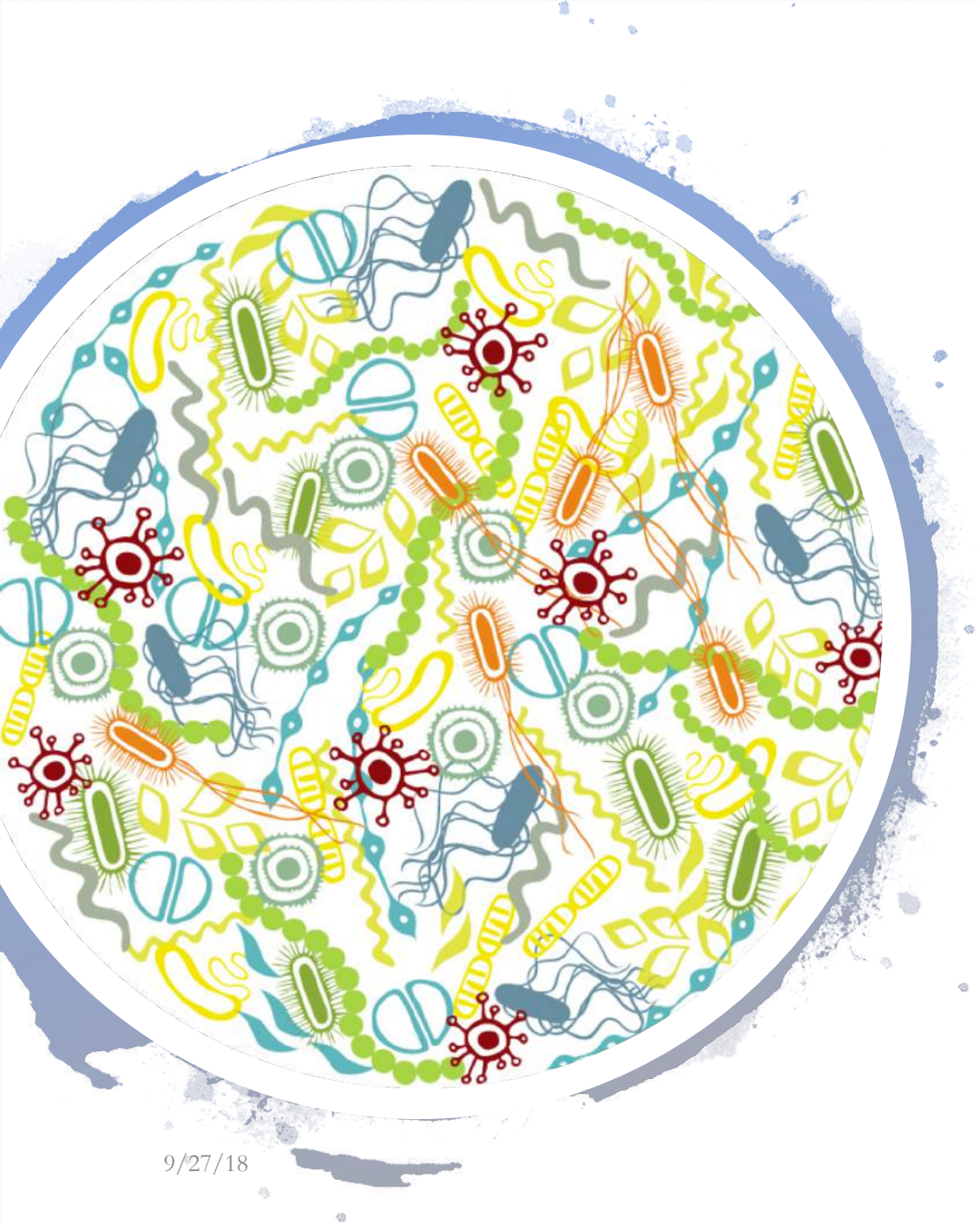
From Bioinformatics to Machine Learning

a Microbiome story

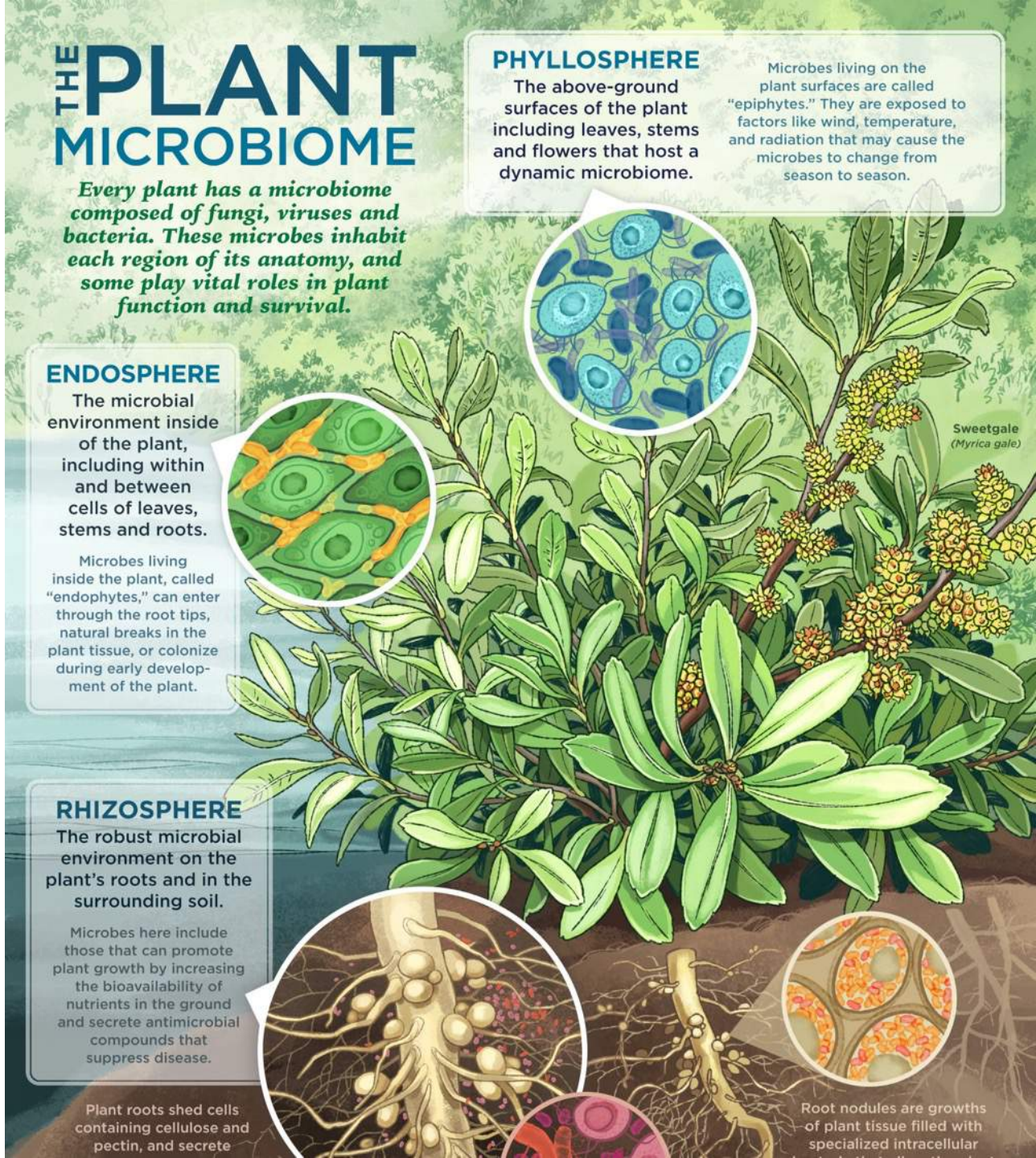


What's the microbiome

- A Microbial biome:
  - Community of microbes
  - Community of genes
- Microbiota → microbes
- Microbiome → microbial genes
- Includes: Bacteria, Archaea, Fungi, Viruses, Algae, Protozoa

Why the Microbiota…

- In Plants?
  - Linked to good things:
    - Yields
    - Protection from diseases
    - Provide nutrients
  - Linked to bad things:
    - Diseases
    - Loss of functions
    - Loss of diversity
  - Multi-organ influence!!!



How do we approach the problem?

- Disentangle complexity:
  - Who is there? (phylogenetic affiliation)
  - What does he do? (functionalities)
  - Why? (expression patterns)
  - How? (metabolic pathways)
  - When? (temporal resolution)
  - Where? (spatial resolution)
- Interactions between members
- Interactions with environment
- Interaction with host

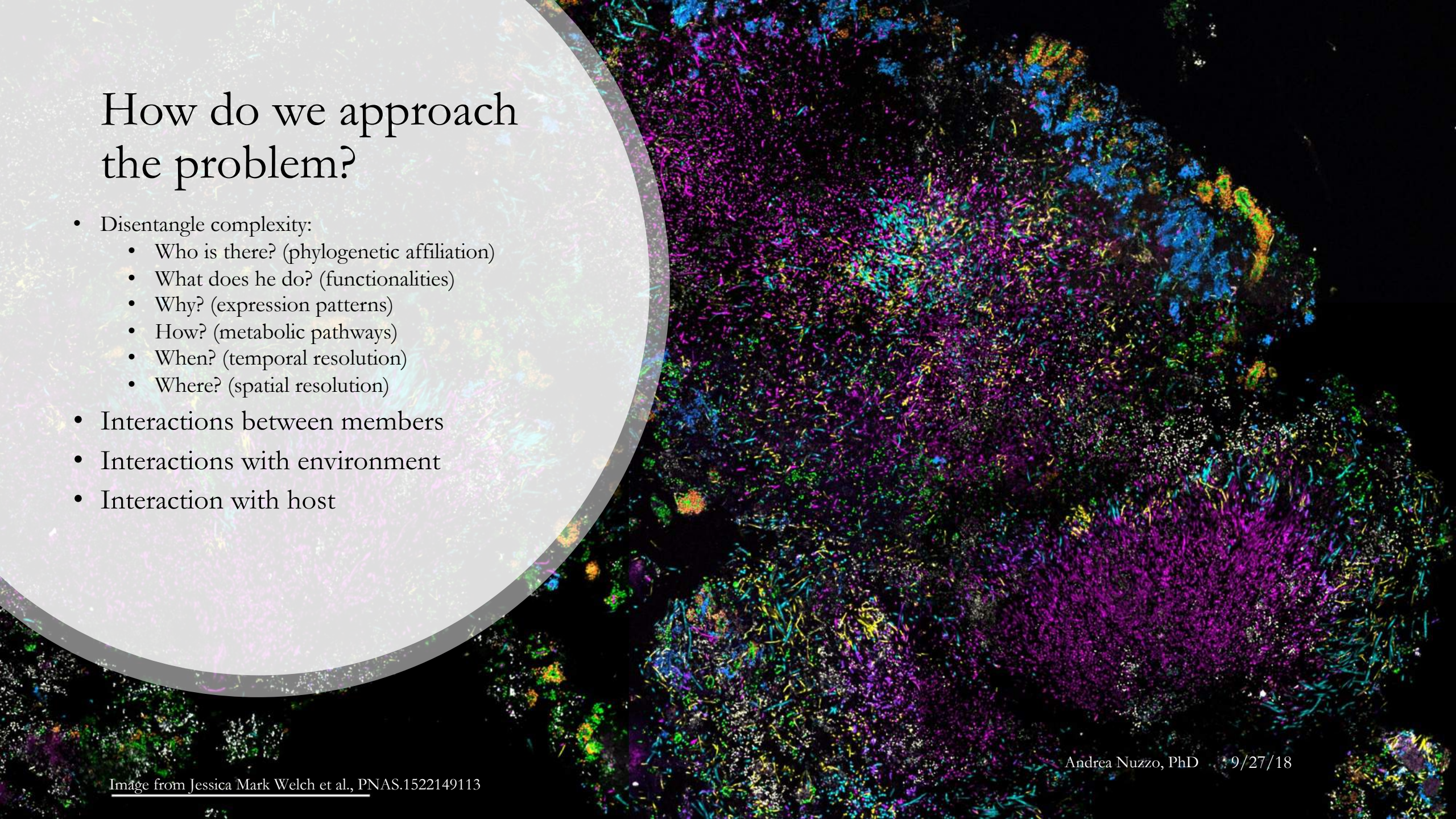
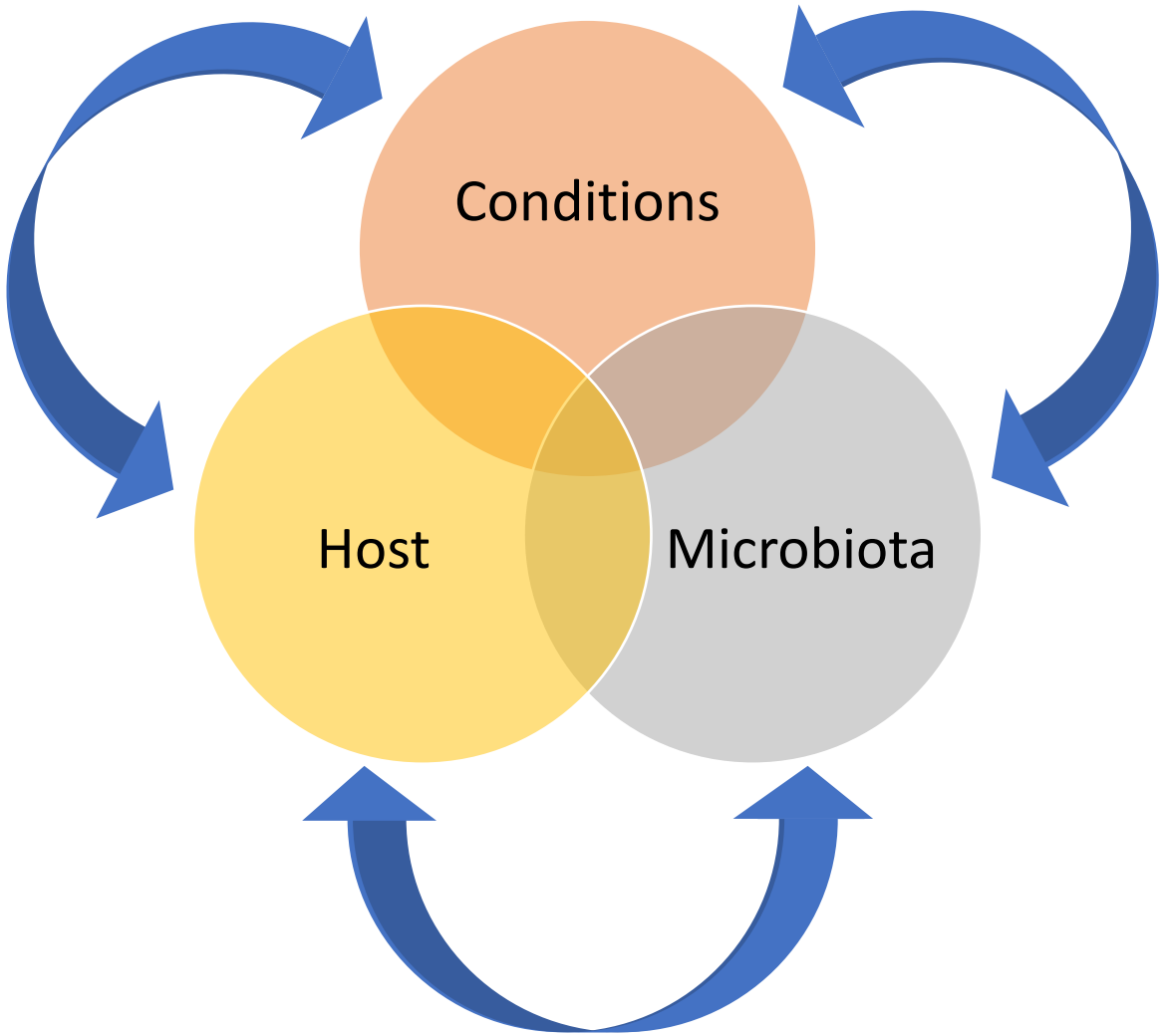


Image from Jessica Mark Welch et al., PNAS.1522149113

OUR MISSION

- Untangle complex system interactions
- Interactions are bidirectional
- There is always a three-way influence (at least)



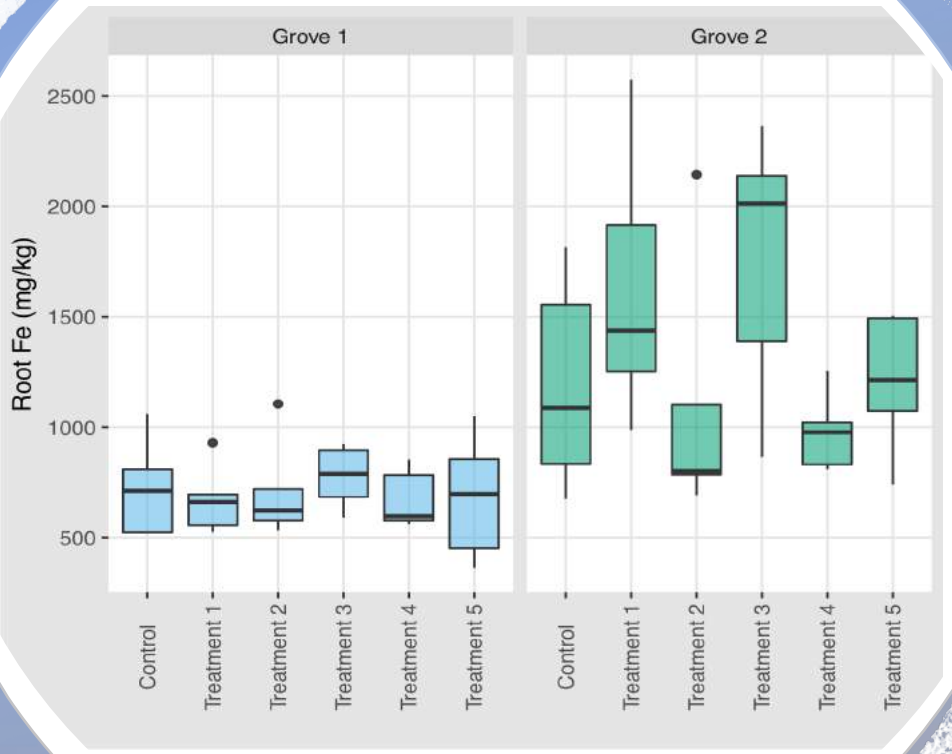
What's the matter?

