning Network）
- 混合式学习网络（Hybrid Learning Network）
- 联想式学习网络（Associate Learning Network）
- 最适化学习网络（Optimization Application Network）

2.依网络架构（Connectionism）分类主要有：

- 前馈神经网络（Feed Forward Network）
- 循环神经网络（Recurrent Network）
- 强化式架构（Reinforcement Network）

人工神经网络维基百科

七、 深度学习

1. 深度学习维基百科定义

深度学习（英语：deep learning）是机器学习的分支，是一种以人工神经网络为架构，对资料进行表征学习的算法。

深度学习是机器学习中一种基于对数据进行表征学习的算法。观测值（例如一幅图像）可以使用多种方式来表示，如每个像素强度值的向量，或者更抽象地表示成一系列边、特定形状的区域等。而使用某些特定的表示方法更容易从实例中学习任务（例如，人脸识别或面部表情识别）。深度学习的好处是用非监督式或半监督式的特征学习和分层特征提取高效算法来替代手工获取特征。

表征学习的目标是寻求更好的表示方法并创建更好的模型来从大规模未标记数据中学习这些表示方法。表示方法来自神经科学，并松散地创建在类似神经系统中的信息处理和对通信模式的理解上，如神经编码，试图定义拉动神经元的反应之间的关系以及大脑中的神经元的电活动之间的关系。

至今已有数种深度学习框架，如深度神经网络、卷积神经网络和深度置信网络和循环神经网络已被应用在计算机视觉、语音识别、自然语言处理、音频识别与生物信息学等领域并获取了极好的效果。

（在机器学习中，**特征学习**或**表征学习**是学习一个特征的技术的集合：将原始数据转换成为能够被机器学习来有效开发的一种形式。它避免了手动提取特征的麻烦，允许计算机学习使用特征的同时，也学习如何提取特征：学习如何学习。）

深度学习维基百科

八、 物联网

1. 物联网维基百科定义

物联网（英语：Internet of Things，简称**IoT**）是一种计算设备、机械、数字机器相互关系的系统，具备通用唯一识别码（UID），并具有通过网络传输数据的能力，无需人与人、或是人与设备的交互。

物联网将现实世界数字化，应用范围十分广泛。物联网可拉近分散的资料，统整物与物的数字信息。物联网的应用领域主要包括以下方面：运输和物流、工业制造、健康医疗、智能环境（家庭、办公、工厂）、个人和社会领域等。

物联网为受各界瞩目的新兴领域，但安全性是物联网应用受到各界质疑的主要因素，主要的质疑在于物联网技术正在快速发展中，但其中涉及的安全性挑战，与可能需要的法规变更等，目前均相当欠缺。

2. 技术

技术路线

技术路线（Technology Roadmap）指对于技术未来发展方向的预测。在物联网领域，广泛被各国政府与机构引用的技术路线为顾问公司SRI Consulting描绘之物联网技术路线，其依据时间轴可分为四个阶段：供应链辅助、垂直市场应用、无所不在的寻址（Ubiquitous positioning），最后可以达到“The Physical Web”（意即让物联网上的每一个智能设备都以URL来标示）。

架构

物联网的架构一般分为三层或四层。三层之架构由底层至上层依序为感测层、网络层与应用层；四层之架构由底层至上层依序为感知设备层（或称感测层）、网络连接层（或称网络层）、平台工具层与应用服务层。三层与四层架构之差异，在于四层将三层之“应用层”拆分成“平台工具层”与“应用服务层”，对于软件应用做更细致的区分。

感测层

寻址资源

物联网的实现，需要给每一个连上物联网的对象分配唯一的标识或地址。最早的概念是由无线射频识别标签和电子产品代码所发展出来的。现在物联网与互联网链接后，由于预估需要大量的IP地址，目前主流的IPv4地址空间有限，因此物联网中的对象倾向使用下一代互联网协议（IPv6），以提供足够的地址空间，IPv6对于物联网的发展扮演重要角色。

网络层

物联网有多种联网技术可供选择，依照有效传输距离可区分为短距离无线、中距离无线、长距离无线，以及有线技术：

短距离无线

- 蓝牙网状网络（Bluetooth mesh networking）– 规范采用蓝牙技术的网状网络，可增加节点数，并提供标准化的应用层。

- 光照上网技术（Li-Fi）– 与Wi-Fi标准相似的无线通信技术，但使用可见光通信以增加带宽。
- 近场通信（Near-field communication，NFC）– 使两个电子设备能够在4公分范围内进行通信的通信协议。
- 无线射频识别（Radio-frequency identification，RFID）– 使用电磁场访问无线射频识别（RFID）标签中数据的技术。
- Wi-Fi – 基于IEEE 802.11标准的无线局域网技术。
- ZigBee – 基于IEEE 802.15.4标准的个人网通信协议，具有低功耗，低数据速率，低成本的特性。
- Z-Wave – 主要应用于智能家庭和安全应用的无线通信协议。

中距离无线

- 高级长期演进技术（LTE-Advanced）– 高速蜂窝网络的通信规范。通过扩展的覆盖范围，提供更高的数据传输量和更低的延迟。
- 5G - 新一代移动通信技术，提供高资料速率、减少延迟、节省能源、提高系统容量和大规模设备连接。

长距离无线

- 低功率广域网（Low-Power Wide-Area Network，LPWAN）– 提供低资料速率与远程通信，降低功耗和传输成本。可用的LPWAN技术和协议分为使用授权频段的NB-IoT，以及使用非授权频段的LoRa、Sigfox、Weightless、Random Phase Multiple Access（RPMA）、IEEE 802.11ah等。
- 甚小孔径终端（Very Small Aperture Terminal, VSAT）– 使用小型碟型天线，透过人造卫星传输之通信技术。

有线

- 以太网（Ethernet）– 基于IEEE 802.3标准的技术，可使用双绞线、光纤连接至集线器或网络交换器。
- 电力线通信（Power Line Communication，PLC）– 以电缆传输电力和数据的通信技术，有HomePlug或G.hn等标准。

应用层

应用层在物联网四层架构中可再细分为“平台工具层”与“应用服务层”。平台工具层为底层的软件平台，作为应用服务层与网络层的接口，以支持各类的软件应用。可归类于“平台工具层”包括大数据、区块链、软件定义网络、软件定义存储、软件定义数据中心、安全通信、杀毒软件、人工智能相关（如自然语言处理、深度学习、语音识别、模式识别、电脑视觉...）等；应用服务层针对不同的应用需求，直接呈现原始资料，或经过加值处理，借由人机界面提供用户，或是对应的硬件/软件目标得到想要的信息。可归类于“应用服务层”包括虚

拟现实/增强现实、人机交互、服务导向架构、永续发展相关（生命周期评估、节能、碳足迹...）等。

在应用层中，通常使用多种编程语言撰写应用程序，使用HTTPS与OAuth之协议。在平台后端使用各种形式的数据库系统，例如时间序列数据或是后端数据存储系统（如Cassandra、PostgreSQL等）。

大多数的物联网系统均是建构在云计算之上，在云当中具备事件队列（event queuing）与消息传递系统，这些系统可以处理在各层级中所需要的通信。一些专家将工业物联网（IIoT）中的三层分类为边缘、平台和企业，它们分别透过邻近网络、接入网络和服务网络来连接。

美国国家标准暨技术研究院（NIST）对于云计算的定义中，将服务模式分为软件即服务（SaaS）、平台即服务（PaaS）、基础设施即服务（IaaS）三种。

智能物联网（AIoT）

智能物联网（AIoT）为物联网与人工智能的结合，以实现更高效率的物联网运作，改善人机交流、增强数据管理和分析。人工智能可用于将物联网数据转化为有用的信息，以改善决策流程，从而为“物联网资料即服务”（IoT Data as a Service，IoTDaaS）的模式奠定基础。

智能物联网的出现，对于物联网与人工智能两者均会产生变革，增加彼此之间的价值。因为人工智能通过机器学习功能，使得物联网变得更有价值；而物联网通过连接、信号和数据交换，使得人工智能可以获得更丰富的数据源。随着物联网遍及许多行业，将有越来越多的人为的、以及机器生成的非结构化资料，智能物联网可在资料分析中提供有力的支持，在各行各业中创造新的价值。

3. 应用

消费者应用

Google Nest的自动调温器，可报告能源使用和当地天气情况

August Home公司的智能门锁，支持HomeKit、Google个人助理、Amazon Alexa等多平台

LG的智能冰箱Internet Digital DIOS

有越来越多的物联网设备可供消费者选用，包括联网的车辆、家庭自动化、联网的可穿戴设备、联网的健康监控设备，以及远程监控设备。

苹果公司的HomeKit为该公司之智能家庭平台，用户可以透过iPhone、iPad、Apple Watch等设备的APP接口，或是由Siri语音控制支持Apple HomeKit标准的家用设备，如电视、电灯、空调、水龙头等，目前支持28类设备。其他类似、但功能与范围不尽相同的产品包括Google的Google Nest与Google个人助理、Amazon的Amazon Echo与Amazon Alexa、三星的SmartThings、小米的小爱同学、联想的Lenovo Smart Assistant等。另外还有一些开放平台如OpenHAB、Domoticz等。

另一项主要的应用为辅助老年人与残疾人士，例如语音控制可以帮助行动不便人士，警报系统可以连接至听障人士的人工耳蜗，另外还有监控跌倒或癫痫等紧急情况的传感器，这些智能家庭技术可以提供用户更多的自由和更高的生活质量。

工业应用

物联网在工业的应用称为工业物联网（Industrial internet of things，IIoT）。工业物联网专注于机器对机器（Machine to Machine，M2M）的通信，利用大数据、人工智能、云计算等技术，让工业运作有更高的效率和可靠度。工业物联网涵盖了整个工业应用，包括了机器人、医疗设备和软件定义生产流程等，为第四次工业革命中，产业转型至工业4.0中不可或缺的一部分。

大数据分析在生产设备的预防性维护中扮演关键角色，其核心为网宇实体系统。可透过5C“连接（Connection）、转换（Conversion）、联网（Cyber），认知（Cognition）、配置（Configuration）”之架构来设计网宇实体系统，将收集来的数据转化为有用的资料，并藉以优化生产流程。

农业应用

物联网在农业中的应用包括收集温度、降水、湿度、风速、病虫害和土壤成分的数据，并加以分析与运用。这样的方式称为精准农业，其利用决策支持系统，将收集来的数据做出精准分析，藉以提高产出的质量和数量，并减少浪费。

