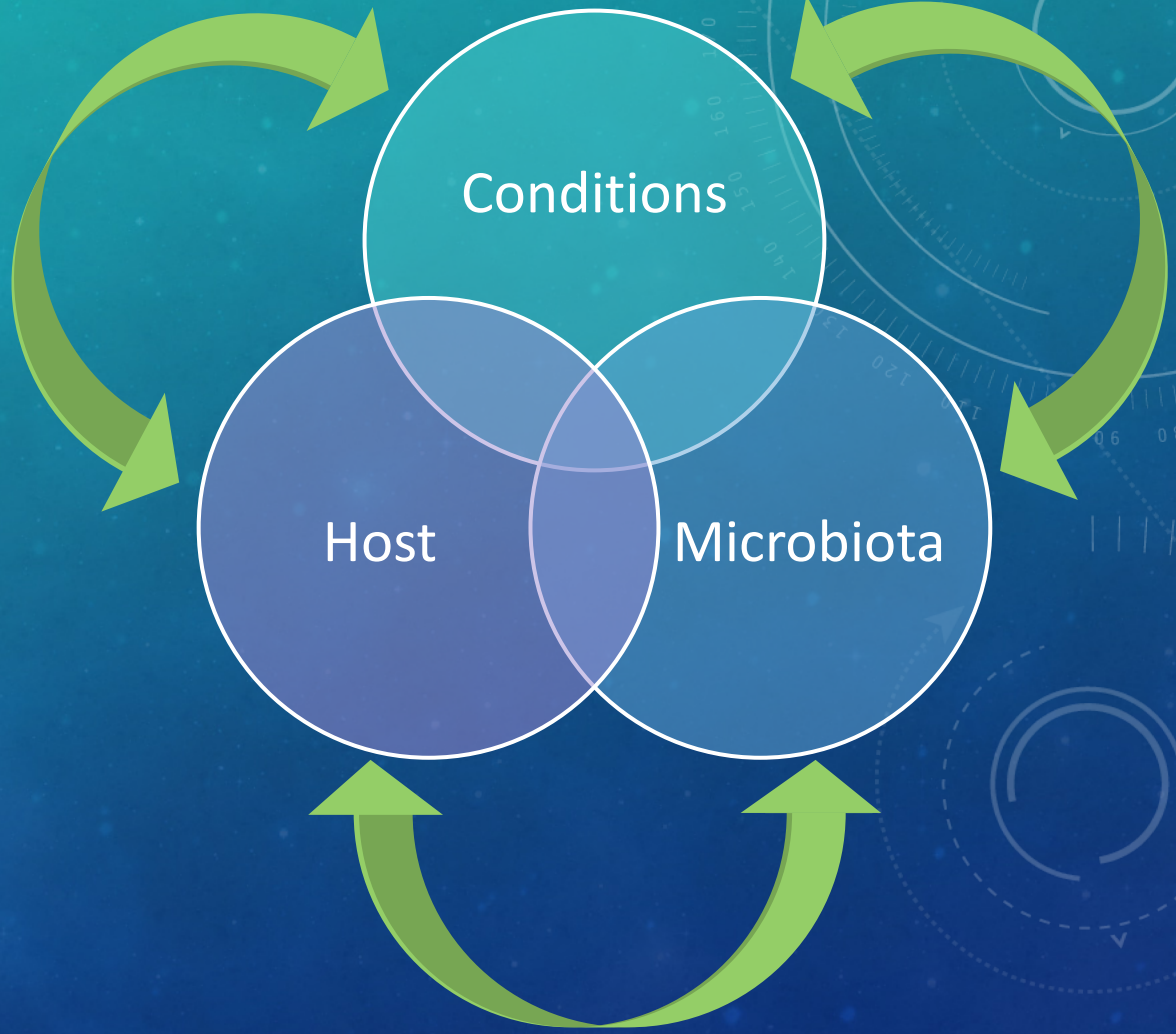
MICROBIOME ANALYSIS

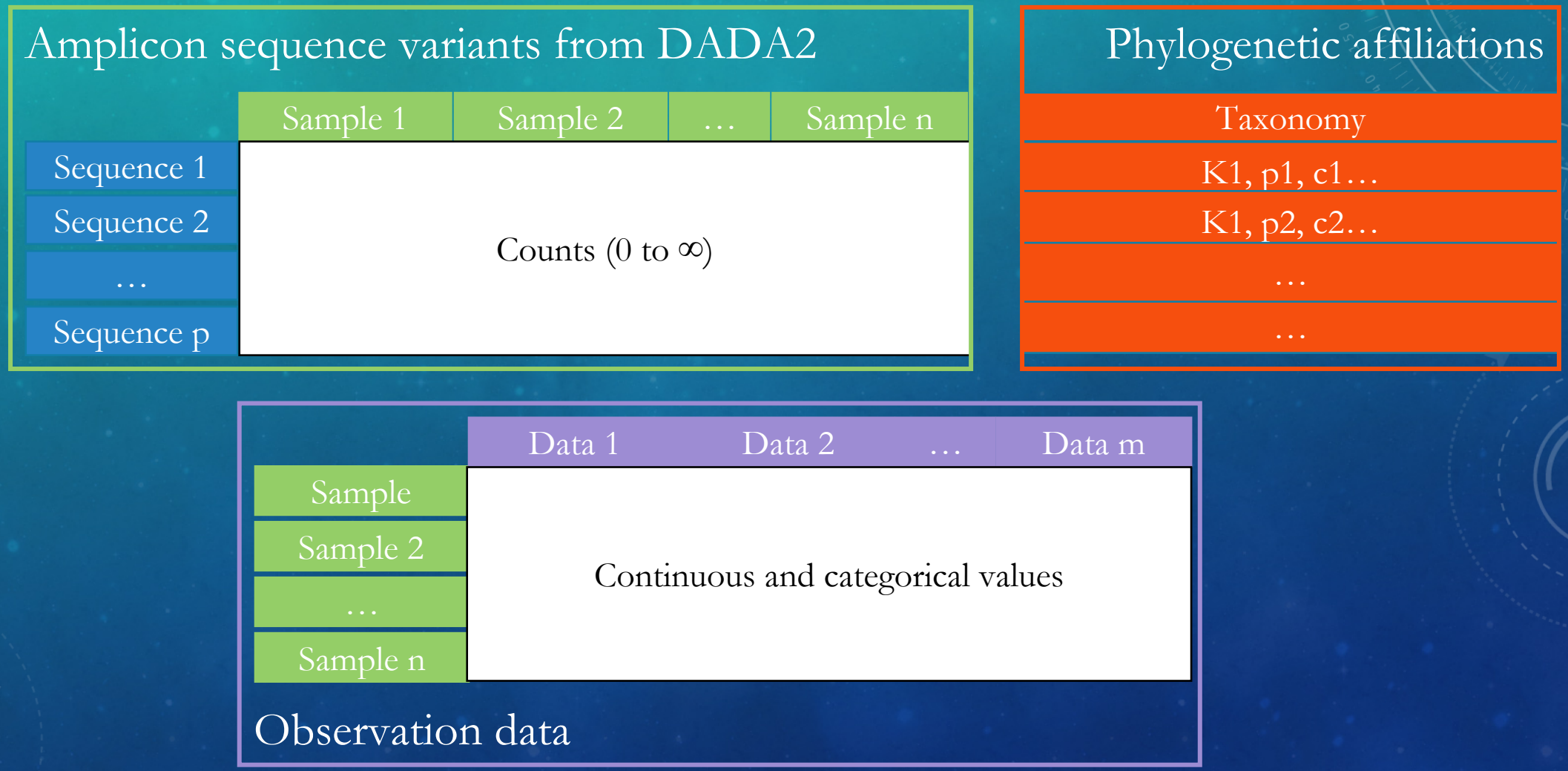
A WORKSHOP WITH CODE

OUR MISSION

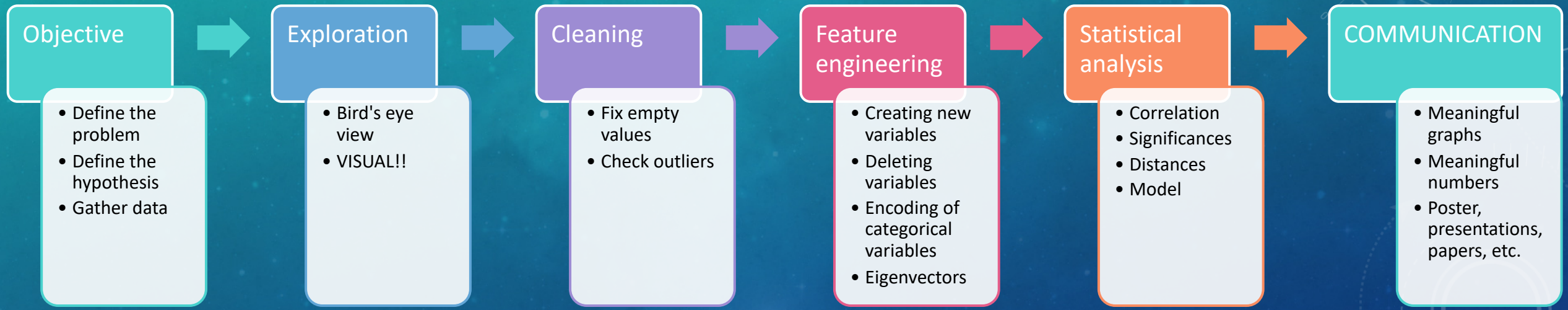
- Untangle complex system interactions
- Interactions are bidirectional
- There is always a three-way (at least) influence



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

A GENERIC DATA SCIENCE WORKFLOW



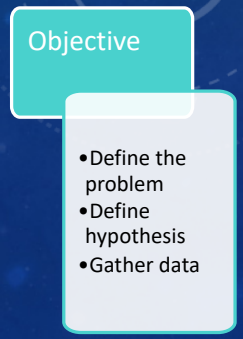
