Neural Embeddings for 16S Microbiome Classification

Advisors: Rob Knight & Daniel McDonald, Prof. Ilkay Altintas
Ryan Conrad, Ryan Inghilterra, Brandon Westerberg, Sean Rowan

DSE CAPSTONE - Cohort 4, Group 1

Apr 09, 2019

Gut Microbiome Signature to Be Developed as Nonalcoholic Fatty Liver Disease Diagnostic

ARTICLE: IN-DEPTH—in Molecular Diagnostics

The researchers said a microbiome-based test could expand the number of individuals who are screened for the disease.

Apr 19, 2019

Active Celiac Disease Processes Informed by Intestinal Expression Comparison

ARTICLE: BREAKING NEWS—in Sequencing

A team compared expression in duodenum samples from active celiac cases, cases in remission, and unaffected controls, identifying active disease-related expression shifts.

Apr 01, 2019

Colorectal Cancer Gut Microbial Signatures May Lead to New Diagnostic Tests

ARTICLE: BREAKING NEWS—in Sequencing

A pair of new fecal metagenomics studies pointed to gut microbial community shifts, related functional changes, and specific signatures for colorectal cancer.

May 09, 2019

Healthy Big Data Firehose?

BLOG POST—in The Scan

New research, and a burgeoning company, point to the possible benefits of deep data profiling on healthy individuals, but critics aren't convinced.

Science Home News Journals Topics Careers

Log in | My accou

SHARE

25K

2K

Among the many microbes in the gut (above), some may influence mood. V. ALTOUNIAN/*SCIENCE*

Evidence mounts that gut bacteria can influence mood, prevent depression

By Elizabeth Pennisi | Feb. 4, 2019 , 11:00 AM

May 30, 2019

Inflammatory Bowel Disease, Prediabetes, Preterm Birth Studies Reveal Microbiome-Host Interactions

ARTICLE: BREAKING NEWS—in Sequencing

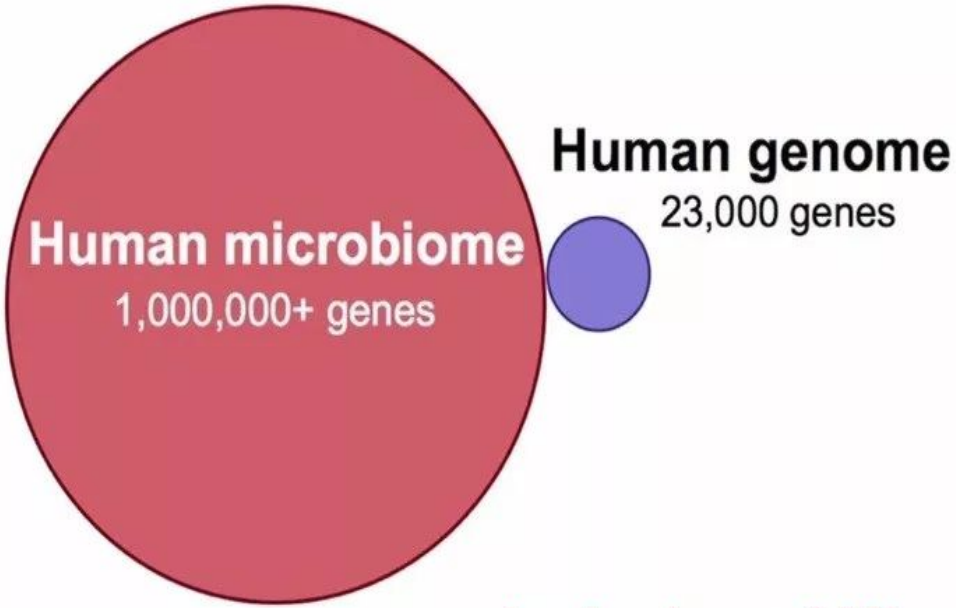
The studies, which used multi-omic approaches, are part of the integrative Human Microbiome Project (iHMP) — the second phase of the Human Microbiome Project.

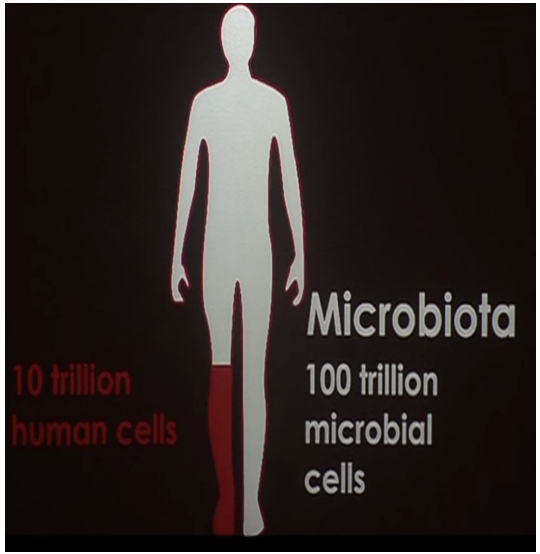
Apr 15, 2019

How You Get the Microbes

BLOG POST—in The Scan

The *Guardian* reports that delivery mode seems to affect the makeup of infants' gut microbiomes.





