調查資料去識別化過程與驗證

－以身心障礙者生活狀況調查為例

衛生福利部統計處
李美鈴
106/3/15

大綱

一、前言

二、去識別化驗證辦理程序

(一)建立管理文件

(二)建立去識別化工具

(三)第三方驗證

三、結語

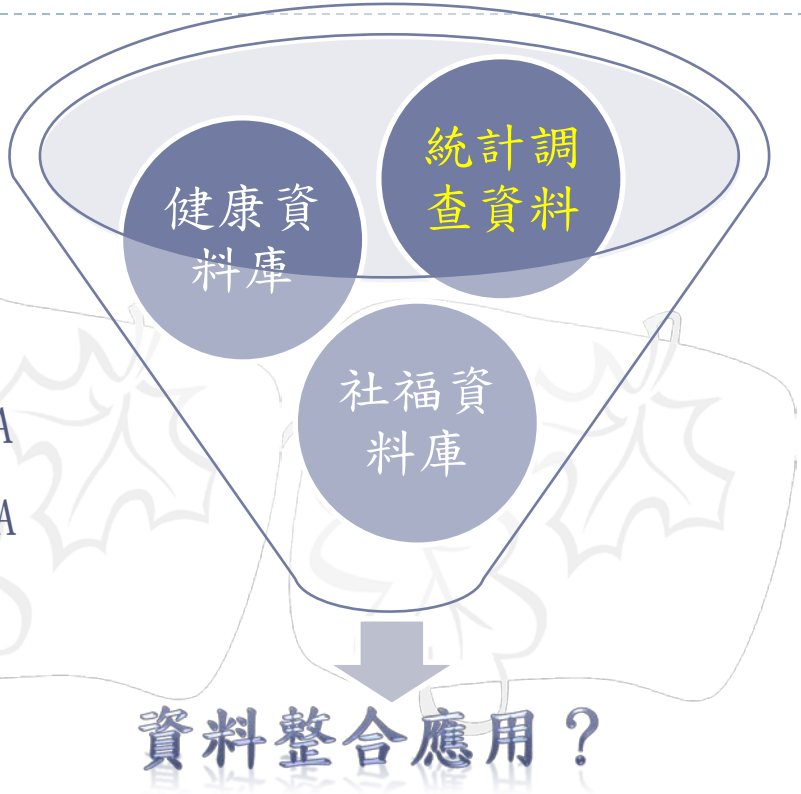
前言(1/4)

資訊進步　　網路發達

社群網站　　APP

BIG DATA

OPEN DATA

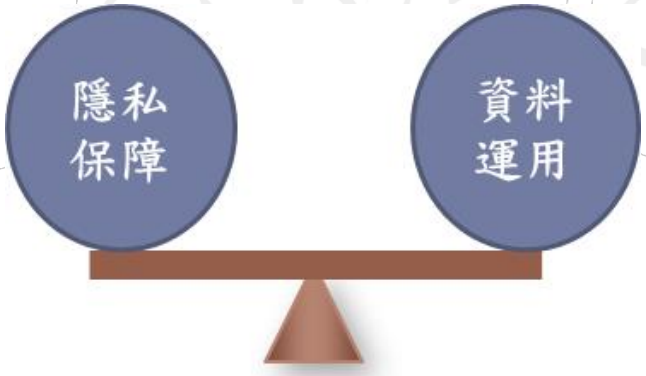


隱私安全的疑慮

資料合理運用？

前言(2/4)

- 全民健康保險資料庫訴訟案：101年人權團體針對健保署將健保資料給國衛院及衛生福利部資料科學中心供學術使用，提起訴訟。



前言(3/4)

美國

. 法規散見不同部門。

. 以個人醫療資料為例，健康保險可攜及責任法案(HIPAA)，係將18種識別資料移除，來判定是否去識別化。

歐盟

. 去識別要達到「匿名」狀態(anonymous)。

. 以資料保護工作小組(The Article 29 Working Party)之指引文件(WP 216)，作為判定是否完成「去識別化」之參考。

日本

. 2015年個資法修法，增訂「匿名化處理資料」概念，並增設「個人資料保護委員會」專責機關。

. 未來由該單位負責「匿名化處理資料」之監督管理與相關配套規範之訂定。

前言(4/4)

