
# 資料探勘:
# 從關聯規則分析到子群組探勘

KNOWLEDGE-BASED ANNOTATION LEARNING SYSTEM

布丁布丁吃布丁

2018/1/2

pudding@nccu.edu.tw

Actionable Rule　　Subgroup Discovery

# 能夠付諸行動的規則:子群組探勘

## 吃藥吃出人命啦！

藥商快想想辦法啊！

### 敘述統計

吃了你藥的病患中,
有10個人
都病得更重了！

### 子群組探勘

吃了你藥的病患中,
**有吸菸的大多數病患們**
都病得更重了！

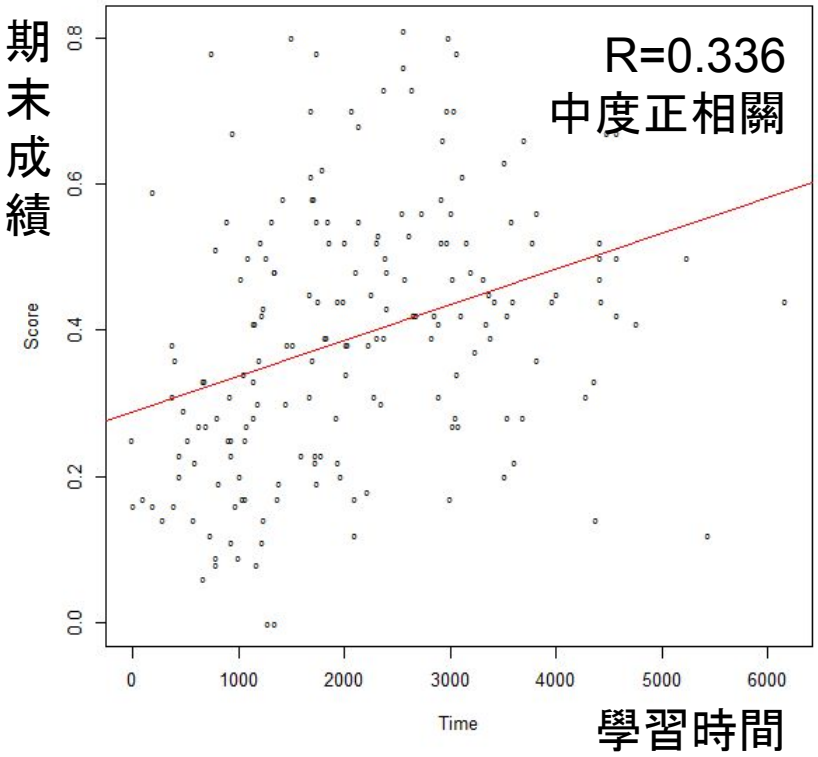
Duivesteijn, W., Feelders, A. J., & Knobbe, A. (2016). Exceptional Model Mining. *Data Mining and Knowledge Discovery*, *30*(1), 47-98. doi:10.1007/s10618-015-0403-4

# 找出隱含在整體資料的潛在規則

子群組探勘

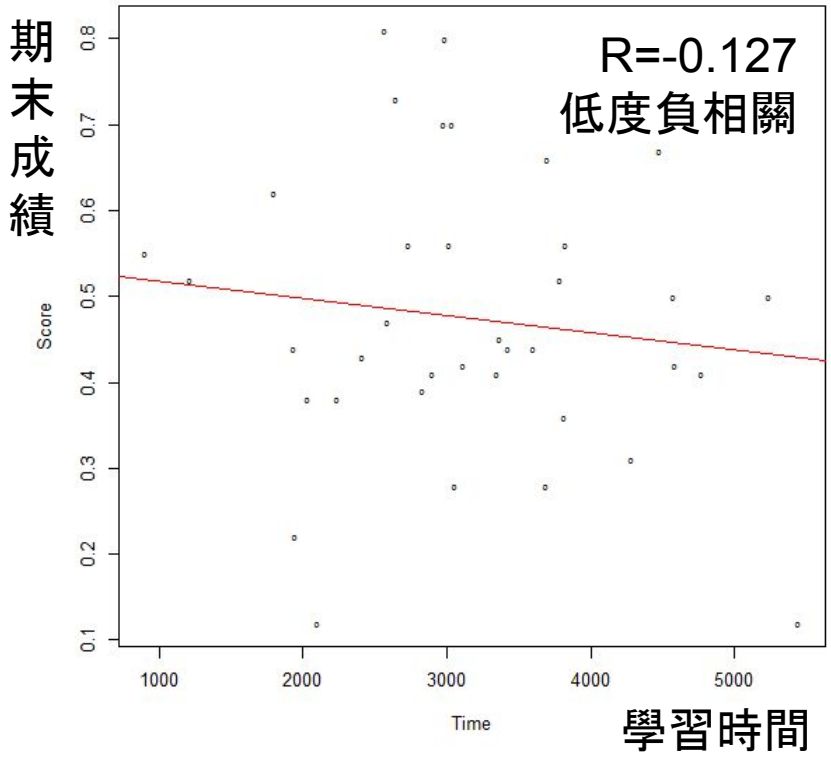
## 整體狀況

學習時間越長，期末成績越好




## 某些課程

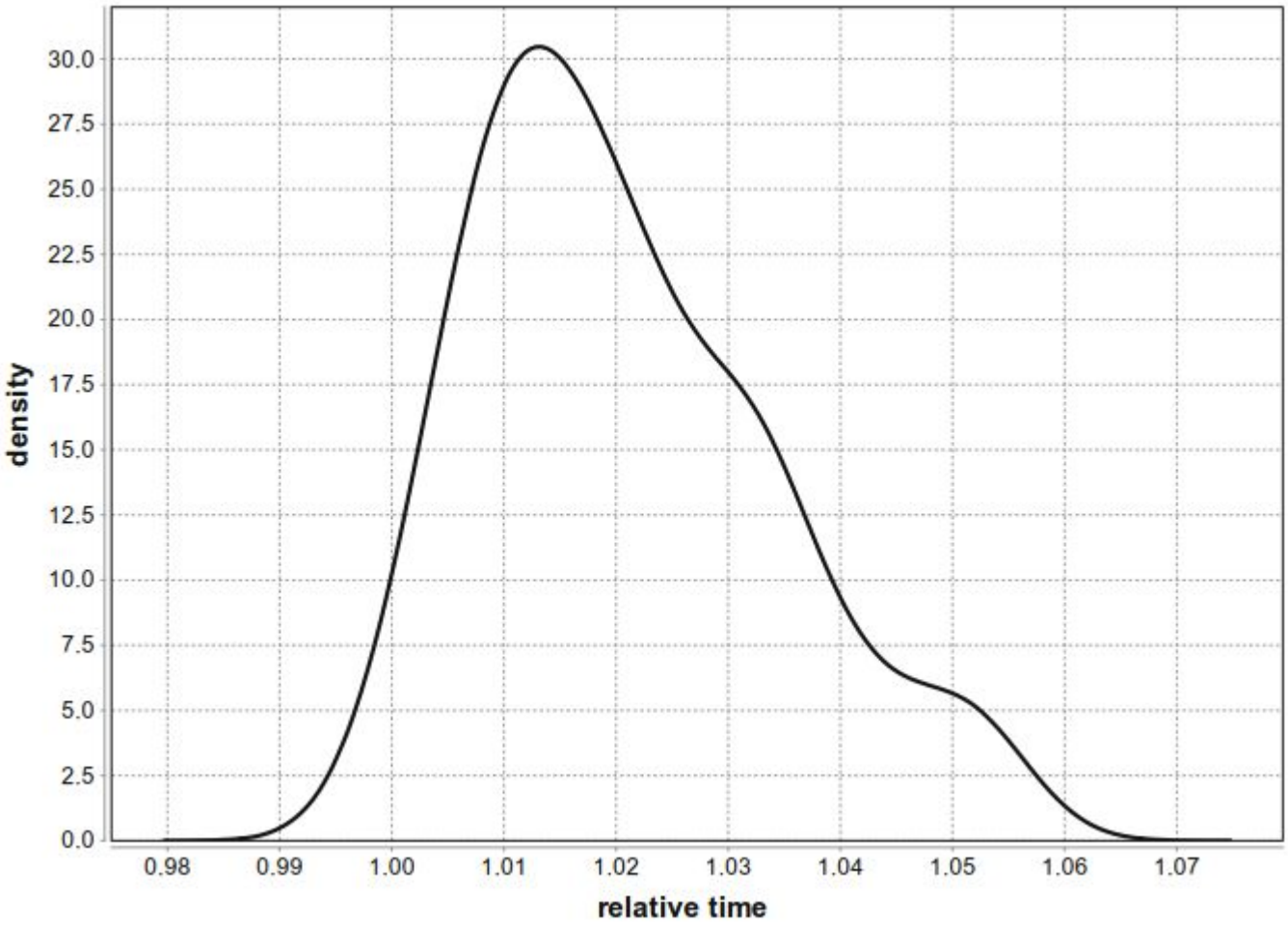
學習時間跟期末成績無關




Lemmerich, F. (2014). *Novel Techniques for Efficient and Effective Subgroup Discovery, Neue Techniken für effiziente und effektive Subgruppenentdeckung*. Retrieved from https://opus.bibliothek.uni-wuerzburg.de/frontdoor/index/index/docId/9781

# 行為不是線性的

運動成效




Knobbe, A., Orie, J., Hofman, N., van der Burgh, B., & Cachucho, R. (2017). Sports analytics for professional speed skating. Data Mining and Knowledge Discovery, 1–31.

# 大綱

1. 什麼是子群組探勘？
2. 子群組探勘的構成
3. 子群組探勘工具
4. 子群組探勘的使用案例
   a. 使用者互動資料的子群組探勘 (2016)
   b. 學習滿意度分析 (2014)
5. Cortana 子群組探勘實戰
   a. 學生成效分析: 什麼樣的學生成績會比較好呢？
   b. 學生成效分析: 找出學校跟成績有差異的子群組


