Machine Learning For Design

Lecture 8/bis - Designing And Develop Machine Learning Models / Part 3

Alessandro Bozzon
30/03/2022

mlfd-io@tudelft.nl
www.ml4design.com

Admin

Assignments

- Deadline 3rd assignment: **Friday, April 1st, end of the day**
  - We will do our best to give you feedback by Tuesday 5th

- Deadline for the *final portfolio of group work* (final assignment) is **Monday, April 11th, end of the Day**
  - Simple packaging of the 3 assignments
  - You are allowed to update the content of the original assignments, to improve them based on feedback (if needed)
    - If you do, add a new section at the end of each assignment “**Improvement of Original Assignment**”
  - Evaluation of *final portfolio of group work* after the exam

- Also, fill in a *peer evaluation of group work* Excel sheet: deadline is **Friday, April 15th, end of the Day**
  - Needed for final grade of individual work

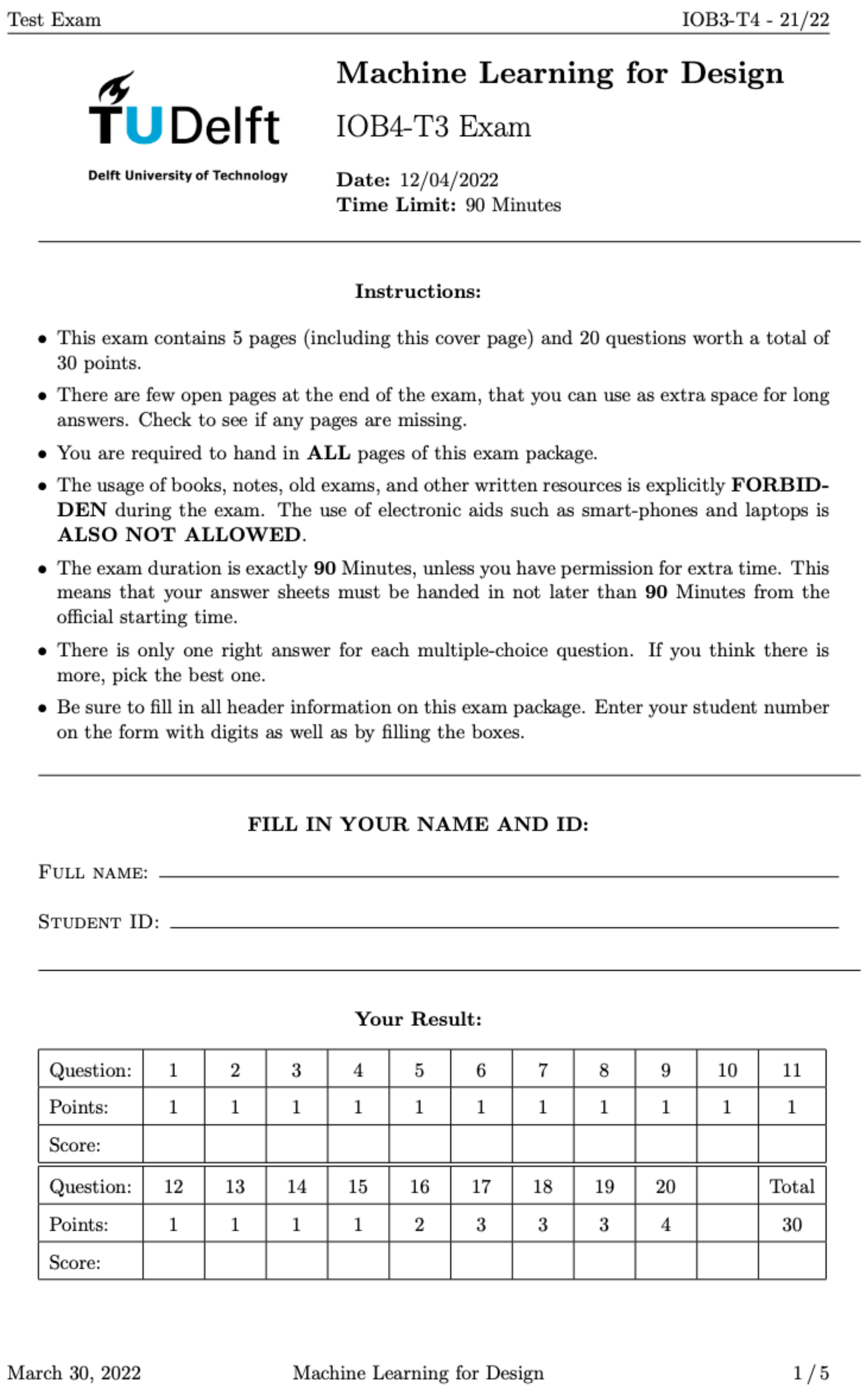
Exam

- Wednesday April 13th, 18.30 - Hall 2 Drebbelweg - 35
  - https://esviewer.tudelft.nl/space/47/
  - Check your timetable!
  - **Register!!!**

