**Data description:**

1. All relevant materials are in a shared folder (link: [Nutriomics challenge and data](https://16515-my.sharepoint.com/:f:/g/personal/osterman_sbpdiscovery_org/Ek85l8Bu0fVPqGj9Jp89evsBvhvdr4RuQRMrYjElVEo8XQ?e=jdOOaG))

Graphical user interface, text, application, email

Description automatically generated

1. All data tables are located in a sub-folder “data”

Graphical user interface, application

Description automatically generated

The files contain the following information:

****1. File “metadata.csv”.** It helps you to choose samples (from the total of 2,827) that you would use as RSC (group “NORMAL”, total 1,414) and as TS (group “DEVIANT”, total of 142). At this point, we stay away from group “OTHER” (just ignore these samples). *Note that:*

* 1. All 2,827 samples were selected from American Gut Project (AGP) by stringent data quality criteria and metadata.
  2. For the purpose of this assignment, the 3 groups were defined based on the computed score as described below.

The score (W) is an abundance-weighted deviation of each taxon's (species or grouped ASVs - rows in the table “taxonomy.csv”) percentile in the sample (calculated for top 200 species). Specifically:

1. For each taxon's abundance, calculate its percentile P\_i.

2. Next, calculate the deviation from the median, i.e., S\_i = abs(P\_i-50) / 50. This will give a uniqueness/abnormality score S\_i for each taxon.

3. Using taxa abundances A\_i, calculate weighted score W, i.e., W = S\_i \* A\_i summed over i.

Groups:”NORMAL”, for W below the median (of all W's); “DEVIANT”, for W above 95th percentile (of all W's) “OTHER” all other samples

**2. File “taxonomy\_400.csv”.** Provides RA% data for all 2,827 samples (columns) and the total of 400 **taxons** (*for simplicity, I suggest to use this term instead of ASV just to be more accurate, however, it means exactly the same – ASV cluster*).   
…

*Note that this is a subset of top 400 taxons from the full contained 2,120 taxons. As a result the tital RA% in some samples is <100% (they are sorted left to right in the descending order of the total RA%). Feel free to exclude 192 rightmost samples where the total RA% is <75%. Others can be used safely (and you may renormalize them all to total 100% if you want to).*

…..

**3. 4 files starting with “nim-…”** provide Nutrient Impact Matrix for four types of nutrients (they could be all combined into one table). Each table has the same 400 taxons (rows) and NI values for respective nutrient shown by abbreviations (all unique across 4 tables). For practical purposes of this assignment, please include the following nutrients:

- **Table “nim-aminoacids\_400.csv”** reflects the requirement for amino acids (auxotrophy, eg inability to synthesize needed amino acids) as protein building blocks. You may see that the average NI of most amino acids is very low (meaning low impact since most taxons are prototrophs for these amino acids, eg they can synthesize them), **I suggest to limit the choice of amino acids as nutrients by the first 6:**



* - **Table “**- **Table “nim-aminoacidsD\_400.csv”** reflects the ability to degrade and utilize (catabolize) amino acids as a source of carbon, nitrogen and energy (not as protein blocks).



- **Table “**- **Table “nim-vitamins\_400.csv”** reflects the requirement for vitamins (auxotrophy, eg inability to synthesize these vitamins) as precursors of essential metabolic cofactors. As you the average impact and variations of all these miucronutrients are quite substantial. Use all of them.



- **Table “**- **Table “nim-sugars\_400.csv”** reflects the ability to utilize (catabolize) carbohydrates (mono-, oligo- and polysaccharides) to use them as a source of carbon and energy. There is total of 56, and I suggest to include them all, even those that have low average impact as they may provide selective solutions, while the most impactful (like glucose or fructose) may not be variable enough to help with differential effects on taxons.



I also added a file **“Phenotypes\_legend.xlsx”,** which is not essential for the algorithms and computing. It provides explanation of binary phenotypes and abbreviations for all nutrients. Just if you are curious to know what they are and why. Looks like that:



*Email me with any questions and good luck! Andrei*