Final report: High-Performance Data Mining for Drug Effect Detection

- a project funded by the Swedish Foundation for Strategic Research during 2012-2017 under grant IIS11-0053

Principal investigator: Prof. Henrik Boström, bostromh@kth.se, now at KTH Royal Institute of Technology, School of Electrical Engineering and Computer Science, and at Stockholm University, Dept. of Computer and Systems Sciences, during the project

Summary

The main goal of the project *High-Performance Data Mining for Drug Effect Detection* was to develop techniques and software tools for large-scale analysis of massive, heterogeneous and continuously growing data sets to support decision-making for drug development and patient monitoring by analyzing drug effects in patient records, individual case safety reports, drug registries and experimental data. Motivated by the specific requirements of the drug development and patient monitoring tasks, the scientific goal of the project was to extend the capabilities and applicability of data and text mining techniques and algorithms, something which was expected to be of importance also for many other areas of application. A long-term goal of the project was to produce qualified researchers (PhDs) in the area of data and text mining, with a focus on applications within drug development and health-care. The objectives of the project were clustered in four main work packages: adverse drug event (ADE) detection (WP1), with a focus on the development of predictive models for suggesting missing diagnosis codes in electronic health records (EHRs) and on techniques and tools for discovery of previously unknown ADE candidates; clinical text mining (WP2), with a focus on exploitation of clinical notes for enhancing the ADE detection and hypothesis generation tasks; conformal prediction (WP3), with a focus on methods for providing confidence-based predictions; and parallel data mining (WP4), with a focus on efficient implementation of learning algorithms, exploiting current developments in multi-core and GPU computing. The project has funded five senior researchers on 10-20% and six PhD students on 40-80%. In addition, three researchers and one PhD student have been contributing to the project, e.g., as co-supervisors, with funding external to the project. An approval from the regional ethical board in Stockholm was obtained to allow for analyzing electronic patient records within the project, and access was provided to anonymized records from more than one million patients. The project has been collaborating with clinical pharmacologists and healthcare providers, as well as with researchers in the pharmaceutical industry, which have resulted in joint publications and efforts to explore and exploit the potential in the available large-scale, heterogeneous datasets as well as novel projects. The outcome of the project comprises 83 scientific publications; 21 journal papers, 45 conference papers, 12 workshop papers, 4 PhD theses, and 1 licentiate thesis. In addition, a large number of software packages have been developed that have been made publicly available.

1. Background, objectives and organization of the project

1.1 Background, motivation and long term vision

The main goal of the project was to develop techniques and software tools for large-scale analysis of massive, heterogeneous and continuously growing data sets to support decision-making for drug development and patient monitoring by analyzing drug effects in patient records, individual case safety reports, drug registries and experimental data. The long-term goal of developing these techniques and tools was to provide researchers within the pharmaceutical industry and academia with support for developing predictive and descriptive models of drug effects, and to provide practitioners within the healthcare sector with support for detecting drug effects on individual patients. Successful use of such tools is expected to result in new findings that could support the development of new effective drugs with fewer or less severe side-effects and also for avoiding, or earlier detection of, adverse drug effects during treatment. Information on drug effects on large numbers of observations is expected to ultimately lead to more effective and safer drugs, which are less expensive to develop, since development costs can be significantly reduced by employing predictive models that early-on can signal for e.g., adverse events. Furthermore, health-care personnel can get improved decision-support from systems that indicate risk for, or observations of, adverse drug events. In addition to improving the quality of the provided health-care, such monitoring systems would increase the quality and quantity of individual case safety reports submitted to the Medical Products Agency and the International Drug Monitoring center.

Motivated by the specific requirements of the drug development and patient monitoring tasks, the scientific goal of the project was to extend the capabilities and applicability of data and text mining techniques and algorithms, something which was expected to be of importance also for many other potential areas of application. A long-term goal of the project was to produce qualified researchers (PhDs) in the area of data and text mining, with a focus on applications within drug development and health-care. Such researchers are expected to enable knowledge-transfer from their areas of expertise to practitioners in the pharmaceutical industry and the health-care sector, in particular by leading the development of novel solutions to support decision-making by analyzing massive amounts of continuously growing data.

1.2 Goals and objectives

A main, concrete goal of the project was to communicate findings in scientific venues within the areas of data and text mining, health informatics and medicinal chemistry. Another main, concrete goal was to develop software for supporting discovery and decision-making by analyzing massive patient records, drug prescriptions, individual case safety reports and chemical compound data.

The objectives of the project were clustered in four main work packages:

WP1. Adverse drug event detection
Adverse drug event (ADE) detection with longitudinal clinical data, such as electronic health records (EHRs), is a crucial component in post-marketing drug safety surveillance. EHRs contain massive and heterogeneous data including both structured data and clinical notes in free text. Since the EHRs are not designed for performing analyses, some of their special characteristics, e.g., the large volume, high dimensionality, sparsity and time dependencies, require the tools and techniques to be able to tackle such data properly. The objectives of this work package have been to:

- enable analysis of structured EHR data to identify, classify and describe potential ADEs
- combine structured and unstructured EHR data to improve the ADE detection performance
- develop techniques and tools for effective and efficient handling of big data with high dimensional and sparse features, unbalanced class distributions, missing values and temporal dependencies.

WP2. Clinical text mining

Perceived drug effects are often reported in natural language, e.g., in the free-text notes of health records. In order to exploit textual data sources for automated drug effect detection, methods are needed that are able to analyze massive amounts of, typically noisy and incomplete, text. To this end, the main objectives of this work package were to:

- enable discovery of relationships between drugs and disorders in clinical notes
- provide terminological and annotated resources for ADE detection
- facilitate learning of effective predictive models from natural language data.

WP3. Conformal prediction

Computational chemists in the pharmaceutical industry as well as healthcare personnel have a demand for confidence-based predictions, i.e., that the error rate can be guaranteed to be below a specified level. Conformal prediction is a new framework that enables predicting with confidence. The main objectives of this work package were to:

- extend and improve the conformal prediction framework and adapt it for state-of-the-art learning algorithms
- develop tools to enable prediction with confidence.

WP4. Parallel data mining

When analyzing large-scale and continuously growing data, it is of high importance that the algorithms scale with the growing problem sizes. To ensure that the algorithms used are able to handle the massive data sizes managed in the project, the main objectives of this work package are to:

- enable efficient, scalable and parallel solutions for training and evaluation of data and text mining algorithms on big data
- utilize cutting-edge parallel platforms such as GPUs and multi-core CPUs to achieve maximum efficiency.

1.3 The basic organization; leadership, research environment, relation to other grants etc.

The project has been organized into four main work packages, each involving at least one responsible senior researcher and one PhD student. Two of the work packages have been coordinated from Stockholm University (SU) and two from University of Borås (UB).

The researchers at SU contribute to the research area data science, which concerns methods, techniques and tools for organizing and analyzing data in order to support decision making. A particular focus of the research within data science at SU is on ensemble methods, i.e., techniques for generating sets of models that collectively form predictions by voting, and on methods for generating interpretable models, e.g., rule learning. Another focus is on text mining, in particular efficient and resource lean methods using language technology for very large text sets. The research also focuses on semantic analysis, e.g., negation, speculation and temporality, in order to be able to extract situation-specific, accurate and relevant information from texts.

One main application area for the research at SU is healthcare analytics, which aims for providing efficient and effective decision support for healthcare and pharmaceutical research. The research group has collaborated for several years with computational chemists in the pharmaceutical industry. This has resulted in new techniques and tools for building predictive models from observed biological activities, e.g., toxicity, of chemical compounds, which are currently being used in the industry. The experience and contacts developed during the project *High-Performance Data Mining for Drug Effect Detection* has contributed to enabling of two additional projects on data mining using health records. The first was a one year pilot study during 2016 which later lead to a three year project 2017-2019 (Coril), both aimed at studying and understanding the treatment of heart failure. Both projects are co-funded by Stockholm University and Stockholm County Council. The main focus of these projects is to use registry data from a regional healthcare data warehouse (Vårdanalysdatabasen, VAL) to analyse and better understand the treatment of heart failure patients. Members of the group also participate in Nordic Center of Excellence in Health-Related e-Sciences (NIASC), which is lead by Karolinska Institutet and funded by Nordforsk during 2014-2018.

Another application area that the group is involved in is modeling of component wear in heavy trucks using data mining and to provide decision support for optimizing heavy truck fleet utilization. Members of the group participate in the project Integrated Dynamic Prognostic Maintenance Support (IRIS), which is lead by Scania AB, and supported by Swedish Governmental Agency for Innovation Systems (VINNOVA) during 2012-2017.

