Welcome!

Benelearn is the annual machine learning conference of the Benelux. It serves as a forum for researchers to exchange ideas, present recent work, and foster collaboration in the broad field of Machine Learning and its applications. It is our pleasure to welcome you to Eindhoven to partake in the 26th edition: welcome to Benelearn 2017!

Benelearn 2017 takes place largely on the campus of the Technische Universiteit Eindhoven, De Zaale, Eindhoven. The Friday programme is located in De Zwarte Doos (see https://goo.gl/maps/XgKE07JxyTC2), and the Saturday programme in Auditorium (see https://goo.gl/maps/B3PnpucJgMJ2). The conference dinner on Friday evening is the only off-campus event; this takes place in the DAF Museum, Tongelresestraat 27, 5613 DA Eindhoven (see https://goo.gl/maps/zNLrhpSqimk).

On campus, and also in parts of the city center, you can access the internet through the Eduroam network with your own university account. Conference proceedings can be downloaded from the conference website in PDF form, also directly accessible via the following link: http://wwwis.win.tue.nl/~benelearn2017/proceedings.pdf.

As part of the main conference programme, we organize three special tracks: one on Complex Networks, one on Deep Learning, and one Industry Track. Distributed over all tracks, contributing researchers not only span all three Benelux countries, but also include affiliations from ten additional countries. The updated three-continent spanning map of the newly expanded Benelux can be found on the front page of this booklet.

We hope you enjoy the conference!

Kind regards,
The Benelearn 2017 organizers.

Organization

Conference Chairs: Wouter Duivesteijn, Mykola Pechenizkiy
Complex Networks Track Chair: George Fletcher
Deep Learning Track Chairs: Vlado Menkovski, Eric Postma
Industry Track Chairs: Joaquin Vanschoren, Peter van der Putten
Local Organization: Riet van Buul
Contents

Programme at a Glance 4

Session Details 5

Abstracts 10
  Invited Talks ......................................................... 10
  Conference Track .................................................... 11
  Complex Networks Track .......................................... 20
  Deep Learning Track ................................................ 22
  Industry Track ........................................................ 24

History of Benelearn 26
# Programme at a Glance

## Friday, June 9 (location: Zwarte Doos)

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>09.00-09.30</td>
<td>Welcome with coffee and tea</td>
</tr>
<tr>
<td>09.30-09.45</td>
<td>Opening Benelearn 2017</td>
</tr>
<tr>
<td>09.45-10.30</td>
<td>Keynote Toon Calders — Data mining, social networks and ethical implications</td>
</tr>
<tr>
<td>10.30-11.15</td>
<td>Conference Track session 1</td>
</tr>
<tr>
<td>11.15-11.30</td>
<td>Coffee break</td>
</tr>
<tr>
<td>11.30-12.15</td>
<td>Conference Track session 2</td>
</tr>
<tr>
<td>12.15-12.20</td>
<td>Sponsor talk Itility</td>
</tr>
<tr>
<td>12.30-13.30</td>
<td>Lunch</td>
</tr>
<tr>
<td>13.30-14.15</td>
<td>Keynote Jean-Charles Delvenne — Dynamics and mining on large networks</td>
</tr>
<tr>
<td>14.15-15.15</td>
<td>Complex Networks Track session</td>
</tr>
<tr>
<td>15.15-15.30</td>
<td>Coffee break</td>
</tr>
<tr>
<td>15.30-16.45</td>
<td>Conference Track session 3</td>
</tr>
<tr>
<td>16.45-17.00</td>
<td>Break</td>
</tr>
<tr>
<td>17.00-18.00</td>
<td>Conference Track session 4</td>
</tr>
<tr>
<td>18.00-18.30</td>
<td>Walk to DAF Museum</td>
</tr>
<tr>
<td>18.30-22.00</td>
<td>Conference dinner</td>
</tr>
</tbody>
</table>

## Saturday, June 10 (location: Auditorium)

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>09.00-09.30</td>
<td>Welcome with coffee and tea</td>
</tr>
<tr>
<td>09.30-10.15</td>
<td>Keynote Holger Hoos — The transformative impact of automated algorithm design: ML, AutoML and beyond</td>
</tr>
<tr>
<td>10.15-11.00</td>
<td>Industry Track session</td>
</tr>
<tr>
<td>11.00-11.15</td>
<td>Coffee break</td>
</tr>
<tr>
<td>11.15-12.30</td>
<td>Poster spotlights</td>
</tr>
<tr>
<td>12.30-13.30</td>
<td>Lunch</td>
</tr>
<tr>
<td>13.30-14.15</td>
<td>Keynote Max Welling — Generalizing Convolutions for Deep Learning</td>
</tr>
<tr>
<td>14.15-15.30</td>
<td>Deep Learning Track session</td>
</tr>
<tr>
<td>15.30-17.00</td>
<td>Poster session</td>
</tr>
<tr>
<td>17.00-18.00</td>
<td>Drinks</td>
</tr>
</tbody>
</table>
Oral Presentation Sessions

Conference Track session 1 (Friday)

10.30-10.45 Marcia Fissette, Bernard Veldkamp and Theo de Vries: *Text mining to detect indications of fraud in annual reports worldwide*

10.45-11.00 Veronika Cheplygina, Lauge Sorensen, David Tax, Marleen De Bruijne and Marco Loog: *Do you trust your multiple instance learning classifier?*

11.00-11.15 Marco Cox and Bert de Vries: *A Gaussian process mixture prior for hearing loss modeling*

Conference Track session 2 (Friday)

11.30-11.45 Piotr Antonik, Marc Haelterman and Serge Massar: *Predicting chaotic time series using a photonic reservoir computer with output feedback*

11.45-12.00 Sibylle Hess, Katharina Morik and Nico Piatkowski: *PRIMPing Boolean Matrix Factorization through Proximal Alternating Linearized Minimization*

12.00-12.15 Niek Tax, Natalia Sidorova and Wil M.P. van der Aalst: *Local Process Models: Pattern Mining with Process Models*

Conference Track session 3 (Friday)

