CLUSTERING

"Data science does not replace statistical modeling and data analysis; it augments them."

(P. Boily)

"Data is not information, information is not knowledge,
knowledge is not understanding, understanding is not wisdom."

(attributed to Cliff Stoll in Keeler's *Nothing to Hide: Privacy in the 21st Century*, 2006)

LEARNING OBJECTIVES

Become familiar with the basic concepts of clustering; as well as some of the common algorithms.

Become familiar with one variant of partition clustering (k-means).

Become familiar with clustering validation criteria.

CONTENTS

CASE STUDY: OK CUPID DATA

CLUSTERING

Finding true love *via* clustering analysis

(K.Poulsen, *How a Math Genius Hacked OK Cupid to Find True Love*, WIRED)

CONTEXT

Chris McKinlay, a 35 year old UCLA Math PhD Student, was looking for a romantic partner online with little luck

- *OK Cupid* algorithms use only the questions that both potential matches decide to answer, and the questions he had chosen (more or less at random up to that point) were not popular

Between June 2012 and December 2013, he

- used statistical sampling to find questions which mattered to the kind of partner he had in mind;
- constructed a new profile that answered only those questions;
- matched only with women in LA who might be right for him.

PROCESS

This story provides a great example of the data mining process, from start to finish:

1. **Collect** data
2. Collect **more** and **slightly better** and **different** data
3. Collect **still more** data
4. Figure out a data mining technique that would be **relevant** to what he wanted to know (clustering)
5. **Validate** the results of the analysis

PROCESS

This story provides a great example of the data mining process, from start to finish (continued):

6. **Investigate** the results, and narrow down which results were actually interesting
7. Analyze the interesting results **some more**, and use this to solve the original problem
8. Use the data to **improve other areas** of his profile as well
9. Sit back and reap the benefits of data mining?

METHODOLOGY AND RESULTS

Used k-mode to cluster 20,000 women into seven statistically distinct clusters based on their questions and answers.

Validated the clustering with another 5,000 profiles from the site.

Analyzed the clusters to find two that interested him

- women in their mid-twenties who looked like indie types, musicians and artists
- slightly older women who held professional creative jobs, like editors and designers.

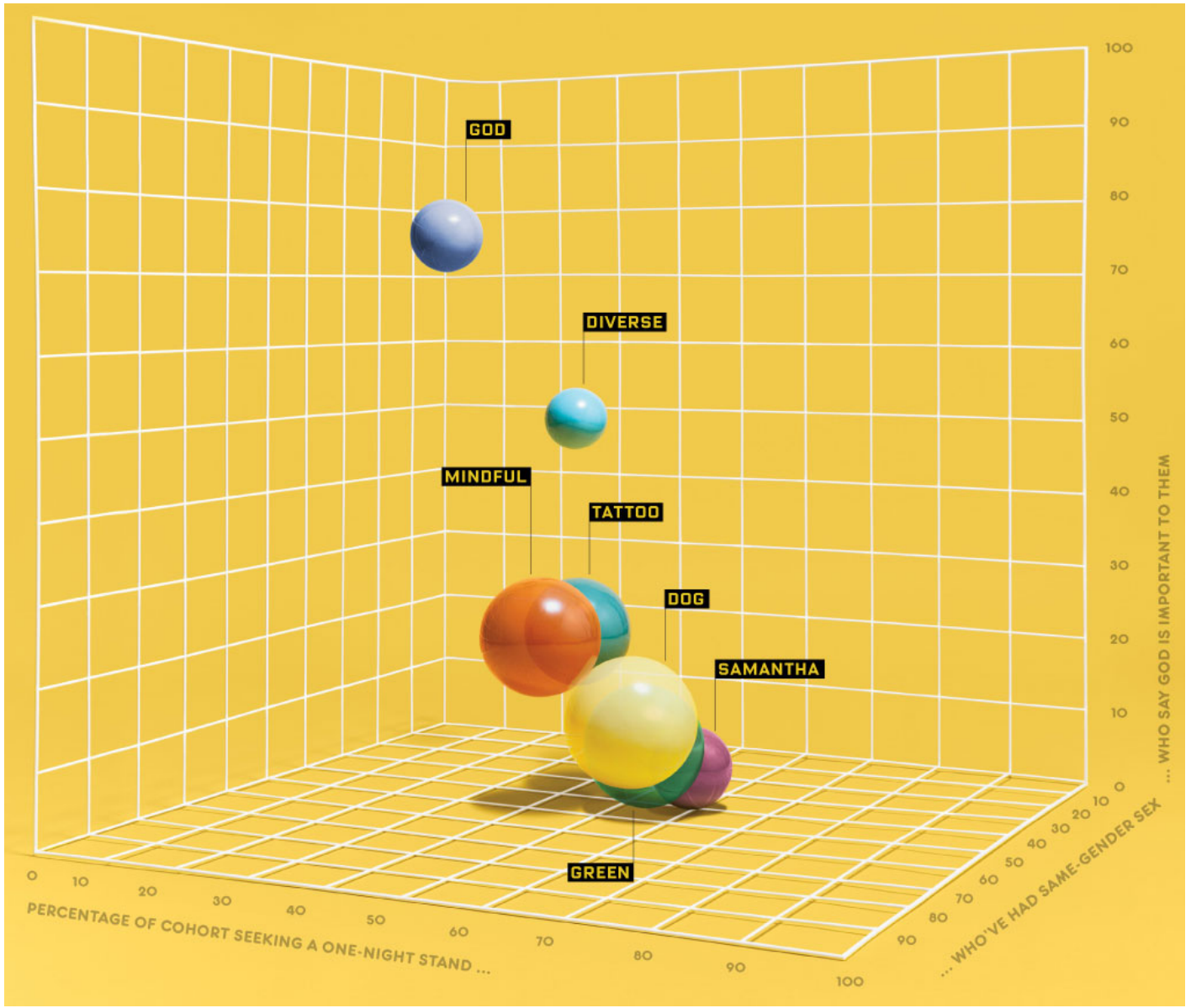
METHODOLOGY AND RESULTS

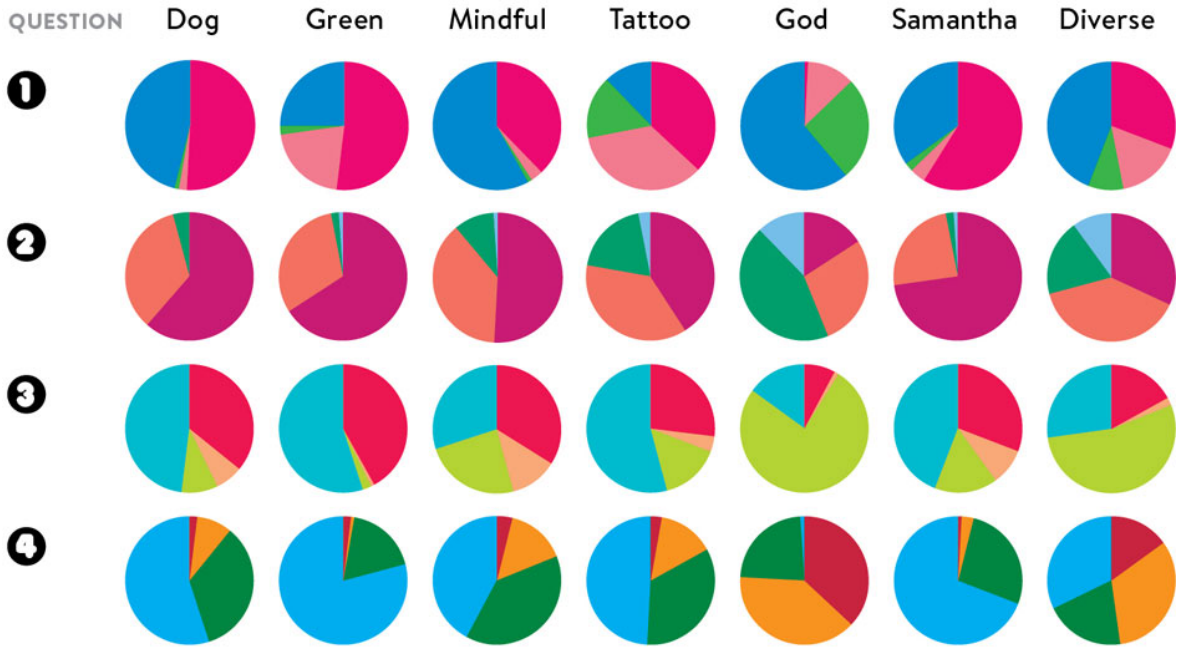
Used adaptive boosting (a machine-learning algorithm) to derive which questions he should answer in his profile.

More people appeared as matches based on his profile, leading to more first dates, some second dates, and a lone third date.

In the end, he was contacted by someone who was intrigued by his profile.

She asked him out and they were living together when the article was written.


GOD
DIVERSE
MINDFUL
TATTOO
DOG
SAMANTHA
GREEN
0 10 20 30 40 50 60 70 80 90 100
PERCENTAGE OF COHORT SEEKING A ONE-NIGHT STAND ...
0 10 20 30 40 50 60 70 80 90 100
... WHO'VE HAD SAME-GENDER SEX
0 10 20 30 40 50 60 70 80 90 100
... WHO SAY GOD IS IMPORTANT TO THEM



1. About how long do you want your next relationship to last?

- One night
- A few months to a year
- Several years
- The rest of my life

2. Say you've started seeing someone you really like. As far as you're concerned, how long will it take before you have sex?

- 1-2 dates
- 3-5 dates
- 6 or more dates
- Only after the wedding

3. Have you ever had a sexual encounter with someone of the same sex?

- Yes, and I enjoyed myself
- Yes, and I did not enjoy myself
- No, and I would never
- No, but I'd like to

4. How important is religion/God in your life?

- Extremely important
- Somewhat important
- Not very important
- Not important at all

DISCUSSION

How do you feel about this use of machine learning?

CLUSTERING BASICS

CLUSTERING

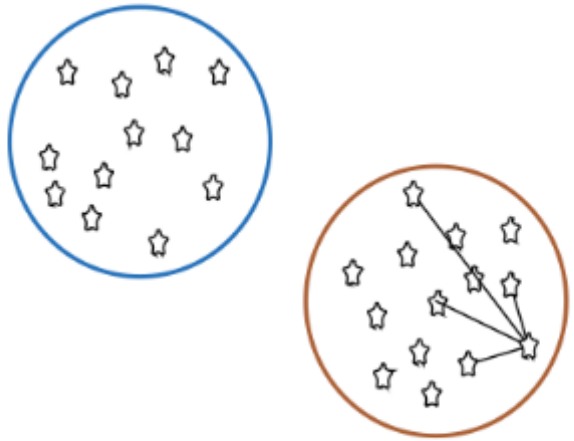
"The Milky Way is nothing but a mass of innumerable stars planted together in clusters."
(Galileo Galilei, *Sidereus Nuncius*)

CLUSTERING OVERVIEW

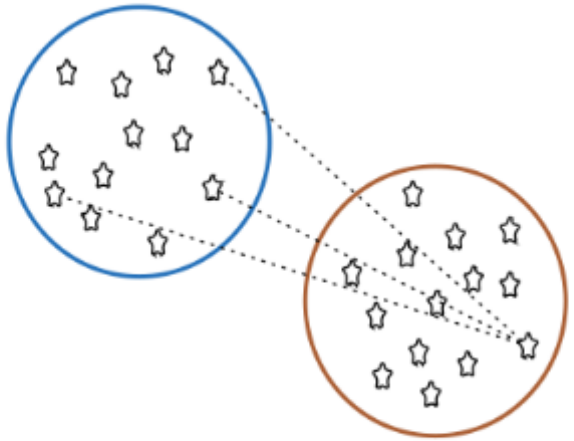
In **clustering**, the data is divided into **naturally occurring groups**. Within each group, the data points are **similar**; from group to group, they are **dissimilar**.

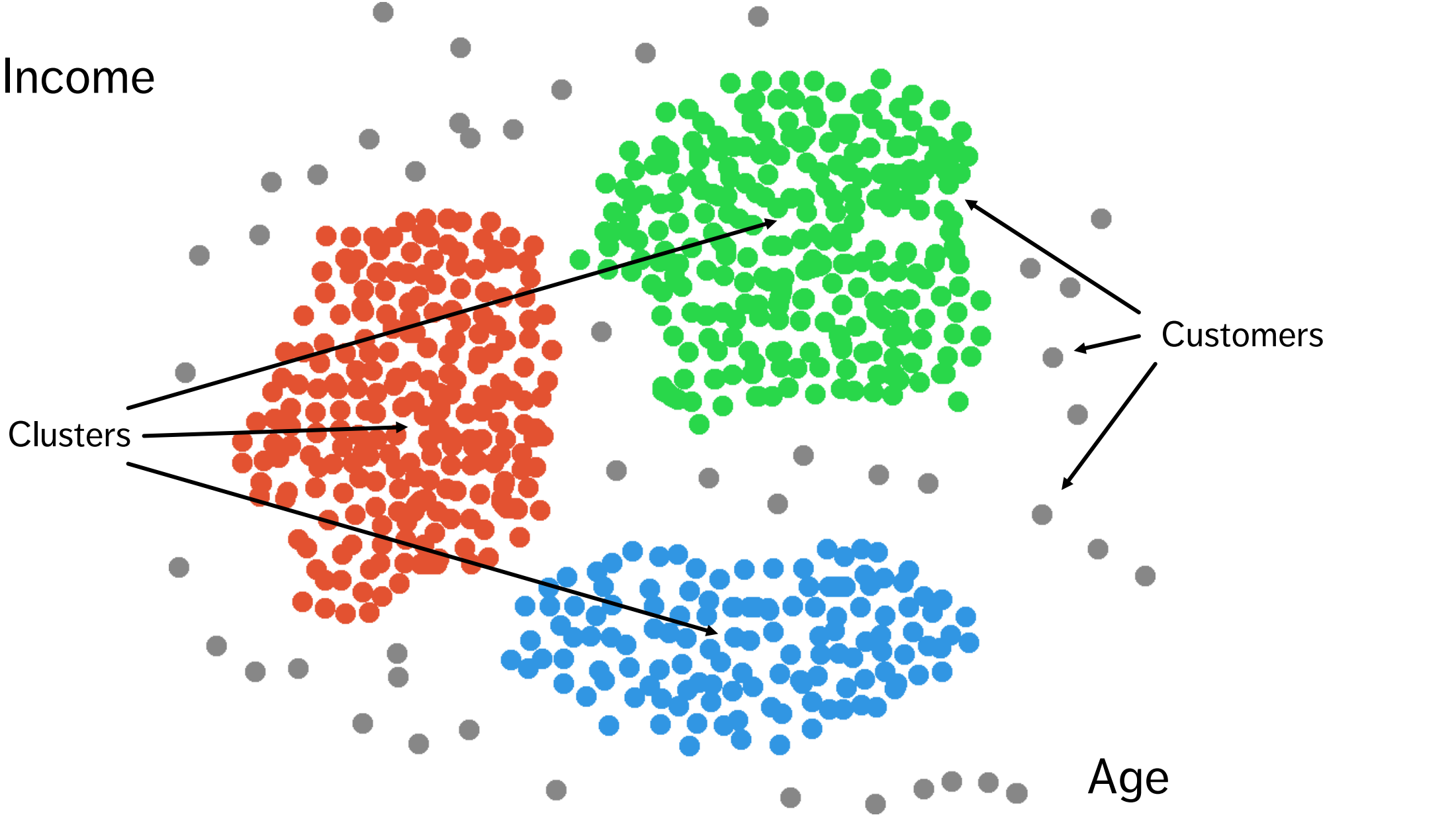
The grouping labels are not determined ahead of time, so clustering is an example of **unsupervised** learning.

average distance to points in own cluster (**low is good**)



average distance to points in neighbouring cluster (**high is good**)




Income
Clusters
Customers
Age

CLUSTERING OVERVIEW

Clustering is a relatively **intuitive** concept for human beings as our brains do it unconsciously

- facial recognition
- searching for patterns, etc.

In general, people are very good at **messy** data, but computers and algorithms have a harder time.

Part of the difficulty is that there is **no agreed-upon definition of what constitutes a cluster:**

- "I may not be able to define what it is, but I know one when I see one"

CLUSTERING OVERVIEW

Clustering algorithms can be **complex** and **non-intuitive**, based on varying notions of similarities between observations.

- in spite of that, the temptation to explain clusters *a posteriori* is **strong**

They are also (typically) **non-deterministic**:

- the same algorithm, applied twice (or more) to the same dataset, can discover completely different clusters
- the order in which the data is presented can play a role
- so can starting configurations

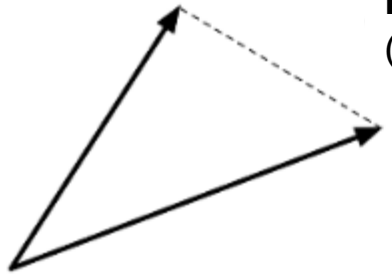
DISCUSSION

What does this (potential) non-repeatability imply for validation?

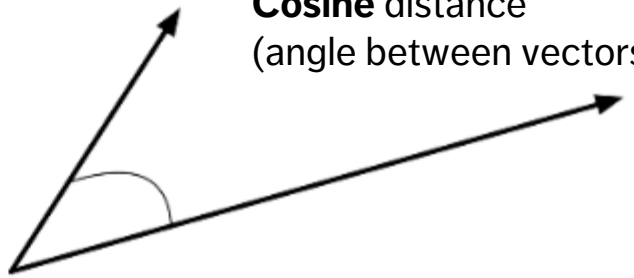
IMPORTANT: data must be scaled before it is fed into clustering algorithms.

CLUSTERING REQUIREMENT

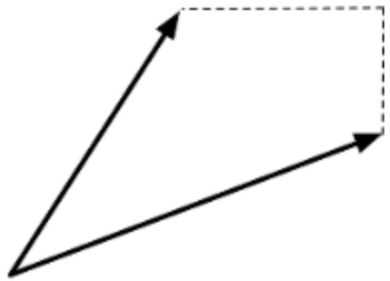
A measure of **similarity** w (or a distance d) between observations.



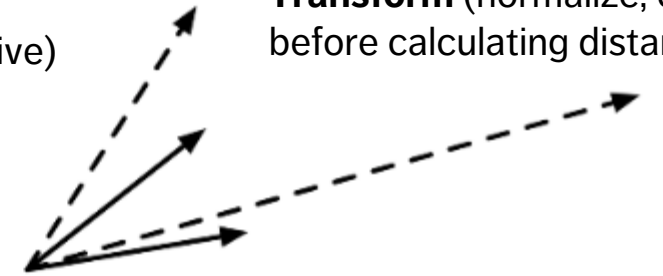
Euclidean distance
(as the crow flies)



Cosine distance
(angle between vectors)



Manhattan distance
(you might have to drive)



Transform (normalize, center)
before calculating distance

Typically, $w \to 1$ as $d \to 0$, and $w \to 0$ as $d \to \infty$.

* may need to be dichotomized

DISTANCE MEASURES (METRICS)

Categorical Variables*

- Hamming distance
- Russel/Rao index
- Jaccard
- Matching coefficient
- Dice's coefficient
- etc.

Numerical Variables*

- Euclidean
- Manhattan
- Correlation
- Cosine
- Pearson
- etc.

No steadfast rule to determine which distance to use in k-means

Competing schemes are often produced using different metrics.