2018年8月，丰田通商与微软、近畿大学水产研究所合作，利用Microsoft Azure的物联网应用包，开发出于水产养殖辅助系统。水产养殖为劳力密集的工作，鱼苗必须由人工进行分类，以确保每条鱼的大小适当且无畸形。借由辅助系统的导入，可以大幅减轻人力负担，将有经验的人移至更高附加价值的工作。

商业应用

医疗保健

医疗物联网（Internet of Medical Things，IoMT）为物联网应用于医疗保健，包括数据收集、分析、研究与监控方面的应用，用以创建数字化的医疗保健系统。物联网设备可用于激活远程健康监控和紧急情况通知系统，包括简易的设施如血压计、便携式生理监控器，至可监测植入人体的设备，如心律调节器、人工耳蜗等[。世界卫生组织规划利用移动设备收集医疗保健数据，并进行统计、分析，创建“m-health”体系。

由于塑料与电子纺织品制造技术的进步，使得一次性使用的IoMT传感器已达到相当低的成本。对于即时医疗诊断应用的创建，可携性与低系统复杂性是不可或缺的要素。物联网在医疗保健的应用，于监测慢性病、以及疾病的预防和控制中产生很大的功用，透过远程监控，医院与卫生相关机构可以获得患者的数据，并可做进一步分析。

交通

物联网可以帮助集成通信、控制与信息处理。物联网的应用可以扩展至运输系统个层面，包括载具、基础设施，以及驾驶人。物联网组件之间的信息传递，使得载具内以及不同载具之间可以互相通信，达成智能交通灯号、智能停车、电子道路收费系统、物流和车队管理、主动巡航控制系统，以及安全和道路辅助等应用。

例如，在物流和车队管理中，物联网平台可以通过无线传感器持续监控货物和资产的位置和状况，并在发生异常事件（延迟、损坏、失窃等）时发送特定警报。这必须借助物联网与设备之间的无缝连接才可能实现。利用GPS、湿度、温度等传感器将数据发送至物联网平台，随后对数据进行分析，并将结果发送给用户。如此，用户可以跟踪载具的即时状态，并做出适当的处置。如果与机器学习结合，还可以进行驾驶睡意侦测，以及提供自动驾驶汽车等来帮助减少交通事故。

基础设施应用

物联网在基础设施的运用主要在监控与控制各类基础设施，例如铁轨、桥梁，海上与陆上的风力发电厂、废弃物管理等。透过监控任何事件或结构状况的变化，以便高效地安排维修和保养活动。

目前全球有数个大规模部署的案例正在进行中，例如韩国松岛国际都市。这是一座设备齐全的智慧城市，对于能源使用、交通流量进行精密的控制，各家户垃圾透过管道集中至废物处理中心，然后在这里进行自动分类，与再回收利用。截至2018年6月约70%的商业区已竣工。

西班牙桑坦德为另一个应用案例。这一座人口约18万的都市，安装了超过两万个传感器，主要应用于三方面：(1) 交通：透过手机APP可以即时获得停车位信息，并引导至该处停车；(2) H2O 2.0：可即时获得用水信息；(3) 公园智能空间：可随温度、湿度调整洒水系统，并检查公园内垃圾桶的垃圾量。

军事应用

军事物联网（Internet of Military Things，IoMT）是物联网在军事领域中的应用，目的是侦察、监控与战斗有关的目标，主要受到未来将于城市环境中战斗影响。军事物联网相关领域包括传感器、车辆、机器人、武器、可穿戴式智能产品，以及在战场上相关智能技术的使用。

战地物联网（The Internet of Battlefield Things，IoBT）是一个美国陆军研究实验室（ARL）的研究项目，着重研究与物联网相关的基础科学，以增强陆军士兵的能力。2017年，ARL启动了战地物联网协作研究联盟（Internet of Battlefield Things Collaborative Research Alliance，IoBT-CRA），创建了产业、大学和陆军研究人员之间的工作合作关系，以推展物联网技术及其在陆军作战中的应用的理论基础。

4. 批评、问题与争议

安全性

安全性是物联网应用受到各界质疑的主要因素，质疑之处在于物联网技术正在快速发展中，但其中涉及的安全性挑战，与可能需要的法规变更等，目前均相当欠缺。

物联网面对的大多数技术安全问题类似于一般服务器、工作站与智能手机，包括密码太短、忘记更改密码的默认值、设备之间传输采用未加密信号、SQL注入、未将软件更新至最新版本等。另外，由于多数物联网设备计算能力相当有限，无法使用常见的安全措施例如防火墙、或是高强度的密码；许多物联网设备因为价格低廉，因此无法有人力与经费支持，将软件更新至最新版本。

安全性较差的物联网设备可能被当作跳板以攻击其他设备。2016年时发生恶意程序Mirai（辞源：日文“未来”）感染物联网设备，以分布式拒绝服务攻击（DDoS）攻击DNS服务器与许多网站。在20小时内，Mirai感染了大约65,000台物联网设备，最终感染数量为20~30万台。感染设备之国家分布以巴西、哥伦比亚和越南居前三位，设备包括数字视频录影机、网络监控摄影机、路由器、打印机等，以厂商区分依序为大华股份、华为、中兴通信、思科、合勤。2017年5月，Cloudflare的计算机科学家Junade Ali指出，由于发布/订阅（Publish–subscribe pattern）的不当设计，许多物联网设备存在DDoS漏洞。利用这些漏洞的将物联网设备作为跳板的攻击，是互联网服务的真正威胁。

产业界对各界质疑安全性问题做出了回应，“物联网安全基金会”（IoTSF）于2015年9月23日成立，期借由倡导知识与最佳实践使得物联网更加安全。此外，一些公司也推出创新解决方案，以确保物联网设备的安全性。2017年，Mozilla公司推出了“Project Things”，该项目可以通过安全的“Web of Things”网关与物联网设备创建加密连线[88]。美国信息安全专家布鲁斯·施奈尔（Bruce Schneier）认为将物联网纳入政府监管业务是有必要的，以确保产业界生产的物联网设备可以遵守安全规范，以及出事的时候有人负责。

平台分散

物联网的一大问题为平台分散、跨平台之可操作性低，以及欠缺通用技术标准。物联网设备种类繁多，以及硬件与在其上运作的软件之间的差异，使得开发系统时，各应用程序保持一致变得很困难。

物联网无定形（amorphous）的计算特性往往会造成安全性问题，因为在核心操作系统中发现的错误修补，通常无法涵盖较早期且入门级的设备，一组研究人员表示，设备供应商未能通过补丁和更新支持较旧的设备，导致超过87%的现行Android设备容易受到攻击。

物联网维基百科

九、网络安全

1. 网络安全维基百科定义

网络安全（英语：network security）包含网络设备安全、网络信息安全、网络软件安全。

黑客通过基于网络的入侵来达到窃取敏感信息的目的，也有人以基于网络的攻击见长，被人收买通过网络来攻击商业竞争对手企业，造成网络企业无法正常营运，网络安全就是为了防范这种信息盗窃和商业竞争攻击所采取的措施。

2. 移动代码

移动代码（Mobile code）是一种软件技术可由远程系统透过另一个网络转存入本机端进行代理作业，可进行下载或在本机端上执行没有明确安装或者接受者的作业。移动代码的例子包括嵌入型脚本（JavaScript、VBScript）、Java小应用程序、ActiveX 控制、flash动画，并且在一般文书文件资料内嵌入。

移动代码也能透过电子邮件方式自动下载并且在客户端执行。移动代码透过电子邮件下载可能附件(例如，大总之文件) 或者透过一个HTML电子邮件内容(例如JavaScript)。例如，ILOVEYOU、TRUELOVE和AnnaK电子邮件电脑病毒/蠕虫病毒全部被作为移动代码实现(VBScript，在Windows为主机写稿子过程中执行的一个.vbs电子邮件附件里)。

在几乎所有现实状况中，用户不会意识到移动代码正下载和在他们的本机电脑中执行。

网络安全维基百科

十、模式识别

1. 模式识别维基百科定义

模式识别（英语：Pattern recognition），就是通过计算机用数学技术方法来研究模式的自动处理和判读。我们把环境与客体统称为“模式”。随着计算机技术的发展，人类有可能研究复杂的信息处理过程。信息处理过程的一个重要形式是生命体对环境及客体的识别。对人类来说，特别重要的是对光学信息（通过视觉器官来获得）和声学信息（通过听觉器官来获得）的识别。这是模式识别的两个重要方面。市场上可见到的代表性产品有光学字符识别、语音识别系统。

计算机识别的显著特点是速度快、准确性高、效率高，在将来完全可以取代人工录入。

识别过程与人类的学习过程相似。以光学字符识别之“汉字识别”为例：首先将汉字图像进行处理，抽取主要表达特征并将特征与汉字的代码存在计算机中。就像老师教我们“这个字叫什么、如何写”记在大脑中。这一过程叫做“训练”。识别过程就是将输入的汉字图像经处理后与计算机中的所有字进行比较，找出最相近的字就是识别结果。这一过程叫做“匹配”。

2. 应用领域

- 计算机视觉
    - 医学影像分析
    - 光学文字识别
- 自然语言处理
    - 语音识别
    - 手写识别
- 生物特征识别
    - 人脸识别
    - 指纹识别
    - 虹膜识别
- 文件分类

- 互联网搜索引擎
- 信用评分
- 测绘学
    - 摄影测量与遥感学

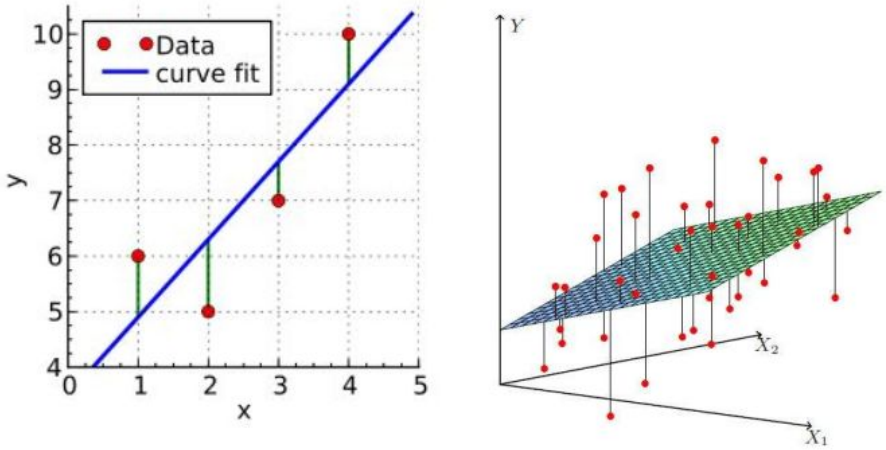
模式识别维基百科

十一、 评价指标(模式识别与机器学习评价指标类似)

