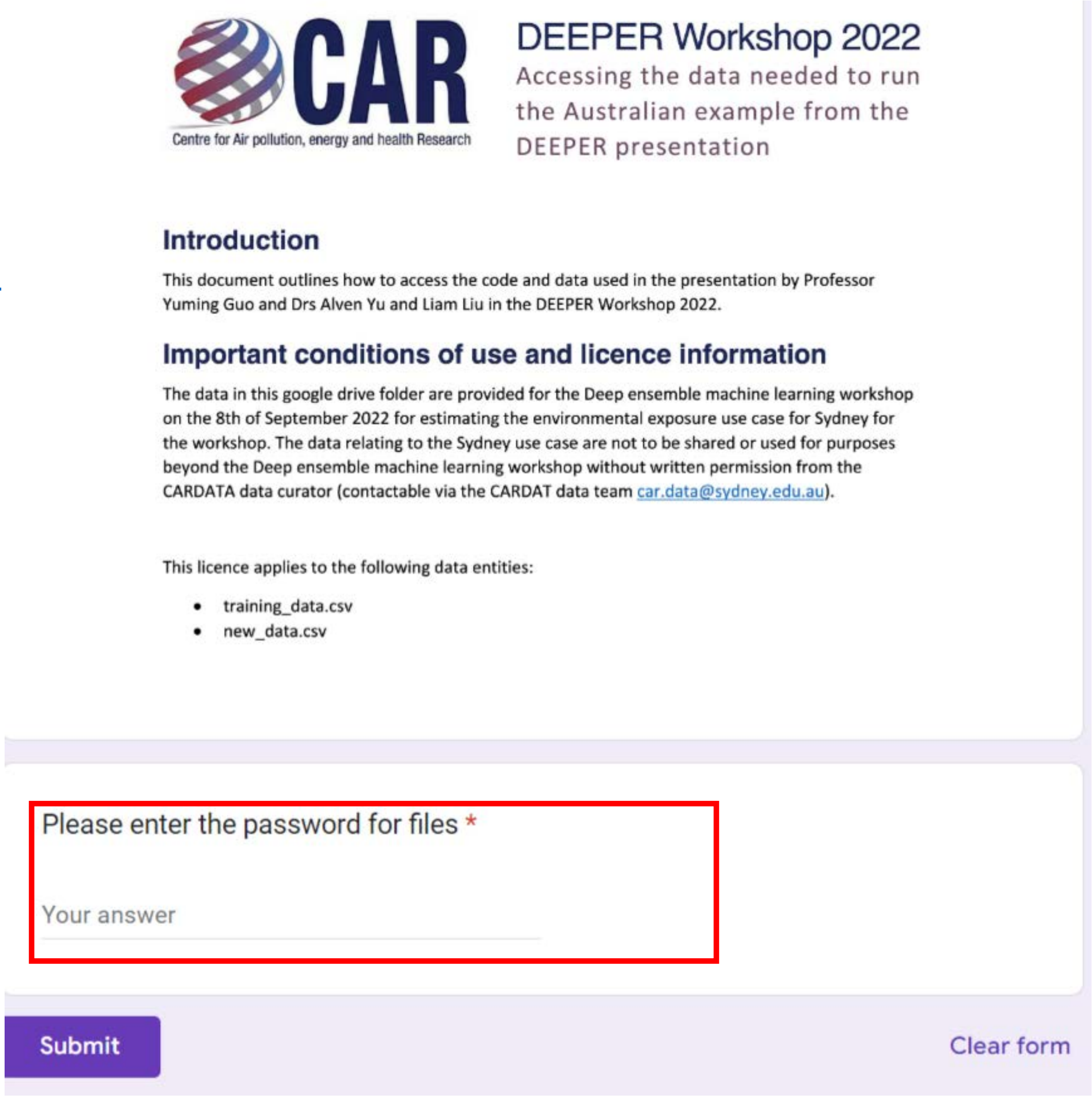
• Before we start…

1. Please download the example CARDAT data and the deeper GUI from: https://forms.gle/4bXqaj2zUsT2mMMA9

   Password: **DEEPERworkshop2022**

- ‘training_data.csv’, ‘new_data.csv’ for training models and estimating new grid data
- Please decompress the zipped 'deeper GUI'



CAR
Centre for Air pollution, energy and health Research

DEEPER Workshop 2022

Accessing the data needed to run the Australian example from the DEEPER presentation

Introduction

This document outlines how to access the code and data used in the presentation by Professor Yuming Guo and Drs Alven Yu and Liam Liu in the DEEPER Workshop 2022.