# 什麼是子群組探勘？

# 原本對於資料探勘的認知

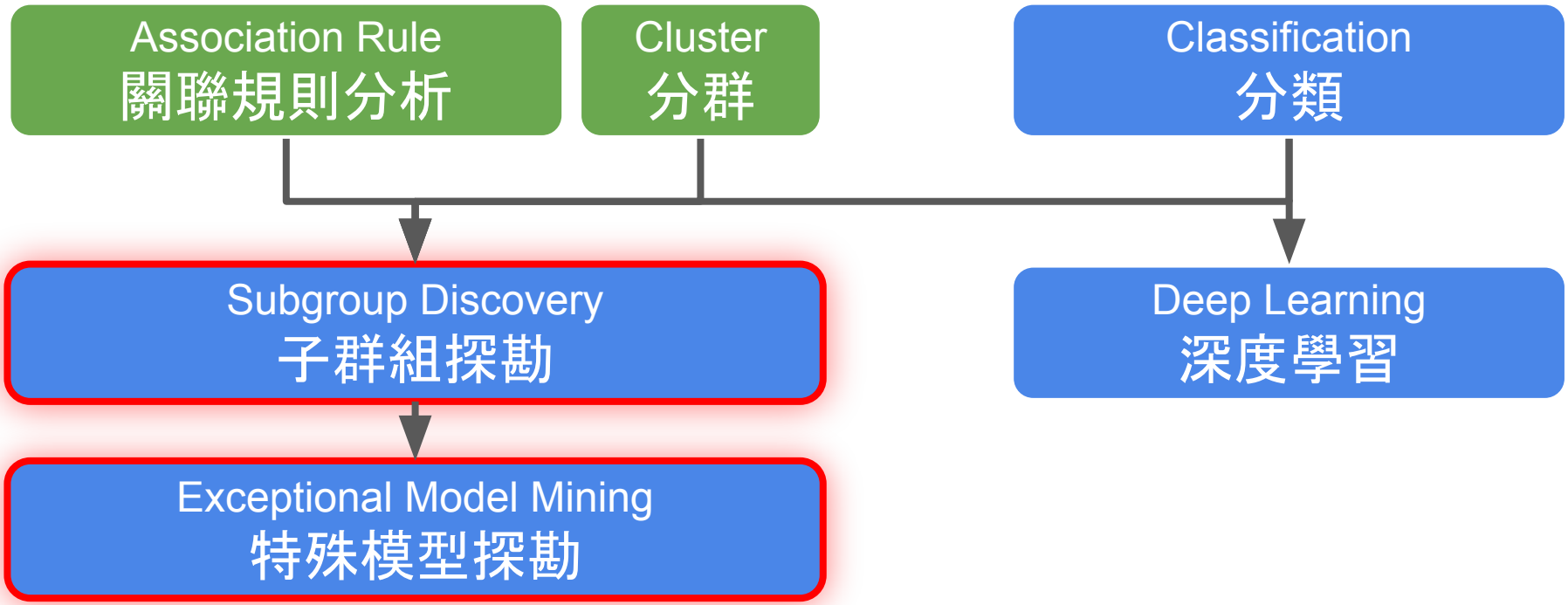
# 資料探勘的領域

**Data Mining**
**資料探勘**

**Machine Learning**
**機器學習**

無監督學習 Unsupervised Learning

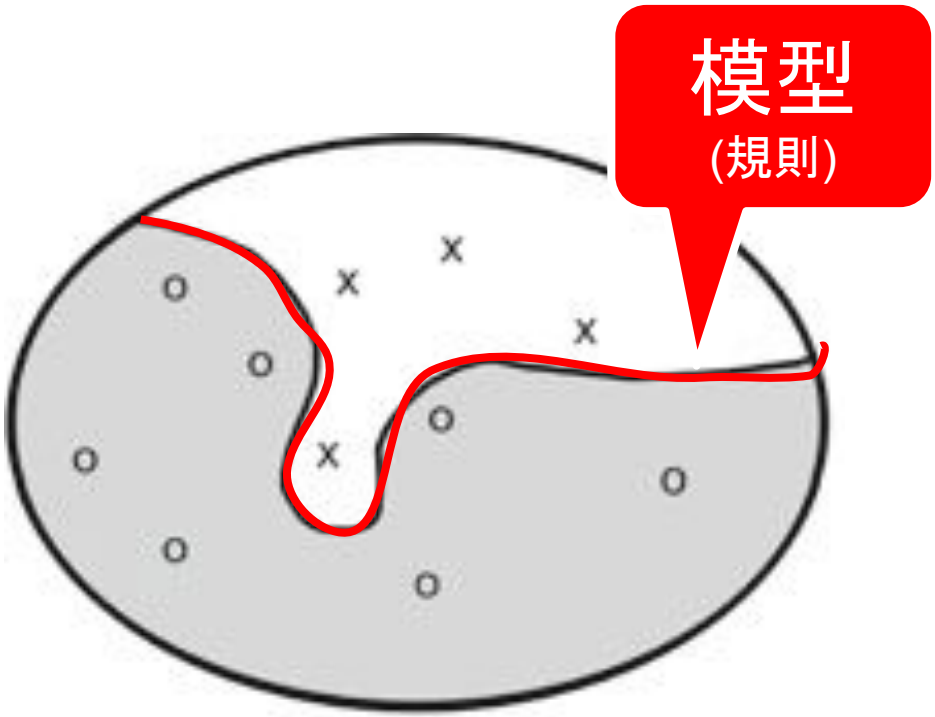
Supervised Learning 監督學習




敘述型歸納 Descriptive induction

Predictive induction 預測型歸納

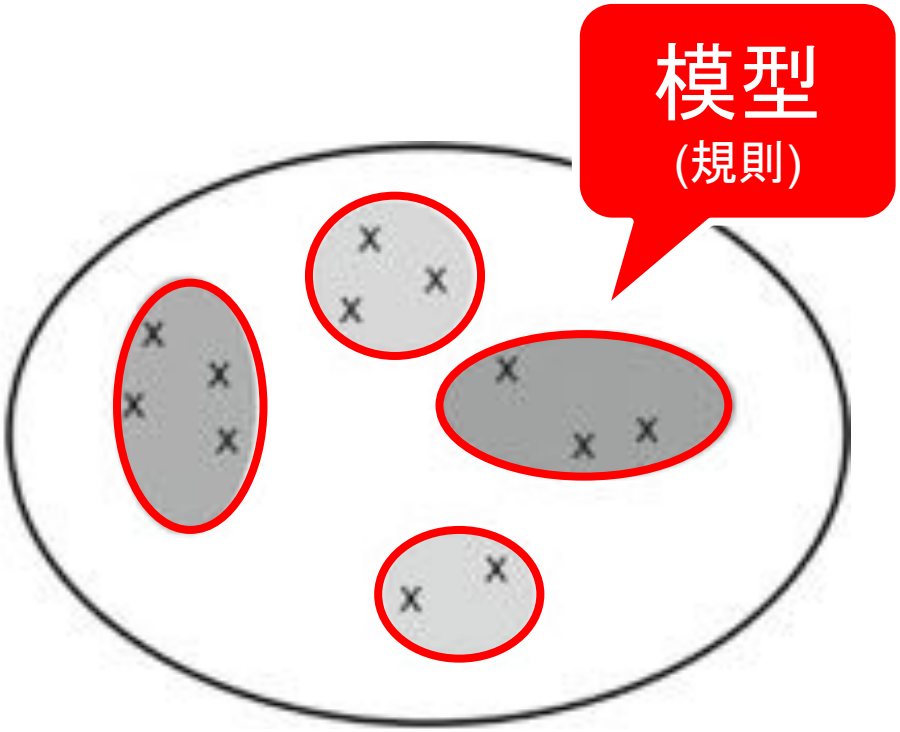
# 分類




- 模型建立目標：預測未知的新案例
  - 找出規則可以區分O跟X
  - 著重正確率跟可解釋性
  - 未來遇到未知案例時，就能用這個規則來辨識它為O或X
- 代表性技術：
  - 決策樹
  - SVM 支持向量機
  - 類神經網路=深度學習

Herrera, F., Carmona, C. J., González, P., & Jesus, M. J. del. (2011). An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, *29*(3), 495-525. doi:10.1007/s10115-010-0356-2

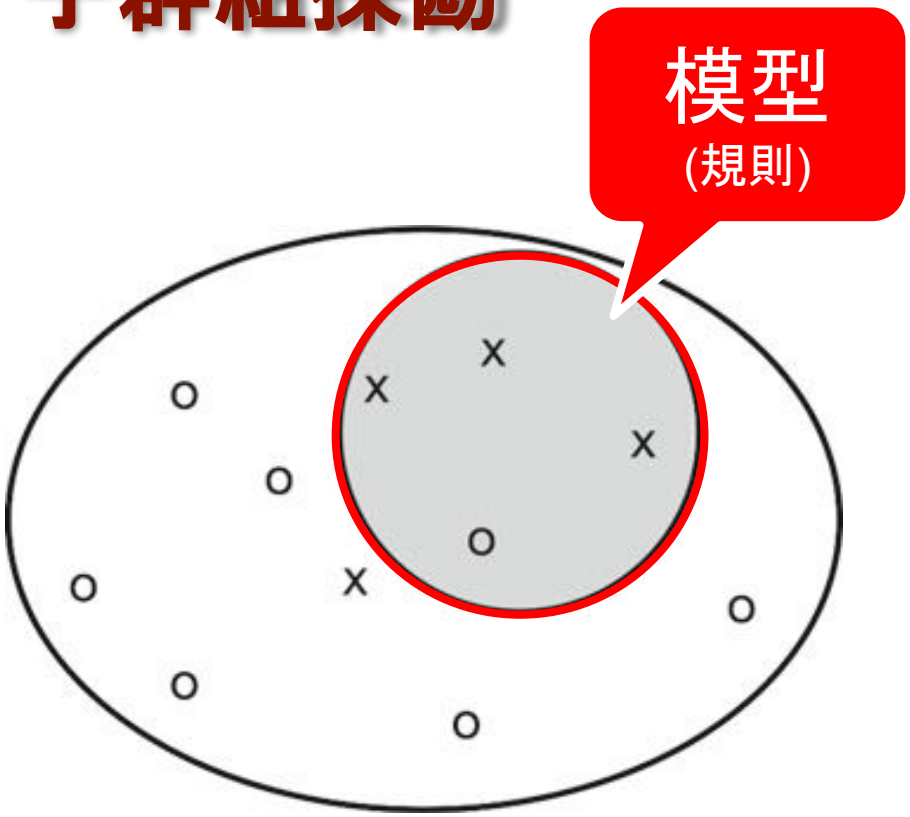
# 分群 & 關聯規則分析




- 模型建立目標：從已知的舊案例找出整體規律模式
  - 大量的混亂資料難以解讀，我們需要找出資料的規律模式
  - 著重支持率跟信心度(有多少案例在這個分群或規則中)
  - 有了規律模式之後，我們就能更容易解釋整體資料
- 代表性技術：
  - K-Means K平均法
  - Apriori 關聯規則探勘

Herrera, F., Carmona, C. J., González, P., & Jesus, M. J. del. (2011). An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, *29*(3), 495-525. doi:10.1007/s10115-010-0356-2

# 子群組探勘




- 模型建立目標：根據已知的舊案例中某些目標變項，找出單一且可解釋的規則
  - 目標變項：X
  - Q: 大部分X都有什麼特徵？
  - A: 都在圈圈右上角
  - 著重描述資料中的潛規則(知識)，而非預測
- 代表性技術：
  - PRIM (HotSpot)
  - Exceptional Model Mining (EMM)

Herrera, F., Carmona, C. J., González, P., & Jesus, M. J. del. (2011). An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, *29*(3), 495-525. doi:10.1007/s10115-010-0356-2

# 子群組探勘近年來的主要研究者


Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

德國 維爾茲堡大學


Universiteit
Leiden

荷蘭 萊頓大學

# 子群組探勘的構成

# 子群組探勘的元件

目標
研究者感興趣的變項

Rule
規則

R: Cond → $Target_{value}$

Condition
條件
聯合多個屬性值配對

Herrera, F., Carmona, C. J., González, P., & Jesus, M. J. del. (2011). An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, *29*(3), 495-525. doi:10.1007/s10115-010-0356-2

# 子群組探勘的例子

## 資料集

特徵值域:

- 年齡 = 小於25, 25至60, 大於60
- 性別 = 男, 女
- 國家 = 美國, 法國, 德國

目標變項值域

- 財富 = 窮, 普通, 富裕

## 子群組 (探勘結果)

- $R_1$: (年齡=小於25 & 國家=德國) → 財富 = 富裕
- $R_2$: (年齡=大於60 & 性別 = 女) → 財富 = 普通

Herrera, F., Carmona, C. J., González, P., & Jesus, M. J. del. (2011). An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, *29*(3), 495-525. doi:10.1007/s10115-010-0356-2

# 子群組探勘的元件:目標

目標
研究者感興趣的變項

R: Cond → $Target_{value}$

目標變項的資料類型

- 類別變項
  - 二元變項:True or False
  - 多個不同的選項
  - Ex: 實驗組/控制組
- 連續變項 (數值):最難分析
  - 離散化 (裝箱法)
  - 最大差異法:找尋平均數有別於整體的子群組
  - Ex: 後測成績
- 目標概念 (EMM演算法)
  - 以多個變項組成一個概念
  - 時間→成績 迴歸模型

Herrera, F., Carmona, C. J., González, P., & Jesus, M. J. del. (2011). An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, *29*(3), 495-525. doi:10.1007/s10115-010-0356-2

# 子群組探勘的元件:敘述語言

R: Cond → $Target_{value}$

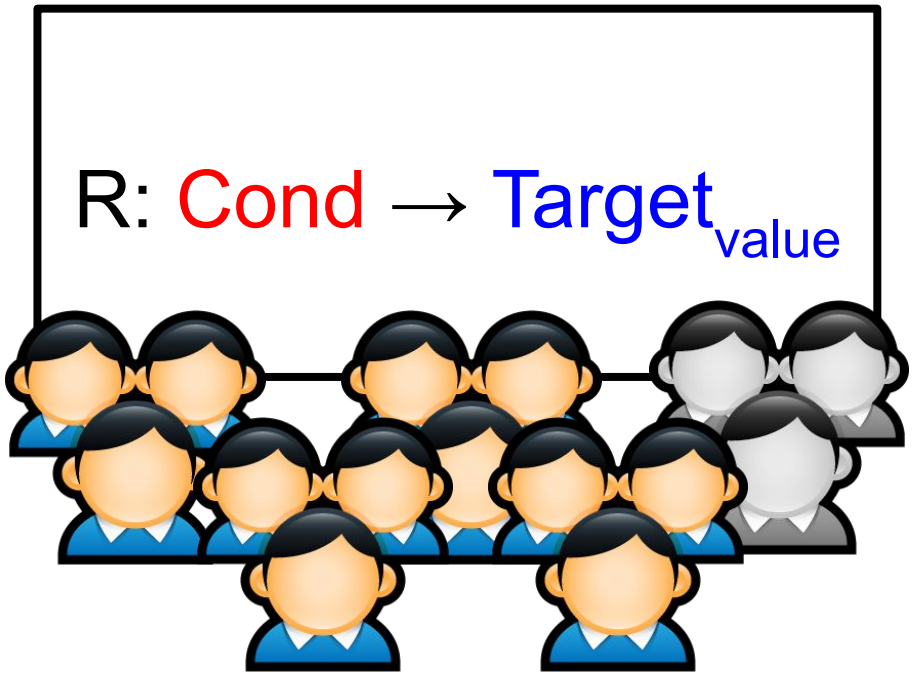
Condition
條件
聯合多個屬性值配對

如何呈現規則？

- 方式:屬性值配對
  - 屬性=值
  - 用 & 聯合多個屬性值配對
- 屬性值配對的運算子
  - 等於 =
  - 大於、小於、或等於 > < ≥ ≤
  - 模糊邏輯 (接近高、中、低)
- 屬性的資料類型:類別, 連續
- 限制
  - 深度:多少組屬性值配對？
  - 廣度:只呈現最重要的多少規則？

