Anonymisation, Risk and Privacy

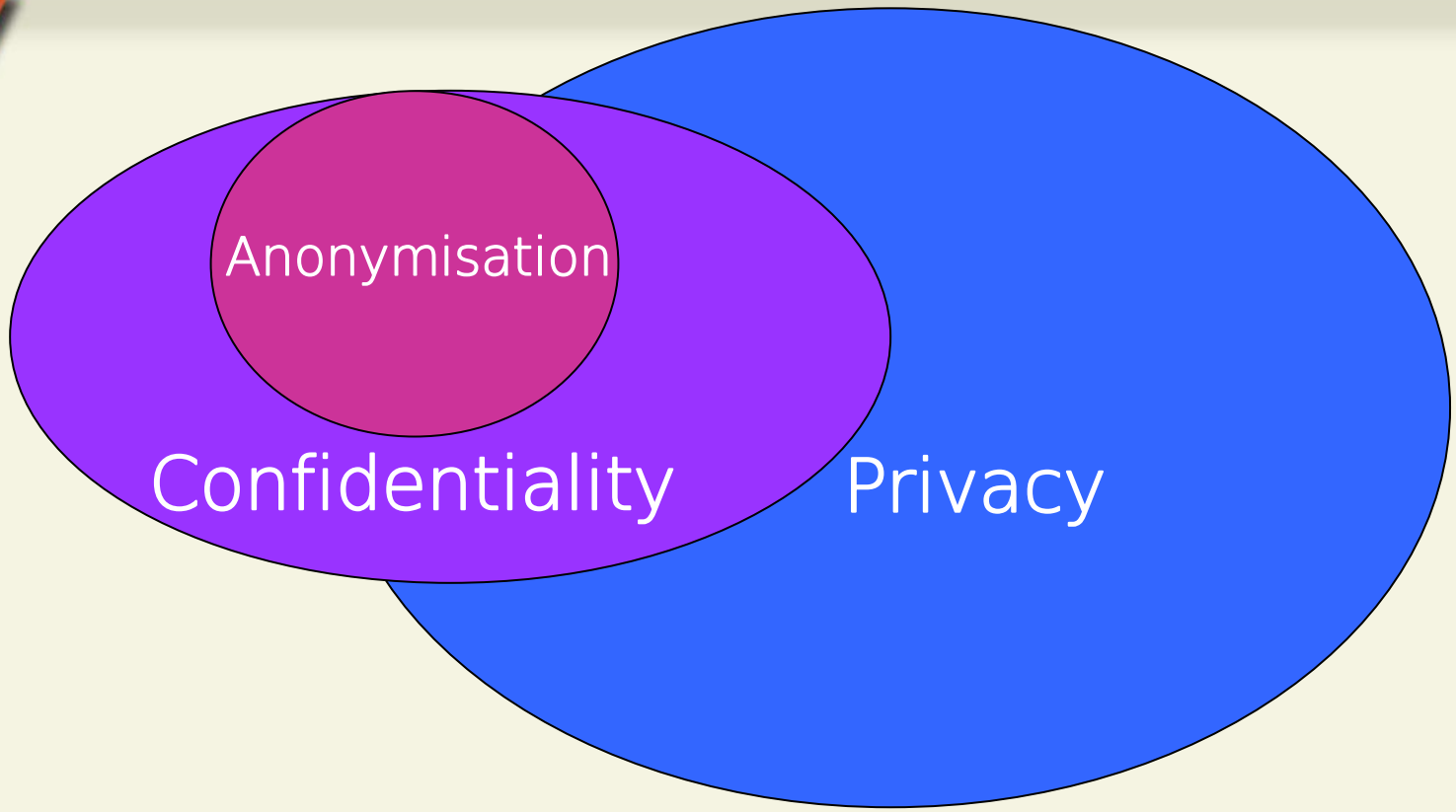
Mark Elliot, University Of Manchester

Presentation to the *New Approaches to data Privacy* Workshop. Isaac Newton Institute, Cambridge December 2016

