**Expected format of the lab report**

**General Instructions** – Keep your Jupyter notebooks short and to the point. Feel free to explore and try lots of different things, but only turn in the "final" product of your exploration of the dataset. The only require sections are (1) introduction, (2) coding section, and (3) a summary. Feel free to break up the coding section with markdown text and comments to increase our understanding of what you did and why. Target a printed length of ~5 pages, including results. Excessively long report will result in penalty. The deliverables must be in the form of jupyter notebooks, **do not** convert it to PDF files. Please remember to cite inline any place where you have borrowed code from stackexchange or made use of AI tools such as Github Copilot and ChatGPT.

**Submission introduction** – Upload and commit (if any) your files onto your GitHub Repo. Make sure you arrange your repo in a clean and well-organized way, so anyone could easily navigate through your repo.

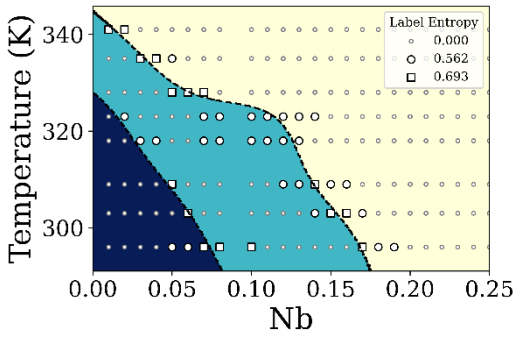
**Introduction** (markdown) – You will be performing clustering on a high-throughput x-ray diffraction dataset for a binary composition spread of $VO_2$ – $Nb_2O_3$ as a function of temperature. The format of the data is that it comes in two files:

The first file, labeled "VO2-Nb2O3 Composition and temp Combiview.txt", contains 2 columns and 353 rows. The first column is labeled "V" and contains the at.% of V in the film, the second column is labeled "temp" and is the temperature at which diffraction was measured.

The second file, labeled "VO2-Nb2O3 XRD Combiview" contains 3841 columns and 353 rows. The row 1 here provides the 2-θ angle for diffraction and the subsequent rows are the diffraction intensity at each 2-θ.

The files are coordinated such that the first rows are the column names and each subsequent row aligns the composition and temperature of a measurement with its diffraction pattern.

$VO_2$ is known to undergo a phase transformation from monoclinic to tetragonal as a function of alloying and temperature. Your goal is **<u>to use the unsupervised tools we developed in class to nail down the different phase regions</u>**. For your reference, the average of 5 human experts is summarized in the figure below (where the different colored regions represent monoclinic, monoclinic + tetragonal, and tetragonal):




**Coding Section** (Combination of markdown and code blocks) –

1. Open the data, perform basic data preparation

2. Demonstrate your ability to plot and navigate the data by:
    a. Generating a composition versus temperature plot with each measurement point clearly marked
    b. Generate an x-ray diffraction versus temperature plots for a constant V composition of 99 at.%
    c. Generate an x-ray diffraction versus composition plots for a constant temperature of 23
3. Perform some form of dimensional reduction on the diffraction data
    a. PCA if you like or t-SNE either is fine
    b. Use total explained variance to drop noise
4. Plot the diffraction patterns in reduced dimensional space with a false color plot (color the spots by composition or temperature)
5. Perform a series of clustering analyses to attempt to get phase regions as similar as possible to the human expert phase mapping exercise
    a. Summarize your results by generating a figure similar to the human expert phase map above
    b. Rationalize based on the image from the lab documents, why a specific tool was selected

**Summary** – Tell us what worked (and maybe what didn't) and give a guess as to why. Teach us 2 lessons you learned about your data set. (E.g. when performing PCA for dimensional reduction, if we plot PC1 and PC2 with the dielectric constant as a color map we observe that there is a cluster that mostly has a value of 0). Provide a physically plausible explanation for what material "archetypes" are contained within the clusters and

**Things to keep in mind:**

**Readability of Code**

- Used Markdown to separate big steps (e.g. PCA for dimensional reduction to clustering)
- Used reasonable in-line comments
    - Concise descriptions
- Notebook follows a logical ordering
    - The cells are run in order
- Thought was put into making certain the markdown text was well formatted
- Plots used are readable and appropriate to the learning task

**Correctness of Python Coding**

- The code can be run w/a single click w/o errors
- Used appropriate (and not excessively complex) methods to transform data and fit models

**Materials Science Machine Learning (semester long objective, just keep them in mind)**

- Showed thoughtfulness in parsing the data set
- Chose an appropriate featurization of the materials dataset
- Successfully performed dimensional reduction on their dataset
    - Checked to make sure dataframes were appropriately formatted

    - Ensured that pre-transformed and post-transformed dataframes were consistently ordered (or addressable via dataframe/array operations)
- Successfully performed a series of clustering operations on the reduced dataset
    - Rationalized the technique used and the number of clusters arrived at
- Came up with a creative method of featurizing and/or performing dimensional reduction and/or clustering
- Identified a materials science reason for the realized clusters
    - E.g. Split on metal vs. non-metal