- 90 minutes, multiple choice and open answers
  - I will publish an example of exam tomorrow

- Content: everything discussed duruing lectures
  - All lectures and tutorial recording available on Brightspace
  - Additional reading material is useful, but not mandatory
  - Assignments are obviously addressing topics discussed during lectures

- Quizzes Week1 to Week 7 are on brightspace - useful to prepare the exam



Test Exam IOB3-T4 - 21/22

TU Delft
Delft University of Technology

Machine Learning for Design

IOB4-T3 Exam

Date: 12/04/2022
Time Limit: 90 Minutes

Instructions:

- This exam contains 5 pages (including this cover page) and 20 questions worth a total of 30 points.
- There are few open pages at the end of the exam, that you can use as extra space for long answers. Check to see if any pages are missing.
- You are required to hand in **ALL** pages of this exam package.
- The usage of books, notes, old exams, and other written resources is explicitly **FORBIDDEN** during the exam. The use of electronic aids such as smart-phones and laptops is **ALSO NOT ALLOWED**.
- The exam duration is exactly **90** Minutes, unless you have permission for extra time. This means that your answer sheets must be handed in not later than **90** Minutes from the official starting time.
- There is only one right answer for each multiple-choice question. If you think there is more, pick the best one.
- Be sure to fill in all header information on this exam package. Enter your student number on the form with digits as well as by filling the boxes.

FILL IN YOUR NAME AND ID:

Full name: ______________________

Student ID: ______________________

Your Result:

Question:	1	2	3	4	5	6	7	8	9	10	11
Points:	1	1	1	1	1	1	1	1	1	1	1
Score:											

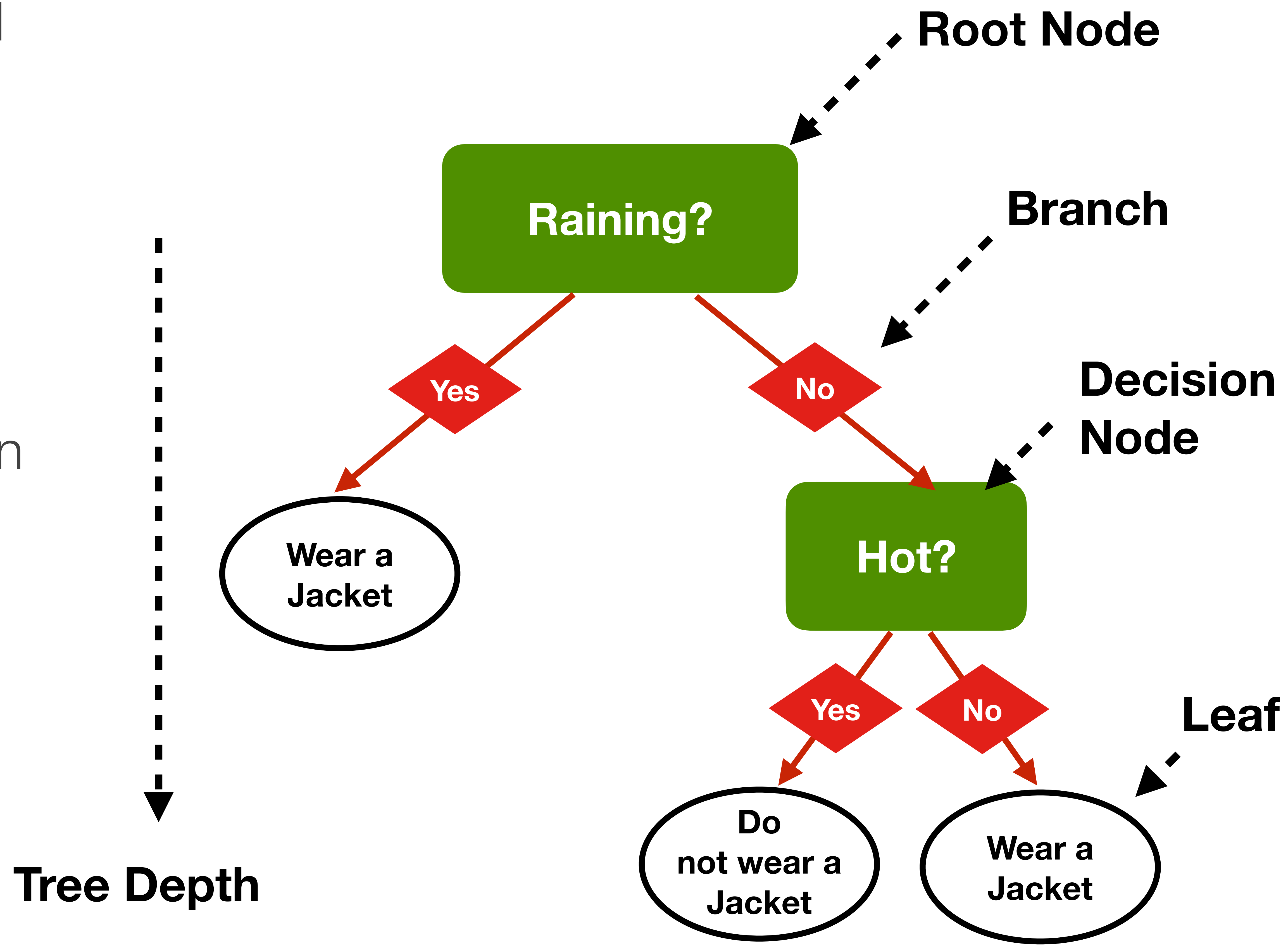
Question:	12	13	14	15	16	17	18	19	20		Total
Points:	1	1	1	1	2	3	3	3	4		30
Score:											

