NYU CDS, SPRING 2018

INTRODUCTION TO DATA SCIENCE

OFFICE HOURS AND CONTACT INFORMATION

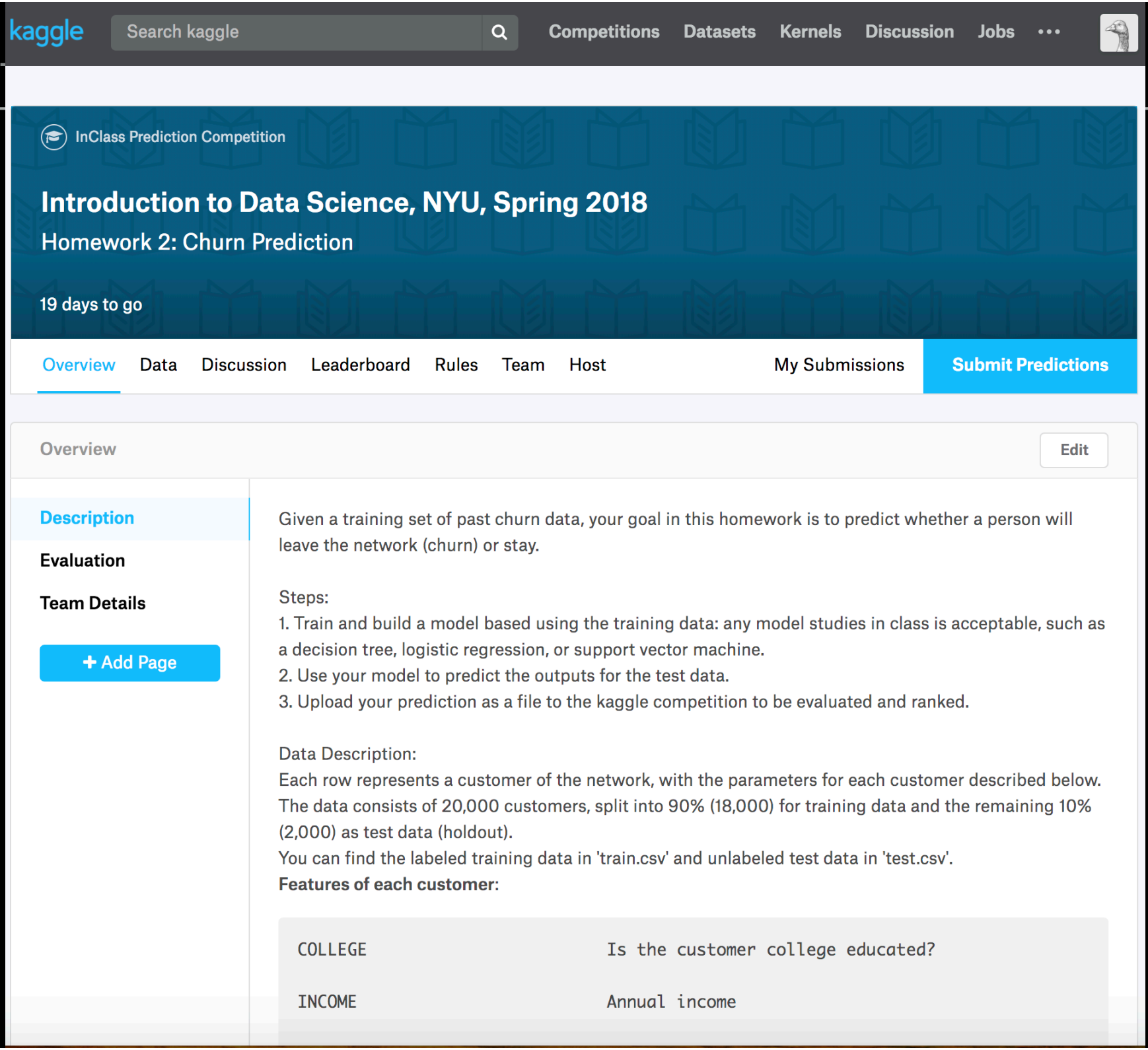
- CDS, 60 5th Avenue, 6th floor, room 620
- Wednesday 12-2pm: Lecturer, Iddo Drori, idrori@nyu.edu
- Tuesday 11am-1pm: Section Leader, Datta Sainath Dwarampudi, ddattasainath@nyu.edu
- Friday 2-4pm: Grader, Samhita Damotharan, sd2941@nyu.edu
- Thursday 2-4pm: Grader, Sai Anirudh Kondaveeti, sak797@nyu.edu

NYU CLASSES

- Resources: lecture and lab notes
- Assignments: homework
- Forums: discussions

HOMEWORK 2

- Kaggle competition: churn prediction
- Submission due Feb. 19.


kaggle
Search kaggle
Competitions
Datasets
Kernels
Discussion
Jobs
InClass Prediction Competition
Introduction to Data Science, NYU, Spring 2018
Homework 2: Churn Prediction
19 days to go
Overview
Data
Discussion
Leaderboard
Rules
Team
Host
My Submissions
Submit Predictions
Overview
Edit
Description
Evaluation
Team Details
+ Add Page
Given a training set of past churn data, your goal in this homework is to predict whether a person will leave the network (churn) or stay.
Steps:
1. Train and build a model based using the training data: any model studies in class is acceptable, such as a decision tree, logistic regression, or support vector machine.
2. Use your model to predict the outputs for the test data.
3. Upload your prediction as a file to the kaggle competition to be evaluated and ranked.
Data Description:
Each row represents a customer of the network, with the parameters for each customer described below.
The data consists of 20,000 customers, split into 90% (18,000) for training data and the remaining 10% (2,000) as test data (holdout).
You can find the labeled training data in 'train.csv' and unlabeled test data in 'test.csv'.
Features of each customer:
COLLEGE Is the customer college educated?
INCOME Annual income

UNSUPERVISED LEARNING: CLUSTERING

- Similarity, distance, KNN classifier
- K-means, hierarchical clustering
- DSB Ch 6

DIMENSIONALITY REDUCTION

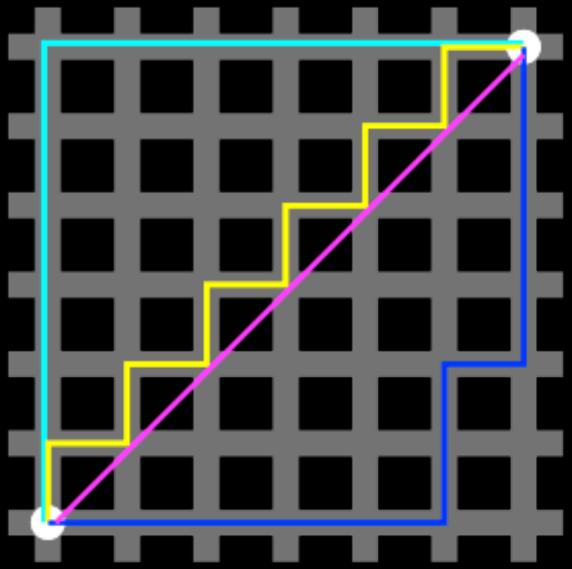
- t-SNE

SIMILARITY AND DISTANCE

- Data as points in high dimensional space
- Distance between points:
  - Euclidean $\sqrt{\sum (x_i - y_i)^2}$
  - Manhattan $\sum |x_i - y_i|$
  - Jaccard $d_J(X,Y) = 1 - J(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$
  - Cosine $d_C(X,Y) = 1 - C(X,Y) = 1 - \cos(\theta) = 1 - \frac{X \cdot Y}{\|X\|_2 \|Y\|_2}$

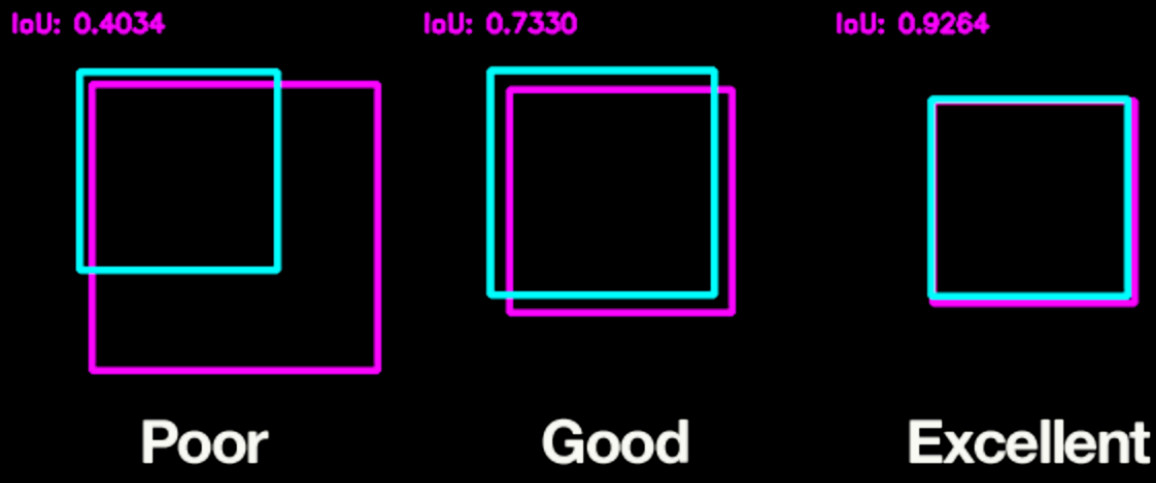
K NEAREST NEIGHBORS

- Euclidean distance $\sqrt{\sum(x_i - y_i)^2}$
- Manhattan distance $\sum|x_i - y_i| = \sum|x_i - y_i| = \sum|x_i - y_i|$