UF | IFAS
UNIVERSITY of FLORIDA

Soil microbiome characterization and manipulation

- Protection from diseases
- Increase in plant yields and health
- Bio-fertilizers

Plant-microbiome interactions



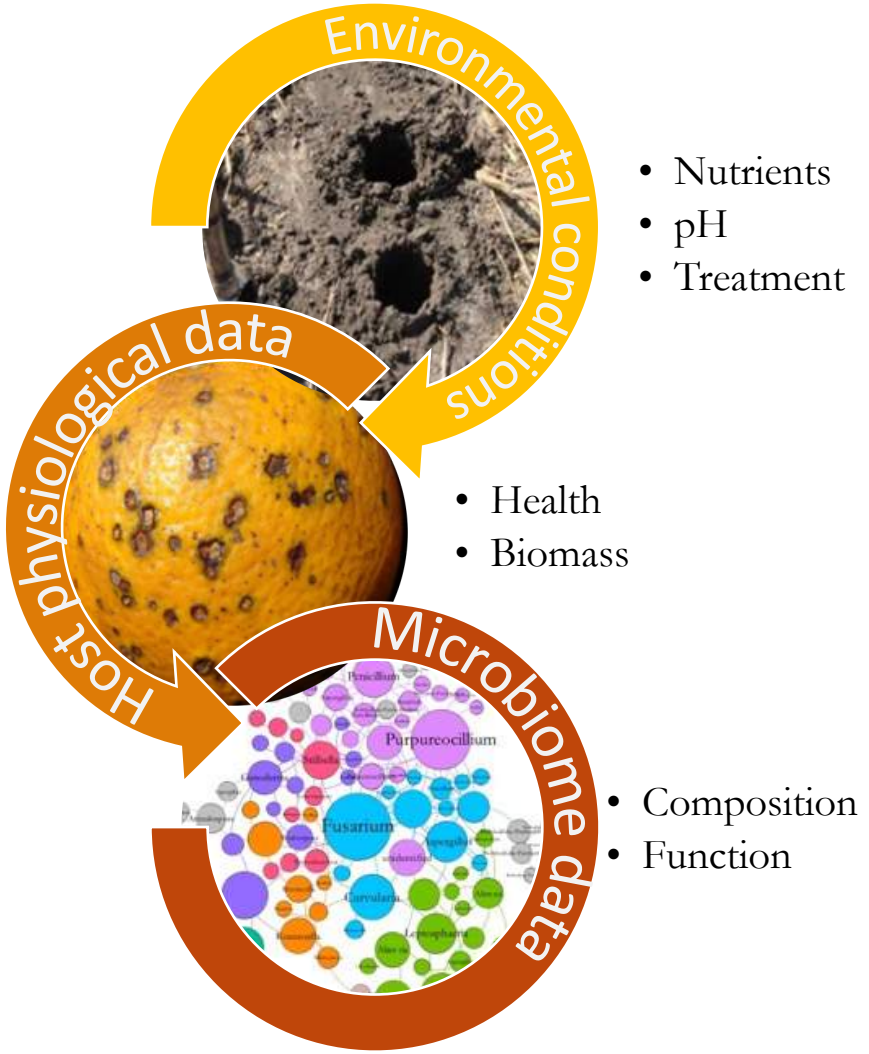
Case study: HLB

- Citrus greening (HLB) is an epidemic, lethal, incurable disease of citrus
  - Pathogen known but uncharacterized
  - Degrades roots first
- Two groves with different management strategy
  - Fertilizer, pesticides, etc.
- 4 treatments applied to try to alleviate roots
  - 2 of them include microbes
- Data show differences!

Untangling microbiome knots

- Soil data
- Plant data
- Metagenomic DNA

NGS of 16S and ITS rRNA genes



- Nutrients
- pH
- Treatment

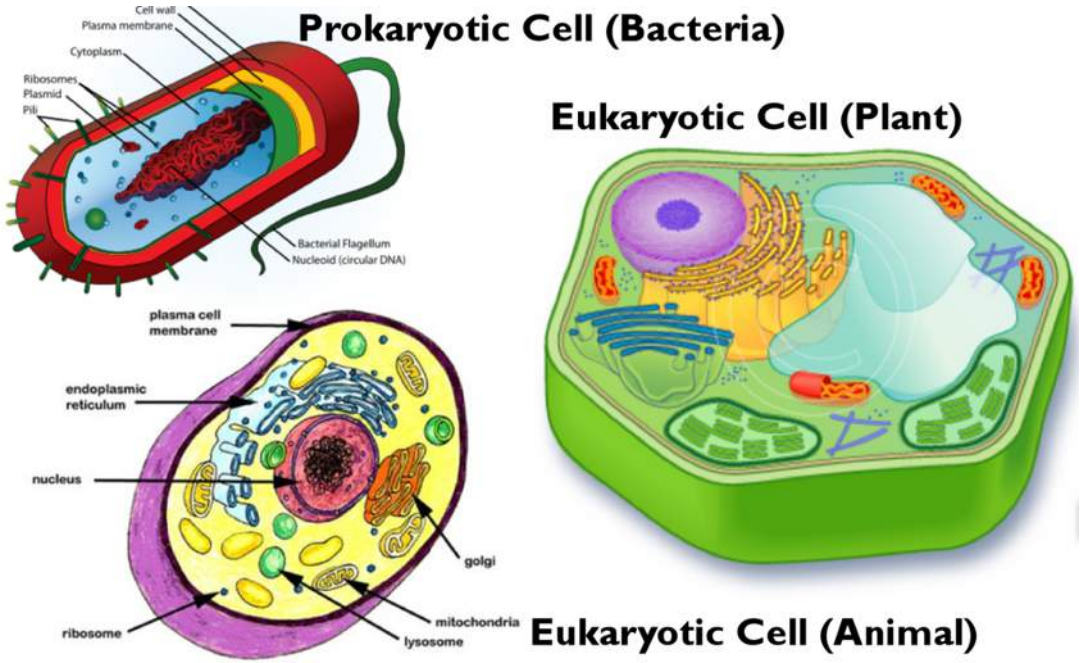
- Health
- Biomass

- Composition
- Function

Host-microbiome interactions

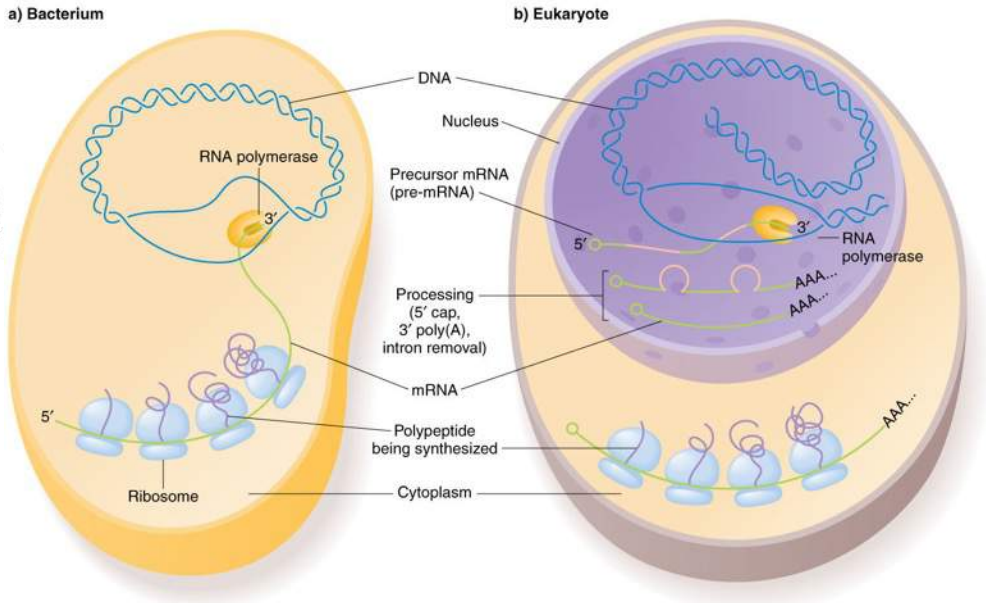
Quick recap of biology basics

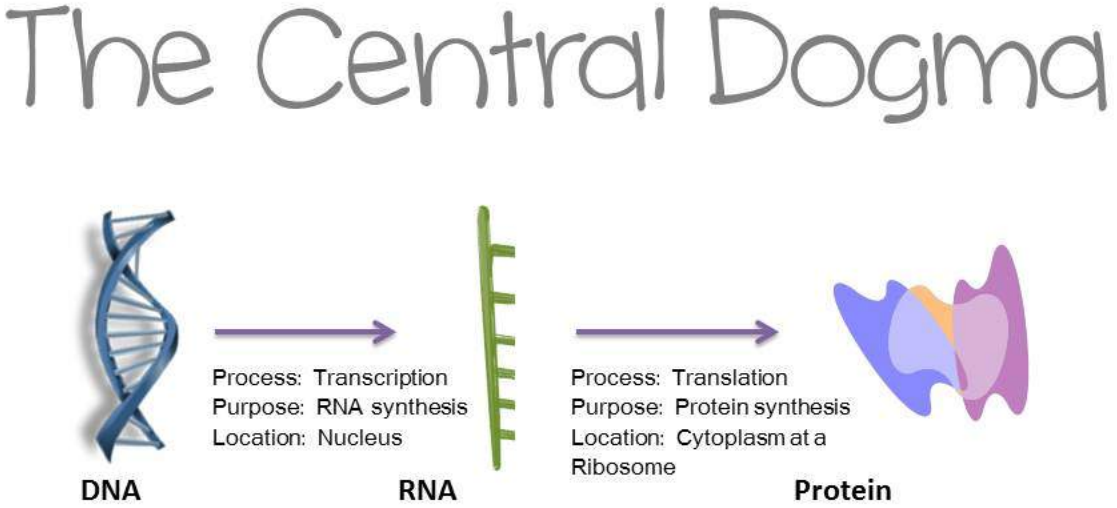
- Bacteria (Prokaryotes)
  - No nucleus, DNA is free inside cell space
- Plant, Fungi, Animals (Eukaryotes)
  - Several cell compartments including nucleus that contains DNA
- Archaea
  - cell compartments but no nucleus
- Viruses?
  - oh well… it's complicated



Quick recap of molecular biology basics

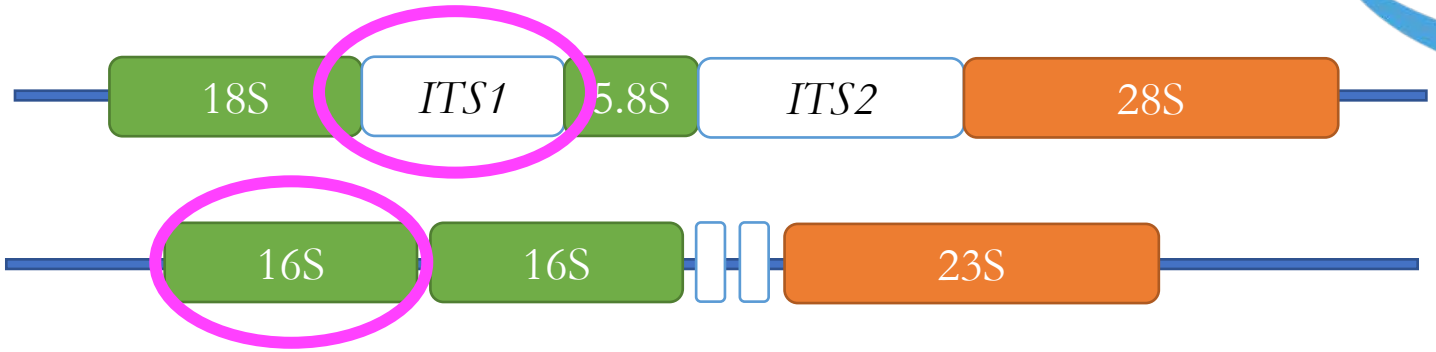
- Central dogma:
  - DNA is the code of life
    - Double-strand helix (forward+reverse)
    - Sequences of 4 nitrogen bases (ATCG)
  - GENES are transcribed into mRNA
    - processed (only in Eukaryotes)
  - mRNA is translated into proteins by the ribosomes
    - Sequence of aminoacids that has several functions (metabolism, structure, etc.)

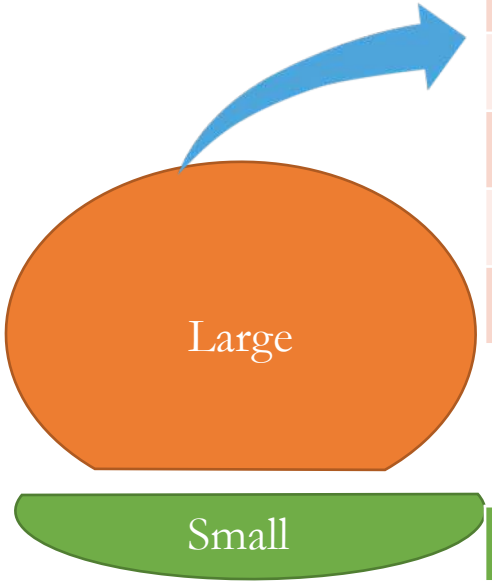




Ribosomial genes

- Ribosomes are the basis of (mostly) all life
  - Made of rRNA and proteins
  - rRNA genes are EXTREMELY conserved
    - Can be used to ”name” life
  - Small unit and large unit



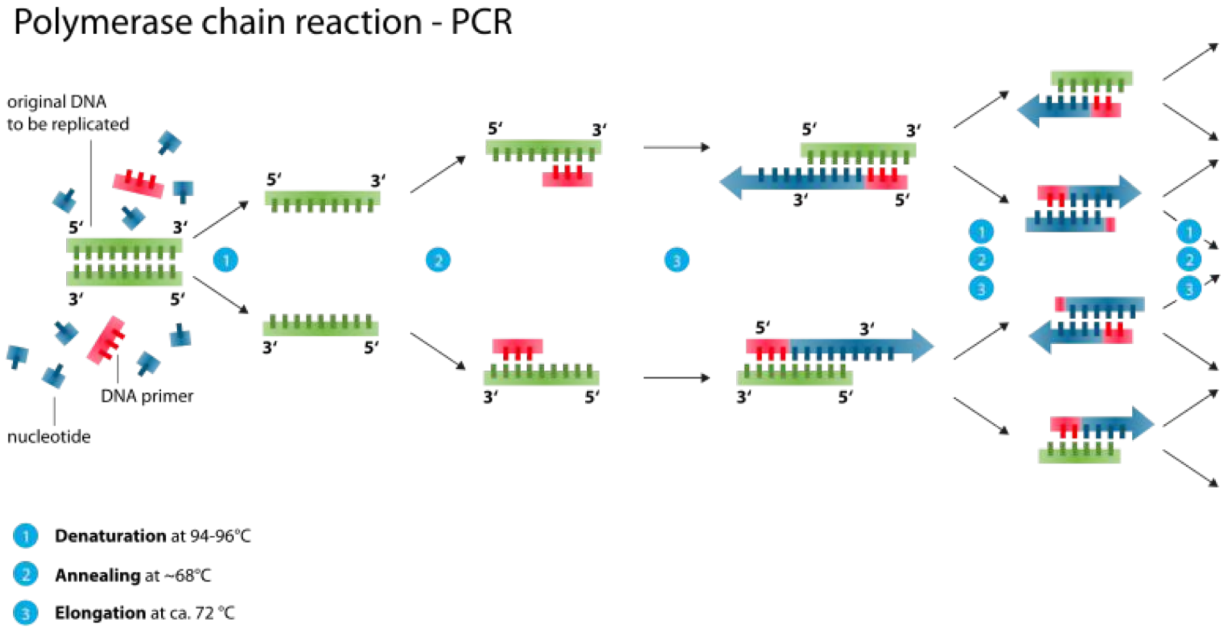


Eukaryotes	Prokaryotes
60S	**50S**
5.8S rRNA	5S rRNA
5S rRNA	23S rRNA
28S rRNA	-
49 proteins	34 proteins

Eukaryotes	Prokaryotes
40S	**30S**
18S rRNA	16S rRNA
33 proteins	21 proteins

Polymerase Chain Reaction (PCR)

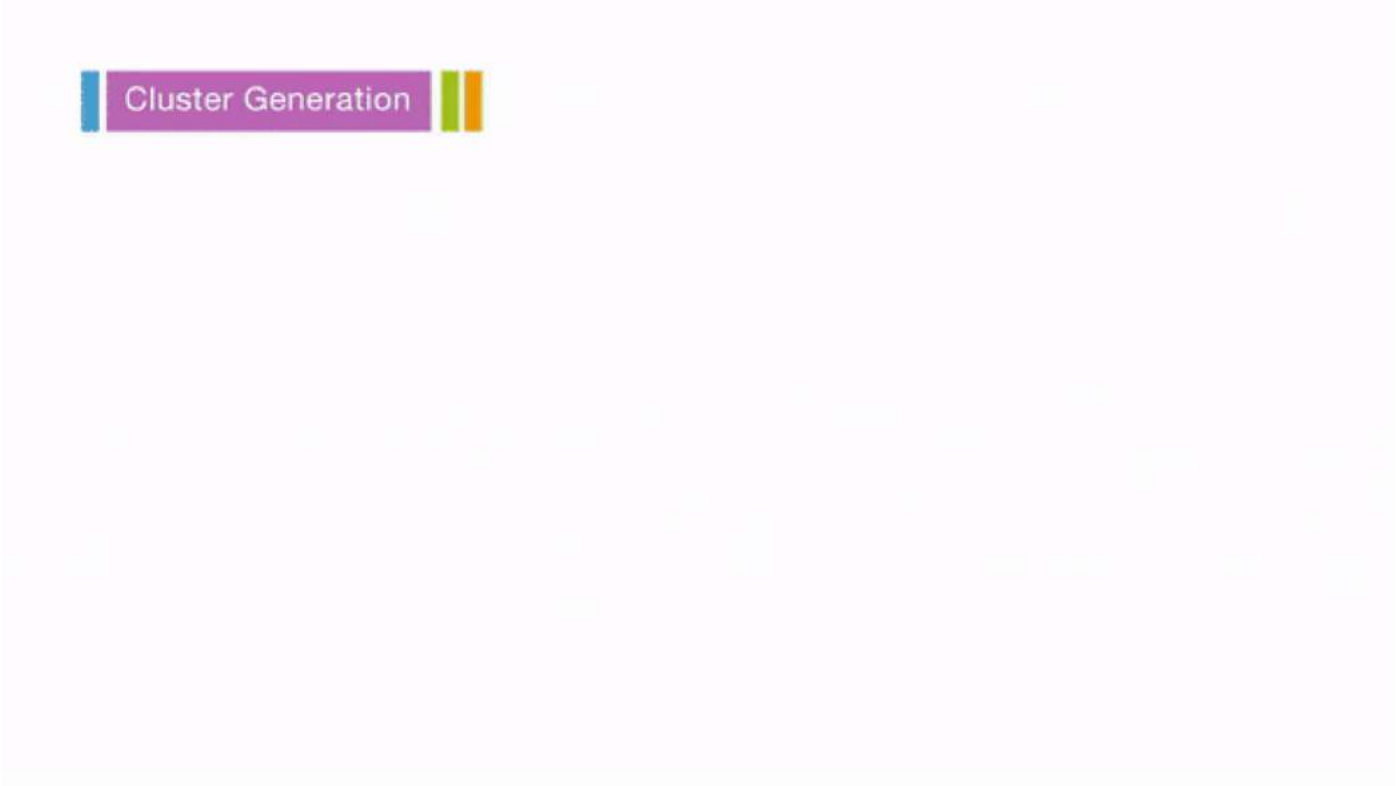
- Replicates multiple copies of a DNA fragment
  - Requires primers, enzymes, nucleotides
  - We chose the primers which are our keys to select which region to amplify
  - Multiple cycles
  - After the first cycle we get 2^n copies of the *SELECTED* fragment every n cycle
- *It allows to increase the signal of our selected DNA fragment*
  - *mistakes included.*



