云计算在统计及Data Mining研究应用及前沿综述

CLOUD-R 雲計算下的R 前沿探討 - 兼談 數據分析的未來

谢邦昌

辅仁大学商学研究所博士班 所长

统计资讯学系教授

中华资料采矿协会 理事长

2011年9月28日

stat1001@mails.fju.edu.tw

WWW.CDMS.ORG.TW

马院士：*统计与数学要相互欣赏*

统计要跟各个领域
做朋友

统计趋势Statistics Trend 趋势统计 Trend Statistics

数据分析科学的过去、现在、未来

--统计是数据科学

Data Science

Dataology

Science

-- (Dealing with DATA) 11 FEBRUARY

2011 VOL 331 SCIENCE www.sciencemag.org

INTRODUCTION

Challenges and Opportunities

SCIENTIFIC INNOVATION HAS BEEN CALLED ON TO SPUR ECONOMIC recovery; science and technology are essential to improving public health and welfare and to inform sustainability; and the scientific community has been criticized for not being sufficiently accountable and transparent. Data collection, curation, and access are central to all of these issues. For this reason, *Science* has joined with colleagues from our sister publications *Science Signaling*, *Science Translational Medicine*, and *Science* Careers to provide a broad look at the issues surrounding the increasingly huge influx of research data. The entire collection is compiled online at www.sciencemag.org/special/data/. As you will discover, two themes appear repeatedly: Most scientific disciplines are finding the data deluge to be extremely challenging, and tremendous opportunities can be realized if we can better organize and access the data.

Our authors explore data issues that apply to specific fields as well as challenges shared between fields. These articles clearly show that the challenges are difficult and growing. We have recently passed the point where more data is being collected than we can physically store (see Hilbert *et al.*, published online). This storage gap will widen rapidly in data-intensive fields. Thus, decisions will be needed on which data to archive and which to discard. A separate problem is how to access and use these data. Many data sets are becoming too large to download. Even fields with well-established data archives, such as genomics, are facing new and growing challenges in data volume and management. And even where accessible, much data in many fields is too poorly organized to enable it to be efficiently used.

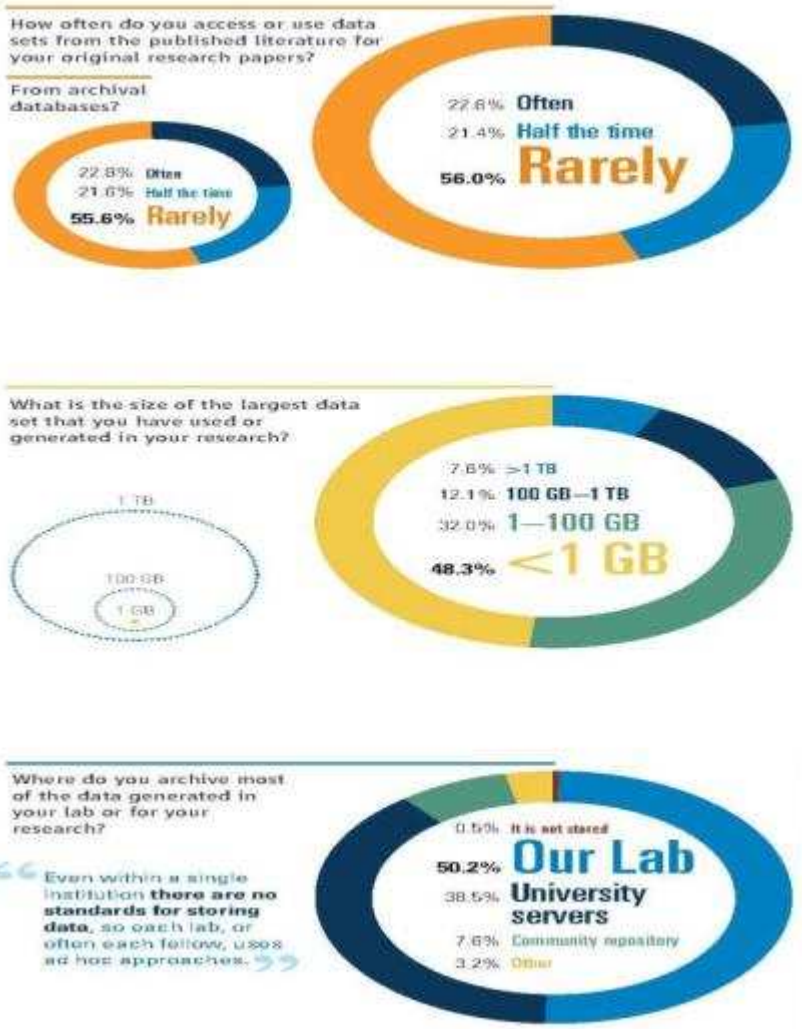
To delve deeper into these issues, *Science* polled our peer reviewers from last year about the availability and use of data. We received about 1700 responses, representing input from an international and interdisciplinary group of scientific leaders. About 20% of the respondents regularly use or analyze data sets exceeding 100 gigabytes, and 7% use data sets exceeding 1 terabyte. About half of those polled store their data only in their laboratories—not an ideal long-term solution. Many bemoaned the lack of common metadata and archives as a main impediment to using and storing data, and most of the respondents have no funding to support archiving.

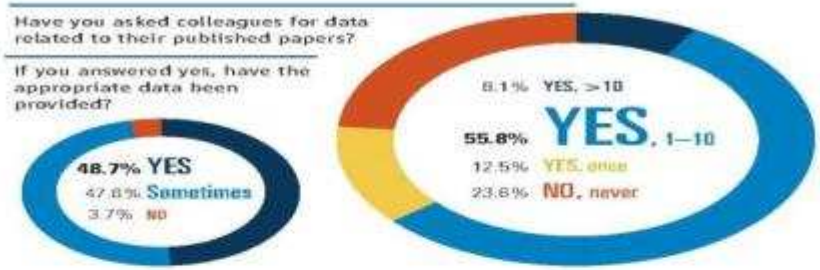
Many of the responders indicated that they seek or would like additional help in analyzing the data that they had collected. If we can use and reuse scientific data better, the opportunities, as indicated in many examples in this special section, are myriad. Large integrated data sets can potentially provide a much deeper understanding of both nature and society and open up many new avenues of research. And they are critical for addressing key societal problems—from improving public health and managing natural resources intelligently to designing better cities and coping with climate change.

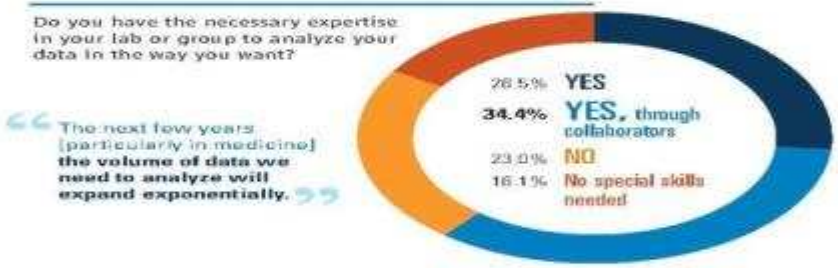
To realize these opportunities, many of the articles in this collection speak of changing the culture of science and the practices of scientists, as well as recognizing the growing responsibility for much better data stewardship. Several of the pieces illustrate steps toward these goals. But it is clear that organized effort and leadership are needed from funders, societies, journals, educators, and individual scientists—and from society at large.

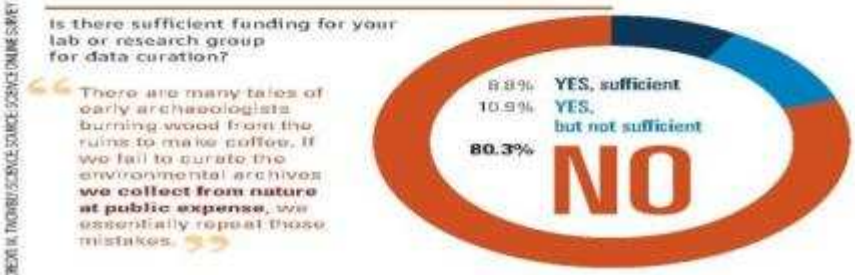
We hope that this collection spurs additional thinking and catalyzes new efforts in dealing with these critical issues. As a start, we invite you to share your thoughts at talk.sciencemag.org, where you can also contribute to our poll.

– *SCIENCE* STAFF









CREDIT: M. THOMAS/SCIENCE; SOURCE: SCIENCE'S COMPASS SURVEY

CONTENTS

*These items, plus a related podcast and online discussion, are available at www.sciencemag.org/special/data/

雲端運算的演化

Super Computer

Cluster Computing

Distributed Computing

Grid Computing

Utility Computing

Cloud Computing

雲端運算

- 透過網路將龐大的運算處理常式自動分拆成無數個較小的副程式，再交由多部伺服器所組成的龐大系統經搜尋、運算分析之後將處理結果回傳給用戶
- 雲～＝網路
- Google：MapReduce、GFS及BigTable

雲端運算產業類型

IIaaS I^2aaS

Information & intelligence as a Service

SaaS

Software as a Service

PaaS

Platform as a Service

IaaS

Infrastructure as a Service

Waves of Innovation

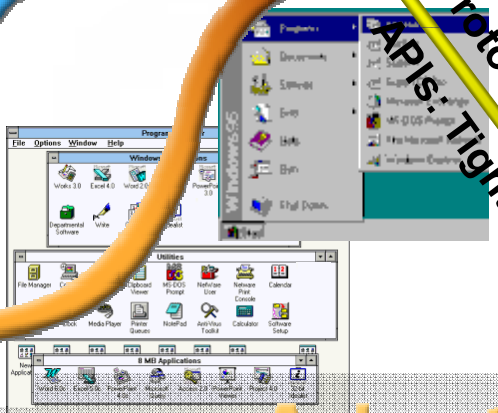
Development

Hard Engineering ➔ Intellectual Property

- PC Architecture
- DOS
- Spreadsheets
- Word Processors
- Mouse
- GUI
- LANs
- XML/SOAP
- HTTP/HTML
- SMTP
- Email Clients
- Web Browsers
- Speech/Writing
- Devices
- Wi-Fi/Broadband
- Web Services
- Trusted Computing Hardware
- Rights Management
- Cloud Computing

APIs: Tightly Coupled

Protocols: Loosely Coupled



Adoption

Intellectual Property ➔ Consumer Benefit

Today

PC	Applications	Internet	Web Apps
Mid 80s	Late 80s-Mid 90s	Mid 90s	Mid 00s - . .

SUBSCRIBER COPY NOT FOR RESALE FEBRUARY 21, 2011

Revolution in Egypt

Joe Klein: What the U.S. should do
On the Street: Hope meets anxiety
Muslim Brotherhood: What it wants

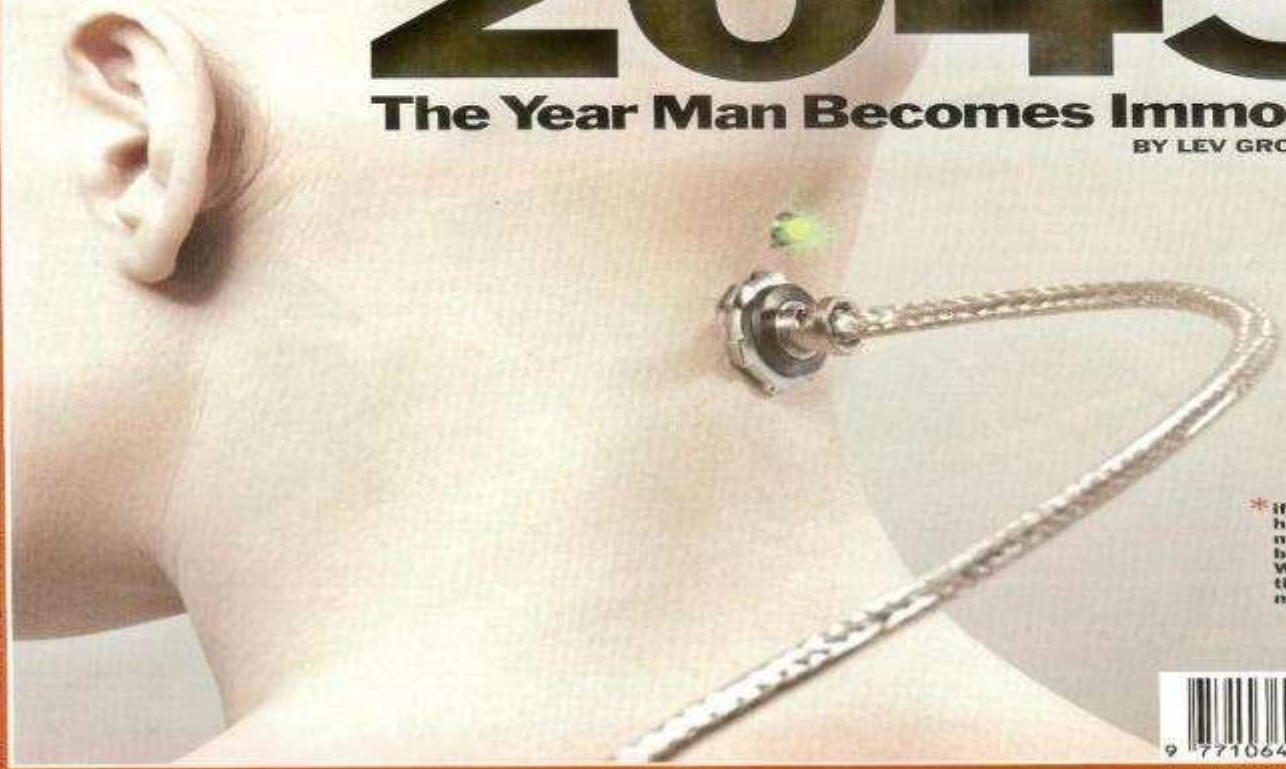
Oscars: Portraits of star power

TIME

2045

The Year Man Becomes Immortal*

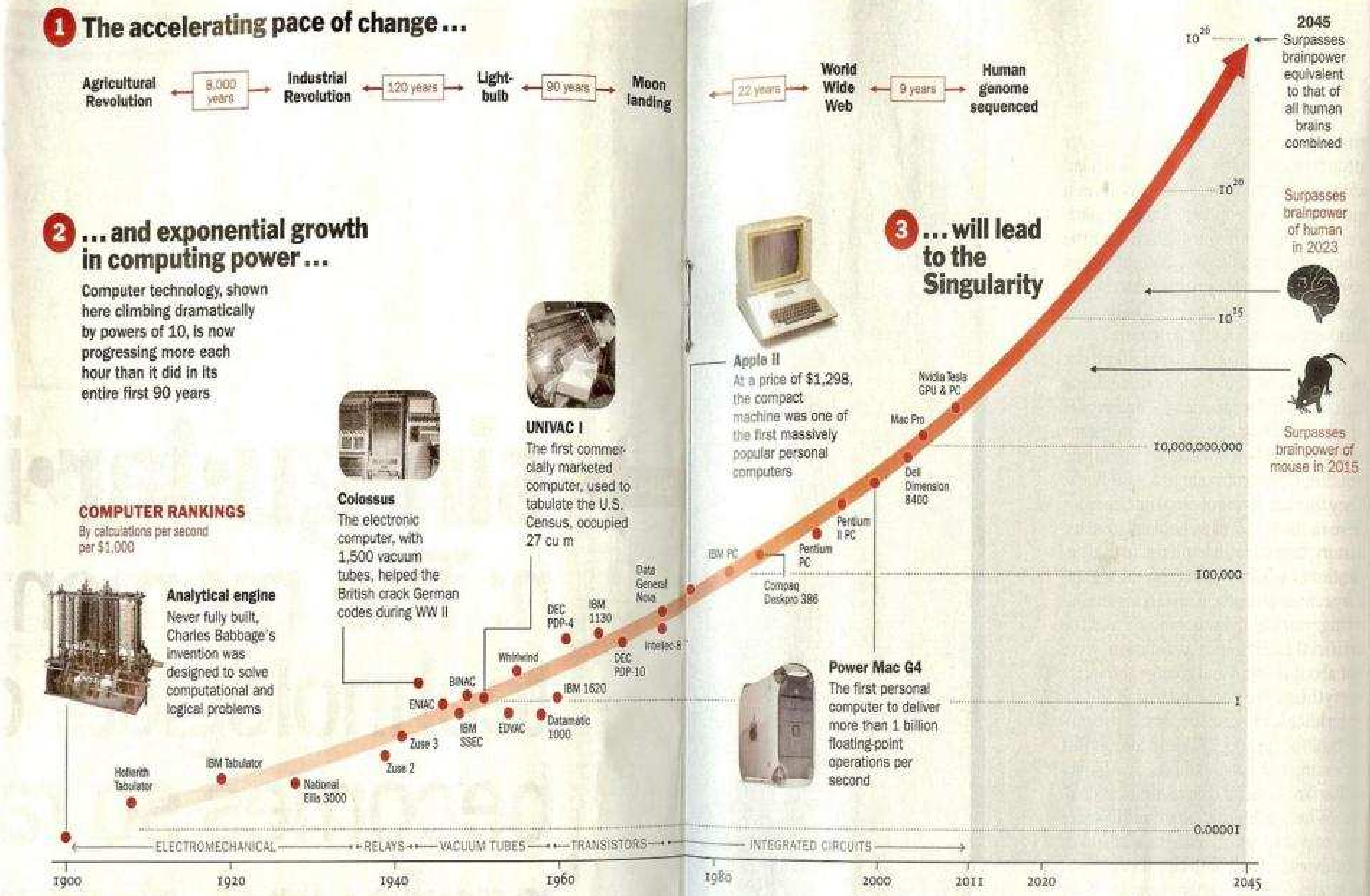
BY LEV GROSSMAN



*if you believe humans and machines will become one. Welcome to the Singularity movement

9 771064 030005 07

www.time.com



on, there's no reason to think computers

Probably. It's impossible to predict the

idea; it's a serious hypothesis about the

he called an "intelligence explosion":

