CLUSTERING

SETTING THE STAGE

"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)

CONTENTS

[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:

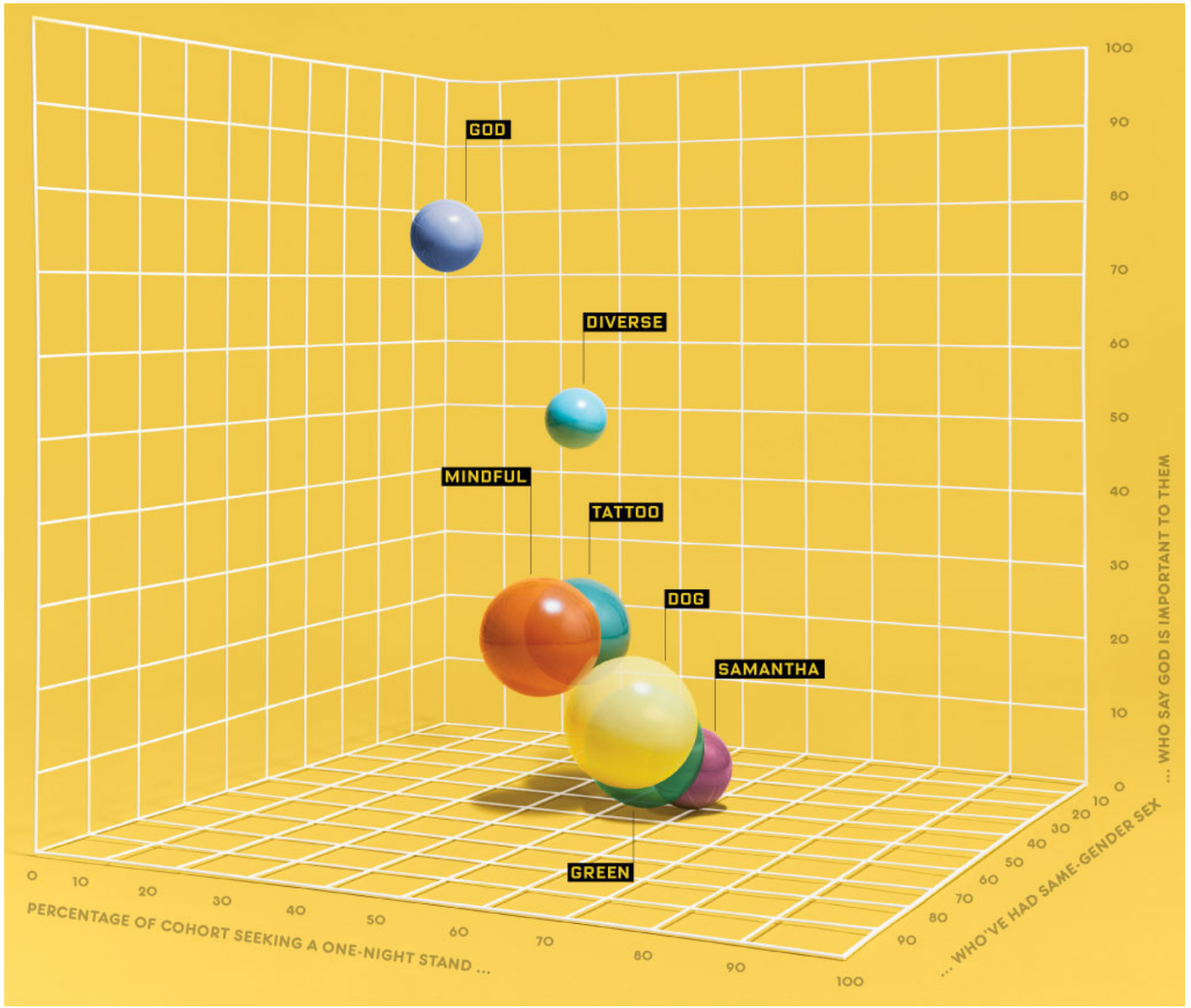
1. **Collect** data
2. Collect **more** and **slightly better** and **different** data
3. Collect **still more** data
4. Figure out a data mining technique that would be **relevant** to what he wanted to know (clustering)
5. **Validate** the results of the analysis

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?

How do you feel about this use of machine learning?