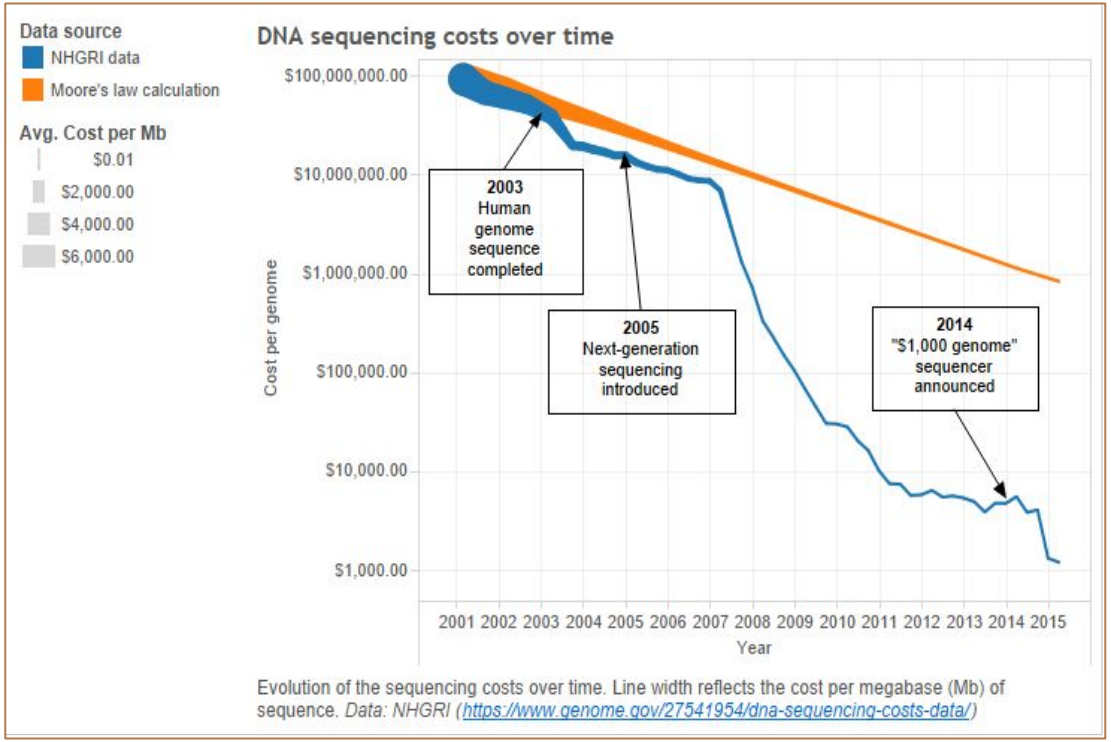
Inclusion of Microbiome
Will Radically Change Medicine

Image attributed to allergiesandyourgut.com

Why is this such a hot topic?

Cheap DNA Sequencing

- Advances in DNA sequencing have made sequencing inexpensive and accessible
- First reference human genome was mapped in 2003 for $3 billion taking 13 years to complete
- Today a whole human genome can be sequenced in one day for less than $800



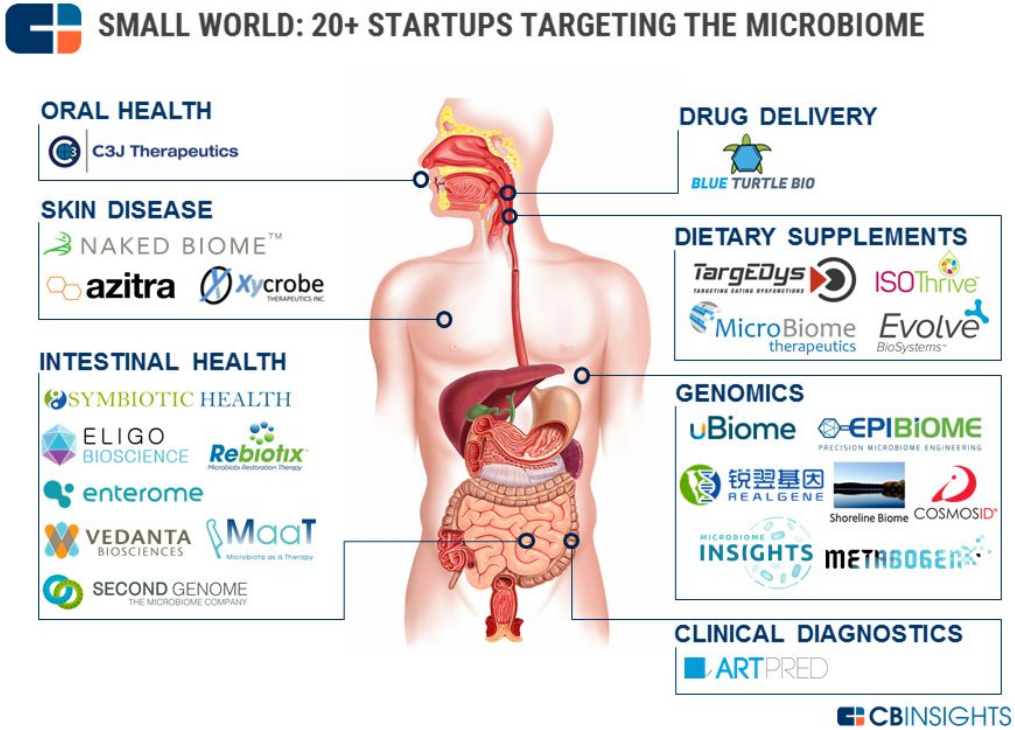
Scientists and companies are taking advantage of cheaper DNA sequencing and increased microbiome analysis capabilities

KNIGHTLAB

biota® TECHNOLOGY

COSMOSID

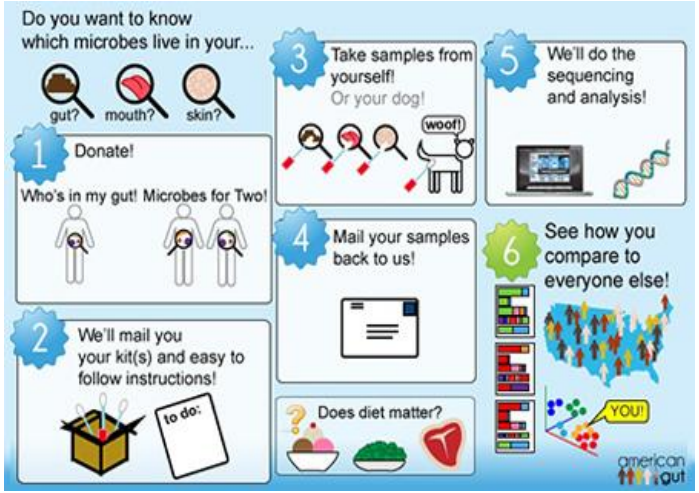
illumina



The American Gut Project

What is the American Gut Project?

