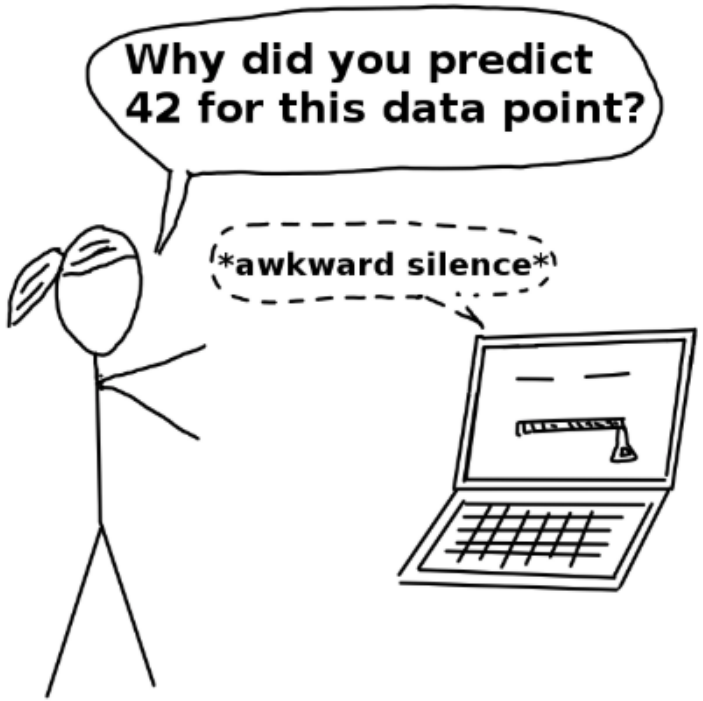
Outline

- Interpretable machine learning
- Overview of the Bokulich 2016 study
- Selecting interpretable features
  - MDITRE model as an example
- Example MDITRE results for Bokulich 2016 study
- General overview of the datasets and what the data looks like

Interpretability

- Notion of interpretability difficult to define in general and is domain specific
- Non-mathematical definitions:
  - “Interpretability is the degree to which a human can understand the cause of a decision” (Miller 2017)
  - “Interpretability is the degree to which a human can consistently predict the model’s result” (Kim et al 2016)



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

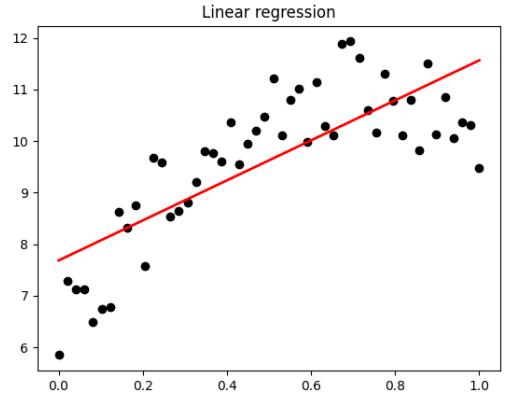
Miller, Tim. “Explanation in artificial intelligence: Insights from the social sciences.” arXiv (2017)
Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. “Examples are not enough, learn to criticize! Criticism for interpretability.” Advances in Neural Information Processing Systems (2016)

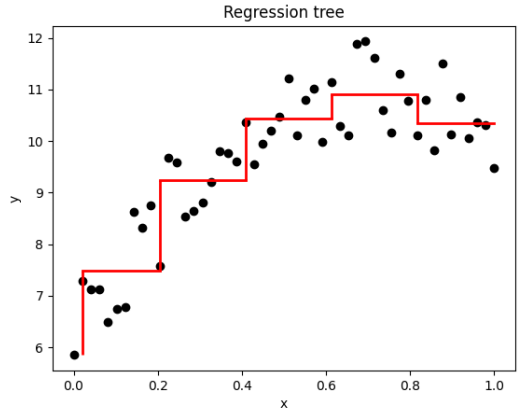
Importance of interpretability

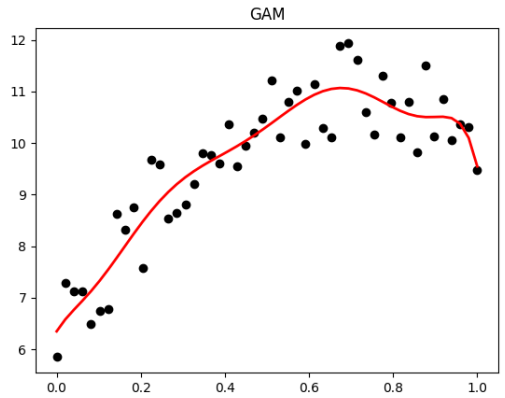
- Why not just trust the model predictions?
  - Ok for low stakes situations (e.g. recommending a movie to watch)
- Model trust and safety: can better trust the model if we understand **why** a decision was made
- Interpretability makes it possible to extract knowledge from the learned model
  - For many microbiome applications, the critical tasks are discovering relationships between the microbiome and the host or finding clinically useful biomarkers
  - Want to know ***why*** a prediction was made more so than just ***what*** the prediction is
- Want to validate predictions, need to be able to generate specific testable hypotheses and clinical tests
  - Favor specificity over sensitivity
  - Ok taking a hit on predictability; want to pick out few interactions we can understand clearly

