CPSC 340: Machine Learning and Data Mining

Non-Parametric Clustering

Summer 2021

Admin

- Assignment 2 due Monday
- Assignment 3 out Friday
- Midterm on Tuesday, June 1, 2021
  - Auto-graded portion and manually graded portion
  - Do each portion in single seating
    - E.g.
      do auto-graded potion at 10am
      do manually-graded portion at 7pm
- Practice Midterm will be out next Tuesday.
  - (Loosely) based on previous term's exams
- No class Monday (Victoria Day)

In This Lecture

- K-Means Discussion (10 minutes)
- Density-Based Clustering (15 minutes)
- Hierarchical Clustering (15 minutes)

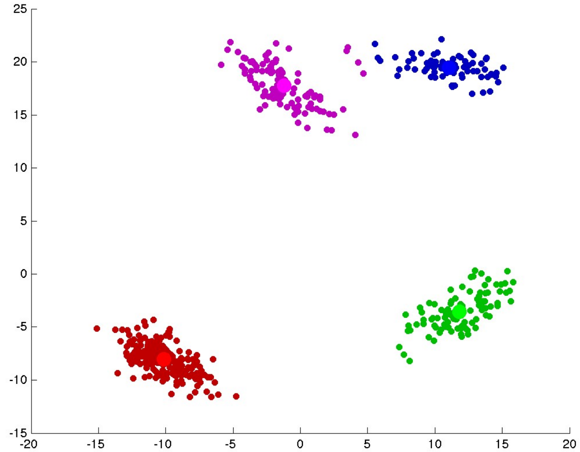
Last Time: K-Means Clustering

- We want to cluster data:
  – Assign examples to groups.

- K-means clustering:
  – Define groups by “means”
  – Assigns examples to nearest mean.
  (And updates means during training.)

- Also used for vector quantization:
  – Learned compression method.
  – “Replace all breakfast with canonical breakfast”

- Issues with k-means:
  – Fast but sensitive to initialization.
  – Choosing ‘k’ is annoying.



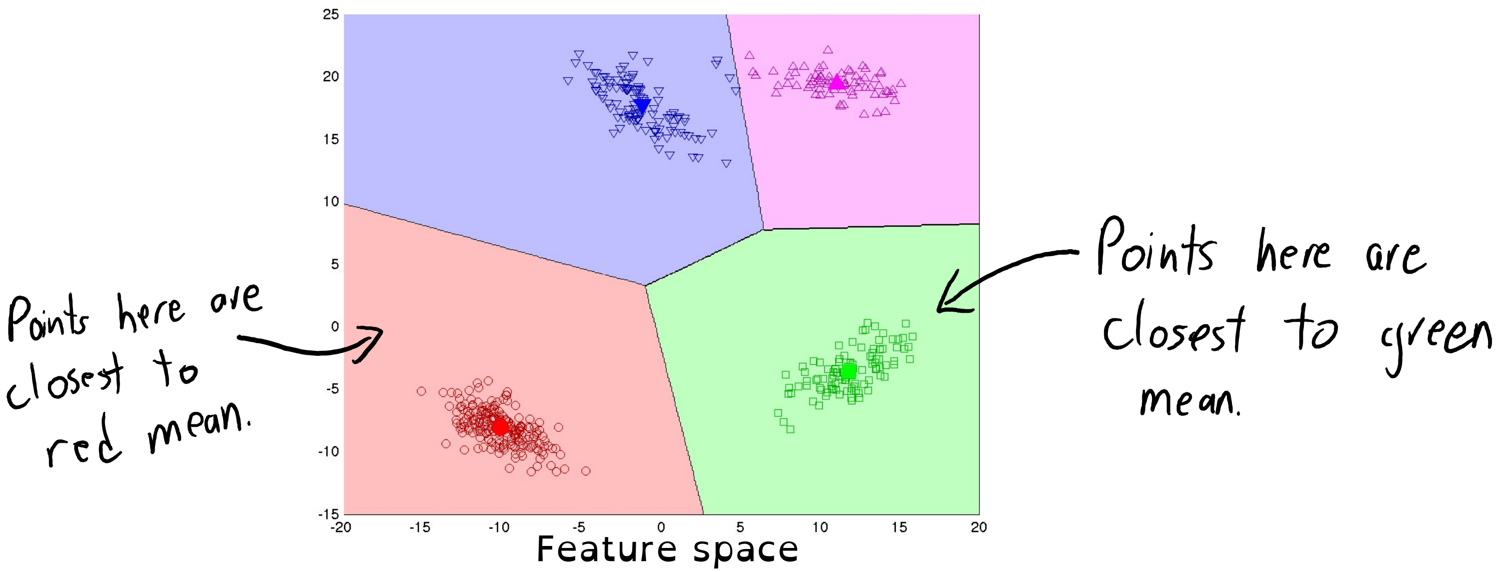
Feature space

Coming Up Next

K-MEANS AND CONVEX SHAPES

Shape of K-Means Clusters

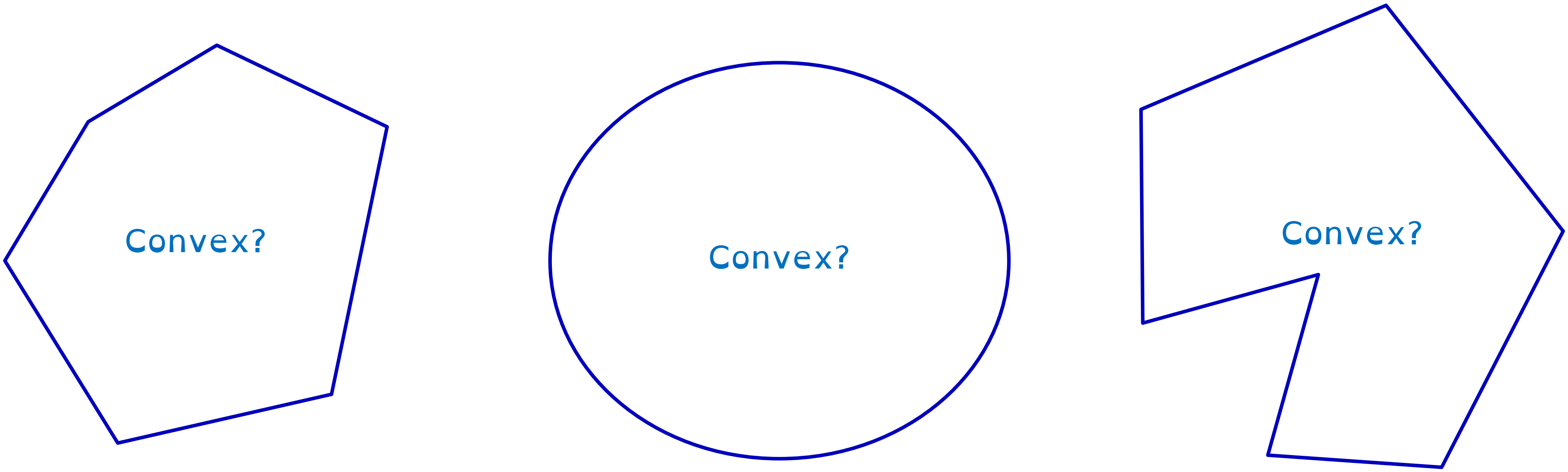
- Recall: k-means assigns cluster labels based on distance to nearest mean.
- Equivalent: partition the feature space based on the "closest mean":



- Observe that the clusters are convex regions (proof in bonus).

Convex Sets

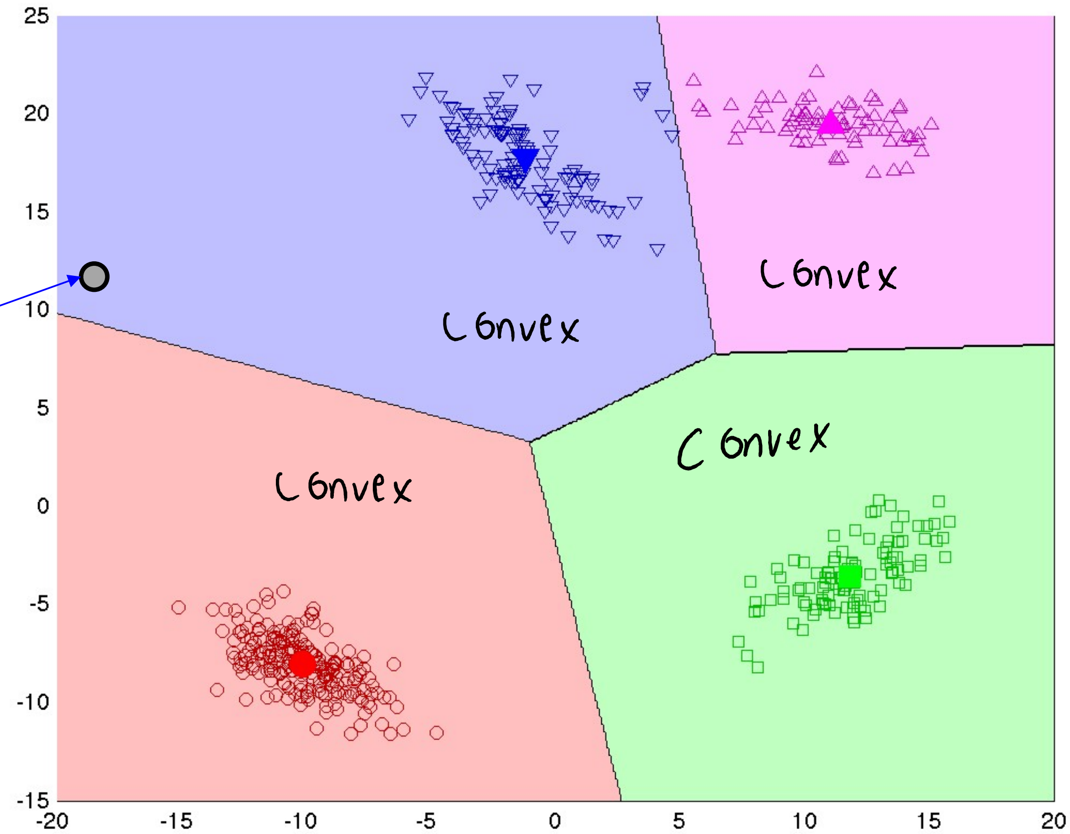
- A set is convex if line between two points in the set ____________________.



Shape of K-Means Clusters

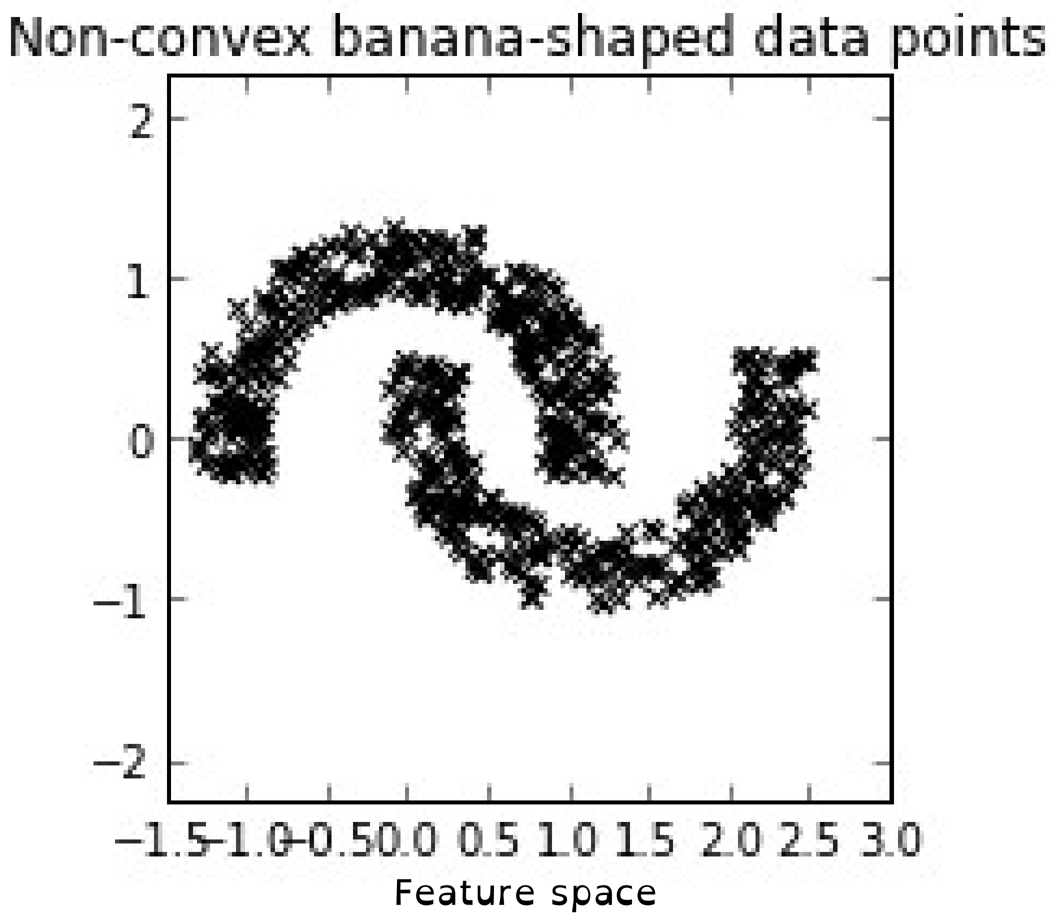
Q: Are clusters always convex?

Q: Should this really be
in the blue cluster?



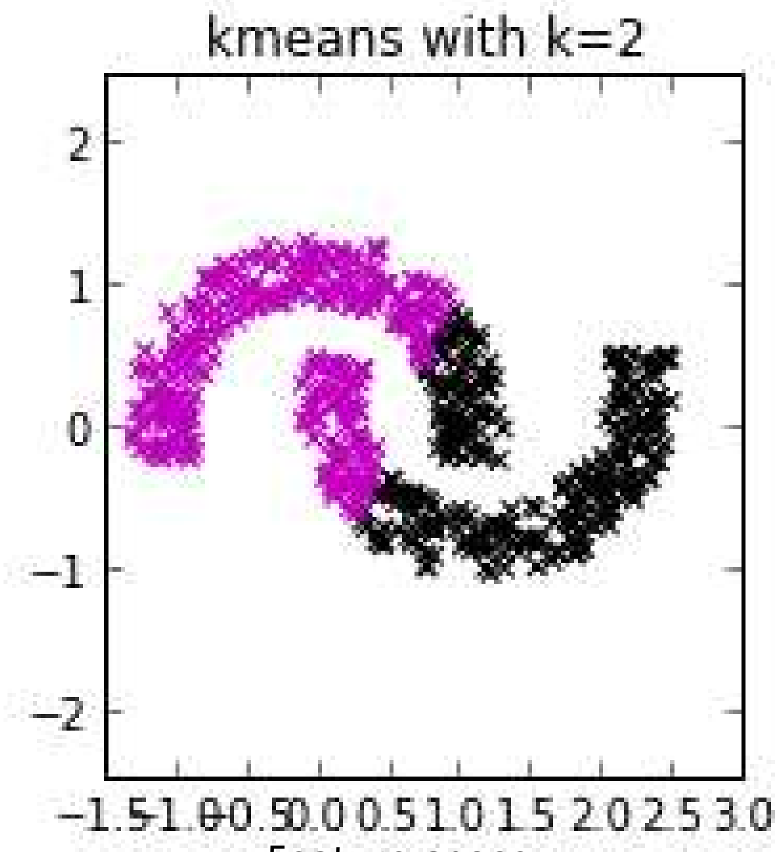
Animation

K-Means with Non-Convex Clusters



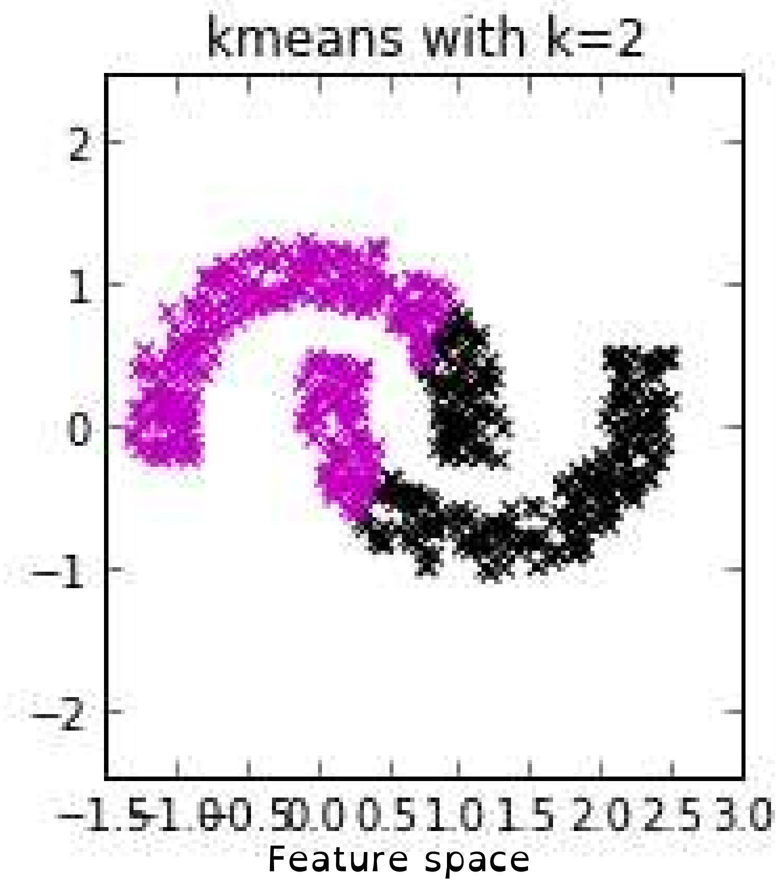
https://corelifesciences.com/human-long-non-coding-rna-expression-microarray-service.html

K-Means with Non-Convex Clusters



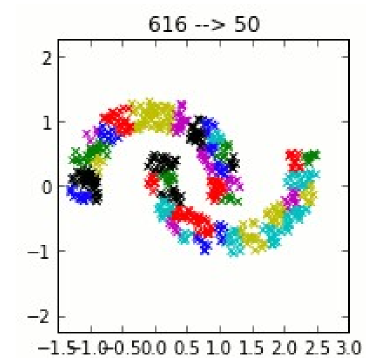
K-means cannot separate some non-convex clusters

K-Means with Non-Convex Clusters

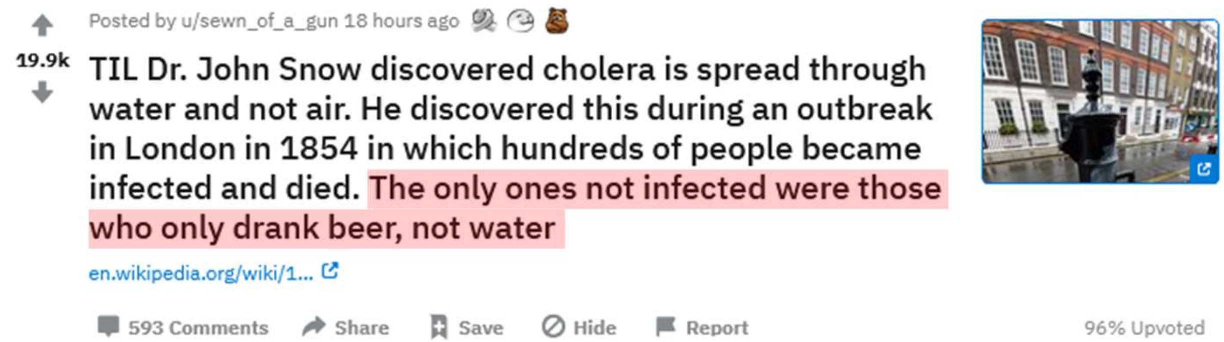


K-means cannot separate some non-convex clusters

Though over-clustering can help ("hierarchical")

