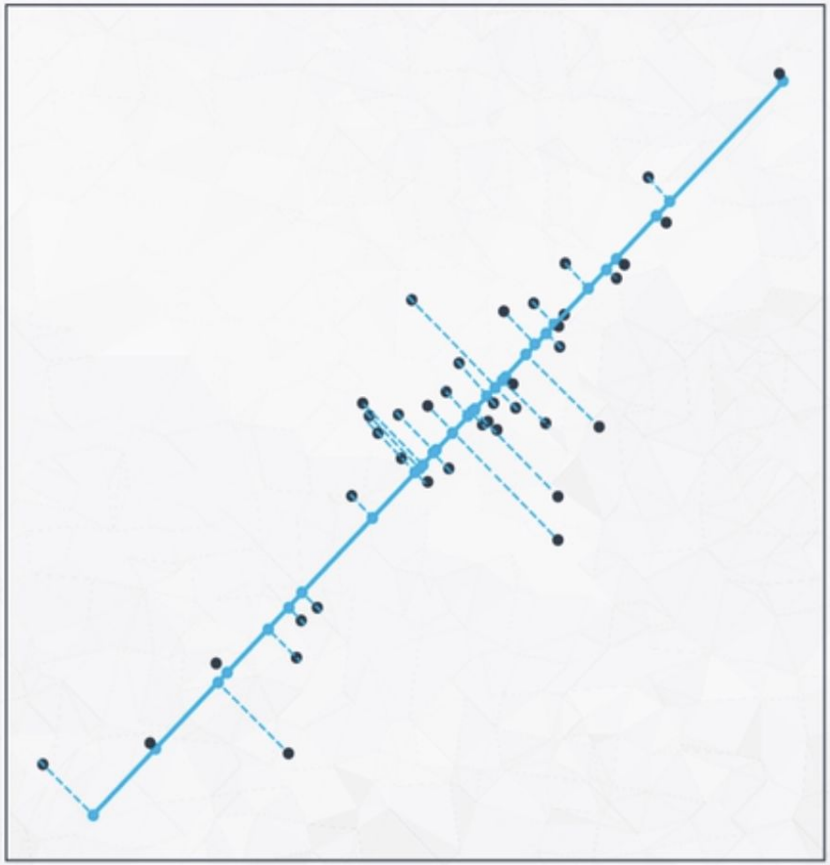
Two Types of Unsupervised Learning

Clustering

Unsupervised learning task concerned with putting similar data into groups

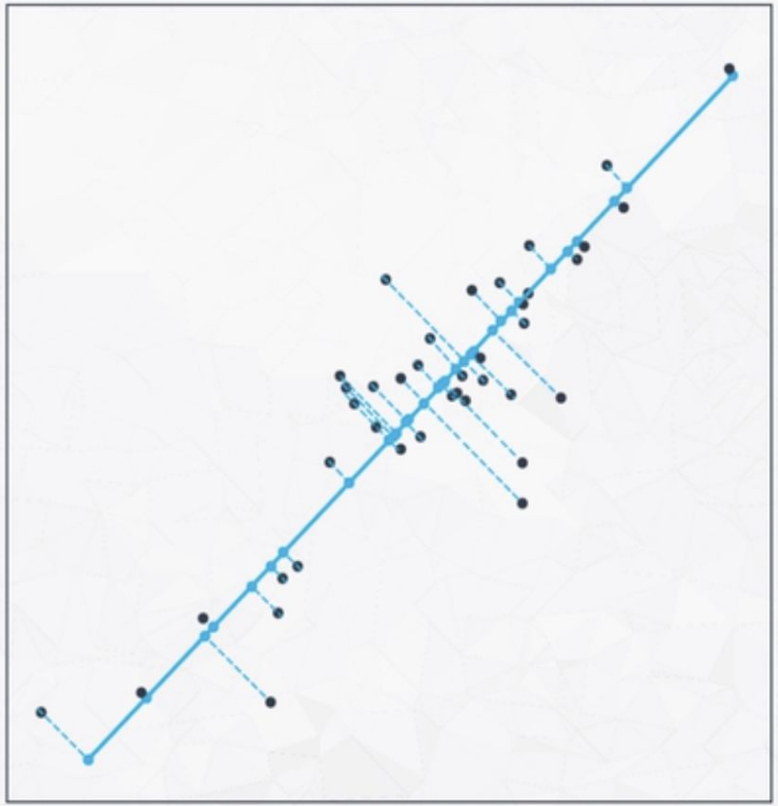
Dimensionality Reduction

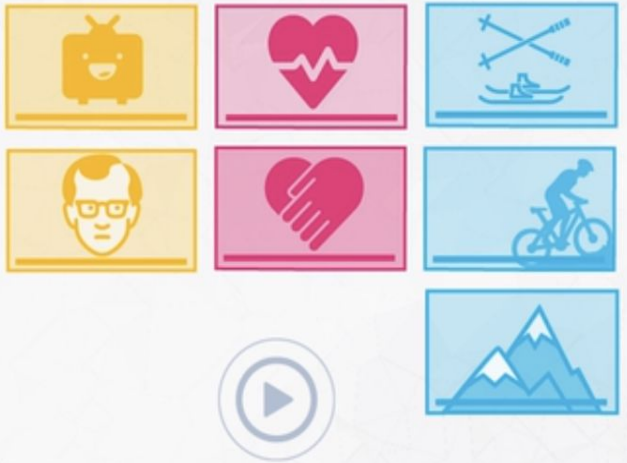
Unsupervised learning task concerned with reducing the number of variables used



Clustering

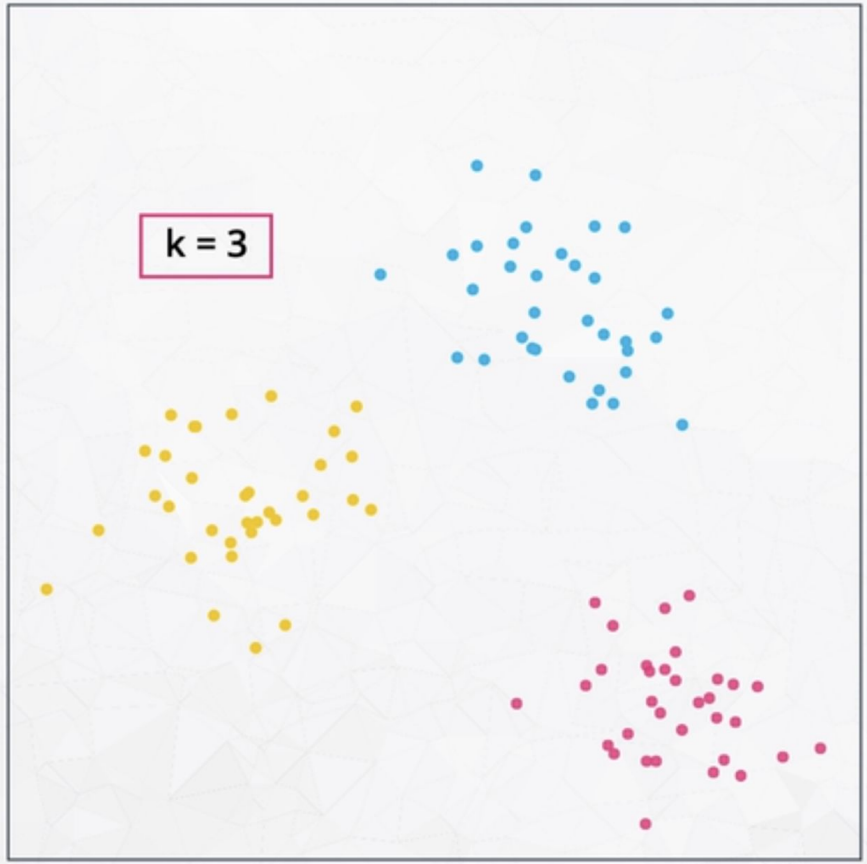
Dimensionality Reduction



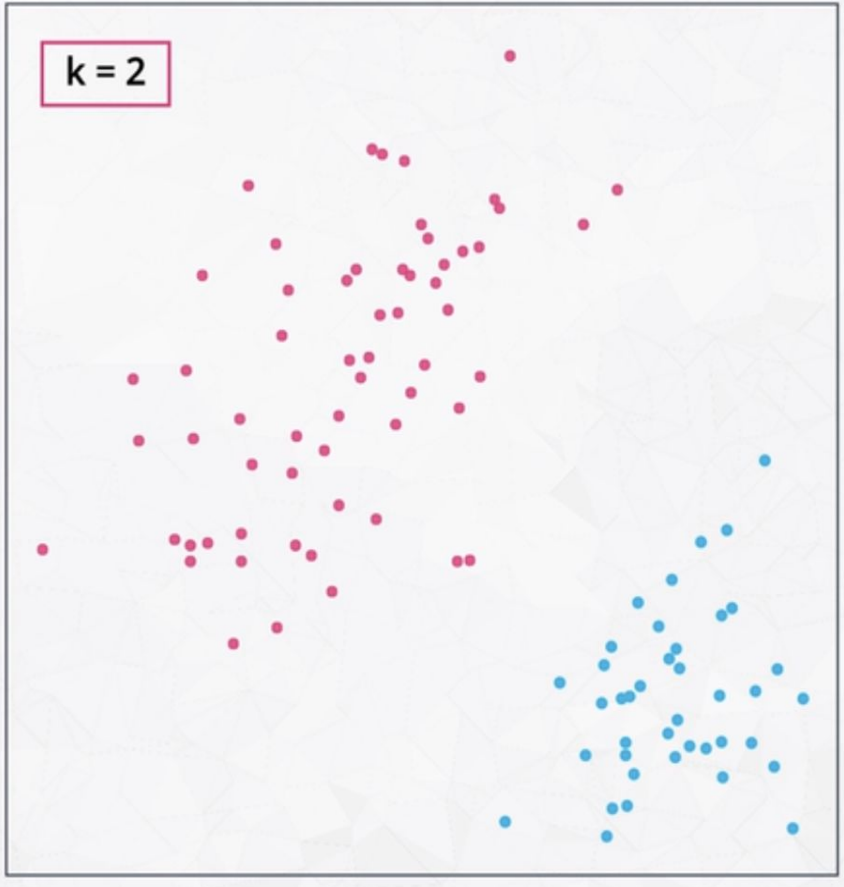



Awesomeness
Travel Distance

Changing k


k = 3


k = 3

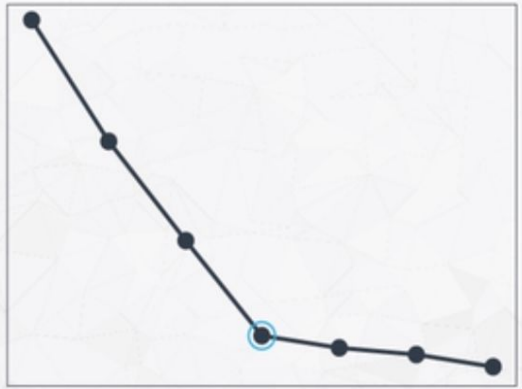

k = 2

x_1	x_2	x_3	...	x_n

Prior Knowledge

$k = ?$

Decision Methods



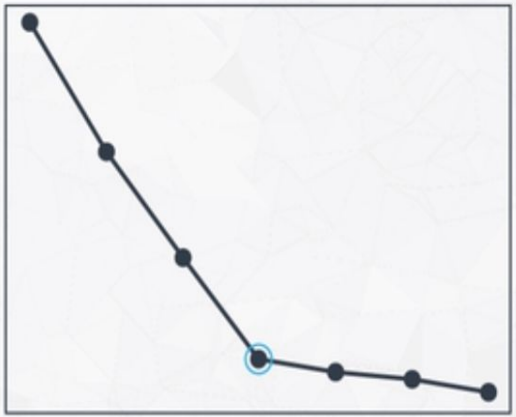
Elbow Method for Finding k

k = ?

