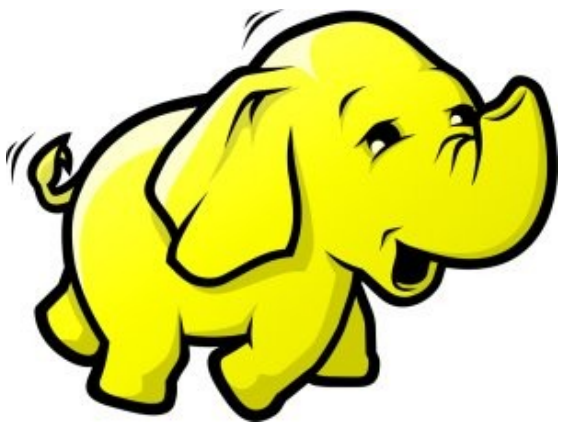
巨量資料的奇幻漂流

Life of Big Data

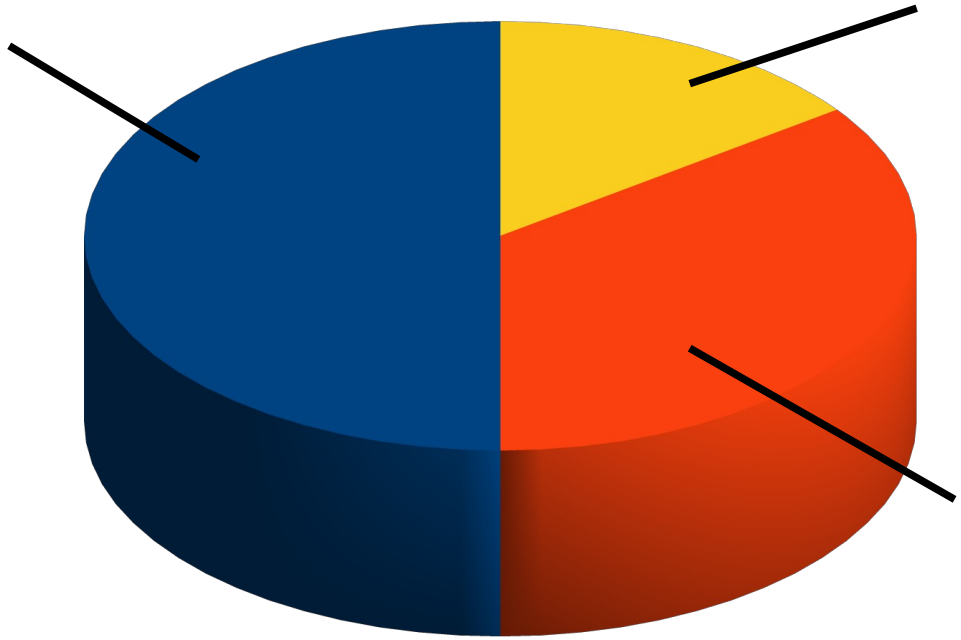
Jazz Wang
Yao-Tsung Wang
jazz@nchc.org.tw



WHO AM I？這傢伙是誰啊？JAZZ？

- 講者介紹：
  - 國網中心　王耀聰　副研究員 / 交大電控八九級碩士
  - jazz@nchc.org.tw
- 所有投影片、參考資料與操作步驟均在網路上
  - http://trac.nchc.org.tw/cloud
  - 由於雲端資訊變動太快，愛護地球，請減少不必要之列印。

FOSS 使用者
Debian/Ubutnu
Access Grid
Motion/VLC
Red5
Debian Router
DRBL/Clonezilla
Hadoop



行動力薄弱的開發者
TRTC WSU/
Haduzilla /
Hadop4Win / Ezilla

推廣者
DRBL/Clonezilla
Partclone/Tuxboot
Hadoop Ecosystem

【資料？ Data ？】

各位數位原住民

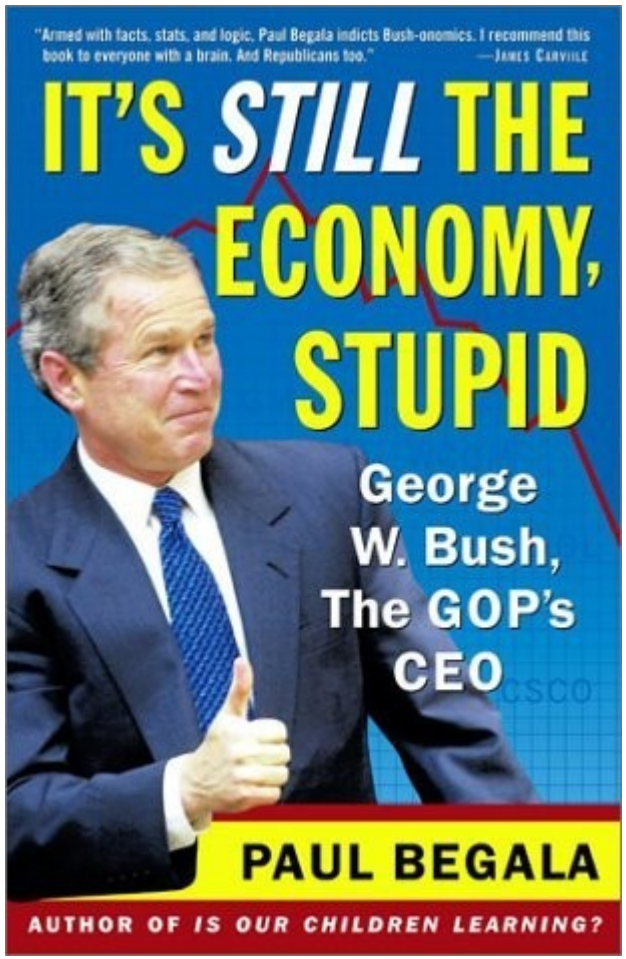
請不要忽視資料的重要性

資料會是下個世紀的黑金

淘金前，先知道資料在哪

接著學習用哪種工具採礦

誰掌握了你的資料，就有機會掌握你的荷包



「笨蛋！重點在經濟」

（**"It's the economy, stupid"**）

卡維爾（**James Carville**）自創這句標語，

促使柯林頓當上美國第 **42** 屆總統。

- 1992 年

「笨蛋！重點還是在經濟」

（**"It's STILL the economy, stupid"**）

卻讓小布希嘲笑是幼稚的總統。

- 2002 年

雲端時代，○○會說：「笨蛋！重點在資料」

（**"It's the data, stupid"**）

誰掌握了你的資料，就有機會掌握你的荷包

想想看，電腦、手機掉了，您心疼的是甚麼呢？

- 2007 年

【問題】

如果○○把您的信件、照片通通搞丟了，該賠償您多少錢呢？用 MB 計算嘛？

Agenda 演講大綱

- From WSN to IoT? 物聯網的崛起
- What is Smart City ? 淺談智慧城市
- What is Open Data ? 何謂開放資料
- What is Big Data ? 何謂海量資料
- Join Communities ! 加入國內社群

資料從哪裡來？往哪裡去？

如何解開資料流的限制？如何處理龐大的資料呢？

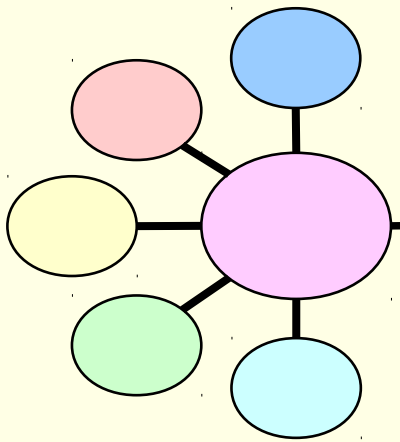
This is a story about Data Flow

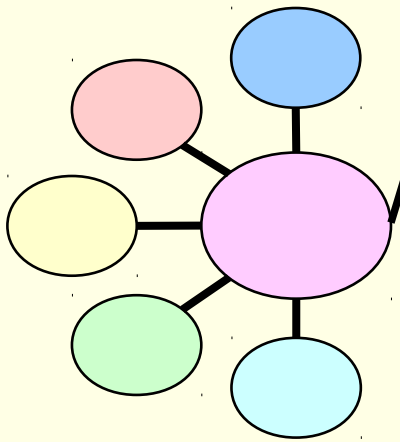
今天要講一個關於資料的漫長旅程

巨量資料的奇幻漂流 Life of Big Data

Internet of Things 物聯網

Sensor Network





Smart Grid

雲 資料中心 提供服務

Public Data Hub
Data as a Service

開放資料
Open Data

Web 2.0

Query

Map Reduce

Storage

Big Data

端 各類裝置 存取服務

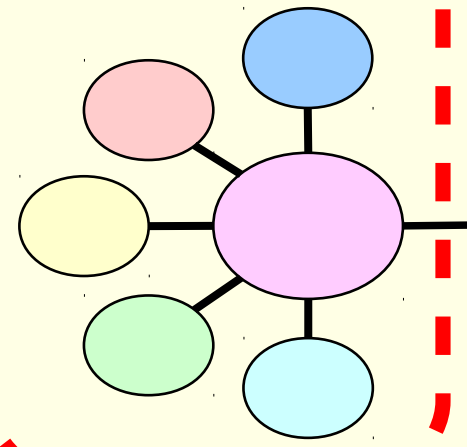
Mobile Computing

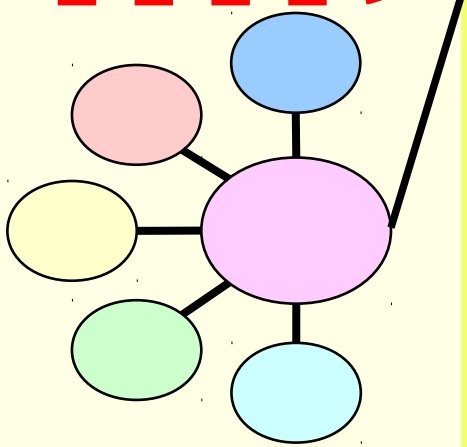
Cloud Computing 雲端運算

巨量資料的奇幻漂流 Life of Big Data

Internet of Things 物聯網

Sensor Network





Smart Grid

雲 資料中心 提供服務

Public Data Hub
Data as a Service
開放資料
Open Data

Web 2.0

Query

Map Reduce

Storage

Big Data

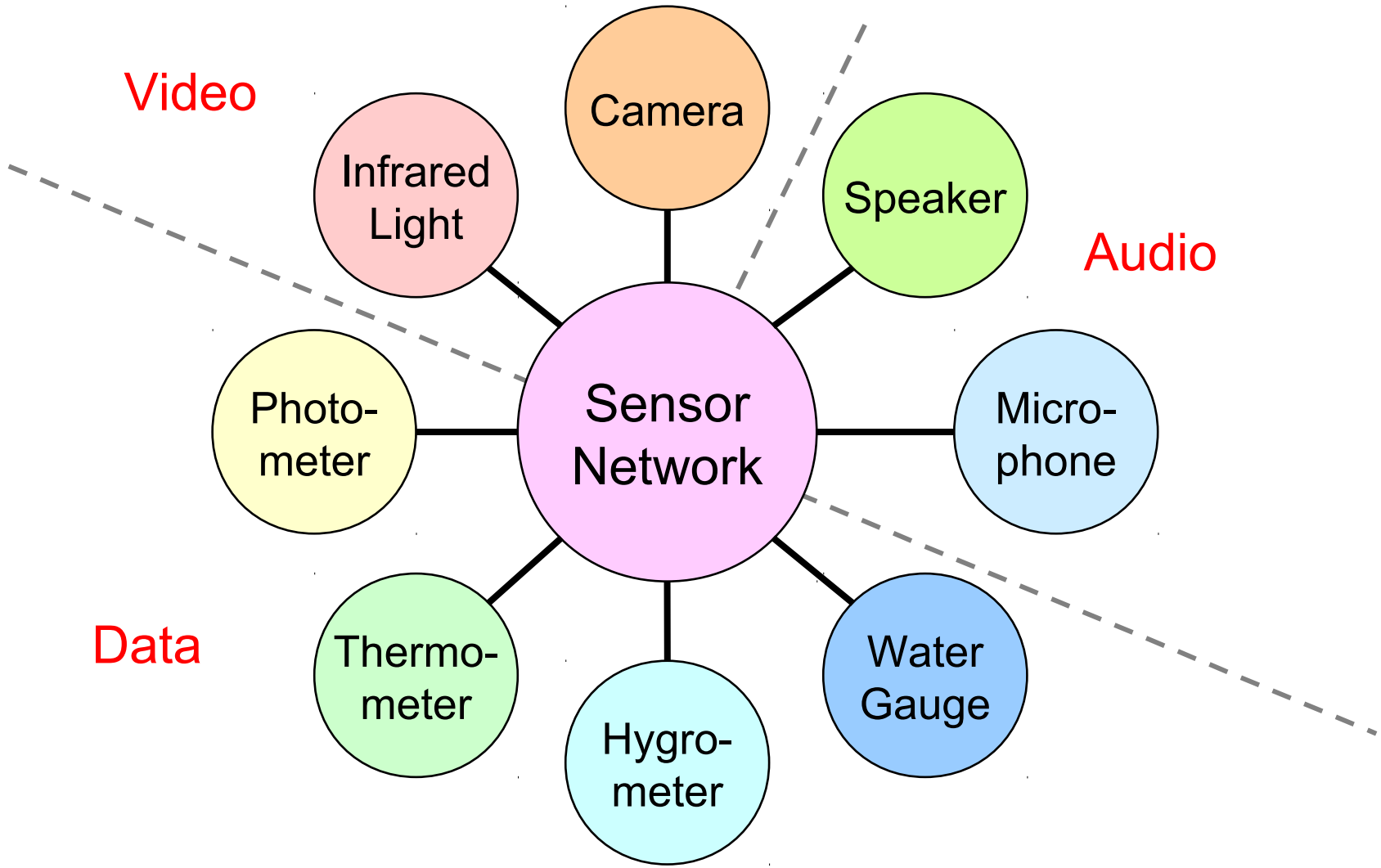
端 各類裝置 存取服務

Mobile Computing

Cloud Computing 雲端運算

無線感測器網路 (Wireless Sensor Network)

感測器網路泛指由各式各樣不同的感測器所組成，透過網路鏈結成一大型的資訊蒐集網。



生態觀測格網 (Ecology Grid)

NCHC

KNG Knowledge Innovation National Grid
ECOGRID

Ecogrid Insight | Events | Ecosites | Ecogrid on Media | Data Query |

Ecogrid project is a joint effort among National Center for High-Performance Computing (NCHC), Taiwan Ecological Research Network (TERN) and Taiwan Forestry Research Institute (TFRI).

Using state-of-the-art grid technology, scientists at NCHC established a cyberinfrastructure for ecological research, the Ecogrid, which integrates geographically distributed sensors, computing power, and storage resources into a uniform and secure platform. Ecologists can conduct researches on this platform, including data acquisition, data analysis, and data sharing. The real time intervention capability of the Ecogrid system allows researchers to carry out interactive/control experiments that have been impossible until now.

Global Lake Observatories Network

With the support from PRAGMA, Ecogrid is now extended to international. By the cooperation with scientists from Center for Limnology of University of Wisconsin, San Diego Supercomputing Center (SDSC), and Center for Biodiversity of Academia Sinica, we deployed a lake observatories network connected LTER Trout Lake station in Wisconsin and TERN Yuan Yang Lake Station in Taiwan in April 2004. Through the ubiquitous web interface, scientists can easily access and compare data from both research stations in near real-time. The pioneer network was used as an example to call for broader partnership when scientists gathered at Scripp Institute of Oceanography in March, 2005 to discuss next move upon the global network.

Underwater Reef Monitoring

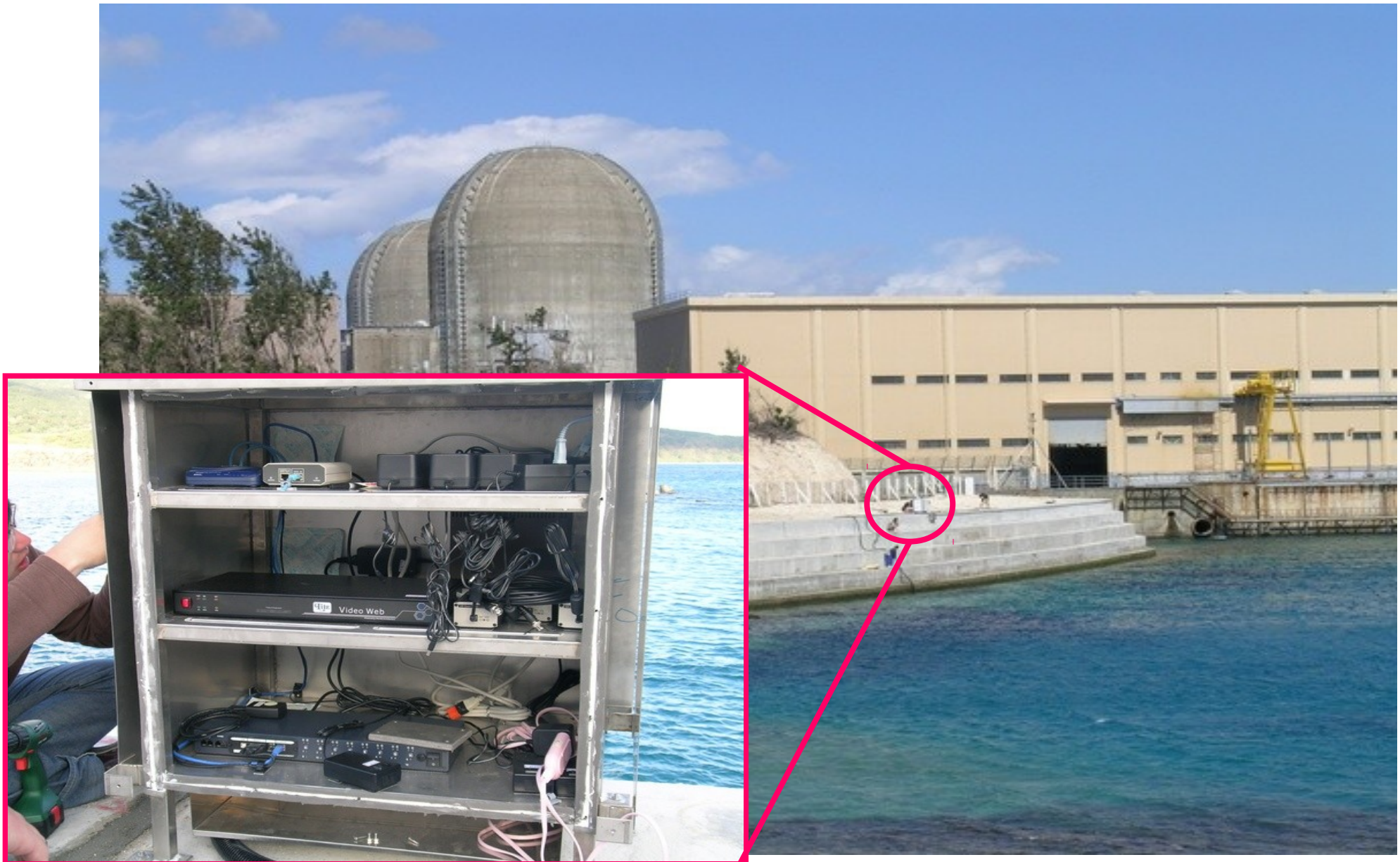
Coral reefs are precious because they are the most biologically diverse and productive marine ecosystems on earth. In addition, revenue from tourists attracted to the beauty of coral reefs can be a significant source of income for human communities in these areas. Because of the important ecological and economic roles coral reef communities fulfill, resources are pulling in to study the stresses and dangers to the reefs.

