Agenda BIRS 15w2181 (consolidated version as of 22.07.2015)

Friday, 24.7.2015 to Sunday, 26.7.2015

Advances in interactive Knowledge Discovery and Data Mining in complex and big data sets

"Beyond Data Mining"

Arrival Friday, 24.7.2015

18:00 – 21:00 Welcome to Banff - come together, informal information exchange amongst participants

DAY 1: Saturday, 25.7.2015

Note: This special workshop is dedicated to stimulate the cross-domain integration and appraisal of different fields and to provide an atmosphere to foster different perspectives and opinions; it will offer a platform for novel ideas and a fresh look on the methodologies to solve our grand common goal: to discover unknown unknowns in complex data sets with a focus on biomedicine. The output will be documented by a Volume of Springer Lecture Notes in Artificial Intelligence (LNAI) edited by the organizers: Andreas HOLZINGER, Randy GOEBEL, Vasile PALADE and Massimo FERRI.

Machine Learning and Biomedicine are **the** topics of the future. The central goal is to outline a roadmap for future and hot research topics, to tackle grand challenges, useful as input for joint grant proposals - to enable tackling grand challenges in a concerted effort and carrying out research which none of us could do alone.

Note: Each of the following presentations is flexible from the timing, it is just a framework - interactive discussions are encouraged! Each sessions will end with a wrap-up session.

S1 - 09:00 to 11:00 FIRST SESSION WITH OPEN DISCUSSION ROUND

Welcome from the organizers (short welcome statements by Andreas, Randy, Vasile and Massimo)

Andreas HOLZINGER, Research Unit HCI-KDD, Institute of Medical Informatics, Medical University Graz and CBmed Competence Center for Biomarker Research in Medicine, Austria

Title: Challenges of biomedicine, health and the life sciences and the chances of Interactive Machine Learning for Knowledge Discovery

Abstract: In this presentation I will provide an overview of the variations and complexity of data sets from biomedicine, health care and the life sciences and the problems and challenges biomedical researchers of today are faced, when trying to gain insight into their data to discover unknown unknowns. Machine learning algorithms may be of help here, and a best practice today is demonstrated by autonomous vehicles ("Google car"). However, in complex domains such as biomedicine, where we deal with uncertain, probabilistic, and weakly structured data the application of fully automatic machine learning algorithms endangers the modelling of artifacts. Therefore I will emphasize in my presentation the importance of supporting human intelligence with interactive

machine learning by putting the human-in-the-loop. Our long term goal is to contribute towards cognitive computing systems, that learn and interact naturally with experts together to extend what neither a human nor a computer could do on its own.

Bio: Andreas Holzinger, head of the Research Unit HCI-KDD, Institute for Medical Informatics at the Medical University Graz, lead of Area 1 Data of the newly established national competence center for Biomarker Research; lead of the international expert network HCI-KDD, Associate Professor of Applied Informatics at the Faculty of Computer Science and Biomedical Engineering at Graz University of Technology. Andreas serves as consultant for the Canadian, Swiss, French and Dutch Government and for the German Excellence Initiative. Andreas was Visiting Professor in Berlin, Innsbruck, Vienna, 2 times in London, and Aachen. Andreas and his group is passionate on supporting human intelligence with machine learning – to discover new, previously unknown insights into complex biomedical data. Andreas holds a PhD (1998) in Cognitive Science from Graz University and a Habilitation (second PhD, 2003) in Computer Science from Graz University of Technology.

Technical Area: Knowledge Discovery/Data Mining, interactive Machine Learning, Cognitive computing

Application Area: Biomedical Informatics, Smart Health, Personalized Medicine

Randy GOEBEL, Machine Learning Group, University of Edmonton, Canada

Title: The role of logic and machine learning within a general theory of visualization

Abstract: The role of logic and machine learning in visualization is not familiar to many, but the idea of visual inference requires inductive transformations from base data to visual data. These transformations need to be constrained by inference principles, including the construction of layers of knowledge, which generally are difficult to construct by hand. The idea is to describe how logic, learning, and visualization are connected, in order to help enable humans to make better inferences from growing volumes of data in every area of application.