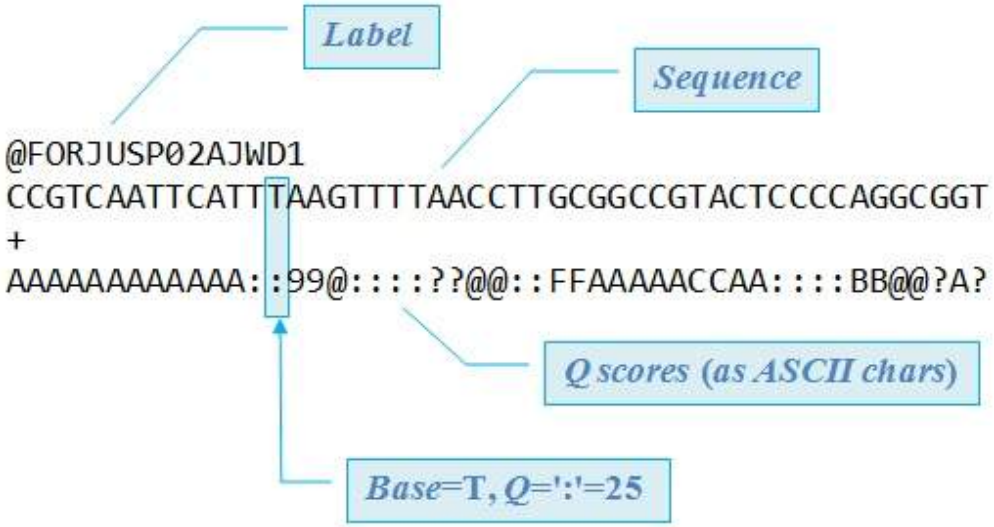
Next Generation Sequencing (NGS)

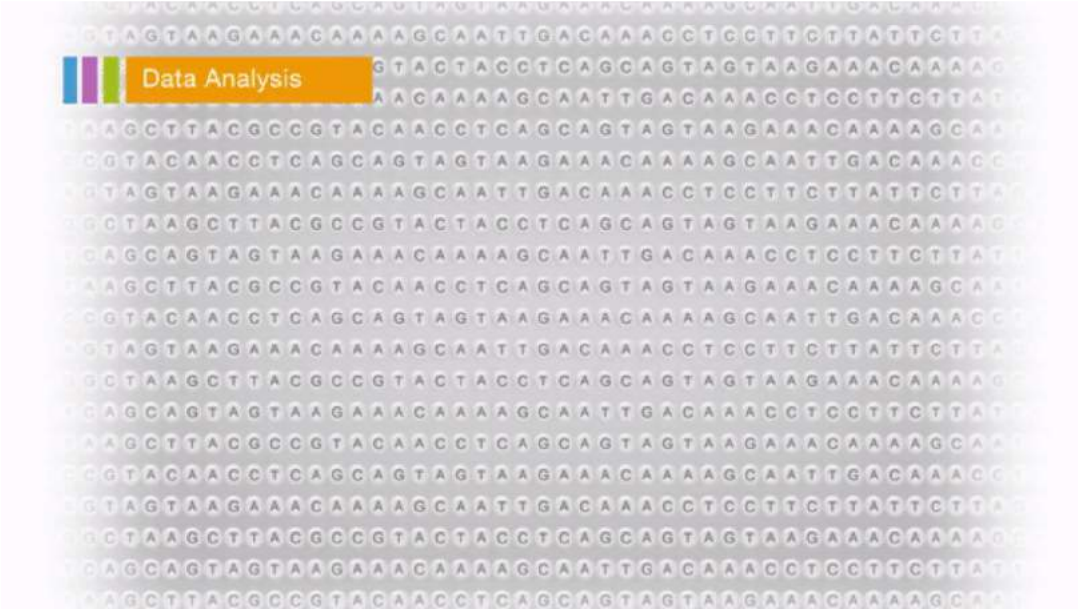
Next Generation Sequencing (NGS)

Next Generation Sequencing (NGS)



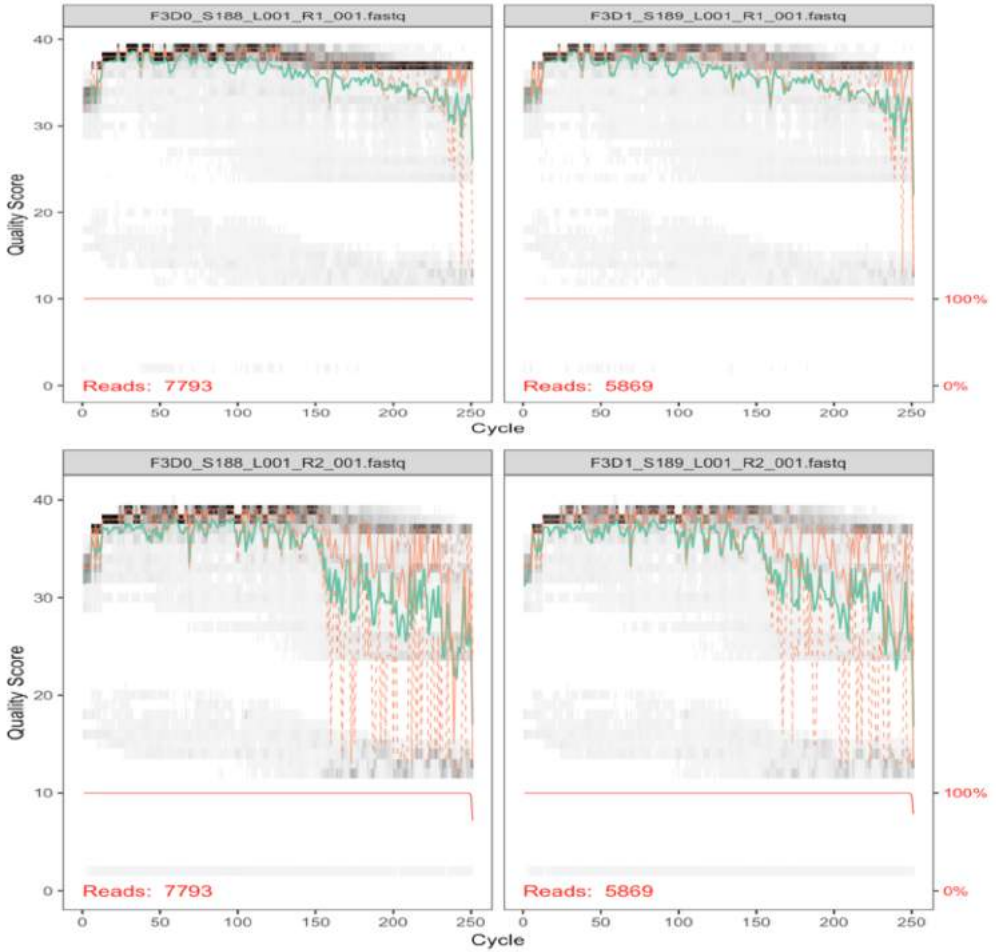
FASTQ files: from molecular biology to (bio)informatics





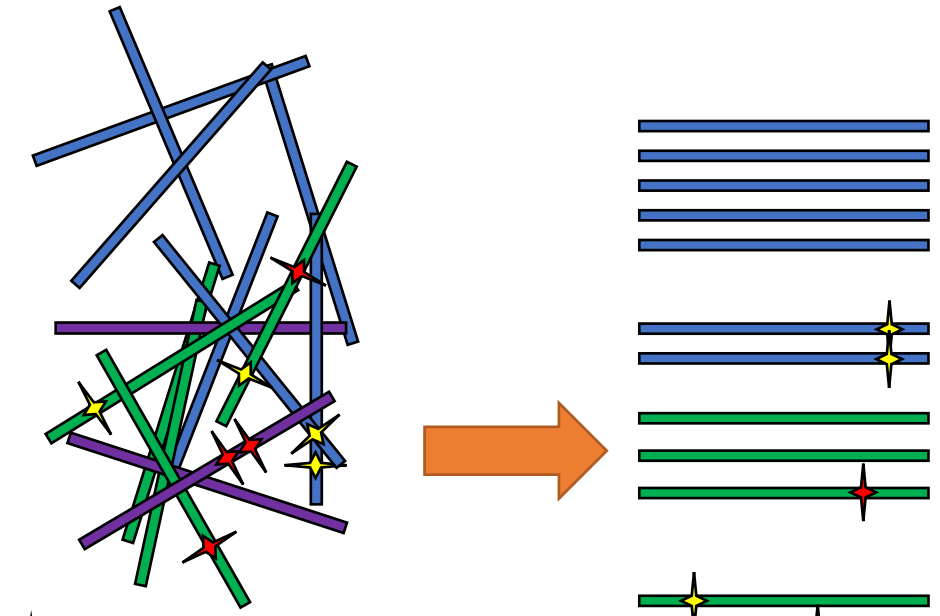
First challenge: how to get good quality paired sequences?

- Forward reads generally better than reverse reads
- How do I know that the base in a certain position is good or a misreading?
- How can I avoid PCR mistakes?
- How can I avoid merging amplicons from different sequences just because they share a common region? (chimeras)



Generating amplicon sequence variants with DADA2*

- Detect error rates
- Inferring a parametrized model of substitution
- Discriminate biological mutations from PCR errors
- Merge overlapping regions
- Infer chimeras by eliminating sequences with more than one other sequence in common



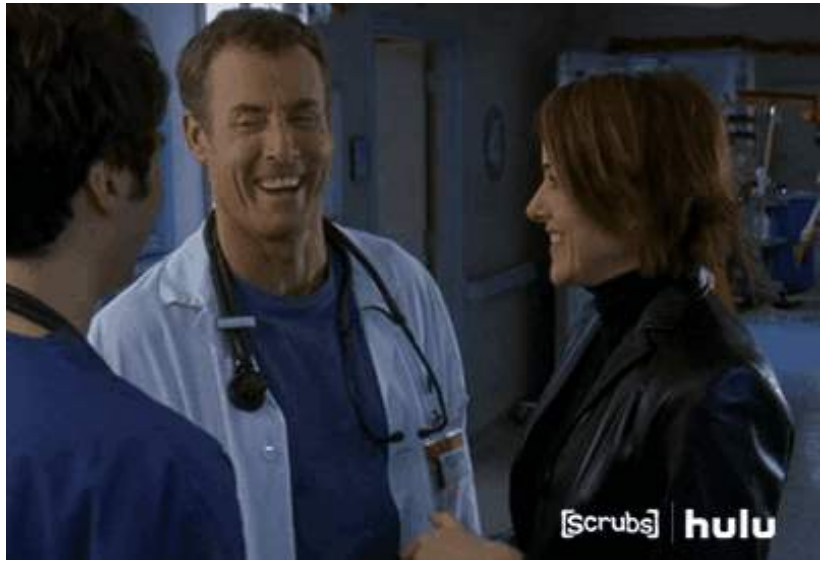
Actual biological mutation (i.e. high frequency)

PCR error (i.e. low frequency)

$$p_A(j \to i) = \frac{1}{1 - \rho_{pois}(n_j \lambda_{ji}, 0)} \sum_{a=a_i}^{\infty} \rho_{pois}(n_j \lambda_{ji}, a)$$

*Callahan, B.J. et al. (2016). Nat. Methods, 13, 581

That was just the first step…

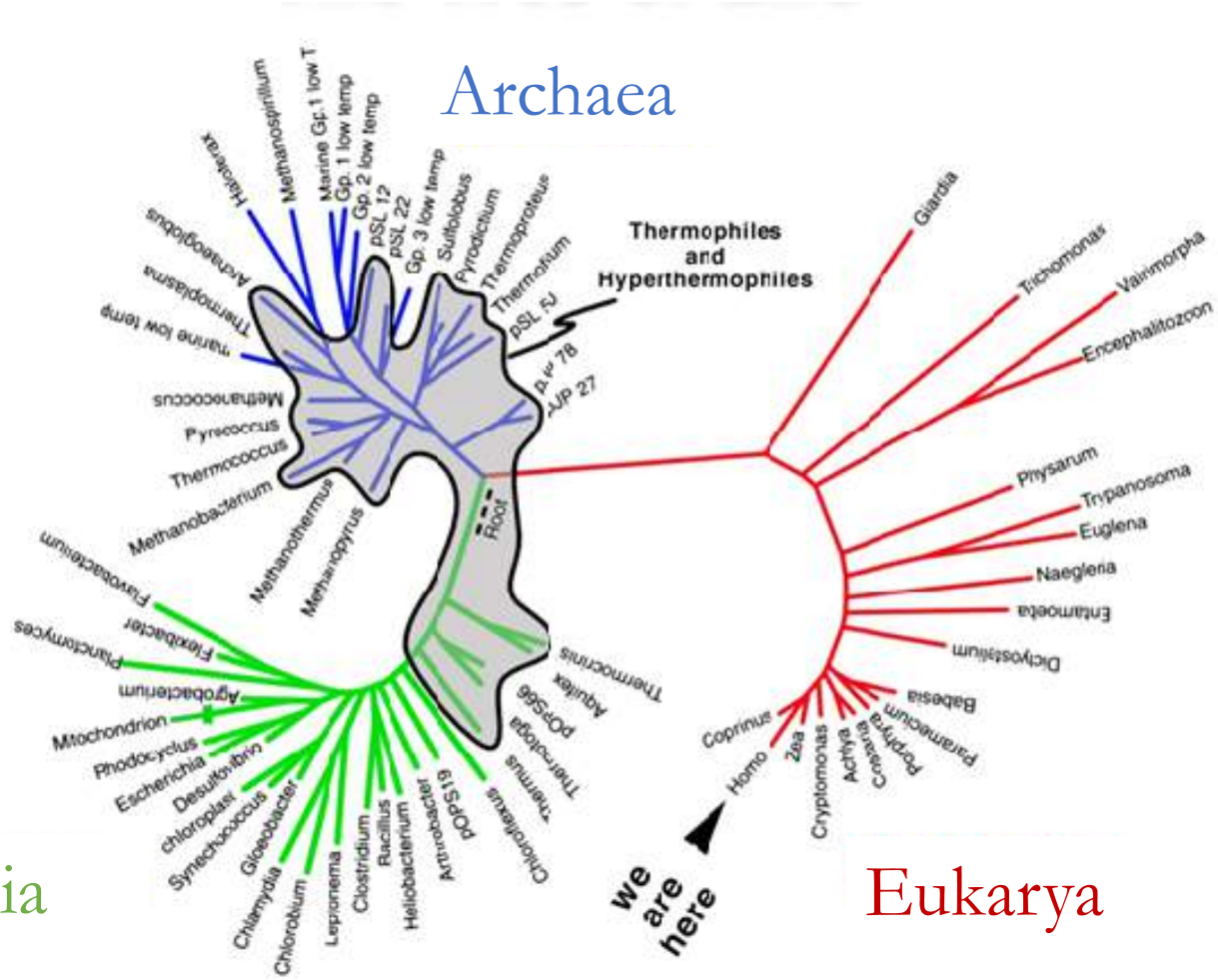


16S		ITS
	Paired-end FASTQ	
DADA2 (calling Amplicon sequence variants)	Dereplication, quality filtering	DADA2 (calling Amplicon sequence variants)
Database: SILVA 128 Naïve Bayes classifier	Phylogenetic Affiliation	Database: UNITE 7.2 Naïve Bayes classifier
PICRUSt	Metabolic predictions*	FunGuild

Diversity	Phyloseq (NMDS, PCoA, CAP, PERMANOVA, DESeq2, etc.)
Network	SPIEC-EASI
Modeling	OLS, LME, Random Forests, GBoost

Phylogenetic affiliation

- Placing organisms on the tree of life: taxonomy
  - A hierarchical clustering of life characteristics
  - Based on similarities in the rRNA sequence (debated)
    - Species = 97%
    - Genus = 95%
    - Family…
    - Order…
    - Class…
    - Phylum…
    - Kingdom…

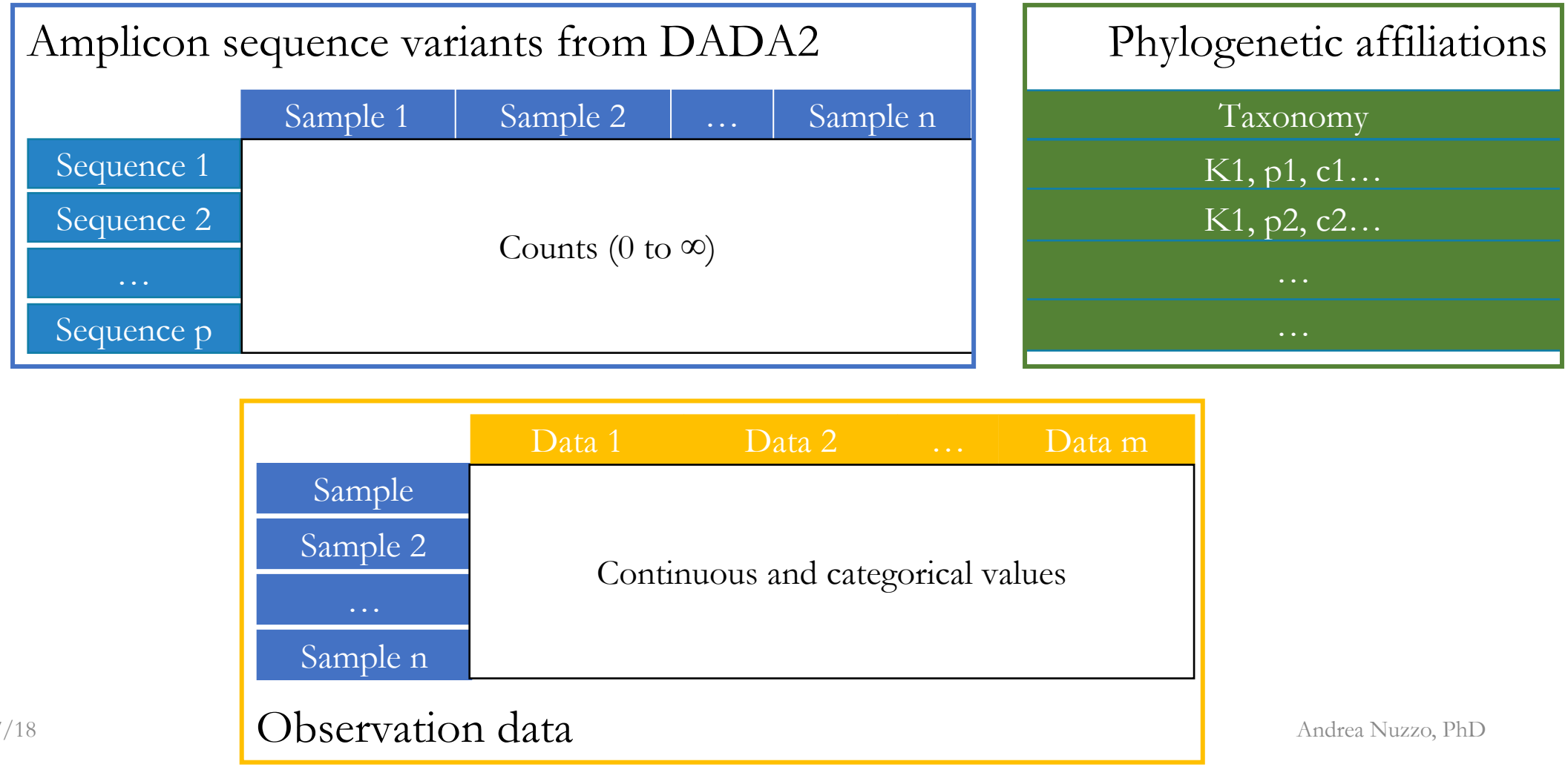


Phylogenetic affiliation

- Uses a pretrained Naïve Bayes Multinomial classifier where:
  - Sequence of ATCG are the x
  - Taxonomical levels are the multiple classes
- Predict classification against a database based on the sequence

