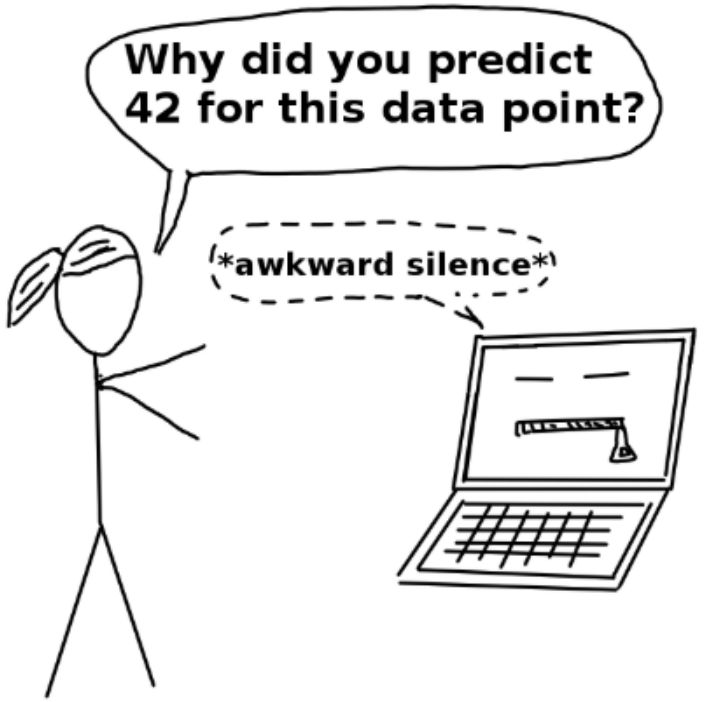
# Outline

- Interpretable models in general
  - Pros/cons of models with built in interpretability vs post-hoc methods
- What the data looks like
- Overview of the datasets
- In depth overview of the Bokulich 2016 study
- Selecting interpretable features
  - MDITRE model as an example
- Example MDITRE results for Bokulich 2016 study

# Interpretability in general

- Notion of interpretability difficult to define in general, is domain specific
- Qualitative metrics: can a human understand why prediction was made
- Why it’s important



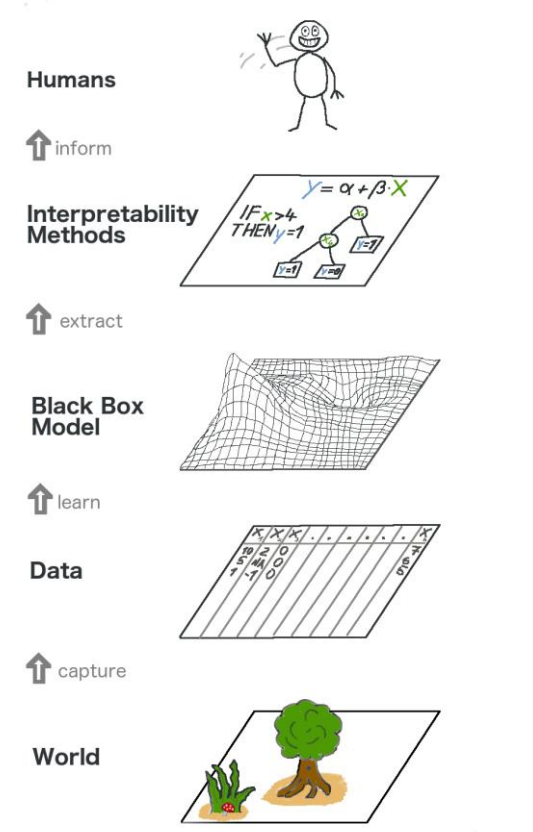

https://christophm.github.io/interpretable-ml-book/terminology.html

# Interpretability in time-series microbiome data

- High stakes medical applications
- Specificity over sensitivity
- For many microbiome applications, the critical tasks are discovering relationships between the microbiome and the host or finding clinically useful biomarkers, rather than pure prediction, care more about model interpretability than predictive power

