Example Table

↓sensitive

name	sex	age	passport	salary
Alice	F	20-29	96273697	0-20k
Bob	M	20-29	53331879	80k-90k
Carol	F	30-39	27429029	40k-50k
Dave	M	30-39	53286237	50k-60k
Eve	F	40-49	43485839	30k-40k

IDentifier, as Intended

name	sex	age	passport	salary
Alice	F	20-29	96273697	0-20k
Bob	M	20-29	53331879	80k-90k
Carol	F	30-39	27429029	40k-50k
Dave	M	30-39	53286237	50k-60k
Eve	F	40-49	43485839	30k-40k

IDentifier, as Intended

name	sex	age	passport	salary
*****	F	20-29	*****	0-20k
*****	M	20-29	*****	80k-90k
*****	F	30-39	*****	40k-50k
*****	M	30-39	*****	50k-60k
*****	F	40-49	*****	30k-40k

Quasi-IDentifier

name	sex	age	passport	salary
*****	F	20-29	*****	0-20k
*****	M	20-29	*****	80k-90k
*****	F	30-39	*****	40k-50k
*****	M	30-39	*****	50k-60k
*****	F	40-49	*****	30k-40k

Scenario. I know my friend Alice is a female in her 20s. Then I know how much she makes.

QID and k-anonymity

name	sex	age	passport	salary
*****	F	20-29	*****	0-20k
*****	M	20-29	*****	80k-90k
*****	F	30-39	*****	40k-50k
*****	M	30-39	*****	50k-60k
*****	F	40-49	*****	30k-40k

A *quasi-identifier*(QID), a là Tore Dalenius(1986), is a set of attributes that **uniquely identifies** entries. In this example, there are two sets of QIDs: $\{sex, age\}$ and $\{salary\}$.
A privacy property that protects against this type of re-identification is *k-anonymity*, which, per Latanya Sweeney(2002), holds if for every entry, there are at least $k-1$ other entries with the same values for some QID.

k-anonymity Example

name	sex	age	salary
Frederick	M	20-29	0-20k
Gabriel	M	20-29	80k-90k
Helen	F	30-39	40k-50k
Ingrid	F	30-39	50k-60k

This table conforms to 2-anonymity with respect to QID $\{sex, age\}$.
Scenario. Even if I know Frederick is a male in his 20s, I still can't be sure how much he makes.

2010 年人口及住宅普查

- We analyzed 全臺閩地區人口狀況資料檔 (/資料檔/p99c1.txt).
- A total of 3,618,831 entries.
- 48 fields, whereof 39 are relavant to the analysis.
  - Such as age, sex, nationality, marital status, education status...
- Fields are mostly boolean or categorical.
- Specifications are in /附件/99年人口及住宅普查-全臺閩地區格式.

- We found that of the 3,618,831 entries, there are only 1,364,624 distinct values.
  - On average, for every individual, there are 2 other people that share the same values as that individual.

What We Found

- The "minimum" QID, the smallest set of attributes that uniquely identifies all 1,364,624 distinct entries, is of size 35.
  - C010:性別、C020:年齡、C032:國籍、C050:婚姻狀況、C060: 與戶長關係、C071:教育狀況、C072:教育程度、C073:是否 上幼稚園(學齡前兒童)、C081:國語、C082:閩南語、C083:客 家語、C084:原住民族語、C085:其他、C091:國語、C092: 閩南語、C093:客家語、C094:原住民族語、C095:其他、C100: 五年前居住地點、C110:主要家計負責人、C121:工作狀況、C122: 行業、C123:職業、C132:工作或求學地點、C141:子女人數、C142: 子女居住地點、C151:吃飯、C152:上下床、C153:更換衣服、C154: 上廁所、C155:洗澡、C156:在室內外走動、C157:家事活動 能力、C160:是否為原住民族、C170:是否為身心障礙.
- This was found and verified in days.

What We Found

- We also found QIDs that distinguishes some percentage of the entries, that is, the QIDs that picks out some percentage of entries with distinct values.
- For 10%, there are two minimum sets of QIDs of size 5. This was found and verified in days.
  - C020(年齡), C060(與戶長關係), C072(教育程度), C122(行業), C123(職業) and
  - C020(年齡), C072(教育程度), C122(行業), C123(職業), C141(子女人數)
- Some small QIDs for other percentages (unverified):

percentage	QID size
25	8
50	10
75	15