A GENERIC DATA SCIENCE WORKFLOW

OBJECTIVE

BIOMICS project

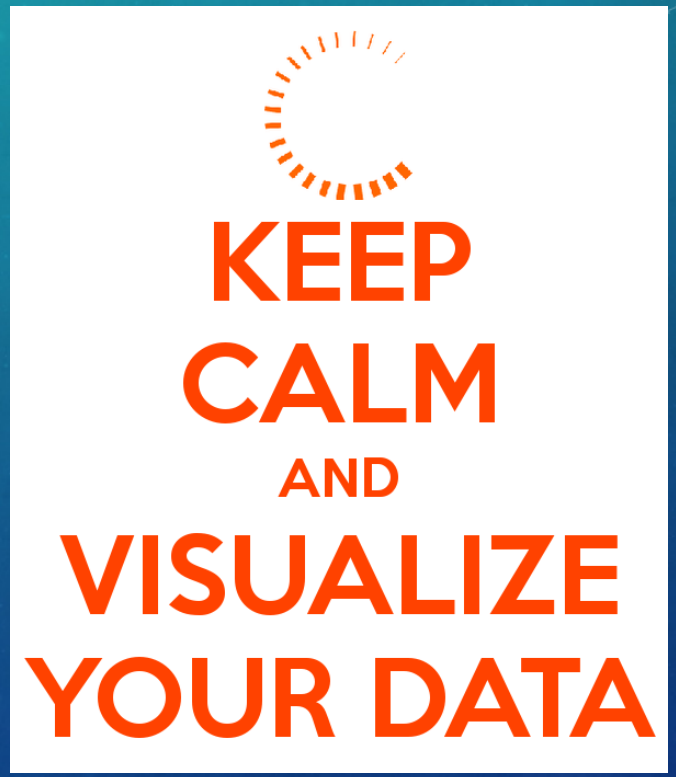
- Tomato plants in the greenhouse treated with 4 different biostimulant products containing different types of ”beneficial microbes”
  - Endomaxx, Inocucor, Pathway, (Earthcare with) Sumagrow
- Sacrificial sampling (tot 120 samples)
  - 4 time points (0, 3, 6, 10 weeks)
  - 5 treatments
  - 6 replicate per treatment

