Machine Learning CS 4641-7641

Clustering Evaluation

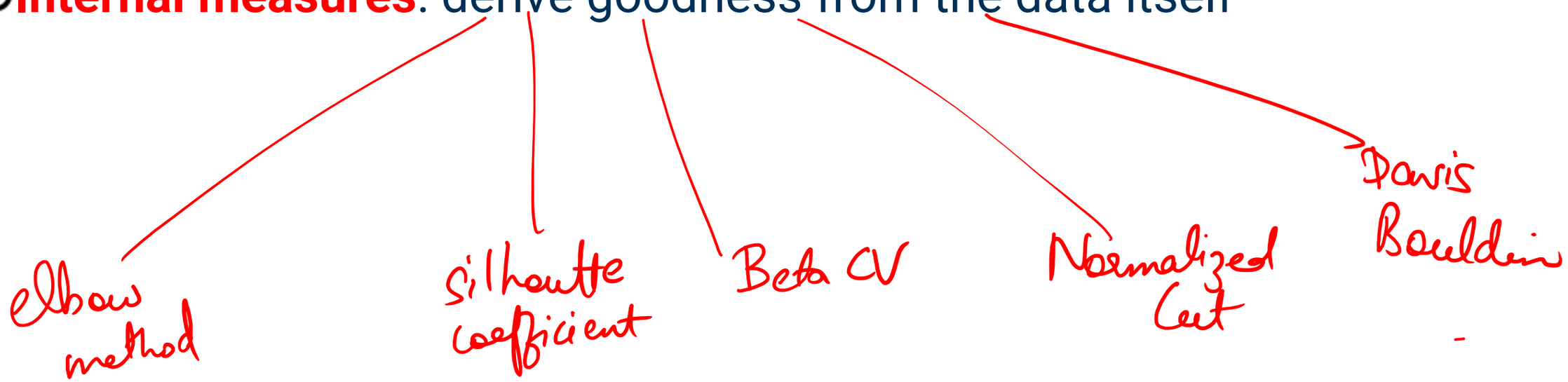
Dr. Nimisha Roy
Lecturer
Director of Online Undergraduate Initiatives
Georgia Tech

These slides are based on slides from Mahdi Roozbahani

Clustering Evaluation

- Clustering evaluation aims at quantifying the goodness or quality of the clustering.

- Two main categories of measures:
  - **External measures**: employ external ground-truth
  - **Internal measures**: derive goodness from the data itself



Outline

- External measures for clustering evaluation
  - Matching-based measures
  - Entropy-based measures
  - Pairwise measures
- Internal measures for clustering evaluation
  - Graph-based measures
  - Davies-Bouldin Index
  - Silhouette Coefficient

External Measures

External measures assume that the correct or ground-truth clustering is known *a priori*, which is used to evaluate a given clustering.

Let $\mathbf{D} = \{\mathbf{x}_i\}_{i=1}^{n}$ be a dataset consisting of n points in a d-dimensional space, partitioned into k clusters. Let $y_i \in \{1, 2, \ldots, k\}$ denote the ground-truth cluster membership or label information for each point.

The ground-truth clustering is given as $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$, where the cluster T_j consists of all the points with label j, i.e., $T_j = \{\mathbf{x}_i \in \mathbf{D} | y_i = j\}$. We refer to $\mathcal{T}$ as the ground-truth *partitioning*, and to each T_i as a *partition*.

Let $\mathcal{C} = \{C_1, \ldots, C_r\}$ denote a clustering of the same dataset into r clusters, obtained via some clustering algorithm, and let $\hat{y}_i \in \{1, 2, \ldots, r\}$ denote the cluster label for $\mathbf{x}_i$.

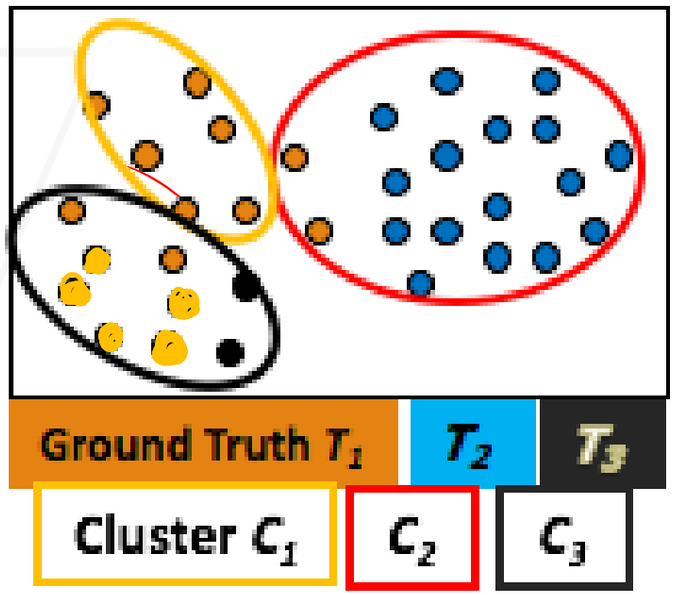
So **k** is the number of ground truth partitions (T) and **r** is the number of clusters (C) obtained by algorithm

n_{ij} = Number of data points in cluster **i** which are also in ground truth partition **j**

Matching-Based Measures (I): Purity

n_{ij} no. of datapoints in cluster i of partition j.

- **Purity**: Quantifies the extent that cluster C_i contains points only from one (ground truth) partition:



Cluster 1: n_{11}, n_{12}, n_{13} → 6 0 0

$purity_1 = \frac{\max(6,0,0)}{6} = 1$

$purity_3 = \frac{7,2,0}{9} = 7/9$

$$purity_i = \frac{1}{n_i} \max_{j=1}^{k} \{n_{ij}\}$$

cluster i; no. of datapoints in cluster i

$$purity_3 = \frac{1}{n_3} \max(n_{31}, n_{32}, n_{33})$$

$$= \frac{1}{9} \max(2,0,7) = \frac{7}{9}$$

The Total purity of clustering C is the weighted sum of the cluster-wise purity:

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

What is purity value for a perfect clustering? Purity = 1

$$purity_i = \frac{1}{n_i} \max_{j=1}^{k} \{n_{ij}\}$$

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

$purity_1 = \frac{1}{50} \max(n_{11}, n_{12}, n_{13})$

$= \frac{1}{50} \max(0, 20, 30)$

$= 30/50$

n_{ij} table

n_{13}

Partition

Cluster

C\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

Example:

purity1 = 30/50;
purity2 = 20/25;
purity3 = 25/25;
purity = (30 + 20 + 25)/100 = 0.75

$\frac{30}{50} \times \frac{50}{100} + \frac{20}{25} \times \frac{25}{100} + \frac{25}{25} \times \frac{25}{100}$

