



# Lecture 7. Methods for Clustering Validation

# Lecture 7. Clustering Validation

---

- ❑ Clustering Validation: Basic Concepts
- ❑ Clustering Evaluation: Measuring Clustering Quality
- ❑ External Measures for Clustering Validation
  - ❑ I: Matching-Based Measures
  - ❑ II: Entropy-Based Measures
  - ❑ III: Pairwise Measures
- ❑ Internal Measures for Clustering Validation
- ❑ Relative Measures
- ❑ Cluster Stability
- ❑ Clustering Tendency
- ❑ Summary



The background of the slide is a complex, abstract composition. It features a dark, muted purple or brownish background. Overlaid on this are several geometric and data-like elements: a network of thin, light-colored lines forming a mesh or web; numerous small, green and blue dots scattered across the space; and a series of faint, light-colored plus signs arranged in a grid-like pattern. In the lower-left corner, there is a small, rectangular inset image showing a cluster of orange and red dots, possibly representing a specific data set or a visualization of a cluster.

# **Session 1: Clustering Validation: Basic Concepts**

# Clustering Validation and Assessment

---

- Major issues on clustering validation and assessment
  - **Clustering evaluation**
    - Evaluating the goodness of the clustering
  - **Clustering stability**
    - To understand the sensitivity of the clustering result to various algorithm parameters, e.g., # of clusters
  - **Clustering tendency**
    - Assess the suitability of clustering, i.e., whether the data has any inherent grouping structure





# **Session 2: Clustering Evaluation: Measuring Clustering Quality**

# Measuring Clustering Quality

---

- ❑ **Clustering Evaluation:** Evaluating the goodness of clustering results
  - ❑ No commonly recognized best suitable measure in practice
- ❑ **Three categorization of measures:** External, internal, and relative
  - ❑ **External:** Supervised, employ criteria not inherent to the dataset
    - ❑ Compare a clustering against prior or expert-specified knowledge (i.e., the ground truth) using certain clustering quality measure
  - ❑ **Internal:** Unsupervised, criteria derived from data itself
    - ❑ Evaluate the goodness of a clustering by considering how well the clusters are separated and how compact the clusters are, e.g., silhouette coefficient
  - ❑ **Relative:** Directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm



The background of the slide is a complex, abstract composition. It features a network of thin, light-colored lines forming a web-like structure. Overlaid on this are various data points and clusters. In the upper left, there's a grid of small, light-colored plus signs. In the lower left, there's a rectangular area with a pixelated, orange and brown pattern. The overall color palette is muted, with shades of brown, beige, and light blue.

# **Session 3: External Measures for Clustering Validation**

# Measuring Clustering Quality: External Methods

---

- ❑ Given the **ground truth**  $T$ ,  $Q(C, T)$  is the **quality measure** for a clustering  $C$
- ❑  $Q(C, T)$  is good if it satisfies the following **four** essential criteria
  - ❑ **Cluster homogeneity**
    - ❑ The purer, the better
  - ❑ **Cluster completeness**
    - ❑ Assign objects belonging to the same category in the ground truth to the same cluster
  - ❑ **Rag bag better than alien**
    - ❑ Putting a heterogeneous object into a pure cluster should be penalized more than putting it into a *rag bag* (i.e., “miscellaneous” or “other” category)
  - ❑ **Small cluster preservation**
    - ❑ Splitting a small category into pieces is more harmful than splitting a large category into pieces



# Commonly Used External Measures

## ❑ Matching-based measures (To be covered)

- ❑ Purity, maximum matching, F-measure

## ❑ Entropy-Based Measures

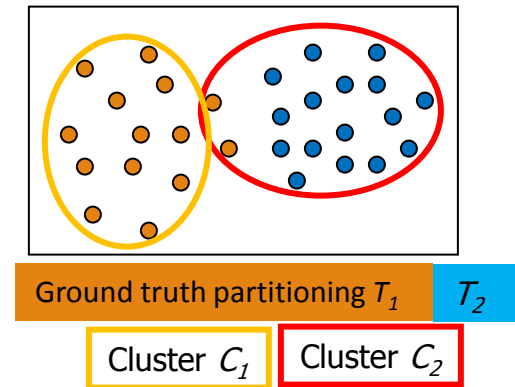
- ❑ Conditional entropy (To be covered)
- ❑ Normalized mutual information (NMI) (To be covered)
- ❑ Variation of information

## ❑ Pairwise measures (To be covered)

- ❑ Four possibilities: True positive (TP), FN, FP, TN
- ❑ Jaccard coefficient, Rand statistic, Fowlkes-Mallow measure

## ❑ Correlation measures

- ❑ Discretized Huber static, normalized discretized Huber static



The background of the slide is a complex, abstract composition. It features a network graph with numerous green nodes and red edges, overlaid on a light blue and white geometric pattern. A prominent white diagonal band runs across the center, serving as a backdrop for the title. In the bottom-left corner, there is a small inset image showing a cluster of orange and red dots on a white background, with a horizontal bar chart overlaid on it.

