**PSPG Workshop -- 2020 -- Day 1**

The overall goal of this workshop is to apply machine learning methods to the DTP-NCI60 cancer drug screening dataset (<https://dtp.cancer.gov/discovery_development/nci-60>) to predict which compounds are most relevant for treatment given the tumor data provided to you.

The NCI60 dataset consists of approximately 27k compounds screened against 60 individual cancer cell lines, originating from 10 tissue subtypes.

In order to determine which compounds are optimal for treatment, you will need to identify which cell line is most relevant to your tumor. Before the second portion of the workshop, your goal is to explore the datasets provided by the NCI60 and to make a decision on which cell line(s?) to focus on.

You will present 1-3 slides (max) on how you made your decision (eg. What you tried, what didn't work and why, and what you ended up doing) at the end of the workshop.

**Exploring the Data**

The NCI's genomics and bioinformatics group has created a web portal consolidating all the data and metadata related to the screen that is publicly available and convenient to access.

The sites homepage is located here: <https://discover.nci.nih.gov/cellminer/home.do>

We encourage you to examine the upper blue tabs, especially the ones labeled "**Data Set Metadata**", "**Cell Line Metadata**", and "**Download Data Sets**"

Some considerations to make while exploring these data that may help in preparing your slides might be:

* What does the data convey in a general sense?
* Is the data lacking anything crucial?
* Is a particular data type more useful raw or pre-processed. If preprocessed, what did the preprocessing do?
* Is all the data there, but the labels provided require conversion to be of any use, and conversion would be such a huge time-suck that it’s kind of stupid this data exists the way it exists?
* What data (tumor or NCI60) do we wish we had, in order to make something we tried work?
* What are the properties of the cell line you chose?