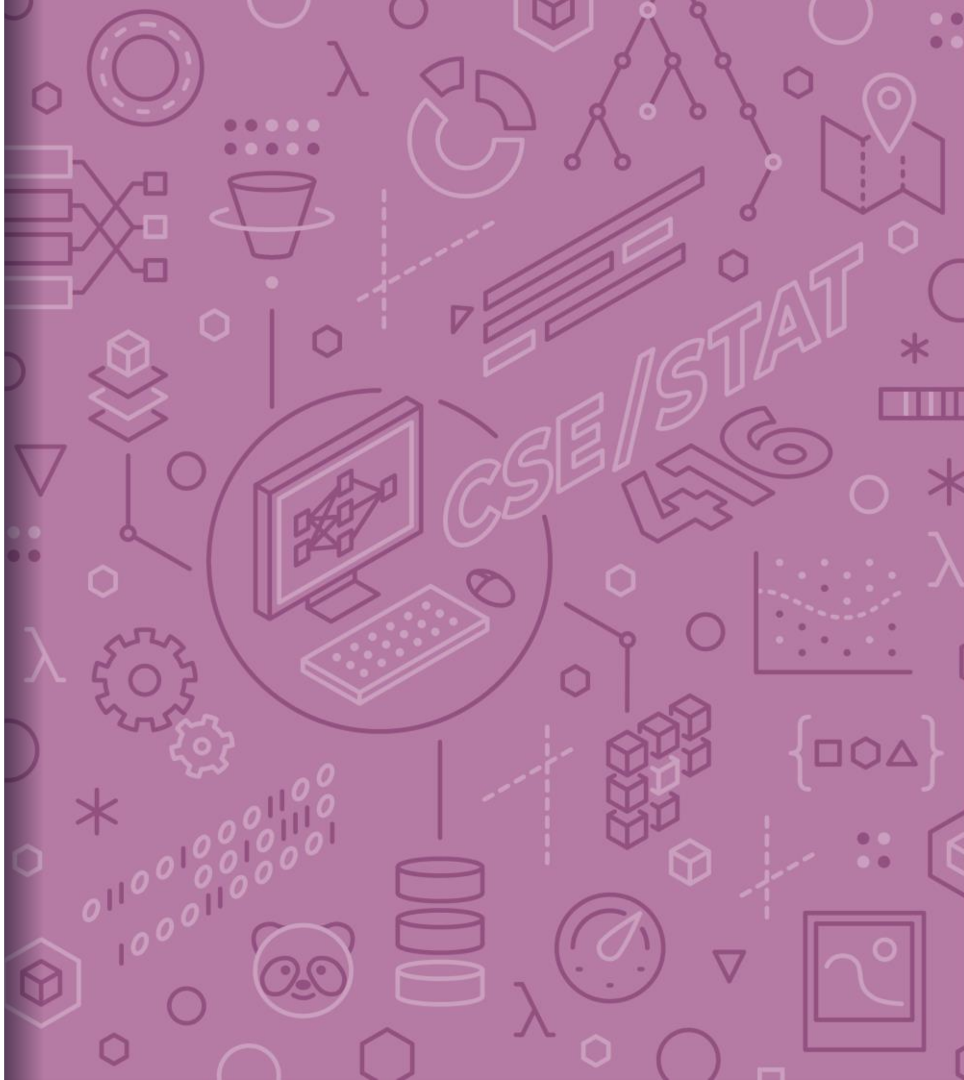
CSE/STAT 416

Other Clustering Methods

Amal Nanavati
University of Washington
Aug 3, 2022

Adapted from Hunter Schafer’s slides



Administrivia

- Last lecture on clustering!
- **Next Week**: Dimensionality Reduction, Recommender Systems
- **Next-Next Week**: Course Wrap-Up & Final
- Deadlines:
  - HW5 late deadline TOMORROW, Thurs 8/4 11:59PM
    - Uses two late days
    - Submit Concept & Programming on Gradescope
  - HW6 Released TODAY, due Tues 8/9 11:59PM
  - LR 7 due Fri 8/5 11:59PM
- Notes on the end of the quarter
  - HW7 due Tues 8/16, **NO LATE DAYS**
  - Take-Home Final Exam: Wed 8/17 9AM – Thurs 8/18 11:59PM

HW6 Walkthrough

Recap

K-Means Clustering

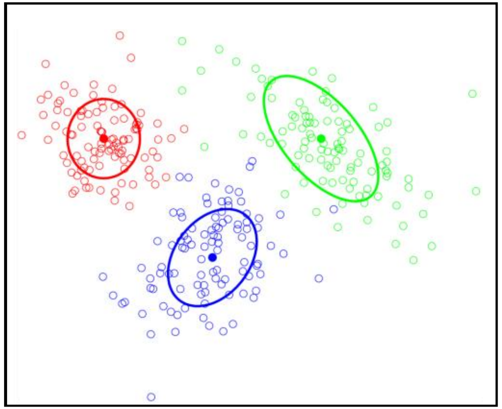
Define Clusters

In their simplest form, a **cluster** is defined by

- The location of its center (**centroid**)
- Shape and size of its **spread**

Clustering is the process of finding these clusters and **assigning** each example to a particular cluster.

- x_i gets assigned $z_i \in [1, 2, \ldots, k]$
- Usually based on closest centroid



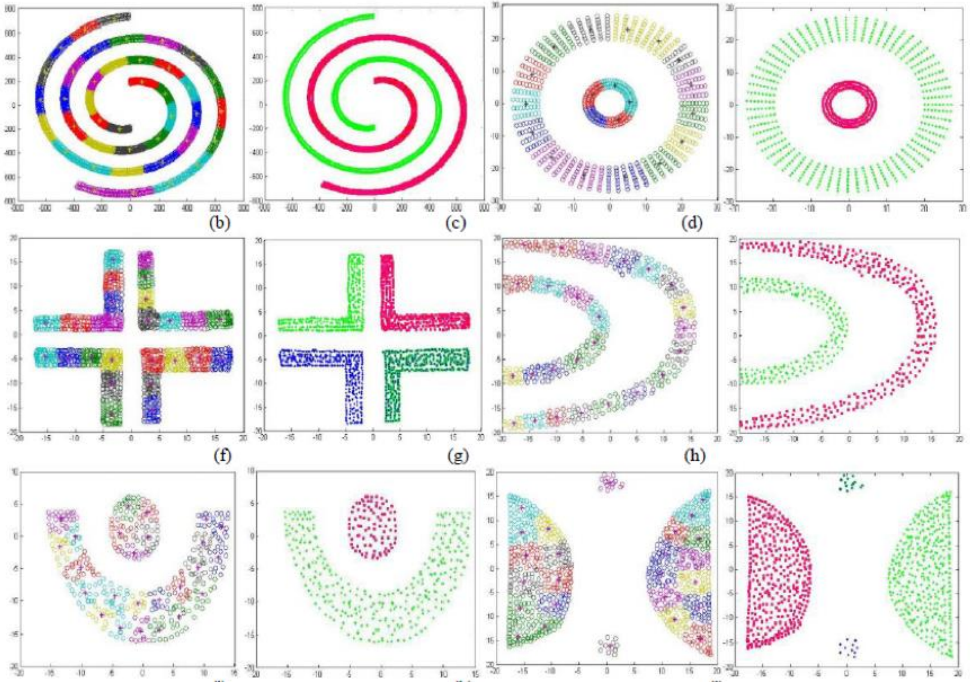
Will define some kind of objective function for a clustering that determines how good the assignments are

- Based on distance of assigned examples to each cluster.
- Close distance reflects strong similarity between datapoints.

Not Always Easy

There are many clusters that are harder to learn with this setup

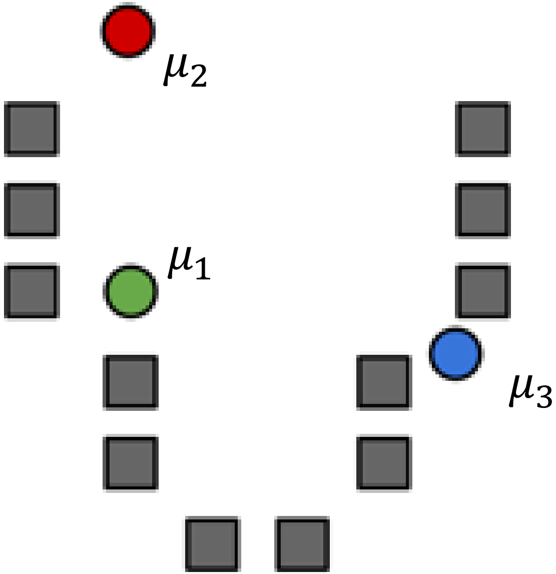
- Distance does not determine clusters



Step 0

Start by choosing the initial cluster centroids

- A common default choice is to choose centroids $\mu_1, \dots, \mu_k$ randomly
- Will see later that there are smarter ways of initializing



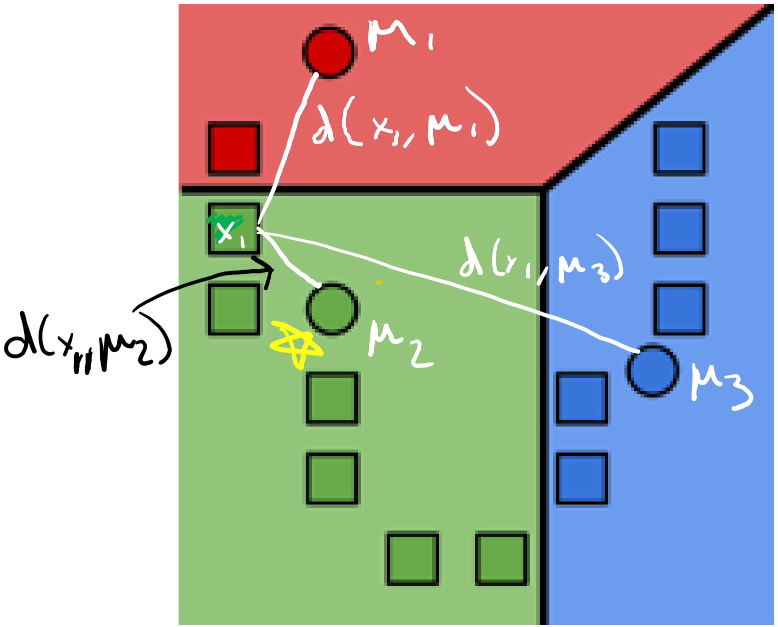
Step 1

Voronoi Tesselation

Assign each example to its closest cluster centroid

For i = 1 to n

$$z_i \leftarrow \underset{j \in [k]}{\operatorname{argmin}} \left|\left|\mu_j - x_i\right|\right|_2^2$$



Step 2

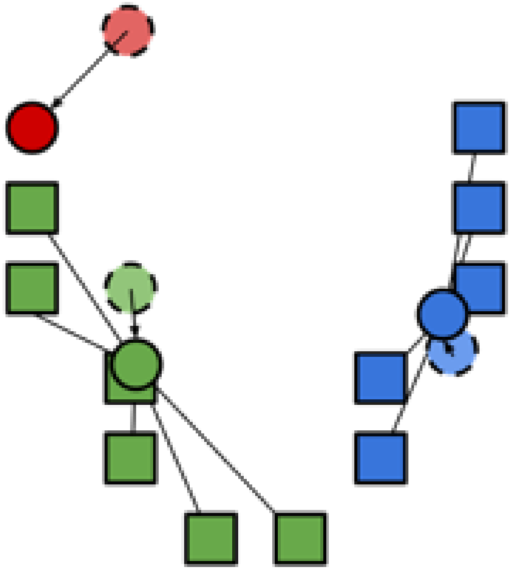
Update the centroids to be the mean of points assigned to that cluster.

$$\mu_j = \frac{\sum_{i=1}^{n} \mathbf{1}\{z_i = j\} x_i}{\sum_{i=1}^{n} \mathbf{1}\{z_i = j\}}$$

= sum of datapoints assigned to cluster j

= number of datapoints assigned to cluster j

Computes center of mass for cluster!



Visualizing k-means

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Smart Initializing w/ k-means++

Making sure the initialized centroids are “good” is critical to finding quality local optima. Our purely random approach was wasteful since it’s very possible that initial centroids start close together.

Idea: Try to select a set of points farther away from each other.

k-means++ does a slightly smarter random initialization

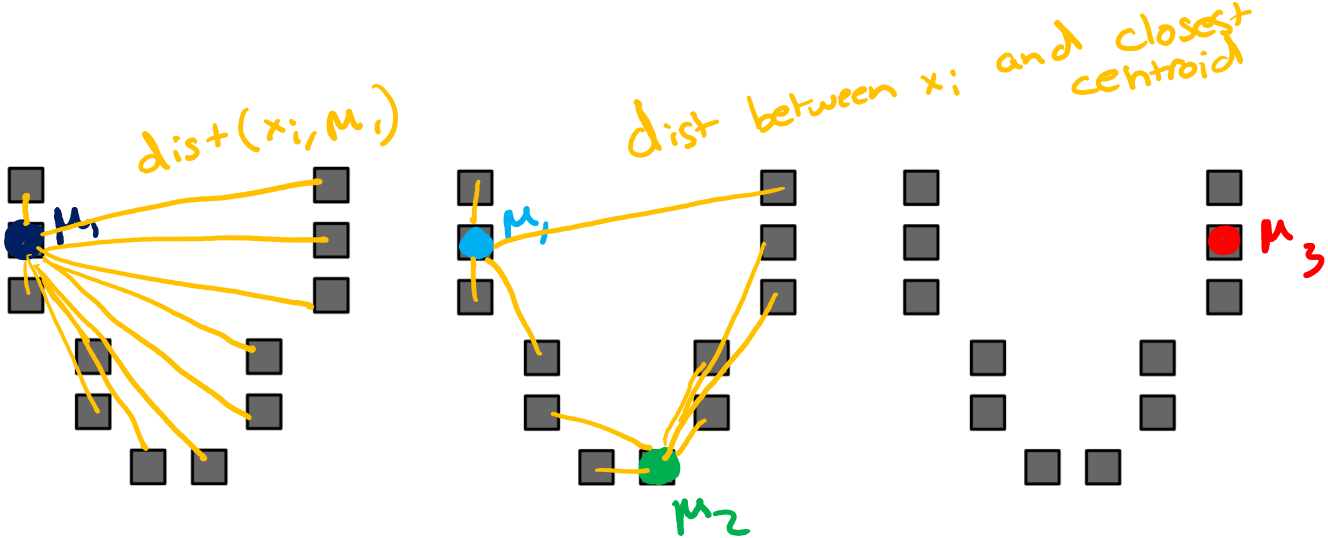
1. Choose first cluster μ_1 from the data uniformly at random
2. For each datapoint x_i, compute the distance between x_i and the closest centroid from the current set of centroids (starting with just μ_i). Denote that distance $d(x_i)$.
3. Choose a new centroid from the remaining data points, where the probability of x_i being chosen is proportional to $d(x_i)^2$.
4. Repeat 2 and 3 until we have selected k centroids.

k-means++ Example

Start by picking a point at random

Then pick points proportional to their distances to their centroids

This tries to maximize the spread of the centroids!