法務部函釋
（法務部103年11月17日法律字第10303513040號函）

- 若資料經過「去識別化」處理，該資料即非個資，而能開放使用。

行政院
去識別化會議

- 責成經濟部標準檢驗局訂定相關程序。
- 希望建立中立第三方驗證機制。
- 由政府機關帶頭示範。

標準局制定
驗證標準規範

- CNS 29100及CNS 29191為驗證標準。
- 研訂「個人資料去識別化過程驗證要求及控制措施」。
- 由財政資訊中心辦理第一波示範。

去識別化驗證辦理程序

104年12月接獲行政院指示

105年1月

確認驗證範圍

4~10月 文件檢視及討論會議(11次)

10.26 提出驗證申請

11.4 內稽

11.21 ~23 外稽

12.5 取得驗證

召開研商會議

成立管理小組

建立管理文件

隱私風險評鑑

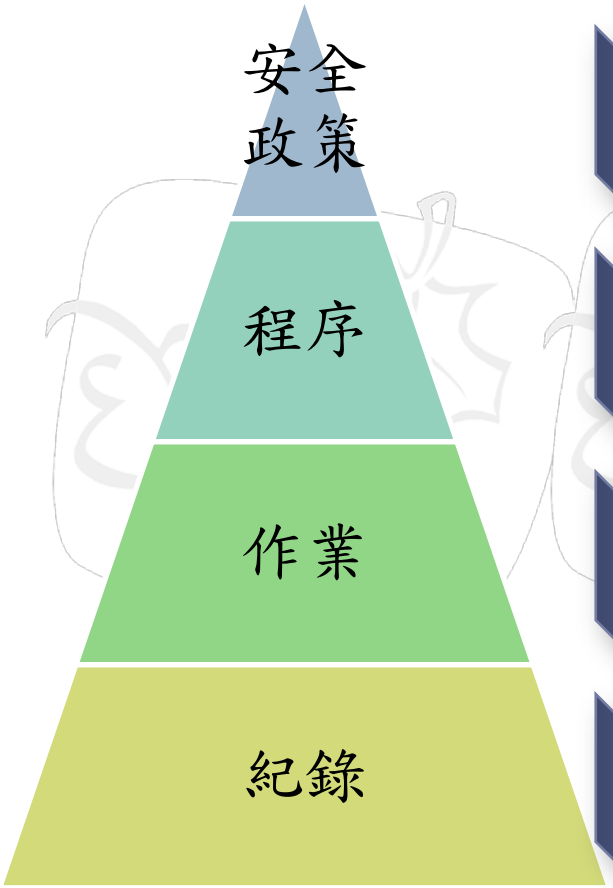
去識別化處理

建立去識別化工具

教育訓練

第三方驗證

去識別化驗證辦理程序
－建立管理文件



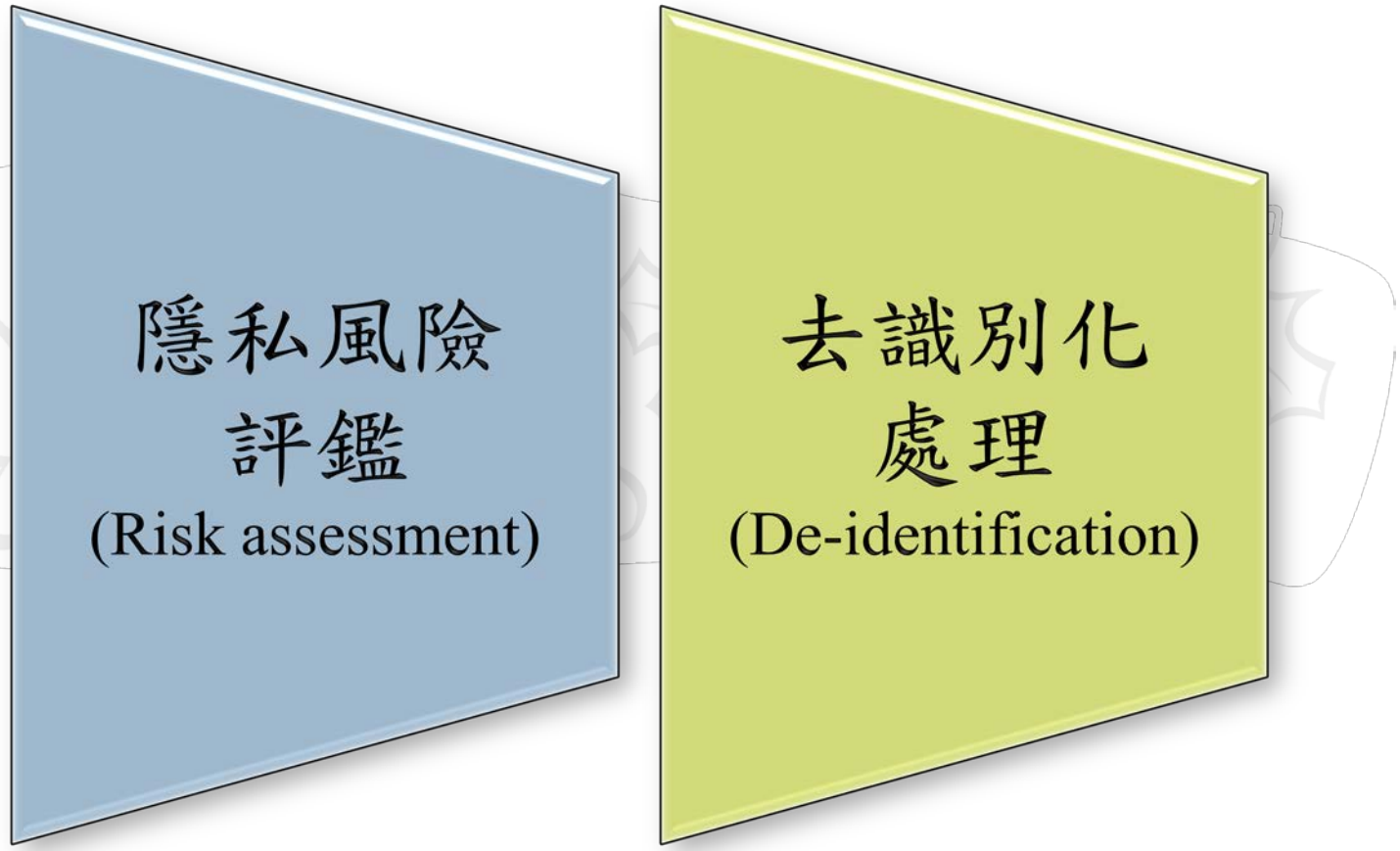
一階(2)
- 隱私權政策
- 適用性聲明書

二階(8)
- 8份管理程序書：文件與紀錄、組織全景評鑑、隱私風險、稽核、組織、人力資源安全、委外作業、非預期資料揭露

三階(1)
- 去識別化處理標準書

四階(48)
- 各文件紀錄表單

去識別化驗證辦理程序－建立去識別化工具



