Eugene Wigner Colloquium

joint event of GRK 1558 and SFB 910

Dr. Philipp Kanehl

IBM Deutschland GmbH

"Data Science at IBM - The Toxic Comment Classification Challenge"

Machine Learning, Neural Networks and Big Data have been omnipresent buzzwords of recent years whose implications are leading to an advance of automated, data-driven decision making in industry. In my talk I will report to you from my ongoing odyssey of becoming a data scientist. I will elucidate how holding a PhD in physics is aiding me along the way, how validation, overfitting and entropy relate to machine learning and why "clean" data is not always easy to come by.

Toxic Comment Classification is the automated labeling of insulting, obscene or threatening comments in forums and was a recent project of mine. Within a Jupyter Notebook, we will successively build from simple classical Natural Language models such as TFIDF to more elaborated convolutional and recurrent neural networks to increase the accuracy of our model. It will become clear why blending both simple and complicated models will generalize the best to unseen data. Finally, we will see how hyper parameter tuning and ensembling yield ready-to-deploy models.

GRK1558 research training group

Thursday, 28.06.18 · 16:15h · EW 202

Technische Universität Berlin · Institut für Theoretische Physik · Hardenbergstraße 36 · 10623 Berlin
www.itp.tu-berlin.de/grk1558 · www.itp.tu-berlin.de/sfb910