APPLICATIONS

Text Documents

- grouping similar documents according to their topics, based on the patterns of common and unusual words

Product Recommendations

- grouping online purchasers based on the products they have viewed, purchased, liked, or disliked
- grouping products based on customer reviews

Marketing and Business

- grouping client profiles based on their demographics and preferences

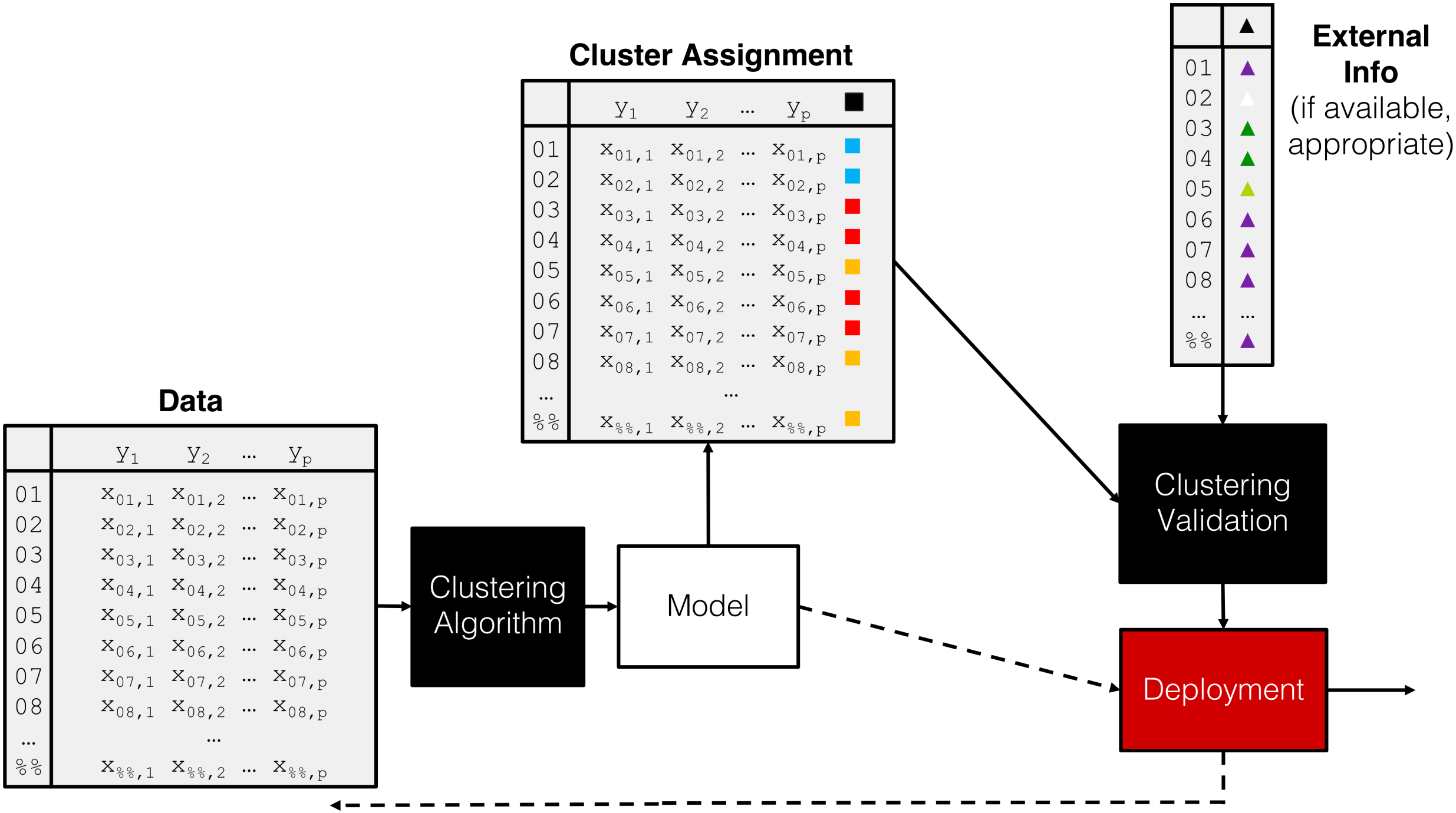
OTHER USES

Dividing a larger group (or area, or category) into **smaller** groups, with members of the smaller groups guaranteed to have similarities of some kind.

- tasks may then be solved separately for each of the smaller groups
- this may lead to increased accuracy once the separate results are aggregated

Creating (new) taxonomies **on the fly**, as new items are added to a group of items

- this would allow for easier product navigation on a website like Netflix, for instance.


Cluster Assignment
External Info
(if available, appropriate)
Data
Clustering Algorithm
Model
Clustering Validation
Deployment

CLUSTERING ALGORITHMS

CLUSTERING

"Clustering is in the eye of the beholder, and as such, researchers have proposed many induction principles and models whose corresponding optimization problem can only be approximately solved by an even larger number of algorithms."

(V. Estivill-Castro, *Why So Many Clustering Algorithms?*)

CLUSTERING SCHEMES

k-Means

- classical (and over-used) model
- assumptions made about the shape of clusters

Hierarchical Clustering

- easy to interpret, deterministic

Latent Dirichlet Allocation

- used for topic modeling

Expectation Maximization

CLUSTERING SCHEMES

Balanced Iterative Reducing and Clustering using Hierarchies

- aka BIRCH

Density-Based Spatial Clustering of Applications with Noise

- graph-based

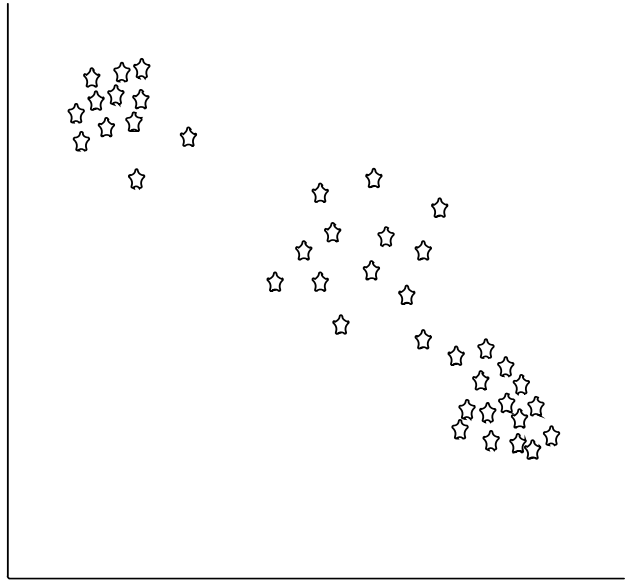
Affinity Propagation

- selects the optimal number of clusters automatically

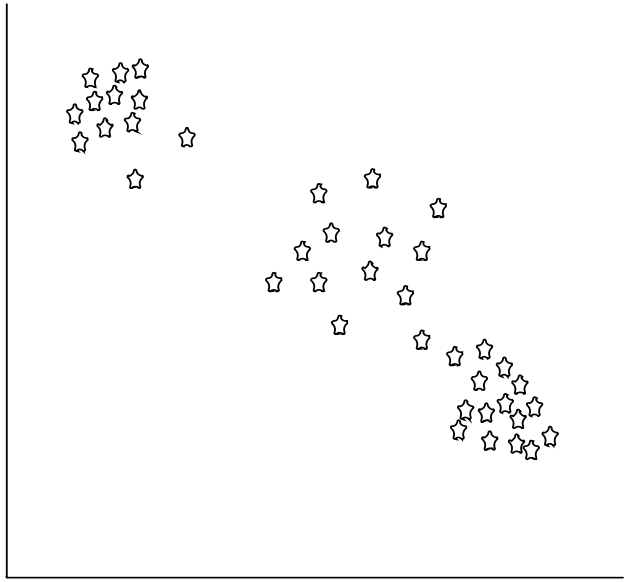
Spectral Clustering

- recognizes non-blob clusters

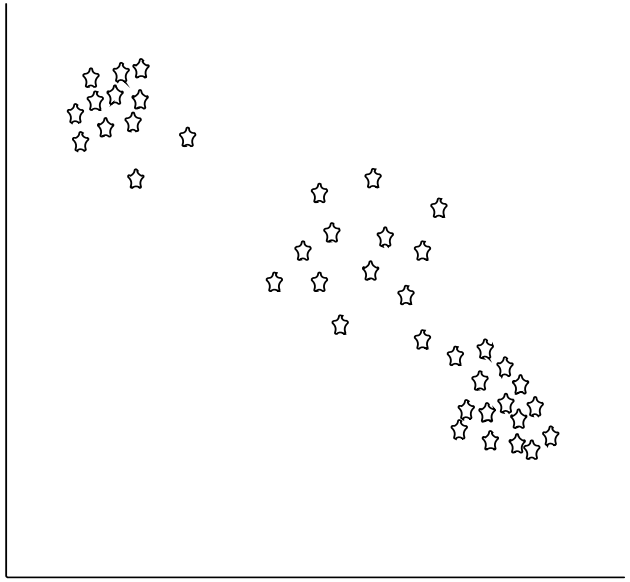
THE GENERAL FORM OF A CLUSTERING ALGORITHM



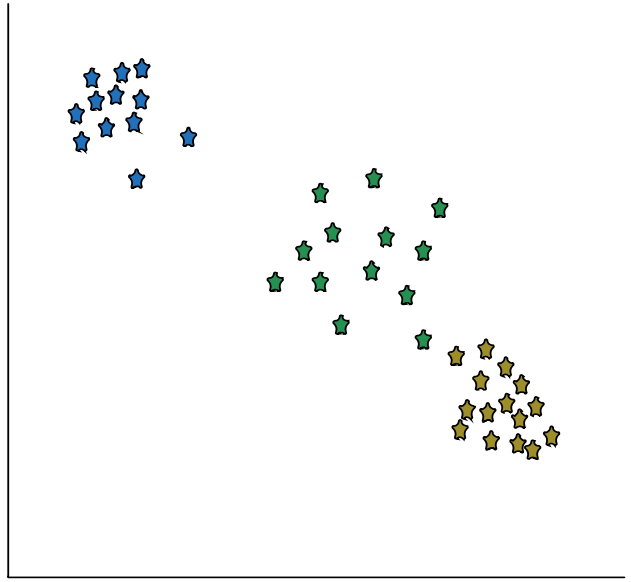
Initialization



Clustering Step A (Usually Repeated, Possibly in Conjunction with Next Step)



Clustering Step B (Usually Repeated, Possibly in Conjunction with Previous Step)

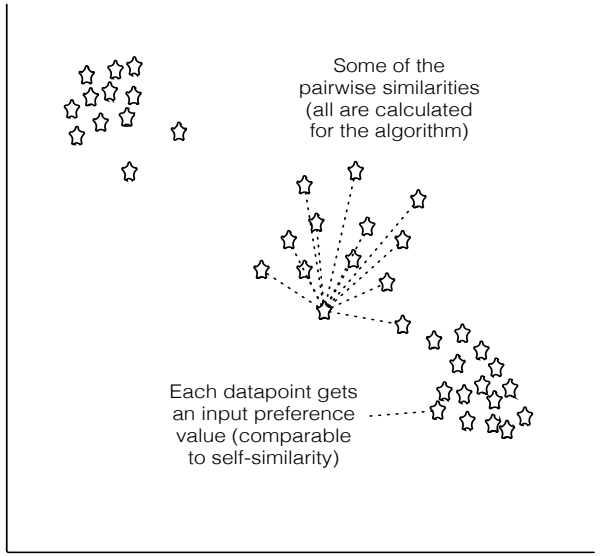


End Condition (Usually When Iterations of Steps A and B Produce Stable Results)

A COMPARISON OF START STATES OF SOME COMMON CLUSTERING

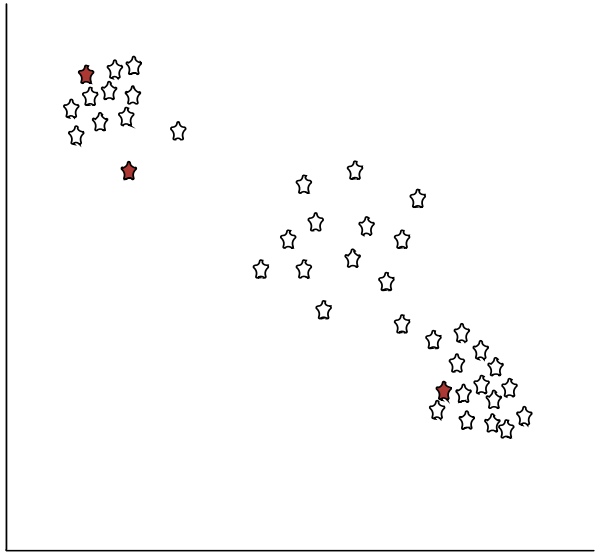
This gives a sense of why clustering results can be very different for different algorithms.

Affinity Propagation Initialization



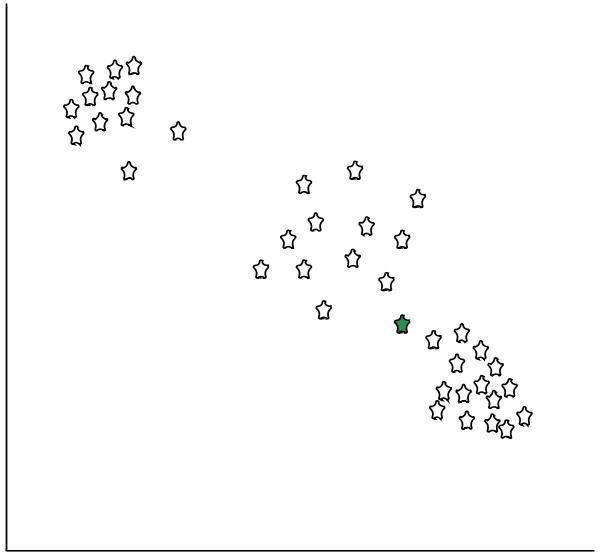
Calculate All Pairwise Similarities, Set Input Preference Values

K-Means Initialization



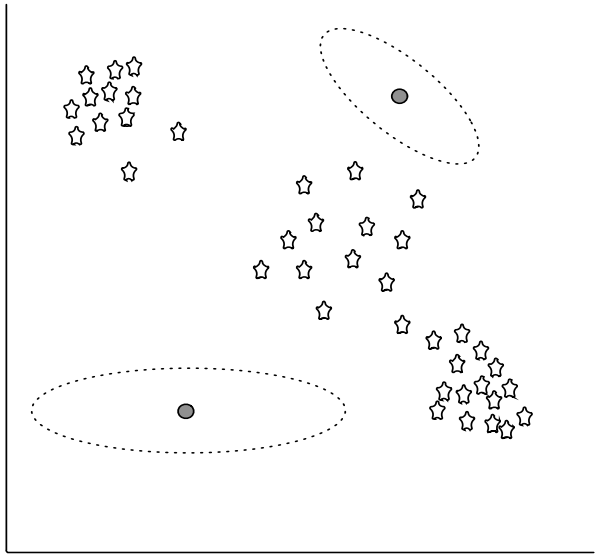
Randomly Pick k Centers

DBSCAN Initialization



Randomly Pick a DataPoint

Expectation Maximization Initialization

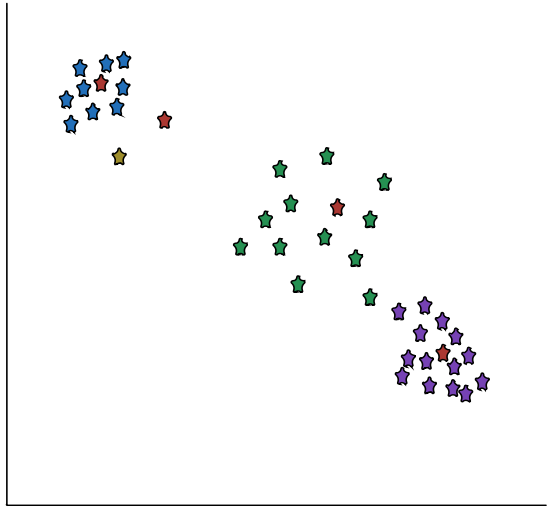


Set Initial Statistical Models

A COMPARISON OF INTERMEDIATE STATES OF COMMON CLUSTERING ALGORITHMS

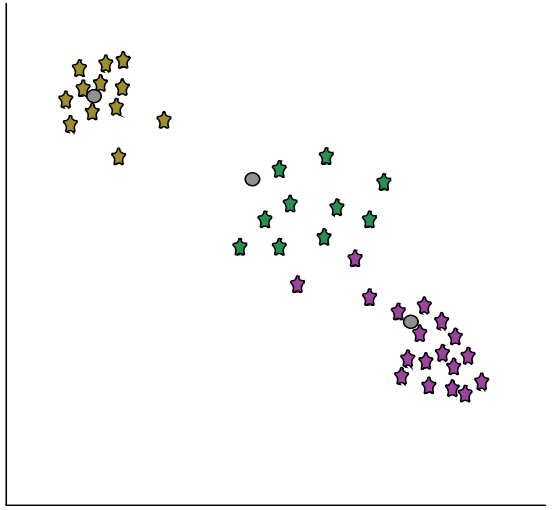
This gives a sense of why clustering results can be very different for different algorithms.

Affinity Propagation



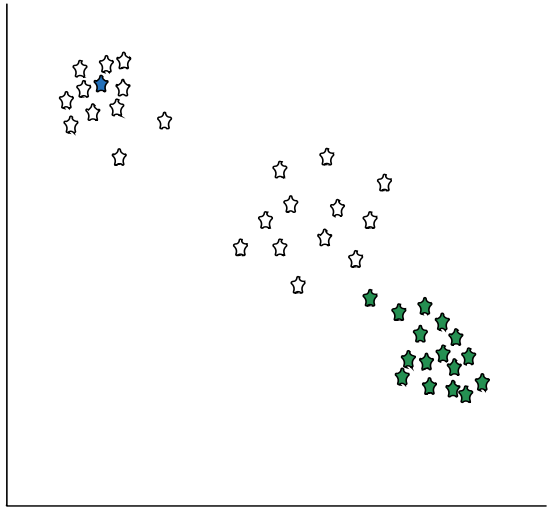
Select Good Over-All Exemplars (Based on Responsibility and Availability Scores). Assign Points to Clusters Based on Which Exemplars Are Most Suitable for Each Point

K-Means



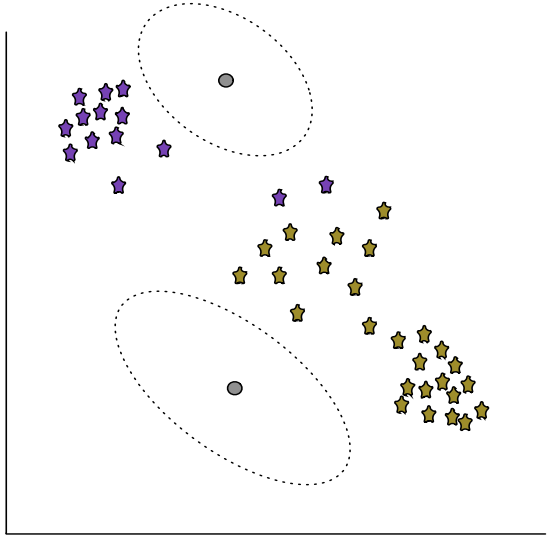
Re-assign Points Based on Centroids. Repeat from Previous Step (Calculate New Centroids) Until Stable

DBSCAN



Randomly Pick a New Unclustered Point and Try to Grow Another Cluster.

