大數據的解決方案及其應用

雲計算在統計及Data Mining研究
應用及前沿綜述
雲計算下的R前沿探討－兼談 數據分析的未來

謝邦昌

輔仁大學商學研究所博士班 所長
統計資訊學系教授
中華資料採礦協會　榮譽理事長

2014年10月02日

025674@mails.fju.edu.tw
WWW.CDMS.ORG.TW

刁院士：*統計與數學要相互欣賞*

統計要跟各個領域做朋友

統計趨勢Statistics Trend趨勢統計Trend Statistics

數據分析科學的過去、現在、未來

--統計是數據科學

Data Science

Dataology

Science

-- (Dealing with DATA) 11 FEBRUARY 2011 VOL 331 SCIENCE www.sciencemag.org

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

BRUNEI B$ 9.00 HONG KONG HK$ 55.00 KOREA W7,500.00 MONGOLIA US$ 5.00 PHILIPPINES P 130.00 THAILAND B160.00
CAMBODIA US$ 5.00 INDIA ₹ 110.00 MACAU HK$ 55.00 MYANMAR US$ 5.00 SINGAPORE S$9.00 U.S. ARMED FORCES US$ 5.00
CHINA RMB 50.00 INDONESIA Rp 35,000 MALAYSIA Rm 10.00 NEPAL N Rs 200.00 SRI LANKA Rs 290.00 VIETNAM US$5.00

9 771064 030005 07

www.time.com

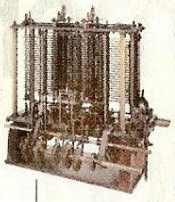
1 The accelerating pace of change...

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

Hollerith Tabulator
IBM Tabulator
National Ellis 3000
Zuse 2
Zuse 3
ENIAC
IBM SSEC
BINAC
Whirlwind
EDVAC
DEC PDP-4
IBM 1130
IBM 1620
Datamatic 1000
DEC PDP-10
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

0.00001 I 100,000 10,000,000,000

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

10^{26} **2045** Surpasses brainpower equivalent to that of all human brains combined

10^{20} Surpasses brainpower of human in 2023

10^{15} Surpasses brainpower of mouse in 2015

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":

「電腦vs.人腦」益智遊戲大賽經過三天激戰後，IBM的超級電腦華生（Watson，圖中）最終擊敗人類，獲頒100萬美元獎金。IBM將把這筆獎金捐給世界展望會等慈善機構。（美聯社）