15.30-15.45 Christina Papagiannopoulou, Stijn Decubber, Willem Waegeman, Matthias Demuuzere, Niko Verhoest and Diego Miralles: *A non-linear Granger causality approach for understanding climate-vegetation dynamics*

15.45-16.00 Michiel Stock, Bernard De Baets and Willem Waegeman: *An Exact Iterative Algorithm for Transductive Pairwise Prediction*

16.00-16.15 Jacopo De Stefani, Gianluca Bontempi, Olivier Caelen and Dalila Hattab: *Multi-step-ahead prediction of volatility proxies*

16.15-16.30 Jesse Krijthe and Marco Loog: *Projected Estimators for Robust Semi-supervised Classification*

16.30-16.45 Björn Brodén, Mikael Hammar, Bengt J. Nilsson and Dimitris Paraschakis: *An ensemble recommender system for e-commerce based on Thompson Sampling*
Conference Track session 4 (Friday)

17.00-17.15 Sara Magliacane, Tom Claassen and Joris Mooij: *Ancestral Causal Inference*

17.15-17.30 Martin Atzmueller: *Exceptional Model Mining in Ubiquitous and Social Environments*

17.30-17.45 Martijn Van Otterlo and Martin Warnaar: *Towards Optimizing the Public Library: Indoor Localization in Semi-Open Spaces and Beyond*

17.45-18.00 Sebastijan Dumancic and Hendrik Blockeel: *An expressive similarity measure for relational clustering using neighbourhood trees*

Complex Networks Track session (Friday)

14.15-14.30 Dounia Mulders, Cyril de Bodt, Johannes Bjelland, Alex Pentland, Michel Verleysen and Yves-Alexandre de Montjoye: *Improving Individual Predictions using Social Networks Assortativity*

14.30-14.45 Xander Wilcke, Victor de Boer and Frank Van Harmelen: *User-Driven Pattern Mining on Knowledge Graphs: an Archaeological Case Study*

14.45-15.00 Leto Peel: *Graph-based semi-supervised learning for complex networks*

15.00-15.15 Martin Atzmueller, Lisa Thiele, Gerd Stumme and Simone Kauffeld: *Contact Patterns, Group Interaction and Dynamics on Socio-Behavioral Multiplex Networks*

Industry Track session (Saturday)

10.15-10.30 Lodewijk Nauta and Max Baak: *Eskapade: a lightweight, python based, analysis framework*

10.30-10.45 Dejan Radosavljevik and Peter van der Putten: *Service Revenue Forecasting in Telecommunications: A Data Science Approach*

10.45-11.00 Michiel van Wezel: *Predicting Termination of Housing Rental Agreements with Machine Learning*
Deep Learning Track session (Saturday)


14.30-14.45 Nanne van Noord: *Normalisation for painting colourisation*

14.45-15.00 Julia Berezutskaya, Zachary V. Freudenburg, Nick F. Ramsey, Umut Güçlü and Marcel A.J. van Gerven: *Modeling brain responses to perceived speech with LSTM networks*

15.00-15.15 Niek Tax, Ilya Verenich, Marcello La Rosa and Marlon Dumas: *Predictive Business Process Monitoring with LSTMs*

15.15-15.30 Stefan Thaler, Vlado Menkovski and Milan Petkovic: *Towards unsupervised signature extraction of forensic logs*

Poster Session (Saturday)

Conference Track

Thijs van de Laar and Bert de Vries: *A Probabilistic Modeling Approach to Hearing Loss Compensation*

L.F.J.M. Kanters: *Extracting relevant discussion from Reddit Science AMAs*

Anouk van Diepen, Marco Cox and Bert de Vries: *An In-situ Trainable Gesture Classifier*

Piotr Antonik, Marc Haelterman and Serge Massar: *Towards high-performance analogue readout layers for photonic reservoir computers*

Ivan Veul: *Locally versus Globally Trained Word Embeddings for Automatic Thesaurus Construction in the Legal Domain*

Dounia Mulders, Michel Verleysen and André Mouraux: *Characterizing Resting Brain Activity to Predict the Amplitude of Pain-Evoked Potentials in the Human Insula*

Quan Nguyen, Bert de Vries and Tjalling Tjalkens: *Probabilistic Inference-based Reinforcement Learning*

Veselka Boeva, Milena Angelova and Elena Tsiporkova: *Identifying Subject Experts through Clustering Analysis*

Rianne Conijn and Menno Van Zaanen: *Identifying writing tasks using sequences of keystrokes*
Towards an automated method based on Iterated Local Search optimization for tuning the parameters of Support Vector Machines

Increasing the Margin in Support Vector Machines through Hyperplane Folding

Generalization Bound Minimization for Active Learning

Towards an Ethical Recommendation Framework

Constraint-based measure for estimating overlap in clustering

Dynamics Based Features for Graph Classification

Harvesting the right tweets: Social media analytics for the Horticulture Industry

Unsupervised region of interest detection in sewer pipe images: Outlier detection and dimensionality reduction methods

Anomaly Analytics and Structural Assessment in Process Industries

Comparison of Syntactic Parsers on Biomedical Texts

The use of shallow convolutional neural networks in predicting promotor strength in Escherichia coli

Big IoT data mining for real-time energy disaggregation in buildings
UTILITY DATA SCIENCE TEAM

BIG DATA PLATFORM

- Dynamic Time Warping
- Mathematical Optimization
- Machine Learning
- Process Mining
- Predictive Statistics
- Text Mining Algorithms
- Anomaly Detection

OUR YOUNG PROFESSIONAL DATA SCIENCE PROGRAM

“6-MONTH PROGRAM FOR YOUNG TALENTED SCIENTISTS.” RECRUITMENT@UTILITY.NL

<table>
<thead>
<tr>
<th>Scrum</th>
<th>Splunk class</th>
<th>R class</th>
<th>Python class</th>
<th>Individual assignment (topic based on interest)</th>
<th>Customer assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>Sprint</td>
<td>Team challenge</td>
<td>Individual challenge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 2, 3, 4</td>
<td>Sprint</td>
<td>Team challenge</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 5, 6, 7</td>
<td>Sprint</td>
<td>Team challenge</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 8, 9, 10</td>
<td>Sprint</td>
<td>Team challenge</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 11, 12, 13</td>
<td>Sprint</td>
<td>Team challenge</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 14 - 28</td>
<td>Sprint</td>
<td>Team challenge</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

PERSONAL GUIDANCE AND COACHING
Abstracts Invited Talks

(ordered alphabetically on the last name of the first author, per track)

Toon Calders — Data mining, social networks and ethical implications. Recently we have seen a remarkable increase of awareness of the value of data. Whereas companies and governments mainly used to gather data about their clients just to support their operations, nowadays they are actively exploring new applications. For instance, a telecom operator may use call data not only to bill its customers, but also to derive social relations between its customers which may help to improve churn models, and governments use mobility data to chart mobility patterns that help to assess the impact of planned infrastructure works. I will give an overview of my research in this fascinating area, including pattern mining, the analysis of influence propagation in social networks, and ethical challenges such as models that discriminate.

