Session 4: Data-driven Modeling and Learning
14:40 – 15:00, 15th, July

The Hashing Trick and Other Related Methods for Speeding-up Machine Learning
Ricardo Nanculef, Universidad Tecnica Federico Santa Maria

The Johnson-Lindenstrauss lemma (JML) states that every set of n points in a Hilbert Space of any dimension can be mapped into a space of dimension only $\mathcal{O}(\epsilon^2 \log n)$ in such a way that all distances between the points are preserved up to a multiplicative factor $1 + \epsilon$. This fact is very interesting in the context of current machine learning applications that need to manipulate lots of high dimensional objects such as images and text. It suggests that there exist methods to alleviate the curse of dimensionality and has indeed inspired powerful methods for dimensionality reduction and approximate computations in learning algorithms.

Among these methods, we have the random projections approach (RP), which is a surprisingly simple implementation of the JML, and the closely related *hashing trick*, which has been successfully employed to speed up support vector machines and other kernel methods. It can be shown that RP and the hashing trick are also closely related to some data sketches, that is, randomized data structures designed to obtain compact representation of massive data streams which approximately preserve some desired statistical properties.

This presentation provides a short introduction to these methods, explaining how they relate each other and highlighting several relevant considerations for putting them to use in machine learning.