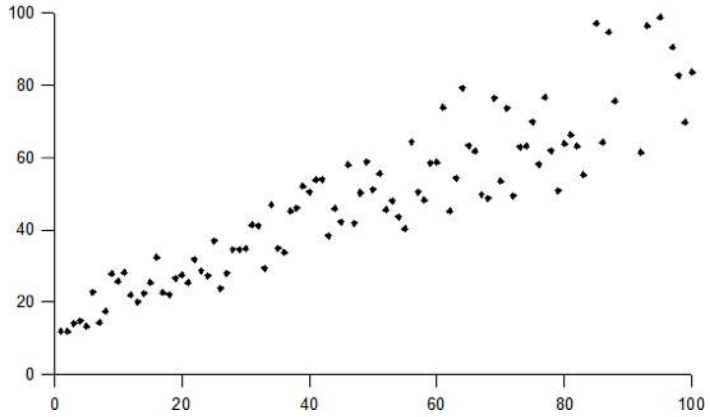
- Depends on the database on which it's trained!!!
  - Considering that ~90% of microbes in the world are estimated to be unknown…
  - And ~99% (some say 60%) of them are not culturable (so you can't easily obtain new sequences to update your database…)

OUR DATA

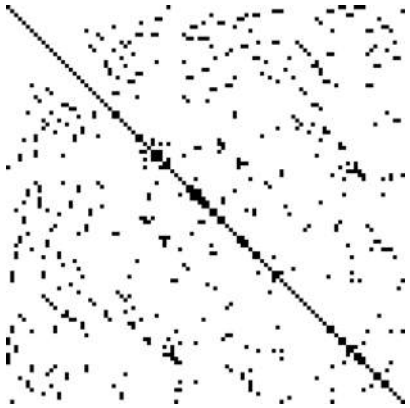

Amplicon sequence variants from DADA2
Sample 1
Sample 2
…
Sample n
Sequence 1
Sequence 2
…
Sequence p
Counts (0 to ∞)
Phylogenetic affiliations
Taxonomy
K1, p1, c1…
K1, p2, c2…
…
…
Data 1
Data 2
…
Data m
Sample
Sample 2
…
Sample n
Continuous and categorical values
Observation data

Challenges

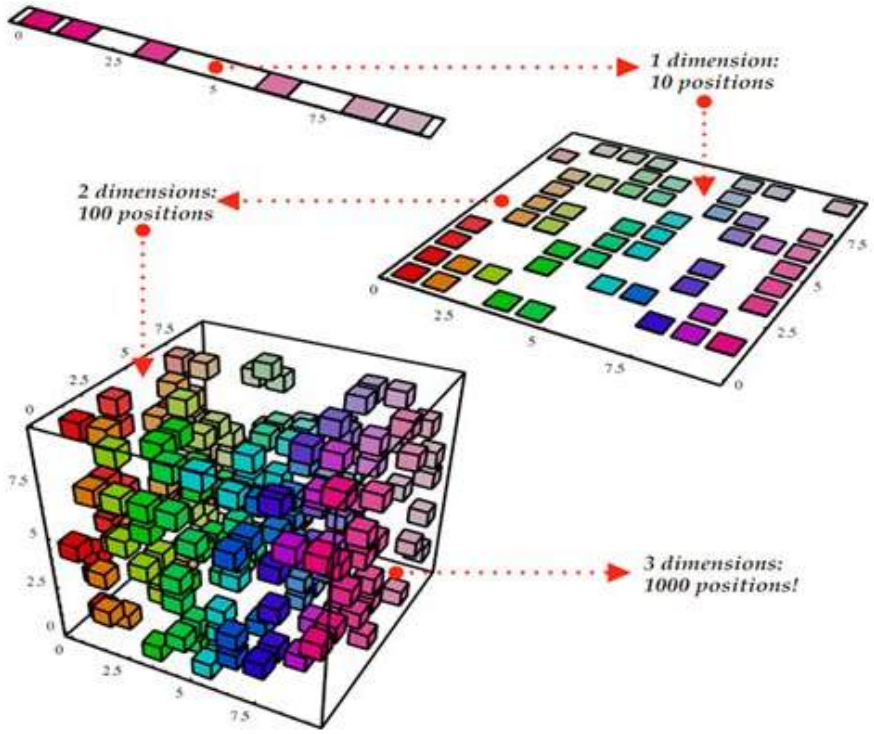
Heteroscedasticity



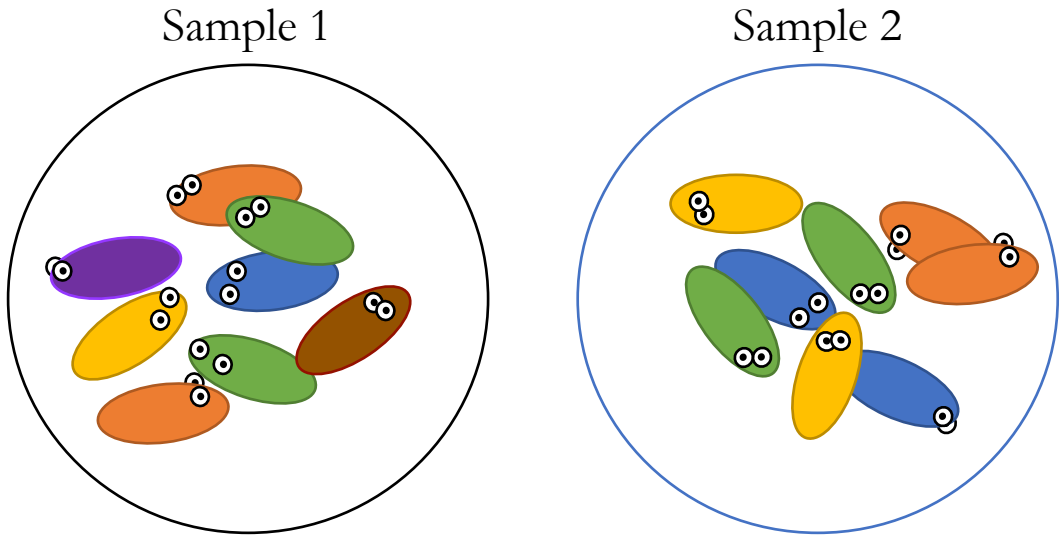
Sparsity



Dimensionality



Diversity Analysis



α_1

Alpha diversity of sample 1: 6 species with different abundances

α_2

Alpha diversity of sample 1: 4 species with similar abundance

$\beta_{1\text{-}2}$

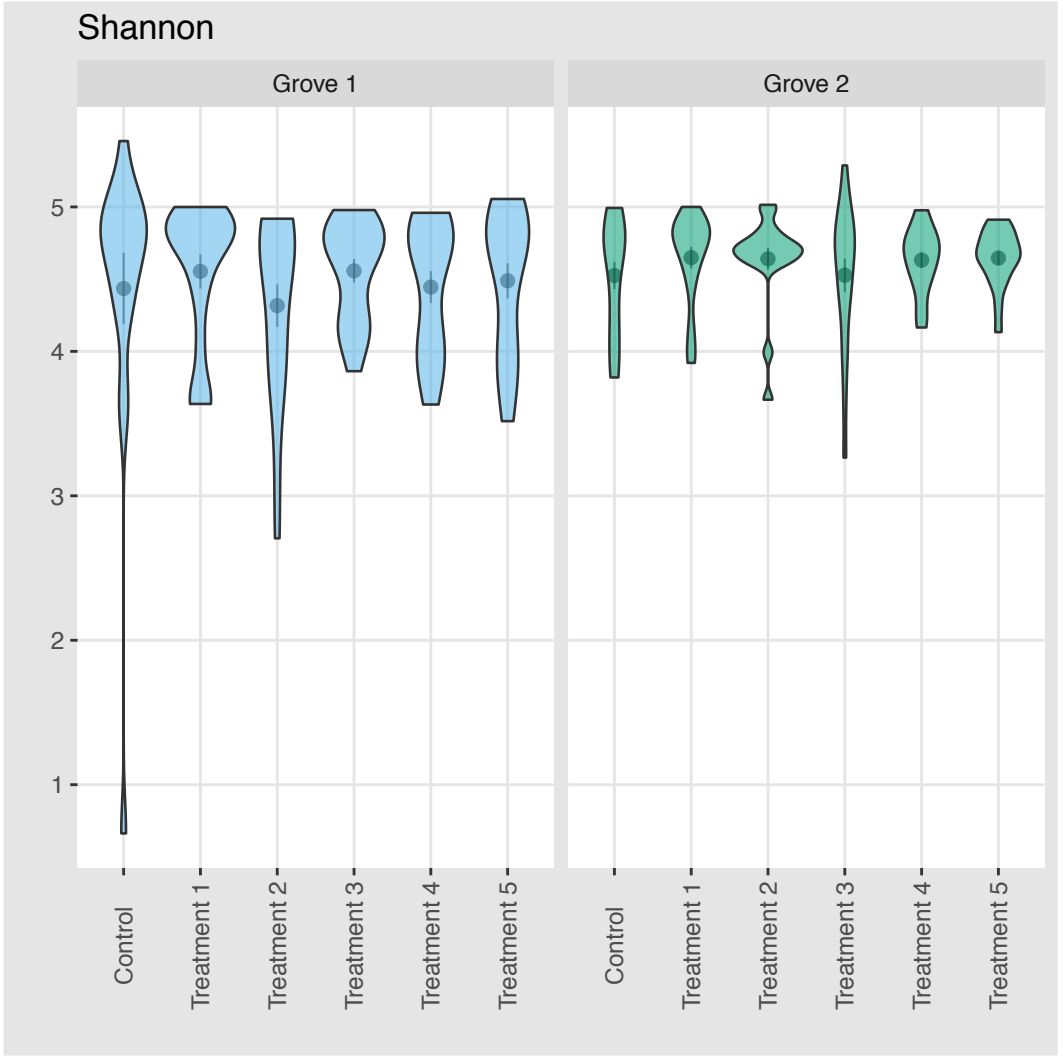
Beta diversity: differences between the two samples (i.e. distances)

ALPHA DIVERSITY

- DOES NOT NEED distribution assumptions:
  - DO NOT normalize your data
  - DO NOT filter your data
- Several indexes are possible (and often different names mean the same index), but they do different things, with different power, and measure different aspects:
  - Shannon, chao1, observed: measures *Richness*
  - Simpson: measures *Dominance*
  - Inverse Simpson: measures *Evenness*
- Rarefaction curves are a good proxy for data quality check!
  - Split the dataset into bins and plot the cumulative sums of the selected α diversity index
  - Look for plateau

Does it help?

```
library(phyloseq)
library(ggplot2)
library(tidyverse)

biomfile < file.path(”dada2_output.biom")
treefile <- file.path("tree.nwk")
mapfile <- file.path(”plant_data.txt")

#Import into Phyloseq
tree <- read_tree(treefile)
table <- import_biom(BIOMfilename = biomfile,
          parseFunction = parse_taxonomy_default,
          parallel = T)
metadata <- import_qiime_sample_data(mapfile)

#Create Phyloseq object
phylo <- merge_phyloseq(table, metadata)

#Plot alpha diversity
plot_richness(phylo, x = "Treatment", measure=“Shannon”,
          color = ”Management") + geom_violin()
```



Beta diversity

- Due to the nature of the data, making comparisons between samples is pretty difficult
- To compare samples diversities, you NEED to make them comparable, which means:
  - Normalize (i.e. fitting counts in a Gaussian curve within same range for all samples)
  - Reduce number of outliers
  - Etc.
- Introducing DISTANCES:
  - Indexes representing "dissimilarity" between samples
  - Make lots of assumptions
- Several indexes existing, including Bray-Curtis (works better on log- or root-transformed data), Jaccard (intersection over union), Manhattan (Euclidean distances), UniFrac (includes phylogeny)
- Overall, when plotting beta diversity, samples close together are more similar with each other

Multivariate analyses

- You want to find out how the WHOLE microbiome is related to your conditions or plant data
  - Possibly, which part of the microbiome is the most important
- AVOID SUBSAMPLING (rarefaction)!
  - To make samples comparable it's better to log-transform

PCoA

- Collapses β-diversity distances in 2 axes
- Useful to visualize dissimilarity matrix

NMDS

- More powerful than PCoA
- Shows how your samples can be clustered together
- NMDS axes are NOT your variables
- PERMANOVA and ANOSIM to check group differences

CCA

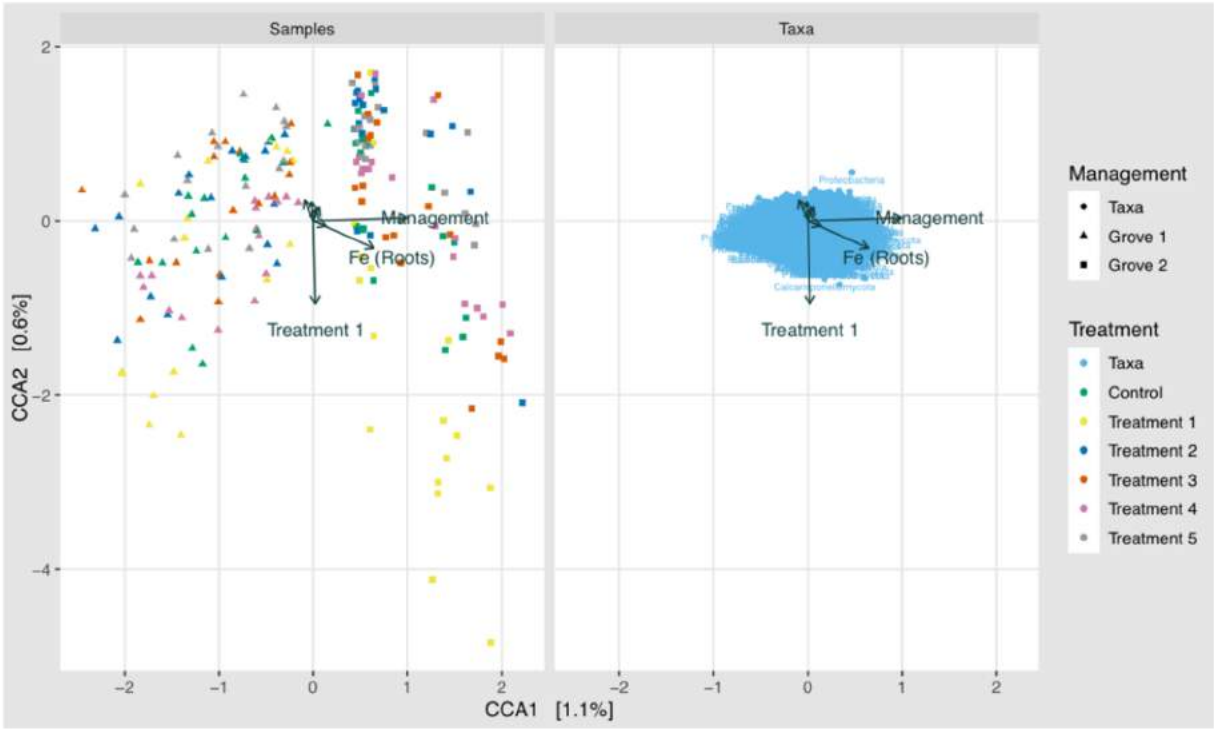
- Correlate counts to your variables
- Sensitive to multicollinearity
- ANOVA to check group differences

CAP

- Forces a PCA to explain variance of your samples with your variables
- Sensitive to multicollinearity
- ANOVA to check group differences

Does it help?

```
source('vif.cca.bw_sel.R')

pslog <- transform_sample_counts(phylo,function(x){log(1 + x)})

cca_vif <- vif.cca.bw_sel(pslog, vifvariables,
                          threshold = 5)

cca_plot <- plot_ordination(physeq = pslog,
                        ordination = cca_vif,
                        type= 'split', color = "Treatment",
                        label = 'Phylum' ) +
            aes(shape = TimePoint) +
            geom_point(aes(colour = Treatment))

cca_arrowmat <- scores(cca_vif, display = "bp")
cca_arrowdf <- data.frame(labels = rownames(cca_arrowmat),
                        cca_arrowmat)

cca_arrow_map <- aes(xend = CCA1, yend = CCA2,
                     x = 0, y = 0, color = NULL, shape = NULL)

cca_label_map <- aes(x = 1.3 * CCA1, y = 1.3 * CCA2,
                     color = NULL, label = labels, shape = NULL)

cca_arrowhead = arrow(length = unit(0.02, "npc"))

cca_plot +
geom_segment(mapping = cca_arrow_map, size = .5, data =
            cca_arrowdf, color = "black", arrow = cca_arrowhead) +
geom_text(mapping = cca_label_map, size = 2, data = cca_arrowdf)+
scale_color_pander()
```



Network analyses in the microbiome

- Borrows from non-bio sciences (graph theory, social networks)
- Biological networks are usually:
  - Scale-free networks (i.e. follow Poissonian distributions)
  - Small-world networks (every node is accessible with a relatively short path)
- Can be obtained via
  - Pairwise dissimilarities
  - Correlations
  - Regressions
  - Probabilistic Graphs Models (i.e. Bayes, Markov, etc.)
- Reveal hidden patterns
- Cluster detection
- Hub-species detection
  - i.e. keystone species
- Microbiome dynamics