Clustering vs Classification

- Clustering looks like we assigned labels (by coloring or numbering different groups) but we didn’t use any **labeled** data.
- In clustering, the “labels” don’t have meaning. To give meaning to the labels, human inputs is required
- Classification learns from minimizing the error between a prediction and an actual **label**.
- Clustering learns by minimizing the distance between points in a cluster.
- Classification quality metrics (accuracy / loss) do not apply to clustering (since there is no label).
- You can’t use validation set / cross-validation to choose the best choice of k for clustering.

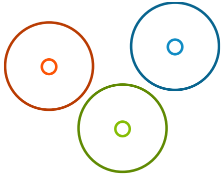
Problems with k-means & Mixture Models

Problems with k-means

In real life, cluster assignments are not always clear cut

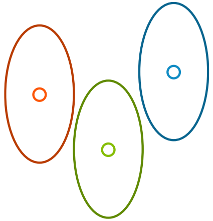
- E.g. The moon landing: Science? World News? Conspiracy?

Because we minimize Euclidean distance, k-means assumes all the clusters are spherical



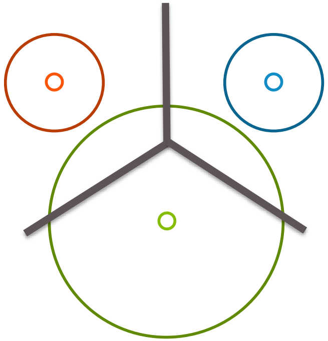
We can change this with weighted Euclidean distance

- Still assumes every cluster is the same shape/orientation

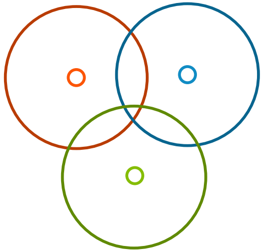


Failure Modes of k-means

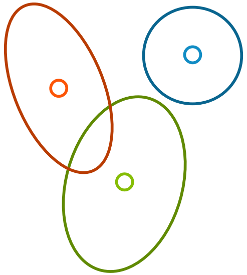
If we don’t meet the assumption of spherical clusters, we will get unexpected results



disparate cluster sizes



overlapping clusters



different shaped/oriented clusters

Mixture Models

A much more flexible approach is clustering with a **mixture model**

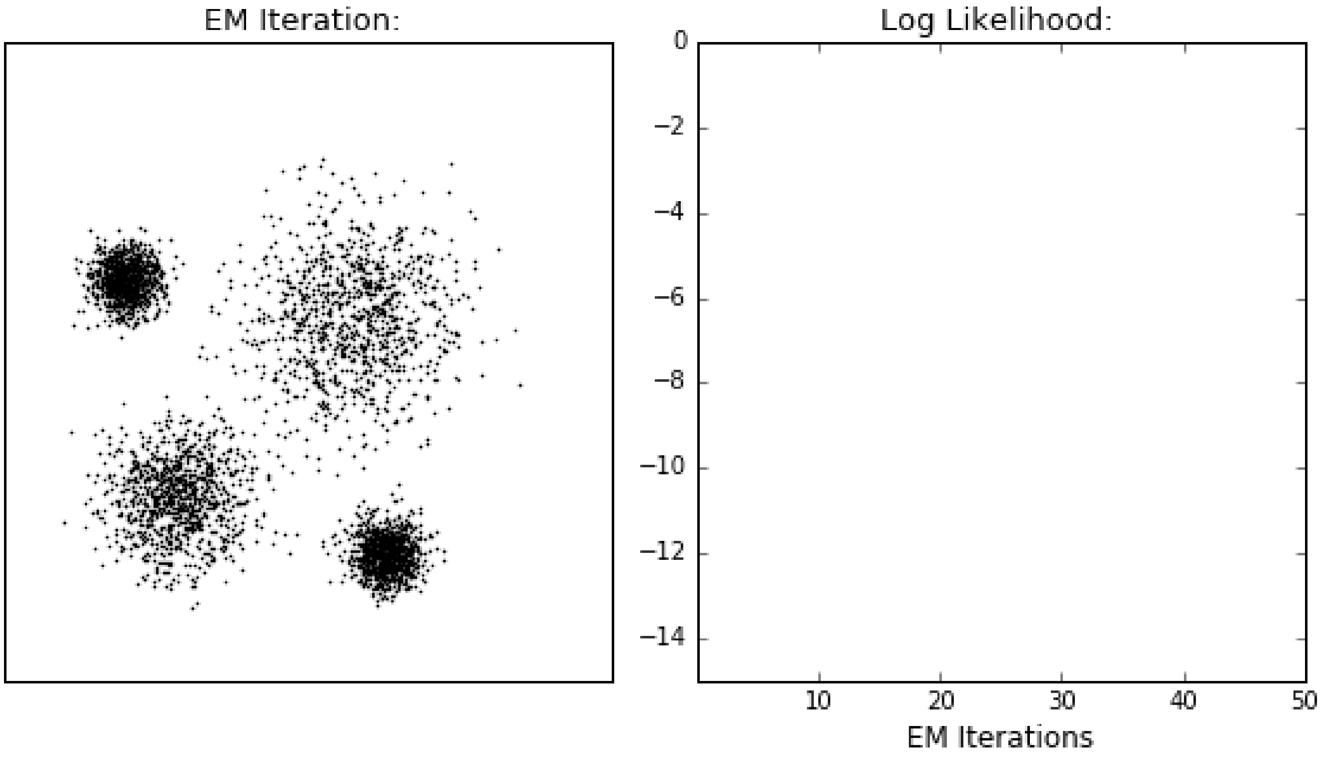
Model each cluster as a different probability distribution and learn their parameters

- One example is Gaussian Mixture Models
- Allows for different cluster shapes and sizes
- Typically learned using Expectation Maximization (EM) algorithm

Allows **soft assignments** to clusters

- Example: A news article: 54% chance is about world news, 45% science, 1% conspiracy theory, 0% other

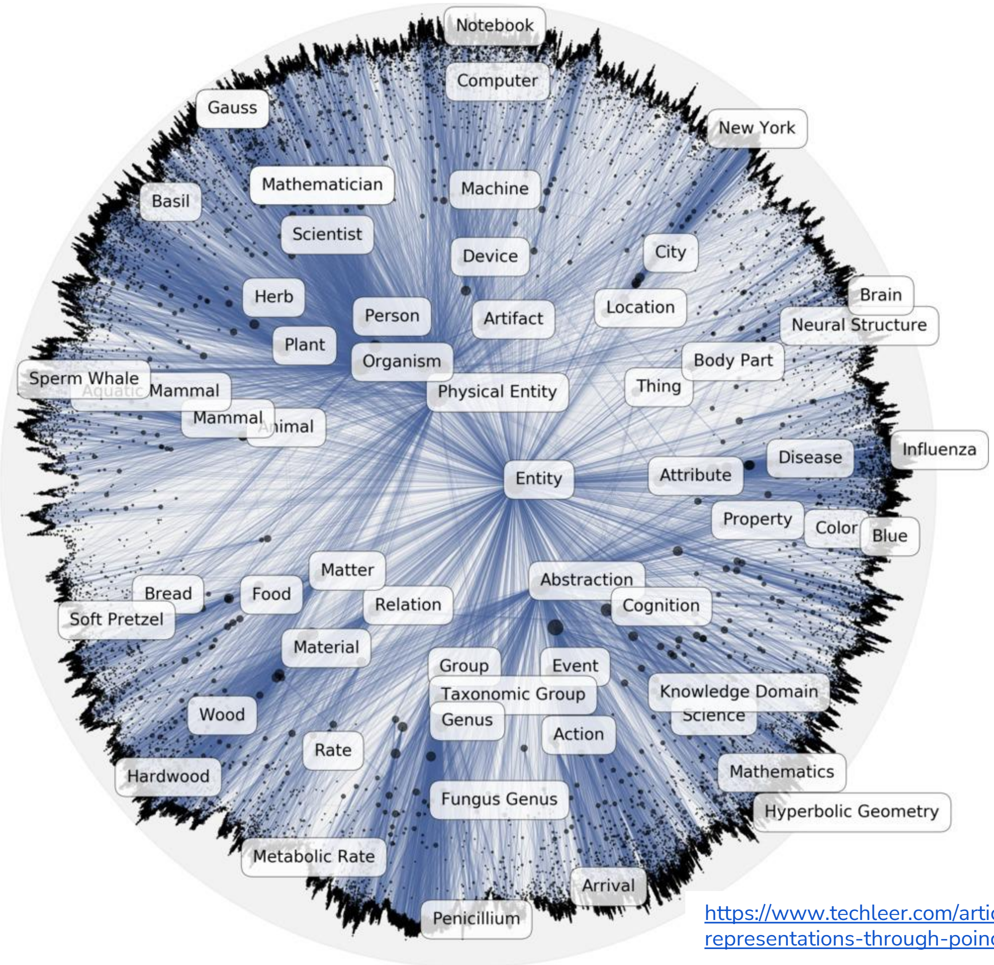
Visualizing Gaussian Mixture Models



Hierarchical Clustering

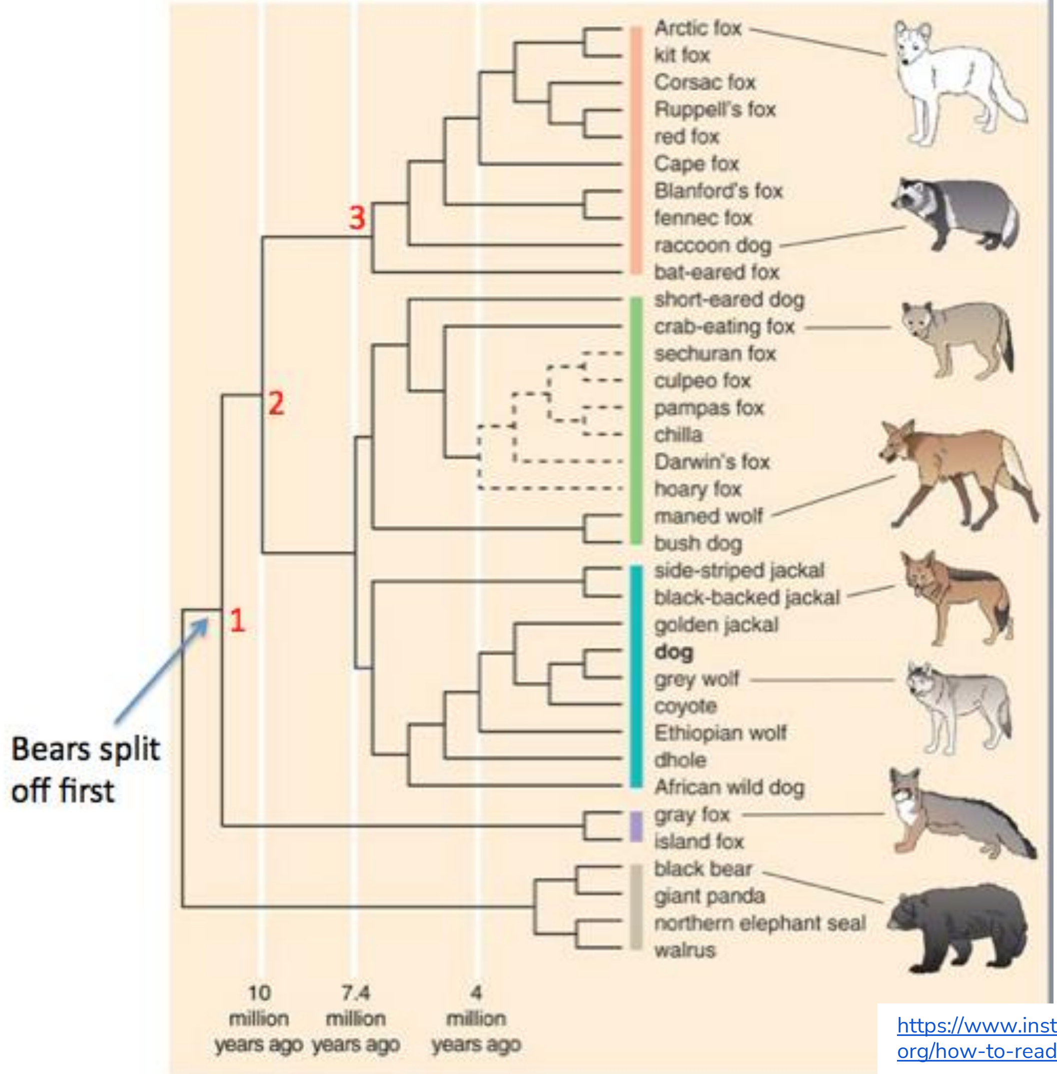
Nouns

Lots of data is hierarchical by nature



https://www.techleer.com/articles/478-insight-into-hierarchical-representations-through-poincare-embedding/

Species

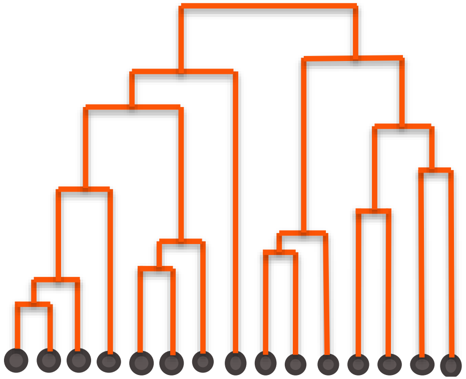


https://www.instituteofcaninebiology.org/how-to-read-a-dendrogram.html

Motivation

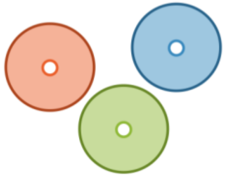
If we try to learn clusters in hierarchies, we can

- Avoid choosing the # of clusters beforehand
- Use **dendrograms** to help visualize different granularities of clusters
- Allow us to use any distance metric
  - K-means requires Euclidean distance
- Can often find more complex shapes than k-means

