出國報告(出國類別：研究)

「短距無線即時通訊及特徵影像辨識技術短期研究」出國報告

服務機關：國防部軍備局中山科學研究院
姓名職稱：廖人吉 聘用技士
派赴國家：美國
出國時間：民國 102 年 05 月 09 日至 102 年 08 月 26 日
報告日期：民國 102 年 09 月 04 日

國防部軍備局中山科學研究院出國報告建議事項處理表			
報告名稱	短距無線即時通訊及特徵影像辨識技術短期研究出國報告		
出國單位	工程測試組	出國人員級職/姓名	聘用技士/廖人吉
公差地點	美國加州	出/返國日期	102.05.09/102.08.26
建議事項	一、大量數據(Big Data)技術為美國歐巴馬總統大舉投入發展的下一個新科技技術，美國白宮宣布，將大量數據喻為「未來的新石油」，是國家發展的戰略性資產，是下一個世紀的重要產業，可促進美國未來數十年的經濟發展。此技術，在台灣仍只是萌芽期，與美國目前擁有的技術，相差甚遠，國內產官學也甚少投入，或投入不多，建議本院提早規劃及研究，以利未來科專建案。 二、無線穿戴式各種應用的開發與研究，例如：眼鏡的各式加值服務及應用、錶的各式加值服務及應用，相關的想法、需求、規格及技術，在美國被視為商業的極端機密，建議本院宜提早規劃及研究相關的技術，以利科專建案或產品研發。 三、以上都是要透過與教授的實際討論與參與，才能了解未來科技及市場的趨勢，這並不是在國內上上網即可獲得或吸收到的想法及觀念，建議本院應有計畫地、持續地派人至先進國家進修研究以了解未來科技及市場趨勢，才能更有助於本院投入相關技術的發展及未來科專建案。		
處理意見	一、電子所空電組已開始進行大量數據(Big Data)之相關技術，如分散式處理系統及 MapReduce 技術之研究，此項技術有助於影像處理之效能增進。 二、無線穿戴式各種應用的開發與研究已放入科專計畫 103-106「人本感知與智慧生活整合服務發展計畫」研發項目。 三、本院未來應持續並有規劃地派員至國外短期研究或邀請國外專家學者來本院進行科技研討，以增進本院之技術與國外同步。		

國防部軍備局中山科學研究院
102年度出國報告審查表

出國單位	電子系統研究所	出國人員級職姓名	聘用技士/廖人吉

單位	審查意見	簽章
一級單位	請掌握會辦期程，至遲應於返國後25天內將核定之出國報告送達企劃處報局；3個月內完成本院工作資訊網。	電子系統研究所 計管組技士 蔣國安 1020905161O 電子系統研究所 計管組副組長 趙俊聲 1020905164O 中山科學研究院 電子所代理所長 李明家 1020905174O
計品會	1.本案之研究內容，有助於解決"智慧感測網路技術與服務發展計畫"之干擾、定位及大量影像資料處理問題，可增進計畫執行成效。 2.本案短期研究之課題有利本院未來技術發展、市場規劃及科專建案。	計品會 秘書組資訊員 江秀芬 1020906110O 中山科學研究院 計品會副主委 萬紹正 1020909164O
保防安全處	案內出國報告（短距無線即時通訊及特徵影像辨識技術）已完成保密檢審作業，對於貴所將本件列為一般性資訊，本處敬表同意，無附加審查意見。	保防安全處 保防官 洪晢惟 1020910163O 中山科學研究院 保防安全處副處長 呂弘文 1020910190O
企劃處	一、案列本院102年出國計畫第102010案，依智慧感測網路技術與服務發展等計畫需求，赴美研習網路工程與管理、大量數據及資料探索課程，符合出國計畫主旨。請依規定辦理知識分享，擴大訓效。 二、請將奉核報告電子檔及紙本裝訂5份送本處續辦。另請於返國後3個月內，將報告電子檔登錄行政院資訊網及本院圖書館工作報告資訊網。	企劃處 科技組秘書 梁瓊真 1020913160O 企劃處 科技組技正 康來利 1020613163O 企劃處 科技組副組長 吳銘燦 1020913164O 中山科學研究院 企劃處處長 洪惠明 1020913170O

批示
1020917152O

國外公差人員出國報告主官（管）審查意見表

一、此次派員至美國西北理工大學短期研究，透過實際的研究及與教授討論，同仁不但可學習相關知識，並可實際了解國外最新科技發展方向及市場趨勢，有利本院未來技術發展規劃及建案。

二、本次出國短期研究內容，網路工程與管理(Network Engineering and management)、大量數據(Big data)、資料探勘(Data Mining)，有助於「智慧感測網路技術與服務發展計畫」之無法精準定位、無線通訊干擾之問題解決及影像資料處理效能之提升，並可增進計畫之執行效益。

附件二

出國報告審核表

出國報告名稱：短距無線即時通訊及特徵影像辨識技術短期研究出國報告		
出國人姓名（2人以上，以1人為代表）	職稱	服務單位
廖人吉	聘用技士	電子研究所工程測試組
出國類別	□考察 □進修 ■研究 □實習 □其他＿＿＿＿＿＿＿＿（例如國際會議、國際比賽、業務接洽等）	
出國期間：102年05月09日至102年08月26日	報告繳交日期：102年09月04日	

出國人員自我檢核	計畫主辦機關審核	審核項目
■	■	1.依限繳交出國報告。
■	■	2.格式完整（本文必須具備「目的」、「過程」、「心得」及「建議事項」）。
■	■	3.無抄襲相關資料。
■	■	4.內容充實完備。
■	■	5.建議具參考價值。
■	■	6.送本機關參考或研辦。
■	■	7.送上級機關參考。
□	□	8.退回補正，原因：
□	□	（1）不符原核定出國計畫。
□	□	（2）以外文撰寫或僅以所蒐集外文資料為內容。
□	□	（3）內容空洞簡略或未涵蓋規定要項。
□	□	（4）抄襲相關資料之全部或部分內容。
□	□	（5）引用其他資料未註明資料來源。
□	□	（6）電子檔案未依格式辦理。
□	□	（7）未於資訊網登錄提要資料及傳送出國報告電子檔。
■	■	9.本報告除上傳至出國報告資訊網外，將採行之公開發表：
■	■	（1）辦理本機關出國報告座談會（說明會），與同仁進行知識分享。
□	□	（2）於本機關業務會報提出報告。
■	■	（3）其他＿＿＿＿本報告已(將)於102年9月13日辦理知識分享。
■	■	10.其他處理意見(凡勾選項3者，請於「建議事項」明確說明不予刊登理由)：
□	□	（1）報告內容屬＿＿＿＿(機密、密)件，嚴禁上傳出國報告資訊網。
■	■	（2）報告內容屬普通件，不涉機敏，資料可對外公開。
□	□	（3）報告內容屬普通件，唯部分章節述及限閱資訊，為避免遭有心人士不當運用而產生後遺，請准比照機密資訊，不予刊登出國報告資訊網。
		請加會保防官及其主管核章 電子系統研究所 保防官 夏俊偉 中山科學研究院 電子所代理所長 李明家 1020905194O

出國人簽章（2人以上，得以1人為代表）	計畫主辦機關審核人	一級單位主管簽章	機關首長或其授權人員簽章
廖人吉		電子系統研究所 計管組組長 汪效聖 1020905194O	中山科學研究院 電子所代理所長 李明家 1020905194O

<table>
<tr><td colspan="4">報　　告　　資　　料　　頁</td></tr>
<tr><td>1.報告編號：</td><td>2.出國類別：
研究</td><td>3.完成日期：
102.09.04</td><td>4.總頁數：
82</td></tr>
<tr><td colspan="4">5.報告名稱：短距無線即時通訊及特徵影像辨識技術短期研究出國報告</td></tr>
<tr><td rowspan="2">6.核准
文號</td><td>人令文號</td><td colspan="2">102.04.23 國人管理字第 1020006627 號</td></tr>
<tr><td>部令文號</td><td colspan="2">102.04.19 國備獲管字第 1020005378 號</td></tr>
<tr><td colspan="2">7.經　　費</td><td colspan="2">新台幣：　405,084　元</td></tr>
<tr><td colspan="2">8.出(返)國日期</td><td colspan="2">102 年 05 月 9 日至 102 年 08 月 26 日</td></tr>
<tr><td colspan="2">9.公 差 地 點</td><td colspan="2">美國，佛立蒙</td></tr>
<tr><td colspan="2">10.公 差 機 構</td><td colspan="2">西北理工大學
(Northwestern Polytechnic University)</td></tr>
<tr><td colspan="2">11.附　　記</td><td colspan="2"></td></tr>
</table>

系統識別號

行政院及所屬各機關出國報告提要

出國報告名稱：短距無線即時通訊及特徵影像辨識技術短期研究出國報告

頁數 82　含附件：■是□否

出國計畫主辦機關/聯絡人/電話

中山科學研究院/廖人吉/(03)4712201 轉 353189

出國人員姓名/服務機關/單位/職稱/電話

廖人吉/國防部軍備局中山科學研究院/電子所測試組/聘用技士/353189

出國類別：□1 考察□2 進修■3 研究□4 實習□5 其他

出國期間：　　出國地區：

102 年 05 月 09 日至 08 月 26 日　　美國佛立蒙

報告日期：

102 年 09 月 4 日

分類號/目

關鍵詞：

無線通訊技術、影像資料處理、無線傳輸干擾

內容摘要：(二百至三百字)

此次出國至美國加州佛利蒙市之西北理工大學進行為期三個半月的短期進修研究，主要進修研究的課題有：網路工程與管理(Network Engineering and management)、大量數據(Big data)、資料探勘(Data Mining)以及計畫管理(Project Management)。學習研究的方向有：無線網路通訊技術、通訊路由(routing)技術、無線網路實體層抗干擾技術、網路安全(security)技術、資料探勘技術、大量資料/影像處理技術及計畫管理。這些技術對目前智慧感測網路技術與服務發展計畫所面臨的問題，如大量影像資料處理效能、無線通訊干擾問題及無法精準定位問題，提供一些方法及解決方向，對本院後續建案及無線通訊技術發展有極大的助益。

目　　次

短距無線即時通訊及特徵影像辨識技術短期研究出國報告

壹、 目的

本所執行經濟部「智慧感測網路技術與服務發展計畫」科專計畫案，開發適用於智慧居家、大樓空間等之緊急呼救相關的動靜態感測網路與無線 Zigbee 語音整合抗干擾應用裝置以及動態網路 Broadcast Protocol 等相關技術開發，當有緊急呼救需求時，能即時就地取材，提供救援通訊網路服務，提升救援效率，並研發視覺化整合服務技術，展示訊息記錄、位置標定，動態計算移動軌跡及視覺化呈現位置圖及 Zigbee 主動式電子封條與貨櫃特徵影像辨識技術等，以擴展安心保全智慧空間產業及貨櫃運輸業。目前仍有些技術尚待增進與突破，例如定位、大量影像處理及無線干擾等問題，派員至國外實際研究相關課題、學習相關知識及技術，可協助問題之解決並增進計畫執行成效。

貳、 過程

102 年 5 月 9 日至美國西北理工大學報到，並於次日至學校選課，與學校課程諮詢人員及教授討論，選定與計畫有關的四門課，分別是網路工程與管理(Network Engineering and management)、大量數據(Big data)、資料探勘(Data Mining)及計畫管理(Project Management)。大量數據(Big data)、資料探勘(Data Mining)，每週六上課，各三小時，10:00~17:00。網路工程與管理與計畫管理，每週四上課，分別為三小時，13:00~16:00 及 18:00~21:00。另外，網路工程與管理，每週一下午有二小時的上機實習課，16:00~18:00。此次研究課程共計 15 週時間，05/12 至 08/24，每次上課完畢，均有課業(homework)需繳交，每門課都有期中考(07/01~07/06)及期末考(08/19~08/25)。個人隨即開始這非常緊湊的學習(learning)之旅，上課、研究(study)、上機實習(drill)、討論(discussion)、課業(Homework)撰寫、準備考試及參加期中/期末考等過程，持續三個半月至 08/24 考完期末考，並於 08/24 晚上搭機返台，完成整個學習研究(study)過程。

茲將四門課程之學習研究(study)內容，簡述如下：

一、 網路工程與管理(Network engineering and Management)

研究目的(The purpose of study)：

1. 網路基礎(Network basics)：網路層(network layers), 協定及應用(protocols and well-know applications)
2. 網路架構(Network architecture), 實作(implementations), 服務模式(service models)
3. 近代網路課題及方法(Modern network topics and approaches)
4. 研究學校網路及網際網路結構(Understand school networks and internet structures)
5. 研究網路基礎元件，如路由器、交換器、存取點及伺服器之功能、運作、及管理。(A detailed understanding of the functions, operations and management of network infrastructure components, including routers, switches, access points, and servers.)

網路工程(network engineering)是利用通訊裝置和線路將地理位置不同的、功能獨立的多個電腦系統連線起來，以功能完善的網路軟體實作網路的硬體、軟體及資源共享和訊息傳遞的系統。簡單的說即連線兩台或多台電腦進行通訊的系統。

電腦網路可以按照其覆蓋範圍分成以下類別:

1.個人區域網絡無線個人區域網絡

2.區域網路

有線區域網路：乙太網路、令牌環、光纖分散式數據介面

無線區域網路：藍芽、Wi-Fi、ZigBee、虛擬區域網路(VLAN, Virtual Local Area Network)

3.校園網路(Campus Area Network, CAN)

4.都會網路

5.廣域網路 非同步傳輸模式、SDH(Synchronous Digital Hierarchy)

本課程主要研究(Study)TCP/IP 協定架構，各層之安全防護機制(security)、干擾(interference)、路徑(routing)、無線網路架構等主題。

OSI 7 layer

7	應用層	HTTP、SMTP、FTP、Telnet、SSH、NFS、RTSP、XMPP、ENRP
6	表示層	XDR、ASN.1、SMB、AFP、NCP

5	會話層	ASAP、ISO 8327、RPC、NetBIOS、Winsock、BSD sockets
4	傳輸層	TCP、UDP、TLS、RTP、SCTP、ATP
3	網路層	IP、ICMP、IGMP、IPX、BGP、OSPF、RIP、IGRP、EIGRP、ARP、RARP
2	連結層	乙太網(Ethernet)、令牌環(token ring)、HDLC、ISDN、ATM、IEEE 802.11、FDDI、PPP
1	實體層	線路、無線電、光纖

TCP/IP protocol：

5	應用層	HTTP、FTP、DNS
4	傳輸層	TCP、UDP、RTP、SCTP
3	網路層	IP
2	連結層	乙太網、Wi-Fi、MPLS
1	實體層	線路、無線電、光纖

1. 應用層(Application-Layer)：應用程式間溝通的協定，如簡易電子郵件傳送（Simple Mail Transfer Protocol, SMTP）、檔案傳輸協定(File Transfer Protocol, FTP）、網路終端機模擬協定（TELNET）等。
2. 傳輸層(Transport Layer)：提供端點間的資料傳送服務，如傳輸控制協定（Transmission Control Protocol, TCP）、使用者資料協定（User Datagram Protocol, UDP）等，負責傳送資料，並且確定資料已被送達並接收。
3. 網路層(Network Layer)：負責提供基本的封包傳送功能，讓每一塊資料封包都能夠到達目的端主機(但不檢查是否被正確接收)，如網際協定(Internet Protocol, IP)。
4. 連結層(Data link Layer)：實質網路媒體的管理協定，定義如何使用實際網路（如 Ethernet, Serial Line 等）來傳送資料。
5. 實體層(Physical layer)：實際電氣訊號。

傳輸層(Transport Layer)服務：

1. 多工及解多工(Multiplexing/De-multiplexing)
2. 可靠的資料傳輸(Reliable Data Transfer)
3. 流量控制(Flow Control)
4. 擁塞控制(Congestion Control)

傳輸層(Transport Layer)的協定：

1. 使用者資料協定(UDP): 非連結導向傳輸(connectionless transport)
2. 傳輸控制協定(TCP): 連結導向傳輸(connection-oriented transport)
3. 傳輸控制協定之擁塞控制(TCP congestion control)

路由演算法需求(Requirement of routing algorithm)

-正確簡易(Correctness and Simplicity)

-系統強固性(Robustness)

-穩定性(Stability)

-公平性(Fairness)

-最佳化(Optimality)

路由演算法(Routing algorithms)

1. 連結狀態(Link state)
2. 距離向量(Distance Vector)
3. 層階式路由(Hierarchical routing)
4. 廣播路由(broadcast routing)
5. 多向廣播路由(multicast routing)
6. 移動式路由(routing for mobile host)
7. 隨插即用網路路由(Routing in Ad Hoc Networks(AODV, Ad hoc On-demand Distance Vector))

路由協定(Routing in the Internet)

1. 路由訊息協定(RIP, Routing Information Protocol)
2. 開放式最短優先路徑(OSPF, Open Shortest Path First)
3. 邊界閘道器協定(BGP, Border Gateway Protocol)

連結層功能：

1. 錯誤偵測及修正(Error detection and correction)
2. 多重存取協定(Multiple access protocols)
3. 連結層定址(Link-Layer Addressing)
4. 乙太網路(Ethernet)

網路安全(security)：

5	應用層	身份認證、訪問控制、數據保密、數據完整性
4	傳輸層	端到端加密
3	網路層	防火牆、IP 加密通道
2	連結層	點到點鏈路加密
1	實體層	熱雜訊(thermal noise), 干擾(interference), the time-varying nature of fading channels

第二層安全機制(Layer 2 security mechanism is as follow and would be more safety than WEP)

第二層安全機制(**Layer 2 Security Mechanism)**		
參數(Parameter)		**說明(Description)**
第二層安全(Layer 2 Security)	WPA+WPA2	Wi-Fi 保護存取之設定
	802.1X	802.1x 之認證設定

	思科關鍵整合協定(CKIP)	設定 Cisco Key Integrity Protocol (CKIP)功能，具 16 字元的加密金鑰。
媒體存取控制過濾(MAC Filtering)	使用媒體存取控制過濾器過濾 client 端的 MAC 位址。	

媒體存取控制(Media access control, MAC)

WPA：Wi-Fi Protected Access, wi-fi 保護存取機制

廣域網路之第三層安全機制(Layer 3 security mechanism for WAN as follow)

廣域網路之第三層安全機制(**Layer 3 Security Mechanism for WLAN)**		
參數(Parameter)		**說明(Description)**
第三層安全(Layer 3 Security)	網路安全協定(IPSec)	啟動 IPSec 功能。請先確認軟體是否正確及客戶端硬體的相容性，再啟動 IPSec。 具 VPN 安全進階模組。
	虛擬私人網路(VPN Pass-Through)	虛擬私人網路設定.