- Began in 2012 as a platform for citizen scientist and professional researchers to understand the microbes that inhabit our bodies using DNA sequencing
- Open source and open data access
- Analyzed over 20,000 samples
- Operated out of UCSD School of Medicine
- For $99 you can contribute and have your gut microbiome analyzed



Problem Statement

The gut microbiome contains a wide feature space containing diverse organisms making it difficult to interpret and model

- Difficult to interpret and model even in its post-sequence form, especially in studies with small sample sizes
- Providing solutions to this problem will aide in future interpretability of microbiome space for researchers

Approach

Goal: Evaluate embedding techniques not traditionally used for representing 16S sequencing data to determine if they are able to retain the semantics of the original data set.

The embeddings will be validated by using them as features for body site classification where the de facto standard embedding technique, principal coordinates analysis, has been shown to be highly effective.

Reduction of Microbiome OTU Data

- OTU - Operational Taxonomic Unit or “The thing(s) being studied”
- In the AGP project, these “things” are 100 nucleotide truncated V4 regions of the 16S SSU rRNA gene referenced using Greengenes taxonomy database
- Results in ~ 19,100 OTUs (i.e. our feature columns) for each sample

Greengenes (GG) Taxonomy Tree

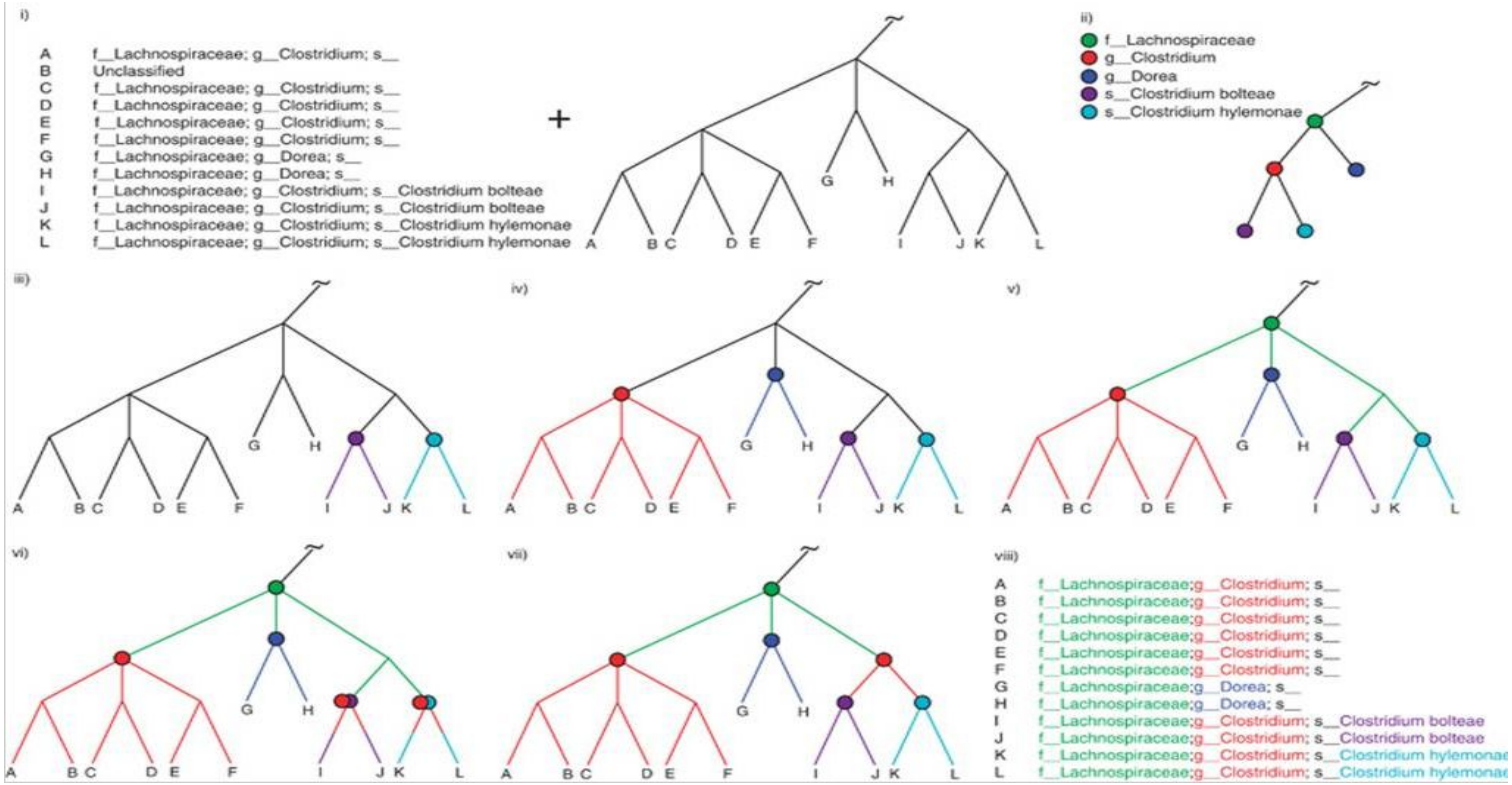


Image from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3280142/

Embedding Techniques

Principal Coordinates Analysis (PCoA)