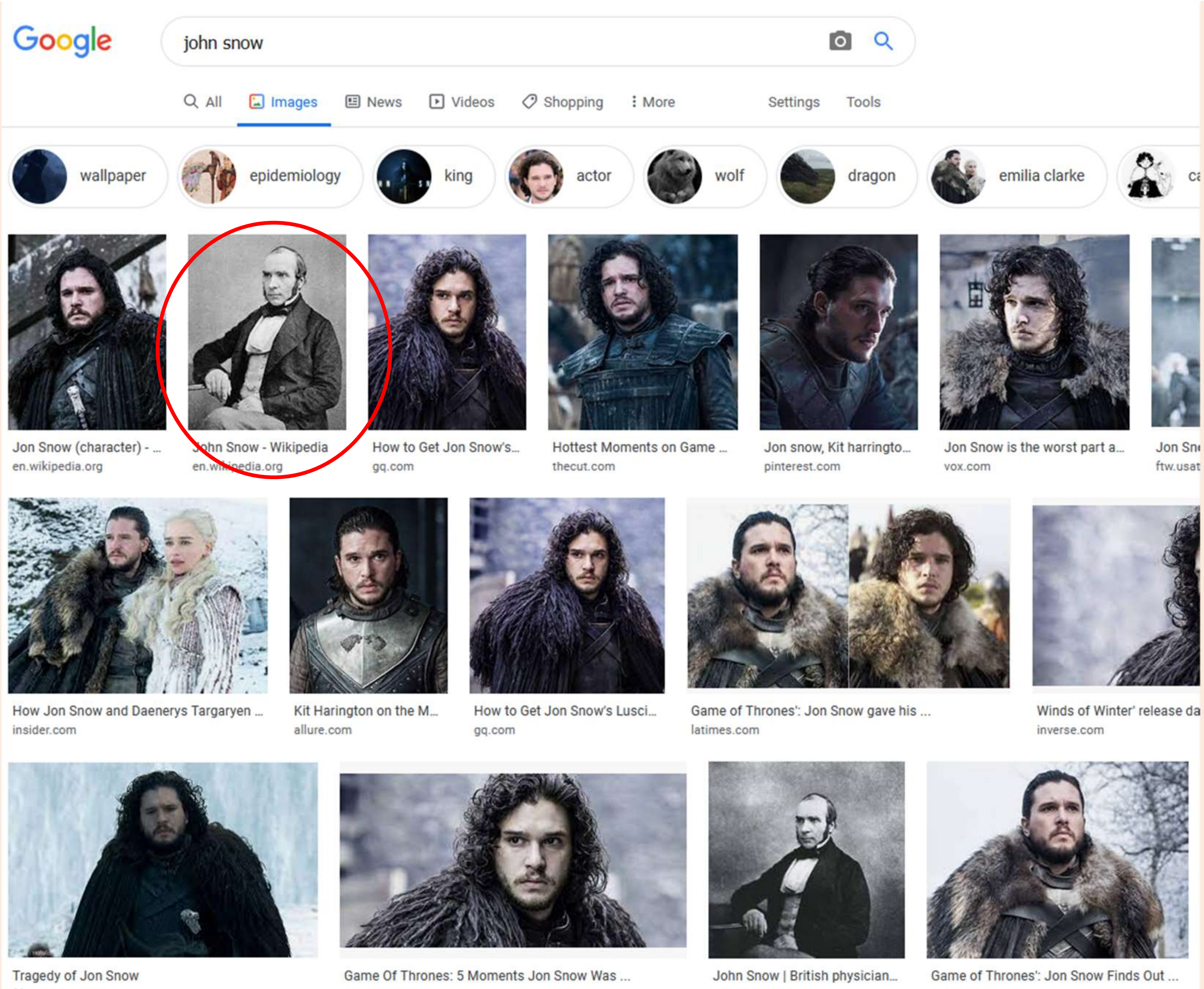


https://corelifesciences.com/human-long-non-coding-rna-expression-microarray-service.html



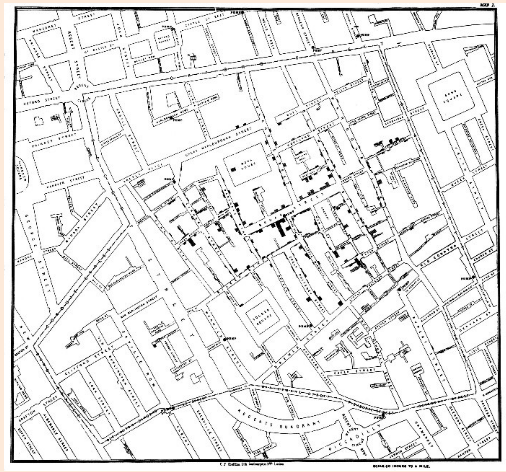
Coming Up Next

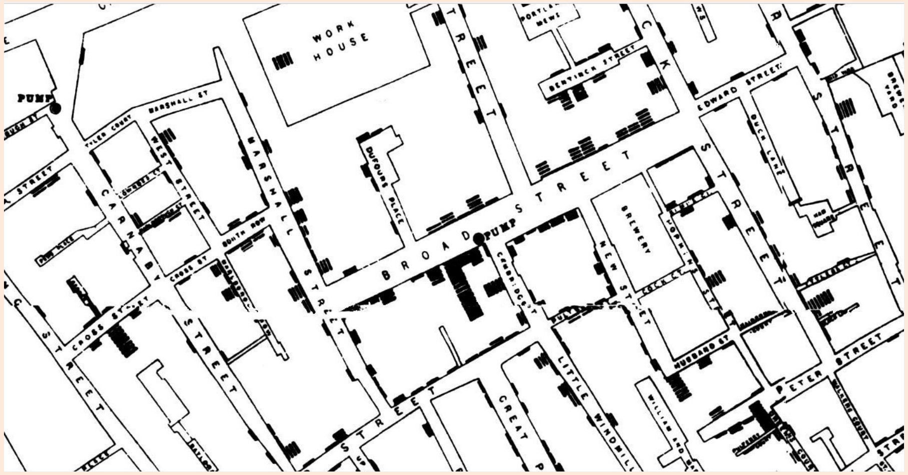
DENSITY-BASED CLUSTERING


john snow
All
Images
News
Videos
Shopping
More
Settings
Tools
wallpaper
epidemiology
king
actor
wolf
dragon
emilia clarke
Jon Snow (character) - ...
en.wikipedia.org
John Snow - Wikipedia
en.wikipedia.org
How to Get Jon Snow's...
gq.com
Hottest Moments on Game ...
thecut.com
Jon snow, Kit harringto...
pinterest.com
Jon Snow is the worst part a...
vox.com
How Jon Snow and Daenerys Targaryen ...
insider.com
Kit Harington on the M...
allure.com
How to Get Jon Snow's Lusci...
gq.com
Game of Thrones': Jon Snow gave his ...
latimes.com
inverse.com
Tragedy of Jon Snow
Game Of Thrones: 5 Moments Jon Snow Was ...
John Snow | British physician...
Game of Thrones': Jon Snow Finds Out ...

John Snow and Cholera Epidemic

- John Snow's 1854 spatial histogram of deaths from cholera:

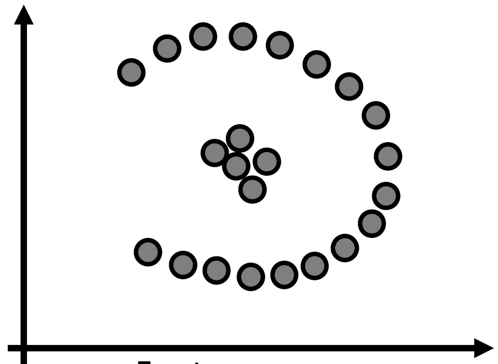




- Found cluster of cholera deaths around a particular water pump.
  - Went against airborne theory, but pump later found to be contaminated.
  - "Father" of epidemiology.

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

Why Would Clusters Be Non-Convex?



- Interaction between features can make regions non-convex
- Incomplete features can make regions non-convex
- Process of generating data can make regions non-convex

- Assumption: well-defined clusters have high ________________.

Why Density-Based Clustering?

- Density-based clustering:
  - Clusters are defined by "dense" regions.
  - Examples in non-dense regions don't get clustered.
    - Not trying to "partition" the space.

- Clusters can be non-convex:
  - Elephant clusters affected by vegetation, mountains, rivers, water access, etc.

- It's a non-parametric clustering method:
  - No fixed number of clusters 'k'.
  - Clusters can become more complicated with more data.

Density and Social Distancing

- Clusters are defined by density
- High density leads to chain reaction
- E.g. COVID-19
  - Social distancing => avoid clusters

Large crowd gathers at Vancouver’s Sunset Beach on 4-20 despite COVID-19 restrictions

By Jon Azpiri • Global News
Posted April 21, 2021 8:39 am · Updated April 21, 2021 11:53 am

Despite the COVID-19 restrictions, the haze was still hanging around Sunset Beach for people celebrating 4-20 Tuesday night. Andrea MacPherson has more on the unofficial event and the who showed up – Apr 21, 2021

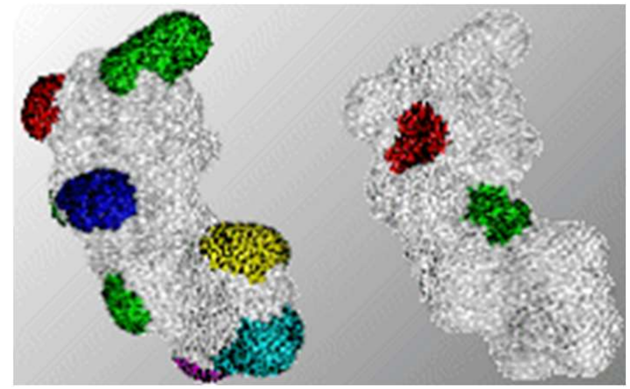
COVID-19 cluster reported at Port Coquitlam Costco; outbreak at VGH declared

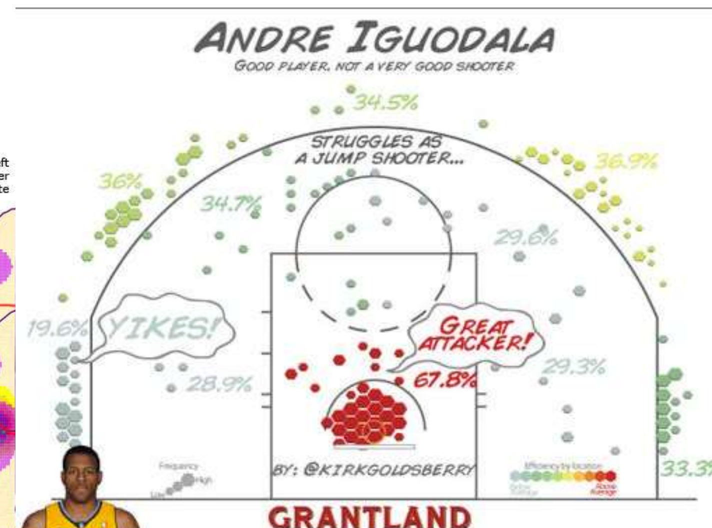
BY DEAN RECKSIEDLER AND HANA MAE NASSAR
Posted Mar 15, 2021 9:38 am PDT Last Updated Mar 16, 2021 at 1:09 am PDT

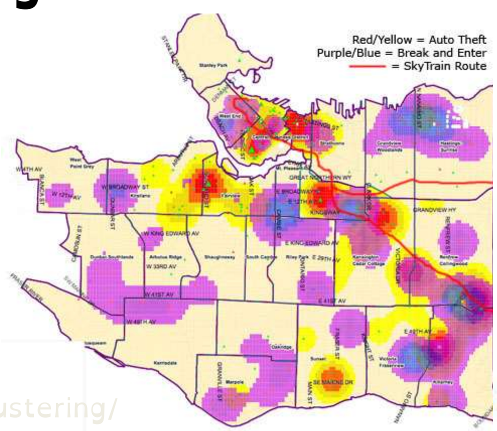
High-density infection event is called CLUSTER!

Other Potential Applications

- Where are high crime regions of a city?
- Where should taxis patrol?
- Where does player make/miss shots?
- Which products are similar to this one?
- Which pictures are in the same place?
- Where can proteins ‘dock’?
- Where are people tweeting?



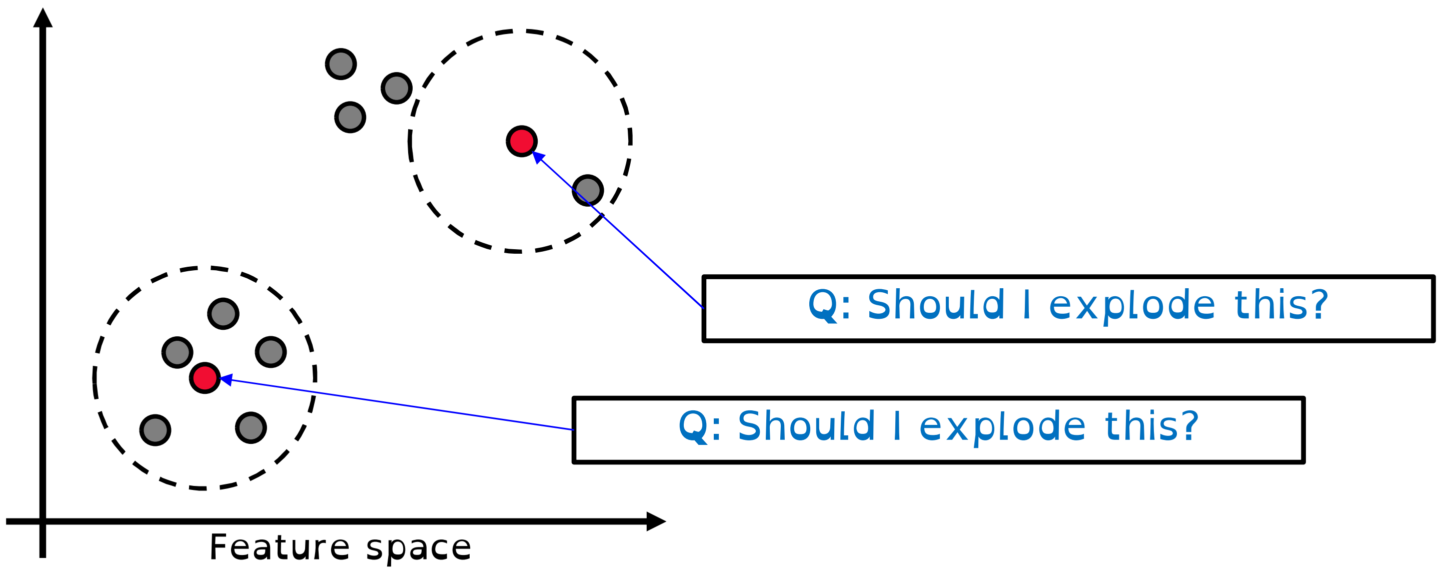




https://en.wikipedia.org/wiki/Cluster_analysis
https://www.flickr.com/photos/dbarefoot/420194128/
http://letsgowarriors.com/replacing-jarrett-jack/2013/10/04/
http://www.dbs.informatik.uni-muenchen.de/Forschung/KDD/Clustering/

What is Density-Based Clustering?

- (I'd call it "chain reaction clustering")
- Enough data points in neighbourhood → "explode"
  - Then see if the neighbours should explode
- Not enough data points in neighbourhood → don't "explode"
- (Dense/sparse) regions → lots of explosion, (dense/sparse) regions → not much explosion
- Data points "burning" together form a cluster (like COVID-19 infections!)

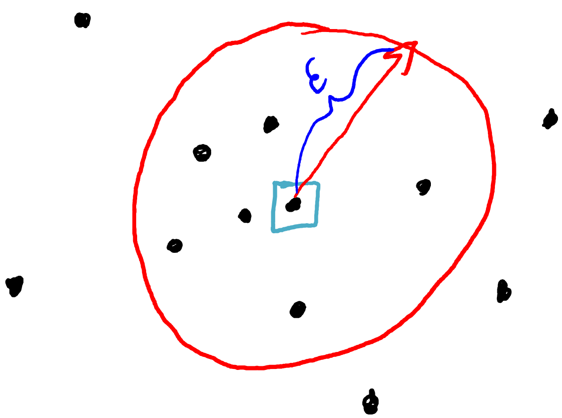


Coming Up Next

MORE FORMAL DESCRIPTION OF DENSITY-BASED CLUSTERING

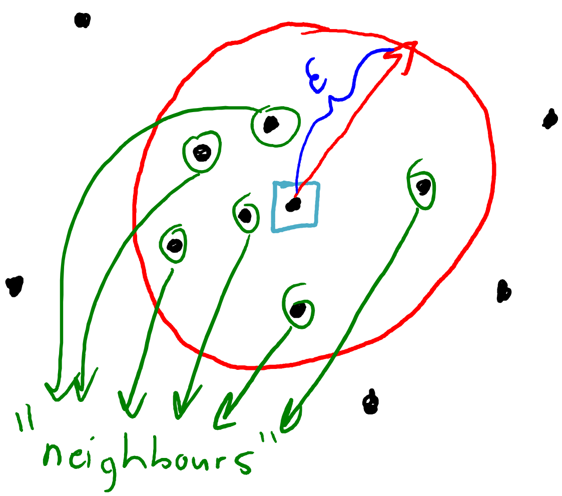
Density-Based Clustering

- Density-based clustering algorithm (DBSCAN) has two hyperparameters:
  - Epsilon (ε): distance we use to decide if another point is a "neighbour".



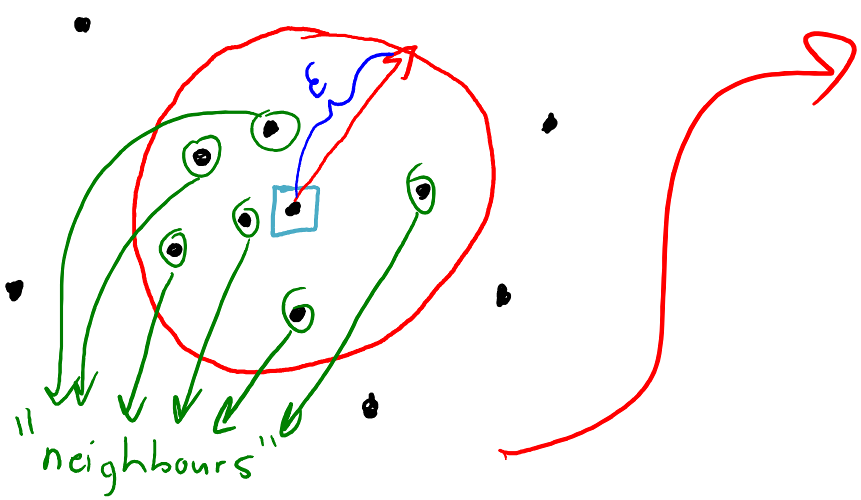
Density-Based Clustering

- Density-based clustering algorithm (DBSCAN) has two hyperparameters:
  - Epsilon (ε): distance we use to decide if another point is a "neighbour".



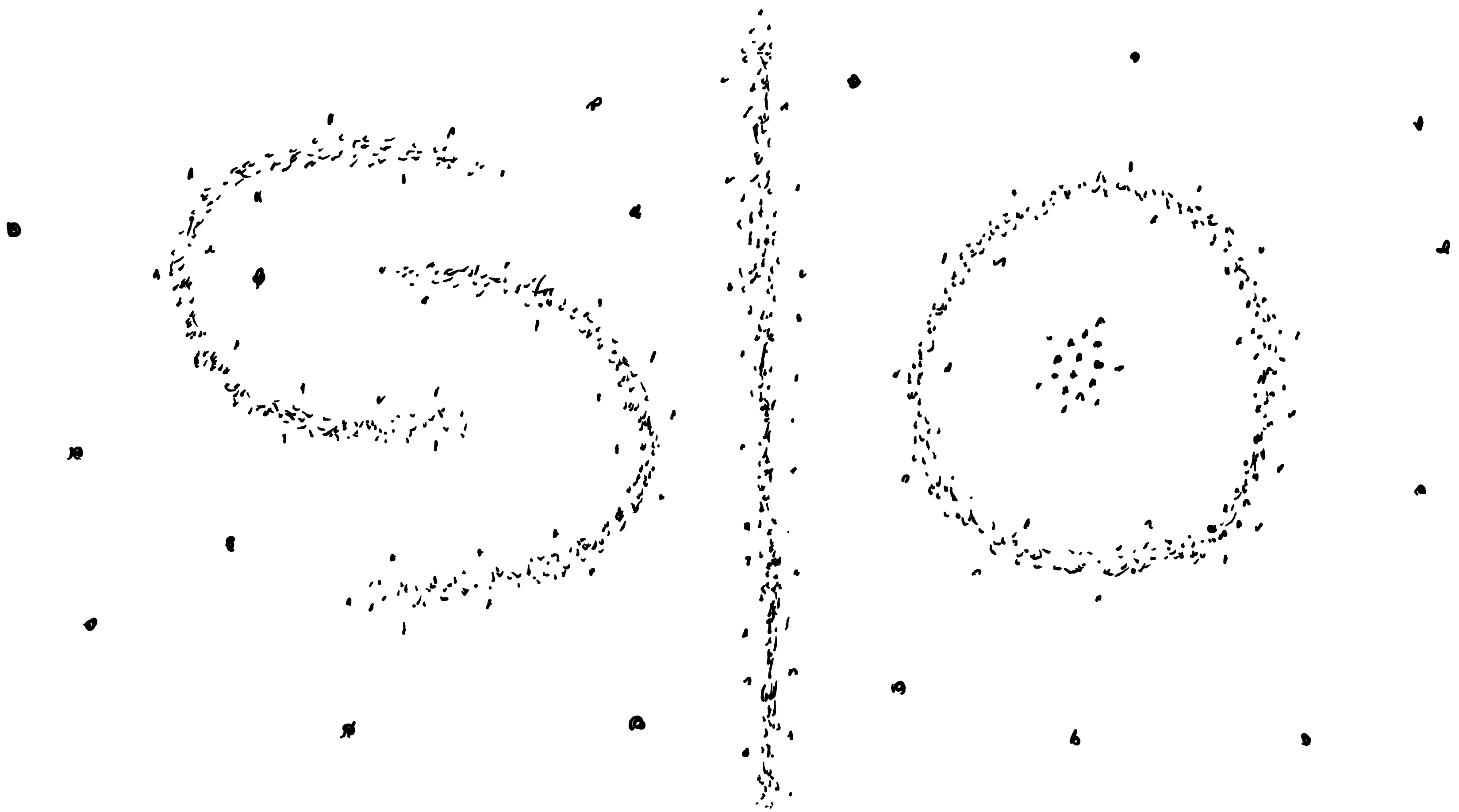
Density-Based Clustering

- Density-based clustering algorithm (DBSCAN) has two hyperparameters:
  - Epsilon (ε): distance we use to decide if another point is a "neighbour".
  - MinNeighbours: number of neighbours needed to say a region is "dense".
    - If you have at least MinNeighbours "neighbours", you are called a "core" point.
- Main idea: merge all neighbouring core points to form clusters.