Two clusters may be matched to the same partition

C1 is more paired with T3
C2 is more paired with T2

C\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

purity = (30 + 20 + 25)/100 = 0.75

C1 is more paired with T2
C2 is more paired with T2

C\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

purity = (30 + 20 + 25)/100 = 0.75

Georgia Tech

Maximum weight matching: Only one cluster can match one partition

Ex. If C1 is more paired with T2 **THEN** C2 and C3 cannot paired with T2

C\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

Purity C_1 = 30/50

Purity C_2 = 5/25

Purity C_3 = 25/25

Purity

MAX

C1 is more paired with T2 = $\frac{30+5+25}{100} = 0.6$

C1 is more paired with T3 = $\frac{20+20+25}{100} = 0.65$

Purity = 0.65

Purity C_1 = 20/50

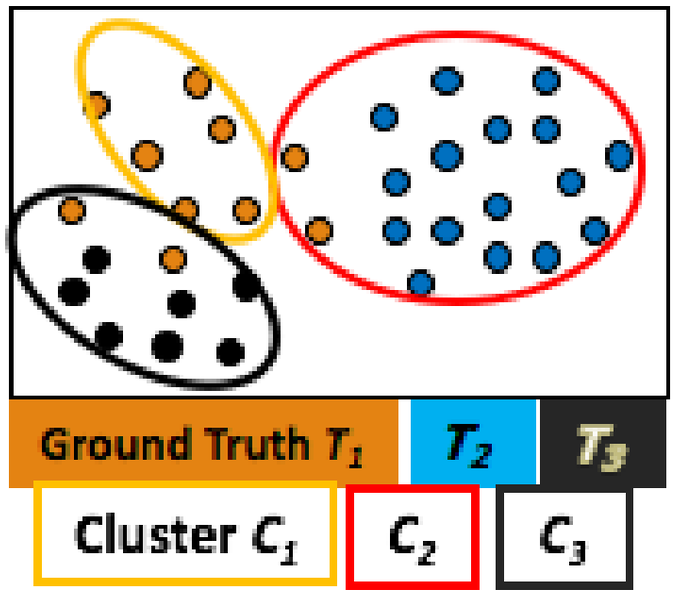
Purity C_2 = 20/25

Purity C_3 = 25/25

Matching-Based Measures (II): Maximum Matching

- **Drawback of purity**: two clusters may be matched to the same partition.
- **Maximum matching**: the maximum purity under the one-to-one matching constraint.
  - Examine all possible pairwise matching between C and T and choose the best (the maximum)

Example:
Maximum matching = 0.65 > 0.6



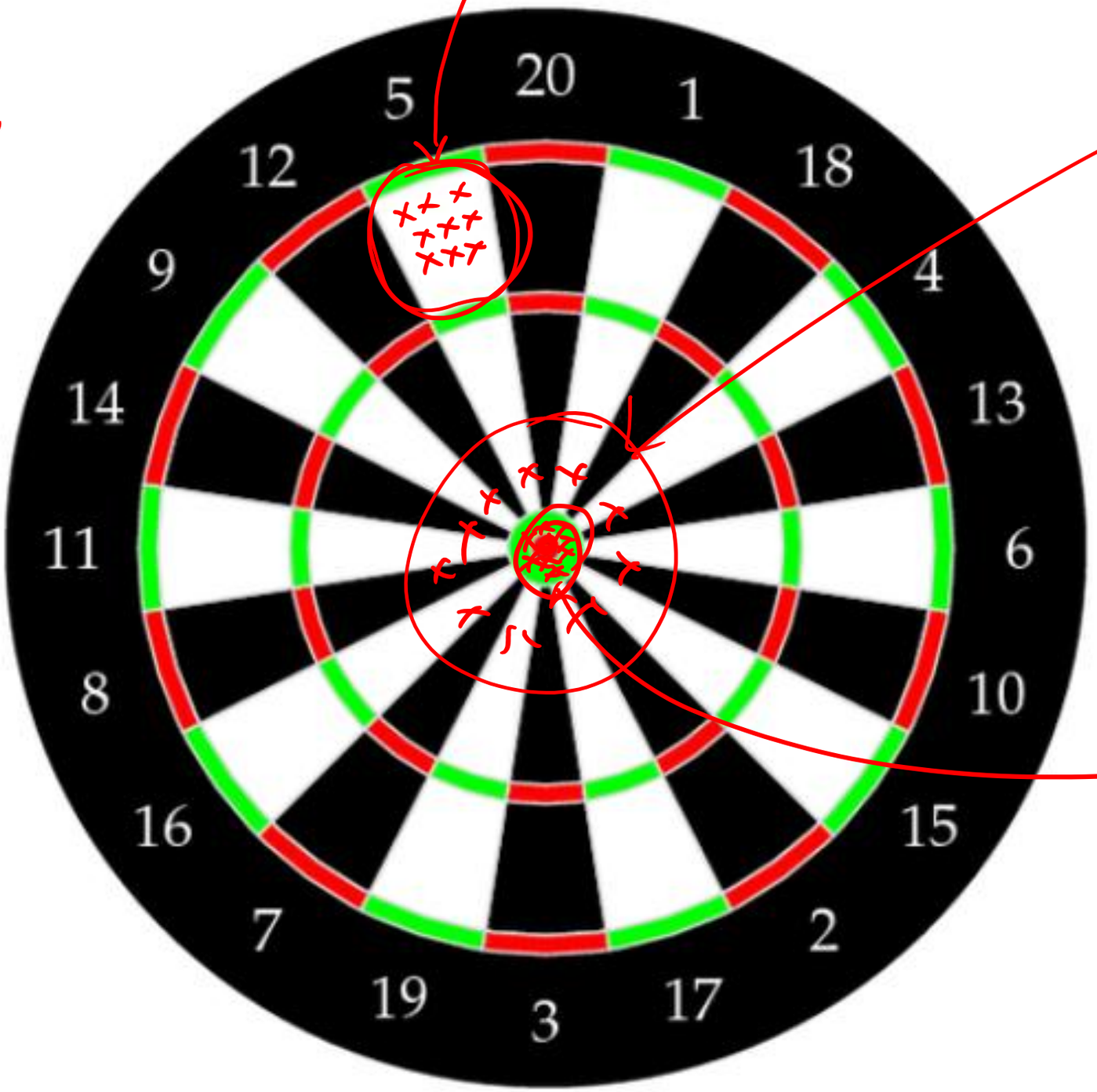
C\T	T_1	T_2	T_3	Sum
C_1	0	30	20	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	50	25	100

Precision → LOCAL MEASURE

Accuracy → GLOBAL MEASURE

high precision
low accuracy

lower precision
High accuracy



high precision
high accuracy

In a general context: Precision, Recall and Accuracy

Per point Scale

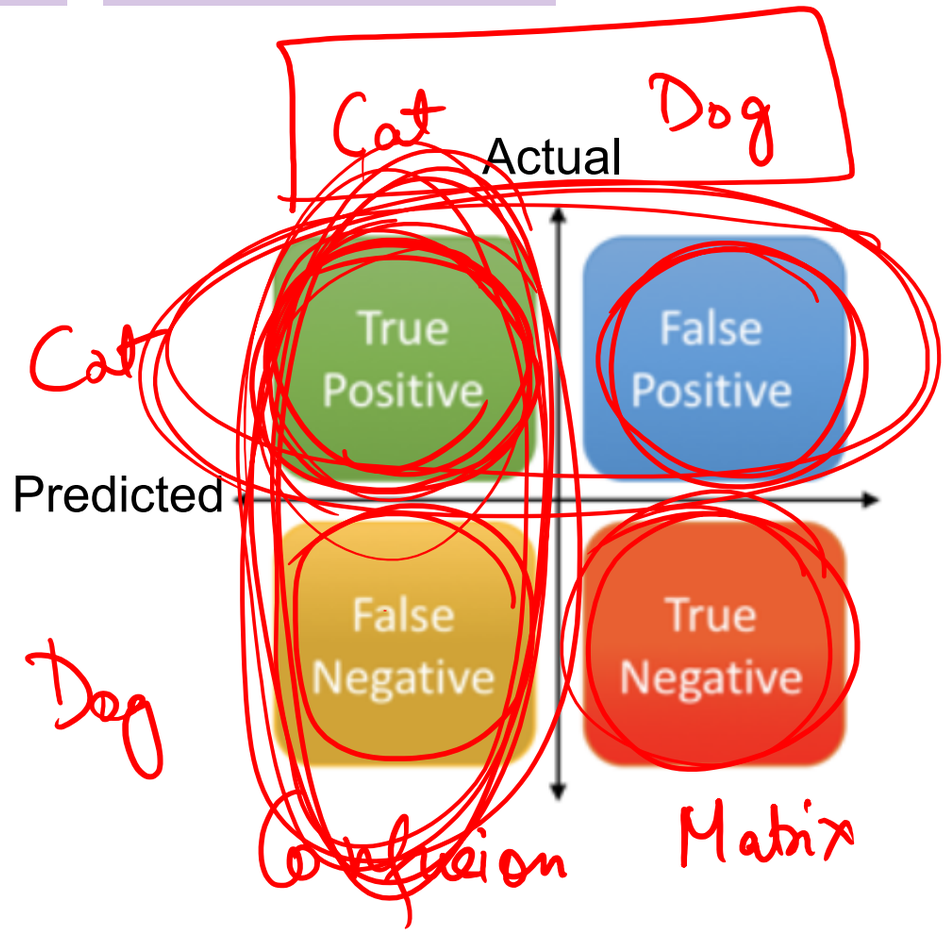
	Correct prediction	Wrong prediction
Number of predicted “positive” labeled data =	True Positive	+ False Positive
Number of predicted “negative” labeled data =	True Negative	+ False Negative

