PHT-meDIC - Federated privacy-preserving machine learning on medical data

Prof. Dr. Oliver Kohlbacher
Dept. of Computer Science, University of Tübingen
Center for Bioinformatics and Medical Informatics, University of Tübingen
Institute for Translational Bioinformatics, University Hospital Tübingen

Person-related medical data has a high need for privacy – patients and citizens need to know that their data will not be abused. At the same time, an overwhelming majority of patients wants their data to be used for research. While the European General Data Protection Regulation (GDPR) protects a data subject's right to privacy, it does grant certain privileges for research. Nevertheless, the frustrating reality is that data protection concerns and the effort required to ensure GDPR compliance often prevent the (re-)use of medical data for research across hospitals, state borders, or national borders.

Leaving the data where it is (safely) stored and bringing the análisis algorithms to the data ('code to data') is a paradigm that is frequently applied to solve data protection hurdles preventing the central pooling of distributed data. In particular for machine learning approaches (ML) that require large data sets, this approach - often termed federated or distributed ML - has to potential to overcome some of the legal hurdles by technical means.

The Personal Health Train (PHT) is a paradigm that has been suggested as part of the GO-FAIR initiative. The core idea is that analysis algorithms ('trains' in this metaphor) pass from hospital to hospital ('train stations') and can incrementally learn on each data point. The data remains at each site, hence the privacy of the patients is ensured. With the PHT-meDIC platform we present a software platform for federated analytics with a focus on clinical, omics, and imaging data. Lightweight container-based virtualization enables the execution of complex software pipelines without additional software installation. Through the adoption of the HL7/FHIR standard used in the medical data integration centers (meDIC) at all German university hospitals, the platform can achieve a high degree of interoperability. We will discuss the current status of the platform, its limitations as well as the plans for extending the platform in the future.