1 回归（Regression）算法指标

- Mean Absolute Error 平均绝对误差
- Mean Squared Error 均方误差
- Root Mean Squared Error：均方根误差

  以下为一元变量和二元变量的线性回归示意图：



怎样来衡量回归模型的好坏呢？ 我们第一眼自然而然会想到采用残差（实际值与预测值差值）的均值来衡量，即：

$$\text{residual}(y, \hat{y}) = \frac{1}{m}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

问题 1：用残差的均值合理吗？

当实际值分布在拟合曲线两侧时，对于不同样本而言 有正有负，相互抵消，因此我们想到采用预测值和真实值之间的距离来衡量。

1.1 平均绝对误差 MAE

平均绝对误差MAE（Mean Absolute Error）又被称为 L1范数损失。

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{m}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

问题 2：MAE有哪些不足？

MAE虽能较好衡量回归模型的好坏，但是绝对值的存在导致函数不光滑，在某些点上不能求导，可以考虑将绝对值改为残差的平方，这就是均方误差。

1.2 均方误差 MSE

均方误差MSE（Mean Squared Error）又被称为 L2范数损失 。

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

问题 3： 还有没有比MSE更合理一些的指标？

由于MSE与我们的目标变量的量纲不一致，为了保证量纲一致性，我们需要对MSE进行开方 。

1.3 均方根误差 RMSE

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$$

问题 4： RMSE有没有不足的地方？有没有规范化（无量纲化的指标）？

上面的几种衡量标准的取值大小与具体的应用场景有关系，很难定义统一的规则来衡量模型的好坏。比如说利用机器学习算法预测上海的房价RMSE在2000元，我们是可以接受的，但是当四五线城市的房价RMSE为2000元，我们还可以接受吗？下面介绍的决定系数就是一个无量纲化的指标。

2 分类（Classification）算法指标

- 精度 Accuracy
- 混淆矩阵 Confusion Matrix
- 准确率（查准率） Precision
- 召回率（查全率）Recall
- AUC Area Under Curve

2.1 精度 Acc

预测正确的样本的占总样本的比例，取值范围为[0,1]，取值越大，模型预测能力越好。

$$\mathrm{Acc}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{m}\mathrm{sign}(\hat{y}_i, y_i)$$

其中：

$$\mathrm{sign}(\hat{y}_i, y_i) = \begin{cases} 1 & \hat{y}_i = y_i \\ 0 & \hat{y}_i \neq y_i \end{cases}$$

精度评价指标对平等对待每个类别，即每一个样本判对 (0) 和判错 (1) 的代价都是一样的。

问题 6：精度有什么缺陷？什么时候精度指标会失效？

- 对于有倾向性的问题，往往不能用精度指标来衡量。
- 比如，判断空中的飞行物是导弹还是其他飞行物，很显然为了减少损失，我们更倾向于相信是导弹而采用相应的防护措施。此时判断为导弹实际上是其他飞行物与判断为其他飞行物实际上是导弹这两种情况的重要性是不一样的；
- 对于样本类别数量严重不均衡的情况，也不能用精度指标来衡量。
- 比如银行客户样本中好客户990个，坏客户10个。如果一个模型直接把所有客户都判断为好客户，得到精度为99%，但这显然是没有意义的。

对于以上两种情况，单纯根据Accuracy来衡量算法的优劣已经失效。这个时候就需要对目标变量的真实值和预测值做更深入的分析。

2.2 混淆矩阵 Confusion Matrix

	预测值 正	预测值 负
真实值 正	TP	FN
真实值 负	FP	TN

这里牵扯到三个方面：真实值，预测值，预测值和真实值之间的关系，其中任意两个方面都可以确定第三个。

通常取预测值和真实值之间的关系、预测值对矩阵进行划分：

- True positive (TP)
- 真实值为Positive，预测正确（预测值为Positive）
- True negative (TN)
- 真实值为Negative，预测正确（预测值为Negative）
- False positive (FP)
- 真实值为Negative，预测错误（预测值为Positive），第一类错误，Type I error。
- False negative (FN)
- 真实值为Positive，预测错误（预测值为 Negative），第二类错误，Type II error。

2.3 准确率（查准率） Precision

Precision 是分类器预测的正样本中预测正确的比例，取值范围为[0,1]，取值越大，模型预测能力越好。

$$\mathrm{P} = \frac{TP}{TP + FP}$$

2.4 召回率（查全率）Recall

Recall 是分类器所预测正确的正样本占所有正样本的比例，取值范围为[0,1]，取值越大，模型预测能力越好。

$$\mathbf{R} = \frac{TP}{TP + FN}$$

应用场景：

1. 地震的预测 对于地震的预测，我们希望的是Recall非常高，也就是说每次地震我们都希望预测出来。这个时候我们可以牺 牲Precision。情愿发出1000次警报，把10次地震都预测正确了；也不要预测100次对了8次漏了两次。

- “宁错拿一万，不放过一个”，分类阈值较低

1. 嫌疑人定罪 基于不错怪一个好人的原则，对于嫌疑人的定罪我们希望是非常准确的。即使有时候放过了一些罪犯，但也 是值得的。因此我们希望有较高的Precision值，可以合理地牺牲Recall。

- “宁放过一万，不错拿一个”，“疑罪从无”，分类阈值较高

问题 7：某一家互联网金融公司风控部门的主要工作是利用机器模型抓取坏客户。互联网金融公司要扩大业务量，尽量多的吸引好客户，此时风控部门该怎样调整Recall和Precision？如果公司坏账扩大，公司缩紧业务，尽可能抓住更多的坏客户，此时风控部门该怎样调整Recall和Precision？

如果互联网公司要扩大业务量，为了减少好客户的误抓率，保证吸引更多的好客户，风控部门就会提高阈值，从而提高模型的查准率Precision，同时，导致查全率Recall下降。如果公司要缩紧业务，尽可能抓住更多的坏客户，风控部门就会降低阈值，从而提高模型的查全率Recall，但是这样会导致一部分好客户误抓，从而降低模型的查准率 Precision。

根据以上几个案，我们知道随着阈值的变化Recall和Precision往往会向着反方向变化，这种规律很难满足我们的期望，即Recall和Precision同时增大。

问题 8：有没有什么方法权衡Recall和Precision 的矛盾？

我们可以用一个指标来统一Recall和Precision的矛盾，即利用Recall和Precision的加权调和平均值作为衡量标准。

2.6 ROC 和 AUC

AUC是一种模型分类指标，且仅仅是二分类模型的评价指标。AUC是Area Under Curve的简称，那么Curve就是 ROC（Receiver Operating Characteristic），翻译为"接受者操作特性曲线"。也就是说ROC是一条曲线，AUC是 一个面积值。

2.6.1 ROC

ROC曲线为 FPR 与 TPR 之间的关系曲线，这个组合以 FPR 对 TPR，即是以代价 (costs) 对收益 (benefits)，显然收益越高，代价越低，模型的性能就越好。

- x 轴为假阳性率（FPR）：在所有的负样本中，分类器预测错误的比例

$$FPR = \frac{FP}{FP + TN}$$

- y 轴为真阳性率（TPR）：在所有的正样本中，分类器预测正确的比例（等于Recall）

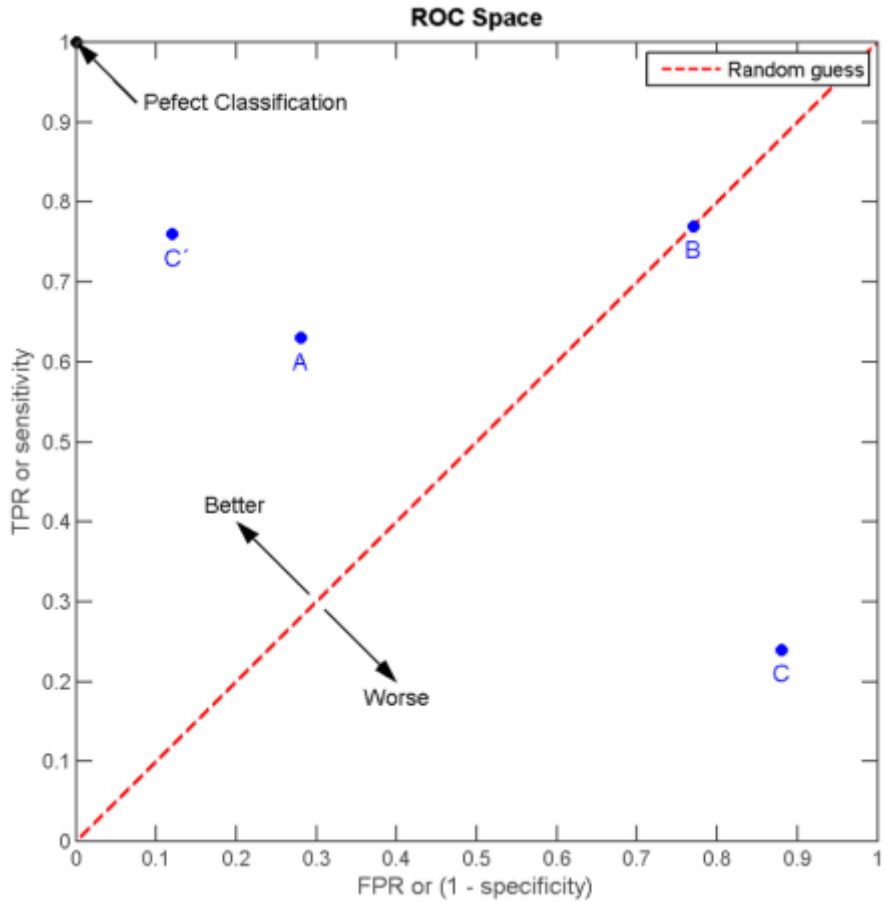
$$TPR = \frac{TP}{TP + FN}$$

为了更好地理解ROC曲线，我们使用具体的实例来说明：

如在医学诊断的主要任务是尽量把生病的人群都找出来，也就是TPR越高越好。而尽量降低没病误诊为有病的人数，也就是FPR越低越好。

不难发现，这两个指标之间是相互制约的。如果某个医生对于有病的症状比较敏感，稍微的小症状都判断为有病，那么他的TPR应该会很高，但是FPR也就相应地变高。最极端的情况下，他把所有的样本都看做有病，那么TPR达到1，FPR也为1。

