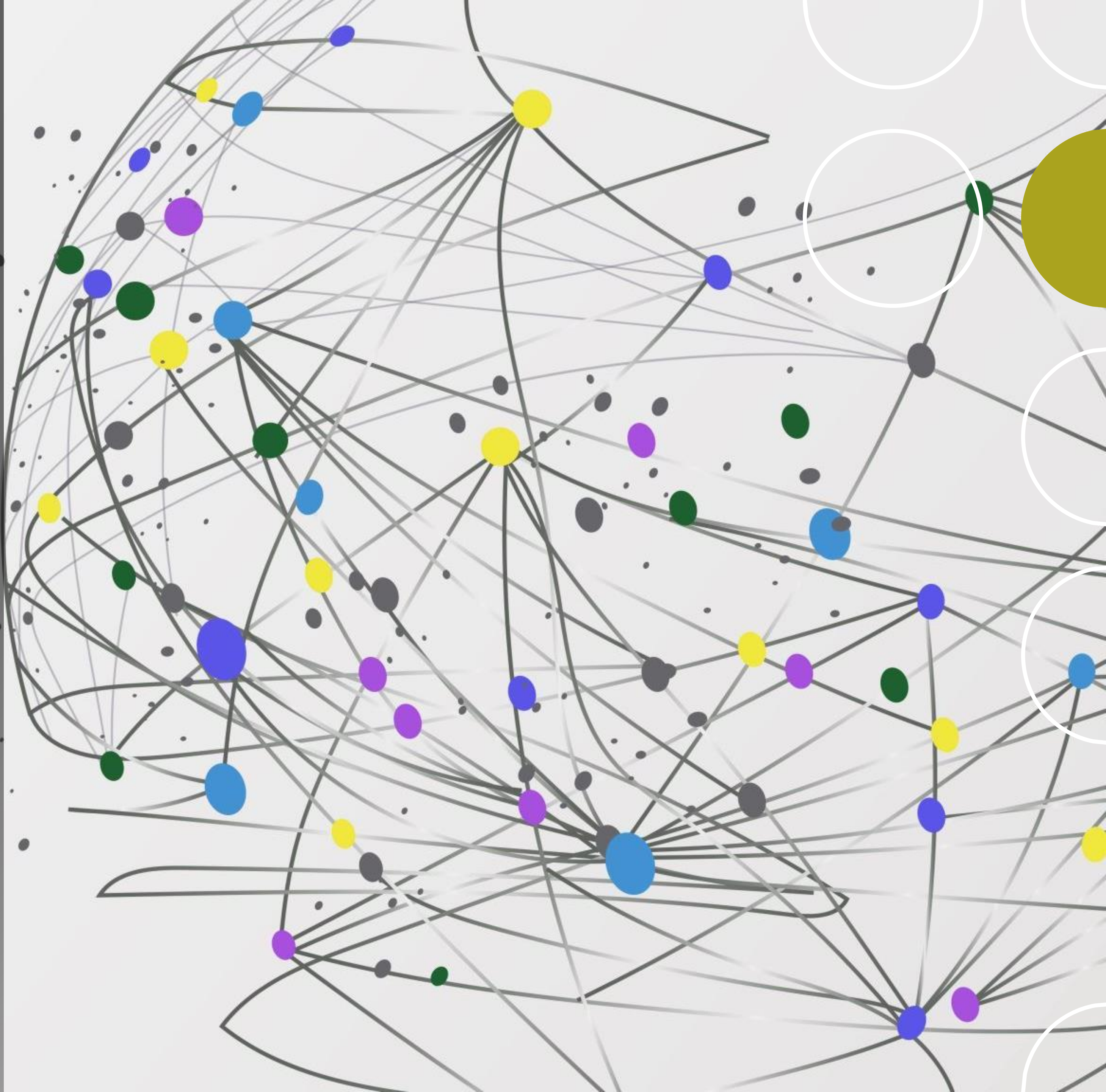
A Data Anonymisation Case Study

Dan Huntley, Divya Shridar, Nicole Cizauskas, Sreenidhi Venkatesh



Overview

1. The Problem
2. Data Anonymisation Considerations
3. The Solution
4. Methods
5. K-anonymity Calculation
6. Sharing of the Anonymized Data
7. Limitations
8. Conclusion and Looking Forward

The Problem

- Protecting the privacy of the customers
- Maximizing the useful information that can be given to the CEO and government research teams

The CEO	The Government	The Data
		
The CEO wants to: ➢ Use her customer's data ➢ Pass this data to research teams ➢ Investigate the travel habits of people with the Wanderlust gene ➢ Potentially increase the insurance policy for customers with the gene	The government wants to: ➢ Investigate people with the Wanderlust gene ➢ Check for educational or geographical similarities	The data includes: ➢ Personal info ➢ Geographical info ➢ Identification numbers ➢ Social habits ➢ Genomic info for the Wanderlust gene

Data Sharing Considerations

Benefits of data sharing:

Enable the community to confirm published results.