Visualizations

Prior Knowledge

Decision Methods



Elbow Method

k = 1: avg. dist = 1.261

k = 2: avg. dist = 0.923

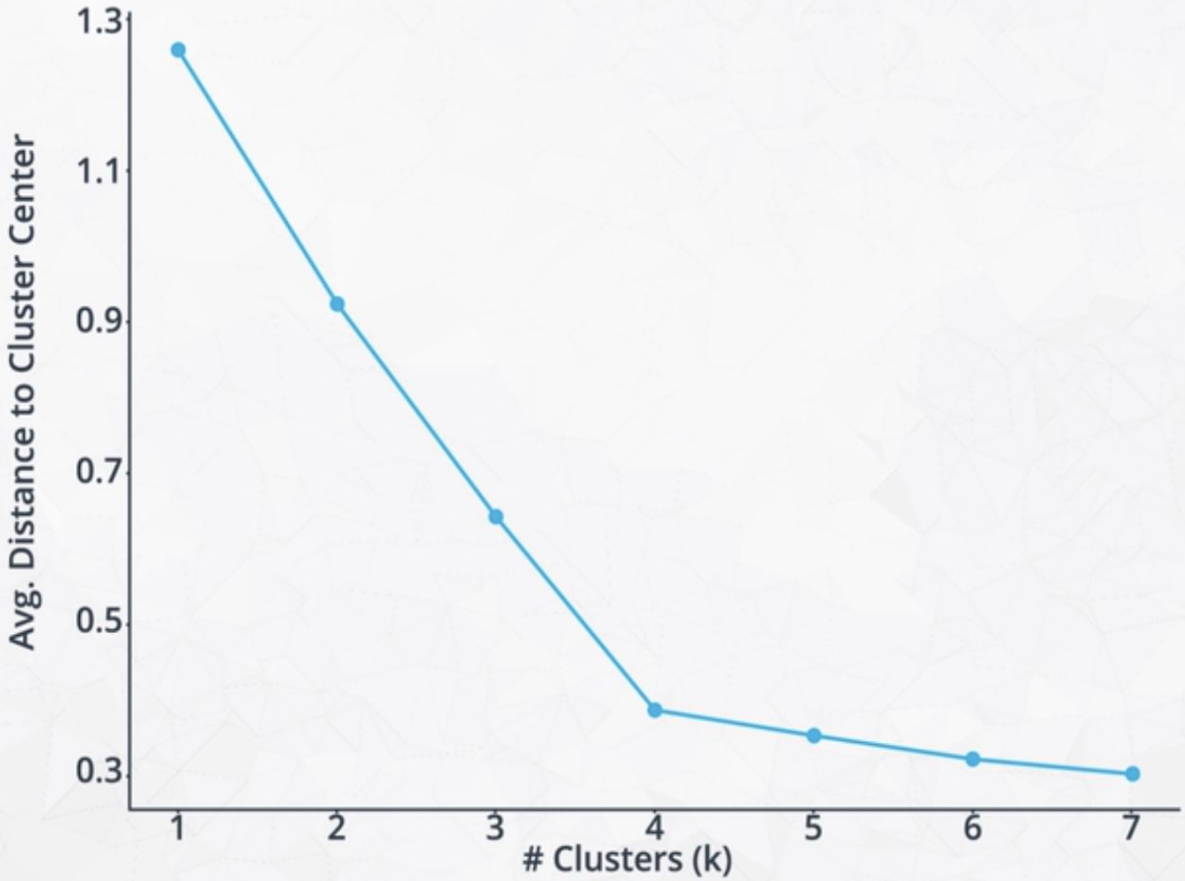
k = 3: avg. dist = 0.639

k = 4: avg. dist = 0.382

k = 5: avg. dist = 0.348

k = 6: avg. dist = 0.318

k = 7: avg. dist = 0.298

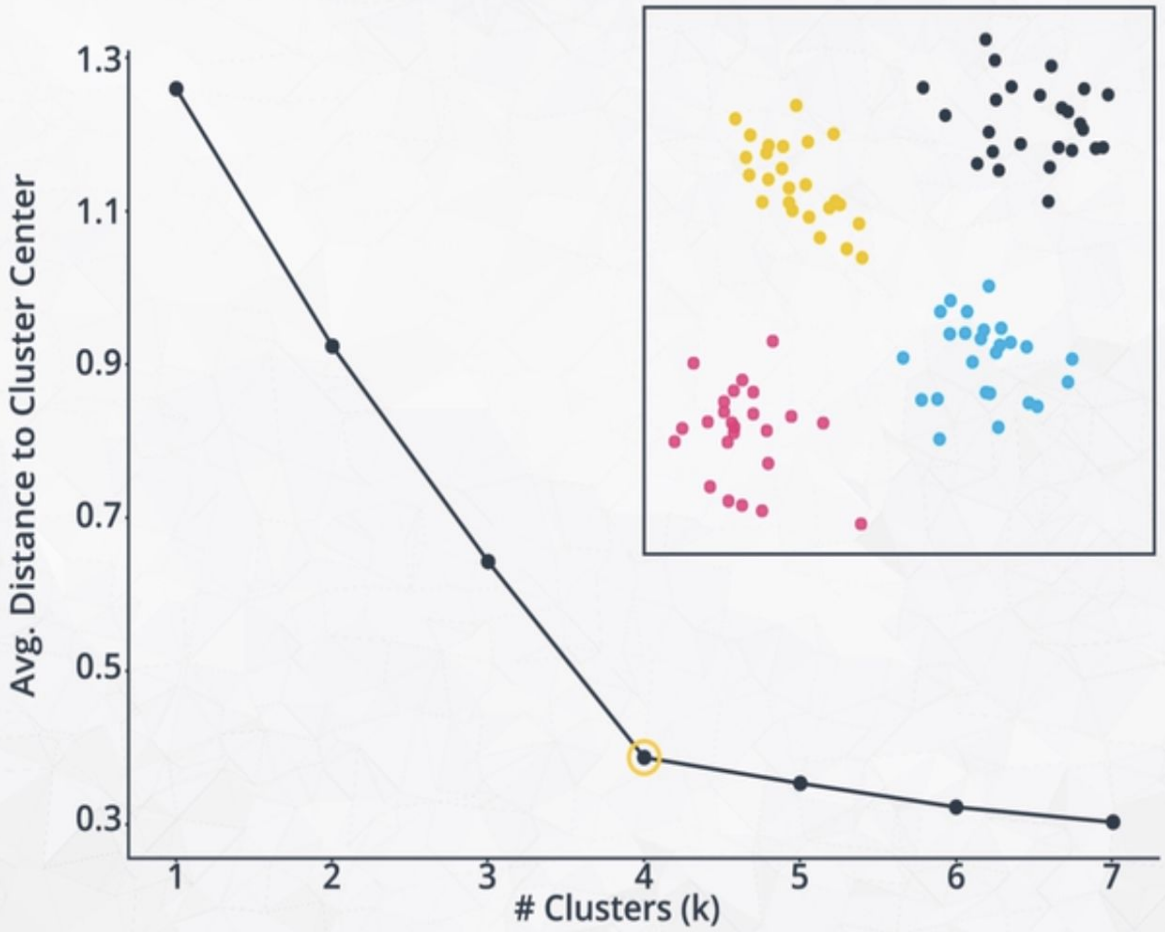


Elbow Method

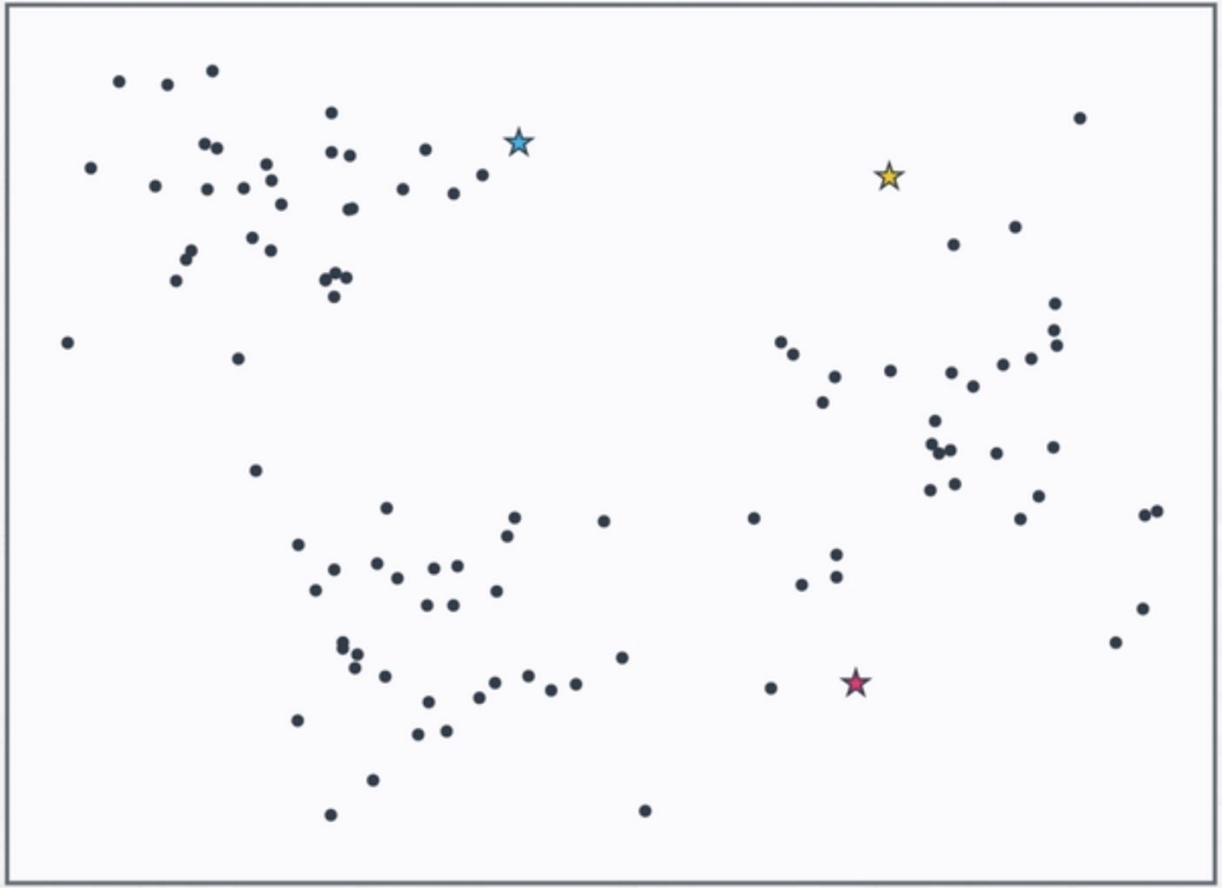
Large decreases at small k

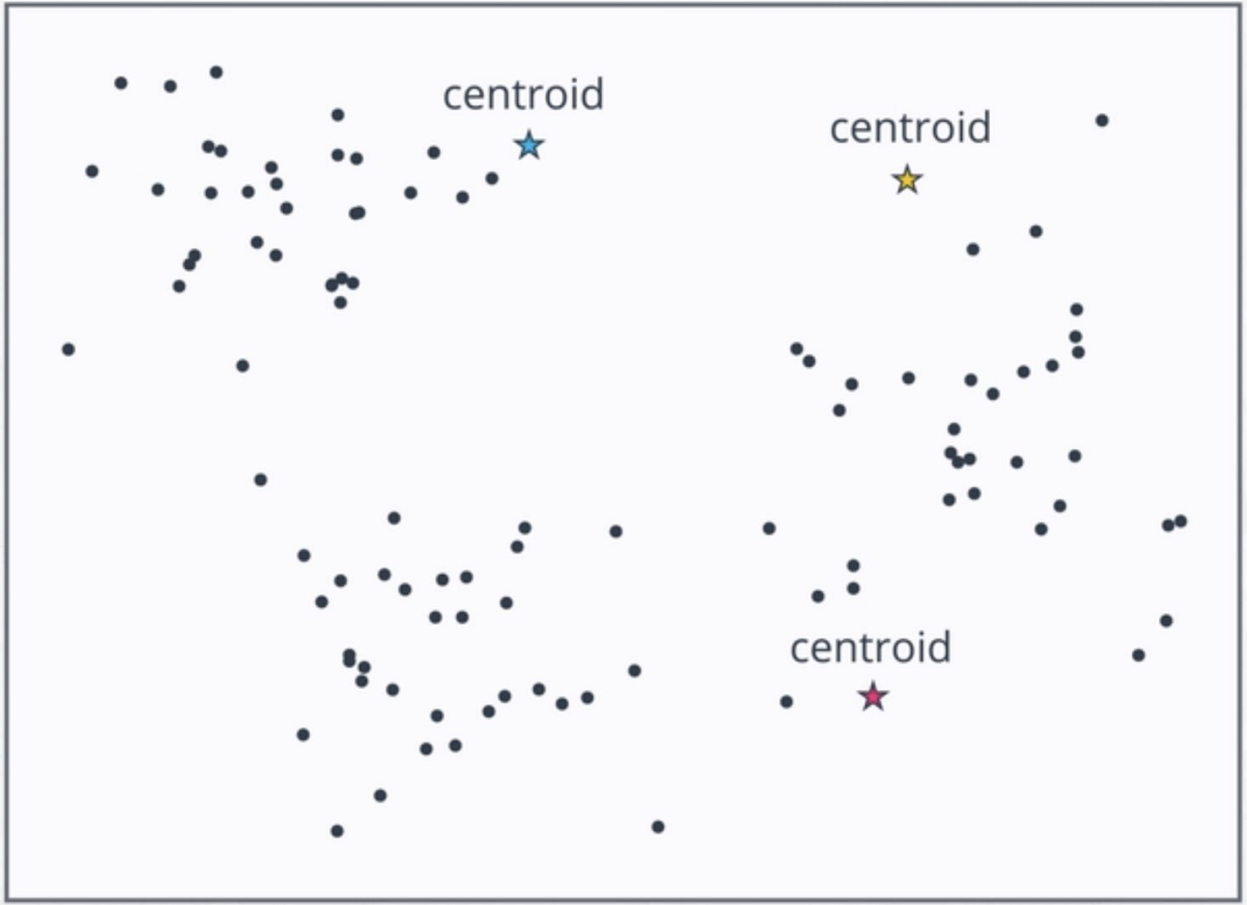
Small decreases at large k

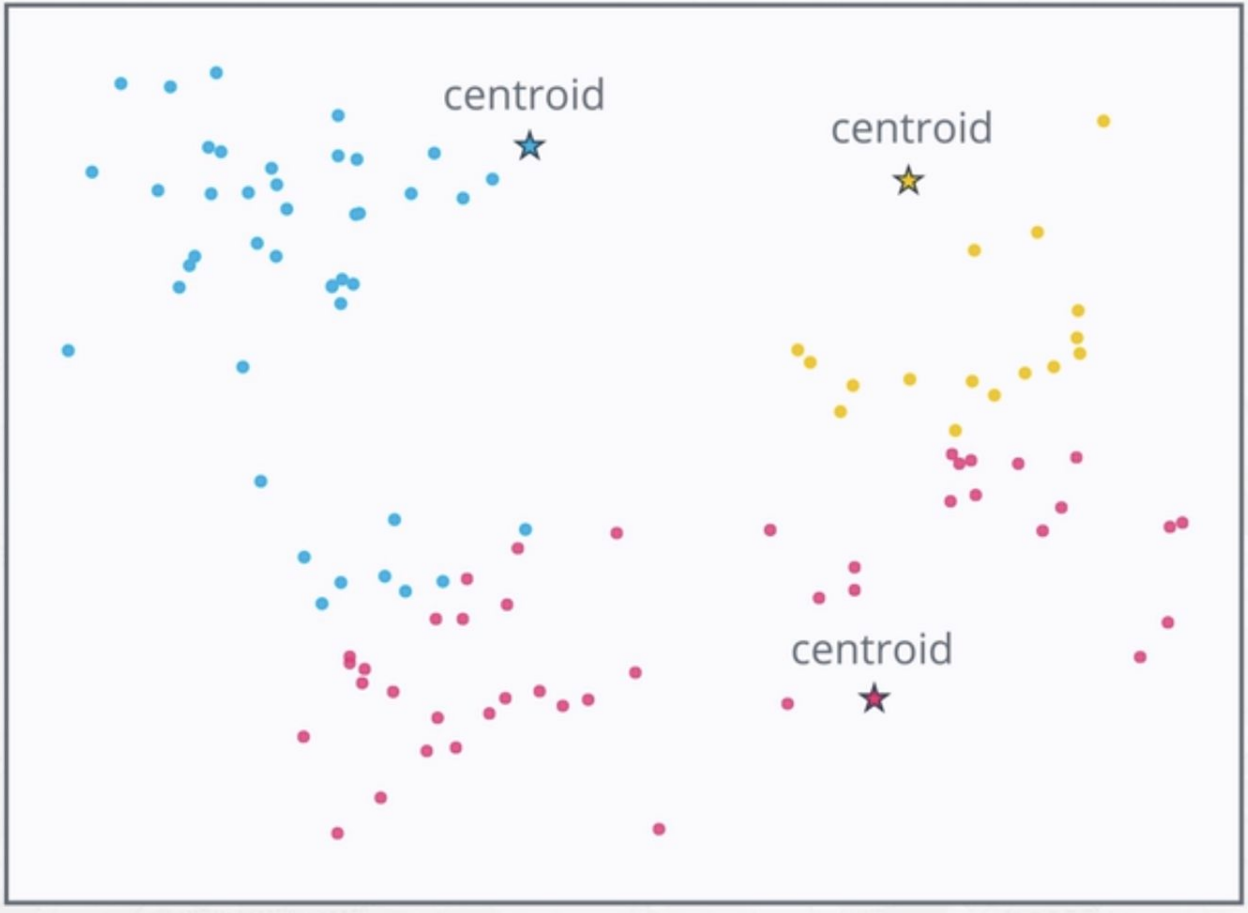
Best choice for k around the 'elbow' of the curve

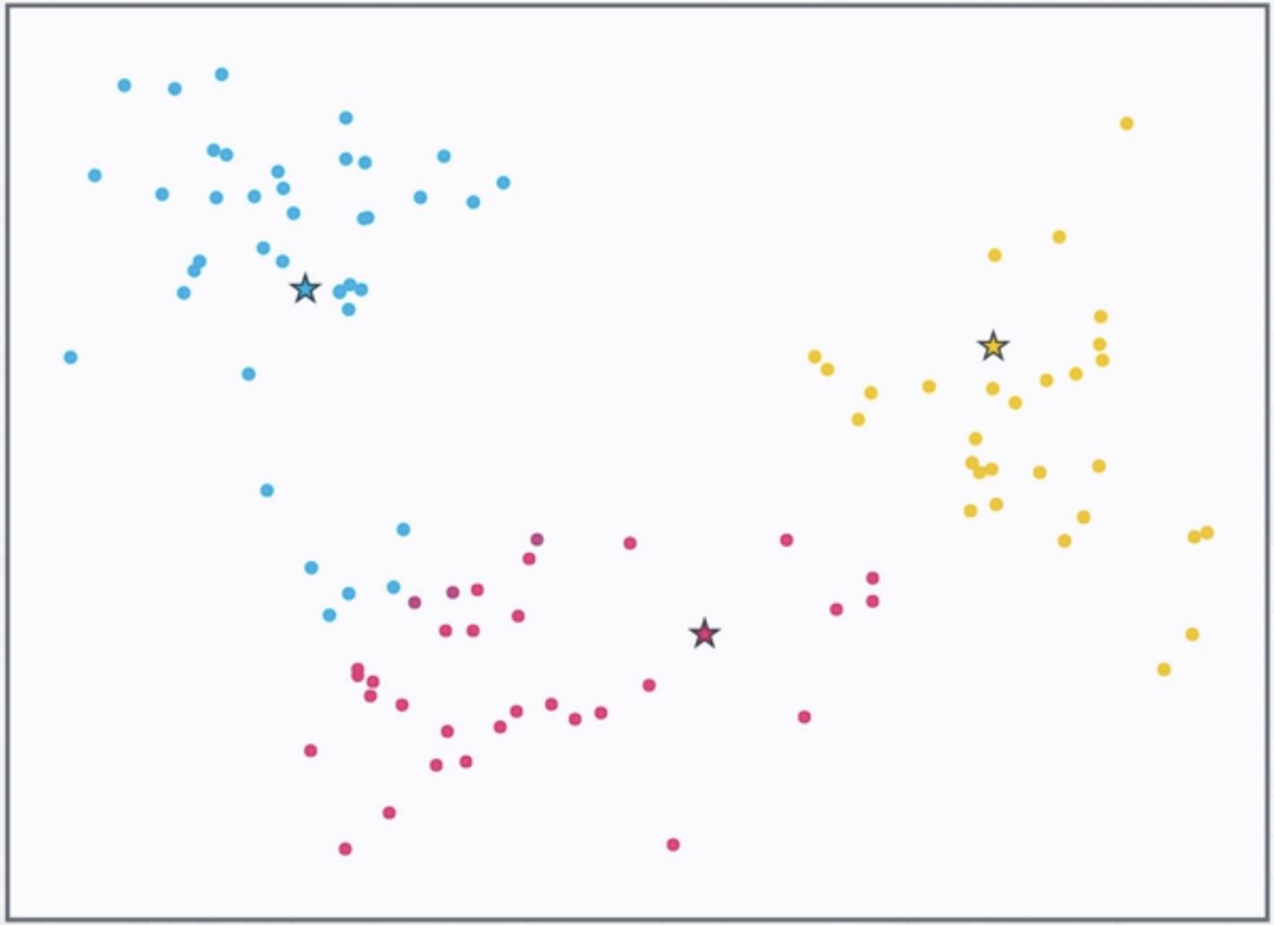


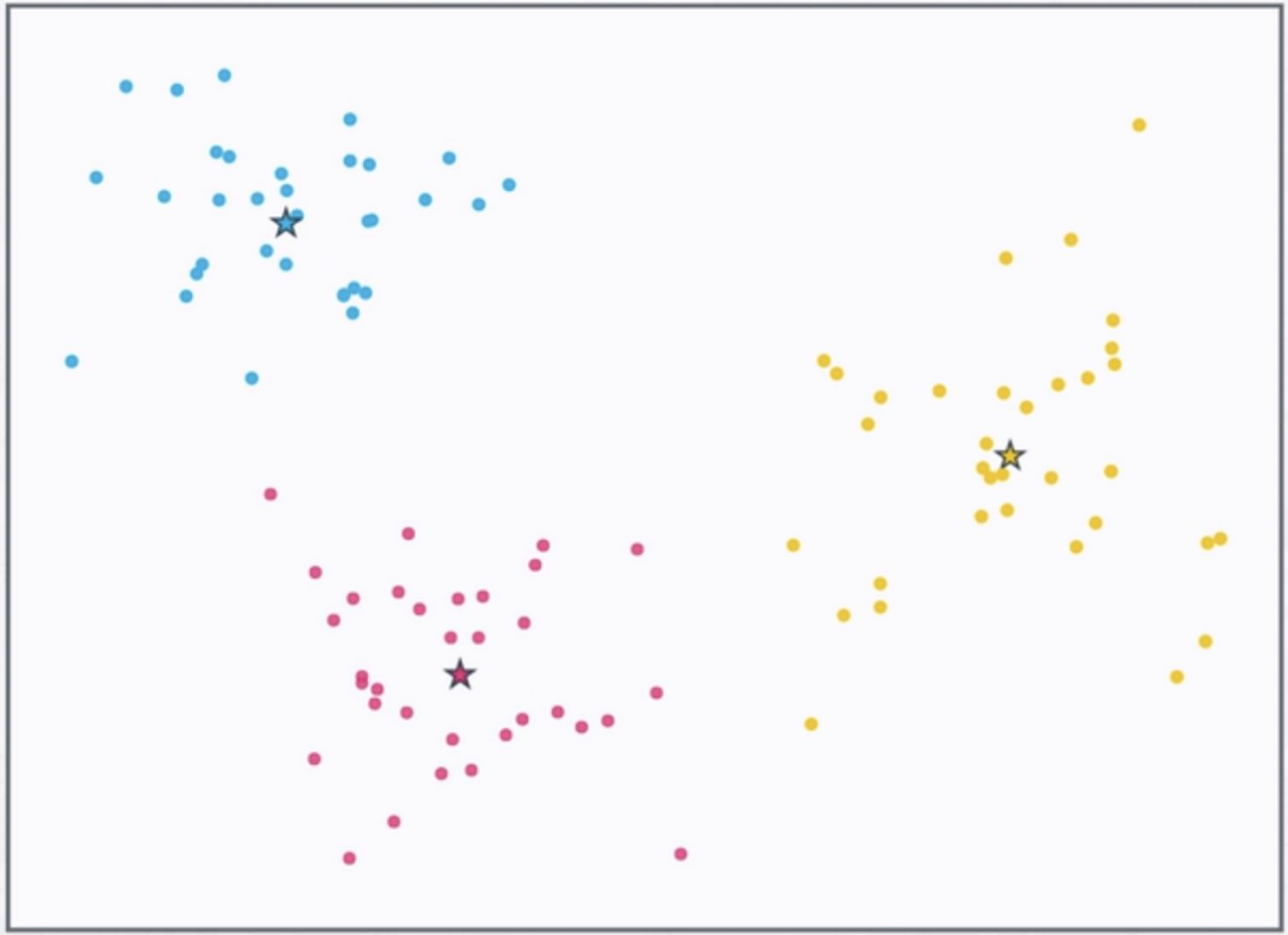
How Does k-means Work?




centroid
centroid
centroid

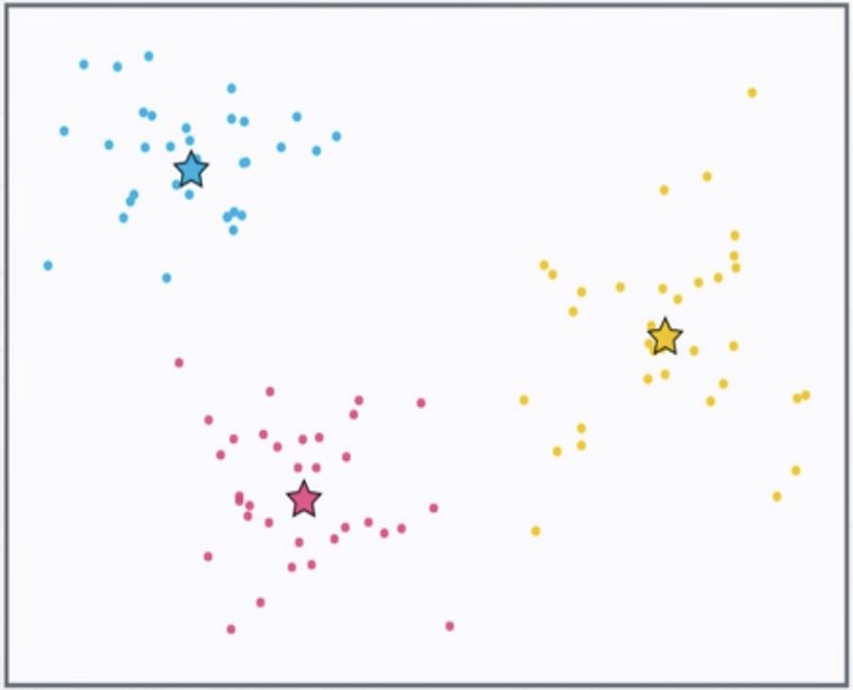

centroid
centroid
centroid

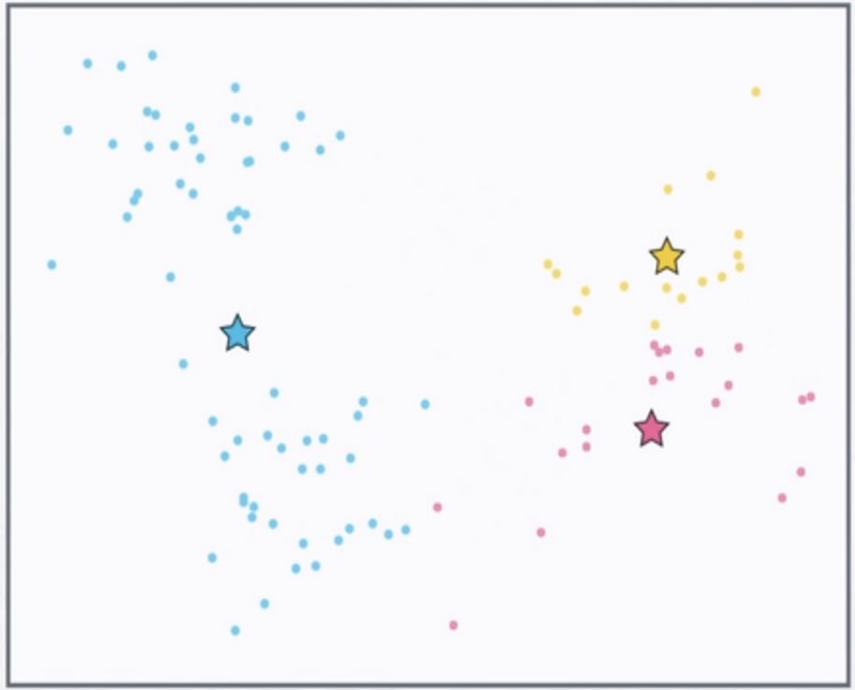




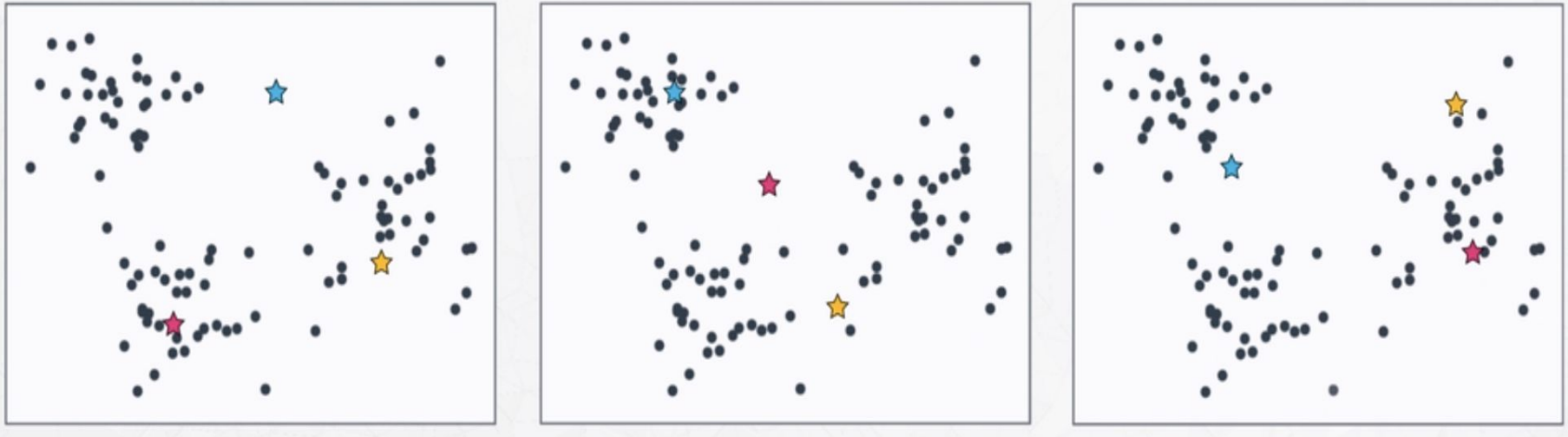
Is That the Optimal Solution?

Example: Different Starting Points May Lead to Different Final Clustering Results





Use **repeated runs** to protect against **local minima**

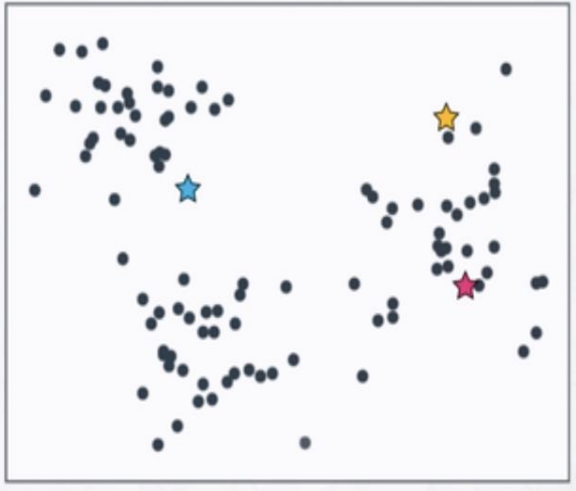


Use **repeated runs** to protect against **local minima**

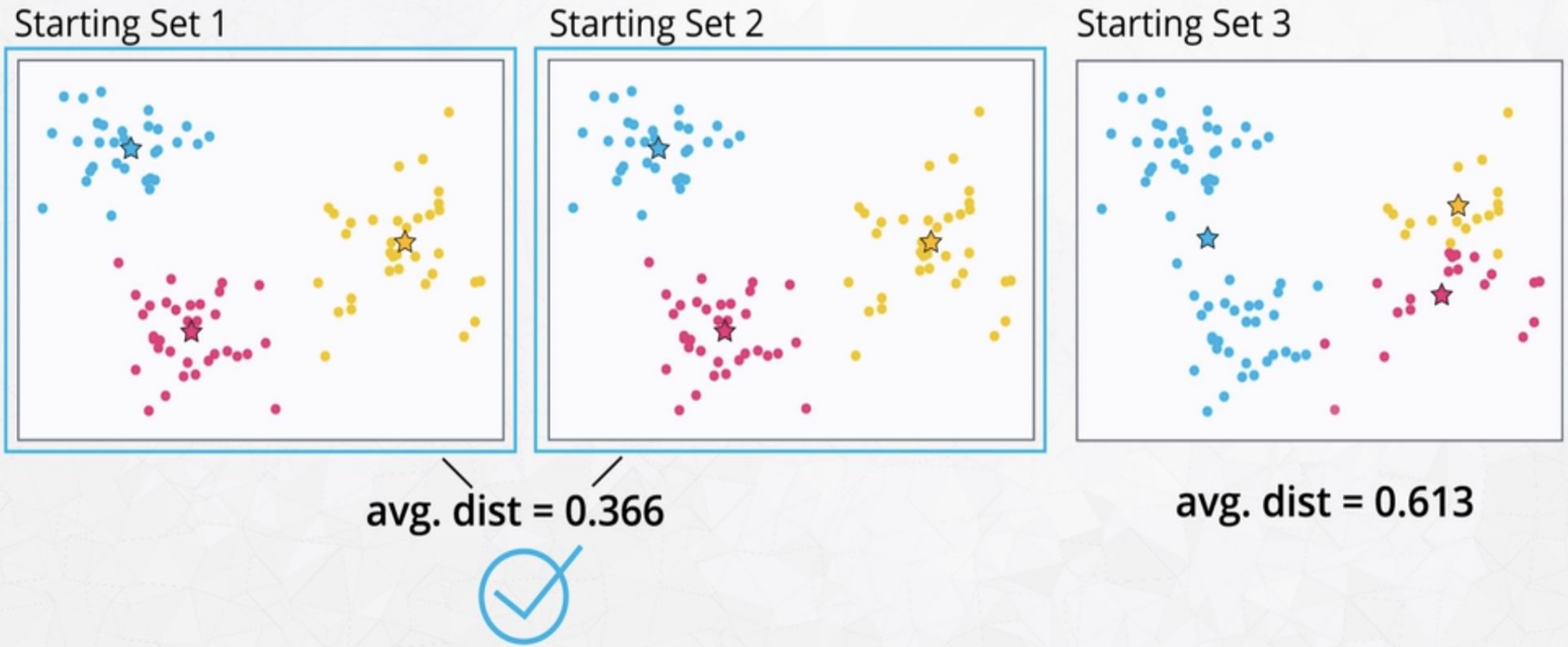
Starting Set 1

Starting Set 2

Starting Set 3



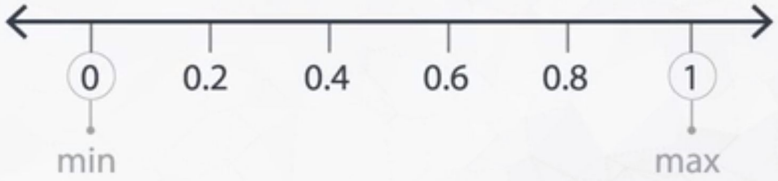
Use **repeated runs** to protect against **local minima**