With the funding from Taipower, NCHC Ecogrid team worked together with Research for Biodiversity Center, Academic Sinica implemented an underwater reef monitoring system. The system composed with 9 cameras, located at 3 sites, lively streaming bioactivities undergoing in reef communities to marine scientists' desktop.

http://ecogrid.nchc.org.tw

Underwater Camera

Site A



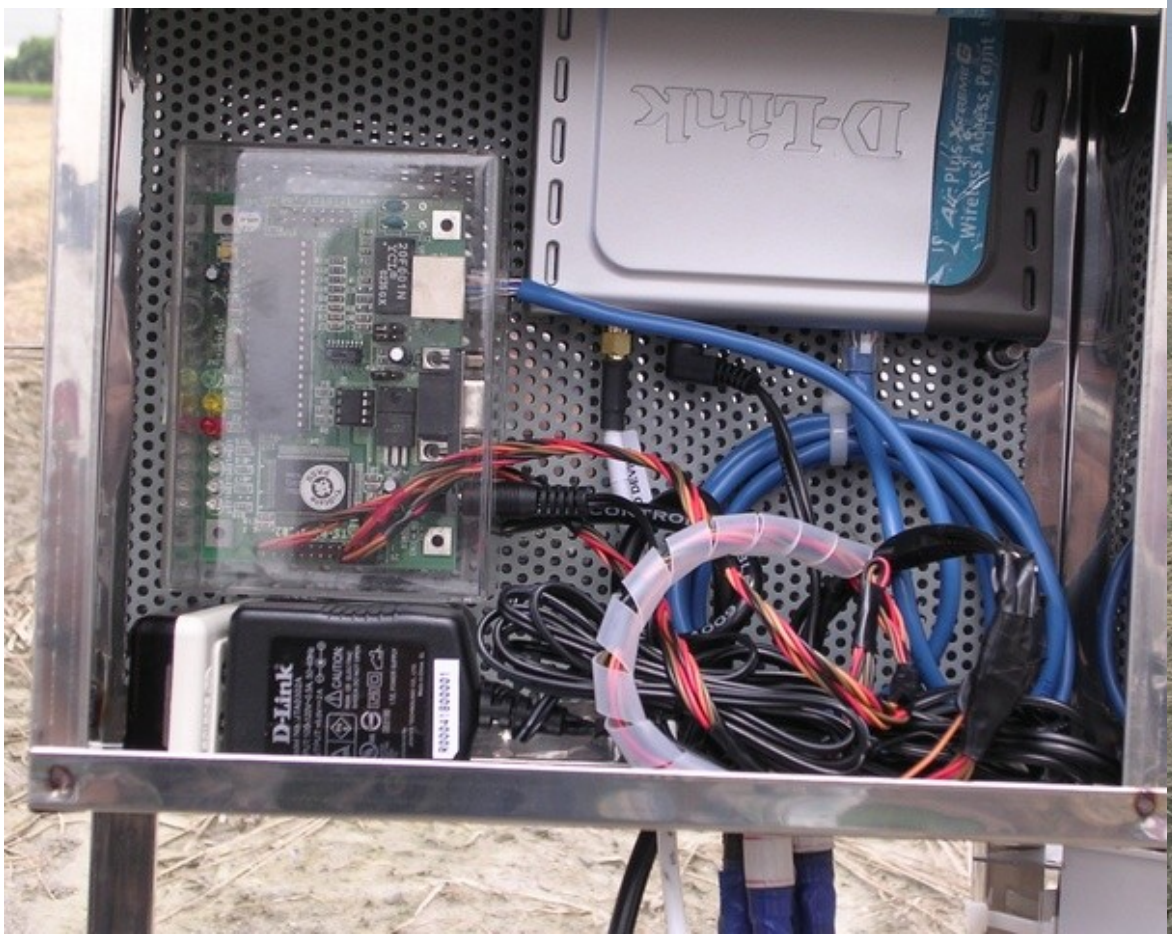
Site B & C

農業格網 (Agriculture Grid)

http://sensor.nchc.org.tw/tounan

Electric Wire Post

Passive Field Server

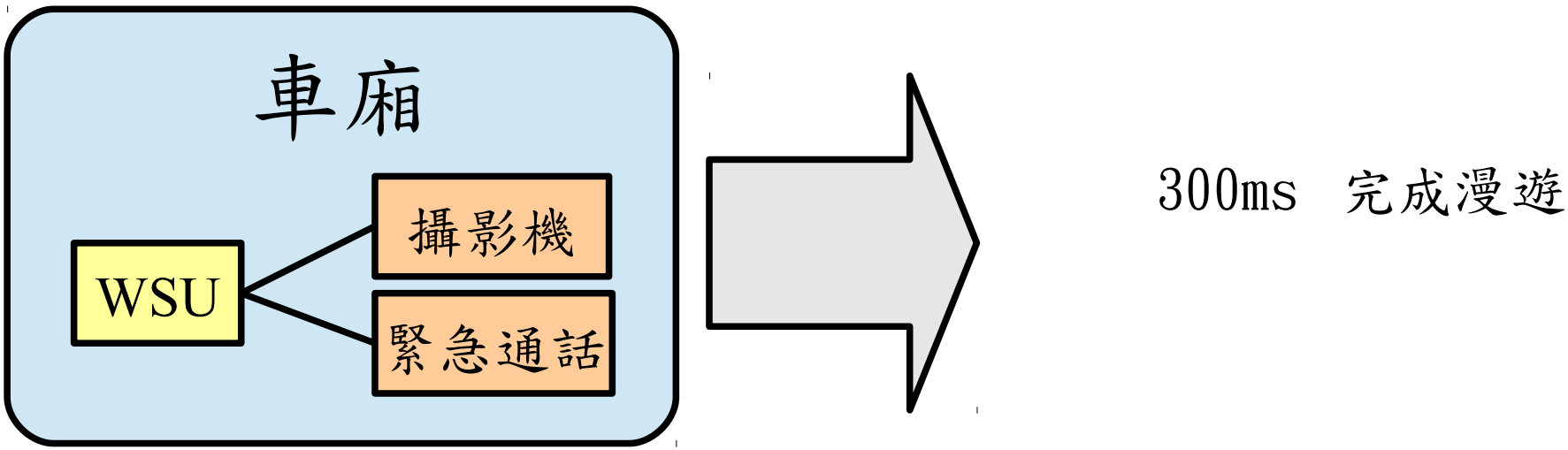


文湖線無線管理單元 (TRTC Neihu WSU)

http://www.flickr.com/photos/rail02000/5292653579/lightbox/

Wireless Fast Roaming

GPRS / 3G 太貴，得自建 WiFi
間隔 400 公尺，時速 60 公里，
如何快速地在無線基地台切換呢？

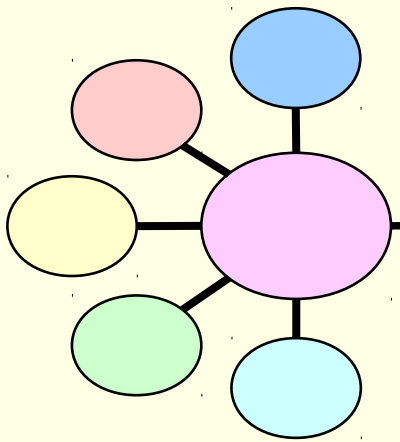


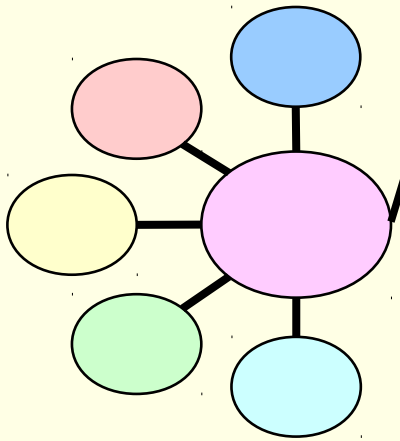
切換時聲音跟影像資料不能中斷！

巨量資料的奇幻漂流 Life of Big Data

Internet of Things 物聯網

Sensor Network





Smart Grid

雲 資料中心 提供服務

Public Data Hub
Data as a Service
開放資料
Open Data

Web 2.0

Query

Map Reduce

Storage

Big Data

端 各類裝置 存取服務

Mobile Computing

Cloud Computing 雲端運算

物聯網的時代來臨 Internet of Things

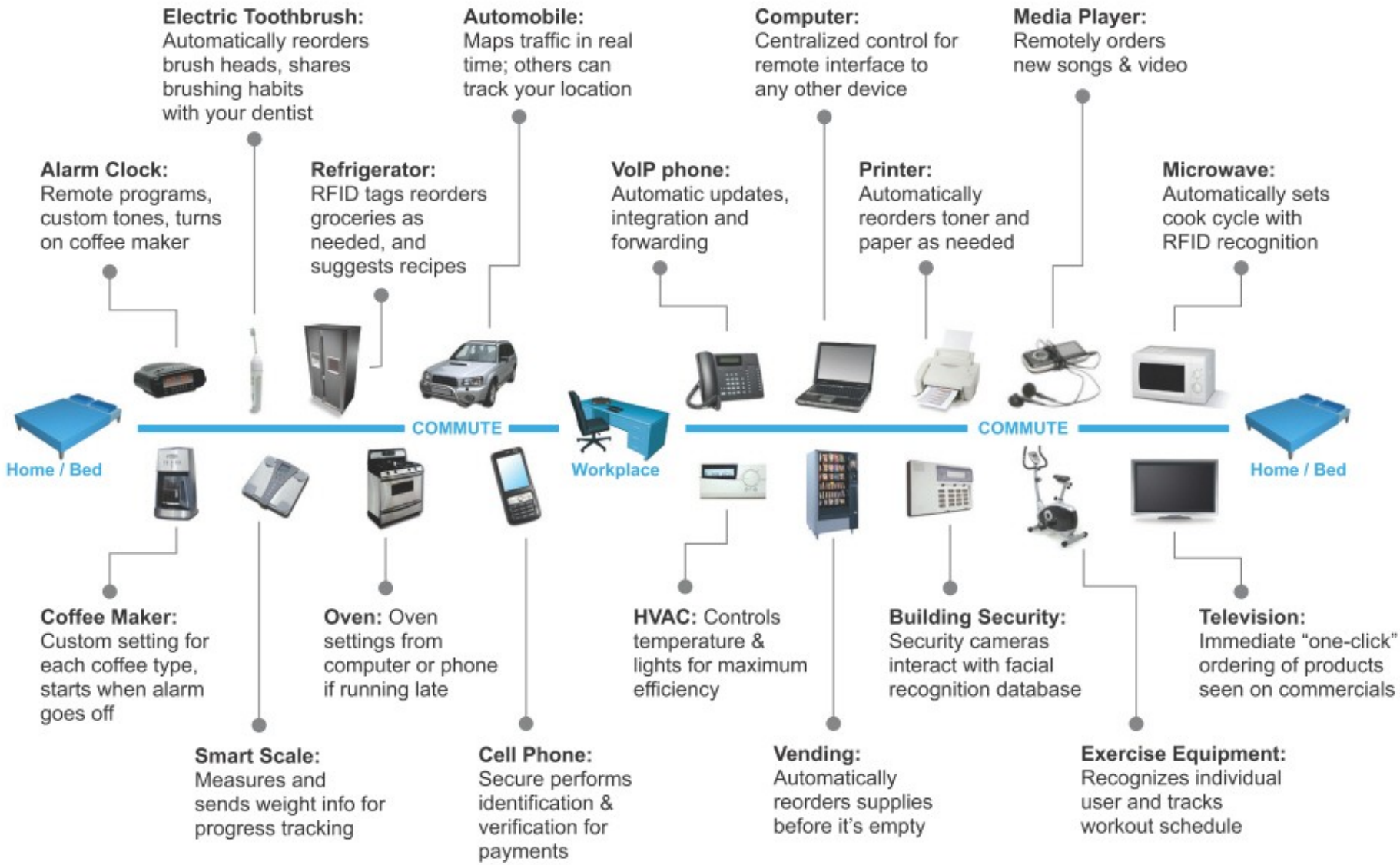


Figure 3. The Internet of Things

圖片來源： Attacks on Mobile and Embedded Systems: Current Trends by Mocana

Wifi體重計：透過網路管理你的體重

無敵小恩恩 發表於 2009年10月05日 13:54 | 收藏此文

讚 0

+1 0

網路入侵到家電用品只是時間的問題，現在體重計也有內建無線網路，透過網路能將你每天量的體重都紀錄下來，不用筆記也能做好更準確的體重管理。有了這麼方便的上網體重計，這下沒有藉口懶得管理你的體重了。 懶人減肥專用機

不只是記錄體重，這台體重計還能算出你的BMI，不但能將資料傳到網站上去，也能透過Wifi傳到iPhone上，經過每天的測量之後，能在網路上看到曲線圖，進而管理你的體重，如果上升的太快就表示最近吃太多動太少，對於懶得記錄自己體重的人來說，這是一台不錯的監督機器，而一台要價159美金，約5000元新台幣，並不便宜，昂貴的價格要普及化還是有一段很長的時間，而且台灣也沒販賣，要買到也不容易。

http://www.techbang.com/posts/1258-wifi-scales-through-the-internet-to-manage-your-weight

【另類解釋】

物聯網＝

連結多個感測網路

Ex. 智慧家庭

氣象局＋鬧鐘＋窗簾

＋咖啡機＋智慧車輛

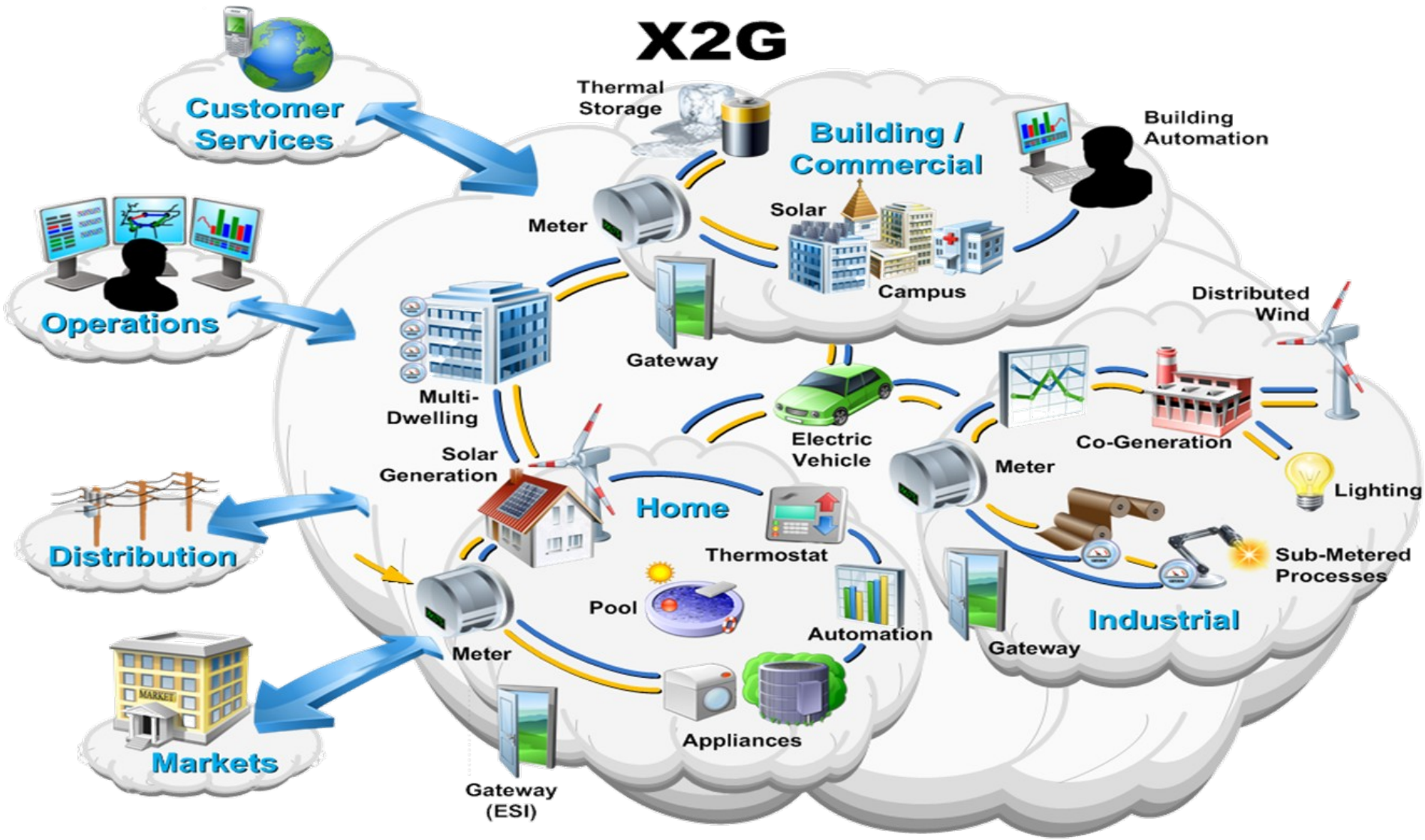
【關鍵技術】

M2M Protocol

(Machine To Machine)

哪一個才是跨感測裝置的通訊格式標準？

智慧電網的新政策 Smart Grid



圖片來源： NIST ，美國國家標準局

【規模擴大到城市】

智慧城市
Smart City

Agenda 演講大綱

From WSN to IoT? 物聯網的崛起

What is Smart City ? 淺談智慧城市

What is Open Data ? 何謂開放資料

What is Big Data ? 何謂海量資料

Join Communities ! 加入國內社群

十九世紀是帝國的世紀
二十世紀是國家的時代

ANATOMY OF A SMART CITY

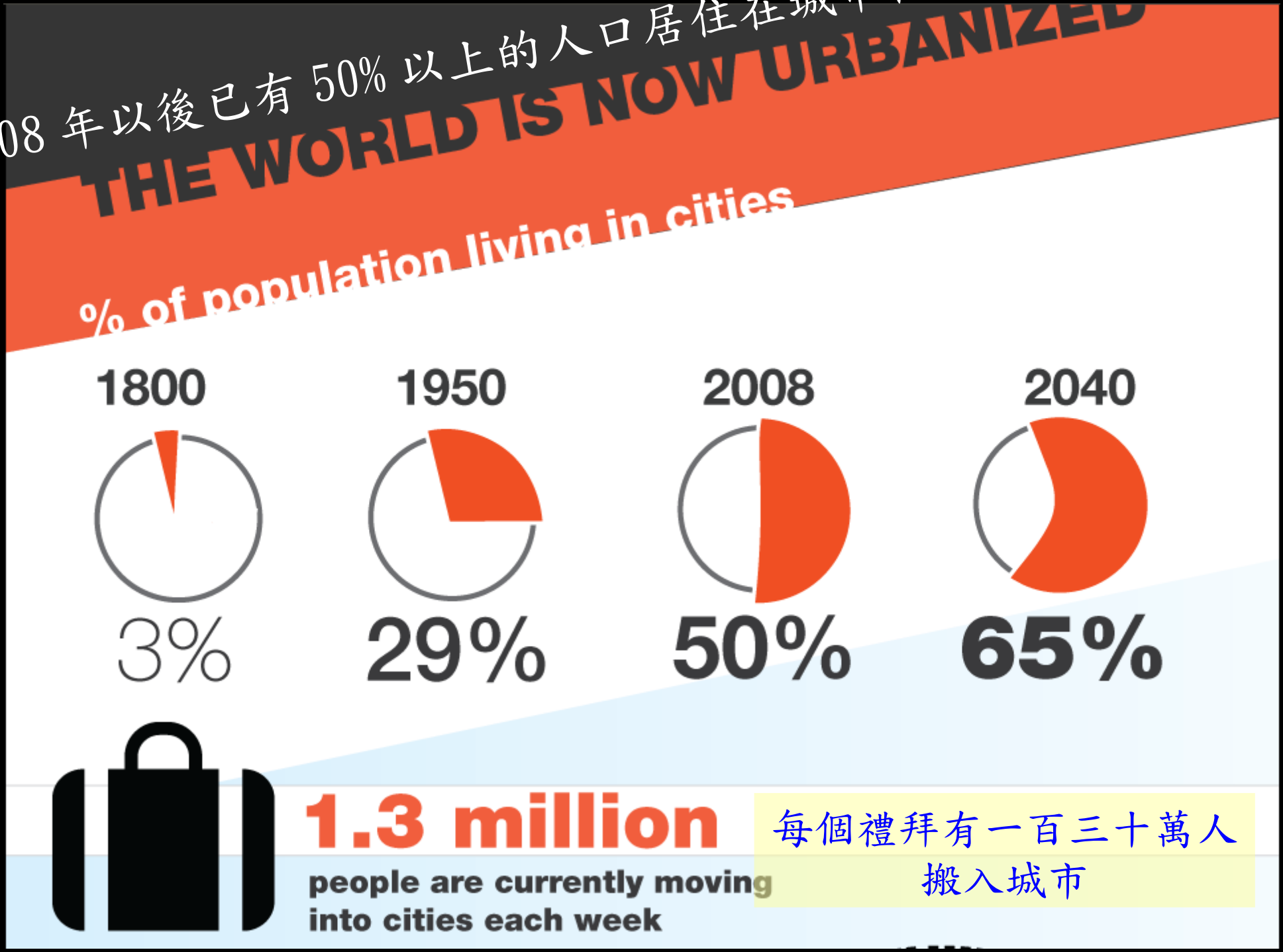
"The 19th century was a century of empires, 20th century was a century of nation states and the 21st century will be a century of cities."