# **Session 4: External Measures I: Matching-Based Measures**

# Matching-Based Measures (I): Purity vs. Maximum Matching

❑ **Purity:** Quantifies the extent that cluster  $C_i$  contains points only from one (ground truth) partition: 
$$purity_i = \frac{1}{n_i} \max_{j=1}^k \{n_{ij}\}$$

❑ Total purity of clustering  $C$ :

$$purity = \sum_{i=1}^r \frac{n_i}{n} purity_i = \frac{1}{n} \sum_{i=1}^r \max_{j=1}^k \{n_{ij}\}$$

❑ Perfect clustering if  $purity = 1$  and  $r = k$  (the number of clusters obtained is the same as that in the ground truth)

❑ Ex. 1 (green or orange):  $purity_1 = 30/50$ ;  $purity_2 = 20/25$ ;  $purity_3 = 25/25$ ;  $purity = (30 + 20 + 25)/100 = 0.75$

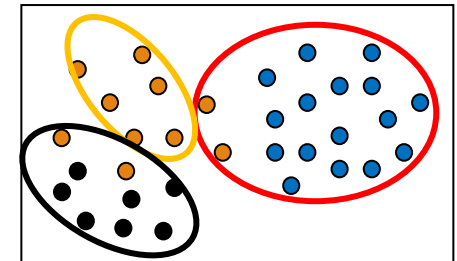
❑ Two clusters may share the same majority partition

❑ **Maximum matching:** Only one cluster can match one partition

❑ Match: Pairwise matching, weight  $w(e_{ij}) = n_{ij}$   $w(M) = \sum_{e \in M} w(e)$

❑ Maximum weight matching:  $match = \arg \max_M \left\{ \frac{w(M)}{n} \right\}$

❑ Ex2. (green)  $match = purity = 0.75$ ; (orange)  $match = 0.65 > 0.6$



Ground Truth $T_1$	$T_2$	$T_3$
Cluster $C_1$	$C_2$	$C_3$

$C \backslash T$	$T_1$	$T_2$	$T_3$	Sum
$C_1$	0	20	30	50
$C_2$	0	20	5	25
$C_3$	25	0	0	25
$m_j$	25	40	35	100

$C \backslash T$	$T_1$	$T_2$	$T_3$	Sum
$C_1$	0	30	20	50
$C_2$	0	20	5	25
$C_3$	25	0	0	25
$m_j$	25	50	25	100



# Matching-Based Measures (II): F-Measure

- Precision:** The fraction of points in  $C_i$  from the majority partition  $T_{j_i}$  (i.e., the same as purity), where  $j_i$  is the partition that contains the maximum # of points from  $C_i$

- Ex. For the green table

$$prec_1 = 30/50; prec_2 = 20/25; prec_3 = 25/25$$

- Recall:** The fraction of point in partition  $T_{j_i}$  shared in common with cluster  $C_i$ , where  $m_{j_i} = |T_{j_i}|$

- Ex. For the green table

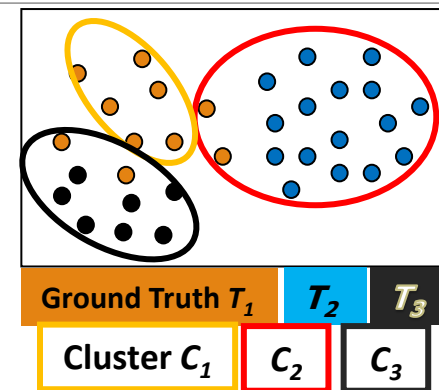
$$recall_1 = 30/35; recall_2 = 20/40; recall_3 = 25/25$$

- F-measure** for  $C_i$ : The harmonic means of  $prec_i$  and  $recall_i$ :  $F_i = \frac{2n_{ij_i}}{n_i + m_{j_i}}$

- F-measure** for clustering  $C$ : average of all clusters:  $F = \frac{1}{r} \sum_{i=1}^r F_i$

- Ex. For the green table

$$F_1 = 60/85; F_2 = 40/65; F_3 = 1; F = 0.774$$



$C \backslash T$	$T_1$	$T_2$	$T_3$	Sum
$C_1$	0	20	30	50
$C_2$	0	20	5	25
$C_3$	25	0	0	25
$m_j$	25	40	35	100

The background of the slide is a complex, abstract composition. It features a dark, muted purple or brownish background. Overlaid on this are several geometric and data-like elements: a network of thin, light-colored lines forming a mesh or web; numerous small, colored dots (green, blue, yellow) scattered across the space; and a grid of small white plus signs. In the upper left, there's a horizontal band with some faint, stylized text or symbols. In the lower left, there's a rectangular inset showing a cluster of orange and red dots, possibly representing a specific data set or a visualization of a concept like entropy.