「计算机vs.人脑」益智游戏大赛经过三天激战后，IBM的超级计算机华生（Watson，图中）最终击败人类，获颁100万美元奖金。IBM将把这笔奖金捐给世界展望会等慈善机构。（美联社）

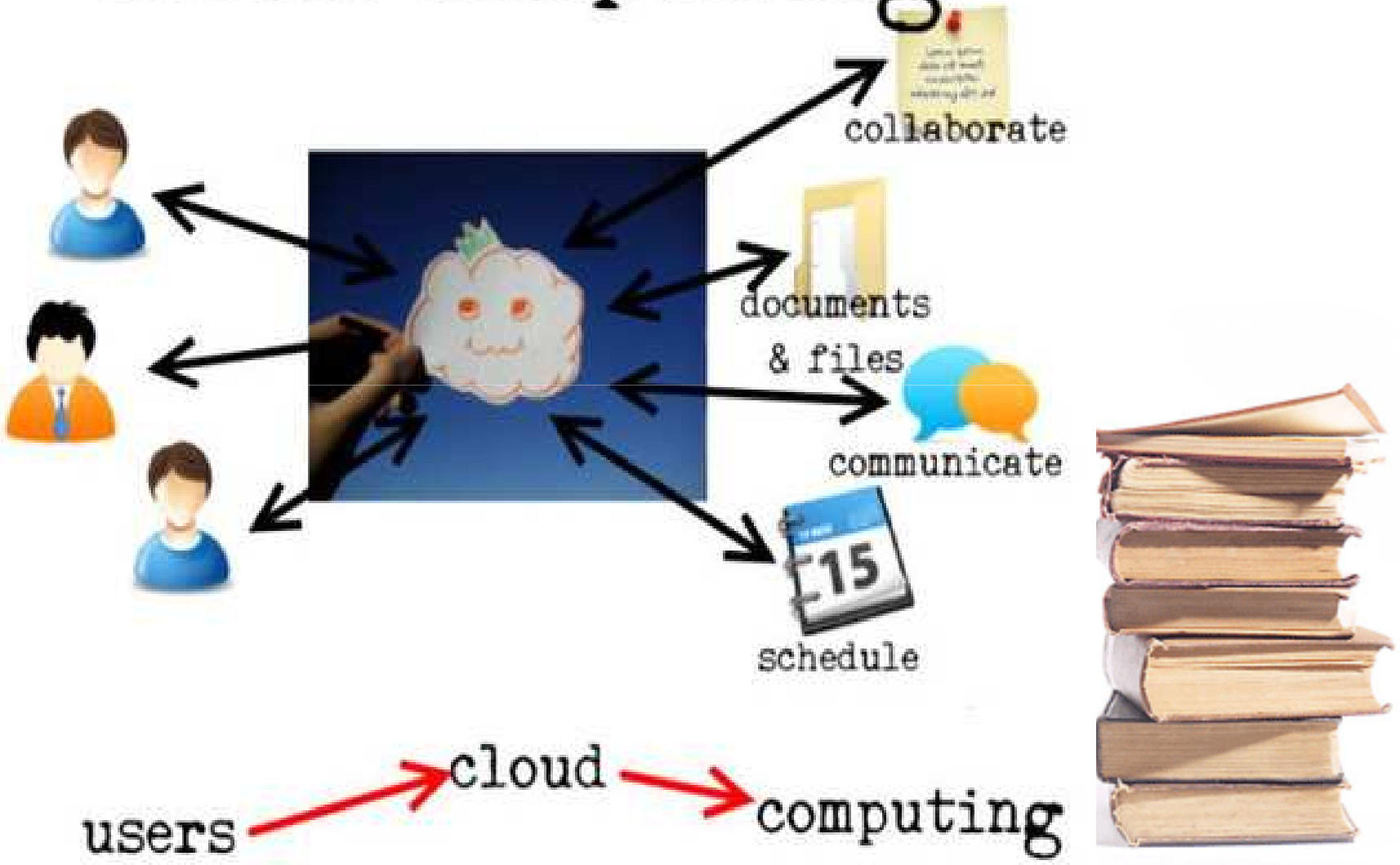
- IBM研究人员花了四年打造华生，它每秒运算能力达80兆次。由2,800个处理器核心、16兆字节工作内存运转。为了建立华生的知识库，研究人员在4兆字节磁盘累积了2亿页的内容，它能运用600万条逻辑规则来确定哪一个是最好的答案。当华生被问问题时，软件就会对名字、数据、地理位置或其他条件开始进行分析。它甚至能对问题的暗示进行语句结构或文法分析。
- 虽然IBM还没有计划要参加第二度决赛，或是创造第二代华生，但的确有计划要在许多领域推销华生计算机技术。像是在健康照护领域，能藉此回答许多关于人体知识的困难问题。
-

云端运算简介

说文解字

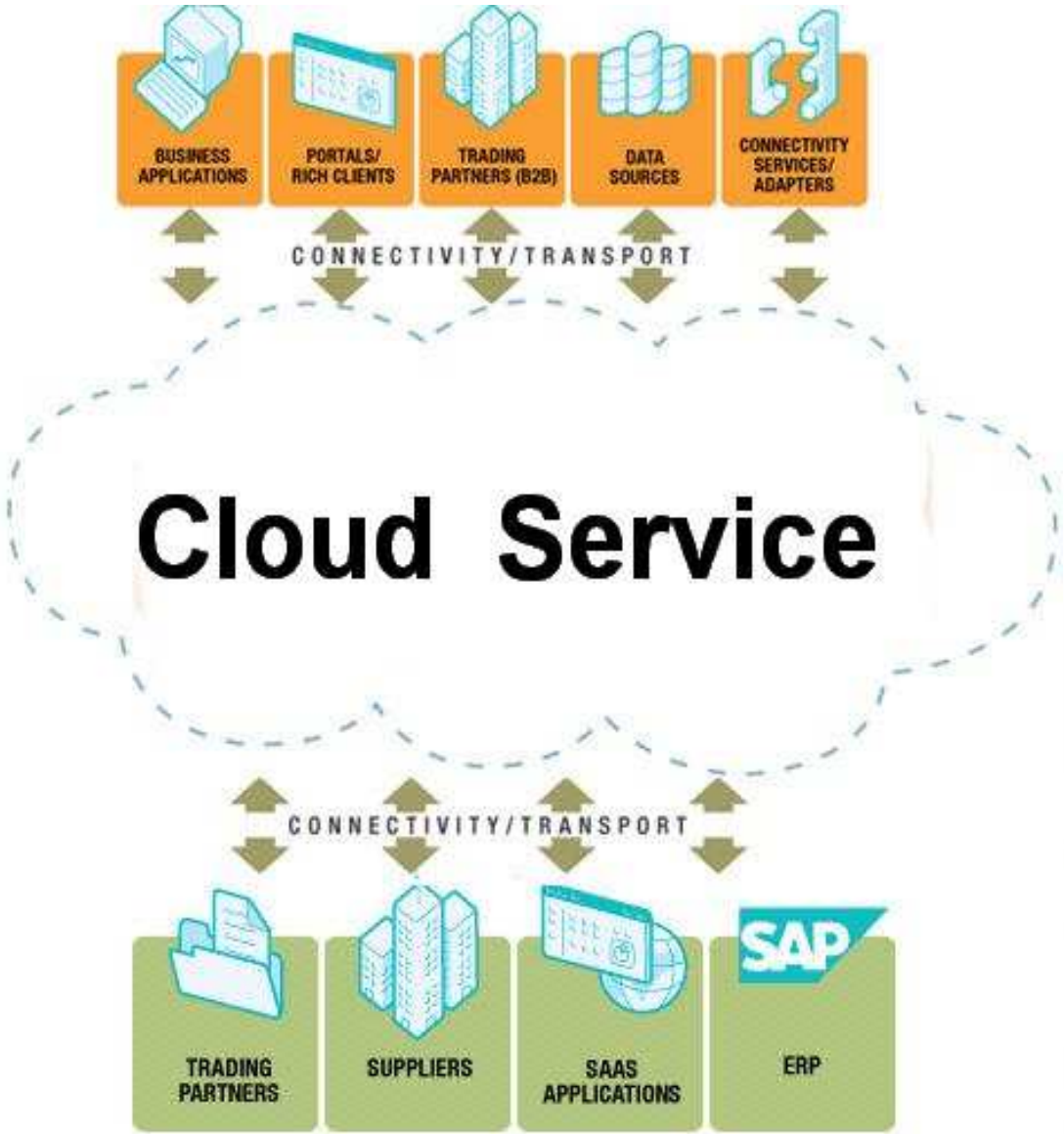
- ***云端运算***(Cloud Computing)：***将庞大运算操作拆成千百个较小的操作，再交给远程、多台服务器，同时运算***。透过此种技术，网络服务提供者可以在数秒之内，处理数以千万计的信息，并提供和「超级计算机」一样强大效能的网络服务，以符合网络用户日增的各种需求。
- Google搜寻服务，Gmail，YouTube，Google Docs，Google Talk，iGoogle，Google Calendar充份使用这种技术。其它如微软，YAHOO，AMAZON亦采用这种技术，提升网络服务功能。

cloud computing



云端服务

- Google搜索
- Web Email
- 在线扫毒
- YouTube
- 在线文件
- 部落格
- …



Super Computer

云端运算的演化

走鹃（Roadrunner）是一套由IBM为美国能源部（Department of Energy）所属的国家核能安全管理部（National Nuclear Security Administration，NNSA）建立的超级计算机

Jaguar—全球最快的Jaguar超级计算机，它采用的Cray XT5超级计算机由原先的Opteron四核心处理器升级为AMD六核心的Istanbul，因而取代了IBM的Roadrunner，成为当今全球最快的超级计算机，实际值的效能达到1.75 petaflop/s。

最快的超级计算机"天河一号"


“天河一号”千万亿次超级计算机系统研制成功
新 闻 发 布 会
国防科学技术大学


天河高效能计算机系统
CRI online 新闻 国际在线
©www.cfp.cn 版权作品 请勿转载

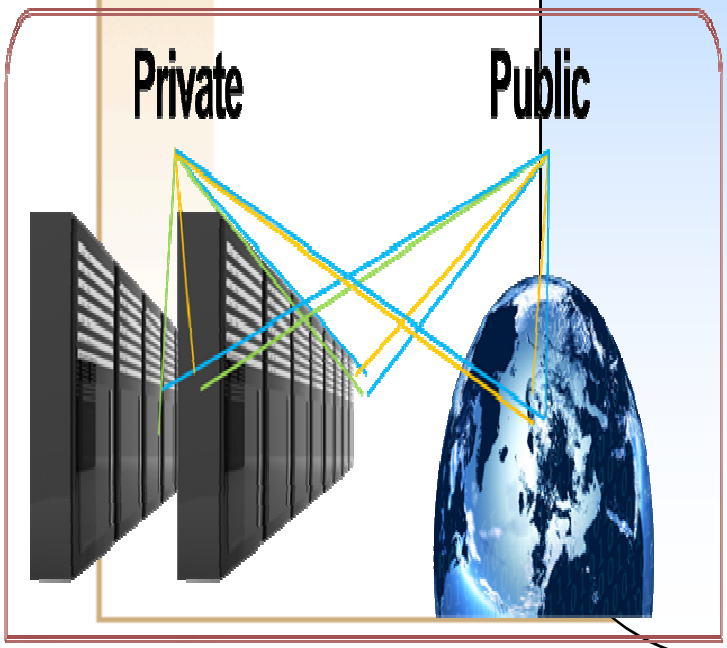
云端运算的演化

Super Computer

Cluster Computing

丛集运算

- 通过一组松散集成的计算器软件和/或硬件连接起来，紧密地协作完成计算工作
- PVM、MPI
- 1960~
- 相对于超级计算机有高的性价比



云端运算的演化

Super Computer

Cluster Computing

Distributed Computing

分布式计算

- 把需要进行大量计算的工程数据分割成小块，由多台计算器分别计算，在上传运算结果后，将结果统一合并得出数据结论的科学。
- 找到新药
- **E-MAIL STAT1001@gmail.com**

云端运算的演化

Super Computer

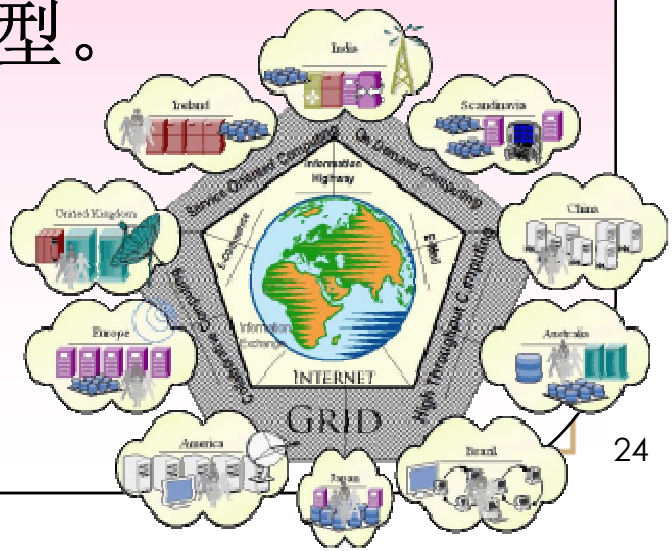
Cluster Computing

Distributed Computing

Grid Computing

格网运算

- 大量异构计算器（通常为桌面）的未用资源（CPU周期和磁盘存储），将其作为嵌入在分布式电信基础设施中的一个虚拟的计算器集群，为解决大规模的计算问题提供了一个模型。
- Globus
- 1990~

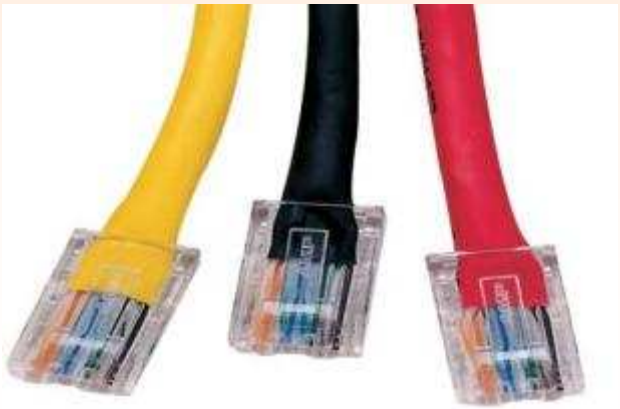


云端运算的演化

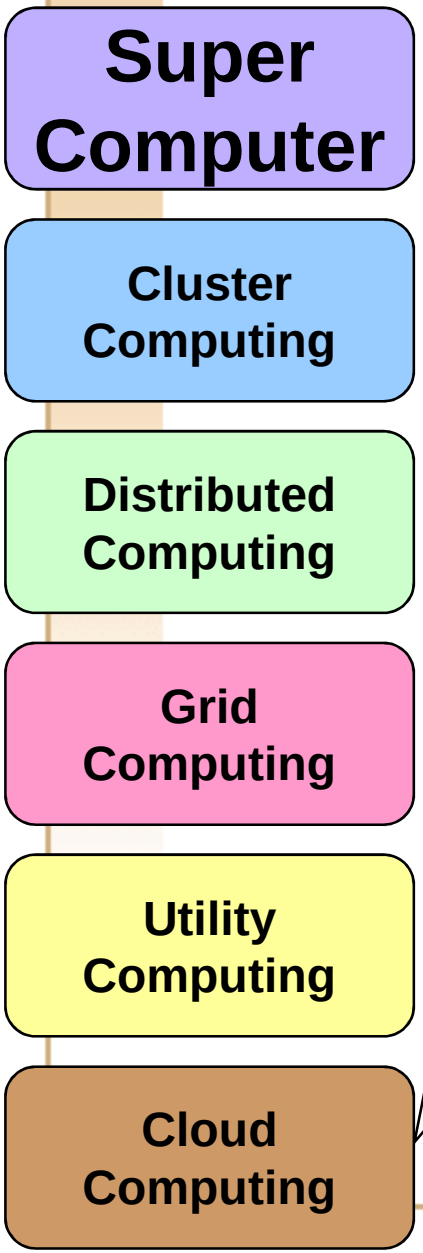
- Super Computer
- Cluster Computing
- Distributed Computing
- Grid Computing
- Utility Computing

公用运算

- 主要提倡一种理想的企业信息架构，让IT服务模仿公用服务的方式进行，如供应水、电力、瓦斯。”用多少付多少”以及”随需即用”
- From IBM,



云端运算的演化



云端运算

- 透过网络将庞大的运算处理程序自动分拆成无数个较小的子程序，再交由多部服务器所组成的庞大系统经搜寻、运算分析之后将处理结果回传给用户
- 云～＝网络
- Google：MapReduce、GFS及BigTable

