CLUSTER ANALYSIS

Andrea Giansanti
Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

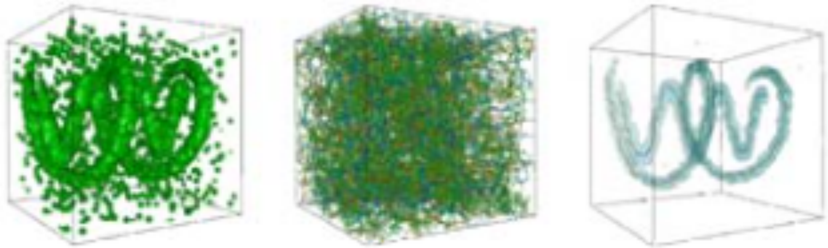
CB_20_21_L15 , Rome 17th Nov 2020

Using slides by Manolis Kelly (MIT)

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

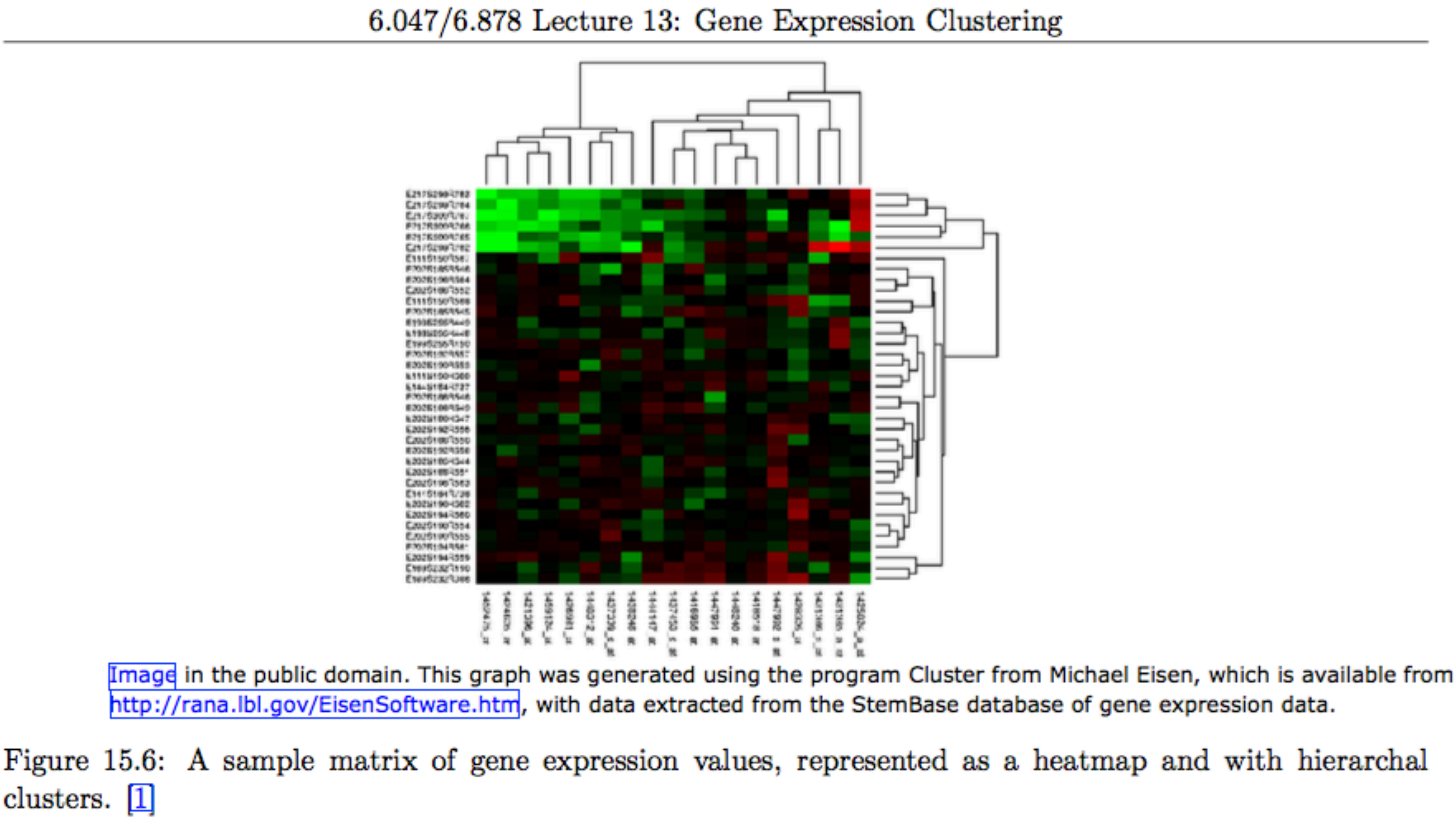
Structure in High-Dimensional Data



Topologically Clean Distance Fields. A. Gyulassy, V. Natarajan, Mark Duchaineau , Valerio Pascucci, Eduardo M. Bringa, Andrew Higginbotham, and Bernd Hamann, IEEE Transactions on Visualization and Computer Graphics.