Bio: R.G. (Randy) Goebel is currently professor of Computing Science in the Department of Computing Science at the University of Alberta, and principle investigator in the Alberta Innovates Centre for Machine Learning (AICML) Randy's theoretical work on abduction, hypothetical reasoning and belief revision is internationally well know, and his recent application of practical belief revision and constraint programming to scheduling, layout, and web mining is now having industrial impact. His recent research is focused on the formalization of visualization, with applications in web mining, optimization, and nanotechnology. Randy has previously held faculty appointments at the University of Waterloo and the University of Tokyo, and is actively involved in academic and industrial collaborative research projects in Canada, Japan, China, and Germany.

Technical Area: Information extraction from text, logic of visualization

Application Area: Bioinformatics, natural language processing

Vasile PALADE, Cogent Computing Applied Research Centre at Faculty of Engineering and Computing, Coventry University and University of Oxford, UK

Title: Class Imbalance Learning

Abstract: Class imbalance of data is commonly found in many data mining tasks and machine learning applications to real-world problems. When learning from imbalanced data, the performance measure used for model selection plays a vital role. The existing and popular performance measures used in class imbalance learning, such as the Gm and Fm, can still result in sub-optimal classification models. The talk will first present a new performance measure, called the Adjusted Geometric-mean (AGm), which overcomes the problems of the existing performance measures when learning from imbalanced data. Support Vector Machines (SVMs) has become a very popular and effective machine learning technique, but which can still produce sub-optimal models when it comes to imbalanced datasets. The talk will then present FSVM-CIL (Fuzzy SVM for Class Imbalance Learning), an effective method to train FSVMs with imbalanced data in the presence of outliers and noise in the data. Finally, some efficient resampling methods for training SVMs with imbalance data will also be discussed in the context of applications.

Bio: Dr. Vasile Palade has joined the Department of Computing at Coventry University, United Kingdom, in September 2013, after working for several years with the Department of Computer Science at the University of Oxford. His research interests spans across several machine learning and computational intelligence domains, and include neural networks and neuro-fuzzy systems, different nature inspired learning and optimization algorithms, hybrid intelligent systems, class imbalance learning. Main application areas are bioinformatics and computational biology problems, fault diagnosis, process modelling and control, web usage mining and social network data analysis, image processing. He has published several books and 120 papers in machine learning journals and conference proceedings. He is acting as an Associate Editor to several journals, e.g., Knowledge and Information Systems, Neurocomputing, International Journal of Artificial Intelligence Tools, International Journal of Hybrid Intelligent Systems.

Technical Area: Machine Learning

Application Area: real-world problems

11:00 to 11:30 COFEE BREAK

S2 - 11:30 to 13:00 SECOND SESSION WITH OPEN DISCUSSION ROUND

Katharina MORIK, Artificial Intelligence Group, Technical University Darmstadt, Germany

Title: Big Data and Small Devices

Abstract: Big data are produced by various sources. Most often, they are distributedly stored at computing farms or clouds. Analytics on the Hadoop Distributed File System (HDFS) then follows the MapReduce programming model (batch layer). It is complemented by the speed layer, which aggregates and integrates incoming data streams in real time. When considering big data and small devices, obviously, we imagine the small devices being hosts of the speed layer, only. Analytics on the small devices is restricted by memory and computation resources. The interplay of streaming and batch analytics offers a multitude of configurations. The collaborative research center SFB 876 investigates data analytics for and on small devices regarding runtime, memory and energy consumption. In this talk, we investigate graphical models, which generate the probabilities for connected (sensor) nodes. Resource-restricted methods deliver insights fast enough for a more interactive analysis.

Bio: Katharina Morik received her Ph D at the University of Hamburg 1981 and became full professor at the University of Dortmund in 1991. Katharina started the IEEE International Conference on Data Mining together with Xindong Wu, being the program chair person in 2004. She organized the summer schools on "Ubiquitous Knowledge Discovery" 2006 and 2008 and co-chaired the European Conference on Machine Learning and Data Mining ECML/PKDD 2008. She is a member of the Scientific Advisory Board of Rapid-I and scientific council of The European Institute for Participatory Media. She was a full partner in several European research projects and coordinator of one. She has published more than 200 papers in acknowledged journals and conferences and is an editorial board member of the journals “Knowledge and Information Systems” and “Data Mining and Knowledge Discovery”. Katharina Morik is the speaker of the collaborative research center SFB876, which started in 2011 and began its second phase of 4 yours in 2015. In 14 projects, 20 professors and about 50 Ph D students and PostDocs work on data analysis under resource constraints.