Jean-Charles Delvenne — Dynamics and mining on large networks. A network, i.e. the data of nodes connected by edges, often comes as the support of dynamical interactions. For example a social network is often measured as the trace of an information flow (phone calls, messages), energy and phase information flow through power networks, biochemical networks are the skeleton of complex reaction systems, etc. It is therefore natural to mine network-shaped data jointly with a real or modelled dynamics taking place on it. In this talk we review how dynamics can provide efficient and accurate methods for community detection, classification, centrality and assortativity measures.

Holger Hoos — The transformative impact of automated algorithm design: ML, AutoML and beyond. Techniques from artificial intelligence — and especially, machine learning — are fundamentally changing the way we solve challenging computational problems, and recently, automated machine learning (AutoML) has begun to take this to a new level. In this talk, I will share my perspective on the success of ML and AutoML, and discuss how the fundamental concepts and tools that enable both have a much broader impact than commonly perceived. In particular, I will highlight the role of a fruitful interplay between machine learning and optimisation in this context, comment on general approaches to automated algorithm design, and share my thoughts on the next big challenge.

Max Welling — Generalizing Convolutions for Deep Learning. Arguably, most excitement about deep learning revolves around the performance of convolutional neural networks and their ability to automatically extract useful features from signals. In this talk I will present work from AMLAB where we generalize these convolutions. First we
study convolutions on graphs and propose a simple new method to learn embeddings of graphs which are subsequently used for semi-supervised learning and link prediction. We discuss applications to recommender systems and knowledge graphs. Second we propose a new type of convolution on regular grids based on group transformations. This generalizes normal convolutions based on translations to larger groups including the rotation group. Both methods often result in significant improvements relative to the current state of the art.

Joint work with Thomas Kipf, Rianne van den Berg and Taco Cohen.

Abstracts Conference Track

Antoine Adam, Hendrik Blockeel — Constraint-based measure for estimating overlap in clustering. Different clustering algorithms have different strengths and weaknesses. Given a dataset and a clustering task, it is up to the user to choose the most suitable clustering algorithm. In this paper, we study to what extent this choice can be supported by a measure of overlap among clusters. We propose a concrete, efficiently computable constraint-based measure. We show that the measure is indeed informative: on the basis of this measure alone, one can make better decisions about which clustering algorithm to use. However, when combined with other features of the input dataset, such as dimensionality, it seems that the proposed measure does not provide useful additional information.

Piotr Antonik, Marc Haelterman, Serge Massar — Predicting chaotic time series using a photonic reservoir computer with output feedback. Reservoir Computing is a bio-inspired computing paradigm for processing time dependent signals (Jaeger & Haas, 2004; Maass et al., 2002). The performance of its hardware implementations matches digital algorithms on a series of benchmark tasks (see e.g. (Soriano et al., 2015) for a review). Their capacities could be extended by feeding the output signal back into the reservoir, which would allow them to be applied to various signal generation tasks (Antonik et al., 2016b). In practice, this requires a high-speed readout layer for real-time output computation. Here we achieve this by means of a field-programmable gate array (FPGA), and demonstrate the first photonic reservoir computer with output feedback. We test our setup on the Mackey-Glass chaotic time series generation task and obtain interesting prediction horizons, comparable to numerical simulations, with ample room for further improvement. Our work thus demonstrates the potential offered by the output feedback and opens a new area of novel applications for photonic reservoir computing. A more detailed description of this work can be found in (Antonik et al., 2017a; Antonik et al., 2017b).
Piotr Antonik, Marc Haelterman, Serge Massar — Towards high-performance analogue readout layers for photonic reservoir computers. Reservoir Computing is a bio-inspired computing paradigm for processing time-dependent signals (Jaeger & Haas, 2004; Maass et al., 2002). The performance of its hardware implementations (see e.g. (Soriano et al., 2015) for a review) is comparable to state-of-the-art digital algorithms on a series of benchmark tasks. The major bottleneck of these implementation is the readout layer, based on slow offline post-processing. Several analogue solutions have been proposed (Smerieri et al., 2012; Duport et al., 2016; Vinckier et al., 2016), but all suffered from noticeable decrease in performance due to added complexity of the setup. Here we propose the online learning approach to solve these issues. We present an experimental reservoir computer with a simple analogue readout layer, based on previous works, and show numerically that online learning allows to disregard the added complexity of an analogue layer and obtain the same level of performance as with a digital layer. This work thus demonstrates that online training allows building high-performance fully-analogue reservoir computers, and represents an important step towards experimental validation of the proposed solution. A more detailed description of this work can be found in (Antonik et al., 2017a; Antonik et al., 2017b).

Martin Atzmueller — Exceptional Model Mining in Ubiquitous and Social Environments. Exceptional model mining in ubiquitous and social environments includes the analysis of resources created by humans (e.g., social media) as well as those generated by sensor devices in the context of (complex) interactions. This paper provides a structured overview on a line of work comprising a set of papers that focus on local exceptionality detection in ubiquitous and social environments and according complex social interaction networks.