- Structure can be used to reduce dimensionality of data
- Structure can tell us something useful about the underlying phenomena
- Structure can be used to make inferences about new data

MOTIVATION: GENE EXPRESSION CLUSTERING



6.047/6.878 Lecture 13: Gene Expression Clustering

Image in the public domain. This graph was generated using the program Cluster from Michael Eisen, which is available from http://rana.lbl.gov/EisenSoftware.htm, with data extracted from the StemBase database of gene expression data.

Figure 15.6: A sample matrix of gene expression values, represented as a heatmap and with hierarchal clusters. [1]

DEFINITION

Data clustering aims to extract the natural structure of a set of data

Given ***N*** objects, they are clustered into ***K*** groups so that objects belonging to the same group are more "similar" than objects of different groups

Objects can be D-dimensional vectors $\vec{x}_i = \{x_i(d)\}$

- There's not a unique definition of similarity
- The number of cluster is not fixed and depends on the level of knowledge of objects

→ Clustering is an ill-posed problem

15.3.1 *K*-Means Clustering

The k-means algorithm clusters n objects based on their attributes into k partitions. This is an example of partitioning, where each point is assigned to exactly one cluster such that the sum of distances from each point to its correspondingly labeled center is minimized. The motivation underlying this process is to make the most compact clusters possible, usually in terms of a Euclidean distance metric.

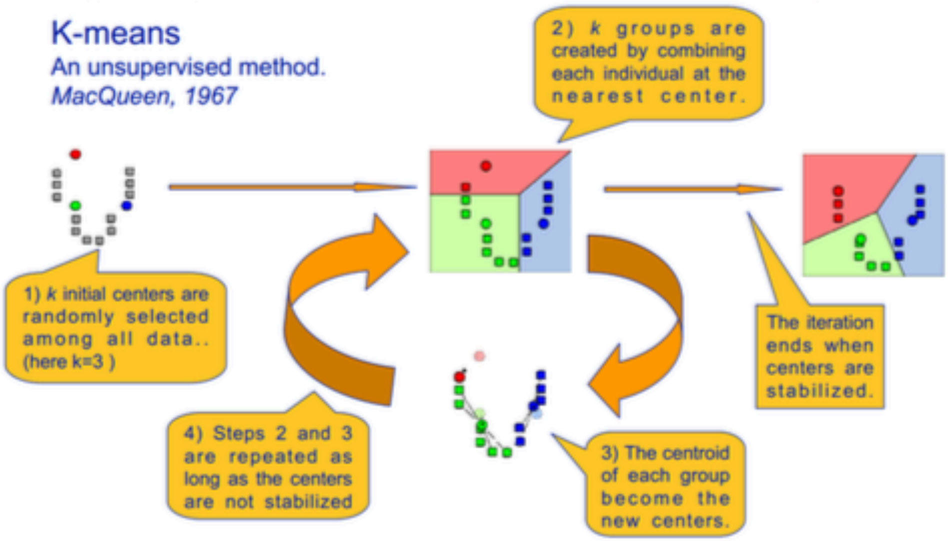

K-means
An unsupervised method.
MacQueen, 1967
1) k initial centers are randomly selected among all data.. (here k=3)
2) k groups are created by combining each individual at the nearest center.
3) The centroid of each group become the new centers.
4) Steps 2 and 3 are repeated as long as the centers are not stabilized
The iteration ends when centers are stabilized.

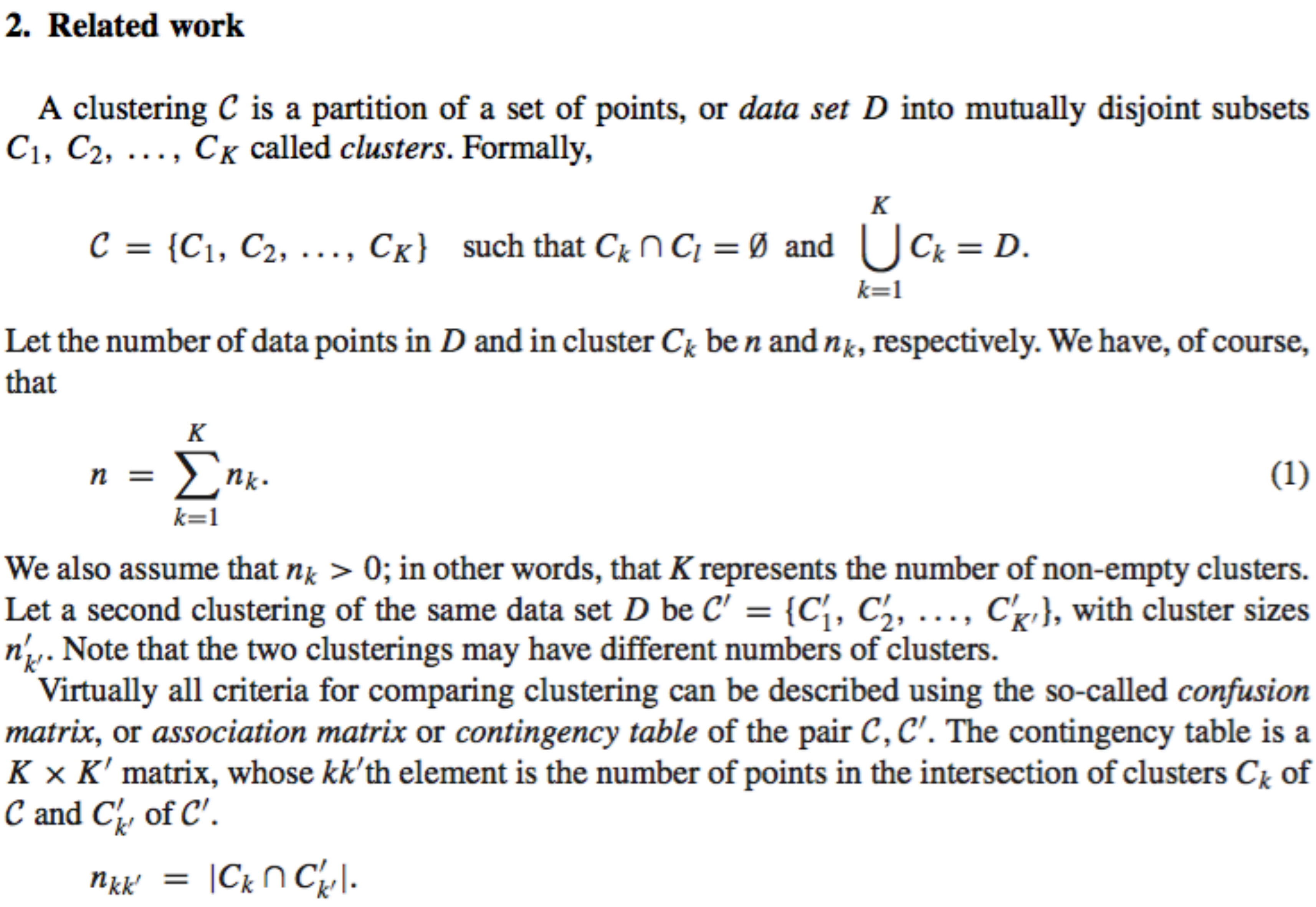
Figure 15.8: The k-means clustering algorithm

