CPSC 340:
Machine Learning and Data Mining

More Clustering

Bonus slides

john snow - Google Sea

google.ca/search?safe=off&tbm=isch&q=john+snow&nfpr=1&sa=X&ved=0ahUKEwj9gsnWsLXWAhUJ8mMKHUAQBdQQvgUIkAIoAQ&biw=1440&bih=750&d

CPSC 540 | CPSC 340 | Old CPSC 340 | 340 TA | MLAPP | Calendar | My Page | 540 - Google Docs | Gmail - Inbox | Trips | Reviews | MLRG | Grants | Vball | Slack

Google

john snow

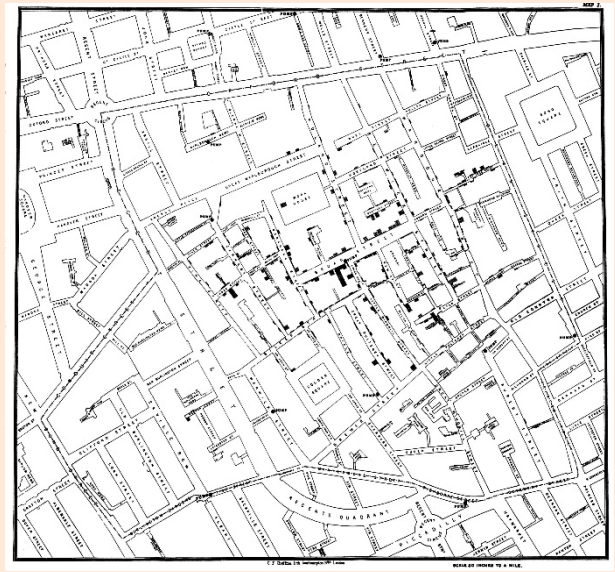
All | Images | News | Videos | Maps | More | Settings | Tools | View saved | SafeSearch

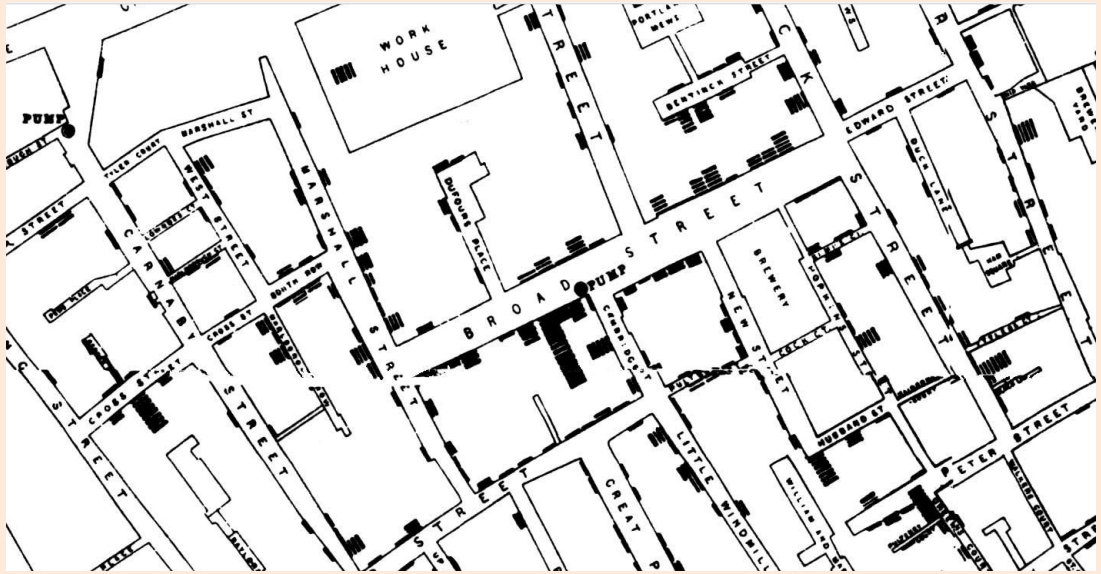
wallpaper | funny | season 6 | winter is coming meme | genderbent | let it snow | real life | ghost | wallpaper hd | the watch meme | his wolf | game throne | simpsons

Did you mean: *jon* snow

John Snow and Cholera Epidemic

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





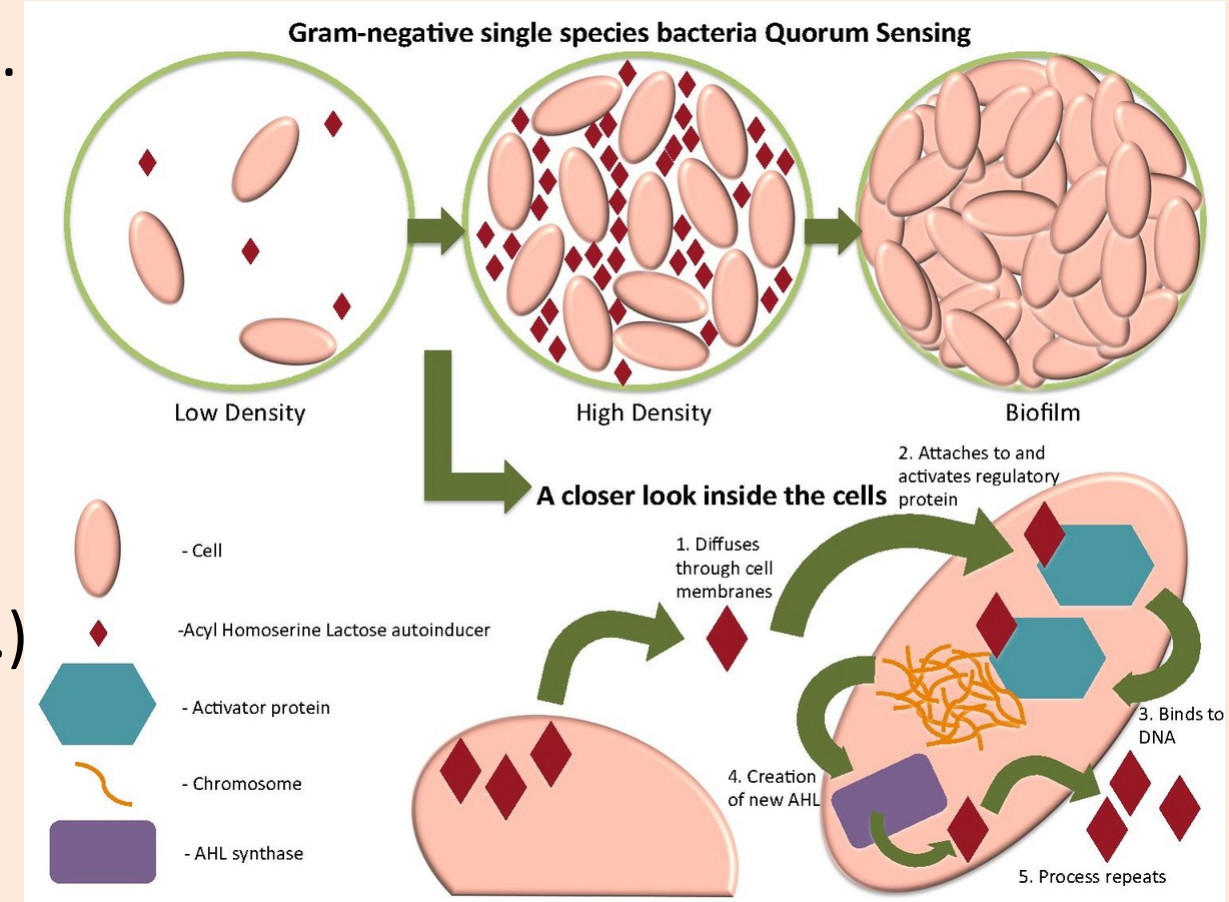
- Found cluster of cholera deaths around a particular water pump.
  - Went against airborne theory, but pump later found to be contaminated.
  - "Father" of epidemiology.

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

Density-Based Clustering in Nature

- Quorum sensing:
  - Bacteria continuously release a particular molecule.
  - They 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”.)



