Realism in Synthetic Data Generation

A thesis presented in fulfilment of the requirements for the degree of:

Master of Philosophy
in
Science

Scott McLachlan
(MCSE, MCT, DipSysEng, GradDipInfSc, GradDipLaw, GradDipBus, MIITP, MBCS)
School of Engineering and Advanced Technology
Massey University
Palmerston North, New Zealand

Supervised by:

Dr. Kudakwashe Dube
School of Engineering and Advanced Technology
Massey University
Palmerston North, New Zealand

Prof. Thomas Gallagher
Applied Computing and Engineering Technology
Missoula College
University of Montana
Missoula, USA

2017

Abstract

There are many situations where researchers cannot make use of real data because either the data does not exist in the required format or privacy and confidentiality concerns prevent release of the data. The work presented in this thesis has been undertaken in the context of security and privacy for the Electronic Healthcare Record (EHR). In these situations, synthetic data generation (SDG) methods are sought to create a replacement for real data. In order to be a proper replacement, that synthetic data must be *realistic* yet no method currently exists to develop and validate realism in a unified way. This thesis investigates the problem of characterising, achieving and validating realism in synthetic data generation. A comprehensive domain analysis provides the basis for new characterisation and classification methods for synthetic data, as well as a previously undescribed but consistently applied generic SDG approach. In order to achieve realism, an existing knowledge discovery in databases approach is extended to discover realistic elements inherent to real data. This approach is validated through a case study. The case study demonstrates the realism characterisation and validation approaches as well as establishes whether or not the synthetic data is a realistic replacement. This thesis presents the ATEN framework which incorporates three primary contributions: (1) the THOTH approach to SDG; (2) the RA approach to characterise the elements and qualities of realism for use in SDG, and finally; (3) the HORUS approach for validating realism in synthetic data. The ATEN framework presented is significant in that it allows researchers to substantiate claims of success and realism in their synthetic data generation projects. The THOTH approach is significant in providing a new structured way for engaging in SDG. The RA approach is significant in enabling a researcher to discover and specify realism characteristics that must be achieved synthetically. The HORUS approach is significant in providing a new practical and systematic validation method for substantiating and justifying claims of success and realism in SDG works. Future efforts will focus on further validation of the ATEN framework through a controlled multi-stream synthetic data generation process.

Publications related to this thesis:

McLachlan, S., Dube, K., & Gallagher, T. (2017). Managing Realism in Synthetic Data Generation. *Manuscript submitted to JAMIA.*

McLachlan, S., Dube, K., & Gallagher, T. (2017). THOTH: The generic approach to and characterisation of Synthetic Data. *Manuscript submitted to JAMIA.*

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S. (2017). Synthea: An approach, method and software mechanism for generating synthetic patients and the synthetic electronic healthcare record. *Manuscript submitted to JAMIA*.

McLachlan, S., Dube, K., & Gallagher, T. (2017). The Realistic Synthetic Electronic Health Record: Challenges, rationale and future directions. *Manuscript submitted to JAMIA*.

McLachlan, S., Dube, K., & Gallagher, T. (2016). Using the CareMap with health incidence statistics for generating the realistic synthetic electronic health record. *IEEE International Conference on Healthcare Informatics, ICHI'16*.

Glossary

ATEN	The ATEN framework is an SDG lifecycle incorporating the THOTH, RA and HORUS approaches.
AU DoH	Australian Department of Health
CPG	Clinical Practice Guideline
HiKER Group	Heath Informatics and Knowledge Engineering Research Group
HIS	Health Incidence Statistics
HORUS	Uses the knowledge developed by RA as the basis for validating realism in synthetic data and justifying success in SDG.
NZ MoH	New Zealand Ministry of Health
RA	A systematic approach used to discover realistic elements, characteristics and rules necessary to the creation of realistic synthetic data.
PK	Primary Key
SDG	Synthetic Data Generation
THOTH	The generic approach for SDG

Dedicated for Danika, Thomas, Liam and James.

Acknowledgements

I acknowledge with the greatest of appreciation the assistance of my supervisors and the wider members of the Health Informatics and Knowledge Engineering Research (HiKER) Group who supported my development as a researcher in the tradition of the scientific method. The support of my proof reader and sometimes editor who every day pointed out when my references were out of order and when what I had written didn't actually say what I thought I had said.

And I can't leave out Master 4, who recognised that I focus and work better when I have multiple streams of input and things to think about. Using this as only a four-year old can; as justification for continually distracting me with games, puzzles, stories and an insatiable need for me to join him as he played with his vast collection of toy trains. My hope is that I live to see the day when my encouragement of you culminates in my receiving a copy of your own thesis. I especially look forward to discussions about the distractions you had to deal with.

There are scores of others with whom I have interacted during the eight months spent researching and writing this thesis. But for the fact that it would take vast amounts of time and far more space than I am given on this page to single you all out, I offer my best wishes and thanks.

Scott

February, 2017.
Sydney, Australia.

To the reader;

The fact that you have chosen to pick up or download this thesis is an act that in and of itself deserves thanks. If nothing else, and in deference to the content, this single act justifies this thesis' existence.

Thank you.

This thesis is also a tribute to the late bloomers. People like Nikola Tesla, Charles Darwin, Samuel Jackson and Richard Adams. To all those who didn't even begin to realise their vast potential until later in life.

Table of Contents

Tables

Figures

"*Behind every algorithm there is always a person. A person with a set of personal beliefs that no code can ever completely eradicate. You must identify your own personal bias. You need to understand that you are human and take responsibility accordingly*."

(Ekstrom, 2015)