Juryrapport
Philips Afstudeerprijs Data Science and Artificial Intelligence in Health Care 2019

J.T. (Jan) Trienes MSc, University of Twente
Comparing Rule-based, Feature-based and Deep Neural Methods for De-identification of Dutch Medical Records

The Royal Holland Society of Sciences and Humanities awards Jan Trienes the Philips Graduation Award for Data Science and Artificial Intelligence in Health Care 2019 for his master thesis completed at the Faculty of Electrical Engineering, Mathematics and Computer Science, Department of Computer Science, Specialization Data Science of the University of Twente.

The thesis of Jan Trienes addresses the important and currently relevant data-science problem of making available data about individuals without compromising their privacy. This is especially relevant in health where one would like to provide access to medical records to learn effective diagnoses and therapies from their history. To do so from digital patient records enables learning from rare diseases or personalized conditions, unheard of when using paper records.
For digital records, it is essential to protect the privacy of patients better than before. The thesis aims to anonymize Dutch medical records by removing all data that may reveal the identity of the patient: removing a *name* and *address* is obvious, but also *age* and even *weight* may carry uniquely identifiable information when the disease is rare. The anonymization procedure is generally referred to as *de-identification*.
Trienes has collected and organized the medical records of 1260 patients, originating from nine health care institutes in three different domains of Dutch health care. Using this data, he has compared AI-methods to detect privacy-sensitive data as their detectability determines the degree to which automatic de-identification is feasible. The thesis focusses on Dutch records and compares the results to those obtained from English records.
The jury is impressed by the efficient organization of such large amounts of data. As a consequence, the thesis reveals that an existing rule-based de-identification method for the Dutch language does not work very well on the newly collected data, where deep learning methods perform much better on Dutch and English records alike, even when data is scarce. Fully automatic de-identification on new data still requires manual intervention by practitioners, but the thesis of Trienes is a significant step forward to fully automatic de-identification of privacy-sensitive data. The jury acknowledges the important contribution to the domains of digital data science in health.

Prof. dr. E.O. (Eric) Postma, hoogleraar kunstmatige intelligentie Tilburg University
Prof. dr. ir. A.W.M. (Arnold) Smeulders, hoogleraar computer vision Universiteit van Amsterdam, voorzitter directie COMMIT

De jury vergaderde op 18 oktober 2019 onder leiding van Ir. B.M.Th. Dortland-Bier, directeur KHMW. Daarnaast waren ter vergadering aanwezig Prof. dr. A.P. IJzerman, secretaris natuurwetenschappen en Drs. S. van Manen, secretaris.