Grascomp Doctoral Day, Namur
Faculté d'Informatique, UNamur, rue Grandgagnage 21, 5000 Namur
List of Papers with Abstracts

Session 1 - Open Source and Open Data	
Damien Legay, Alexandre Decan and Tom Mens *Towards understanding the relation between badges and contributions in GitHub repositories*	Continuously attracting contributors is key to the health of open source software projects. The appearance of badges in online collaborative development platforms affords maintainers the opportunity to advertise the quality of their project to potential contributors.We speculate that contributors rely on those badges to evaluate which projects they should contribute to. In this preliminary research, we measure how prevalent badges are in open source projects, which badges are used, when and how they are introduced, and which combinations of badges co-occur.We find that the most widespread badges convey static information or relay information about the build status of a project. Those badges are typically added early in projects and prior to or at the same time as other badges.
Mehdi Golzadeh, Alexandre Decan and Tom Mens *On the effect of discussions on pull request decisions*	Open-source software (OSS) relies on contributions from different types of contributors. Online collaborative development platforms, such as GitHub, usually provide explicit support for these contributions through the mechanism of pull requests, allowing project members and external contributors to discuss and evaluate the submitted code. These discussions can play an important role in the decision-making process leading to the acceptance or rejection of a pull request. We empirically examine in this paper 183K pull requests and their discussions, for almost 4.8K GitHub repositories for the Cargo ecosystem. We investigate the prevalence of such discussions, their participants and their size in terms of messages and durations, and we show how these aspects relate to the acceptance and rejection of pull requests.
Rabeb Abida, Emna Hachicha Belghith and Anthony Cleve *Supporting Semantics-aware Analysis in Linked Open Data-enabled Information Systems*	In recent years, open data is becoming an important trend, especially in government data context (i.e., Open Government Data), since it offers citizens and public institutions a transparent, free and customized access to such information as well as more public services and commercial re-use [1, 2]. Open Government Data (OGD) may contain multiple datasets, including budget and spending, population, census, geographical, parliament minutes, etc. It also includes data that is indirectly ‘‘owned’’ by public administrations (e.g. through subsidiaries or agencies), such as data related to climate/pollution, public transportation, congestion/traffic, child care/education. These data are considered as a lever to achieve: (i) more transparency in decision-making, (ii) increased collaboration of citizens and organizations in government policies, and (iii) innovation leading to a public and private value [3]. Hence, the appearance of Linked Open Data (LOD) that refers to data, which is published on the Web for use by public administrations, business and citizens. Such data apart from being machine-readable, is also linked to other external datasets. LOD is the process of following a set of best practices for publishing and connecting structured data on the Web [4]
Session 2 - Applications of Machine Learning	
Paul Vanabelle, Pierre De Handschutter, Riëm El Tahry, Mohammed Benjelloun and Mohamed Boukhebouze *Towards interpretable automated seizure detection through Fast Gradient Boosting applied on EEG signals*	The problem of automated seizure detection is treated using clinical electroencephalograms (EEG) and machine learning algorithms on the Temple University Hospital EEG Seizure Corpus (TUSZ). The purpose of this work is to determine to which extent the use of larger amount of data can help to improve the performances. Two methods are explored on a recent and large version of the TUSZ: a standard partitioning and a Leave-One-Out approach used to increase the amount of data in the training set. XGBoost, a fast implementation of the gradient boosting classifier, is used. The performances obtained are in the range of what is reported until now in the literature with deep learning models. We give interpretation to our results by identifying the most relevant features and analyzing performances by types of seizure.
Yassine Amkrane, Mohammed El Adoui and Mohammed Benjelloun *Analysing Breast cancer reaction to chemotherapy using Radiomics*	Breast cancer presents one of the most important diseases of women in the world. Indeed, Breast cancer presents is the second cause of death in the world with 8,8 million of deaths only in 2015. On the other hand, more than 70% of deaths, caused by breast cancer, occur in low-income countries since they are not equipped provided by efficient solutions for breast cancer diagnosis and prediction. The main challenge is diagnosis breast cancer as soon as it appears in other order to have the possibility ability to provide convenient and efficient treatments. In this context, several image modalities are used for breast tumor diagnosis such as echography, mammography and, PET (positron emission tomography) scans and MRI (Magnetic Resonance Images). One of the main treatments of this kind of pathologies is neoadjuvant chemotherapy, which attacks cancer cells and reduce the breast tumor’s size to facilitate the surgery. However, chemotherapy is hampered by several secondary effects (hair loss, osteoporosis, vomiting, etc.) and the cancer may not respond to it after several years of treatment. In this paper, we study propose a new method for breast cancer response prediction using bases on three steps: 1. Breast cancer segmentation from MR images. 2. Features extraction from segmented tumors in order to generate a complete and exploitable database. 3. Exploitation of deep learning architectures inf order to compute models allowing to the prediction of the tumor response. Experimental results will be conducted using a dataset of 42 breast cancer patients having local breast cancer, provided by our collaborator, Jules Bordet Institute - (Brussels) – Belgium.