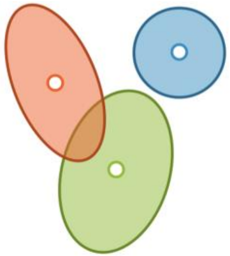


Finding Shapes

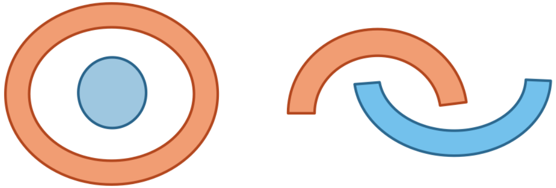
k-means



Mixture Models



Hierarchical Clustering



Types of Algorithms

Divisive, a.k.a. *top-down*

- Start with all the data in one big cluster and then recursively split the data into smaller clusters
  - Example: **recursive k-means**

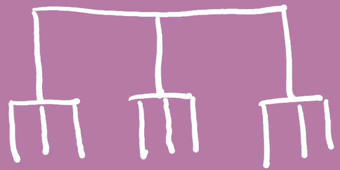
Agglomerative, a.k.a. *bottom-up*:

- Start with each data point in its own cluster. Merge clusters until all points are in one big cluster.
  - Example: **single linkage clustering**

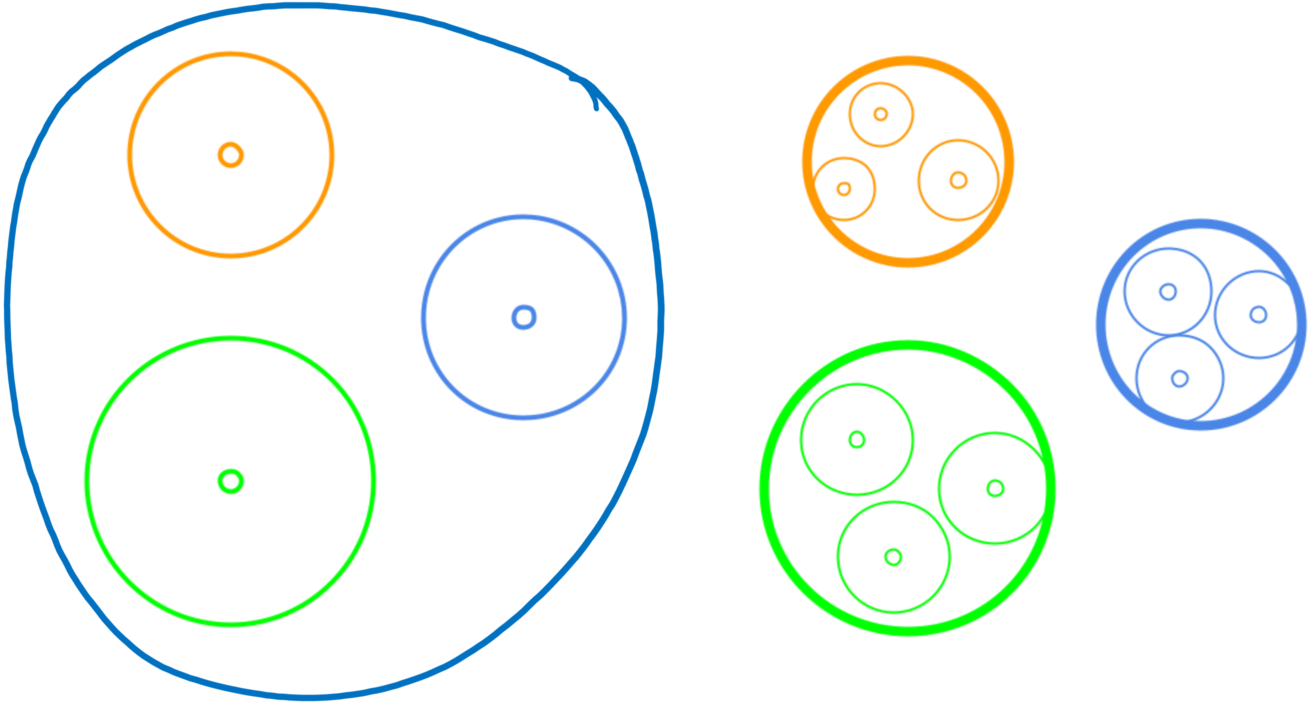
→ how we measure distance between clusters

Divisive Clustering

k=3



Start with all the data in one cluster, and then repeatedly run k-means to divide the data into smaller clusters. Repeatedly run k-means on each cluster to make sub-clusters.

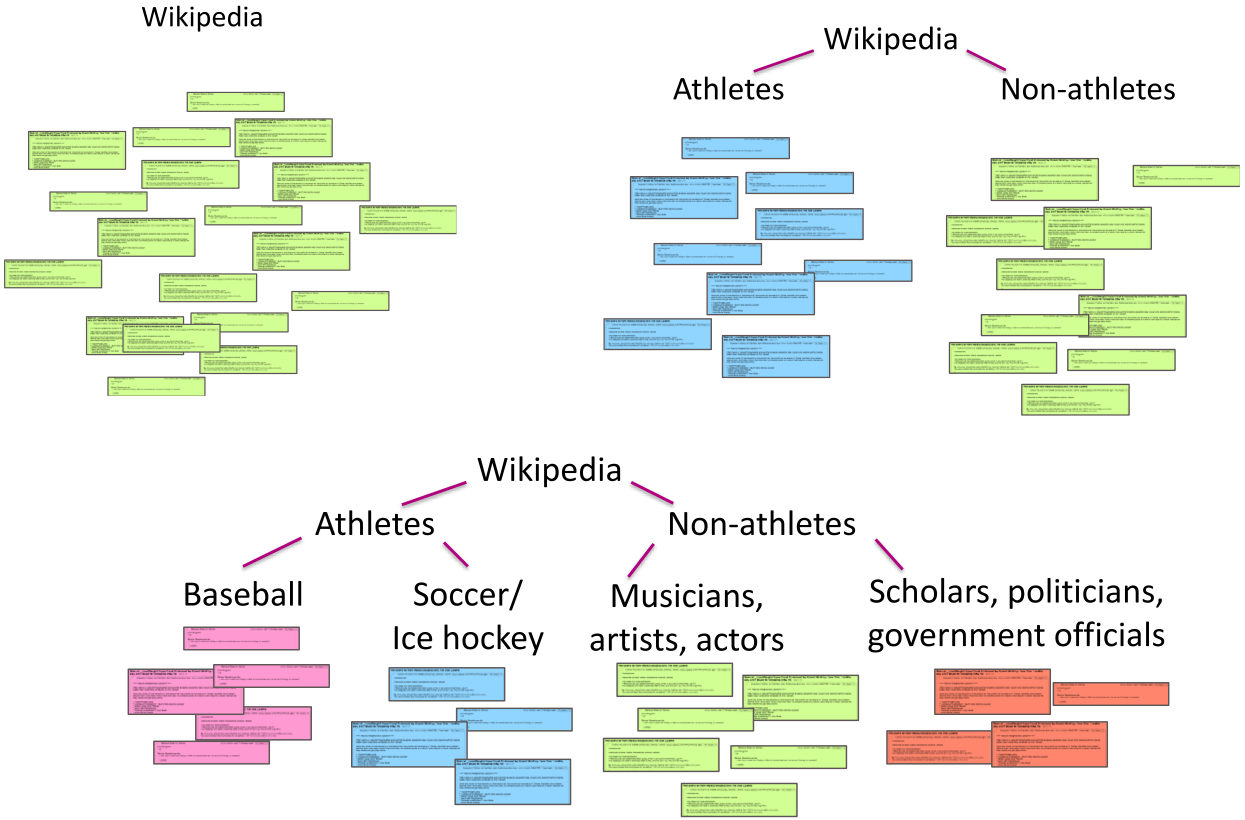


Example

K=2

Bisecting K-means

Using Wikipedia



Wikipedia

Wikipedia

Athletes

Non-athletes

Wikipedia

Athletes

Non-athletes

Baseball

Soccer/ Ice hockey

Musicians, artists, actors

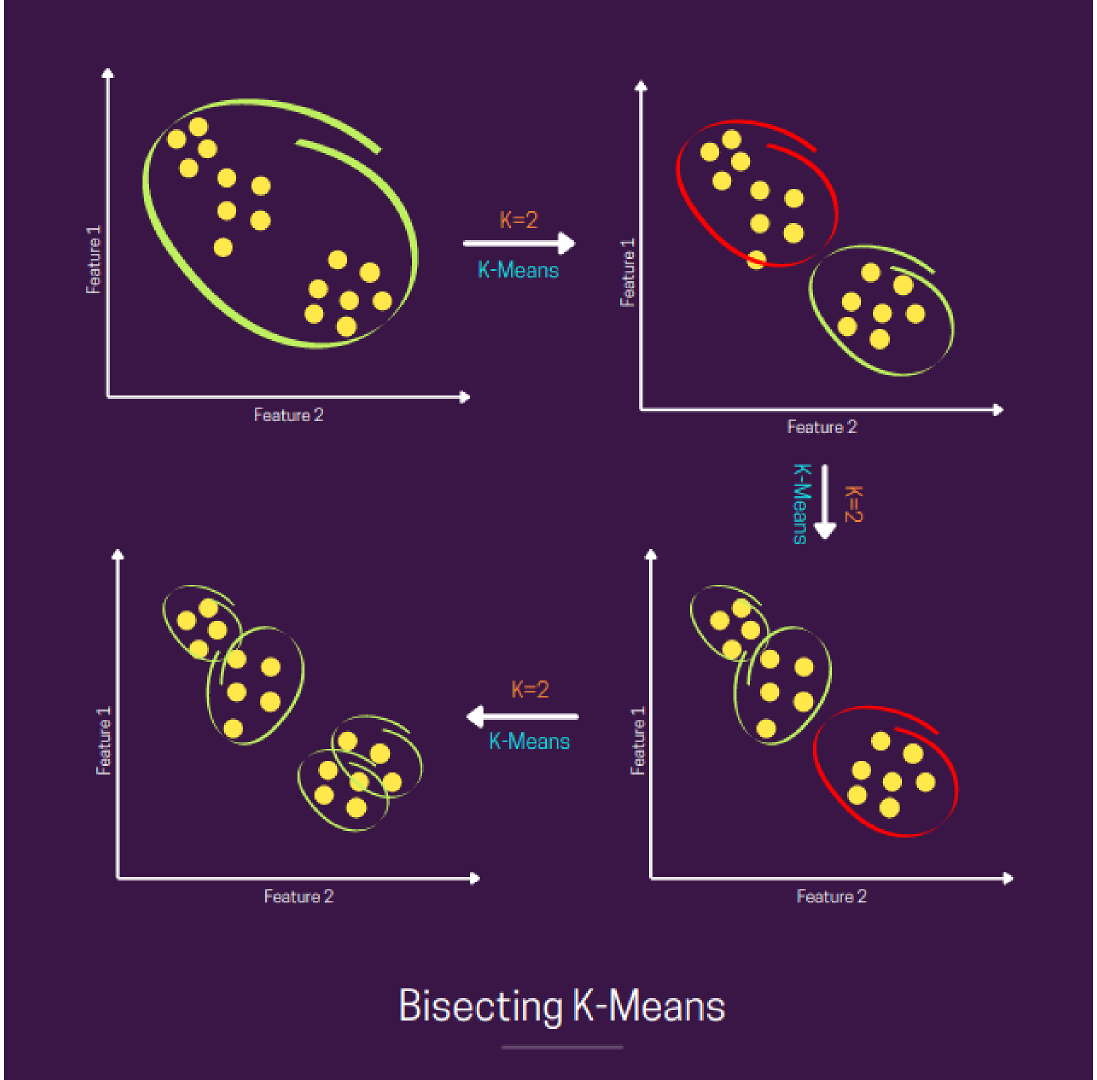
Scholars, politicians, government officials

Choices to Make

For divisive clustering, you need to make the following choices:

- Which algorithm to use (e.g., k-means)
- How many clusters per split
- When to split vs when to stop
  - **Max cluster size**
    Number of points in cluster falls below threshold
  - **Max cluster radius**
    distance to furthest point falls below threshold
  - **Specified # of clusters**
    split until pre-specified # of clusters is reached

Bisecting K-Means

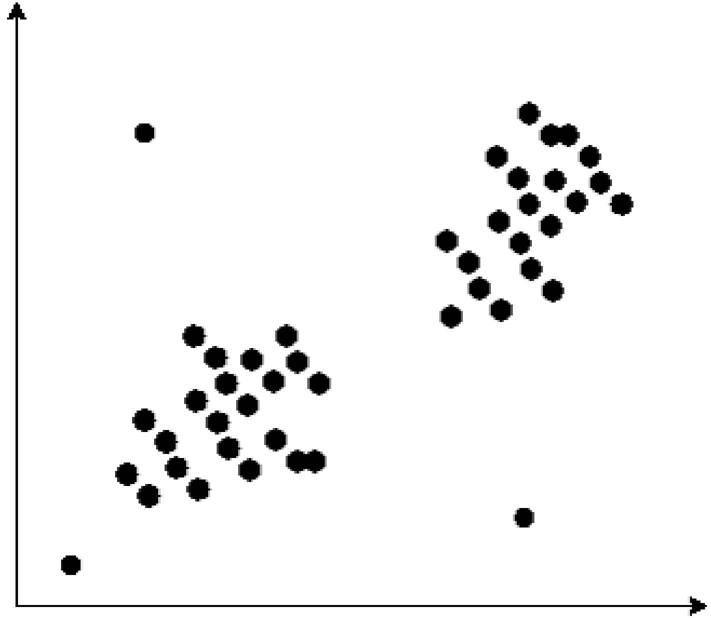


Poll Everywhere

Think

1 min

- You want to detect outliers in a dataset (shown below).
  - How would you use k-means clustering to detect outliers?
  - How would you use divisive clustering to detect outliers?