- HYPOTHESIS:
  - Biostimulant products have beneficial effects on the plants through impacts on the microbiome
  - ”Beneficial microbes” in the products survive and/or thrive in the soil
- GATHERED DATA
  - Leaf and root area and dry weight
  - Chlorophyll content of the leaves (SPAD)
  - Soil microbiota (bulk and rhizosphere not differentiated) -> 16S and ITS rRNA genes



EXPLORATION

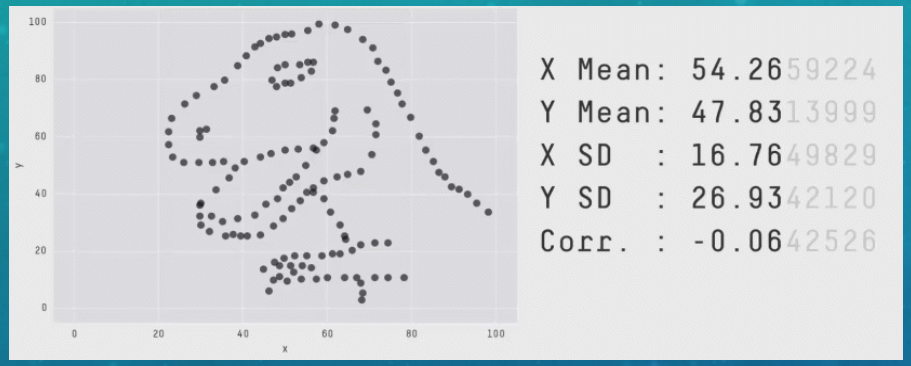
- Boxplots or violin plots for numeric variables
- Preliminary analysis of distributions

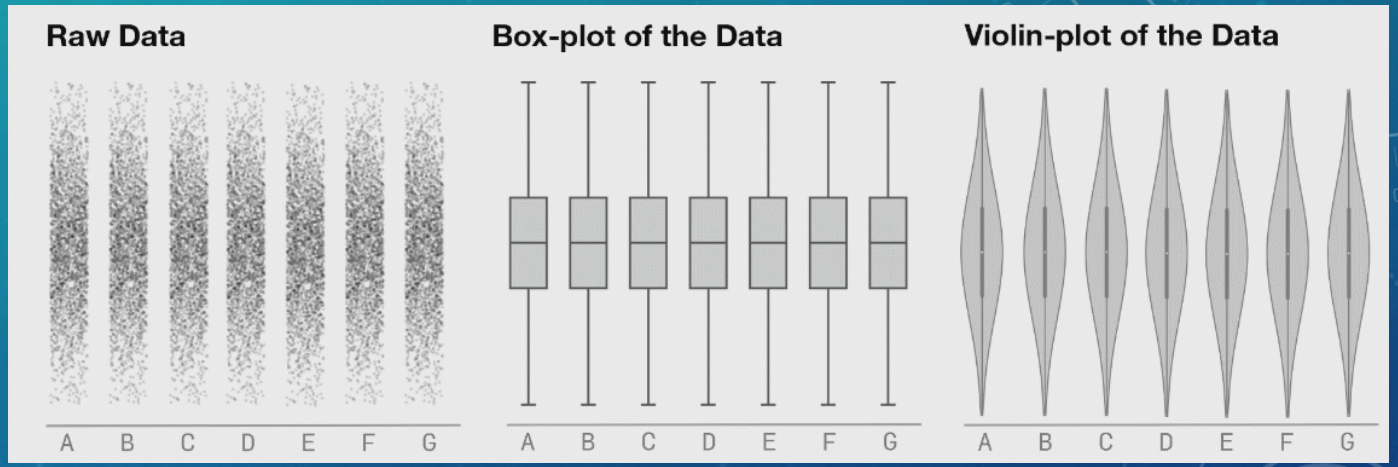


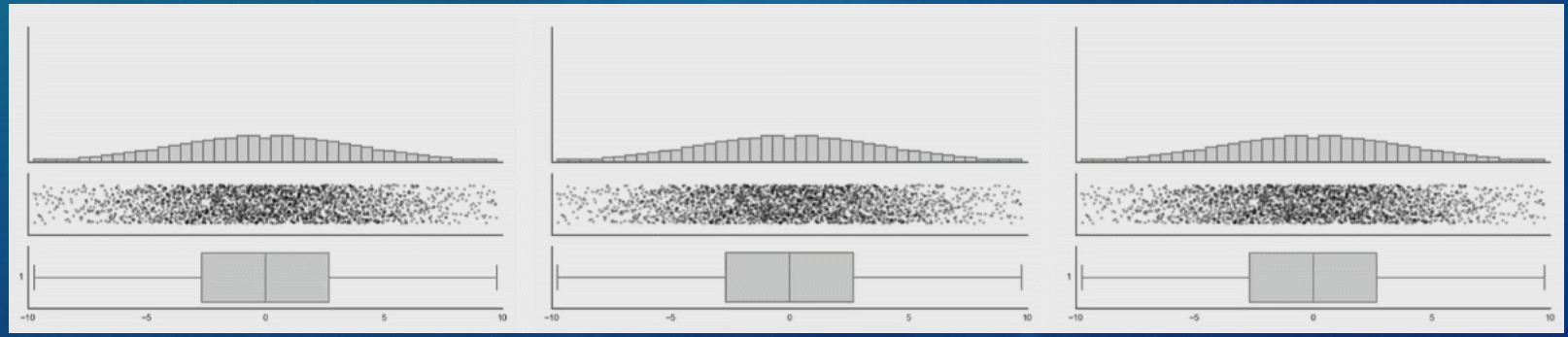
Exploration

- Bird's eye view of your data
- VISUAL!!!