Gaulthier Gain, Cyril Soldani and Laurent Mathy *UNICORE: A toolkit to automatically build unikernels*	Recent years have seen the IT industry move massively towards the use of virtualization for the deployment of applications. However, the two most prominent virtualization technologies, i.e. virtual machines (VMs) and containers, both present serious drawbacks. Full-blown VMs provide a good level of isolation, but are generally heavyweight. On the other hand, containers are generally more lightweight, but offer less isolation and thus a much greater attack surface. Unikernels have been proposed to virtualize applications in a way that is both safe, and efficient. They are specialized operating systems, tailored for a specific application, which allows to build minimalist VMs with tiny memory footprints. They keep the increased security of VMs, but with performance equivalent to or even better than equivalent containers. Unfortunately, porting an application to the unikernel paradigm currently requires expert knowledge, and can be very time-consuming. In this paper, we introduce UNICORE, a common code base and toolkit to automate the building of efficient unikernels from existing off-the-shelf applications. Although UNICORE is still in the early stages, we present early results showing that UNICORE images are able to yield performance similar or better than lightweight virtualization technologies such as containers.

Posters Presentation	
Boris Ndjia Njike and Xavier Siebert *Nonparametric adaptive active learning under local smoothness condition*	The paradigm of passive learning consists in providing a classifier based on a labelled raw data, chosen in a random way from a large pool of data. Due to a huge increase in the volume of the data available, we are sometimes constrained, from the point of view of the process of labeling data only, to look beyond the standard passive learning. In this situation, one of the most studied technique is active learning, where the algorithm is presented with a large unlabelled pool of data and we can iteratively request at a certain cost (to a so-called oracle) a given number n (called budget) of samples from the pool. The goal is to use this interaction to drastically reduce the number of label needed to provide a classifier whose excess error is as small as possible. Over the past decade, there has been a large body of work on understanding the benefit of active learning over the passive learning. One of the seminal works due to Castro and Nowak [4] analysed various scenarios and provided one in which active learning outperforms passive learning. This situation corresponds to a common assumption called Tsybakov noise assumption [5] that caracterizes (with parameter β) the noise near to the boundary decision; together with a smoothness assumption (with parameter α) related to the boundary decision (and that ensures for example that two closest points, with respect to a specific metric ”tend” to have the same label), they provided an active learning strategy that is better than the passive learning in the sense that it uses fewer labels request to reach a low-error. One of the main inconvenients of this strategy is that it is non adaptive in the sense that it requires the knowledge of the noise and smoothness parameter α and β that are unknown in practice. This drawback has been bypassed in [1] and [2] where adaptive (with respect to the smoothness and noise parameter) active algorithms are provided. Motivated by the real applications, we use a more general and realistic smoothness assumption [3], [6] instead of the one used in [1], [2], [4]. This smoothness assumption is caracterized by a parameter α ’. Together with Tsybakov noise assumption, we provide an active learning strategy that does not require the knowledge of α ’ and β.
James Ortiz and Pierre-Yves Schobbens *Specification and Modeling of Distributed Real-Time Systems*	Distributed Real-Time Systems (DRTS) can be characterized by a set of timing constraints, clocks and reactive interactions with several components or pro- cesses. In a DRTS the components run at different locations distributed geo- graphically over a communication network. Also, a DRTS can be classified as: (1) synchronous, if each of its components use the same clock and (2) asyn- chronous, if each of its components have their own independent clock which is subject to clock drifts. Synchronous and asynchronous models depict two forms in modeling and implementation of DRTS. However, the majority of current im- plementation of DRTS combines the advantages of these two models, which is known as timed asynchronous model. In a timed asynchronous DRTS, each pro- cess has access to its local clock that runs at the rate of global time. Formal verification methods and formalisms have been used to verify the correctness of DRTS. The traditional formalisms for reasoning about Real-time systems (RTS) are not always adequate for reasoning about DRTS. The most successful tech- niques for modeling RTS are Timed Automata (TA). A TA is a finite automaton augmented with real-valued clocks. The model of TA assumes perfect clocks: all clocks have infinite precision and are perfectly synchronized. This causes TA to have an undecidable language inclusion problem. Some variants of TA called Timed Automata with Independent Clocks (icTA) and Distributed Recursive Event Clock Automata (DECA) are proposed to model DRTS, where the clocks are not necessarily synchronized. In this poster we will present formal methods for the specification and model- ing of DRTS based on TA, icTA and Timed Labelled Transition Systems (TLTS). Here, we will extend the TLTS and icTA semantics to work with the new notion of distributed clocks (multi-timed automata (MTA)). Furthermore, we will extend the classical theory of timed bisimulation with the new notion of distributed clocks and multi-timed bisimulation. Also, we propose a new decision algorithm for multi-timed bisimulation. Additionally, we will propose extensions of the existing Lv and Hennessy- Milner logic with distributed clocks to allow the specification of distributed and timed properties. This gives us the (multi-timed) modal logic Lv, which we will show is PSPACE-complete for the (bounded) satisfiability and validity problem. Also, we show the applicability of MTA, multi-timed bisimulation and MLv over a DRTS. Finally, we will show an automatic formal verification tool which allows to verify DRTS.