Intrinsically interpretable models

- Regression models:
  - Linear
    - Can interpret learned weights
  - Generalized Linear Models
    - E.g. logistic regression
  - Generalized Additive Models
    - Interpret from plots
- Decision trees
  - Capture feature interactions, nonlinearities
  - Regression or classification
- Rule lists/sets
  - IF-THEN statements
  - Features combined with AND's
  - can be as expressive as decision trees, while being more compact

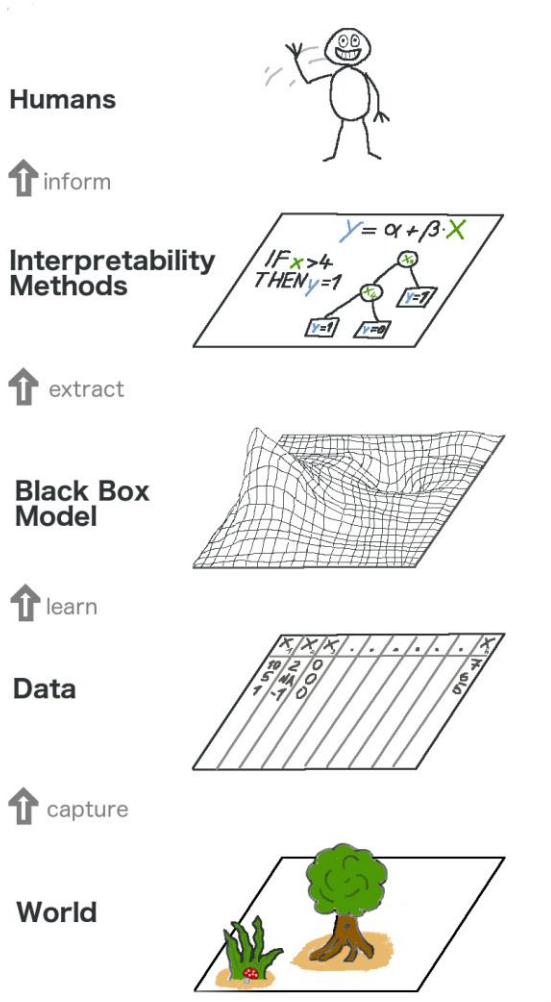






Post-hoc interpretation

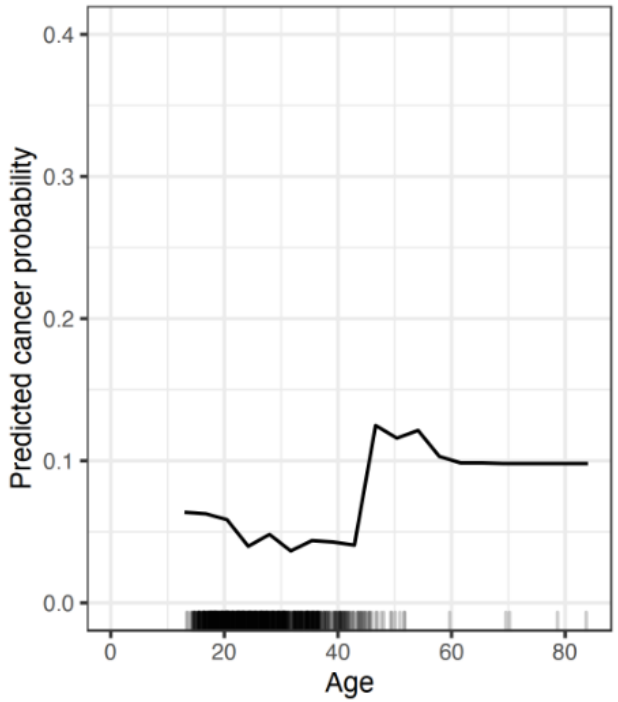
- Model agnostic
  - Can work with any black box model
  - Usually work by analyzing feature input and output pairs
  - Cannot have access to model internals such as weights or structural information
- Global methods explain the entire model behavior
  - Partial dependence plot (PDP)
  - Permutation feature importance
  - Global surrogate models
- Local methods seek to explain individual predictions
  - Local surrogate models (LIME)
  - Individual conditional expectation (ICE)
  - Shapely values



