

SQL:疏理資料好幫手

大數據中心系列課程



陽明大學 台北榮總家庭醫學科陳育群

收納 整理 利用
資料的方法與工具

SQL

常用的統計不過 8 種....

T 檢定

線性迴歸

卡方檢定

羅吉斯回歸

ANOVA

K-M 分析

無母數

存活分析



ANALYTICS IN HEALTHCARE*

醫療照護正經歷著病患資料爆炸性成長，但是只有相關醫療機構可以從中找出有用資訊，並讓這些資料發揮真正價值



1. 自行收集資料

#問卷調查 #實驗動物 #抽血分析

2. 使用現有資料

#健保資料庫 #衛福部加值中心

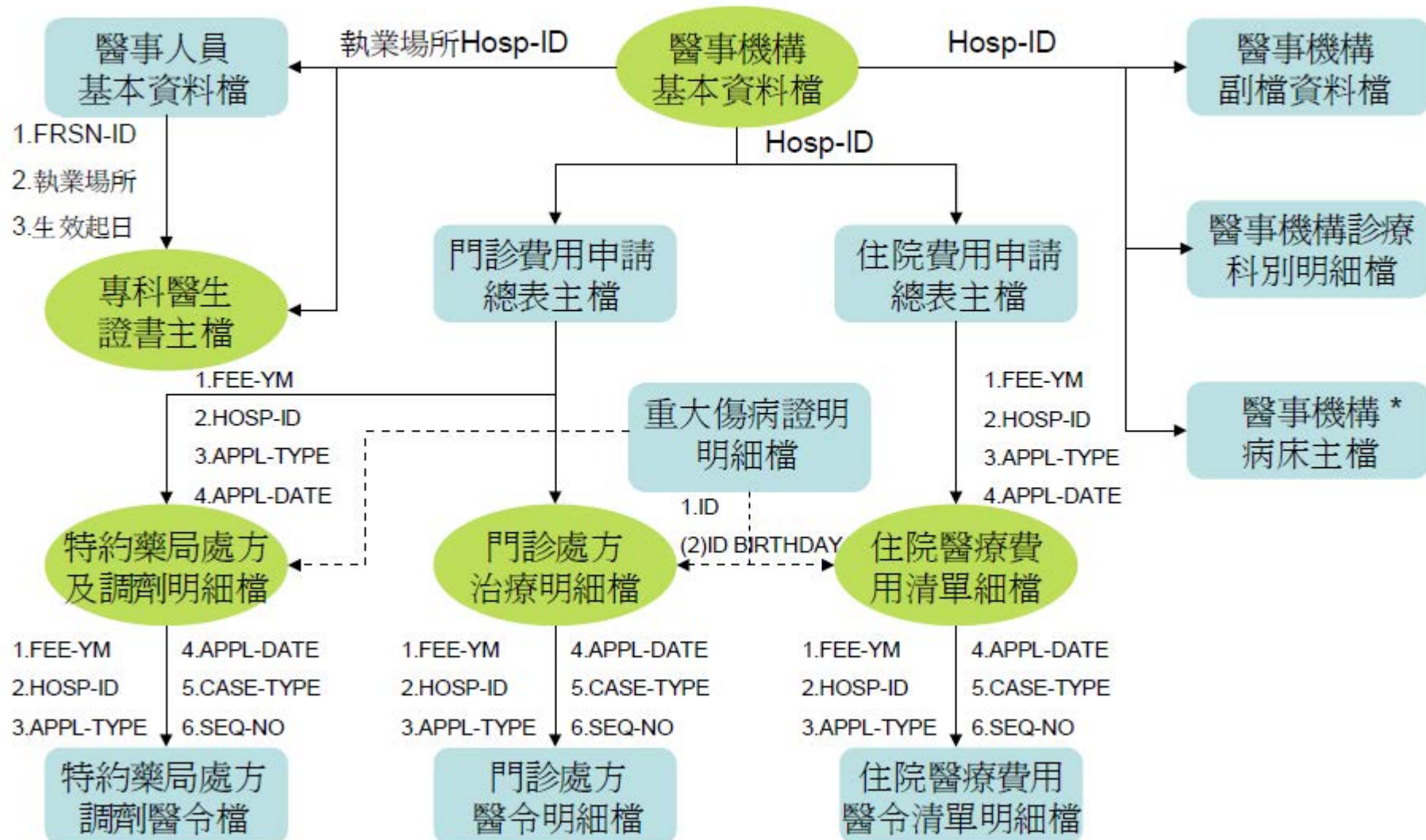
1. 自行收集資料

1.5 整理病歷資料

2. 使用現有資料

資料庫(database)通常由多表格組成

串檔變項說明 (KEY-Relationship)



資料整理

- Table -- 最簡單、直覺的方式
 - 常見於 紙本, 大部分軟體 (EXCEL, SPSS, STATA, MedCalc, R)

TABLE (表格)
/

病患身分證號	生日	性別	住址	住院後追蹤
~~~~~	~~~~~	~~~~~	~~~~~	~~~~~
~~~~~	~~~~~	~~~~~	~~~~~	~~~~~
~~~~~	~~~~~	~~~~~	~~~~~	~~~~~
~~~~~	~~~~~	~~~~~	~~~~~	~~~~~


:: 資料儲存問題 (wide form)

- 常見問題: 沒有彈性, 追蹤時間易混淆

[illegible]

資料儲存- (long form, indexed form)

- 如果追蹤次數不一定怎麼辦？

wide form →

病患身分證號	生日	性別	住址	追蹤1	追蹤2	追蹤3
~~~~~	~~~~~	~~~~~	~~~~~	~~~~~		
-----	-----	-----	-----	-----1	-----2	-----3
=====	=====	=====	=====	=====1	=====2	
+++++	+++++	+++++	+++++			

long form, indexed data ↓

病患身分證號	生日	性別	住址	追蹤
~~~~~	~~~~~	~~~~~	~~~~~	~~~~~
-----	-----	-----	-----	追蹤1
-----	-----	---X---	-----	追蹤2
-----	-----	-----	-----	追蹤3
+++++	+++++	+++++	+++++	

資料儲存- (long form, indexed form)

- 常見問題:
- 容易出錯不一致(inconsistency)