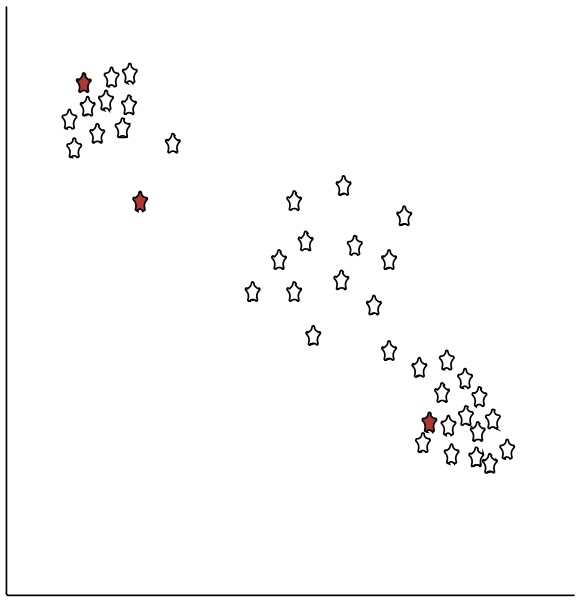
Expectation Maximization



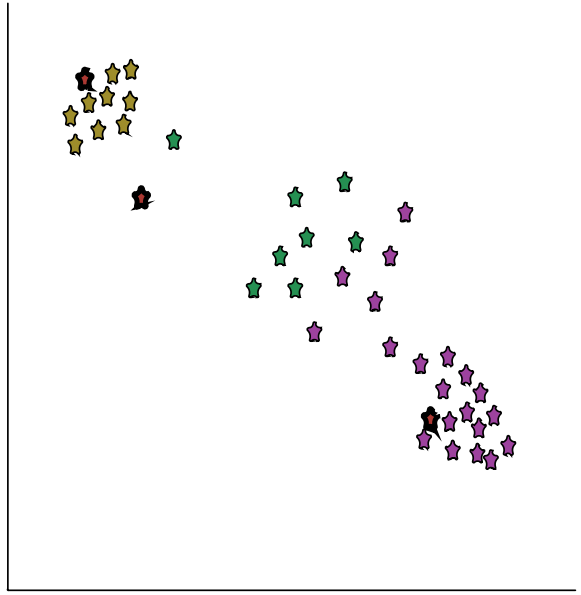
Adjust Clustering Assignment. Repeat from Previous Step (Adjust Models) Until Stable

IDLEWYLD Sysabee DAVHILL uOttaw

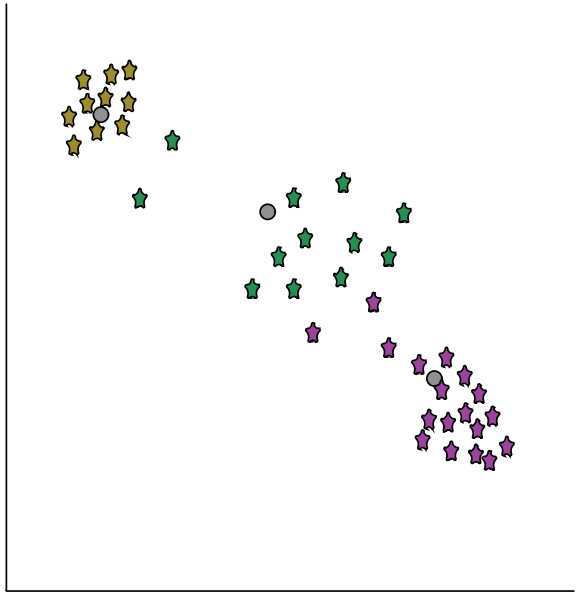
K-MEANS ALGORITHM GLOSS



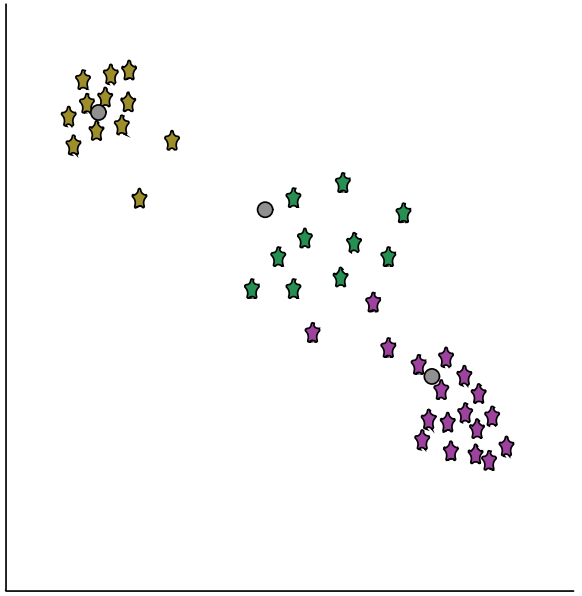
Initialization (e.g. Randomly Pick k Centers)



Assign Initial Clusters (Based on Distance to Centers)



Calculate Centroids of Clusters



Re-assign Points Based on Centroids. Repeat from Previous Step Until Stable

k-MEANS ALGORITHM

1. Select the desired **number of clusters**, say k
2. Randomly choose k instances as initial **cluster centres**
3. Calculate the **distance** from each observation to each centre
4. Place each instance in the cluster whose centre it is **nearest** to
5. Compute the **centroid** for each cluster
6. Repeat steps 3 – 5 with the new centroids
7. Repeat step 6 until the clusters are **stable**

k-MEANS STRENGTHS

Easy to implement (without having to actually compute pairwise distances).

- extremely common as a consequence
- elegant and simple

In many contexts, k-means is a **natural** way to look at grouping observations.

Helps provide a **basic understanding of the data structure** in a first pass.

k-MEANS LIMITATIONS

Data points can only be assigned to **one** cluster.

- this can lead to overfitting
- robust solution: consider the **probability** of belonging to each cluster

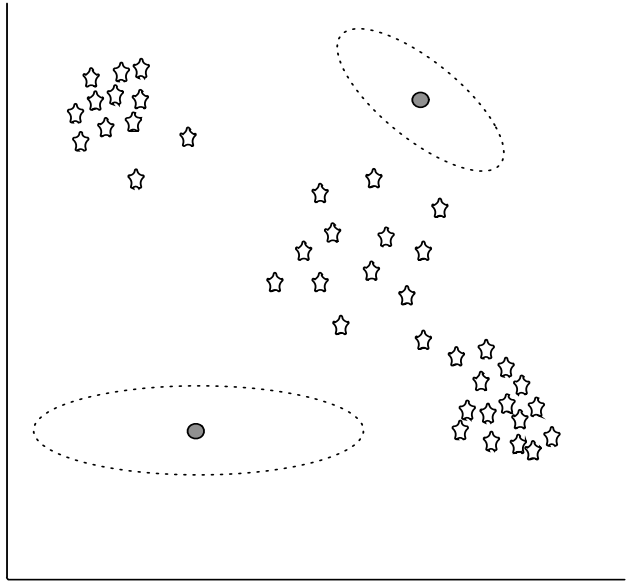
Underlying clusters are assumed to be **blob-shaped**

- k-means will fail to produce useful clusters if that assumption is not met in practice

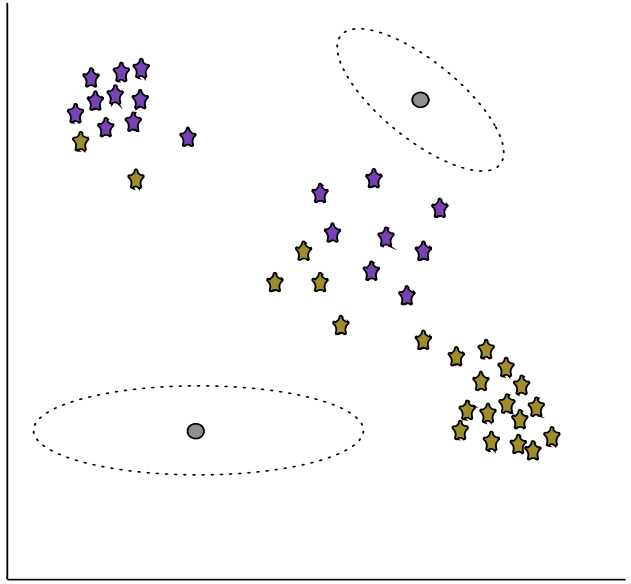
Clusters are assumed to be separate (discrete)

- k-means does not allow for **overlapping** or **hierarchical** groupings

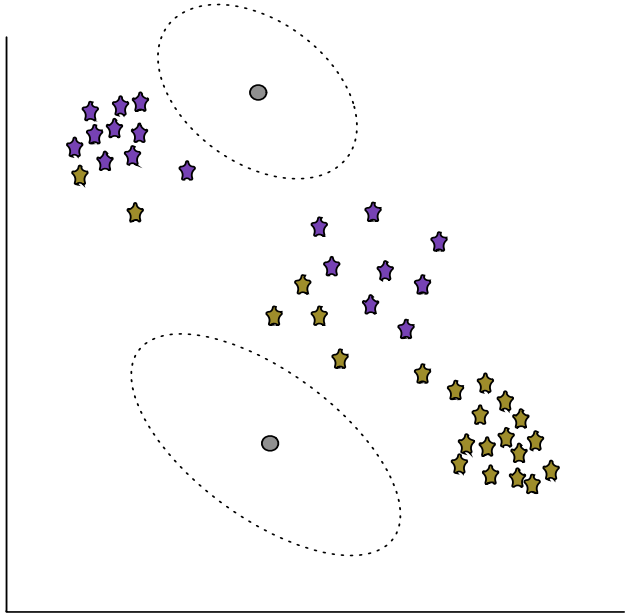
EXPECTATION MAXIMIZATION ALGORITHM GLOSS



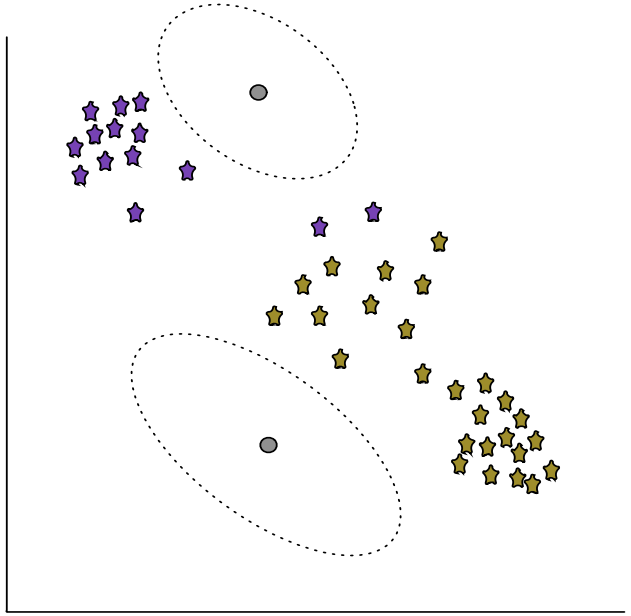
Initialization (Set Initial Statistical Models)



Assign Clusters Based on Models

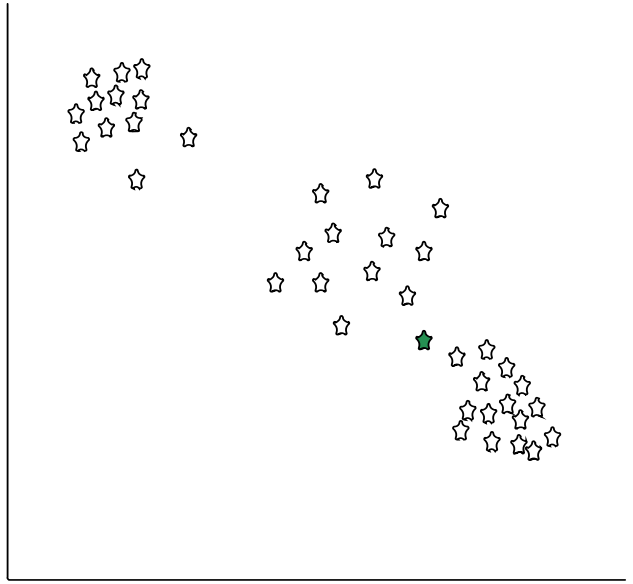


Adjust Statistical Models

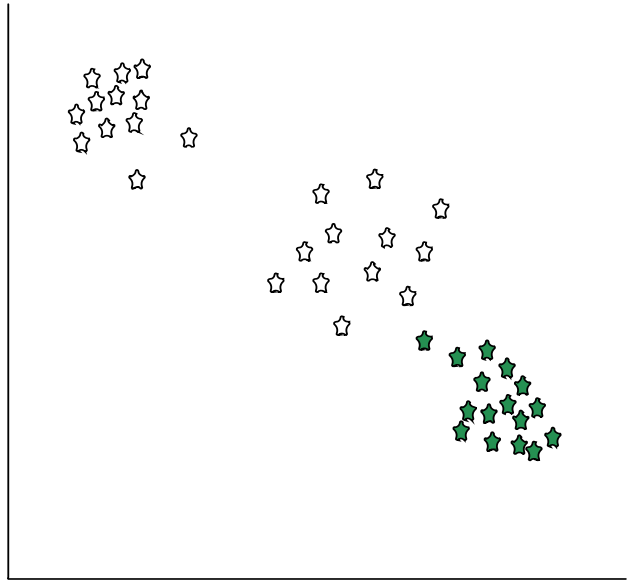


Adjust Clustering Assignment.
Repeat from Previous Step Until Stable

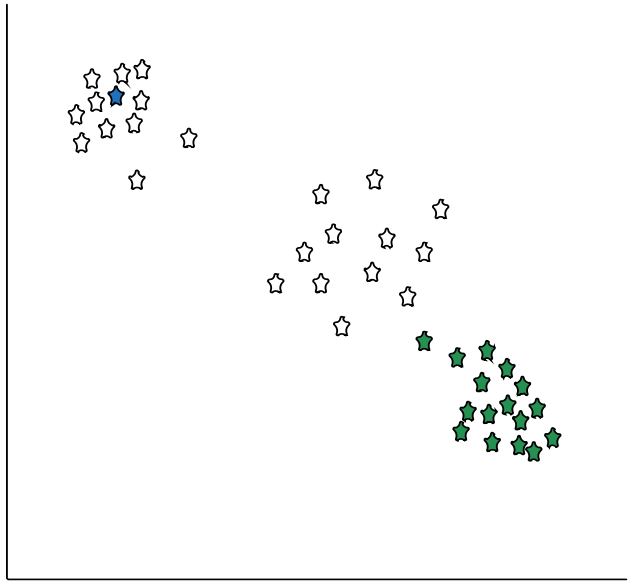
DBSCAN ALGORITHM GLOSS



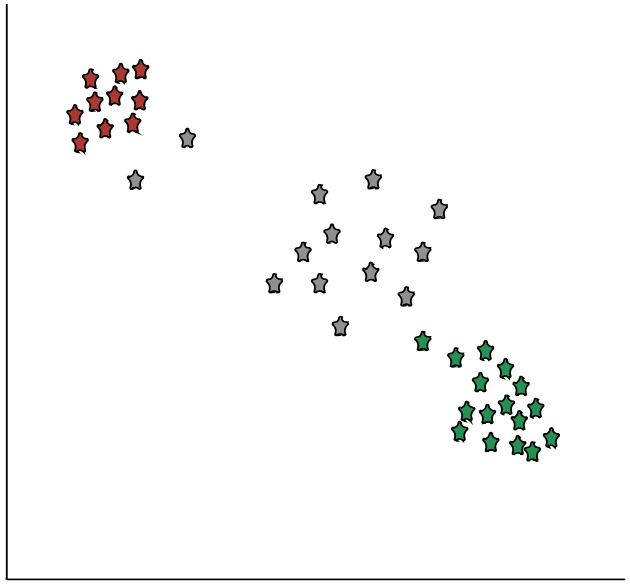
Initialization (Randomly Pick a DataPoint)



(Try to) Grow A Cluster. Stop When There Are No More Close Points

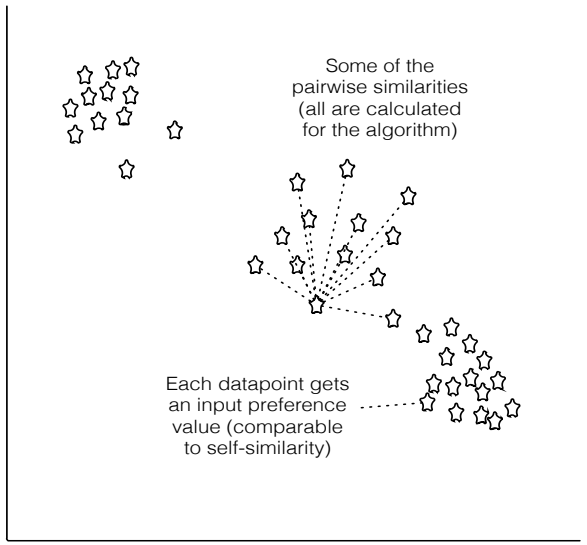


Randomly Pick a New Unclustered Point and Try to Grow Another Cluster.

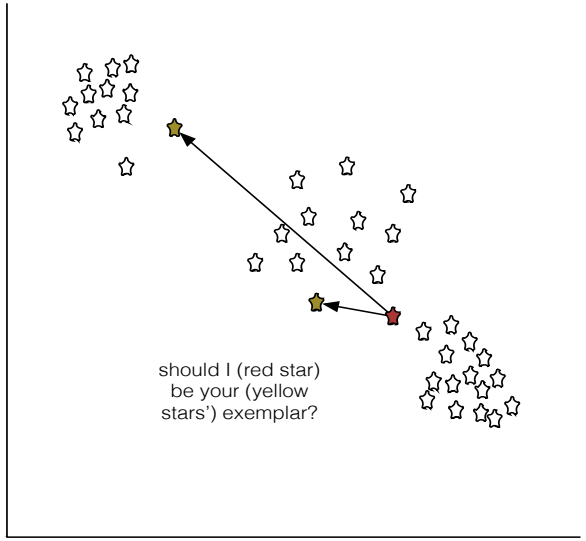


Stop Clustering When All Points Are Clustered (or Marked as Anomalies)

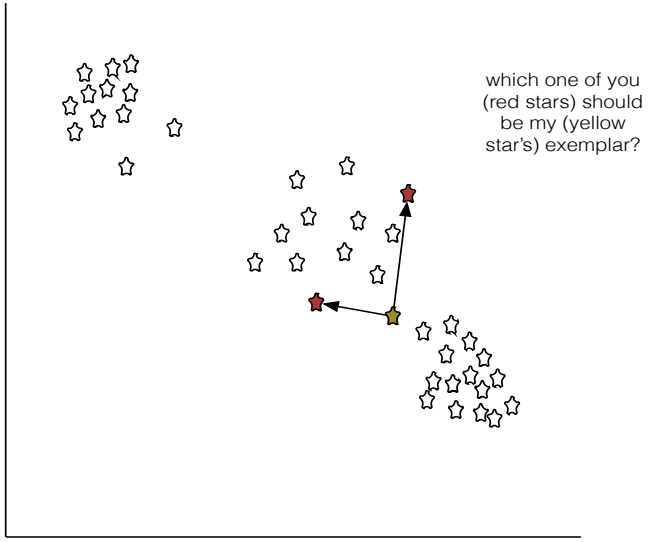
AFFINITY PROPAGATION GLOSS



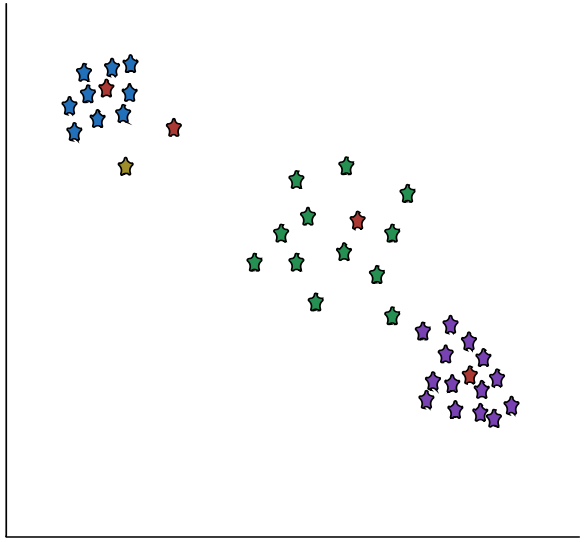
Initialization (Calculate All Pairwise Similarities, Set Input Preference Values)



Consider, for Each Point, its Suitability to Be an Exemplar For Each Other Point (Responsibility Score)



Consider, For Each Point, The Suitability of Every Other Point To Be an Exemplar for that Point (Availability Score - Based Also on Responsibility). Iterate Between This And Previous Step Until Stable



Select Good Over-All Exemplars (Based on Responsibility and Availability Scores). Assign Points to Clusters Based on Which Exemplars Are Most Suitable for Each Point

DISCUSSION

In what way does the choice of the algorithm(s) (and distance metric and associated parameters) depend on the available data and data types?