- Former Denver Mayor W. Webb

二十一世紀將是城市的時代

http://www.postscapes.com/anatomy-of-a-smart-city-full

2008 年以後已有 50% 以上的人口居住在城市裡



每個禮拜有一百三十萬人搬入城市

http://www.postscapes.com/anatomy-of-a-smart-city-full

全球有二十一個《超城市》

東京是目前全球最大的城市

There are

21 MEGACITIES

With over 10 Million people

Up until 1975 there were just 3
New York, Tokyo and Mexico City

Largest City:

Tokyo 36 Million +

If it were a country, it would rank 35th in population size

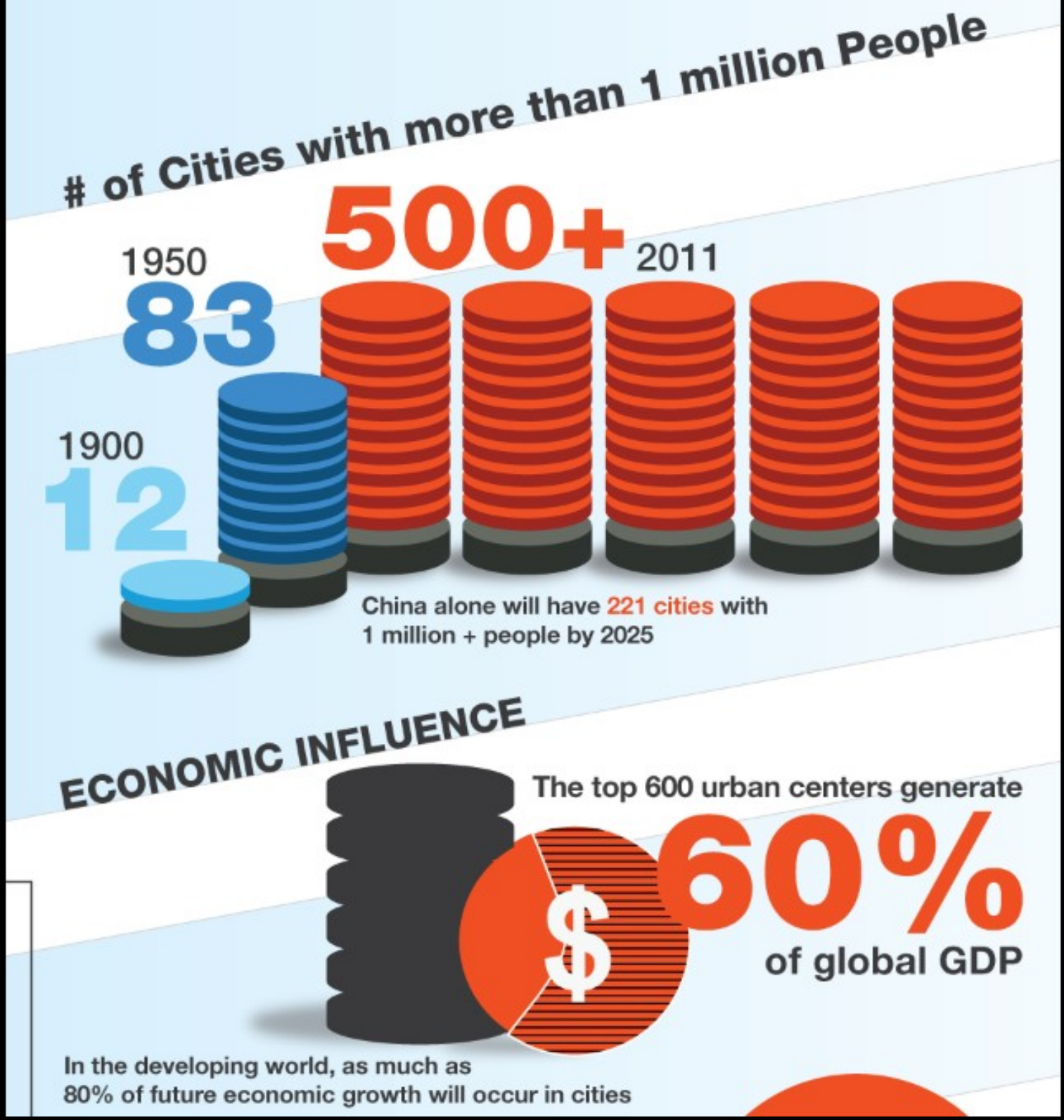
29

By 2025, the number of megacities is expected to reach **with an additional five in Asia** (Shenzhen, Chongqing, Guangzhou, Jakarta and Lahore), **two in Latin America** (Bogotá and Lima), **and one in Africa** (Kinshasa)

http://www.postscapes.com/anatomy-of-a-smart-city-full

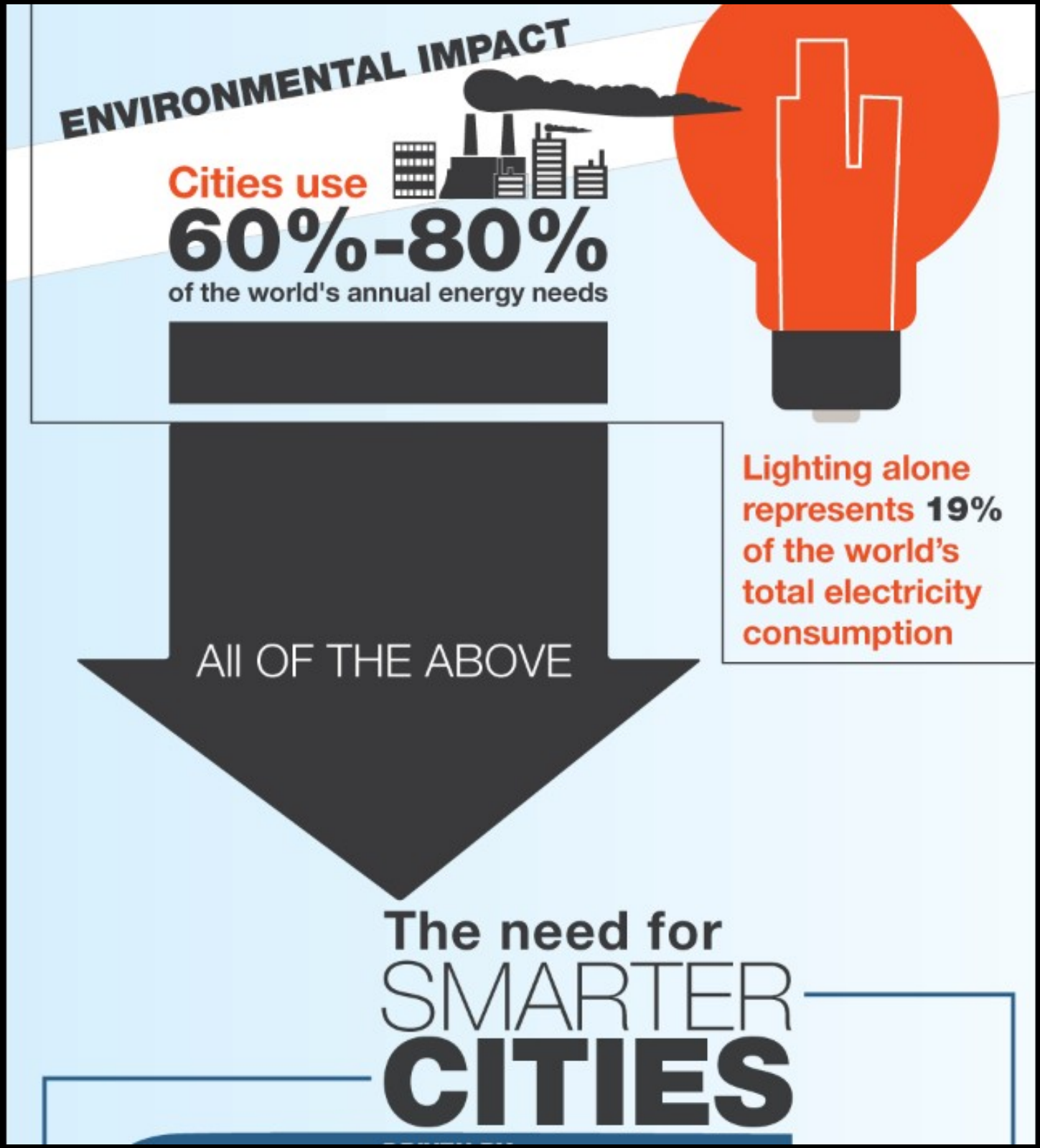
前六百大城市貢獻了百分之六十的生產毛額

2011 年全球已有五百多個城市超過一百萬人口



http://www.postscapes.com/anatomy-of-a-smart-city-full

城市使用了百分之六十至八十的能源

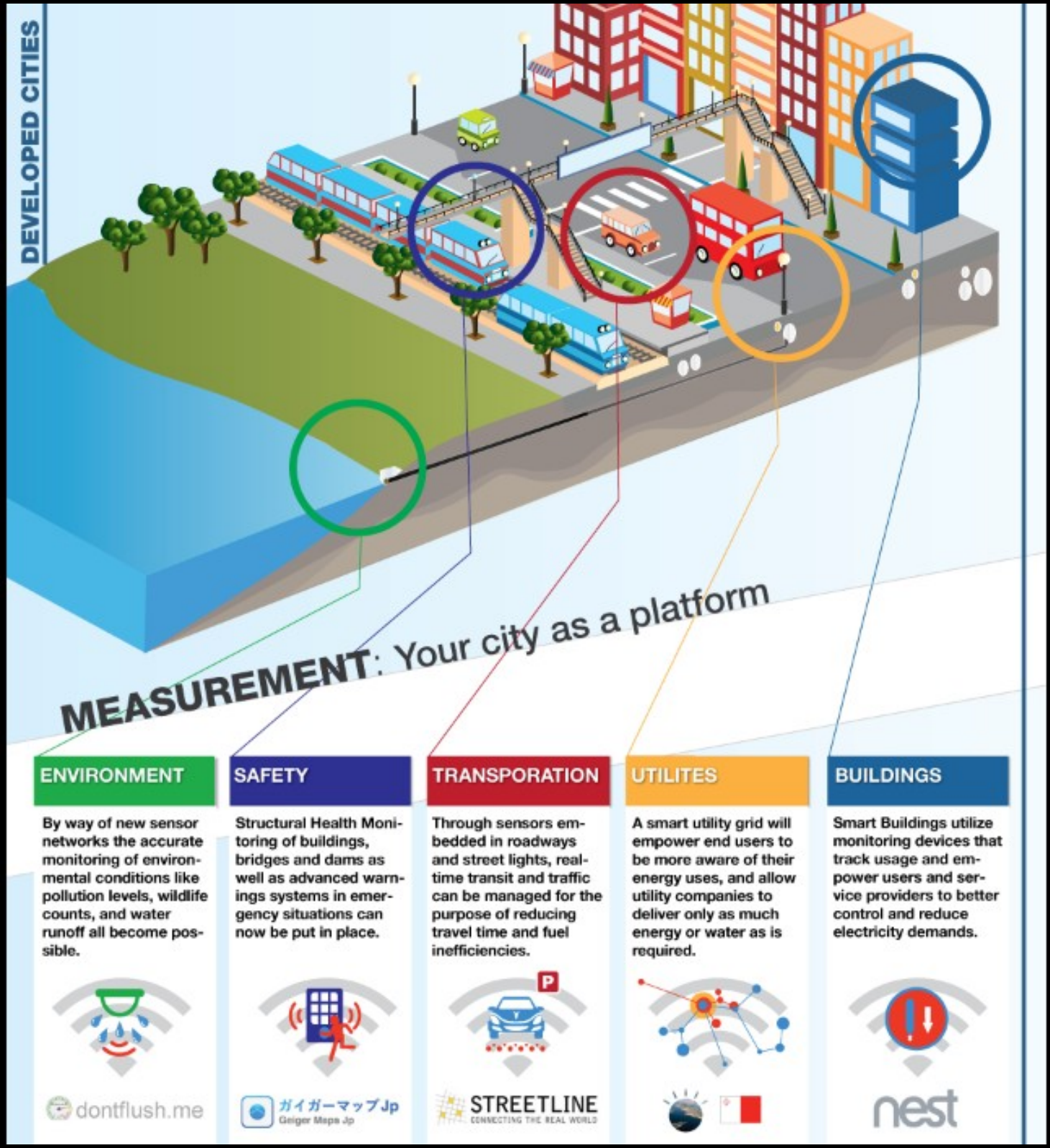


http://www.postscapes.com/anatomy-of-a-smart-city-full

智慧城市主要基於感測器、網路與行動裝置造成了了即時城市資訊系統

http://www.postscapes.com/anatomy-of-a-smart-city-full

量測對象包括環境、安全、交通、電力、建築物



http://www.postscapes.com/anatomy-of-a-smart-city-full

OUTLOOK

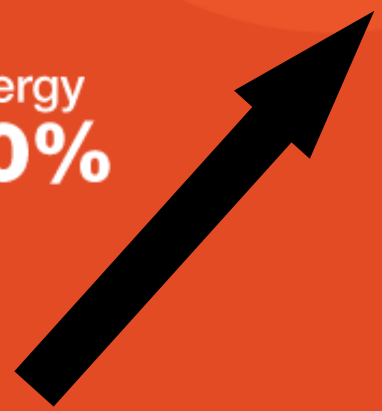
FINANCIAL

The Smarter cities IT market opportunity is expected to be:

$34 Billion annually

CIVIC & ENVIRONMENTAL

Cisco estimates that cities that run on information can improve their energy efficiency by **30%** within 20 years.