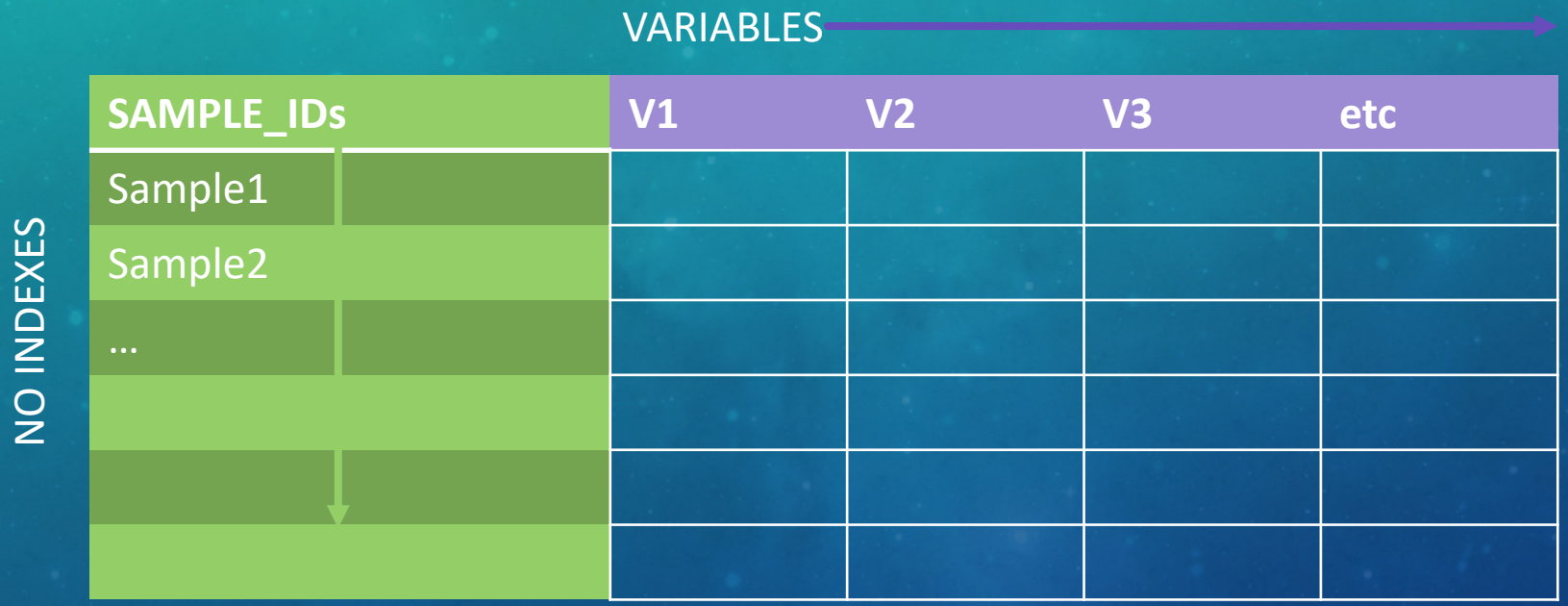
WHY IS IT IMPORTANT?







IMPORTANT!!