Positive
Cat
Dog
Negative

Precision = $\frac{\textit{True Positive}}{\textit{Predicted Results}}$ or $\frac{\text{True Positive}}{\text{True Positive + False Positive}}$

Recall = $\frac{\textit{True Positive}}{\textit{Actual Results}}$ or $\frac{\text{True Positive}}{\text{True Positive + False Negative}}$

Accuracy = $\frac{\text{True Positive + True Negative}}{\text{Total}}$



False positive is also called false alarm

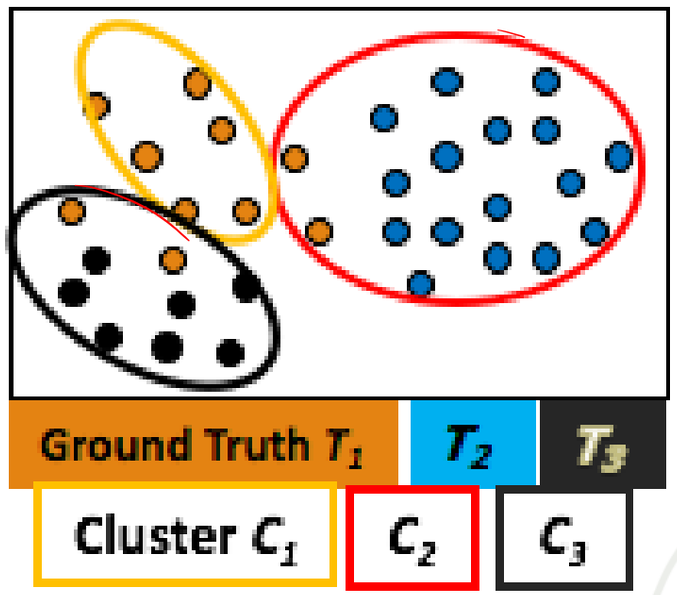
Cluster Scale

Matching-Based Measures (II): F-Measure

- **Precision**: which measure ***quality***, is the same as purity:
  - How precisely does each cluster represent the ground truth?
  - Of all predicted positives, how many are actually positive?
- **Recall**: measures completeness
  - How completely does each cluster recover the ground truth?
  - Of all actual positives, how many were correctly predicted?

$$recall_i = \frac{n_{ij_i}}{|T_{j_i}|} = \frac{n_{ij_i}}{m_{j_i}} \quad Recall_1 = \frac{6}{10} \qquad prec_i = \frac{1}{n_i}\max_{j=1}^{k}\{n_{ij}\} = \frac{n_{ij_i}}{n_i} \quad Prec_1 = \frac{6}{6}$$

- Ci = i-th **cluster** produced by the algorithm
- Tj* = **ground truth partition** that best matches cluster Ci
- Nij*= number of shared data points between Ci and Tj*
- mj* = total number of points in that ground truth partition Tj*



$recall_3 = \frac{7}{7}$ $precision_3 = 7/9$

$recall_2 = \frac{16}{16}$ $precision_2 = \frac{16}{18}$

$recall_1 = \frac{6}{10}$ $precision_1 = \frac{6}{6}$

Georgia Tech

Precision and Recall

(Precision here is same as the purity)

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

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

m_{ij}

C\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

Matching-Based Measures (II): F-Measure

- **F-Measure**: the harmonic mean of precision and recall
  - Take into account both *precision* and *completeness*

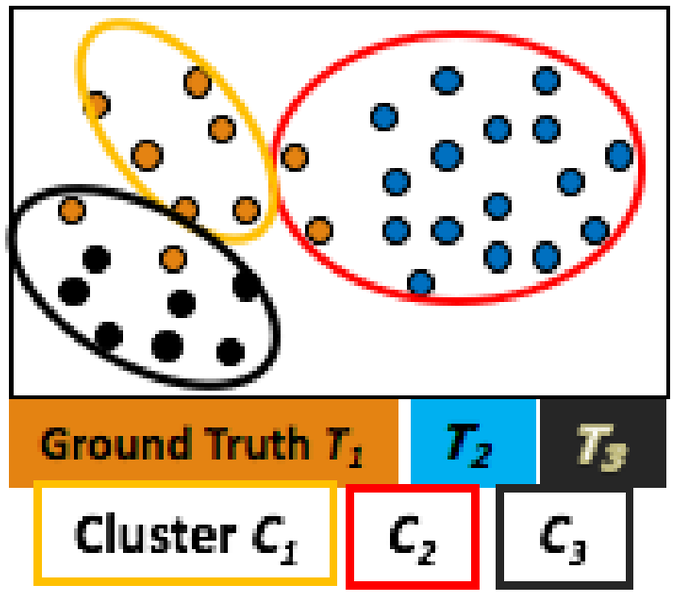
harmonic mean

$$F_i = \frac{2}{\frac{1}{prec_i} + \frac{1}{recall_i}} = \frac{2 \cdot prec_i \cdot recall_i}{prec_i + recall_i} = \frac{2\, n_{ij_i}}{n_i + m_{j_i}}$$

The F-measure for the clustering $\mathcal{C}$ is the mean of clusterwise F-measure values:

$$F = \frac{1}{r} \sum_{i=1}^{r} F_i$$

F1 = 60/85;
F2 = 40/65;
F3 = 1;
F = 0.774



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

Entropy-Based Measures

- Expectation

$$E[g(x)] = \sum p(x)\, g(x)$$

- Information

$$\log_2 \frac{1}{p(x)} = -\log_2 p(x)$$

- Entropy

$$H(x) = E[I(x)] = \sum p(x) I(x) = -\sum p(x) \log_2 p(x)$$

- We want entropy of a cluster to be high or low?

- Conditional Entropy

$$H(Y|X_i) = -\sum_i p(y|x_i) \log_2 p(y|x_i)$$

$$H(Y|X) = -\sum_i p(x)\, H(Y|X_i)$$

Entropy-Based Measures (I): Conditional Entropy

Conditional entropy measures how mixed the true labels are inside clusters.

H(T|C): How uncertain are the true labels T once we know the cluster assignment C?