The research group CSL@BS at UB covers a wide range of topics in the field of computer science, but the main focus are data analytics and high performance computing. More technically, the research is conducted in the areas of parallel and distributed computing, data mining and machine learning. Within the fields of data mining and machine learning, the aim is to develop and improve machine learning methods, techniques and algorithms for use in data analytics. The group conducts both theoretical and applied research, often in close cooperation with industry, and a common theme for all research is algorithm development and implementation. With a pragmatic approach, the research group develops general as well as application or environment centric solutions, where the involvement spans the whole process from algorithm design, analysis, proof and experimentation to deployable implementations and frameworks. Although developed solutions tend to be general-purpose rather than application specific, solutions are continuously evaluated on real-world problems in collaboration with industry.

CSL@BS has during the project *High-Performance Data Mining for Drug Effect Detection* run two research projects funded by the Knowledge foundation: *Big Data Analytics by Online Ensemble Learning* (BOEL) 2013-2015 and *Data Analytics for Research and Development* (DASTARD) 2016-2018. Both these projects focus on developing novel machine learning methods and algorithms for data analytics. BOEL and DASTARD are complementary to the *High-Performance Data Mining for Drug Effect Detection* project, since similar techniques are investigated, but in other domains and for other applications. Specifically, several findings from the BOEL and DASTARD projects about conformal prediction have been adapted and/or further developed in *High-Performance Data Mining for Drug Effect Detection*.

1.4 The changes made to the project during its period, with reference to the original objectives

A major change compared to the original objectives was the introduction of predictions with confidence, or conformal prediction, as a research focus. The research on conformal prediction was actually requested by researchers at AstraZeneca, who need to be able to provide guarantees along with the predictions of the generated models. Furthermore, it soon became evident that confidence in predictions is also a vital aspect to consider when working with EHR data and with predictions of diagnoses.

2. The research of the project

2.1 The scientific approach and the results compared to the scientific objectives

WP: Adverse drug event detection

This work package has developed and evaluated methods for facilitating the secondary use of electronic health records (EHRs) by using machine learning to build predictive models. The research focused on the technical challenges involved in learning predictive models from EHRs and addressed these by investigating issues primarily related to data representation. While EHRs provide large amounts of valuable clinical data, the complexity of the data poses a multitude of challenges for large-scale analysis, including the existence of heterogeneous types of data, the amount of attributes, the sparseness of measurements, the concept hierarchies used for encoding drugs and diagnoses, as well as the inherent temporality of clinical events.

Methods have been developed to facilitate the identification of features from the richly structured and complex EHR data, which is not only of high dimensionality and sparsity but also heterogeneous and embedded with irregular temporality in terms of length and rhythms. The application of these methods to EHR data pushes forward the meaningful secondary use of EHRs by unlocking more possibilities for analyzing and modeling EHR data. At the more practical end, with the help of the predictive models learned as part of the project, under-reporting of ADEs can potentially be mitigated. The methods that have been developed as part of this project can be applied in clinical decision support systems and assist doctors in assigning ADE-specific diagnosis codes.

Representing heterogeneous types of data

Due to the unique characteristics of each type of data in EHRs, the representation of data from each type is also specific [Karlsson I., Zhao J., Boström H. and Asker L., 2013]. For structured EHR data, representing drugs and diagnoses can benefit from exploiting the encoding systems [Zhao J., Henriksson A. and Boström H., 2014], and representing clinical measurements and laboratory tests requires the consideration of missing and repeated events [Zhao J., Henriksson A., Asker L. and Boström H., 2014]. Moreover, in specific cases it is beneficial to combine multiple types of data [Zhao J., Henriksson A., Asker L. and Boström H., 2015]. In other cases, it is important to handle the sparsity and high dimensionality of the data [Karlsson, I. and Zhao J., 2014; Karlsson I. and Boström H., 2014].

Leveraging concept hierarchies within medical data

The impact of using the concept hierarchies of clinical codes ICD-10 for diagnoses and ATC for drugs has been studied from two perspectives: enriching the feature space and decomposing the predictive task. On the one hand, the concept hierarchy of the two coding systems is shown to be useful for constructing a feature space containing clinical codes from different hierarchical levels to improve the predictive performance [Zhao J., Henriksson A. and Boström H., 2014]. On the other hand, the concept hierarchy of ICD-10 codes is shown to be beneficial for decomposing the predictive task into a number of sequential steps, serving different application scenarios such as predicting a disease family or a specific disease [Zhao J., Henriksson A. and Boström H., 2015].

Incorporating temporal information

Temporal information is shown to be valuable for learning effective predictive models. Various ways of incorporating temporality have been investigated in this project by creating a single-point representation or a heterogeneous and multivariate time series representation. In the single-point representation, by assigning temporal weights to clinical events prior to learning the predictive model, the predictive performance is improved compared to ignoring temporality or treating events from different time windows as different features [Zhao J., Henriksson A., Kvist M., Asker L. and Boström H., 2015; Zhao J., 2015;]. Moreover, the temporal weights can also be learned through the random forest algorithm [Zhao J. and Henriksson A., 2016]. In the multivariate time series representation, each clinical event is treated as a time series, which leads to an advanced representation consisting of multivariate time series that take into account both the order of clinical events and their corresponding values [Zhao J., Papapetrou P., Asker L. and Boström H., 2017]. Methods are developed to build predictive models using such a representation, which are useful not only in representing longitudinal EHR [Karlsson I. and Boström H., 2016] data but also for general time series classification problems [Karlsson I., Boström H. and Papapetrou P., 2016].

WP: Clinical text mining

This work package has focused on extracting information and learning predictive models from clinical text. There are numerous challenges involved in analyzing unstructured, high-dimensional and sparse EHR data in the form of clinical text, most of which concern the high degree of syntactic and semantic ambiguity caused by domain-specific vocabularies, frequent misspellings, ad-hoc shorthand and non-adherence to conventional grammar. The methods proposed in the work package address many of these challenges and have focused on unsupervised and semi-supervised learning approaches in order to avoid the bottleneck caused by relying on access to large amounts of labeled data, which is particularly expensive to create in the medical domain.

Ensembles of semantic spaces

Distributional semantics allow models of linguistic meaning to be derived from large-scale observations of language use in a wholly unsupervised manner. Continuous vector representations of words - often referred to as word embeddings - inhabiting a so-called semantic space have proven valuable in numerous downstream natural language processing (NLP) tasks. In this work package, it has been shown that the semantics of a word - as well as other sequential data types - can be better captured in an ensemble of semantic spaces. To that end, various strategies for creating the constituent semantic spaces - by manipulating the underlying data and certain model hyper-parameters - as well as for combining them, have been explored. The semantic space ensembles are used both directly for *k*-nearest neighbors (*k*NN) retrieval and for semi-supervised machine learning [Henriksson A., 2015c].

Handling polysemy of medical terms

Many approaches to the analysis of clinical text rely on terminological resources for handling polysemy of medical terms, in the form of both synonyms and shorthand, i.e. abbreviations and acronyms. Manual construction of wide-coverage terminological resources are both difficult and costly to create and maintain over time. In this work package, it has been shown that distributional semantics can be leveraged for automatically extracting synonym candidates of medical terms from large corpora [Henriksson, A., Conway, M., Duneld, M., Chapman, W.W., 2013]. This approach has also been proposed as a means to help expand existing medical terminologies, such as SNOMED CT, for which synonyms are absent in Swedish [Henriksson, A., Skeppstedt, M., Kvist, M., Duneld, M., Conway, M., 2013]. Improvements can moreover be gained by combining a set of *k*NN rankings for representations obtained using different strategies for handling word order, [Henriksson, A., Moen, H., Skeppstedt, M., Eklund, A-M., Daudaravicius, V. and Hassel, M., 2012], window sizes and types of corpora [Henriksson, A., Moen, H., Skeppstedt, M., Daudaravicius, V., Duneld, M., 2014]. A method for extracting terms that belong to a specific semantic category, e.g. clinical findings or drugs, has also been proposed [Skeppstedt, M., Ahltorp, M., Henriksson, A., 2013], with improvements obtained through category-specific pre-processing and clustering [Ahltorp, M., Skeppstedt, M., Kitajima, S., Henriksson, A., Rzepka, R., Araki, K., 2016]. Methods based on distributional semantics have also been proposed and evaluated for automatically expanding abbreviations [Tengstrand, L., Megyesi, B., Henriksson, A., Duneld, M., Kvist, M., 2014], as well as for expanding dictionaries of marker words for uncertainty and negation [Alfalahi, A., Skeppstedt, M., Ahlbom, R., Baskalayci, R., Henriksson, A., Asker, L., Paradis, C., Kerren, A., 2015].