IPSec(Internal protocol security)：網路安全協定。

VPN(Virtual private network)：虛擬私人網路

客戶區域網路之第三層安全機制(Layer 3 security mechanism for Guess LAN as follow)

參數(Parameter)		**說明(Description)**
第三層安全 Layer 3 Security	上網認證(Web Authentication)	需提供使用者及密碼之上網認證機制。
	上網通過(Web Pass through)	不需提供使用者及密碼即可上網。

二、 資料探勘**(Data Mining)**：

研究目的(Purpose of study)：

1. 研究資料探勘的技術及運算步驟(To study how computational procedures and techniques are employed in mining data)
2. 研究實作機器學習的策略(To study the implementation details of machine learning strategies)

資料探勘（Data mining），亦稱為數據挖掘、資料挖掘、資料採礦。它是資料庫知識發現（Knowledge-Discovery in Databases，KDD）中的一個步驟。資料探勘一般是指從大量的資料中自動搜尋隱藏於其中的有著特殊關聯性（屬於關聯學習規則,Association rule learning）的訊息的過程。資料挖掘通常與電腦科學有關，並通過統計、線上分析處理、情報檢索、機器學習、專家系統（依靠過去的經驗法則）和模式識別等諸多方法來實現上述目標。

資料探勘的定義：

1. 「從資料中提取出隱含的過去未知的有價值的潛在訊息」
2. 「一門從大量資料或者資料庫中提取有用訊息的科學」

儘管通常資料挖掘應用於資料分析，但是像人工智慧一樣，它也是一個具有豐富含義的詞彙，可用於不同的領域。 它與 KDD 的關聯是：KDD 是從資料中辨別有效的、新穎的、潛在有用的、最終可理解的模式的過程；而資料探勘是 KDD 透過特定的演算法在可接受的計算效率限制內生成特定模式的一個步驟。 事實上，在現今的文獻中，這兩個術語經常不加區分的使用。

資料探勘的方法包括監督式學習、非監督式學習、關聯分組（Affinity Grouping，作關聯性的分析）與購物籃分析（Market Basket Analysis）、群聚（Clustering）與描述（Description）。監督式學習包括：分類、估計、預測。

資料探勘在零售行業中的應用：零售公司跟蹤客戶的購買情況，發現某個客戶購買了大量的真絲襯衣，這時資料挖掘系統就在此客戶和真絲襯衣之間建立關聯。銷售部門就會看到此訊

息，直接發送真絲襯衣的當前行情，以及所有關於真絲襯衫的資料發給該客戶。這樣零售商店通過資料挖掘系統就發現了以前未知的關於客戶的新訊息，並且擴大經營範圍。

資料挖掘是因為大量有用資料快速增長的產物，使用計算機進行歷史資料分析，1960 年代數字方式採集資料已經實現。1980 年代，關聯式資料庫隨著能夠適應動態按需分析資料的結構化查詢語言發展起來。資料倉儲開始用來存儲大量的資料。因為面臨處理資料庫中大量資料的挑戰，於是資料挖掘應運而生，對於這些問題，它的主要方法是資料統計分析和人工智慧搜尋技術。

資料預處理一般包括資料清理、資料整合、資料變換和資料規約四個處理過程。

資料探勘過程：

1.問題的定義與主題分析

2.準備資料：資料清理、資料合併、資料選擇、資料變換、資料濃縮、資料品質分析

3.建立模型：何謂模式(what is model)、模型的精確度、模型的驗證

4.模式的評估：模式是什麼、探勘結果的評價和驗證

5.資料視覺化和知識管理：視覺化表示、知識管理

關聯規則：

1.關聯規則：概念分層、興趣度、資料庫中關聯規則的發現

2.關聯規則學習的 Apriori 演算法：使用侯選項集找頻繁項集、頻繁項集產生關聯規則

3.探勘關聯規則的多策略方法：多層關聯規則、多維關聯規則

決策樹：

1.什麼是決策樹

2.決策樹的原理：歸納學習、決策樹的表示、決策樹的學習、ID3 的演算法、修剪決策樹

3.決策樹的應用：規則提取、分類。

4.決策樹的優點

群聚分析：

1.既述：什麼是群聚分析、群聚分析的基本知識、群聚分析的分類

2.基於劃分的群聚演算法：基於劃分的評價函數、K-平均方法、K-中心點方法

3.層次群聚：凝聚方法、分裂方法

4.孤立點分析：基於統計的孤立點檢測、基於距離的孤立點檢測、基於偏移的孤立點檢測

資料探勘(Data Mining)導入企業，其重點在於企業領域方面的知識，而它的特殊領域工具(Domain-specific Tools)要結合企業中使用者的語言和分析過程，才能發揮工具的效能與增進企業的智慧。也就是要顛覆常規和超越平日的想像，展現企業目標與問題的知識，以支援解釋別人看不到、看不出的資訊來。企業必須能夠從巨大資料庫中挖掘到濃縮、先前不知、可理解的資訊，並從使用中獲利。例如，一個發行管理共同基金（mutual funds)的企業體要發掘潛在客戶，它要能整合客戶的帳戶、人口統計、生活型態等資料。也就是說要能把資料庫中人口資料切分成為一些關鍵子集合：都市化情況、婚姻狀態、家庭所得、年齡、風險偏好、高淨值等。最後，依據資料挖寶分析結果，可區分集群和從事推廣促銷活動，成功的把共同基金推展至市場上。

資料探勘(Data Mining)應用的行業包括了金融業、通訊電子、電信業、零售商、直效行銷、製造業、醫療保健及製藥業等等，應用領域如下表：

資料探勘的應用(Applications of Data Mining)		
客戶領域(Customer-focused)	營運(Operations-focused)	研究領域 (Research-focused)
生命週期(Life-time Value) 市場分析(Market-Basket Analysis) 研究及分割(Profiling & Segmentation) 保留(Retention) 目標市場(Target Market) 獲得(Acquisition)	利潤分析(Profitability Analysis) 價格(Pricing) 假貨偵測(Fraud Detection) 風險評估(Risk Assessment) 投資管理(Portfolio Management)	組合化學(Combinatorial Chemistry) 基因研究(Genetic Research) 流行病學(Epidemiology)

知識領域(Knowledge Portal) 交叉銷售(Cross-Selling) 活動管理(Campaign Management) 電子化商業(E-Commerce)	雇員流動(Employee Turnover) 資金管理(Cash Management) 生產效率(Production Efficiency) 網路性能(Network Performance) 製造流程(Manufacturing Processes)	

三、大量數據(**Big Data)**：

研究目的(Purpose of study)：

1. 研究 Apache Hadoop 架構(To study how to use Apache Hadoop framework)
2. 研究 MapReduce 樣型及方法(To understand MapReduce pattern and methods)
3. 分析大量數據及影像, MapReduce 的樣型與方法的應用(To analyze big data (and image) and apply MapReduce pattern and methods)

大量數據（Big data），或稱巨量資料、海量資料，指的是所涉及的資料量規模巨大到無法透過目前主流軟體工具，在合理時間內達到擷取、管理、處理、並整理成為幫助企業經營決策更積極目的的資訊。網路上每一筆搜索，網站上每一筆交易，敲打鍵盤，點擊滑鼠的每一個輸入都是數據，整理起來分析排行，它的功能可不僅僅止於事後被動了解市場，蒐集起來的資料還可以被規劃，引導開發更大的消費力量。

大數據的常見特點是 4V：資料量大(Volume)、輸入和處理速度快(Velocity)、資料多樣性(Variety)、真偽存疑的資料(Veracity)。

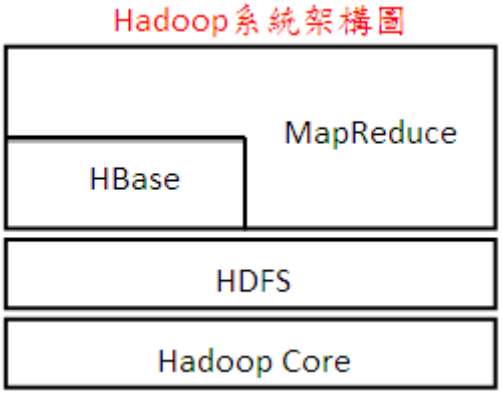
「大數據」是由數量巨大、結構複雜、類型眾多數據構成的數據集合，是基於雲計算的數據處理與應用模式，通過數據的整合共享，交叉復用形成的智力資源和知識服務能力。

大數據由巨型數據集（Data set）組成，這些數據集大小常超出常用軟體在可接受時間下的收集（data acquisition）、應用、管理和處理能力。決定大數據大小的指標永遠在變，截至 2012

年，大數據中的數據集可以由幾十兆位元組至數拍位元組(Petabyte, PB)的數據組成。這指標不固定是因為傳統資料庫管理系統以至 NoSQL 等新型數據庫，它們的科技和處理大容量數據的能力不斷在改進。在這前題下，新的平台正被開發去處理這些海量資料。美國在 2012 年就開始著手大數據，歐巴馬更在同年投入 2 億美金在大數據的開發中，更強調大數據會是之後的未來石油。1PB=1,000,000GB

對應的，大數據分析技術是對大數據的產生、存儲、挖掘和展現的全生命周期進行綜合分析處理的過程。

要提供要求(on-demand)動態擴充的運算能力，除了需要靠虛擬化技術以外，也必須建構在分散式運算 (Hadoop)以及分散式檔案系統 (HDFS)之上。如果僅能透過虛擬化技術作到快速的新增虛擬機器(Virtual machine, VM)，並且設定好相關的網路組態，但是資料無法有效的存放到分散式的檔案系統中，以提昇存取效率，對於超大量資料 (PB 等級) 的運算是沒有幫助的。而若只是將資料由集中改為分散存放，程式是不會自動變成以分散式的方式去執行的，程式架構必須重新設計(也就是利用 MapReduce 演算法改寫)，這就和單執行緒程式如果不經過調整，只是把 CPU 從單核換成多核，程式也不會自動變成多執行緒，是一樣的道理。



Hadoop 可分為運算及儲存兩大部份，前者由 MapReduce 負責，後者則由 HDFS 分散式檔案系統(Hadoop Distributed File System)負責

HDFS 分散式檔案系統：

以機率的觀點出發，將資料平均分散儲存在 HDFS 的成員(Data node)中，以提高存取效率，存入 HDFS 中的資料至少都會有3 份副本 (Replication)，存放於(理想上位於不同機架<rack>)的機器上。把一個檔案存入 HDFS 時，HDFS (更精確的說是 Name node) 會把檔案切割成固

定大小的區塊(block)，而後將各區塊(block)分散儲存到不同的 Data nodes 上，由於每個檔案的儲存都是跨實體機器的，因此 HDFS 可視為一個虛擬的分散式檔案系統 (傳統的檔案系統一樣會將檔案切割為區塊(block)，但都儲存到同一台實體機器的硬碟上)，或者說是一個 Logical File System，而 Name node 就負責扮演 Linux file system 中 inode 的角色，要知道組成某個檔案的所有區塊(block) 被儲存在哪些資料節點(Data node)? 問命名節點(Name node)就對了。

為了盡可能的提昇 HDFS 的存取效能(特別是讀取速度)，HDFS 在儲存資料時必須將資料根據機率平均的分佈在組成 Hadoop cluster 的成員硬碟上(平衡的工作)。由於資料是平均分佈的，因此存取時就可以多管齊下。

當 Client 需要讀取資料時，可以同時分別從五台伺服器(Server)各讀取一個區塊，讀到以後再組合成完整的檔案即可，如此一來存取效率會比循序的從單一 Server 上依序讀取區塊 1~5 來的快速許多，也可以降低每一台伺服器(Server)的 read lock 時間。

HDFS(Hadoop Distributed File System)的優點：

1.資料存取時間可以控制在一定的範圍內，硬碟的損耗也更平均，進一步減少檢修硬體的成本。

2.儲存於 HDFS 中的檔案大小可超過一顆實體硬碟的容量。由於 HDFS 是一個虛擬的分散式檔案系統，每個檔案的 block 本身就會跨實體機器儲存，因此自然不會受限於單一機器的硬碟大小了。例如雖然一個 Hadoop cluster 中的每台主機的硬碟都只有 500G，但這個 cluster 還是可以用來儲存 1TB/1PB 甚至更大的單一檔案，只要 cluster 內硬碟的總容量夠大就好。

3.整個 HDFS 系統是可以熱插拔的，當某一台 Server 的硬體壞掉時，HDFS 仍可正常運作(因為資料至少還會有另外 2 份副本)，並且由於資料的副本數低於 Policy 所規範的量，HDFS 會立即開始找尋另外正常運做的 Server 來維持副本的數量。接下來的維修只要直接把那台 Server 關機，換一台新的 Server 上去加入服務就好。因此機房管理人員就不需要利用三更半夜進行硬體的修復了。

由於分散式架構可大幅提昇檔案的讀取速度，因此有雲端運算需求的企業(如 Google) 就不用

再耗費鉅資購買高階的 Server，只要以一般的消費型機種就可以架構出高效能的運算平台(以 1 台高階伺服器的價格可以輕易的買到 10 台以上的消費型機種，但 1 台高階伺服器的效能不會比 10 台消費型機種的總和來的強)，如此可進一步降低 Data Center 的成本。量少質精的資料不如超大量、品質還 OK 的資料，這是大企業，如 google 奉行的原則

四、計畫管理(Project Management)

研究目的(objectives of study)：

1. 研究計畫管理的觀念及知識(To provide a good understanding of the project management terms and concept)
2. 研究計畫的建案與發展及如何於計畫執行階段制定策略(To provide a good understanding point for creating and developing projects and making strategy to implement during project execution phase)

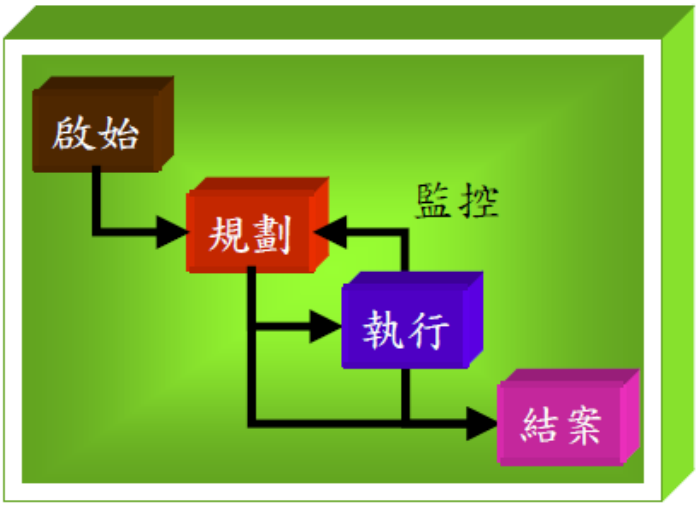
研究計畫管理的五大流程及九大知識領域、研究如何建立計畫的要徑(critical path)及如何用 excel solver 來規劃求得實際商業交易(business)的最佳解。其目的就是運用上述知識、技術、工具和方法來組織計畫活動，使能符合計畫的需求」。

摘要如下：

計畫管理的五大流程：

計畫管理是經由計畫起始、計劃、執行、監控及結案等五大程序的運作，方得以完成，每一個計畫都有五個主要程序組，包括：

1. 起始程序組（Initiating Processes Group）
2. 計畫程序組（Planning Processes Group）
3. 執行程序組（Executing Processes Group）
4. 監控程序組（Controlling Processes Group）
5. 結案程序組（Closing Processes Group）

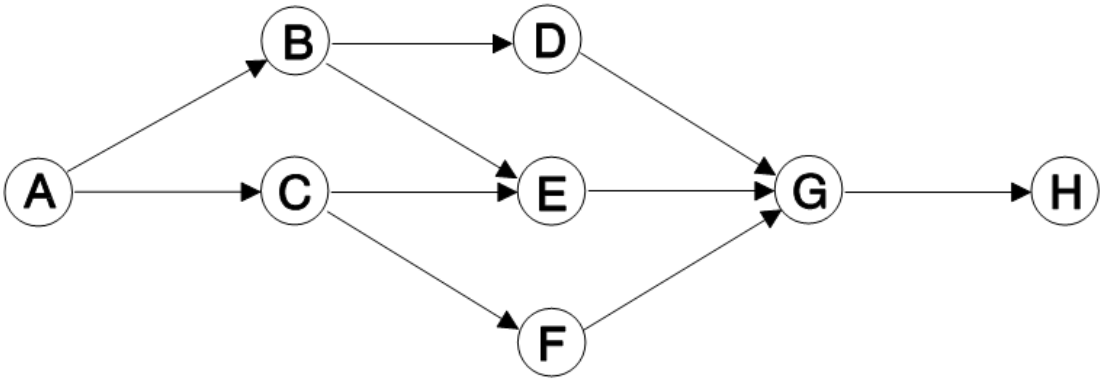


計畫管理的九大知識領域：

1. 計畫整合管理（Project Integration Management）
2. 計畫範疇管理（Project Scope Management）
3. 計畫時間管理（Project Time Management）
4. 計畫成本管理（Project Cost Management）
5. 計畫品質管理（Project Quality Management）
6. 計畫人力資源管理（Project Human Resource Management）
7. 計畫溝通管理（Project Communications Management）
8. 計畫風險管理（Project Risk Management）
9. 計畫採購管理（Project Procurement Management）

了解計畫管理的九大知識領域，並實際運作於計畫當中，可增進計畫執行效益。

如何建立計畫要徑(Critical path)？



工作	工作週數	前項工作	最早時間		最晚時間		寬裕時間
			開工	完工	開工	完工	
A	4	-	0	4	0	4	0
B	7	A	4	11	7	14	3
C	10	A	4	14	4	14	0
D	8	B	11	19	18	26	7
E	12	B、C	14	26	14	26	0
F	7	C	14	21	19	26	5
G	5	D、E、F	26	31	26	31	0
H	4	G	31	35	31	35	0

可算出計畫完工的時間

本例中最後工作為 H，其最早(或最晚)完工時間為 35 週(末)

計算每項工作的寬裕時間

表中工作 F 的寬裕時間為 5 週

表示該項工作最早可在第 14 週開工，最晚可在第 19 週開工

決定計畫的要徑(critical path)

本例中工作 A、C、E、G、H 為緊要工作

路線 A→C→E→G→H 為緊要路徑

如何用 excel solver 來規劃求得實際 business 的最佳解?