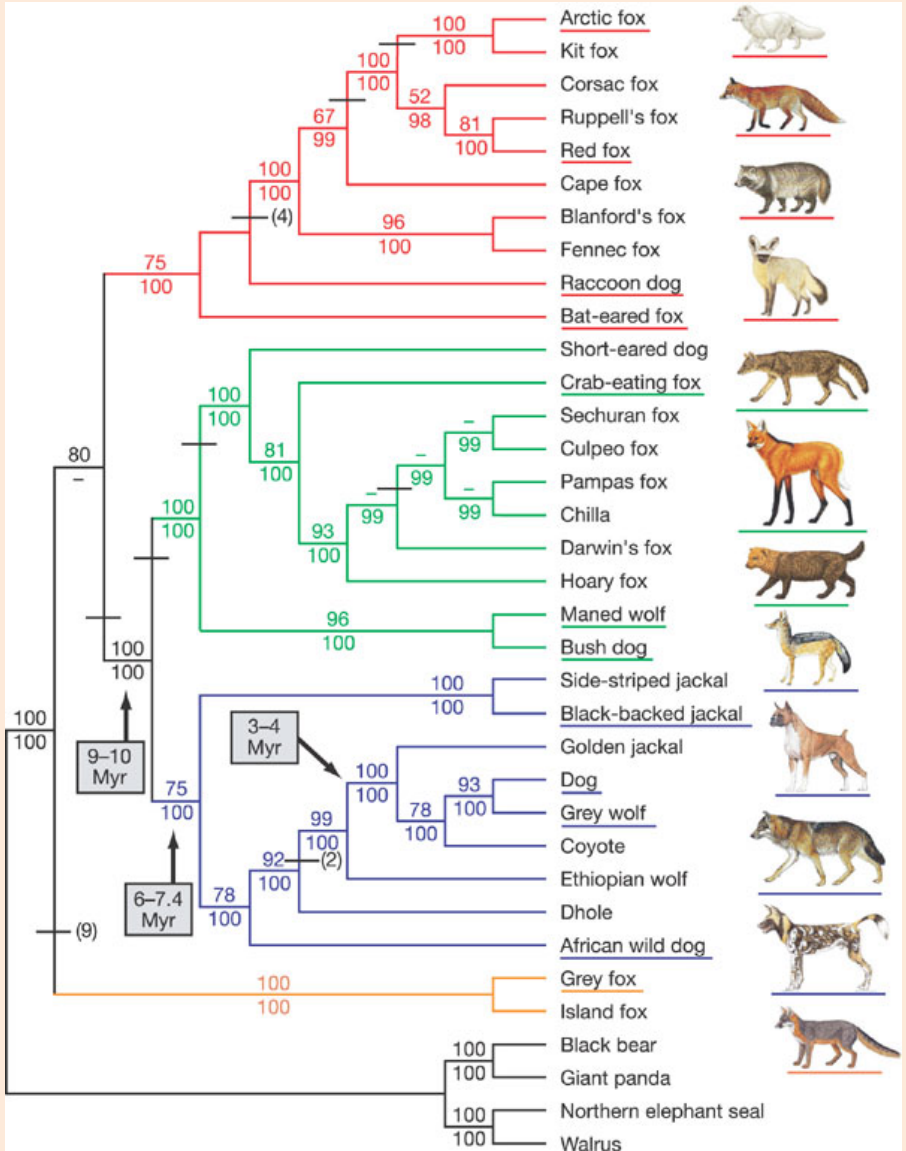
https://en.wikipedia.org/wiki/Quorum_sensing

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’.

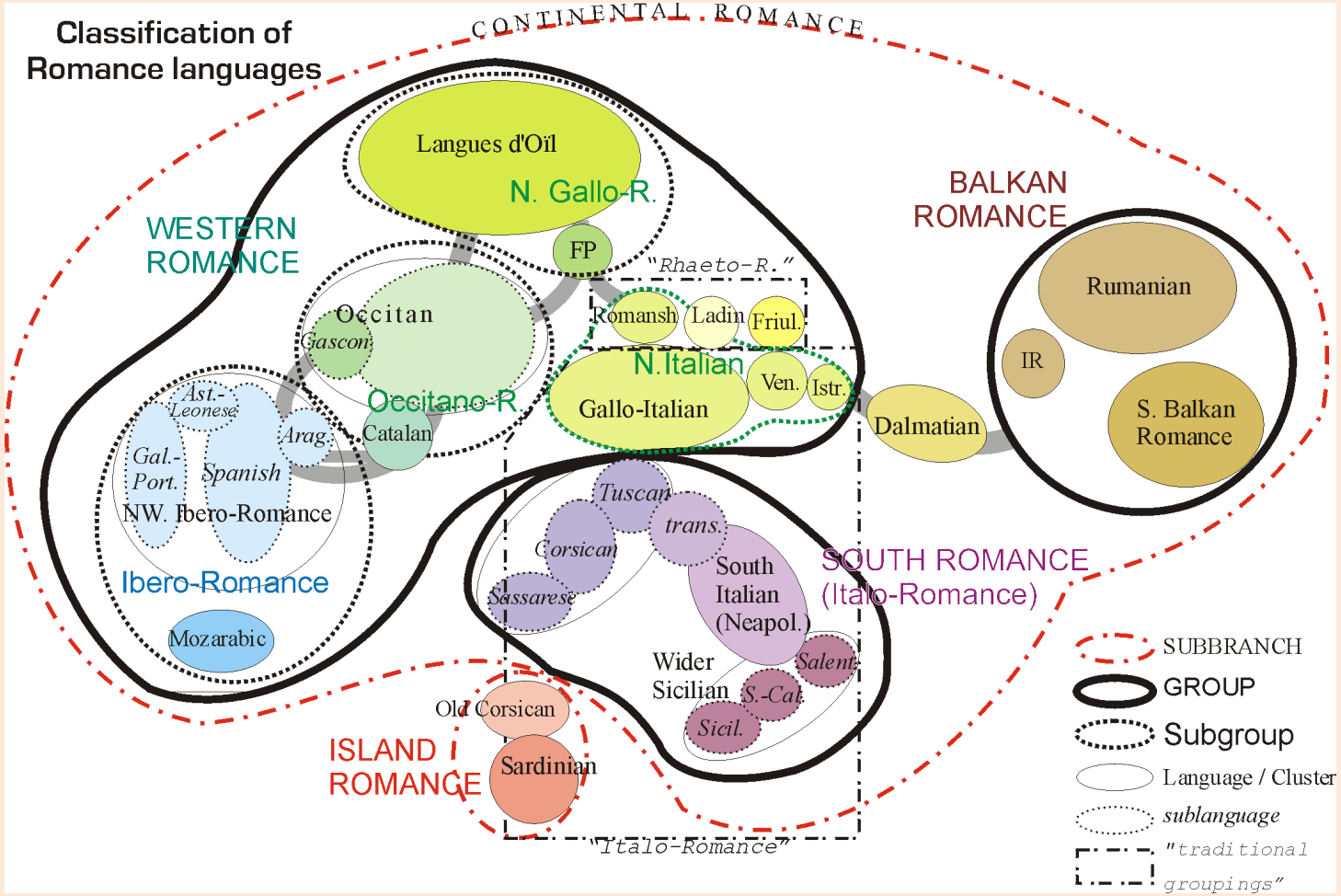
- More useful information in the plot:
  - Line lengths: chosen here to approximate time.
  - Numbers: #clusterings across bootstrap samples.
  - ‘Outgroups’ (walrus, panda) are a sanity check.



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

Application: Phylogenetics

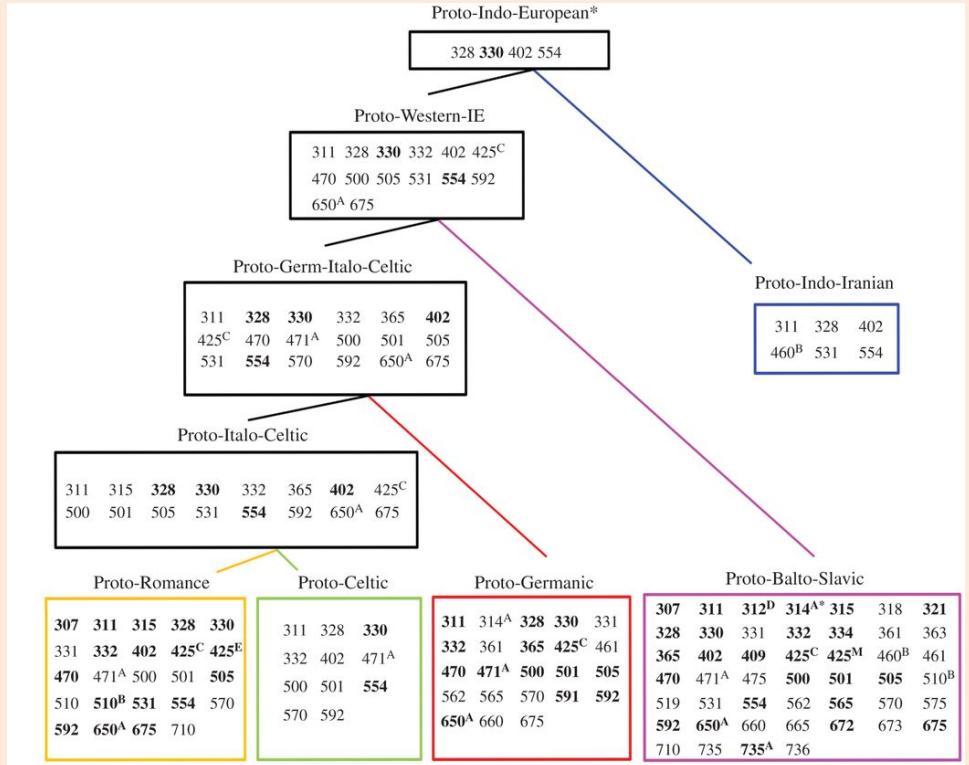
- Comparative method in linguistics studies evolution of languages:



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.