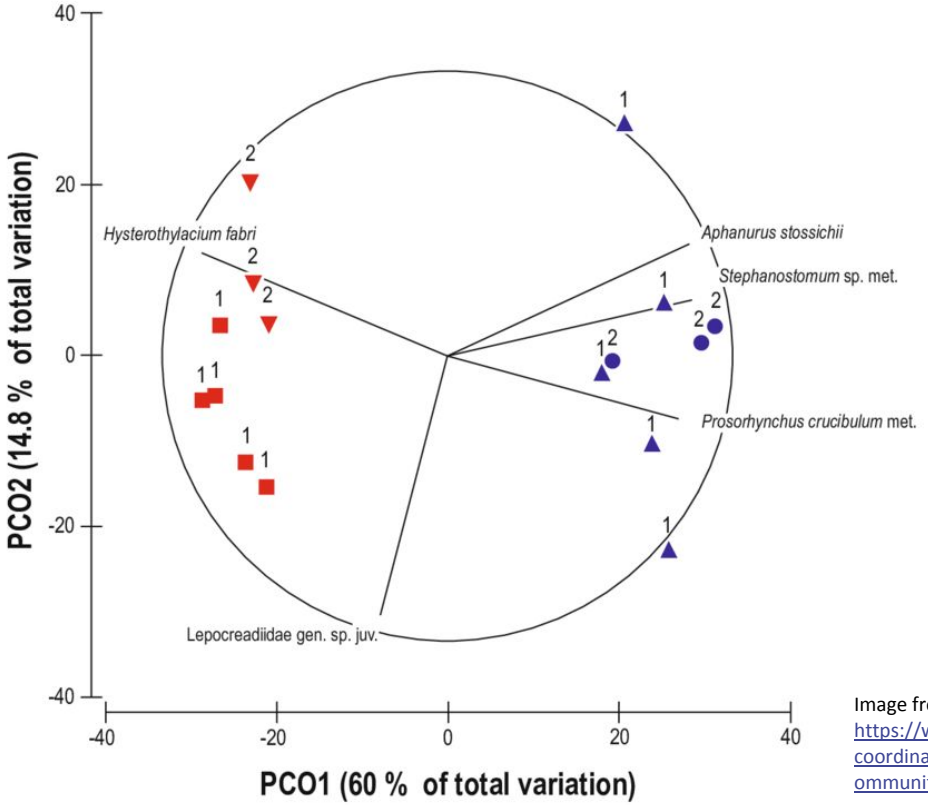


Image from https://www.researchgate.net/figure/Principal-coordinates-analysis-PCO-plot-of-component-communities-in-Boops-boops_fig2_313735563

PCoA Embeddings

- UniFrac is a β-diversity measure (distance metric) that uses phylogenetic information to compare samples
- Principal Coordinates Analysis (pcOa) can be applied to the resulting unifrac distance matrix
- This approach is used frequently within the metagenomics field to identify factors explaining differences among microbial communities

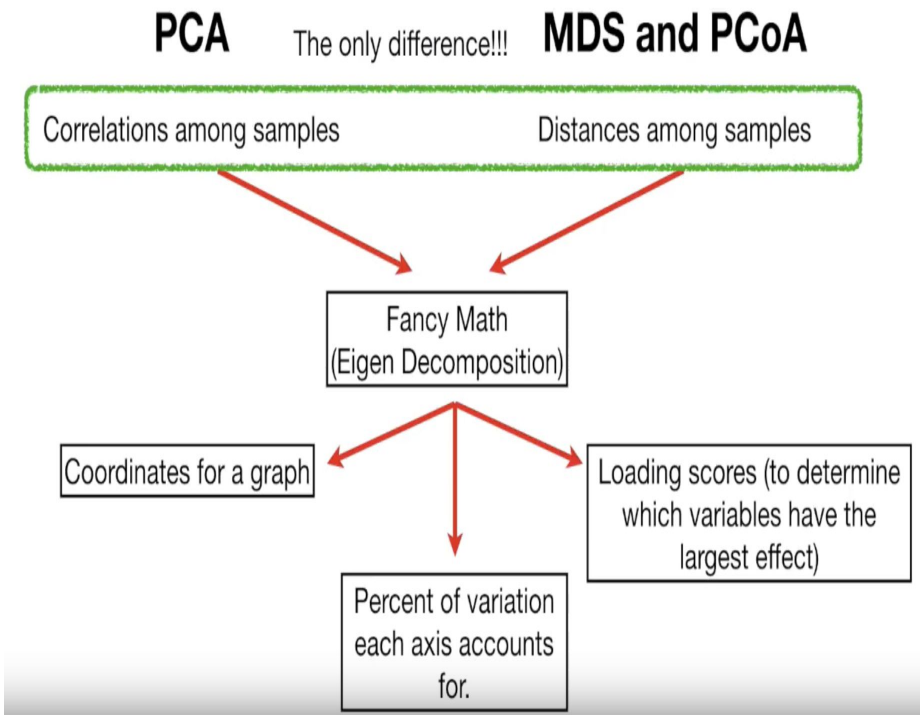
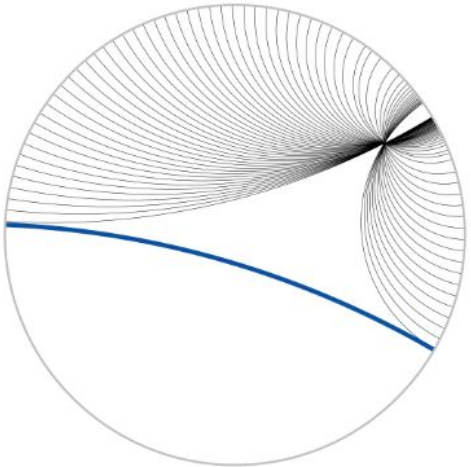