Learning semantic prototypes for named entity recognition

Named entity recognition is a critical component of many NLP systems and involves identifying references in free-text to specific semantic categories. This task is typically approached as a sequence to sequence learning task, relying on annotated examples of named entities. In this work package, it has been shown how named entity recognition in clinical text can be improved by learning prototypical representations of each semantic category from a small set of examples in conjunction with a semantic space obtained using a model of distributional semantics on a large, unannotated corpus [Henriksson, A., Dalianis, H., Kowalski, S., 2014.]. Further improvements were obtained by exploiting the notion of semantic space ensembles, i.e. creating and combining multiple prototypes of each semantic category. This method has been evaluated for de-identification of clinical text, i.e., by recognizing and obscuring sensitive information [Henriksson, A., 2015a] - as well as for identifying information pertaining to adverse drug events, e.g. references to drugs and symptoms [Henriksson, A., Kvist, M., Dalianis, H., Duneld, M., 2015].

Detecting adverse drug events with distributed representations

Distributed representations of clinical notes have also been shown to be effective for assigning diagnosis codes, both generally [Henriksson, A. and Hassel, M., 2013] and specifically for detecting adverse drug events in EHRs [Henriksson, A., 2015b], outperforming the commonly used but shallow bag-of-words representation. In this work package, the distributional semantics framework was moreover extended to other types of sequential data, allowing effective representations to be created for heterogeneous data types in EHRs. These dense, reduced-dimensional representations outperform their sparse and high-dimensional counterparts and also facilitate the combination of structured and unstructured EHR data for ADE detection [Henriksson, A., Zhao, J., Boström, H., Dalianis, H., 2015a]. The notion of semantic space ensembles is moreover shown also to be effective in a classification setting, favoring early (feature) fusion over late (classifier) fusion [Henriksson, A., Zhao, J., Boström, H., Dalianis, H., 2015b]. Different ways of utilizing a set of semantic spaces built over heterogeneous data are investigated, demonstrating how it can be exploited for creating effective ensembles of randomized trees by randomly sampling distributed representation of clinical events in the tree-building procedure [Henriksson, A., Zhao, J., Dalianis, H., Boström, H., 2016]. Distributed representations are also investigated for detecting information pertaining to adverse drug events in clinical notes, such as drugs and symptoms, as well as relations that may hold between such mentions, for instance whether it expresses an indication for prescribing a drug or a side-effect of a drug [Henriksson, A., Kvist, M., Dalianis, H., Duneld, M., 2015].

WP: Conformal prediction

When creating confidence-based predictors using conformal prediction, there are several open questions. One of the questions addressed relates to how data can be utilized effectively to achieve more efficient confidence based predictions using ensembles? It was shown in an initial study that the use of out-of-bag estimates when using bagging ensembles results in more effective conformal predictors [T. Löfström, U. Johansson and H. Boström, 2013]. This was later followed up in additional studies, where it was shown that the calculation of the employed normalization procedure using kNN can be significantly sped up by instead using a variance-based metric [H. Boström, H. Linusson, T. Löfström and U. Johansson. 2016] and that the use of traditional out-of-bag estimates violates the exchangeability requirement and a slightly altered procedure allowing the use of out-of-bag estimates, which does not violate the requirements, was proposed [H. Boström, U. Johansson, H. Linusson, and T. Löfström, 2017]. Another question that were addressed relates to how problems with class imbalance affect the confidence based predictions when using conformal prediction. It was shown in an initial study that a conformal predictor conditioned on the class labels to avoid a strong bias towards the majority class is more effective on problems with class imbalance [Löfström, T., Boström, H., Linusson, H., Johansson, U. 2015]. In a second study, using EHR data generated within the project, the results were improved by using different models optimized for each class [Löfström, T., Zhao, J., Linusson, H., and Jansson, K. 2015].

In addition, efforts have been made to advance the field of conformal prediction, resulting in several studies providing theoretical progressions, including a novel method for performing conformal regression with stronger theoretical guarantees [Linusson, H., Johansson, U., & Löfström, T. 2014] and producing asymptotically valid p-values under slightly relaxed conditions [Carlsson, L. Ahlberg, E., Boström, H., Johansson, U. and Linusson, H. 2015]. Practical details pertaining to the usage of conformal predictors have also been studied, including a study on how to handle small calibration sets in mondrian inductive conformal regressors [Johansson, U, Ahlberg, E., Boström, H. Carlsson, L, Linusson, H. and Sönströd, C. 2015], an investigation into assessing the a posteriori error rates of binary conformal classifiers [H. Linusson, U. Johansson, H. Boström and T. Löfström. 2016], a comparison of the informational efficiency of various types of conformal predictors [H Linusson, U Johansson, H Boström, T Löfström. 2014], as well as studies on how to define optimal conformal predictors using different kinds of machine learning algorithms [Johansson, U., Boström, H. and Löfström, T. 2013; Johansson, U., Boström, H., Löfström, T. and Linusson, H. 2014].

WP: Parallel data mining

This work package has developed and implemented parallel machine learning algorithms for graphical processing units (GPU) and multi-core processors. The research has focused on parallelization of different types of decision tree ensemble techniques for classification and regression tasks, mainly for the massively parallel GPU platform. The purpose with the research has been to enable bigger data sets to be processed in less time thus increasing workflow speed, and or, the ability to handle much larger data sets than previously possible.

Random Forests and Extremely Randomized Trees for GPUs

This part of the work package focused on overcoming the challenges that comes when parallelizing Random Forests and Extremely Randomized Trees for the GPU architecture. For training, the developed algorithm takes a breadth first approach where an entire node-level of an ensemble can be built in parallel on the available GPU's on the system. Furthermore, each node is processed in parallel on a GPU using two GPU thread-blocks. With these two levels of parallelization the algorithm can drill down to a more than sufficient level of granularity to saturate the multiple thousands of processing cores on the GPU. The implementations are denoted as - *gpuRF* for Random Forests and *gpuERT* for Extremely Randomized Trees.

Random Forests and Extremely Randomized Trees for multi-core CPUs
This part of the work package focused on implementing and optimizing efficient multi-core CPU algorithms of Random Forests and Extremely Randomized Trees. The resulting implementations are used as a baseline comparisons with the GPU implementations, as well as being efficient and fast alternatives for computers that lack a powerful GPU. The implementations are denoted as - *cpuRF* for Random Forests and *cpuERT* for Extremely Randomized Trees.

The results from the studies into the GPU and multi-core CPU algorithms are presented in [Jansson, K., Sundell, H., Boström, H., 2014] with early results being presented in [Jansson, K., Sundell, H., & Boström, H., 2013]. Extended versions of the multi-core CPU implementations, with balanced under-sampling, were also used in another study within the project [Löfström, T., Zhao, J., Linusson, H., and Jansson, K., 2015].

CPU-GPU hybrid implementations
This part of the work package focused on combining the results from the previous studies. Enabling further training and prediction acceleration through usage of the GPU and CPU implementations concurrently on a single machine. The main challenges here was the balancing between the CPU and GPU implementations, and making sure that the different implementations could sync their individual output into one resulting model or prediction. The results from this work do not have an article related to them, but the implementations are present in the final software output of the work package.

The final software output of this work package is published as open source on GitHub under the MIT license. It includes improved implementations of all the multi-GPU and multi-core CPU algorithms that have been presented in earlier articles. In addition, the multi-core CPU implementations have been extended to support Non-uniform memory access (NUMA) for usage on server computers.

2.2 Participating researchers

Detailed information about the participating researchers is provided in appendix A.3.

2.3 Publications

The publications of the project are listed in appendix A.4.

2.4 Activities

Members of the project organized the Second Swedish Data Science Workshop in 2014, gathering around 60 researchers and practitioners from industry and academia.

Members of the project hosted the 15th International Symposium on Intelligent Data Analysis (IDA) 2016.

Members of the group organized the Fifth International Workshop on Health Text Mining and Information Analysis (LOUHI) 2014.

The full list of events appears in appendix A5.

3. Strategic relevance

3.1 How have the project members ensured that PhDs and research results from the project are incorporated or utilized by industry and society?

The researchers and PhD students in the project have been working with clinicians and researchers in the pharmaceutical industry, resulting in several joint publications and project applications. Software developed in the project has been utilized by the collaborating partners.

3.2 Describe the collaboration with industry and other parts of society

The collaboration has been manifested through meetings and workshops, in which challenges within the application domains and results from applying developed methods have been discussed. As indicated above, this has led to joint publications (see Appendix A.4) and also project applications (see section 1.3 for ongoing projects). The collaborating partners have also acted as testers of the developed software (described in Appendix A.6).

3.3 Describe the industrially or societally relevant results of the project.

Software that has been developed is listed in appendix A.6.

3.4 Describe the intellectual assets and property rights developed by the project.

Intellectual assets, in addition to the software listed in Appendix A.6, are listed in appendix A.7.

3.5 Which research results of the project have been [or will be within six months of the project's contractual expiration] implemented by industry/society?

There are currently no known plans for external partners to include the software developed in the project into products or to put it into production. However, several of the developed packages are made public under, e.g., the MIT license, effectively allowing external partners to freely exploit the software.

