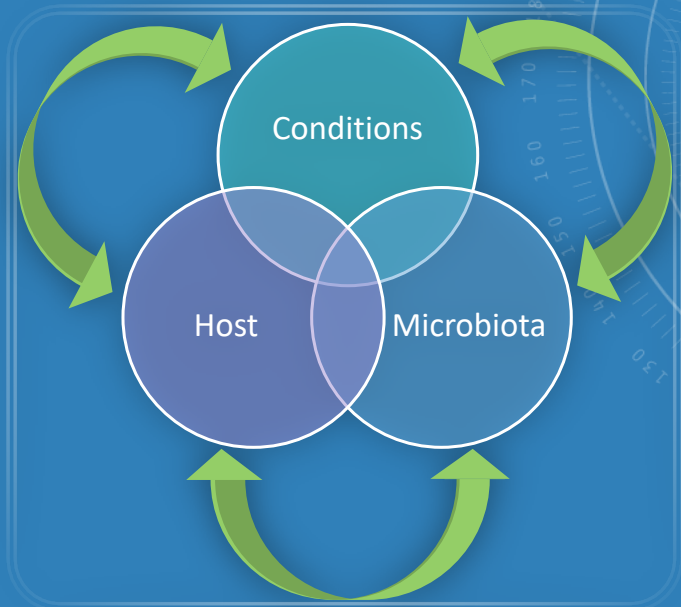
WHERE WE LEFT OFF

- We took our environmental/host data
  - Did we found which condition impact the host?
- We took our microbiome data
- We produced count table, taxonomy, trees
- We made a bunch of multivariate analysis
  - Did we understand which condition impacts the microbiome? Where?
  - Did we understand if microbiome impacts host? Why?

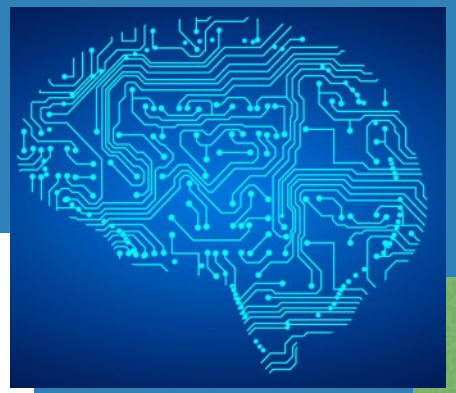


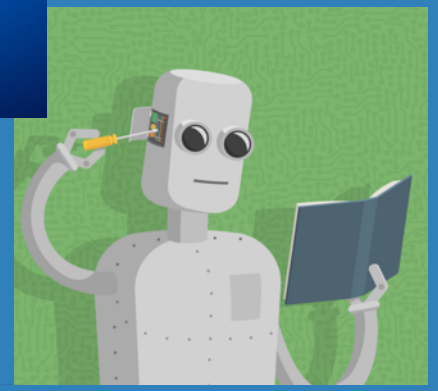
AN ALTERNATIVE METHOD

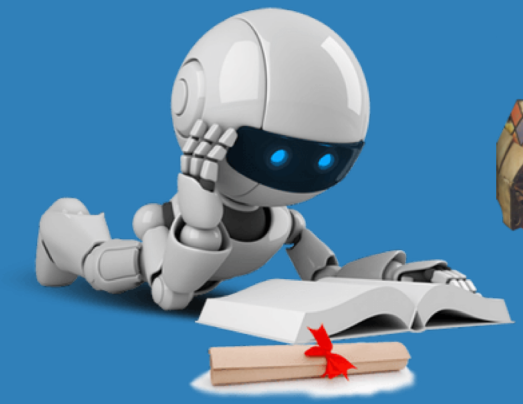
- Generalized
- Powerful
- Advanced
- Sensitive

