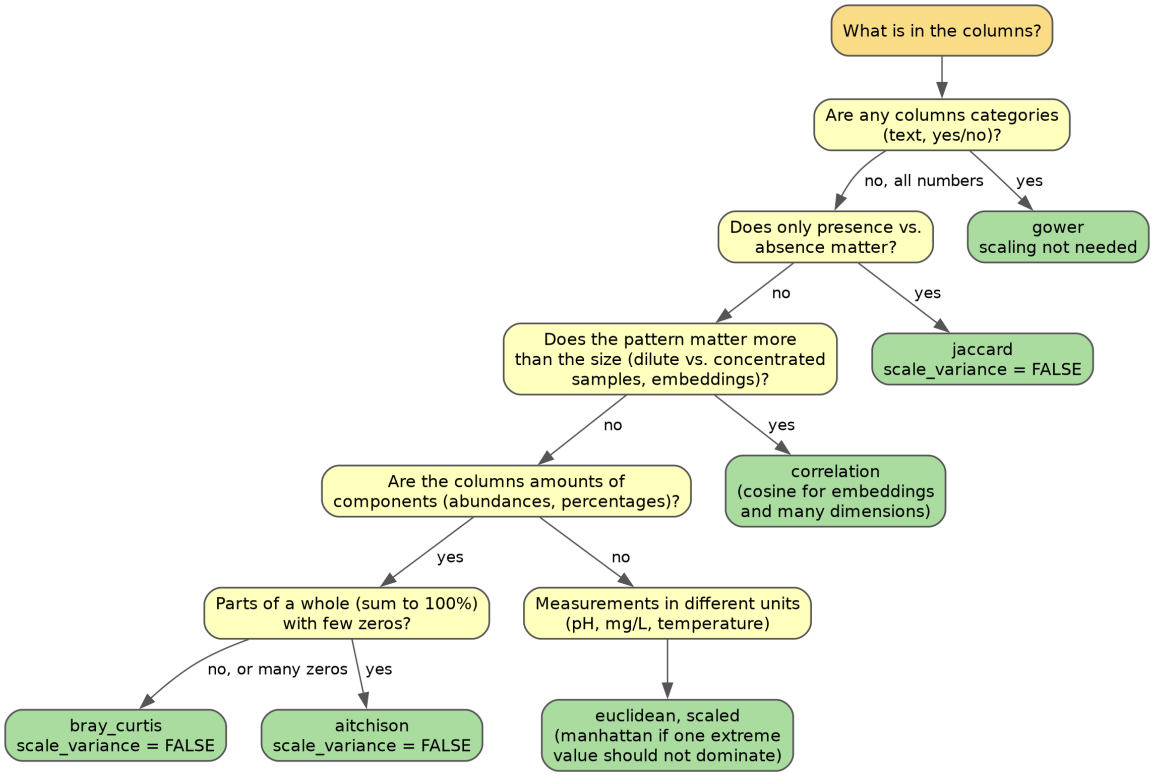
distances and scaling: summary

Integrated Bioanalytical Chemistry · thebustalab.github.io/integrated_bioanalytics



Two decisions go into every distance matrix. **Scaling decides which variables count**: scale measurements that are in different units so that no single variable dominates, and leave raw amounts unscaled when you want distances to reflect potential differences in the scales that your variables occupy. **The distance metric decides what "similar" means**: the same size and values, the same pattern, the same composition, the same things present, or the same mix of measurements and categories. The flowchart and the table below summarize the choices.

`distance_method`	Similar means…	Typical data	Scale first?
`"euclidean"`	close values in every variable, with large single differences counting heavily	measurements in different units (pH, mg/L, temperature)	yes, when units differ
`"manhattan"`	close values in every variable, with one extreme difference counting less	as euclidean, when outliers or spikes are a concern	yes, when units differ
`"correlation"`	the same pattern across the variables, regardless of size	profiles at different concentrations or totals	optional
`"cosine"`	pointing in the same direction, regardless of length	embeddings and other high-dimensional data	optional
`"bray_curtis"`	sharing the same amounts of the same components	abundances and percentages, especially with many zeros	no
`"aitchison"`	the same proportions between components	compositions summing to 100% with few zeros	no
`"jaccard"`	containing the same things, amounts ignored	presence/absence, detected or not	no
`"gower"`	matching across a mix of numbers and categories	tables with categorical columns	handled internally