Desired Needs

- The project should be able to produce a proof-of-concept demonstration of machine-learning based classification of flow cytometry data into cancer/non-cancer
- The project should assess the performance of normalization tools used for flow cytometry data
- The project should assess the feasibility of using UMAP and dimensionality reduction techniques to visualize sample heterogeneity

Constraints

- The project is limited by compute time and resources, as some subprojects require significant time and/or computational resources to run analyses
- Since the project works with health-based data, the team interacted with only de-identified data that was pre-transformed

Engineering Standards

- Several standards related to coding quality, documentation, reproducibility, and availability were adhered to, allowing the project code to be saved, used, and iterated upon reliably
- Accessibility and acknowledgement was emphasized for the project code

Ethical, Environmental, or Societal Concerns

- Due to excessively long computational times, the team limited analyses to subsampled data for large data subprojects
- The data was already de-identified, so there is no risk of data leak and patient re-identification

Active Teamwork and Leadership

- The team met several times per week to discuss the independent subprojects assigned to each member, collaborating and providing feedback for deliverables
- The team used a calendar-based approach to accommodate deadlines and keep track of goals

Motivating Factors

- New knowledge was required in every aspect of the project, including understanding the mechanistic behavior of machine learning models in order to apply models to data, and understanding the limitations of tools such as cyCombine for normalization.
- Since the project consisted of independent subprojects, self-initiation and persistence in development was required and made easier by consistent meetings with the team mentor
- As the project had the potential to impact the lives of patients with cancer, utmost care was taken to ensure accountability in reporting and documentation

Innovation

- The most innovative idea for this project is the development of the hands-free distribution statistic machine learning model, which works with raw data and can achieve relatively good performance on minimal information under minimal interaction