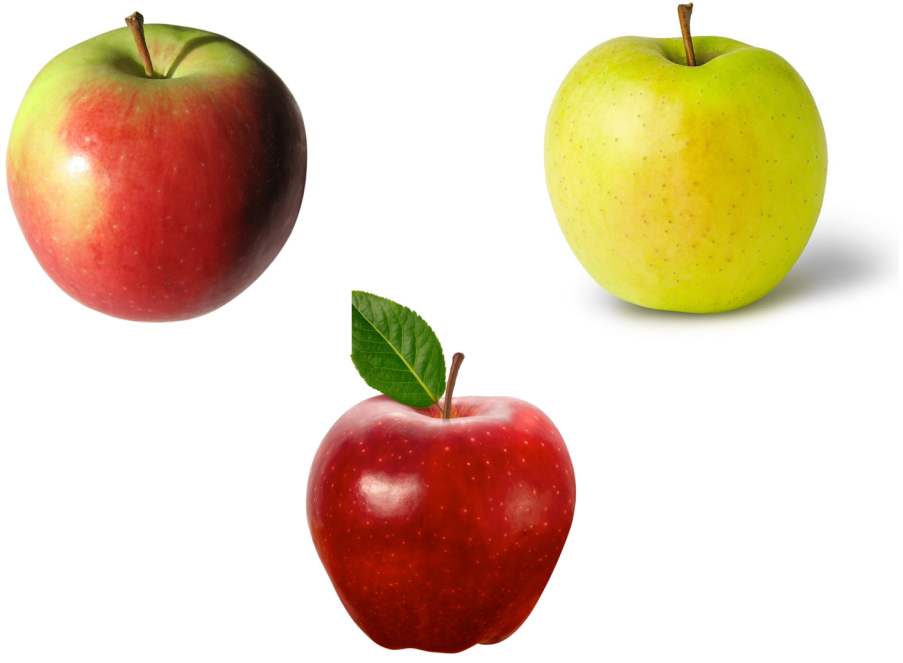
Clustering Validation

Cluster Validation – Part 1

INTRODUCTION

Clustering

- In machine learning, ***clustering*** is defined as grouping objects based on their **over-all** similarity (or dissimilarity) to each other
- Note that each object has *multiple dimensions*, or attributes available for comparison
- It's tempting to focus on just one or two attributes, but that is typically **not** what we are doing in (machine learning) clustering!
- When we cluster, even if we were to focus on one particular attribute, all of the other attributes would still come along for the ride



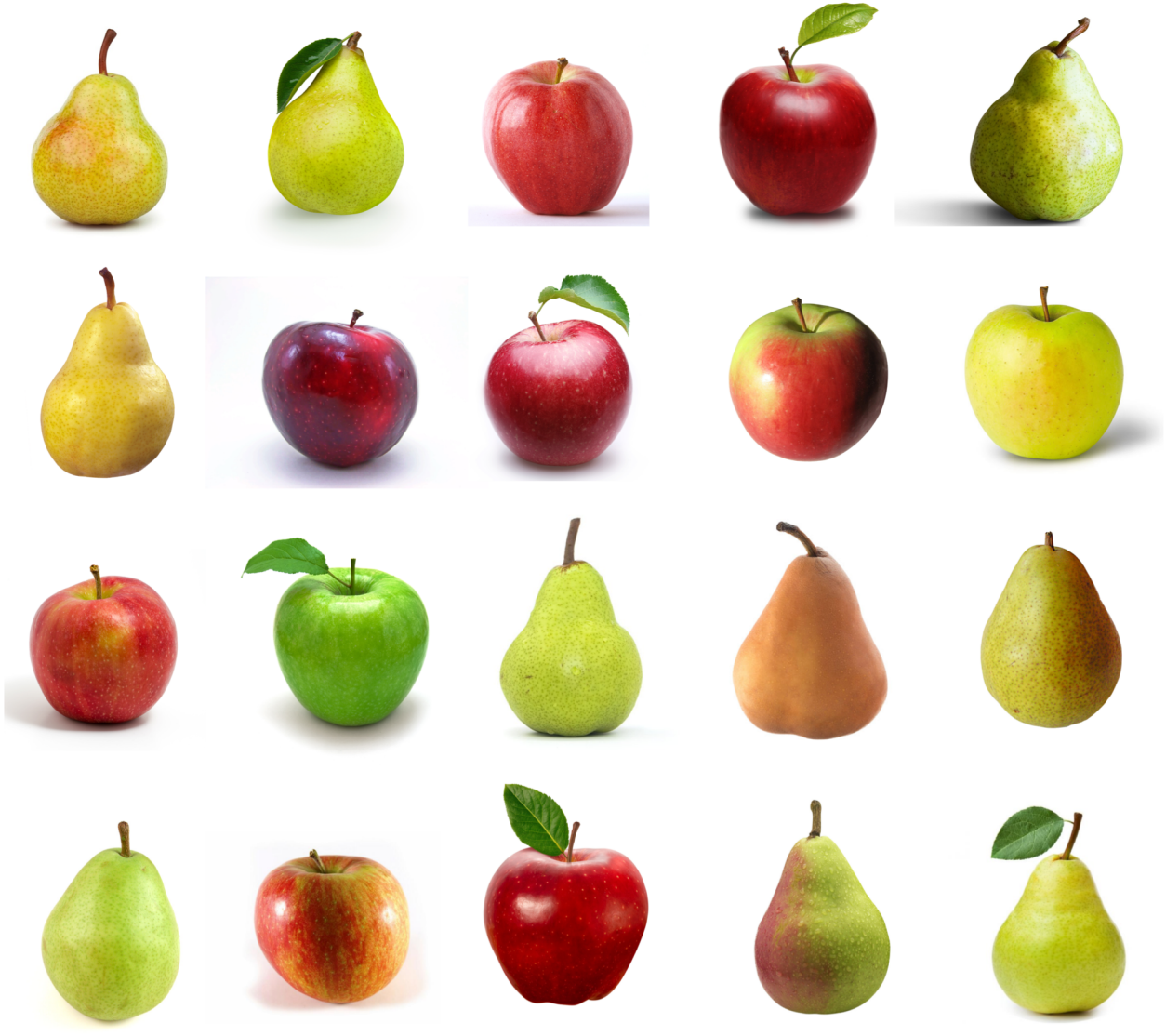
What is the same about these objects?
What is different?
Do they belong in the same group?
How many groups? How many classes?

Fruit Image Dataset

- 20 images of fruit
- Are there right or wrong groupings of this dataset?
- Are there multiple possible ‘natural’ clusterings?
- Could different clusterings be used differently?
- Will some clusterings be of (objectively) higher ***quality*** than others?

Making Concepts Concrete

- To appreciate clustering validation, it helps to relate the concepts to something tangible
- In what follows, take the time to think about how the presented concepts can be related to the images from this small dataset

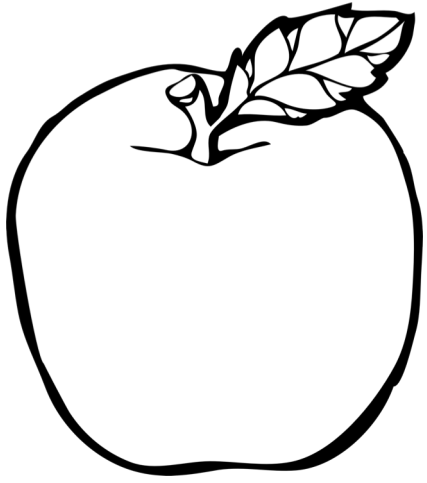


Clustering Validation – Part 2

KEY CONCEPTS ILLUSTRATED

Concept vs. Instance

- We group instances of objects into larger categories (clusters, classes, types)
- These larger categories can be represented by a concept, exemplar, representative or definition
- The concept (exemplar/definition) is a generalized representation - it captures something about all of the instances
- **For a given grouping – *can* we come up with a clear concept that captures the 'essence' of that grouping?**
- If yes, does that make it a good clustering?



Exemplar,
Concept,
Representative

Definition: "the fleshy, usually rounded red, yellow, or green edible pome fruit of a usually cultivated tree (genus Malus) of the rose family" Mirriam-Webster

Instances

Instance Properties

- For machine learning purposes, we represent properties of object instances using vectors
- Each vector element represents an attribute of the object.
- The value of the vector element represents the value of that property (e.g. the colour) of that object
- Vector Properties:
  - Length
    - (= number of dimensions/attributes)
  - For each dimension
    - Continuous/Discrete
    - Numeric/Categorical
    - Range/Possible Values

[12, 9.12, round, golden delicious]

Does this vector sufficiently describe this object?

Instance-Instance Relationships

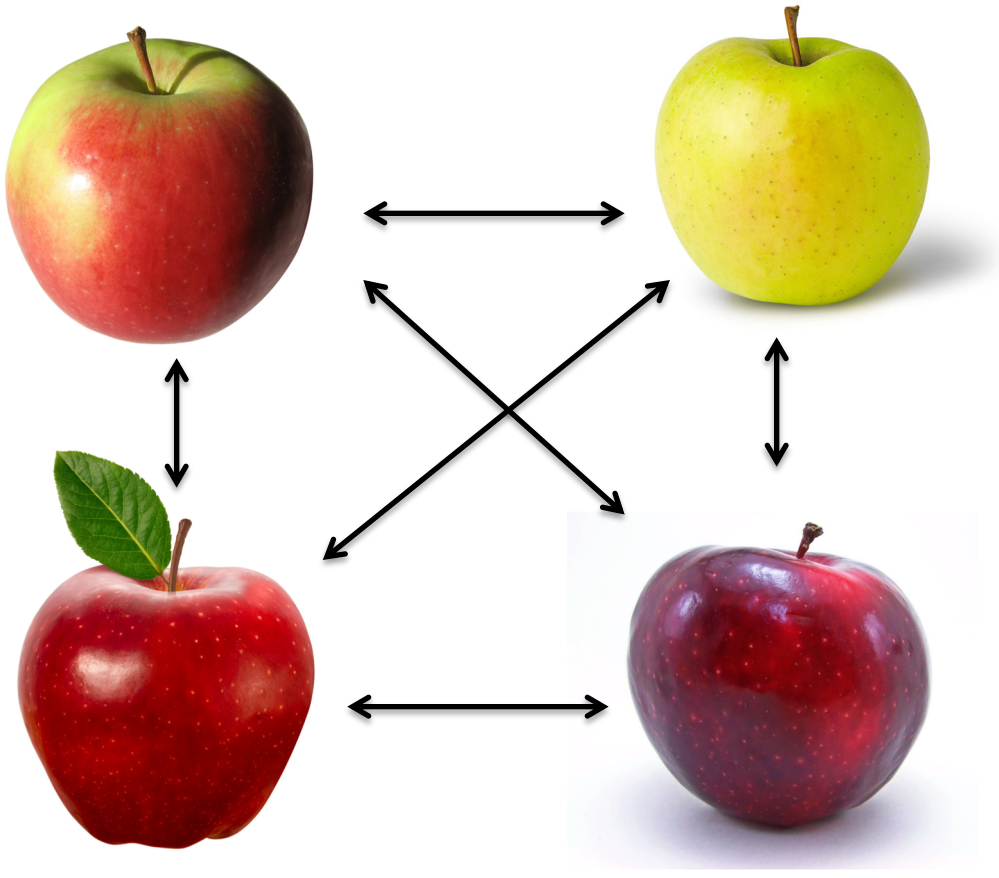
- Defined relationships between instances
- Comparison functions between instances:
  - Take as input vectors or parts of vectors
  - Might only take certain types of input (e.g. numeric)
  - Outputs a comparison result
- Similarity
  - Similarity as defined on a single dimension? Multiple dimensions?
  - Can we come up with functions that give us an overall similarity measure, across all dimensions?