# **Session 5: External Measures II: Entropy-Based Measures**

# Entropy-Based Measures (I): Conditional Entropy

□ Entropy of clustering  $\mathcal{C}$ :  $H(\mathcal{C}) = - \sum_{i=1}^r p_{C_i} \log p_{C_i}$   $p_{C_i} = \frac{n_i}{n}$  (i.e., the probability of cluster  $C_i$ )

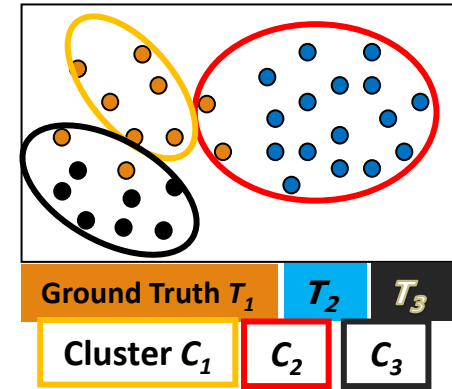
□ Entropy of partitioning  $\mathcal{T}$ :  $H(\mathcal{T}) = - \sum_{j=1}^k p_{T_j} \log p_{T_j}$

□ Entropy of  $\mathcal{T}$  with respect to cluster  $C_i$ :  $H(\mathcal{T}|C_i) = - \sum_{j=1}^k \left(\frac{n_{ij}}{n_i}\right) \log \left(\frac{n_{ij}}{n_i}\right)$

□ Conditional entropy of  $\mathcal{T}$  with respect to clustering  $\mathcal{C}$ :  $H(\mathcal{T}|\mathcal{C}) = - \sum_{i=1}^r \left(\frac{n_i}{n}\right) H(\mathcal{T}|C_i) = - \sum_{i=1}^r \sum_{j=1}^k p_{ij} \log \left(\frac{p_{ij}}{p_{C_i}}\right)$

□ The more a cluster's members are split into different partitions, the higher the conditional entropy

□ For a perfect clustering, the conditional entropy value is 0, where the worst possible conditional entropy value is  $\log k$



$$\begin{aligned}
 H(\mathcal{T}|\mathcal{C}) &= - \sum_{i=1}^r \sum_{j=1}^k p_{ij} (\log p_{ij} - \log p_{C_i}) = - \sum_{i=1}^r \sum_{j=1}^k p_{ij} \log p_{ij} + \sum_{i=1}^r (\log p_{C_i} \sum_{j=1}^k p_{ij}) \\
 &= - \sum_{i=1}^r \sum_{j=1}^k p_{ij} \log p_{ij} + \sum_{i=1}^r (p_{C_i} \log p_{C_i}) = H(\mathcal{C}, \mathcal{T}) - H(\mathcal{C})
 \end{aligned}$$



# Entropy-Based Measures (II): Normalized Mutual Information (NMI)

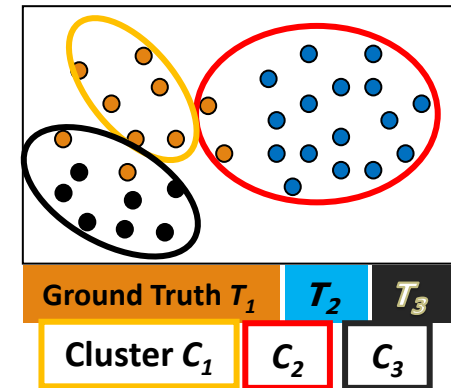
## □ Mutual information:


- Quantifies the amount of shared info between the clustering  $C$  and partitioning  $T$ 
$$I(C, T) = \sum_{i=1}^r \sum_{j=1}^k p_{ij} \log\left(\frac{p_{ij}}{p_{C_i} \cdot p_{T_j}}\right)$$
- Measures the dependency between the observed joint probability  $p_{ij}$  of  $C$  and  $T$ , and the expected joint probability  $p_{C_i} \cdot p_{T_j}$  under the independence assumption
- When  $C$  and  $T$  are independent,  $p_{ij} = p_{C_i} \cdot p_{T_j}$ ,  $I(C, T) = 0$ . However, there is no upper bound on the mutual information

## □ Normalized mutual information (NMI)

$$NMI(C, T) = \sqrt{\frac{I(C, T)}{H(C)} \cdot \frac{I(C, T)}{H(T)}} = \frac{I(C, T)}{\sqrt{H(C) \cdot H(T)}}$$

- Value range of NMI:  $[0, 1]$ . Value close to 1 indicates a good clustering



The background of the slide is a complex, abstract composition. It features a network graph with numerous green nodes and red edges, overlaid on a light blue and white geometric pattern. A prominent white diagonal band runs across the center, serving as a backdrop for the title. In the bottom-left corner, there is a small inset image showing a cluster of orange and red dots, possibly representing a data visualization or a network component.

