Data De-Identification Support System

Sudha Harikrishnan, SingHealth, HSRC
Lam Shao Wei, SingHealth, HSRC

On Behalf of the SingHealth Data Deidentification Workgroup

Background
With the implementation of Human Biomedical Research Act (HBRA) (2017), the researchers who conduct HBR were to de-identify the data if the patient consent for the study is not obtained or the informed consent doesn’t meet the relevant requirement specification. Thus, the SingHealth HSRC (Health Science Research Center) was tasked to develop a decision support system to de-identify the direct identifiers and the residual risks associated with the data in compliance with HBRA and existing policies and processes in a healthcare industry.

Aim
To provide the seamless, one-stop platform for the end user, in order to complete the data de-identification process with the maximum automation.

Methodology
A synthetic dataset that is representative of medical data was used for this project. Electronic medical data mostly contains personal identifiers, quasi identifiers and sensitive data. The confidentiality and sensitivity of the data is highly correlated with the data type. Based on the data type, de-identification techniques like cryptographic hashing, masking, generalization, suppression, K-anonymity, etc. were applied to render the data non-identifiable before it is realized to researchers and collaborators.

Conclusion
In this project, we conceptualize, decide and implement a data de-identification system for Singhealth. The system aims to provide a seamless solution to current data requesting process. With the goal in mind, we develop a Python-based solution for data de-identification practice, which could achieve:
1) Field type and method auto suggestions
2) Risk Assessment using K Anonymity
3) Grid Search for partial optimal solution
4) Report Generating

We build our UI under web-based flask framework, to serve a better flexibility for our client. The solution solves most of the current pain point in the data requesting process, and could help the healthcare to both protect users privacy and enhance data-driven research capacity.

Acknowledgement: This project was supported by the NUS School of Computing. Special thanks to He YingXu, Sheng Yu, Xiao ZuoLing, Yu ZongDong.