[3, 10.43, round, macintosh]

[12, 9.12, round, golden delicious]

Distance

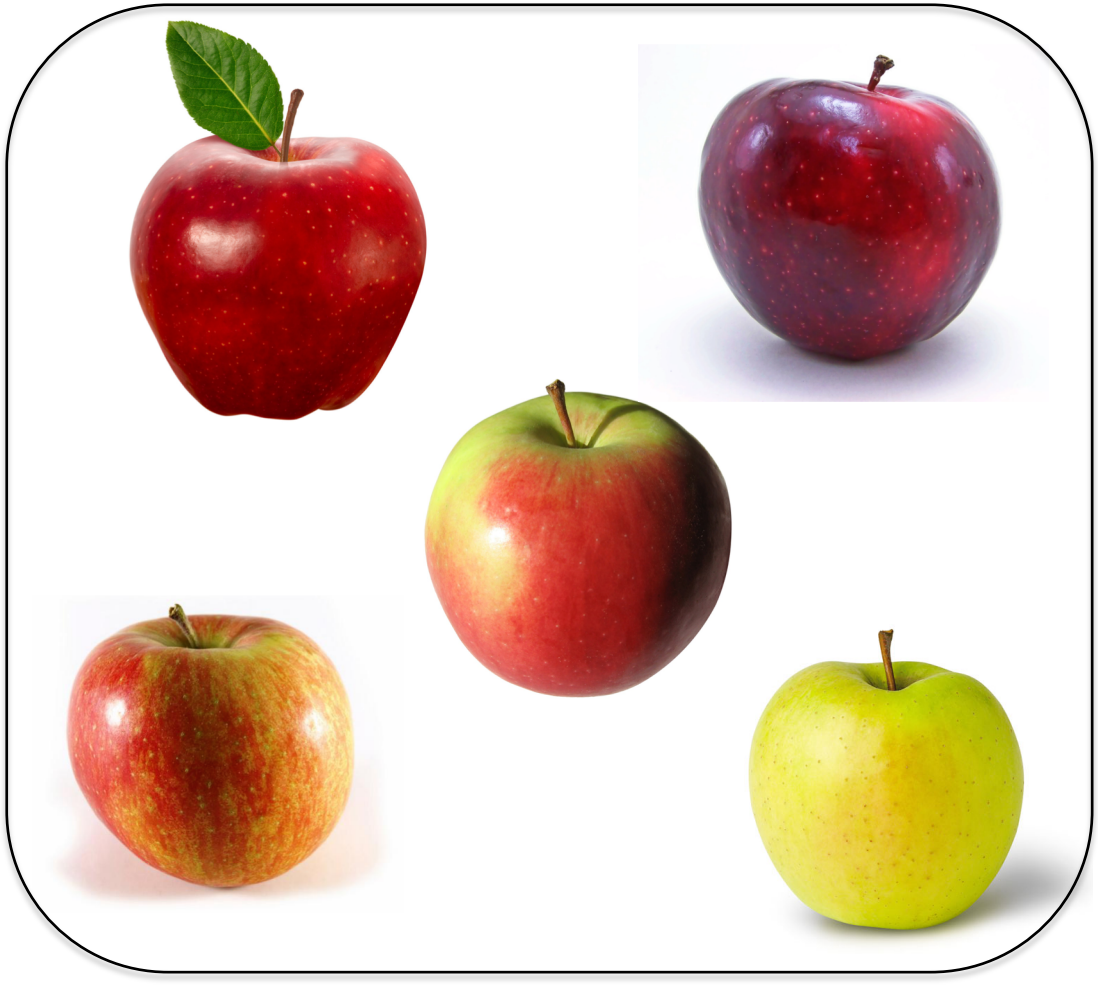
- Distance is a popular strategy for defining how similar to objects are to each other
- It is called distance because it is calculated in the same manner as Euclidean distance
- Importantly, distance takes into account *all* of the properties of the objects in question – it doesn't just focus on one or two
- Only numeric attributes are allowed as input, but it is technically possible to convert categorical attributes to numeric ones
- This only works as long as the categorical concepts are in some sense equidistant from each other, conceptually. Consider as an example where they are not - [apple, pear, vegetable].



How far apart are these apples?

Cluster Properties

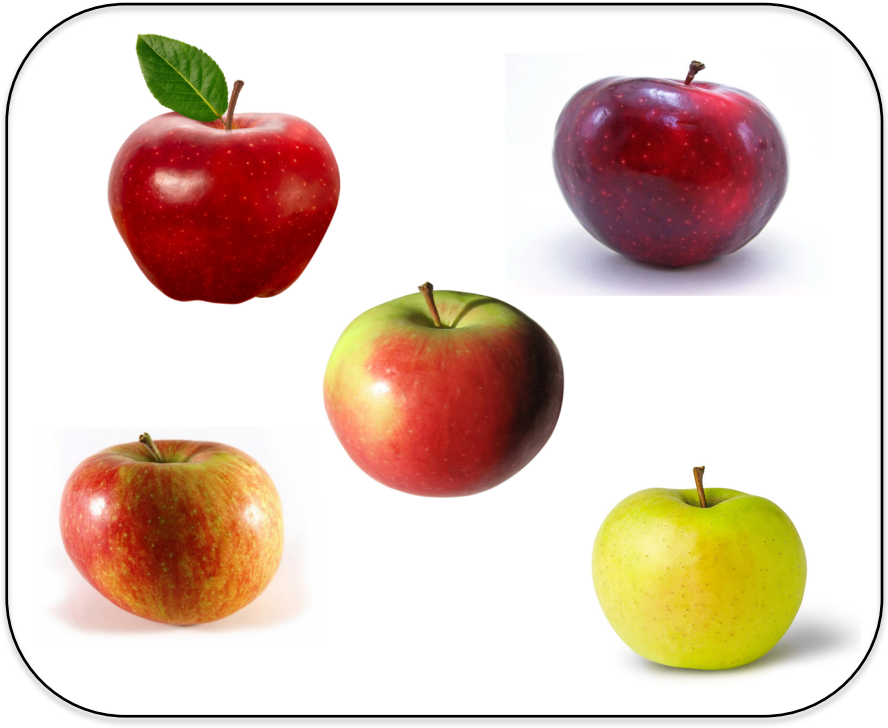
- Number of instances
- Similarity measures across instances within cluster
  - minimum similarity
  - maximum similarity
  - average similarity
- Cluster Representative:
  - may be an instance
  - may be an amalgamation of multiple instances (e.g. exemplar)



Which are the most similar? Which are the least? Which is the best representative?

Cluster – Instance Relationship

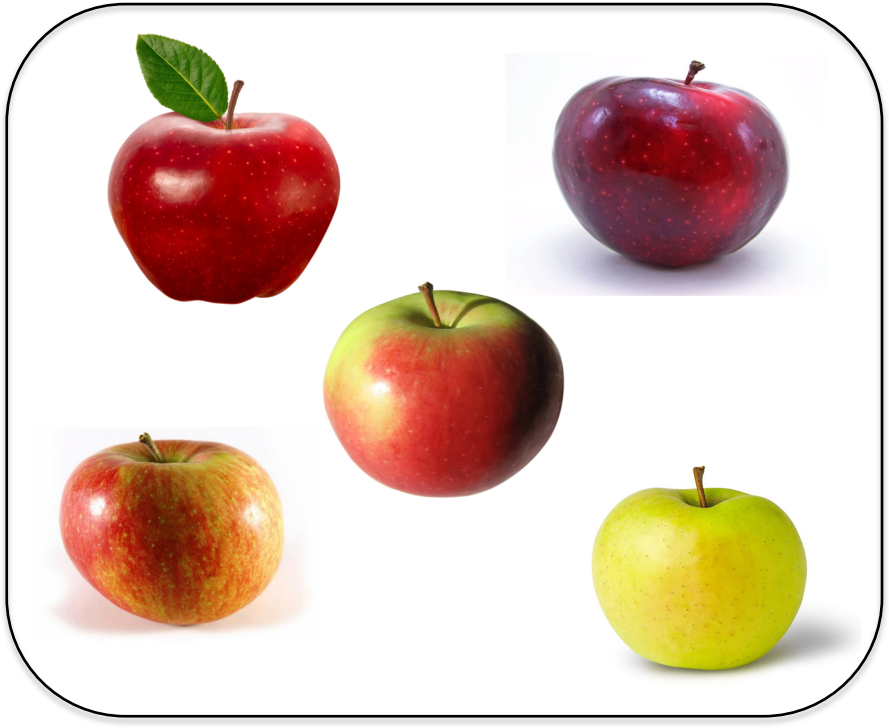
- Comparison of instance to cluster
- Might compare with representative instance
- See also instance instance relationships for comparison between the instance and specific instances within the cluster:
  - instance with cluster instance the greatest distance away from it
  - instance with cluster instance that is most similar



Is this instance similar to this cluster? Does it belong in this cluster?