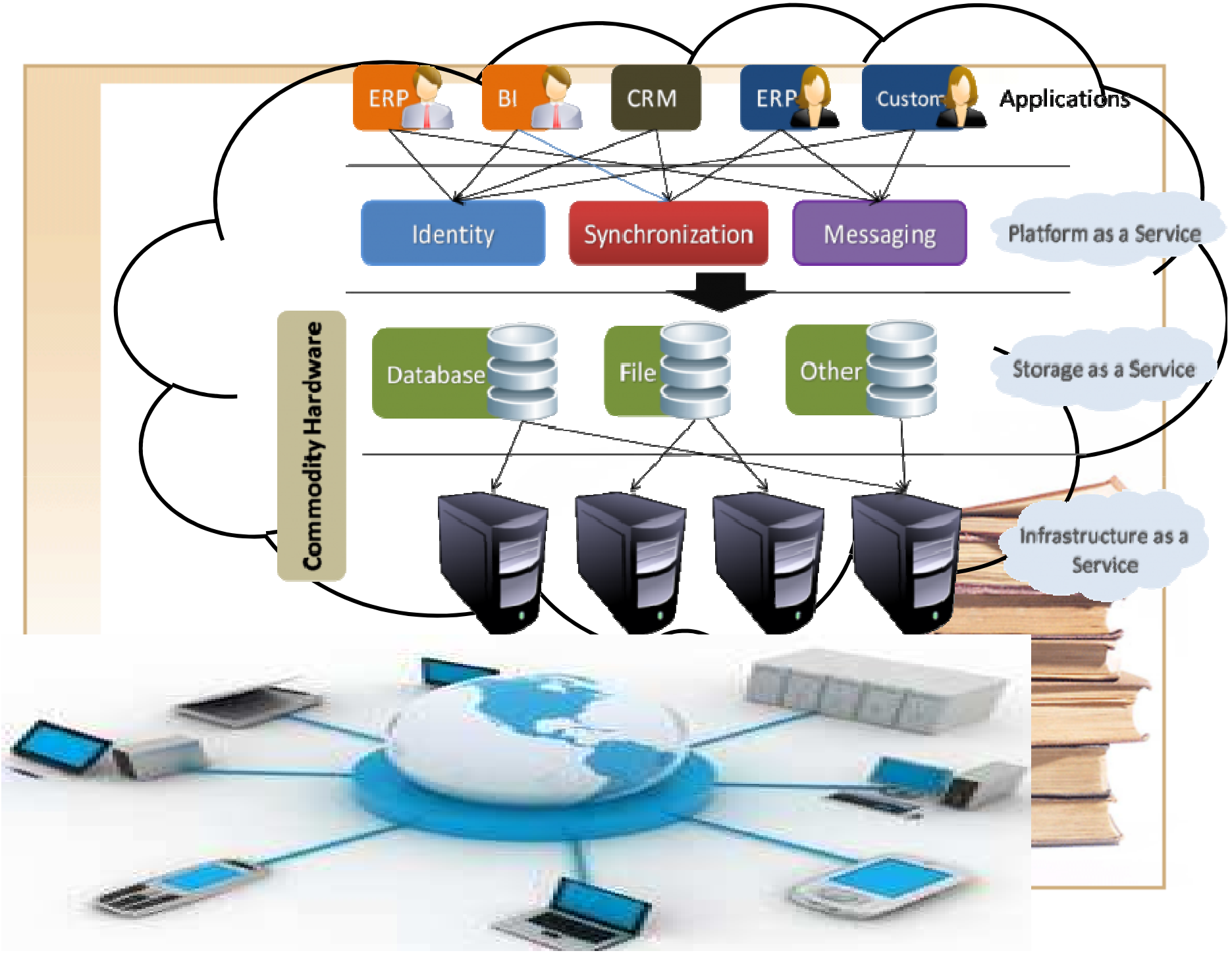
国家统计局

统计四大工程的主要内容

（一）统计四大工程的定义和主要内容

- “十二五”时期，要加快建设并初步建成基本单位名录库、企业一套表制度、数据采集处理软件系统和联网直报系统等互相联系、共为整体的四大工程。
- 我们要站在时代的高度，全面领会与把握推进四大工程建设的重要作用和深刻内涵。

这就是云计算的观念–快速运算&大量储存


ERP
BI
CRM
ERP
Custom
Applications
Identity
Synchronization
Messaging
Platform as a Service
Commodity Hardware
Database
File
Other
Storage as a Service
Infrastructure as a Service

DATA Center -- >货柜云?

1. 台湾PC硬件的第二春?
2. 小虾米V.S.大鲸鱼?
3. 软硬通吃?
4. 众志成城?

Data Center虚拟现实


535
530
525

Data Center Growth

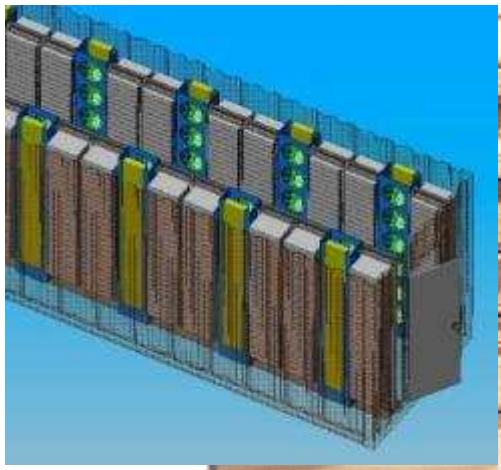
An Idea Whose Time Has Come

Nortel Steel Enclosure
Containerized telecom equipment

Sun Project Black Box
242 systems in 20’

Caterpillar
Portable Power

Rackable Systems
1,152 Systems in 40’



Rackable Systems Container
Cooling Model

Shipping Container as Data Center Module

Unit of Data Center Growth

Manufacturing & H/W Admin.

Systems & Power Density

云端运算特色

高可靠度

超大规模

虚拟化

高通用性

使用者付费

成本低

高扩充性

现有的云端运算服务

- Windows
- Google
- Amazon
- Yahoo
- ….

云端运算产业类型

SaaS
Software as a Service

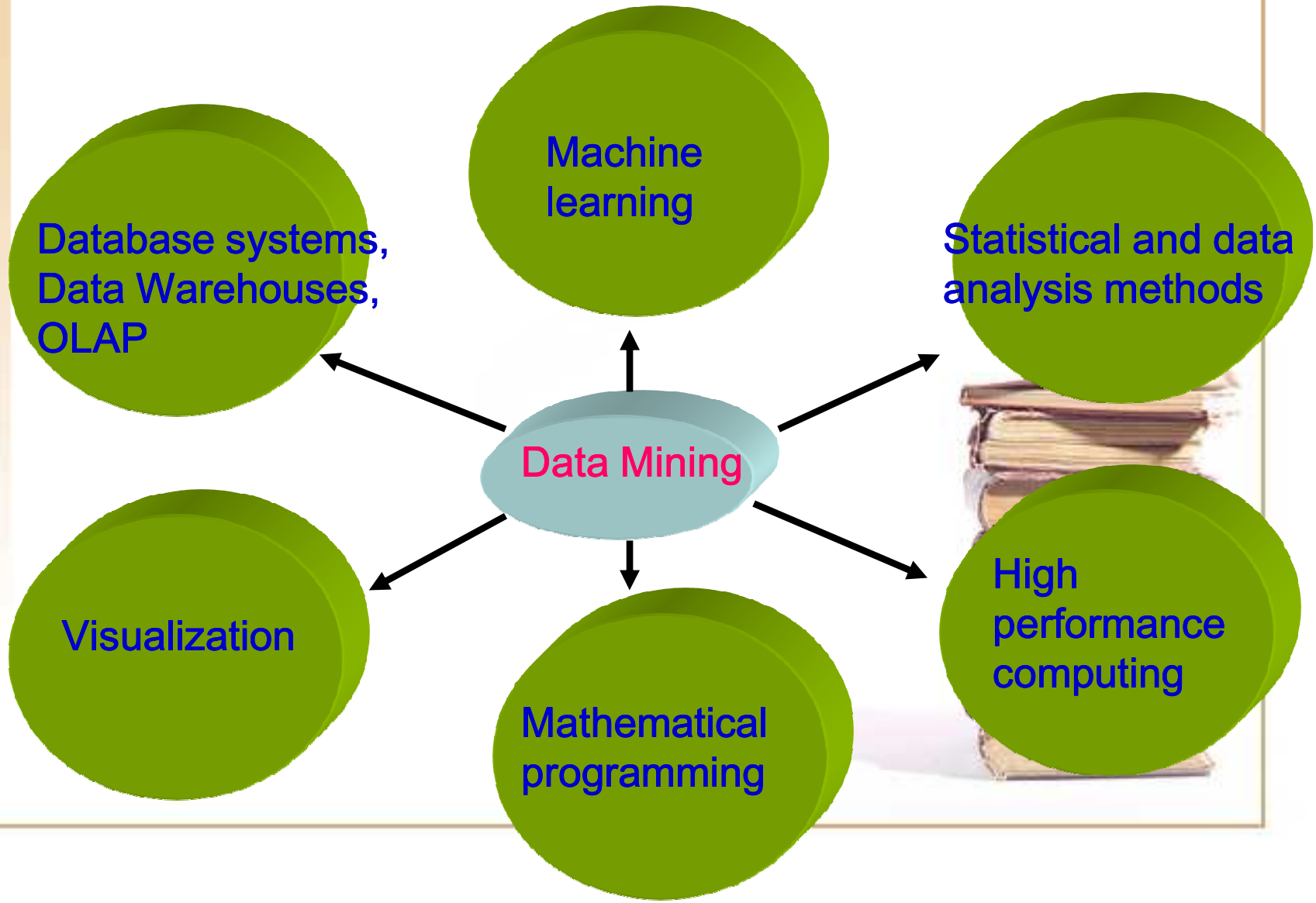
PaaS
Platform as a Service

IaaS
Infrastructure as a Service

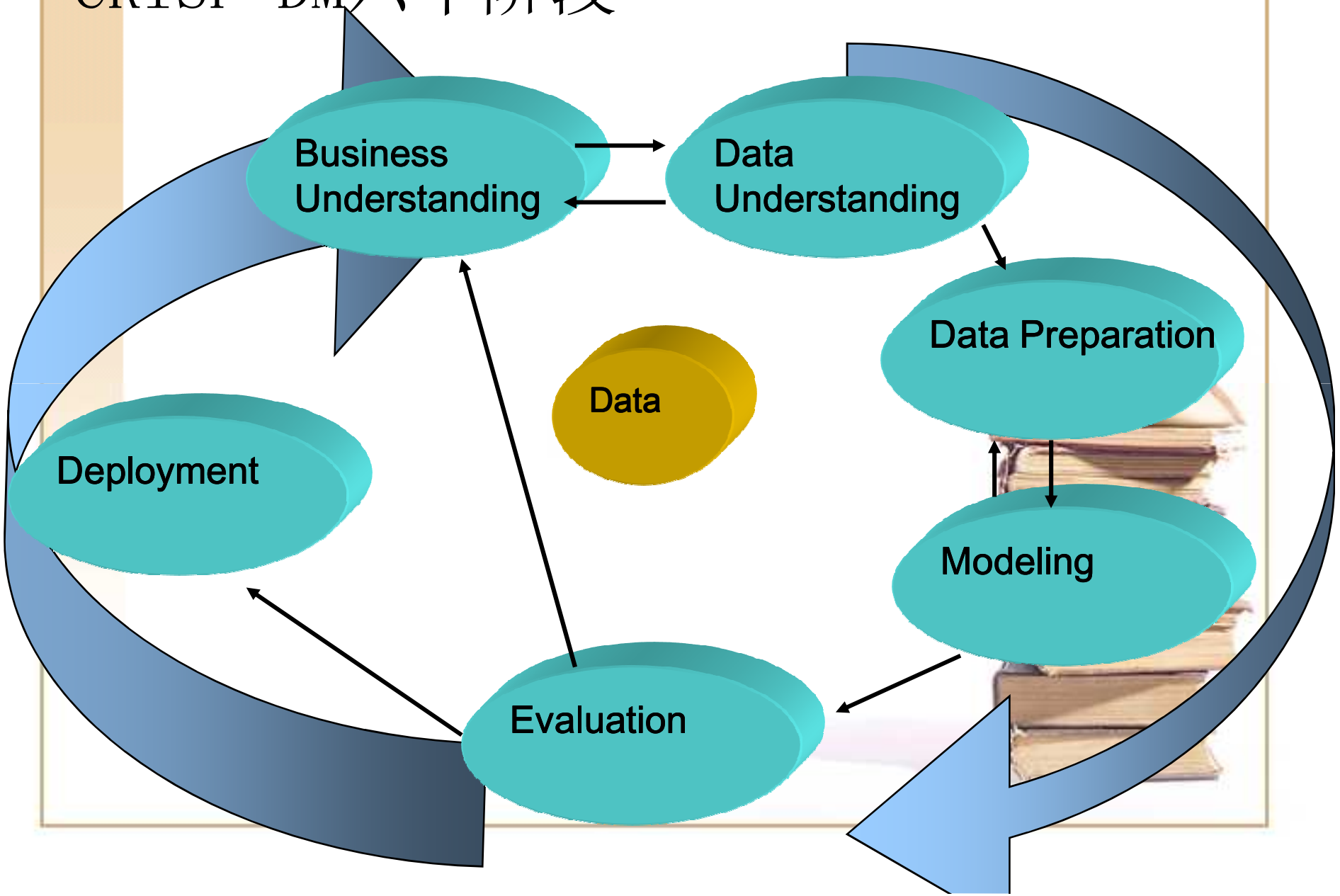
比较表

服务 / 属性	Amazon EC2	Google App Engine	Microsoft Azure	Yahoo Hadoop
架构	Iaas/Paas	Paas	Paas	Software
服务型态	Compute/ Storage	Web application	Web and non-web	Software
管理技术	OS on Xen hypervisor	Application container	OS through Fabric controller	Map / Reduce Architecture
用户接口	EC2 Command-line tools	Web-based Administration console	Windows Azure portal	Command line and web
APIs	yes	yes	yes	yes
收费	yes	maybe	yes	no
程序语言	AMI(Amazon Machine Image)	Python	.NET framework	Java,

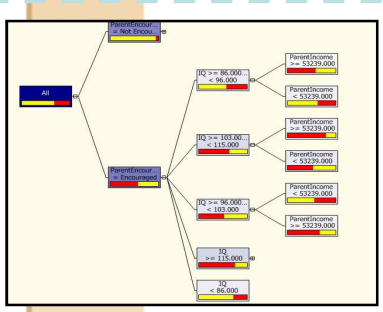
Data Mining包含六大领域



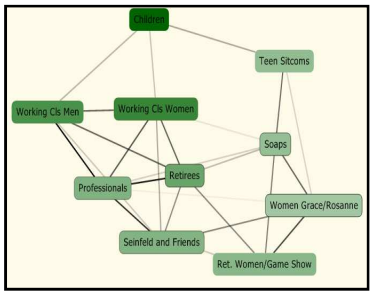
CRISP-DM六个阶段



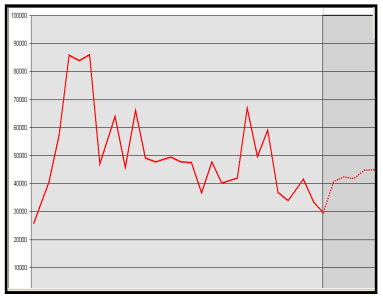
完整的算法 *SQL Server 2008已提供*



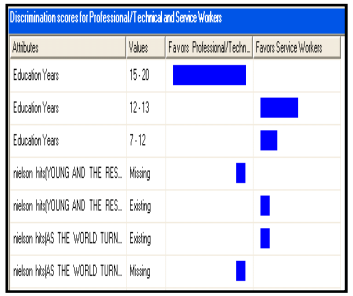
决策树



群集

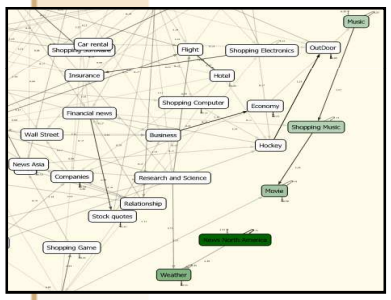


时间序列

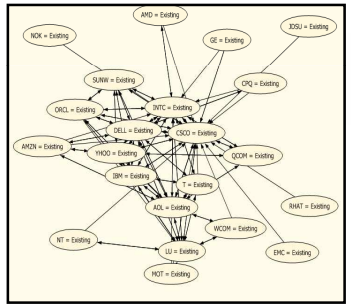


贝氏机率分类

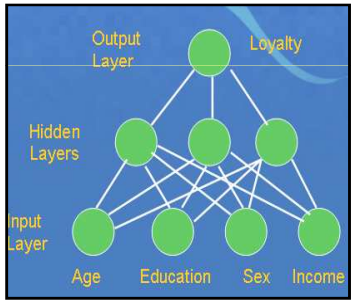
SQL Server 2000已提供



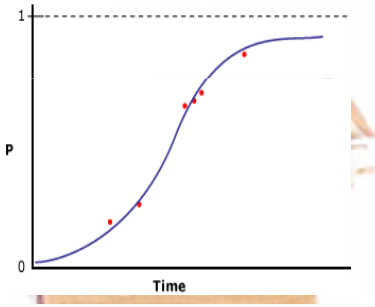
时序群集



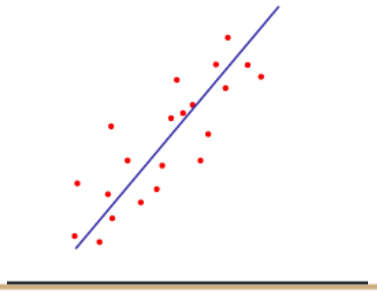
关联规则



类神经网络



罗吉斯回归



线性回归

文字数据挖掘

常用的Data Mining及统计学习方法-1

Binary Classifier (二元分类)
Numeric Predictor (数值预测)
Time Series (时间序列)
C&R TREE (分类回归树)
Quick Unbiased Efficient Statistical Tree (QUEST判定树模型)
CHAID (分类树)
Decision List (判定树列表)
Regression (线性回归分析)
PCA/Factor (主成分分析)
Neural Net (类神经网络)
C5.0 (判定树)
Feature Selection (特征选取)
Discriminant Analysis (判别分析)
Logistic (罗吉斯回归)
Generalize Linear Model (广义线性模型)
Cox Regression

常用的Data Mining及统计学习方法-2

Support Vector Machine (SVM支持向量机)
Bayes Net (贝氏分类器)
SLRM (自我学习反应模型)
GRI关联
Apriori关联
CARMA关联(连续交易)
Sequence Clusterc序列关联
K-Means (K-Means分群)
Kohonen (自我组织化)
Two-Step (二阶段)
Anomaly (异常检测)
Random Forests (随机森林)
ICA (独立成分分析)
Multivariate adaptive regression spline (MARS多元适应性回归平滑)
Pmml(预测模型标记语言)
Boosting