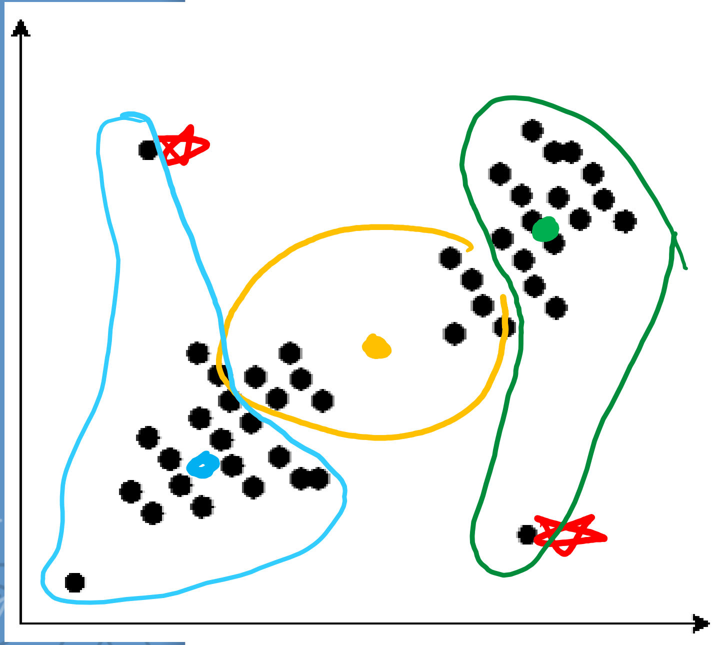


Poll Everywhere

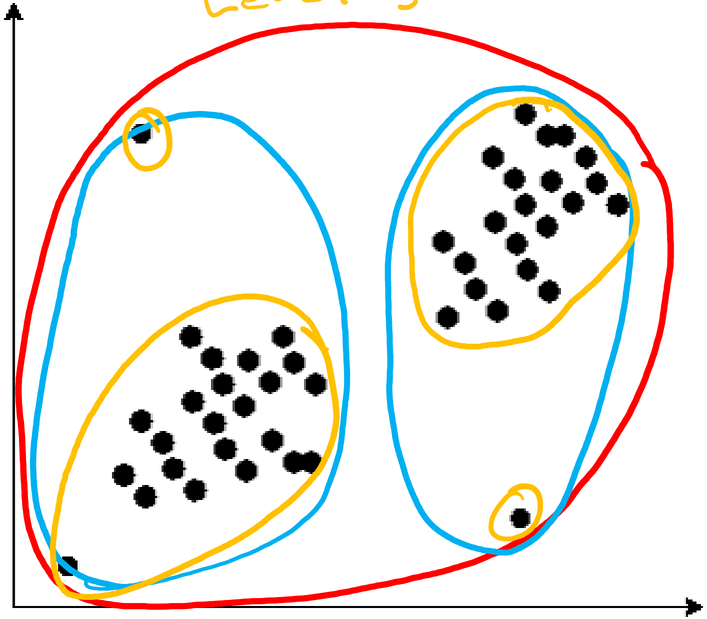
Group

2 min

- You want to detect outliers in a dataset (shown below).
  - How would you use k-means clustering to detect outliers?
  - How would you use divisive clustering to detect outliers?



Level 1
Level 2
Level 3



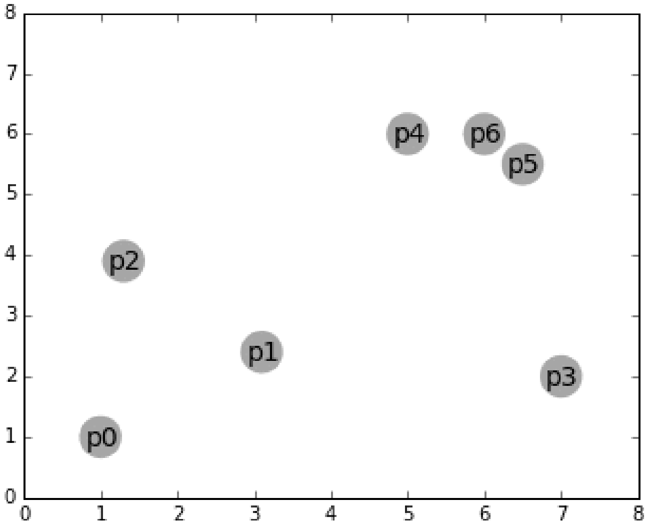
Brain Break

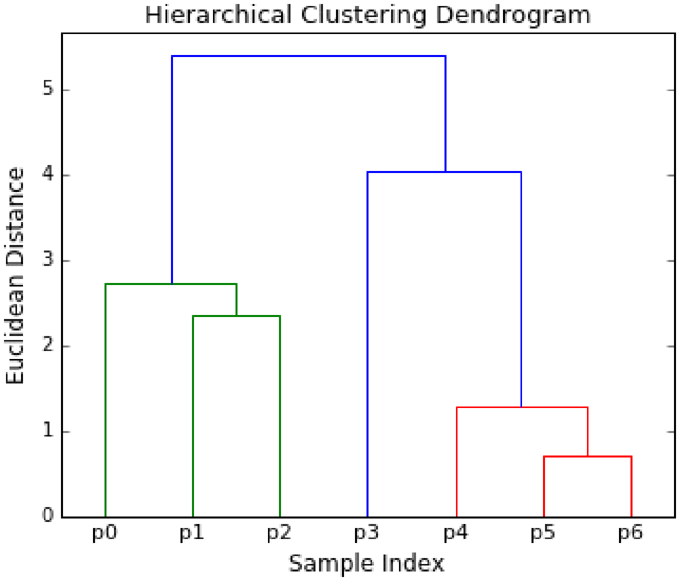
3:12

Agglomerative Clustering

Agglomerative Clustering

Merge closest pair of clusters





Agglomerative Clustering

Algorithm at a glance

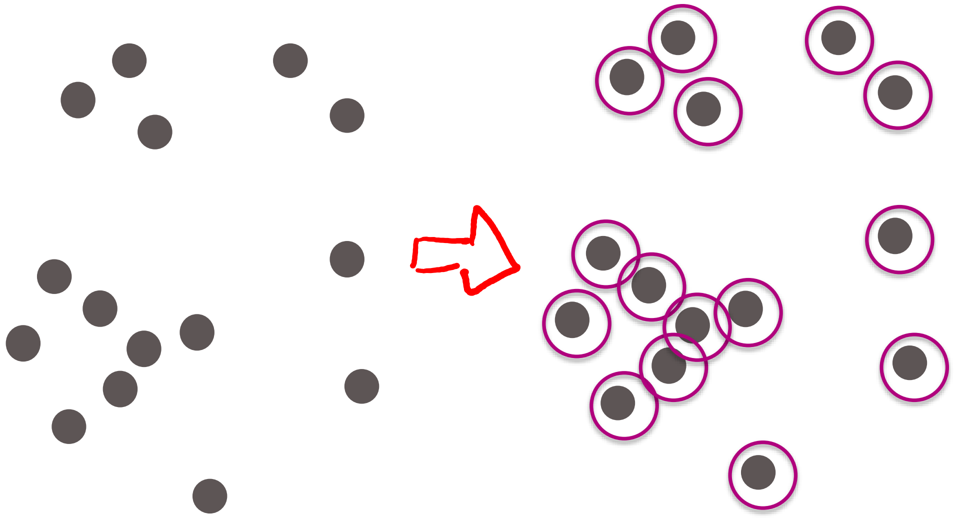
1. Initialize each point in its own cluster
2. Define a distance metric between clusters ← Hyperparameter

While there is more than one cluster

3. Merge the two closest clusters (and add it to dendrogram)

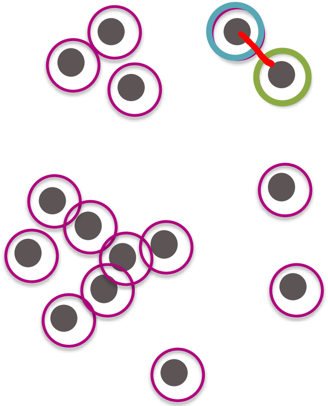
Step 1

1. Initialize each point to be its own cluster



Step 2

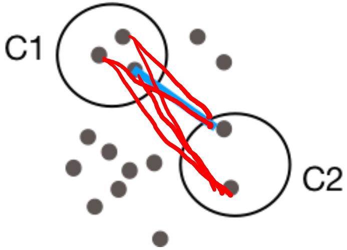
2. Define a distance metric between clusters



Single Linkage

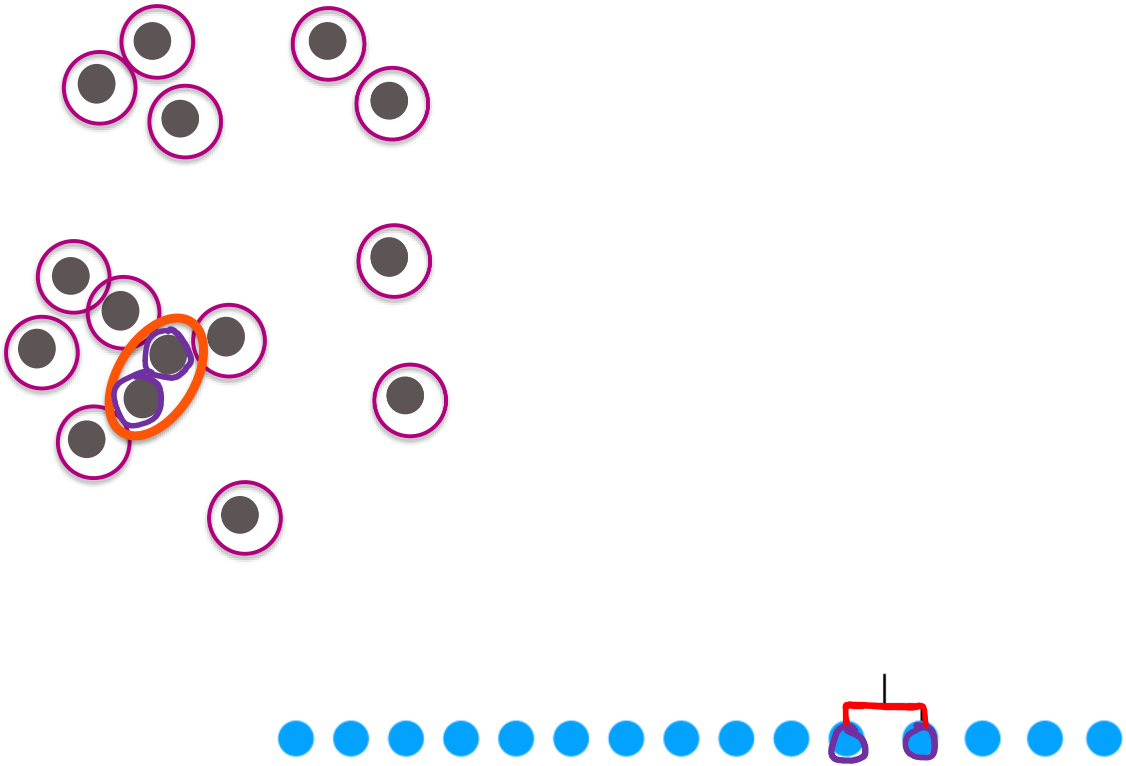
$$distance\left(C^{(1)}, C^{(2)}\right) = \min_{x^{(i)} \in C^{(1)}, x^{(j)} \in C^{(2)}} d\left(x^{(i)}, x^{(j)}\right)$$

This formula means we are defining the distance between two clusters as the smallest distance between any pair of points between the clusters.