Avoids duplicating work

Reduces cost

Facilitates further analysis on the same dataset

Encourages collaborative work

Issues of data sharing:

Data privacy

- Confidentiality
- Ideas could be stolen
- Malicious misuse of data
- Accidental misuse of data

The Solution

Data anonymisation:

- The process of cleaning personal identifiers within a dataset that could potentially identify unwilling individuals

Removal of direct identifiers

- Taking out values in the data that could identify a specific individual

Pseudonymisation

- Replacing personal, identifiable data with artificial identifiers

Banding

- Classifying data into buckets with numeric ranges or representative categories

Aggregation

- Gathering data to express in a broader, summarised form

K-anonymity

- First described by Latanya Sweeney in 1998.
- It tells us the likelihood of individuals being identified from other individuals within the dataset via the combination of quasi-identifiers.
- Each record should be similar to at least k-1 other records based on the potentially identifiable variables (quasi-identifiers).

Direct -identifiers

- Can directly identify an individual
- Ex: name

Quasi-identifiers

- Can indirectly identify an individual through combination
- Ex: country of birth

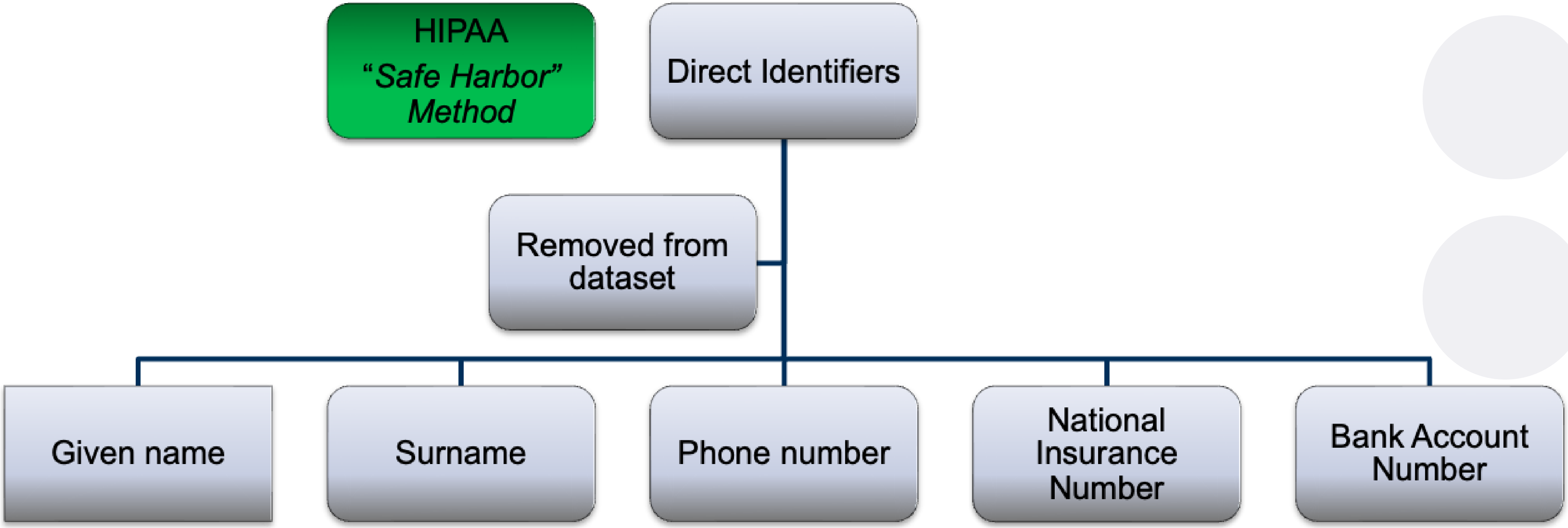
Other values

- Values that are not a direct identifier and are not able to be combined to identify individuals
- Ex: weight and height