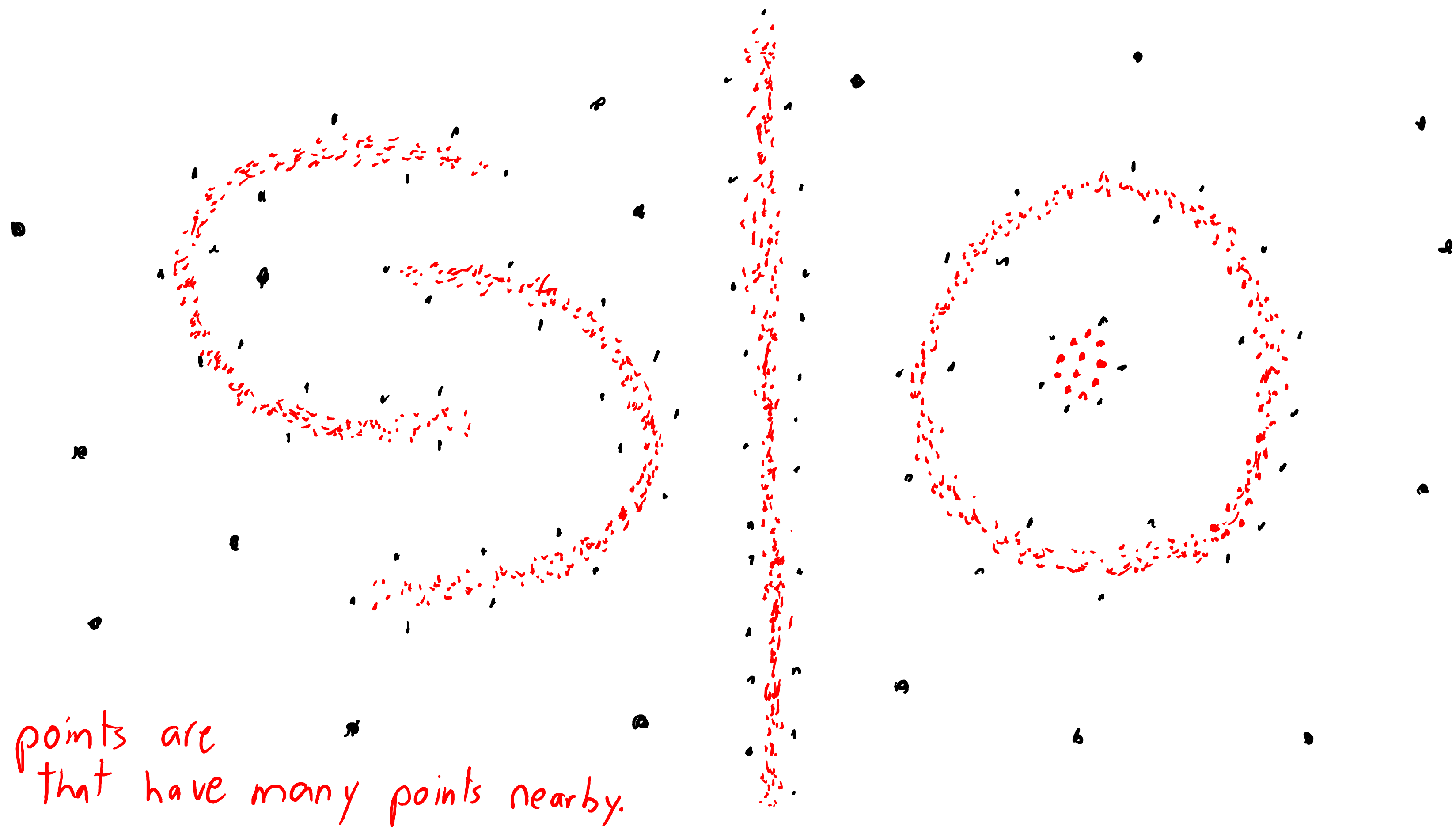


E.g., if minNeighbours = 3 then this is a "core" point since 6 points are "neighbours"

Density-Based Clustering

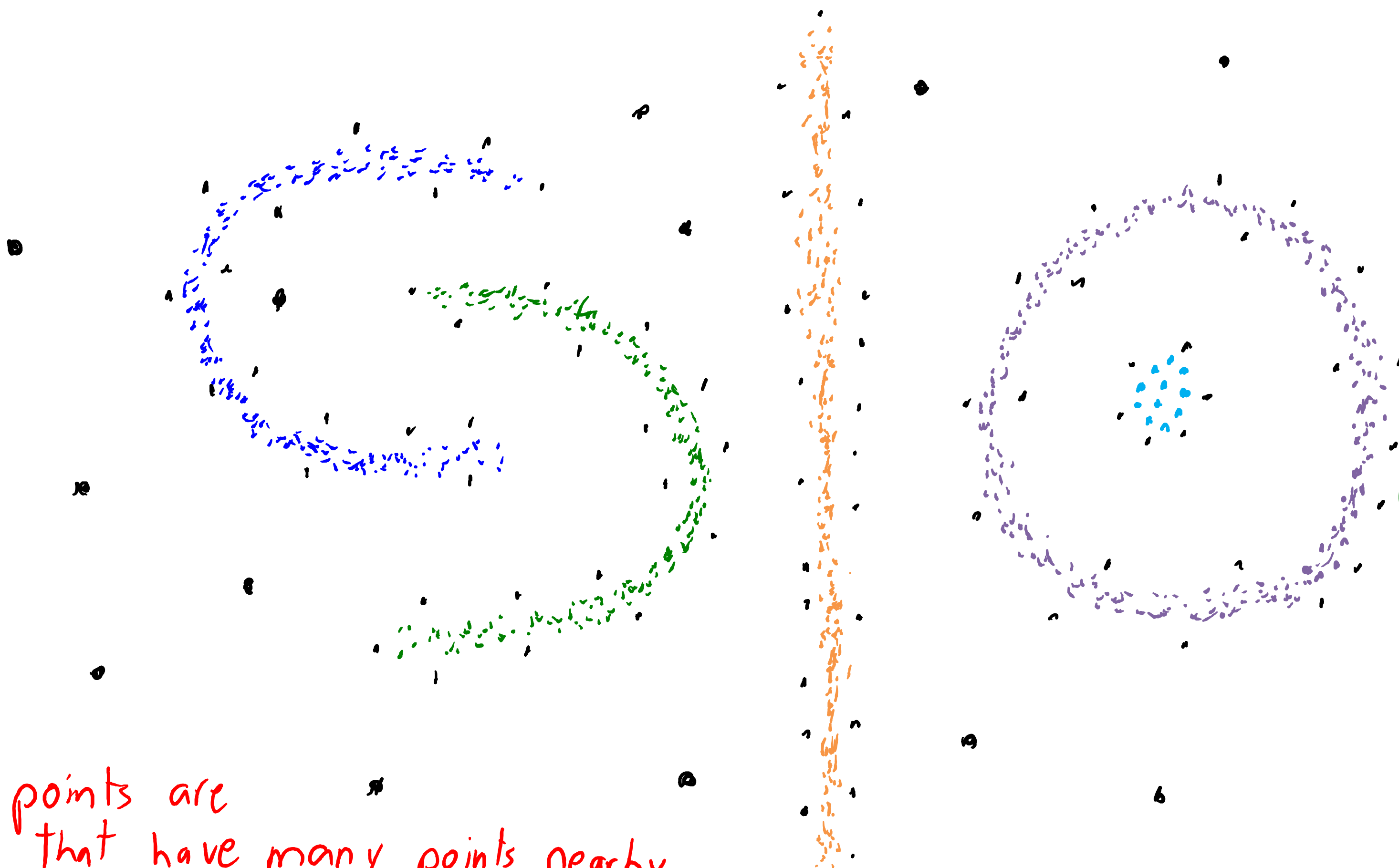


Density-Based Clustering



"Core" points are points that have many points nearby.

Density-Based Clustering

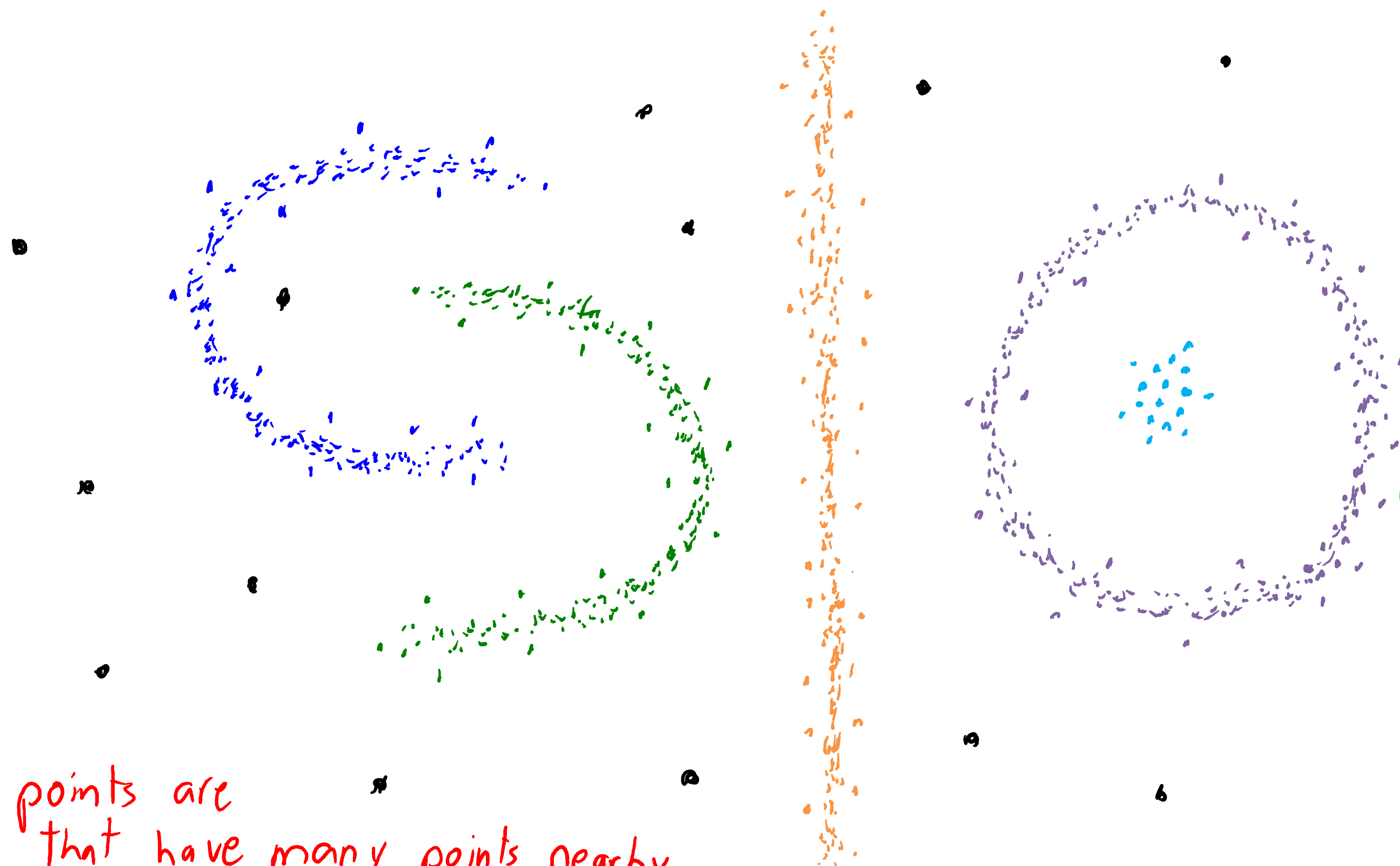


"Core" points are points that have many points nearby.

Clusters contain:

1. All "core" points that can be reached by following a <u>sequence</u> of core points.

Density-Based Clustering



"Core" points are points that have many points nearby.

Clusters contain:

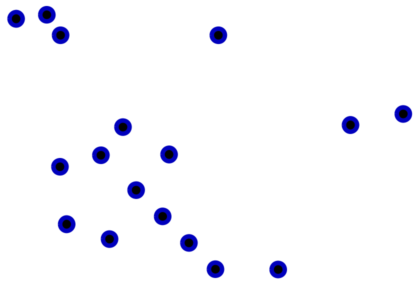
1. All "core" points that can be reached by following a sequence of core points.
2. All non-core neighbours of core points (boundary points)

Density-Based Clustering Pseudo-Code

- For each example x_i:
  - If x_i is already assigned to a cluster, do nothing.
  - Test whether x_i is a 'core' point ($\geq$ MinNeighbours examples within 'ε').
    - If x_i is not core point, do nothing (this could be an outlier).
    - If x_i is a core point, make a new cluster and call the "expand cluster" function.

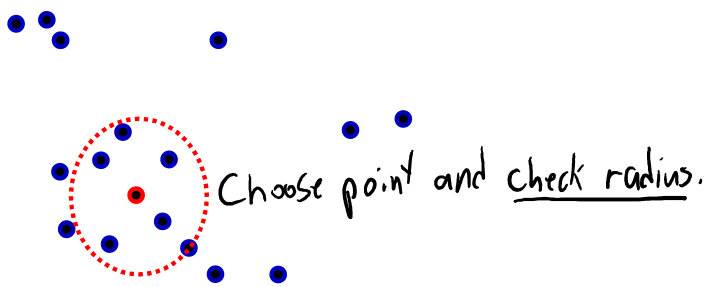
Density-Based Clustering Pseudo-Code

- "Expand cluster" function:
  - Assign to this cluster all x_j within distance 'ε' of core point x_i to this cluster.
  - For each new "core" point found, call "expand cluster" (recursively).



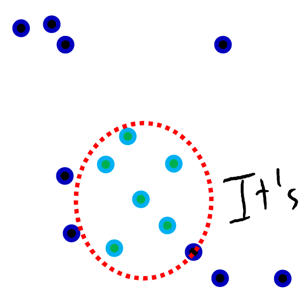
Density-Based Clustering Pseudo-Code

- "Expand cluster" function:
  - Assign to this cluster all x_j within distance 'ε' of core point x_i to this cluster.
  - For each new "core" point found, call "expand cluster" (recursively).



Density-Based Clustering Pseudo-Code

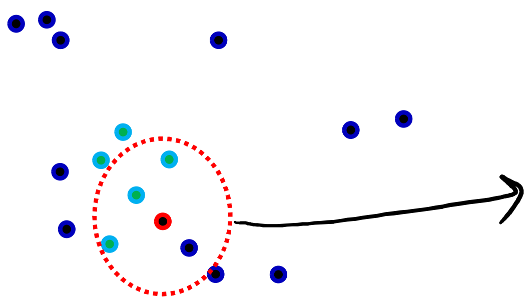
- "Expand cluster" function:
  - Assign to this cluster all x_j within distance 'ε' of core point x_i to this cluster.
  - For each new "core" point found, call "expand cluster" (recursively).



It's a "core" point, so make a new cluster with points within epsilon.

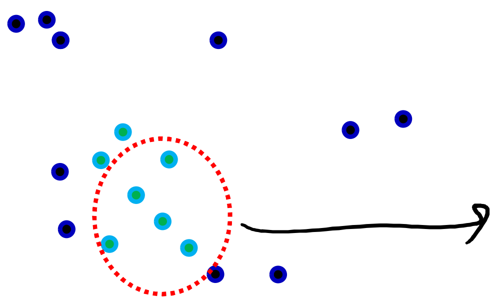
Density-Based Clustering Pseudo-Code

- "Expand cluster" function:
  - Assign to this cluster all x_j within distance 'ε' of core point x_i to this cluster.
  - For each new "core" point found, call "expand cluster" (recursively).



Density-Based Clustering Pseudo-Code

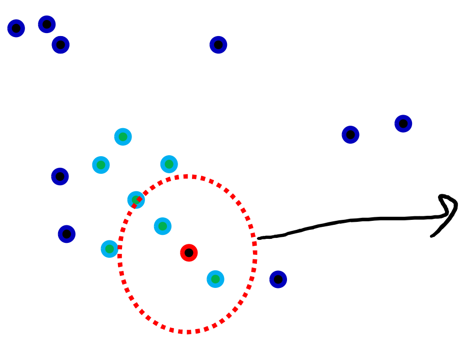
- "Expand cluster" function:
  - Assign to this cluster all x_j within distance 'ε' of core point x_i to this cluster.
  - For each new "core" point found, call "expand cluster" (recursively).



It's a "core" point, so add its neighbours.

Density-Based Clustering Pseudo-Code

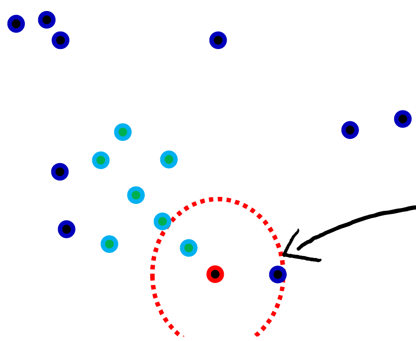
- “Expand cluster” function:
  - Assign to this cluster all x_j within distance ‘ε’ of core point x_i to this cluster.
  - For each new “core” point found, call “expand cluster” (recursively).



This is a "core" point too, so add its neighbours.

Density-Based Clustering Pseudo-Code

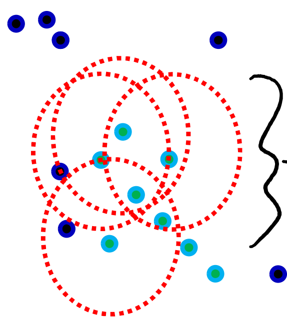
- "Expand cluster" function:
  - Assign to this cluster all x_j within distance 'ε' of core point x_i to this cluster.
  - For each new "core" point found, call "expand cluster" (recursively).



Not a "core" point, so do not add its neighbour.

Density-Based Clustering Pseudo-Code

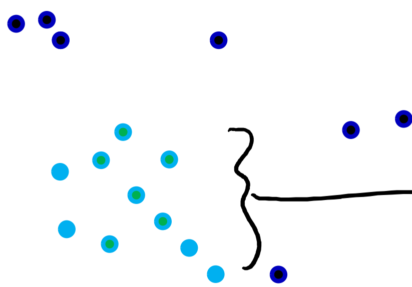
- "Expand cluster" function:
  - Assign to this cluster all x_j within distance 'ε' of core point x_i to this cluster.
  - For each new "core" point found, call "expand cluster" (recursively).



Continue testing whether the other cluster point are "core" points.

Density-Based Clustering Pseudo-Code

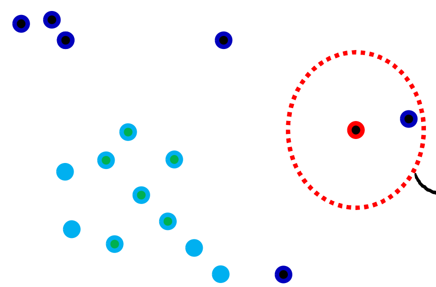
- "Expand cluster" function:
  - Assign to this cluster all x_j within distance 'ε' of core point x_i to this cluster.
  - For each new "core" point found, call "expand cluster" (recursively).



Eventually, there will be no new "core" points to "expand."

Density-Based Clustering Pseudo-Code

- "Expand cluster" function:
  - Assign to this cluster all x_j within distance 'ε' of core point x_i to this cluster.
  - For each new "core" point found, call "expand cluster" (recursively).

