# Summary of our SVM models

| SVM Model Name (#) | Classifier | Sample Set | What question(s) does this model allow us to explore? |
|---|---|---|---|
| **Disease-All (1)** | NTM disease | Comprehensive | Which microbiome features distinguish samples from subjects with/without NTM disease? |
| **Disease-Main (2)** | NTM disease | All "no disease" samples, plus MAC/Mab. (+) samples only for NTM disease group | Which microbiome features distinguish samples from subjects with/without MAC/Mab. (+) samples? |
| **Disease-Index (3)** | NTM disease | Index sample (or nearest) and prior sample only | What community changes precede NTM first (+) culture? |
| **Persist-All (4)** | Persistent/transient | Comprehensive | What microbiome features distinguish samples from subjects with persistent/transient infection? |

- **Comprehensive:** All samples listed in: https://github.com/caverlyl/NTM/blob/master/analysis/sample_list.csv
- **NTM disease:** Sample came from a subject with at least one (+) NTM culture
- **Persistent/Transient**: Persistent: Sample comes from a subject that has more than one (+) culture in first year)

# How do we interpret our SVM results?

| SVM Model | What question does this model allow us to explore? | How does the SVM output allow us to explore this question? |
|---|---|---|
| **Disease-All (1)** | Which microbiome features distinguish samples from subjects with/without NTM disease? | ***Microbiome*** features with the ***highest F-scores*** distinguish samples with/without NTM disease |
| **Disease-Main (2)** | Which microbiome features distinguish samples from subjects with/without MAC/Mab. (+) samples? | ***Microbiome*** features with the ***highest F-scores*** distinguish samples with/without NTM disease |
| **Disease-Index (3)** | What community changes precede NTM first (+) culture? | ***Regression*** features with the ***highest F-scores*** are community changes that distinguish samples from subjects with/without (+) NTM culture |
| **Persist-All (4)** | What microbiome features distinguish samples from subjects with persistent/transient infection? | ***Microbiome*** features with the ***highest F-scores*** distinguish samples from subjects with/without persistent infection |

# Microbiome features in our SVM models

**Ecological features**

- Relative abundance at varying taxonomic ranks (OTU, genus, family, order)
- Total relative abundance of OTUs belonging to the following groups:
  - Strict and facultative anaerobes
  - "CF pathogen"
  - Oral microbiota
- All *mothur* α-diversity measures and community type
- Bray-Curtis/Shannon-Beta (β-diversity measures from *mothur/entropart*, respectively)
- γ-diversity and α/β/γ entropy (Defaults in *entropart*)

**Regression**

- The following subject-specific linear regression results for all subjects with >1 sample, using all structural features listed at left:
  - Slope
  - y-intercept
  - $R^2$
  - P-value

**Sample-set-specific regression**

- Regression as above with only the sample prior to first (+) NTM culture and the nearest sample to the index date

# Remaining features in our SVM model

- Year of index NTM culture (to evaluate periodic effects of “NTM disease” diagnosis)
- Age of sample relative to date of first (+) NTM index culture
- BETR

# Feature labels/definitions explained

- **Strict/facultative anaerobes:** Extracted from *Bergey's Manual of Determinative Bacteriology*, 6th Edition, 1948.
- **"CF pathogens":** Pseudomonas, Achromobacter, Staphylococcus, Burkholderia, and Stenotrophomonas
- **Oral microbiota:** List from Fig. 1B of Welch, J. L., et al, *PNAS*, **2016**, E791-E800.
- **α-diversity:** Community distribution of specified taxonomic rank within a single sputum sample
- **β-diversity:** Distribution of community structures when comparing multiple samples from the same subject
- **γ-diversity:** The product of α- and β- diversity measures