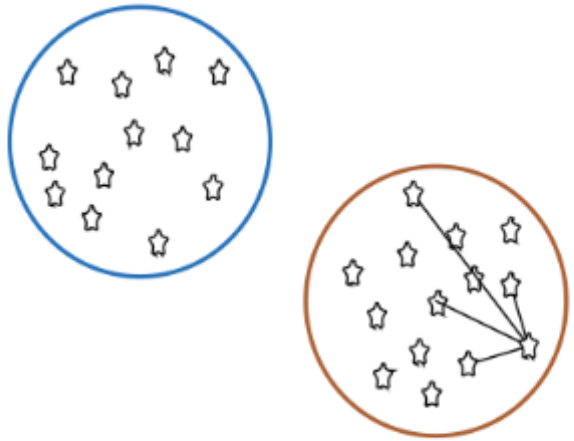

GOD
DIVERSE
MINDFUL
TATTOO
DOG
SAMANTHA
GREEN
0 10 20 30 40 50 60 70 80 90 100
PERCENTAGE OF COHORT SEEKING A ONE-NIGHT STAND ...
0 10 20 30 40 50 60 70 80 90 100
... WHO'VE HAD SAME-GENDER SEX
0 10 20 30 40 50 60 70 80 90 100
... WHO SAY GOD IS IMPORTANT TO THEM

CLUSTERING OVERVIEW

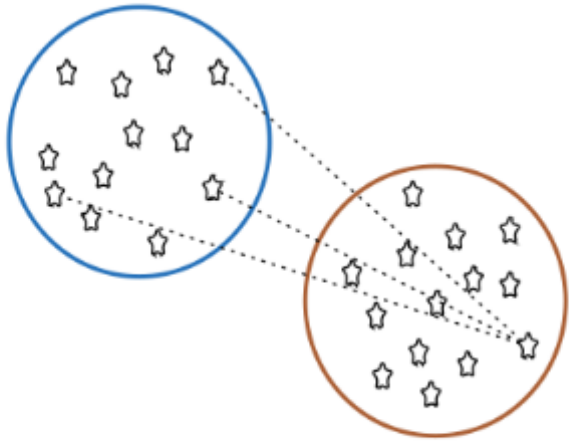
In **clustering**, the data is divided into **naturally occurring groups**. Within each group, the data points are **similar**; from group to group, they are **dissimilar**.

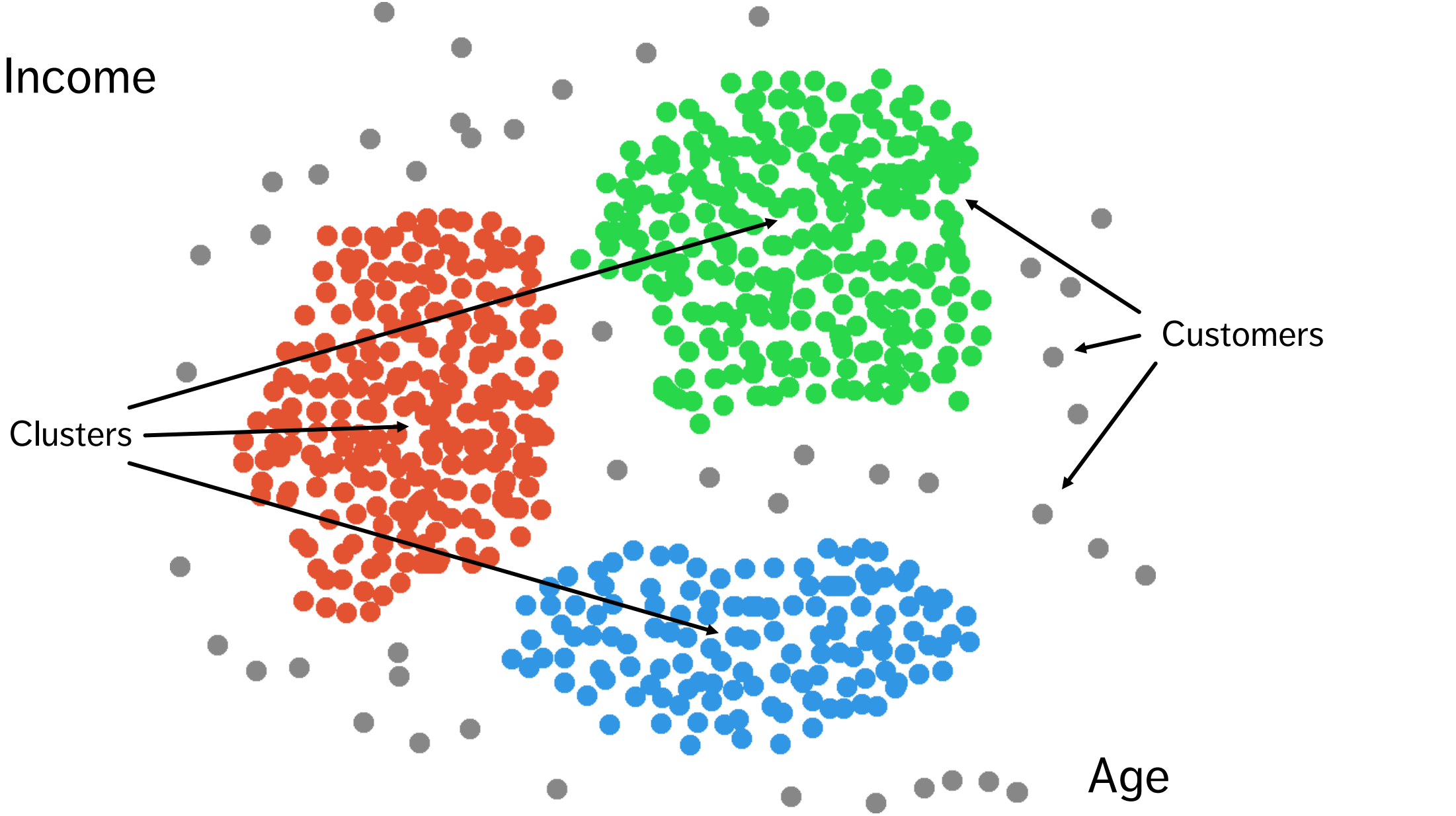
The grouping labels are not determined ahead of time, so clustering is an example of **unsupervised** learning.

