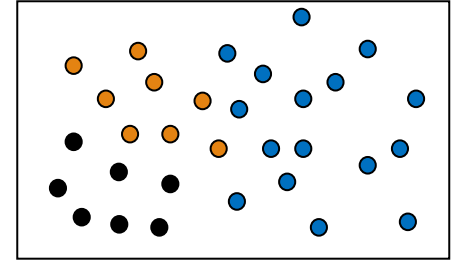


The background is a complex collage of abstract elements. It features a grid of small grey plus signs, a network of red lines connecting green dots, and a dense field of small, multi-colored dots. A large, white, angular shape is positioned in the center, serving as a backdrop for the title.

Clustering Tendency

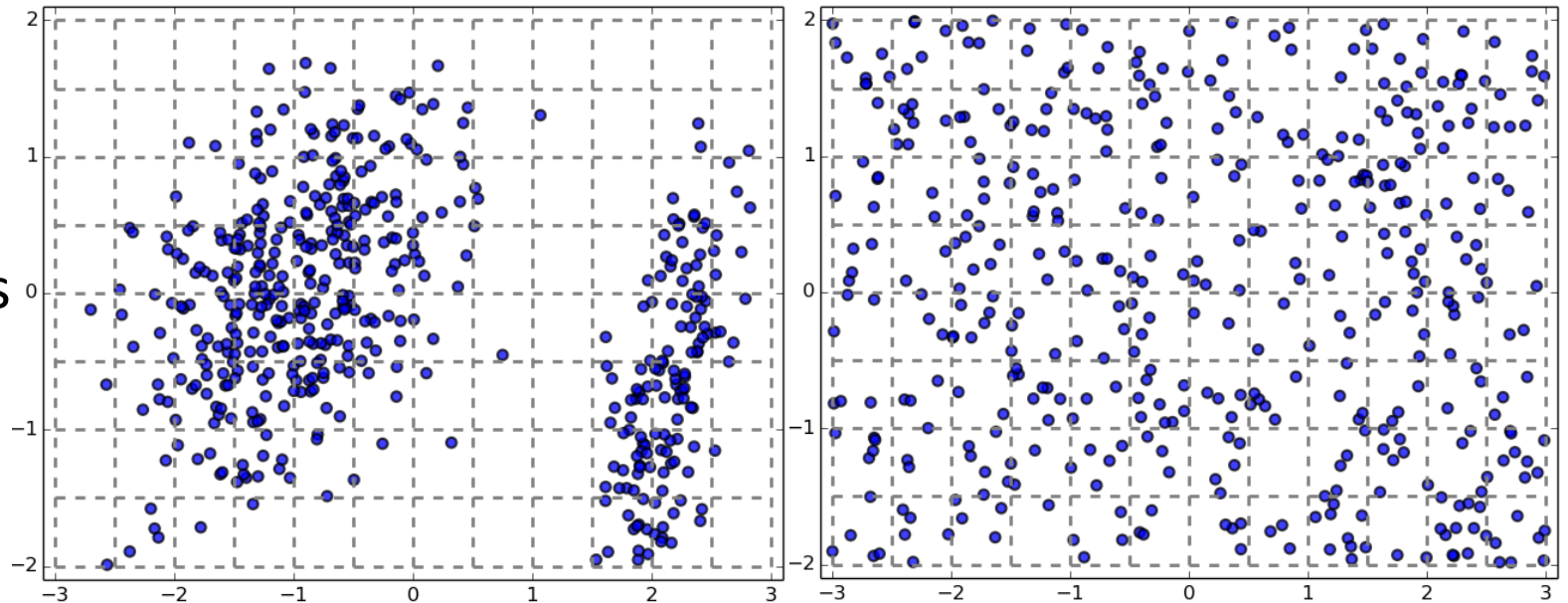
Clustering Tendency: Whether the Data Contains Inherent Grouping Structure

- ❑ Assessing the **suitability of clustering**
 - ❑ (i.e., whether the data has any inherent grouping structure)
- ❑ Determining ***clustering tendency*** or ***clusterability***
 - ❑ **A hard task** because there are so many different definitions of clusters
 - ❑ E.g., partitioning, hierarchical, density-based, graph-based, etc.
 - ❑ Even fixing cluster type, still hard to define an appropriate null model for a data set
- ❑ Still, there are some **clusterability assessment methods**, such as
 - ❑ **Spatial histogram**: Contrast the histogram of the data with that generated from random samples To be covered here
 - ❑ **Distance distribution**: Compare the pairwise point distance from the data with those from the randomly generated samples
 - ❑ **Hopkins Statistic**: A sparse sampling test for spatial randomness



Testing Clustering Tendency: A Spatial Histogram Approach

- ❑ **Spatial Histogram Approach:** Contrast the d -dimensional histogram of the input dataset D with the histogram generated from random samples
 - ❑ Dataset D is clusterable if the distributions of two histograms are rather different
- ❑ Method outline
 - ❑ Divide each dimension into equi-width bins, count how many points lie in each cells, and obtain the empirical joint probability mass function (EPMF)
 - ❑ Do the same for the randomly sampled data
 - ❑ Compute how much they differ using the *Kullback-Leibler (KL) divergence* value



Recommended Readings

- ❑ M. J. Zaki and W. Meira, Jr.. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, 2014
- ❑ L. Hubert and P. Arabie. Comparing Partitions. *Journal of Classification*, 2:193–218, 1985
- ❑ A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Printice Hall, 1988
- ❑ M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On Clustering Validation Techniques. *Journal of Intelligent Info. Systems*, 17(2-3):107–145, 2001
- ❑ J. Han, M. Kamber, and J. Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011
- ❑ H. Xiong and Z. Li. Clustering Validation Measures. in (Chapter 23) C. Aggarwal and C. K. Reddy (eds.), Data Clustering: Algorithms and Applications. CRC Press, 2014