CLUSTERING VALIDATION

CLUSTERING

CLUSTERING VALIDATION

What does it mean for a clustering scheme to be **better** than another?

What does it mean for a clustering scheme to be **valid**?

What does it mean for a single cluster to be **good**?

How many clusters are there in the data, really?

Right vs. wrong is meaningless: seek **optimal vs. sub-optimal**.

CLUSTERING VALIDATION

Optimal clustering scheme:

- maximal separation between clusters
- maximal similarity within groups
- agrees with human eye test
- useful at achieving its goals

Validation types

- external (uses additional information)
- internal (uses only the clustering results)
- relative (compares across clustering attempts)

DISCUSSION

The main clustering challenge is that we don't know what we are comparing the resulting clustering scheme **against** (versions of this problem plague unsupervised tasks).

So why bother with clustering in the first place?

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?

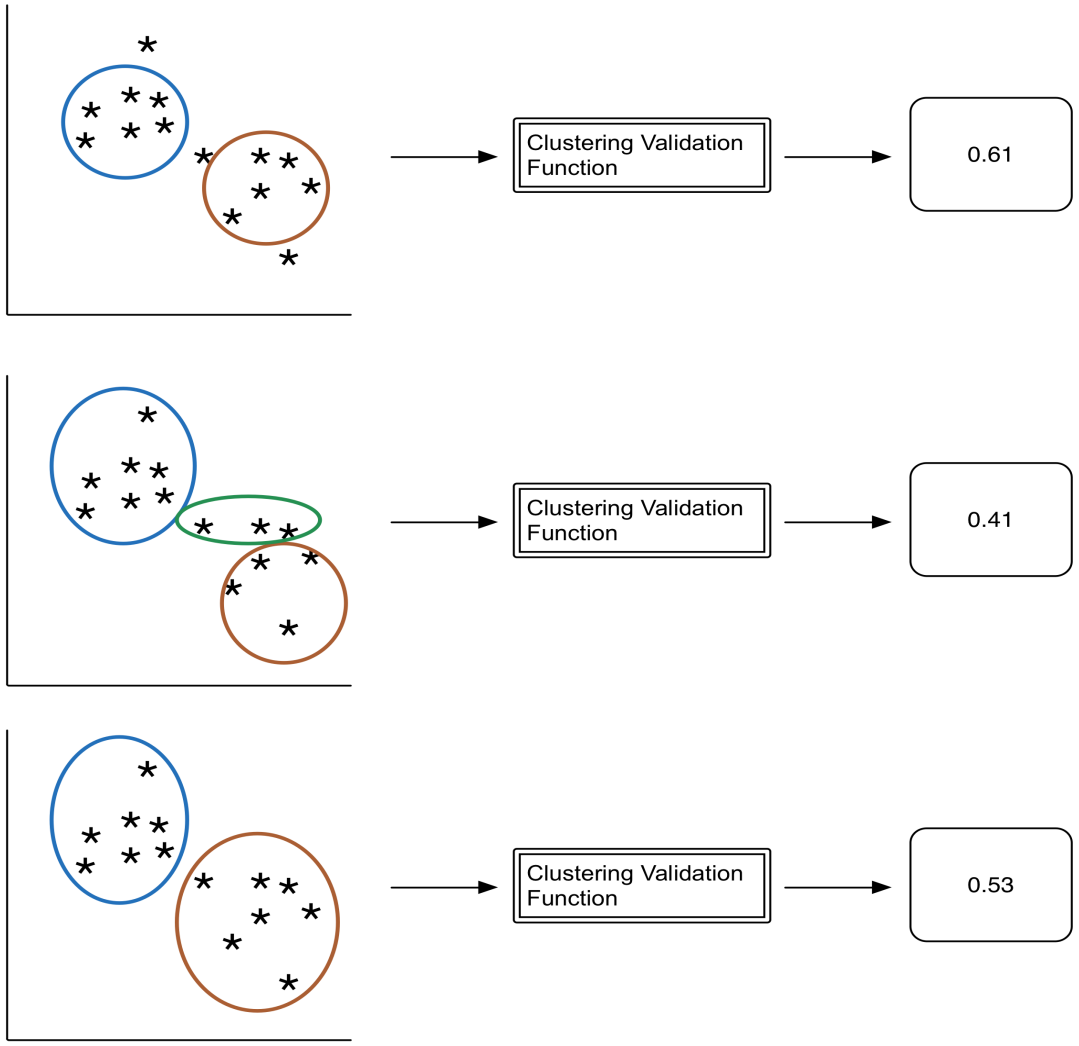
Will some clusterings be of (objectivel higher ***quality*** than others?

MAKING CONCEPTS CONCRE

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



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

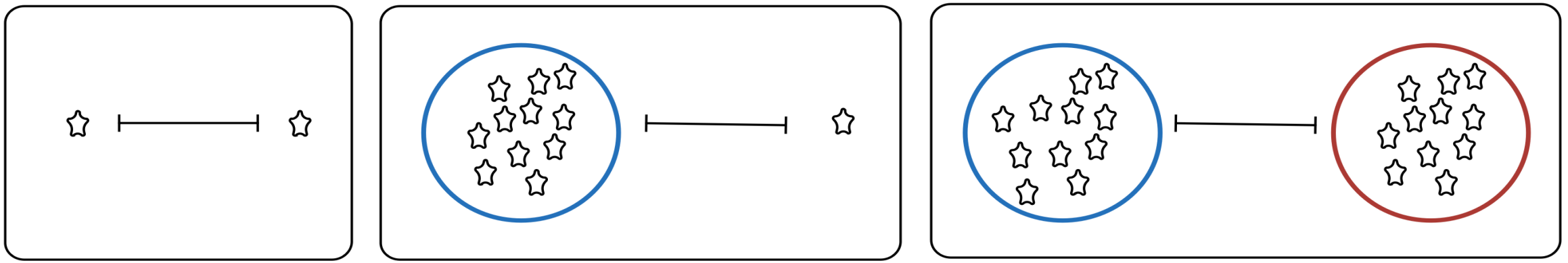
- Input: Instances + Cluster Assignments (+ similarity matrix, usually)
- Output: A numeric value

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**

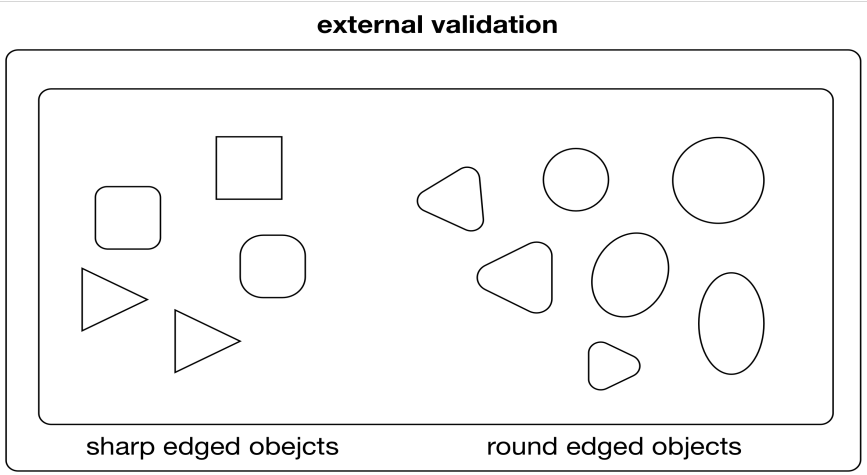


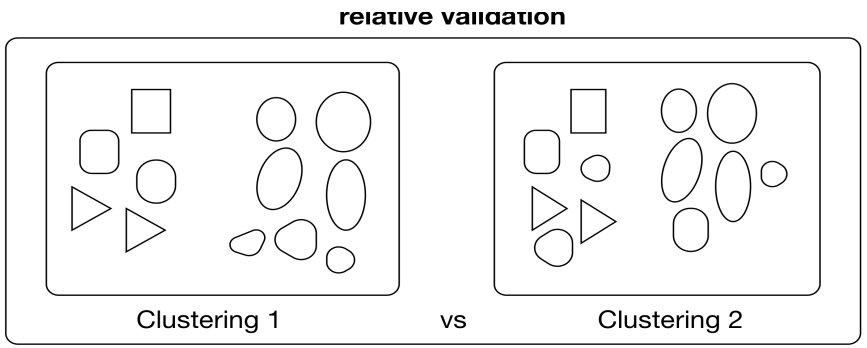
THREE TYPES OF VALIDATION

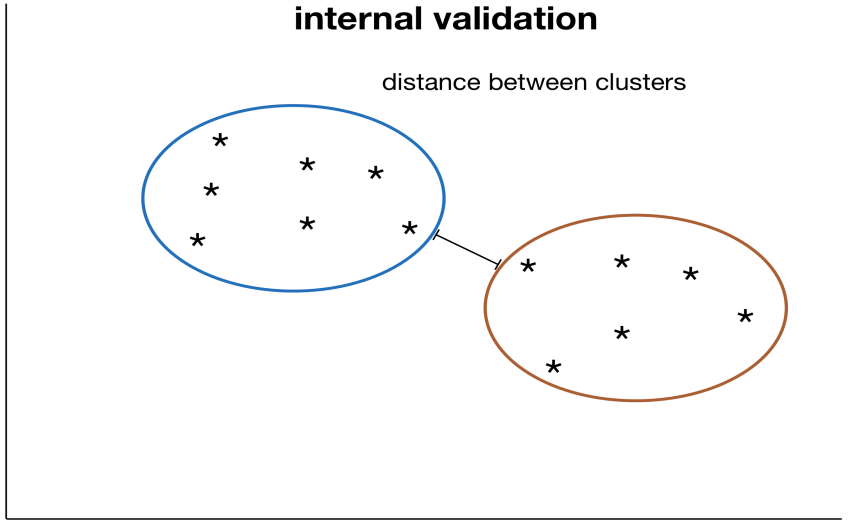
Internal Validation: Based only on properties available within a single clustering result (note that this comprises multiple clusters)

Relative Validation: Comparison of one (entire) clustering result with another

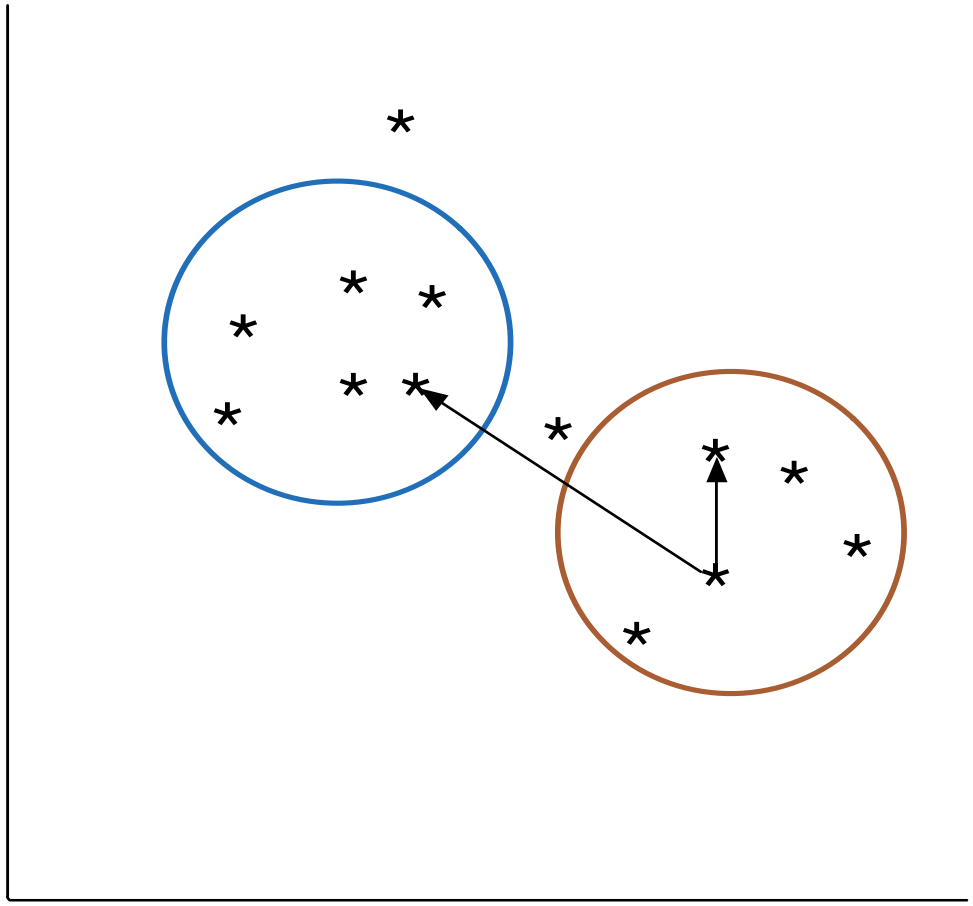
External Validation: Comparison of a (single) clustering result with some external standard







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

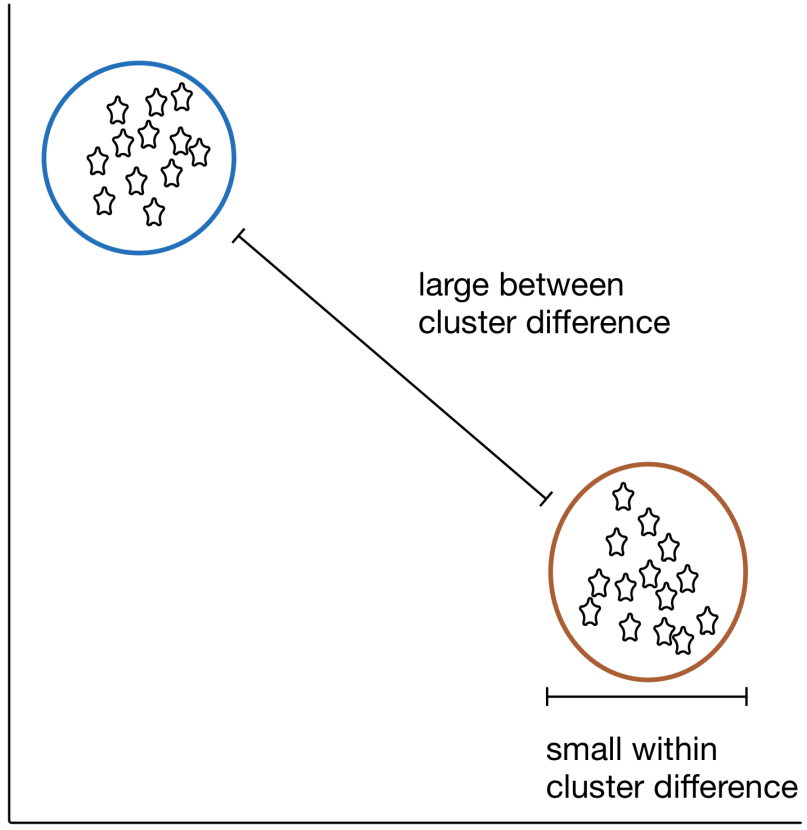
VERY BROAD GOALS

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).

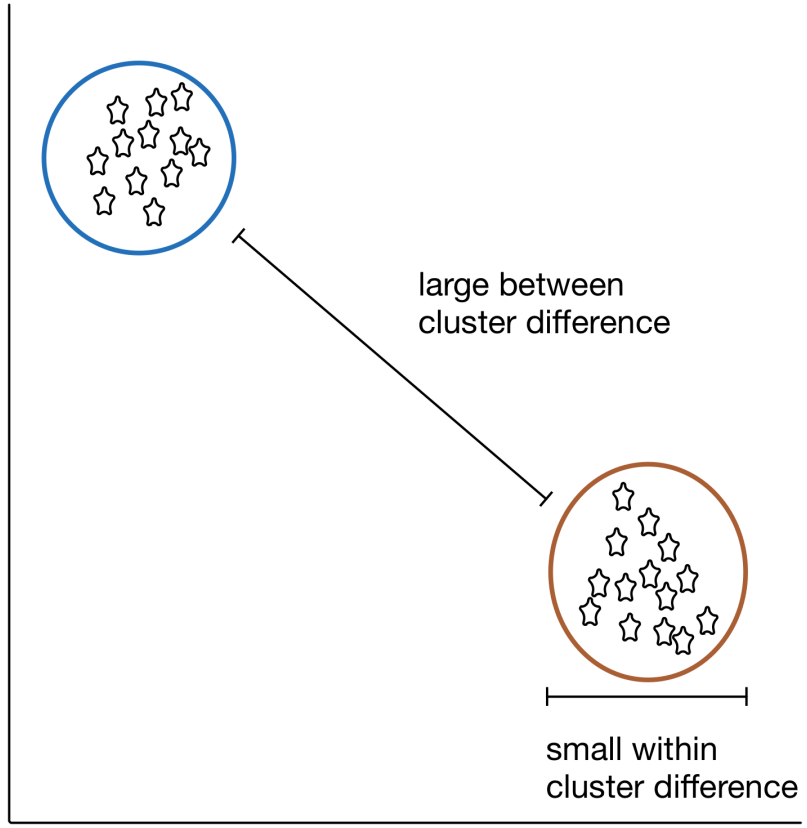
Thus the large number of CQMs.



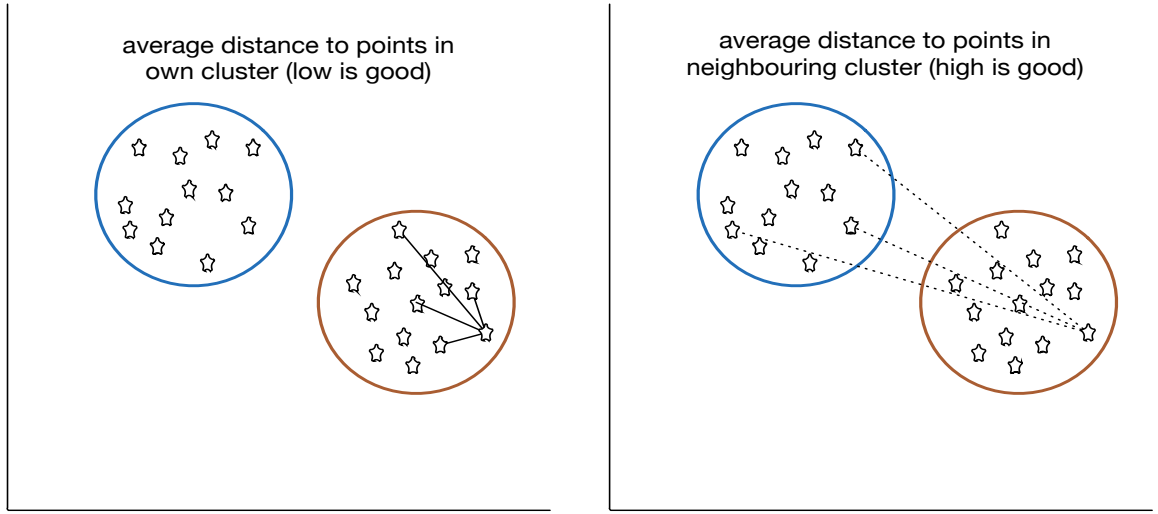
VERY BROAD GOALS

Question: is this trade-off (and the resulting CQMs) really context independent?