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**

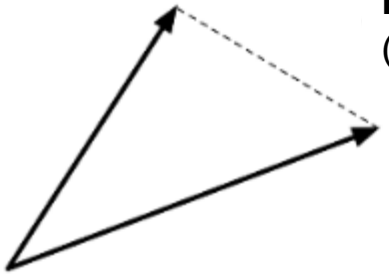
They are also (typically) **non-deterministic**:

- 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

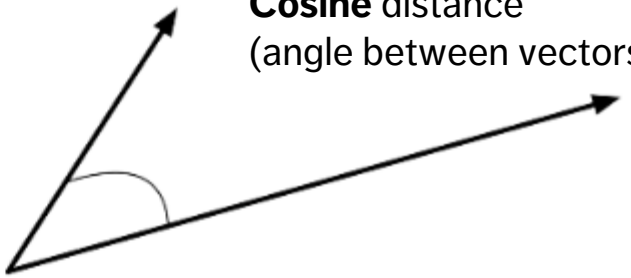
IMPORTANT: data must be scaled before it is fed into clustering algorithms.

CLUSTERING REQUIREMENT

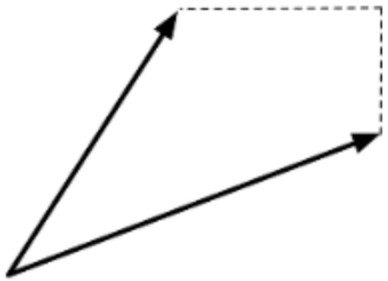
A measure of **similarity** w (or a distance d) between observations.



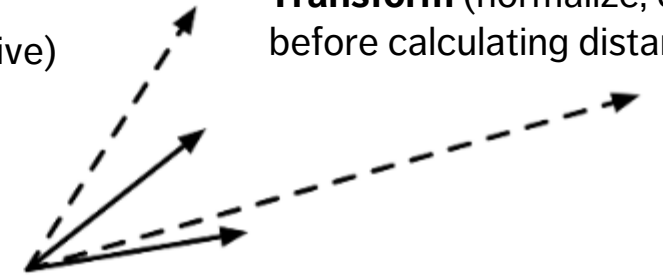
Euclidean distance
(as the crow flies)



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 $d \to \infty$.