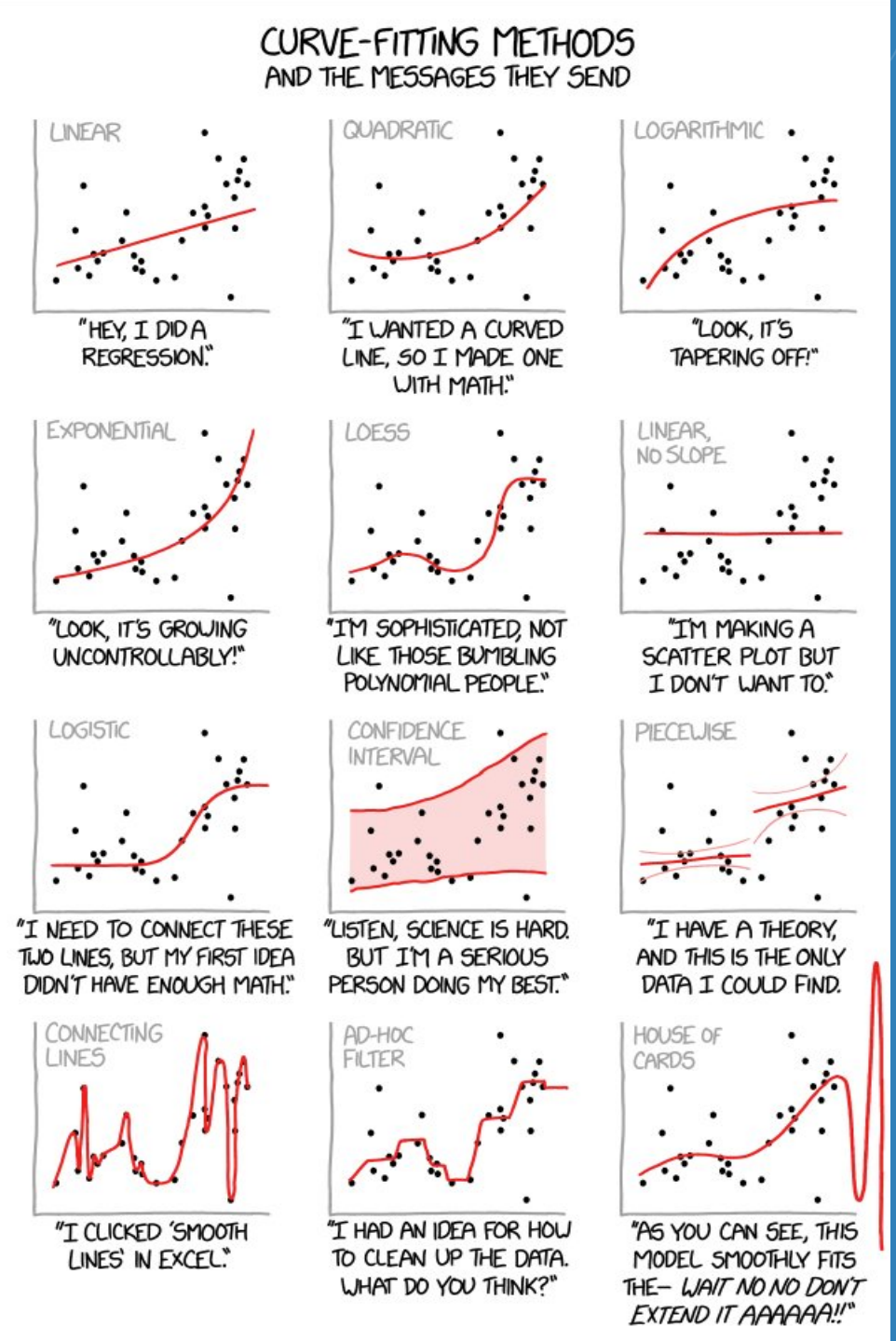
WHAT IS MACHINE LEARNING?