JACCARD INDEX

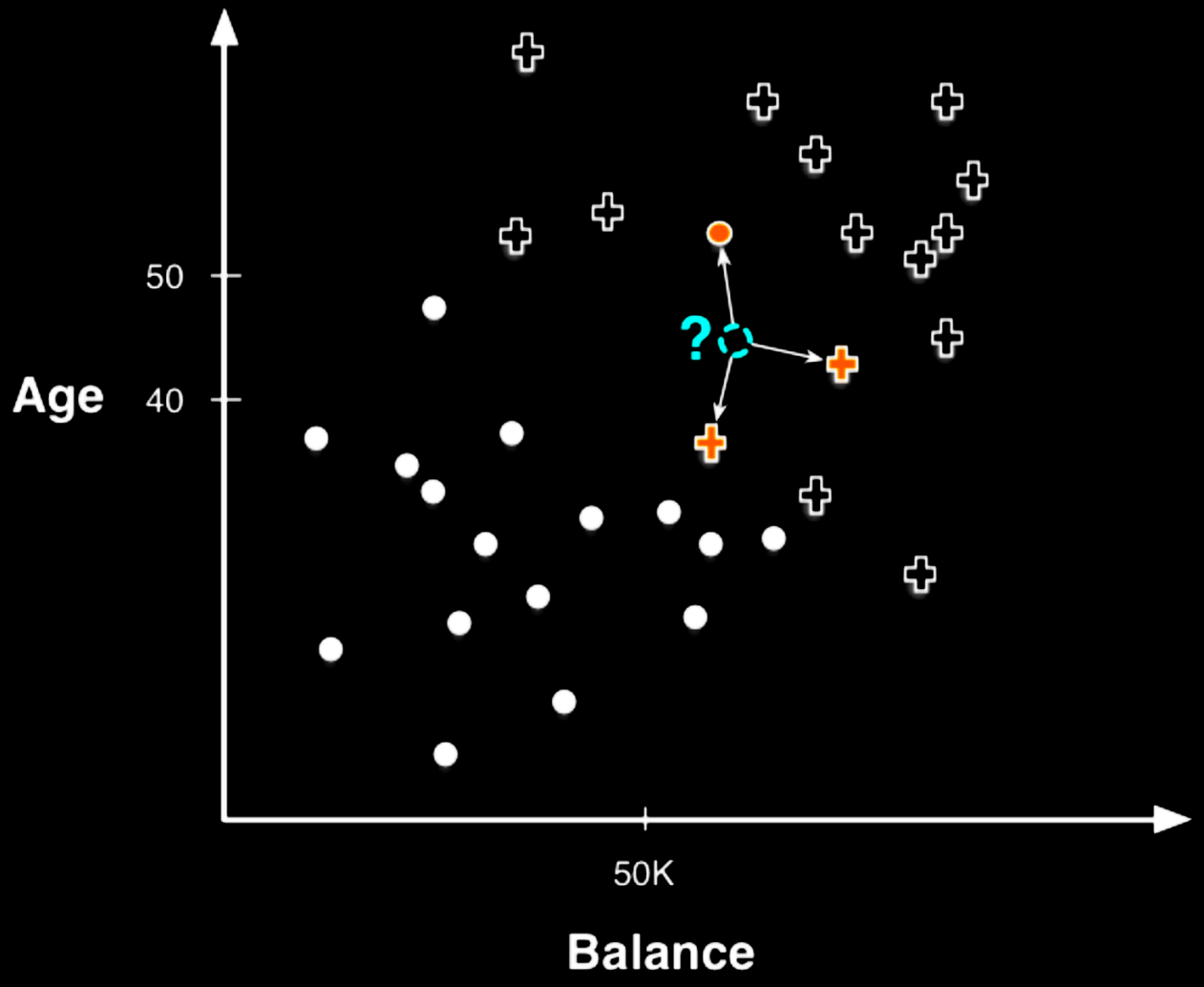
$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$



K NEAREST NEIGHBORS

- Classification
- Regression

K NEAREST NEIGHBORS

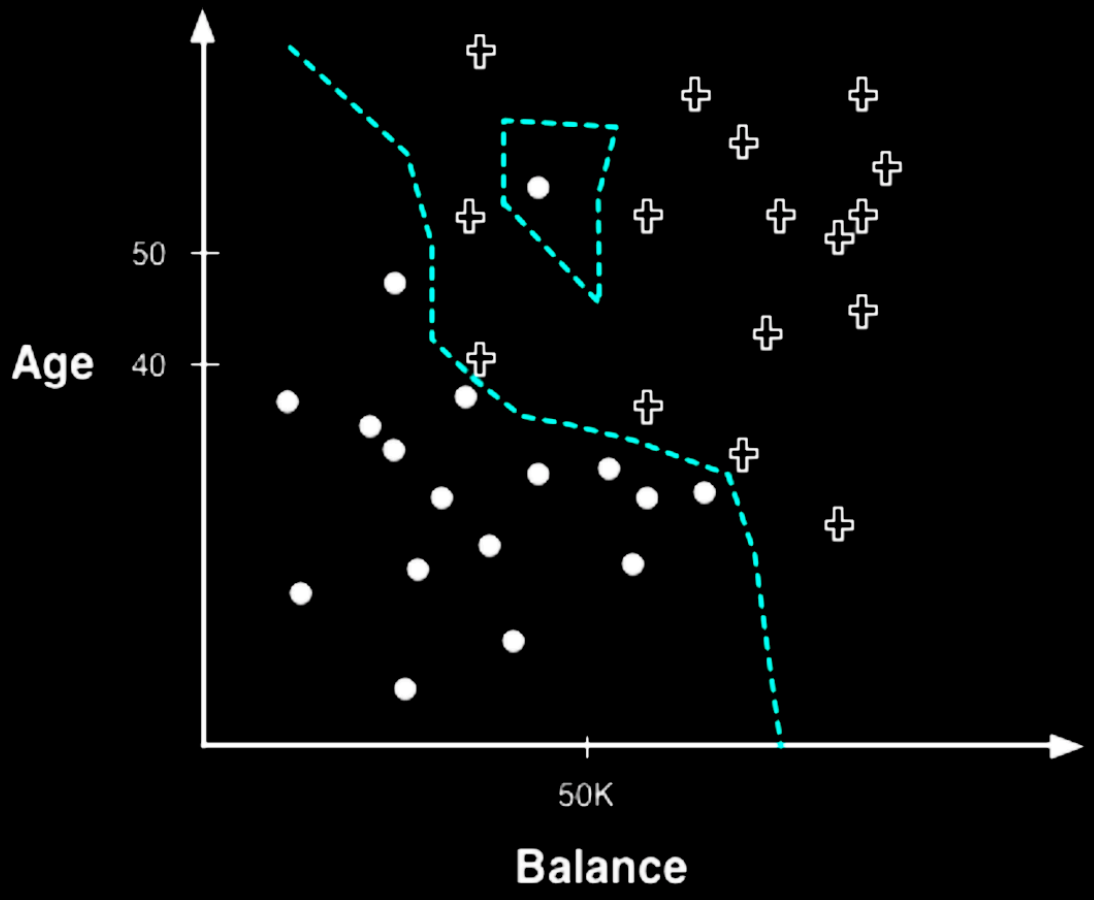


K NEAREST NEIGHBORS CLASSIFICATION

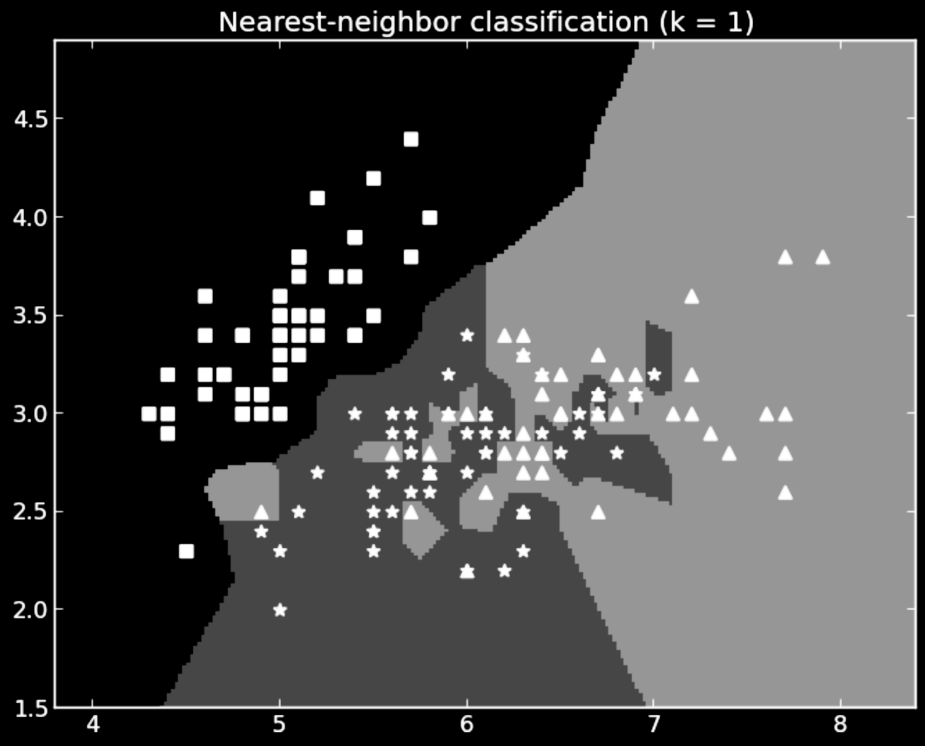
- Majority vote

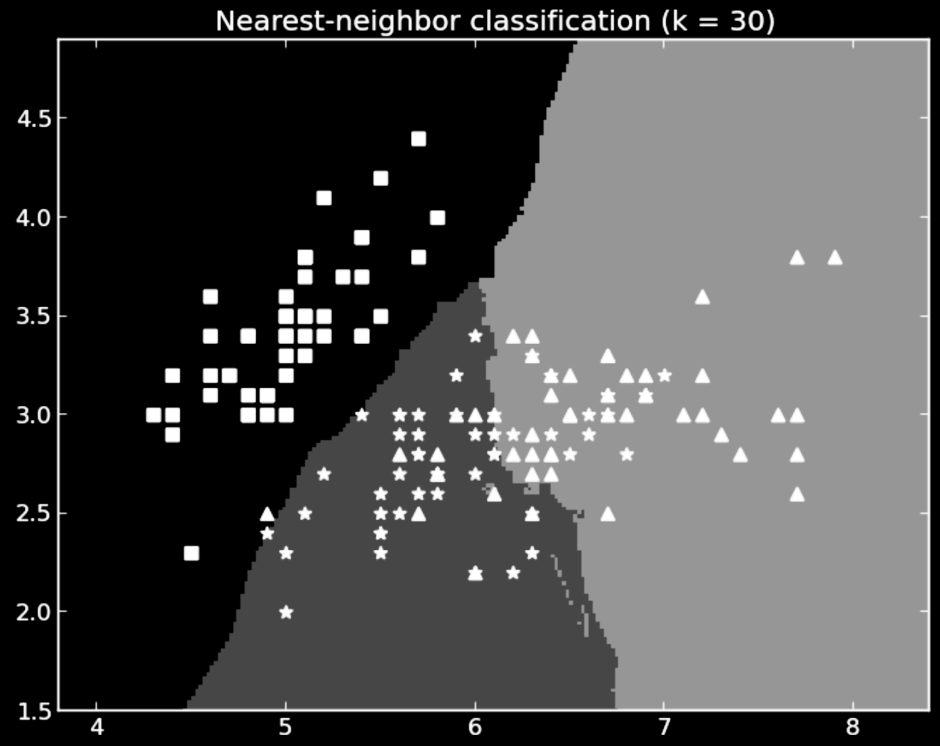
$$c(x) = \arg\max_{c} \sum_{y \in KNN(x)} \left[class(y) = c \right]$$

K NEAREST NEIGHBORS



K NEAREST NEIGHBORS





K NEAREST NEIGHBORS CLASSIFICATION

- Weighted by similarity

$$w(x,y) = \frac{1}{dist(x,y)^2}$$

$$p(c \mid x) = \frac{\sum_{y \in KNN(x)} w(x,y) \cdot [class(y) = c]}{\sum_{y \in KNN(x)} w(x,y)}$$