# **Session 6: External Measures III: Pairwise Measures**

# Pairwise Measures: Four Possibilities for Truth Assignment

❑ **Four possibilities** based on the agreement between cluster label and partition label

❑ **TP: true positive**—Two points  $\mathbf{x}_i$  and  $\mathbf{x}_j$  belong to the same partition  $T$ , and they also in the same cluster  $C$

$$TP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$$

where  $y_i$ : the true partition label, and  $\hat{y}_i$ : the cluster label for point  $\mathbf{x}_i$

❑ **FN: false negative:**  $FN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}|$

❑ **FP: false positive**  $FP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$

❑ **TN: true negative**  $TN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}|$

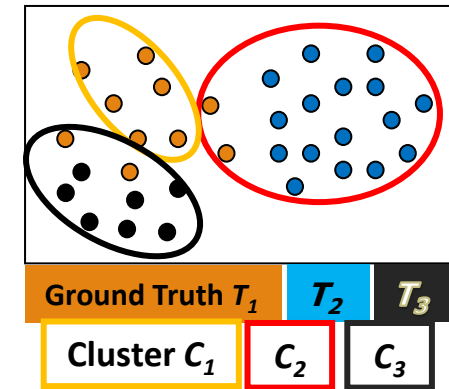
❑ Calculate the four measures:

$$N = \binom{n}{2}$$

Total # of pairs of points

$$TP = \sum_{i=1}^r \sum_{j=1}^k \binom{n_{ij}}{2} = \frac{1}{2} \left( \left( \sum_{i=1}^r \sum_{j=1}^k n_{ij}^2 \right) - n \right) \quad FN = \sum_{j=1}^k \binom{m_j}{2} - TP$$

$$FP = \sum_{i=1}^r \binom{n_i}{2} - TP \quad TN = N - (TP + FN + FP) = \frac{1}{2} \left( n^2 - \sum_{i=1}^r n_i^2 - \sum_{j=1}^k m_j^2 + \sum_{i=1}^r \sum_{j=1}^k n_{ij}^2 \right)$$





# Pairwise Measures: Jaccard Coefficient and Rand Statistic

- ❑ **Jaccard coefficient:** Fraction of true positive point pairs, but after ignoring the true negatives (thus asymmetric)

- ❑  $Jaccard = TP / (TP + FN + FP)$  [i.e., denominator ignores  $TN$ ]

- ❑ Perfect clustering:  $Jaccard = 1$

- ❑ **Rand Statistic:**

- ❑  $Rand = (TP + TN) / N$

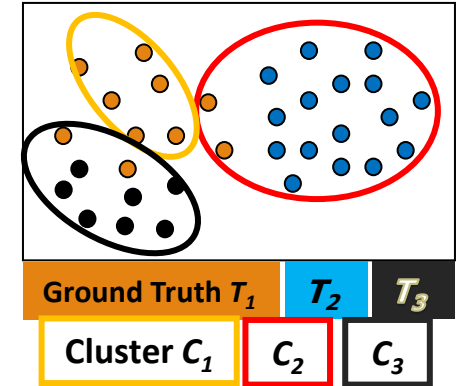
- ❑ Symmetric; perfect clustering:  $Rand = 1$