- 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





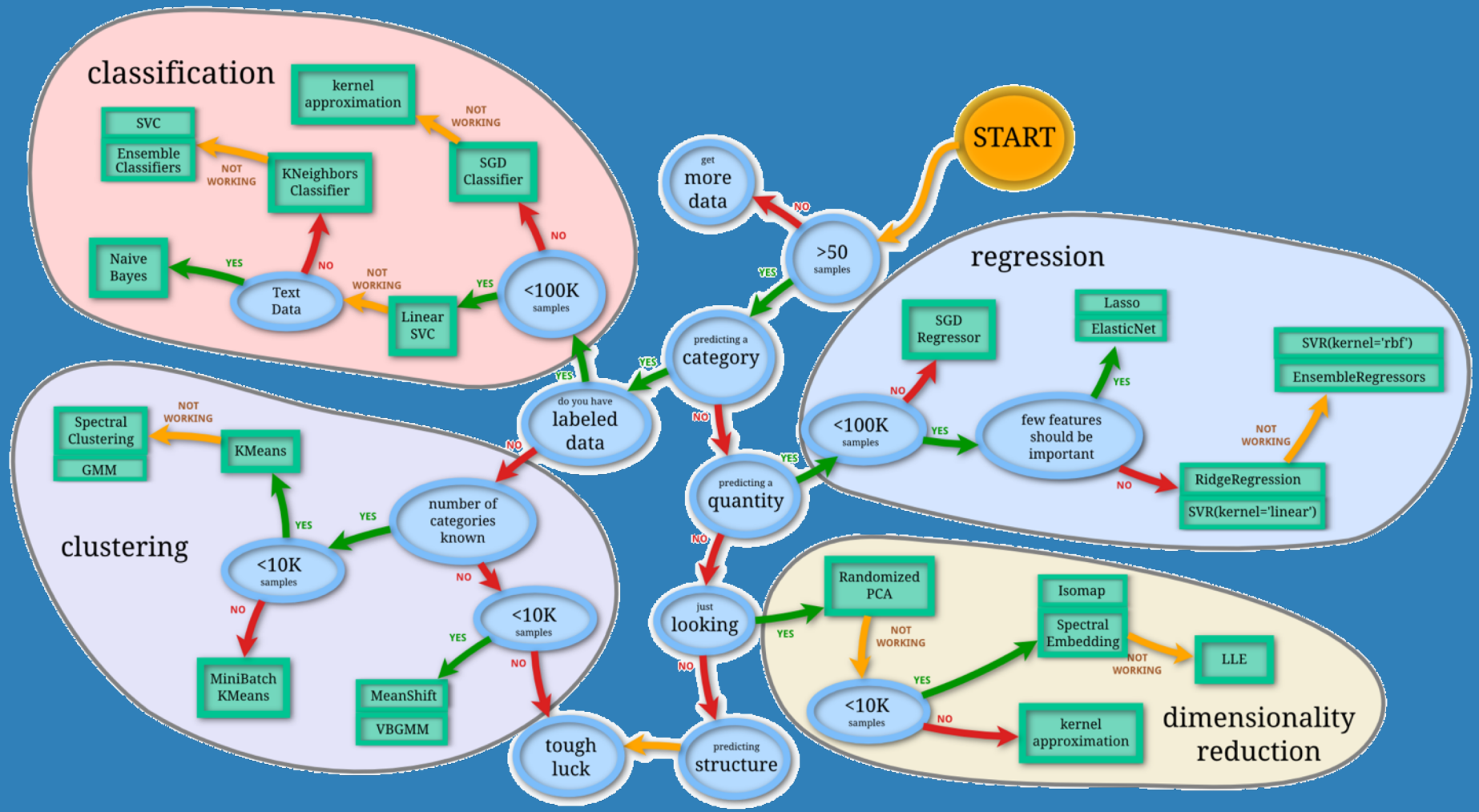




MACHINE LEARNING CATEGORIES

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?



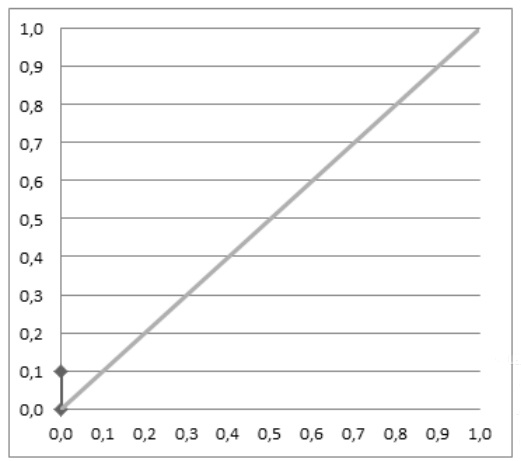
HOW DO I KNOW MY CLASSIFICATION IS GOOD?

- USE THE CONFUSION MATRIX
  - Precision $= \frac{TP}{TP+FP}$
  - Recall $= \frac{\mathrm{TP}}{\mathrm{TP+FN}}$
  - Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
  - $F1\ score = 2\frac{Precision*Recall}{Precision+Recall} = \frac{2TP}{2TP+FP+FN}$
  - AUC (Area Under the Curve)
    - What curve? ROC = Recall over FP/N
  - Etc…

"Essentially, all models are wrong, but some are useful"
George Box

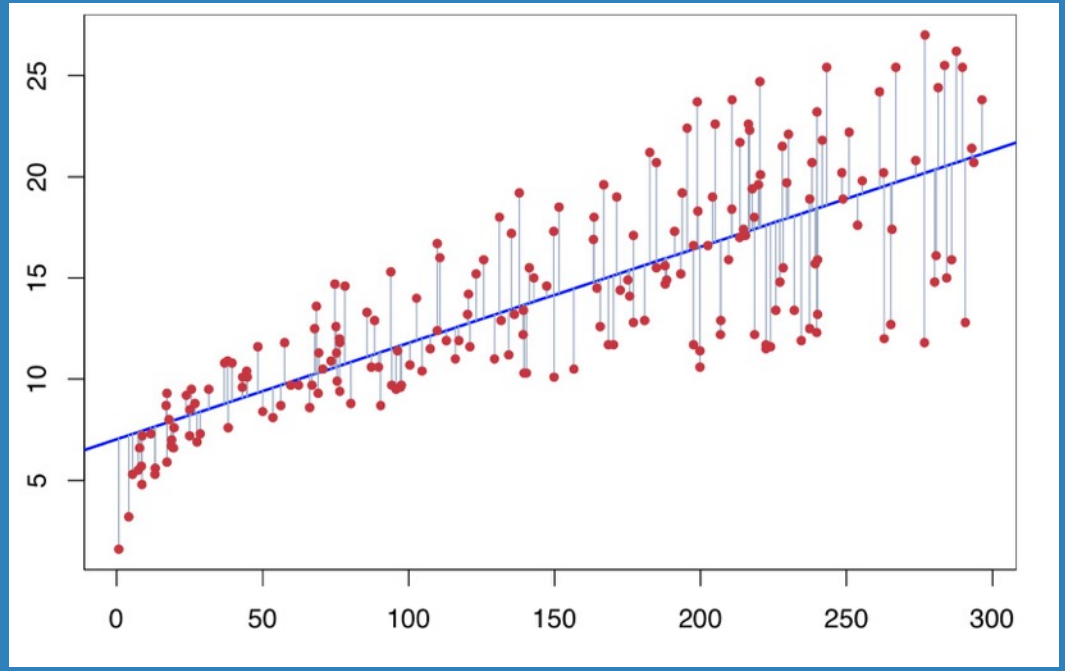
What model says	What reality is: Positive	What reality is: Negative
Positive	True Positive (TP)	False Positive (FP) !1
Negative	False Negative	True Negative