K NEAREST NEIGHBORS REGRESSION

- Weighted by similarity

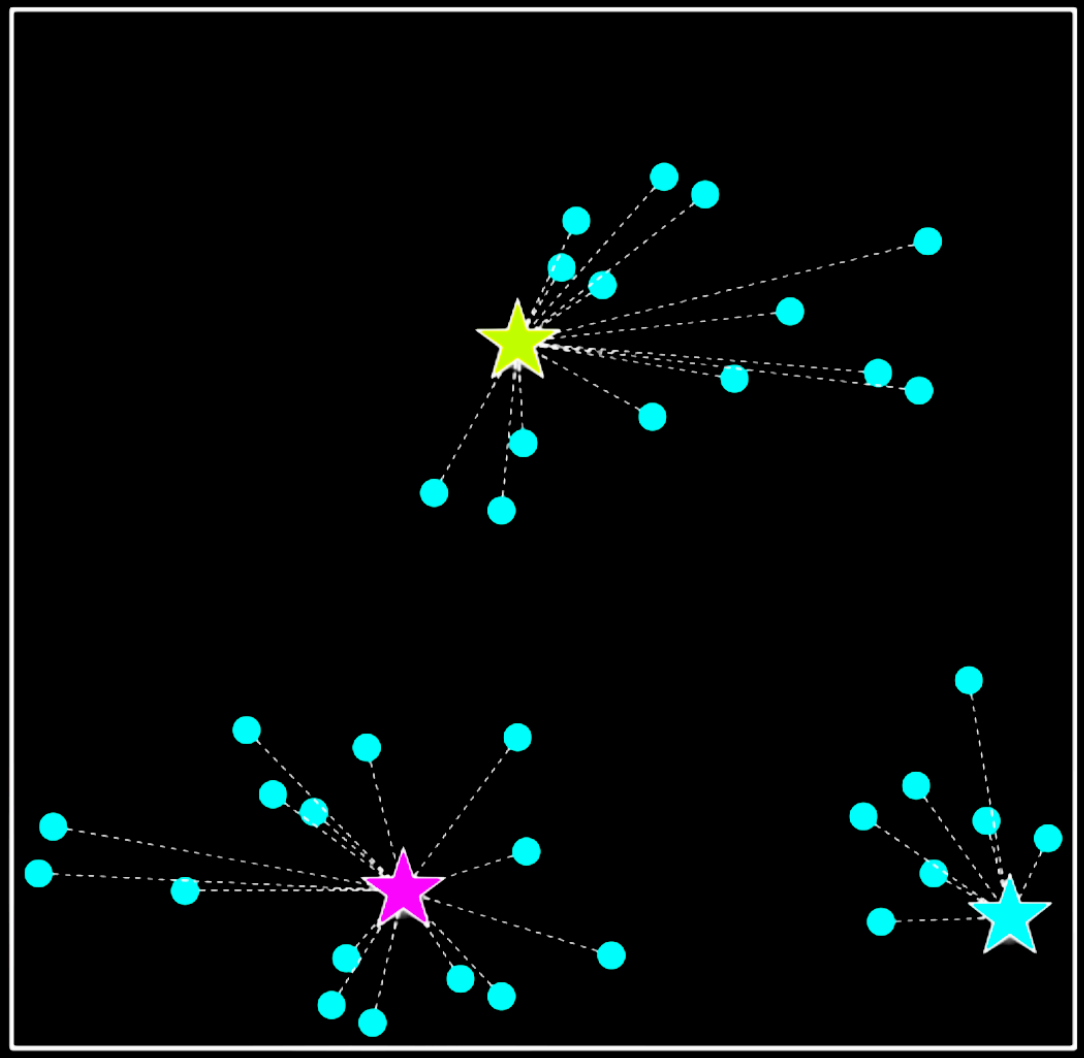
$$w(x,y) = \frac{1}{dist(x,y)^2}$$

$$f(x) = \frac{\sum_{y \in KNN(x)} w(x,y) \cdot t(y)}{\sum_{y \in KNN(x)} w(x,y)}$$

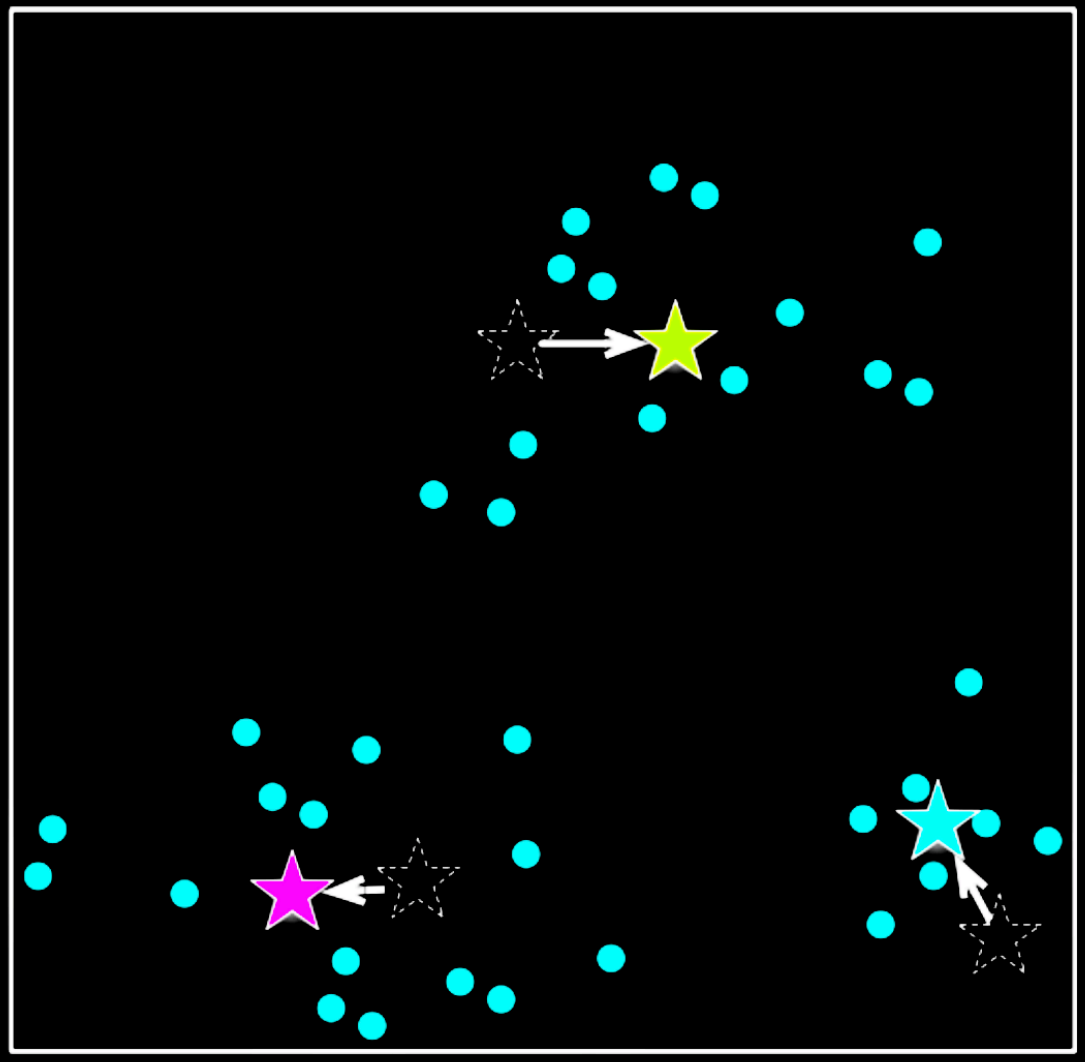
CLUSTERING

- Basic experimental procedure in data analysis
- Used in all natural and social sciences: biology, astronomy, psychology, medicine, chemistry.
- Goals:
  - Reliably achieve high accuracy across domains
  - Handle high data dimensionality
  - Scale to large datasets