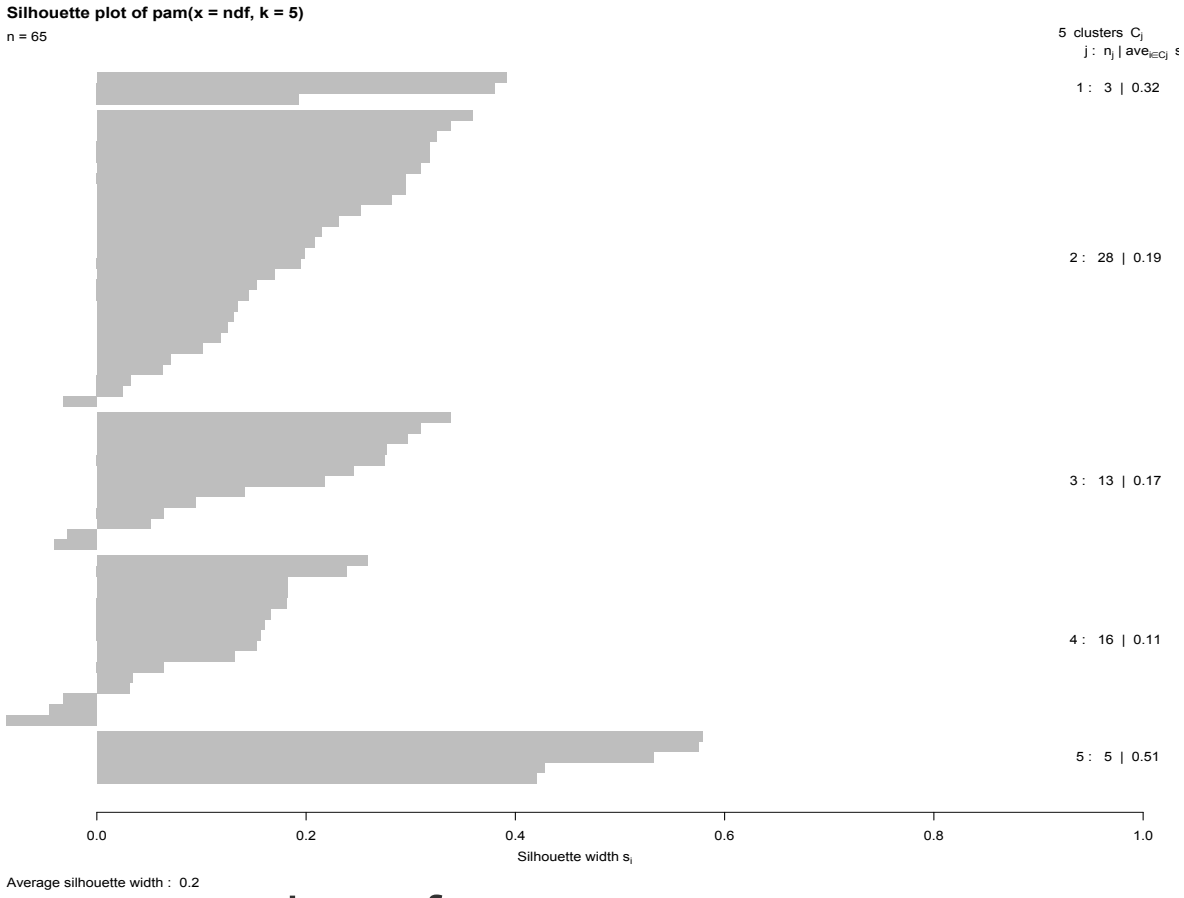
Maybe different weightings are more relevant in different contexts?



SILHOUETTE INDEX



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



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

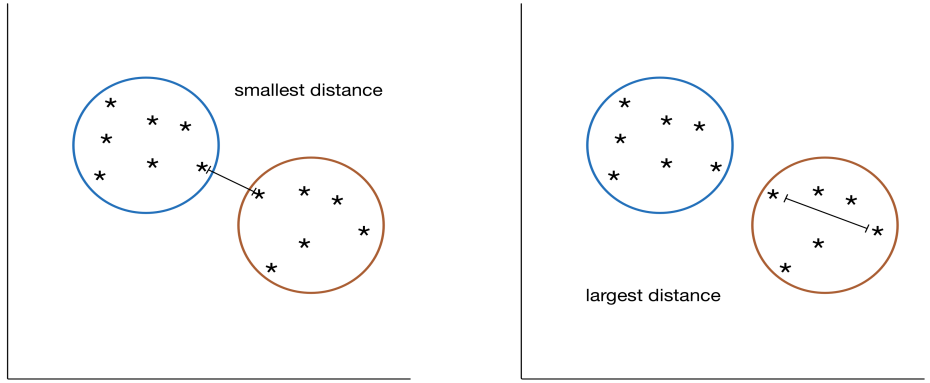
DUNN'S INDEX

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.



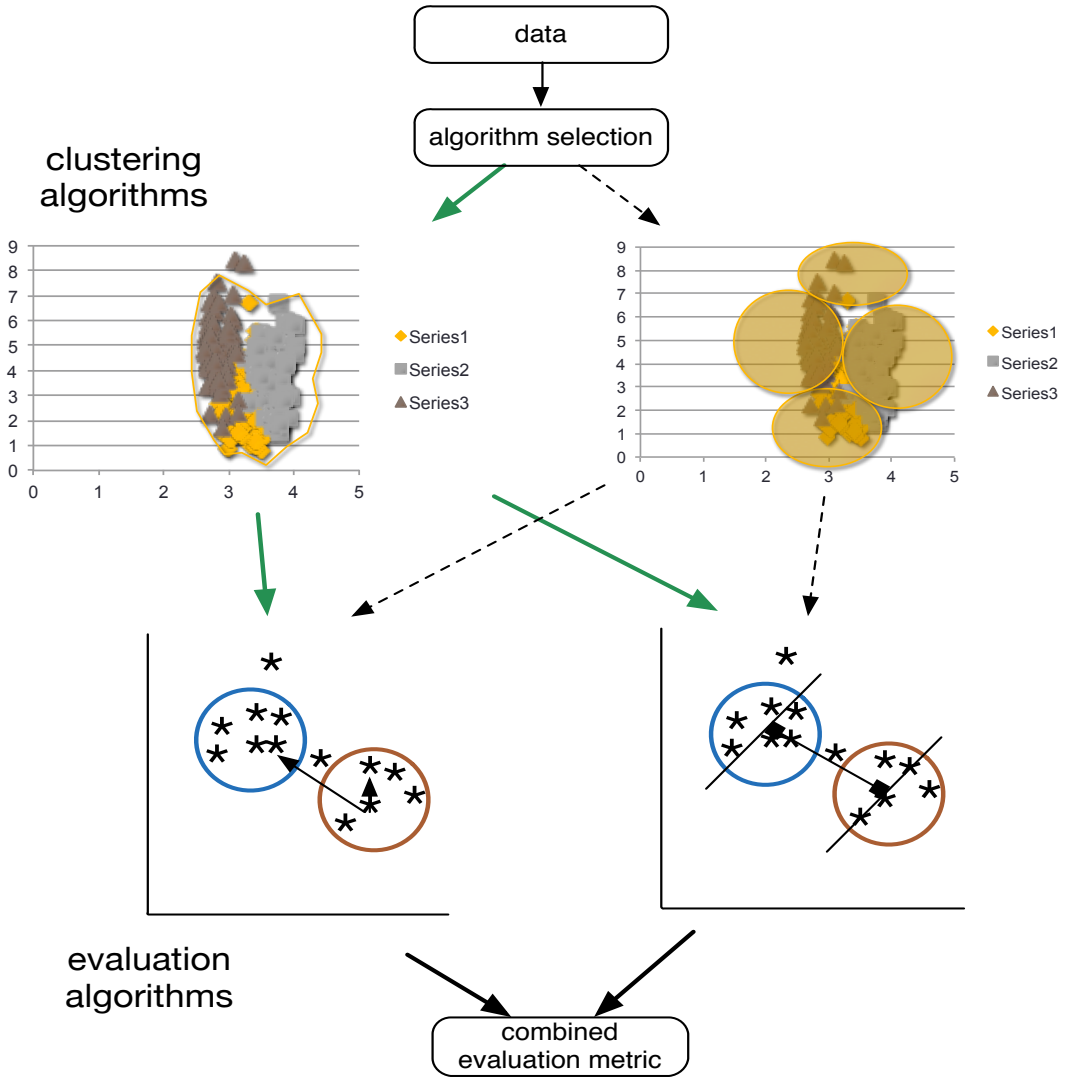
Comparison with Silhouette Index: In a sense, a simpler measure. More of a whole cluster measure, rather than a point by point measure. Evaluates based on extremes (max, min).

MORE IS BETTER (RELATIVE)?

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.



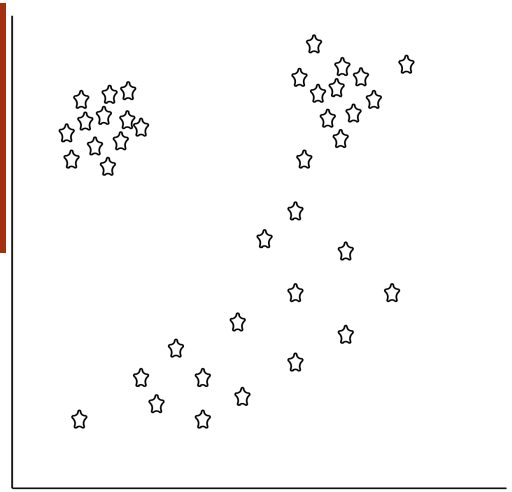
STABILITY

Some options:

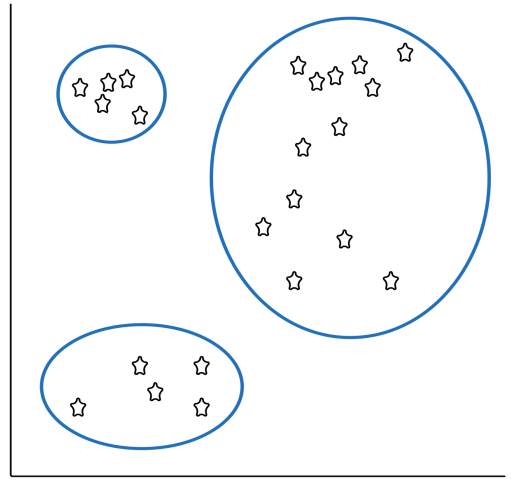
- multiple datasets sampled from same source
- different columns used to generate clusters (i.e. drop a different column each time)

Similarity of results is measured.

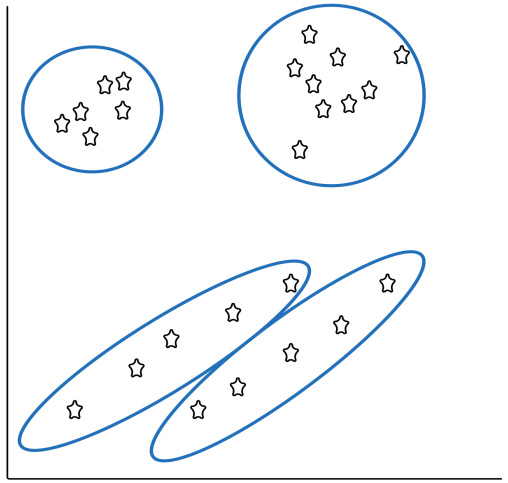
If results are not stable across clusterings, further investigation required.



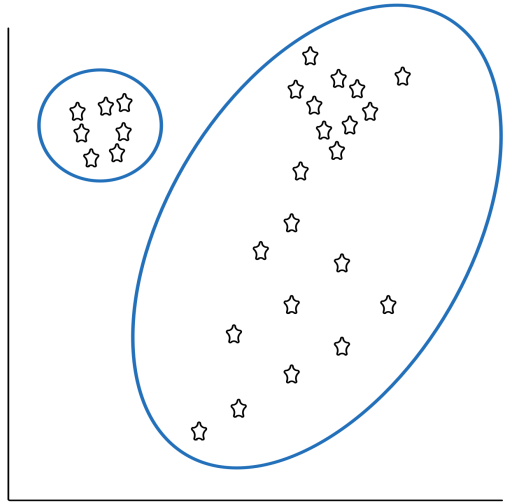
Full dataset



Sample 1 clustering



Sample 2 clustering



Sample 3 clustering

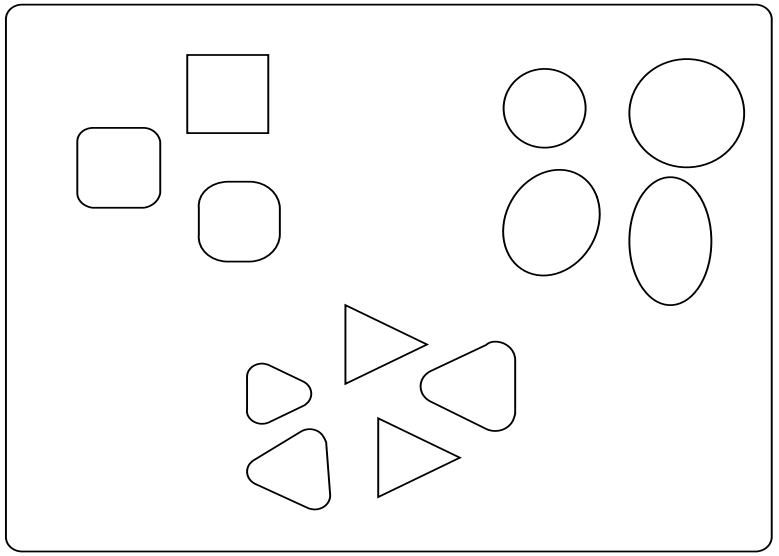
BACK TO CONTEXT (EXTERNAL)

Brings in outside information to evaluate the clusters.

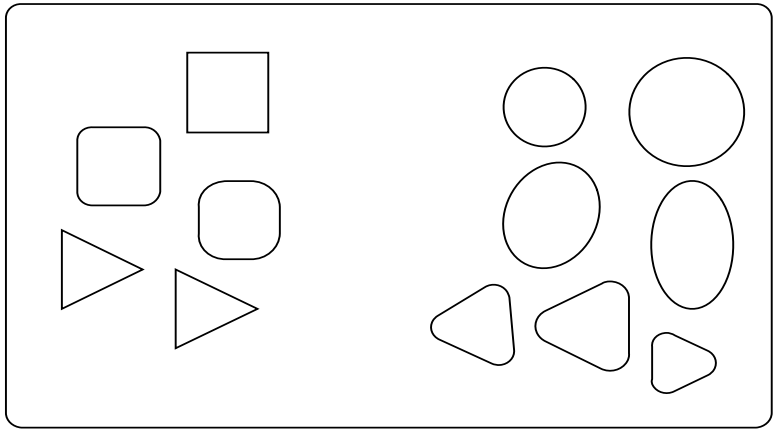
Outside information is typically the ‘correct’ class.

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

PURITY (EXTERNAL)

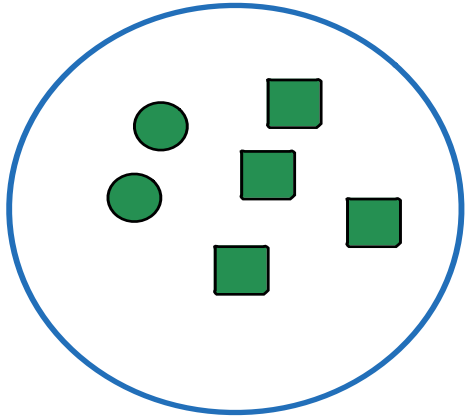
For this metric each cluster is assigned to the class which is most frequent in the cluster.

To calculate the purity: number of correctly assigned points / number of points in the cluster.

Some other options: precision, recall.

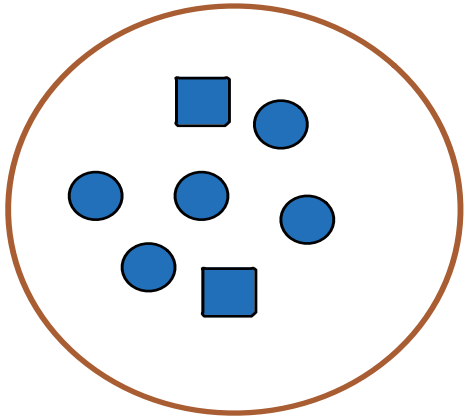
Assuming we are interested in shape…

SQUARE CLUSTER



purity = 66%

CIRCLE CLUSTER



purity = 71%

TRY AND TRY AGAIN

Diversity in clustering validation techniques.

Be aware of the types of validation, and variations within types.

Seek agreement across techniques.

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**.