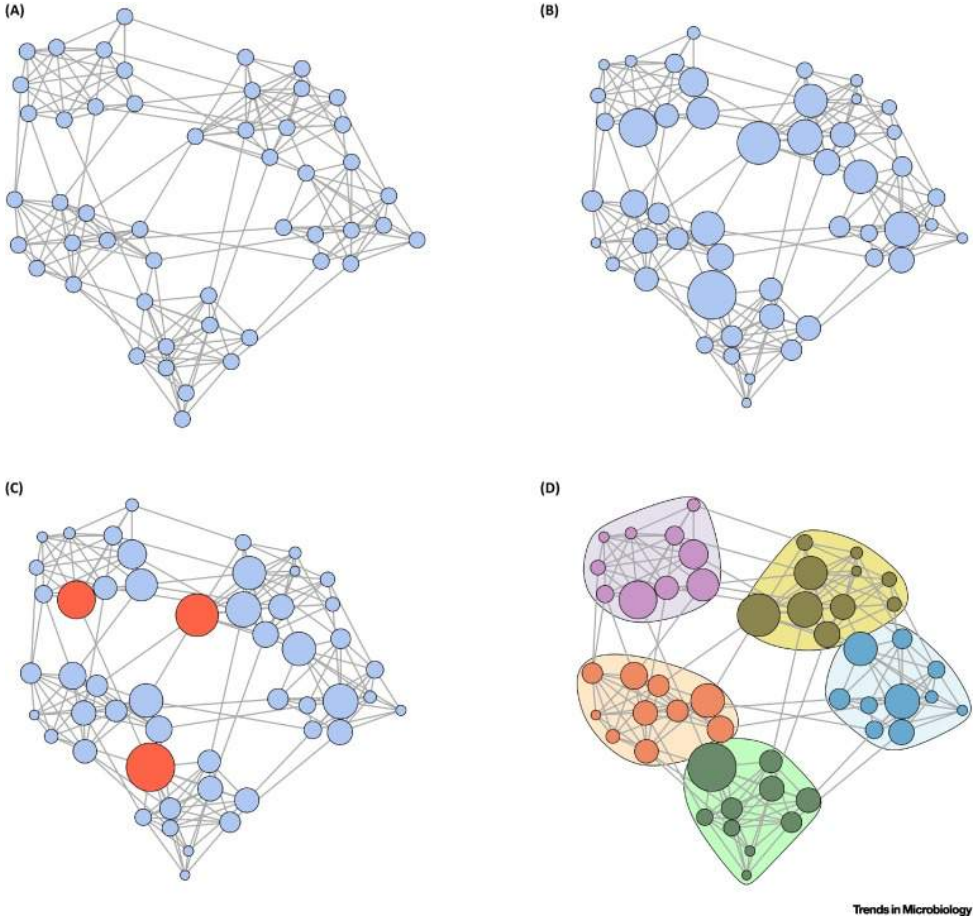
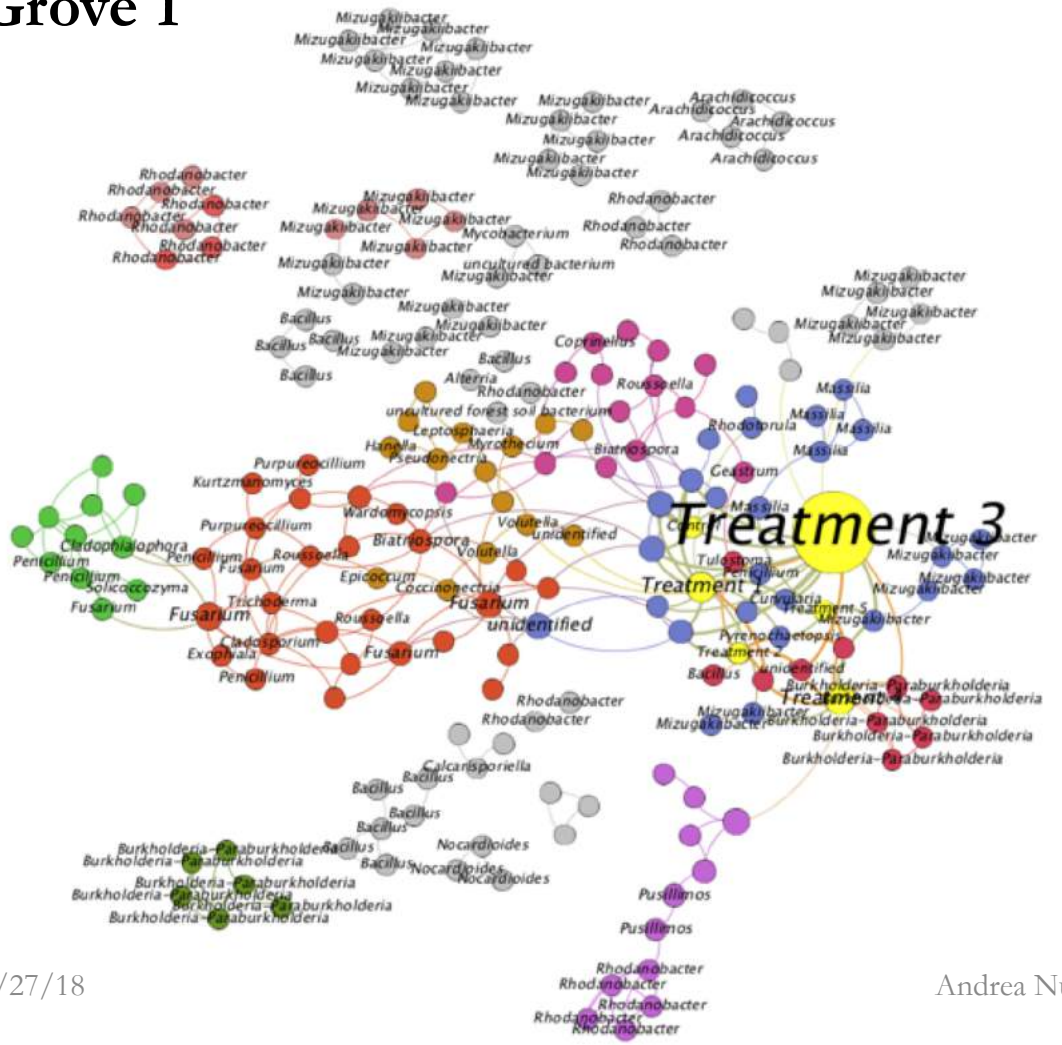


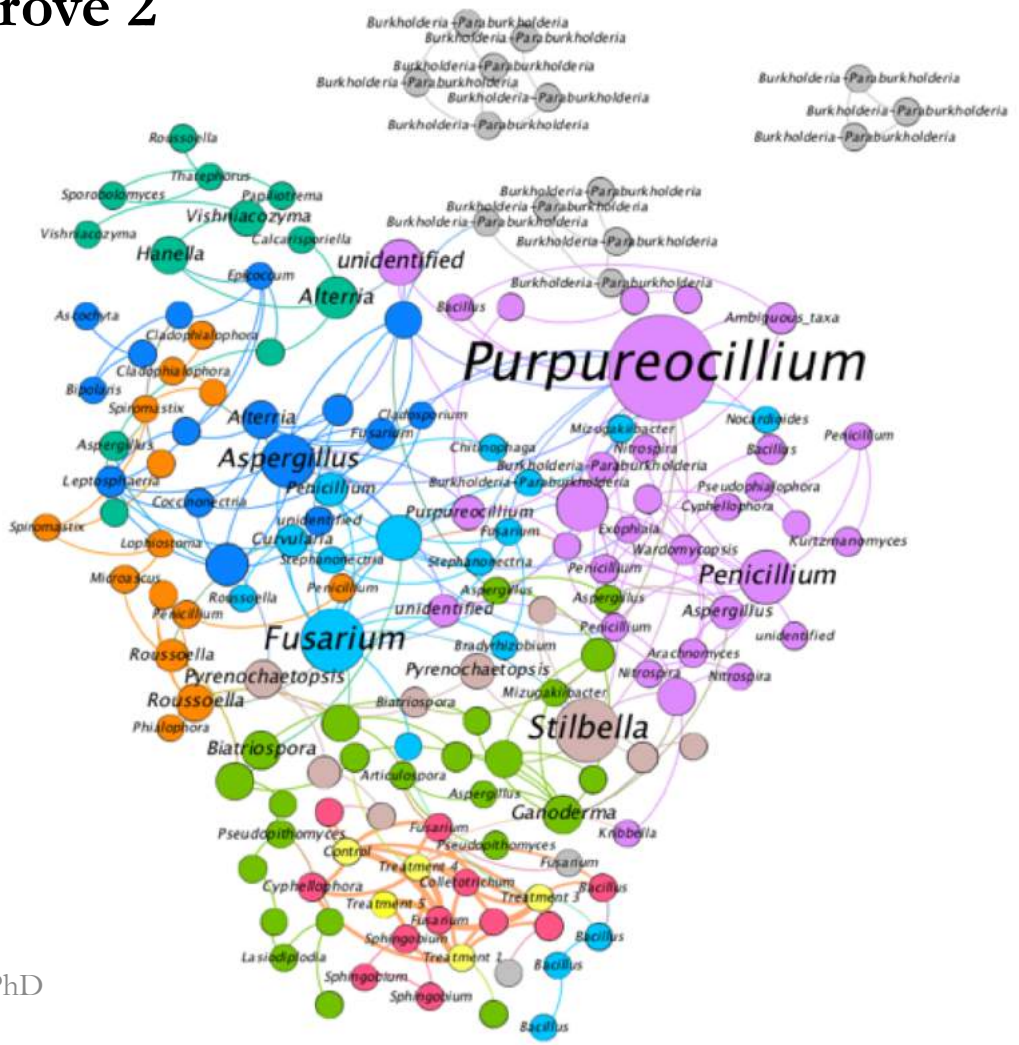
Image from Layeghifard, M. et al. (2016). Trend. Micro., 25, 3, 217-228

Does it help?

Grove 1



Grove 2

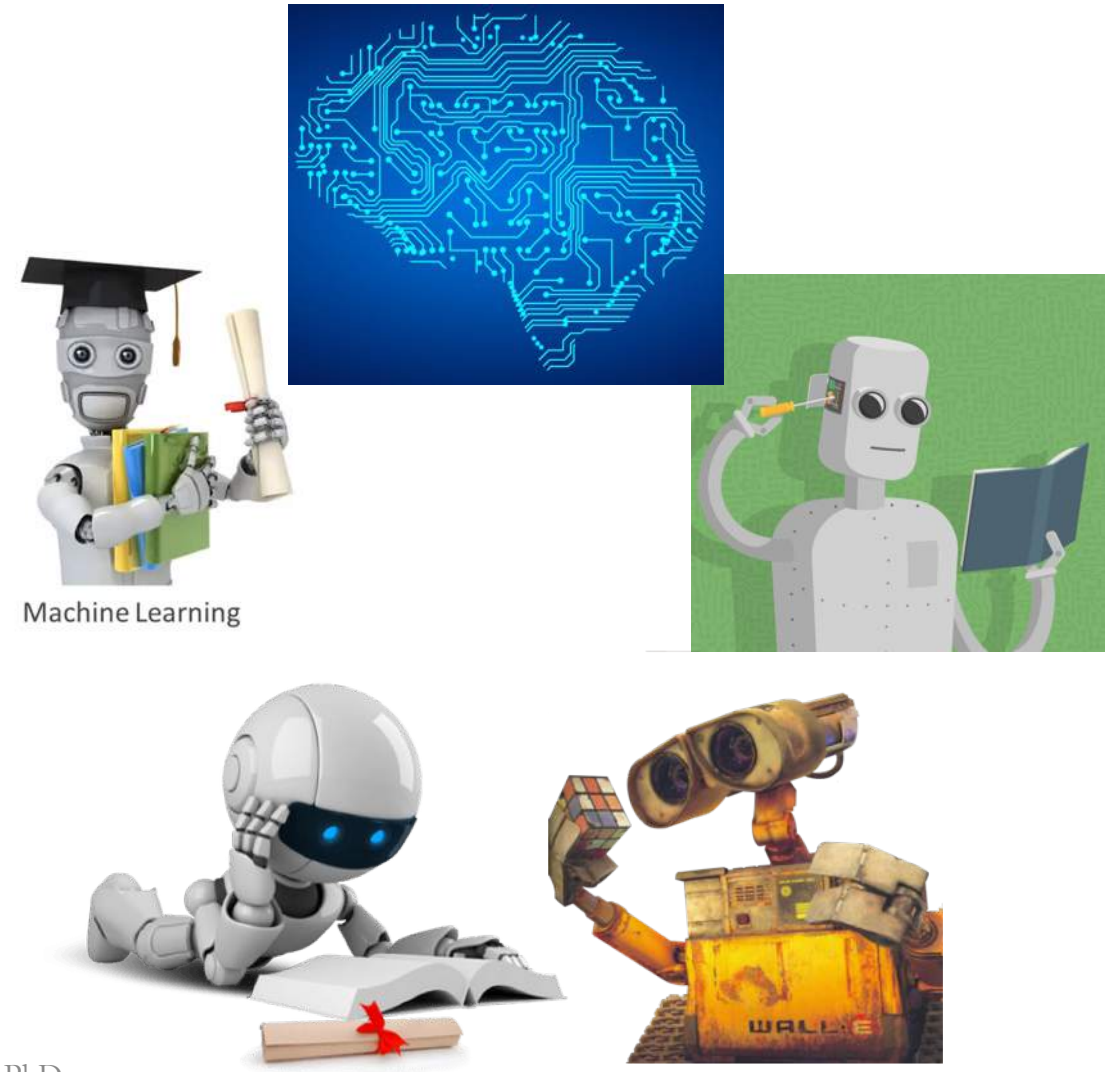


AN ALTERNATIVE METHOD

- Generalized
- Powerful
- Advanced
- Sensitive

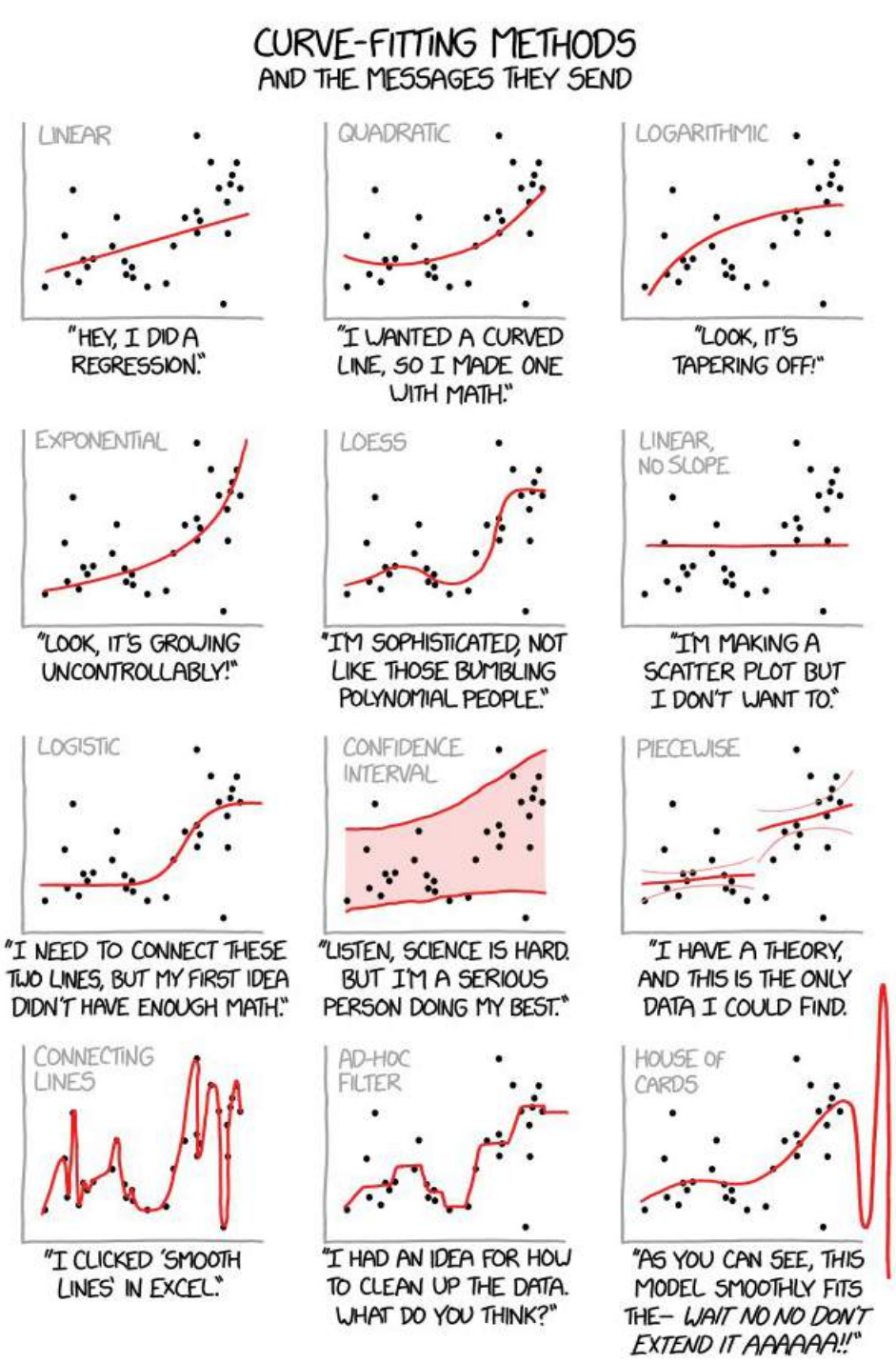
What's Machine Learning?