建立去識別化工具
－隱私風險評鑑（1/10）

- 平衡資料隱私與資料可用性。
- 建置風險模型，量化評估「重新識別」風險。
- 訂定可容忍之風險門檻值。

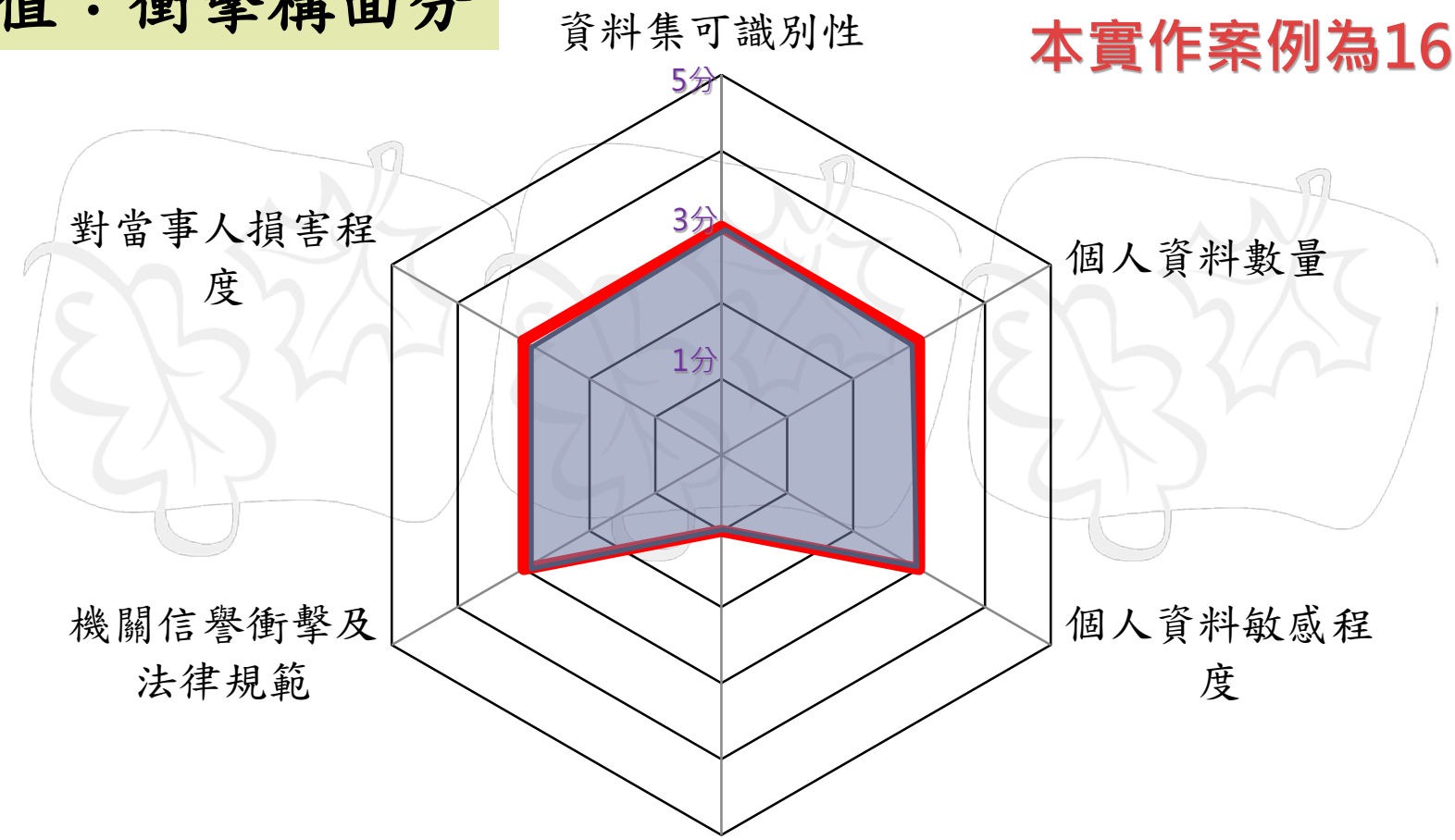
建立去識別化工具
－隱私風險評鑑（2/10）

風險值= 衝擊值 X 重新識別可能性

建立去識別化工具
－隱私風險評鑑（3/10）

衝擊值：衝擊構面分

本實作案例為16分



建立去識別化工具
－隱私風險評鑑（4/10）

重新識別可能性

取最高之機率值		
重現性 (Replicability)	資源可用性 (Resource Availability)	區別性 (Distinguish)