Herrera, F., Carmona, C. J., González, P., & Jesus, M. J. del. (2011). An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, *29*(3), 495-525. doi:10.1007/s10115-010-0356-2

# 子群組探勘的元件:品質評估指標




有12/15=80%的人
都符合這條規則 (涵蓋率)

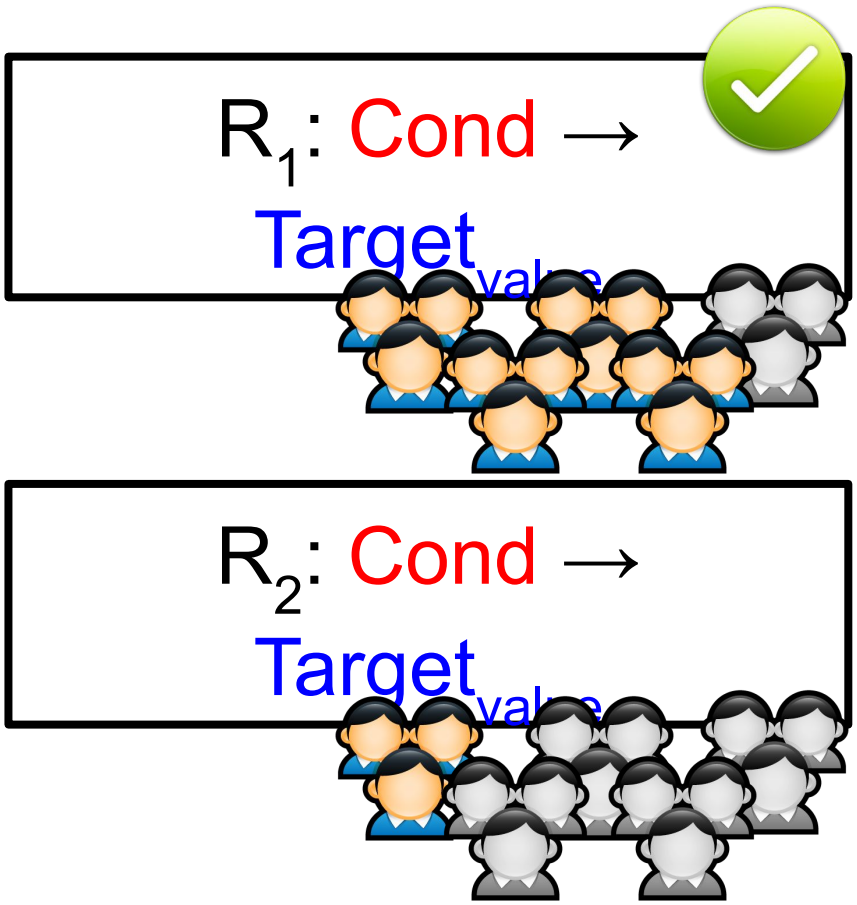
何謂「好的規則」?

- 泛用指標:
  - 涵蓋率, 支持率
- 精準指標:
  - 信心度, 正確率$Q_c$, 正確率$Q_g$
- 值得關注指標:
  - 興趣度, 新穎度, 顯著度
- 混合指標:
  - 敏感度, 錯誤警示, 專指度, 少見度 WRAcc

沒有最好的指標!

Herrera, F., Carmona, C. J., González, P., & Jesus, M. J. del. (2011). An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, *29*(3), 495-525. doi:10.1007/s10115-010-0356-2

# 子群組探勘的元件:搜尋策略




如何找到「好的規則」?

- 窮盡演算法 & 篩選
- 啟發式演算法
- 收束搜尋 Beam search
- 演化式演算法:基因演算法

如何處理連續變項?

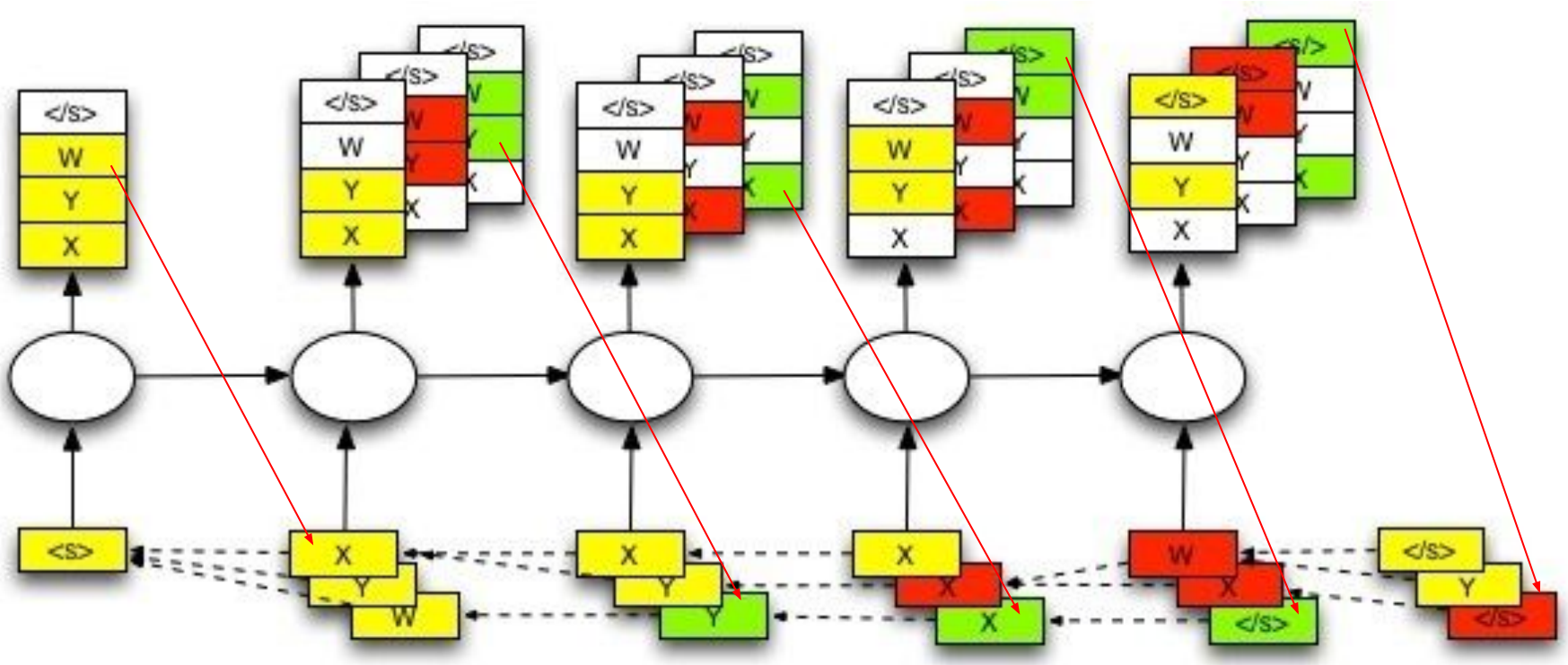
- 離散化
- 模糊邏輯

指標改進多少才是好的規則?

- 最小進步門檻

Herrera, F., Carmona, C. J., González, P., & Jesus, M. J. del. (2011). An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, *29*(3), 495-525. doi:10.1007/s10115-010-0356-2

# 收束搜尋 Beam search



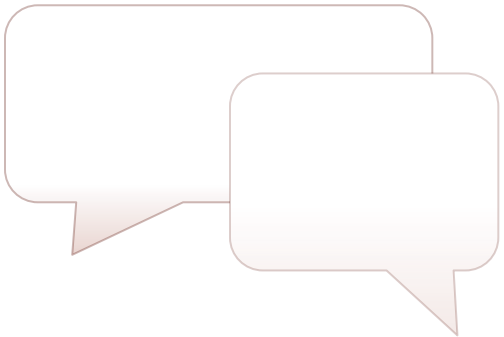

- 設定候選規則欄位:3
- 每次搜尋過程中，不斷選擇最佳的3個候選規則，剔除不佳者

https://www.quora.com/Why-is-beam-search-required-in-sequence-to-sequence-transduction-using-recurrent-neural-networks

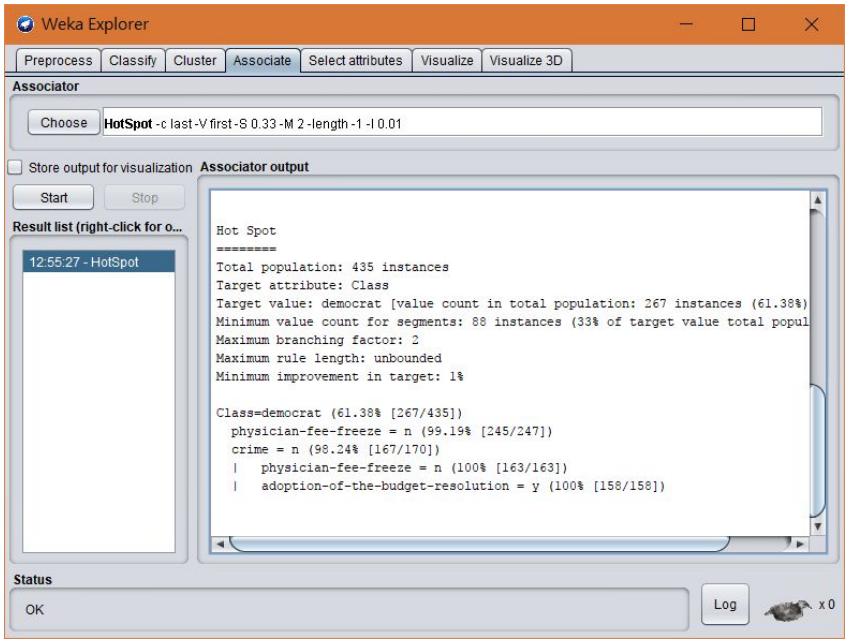
# 子群組探勘工具

# 工具列表

- Explora
- KEPLER
- Subgroup Miner
- RapidMiner
- Knime
- Weka的HotSpot
- Orange
- KEEL
- VIKAMIME
- Cortana



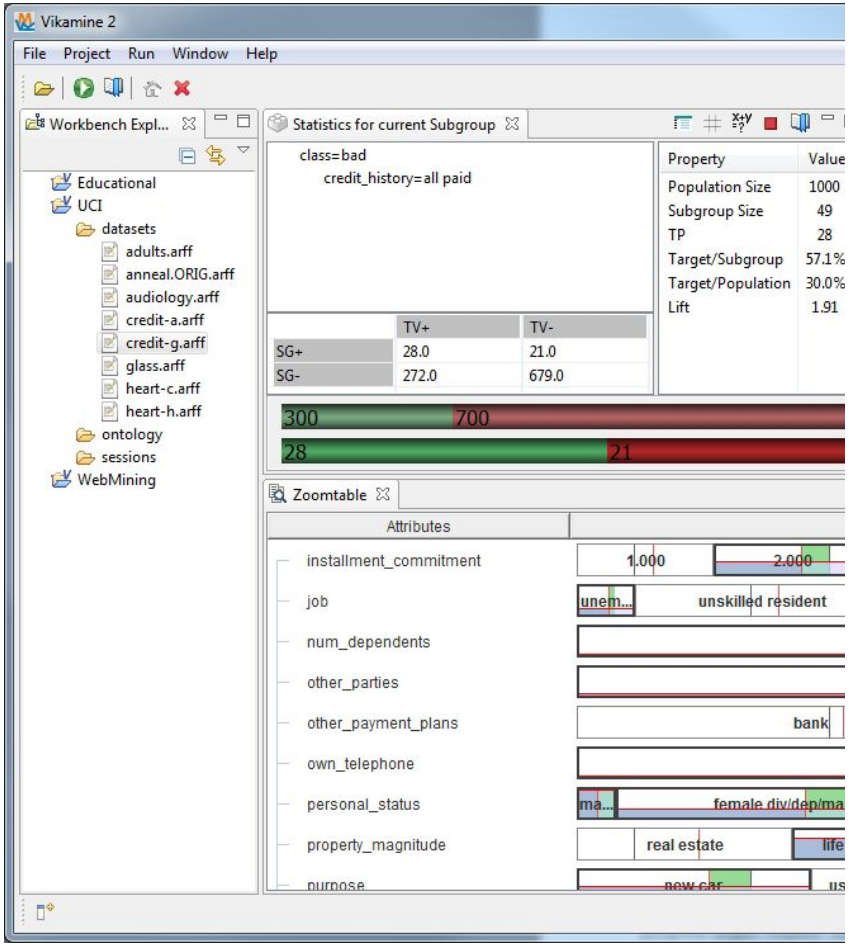
# Weka的HotSpot (2013)