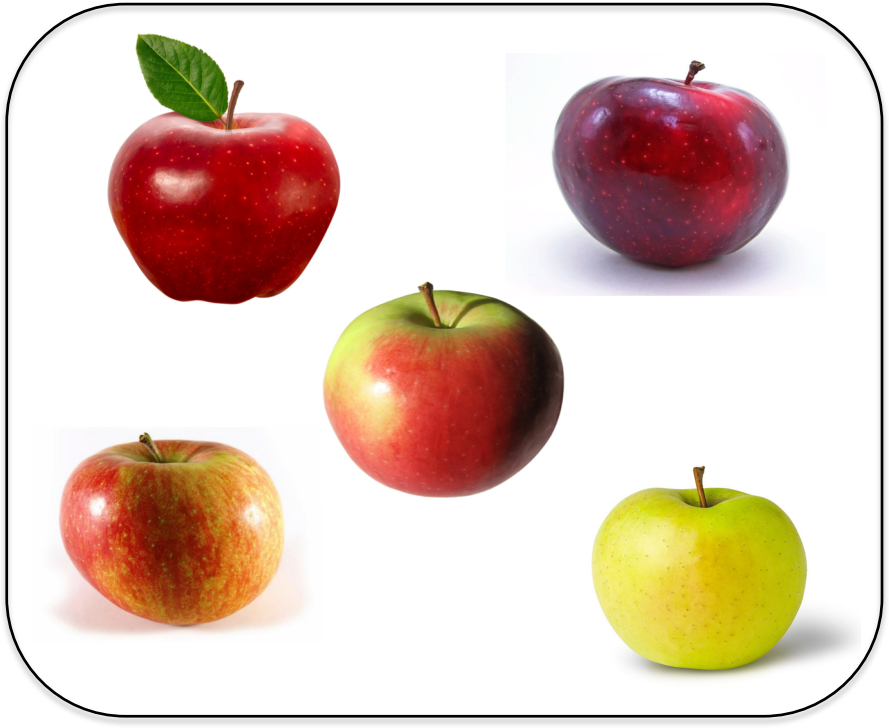
Cluster – Cluster Relationship

- Comparison of cluster level properties:
  - number of instances
  - max or min similarity
  - cluster representatives



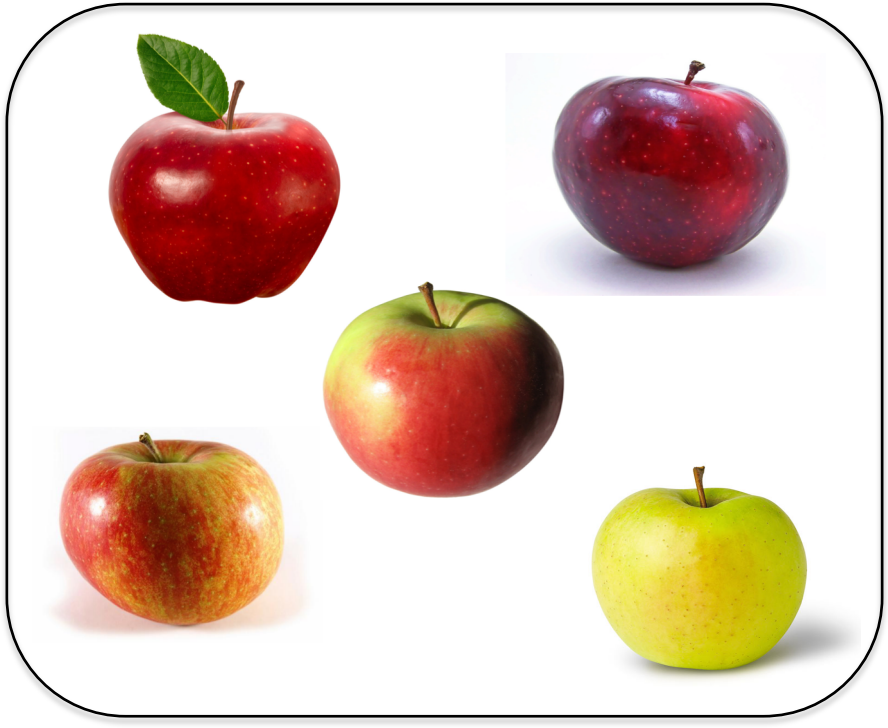
Comparisons Summary

- Comparison of cluster level properties
  - number of instances
  - max or min similarity
  - cluster representatives
- Comparison of cluster to instance properties
  - instance vector to cluster representative vector
- Comparison of instance to instance properties
  - similarity measures
- Comparisons may occur both within cluster and across clusters



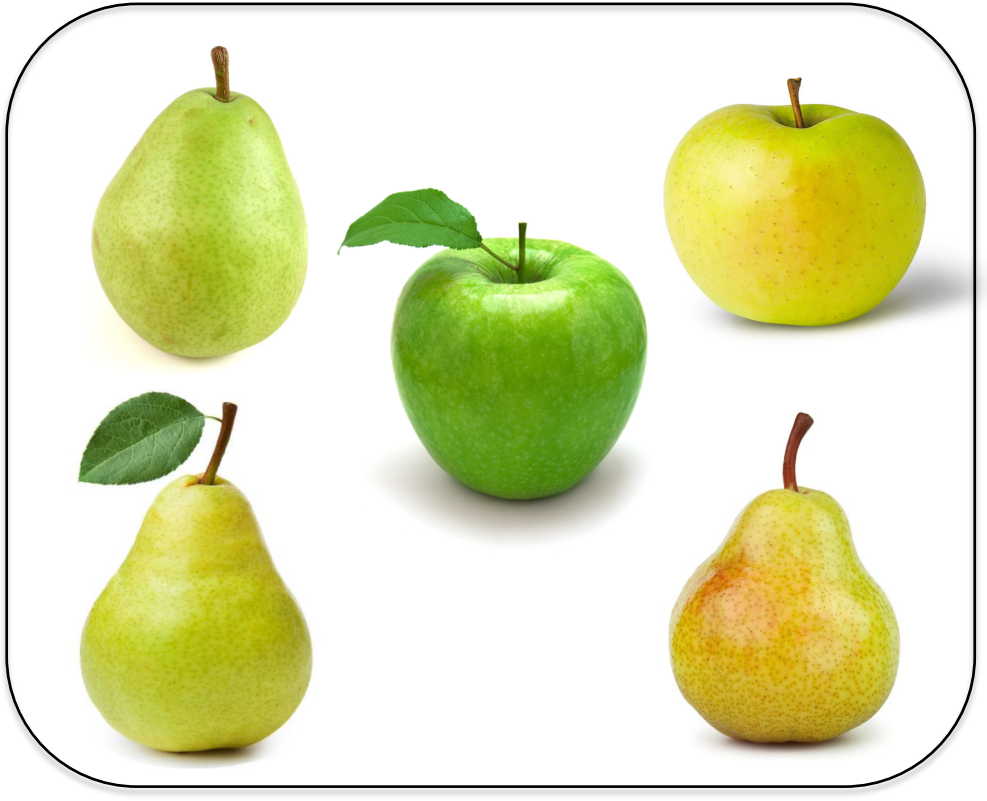
Getting to quality clusters

- Cluster and instance comparisons can be combined in many different ways.
- These can be used to generate a vast number of different cluster validation functions
- What do these tell us about the **quality** of a particular clustering:
  - relative to some objective criteria about good clustering schemes
  - relative to another clustering option
  - relative to external information (e.g. functionality, natural classes)



?

A Quality Clustering? Natural?



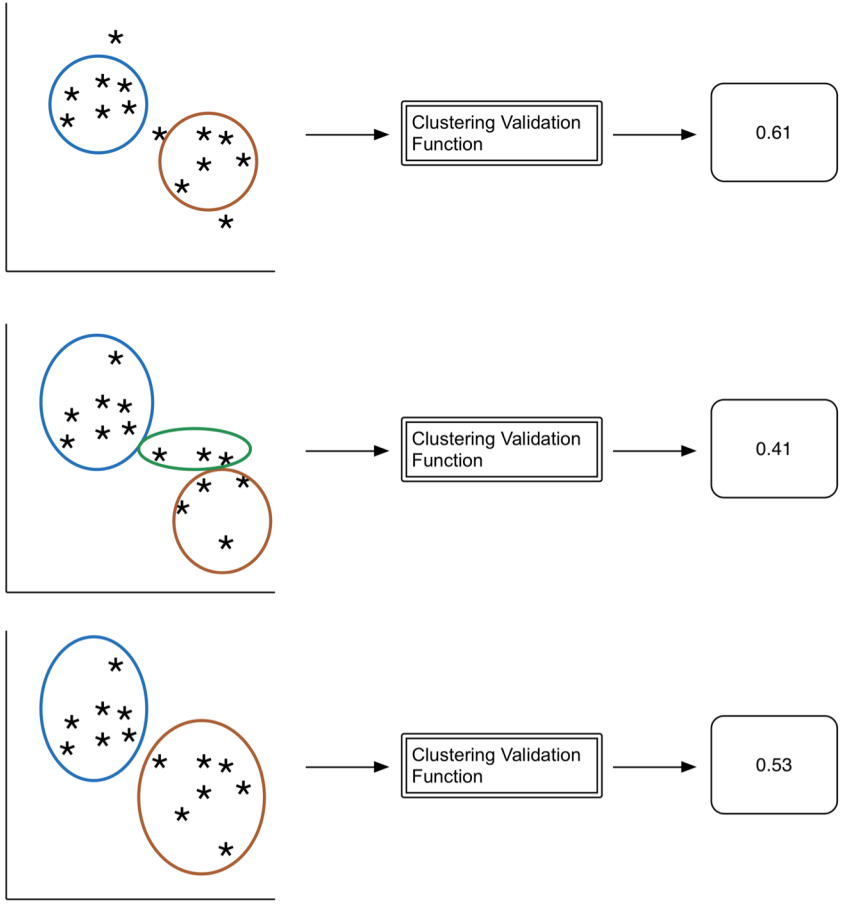
What level of quality is this clustering? Are there higher quality clusterings? Lower? How would you quantify this? Use some of the introduced concepts?