- ❑ **Fowlkes-Mallow Measure:**

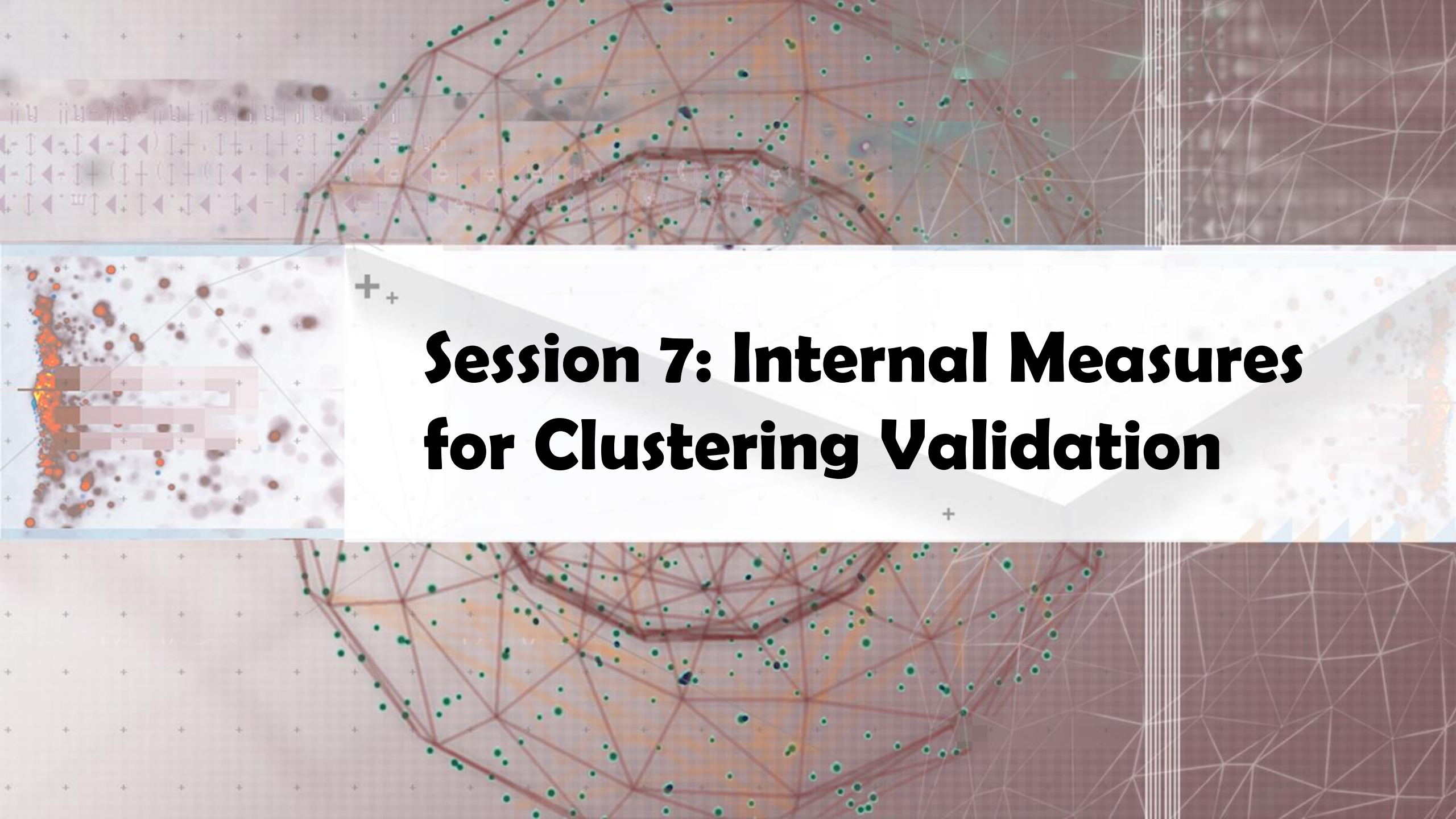
- ❑ Geometric mean of precision and recall

$$FM = \sqrt{prec \times recall} = \frac{TP}{\sqrt{(TP + FN)(TP + FP)}}$$

- ❑ Using the above formulas, one can calculate all the measures for the green table (leave as an exercise)



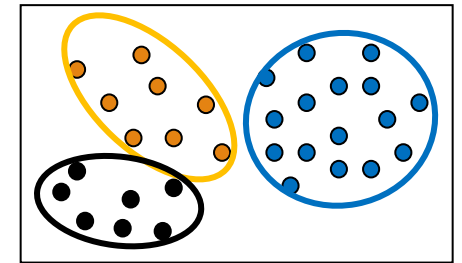
$C \backslash T$	$T_1$	$T_2$	$T_3$	Sum
$C_1$	0	20	30	50
$C_2$	0	20	5	25
$C_3$	25	0	0	25
$m_j$	25	40	35	100

The background of the slide is a complex, abstract composition. It features a network of thin, light-colored lines forming a web-like structure. Overlaid on this are various data visualization elements: a grid of small grey plus signs, clusters of green and blue dots, and a prominent orange and red cluster on the left side. The overall color palette is muted, with earthy tones and soft pastels.