- 演算法：PRIM (HotSpot)
- 目標：
  - 只能一個變項
  - 類別變項
- 連續變項特徵處理：
  - 連續變項 (平均值二分法)
  - 手動離散化
- 品質評估指標：
  - 涵蓋率, 信賴度, 增益度, 影響度, 肯定度 (關聯規則分析的指標)
- 互動性跟解讀性差

http://blog.pulipuli.info/2017/08/wekahotspot-association-rule-mining.html

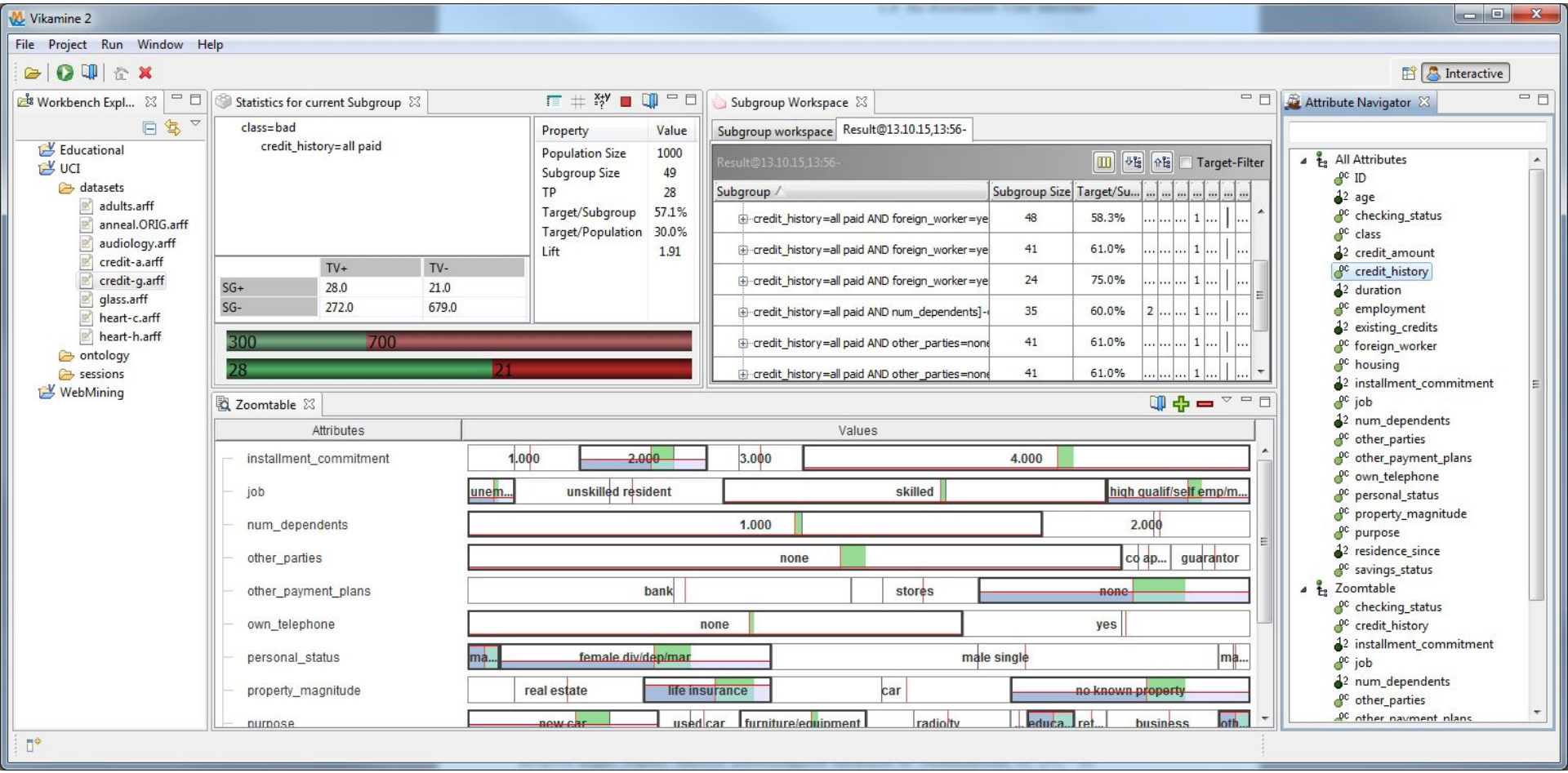
# VIKAMIME (2016)



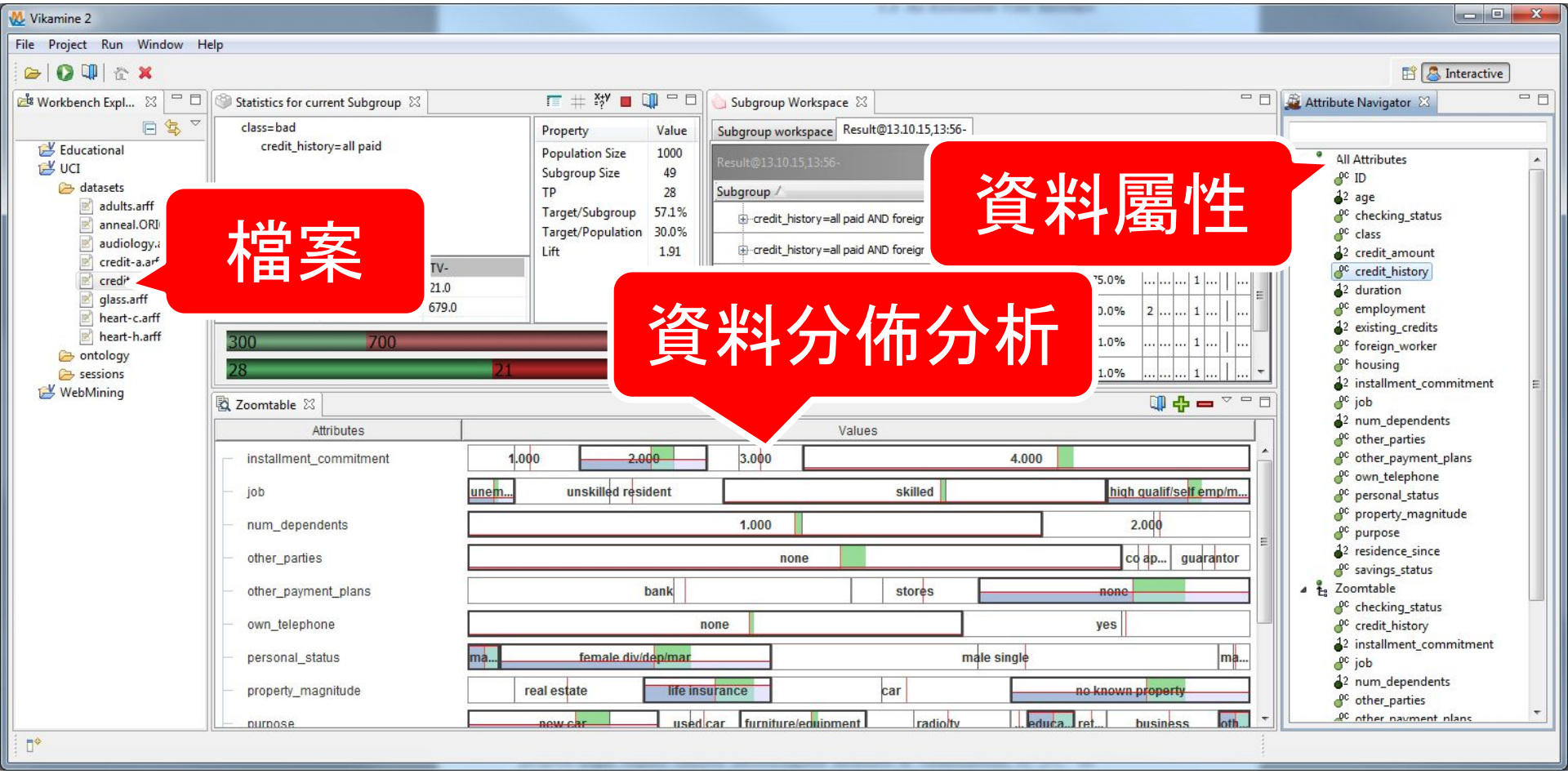

- 演算法:
  - SDMAP (廣度最佳), BitSetSD, Beam Search, Simple DFS (深度最佳)
- 目標:
  - 只能一個變項
  - 類別/連續變項
- 連續變項特徵處理
  - 自動離散化(有待調查?
  - 手動離散化:裝箱/分群
- 品質評估指標:
  - 涵蓋率, 正確率, 影響度, 顯著度
- 互動性跟解讀性極佳

http://vikamine.org/

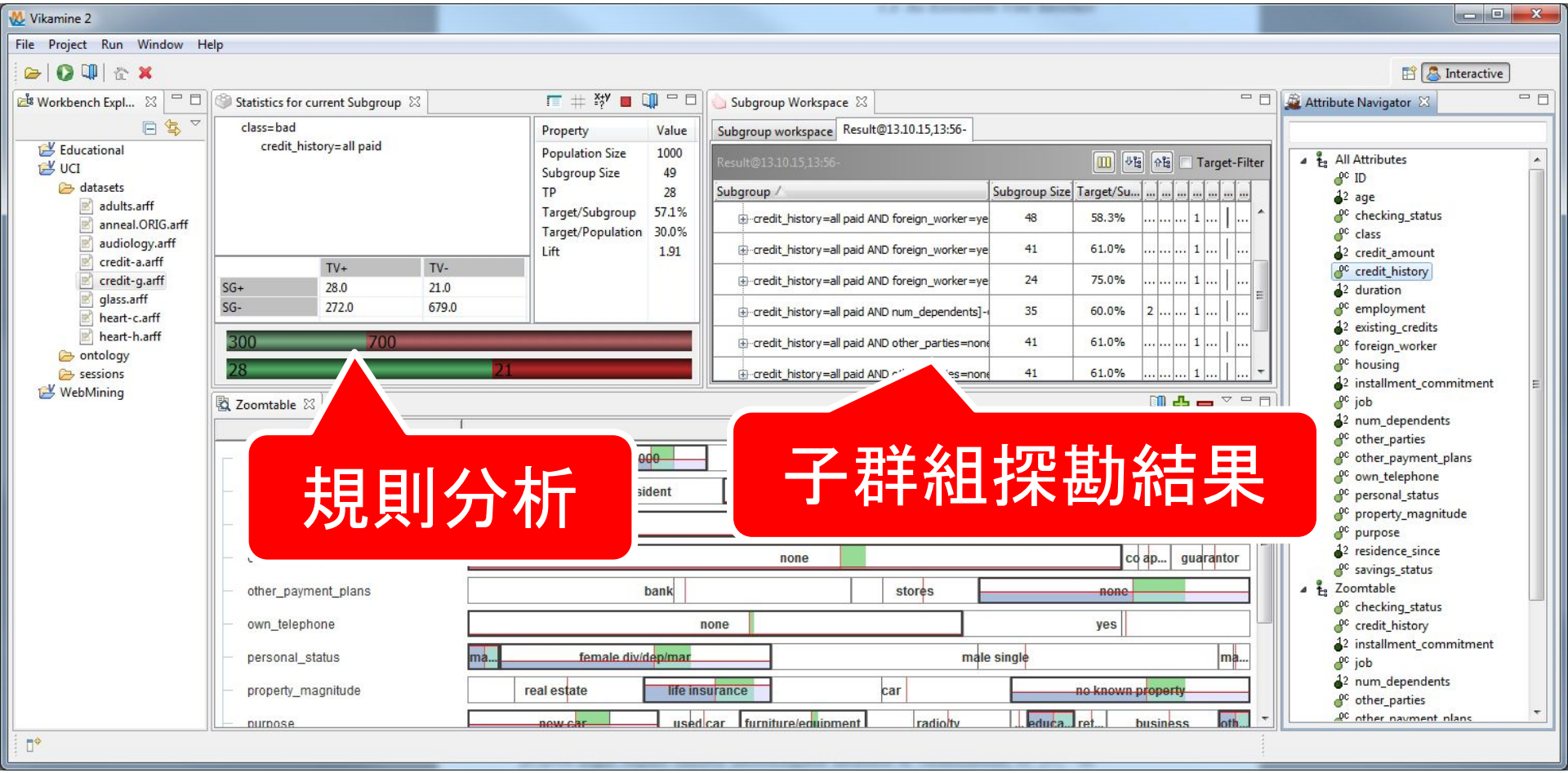
# VIKAMIME 操作介面 (1/3)