K-MEANS CLUSTERING



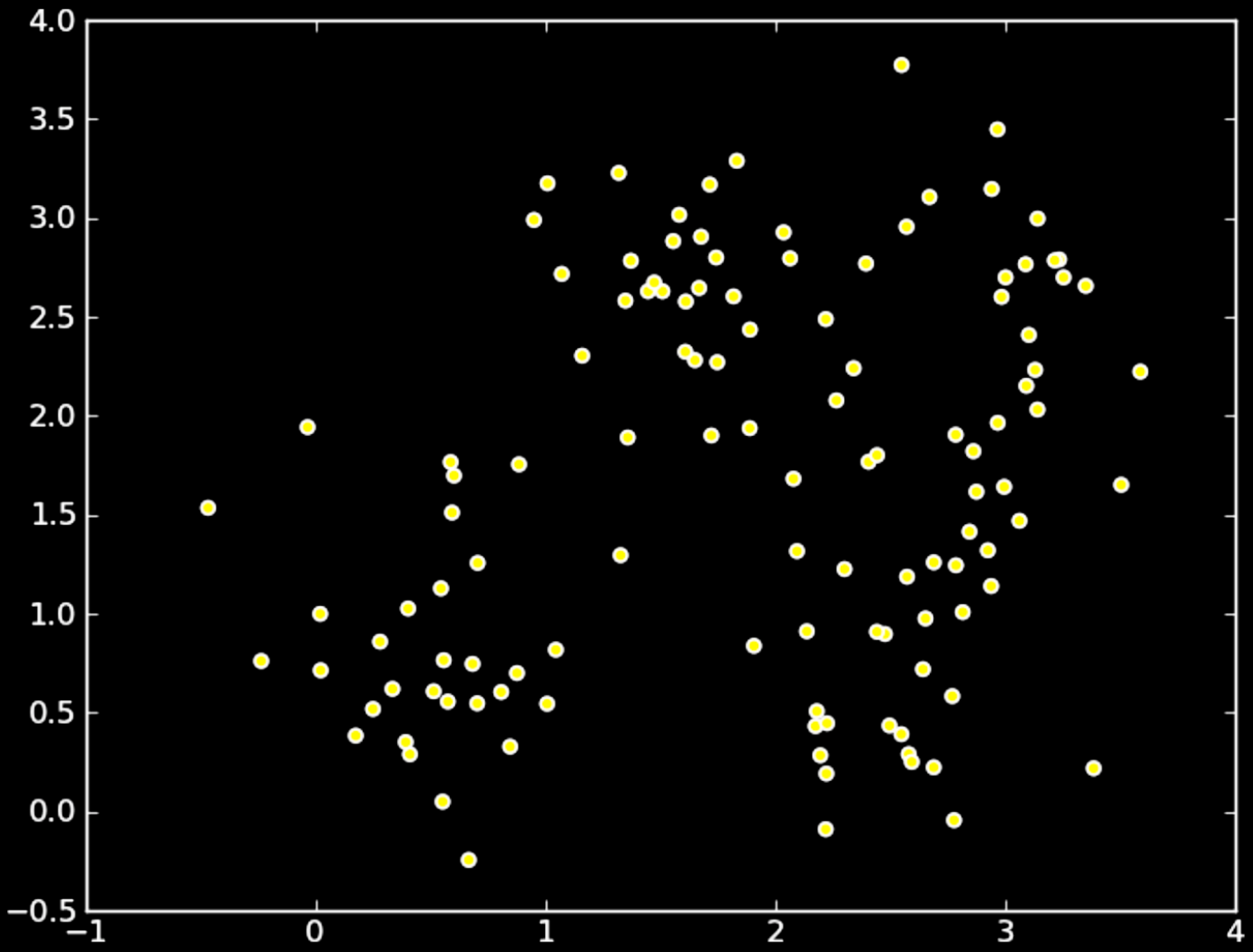
K-MEANS CLUSTERING



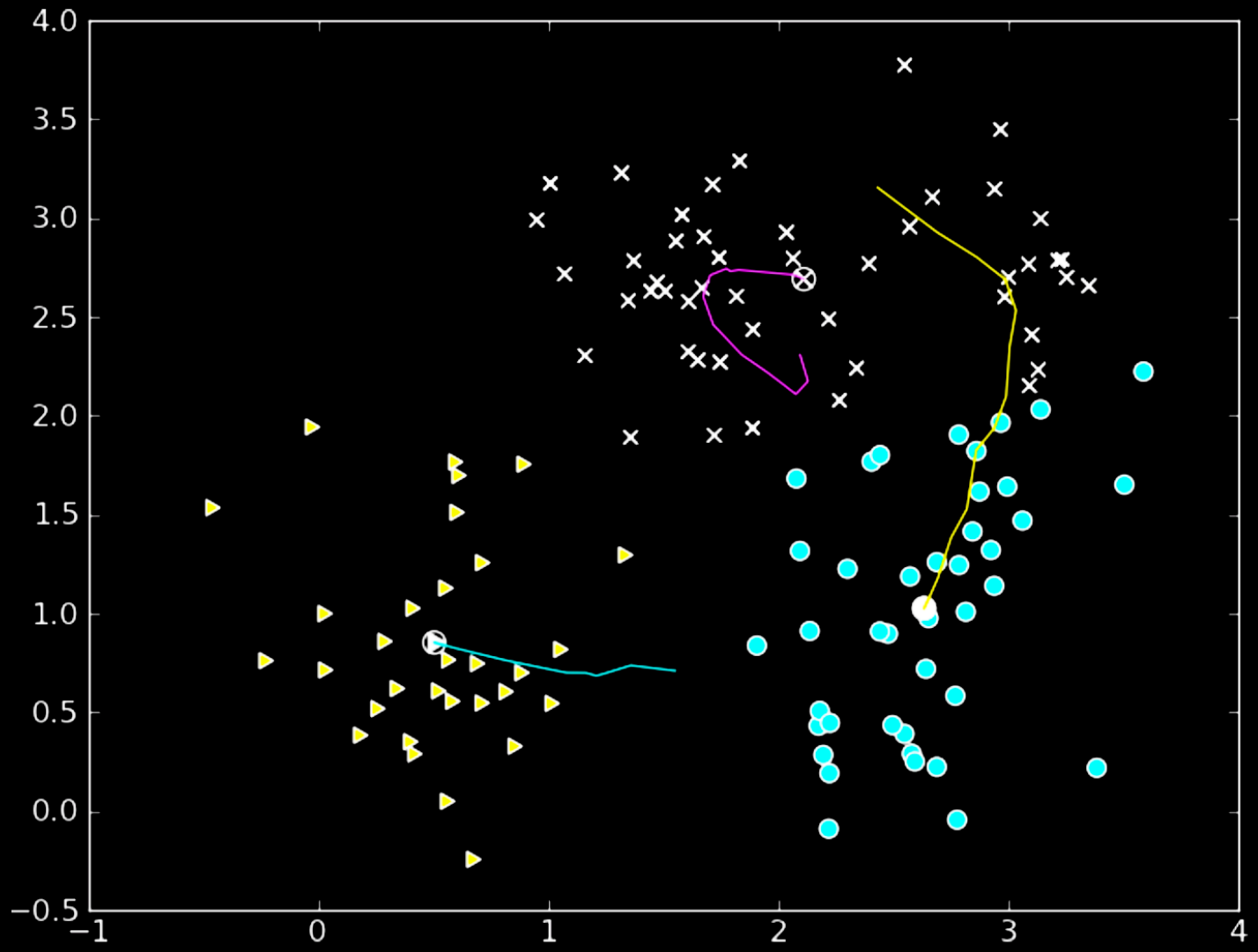
K-MEANS CLUSTERING

- Choose k
- Initialize cluster centroids as k random examples
- Repeat until convergence
  - For each example: find its nearest cluster centroid and label example as belonging to that cluster.
  - For each cluster centroid: update to mean of its labeled examples.

K-MEANS CLUSTERING



K-MEANS CLUSTERING



K-MEANS CLUSTERING

$$\min_{c(x_i)} \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \mu_{c(x_i)} \right)^2$$

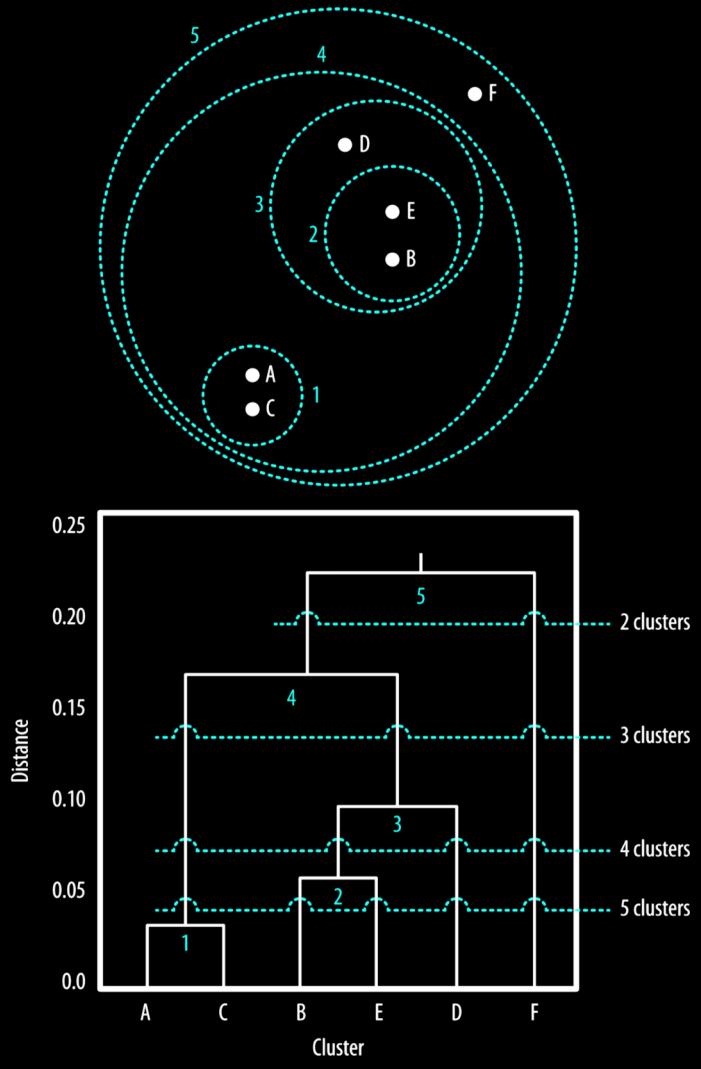
K-MEANS CLUSTERING

- Choose number of clusters k
  - Choose number of tests t
  - For each test run k-means
  - Select best clustering among t test
- Select k

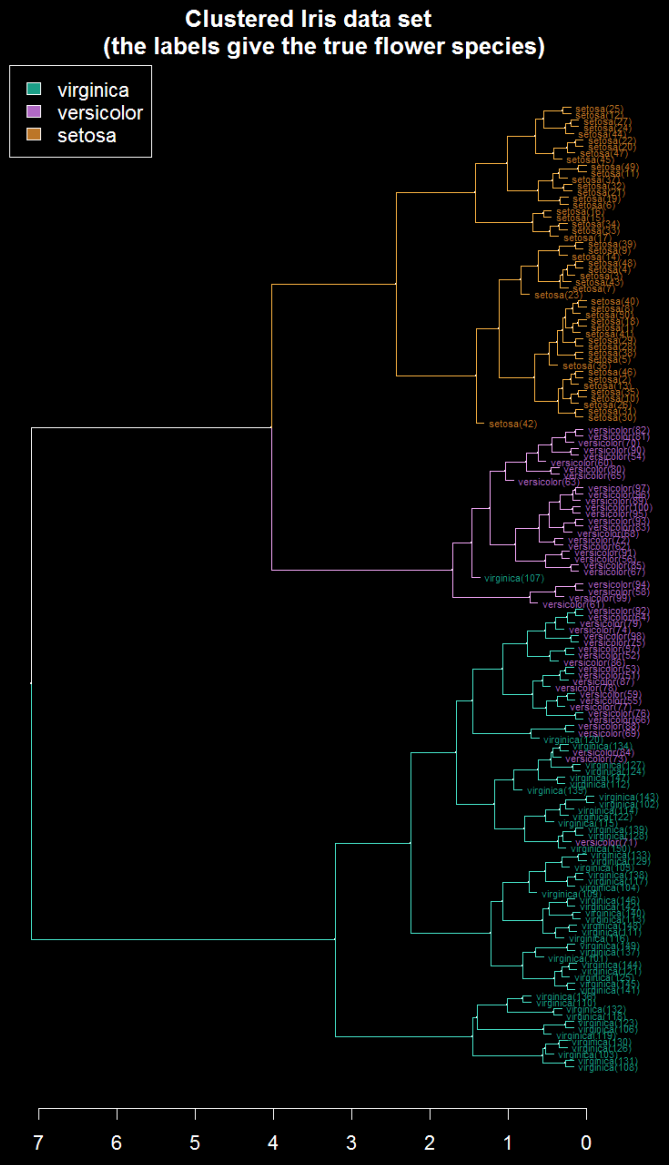
HIERARCHICAL CLUSTERING

- Bottom up: agglomerative
  - Each sample starts in its own cluster, and pairs of clusters are merged
- Top down: divisive
  - All samples start in one cluster, and splits are performed recursively