使用软件 常用

SQL server 2008
SPSS 17 (PAWS) --IBM
SAS
SQL 2008+Excel (2008)-Data Mining Add-in
Clementine 12.0
Statistica 7.0
WEKA
R → Cloud R
R+Excel ADD-IN ……*还有更多云端软件*

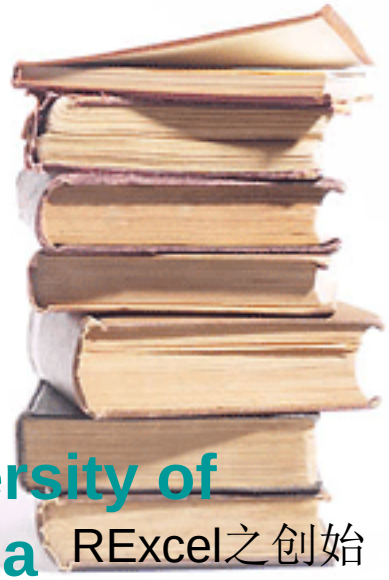
R Excel
数据挖掘发展趋势

- RExcel
  - RExcel之创始
  - RExcel之启动
  - RExcel之应用
    - 数据导入
    - 资料分析
    - 结果保存

statconn之“幕后黑手”（The masterminds behind statconn）

- Thomas Baier（1971-）
- 在不同环境中应用R
  - R/Scilab (D)COM Server
    - RExcel (1998)

- Erich Neuwirth（1948-）
- RExcel的主要作者

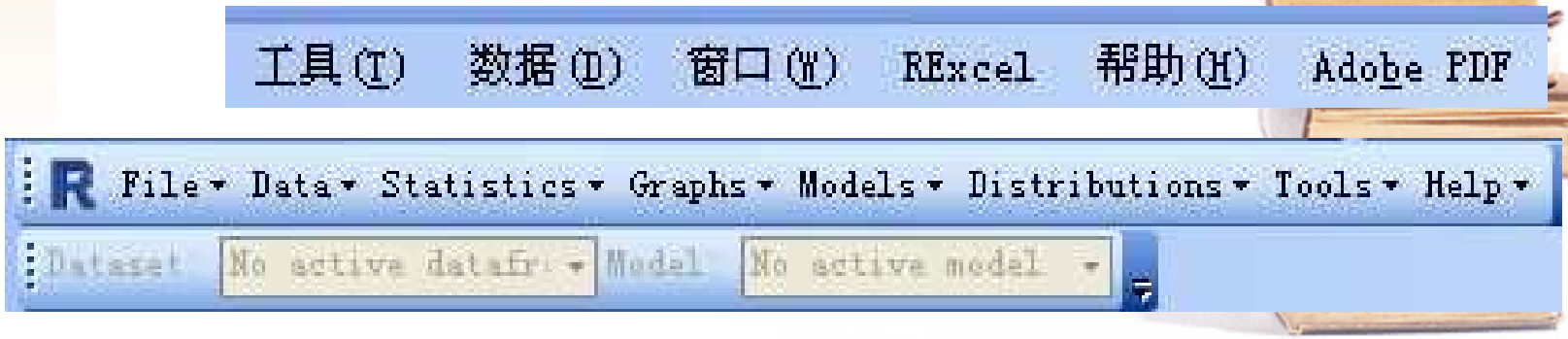


University of Vienna RExcel之创始

•http://rcom.univie.ac.at/

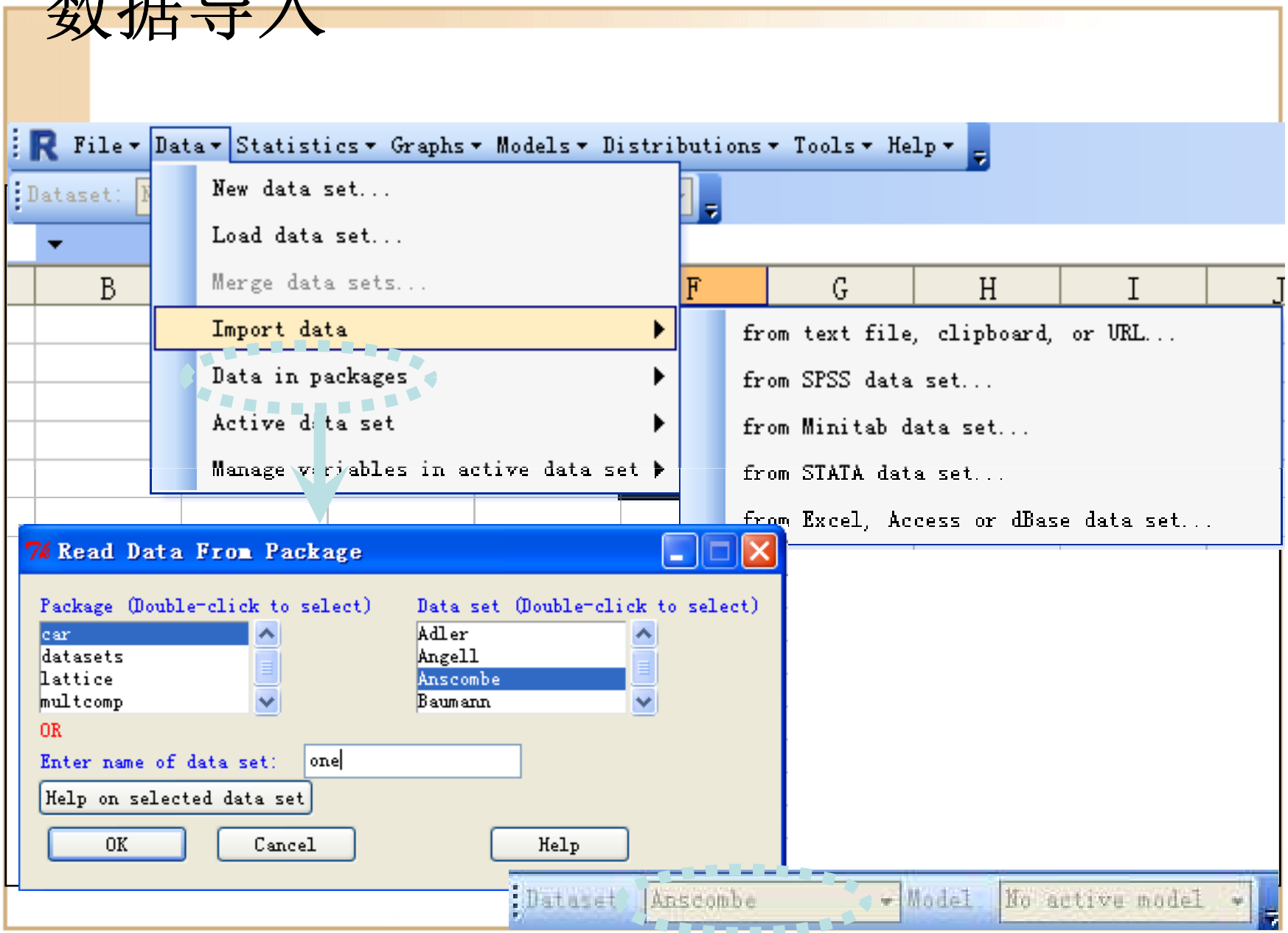
安装：

- 安装：http://rcom.univie.ac.at/
  - 据网页提示手动逐步安装
  - 直接下载RAndFriends压缩包
- 安装须知
  - R的版本2.9.0以上
  - Excel的版本03、07均可



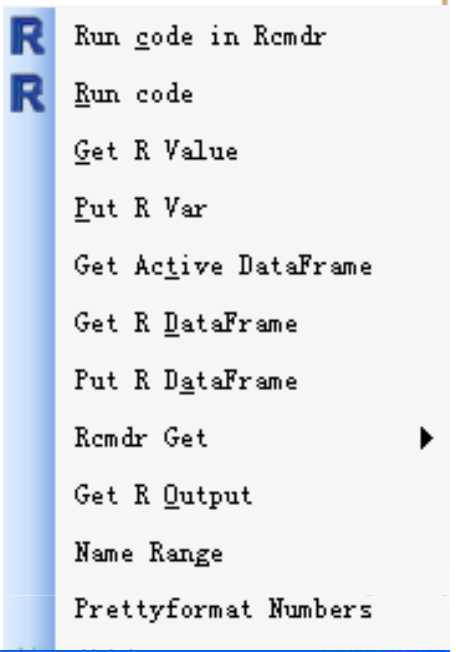
RExcel之启动

数据导入


File
Data
Statistics
Graphs
Models
Distributions
Tools
Help
Dataset:
New data set...
Load data set...
Merge data sets...
Import data
Data in packages
Active data set
Manage variables in active data set
from text file, clipboard, or URL...
from SPSS data set...
from Minitab data set...
from STATA data set...
from Excel, Access or dBase data set...
Read Data From Package
Package (Double-click to select)
car
datasets
lattice
multcomp
Data set (Double-click to select)
Adler
Angell
Anscombe
Baumann
OR
Enter name of data set:
one
Help on selected data set
OK
Cancel
Help
Dataset
Anscombe
Model
No active model

资料分析

- 任何程序均可写在单元格、Commander、R console中
- 右图为右键功能



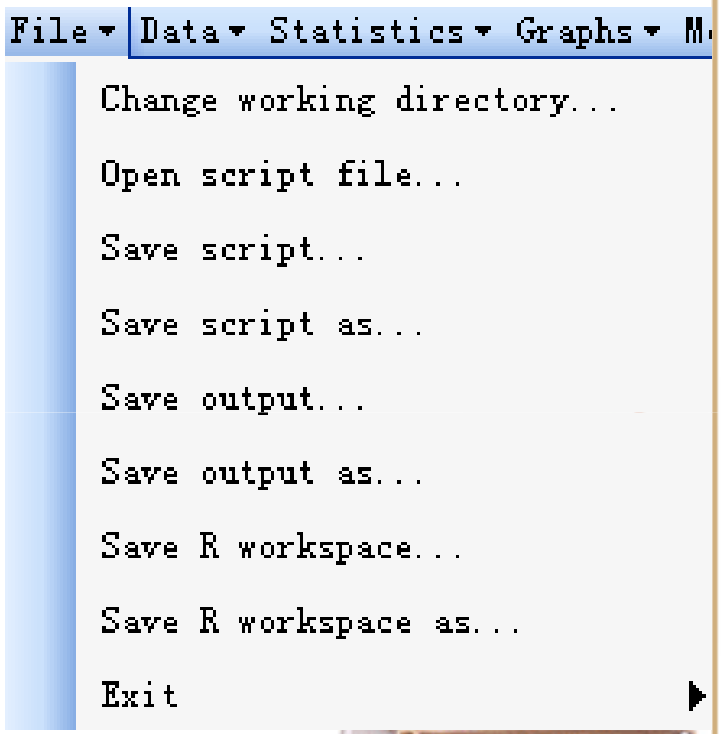
三个部分 RCommander

资料分析

- Statistics
  - 描述统计，简单参数和非参数检验、线性模型
- Graphs
  - 各种统计图表
- Models
  - 经典统计模型
- Distributions
  - 各种分布的分位数、分布图、抽样、尾概率等

结果保存

- 可直接储存在Excel中
- 其它储存方法如右图



数据挖掘在各产业的应用

- 金融服务业

  客户贡献度分析、信用评分、风险评估、客户区隔、交叉营销等。

- 保险业

  顾客贡献度分析、信用评分、风险评估、客户区隔、交叉营销、客户流失分析和诈欺侦测等。

- 电信业

  顾客贡献度分析、信用评分、客户区隔、交叉营销、客户流失分析、销售预测和诈欺侦测等。

数据挖掘在各产业的应用

- 制造业

  客户贡献度分析、质量管理、营销绩效分析、生产分析和存货分析等。

- 零售业

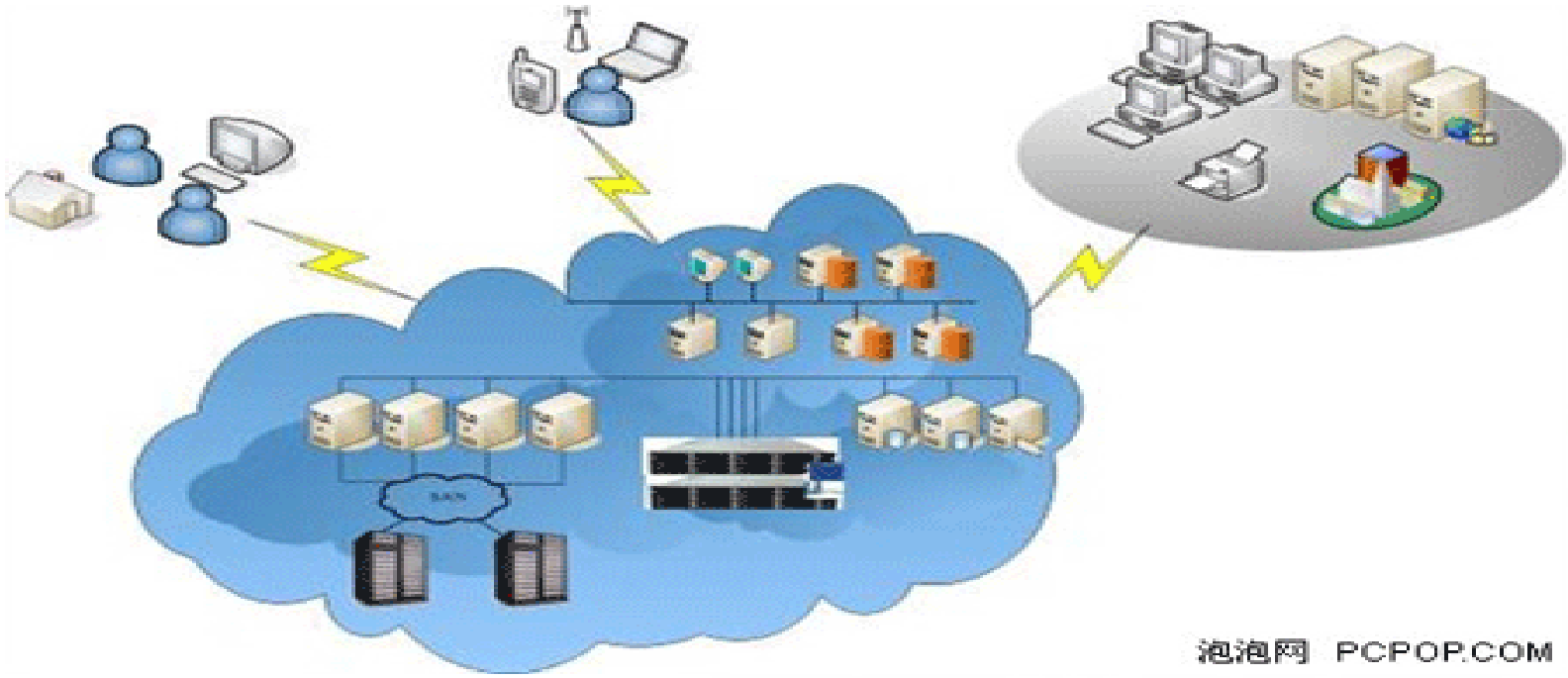
  客户忠诚度、客户区隔、购物篮分析、定价分析、交叉营销和销售预测等。

- 生物科技、医疗保健、航天空业、环境、法律等

商业智慧的核心

- 如何收集资料
  - 营运数据，市场调查资料，固定**Panel**追踪
- 如何管理数据
  - ETL，Data warehousing
- 如何从数据中获取智能
  - Data Mining，OLAP，Statistics
- 如何应用智能
  - 营销策略，主管决策，互动化**CRM**机制

- 云端运算可以实现适应端通过在在线传数据或购买数据，通过云数据仓库，进行数据仓库建模或数据抽取，在线支付使用资料采矿工具和商业智能相关处理软件



IIaaS是SaaS的延伸 I^2aaS

- 数据挖掘和商业智能的原理相似，均由数据提供信息、产生知识，再由知识累积智能。而云端运算可以使这个过程在因特网上得以实现。也就是说云端运算可以提供基于SaaS的知识与智能分析的服务（Information&Intelligence as a Service），简称IIaaS；I^2aaS，它是SaaS的延伸。

云端运算产业类型

IIaaS I^2aaS

Information & intelligence as a Service

SaaS

Software as a Service

PaaS

Platform as a Service

IaaS

Infrastructure as a Service

Waves of Innovation

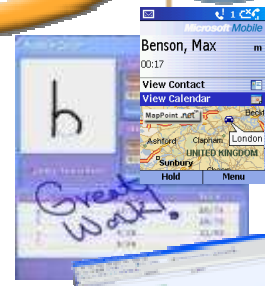
- Speech/Writing
- Devices
- Wi-Fi/Broadband
- Web Services
- Trusted Computing Hardware
- Rights Management

Development

Hard Engineering ➔ Intellectual Property

- XML/SOAP
- HTTP/HTML
- SMTP

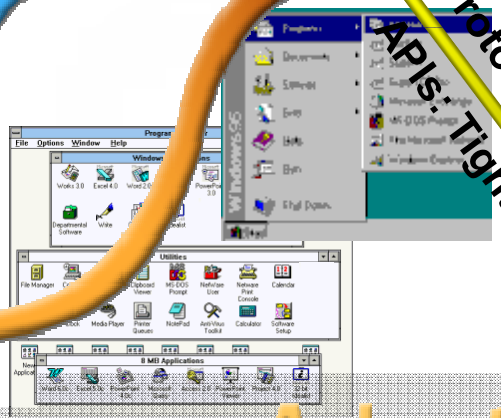
- Cloud Computing
- Email Clients
- Web Browsers



- Mouse
- GUI
- LANs

APIs: Tightly Coupled

Protocols: Loosely Coupled



- PC Architecture
- DOS

- Spreadsheets
- Word Processors

Adoption

Intellectual Property ➔ Consumer Benefit

Today

PC	Applications	Internet	Web Apps
Mid 80s	Late 80s-Mid 90s	Mid 90s	Mid 00s - . .