Clustering Validation – Part 3

TYPES OF CLUSTERING VALIDATION

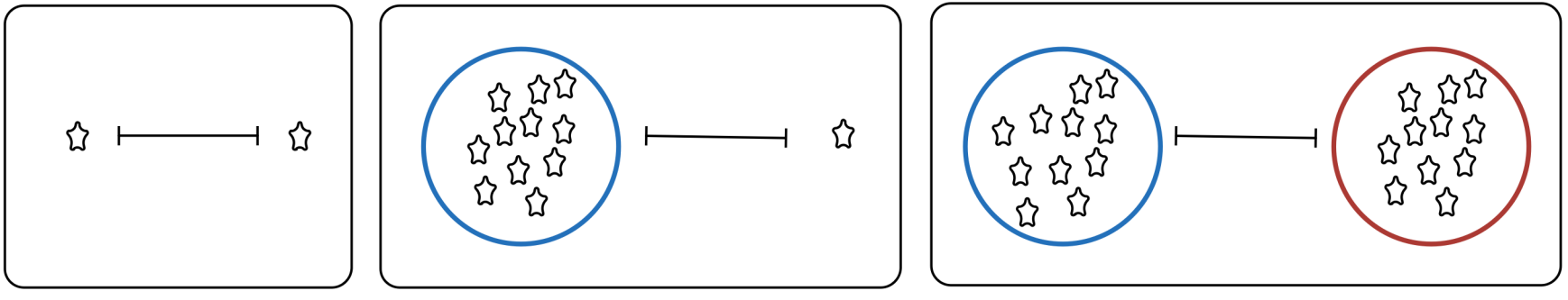
Clustering Operations

- Clustering involves two main activities
  - Creating clusters
  - Assessing cluster quality
- We create functions to carry out both of these activities
- Clustering functions
  - Input: Instances (vectors)
  - Output: Cluster assignment to each instance
- Assessing cluster quality
  - Input: Instances + Cluster Assignments (+ similarity matrix, usually)
  - Output: A numeric value



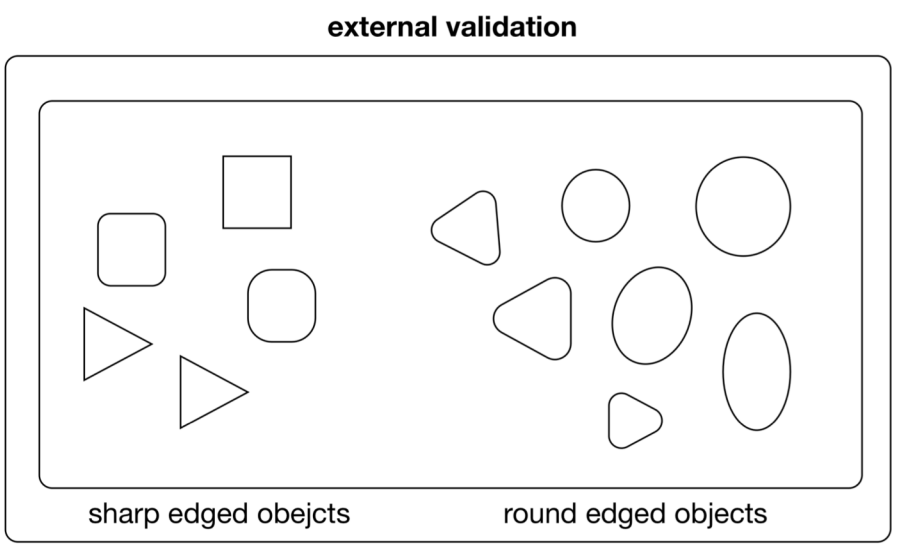
Clustering Validation Function Components

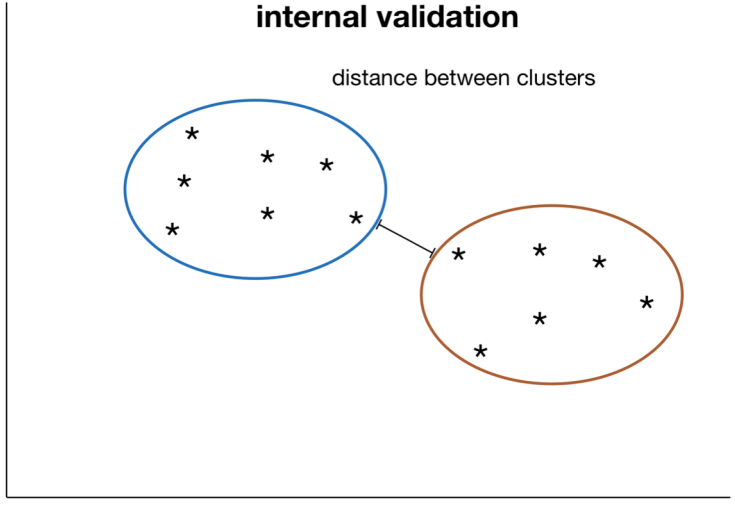
- There are a huge number of both of clustering and cluster validation functions
- However, all are built up out of the basic measures relating to instance or cluster properties we have already reviewed:
  - **Instance Properties**
  - **Cluster Properties**
  - **Instance – Instance relationship properties**
  - **Cluster – Instance Relationship Properties**
  - **Cluster – Cluster Relationship Properties**

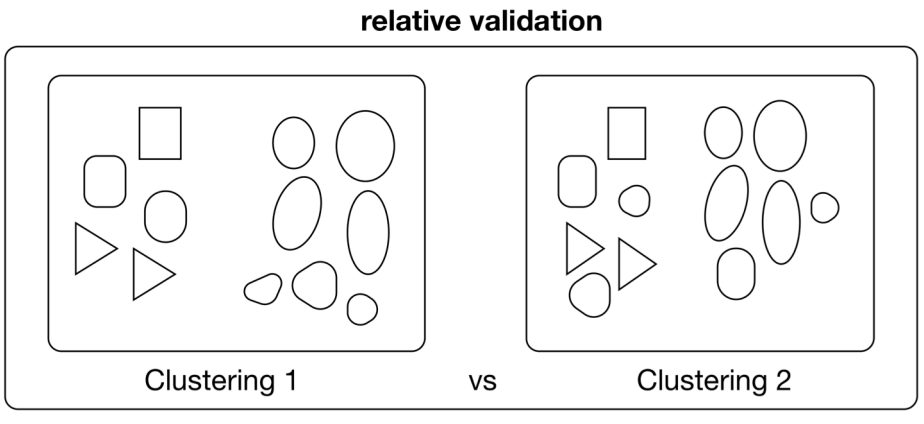


Three types of validation

- **Internal Validation**: Based only on properties available within a single clustering result (note that this comprises multiple clusters)
- **Relative Validation**: Comparison of one (entire) clustering result with another
- **External Validation**: Comparison of a (single) clustering result with some external standard





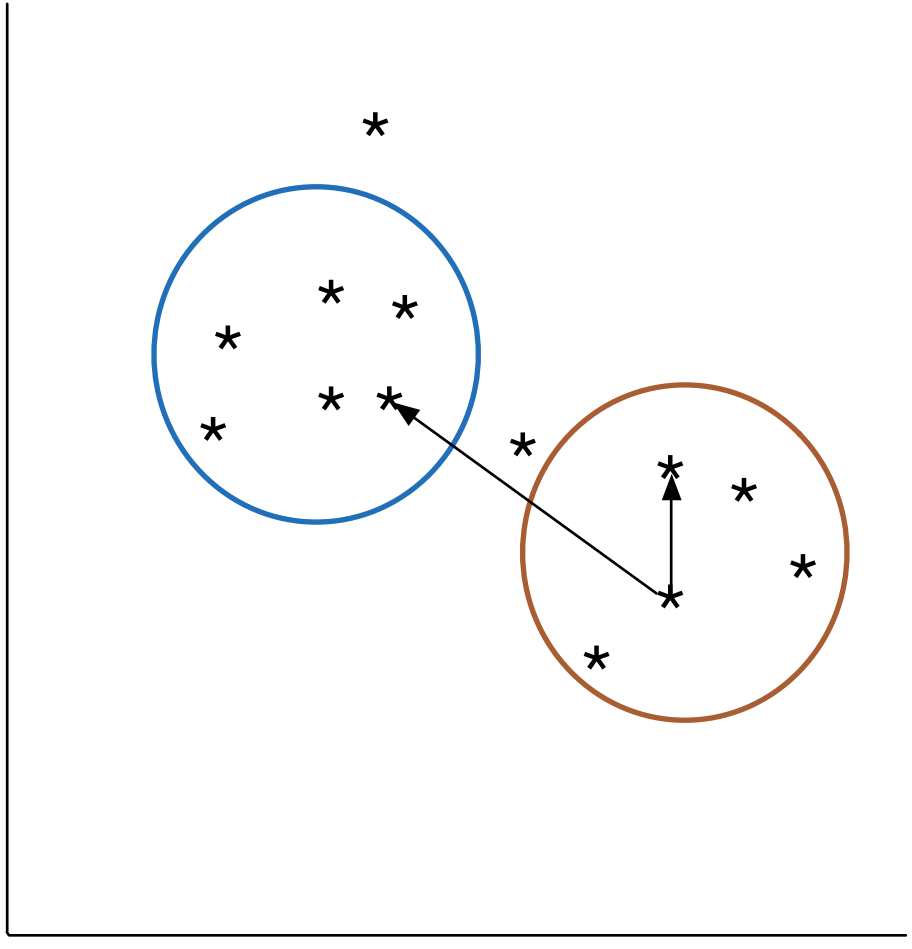


Clustering Validation – Part 4

INTERNAL VALIDATION

Validity vs. Quality

- Context is very relevant to the quality of a given clustering
- BUT what if we have no context?
- Is there a way to objectively measure cluster quality without any specific context?
- The term ‘validity’ suggests there is a **correct** clustering, and all we need to do is see how close we are to that
- Alternatively Lewis, Ackerman and de Sa (2012) use the term **Clustering Quality Measures** (CQM) instead