#	C	Score
1	P	0,9
2	P	0,8
3	N	0,7
4	P	0,6
5	P	0,55
6	P	0,54
7	N	0,53
8	N	0,52
9	P	0,51
10	N	0,505
11	P	0,4
12	N	0,39
13	P	0,38
14	N	0,37
15	N	0,36
16	N	0,35
17	P	0,34
18	N	0,33
19	P	0,3
20	N	0,1



HOW DO I KNOW MY REGRESSION IS GOOD?

- $\text{RSS} = \sum(y_i - \hat{y}_i)^2$ (error of the model)
- $\text{TSS} = \sum(y_i - \bar{y})^2$ (variance in the sample)
- $R^2 = 1 - \frac{RSS}{TSS}$ (NB: never decreases when increasing number of predictors, regardless of their significance)
- $MAE = \frac{1}{n}\sum|y_i - \hat{y}_i|$
- $RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$
- $AdjR^2 = 1 - \frac{RSS}{TSS} \times \frac{n-1}{n-k-1}$

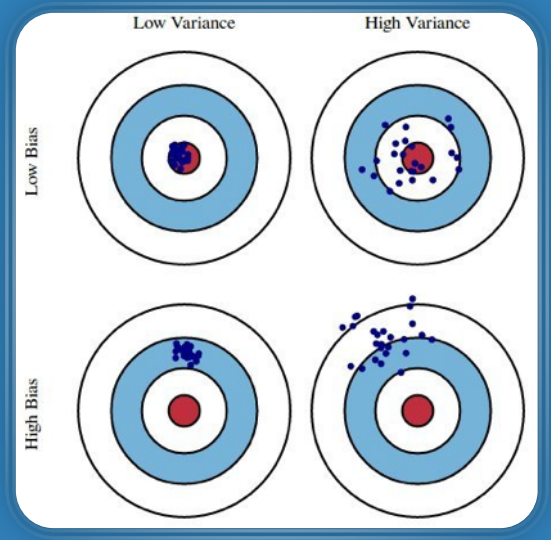


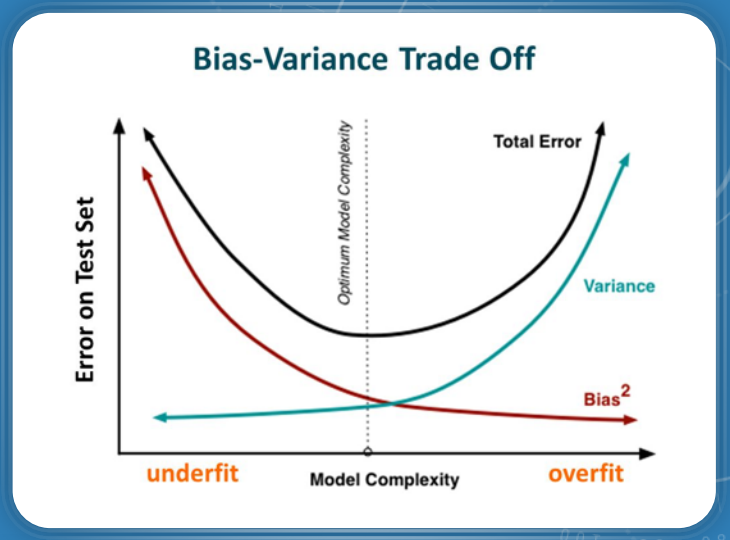
"Essentially, all models are wrong, but some are useful"

