HOW DOES HIERARCHICAL CLUSTERING WORK?

Clustering things into groups is helpful for different kinds of design research work. When doing information architecture it can inform where a specific piece of information belongs in an overall hierarchy. When doing other kinds of analysis it can be helpful for creating high level groupings to summarize a set of item.

Using an algorithm to cluster will give you a way to check your intuition regarding the way you group things. They can produce dendrograms and similarity matrices which can be useful visualizations for summarizing and presenting your work.

GETTING SIMILARITY DATA

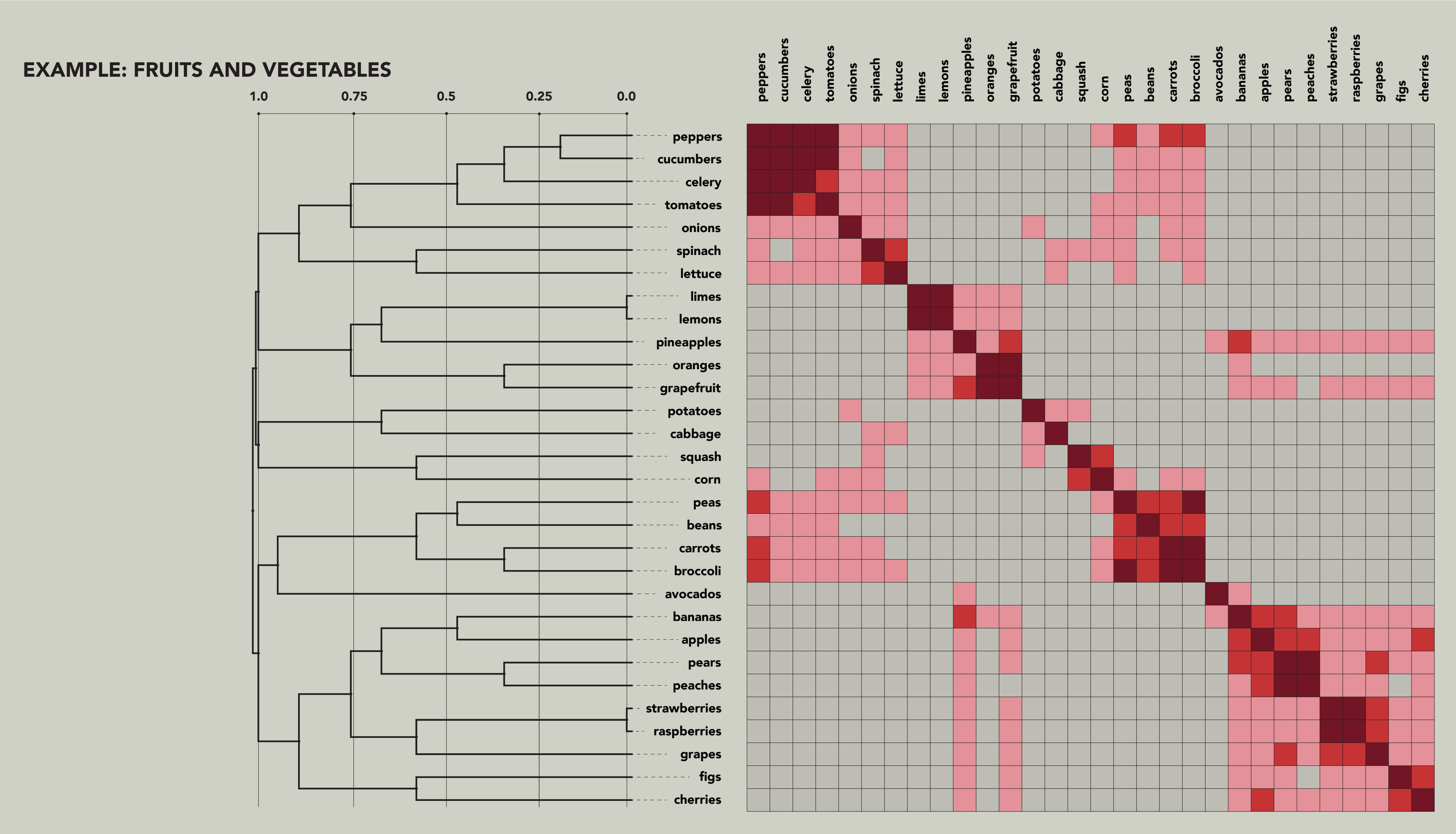
To start, you will need to collect data that shows how each possible pairing of elements in the set you're clustering compares to each other.

You can do pairwise comparisons for up to about 30 items manually. Label the rows and columns in a spreadsheet with the elements you want to cluster; so for 30 items, you'll have 30 rows and 30 columns. Then, in each square, decide how similar each pair of items are on a scale of 0.0 to 1.0. (It helps to use a limited number of values, like 0.0, 0.33, 0.66, and 1.0.) In my example involving fruits and vegetables apples and applies are perfectly similar and score a 1.0, while apples and oranges score a 0.66, and apples and lettuce score a 0.0.

A card sort can help you cluster up to about 100 elements. Run an appropriate number of tests, and make lists containing the items in each group that participants created. Next calculate the Jaccard Index for each pair of items: this is the number of groups where both items appeared together divided by the number of groups where at least one item in the pair appeared. Record each score of 0.0 to 1.0 in the matrix. It helps to use a script to do the math, and more involved scripts can compare even larger sets.

"BOTTOM-UP" CLUSTERING

Once you have collected similarity data you will need to produce a chart called a dendrogram: it's the tree-shaped chart on the left in my fruits and vegetables example.



EXAMPLE: FRUITS AND VEGETABLES

Using clustering to sort a matrix of items, to identify groupings. See https://github.com/johnjung/planning_tools for source code.

To make the chart, start by placing each item in it's own group. If you are clustering a set of 30 items, you will start with 30 groups. Next, compare each item against each other, looking for the items with the highest similarity score. If there are multiple pairs of items with equal similarity just pick any pair for this round.

Repeat this process, looking for the next most similar groups. To calculate the similarity for groups with multiple items you will need to choose a "linkage method". Compare each possible pair of elements in each set. For single-linkage clustering, the similarity of the groups is equal to the lowest similarity of every possible pairing of items in each group. For complete-linkage clustering, group similarity is the highest similarity score of each possible pairing of items. You might want to experiment with both of these options because they produce different kinds of clusters. Repeat the process, comparing bigger and bigger groups, until all groups have been consolidated into one.

SORTING THE MATRIX

If you take the order of elements in the dendrogram you can sort the similarity matrix. You should see clusters emerge in the resulting chart. The dendrogram contains the Jaccard Index where each cluster and sub-cluster formed, and the similarity matrix will show primary clusters along the diagonal from top left to lower right, along with secondary clusters in other places in the chart.