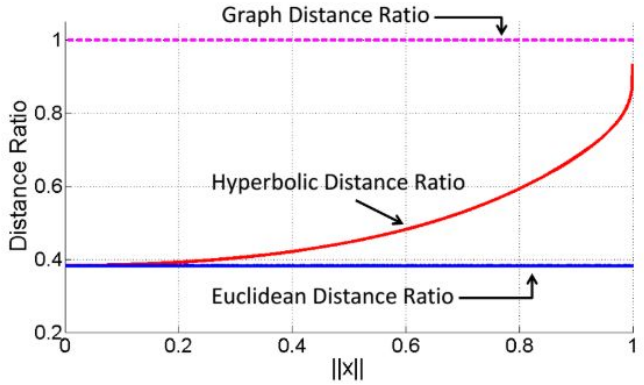
Image attributed to
StatQuest with Josh Starmer

Hyperbolic Embeddings

- Data with a latent hierarchy is challenging to accurately represent in low dimensional Euclidean space
- Projecting hierarchical data into hyperbolic space can overcome the issues of embedding in Euclidean space
- Proposed by Maximilian Nickel and Douwe Kiela of Facebook AI Research in 2017

The nonlinear properties of hyperbolic space allow for excellent embedding of trees and graphs

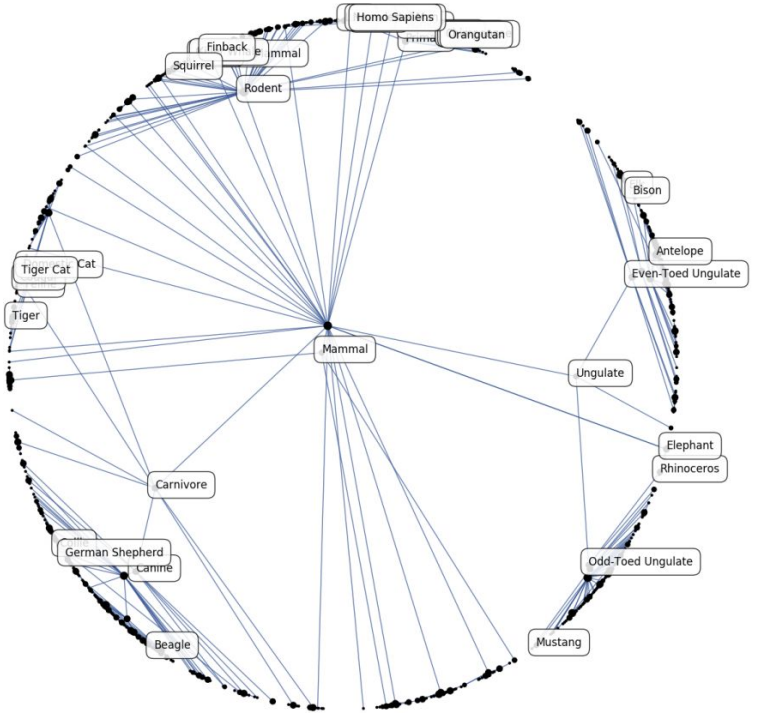




https://dawn.cs.stanford.edu/2018/03/19/hyperbolics/

Hyperbolic Embeddings for Microbial Taxonomies

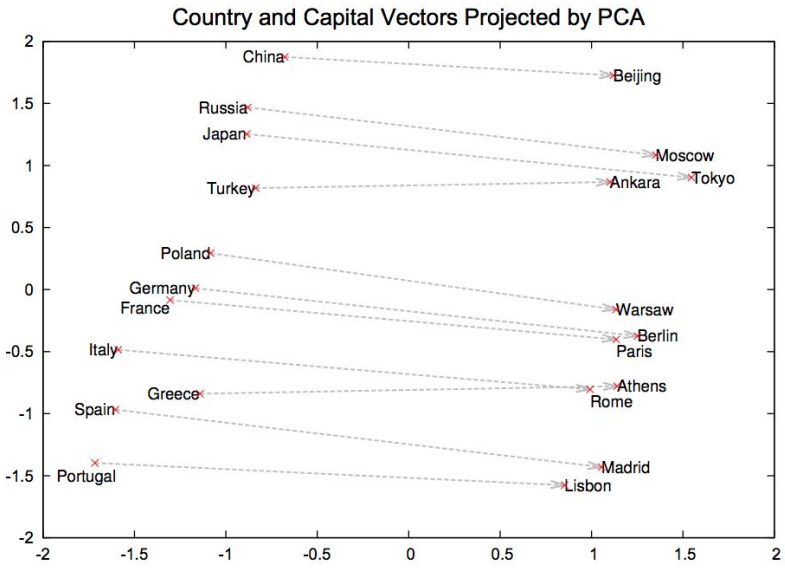
- All life can be classified into a phylogenetic tree describing its evolutionary history
- Nodes in a phylogenetic tree that share an edge share a common ancestry, expected to be close to each other in hyperbolic embedding
- Embedding allows for each taxonomic units location in the tree to be described as a vector
- Vectors for individual microbes can be combined to create a single vector that describes a microbiome semantic structure