Feature Scaling

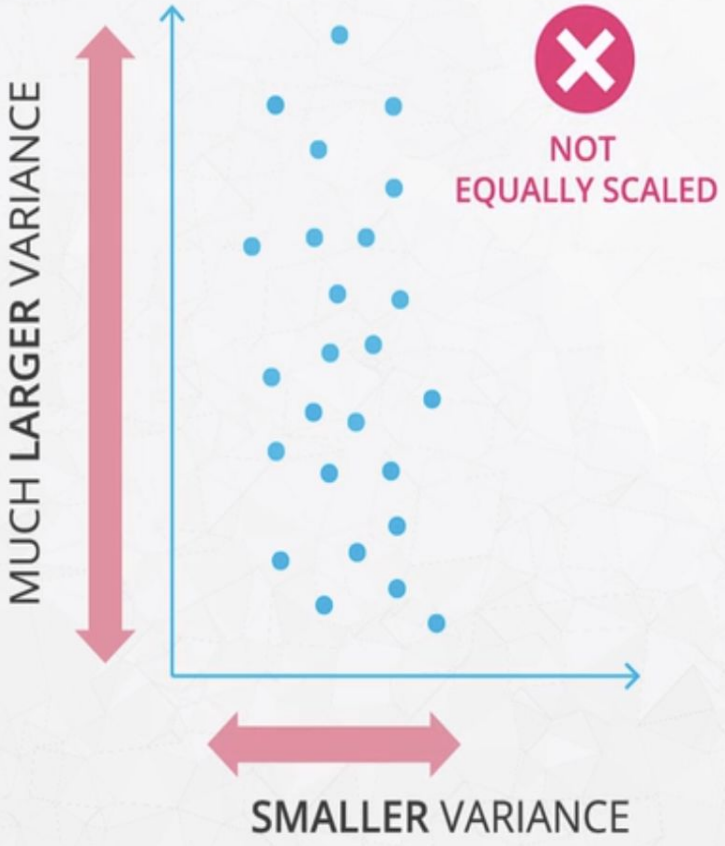

STANDARDIZING
-2
-1
0
1
2
z-scores

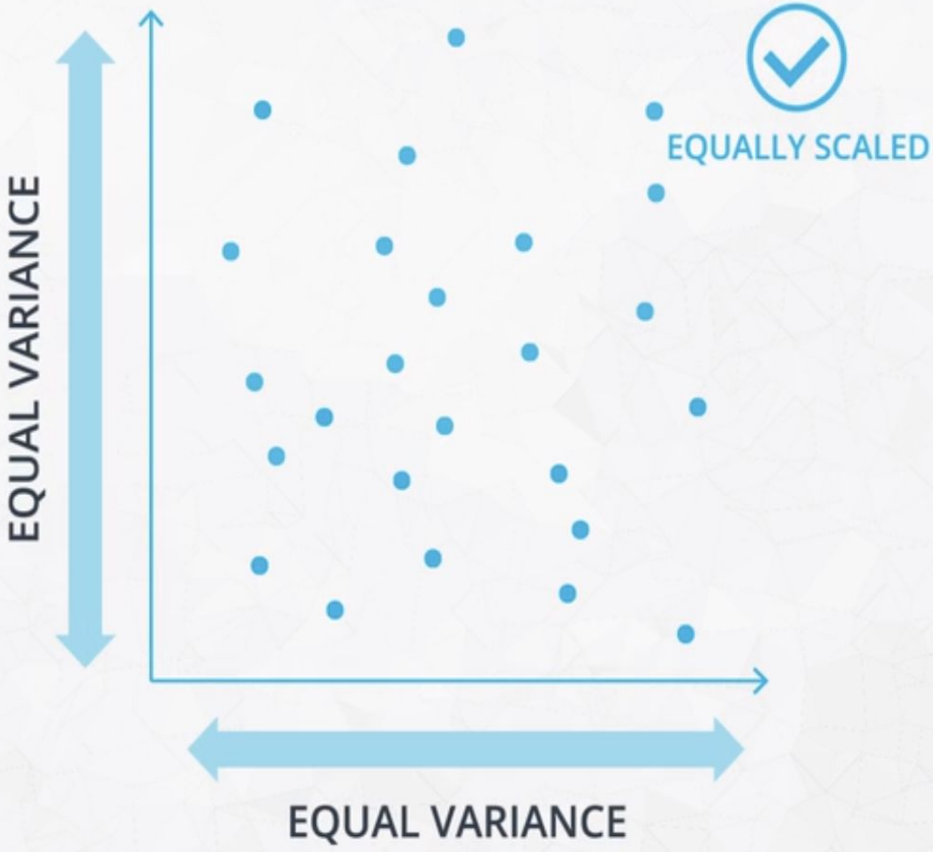
NORMALIZING



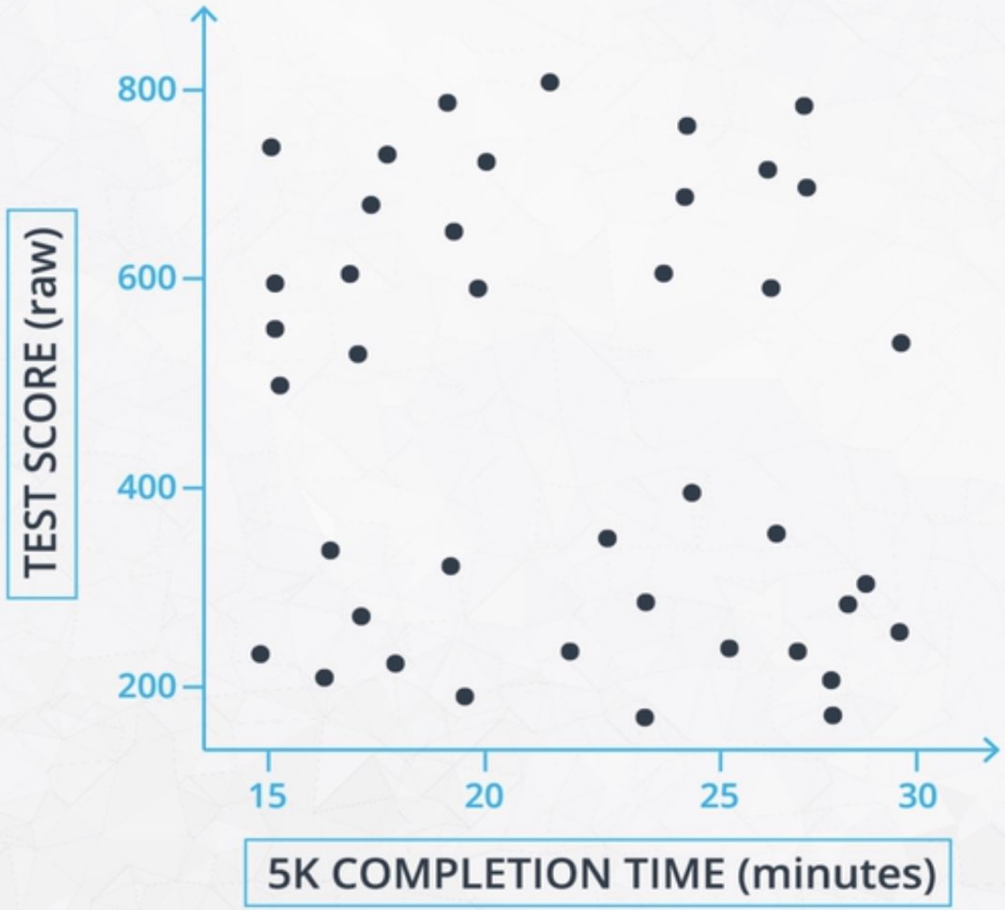
original

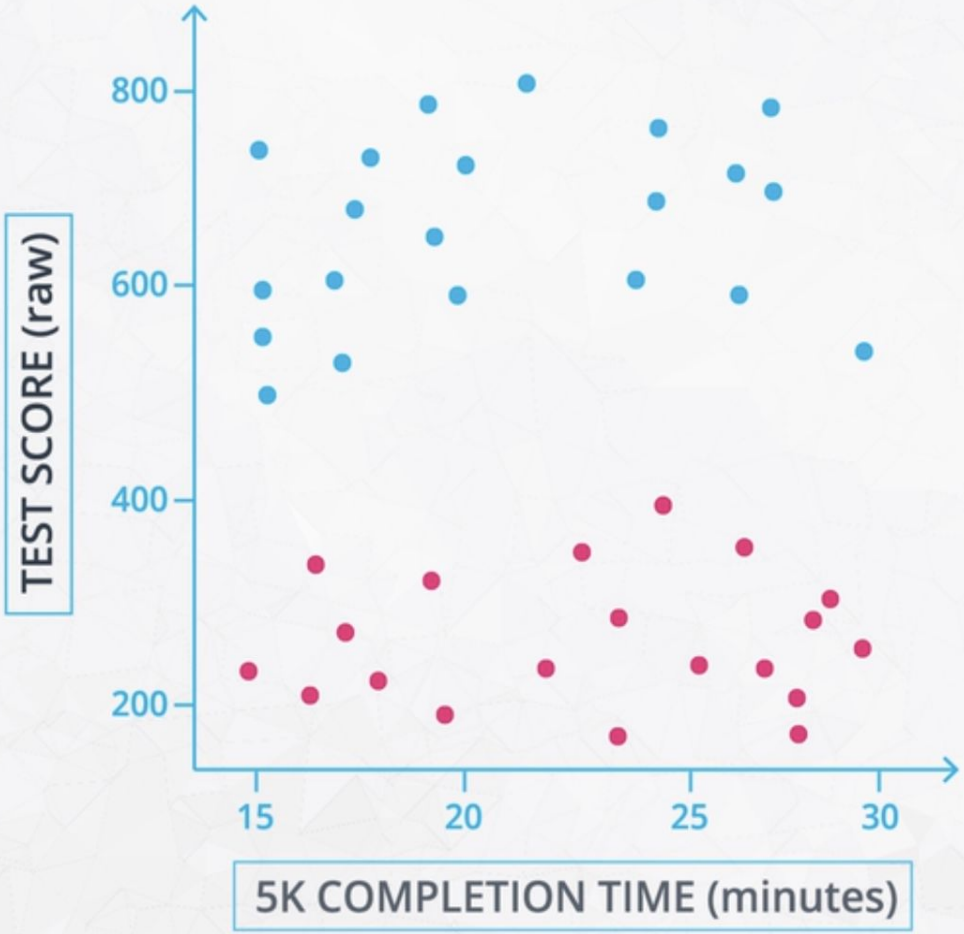
color scaled

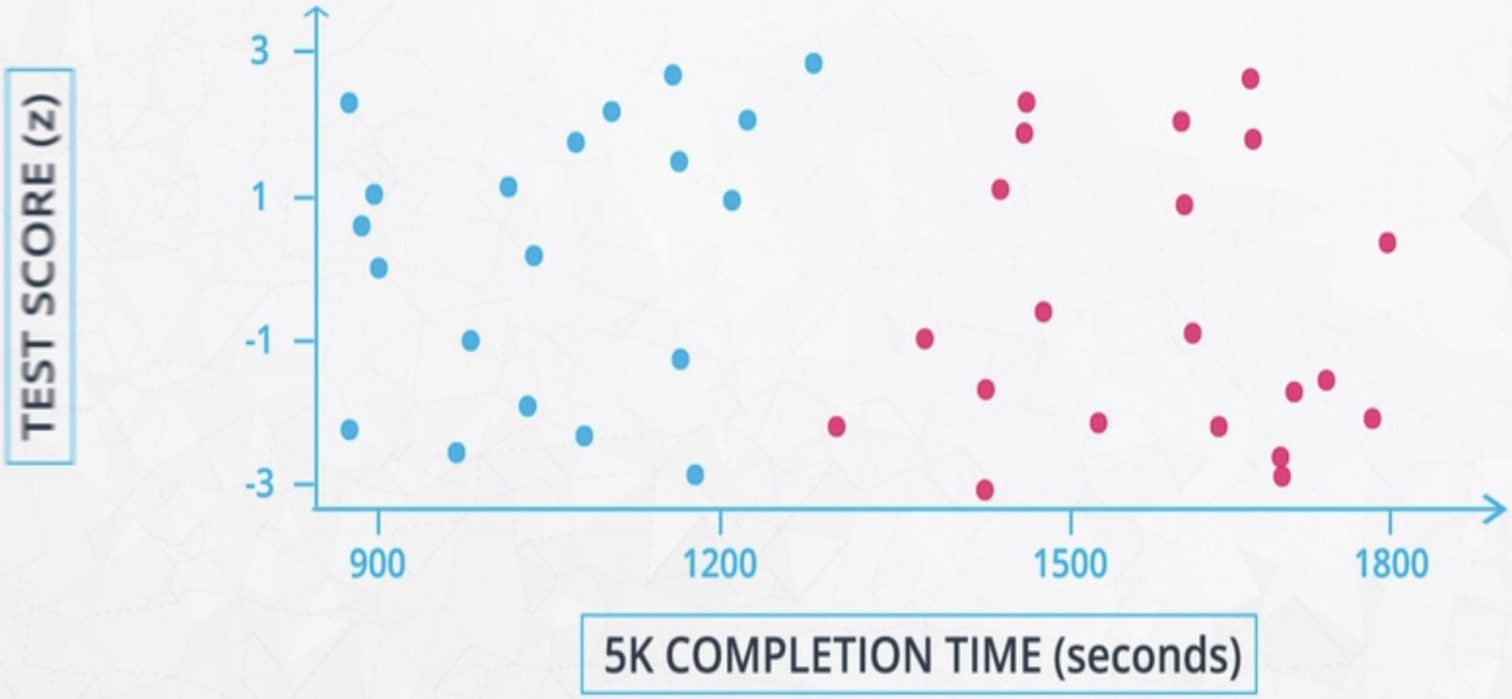

MUCH LARGER VARIANCE
NOT
EQUALLY SCALED
SMALLER VARIANCE


EQUAL VARIANCE
EQUALLY SCALED
EQUAL VARIANCE

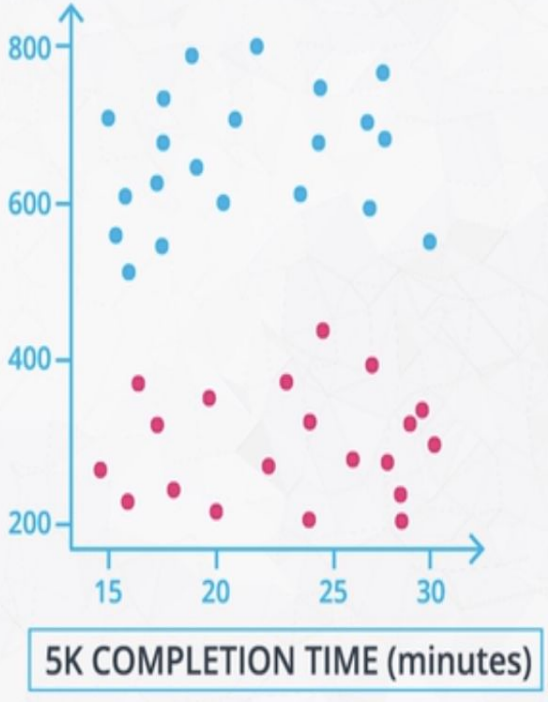
Feature Scaling Example

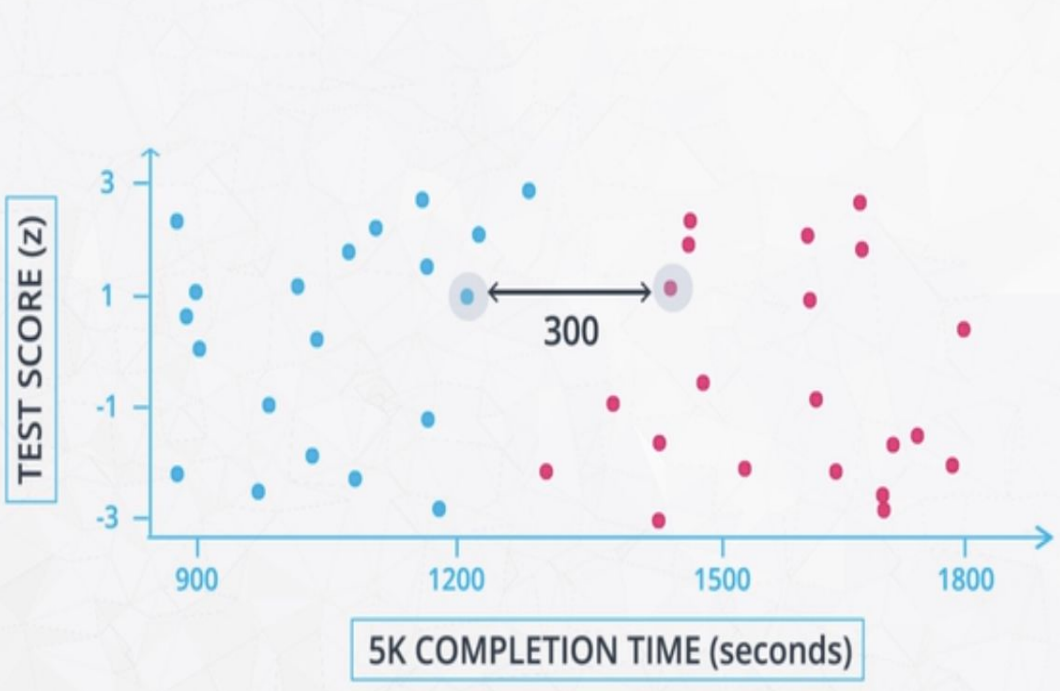

800
600
400
200
TEST SCORE (raw)
15
20
25
30
5K COMPLETION TIME (minutes)

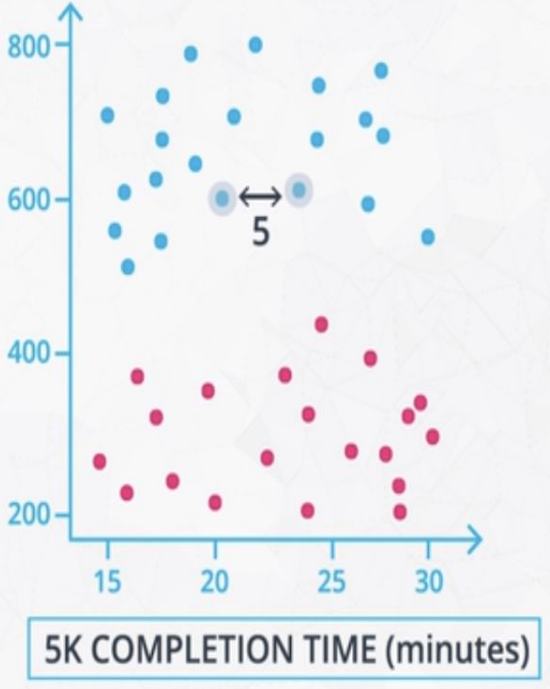

800
600
400
200
TEST SCORE (raw)
15
20
25
30
5K COMPLETION TIME (minutes)

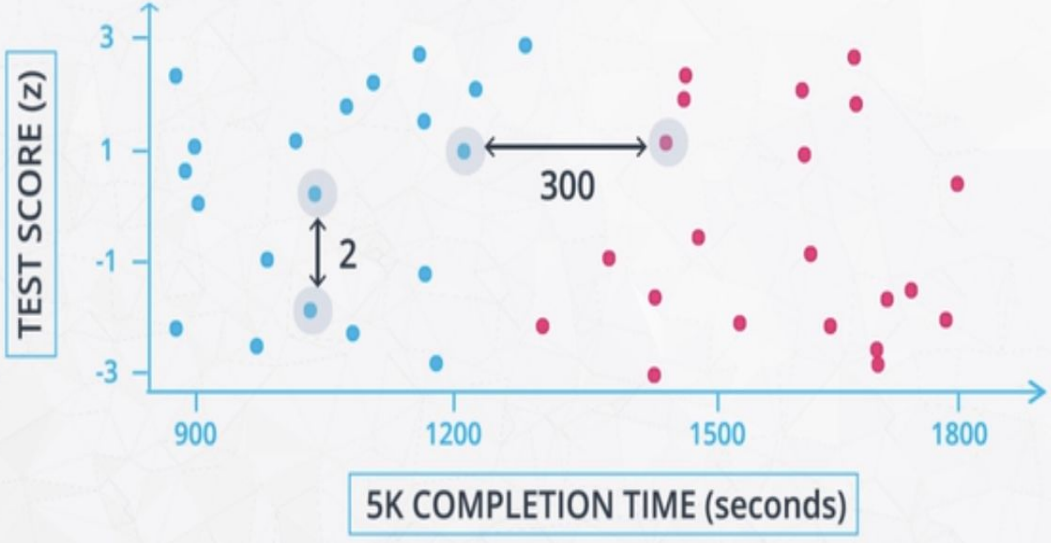

TEST SCORE (z)
3
1
-1
-3
900
1200
1500
1800
5K COMPLETION TIME (seconds)

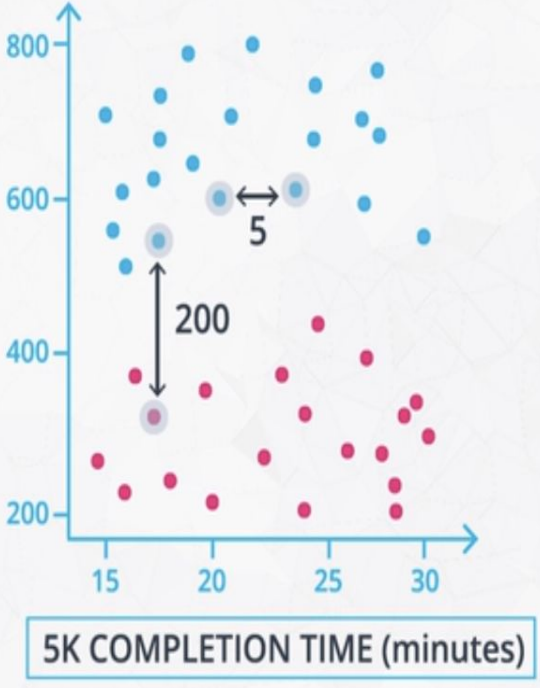

TEST SCORE (z)
3
1
-1
-3
900
1200
1500
1800
5K COMPLETION TIME (seconds)


800
600
400
200
15
20
25
30
5K COMPLETION TIME (minutes)


TEST SCORE (z)
3
1
-1
-3
300
900
1200
1500
1800
5K COMPLETION TIME (seconds)


800
600
400
200
5
15
20
25
30
5K COMPLETION TIME (minutes)