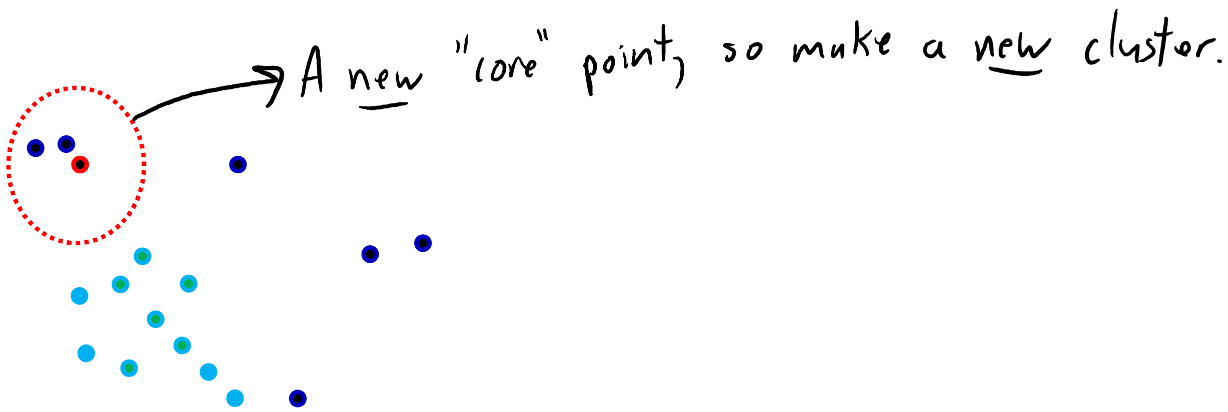


Not enough points in radius => not a "core" point.

Don't assign cluster

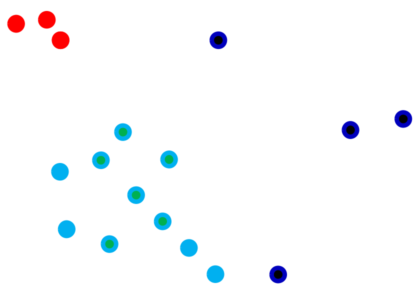
Density-Based Clustering Pseudo-Code

- "Expand cluster" function:
  - Assign to this cluster all x_j within distance 'ε' of core point x_i to this cluster.
  - For each new "core" point found, call "expand cluster" (recursively).



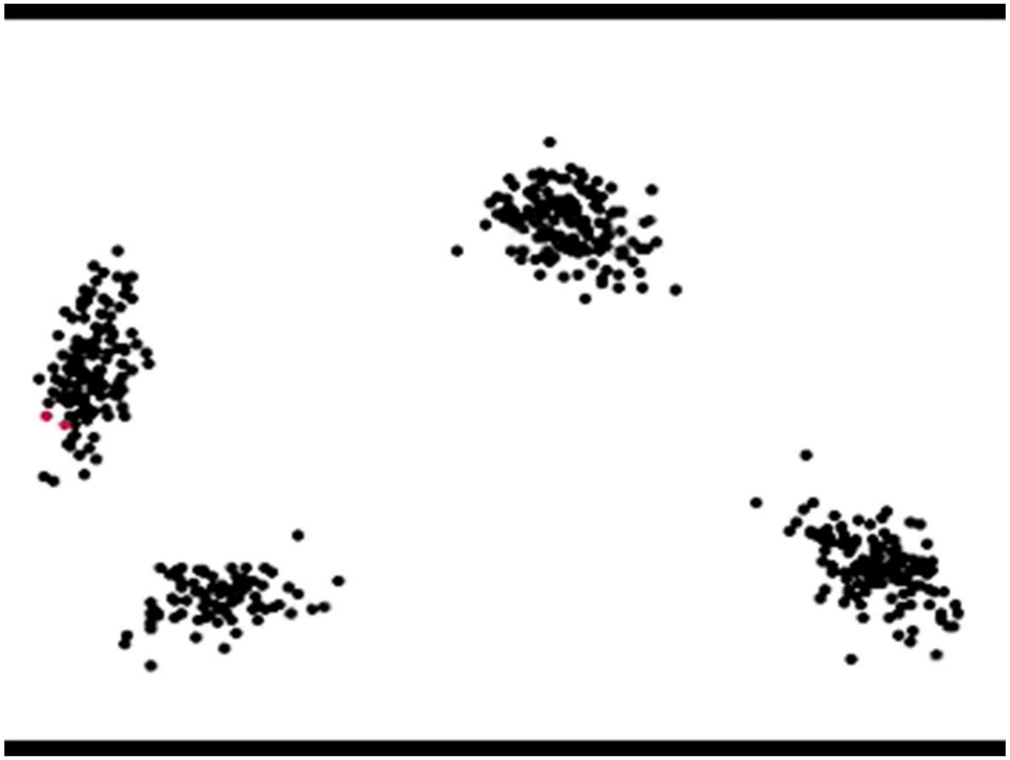
Density-Based Clustering Pseudo-Code

- "Expand cluster" function:
  - Assign to this cluster all x_j within distance 'ε' of core point x_i to this cluster.
  - For each new "core" point found, call "expand cluster" (recursively).



Final result.

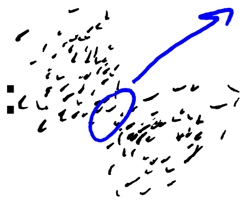
Density-Based Clustering in Action



[Interactive demo]

Density-Based Clustering Issues

- Some points are not assigned to a cluster.
  - Good or bad, depending on the application.
- Ambiguity of "non-core" (boundary) points:



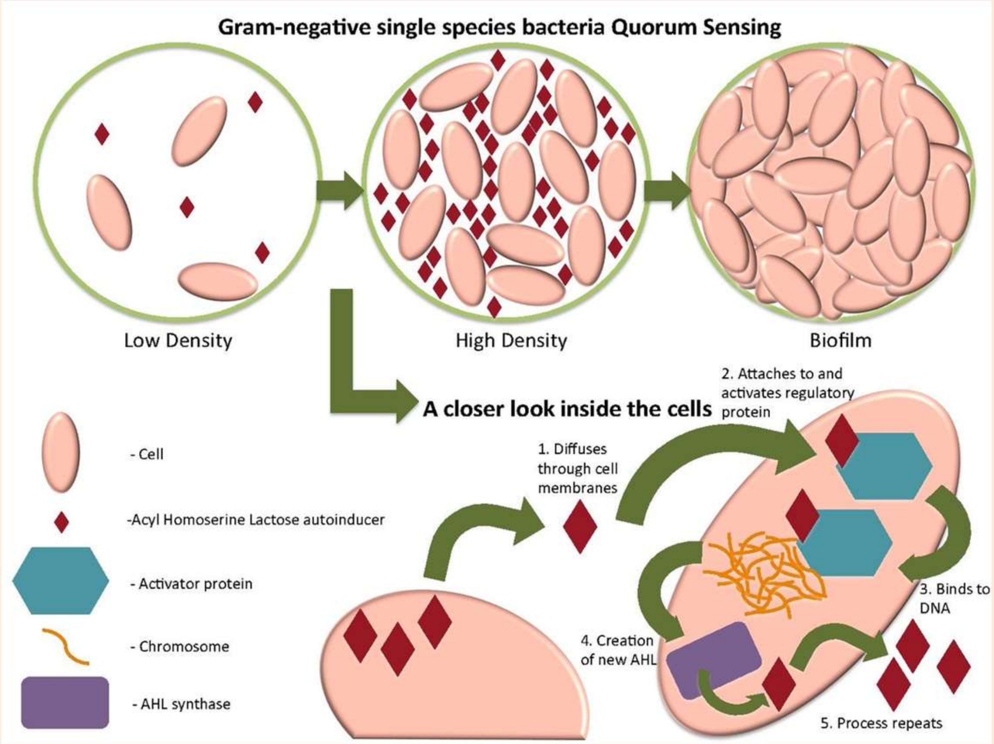
Points at "boundary" between clusters.

- Sensitive to the choice of ε and MinNeighbours.
  - Original paper proposed an "elbow" method (see bonus slide).
  - Otherwise, not sensitive to initialization (except for boundary points).

- If you get a new example, finding cluster is expensive.
  - Need to compute distances to core points (or maybe all training points).
- In high-dimensions, need a lot of points to 'fill' the space.

Density-Based Clustering in Nature

- Quorum sensing:
  - Bacteria continuously release a particular molecule.
  - Bacteria have sensors for this molecule.
- If sensors become very active:
  - It means cell density is high.
  - Causes cascade of changes in cells. (Some cells "stick together" to form a physical cluster via "biofilm".)



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

Coming Up Next

ENSEMBLE CLUSTERING AND LABEL SWITCHING

Ensemble Clustering

question

stop following 23 views

Multiple random runs of K means

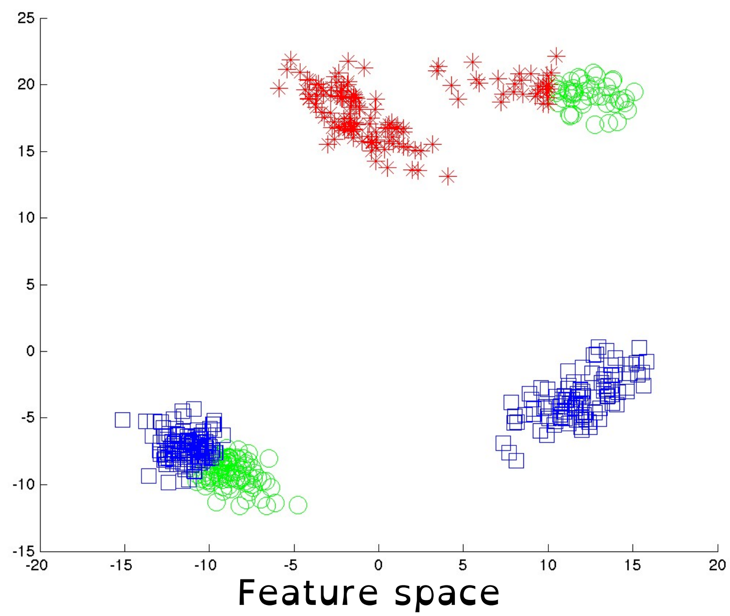
I was wondering how running K Means (original version, not K means ++) several times with random initializations can help us make an accurate model. K Means outputs the class labels of all the samples. We definitely can't use mode of all the labels it got in different runs because class labels from different runs don't make any sense when compared. We somehow have to see what points are coming in the same cluster in a lot of runs..I am not sure, how do we do it?

- We can consider ensemble methods for clustering.
  - "Consensus clustering"
- It's a good/important idea:
  - Bootstrapping is widely-used.
  - "Do clusters change if the data was slightly different?"

Q: What can go wrong with ensemble clustering?

Ensemble Clustering

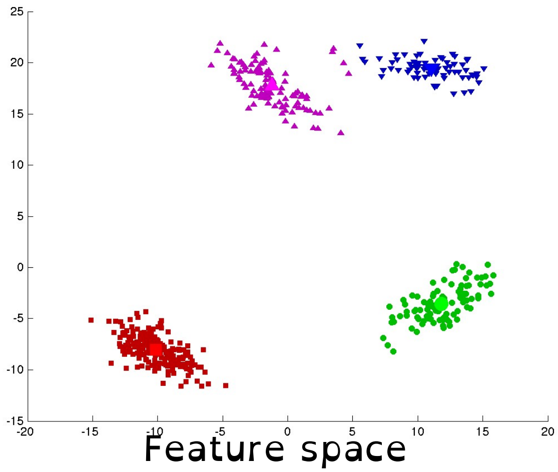
- E.g., run k-means 20 times and then cluster using the mode of each $\hat{y}_i$.
- Normally, averaging across models doing different things is good.

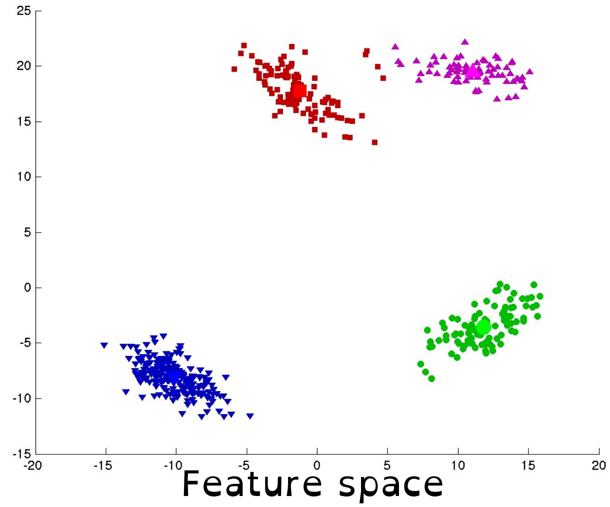


- But this is a bad ensemble method: worse than k-means on its own.

Label Switching Problem

- This doesn't work because of "label switching" problem:
  - The cluster labels $\hat{y}_i$ are meaningless.
  - We could get same clustering with _____________ labels

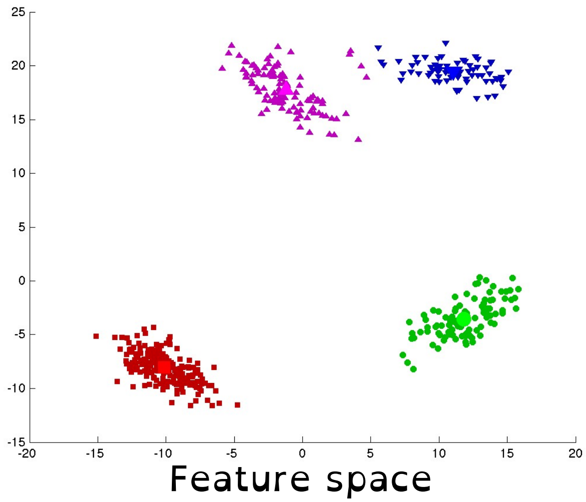


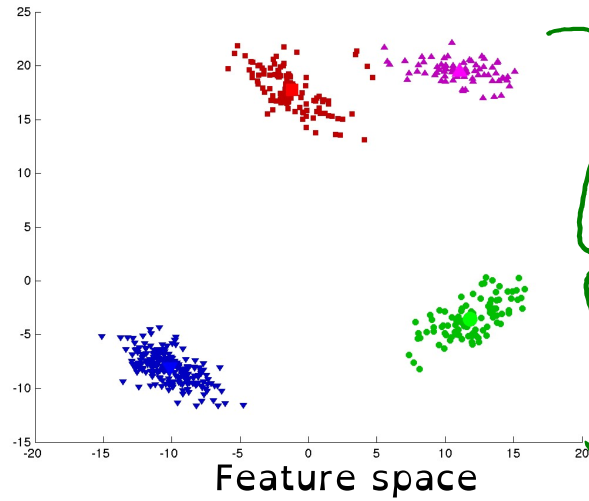


  - All $\hat{y}_i$ become equally likely as number of initializations increases.

Addressing Label Switching Problem

- Ensembles can't depend on label "meaning":
  - Don't ask "is point x_i in red square cluster?", which is meaningless.
  - Ask "is point x_i in the same cluster as x_j?", which is meaningful.





Different permutation of labels but same groups of points.

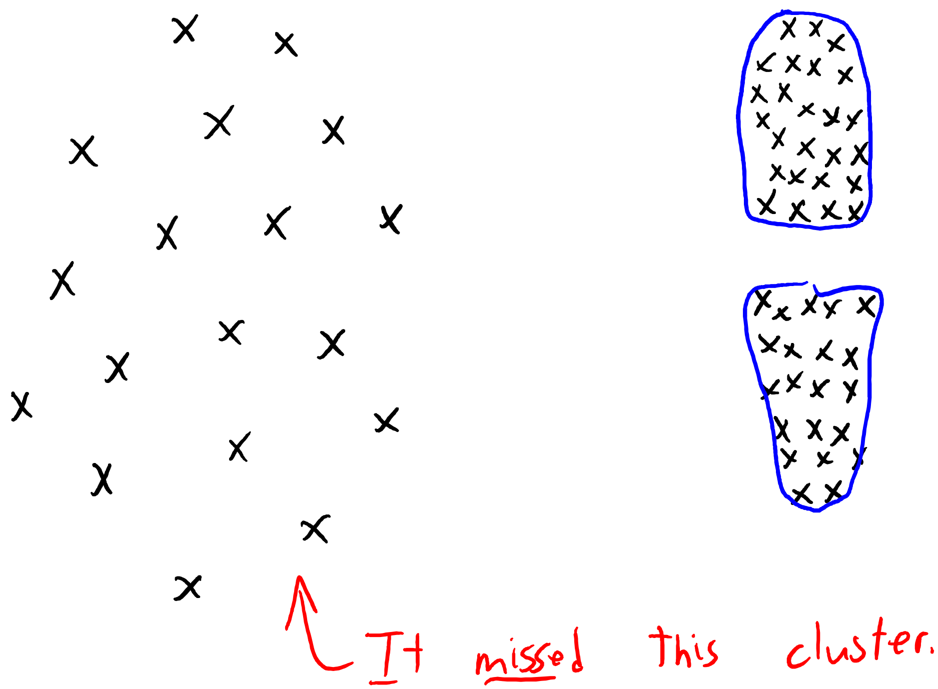
  - Bonus slides give an example method ("UBClustering").

Coming Up Next

HIERARCHICAL CLUSTERING

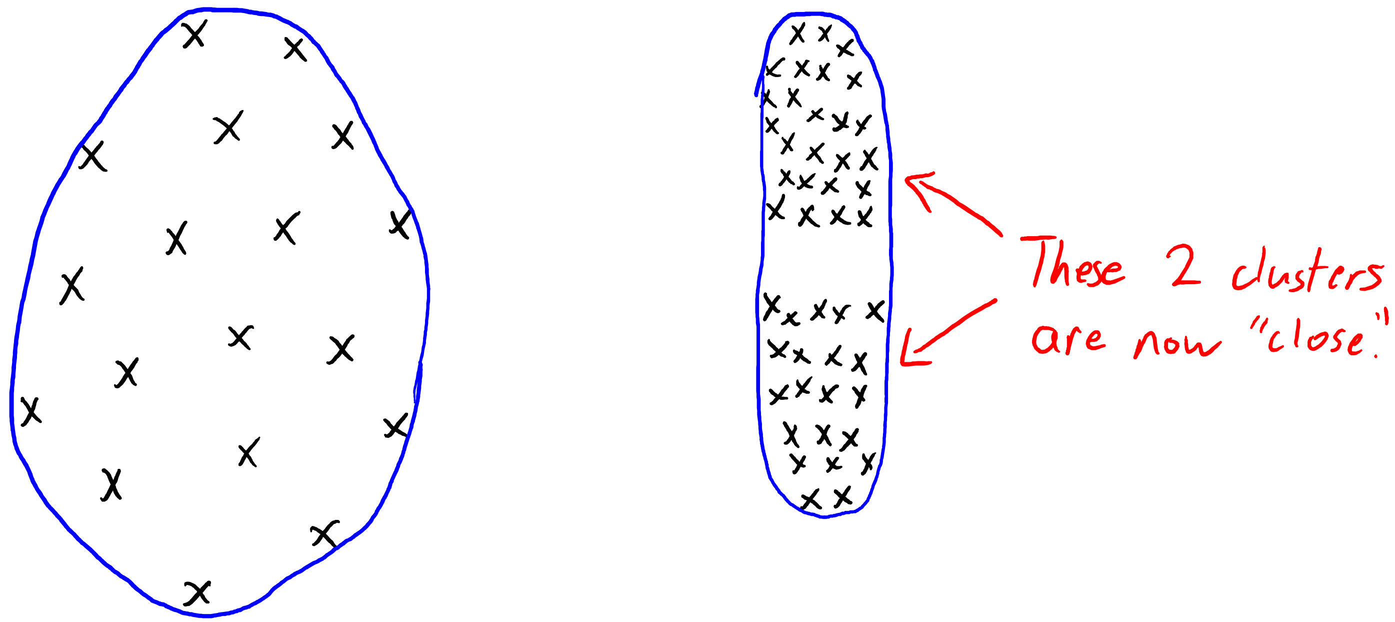
Differing Densities

- Consider density-based clustering on this data:



Differing Densities

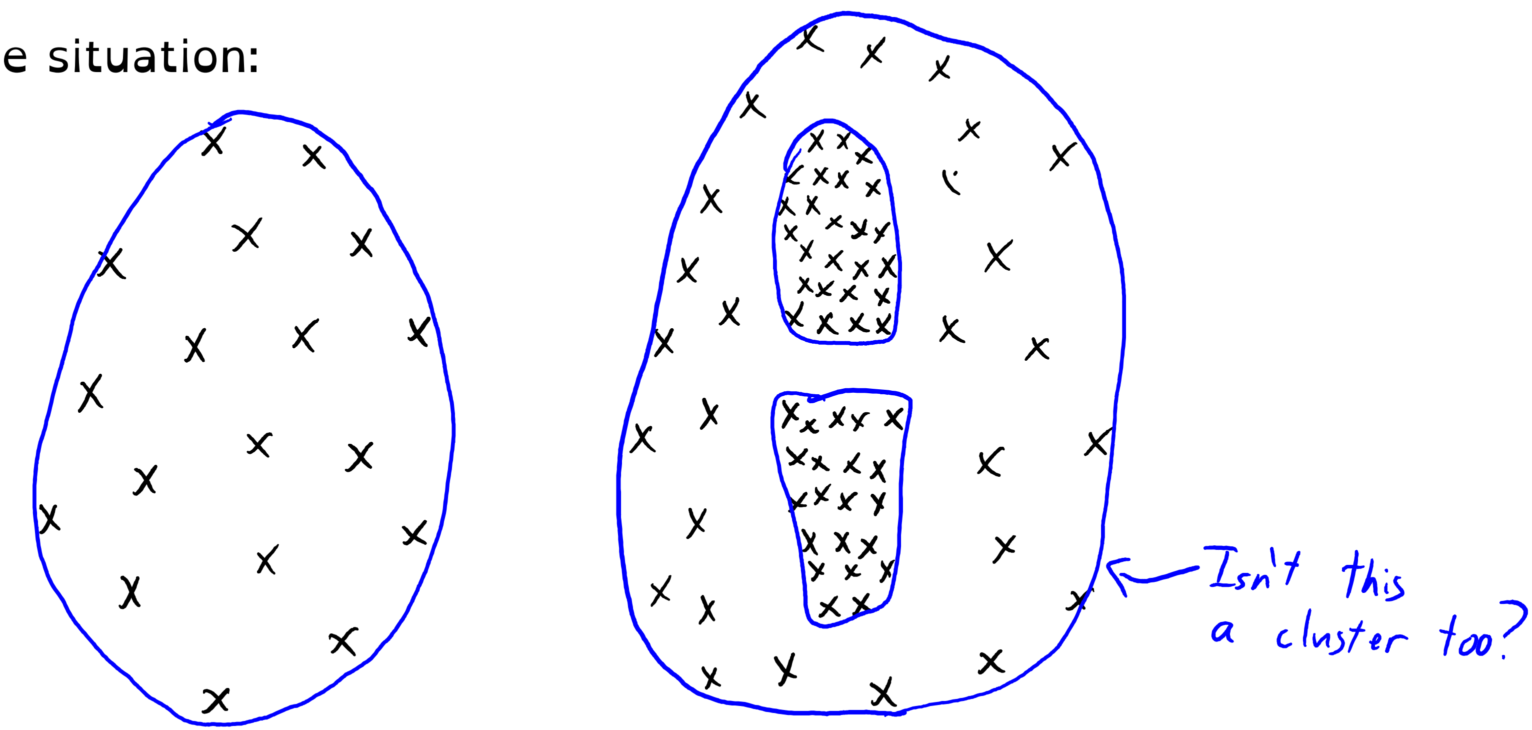
- Increase epsilon and run it again:



- There may be no density-level that gives you 3 clusters.

Differing Densities

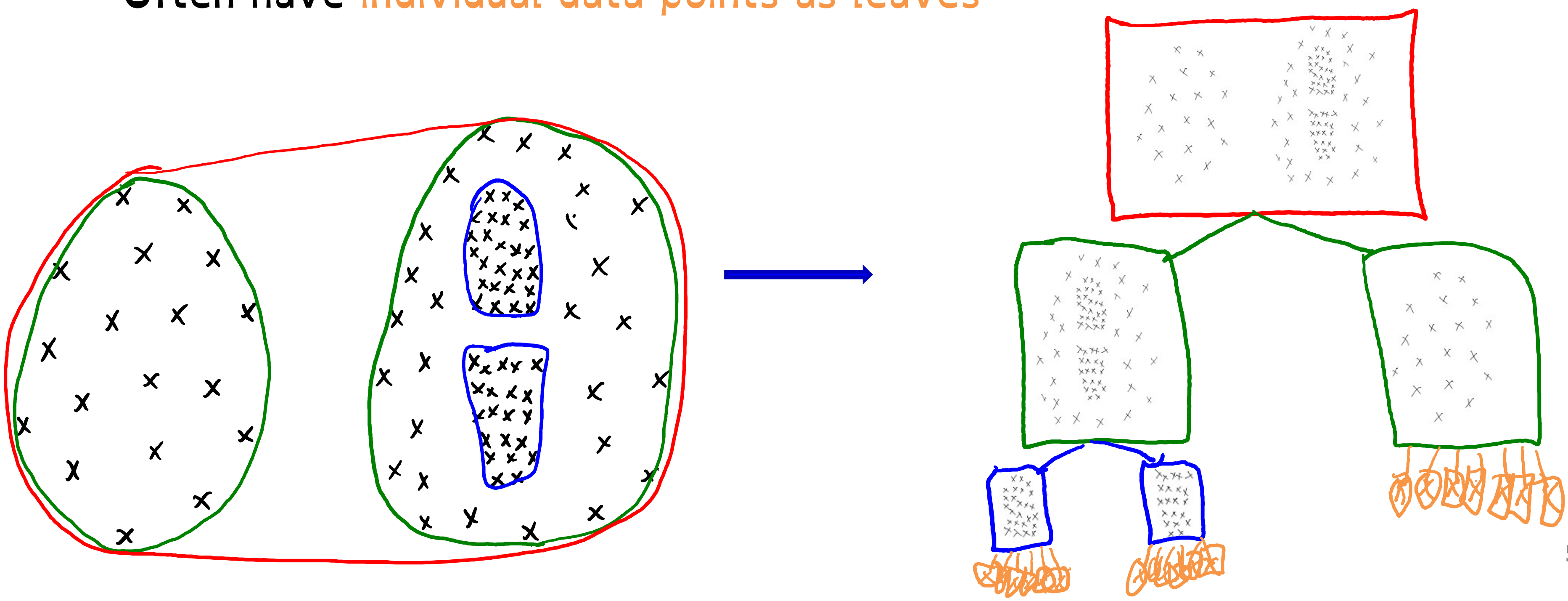
- Here is a worse situation:



- Now you need to choose between coarse/fine clusters.
- Instead of fixed clustering, we often want hierarchical clustering.

Hierarchical Clustering

- Hierarchical clustering produces a _____ of clusterings.
  - Each node in the tree splits the data into 2 or more clusters.
  - Much more information than using a fixed clustering.
  - Often have individual data points as leaves



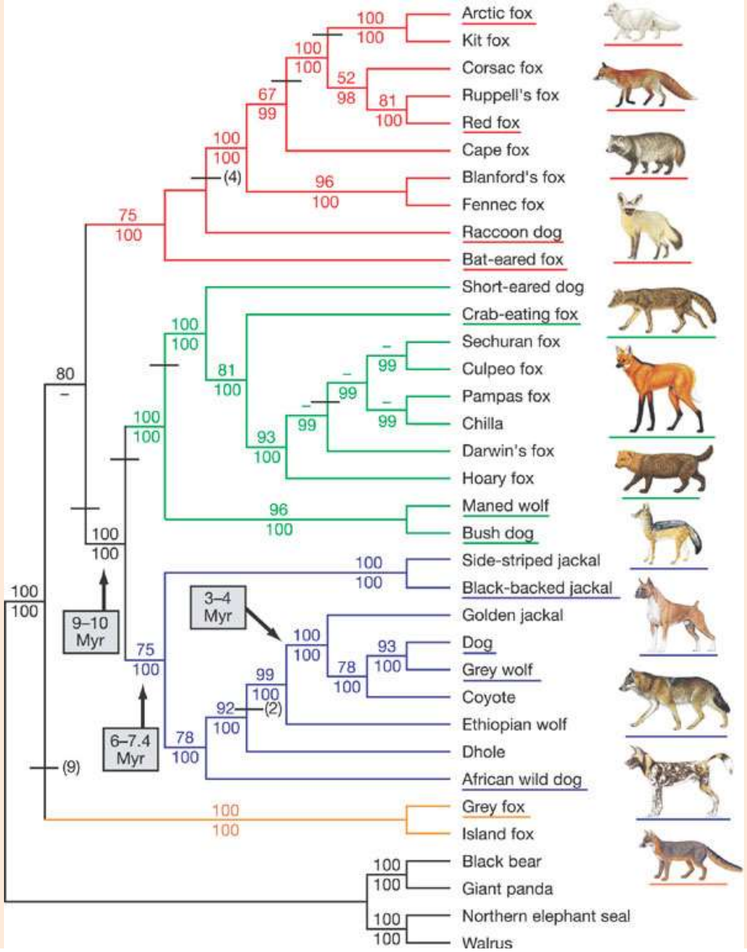
GIF

Application: Phylogenetics

- We sequence genomes of a set of organisms.
- Can we construct the "tree of life"?

- Comments on this application:
  - On the right are individuals.
  - As you go left, clusters merge.
  - Merges are 'common ancestors'.

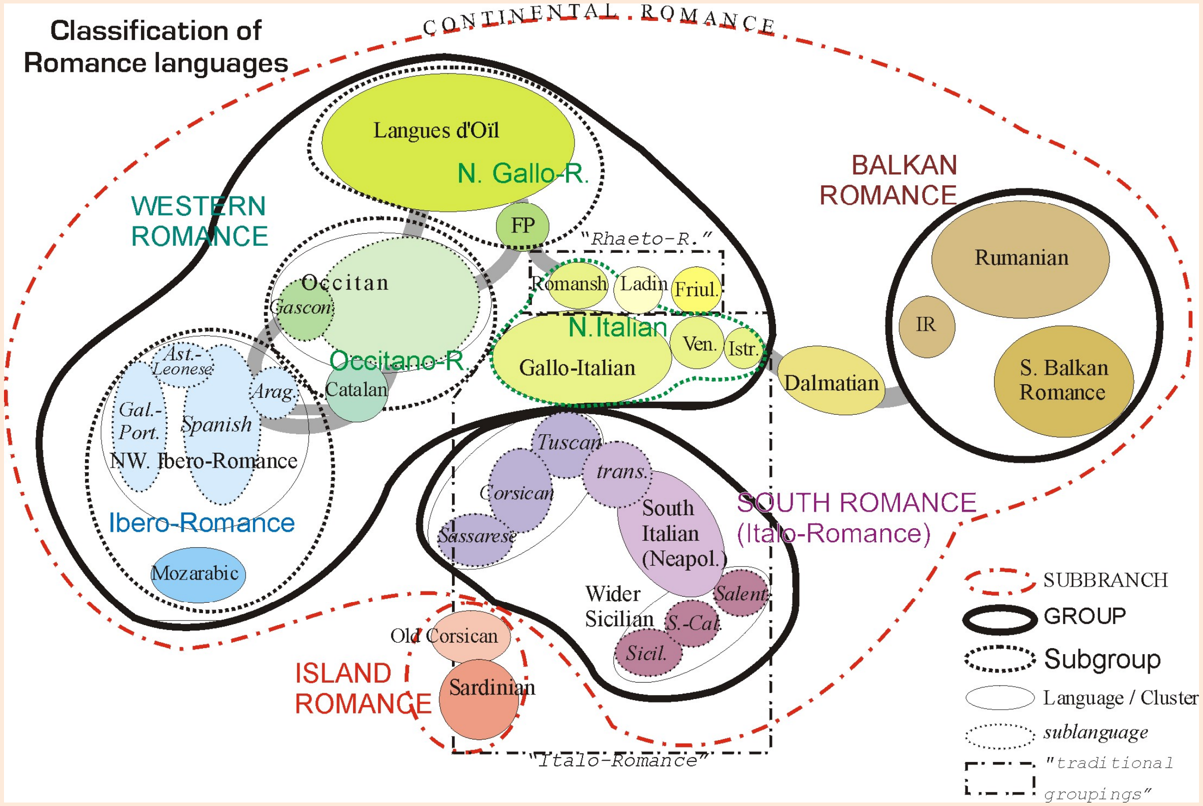
- More useful information in the plot:
  - Line lengths: chosen here to approximate time.
  - Numbers: #clusterings across bootstrap samples.
  - 'Outgroups' (walrus, panda) are a sanity check.



http://www.nature.com/nature/journal/v438/n7069/fig_tab/nature04338_F10.html

Application: Phylogenetics

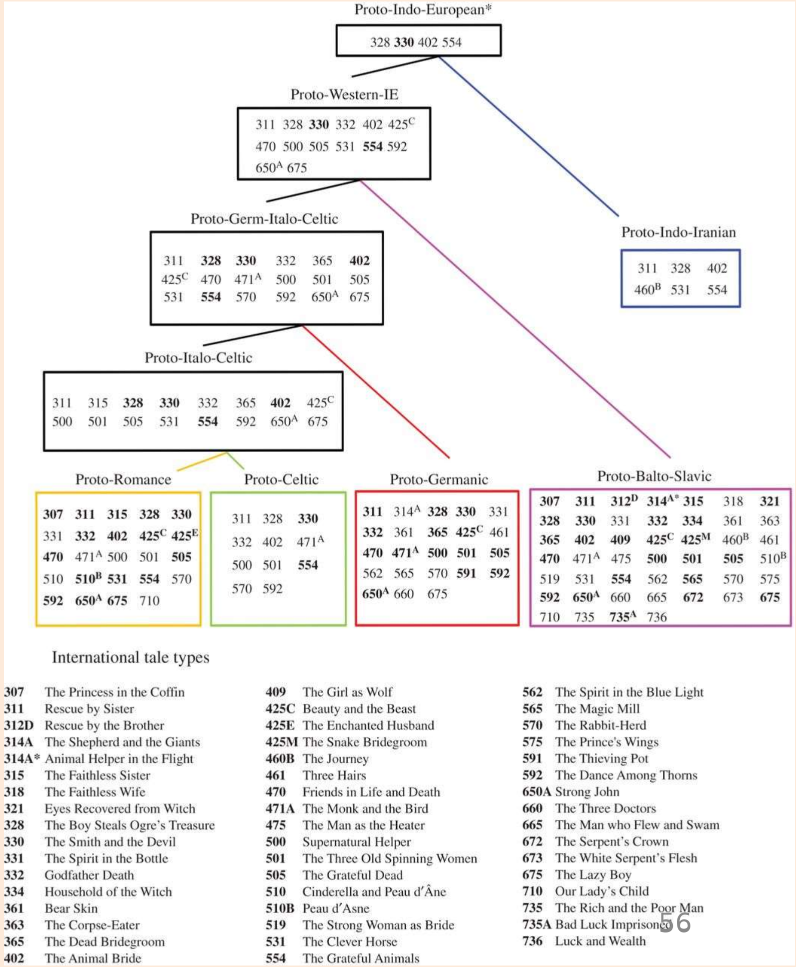
- Comparative method in linguistics studies evolution of languages:



https://en.wikipedia.org/wiki/Comparative_method_(linguistics)

Application: Phylogenetics

- January 2016: evolution of fairy tales.
  - Evidence that “Devil and the Smith” goes back to bronze age.
  - “Beauty and the Beast” published in 1740, but might be 2500-6000 years old.

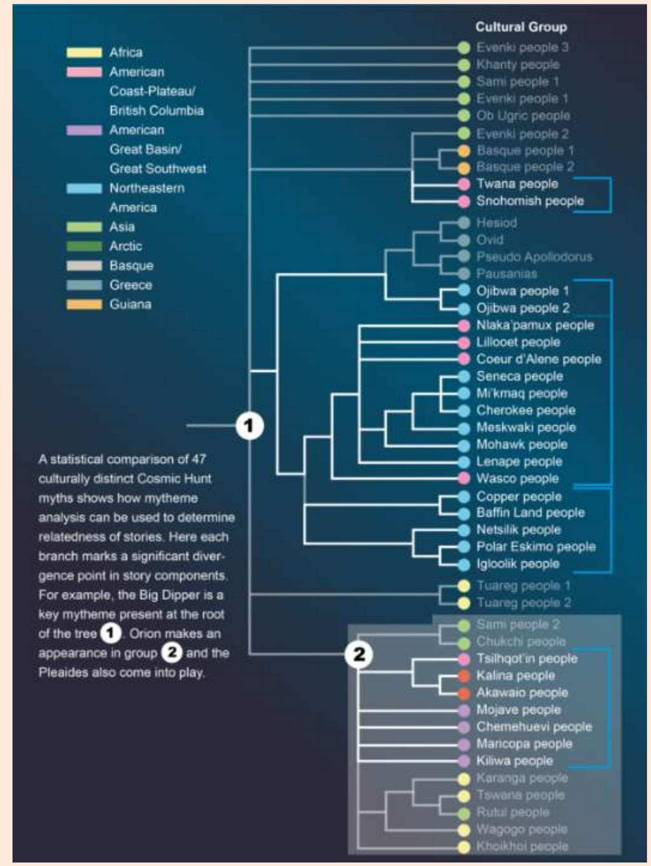


http://rsos.royalsocietypublishing.org/content/3/1/150645

Application: Phylogenetics

- January 2016: evolution of fairy tales.
  - Evidence that “Devil and the Smith” goes back to bronze age.
  - “Beauty and the Beast” published in 1740, but might be 2500-6000 years old.

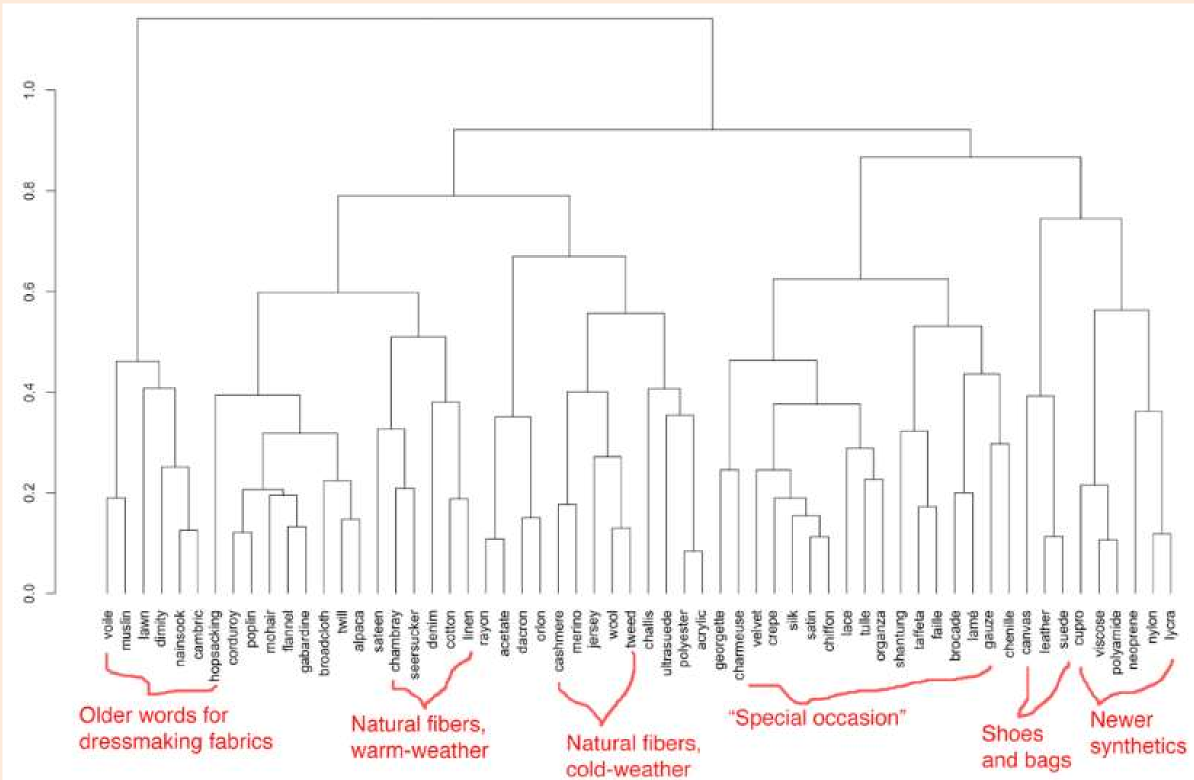
- September 2016: evolution of myths.
  - “Cosmic hunt” story:
    - Person hunts animal that becomes constellation.
      - Previously known to be at least 15,000 years old.
    - May go back to paleololithic period.



http://www.nature.com/nature/journal/v438/n7069/fig_tab/nature04338_F10.html

Application: Fashion?