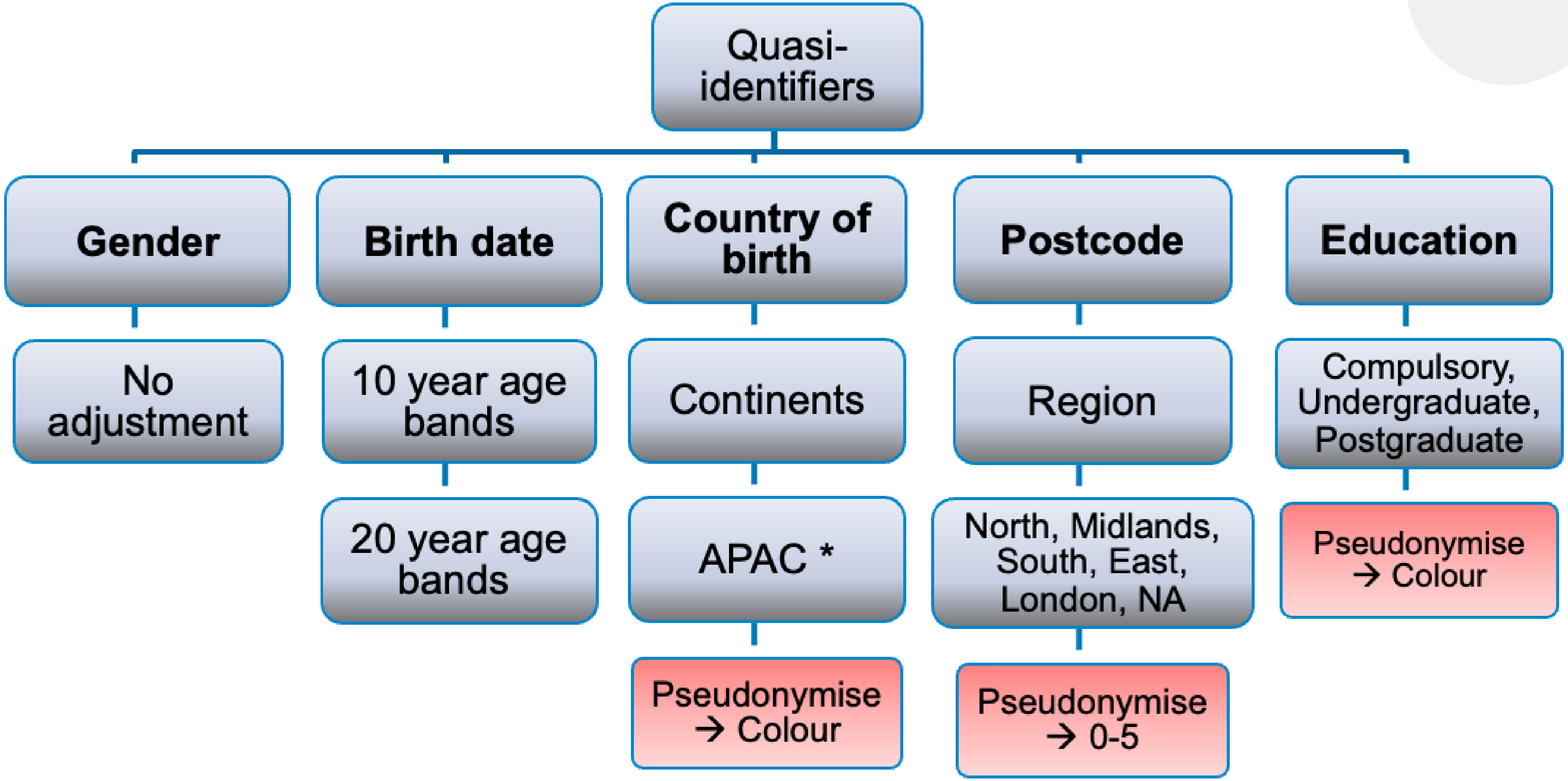
Methods

1. **Identification of direct identifiers**
   Removed from the dataset
2. **Identification of quasi-identifiers**
   Pseudonymisation to turn some quasi-identifiers into other value type
3. **Creation of cleaned dataframe**
   Fully anonymised data
4. **K-anonymity calculated**
5. **Adjustments made to quasi-identifiers**
   Until k-anonymity >1
   Trial and error
6. **Read only CSV file**
7. **Password protected zip folder**
   Optional keys for pseudonymised data