APPLICATIONS

Text Documents

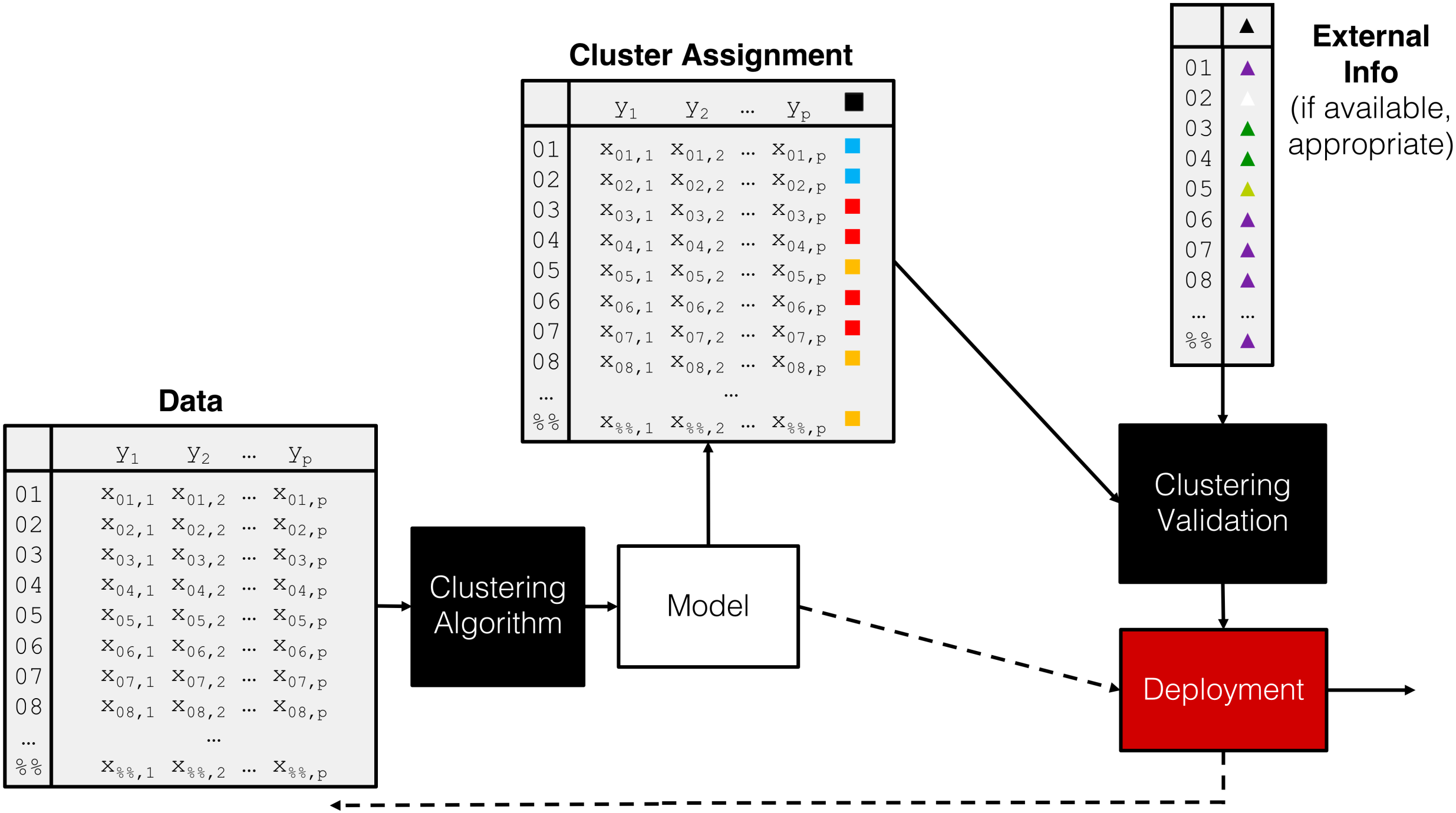
- grouping similar documents according to their topics, based on the patterns of common and unusual words

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


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

CLUSTERING SCHEMES

k-Means

Hierarchical Clustering

Latent Dirichlet Allocation

Expectation-Maximization

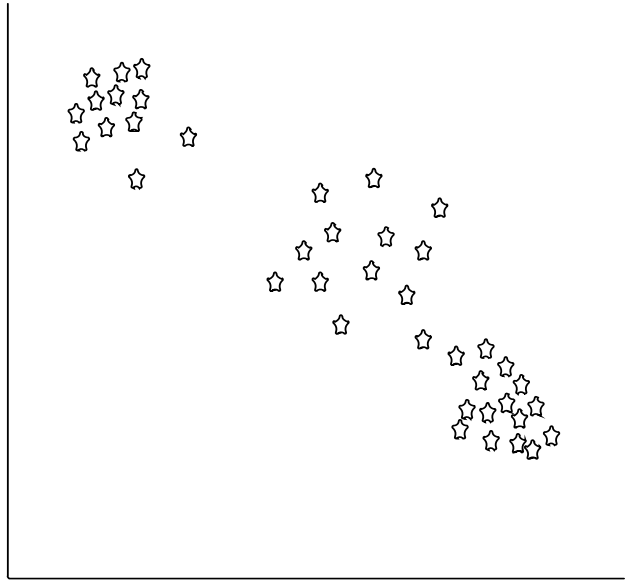
Balanced Iterative Reducing and Clustering using Hierarchies