Technical Area: Feature extractions and selection, modeling under resource constraints, spatio-temporal modeling

Application Area: Information Extraction from Texts and Text Classification, Data Mining in Industry 4.0, Big Data Analysis in Astroparticle Physics

13:00 – 14:30 LUNCH

S3 - 14:30-15:30 3rd SESSION (STUDENT SESSION) WITH OPEN DISCUSSION ROUND

Sibylle HESS, PhD Student at the SFB 876, Technical University Dortmund, Germany

Title: Investigation of Code Tables to compress and describe the underlying characteristics of binary databases.

Abstract: We inspect the spectrum of methods (from frequent pattern mining to numerical optimization) to extract the pattern set that describes a binary database best. Invoking the Minimum Description Length (MDL) principle, this objective can be stated as: find the code table that compresses the database most. A particularly interesting interpretation of this task, relating it to biclustering, arises from the formulation as a matrix factorisation problem. Biclustering has a variety of applications in research fields such as collaborative filtering, gene expression analysis and text mining. The derived matrix factorisation analogy provides a new perspective on distinct data mining subfields (unifying biclustering and pattern mining concepts such as Krimp), initialising a cross-over of their applications and interpretations of derived models.

Katharina HOLZINGER, Student of Natural Sciences, Karl-Franzens University Graz

Title: Darwin, Lamarck, Baldwin, Mendel: What can we learn from them?

Abstract: Evolutionary Algorithms, inspired by biological mechanisms observed in nature, such as selection and genetic changes, have much potential to find the best solution for a given optimisation problem. Contrary to Darwin, and according to Lamarck and Baldwin, organisms in natural systems learn to adapt over their lifetime and allow to adjust over generations. Whereas earlier research was rather reserved, more recent research underpinned by the work of Lamarck and Baldwin, finds that these theories have much potential, particularly in upcoming fields such as epigenetics. Particularly the integration of the Theories of Gregor Mendel could help knowledge discovery.

15:30-16:00 COFFEE BREAK

S4 - 16:00 - 18:00 FOURTH SESSION WITH OPEN DISCUSSION ROUND

Nitesh CHWALA, Interdisciplinary Center for Network Science (iCeNSA), University of Notre Dame, US

Title: Big Data and Small Data for Personalized and Population Healthcare

Abstract: Proactive personalized medicine can bring fundamental changes in healthcare. Can we then take a data-driven approach to discover nuggets of knowledge and insight from the big data in healthcare for patient-centered outcomes and personalized healthcare? Can we answer the question: What are my disease risks and how to best manage it? How to scale this at the population level? I will discuss our work that takes the data and networks driven thinking to personalized healthcare and

patient-centered outcomes. It demonstrates the effectiveness of population health data to drive personalized disease management and wellness strategies, and in effect impacting population health. I will also share various pilots under-way that take the algorithms and tools on a "road-show".

Bio: Nitesh Chawla, PhD is the Professor of Computer Science and Engineering and the Frank Freimann Collegiate Chair of Engineering at the University of Notre dame. He directs the Notre Dame Interdisciplinary Center for Network Science (iCeNSA), which is at the frontier of network and data science and transformative interdisciplinary applications (Big Data). His work has received and been nominated for a number of best paper awards, and has been recognized by various publications and avenues. He is the recipient of the 2015 IEEE CIS Outstanding Early Career Award. He received the IBM Watson Faculty Award in 2012, and the IBM Big Data and Analytics Faculty Award in 2013. He was also highlighted as IBM Big Data and Analytics Hero in 2014. He is the recipient of the National Academy of Engineering New Faculty Fellowship. He received the Outstanding Teacher Award in Computer Science and Engineering in 2008 and 2011. In recognition of the societal impact of his research, he was recognized with the Rodney Ganey Award, and the Michiana 40 Under 40 honor in 2014. He is also the director of ND-GAIN Index, Fellow of the Reilly Center for Science, Technology, and Values, Fellow of the Institute of Asia and Asian Studies, and Fellow of the Kroc Institute for International Peace Studies at Notre Dame. He is a co-founder of Aunalytics, Inc., a big data analytics start-up.