Direct Identifiers

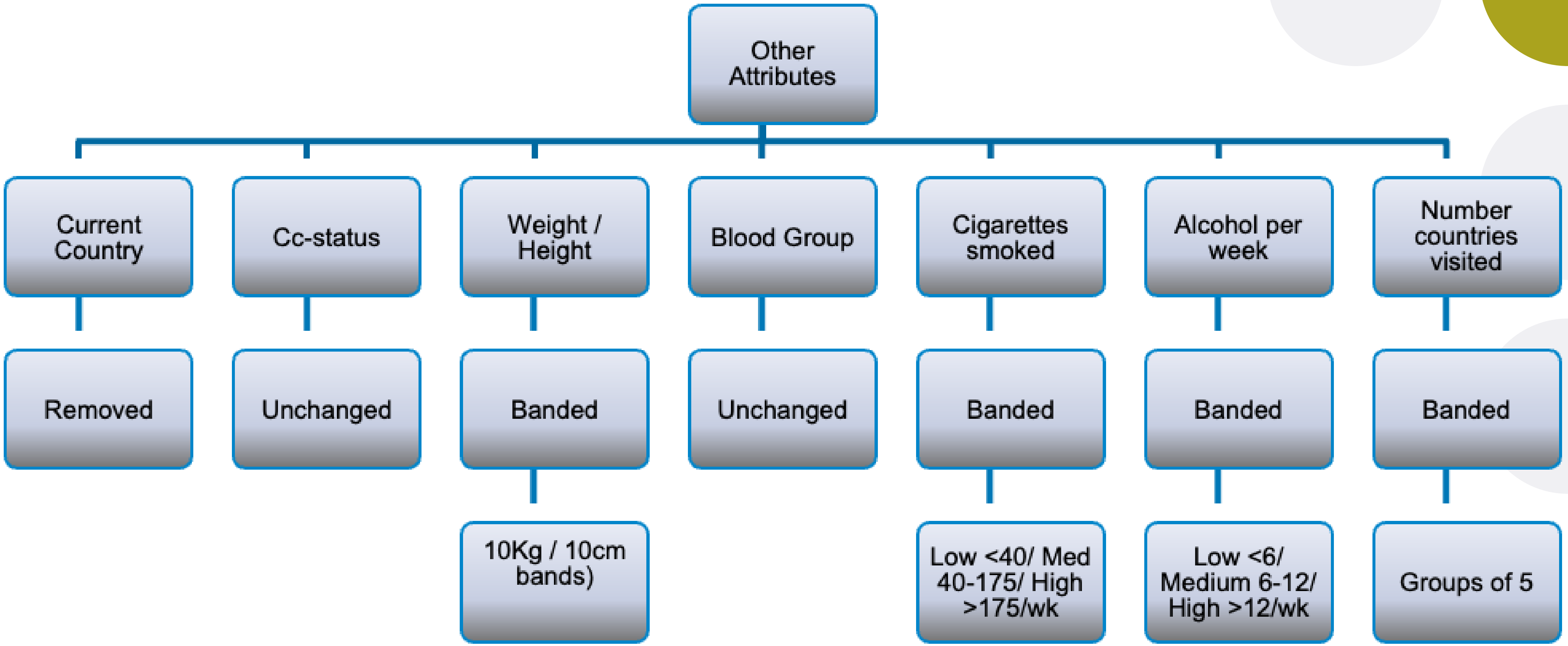


Quasi-identifiers



* APAC - Asia + Oceania + Antartica

Other Values



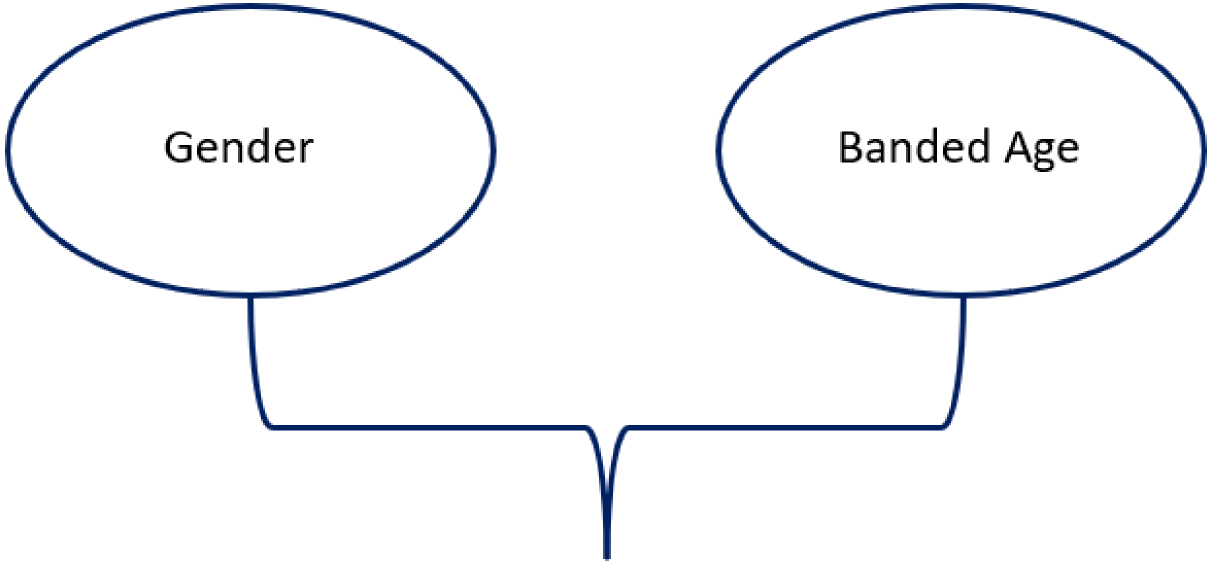
Final Cleaned Dataset

K-anonymity Calculation

- K-anonymity is calculated by finding the minimum matches of rows of quasi-identifiers

- Our two quasi-identifiers after cleaning the data were gender and banded age

- K-anonymity = 15



Each row has at least 14 others matched to it

K-anonymity = 15

Sharing Anonymised Data

Deliverable
For Imperial Researchers and Government

Anonymised_data
Final Folder with various components

Password Protected Zip Folder: data.zip
Raw-anonymised data in read-only csv format in a password protected folder

README File: README.MD
Relevant data documentation for future users

Password: wanderlust.txt
20-character alpha-numeric string password to access protected zipped folder with data

- **Used csv files** – industry standard to share non-complex data
- **Password protected zipped folders** instead of files – ensures files within are not corrupted
- **Zipped folders** – data is compressed and shareable
- **Read-Me file** – documentation for future users to access and use data

Limitations

- Potential over-aggregation of country of birth and postcode data
- Banding reduces specificity of research
- Certain circumstances when other information could be used to identify an individual – extreme outliers
- Pseudo-anonymisation – ratios of different groupings could be used to determine true values
- Still potential for misuse from researchers

Conclusion and Looking Forward

Challenges

- Lots of trial and error required to reach K > 1