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

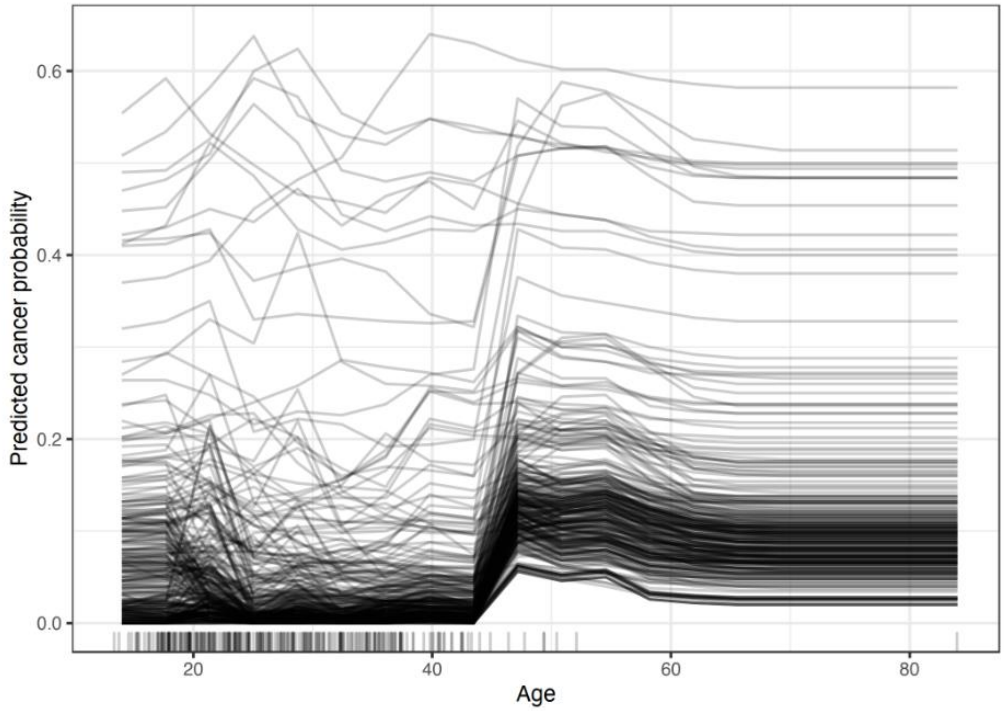
Global vs local methods

Partial dependence plots (PDP)



GLOBAL: PDP shows the marginal effect one or two features have on the predicted outcome of a machine learning model

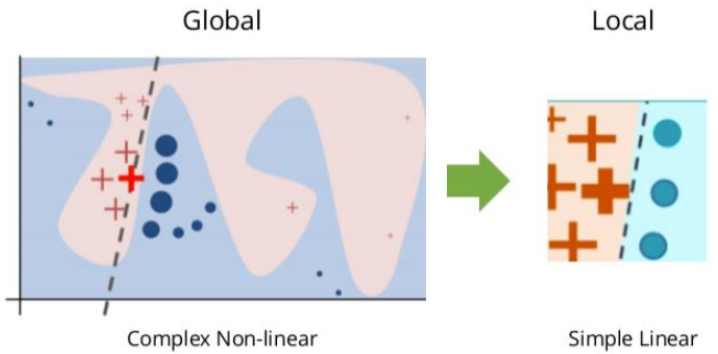
Individual conditional expectation (ICE)