Veselka Boeva, Milena Angelova, Elena Tsiporkova — Identifying Subject Experts through Clustering Analysis. In this work, we discuss an approach for identifying subject experts via clustering analysis of the available online information. Initially, the domain of interest is partitioned to a number of subject areas. Next each extracted expert is represented by a vector of contributions of the expert to the different areas. Finally, the set of all extracted experts is grouped into a set of disjoint expert areas by applying Formal Concept Analysis (FCA). The produced grouping is further shown to facilitate the location of experts with the required competence.

Björn Brodén, Mikael Hammar, Bengt J. Nilsson, Dimitris Paraschakis — An Ensemble Recommender System for e-Commerce. In our ongoing work we extend the Thompson Sampling (TS) bandit policy for orchestrating the collection of base recommendation algorithms for e-Commerce. We focus on the problem of item-to-item recommendations, for which multiple behavioral and content-based predictors are provided to an ensemble learner.
The extended TS-based policy must be able to handle situations when bandit arms are non-stationary and non-answering. Furthermore, we investigate the effects of priming the sampler with pre-set parameters of reward distributions by analyzing the product catalog and/or event history, when such information is available. We report our preliminary results based on the analysis of two real-world e-Commerce datasets.

Veronika Cheplygina, Lauge Sørensen, David M.J. Tax, Marleen de Bruijne, Marco Loog — Do you trust your multiple instance learning classifier?. Multiple instance learning (MIL) is a weakly supervised learning scenario where labels are given only for groups (bags) of examples (instances). Because some MIL classifiers can provide instance labels at test time, MIL is popular in applications where labels are difficult to acquire. However, MIL classifiers are frequently only evaluated on their bag-level, not instance-level performance. In this extended abstract, which covers previously published work, we demonstrate why this could be problematic and discuss this open problem.

Rianne Conijn, Menno van Zaanen — Identifying writing tasks using sequences of keystrokes. The sequences of keystrokes that are generated when writing texts contain information about the writer as well as the writing task and cognitive aspects of the writing process. Much research has been conducted in the area of writer identification. However, research on the analysis of writing processes based on sequences of keystrokes has received only a limited amount of attention. Therefore, in this study we try to identify properties of keystrokes that indicate cognitive load of the writing process. Moreover, we investigate the influence of these properties on the classification of texts written during two different writing tasks: copying a text and free-form generation of text. We show that we can identify properties that allow for the correct classification of writing tasks, which at the same time do not describe writer-specific characteristics. However, some properties are the result of an interaction between the typing characteristics of the writer and the writing task.

Sergio Consoli, Jacek Kustra, Pieter Vos, Monique Hendriks, Dimitrios Mavroeidis — Towards an automated method based on Iterated Local Search optimization for tuning the parameters of Support Vector Machines. We provide preliminary details and formulation of an optimization strategy under current development that is able to automatically tune the parameters of a Support Vector Machine over new datasets. The optimization strategy is a heuristic based on Iterated Local Search, a modification of classic hill climbing which iterates calls to a local search routine.

Marco Cox, Bert de Vries — A Gaussian process mixture prior for hearing loss modeling. In this abstract we introduce a prior distribution for hearing thresholds learned from a large database containing the hearing thresholds, ages and genders of around 85,000
people. Almost all existing work is based on very simple and/or uninformative GP priors; simply selecting a suitable type of kernel that assumes the threshold curve to be smooth is already sufficient to yield a well working system. However, by fitting a slightly more complex model to a vast database of measured thresholds, we obtain a prior that is more informative and empirically justified.

Jacopo De Stefani, Gianluca Bontempi, Olivier Caelen, Dalila Hattab — Multi-step-ahead prediction of volatility proxies. Though machine learning techniques have been often used for stock prices forecasting, few results are available for market fluctuation prediction. Nevertheless, volatility forecasting is an essential tool for any trader wishing to assess the risk of a financial investment. The main challenge of volatility forecasting is that, since this quantity is not directly observable, we cannot predict its actual value but we have to rely on some observers, known as volatility proxies (Poon & Granger, 2003) based either on intraday (Martens, 2002) or daily data. Once a proxy is chosen, the standard approach to volatility forecasting is the well-known GARCH-like model (Andersen & Bollerslev, 1998). In recent years several hybrid approaches are emerging (Kristjanpoller et al., 2014; Dash & Dash, 2016; Monfared & Enke, 2014) which combine GARCH with a non-linear computational approach. What is common to the state-of-the art is that volatility forecasting is addressed as an univariate and one-step-ahead auto-regressive (AR) time series problem. The purpose of our work is twofold. First, we aim to perform a statistical assessment of the relationships among the most used proxies in the volatility literature. Second, we explore a NARX (Nonlinear Autoregressive with eXogenous input) approach to estimate multiple steps of the output given the past output and input measurements, where the output and the input are two different proxies. In particular, our preliminary results show that the statistical dependencies between proxies can be used to improve the forecasting accuracy.

Anouk van Diepen, Marco Cox, Bert de Vries — An In-situ Trainable Gesture Classifier. Gesture recognition, i.e., the recognition of pre-defined gestures by arm or hand movements, enables a natural extension of the way we currently interact with devices (Horsley, 2016). Commercially available gesture recognition systems are usually pre-trained: the developers specify a set of gestures, and the user is provided with an algorithm that can recognize just these gestures. To improve the user experience, it is often desirable to allow users to define their own gestures. In that case, the user needs to train the recognition system herself by a set of example gestures. Crucially, this scenario requires learning gestures from just a few training examples in order to avoid overburdening the user. We present a new in-situ trainable gesture classifier based on a hierarchical probabilistic modeling approach. Casting both learning and recognition as probabilistic inference tasks yields a principled way to design and evaluate algorithm candidates. Moreover, the Bayesian approach facilitates learning of prior knowledge about gestures, which leads to fewer needed examples for training new gestures.
Sebastijan Dumančić, Hendrik Blockeel — An expressive similarity measure for relational clustering using neighbourhood trees. In this paper, we introduce a novel similarity measure for relational data. It is the first measure to incorporate a wide variety of types of similarity, including similarity of attributes, similarity of relational context, and proximity in a hypergraph. We experimentally evaluate how using this similarity affects the quality of clustering on very different types of datasets. The experiments demonstrate that (a) using this similarity in standard clustering methods consistently gives good results, whereas other measures work well only on datasets that match their bias; and (b) on most datasets, the novel similarity outperforms even the best among the existing ones.