Density-Based Spatial Clustering of Applications with Noise

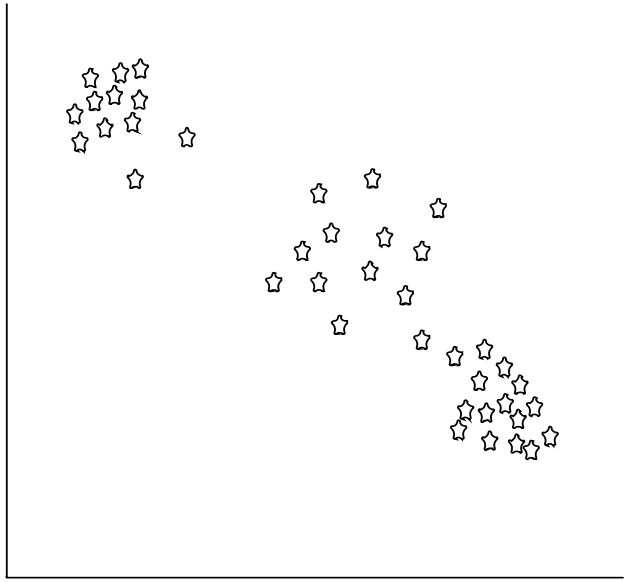
Affinity Propagation

Spectral Clustering

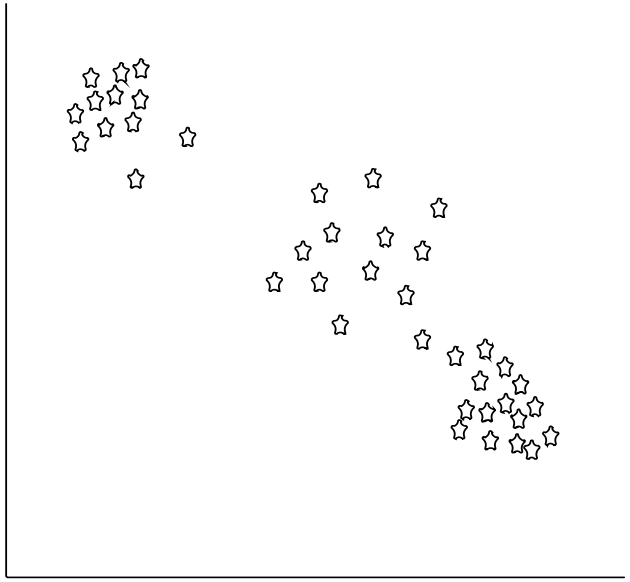
THE GENERAL FORM OF A CLUSTERING ALGORITHM



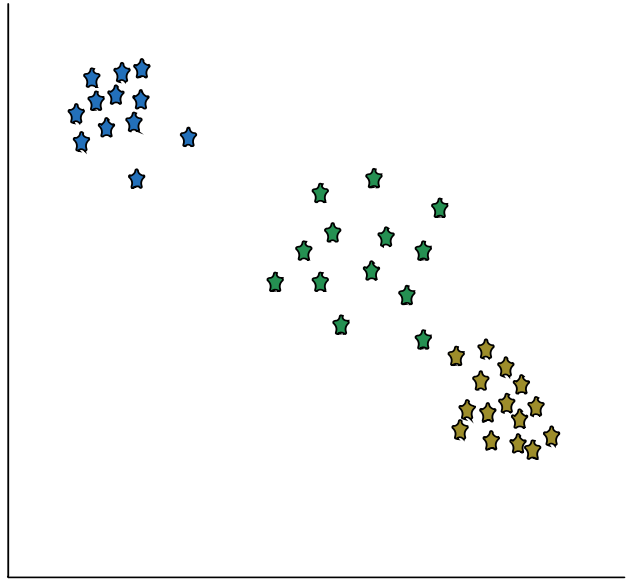
Initialization



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

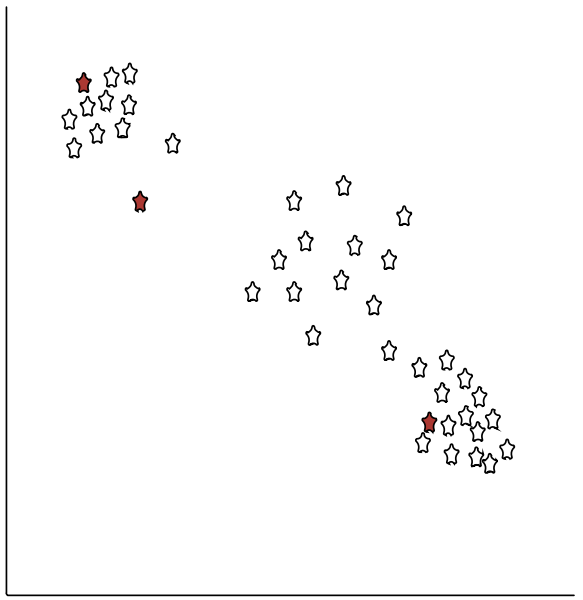


Clustering Step B (Usually Repeated, Possibly in Conjunction with Previous Step)

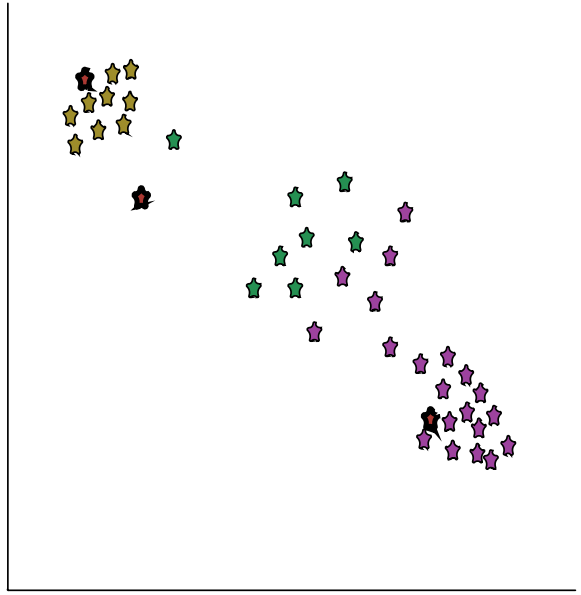


End Condition (Usually When Iterations of Steps A and B Produce Stable Results)