Technical Area: Machine Learning, Data Science, Network Science

Application Area: Healthcare Analytics, Smart Health, Personalized Medicine

Yuzuru TANAKA, Meme Media Laboratory, Hokkaido University, Sapporo, Japan

Title: Exploratory Visual Analytics for the Discovery of Complex Analysis Scenarios for Big Data

Abstract: Data-centric approach is increasing its significance in varieties of scientific research areas and large-scale social cyber-physical systems. Biomedical research area and the urban-scale winter road management are such examples. Through his involvement in three major projects on these subjects, the current author recognized a big gap between the state-of-the-art big data core technologies and both the data-centric research for the analysis of clinico-genomic trial data and the big data approach to the optimization of social system services. During the last couple of decades, the enabling core technologies for big data analysis have made remarkable advances in both analysis and management technologies. However, we still lack methodologies to find out the best analysis scenario for finding out such solutions as personalized medicines or optimized snow removal plans from a given clinico-genomic trial data set or from given traffic and weather data sets. This talk will propose exploratory visual analytics to support analysts to find out complex analysis scenarios, and the coordinated multiple views and analyses framework as its application framework.

Bio: Yuzuru Tanaka has been a full professor of computer architecture at the Department of Electrical Engineering (1990-2003), then of knowledge media at the Department of Computer Science,

Graduate School of Information Science and Technology (2004-), Hokkaido University, and the founding director of Meme Media Laboratory (1995-2013), Hokkaido University. He was also a full professor of Digital Library, Graduate School of Informatics, Kyoto University (1998-2000) in parallel, and has been an adjunct professor of National Institute of Informatics (2004-). His current research areas cover meme media architectures, knowledge federation frameworks, proximity-based federation of smart objects, and their application to digital libraries, e-Science, clinical trials on cancer, and social cyber-physical systems for the optimization of social system services. He worked as a visiting research fellow at IBM T.J. Watson Research Center (1985-1986), an affiliated scientist of FORTH in Crete (2010-), and the program officer of JST'S eight year CREST Program on Big Data Application Technologies (2013-).

Technical Area: Knowledge Media Architecture Knowledge Federation Exploratory Visual Analytics

Application Area: personalized medicine social cyber-physical systems

DAY 2: Sunday, 26.7.2015 ----------

S5 - 09:00 – 11:00 FIFTH SESSION WITH OPEN DISCUSSION ROUND

Mateusz JUDA, Mrozek Group, Jagiellonian University, Krakow, Poland

Title: Homology of big data - algorithms and applications

Abstract: Homology is a well known and powerful tool in pure mathematics. For many years it was impossible to use this tool in applied science because of data size and cubical algorithms for computing homology. New preprocessing methods give us a possibility to apply homology for real data. Discrete Morse theory is an example of a tool, which simplifies data without changing its topological information. During this talk I introduce discrete Morse theory and its application to homology computations. I show how to construct a discrete vector field (Morse matching) using parallel and distributed algorithms. I also show an application of this tool to knots detection and classification in a biological context.

Bio: Mateusz Juda, post-doc at Jagiellonian University, Krakow, Poland. Mateusz holds PhD (2013) in Computational Mathematics, MSc (2008) in Computer Science. Mateusz works on computational topology, especially on algorithms for homology computations. Now he is working on Topological Complex Systems (Toposys project funded by EU 7FP). He is main developer of CAPD::RedHom software - a tool for homology computations. He also works on parallel and distributed algorithms for homology of big data sets. Mateusz is also a professional software developer with 5 years of commercial experience.

Technical Area: computational topology, scientific computations, software development

Application Area: data mining, complex systems

Massimo FERRI, Vision Mathematics Group, Department of Mathematics, University of Bologna, Italy

Title: Persistent topology for natural shape analysis and image retrieval.

Abstract: Data are more and more often of "natural" origin (pictures or *3D* meshes representing living beings, faces, handwritten words, hand-drawn sketches etc.). Classical mathematical techniques do not fit well the task of analyzing, comparing, classifying, retrieving such data. On the contrary topology (and in particular algebraic topology) is, by its very nature, the part of mathematics which formalizes qualitative aspects of objects; therefore topological data processing and topological data mining well integrate with more classical mathematical tools.

Persistent homology combines geometry and algebraic topology in the study of pairs (X,f) where X is an object (typically a topological space) and f is a continuous function defined on X (typically with real values). One application is the extraction of topological features of an object out of a cloud of sample points. Another class of applications uses f as a formalization of a classification criterion; in this case various functions can give different criteria, cooperating in a complex classifier. Persistent

homology is studied by several teams throughout the world both from the theoretical and computational viewpoints and has already given rise to several applications: dermatological diagnosis, evolution of hurricanes, signature recognition, gesture recognition; retrieval of trademarks, 3D meshes, hand-drawn sketches etc.