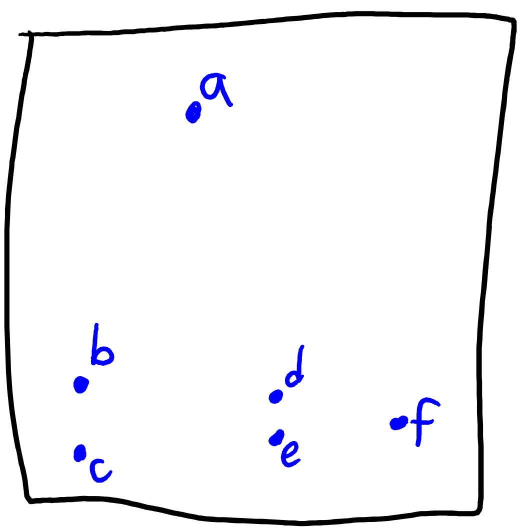
- Hierarchical clustering of clothing material words in Vogue:



http://dh.library.yale.edu/projects/vogue/fabricspace/

Agglomerative (Bottom-Up) Clustering

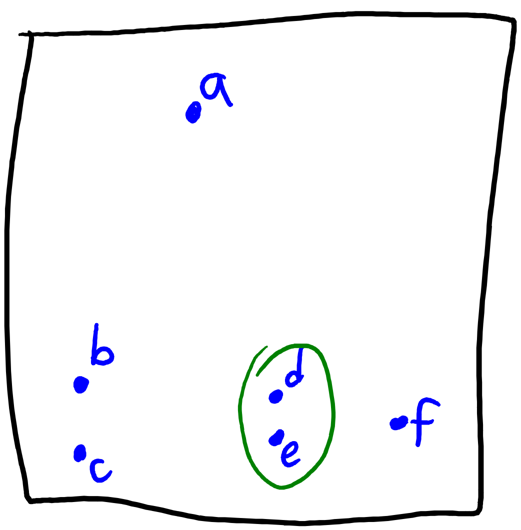
- Most common hierarchical method: agglomerative clustering.
    1. Starts with each point in its own cluster.

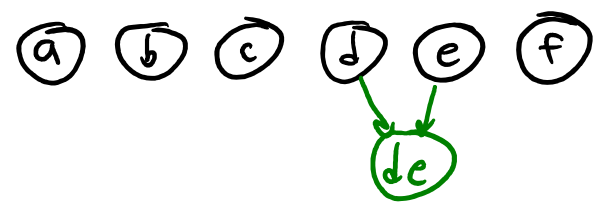


https://en.wikipedia.org/wiki/Hierarchical_clustering

Agglomerative (Bottom-Up) Clustering

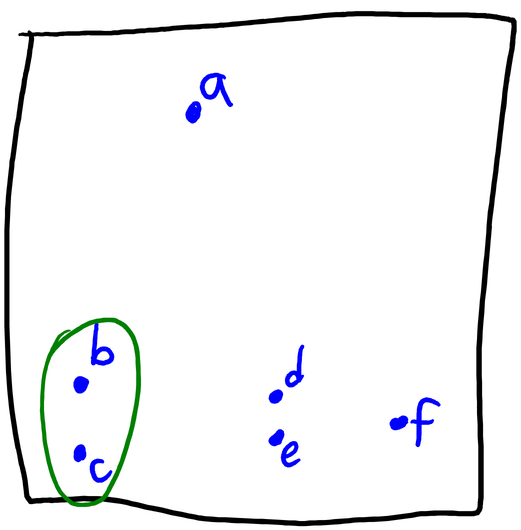
- Most common hierarchical method: agglomerative clustering.
  1. Starts with each point in its own cluster.
  2. Each step merges the two "closest" clusters.

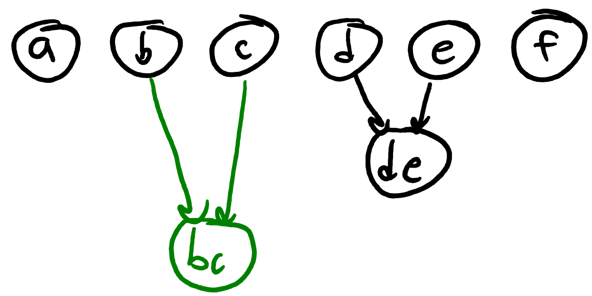




Agglomerative (Bottom-Up) Clustering

- Most common hierarchical method: agglomerative clustering.
  1. Starts with each point in its own cluster.
  2. Each step merges the two "closest" clusters.

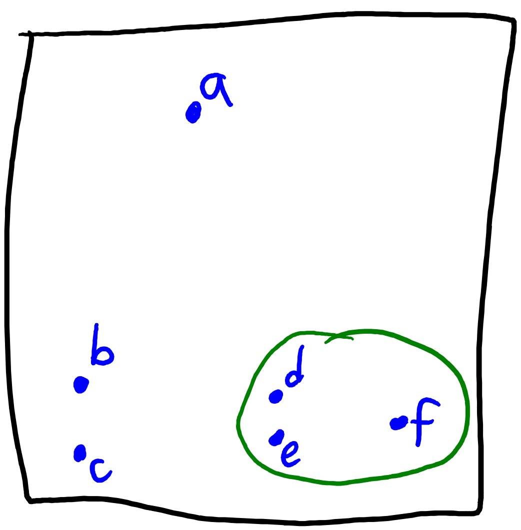


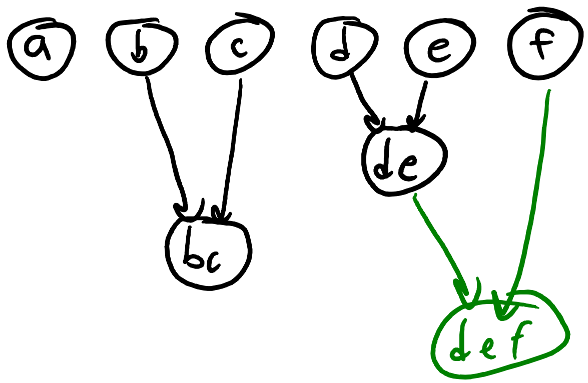


https://en.wikipedia.org/wiki/Hierarchical_clustering

Agglomerative (Bottom-Up) Clustering

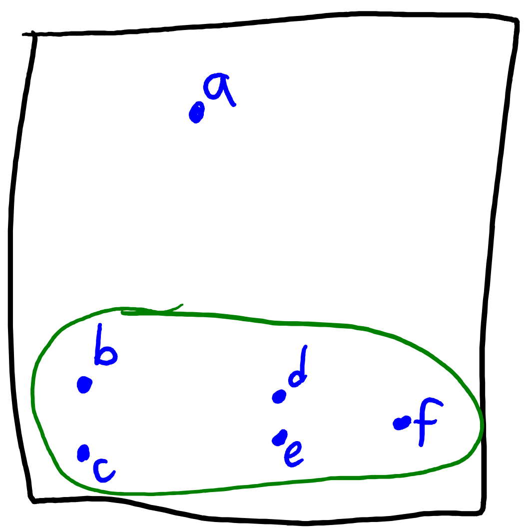
- Most common hierarchical method: agglomerative clustering.
  1. Starts with each point in its own cluster.
  2. Each step merges the two "closest" clusters.

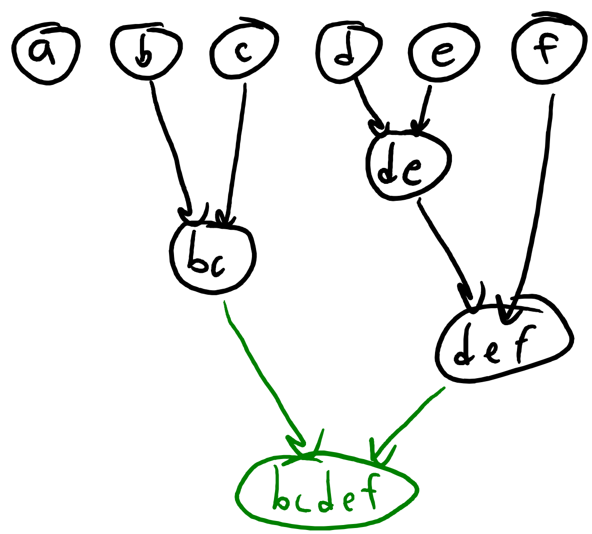




Agglomerative (Bottom-Up) Clustering

- Most common hierarchical method: agglomerative clustering.
  1. Starts with each point in its own cluster.
  2. Each step merges the two “closest” clusters.

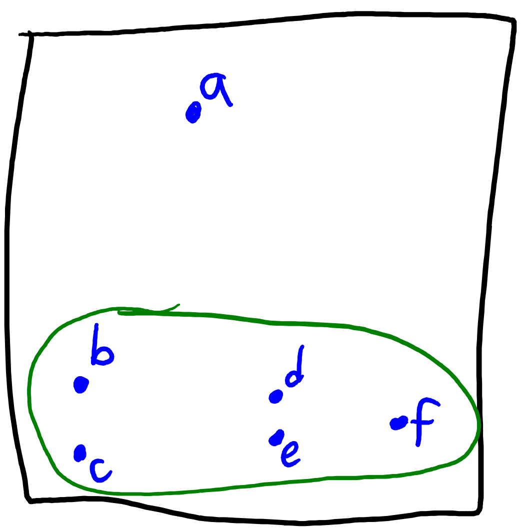


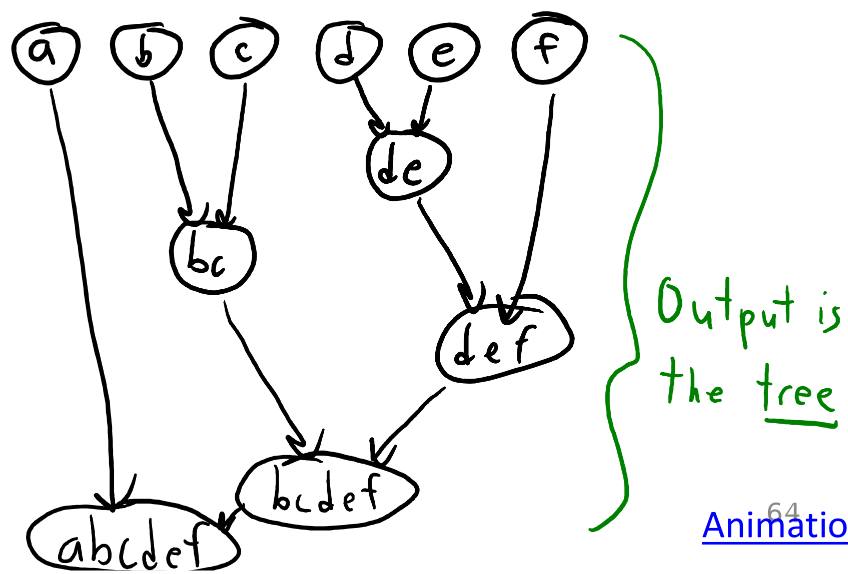


https://en.wikipedia.org/wiki/Hierarchical_clustering

Agglomerative (Bottom-Up) Clustering

- Most common hierarchical method: agglomerative clustering.
  1. Starts with each point in its own cluster.
  2. Each step merges the two "closest" clusters.
  3. Stop with one big cluster that has all points.





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

Animation

Agglomerative (Bottom-Up) Clustering

- Reinvented by different fields under different names ("UPGMA").
- Needs a "distance" between two clusters.
- A standard choice: distance between means of the clusters.
  – Not necessarily the best, many choices exist (bonus slide).
- Cost is O(___) for basic implementation.
  – Each step costs O(___), and each step might only cluster 1 new point.

Summary

- Shape of K-means clusters:
  - Partitions space into convex sets.
- Density-based clustering:
  - "Expand" and "merge" dense regions of points to find clusters.
  - Not sensitive to initialization or outliers.
  - Useful for finding non-convex connected clusters.
- Ensemble clustering: combines multiple clusterings.
  - Can work well but need to account for label switching.
- Hierarchical clustering: more informative than fixed clustering.
- Agglomerative clustering: standard hierarchical clustering method.
  - Each point starts as a cluster, sequentially merge clusters.

- Next time:
  - Discovering (and then ignoring) a hole in the ozone layer.

Review Questions

- Q1: Does k-means with k=2 still produce convex partitions of feature space?
- Q2: What is the main problem of density-based clustering with fixed "core" point criteria?
- Q3: Why do we need $O(n^2)$ distances for each step of hierarchical clustering?

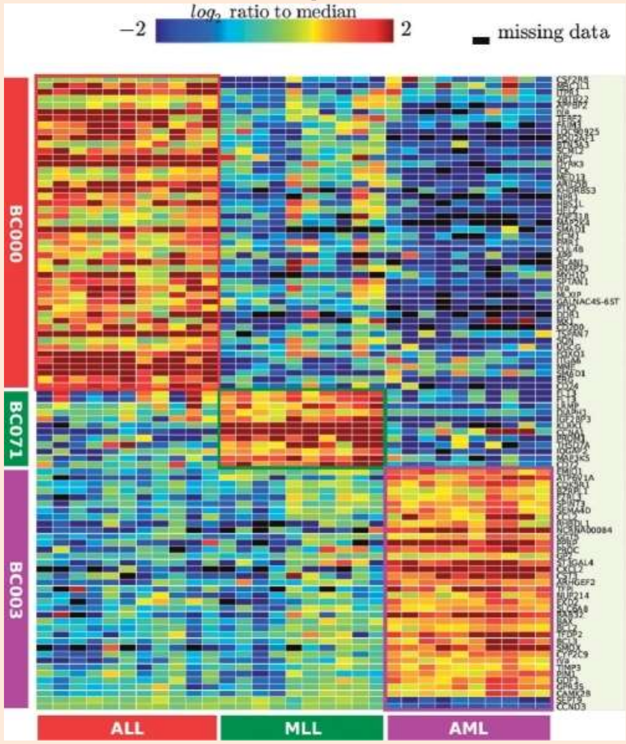
Bi-clustering

- Bi-clustering:
  - Cluster the training examples and features.
  - Also gives feature relationship information.

- Simplest and most popular method:
  - Run clustering method on ‘X’ (examples).
  - Run clustering method on ‘X^T’ (features).

- Often plotted with ‘X’ as a heatmap.
  - Where rows/columns arranged by clusters.
  - Helps you ‘see’ why things are clustered.

Matrix plot



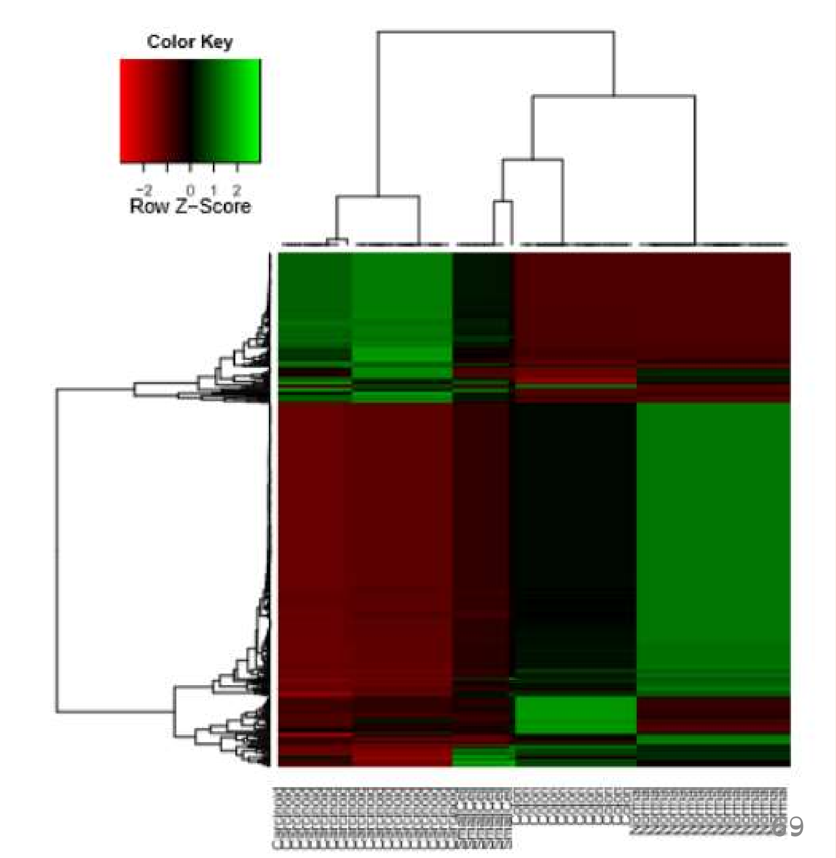
Sorted by feature cluster index

Sorted by example cluster index

http://openi.nlm.nih.gov/detailedresult.php?img=2731891_gkp491f3&req=4

Bi-clustering

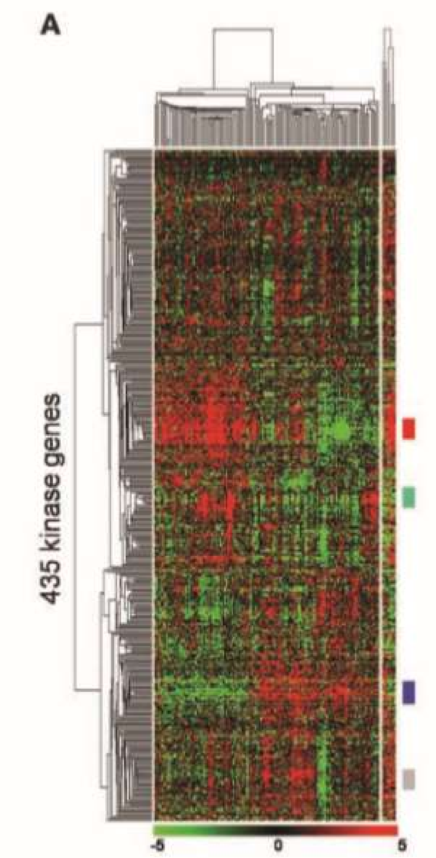
- Visualization: hierarchical bi-clustering + heatmap + dendrograms.
  - Popular in biology/medicine.



https://arxiv.org/pdf/1408.0856v1.pdf

Application: Medical data

- Hierarchical clustering is very common in medical data analysis.
  - Bi-clustering different samples of breast cancer:



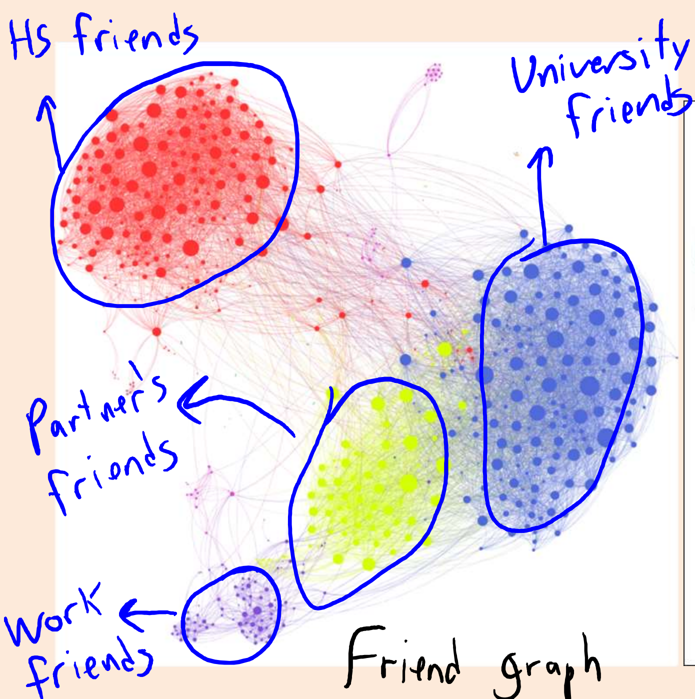
http://members.cbio.mines-paristech.fr/~jvert/svn/bibli/local/Finetti2008Sixteen-kinase.pdf

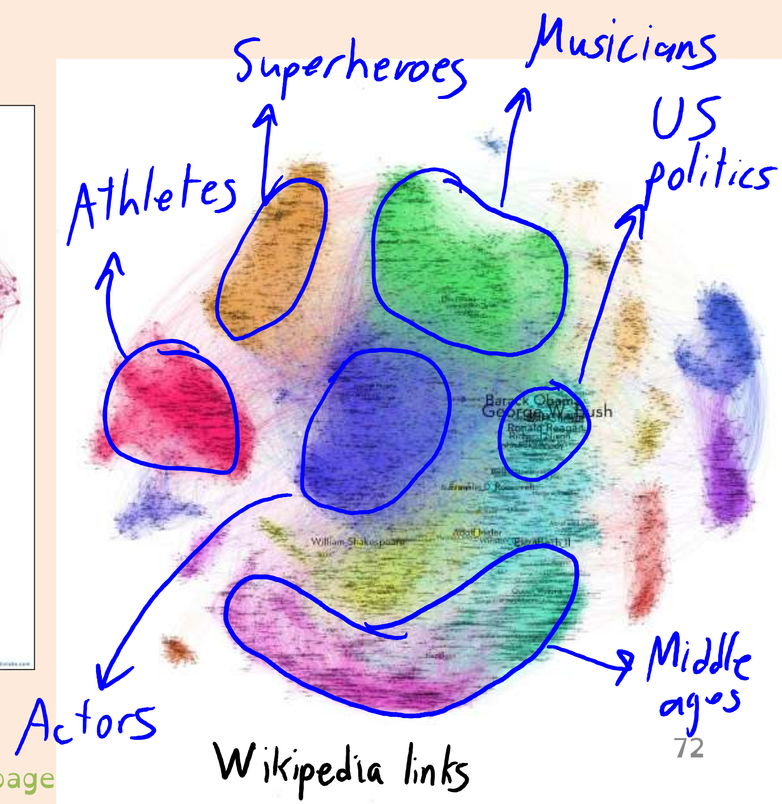
Other Clustering Methods

- Mixture models (540):
  - Probabilistic clustering.
- Mean-shift clustering:
  - Finds local “modes” in density of points.
  - Alternative approach to vector quantization.
- Bayesian clustering:
  - A variant on ensemble methods.
  - Averages over models/clusterings, weighted by “prior” belief in the model/clustering.