YOU
+
SERVICES
+
NETWORKS
+
OPEN
DATA
+
SENSORS

=

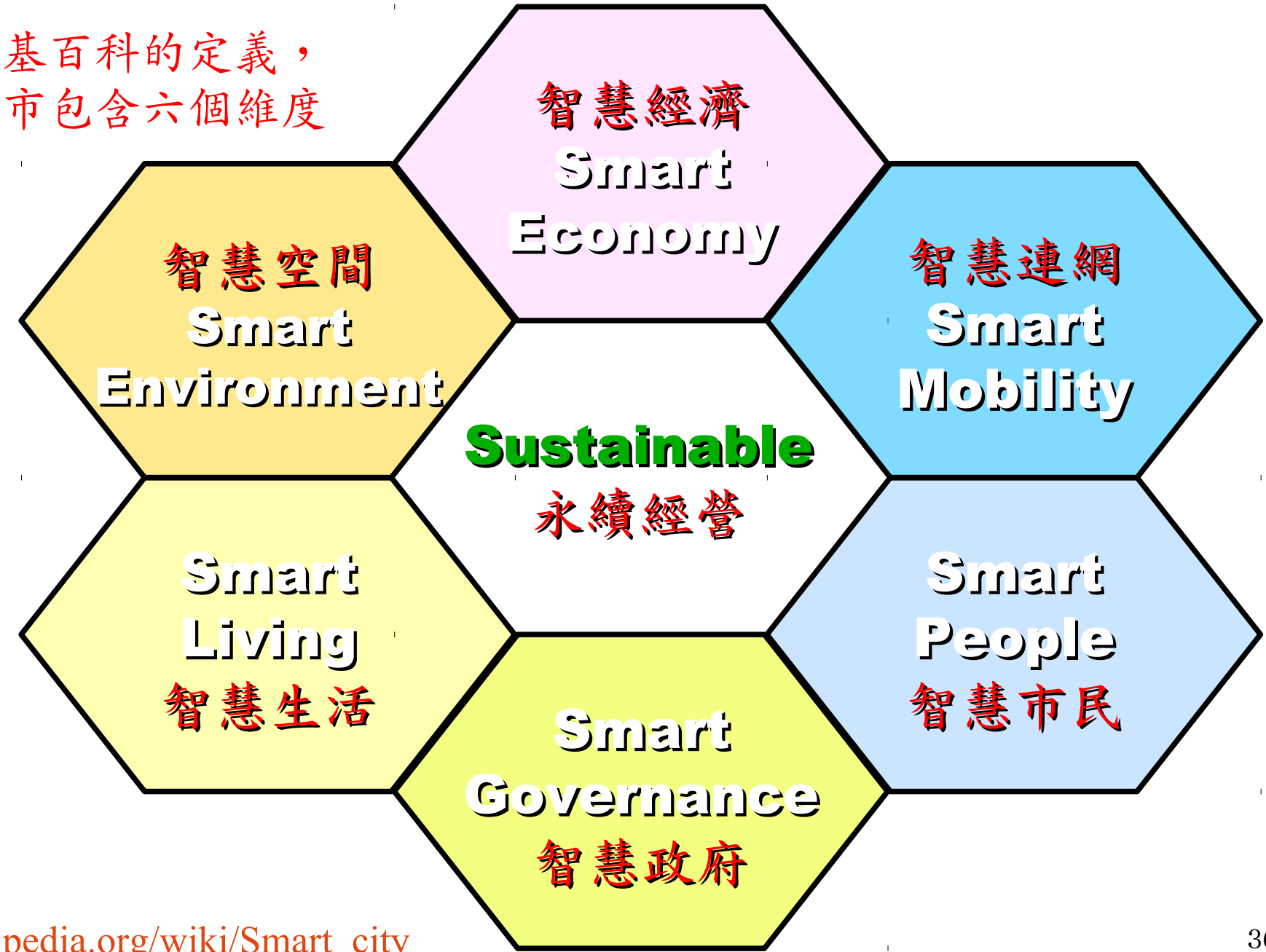
ADAPTIVE SENSEABLE EFFICIENT Cities

Postscapes™

http://www.postscapes.com/anatomy-of-a-smart-city-full

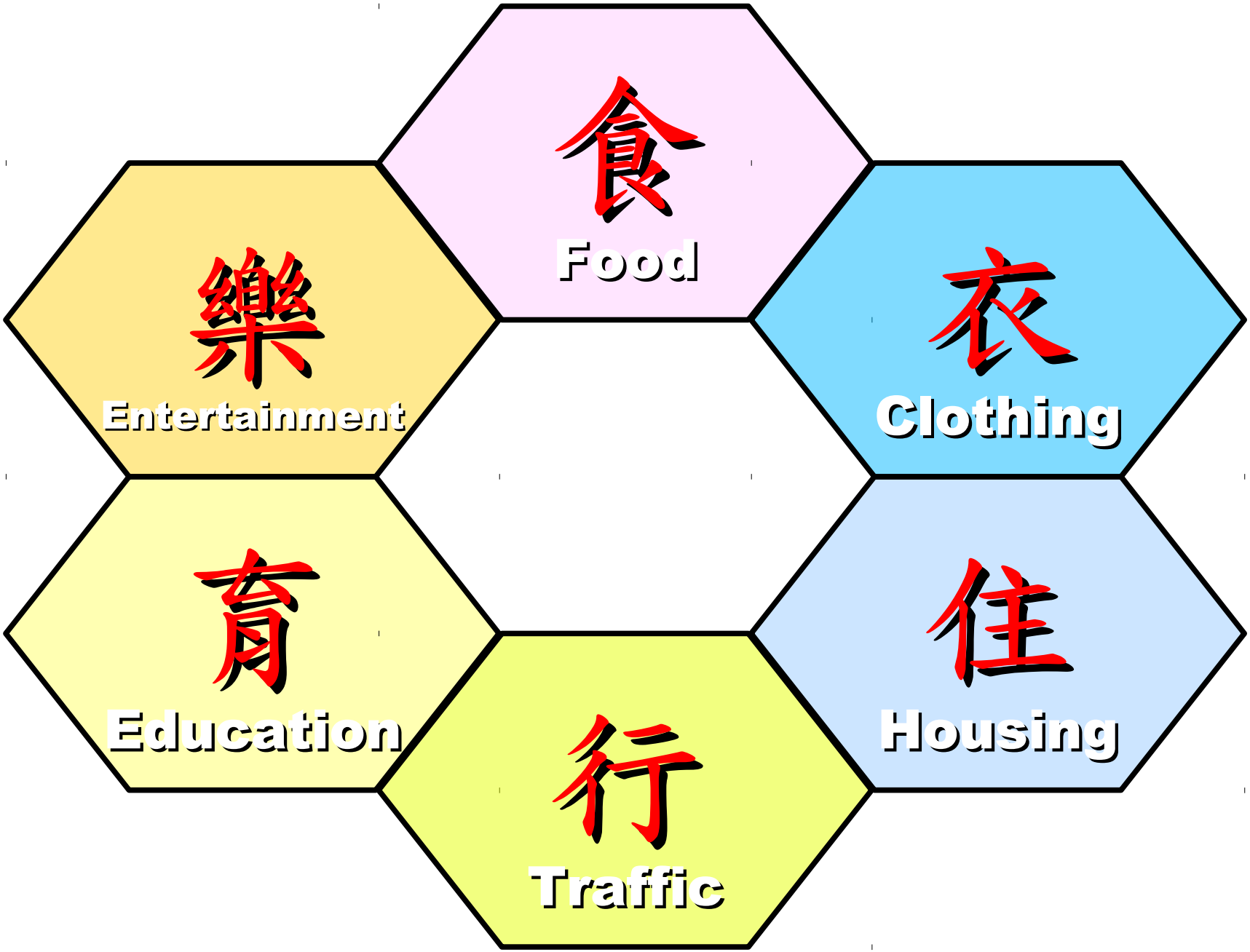
智慧城市的六個維度 6D of Smart Cities

根據維基百科的定義，
智慧城市包含六個維度



http://en.wikipedia.org/wiki/Smart_city

讓人民有感的智慧城市 Ubiquitous Computing



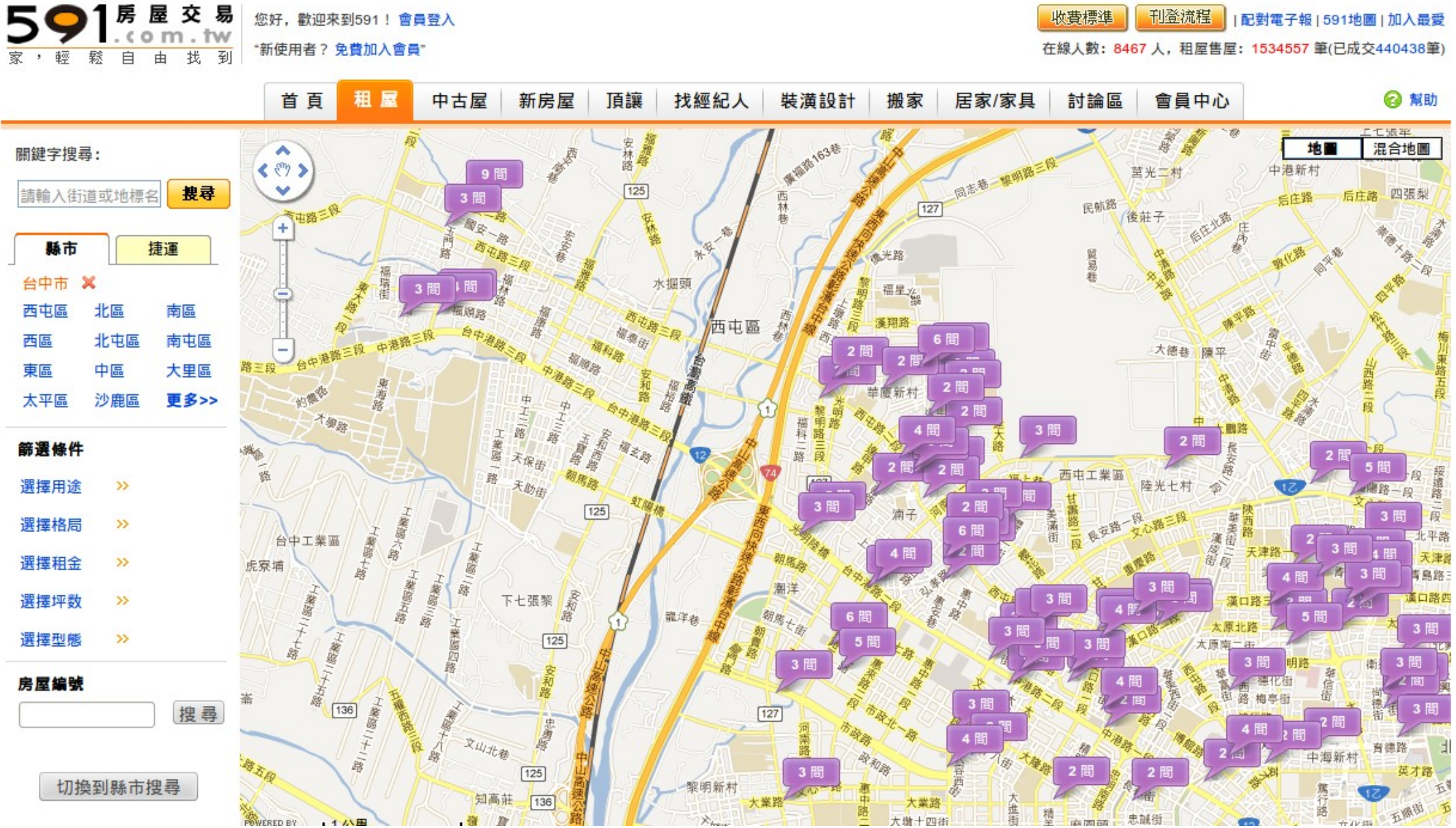
幫忙您找食物的地理資訊系統服務

http://www.ipeen.com.tw/map/#loc=台中市

幫忙您決定該怎麼穿衣服的氣象服務

http://www.cwb.gov.tw/township/

幫忙您找房子的地理資訊服務

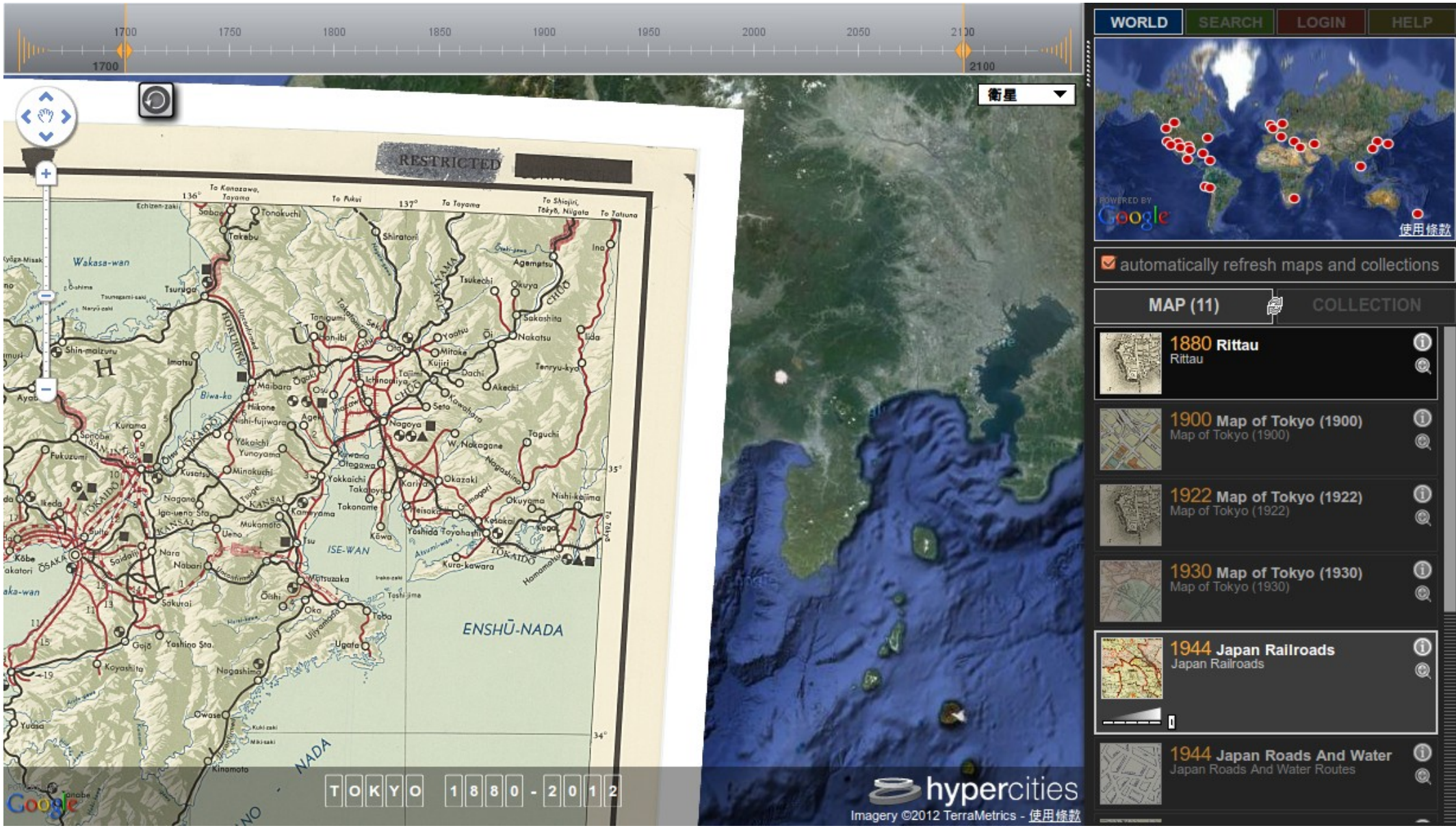


http://rent.591.com.tw/

幫忙您查即時交通路況的地理資訊服務

http://1968.freeway.gov.tw/

用地理資訊服務學習一個城市的歷史



http://hypercities.ats.ucla.edu/

Agenda 演講大綱

From WSN to IoT? 物聯網的崛起

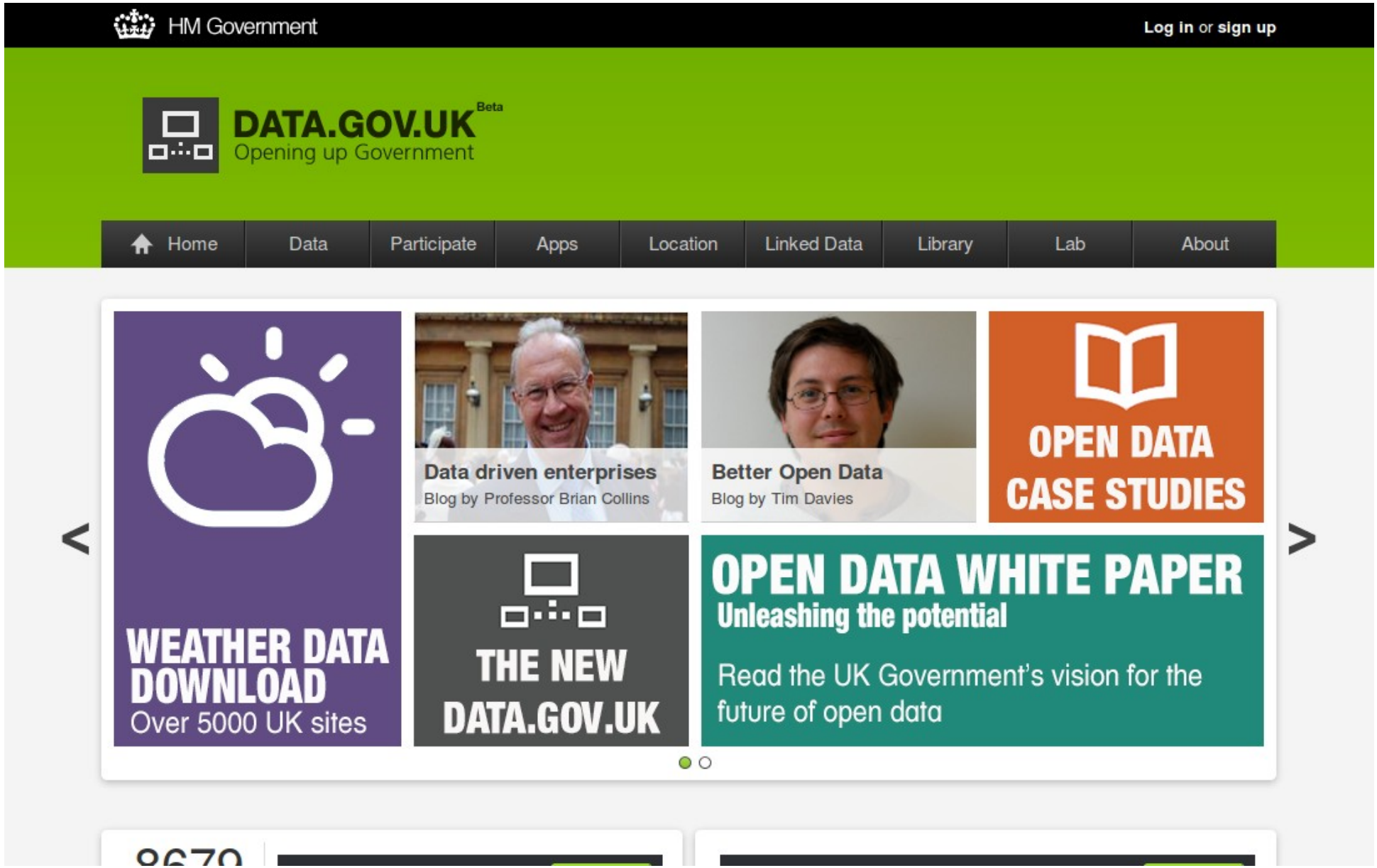
What is Smart City ? 淺談智慧城市

What is Open Data ? 何謂開放資料

What is Big Data ? 何謂海量資料

Join Communities ! 加入國內社群

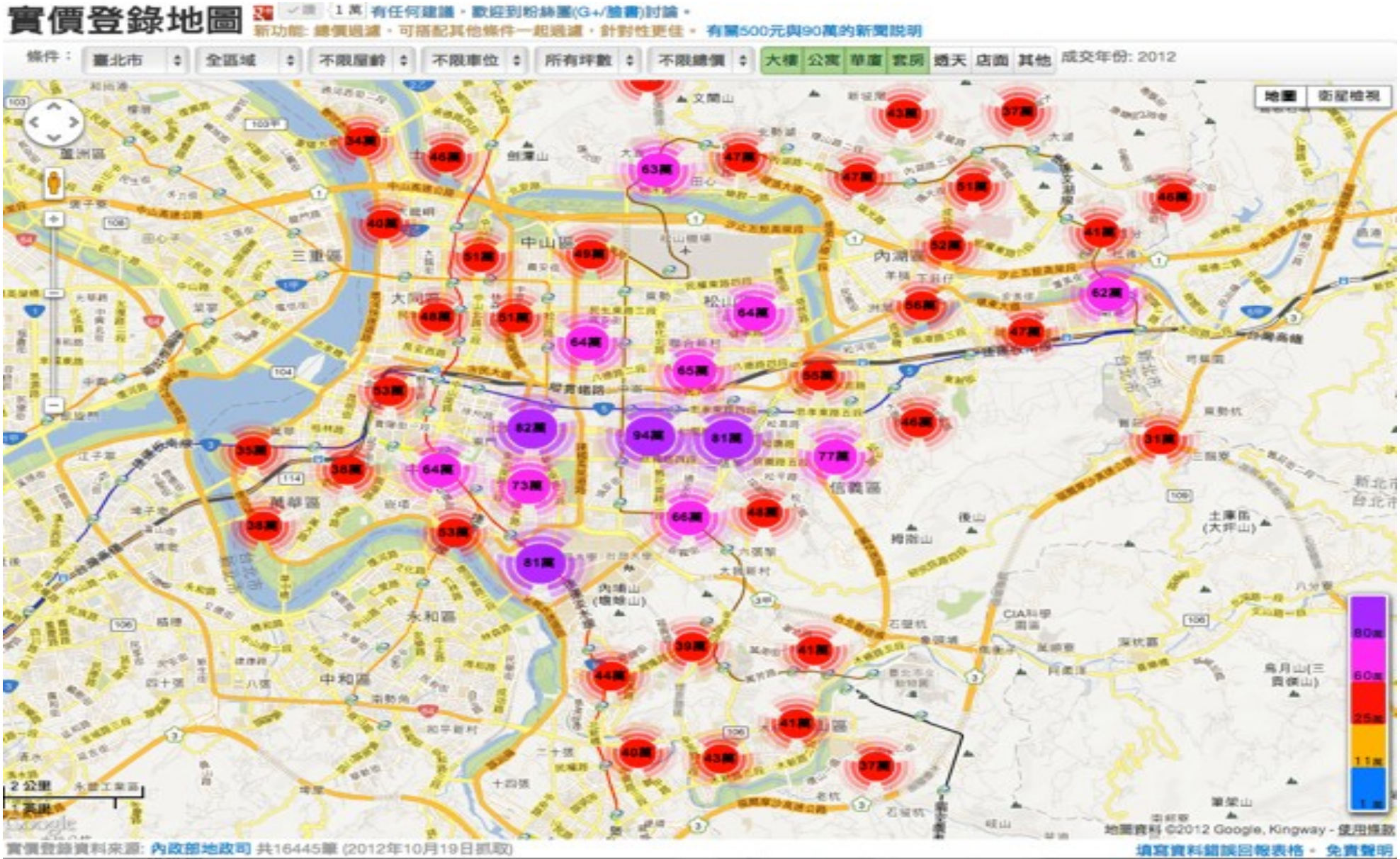
開放資料的新趨勢 Open Data



http://data.gov.uk/

開放資料的實例 Open Data

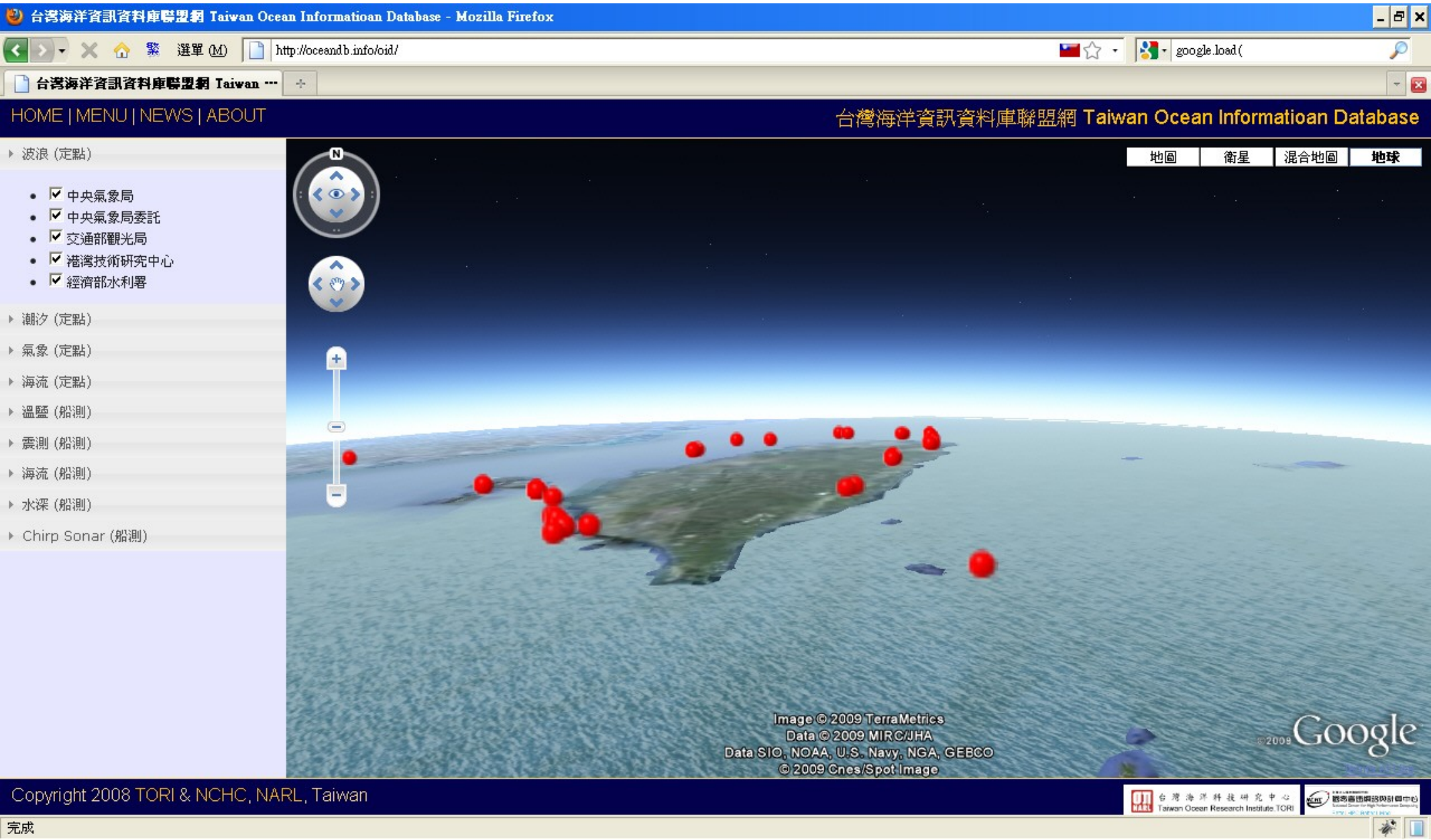
http://real-price.appspot.com/



開放資料的民間推動 Code for Tomorrow

http://codefortomorrow.org/

我個人的開放資料慘痛經驗 Ocean Database



http://oceandb.info

Agenda 演講大綱

From WSN to IoT? 物聯網的崛起

What is Smart City ? 淺談智慧城市

What is Open Data ? 何謂開放資料

What is Big Data ? 何謂海量資料

Join Communities ! 加入國內社群

Why are they named by “ SMART ” ?!
智慧打哪兒來？！

Smart Phone
智慧手機

Smart Car
智慧車輛

Smart Grid
智慧電網

SMART ?
哪裡長
智慧了？

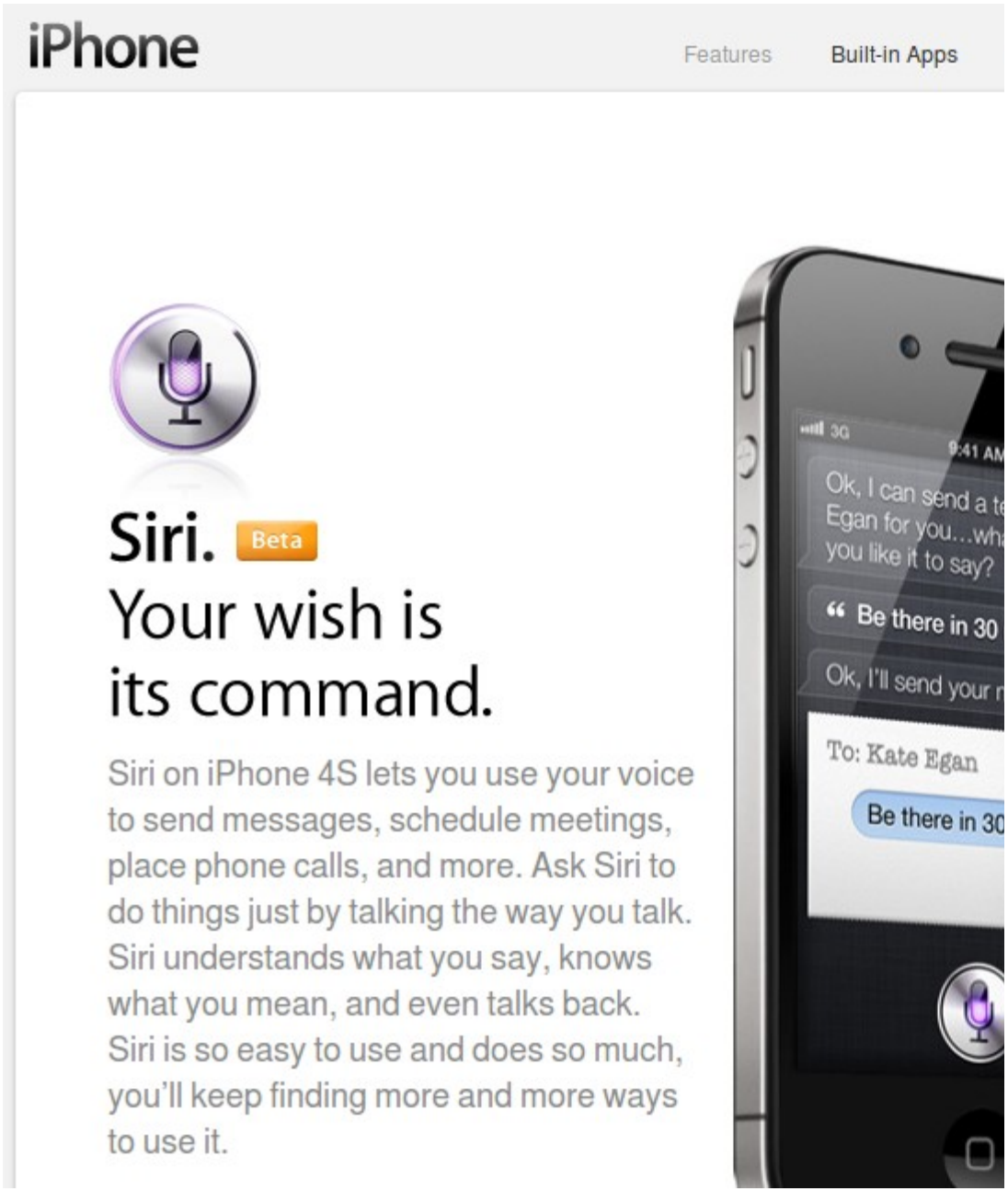
Smart City
智慧城市

Smart Home
智慧家庭

Smart Meter
智慧電錶

資料 Data → 知識 Knowledge → 智慧 Wisdom

Can Machine understand You? 讓機器更懂你？



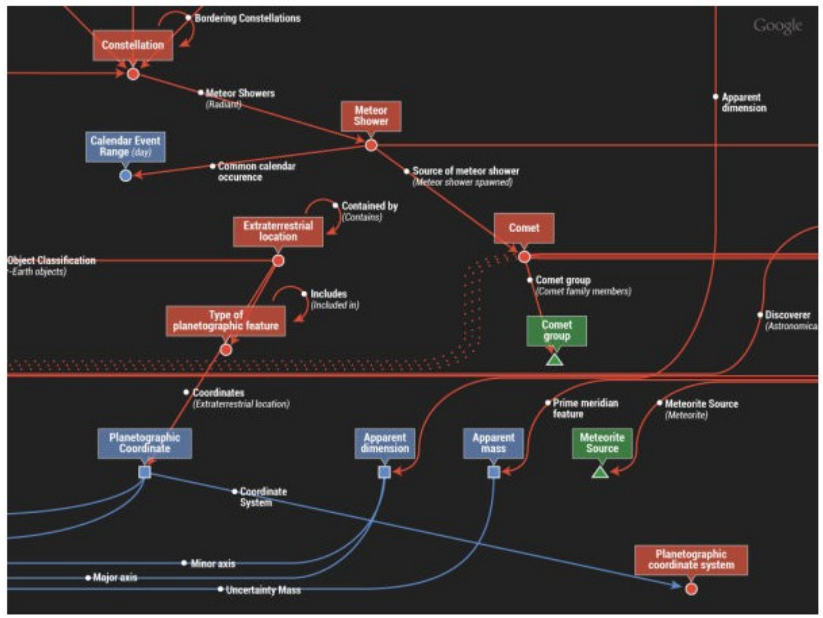
Google將發展「人工智慧」　永久改變搜尋引擎

2012年02月15日 00:11

點評：*超級阿斯拉，衝啊！（阿斯拉：好的，隼人！）*