3.6 Which activities, publications, etc. have been directed towards the general public or to younger people?

Open Seminar held by Henrik Boström and Hercules Dalianis in conjunction with 50 year jubilee of the Dept. of Computer and Systems Sciences at SU on Sep. 9, 2016 (in Swedish): *Artificiell intelligens för analys av patientjournaler*, link

Talk by Hercules Dalianis at Stockholm NLP Meetup, at the news monitoring company Meltwater, February 29, 2016, *Clinical text retrieval – some challenges, methods and applications using Swedish patient records to improve healthcare*, link

Talk by Hercules Dalianis, at Meetings at Health 2.0 at HUB, March 6, 2013, *How can advanced language technology be used to extract valuable information from free text in collections of patient records?* link

Invited talk by Hercules Dalianis at Global Forum Nov 2012, Grand Hotel, Stockholm, *Clinical Text Mining for Health Care Managers using aggregated data in the Cloud,* link

Ulf Johansson and Henrik Linusson gave a tutorial about conformal prediction at the International Joint Conference of Neural Networks (IJCNN) 2015, link

3.7 If your project was eligible for 3% "nyttiggörande" please describe here how these resources were used, the effect of them, and future plans

Through a public procurement, Novelari, was contracted to package and optimize the Julia code for random forests for classification, regression and survival analysis with conformal prediction that has been developed in the project, and furthermore to integrate the package with the also developed CPU-GPU hybrid implementation of random forests, and to implement a web-based GUI. Test suits, examples and documentation were also developed, which together with the package can be found at: https://github.com/henrikbostrom/RandomForest
The plan is to reach a wider audience for the software by releasing the Julia package through the Julia ecosystem under the MIT licence.

4. The graduate training of the project

4.1 Has the project contributed to an improved graduate training?

No courses have been specifically developed for the project.

4.2 Which younger researchers have been able to establish themselves as independent group leaders in academy or research leaders in industry as a result of the project?

Dr. Maria Skeppstedt who finished her PhD during the project, but without direct financial support from the project, has after her PhD become postdoc both at Gavagai AB and Linnaeus University in Sweden and now 2016 with her own funding from Swedish Research Council, Vetenskapsrådet, VR, at University of Potsdam, Germany.

4.3 Exams

The students and their exams (or lack of) are listed in appendices A.9-A.12

5. Collaborations

5.1 Scientific collaborations within the project between participating groups

The research groups at Stockholm University and University of Borås are collaborating closely within (and outside) the project through joint publications, supervision and discussions of both scientific and strategic issues. All three PhD students located at University of Borås have been jointly supervised by a supervisor from the group in Borås and a supervisor from the group in Stockholm. Several of the scientific publications are jointly authored from researchers and PhD students assigned to different work packages. The collaboration between researchers and PhD students in different work packages is further strengthened by the joint efforts to develop the demonstrator, in which all project members participated, either directly by designing and coding, or indirectly, by providing requirements and feedback from evaluations. The joint study program further contributes to strengthening the collaboration between the PhD students in the project.

5.2 Scientific collaboration between different disciplines and departments

The setup of the project has required true interdisciplinary collaboration, involving researchers from computer science, medicine and chemistry, in particular involving the areas of data science, computational linguistics, medicinal chemistry, clinical pharmacology and pharmacoepidemiology. The collaboration has resulted in several joint publications with researchers from different departments and disciplines as co-authors. The departments include Dept. of Linguistics at Stockholm University, Dept. of Computational Linguistics at Uppsala University, Dept. of Medicine, Dept. of Neurobiology, Care Sciences and Society, and Centre for Pharmacoepidemiology at the Karolinska Institute.

5.3 International collaboration

Members of the project have collaborated with researchers at the Finnish Institute of Occupational Health in Helsinki, Finland, including several research visits resulting several joint publications. The collaboration has also resulted in a joint project application to the Nordic Research Council on *Social Dynamics of Health and Well-being*, including co-applicants from Aalborg University, Denmark, and Aalto University, Finland. A workshop to discuss research collaboration among the applicants was held in Stockholm.

The groups at Stockholm University and University of Borås participated in an EU application on a joint PhD training network with a focus on conformal prediction for *in silico* modeling, which was to be headed by AstraZeneca AB, and also involved Royal Holloway, University of London, and Lundbeck A/S, Denmark.

Aron Henriksson and Maria Skeppstedt have collaborated with researchers at the University of California, San Diego (UCSD). Aron and Maria have spent one month on a research visit at the Division of Biomedical Informatics at UCSD, working with Professor Wendy Chapman and Dr. Mike Conway. This collaboration has resulted in a number of joint publications. Aron spent another two weeks visiting Dr. Mike Conway, initiating a small project to detect perceived adverse drug events in social media, in particular Twitter. Aron has also spent two weeks visiting Professor Marco Baroni's CLIC group at the Center for Mind/Brain Sciences of the University of Trento. The CLIC group is prominent in the area of distributional semantics, which Aron has been working on and applying to medical data.

Professor Hercules Dalianis obtained funds from *Riksbankens Jubileumsstiftelse* to carry out a sabbatical stay at CSIRO/Macquarie University, Sydney, Australia 2016-2017, to author a text book in clinical text mining.

5.4 Collaboration with industry and other parts of society

The project members have been collaborating closely with researchers from the pharmaceutical industry, primarily at AstraZeneca, Mölndal, and H. Lundbeck A/S, Denmark, with frequent visits and teleconferences. The collaboration has resulted in several joint publications as well as joint work on technological developments and an EU application for a PhD training network.

Meetings have been held at several occasions with researchers from WHO Uppsala Monitoring Centre on pharmacovigilance and they have expressed a strong interest in the results of the project. The collaboration has mainly been on the exchange of ideas and experience.

The collaboration with healthcare providers within the project has resulted in three additional projects on data and text mining of health records, co-funded by Stockholm University and Stockholm County Council. The main focus of two of these projects is to use registry data from a regional healthcare data warehouse (Vårdanalysdatabasen, VAL) to analyse and better understand the treatment of heart failure patients. A position as Postdoctoral Fellow in Data Science has been announced as part of one of these projects. Another project focused on developing methods for de-identification and pseudonymization of health records as a means to facilitate the secondary use of clinical data.

The participation in the Nordic Center of Excellence in Health-Related e-Sciences (NIASC) has led to funding of a project on *Data and text mining of cancer symptoms and comorbidities in electronic patient records in the Nordic languages* (MINECAN). The project has recruited a PhD student to study early symptoms of cancers and to extract information from pathology reports for cancer registries. Collaborating partners are Karolinska Institutet, CBS, University of Copenhagen and Cancer Registry of Norway.

Members of the project are currently participating in a research project funded by Vinnova, *HAI Proactive - New Workflows and IT Tools to Combat Healthcare-Associated Infections,* during 2016-2018. The project partly focuses on developing algorithms for analyzing health records in order to detect another type of adverse healthcare event, namely healthcare-associated infections. The project is led by Karolinska University Hospital and is a collaboration between several county councils, academia (Stockholm University and Karolinska Institute) and industry (SAS Institute, Tieto, and Treat Systems).

Members of the project have participated in two research projects funded by the Knowledge foundation: *Big Data Analytics by Online Ensemble Learning* (BOEL) 2013-2015 and *Data Analytics for Research and Development* (DASTARD) 2016-2018. Both these projects focus on developing novel machine learning methods and algorithms for data analytics and involve AstraZeneca and Scania as partners. The latter project is currently funding a postdoctoral researcher working on-site at AstraZeneca.

5.5 What has the project meant to the researchers in the project?

The probably most important outcome of the project is that it has contributed to creating a critical mass of expertise within data science with extensive experience in analyzing electronic patient records and medicinal chemistry data.

The demand for confidence-based predictions from the pharmaceutical industry has led to an entirely new focus on conformal prediction, which now is a central component of the research conducted within the research groups at Stockholm University and University of Borås.

Several project members have, through the project, been able to establish closer contacts with a wide range of expertise from other disciplines. These include clinical pharmacologists and researchers from the pharmaceutical industry as well as medical doctors working with clinical healthcare on a daily basis.

6. Continued work after the project is finished

6.1 What are the main lessons learned from the project?

The perhaps most important lesson from the project is that the needs and requirements that stem from domain experts and stakeholders can sometimes give rise to a completely new focus for the research, as exemplified in this project by the need for predictions with confidence (conformal prediction) as requested by researchers in the pharmaceutical industry. By keeping an open mind from the side of the researchers, rather than trying to enforce tools and techniques onto the end-users, completely new research findings and directions may come out as a result.

One main lesson learned from the project concerns the importance of involving domain experts as early as possible in the process to identify relevant problems within the application area, as well as for choosing relevant criteria of success. It is often hard to rely on their interest only, as they frequently are highly requested in their daily work. One option would instead be to, if possible, contract them to participate actively in the project. Although not really a novel lesson, one should always keep in mind that data can rarely be easily obtained from end-users, and the complexity of the task of getting the data into a workable format is often underestimated.