我们以FPR为横轴，TPR为纵轴，得到如下ROC空间：

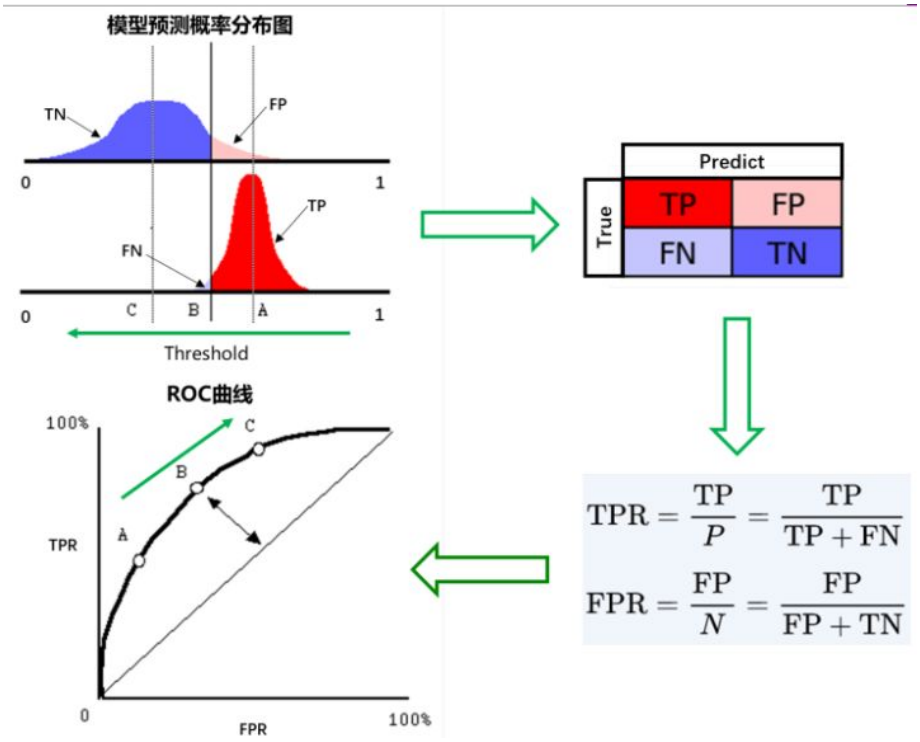


我们可以看出，左上角的点(TPR=1，FPR=0)，为完美分类，也就是这个医生医术高明，诊断全对。点A(TPR>FPR),医生A的判断大体是正确的。中线上的点B(TPR=FPR),也就是医生B全都是蒙的，蒙对一半，蒙错一半；下半平面的点C(TPR<FPR)，这个医生说你有病，那么你很可能没有病，医生C的话我们要反着听，为真庸医。上图中一个阈值，得到一个点。现在我们需要一个独立于阈值的评价指标来衡量这个医生的医术如何，也就是遍历所有的阈值,得到 ROC 曲线。

假设下图是某医生的诊断统计图，为未得病人群（上图）和得病人群（下图）的模型输出概率分布图（横坐标表示模型输出概率，纵坐标表示概率对应的人群的数量），显然未得病人群的概率值普遍低于得病人群的输出概率值（即正常人诊断出疾病的概率小于得病人群诊断出疾病的概率）。

竖线代表阈值。显然，图中给出了某个阈值对应的混淆矩阵，通过改变不同的阈值，得到一系列的混淆矩阵，进而得到一系列的TPR和FPR，绘制出ROC曲线。

阈值为1时，不管你什么症状，医生均未诊断出疾病（预测值都为N），此时，位于左下。阈值为 0 时，不管你什么症状，医生都诊断结果都是得病（预测值都为P），此时，位于右上。



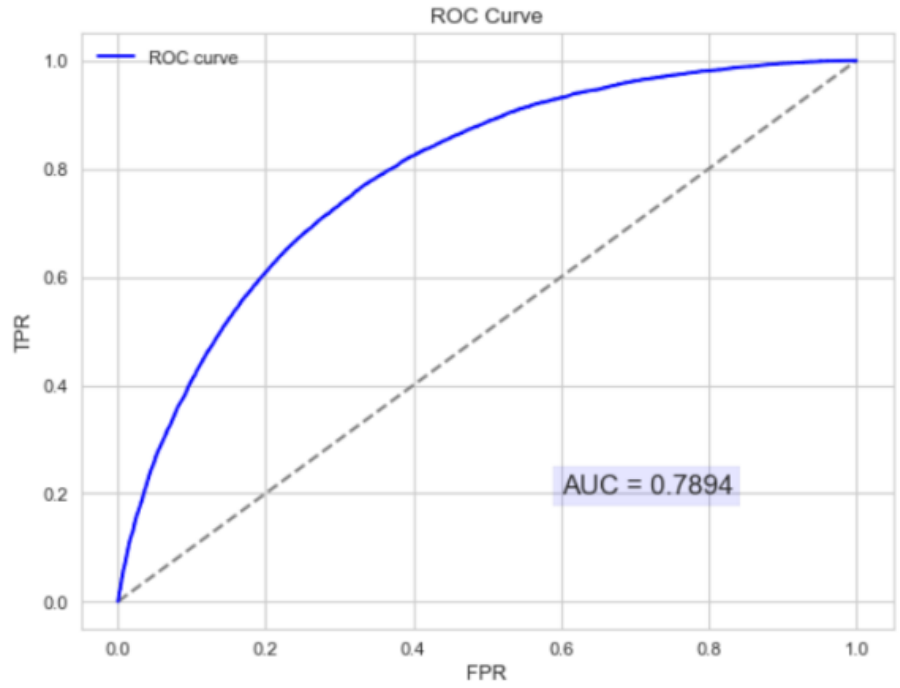
2.6.2 AUC

AUC定义：

- AUC 值为 ROC 曲线所覆盖的区域面积，显然，AUC越大，分类器分类效果越好。
- AUC = 1，是完美分类器。
- 0.5 < AUC < 1，优于随机猜测。有预测价值。
- AUC = 0.5，跟随机猜测一样（例：丢铜板），没有预测价值。
- AUC < 0.5，比随机猜测还差；但只要总是反预测而行，就优于随机猜测。

注：对于AUC小于0.5的模型，我们可以考虑取反（模型预测为positive，那我们就取negtive），这样就可以保证模型的性能不可能比随机猜测差。

以下为ROC曲线和AUC值的实例：



AUC的物理意义 AUC的物理意义正样本的预测结果大于负样本的预测结果的概率。所以AUC反应的是分类器对样本的排序能力。另外值得注意的是，AUC对样本类别是否均衡并不敏感，这也是不均衡样本通常用AUC评价分类器性能的一个原因。

问题 13 ：小明一家四口，小明5岁，姐姐10岁，爸爸35岁，妈妈33岁，建立一个逻辑回归分类器，来预测小明家人为成年人概率。

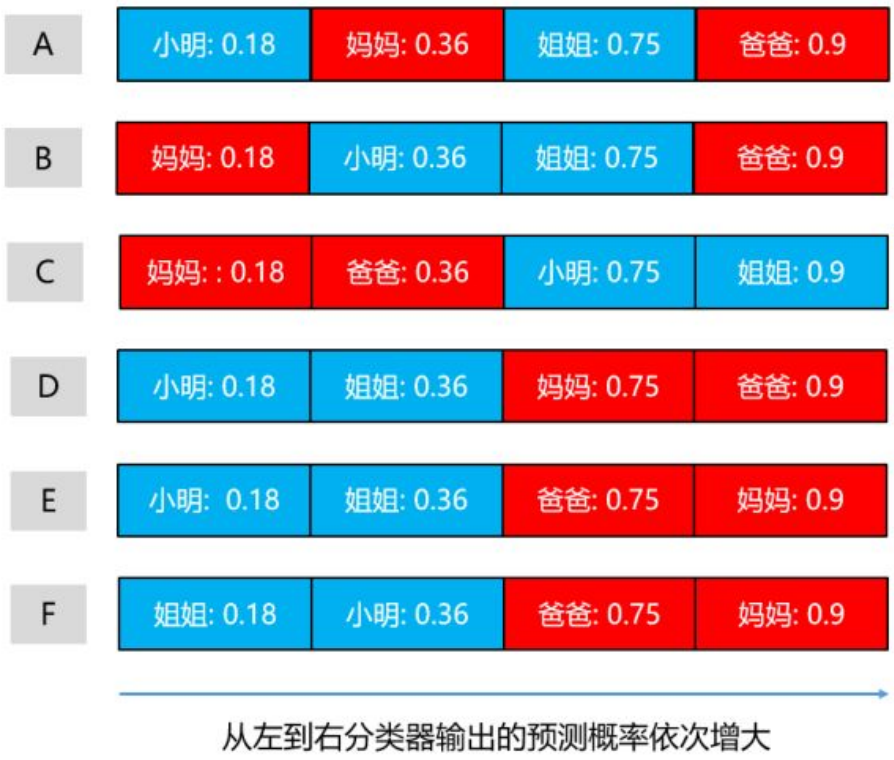
以下为三种模型的输出结果，求三种模型的 AUC ：

	小明	姐姐	妈妈	爸爸
a	0.12	0.35	0.76	0.85
b	0.12	0.35	0.44	0.49
c	0.52	0.65	0.76	0.85

1. AUC更多的是关注对计算概率的排序，关注的是概率值的相对大小，与阈值和概率值的绝对大小没有关系 例子中并不关注小明是不是成人，而关注的是，预测为成人的概率的排序。
2. AUC只关注正负样本之间的排序，并不关心正样本内部，或者负样本内部的排序。这也体现了AUC的本质： 任意个正样本的概率都大于负样本的概率的能力。

例子中AUC只需要保证（小明和姐姐）（爸爸和妈妈），小明和姐姐在前2个排序，爸爸和妈妈在后2个排序，而不会考虑小明和姐姐谁在前，或者爸爸和妈妈谁在前 。AUC只与概率的相对大小（概率排序）有关，和绝对大小没关系。由于三个模型概率排序的前两位都是未成年人（小明，姐姐），后两位都是成年人（妈妈，爸爸），因此三个模型的AUC都等于1。

问题 14：以下已经对分类器输出概率从小到大进行了排列，哪些情况的AUC等于1， 情况的AUC为0（其中背景色表示True value，红色表示成年人，蓝色表示未成年人）。