案例一，如下

相關參數設定如下：目標欄位(Target cell)為總利潤(total profit)且為最大值(max.)

限制條件(constraints)有二個->amt left >=0 及 no. to make >=0

by changing variable cells 為 no. to make

再利用 excel solver 規劃求解，即可得最佳利潤為 12360 元

	A	B	C	D	E	F	G	H	I	J
1					npu toys inc					
2					materials needed					
3								amt.	amt	amt
4		material	toy A	toy B	toy C	toy D	toy E	avail	used	left
5		red paint	0	1	0	1	3	625	625	0
6		blue paint	3	1	0	1	0	640	637	3
7		white paint	2	1	2	0	2	1100	1095	5
8		plastic	1	5	2	2	1	875	875	0
9		wood	3	0	3	5	5	2200	2200	0
10		glue	1	2	3	2	3	1500	1352	148
11		unit profit	15	30	20	25	25			
12		no.to make	192	19	158	42	188			
13		profit	2880	570	3160	1050	4700			
14		**total profit**	12360							

案例二、如下：

target cell 為 shipping costs: 並設 min

by changing cell 為所有 no. to ship from...

還要設三個 constraints 如下：

no. to be shipped >=0

no. remaining >=0

number needed 要等於 no. to be shipped.

再利用 excel solver 規劃求解，即可得最佳利潤為$55,515 元

shipping cost table

	la	st louis	boston
denver	58	47	108
houston	87	46	100
atlanta	121	30	57
miami	149	66	83
seattle	62	115	164
detroit	128	28	38

store	number needed	no. to ship from… la	st.louis	boston	no.to be shipped
denver	150	150	0	0	150
houston	225	0	225	0	225
atlanta	100	0	100	0	100
miami	250	0	25	225	250
seattle	120	120	0	0	120
detroit	150	0	0	150	150
total		270	350	375	

warehouse inventory

starting inventory	400	350	500
no.remaining	130	0	125

shipping costs	55515

參、 心得

此次來美進行短期進修研究三個半月，深刻感覺到美國是制定世界標準的地方，也是世界頂尖人才聚集之所在，各種先進技術、觀念、想法及未來的市場趨勢，都可以在這裡透過與教授實際討論、溝通而得到了解與啟發。例如：最新的技術，大量數據(Big Data)，這是美國歐巴馬總統已大舉投入發展的下一個新科技技術，美國白宮宣布，將大量數據喻為「未來的新石油」，是國家發展的戰略性資產，是下一個世紀的重要產業，可促進美國未來數十年的經濟發展。此技術，在台灣仍只是萌芽期，與美國目前擁有的技術，相差甚遠，國內產官學也甚少投入，或投入不多。另外，與教授實際討論中，也發現美國業者也大舉投入無線穿戴式各種應用的開發與研究，例如：眼鏡的各式加值服務及應用、錶的各式加值服務及應用，相關的想法、需求、規格及技術，在這裡被視為商業的極端機密。以上，都是要透過與教授的實際討論與參與，才能了解未來科技及市場的趨勢，這並不是在國內上上網即可獲得或吸收到的想法及觀念，本院應有計畫地、持續地派人至先進國家進修研究以了解未來科技及市場趨勢，才能更有助於本院投入相關技術的發展及未來科專建案。

成果：

一、此次來美所進修研究的四門課，分別為：網路工程與管理(Network Engineering and

management)、大量數據(Big data)、資料探勘(Data Mining)以及計畫管理(Project Management)，均是與目前正在執行的計畫有關，透過實際的學習與討論，了解最新的技術及市場趨勢，有助於未來投入新技術或新產品之發展規劃及建案。

二、網路工程與管理之各層(Layer)安全(Security)機制及技術，有助於「智慧感測網路技術與服務發展計畫」之無線(Zigbee)被/主動式電子封條安全(Security)技術之開發及增進。

三、網路工程與管理之實體層之安全(Security)想法及觀念，有助於「智慧感測網路技術與服務發展計畫」之無線(Zigbee)雙向語音傳輸裝置之干擾問題之解決。

四、網路工程與管理之路由(Routing)機制及技術，有助於「智慧感測網路技術與服務發展計畫」之可移動式叢集分散式架構、雙向語音傳輸、感測訊號與室內定位安全救援關鍵技術開發之增進。

五、大量數據(Big data)之 Mapreduce 觀念及技術，有助於「智慧感測網路技術與服務發展計畫」之影像資料處理效能之增進。

六、資料探勘(Data Mining)的觀念及技術，有助於「智慧感測網路技術與服務發展計畫」之大量影像資料分類、精簡、整合與處理。

七、大量數據(Big data)觀念及技術，為未來新興科技的趨勢，國內仍在萌芽期，可提早規劃及投入該項技術開發，以利科專建案。

八、計畫管理(project management)有助於「智慧感測網路技術與服務發展計畫」及各項計畫之管理。

九、以上四課程資料，均與「智慧感測網路技術與服務發展計畫」之短距無線即時通訊及特徵影像辨識技術短期研究工作目標有關，亦有蒐集相關之電子檔資料，以利後續研究。

肆、 建議事項

一、大量數據(Big Data)技術為美國歐巴馬總統大舉投入發展的下一個新科技技術，美國白宮宣布，將大量數據喻為「未來的新石油」，是國家發展的戰略性資產，是下一個世紀的重要產業，可促進美國未來數十年的經濟發展。此技術，在台灣仍只是萌芽期，與美國目前擁有的技術，相差甚遠，國內產官學也甚少投入，或投入不多，建議本院提早規劃及研究，以利未來科專建案。

二、無線穿戴式各種應用的開發與研究，例如：眼鏡的各式加值服務及應用、錶的各式加值服務及應用，相關的想法、需求、規格及技術，在美國被視為商業的極端機密，建議本院宜提早規劃及研究相關的技術，以利科專建案或產品研發。

三、以上都是要透過與教授的實際討論與參與，才能了解未來科技及市場的趨勢，這並不是在國內上上網即可獲得或吸收到的想法及觀念，建議本院應有計畫地、持續地派人至先進國家進修研究以了解未來科技及市場趨勢，才能更有助於本院投入相關技術的發展及未來科專建案。

附件

附件一

Description of Network Engineering and Management

摘要：網路工程(network engineering)是利用通訊裝置和線路將地理位置不同的、功能獨立的多個電腦系統連線起來，以功能完善的網路軟體實作網路的硬體、軟體及資源共享和訊息傳遞的系統，也就是即連線兩台或多台電腦進行通訊的系統。電腦網路可以按照其覆蓋範圍可分為：個人區域網絡無線個人區域網絡、區域網路、有線區域網路、校園網路、都會網路、廣域網路。主要探討 TCP/IP 協定架構，各層之安全防護機制(security)、干擾(interference)、路徑(routing)、無線網路架構。

A network engineering (or computer network) is a telecommunications network that allows computers to exchange data. The connections (network links) between networked computing devices (network nodes) are established using either cable media or wireless media. The best-known computer network is the Internet.

Network devices that originate, route and terminate the data are called network nodes.Nodes can include hosts such as servers and personal computers, as well as network hardware. Two devices are said to be networked when a process in one device is able to exchange information with a process in another device.

Computer networks support applications such as access to the World Wide Web, shared use of application and storage servers, printers, and fax machines, and use of email and instant messaging applications. The remainder of this article discusses local area network technologies and classifies them according to the following characteristics: the physical media used to transmit signals, the communications protocols used to organize network traffic, along with the network's size, its topology and its organizational intent.

Computer network may be considered a branch of electrical engineering, telecommunications, computer science, information technology or computer engineering, since it relies upon the theoretical and practical application of the related disciplines.

A computer network has the following properties:

Facilitates interpersonal communications

People can communicate efficiently and easily via email, instant messaging, chat rooms, telephone, video telephone calls, and video conferencing.

Allows sharing of files, data, and other types of information

Authorized users may access information stored on other computers on the network. Providing access to information on shared storage devices is an important feature of many networks.

Allows sharing of network and computing resources

Users may access and use resources provided by devices on the network, such as printing a document on a shared network printer. Distributed computing uses computing resources across a network to accomplish tasks.

May be insecure

A computer network may be used by computer Hackers to deploy computer viruses or computer worms on devices connected to the network, or to prevent these devices from accessing the network (denial of service).

May interfere with other technologies

Power line communication strongly disturbs certain[5] forms of radio communication, e.g., amateur radio. It may also interfere with last mile access technologies such as ADSL and VDSL.

May be difficult to set up

A complex computer network may be difficult to set up. It may be costly to set up an effective computer network in a large organization.

Network links

The communication media used to link devices to form a computer network include electrical cable (HomePNA, power line communication, G.hn), optical fiber (fiber-optic communication), and radio waves (wireless network). In the OSI model, these are defined at layers 1 and 2 — the physical layer and the data link layer.

A widely-adopted family of communication media used in local area network (LAN) technology is collectively known as Ethernet. The media and protocol standards that enable communication between networked devices over Ethernet is defined by IEEE 802. Ethernet encompasses both wired and wireless LAN technologies. Wired LAN devices transmit signals over cable media. Wireless LAN devices use radio waves or infrared signals as a transmission medium.

Wired technologies

The orders of the following wired technologies are, roughly, from slowest to fastest transmission speed.

- *Twisted pair wire* is the most widely used medium for all telecommunication. Twisted-pair cabling consist of copper wires that are twisted into pairs. Ordinary telephone wires consist of two insulated copper wires twisted into pairs. Computer network cabling (wired Ethernet as defined by IEEE 802.3) consists of 4 pairs of copper cabling that can be utilized for both voice and data transmission. The use of two wires twisted together helps to reduce crosstalk and electromagnetic induction. The transmission speed ranges from 2 million bits per second to 10 billion bits per second. Twisted pair cabling comes in two forms: unshielded twisted pair (UTP) and shielded twisted-pair (STP). Each form comes in several category ratings, designed for use in various scenarios.

- *Coaxial cable* is widely used for cable television systems, office buildings, and other work-sites for local area networks. The cables consist of copper or aluminum wire surrounded by an insulating layer (typically a flexible material with a high dielectric constant), which itself is surrounded by a conductive layer. The insulation helps minimize interference and distortion. Transmission speed ranges from 200 million bits per second to more than 500 million bits per second.

- ITU-T G.hn technology uses existing home wiring (coaxial cable, phone lines and power lines) to create a high-speed (up to 1 Gigabit/s) local area network.

- An optical fiber is a glass fiber. It uses pulses of light to transmit data. Some advantages of optical fibers over metal wires are less transmission loss, immunity from electromagnetic radiation, and very fast transmission speeds of up to trillions of bits per second. One can use different colors of lights to increase the number of messages being sent over a fiber optic cable.

Wireless technologies

- *Terrestrial microwave* – Terrestrial microwave communication uses Earth-based transmitters and receivers resembling satellite dishes. Terrestrial microwaves are in the low-gigahertz range, which limits all communications to line-of-sight. Relay stations are spaced approximately 48 km (30 mi) apart.

- *Communications satellites* – Satellites communicate via microwave radio waves, which are not deflected by the Earth's atmosphere. The satellites are stationed in space, typically in geosynchronous orbit 35,400 km (22,000 mi) above the equator. These Earth-orbiting systems are capable of receiving and relaying voice, data, and TV signals.

- *Cellular and PCS systems* use several radio communications technologies. The systems divide the region covered into multiple geographic areas. Each area has a low-power transmitter or radio relay antenna device to relay calls from one area to the next area.

- *Radio and spread spectrum technologies* – Wireless local area networks use a high-frequency radio technology similar to digital cellular and a low-frequency radio technology. Wireless LANs use spread spectrum technology to enable communication between multiple devices in a limited area. IEEE 802.11 defines a common flavor of open-standards wireless radio-wave technology known as Wifi.

- Infrared communication can transmit signals for small distances, typically no more than 10 meters. In most cases, line-of-sight propagation is used, which limits the physical positioning of communicating devices.

- A global area network (GAN) is a network used for supporting mobile across an arbitrary number of wireless LANs, satellite coverage areas, etc. The key challenge in mobile communications is handing off user communications from one local coverage area to the next. In IEEE Project 802, this involves a succession of terrestrial wireless LANs.

Exotic technologies

There have been various attempts at transporting data over exotic media:

IP over Avian Carriers was a humorous April fool's Request for Comments, issued as RFC 1149. It was implemented in real life in 2001.

Extending the Internet to interplanetary dimensions via radio waves.

Both cases have a large round-trip delay time, which gives slow two-way communication, but doesn't prevent sending large amounts of information.

Network nodes

Apart from the physical communications media described above, networks comprise additional basic hardware building blocks, such as network interface controller cards (NICs), repeaters, hubs, bridges, switches, routers, and firewalls.

Network interfaces

A network interface controller (NIC) is a hardware accessory that provides a computer with both a physical interface for accepting a network cable connector and the ability to process low-level network information.

In Ethernet networks, each network interface controller has a unique Media Access Control (MAC) address which is usually stored in the card's permanent memory. MAC address uniqueness is maintained and administered by the Institute of Electrical and Electronics Engineers (IEEE) in order to avoid address conflicts between devices on a network. The size of an Ethernet MAC address is six octets. The 3 most significant octets are reserved to identify card manufacturers. The card manufacturers, using only their assigned prefixes, uniquely assign the 3 least-significant octets of every Ethernet card they produce.

Repeaters and hubs

A repeater is an electronic device that receives a network signal, cleans it of unnecessary noise, and regenerates it. The signal is retransmitted at a higher power level, or to the other side of an obstruction, so that the signal can cover longer distances without degradation. In most twisted pair Ethernet configurations, repeaters are required for cable that runs longer than 100 meters. A repeater with multiple ports is known as a hub. Repeaters work on the physical layer of the OSI model. Repeaters require a small amount of time to regenerate the signal. This can cause a propagation delay which can affect network performance. As a result, many network architectures limit the number of repeaters that can be used in a row, e.g., the Ethernet 5-4-3 rule.

Repeaters and hubs have been mostly obsoleted by modern switches.

Bridges

A network bridge connects multiple network segments at the data link layer (layer 2) of the OSI model to form a single network. Bridges broadcast to all ports except the port on which the broadcast was received. However, bridges do not promiscuously copy traffic to all ports, as hubs do. Instead, bridges learn which MAC addresses are reachable through specific ports. Once the bridge associates a port with an address, it will send traffic for that address to that port only.

Bridges learn the association of ports and addresses by examining the source address of frames that it sees on various ports. Once a frame arrives through a port, the bridge assumes that the MAC address is associated with that port and stores its source address. The first time a bridge sees a previously unknown destination address, the bridge will forward the frame to all ports other than the one on which the frame arrived.

Bridges come in three basic types:

- Local bridges: Directly connect LANs
- Remote bridges: Can be used to create a wide area network (WAN) link between LANs. Remote bridges, where the connecting link is slower than the end networks, largely have been replaced with routers.
- Wireless bridges: Can be used to join LANs or connect remote devices to LANs.

Switches

A network switch is a device that forwards and filters OSI layer 2 datagrams between ports based on the MAC addresses in the packets. A switch is distinct from a hub in that it only forwards the frames to the ports involved in the communication rather than all ports connected. A switch breaks the collision domain but represents itself as a broadcast domain. Switches make decisions about where to forward frames based on MAC addresses. A switch normally has numerous ports, facilitating a star topology for devices, and cascading additional switches. Multi-layer switches are capable of routing based on layer 3 addressing or additional logical levels. The term switch is often used loosely to include devices such as routers and bridges, as well as devices that may distribute traffic based on load or based on application content (e.g., a Web URL identifier).

Routers

A router is an internetwork device that forwards packets between networks by processing the routing information included in the packet or datagram (Internet protocol information from layer 3). The routing information is often processed in conjunction with the routing table (or forwarding table). A router uses its routing table to determine where to forward packets. (A destination in a routing table can include a "null" interface, also known as the "black hole" interface because data can go into it, however, no further processing is done for said data.)

Firewalls

A firewall is a network device for controlling network security and access rules. Firewalls are typically configured to reject access requests from unrecognized sources while allowing actions from recognized ones. The vital role firewalls play in network security grows in parallel with the constant increase in cyber attacks.

Communications protocols

Internet map. The Internet is a global system of interconnected computer networks that use the standard Internet Protocol Suite (TCP/IP) to serve billions of users worldwide.

A communications protocol is a set of rules for exchanging information over a network. In a protocol stack (also see the OSI model), each protocol leverages the services of the protocol below it. An important example of a protocol stack is HTTP running over TCP over IP over IEEE 802.11. (TCP and IP are members of the Internet Protocol Suite. IEEE 802.11 is a member of the Ethernet protocol suite.) This stack is used between the wireless router and the home user's personal computer when the user is surfing the web.

Communication protocols have various characteristics. They may be connection-oriented or connectionless, they may use circuit mode or packet switching, and they may use hierarchical addressing or flat addressing.

There are many communication protocols, a few of which are described below

Ethernet

Ethernet is a family of protocols used in LANs, described by a set of standards together called IEEE 802 published by the Institute of Electrical and Electronics Engineers. It has a flat addressing scheme. It operates mostly at levels 1 and 2 of the OSI model. For home users today, the most well-known member of this protocol family is IEEE 802.11, otherwise known as Wireless LAN (WLAN). The complete IEEE 802 protocol suite provides a diverse set of network capabilities. For example, MAC bridging (IEEE 802.1D) deals with the routing of Ethernet packets using a Spanning Tree Protocol, IEEE 802.1Q describes VLANs, and IEEE 802.1X defines a port-based Network Access Control protocol, which forms the basis for the authentication mechanisms used in VLANs (but it is also found in WLANs) – it is what the home user sees when the user has to enter a "wireless access key".

