ASM microbe ONLINE
summer of science

How to Machine Learn

Best practices in applying machine learning to microbiome data

Begüm D. Topçuoğlu

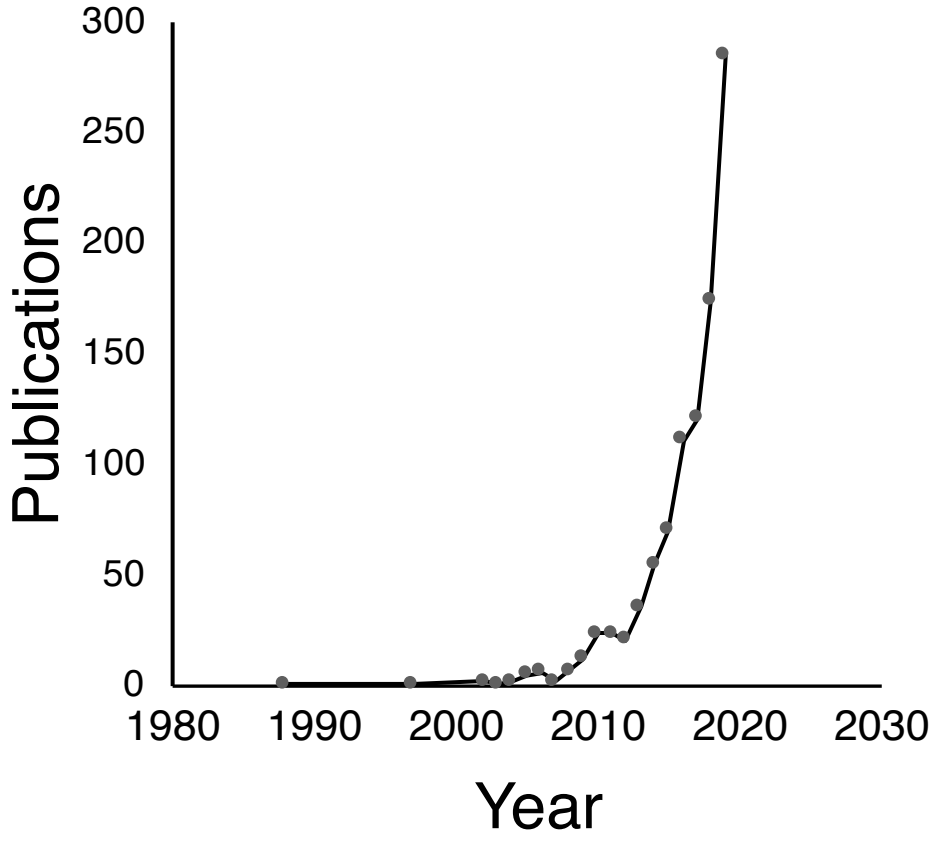
Senior Computational Biologist,
Exploratory Science Center, Merck & Co., Inc., Cambridge, Massachusetts, USA.

github.com/BTopcuoglu

Personal website: btopcuoglu.github.io/

Machine Learning

- Computer systems learning input to produce predictions on never-before-seen data using statistical techniques.
- Now commonly used in the microbiology field.
- Machine learning can help us utilize heterogenous and complex microbiology data.



Use of ML in microbiology

- Microbial Diagnostics:

Research Article | Host-Microbe Biology

Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome

Geoffrey D. Hannigan, Melissa B. Duhaime, Mack T. Ruffin IV, Charlie C. Koumpouras, Patrick D. Schloss

- Modeling microbe-microbe and microbe-host interactions:

Machine Learning Reveals Missing Edges and Putative Interaction Mechanisms in Microbial Ecosystem Networks