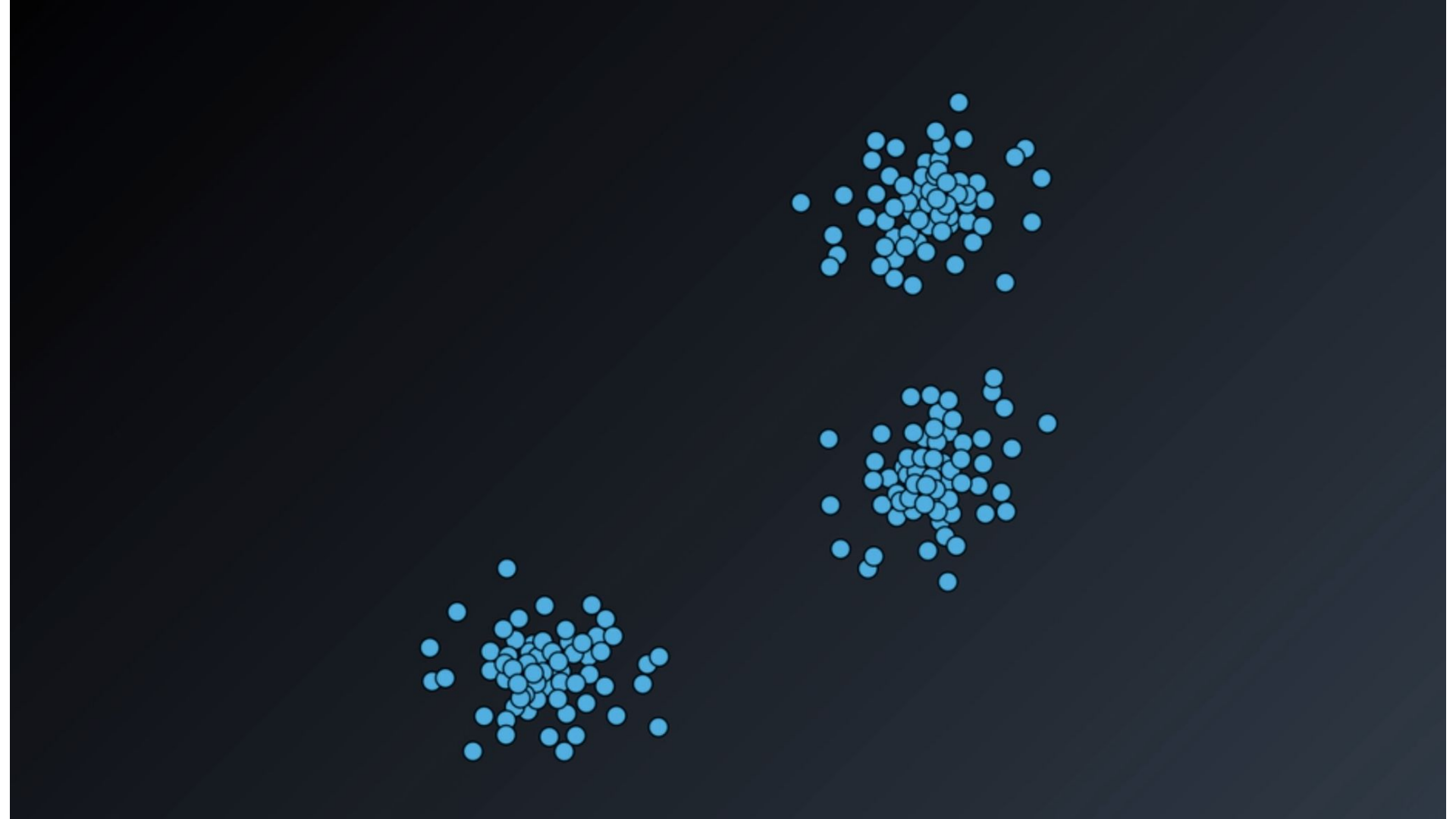

TEST SCORE (z)
3
1
-1
-3
2
300
900
1200
1500
1800
5K COMPLETION TIME (seconds)

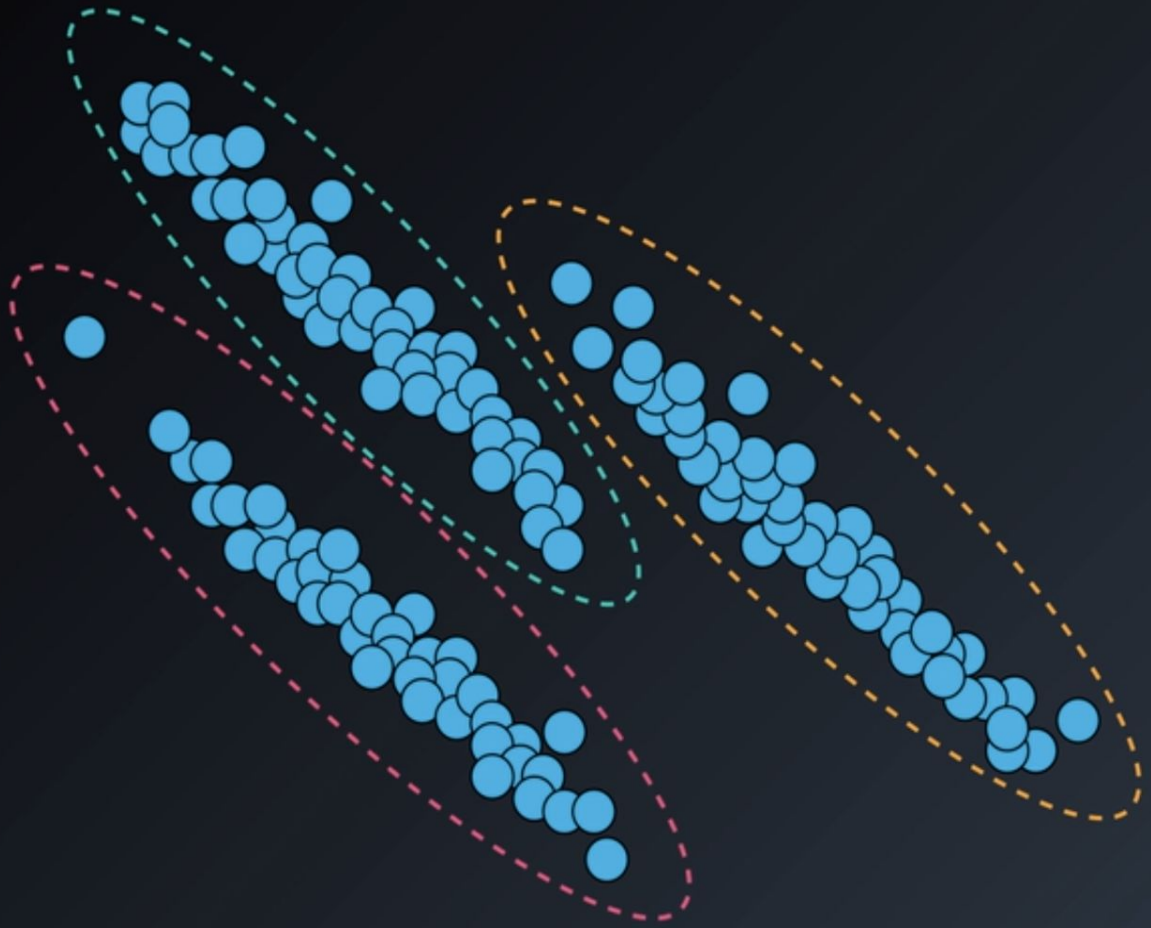

800
600
400
200
5
200
15
20
25
30
5K COMPLETION TIME (minutes)