Poincare Embeddings for Learning Hierarchical Representations, Nickel and Kiela

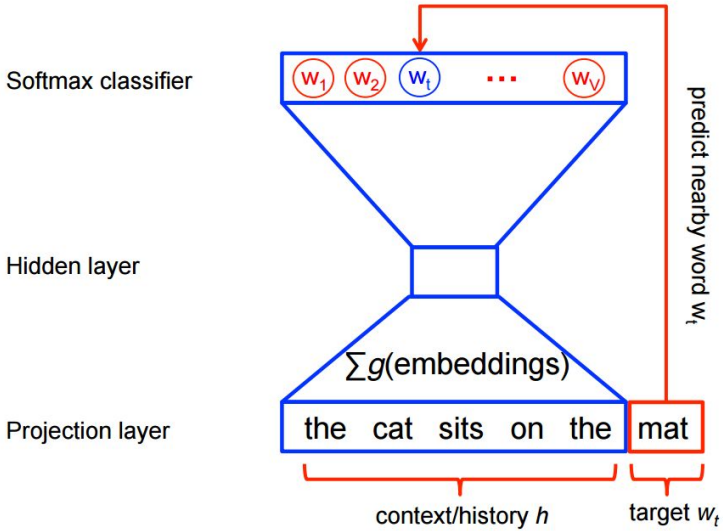
Word2Vec

- NLP Technique, Mikolov et al. 2013
- Create word vectors that represent the semantic meaning of words based on their occurrence in the training corpus
- Ex. Dog and cat should be close in proximity in vector space while dog and queen should be further away



AGP Word2Vec Embeddings

- Built corpus of samples using sequence data
  - Treat each sample as a sentence, OTU as a word, and OTU value as the number of times the word appears in the sentence
- Hyperparameter search for optimal Word2Vec model
  - Number of output dimensions
  - Epochs
  - Minimum times an OTU appears in corpus
- Optimal model embedding each sample into 80 dimensions