D 模型, E模型和F模型的AUC值为1，C 模型的AUC值为0（爸妈为成年人的概率小于小明和姐姐，显然这个模型预测反了）。

AUC的计算

- 法1：AUC为ROC曲线下的面积，那我们直接计算面积可得。面积为一个个小的梯形面积（曲线）之和。计算的精度与阈值的精度有关 。
- 法2：根据AUC的物理意义，我们计算*正样本预测结果大于负样本预测结果*的概率。取n1* n0(n1为正样本数，n0为负样本数)个二元组，每个二元组比较正样本和负样本的预测结果，正样本预测结果高于负样本预测结果则为预测正确，预测正确的二元组占总二元组的比率就是最后得到的AUC。时间复杂度为O(N* M)。
- 法3：我们首先把所有样本按照score排序，依次用rank表示他们，如最大score的样本，rank=n (n=n0+n1，其中n0为负样本个数，n1为正样本个数)，其次为n-1。那么对于正样本中rank最大的样本，rank_max，有n1-1个其他正样本比他score小,那么就有(rank_max-1)-(n1-1)个负样本比他score小。其次为(rank_second-1)-(n1-2)。最后我们得到正样本大于负样本的概率为 :

$$AUC = \frac{\sum_{正样本} \text{rank}(score) - \frac{n_1*(n_1+1)}{2}}{n_0 * n_1}$$

其计算复杂度为O(N+M)。

下面有一个简单的例子：

真实标签为 (1, 0, 0, 1, 0) 预测结果1（0.9, 0.3, 0.2, 0.7, 0.5） 预测结果2（0.9, 0.3, 0.2, 0.7, 0.8）)

分别对两个预测结果进行排序，并提取他们的序号 结果1 (5, 2, 1, 4, 3) 结果2 (5, 2, 1, 3, 4)

对正分类序号累加 结果1：SUM正样本（rank(score))=5+4=9 结果2: SUM正样本（rank(score))=5+3=8

计算两个结果的AUC: 结果1：AUC= (9-2*3/2)/6=1 结果2：AUC= (8-3/2)/6=0.833*

问题 15：为什么说 ROC 和AUC都能应用于非均衡的分类问题？

ROC曲线只与横坐标 (FPR) 和 纵坐标 (TPR) 有关系 。我们可以发现TPR只是正样本中预测正确的概率，而FPR只是负样本中预测错误的概率，和正负样本的比例没有关系。因此 ROC 的值与实际的正负样本比例无关，因此既可以用于均衡问题，也可以用于非均衡问题。而 AUC 的几何意义为ROC曲线下的面积，因此也和实际的正负样本比例无关。

机器学习评价指标

十二、 图像处理

1. 图像维基百科定义

图像处理是指对图像进行分析、加工、和处理，使其满足视觉、心理或其他要求的技术。图像处理是信号处理在图像领域上的一个应用。目前大多数的图像均是以数字形式存储，因而图像处理很多情况下指数字图像处理。此外，基于光学理论的处理方法依然占有重要的地位。

图像处理是信号处理的子类，另外与计算机科学、人工智能等领域也有密切的关系。

传统的一维信号处理的方法和概念很多仍然可以直接应用在图像处理上，比如降噪、量化等。然而，图像属于二维信号，和一维信号相比，它有自己特殊的一面，处理的方式和角度也有所不同。

2. 解决方案

影像强化

几十年前，图像处理大多数由光学设备在模拟模式下进行。由于这些光学方法本身所具有的并行特性，至今他们仍然在很多应用领域占有核心地位，例如全息摄影。但是由于计算机速度的大幅度提高，这些技术正在迅速的被数字图像处理方法所替代。

从通常意义上讲，数字图像处理技术更加普适、可靠和准确。比起模拟方法，它们也更容易实现。专用的硬件被用于数字图像处理，例如，基于流水线的计算机体系结构在这方面获取了巨大的商业成功。今天，硬件解决方案被广泛的用于视频处理系统，但商业化的图像处理任务基本上仍以软件形式实现，运行在通用个人电脑上。

3. 常用的信号处理技术

大多数用于一维信号处理的概念都有其在二维图像信号领域的延伸，它们之中的一部分在二维情形下变得十分复杂。 同时图像处理自身也具有一些新的概念， 例如，连通性、旋转不变性，等等。这些概念仅对二维或更高维的情况下才有非平凡的意义。

图像处理中常用到快速傅立叶变换，因为它可以减小数据处理量和处理时间。

从一维信号处理扩展来的技术和概念

- 分辨率
- 动态范围
- 带宽
- 滤波器设计
- 微分算子
- 边缘检测
- Domain modulation
- 降噪（Noise reduction）

专用于二维（或更高维）的技术和概念

- 连通性
- 旋转不变性

4. 典型问题

- 几何变换（geometric transformations）：包括放大、缩小、旋转等。
- 颜色处理（color）：颜色空间的转化、亮度以及对比度的调节、颜色修正等。
- 图像融合（image composite）：多个图像的加、减、组合、拼接。
- 降噪（image denoising）：研究各种针对二维图像的去噪滤波器或者信号处理技术。
- 边缘检测：进行边缘或者其他局部特征提取。
- 分割：依据不同标准，把二维图像分割成不同区域。
- 图像编辑：和计算机图形学有一定交叉。
- 图像配准：比较或集成不同条件下获取的图像。
- 图像增强（image enhancement）：
- 图像数字水印：研究图像域的数据隐藏、加密、或认证。
- 图像压缩：研究图像压缩。

5. 应用

- 摄影及印刷
- 卫星图像处理（Satellite image processing）
- 医学图像处理（Medical image processing）
- 面孔识别,特征识别（Face detection, feature detection, face identification）
- 显微图像处理（Microscope image processing）
- 汽车障碍识别（Car barrier detection）

软件工具

- Adobe Photoshop
- Aphelion
- ImageJ
- OpenCV
- Ulead PhotoImpact
- Rapidminer图像处理扩展[永久失效链接] -工具，图像处理和图像挖掘

相关相近领域

- 分类
- 特征提取
- 模式识别
- 投影
- 多尺度信号分析
- 离散余弦变换

图像处理维基百科

十三、 软件架构

1. 软件架构维基百科定义

   软件架构是有关软件整体结构与组件的抽象描述，用于指导大型软件系统各个方面的设计。软件架构会包括软件组件、组件之间的关系，组件特性以及组件间关系的特性。软件架构可以和建筑物的架构相比拟。软件架构是构建计算机软件，开发系统以及计划进行的基础，可以列出开发团队需要完成的任务。

   软件架构是在软件的基础架构上进行决策，一但决定后，再修改的代价很大。软件架构中的决策包括在软件设计时的一些特殊结构性选项，例如要控制太空船登陆艇的系统需要快速而且可靠,因此需要选择适合实时计算的语言，而且为了满足可靠度的需求，程序需要有数个冗余的复本，各复本运作在不同的硬件上，以便比对各程序的结果。

   将软件架构文档化有助于和项目关系人之间的沟通，在高层设计时就可以提早进行决策，也可以在各项目之间复用设计组件

2. 介绍

软件体系结构是构建计算机软件实践的基础。与建筑师设定建筑项目的设计原则和目标，作为绘图员画图的基础一样，软件架构师或者系统架构师陈述软件架构以作为满足不同客户需求的实际系统设计方案的基础。从和目的、主题、材料和结构的联系上来说，软件架构可以和建筑物的架构相比拟。一个软件架构师需要有广泛的软件理论知识和相应的经验来实施和管理软件产品的高级设计。软件架构师定义和设计软件的模块化，模块之间的交互，用户界面风格，对外接口方法，创新的设计特性，以及高层事物的对象操作、逻辑和流程。

软件架构师与客户商谈概念上的事情，与经理商谈广泛的设计问题，与软件工程师商谈创新的结构特性，与程序员商谈实现技巧，外观和风格。

软件架构是一个系统的草图。软件架构描述的对象是直接构成系统的抽象组件。各个组件之间的连接则明确和相对细致地描述组件之间的通讯。在实现阶段，这些抽象组件被细化为实际的组件，比如具体某个类或者对象。在面向对象领域中，组件之间的连接通常用接口来实现。

3. 范围

软件架构的范围有许多不同的定义：

- 巨观系统架构：这是指高端的软件系统抽象化，其中包括了许多的组件（component），以及描述各模块之间关系的“连接器”（connector）。
- 重要的东西，无论是什么都可以：这是指软件架构师需要根据项目判断，哪些决策对系统以及项目关系人有高度影响。
- 了解系统环境的基础。
- 一些人们认为不容易改变的事务：设计架构是在软件生命周期一开始就要进行的，软件架构师需专注在一些“一开始就要正确”的决策，依照这个思路，若有些问题是可逆的，软件架构上的问题就可以转换为非架构性的问题。
- 许多的架构设计决策：软件架构不能只考虑许多的模型及结构，也要考虑造成这些特殊结构的决策，以及背后的原因。此见解引发了大量有关软件架构知识管理的研究。

在软件架构、设计、需求工程之间，没有具体明显的分界。这些是“一连串意图的结合”，从高端的设计意向到低端的设计细节。

4. 特点

软件架构有以下这些特点：

众多的关系人：软件架构需配合许多的关系人（stakeholder），例如业务经理、部门主管、用户及运营商。每一个关系人都有各自关注的内容。在设计系统中，如何平衡这些关注，并展示他们所关注的消息，也是一个重点。因此，软件架构中就包括了处理众多的关注及关系人，因此在本质上就是跨领域的。

关注点分离：架构师降低复杂度的可行方式，就是将驱动设计的各关注分开。架构文件会呈现相关者关注的所有内容，会以建构的方式表示，另外也会用各相关者关注的角度来描述软件的架构。这种分开来的说明称为架构视图，例如4+1架构视图。

质量导向：传统的软件设计方法（例如杰克逊结构化编程）是依需求的机能以及资料在系统中流动的方式所驱动，不过目前的见解[4]:26–28是软件系统的架构和其质量属性（例如故障容许度、向下兼容、可扩展性、可靠度、可维护性、可用性、资料安全等）的关系更高。相

关者的关注可以转换为有关这些质量属性上的需求，一般会称为非功能性需求、额外功能性需求、行为需求或质量属性需求。

重复的风格：软件架构和建筑类似，在处理一些重复出现的事务时会发展出标准化的作法。标准化作法有许多不同的名称，其中也有不同程度的抽象化。常见的术语有架构风格、tactic、参考架构及架构模式。

概念完整性：这是佛瑞德·布鲁克斯在写作《人月神话》一书时提及：软件系统的架构是有关软件系统该作什么以及不该作什么的实体观点。这些观点应和软件的实现分开。架构师的角色是“观点的看守者”，确认系统中增加的部分是符合此架构，因此可以保有概念完整性。

认知制约：程序员马尔文·康威在1967年论文发表了康威定律，其中提到一个组织开发的软件，其架构会反映其组织架构。佛瑞德·布鲁克斯在写作《人月神话》一书时，就在书上时提到此例子，命名为“康威定律”。