建立去識別化工具
－隱私風險評鑑（5/10）

- 不變動型資料集：1
- 部分變動型資料集：0.75
- 變動型資料集：0.5

- 資料接收者團隊恰巧認識資料集中個體的數值。
- 通常150~230。
- 本實作案例為200 X 身障人口占全國人口比率

$$\text{重現性} = \text{權重} \times \frac{\text{鄧巴數•身障人口比}}{\text{資料集提供筆數}}$$

$$= 1 \times \frac{200 \times \text{身障人口占全國人口比率}}{\text{調查人口}}$$

$$= 1 \times (200 * 0.049 / 19{,}301) = 0.000508$$

建立去識別化工具
－隱私風險評鑑（6/10）

- 利用外部4種類型網站測試

$$\text{資源可用性} = \frac{\text{可重新識別資料筆數}}{\text{重新識別測試資料筆數}}$$

本實作案例為0

抽出96筆測試

- 樣本為19,301筆。
- 基於95%信心水準,誤差為正負10%。
- 共抽樣96筆進行重新識別測試。

建立去識別化工具
－隱私風險評鑑（7/10）

公開/申請外釋
攻擊機率=1

區別性=	攻擊機率x	猜中的最高機率

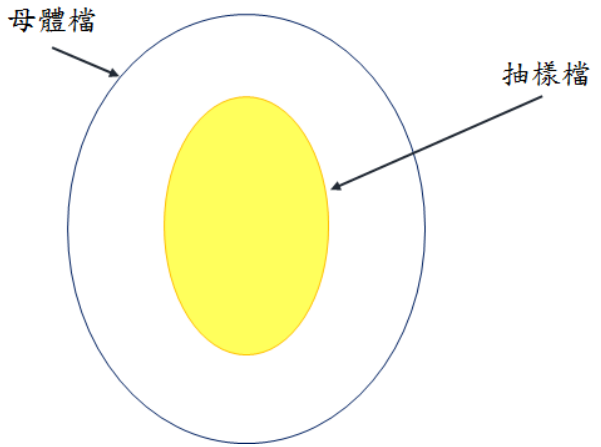
本實作案例為0.006

猜中個體的最高機率＝資料正確性x抽樣率 x 抽樣檔