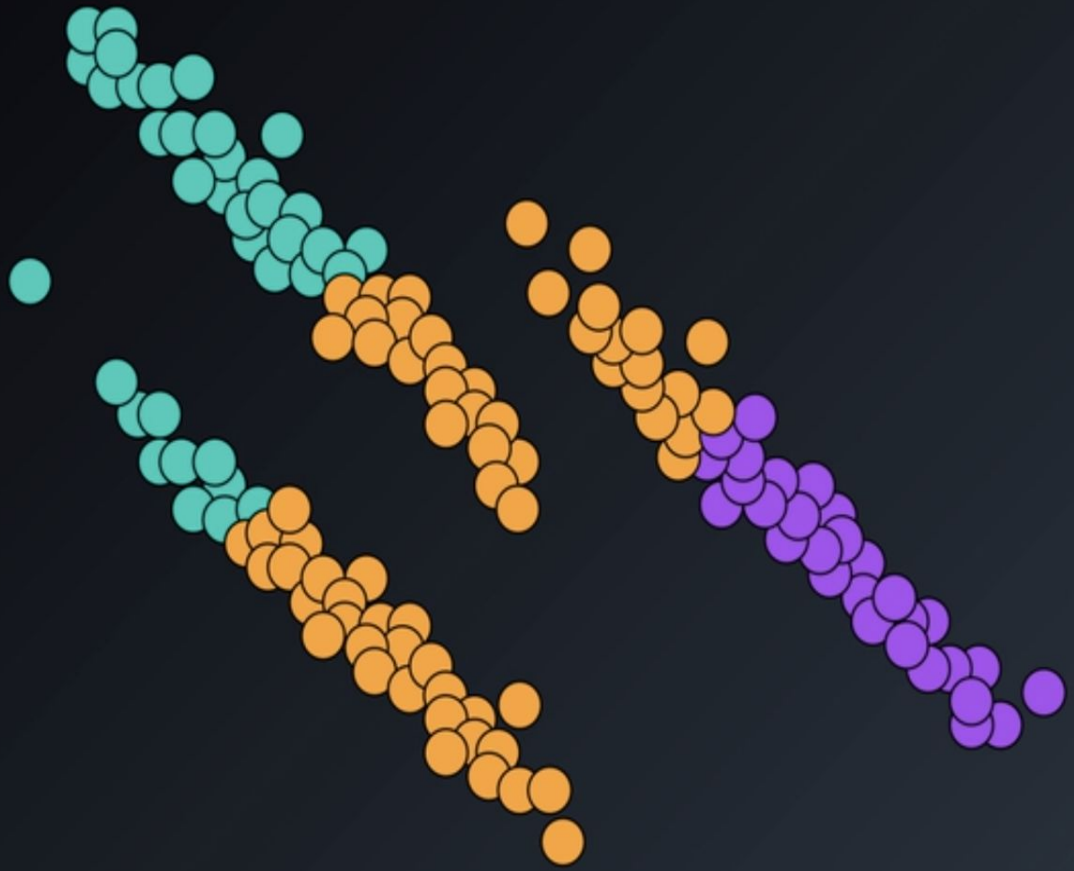
MLND - Unsupervised Learning

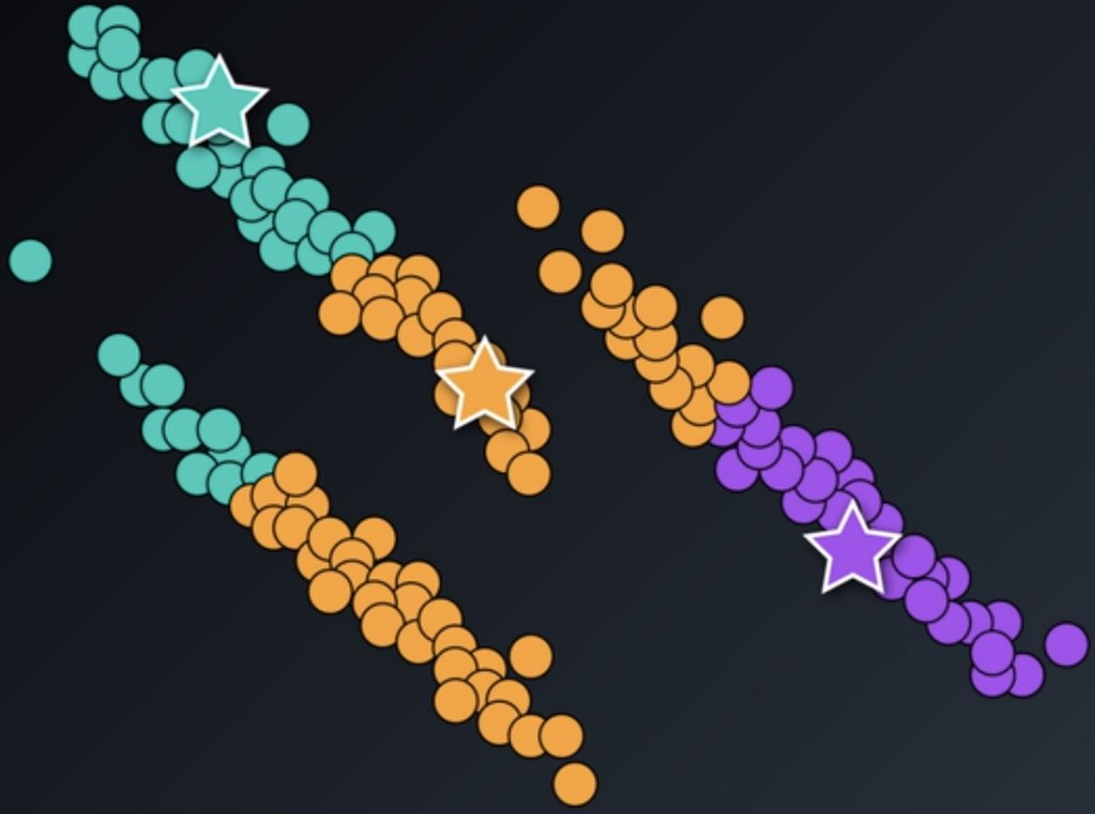
LESSON 2 - CLUSTERING 2

1- K-MEANS CONSIDERATIONS







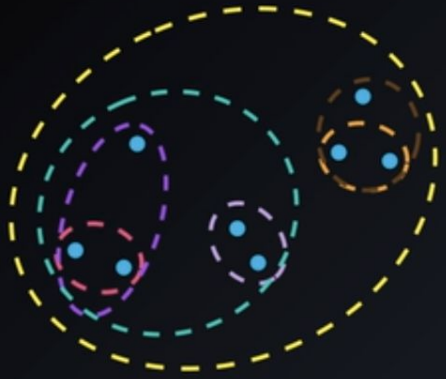


DATASET

K-MEANS CLUSTERING

k = 3 k = 2 k = 2 k = 3 k = 3 k = 3

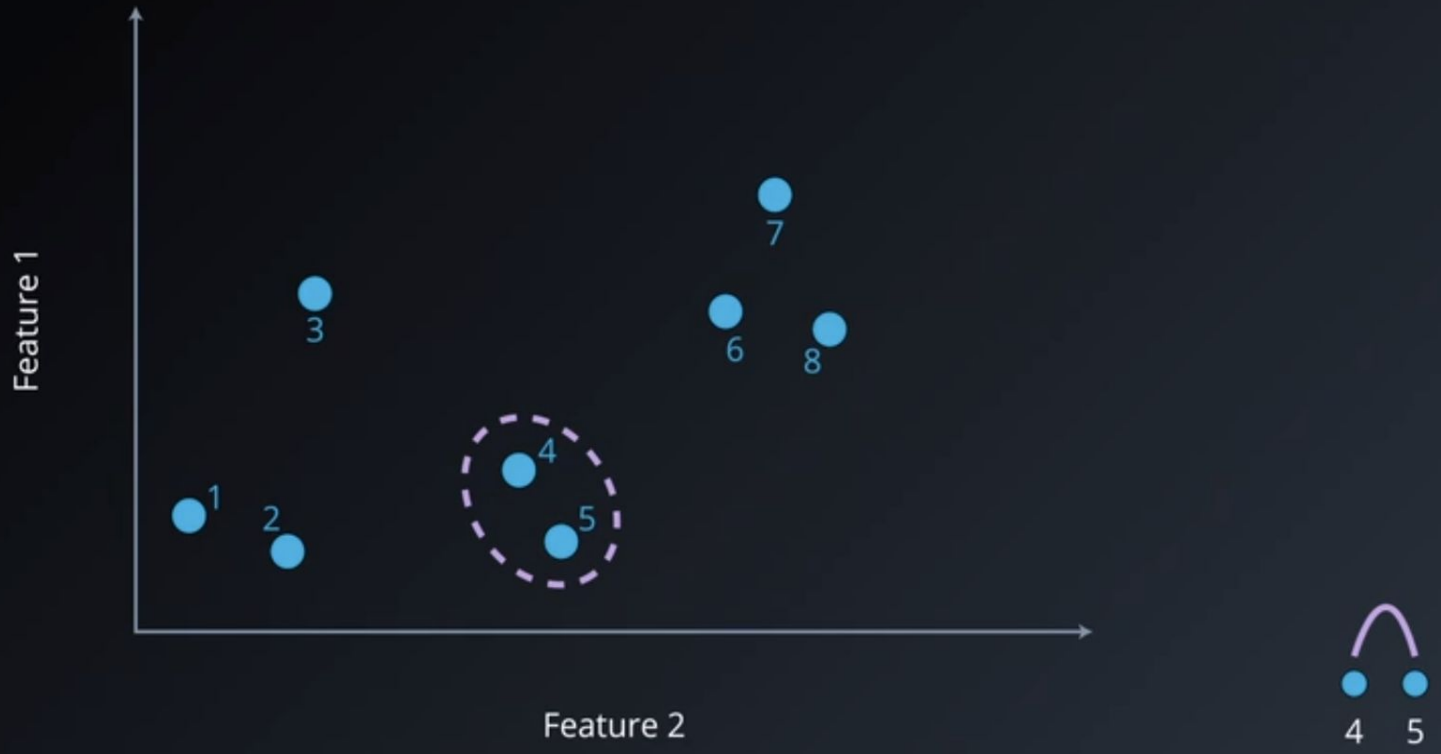
HIERARCHICAL CLUSTERING



DENSITY CLUSTERING

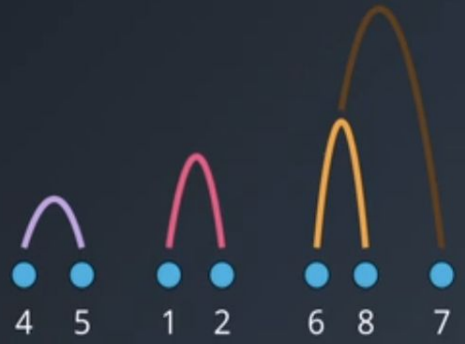
DBSCAN

SINGLE-LINK CLUSTERING

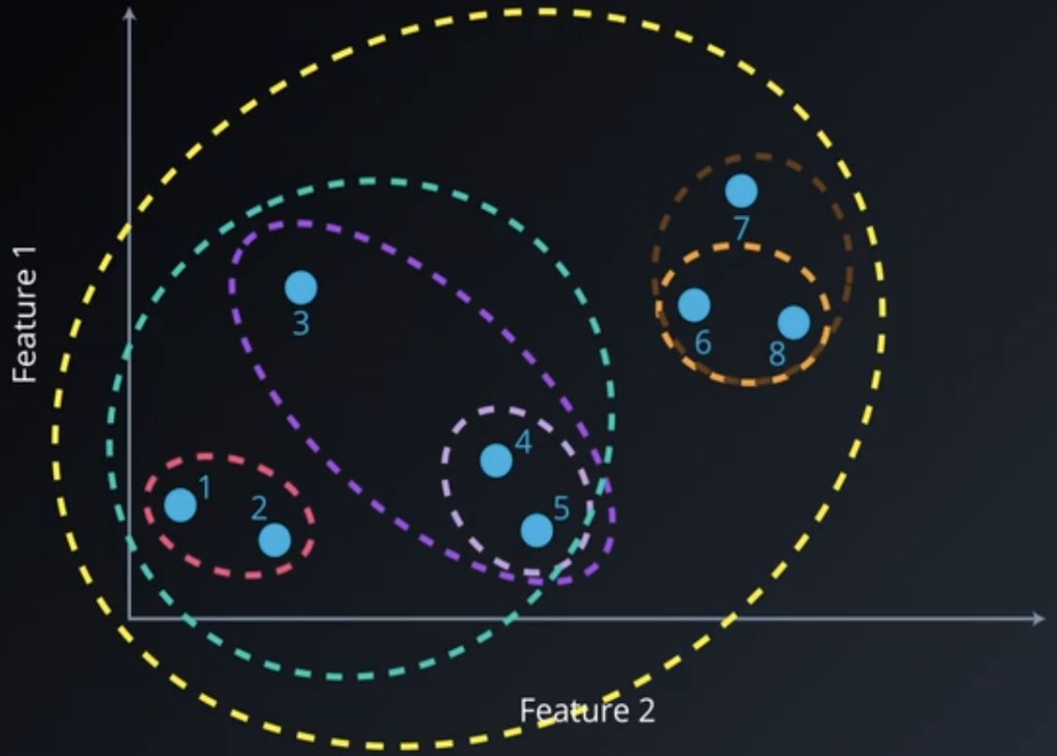


SINGLE-LINK CLUSTERING

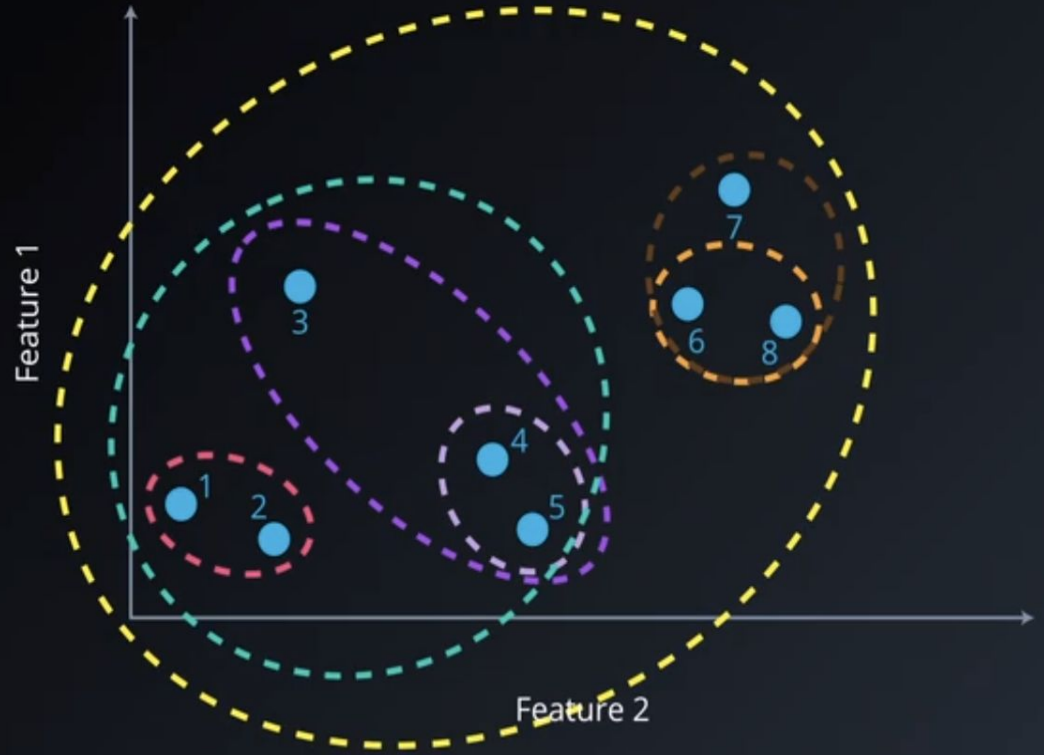
SINGLE-LINK CLUSTERING

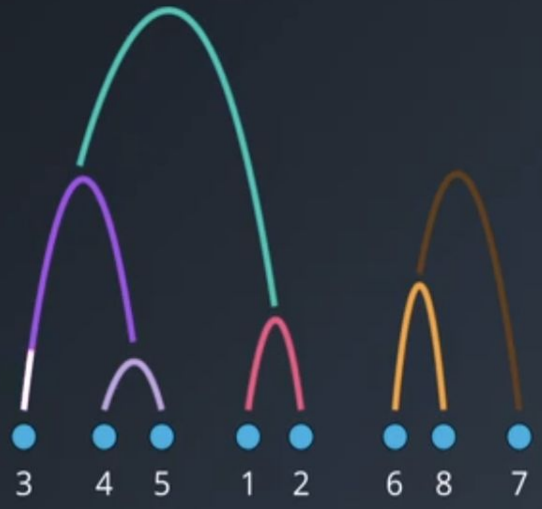


SINGLE-LINK CLUSTERING

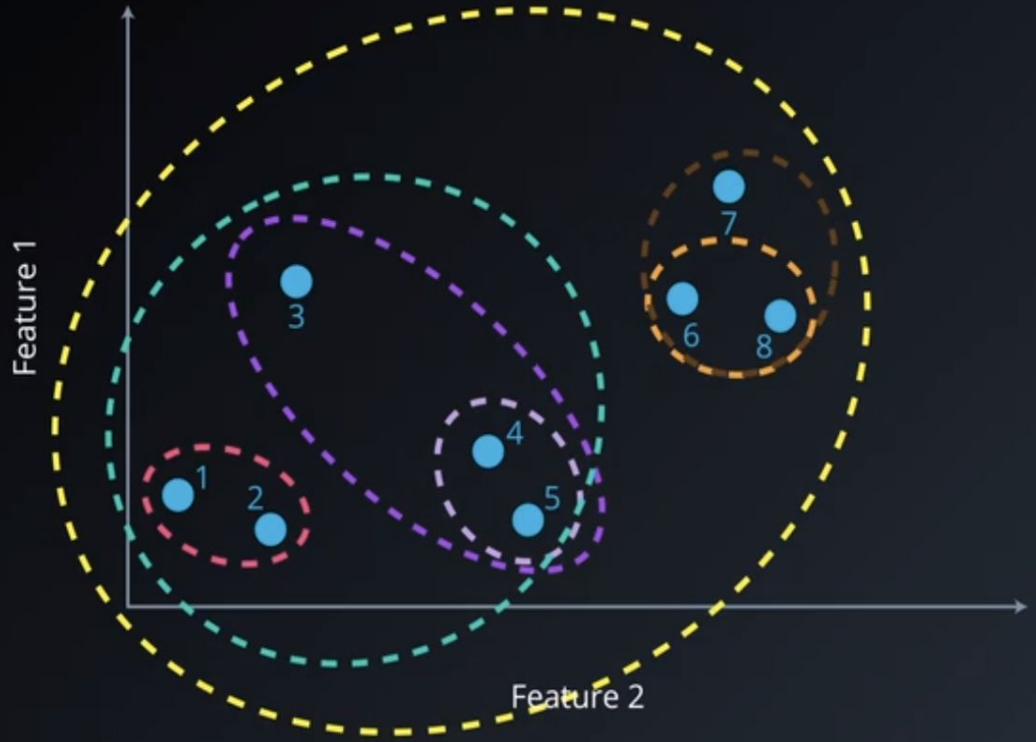


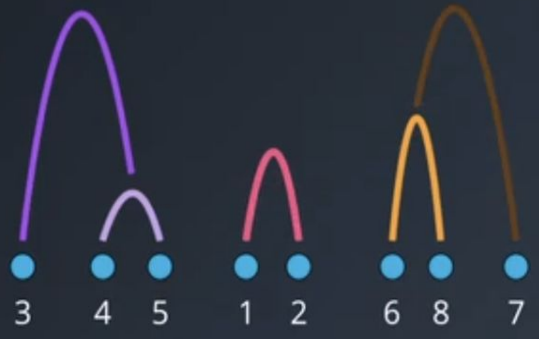
SINGLE-LINK CLUSTERING





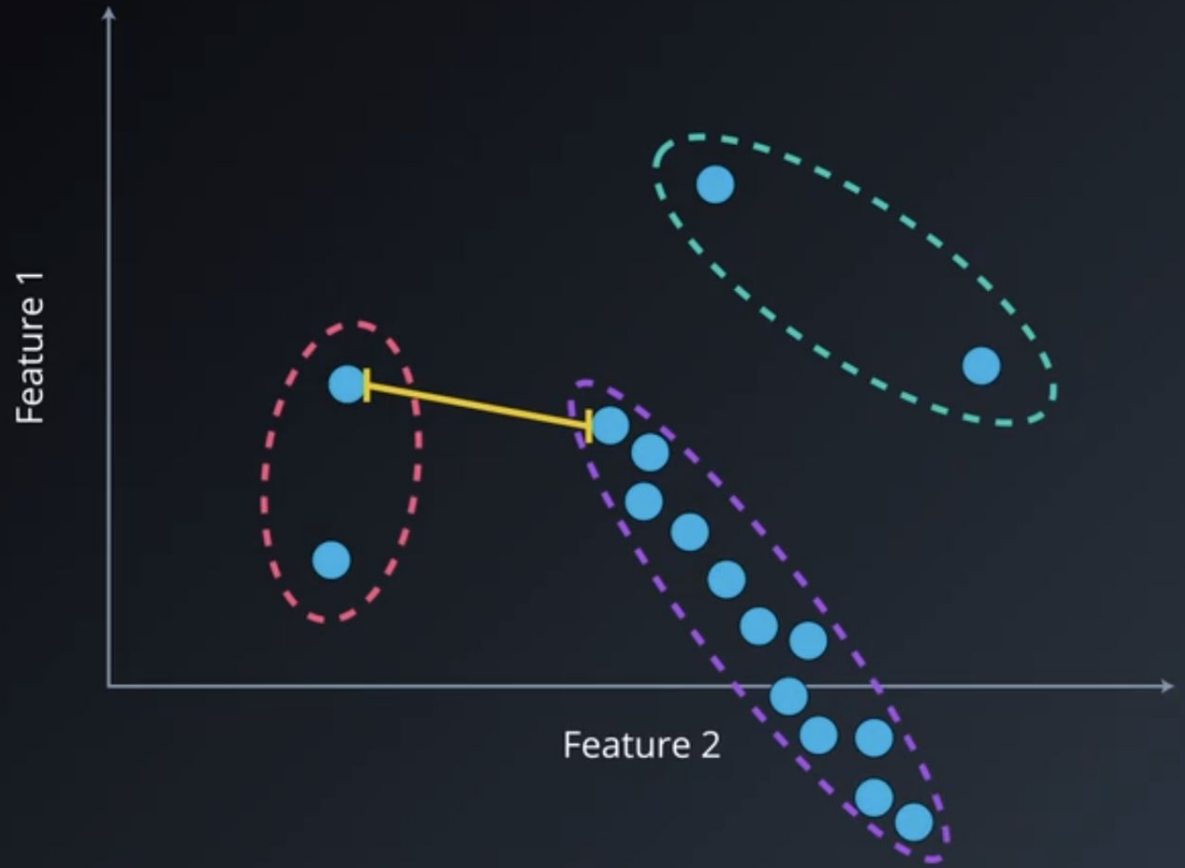
SINGLE-LINK CLUSTERING





DISTANCE MEASURE | SINGLE LINK

DISTANCE MEASURE | SINGLE LINK

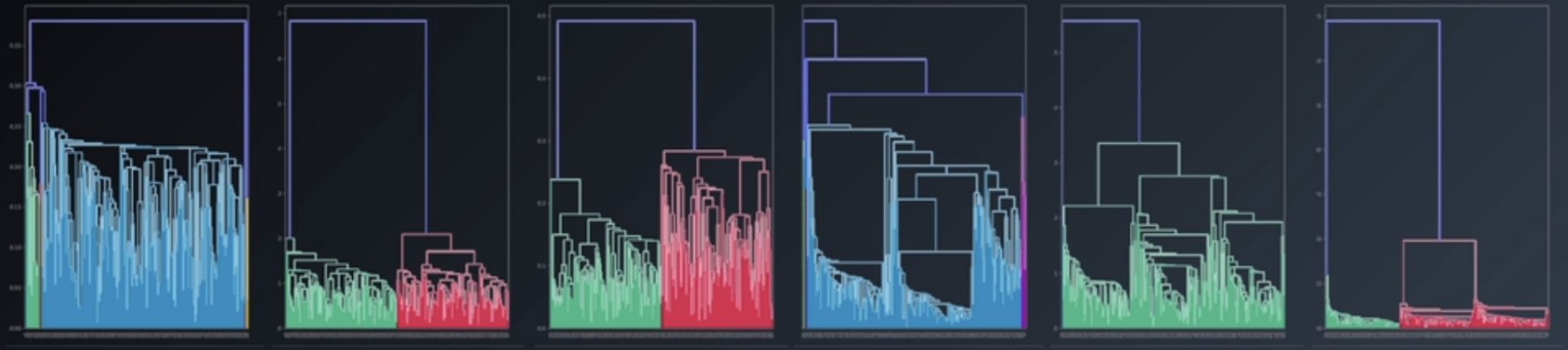


K-MEANS CLUSTERING

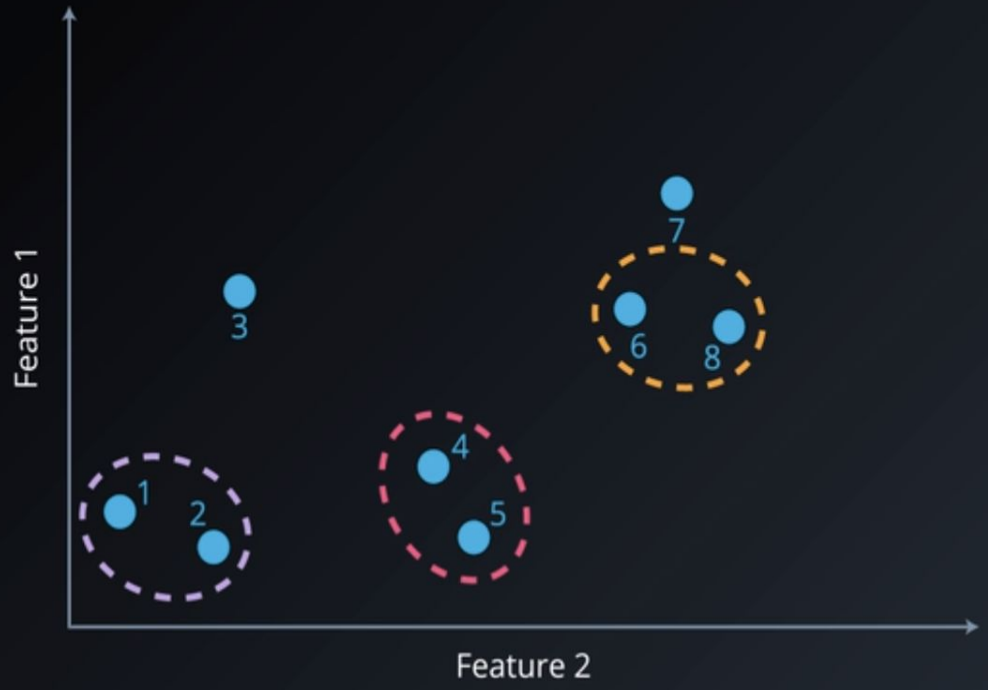
SINGLE LINK HIERARCHICAL CLUSTERING

SINGLE LINK HIERARCHICAL CLUSTERING

LINKAGE DENDROGRAMS

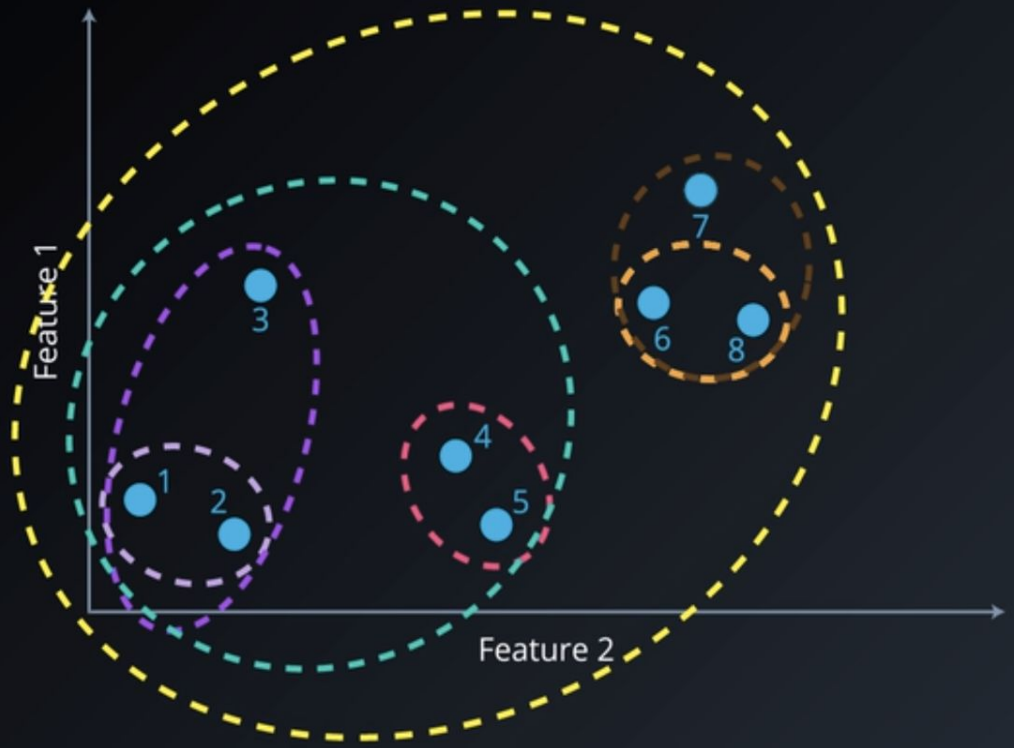


COMPLETE LINK CLUSTERING

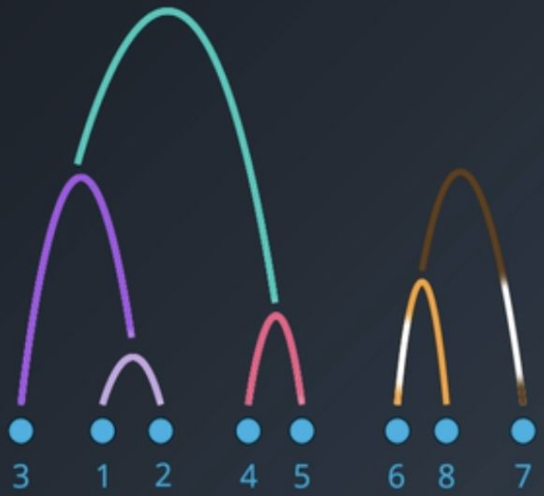


AGGLOMERATIVE CLUSTERING

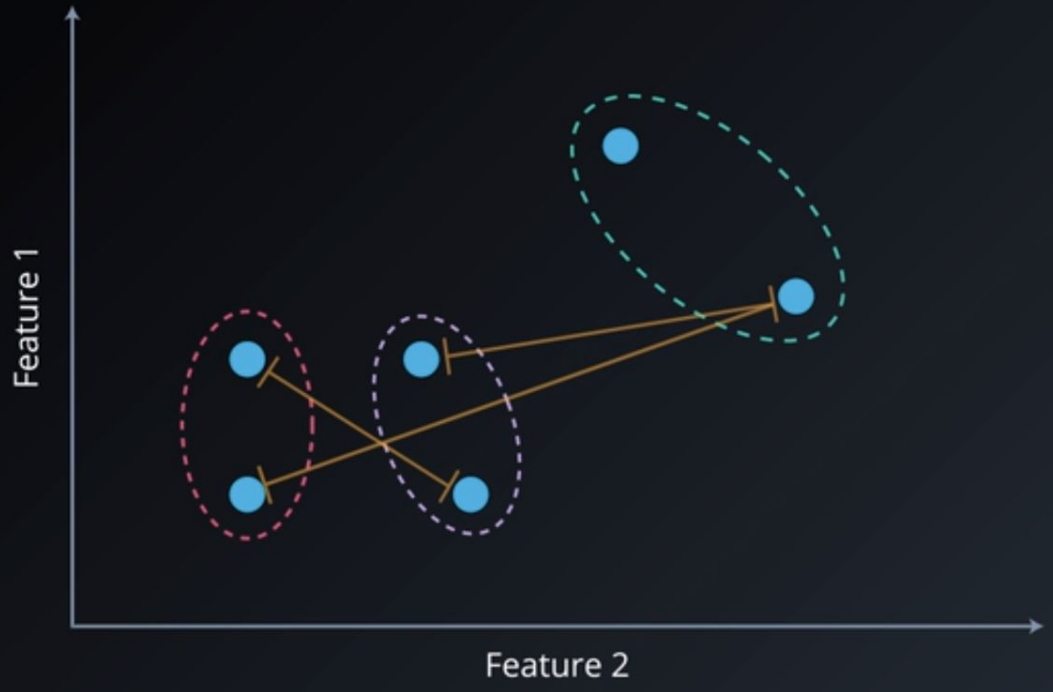
COMPLETE LINK CLUSTERING



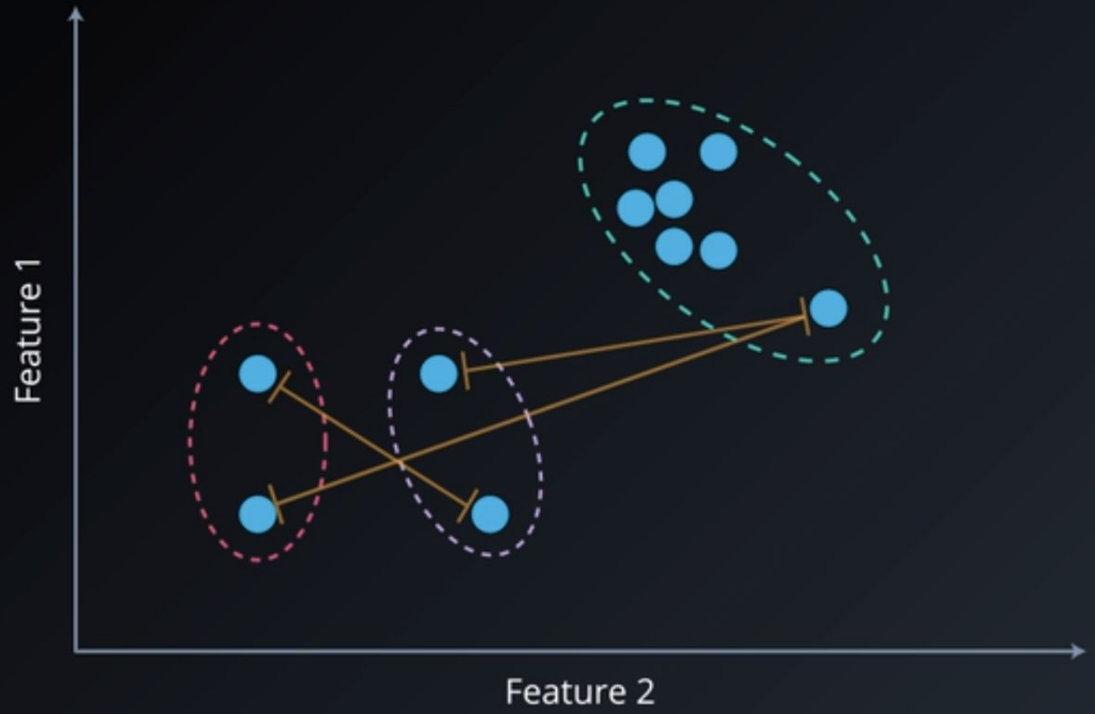
AGGLOMERATIVE CLUSTERING



DISTANCE MEASURE | COMPLETE LINK



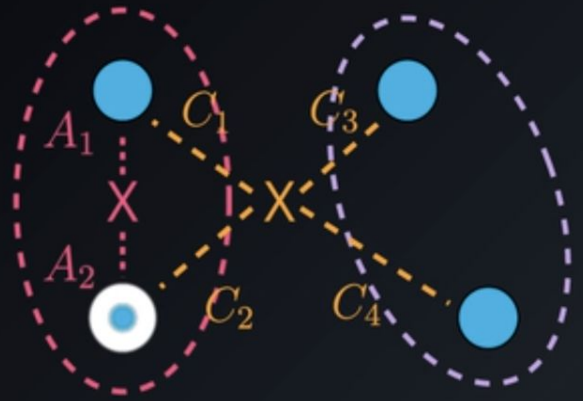
DISTANCE MEASURE | COMPLETE LINK



WARD'S METHOD

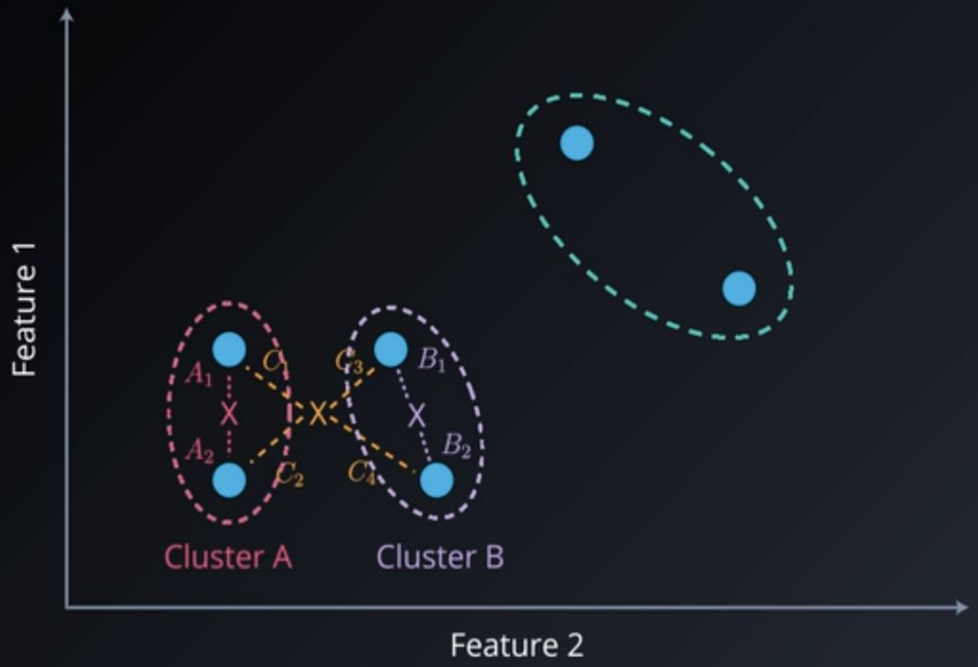
DISTANCE BETWEEN CLUSTERS A AND B

$$\Delta(A,B) = C_1^2 + C_2^2 + C_3^2 + C_4^2 - A_1^2 - A_2^2$$



Feature 2

WARD'S METHOD



DISTANCE BETWEEN CLUSTERS A AND B

$$\Delta(A,B) = C_1^2 + C_2^2 + C_3^2 + C_4^2 - A_1^2 - A_2^2 - B_1^2 - B_2^2$$

HIERARCHICAL CLUSTERING

ADVANTAGES:

- Resulting hierarchical representation can be very informative
- Provides an additional ability to visualize
- Especially potent when the dataset contains real hierarchical relationships (e.g. Evolutionary biology)

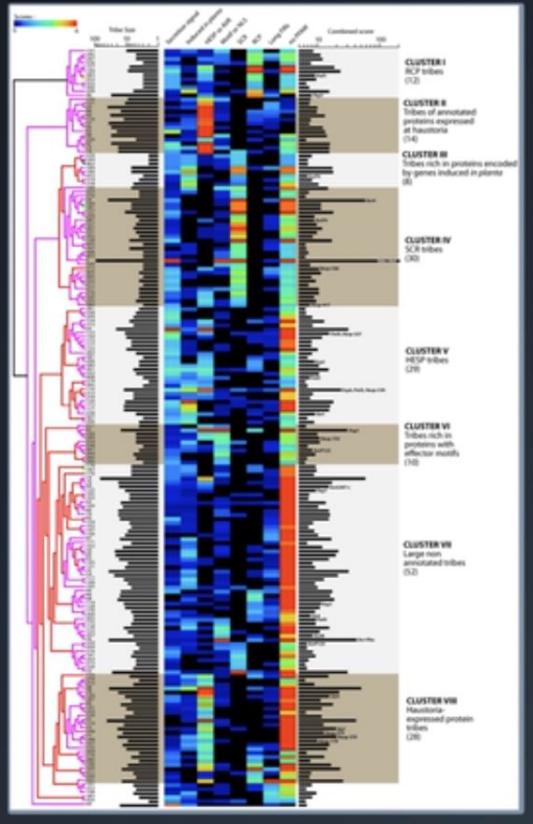
DISADVANTAGES:

- Sensitive to noise and outliers
- Computationally intensive $O(N^2)$

HIERARCHICAL CLUSTERING | APPLICATIONS

Using Hierarchical Clustering of Secreted Protein Families to Classify and Rank Candidate Effectors of Rust Fungi