- An excuse for fantasy android/brain images that don't really explain anything
  - (please use Wall-E).
- Not statistics!
  - Statistics focuses on describing observations (mean, variance, etc)
  - ML focuses on predicting NEW observations BY generalizing observations into a model
  - Like all good things, this is just a simplification
- Do we care about predictions? No, but we care about understanding which are the best predictors (variables) for our model, i.e. understanding relationships in our data

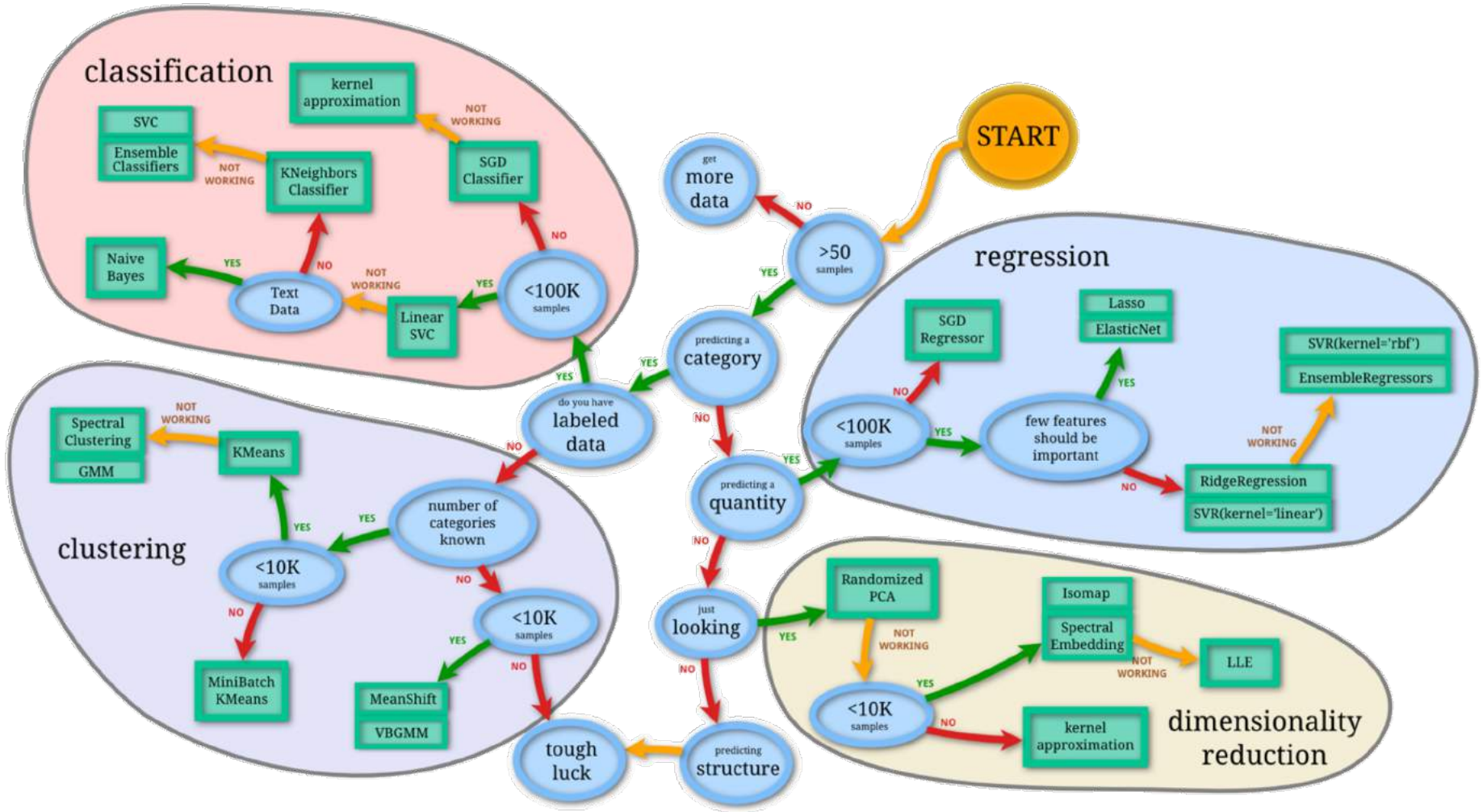


Machine learning categories

Based on the type of variables	Categorical → Classification Numerical →Regression
Based on the type of response	We know what we want → Supervised We want to get insights → Unsupervised
Based on how they optimize the model fitting:	Minimize distance between points and a line → Linear/Logistic regression Find partitions in the data→ tree-based methods, support vector machines Reduce dimensionality → PCA and similar Clustering, Bayesian, etc



HOW DO I CHOOSE?



Gneiss (Qiime2)

- Method to transform count tables
  - Applies logarithmic transformation on species ratio
  - Reduces heteroscedasticity
  - Infers relationships
  - Clusters those log-ratio using hierarchical clustering
  - Results in a weighted tree of balances
- Weighted log-ratios change between pair of species BUT NOT in the whole dataset

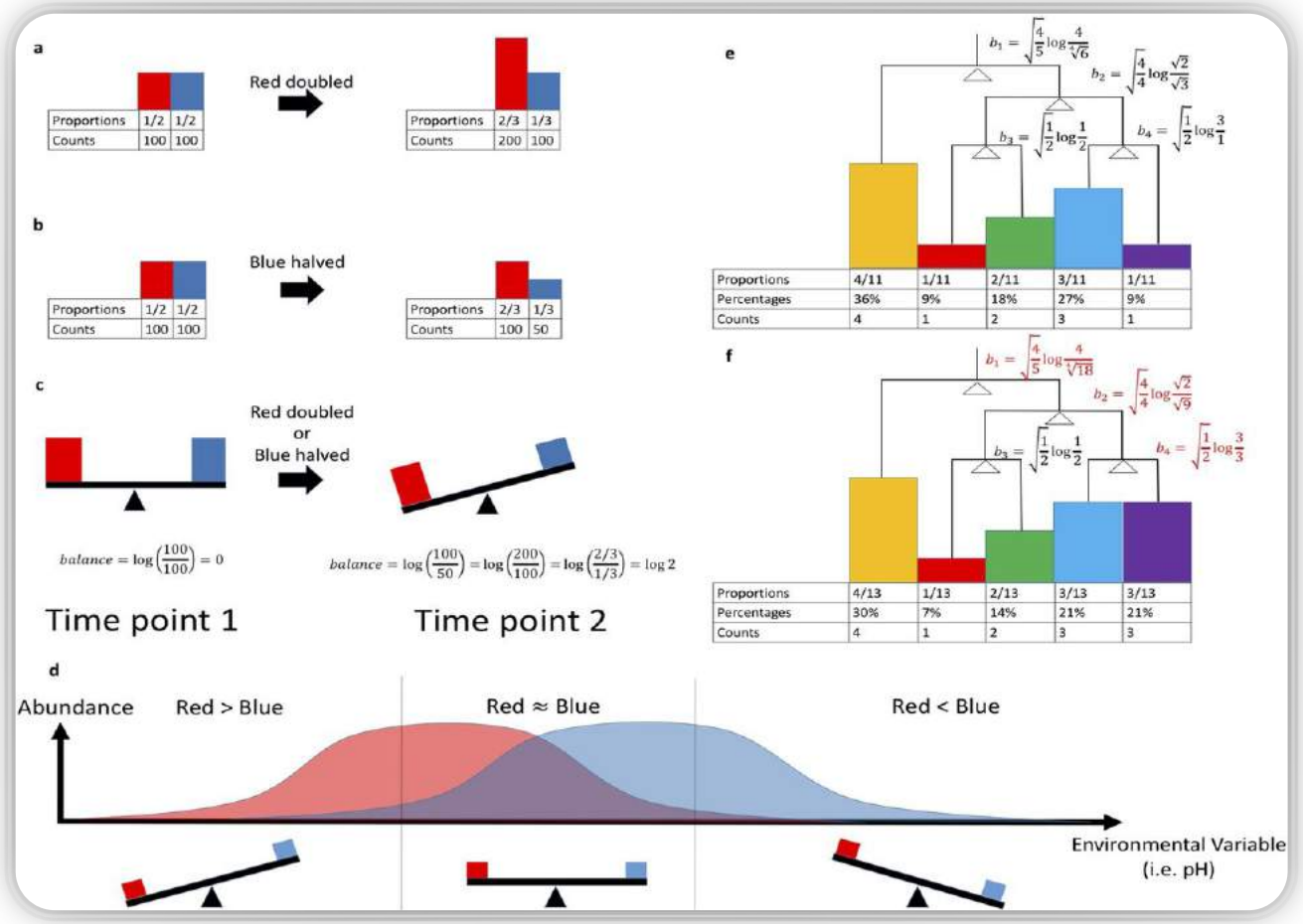


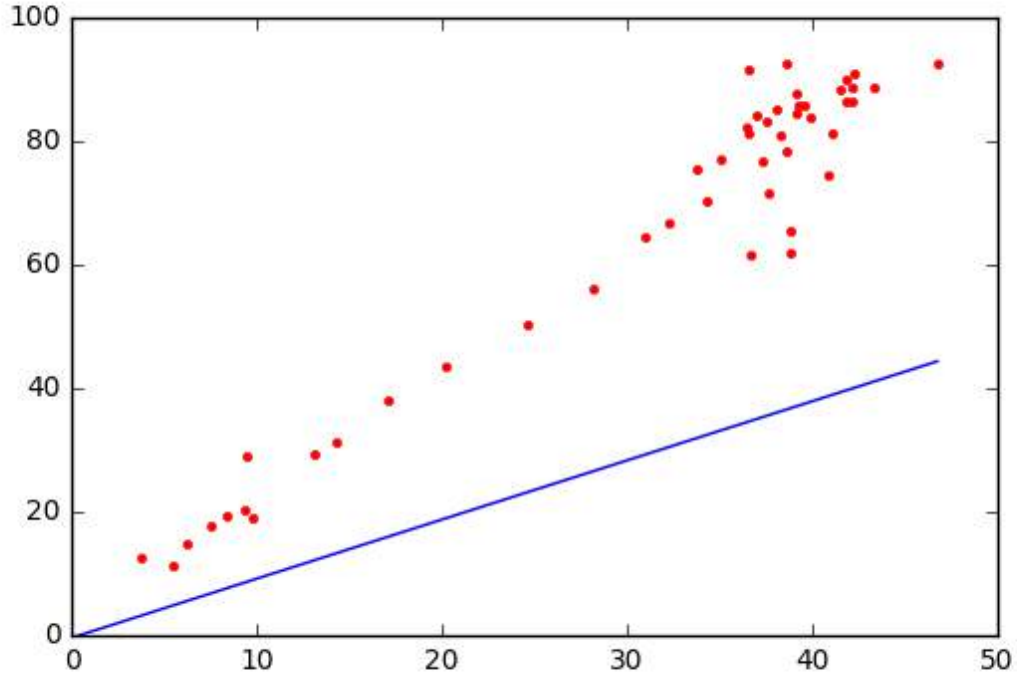
Image from James T. Morton et al. mSystems 2017; 2:e00162-16

Ordinary Least Squares (OLS) with balances

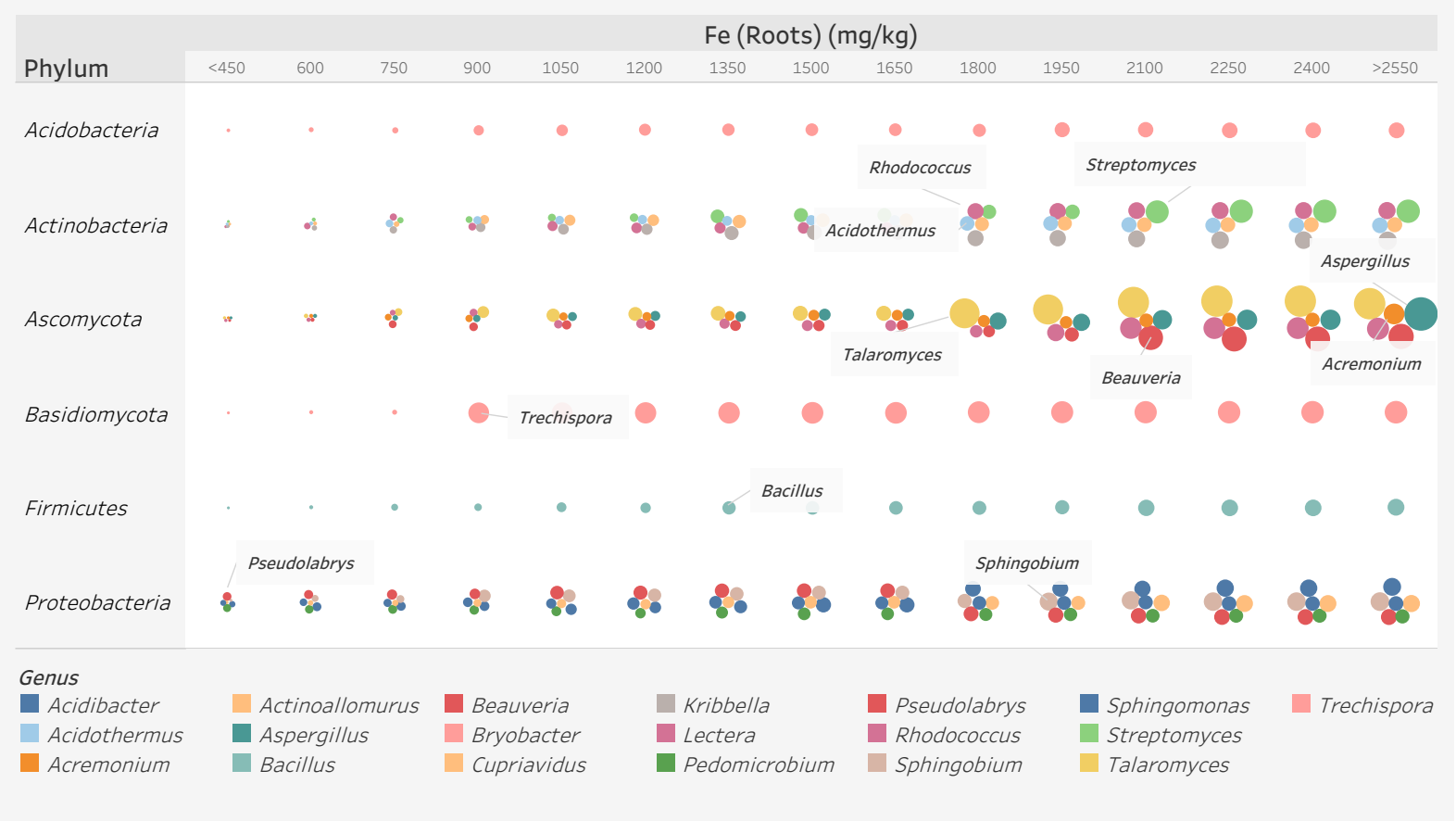
- Finds linear fit to the variables
  - In our case equation is

$$balance_i \sim var_1 + var_2 + etc$$

- We use balances to solve heteroscedasticity and magnitude issues
- Results: tells us which balance is significantly correlated with each variables



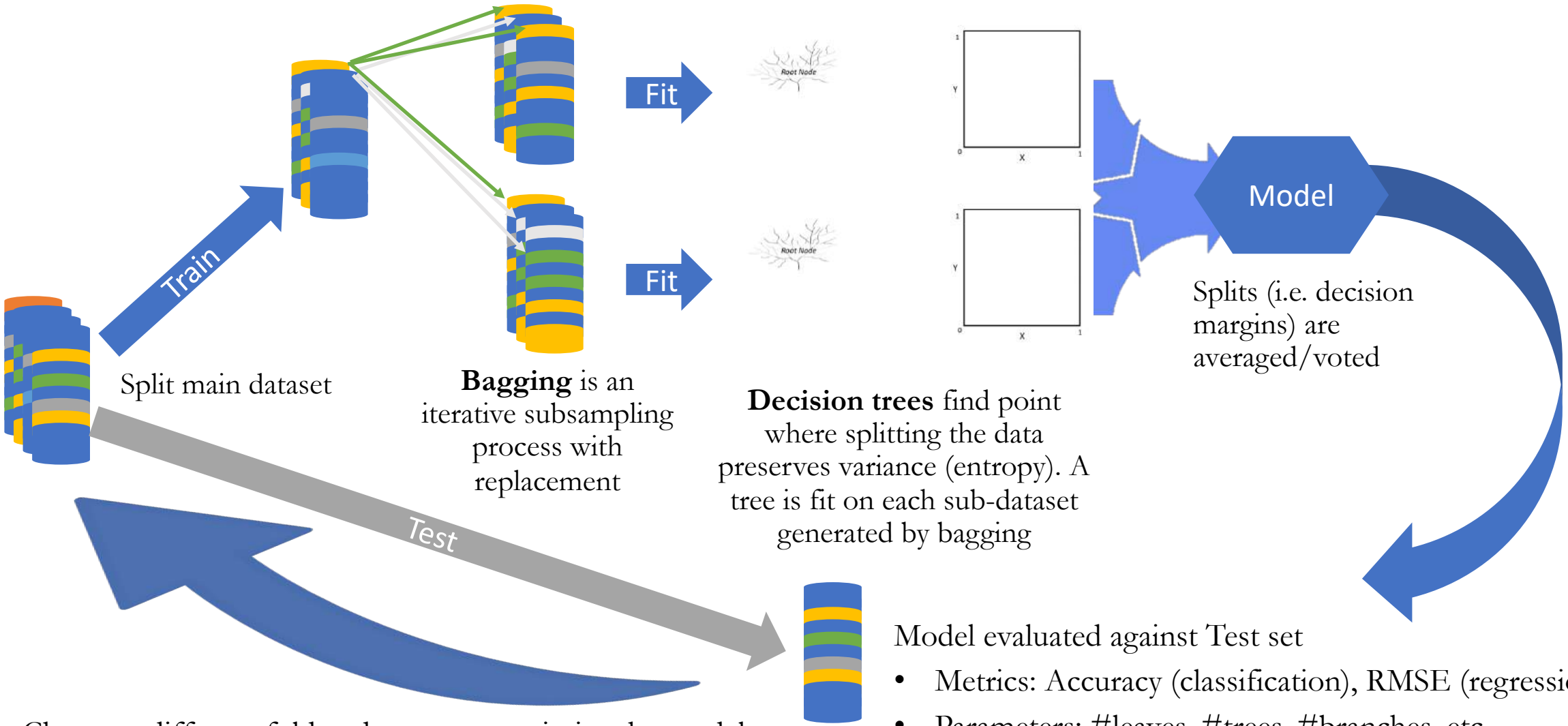
Does it help?


Fe (Roots) (mg/kg)
Phylum
<450
600
750
900
1050
1200
1350
1500
1650
1800
1950
2100
2250
2400
>2550
Acidobacteria
Actinobacteria
Ascomycota
Basidiomycota
Firmicutes
Proteobacteria
Rhodococcus
Streptomyces
Acidothermus
Aspergillus
Talaromyces
Beauveria
Acremonium
Trechispora
Bacillus
Pseudolabrys
Sphingobium
Genus
Acidibacter
Acidothermus
Acremonium
Actinoallomurus
Aspergillus
Bacillus
Beauveria
Bryobacter
Cupriavidus
Kribbella
Lectera
Pedomicrobium
Pseudolabrys
Rhodococcus
Sphingobium
Sphingomonas
Streptomyces
Talaromyces
Trechispora

Random Forest Analysis

- Does not require log-transform
- Can optimally discriminate rare species
- Produces feature importances
- Can be optimized by collapsing the taxa at different levels

Random forest analysis



Split main dataset

Bagging is an iterative subsampling process with replacement

Decision trees find point where splitting the data preserves variance (entropy). A tree is fit on each sub-dataset generated by bagging

Splits (i.e. decision margins) are averaged/voted

Model evaluated against Test set

- Metrics: Accuracy (classification), RMSE (regression)
- Parameters: #leaves, #trees, #branches, etc

Choose a different fold and repeat to optimize the model

Does it help?


