Background: Herbal supplements are very popular in the US and elsewhere, yet these bioactive food additives are not regulated (compared to food as well as prescription and over-the-counter drugs). United States Government Accountability Office Report on Contaminants found in Herbal Supplements (https://www.gao.gov/assets/130/124769.pdf) found that many herbal supplements sold in US markets contain traces of harmful contaminants (arsenic, lead, cadmium). They also note that many health claims are bogus.

Idea: Use DL tools to estimate the contribution of dietary supplements into the body burden of these contaminants in the US population. (Note: Body burdens of contaminants are measured by concentrations of these contaminants in blood, urine, hair, etc. These data are referred to as biomonitoring data.)

Why is this a deep learning project? Because dietary and biomonitoring data are multivariate and sparse, and the relationships between diet/supplements/biomonitoring test results is highly non-linear. Here is last year's ICLR paper on a related topic:

- Diet Networks: Thin Parameters for Fat Genomics

More specifically, the idea is to build NN that takes sociodemographic and diet data, and predict a vector of real-valued contaminant concentration measurements. So this is a regression problem, not a classification problem. With this model, we could create a simulation experiment, in which we predict contaminants sans herbal supplement for each individual, and report on the aggregated effect at the US population level.

Data: National Health and Nutrition Examination Survey (NHANES, 1999-2016) collected by US Centers for Disease Control and Prevention. Every two years the NHANES program collects data on demographic and socioeconomic characteristics, biometrics, diet (including Rx and herbal supplements), health status, and biomonitoring results, including tests for contaminants (pesticides, heavy metals, dioxins). The number of participants in each survey cycle is 5000, so there are at most 45,000 data points for the cycles covering 1999-2016. Missing data are likely, so the actual number of usable data points may be lower.

Notes:

- Expected modeling challenges will include:
    - Tackling missing data (e.g., many people have BMI and income information missing) and partial data on contaminant body burdens in a way that does not shrink our dataset to an unusable size.
    - In terms of modeling, we at minimum should experiment with appropriate loss functions for the data. Contaminant concentrations are highly skewed, noisy, and very low concentrations can't be detected. So MSE loss may be too simplistic. I had worked with a censored log-normal loss for these data, but log-normal could be a very restrictive assumption.

    - We could also explore variational inference or GANs, to have an NN tackle the question of proper loss. But I would explore this only after we had done a simple MLP version.
    - To generate estimates that a representative at the population-level we would need to take into account NHANES survey design characteristics. For us, this boils down to the proper sampling/aggregation strategy using survey weights.
- Dataset preparation will take work. NHANES data comes in multiple smaller datasets that correspond to portions of the questionnaire and/or lab testing. There will be variations in the format over time, but hopefully this can be managed. For reference, see this document that summarizes which parts of the data are available for what years.
- Producing a revised literature review for the mid-term report. NHANES dataset has been studied extensively in public health, but not with DL tools.
- Anna has a high degree of familiarity with most parts of NHANES data (authored two papers on NHANES biomonitoring topics in public health journals).
    - If we do get interesting results, we could submit the work to a public health journal (if not ICLR).

Possible project task structure/sequence:

1. Download/process data for one survey cycle to get a dataset for practice. We can split who processes what part of the survey.
2. Develop/train version 0 NN model, including simulations.
3. Write the progress report (with a very light-touch related work section)
4. Process and normalize the remaining data.
5. Revise/revit/enhance the version 0 model
6. Finalize the report
