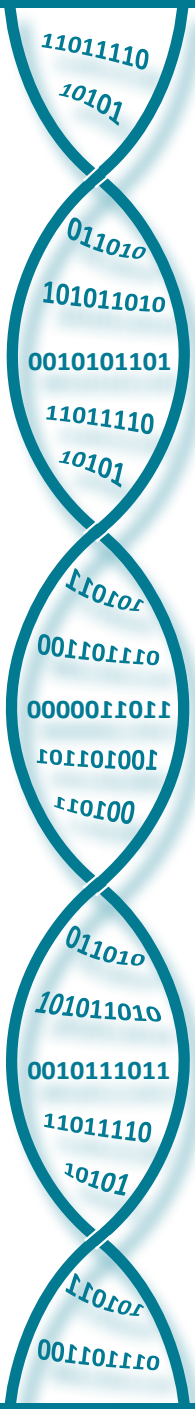

Karsten Borgwardt

KRUPP SYMPOSIUM

From Machine Learning to Personalized Medicine

Max Planck Institute of Psychiatry
Munich, Germany
October 21, 2016

ETH zürich
Alfried Krupp von Bohlen und Halbach-Stiftung

Alfried Krupp von Bohlen und Halbach-Stiftung

The non-profit **Alfried Krupp von Bohlen und Halbach-Stiftung** is the bequest of Dr.-Ing. E. h. Alfried Krupp von Bohlen und Halbach, the last sole proprietor of the Fried. Krupp company. Upon his death in 1967, his entire assets passed to the foundation which he had previously established. This foundation was only made possible by his only son, Arndt von Bohlen und Halbach, who renounced his inheritance.

The Alfried Krupp von Bohlen und Halbach-Stiftung became active in January 1968, less than half a year after Krupp's death. It supports national and international projects in the fields of science in research and teaching, education and training, health services, sports, as well as literature, music and fine art.

The **Alfried Krupp-Förderpreis für junge Hochschullehrer** (Alfried Krupp Prize for Young University Teachers) was established in 1986 and has since been awarded to 36 young professors teaching and researching in all fields of natural and engineering sciences. The award money of 1 million Euros is intended to support the young scientists in pursuing their individual research projects and goals independently from university or government funds.

(October 2016)

http://www.krupp-stiftung.de

Program

13:00-13:10	**Welcome** Karsten Borgwardt *ETH Zürich*
13:10-13:50	**Machine Learning for Precision Medicine** Jean-Philippe Vert *Mines ParisTech, Institut Curie and Ecole Normale Supérieure*
13:50-14:30	**The Case for Non-Linearity in Statistical Genetics and Beyond** Bertram Müller-Myhsok *MPI of Psychiatry, Munich and University of Liverpool*
14:30-15:10	**Multi-Locus Genome-Wide Association Studies** Chloé-Agathe Azencott *Mines ParisTech, Institut Curie and INSERM*
15:10-15:40	**Coffee Break**
15:40-16:20	**Significant Pattern Mining for Biomedical Applications** Koji Tsuda *University of Tokyo*
16:20-17:00	**Significant Pattern Mining for Biomarker Discovery** Felipe Llinares-López *ETH Zürich*
17:00-17:40	**Exploiting Spatial Features in the Analysis of DNA-Methylation Data** Guido Sanguinetti *University of Edinburgh*
17:40-18:00	**Closing Remarks** Karsten Borgwardt *ETH Zürich*
18:00	**Transfer to Symposium Dinner**
18:30	**Dinner at Seehaus im Englischen Garten**

Abstracts of the talks

Machine Learning for Precision Medicine

Jean-Philippe Vert
Mines ParisTech, Institut Curie and Ecole Normale Supérieure

The development of DNA sequencing technologies allows us to collect large amounts of molecular data about the genome of each individual, and opens the possibility to predict drug response or evaluate the risk of various diseases from one's molecular identity. It also raises statistical and computational challenges, as the quantity of data collected per sample is usually far larger than the number of samples available to estimate predictive models. In this talk I will discuss some of these challenges and describe a few methods to attack them through regularization or change of representation, illustrated on applications in cancer prognosis from gene expression and somatic mutations.