International tale types

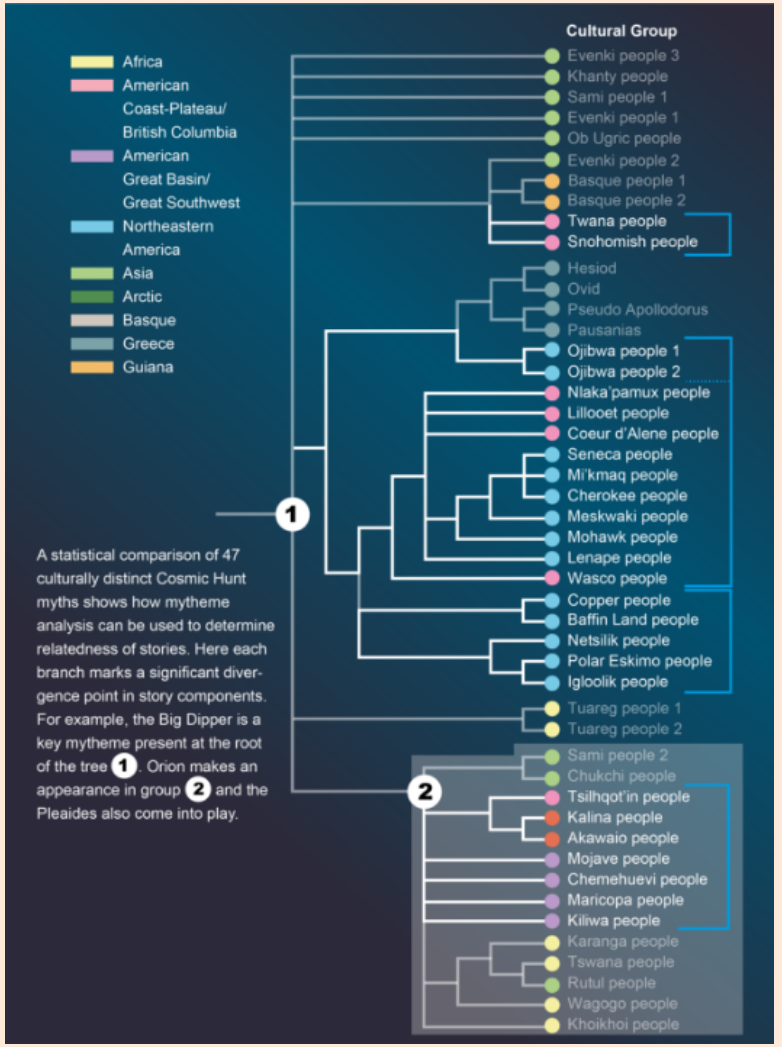
307	The Princess in the Coffin	**409**	The Girl as Wolf	**562**	The Spirit in the Blue Light
311	Rescue by Sister	**425C**	Beauty and the Beast	**565**	The Magic Mill
312D	Rescue by the Brother	**425E**	The Enchanted Husband	**570**	The Rabbit-Herd
314A	The Shepherd and the Giants	**425M**	The Snake Bridegroom	**575**	The Prince's Wings
314A*	Animal Helper in the Flight	**460B**	The Journey	**591**	The Thieving Pot
315	The Faithless Sister	**461**	Three Hairs	**592**	The Dance Among Thorns
318	The Faithless Wife	**470**	Friends in Life and Death	**650A**	Strong John
321	Eyes Recovered from Witch	**471A**	The Monk and the Bird	**660**	The Three Doctors
328	The Boy Steals Ogre's Treasure	**475**	The Man as the Heater	**665**	The Man who Flew and Swam
330	The Smith and the Devil	**500**	Supernatural Helper	**672**	The Serpent's Crown
331	The Spirit in the Bottle	**501**	The Three Old Spinning Women	**673**	The White Serpent's Flesh
332	Godfather Death	**505**	The Grateful Dead	**675**	The Lazy Boy
334	Household of the Witch	**510**	Cinderella and Peau d'Âne	**710**	Our Lady's Child
361	Bear Skin	**510B**	Peau d'Asne	**735**	The Rich and the Poor Man
363	The Corpse-Eater	**519**	The Strong Woman as Bride	**735A**	Bad Luck Imprisoned
365	The Dead Bridegroom	**531**	The Clever Horse	**736**	Luck and Wealth
402	The Animal Bride	**554**	The Grateful Animals		

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.

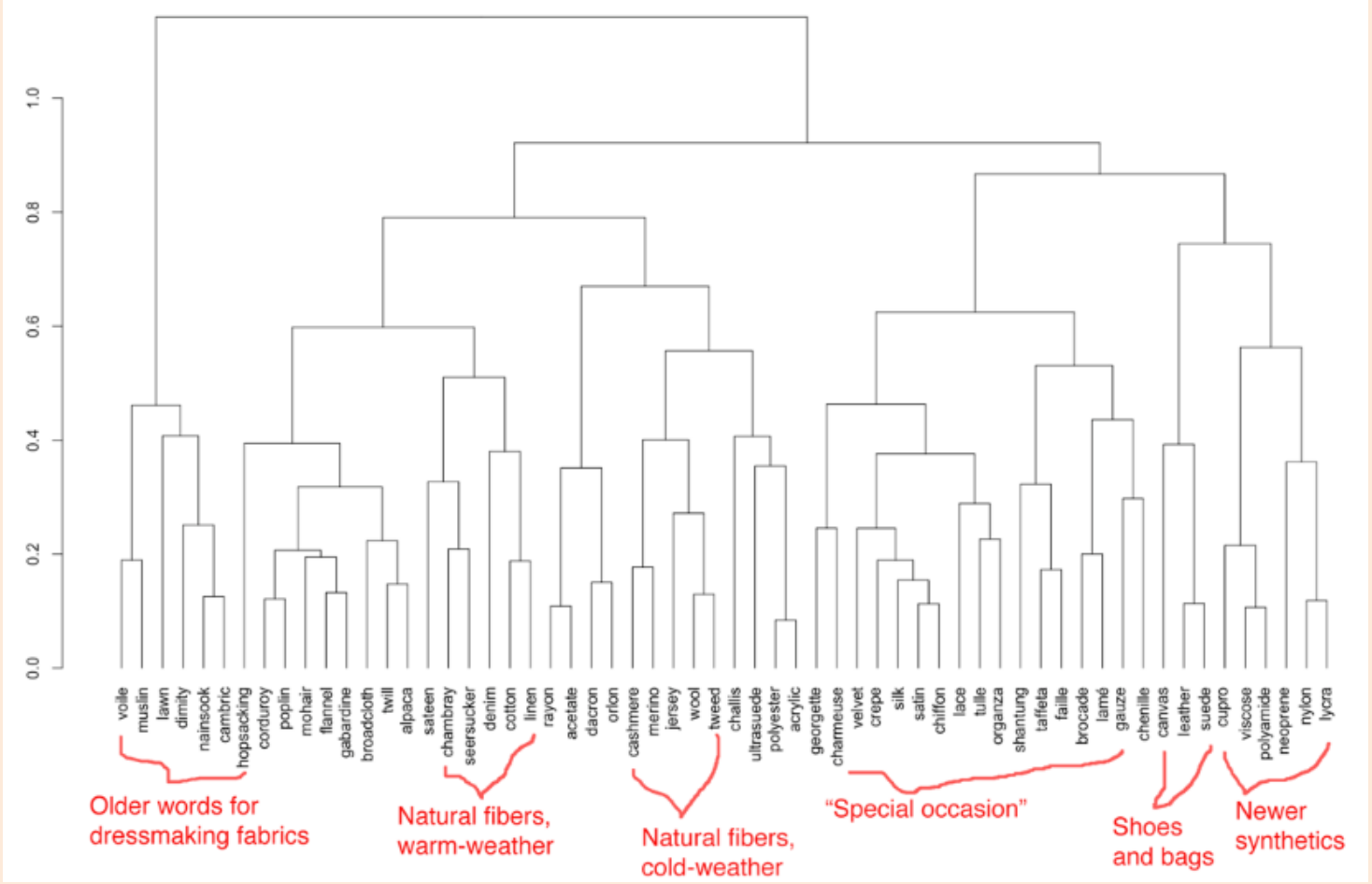
- September 2016: evolution of myths.
  - “Cosmic hunt” story:
    - Person hunts animal that becomes constellation.
      - Previously known to be at least 15,000 years old.
    - May go back to paleololithic period.