Internet Protocol Suite

The Internet Protocol Suite, also called TCP/IP, is the foundation of all modern internetwork. It offers connection-less as well as connection-oriented services over an inherently unreliable network traversed by datagram transmission at the Internet protocol (IP) level. At its core, the protocol suite defines the addressing, identification, and routing specifications for Internet Protocol Version 4 (IPv4) and for IPv6, the next generation of the protocol with a much enlarged addressing capability.

SONET/SDH

Synchronous optical network (SONET) and Synchronous Digital Hierarchy (SDH) are standardized multiplexing protocols that transfer multiple digital bit streams over optical fiber using lasers. They were originally designed to transport circuit mode communications from a variety of different sources, primarily to support real-time, uncompressed, circuit-switched voice encoded in PCM(Pulse-Code Modulation) format. However, due to its protocol neutrality and transport-oriented features, SONET/SDH also was the obvious choice for transporting Asynchronous Transfer Mode (ATM) frames.

Asynchronous Transfer Mode

Asynchronous Transfer Mode (ATM) is a switching technique for telecommunication networks. It uses asynchronous time-division multiplexing and encodes data into small, fixed-sized cells. This differs from other protocols such as the Internet Protocol Suite or Ethernet that use variable sized packets or frames. ATM has similarity with both circuit and packet switched network. This makes it a good choice for a network that must handle both traditional high-throughput data traffic, and real-time, low-latency content such as voice and video. ATM uses a connection-oriented model in which a virtual circuit must be established between two endpoints before the actual data exchange begins.

While the role of ATM is diminishing in favor of next-generation networks, it still plays a role in the last mile, which is the connection between an Internet service provider and the home user. For an interesting write-up of the technologies involved, including the deep stacking of communications protocols used, see.

Scale

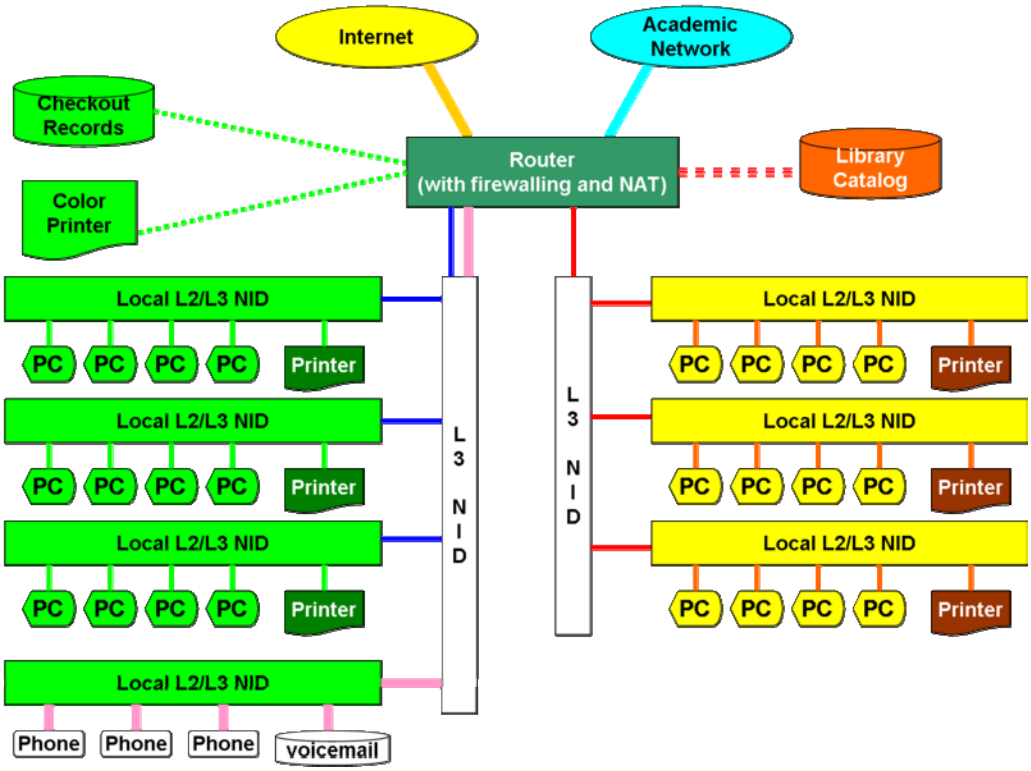
A network can be characterized by its physical capacity or its organizational purpose. Use of the network, including user authorization and access rights, differ accordingly.

Personal area network

A personal area network (PAN) is a computer network used for communication among computer and different information technological devices close to one person. Some examples of devices that are used in a PAN are personal computers, printers, fax machines, telephones, PDAs, scanners, and even video game consoles. A PAN may include wired and wireless devices. The reach of a PAN typically extends to 10 meters. A wired PAN is usually constructed with USB and Firewire connections while technologies such as Bluetooth and infrared communication typically form a wireless PAN.

Local area network

A local area network (LAN) is a network that connects computers and devices in a limited geographical area such as a home, school, office building, or closely positioned group of buildings. Each computer or device on the network is a node. Wired LANs are most likely based on Ethernet technology. Newer standards such as ITU-T G.hn also provide a way to create a wired LAN using existing wiring, such as coaxial cables, telephone lines, and power lines.



Typical library network, in a branching tree topology with controlled access to resources

A LAN is depicted in the accompanying diagram. All interconnected devices use the network layer (layer 3) to handle multiple subnets (represented by different colors). Those inside the library have 10/100 Mbit/s Ethernet connections to the user device and a Gigabit Ethernet connection to the central router. They could be called Layer 3 switches, because they only have Ethernet interfaces and support the Internet Protocol. It might be more correct to call them access routers, where the router at the top is a distribution router that connects to the Internet and to the academic networks' customer access routers.

The defining characteristics of a LAN, in contrast to a wide area network (WAN), include higher data transfer rates, limited geographic range, and lack of reliance on leased lines to provide connectivity. Current Ethernet or other IEEE 802.3 LAN technologies operate at data transfer rates up to 10 Gbit/s. The IEEE investigates the standardization of 40 and 100 Gbit/s rates.[13] A LAN can be connected to a WAN using a router.

Home area network

A home area network (HAN) is a residential LAN which is used for communication between digital devices typically deployed in the home, usually a small number of personal computers and accessories, such as printers and mobile computing devices. An important function is the sharing of Internet access, often a broadband service through a cable TV or digital subscriber line (DSL) provider.

Storage area network

A storage area network (SAN) is a dedicated network that provides access to consolidated, block level data storage. SANs are primarily used to make storage devices, such as disk arrays, tape libraries, and optical jukeboxes, accessible to servers so that the devices appear like locally attached devices to the operating system. A SAN typically has its own network of storage devices that are generally not accessible through the local area network by other devices. The cost and complexity of SANs dropped in the early 2000s to levels allowing wider adoption across both enterprise and small to medium sized business environments.

Campus area network

A campus area network (CAN) is made up of an interconnection of LANs within a limited geographical area. The network equipment (switches, routers) and transmission media (optical fiber, copper plant, Cat5 cabling, etc.) are almost entirely owned by the campus tenant / owner (an enterprise, university, government, etc.).

For example, a university campus network is likely to link a variety of campus buildings to connect academic colleges or departments, the library, and student residence halls.

Backbone network

A backbone network is part of a computer network infrastructure that provides a path for the exchange of information between different LANs or sub-networks. A backbone can tie together diverse networks within the same building, across different buildings, or over a wide area.

For example, a large company might implement a backbone network to connect departments that are located around the world. The equipment that ties together the departmental networks constitutes the network backbone. When designing a network backbone, network performance and network congestion are critical factors to take into account. Normally, the backbone network's capacity is greater than that of the individual networks connected to it.

Another example of a backbone network is the Internet backbone, which is the set of wide area networks (WANs) and core routers that tie together all networks connected to the Internet.

Metropolitan area network

A Metropolitan area network (MAN) is a large computer network that usually spans a city or a large campus.

Wide area network

A wide area network (WAN) is a computer network that covers a large geographic area such as a city, country, or spans even intercontinental distances. A WAN uses a communications channel that combines many types of media such as telephone lines, cables, and air waves. A WAN often makes use of transmission facilities provided by common carriers, such as telephone companies. WAN technologies generally function at the lower three layers of the OSI reference model: the physical layer, the data link layer, and the network layer.

Enterprise private network

An enterprise private network is a network built by a single organization to interconnect its office locations (e.g., production sites, head offices, remote offices, shops) in order to share computer resources.

Virtual private network

A virtual private network (VPN) is a computer network in which some of the links between nodes are carried by open connections or virtual circuits in some larger network (e.g., the Internet) instead of by physical wires. The data link layer protocols of the virtual network are said to be tunneled through the larger network when this is the case. One common application is secure communications through the public Internet, but a VPN need not have explicit security features, such as authentication or content encryption. VPNs, for example, can be used to separate the traffic of different user communities over an underlying network with strong security features.

VPN may have best-effort performance, or may have a defined service level agreement (SLA) between the VPN customer and the VPN service provider. Generally, a VPN has a topology more complex than point-to-point.

Organizational scope

Networks are typically managed by the organizations that own them. Private enterprise networks may use a combination of intranets and extranets. They may also provide network access to the Internet, which has no single owner and permits virtually unlimited global connectivity.

Intranets and extranets[edit source | edit]

Intranets and extranets are parts or extensions of a computer network, usually a LAN.

An intranet is a set of networks that are under the control of a single administrative entity. The intranet uses the IP protocol and IP-based tools such as web browsers and file transfer applications. The administrative entity limits use of the intranet to its authorized users. Most commonly, an intranet is the internal network of an organization. A large intranet will typically have at least one web server to provide users with organizational information.

An extranet is a network that is also under the administrative control of a single organization, but supports a limited connection to a specific external network. For example, an organization may provide access to some aspects of its intranet to share data with its business partners or customers. These other entities are not necessarily trusted from a security standpoint. Network connection to an extranet is often, but not always, implemented via WAN technology.

Internetwork

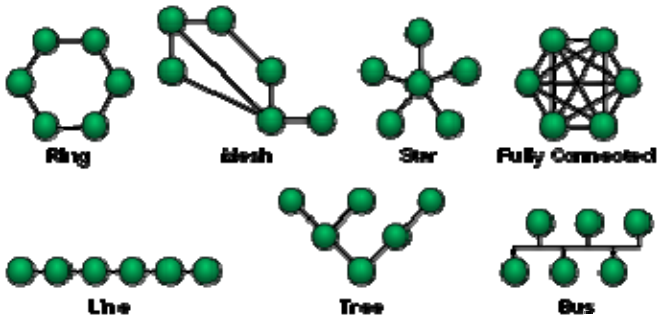
An internetwork is the connection of multiple computer networks via a common routing technology using routers.

Internet

The Internet is the largest example of an internetwork. It is a global system of interconnected governmental, academic, corporate, public, and private computer networks. It is based on the network technologies of the Internet Protocol Suite. It is the successor of the Advanced Research Projects Agency Network (ARPANET) developed by DARPA of the United States Department of Defense. The Internet is also the communications backbone underlying the World Wide Web (WWW).

Participants in the Internet use a diverse array of methods of several hundred documented, and often standardized, protocols compatible with the Internet Protocol Suite and an addressing system (IP addresses) administered by the Internet Assigned Numbers Authority and address registries. Service providers and large enterprises exchange information about the reachability of their address spaces through the Border Gateway Protocol (BGP), forming a redundant worldwide mesh of transmission paths.

Network topology



Network Topologies examples

Network topology is the layout or organizational hierarchy of interconnected nodes of a computer network.

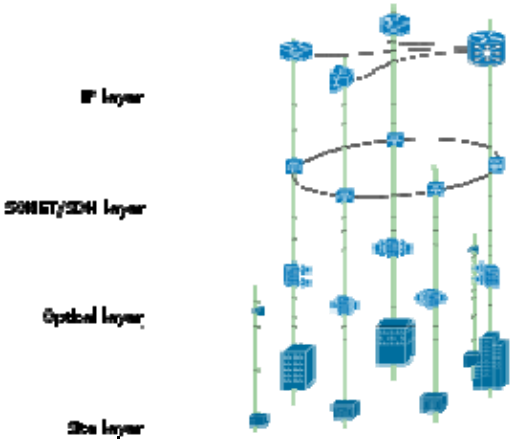
Common layouts are:

- A bus network: all nodes are connected to a common medium along this medium. This was the layout used in the original Ethernet, called 10BASE5 and 10BASE2.
- A star network: all nodes are connected to a special central node. This is the typical layout found in a Wireless LAN, where each wireless client connects to the central Wireless access point.
- A ring network: each node is connected to its left and right neighbour node, such that all nodes are connected and that each node can reach each other node by traversing nodes left- or rightwards. The Fiber Distributed Data Interface (FDDI) made use of such a topology.
- A mesh network: each node is connected to an arbitrary number of neighbours in such a way that there is at least one traversal from any node to any other.
- A fully connected network: each node is connected to every other node in the network.

Note that the physical layout of the nodes in a network may not necessarily reflect the network topology. As an example, with FDDI, the network topology is a ring (actually two counter-rotating rings), but the physical topology is a star, because all neighboring connections are routed via a central physical location.

Overlay network

An overlay network is a virtual computer network that is built on top of another network. Nodes in the overlay network are connected by virtual or logical links. Each link corresponds to a path, perhaps through many physical links, in the underlying network. The topology of the overlay network may (and often does) differ from that of the underlying one.



A sample overlay network: IP over SONET over Optical

For example, many peer-to-peer networks are overlay networks. They are organized as nodes of a virtual system of links that run on top of the Internet. The Internet was initially built as an overlay on the telephone network.

The most striking example of an overlay network is the Internet itself. At the network layer, each node can reach any other by a direct connection to the desired IP address, thereby creating a fully connected network. The underlying network, however, is composed of a mesh-like interconnect of sub-networks of varying topologies (and technologies). Address resolution and routing are the means that allow mapping of a fully connected IP overlay network to its underlying network.

Overlay networks have been around since the invention of network when computer systems were connected over telephone lines using modems, before any data network existed.

Another example of an overlay network is a distributed hash table, which maps keys to nodes in the network. In this case, the underlying network is an IP network, and the overlay network is a table (actually a map) indexed by keys.

Overlay networks have also been proposed as a way to improve Internet routing, such as through quality of service guarantees to achieve higher-quality streaming media. Previous proposals such as IntServ, DiffServ, and IP Multicast have not seen wide acceptance largely because they require modification of all routers in the network.[citation needed] On the other hand, an overlay network can be incrementally deployed on end-hosts running the overlay protocol software, without cooperation from Internet service providers. The overlay network has no control over how packets are routed in the underlying network between two overlay nodes, but it can control, for example, the

sequence of overlay nodes that a message traverses before it reaches its destination.
For example, Akamai Technologies manages an overlay network that provides reliable, efficient content delivery (a kind of multicast). Academic research includes end system multicast, resilient routing and quality of service studies, among others.

Network service

Network services are hosted by servers on a computer network, to provide some functionality for members or users of the network, or to help the network itself to operate. Services are usually based on a service protocol which defines the format and sequencing of messages between clients and servers of that network service.
Network services such as DNS (Domain Name System) give names for IP and MAC addresses (people remember names like “nm.lan” better than numbers like “210.121.67.18”), and DHCP to ensure that the equipment on the network has a valid IP address.
E-mail, printing and network file sharing services are also very common network services.

Network congestion

Network congestion occurs when a link or node is carrying so much data that its quality of service deteriorates. Typical effects include queueing delay, packet loss or the blocking of new connections. A consequence of these latter two is that incremental increases in offered load lead either only to small increase in network throughput, or to an actual reduction in network throughput.
Network protocols which use aggressive retransmissions to compensate for packet loss tend to keep systems in a state of network congestion even after the initial load has been reduced to a level which would not normally have induced network congestion. Thus, networks using these protocols can exhibit two stable states under the same level of load. The stable state with low throughput is known as congestive collapse.
Modern networks use congestion control and congestion avoidance techniques to try to avoid congestion collapse. These include: exponential backoff in protocols such as 802.11's CSMA/CA and the original Ethernet, window reduction in TCP, and fair queueing in devices such as routers. Another method to avoid the negative effects of network congestion is implementing priority schemes, so that some packets are transmitted with higher priority than others. Priority schemes do not solve network congestion by themselves, but they help to alleviate the effects of congestion for some services. An example of this is 802.1p. A third method to avoid network congestion is the explicit allocation of network resources to specific flows. One example of this is the use of Contention-Free Transmission Opportunities (CFTXOPs) in the ITU-T G.hn standard, which provides high-speed (up to 1 Gbit/s) Local area network over existing home wires (power lines, phone lines and coaxial cables).
RFC 2914 addresses the subject of congestion control in detail.

Network performance

Network performance refers to the measures of service quality of a telecommunications product as seen by the customer.
The following list gives examples of network performance measures for a circuit-switched network and one type of packet-switched network, viz. ATM:

- Circuit-switched networks: In circuit switched networks, network performance is synonymous with the grade of service. The number of rejected calls is a measure of how well the network is performing under heavy traffic loads. Other types of performance measures can include the level of noise and echo.

- ATM: In an Asynchronous Transfer Mode (ATM) network, performance can be measured by line rate, quality of service (QoS), data throughput, connect time, stability, technology, modulation technique and modem enhancements.

There are many ways to measure the performance of a network, as each network is different in nature and design. Performance can also be modelled instead of measured. For example, state transition diagrams are often used to model queuing performance in a circuit-switched network. These diagrams allow the network planner to analyze how the network will perform in each state, ensuring that the network will be optimally designed.