Demetrius DiMucci, Mark Kon, Daniel Segrè

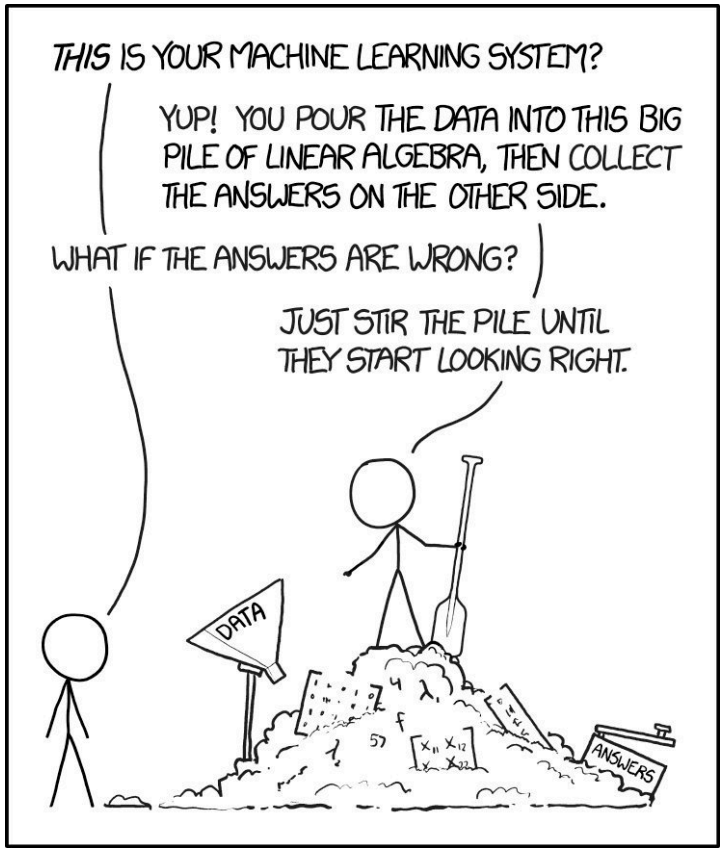
- Identification of genomic features:

A deep learning genome-mining strategy for biosynthetic gene cluster prediction

Geoffrey D Hannigan,[1] David Prihoda,[2,3] Andrej Palicka,[4] Jindrich Soukup,[5] Ondrej Klempir,[6] Lena Rampula,[7] Jindrich Durcak,[6] Michael Wurst,[4] Jakub Kotowski,[4] Dan Chang,[8] Rurun Wang,[1] Grazia Piizzi,[1] Gergely Temesi,[6] Daria J Hazuda,[1,9] Christopher H Woelk,[1] and Danny A Bitton[6]

New avenue in the microbiome field

- Gap in knowledge and technical expertise
- Common methodological problems.
- We need reliable and responsible ML models.

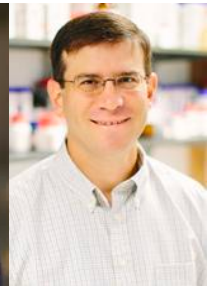


xkcd.com

ASM microbe ONLINE summer of science

CSE COMPUTER SCIENCE AND ENGINEERING UNIVERSITY OF MICHIGAN

MEDICAL SCHOOL UNIVERSITY OF MICHIGAN



mBio AN OPEN THE AME

Research Article | Host-Microbe Biology

A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems

Begüm D. Topçuoğlu, Nicholas A. Lesniak, Mack T. Ruffin IV, Jenna Wiens, Patrick D. Schloss

Today we will learn basics of ML and do some live coding to build ML pipelines for microbiome based classification problems.

Key terminology

Label: What we are predicting

- IMDB score of a movie
- If a passenger survived Titanic
- If someone has colon cancer

Feature: Input variable

- # of Instagram likes of lead actor
- Age and gender of passenger
- Species abundances in the stool

- x_1 Maybe just on feature in a simple machine learning project.
- $x_1\, x_2\, x_3\, \ldots\, x_N$ (many features) for a sophisticated machine learning project.

Model learns the relationship between the features and the label

A simple example:
Classify an email as spam or not?

Labels:

- Spam or Not Spam

Features:

- Sender's email address.
- Time of the day it was sent.
- Subject line has "loan repayment".
- Subject line has a receipt number.
- Emojis in Subject Line.

