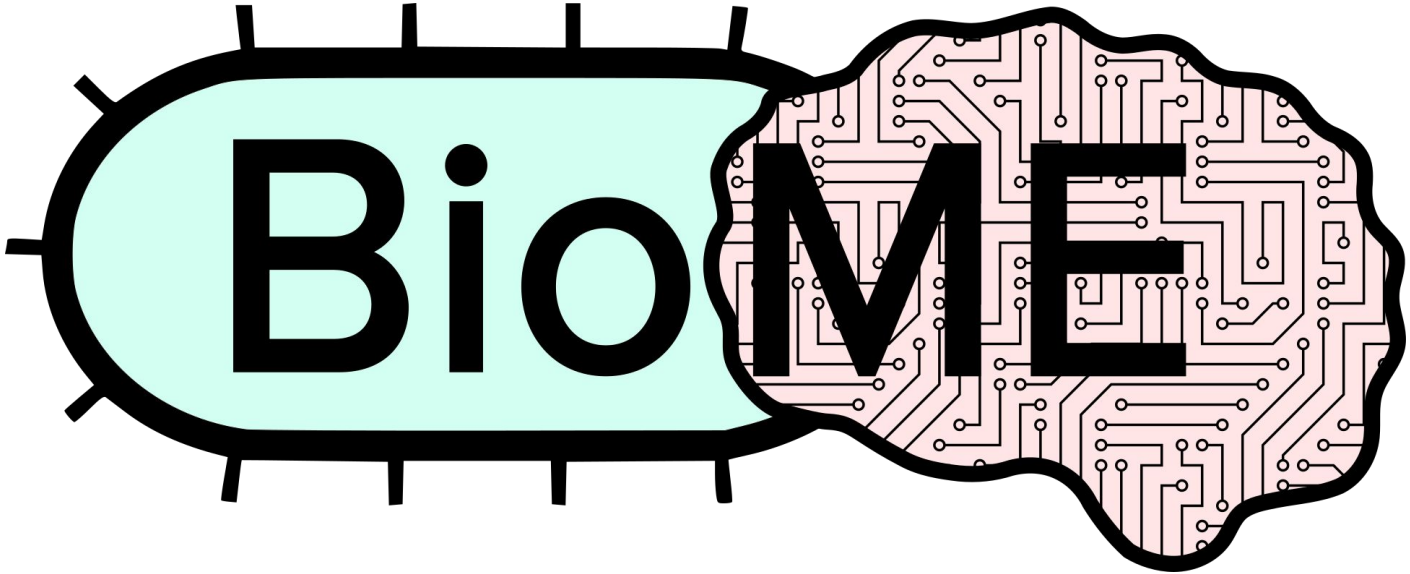


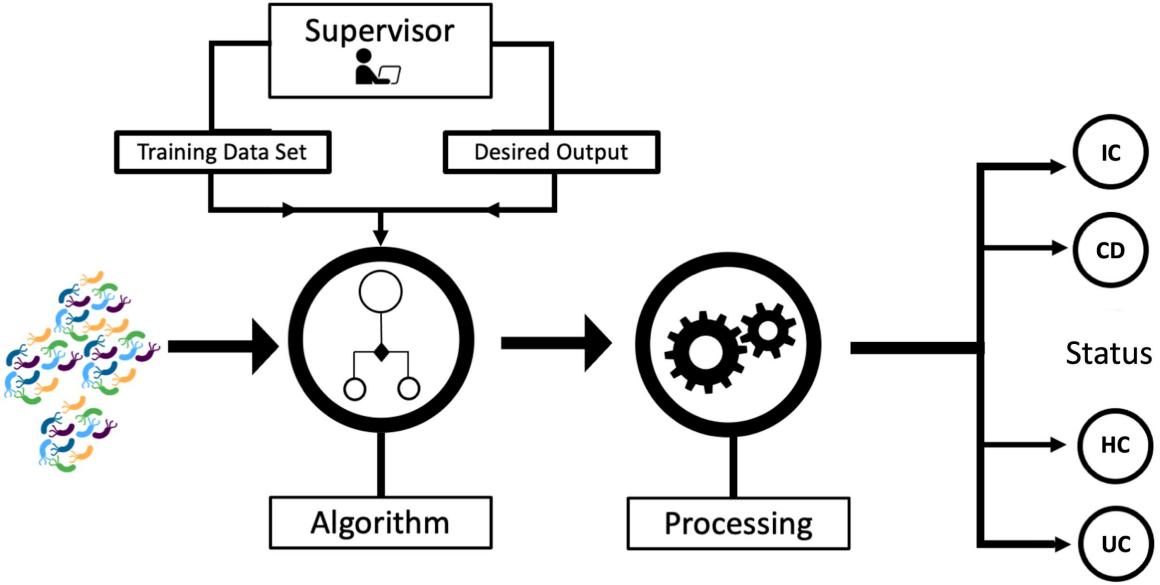
Supervised Machine Learning for Microbiome Data