Important conditions of use and licence information

The data in this google drive folder are provided for the Deep ensemble machine learning workshop on the 8th of September 2022 for estimating the environmental exposure use case for Sydney for the workshop. The data relating to the Sydney use case are not to be shared or used for purposes beyond the Deep ensemble machine learning workshop without written permission from the CARDATA data curator (contactable via the CARDAT data team car.data@sydney.edu.au).

This licence applies to the following data entities:

- training_data.csv
- new_data.csv

Please enter the password for files *

Your answer

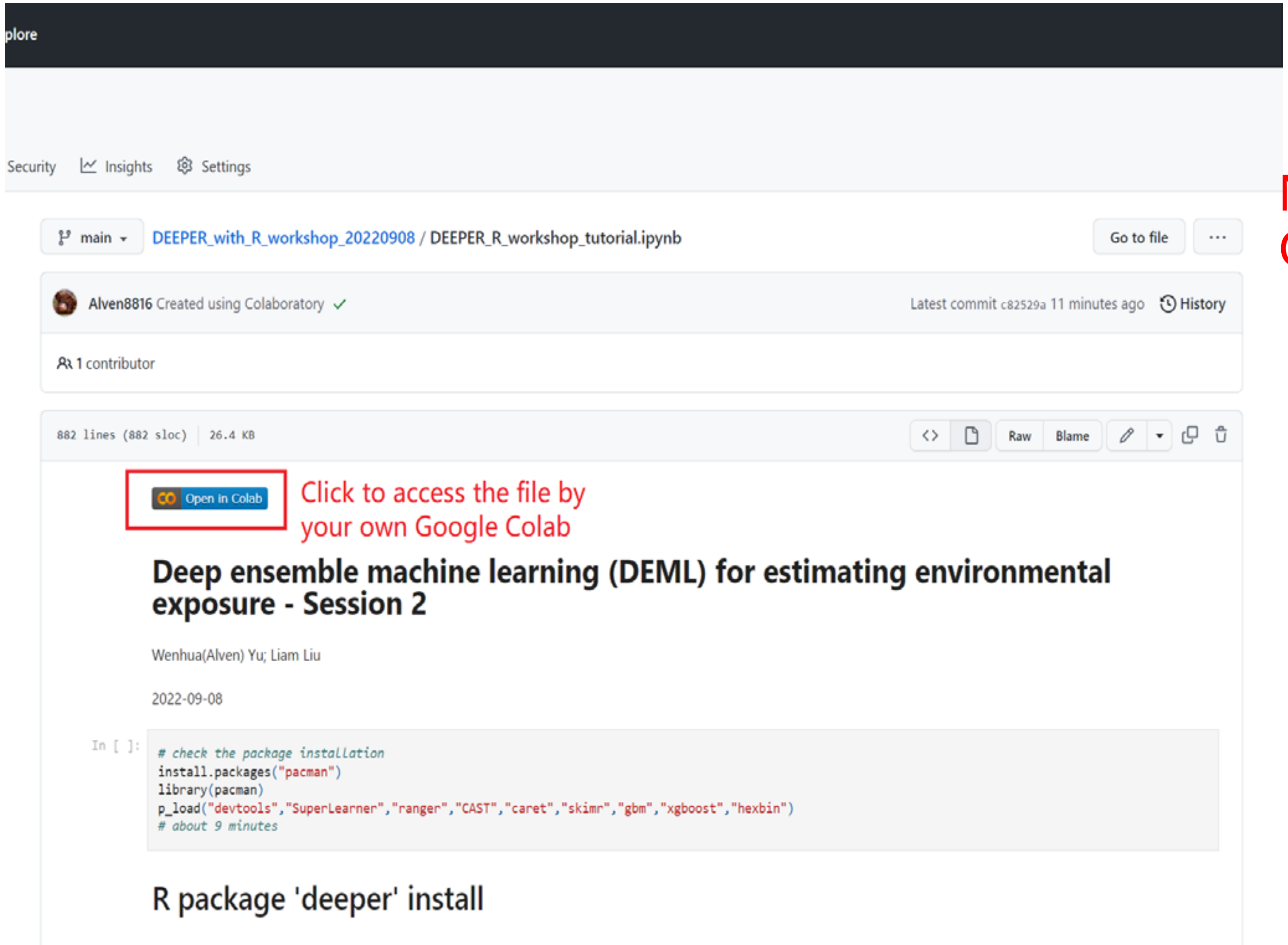
Submit

Clear form

MONASH University

• Before we start…

2. Get access to the tutorial code from the Google Colab: https://github.com/Alven8816/DEEPER_with_R_workshop_20220908/blob/main/DEEPER_R_workshop_tutorial.ipynb



Note: Please resave a copy to your Google drive to run the tutorial code.