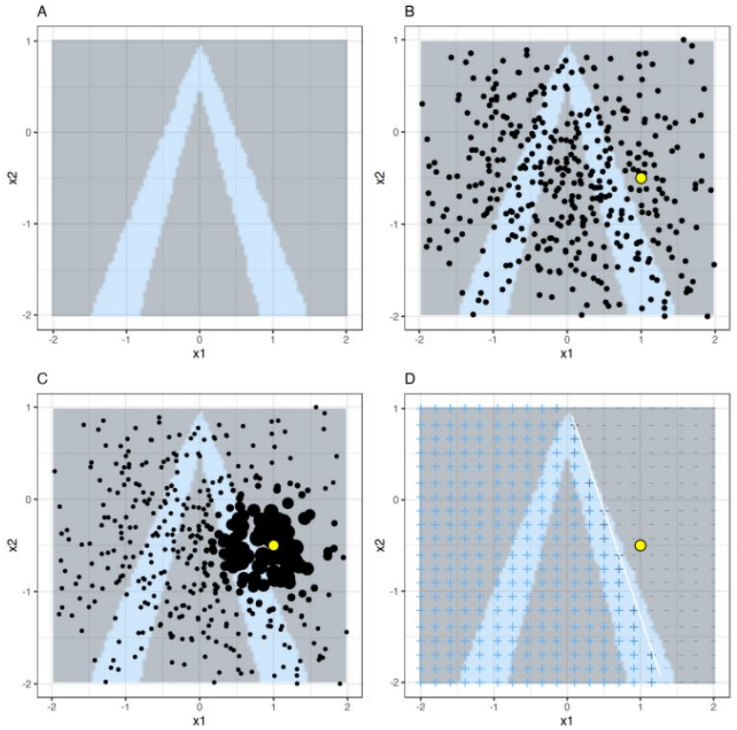


LOCAL: ICE plots display one line per instance, showing how each instance's prediction changes when a feature changes

Local surrogate models (LIME)

- LIME generates a new dataset consisting of perturbed samples around an instance of interest and the corresponding predictions of the black box model
- Train a weighted, interpretable model on the generated dataset
- Can then explain the prediction by interpreting the local model



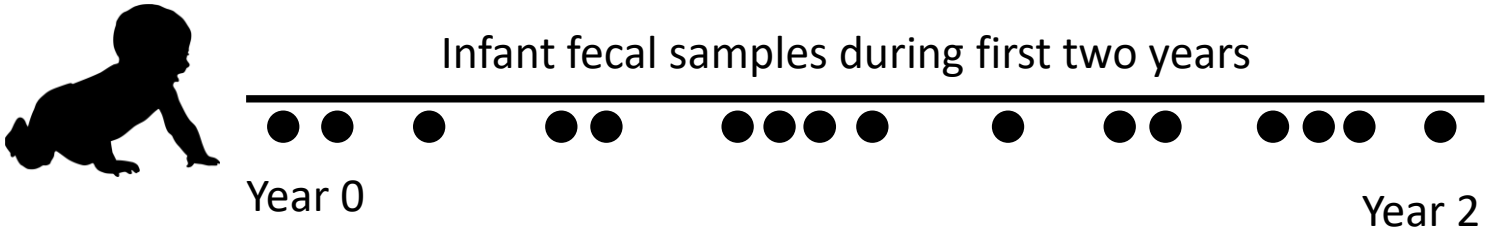


https://christophm.github.io/interpretable-ml-book/lime.html

LIME advantages and disadvantages

- Advantages:
  - Black box models might perform better for prediction; can maybe pick out weaker signals
  - Model agnostic, can be applied to any model in general, even on already interpretable models
  - Explanations created with local surrogate models can use other (interpretable) features than the original model was trained on
    - E.g. regression model could be trained on components of a principal component analysis (PCA) of answers to a survey, but LIME might be trained on the original survey questions
- Disadvantages:
  - Might pick out more false positives, may not be as sparse
  - For LIME, complexity of the explanation model has to be defined in advance
  - Explanations can be unstable, explanations of close points can potentially vary greatly

Bokulich 2016 study



- 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 and pplacer to obtain tables of OTU abundances and phylogenetic placements for each OTU on a reference tree
- Host labels for antibiotic exposer, mode of birth, and diet