Marcia Fissette, Bernard Veldkamp, Theo de Vries — Text mining to detect indications of fraud in annual reports worldwide. Fraud affects the financial results presented in the annual reports of companies worldwide. Analysis performed on annual reports focuses on the quantitative data in these reports. However, the amount of textual information in annual reports increased in the past decade with companies using the reports to project themselves. The texts provide information that is complementary to the financial results. Therefore, the analysis of the textual information in annual reports may provide indications of the presence of fraud within a company. This piece of research uses the extensive and reality approaching data set containing annual reports of companies worldwide to answer the research question: Can a text mining model be developed that can detect indications of fraud in the management discussion and analysis section of annual reports of companies worldwide?

Sibylle Hess, Katharina Morik, Nico Piatkowski — PRIMPing Boolean Matrix Factorization through Proximal Alternating Linearized Minimization. We propose a novel Boolean matrix factorization algorithm to solve the tiling problem, based on recent results from optimization theory. We demonstrate the superior robustness of the new approach in the presence of several kinds of noise and types of underlying structure. Experimental results on image data show that the new method identifies interpretable patterns which explain the data almost always better than the competing algorithms.

L.F.J.M. Kanters — Extracting relevant discussion from Reddit Science AMAs. The social network and content aggregation website Reddit occasionally hosts Q&A sessions with scientists called science AMA (Ask Me Anything). These science AMAs are conducted through the comment system of Reddit which has a tree structure, mark-up and community driven feedback on both users and comments in the form of “karma” scores. Most of the actual discussion in these science AMAs tends to be of high quality. However a large number of the comments are superfluous and not really part of the conversation with the scientist. The goal of this project is to determine if text mining
methods can be used to filter out the unwanted comments. A secondary goal is to determine the relative importance of Reddit meta-data (tree structure, karma scores, etc) compared to the actual content of the comments. The python Reddit API was used to retrieve the AMAs. The CoreNLP tools were used to extract tokens, sentences, named entities and sentiment. These were combined with other information, like Reddit meta-data and WordNet, and used to extract features. The classification was done by a Gaussing naive Bayes classifier using the scikit-learn toolbox. Classification using all features or only text-based features was effective both yielding a precision/recall/f1-score of 0.84/0.99/0.91. Only using Reddit based features was slightly less effective, yielding 0.89/0.63/0.74. Only using a single WordNet based similarity feature still worked, yielding 0.81/0.99/0.89.

Jesse H. Krijthe, Marco Loog — Projected Estimators for Robust Semi-supervised Classification. For semi-supervised techniques to be applied safely in practice we at least want methods to outperform their supervised counterparts. We study this question for classification using the well-known quadratic surrogate loss function. Unlike other approaches to semisupervised learning, the procedure proposed in this work does not rely on assumptions that are not intrinsic to the classifier at hand. Using a projection of the supervised estimate onto a set of constraints imposed by the unlabeled data, we find that it is possible to safely improve over the supervised solution in terms of this quadratic loss.

Thijs van de Laar, Bert de Vries — A Probabilistic Modeling Approach to Hearing Loss Compensation. This short paper summarizes ongoing work on a probabilistic modeling approach to the design of personalized hearing aid algorithms (van de Laar & de Vries, 2016). In this framework, we first specify a probabilistic generative model that includes an explicit description of the hearing loss problem. Given the model, hearing aid signal processing relates to on-line Bayesian state estimation (similar to Kalman filtering). Estimation of the tuning parameters (known as the ‘fitting’ task in hearing aid parlance) corresponds to Bayesian parameter estimation. The innovative aspect of the framework is that both the signal processing and fitting tasks can be automatically inferred from the probabilistic model in conjunction with patient appraisals (the data).

Lars Lundberg, Håkan Lennerstad, Eva Garcia-Martin, Niklas Lavesson, Veselka Boeva — Increasing the Margin in Support Vector Machines through Hyperplane Folding. We present a method, called hyperplane folding, that increases the margin in a linearly separable binary dataset by replacing the SVM hyperplane with a set of hinging hyperplanes. Based on the location of the support vectors, the method splits the dataset into two parts, rotates one part of the dataset and then merges the two parts again. This procedure increases the margin in each iteration as long as the margin is smaller than half of the shortest distance between any pair of data points from the two different
We provide an algorithm for the general case with \( n \)-dimensional data points. A small experiment with three folding iterations on 2-dimensional data points shows that the margin does indeed increase and that the accuracy improves with a larger margin, i.e., the number of misclassified data points decreases when we use hyperplane folding. The method can use any standard SVM implementation plus some additional basic manipulation of the data points, i.e., splitting, rotating and merging.

**Sara Magliacane, Tom Claassen, Joris M. Mooij — Ancestral Causal Inference.** Discovering causal relations from data is at the foundation of the scientific method. Traditionally, cause-effect relations have been recovered from experimental data in which the variable of interest is perturbed, but seminal work like the do-calculus (Pearl, 2009) and the PC/FCI algorithms (Spirtes et al., 2000; Zhang, 2008) demonstrate that, under certain assumptions, it is already possible to obtain significant causal information by using only observational data. We propose Ancestral Causal Inference (ACI), a logic-based method that provides a comparable accuracy to the best state-of-the-art constraint-based methods as HEJ (Hyttinen et al., 2014), but improves on the scalability by using a more coarse-grained representation.

**Dounia Mulders, Michel Verleysen, Giulia Liberati, André Mouraux — Characterizing Resting Brain Activity to Predict the Amplitude of Pain-Evoked Potentials in the Human Insula.** How the perception of pain emerges from human brain activity remains largely unknown. Apart from inter-individual variations, this perception depends not only on the physical characteristics of the painful stimuli, but also on other psycho-physiological aspects. Indeed a painful stimulus applied to an individual can sometimes evoke very distinct sensations from one trial to the other. Hence the state of a subject receiving such a stimulus should (at least partly) explain the intensity of pain elicited by that stimulus. Using intracranial electroencephalography (iEEG) from the insula to measure this cortical "state", our goal is to study to which extent ongoing brain activity in the human insula, an area thought to play a key role in pain perception, may predict the magnitude of pain-evoked potentials and, more importantly, whether it may predict the perception intensity. To this aim, we summarize the ongoing insular activity by defining frequency-dependent features, derived using continuous wavelet and Fourier transforms. We then take advantage of this description to predict the amplitude of the insular responses elicited by painful (heat) and non-painful (auditory, visual and vibrotactile) stimuli, as well as to predict the intensity of perception.