MONASH University

• Before we start…

3. Install all packages required for this tutorial (10 mins)

DEEPER_R_workshop_tutorial.ipynb

File Edit View Insert Runtime Tools Help Cannot save changes

+ Code + Text Copy to Drive

Deep ensemble machine learning (DEML) for estimating environmental exposure - Session 2

Wenhua(Alven) Yu; Liam Liu

2022-09-08

Click here to run code

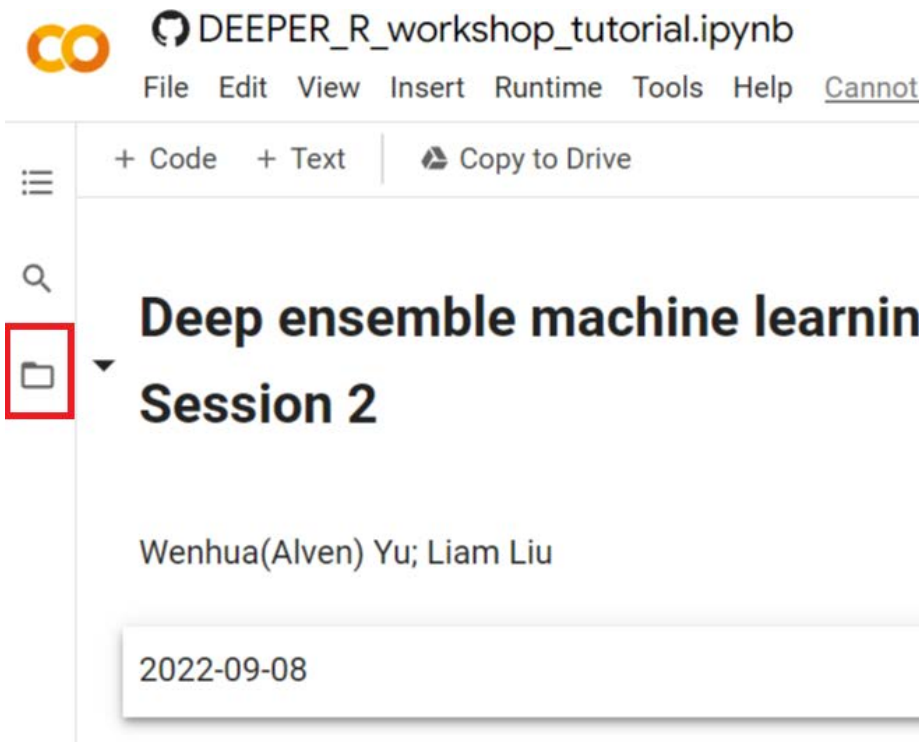
```
# check the package installation
install.packages("pacman")
library(pacman)
p_load("devtools","SuperLearner","ranger","CAST","caret","skimr","gbm","xgboost","hexbin")
# about 9 minutes
```