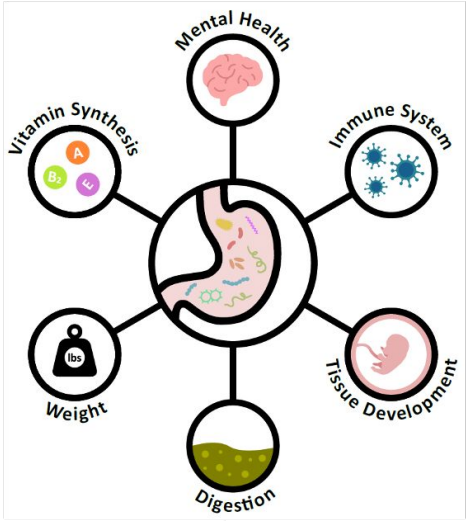
Binnan Yu, Lillian Tatka, Kristina Herman, David Lee, Sierra Gillman

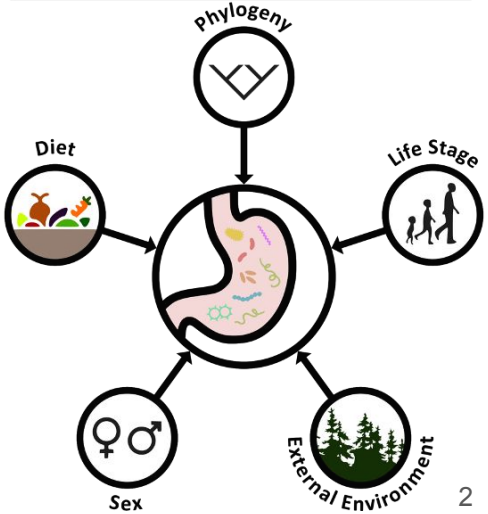
Introduction

The problem being addressed

Increasing accessibility to machine learning for microbiome research







The Data

4 tabundance ables of Operational Taxonomic Units (OTUs) and their metadata files

Cell Host & Microbe
Resource

The Treatment-Naive Microbiome in New-Onset Crohn's Disease

Dirk Gevers,[1] Subra Kugathasan,[2,23] Lee A. Denson,[3,24] Yoshiki Vázquez-Baeza,[4] Will Van Treuren,[1] Boyu Ren,[5] Emma Schwager,[6] Dan Knights,[7,8,9] Se Jin Song,[4] Moran Yassour,[10] Xochitt C. Morgan,[4] Aleksandar D. Kostic,[1] Chengwei Luo,[1] Antonio González,[4] Daniel McDonald,[4] Yael Haberman,[3] Thomas Walters,[11] Susan Baker,[12] Joel Rosh,[13] Michael Stephens,[14] Melvin Heyman,[15] James Markowitz,[16] Robert Baldassano,[17] Anne Griffiths,[18] Francisco Sylvester,[19] David Mack,[20] Sandra Kim,[21] Wallace Crandall,[21] Jeffrey Hyams,[18] Curtis Huttenhower,[1,4] Rob Knight,[4,22,23] and Ramnik J. Xavier[1,2,3,*]

