I am currently a semester project student at ETH Zurich's Computational Genomics Lab, where I am exploring augmentation techniques for single-cell RNA-sequence data (scRNA-seq). This type of data is hindered by several challenges like limited samples, batch biases, and biases due to sequencing technologies. Several augmentation methods have gained traction in recent years, but they primarily rely on deep learning frameworks such as GANs or VAEs. In computer vision and natural language processing, some of the most successful augmentation techniques are grounded and non-ML based (e.g. cropping, rotation, paraphrasing etc), so the aim of my project is to develop biologically-grounded augmentation techniques to help in downstream tasks. Addressing biases and providing high quality data is a ubiquitous topic in machine learning, including for applications like content recommendation at Netflix.

I was previously a technical student at CERN, where I worked on different projects. I spearheaded the development of FITCompress, a new algorithm that combines quantization and pruning of neural networks for deployment in constrained environments (e.g. mobile devices, FPGAs). Very few methods in literature aim to combine quantization and pruning. The number of combinations grows exponentially. Current methods rely either on Bayesian learning techniques or deep reinforcement learning. These are typically slow, memory-hungry and/or rely on arbitrary priors. In contrast, I proposed an algorithm that uses the Fisher Information Trace as a heuristic (previously shown to be a robust estimate of how much damage quantization does to a network compared to the Hessian) and extended this heuristic to encompass pruning. Given a pre-trained 32bit model, a memory constraint, quantization and pruning schedules, the algorithm chooses whether to quantize a single layer (e.g. 32 bits to 8 bits) or to keep pruning the network globally (e.g. increase sparsity). Each step is taken iteratively to "damage" the network minimally and we only retrain the model once. I showed this method to provide state of the art compression-performance tradeoff in image classification, object detection and natural language processing tasks. Deploying deep learning models in resource-constrained environments is likely a big challenge Netflix is working on as I saw in the blogpost titled "**Reinforcement Learning for Budget Constrained Recommendations**".

I also worked on a software package, where I leveraged DBSCAN (temporal clustering) and K-Means (spatial clustering) to cluster beam events and facilitate analysis of data at the Proton Synchrotron Beam Gas Ionization experiment. Finally, as part of the GPT for Accelerators project, I provided the first functioning prototype of a LLM (Meta's Code LLaMa) fine-tuned on CERN's own generic optimization framework. Clustering algorithms can be immensely useful in recommender systems, which are at the heart of Netflix research. While my project with LLMs focused on code generation, LLMs can certainly be used for natural language understanding to drive context-aware recommendations, such as those provided by Netflix.

As a summer student at CERN, I explored generative models (VAE, GANs and normalizing flows) for fast and accurate simulation of proton-proton collision events, necessary for the upcoming High Luminosity Large Hadron Collider. It was determined that although GANs typically provide more accurate synthetic data, VAEs lead to more practical (e.g. simpler and quicker training in constrained environments) that is sufficient enough for this particular use case. I imagine Netflix is deeply interested in generative models like autoencoders, as evidenced by several articles: "**Autoencoders that don't overfit towards the Identity**", "**Embarrassingly Shallow Autoencoders for Sparse Data**" and "**Variational Autoencoders for Collaborative Filtering".**