Bio: Massimo Ferri is full professor of Geometry at the University of Bologna, where he leads the Vision Mathematics Group. He is the coordinator of Research Project 2 "Computer vision and image processing systems" of the Research Centre on Electronic Systems for the Information and Communication Technology "E. De Castro" of the University of Bologna. His initial research interest was in the topology of low-dimensional combinatorial manifold, but then he drifted towards applications of geometry and topology to computer vision and pattern recognition. He worked at geometrical methods in robot navigation and at the development of aids for the visually impaired. He follows the development of persistent homology since its birth.

Technical Area: Topological methods in shape analysis.

Application Area: Pattern recognition, shape analysis, image retrieval.

11:00 – 11:30 COFFEE BREAK

S6 - 11:30 – 13:00 SIXTH SESSION WITH OPEN DISCUSSION ROUND

Mirko CESARINI, Department of Statistics and Quantitative Methods, University of Milano Bicocca, Italy

Title: Data Quality in Schema free (big) data.

Abstract: The presentation will focus on the challenges and open problems emerging when complex data sets are used to obtain insights about a population e.g., analysing job offers using data from web job boards, inspecting the job history of the working population (starting from administrative records), and analysing cellular network traffic. A huge set of weakly structured data can be derived from information sources containing a variety of data types. In such a context, techniques ranging from formal methods to machine learning can identify and exploit information structures (both hidden and visible) to check data consistency, to ameliorate the data (e.g., fixing inconsistencies), and to create synthetic representations of the original data.

Bio: Mirko Cesarini is currently working as professor assistant at the Department of Statistics and Quantitative Methods, University of Milan Bicocca. His research activities focus on Information Systems design, checking and improving Data Quality over large data sets, Information extraction from unstructured data, and Machine Learning. Mirko Cesarini holds a Ph.D. in Computer Engineering awarded by the Polytechnic University of Milan in 2005. He is affiliated to the CRISP research Centre, University of Milan Bicocca. He participated to several research projects funded by the European Commission, the Italian national and regional governments. He serves also as consultant

for several institutions, organizations, and governmental offices. He published several scientific papers on international reviews, conference proceedings, and books.
Technical Area: Data Quality / Machine Learning / Unstructured Data
Application Area: Job Market Place, Healthcare data, Mobile (Cellular) Network Data

S7 - 11:30 – 13:00 SEVENTH SESSION WITH OPEN DISCUSSION ROUND

Sou-Cheng CHOI, NORC at the University of Chicago and Illinois Institute of Technology, US
Title: Machine Learning for Machine Data in Computational Social Sciences
Abstract: We present machine learning and high-accuracy prediction methods of rare events in semi-structured or unstructured log files produced at high velocity and high volume by NORC's computer-assisted telephone interviewing network. These machine log files are generated by our internal Voxco Servers for a telephone survey. We adapt natural language processing (NLP) techniques and data-mining methods to train powerful learning and prediction models for error messages in the absence of source code, updated documentation, and relevant dictionaries.
Bio: Sou-Cheng Choi is a Senior Statistician with the Statistics and Methodology, Department at NORC and Research Assistant Professor in the Department of Applied Mathematics, Illinois Institute of Technology (IIT). At NORC, she works in the areas of computational social sciences that involve processing, analysis, and mining of big data. She is currently conducting research and development work on probabilistic record linkage of massive administrative data entries with missing unique identifiers. Choi also provides technical expertise on performance management of a vendor phone surveying network deployed in the continental call centers of NORC, applying efficient natural-language processing techniques and machine-learning methods on high-velocity network and server-log data. With other NORC experts, she has developed computational methods for analyzing and de-identifying structured as well as unstructured data from large repositories of confidential medical transcription. In NORC, she has worked on the production system of National Immunization Survey for Children and Teenagers. With her IIT collaborators, Sou-Cheng co-creates Guaranteed Automatic Integration Library (GAIL), a suite of algorithms for function approximation, global optimization, integration problems in one or many dimensions, and whose answers are guaranteed to be correct.
Technical Areas: Applied and computational mathematics and statistics. Numerical analysis and algorithms. Data sciences.
Application Area: Networks

Sayan MUKHERJEE, Duke University, US
Monica NICOLAU, Stanford School of Medicine, US
Lek Heng-LIM, University of Chicago, US
Title, Abstract and Bio follows