Model learns the relationship between the features and the label

Steps in building a ML model

1. Gather Data: Look at hundreds of thousands of emails
2. Prepare/clean data: e.g. correct or remove SPAM labeled .edu emails, emails from ASM
3. Separate data into train and test sets

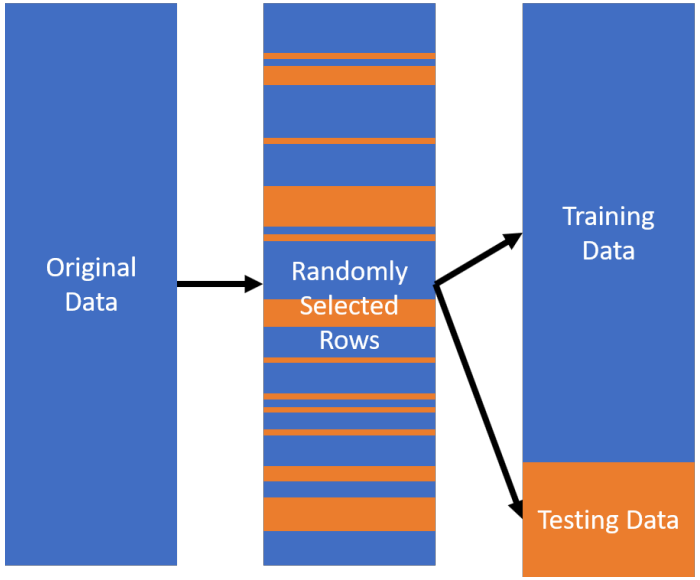
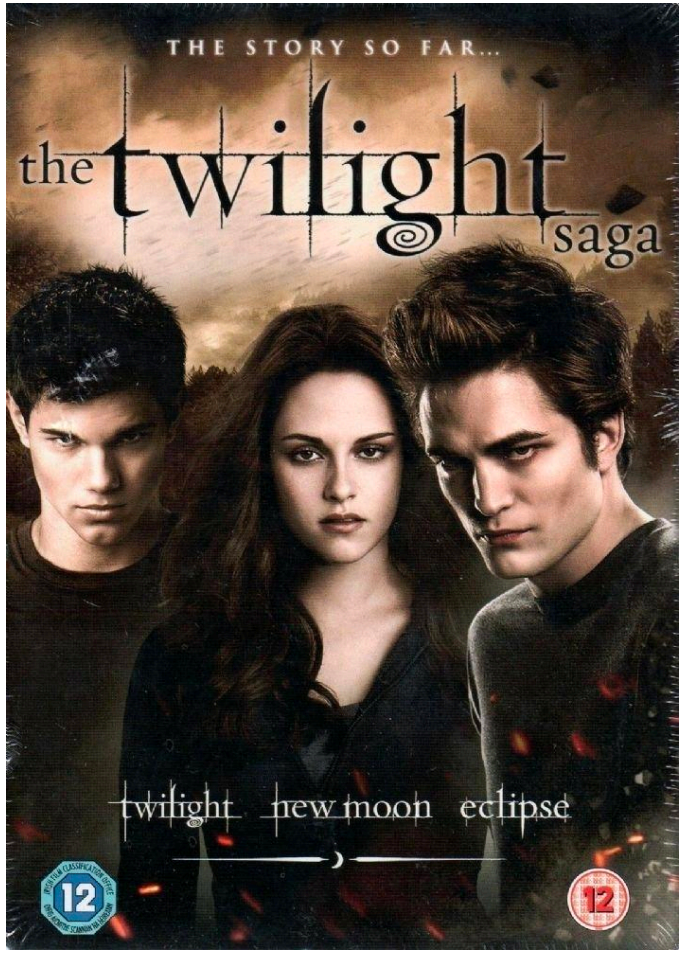


Figure by Zena Lapp

Steps in building a ML model

4. Choose a model: e.g. logistic regression
5. Tune parameters: e.g. penalty for getting things too right
6. Train (learning): show your model spam email examples, and enable the model to gradually learn the relationships between features and label
7. Evaluate: Are we doing well, should we change model/parameters?
8. Predict: Apply the trained model to unlabeled emails

Example: A ML model that can predict the IMDB score of a movie



Can I predict the IMDB score of a movie by looking at Instagram likes of the lead actor?