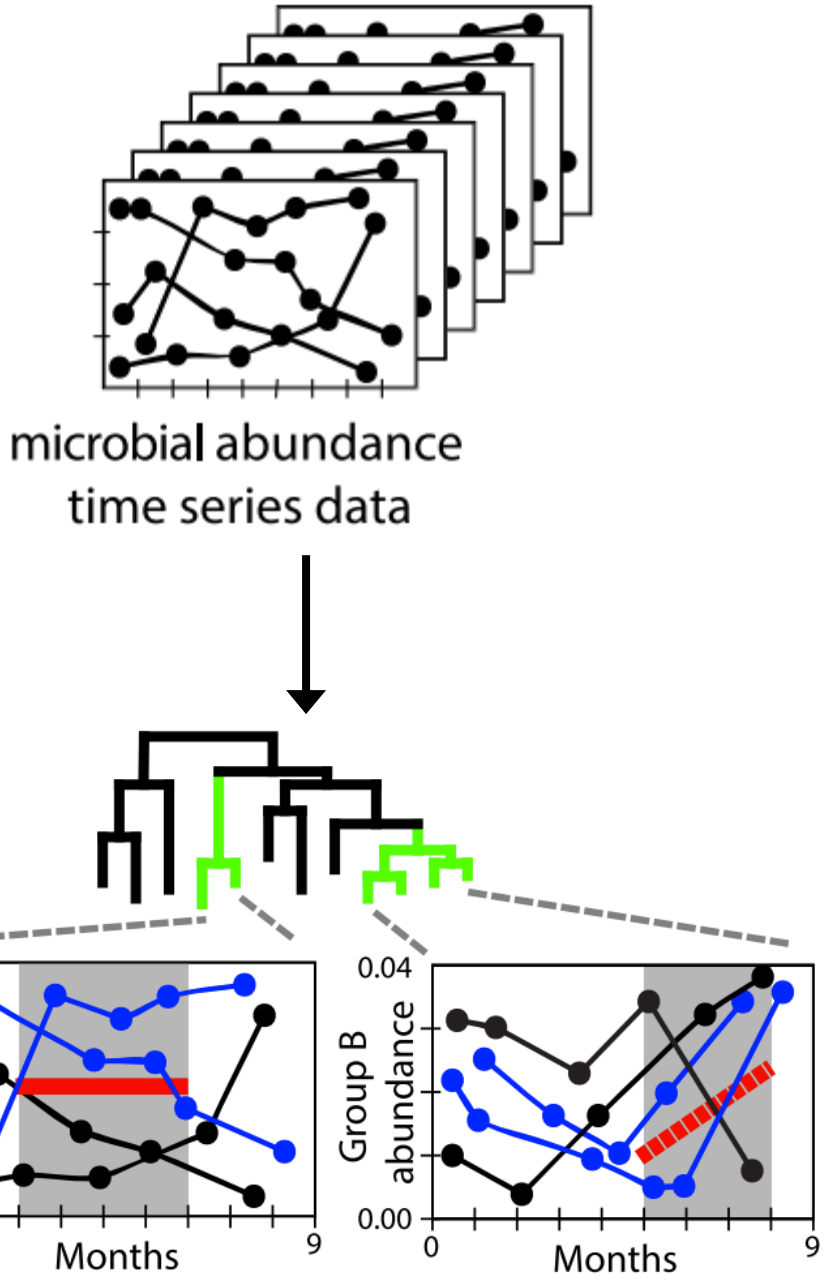
- Antibiotic exposure
  - Could be at any point over the 2 years
- Mode of birth
- Diet (Breast fed vs formula)
  - during the first 3 months of life

Bokulich 2016 study

- Motivation:
  - Disruptions to microbiome development has been associated with conditions emerging later in life, including obesity, diabetes, and allergies
  - Microbial taxa that best define “microbial age” can be used as biomarkers to track infant microbiota progress, paralleling how weight-for-height tracks child development, both affected by disruptions to health
  - Aimed to characterize early-life microbial development in the context of antibiotic use, cesarean section, and formula feeding
- Example research question:
  - Can we predict which subjects were breast fed or formula fed from the microbial time series data?
  - From our model, can we interpret why the prediction was made?
    - E.g.: What are the relevant microbes and time windows?