The k-means algorithm, as illustrated in figure 15.8, is implemented as follows:

1. Assume a fixed number of clusters, k
2. *Initialization*: Randomly initialize the k means μ_k associated with the clusters and assign each data point x_i to the nearest cluster, where the distance between x_i and μ_k is given by $d_{i,k} = (x_i - \mu_k)^2$.
3. *Iteration*: Recalculate the centroid of the cluster given the points assigned to it: $\mu_k(n+1) = \sum\limits_{x_i \in k} \frac{x_i}{|x^k|}$ where x_k is the number of points with label k. Reassign data points to the k new centroids by the given distance metric. The new centers are effectively calculated to be the average of the points assigned to each cluster.
4. *Termination*: Iterate until convergence or until a user-specified number of iterations has been reached. Note that the iteration may be trapped at some local optima.

COMPARING CLUSTERINGS

(see Maila2007)



2. Related work

A clustering $\mathcal{C}$ is a partition of a set of points, or *data set* D into mutually disjoint subsets $C_1, C_2, \ldots, C_K$ called *clusters*. Formally,

$$\mathcal{C} = \{C_1, C_2, \ldots, C_K\} \quad \text{such that } C_k \cap C_l = \emptyset \text{ and } \bigcup_{k=1}^{K} C_k = D.$$

Let the number of data points in D and in cluster C_k be n and n_k, respectively. We have, of course, that

$$n = \sum_{k=1}^{K} n_k. \tag{1}$$

We also assume that $n_k > 0$; in other words, that K represents the number of non-empty clusters. Let a second clustering of the same data set D be $\mathcal{C}' = \{C'_1, C'_2, \ldots, C'_{K'}\}$, with cluster sizes $n'_{k'}$. Note that the two clusterings may have different numbers of clusters.

Virtually all criteria for comparing clustering can be described using the so-called *confusion matrix*, or *association matrix* or *contingency table* of the pair $\mathcal{C}, \mathcal{C}'$. The contingency table is a $K \times K'$ matrix, whose kk'th element is the number of points in the intersection of clusters C_k of $\mathcal{C}$ and $C'_{k'}$ of $\mathcal{C}'$.

$$n_{kk'} = |C_k \cap C'_{k'}|.$$

2.1. *Comparing clusterings by counting pairs*

An important class of criteria for comparing clusterings is based on counting the pairs of points on which two clusterings agree/disagree. A pair of points from D can fall under one of four cases described below.

N_{11}	the number of point pairs that are in the same cluster under both $\mathcal{C}$ and $\mathcal{C}'$
N_{00}	number of point pairs in different clusters under both $\mathcal{C}$ and $\mathcal{C}'$
N_{10}	number of point pairs in the same cluster under $\mathcal{C}$ but not under $\mathcal{C}'$
N_{01}	number of point pairs in the same cluster under $\mathcal{C}'$ but not under $\mathcal{C}$

