# Healthy Aging Signal Research
# Summary Report Milestone 2

**Partner:**
**Merck & Co.**

**Group Members:**
**Eleonora Shantsila**: eshantsila@g.harvard.edu,
**Yaxin Lei**: yaxin_lei@g.harvard.edu,
**Aaron Jacobson**: aaronjacobson@g.harvard.edu ,
**Daniel Cox**: daniel_cox@g.harvard.edu

Hello Antong and Greg,

As we may have mentioned at our last meeting, at this point in the semester we are required to submit a technical report for Milestone 2 that provides a description of our work since Milestone 1 (March 2nd). The new report is attached. In its section headed Milestone 2: Modeling with DNA Methylation Data you will find a description of much of our recent work, which as you know has focused on predicting age from DNA-methylation data. More specifically since Milestone 1 we have accomplished the following.

- Identified a good source of DNA methylation data, the EWAS datahub.
- Developed methods to download and analyze the large datasets found at EWAS.
- Examined various ways of imputing the missing values in our data.
- Built linear models that predict age from DNA methylation data from blood using the entire ~400, 000 methylation sites (cpgs) in the data.
- Investigated partial least squares regression as a tool for feature selection.
- Identified the methylation sites that appear to be the most useful predictors of age via a XGboost-cross-validation-based feature importance analysis and by statistical testing.
- Demonstrated that models built with just 100 cpgs work slightly better than those built with ~ 400,000.
- Developed a neural network model that can predict age from blood data to an rMSE of less than 5 years.
- Determined that models built with blood data can not be used effectively with methylation data from other tissues, including brain, and breast tissue.
- Built separate models for brain and breast tissue.
- Started to investigate to what extent cpg features can be shared across tissues.
- Started to investigate differences in model performance between healthy and unhealthy individuals.

Currently, we are also working to make our methods of data manipulation more efficient. We are making some of our previous analyses more rigorous, and we are turning our attention to examining which methylation sites might be most predictive of unhealthy aging. We thank you very much for your guidance over the past several weeks, and we look forward to our next meeting

Sincerely,

Daniel Cox
Eleonora Shantsilla
Yaxin Lei
Aaron Jacobson