A (Small?) Sample of Internal CQMs

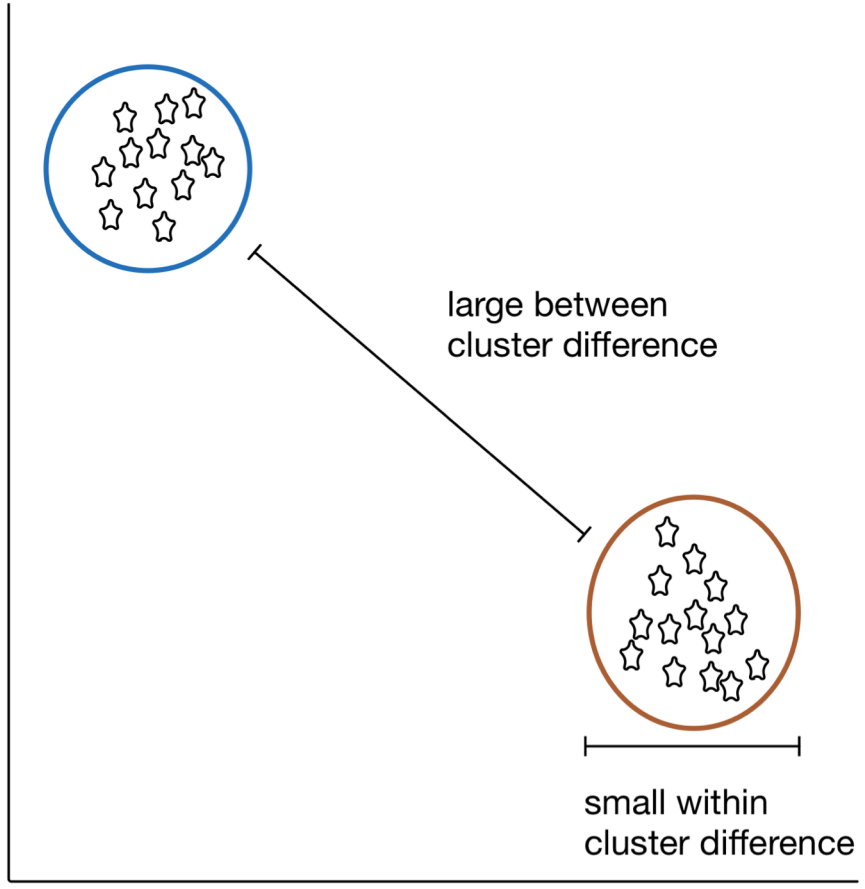
- The Ball-Hall index
- The Banfeld-Raftery index
- The C index
- The Calinski-Harabasz index
- The Davies-Bouldin index
- The Det Ratio index
- The Dunn index
- The Baker-Hubert Gamma index
- The GDI index
- The Gplus index
- The KsqDetW index
- The LogDetRatio index
- The LogSSRatio index
- The McClain-Rao index
- The PBM index
- The Point-Biserial index
- The Ratkowsky-Lance index
- The Ray-Turi index
- The Scott-Symons index
- The SD index
- The SDbw index
- The Silhouette index
- The Tau index
- The TraceW index
- The TraceWiB index
- The Wemmert-Gançarski index
- The Xie-Beni index

(These are all defined and available in the clusterCrit package for R)

What are we to make of all these different, supposedly context free measures of clustering quality?

Very Broad Goals

- Within clusters, everything is very similar.
- Between clusters, there is a lot of difference.
- The problem: there are many ways for clusters to deviate from this ideal.
- In specific clustering cases, how do we weigh the good aspects (e.g. high within cluster similarity) relative to the bad (e.g. low between cluster separation)
- Thus the large number of CQMs
- Question: is this trade-off (and the resulting CQMs) really context independent?
- Maybe different weightings are more relevant in different contexts?



Comparing measures across datasets

Vendramin et al 2010 used a number of benchmark tests to compared a large number of intrinsic validation measures

Broad conclusion: variants of Silhouette performed well across tests

		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Point-Biserial	A	0.000	0.046	0.226	0.247	0.262	0.289	0.306	0.373	0.390	0.408	0.488	0.555	0.566	0.571	0.584	0.636	0.642	0.645	0.694	0.705	0.729	0.736	0.768	0.822	0.837	1.107
Tau	B	-0.046	0.000	0.180	0.201	0.216	0.243	0.260	0.327	0.344	0.362	0.442	0.509	0.520	0.525	0.538	0.590	0.597	0.599	0.649	0.659	0.683	0.690	0.722	0.776	0.791	1.061
C/k$^{1/2}$	C	-0.226	-0.180	0.000	0.021	0.036	0.063	0.080	0.147	0.164	0.182	0.263	0.329	0.340	0.345	0.358	0.410	0.417	0.419	0.469	0.479	0.504	0.510	0.542	0.596	0.611	0.881
ASWC	D	-0.247	-0.201	-0.021	0.000	0.015	0.042	0.060	0.127	0.143	0.161	0.242	0.308	0.319	0.324	0.338	0.390	0.396	0.398	0.448	0.458	0.483	0.489	0.521	0.575	0.590	0.860
ASSWC	E	-0.262	-0.216	-0.036	-0.015	0.000	0.027	0.045	0.112	0.128	0.146	0.227	0.293	0.304	0.309	0.323	0.375	0.381	0.384	0.433	0.443	0.468	0.474	0.506	0.560	0.575	0.846
PBM	F	-0.289	-0.243	-0.063	-0.042	-0.027	0.000	0.017	0.084	0.101	0.119	0.199	0.266	0.277	0.282	0.295	0.347	0.353	0.356	0.406	0.416	0.440	0.447	0.479	0.533	0.548	0.818
SWC	G	-0.306	-0.260	-0.080	-0.060	-0.045	-0.017	0.000	0.067	0.083	0.102	0.182	0.249	0.260	0.265	0.278	0.330	0.336	0.339	0.388	0.399	0.423	0.430	0.462	0.516	0.530	0.801
SSWC	H	-0.373	-0.327	-0.147	-0.127	-0.112	-0.084	-0.067	0.000	0.016	0.035	0.115	0.181	0.193	0.198	0.211	0.263	0.269	0.272	0.321	0.332	0.356	0.363	0.395	0.449	0.463	0.734
Dunn12	I	-0.390	-0.344	-0.164	-0.143	-0.128	-0.101	-0.083	-0.016	0.000	0.018	0.099	0.165	0.176	0.181	0.195	0.247	0.253	0.255	0.305	0.315	0.340	0.346	0.378	0.432	0.447	0.717
Dunn62	J	-0.408	-0.362	-0.182	-0.161	-0.146	-0.119	-0.102	-0.035	-0.018	0.000	0.080	0.147	0.158	0.163	0.176	0.228	0.234	0.237	0.287	0.297	0.321	0.328	0.360	0.414	0.429	0.699
Dunn13	K	-0.488	-0.442	-0.263	-0.242	-0.227	-0.199	-0.182	-0.115	-0.099	-0.080	0.000	0.066	0.078	0.082	0.096	0.148	0.154	0.157	0.206	0.217	0.241	0.248	0.280	0.334	0.348	0.619
VRC	L	-0.555	-0.509	-0.329	-0.308	-0.293	-0.266	-0.249	-0.181	-0.165	-0.147	-0.066	0.000	0.011	0.016	0.030	0.082	0.088	0.090	0.140	0.150	0.175	0.181	0.213	0.267	0.282	0.552
Ball and Hall	M	-0.566	-0.520	-0.340	-0.319	-0.304	-0.277	-0.260	-0.193	-0.176	-0.158	-0.078	-0.011	0.000	0.005	0.018	0.070	0.076	0.079	0.129	0.139	0.163	0.170	0.202	0.256	0.271	0.541
Trace(W)	N	-0.571	-0.525	-0.345	-0.324	-0.309	-0.282	-0.265	-0.198	-0.181	-0.163	-0.082	-0.016	-0.005	0.000	0.013	0.065	0.072	0.074	0.124	0.134	0.159	0.165	0.197	0.251	0.266	0.536
DB	O	-0.584	-0.538	-0.358	-0.338	-0.323	-0.295	-0.278	-0.211	-0.195	-0.176	-0.096	-0.030	-0.018	-0.013	0.000	0.052	0.058	0.061	0.110	0.121	0.145	0.152	0.184	0.238	0.252	0.523
Nlog(|T|/|W|)	P	-0.636	-0.590	-0.410	-0.390	-0.375	-0.347	-0.330	-0.263	-0.247	-0.228	-0.148	-0.082	-0.070	-0.065	-0.052	0.000	0.006	0.009	0.058	0.069	0.093	0.100	0.132	0.186	0.200	0.471
Trace(CovW)	Q	-0.642	-0.597	-0.417	-0.396	-0.381	-0.353	-0.336	-0.269	-0.253	-0.234	-0.154	-0.088	-0.076	-0.072	-0.058	-0.006	0.000	0.003	0.052	0.063	0.087	0.094	0.126	0.180	0.194	0.465
k^2|W|	R	-0.645	-0.599	-0.419	-0.398	-0.384	-0.356	-0.339	-0.272	-0.255	-0.237	-0.157	-0.090	-0.079	-0.074	-0.061	-0.009	-0.003	0.000	0.049	0.060	0.084	0.091	0.123	0.177	0.192	0.462
log(SSB/SSW)	S	-0.694	-0.649	-0.469	-0.448	-0.433	-0.406	-0.388	-0.321	-0.305	-0.287	-0.206	-0.140	-0.129	-0.124	-0.110	-0.058	-0.052	-0.049	0.000	0.010	0.035	0.041	0.074	0.128	0.142	0.413
Dunn11	T	-0.705	-0.659	-0.479	-0.458	-0.443	-0.416	-0.399	-0.332	-0.315	-0.297	-0.217	-0.150	-0.139	-0.134	-0.121	-0.069	-0.063	-0.060	-0.010	0.000	0.024	0.031	0.063	0.117	0.132	0.402
Gamma	U	-0.729	-0.683	-0.504	-0.483	-0.468	-0.440	-0.423	-0.356	-0.340	-0.321	-0.241	-0.175	-0.163	-0.159	-0.145	-0.093	-0.087	-0.084	-0.035	-0.024	0.000	0.007	0.039	0.093	0.107	0.378
McClain and Rao	V	-0.736	-0.690	-0.510	-0.489	-0.474	-0.447	-0.430	-0.363	-0.346	-0.328	-0.248	-0.181	-0.170	-0.165	-0.152	-0.100	-0.094	-0.091	-0.041	-0.031	-0.007	0.000	0.032	0.086	0.101	0.371
C-Index	W	-0.768	-0.722	-0.542	-0.521	-0.506	-0.479	-0.462	-0.395	-0.378	-0.360	-0.280	-0.213	-0.202	-0.197	-0.184	-0.132	-0.126	-0.123	-0.074	-0.063	-0.039	-0.032	0.000	0.054	0.069	0.339
|T|/|W|	X	-0.822	-0.776	-0.596	-0.575	-0.560	-0.533	-0.516	-0.449	-0.432	-0.414	-0.334	-0.267	-0.256	-0.251	-0.238	-0.186	-0.180	-0.177	-0.128	-0.117	-0.093	-0.086	-0.054	0.000	0.015	0.285
Trace(W^{-1}B)	Y	-0.837	-0.791	-0.611	-0.590	-0.575	-0.548	-0.530	-0.463	-0.447	-0.429	-0.348	-0.282	-0.271	-0.266	-0.252	-0.200	-0.194	-0.192	-0.142	-0.132	-0.107	-0.101	-0.069	-0.015	0.000	0.270
G(+)	Z	-1.107	-1.061	-0.881	-0.860	-0.846	-0.818	-0.801	-0.734	-0.717	-0.699	-0.619	-0.552	-0.541	-0.536	-0.523	-0.471	-0.465	-0.462	-0.413	-0.402	-0.378	-0.371	-0.339	-0.285	-0.270	0.000
	Mean	0.959	0.913	0.733	0.712	0.697	0.670	0.653	0.586	0.569	0.551	0.471	0.404	0.393	0.388	0.375	0.323	0.316	0.314	0.264	0.254	0.230	0.223	0.191	0.137	0.122	-0.148

Fig. 10 Mean values (bottom bar) and their differences (cells) for Pearson correlation between relative and external (Jaccard) criteria: $k_{\max} = 25$.