雲端運算的演化

Super Computer

Cluster Computing

Distributed Computing

Grid Computing

Utility Computing

Cloud Computing

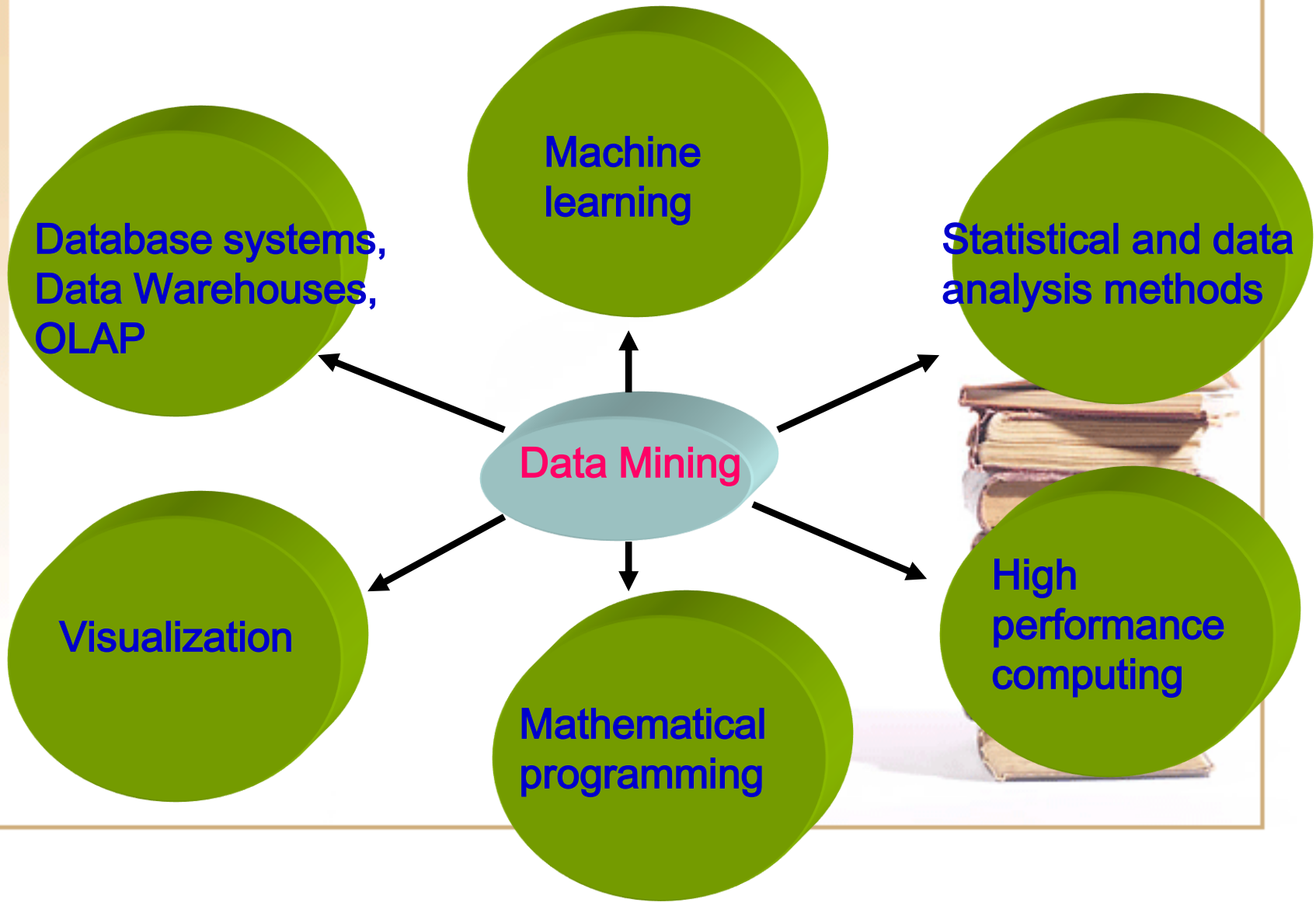
雲端運算

- 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

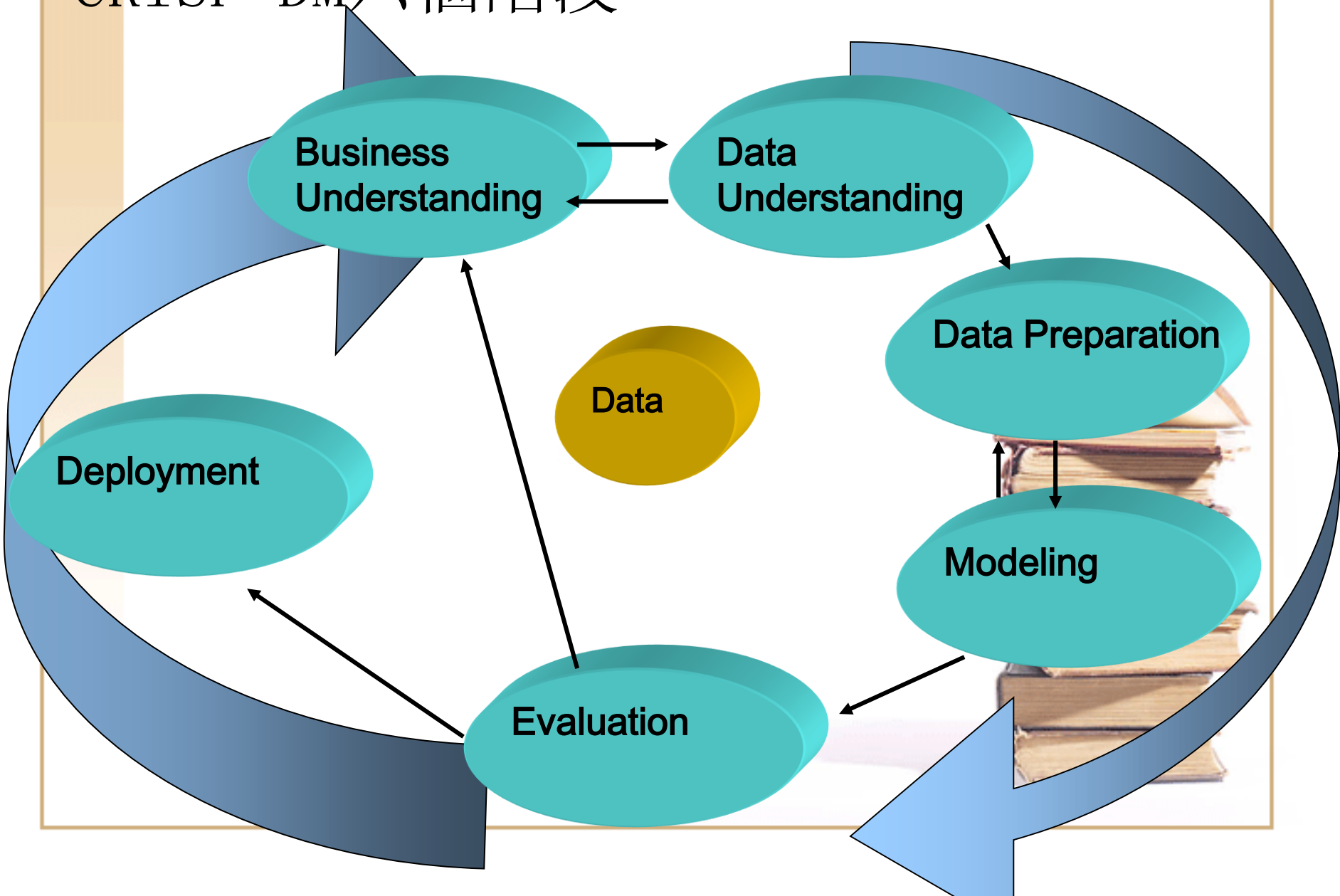
比較表

服務 / 屬性	Amazon EC2	Google App Engine	Microsoft Azure	Yahoo Hadoop
架構	Iaas/Paas	Paas	Paas	Software
服務型態	Compute/ Storage	Web application	Web and non-web	Software
管理技術	OS on Xen hypervisor	Application container	OS through Fabric controller	Map / Reduce Architecture
使用者介面	EC2 Command-line tools	Web-based Administration console	Windows Azure portal	Command line and web
APIs	yes	yes	yes	yes
收費	yes	maybe	yes	no
程式語言	AMI(Amazon Machine Image)	Python	.NET framework	Java,