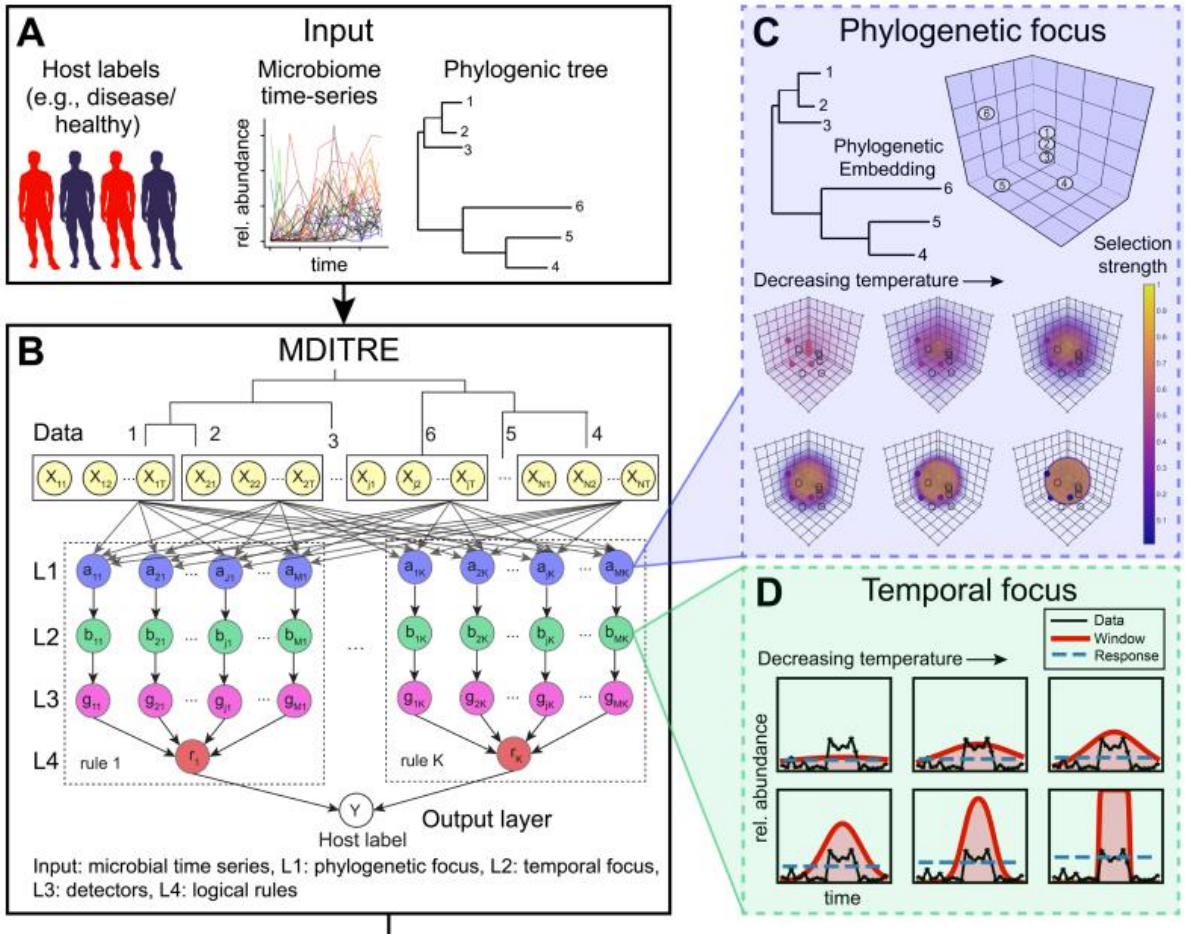
Selecting relevant features

- Data consists of many (~ 1000+) OTUs (operational taxonomic units)
- Possibly many time points (100s)
- Want few features that are easy to understand
- Reducing number of OTUs:
  - Can pick top few in terms of abundance
  - Can aggregate by taxonomic rank
  - Can work with sequences directly; e.g. group by hamming distance
  - Have access to reference trees, use distance metric
- Reducing time points:
  - Group together time points into time windows
  - Note: subjects are synchronized in time in these datasets, MDITRE relies on this with time windows
  - Many other time-series methods rely on long time-series, these methods won't apply well here



MDITRE model overview

- Phylogenetic focus layer:
  - Model takes phylogenetic reference tree and computes distances between OTUs
  - Embed distance matrix into latent space of reduced dimensionality via PCoA
  - Learn groups of closely related taxa with phylogenetic window
- Aggregating microbes over time windows
  - Learns relevant time windows
  - Computes average abundances and rates of change of abundances of phylogenetically focused groups of microbes
- Possible features missed by MDITRE model:
  - Time lags and features such as “IF A increases X days <u>after</u> B THEN, …”
  - Periodic time features



Example MDITRE prediction for Bokulich study

Rule 1: TRUE for Formula with log-odds -2.39 IF:

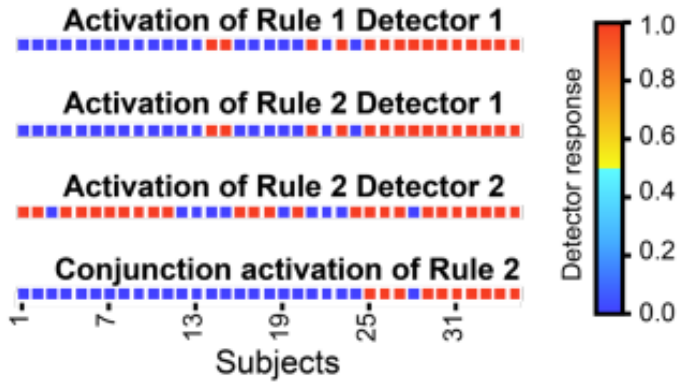
Detector 1: the average abundance of selected taxa between days 118 and 181 is greater than 6.2206%

