Statistical Disclosure Control

Basic Concepts

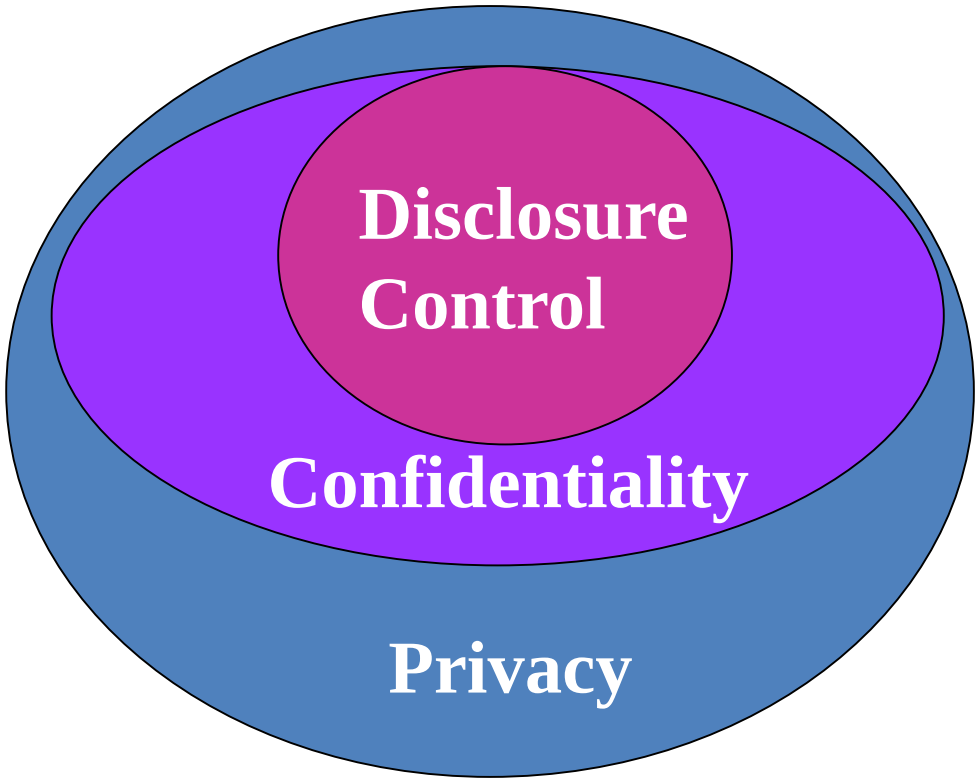
Professor Mark Elliot

Outline

- What is a statistical disclosure
- How might statistical disclosure happen?

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

Privacy, confidentiality and disclosure



What is Statistical Disclosure Control (SDC)?

Statistical disclosure control (SDC) is the practice of reducing the risk of:

finding people (or other entities) in data: ***Re-identification***

and/or

associating data with a person (or entity): ***Association***

What is Statistical Disclosure Control (SDC)?

Need to strike the right balance between maximising data utility (including meeting customer requirements) and management of confidentiality risk.

