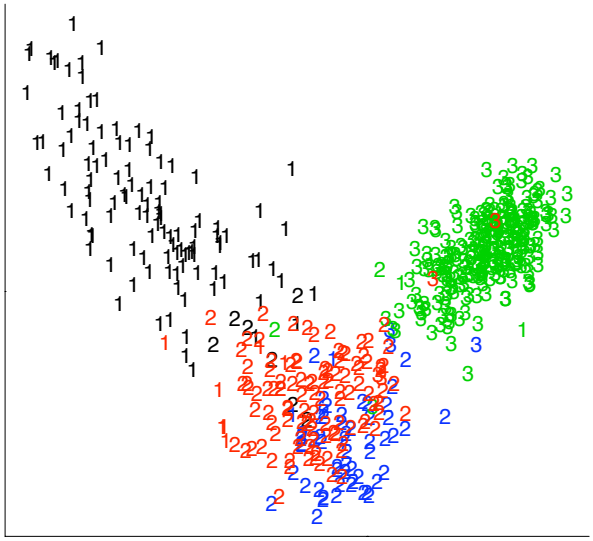

Depicted (top) is a two dimensional representation of a corpus of 579 Science News documents in four classes. The color code is: black for astronomy; red for physics; blue for mathematics; and green for medicine. This representation is obtained via a similarity measurement between documents based on shared important words – two documents that share many important words with each other are considered more similar than are two documents that share few important words.

The symbols represent groups obtained via clustering – performed without knowledge of the class labels. Documents from physics and mathematics are highly intermingled in cluster 2, while the astronomy and medicine clusters are mostly coherent. In addition, physics melts into astronomy (which seems reasonable) while the delineation of medicine is more crisp.

The bottom figures depict just cluster 2. The bottom left is simply an enlarged depiction from the top figure, while the bottom right depicts the results of a re-evaluation of which words are important in the context of the cluster 2 documents only. Because the similarity between documents is evaluated in context (an approach we dub *corpus-dependent feature extraction*) a very different collection of words is deemed important in the former case (considering all four classes of documents) than in the latter case (the predominantly physics and mathematics context) – and in fact the subsequent focused analysis uncovers a much clearer separation between physics and mathematics. This process of iteratively focusing on subsets of the data and re-evaluating feature importance in these focused contexts is called *iterative denoising*.