HIERARCHICAL CLUSTERING: DENDROGRAM

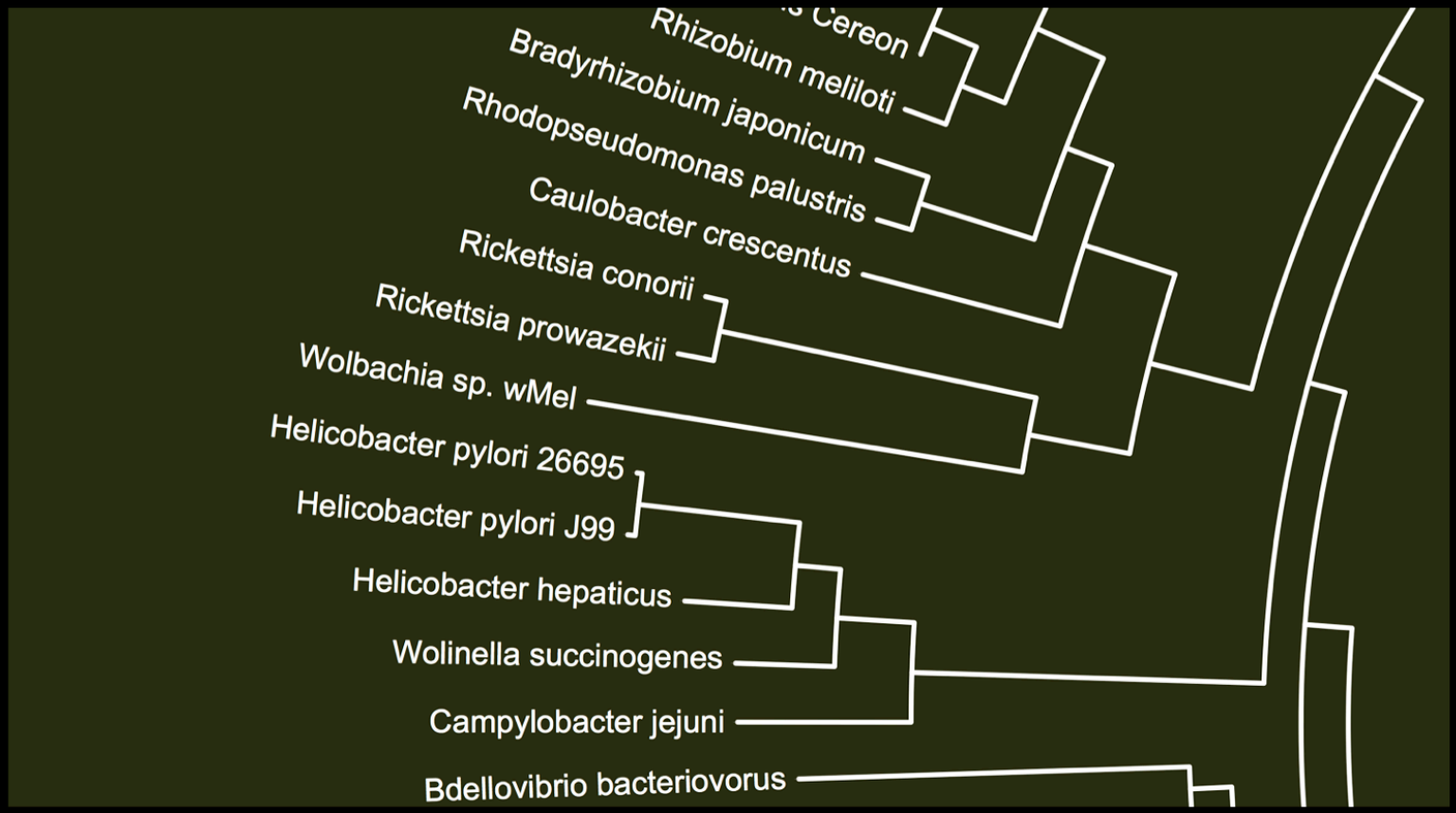


HIERARCHICAL CLUSTERING: DENDROGRAM



HIERARCHICAL CLUSTERING: PHYLOGENETIC TREE OF LIFE

HIERARCHICAL CLUSTERING



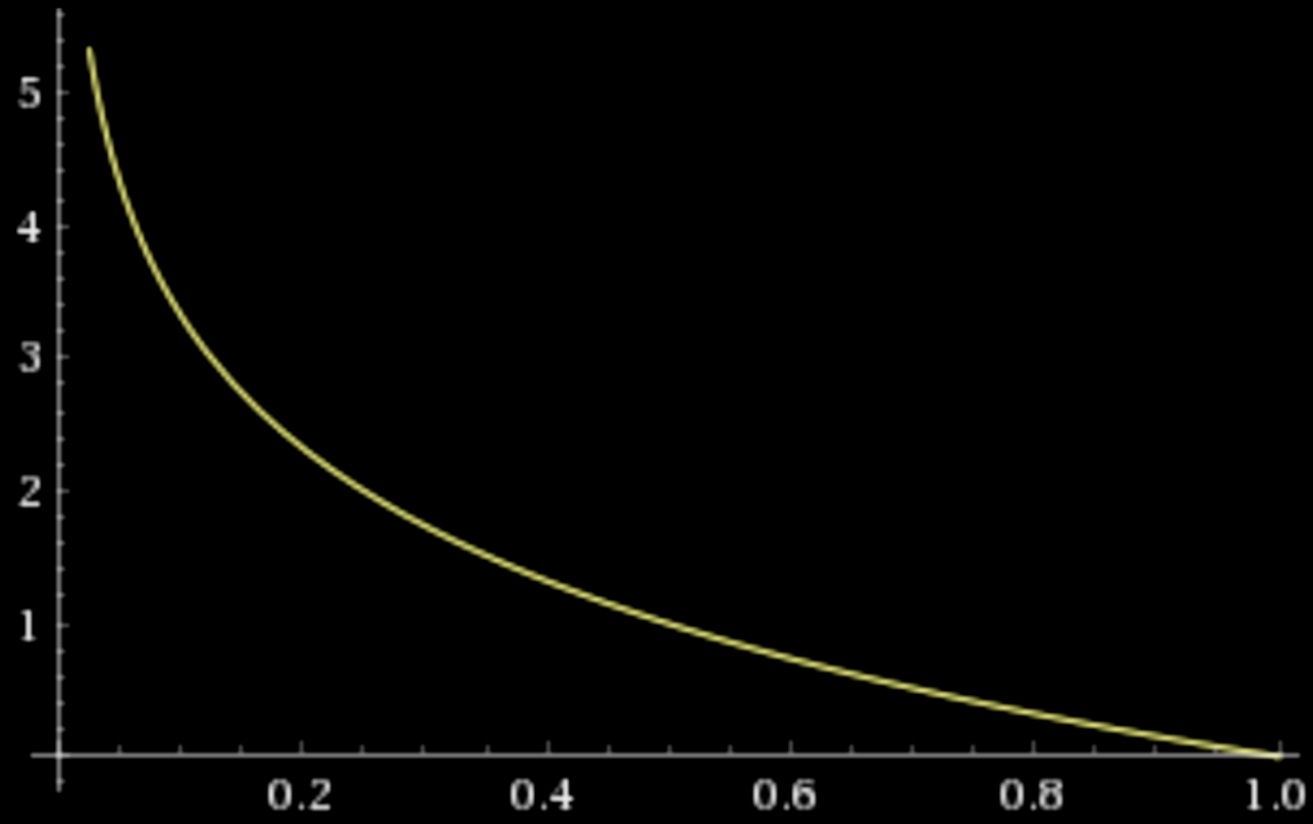
CLUSTERING QUALITY

- Normalized mutual information

$$\frac{I(X;Y)}{\sqrt{H(X)H(Y)}}$$

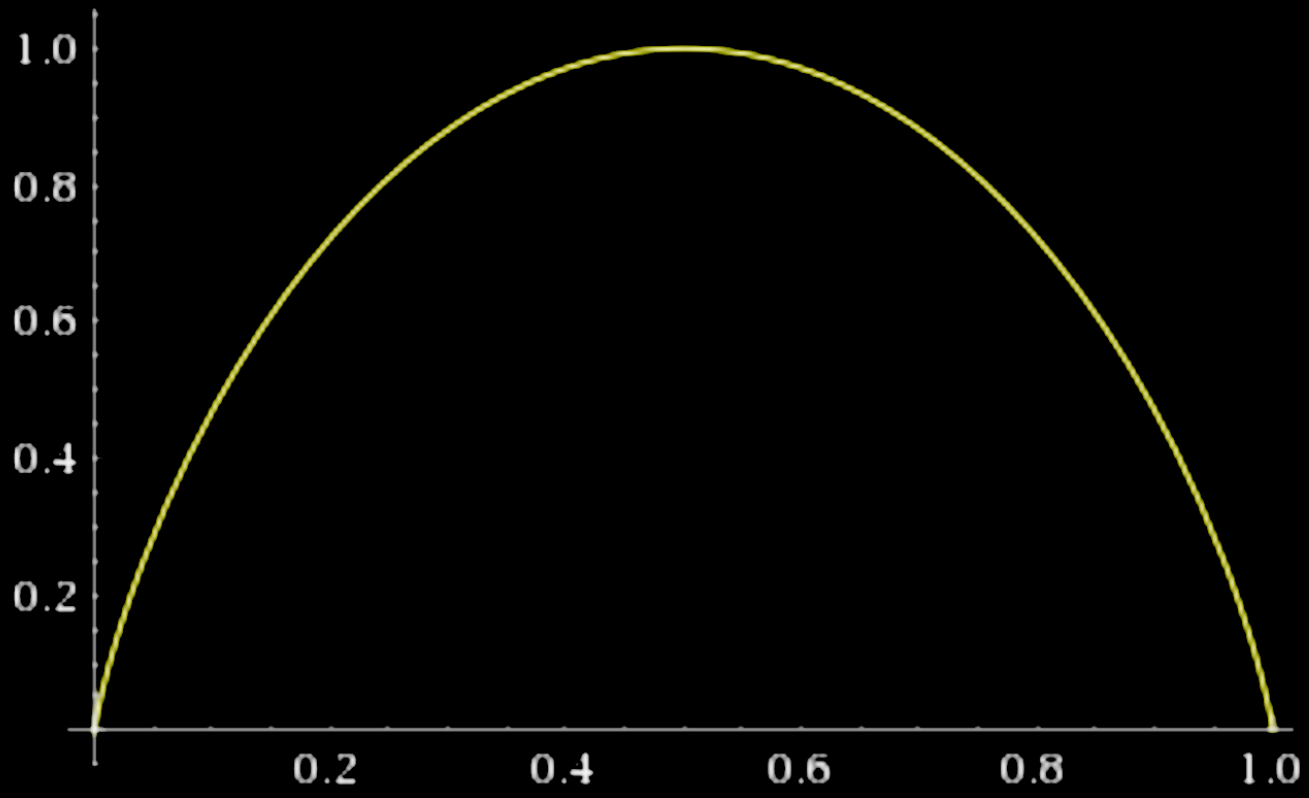
- Adjusted mutual information
- Rand index

INFORMATION



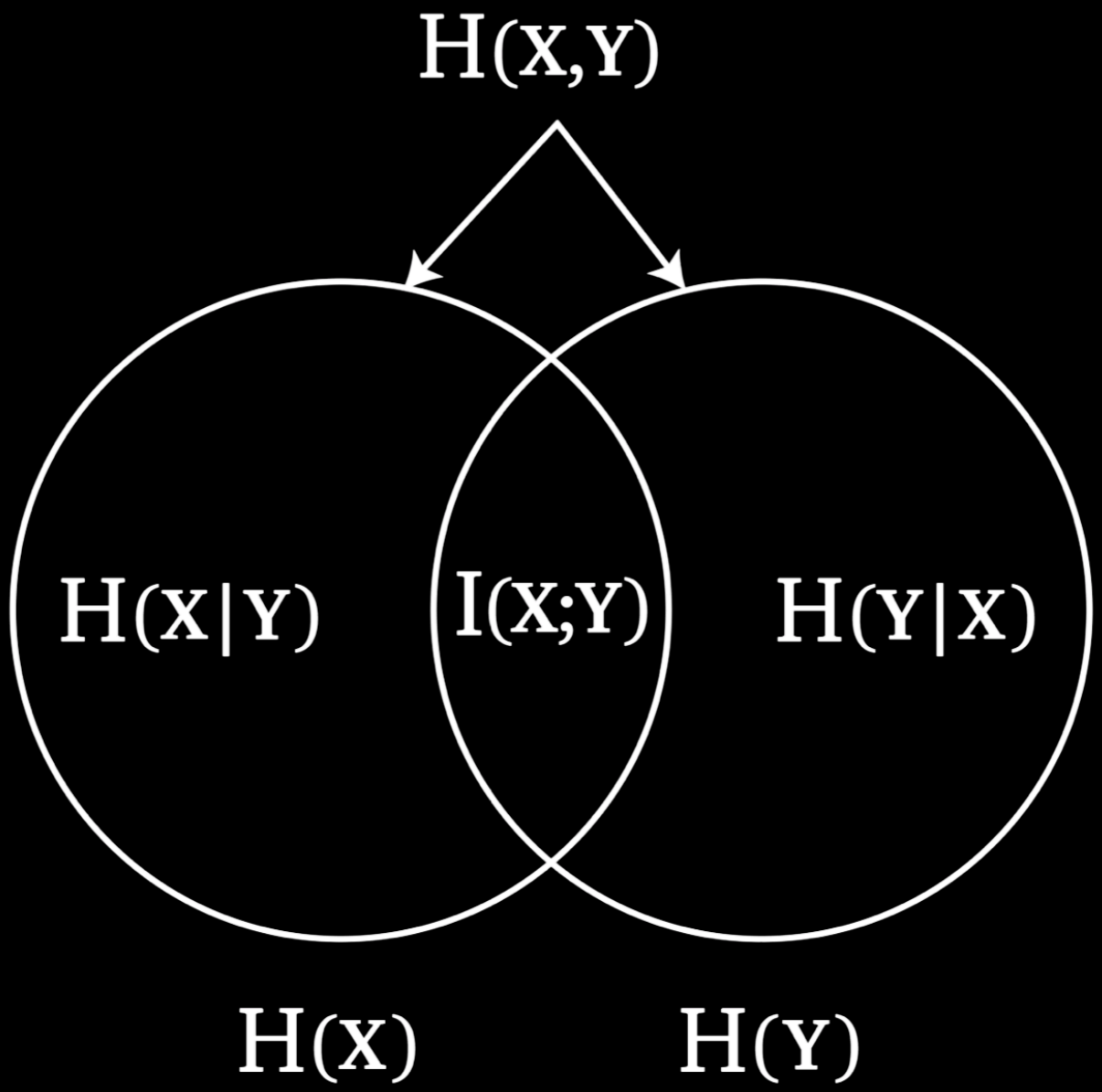
Information: $-\log_2(p)$ where $p \in [0, 1]$