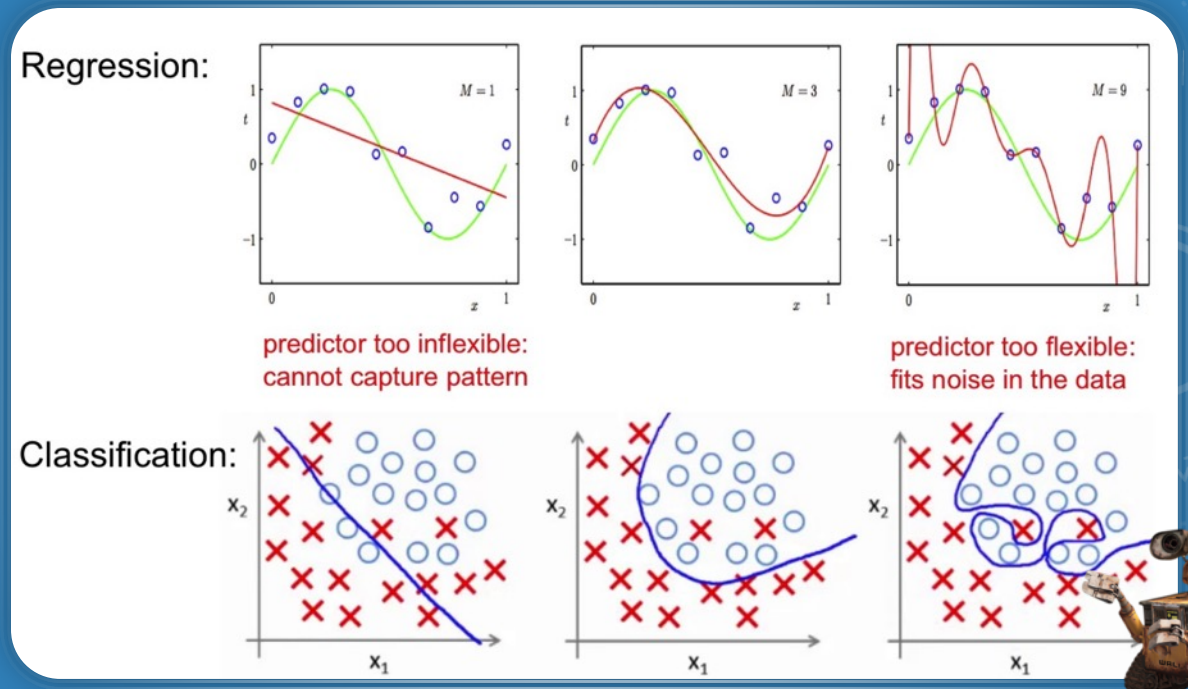
George Box

WHY ALL MODELS ARE WRONG?

- Bias-variance trade-off
  - Bias: distance from real value
  - Variance: dispersion around your mean prediction
- Right fit is hard to choose:
  - Underfitting: does not capture pattern
  - Overfitting: will fit data perfectly but do all predictions wrong
- High bias leads to underfitting
- High variance leads to overfitting

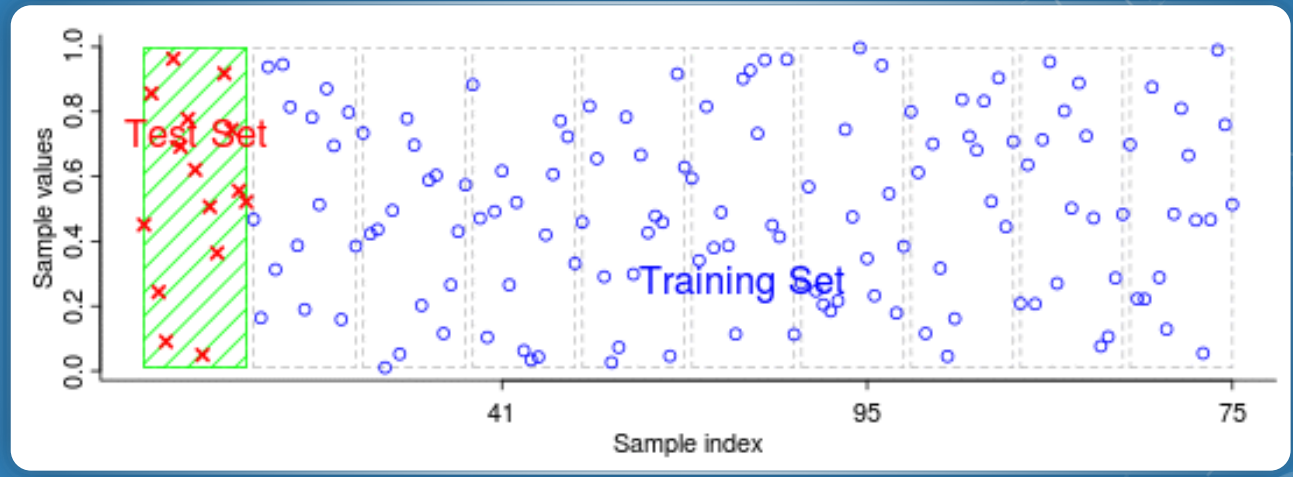






HOW CAN I OPTIMIZE THE MODEL?

- Feature engineering
- Hyperparameter tuning (i.e. parameters of the algorithm)
- Resample your data:
  - Cross-validation (k-fold, leave one out, etc.)
  - Bootstrap, bagging, boosting
- Cross validation is often used to fine-tune hyperparameters and feature selection



LOG-TRANSFORMED BALANCES

- 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
  - Avoids the problem of relative abundances
  - Breaks time-dependence