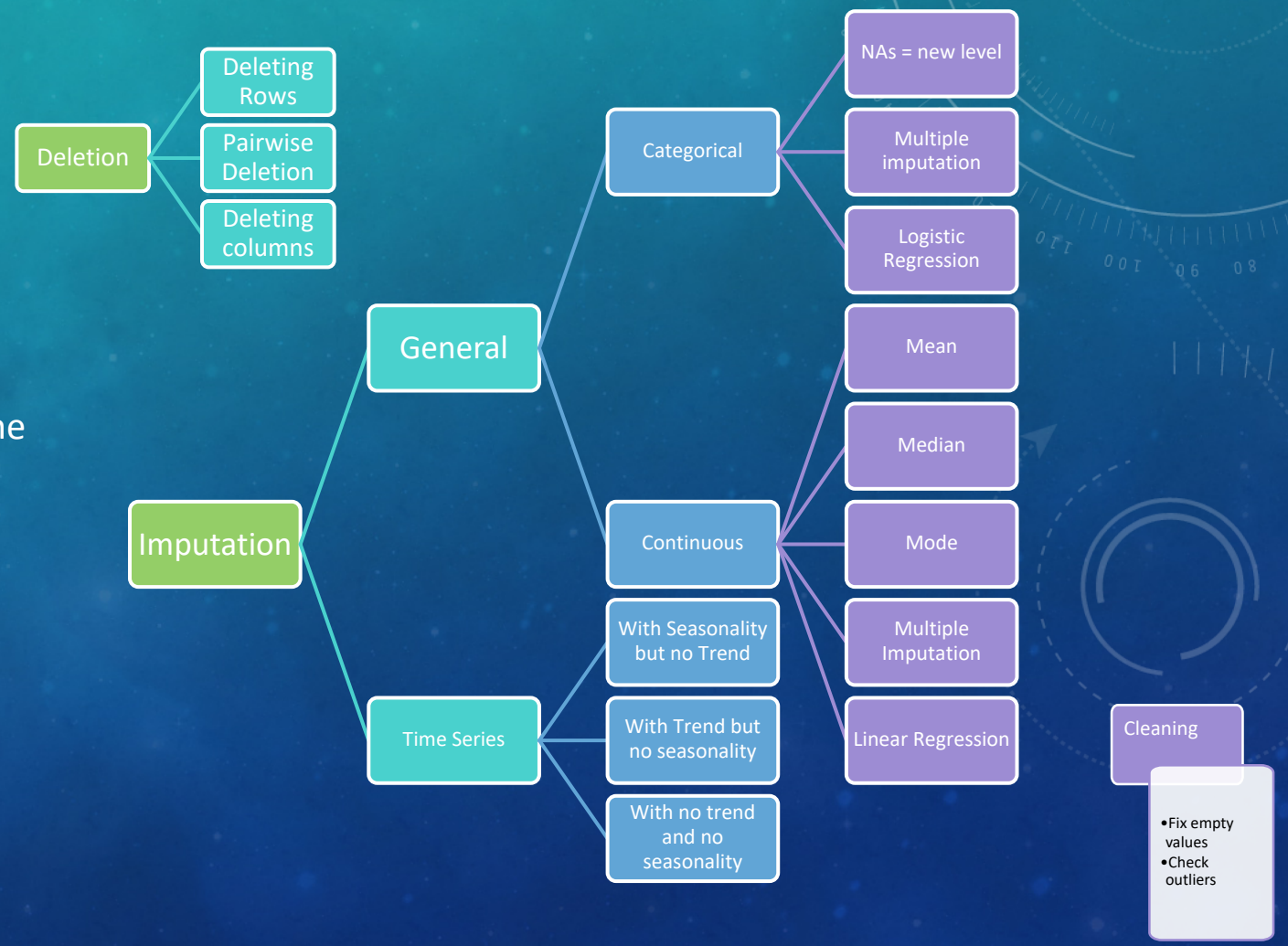


Exploration

- Bird's eye view of your data
- VISUAL!!!

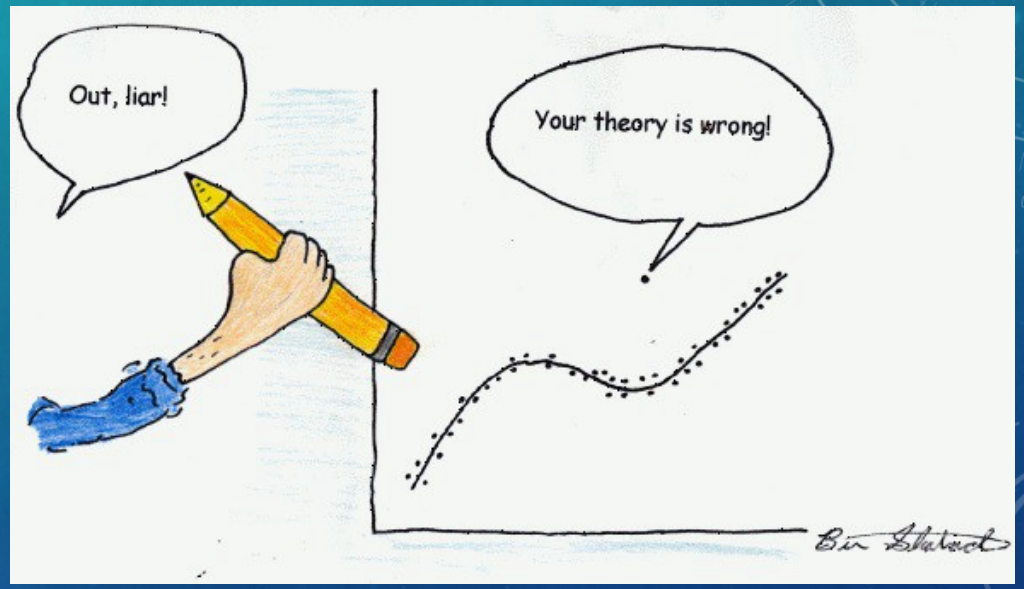
THE PROBLEM OF MISSING DATA

- Empty values (NAs or NaNs)
- NOT ZEROS!! ZERO IS A VALUE!!
- Can have three reasons:
  - *MAR* (missing at random): not related to the origin but related to the observation
  - *MCAR* (completely at random): not related to the sampling nor the observation
  - *MNAR* (not at random): depending on some variables (i.e. rich people do not share their income data)
- NO SINGLE SOLUTION!



OUTLIERS

- Data entry errors (human errors)
- Measurement errors (instrument errors)
- Experimental errors (data extraction or experiment planning/executing errors)
- Intentional (dummy outliers made to test detection methods)
- Data processing errors (data manipulation or data set unintended mutations)
- Sampling errors (extracting or mixing data from wrong or various sources)
- Natural (not an error, novelties in data)
- NO SINGLE SOLUTION!!



Cleaning

- Fix empty values
- Check outliers

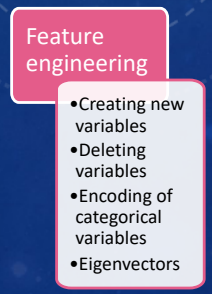
BASICS OF FEATURE ENGINEERING

"Coming up with features is difficult, time-consuming, requires expert knowledge. 'Applied machine learning' is basically feature engineering."

Prof. Andrew Ng

"Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data."

Dr. Jason Brownlee



WHAT CAN WE DO

CATEGORICAL VARIABLES

- Combine variables
- Aggregate variables into superior categories
- Encoding variables