Another lesson concerns the development of demonstrators and software packages, which have to be appropriately budgeted and allocated time for. Although implementations to some extent could be done as part of regular research work, any requirements on user-friendliness and documentation can hardly be satisfied as part of the research work.

6.2 What will happen to the project?

A natural next step after the project has ended, is to exploit the critical mass of expertise that has been built up, through collaboration with authorities and organizations with access to more and additional types of data, such as biobanks and registries for larger parts of the Swedish population. Collaborations with developers and maintainers of electronic patient record systems would also increase the possibilities of exploiting this expertise to improve data collection and organization to support learning over time, not only regarding drug effects, but for any health surveillance task that requires massive data collected over time.

The expertise within high-performance data and text mining will also be exploited in other application areas of interest to the industry and society at large, e.g., through ongoing collaboration with Scania AB, which targets optimizing maintenance and reducing the environmental pressure by learning from massive data collected on-board at trucks. The work on extending the applicability of learning algorithms, e.g., for new types of data, as well as meeting requirements on robustness, efficiency and interpretability will be an effort that does not end with this project.

The work in the ongoing related projects will continue, e.g., the participation in the Nordic Center of Excellence in Health-Related e-Sciences (NIASC) and the project Data Analytics for Research and Development (DASTARD) will continue until end of 2018, while the Coril project, funded by Stockholm County council, will continue until the end of 2019. These projects, as well as additional projects, which have been applied for, e.g., from the Swedish Research council, aim for supporting a continuation for the researchers and graduated PhDs in the project.

6.3 Give a brief long term perspective on the field of the project

Accurate estimates of risk for adverse drug events is of high relevance to authorities, such as the Medical Products Agency, e.g., to allow for valid recommendations and decisions, such as the withdrawal of a product from the market. Currently, such estimates are based on reports from the pharmaceutical industry and organizations, such as WHO Uppsala Monitor Center, which mainly obtain such signals through analysis of clinical trials and spontaneous reports from healthcare personnel, respectively. We expect that the future of the area of pharmacovigilance to a large extent will rely also on the analysis of electronic patient records, as explored in this project.

The ability to detect effects of drug use in large populations may also be very beneficial to pharmaceutical companies, not only for identifying adverse effects, but also for identifying possibly unexpected positive effects, which could result in new markets for already approved drugs.

The technological developments within the project will be of relevance not only for analysis of electronic patient records and chemical compounds, but can be used in any application area with a need for high-performance analysis of both structured and unstructured data. The demand for technology and expertise within data science in industry and society at large has increased significantly during the lifetime of the project, and this development is expected to continue within the decades to come.

7. Costs

The total cost for the project was 18 430 000 SEK, including OH on 4 735 173 SEK. In addition to this, 392 680 SEK, was used for utilization, for which Novelari was contracted through public procurement to package and optimize the Julia code for random forests for classification, regression and survival analysis with conformal prediction that has been developed in the project, and furthermore to integrate the package with the also developed CPU-GPU hybrid implementation of random forests, and to implement a web-based GUI. Test suits, examples and documentation were also developed.

The project was extended from the originally planned ending in December 2016 to March 2017. During the extension the principal investigator was finalizing part of the outcome of the utilization project, i.e., the Julia package, summarizing the project results for the final report, and took part in the organization of two PhD defences, that were the result of the project.

8. External information and other activities

In addition to the invited talks and open seminars listed in Section 3.6, members of the project have disseminated information about the project at the following events:

Aron Henriksson was invited to present results of the project at the annual scientific conference of a similar research project funded by the French National Research Agency, PractiKPharma at LORIA in Nancy, France, http://practikpharma.loria.fr

Henrik Boström presented results from the project at the first Swedish Data Science workshop at University of Borås in 2013 as well as on the third Swedish Data Science workshop at Blekinge Institute of Technology in 2015.

Henrik Boström and Hercules Dalianis presented the project at a meeting for the Network on Structured Patient Data in March 2013, organized by Prof. Ingvar Krakau at Dept. of Medicine at Karolinska Institute with more than 40 attending researchers.

Henrik Boström gave an overview of the project at a seminar on big data analytics at Swedish Institute for Computer Science in October 2012 with over 30 participants from Swedish industry and universities.

The project has been presented for the Regional Cancer Centre Stockholm–Gotland, which has shown interest in the techniques and results of the project. The project has also been presented both for Pygargus AB, a clinical trials brooker, and Capish Knowledge AB, a tools suppliers that have developed QlickView. Representatives from e.g., IBM, Amgen and Microsoft have made visits to discuss collaborations within the project.

9. SWOT analysis

The identified strengths, i.e., the project members constitute one of the strongest groups nationally within data science, with expertise in data mining, text mining and high performance computing, the access to abundant clinical data, the well-established collaborations within the researchers in the pharmaceutical industry and medicine, and extensive experience in software development, can to a large extent explain the successful outcome of the project. The identified threats, i.e., that AstraZeneca would shut down its research activities in Mölndal, similar to what has been done in Södertälje and Lund, and that the debate on privacy issues could make it hard or impossible to get access to and analyze even anonymized healthcare data, turned out not to take place. The identified weaknesses, i.e., lack of expertise within the group on adverse drug reactions and clinical pharmacology and lack of access to healthcare systems in use, were to some extent compensated for by the collaboration with external partners. However, the results of the project would most likely be even more relevant for clinical practice with more direct involvement of experts in the above areas. The first listed opportunity, i.e., to collaborate with the product development and consultancy company Tieto for exploitation of the results of the project in production environments using healthcare data, was discussed but not materialized. The second opportunity, i.e., that large investments in data analysis is expected for the coming years, which will lead to increased awareness and requirements for additional development, is still considered to be valid, and has been manifested through the high demand from industry for well educated and experienced data scientists.

A. Appendices

A.3 A list of the researchers

Henrik Boström, professor, Dept. of Computer and Systems Sciences, Stockholm University, PhD 1993, male
Lars Asker, docent, associate professor, Dept. of Computer and Systems Sciences, Stockholm University, PhD 1994, male.
Hercules Dalianis, professor, Dept. of Computer and Systems Sciences, Stockholm University, PhD 1996, male
Ulf Johansson, professor, Dept. of Computer Science and Informatics, Jönköping University [part-time at Dept. of Information Technology, University of Borås], PhD 2007, male
Håkan Sundell, docent, associate professor, Dept. of Information Technology, University of Borås, PhD 2004, male.

A.4 A list of selected publications

Journal publications

H. Boström, U. Johansson, H. Linusson, and T. Löfström, 2017. Accelerating difficulty estimation for conformal regression forests. *Annals of Mathematics and Artificial Intelligence*, pp. 1-20.

J. Zhao, P. Papapetrou, L. Asker and H. Boström, 2017. Learning from heterogeneous temporal data in electronic health records. *Journal of Biomedical Informatics*. 65, pp. 105-119.

M. Ahltorp, M. Skeppstedt, S. Kitajima, A. Henriksson, R. Rzepka and K. Araki, 2016. Expansion of medical vocabularies using distributional semantics on Japanese patient blogs. *Journal of Biomedical Semantics*, 7:58, pp. 1-18.

I. Karlsson, P. Papapetrou and H. Boström, 2016. Generalized random shapelet forests. *Data Min. Knowl. Discov.* 30(5): 1053-1085

G. Grigonyte, M. Kvist, M. Wiren, S. Velupillai. and A. Henriksson, 2016. Swedification patterns of Latin and Greek affixes in clinical text. *Nordic Journal of Linguistics*, 39(1), pp. 5-37.

A. Henriksson, J. Zhao, H. Dalianis, and H. Boström. 2016. Ensembles of randomized trees using diverse distributed representations of clinical events. *BMC Medical Informatics and Decision Making*. 16(S2):69, pp. 85-95

J. Zhao and A. Henriksson. 2016. Learning temporal weights of clinical events using variable importance. *BMC Medical Informatics and Decision Making*. 16(S2):71, pp. 111-121

Henriksson, A. 2015a. Learning Multiple Distributed Prototypes of Semantic Categories for Named Entity Recognition. *International Journal of Data Mining and Bioinformatics*, Vol. 13, No. 4, pp. 395-411.

Henriksson, A., Kvist, M., Dalianis, H., Duneld, M. 2015. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *Journal of Biomedical Informatics*, 57: 333-349.

Kotsifakos, A., Karlsson, I., Papapetrou, P., Athitsos, V., & Gunopulos, D. 2015. Embedding-based subsequence matching with gaps–range–tolerances: a Query-By-Humming application. *The VLDB Journal*, 24(4), 519-536.

Löfström, T., Boström, H., Linusson, H., Johansson, U. 2015. Bias Reduction through Conditional Conformal Prediction. *Intelligent Data Analysis*, Vol. 9, nr 6, pp. 1355-1375.