Grove 1
Grove 2
Control
Treatment 1
Treatment 2
Treatment 3
Treatment 4
Treatment 5
Fusarium 0.26 %
Fusarium 0.30 %
Fusarium 0.29 %
Exophiala
Fusarium 0.26 %
Fusarium 0.24 %
Exophiala
Fusarium 0.22 %
Fusarium 0.29 %
Unknown
Fusarium 0.25 %
Fusarium 0.17 %
Fusarium 0.25 %
Fusarium 0.20 %
Fusarium 0.16 %
Class
Actinobacteria
Alphaproteobacteria
Anaerolineae
Bacilli
Betaproteobacteria
Dothideomycetes
Elusimicrobia
Eurotiomycetes
Gammaproteobacteria
Gemmatimonadetes
Ktedonobacteria
Nitrospira
Phycisphaerae
Planctomycetacia
Soil Crenarchaeotic Group(SCG)
Solibacteres
Sordariomycetes
Subgroup 6
Subgroup 13
Thermoleophilia
Tremellomycetes

Sampled microbial DNA
- Converted to digital data
- Characterized phylogeny

Combined with other data
- Plant physiology
- Treatments

Multivariate analysis
- β-diversity for relationships between microbiome and plant/treatments

Network analysis
- Determination of metabolic consortia and/or keystone (hub) species

Machine Learning
- Hierarchical clustering and OLS revealed correlations of species and Root Fe conc.
- Random forest revealed chances in less abundant taxa with treatments and management

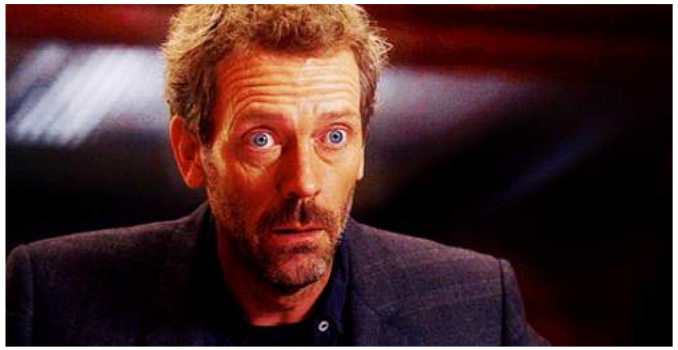
CORRELATIONS ONLY!
- Require experimental validation
- NOT causation

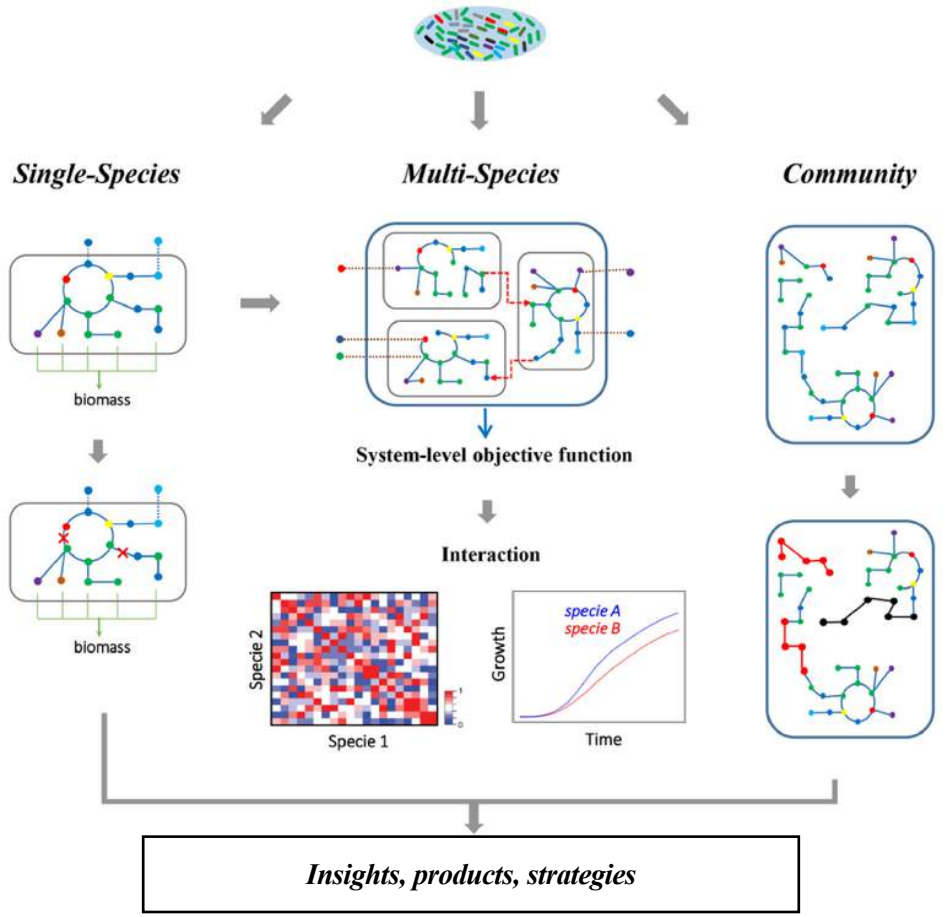
Barely **scratched the surface**
- We only answered the *who*
- What about the function?

Did we find the needle?

Adding complexity

- Dimensionality reduction
  - LASSO
  - Partial least squares
- Whole shotgun metagenome
  - (not limited to rRNA)
- Microbiome-wide association (MWAS)
  - Integration with plant gene expression
- Time-series analysis
  - Dynamic networks
- Metabolome analysis
- State prediction





What for?

- Microbes are everywhere
- Microbes potentially impact anything
- Lot of unsolved questions
  - Correlation vs Causation
  - Functions
  - Metabolites
  - Interactions
  - Manipulation
- From 2011 to 2015:
  - Venture funding up 485% ($114.5 mln)
  - More than $600 mln invested in 2016
  - Includes GSK, Novartis, Indigo, Syngenta. etc

- Huge challenges ahead!!
  - Climate change,
  - Food shortage,
  - Loss of biodiversity
- Bioinformatics *accelerates* discovery
- You can start bioinformatics TOMORROW!
  - Find good online courses
  - Read papers
  - Join meetups
  - Download a dataset and work on it!
  - Join teams!

Team effort!

Acknowledgements

Southwest Research and Education Center, University of Florida

- Prof. Sarah Strauss
- Rachel Berner, BioSci

Miami Machine Learning Meetup

- Mash Zahid, MBA
- Neil Lamarre, PhD
- Guillermo Aure, MSc

University of Bologna, Italy

- Marco Rocca, PhD
- Mariaelena Antinori MSc
- Loredana Baffoni, PhD

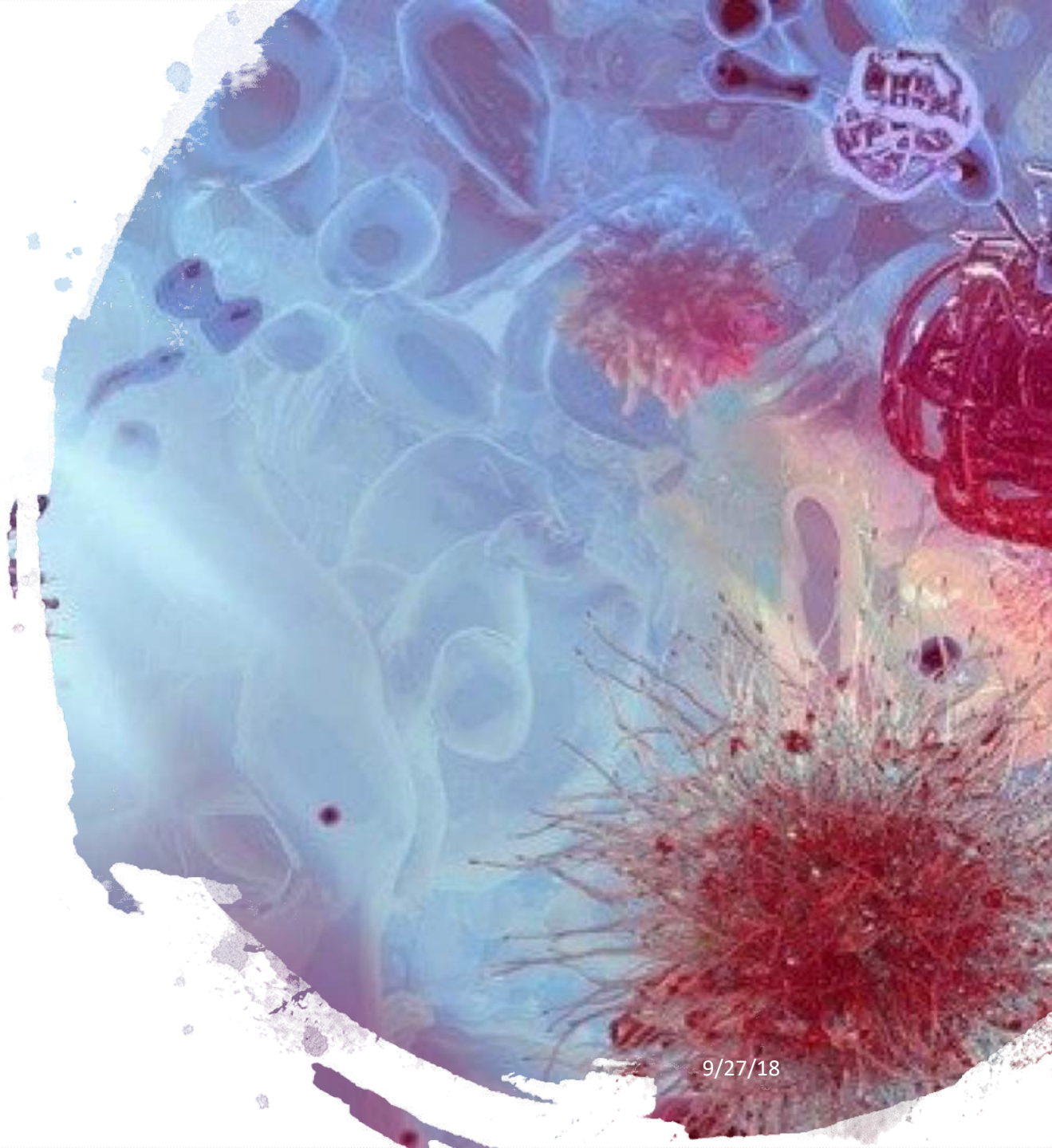
RWTH AAchen, Germany

- Prof. Miriam Rosenbaum
- Bastian Molitor, PhD

"Most people **never ask.** That's what separates the people that **do things** from the people that **just dream** about them."

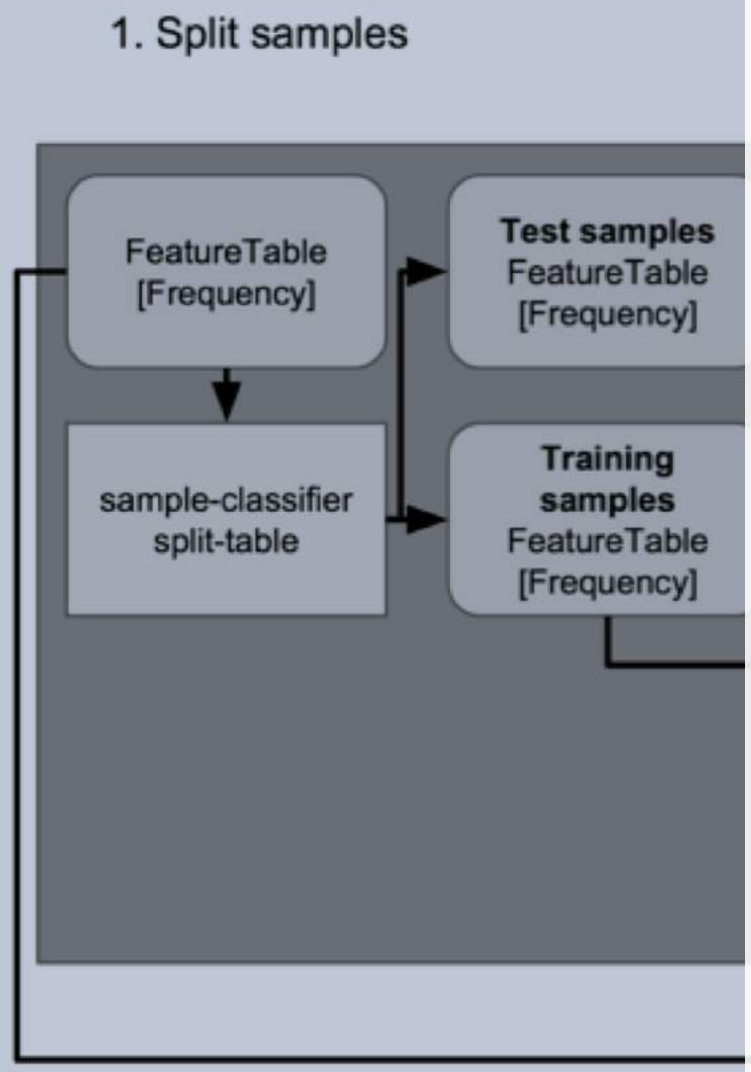
STEVE JOBS

TIME



Andrea Nuzzo, PhD

Slides available at http://andreanuzzo.github.io/Strausslab/UIC.pdf

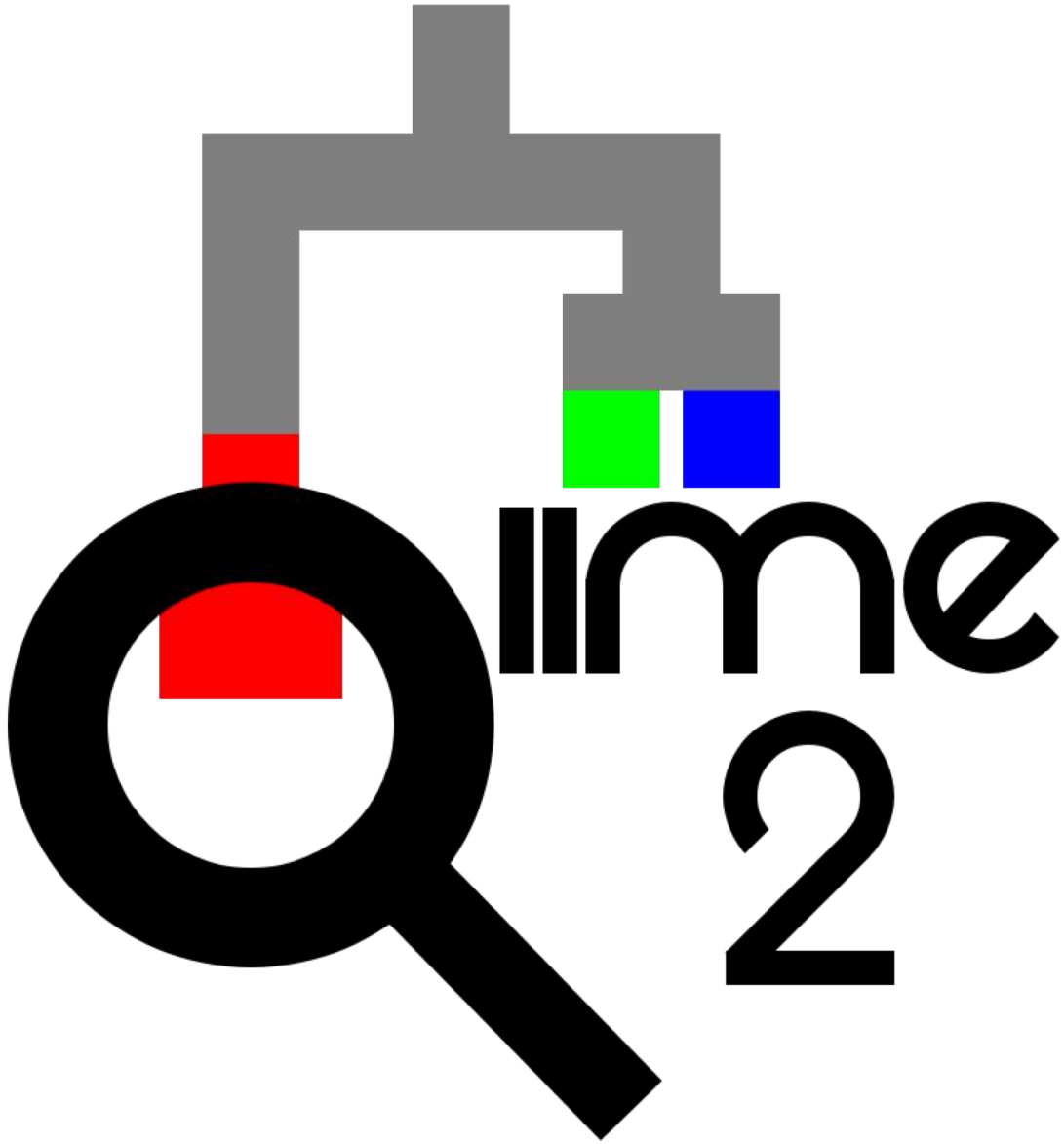


RESOURCES

- GUSTA.ME website for multivariate statistics (https://mb3is.megx.net/gustame)
- Why you don't want to subsample your microbiome data (https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003531)
- Why DADA2 is better than OTU picking (https://www.nature.com/articles/ismej2017119)
- Network analyses in the microbiome (https://www.cell.com/trends/microbiology/fulltext/S0966-842X(16)30185-8#secsect0020)
- Review on normalizations (https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0237-y)
- Bioconductor workflow (https://f1000research.com/articles/5-1492/v2)
- Phyloseq website (https://joey711.github.io/phyloseq/index.html)
- Qiime2 tutorials (https://docs.qiime2.org/2018.8/tutorials/)
- Introduction to statistical learning book (http://www-bcf.usc.edu/~gareth/ISL/) and mooc (https://lagunita.stanford.edu/courses/HumanitiesSciences/StatLearning/Winter2016/about)
- Machine Learning A to Z™ mooc (https://www.udemy.com/share/100034BUYcc1ZRR34=/)
- Wikipedia Confusion Matrix for ML model evaluations (https://en.wikipedia.org/wiki/Confusion_matrix)

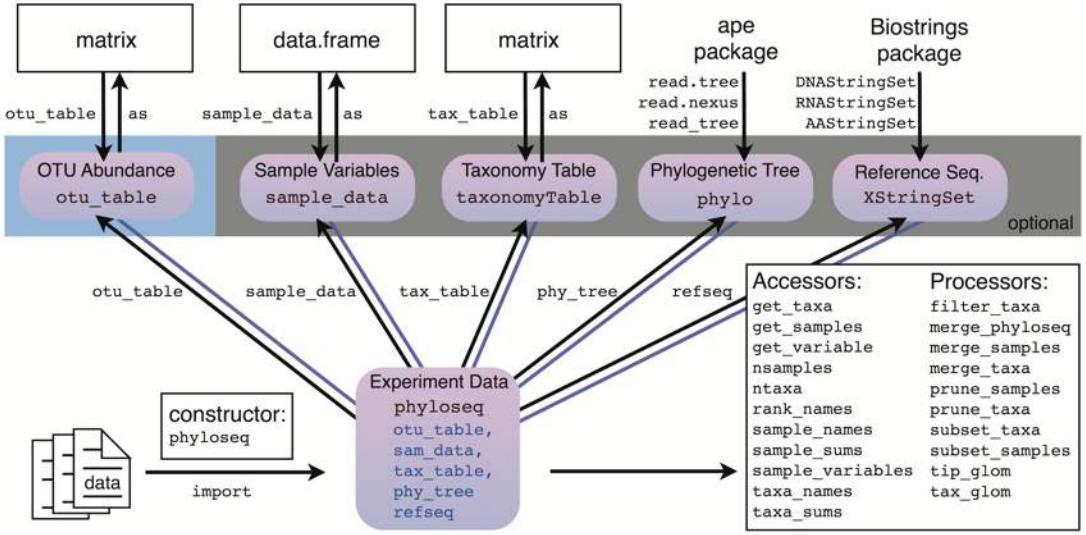
Introducing: Qiime2

- Modular platform for Quantitative insights in microbial ecology
- Written in Python3
  - Calling scripts from bash
  - API in Jupyter
- Plugin-structure
- Performs: DADA2, diversity analyses, machine learning, time series analyses and much more
- Standard for microbiome characterization



