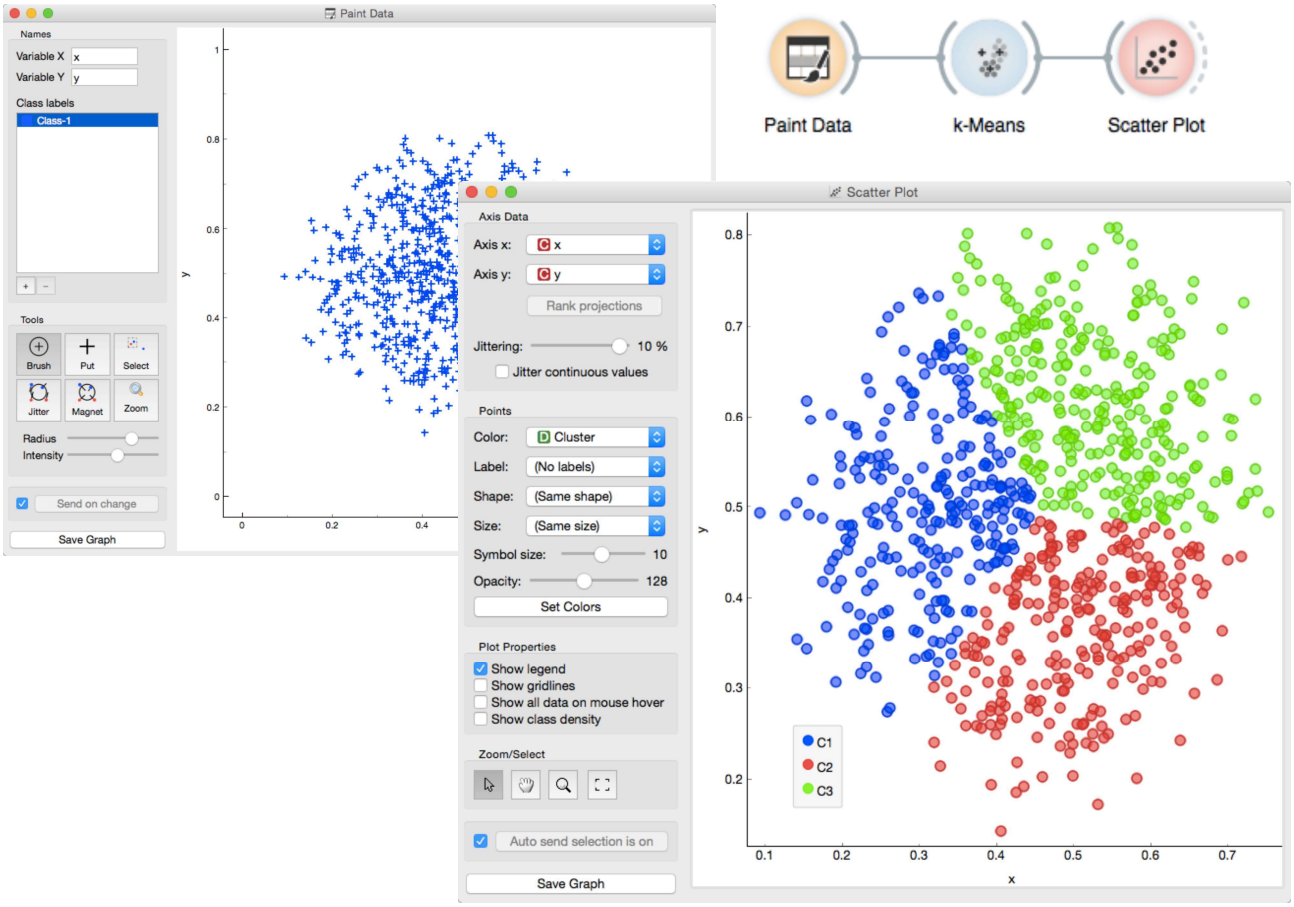
Lesson 28: Finding Clusters When There Are None

We saw how clustering can discover the subgroups in the data. The flip side of this is that algorithms like k-means will always find them even when they do not actually exist.



Playing with Paint Data and k-Means can be quite fun. Try painting the data where there are clusters, but k-means does not find them. Or, actually, finds the wrong ones. What kind of clusters are easy to find for k-means? Are these the kind of clusters we would actually find in real data sets?

It is difficult to verify whether the clusters we found are "real". Data mining methods like clustering can serve only as hints that can help forming new hypotheses, which must make biological sense and be verified on new, independent data. We cannot make conclusions based only on "discovering" clusters.