病患身分證號	生日	性別	住址	追蹤
~~~~~	~~~~~	~~~~~	~~~~~	~~~~~
-----	-----	-----	-----	追蹤1
-----	-----	-----	-----	追蹤2
-----	-----	-----	-----	追蹤3
+++++	+++++	+++++	+++++	



# 關聯式資料概念 (Relational database)

病患身分證號	生日	性別	住址	追蹤	追蹤	追蹤
~~~~~	~~~~~	~~~~~	~~~~~	~~~~~		
-----	-----	-----	-----	-----1	-----2	-----3
=====	=====	=====	=====	=====1	=====2	
+++++	+++++	+++++	+++++			

wide form

病患身分證號	生日	性別	住址	追蹤
~~~~~	~~~~~	~~~~~	~~~~~	~~~~~
-----	-----	-----	-----	-----1
-----	-----	-----	-----	-----2
-----	-----	-----	-----	-----3
+++++	+++++	+++++	+++++	

long form, indexed data

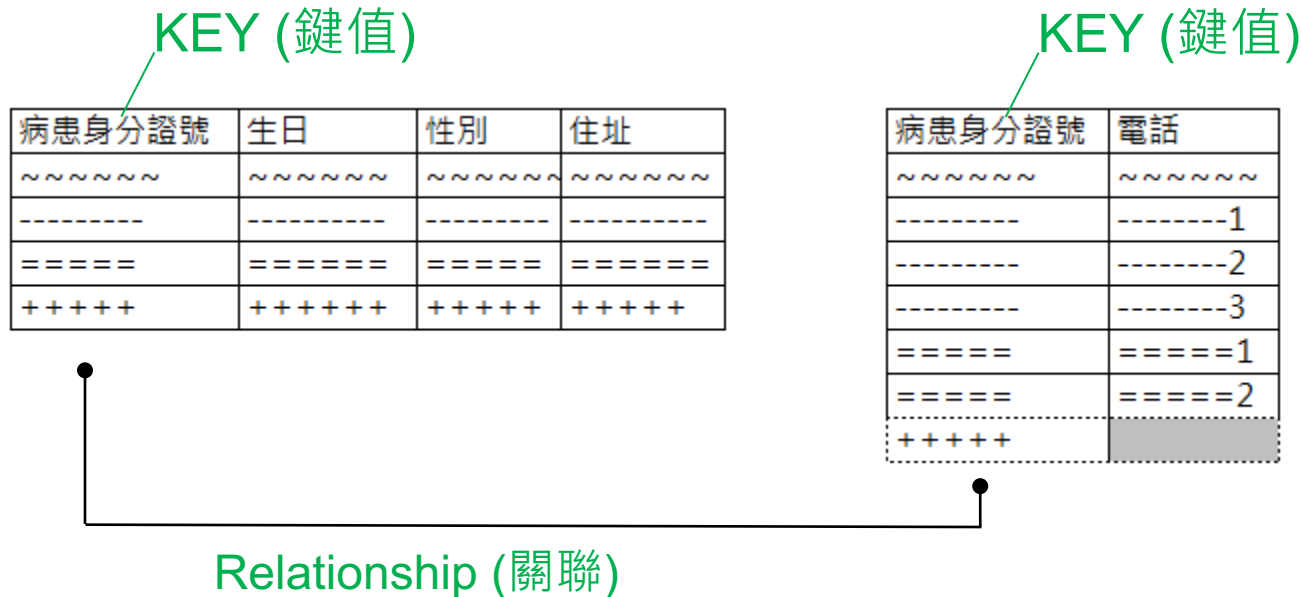
結合兩者的優點: 關聯式資料 – 節省空間、具彈性、保持一致性

病患身分證號	生日	性別	住址
~~~~~	~~~~~	~~~~~	~~~~~
-----	-----	-----	-----
=====	=====	=====	=====
+++++	+++++	+++++	+++++