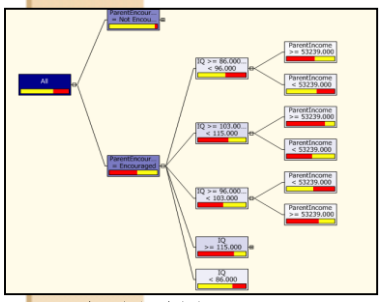
Data Mining包含六大領域



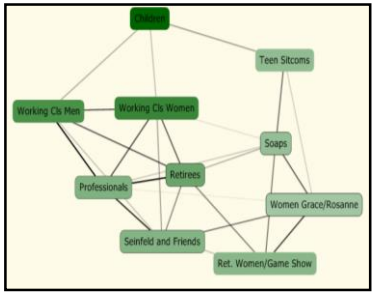
CRISP-DM六個階段



完整的演算法 *SQL Server 2008已提供 SQL Server 2014在雲端*

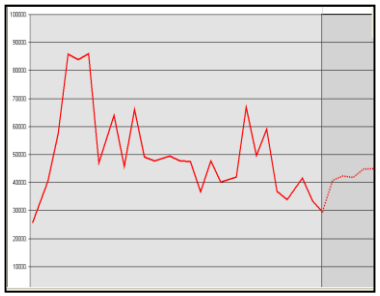


決策樹

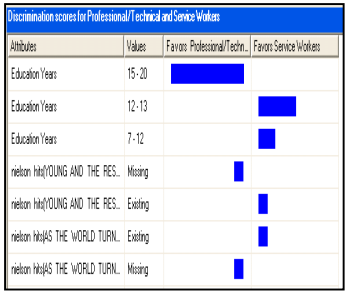


群集

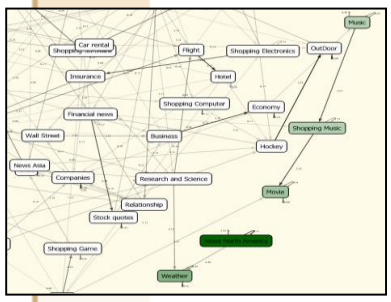
SQL Server 2000已提供



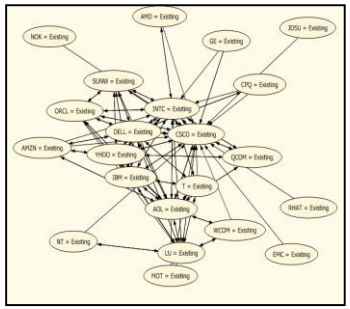
時間序列



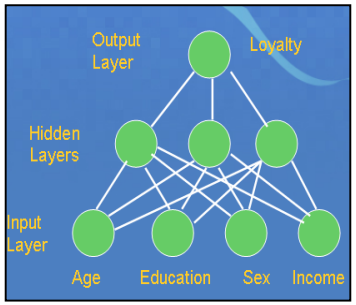
貝氏機率分類



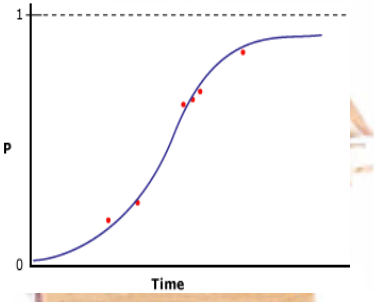
時序群集



關聯規則