Graph-Based Clustering

- Spectral clustering and graph-based clustering:
  - Clustering of data described by graphs.

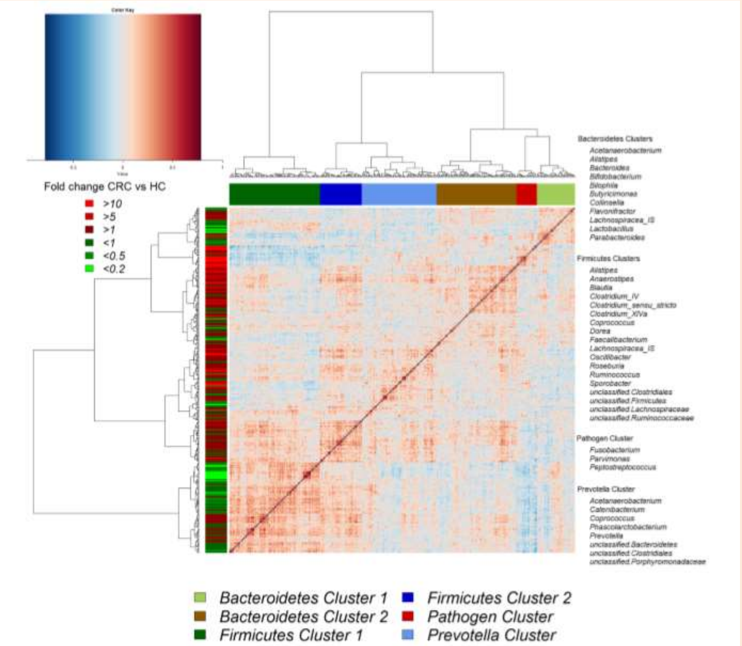




https://griffsgraphs.wordpress.com/tag/clustering/
http://ascr-discovery.science.doe.gov/2013/09/sifting-genomes/
https://www.hackdiary.com/2012/04/05/extracting-a-social-graph-from-wikipedia-people-page

Application: Medical data

- Hierarchical clustering is very common in medical data analysis.
  - Clustering different samples of colorectoral cancer:



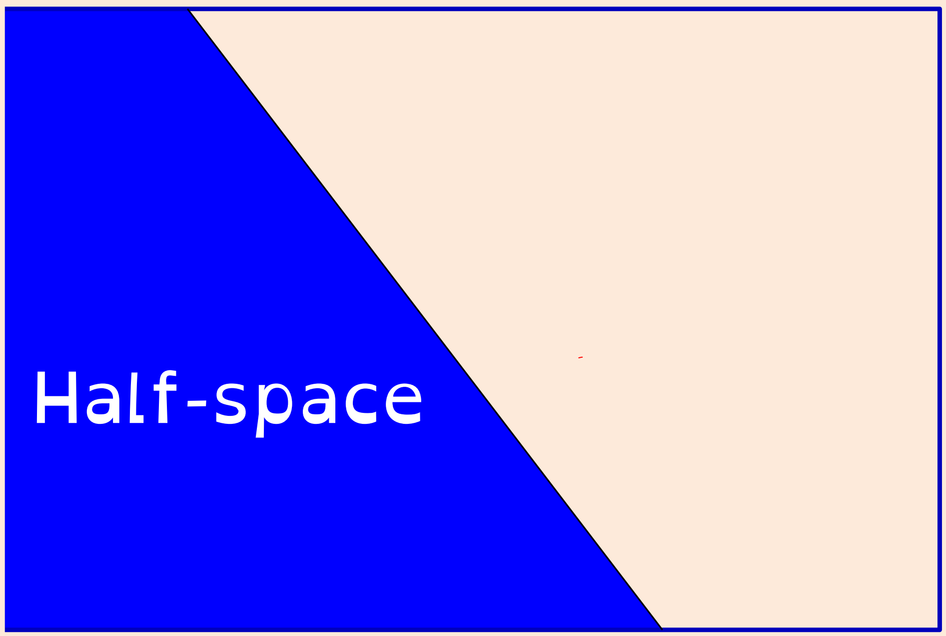
  - This plot is different, it's not a biclustering:
    - The matrix is 'n' by 'n'.
    - Each matrix element gives correlation.
    - Clusters should look like "blocks" on diagonal.
    - Order of examples is reversed in columns.
      - This is why diagonal goes from bottom-to-top.
      - Please don't do this reversal, it's confusing to me.

https://gut.bmj.com/content/gutjnl/66/4/633.full.pdf

Why are k-means clusters convex?

- K-means clusters are formed by the intersection of half-spaces.

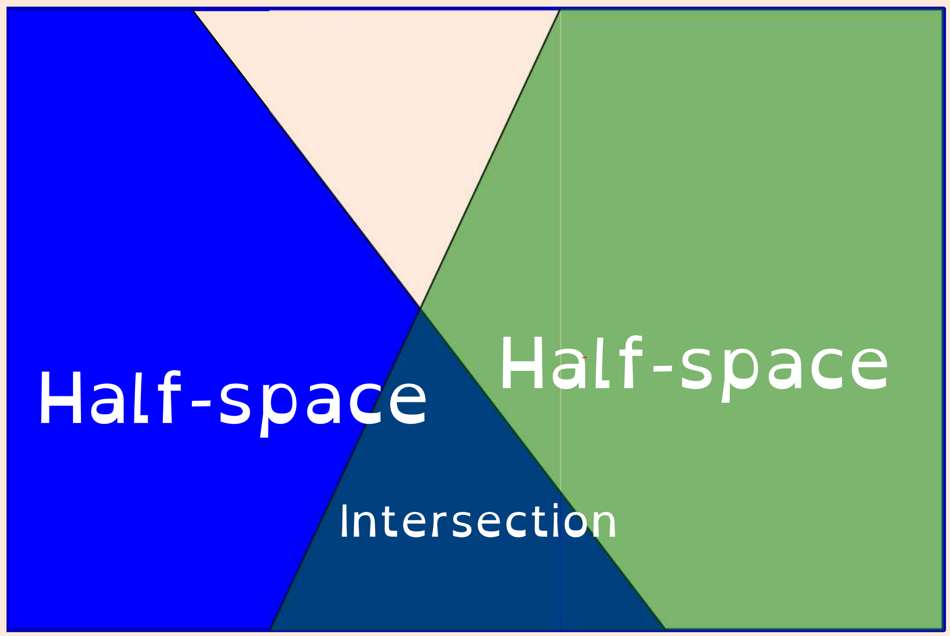
Half-space is Set of points satifying 'x' a linear inequality, like $\sum_{j=1}^{d} a_j x_j \le b$



Why are k-means clusters convex?

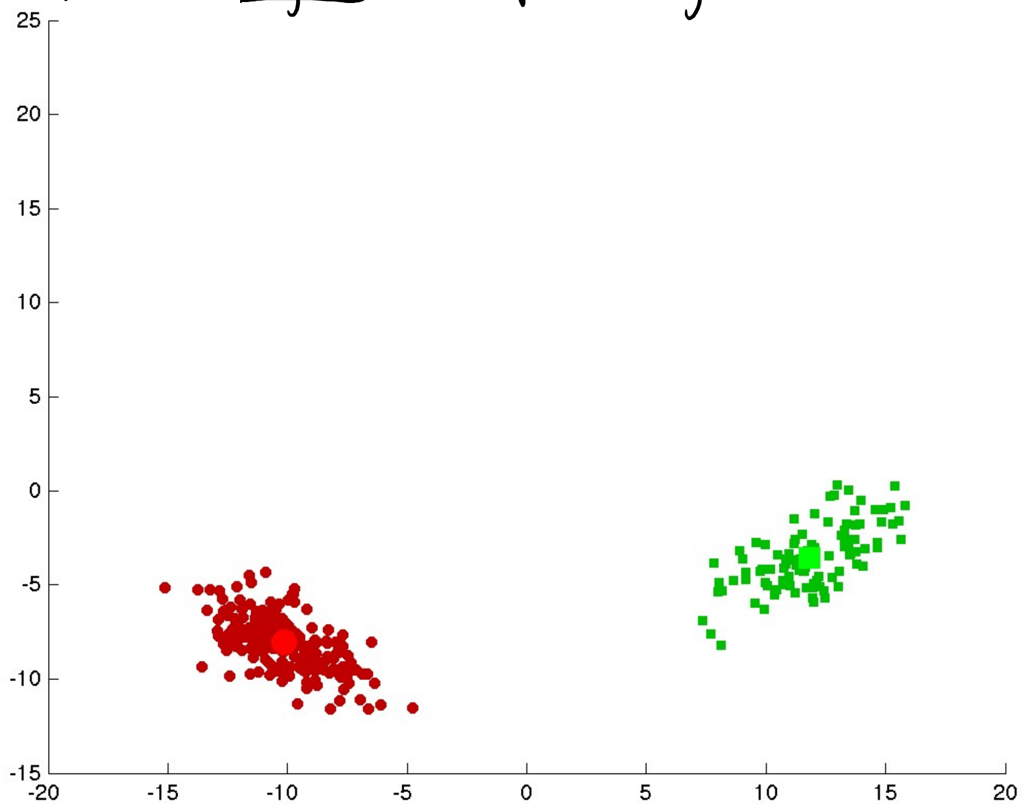
- K-means clusters are formed by the intersection of half-spaces.

Half-space is Set of points 'x' satifying a linear inequality, like $\sum_{j=1}^{d} a_j x_j \le b$

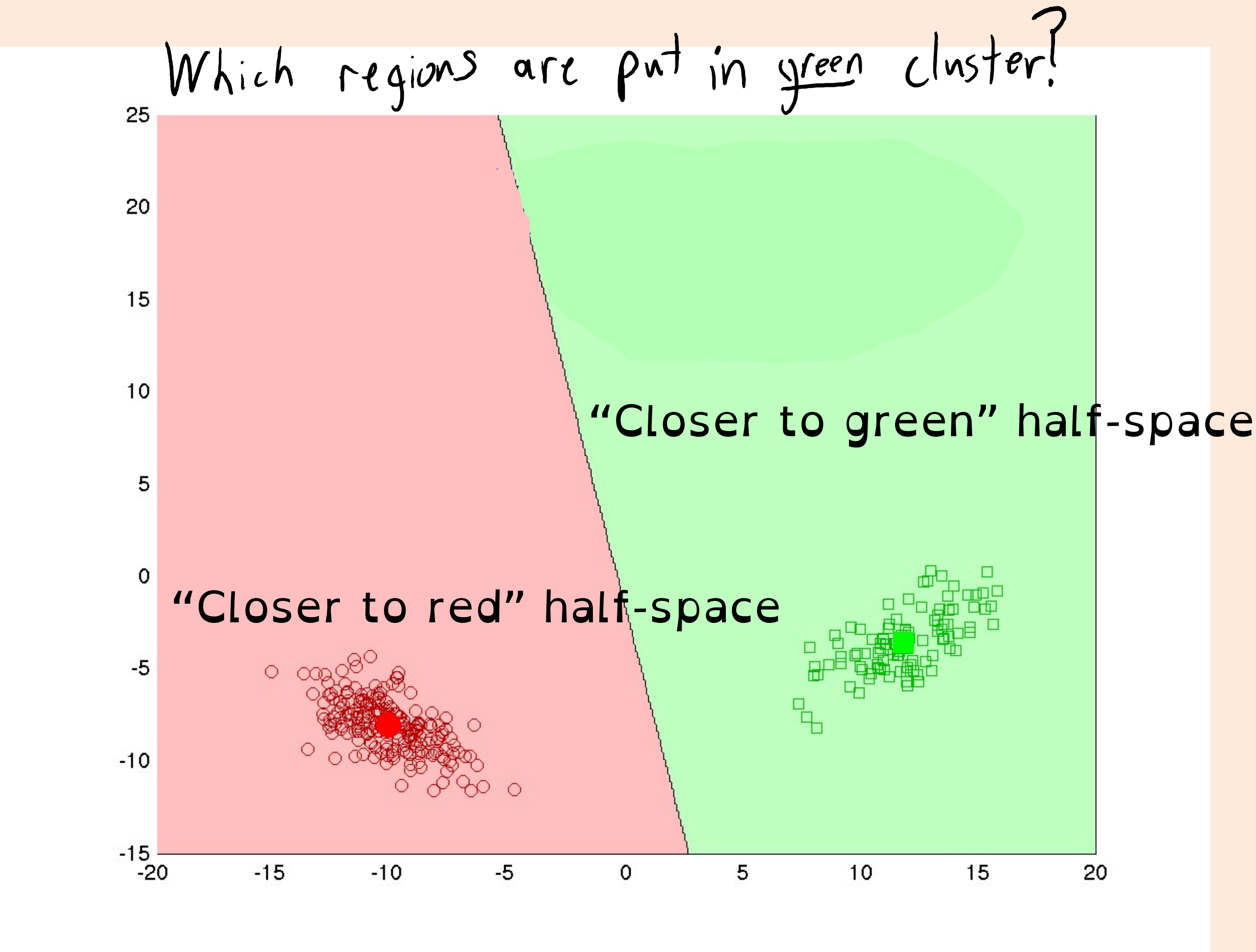


Why are k-means clusters convex?

Which regions are put in green cluster?

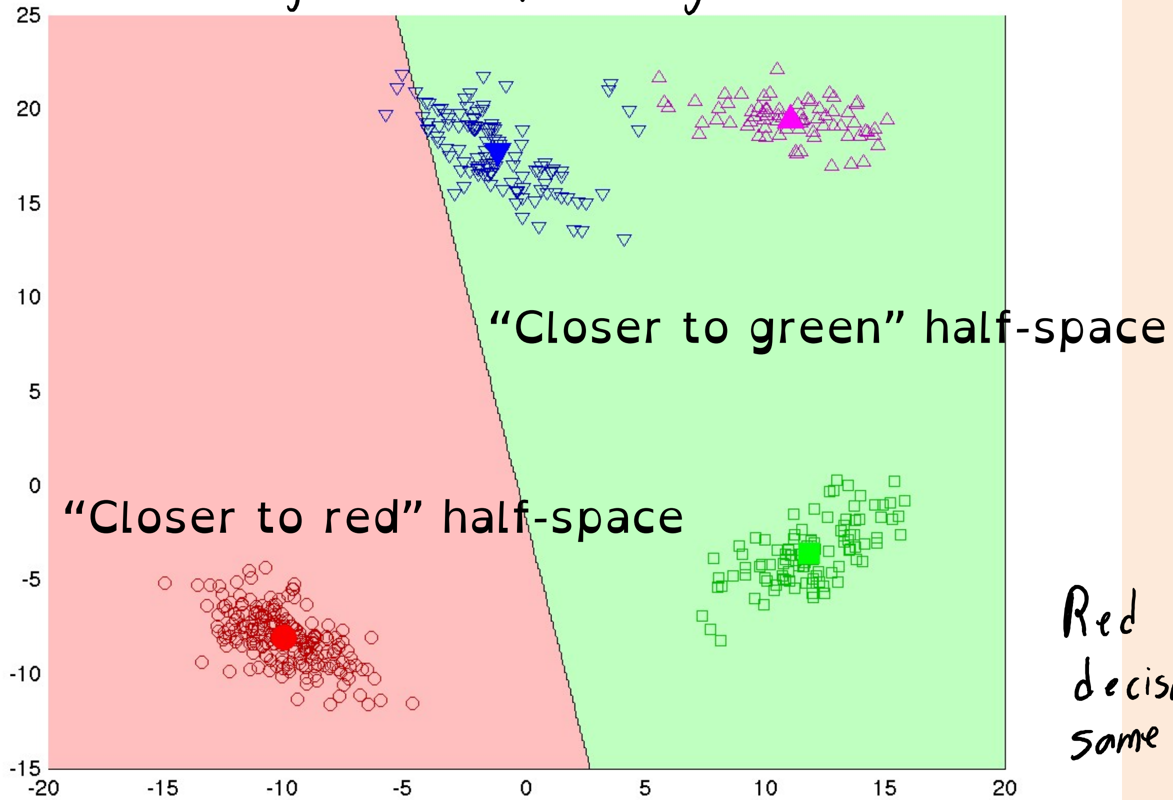


Why are k-means clusters convex?



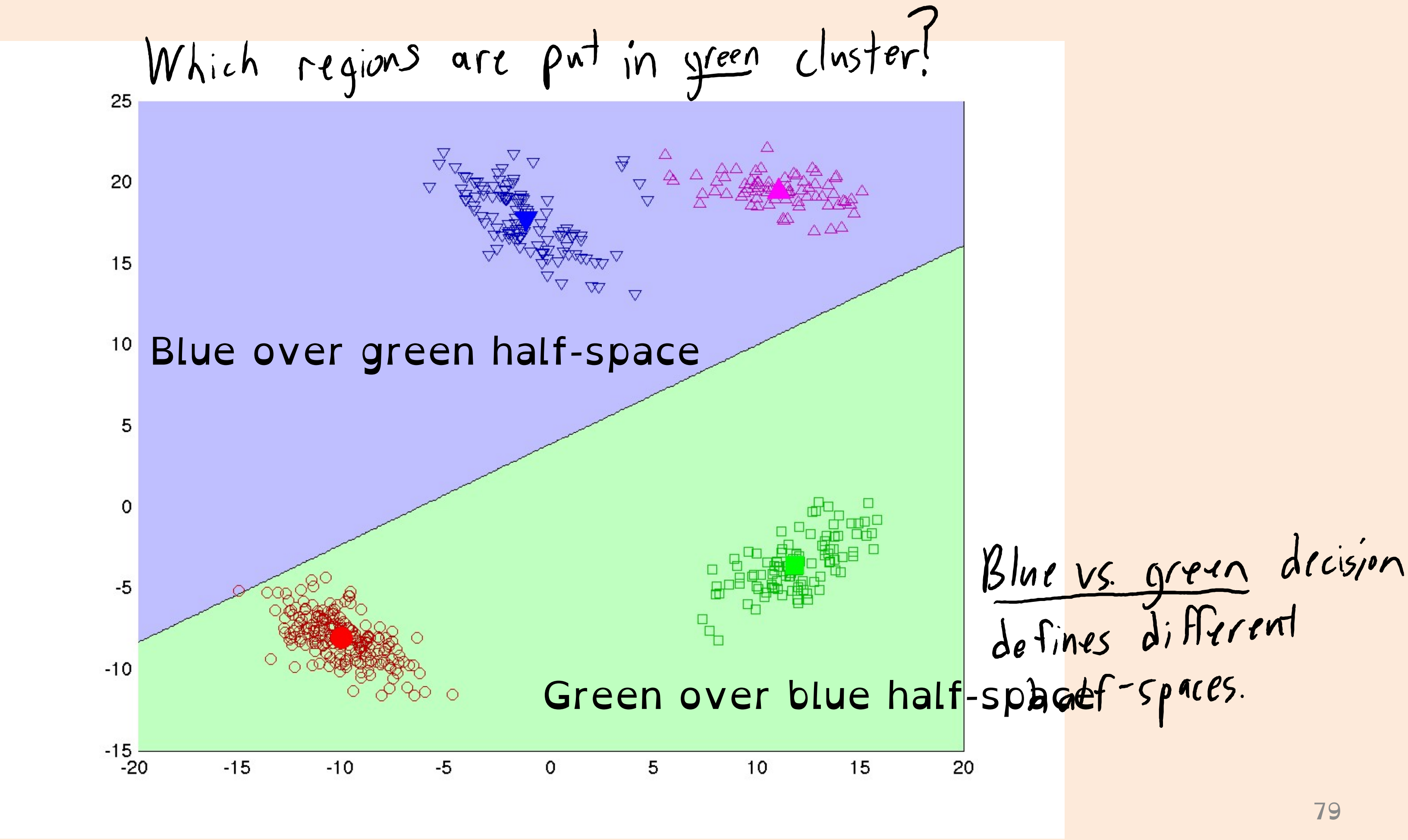
Why are k-means clusters convex?