H(T|C) for best clustering = 0

Interpretation:

- Low value → clusters match true classes well
- High value → clusters mix many different classes

C is independent of T

H(T|C) = H(T)

Entropy-Based Measures (I): Conditional Entropy

Amount of information orderness in different partitions

- The entropy for clustering C and partition T is

$$H(C) = -\sum_{i=1}^{r} p_{C_i} \log p_{C_i} \qquad H(T) = -\sum_{j=1}^{k} p_{T_j} \log p_{T_j}$$

where $p_{C_i} = \frac{n_i}{n}$ and $p_{T_j} = \frac{m_j}{n}$

i.e., The probability of ground truth T_j

$p(C_1) = 50/100$ $p(T_1) = 25/100$

$p(C_2) = 25/100$ $p(T_2) = 40/100$

$p(C_3) = 25/100$ $p(T_3) = 35/100$

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

Entropy-Based Measures (I): Conditional Entropy

- **Conditional Entropy**: The cluster-specific entropy, namely the conditional entropy of T with respect to cluster C_i:

$$H(T|C_i) = -\sum_{j=1}^{k} \left(\frac{n_{ij}}{n_i}\right) \log_2\left(\frac{n_{ij}}{n_i}\right)$$

How ground truth is distributed within each cluster

n_{ij} ← Ground truth (T)

Cluster (C)

$$H(T|C_i) = -\sum p(T|C_i) \cdot \log_2 p(T|C_i)$$

$$p(T|C_i) = \frac{p(T, C_i)}{p(C_i)} = \frac{n_{ij}}{n_i}$$

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

Entropy-Based Measures (I): Conditional Entropy

- The conditional entropy of T given clustering C is defined as the **weighted average of the entropy inside each cluster**.

$$H(\mathcal{T}|\mathcal{C}) = \sum_{i=1}^{r} \frac{n_i}{n} H(\mathcal{T}|C_i) = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log\left(\frac{p_{ij}}{p_{C_i}}\right)$$

$$= H(\mathcal{C},\mathcal{T}) - H(\mathcal{C})$$

$p(C_i) = n_i/n$

$\frac{n_{ij}}{n}$ $\frac{n_i}{n}$

The more clusters members are split into different partitions, the higher the conditional entropy (not a desirable condition and the max value is log k)

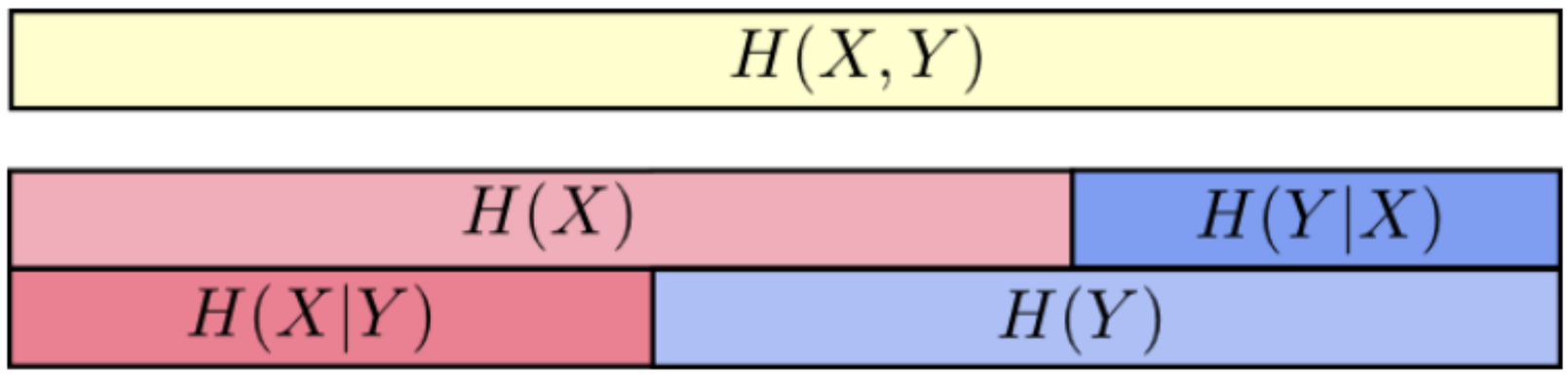
k: no. of ground truth partition

$H(\mathcal{T}|\mathcal{C}) = 0$ if and only if $\mathcal{T}$ is completely determined by $\mathcal{C}$, corresponding to the ideal clustering. If $\mathcal{C}$ and $\mathcal{T}$ are independent of each other, then $H(\mathcal{T}|\mathcal{C}) = H(\mathcal{T})$.

$$H(\mathcal{T}|\mathcal{C}) = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}\, log\, \frac{p_{ij}}{p_{\mathcal{C}_i}}$$

$$= -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}(log\, p_{ij} - \log p_{\mathcal{C}_i}) = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}\big(log\, p_{ij}\big) + \sum_{i=1}^{r}(\log p_{\mathcal{C}_i} \sum_{j=1}^{k} p_{ij}$$

$$-\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}\log p_{ij} + \sum_{i=1}^{r}(p_{\mathcal{C}_i}\log p_{\mathcal{C}_i}) = H(\mathcal{T},\mathcal{C}) - H(\mathcal{C})$$



Georgia Tech

Entropy-Based Measures (I): Mutual Information

The *mutual information* tries to quantify the amount of shared information between the clustering $\mathcal{C}$ and partitioning $\mathcal{T}$, and it is defined as

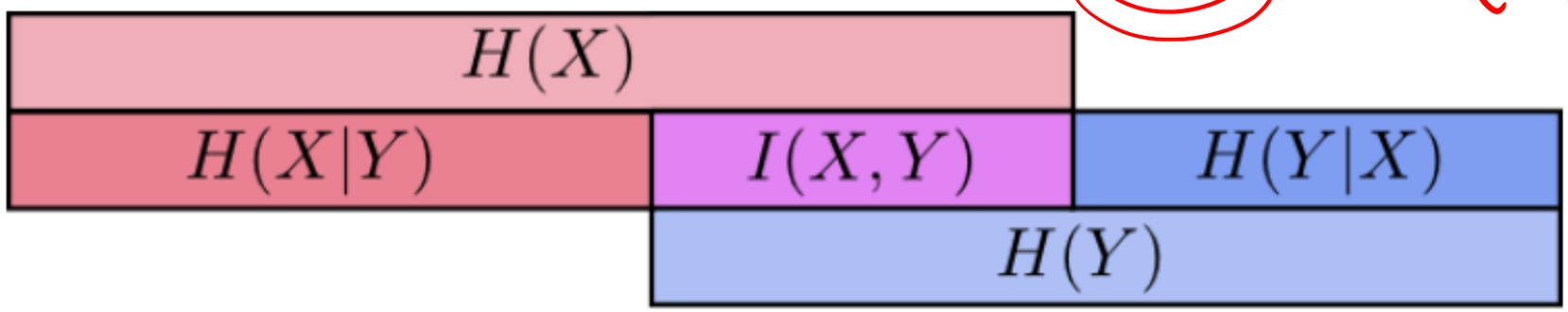
We do not have a fixed upper bound for I(C,T)

$$I(\mathcal{C},\mathcal{T}) = \sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log\left(\frac{p_{ij}}{p_{C_i} \cdot p_{T_j}}\right) = H(\mathcal{T}) - H(\mathcal{T}|\mathcal{C})$$

$H(\mathcal{T}|\mathcal{C}) = 0$

When $\mathcal{C}$ and $\mathcal{T}$ are independent then $p_{ij} = p_{C_i} \cdot p_{T_j}$, and thus $I(\mathcal{C},\mathcal{T}) = 0$.
However, there is no upper bound on the mutual information.