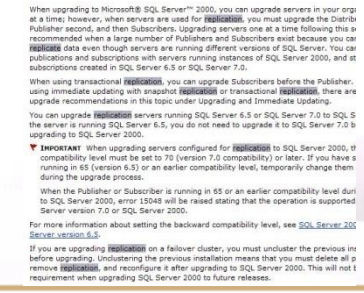


類神經網路



羅吉斯回歸

線性回歸



文字數據採礦

常用的Data Mining及統計學習方法-1

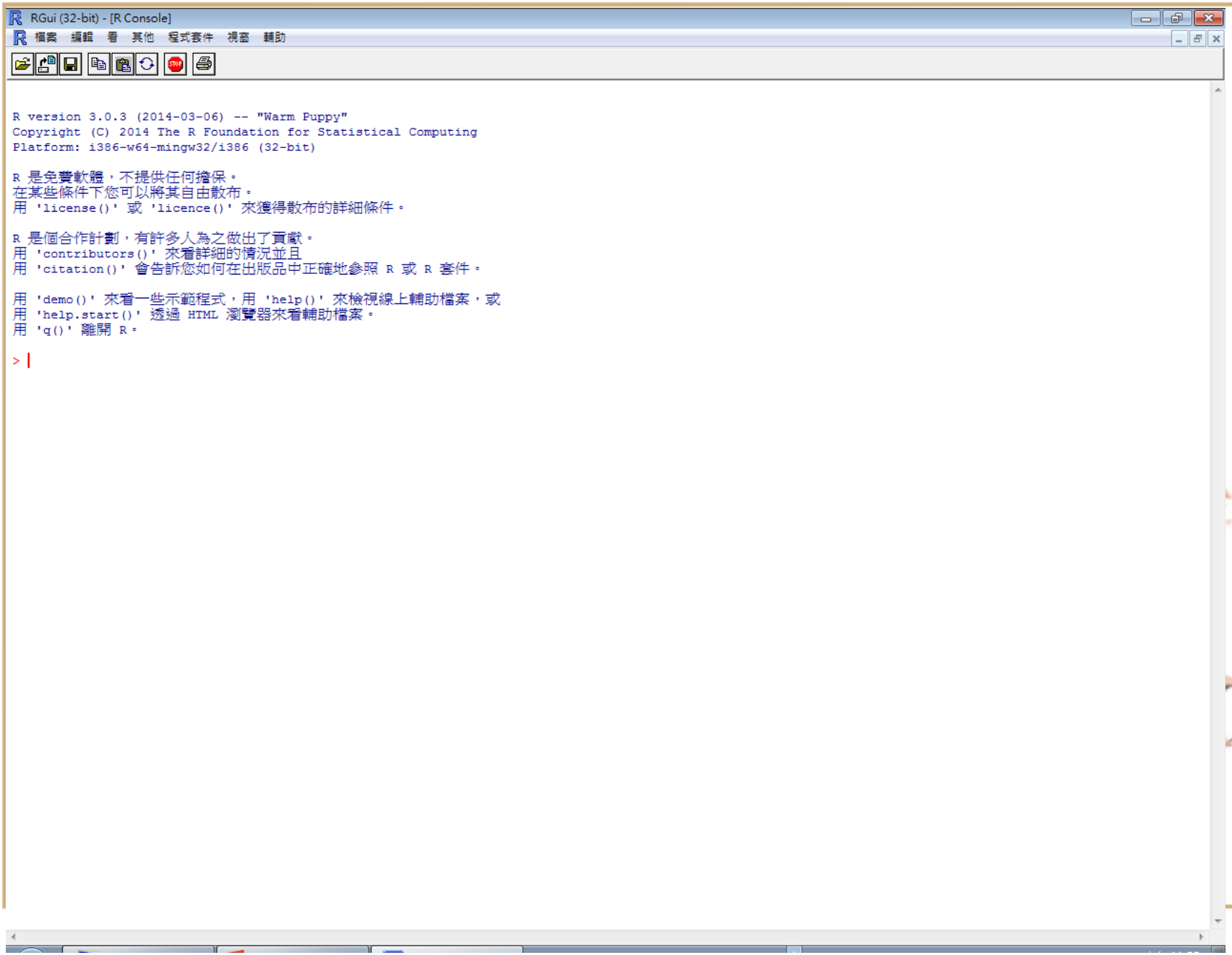
Binary Classifier (二元分類)
Numeric Predictor (數值預測)
Time Series (時間序列)
C&R TREE (分類回歸樹)
Quick Unbiased Efficient Statistical Tree (QUEST決策樹模型)
CHAID (分類樹)
Decision List (決策樹列表)
Regression (線性回歸分析)
PCA/Factor (主成分分析)
Neural Net (類神經網路)
C5.0 (決策樹)
Feature Selection (特徵選取)
Discriminant Analysis (判別分析)
Logistic (羅吉斯回歸)
Generalize Linear Model (廣義線性模型)
Cox Regression