病患身分證號	追蹤
~~~~~	~~~~~
-----	-----1
-----	-----2
-----	-----3
=====	=====1
=====	=====2
+++++	

# 關聯式資料概念 (Relational database)

關聯式資料 – 節省空間、具彈性、保持一致性

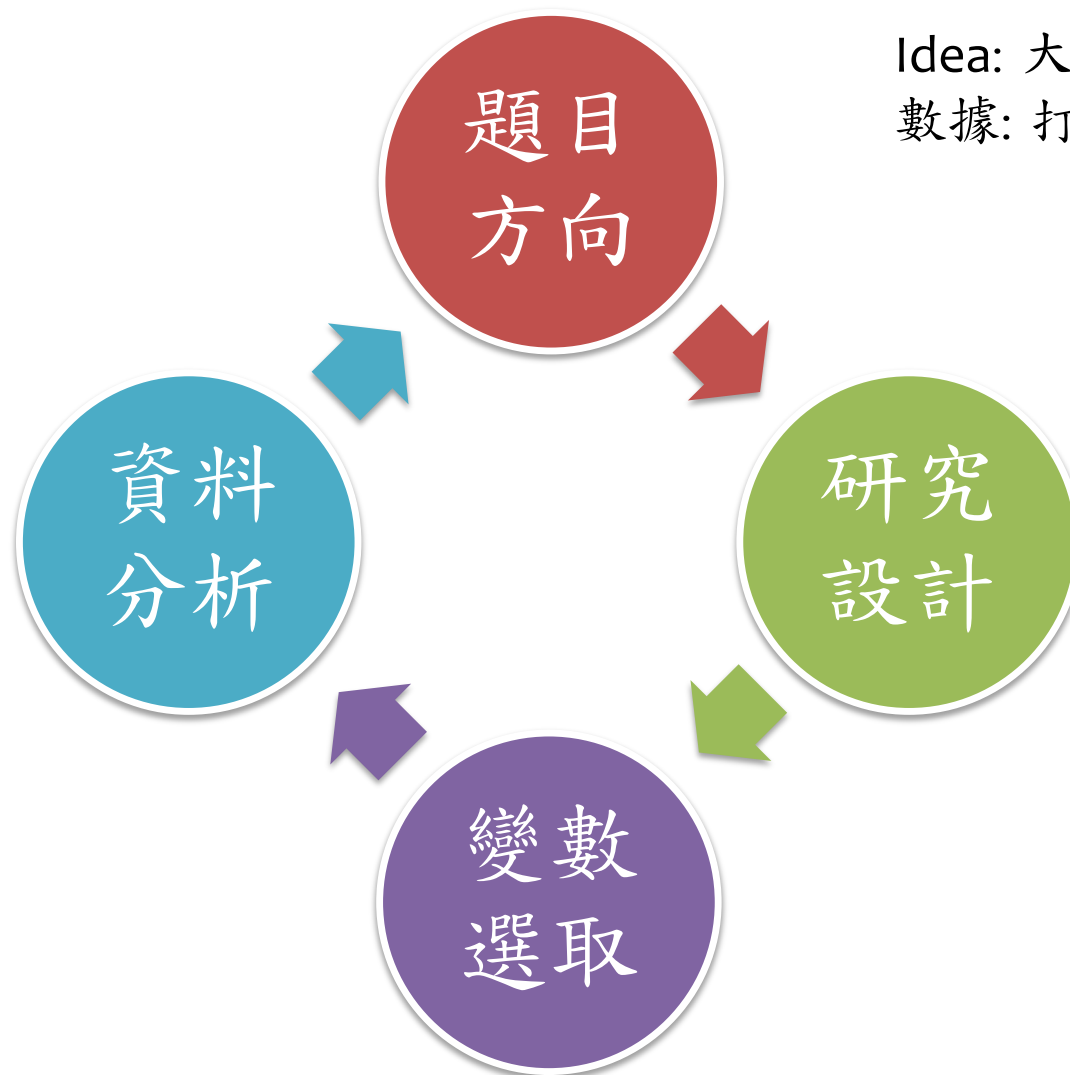


常見問題: 如何串連多個表單? (串檔)

- copy-paste
- EXCEL – vlookup, PowerPivot
- SQL, Pro MERGE

資料

# :::研究=資料收納分析

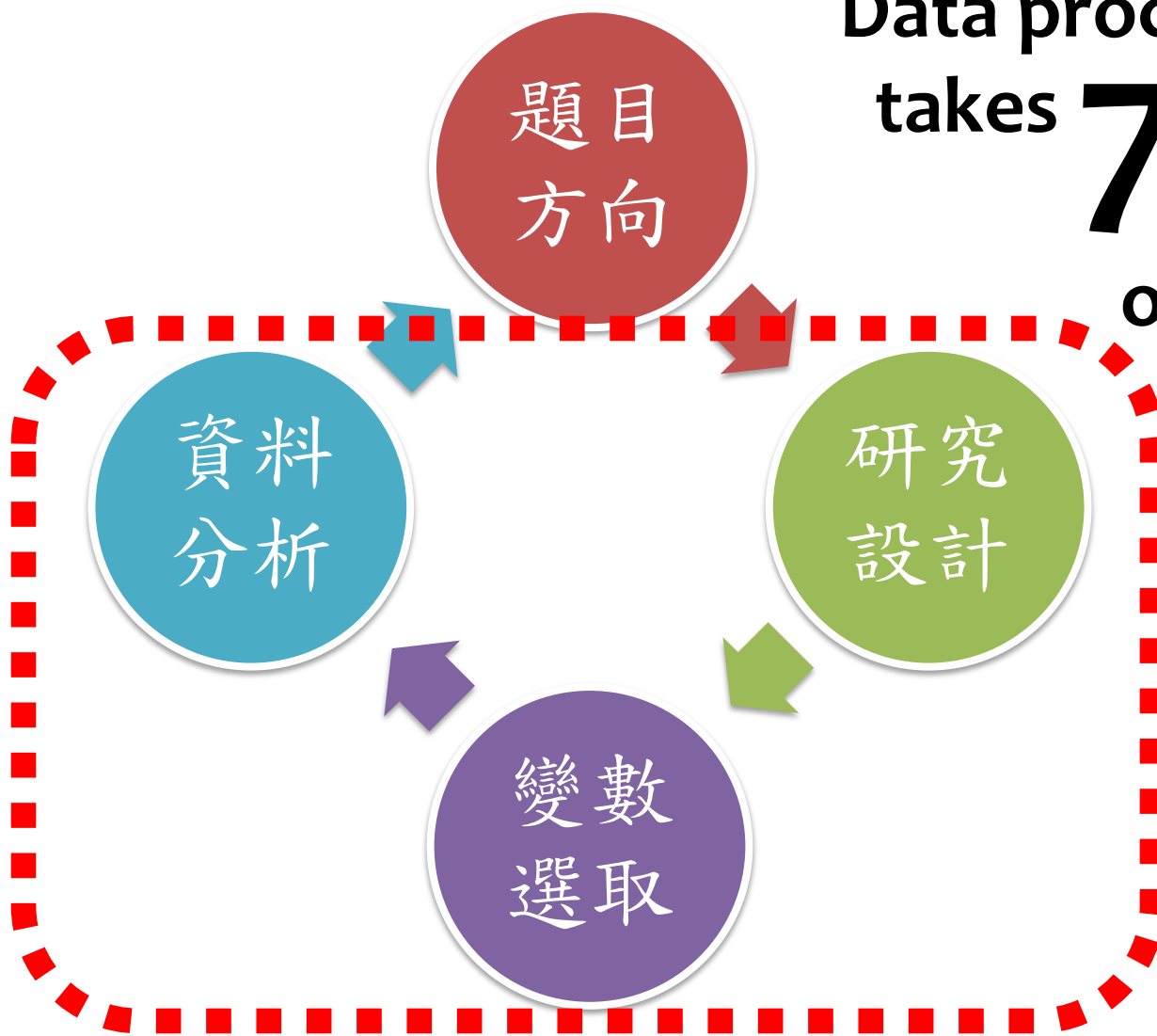


Idea: 大家都有

數據: 打開電腦就可以跑

# 資料性研究

Data processing  
takes **75%**  
of the task



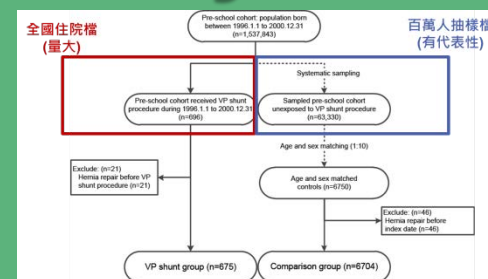


# ::: A journey from idea to result

## 1. Idea



## 2. Study



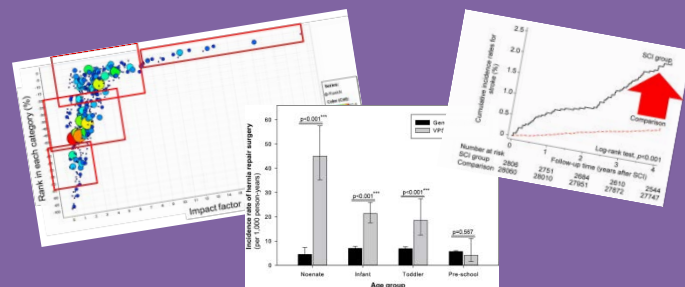
DATA

## 3. Data

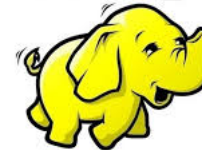
- (1)
- (2)
- (3)



## 4. Result



# ::: 常使用於資料庫研究分析工具



## 統計軟體

- 👍 統計檢定
- 👎 資料處理彈性不足

## EXCEL

- 👍 親善好用—大家的好朋友
- 👎 串連 (join)

## POWERPIVOT

- 👍 親善、可串連, Join (DAX)
- 👎 跨年追蹤 (Union)

## SQL

- 👍 強大串連跨年追蹤
- 👎 學習、安裝 (殺雞用牛刀)