NOTES

CLUSTERING

"Woes clusters. Rare are solitary woes; they love a train, they tread each other's heel"
(Edward Young)

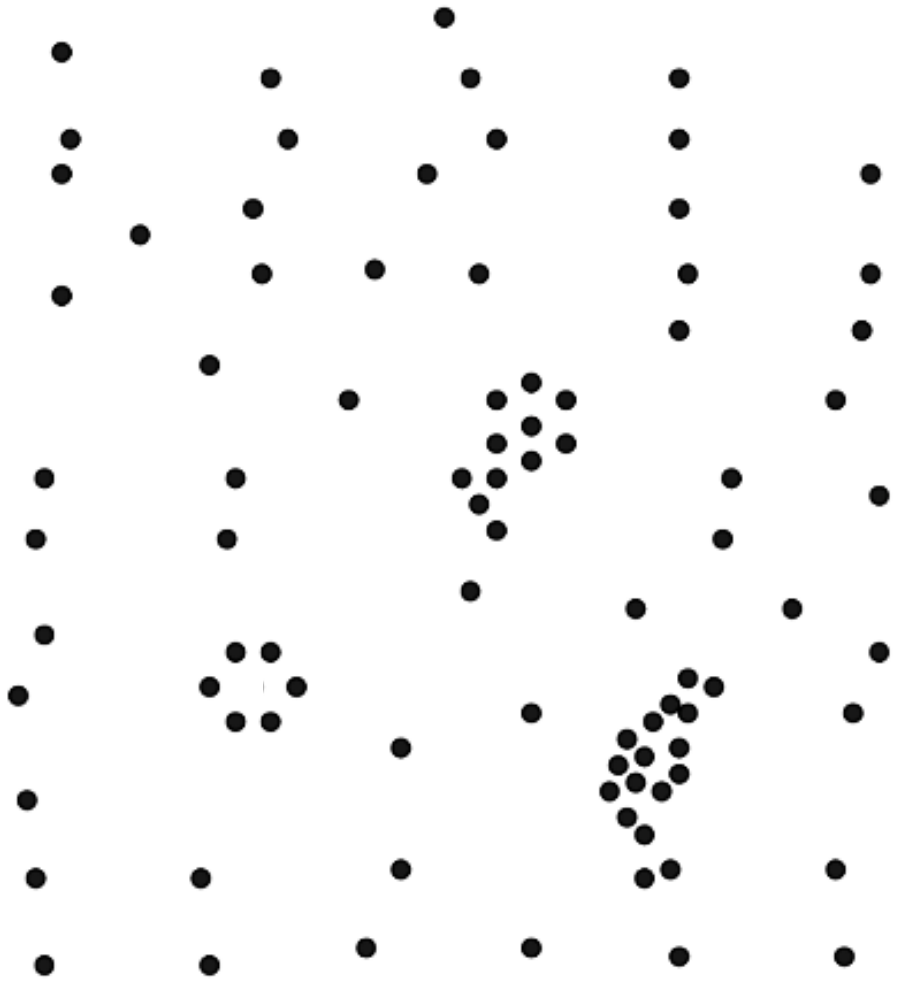
CLUSTERING CHALLENGES

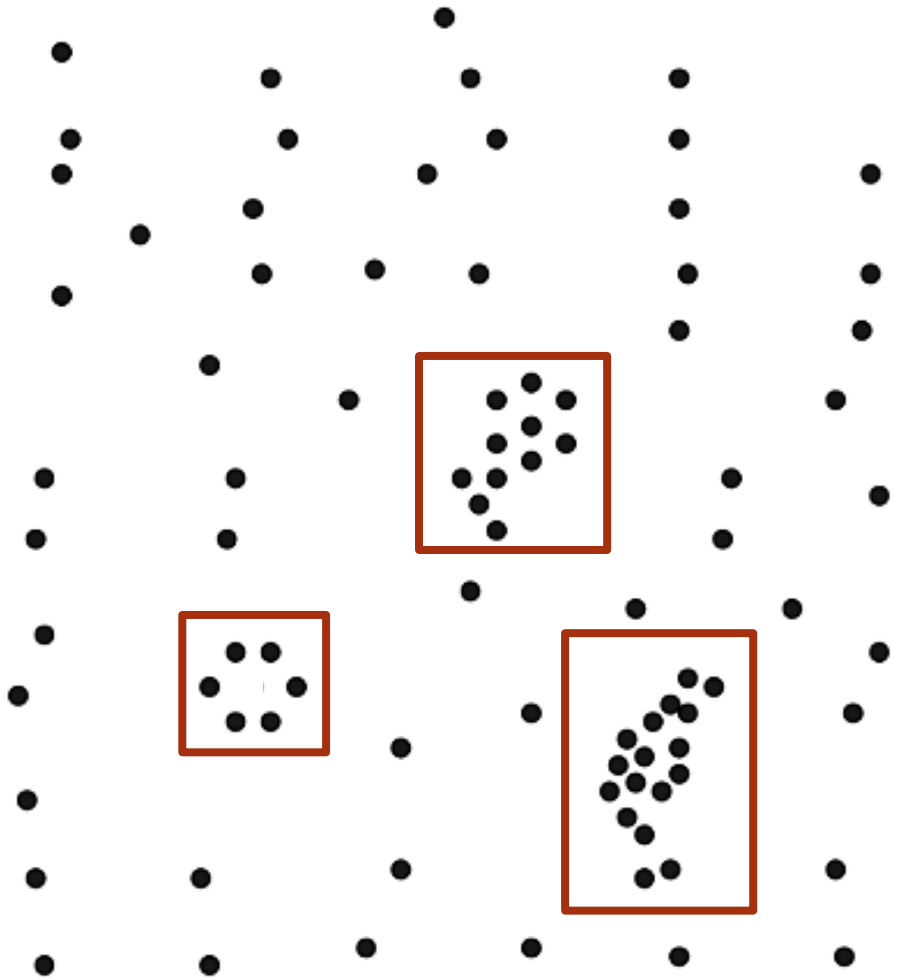
Automation
relatively intuitive for humans, but harder for machines

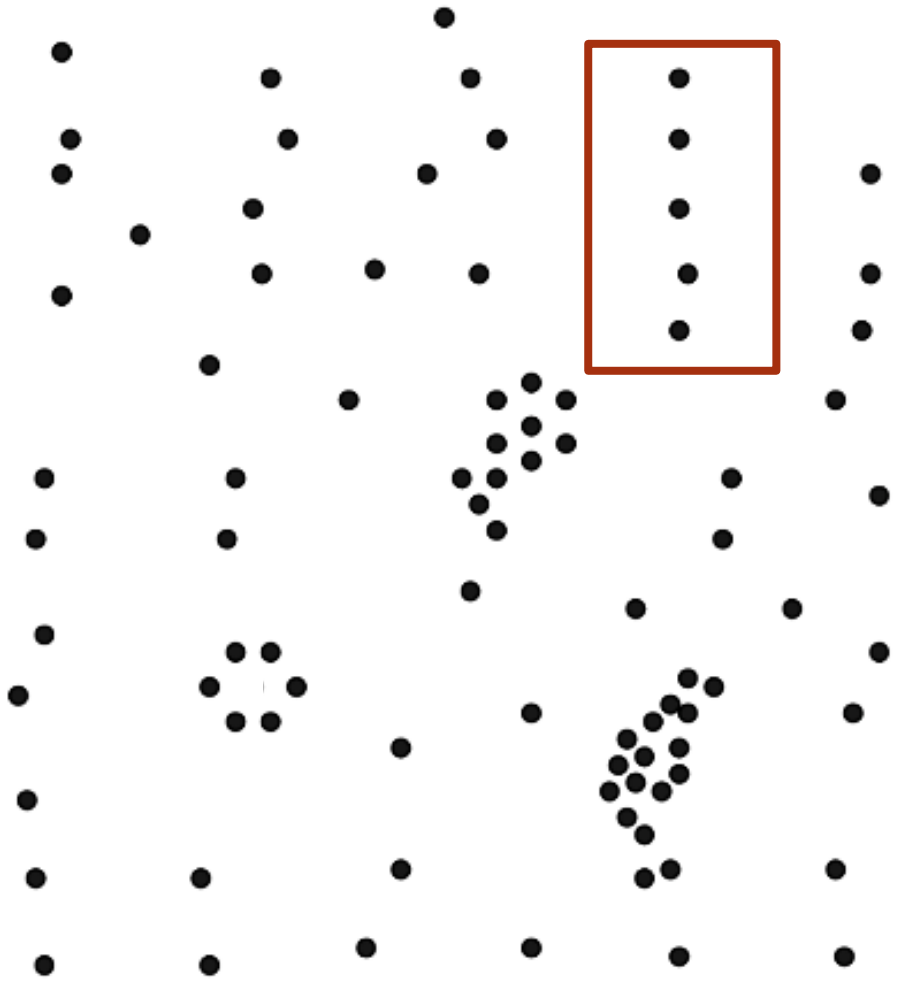
Lack of a clear-cut definition
no universal agreement as to what constitutes a cluster

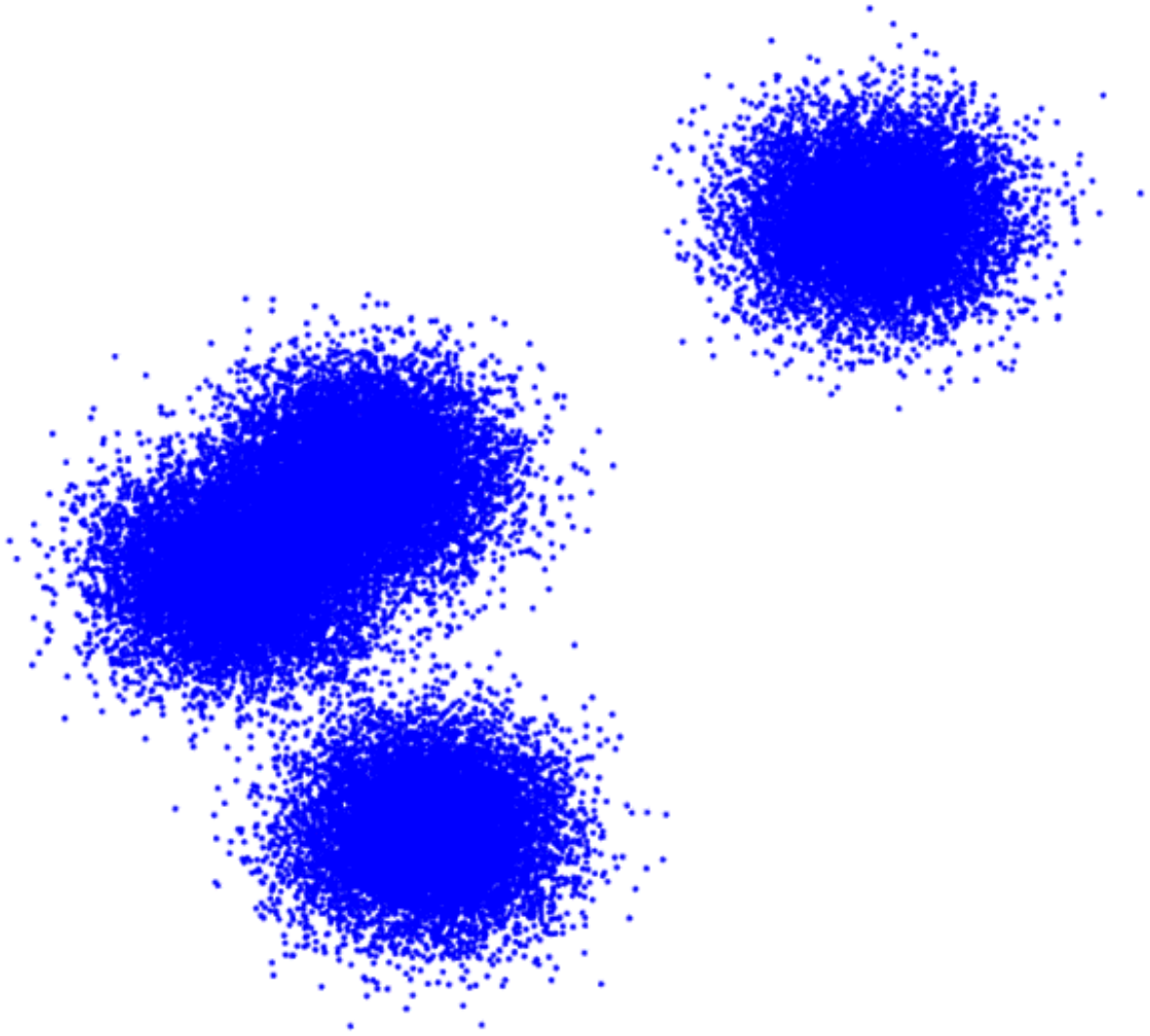
Lack of repeatability
non-deterministic: the same algorithm, applied twice to the same dataset can discover completely different clusters

Number of clusters
optimal number of clusters difficult to determine









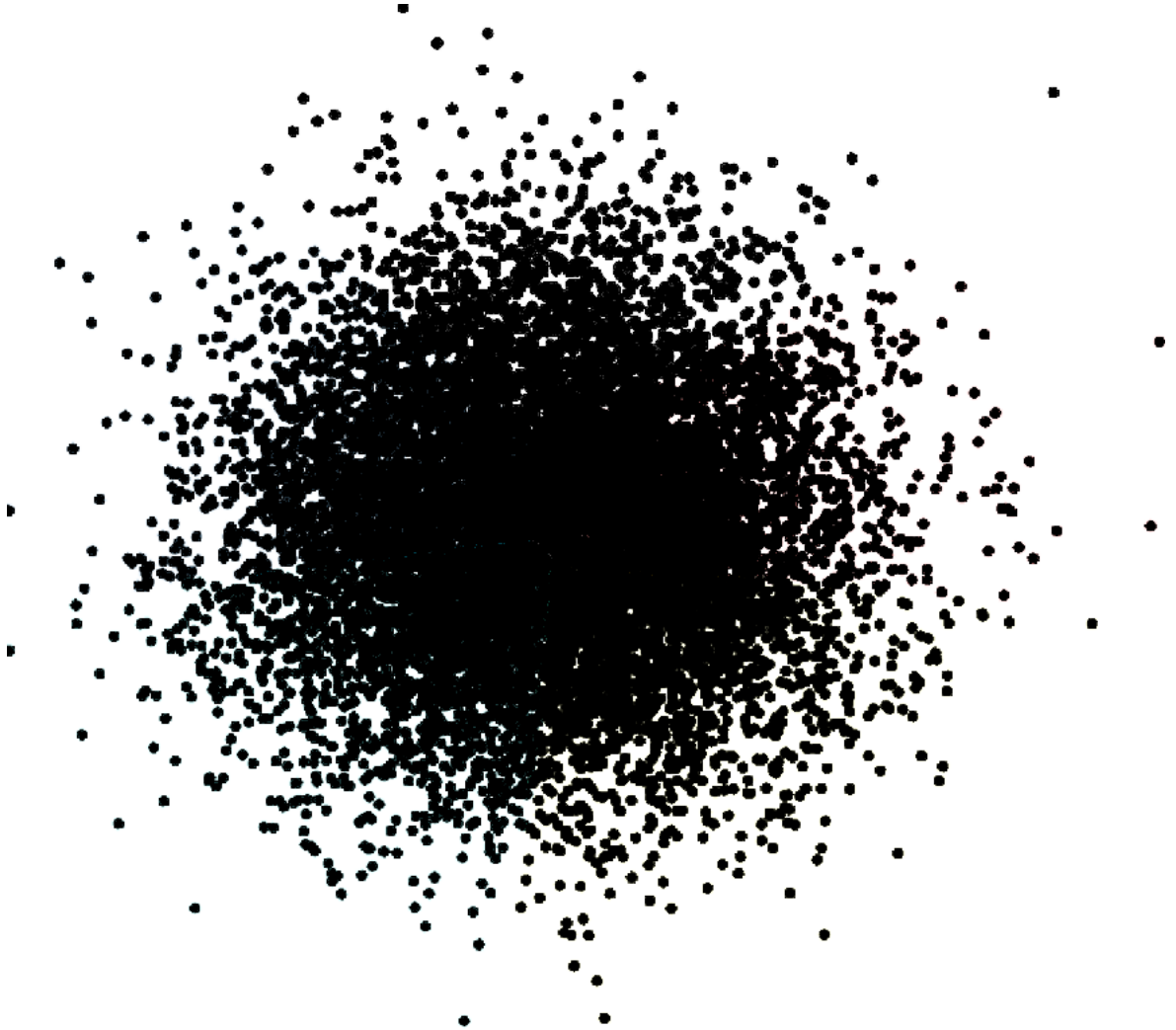
CLUSTERING CHALLENGES

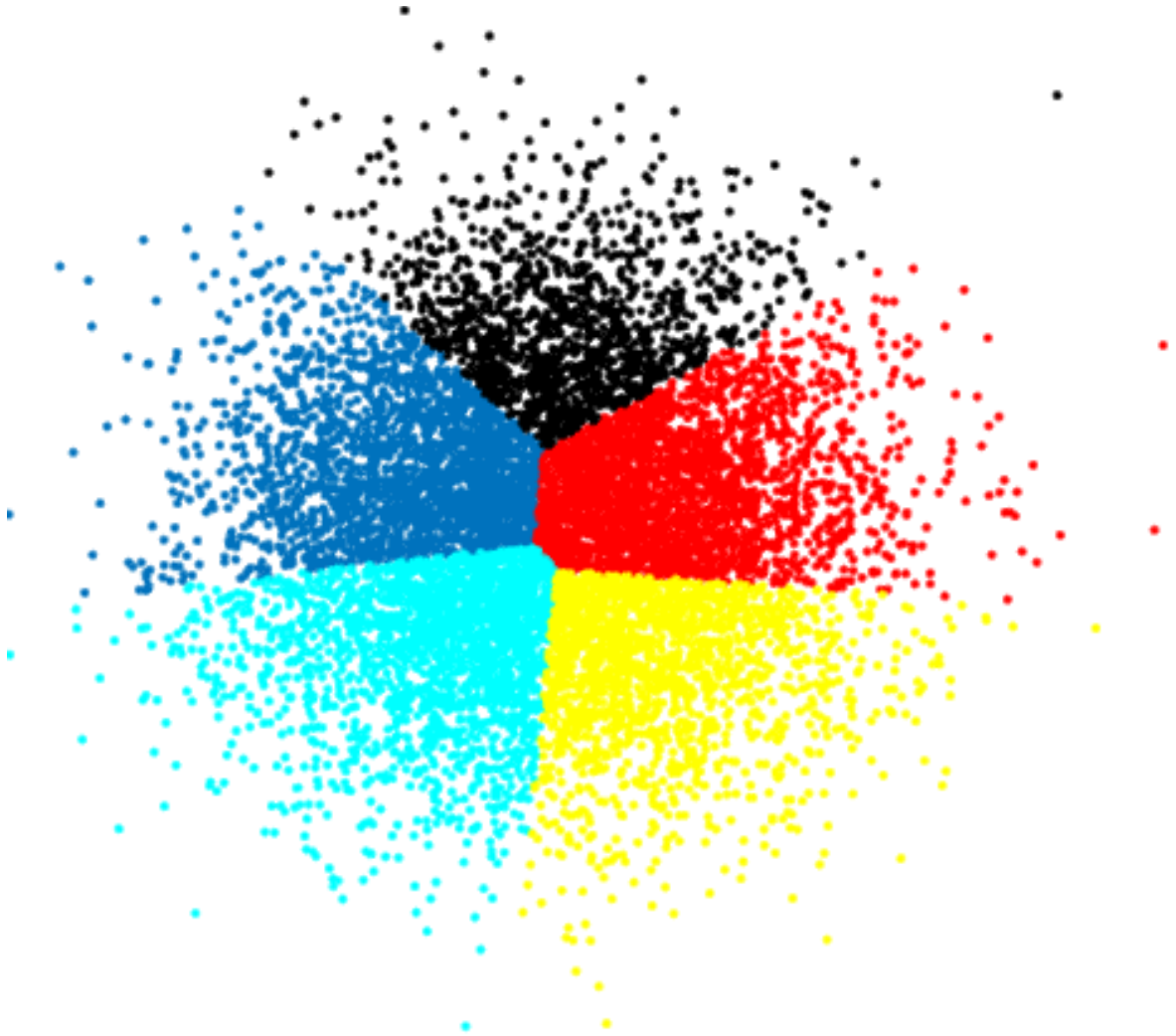
Cluster description
should clusters be described using representative instances or average values?

Model validation
no true clustering information against which to contrast the clustering scheme, so how do we determine if it is appropriate?

Ghost clustering
most methods will find clusters even if there are none in the data

***A posteriori* rationalization**
once clusters have been found, it is tempting to try to "explain" them ...





EXAMPLE: IRISES

CLUSTERING

"Data science students don't have to be gardeners, but it helps."