# VIKAMIME 操作介面 (2/3)



# VIKAMIME 操作介面 (3/3)



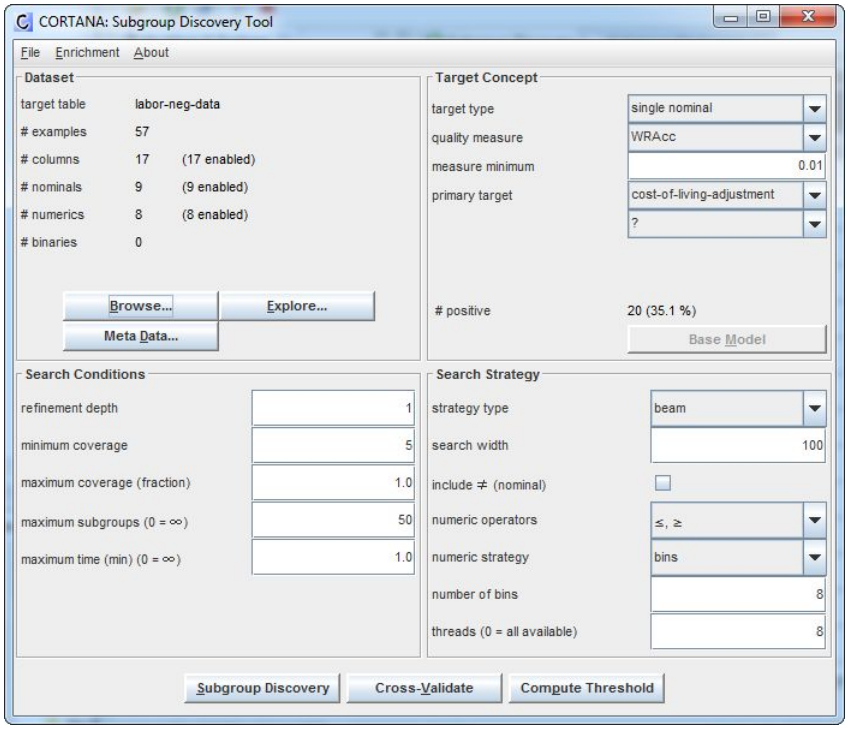
# VIKAMIME 探勘結果

品質=跟整體相比，子群組的指標改進程度

顯著性=規則的不隨機性

| Subgroup | Quality | Popul... | Subgr... | Targe... | Targe... | Lift | Signif... | Cluster | TP / FP |
|---|---|---|---|---|---|---|---|---|---|
| class=good | | | | | | | | | |
| contribution-to-dental-plan=full | 4.561 | 57 | 13 | 100.0% | 64.9% | 1.541 | <=0.01... | 1 | 13 |
| contribution-to-health-plan=full AND | 2.807 | 57 | 8 | 100.0% | 64.9% | 1.541 | <=0.1 ... | 1 | 8 |
| contribution-to-dental-plan=full | | | | | | | | | |
| contribution-to-health-plan=full | | | | | | | | | |
| duration[1.5;2.5[ AND contribution-t | 3.158 | 57 | 9 | 100.0% | 64.9% | 1.541 | <=0.05... | 1 | 9 |
| statutory-holidays[10.5;11.5[ AND co | 2.807 | 57 | 8 | 100.0% | 64.9% | 1.541 | <=0.1 ... | 1 | 8 |
| vacation=generous | 3.614 | 57 | 16 | 87.5% | 64.9% | 1.348 | <=0.05... | 999 | 14 TP, 2 FP |
| wage-increase-first-year[4.15;4.8[ | 2.86 | 57 | 11 | 90.9% | 64.9% | 1.4 | <=0.1 ... | 999 | 10 |
| wage-increase-first-year[4.8;∞[ | 4.211 | 57 | 12 | 100.0% | 64.9% | 1.541 | <=0.05... | 999 | 12 |
| wage-increase-second-year[3.75;4.2[ | 2.807 | 57 | 8 | 100.0% | 64.9% | 1.541 | <=0.1 ... | 999 | 8 |
| wage-increase-second-year[4.2;4.9[ | 3.158 | 57 | 9 | 100.0% | 64.9% | 1.541 | <=0.05... | 999 | 9 |
| working-hours]-∞;36.5[ | 2.86 | 57 | 11 | 90.9% | 64.9% | 1.4 | <=0.1 ... | 999 | 10 TP, 1 FP |

# CORTANA (2013)




- 演算法：EMM
  - Beam Search, 深度優先, 廣度優先
- 目標：
  - 一個/二個變項, 多變項
  - 類別/連續變項
- 連續變項特徵處理
  - 自動離散化：裝箱法
- 品質評估指標：
  - 眾多指標
- 用推論統計決定最佳參數設定
- 互動性跟解讀性普通

http://datamining.liacs.nl/cortana.html

# CORTANA 探勘結果

9 subgroups found; target table = labor-neg-data; quality measure = WRAcc; target value = go...

| Nr. | Depth | Coverage | Quality | Probability | Positives | p-Value | Conditions |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 30 | 0.149584 | 0.933333 | 28 | 9.325873... | pension = '?' |
| 2 | 1 | 39 | 0.134811 | 0.846154 | 33 | 4.589414... | wage-increase-first-year >= '3.0' |
| 3 | 1 | 31 | 0.120652 | 0.870968 | 27 | 4.620461... | wage-increase-second-year >= '4.0' |
| 4 | 1 | 22 | 0.117882 | 0.954545 | 21 | 1.921927... | wage-increase-first-year >= '4.5' |
| 5 | 1 | 29 | 0.108341 | 0.862069 | 25 | 0.0103455 | wage-increase-first-year >= '4.0' |
| 6 | 1 | 37 | 0.104955 | 0.810811 | 30 | 0.030420... | wage-increase-second-year >= '3.0' |
| 7 | 1 | 18 | 0.093259 | 0.944444 | 17 | 0.360048... | shift-differential >= '4.0' |
| 8 | 1 | 37 | 0.087411 | 0.783784 | 29 | 0.655359... | statutory-holidays >= '11.0' |
| 9 | 1 | 45 | 0.084026 | 0.755556 | 34 | 0.7992004 | wage-increase-first-year >= '2.5' |

ROC | Browse Selected | Delete Selected | Gaussian p-Values

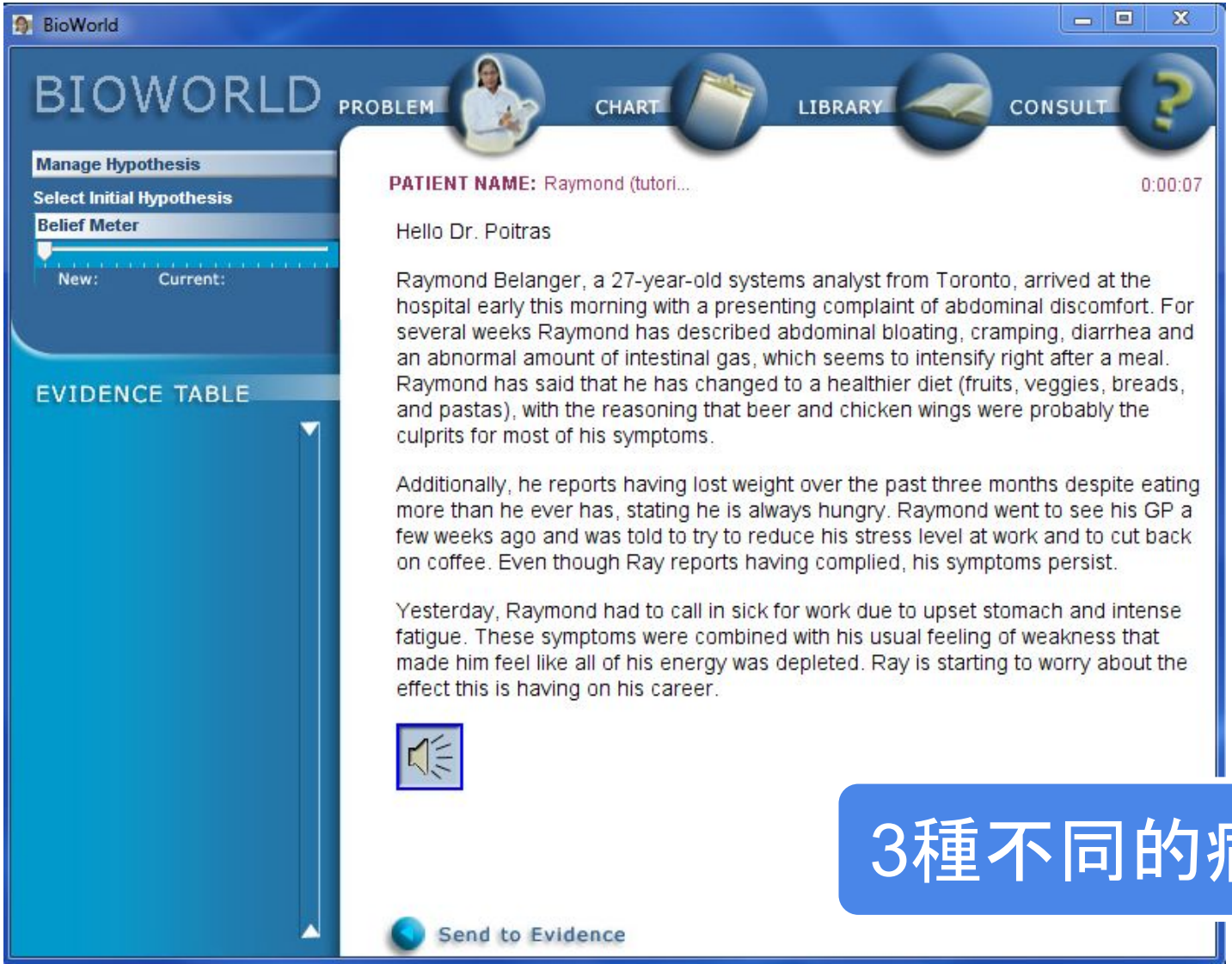
Close

子群組探勘的使用案例A

# 使用者互動資料的子群組探勘 (2016)


Poitras, E. G., Lajoie, S. P., Doleck, T., & Jarrell, A. (2016). Subgroup discovery with user interaction data: An empirically guided approach to improving intelligent tutoring systems. Journal of Educational Technology & Society, 19(2), 204.