Gaël Aglin, Siegfried Nijssen and Pierre Schaus *Learning Optimal Decision Trees using Caching Branch-and-Bound Search*	Several recent publications have studied the use of Mixed Integer Programming (MIP) for finding an optimal decision tree, that is, the best decision tree under formal requirements on accuracy, fairness or interpretability of the predictive model. These publications used MIP to deal with the hard computational challenge of finding such trees. In our work, we introduce a new eficient algorithm, DL8.5, for finding optimal decision trees, based on the use of itemset mining techniques. We show that this new approach outperforms earlier approaches with several orders of magnitude, for both numerical and discrete data, and is generic as well. The key idea underlying this new approach is the use of a cache of itemsets in combination with branch-and-bound search, and the eficient detection of equivalent tests on attributes.
Oussama Bouldjedri, Siegfried Nijssen and Pierre Schaus *Interpretable Soft Decision Trees*	In machine learning, some researchers focused on accuracy oriented models such as neural networks. These models made a good breakthrough in many applications. Others focused on creating algorithms with more interpretability (such as rule learning and decision tree). The question is what the trade-off between interpretability and accuracy is. Recent work on soft decision trees explored this trade-off by studying a form of decision trees that is less interpretable than standard decision trees, but more accurate. We continue the study of this trade-off by exploring whether or not we lose a lot of accuracy by simplifying the tests that are learned by soft decision trees. Decision trees (such as CART or C4.5) are one of the most interpretable machine learning models, every inner node [5] decides to which child an example is going by testing one attribute, and every example is predicted by one leaf. On the other hand, this type of tree has some drawbacks related to accuracy. In recent years, neural networks have obtained success by using many hidden layers and big architectures (VGG16[7], ...). However, these models have the noticeable disadvantage of non interpretable behaviour and unexplainable predictions. This is due to the fact that a single prediction can involve many (millions of) mathematical matrix operations, and non linear transformations, related to the used architecture or algorithm type. Humans have no chance in following this process from the beginning till the end. This pushed the community to create some prediction explanation and local interpretation methods, such as LIME.[6]. Soft decision trees were introduced by [3] and [4] they represent a trade-off between accuracy and interpretability. Compared to standard decision trees, in every node of the tree they use a weights vector to calculate a probability that an example goes in the lefthand or right-hand branch of the node. The prediction is based on all the leafs of the tree, where the predictions in the leafs are weighted by the probabilities of arriving in that leaf. This type of tree is more accurate than a standard decision tree but less interpretable. The inner node uses this formula: pi (x) = σ(xwi + bi) where σ is the sigmoid function, x the input data, w the weights vector, b the bias, and p(x) the probability (decision). This type of tree is more interpretable than a neural network because this model is hierarchical mixture of expert[4], for which every expert is a bigot (leaf), that always produces the same probability distribution. The model learns a hierarchy of filters that are used to map each example to a particular expert. Soft decision trees are more expressive than standard decision trees. Inner nodes can express more complex tests than standard decision trees by using arbitrary linear equations; standard trees are limited to single attributes. Moreover, they offer the ability to sum weighted predictions over all leafs, instead of giving all weight to one leaf which make them more accurate than standard decision trees. However, the fact that each node exploits a larger number of features, and that every prediction is based on all leafs of the tree, makes these predictions harder to understand. Which make them less interpretable than a standard decision tree. Our study explores the trade-off between interpretability and accuracy that was initially explored in soft decision tree[3]. This is achieved by simplifying the tests performed in the tree. The main idea is the discretization of the weights[1], and activation functions[2] through binarization, using 0.5 as threshold to round values to 0 or 1.

Session 3 - Foundations : Logic Programming and Game Theory	
Quentin Meurisse *Local search and game theory applied to an urban planning problem*	In prospect of sustainable urban densification, a tool aiming to assess and to assist the design of compact housing blocks with a target population density was created and tested in the scope of the CoMod Project, a coordinated project supervised by the Faculty of Sciences and the Faculty of Architecture and Urban Planning of the University of Mons. The concept of spatial compactness is here applied, at the architectural scale, on the built environment, the non-built environment and both combined. This approach encourages typo-morphologies which save land and material resources while achieving high energy efficiency. Potential misuse of the concept is prevented by numerous objective criteria notably relative to green areas, projected shadows as well as minimal distances and surfaces to consider.