Network security

Network security consists of the provisions and policies adopted by the network administrator to prevent and monitor unauthorized access, misuse, modification, or denial of the computer network and its network-accessible resources. Network security is the authorization of access to data in a network, which is controlled by the network administrator. Users are assigned an ID and password that allows them access to information and programs within their authority. Network security is used on a variety of computer networks, both public and private, to secure daily transactions and communications among businesses, government agencies and individuals.

Network resilience

Network resilience is "the ability to provide and maintain an acceptable level of service in the face of faults and challenges to normal operation."

Views of networks

Users and network administrators typically have different views of their networks. Users can share printers and some servers from a workgroup, which usually means they are in the same geographic location and are on the same LAN, whereas a Network Administrator is responsible to keep that network up and running. A community of interest has less of a connection of being in a local area, and should be thought of as a set of arbitrarily located users who share a set of servers, and possibly also communicate via peer-to-peer technologies.

Network administrators can see networks from both physical and logical perspectives. The physical perspective involves geographic locations, physical cabling, and the network elements (e.g., routers, bridges and application layer gateways) that interconnect the physical media. Logical networks, called, in the TCP/IP architecture, subnets, map onto one or more physical media. For example, a common practice in a campus of buildings is to make a set of LAN cables in each building appear to be a common subnet, using virtual LAN (VLAN) technology.

Both users and administrators will be aware, to varying extents, of the trust and scope characteristics of a network. Again using TCP/IP architectural terminology, an intranet is a community of interest under private administration usually by an enterprise, and is only accessible

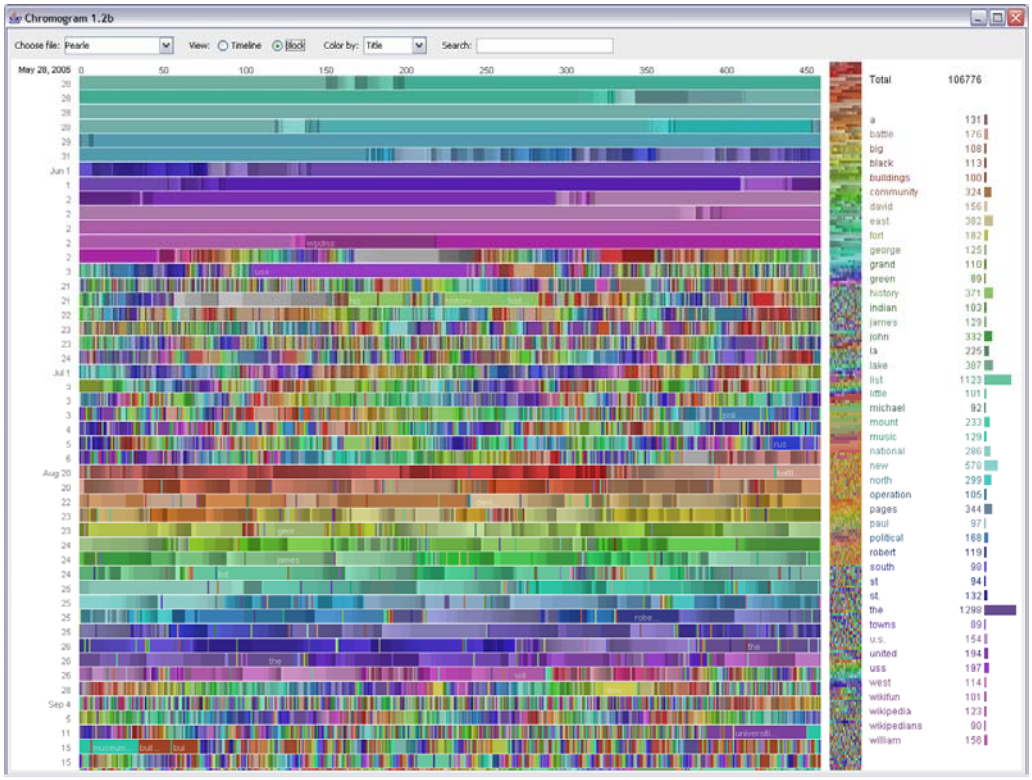
by authorized users (e.g. employees). Intranets do not have to be connected to the Internet, but generally have a limited connection. An extranet is an extension of an intranet that allows secure communications to users outside of the intranet (e.g. business partners, customers).
Unofficially, the Internet is the set of users, enterprises, and content providers that are interconnected by Internet Service Providers (ISP). From an engineering viewpoint, the Internet is the set of subnets, and aggregates of subnets, which share the registered IP address space and exchange information about the reachability of those IP addresses using the Border Gateway Protocol. Typically, the human-readable names of servers are translated to IP addresses, transparently to users, via the directory function of the Domain Name System (DNS).
Over the Internet, there can be business-to-business (B2B), business-to-consumer (B2C) and consumer-to-consumer (C2C) communications. When money or sensitive information is exchanged, the communications are apt to be protected by some form of communications security mechanism. Intranets and extranets can be securely superimposed onto the Internet, without any access by general Internet users and administrators, using secure Virtual Private Network (VPN) technology.

附件二

Description of Big Data

摘要：大量數據（Big data），或稱巨量資料、海量資料，指的是所涉及的資料量規模巨大到無法透過目前主流軟體工具，在合理時間內達到擷取、管理、處理、並整理成為幫助企業經營決策更積極目的的資訊。網路上每一筆搜索，網站上每一筆交易，敲打鍵盤，點擊滑鼠的每一個輸入都是數據，整理起來分析排行，它的功能可不僅僅止於事後被動了解市場，蒐集起來的資料還可以被規劃，引導開發更大的消費力量。

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."



A visualization created by IBM of Wikipedia edits. At multiple terabytes in size, the text and images of Wikipedia are a classic example of big data.

As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of exabytes of data. Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research. The limitations also affect Internet search, finance and business informatics. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification readers, and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012 every day 2.5 quintillion (2.5×1018) bytes of data

were created. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.
Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers".What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."

Definition

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target, as of 2012[update] ranging from a few dozen terabytes to many petabytes of data in a single data set. The target moves due to constant improvement in traditional DBMS technology as well as new databases like NoSQL and their ability to handle larger amounts of data. With this difficulty, new platforms of "big data" tools are being developed to handle various aspects of large quantities of data.
In a 2001 research report and related lectures, META Group (now Gartner) analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Gartner, and now much of the industry, continue to use this "3Vs" model for describing big data.In 2012, Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.". Additionally, a new V "Veracity" is added by some organizations to describe it.
If Gartner’s definition (the 3Vs) is still widely used, the growing maturity of the concept fosters a more sound difference between Big Data and Business Intelligence, regarding data and their use:

- Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends etc.;
- Big Data uses inductive statistics with data with low information density whose huge volume allow to infer laws (regressions…) and thus giving (with the limits of inference reasoning) to Big Data some predictive capabilities.

Examples

Examples include Big Science, RFID, sensor networks, social networks, big social data analysis (due to the social data revolution), Internet documents, Internet search indexing, call detail records, astronomy, atmospheric science, genomics, biogeochemical, biological, and other complex and often interdisciplinary scientific research, military surveillance, forecasting drive times for new home buyers, medical records, photography archives, video archives, and large-scale e-commerce.

Big science

The Large Hadron Collider experiments represent about 150 million sensors delivering data 40 million times per second. There are nearly 600 million collisions per second. After filtering and refraining from recording more than 99.999% of these streams, there are 100 collisions of interest per second.

- As a result, only working with less than 0.001% of the sensor stream data, the data flow from all four LHC experiments represents 25 petabytes annual rate before replication (as of 2012). This becomes nearly 200 petabytes after replication.
- If all sensor data were to be recorded in LHC, the data flow would be extremely hard to work with. The data flow would exceed 150 million petabytes annual rate, or nearly 500 exabytes per day, before replication. To put the number in perspective, this is equivalent to 500 quintillion (5×10^{20}) bytes per day, almost 200 times higher than all the other sources combined in the world.

Science and research

- When the Sloan Digital Sky Survey (SDSS) began collecting astronomical data in 2000, it amassed more in its first few weeks than all data collected in the history of astronomy. Continuing at a rate of about 200 GB per night, SDSS has amassed more than 140 terabytes of information. When the Large Synoptic Survey Telescope, successor to SDSS, comes online in 2016 it is anticipated to acquire that amount of data every five days.
- Decoding the human genome originally took 10 years to process, now it can be achieved in less than a week : the DNA sequencers have divided the sequencing cost by 10 000 in the last ten years, which is a factor 100 compared to Moore's Law.
- Computational social science— Tobias Preis *et al.* used Google Trends data to demonstrate that Internet users from countries with a higher per capita gross domestic product (GDP) are more likely to search for information about the future than information about the past. The findings suggest there may be a link between online behavior and real-world economic indicators. The authors of the study examined Google queries logs made by Internet users in 45 different countries in 2010 and calculated the ratio of the volume of searches for the coming year (‘2011’) to the volume of searches for the previous year (‘2009’), which they call the ‘future orientation index’. They compared the future orientation index to the per capita GDP of each country and found a strong tendency for countries in which Google users enquire more about the future to exhibit a higher GDP. The results hint that there may potentially be a relationship between the economic success of a country and the information-seeking behavior of its citizens captured in big data.
- The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.
- Tobias Preis and his colleagues Helen Susannah Moat and H. Eugene Stanley introduced a method to identify online precursors for stock market moves, using trading strategies based

on search volume data provided by Google Trends. Their analysis of Google search volume for 98 terms of varying financial relevance, published in *Scientific Reports*, suggests that increases in search volume for financially relevant search terms tend to precede large losses in financial markets.

Government

- In 2012, the Obama administration announced the Big Data Research and Development Initiative, which explored how big data could be used to address important problems faced by the government. The initiative was composed of 84 different big data programs spread across six departments.
- Big data analysis played a large role in Barack Obama's successful 2012 re-election campaign.
- The United States Federal Government owns six of the ten most powerful supercomputers in the world.
- The Utah Data Center is a data center currently being constructed by the United States National Security Agency. When finished, the facility will be able to handle yottabytes of information collected by the NSA over the Internet.

Private sector

- Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB.
- Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data – the equivalent of 167 times the information contained in all the books in the US Library of Congress.
- Facebook handles 50 billion photos from its user base.
- FICO Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide.
- The volume of business data worldwide, across all companies, doubles every 1.2 years, according to estimates.
- Windermere Real Estate uses anonymous GPS signals from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day.

International development

Following decades of work in the area of the effective usage of information and communication technologies for development (or ICT4D), it has been suggested that Big Data can make important

contributions to international development. On the one hand, the advent of Big Data delivers the cost-effective prospect to improve decision-making in critical development areas such as health care, employment, economic productivity, crime and security, and natural disaster and resource management. On the other hand, all the well-known concerns of the Big Data debate, such as privacy, interoperability challenges, and the almighty power of imperfect algorithms, are aggravated in developing countries by long-standing development challenges like lacking technological infrastructure and economic and human resource scarcity. "This has the potential to result in a new kind of digital divide: a divide in data-based intelligence to inform decision-making."

Market

"Big data" has increased the demand of information management specialists in that Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC, and HP have spent more than $15 billion on software firms only specializing in data management and analytics. In 2010, this industry on its own was worth more than $100 billion and was growing at almost 10 percent a year: about twice as fast as the software business as a whole.

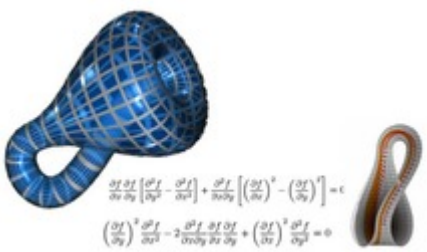
Developed economies make increasing use of data-intensive technologies. There are 4.6 billion mobile-phone subscriptions worldwide and there are between 1 billion and 2 billion people accessing the internet. Between 1990 and 2005, more than 1 billion people worldwide entered the middle class which means more and more people who gain money will become more literate which in turn leads to information growth. The world's effective capacity to exchange information through telecommunication networks was 281 petabytes in 1986, 471 petabytes in 1993, 2.2 exabytes in 2000, 65 exabytes in 2007 and it is predicted that the amount of traffic flowing over the internet will reach 667 exabytes annually by 2013.

Architecture

In 2004, Google published a paper on a process called MapReduce that used such an architecture. MapReduce framework provides a parallel programming model and associated implementation to process huge amount of data. With MapReduce, queries are split and distributed across parallel nodes and processed in parallel (the Map step). The results are then gathered and delivered (the Reduce step). The framework was incredibly successful,[clarification needed] so others wanted to replicate the algorithm. Therefore, an implementation of MapReduce framework was adopted by an Apache open source project named Hadoop.

MIKE2.0 is an open approach to information management. The methodology addresses handling big data in terms of useful permutations of data sources, complexity in interrelationships, and difficulty in deleting (or modifying) individual records.

Technologies



DARPA's Topological Data Analysis program (showing a Klein bottle) seeks the fundamental structure of massive data sets.

Big data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times. A 2011 McKinsey report suggests suitable technologies include A/B testing, association rule learning, classification, cluster analysis, crowdsourcing, data fusion and integration, ensemble learning, genetic algorithms, machine learning, natural language processing, neural networks, pattern recognition, anomaly detection, predictive modelling, regression, sentiment analysis, signal processing, supervised and unsupervised learning, simulation, time series analysis and visualisation. Multidimensional big data can also be represented as tensors, which can be more efficiently handled by tensor-based computation, such as multilinear subspace learning. Additional technologies being applied to big data include massively parallel-processing (MPP) databases, search-based applications, data-mining grids, distributed file systems, distributed databases, cloud based infrastructure (applications, storage and computing resources) and the Internet.[citation needed]

Some but not all MPP relational databases have the ability to store and manage petabytes of data. Implicit is the ability to load, monitor, back up, and optimize the use of the large data tables in the RDBMS.

DARPA's Topological Data Analysis program seeks the fundamental structure of massive data sets and in 2008 the technology went public with the launch of a company called Ayasdi.

The practitioners of big data analytics processes are generally hostile to slower shared storage, preferring direct-attached storage (DAS) in its various forms from solid state drive (SSD) to high capacity SATA disk buried inside parallel processing nodes. The perception of shared storage architectures—SAN and NAS—is that they are relatively slow, complex, and expensive. These qualities are not consistent with big data analytics systems that thrive on system performance, commodity infrastructure, and low cost.

Real or near-real time information delivery is one of the defining characteristics of big data analytics. Latency is therefore avoided whenever and wherever possible. Data in memory is good—data on spinning disk at the other end of a FC SAN connection is not. The cost of a SAN at the scale needed for analytics applications is very much higher than other storage techniques.

There are advantages as well as disadvantages to shared storage in big data analytics, but big data analytics practitioners as of 2011 did not favor it.

Research activities

In March 2012, The White House announced a national "Big Data Initiative" that consisted of six Federal departments and agencies committing more than $200 million to big data research projects. The initiative included a National Science Foundation "Expeditions in Computing" grant of $10 million over 5 years to the AMPLab at the University of California, Berkeley. The AMPLab also received funds from DARPA, and over a dozen industrial sponsors and uses big data to attack a wide range of problems from predicting traffic congestion to fighting cancer.

The White House Big Data Initiative also included a commitment by the Department of Energy to provide $25 million in funding over 5 years to establish the Scalable Data Management, Analysis

and Visualization (SDAV) Institute, led by the Energy Department's Lawrence Berkeley National Laboratory. The SDAV Institute aims to bring together the expertise of six national laboratories and seven universities to develop new tools to help scientists manage and visualize data on the Department's supercomputers.

The U.S. state of Massachusetts announced the Massachusetts Big Data Initiative in May 2012, which provides funding from the state government and private companies to a variety of research institutions. The Massachusetts Institute of Technology hosts the Intel Science and Technology Center for Big Data in the MIT Computer Science and Artificial Intelligence Laboratory, combining government, corporate, and institutional funding and research efforts.

The European Commission is funding a 2-year-long Big Data Public Private Forum through their Seventh Framework Program to engage companies, academics and other stakeholders in discussing Big Data issues. The project aims to define a strategy in terms of research and innovation to guide supporting actions from the European Commission in the successful implementation of the Big Data economy. Outcomes of this project will be used as input for Horizon 2020, their next framework program.

The IBM sponsored 37th annual "Battle of the Brains" student Big Data championship will be held in July 2013. The inaugural professional 2014 Big Data World Championship is to be held in Dallas, Texas.

Critique

Critiques of the Big Data paradigm come in two flavors, those that question the implications of the approach itself, and those that question the way it is currently done.

Critiques of the Big Data paradigm

"A crucial problem is that we do not know much about the underlying empirical micro-processes that lead to the emergence of the[se] typical network characteristics of Big Data". In their critique, Snijders, Matzat, and Reips point out that often very strong assumptions are made about mathematical properties that may not at all reflect what is really going on at the level of micro-processes. Mark Graham has leveled broad critiques at Chris Anderson's assertion that big data will spell the end of theory: focusing in particular on the notion that big data will always need to be contextualized in their social, economic and political contexts. Even as companies invest eight- and nine-figure sums to derive insight from information streaming in from suppliers and customers, less than 40% of employees have sufficiently mature processes and skills to do so. To overcome this insight deficit, "big data", no matter how comprehensive or well analyzed, needs to be complemented by "big judgment", according to an article in the Harvard Business Review.

Much in the same line, it has been pointed out that the decisions based on the analysis of Big Data are inevitably "informed by the world as it was in the past, or, at best, as it currently is". Fed by a large number of data on past experiences, algorithms can predict future development if the future is similar to the past. If the systems dynamics of the future change, the past can say little about the future. For this, it would be necessary to have a thorough understanding of the systems dynamic, which implies theory. As a response to this critique it has been suggested to combine Big Data approaches with computer simulations, such as agent-based models, for example. Those are

increasingly getting better in predicting the outcome of social complexities of even unknown future scenarios through computer simulations that are based on a collection of mutually interdependent algorithms.[citation needed] In addition, use of multivariate methods that probe for the latent structure of the data, such as factor analysis and cluster analysis, have proven useful as analytic approaches that go well beyond the bi-variate approaches (cross-tabs) typically employed with smaller data sets.

