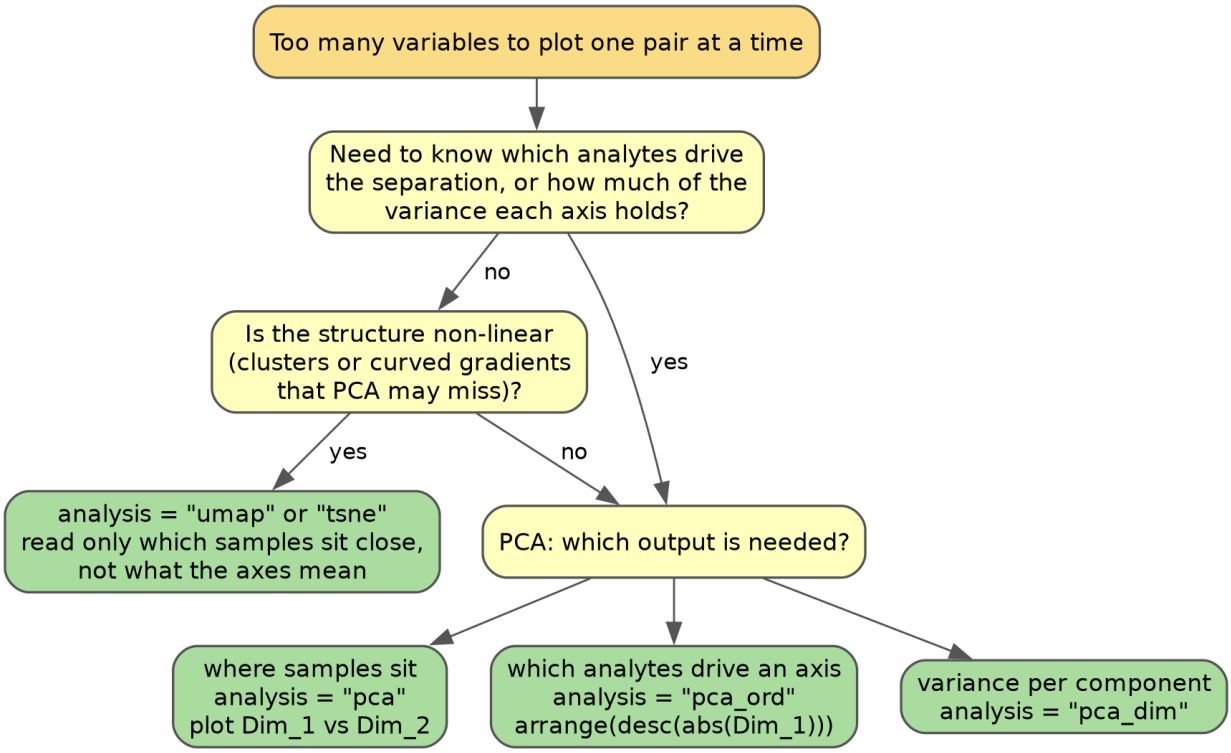
dimensional reduction: summary

Integrated Bioanalytical Chemistry · thebustalab.github.io/integrated_bioanalytics



Dimensionality reduction places samples with many measurements on a two-dimensional plot, where nearby samples have similar profiles. **PCA is the one to use when the axes need explaining**: each principal component is a combination of the analytes, the ordination shows which analytes drive each one (ranked by the size of the loading, ignoring its sign), and the scree plot shows how much of the variance each principal component describes. PCA scales every analyte first, and the ordination plot needs `coord_fixed()` so that distances in the two directions can be compared. **UMAP and t-SNE are non-linear** and can reveal clusters or curved gradients that PCA misses, but their axes have no quantitative meaning and each run can come out a little differently, so these should only be used to interpret local structures, not data set-wide layout.

`analysis` or function	What it returns or does	What to set or watch
`"pca"`	scores: one row per sample, with `Dim_1`, `Dim_2`	`columns_w_values_for_single_analyte`, `columns_w_sample_ID_info`, then `coord_fixed()` on the plot
`"pca_ord"`	loadings: one row per analyte	rank with `arrange(desc(abs(Dim_1)))`, draw with `geom_segment()` and `geom_circle()`
`"pca_dim"`	`percent_variance_explained` for each `principal_component`	scree plot, or `filter(principal_component == 1)` for one value
biplot	scores and loadings on shared axes	`geom_point()` from `"pca"` plus `geom_segment()` from `"pca_ord"`
`pcaVisualizer()`	linked scores, ordination and heat map panels	`data` plus the two column arguments, desktop R only
`"umap"`	non-linear embedding	axes have no meaning, runs vary, does not run in browser cells
`"tsne"`	non-linear embedding that favors local clusters	remove samples with duplicate analyte values first, runs vary