The four counts always satisfy

$$N_{11} + N_{00} + N_{10} + N_{01} \;=\; n(n-1)/2.$$

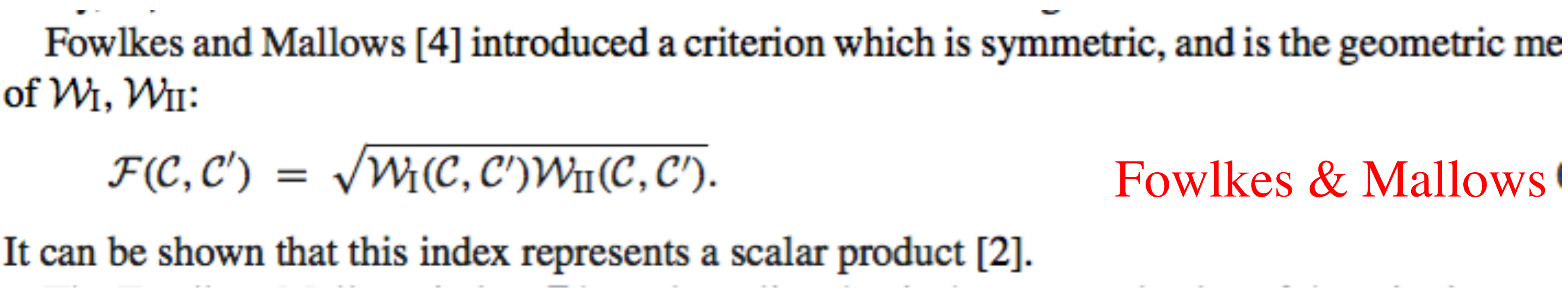
They can be obtained from the contingency table $[n_{kk'}]$. For example $2N_{11} = \sum_{k,k'} n_{kk'}^2 - n$.

Wallace [20] proposed the two asymmetric criteria $\mathcal{W}_\mathrm{I}$, $\mathcal{W}_\mathrm{II}$ below:

$$\mathcal{W}_\mathrm{I}(\mathcal{C},\mathcal{C}') \;=\; \frac{N_{11}}{\sum_k n_k(n_k-1)/2}, \tag{2}$$

$$\mathcal{W}_\mathrm{II}(\mathcal{C},\mathcal{C}') \;=\; \frac{N_{11}}{\sum_{k'} n'_{k'}(n'_{k'}-1)/2}. \tag{3}$$

Wallace criteria

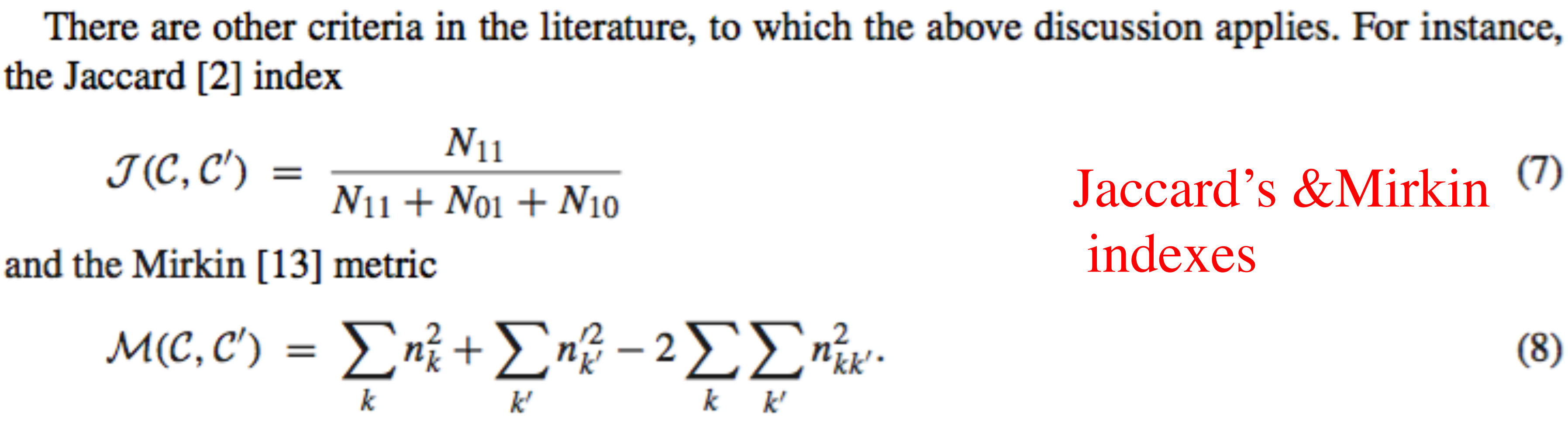


Fowlkes and Mallows [4] introduced a criterion which is symmetric, and is the geometric me
of $\mathcal{W}_{\mathrm{I}}, \mathcal{W}_{\mathrm{II}}$:

$$\mathcal{F}(\mathcal{C}, \mathcal{C}') = \sqrt{\mathcal{W}_{\mathrm{I}}(\mathcal{C}, \mathcal{C}')\mathcal{W}_{\mathrm{II}}(\mathcal{C}, \mathcal{C}')}.$$

It can be shown that this index represents a scalar product [2].