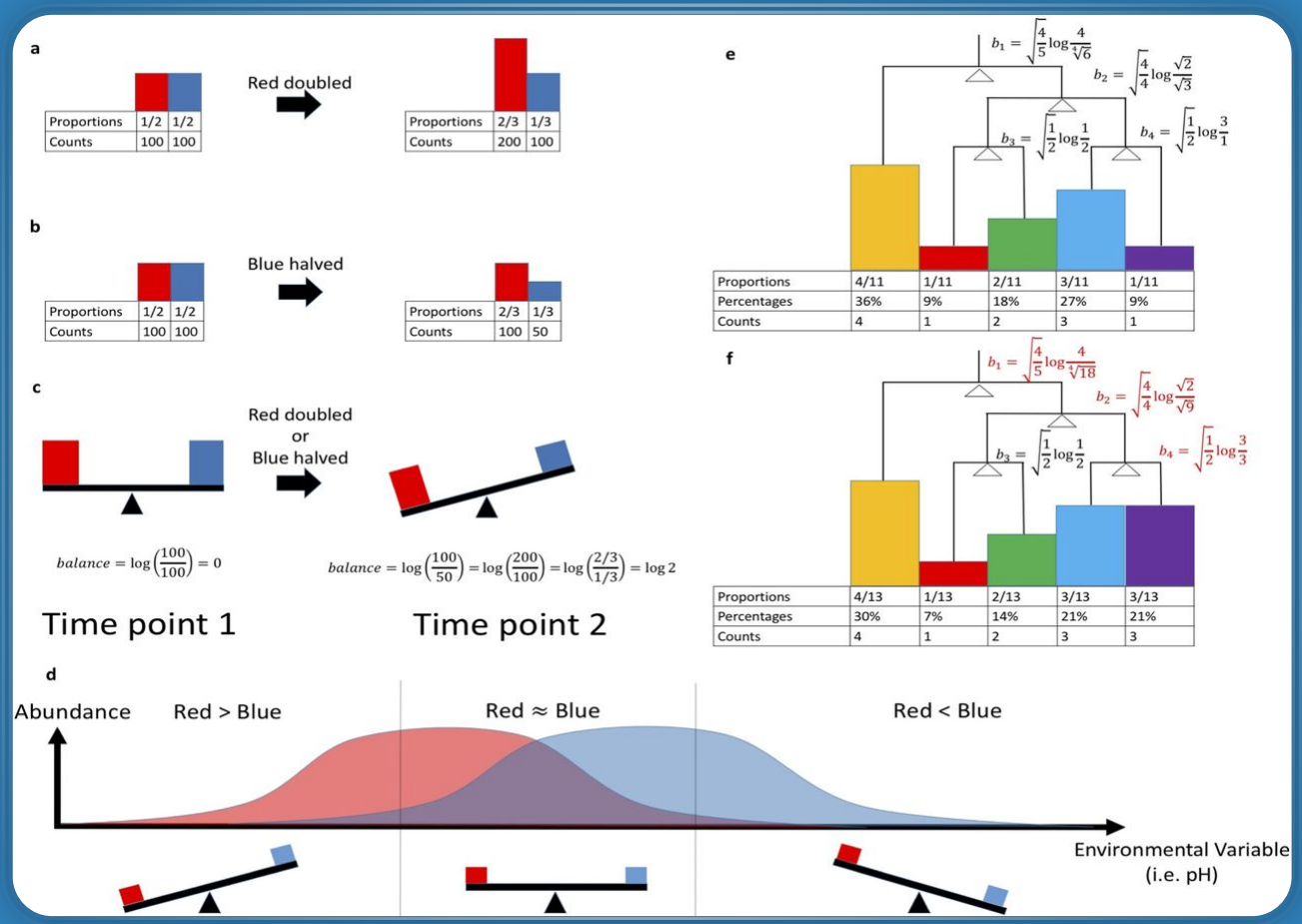


Image from James T. Morton et al. mSystems 2017; 2:e00162-16

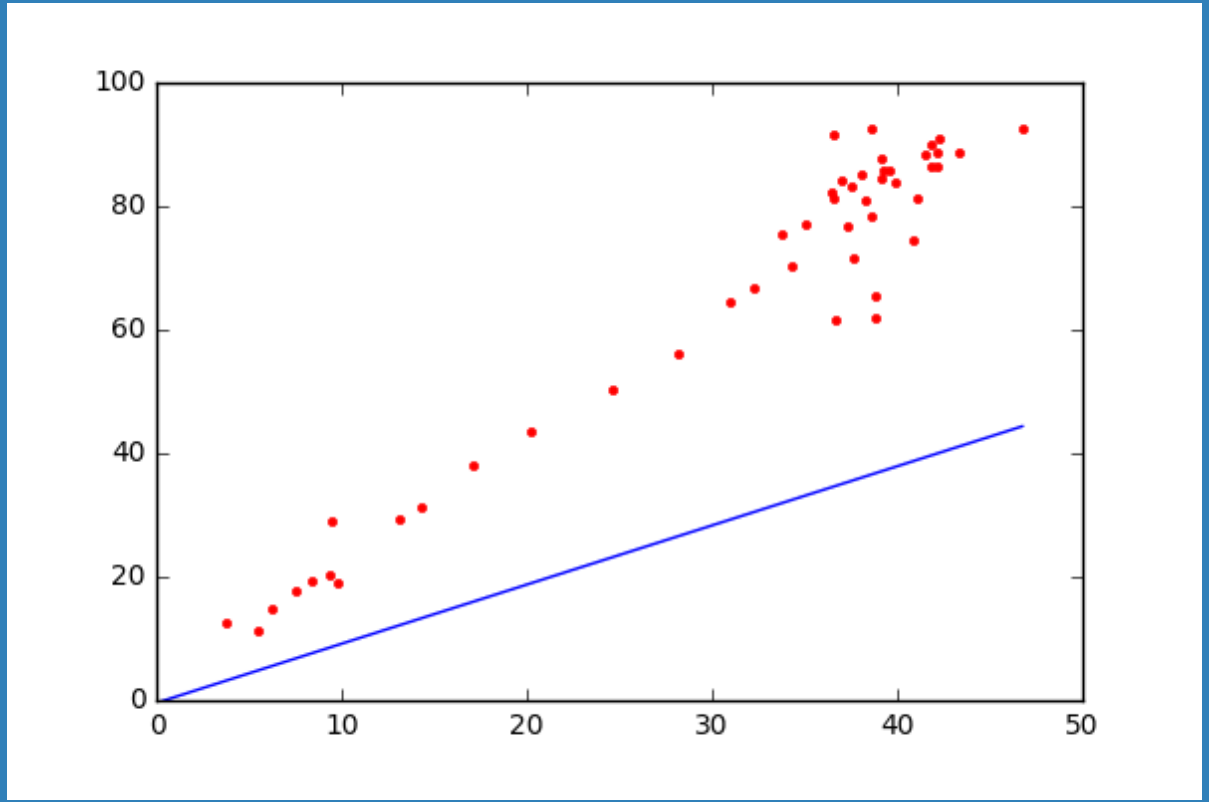
ORDINARY LEAST SQUARES (OLS)