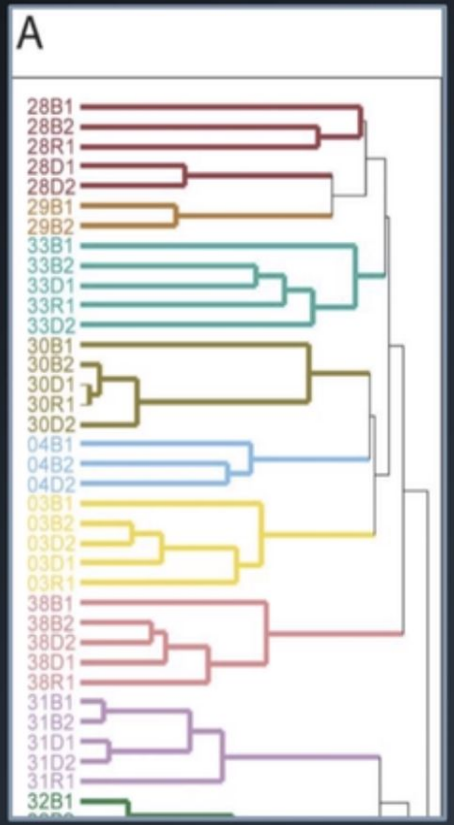
Diane G. O. Saunders, Joe Win, Liliana M. Cano,
Les J. Szabo, Sophien Kamoun, Sylvain Raffaele



HIERARCHICAL CLUSTERING | APPLICATIONS

Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency

Melanie D. Spencer, Timothy J. Hamp,
Robert W. Reid, Leslie M. Fischer,
Steven H. Zeisel, and Anthony A. Fodor



DENSITY-BASED CLUSTERING | DBSCAN

Inputs

Epsilon = 1
Search distance around point
ϵ

MinPts = 5
Minimum number of points
required to form a density cluster

- Noise point

DENSITY-BASED CLUSTERING | DBSCAN

Inputs

Epsilon = 1
Search distance around point
ϵ

MinPts = 5
Minimum number of points
required to form a density cluster

- Noise point

DENSITY-BASED CLUSTERING | DBSCAN

Inputs

Epsilon = 1
Search distance around point
ϵ

MinPts = 5
Minimum number of points
required to form a density cluster

- Noise point
- Core point - Cluster 1

DENSITY-BASED CLUSTERING | DBSCAN

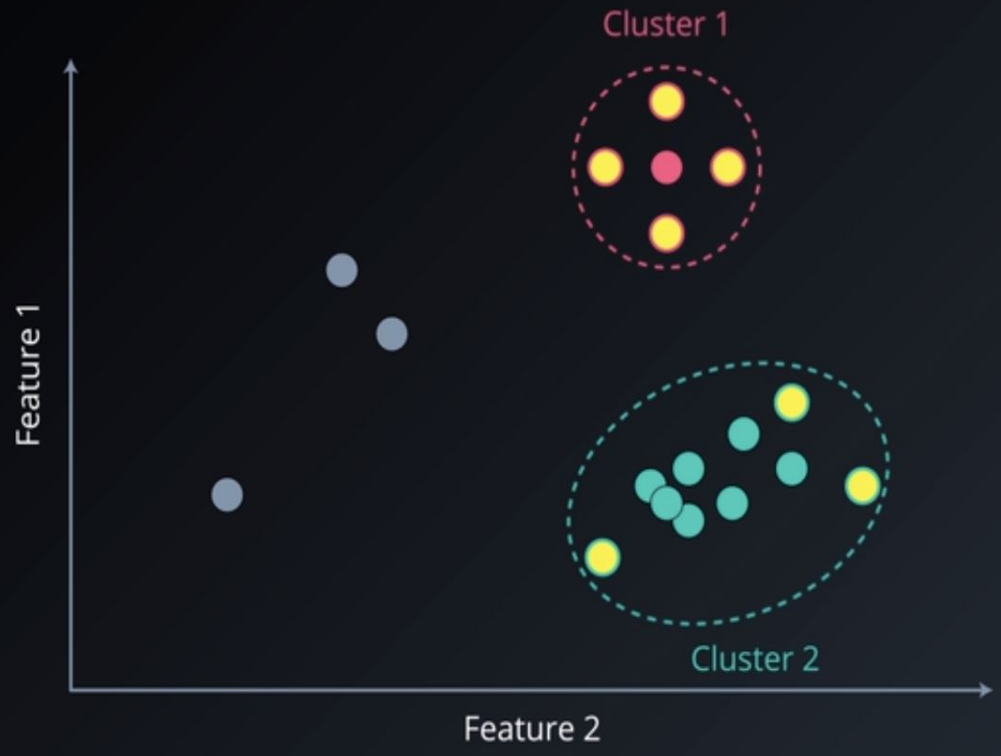
Inputs

Epsilon = 1
Search distance around point
ϵ

MinPts = 5
Minimum number of points
required to form a density cluster

- Noise point
- Core point - Cluster 1
- Border point - Cluster 1
- Core point - Cluster 2

DENSITY-BASED CLUSTERING | DBSCAN



Inputs

Epsilon = 1
Search distance around point

ϵ

MinPts = 5
Minimum number of points required to form a density cluster

- Noise point
- Core point - Cluster 1
- Border point - Cluster 1
- Core point - Cluster 2
- Border point - Cluster 2

K-MEANS CLUSTERING

DBSCAN

DENSITY-BASED CLUSTERING | DBSCAN

ADVANTAGES:

- We don’t need to specify the number of clusters
- Flexibility in the shapes & sizes of clusters
- Able to deal with noise
- Able to deal with outliers

DISADVANTAGES:

- Border points that are reachable from two clusters
- Faces difficulty finding clusters of varying densities

DENSITY-BASED CLUSTERING | APPLICATIONS

Traffic Classification Using Clustering Algorithms

Jeffrey Erman, Martin Arlitt, Anirban Mahanti

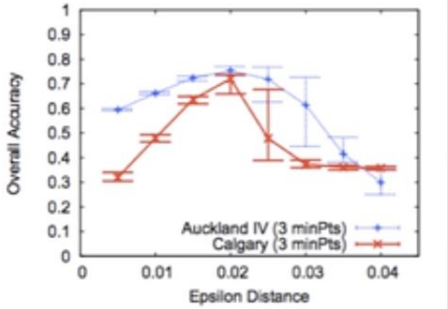


Figure 2: Accuracy using DBSCAN

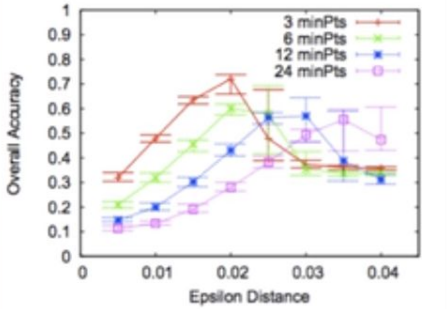


Figure 3: Parametrization of DBSCAN

DENSITY-BASED CLUSTERING | APPLICATIONS

Traffic Classification Using Clustering Algorithms

Jeffrey Erman, Martin Arlitt, Anirban Mahanti

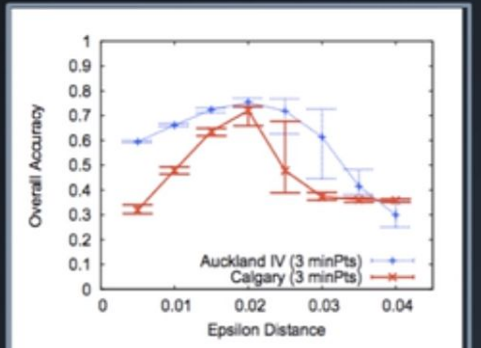


Figure 2: Accuracy using DBSCAN

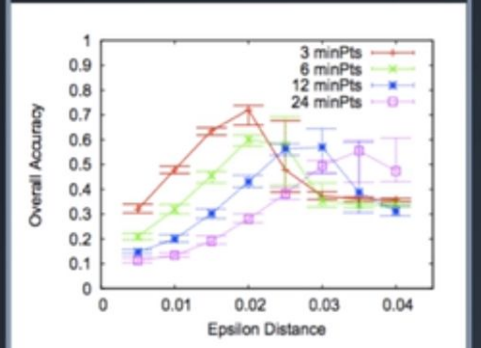
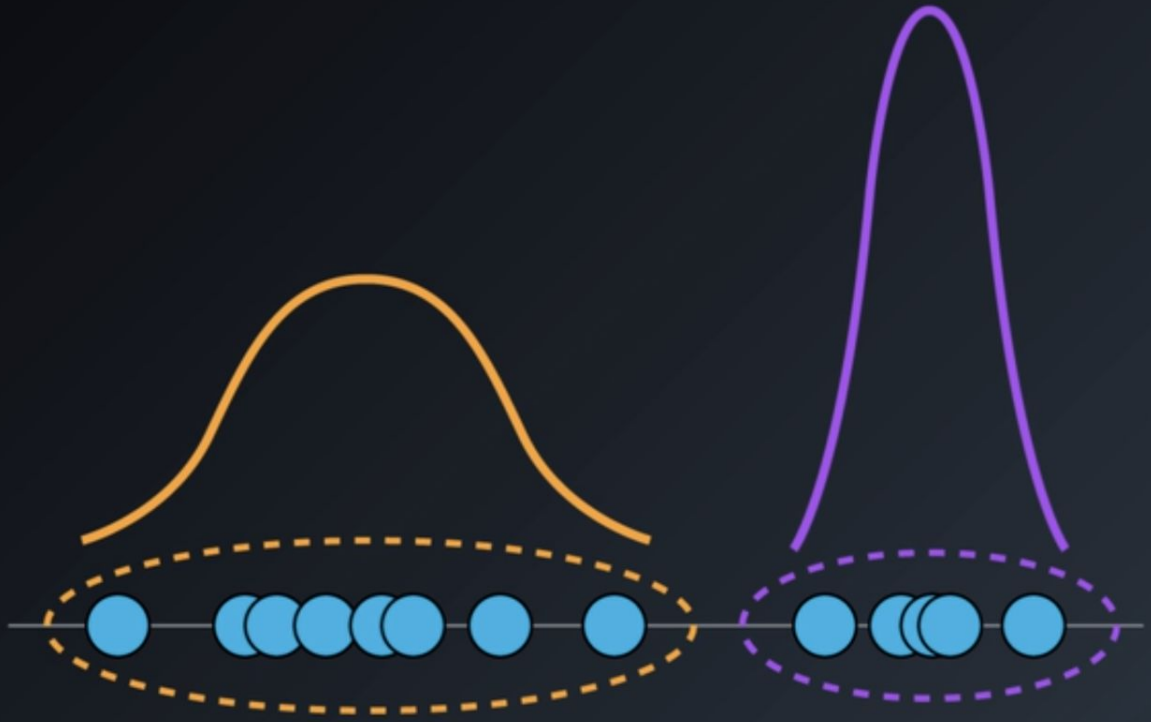