In Health and biology, conventional scientific approaches are based on experimentation. For these approaches, the limiting factor are the relevant data that can confirm or refute the initial hypothesis. A new postulate is accepted now in biosciences : the information provided by the data in huge volumes (omics) without prior hypothesis is complementary and sometimes necessary to conventional approaches based on experimentation. In the massive approaches it is the formulation of a relevant hypothesis to explain the data that is the limiting factor. The search logic is reversed and the limits of induction ("Glory of Science and Philosophy scandal", C. D. Broad, 1926) to be considered.

Consumer privacy advocates are concerned about the threat to privacy represented by increasing storage and integration of personally identifiable information; expert panels have released various policy recommendations to conform practice to expectations of privacy.

Critiques of Big Data execution

Danah Boyd has raised concerns about the use of big data in science neglecting principles such as choosing a representative sample by being too concerned about actually handling the huge amounts of data. This approach may lead to results bias in one way or another. Integration across heterogeneous data resources— some that might be considered "big data" and others not — presents formidable logistical as well as analytical challenges, but many researchers argue that such integrations are likely to represent the most promising new frontiers in science.

附件三

Description of Data Mining

摘要：資料探勘（Data mining），亦稱為數據挖掘、資料挖掘、資料採礦。它是資料庫知識發現（Knowledge-Discovery in Databases，KDD）中的一個步驟。資料探勘一般是指從大量的資料中自動搜尋隱藏於其中的有著特殊關聯性（屬於 Association rule learning）的訊息的過程。資料挖掘通常與電腦科學有關，並通過統計、線上分析處理、情報檢索、機器學習、專家系統（依靠過去的經驗法則）和模式識別等諸多方法來實現上述目標。

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The term is a buzzword, and is frequently misused to mean any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) but is also generalized to any kind of computer decision support system, including artificial intelligence, machine learning, and business intelligence. In the proper use of the word, the key term is discovery[citation needed], commonly defined as "detecting something new". Even the popular book "Data mining: Practical machine learning tools and techniques with Java" (which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the term "data mining" was only added for marketing reasons. Often the more general terms "(large scale) data analysis", or "analytics" – or when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms data dredging, data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

Data mining uses information from past data to analyze the outcome of a particular problem or situation that may arise. Data mining works to analyze data stored in data warehouses that are used to store that data that is being analyzed. That particular data may come from all parts of business, from the production to the management. Managers also use data mining to decide upon marketing

strategies for their product. They can use data to compare and contrast among competitors. Data mining interprets its data into real time analysis that can be used to increase sales, promote new product, or delete product that is not value-added to the company.

Etymology

In the 1960s, statisticians used terms like "Data Fishing" or "Data Dredging" to refer to what they considered the bad practice of analyzing data without an a-priori hypothesis. The term "Data Mining" appeared around 1990 in the database community. At the beginning of the century, there was a phrase "database mining"™, trademarked by HNC, a San Diego-based company (now merged into FICO), to pitch their Data Mining Workstation; researchers consequently turned to "data mining". Other terms used include Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, etc. Gregory Piatetsky-Shapiro coined the term "Knowledge Discovery in Databases" for the first workshop on the same topic (1989) and this term became more popular in AI and Machine Learning Community. However, the term data mining became more popular in the business and press communities.Currently, Data Mining and Knowledge Discovery are used interchangeably.

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Bayes' theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. As data sets have grown in size and complexity, direct "hands-on" data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees (1960s), and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns in large data sets. It bridges the gap from applied statistics and artificial intelligence (which usually provide the mathematical background) to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets.

The premier professional body in the field is the Association for Computing Machinery's (ACM) Special Interest Group (SIG) on Knowledge Discovery and Data Mining (SIGKDD). Since 1989 this ACM SIG has hosted an annual international conference and published its proceedings, and since 1999 it has published a biannual academic journal titled "SIGKDD Explorations".

Computer science conferences on data mining include:

- CIKM Conference – ACM Conference on Information and Knowledge Management
- DMIN Conference – International Conference on Data Mining
- DMKD Conference – Research Issues on Data Mining and Knowledge Discovery
- ECDM Conference – European Conference on Data Mining
- ECML-PKDD Conference – European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
- EDM Conference – International Conference on Educational Data Mining

- ICDM Conference – IEEE International Conference on Data Mining
- KDD Conference – ACM SIGKDD Conference on Knowledge Discovery and Data Mining
- MLDM Conference – Machine Learning and Data Mining in Pattern Recognition
- PAKDD Conference – The annual Pacific-Asia Conference on Knowledge Discovery and Data Mining
- PAW Conference – Predictive Analytics World
- SDM Conference – SIAM International Conference on Data Mining (SIAM)
- SSTD Symposium – Symposium on Spatial and Temporal Databases
- WSDM Conference – ACM Conference on Web Search and Data Mining

Data mining topics are also present on many data management/database conferences such as the ICDE Conference, SIGMOD Conference and International Conference on Very Large Data Bases

Process

The Knowledge Discovery in Databases (KDD) process is commonly defined with the stages:

(1) Selection
(2) Pre-processing
(3) Transformation
(4) *Data Mining*
(5) Interpretation/Evaluation.

It exists, however, in many variations on this theme, such as the Cross Industry Standard Process for Data Mining (CRISP-DM) which defines six phases:

(1) Business Understanding
(2) Data Understanding
(3) Data Preparation
(4) Modeling
(5) Evaluation
(6) Deployment

or a simplified process such as (1) pre-processing, (2) data mining, and (3) results validation. Polls conducted in 2002, 2004, and 2007 show that the CRISP-DM methodology is the leading methodology used by data miners. The only other data mining standard named in these polls was SEMMA. However, 3-4 times as many people reported using CRISP-DM. Several teams of researchers have published reviews of data mining process models, and Azevedo and Santos conducted a comparison of CRISP-DM and SEMMA in 2008.

Pre-processing

Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a data mart or data warehouse. Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data.

Data mining

Data mining involves six common classes of tasks:

- Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – Attempts to find a function which models the data with the least error.
- Summarization – providing a more compact representation of the data set, including visualization and report generation.
- Sequential pattern mining – Sequential pattern mining finds sets of data items that occur together frequently in some sequences. Sequential pattern mining, which extracts frequent subsequences from a sequence database, has attracted a great deal of interest during the recent data mining research because it is the basis of many applications, such as: web user analysis, stock trend prediction, DNA sequence analysis, finding language or linguistic patterns from natural language texts, and using the history of symptoms to predict certain kind of disease.

The final step of knowledge discovery from data is to verify that the patterns produced by the data mining algorithms occur in the wider data set. Not all patterns found by the data mining algorithms are necessarily valid. It is common for the data mining algorithms to find patterns in the training set which are not present in the general data set. This is called overfitting. To overcome this, the evaluation uses a test set of data on which the data mining algorithm was not trained. The learned patterns are applied to this test set and the resulting output is compared to the desired output. For example, a data mining algorithm trying to distinguish "spam" from "legitimate" emails would be trained on a training set of sample e-mails. Once trained, the learned patterns would be applied to the test set of e-mails on which it had not been trained. The accuracy of the patterns can then be measured from how many e-mails they correctly classify. A number of statistical methods may be

used to evaluate the algorithm, such as ROC curves.
If the learned patterns do not meet the desired standards, then it is necessary to re-evaluate and change the pre-processing and data mining steps. If the learned patterns do meet the desired standards, then the final step is to interpret the learned patterns and turn them into knowledge.
There have been some efforts to define standards for the data mining process, for example the 1999 European Cross Industry Standard Process for Data Mining (CRISP-DM 1.0) and the 2004 Java Data Mining standard (JDM 1.0). Development on successors to these processes (CRISP-DM 2.0 and JDM 2.0) was active in 2006, but has stalled since. JDM 2.0 was withdrawn without reaching a final draft.
For exchanging the extracted models – in particular for use in predictive analytics –the key standard is the Predictive Model Markup Language (PMML), which is an XML-based language developed by the Data Mining Group (DMG) and supported as exchange format by many data mining applications. As the name suggests, it only covers prediction models, a particular data mining task of high importance to business applications. However, extensions to cover (for example) subspace clustering have been proposed independently of the DMG.

Notable uses

Science and engineering

In recent years, data mining has been used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering.
In the study of human genetics, sequence mining helps address the important goal of understanding the mapping relationship between the inter-individual variations in human DNA sequence and the variability in disease susceptibility. In simple terms, it aims to find out how the changes in an individual's DNA sequence affects the risks of developing common diseases such as cancer, which is of great importance to improving methods of diagnosing, preventing, and treating these diseases. The data mining method that is used to perform this task is known as multifactor dimensionality reduction.
In the area of electrical power engineering, data mining methods have been widely used for condition monitoring of high voltage electrical equipment. The purpose of condition monitoring is to obtain valuable information on, for example, the status of the insulation (or other important safety-related parameters). Data clustering techniques – such as the self-organizing map (SOM), have been applied to vibration monitoring and analysis of transformer on-load tap-changers (OLTCS). Using vibration monitoring, it can be observed that each tap change operation generates a signal that contains information about the condition of the tap changer contacts and the drive mechanisms. Obviously, different tap positions will generate different signals. However, there was considerable variability amongst normal condition signals for exactly the same tap position. SOM has been applied to detect abnormal conditions and to hypothesize about the nature of the abnormalities.
Data mining methods have also been applied to dissolved gas analysis (DGA) in power transformers. DGA, as a diagnostics for power transformers, has been available for many years. Methods such as SOM has been applied to analyze generated data and to determine trends which

are not obvious to the standard DGA ratio methods (such as Duval Triangle).
Another example of data mining in science and engineering is found in educational research, where data mining has been used to study the factors leading students to choose to engage in behaviors which reduce their learning, and to understand factors influencing university student retention.A similar example of social application of data mining is its use in expertise finding systems, whereby descriptors of human expertise are extracted, normalized, and classified so as to facilitate the finding of experts, particularly in scientific and technical fields. In this way, data mining can facilitate institutional memory.
Other examples of application of data mining methods are biomedical data facilitated by domain ontologies, mining clinical trial data, and traffic analysis using SOM.
In adverse drug reaction surveillance, the Uppsala Monitoring Centre has, since 1998, used data mining methods to routinely screen for reporting patterns indicative of emerging drug safety issues in the WHO global database of 4.6 million suspected adverse drug reaction incidents. Recently, similar methodology has been developed to mine large collections of electronic health records for temporal patterns associating drug prescriptions to medical diagnoses.
Data mining has been applied software artifacts within the realm of software engineering: Mining Software Repositories.

Human rights

Data mining of government records – particularly records of the justice system (i.e. courts, prisons) – enables the discovery of systemic human rights violations in connection to generation and publication of invalid or fraudulent legal records by various government agencies.

Medical data mining

In 2011, the case of Sorrell v. IMS Health, Inc., decided by the Supreme Court of the United States, ruled that pharmacies may share information with outside companies. This practice was authorized under the 1st Amendment of the Constitution, protecting the "freedom of speech."

Spatial data mining

Spatial data mining is the application of data mining methods to spatial data. The end objective of spatial data mining is to find patterns in data with respect to geography. So far, data mining and Geographic Information Systems (GIS) have existed as two separate technologies, each with its own methods, traditions, and approaches to visualization and data analysis. Particularly, most contemporary GIS have only very basic spatial analysis functionality. The immense explosion in geographically referenced data occasioned by developments in IT, digital mapping, remote sensing, and the global diffusion of GIS emphasizes the importance of developing data-driven inductive approaches to geographical analysis and modeling.
Data mining offers great potential benefits for GIS-based applied decision-making. Recently, the task of integrating these two technologies has become of critical importance, especially as various public and private sector organizations possessing huge databases with thematic and geographically referenced data begin to realize the huge potential of the information contained therein. Among those organizations are:

- offices requiring analysis or dissemination of geo-referenced statistical data
- public health services searching for explanations of disease clustering
- environmental agencies assessing the impact of changing land-use patterns on climate change
- geo-marketing companies doing customer segmentation based on spatial location.

Challenges in Spatial mining: Geospatial data repositories tend to be very large. Moreover, existing GIS datasets are often splintered into feature and attribute components that are conventionally archived in hybrid data management systems. Algorithmic requirements differ substantially for relational (attribute) data management and for topological (feature) data management. Related to this is the range and diversity of geographic data formats, which present unique challenges. The digital geographic data revolution is creating new types of data formats beyond the traditional "vector" and "raster" formats. Geographic data repositories increasingly include ill-structured data, such as imagery and geo-referenced multi-media.
There are several critical research challenges in geographic knowledge discovery and data mining. Miller and Han offer the following list of emerging research topics in the field:

- **Developing and supporting geographic data warehouses (GDW's)**: Spatial properties are often reduced to simple aspatial attributes in mainstream data warehouses. Creating an integrated GDW requires solving issues of spatial and temporal data interoperability – including differences in semantics, referencing systems, geometry, accuracy, and position.

- **Better spatio-temporal representations in geographic knowledge discovery**: Current geographic knowledge discovery (GKD) methods generally use very simple representations of geographic objects and spatial relationships. Geographic data mining methods should recognize more complex geographic objects (i.e. lines and polygons) and relationships (i.e. non-Euclidean distances, direction, connectivity, and interaction through attributed geographic space such as terrain). Furthermore, the time dimension needs to be more fully integrated into these geographic representations and relationships.

- **Geographic knowledge discovery using diverse data types**: GKD methods should be developed that can handle diverse data types beyond the traditional raster and vector models, including imagery and geo-referenced multimedia, as well as dynamic data types (video streams, animation).

Sensor data mining
Wireless sensor networks can be used for facilitating the collection of data for spatial data mining for a variety of applications such as air pollution monitoring.A characteristic of such networks is that nearby sensor nodes monitoring an environmental feature typically register similar values. This kind of data redundancy due to the spatial correlation between sensor observations inspires the techniques for in-network data aggregation and mining. By measuring the spatial correlation

between data sampled by different sensors, a wide class of specialized algorithms can be developed to develop more efficient spatial data mining algorithms.

Visual data mining

In the process of turning from analogical into digital, large data sets have been generated, collected, and stored discovering statistical patterns, trends and information which is hidden in data, in order to build predictive patterns. Studies suggest visual data mining is faster and much more intuitive than is traditional data mining. See also Computer vision.

Music data mining

Data mining techniques, and in particular co-occurrence analysis, has been used to discover relevant similarities among music corpora (radio lists, CD databases) for the purpose of classifying music into genres in a more objective manner.

Surveillance

Data mining has been used to fight terrorism by the U.S. government. Programs include the Total Information Awareness (TIA) program, Secure Flight (formerly known as Computer-Assisted Passenger Prescreening System (CAPPS II)), Analysis, Dissemination, Visualization, Insight, Semantic Enhancement (ADVISE), and the Multi-state Anti-Terrorism Information Exchange (MATRIX). These programs have been discontinued due to controversy over whether they violate the 4th Amendment to the United States Constitution, although many programs that were formed under them continue to be funded by different organizations or under different names.

In the context of combating terrorism, two particularly plausible methods of data mining are "pattern mining" and "subject-based data mining".

Pattern mining

"Pattern mining" is a data mining method that involves finding existing patterns in data. In this context patterns often means association rules. The original motivation for searching association rules came from the desire to analyze supermarket transaction data, that is, to examine customer behavior in terms of the purchased products. For example, an association rule "beer $\Rightarrow$ potato chips (80%)" states that four out of five customers that bought beer also bought potato chips.

In the context of pattern mining as a tool to identify terrorist activity, the National Research Council provides the following definition: "Pattern-based data mining looks for patterns (including anomalous data patterns) that might be associated with terrorist activity — these patterns might be regarded as small signals in a large ocean of noise." Pattern Mining includes new areas such a Music Information Retrieval (MIR) where patterns seen both in the temporal and non temporal domains are imported to classical knowledge discovery search methods.

Subject-based data mining

"Subject-based data mining" is a data mining method involving the search for associations between individuals in data. In the context of combating terrorism, the National Research Council provides the following definition: "Subject-based data mining uses an initiating individual or other datum that is considered, based on other information, to be of high interest, and the goal is to determine what other persons or financial transactions or movements, etc., are related to that initiating datum."

Knowledge grid

Knowledge discovery "On the Grid" generally refers to conducting knowledge discovery in an open environment using grid computing concepts, allowing users to integrate data from various online data sources, as well make use of remote resources, for executing their data mining tasks. The earliest example was the Discovery Net, developed at Imperial College London, which won the "Most Innovative Data-Intensive Application Award" at the ACM SC02 (Supercomputing 2002) conference and exhibition, based on a demonstration of a fully interactive distributed knowledge discovery application for a bioinformatics application. Other examples include work conducted by researchers at the University of Calabria, who developed a Knowledge Grid architecture for distributed knowledge discovery, based on grid computing.

Data mining can be misused, and can also unintentionally produce results which appear significant but which do not actually predict future behavior and cannot be reproduced on a new sample of data.

Some people believe that data mining itself is ethically neutral. While the term "data mining" has no ethical implications, it is often associated with the mining of information in relation to peoples' behavior (ethical and otherwise). To be precise, data mining is a statistical method that is applied to a set of information (i.e. a data set). Associating these data sets with people is an extreme narrowing of the types of data that are available. Examples could range from a set of crash test data for passenger vehicles, to the performance of a group of stocks. These types of data sets make up a great proportion of the information available to be acted on by data mining methods, and rarely have ethical concerns associated with them. However, the ways in which data mining can be used can in some cases and contexts raise questions regarding privacy, legality, and ethics. In particular, data mining government or commercial data sets for national security or law enforcement purposes, such as in the Total Information Awareness Program or in ADVISE, has raised privacy concerns.

Data mining requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. A common way for this to occur is through data aggregation. Data aggregation involves combining data together (possibly from various sources) in a way that facilitates analysis (but that also might make identification of private, individual-level data deducible or otherwise apparent). This is not data mining per se, but a result of the preparation of data before – and for the purposes of – the analysis. The threat to an individual's privacy comes into play when the data, once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when the data were originally anonymous.

It is recommended that an individual is made aware of the following before data are collected:

- the purpose of the data collection and any (known) data mining projects
- how the data will be used
- who will be able to mine the data and use the data and their derivatives
- the status of security surrounding access to the data
- how collected data can be updated.

