**Title:** Learning and in high dimensions when most features are useless

**Mentors:** The project will be supervised by Dr. Alon Kipnis

**Description:**
This project considers classification problems with 2 or more classes when the number of features in each data point is large but most features are useless; we do not know in advance which features those might be. In this project, we explore two variants of feature heterogeneity in multiclass tasks denoted as one-vs-many and diversity pursuit. In one-vs-many, we assume that each class has a characteristic set of features that discriminate it from the rest. For example, birds have feathers and are thus different than dogs and cats. In diversity pursuit, we assume that certain features discriminate against all classes while others do not. For example, dogs, cats, and birds are all different in the shape of their feet, but not different in the number of their eyes. For each variant, we propose a method for feature selection that is optimal under a certain generative model in the sense that the resulting features minimize the generalization error.

**Background:**
Scarcity in training data for learning and classification is inherent in most applications. For example, suppose that we would like to discriminate between two breeds of dogs, but we only have limited new labeled data of the two breeds. A standard procedure is to use the transfer learning principle: build your model on top of a pre-trained model that performs a similar classification task, e.g., a neural network that was trained to discriminate between images of dogs, cats, and birds. If the new data is very scarce and the number of useful features is very small, then a fine-tuning supervised approach would likely fail. This is where feature selection techniques like those based on Higher Criticism thresholding (Donoho & Jin 2008) and others (Abramovich et. al. 2021) come in.

One interesting observation is that some features are class characteristics (beak for the “birds” class) while other features may promote diversity but are not strongly associated with a specific class (e.g., animal size). We aim to develop methods that can handle both cases.

**Requirements:** The scope of this project is relatively large and involves multiple layers. The most demanding part concerns the theoretical analysis of the methods, which requires strong mathematical statistics and high-dimensional probability skills. Considering the training process in these areas, this part is likely to be in the scope of a

Ph.D. Other layers of this project concern data science and experimental work involving real and simulated data.

**References**

Donoho, D., & Jin, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. Proceedings of the National Academy of Sciences, 105(39), 14790-14795.

Abramovich, F., Grinshtein, V., & Levy, T. (2021). Multiclass classification by sparse multinomial logistic regression. IEEE Transactions on Information Theory, 67(7), 4637-4646.