# **Session 7: Internal Measures for Clustering Validation**

# Internal Measures (I): BetaCV Measure

- A trade-off in maximizing intra-cluster compactness and inter-cluster separation
- Given a clustering  $C = \{C_1, \dots, C_k\}$  with  $k$  clusters, cluster  $C_i$  containing  $n_i = |C_i|$  points
  - Let  $W(S, R)$  be sum of weights on all edges with one vertex in  $S$  and the other in  $R$
  - The sum of all the intra-cluster weights over all clusters:  $W_{in} = \frac{1}{2} \sum_{i=1}^k W(C_i, C_i)$
  - The sum of all the inter-cluster weights:  $W_{out} = \frac{1}{2} \sum_{i=1}^k W(C_i, \overline{C_i}) = \sum_{i=1}^{k-1} \sum_{j>i}^k W(C_i, C_j)$
  - The number of distinct intra-cluster edges:  $N_{in} = \sum_{i=1}^k \binom{n_i}{2}$
  - The number of distinct inter-cluster edges:  $N_{out} = \sum_{i=1}^{k-1} \sum_{j=i+1}^k n_i n_j$
- **Beta-CV measure:**  $BetaCV = \frac{W_{in} / N_{in}}{W_{out} / N_{out}}$ 
  - The ratio of the mean intra-cluster distance to the mean inter-cluster distance
  - The smaller, the better the clustering



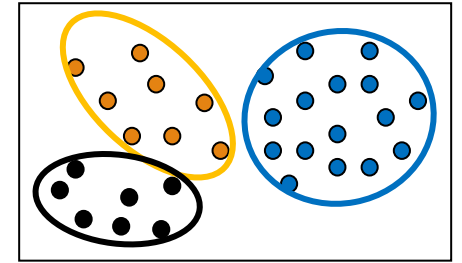


# Internal Measures (II): Normalized Cut and Modularity

□ **Normalized cut:** 
$$NC = \sum_{i=1}^k \frac{W(C_i, \bar{C}_i)}{vol(C_i)} = \sum_{i=1}^k \frac{W(C_i, \bar{C}_i)}{W(C_i, V)} = \sum_{i=1}^k \frac{W(C_i, \bar{C}_i)}{W(C_i, C_i) + W(C_i, \bar{C}_i)} = \sum_{i=1}^k \frac{1}{\frac{W(C_i, C_i)}{W(C_i, \bar{C}_i)} + 1}$$

where  $vol(C_i) = W(C_i, V)$  is the volume of cluster  $C_i$

- The higher normalized cut value, the better the clustering



□ **Modularity** (for graph clustering) 
$$Q = \sum_{i=1}^k \left( \frac{W(C_i, C_i)}{W(V, V)} - \left( \frac{W(C_i, V)}{W(V, V)} \right)^2 \right)$$

- Modularity  $Q$  is defined as

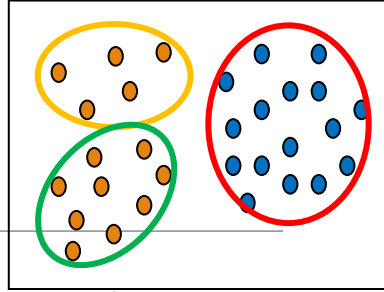
where 
$$W(V, V) = \sum_{i=1}^k W(C_i, V) = \sum_{i=1}^k W(C_i, C_i) + \sum_{i=1}^k W(C_i, \bar{C}_i) = 2(W_{in} + W_{out})$$

- Modularity measures the difference between the observed and expected fraction of weights on edges within the clusters.
- The smaller the value, the better the clustering—the intra-cluster distances are lower than expected

The background of the slide is a complex, abstract composition. It features a dark, reddish-brown base with a network of thin, light-colored lines forming a triangular mesh. Overlaid on this are various data visualizations: a grid of small, light-colored plus signs, a series of small, colorful dots (green, blue, yellow) connected by lines, and a large, semi-transparent white triangle in the center. The text 'Session 8: Relative Measures' is prominently displayed in the center of this white triangle.

# **Session 8: Relative Measures**

# Relative Measure



- Relative measure: Directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm

- Silhouette coefficient as an internal measure:** Check cluster cohesion and separation

- For each point  $\mathbf{x}_i$ , its silhouette coefficient  $s_i$  is: 
$$s_i = \frac{\mu_{out}^{\min}(\mathbf{x}_i) - \mu_{in}(\mathbf{x}_i)}{\max\{\mu_{out}^{\min}(\mathbf{x}_i), \mu_{in}(\mathbf{x}_i)\}}$$
 where  $\mu_{in}(\mathbf{x}_i)$  is the mean distance from  $\mathbf{x}_i$  to points in its own cluster

$\mu_{out}^{\min}(\mathbf{x}_i)$  is the mean distance from  $\mathbf{x}_i$  to points in its closest cluster

- Silhouette coefficient (SC) is the mean values of  $s_i$  across all the points: 
$$SC = \frac{1}{n} \sum_{i=1}^n s_i$$