SUBSCRIBER COPY NOT FOR RESALE FEBRUARY 21, 2011

Revolution in Egypt

Joe Klein: What the U.S. should do
On the Street: Hope meets anxiety
Muslim Brotherhood: What it wants

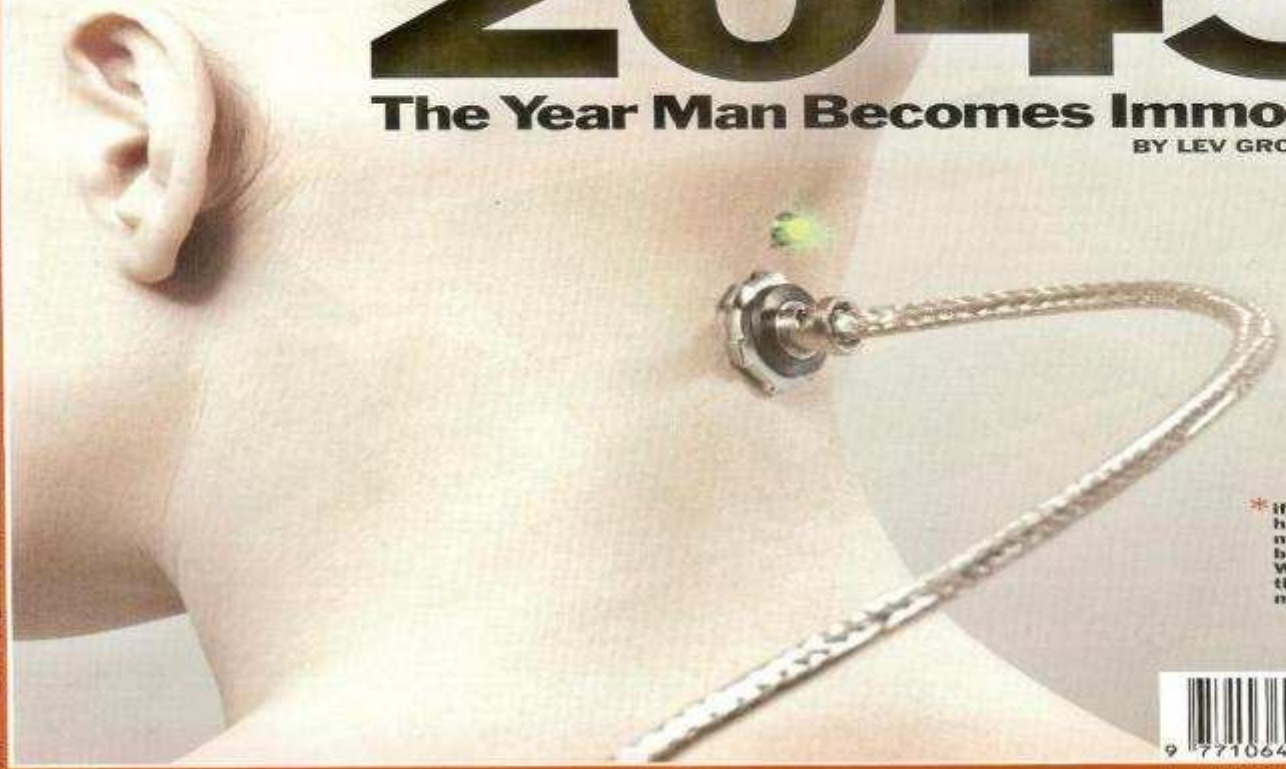
Oscars: Portraits of star power

TIME

2045

The Year Man Becomes Immortal*

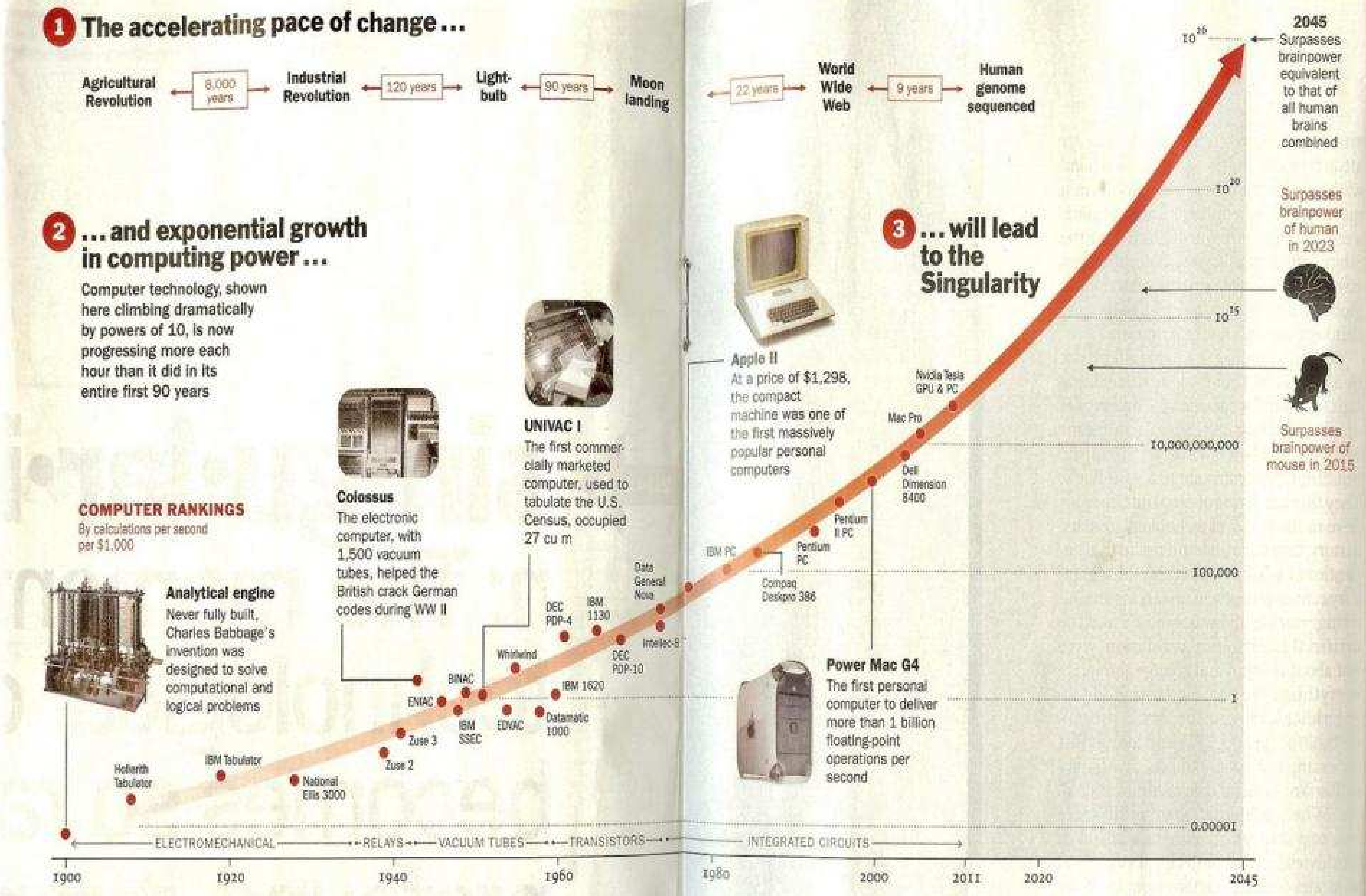
BY LEV GROSSMAN



*if you believe humans and machines will become one. Welcome to the Singularity movement

9 771064 030005 07

www.time.com



on, there's no reason to think computers

Probably. It's impossible to predict the

idea; it's a serious hypothesis about the

he called an "intelligence explosion":

云端运算在商业智慧上的运用

- 一方面，可以通过云的数据仓库实现海量数据的高效计算。
- 另一方面，可以实现在线支付使用资料采矿工具和商业智能相关分析处理软件。

以云端为基础的医疗健康照护服务平台

云端健康照护服务

数据源：

http://www.guidertech.com/02_produter_15.htm

数据源：

http://www.guidertech.com/02_produter_15.htm

云端健康照护—情意计算

- 情意计算是负责与人类情意有关的资料计算，注重在情意的辨识与表达，正确判断出使用者可能知情意，才能了解影响其行为举止的原因。
- 通常情意是对外在周围的感受，较为明显容易被观察及发现，藉由有限的传感器、摄影设备及处理数据能力的计算机，即可辨识感知使用者的情绪，如喜、怒、哀、乐。情意较为明显观察，以非语言的方式来表达。

情义计算方式	内容	辨识正确率
语音识别	根据声音电图起伏，辨识具有情意之声音语调，一般还会搭配侦测分析生理讯号，提高辨识率。	50%至87.5%
解读肢体语言	主要根据使用者头部、手部及脚部之动作及姿势，来判断其意向及情意。	比脸部辨识及侦测分析生理讯号低
脸部辨识	将脸部分成几个特征区域，如眉毛、眼睛、嘴巴等等，观察分析其表情变化，脸部表情是最直接的情意表达，它的辨识率最高。	88%至89%
侦测分析生理讯号	透过侦测用户生理讯号，如心跳、体温、肌动电流图、血压、皮肤导电度、呼吸率等等，利用医学分析可辨识出其可能的情意。	81%

以云端为基础的智能型老人居家照护软件服务平台

- 以云端为基础的智能型老人居家照护软件服务平台(图15)，它包括三个主要模块：**数据管理员、云端健康管理员及接口管理员**。
- 数据管理员将收集老年人居家生理讯号及影像数据，透过特征撷取及数据融合技术，以取出重要的健康信息。
- 云端健康管理员将负责情境分析、行为辨识、睡眠分析及情意分析。
- 接口管理员负责用户授权、云端共享、紧急医疗及普适服务等功能。

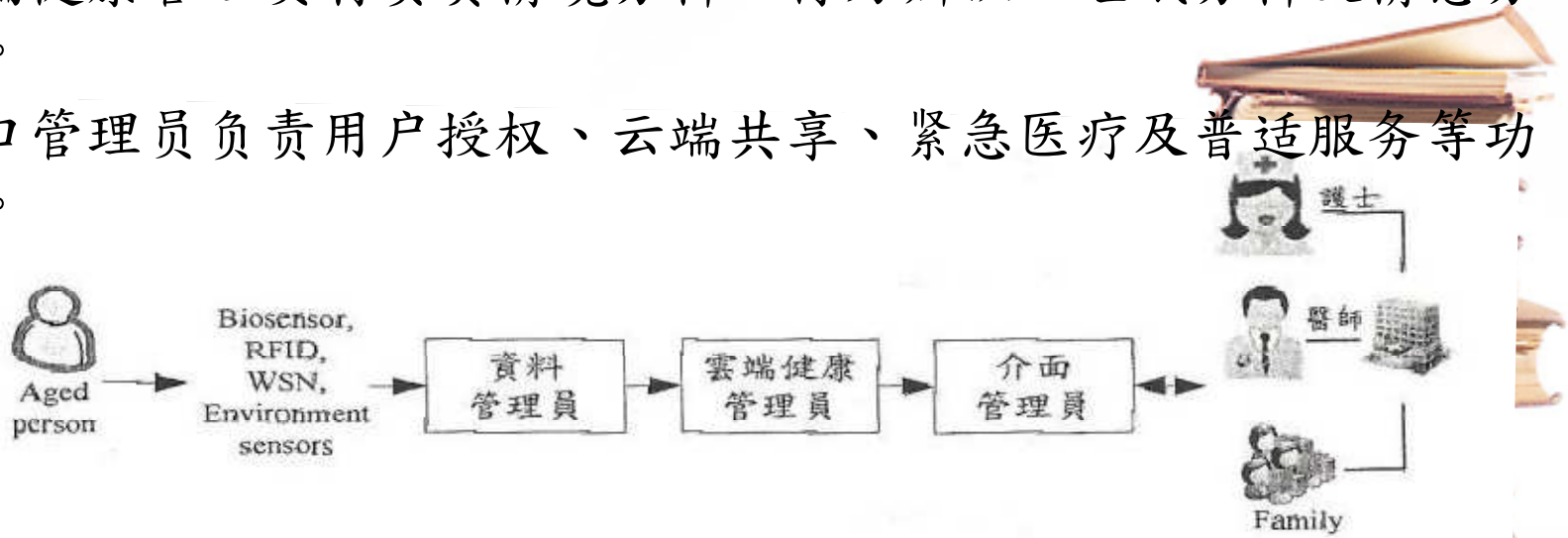


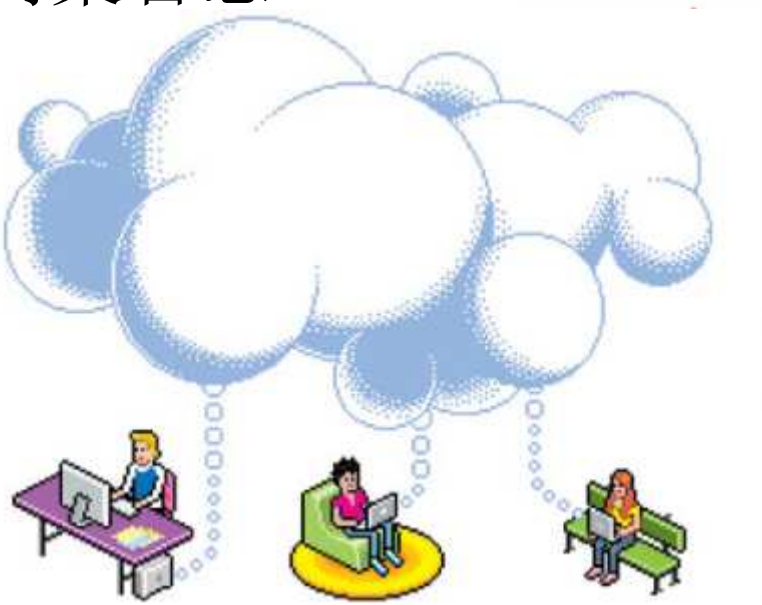
圖 15 以雲端為基礎的智慧型老人居家照護軟體服務平台

云端健康照护平台—功能及技术

模块名称	功能	应用技术
数据管理员	资料收集	RFID技术、无限感测网络、加速度计
	特征撷取	讯号分析、图像处理
	资料融合	智能型系统技术
云端健康管理员	情境分析	情境模型、本体论
	行为辨识	类神经网络
	睡眠分析	机器学习、推理方法
	情意分析	模糊理论
接口管理员	使用者授权	用户模型、授权认证机制
	云端共享	云端网关、HL7协定
	紧急医疗	智能型代理人
	普适服务	行动装置通用接口、服务搜寻

雲端運算在商業智慧上的運用

- 一方面，可以通過雲的資料倉庫實現海量資料的高效計算。
- 另一方面，可以實現線上支付使用資料採礦工具和商業智慧相關分析處理軟體。



臺灣的導入客戶成功案例

Quanta Computer

DATA MINING大幅提升的效能與新增強化的功能，吸引我們全面升級以發揮新技術的效益。廣達以Intel搭配SQLServer，所耗費的成本卻低於Unix的三分之一，廣達創造了絕對的成本優勢」

PSC

DATA MINING在效能的顯著提升。原先的Unix資料庫執行某項資料處理作業，約5分鐘，但在SQL Server2008應用UDM模型，於32位只需23秒，64位則僅需11秒。過去每支報表平均開發時間是一星期，但在新環境，則只需要半天。