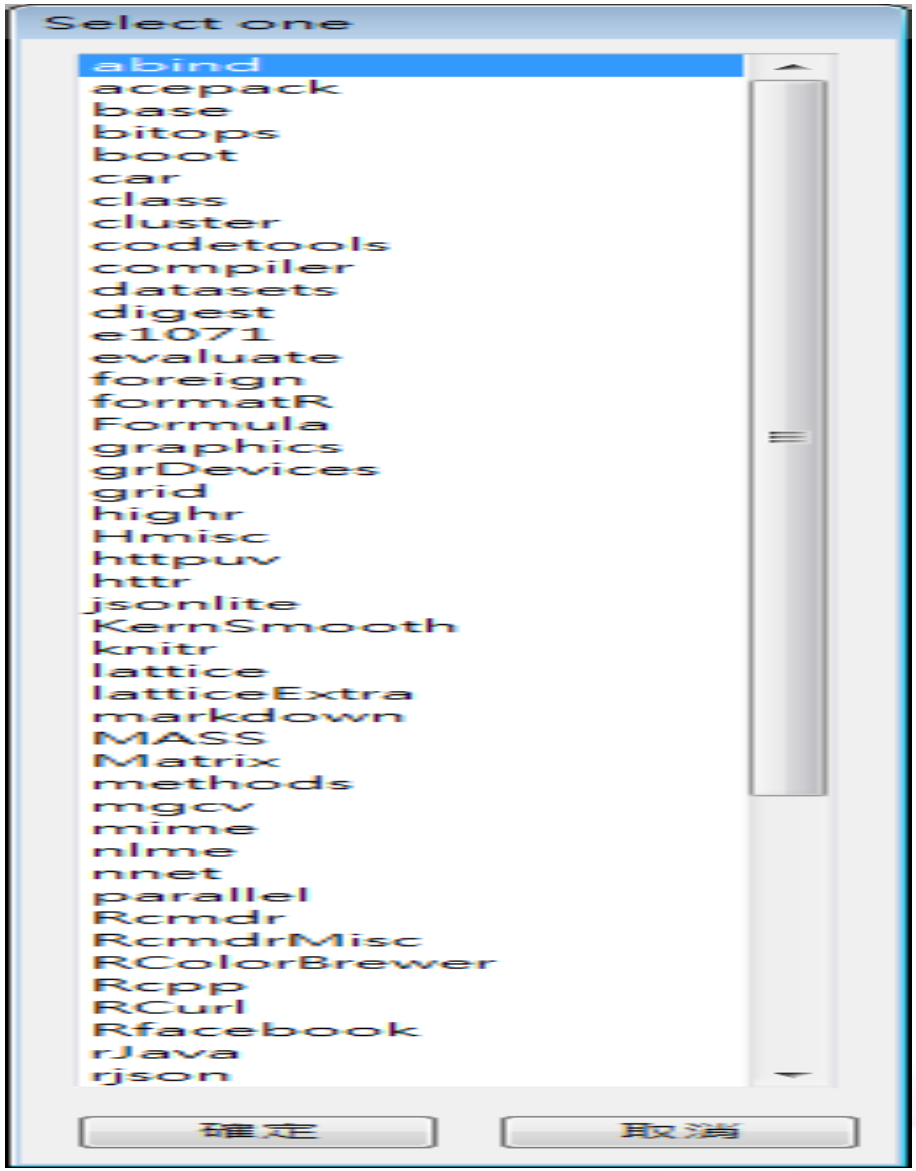
常用的Data Mining及統計學習方法-2

Support Vector Machine (SVM支持向量機)
Bayes Net (貝氏分類器)
SLRM (自我學習反應模型)
GRI關聯
Apriori關聯
CARMA關聯(連續交易)
Sequence Clusterc序列關聯
K-Means (K-Means分群)
Kohonen (自我組織化)
Two-Step (二階段)
Anomaly (異常檢測)
Random Forests (隨機森林)
ICA (獨立成分分析)
Multivariate adaptive regression spline (MARS多元適應性回歸平滑)
Pmml(預測模型標記語言)
Boosting

使用軟體 常用

SQL server 2014
SPSS 22 (PAWS) --IBM
SAS
SQL 2014+Excel (2013)-Data Mining Add-in
Clementine 12.0
Statistica 14.0
WEKA
R → Cloud R
R+Excel ADD-IN …….*還有更多雲端軟體*


RGui (32-bit) - [R Console]
檔案 編輯 看 其他 程式套件 視窗 輔助
R version 3.0.3 (2014-03-06) -- "Warm Puppy"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)
R 是免費軟體，不提供任何擔保。
在某些條件下您可以將其自由散布。
用 'license()' 或 'licence()' 來獲得散布的詳細條件。
R 是個合作計劃，有許多人為之做出了貢獻。
用 'contributors()' 來看詳細的情況並且
用 'citation()' 會告訴您如何在出版品中正確地參照 R 或 R 套件。
用 'demo()' 來看一些示範程式，用 'help()' 來檢視線上輔助檔案，或
用 'help.start()' 透過 HTML 瀏覽器來看輔助檔案。
用 'q()' 離開 R。
>


Select one
abind
acepack
base
bitops
boot
car
class
cluster
codetools
compiler
datasets
digest
e1071
evaluate
foreign
formatR
Formula
graphics
grDevices
grid
highr
Hmisc
httpuv
httr
jsonlite
KernSmooth
knitr
lattice
latticeExtra
markdown
MASS
Matrix
methods
mgcv
mime
nlme
nnet
parallel
Rcmdr
RcmdrMisc
RColorBrewer
Rcpp
RCurl
Rfacebook
rJava
rjson
確定
取消