Fowlkes & Mallows



There are other criteria in the literature, to which the above discussion applies. For instance, the Jaccard [2] index

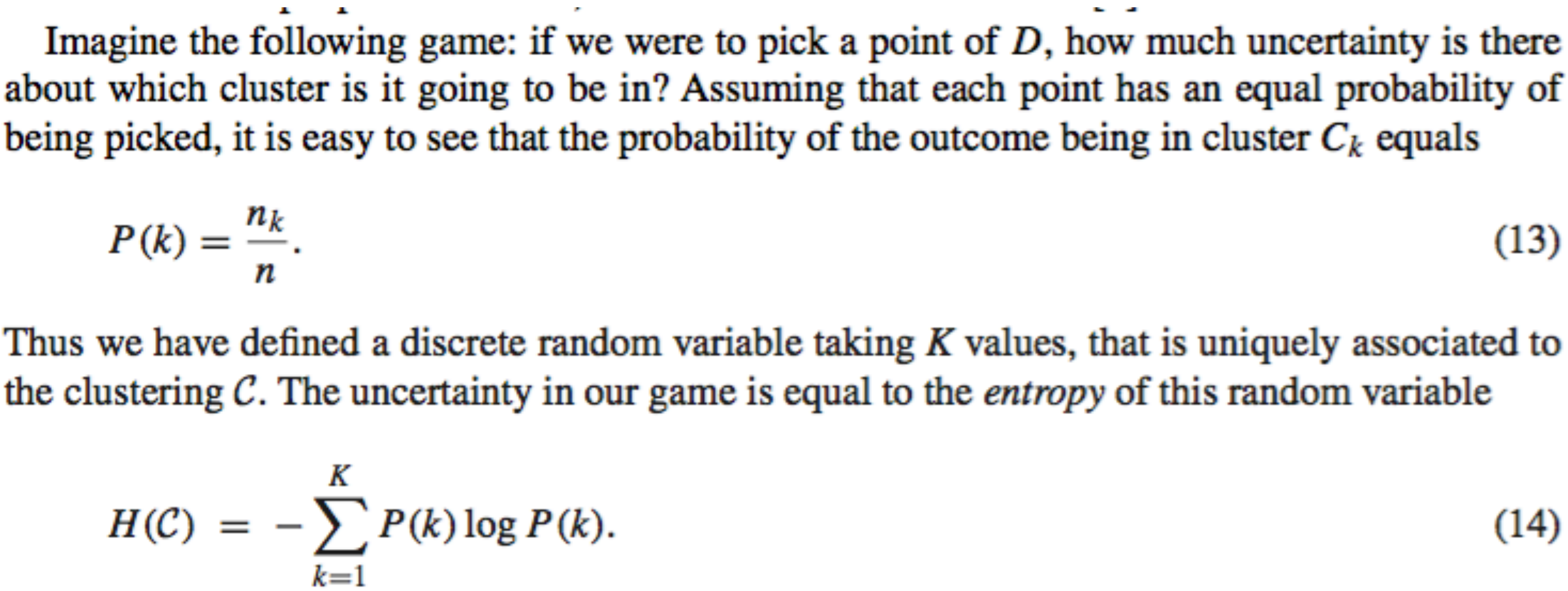
$$\mathcal{J}(\mathcal{C}, \mathcal{C}') = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \quad (7)$$

and the Mirkin [13] metric

$$\mathcal{M}(\mathcal{C}, \mathcal{C}') = \sum_k n_k^2 + \sum_{k'} n_{k'}'^2 - 2 \sum_k \sum_{k'} n_{kk'}^2. \quad (8)$$

Jaccard's &Mirkin
indexes

INFORMATION THEORETIC APPROACH



Imagine the following game: if we were to pick a point of D, how much uncertainty is there about which cluster is it going to be in? Assuming that each point has an equal probability of being picked, it is easy to see that the probability of the outcome being in cluster C_k equals

$$P(k) = \frac{n_k}{n}. \tag{13}$$

Thus we have defined a discrete random variable taking K values, that is uniquely associated to the clustering $\mathcal{C}$. The uncertainty in our game is equal to the *entropy* of this random variable

$$H(\mathcal{C}) = -\sum_{k=1}^{K} P(k) \log P(k). \tag{14}$$

relative proportions of the clusters.

We now define the *mutual information* between two clusterings, i.e. the information that one clustering has about the other. Denote by $P(k)$, $k = 1, \ldots, K$ and $P'(k')$, $k' = 1, \ldots, K'$ the random variables associated with the clusterings $\mathcal{C}$, $\mathcal{C}'$. Let $P(k, k')$ represent the probability that a point belongs to C_k in clustering $\mathcal{C}$ and to $C'_{k'}$ in $\mathcal{C}'$, namely the joint distribution of the random variables associated with the two clusterings:

$$P(k, k') = \frac{|C_k \bigcap C'_{k'}|}{n}. \tag{15}$$

We define $I(\mathcal{C}, \mathcal{C}')$ the mutual information between the clusterings $\mathcal{C}$, $\mathcal{C}'$ to be equal to the mutual information between the associated random variables

$$I(\mathcal{C}, \mathcal{C}') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') \log \frac{P(k, k')}{P(k) P'(k')}. \tag{16}$$

Intuitively, we can think of $I(\mathcal{C}, \mathcal{C}')$ in the following way: we are given a random point in D. The uncertainty about its cluster in $\mathcal{C}'$ is measured by $H(\mathcal{C}')$. Suppose now that we are told which cluster the point belongs to in $\mathcal{C}$. How much does this knowledge reduce the uncertainty about $\mathcal{C}'$? This reduction in uncertainty, averaged over all points, is equal to $I(\mathcal{C}, \mathcal{C}')$.