在母體中被猜中的最高機率
$=1x0.0179x\frac{1}{3}=0.006$

建立去識別化工具
－隱私風險評鑑（8/10）

- Journalist模式之K-匿名處理，且K至少為3

- K=MIN(抽樣檔在母體之間接識別資料組合數)。



建立去識別化工具
－隱私風險評鑑（9/10）

樣本資料

年齡別	障礙類別	性別	人數
15	視覺障礙	女	1
16	視覺障礙	女	1
30	失智症	男	1
32	失智症	男	1

母體資料

年齡別	障礙類別	性別	人數
15	視覺障礙	女	2
16	視覺障礙	女	1
30	失智症	男	2
32	失智症	男	5

去識別化處理

樣本資料

年齡別	障礙類別	性別	人數
15-19 歲	視覺障礙	女	2
30-34 歲	失智症	男	2

母體資料

年齡別	障礙類別	性別	人數
15-19 歲	視覺障礙	女	3
30-34 歲	失智症	男	7

建立去識別化工具
－隱私風險管理（10/10）

定義可接受風險值為0.2 ↓

衝擊構面評分6～30分 ↓

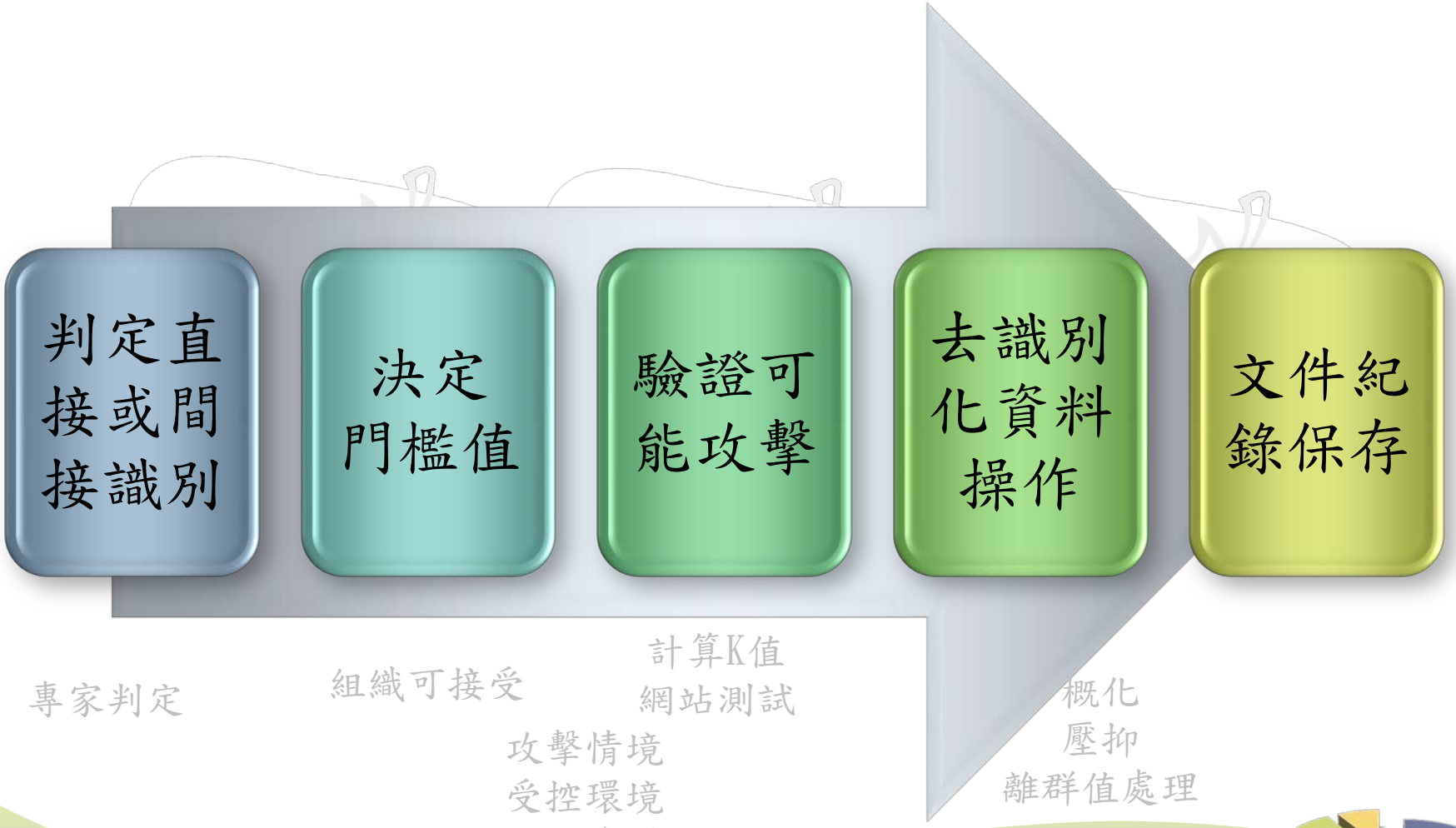
重現性、資源可用性及區別性三者擇其最高者 ↓

風險值=衝擊值X重新識別可能性

本實作案例：

0.096 = 16 X 0.006

建立去識別化工具環境 –去識別化處理(1/4)



建立去識別化工具 －去識別化處理(2/4)