## 統計軟體

- 👍 統計檢定
- 👎 資料處理彈性不足



## EXCEL

- 👍 親善好用—大家的好朋友
- 👎 串連 (join)

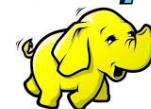
## POWERPIVOT

- 👍 親善、可串連 (DAX)
- 👎 跨年追蹤 (Union)



PostgreSQL

hadoop



## SQL

- 👍 強大串連跨年追蹤
- 👎 學習、安裝 (殺雞用牛刀)



Microsoft®  
SQL Server®



斷捨離

Patients

Collections  
(claims, visits,  
consultations)

Contexts  
(provenance)

Clinical Codes  
(patient, dates, code,  
sequence)

Measurement  
Details

Drug Exposure  
Details

Costs

Payer  
Reimbursements

Contexts  
Practitioners

Admission  
Details

Facilities

Information  
Periods

Patient Details

Deaths

# Common Data Model

## 通用醫療資料模型

### Lookup Tables

Vocabularies

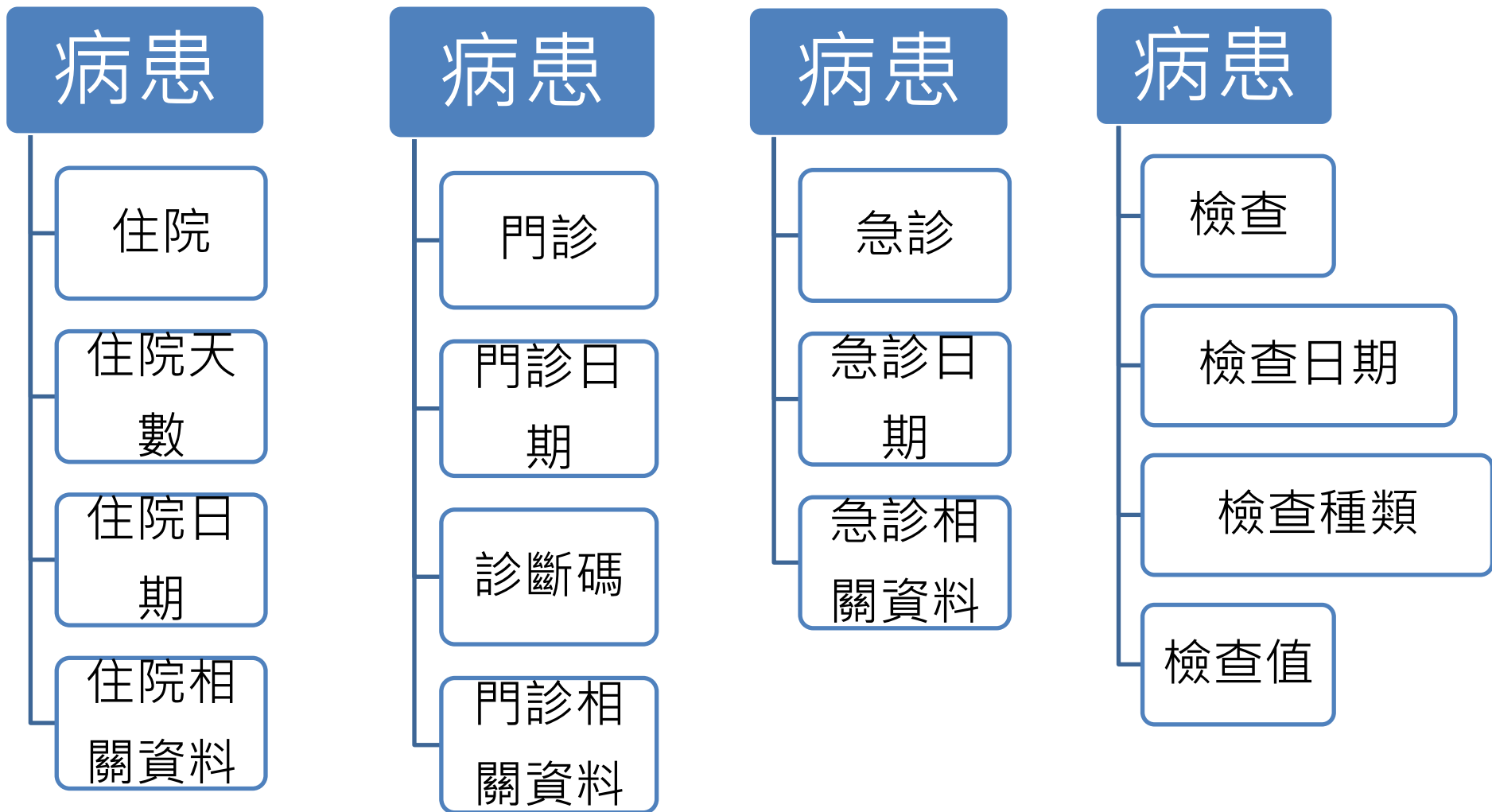
Concepts

Mappings

Practitioners

Addresses

# 斷：將資料依照種類切分 存放



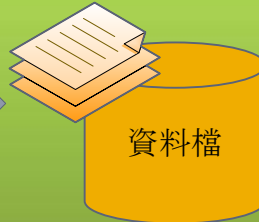




斷捨離

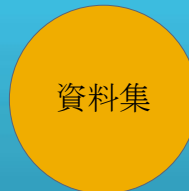