Regression: one feature (simple linear ML model)

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon$$

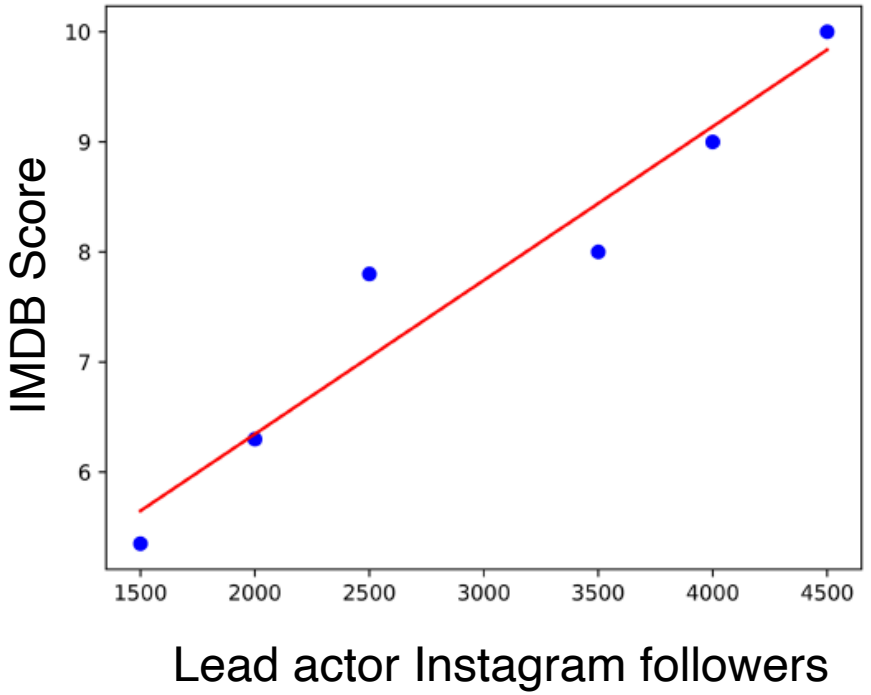
Intercept

Error

Label of ith sample:
IMDB score of movie

Weight of Feature 1

Feature 1:
Lead actor Instagram followers

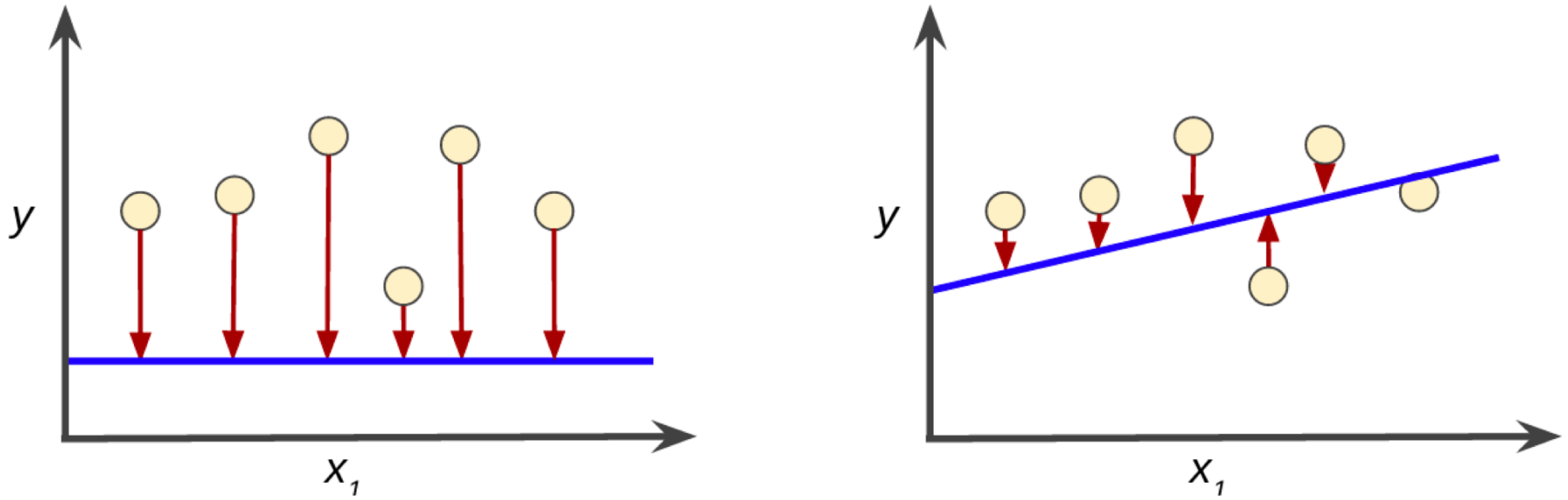


Regression: many features (more sophisticated ML model)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \epsilon$$

- **Features:** Country of the movie, genre, duration, plot keywords, budget etc.
- Each of these features will have different weights.
- Training a model simply means learning (determining) good values for all the weights from labeled examples.
- How does a model do that?

The model learns by making bad predictions and getting a penalty for the bad prediction

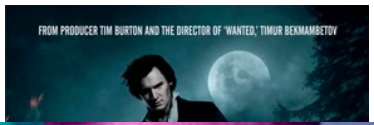


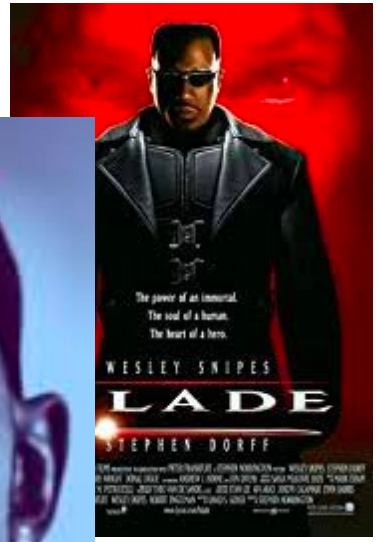
The goal of training a model is to find a set of weights that have low loss, on average, across all examples (minimize the total error).

Be careful of overfitting!

Memorizing is not learning

Be careful of overfitting!





Potential ML problems with microbiological impact

- Genome-wide association studies (GWAS)
  - Relationship between genetic variants and trait (ex. disease)
- Metagenome level information
  - Relationship between metaG features and response to a medication.
- 16 rRNA gene surveys
  - Relationship between bacterial abundances and trait (ex. disease)

Live-coding where we will use a published dataset:

Predict if a patient has colorectal cancer using bacterial abundances in their stool.

ML problem:

Predict if a patient has colorectal cancer using bacterial abundances in their stool.

Labels:

Colorectal lesions of patients - defined as cancer or normal.