mSystems

American Gut: an Open Platform for Citizen Science Microbiome Research

qiime2

Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases

Jason Lloyd-Price[1,2], Cesar Arze[2], Ashwin N. Ananthakrishnan[3], Melanie Schirmer[1,2], Julian Avila-Pacheco[4], Tiffany W. Poon[5], Elizabeth Andrews[6], Nadim J. Ajami[6], Kevin S. Bonham[1,7], Colin J. Brislawn[8], David Casero[9], Holly Courtney[3], Antonio Gonzalez[4], Thomas G. Graeber[10], A. Brantley Hall[11], Kathleen Lake[12], Carol J. Landers[13], Himel Mallick[1,2], Damian R. Plichta[1], Mahadev Prasad[12], Ghodamrih Rahnavard[1,2], Jenny Sauk[14], Dmitry Shungin[1,4], Yoshiki Vázquez-Baeza[4,15], Richard A. White III[4], IBDMDB Investigators[17], Jonathan Braun[3], Lee A. Denson[16,18], Janet K. Jansson[8], Rob Knight[8,19,20], Subra Kugathasan[12], Dermot P. B. McGovern[13], Joseph F. Petrosino[7], Thaddeus S. Stappenbeck,[16], Harland S. Winter[14,17], Clary B. Clish[4], Eric A. Franzosa[1], Hera Vlamakis[1], Ramnik J. Xavier[1,2,3,21] & Curtis Huttenhower[1,2,24,*]

BRIEF COMMUNICATION
https://doi.org/10.1038/s41592-018-0141-9

nature methods

Qiita: rapid, web-enabled microbiome meta-analysis

Antonio Gonzalez[1,12], Jose A. Navas-Molina[1,2,10,12], Tomasz Kosciolek[1], Daniel McDonald[1], Yoshiki Vázquez-Baeza[1], Gail Ackermann[1], Jeff DeReus[1], Stefan Janssen[1], Austin D. Swafford[3], Stephanie B. Orchanian[3], Jon G. Sanders[1], Joshua Shorenstein[1,11], Hannes Holste[1,2], Semar Petrus[4], Adam Robbins-Pianka[5], Colin J. Brislawn[6], Mingxun Wang[7], Jai Ram Rideout[8], Evan Bolyen[8], Matthew Dillon[8], J. Gregory Caporaso[8,9], Pieter C. Dorrestein[1,3,7] and Rob Knight[1,2,3*]

Qiita

nature microbiology

ARTICLES
PUBLISHED: 13 FEBRUARY 2017 | VOLUME: 2 | ARTICLE NUMBER: 17004

Dynamics of the human gut microbiome in inflammatory bowel disease

Jonas Halfvarson[1], Colin J. Brislawn[2], Regina Lamendella[3], Yoshiki Vázquez-Baeza[4], William A. Walters[5], Lisa M. Bramer[6], Mauro D'Amato[7,8], Ferdinando Bonfiglio[9], Daniel McDonald[10], Antonio Gonzalez[11], Erin E. McClure[3], Mitchell F. Dunklebarger[3], Rob Knight[4,10,11] and Janet K. Jansson[2*]

The Users

The Ecologist: Loves R
Is interested in seeing if habitat quality can be predicted based on fecal microbiome community composition in American marten (*Martes americana*). They want to determine if they can classify individuals from primary or disturbed habitat as this could be a powerful tool for conservation and management!

The Medical Clinician: Can google
Wants to be able to determine if suspected patients have inflammatory bowel disease so they can begin effective treatment before their health deteriorate, and will use BioME to classify people!

The Microbiologist: bioinformatic buff
Is here for the preliminary results/confirmation. The lab intern might have mislabeled some (~100) of the samples. Does say that a *Turdis turdis* or just *Turtle's turds*? They don't want to have to throw out all those samples..