Velupillai, S., M. Duneld, A. Henriksson, M. Kvist, M. Skeppstedt and H. Dalianis. 2015.
Louhi 2014: Special issue on health text mining and information analysis, *BMC Medical Informatics and Decision Making*, 2015, 15(Suppl 2):S1

Zhao J., Henriksson A., Asker L., Boström H. Predictive modeling of structured electronic health records for adverse drug event detection. *BMC Med Inform Decis Mak.* 2015 Nov 25;15 (Suppl 4):S1. doi: 10.1186/1472-6947-15-S4-S1. Epub 2015 Nov 25.

Henelius, A., Puolamäki, K., Boström, H., Asker, L., and Papapetrou, P. 2014. A peek into the black box: exploring classifiers by randomization. *Data Min. Knowl. Discov.* 28, 5-6 (September 2014), 1503-1529

Henriksson, A., Moen, H., Skeppstedt, M., Daudaravicius, V. and Duneld, M. Synonym Extraction and Abbreviation Expansion with Ensembles of Semantic Spaces. *Journal of Biomedical Semantics*, 5:6, 2014.

Johansson, U., Boström, H., Löfström, T. and Linusson, H. Regression Conformal Prediction with Random Forests, *Machine Learning*, 97(1-2): 155-176 2014.

Skeppstedt, M., M. Kvist, H. Dalianis and G.H. Nilsson. 2014. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics*, DOI: 10.1016/j.jbi.2014.01.012.

Velupillai, S., M. Skeppstedt, M. Kvist, D. Mowery, B. Chapman, H. Dalianis and W. Chapman, W. 2014. Cue-based assertion classification for Swedish clinical text - developing a lexicon for pyConTextSwe. Special issue: Text Mining and Information Analysis. *Artificial Intelligence In Medicine*, DOI: 10.1016/j.artmed.2014.01.001.

Karunaratne, T. and H. Boström and U. Norinder, Comparative analysis of the use of chemoinformatics-based and substructure-based descriptors for quantitative structure-activity relationship (QSAR) modeling. *Intelligent Data Analysis*, Vol. 17, No. 2, IOS press, 2013.

Norinder, U. and Boström H., Representing descriptors derived from multiple conformations as uncertain features for machine learning. *Journal of Molecular Modeling*, Vol. 19, No. 6, pp. 2679-2685, Springer, 2013.

Norinder, U. and Boström, H. (2012). Introducing Uncertainty in Predictive Modeling - Friend or Foe?. *Journal of Chemical Information and Modeling*, vol. 52, pp. 2815-2822

Conference publications

I. Karlsson, P. Papapetrou and H. Boström, 2016. Early Random Shapelet Forest. Proc. of the 19th International Conference on Discovery Science. Springer, pp. 261-276 [Carl H. Smith Award for best paper]

I. Karlsson and H. Boström, 2016. Predicting Adverse Drug Events using Heterogeneous Event Sequences. Proc. of the IEEE International Conference on Healthcare Informatics (ICHI 2016), IEEE

H. Boström, H. Linusson, T. Löfström and U. Johansson. 2016. Evaluation of a Variance-Based Nonconformity Measure for Regression Forests. Proc. of the International Symposium on Conformal and Probabilistic Prediction with Applications, Springer, pp. 75-89.

L. Asker, H. Boström, P. Papapetrou and H. Persson. 2016. Identifying Factors for the Effectiveness of Treatment of Heart Failure: A Registry Study. CBMS, pp. 205-206

L. Asker, P. Papapetrou and H. Boström. 2016. Learning from Swedish Healthcare Data. PETRA, ACM.

Carlsson, L. Ahlberg, E., Boström, H., Johansson, U. and Linusson, H. 2015. Modifications to p-Values of Conformal Predictors, International Symposium on Learning and Data Sciences (SLDS), 2015.

Dalianis, H., A. Henriksson, M. Kvist, S. Velupillai and R. Weegar. 2015. HEALTH BANK - A Workbench for Data Science Applications in Healthcare. Proceedings of the CAiSE-2015 Industry Track co-located with 27th Conference on Advanced Information Systems Engineering (CAiSE 2015), J. Krogstie, G. Juell-Skielse and V. Kabilan, (Eds.), Stockholm, Sweden, June 11, 2015, CEUR, Vol-1381, urn:nbn:de:0074-1381-0, pp 1-1

Henelius, A., Puolamäki, K., Karlsson, I., Zhao, J., Asker, L., Boström, H., & Papapetrou, P. (2015). GoldenEye^{++}: A Closer Look into the Black Box. In Statistical Learning and Data Sciences (pp. 96-105). Springer International Publishing.

Henriksson, A., Zhao, J., Boström, H., and Dalianis, H. 2015a. Modeling Heterogeneous Clinical Sequence Data in Semantic Space for Adverse Drug Event Detection. In Proceedings of IEEE International Conference on Data Science and Advanced Analytics (DSAA).

Henriksson, A., Zhao, J., Boström, H., and Dalianis, H. 2015b. Modeling Electronic Health Records in Ensembles of Semantic Spaces for Adverse Drug Event Detection. In Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM).

Johansson, U, Ahlberg, E., Boström, H. Carlsson, L, Linusson, H. and Sönströd, C. (2015), Handling Small Calibration Sets in Mondrian Inductive Conformal Regressors, International Symposium on Learning and Data Sciences (SLDS), 2015.

Karlsson, I., Papapetrou, P., and Boström, H., Forests of Randomized Shapelet Trees. International Symposium on Learning and Data Sciences (SLDS), 2015.

Karlsson, I., Papapetrou, P., & Asker, L. (2015, July). Multi-channel ECG classification using forests of randomized shapelet trees. In Proceedings of the 8th ACM International Conference on Pervasive Technologies Related to Assistive Environments (p. 43). ACM.

Löfström, T., Zhao, J., Linusson, H., and Jansson, K. Predicting Adverse Drug Events with Confidence. In Proceedings of the 13th Scandinavian Conference on Artificial Intelligence (SCAI). 2015

Velupillai, S., D. Mowery, B.R. South, M. Kvist and H Dalianis. (2015). Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis. Yearbook of medical informatics, 10(1), 183-193

Zhao, J. Temporal Weighting of Clinical Events in Electronic Health Records for Pharmacovigilance. In Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2015.

Zhao, J., Henriksson, A., Kvist, M., Asker, L., and Boström, H., Handling Temporality of Clinical Events for Drug Safety Surveillance, In Annual Symposium of American Medical Informatics Association, 2015 [Distinguished Paper Award].

Zhao, J., Henriksson, A., and Boström, H., Cascading Ensemble Classifiers for the Detection of Adverse Drug Events in Electronic Health Records, In Proc of IEEE International Conf. on Data Science and Advanced Analytics (DSAA), 2015.

Zhao J., Henriksson A., Asker L. and Boström H. Detecting adverse drug events with multiple representations of clinical measurements. In Proceedings of *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 536–543, November 2-5, 2014, Belfast, UK.

Zhao J., Henriksson A. and Boström H. Detecting adverse drug events using concept hierarchies of clinical codes. In Proceedings of *IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 285–293, September 15-17, 2014, Verona, Italy

Asker, L., Boström, H., Karlsson, I., Papapetrou, P. and Zhao, J. Mining Candidates for Adverse Drug Interactions in Electronic Patient Records. In Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments, PETRA'14, May 27-30, 2014, Island of Rhodes, Greece.

Henriksson, A., Dalianis, H., Kowalski, S. Generating Features for Named Entity Recognition by Learning Prototypes in Semantic Space: The Case of De-Identifying Health Records. In Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2014.

Henriksson, A., Conway, M., Duneld, M. and Chapman, W. Identifying Synonymy between SNOMED Clinical Terms of Varying Length Using Distributional Analysis of Electronic Health Records. In Proceedings of the Annual Symposium of the American Medical Informatics Association, AMIA 2013, pp. 600-609, American Medical Informatics Association, 2013, Washington DC, USA.

Henriksson, A., Skeppstedt, M., Kvist, M., Duneld, M. and Conway, M. Corpus-Driven Terminology Development: Populating Swedish SNOMED CT with Synonyms Extracted from Electronic Health Records. In Proceedings of BioNLP, pp. 36-44, Association for Computational Linguistics, 2013, Sofia, Bulgaria.

Johansson, U., Boström, H. and Löfström, T. (2013), Conformal Prediction Using Decision Trees, IEEE International Conference on Data Mining (ICDM), pp. 330-339, Dallas, TX.

Johansson, U., Löfström, T. and Boström, H. (2013), Random Brains, The International Joint Conference on Neural Networks (IJCNN), Dallas, TX, IEEE.

Johansson, U., König, R., Löfström, T. and Boström, H. (2013), Evolved Decision Trees as Conformal Predictors, IEEE Congress on Evolutionary Computation (CEC), pp. 1794-1801, Cancun, Mexico.