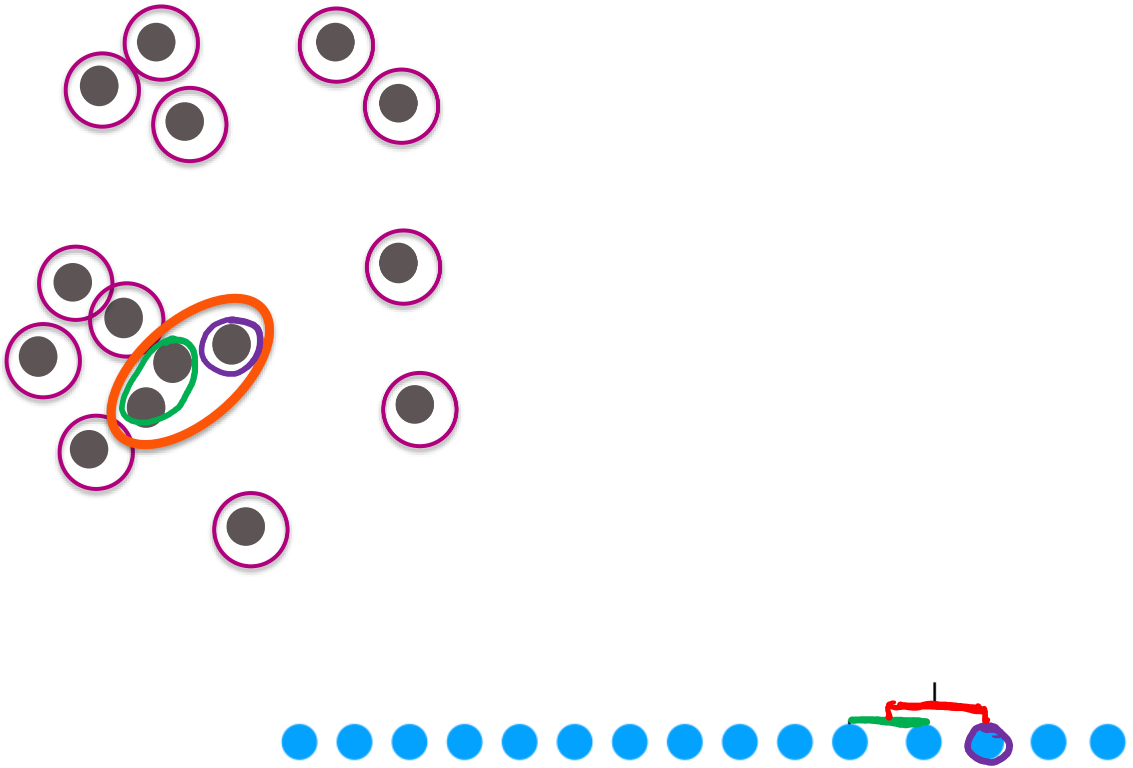


Step 3

Merge closest pair of clusters

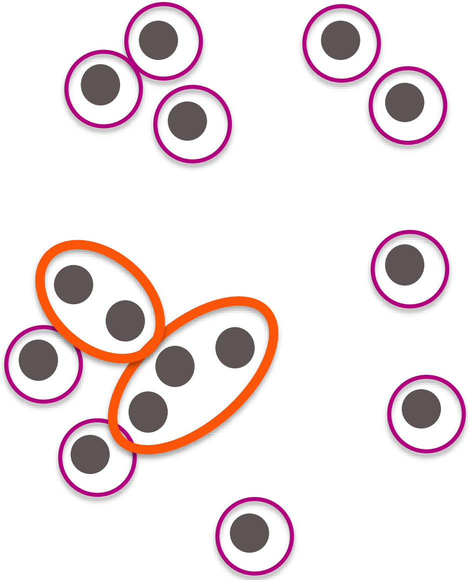


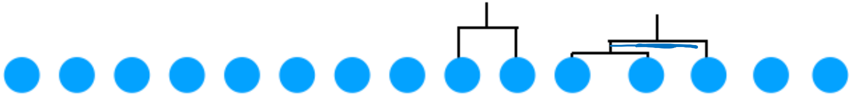
Repeat



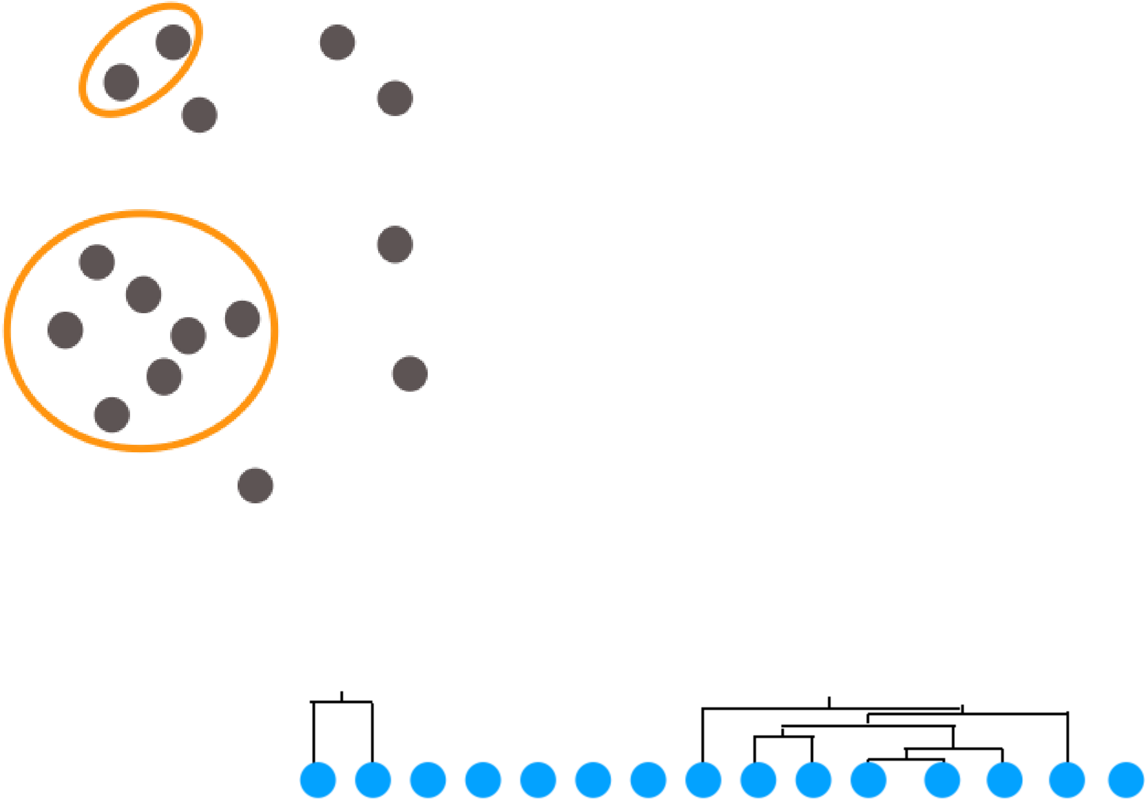
Repeat

Notice that the height of the dendrogram is growing as we group points farther from each other





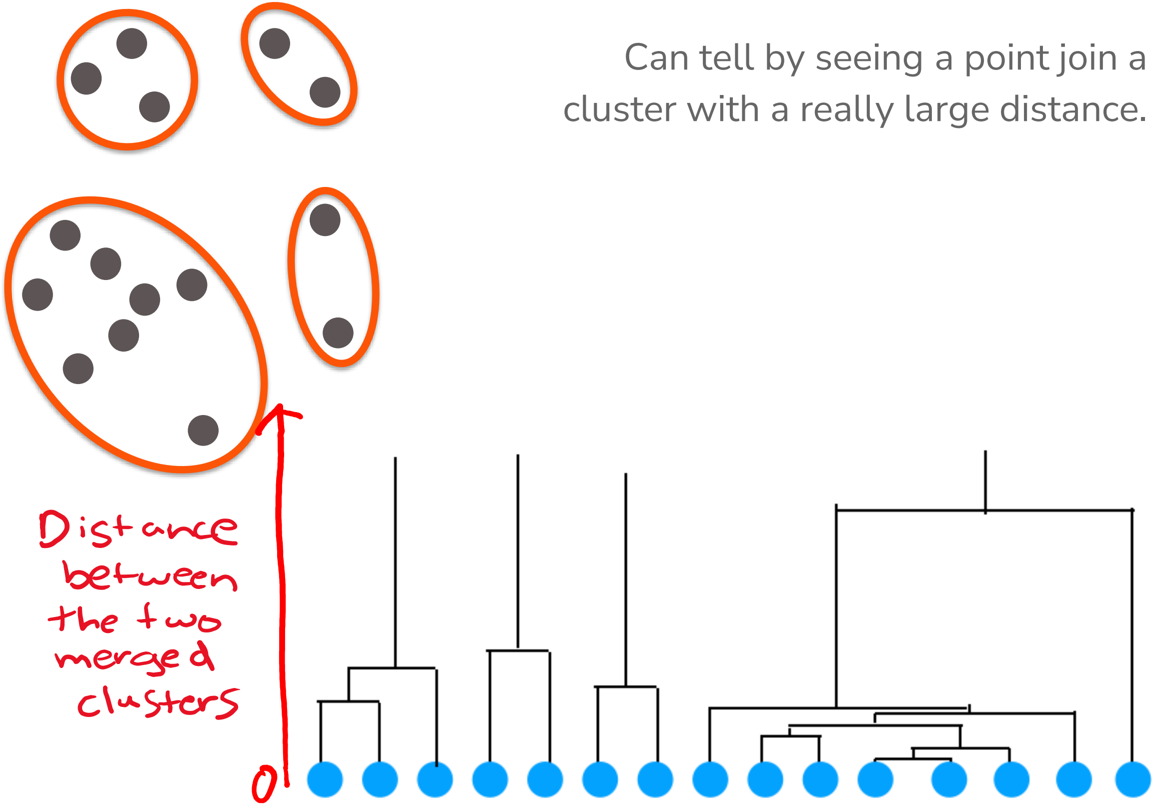
Repeat



Repeat

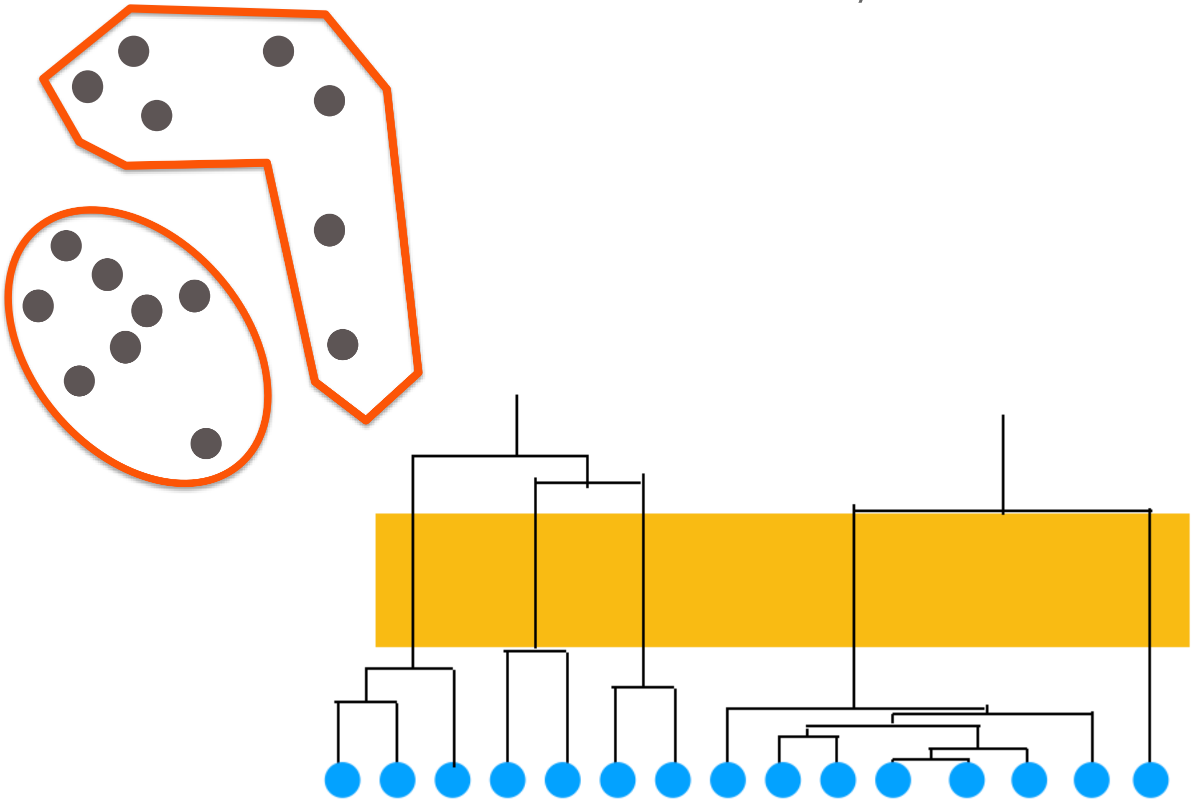
Looking at the dendrogram, we can see there is a bit of an outlier!

Can tell by seeing a point join a cluster with a really large distance.



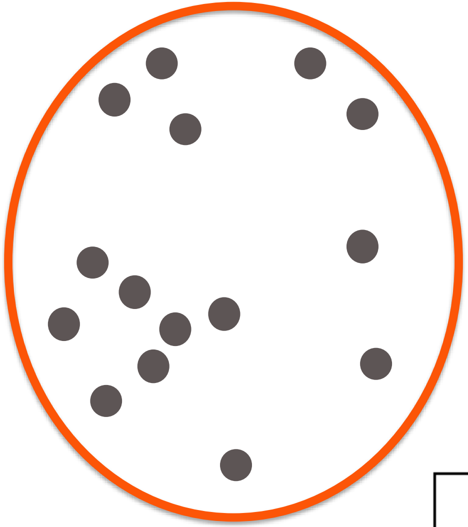
Repeat

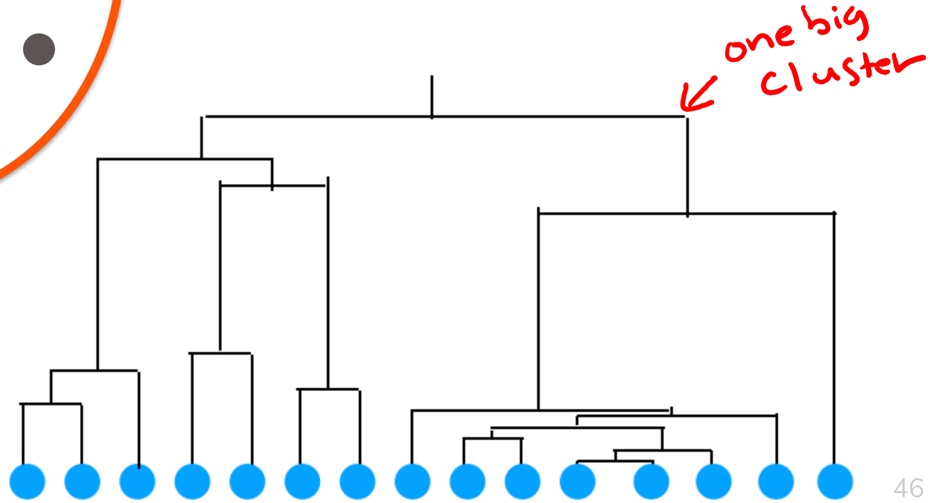
The tall links in the dendrogram show us we are merging clusters that are far away from each other