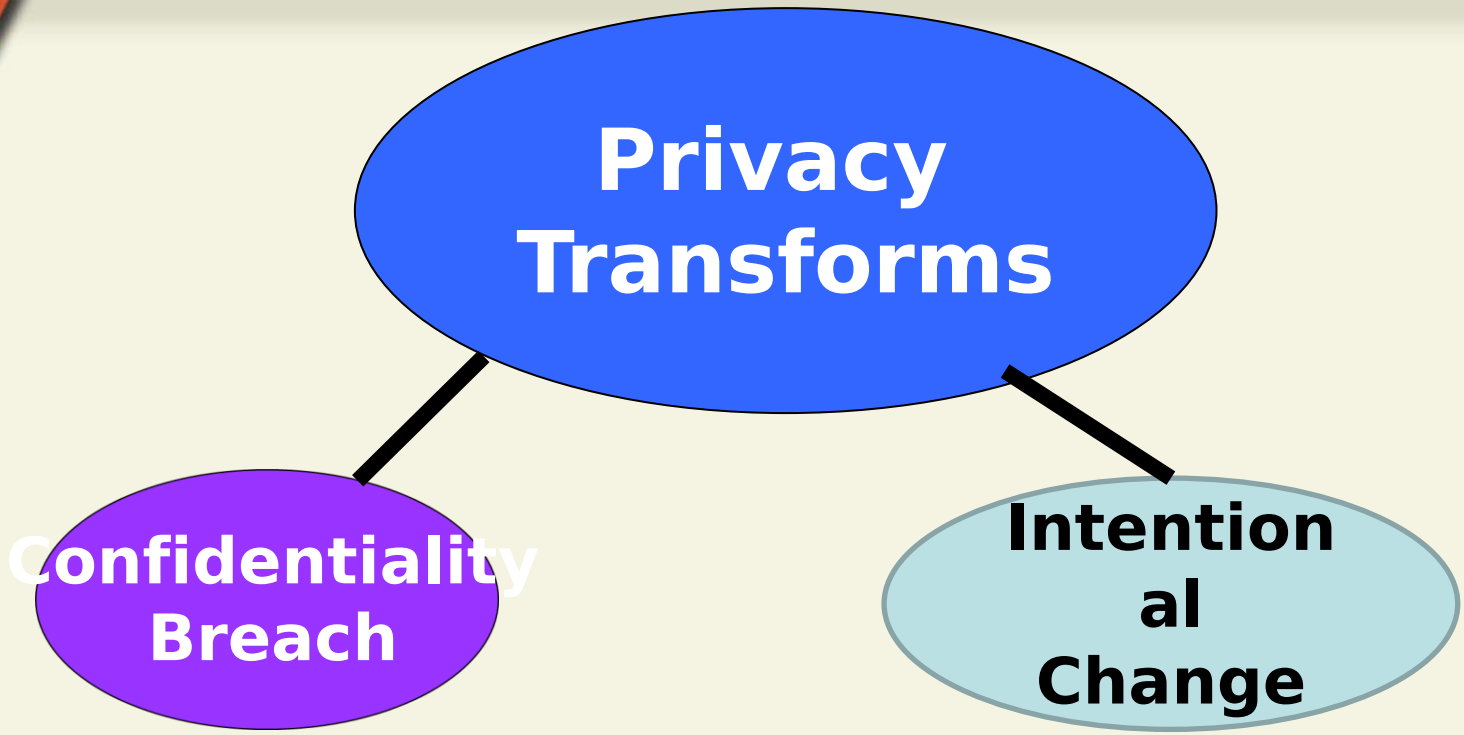
Outline

- Conceptual Clarification
- Privacy and Risk
- The Meaning of Anonymisation
- Risk Assessment
- Controlling Risk