The Users

The Ecologist: Loves R

The ecologist collects fecal samples from wild martens and puts the data into BioME where he sees that support vector classification is able to sort out the samples with high accuracy.

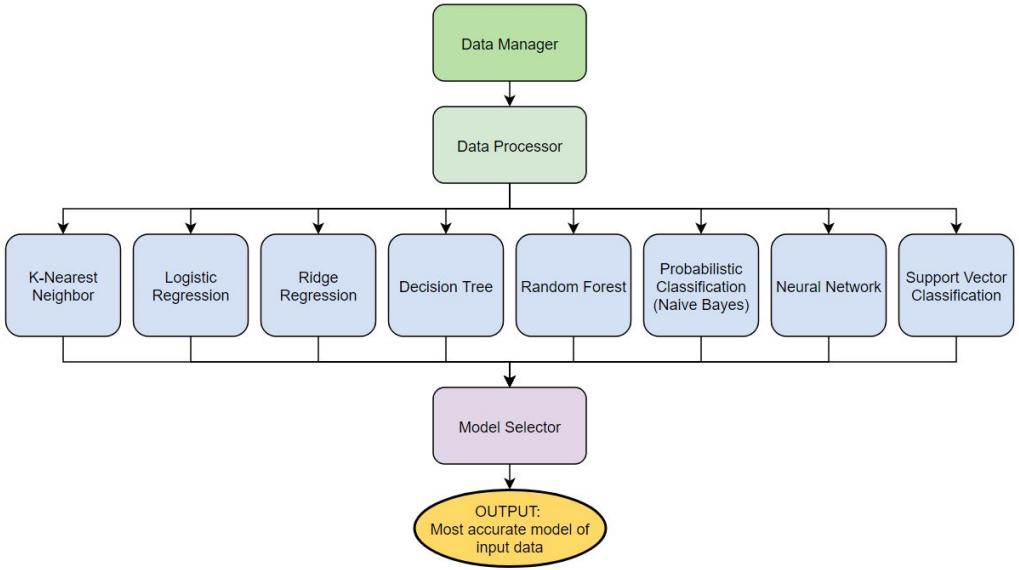
The Medical Clinician: Can google

The clinician gives BioME data from healthy patients and patients with inflammatory bowel disease. She then inputs the patient's data for BioME to classify as healthy or diseased.

The Microbiologist: bioinformatic buff

The microbiologist gives sample data to BioME to verify the identity of mislabeled samples. Later, he gives his verified data to BioME to compare models.

Functional Specifications




```
BioME/
    | - README.md
    | - biome/
        | - __init__.py
        | - prep_split_data.py
        | - select_model.py
        | - train_mlp.py
        | - knn.py
        | - dtree.py
        | - logistic.py
        | - ridge.py
        | - random.py
        | - SVC.py
        | - naive_bayes.py
        | - scripts/
            | - biome_run.py
        | - tests/
            | - …
        | - Data/
            | - bug_OTU_rel.tsv
            | - bug_OTU_raw.tsv
            | - FecesMeta.txt
            | - query_point.tsv
    | - docs/
        | - FunctionalSpec.md
        | - ComponentSpec.md
        | - Technology Review Presentation
        | - images/
            | - …
    | - setup.py
    | - LICENSE
    | - BioME_environment.yml
```