# interpretability in models

- Want features that are easier to understand/use:
  - Grouping together relevant sets of taxa
  - Focusing on relevant time windows
- Human interpretable model output…
  - Models with interpretability baked-in
    - Regression models: linear, logistic, GLM, GAM
    - Decision trees
    - Rule lists/sets
  - Model agnostic post-hoc methods
    - LIME




https://christophm.github.io/interpretable-ml-book/terminology.html

# Interpretable models vs post-hoc interpretability

- Pros and cons of each
- Post-hoc pro, more general model can maybe pick out weaker signals, con, might be false positive, or not sure why its being picked out

# Abundance data (16S amplicon)

## Abundance data

A CSV file containing the microbial abundances, formatted with the first row providing OTU IDs and the first column providing sample IDs.

```
abundances = pd.read_csv(os.path.join(dataset_path, "abundance.csv"), index_col=0)
```

```
abundances
```

| | Otu000001 | Otu000002 | Otu000003 | Otu000004 | Otu000005 | Otu000006 | Otu000007 | Otu000008 | Otu000009 | Otu000010 | ... | Otu017301 | Otu017302 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **DD10** | 5629 | 0 | 623 | 0 | 291 | 0 | 0 | 1263 | 1961 | 515 | ... | 0 | 0 |
| **DD102** | 5194 | 0 | 218 | 0 | 674 | 0 | 0 | 2307 | 560 | 0 | ... | 0 | 0 |
| **DD104** | 5292 | 0 | 81 | 634 | 2518 | 0 | 1938 | 2009 | 0 | 691 | ... | 0 | 0 |
| **DD106** | 1780 | 0 | 164 | 0 | 384 | 0 | 0 | 934 | 1798 | 865 | ... | 0 | 0 |
| **DD107** | 6046 | 0 | 811 | 0 | 69 | 3 | 0 | 0 | 234 | 459 | ... | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **ID92** | 5963 | 0 | 815 | 0 | 29 | 0 | 2458 | 674 | 386 | 8999 | ... | 0 | 0 |
| **ID95** | 11834 | 0 | 1650 | 467 | 1184 | 0 | 0 | 0 | 1327 | 0 | ... | 0 | 0 |
| **ID97** | 538 | 30569 | 179 | 60 | 396 | 19910 | 0 | 0 | 1398 | 20 | ... | 0 | 0 |
| **ID98** | 9981 | 0 | 5211 | 0 | 1602 | 0 | 7275 | 0 | 0 | 898 | ... | 5 | 0 |
| **ID99** | 32485 | 0 | 2763 | 0 | 395 | 7 | 6 | 409 | 3999 | 464 | ... | 0 | 0 |

236 rows × 17310 columns

Will probably want to do some data cleaning before hand; remove otus with low counts or that are only present in a few samples/time points;

- might want to only focus on some time periods (e.g. if most subjects do not have samples outside of a given time window)
- Discard subjects with too few samples, etc…

# Abundance (shotgun metagenomics)

## MetaPhlAn abundance tables

Note that MetaPhlAn outputs organism relative abundances (out of 100%), listed as one clade per line. The first column lists clades, ranging from taxonomic kingdoms (Bacteria, Archaea, etc.) through species. The taxonomic level of each clade is prefixed to indicate its level: `Kingdom: k__, Phylum: p__, Class: c__, Order: o__, Family: f__, Genus: g__, Species: s__`. The total sum of relative abundances for each clade should then sum to 100.0.

```
abundances = pd.read_csv(os.path.join(dataset_path, "diabimmune_t1d_metaphlan_table.txt"), sep="\t")
```

```
abundances
```

