# Recognition of adversarial inputs

**Background**: Adversarial perturbations were first discovered in the context of deep neural networks (DNNs), where the networks' gradients were used to produce small bounded-norm perturbations of the input that significantly altered their output. Methods for producing such perturbations and the resulting perturbed inputs are referred to as adversarial attacks and adversarial inputs. Such attacks target the increase of the model's loss or the decrease of its accuracy and were shown to undermine the impressive performance of DNNs in multiple fields. Relevant papers: “Explaining and harnessing adversarial examples”.  
**Project Description**: In this project, we aim to differentiate between regular and adversarial inputs by processing the inputs' signals.  
**Prerequisites**: Deep learning course  
**Supervisor**: Yaniv Nemcovsky (yanemcovsky@gmail.com)