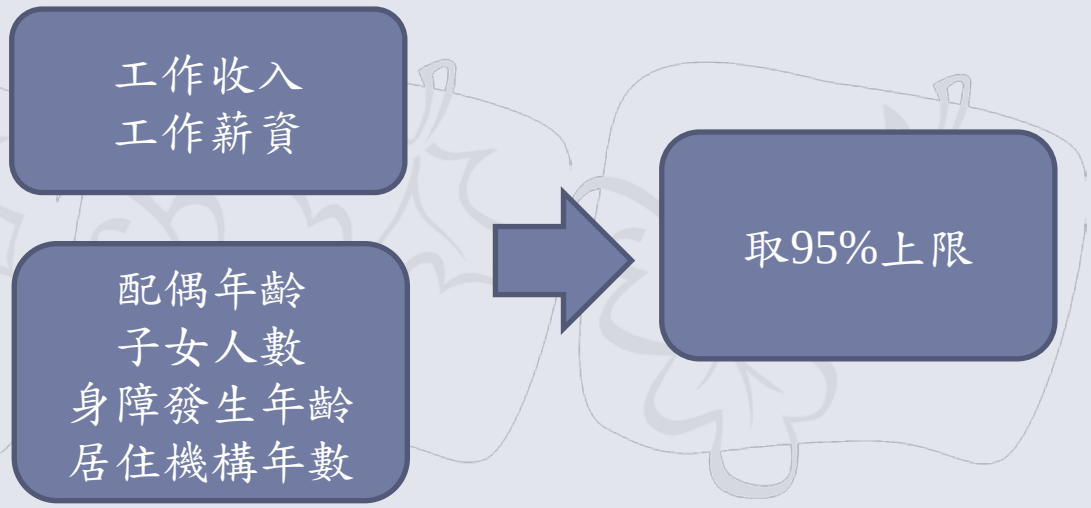
- 個資去識別化處理的方法，不是單選題，處理方式可能包括以下組合：
  - 遮罩(masking)
  - 概化(Generalization)
  - 壓抑(Suppression)
  - 次抽樣(Subsampling)

視去識別化目的、資料欄位、所要串接的資料，是否可能連結導致重新識別等因素，綜合考量後，再決定去識別化方法。

建立去識別化工具
－去識別化處理(3/4)

去識別化處理	說　明
概化 (Generalization)	.障礙類別：罕見疾病併入其他 .身障者年齡：轉換5歲年齡組。
壓抑 (Suppression)	將5筆未符合K=3之部分欄位資料刪除。

建立去識別化工具
－去識別化處理(4/4)

去識別化處理	說　　明
離群值	

去識別化驗證辦理程序－第三方驗證(1/3)

- 驗證(Certification)：係透過中立第三方，證明資料經去識別化後，是否達到隱私風險的標準。
- 協助機關檢視隱私保障相關作業管理措施及流程是否完備。

去識別化驗證辦理程序－第三方驗證(2/3)

提出驗證申請準備工作：

- **備妥驗證標的實施過程文件**：已依「個人資料去識別化過程驗證要求及控制措施」實施個人資料去識別化過程。
- **完成人員能力訓練**：相關人員已接受相關訓練活動。
- **完備監控紀錄**：各項個人資料去識別化過程相關流程、改善方案等已有執行或監督量測等紀錄。
- **完成事前審查**：至少執行過1次（含）以上之內部稽核及管理審查。

去識別化驗證辦理程序－第三方驗證(3/3)

105.12.5
通過驗證


個人資料去識別化過程驗證證書
ETC
證明
衛生福利部統計處
地址：11558 台北市南港區忠孝東路6段488號4樓
財團法人台灣電子檢驗中心
執行長 李海清

- **驗證機關**：財團法人台灣電子檢驗中心
  (Electronics Testing Center,Taiwan)
- **驗證流程**：

驗證申請 → 訪談 → 實地稽核 → 審查 → 驗證通過

結語

- **知能培力**：累積經驗培養同仁辦理去識別化工作之基礎知識與能力，已應用於死因開放資料集之去識別化檢視。
- **擴大推動**其他資料集：
  - 其他外釋單一橫斷面抽樣調查資料集
  - 世代追蹤型抽樣調查資料集
  - 世代追蹤健康抽樣資料集
- **開發多元**去識別化**工具**

謝謝聆聽，敬請指教