記者黃郁棋／綜合報導

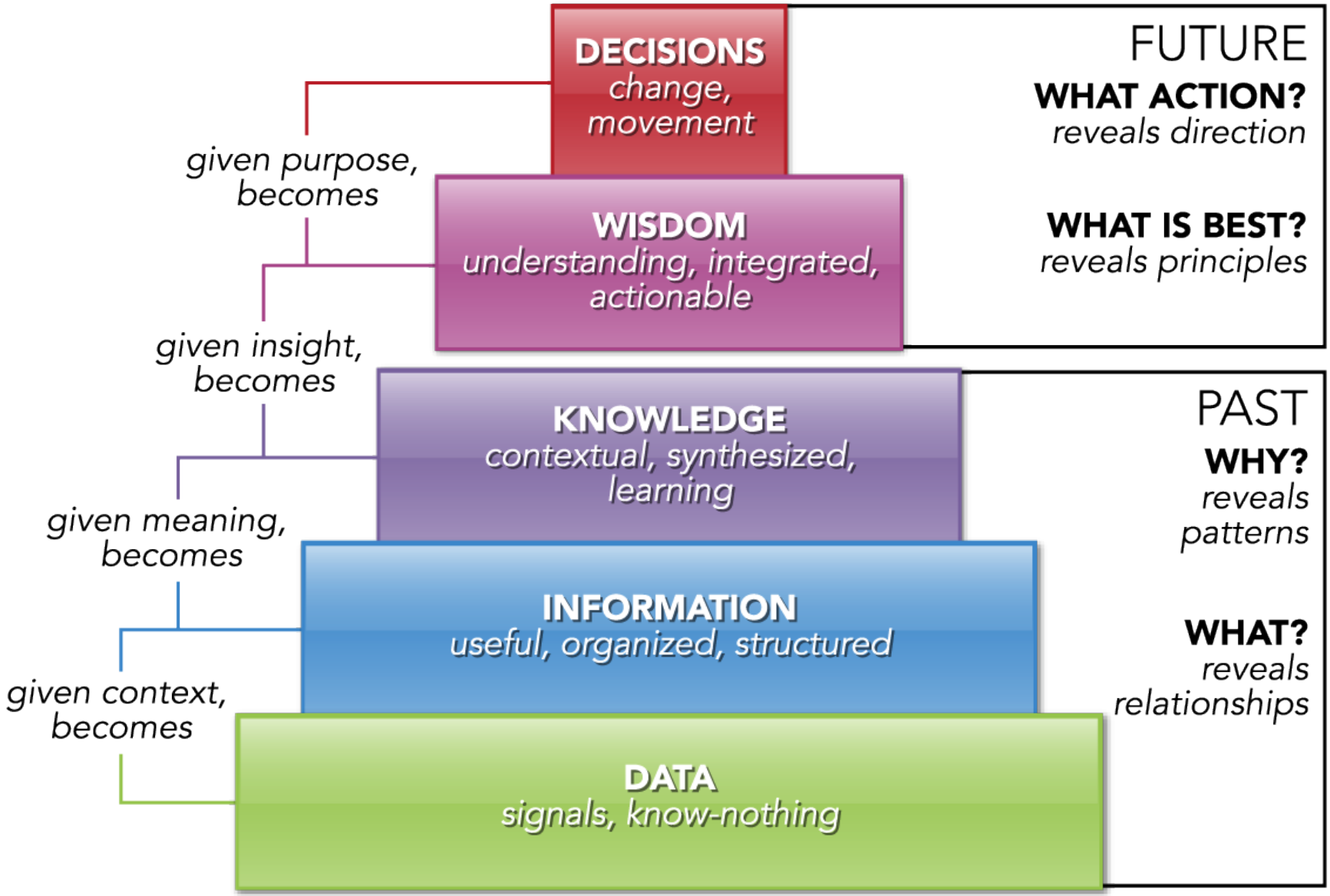
每個人都在猜，下一波網路革命是什麼？每個人都在猜，未來的世界會如何運作？Google的資深副總Amit Singhai透露了一點訊息。「Google正努力從『單字』層面進展到『意義』層面，未來搜尋引擎提供的不只是關鍵字搜尋，搜尋引擎甚至會『明白』你到底要什麼。」



▲Google未來將會朝「人工智慧」前進。(圖／取自mashable.com)

http://www.ettoday.net/news/20120215/25085.htm

Knowledge is from the Past, Wisdom aim for Future
知識源自於過去，智慧在能預測未來



http://www.pursuantgroup.com/blog/tag/dikw-model/

What is Big Data?! 何謂『海量資料』？

海量資料泛指資料大小已無法用一般軟體擷取、管理與處理；
單一資料集大小介於數十 TB 至數 PB 的資料。

'Big Data' = few dozen TeraBytes to PetaBytes in single data set.

Definition [edit]

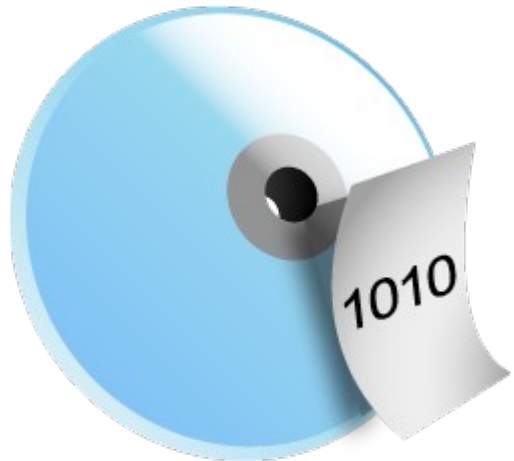
Big data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes of data in a single data set.

In a 2001 research report[14] and related conference presentations, then 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/out), and variety (range of data types, sources). Gartner continues to use this model for describing big data.[15]

出處：http://en.wikipedia.org/wiki/Big_data

多個檔案，容量 100TB

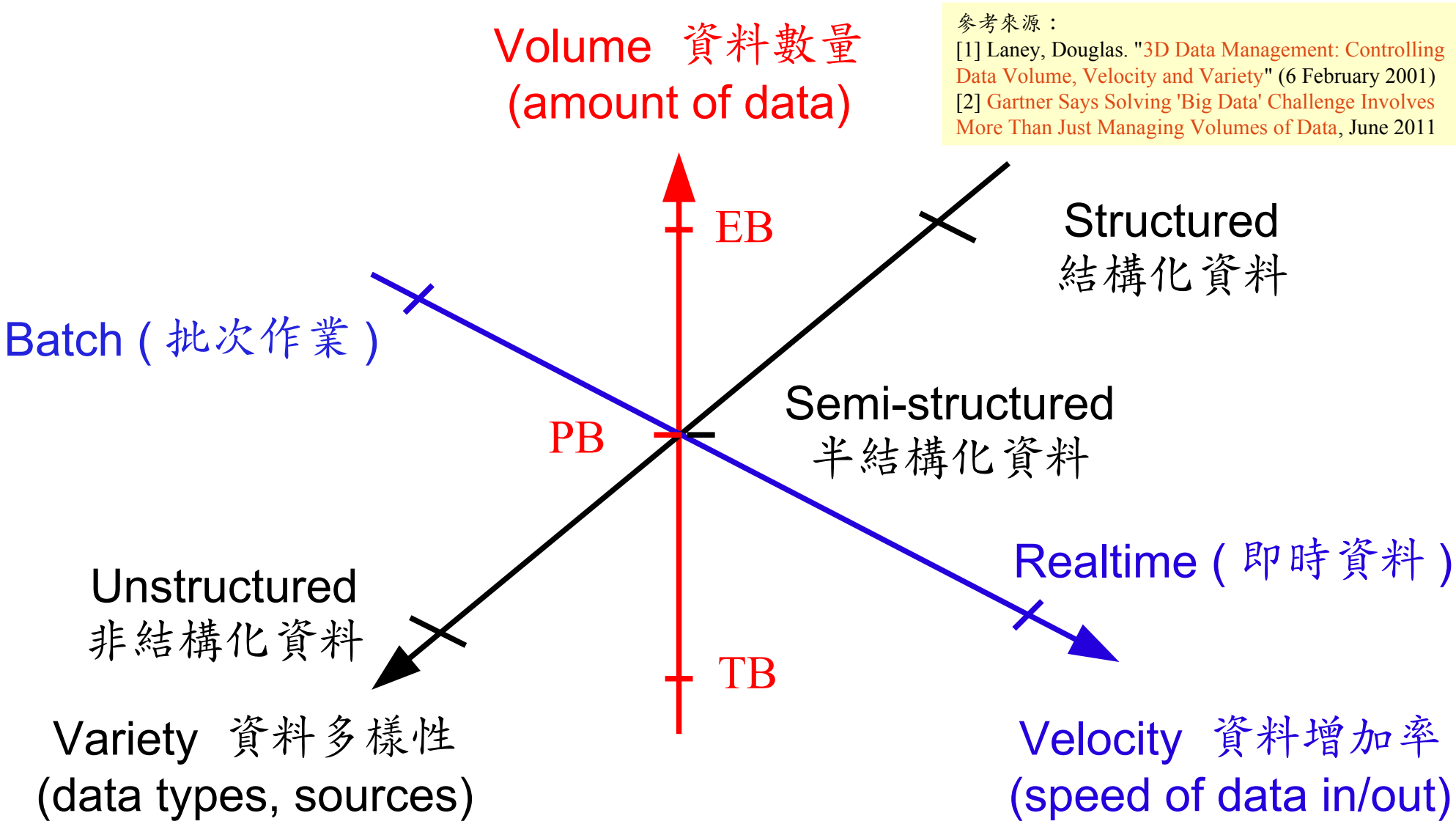
一個資料庫，容量 100TB



一個檔案，容量 100TB

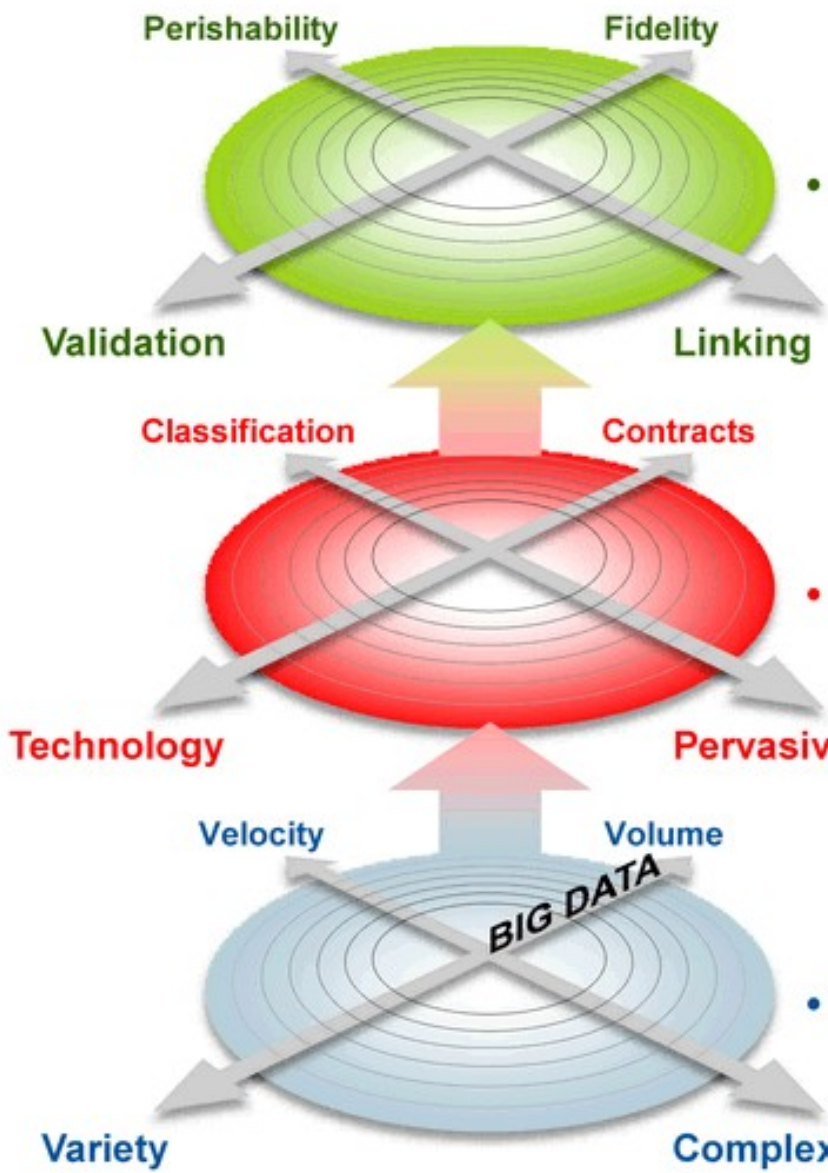
Gartner Big Data Model？ 海量資料的模型？

海量資料的挑戰在於如何管理「數量」、「增加率」與「多樣性」



參考來源：
[1] Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety" (6 February 2001)
[2] Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data, June 2011

12D of Information Management? 12 個維度？



品質管控

- Qualification and Assurance

權限管控

- Access Enablement and Control

數量管控

- Quantification

Big Data 只是終極資訊管理的開端！

Source: Gartner (March 2011), 'Big Data' Is Only the Beginning of Extreme Information Management, 7 April 2011, http://www.gartner.com/id=1622715

New IT Architecture toward Cloud Computing !!