- Difficult to intuitively determine what a quasi-identifier is
  - Especially with medical data
- Difficult to balance needs of CEO researchers and government
- Unsure of which information is valuable to include

Takeaways

- Hashing is best for data with lots of unique values
- Sorting data types (direct, quasi, other) first helps
- Include as much info as you can

References

- Devane, H. (2022) *Everything you need to know about K-anonymity*, *Immuta*. Available at: https://www.immuta.com/blog/k-anonymity-everything-you-need-to-know-2021-guide/ (Accessed: December 14, 2022).
- *Data Anonymization* (2022) *Corporate Finance Institute*. Available at: https://corporatefinanceinstitute.com/resources/business-intelligence/data-anonymization/ (Accessed: December 14, 2022).
- Ucl (2019) *Anonymisation and Pseudonymisation*, *Data Protection*. Available at: https://www.ucl.ac.uk/data-protection/guidance-staff-students-and-researchers/practical-data-protection-guidance-notices/anonymisation-and (Accessed: December 14, 2022).
- Sweeney L. K-anonymity: A Model For Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems.* 2002; 10(5): 557-570. https://doi.org/10.1142/S0218488502001648
- El Emam K, Dankar FK. Protecting privacy using k-anonymity. J Am Med Inform Assoc. 2008 Sep-Oct;15(5):627-37. doi: 10.1197/jamia.M2716. Epub 2008 Jun 25. PMID: 18579830; PMCID: PMC2528029. doi: 10.1197/jamia.M2716
- US Department of Health and Human Services. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Available at: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#safeharborguidance (Accessed: December 14th 2022)