5. 动机

软件架构是复杂系统“在智力上能理解”（intellectually graspable）的抽象，此抽象有以下的好处：

- 软件架构是在系统实现之前，分析软件系统行为的基础。不需要实际实现系统，就可确认某一软件系统符合关系人的需求，这在降低成本以及风险减轻上都很有助益。已针对这类的分析开发了许多的技术，例如软件架构分析方法（SAAM）、架构权衡分析方法（ATAM），或是针对软件系统以可视化的方式来呈现。
- 软件架构是软件复用以及决策的基础。不论是软件的软件架构，或是在软件架构上的个别策略及决策，若关系人在其他系统中也需要类似的属性或是机能，就可以重复使用，因此可以减少设计成本，也减少设计错误产生的风险。
- 可以在提早就进行会影响系统开发、布署以及维护的设计决策。若要避免时程逾期或是费用超支，提早做出正确的，高影响性的决策非常重要。
- 有助于和关系人之间的沟通，可以产出一个比较符合各方需求的系统。在有关复杂系统的沟通时，以关系人的观点来沟通有助于他们了解其提出需求和以此产生的设计决策之间的关系。透过架构，可以在系统实现之前（也比较容易调整的时候）就进行设计决策的沟通。
- 有助于风险管理。软件架构可以减少风险以及失败的几率。
- 可以降低成本。软件架构是一种管理复杂IT计划风险以及成本的方式。

软件架构维基百科

十四、 软件工程

1. 软件工程维基百科定义

软件工程（英语：software engineering[1]），是软件开发领域里对工程方法的系统应用。

1968年秋季，NATO（北约）的科技委员会召集了近50名一流的编程人员、计算机科学家和工业界巨头，讨论和制定摆脱“软件危机”的对策。在那次会议上第一次提出了软件工程（software engineering）这个概念，研究和应用如何以系统性的、规范化的、可定量的过程化方法去开发和维护软件，以及如何把经过时间考验而证明正确的管理技术和当前能够得到的最好的技术方法结合起来的学科。它涉及到程序设计语言、数据库、软件开发工具、系统平台、标准、设计模式等方面。其后的几十年里，各种有关软件工程的技术、思想、方法和概念不断被提出，软件工程逐步发展为一门独立的科学。

1993年，电气电子工程师学会（IEEE）给出了一个更加综合的定义："将系统化的、规范的、可度量的方法用于软件的开发、运行和维护的过程，即将工程化应用于软件开发中"。此后，IEEE多次给出软件工程的定义。

在现代社会中，软件应用于多个方面。典型的软件比如有电子邮件、嵌入式系统、人机界面、办公包、操作系统、网页、编译器、数据库、游戏等。同时，各个行业几乎都有计算机软件的应用，比如工业、农业、银行、航空、政府部门等。这些应用促进了经济和社会的发展，提高人们的工作效率，同时提升了生活质量。

软件工程师是对应用软件创造软件的人们的统称，软件工程师按照所处的领域不同可以分为系统分析师、系统架构师、前端和后端工程师、程序员、测试工程师、用户界面设计师等等。各种软件工程师人们俗称程序员。

2. 名称由来和定义

软件工程包括两种构面：软件开发技术和软件项目管理。

   a. 软件开发技术：软件开发方法学、软件工具和软件工程环境。
   b. 软件项目管理：软件度量、项目估算、进度控制、人员组织、配置管理、项目项目等。

软件危机

1970年代和1980年代的软件危机。在那个时代，许多软件最后都得到了一个悲惨的结局，软件项目开发时间大大超出了规划的时间表。一些项目导致了财产的流失，甚至某些软件导致了人员伤亡。同时软件开发人员也发现软 软件开发的难度越来越大。在软件工程界被大量引用的案例是Therac-25的意外：在1985年六月到1987年一月之间，六个已知的医疗事故来自于Therac-25错误地超过剂量，导致患者死亡或严重辐射灼伤[2]。

由来

鉴于软件开发时所遭遇困境，北大西洋公约组织（NATO）在1968年举办了首次软件工程学术会议[3]，并于会中提出“软件工程”来界定软件开发所需相关知识，并建议“软件开发应该是类似工程的活动”。软件工程自1968年正式提出至今，这段时间累积了大量的研究成果，广泛地进行大量的技术实践，借由学术界和产业界的共同努力，软件工程正逐渐发展成为一门专业学科。

定义

> *关于软件工程的定义，在GB/T11457-2006《信息技术 软件工程术语》中将其定义为"应用计算机科学理论和技术以及工程管理原则和方法，按预算和进度，实现满足用户要求的软件产品的定义、开发、和维护的工程或进行研究的学科"。*

包括：

- 创立与使用健全的工程原则，以便经济地获得可靠且高效率的软件。
- 应用系统化，遵从原则，可被计量的方法来发展、操作及维护软件；也就是把工程应用到软件上。
- 与开发、管理及更新软件产品有关的理论、方法及工具。
- 一种知识或学科，目标是生产质量良好、准时交货、符合预算，并满足用户所需的软件。
- 实际应用科学知识在设计、建构计算机程序，与相伴而来所产生的文件，以及后续的操作和维护上。
- 使用与系统化生产和维护软件产品有关之技术与管理的知识，使软件开发与修改可在有限的时间与费用下进行。
- 建造由工程师团队所开发之大型软件系统有关的知识学科。
- 对软件分析、设计、实施及维护的一种系统化方法。
- 系统化地应用工具和技术于开发以计算机为主的应用。
- 软件工程是关于设计和开发优质软件。

软件工程维基百科

十五、 大数据

1. 大数据维基百科定义

大数据（英语：**Big data**），指的是传统数据处理应用软件不足以处理的大或复杂的数据集的术语。

大数据也可以定义为来自各种来源的大量非结构化或结构化数据。从学术角度而言，大数据的出现促成广泛主题的新颖研究。这也导致各种大数据统计方法的发展。大数据并没有统计学的抽样方法；它只是观察和追踪发生的事情。因此，大数据通常包含的数据大小超出传统软件在可接受的时间内处理的能力。由于近期的技术进步，发布新数据的便捷性以及全球大多数政府对高透明度的要求，大数据分析在现代研究中越来越突出。

2. 概述

截至2012年，技术上可在合理时间内分析处理的数据集大小单位为艾字节（EB）。在许多领域，由于数据集过度庞大，科学家经常在分析处理上遭遇限制和阻碍；这些领域包括气象学、基因组学、神经网络体学、复杂的物理模拟，以及生物和环境研究。这样的限制也对网络搜索、金融与经济信息学造成影响。数据集大小增长的部分原因来自于信息持续从各种来源被广泛收集，这些来源包括搭载感测设备的移动设备、高空感测科技（遥感）、软件记录、相机、麦克风、无线射频识别（RFID）和无线感测网络。自1980年代起，现代科技可存储数

据的容量每40个月即增加一倍；截至2012年，全世界每天产生2.5艾字节（2.5×1018字节）的数据。

大数据几乎无法使用大多数的数据库管理系统处理，而必须使用“在数十、数百甚至数千台服务器上同时平行运行的软件”（电脑集群是其中一种常用方式）。大数据的定义取决于持有数据组的机构之能力，以及其平常用来处理分析数据的软件之能力。“对某些组织来说，第一次面对数百GB的数据集可能让他们需要重新思考数据管理的选项。对于其他组织来说，数据集可能需要达到数十或数百TB才会对他们造成困扰。”

随着大数据被越来越多的提及，有些人惊呼大数据时代已经到来了，2012年《纽约时报》的一篇专栏中写到，“大数据”时代已经降临，在商业、经济及其他领域中，决策将日益基于数据和分析而作出，而并非基于经验和直觉。但是并不是所有人都对大数据感兴趣，有些人甚至认为这是商学院或咨询公司用来哗众取宠的时髦术语(buzzword)，看起来很新颖，但只是把传统重新包装，之前在学术研究或者政策决策中也有海量数据的支撑，大数据并不是一件新兴事物。

大数据时代的来临带来无数的机遇，但是与此同时个人或机构的隐私权也极有可能受到冲击，大数据包含各种个人信息数据，现有的隐私保护法律或政策无力解决这些新出现的问题。有人提出，大数据时代，个人是否拥有“被遗忘权”，被遗忘权即是否有权利要求数据商不保留自己的某些信息，大数据时代信息为某些互联网巨头所控制，但是数据商收集任何数据未必都获得用户的许可，其对数据的控制权不具有合法性。2014年5月13日欧盟法院就“被遗忘权”（Case of Right to be Forgotten）一案作出裁定，判决谷歌应根据用户请求删除不完整的、无关紧要的、不相关的数据以保证数据不出现在搜索结果中。这说明在大数据时代，加强对用户个人权利的尊重才是时势所趋的潮流。

3. 定义

大数据由巨型数据集组成，这些数据集大小常超出人类在可接受时间下的收集、废用、管理和处理能力。大数据的大小经常改变，截至2012年，单一数据集的大小从数太字节（TB）至数十兆亿字节（PB）不等。

在一份2001年的研究与相关的演讲中，麦塔集团（META Group，现为高德纳）分析员道格·莱尼（Doug Laney）指出数据长的挑战和机遇有三个方向：量（Volume，数据大小）、速（Velocity，数据输入输出的速度）与多变（Variety，多样性），合称“3V”或“3Vs”。高德纳与现在大部分大数据产业中的公司，都继续使用3V来描述大数据[18]。高德纳于2012年修改对大数据的定义：“大数据是大量、高速、及/或多变的信息资产，它需要新型的处理方式去促成更强的决策能力、洞察力与优化处理。”另外，有机构在3V之外定义第4个V：真实性（Veracity）为第四特点。

大数据必须借由计算机对数据进行统计、比对、解析方能得出客观结果。美国在2012年就开始着手大数据，奥巴马更在同年投入2亿美金在大数据的开发中，更强调大数据会是之后的未来石油。

数据挖掘（data mining）则是在探讨用以解析大数据的方法。

大数据需要特殊的技术，以有效地处理大量的容忍经过时间内的数据。适用于特殊大数据的技术，包括大规模并行处理（MPP）数据库、数据挖掘、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统。

4. 数据来源类型

大数据获取之来源影响其应用之效益与质量，依照获取的直接程度一般可分为三种：

- 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，拥有者可弹性地使用于分析研究、营销推广等。
- 第二方数据(Second Party Data)：取自第一方的数据，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。如：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，另一单位即可推荐他相关的旅游产品；或是已知某单位具有己方想要的数据，透过议定采购，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者，该数据即为第三方数据。通常提供第三方数据的单位为数据供应商，其广泛搜集各式数据，并贩售给数据需求者，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。