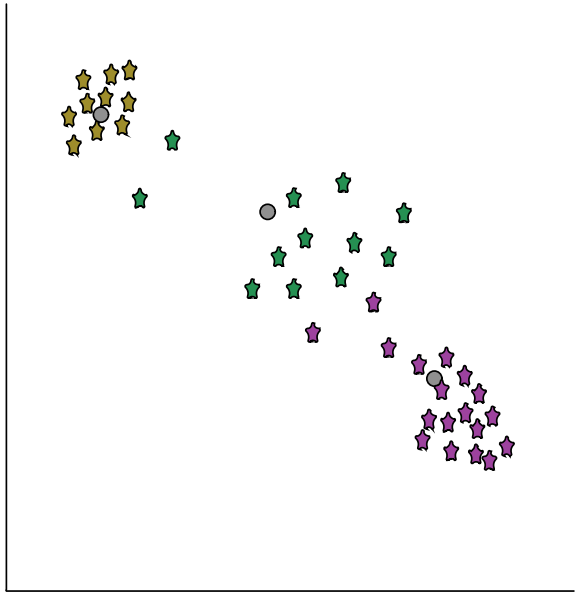
K-MEANS ALGORITHM GLOSS



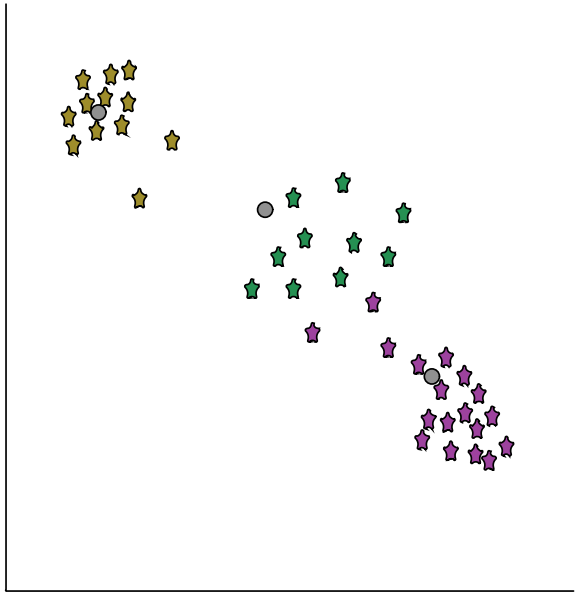
Initialization (e.g. Randomly Pick k Centers)



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 AND LIMITATIONS

Easy to implement

Often a **natural** way to group observations.

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

Points can only be assigned to **one** cluster

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

Clusters are assumed to be separate

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

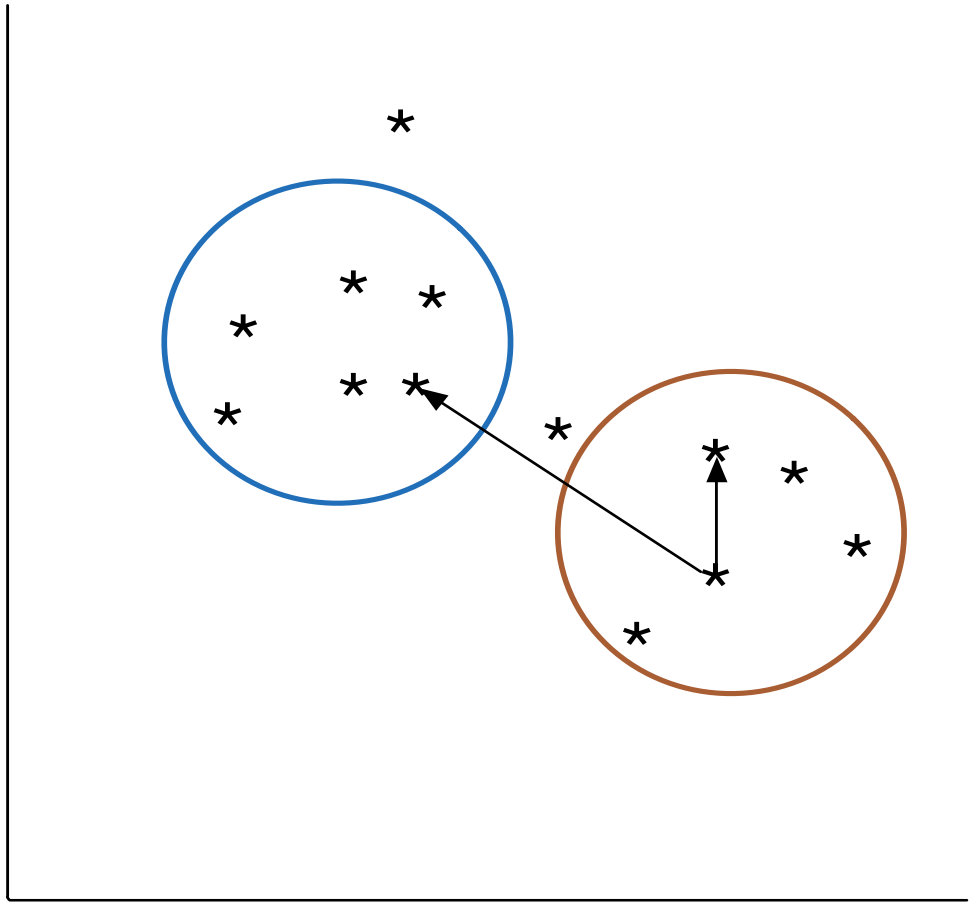
Are there right or wrong groupings of this dataset?