# Three main components of data processing



## 1. 資料轉換

READ



## 2. 變數處理

SELECT, FILTER  
APPEND, UNION  
JOIN, SORT



## 3. 結果分析

AGGREGATE  
OUTPUT  
STATISTICS

# ∴ 1. 資料轉換



- 由資料來源讀取出資料，將它們轉換成適合分析的型態，並且將它們匯入資料庫內。
- 通常還要搭配著資料清潔 ( **Data Cleaning** ) 將系統源頭許多未經整合的、不允許的、遺失的或者錯誤的資料，在匯入之前重新整頓 (Garbage in Garbage out) 。

# 確保資料正確

a1587_00200412.DAT x			
200	36424A015279209	100TID	0000000.0000001.500000000↓
200	36421M03	100	0000075.0000001.000000075↓
200	31084A022124100	200HS	0000000.0000014.000000000↓
200	31084A009250500	100PRN	0000000.0000004.000000000↓
200	31084A022105100	100QID	0000000.0000012.000000000↓
200	08164B011254100	100TID	0000000.0000009.000000000↓
200	08164A016489100	100TID	0000000.0000009.000000000↓
200	08164A014277100	100TID	0000000.0000009.000000000↓
200	08164A020570100	100TID	0000000.0000009.000000000↓
200	08161M03	001	0000075.0000001.000000075↓
200	00274A027040100	300TID0X 3	0000000 0000000 000000000↓
200	00274A027040100	300TID0X 3	0000000 0000000 000000000↓
200	03931M03	100100	0000075.0000001.000000075↓
200	03934A040130100	300TID	0000000.0000009.000000000↓
200	03934A021981499	300Q6H	0000000.0000003.000000000↓
200	03934A026643100	150TID	0000000.0000004.500000000↓
200	03934N000168100	300TID	0000000.0000009.000000000↓
200	36244PTS5		0000000.0000001.000000000↓
200	36244PTS3		0000000.0000001.000000000↓
200	3624242006C		0000190.0000001.000000190↓
200	36244PTS9		0000000.0000001.000000000↓
200	36244PTS8		0000000.0000001.000000000↓
200	36241A040358345	100QD	0000015.0000008.000000120↓
1.81 GB (1,946,006,640 bytes), 16,216,723 lines.			
Text Ln 4865021, Col 37			

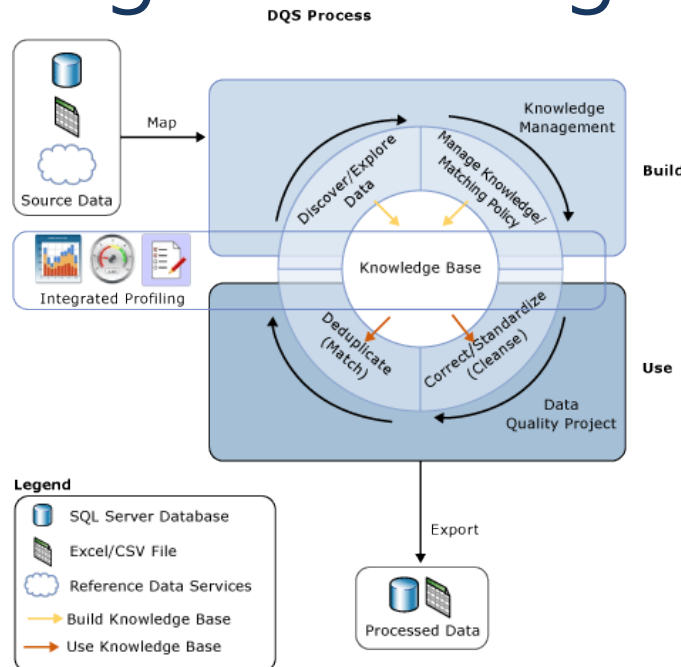
Line: 4,865,021

$\frac{1}{16,216,723}$  的錯誤