5. 应用示例

大数据的应用示例包括大科学、RFID、感测设备网络、天文学、大气学、交通运输、基因组学、生物学、大社会数据分析、互联网文件处理、制作互联网搜索引擎索引、通信记录明细、军事侦查、金融大数据，医疗大数据，社交网络、通勤时间预测、医疗记录、照片图像和影像封存、大规模的电子商务等。

应用于运动界

巨大科学

大型强子对撞机中有1亿5000万个传感器，每秒发送4000万次的数据。实验中每秒产生将近6亿次的对撞，在过滤去除99.999%的撞击数据后，得到约100次的有用撞击数据。

将撞击结果数据过滤处理后仅记录0.001%的有用数据，全部四个对撞机的数据量复制前每年产生25拍字节（PB），复制后为200拍字节。

如果将所有实验中的数据在不过滤的情况下全部记录，数据量将会变得过度庞大且极难处理。每年数据量在复制前将会达到1.5亿拍字节，等于每天有近500艾字节（EB）的数据量。这个数字代表每天实验将产生相当于500垓（5×1020）字节的数据，是全世界所有数据源总和的200倍。

科学研究

卫生学

国际卫生学教授汉斯·罗斯林使用“Trendalyzer”工具软件呈现两百多年以来全球人类的人口统计数据，跟其他数据交叉比对，例如收入、宗教、能源使用量等。

公共部门

目前，发达国家的政府部门开始推广大数据的应用。2012年奥巴马政府投资近两亿美元开始推行《大数据的研究与发展计划》，本计划涉及美国国防部、美国卫生与公共服务部门等多个联邦部门和机构，意在通过提高从大型复杂的的数据中提取知识的能力，进而加快科学和工程的开发，保障国家安全。

信息审查

参见：大规模监控和社会信用体系

中国政府计划创建全面的个人信用评分体系，其包含不少对个人行为的评定，有关指标会影响到个人贷款、工作、签证等生活活动。高科技公司在被政治介入为其目的服务，个人的大部分行为和社交关系受掌控，几乎无人可免于被纳入个人信用评价体系的监控中。除获取网络数据外，中国政府还希望从科技公司获得分类和分析信息的云端计算能力，透过城市监控摄像机、智能手机、政府数据库等搜集数据，以建造智慧城市和安全城市。人权观察驻香港研究员王松莲指出，整个安全城市构想无非是一个庞大的监控项目。

民间部门

- 亚马逊公司，在2005年的时点，这间公司是世界上最大的以LINUX为基础的三大数据库之一。
- 沃尔玛可以在1小时内处理百万以上顾客的消费处理。相当于美国国会图书馆所藏的书籍之167倍的情报量。
- Facebook，处理500亿枚的用户照片。
- 全世界商业数据的数量，统计全部的企业全体、推计每1.2年会倍增。
- 西雅图文德米尔不动产分析约1亿匿名GPS信号，提供购入新房子的客户从该地点使用交通工具(汽车、脚踏车等)至公司等地的通勤时间估计值。
- 软银，每个月约处理10亿件（2014年3月现在）的手机LOG情报，并用其改善手机信号的信号强度。
- 大企业对大数据技能需求量大，吸引了许多大学诸如伯克利大学开专门提供受过大数据训练的毕业者的大学部门。硅谷纽约为主《The Data Incubator》公司，2012年成立，焦点是数据科学与大数据企业培训，提供国际大数据培训服务。

社会学

大数据产生的背景离不开Facebook等社交网络的兴起，人们每天通过这种自媒体传播信息或者沟通交流，由此产生的信息被网络记录下来，社会学家可以在这些数据的基础上分析人类的行为模式、交往方式等。美国的涂尔干计划就是依据个人在社交网络上的数据分析其自杀倾向，该计划从美军退役士兵中拣选受试者，透过Facebook的行动app收集资料，并将用户的活动数据传送到一个医疗资料库。收集完成的数据会接受人工智能系统分析，接着利用预测程序来即时监视受测者是否出现一般认为具伤害性的行为。

商业

运用数据挖掘技术，分析网络声量，以了解客户行为、市场需求，做营销策略参考与商业决策支持，或是应用于品牌管理，经营网络口碑、掌握负面事件等。如电信运营商透过品牌的网络讨论数据，即时找出负面事件进行处理，减低负面讨论在网络扩散后所可能引发的形象危害。

市场

大数据的出现提升了对信息管理专家的需求，Software AG、Oracle、IBM、微软、SAP、易安信、惠普和戴尔已在多间数据管理分析专门公司上花费超过150亿美元。在2010年，数据管理分析产业市值超过1,000亿美元，并以每年将近10%的速度成长，是整个软件产业成长速度的两倍。

经济的开发成长促进了密集数据科技的使用,也促成发展了相对于信息经济的数据经济。全世界共有约46亿的移动电话用户，并有10至20亿人链接互联网。自1990年起至2005年间，全世界有超过10亿人进入中产阶级，收入的增加造成了识字率的提升，更进而带动信息量的成长。全世界透过电信网络交换信息的容量在1986年为281兆亿字节（PB），1993年为471兆亿字节，2000年时增长为2.2艾字节（EB），在2007年则为65艾字节。根据预测，在2013年互联网每年的信息流量将会达到667艾字节。

大数据维基百科

十六、 数据挖掘

1. 数据挖掘维基百科定义

数据挖掘（英语：**data mining**）是一个跨学科的计算机科学分支 。它是用人工智能、机器学习、统计学和数据库的交叉方法在相对较大型的数据集中发现模式的计算过程。

数据挖掘过程的总体目标是从一个数据集中提取信息，并将其转换成可理解的结构，以进一步使用。除了原始分析步骤，它还涉及到数据库和数据管理方面、数据预处理、模型与推断方面考量、兴趣度度量、复杂度的考虑，以及发现结构、可视化及在线更新等后处理。数据挖掘是“数据库知识发现”（Knowledge-Discovery in Databases, KDD）的分析步骤，本质上属于机器学习的范畴。

类似词语“数据捕捞”、“数据捕鱼”和“数据探测”指用数据挖掘方法来采样（可能）过小以致无法可靠地统计推断出所发现任何模式的有效性的更大总体数据集的部分。不过这些方法可以创建新的假设来检验更大数据总体。

定义

数据挖掘有以下这些不同的定义：

a. “从资料中提取出隐含的过去未知的有价值的潜在信息”
b. “一门从大量资料或数据库中提取有用信息的科学”

尽管通常数据挖掘应用于资料分析，但是像人工智能一样，它也是一个具有丰富含义的词汇，可用于不同的领域。 它与KDD(Knowledge discovery in databases)的关系是：KDD是从数据中辨别有效的、新颖的、潜在有用的、最终可理解的模式的过程；而数据挖掘是KDD通过特定的算法在可接受的计算效率限制内生成特定模式的一个步骤。 事实上，在现今的文献中，这两个术语经常不加区分的使用。

2. 本质

数据挖掘本质上属于机器学习的内容。

例如《数据挖掘：实用机器学习技术及Java实现》一书大部分是机器学习的内容。这本书最初只叫做“实用机器学习”，“数据挖掘”一词是后来为了营销才加入的。通常情况下，使用更为正式的术语，（大规模）数据分析和分析学，或者指出实际的研究方法（例如人工智能和机器学习）会更准确一些。

3. 过程

数据挖掘的实际工作是对大规模数据进行自动或半自动的分析，以提取过去未知的有价值的潜在信息，例如数据的分组（通过聚类分析）、数据的异常记录（通过异常检测）和数据之间的关系（通过关联式规则挖掘）。这通常涉及到数据库技术，例如空间索引。这些潜在信息可通过对输入数据处理之后的总结来呈现，之后可以用于进一步分析，比如机器学习和预测分析。举个例子，进行数据挖掘操作时可能要把数据分成多组，然后可以使用决策支持系统以获得更加精确的预测结果。不过数据收集、数据预处理、结果解释和撰写报告都不算数据挖掘的步骤，但是它们确实属于“数据库知识发现”（KDD）过程，只不过是一些额外的环节。

数据库知识发现（KDD）过程通常定义为以下阶段：

预处理

在运用数据挖掘算法之前，必须收集目标数据集。由于数据挖掘只能发现实际存在于数据中的模式，目标数据集必须大到足以包含这些模式，而其余的足够简洁以在一个可接受的时间范围内挖掘。常见的数据源如资料超市或数据仓库。在数据挖掘之前，有必要预处理来分析多变量数据。然后要清理目标集。数据清理移除包含噪声和含有缺失数据的观测量。

数据挖掘

数据挖掘涉及六类常见的任务：

- 异常检测（异常/变化/偏差检测）– 识别不寻常的数据记录，错误数据需要进一步调查。
- 关联规则学习（依赖建模）– 搜索变量之间的关系。例如，一个超市可能会收集顾客购买习惯的数据。运用关联规则

学习，超市可以确定哪些产品经常一起买，并利用这些信息帮助营销。这有时被称为市场购物篮分析。
- 聚类 – 是在未知数据的结构下，发现数据的类别与结构。
- 分类 – 是对新的数据推广已知的结构的任务。例如，一个电子邮件程序可能试图将一个电子邮件分类为“正常邮件”或“垃圾邮件”。
- 回归 – 试图找到能够以最小误差对该数据建模的函数。
- 汇总 – 提供了一个更紧凑的数据集表示，包括生成可视化和报表。

结果验证

数据挖掘的价值一般带着一定的目的，而这目的是否得到实现一般可以通过结果验证来实现。验证是指“通过提供客观证据对规定要求已得到满足的认定”，而这个“认定”活动的策划、实施和完成，与“规定要求”的内容紧密相关。数据挖掘过程中的数据验证的“规定要求”的设定，往往与数据挖掘要达到的基本目标、过程目标和最终目标有关。验证的结果可能是“规定要求”得到完全满足，或者完全没有得到满足，以及其他介于两者之间的满足程度的状况。验证可以由数据挖掘的人自己完成，也可以通过其他人参与或完全通过他人的项目，以与数据挖掘者毫无关联的方式进行验证。一般验证过程中，数据挖掘者是不可能不参与的，但对于认定过程中的客观证据的收集、认定的评估等过程如果通过与验证提出者无关的人来实现，往往更具有客观性。通过结果验证，数据挖掘者可以得到对自己所挖掘的数据价值高低的评估。

4. 隐私问题及伦理

与数据挖掘有关的，还牵扯到隐私问题，例如：一个雇主可以透过访问医疗记录来筛选出那些有糖尿病或者严重心脏病的人，从而意图削减保险支出。然而，这种做法会导致伦理和法律问题。