Figure 3: Parametrization of DBSCAN

GAUSSIAN MIXTURE MODELS

GAUSSIAN MIXTURE MODELS



CLUSTER VALIDATION

K-means

silhouette score: 0.637

Single Link

silhouette score: -0.08

Complete Link

silhouette score: 0.589

Ward

silhouette score: 0.613

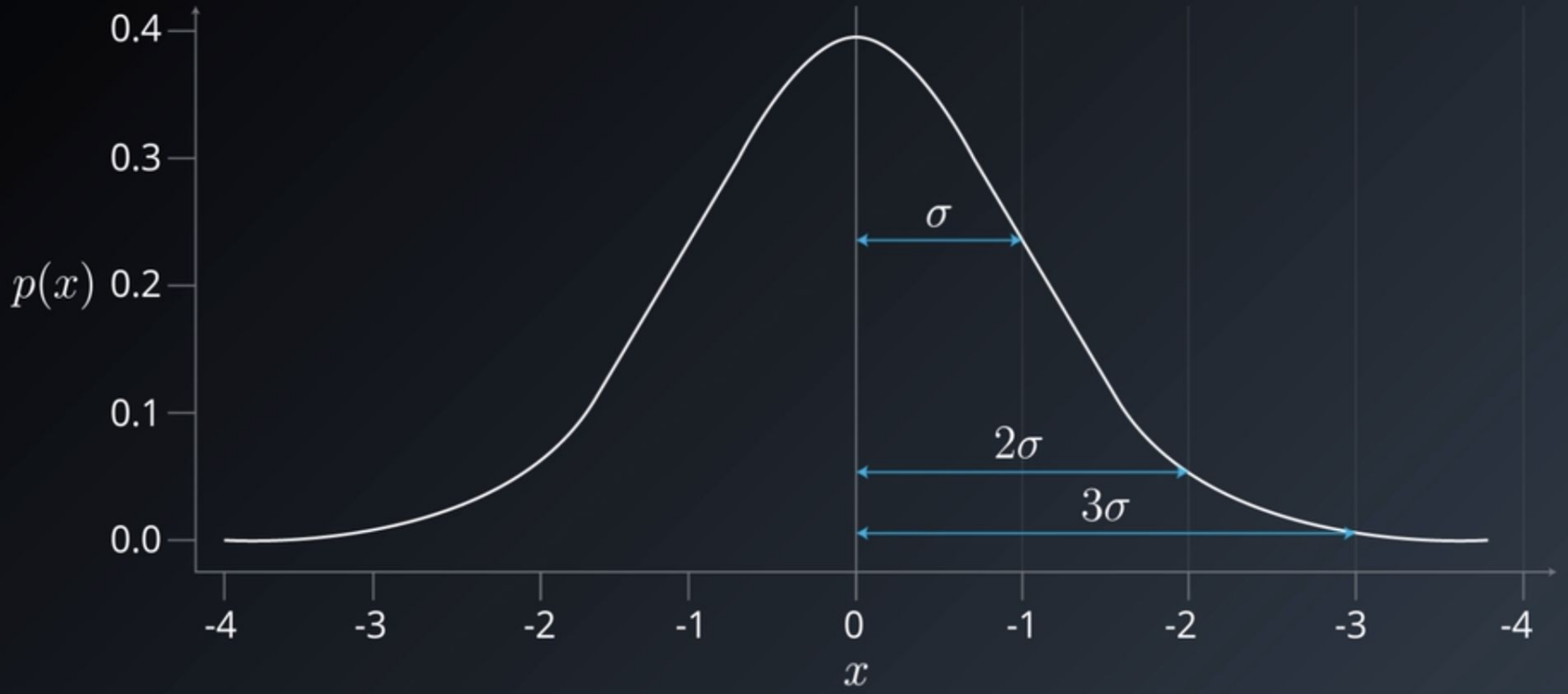
DBSCAN

silhouette score: 0.397

GMM

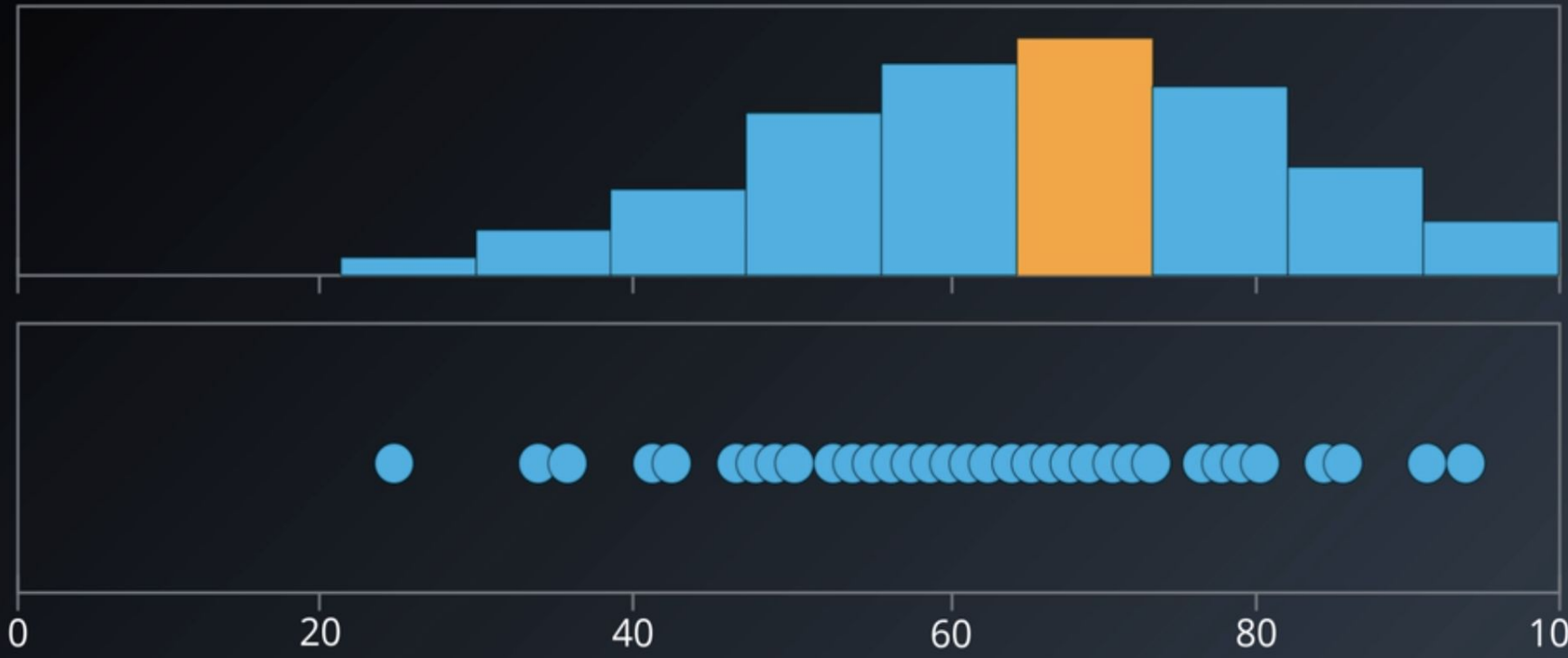
silhouette score: 0.575

NORMAL DISTRIBUTION WHEN $\mu = 0$ AND $\sigma = 1$



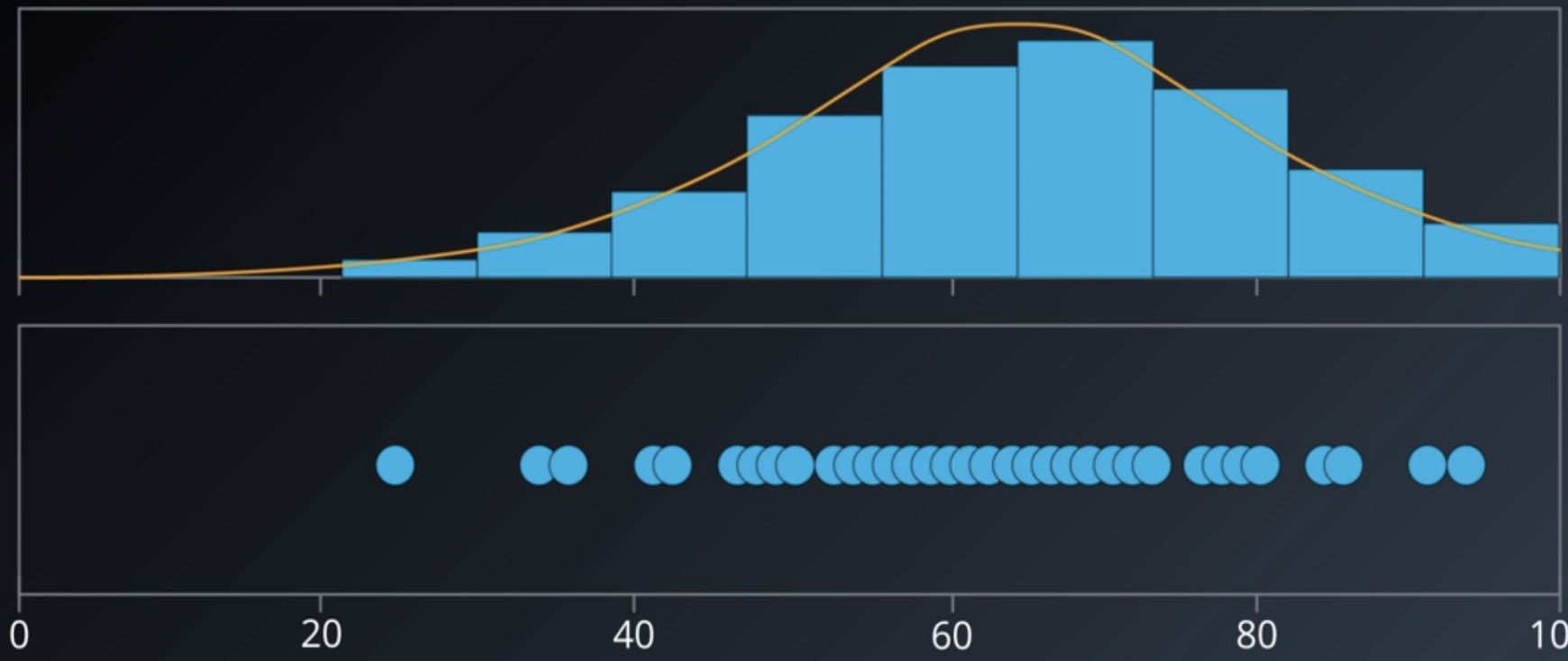
GAUSSIAN DISTRIBUTION

EXAMPLE: TEST SCORES



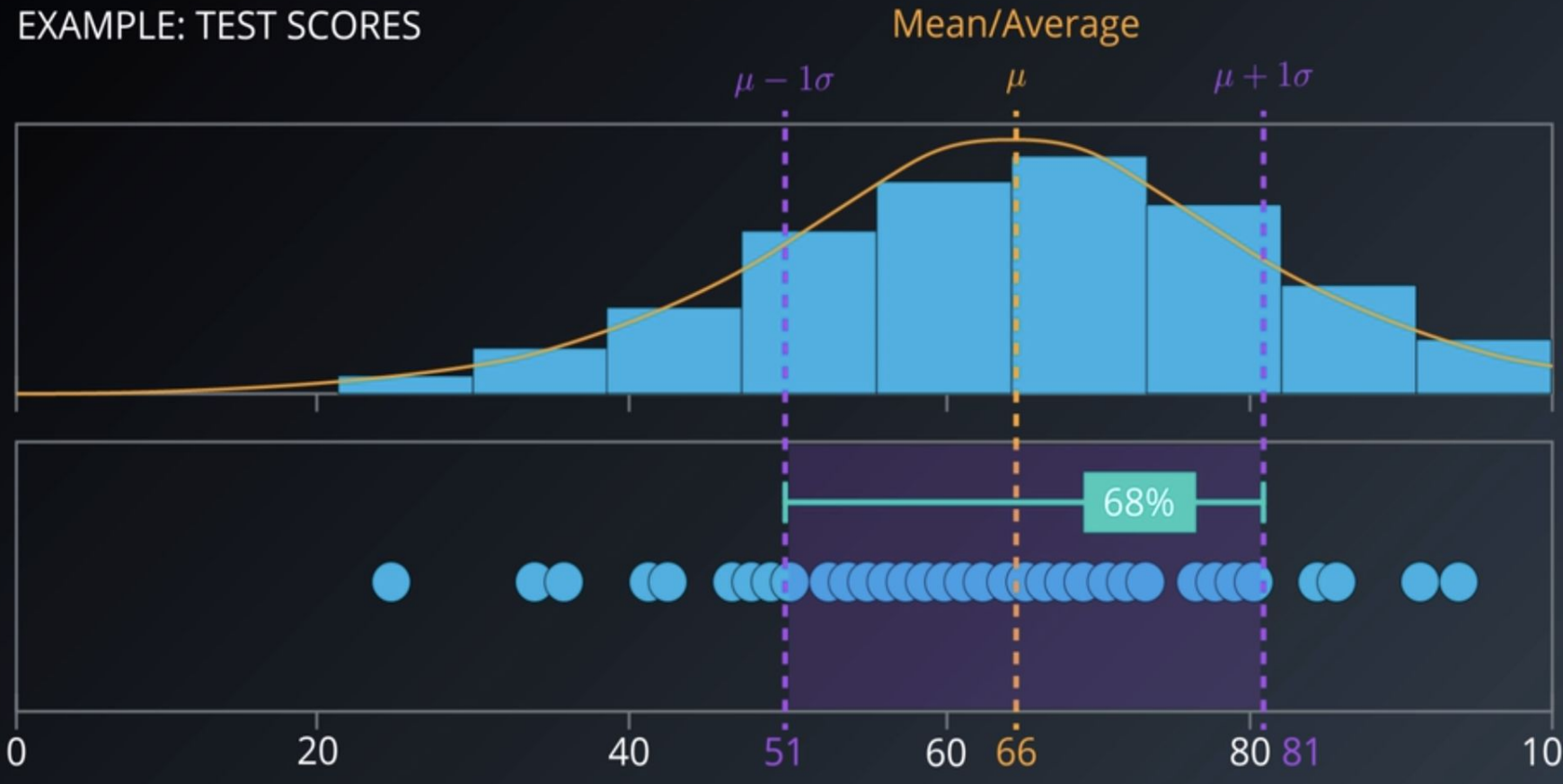
GAUSSIAN DISTRIBUTION

EXAMPLE: TEST SCORES



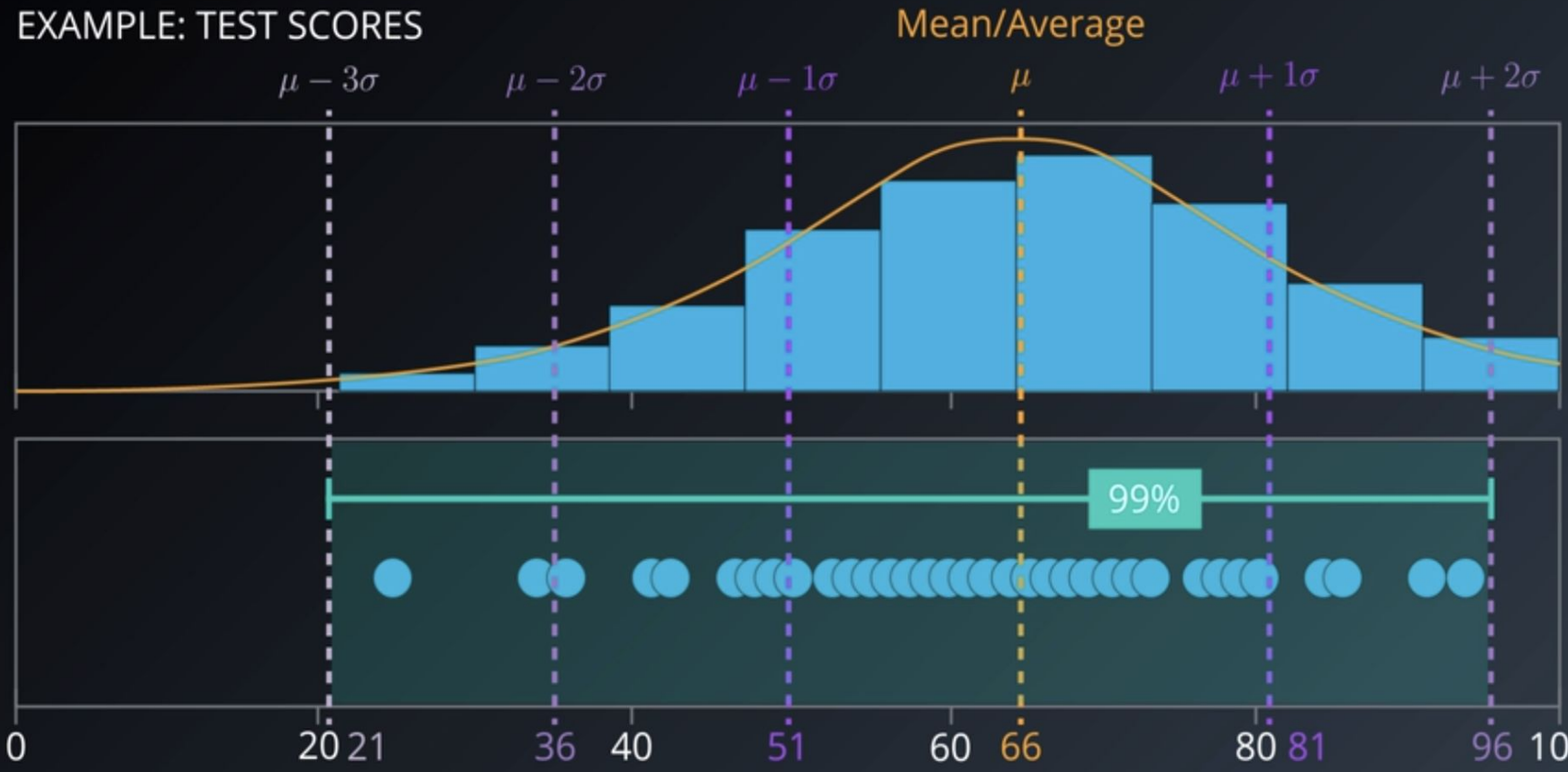
GAUSSIAN DISTRIBUTION

EXAMPLE: TEST SCORES

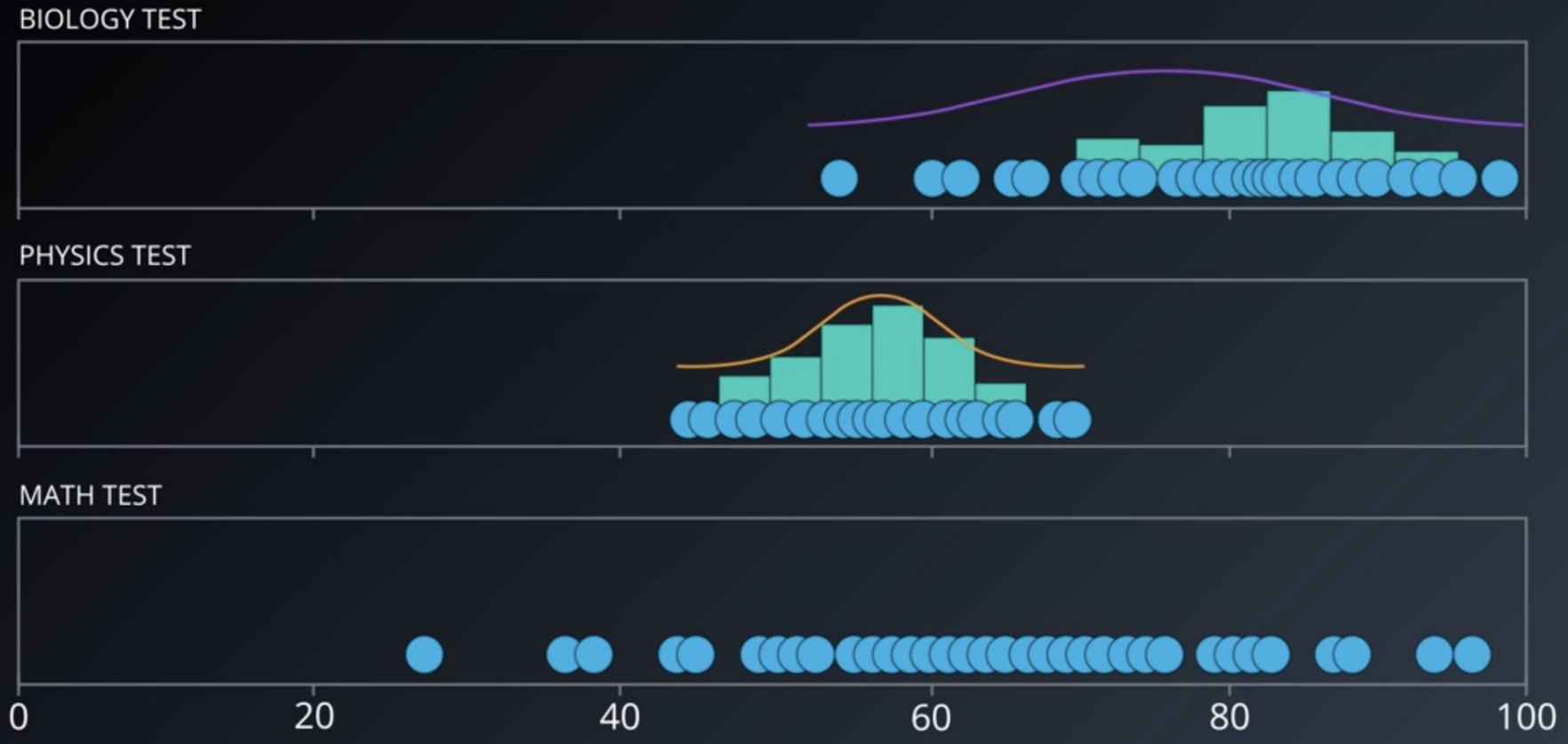


GAUSSIAN DISTRIBUTION

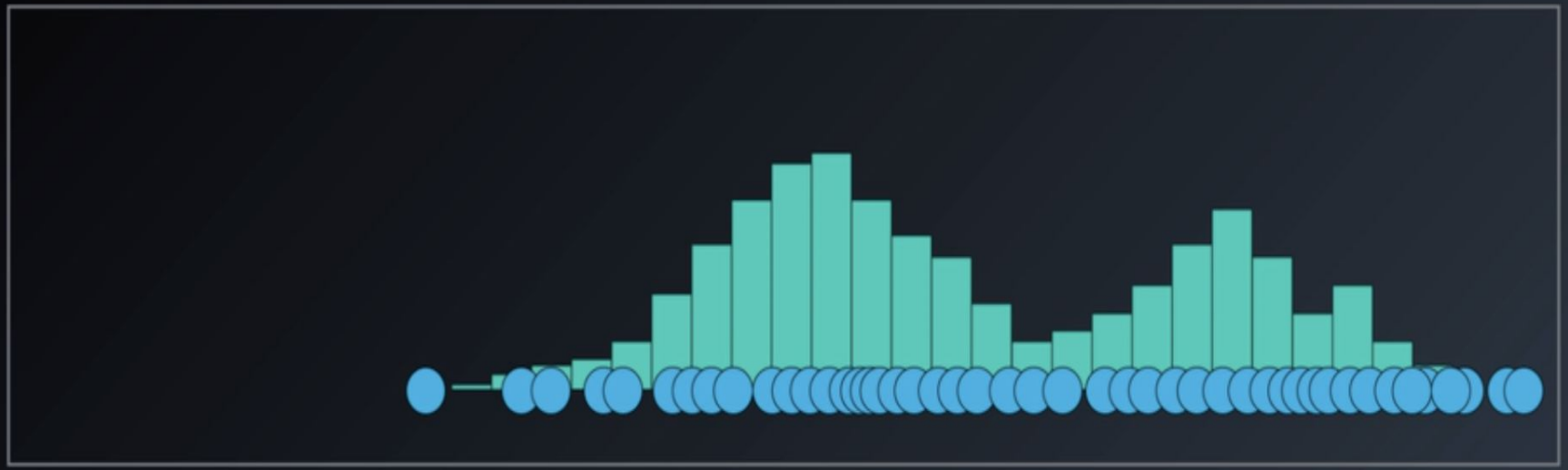
EXAMPLE: TEST SCORES



GAUSSIAN DISTRIBUTION



GAUSSIAN MIXTURE MODEL CLUSTERING

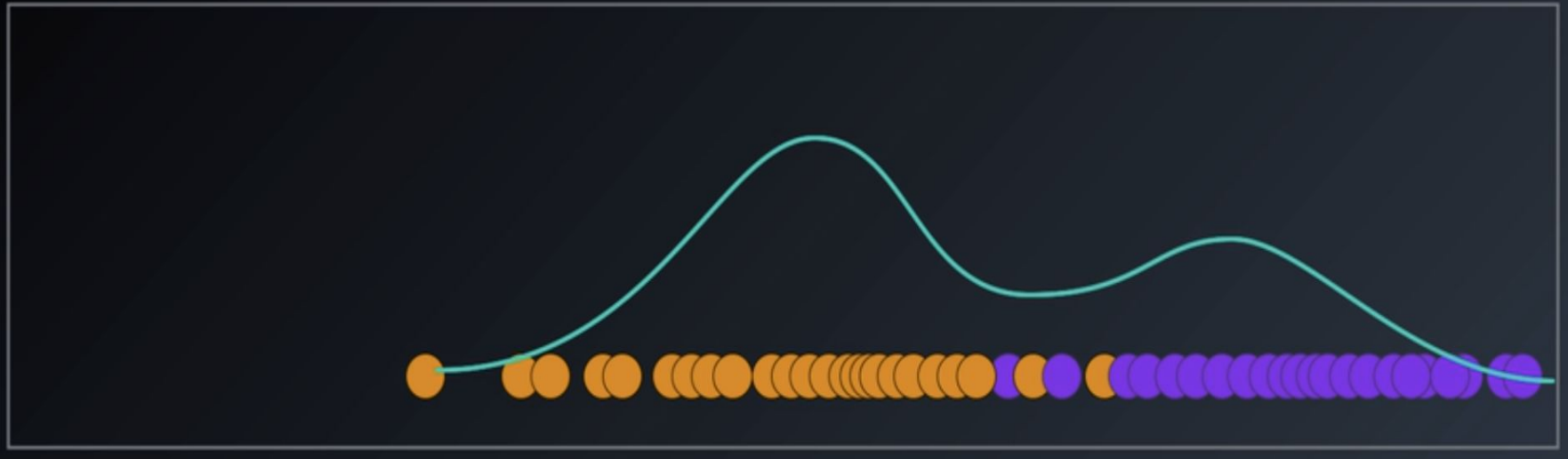


GAUSSIAN MIXTURE MODEL CLUSTERING

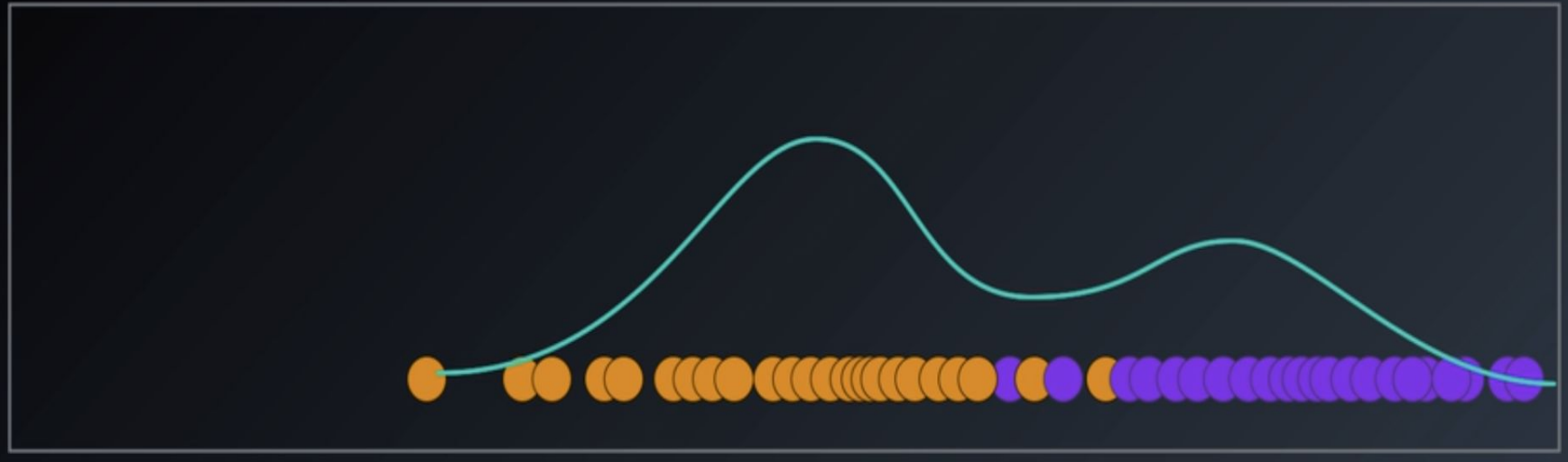
BIOLOGY TEST

PHYSICS TEST

GAUSSIAN MIXTURE MODEL CLUSTERING



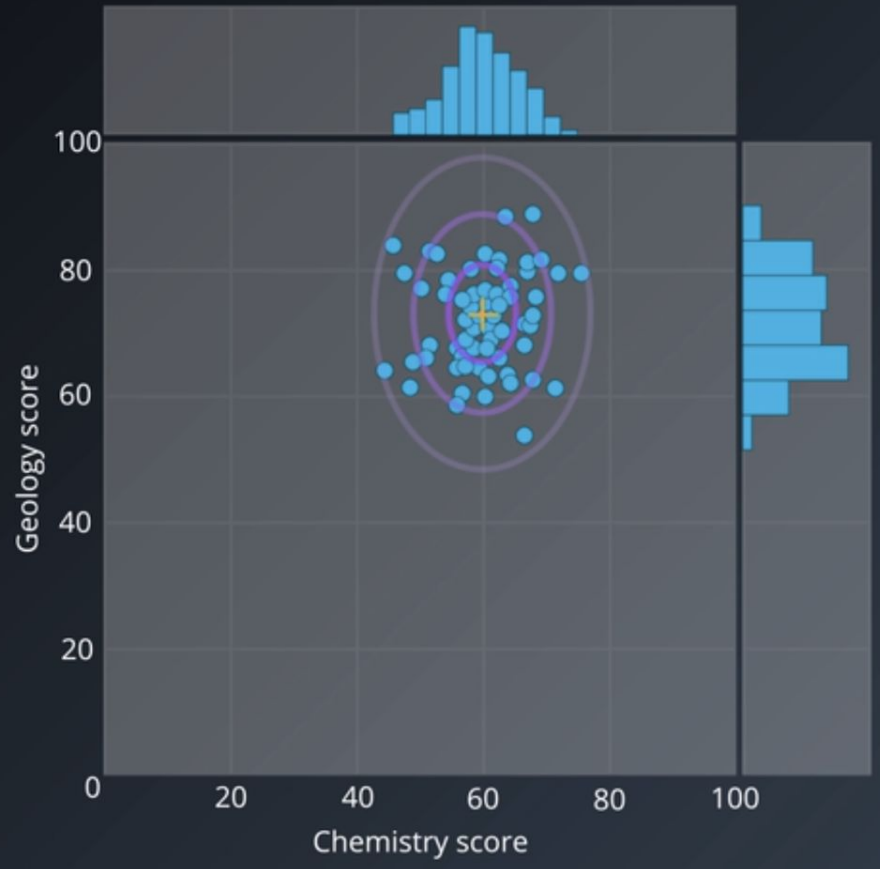
GAUSSIAN MIXTURE MODEL CLUSTERING



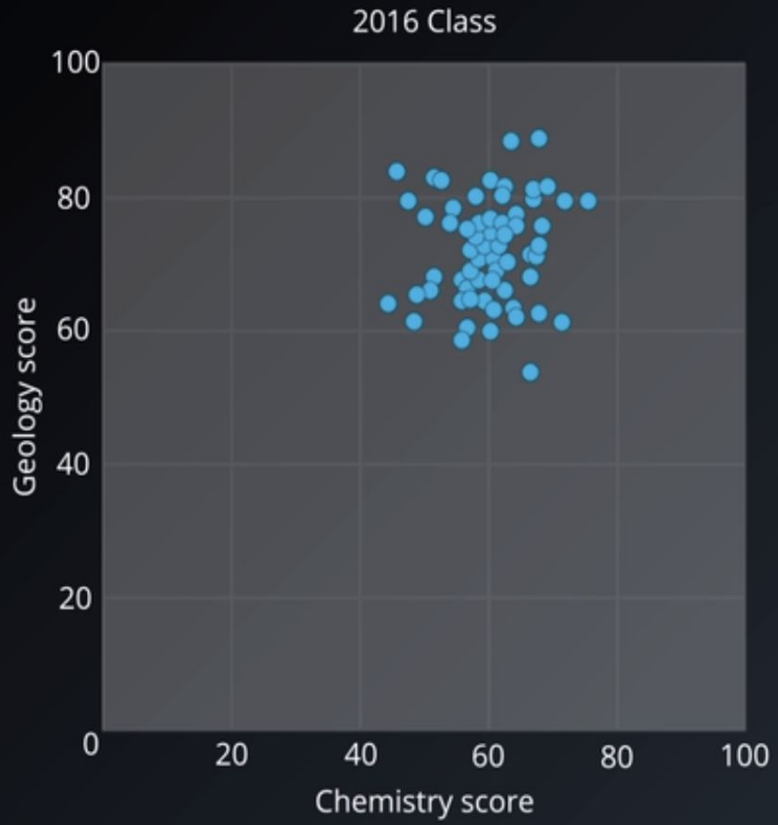
GAUSSIAN DISTRIBUTION | TWO DIMENSIONS

MULTIVARIATE GAUSSIAN DISTRIBUTION

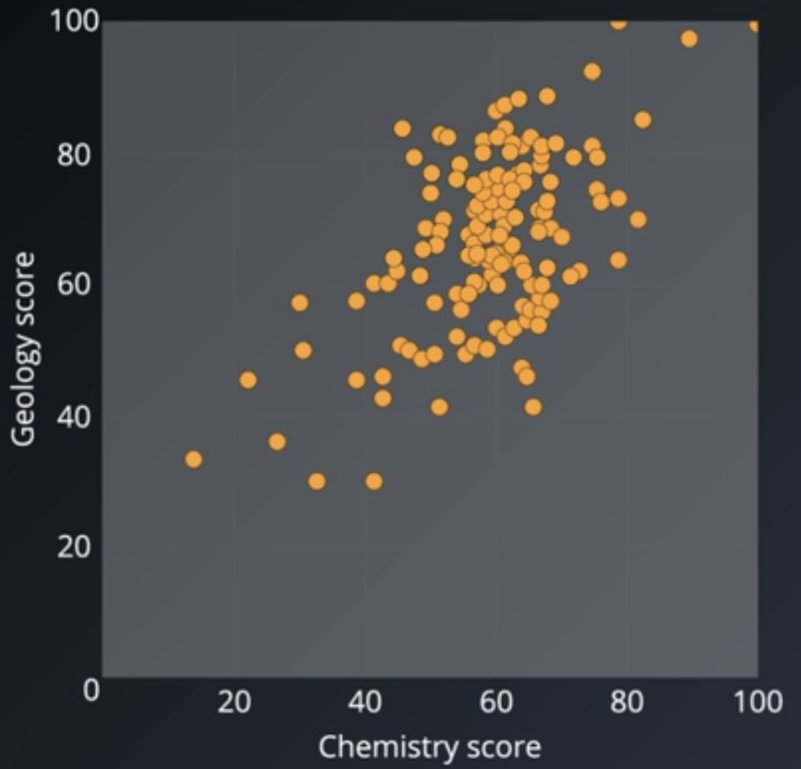
Student ID	Chemistry Score	Geology Score
1	66	83
2	61	73
3	57	87
4	61	72
...	...	...
99	60	76
MEAN	59	78
STD DEVIATION	5	8



GAUSSIAN DISTRIBUTION | TWO DIMENSIONS

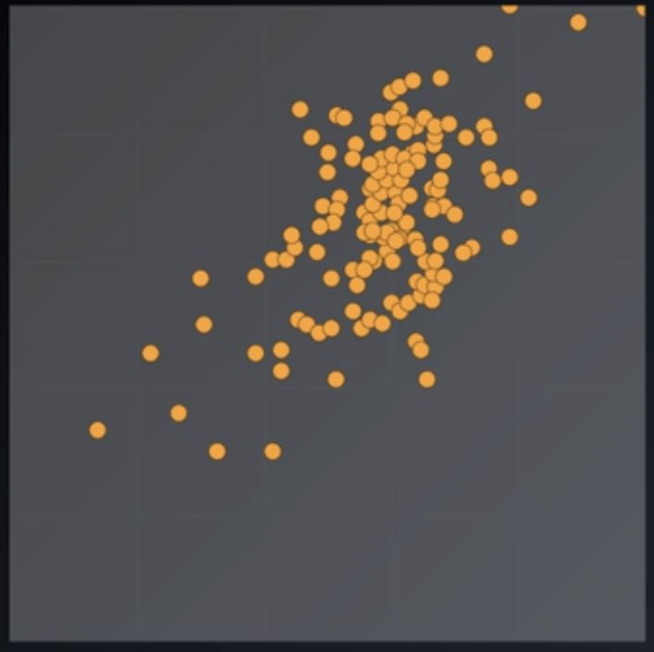


GAUSSIAN DISTRIBUTION | TWO DIMENSIONS



GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Expectation - Maximization For Gaussian Mixtures:



Dataset to cluster into two clusters

STEP #1: INITIALIZE K GAUSSIAN DISTRIBUTIONS

STEP #2: SOFT-CLUSTER DATA - “EXPECTATION”

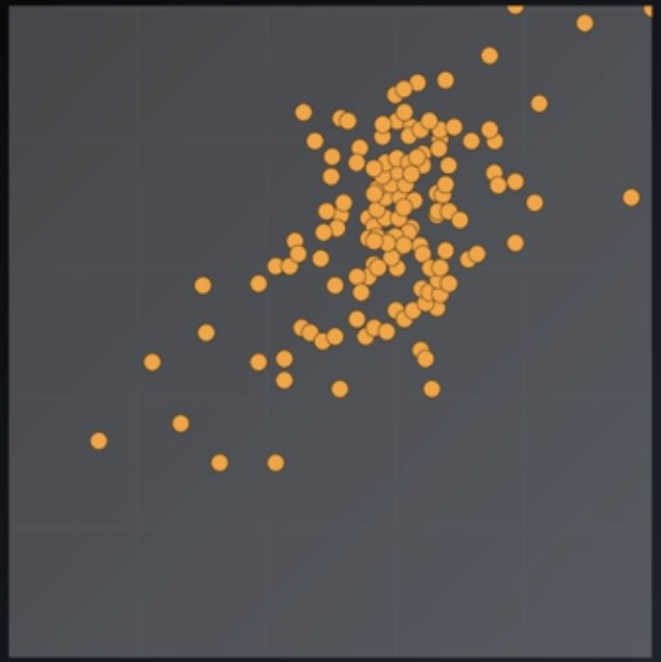
STEP #3: RE-ESTIMATE THE GAUSSIANS - “MAXIMIZATION”

STEP #4: EVALUATE LOG-LIKELIHOOD TO CHECK FOR CONVERGENCE

REPEAT FROM STEP #2 UNTIL CONVERGED

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Step 1 - Initialize Gaussian Distributions



Initial Gaussian Parameters

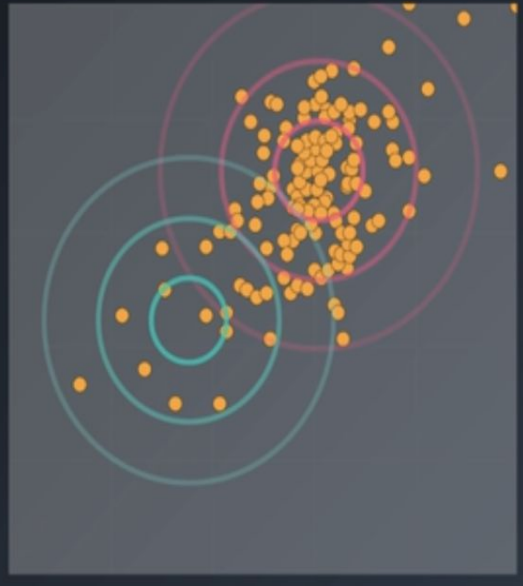
Cluster	μ	σ^2
A	(64.63, 76.30)	100
B	(46.02, 51.30)	57

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Step 2 - Soft-cluster the data points - “Expectation” step

SOFT CLUSTERING (”RESPONSIBILITIES”)

Point #	feature 1	feature 2	Cluster A	Cluster B
1	62	71		
2	58	81		
3	52	74		
...	...	...		
N	52	78		

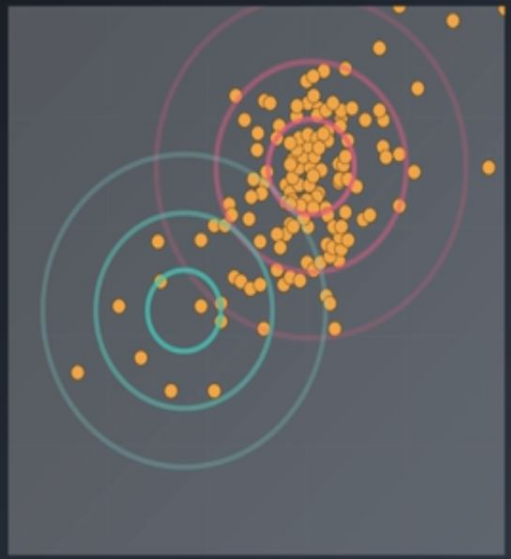


GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Step 2 - Soft-cluster the data points - “Expectation” step

SOFT CLUSTERING (”RESPONSIBILITIES”)

Point #	feature 1	feature 2	Cluster A	Cluster B
1	62	71		
2	58	81		
3	52	74		
...	...	...		
N	52	78		



$$E[Z_{1A}] = \frac{N(X_i|\mu_A, \sigma_A^2)}{N(X_i|\mu_A, \sigma_A^2) + N(X_i|\mu_B, \sigma_B^2)}$$

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Step 2 - Soft-cluster the data points - “Expectation” step

SOFT CLUSTERING (”RESPONSIBILITIES”)

Point #	feature 1	feature 2	Cluster A	Cluster B
1	62	71	0.99976	
2	58	81		
3	52	74		
...	...	...		
N	52	78		

Cluster	μ	σ^2
A	(64.63, 76.30)	100
B	(46.02, 51.30)	57

$$E[Z_{1A}] = \frac{N(X_i|\mu_A, \sigma_A^2)}{N(X_i|\mu_A, \sigma_A^2) + N(X_i|\mu_B, \sigma_B^2)} = \frac{0.001288}{0.001288 + 0.0000038}$$

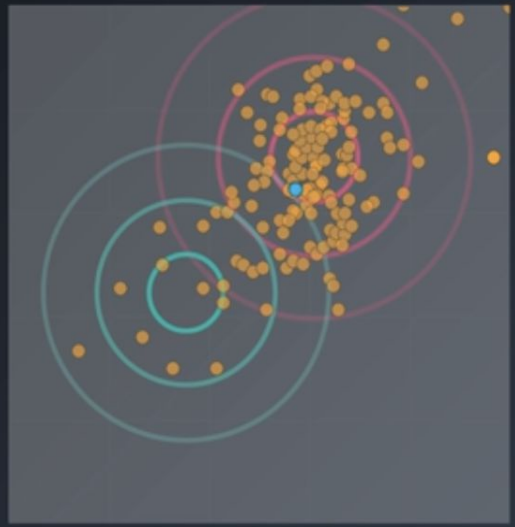
$$N(X|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Step 2 - Soft-cluster the data points - “Expectation” step

SOFT CLUSTERING (”RESPONSIBILITIES”)

Point #	feature 1	feature 2	Cluster A	Cluster B
1	62	71	0.99976	0.00024
2	58	81		
3	52	74		
...	...	...		
N	52	78		



Cluster	μ	σ^2
A	(64.63, 76.30)	100
B	(46.02, 51.30)	57

$$E[Z_{1A}] = \frac{N(X_i|\mu_A, \sigma_A^2)}{N(X_i|\mu_A, \sigma_A^2) + N(X_i|\mu_B, \sigma_B^2)} = \frac{0.001288}{0.001288 + 0.0000038}$$

$$N(X|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Step 3 - Re-estimate parameters of Gaussians - “Maximization” step

NEW GAUSSIAN PARAMETERS

Cluster	new μ	new σ^2
A	(64.4872457, 76.3074590)	
B		

$$\text{new } \mu_A = \frac{\sum_{i=1}^{N} E[Z_{ij}]X_i}{\sum_{i=1}^{N} E[Z_{ij}]} = \frac{0.99937\text{x}(62\ 71) + 0.9998\text{x}(58\ 81) + \ldots}{0.99937 + 0.9998 + 0.55818 + \ldots}$$

$$= (64.4872457, 76.3074590)$$

Point #	feature 1	feature 2	Cluster A	Cluster B
1	62	71	0.99937	0.00063
2	58	81	0.9998	0.0002
3	52	74	0.55818	0.44182
...	...	...	...	...
N	52	78	0.99133	0.00867

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Step 3 - Re-estimate parameters of Gaussians - “Maximization” step

NEW GAUSSIAN PARAMETERS

Cluster	new μ	new σ^2
A	(64.4872457, 76.3074590)	103.92494596
B	(46.0271498, 51.3087720)	67.10773268

$$\text{new } \sigma_A^2 = \frac{\sum_{i=1}^{N} E[Z_{iA}](X_i - \mu_A^{new})(X_i - \mu_A^{new})^T}{\sum_{i=1}^{N} E[Z_{iA}]}$$

$$= 103.92494596$$

Point #	feature 1	feature 2	Cluster A	Cluster B
1	62	71	0.99937	0.00063
2	58	81	0.9998	0.0002
3	52	74	0.55818	0.44182
...	...	...	...	...
N	52	78	0.99133	0.00867

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Step 4 - Evaluate log-likelihood

$$1\mathrm{n}\ p(X|\mu,\sigma^2) = \sum_{i=1}^{N} 1\mathrm{n}(\sum_{k=1}^{K} \pi_k N(X_i|\mu_k,\sigma_k^2))$$

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

covariance_type:
Spherical

Initialization:
Manual

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

covariance_type:
Spherical

Initialization:
Default (k-means)

GAUSSIAN MIXTURE MODEL CLUSTERING

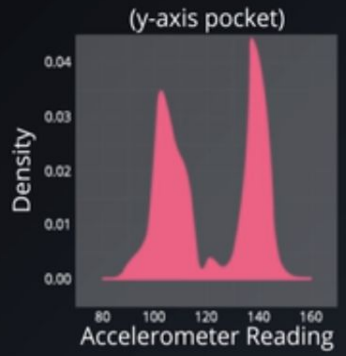
Advantages:
- Soft-clustering (sample membership of multiple clusters)
- Cluster shape flexibility

Disadvantages:
- Sensitive to initialization values
- Possible to converge to a local optimum
- Slow convergence rate

GAUSSIAN MIXTURE MODEL CLUSTERING | APPLICATIONS

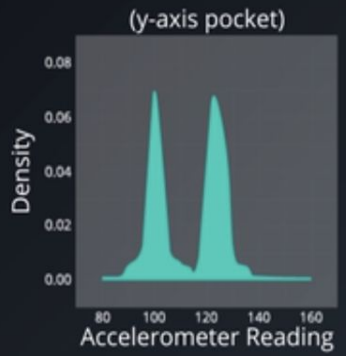
Nonparametric Discovery of Human Routines from Sensor Data

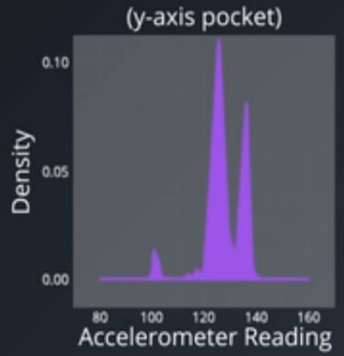
commuting



office work

lunch routine



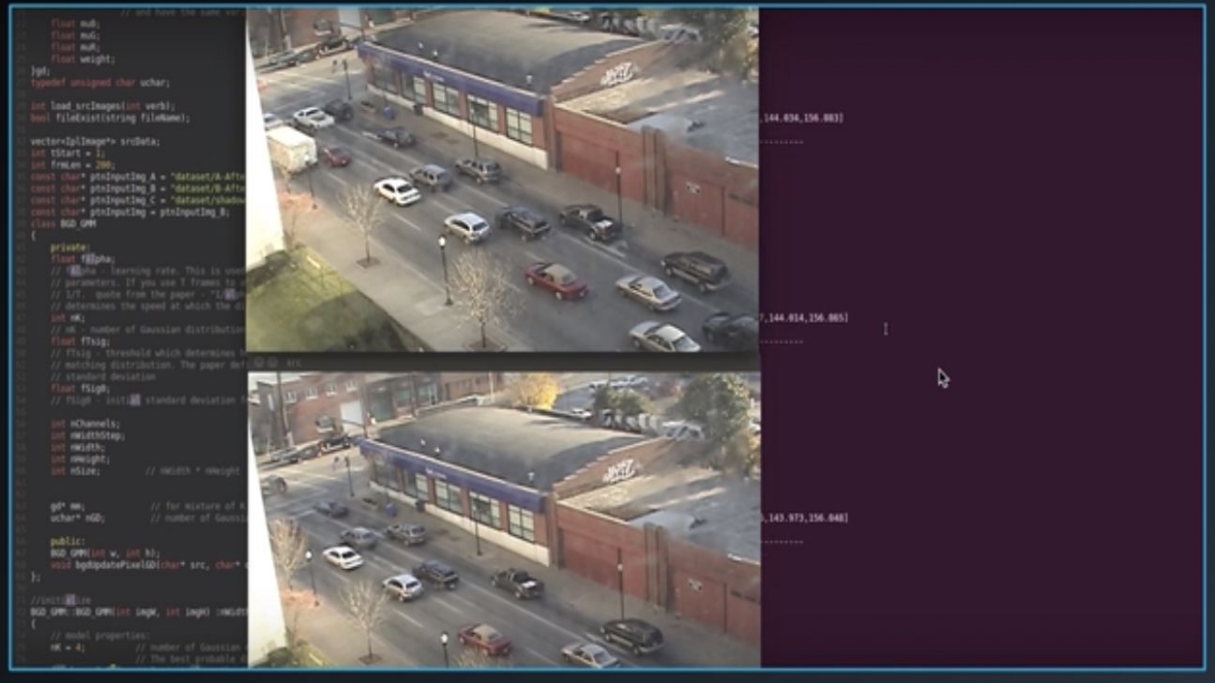


dinner activities

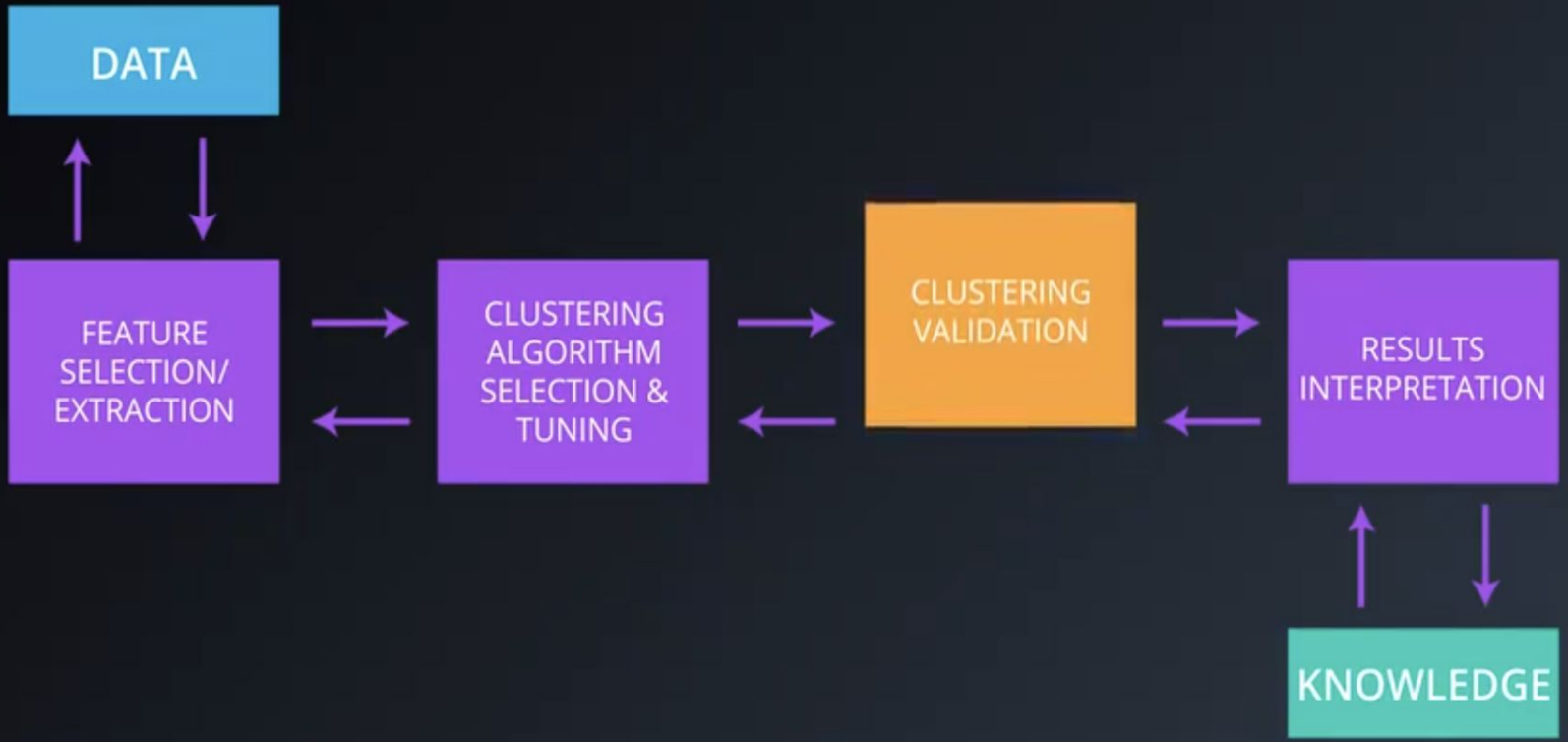
(a) Density distributions of mean of accelerometer data (y-axis pocket) from the daily routine dataset

Feng-Tso Sun
Yi-TIng Yeh
Heng-Tze Cheng
Cynthia Kuo
Martin Griss

GAUSSIAN MIXTURE MODEL CLUSTERING | APPLICATIONS



CLUSTER ANALYSIS



CLUSTER VALIDATION

Categories of cluster validation indices:

- External indices
- Internal indices
- Relative indices

Compactness

Separability

CLUSTER VALIDATION | EXTERNAL INDICES

Matching a clustering structure to information we know beforehand.

Index	Range	Available in sklearn
Adjusted Rand Score	[-1,1]	✓
Fawlks and Mallows	[0,1]	✓
NMI measure	[0,1]	✓
Jaccard	[0,1]	✓
F-measure	[0,1]	✓
Purity	[0,1]	

CLUSTER VALIDATION | EXTERNAL INDICES

Adjusted Rand Index

Between -1 and 1

$$Rand\ Index = \frac{a+b}{\binom{n}{2}}$$

Actual label (c)

Clustering result (K)

a: number of pairs in the same cluster C and in the same cluster in K

CLUSTER VALIDATION | EXTERNAL INDICES

Adjusted Rand Index

Between -1 and 1

Actual label (c)

Clustering result (K)

$$Rand\ Index = \frac{a+b}{\binom{n}{2}} = \frac{2+4}{\binom{5}{2}} = \frac{6}{10}$$

a: number of pairs in the same cluster C and in the same cluster in K	2
b: number of pairs in a different cluster C and in a different cluster in K	4
n: number of samples/points	5

$$ARI = \frac{RI - ExpectedIndex}{max(RI) - ExpectedIndex}$$

CLUSTER VALIDATION | EXTERNAL INDICES

Adjusted Rand Index

Ground truth

K-means 1

ARI = 0.8427

K-means 2

ARI = 0.8693

Random Assignment

ARI = 0.00014

CLUSTER VALIDATION | INTERNAL INDICES

Silhouette coefficient

Between -1 and 1

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

Clustering result

a: average distance to other samples in the same cluster

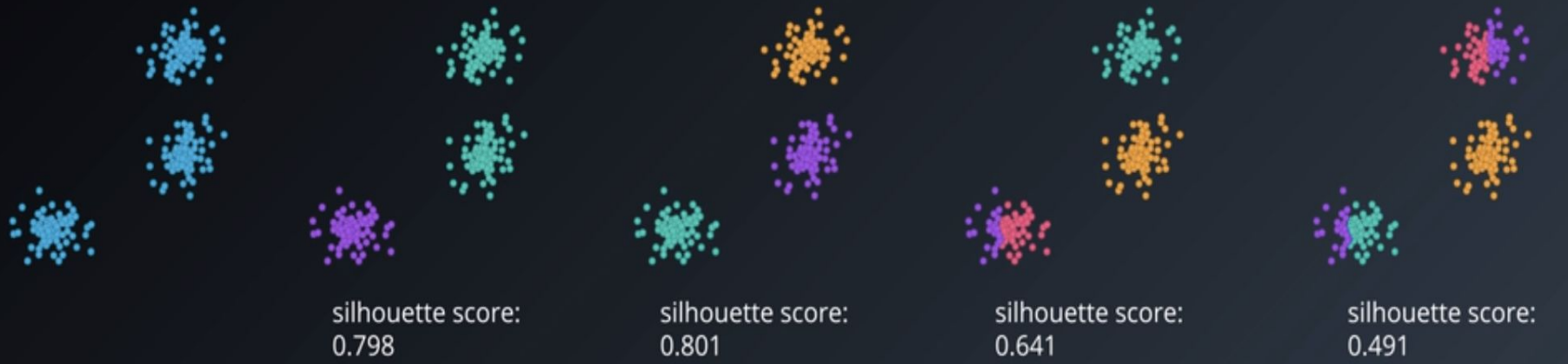
b: average distance to samples in the *closest neighboring* cluster

$$S = average(S_1, S_2, ..., S_n)$$

CLUSTER VALIDATION | INTERNAL INDICES

Silhouette coefficient

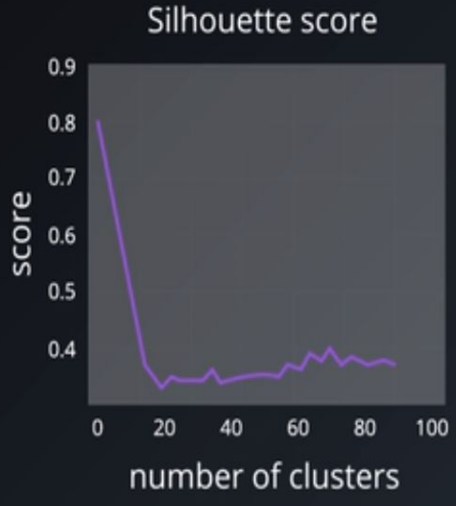
Finding K

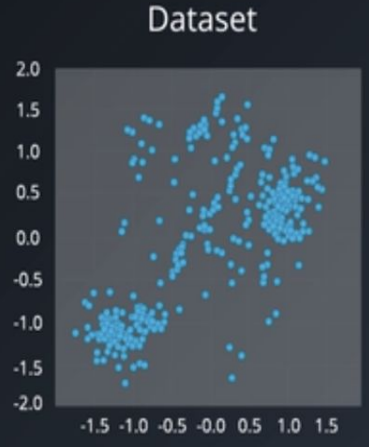


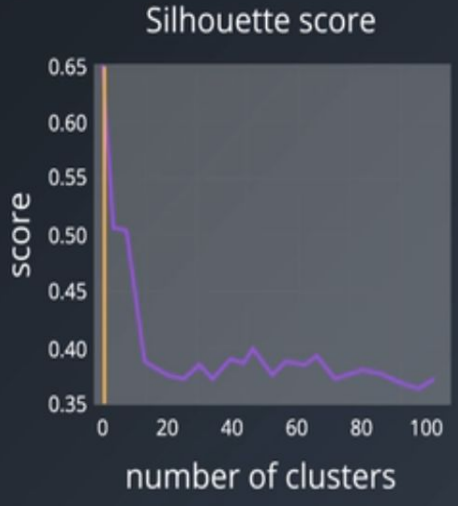
CLUSTER VALIDATION | INTERNAL INDICES

Silhouette coefficient

Finding K







CLUSTER VALIDATION | INTERNAL INDICES

Silhouette coefficient

Comparing clustering algorithm results on a certain dataset

K-means

silhouette score: 0.637

Single Link

silhouette score: -0.08

Complete Link

silhouette score: 0.589

Ward

silhouette score: 0.613

DBSCAN

silhouette score: 0.397

GMM

silhouette score: 0.575

CLUSTER VALIDATION | INTERNAL INDICES

Silhouette coefficient

Comparing Clustering algorithms

D B C V

K-means

silhouette score:
0.355

Single Link

silhouette score:
0.109

Complete Link

silhouette score:
0.294

Ward

silhouette score:
0.334

DBSCAN

silhouette score:
-0.005

GMM

silhouette score:
0.354