Big Data and Data Analytics

WIC Midwintermeeting, February 1st, 2016
Eindhoven, TU/e, Zwarte Doos

Organized by:
Werkgemeenschap voor Informatie- en Communicatietheorie,http://www.w-i-c.org/
IEEE Benelux Chapter on Information Theory, http://ewh.ieee.org/r8/benelux/it/

Big Data refers to data sets that are so large that traditional data processing applications are inadequate. The challenge is to extract relevant information from these data sets and to analyse the data, which is the objective in the research area Data Analytics. At the 2016 WIC Midwintermeeting experts from different scientific domains will share with us their views on recent advances and innovative technologies involving big bata. The connection to Information Theory is also discussed during the meeting. We are inviting you to participate in the Midwintermeeting in the Zwarte Doos in Eindhoven on February 1st.

Program:

09:30 - 10:00	Registration and Coffee
10:00 - 10.15	Opening
10.15 - 11.00	**Joel Veness (Google DeepMind, London)**, "Compression to Control"
11.00 - 11.45	**Joris Mooij (UVA, Amsterdam)**, "Causal Discovery and Prediction from Big Data"
11.45 - 12.00	Short Break
12.00 - 12.45	**Bert de Vries (GN Resound & TU/e, Eindhoven)**, "Design of Signal Processing Algorithms through Probabilistic Inference"
12.45 - 13.45	Lunch Break
13.45 - 14.30	**Elena Marchiori (Radboud Universiteit, Nijmegen)**, "Network Community Detection by Seed Expansion"
14.30 - 15.15	**Tim van Erven (Leiden University, Leiden)**, "From Data Compression to Online Machine Learning"
15.15 - 15.30	Short Break
15.30 - 16.15	**Jeroen Laros (Leids Universitair Medisch Centrum),** "Applications of K-Mer Profiling in Genomics"
16.15 - 17.00	Closing and Drinks

Registration: by sending an e-mail to f.m.j.willems@tue.nl with your name and affiliation and subject line "Registration WIC MWM 2016"; registration deadline: January 25, 2016. The maximum number of registrations is one hundred (FCFS) due to the room capacity. The registration fee (including lunch and coffee/tea) is 50 euro, to be paid in cash upon entering the Zwarte Doos.

Directions to the venue (TU/e, Zwarte Doos): http://www.dezwartedoos.nl/nl/filmhuis/routebeschrijving

In case you have further **questions**, please contact one of the organizers:

Maurice Groten, TU/e, m.groten@tue.nl,
Dimitrios Mavroeidis, Philips Research, dimitrios.mavroeidis@philips.com, and
Frans Willems, TU/e, f.m.j.willems@tue.nl.

Abstracts

Joel Veness, "Compression to control"
This talk provides an overview of a recently introduced technique for policy evaluation in reinforcement learning, that converts any adaptive coding distribution for states into a value estimate for a given policy. Some of examples of this technique will be shown for real-time control of Atari video games from raw pixel input.

Joris Mooij, "Causal Discovery and Prediction from Big Data"
The discovery of causal relationships from experimental data and the construction of causal theories to describe phenomena are fundamental pillars of the scientific method. How to reason effectively with causal models, how to learn these from data, and how to obtain causal predictions has been traditionally considered to be outside of the realm of statistics. Therefore, most empirical scientists still perform these tasks informally, without the help of mathematical tools and algorithms. This traditional informal way of causal inference does not scale, and this is becoming a serious bottleneck in the analysis of the outcomes of large-scale experiments nowadays. In this lecture I will describe formal causal reasoning methods and algorithms, that can help to automate the process of scientific discovery from data.

Bert de Vries, "Design of Signal Processing Algorithms through Probabilistic Inference"
Abstract: We propose a probabilistic modeling approach to the design of signal processing algorithms. The proposed method relies on a generative probabilistic model that reflects the problem statement as well as any solution constraints. Through (automatable) probabilistic inference, solutions are derived for (1) the executable signal processing algorithm, (2) the tuning parameter estimation problem, and (3) a (Bayesian) performance evaluation metric. All three solutions are realized as message passing algorithms in a factor graph representation of the generative model, which in principle allows for fast implementation on mobile device hardware. The methods are illustrated in the context of the design of a hearing aid algorithm.

Elena Marchiori, "Network Community Detection by Seed Expansion"
Abstract: The enormous growth of network data from diverse disciplines such as social, information and technological science as well as biology has boosted research on network community detection. In particular network community detection by seed expansion, also called local community detection, amounts to finding a group of cohesive nodes concentrated around a given node, called seed. In this talk we describe the problem and outline ongoing research on this topic.

Tim van Erven, "From Data Compression to Online Machine Learning"
Abstract: Machine learning algorithms are able to understand the voice commands you speak into your smart phone, to filter out spam from your regular e-mail, and to predict how much electricity an electricity company needs to produce early enough to adjust the production levels of their power plants. The more data becomes available to these algorithms, the better they should work... in theory. In practice, however, many of these algorithms need to solve optimization problems that take too long to solve or that no longer fit into memory when there is too much data, or they need to update their predictions every day to take into account more data. In recent years, this has lead to the rise of so-called online machine learning algorithms, which process the data one item at a time, in a streaming fashion. An important characteristic of these methods is that they do not slow down over time, and can therefore keep running indefinitely as more data become available. I will explain how these online learning methods are related to information theory, and how, in particular, they may be viewed as extensions of methods for data compression, which solve the (seemingly unrelated) problem of turning a large file on your computer into a smaller file that can be stored or downloaded more easily.

Jeroen Laros, "Applications of K-Mer Profiling in Genomics"
We describe a number of applications of a method that utilises profiles of k-mer frequencies made from raw sequencing data. This method was originally designed for the comparison of raw datasets without the need for a common reference sequence. Additionally, by analysis of the characteristics of the profile itself, technical artefacts, e.g., various problems with the sample preparation, contamination, etc., can be detected. Moreover, this method can be used for the deconvolution of metagenomic datasets. These datasets (mostly microbial communities) can be extremely complex as they can consist of several hundreds of species, many of which are unknown, so alignment free methods like ours are highly desired.