ENTROPY



Entropy: $-p\log_2(p) - (1-p)\log_2(1-p)$ where $p \in [0,1]$

MUTUAL INFORMATION



MUTUAL INFORMATION

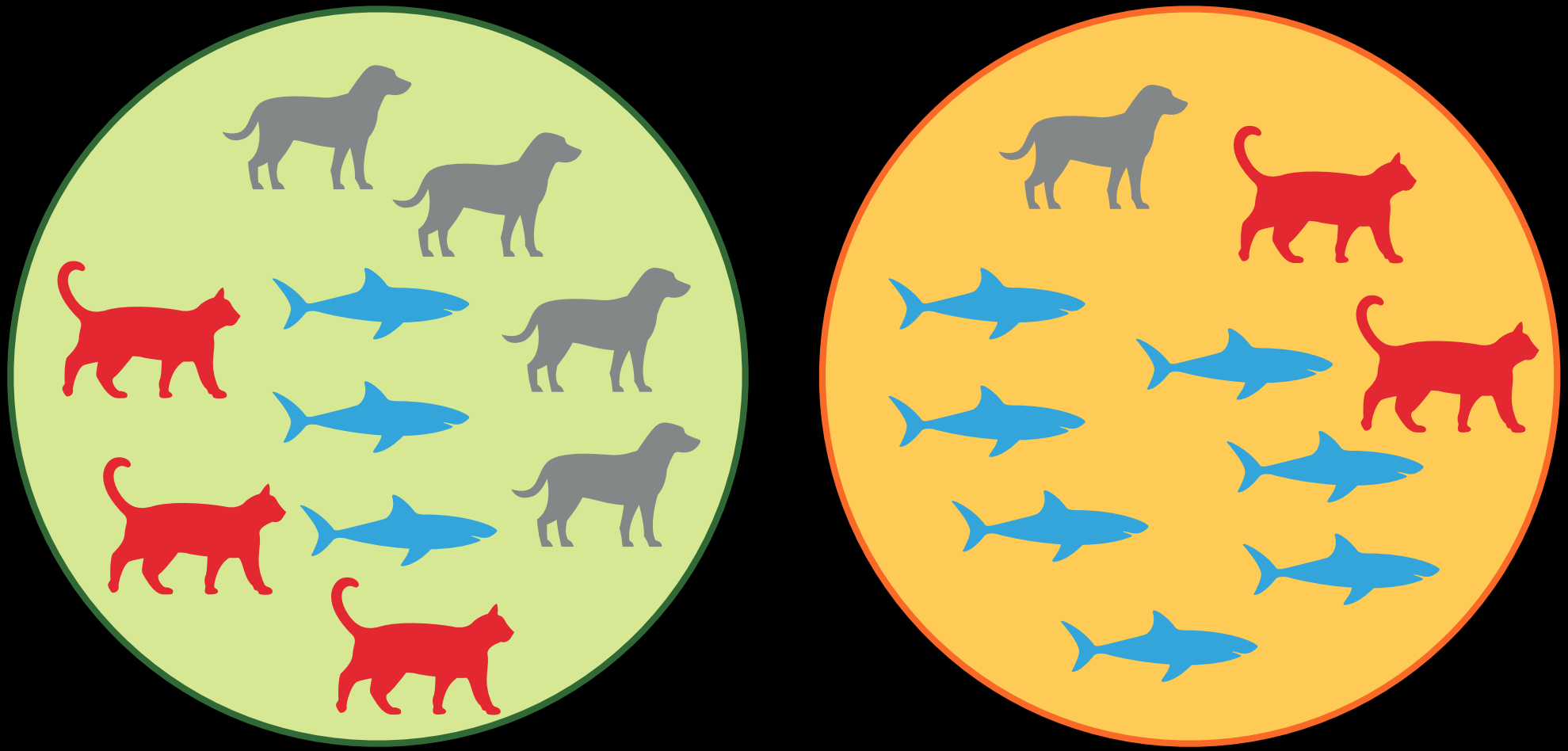
- $I(X;Y) = H(X) - H(X|Y)$
- Reduction in uncertainty of X due to knowledge of Y

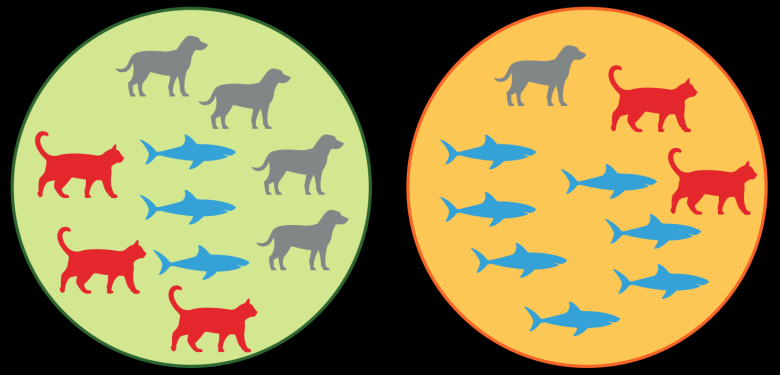
CLUSTERING QUALITY

- Normalized mutual information
- Between class labels Y and cluster labels C

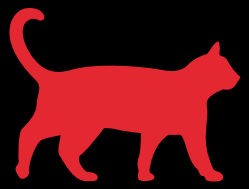
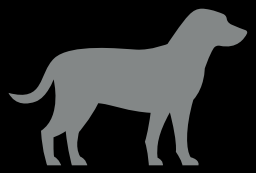
$$\frac{I(Y;C)}{\sqrt{H(Y)H(C)}}$$

CLUSTERING QUALITY

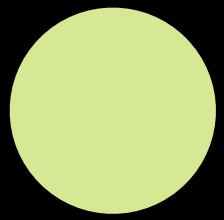
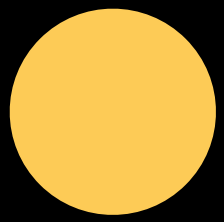


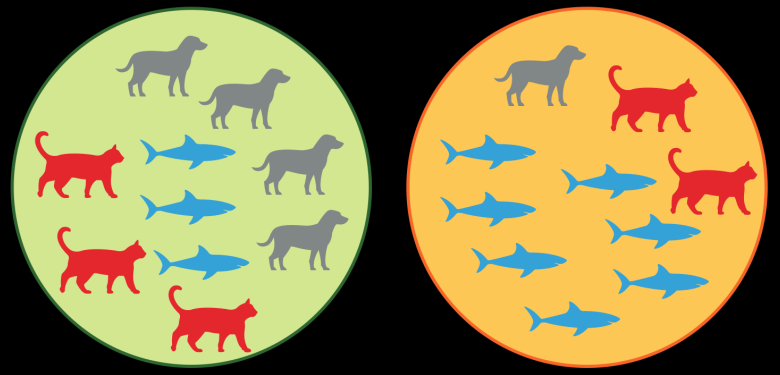


CLUSTERING QUALITY

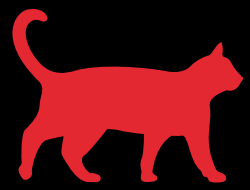
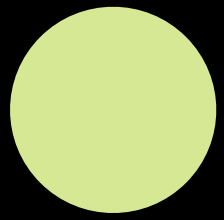
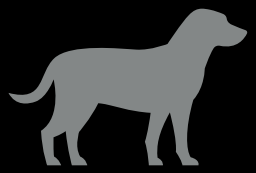
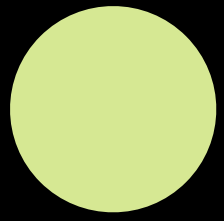
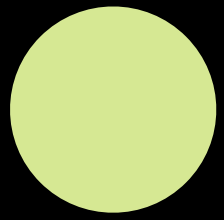
- Probability () = 5/20 = 1/4
- Probability () = 5/20 = 1/4
- Probability () = 10/20 = 1/2
- H(Y) = - 1/4log(1/4) - 1/4log(1/4) - 1/2log(1/2) = 3/2

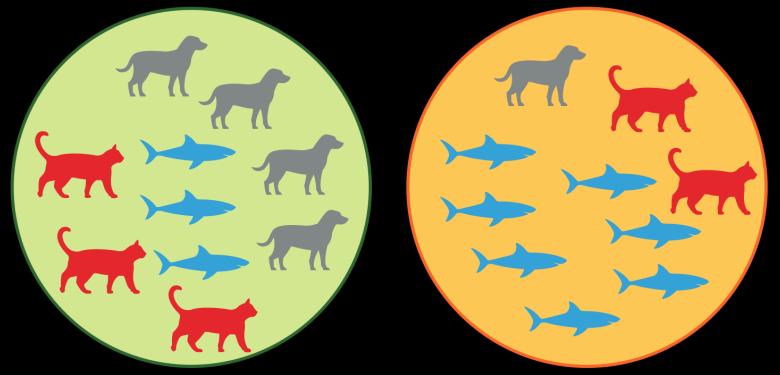
CLUSTERING QUALITY

- Probability () = 10/20 = 1/2
- Probability () = 10/20 = 1/2
- H(C) = - 1/2log(1/2) - 1/2log(1/2) = 1

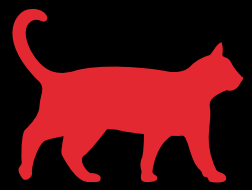
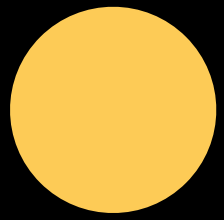
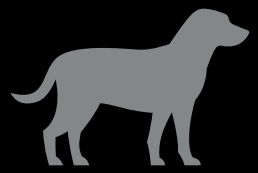
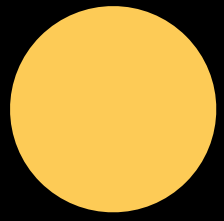
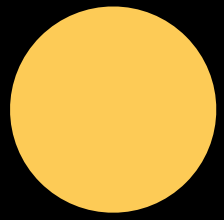