**Quan Nguyen, Bert de Vries, Tjalling J. Tjalkens — Probabilistic Inference-based Reinforcement Learning.** We introduce probabilistic inference-based reinforcement learning (PIReL), an approach to solve decision making problems by treating them as probabilistic inference tasks. Unlike classical reinforcement learning, which requires explicit reward
functions, in PIReL they are implied by probabilistic assumptions of the model. This would enable a fundamental way to design the reward function by model selection as well as bring the potential to apply existing probabilistic modeling techniques to reinforcement learning problems.

Martijn van Otterlo, Martin Warnaar — Towards Optimizing the Public Library: Indoor Localization in Semi-Open Spaces and Beyond. We report on the BLIIPS project which aims at the digitalization and optimization of physical, public libraries through the use of artificial intelligence combined with sensor technology. As a first step we introduce FLib, a localization application, and additional developments for interaction with physical books. The contributions of this paper are the introduction of the public library as an interesting testbed for smart technologies, a novel localization application with an experimental evaluation, and a compact research agenda for smart libraries.

Christina Papagiannopoulou, Stijn Decubber, Willem Waegeman, Matthias Demuzere, Niko E.C. Verhoest, Diego G. Miralles — A non-linear Granger causality approach for understanding climate-vegetation dynamics. Satellite Earth observation provides new means to unravel the drivers of long-term changes in climate. Global historical records of crucial environmental and climatic variables, which have the form of multivariate time series, now span up to 30 years. In this abstract we present a non-linear Granger causality approach to detect causal relationships between climatic time series and vegetation. Our framework consists of several components, including data fusion from various databases, time series decomposition techniques, feature construction methods and Granger causality analysis by means of machine learning algorithms. Experimental results on large-scale entire-globe datasets indicate that, with this framework, it is possible to detect non-linear patterns that express the complex relationships between climate and vegetation.

Dimitris Paraschakis — Towards an Ethical Recommendation Framework. This study provides an overview of various ethical challenges that complicate the design of recommender systems (RS). The articulated ethical recommendation framework maps RS design stages to the corresponding ethical concerns, and further down to known solutions and the proposed user-adjustable controls. This framework aims to aid RS practitioners in staying ethically alert while taking morally charged design decisions. At the same time, it would give users the desired control over the sensitive moral aspects of recommendations via the proposed “ethical toolbox”. The idea is embraced by the participants of our feasibility study.
Michael Stock, Bernard De Baets, Willem Waegeman — An Exact Iterative Algorithm for Transductive Pairwise Prediction. Imputing missing values of a matrix when side-features are available can be seen as a special case of pairwise learning. In this extended abstract we present an exact iterative algorithm to impute these missing values efficiently.

Niek Tax, Natalia Sidorova, Wil M.P. van der Aalst — Local Process Models: Pattern Mining with Process Models. Process mining aims to extract novel insights from event data (van der Aalst, 2016). Process discovery plays a prominent role in process mining. The goal is to discover a process model that is representative for the set of event sequences in terms of start-to-end behavior, i.e. from the start of a case till its termination. Many process discovery algorithms have been proposed and applied to a variety of real life cases. A more conventional perspective on discovering insights from event sequences can be found in the areas of sequential pattern mining (Agrawal & Srikant, 1995) and episode mining (Mannila et al., 1997), which focus on finding frequent patterns, not aiming for descriptions of the full event sequences from start to end. Sequential pattern mining is limited to the discovery of sequential orderings of events, while process discovery methods aim to discover a larger set of event relations, including sequential orderings, (exclusive) choice relations, concurrency, and loops, represented in process models such as Petri nets (Reisig, 2012), BPMN (Object Management Group, 2011), or UML activity diagrams. Process models distinguish themselves from more traditional sequence mining approaches like Hidden Markov Models (Rabiner, 1989) and Recurrent Neural Networks with their visual representation, which allows them to be used for communication between process stakeholders. However, process discovery is normally limited to the discovery of a complete model that captures the full behavior of process instances, and not local patterns within instances. Local Process Models (LPMs) allow the mining of patterns positioned in-between simple patterns (e.g. subsequences) and end-to-end models, focusing on a subset of the process activities and describing frequent patterns of behavior.

I.G. Veul — Locally versus Globally Trained Word Embeddings for Automatic Thesaurus Construction in the Legal Domain. In this paper two different word embedding methods are explored for the automatic construction of a thesaurus for legal texts. A word embedding maps every word to a relatively low dimensional vector, which is then used to compare similarities between words. We use Word2Vec for the word embedding, which is an unsupervised learning method that defines a word based on its context. Words with similar contexts will then be considered similar. The unsupervised nature of Word2Vec allows for the construction of the thesaurus without requiring relevance feedback. A downside with the standard Word2Vec approach though is that the resulting word embeddings tend to be too general, when trained on an entire corpus. This paper studies whether training the word embeddings separately for different jurisdictions results in a better thesaurus. The thesauri are trained on the text of 300,000 Dutch
legal rulings. To assess the performance of the globally and locally trained thesauri, they are compared to a manually constructed thesaurus, which is already being used for query expansion in the legal domain. The results show that there is a significant difference between the global and local thesauri, but that the global thesaurus actually outperforms the local thesaurus.

Tom Viering, Jesse Krijthe, Marco Loog — Generalization Bound Minimization for Active Learning. We studied the state-of-the-art Maximum Mean Discrepancy (MMD) active learner that minimizes a generalization bound (Chattopadhyay et al., 2012; Wang & Ye, 2013). The MMD is a divergence measure (Gretton et al., 2012) which is closely related to the Discrepancy measure (Mansour et al., 2009). Our study illustrates that tighter bounds do not guarantee improved active learning performance and that a probabilistic analysis is essential: active learners should optimize their strategy for scenarios that are likely to occur in order to perform well in practice.