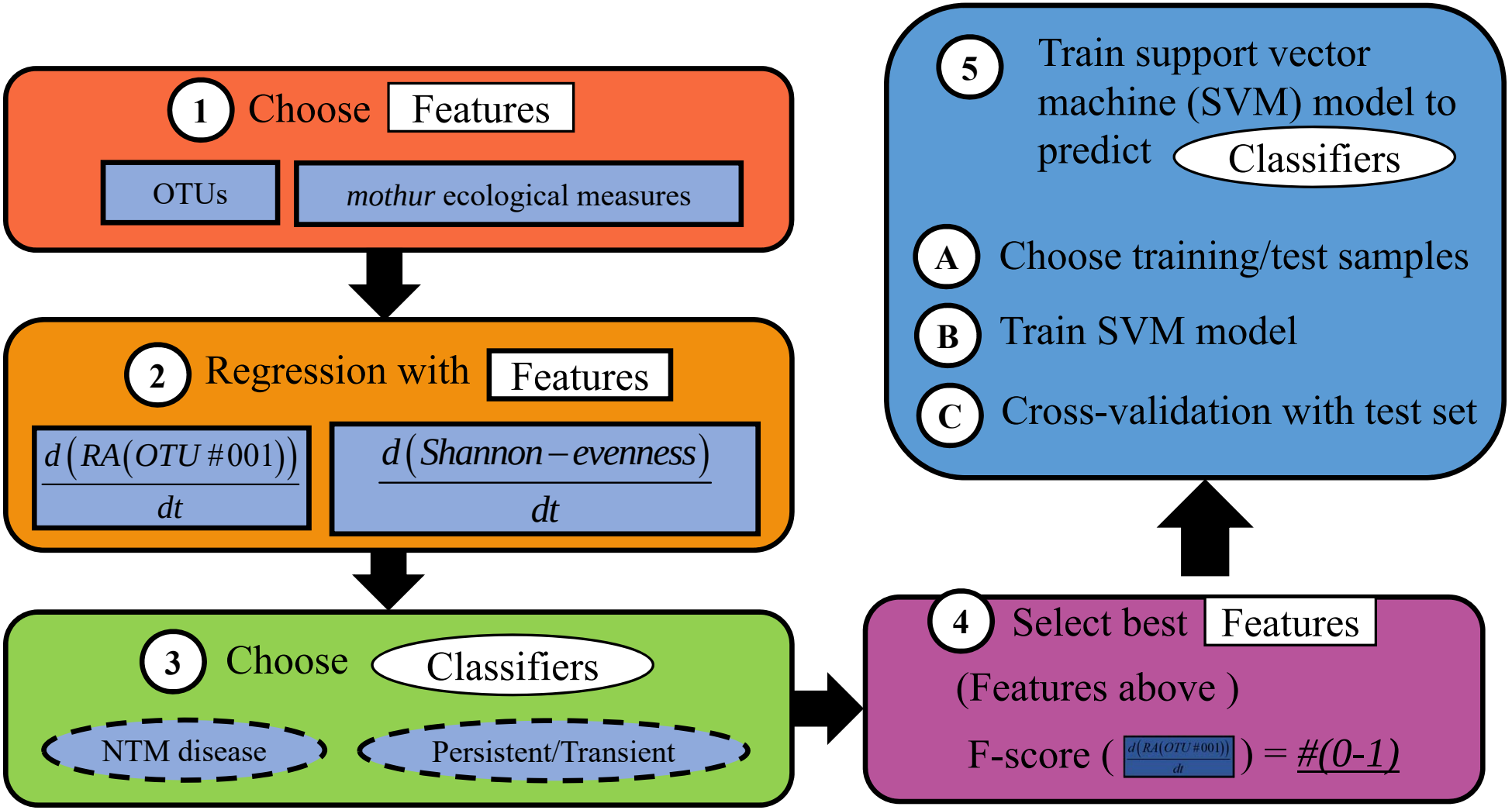
# Support vector machine methods explained

**SVM:** F-score plus optimized feature selection SVM model.

- **Source files:** Extension of "fselect.py" (Kris' idea, from *libsvm* CRAN package)
- **Reference:** W.; Lin, C.-J., "Combining SVMs with Various Feature Selection Strategies." In *Feature Extraction: Foundations and Applications*, Guyon, I.; Nikravesh, M.; Gunn, S.; Zadeh, L. A., Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, 2006; pp 315-324.**)**

**CW-SVM:** Class-weighted SVM (corrects for imbalanced NTM disease group sizes)

- **Source files:**


1 Choose Features
OTUs
mothur ecological measures
2 Regression with Features
$\frac{d(RA(OTU\#001))}{dt}$
$\frac{d(Shannon-evenness)}{dt}$
3 Choose Classifiers
NTM disease
Persistent/Transient
4 Select best Features
(Features above )
F-score ( $\frac{d(RA(OTU\#001))}{dt}$ ) = #(0-1)
5 Train support vector machine (SVM) model to predict Classifiers
A Choose training/test samples
B Train SVM model
C Cross-validation with test set

# How *accurately* do SVM models classify sputum samples from subjects with NTM disease?

| SVM Model | SVM | CW-SVM | SVR |
|---|---|---|---|
| **Disease-All (1)** | | | |
| **Disease-Main (2)** | | | |
| **Disease-Index (3)** | | | |
| **Persist-All (4)** | | | |

**Cross-Validation Accuracy (%)**

# Which microbiome features *distinguish* sputum from subjects with NTM disease?

| SVM | #1 | #2 | #3 | #4 | #5 |
|---|---|---|---|---|---|
| **Disease-All (1)** | | | | | |
| **Disease-Main (2)** | | | | | |
| **Disease-Index (3)** | | | | | |
| **Persist-All (4)** | | | | | |

Top 5 features (F-score)

| CW-SVM | #1 | #2 | #3 | #4 | #5 |
|---|---|---|---|---|---|
| **Disease-All (1)** | | | | | |
| **Disease-Main (2)** | | | | | |
| **Disease-Index (3)** | | | | | |
| **Persist-All (4)** | | | | | |