$I(C,T) = H(T) - H(T|C)$
$= H(C) - H(C|T)$

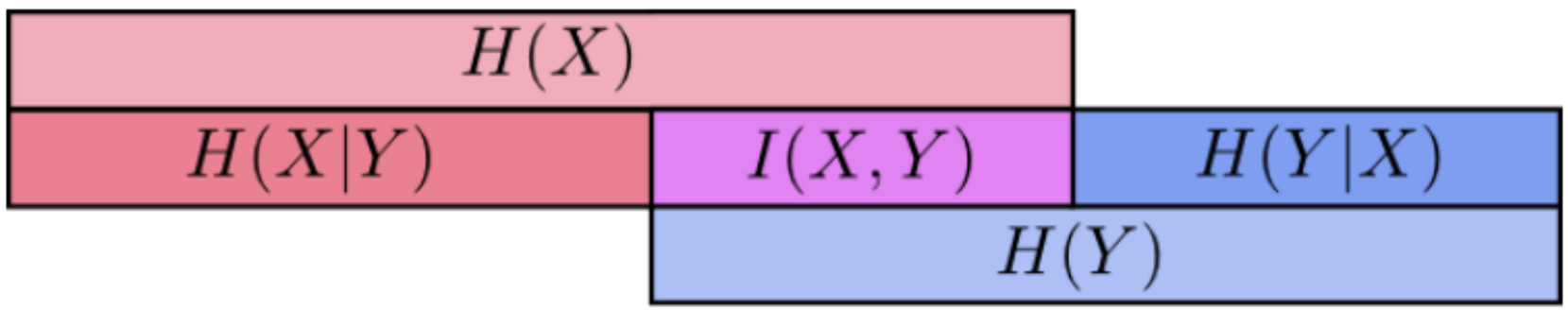


Entropy-Based Measures (I): Mutual Information

The *normalized mutual information* (NMI) is defined as the geometric mean:

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

The NMI value lies in the range $[0, 1]$. Values close to 1 indicate a good clustering.

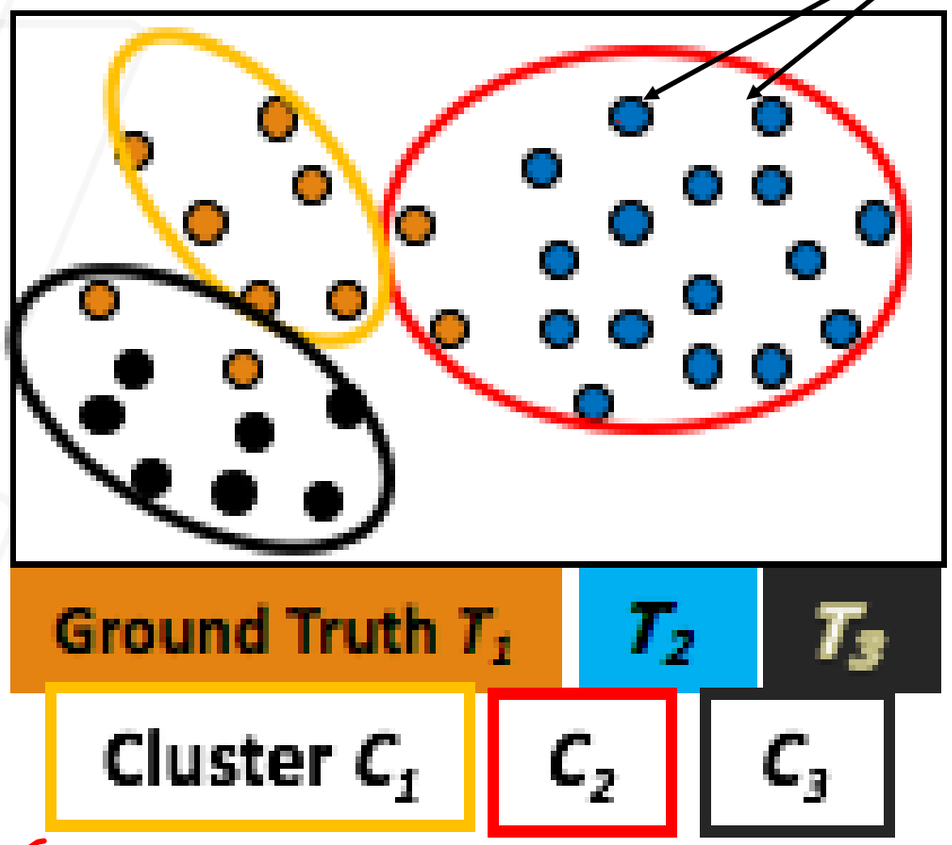


Pairwise Measures

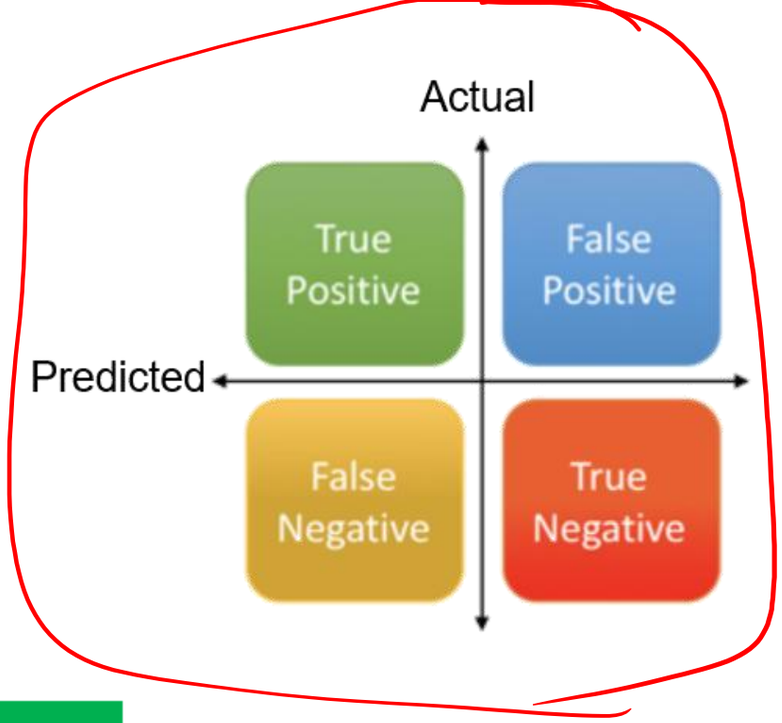
These metrics evaluate clustering based on how well it groups **pairs of points** that should (or should not) be together.

Given clustering $\mathcal{C}$ and ground-truth partitioning $\mathcal{T}$, let $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{D}$ be any two points, with $i \neq j$. Let y_i denote the true partition label and let $\hat{y}_i$ denote the cluster label for point $\mathbf{x}_i$.