用雲掌握資料，加以分析，形成智能給端用

雲

資料中心
提供服務

雲端設計新思維：端的智能來自於雲的服務

Devices share the wisdom of Cloud

端

各類裝置
存取服務

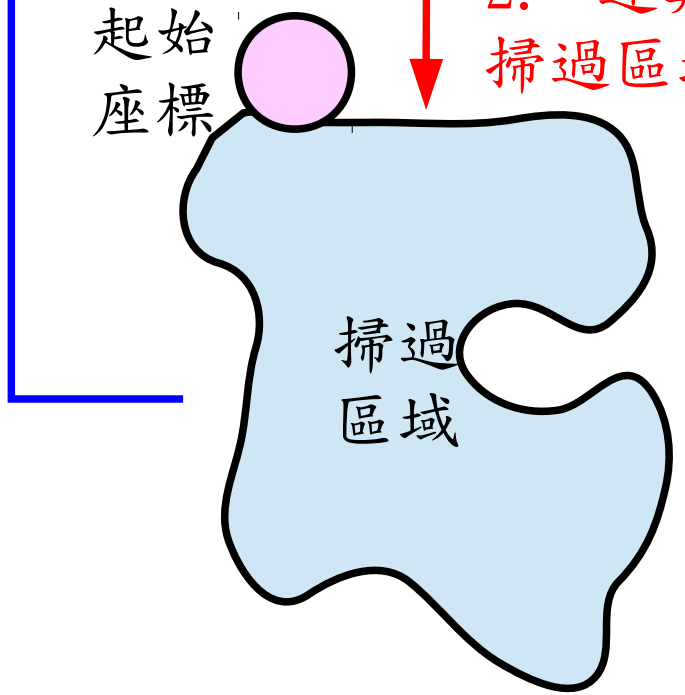
4. 下載最佳路徑

1. 上傳行進軌跡

6. 上傳最佳路徑執行狀態
（判斷是否需要修正）

3. 運算
最佳路徑

2. 運算
掃過區域



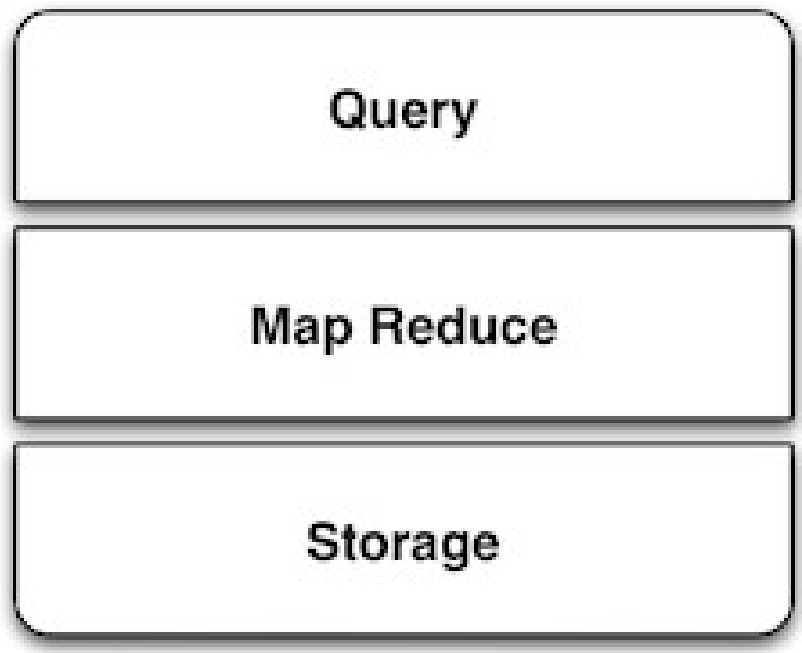
5. 執行
最佳路徑

曾經有人問道：我是學電機的，那電機與控制跟雲端有甚麼關聯性？

The SMAQ stack for big data
海量資料處理的資訊架構

做網頁相關的人可能聽過 LAMP

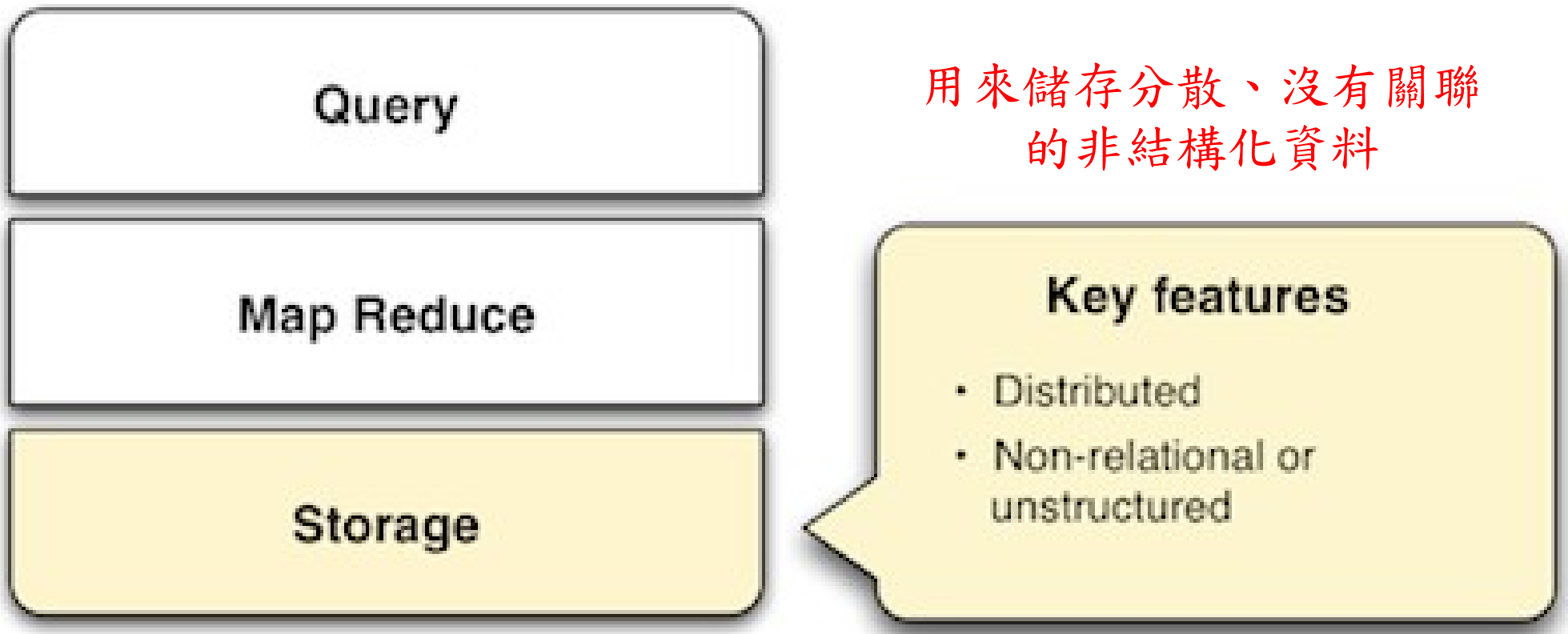
未來處理海量資料的人必需知道 SMAQ（Storage, MapReduce and Query）



參考來源：The SMAQ stack for big data，Edd Dumbill，22 September 2010，http://radar.oreilly.com/2010/09/the-smaq-stack-for-big-data.html

圖片來源：http://smashingweb.ge6.org/wp-content/uploads/2011/10/apache-php-mysql-ubuntu.png

The SMAQ stack for big data
海量資料處理的資訊架構

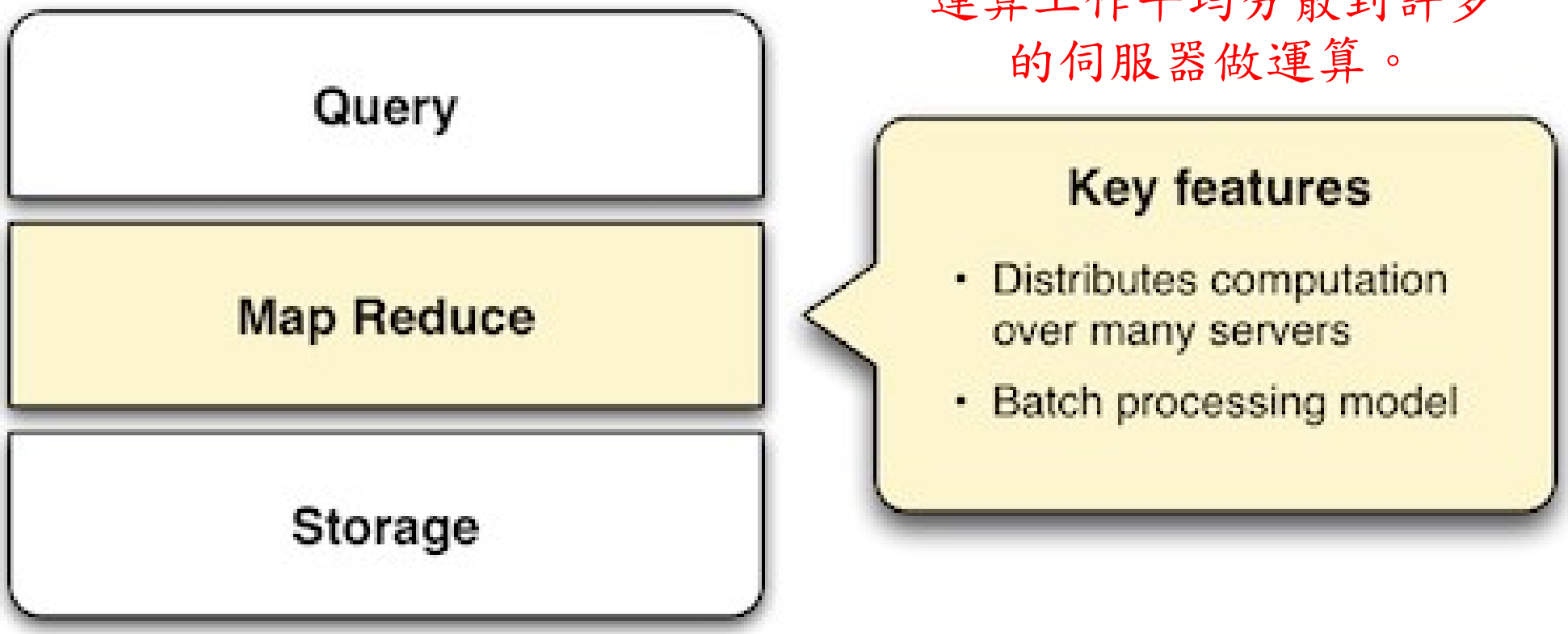


用來儲存分散、沒有關聯的非結構化資料

參考來源：The SMAQ stack for big data，Edd Dumbill，22 September 2010，http://radar.oreilly.com/2010/09/the-smaq-stack-for-big-data.html

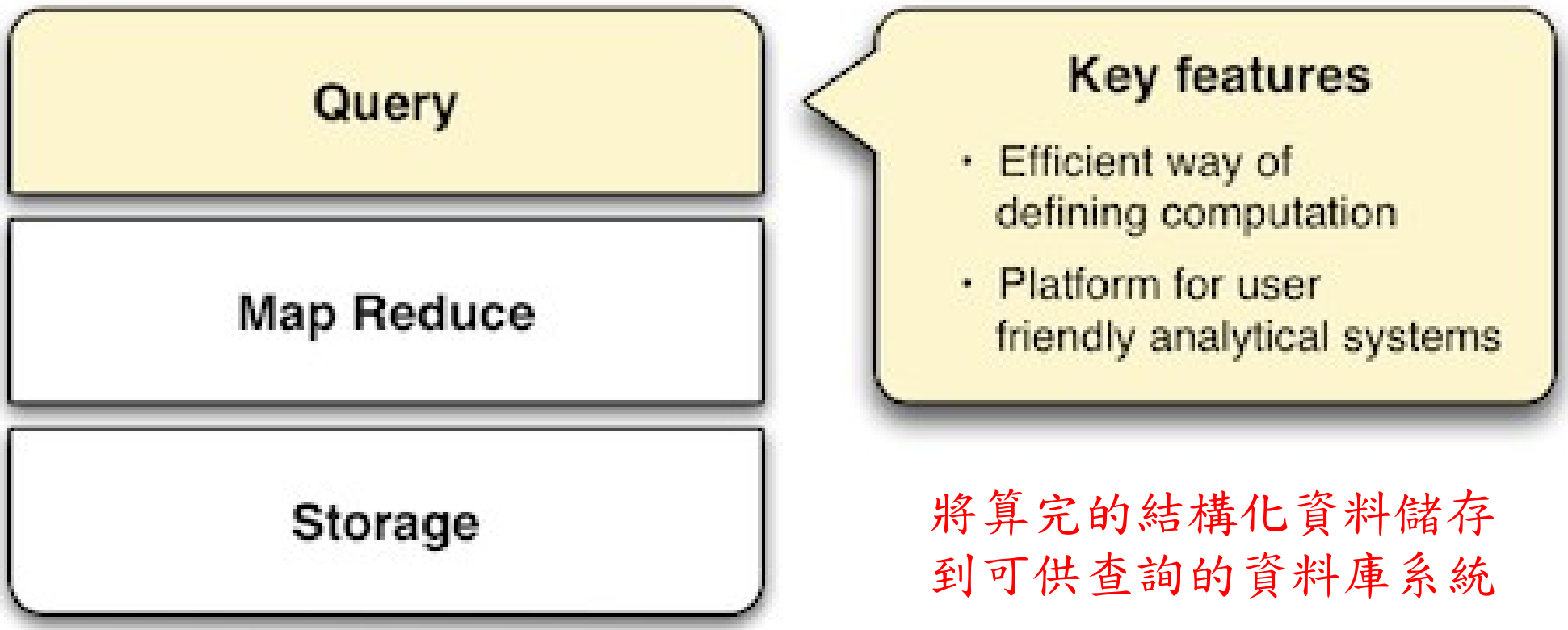
The SMAQ stack for big data
海量資料處理的資訊架構

運用批次處理的方式，將運算工作平均分散到許多的伺服器做運算。



參考來源：The SMAQ stack for big data，Edd Dumbill，22 September 2010，http://radar.oreilly.com/2010/09/the-smaq-stack-for-big-data.html

The SMAQ stack for big data
海量資料處理的資訊架構



將算完的結構化資料儲存到可供查詢的資料庫系統

參考來源：The SMAQ stack for big data，Edd Dumbill，22 September 2010，http://radar.oreilly.com/2010/09/the-smaq-stack-for-big-data.html

Three Core Technologies of Google ....
Google 的三大關鍵技術 ....