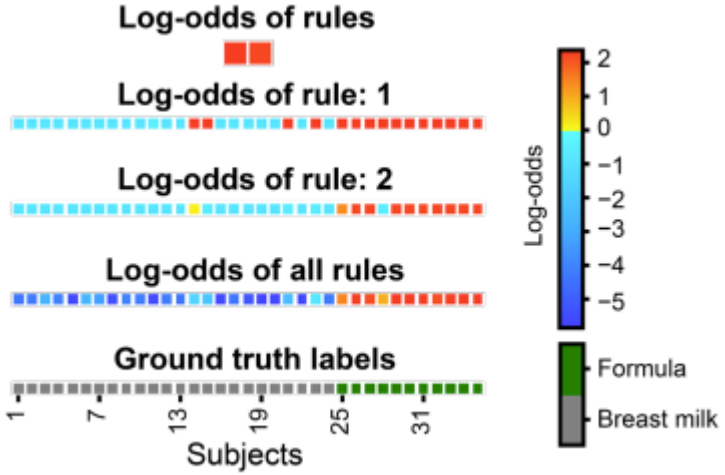
Rule 2: TRUE for Formula with log-odds 2.25 IF:

Detector 1: the average abundance of selected taxa between days 118 and 183 is greater than 7.1523%

AND

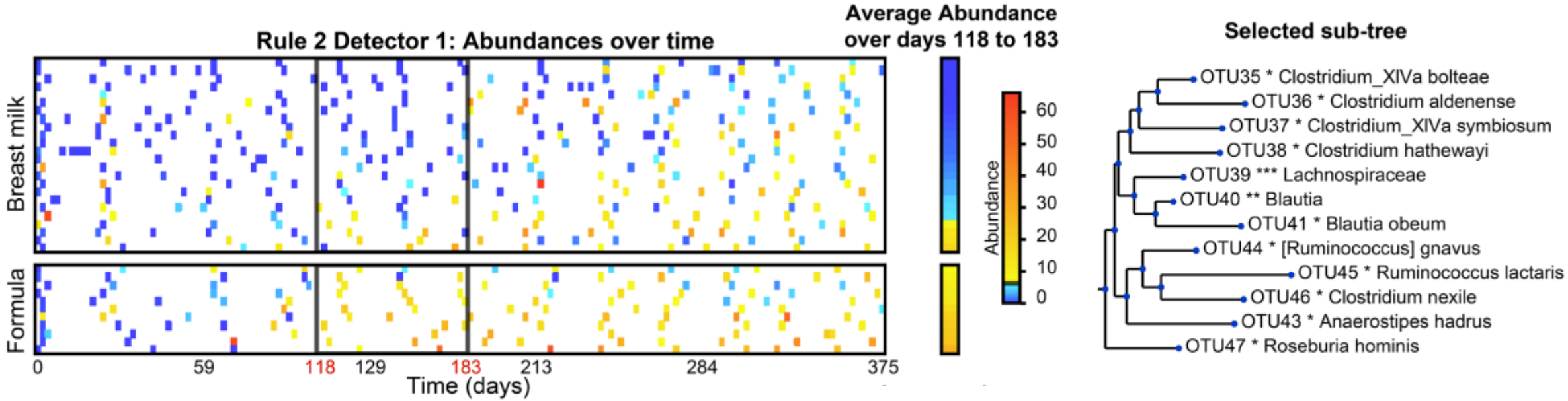
Detector 2: the average slope of selected taxa between days 118 and 190 is greater than -0.0064% per day