(author unknown)

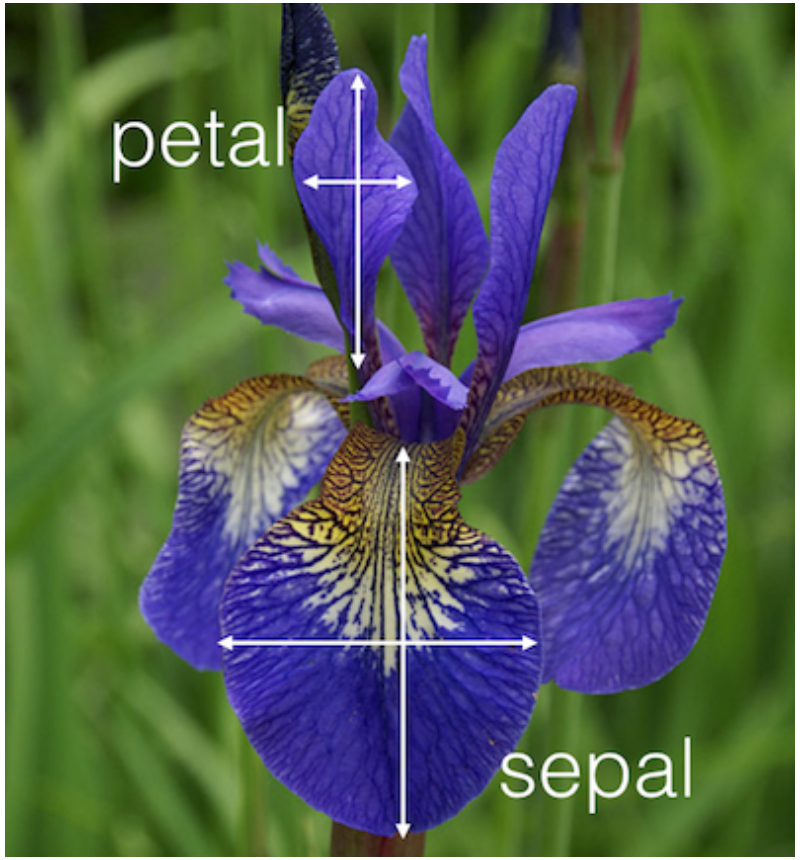
[http://blog.kaggle.com/wp-content/uploads/2015/04/iris_petal_sepal.png]

EXAMPLE – IRIS DATASET

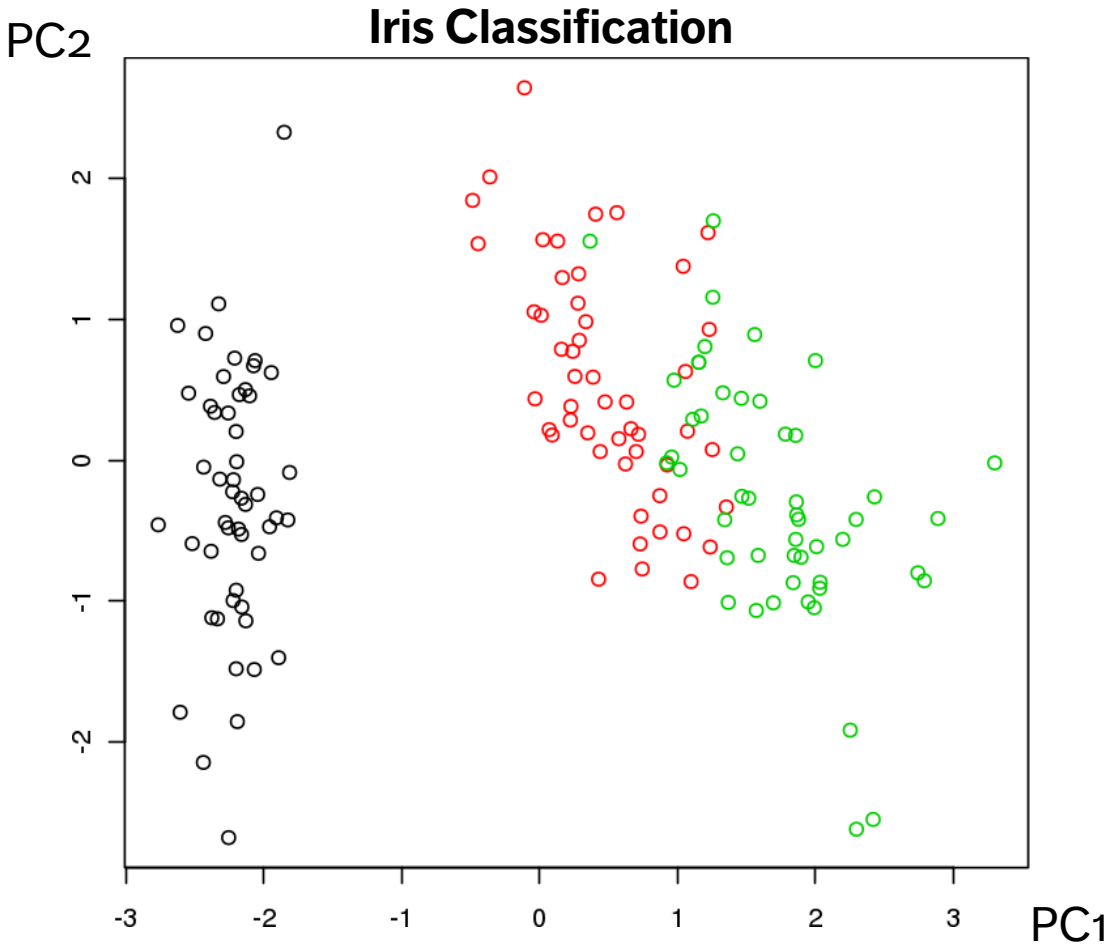
Iris is a genus of plants with showy flowers.

Fisher's iris dataset contains 150 observations of 5 attributes for specimens collected by Anderson, mostly from a Gaspé peninsula's pasture in the 1930s:

- **petal width**
- **petal length**
- **sepal width**
- **sepal length**
- **species**



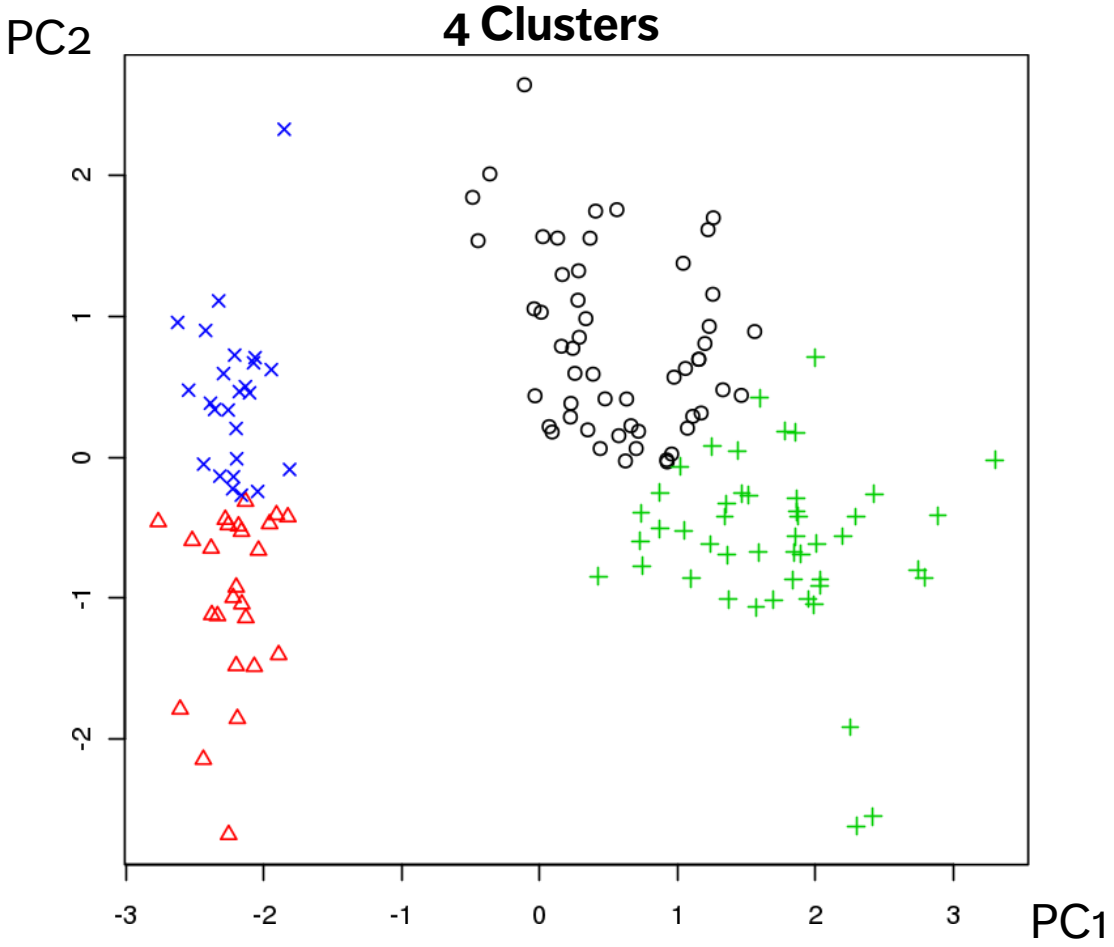
EXAMPLE – IRIS DATASET



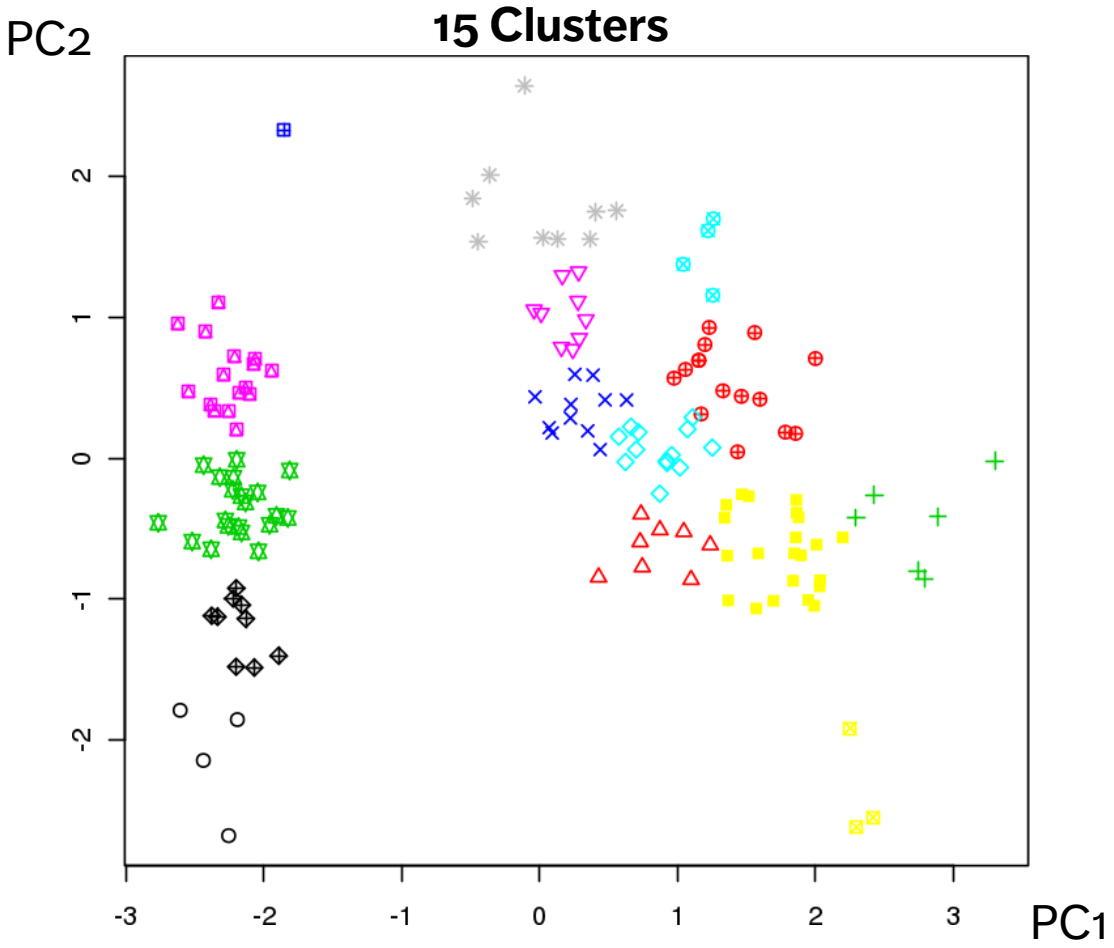
EXAMPLE – IRIS DATASET

EXAMPLE – IRIS DATASET

EXAMPLE – IRIS DATASET



EXAMPLE – IRIS DATASET



A (SMALL?) SAMPLE OF INTERNAL CQMS

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

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

(available in R via `clusterCrit`)

INTERNAL CLUSTERING VALIDATION

Davies-Bouldin Index

$$DB = \frac{1}{N}\sum_{i=1}^{N} \max_{j \neq i} \frac{s_i + s_j}{d(c_i, c_j)},$$

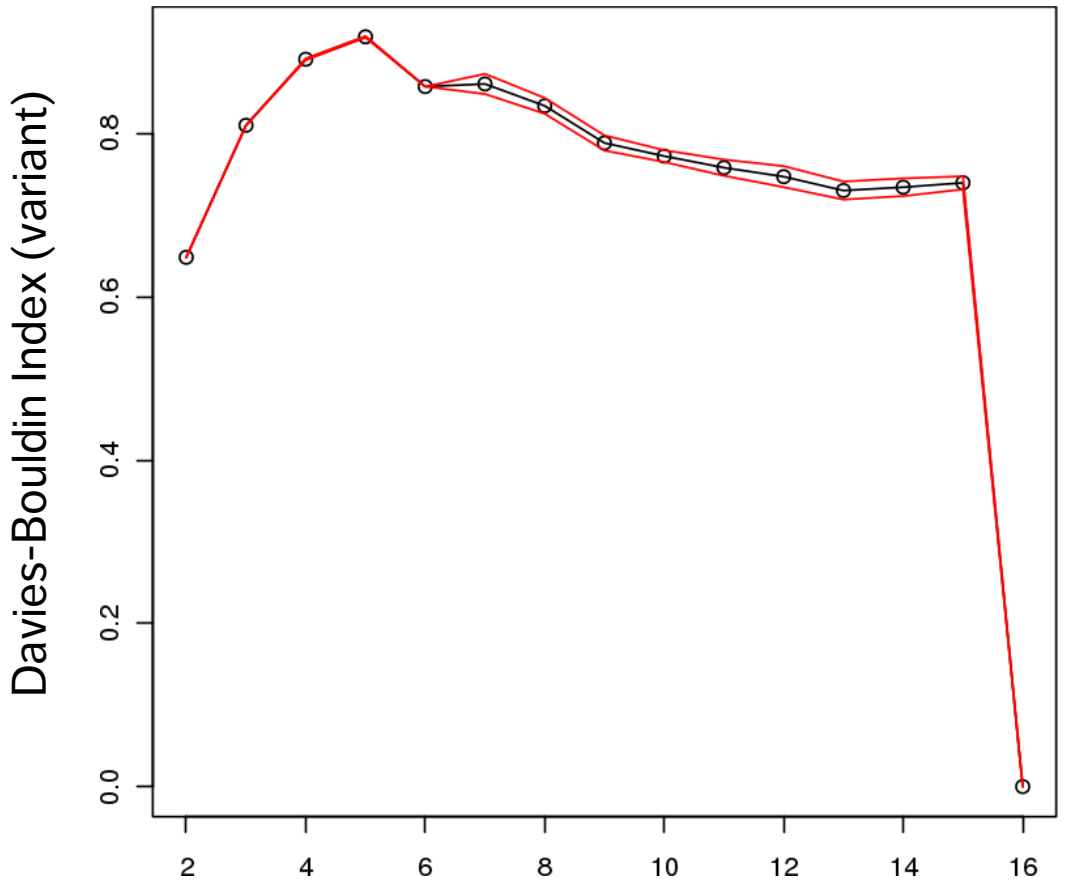
where N is the number of clusters, c_m is the centroid of the m^{th} cluster, and s_m is the average distance of the points in the m^{th} cluster to c_m;

can be used to determine the number of clusters in k-means

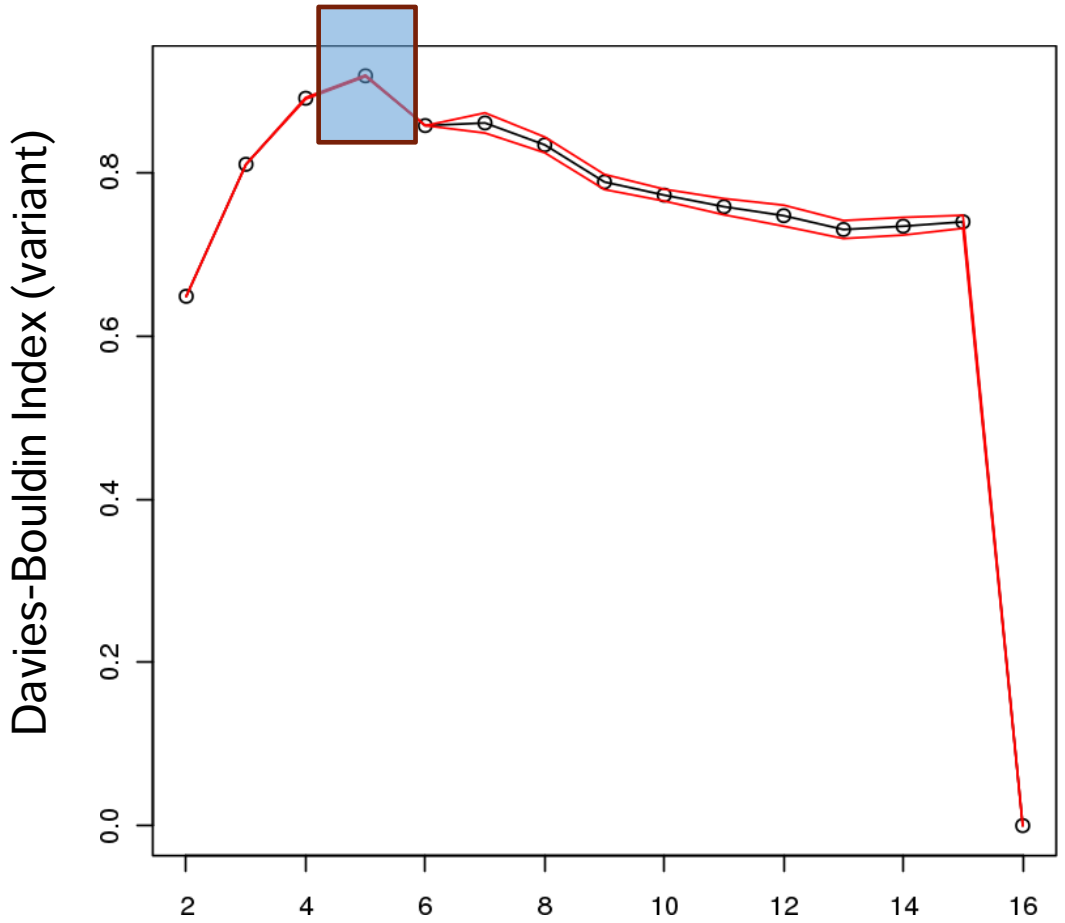
Other Methods

- Sum of Squared Errors, Dunn's Index, Silhouette Metric, etc.