# BIOWORLD: 醫生模擬訓練系統




3種不同的病患

# BIOWORLD操作步驟

## 主要操作

1. 閱讀病患自述狀況
   a. 閱讀一連串病患的自述
   b. 選擇自述中值得注意的重點 (標註) + 信心程度加入到證據表中
2. 檢查病患狀況
   a. 選擇要檢測的項目
   b. 獲得檢測結果
3. 診斷疾病

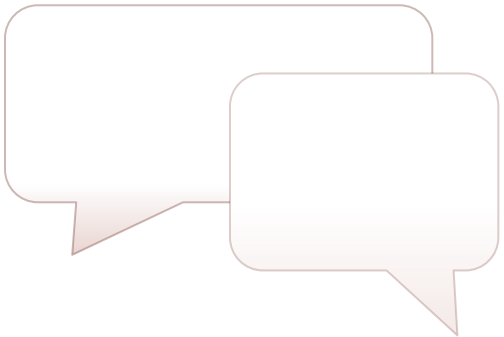
## 輔助功能

- 疾病資料庫
  - 可以查詢某些疾病可能有的特徵
- 求助工具
  - 取得提示

## Novice-Expert overlay Model

# 新手-專家重疊模型:如何改進?

- 先前的研究指出:新手跟老手的操作是有差別的
- Bioworld想要結合新手專家重疊模型來作為指引新手的指示
- 問題聚焦
  - 什麼樣的檢查結果，會導向錯誤的診斷？
  - 這些困難問題的影響是什麼？



# 子群組分析結果 罕見疾病 嗜鉻細胞瘤 案例誤判規則

*Table 2.* Patterns extracted from the subgroup discovery algorithm

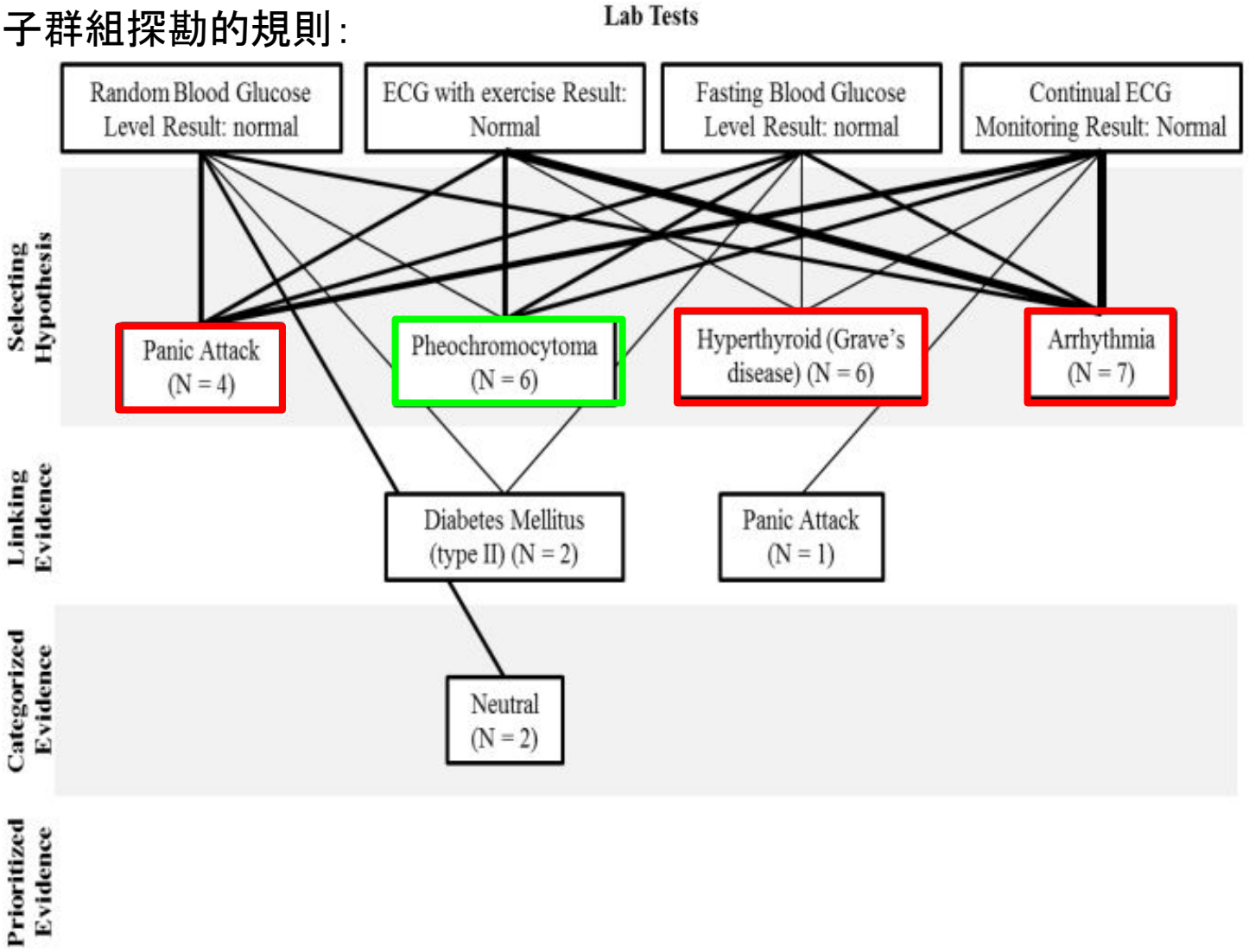
| Antecedent | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| *Incorrect diagnosis for the Cynthia case (i.e., exhibiting signs of Pheochromocytoma)* | | | | | | |
| Random Blood Glucose Level Result: normal=true | 5 | 1 | 6 | 6.1% | 83.3% | 43.9% |
| Continual ECG Monitoring Result: Normal=true | 9 | 2 | 11 | 11.2% | 81.8% | 46.9% |
| Fasting Blood Glucose Level Result: normal=true | 5 | 2 | 7 | 7.1% | 71.4% | 42.9% |
| ECG with exercise Result: Normal=true | 7 | 3 | 10 | 10.2% | 70.0% | 43.9% |

*Note.* 1 = Positive instances; 2 = Negative instances; 3 = Size; 4 = Coverage; 5 = Precision; 6 = Accuracy.

- 看到檢查結果之後誤判疾病的規則
- Random Blood Glucose Level Result: normal=true
  - 6個人進行此檢查
  - 5個人判斷疾病錯誤 (正面案例)
  - 1個人判斷疾病正確 (負面案例)
- 可以據此改善新手-專家重疊模型，加入更多提示

# 規則細部分析 診斷錯的情況還做了什麼事情?

子群組探勘的規則:



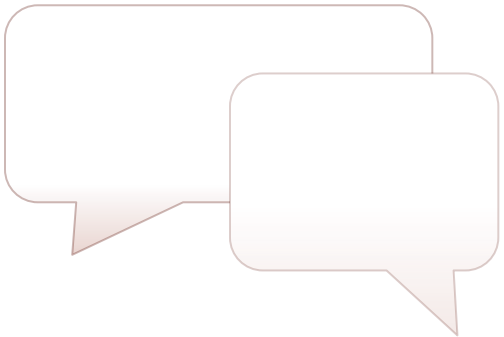
子群組探勘的使用案例B

# 學習滿意度分析 (2014)

Lemmerich, F. (2014). Novel Techniques for Efficient and Effective Subgroup Discovery, Neue Techniken für effiziente und effektive Subgruppenentdeckung. Würzburg, Univ., Germany. Retrieved from https://opus.bibliothek.uni-wuerzburg.de/frontdoor/index/index/docId/9781

# 資料蒐集

- 學生的期末學習課程滿意度調查
- 類別變項
  - 課程
  - 學院
  - 學位
- 連續變項
  - 各種滿意度：1最佳、6最不滿意
  - 學生背景調查：學習時間、是否工讀等等
- 共有119個特徵，2800份填答



# 整體分析

| Description | # Students | Mean satisfaction | |
|---|---|---|---|
| *Overall* | *2799* | *2.61* | |
| Matching requirem. = [4-6] | 258 | 3.46 | |
| Matching requirem. = [4-6] ∧ Information = [4-6] | 55 | 3.93 | 不滿意 |
| Overcrowded courses = true | 887 | 2.89 | |
| Overlapping courses = true | 1030 | 2.85 | |
| Degree program = Medicine | 321 | 2.24 | 很滿意 |
| Degree program = Medicine ∧ Semester= [1-2] | 70 | 1.93 | |

- 比整體更不滿意的子群組規則
  - 不符合要求條件、課程未充分告知、課堂擁擠、衝堂
- 比整體更滿意的子群組規則
  - 醫學系、第1-2學期

# 性別x學習滿意度:指標定義

- 定義新的品質評估指標:

$$q^{a}_{meanDiff}(P) = i_P{}^{a} \cdot ((\mu_{F \wedge P} - \mu_{M \wedge P}) - (\mu_F - \mu_M))$$

- 考慮變數
  - 整體的男女學習滿意度平均值差異
  - 符合子群組規則的男女學習滿意度平均值差異
  - 子群組符合的案例數量

# 性別x學習滿意度:探勘結果

- 某自然科學的課堂
  - 整體滿意度 2.5
  - 40位男生滿意度 2.25
  - 24位女生滿意度 2.96
  - 但不是每堂自然科學課程都是如此
- 需要每週20小時讀書 & 交通時間1小時
  - 10位男生滿意度 2.2
  - 27位女生滿意度 2.93

# 對實驗室分析方法的啟示

某些背景變項

- 需要每週20小時讀書 & 交通時間1小時
  - 10位男生滿意度 2.2
  - 27位女生滿意度 2.93

學習成效

實驗組/控制組

Cortana 子群組探勘實戰A


# 學生成效分析
## 什麼樣的學生成績會比較好呢？

# 資料集

- 取自UCI機器學習保存庫中的「Student Performance Data Set」
- 葡萄牙的兩所中學的學生成績跟學生的背景變項
  - 學生的家庭
  - 學生的學校表現
  - 學生的最後一學期的成績

為了方便解說，我修改了部分資料內容

https://docs.google.com/spreadsheets/d/1aHRvPB2RJFseQYMS9j-sFQHlcqNY3os8lxwiEaPo9iA/edit?usp=sharing

http://archive.ics.uci.edu/ml/datasets/Student+Performance

# 特徵說明 (1/3)

- school 學校: 實驗組學校(2), 控制組學校(1)
- sex 性別: F, M
- age 年齡: 15~22
- address 居住處: U都市, R鄉村
- famsize 家庭人數: ≤ 3, > 3
- Pstatus 雙親狀態: 同居, 分居
- Medu 母親教育: 未受教育0~4高等教育
- Fedu 父親教育: 未受教育0~4高等教育
- Mjob 母親職業: teacher, health, civil, at_home, other
- Fjob 母親職業: teacher, health, civil, at_home, other
- reason 選校理由: home 近家, reputation 學校名聲, course 優質課程, other

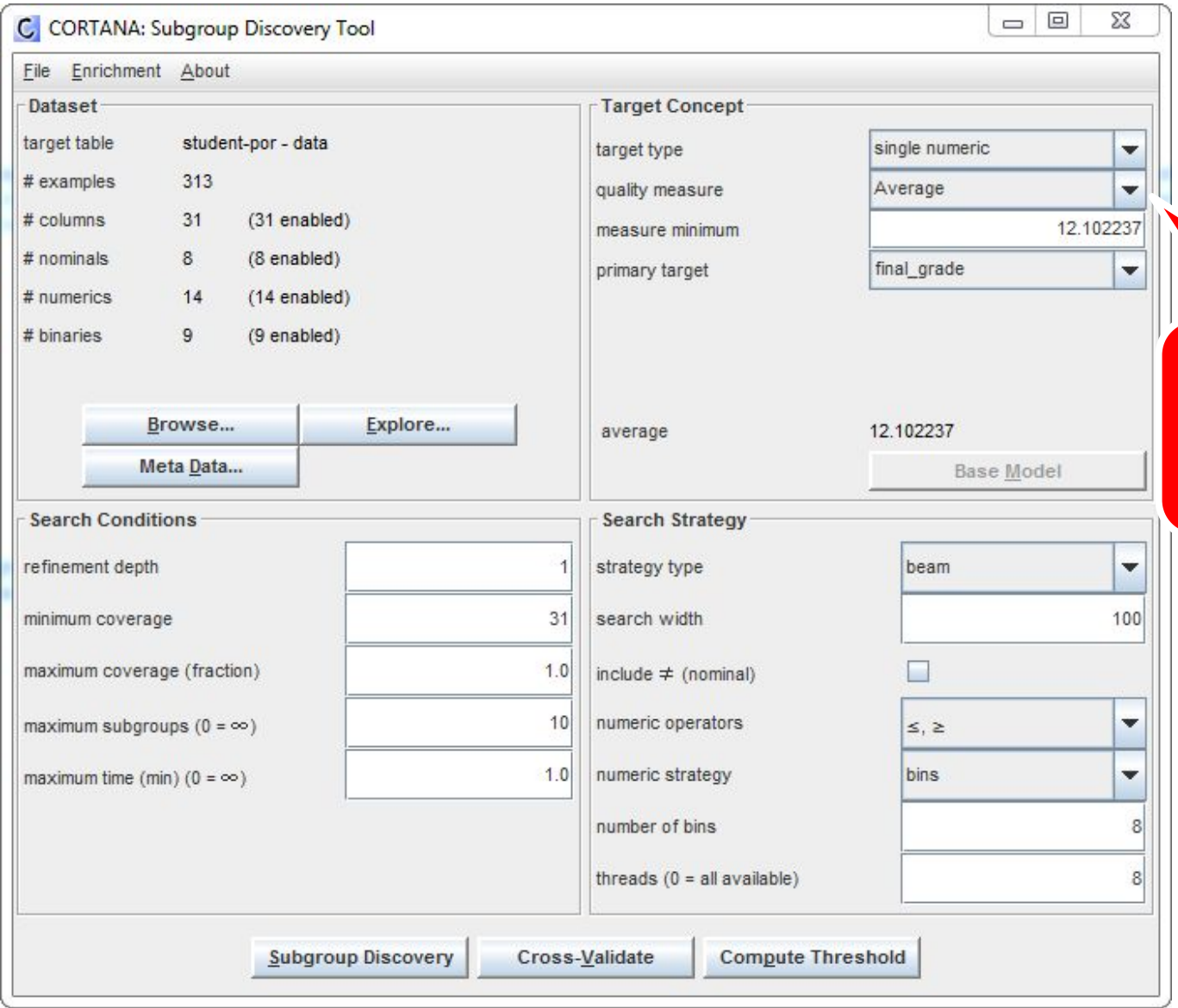
# 特徵說明 (2/3)