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

Application: Fashion?

- Hierarchical clustering of clothing material words in Vogue:

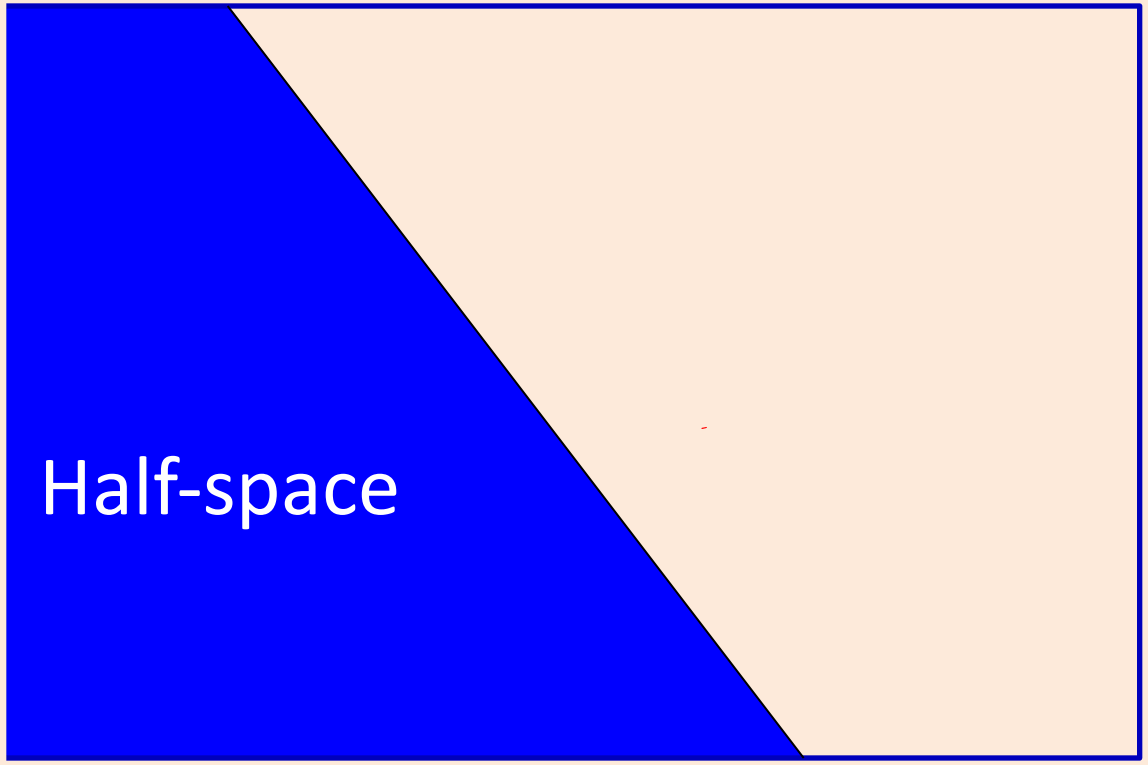


http://dh.library.yale.edu/projects/vogue/fabricspace/

Why are k-means clusters convex?

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

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

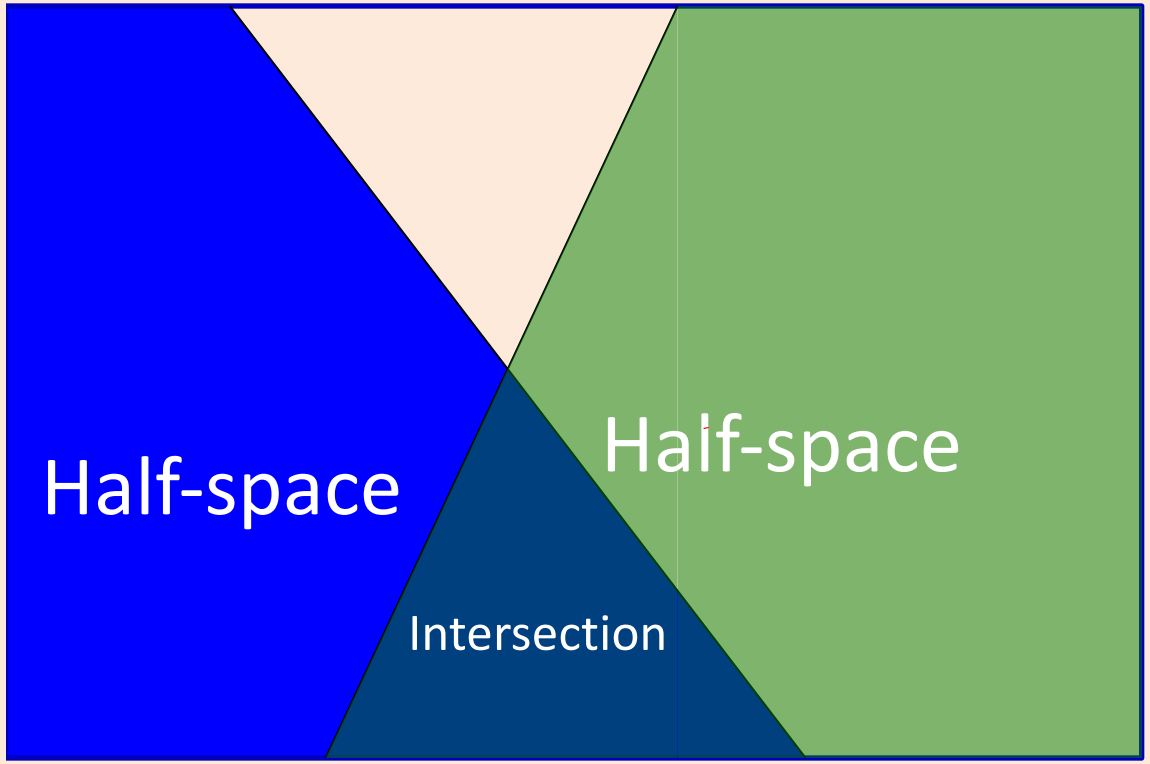


Why are k-means clusters convex?

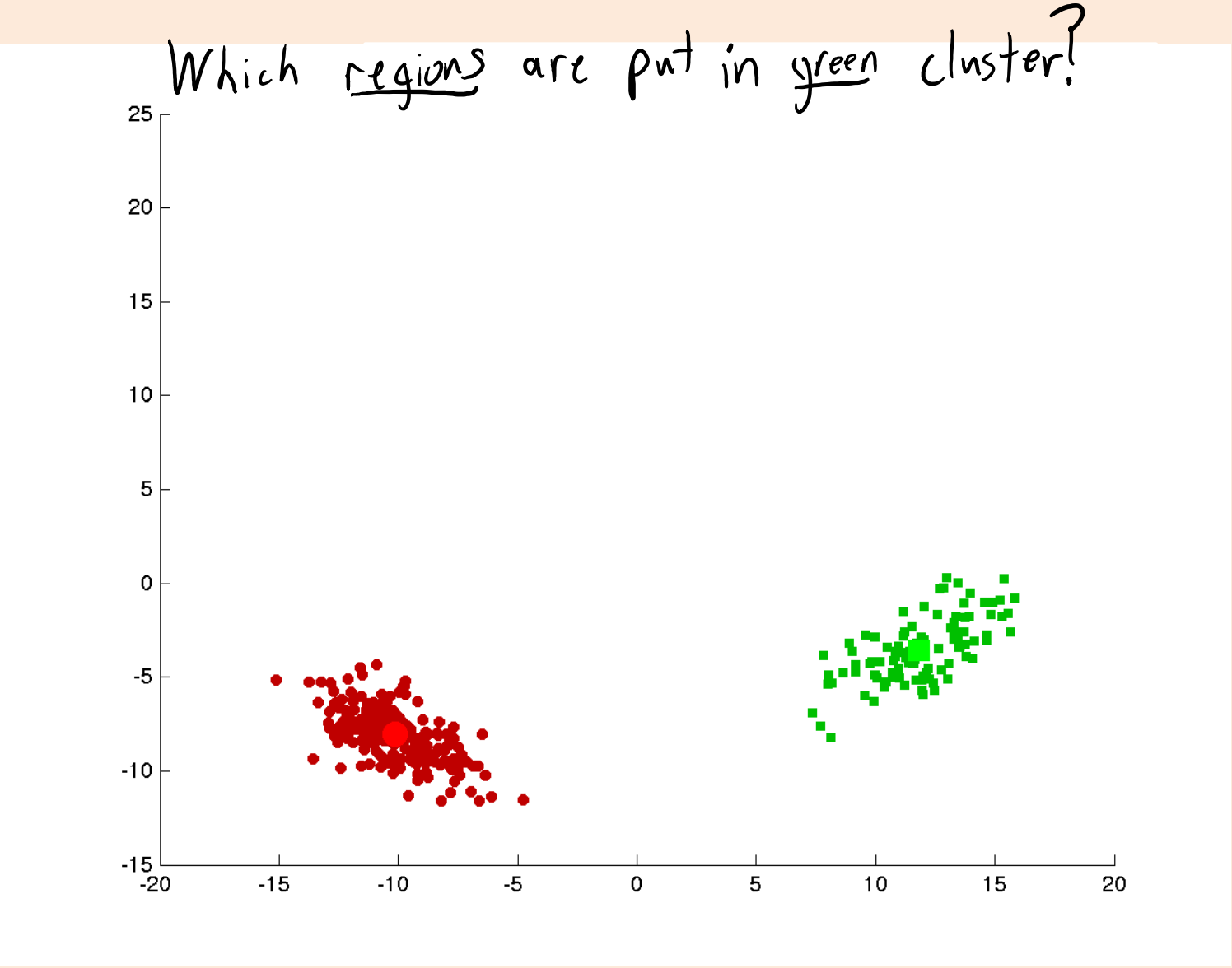
- K-means clusters are formed by the intersection of half-spaces.