- SC close to +1 implies good clustering

- Points are close to their own clusters but far from other clusters

- Silhouette coefficient as a relative measure:** Estimate the # of clusters in the data

$$SC_i = \frac{1}{n_i} \sum_{x_j \in C_i} s_j$$

Pick the  $k$  value that yields the best clustering, i.e., yielding high values for  $SC$  and  $SC_i$  ( $1 \leq i \leq k$ )

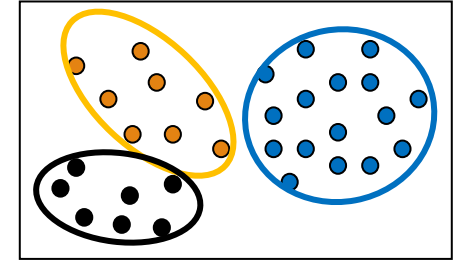


The background of the slide is a complex, abstract composition. It features a dark, reddish-brown base with a network of thin, light-colored lines forming a mesh or web-like structure. Overlaid on this are various data visualizations: a grid of small, light-colored plus signs, a series of small, colorful dots (green, blue, yellow) connected by lines, and a large, semi-transparent white triangle that points downwards. In the top left corner, there is a small, rectangular inset showing a cluster of orange and red dots. The overall aesthetic is technical and data-driven.

# Session 9: Cluster Stability

# Cluster Stability

- ❑ Clusterings obtained from several datasets sampled from the same underlying distribution as  $\mathbf{D}$  should be similar or “stable”
- ❑ Typical approach:
  - ❑ Find good parameter values for a given clustering algorithm
- ❑ Example: Find a good value of  $k$ , the correct number of clusters
- ❑ A **bootstrapping approach** to find the best value of  $k$  (judged on stability)
  - ❑ Generate  $t$  samples of size  $n$  by sampling from  $\mathbf{D}$  with replacement
  - ❑ For each sample  $\mathbf{D}_i$ , run the same clustering algorithm with  $k$  values from 2 to  $k_{max}$
  - ❑ Compare the distance between all pairs of clusterings  $C_k(\mathbf{D}_i)$  and  $C_k(\mathbf{D}_j)$  via some distance function
    - ❑ Compute the expected pairwise distance for each value of  $k$
  - ❑ The value  $k^*$  that exhibits the least deviation between the clusterings obtained from the resampled datasets is the best choice for  $k$  since it exhibits the most stability



# Other Methods for Finding K, the Number of Clusters

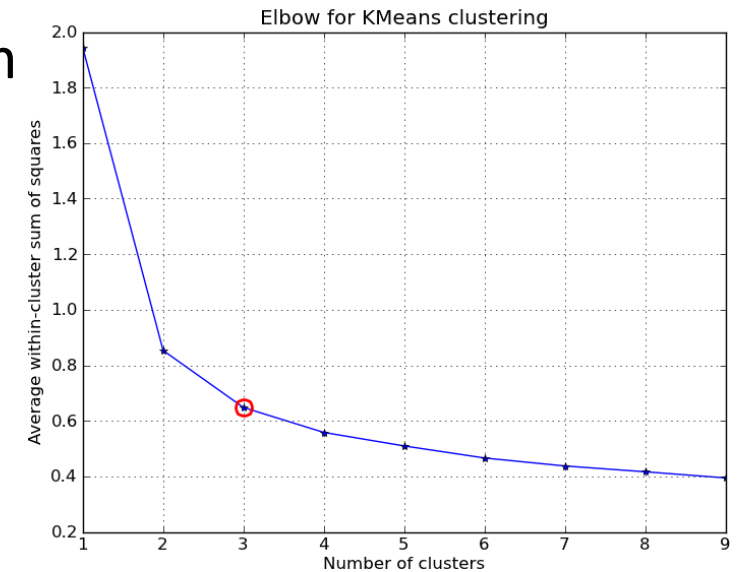
## □ Empirical method

- # of clusters:  $k \approx \sqrt{n/2}$  for a dataset of  $n$  points (e.g.,  $n = 200$ ,  $k = 10$ )

## □ Elbow method: Use the turning point in the curve of the sum of within cluster variance with respect to the # of clusters

## □ Cross validation method

- Divide a given data set into  $m$  parts
- Use  $m - 1$  parts to obtain a clustering model
- Use the remaining part to test the quality of the clustering
  - For example, for each point in the test set, find the closest centroid, and use the sum of squared distance between all points in the test set and the closest centroids to measure how well the model fits the test set
- For any  $k > 0$ , repeat it  $m$  times, compare the overall quality measure w.r.t. different  $k$ 's, and find # of clusters that fits the data the best



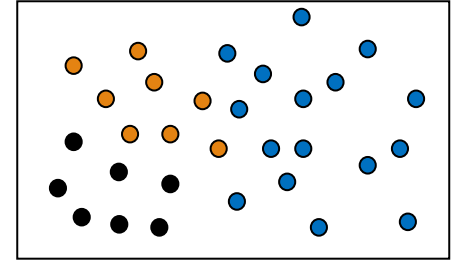


The background features a complex, abstract design. It includes a grid of small grey plus signs, a network of thin red lines connecting green dots, and a large, light grey geometric shape resembling a stylized 'V' or a folded piece of paper. The overall color palette is muted, with greys, reds, and greens.