- guadian 學生監護人: mother, father, other
- traveltime 學校交通時間(min): (1)1~15, (2)15~30, (3)30~60, (4)>60
- studytime 每週學習時間(hour): (1)<2, (2)2~5, (3)5~10, (4)>10
- failures 課程犯錯次數: 0~4
- schoolsup 學校教育補助: yes,no
- famsup 家庭教育補助: yes,no
- paid 其他補習課程: yes, no
- activities 其他課程活動: yes, no
- nursery 是否去托兒所: yes, no
- higher 是否想上高等教育: yes, no

# 特徵說明 (3/3)

- internet 家裡能上網嗎: yes, no
- romantic 戀愛中: yes, no
- famrel 與家人的關係: 非常糟1~5非常好
- freetime 課後空閒時間: 非常少1~5非常多
- goout 跟朋友出遊: 非常少1~5非常多
- Dalc 工作日喝酒: 非常少1~5非常多
- Walc 週末喝酒: 非常少1~5非常多
- health 現在健康狀況: 非常糟1~5非常好
- absences 缺席次數: 0~93
- grade3 最後一學期成績: 0~20

# 什麼樣特徵的學生，成績會較好呢?




目標：要找出比整體平均值高的子群組

# 什麼樣特徵的學生，成績會較好呢？

10 subgroups found; target table = student-por - data; quality measure = Z-Score

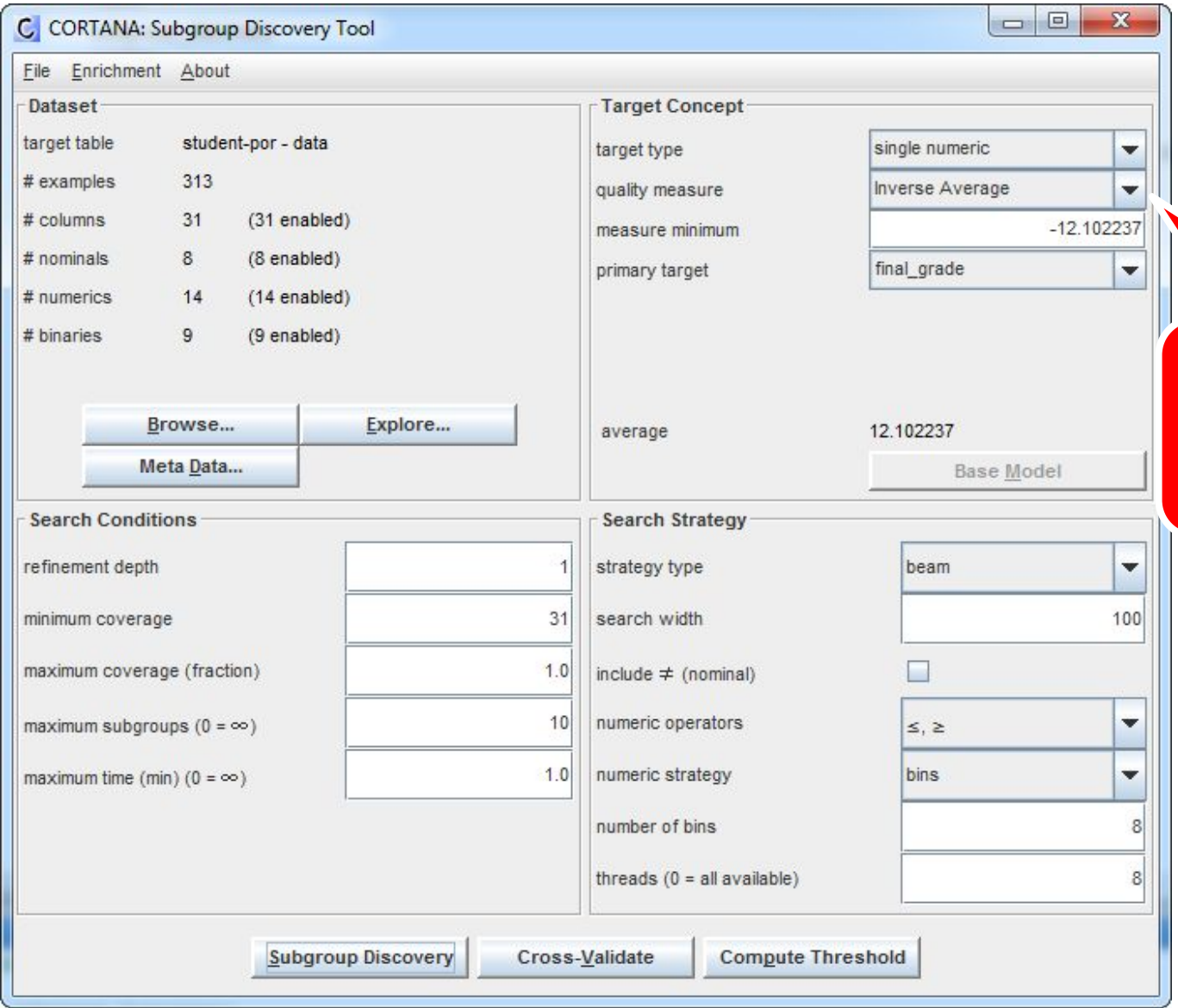
| Nr. | Depth | Coverage | Quality | Average | St. Dev. | p-Value | Conditions |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 88 | 4.133197 | 13.147... | 2.395734 | - | Medu >= '4.0' |
| 2 | 1 | 34 | 3.868419 | 13.676... | 2.284482 | - | Mjob = 'teacher' |
| 3 | 1 | 147 | 2.953506 | 12.680... | 2.291618 | - | Medu >= '3.0' |
| 4 | 1 | 124 | 2.926311 | 12.725... | 2.631505 | - | absences <= '0.0' |
| 5 | 1 | 60 | 2.713013 | 12.933... | 2.462158 | - | Fedu >= '4.0' |
| 6 | 1 | 62 | 2.55091 | 12.870... | 2.225105 | - | reason = 'reputation' |
| 7 | 1 | 273 | 2.425353 | 12.450... | 2.211842 | - | failures <= '0.0' |
| 8 | 1 | 69 | 2.331014 | 12.768... | 2.233765 | - | studytime >= '3.0' |
| 9 | 1 | 127 | 2.244345 | 12.574... | 2.253675 | - | Fedu >= '3.0' |
| 10 | 1 | 197 | 2.217673 | 12.477... | 2.283407 | - | Walc <= '2.0' |

Browse Selected | Delete Selected | Save | Print

Close

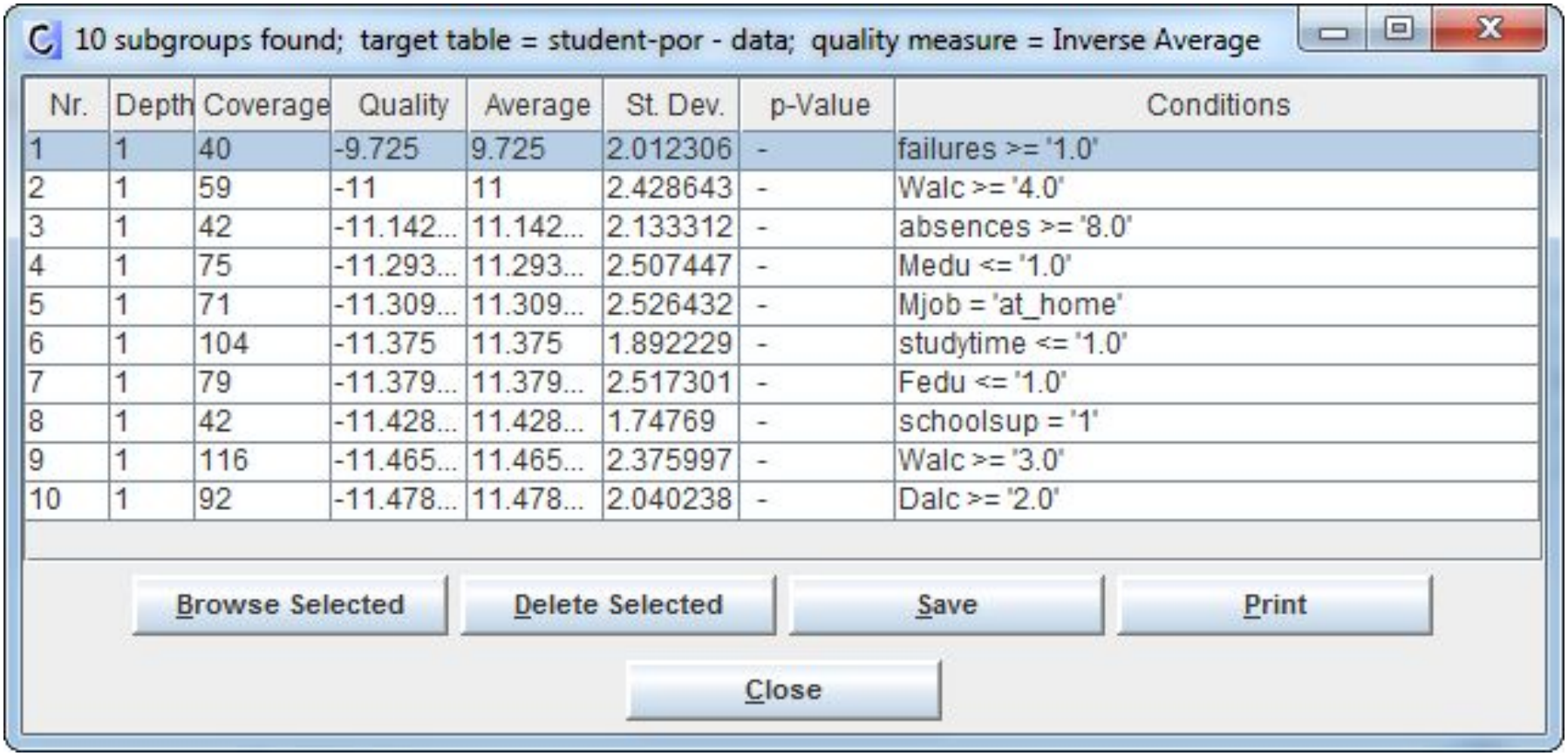
母親教育在大學以上時，子群組的成績平均值為較高的13.14！

# 什麼樣特徵的學生，成績會較差呢？




目標：要找出比整體平均值低的子群組

# 什麼樣特徵的學生，成績會較差呢？



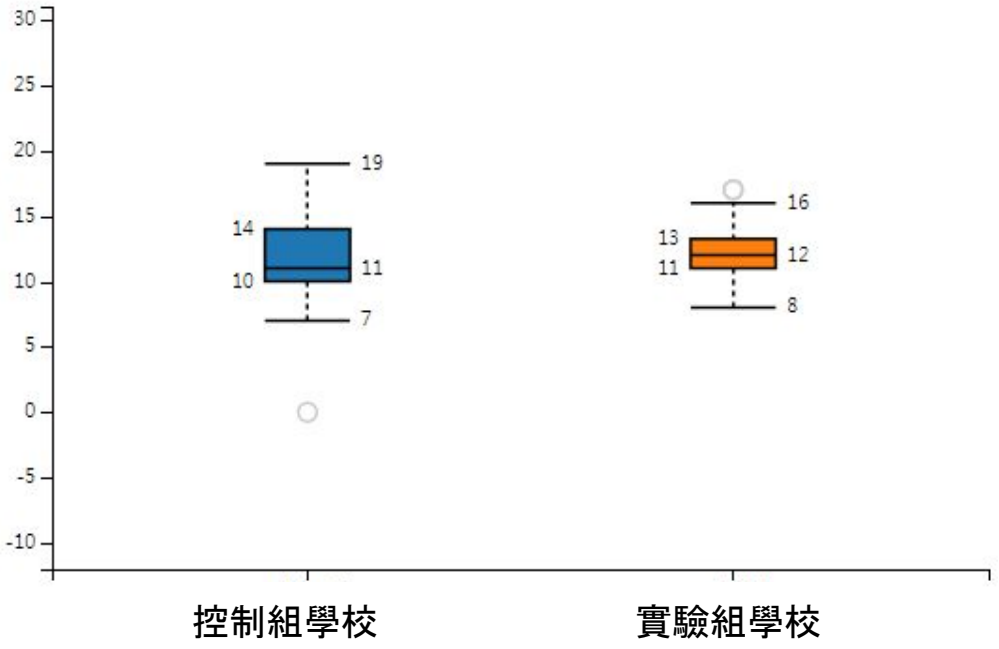

| Nr. | Depth | Coverage | Quality | Average | St. Dev. | p-Value | Conditions |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 40 | -9.725 | 9.725 | 2.012306 | - | failures >= '1.0' |
| 2 | 1 | 59 | -11 | 11 | 2.428643 | - | Walc >= '4.0' |
| 3 | 1 | 42 | -11.142... | 11.142... | 2.133312 | - | absences >= '8.0' |
| 4 | 1 | 75 | -11.293... | 11.293... | 2.507447 | - | Medu <= '1.0' |
| 5 | 1 | 71 | -11.309... | 11.309... | 2.526432 | - | Mjob = 'at_home' |
| 6 | 1 | 104 | -11.375 | 11.375 | 1.892229 | - | studytime <= '1.0' |
| 7 | 1 | 79 | -11.379... | 11.379... | 2.517301 | - | Fedu <= '1.0' |
| 8 | 1 | 42 | -11.428... | 11.428... | 1.74769 | - | schoolsup = '1' |
| 9 | 1 | 116 | -11.465... | 11.465... | 2.375997 | - | Walc >= '3.0' |
| 10 | 1 | 92 | -11.478... | 11.478... | 2.040238 | - | Dalc >= '2.0' |