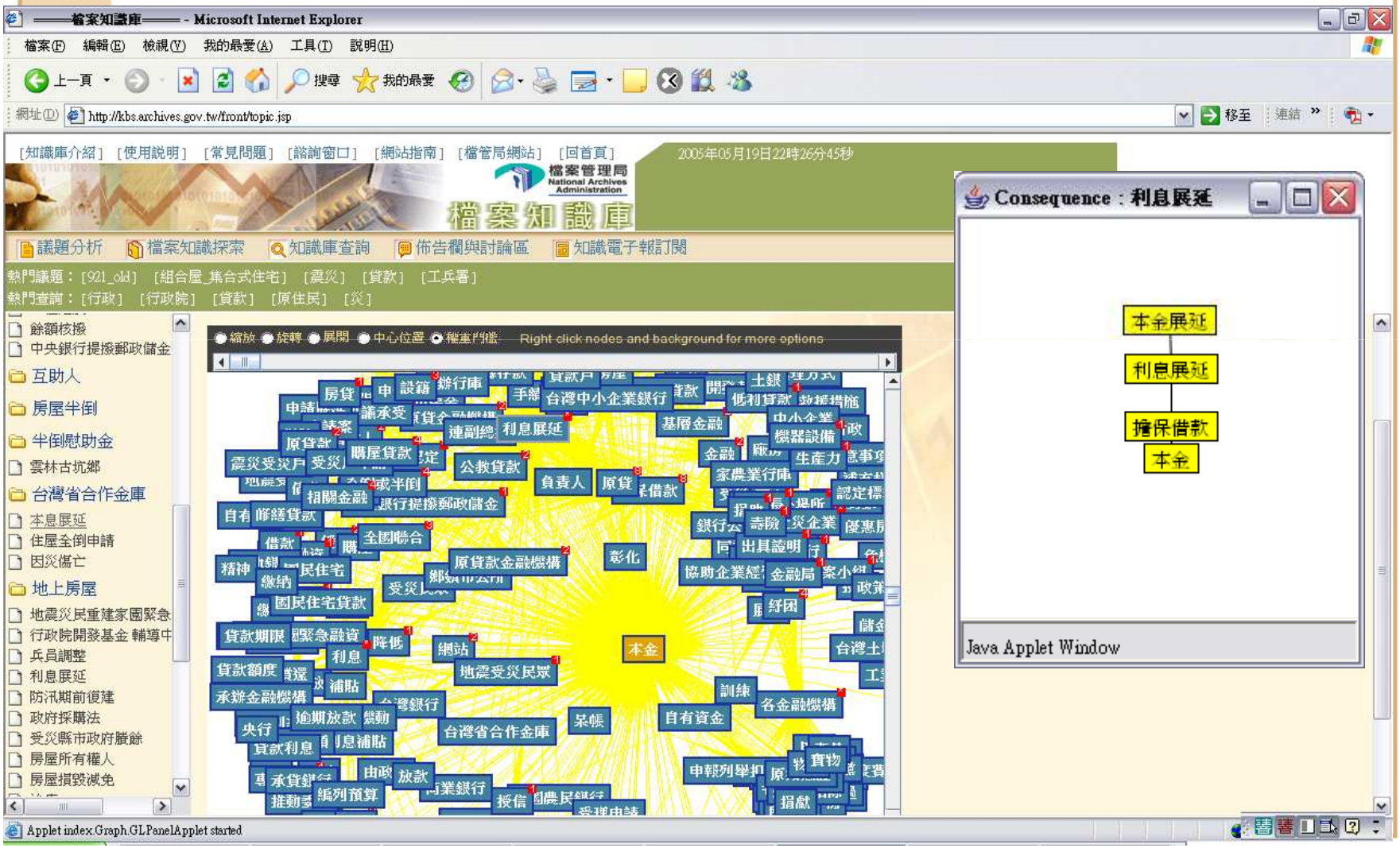
新光金控 MACOTO BANK 誠泰銀行

將以應用面的強化為重點，善加利用DATA MINING新增的多項進階功能，以配合持續進化中的業務需求，例如：交叉行銷、巴塞爾協定（BASELII）」

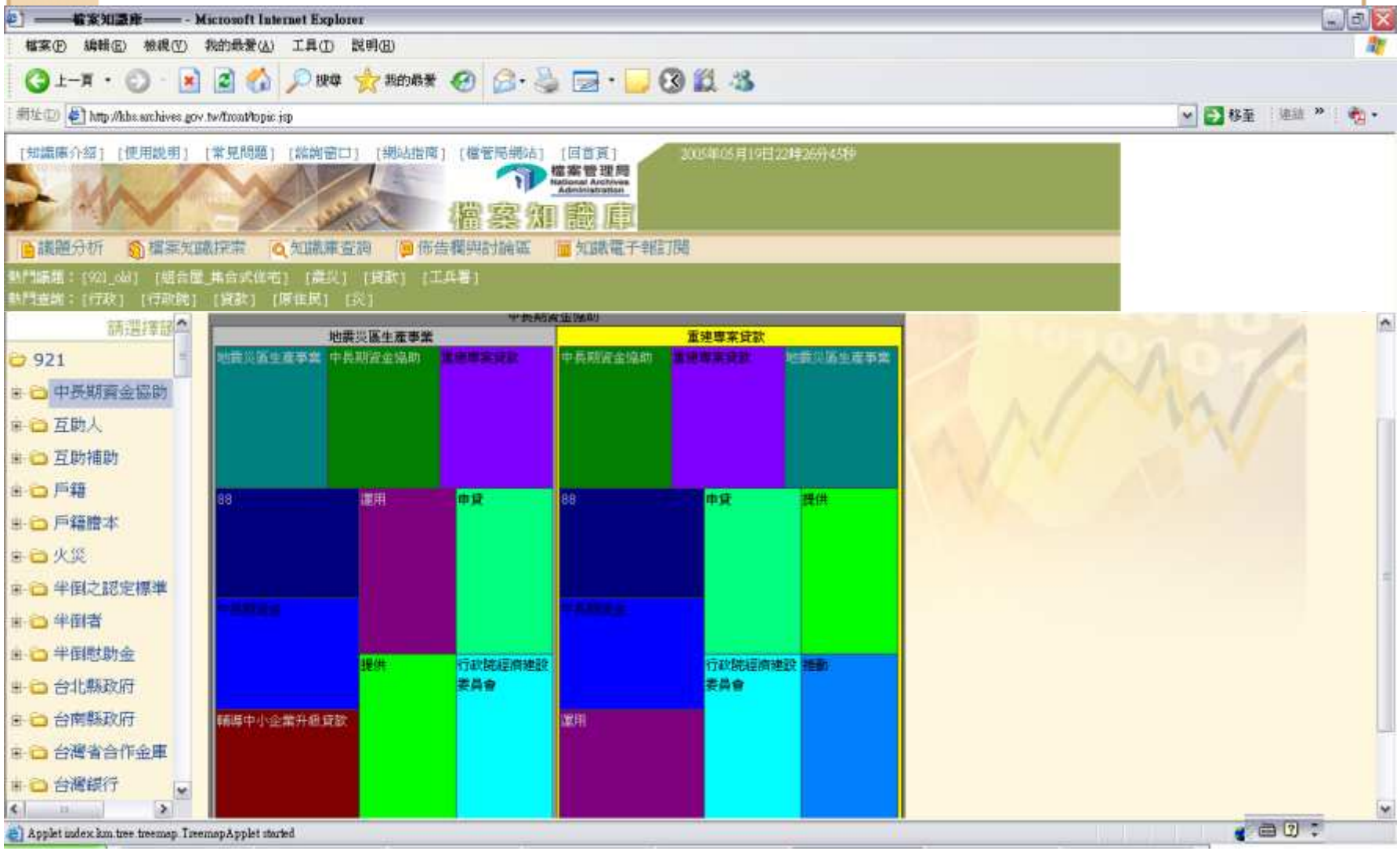
威達電股份有限公司 ICP Electronics Inc.

DATA MINING具備更多的工具與更成熟的功能，與領導品牌的BI解決方案並駕其驅，但投資成本卻更低，最能滿足威達電以最少投資達成最大效益的目標

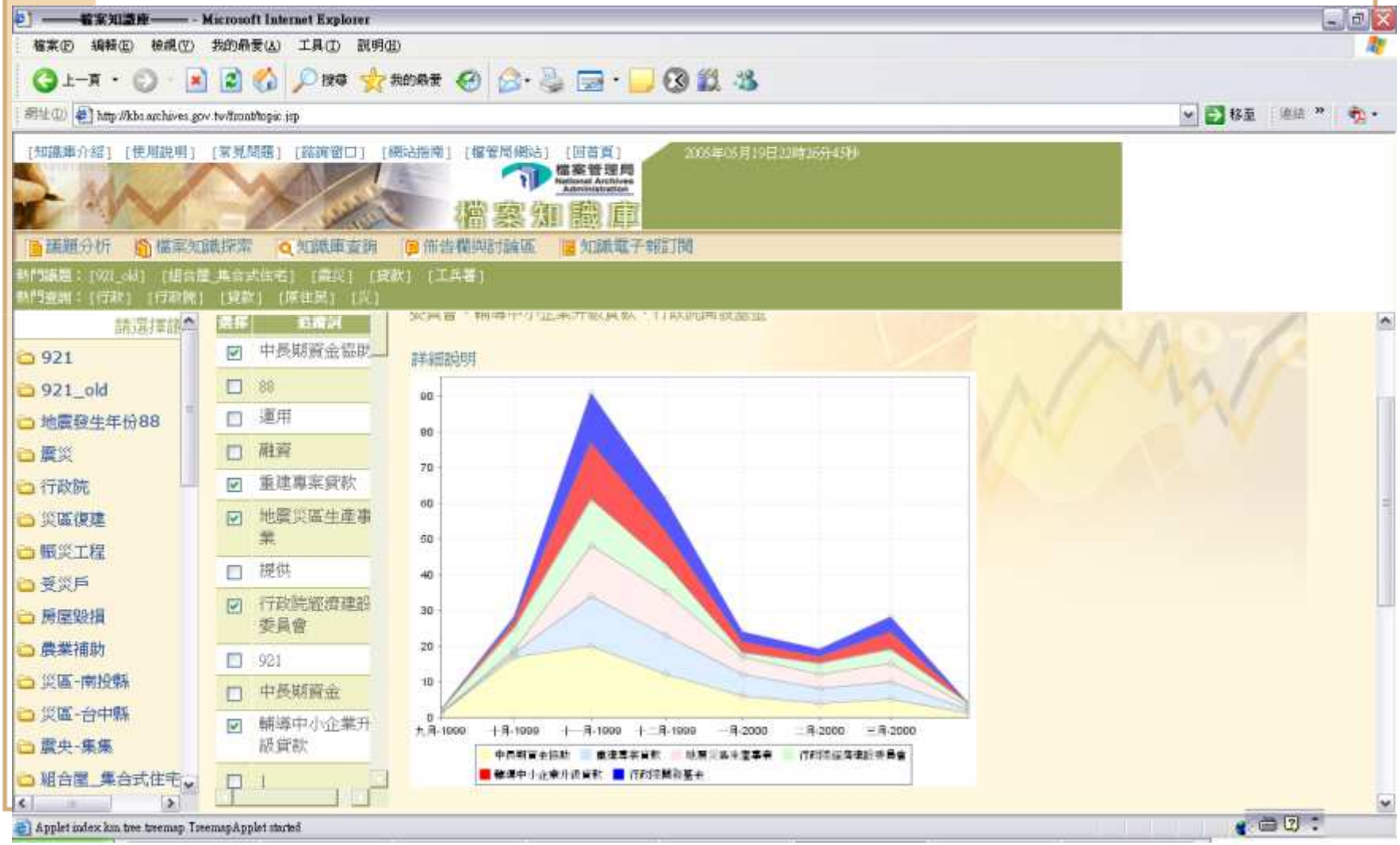
知識脈絡


檔案知識庫 - Microsoft Internet Explorer
http://kbs.archives.gov.tw/front/topic.jsp
檔案知識庫
議題分析
檔案知識探索
知識庫查詢
佈告欄與討論區
知識電子報訂閱
Consequence : 利息展延
本金展延
利息展延
擔保借款
本金
Java Applet Window
Applet index.Graph.GLPanelApplet started

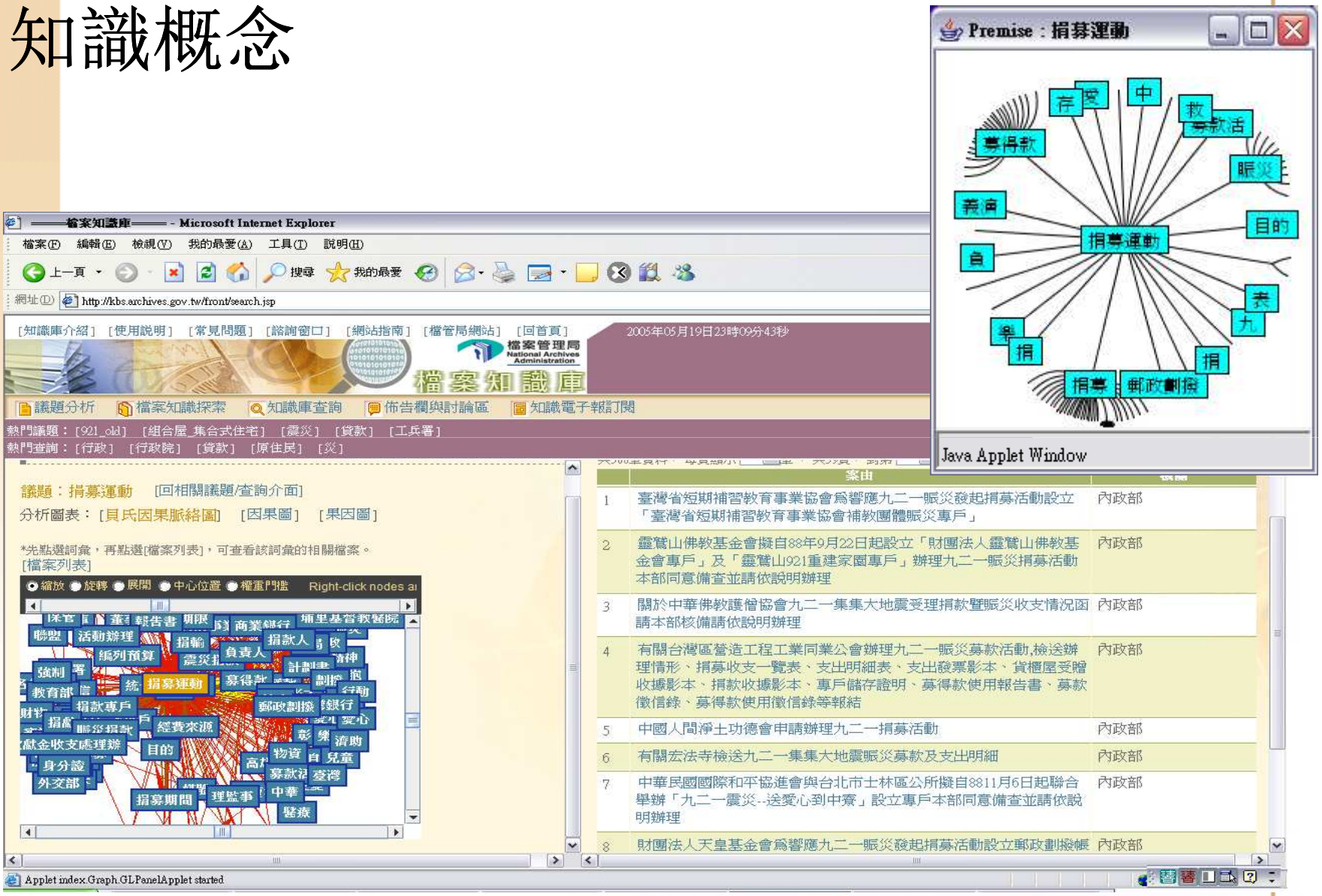
知識地圖


檔案知識庫 - Microsoft Internet Explorer
http://khs.archives.gov.tw/front/topic.jsp
檔案知識庫
921
中長期資金協助
互助人
互助補助
戶籍
戶籍謄本
火災
半倒之認定標準
半倒者
半倒慰助金
台北縣政府
台南縣政府
台灣省合作金庫
台灣銀行
地震災區生產事業
重建專案貸款
中長期資金協助
提供
行政院經濟建設委員會
輔導中小企業升級貸款
Applet index.km.tree.treemap.TreemapApplet started

事件追蹤


Microsoft Internet Explorer
檔案知識庫
921
921_old
地震發生年份88
震災
行政院
災區復建
賑災工程
受災戶
房屋毀損
農業補助
災區-南投縣
災區-台中縣
震央-集集
組合屋_集合式住宅
中長期資金
88
運用
融資
重建專案貸款
地震災區生產事業
提供
921
中長期資金
輔導中小企業升級貸款
詳細說明

知識概念


Premise：捐募運動
存
愛
中
救
募得款
賑災
義舉
目的
自
捐募運動
表
九
舉
捐
捐
捐募
郵政劃撥
Java Applet Window
檔案知識庫 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D) http://kbs.archives.gov.tw/front/search.jsp
[知識庫介紹] [使用說明] [常見問題] [諮詢窗口] [網站指南] [檔管局網站] [回首頁]
2005年05月19日23時09分43秒
檔案管理局 National Archives Administration
檔案知識庫
議題分析 檔案知識探索 知識庫查詢 佈告欄與討論區 知識電子報訂閱
熱門議題：[921_old] [組合屋_集合式住宅] [震災] [貸款] [工兵署]
熱門查詢：[行政] [行政院] [貸款] [原住民] [災]
議題：捐募運動 [回相關議題/查詢介面]
分析圖表：[貝氏因果脈絡圖] [因果圖] [果因圖]
*先點選詞彙，再點選[檔案列表]，可查看該詞彙的相關檔案。
[檔案列表]
縮放 旋轉 展開 中心位置 權重門檻 Right-click nodes ar
捐募運動
捐款人
負責人
編列預算
活動辦理
聯盟
強制
教育部
捐款專戶
經費來源
目的
身分證
外交部
捐募期間
理監事
中華
醫療
郵政劃撥
物資
兒童
濟助
案由
1 臺灣省短期補習教育事業協會爲響應九二一賑災發起捐募活動設立「臺灣省短期補習教育事業協會補教團體賑災專戶」 內政部
2 靈鷲山佛教基金會擬自88年9月22日起設立「財團法人靈鷲山佛教基金會專戶」及「靈鷲山921重建家園專戶」辦理九二一賑災捐募活動本部同意備查並請依說明辦理 內政部
3 關於中華佛教護僧協會九二一集集大地震受理捐款暨賑災收支情況函請本部核備請依說明辦理 內政部
4 有關台灣區營造工程工業同業公會辦理九二一賑災募款活動,檢送辦理情形、捐募收支一覽表、支出明細表、支出發票影本、貨櫃屋受贈收據影本、捐款收據影本、專戶儲存證明、募得款使用報告書、募款徵信錄、募得款使用徵信錄等報結 內政部
5 中國人間淨土功德會申請辦理九二一捐募活動 內政部
6 有關宏法寺檢送九二一集集大地震賑災募款及支出明細 內政部
7 中華民國國際和平協進會與台北市士林區公所擬自8811月6日起聯合舉辦「九二一震災--送愛心到中寮」設立專戶本部同意備查並請依說明辦理 內政部
8 財團法人天皇基金會爲響應九二一賑災發起捐募活動設立郵政劃撥帳 內政部
Applet index.Graph.GLPanelApplet started