对于政府和商业资料的挖掘，可能会涉及到的，是国家安全或者商业机密之类的问题。这对于保密也是个不小的挑战。

数据挖掘有很多合法的用途，例如可以在患者群的数据库中查出某药物和其副作用的关系。这种关系可能在1000人中也不会出现一例，但药物学相关的项目就可以运用此方法减少对药物有不良反应的病人数量，还有可能挽救生命；但这当中还是存在着数据库可能被滥用的问题。

数据挖掘实现了用其他方法不可能实现的方法来发现信息，但它必须受到规范，应当在适当的说明下使用。

如果资料是收集自特定的个人，那么就会出现一些涉及保密、法律和伦理的问题。

2018年5月25日，欧盟一般资料保护规范(General Data Protection Regulation，GDPR)正式上路，保障个人资料搜集的同意权与删除要求，在进入网站时会进行个人资料搜集、处理及利用之告知，并在当事人同意之下做搜集。

5. 方法

数据挖掘的方法包括监督式学习、非监督式学习、半监督学习、增强学习。监督式学习包括：分类、估计、预测。非监督式学习包括：聚类，关联规则分析。

6. 例子

数据挖掘在零售行业中的应用：零售公司跟踪客户的购买情况，发现某个客户购买了大量的真丝衬衣，这时数据挖掘系统就在此客户和真丝衬衣之间创建关系。销售部门就会看到此信息，直接发送真丝衬衣的当前行情，以及所有关于真丝衬衫的资料发给该客户。这样零售商店通过数据挖掘系统就发现了以前未知的关于客户的新信息，并且扩大经营范围。

7. 数据捕捞

通常作为与资料仓库和分析相关的技术，数据挖掘处于它们的中间。然而，有时还会出现十分可笑的应用，例如发掘出不存在但看起来振奋人心的模式（特别的因果关系），这些根本不相关的、甚至引人误入歧途的、或是毫无价值的关系，在统计学文献里通常被戏称为“资料挖泥”（Data dredging, data fishing, or data snooping）。

数据挖掘意味着扫描可能存在任何关系的资料，然后筛选出符合的模式，（这也叫作“过度匹配模式”）。大量的数据集中总会有碰巧或特定的资料，有着“令人振奋的关系”。因此，一些结论看上去十分令人怀疑。尽管如此，一些探索性资料分析 还是需要应用统计分析查找资料，所以好的统计方法和数据资料的界限并不是很清晰。

更危险是出现根本不存在的关系性。投资分析家似乎最容易犯这种错误。在一本叫做《顾客的游艇在哪里？》的书中写道：“总是有相当数量的可怜人，忙于从上千次的赌轮盘的轮子上查找可能的重复模式。十分不幸的是，他们通常会找到。”

多数的数据挖掘研究都关注于发现大量的资料集中，一个高度详细的模式。在《大忙人的数据挖掘》一书中， 西弗吉尼亚大学和不列颠哥伦比亚大学研究者讨论了一个交替模式，用来发现一个资料集当中两个元素的最小区别，它的目标是发现一个更简单的模式来描述相关数据。

数据挖掘维基百科

十七、 数据库

1. 数据库维基百科定义

数据库，又称为数据管理系统，简而言之可视为电子化的文件柜——存储电子文件的处所，用户可以对文件中的资料执行新增、截取、更新、删除等操作。

所谓“数据库”是以一定方式储存在一起、能予多个用户共享、具有尽可能小的冗余度、与应用程序彼此独立的数据集合。一个数据库由多个表空间（Tablespace）构成

2. 背景/技术初衷

在操作系统出现之后，随着计算机应用范围的扩大、需要处理的数据迅速膨胀。最初，数据与程序一样，以简单的文件作为主要存储形式。以这种方式组织的数据在逻辑上更简单，但可扩展性差，访问这种数据的程序需要了解数据的具体组织格式。当系统数据量大或者用户访问量大时，应用程序还需要解决数据的完整性、一致性以及安全性等一系列的问题。因此，必须开发出一种系统软件，它应该能够像操作系统屏蔽了硬件访问复杂性那样，屏蔽数据访问的复杂性。由此产生了数据管理系统，即数据库。

3. 数据库管理系统

数据库管理系统（英语：Database Management System，简称DBMS）是为管理数据库而设计的电脑软件系统，一般具有存储、截取、安全保障、备份等基础功能。数据库管理系统可以依据它所支持的数据库模型来作分类，例如关系式、XML；或依据所支持的电脑类型来作分类，例如服务器聚类、移动电话；或依据所用查询语言来作分类，例如SQL、XQuery；或依据性能冲量重点来作分类，例如最大规模、最高执行速度；亦或其他的分类方式。不论使用哪种分类方式，一些DBMS能够跨类别，例如，同时支持多种查询语言。

4. 数据库的分类

随着数据库技术与其他分支学科技术的结合，出现了多种新型数据库，例如：与分布处理技术结合产生的分布式数据库、与并行处理技术结合产生的并行数据库、与人工智能结合产生的演绎数据库、与多媒体技术结合产生的多媒体数据库。另外，数据库技术应用于特定的领域，出现了工程数据库、 地理数据库、统计数据库、空间数据库等特定领域数据库。

关系数据库

- MySQL
    - MariaDB（MySQL的代替品，维基媒体基金会项目已从MySQL转向MariaDB）
    - Percona Server（MySQL的代替品）
- PostgreSQL
- Microsoft Access
- Microsoft SQL Server
- Google Fusion Tables
- FileMaker
- Oracle数据库
- Sybase
- dBASE
- Clipper
- FoxPro
- foshub

几乎所有的数据库管理系统都配备了一个开放式数据库连接（ODBC）驱动程序，令各个数据库之间得以互相集成。

非关系型数据库（NoSQL）

- BigTable（Google）
- Cassandra
- MongoDB
- CouchDB
- Redis

键值数据库

- Apache Cassandra（为Facebook所使用）：高度可扩展
- Dynamo
- LevelDB（Google）

5. 数据库技术的发展

随着互联网的普及，数据库使用环境也随之发生变化，这种变化主要体现为XML和Java技术的大量使用、要求支持各种互联网环境下的应用服务器、极容易出现大量用户同时访问数据库、要求支持7x24小时不间断运行和高安全性等。

为解决由于这些变化所带来的新问题，数据库管理系统也逐渐产生变化，包括：

a. 网络化的大型通用数据库管理系统的出现

由于互联网应用的用户数量无法预测，这就要求数据库相比以前拥有能处理更大量的数据以及为更多的用户提供服务的能力，即更好的可伸缩性及高可用性，因此，能够支持Internet的数据库应用已经成为数据库系统的重要方面，学术界及各主流数据库公司都将大型通用数据管理系统作为主要发展方向。例如Oracle公司从 8 版起全面支持互联网应用，微软公司更是将 SQL Server 作为 其整个 .NET计划中的一个重要的成分。

b. 数据库安全系统及技术的提升

由于数据库系统在现代计算机系统中的地位越来越趋于核心的地位，数据库系统的安全问题自然受到越来越多的关注。在目前各国所引用或制定的 一系列安全标准中，最重要的两个是由美国国防部制定的《可信计算机系统的评估标准》(简称TCSEC)和《可信计算机系统的评估标准关于可信数据库系统的解释》(简称 TDI)。目前，所有数据库的开发必须遵从相应的安全标准。

c. XML及Web数据管理技术的普及

随着越来越多的Web应用，如电子商务、数字图书馆、信息服务等采用XML作为数据表现形式、越来越多网站采用XML作为信息发布的语言，以XML格式数据为主的半结构化数据逐步成为网上数据交换和数据表示的标准。而XML具有如下的一些特征：面向显示、半结构化和无结构、不同形式的数据源，动态变化以及数据海量等。因此，支持这种结构松散、形式多样、动态变化的海量数据的存储、共享、管理、检索，成了数据库技术的大势所趋。

Web数据管理是一个很松散的概念，大体上它是指在Web环境下对各种复杂信息的有效组织与集成，进行方便而准确的信息查询和发布。当前Web数据管理的研究开发方向主要包括：半结构化数据管理、Web数据查询、Web信息集成、XML数据管理等。到目前为止，XML 与 Web 数据管理的研究工作中主要集中在如下的一些方面。

- 半结构化数据
- Web数据查询
- XML相关标准
- XML数据管理

d. 嵌入式移动数据库技术

随着移动通信技术的迅速发展和投入使用，加上移动智能电话、移动计算机的大量普及，国内外许多研究机构都展开了对移动数据库的研究，并获取了许多有价值的成果。移动数据库技术涉及数据库技术、分布式计算技术以及移动通信技术等多个学科领域，具有较高的学术起点。

6. 数据库模型

- 对象模型
- 层次模型（轻量级数据访问协议）
- 网状模型（大型数据储存）
- 关系模型
- 面向对象模型
- 半结构化模型
- 平面模型（表格模型，一般在形式上是一个二维数组。如表格模型数据Excel)

架构 (Schema)

数据库的架构可以大致区分为三个概括层次：内层、概念层和外层。

- 内层：最接近实际存储体，亦即有关资料的实际存储方式。
- 外层：最接近用户，即有关个别用户观看资料的方式。
- 概念层：介于两者之间的间接层。

数据库索引

资料索引的观念由来已久，像是一本书前面几页都有目录，目录也算是索引的一种，只是它的分类较广，例如车牌、身份证字号、条码等，都是一个索引的号码，当我们看到号码时，可以从号码中看出其中的端倪，若是要找的人、车或物品，也只要提供相关的号码，即可迅速查到正确的人事物。

另外，索引跟字段有着相应的关系，索引即是由字段而来，其中字段有所谓的关键字段（Key Field），该字段具有唯一性，即其值不可重复，且不可为"空值（null）"。例如：在合并资料时，索引便是扮演欲附加字段资料之指向性用途的角色。故此索引为不可重复性且不可为空。

数据库事务

事务（transaction）包含一组数据库操作的逻辑工作单元，在事务中包含的数据库操作是不可分割的整体，这些操作要么一起做，要么一起回滚（Roll Back）到执行前的状态。 事务的ACID特性：

- 原子性（atomicity）
- 一致性（consistency）
- 隔离性（isolation）
- 持续性（durability）

事务的并发性是指多个事务的并行操作轮流交叉运行，事务的并发可能会访问和存储不正确的数据，破坏交易的隔离性和数据库的一致性。

网状数据模型的数据结构

网状模型

满足下面两个条件的基本层次联系的集合为网状模型。

a. 允许一个以上的结点无双亲；
b. 一个结点可以有多于一个的双亲。

数据库维基百科