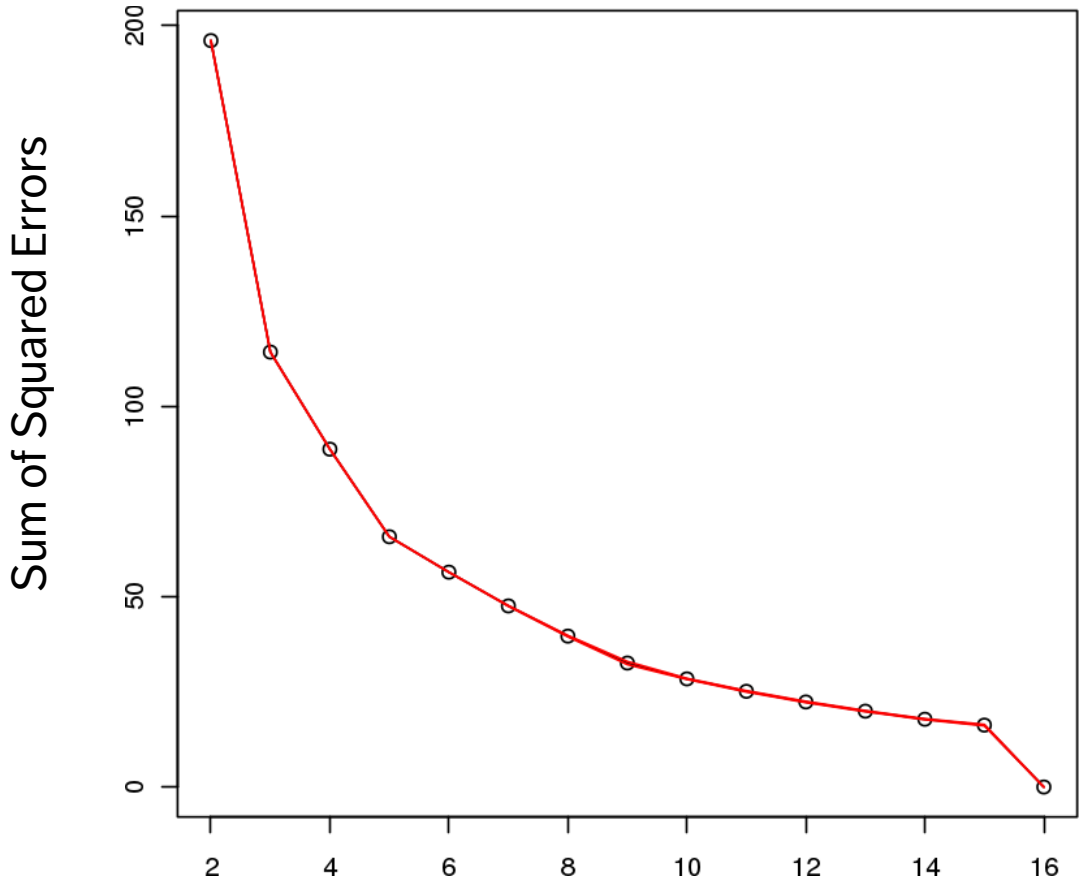
EXAMPLE – IRIS DATASET



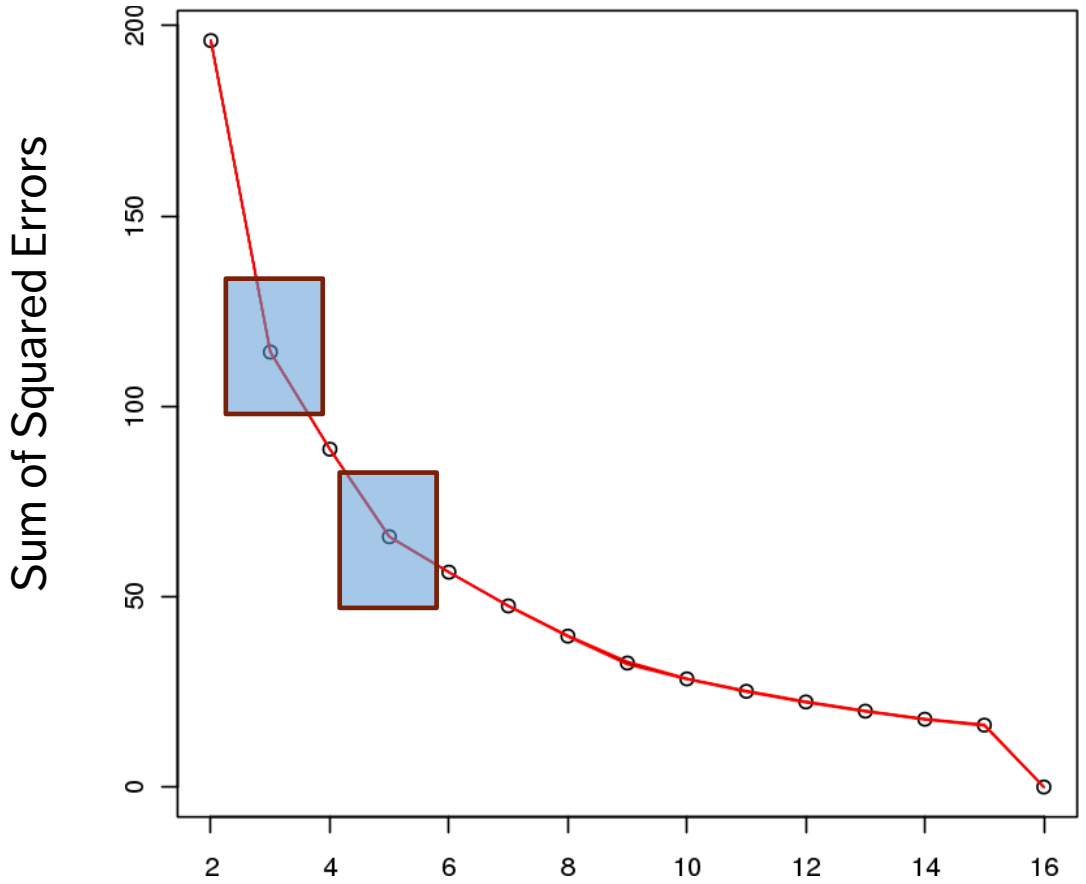
EXAMPLE – IRIS DATASET



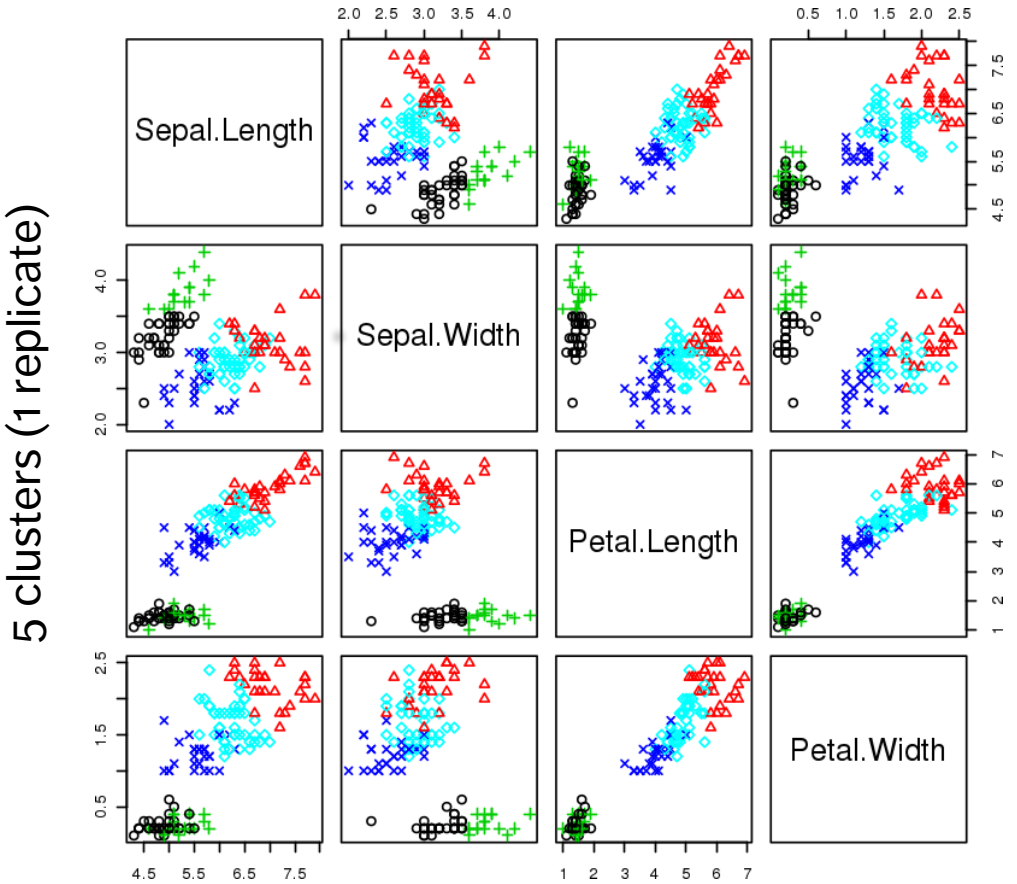
EXAMPLE – IRIS DATASET



EXAMPLE – IRIS DATASET



EXAMPLE – IRIS DATASET



DISCUSSION

Is this a “good” clustering model?

REFERENCES

CLUSTERING

SUPPLEMENTAL MATERIAL

Hierarchical Clustering
https://www.data-action-lab.com/wp-content/uploads/2019/03/Hierarchical-Clustering.pdf

DBSCAN
https://www.data-action-lab.com/wp-content/uploads/2019/03/Density-Based-Clustering.pdf

Spectral Clustering
https://www.data-action-lab.com/wp-content/uploads/2019/03/Spectral-Clustering.pdf

Clustering Notebooks
https://www.data-action-lab.com/wp-content/uploads/2019/03/ClusteringNotebooks.zip

REFERENCES

https://en.wikipedia.org/wiki/Davies–Bouldin_index

https://algobeans.com/2015/11/30/k-means-clustering-laymans-tutorial/

http://www.cs.umd.edu/~samir/498/10Algorithms-08.pdf

Aggarwal, C.C., Reddy, C.K. (eds.) [2014], *Data Custering: Algorithms and Applications*, CRC Press.

Torgo, L. [2017], *Data Mining with R: Learning with Case Studies* (2nd ed.), CRC Press

Aggarwal, C.C. [2015], *Data Mining: the Textbook*, Springer

Maheshwari, A.K. [2015], Business Intelligence and Data Mining, Business Expert Press.

Leskovec, J., Rajamaran, A., Ullman, J.D. [2014], *Mining of Massive Datasets*, Cambridge Press.

REFERENCES

Provost, F., Fawcett, T. [2013], Data Science for Business, O'Reilly.

Hastie, T., Tibshirani, R., and J. Friedman [2008], The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer.

Frank, E., Witten, I.H. [2005], Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Elsevier.

https://en.wikipedia.org/wiki/Cluster_analysis

REFERENCES

hclust {stats}, R Documentation: Hierarchical Clustering, from Package stats version 3.3.0., by R Core Team and contributors worldwide. Retrieved 2016.10.11

Wikipedia: *Hierarchical clustering*. Last edited 2016-09-16. Retrieved 2016.10.11.

Hierarchical agglomerative clustering, in Introduction to Information Retrieval, by C.D. Manning, P. Raghavan and H. Schütze. Published (online version) 2009.04.07

Hierarchical Clustering with R (feat. D3.js and Shiny), by R. Vogler. Published 2014-12-14. Retrieved 2016.10.11.

Introduction to Visualizing Hierarchies, by R. Mazza, D. Brodbeck, M. Lanza and R. Wettel. Retrieved 2016.10.11.

Hierarchical cluster analysis, in Multivariate Analysis of Ecological Data, by M. Greenacre and R. Primicerio. Published 2013, by Fundación BBVA, 2013, Plaza de San Nicolás, 4 48005 Bilbao.

TIBCO Spotfire Documentation: What is a Treemap? Last edited: 2015-02-12. Retrieved 2016.10.11.

REFERENCES

Wikipedia: *Silhouette (clustering)*. Last edited 2016-07-15. Retrieved 2016.10.11.

silhouette {cluster} Compute or Extract Silhouette Information from Clustering, from Package cluster version 2.0.3, by M. Maechler, P. Rousseeuw, A. Struyf and M. Hubert. Published 2016-10-08.

Relative Clustering Validity Criteria: A Comparative Overview by L. Vendramin, R.J.G.B. Campello and E.R. Hruschka. Published online 2010-06-30, by Wiley InterScience.

A General Coefficient of Similarity and Some of Its Properties, by J. C. Gower. Biometrics, 1971.

Dissimilarity Measures, in A Guide to Statistical Analysis in Microbial Ecology: a community-focused, living review of multivariate data analyses. by P.L. Buttigieg and A. Ramette.

Cosine similarity, Pearson correlation, and OLS coefficients, by B. O'Connor.

Data Mining Portfolio Similarity and Dissimilarity Measures, by G. Benoît.

Measures of distance and correlation between variables, in Multivariate Analysis of Ecological Data, by M. Greenacre and R. Primicerio. Published 2013.

REFERENCES

Scientists Trace Society's Myths to Primordial Origins, by J. d'Huy. In Scientific American (Online). Published 2016-09-29. Retrieved 2016-10-11.

Complex building's energy system operation patterns analysis using bag of words representation with hierarchical clustering, by U. Habib, K. Hayat and G. Zucker. Complex Adaptive Systems Modeling, 2016, 4:8.

A Comparison of Antioxidant, Antibacterial, and Anticancer Activity of the Selected Thyme Species by Means of Hierarchical Clustering and Principal Component Analysis, by M. Orłowska, K. Pytlakowskae, A. Mrozek-Wilczkiewicz, R. Musioł, M. Waksmundzka-Hajnos, M. Sajewicz, T. Kowalska.Acta Chromatographica, 2016, 28.

Use of hierarchical cluster analysis to classify prisons in Ireland into mutually exclusive drug-use risk categories, by M. Codd, J. Mehegan, C. Kelleher, A. Drummond.

Divisive Analysis (DIANA) of hierarchical clustering and GPS data for level of service criteria of urban streets, by A.K. Patnai, P.K. Bhuyan, K.V.K. Rao.

REFERENCES

Clusters and DBScan, by Jesse Johnson.Published 2013-08-20. Retrieved 2016.10.11.

Scikit Documentation: Comparing different clustering algorithms on toy datasets. By scikit-learn developers. Retrieved 2016.10.11.

Data Mining TNM033 Notes: DBSCAN A Density-Based Spatial Clustering of Application with Noise, by Henrik Bäcklund, Anders Hedblom and Niklas Neijman. Published 2011-11-30. Retrieved 2016.10.11.

Scatterplot3d: an R package for Visualizing Multivariate Data, by Uwe Ligges and Martin Mächler. Published 2003. Retrieved 2016.10.11.

Wikipedia article for DBSCAN

Schubert, Erich; Sander, Jörg; Ester, Martin; Kriegel, Hans Peter; Xu, Xiaowei (July 2017). "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN", *ACM Trans. Database Syst.* **42** (3): 19:1–19:21. doi:10.1145/3068335, ISSN 0362-5915

REFERENCES

Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease, by C. Plant, S.J. Teipel, A. Oswald, C. Böhm, T. Meindl, J. Mourao-Miranda, A.W. Bokde, H. Hampel, M. Ewers.

A Novel Approach for Predicting the Length of Hospital Stay With DBSCAN and Supervised Classification Algorithms, by Panchami V.U., N. Radhika, A.V. Vidyapeetham.

Simulation of DNA damage clustering after proton irradiation using an adapted DBSCAN algorithm, by Z. Francis, C. Villagrasa, I. Clairand.

Where traffic meets DNA: mobility mining using biological sequence analysis revisited, by A. Jawad, K. Kersting, N.V. Andrienko.

Individual movements and geographical data mining. clustering algorithms for highlighting hotspots in personal navigation routes, by G. Schoier, G. Borruso.

REFERENCES

Ulrike von Luxburg, *A Tutorial on Spectral Clustering*, Max Planck Institute for Biological Cybernetics

Aarti Singh, *Spectral Clustering*.

Andrew Y. Ng, Michael I. Jordan, Yair Weiss, *On Spectral Clustering: Analysis and an Algorithm.*

Jing Wang, *An Introduction to Support Vector Machine and Spectral Clustering.*

Lihi Zelnik-Manor, Pietro Perona, *Self-Tuning Spectral Clustering.*

Denis Hamad, Philippe Biela, *Introduction to spectral clustering.*

Marina Meila, *Classic and Modern Data Clustering*

H.T. Kung, Dario Vlah, *A Spectral Clustering Approach to Validating Sensors via Their Peers in Distributed Sensor Networks*

Morteza Chehreghani, Alberto Busetto, Joachim Buhmann, *Information Theoretic Model Validation for Spectral Clustering*

Sepandar D. Kamvar, Dan Klein, Christopher Manning, *Spectral Clustering*

REFERENCES

Vendramin, L. & J. G. B. Campello, R. & Hruschka, E. (2010). Relative clustering validity criteria: A comparative overview. Statistical Analysis and Data Mining. 3. 209-235. 10.1002/sam.10080.

Amigó, E. & Gonzalo, J. & Artiles, J. & Verdejo, M. (2009). Comparison of extrinsic clustering evaluation metrics based on formal constraints. Information Retrieval. 12. 461-486.

M Lewis, J. & Ackerman, M. & de Sa, V. (2012). Human Cluster Evaluation and Formal Quality Measures: A Comparative Study. Proc. 34th Conf. of the Cognitive Science Society.

Bernard Desgraupes (2013). Clustering Indices. Lab Modal'X, University Paris Ouest.

Justin Cranshaw, Raz Schwartz, Jason I. Hong, and Norman Sadeh, "The Livehoods Project: Utilizing Social Media to Understand the Dynamics of A City," *Proceedings of the the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM–12),* Dublin, Ireland, pp. 1–8, 2012.

Kung H. T., Vlah D.A, "Spectral clustering approach to validating sensors via their peers in distributed sensor networks", *Proceedings of the 18th IEEE International Conference on Computer Communications and Networks (ICCCN '09),* 2009.

SUPPLEMENTAL MATERIAL

CLUSTERING

COMPARING MEASURES ACROSS DATASETS

		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: $k_{\max} = 25$.

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.

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?

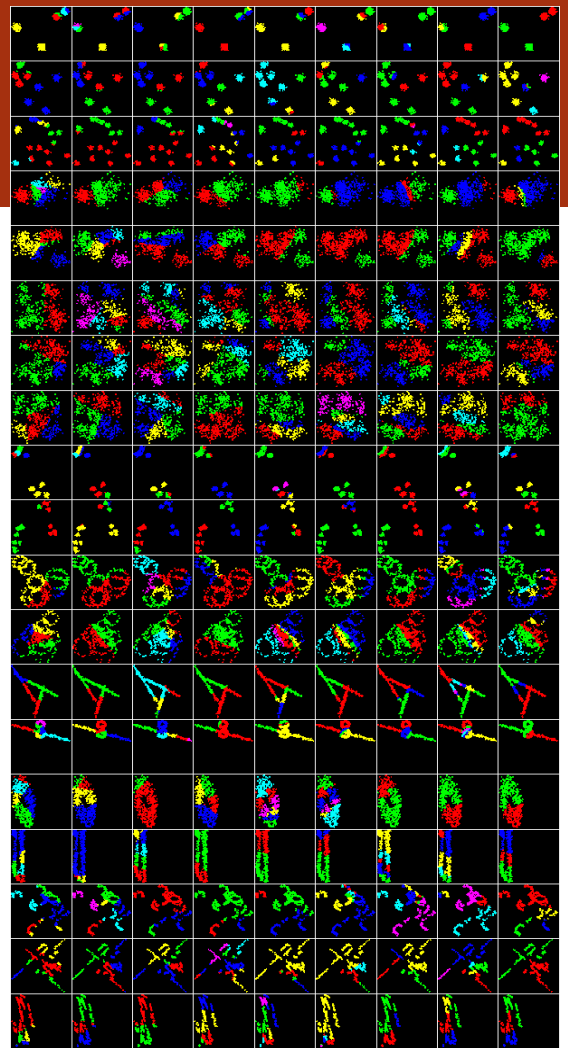


Figure 1: All stimuli. Datasets are in rows; partitions are in columns.

CORRELATION MEASURES

	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).

Look at correlation between clustering assignments

Rand, Jaccard, Gamma

Perfect correlation gives maximum value of the measure