State
Florida
Colorado
Colorado
Washington
Florida
Washington
Florida
Colorado

→

Florida	Washington
1	0
0	0
0	0
1	1
1	0
0	1
1	0
0	0

CONTINUOUS VARIABLES

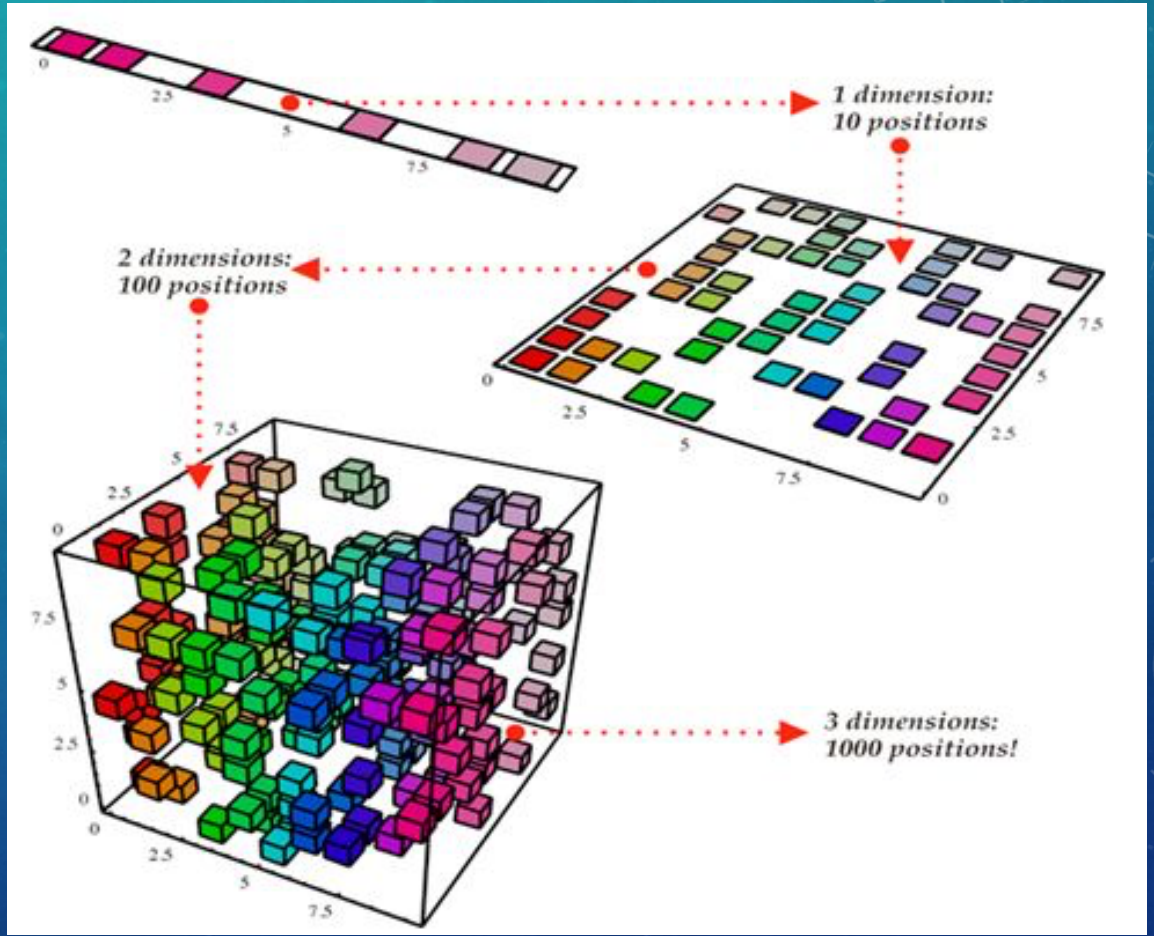
- Extract mean, quartiles, median etc.
- Mathematical combinations
- Transform
  - Normalization
  - Log-transform
  - Rank/Binarize
  - Bin
  - PCA, LDA, etc

Feature engineering

- Creating new variables
- Deleting variables
- Encoding of categorical variables
- Eigenvectors

PRINCIPAL COMPONENT ANALYSIS

- Reduces the number of features (!!!) while preserving as much as the randomness of the original datasets
- Many algorithms fail with too many features
- Dimensionality problem

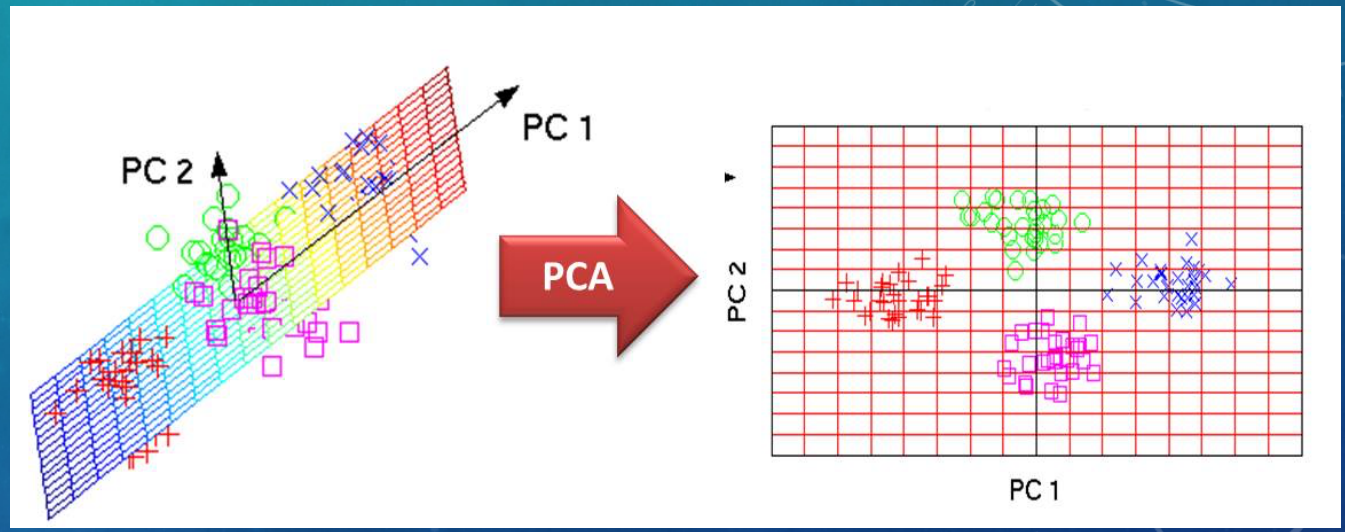


Feature engineering

- Creating new variables
- Deleting variables
- Encoding of categorical variables
- Eigenvectors

PCA: PRINCIPAL COMPONENT ANALYSIS

- Standardize (i.e. normalize) data
- Find the eigenvalues and eigenvectors of covariance matrices
- Transform the data by projecting them on the eigenvectors
- Plots the components