# Data Quality Service: 清理資料

- 不正確的資料可能是因為使用者輸入錯誤、傳輸或儲存中損毀、不符的資料字典定義，以及其他資料品質與處理序問題。

## Garbage in Garbage out (GIGO)



### 需要清理資料:

- 藥物頻次
- 不合理的ICD9 coding
- 病患生日
- 病患性別
- 病患居住地...等等

## 2. 變數處理



- 把分別儲存於不同表格的原始資訊，如醫院層級、藥物分類、疾病分類、重大傷病等等「串連」許多個資料表。
- 選取適當的變數(proxy variable)，進行資料加值。
- 敏感度測試。



斷捨離

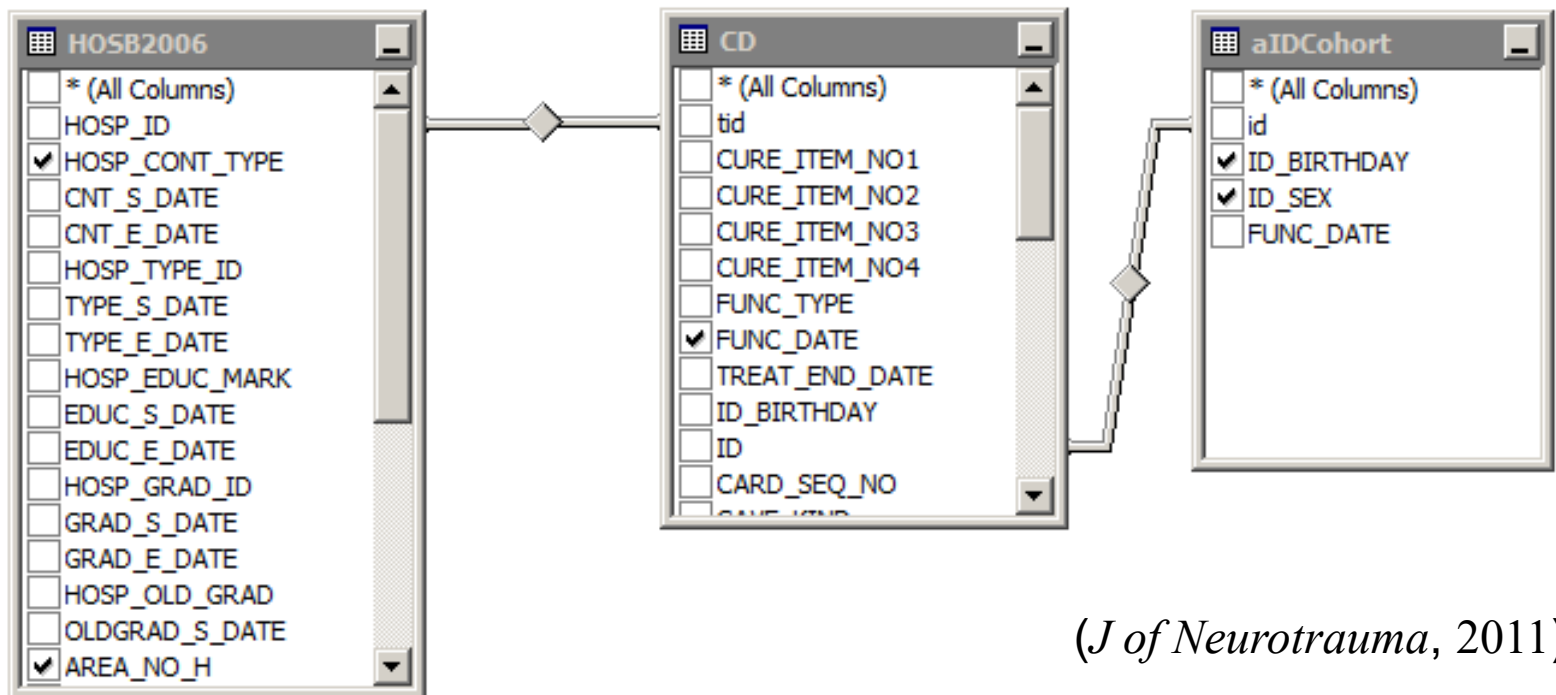
# ∴ 3. 結果分析



- 把整理好的資料，以表格方式 (tabular) 輸入分析軟體 (資料探勘/統計/繪圖軟體)，呈現有意義的結果發現。

# 資料串連加值: 通常是最重要/耗時的步驟

- 已知某個族群(ID_Cohort)，想了解這個族群在2000年到2007年的就診情形(就診科別、醫院層級、醫院所在地點、看診日期)

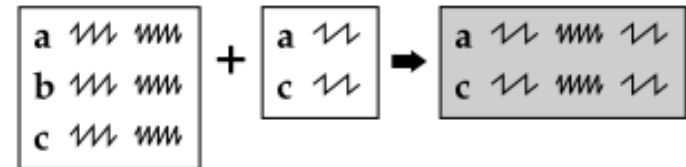


(J of Neurotrauma, 2011)

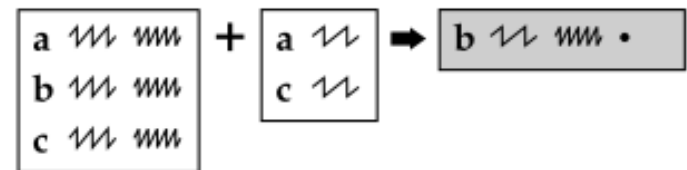


# ::: Basic idea for data join / merge

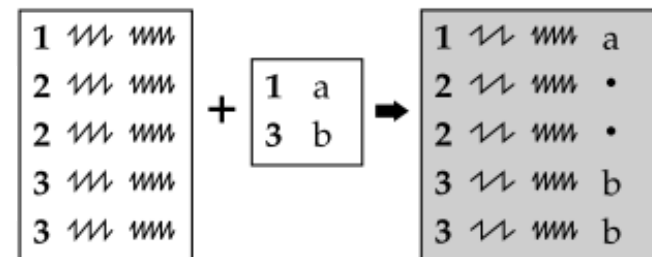
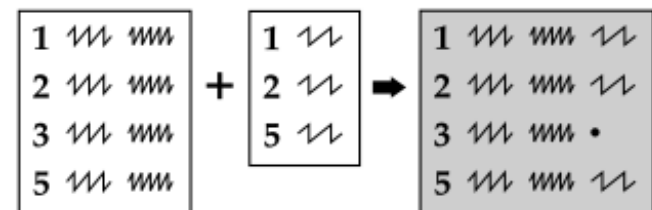
1. Matching obs.



2. Non-matching obs.



3. Combined (Matching + non-matching)

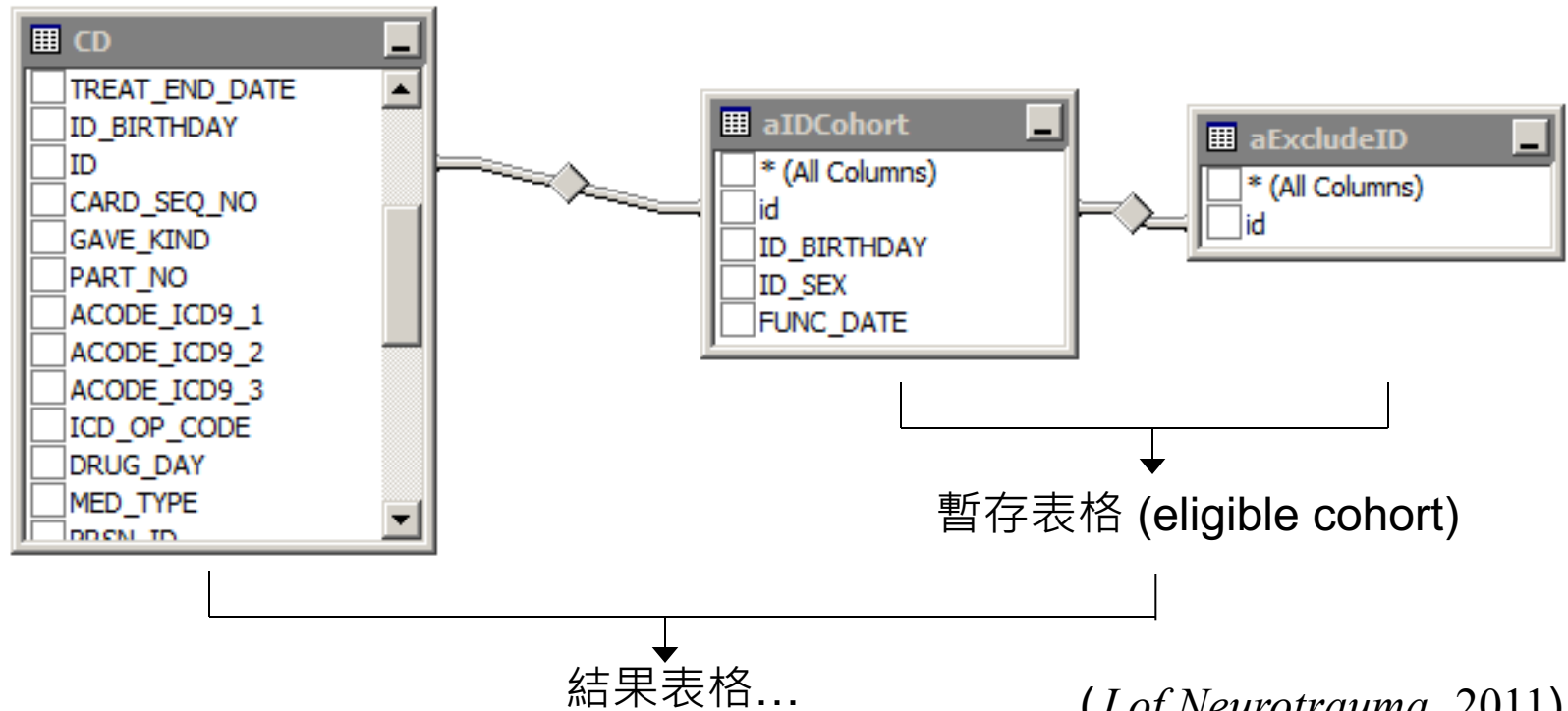