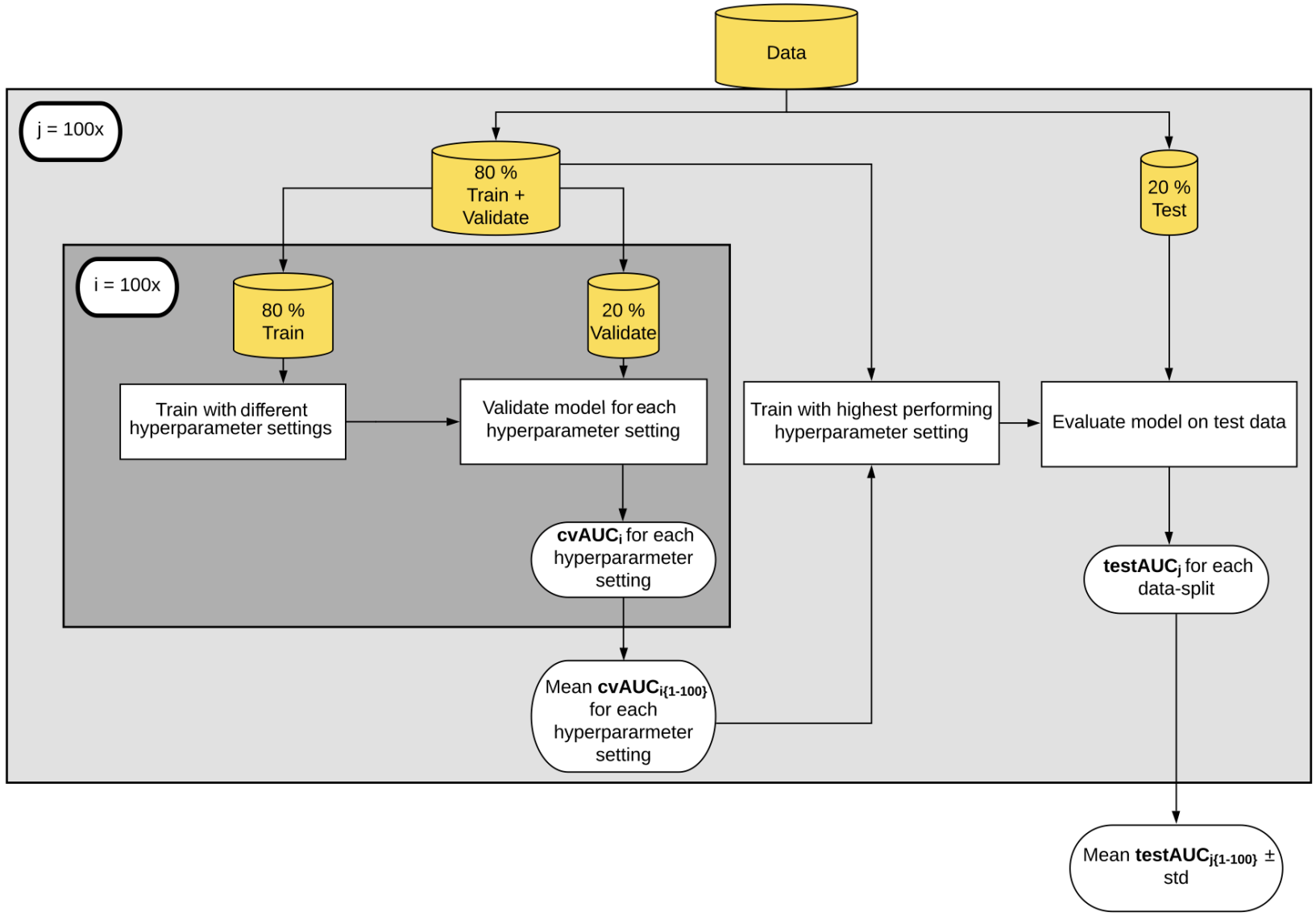
Features:

Bacterial abundances in stool (OTU-level information)

ML algorithm:

Logistic regression

ML pipeline


Data
j = 100x
80 % Train + Validate
20 % Test
i = 100x
80 % Train
20 % Validate
Train with different hyperparameter settings
Validate model for each hyperparameter setting
cvAUC$_i$ for each hyperpararmeter setting
Mean cvAUC$_{i\{1\text{-}100\}}$ for each hyperpararmeter setting
Train with highest performing hyperparameter setting
Evaluate model on test data
testAUC$_j$ for each data-split
Mean testAUC$_{j\{1\text{-}100\}}$ ± std

Live-coding

We will go step by step to build a ML model with a colorectal cancer microbiome dataset.

If you would like to reproduce what I will show today:

https://github.com/BTopcuoglu/machine-learning-pipelines-r

If you want to use the full robust ML pipeline, tune in to our new R package development:

https://github.com/SchlossLab/mikRopML

Zena Lapp and Kelly Sovacool

Acknowledgements

Schloss Lab

Patrick Schloss

Nick Lesniak

Kelly Sovocol

Zena Lapp

Lucas Bishop

Ana Taylor

Will Close

Neil Baxter

Sarah Tomkovich

Joshua Stough

Sarah Wescott

Computer Science and Engineering Department

Jenna Wiens

Penn State Cancer Institute

Mack Ruffin

Great Lakes-New England Early Detection Research Network

PennState Cancer Institute

ASM microbe ONLINE
summer of science

How to Machine Learn

Best practices in applying machine learning to microbiome data

Begüm D. Topçuoğlu

Senior Computational Biologist,
Exploratory Science Center, Merck & Co., Inc., Cambridge, Massachusetts, USA.

github.com/BTopcuoglu

Personal website: btopcuoglu.github.io/