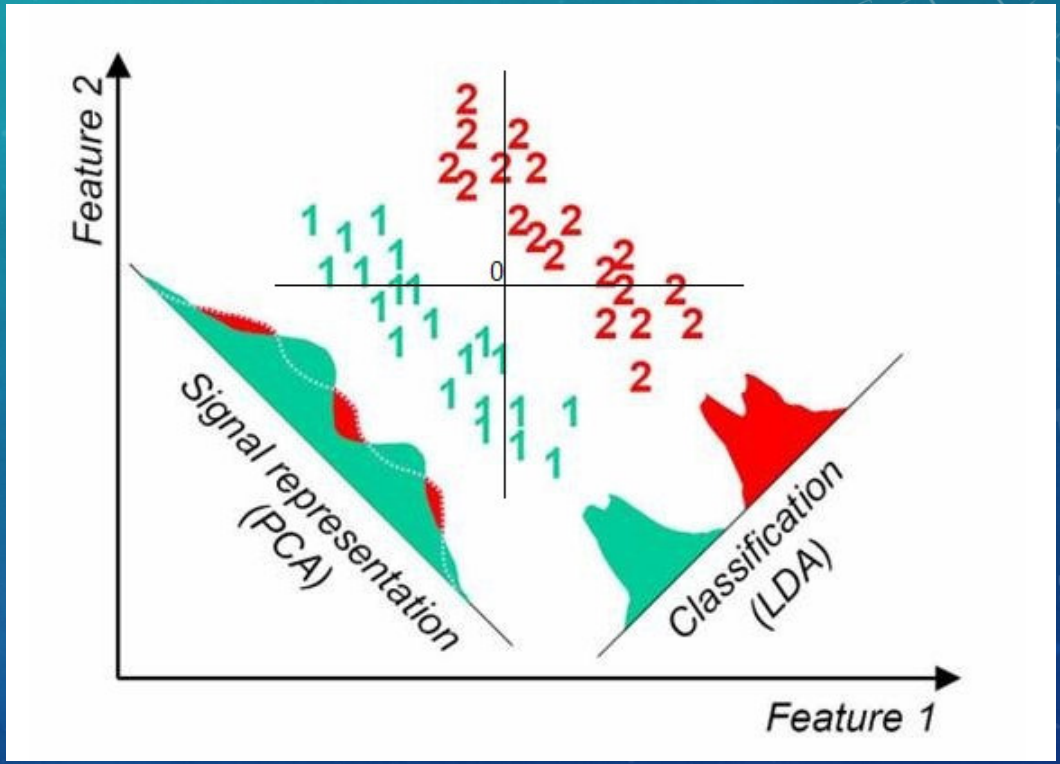


Feature engineering

- Creating new variables
- Deleting variables
- Encoding of categorical variables
- Eigenvectors

LDA: LINEAR DISCRIMINANT ANALYSIS

- CLASSIFICATION method
- Similar to PCA but:
  - Forces the components on determined variables instead of calculating them
  - Maximize distances between categories
- We will use it to combine categorical variables and continuous variables



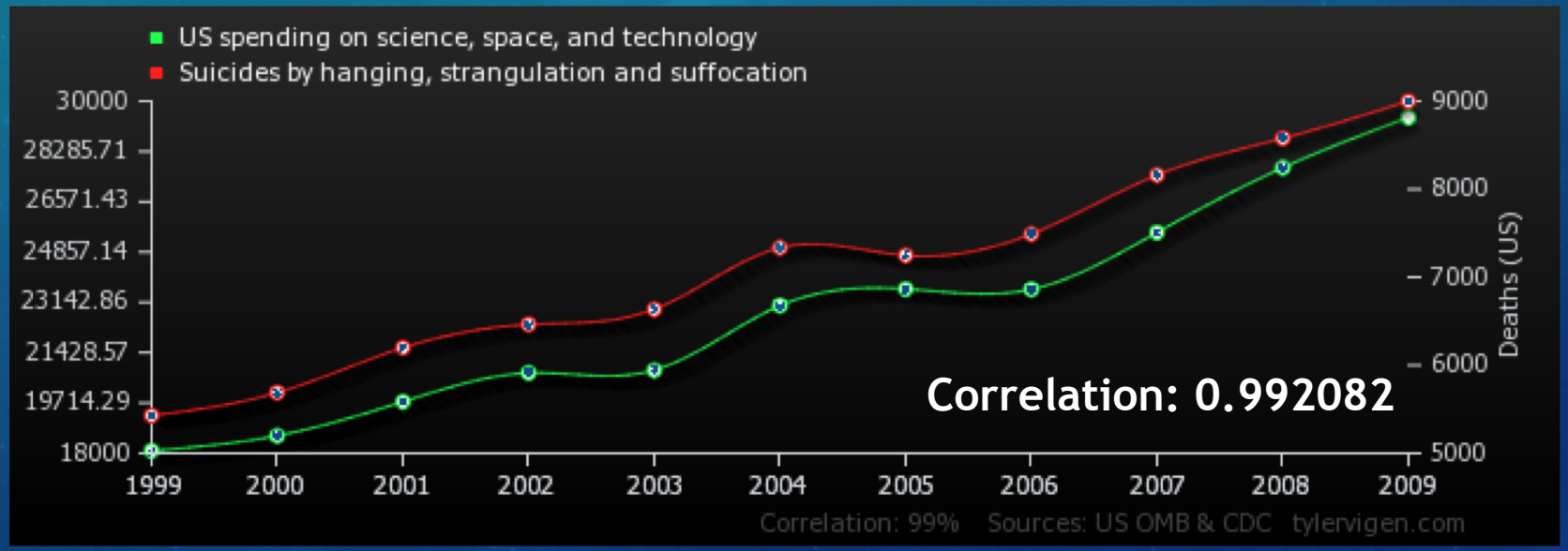
CORRELATIONS

WHY

- Find dependencies between variables in your data
- Strongly correlated variables break models
  - Determine which variable needs to be eliminated