# Session 10: Clustering Tendency

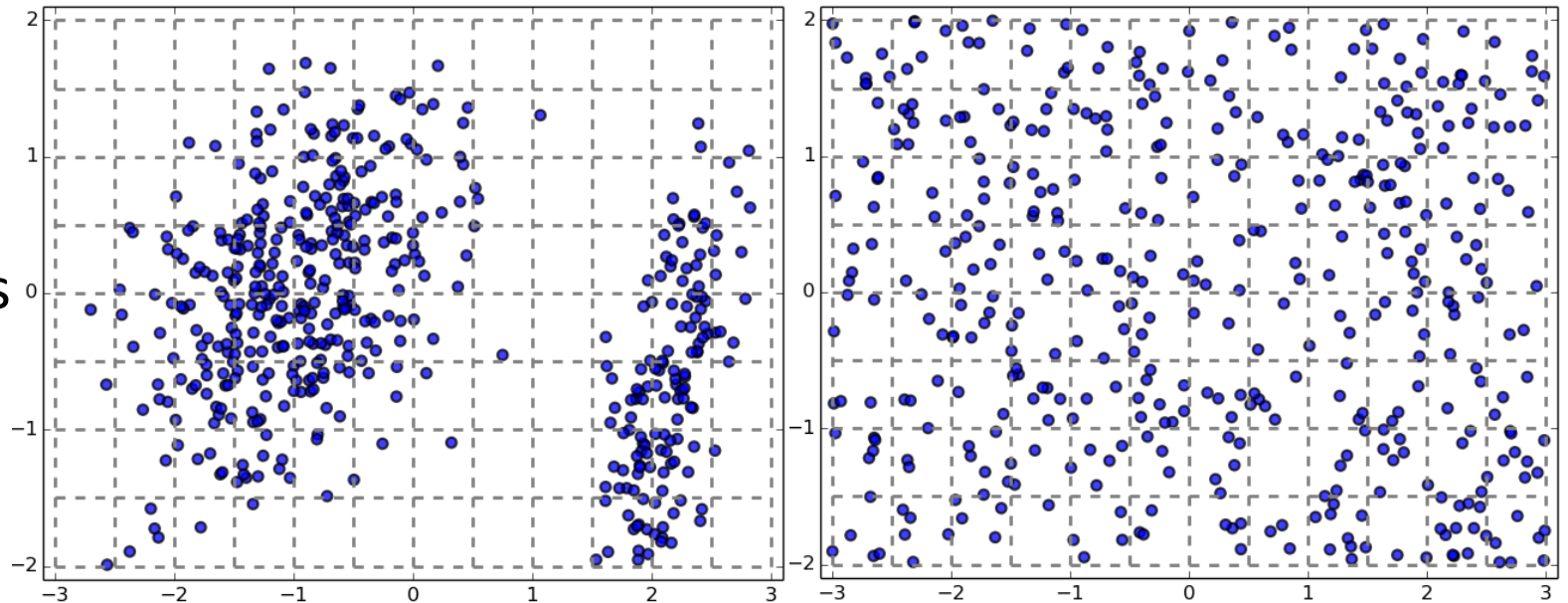
# Clustering Tendency: Whether the Data Contains Inherent Grouping Structure

- ❑ 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





The background of the slide is a complex, abstract composition. It features a central white banner with a subtle geometric pattern. Above and below this banner are sections with a dark, reddish-brown background, overlaid with a network of thin, light-colored lines and small, colorful dots (green, blue, yellow). On the left side, there is a vertical strip with a light blue background, containing a grid of small, colorful dots (orange, red, blue, purple) and a larger, semi-transparent rectangular area with a grid pattern. The overall aesthetic is modern and technical, suggesting a focus on data or science.

# Session 11: Summary

# Summary: Clustering Validation

---

- ❑ Clustering Validation: Basic Concepts
- ❑ Clustering Evaluation: Measuring Clustering Quality
- ❑ External Measures for Clustering Validation
  - ❑ I: Matching-Based Measures
  - ❑ II: Entropy-Based Measures
  - ❑ III: Pairwise Measures
- ❑ Internal Measures for Clustering Validation
- ❑ Relative Measures
- ❑ Cluster Stability
- ❑ Clustering Tendency
- ❑ Summary

# 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, 3<sup>rd</sup> 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