```
DATA new-data-set;
  MERGE data-set-1 data-set-2;
  BY variable-list;
```

```
DATA both;
  MERGE state (IN = InState) county (IN = InCounty);
  BY StateName;
```

# ∴ 資料串連加值:

- 已知某個族群(ID_Cohort)，排除掉(exclusion criteria)，想抓取該族群在2000年到2007年內第一次就診科別

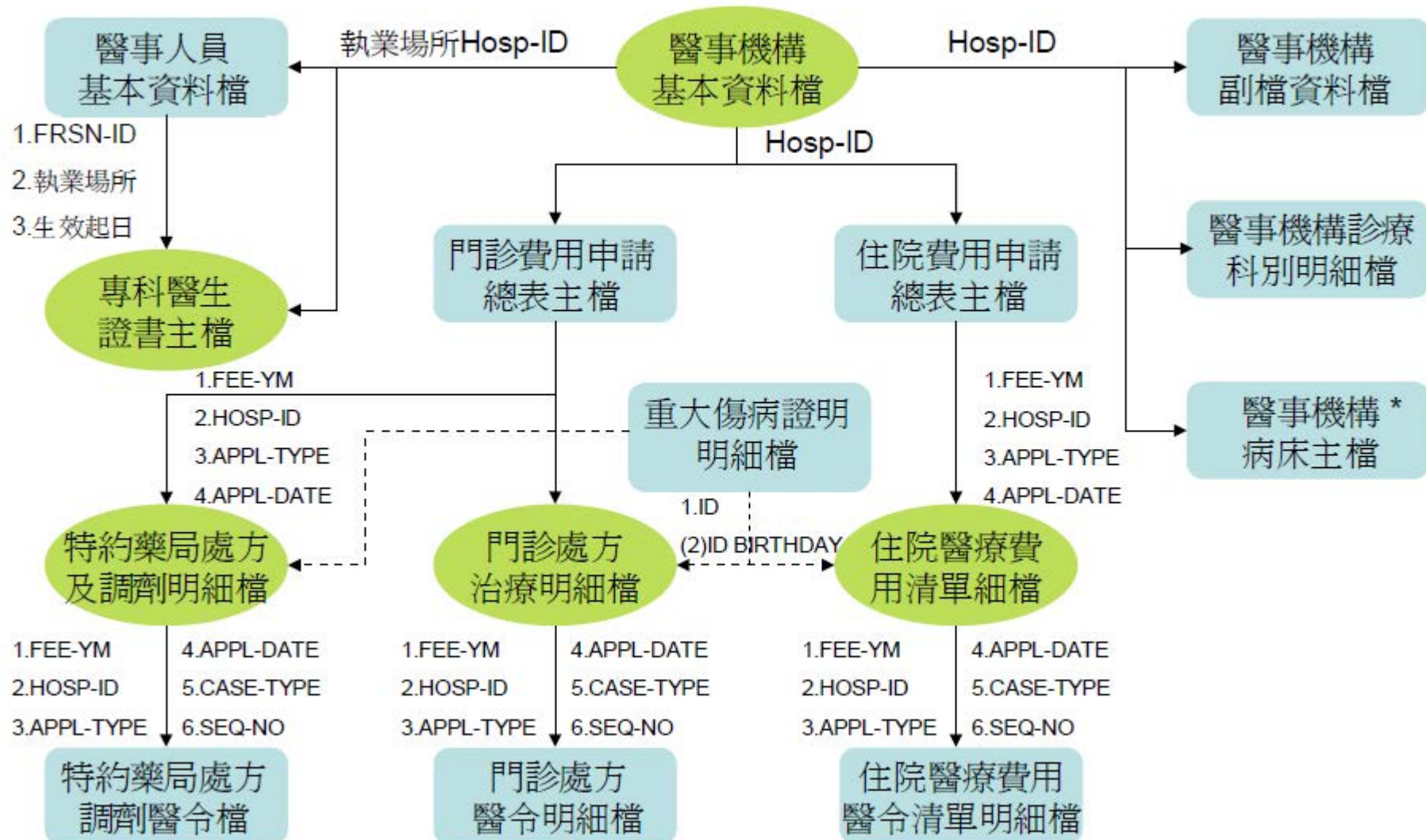


## ∴ 資料串連加值:

- 已知某個族群(ID_Cohort)，扣掉不符合的案例(exclusion criteria)，想了解2000年到2007年內最後一次就診日期及就醫地點科別、治療時期及使用藥物。

# 資料庫(database)通常由多表格組成

## 串檔變項說明 (KEY-Relationship)





# 基本串連功能不足以應付 複雜處理資料需求

- 需要有一種工具能夠儲存大量分散且相關連的資料。
- 能夠有一種指令，像英文一樣直覺將大筆資料互相串連，並擷取所需的部份資料。
- 最好還能夠提供基本的統計加總功能。

→ Invention of RDBMS and SQL

SQL

# 關聯式資料庫管理系統 (RDBMS) 與 結構式查詢語言 (SQL)

1970年代, 因應大量資料處理, IBM 推出:

**RDBMS** (Relational database management system)

- 關聯式資料庫管理系統被發展用來處理/儲存銀行公司等大量錯綜複雜資料。

**SQL** (Structured Query Language)

- 用於大型資料庫中資料的定義/操作/查詢/控制。

→ 成為現代資料庫系統標準。



# SQL:專為資料處理而設計的語言

## → 簡單/易讀/能夠串連大量表格

- Data Definition Language (DDL)
  - 建置資料 / 維護資料 / 定義資料
  - 總共只有三大命令
  - **CREATE**
  - ALTER, DROP [知道即可]
- Data Manipulation Language (DML)
  - 操作 / 查詢 / 控制 資料
  - 總共有四大命令
  - **SELECT**
  - **INSERT**, UPDATE, DELETE [知道即可]



# ::: SQL Syntax

-- 從2009年門診檔中擷取1998年1月1日以後出生的病患紀錄並且依照 生日排序

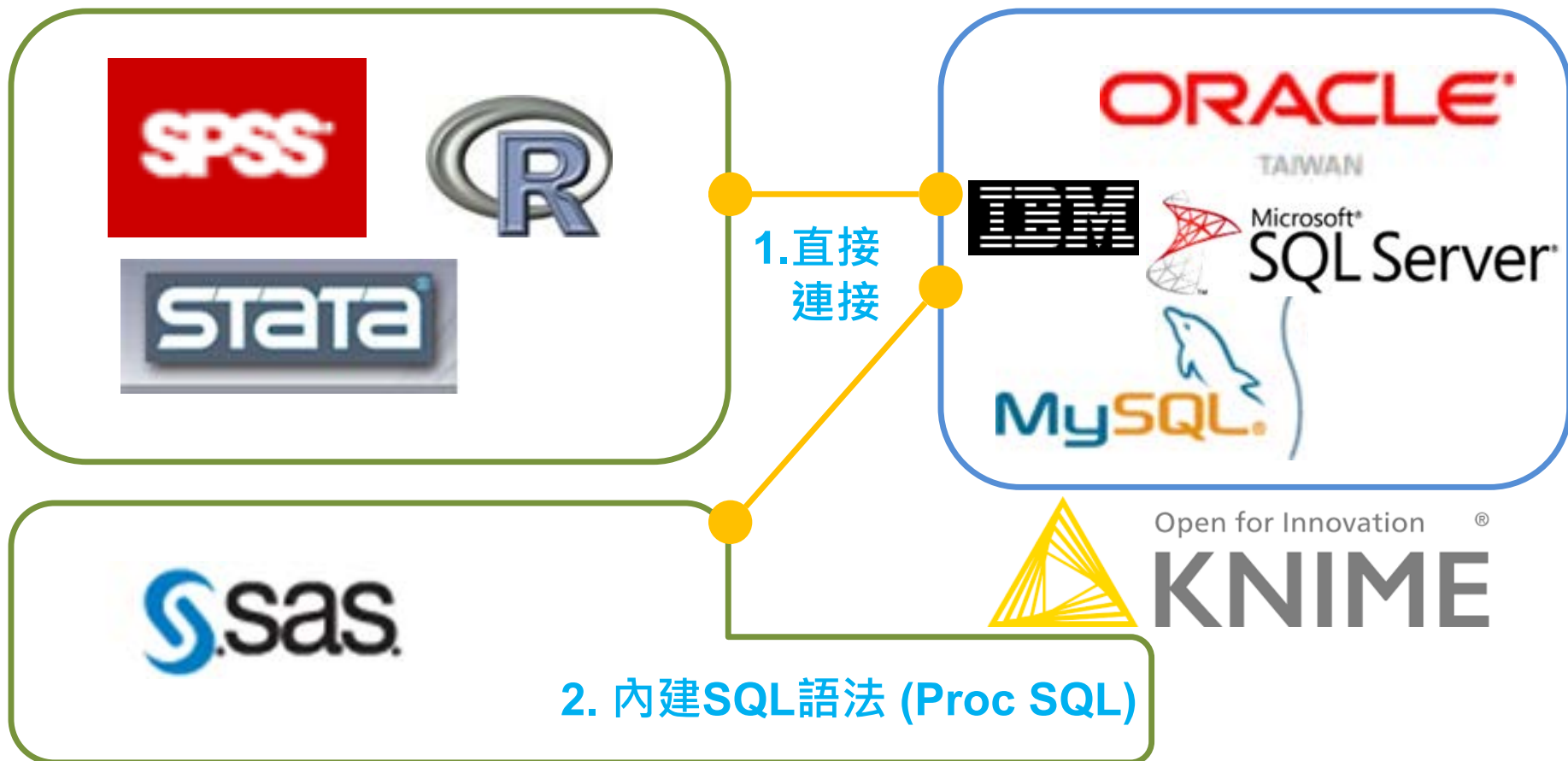
```
SELECT    id, id_sex
FROM      cd2009
WHERE     id_birthday >= '19980101'
ORDER BY id;
```

- SQL statement: 總是由動詞開始, 不分大小寫, 不分換行
- Comment (註解),
- 句後加上分號 (;)

# Current trend: 統計軟體- 結合後端資料庫處理系統

統計軟體 (友善)

資料處理軟體(強悍)



# SQL專注於邏輯

- 已知某個族群(ID_Cohort)，扣掉不符合的案例(exclusion criteria)，想了解2000年到2007年內最後一次就診日期及就醫地點科別、治療時期及使用藥物。

。

自動產生程式碼