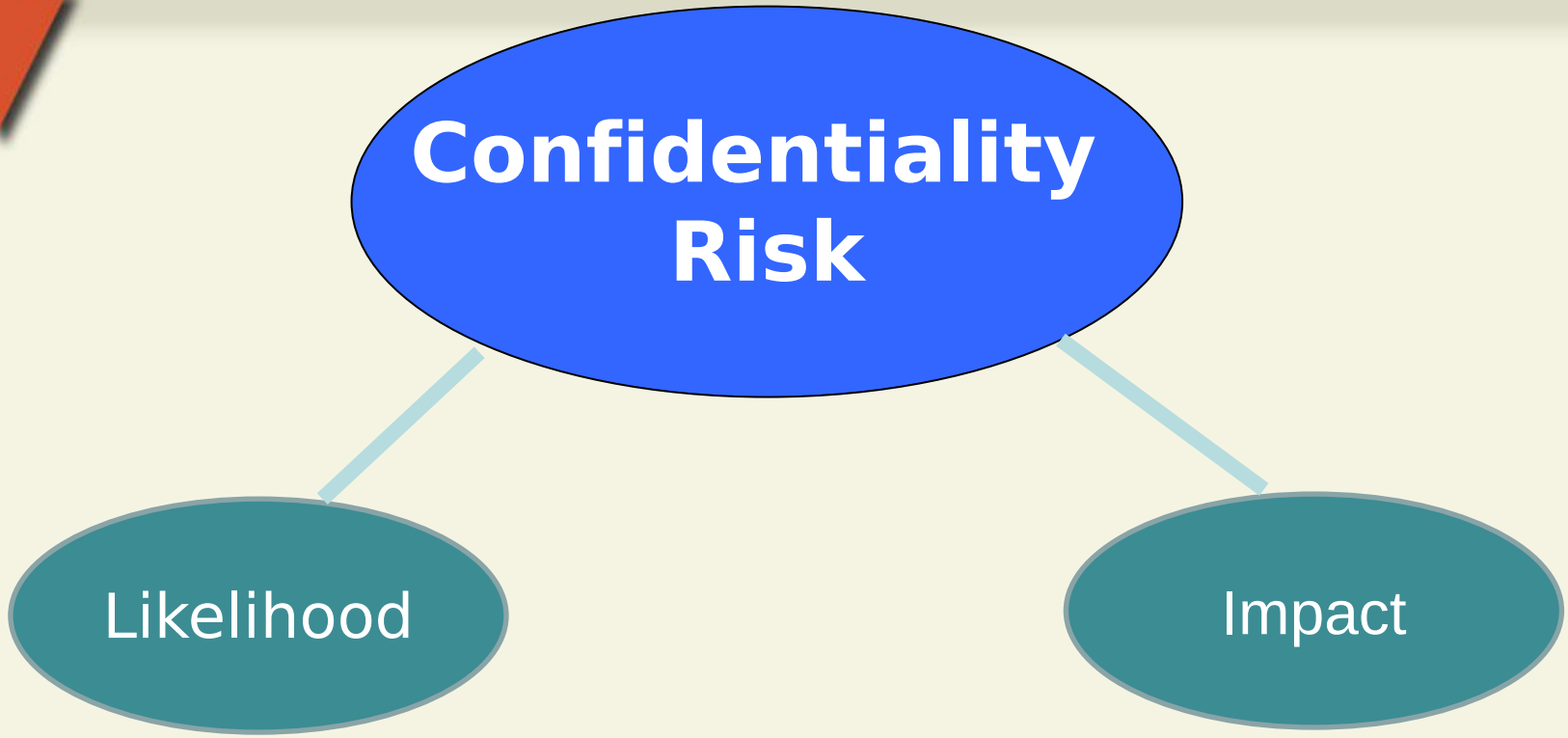
Privacy, Confidentiality and Anonymisation



Privacy and Risk….