The Case for Non-Linearity in Statistical Genetics and Beyond

Bertram Müller-Myhsok
MPI of Psychiatry, Munich and University of Liverpool

Much of genetic analysis in humans currently is focused on the linear model, especially a polygenic model variant that encodes small variations. However, there is ample evidence that the underlying nature of phenotypic traits in biological systems is actually much less linear in nature than currently assumed implicitly. This calls for efficient, powerful and general methods of analysis in the study of dependencies, thus providing a strong case for joining statistics and machine learning in general and statistical genetics and machine learning in particular.

Multi-Locus Genome-Wide Association Studies

Chloé-Agathe Azencott
Mines ParisTech, Institut Curie and INSERM

As an increasing number of genome-wide association studies reveal the limitations of attempts to explain phenotypic heritability by single genetic loci, the need for methods associating complex phenotypes with sets of genetic loci becomes pressing. I will present two such methods. The first one makes use of the computational power of GPUs to enable systematic two-locus mapping

tools; the second relates sets of associated loci of arbitrary sizes with prior biological knowledge about gene pathways and networks.

Significant Pattern Mining for Biomedical Applications

Koji Tsuda
University of Tokyo

Pattern mining techniques such as itemset mining, sequence mining and graph mining have been applied to a wide range of datasets. To convince biomedical researchers, however, it is necessary to show statistical significance of obtained patterns to prove that the patterns are not likely to emerge from random data. The key concept of significance testing is family-wise error rate (FWER), i.e., the probability of at least one pattern is falsely discovered under null hypotheses. In the worst case, FWER grows linearly to the number of all possible patterns. We show that, in reality, FWER grows much slower than the worst case, and it is possible to find significant patterns in biomedical data. The following two properties are exploited to accurately bound FWER and compute small p-value correction factors. 1) Only closed patterns need to be counted. 2) Patterns of low support can be ignored, where the support threshold depends on the Tarone bound. We introduce an efficient depth-first search algorithm for discovering all significant patterns and discuss about parallel implementations.

Significant Pattern Mining for Biomarker Discovery

Felipe Llinares-López
ETH Zürich

Significant pattern mining has recently emerged as a promising approach for biomarker discovery due to its ability to discover discriminative high-order feature interactions despite the enormous number of possible candidate interactions in real-world datasets. At the heart of significant pattern mining is Tarone's concept of testability, which enables pruning a large part of the search space by proving that many candidate feature interactions cannot possibly achieve significance nor cause false positives. The use of Tarone's testability criterion allows controlling the Family Wise Error Rate while retaining high statistical power to discover truly discriminative feature interactions, as well as being essential for computational efficiency.

Despite their success, the original significant pattern mining algorithms inher-

ited some fundamental limitations from Tarone's testability criterion. In this talk, we will introduce two extensions of the existing framework of special importance for biomarker discovery. First, we will discuss how to take into account the dependence between patterns by using permutation testing, hereby increasing statistical power by accounting for redundancy. Secondly, we will show how the original model can be extended to correct for confounding by incorporating an observed categorical covariate without sacrificing power nor computational efficiency.

Exploiting Spatial Features in the Analysis of DNA-Methylation Data

Guido Sanguinetti
University of Edinburgh

Epigenetic modifications such as histone modifications and DNA methylation play a central role in the regulation of gene expression. Next generation sequencing technologies are now enabling genome-wide measurements of epigenetic marks, yet the data returned by such technologies is difficult to interpret. Here, we start from the observation that such measurements often return broad, spatially correlated patterns of modification which are highly reproducible between replicate experiments. We then use machine learning methodologies to exploit such spatial correlations to define stronger prediction methods. In particular, I will illustrate a novel statistical hypothesis testing methodology, M3D, for BS-Seq data, which exploits spatial features to yield more powerful tests. I will also illustrate how higher-order spatial features may be used to predict gene expression from DNA methylation alone.

About the symposium

Karsten Borgwardt is the 2013 laureate of the Alfried Krupp Prize for Young University Teachers worth 1 million Euros. By tradition, all laureates host a scientific symposium with leading international experts from their research field three years after receiving the award.

More information on Karsten, his lab and work is available on:

https://www.bsse.ethz.ch/mlcb

More information and news articles on the Krupp Award 2013 can be found on:

https://www.bsse.ethz.ch/mlcb/krupp-award-2013