The mutual information between two random variables is always non-negative and symmetric:

$$I(\mathcal{C}, \mathcal{C}') = I(\mathcal{C}', \mathcal{C}) \geqslant 0. \tag{17}$$

Also, the mutual information can never exceed the total uncertainty in a clustering, so

$$I(\mathcal{C}, \mathcal{C}') \leqslant \min(H(\mathcal{C}),\ H(\mathcal{C}')). \tag{18}$$

Equality in the above formula occurs when one clustering completely determines the other. For example, if $\mathcal{C}'$ is obtained from $\mathcal{C}$ by merging two or more clusters, then

$$I(\mathcal{C}, \mathcal{C}') = H(\mathcal{C}') < H(\mathcal{C}).$$

When the two clusterings are equal, and only then, we have

$$I(\mathcal{C}, \mathcal{C}') = H(\mathcal{C}') = H(\mathcal{C}).$$

We propose to use as a comparison criterion for two clusterings $\mathcal{C}, \mathcal{C}'$ the quantity

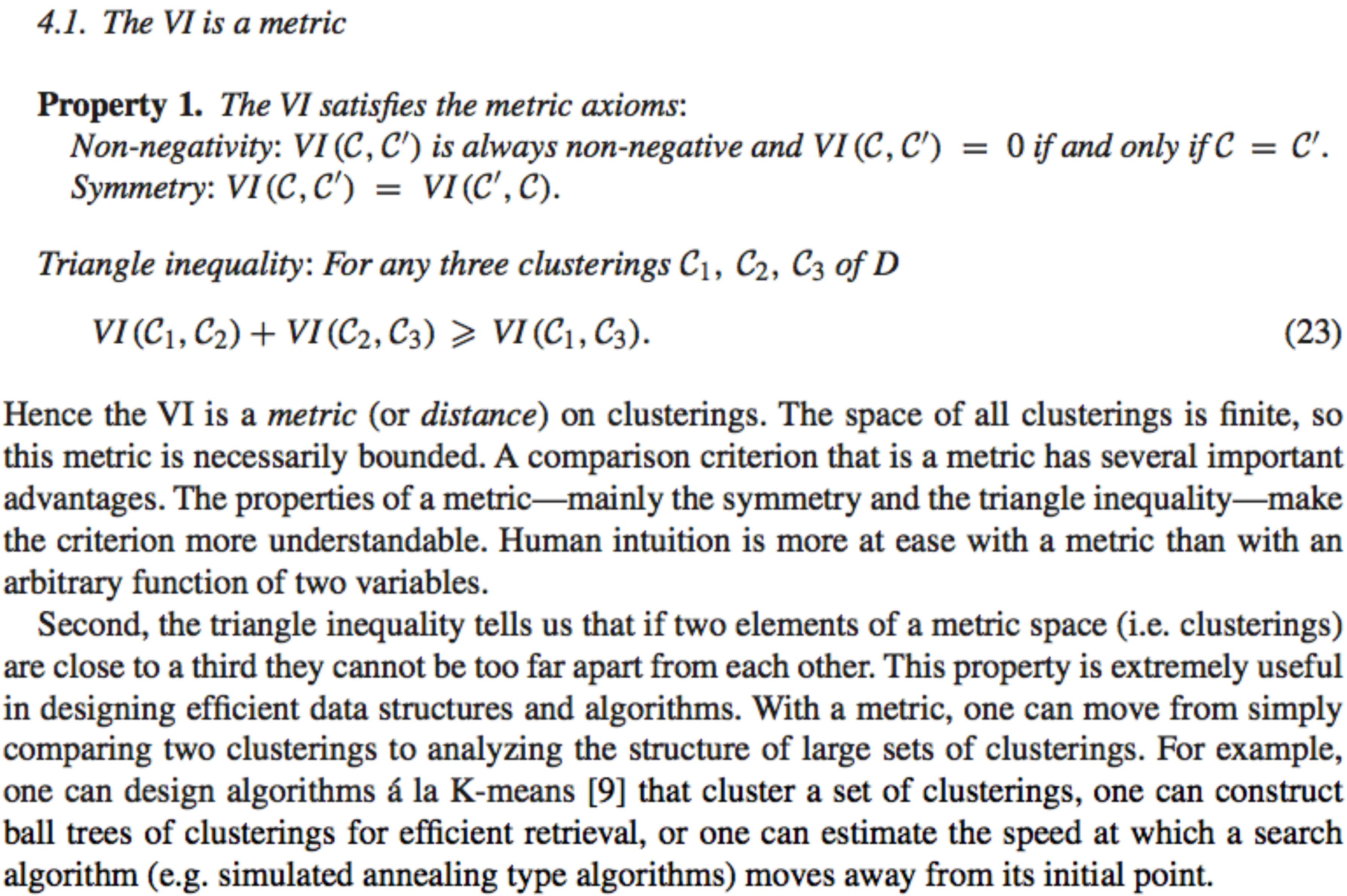
$$VI(\mathcal{C}, \mathcal{C}') = H(\mathcal{C}) + H(\mathcal{C}') - 2I(\mathcal{C}, \mathcal{C}'). \tag{19}$$

At a closer examination, this is the sum of two positive terms

$$VI(\mathcal{C}, \mathcal{C}') = [H(\mathcal{C}) - I(\mathcal{C}, \mathcal{C}')] + [H(\mathcal{C}') - I(\mathcal{C}, \mathcal{C}')]. \tag{20}$$

By analogy with the total variation of a function, we call it VI between the two clusterings. The two terms represent the conditional entropies $H(\mathcal{C}|\mathcal{C}')$, $H(\mathcal{C}'|\mathcal{C})$. The first term measures the amount

VI inducesc a metric in the space of clusterings



4.1. The VI is a metric

Property 1. *The VI satisfies the metric axioms*:

Non-negativity: $VI(\mathcal{C}, \mathcal{C}')$ *is always non-negative and* $VI(\mathcal{C}, \mathcal{C}') = 0$ *if and only if* $\mathcal{C} = \mathcal{C}'$.

