**Lab 2: High-dimensional exposure data in the Exposome dataset: Gradient Boosting Machines and XGBoost**

We continue with the same data from the previous lab. The aim of the second lab is to introduce you to the practical parts of supervised machine learning techniques that were discussed on day 2/3. Specifically:

* Tune, fit, and interpret tree-based ensemble models (GBM and XGBoost) for our continuous outcome (birthweight)
* Interpretable machine learning for variables importance (VIP), partial dependence (PDP), individual conditional expectation (ICE), and interaction diagnostics

The lab consists of two parts, each with multiple steps.

1. **Part 1: Gradient Boosting** (**gbm, vip, pdp, iml)**

* Prepare data and outcome formula
* Tune hyperparameters (learning rate, tree depth, number of trees)
* Fit a final GBM model for birthweight
* Explore variables importance (VIP)
* Visualize exposure–response patterns using PDP/ICE plots
* Assess exposure–exposure interactions

1. **Part 2: XGBoost** (**xgboost, vip)**

* Split data into training and validation sets
* Tune hyperparameters and select best model
* Fit a final XGBoost model for birthweight
* Examine variables importance

The final part of the lab will:

* compare results between GBM and XGBoost
* reflect on differences vs penalized regression (from Lab 1) and identify a potential final model and set of conclusions
* discuss strengths and limitations of boosting methods in epidemiology