Data Pipeline

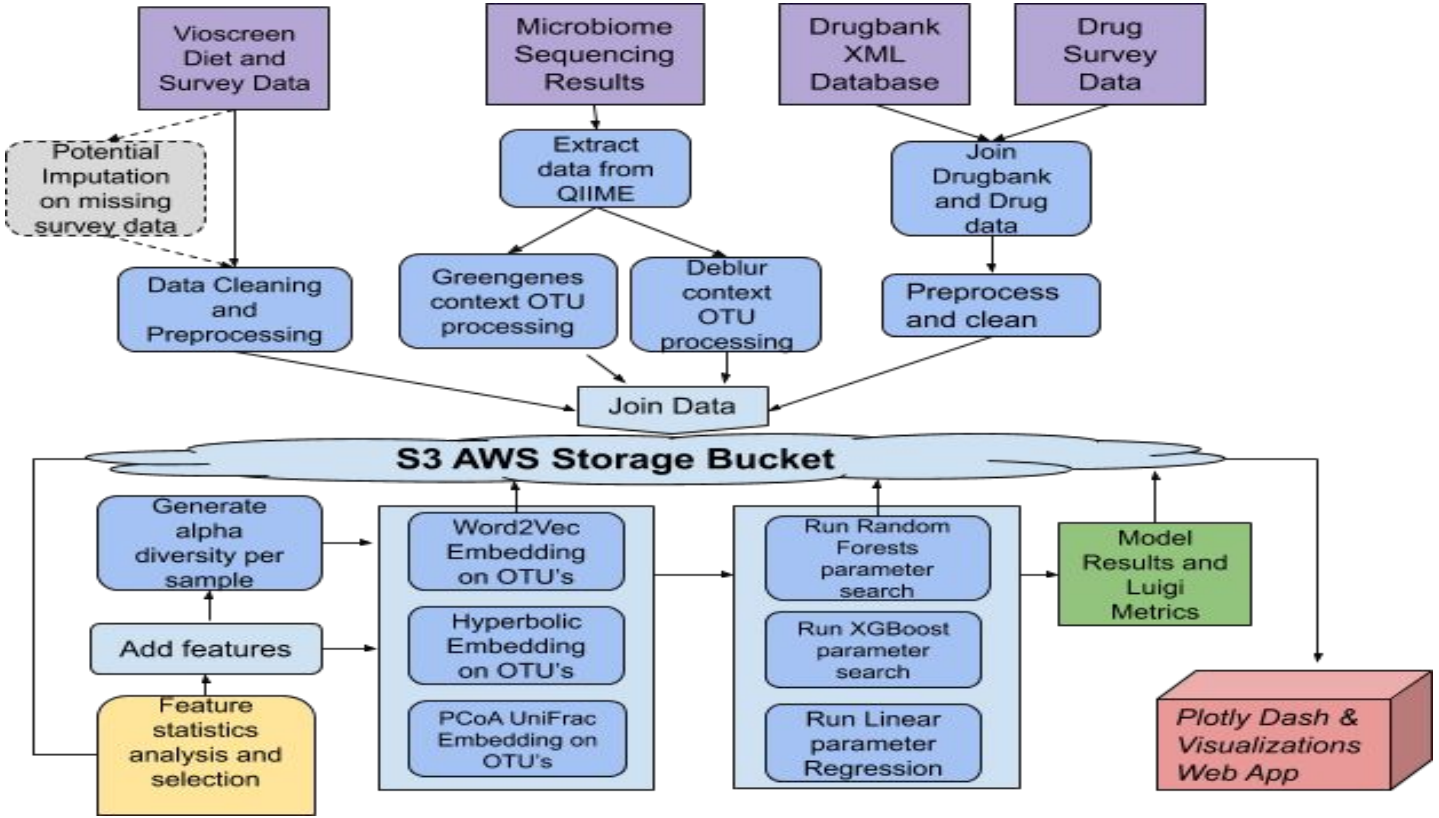
Raw Data Sources

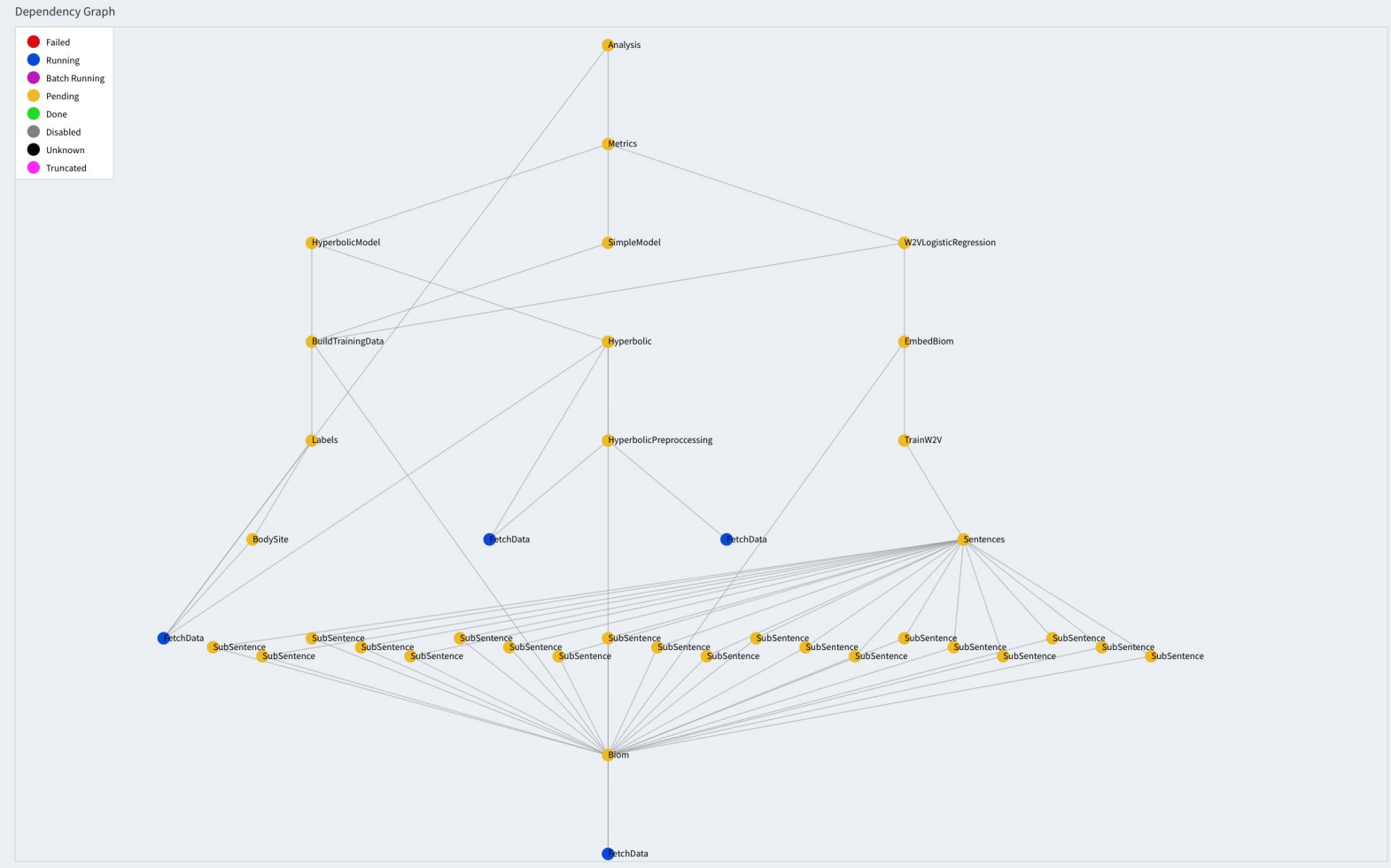
- **16S Metagenomic Sequencing Data**
  - American Gut Projects crowd-sourced citizen scientist collection
- **Vioscreen Food Survey**
  - 2,000 out of 19,000 citizen-scientists filled these surveys, providing specific diet information
- **Drug Use Questionnaire**
  - 19,000+ filled surveys with natural language processed medical drug names
- **Metadata Sample Surveys**
  - Varying results of ~200 questions about person's life, living, and well-being

Data Pipeline

- Spotify Luigi
  - “Luigi is a Python module that helps you build complex pipelines of batch jobs. It handles dependency resolution, workflow management, visualization etc.”
  - https://github.com/spotify/luigi
- Leverage a framework for parallelization and dependency management

Data Integration of Sources




Dependency Graph
Failed
Running
Batch Running
Pending
Done
Disabled
Unknown
Truncated
Analysis
Metrics
HyperbolicModel
SimpleModel
W2VLogisticRegression
BuildTrainingData
Hyperbolic
EmbedBiom
Labels
HyperbolicPreproccessing
TrainW2V
BodySite
FetchData
FetchData
Sentences
FetchData
SubSentence
Biom
FetchData

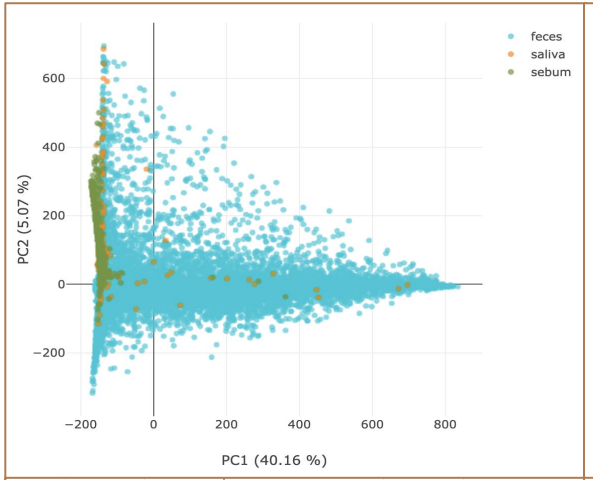
Results

Body Site Classification Results

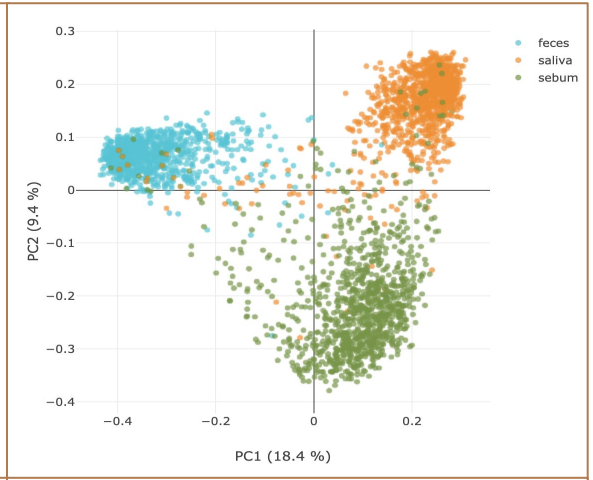
Embedding	Dimensions	Test F1 Score
No embedding	19105	.982
PCoA	3	.958
Hyperbolic	10	.961
Word2Vec	80	.848