資料採礦範例–推薦知識社群

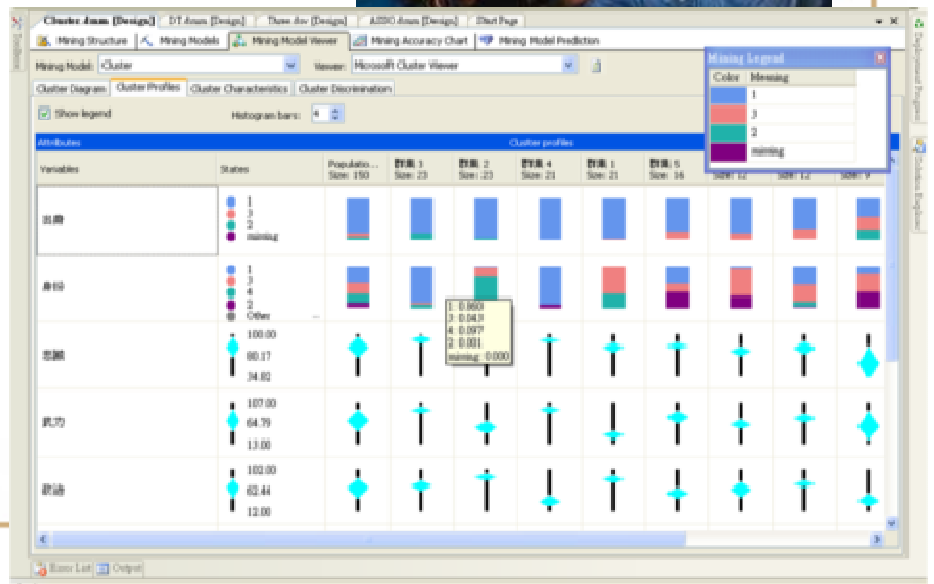
Keyword &
流覽行為
統計

推薦知識社群

Data Mining
Center

比對相似度

知識社群



完整涵蓋主要國家級數位典藏重要計畫

臺灣師範大學
尋根網
2002.3

文建會
國家文化
資料庫
2003.1

國家圖書館
走讀臺灣
2004.1

外交部
數位影像
資料庫
2004.12

原民會
臺灣原住民
影音資料庫
2005.5

故宮
數位圖檔資
料管理系統
2006.5

2002 — 2006

國家圖書館
臺灣記憶網
2002.12

檔案管理局
檔案知識庫
2003.9

國立傳藝中心
典藏系統
2004.11

文建會
圖書與影音
出版品系統
2005.3

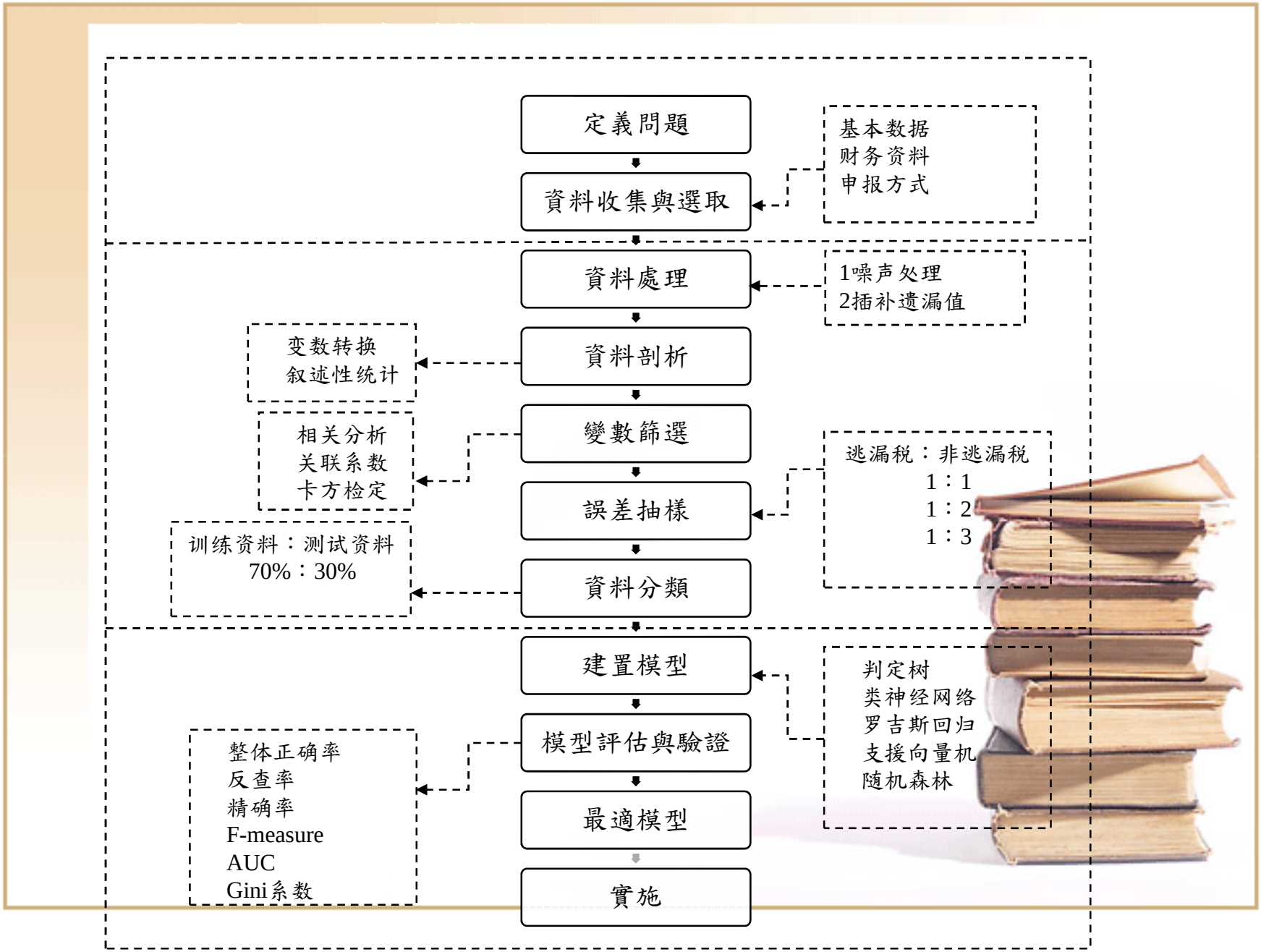
臺灣文學館
文學文物數
字典藏系統
2006.5

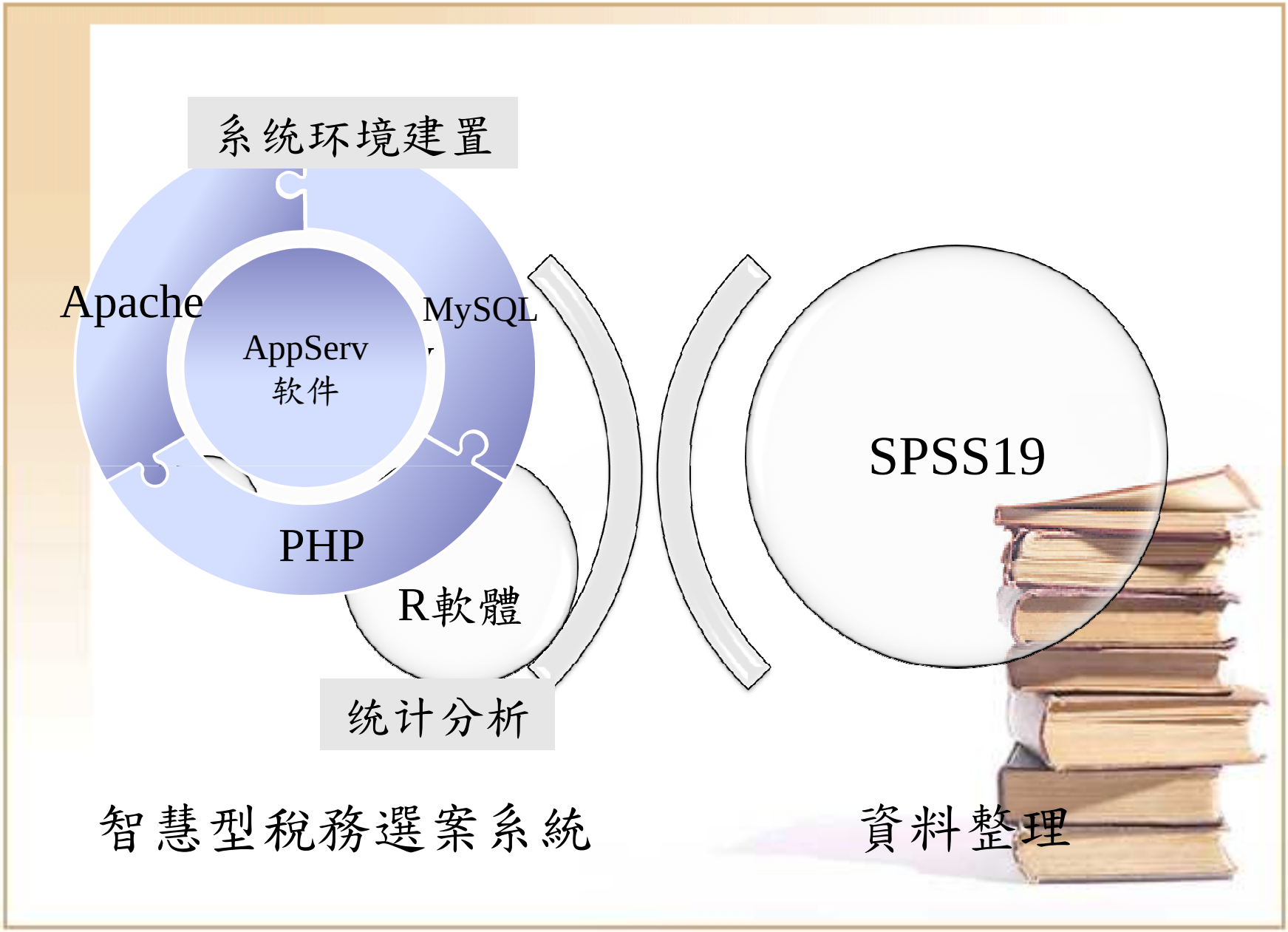
國立傳統藝術中心
主題知識網
2006.12

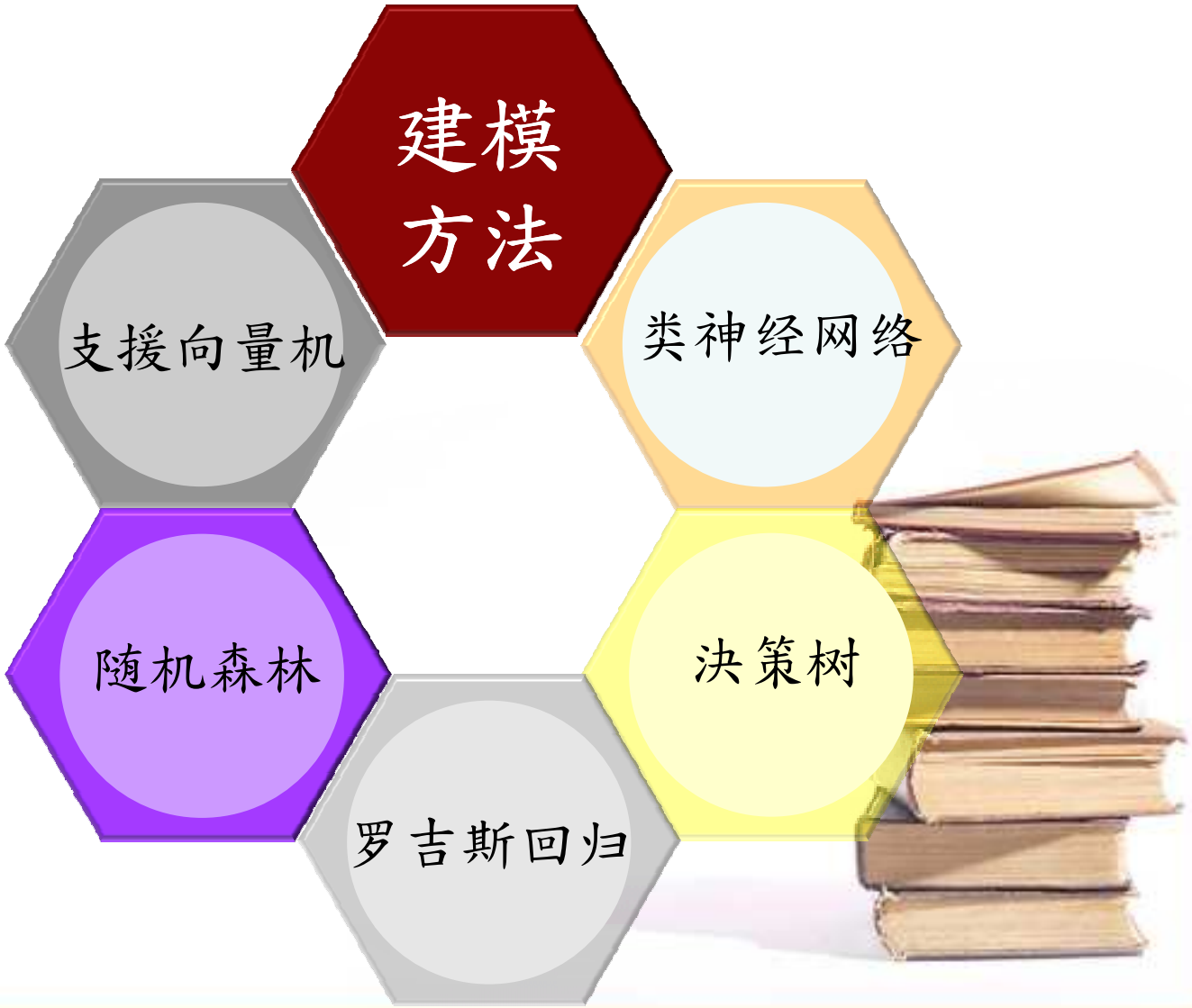
DataMining + BI 的應用 ***(IN Cloud Computing)***

Customer-focused	Operations-focused	Research-focused
●Life-time Value ●Market-Basket Analysis ●Profiling & Segmentation ●Retention ●Target Market ●Acquisition ●Knowledge Portal ●Cross-Selling ●Campaign Management ●E-Commerce	●Profitability Analysis ●Pricing ●Fraud Detection ●Risk Assessment ●Portfolio Management ●Employee Turnover ●Cash Management ●Production Efficiency ●Network Performance ●Manufacturing Processes	●Combinatorial Chemistry ●Genetic Research ●Epidemiology

应用数据挖掘与云端技术于智能型税务选案


定義問題
資料收集與選取
基本数据
财务资料
申报方式
資料處理
1噪声处理
2插补遗漏值
資料剖析
变数转换
叙述性统计
變數篩選
相关分析
关联系数
卡方检定
誤差抽樣
逃漏税：非逃漏税
1：1
1：2
1：3
資料分類
训练资料：测试资料
70%：30%
建置模型
判定树
类神经网络
罗吉斯回归
支援向量机
随机森林
模型評估與驗證
整体正确率
反查率
精确率
F-measure
AUC
Gini系数
最適模型
實施


系统环境建置
Apache
AppServ
软件
MySQL
PHP
R軟體
統计分析
SPSS19
智慧型稅務選案系統
資料整理


建模方法
支援向量机
类神经网络
随机森林
决策树
罗吉斯回归

智能型税务选案系统介绍

- 资料准备
- 智能型税务选案系统介绍
- 抽样比例与模型比较
- 建置模型
- 模模型评估
- 及验证

http://140.136.134.52/ias

首页

- 资料准备
- 智能型税务选案系统介绍
- 抽样比例与模型比较
- 建置模型
- 模型评估
- 及验证

ITAS系统相关数据

- 资料准备
- 智能型税务选案系统介绍
- 抽样比例与模型比较
- 建置模型
- 模型评估
- 及验证

首页

- 资料准备
- 智能型税务选案系统介绍
- 抽样比例与模型比较
- 建置模型
- 模型评估
- 及验证

统计方法介绍

- 资料准备
- 智能型税务选案系统介绍
- 抽样比例与模型比较
- 建置模型
- 模型评估
- 及验证

建模方法介绍

二元羅吉斯迴歸	羅吉斯迴歸觀念可由線性迴歸的方式引入，主要在分析二分類或有次序的依變數與解釋變數間的關係，(例如：「逃漏稅」與「未逃漏稅」)，可解決傳統線性迴歸模型不能處理反應變數是兩個類別變數的缺點。
類神經網路	主要嘗試著模仿人類的神經系統，與人腦功能相似，具有平行處理特性、容錯特性、結合式記憶的特性、解決最佳化問題。類神經網路是由很多非線性的運算單元和位於這些運算單元間的眾多連結所組成，而這些運算單元通常是以平行且分散的方式在作運算，如此就可以同時處理大量的資料。
隨機森林	是一種拔靴集合為基礎所發展出的決策樹，其運算原理乃是一種整體學習方法，將測試資料的拔靴樣本，來建構整棵決策樹中的分支，並以簡單多數表決最後之結果，所以隨機森林是由許多決策樹子集合而形成的決策樹。決策樹會依照大數法則與決策樹結構進行收斂，因此即使隨機森林有許多決策樹子集合也不會有過度配適的情況產生。
決類樹	主要功能是藉由分類已知的事情來建立樹狀結構，採用樹狀分岔的架構從中歸納出分類規則，並利用樣本進行預測。決策樹模型透過不斷地劃分資料，使依賴變數的差別最大，最終目的是將資料分類到不同的組織或不同的分枝，在依賴變數的值上建立最強的歸類。
支援向量機	主要是利用超平面機器學習法來分隔兩個或多個不同類別的資料， 處理資料探勘中分類的問題。其主要理論是建立在統計學習理論的 VC 維度理論與結構化風險最小誤差法的基礎上，能較好地解決小樣本、非線性、高維度和局部及小點等實際問題。