Input Known Data

OTUs "bugs" as column 1

SampleID as column 1 in metadata & Second row in OUT file

Column for prediction

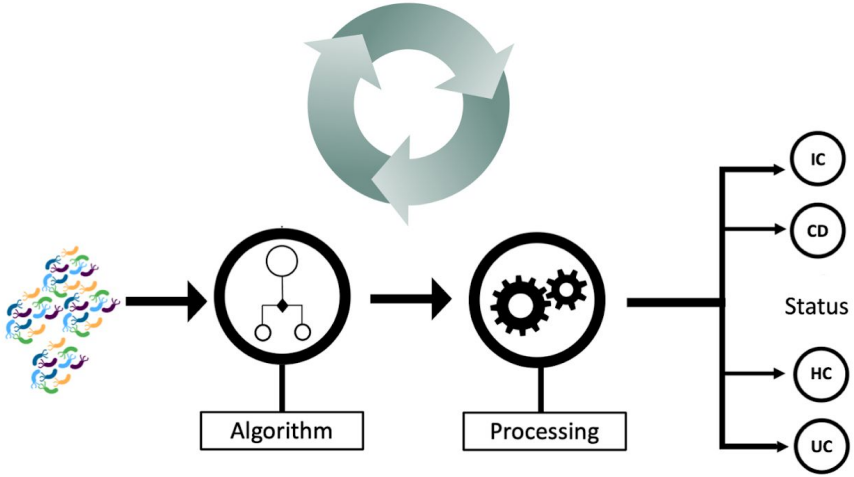
# Constructed from biom file					
#OTU ID	1939.10006	1939.122117	1939.122163	1939.12128	1939.12207
AACATAGGGGGCAAGCGTTGTCCGGA	0	0	0	0	0
AACATAGGGGGCAAGCGTTGTCCGGA	0	0	0	0	0
AACATAGGGGGCAAGCGTTGTCCGGA	0	0	0	0	0
AACATAGGGGGCAAGCGTTGTCCGGA	0	0	0	0	0

#SampleID	qiita_study_id	ML_Diagnosis	b_cat	biopsy_locat	birthdate
1939.100058	S1939	CD	Missing: Res	stool	Missing: Res
1939.122117.a	S1939	CD	B1	stool	8/2/91
1939.122163.a	S1939	UC	no	stool	2/14/91
1939.121281	S1939	CD	B3	stool	3/30/87
1939.122074	S1939	CD	no	stool	10/8/82

Data is prepped and split into four groups:
X_train- OTU table to train models Y_train metadata to train models
X_test: OUT table to get predictions/test accuracy Y_test metadata for predictions/accuracy

Choose which models to run

K-Nearest Neighbor | Neural Networks | Logistic Regression | Support Vector Machine | Decision Tree| Random Forest



Design

```
                                                          ***********
          \           |              |                **********************
        .%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% *********************************
     -%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% *************************************
  \-%%%%%% @@@@@@@@@ %%%% @@ %%%%%%%%%%%%%%%%%  @@@ ******** @@ *** @@@@@@@@@@ *********
  *%%%%%%% @ %%%%%% @@ %%%%%%%%%%%%%%%%%%%%% ** @@@ ******* @@@ *** @ *******************
 +%%%%%%%% @ %%%%%% @@ %% @@ %%%%% @@@@ %%%% ** @@@@ ***** @@@@ *** @ *******************
_%%%%%%%%% @@@@@@@@@@ %%% @@ %% @@ %%%% @@ % ** @@ @@ *** @@ @@ *** @@@@@@@@@@************
 *%%%%%%%% @ %%%%%%% @ %% @@ % @@ %%%%%% @ % ** @@ * @@ @@ * @@ *** @ *********************
  #%%%%%%% @ %%%%%% @@ %% @@ %% @@ %%%% @@ %% * @@ ** @@@ ** @@ *** @ *********************
  /+%%%%%% @@@@@@@@@@ %%% @@ %%%% @@@@@@  %%% * @@ *** @ *** @@ *** @@@@@@@@@@ ************
     .%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% *******************************************
        .%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%  ***************************************
           /            |              |                 *******************************
                                                                 ********************
                                                                      ************
                                                                        ******

Please enter the relative path to the OTU data: biome/Data/bug_OTU_rel.tsv

Please enter the relative path to the categorical data: biome/Data/FecesMeta.txt

Please list the categorical variables of interest: CD,UC,IC,HC,CC

What models would you like to test?
See README.md for abbreviations. Type all if all models should be tested: mlp1,mlp3,dtree,knn,gnb

enter number of neighbors: 5
Model: Accuracy score
MLP (single hidden layer): 0.738
MLP (three hidden layers): 0.705
Gaussian Naive Bayes: 0.656
Decision Tree: 0.623
k-nearest neighbors: 0.5493

The best performing model is: MLP (single hidden layer)

Would you like to make a prediction with the best model? yes

Please enter the path to the data that you would like to make a prediction for: biome/Data/query_point.tsv

The predicted category is: CD

Would you like to make another prediction? no
```