Are there multiple possible ‘natural’ clusterings?

Could different clusterings be used differently?

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

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

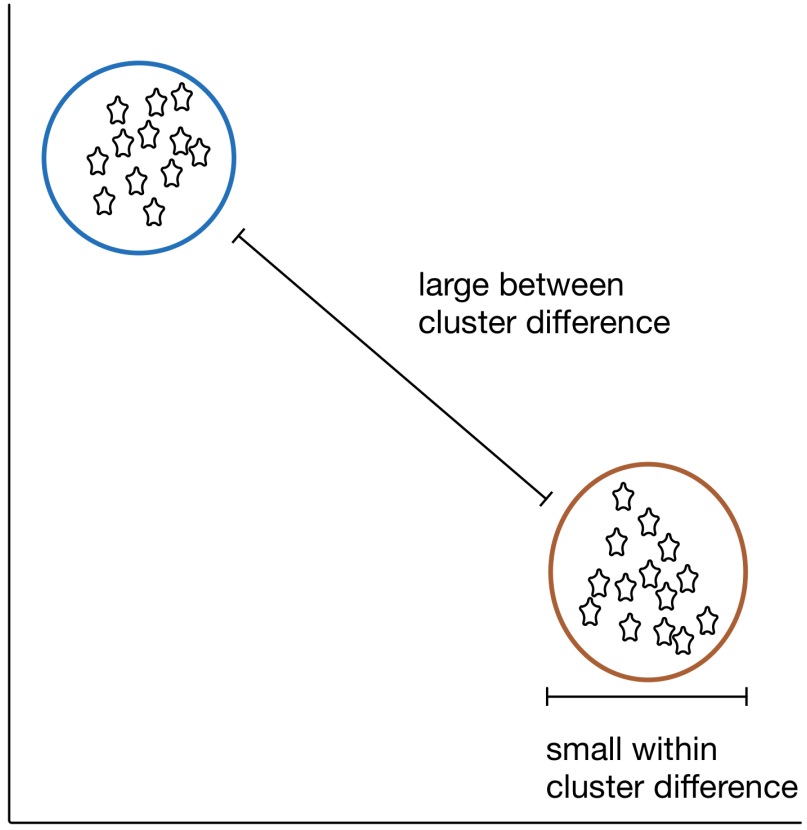
VERY BROAD GOALS

Within clusters, everything is very similar. Between clusters, there is a lot of difference.

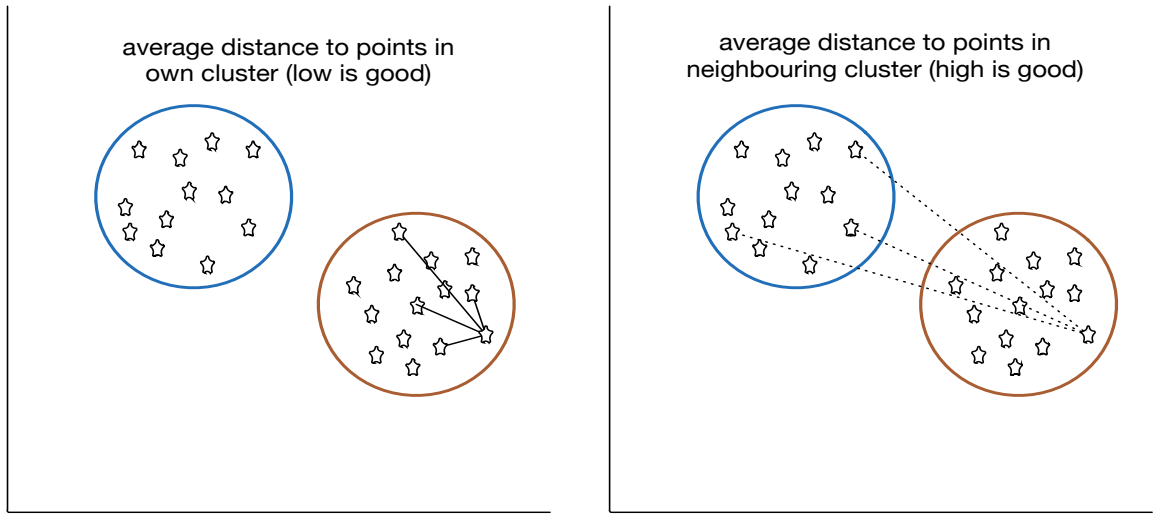
The problem: there are many ways for clusters to deviate from this ideal.