課堂犯錯1次以上的學生，子群組的成績平均值為較低的9.725！

Cortana 子群組探勘實戰B


# 學生成效分析
## 找出學校跟成績有差異的子群組

# 兩所學校的學生成績比較

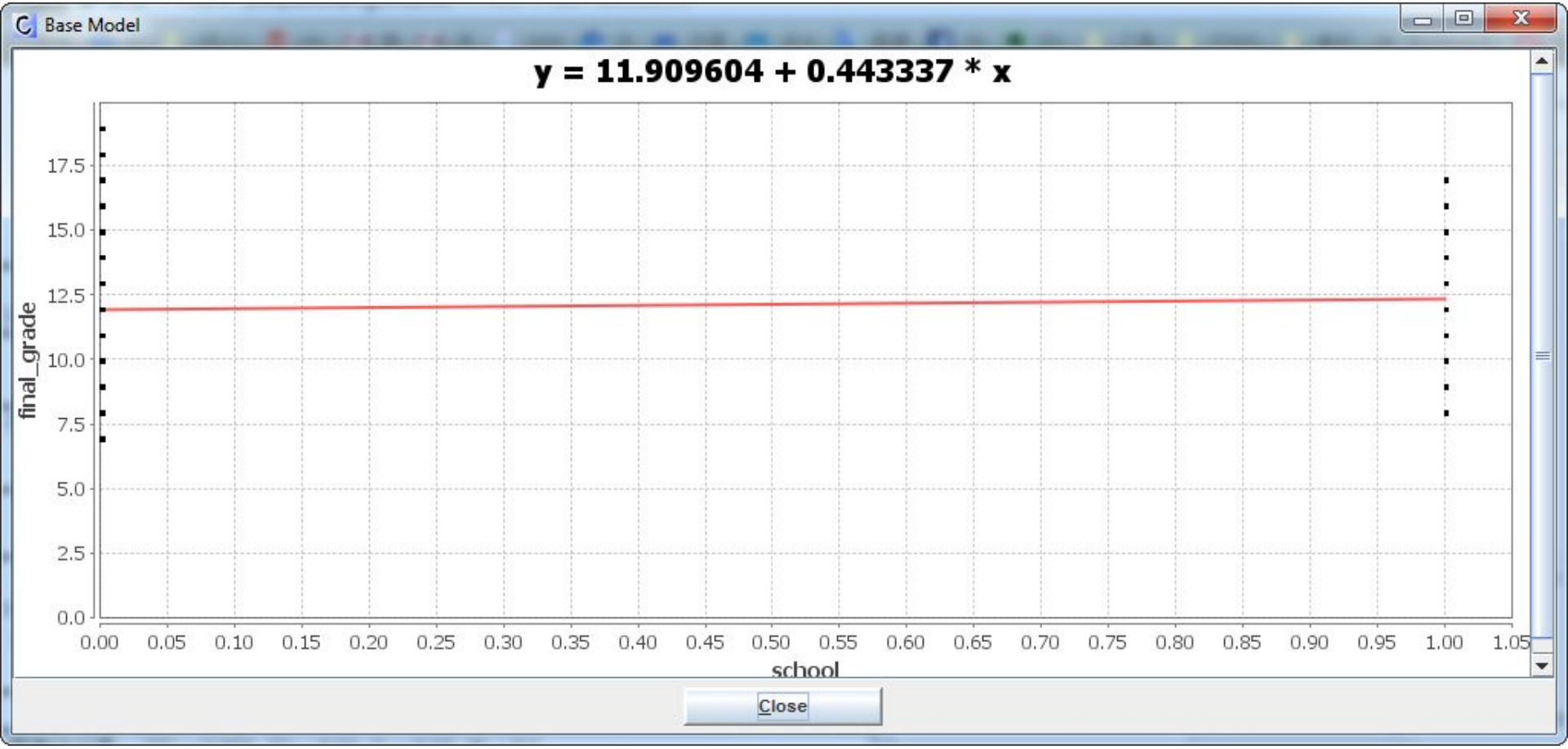



實驗組學校:136位學生，平均成績12.3529

控制組學校:177位學生，平均成績11.9096

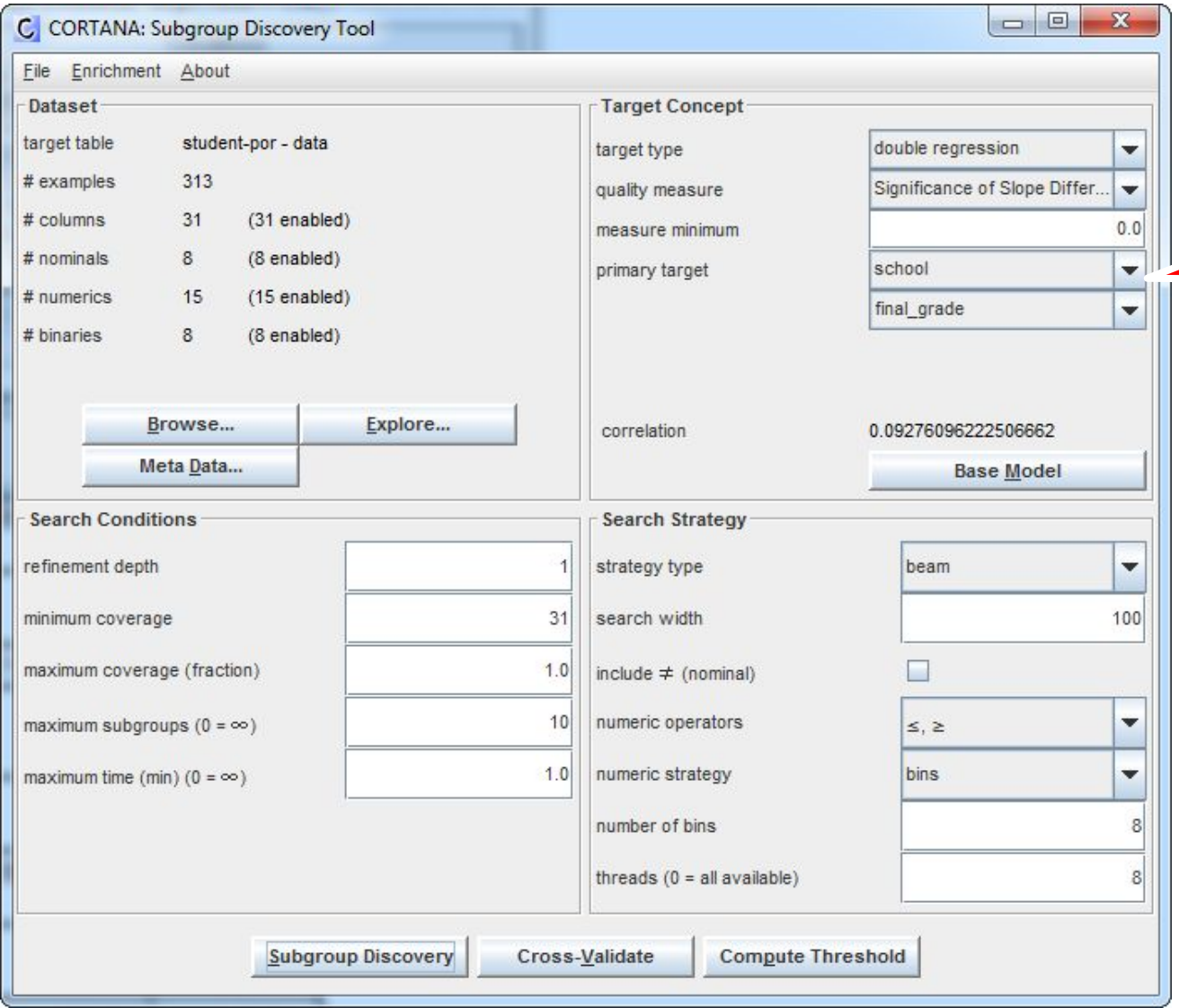
t值1.64，p值0.1，差異未達顯著

# 基礎模型




看不出什麼差別

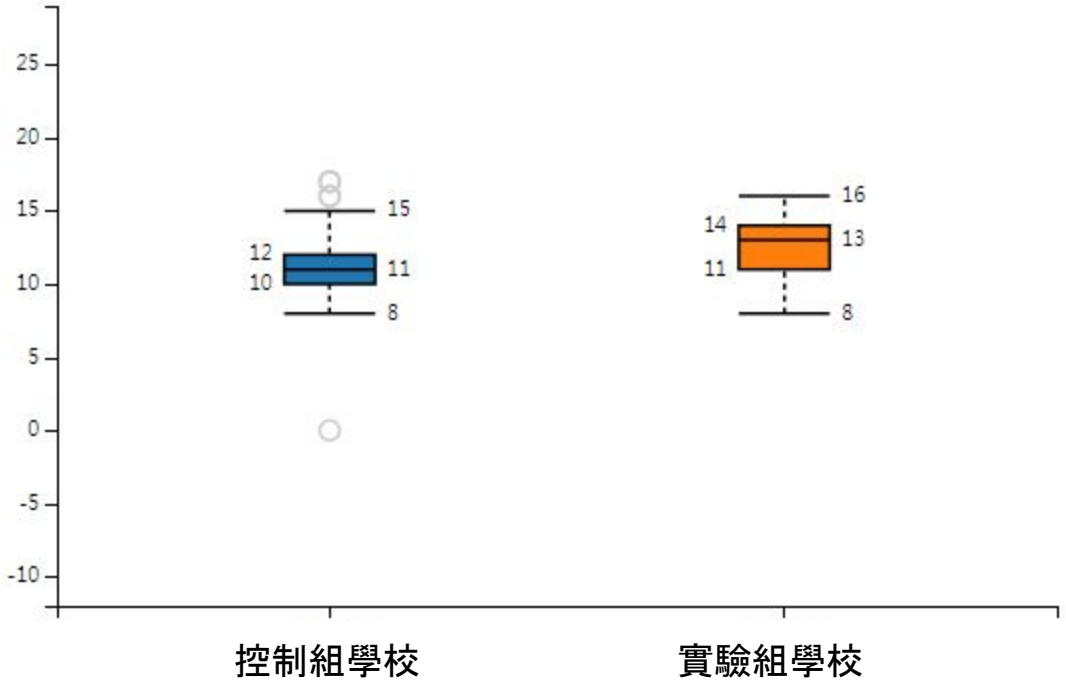
# 從學校能否預測成績呢? (迴歸)




用學校來預測成績

# 自由時間多寡似乎容易影響成績

# 僅選出freetime>=4的子群組




實驗組學校(school-2)：47位學生，平均成績12.5106

控制組學校(school-1)：76位學生，平均成績11.25

t值-2.98，p值0.0034，差異達到顯著

分析技術就這樣了

案例整理還有待處理

# 報告完畢