Abstracts Complex Networks Track

Martin Atzmueller, Lisa Thiele, Gerd Stumme, Simone Kauffeld — Contact Patterns, Group Interaction and Dynamics on Socio-Behavioral Multiplex Networks. The analysis of social interaction networks is essential for understanding and modeling network structures as well as the behavior of the involved actors. This paper summarizes an analysis at large scale using (sensor) data collected by RFID tags complemented by self-report data obtained using surveys. We focus on the social network of a students' freshman week, and investigate research questions concerning group behavior and structure, gender homophily, and interrelations of sensor-based (RFID) and self-report social networks. Such analyses are a first step for enhancing interactions and enabling proactive guidance.

Leonardo Gutiérrez Gómez, Jean-Charles Delvenne — Dynamics Based Features for Graph Classification. In this paper we propose a new feature based approach to network classification. We show how a dynamics on a network can be useful to reveal patterns about the organization of the components of the underlying graph where the process takes place. Measuring the autocovariance along a random path on the network of a suitable set of network attributes including node labels, allows us to define generalized features across different time scales. These dynamic features turn out to be an appropriate discriminative signature of the network suitable for classification and recognition purposes. The method is tested empirically on established network benchmarks. Results show that our dynamic-based features are competitive and often outperform state of the art graph kernel based methods.
Dounia Mulders, Cyril de Bodt, Michel Verleysen, Johannes Bjelland, Alex Pentland, Yves-Alexandre de Montjoye — Improving Individual Predictions using Social Networks Assortativity.

Social networks are known to be assortative with respect to many attributes such as age, weight, wealth, ethnicity and gender. Independently of its origin, this assortativity gives us information about each node given its neighbors. It can thus be used to improve individual predictions in many situations, when data are missing or inaccurate. This work presents a general framework based on probabilistic graphical models to exploit social network structures for improving individual predictions of node attributes. We quantify the assortativity range leading to an accuracy gain. We also show how specific characteristics of the network can improve performances further. For instance, the gender assortativity in mobile phone data changes significantly according to some communication attributes.

Leto Peel — Graph-based semi-supervised learning for complex networks. We address the problem of semi-supervised learning in relational networks, networks in which nodes are entities and links are the relationships or interactions between them. Typically this problem is confounded with the problem of graph-based semi-supervised learning (GSSL), because both problems represent the data as a graph and predict the missing class labels of nodes. However, not all graphs are created equally. In GSSL a graph is constructed, often from independent data, based on similarity. As such, edges tend to connect instances with the same class label. Relational networks, however, can be more heterogeneous and edges do not always indicate similarity. In this work (Peel, 2017) we present two scalable approaches for graph-based semi-supervised learning for the more general case of relational networks. We demonstrate these approaches on synthetic and real-world networks that display different link patterns within and between classes. Compared to state-of-the-art baseline approaches, ours give better classification performance and do so without prior knowledge of how classes interact.

Marijn ten Thij, Sandjai Bhulai — Harvesting the right tweets: Social media analytics for the Horticulture Industry. In our current society, data has gone from scarce to superabundant: huge volumes of data are being generated every second. A big part of this owes is due to social media platforms, which provide a very volatile flow of information. Leveraging this information, which is buried in this fast stream of messages, poses a serious challenge. A vast amount of work is devoted to tackle this challenge in different business areas. In our work, we address this challenge for the horticulture sector, which has not received a lot of attention in the literature. Our aim is to extract information from the social data flow that can empower the horticulture sector. In this abstract, we present our first steps towards this goal.
W.X. Wilcke, V. de Boer, F.A.H. van Harmelen — User-Driven Pattern Mining on knowledge graphs: an Archaeological Case Study. In this work, we investigate to what extent data mining can contribute to the understanding of archaeological knowledge, published as knowledge graph, and which form would best meet the communities’ needs. A case study was held which involved the user-driven mining of generalized association rules. Experiments have shown that the approach yielded mostly plausible patterns, some of which were rated as highly relevant by domain experts.

Abstracts Deep Learning Track

Julia Berezutskaya, Zachary V. Freudenburg, Nick F. Ramsey, Umut Güçlü, Marcel A.J. van Gerven — Modeling brain responses to perceived speech with LSTM networks. We used recurrent neural networks with long-short term memory units (LSTM) to model the brain responses to speech based on the speech audio features. We compared the performance of the LSTM models to the performance of the linear ridge regression model and found the LSTM models to be more robust for predicting brain responses across different feature sets.

Jim Clauwaert, Michiel Stock, Marjan De Mey, Willem Waegeman — The use of shallow convolutional neural networks in predicting promoter strength in Escherichia coli. Gene expression is an important factor in many processes of synthetic biology. The use of well-characterized promoter libraries makes it possible to obtain reliable estimates on the transcription rates in genetic circuits. Yet, the relation between promoter sequence and transcription rate is largely undiscovered. Through the use of shallow convolutional neural networks, we were able to create models with good predictive power for promoter strength in E. coli.

Decebal Constantin Mocanu, Elena Mocanu, Phuong H. Nguyen, Madeleine Gibescu, Antonio Liotta — Big IoT data mining for real-time energy disaggregation in buildings. In the smart grid context, the identification and prediction of building energy flexibility is a challenging open question. In this paper, we propose a hybrid approach to address this problem. It combines sparse smart meters with deep learning methods, e.g. Factored Four-Way Conditional Restricted Boltzmann Machines (FFW-CRBMs), to accurately predict and identify the energy flexibility of buildings unequipped with smart meters, starting from their aggregated energy values. The proposed approach was validated on a real database, namely the Reference Energy Disaggregation Dataset.
Nanne van Noord — Normalisation for painting colourisation. In this work we use a painting colourisation model capable of producing visually appealing colourisations to compare three normalisation techniques. We conclude that using an instance-based normalisation techniques is beneficial for painting colourisation and that conditioning the shifting and scaling parameters on the painter only leads to minimal improvements.