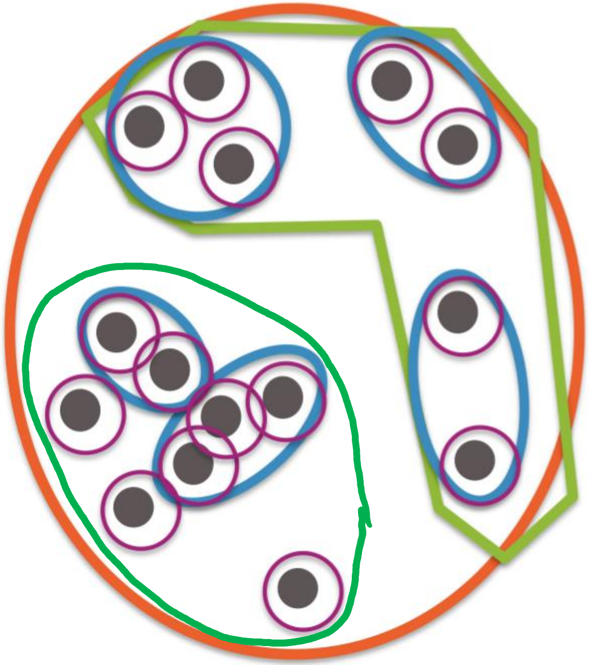
Repeat

Final result after merging all clusters





Final Result

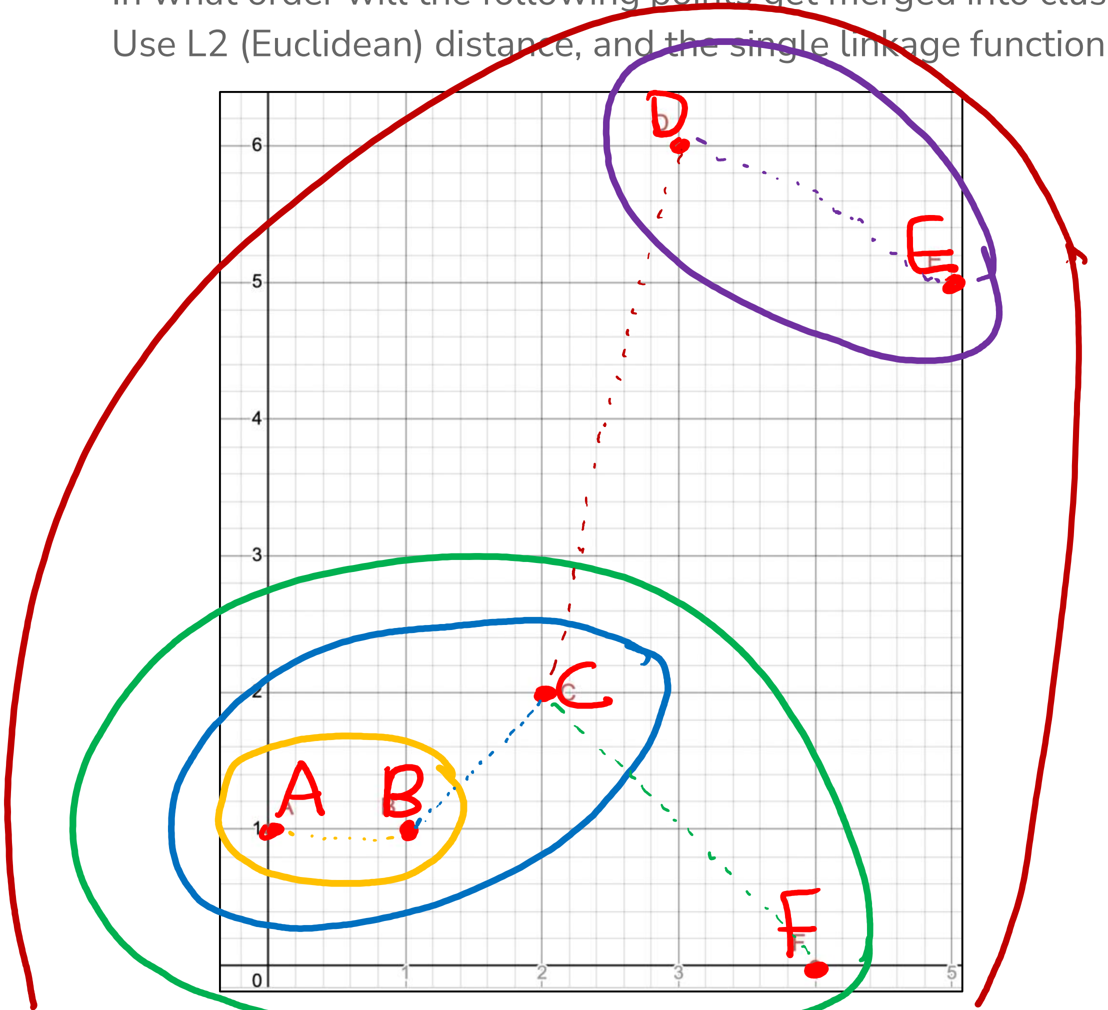


Poll Everywhere

Think

1 min

- In what order will the following points get merged into clusters? Use L2 (Euclidean) distance, and the single linkage function.

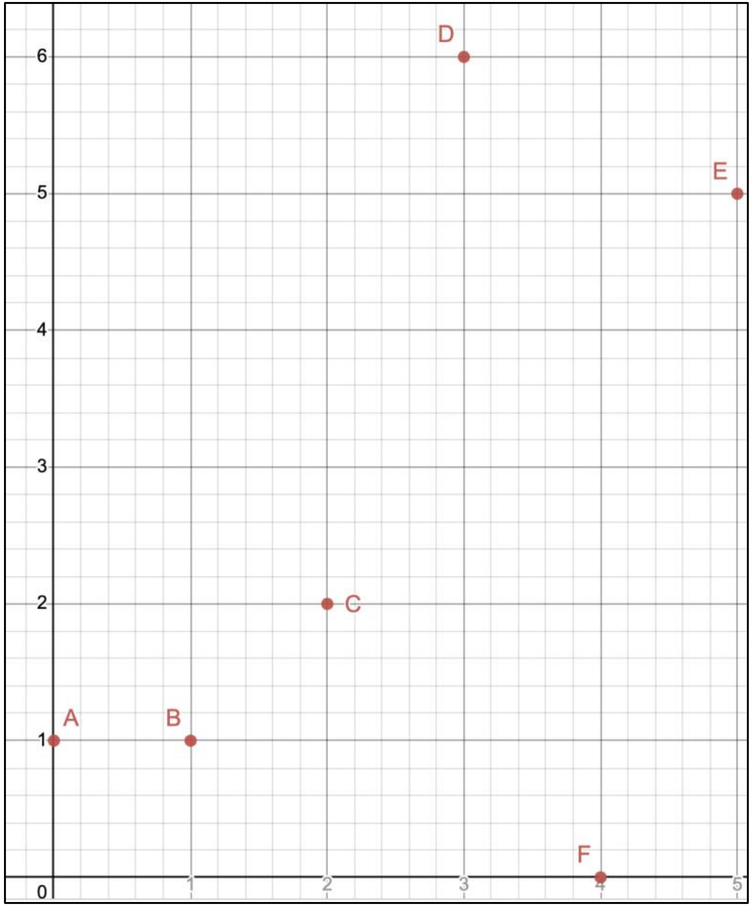


Poll Everywhere

Group

2 min

- In what order will the following points get merged into clusters? Use L2 (Euclidean) distance, and the single linkage function.

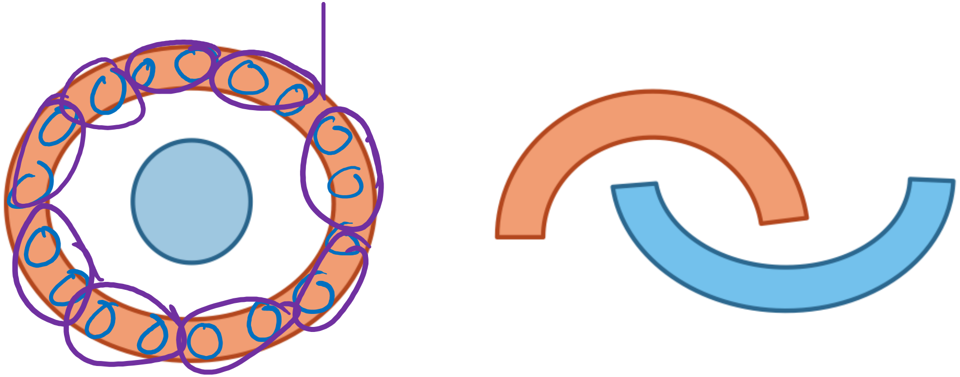


Dendrograms

Agglomerative Clustering

With agglomerative clustering, we are now very able to learn weirder clusterings like

Single Linkage: $\min_{x_i \in C_1, x_j \in C_2} dist(x_i, x_j)$

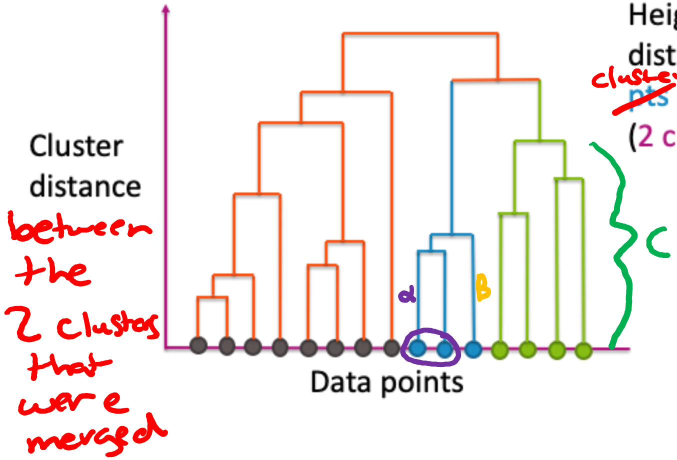


Single Linkage can merge long chains

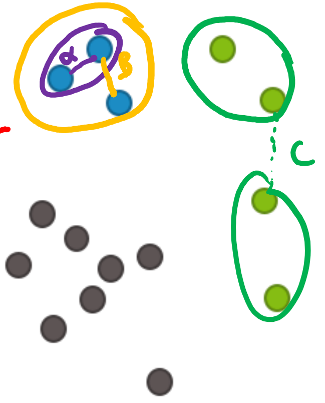
Dendrogram

x-axis shows the datapoints (arranged in a very particular order)

y-axis shows distance between pairs of clusters

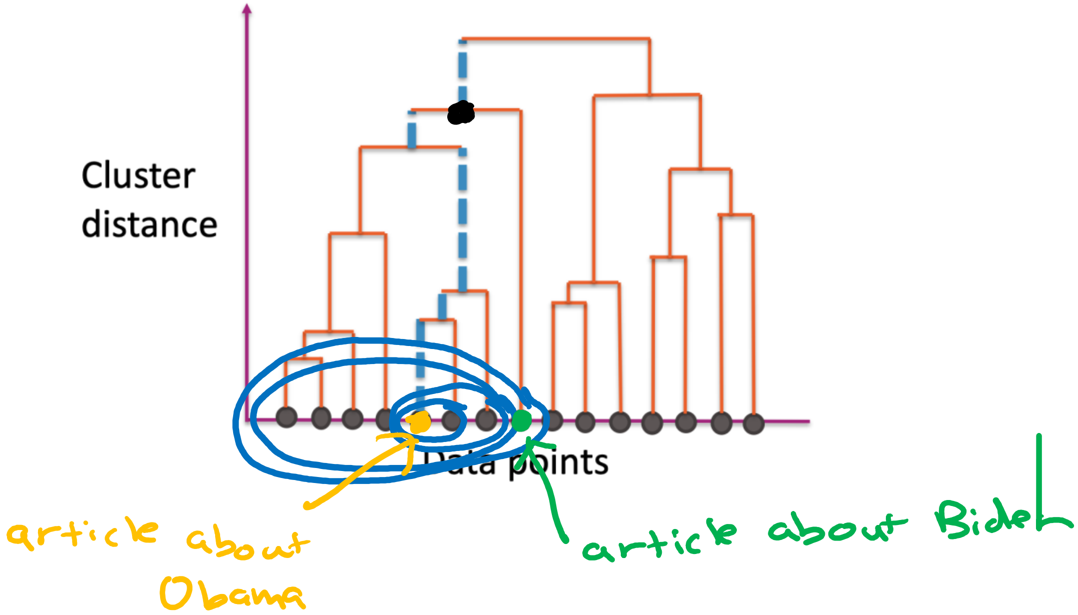


Height here indicates min distance between blue ~~pts~~ cluster and green ~~pts~~ cluster (2 clusters)



Dendrogram

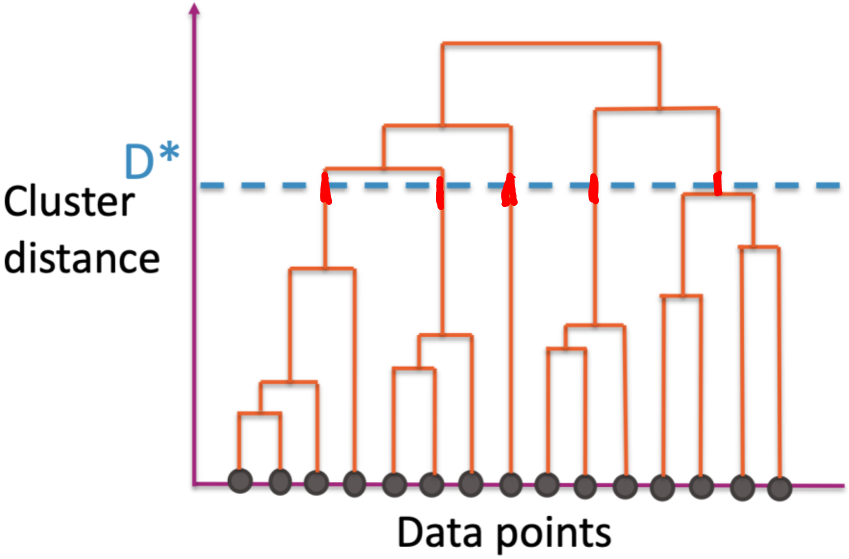
The path shows you all clusters that a single point belongs and the order in which its clusters merged



Cut Dendrogram

Choose a distance D^* to “cut” the dendrogram

- Use the largest clusters with distance < D^*
- Usually ignore the idea of the nested clusters after cutting

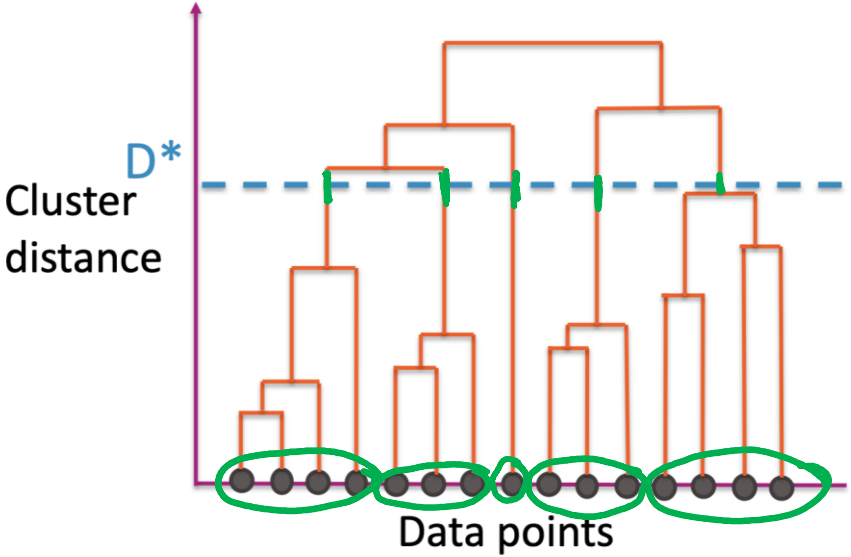


Poll Everywhere

Think 👥

1 min

How many clusters would we have if we use this threshold?