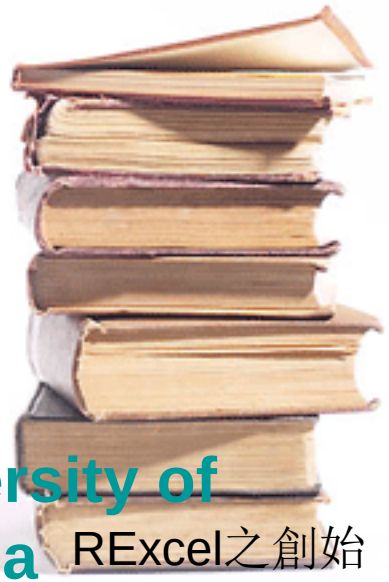
R Excel
數據採礦發展趨勢

- **RExcel**
  - **RExcel之創始**
  - RExcel之啟動
  - RExcel之應用
    - 數據導入
    - 數據分析
    - 結果保存

statconn之“幕後黑手”（The masterminds behind statconn）

- Thomas Baier (1971-)
- 在不同環境中應用R
  - R/Scilab (D)COM Server
  - RExcel (1998)

- Erich Neuwirth (1948-)
- RExcel的主要作者



University of Vienna RExcel之創始

•http://rcom.univie.ac.at/

數據採礦在各產業的應用

- **金融服務業**

  客戶貢獻度分析、信用評分、風險評估、客戶區隔、交叉行銷等。

- **保險業**

  顧客貢獻度分析、信用評分、風險評估、客戶區隔、交叉行銷、客戶流失分析和詐欺偵測等。

- **電信業**

  顧客貢獻度分析、信用評分、客戶區隔、交叉行銷、客戶流失分析、銷售預測和詐欺偵測等。

數據採礦在各產業的應用

- **製造業**

  客戶貢獻度分析、品質管制、行銷績效分析、生產分析和存貨分析等。

- **零售業**

  客戶忠誠度、客戶區隔、購物籃分析、定價分析、交叉行銷和銷售預測等。

- **生物科技、醫療保健、航太空業、環境、法律等**

商業智慧的核心

- 如何收集數據
  - 營運數據，市場調查數據，固定**Panel**追蹤
- 如何管理數據
  - ETL，Data warehousing
- 如何從數據中獲取智慧
  - Data Mining，OLAP，Statistics
- 如何應用智慧
  - 行銷策略，主管決策，互動化**CRM**機制

- 雲端運算可以實現適應端通過在線上傳數據或購買數據，通過雲數據倉庫，進行數據倉庫建模或數據抽取，線上支付使用數據採礦工具和商業智慧相關處理軟體

IIaaS是SaaS的延伸 I^2aaS

- 數據採礦和商業智慧的原理相似，均由數據提供資訊、產生知識，再由知識累積智慧。而雲端運算可以使這個過程在網際網路上得以實現。也就是說雲端運算可以提供基於SaaS的知識與智慧分析的服務（Information&Intelligence as a Service），簡稱IIaaS；I^2aaS，它是SaaS的延伸。