Machine Learning vs Human Learning

- Lewis et al compare 6 common CQMs with human evaluation of clustering results
- Main finding: Human clustering evaluation was most similar to Silhouette and Calinski-Harabasz
- Maybe internal validation/CQM is saying something about clustering across all contexts?
- Maybe easier to identify the clearly bad than all the variations of good?

Table 1: Correlation coefficients between human responses and CQMs with k factored out (except for the k column). Text in bold (excluding k column) if $p < .0025$ after Bonferroni correction for $n = 20$ comparisons per subject group and $\alpha = .05$.

ρ	Expert Positive	Expert Negative	Novice Positive	Novice Negative	Gamma	Silhouette	Dunn	Avg Within	Avg Btw	CH	W-Inter/Intra	k
Expert Pos	1	**-.35**	**.56**	-.19	-.15	**.46**	**.40**	**-.39**	**.34**	**.44**	.19	-.43
Expert Neg		1	-.13	**.44**	.09	**-.27**	-.12	**.44**	-.18	**-.36**	**-.30**	.32
Novice Pos			1	-.04	-.13	**.39**	**.40**	-.20	.23	**.30**	.04	-.73
Novice Neg				1	.08	**-.27**	.01	**.30**	-.07	**-.25**	**-.27**	.71

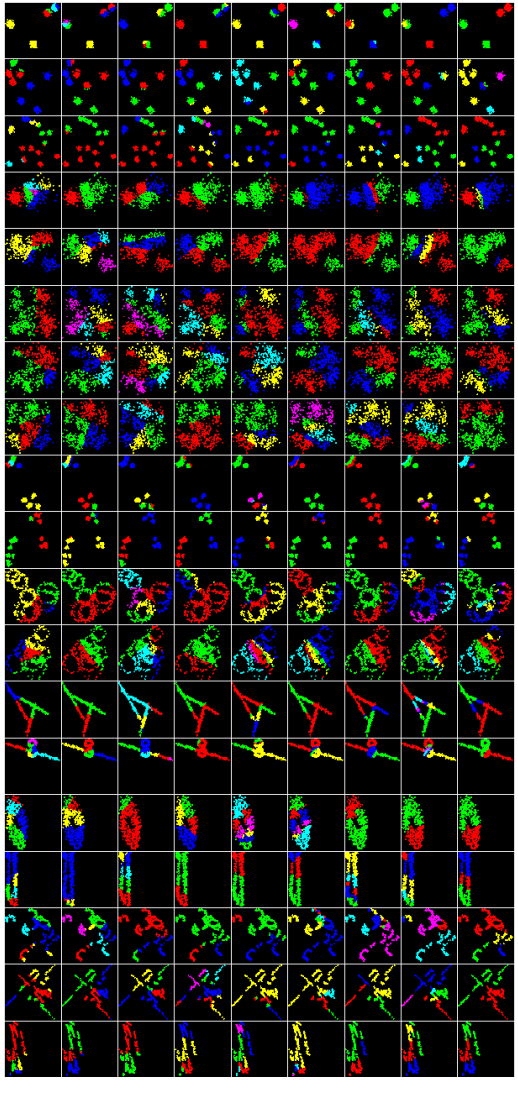
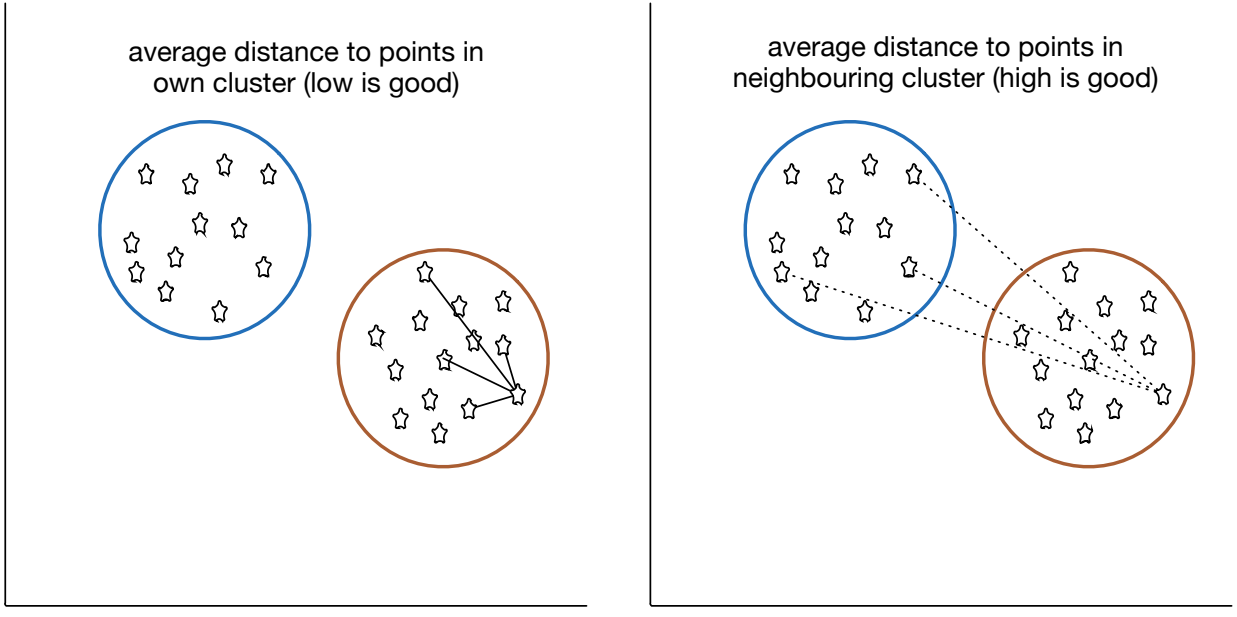


Figure 1: All stimuli. Datasets are in rows; partitions are in columns.

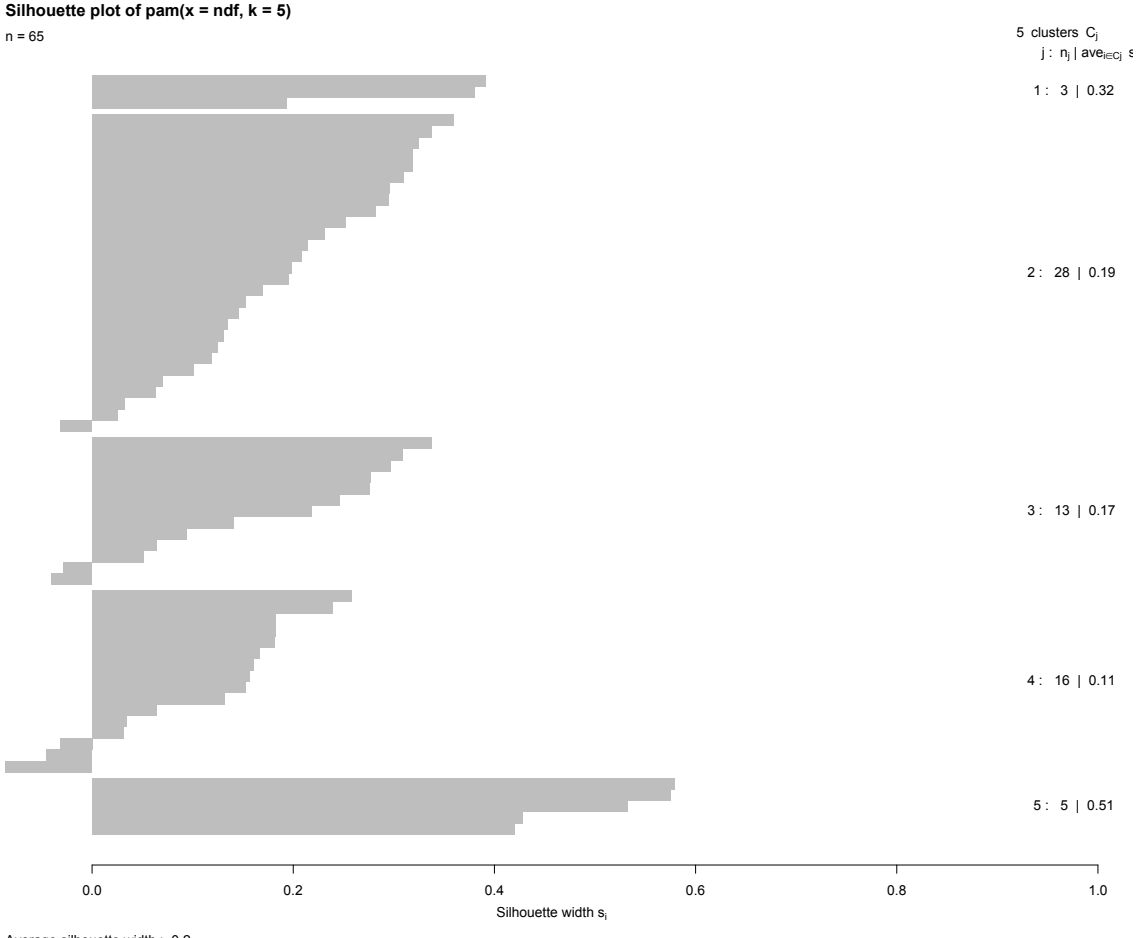
Silhouette Index: Algorithm



$$\text{silhouette metric for a point} = \frac{\text{(average dissimiliarity with neighbouring cluster - average dissimilarity with own cluster)}}{\text{maximum dissimilarity value (own or neighbour)}}$$

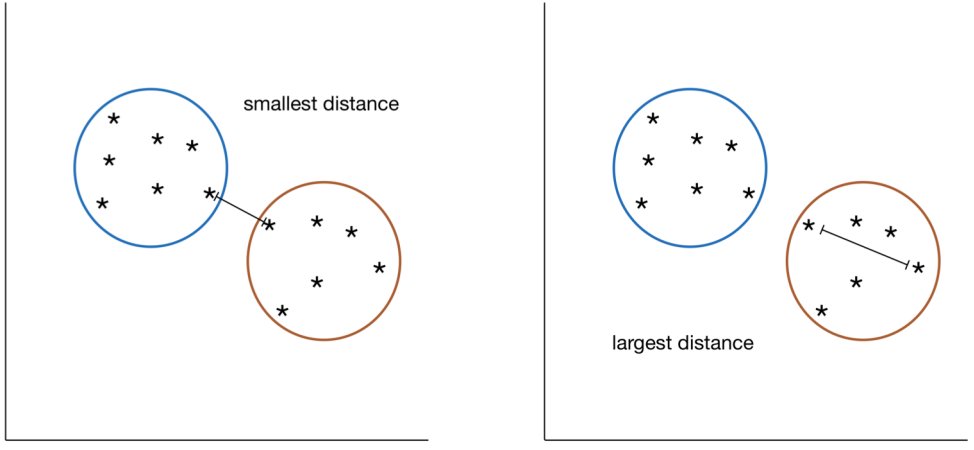
A strong internal validation metric that incorporates a number of measures.