Pairwise Measures



$^0C_2 + {}^{20}C_2 + {}^{30}C_2 + {}^0C_2 + {}^{20}C_2$
$+ {}^5C_2 + {}^{25}C_2 + 0 + 0$



n_{ij}

C\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

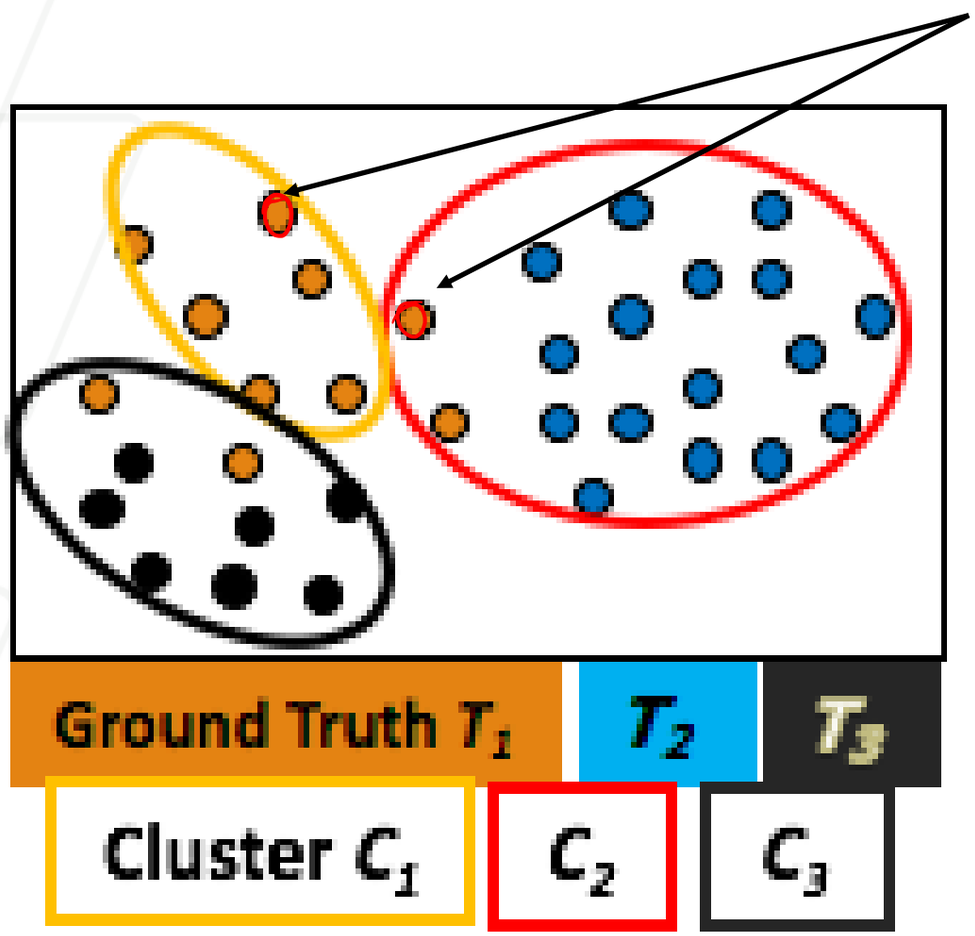
n datapoints
↓
how many edges to create pairs

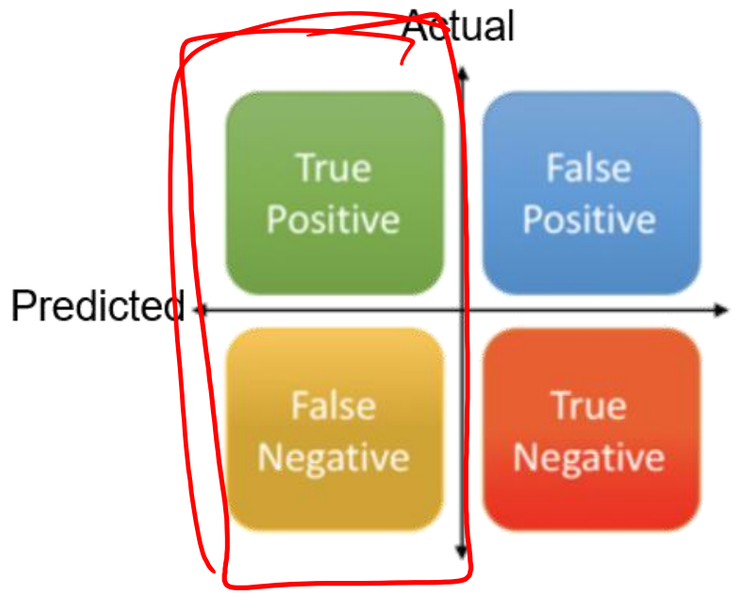
$^nC_2$ $\binom{n}{2}$

True Positives: $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same partition in $\mathcal{T}$, and they are also in the same cluster in $\mathcal{C}$. The number of true positive pairs is given as

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

Pairwise Measures





$C \backslash T$	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

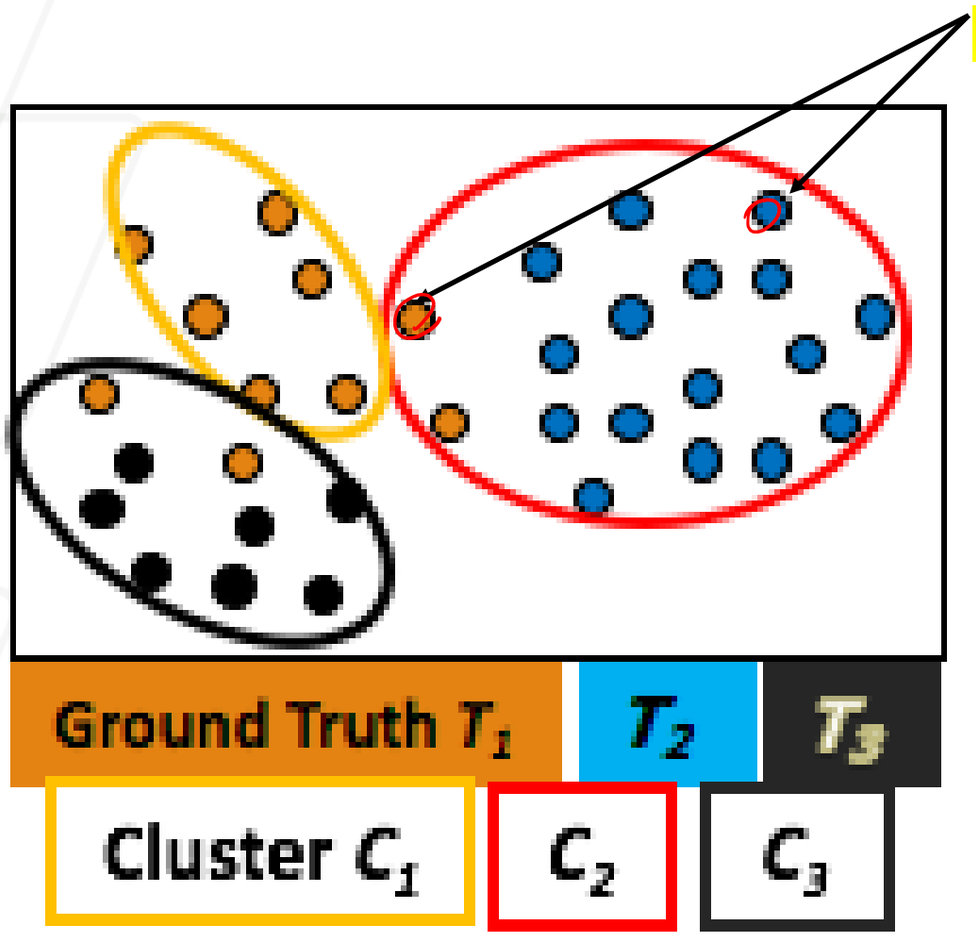
$TP + FN = {}^{25}C_2 + {}^{40}C_2 + {}^{35}C_2$