Niek Tax, Ilya Verenich, Marcello La Rosa, Marlon Dumas — Predictive Business Process Monitoring with LSTMs. The foremost contribution of this paper is a technique to predict the next activity of a running case and its timestamp using LSTM neural networks. We show that this technique outperforms existing baselines on real-life data sets. Additionally, we find that predicting the next activity and its timestamp via a single model (multi-task learning) yields a higher accuracy than predicting them using separate models. We then show that this basic technique can be generalized to address two other predictive process monitoring problems: predicting the entire continuation of a running case and predicting the remaining cycle time.

Stefan Thaler, Vlado Menkovski, Milan Petković — Towards unsupervised signature extraction of forensic logs. Log signature extraction is the process of finding a set of templates generated a set of log messages from the given log messages. This process is an important pre-processing step for log analysis in the context of information forensics because it enables the analysis of event sequences of the examined logs. In earlier work, we have shown that it is possible to extract signatures using recurrent neural networks (RNN) in a supervised manner (Thaler et al., 2017). Given enough labeled data, this supervised approach works well, but obtaining such labeled data is labor intensive. In this paper, we present an approach to address the signature extraction problem in an unsupervised way. We use an RNN autoencoder to create an embedding for the log lines and we apply clustering in the embedded space to obtain the signatures. We experimentally demonstrate on a forensic log that we can assign log lines to their signature cluster with a V-Measure of 0.94 and a Silhouette score of 0.75.

Jakub M. Tomczak, Max Welling — Improving Variational Auto-Encoders using convex combination linear Inverse Autoregressive Flow. In this paper, we propose a new volume-preserving flow and show that it performs similarly to the linear general normalizing flow. The idea is to enrich a linear Inverse Autoregressive Flow by introducing multiple lower-triangular matrices with ones on the diagonal and combining them using a convex combination. In the experimental studies on MNIST and Histopathology data we show that the proposed approach outperforms other volume-preserving flows and is competitive with current state-of-the-art linear normalizing flow.
Abstracts Industry Track

Martin Atzmueller, David Arnu, Andreas Schmidt — Anomaly Analytics and Structural Assessment in Process Industries. Detecting anomalous behavior can be of critical importance in an industrial application context: While modern production sites feature sophisticated alarm management systems, they mostly react to single events. In the context of process industries and heterogeneous data sources, we model sequential alarm data for anomaly detection and analysis, based on first-order Markov chain models. We outline hypothesis-driven and description-oriented modeling and provide an interactive dashboard for exploration and visualization.

Maria Biryukov — Comparison of Syntactic Parsers on Biomedical Texts. Syntactic parsing is an important step in the automated text analysis which aims at information extraction. Quality of the syntactic parsing determines to a large extent the recall and precision of the text mining results. In this paper we evaluate the performance of several popular syntactic parsers in application to the biomedical text mining.

Dirk Meijer, Arno Knobbe — Unsupervised region of interest detection in sewer pipe images: Outlier detection and dimensionality reduction methods. Sewer pipes require regular inspection to determine the deterioration state and performance, before deciding whether repair or replacement is necessary. Inspections are still mostly performed manually, which leads to subjective and inconsistent ratings for deterioration and urgency, differing between inspectors and even within inspectors (Dirksen et al., 2013). The SewerSense project aims to investigate the possibilities of automating the analysis of sensory data from sewer pipe inspections and the consequences this would have for the sewer asset management industry.

Lodewijk Nauta, Max Baak — Eskapade: a lightweight, python based, analysis framework. Eskapade is a python framework that accelerates development of advanced analytics workflows in big data environments. The modular set-up allows scalable designs of analysis chains based on commonly available open-source libraries and custom built algorithms using single configuration files with simple syntax: from data ingestion and transformations to ML models including feedback processing.
Dejan Radosavljevik, Peter van der Putten — Service Revenue Forecasting in Telecommunications: A Data Science Approach. This paper discusses a real-world case of revenue forecasting in telecommunications. Apart from the method we developed, we will describe the implementation, which is the key factor of success for data science solutions in a business environment. We will also describe some of the challenges that occurred and our solutions to them. Furthermore, we will explain our unorthodox choice for the error measure. Last, but not least we will present the results of this process.

Michiel van Wezel — Predicting Termination of Housing Rental Agreements with Machine Learning. Terminations of rental agreements (tenancy endings) are important in the business processes of housing associations. I describe the results of a model that predicts tenancy endings using Machine Learning.
History of Benelearn

This booklet concerns the 26th annual machine learning conference of the Benelux. This page contains some data about all editions; see if you can spot the pattern.

First: Leuven, Belgium, 1990
Second: Amsterdam, the Netherlands, 1991
Third: Brussel, Belgium, 1993
Fourth: Rotterdam, the Netherlands, 1994
Fifth: Brussel, Belgium, 1995
Sixth: Maastricht, the Netherlands, 1996
Seventh: Tilburg, the Netherlands, 1997
Eighth: Wageningen, the Netherlands, 1998
Ninth: Maastricht, the Netherlands, 1999 (organized by Leuven)
Tenth: Tilburg, the Netherlands, 2000
Eleventh: Antwerpen, Belgium, 2001
Twelfth: Utrecht, the Netherlands, 2002
Thirteenth: Brussel, Belgium, 2004
Fourteenth: Enschede, the Netherlands, 2005
Fifteenth: Gent, Belgium, 2006
Sixteenth: Amsterdam, the Netherlands, 2007
Seventeenth: Spa, Belgium, 2008
Eighteenth: Tilburg, the Netherlands, 2009
Nineteenth: Leuven, Belgium, 2010
Twentieth: Den Haag, the Netherlands, 2011
Twenty-first: Gent, Belgium, 2012
Twenty-second: Nijmegen, the Netherlands, 2013
Twenty-third: Brussel, Belgium, 2014
Twenty-fourth: Delft, the Netherlands, 2015
Twenty-fifth: Leuven, Belgium, 2016
Twenty-sixth: Eindhoven, the Netherlands, 2017
Room for Creativity. Maybe Write a Haiku or a Sonnet...
Many Thanks to our Sponsors!