- Finds linear fit to the variables
- Minimizes RMSE
  - Sensitive to magnitude and heteroscedasticity

- In our case equation is

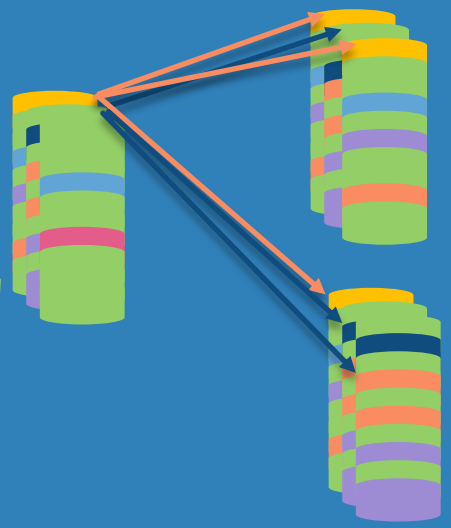
  $balance_i \sim var_1 + var_2 + etc$

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



RANDOM FOREST ANALYSIS

- **Feature importance** is a measure the weight of each feature in the split → this gives us what we want
- In our data, a good "feature engineering" is to collapse the taxa to higher levels
- No data transformation required

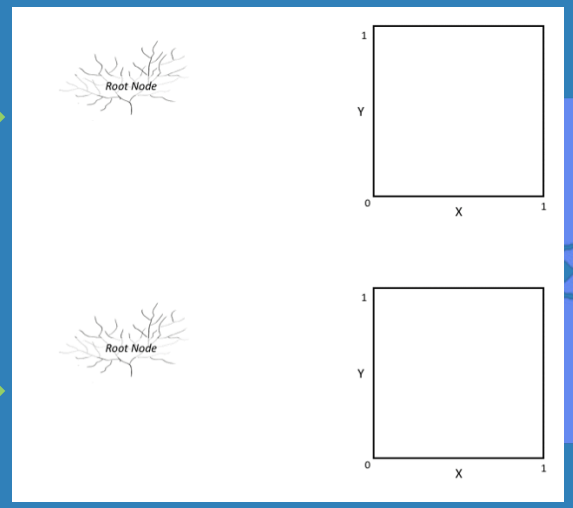


Train

Split main dataset

Fit

Fit



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 subdataset generated by bagging

