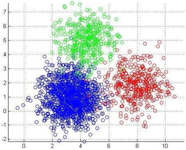
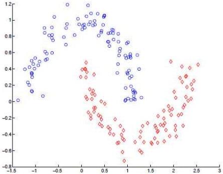
Notions of Similarity

- Choice of the **similarity measure** is very important for clustering
- Similarity is inversely related to distance
- Different ways exist to measure distances. Some examples:
  - Euclidean distance: $d(\mathbf{x},\mathbf{z}) = ||\mathbf{x}-\mathbf{z}|| = \sqrt{\sum_{d=1}^{D}(x_d - z_d)^2}$
  - Manhattan distance: $d(\mathbf{x},\mathbf{z}) = \sum_{d=1}^{D}|x_d - z_d|$
  - Kernelized (non-linear) distance: $d(\mathbf{x},\mathbf{z}) = ||\phi(\mathbf{x}) - \phi(\mathbf{z})||$





  - For the left figure above, Euclidean distance may be reasonable
  - For the right figure above, kernelized distance seems more reasonable