Symmetry: $VI(\mathcal{C}, \mathcal{C}') = VI(\mathcal{C}', \mathcal{C})$.

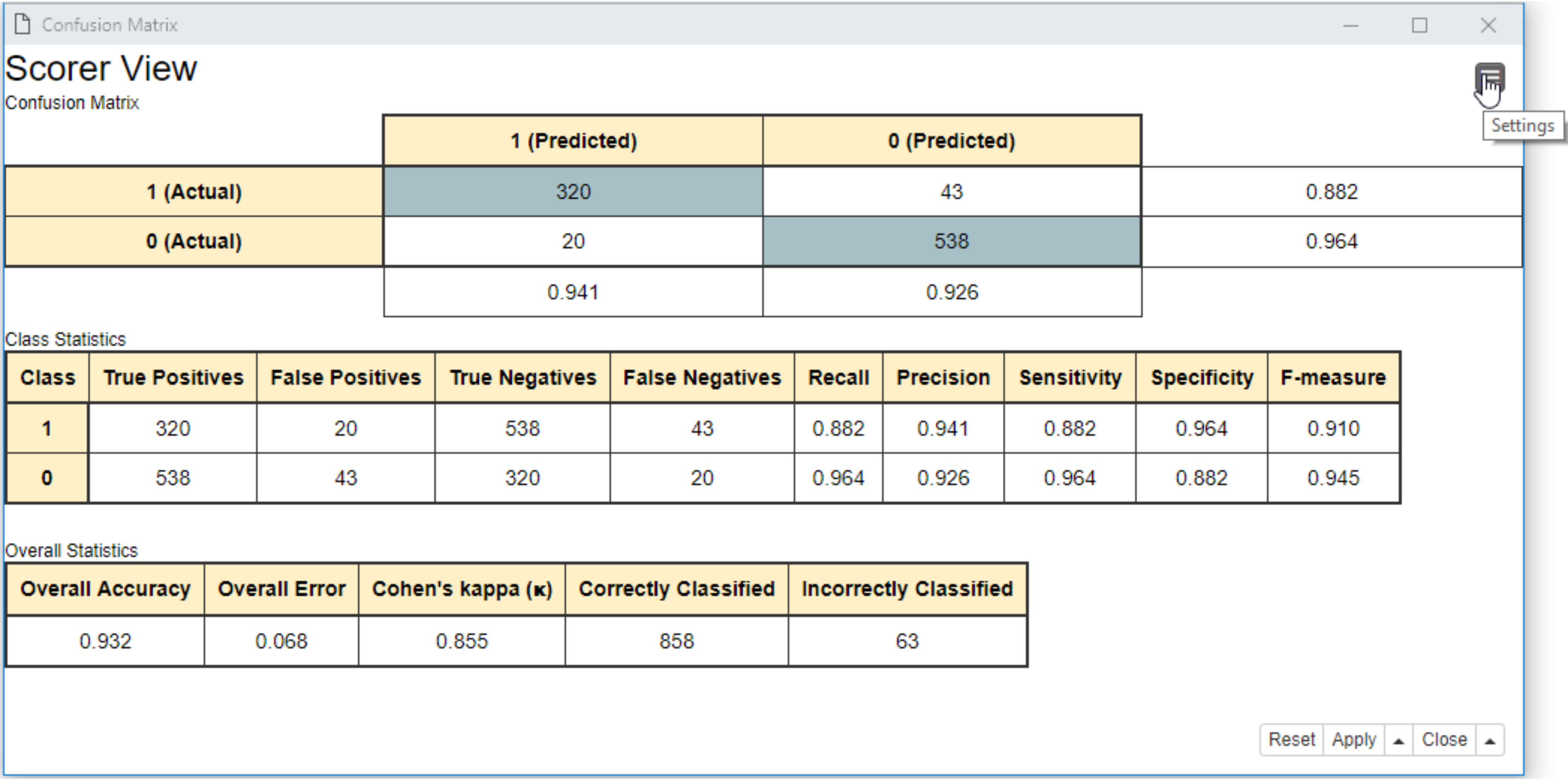
Triangle inequality: *For any three clusterings* $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$ *of* D

$$VI(\mathcal{C}_1, \mathcal{C}_2) + VI(\mathcal{C}_2, \mathcal{C}_3) \geqslant VI(\mathcal{C}_1, \mathcal{C}_3). \tag{23}$$

Hence the VI is a *metric* (or *distance*) on clusterings. The space of all clusterings is finite, so this metric is necessarily bounded. A comparison criterion that is a metric has several important advantages. The properties of a metric—mainly the symmetry and the triangle inequality—make the criterion more understandable. Human intuition is more at ease with a metric than with an arbitrary function of two variables.