Model

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

Test

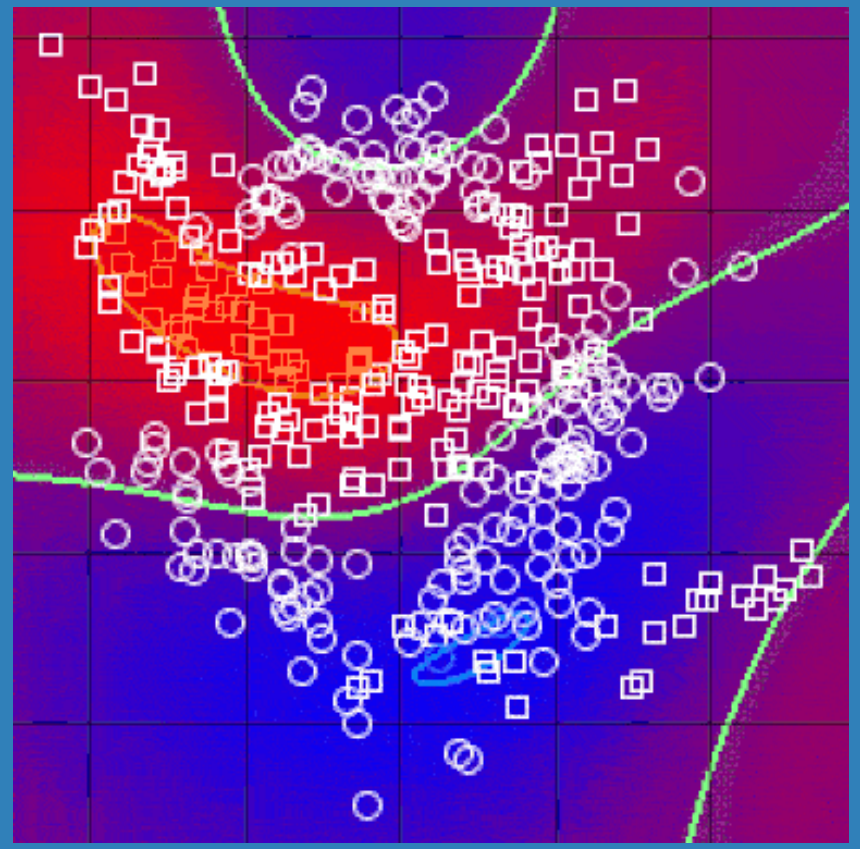
Model evaluated against Test set

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

Choose a different fold and repeat to optimize the model

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 $AdjR^2$)



HOW ARE WE DOING?

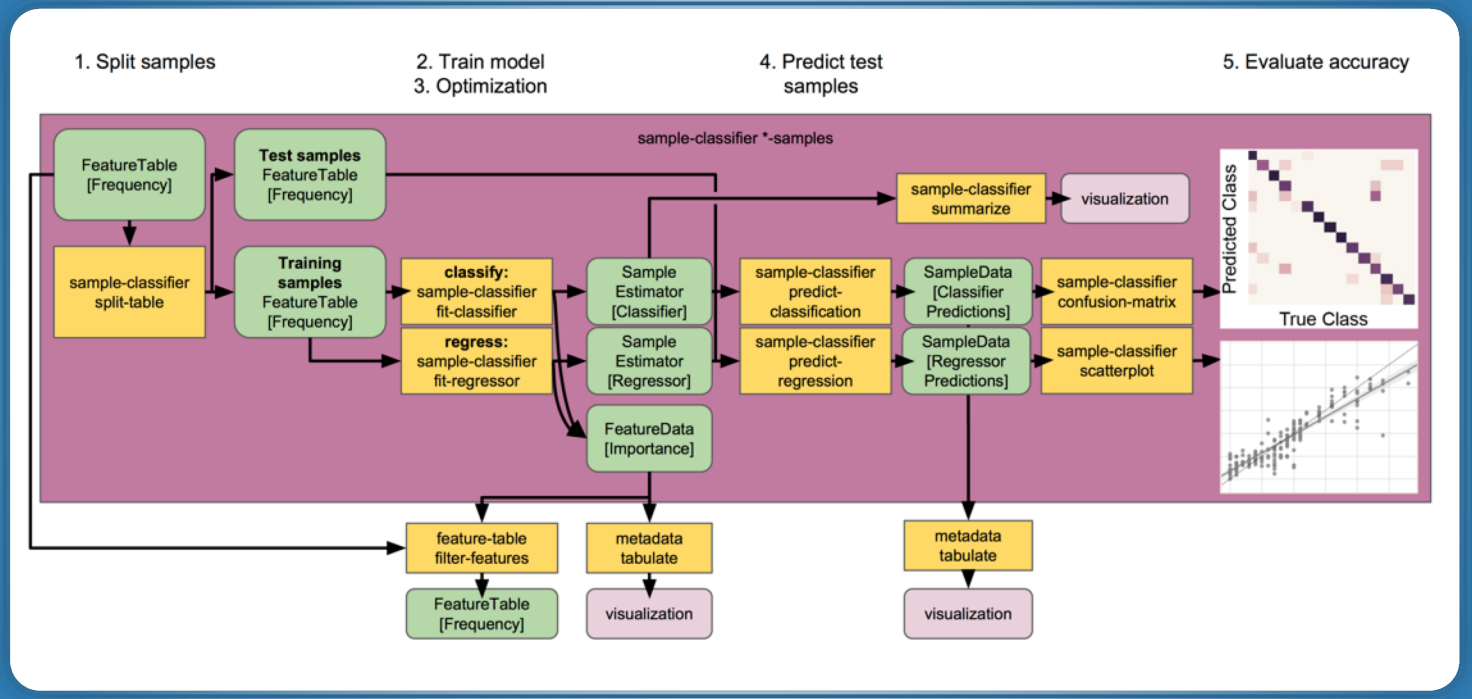
A

B

C

CODE

OTHER RESOURCES



- GUSTA.ME website for multivariate statistics (https://mb3is.megx.net/gustame)
- Phyloseq website (https://joey711.github.io/phyloseq/index.html)
- Waste Not Want Not paper (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 review (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)
- Qiime2 tutorials (https://docs.qiime2.org/2018.8/tutorials/)
- Wikipedia Confusion Matrix (https://en.wikipedia.org/wiki/Confusion_matrix)
- 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=/)