Half-space is Set of points (satisfying a linear inequality, like $\sum_{j=1}^{d} a_j x_j \leq b$

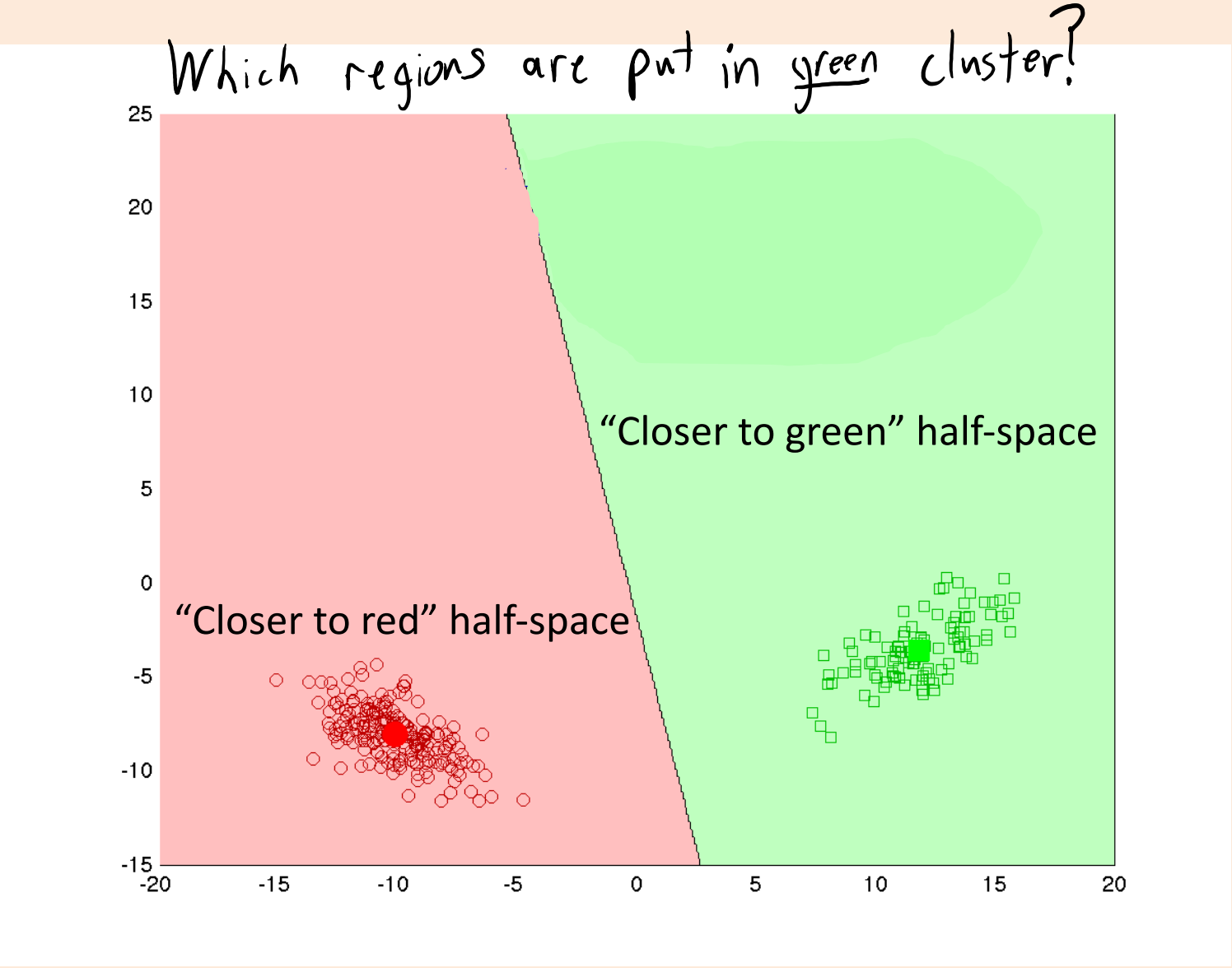
'x'



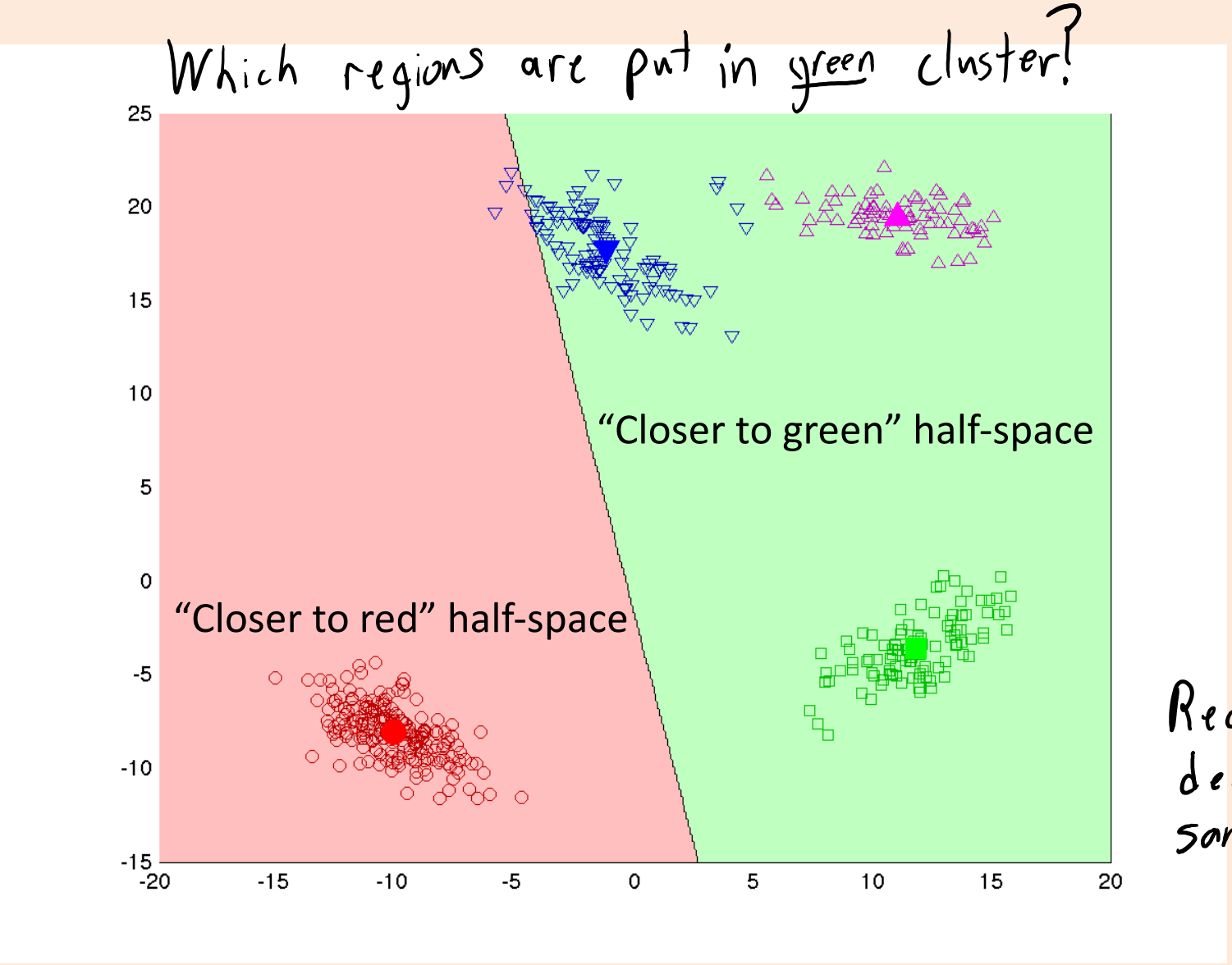
Why are k-means clusters convex?