| | Taxonomy | G35421 | G35451 | G35893 | G35464 | G35465 | G35474 | G35488 | G35906 | G35951 | ... | G36267 | G36268 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | k__Bacteria | 100.00000 | 100.00000 | 100.00000 | 100.00000 | 100.00000 | 100.00000 | 100.00000 | 100.00000 | 100.00000 | ... | 100.00000 | 100.00000 |
| **1** | k__Bacteria\|p__Actinobacteria | 36.48375 | 3.45029 | 0.97899 | 46.08772 | 17.65351 | 72.96490 | 59.17006 | 2.76232 | 0.49062 | ... | 46.72820 | 1.95127 |
| **2** | k__Bacteria\|p__Actinobacteria\|c__Actinobacteria | 36.48375 | 3.45029 | 0.97899 | 46.08772 | 17.65351 | 72.96490 | 59.17006 | 2.76232 | 0.49062 | ... | 46.72820 | 1.95127 |
| **3** | k__Bacteria\|p__Actinobacteria\|c__tinobacteri... | 0.00941 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00895 | 0.07594 | 0.00000 | ... | 0.02479 | 0.00000 |
| **4** | k__Bacteria\|p__Actinobacteria\|c__Actinobacteri... | 0.00941 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00895 | 0.00000 | 0.00000 | ... | 0.02479 | 0.00000 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **375** | k__Bacteria\|p__Verrucomicrobia\|c__Verrucomicro... | 4.62994 | 0.00000 | 0.08629 | 0.00000 | 0.02831 | 0.01743 | 0.00000 | 4.35279 | 0.00000 | ... | 3.32411 | 0.00000 |
| **376** | k__Bacteria\|p__Verrucomicrobia\|c__Verrucomicro... | 4.62994 | 0.00000 | 0.08629 | 0.00000 | 0.02831 | 0.01743 | 0.00000 | 4.35279 | 0.00000 | ... | 3.32411 | 0.00000 |
| **377** | k__Bacteria\|p__Verrucomicrobia\|c__Verrucomicro... | 4.62994 | 0.00000 | 0.08629 | 0.00000 | 0.02831 | 0.01743 | 0.00000 | 4.35279 | 0.00000 | ... | 3.32411 | 0.00000 |
| **378** | k__Bacteria\|p__Verrucomicrobia\|c__Verrucomicro... | 4.62994 | 0.00000 | 0.08629 | 0.00000 | 0.02831 | 0.01743 | 0.00000 | 4.35279 | 0.00000 | ... | 3.32411 | 0.00000 |
| **379** | k__Bacteria\|p__Verrucomicrobia\|c__Verrucomicro... | 4.62994 | 0.00000 | 0.08629 | 0.00000 | 0.02831 | 0.01743 | 0.00000 | 4.35279 | 0.00000 | ... | 3.32411 | 0.00000 |

380 rows × 125 columns

# Metadata

**Sample metadata**

| | sample_ID | subject_ID | time |
|---|---|---|---|
| 0 | DD2 | Plant5 | 3.0 |
| 1 | DD3 | Plant7 | 4.0 |
| 2 | DD4 | Plant7 | 3.0 |
| 3 | DD5 | Plant4 | 2.0 |
| 4 | DD6 | Plant8 | -1.0 |
| ... | ... | ... | ... |
| 231 | ID262 | Animal3 | 8.0 |
| 232 | ID263 | Animal4 | 10.0 |
| 233 | ID264 | Animal5 | 5.0 |
| 234 | ID265 | Animal1 | -2.0 |
| 235 | ID266 | Animal8 | 8.0 |

A CSV file that specifies an associated subject ID and timepoint for each sample ID.