系统实例操作

- 资料准备
- 智能型税务选案系统介绍
- 抽样比例与模型比较
- 建置模型
- 模型评估
- 及验证

• 上传档案之页面

系统实例操作

- 资料准备
- 智能型税务选案系统介绍
- 抽样比例与模型比较
- 建置模型
- 模型评估
- 及验证

• 检视上传资料

No.	x1	x2	x3	x4	x5	x6	Y	x9	x23
1	498540	310199	498540	4187	15510	15510	0	0.62	3672
2	1743059	1418945	1743059	1637	80141	70950	0	0.81	7994
3	1743059	1418945	1743059	1637	80141	70950	0	0.81	25280
4	302309	76516	317825	999	15115	3825	0	0.24	7303
5	718201	609576	718201	2234	30479	30479	0	0.85	19631
6	177714	35484	177714	6484	8886	1775	0	0.2	9561
7	1948432	1877670	1948432	29720	94354	93883	0	0.96	3673
8	1948432	1877670	1948432	29720	94354	93883	0	0.96	22521
9	1008212	1855753	1008212	19949	50411	92788	0	1.84	5223
10	2725039	3305071	2725039	37791	136252	165255	0	1.21	2661
11	1144894	1044595	1144894	365	52231	52231	0	0.91	19592
12	133427	24638	133427	1661	6671	1234	0	0.18	14221
13	133427	24638	133427	1661	6671	1234	0	0.18	35431
14	2425226	2473484	2515226	18859	119176	123676	0	0.98	3281
15	9061156	8223700	9333546	76	394997	404913	0	0.88	50364
16	145950	138357	145950	754	7298	6919	0	0.95	7955
17	145950	138357	145950	754	7298	6919	0	0.95	37330
18	152490	119708	152490	1343	5985	5985	0	0.79	5040
19	408518	207160	440326	1538	8770	10360	0	0.47	4375
20	408518	207160	440326	1538	8770	10360	0	0.47	15341
21	297311	189609	297311	1996	9028	9028	0	0.64	28113
22	2025375	1642309	2025375	3236	81490	81490	0	0.81	14751
23	731214	539907	731214	3367	26997	26997	0	0.74	4443
24	2393381	1569818	2393381	3536	78491	78491	0	0.66	2182

系统实例操作

- 资料准备
- 智能型税务选案系统介绍
- 抽样比例与模型比较
- 建置模型
- 模型评估
- 及验证

- 选择欲分析之建模方法

- 资料准备
- 智能型税务选案系统介绍
- 抽样比例与模型比较
- 建置模型
- 模型评估
- 及验证

系统实例操作

选择欲分析之应变量与自变量

系统实例操作

- 资料准备
- 智能型税务选案系统介绍
- 抽样比例与模型比较
- 建置模型
- 模型评估
- 及验证

- 选择测试样本占全部样本的比例

系统实例操作

- 资料准备
- 智能型税务选案系统介绍
- 抽样比例与模型比较
- 建置模型
- 模型评估
- 及验证

• 分析结果

系统实例操作

- 资料准备
- 智能型税务选案系统介绍
- 抽样比例与模型比较
- 建置模型
- 模型评估
- 及验证

• 模型评估

系统实例操作

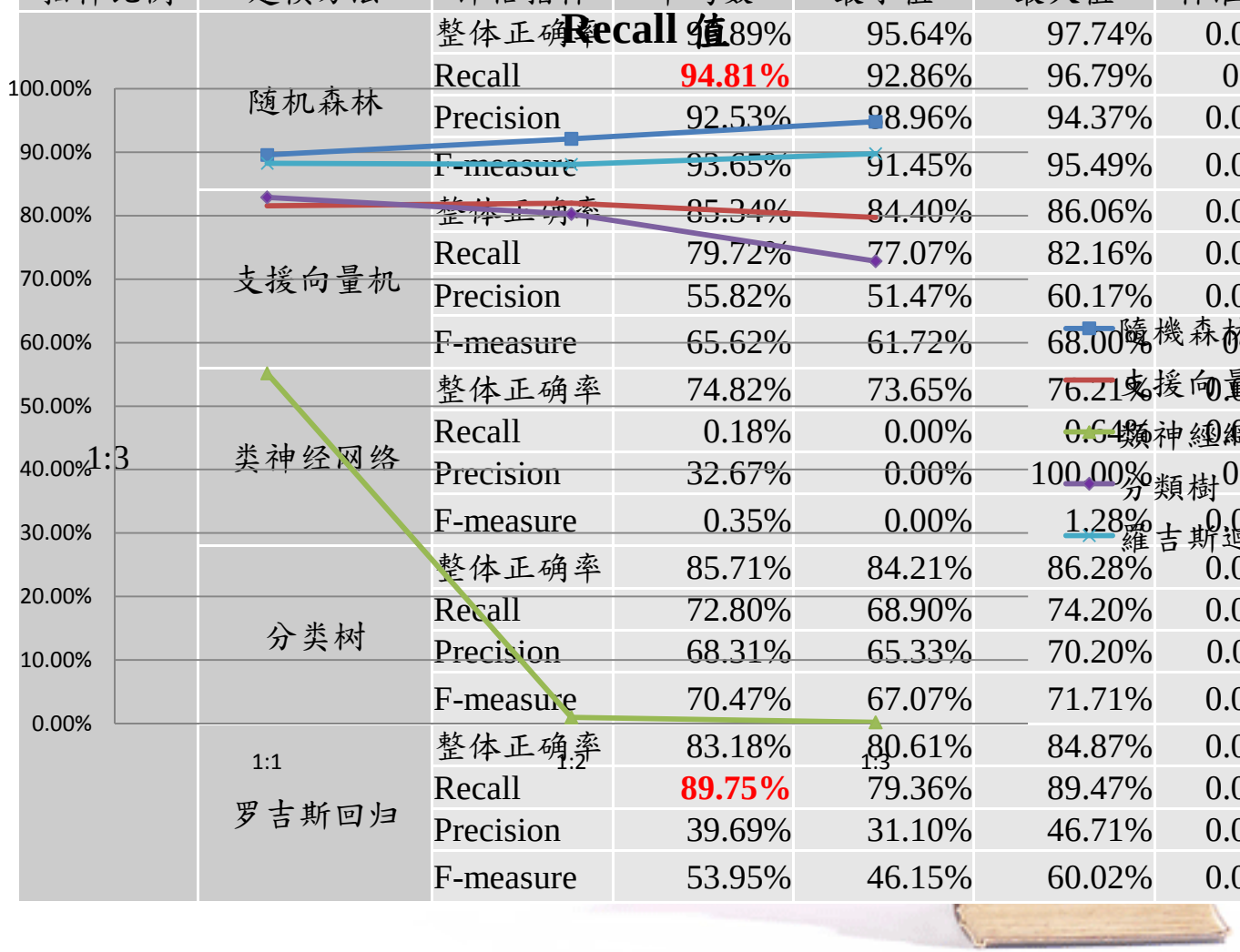
- 资料准备
- 智能型税务选案系统介绍
- 抽样比例与模型比较
- 建置模型
- 模型评估
- 及验证

• 罗吉斯回归预测模型

抽样比例与模型比较

- 资料准备
- 智能型税务选案系统介绍
- 抽样比例与模型比较
- 建置模型
- 模型评估
- 及验证

抽样比例	建模方法	评估指标	平均数	最小值	最大值	标准偏差
1:3	随机森林	整体正确率	96.89%	95.64%	97.74%	0.000022
		Recall	**94.81%**	92.86%	96.79%	0.00005
		Precision	92.53%	88.96%	94.37%	0.000158
		F-measure	93.65%	91.45%	95.49%	0.000085
	支援向量机	整体正确率	85.34%	84.40%	86.06%	0.000018
		Recall	79.72%	77.07%	82.16%	0.000133
		Precision	55.82%	51.47%	60.17%	0.000397
		F-measure	65.62%	61.72%	68.00%	0.00023
	类神经网络	整体正确率	74.82%	73.65%	76.21%	0.000028
		Recall	0.18%	0.00%	0.54%	0.000003
		Precision	32.67%	0.00%	100.00%	0.09652
		F-measure	0.35%	0.00%	1.28%	0.000012
1:2	分类树	整体正确率	85.71%	84.21%	86.28%	0.000022
		Recall	72.80%	68.90%	74.20%	0.000148
		Precision	68.31%	65.33%	70.20%	0.000112
		F-measure	70.47%	67.07%	71.71%	0.000102
1:1	罗吉斯回归	整体正确率	83.18%	80.61%	84.87%	0.000087
		Recall	**89.75%**	79.36%	89.47%	0.000496
		Precision	39.69%	31.10%	46.71%	0.001618
		F-measure	53.95%	46.15%	60.02%	0.001217



结论

只须连上网络

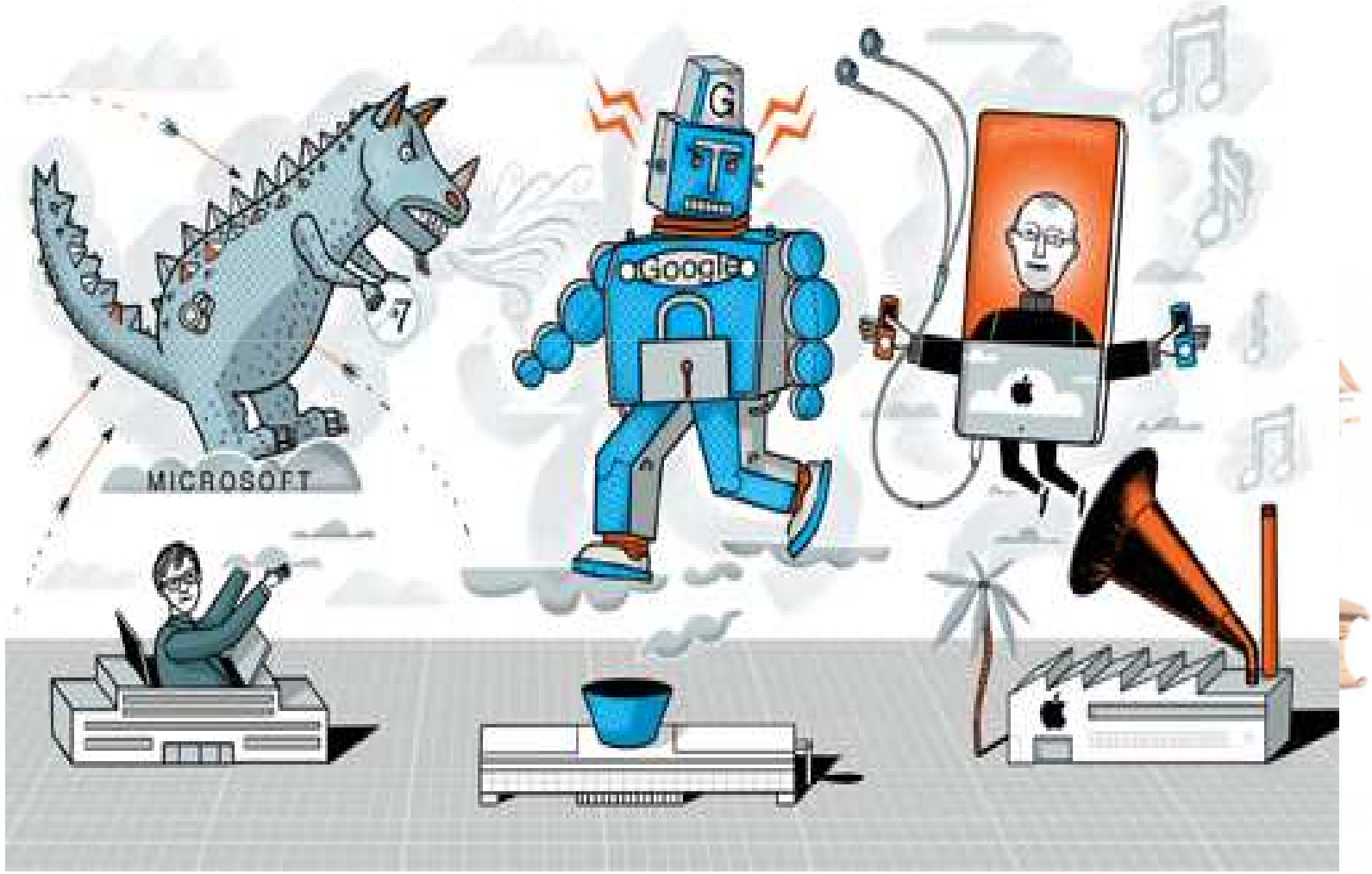
智能型税务选案系统

界面操作简单

- 从2006年3月，亚马逊(Amazon.com)首开市场先例，推出云端服务；来年，Google正式提出云端运算一词之后，「愈来愈多企业紧追云的踪迹，无不希望飞上云端，」前中国网通执行长，现为中国宽带产业基金董事长田溯宁观察。

- 最新一期的《经济学人》（The Economist）以「云端上的战争」为题，点出Google、微软(Microsoft)、苹果（Apple）三巨头在争抢云端运算商机的激烈程度。以目前的态势来看，以搜索引擎起家的网络巨擘Google居于领先地位，但过去沉溺于垄断个人计算机（PC）作业平台甜头的微软，也意识到这样的改变，正投入大量的资源准备急起直追。

云端大战



云计算技术给数据仓库的高效计算带来一次革命性的影响

- 数据仓库是数据挖掘得以发展的基础，也是商业智能的支撑，同时数据挖掘也是商业智能的重要环节，由此可见数据仓库对于商业智慧来言具有很重要的作用。它集成了企业的最核心的数据，随着企业对数据的再次利用和深入挖掘，海量数据的高效计算问题成为企业最为关注的一个问题之一。

- 政府端〉欧巴马用它建新政府网站

云端运算的火红程度，就连美国总统欧巴马（Barack Obama）也要参一脚。今年九月，白宫信息长坎卓（Vivek Kundra）宣布，运用云端科技开发出来的新政府网站Apps.Gov.，将取代旧有的IT信息系统。据《纽约时报》（New York Times）报导，Apps.Gov预估将替美国政府每年省下七百五十亿美元（约合新台币二兆四千亿元）的支出。

生活在云端

- 中华电信与微软签订云端运算策略联盟合作备忘录，将针对客户端设备软件应用服务和云端服务等进行合作，以新的营运模式携手开创云端服务商机。此次策略联盟，中华电信将运用微软的先进平台与技术，提供企业客户与消费者最便利的行动加值与云端服务，让企业客户提升工作效率，并带给消费者便利与智能的数字生活体验。

生活在云端

- 代表着双方在未来的『云端服务』与计算机、手机与电视整合加值服务上将迈向另一里程，相信藉由微软云端平台技术与顾问咨询以及中华电信本身的技术资源与服务平台的整合与分享，很快的就可以让消费者与企业用户体验到任何时间、任何地点、使用任何连网的用户终端设备，即可快速取得所需信息与服务，享受便利的数字生活。」

工作在云端

- 台积电就开始打造内部的云端架构，将公司的DT改成精简计算机（thin client），不但降低采购成本，同时也提升数据的安全性，降低员工外泄与商业间谍的渗透机率。

科技「疯」云，再辉煌十年

- 「云端运算(Cloud Computing)即将引爆商业革命，改写游戏规则，」2009年6月，美国《BusinessWeek》如此写道。「云端运算让企业节流，也可以变得有创意，」2009年7月，美国《哈佛商业评论》制作后风暴时代首部曲专题，撰文分析。「云端运算将是一朵长长的雨云(nimbus)，让企业更灵活，」从去年10月以来，英国《经济学人》也陆续有着相关报导。近期以来，关于云端运算议题不断发烧，不仅全球重要媒体关注，各国企业更是积极投入。

1 The accelerating pace of change...

Agricultural Revolution ← 8,000 years → Industrial Revolution ← 120 years → Light-bulb ← 90 years → Moon landing

← 22 years → World Wide Web ← 9 years → Human genome sequenced

2 ...and exponential growth in computing power...

Computer technology, shown here climbing dramatically by powers of 10, is now progressing more each hour than it did in its entire first 90 years

COMPUTER RANKINGS
By calculations per second per $1,000

Analytical engine
Never fully built, Charles Babbage's invention was designed to solve computational and logical problems

Colossus
The electronic computer, with 1,500 vacuum tubes, helped the British crack German codes during WW II

UNIVAC I
The first commercially marketed computer, used to tabulate the U.S. Census, occupied 27 cu m

Apple II
At a price of $1,298, the compact machine was one of the first massively popular personal computers

Power Mac G4
The first personal computer to deliver more than 1 billion floating-point operations per second

3 ...will lead to the Singularity

2045
Surpasses brainpower equivalent to that of all human brains combined

Surpasses brainpower of human in 2023

Surpasses brainpower of mouse in 2015

10^{26}
10^{20}
10^{15}
10,000,000,000
100,000
1
0.00001

Hollerith Tabulator
IBM Tabulator
National Ellis 3000
Zuse 2
Zuse 3
ENIAC
BINAC
IBM SSEC
EDVAC
Whirlwind
Datamatic 1000
IBM 1620
DEC PDP-10
DEC PDP-4
IBM 1130
Data General Nova
Intellec-8
IBM PC
Compaq Deskpro 386
Pentium PC
Pentium II PC
Dell Dimension 8400
Mac Pro
Nvidia Tesla GPU & PC

ELECTROMECHANICAL — RELAYS — VACUUM TUBES — TRANSISTORS — INTEGRATED CIRCUITS

1900 1920 1940 1960 1980 2000 2011 2020 2045

on, there's no reason to think computers

Probably. It's impossible to predict the

idea; it's a serious hypothesis about the

he called an "intelligence explosion":

站上云端，扭转趋势

- 未来如智能型手机（Smart phone）、卫星导航（GPS）等行动装置，都可以透过云端运算，发展出更多的应用服务。进一步的云端运算，更可应用在生物科学，例如：分析基因结构、基因图谱定序、解析癌症细胞等。利用云端运算架构协助，效率快又准确。也就是当大量信息处理不再昂贵的时候，许多科技就会百花齐放，应运而生。

云端运算-待解决的问题

1. 安全问题

自从“云计算”的概念提出以来，关于其资安全性的质疑就一直不曾平息，这里的安全性主要包括两个方面一是自己的信息不会被泄露从而给自己造成不必要的损失，二是自己在需要时能够保证准确无误地获取这些信息。

2009年3月，世界著名的谷歌公司不得不尴尬地承认了不小心泄露客户私人信息的事实，这也使得人们对于其提供的云计算服务器不得不重新进行审视安全问题

云端运算-待解决的问题

- 2.云运算系统故障问题（Sidekick服务中断、Amazon EC2遭到阻断服务攻击，以及Google电子邮件服务中断）
- 3.推广及应用层面之扩大
- 4.产生之专业伦理及道德方面的问题(法律层面)

云端未来

漫步云端，任重而道远！