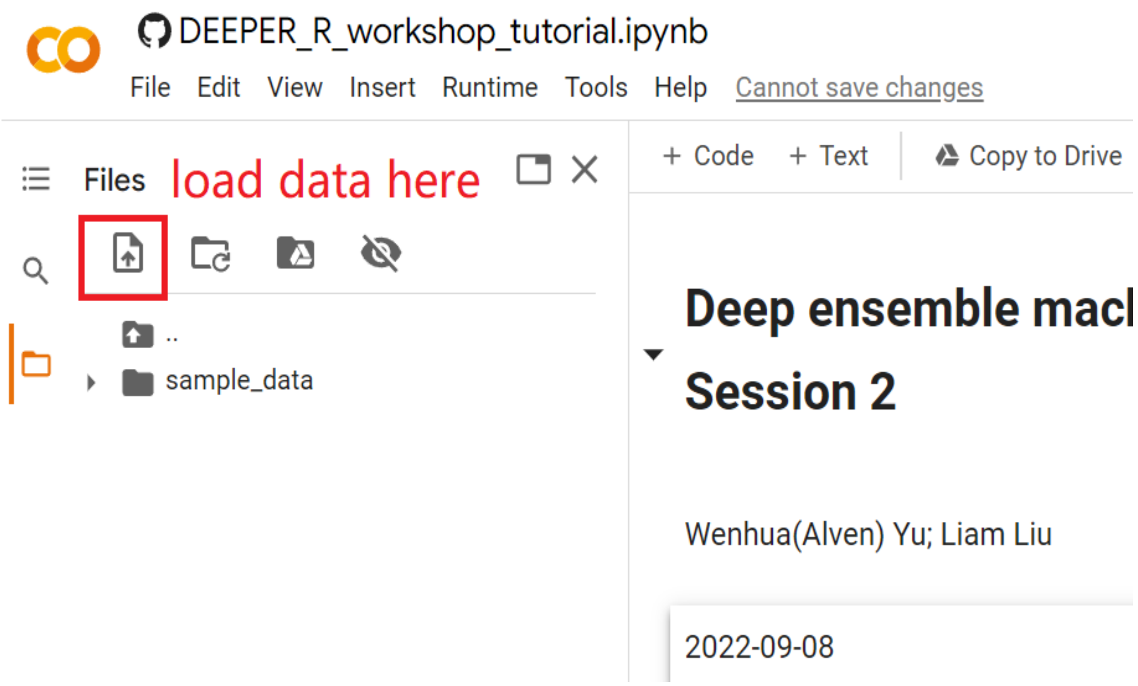
MONASH University

•Before we start...

4. Load the data into the Google Colab

Click file icon in the left side > Upload to session storage > choose the files > OK





MONASH University

DEEP ENSEMBLE MACHINE LEARNING FOR ESTIMATING ENVIRONMENTAL EXPOSURE - SESSION 2

Wenhua (Alven) Yu, Liam Liu

Climate, Air Quality Research Unit,
Monash University

Contents

- R package ‘deeper’ installation
- Basic steps for DEML
- Example and practices
- Deeper graphic user interface introduction

Through the tutorial, you will learn:

- How to use a single ML method to estimate air pollutants
- How to assess ML models with the optimal parameters
- How to use the "deeper" R package to perform DEML model
- How to use a Graphical User Interface to conduct DEML

•R ‘deeper’ Installation

Please make sure:

- using R (\>= 3.5.0)
- install certain dependent R packages: devtools, SuperLearner(\>= 2.0-28)
- install other suggested R packages: caret, skimr, CAST, ranger, gbm, xgboost

Install ‘deeper’ through following syntax:

```
library(devtools)
install_github("Alven8816/deeper")
```


MONASH University

• Algorithms selection

Deeper include 35 algorithms which are based on 'SuperLearner' R package

parameter	algorithm	required package	types
SL.bayesglm	Bayesian generalized linear regression	arm	R
SL.biglasso	Extending Lasso Model Fitting to Big Data	biglasso	R
SL.caret	random Forest as default	caret	R
SL.caret.rpart	decision trees as default	caret	R
SL.cforest	Breiman's random forests	party	R
SL.earth	Multivariate Adaptive Regression Splines	earth	R
SL.gam	generalized additive models	gam	N
SL.gbm	generalized boosting algorithm	gbm	R
SL.glm	generalized linear models	NA	R
SL.glm.interaction	generalized linear models	NA	R
SL.ipredbagg	Bootstrap aggregation (bagging)	ipred	R
SL.kernelKnn	Kernel k Nearest Neighbors	kernelknn	C
SL.ksvm	Kernlab's SVM Algorithm	kernlab	R
SL.lda	Linear discriminant analysis, used for classification	MASS	C
SL.lm	OLS via lm(),be faster than glm()	NA	R
SL.loess	Local Regression is a non-parametric approach that fits multiple regressions in local neighborhood	loess	N
SL.logreg	Logic Regression	LogicReg	N
SL.mean	mean value	NA	R
SL.nnet	Feed-Forward Neural Networks and Multinomial Log-Linear Models	nnet	N
SL.nnls	Non-negative least squares algorithm	nnls	N
SL.polymars	Polynomial Spline Routines	polspline	R
SL.qda	Quadratic discriminant analysis, used for classification	MASS	C
SL.randomForest	random Forest	randomForest	R
SL.ranger	a fast implementation of Random Forest	ranger	R