雲端運算產業類型

IIaaS I^2aaS

Information & intelligence as a Service

SaaS

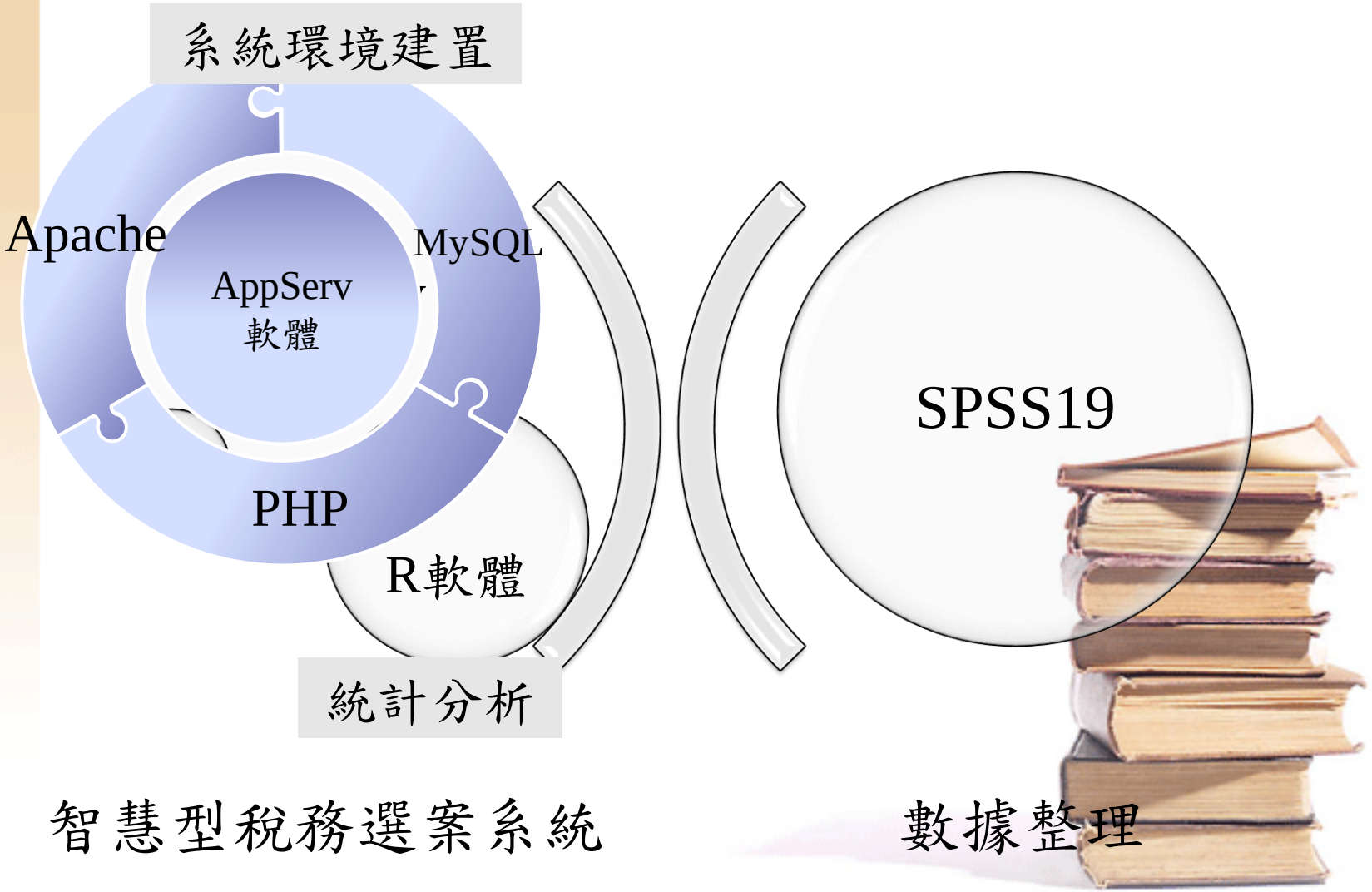
Software as a Service

PaaS

Platform as a Service

IaaS

Infrastructure as a Service


系統環境建置
Apache
AppServ
軟體
MySQL
PHP
R軟體
統計分析
SPSS19
智慧型稅務選案系統
數據整理

科技「瘋」雲，再輝煌十年

- 「雲端運算(Cloud Computing)即將引爆商業革命，改寫遊戲規則，」2009年6月，美國《BusinessWeek》如此寫道。「雲端運算讓企業節流，也可以變得有創意，」2009年7月，美國《哈佛商業評論》製作後風暴時代首部曲專題，撰文分析。「雲端運算將是一朵長長的雨雲(nimbus)，讓企業更靈活，」從去年10月以來，英國《經濟學人》也陸續有著相關報導。近期以來，關於雲端運算議題不斷發燒，不僅全球重要媒體關注，各國企業更是積極投入。

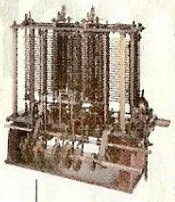
1 The accelerating pace of change...

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

Hollerith Tabulator
IBM Tabulator
National Ellis 3000
Zuse 2
Zuse 3
ENIAC
IBM SSEC
BINAC
Whirlwind
EDVAC
DEC PDP-4
IBM 1130
IBM 1620
Datamatic 1000
DEC PDP-10
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

3 ...will lead to the Singularity

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

10^{26} — **2045** Surpasses brainpower equivalent to that of all human brains combined

10^{20} — Surpasses brainpower of human in 2023

10^{15}

10^{12}

10,000,000,000

100,000

1

0.00001

Surpasses brainpower of mouse in 2015

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":

漫步雲端，任重而道遠！