In America, privacy concerns have been addressed to some extent by the US Congress via the

passage of regulatory controls such as the Health Insurance Portability and Accountability Act (HIPAA). The HIPAA requires individuals to give their "informed consent" regarding information they provide and its intended present and future uses. According to an article in Biotech Business Week', "'[i]n practice, HIPAA may not offer any greater protection than the longstanding regulations in the research arena,' says the AAHC. More importantly, the rule's goal of protection through informed consent is undermined by the complexity of consent forms that are required of patients and participants, which approach a level of incomprehensibility to average individuals." This underscores the necessity for data anonymity in data aggregation and mining practices.
Data may also be modified so as to become anonymous, so that individuals may not readily be identified. However, even "de-identified"/"anonymized" data sets can potentially contain enough information to allow identification of individuals, as occurred when journalists were able to find several individuals based on a set of search histories that were inadvertently released by AOL.

Free open-source data mining software and applications

- Carrot2: Text and search results clustering framework.
- Chemicalize.org: A chemical structure miner and web search engine.
- ELKI: A university research project with advanced cluster analysis and outlier detection methods written in the Java language.
- GATE: a natural language processing and language engineering tool.
- SCaViS: Java cross-platform data analysis framework developed at Argonne National Laboratory.
- KNIME: The Konstanz Information Miner, a user friendly and comprehensive data analytics framework.
- ML-Flex: A software package that enables users to integrate with third-party machine-learning packages written in any programming language, execute classification analyses in parallel across multiple computing nodes, and produce HTML reports of classification results.
- NLTK (Natural Language Toolkit): A suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the Python language.
- SenticNet API: A semantic and affective resource for opinion mining and sentiment analysis.
- Orange: A component-based data mining and machine learning software suite written in the Python language.
- R: A programming language and software environment for statistical computing, data mining, and graphics. It is part of the GNU Project.
- RapidMiner: An environment for machine learning and data mining experiments.
- UIMA: The UIMA (Unstructured Information Management Architecture) is a component framework for analyzing unstructured content such as text, audio and video – originally developed by IBM.
- Weka: A suite of machine learning software applications written in the Java programming language.

Commercial data-mining software and applications

- Angoss KnowledgeSTUDIO: data mining tool provided by Angoss.
- BIRT Analytics: visual data mining and predictive analytics tool provided by Actuate Corporation.
- Clarabridge: enterprise class text analytics solution.
- IBM DB2 Intelligent Miner: in-database data mining platform provided by IBM, with modeling, scoring and visualization services based on the SQL/MM - PMML framework.
- IBM SPSS Modeler: data mining software provided by IBM.
- KXEN Modeler: data mining tool provided by KXEN.
- LIONsolver: an integrated software application for data mining, business intelligence, and modeling that implements the Learning and Intelligent OptimizatioN (LION) approach.
- Microsoft Analysis Services: data mining software provided by Microsoft.
- Oracle Data Mining: data mining software by Oracle.
- Predixion Insight: data mining software by Predixion Software.
- SAS Enterprise Miner: data mining software provided by the SAS Institute.
- STATISTICA Data Miner: data mining software provided by StatSoft.

附件四

Description of Project Management

摘要：探討計畫管理的五大流程及九大知識領域、研究如何建立計畫的 critical path 及如何用 excel solver 來規劃求得實際 business 的最佳解，其目的就是運用上述知識、技術、工具和方法來組織計畫活動，使能符合計畫的需求。

Project management is the discipline of planning, organizing, motivating, and controlling resources to achieve specific goals. A project is a temporary endeavor with a defined beginning and end (usually time-constrained, and often constrained by funding or deliverables), undertaken to meet unique goals and objectives, typically to bring about beneficial change or added value. The temporary nature of projects stands in contrast with business as usual (or operations), which are repetitive, permanent, or semi-permanent functional activities to produce products or services. In practice, the management of these two systems is often quite different, and as such requires the development of distinct technical skills and management strategies.

The primary challenge of project management is to achieve all of the project goals and objectives while honoring the preconceived constraints. The primary constraints are scope, time, quality and budget. The secondary —and more ambitious— challenge is to optimize the allocation of necessary inputs and integrate them to meet pre-defined objectives.

Approaches

There are a number of approaches to managing project activities including lean, iterative, incremental, and phased approaches.

Regardless of the methodology employed, careful consideration must be given to the overall project objectives, timeline, and cost, as well as the roles and responsibilities of all participants and stakeholders

The traditional approach

A traditional phased approach identifies a sequence of steps to be completed. In the "traditional approach",five developmental components of a project can be distinguished (four stages plus control):

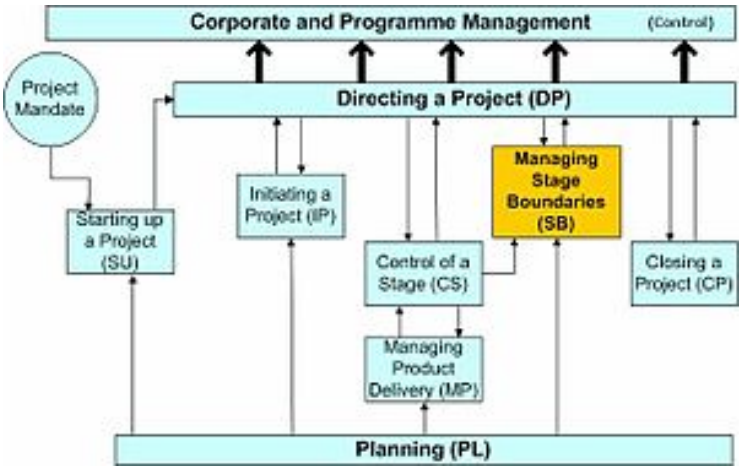
Typical development phases of an engineering project

1. initiation
2. planning and design
3. execution and construction
4. monitoring and controlling systems
5. completion

Not all projects will have every stage, as projects can be terminated before they reach completion. Some projects do not follow a structured planning and/or monitoring process. And some projects will go through steps 2, 3 and 4 multiple times.

Many industries use variations of these project stages. For example, when working on a brick-and-mortar design and construction, projects will typically progress through stages like pre-planning, conceptual design, schematic design, design development, construction drawings (or contract documents), and construction administration. In software development, this approach is often known as the waterfall model, i.e., one series of tasks after another in linear sequence. In software development many organizations have adapted the Rational Unified Process (RUP) to fit this methodology, although RUP does not require or explicitly recommend this practice. Waterfall development works well for small, well defined projects, but often fails in larger projects of undefined and ambiguous nature. The Cone of Uncertainty explains some of this as the planning made on the initial phase of the project suffers from a high degree of uncertainty. This becomes especially true as software development is often the realization of a new or novel product. In projects where requirements have not been finalized and can change, requirements management is used to develop an accurate and complete definition of the behavior of software that can serve as the basis for software development.[20] While the terms may differ from industry to industry, the actual stages typically follow common steps to problem solving—"defining the problem, weighing options, choosing a path, implementation and evaluation."

PRINCE2



The PRINCE2 process model

PRINCE2 is a structured approach to project management, released in 1996 as a generic project management method.[21] It combined the original PROMPT methodology (which evolved into the PRINCE methodology) with IBM's MITP (managing the implementation of the total project) methodology. PRINCE2 provides a method for managing projects within a clearly defined framework. PRINCE2 describes procedures to coordinate people and activities in a project, how to design and supervise the project, and what to do if the project has to be adjusted if it does not develop as planned.

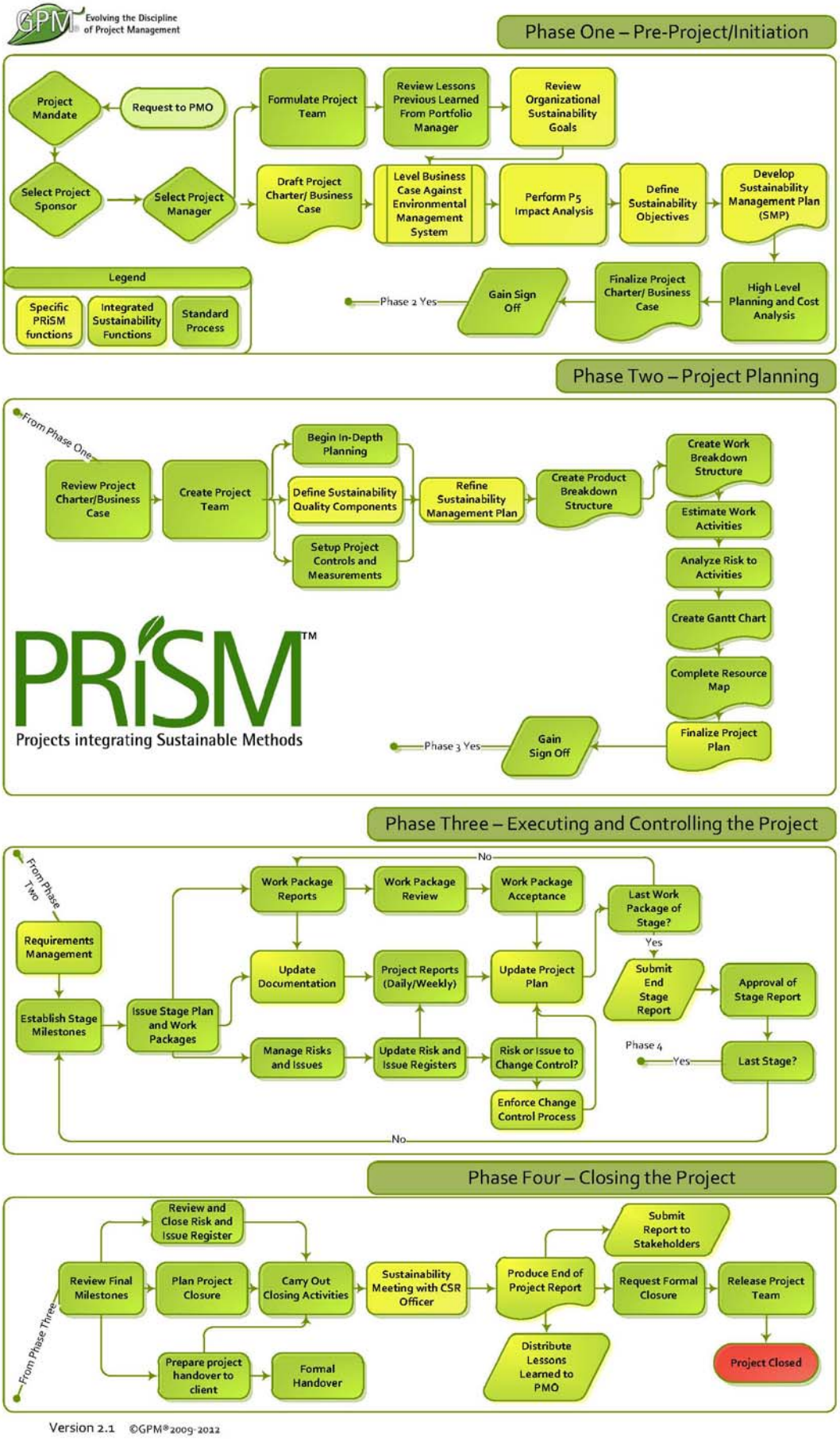
In the method, each process is specified with its key inputs and outputs and with specific goals and activities to be carried out. This allows for automatic control of any deviations from the plan. Divided into manageable stages, the method enables an efficient control of resources. On the basis of close monitoring, the project can be carried out in a controlled and organized way.

PRINCE2 provides a common language for all participants in the project. The various management

roles and responsibilities involved in a project are fully described and are adaptable to suit the complexity of the project and skills of the organization.

PRiSM (Projects integrating Sustainable Methods)

PRiSMis a process-based, structured project management methodology that introduces areas of sustainability and integrates them into four core project phases in order to maximize opportunities to improve sustainability and the use of finite resources.



The PRiSM Flowchart

The methodology encompasses the management, control and organization of a project with consideration and emphasis beyond the project life-cycle and on the five aspects of sustainability, People, Planet, Profit, Process and Product. It derives the framework from ISO:21500 as well as ISO 14001, ISO 26000, and ISO 9001 PRiSM is also used to refer to the training and accreditation of authorized practitioners of the methodology who must undertake accredited qualifications based on competency to obtain the GPM certification.

Critical chain project management

Main article: Critical chain project management

Critical chain project management (CCPM) is a method of planning and managing project execution designed to deal with uncertainties inherent in managing projects, while taking into consideration limited availability of resources (physical, human skills, as well as management & support capacity) needed to execute projects.

CCPM is an application of the theory of constraints (TOC) to projects. The goal is to increase the flow of projects in an organization (throughput). Applying the first three of the five focusing steps of TOC, the system constraint for all projects is identified as are the resources. To exploit the constraint, tasks on the critical chain are given priority over all other activities. Finally, projects are planned and managed to ensure that the resources are ready when the critical chain tasks must start, subordinating all other resources to the critical chain.

The project plan should typically undergo resource leveling, and the longest sequence of resource-constrained tasks should be identified as the critical chain. In some cases, such as managing contracted sub-projects, it is advisable to use a simplified approach without resource leveling.

In multi-project environments, resource leveling should be performed across projects. However, it is often enough to identify (or simply select) a single "drum". The drum can be a resource that acts as a constraint across projects, which are staggered based on the availability of that single resource. One can also use a "virtual drum" by selecting a task or group of tasks (typically integration points) and limiting the number of projects in execution at that stage.

Event chain methodology

Main article: Event chain methodology

Event chain methodology is another method that complements critical path method and critical chain project management methodologies.

Event chain methodology is an uncertainty modeling and schedule network analysis technique that is focused on identifying and managing events and event chains that affect project schedules. Event chain methodology helps to mitigate the negative impact of psychological heuristics and biases, as well as to allow for easy modeling of uncertainties in the project schedules. Event chain methodology is based on the following principles.

- Probabilistic moment of risk: An activity (task) in most real-life processes is not a continuous uniform process. Tasks are affected by external events, which can occur at some point in the middle of the task.
- Event chains: Events can cause other events, which will create event chains. These event

chains can significantly affect the course of the project. Quantitative analysis is used to determine a cumulative effect of these event chains on the project schedule.

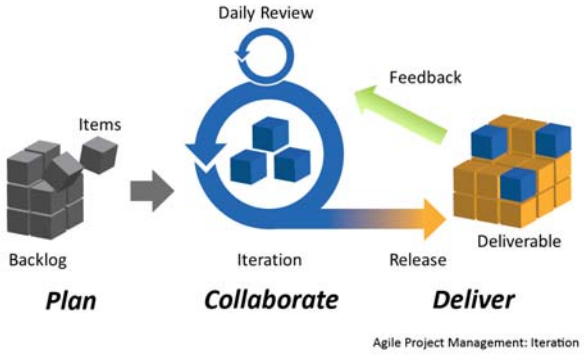
- Critical events or event chains: The single events or the event chains that have the most potential to affect the projects are the “critical events” or “critical chains of events.” They can be determined by the analysis.
- Project tracking with events: Even if a project is partially completed and data about the project duration, cost, and events occurred is available, it is still possible to refine information about future potential events and helps to forecast future project performance.
- Event chain visualization: Events and event chains can be visualized using event chain diagrams on a Gantt chart.

Process-based management

Main article: Process-based management

Also furthering the concept of project control is the incorporation of process-based management. This area has been driven by the use of Maturity models such as the CMMI (capability maturity model integration; see this example of a predecessor) and ISO/IEC15504 (SPICE – software process improvement and capability estimation).

Agile project management



The iteration cycle in agile project management

Agile project management approaches based on the principles of human interaction management are founded on a process view of human collaboration. It is "most typically used in software, website, technology, creative and marketing industries."This contrasts sharply with the traditional approach. In the agile software development or flexible product development approach, the project is seen as a series of relatively small tasks conceived and executed as the situation demands in an adaptive manner, rather than as a completely pre-planned process. It is the most consistent project management technique since it involves frequent testing of the project under development. It is the only technique in which the client will be actively involved in the project development. But the only disadvantage with this technique is that it should be used only if the client has enough time to be actively involved in the project every now and then.

Lean project management

Lean project management uses principles from lean manufacturing to focus on delivering value with less waste and reduced time

Extreme project management

Planning and feedback loops in Extreme programming (XP) with the time frames of the multiple loops.

In critical studies of project management it has been noted that several PERT based models are not well suited for the multi-project company environment of today.Most of them are aimed at very large-scale, one-time, non-routine projects, and currently all kinds of management are expressed in terms of projects.

Using complex models for "projects" (or rather "tasks") spanning a few weeks has been proven to cause unnecessary costs and low maneuverability in several cases Instead, project management experts try to identify different "lightweight" models, such as Extreme Programming and Scrum.

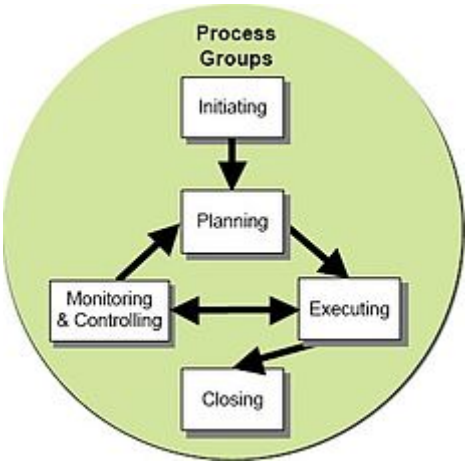
The generalization of Extreme Programming to other kinds of projects is extreme project management, which may be used in combination with the process modeling and management principles of human interaction management.

Benefits realisation management

Benefits realization management (BRM) enhances normal project management techniques through a focus on outcomes (the benefits) of a project rather than products or outputs, and then measuring the degree to which that is happening to keep a project on track. This can help to reduce the risk of a completed project being a failure by delivering agreed upon requirements/outputs but failing to deliver the benefits of those requirements.

An example of delivering a project to requirements might be agreeing to deliver a computer system that will process staff data and manage payroll, holiday and staff personnel records. Under BRM the agreement might be to achieve a specified reduction in staff hours required to process and maintain staff data.