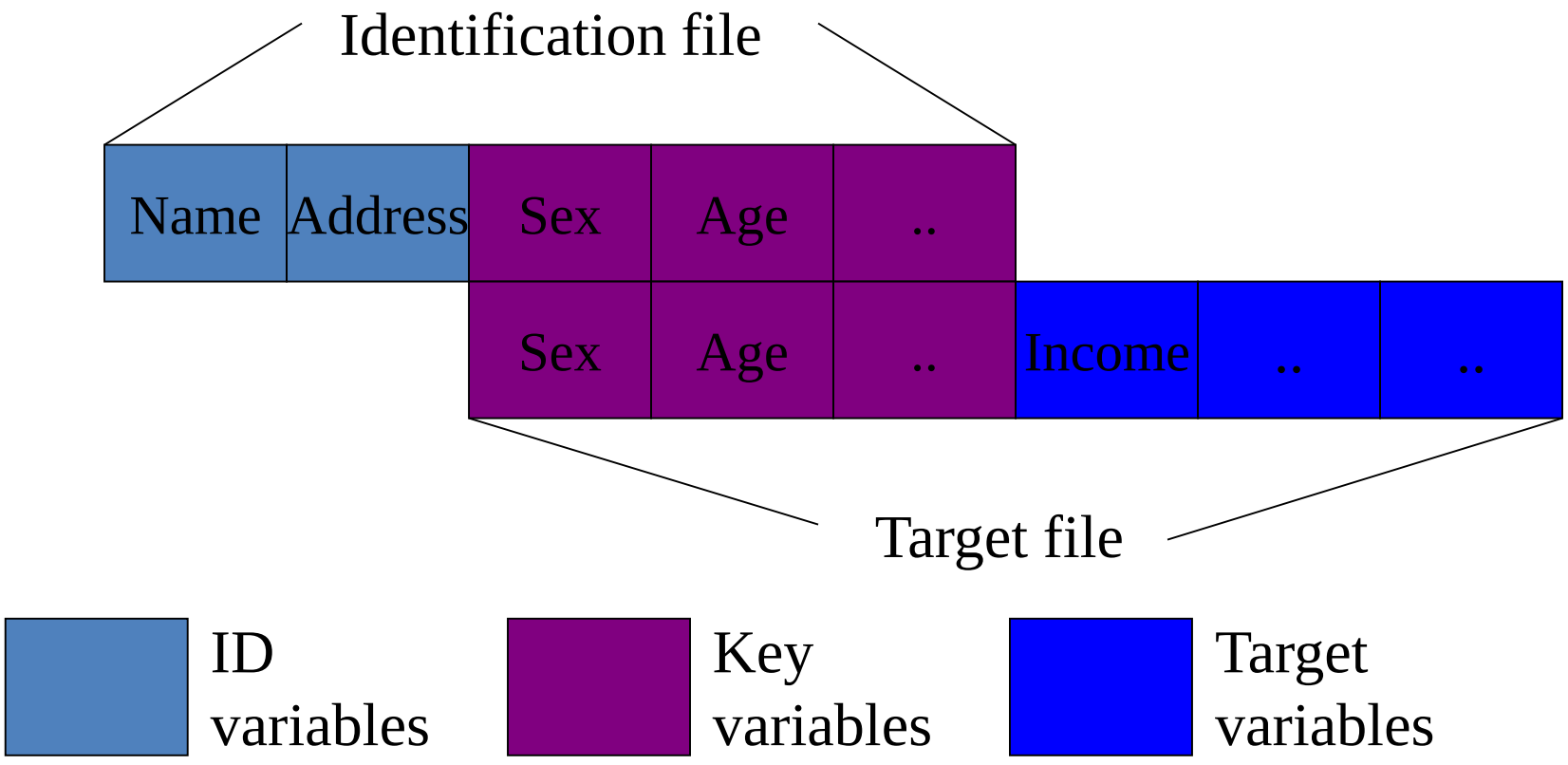
Statistical disclosure is itself an active research area

- Sub fields
  - **Disclosure risk assessment**
  - Disclosure control methodology
  - Measurement of analytical validity
  - Data Environment Analysis
- All data types
  - **Typically Microdata and Aggregate data**
  - Business and Personal data
  - Intentional and Consequential data

How might a disclosure happen?

- Imagine you are a “data intruder”
  - What would you need to do in order to identify information about individuals within anonymised data?
  - What might be your motivations?
- In what other ways might a statistical disclosure happen other than malicious intrusion?

The Disclosure Risk Problem: Type I: Identification



The Disclosure Risk Problem
Type II: Attribution

Income levels for two occupations				
	High	Medium	Low	Total
Professors	0	100	50	150
Pop stars	100	50	5	155
Total	100	150	55	305

The Disclosure Risk Problem: Type III: Subtraction

Income levels for two occupations				
	High	Medium	Low	Total
Professors	1	100	50	151
Pop Stars	100	50	5	155
Total	101	150	55	306

The Disclosure Risk Problem
Type III: After subtraction

Income levels for two occupations				
	High	Medium	Low	Total
Professors	0	100	50	150
Pop Stars	100	50	5	155
Total	100	150	55	305

The Disclosure Risk Problem
Type IV: Table linkage

	Tenure	
Age	O	R
Young	3	9
Old	2	2

	Tenure	
HIV?	O	R
N	1	10
Y	4	1

	Age	
HIV?	Y	O
N	8	3
Y	4	1

The Disclosure Risk Problem
Type IV: Table linkage

	Age and Tenure			
HIV?	O,R	O,O	Y,R	Y,O
N	2	8	1	0
Y	0	1	1	3

- Original cell counts can be recovered from the marginal tables

The Disclosure Risk Problem: Other data types

- Network data
- Qualitative data
- Genomics Data
- Stream Data
- Mixed data – Jigsaw identification

Summary

- Statistical disclosure is a complex topic
  - Still an active research field
- As researchers using sensitive/personal data you will need to:
  - Be aware of the issues and considerations of statistical disclosure
  - Be able to make principled judgements about the disclosiveness of your output