CLUSTERING QUALITY

- Probability ( | ) = 3/10
- Probability ( | ) = 4/10 = 2/5
- Probability ( | ) = 3/10
- H(Y|C=1) = -P(C=1) sum_y P(Y=y|C=1)log(P(Y=y|C=1))



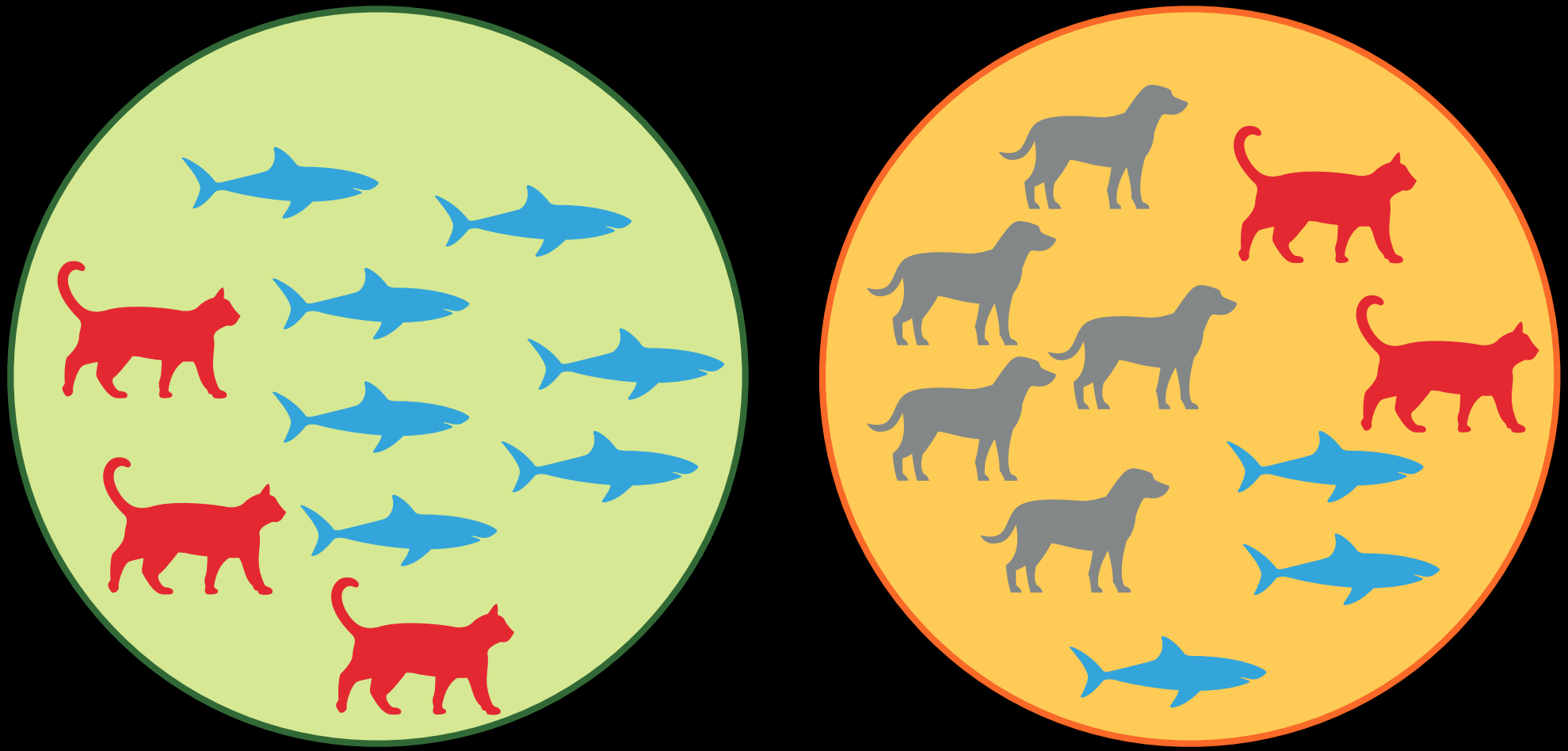
CLUSTERING QUALITY

- Probability ( | ) = 2/10
- Probability ( | ) = 1/10
- Probability ( | ) = 7/10
- H(Y|C=2) = -P(C=2) sum_y P(Y=y|C=2)log(P(Y=y|C=2))

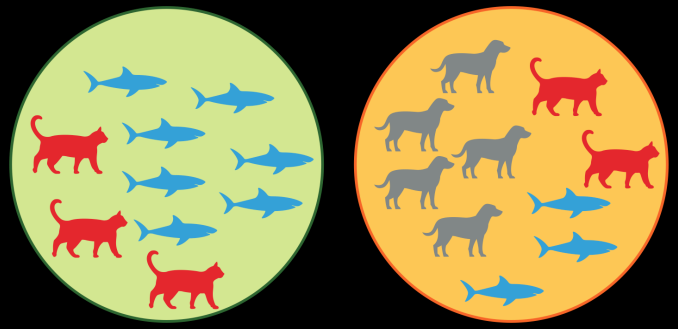
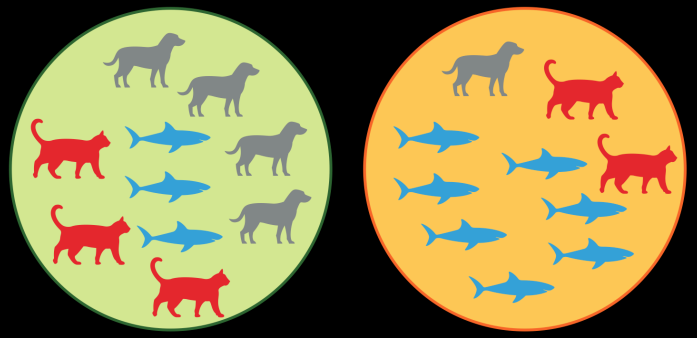
CLUSTERING QUALITY

- $H(Y|C) = H(Y|C=1) + H(Y|C=2)$
- $I(Y;C) = H(Y) - H(Y|C)$

CLUSTERING QUALITY



CLUSTERING QUALITY

- NMI() > NMI ()

DIMENSIONALITY REDUCTION

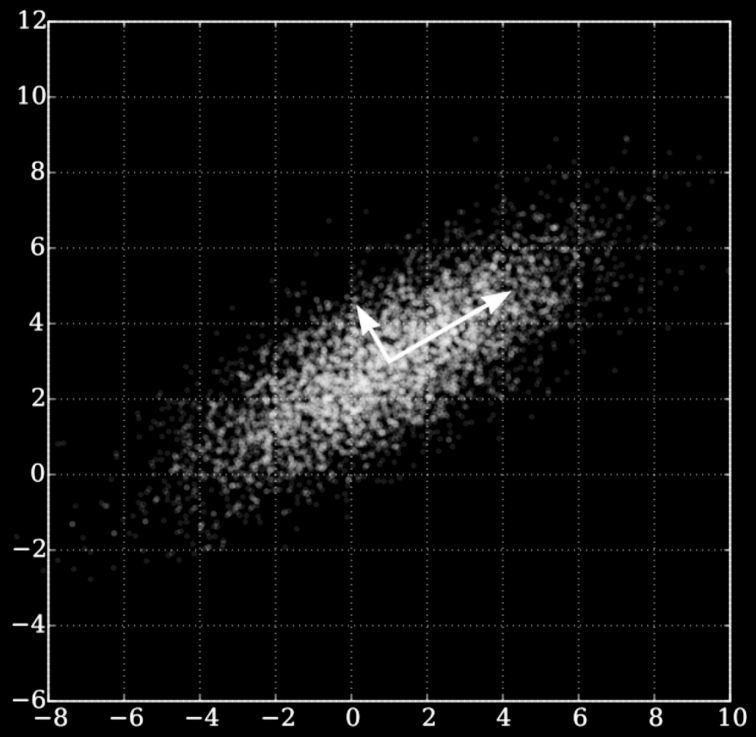
- Computation, storage
- Visualization

DIMENSIONALITY REDUCTION: PCA

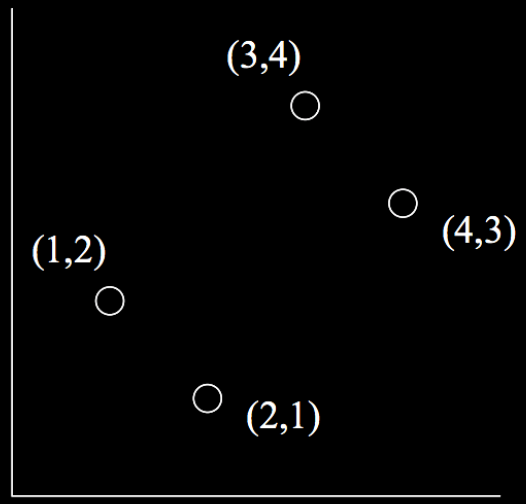
- Principle component analysis (PCA)
  - Maximize variance
  - Minimize mean squared error
- Find new basis in which vectors maximize variance
- Orthogonal linear transformation of data to new coordinate system such that greatest variance by projection of data is on 1st coordinate (first principal component), 2nd greatest variance is on 2nd coordinate, and so on

DIMENSIONALITY REDUCTION: PCA

- Subtract mean, so new mean is 0
- Variance is then $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$



DIMENSIONALITY REDUCTION: PCA



$$M = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix}$$

DIMENSIONALITY REDUCTION: PCA

$$M^{\mathrm{T}}M = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix}$$

$$(30 - \lambda)(30 - \lambda) - 28 \times 28 = 0$$

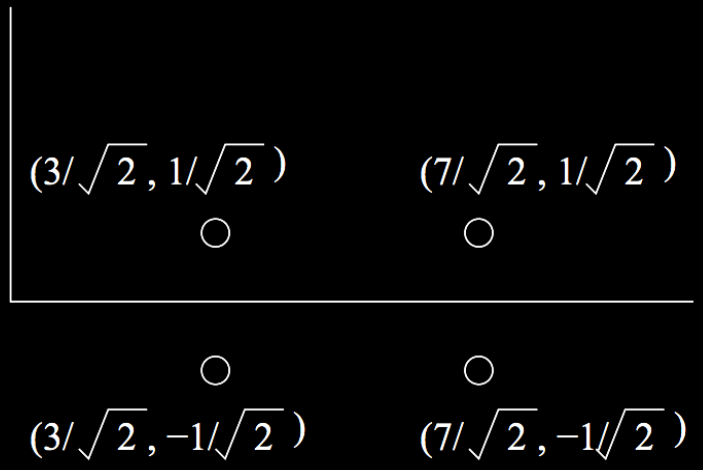
$$\lambda = 58 \text{ and } \lambda = 2$$

$$\begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 58 \begin{bmatrix} x \\ y \end{bmatrix} \qquad \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

$$\begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 2 \begin{bmatrix} x \\ y \end{bmatrix} \qquad \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

$$E = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

DIMENSIONALITY REDUCTION: PCA



$$ME = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 3/\sqrt{2} & 1/\sqrt{2} \\ 3/\sqrt{2} & -1/\sqrt{2} \\ 7/\sqrt{2} & 1/\sqrt{2} \\ 7/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

DIMENSIONALITY REDUCTION: PCA

E_k be the first k columns of E

ME_k is a k-dimensional representation of M

$$ME_1 \quad \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 3/\sqrt{2} \\ 3/\sqrt{2} \\ 7/\sqrt{2} \\ 7/\sqrt{2} \end{bmatrix}$$