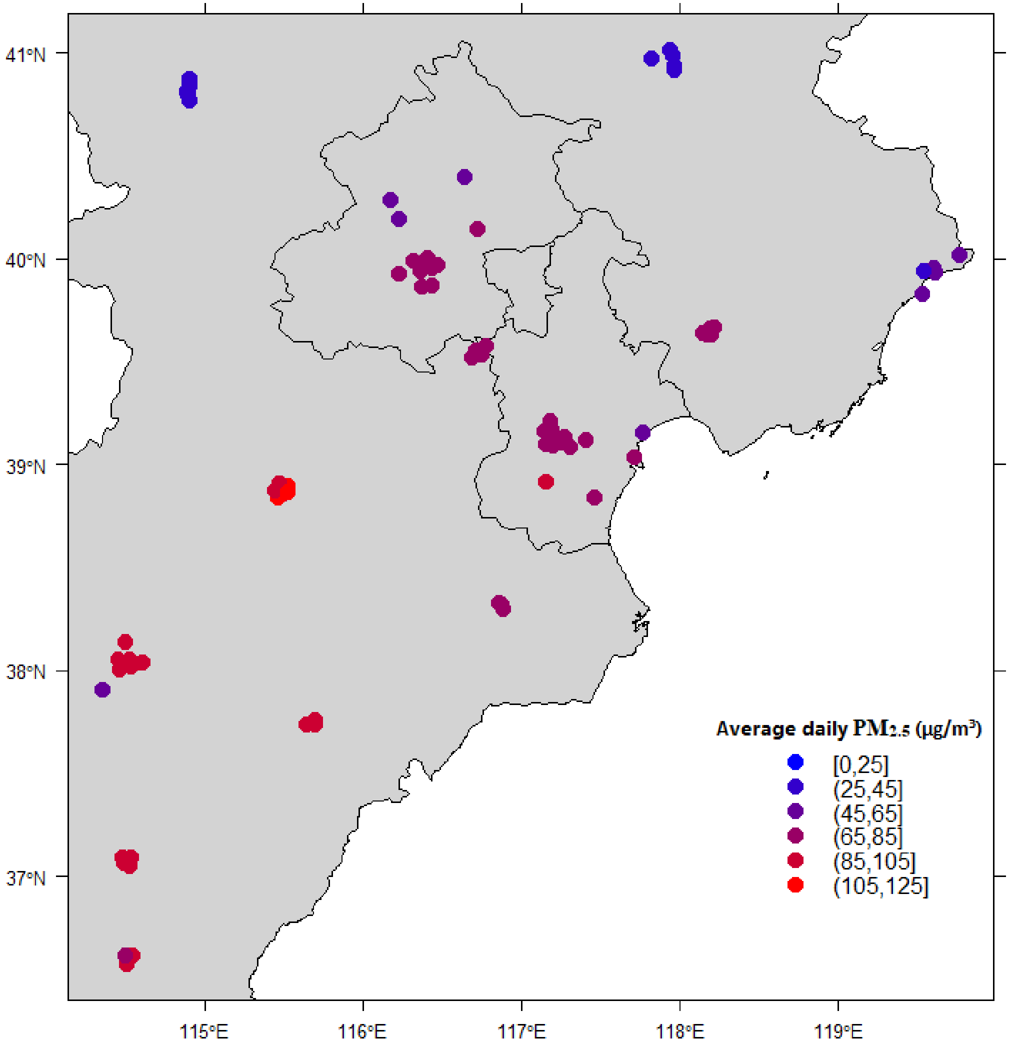
• Basic steps for DEML

- Step 1. Data preparation
- Step 2. Establish base models
- Step 3. Stacking meta models
- Step 4. Prediction based on new data set

Example:

To estimate the daily ambient $PM_{2.5}$ in the northeast of China in 2015-2016

- Remote sensing aerosol optical depth (AOD)
- Daily climate data: Temperature, Relative humidity, precipitation, pressure…
- Land cover information
- Elevation
- Fire information



MONASH University

Wenhua.yu@monash.edu

Presentation Materials can be downloaded here:
https://github.com/Alven8816/DEEPER_with_R_workshop_20220908