Johansson, U., Löfström, T. and Boström, H. (2013), Overproduce-and-Select: The Grim Reality, Computational Intelligence and Ensemble Learning, IEEE Symposium Series on Computational Intelligence (SSCI), pp. 52-59, Singapore.

Johansson, U. and Löfström, T. (2012). Producing Implicit Diversity in ANN Ensembles, The International Joint Conference on Neural Networks,pp. 1-8, Brisbane, Australia.

Karlsson I., J. Zhao, L. Asker and H. Boström, Predicting Adverse Drug Events by Analyzing Electronic Patient Records. Proc. of the 14th Conference on Artificial Intelligence in Medicine (*AIME*), Lecture Notes in Computer Science, Vol. 7885, pp. 125-129, Springer Publishing Company, 2013.

Karlsson, I. and Zhao, J. Dimensionality Reduction with Random Indexing: an Application on Adverse Drug Event Detection using Electronic Health Records. In Proceedings of the 27th International Symposium on Computer-Based Medical Systems (CBMS), May 27-29, 2014, New York, USA.

Karlsson, I. and Boström, H. Handling Sparsity with Random Forests when Predicting Adverse Drug Events from Electronic Health Records. In Proceedings of IEEE International Conference on Healthcare Informatics (ICHI), September 15-17, 2014 (to appear), Verona, Italy.

Karunaratne, T. and Boström, H. (2012). Can frequent itemset mining be efficiently and effectively used for learning from graph data?, In Proc. of 11th International Conference on Machine Learning and Applications, pp. 409-414

Kvist, M. and S. Velupillai. 2013. Professional Language in Swedish Radiology Reports – Characterization for Patient-Adapted Text Simplification. In Proc. of Scandinavian Conference on Health Informatics 2013 pp. 55-59, Linköping University Electronic Press.

Kvist, M. and S. Velupillai. 2014. SCAN: A Swedish Clinical Abbreviation Normalizer. Further Development and Adaptation to Radiology. To appear in: Lecture Notes in Computer Science, Springer. Conference and Labs of the Evaluation Forum (CLEF 2014), Sheffield, UK, sept 2014.

Linusson, H., Johansson, U., Boström, H., and Löfström, T. Efficiency comparison of unstable transductive and inductive conformal classifiers. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 261-270. Springer Berlin Heidelberg, 2014.

Linusson, H., Johansson, U. and Löfström, T.. Signed-error conformal regression. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 224-236. Springer International Publishing, 2014.

Linusson, H., Johansson, U., Boström, H., and Löfström, T. Reliable Confidence Predictions Using Conformal Prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 77-88. Springer International Publishing, 2016.

T. Löfström, U. Johansson and H. Boström, Effective Utilization of Data in Inductive Conformal Prediction using Ensembles of Neural Networks. pp. 1-8, in IEEE conference proceedings of the 2013 International Joint Conference on Neural Networks (IJCNN), 2013.

S. Meystre, H. Dalianis, J. Aberdeen and B. Malin. Automatic clinical text de-identification: is it worth it, and could it work for me?. In Studies in Health Technology and Informatics, Vol. 192, pp. 1242-1242, IOS Press, 2013.

Skeppstedt, M., Kvist, M. and Dalianis, H. .2012. Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text. In Proceedings of International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey.

Skeppstedt, M. Ahltorp, M. and Henriksson, A. Vocabulary Expansion by Semantic Extraction of Medical Terms. In Proceedings of Languages in Biology and Medicine, LBM 2013, December 12-13, 2013, Tokyo, Japan.

Tanushi, H., H. Dalianis, M. Duneld, M. Kvist, M. Skeppstedt and S. Velupillai. Negation Scope Delimitation in Clinical Text Using Three Approaches: NegEx, PyConTextNLP and SynNeg." *19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. Linköping University Electronic Press, pp. 387-397, 2013.

ul Muntaha S., Skeppstedt M., Kvist M and H. Dalianis. 2012. Entity Recognition of Pharmaceutical Drugs in Swedish Clinical Text. In Proceedings of Swedish Language Technology Conference (SLTC 2012), Lund, Sweden.

Henriksson, A., Moen, H., Skeppstedt, M., Eklund, A-M., Daudaravicius, V. and Hassel, M. 2012. Synonym Extraction of Medical Terms from Clinical Text Using Combinations of Word Space Models. In Proceedings of Semantic Mining in Biomedicine, SMBM 2012, Zurich, Switzerland.

Workshop publications

Alfalahi, A., Skeppstedt, M., Ahlbom, R., Baskalayci, R., Henriksson, A., Asker, L., Paradis, C., Kerren, A. 2015. Expanding a dictionary of marker words for uncertainty and negation using distributional semantics. In Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), Lisbon, Portugal.

Friedrich, S. and H. Dalianis. 2015. Adverse drug event classification of health records using dictionary-based pre-processing and machine learning. In the proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis, Louhi, held in conjunction with EMNLP 2015, Lisbon, Portugal, pp 121-130,

Henriksson, A. 2015b. Representing Clinical Notes for Adverse Drug Event Detection. In the proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis, Louhi, held in conjunction with EMNLP 2015, Lisbon, Portugal.

Alfalahi, A., S. Brissman and H. Dalianis. 2012. Pseudonymisation of person names and other PHIs in an annotated clinical Swedish corpus. In Proc. of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC 2012, May 26, Istanbul, pp 49-54

Boström, H. and H. Dalianis, 2012. De-identifying health records by means of active learning. In Proc. of ICML Workshop on Machine Learning for Clinical Data Analysis, Edinburgh, UK.

Dalianis, H. and Boström, H. 2012. Releasing a Swedish Clinical Corpus after Removing all Words – De-identification Experiments with Conditional Random Fields and Random Forests. In Proc. of Third LREC Workshop on Building and Evaluating Resources for Biomedical Text Mining

Henriksson, A., Kvist, M., Hassel, M. and Dalianis, H. 2012. Exploration of Adverse Drug Reactions in Semantic Vector Space Models of Clinical Text. In Proceedings of the ICML Workshop on Machine Learning for Clinical Data Analysis, Edinburgh, UK,

Henriksson, A. and Duneld, M. 2013. Optimizing the Dimensionality of Clinical Term Spaces for Improved Diagnosis Coding Support. In Proceedings of Louhi Workshop on Health Document Text Mining and Information Analysis, NICTA, Sydney, Australia.

Jansson, K., Sundell, H., and Boström, H. (2013, November). Parallel tree-ensemble algorithms for GPUs using CUDA. In *Sixth Swedish Workshop on Multicore Computing*., Halmstad, Sweden.

Jansson, K., Sundell, H., and Boström, H. (2014), gpuRF and gpuERT: Efficient and Scalable GPU Algorithms for Decision Tree Ensembles. IEEE 28th International Parallel & Distributed Processing Symposium Workshops (IPDPSW), pp. 1612-1621, Phoenix, USA.

Tengstrand, L., Megyesi, B., Henriksson, A., Duneld, M. and Kvist, M. EACL – Expansion of Abbreviations in CLinical text. In Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations, PITR 2014, Association of Computational Linguistics, pp 94-103, Göteborg, Sweden.

Zhao, J., Karlsson, I., Asker, L. and Boström, H. Applying Methods for Signal Detection in Spontaneous Reports to Electronic Patient Records. In 19th Knowledge Discovery and Data Mining (KDD) Conference’s Workshop on Data Mining for Healthcare (DMH), August 11-14, 2013, Chicago, USA.

PhD theses

Karlsson, I. 2017. Order in the Random Forest, Stockholm University.

Zhao, J. 2017. Learning Predictive Models from Electronic Health Records, Stockholm University.

Henriksson, A. 2015. Ensembles of Semantic Spaces: On Combining Models of Distributional Semantics with Applications in Healthcare, Stockholm University.

Löfström, T. 2015. On Effectively Creating Ensembles of Classifiers: Studies on Creation Strategies, Diversity and Predicting with Confidence, Stockholm University.

Skeppstedt, M. 2015. Extracting Clinical Findings from Swedish Health Record Text, Stockholm University. *[Not funded by the project, but associated to the project]*

Licentiate thesis

Henriksson, A. 2013. Semantic Spaces of Clinical Text – Leveraging Distributional Semantics for Natural Language Processing of Electronic Health Records. Licentiate Thesis of Philosophy, Stockholm University.

A.5 A full list of events organised by the project

Members of the project organized the Second Swedish Data Science Workshop in 2014, gathering around 60 researchers and practitioners from industry and academia.

Members of the project hosted the 15th International Symposium on Intelligent Data Analysis (IDA) 2016.

Members of the group organized the Fifth International Workshop on Health Text Mining and Information Analysis (LOUHI) 2014.