DIMENSIONALITY REDUCTION: T-SNE

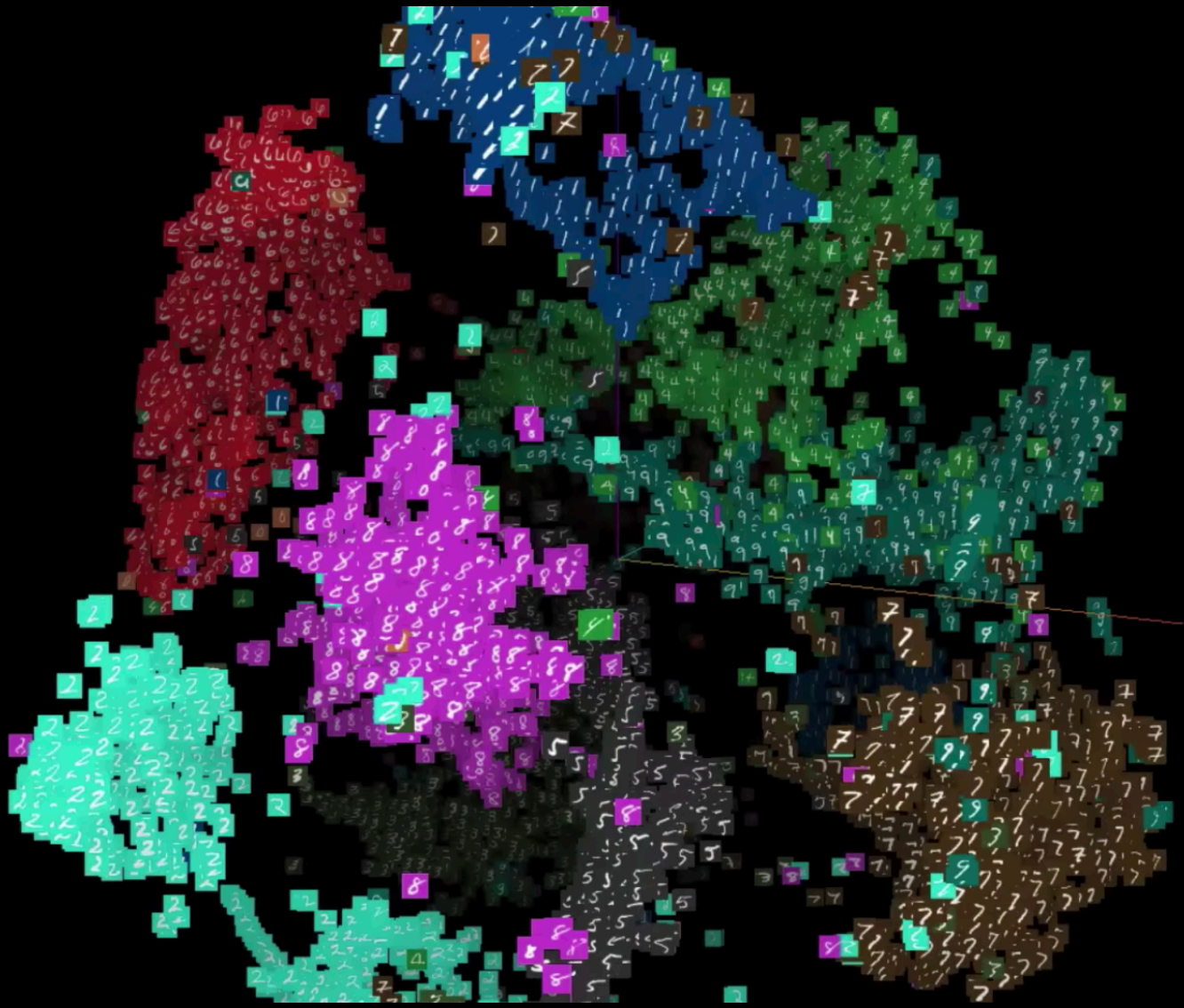
- t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Map high-dimensional points to 2D/3D points such that:
  - Similar points map to nearby points
  - Dissimilar points map to distant points

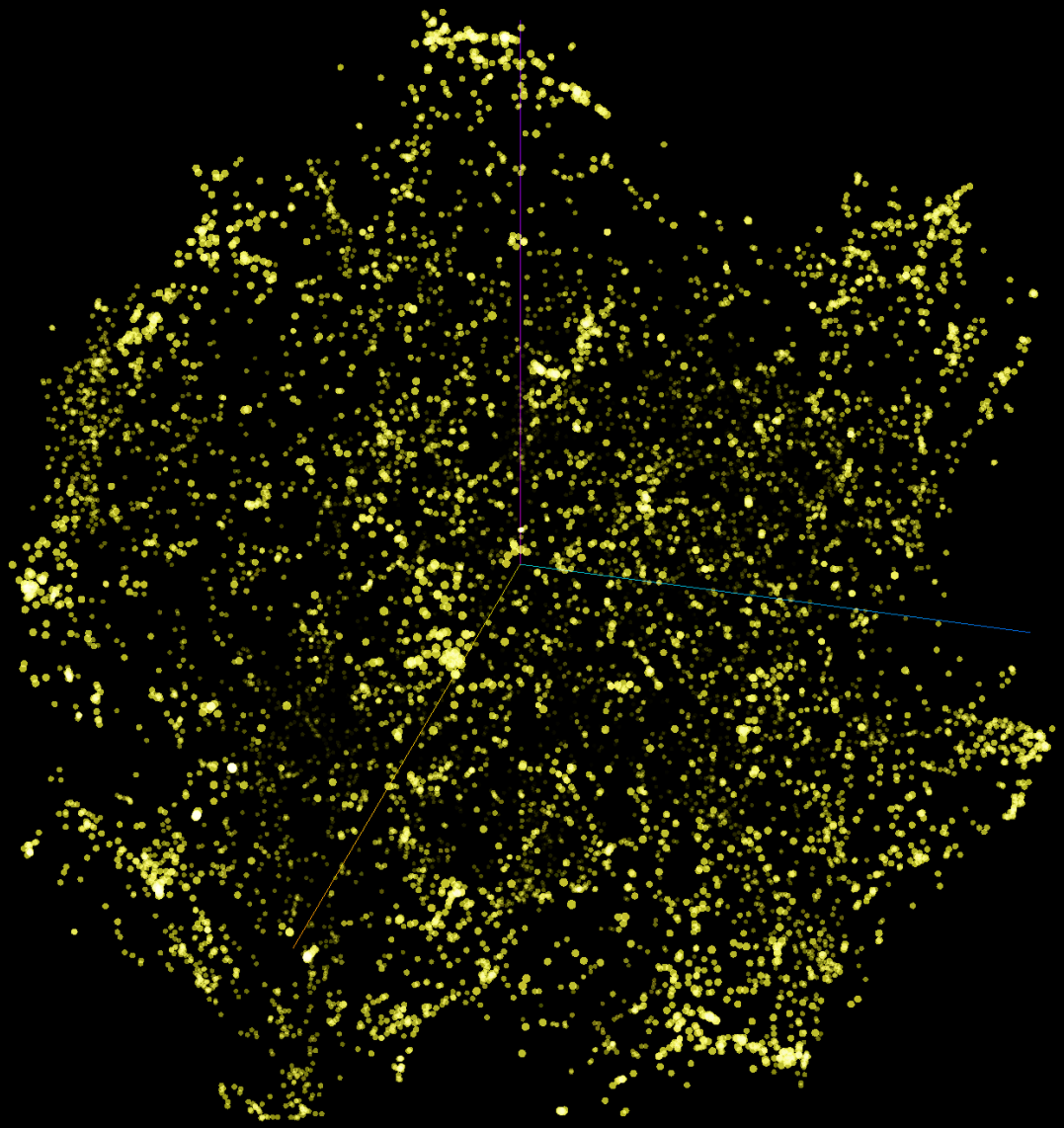
DIMENSIONALITY REDUCTION: T-SNE

- t-Distributed Stochastic Neighbor Embedding (t-SNE)
- pij measures similarity between xi and xj
- qij measures similarity between yi and yj
- Minimize Kullback Leibler divergence between p and q.
- Low dimensional map reflects similarities between high dimensional points.

DIMENSIONALITY REDUCTION: T-SNE

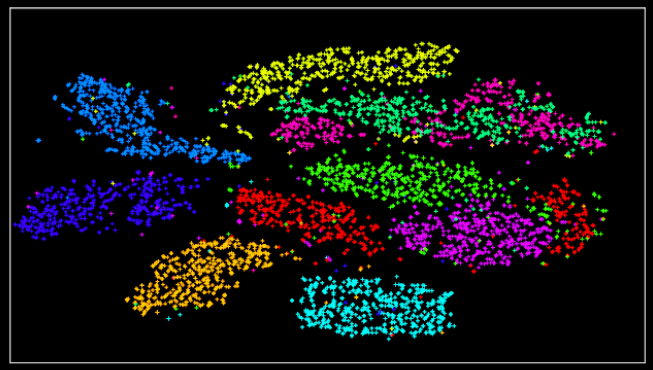


DIMENSIONALITY REDUCTION: T-SNE

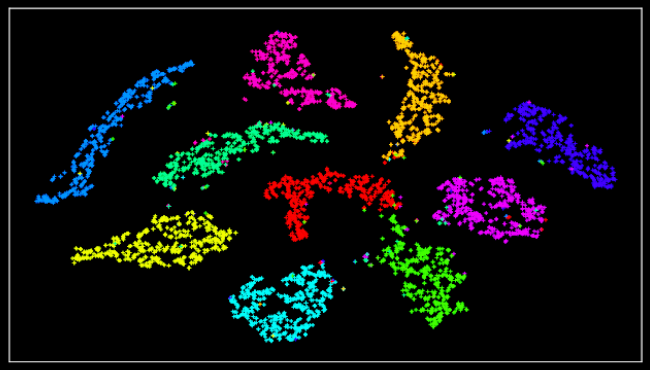


ROBUST CONTINUOUS CLUSTERING, SHAH & KULTON, ENAS 2017

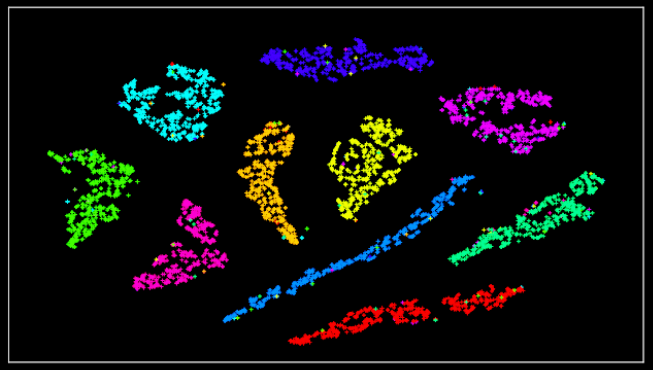
- Given $X = \left[x_1, x_2, \ldots, x_n \right] \quad x_i \in R^d$
- Generate KNN graph
- Minimize $C(U) = \frac{1}{2} \sum_{i=1}^{n} \left\| x_i - u_i \right\|_2^2 + \frac{\lambda}{2} \sum_{(p,q) \in \varepsilon} w_{p,q} \rho \left(\left\| u_p - u_q \right\|_2 \right)$
- t-SNE visualization



Initialization



Iteration 20



Final