DEML Framework

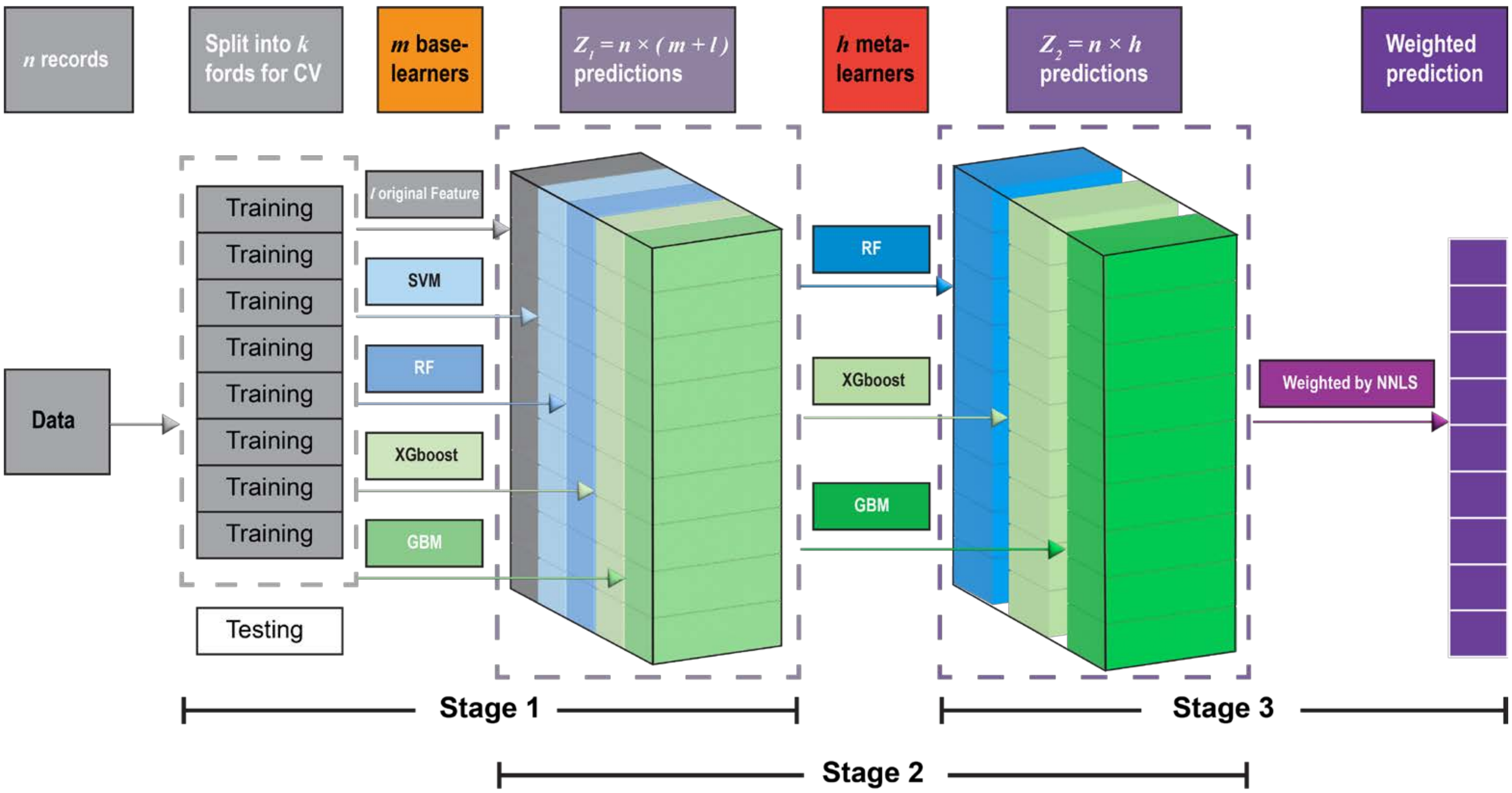


Fig 1. The overview of DEML framework

Note: SVM: Support Vector Machine; RF: Random Forest; XGBoost: Extreme Gradient Boosting; GBM: Gradient Boosting Machine; NNLS: Non-negative least squares

MONASH University

DEML Advantages

1. Outstanding model performance

2. Customizing diverse hierarchy structure

3. Minimize the extent of the empirical model selection

4. Automatically provide an optimal set of weights)

5. Easy to use and extent

DEML LIMITATIONS

1. Missing value sensitive

2. Computational complexity

3. ‘large’ sample size