Future Work

Create a webapp that allows user to select a model they want to run

Test Page

kmherman.github.io/BioME/

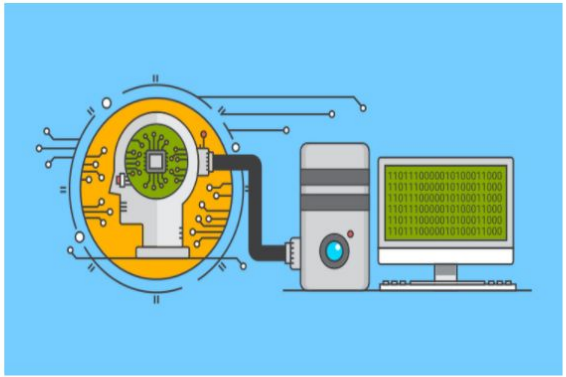
Why Our Solution is Trustworthy?

1. Algorithm

Our product is developed based on 7 different algorithms, each of it have different characteistics so that our product can handle different dataset.

Click the dropdown list to see algorithms option we provide

Logistic Regression
K-nearest Neighbor
Logistic Regression
Ridge Regression
Dicision Tree
Random Forest
Naives Bayes
Neural Network



Test Page

kmherman.github.io/BioME/

Supervised Machine Learning for Microbiome Data

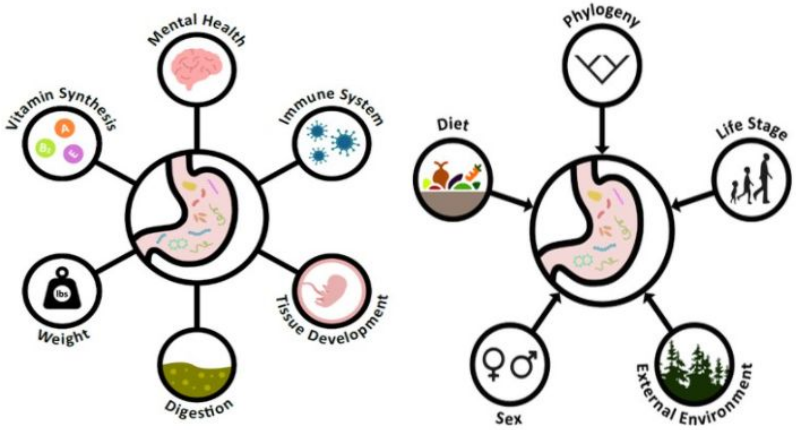
Kristina Herman[1] Lillian Tatka[2] David Lee[2] Sierra Gillman[3] Binnan Yu[4,5]

[1]Department of Chemistry; [2]Department of Bioengineering; [3]School of Environmental and Forest Sciences; [4]Department of Rehabilitation Medicine; [5]Department of Applied Mathematics, University of Washington, Seattle, WA, 98105

Statement of The Problem

Once thought only to be pathogenic, the microorganisms living on and within an animal host are now recognized as playing critical roles in host health. Factors that shape gut microbial communities are multifaceted and include the host's diet and life-stage.

While microbial shifts have been implicated in numerous human ailments (e.g., obesity, anxiety, inflammatory bowel disease), research has thus far been limited to differentiating microbial communities between groups and less for predictive uses. As a result of an ever expanding microbiome data availability, microbiome research lends itself to advancement with machine learning and data science. However, the implementation of machine learning in microbiome research might feel daunting and time consuming to those outside of the realm of data science.



Thank You!