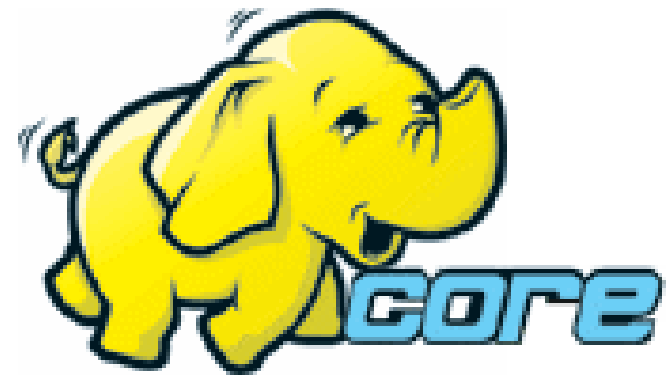
- Google 在一些會議分享他們的三大關鍵技術
- Google shared their design of web-search engine
  - SOSP 2003 :
  - "The Google File System"
  - http://labs.google.com/papers/gfs.html
  - OSDI 2004 :
  - "MapReduce : Simplifed Data Processing on Large Cluster"
  - http://labs.google.com/papers/mapreduce.html
  - OSDI 2006 :
  - "Bigtable: A Distributed Storage System for Structured Data"
  - http://labs.google.com/papers/bigtable-osdi06.pdf

Open Source Mapping of Google Core Technologies
Google 三大關鍵技術對應的自由軟體

Google 三大關鍵技術	自由軟體對應解決方案
Q = Query **BigTable** A huge key-value datastore	HBase, Hypertable Cassandra, ...
MapReduce To parallel process data	Hadoop MapReduce API Sphere MapReduce API, ...
S = Storage **Google File System** To store petabytes of data	Hadoop Distributed File System (HDFS) Sector Distributed File System

Hadoop

- http://hadoop.apache.org
- Hadoop 是 Apache Top Level 開發專案
- **Hadoop is Apache Top Level Project**
- 目前主要由 Yahoo! 資助、開發與運用
- **Major sponsor is Yahoo!**
- 創始者是 Doug Cutting ，參考 Google Filesystem
- **Developed by Doug Cutting, Reference from Google Filesystem**
- 以 Java 開發，提供 HDFS 與 MapReduce API 。
- **Written by Java, it provides HDFS and MapReduce API**
- 2006 年使用在 Yahoo 內部服務中
- **Used in Yahoo since year 2006**
- 已佈署於上千個節點。
- **It had been deploy to 4000+ nodes in Yahoo**
- 處理 Petabyte 等級資料量。
- **Design to process dataset in Petabyte**



Facebook、**Last.fm**、**Joost** are also powered by Hadoop

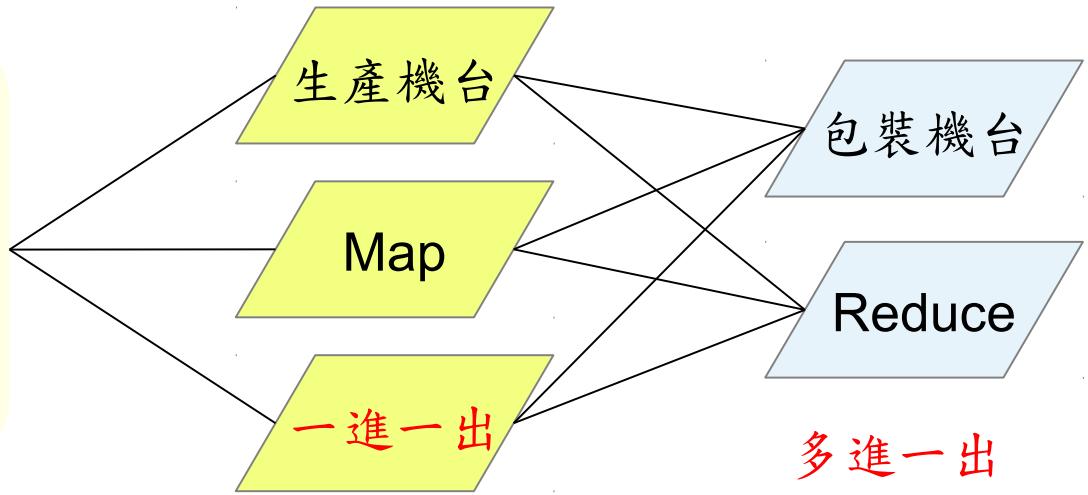
Hadoop 簡介

Hadoop 是一個讓使用者簡易撰寫並執行處理海量資料應用程式的軟體平台。

亦可以想像成一個處理海量資料的生產線，只須學會定義 **map** 跟 **reduce** 工作站該做哪些事情。

就像工廠的倉庫
存放生產原料跟待售貨物

HDFS 存放
待處理的非結構化資料
與處理後的結構化資料



Sector / Sphere

- http://sector.sourceforge.net/
- 由美國資料探勘中心研發的自由軟體專案。
- **Developed by National Center for Data Mining, USA**
- 採用 C/C++ 語言撰寫，因此效能較 Hadoop 更好。
- **Written by C/C++, so performance is better than Hadoop**
- 提供「類似」 Google File System 與 MapReduce 的機制
- **Provide file system similar to Google File System and MapReduce API**
- 基於UDT高效率網路協定來加速資料傳輸效率
- **Based on UDT which enhance the network performance**
- Open Cloud Testbed有提供測試環境，並開發MalStone效能評比軟體
- **Open Cloud Consortium provide Open Cloud Testbed and develop MalStone toolkit for benchmark**

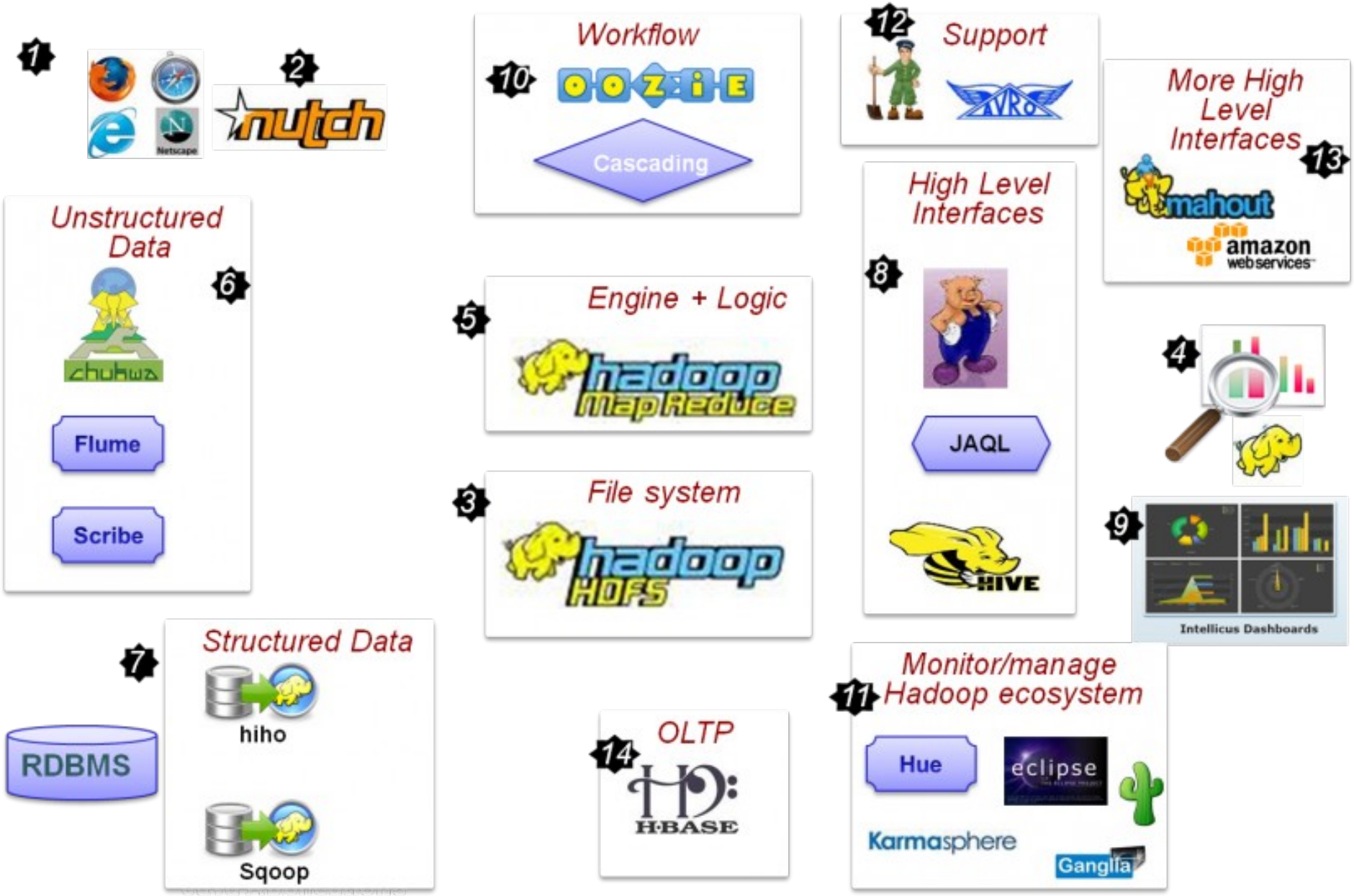
Sector-Sphere

National Center for Data Mining
University of Illinois at Chicago
UIC

open data
Open Data Group
http://www.opendatagroup.com/

Why we choice Hadoop? Good Ecosystem!

豐富的生態系建構出處理海量資料的工具庫



http://rationalintelligence.com/wp_log/?p=104

巨量資料的未來？ Future of Big Data ？

http://www.huffingtonpost.com/deb-roy/the-birth-of-a-word_b_2639625.html

個人化的巨量資料

Big Data Can Get Very Personal

http://www.ted.com/talks/deb_roy_the_birth_of_a_word.html

Agenda 演講大綱

From WSN to IoT? 物聯網的崛起

What is Smart City ? 淺談智慧城市

What is Open Data ? 何謂開放資料

What is Big Data ? 何謂海量資料

Join Communities ! 加入國內社群

【建議】

現在做研究也要打團體戰

加入技術討論社群

增加學習知識的觸角！

雲端產業供應鏈 Supply Chain of Cloud Computing

應用軟體供應商

端

各類裝置存取服務

行動裝置共通平台

htc quietly brilliant

軟體服務供應商

Google

amazon web services™

雲

資料中心提供服務

資料中心機房維運

acer

基本硬體建設組件

FOXCONN 鴻海科技集團

英業達集團 Inventec

廣達電腦 Quanta Computer

雲端產業供應鏈 Supply Chain of Cloud Computing

應用軟體供應商

端

各類裝置存取服務

行動裝置共通平台

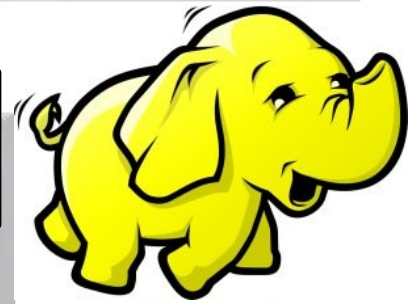
htc quietly brilliant

軟體服務供應商

Google

amazon web services™

大象擺這裡



資料中心機房維運

acer

資料中心提供服務

基本硬體建設組件

FOXCONN 鴻海科技集團

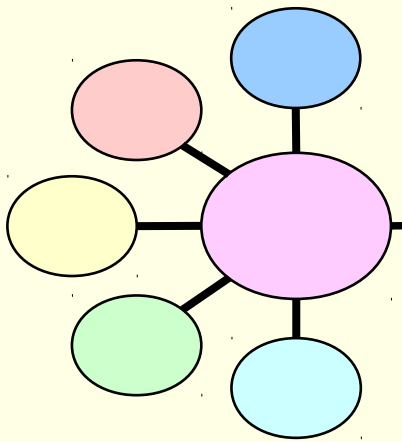
英業達集團 Inventec

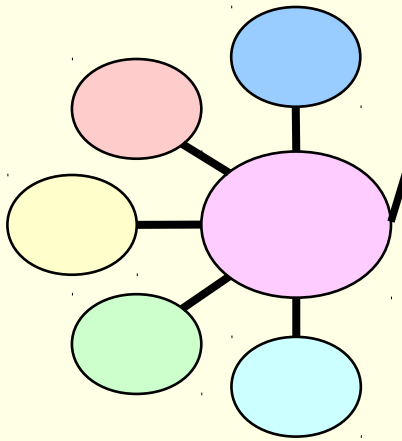
廣達電腦 Quanta Computer

巨量資料的奇幻漂流 Life of Big Data

Internet of Things 物聯網

Sensor Network





Smart Grid

雲 資料中心 提供服務

Public Data Hub
Data as a Service

開放資料
Open Data

Web 2.0

Query

Map Reduce

Storage

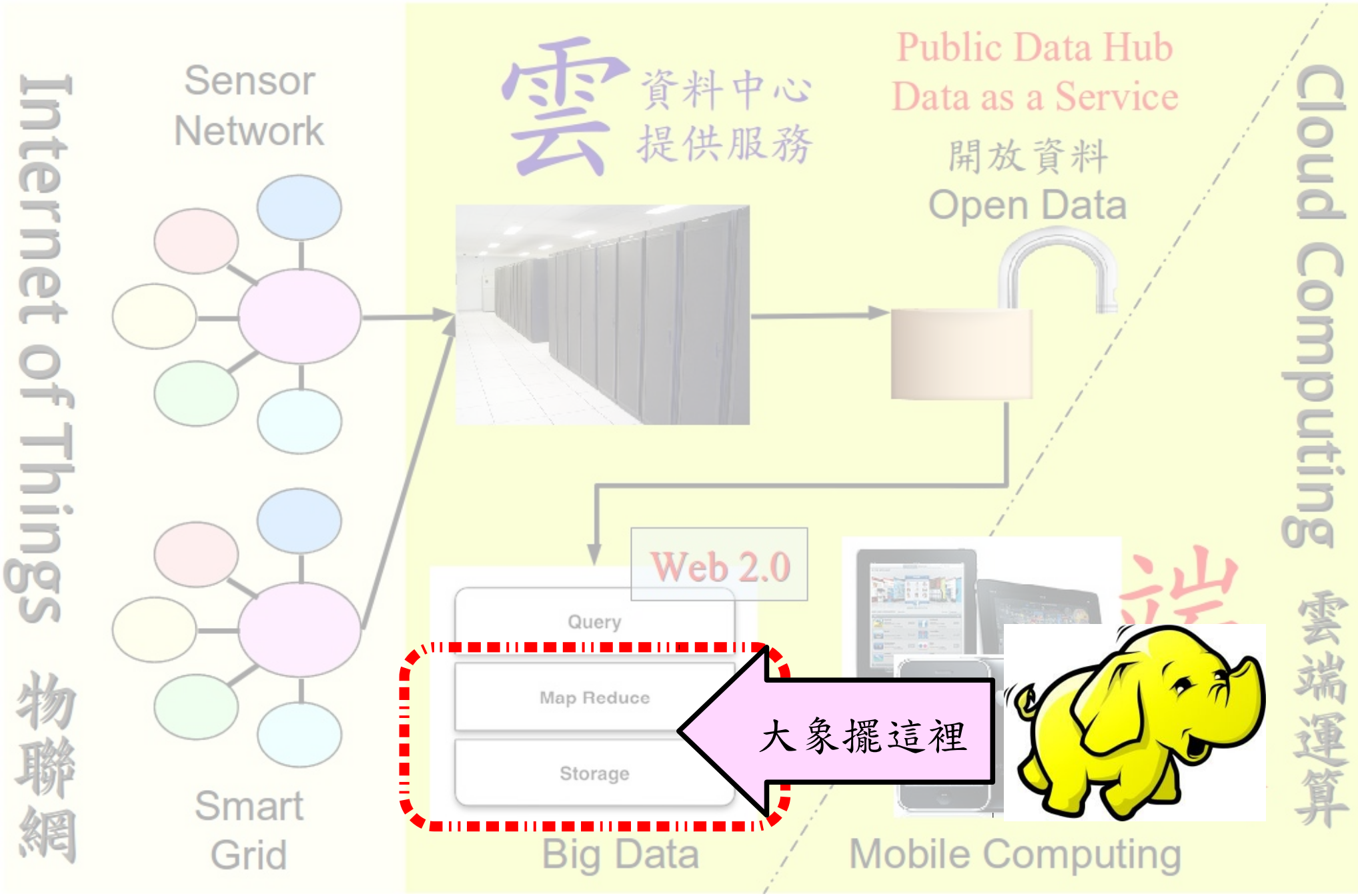
Big Data

端 各類裝置 存取服務

Mobile Computing

Cloud Computing 雲端運算

巨量資料的社群關聯 Relations between Communities



Hadoop related Facebook Group in Taiwan

Hadoop.TW

https://www.facebook.com/groups/hadoop.tw

2013/03/22 Hadoop.TW 533 members

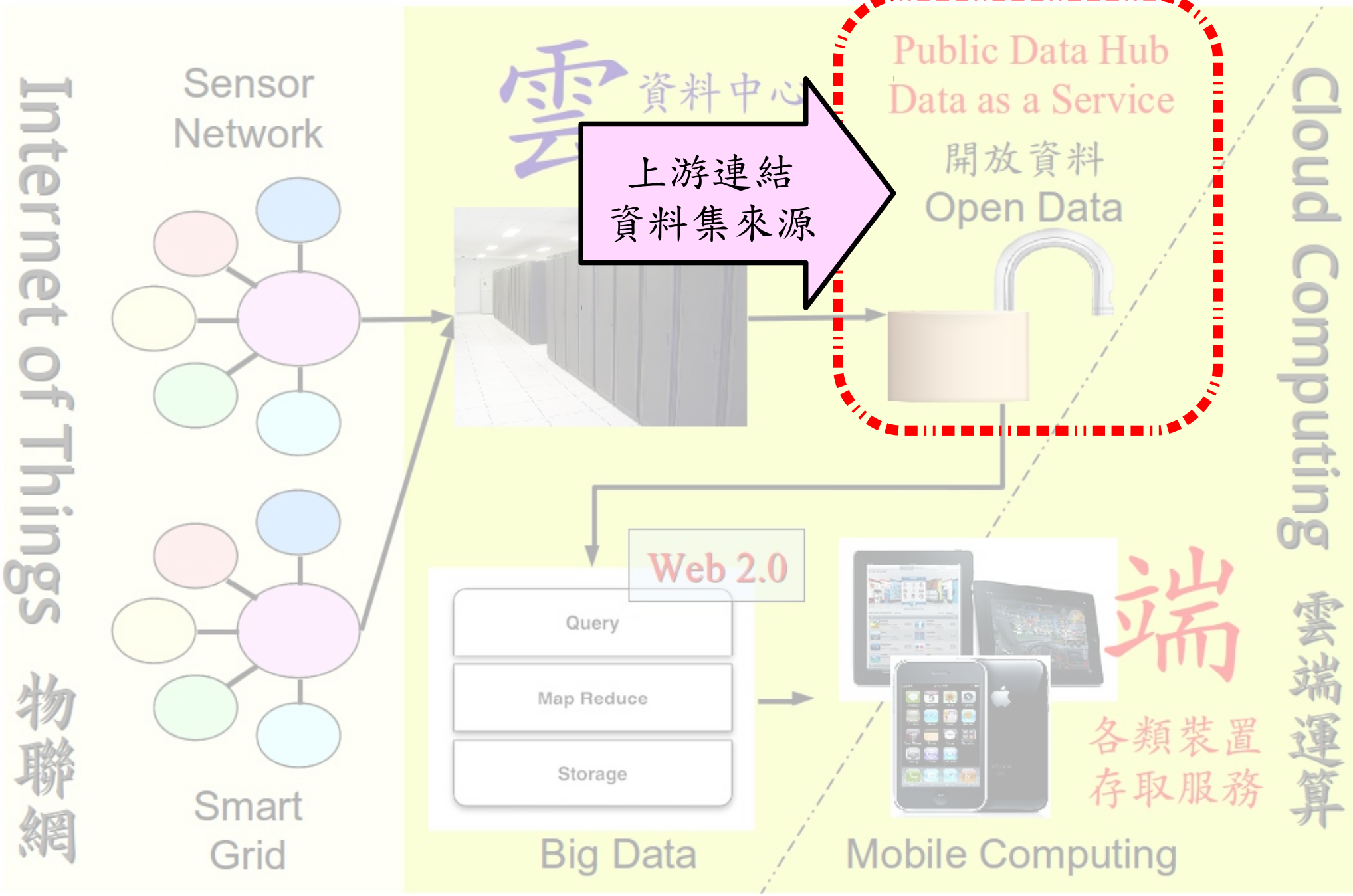
Hadoop related Facebook Group in Taiwan

Hadoop in Taiwan

https://www.facebook.com/groups/hadoopintaiwan/

2013/03/22 Hadoop in Taiwan 59 members

巨量資料的社群關聯 Relations between Communities



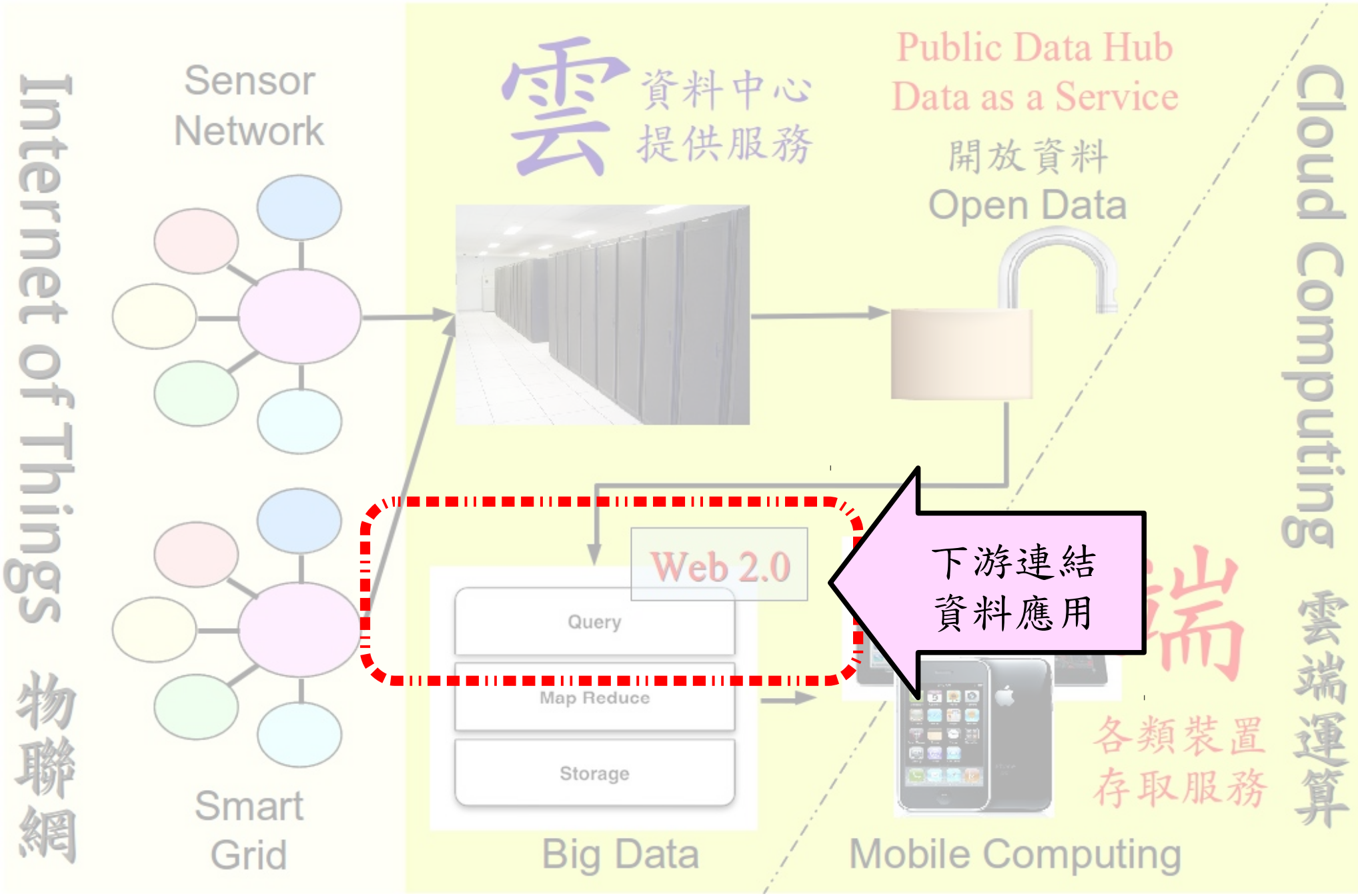
Open Data Taiwan Facebook Group

OpenData / Taiwan (ODTWN)