WHY NOT

- NOT CAUSATION
  - May miss unknown (confounding) variables
  - Require normal distributions

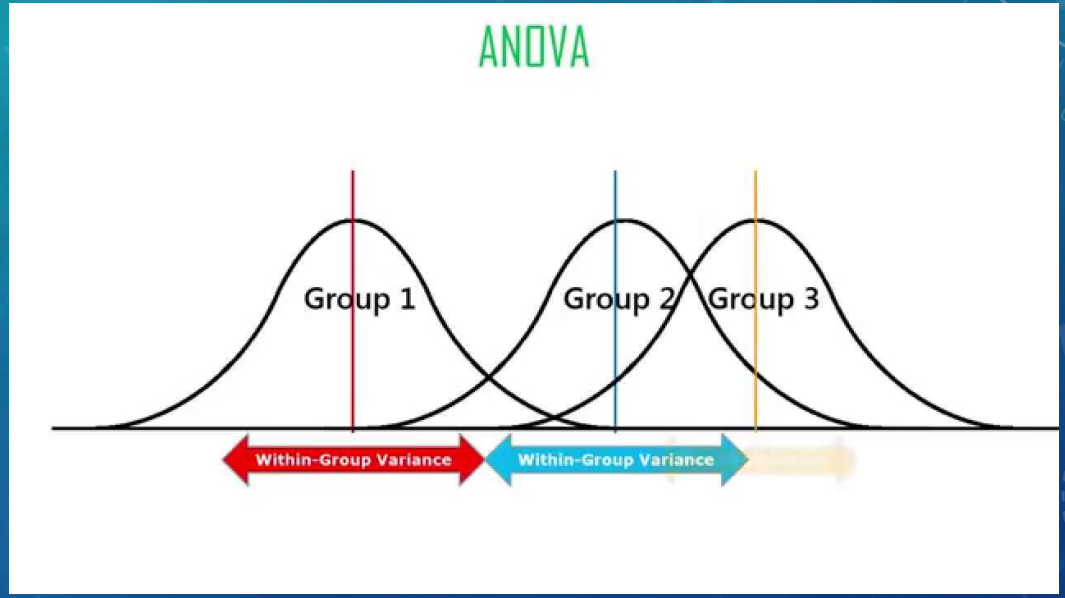


Statistical analysis

- Correlations
- Significances
- Distances
- Modeling

ANOVA

- Compares differences between groups
  - Normalizes distribution of each group into t-Student distribution
  - Applies two-way t-test for each pair:
    - i.e. is the overlapping region of the distribution greater or lesser than the significance value?
  - *F* = variability between groups/ within groups

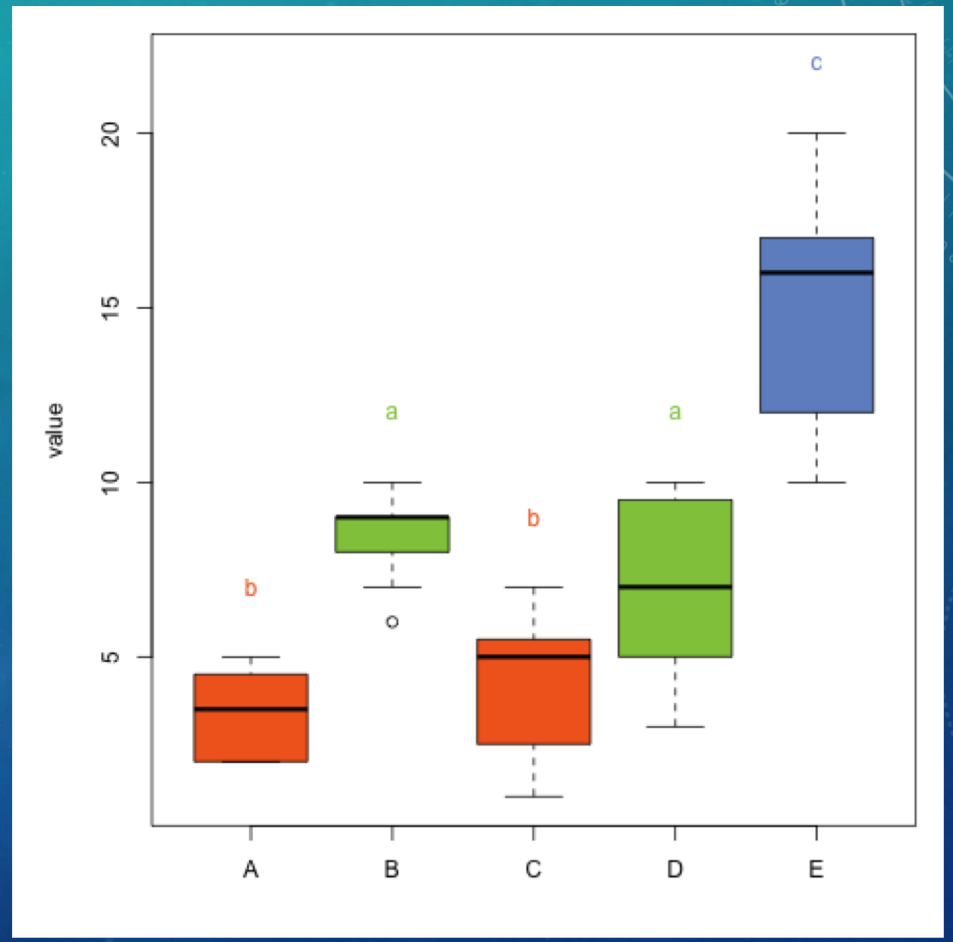


Statistical analysis

- Correlations
- Significances
- Distances
- Modeling

TUKEY'S HSD TEST

- ANOVA tells you IF there are differences. Tukey's tells you WHERE there are differences.
- More general than two-way ANOVA but less powerful



Statistical analysis

- Correlations
- Significances
- Distances
- Modeling

HOW ARE WE DOING?

A

B

C

CODE