Which regions are put in green cluster?



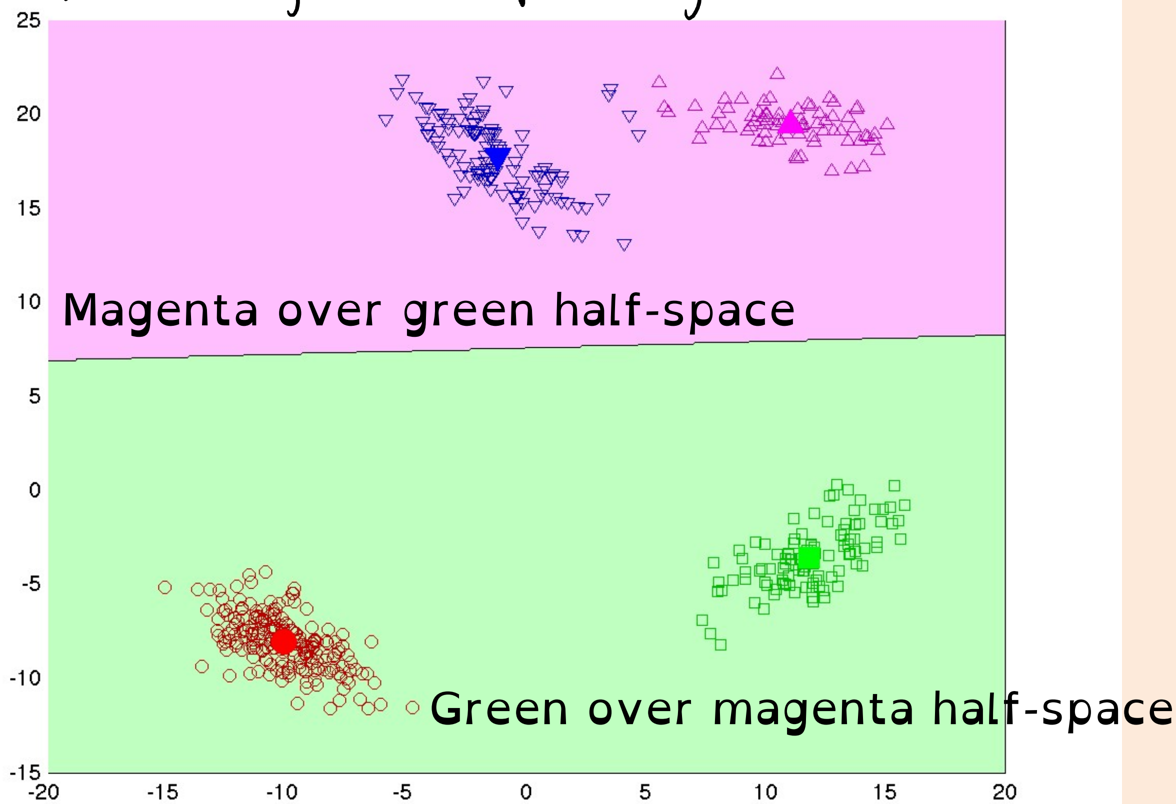
Red vs. green decision stays the same with more clusters.

Why are k-means clusters convex?

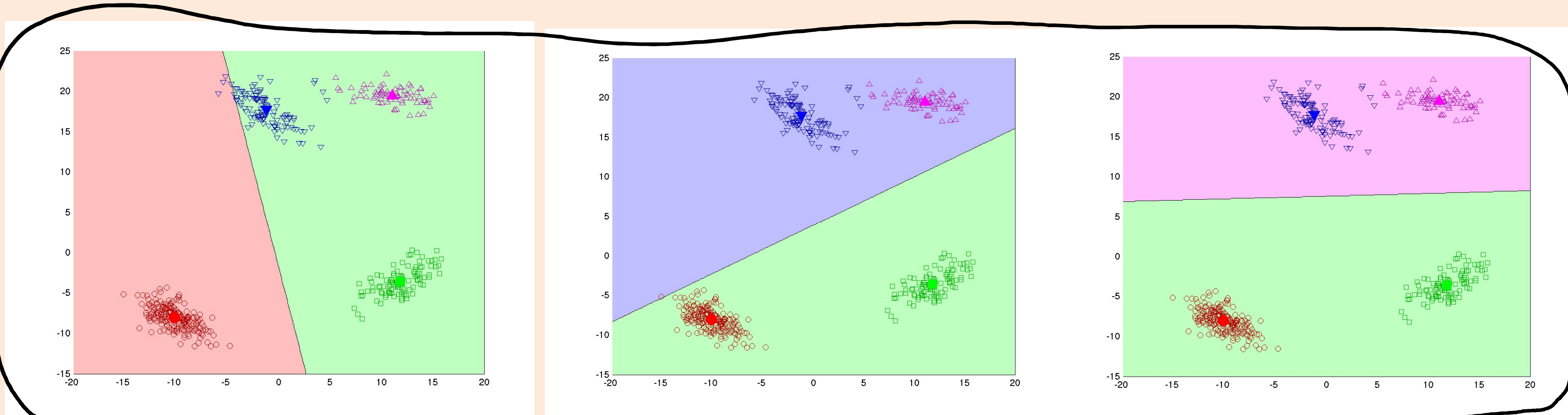


Why are k-means clusters convex?

Which regions are put in green cluster?

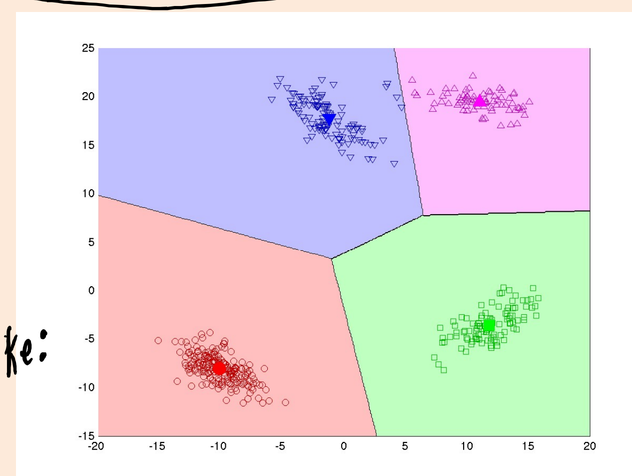


Why are k-means clusters convex?



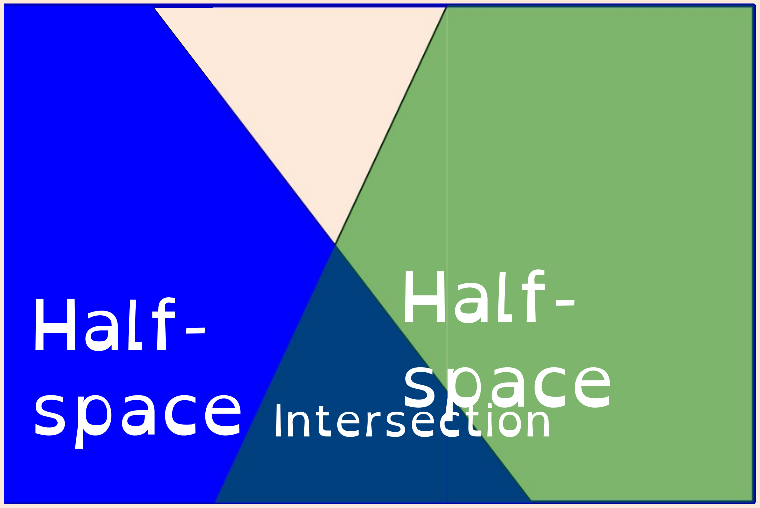
Green "cluster" is the intersection of these three half-spaces.

Here is what the four clusters look like:



Why are k-means clusters convex?

- Half-spaces are convex sets.
- Intersection of convex sets is a convex set.
  - Line segment between points in each set are still in each set.
- So intersection of half-spaces is convex.



Why are k-means clusters convex?

- Formal proof that "cluster 1" is convex (works for other clusters).

Let x_i and x_j be arbitrariy points in cluster 1.

→ By def'n of cluster 1, $\|x_i - w_1\| \leq \|x_i - w_c\|$ for all 'c'
$\|x_j - w_1\| \leq \|x_j - w_c\|$ for all 'c' } equality for $c = 1$

→ Let x_m be an arbitrariy point between x_i and x_j.

→ So we can write it as $x_m = \theta x_i + (1-\theta)x_j$ for some $\theta \in [0,1]$

Then $\|x_m - w_1\| = \|\theta x_i + (1-\theta)x_j - (\theta w_1 + (1-\theta)w_1)\|$ $\quad (w_1 = \theta w_1 + (1-\theta)w_1)$

$\leq \|\theta x_i - \theta w_1\| + \|(1-\theta)x_j - (1-\theta)w_1\|$ (triangle inequality)

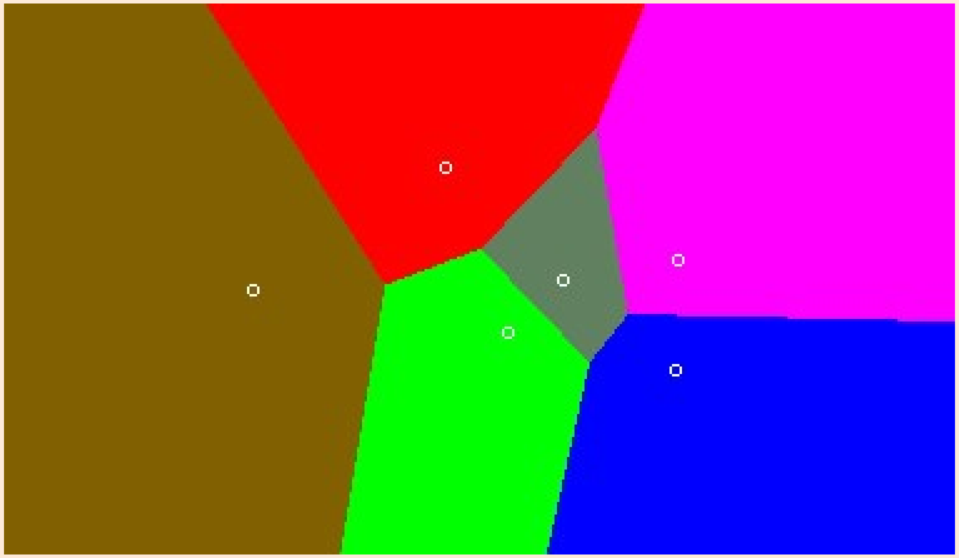
$= \theta\|x_i - w_1\| + (1-\theta)\|x_j - w_1\|$ (homogenity of norms)

x_i and x_j are in cluster 1 {

$\leq \theta\|x_i - w_c\| + (1-\theta)\|x_j - w_c\| = \|x_j - w_c\|$ so x_m is in cluster 1.

Voronoi Diagrams

- The k-means partition can be visualized as a Voronoi diagram:



- Can be a useful visualization of "nearest available" problems.
  - E.g., nearest tube station in London.

http://datagenetics.com/blog/may12017/index.html

Density-Based Clustering Runtime

question ☆ stop following **72 views**

Actions

DBSCAN Training time & Testing time

This is a follow-up inquiry post with Mike about the DBScan, would like to know:

1. Training runtime of DBScan, under k iterations (training set X has n examples and d features)
2. Testing runtime for a single example in DBScan; Testing runtime for test set of size t in DBScan,

the instructors' answer, *where instructors collectively construct a single answer*

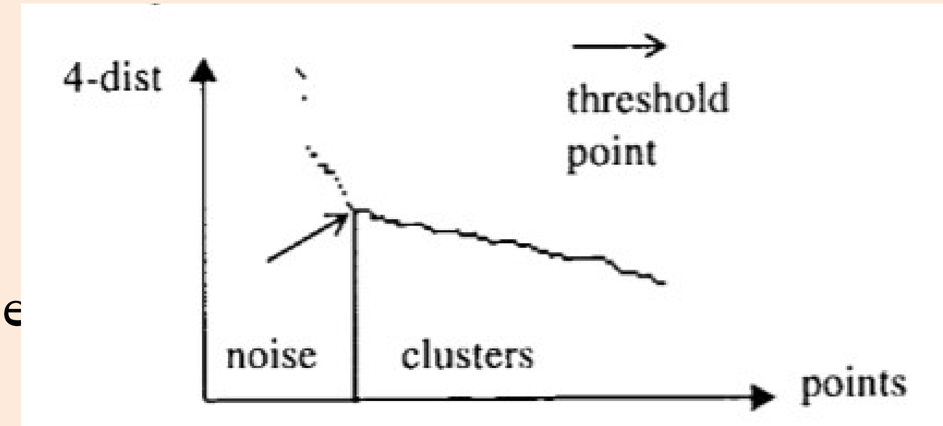
For training, you'll check that each point is a core point exactly once. This check costs $O(nd)$ since you measure the distance to each other point, leading to a total training cost of $O(n^2d)$.

(There are ways to speed this up, like grid-based pruning.)

We didn't define how to apply the DBSCAN model to test data. But a plausible way is to test if the new point is a neighbor of any existing core points. If you have m core points, you would be able to do this in $O(md)$.

“Elbow” Method for Density-Based Clustering

- From the original DBSCAN paper:
  - Choose some ‘k’ (they suggest 4) and set minNeighbours=k.
  - Compute distance of each points to its ‘k’ nearest neighbours.
  - Sort the points based on these distances and plot the distances:

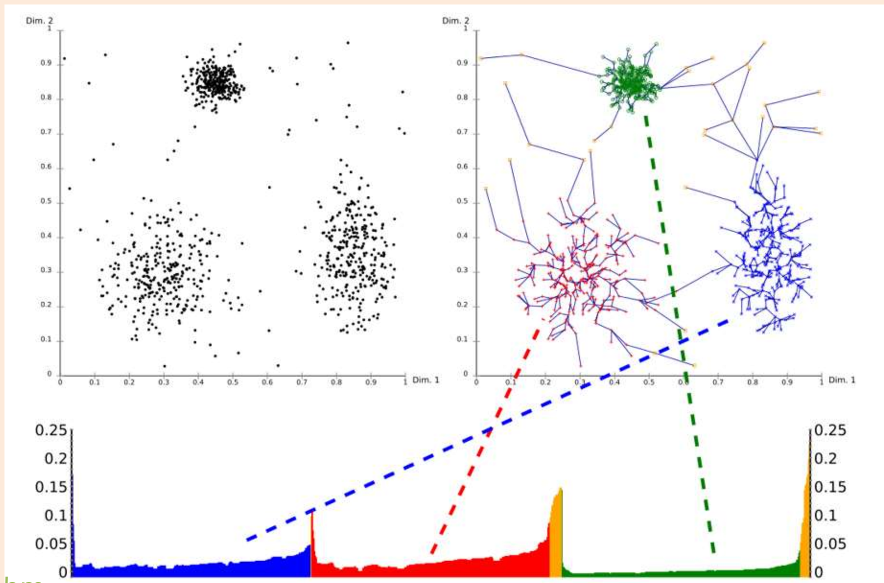


  - Look for an “e

https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

OPTICS

- Related to the DBSCAN “elbow” is “OPTICS”.
  - Sort the points so that neighbours are close to each other in the ordering.
  - Plot the distance from each point to the next point.
  - Clusters should correspond to sequencers with low distance.



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

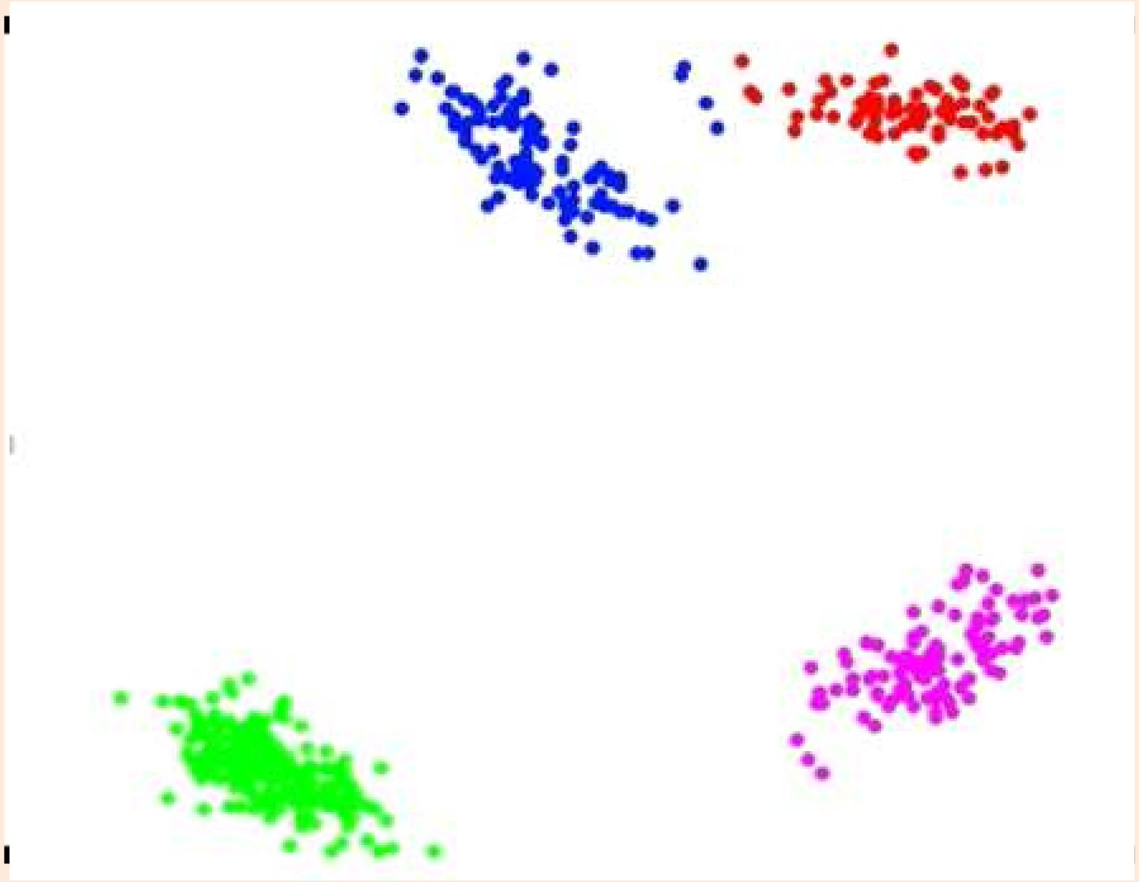
UBClustering Algorithm

- Let's define a new ensemble clustering method: UBClustering.

1. Run k-means with 'm' different random initializations.
2. For each example i and j:
   - Count the number of times x_i and x_j are in the same cluster.
   - Define $p(i,j) = \text{count}(x_i \text{ in same cluster as } x_j)/m$.
3. Put x_i and x_j in the same cluster if $p(i,j) > 0.5$.

- Like DBSCAN merge clusters in step 3 if i or j are already assigned.
  - You can implement this with a DBSCAN code (just changes "distance").
  - Each x_i has an x_j in its cluster with $p(i,j) > 0.5$.
  - Some points are not assigned to any cluster.

UBClustering Algorithm



It looks like DBSCAN, but far-away points will be assigned to a cluster if they always appear in same cluster as other points.

Distances between Clusters

- Other choices of the distance between two clusters:
  - "Single-link": minimum distance between points in clusters.
  - "Average-link": average distance between points in clusters.
  - "Complete-link": maximum distance between points in clusters.
  - Ward's method: minimize within-cluster variance.
  - "Centroid-link": distance between a representative point in the cluster.
    - Useful for distance measures on non-Euclidean spaces (like Jaccard similarity).
    - "Centroid" often defined as point in cluster minimizing average distance to other points.

Cost of Agglomerative Clustering

- One step of agglomerative clustering costs $O(n^2d)$:
  - We need to do the $O(d)$ distance calculation between up to $O(n^2)$ points.
  - This is assuming the standard distance functions.
- We do at most $O(n)$ steps:
  - Starting with 'n' clusters and merging 2 clusters on each step, after $O(n)$ steps we'll only have 1 cluster left (though typically it will be much smaller).
- This gives a total cost of $O(n^3d)$.

- This can be reduced to $O(n^2d \log n)$ with a priority queue:
  - Store distances in a sorted order, only update the distances that change.
- For single- and complete-linkage, you can get it down to $O(n^2d)$.
  - "SLINK" and "CLINK" algorithms.

Bonus Slide: Divisive (Top-Down) Clustering

- Start with all examples in one cluster, then start dividing.
- E.g., run k-means on a cluster, then run again on resulting clusters.
  - A clustering analogue of decision tree learning.