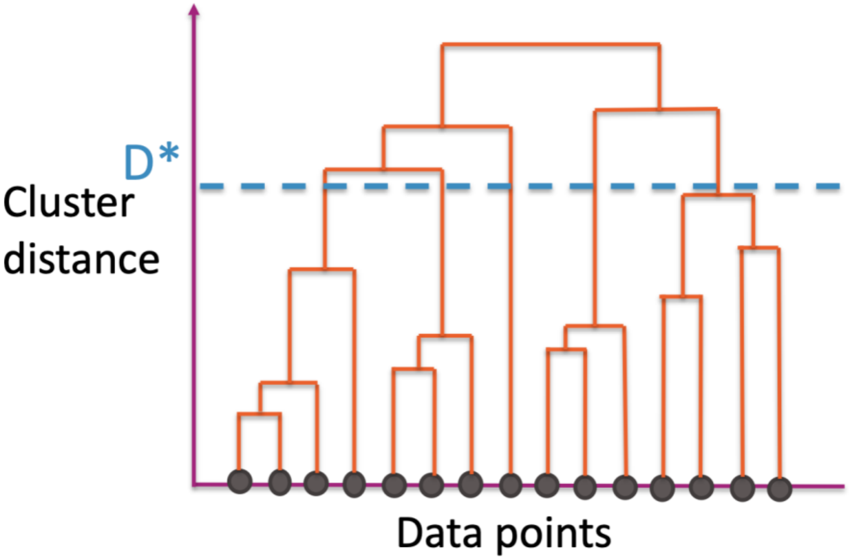


Poll Everywhere

Think

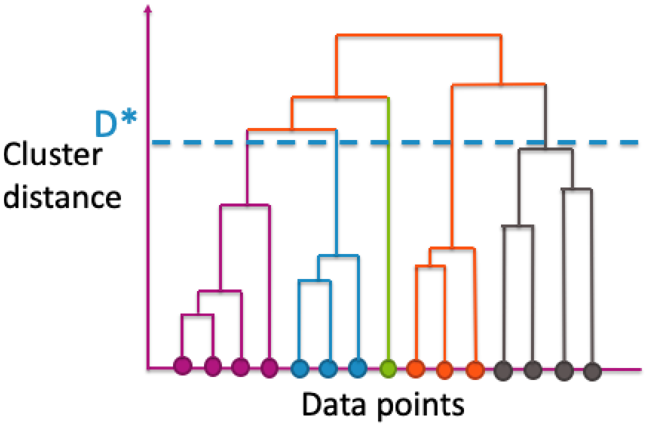
2 min

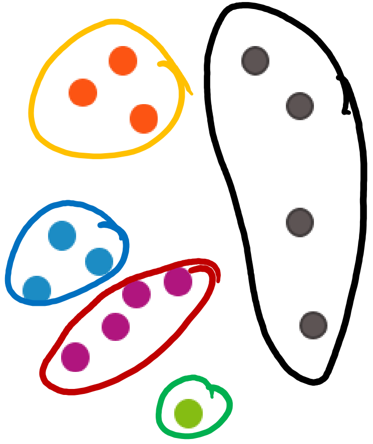
How many clusters would we have if we use this threshold?



Cut Dendrogram

Every branch that crosses D^* becomes its own cluster





Choices to Make

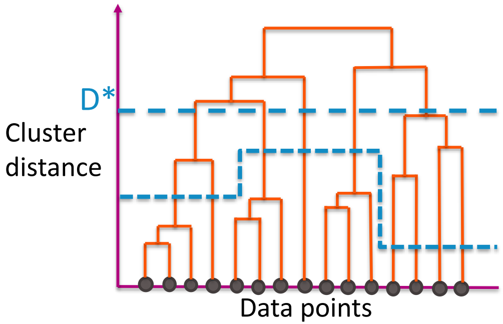
For agglomerative clustering, you need to make the following choices:

- Distance metric $d(x_i, x_j)$
- Linkage function
  - Single Linkage:

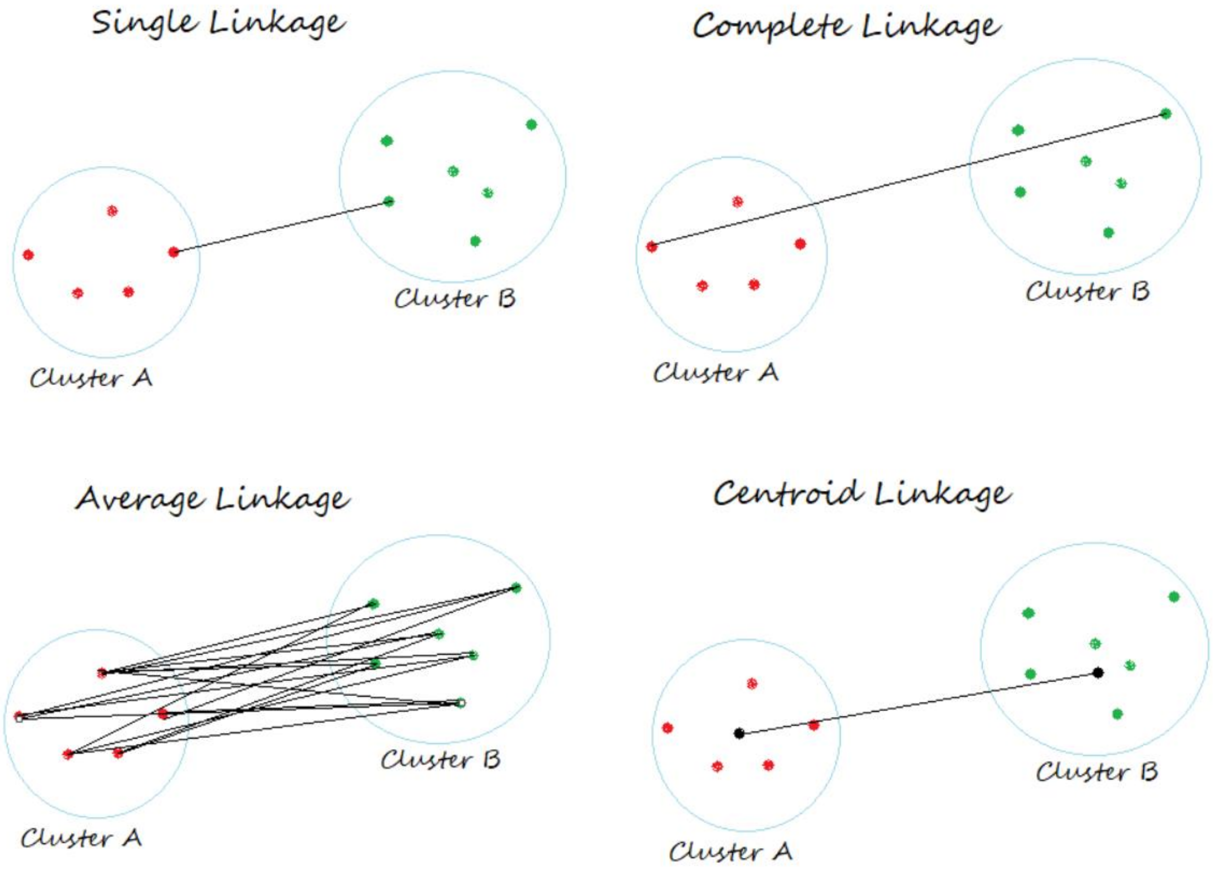
    $$\min_{x_i \in C_1, x_j \in C_2} d(x_i, x_j)$$
  - Complete Linkage:

    $$\max_{x_i \in C_1, x_j \in C_2} d(x_i, x_j)$$
  - Centroid Linkage

    $$d(\mu_1, \mu_2)$$
  - Others
- Where and how to cut dendrogram



Linkage Functions



Practical Notes

For visualization, generally a smaller # of clusters is better

For tasks like outlier detection, cut based on:

- Distance threshold
- Or some other metric that tries to measure how big the distance increased after a merge

No matter what metric or what threshold you use, no method is “incorrect”. Some are just more useful than others.

Computational Cost of Agglomerative Clustering

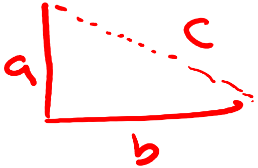
Computing all pairs of distances is pretty expensive!

- A simple implementation takes $\mathcal{O}(n^2 \log(n))$

Can be much implemented more cleverly by taking advantage of the **triangle inequality**

- "Any side of a triangle must be less than the sum of its sides"

Best known algorithm is $\mathcal{O}(n^2)$



$c \leq a+b$

k-means vs. Agglomerative Clustering

- K-means is more efficient on big data than hierarchical clustering.
- Initialization changes results in k-means, not in agglomerative clustering has reproducible results.
- K-means works well only for hyper-spherical clusters, agglomerative clustering can handle more complex cluster shapes.
- K-means requires selecting a number of clusters beforehand. In agglomerative clustering, you can decide on the number of clusters afterwards using the dendrogram.

SKIPPED

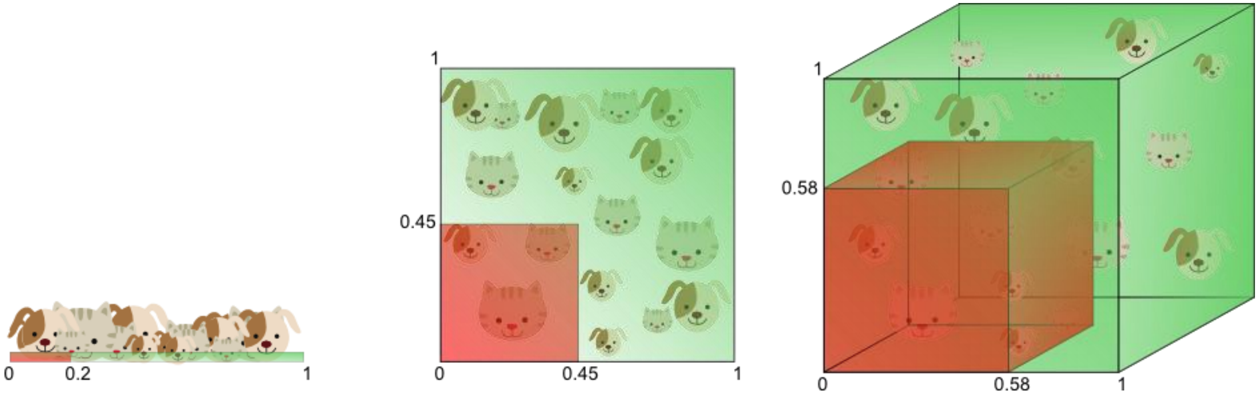
Curse of Dimensionality

High Dimensions

SKIPPED

Methods like k-NN and k-means that rely on computing distances start to struggle in high dimensions.

As the number of dimensions grow, the data gets sparser!

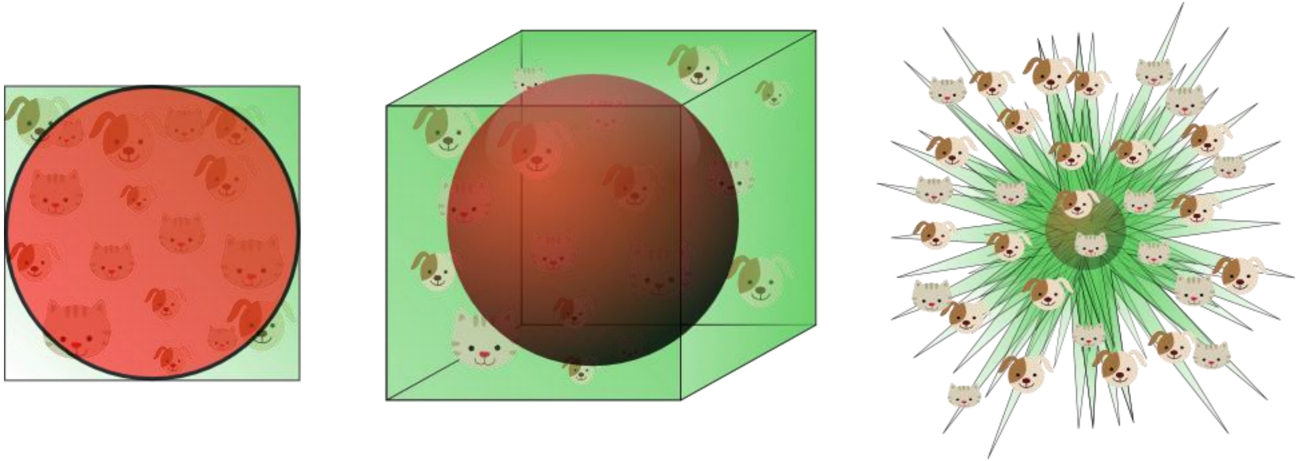


Need more data to make sure you cover all the space in high dim.

SKIPPED

Data Moves Farther Apart in Higher Dimensions

It's believable with more dimensions the data becomes more sparse, but what's even weirder is the sparsity is not uniform!



As D increases, the "mass" of the space goes towards the corners.

- Most of the points aren't in the center.
- The distance between points gets really high!

SKIPPED

Practicalities

Have to pay attention to the number of dimensions with distance-based methods (k-means clustering, also k nearest neighbors).

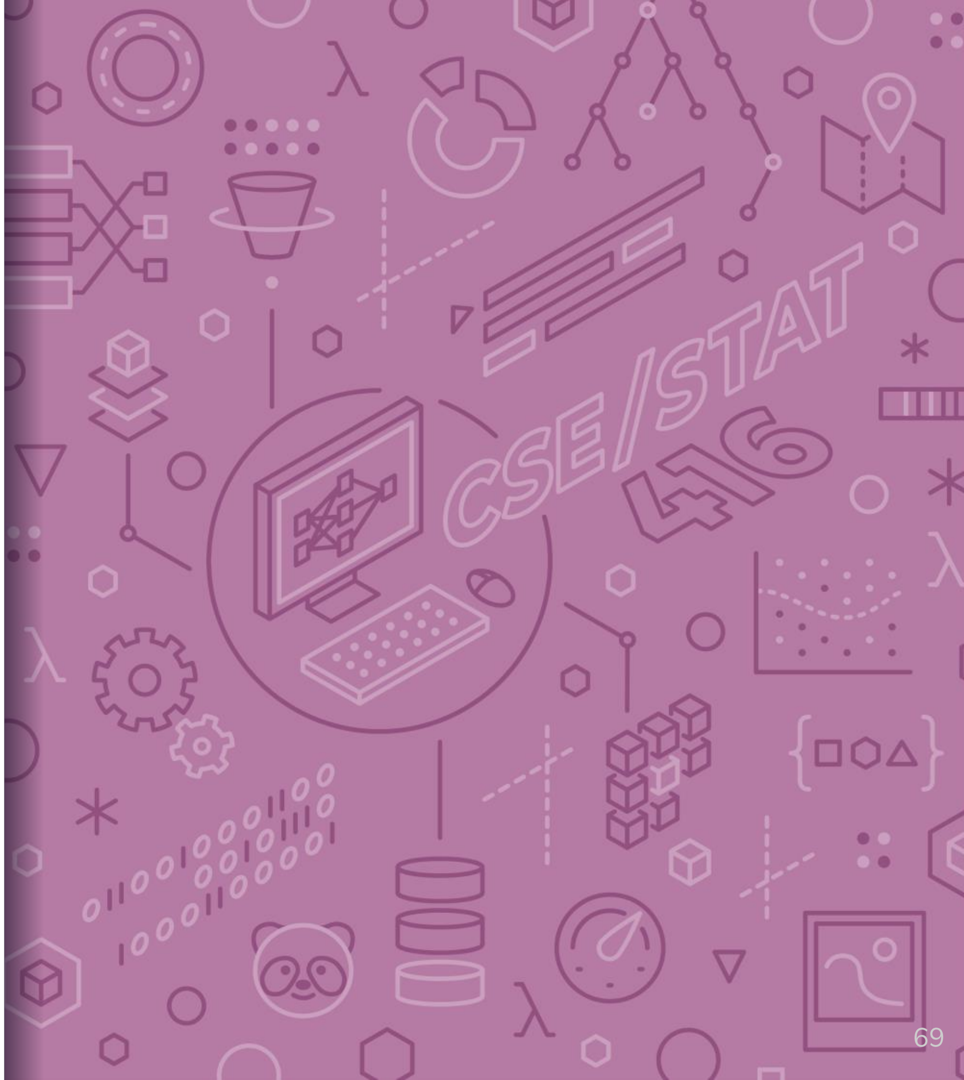
- Very tricky if $n < D$
- Can run into some strange results if D is very large

Later, we will talk about ways of trying to do dimensionality reduction in order to reduce the number of dimensions here.