Confidentiality and Risk



Don’t confuse “could” with “will”.

Marsh et al (1991)

$$pr(identification) =$$

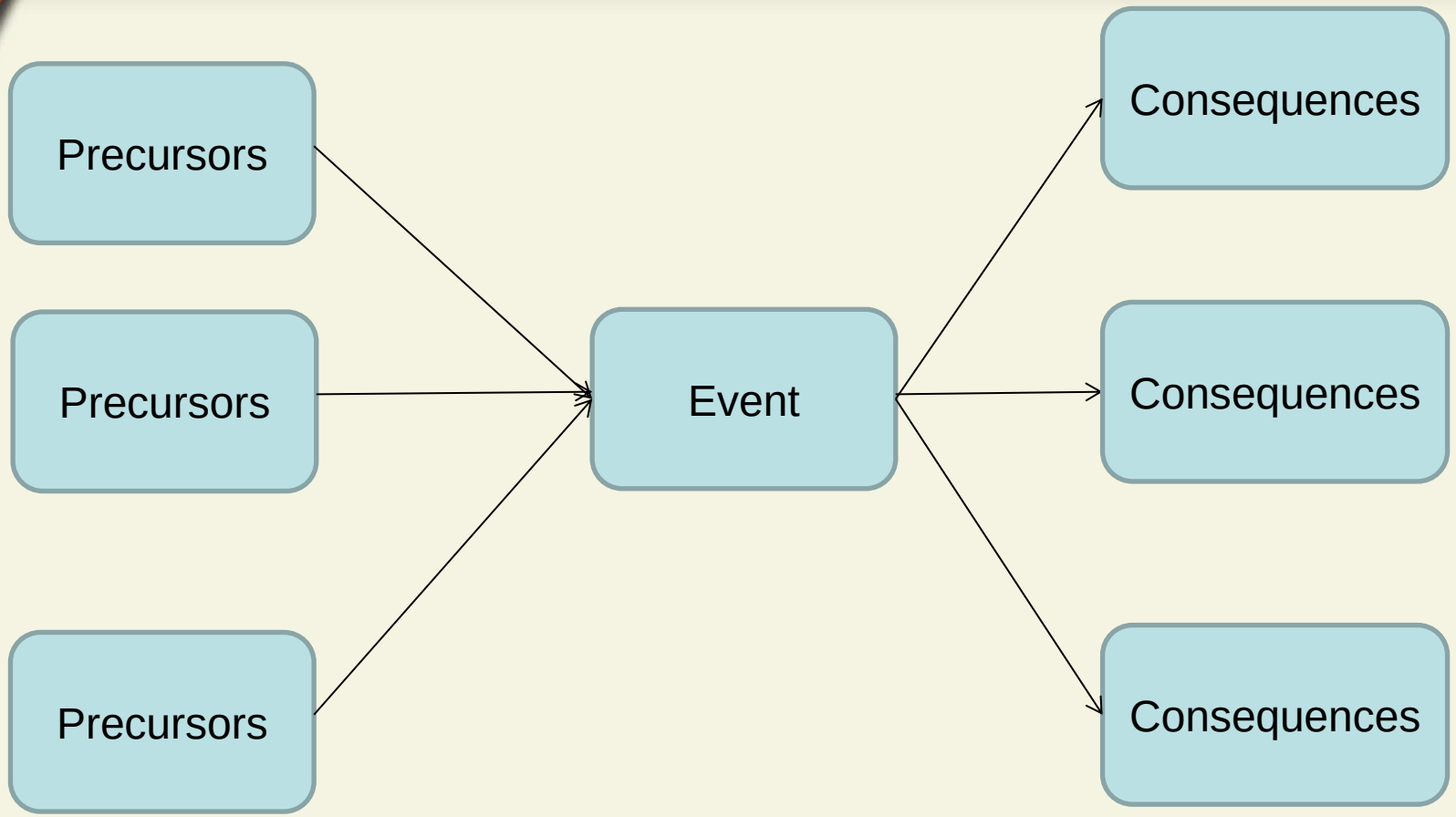
$$pr(attempt) \times pr(identification|attempt)$$

Confidentilaity and risk

Don’t confuse “could” with “will”.