Processes



The project development stages

Traditionally, project management includes a number of elements: four to five process groups, and a control system. Regardless of the methodology or terminology used, the same basic project management processes will be used. Major process groups generally include:

- Initiation
- Planning or design
- Production or execution
- Monitoring and controlling
- Closing

In project environments with a significant exploratory element (e.g., research and development), these stages may be supplemented with decision points (go/no go decisions) at which the project's continuation is debated and decided. An example is the Phase–gate model.

Initiating

Initiating process group processes

The initiating processes determine the nature and scope of the project, If this stage is not performed well, it is unlikely that the project will be successful in meeting the business' needs. The key project controls needed here are an understanding of the business environment and making sure that all necessary controls are incorporated into the project. Any deficiencies should be reported and a recommendation should be made to fix them.

The initiating stage should include a plan that encompasses the following areas:

- analyzing the business needs/requirements in measurable goals
- reviewing of the current operations
- financial analysis of the costs and benefits including a budget
- stakeholder analysis, including users, and support personnel for the project
- project charter including costs, tasks, deliverables, and schedule

Planning and design

After the initiation stage, the project is planned to an appropriate level of detail (see example of a flow-chart).The main purpose is to plan time, cost and resources adequately to estimate the work needed and to effectively manage risk during project execution. As with the Initiation process group, a failure to adequately plan greatly reduces the project's chances of successfully accomplishing its goals.

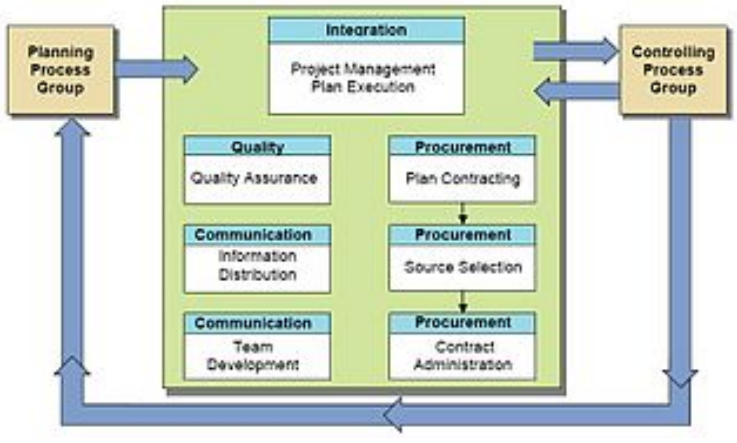
Project planning generally consists of:

- determining how to plan (e.g. by level of detail or rolling wave);
- developing the scope statement;
- selecting the planning team;
- identifying deliverables and creating the work breakdown structure;
- identifying the activities needed to complete those deliverables and network the activities in their logical sequence;
- estimating the resource requirements for the activities;
- estimating time and cost for activities;
- developing the schedule;
- developing the budget;
- risk planning;
- gaining formal approval to begin work.

Additional processes, such as planning for communications and for scope management, identifying roles and responsibilities, determining what to purchase for the project and holding a kick-off meeting are also generally advisable.

For new product development projects, conceptual design of the operation of the final product may be performed concurrent with the project planning activities, and may help to inform the planning team when identifying deliverables and planning activities.

Executing



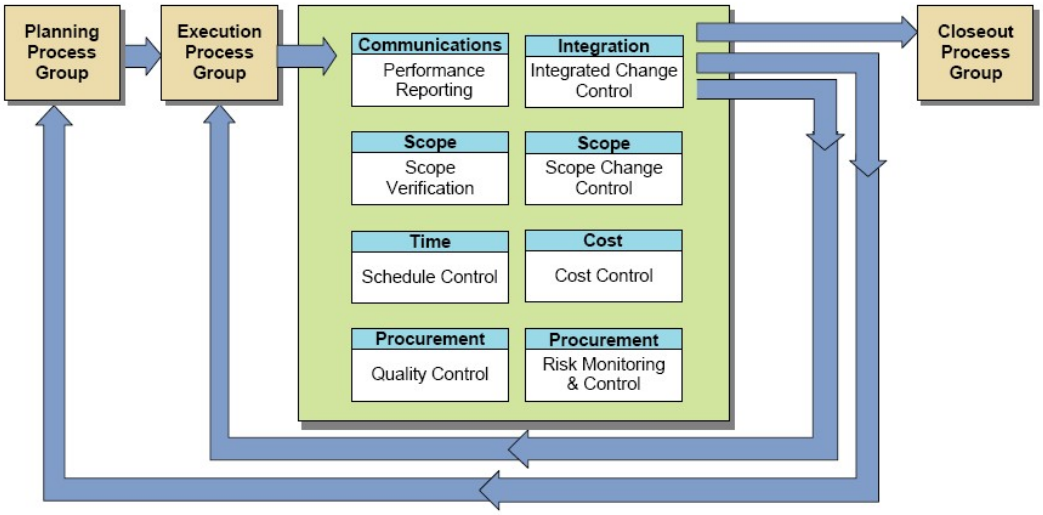
Executing process group processes

Executing consists of the processes used to complete the work defined in the project plan to accomplish the project's requirements. Execution process involves coordinating people and

resources, as well as integrating and performing the activities of the project in accordance with the project management plan. The deliverables are produced as outputs from the processes performed as defined in the project management plan and other frameworks that might be applicable to the type of project at hand.
Execution process group include:

- Direct and manage project execution
- Quality assurance of deliverables
- Acquire, develop and manage Project team
- Distribute information
- Manage stakeholder expectations
- Conduct procurement

Monitoring and controlling



Monitoring and controlling process group processes

Monitoring and controlling consists of those processes performed to observe project execution so that potential problems can be identified in a timely manner and corrective action can be taken, when necessary, to control the execution of the project. The key benefit is that project performance is observed and measured regularly to identify variances from the project management plan.
Monitoring and controlling includes:

- Measuring the ongoing project activities ('where we are');
- Monitoring the project variables (cost, effort, scope, etc.) against the project management plan and the project performance baseline (*where we should be*);
- Identify corrective actions to address issues and risks properly (*How can we get on track again*);
- Influencing the factors that could circumvent integrated change control so only approved changes are implemented.

In multi-phase projects, the monitoring and control process also provides feedback between project phases, in order to implement corrective or preventive actions to bring the project into compliance with the project management plan.
Project maintenance is an ongoing process, and it includes:

- Continuing support of end-users
- Correction of errors
- Updates of the software over time

Monitoring and controlling cycle

In this stage, auditors should pay attention to how effectively and quickly user problems are resolved.
Over the course of any construction project, the work scope may change. Change is a normal and expected part of the construction process. Changes can be the result of necessary design modifications, differing site conditions, material availability, contractor-requested changes, value engineering and impacts from third parties, to name a few. Beyond executing the change in the field, the change normally needs to be documented to show what was actually constructed. This is referred to as change management. Hence, the owner usually requires a final record to show all changes or, more specifically, any change that modifies the tangible portions of the finished work. The record is made on the contract documents – usually, but not necessarily limited to, the design drawings. The end product of this effort is what the industry terms as-built drawings, or more simply, “as built.” The requirement for providing them is a norm in construction contracts.
When changes are introduced to the project, the viability of the project has to be re-assessed. It is important not to lose sight of the initial goals and targets of the projects. When the changes accumulate, the forecasted result may not justify the original proposed investment in the project.
Closing

Closing process group processes.

Closing includes the formal acceptance of the project and the ending thereof. Administrative activities include the archiving of the files and documenting lessons learned.
This phase consists of.

- **Contract closure**: Complete and settle each contract (including the resolution of any open items) and close each contract applicable to the project or project phase.
- **Project close**: Finalize all activities across all of the process groups to formally close the project or a project phase

Project controlling and project control systems

Project controlling should be established as an independent function in project management. It implements verification and controlling function during the processing of a project in order to reinforce the defined performance and formal goals. The tasks of project controlling are also:

- the creation of infrastructure for the supply of the right information and its update
- the establishment of a way to communicate disparities of project parameters
- the development of project information technology based on an intranet or the determination of a project key performance index system (KPI)
- divergence analyses and generation of proposals for potential project regulations
- the establishment of methods to accomplish an appropriate the project structure, project workflow organization, project control and governance
- creation of transparency among the project parameters

Fulfillment and implementation of these tasks can be achieved by applying specific methods and instruments of project controlling. The following methods of project controlling can be applied:

- investment analysis
- cost–benefit analyses
- value benefit Analysis
- expert surveys
- simulation calculations
- risk-profile analyses
- surcharge calculations
- milestone trend analysis
- cost trend analysis
- target/actual-comparison

Project control is that element of a project that keeps it on-track, on-time and within budget.[28] Project control begins early in the project with planning and ends late in the project with post-implementation review, having a thorough involvement of each step in the process. Each project should be assessed for the appropriate level of control needed: too much control is too time consuming, too little control is very risky. If project control is not implemented correctly, the cost to the business should be clarified in terms of errors, fixes, and additional audit fees.
Control systems are needed for cost, risk, quality, communication, time, change, procurement, and human resources. In addition, auditors should consider how important the projects are to the

financial statements, how reliant the stakeholders are on controls, and how many controls exist. Auditors should review the development process and procedures for how they are implemented. The process of development and the quality of the final product may also be assessed if needed or requested. A business may want the auditing firm to be involved throughout the process to catch problems earlier on so that they can be fixed more easily. An auditor can serve as a controls consultant as part of the development team or as an independent auditor as part of an audit. Businesses sometimes use formal systems development processes. These help assure that systems are developed successfully. A formal process is more effective in creating strong controls, and auditors should review this process to confirm that it is well designed and is followed in practice. A good formal systems development plan outlines:

- A strategy to align development with the organization's broader objectives
- Standards for new systems
- Project management policies for timing and budgeting
- Procedures describing the process
- Evaluation of quality of change

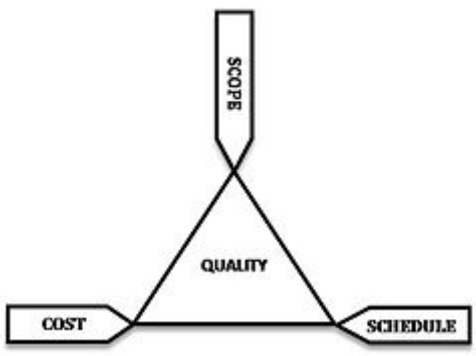
Project managers

A project manager is a professional in the field of project management. Project managers can have the responsibility of the planning, execution, and closing of any project, typically relating to construction industry, engineering, architecture, computing, and telecommunications. Many other fields in production engineering and design engineering and heavy industrial have project managers.

A project manager is the person accountable for accomplishing the stated project objectives. Key project management responsibilities include creating clear and attainable project objectives, building the project requirements, and managing the triple constraint for projects, which is cost, time, and scope.

A project manager is often a client representative and has to determine and implement the exact needs of the client, based on knowledge of the firm they are representing. The ability to adapt to the various internal procedures of the contracting party, and to form close links with the nominated representatives, is essential in ensuring that the key issues of cost, time, quality and above all, client satisfaction, can be realized.

Project management triangle



The project management triangle

Like any human undertaking, projects need to be performed and delivered under certain constraints. Traditionally, these constraints have been listed as "scope," "time," and "cost". These are also referred to as the "project management triangle", where each side represents a constraint. One side of the triangle cannot be changed without affecting the others. A further refinement of the constraints separates product "quality" or "performance" from scope, and turns quality into a fourth constraint.

The time constraint refers to the amount of time available to complete a project. The cost constraint refers to the budgeted amount available for the project. The scope constraint refers to what must be done to produce the project's end result. These three constraints are often competing constraints: increased scope typically means increased time and increased cost, a tight time constraint could mean increased costs and reduced scope, and a tight budget could mean increased time and reduced scope.

The discipline of project management is about providing the tools and techniques that enable the project team (not just the project manager) to organize their work to meet these constraints.

Work breakdown structure

The work breakdown structure (WBS) is a tree structure that shows a subdivision of effort required to achieve an objective—for example a program, project, and contract. The WBS may be hardware-, product-, service-, or process-oriented (see an example in a NASA reporting structure (2001)).

A WBS can be developed by starting with the end objective and successively subdividing it into manageable components in terms of size, duration, and responsibility (e.g., systems, subsystems, components, tasks, sub-tasks, and work packages), which include all steps necessary to achieve the objective.

The work breakdown structure provides a common framework for the natural development of the overall planning and control of a contract and is the basis for dividing work into definable increments from which the statement of work can be developed and technical, schedule, cost, and labor hour reporting can be established.

Project management framework

The program (investment) life cycle integrates the project management and system development life cycles with the activities directly associated with system deployment and operation. By design, system operation management and related activities occur after the project is complete and are not documented within this guide

For example, see figure, in the US United States Department of Veterans Affairs (VA) the program management life cycle is depicted and describe in the overall VA IT Project Management Framework to address the integration of OMB Exhibit 300 project (investment) management activities and the overall project budgeting process. The VA IT Project Management Framework diagram illustrates Milestone 4 which occurs following the deployment of a system and the closing of the project. The project closing phase activities at the VA continues through system deployment and into system operation for the purpose of illustrating and describing the system activities the VA considers part of the project. The figure illustrates the actions and associated artifacts of the VA IT

Project and Program Management process.

International standards

There have been several attempts to develop project management standards, such as:

- Capability Maturity Model from the Software Engineering Institute.
- GAPPS, Global Alliance for Project Performance Standards – an open source standard describing COMPETENCIES for project and program managers.
- A Guide to the Project Management Body of Knowledge from the Project Management Institute (PMI)
- HERMES method, Swiss general project management method, selected for use in Luxembourg and international organizations.
- The ISO standards ISO 9000, a family of standards for quality management systems, and the ISO 10006:2003, for Quality management systems and guidelines for quality management in projects.
- PRINCE2, PRojects IN Controlled Environments.
- Association for Project Management Body of Knowledge[34]
- Team Software Process (TSP) from the Software Engineering Institute.
- Total Cost Management Framework, AACE International's Methodology for Integrated Portfolio, Program and Project Management.
- V-Model, an original systems development method.
- The Logical framework approach, which is popular in international development organizations.
- IAPPM, The International Association of Project & Program Management, guide to project auditing and rescuing troubled projects.

Project portfolio management

An increasing number of organizations are using, what is referred to as, project portfolio management (PPM) as a means of selecting the right projects and then using project management techniques as the means for delivering the outcomes in the form of benefits to the performing private or not-for-profit organization.

Project management software

Project management software has a capacity to help plan, organize, and manage resource pools and develop resource estimates. Depending the sophistication of the software, resource including estimation and planning, scheduling, cost control and budget management, resource allocation, collaboration software, communication, decision-making, quality management and documentation or administration systems. Today, numerous PC-based project management software packages exist, and they are finding their way into almost every type of business. Software may range from the high-end Microsoft Project to a simple spreadsheet in Microsoft Excel.

附件五

英文縮寫對照表(Glossary)：

AFP：Apple Filing Protocol，安德魯檔案系統
AP：Access Point，無線接入點(無線網路基地台)
ARP：Address Resolution Protocol，位址解析協定
ASAP：Aggregate Server Access Protocol，聚集伺服器存取協定
ASN.1：Abstract Syntax Notation One，抽象語法標示
ATM：Asynchronous Transfer Mode，非同步傳輸模式
ATP：Agent Transfer Protocol，代理傳輸協定
BGP：Border Gateway Protocol，邊界閘道協定
BSD sockets：Berkeley socket API，柏克來插槽應用程式
CKIP：Cisco key integrity Protocol，思科關鍵整合協定
EIGRP：Enhanced Interior Gateway Routing Protocol，內部增進型閘道路由協定
ENRP：Endpoint Handlespace Redundancy Protocol，端點處理空間容錯協定
FDDI：Fiber Distributed Data Interface.，光纖分散式資料介面
FTP：File Transfer Protocol，檔案傳輸協定
HDLC：High-Level Data Link Control，高階資料連結協控制
HTTP：HyperText Transport Protocol，超文件傳輸協定
ICMP：Internet Control Message Protocol，網路控制訊息協定
IEEE：Institute of Electrical and Electronics Engineers，電機電子工程師學會
IGMP：Internet Group Management Protocol，網路群組管理協定
IGRP：Interior Gateway Routing Protocol，內部閘道路由協定
IP：Internet Protocol，網路協定
IPSec：Internet Protocol Security，網路協定安全機制
IPX：Internet Packet Exchange，網路封包交換
ISDN：Integrated Services Digital Network，整合服務數位網路
MAC：Medium Access Control，媒體存取控制
NCP：Network Control Program，網路控制程式
NetBIOS (Network Basic Input/Output System)，網路基本輸出入系統
NFS：Network File System，網路檔案系統
OSI：Open Systems Interconnection，開放式互聯系統
OSPF：Open Shortest Path First，開放式第一最短路徑
PPP：Point-to-Point Protocol，點對點協定
RADIUS：Remote Authentication Dial-In User Service，撥入用戶遠程認證服務
RARP：Reverse Address Resolution Protocol，逆向位址解析協定
RIP：Routing Information Protocol，路由資訊協定
RPC：Remote Procedure Call，遠端程序呼叫
RTP：Real-time Transport Protocol，即時傳輸協定
RTSP：Real Time Streaming Protocol，即時串流協定

SCTP：Stream Control Transmission Protocol，串流控制傳輸協定
SMB：Server Message Block，伺服器訊息區塊
SMTP：Simple Mail Transfer Protocol，簡易郵件傳輸協定
SSH：Secure Shell，安全殼協定
TCP：Transmission Control Protocol，傳輸控制協定
Telnet：A user command and TCP/IP protocol used for accessing remote PCs. 遠端登入服務協定
TKIP：Temporal Key Integrity Protocol，臨時鑰匙完整性協定
TLS：Transport Layer Security，傳輸層安全機制
UDP：User Datagram Protocol，使用者資料協定
VPN：Virtual Private Network，虛擬私有網路
WEP：Wired Equivalent Privacy，有線等效加密
WLAN：Wireless LAN，無線區域網路
WPA：Wi-Fi Protected Access, wi-fi 保護存取機制
XDR：External Data Representation，外部數據表示法
XMPP：Extensible Messaging and Presence Protocol，可擴展通訊與表示協定