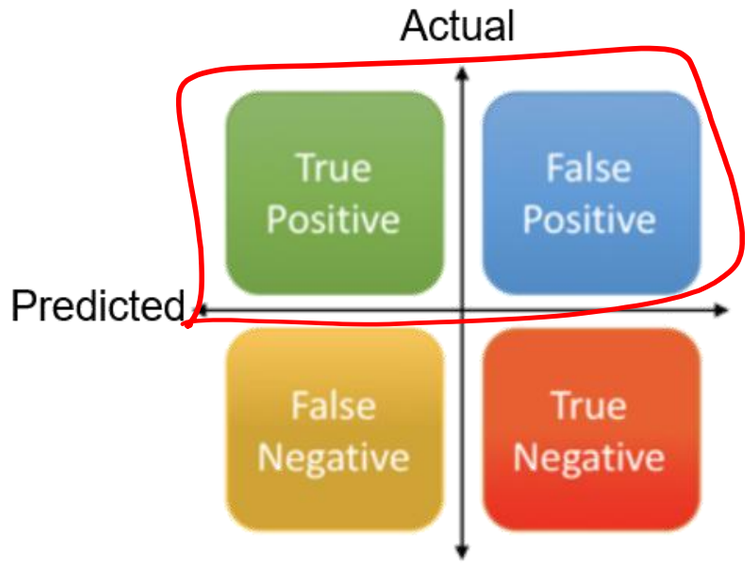
$FN = {}^{25}C_2 + {}^{40}C_2 + {}^{35}C_2 - TP$

False Negatives: $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same partition in $\mathcal{T}$, but they do not belong to the same cluster in $\mathcal{C}$. The number of all false negative pairs is given as

$$FN = \left|\{(\mathbf{x}_i, \mathbf{x}_j) : \; y_i = y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}\right|$$

Pairwise Measures





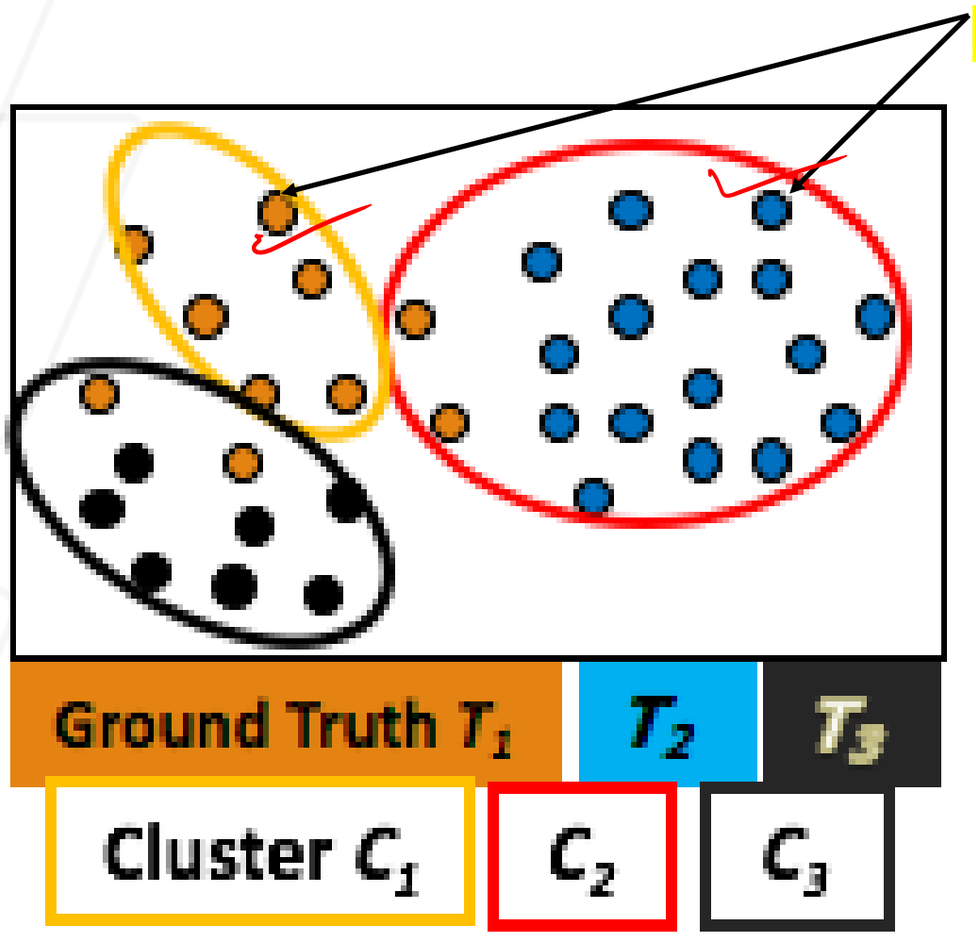
$C \backslash T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

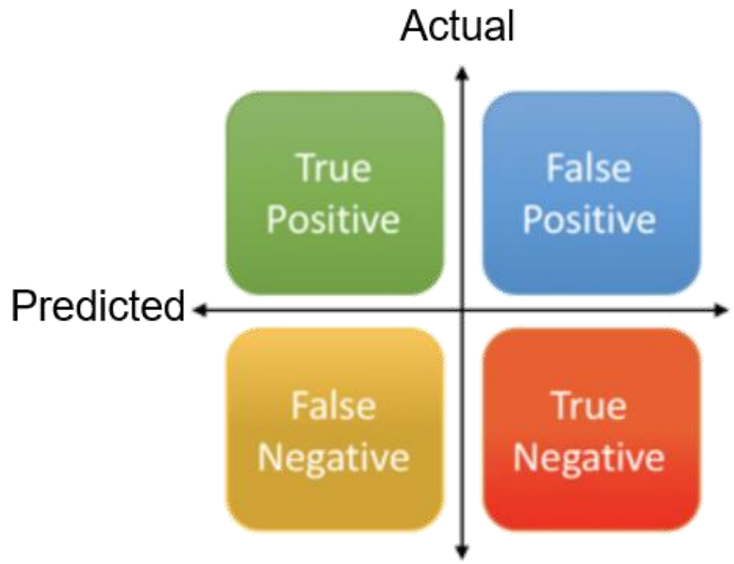
TP + FP
$= {}^{50}C_2 + {}^{25}C_2$
$+ {}^{25}C_2$

False Positives: $\mathbf{x}_i$ and $\mathbf{x}_j$ do not belong to the same partition in $\mathcal{T}$, but they do belong to the same cluster in $\mathcal{C}$. The number of false positive pairs is given as

$$FP = \left|\{(\mathbf{x}_i, \mathbf{x}_j) : \; y_i \neq y_j \text{ and } \hat{y}_i = \hat{y}_j\}\right|$$

Pairwise Measures





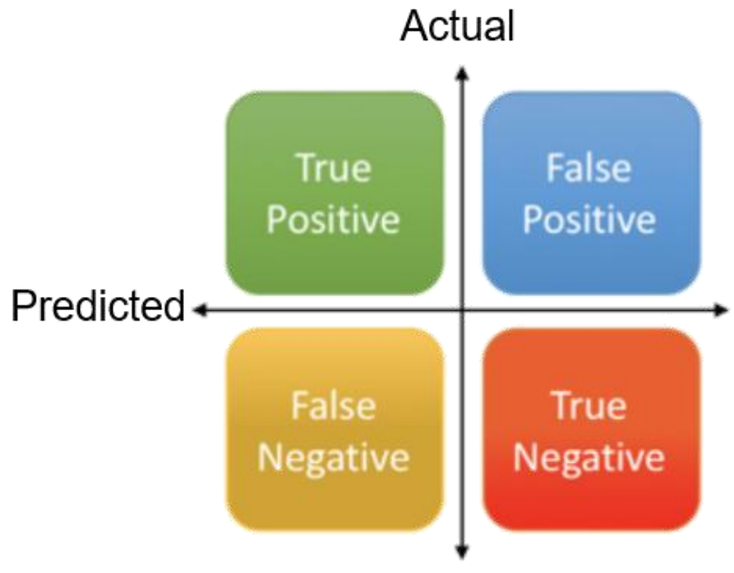
$C \backslash T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

$TP + FP + FN + TN = 100C_2$

True Negatives: $\mathbf{x}_i$ and $\mathbf{x}_j$ neither belong to the same partition in $\mathcal{T}$, nor do they belong to the same cluster in $\mathcal{C}$. The number of such true negative pairs is given as

$$TN = \left|\{(\mathbf{x}_i, \mathbf{x}_j) : \; y_i \neq y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}\right|$$

Pairwise Measures



Because there are $N = \binom{n}{2} = \frac{n(n-1)}{2}$ pairs of points, we have the following identity:

$$N = TP + FN + FP + TN$$

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

$$\text{FP} = \sum_{i=1}^{r}\binom{n_i}{2} - TP$$

$$\text{FN} = \sum_{j=1}^{k}\binom{m_j}{2} - TP$$

$$\text{TN} = N - (TP + FN + FP)$$

Pairwise Measures: Jaccard Coefficient

Jaccard Coefficient: measures the fraction of true positive point pairs, but after ignoring the true negative:

$$Jaccard = \frac{TP}{TP + FN + FP}$$

= 1

Measures **similarity** between clustering and ground truth — focuses on positive pairs only (ignores TN).