Treating Privacy breaches as apocalyptic has lead to some very muddled thinking….

Confidentiality and risk



What is Anonymisation?

Anonymisation is **process** by which personal data are rendered non personal.

What is Anonymisation?

Anonymisation is **process** by which personal data are rendered non personal.

Avoid using success terms:

“anonymised”

What is Anonymisation?

Anonymisation is **process** by which personal data are rendered non personal.

Avoid using success terms

“anonymised”

or worse “truly anonymised”

What is Anonymisation?

Anonymisation is **process** by which personal data are rendered non personal.

Avoid using success terms

“anonymised”

or worse “truly anonymised”

And “really truly anonymised” is right out

Anonymisation and de-identification

- Deal with different parts of the normal definition of personal data
- De-identification tackles:
  - **“Directly from those data”**
- Anonymisation tackles:
  - **“Indirectly from those data and other information** which is in the in the possession of, or is likely to come into the possession of, the data controller…”

Some tenets

- Anonymisation is not about the data.
- Anonymisation is about **data situations**.
- Data situations arise from data interacting with data environments.

Some tenets

- Data environments are:
  - *the set of formal and informal structures, processes, mechanisms and agents that either:*
    i. *act on data;*
    ii. *provide interpretable context for those data or*
    iii. *define, control and/or interact with those data.*

Elliot and Mackey (2014)

Anonymisation types

- Absolute Anonymisation
  - Zero possibility of re-identification under any circumstances
- Formal Anonymisation
  - De-identification (including pseudonymisation)
- Statistical Anonymisation
  - Statistical Disclosure Control
- Functional Anonymisation