Silhouette Metric Sample Results (I)



Dunn's Index: Algorithm

- Within a cluster, the size of the cluster (e.g. greatest distance between points)
- Between two clusters, the distance between the clusters (e.g. minimum distance between points)
- Ratio: The minimum intercluster distance across all pairs of clusters / maximum intracluster distance across all clusters
- A number of possible ways to define inter cluster distance and cluster size.



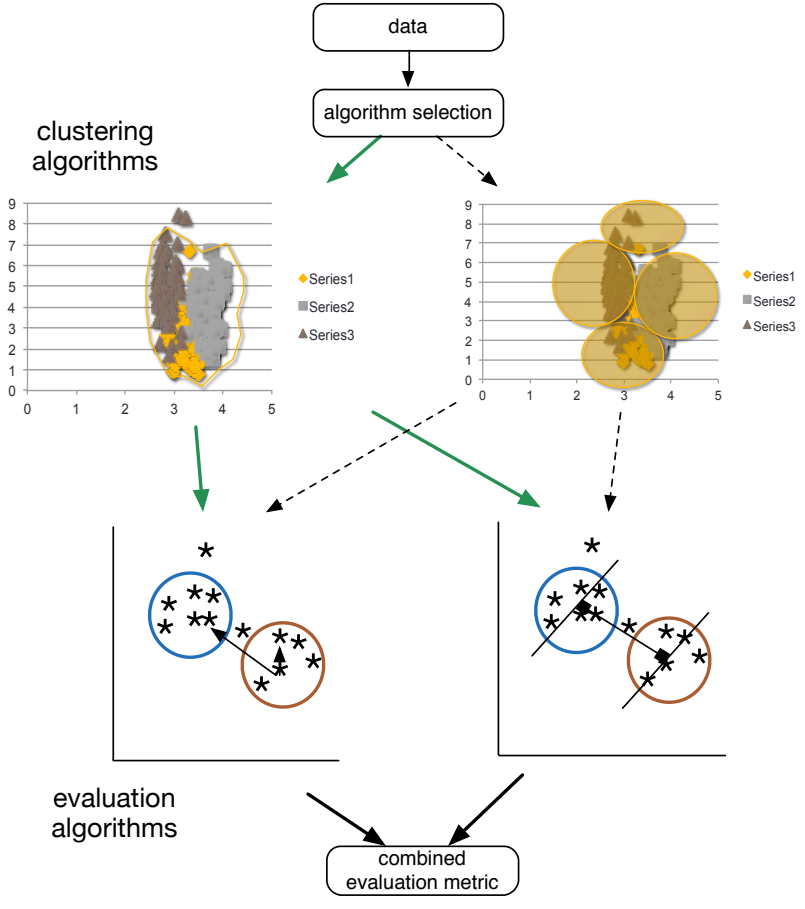
Comparison with Silhouette Index: In a sense, a simpler measure. More of a whole cluster measure, rather than a point by point measure. Evaluates based on extremes (max and min).

Clustering Validation – Part 5

RELATIVE VALIDATION

More is better?

- Getting a single validation measure for a single clustering is not that useful – could the results be better? Is this the best we can hope for?
- How about comparing results across runs or parameter settings?
- Main emphasis with relative validation is how to compare results of individual runs.



Correlation Measures

- Look at correlation between clustering assignments
- Rand, Jaccard, Gamma
- Perfect correlation gives maximum value of the measure

	P1	P2	P3	P4	P5	P6
P1	1					
P2	0	1				
P3	1	0	1			
P4	1	0	1	1		
P5	0	1	0	0	1	
P6	0	0	0	0	0	1

	P1	P2	P3	P4	P5	P6
P1	1					
P2	0	1				
P3	1	0	1			
P4	1	0	1	1		
P5	0	1	0	0	1	
P6	0	1	0	0	1	1

Two very similar clustering results (but notice they vary in the number of clusters)

Rand's Index

- SS- the number of pairs of items belonging to the same cluster in both clusterings (C1 = 1 and C2 = 1)
- SD- the number of pairs together in one clustering but not the other (C1 = 1 and C2 = 0)
- DS- the number of pairs not together in one clustering but together in the other (C1 = 0 and C2 = 1)
- DD- the number of pairs not together in either cluster (C1 = 0 and C2 = 0)
- Note that SS and DD are good, and DS, SD are bad.
- Rand Index is the ratio of SS + DD to the total number of pairs. If Rand Index = 1, the clustering perfectly matches the gold standard.

Clustering 1 (C1)

	P1	P2	P3	P4	P5	P6
P1	1					
P2	0	1				
P3	1	0	1			
P4	1	0	1	1		
P5	0	1	0	0	1	
P6	0	0	0	0	0	1

Clustering 2 (C2)

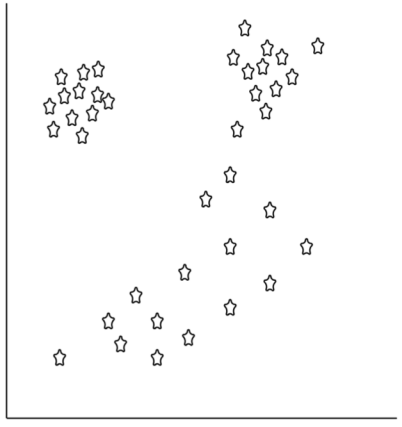
	P1	P2	P3	P4	P5	P6
P1	1					
P2	0	1				
P3	1	0	1			
P4	1	0	1	1		
P5	0	1	0	0	1	
P6	0	1	0	0	1	1

$$(SS + DD)/(SS + DD + SD + DS)$$

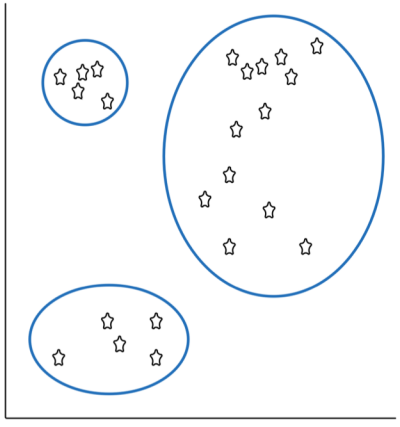
$$(4 + 9)/(4 + 9 + 0 + 2) = 0.87$$

Stability

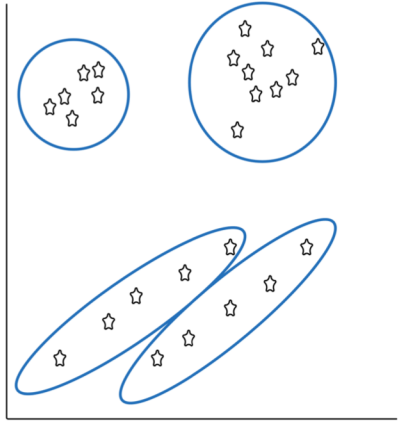
- Some options:
  - multiple datasets sampled from same source
  - different columns used to generate clusters (i.e. drop a different column each time)
- Similarity of results is measured
- If results are not stable across clustering schemes, further investigation required



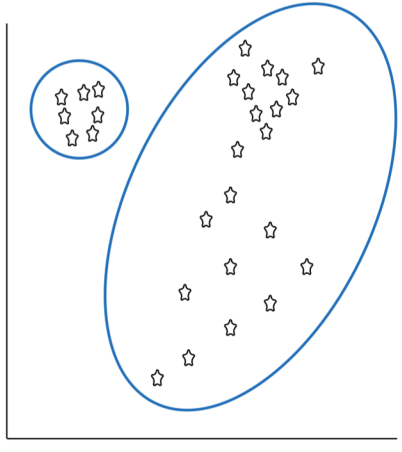
Full dataset



Sample 1 clustering



Sample 2 clustering



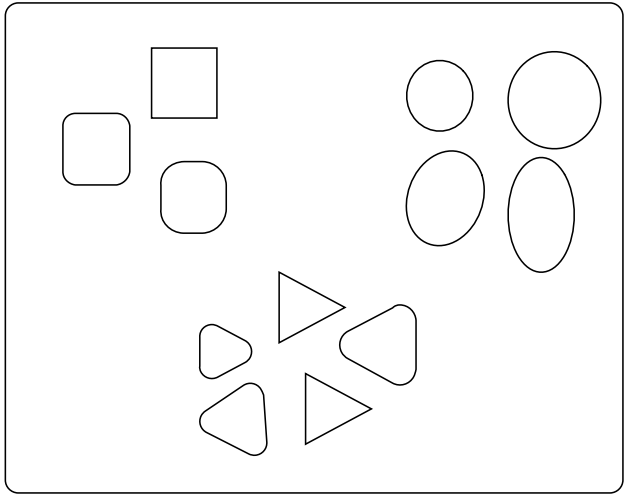
Sample 3 clustering

Clustering Validation – Part 6

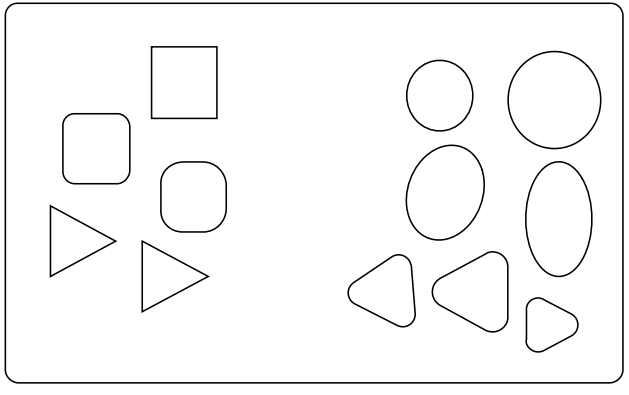
EXTERNAL VALIDATION

Back to Context

- Brings in outside information to evaluate the clusters
- Outside information is typically the ‘correct’ class
- How is this different from classification then?
- Often used to build confidence in the overall approach, based on preliminary or sample results