**Subject metadata**

| | subject_ID | diet |
|---|---|---|
| 0 | Plant5 | Plant |
| 1 | Plant7 | Plant |
| 2 | Plant4 | Plant |
| 3 | Plant8 | Plant |
| 4 | Plant6 | Plant |
| 5 | Plant9 | Plant |
| 6 | Plant3 | Plant |
| 7 | Plant1 | Plant |
| 8 | Plant10 | Plant |
| 9 | Plant2 | Plant |
| 10 | Animal11 | Animal |

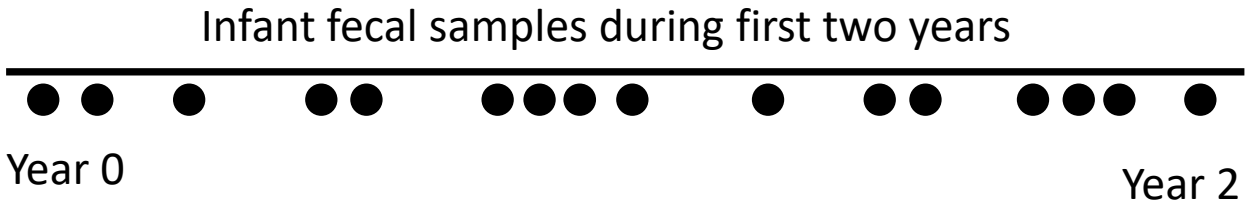
A CSV file that gives information about each subject, (including the value of whatever variable will be used as the host outcome for prediction (e.g., Plant-diet or Animal-diet in the David et al).

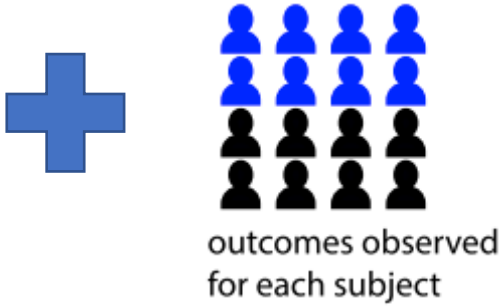
# Datasets overview

| Study | Subjects | Type | Classification tasks |
|---|---|---|---|
| Bokulich 2016 | Gut microbiomes of infants sampled over the first two years of life | 16S | (a) Diet (breast fed vs formula)<br>(b) Mode of birth (vaginal or c-section) |
| Brooks 2017 | Gut microbiomes of 30 infants sampled over 75 days | MAG? | Mode of birth (vaginal versus C-section) |
| David 2014 | Microbiomes of 20 healthy adults receiving dietary interventions | 16S | Diet (plant based vs animal) |
| DiGiulio 2015 | Vaginal microbiomes of 37 pregnant women | 16S | Delivery time (at term vs pre-term) |
| Kostic 2015 | Gut microbiomes of 17 infants sampled over the first 3 years of life | MAG | Normal vs development of type 1 diabetes |
| Shao 2019 | Gut microbiomes of 282 infants (after filtering for subjects with fewer than three timepoints) sampled over 424 days | MAG | Mode of birth (vaginal versus C-section) |
| Vatanen 2016 | Gut microbiomes of 117 children sampled over the first three years of life | 16S | Nationality (Russian versus Estonian/Finnish) |

*Shao dataset has very little time-series, may not want to use

# Bokulich 2016 study




- Aimed to characterize early-life microbial development in the context of antibiotic use, cesarean section, and formula feeding
- total of 43 infants were enrolled for follow-up for up to the age of 2 years
- Stool samples collected and sequenced; 151-bp paired end sequencing on the Illumina MiSeq platform
- Original study: Operational taxonomic units (OTUs) were assigned using QIIME's uclust; data you have has been reprocessed with dada2 to obtain tables of OTU abundances and phylogenetic placements for each OTU on a reference tree

- Antibiotic exposure
  - Could be at any point over the 2 years
- Mode of birth
- Diet (Breast fed vs formula)
  - during the first 3 months of life