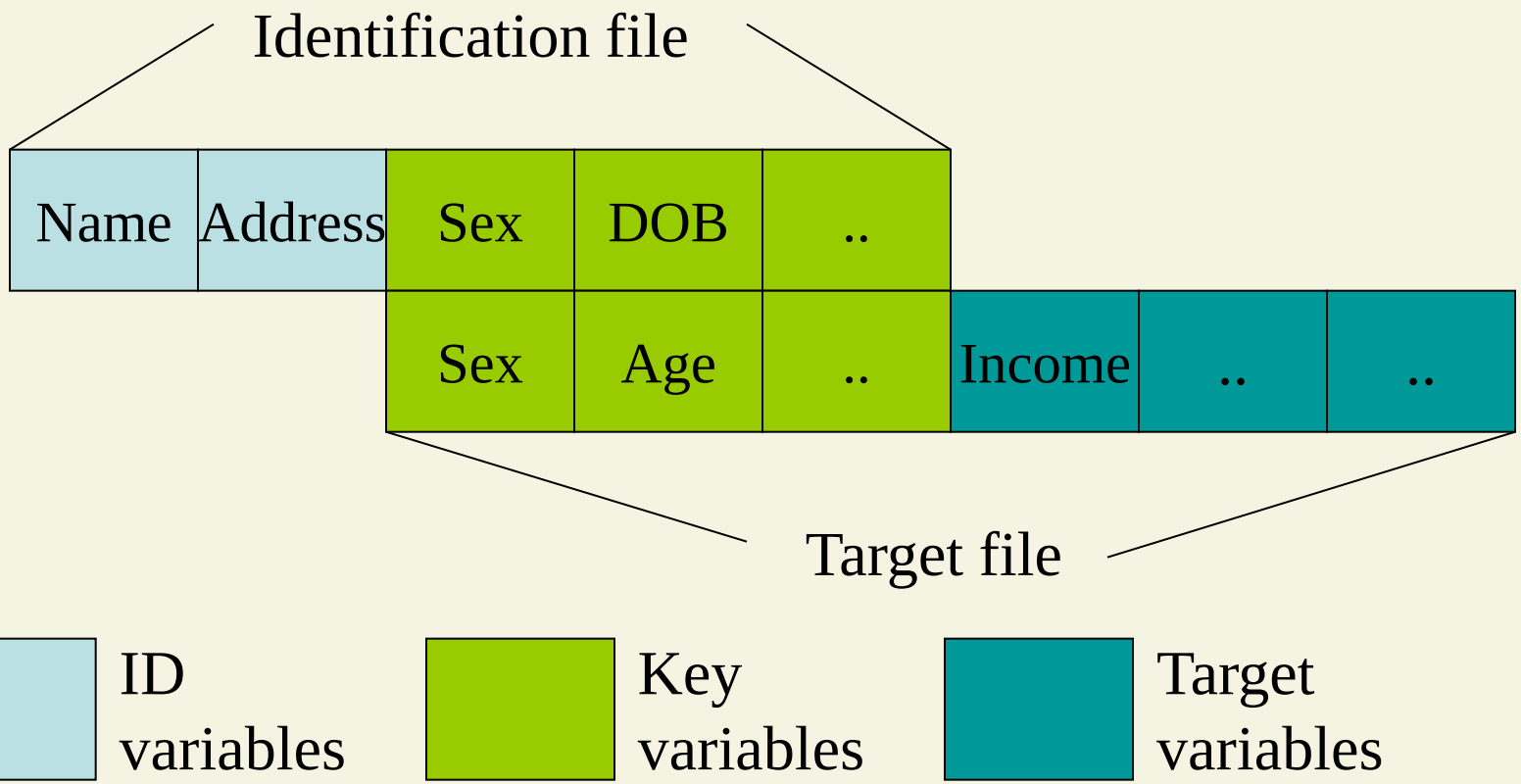
Unintended Disclosure

- Consists of two processes:
  - Identification: the (correct) association of a population unit and a data unit
  - Attribution: the (correct) association or disassociation of an item of data with a population unit

Unintended Disclosure

- Consists of two processes:
  - Identification = I have found you
  - Attribution = I have learnt something about you
- Can occur independently
  - Without the latter there is no disclosure

The Disclosure Risk Problem: Type I: Identification



Risk is present

- At the variable level
  - Power
    - Differentiation
    - Skew
  - Quality: susceptibility to divergence
  - Availability

Risk is present

- At the data unit level
  - Outliers
  - Vulnerable person’s

Risk is present

- At the attribute value level
  - Unusual characteristics
  - Sensitive attribute values

The Disclosure Risk Problem 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 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 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

Many other attack forms

- Table linkage
- Stream linkage
- Mash attacks
  - Cross dataset linkage enhancement
  - Match and search
- Repetitive queries
- Data hiding
- Data manipulation
- Response knowledge
- Response inference
- Etc etc.

Assessing Risk

- Risk modelling
- Penetration testing
- Data Environment Analysis
- Privacy Models
  - Differential Privacy
  - K-anonymisation

Controlling Risk

- Data Focused Controls
  - Statistical Disclosure Control
  - Differential Privacy
- Synthetic Data
- Environmental Controls

How much risk is negligible?

- Policy decision based on risk appetite
- Mature Risk management triangulates across the data situation:
  - Disclosiveness
  - Sensitivity
  - Environment
    - Agents
    - Data
    - Governance
    - Security

Concluding Remarks

- Anonymisation done correctly is a functional process
  - which turns personal data to non-personal data.
- There a variety of techniques and tools which can be used to asses the likelihood of an attack occurring and the likelihood of it succeeding once it has occurred.
- Functional anonymisation requires an evaluation of the totality of a data situation not just the data in question.
- Anonymisation will have a lifespan

Thank you!