Visual Comparison of Embedding Results

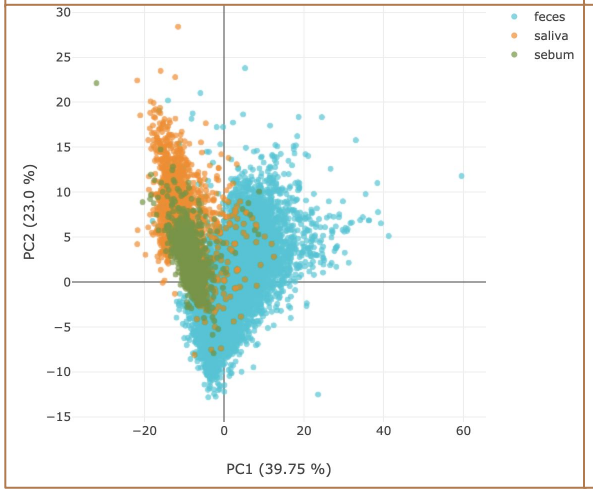
No Embeddings



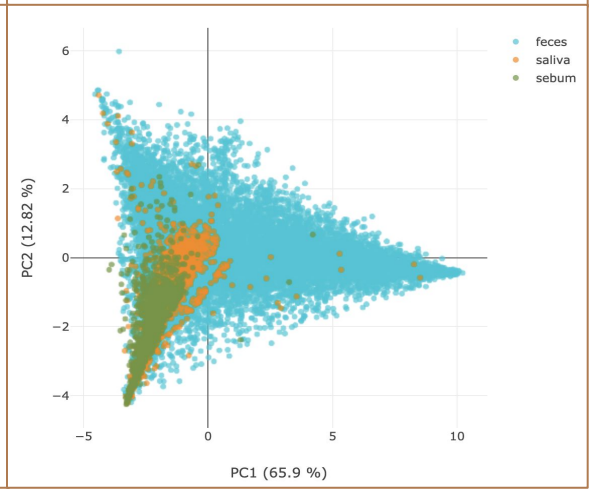
PCoA



Hyperbolic



Word2vec



Visualization Dashboard

The American Gut Project - Body Site Classification

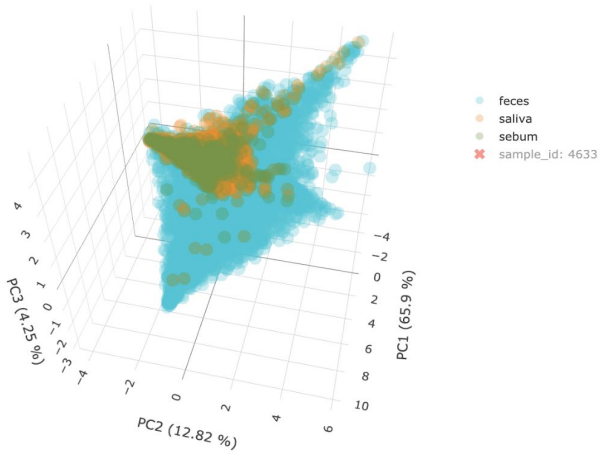
Select Embedding Type:

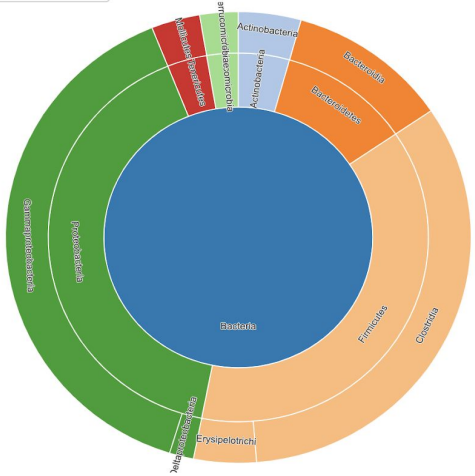
word2vec

2d scatterplot 3d scatterplot

Select Sample ID:

4633





Sample ID: 4633

Body Site: feces

Faith PD Alpha Diversity: 10.56

BMI: Not provided

Age: 73.0

Country: USA

Lat, Lon: 42.0, -87.7

Antibiotics: N/A

Supplements: Vitamins D, E, Calcium CoQ10

Medications: N/A

Interactive Dashboard

- Users can view the difference of embeddings at a granular scale, even down to a single sample and its microbiome taxonomy makeup.
- Interactive and dynamic exploratory analysis of single samples, as well as embedding results
- Implemented with Python, using Plotly’s Dash framework

Demo

Conclusion

- Scalable Ingestion process and pipeline can drastically speed up performing power as well as iterative data analysis of models, features, and targets
- Different embedding techniques were shown to effectively represent the microbiome community structure of each sample in a low dimensional space, despite each one operating on fundamentally different principles
- The semantics of these embeddings are currently not fully appreciated, future work can help to shed light on the differences between them

Acknowledgements

Thank you to our program advisor Ilkay Altintas, our academic advisors Rob Knight and Daniel McDonald, and the thousands of American Gut participants who made this research possible.