Summer Internship Assignment: Data Science at Melio

We are Melio (from the Latin melior, meaning "to make better") — a Bay Area-based, early-stage, VC-funded startup with a bold mission: to transform the way infectious diseases are diagnosed. Our goal is to develop a rapid, affordable, and comprehensive test that can identify all common bloodstream infections. It's an ambitious challenge, but one with the potential to make a huge impact on global health.

We believe solving hard problems requires people who are intellectually curious, resilient, and excited by the unknown. That is why we are reaching out to you.

We are currently recruiting full-time Data Science Interns for Summer 2025, with the potential to continue part-time (20 hrs/week) into the Fall. Our projects span:

- Bioinformatics + AI: marrying domain knowledge with machine learning to unlock novel insights
- Data Engineering: building robust, scalable pipelines for biological and clinical data
- Imaging + Signal Processing: automating and analyzing complex optical outputs from microfluidic chips and workflow optimizations for our microbiology workflow.

We have put together a few assignments for you to work through — not just to showcase your skills, but for us to get a sense of how you approach problems, how you think, and whether this is the right fit on both sides.

Let's build something meaningful together.

Assignment Question 1

You are given five CSV files. Each file contains multiple time series samples that are noisy versions of a typical pattern. Each file corresponds to a distinct “truth” (ground truth category)

What types of exploratory or statistical analyses would you suggest to assess the classifiability of these truth categories? Please describe your approach and justify your choices.

More specifically, how well-separated or distinguishable are the five classes? Are there any patterns or features in the signal shape that appear consistently within a class or distinct across classes? Please don't forget to add any limitation you can think of with this method. What are the limitations of drawing conclusions about separability based solely on the exploratory visualization and descriptive analysis?

Instructions to access data from AWS :

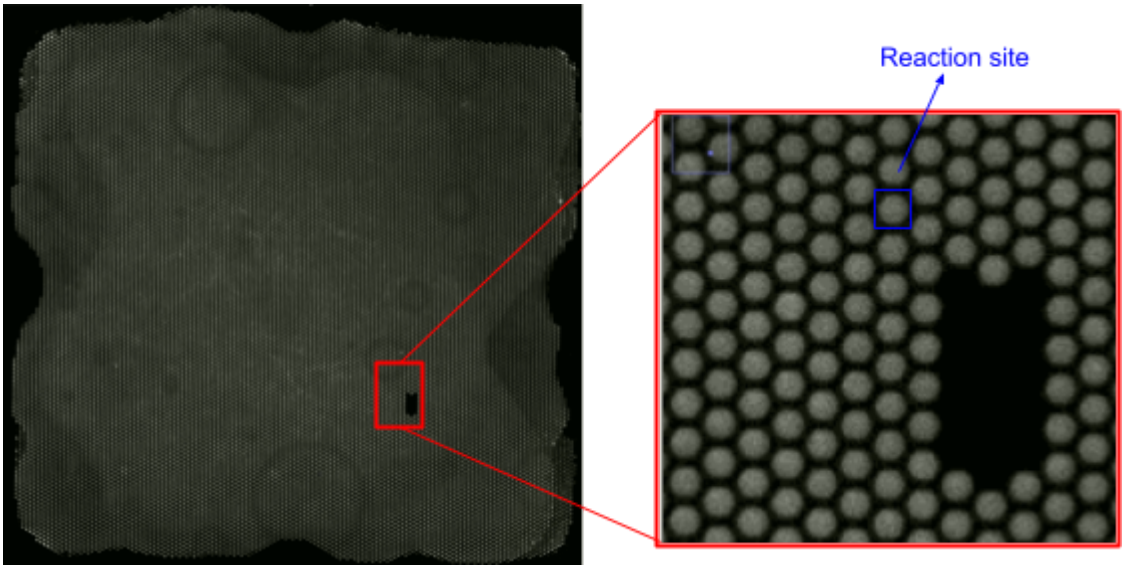
You can find the images in an AWS bucket. The credentials and details of the AWS bucket are provided below.

- AWS bucket name: s3://dna-image-repo
- File in S3 bkt with data for Assignment-1 - s3://dna-image-repo/melt-csv-sample.zip
- User name data_share_1
- Access key ID: AKIAZ4HHHCZXAKXOPSAB

Assignment 2 Registration and Tracking of Reaction Sites on a Microfluidic Chip

Here is a microfluidic chip that contains ~20,000 independent reaction sites, each capable of receiving analytes for biochemical reactions.

Below is an image of an example chip showing these densely packed sites arranged in a regular hexagonal grid.



You are provided with:

- Above described microfluidic chip image containing thousands of reaction sites
- **DXF file** representing the **design mask** of the chip — i.e., the intended physical layout of all reaction sites
- **(Stretch goal)** A **time series of images** of the same chip acquired during an experiment, with minor shifts in:
    - x, y (translation)
    - θ (rotation)
    - z (zoom/focus)

Design a script that:

1. Registers the true mask (from the DXF) to the imaged chip, allowing each reaction site in the image to be mapped to its intended physical identity (e.g., well #0, well #1, …)
2. Identifies filled vs unfilled reaction sites, and produces a map of active wells, preserving spatial identity (i.e., which well in which location on the chip
3. Stretch goal - from timeseries: Aligns each image in the time series to correct for small shifts or rotations over time

Instructions to access data from AWS :

You can find the images in an AWS bucket. The credentials and details of the AWS bucket are provided below.

- AWS bucket name: s3://dna-image-repo
- Folder in S3 bkt with dataset for Assignment-2
  s3://dna-image-repo/sample-control-imgs-E0AIZP/
- User name data_share_1
- Access key ID: AKIAZ4HHHCZXAKXOPSAB