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).

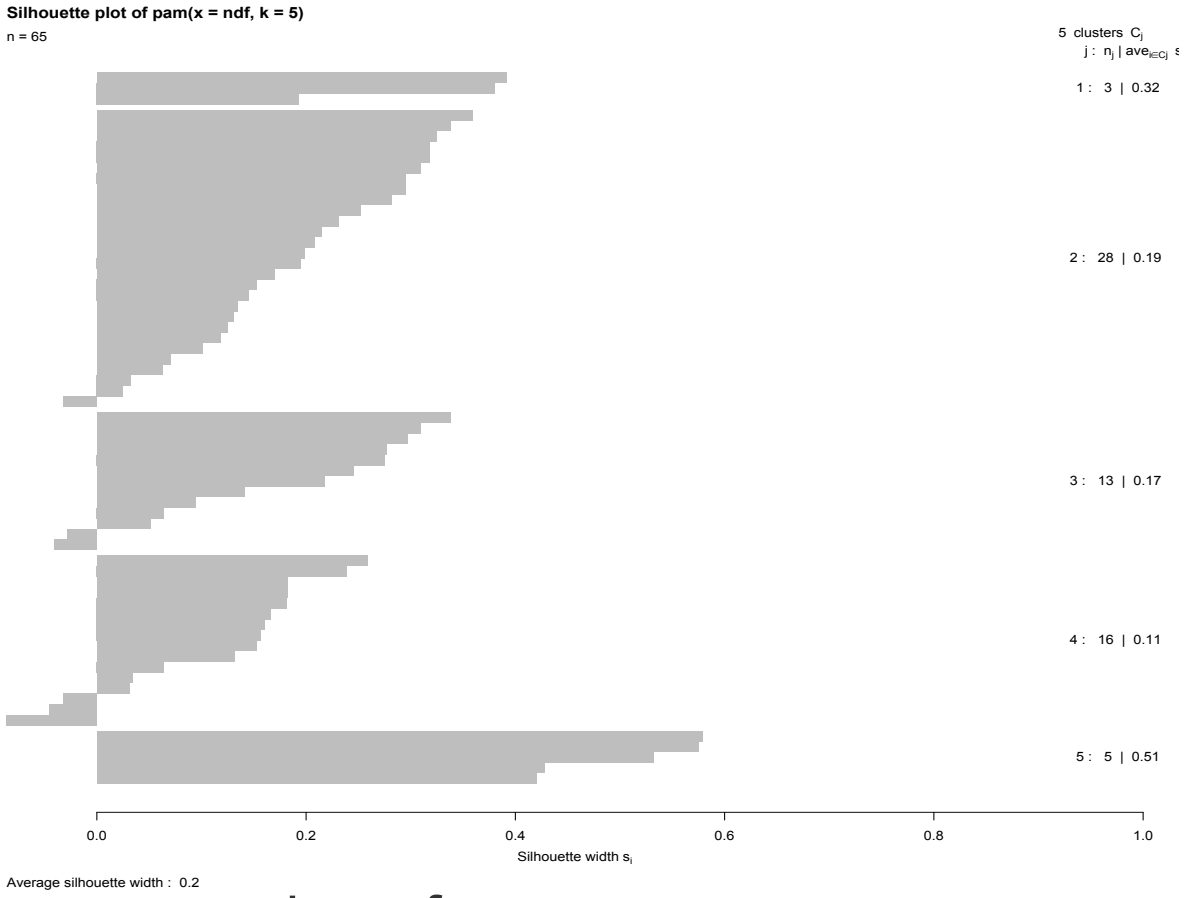
Thus the large number of CQMs.



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.

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`)

CLUSTERING CHALLENGES

Automation

Lack of a clear-cut definition

Lack of repeatability

Number of clusters

Cluster description

Model validation

Ghost clustering

***A posteriori* rationalization**