A **perfect clustering** gives Jaccard = 1.

It answers: “Of all pairs that should or were grouped together, what fraction are correctly grouped?”

Pairwise Measures: Rand

Rand Statistic: measures the fraction of true positives and true negatives over all point pairs:

$$Rand = \frac{TP + TN}{N}$$

- Measures the **fraction of agreements** (both same-cluster and different-cluster pairs) between clustering and ground truth.
- Includes true negatives (pairs correctly not grouped together).
- Conceptually similar to **accuracy** in classification.
- **Perfect clustering** → Rand = 1.
- **Downside:** Because TN pairs dominate (most pairs are in different clusters), Rand Index can appear high even if clustering is mediocre.

Pairwise Measures: FM

pair of points scale

Fowlkes-Mallows Measure: Define the overall *pairwise precision* and *pairwise recall* values for a clustering $\mathcal{C}$, as follows:

$$prec = TP/TP + FP \qquad recall = TP/TP + FN$$

The Fowlkes–Mallows (FM) measure is defined as the geometric mean of the pairwise precision and recall

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

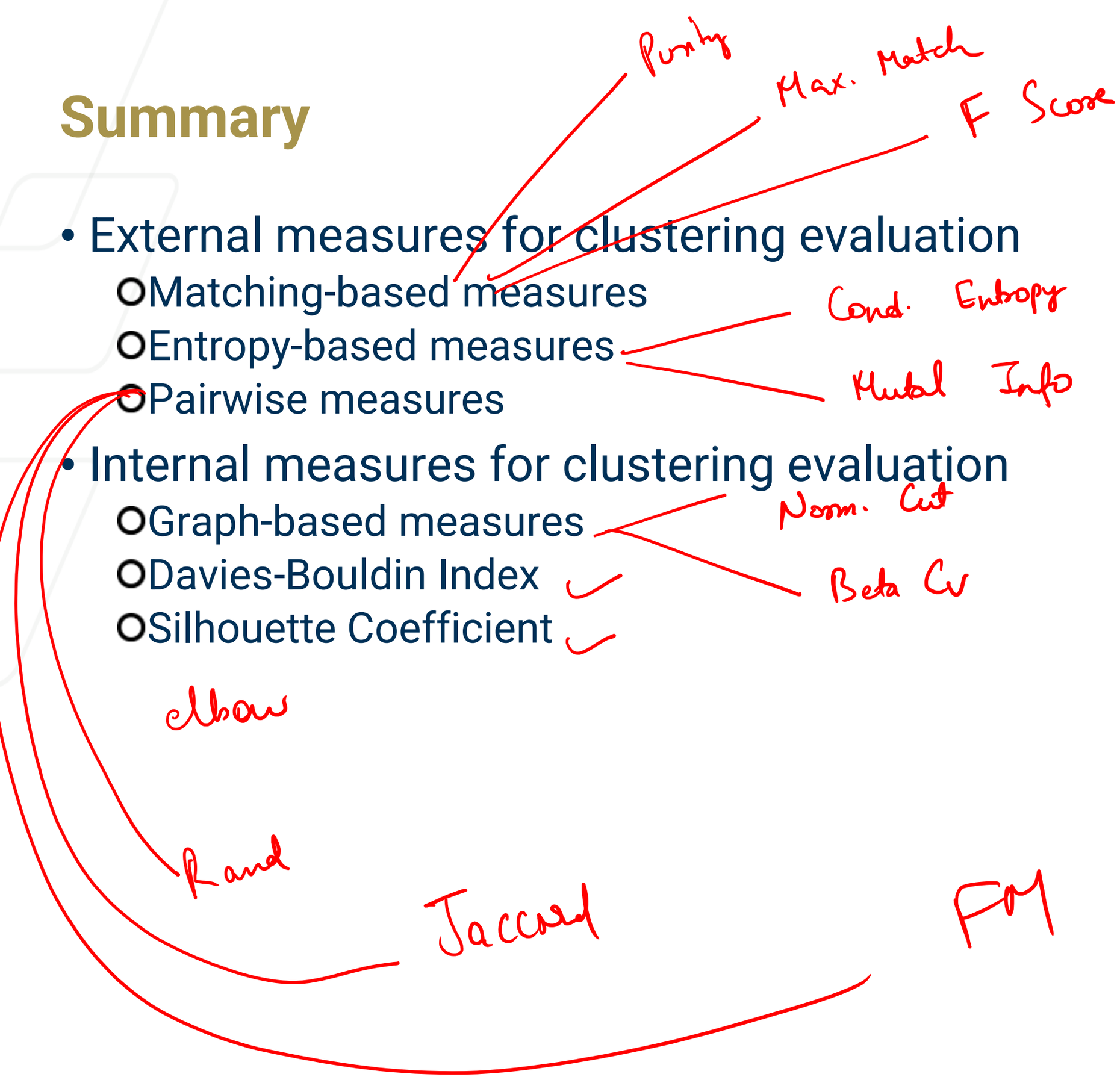
= 1

- It combines **pairwise precision** and **pairwise recall** — just like the F1 score did for classification.
- A **perfect clustering** gives FM = 1.

***Jaccard** and **FM** are preferred when **number of TNs is large**, since they avoid the dominance of negatives. **Rand Index** is simple and interpretable but can be inflated when most pairs are different.*

Summary

- External measures for clustering evaluation
  - Matching-based measures
  - Entropy-based measures
  - Pairwise measures
- Internal measures for clustering evaluation
  - Graph-based measures
  - Davies-Bouldin Index
  - Silhouette Coefficient



Quick Knowledge Check

1. **What does purity measure?** A. Compactness of clusters. B. Homogeneity with respect to ground truth. C. Separation between clusters
2. **What problem does maximum matching fix?** A. Duplicate cluster–partition matches. B. Small cluster bias. C. Missing cluster labels
3. **Why is the harmonic mean used in F-measure?** A. To emphasize the larger value. B. To balance precision and recall fairly. C. To simplify computation
4. **What does conditional entropy $H(T \mid C)$represent?** A. Uncertainty of true labels within clusters. B. Compactness of clusters. C. Distance between centroids
5. **Higher conditional entropy implies:** A. Better clustering. B. Poorer clustering. C. No change in clustering quality
6. **Why doesn't mutual information have a universal upper bound in clustering evaluation?** A. It depends on the number and size of clusters, which can vary. B. It ignores entropy in computation
7. **Jaccard coefficient ignores:** A. True negatives. B. False positives. C. False negatives
8. **Rand Index is most similar to:** A. Accuracy. B. Precision. C. Recall
9. **Why can Rand Index appear high even for poor clustering?** A. TN pairs dominate. B. FP pairs dominate. C. FN pairs dominate

Georgia Tech