2) Written challenge:

Say we are engaging with a new customer, X-Materials. X-materials has hundreds of researchers working on dozens of materials development projects. We have signed a two year contract, the first year of which will be devoted to getting their data onto the Citrination platform.

X-Materials has a few main goals for the 2 year contract:

1. Bring their data onto a centralized platform allowing for easy upload, access and storage of their data
2. Organize their data in an intelligent and meaningful way
    1. This organization should allow them to build contextual pedigree around their data
3. Use the Citrination platform to build machine learning models to guide their materials development efforts

As the Data Engineer on the team, propose a detailed project plan describing how you would approach and execute on understanding, organizing and capturing their data, and the challenges therein.

Answer:

To foster a collaborative partnership with new customer, X-Materials, effective communication with different business units, product engineering teams and research scientists is crucial, as production engineering teams and research teams might have different priorities across the corporation.

After obtaining an overview of the materials categories and research development directions, bringing customer's data onto a centralized platform requires data collection and data representation. Data collection may require initial preprocessing to handle missing elements, error detection and non-digitalized historical documents. Utilizing state of the art deep learning methods in image processing and text mining areas can also help extract information during data collection step. Moreover, certain characterization or property data may be missing in historical data due to instrumentation limitations. Some human and measurement errors exist in documents with experimental data. These all pose some challenges for the dataset when building the machine learning model.

Encoding structures and properties from the raw data in an intelligent and meaningful way requires insight into the underlying scientific problem, analysis of experimental data and the codification of chemical intuition. The encoding for molecular systems such as materials synthesis and characterization will be significantly different from crystalline solids representation. The more suitable the representation of the input data, the more accurately an algorithm can map it to a property of interest. The other complexity to solve is representation of structural, spectral data and building machine learning models with it. For instance, most spectral data such as FTIR, NMR were collected in time domain but presented in frequency domain. When organizing the data, besides the complex physical science models, data related economic factors such as cost, purity, time and toxicity will also be taken into considerations.