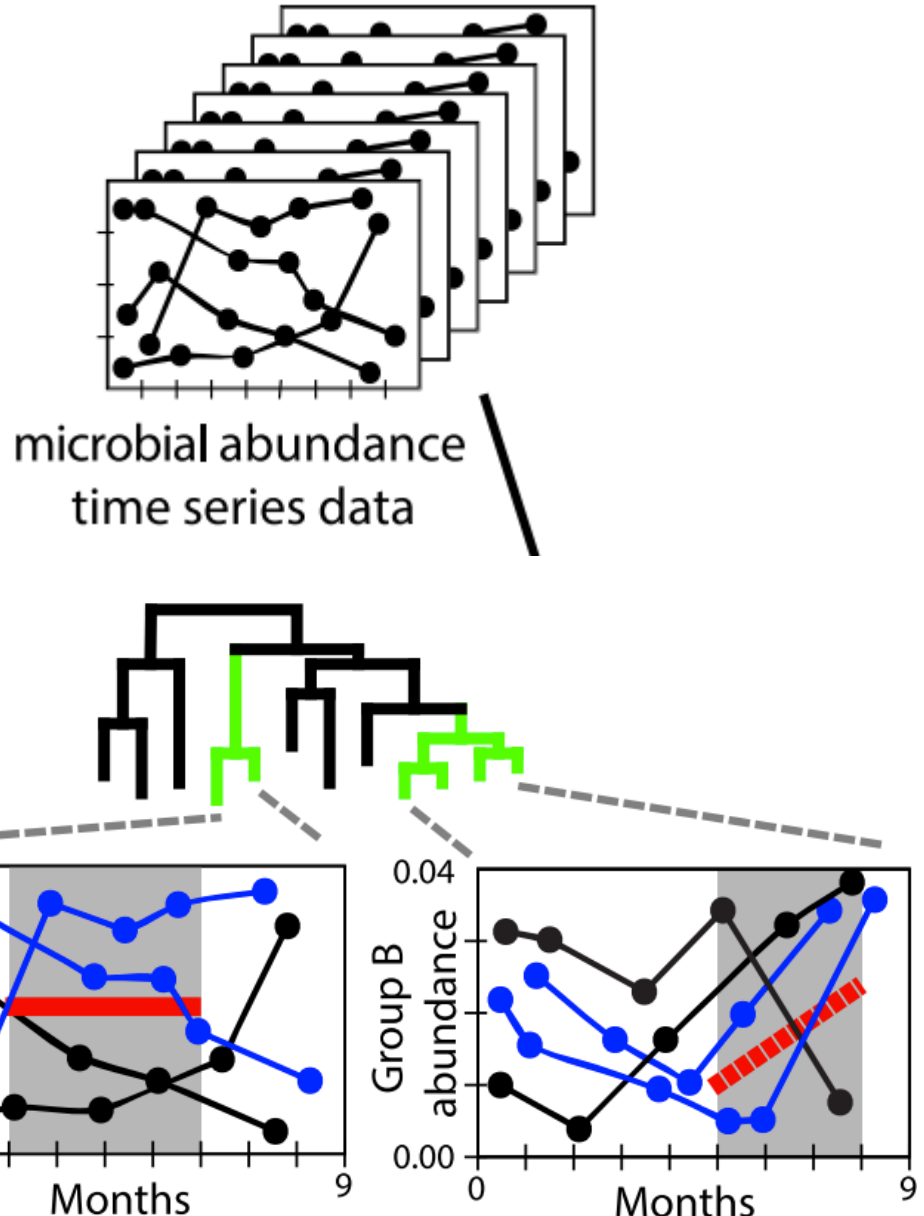
# Example research question

- Can we predict which subjects were breast fed or formula fed from the microbial time series data?

- What features would be relevant predictors?
  - Groups of taxa with expected similar functions [phylogeny]
  - Relevant time windows
  - Want to focus on finding only the essential features (in this case, microbial clades and relevant time windows), for better interpretability and possibly better model performance

# Selecting relevant features

- Can aggregate by taxonomic rank for example
- Can pick top abundance bugs to start; will want to do some sort of filtering in general
- Might be better to work with sequences directly
- Have access to reference trees, use distance metric
- generating time windows



# MDITRE model overview

- Aggregating phylogenetically similar microbes
- Aggregating microbes over time windows

- Possible features missed by mditre model

# Example MDITRE prediction for Bokulich study