Why are k-means clusters convex?

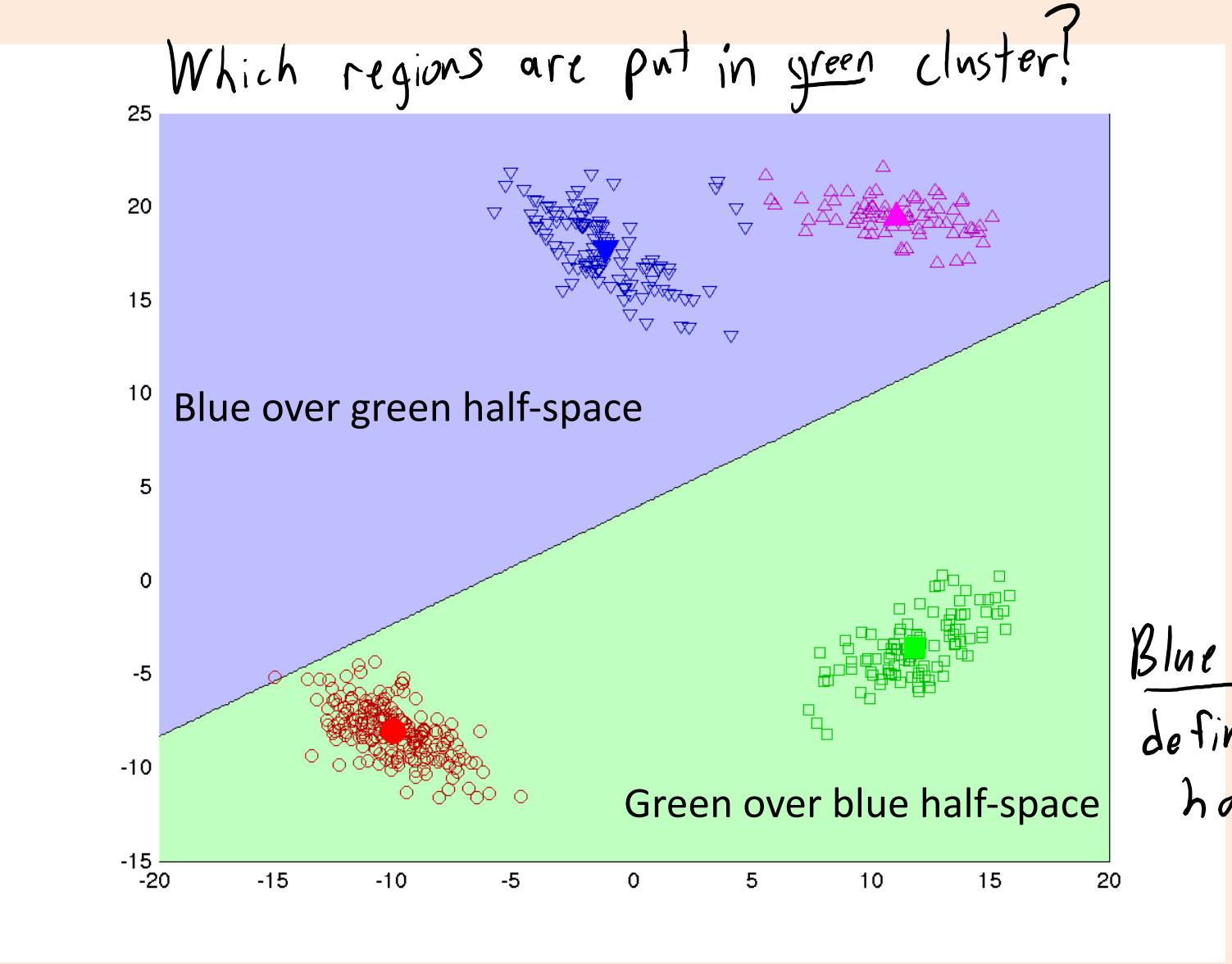


Why are k-means clusters convex?



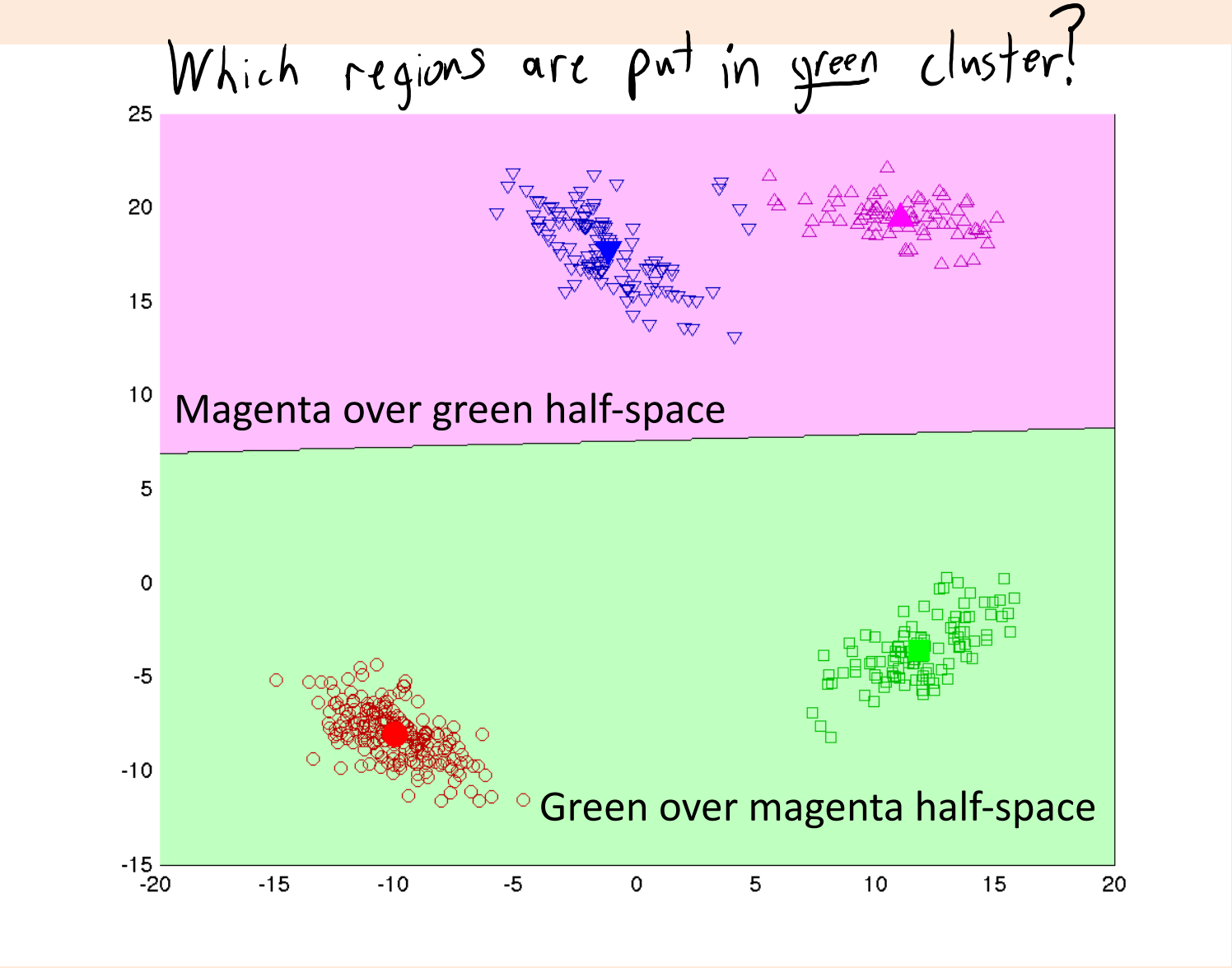
Red vs. green decision stays the same with more clusters.

Why are k-means clusters convex?