Aline Goeminne *Multiplayer reachability games played on graphs*	A reactive system is a system which interacts continuously with the environment in which it evolves. This system aims at satisfying some objective whatever the behaviour of the environment. For example, the autopilot of a plane has to ensure the safe landing of the plane whatever weather conditions (see Figure 1). For this kind of critical systems, it may be dramatic for the system to be subject to bugs. It is the reason why we want to know if a system is correct and satisfy some properties. One way to try to do so is program testing but as Edsger W. Dijkstra said “Program testing can be used to show the presence of bugs, but neither to show their absence”.
Pierre De Handschutter, Nicolas Gillis, Arnaud Vandaele and Xavier Siebert *Near Convex Archetypal Analysis*	Nonnegative matrix factorization (NMF) is a widely used linear dimensionality reduction technique for nonnegative data. NMF requires that each data point is approximated by a convex combination of basis elements. Archetypal analysis (AA), also referred to as convex NMF, is a well-known NMF variant imposing that the basis elements are themselves convex combinations of the data points. AA has the advantage to be more interpretable than NMF because the basis elements are directly constructed from the data points. However, it usually suffers from a high data fitting error because the basis elements are constrained to be contained in the convex cone of the data points. In this letter, we introduce near-convex archetypal analysis (NCAA) which combines the advantages of both AA and NMF. As for AA, the basis vectors are required to be linear combinations of the data points and hence are easily interpretable. As for NMF, the additional flexibility in choosing the basis elements allows NCAA to have a low data fitting error. We show that NCAA compares favorably with a state-of-the-art minimum-volume NMF method on synthetic datasets and on a real-world hyperspectral image.
Gonzague Yernaux *Generalizing Generalization: towards a framework for anti-unification problems in Logic Programming*	Anti-unification (the dual operation of unification) is typically defined as a computing process that outputs a program object (called a generalization) when other program objects of a similar nature are given as inputs. The computed generalization must be more general than all input objects with respect to a given generalization relation, and usually some optimization is required in order to find the "most specific" or "best" possible generalization according to some criterion, rather than any common generalization. In logic programming contexts, many independent anti-unification approaches exist – each with its own interesting specificities – but no effort has yet been done in the literature to devise a synthetic framework for anti-unification. We believe that such an effort would not only provide possibilities for replacing existing anti-unification algorithms by slightly different and perhaps more adequate or more efficient solutions, but also highlight specific configurations for which no algorithm even exists at the moment, although the anti-unification problems falling into said configurations could be numerous in practice. This short paper first aims to formalize the concept of configuration as a foundation for such a theory of generalization. It will pursue related work in the exploration of specific configurations that haven’t been thoroughly studied before (such as unordered goals anti-unification). Throughout the paper we describe and motivate ongoing and potential future work in the vast area of anti-unification techniques in logic programming.
Session 4 - Computing Education and HMI	
Simon Liénardy, Lev Malcev, Laurent Leduc and Benoit Donnet *Graphical Loop Invariant programming in CS1*	This paper introduces the use of Graphical Loop Invariant as a programming methodology in a CS1 course, in which the Loop Invariant is determined prior to writing the code and is meant as a help to find the loop instructions. This paper also introduces two learning tools: GLI, an application helping students to draw Loop Invariant and Café, an on-line platform designed to assess and deliver automatic feedback and feedforward information to students, in particular on their Loop Invariants and the pieces of code based upon them. The paper reports preliminary evaluation on Café usage.
Adrien Coppens *Graph-Based Architectural Modelling in Virtual Reality*	The current integration of immersive technologies in architecture is mostly limited to visualisation purposes. Little work has been done in providing design capabilities from within virtual environments, even though the three-dimensional context associated with those environments matches the dimensionality of the output geometries. The shortage in VR-enabled tooling is particularly striking for parametric architectural modelling and we therefore introduce the research we conducted on filling that gap. Our contributions include different VR prototypes that allow designers to modify parametric models (i.e. interact with nodes and edges in a graph), their parameters, and see the resulting changes on generated geometries.
Cédric Libert and Wim Vanhoof *Analysis of Students’ Preconceptions of Concurrent Programming*	Dans la littérature, plusieurs auteurs encouragent l’enseignement de la concurrence, car beaucoup d’évolutions en informatique font appel à ce concept. Nous rapportons dans cet article le résultat d’une expérience menée auprès de 101 adolescents et adolescentes de 12 à 15 ans à qui nous avons dispensé deux séances de 100 minutes sur la concurrence par passage de messages. Nous y commentons l’évolution des élèves face au concept de concurrence et face à la réalisation d’un problème de synchronisation.