Internship Project Description: Summer 2018

Project Title: How can we learn more from high-dimensional data?
Supervisor Name: Caleb Weinreb
Lab PI Name: Allon Klein

Project Description: Every cell in your body has an almost identical genome, and yet these cells perform wildly different functions and come in many shapes and sizes. How can a single genome give rise to such diverse cell states? New tools for simultaneously measuring the detailed molecular composition of very large numbers of individual cells may bring us closer to answering this question. But this type of data is entirely new in biology, and so the algorithms and computational techniques required to interpret these measurements are still in their infancy. The goal of this summer project is to help develop computational tools that can make sense of high-dimensional single-cell measurements.

The Klein lab uses single-cell RNA sequencing to understand the ways cells change their state over time. In a typical experiment, we measure the expression of thousands of genes in tens of thousands of individual cells. The end product of these measurements can be thought of as a cloud of points in high dimensional gene-expression space. Each dimension of the space corresponds to a gene and each point to a single cell. The way cells are distributed in this space reveals properties of the biological system. For example, dense clusters correspond to groups of cells that are in a coherent state (e.g. a cell type), whereas elongated - or even branching - point clouds correspond to dynamic states across space or time.

Existing computational tools offer a good first look at the shapes that single cells make in high dimensional gene expression space, but they clearly leave a lot of insight on the table. We simply are not very good at visualizing very high dimensional spaces. The specific goal of this project is to understand the principles that cause existing computational methods to fail, and to use this knowledge to develop new tools that better capture the full complexity of high-dimensional datasets.