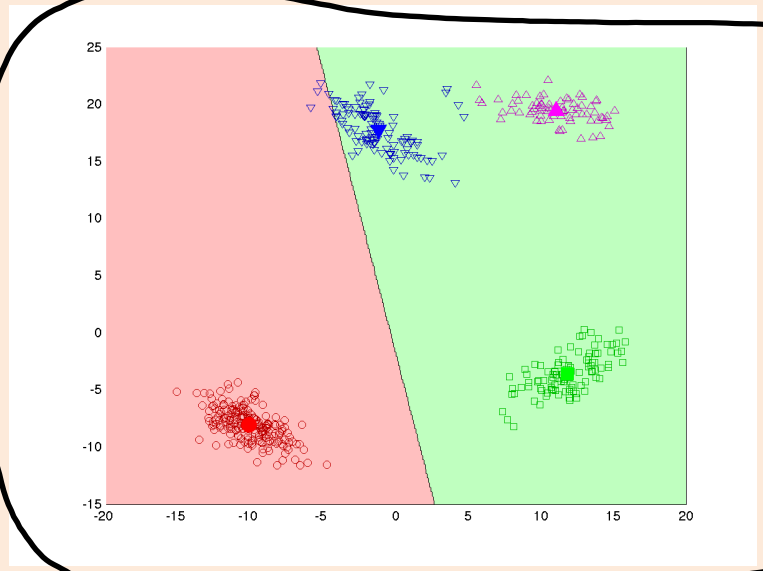


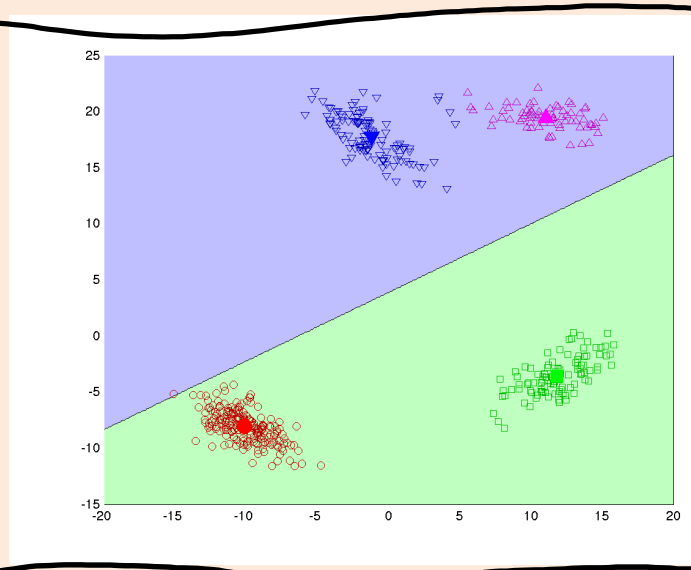
Blue vs. green decision defines different half-spaces.

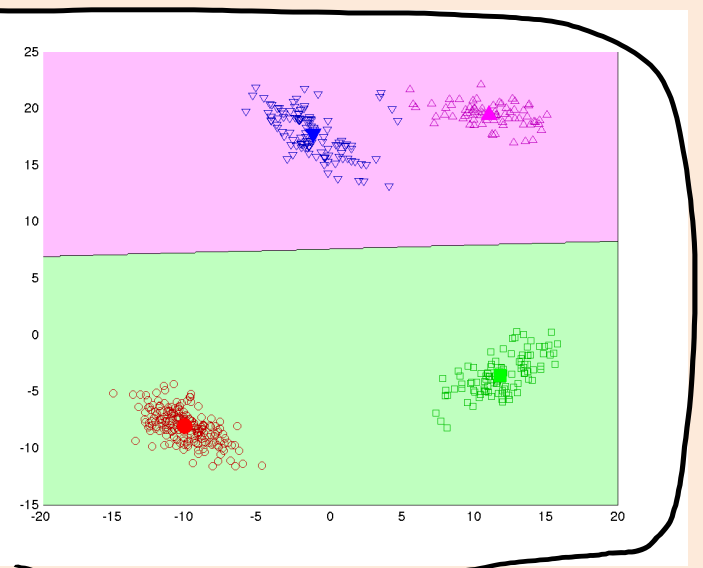
Why are k-means clusters convex?



Why are k-means clusters convex?

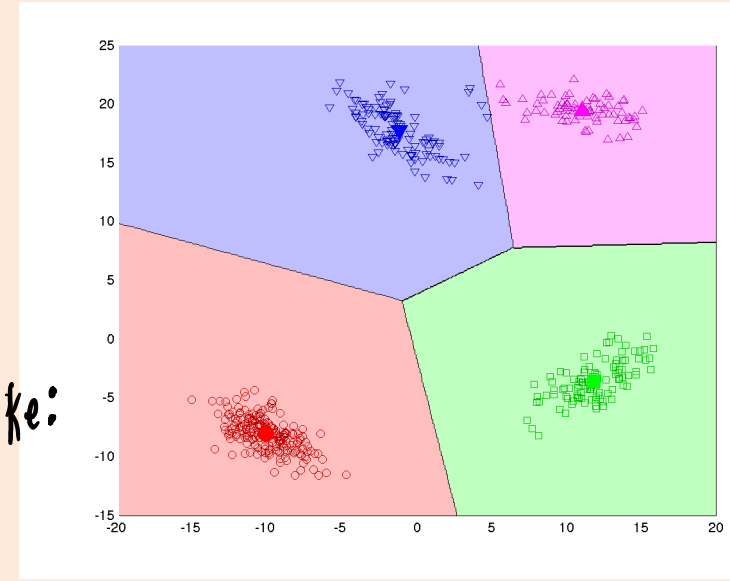






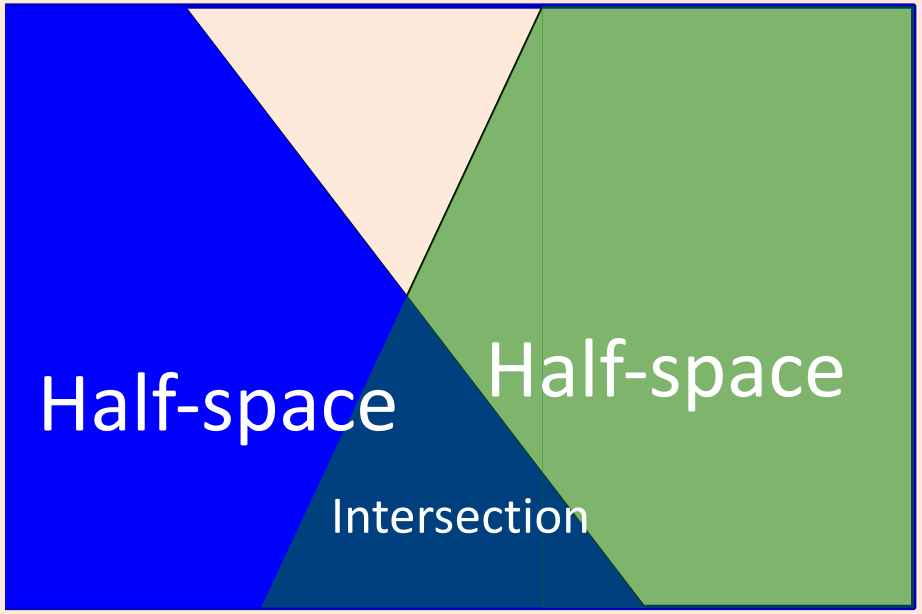
Green "cluster" is the intersection of these three half-spaces.

Here is what the four clusters look like:



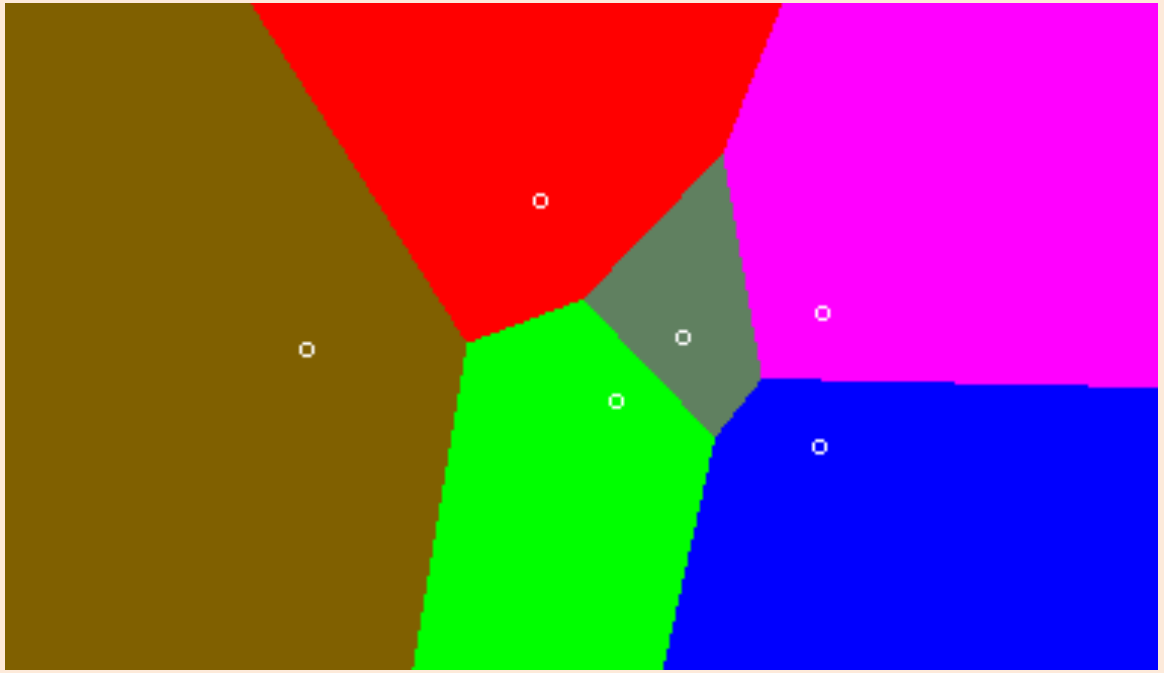
Why are k-means clusters convex?