Recap

- Problems with k-means
- Mixture Models
- Hierarchical clustering
- Divisive Clustering
- Agglomerative Clustering
- Dendrograms

Activity: Promoting Fairness in Machine Learning Models



Ways to Promote Ethical Machine Learning

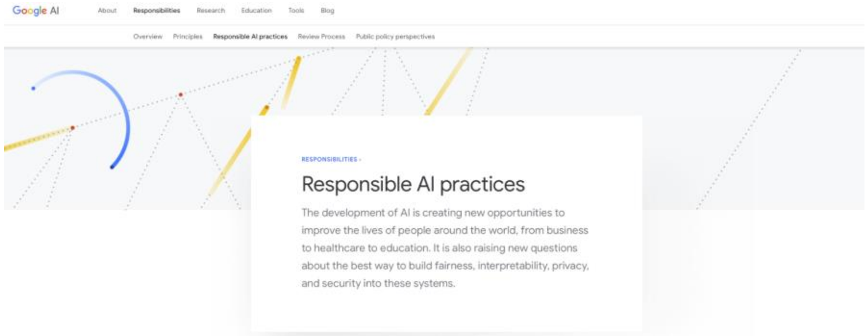
IBM abandons facial recognition

- In the aftermath of killing of George Floyd, IBM decided to abandon its facial recognition for surveillance and profiling
- IBM recognized that its software is biased and hence a hindrance in fight against racism.
- We might not have technological solutions – sometimes, not creating a technology is the best route.

https://www.bbc.com/news/technology-52978191

Education

- Many top tech companies have made guidelines for Responsible AI practices.
- These are prescribed to the AI teams and their engineers
- These practices are (hopefully) utilized when developing their current and future products.



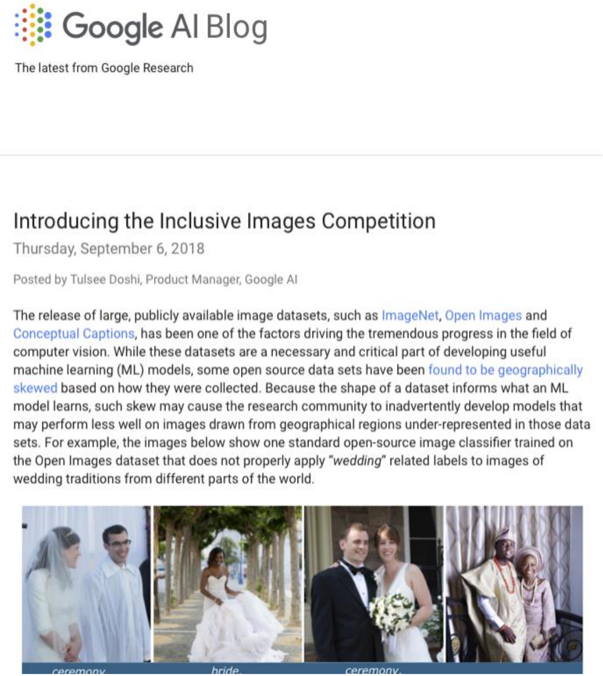
RESPONSIBILITIES ›

Responsible AI practices

The development of AI is creating new opportunities to improve the lives of people around the world, from business to healthcare to education. It is also raising new questions about the best way to build fairness, interpretability, privacy, and security into these systems.

Inclusive Datasets

- In 2018, Google organized the inclusive dataset competition.
- They released a 500K image dataset that had datapoints across the world.



Google AI Blog

The latest from Google Research

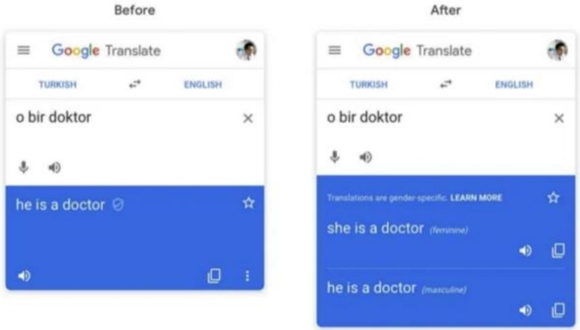
Introducing the Inclusive Images Competition

Thursday, September 6, 2018

Posted by Tulsee Doshi, Product Manager, Google AI

The release of large, publicly available image datasets, such as ImageNet, Open Images and Conceptual Captions, has been one of the factors driving the tremendous progress in the field of computer vision. While these datasets are a necessary and critical part of developing useful machine learning (ML) models, some open source data sets have been found to be geographically skewed based on how they were collected. Because the shape of a dataset informs what an ML model learns, such skew may cause the research community to inadvertently develop models that may perform less well on images drawn from geographical regions under-represented in those data sets. For example, the images below show one standard open-source image classifier trained on the Open Images dataset that does not properly apply "*wedding*" related labels to images of wedding traditions from different parts of the world.

Result: Reduced Gender Bias in Translate



Before

Google Translate

TURKISH ⇄ ENGLISH

o bir doktor

he is a doctor

After

Google Translate

TURKISH ⇄ ENGLISH

o bir doktor

Translations are gender-specific. LEARN MORE

she is a doctor (feminine)

he is a doctor (masculine)

Fairness Definitions

1. “Fairness through Unawareness”
    1. To avoid unfair decisions, prevent the model from every looking at protected attribute (e.g., race, gender).
    2. **Doesn’t work in practice**
2. Statistical Parity
    1. Idea: Equal performance across groups.

       $$\Pr\left(\hat{Y} = +\mid A = \blacksquare\right) = \Pr\left(\hat{Y} = +\mid A = \bigcirc\right)$$
    2. Also phrased as matching demographic statistics (e.g., if 33% of population are Circles, 33% of those admitted should be Circles).
3. Equal Opportunity
    1. Idea: True positive rate should be equal across groups

       $$\Pr\left(\hat{Y} = +\mid A = \blacksquare, Y = +\right) = \Pr\left(\hat{Y} = +\mid A = \bigcirc, Y = +\right)$$

Fairness Through Unawareness

Facebook Ads

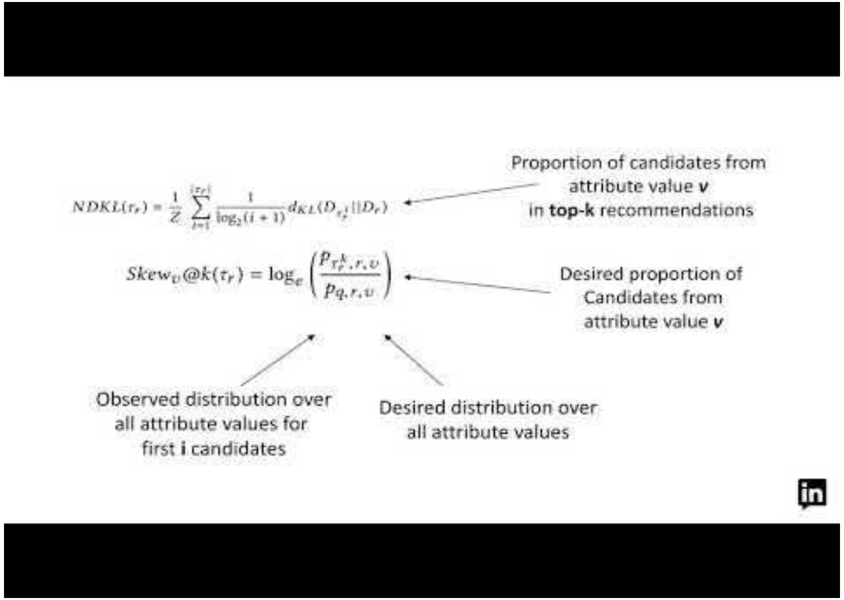
- Facebook gave in to the allegations and agreed to take steps:
  - Advertisers cannot target ads based on users’ age, gender, race, or zip code, and other membership categories.
    - ^^ **Fairness through Unawareness**
  - Require all advertisers to certify compliance with anti-discrimination laws.
  - Meet with ACLU members every 6 months for 3 years to enable them to monitor the reforms Facebook is undertaking.

https://www.technologyreview.com/2019/03/20/1225/facebook-is-going-to-stop-letting-advertisers-target-by-race-gender-or-age/

Statistical Parity & Equal Opportunity

Linkedin Ranking

- Linkedin used fairness constraints (equal opportunity and statistical parity) to rank candidates.
- Online A/B testing resulted in three-fold increase in fairness metrics while not affecting the business metrics.



$$NDKL(\tau_r) = \frac{1}{Z}\sum_{i=1}^{|\tau_r|}\frac{1}{\log_2(i+1)}d_{KL}(D_{\tau_r^i}||D_r)$$

$$Skew_v@k(\tau_r) = \log_e\left(\frac{p_{\tau_r^k,r,v}}{p_{q,r,v}}\right)$$

Proportion of candidates from attribute value **v** in **top-k** recommendations

Desired proportion of Candidates from attribute value **v**

Observed distribution over all attribute values for first **i** candidates

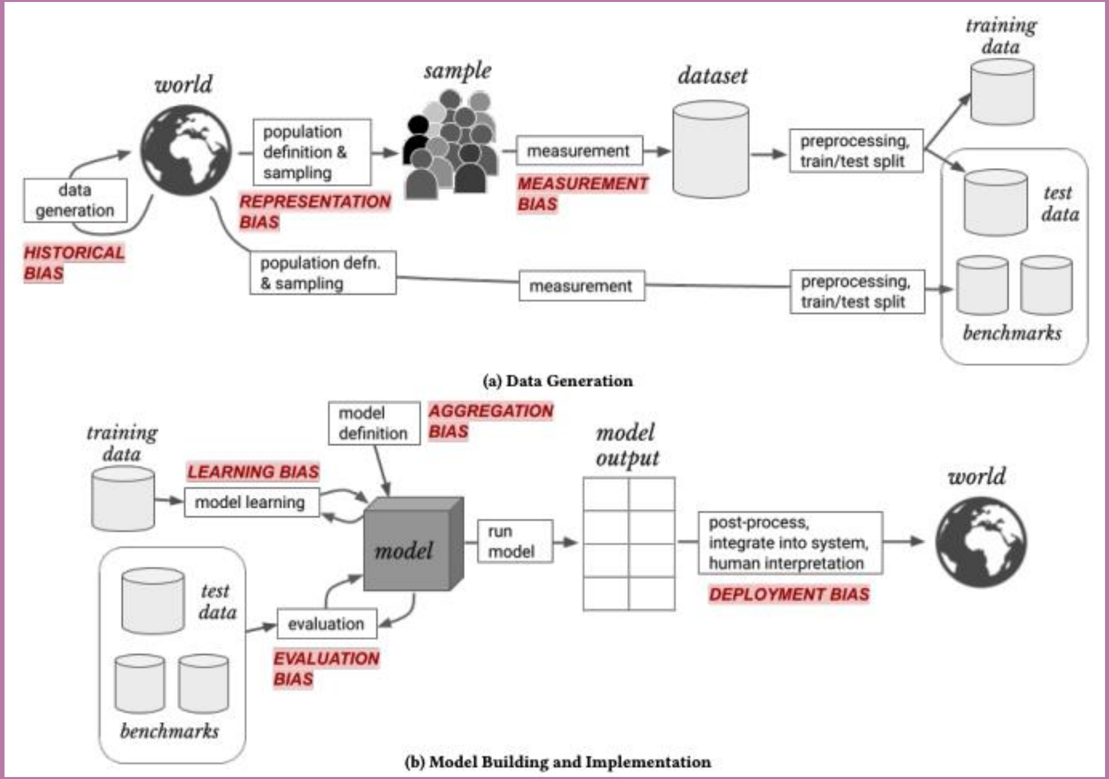
Desired distribution over all attribute values

Regulations

- EU has been championing the law landscape of AI regulation
- GDPR enacted in 2018 requires companies to ask consent for use of personal data and several other regulations.
- US has regulations on equal credit, housing, and job opportunities. But has lot of loopholes. It’s catching up!

Brainstorming Ways to Address Bias:

Case Studies



(a) Data Generation

(b) Model Building and Implementation

Task

In groups of 2-3, do the following for each case study (Suggested time: 5 minutes per):

1. Read through the goal, data, analysis, action, and constraints.
2. Note what biases that might come up in the scenario.
3. Propose solutions that would solve or mitigate the biases.

Case Study 1: Student Support Programs

Goal: Improve graduation rates for students

Data: Student records from different school districts and states, National Student Clearinghouse data (which gives us information about college outcomes)

Analysis: Predict risk of not graduating on time

Actions: Assign after-school programs to most at-risk students

Constraints: Resources are available to target additional tutoring to 10% of students

Case Study 2: Loans

Goal: Provide loans while balancing repayment rates for bank loans

Data: Historical loans and payments, credit reporting data, background checks

Analysis: Build model to predict risk of not repaying on time

Actions: Deny loan or increase interest rate/penalties

Case Study 3: Disaster Relief

Goal: Accurately assess damage and send appropriate relief resources

Data: Twitter posts, Facebook posts geocoded with lat-long within disaster area and keywords/hashtags related to the storm

Analysis: Intensity and type of damage by neighborhood

Actions: Assessment and allocate relief effort (type and amount)

Constraints: Limited resources for relief efforts

Recap

Theme: Thinking about solutions for bias and how to encourage ethical machine learning

Approaches towards mitigating bias:

- Not developing the technology
- Educate software engineers
- Using inclusive datasets
- Applying fairness definitions
  - Fairness through Unawareness
  - Statistical Parity
  - Equal Opportunity