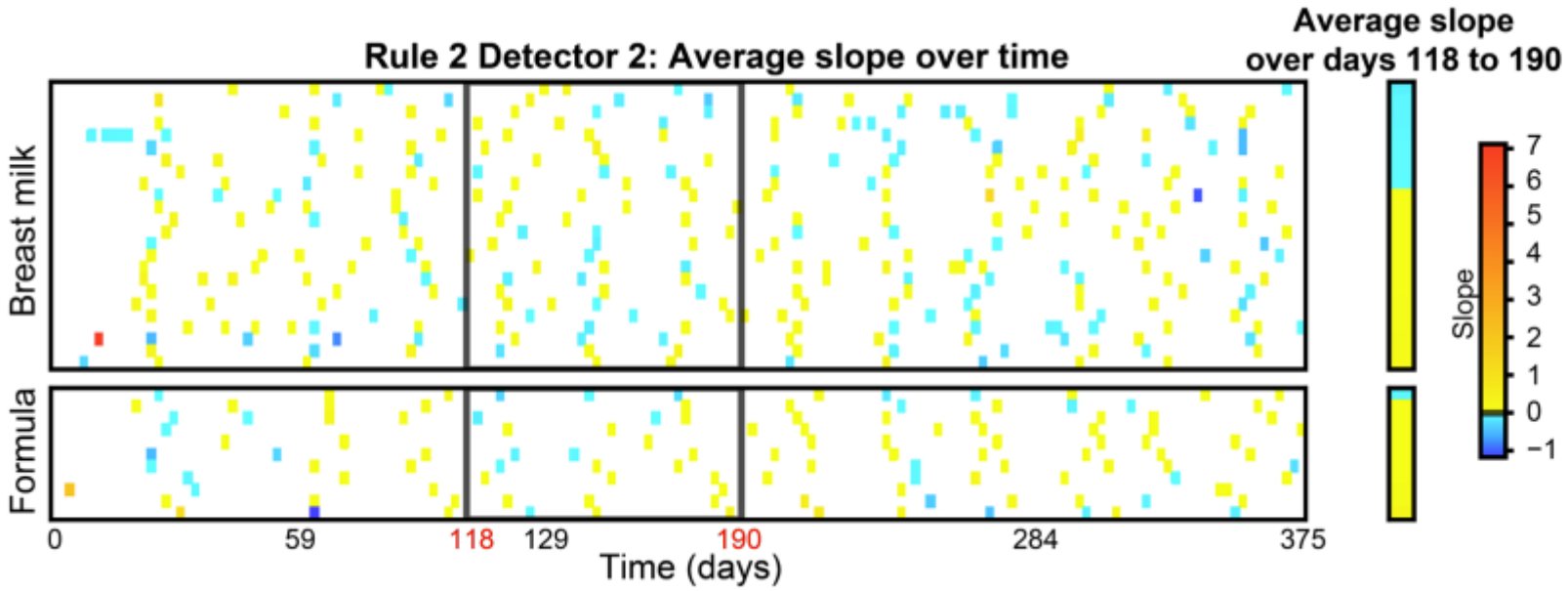
Combining both rules together gives clear separation of both groups with high odds

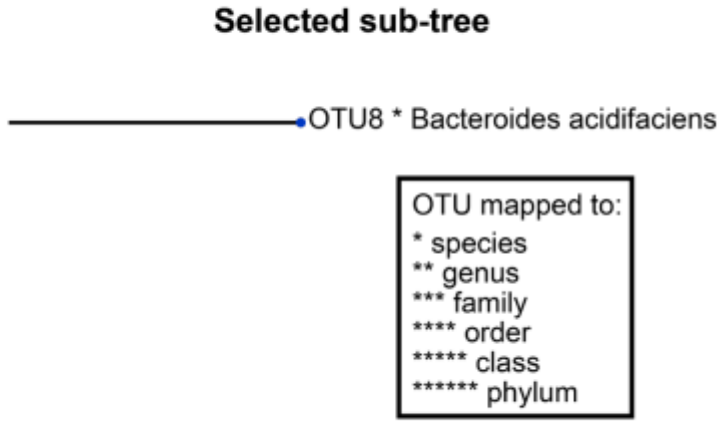
Second rule, first detector



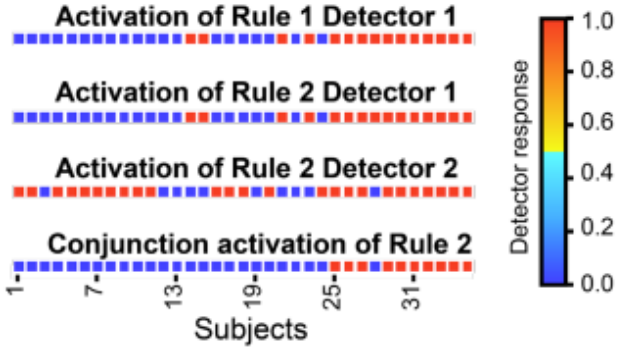
- Both detectors focus on same time window,
  - Around when most infants are introduced to solid foods, which may occur earlier for formula-fed infants
  - From visual of rules, can see that after the window, abundances become more difficult to distinguish; may suggest similar diets post-liquid foods in both groups
- First detector selected selected twelve taxa in the Order Clostridiales
  - Detects all the formula-fed infants and four of the breast-fed infants
  - Many of the selected taxa are strict anaerobes that metabolize more complex nutrient sources, including starches and lipids which may be present in formula

Second rule, second detector





- Second detector focuses on the rate of increase of a single taxon, Bacteroides acidifaciens; TRUE if this taxon is increasing
- Bacteroides acidifaciens has been shown to increase with higher fiber diets
- Detector picks out all but one formula-fed infant but also many breast-fed infants; but is false for the four breastfed infants identified by the first detector



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) (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

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

- Note: Will probably want to do some data preprocessing
- Remove OTUs with low counts or that are only present in a few samples/time points
- Discard subjects with too few samples
- Discard time points with only few subjects
- Can look to MITRE and MDITRE papers for some guidelines

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__Actinobacteri...	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

To filter at the species level, for example, can search for the pattern '**s__**'

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)