- 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.



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*

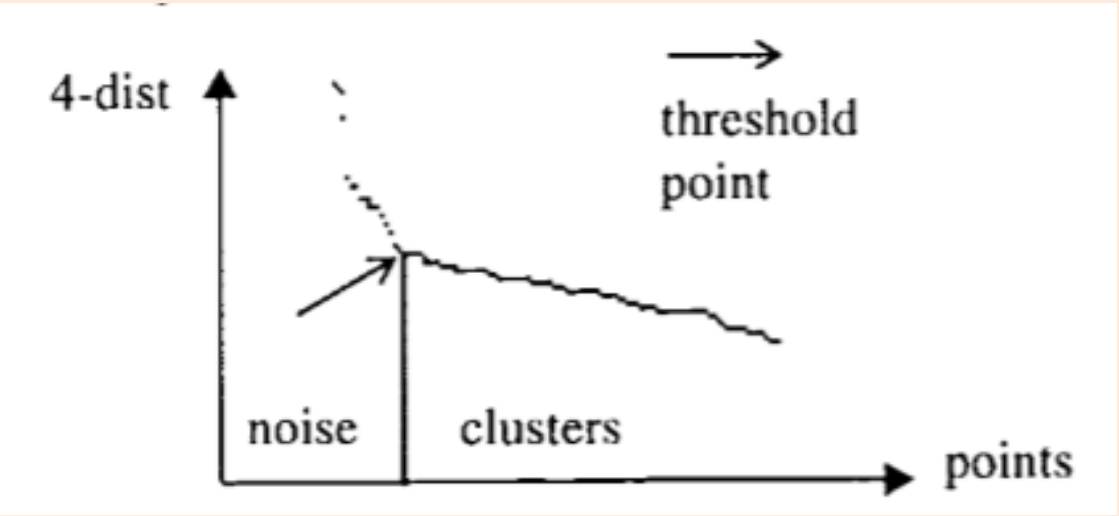
For training, you'll check that each point is a core point exactly once. This check costs $O(nd)$ since you measure the distance to each other point, leading to a total training cost of $O(n^2d)$.

(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:

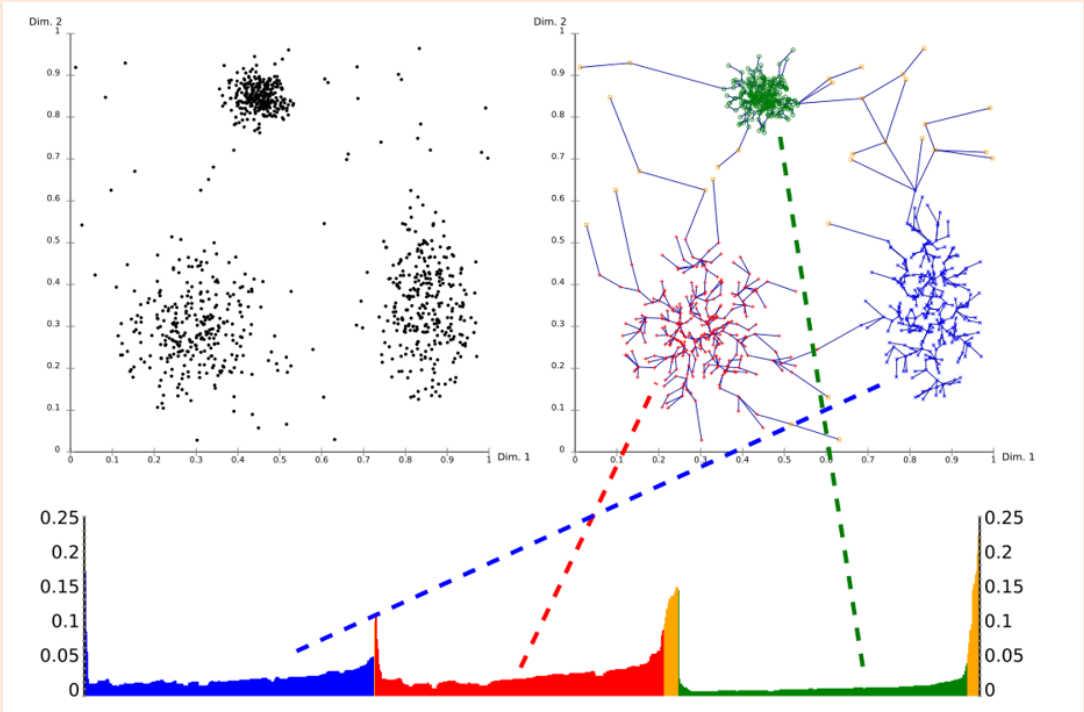


  - Look for an “elbow” to choose ϵ.

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

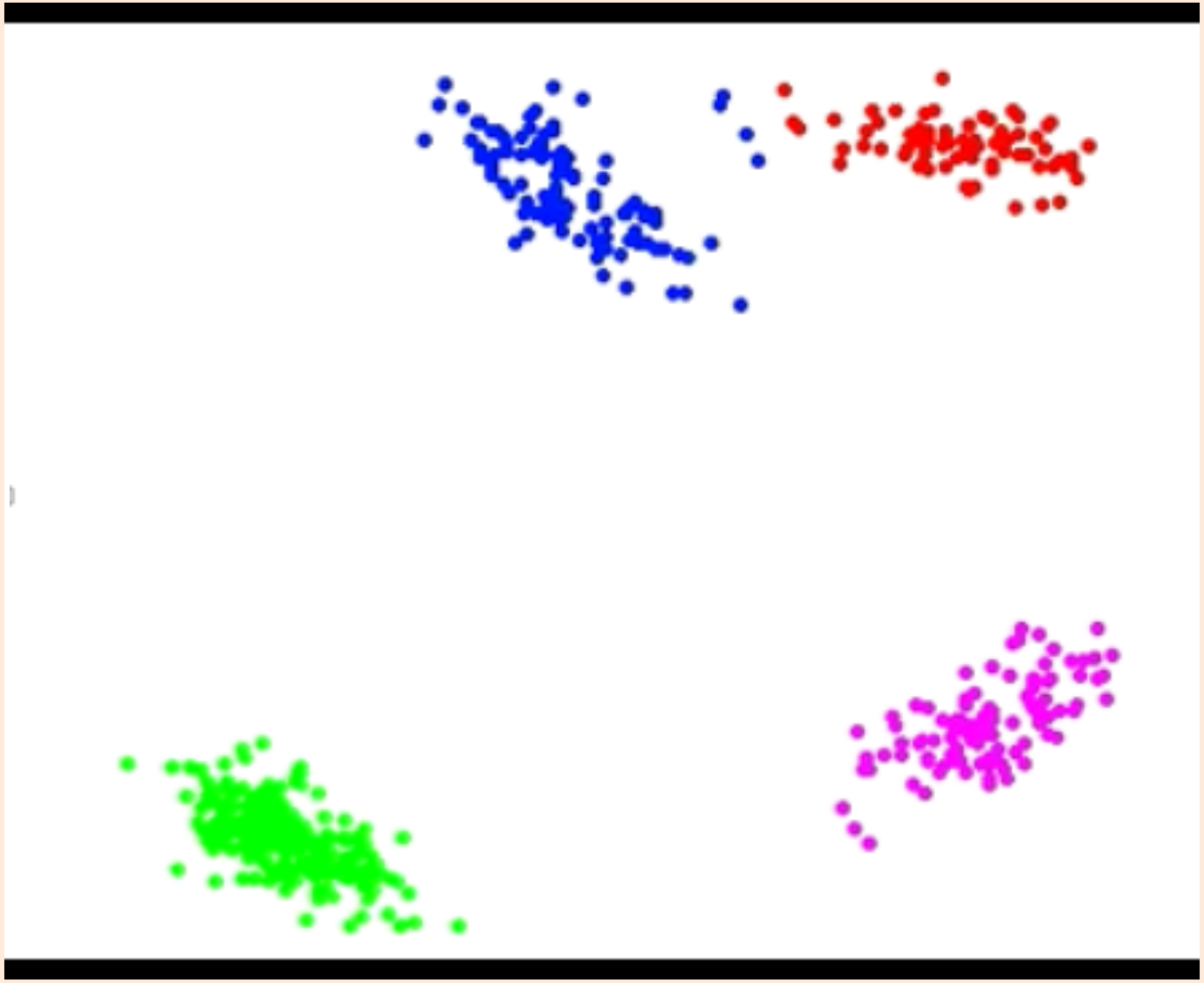
UBClustering Algorithm

- 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) = count(x_i 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.