A.6 A list of innovations and prototypes that have been produced, spin-off companies, etc

Below is a list of prototypes that have been developed within the project:

- A technical data mining and text mining back-end (aDEB) consolidating the software developed by the different parts of the project into one environment. aDEB centralizes shared functionality and acts as a bridge between tools written in different programming languages. Mainly intended to be used within the project to ease the collaboration between the work packages on a software level. No official link to the software. Owners: Isak Karlsson, Karl Jansson, Henrik Linusson
- Implementation of the adverse event exploration tool (aDEX). This module allows for defining the population of interest, as well as case and control groups, and obtaining descriptive and discriminatory characteristics of the groups. The implementation is available here: https://github.com/isakkarlsson/adex. Owners: Henrik Boström, Isak Karlsson

- Implementation of the adverse event detection tool (aDET). This module was aimed at doctors to help them identify potentially missing ADEs. No official link to the software. Owners: Henrik Linusson, Tuve Löfström, Karl Jansson.
- Implementations of Random Forests and Extremely Randomized Trees for graphics processing units (GPUs) and multi-core CPUs. Released as open source under the MIT license. The implementations are published and made available here: https://github.com/KarlJansson/DataminingLibs. Owner: Karl Jansson
- Implementation of the random forest algorithm for sparse and unbalanced data. The license does not restrict commercial use as long as the (possibly edited) source code is available. The implementation is available here: https://github.com/isakkarlsson/erlang-rr. Owner: Isak Karlsson
- Implementation of parts of the Rule Discovery System (RDS) in Erlang, including some additional features such as parallel execution on multi-core platforms, handling of uncertain data, conformal prediction, etc. A version for test trials is available here: https://www.box.com/s/1iyey6s7qbry1bx5gl7e. Owner: Henrik Boström
- A Julia package that implements random forests for classification, regression and survival analysis with conformal prediction. The implementation is available here: https://github.com/henrikbostrom/RandomForest. Owner: Henrik Boström
- A Python implementation of the conformal prediction framework, which primarily is to be used as an extension to the scikit-learn library. The implementation is available here: https://github.com/donlnz/nonconformist. Owner: Henrik Linusson
- Implementation of text mining tools for extracting information from patient record text. The tools are machine learning based using annotated clinical text. http://dsv.su.se/health/clinical-text-mining-tools. Owner: Hercules Dalianis and other researchers
- Implementation of data and matrix handling for time series machine learning written in Java under MIT license. The implementations are available here: https://gitlab.com/briljant/briljant, https://gitlab.com/briljant/mimir, and https://people.dsv.su.se/~isak-kar/grsf/. Owner: Isak Karlsson

A.7 A list of intellectual assets and property rights awarded or pending

The infrastructure HEALTH BANK (Swedish Health Record Research Bank), of which Stockholm EPR Corpus is a part, is a database containing over two million health records, from the years 2006-2014 from Karolinska University Hospital. The corpus encompasses over 500 clinical units. It contains both structured information as age, gender, pseudonymised social security number, ICD-10 diagnosis codes, ATC drug codes,blood and microbiological values, admission and discharge dates of the patient as well as unstructured information as patient records text. Professor Hercules Dalianis is responsible for HEALTH BANK, http://dsv.su.se/healthbank. Research may be carried out according to ethical permission.

A.8 A full list of all graduate/post-graduate courses developed within the project

No courses have been developed.

A.9 PhD exams

Students who have completed their PhD, with year, gender, thesis title, supervisor(s), university department, university of basic academic training, total amount of Foundation funding, and employer six months (or at a later time if available) after exam.

- Tuwe Löfström (male), 2015, “On Effectively Creating Ensembles of Classifiers: Studies on Creation Strategies, Diversity and Predicting with Confidence”, Henrik Boström and Ulf Johansson, Department of Computer and Systems Sciences, Stockholm University, University of Borås, 1 153 613 SEK, University of Borås.
- Aron Henriksson, 2015, male, “Ensembles of Semantic Spaces: On Combining Models of Distributional Semantics with Applications in Healthcare”, Hercules Dalianis and Martin Duneld, Department of Computer and Systems Sciences, Stockholm University, 1 644 553 SEK, Stockholm University.
- Jing Zhao, 2017, female, “Learning Predictive Models from Electronic Health Records”, Lars Asker and Henrik Boström, Department of Computer and Systems Sciences, Stockholm University, 1 610 327 SEK, Stockholm University.
- Isak Karlsson, 2017, male, “Order in the Random Forest”, Henrik Boström and Lars Asker, Department of Computer and Systems Sciences, Stockholm University, 1 430 606 SEK, Stockholm University.

A.10 Lic. exams

Ditto for students who have completed a licentiate exam.

- Aron Henriksson, 2013, male, “Semantic Spaces of Clinical Text: Leveraging Distributional Semantics for Natural Language Processing of Electronic Health Records”, Hercules Dalianis and Martin Duneld, Department of Computer and Systems Sciences, Stockholm University, 1 644 553 SEK, Stockholm University.

A.11 Future exams

- Henrik Linusson, expected completion: 2018-06-30. This is according to plan as he started on 2013-07-01.

A.12 No exams

- Karl Jansson

A.13 A list of awards to participating researchers, etc.

- Carl H. Smith Award for best paper for I. Karlsson, P. Papapetrou and H. Boström, 2016. Early Random Shapelet Forest. Proc. of the 19th International Conference on Discovery Science. Springer, pp. 261-276
- The Börje Langefors Prize awarded by SISA to Aron Henriksson for best PhD thesis in informatics at a Swedish university in 2016.
- Distinguished Paper Award to Jing Zhao et al. at American Medical Informatics Association (AMIA) Annual Symposium, 2015.

B Questions for the Project leader(s)

B.1 If the project had been set up today, what changes would you have made to it given everything that you now know [apart from the research results, of course]?

I would have allocated a part of the budget for clinicians and domain experts within clinical pharmacology and epidemiology to early on get help with identifying clinically relevant targets for the analysis. I would have liked to see the expertise within the project broadened from purely data science to the application areas. I would also have allocated part of the budget for the development of the demonstrator, which now to a large extent had to be developed as part of the PhD projects.

The project would also have benefited from ensuring that all developed components fit into an overall architecture through well-defined APIs. A less liberal approach regarding the choice of programming languages and platforms would most likely have promoted a closer collaboration and integration of the components.

B.2 What – if anything – will ultimately be the main impact of the project on society and academy?

The four PhDs (and a fifth in the pipeline) will bring technological expertise to academia, and at least indirectly to industry, in the field of data science, which is currently very much in demand. The project has contributed with novel findings regarding the analysis of electronic health records that provide stepping stones for future research, and which may in the long term have a large impact on clinical practice. From a data science perspective, the project has contributed with novel algorithms, implementations, empirical findings and theoretical results, in particular regarding random forests, conformal prediction and representations for text mining, which are expected to have a wide applicability also outside healthcare and pharmaceutical research. The latter will be supported by the developed software packages that will continue to be available after the termination of the project.

B.3 What do you expect will happen [What has happened...] to the activities within the project after the Foundation funding has expired?

The project has resulted in both new and deepened collaborations with end-users or stakeholders in the pharmaceutical industry and within the healthcare sector. These collaborations have resulted in now ongoing or planned projects, e.g., with Stockholm county council. All PhDs are planning to apply for postdoc positions to continue their academic careers.

B.4 What were the problems of the project?

The main problem was a lack of expertise within the project regarding what clinically relevant problems to study. Although this was handled through discussions with external partners, a higher degree of direct involvement would have sped up the process substantially.

Another problem concerned the development of the demonstrator. As only a part of this development could fit naturally within the PhD projects, in particular the implementation of novel algorithms, while other parts, such as GUIs, could not, the demonstrator would have benefited from having a part of the budget to cover such other development costs, in addition to the support for utilization.

B.5 What was the most fun with the project?

The most fun part of the project was to see the highly motivated PhD students take on their studies in an excellent way, resulting in numerous publications in top venues, some even awarded, and high-quality PhD theses, again some even rewarded.

B.6 Your main complaints and appreciations of the Foundation?

I think the communication with the program responsible has been excellent, with timely and informative responses on any question that we have had. The reporting and half-time evaluation were quite smooth and did not incur any substantial overhead.

I do not have any direct complaints, but a few suggestions. In addition to the general presentations of the project leaders within the program, it would have been interesting to see also technical presentations from researchers and PhD students from the other projects, at some joint workshop. It would also be very fine if there would be a dedicated meeting to discuss plans for how the critical mass of expertise that the funding from SSF has helped to build can be best exploited and ways of supporting the new PhDs in their future careers.

B.7 Your view of the project board or steering group and its role?

The project did not have any formal project board or steering group. The action plan was continuously updated, based on discussions at regular meetings with the entire project and also requirements and input from SSF, e.g., at the half-time evaluation. A steering group (with members from outside the project) would probably have been quite useful as support to the project leader, in particular to early on spot upcoming challenges and provide advice on how to tackle these.