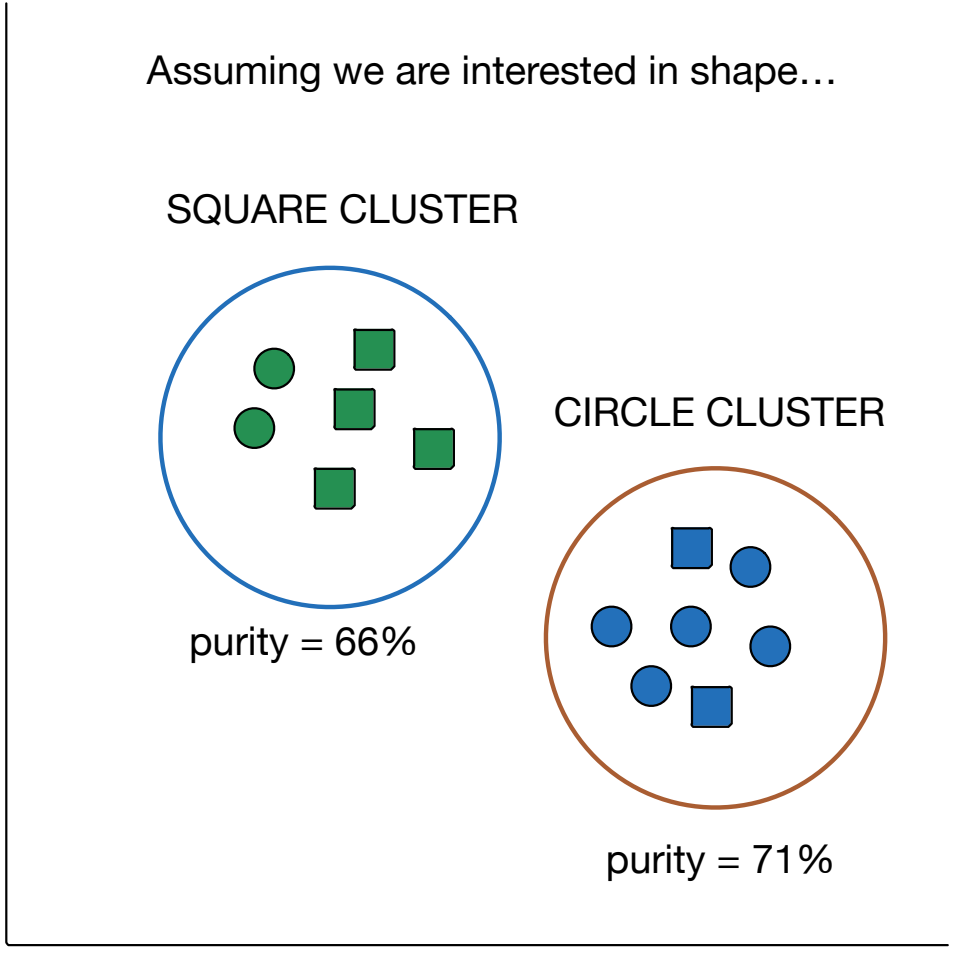
Natural Groupings



Clustering Results

Example Metric: Purity

- For this metric each cluster is assigned to the class which is most frequent in the cluster
- To calculate the purity: number of correctly assigned points / number of points in the cluster
- Some other options: precision, recall



Types of External Validation

- Amigó et al (2009) provide a number of constraints for external validation measures
- They suggest external evaluation strategies can be based on:
  - set matching
  - counting pairs
  - entropy measures
  - edit distance
- Similar to strategies used to evaluate classification
- They recommend using a particular version (Bcubed) of precision and recall for external validation, as these best take into consideration the 4 constraints

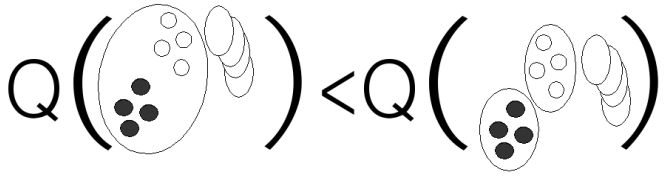


Figure 1: Constraint 1: Cluster Homogeneity

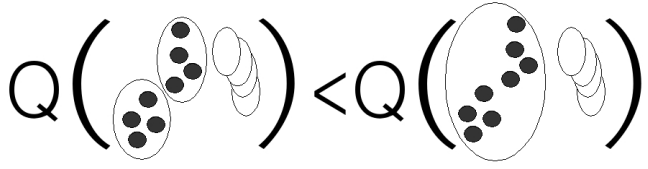


Figure 2: Constraint 2: cluster completeness

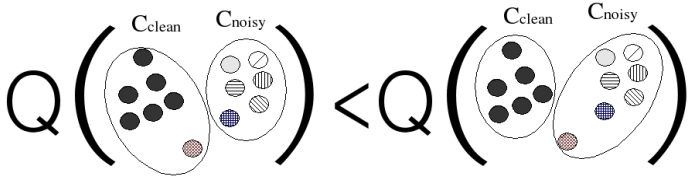


Figure 3: Constraint 3: Rag Bag

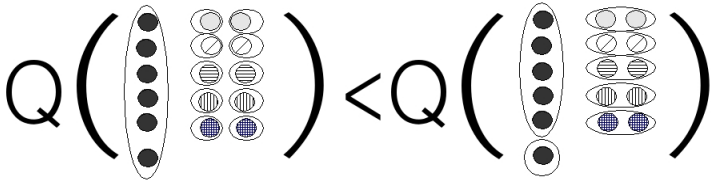


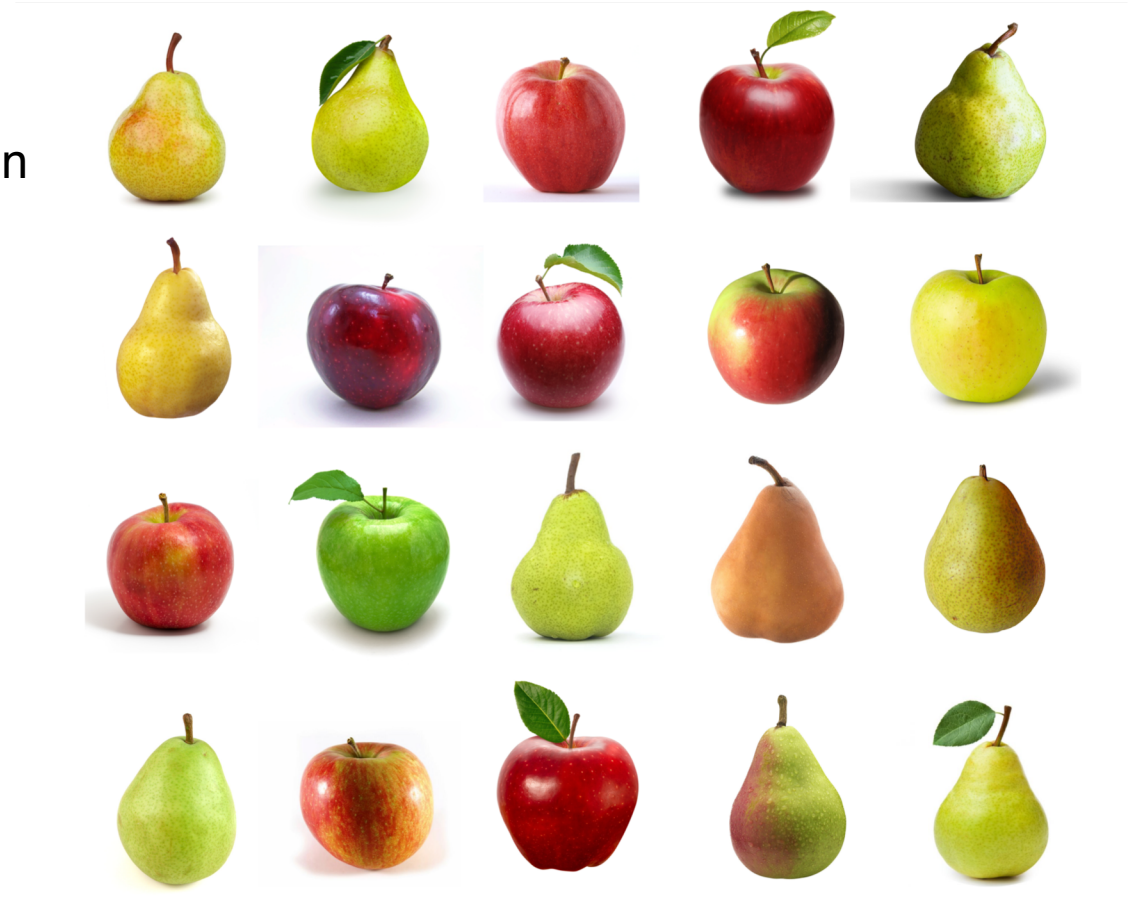
Figure 4: Clusters Size vs. Quantity

Clustering Validation – Part 7

CONCLUDING THOUGHTS

Try and Try Again

- A large amount of diversity in clustering validation techniques
- Be aware of the types of validation, and variations within types
- Seek agreement across techniques, ok to compare
- There are many ways for a clustering to be ‘ok’ – you need to decide what is important and what can be ignored
- A lot depends on context



References

- Vendramin, Lucas & J. G. B. Campello, Ricardo & Hruschka, Eduardo. (2010). Relative clustering validity criteria: A comparative overview. Statistical Analysis and Data Mining. 3. 209-235. 10.1002/sam.10080.
- Amigó, Enrique & Gonzalo, Julio & Artiles, Javier & Verdejo, M. (2009). Comparison of extrinsic clustering evaluation metrics based on formal constraints. Inform Retriev 12:461-486. Information Retrieval. 12. 461-486. 10.1007/s10791-008-9066-8.
- M Lewis, Joshua & Ackerman, Margareta & de Sa, Virginia. (2012). Human Cluster Evaluation and Formal Quality Measures: A Comparative Study. Proc. 34th Conf. of the Cognitive Science Society (CogSci. .
- Bernard Desgraupes (2013). Clustering Indices. Lab Modal'X, University Paris Ouest.