Second, the triangle inequality tells us that if two elements of a metric space (i.e. clusterings) are close to a third they cannot be too far apart from each other. This property is extremely useful in designing efficient data structures and algorithms. With a metric, one can move from simply comparing two clusterings to analyzing the structure of large sets of clusterings. For example, one can design algorithms á la K-means [9] that cluster a set of clusterings, one can construct ball trees of clusterings for efficient retrieval, or one can estimate the speed at which a search algorithm (e.g. simulated annealing type algorithms) moves away from its initial point.

CONFUSION MATRICES& MODEL PERFORMACE TESTYS



Scorer View

Confusion Matrix

	1 (Predicted)	0 (Predicted)	
1 (Actual)	320	43	0.882
0 (Actual)	20	538	0.964
	0.941	0.926	

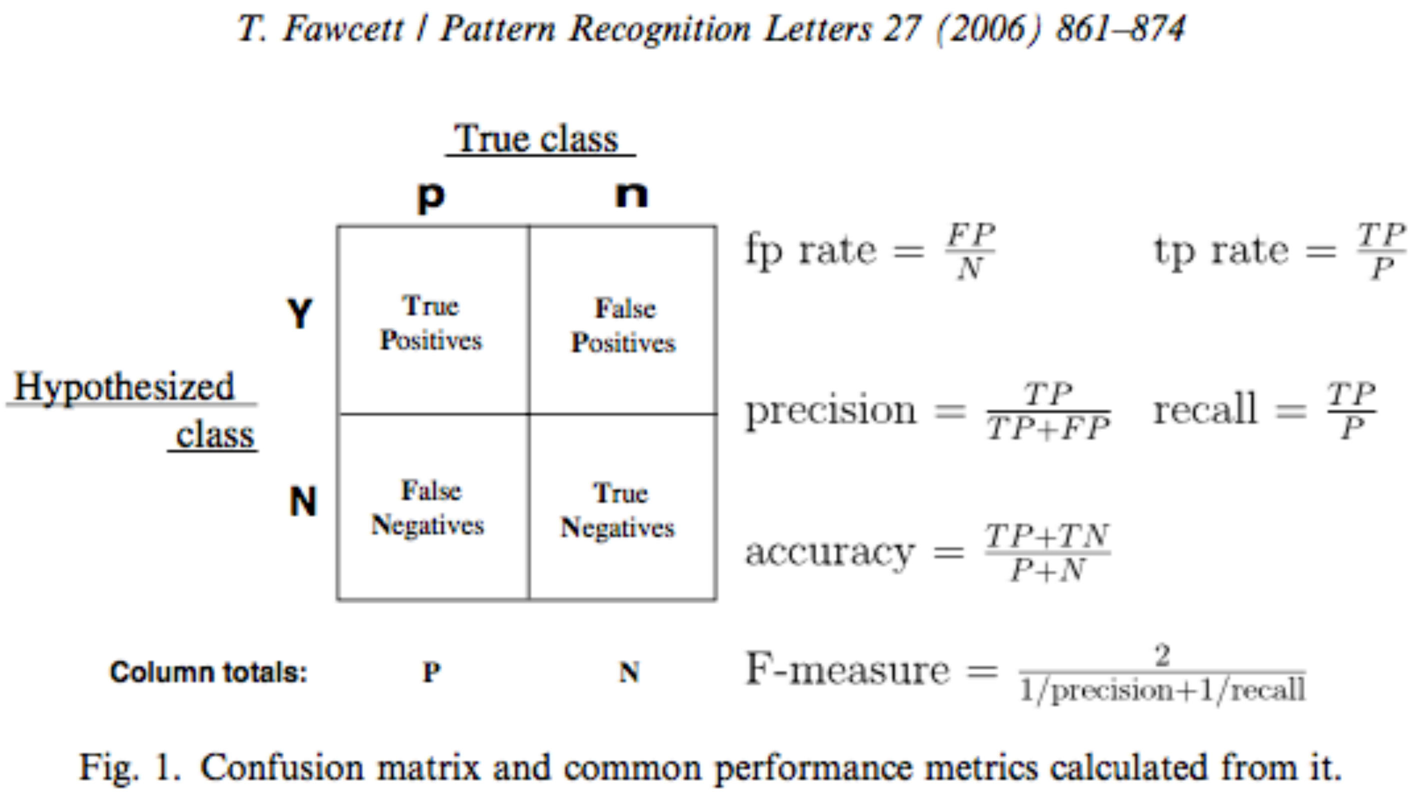
Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
1	320	20	538	43	0.882	0.941	0.882	0.964	0.910
0	538	43	320	20	0.964	0.926	0.964	0.882	0.945

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
0.932	0.068	0.855	858	63

https://towardsdatascience.com/confusion-matrix-and-class-statistics-68b79f4f510b



T. Fawcett / Pattern Recognition Letters 27 (2006) 861–874

Hypothesized class	True class: **p**	True class: **n**
Y	True Positives	False Positives
N	False Negatives	True Negatives
Column totals:	**P**	**N**

$$\text{fp rate} = \frac{FP}{N} \qquad \text{tp rate} = \frac{TP}{P}$$

$$\text{precision} = \frac{TP}{TP+FP} \qquad \text{recall} = \frac{TP}{P}$$

$$\text{accuracy} = \frac{TP+TN}{P+N}$$

$$\text{F-measure} = \frac{2}{1/\text{precision}+1/\text{recall}}$$

Fig. 1. Confusion matrix and common performance metrics calculated from it.