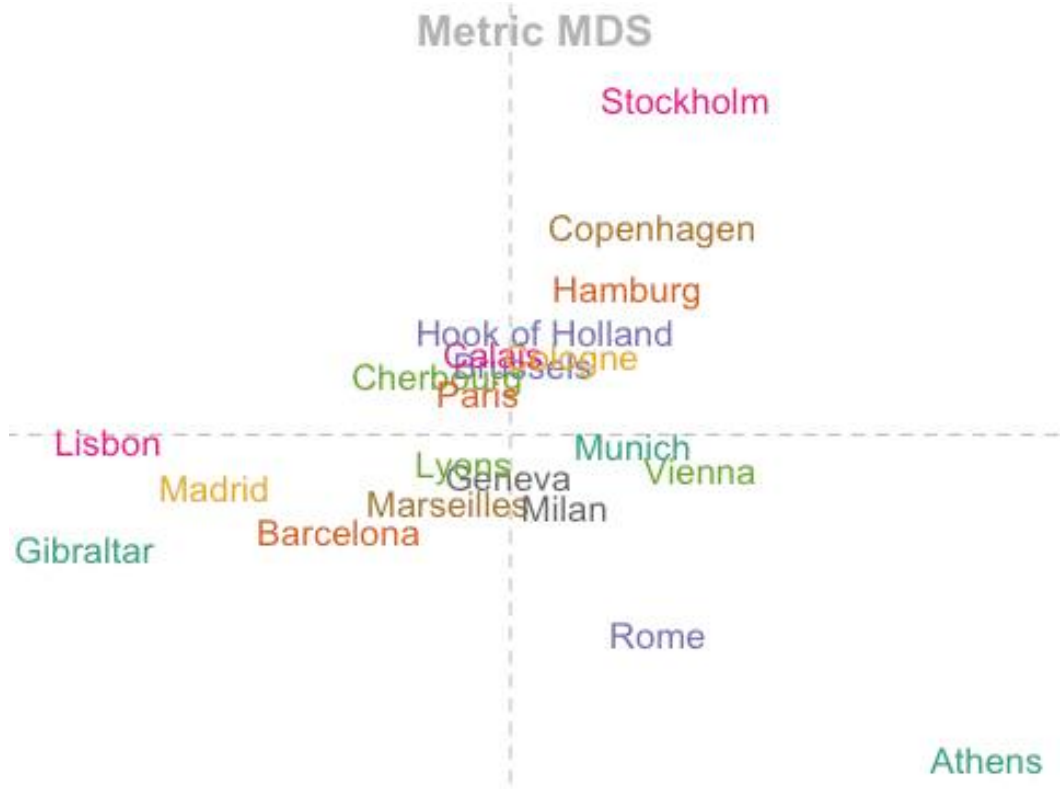
Introducing Phyloseq

- Package in R
- Made by biostatisticians
  - Much more complete and accurate than qiime
- Integrates with ggplot2
  - *Nice graphs*!
- Performs: DADA2, in-depth diversity analyses, DeSeq2, network analysis, and much more



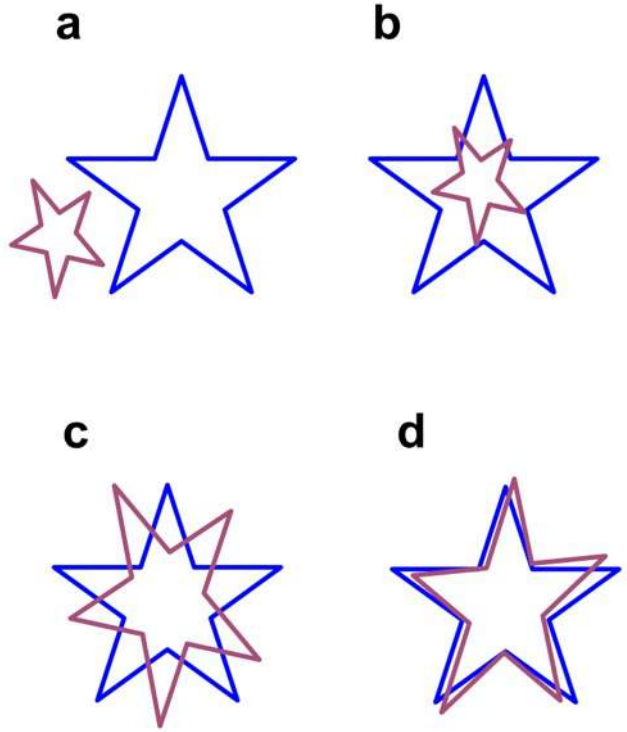
PCoA: Principal coordinate analysis (or MDS)

- NOT PCA
- If I know distance between cities, but not their coordinates, how can I draw a map?
- Your count table is converted into a matrix of dissimilarity (using the diversity index chosen)
- May be impacted by high variance (so, you need to normalize)
- If some results in the dissimilarity matrices are negative, it leads to imaginary numbers in the eigenvectors
- The eigenvectors are not your variables, but are correlated with different percentages



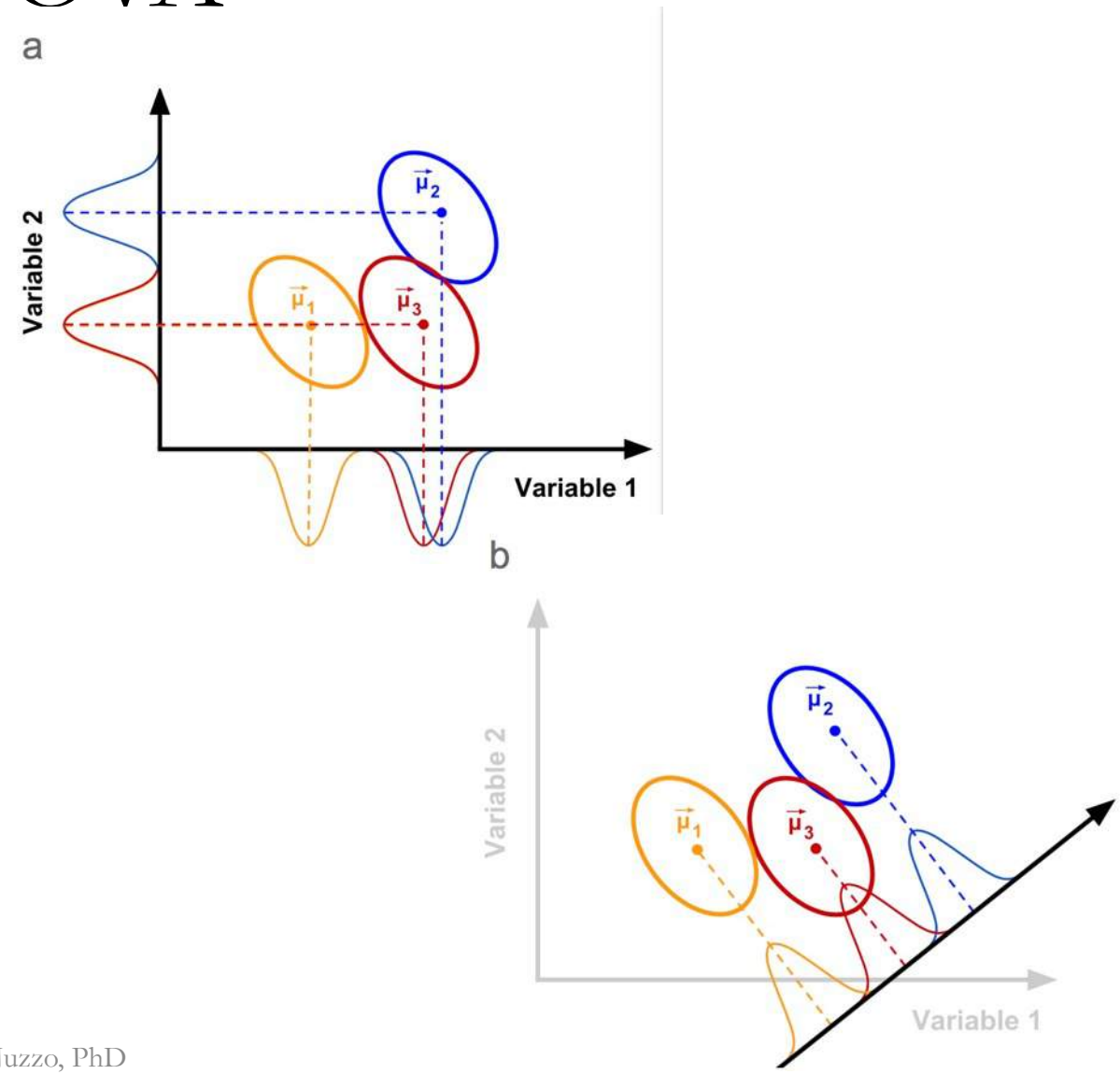
NMDS: NON-metric multidimensional scaling

- Starts from a dissimilarity matrix
  - Ranking
- More robust than PCoA
- Applies Procrustes analysis to modify the matrix so that the eigenvectors are close to the original dimensions
  - Generate a ”stress value”
    - <0.1 good model
    - 0.1 < stress < 0.2 meh model
    - 0.2 < stress < 0.3 bad model
    - >0.3 random
- Eingenvectors are still NOT the original dimensions



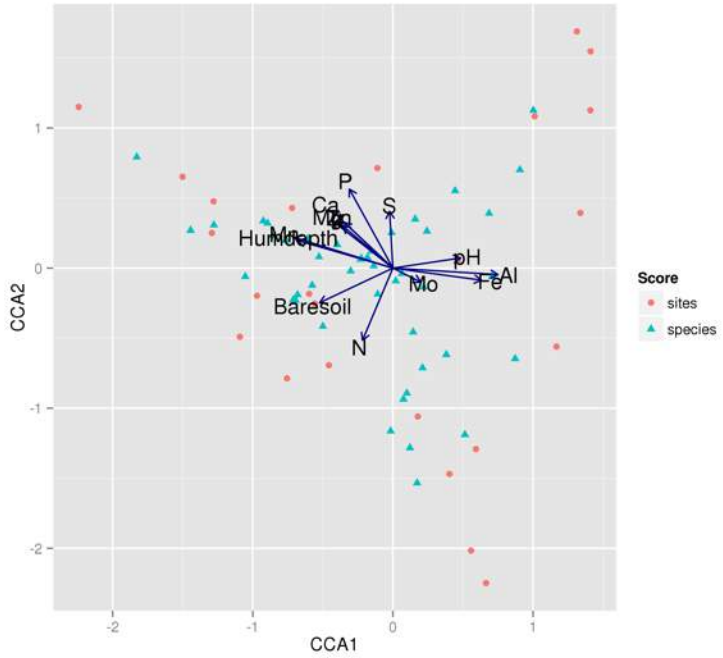
ANOSIM & PERMANOVA

- Both similar to ANOVA
- ANOSIM uses a dissimilarity matrix instead of the raw data
  - Finds differences between groups
  - Highly sensitive to dispersity
- PERMANOVA is a
  - Multivariate ANOVA (i.e. multiple factors influence multiple responses)
  - With PERmutations (solves the problem of limited number of samples)
  - Sensitive to dispersity



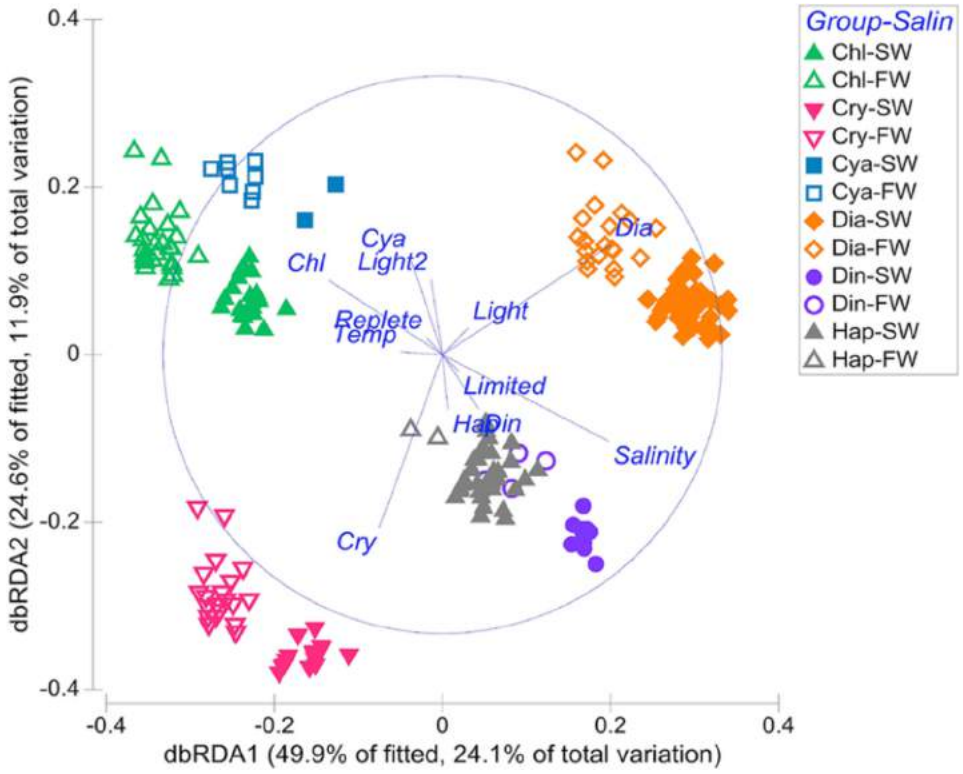
CCA: Canonical Correspondence analysis

- Analyzes correspondences between a matrix of frequencies and a matrix of variables
  - How much the frequencies deviate from random per each variable?
- Correlate counts to variables (finally!)
- Does not try to maximize coverage of variance (unlike PCA)
- You can use ANOVA on CCA
- BEWARE of using only significant variables
  - Avoid collinearity (VIF)



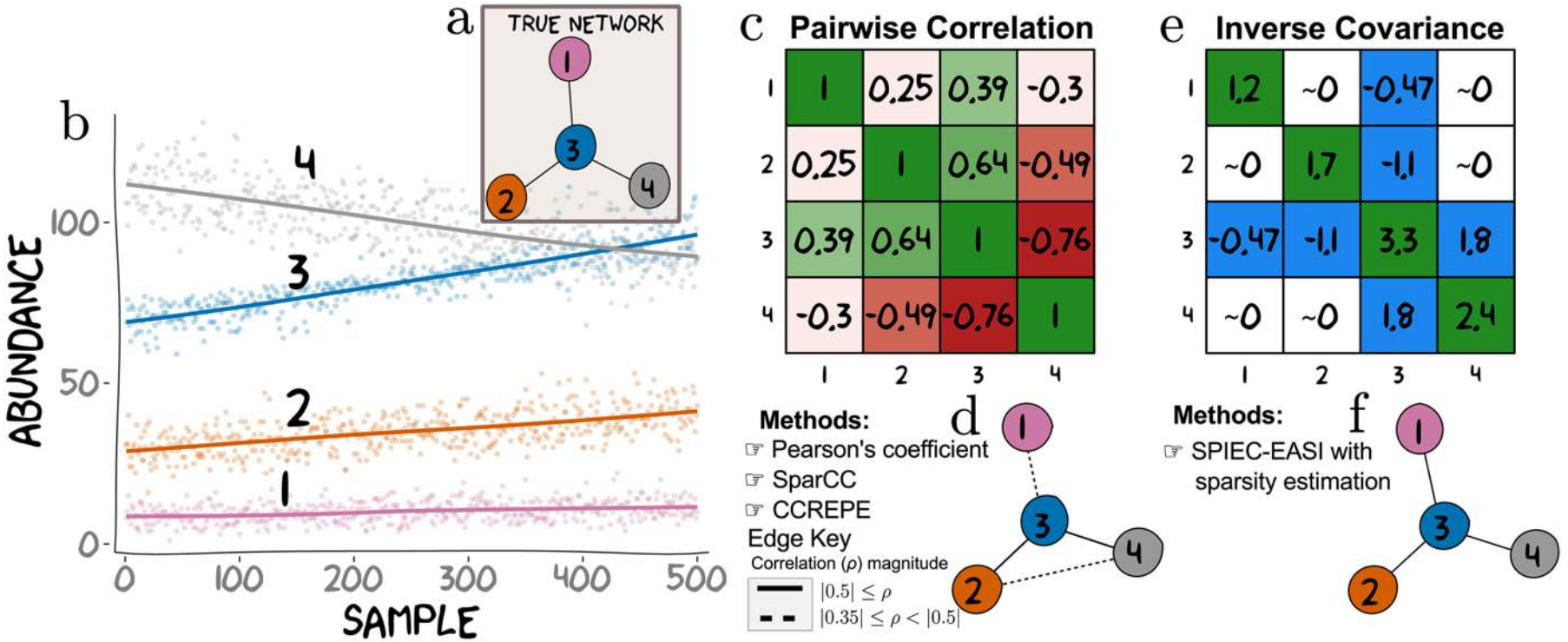
CAP (or db-rda): Constrained analysis of principal coordinates

- Basically a PCA where you constrain your components to your variables
  - Maximizes explanation of variance
- Similar (but different) to CCA
  - Still sensitive to collinearity
- BEWARE of using only significant variables



Network analyses with SPIEC-Easi

- Assume interdependencies of OTUs
- Draws samples from negative binomial distributions
- Sparse data → inverse covariance matrix depends on the conditional states of all available nodes
- Avoids weak or false positive associations



*Kurtz, Z.D. et al. (2015). PLoS Comput. Biol., 11, 1–25

OTHER ML Algorithms that might be useful

- Support Vector Machines: find hyperplanes that separate data into features minimizing the tolerance boundaries
- Logistic regression: the basic of all classification methods (logit function, between 0 and 1)
- K-Nearest Neighbors: classify the new point based on the majority of its closest points
- PCA and LDA for dimensionality reduction
- Ridge and Lasso: two regularization methods for linear regression (discard features to increase $AdjR^2$)