https://www.facebook.com/groups/odtwn/

2013/03/22 Open Data Taiwan 170 members

巨量資料的社群關聯 Relations between Communities



NoSQL Taiwan Facebook Group

NoSQL Taiwan

https://www.facebook.com/groups/306552142710977/

2013/03/22 NoSQL Taiwan 683 members

NoSQL & Big Data Architecture

NoSQL & BigData Architecture

https://www.facebook.com/groups/423848814337101/

2013/03/22 NoSQL-BigData 92 members

HBase Taiwan Facebook Group

HBase 小聚

https://www.facebook.com/groups/289369481132604/

2013/03/22 HBase.TW 89 members

MongoDB Taiwan Facebook Group

MongoDB 小聚

https://www.facebook.com/groups/142553245867411/

2013/03/22 MongoDB.TW 114 members

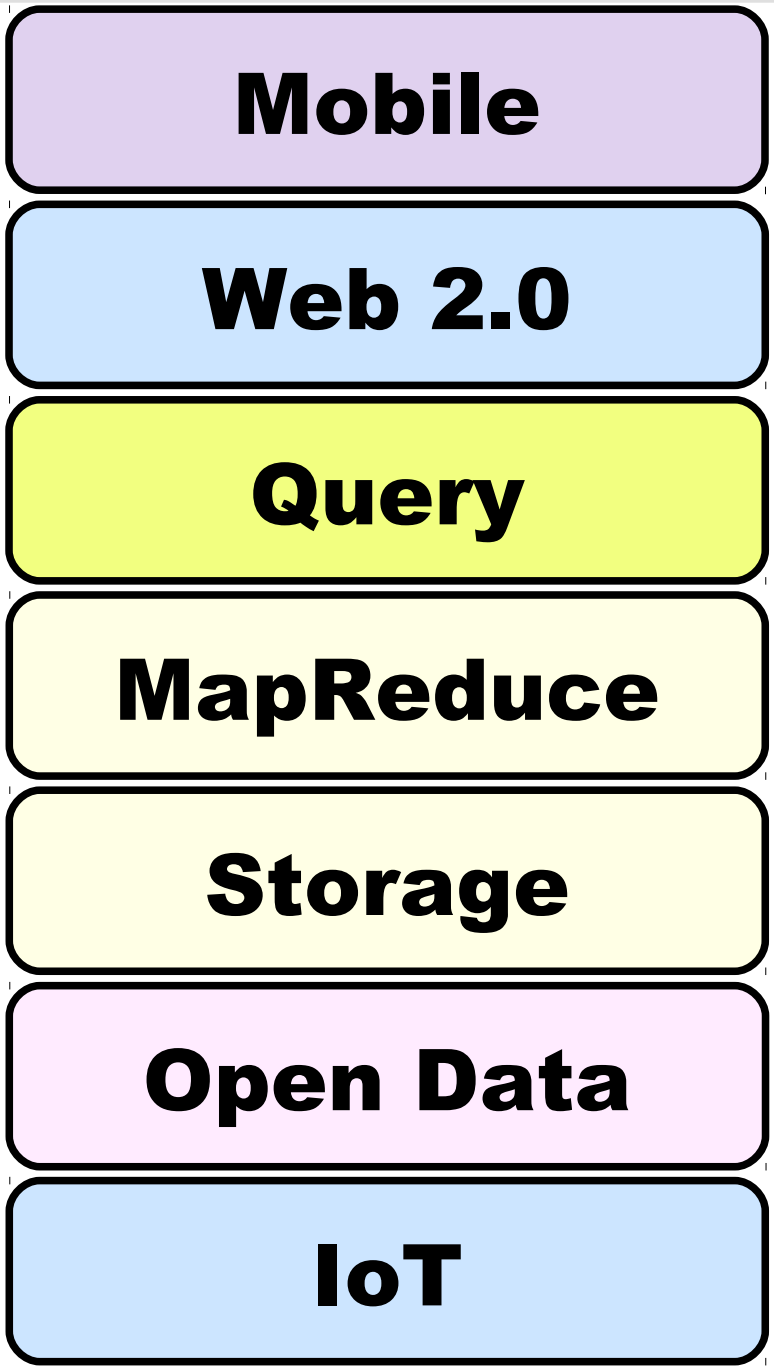
JavaScript Taiwan Facebook Group

JavaScript.TW

https://www.facebook.com/groups/javascript.tw/

2013/03/22 JavaScript.TW 2177 members

Supply Chain of Big Data Industry



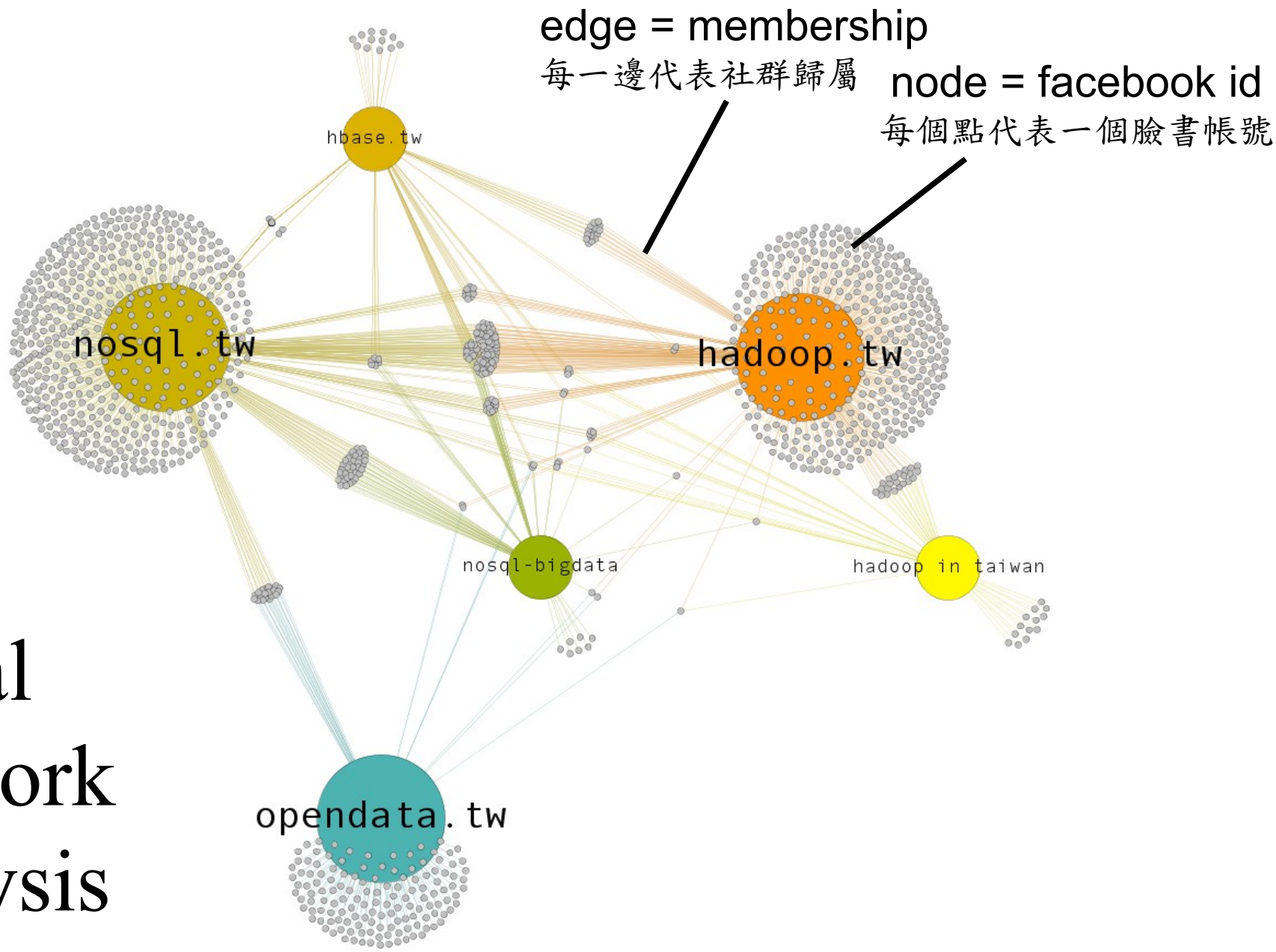
JavaScript.TW 2177 members

NoSQL Taiwan 683 members

Hadoop Taiwan 533 members

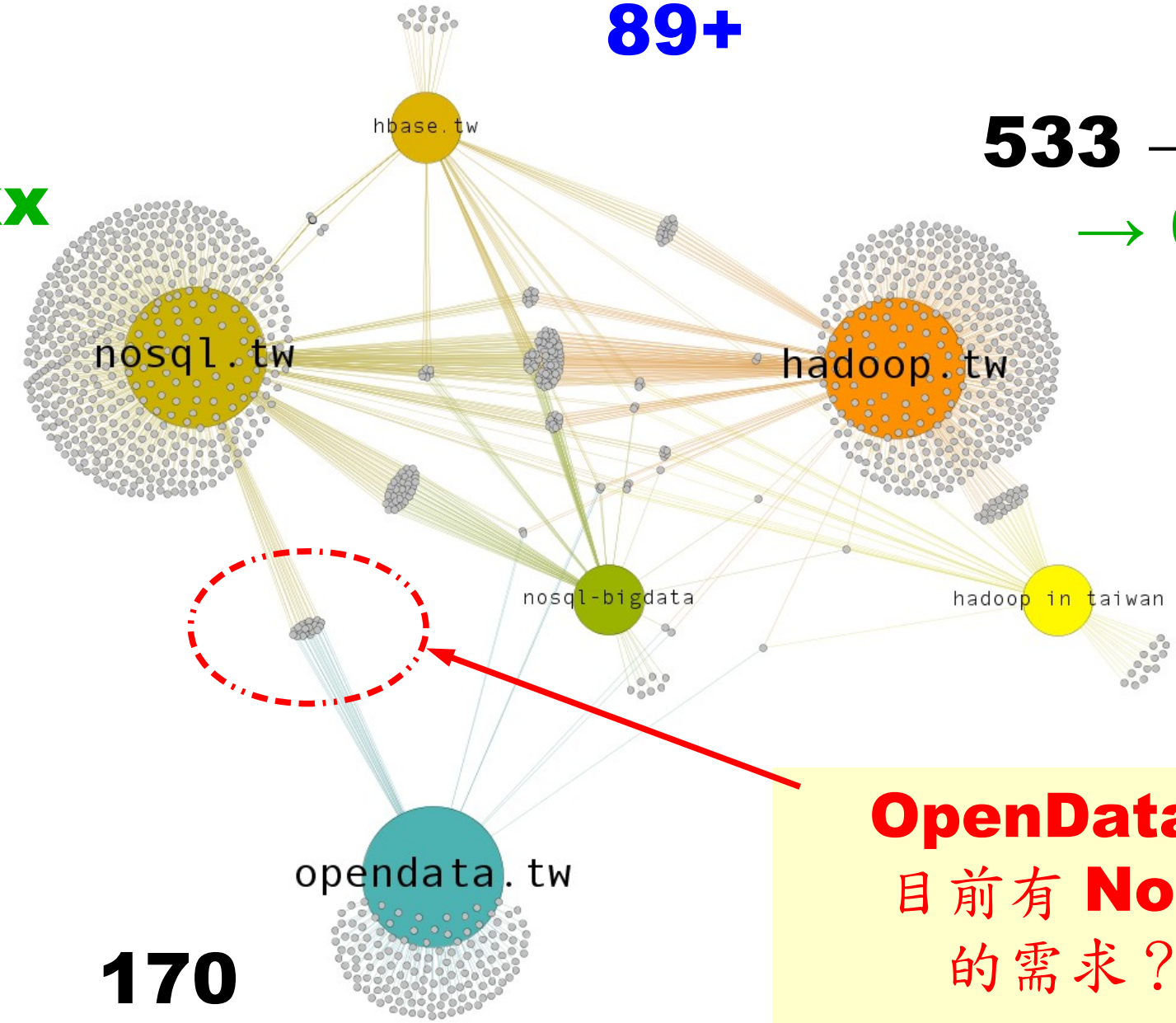
Open Data Taiwan 170 members

相關社群的群聚分析　SNA of Communities



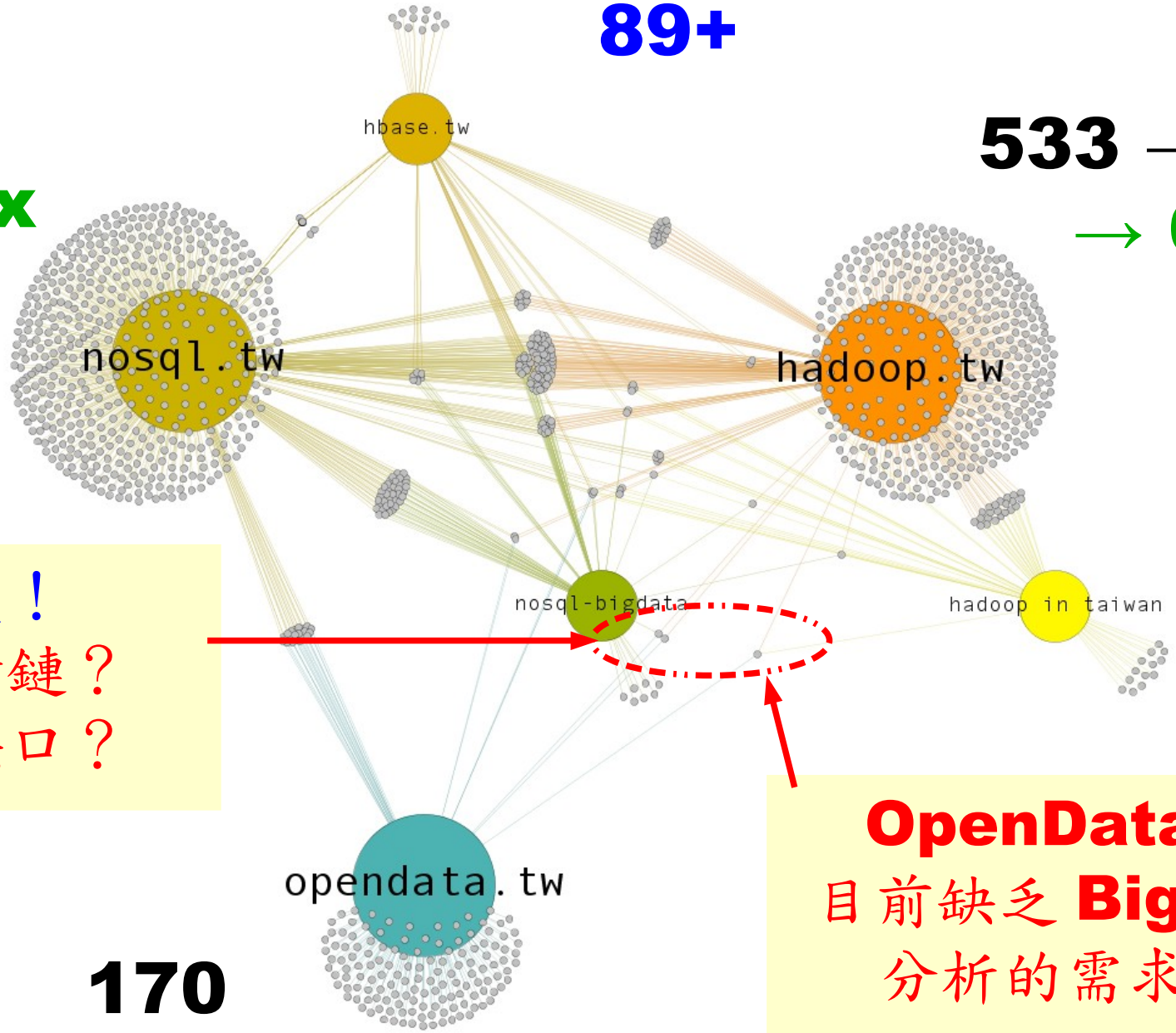
Social
Network
Analysis

從社群談未來的展望 Actions driven from SNA



OpenData.TW
目前有 **NoSQL**
的需求？！

從社群談未來的展望 Actions driven from SNA

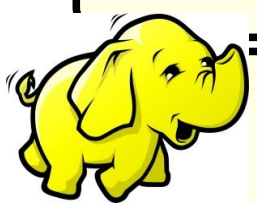


隱憂！
供應斷鏈？
人力缺口？

OpenData.TW
目前缺乏 **Big Data**
分析的需求？！

結　語　Conclusion

- Mobile
- Web 2.0
- Query
- MapReduce
- Storage
- Open Data
- IoT



紅海

需求明確
市場飽和

JavaScript.TW 2177 members

缺乏成功案例

NoSQL Taiwan 683 members

還算健康
持續互動

Hadoop Taiwan 533 members

藍海

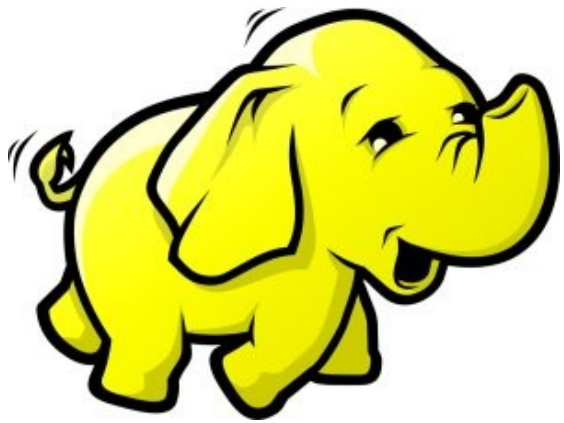
設法找出需求
強化供應鏈

Open Data Taiwan 170 members

有待法令開放

海量資料產業尚未成型
仍有待打通產業供應鏈

Questions?

Slides - http://trac.nchc.org.tw/cloud

Jazz Wang
Yao-Tsung Wang
jazz@nchc.org.tw