March 30, 2022 Machine Learning for Design 1 / 5

Decision Trees

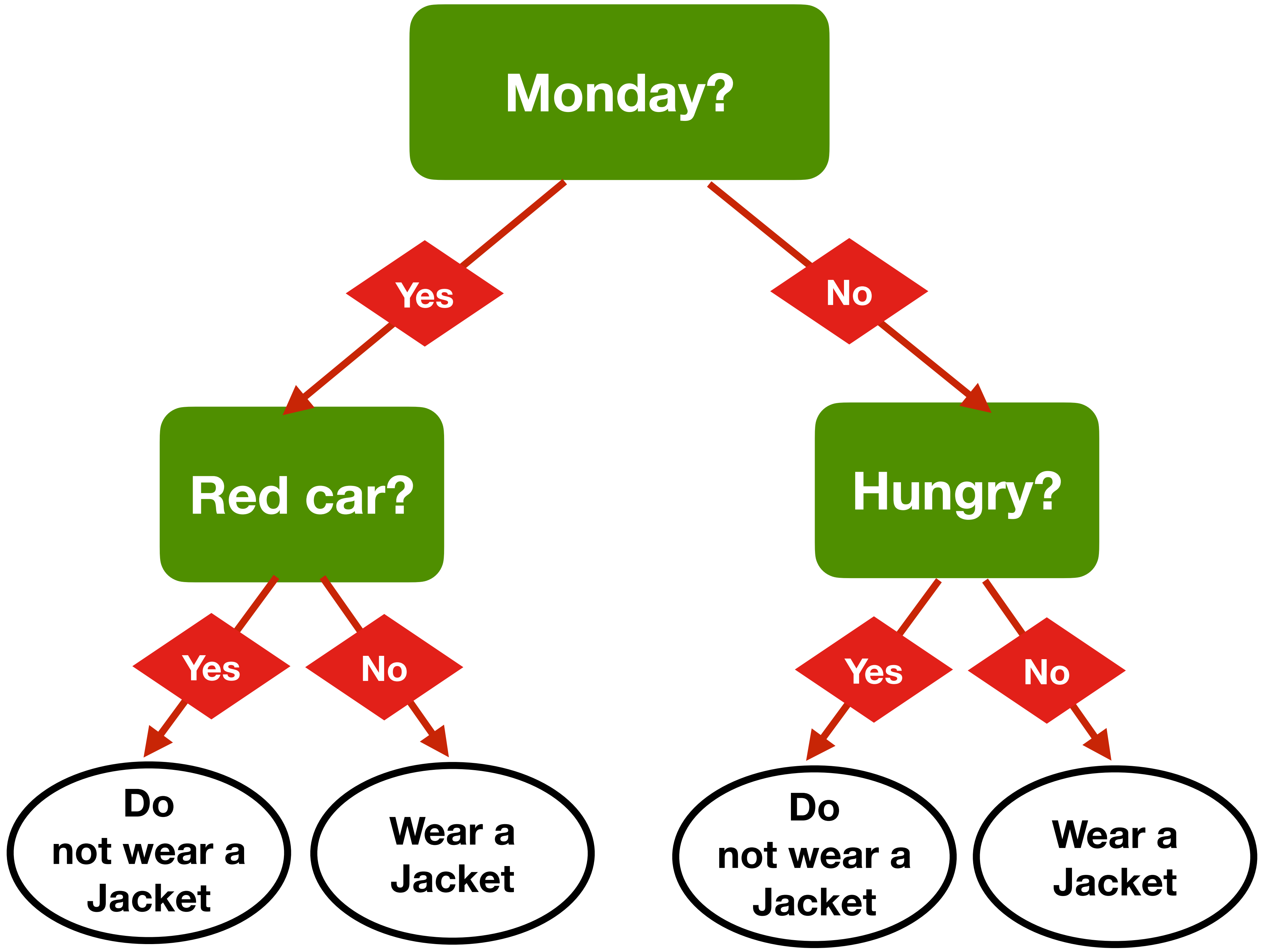
Decision Trees

- Machine learning models used both for **classification** and **regression**
- Trained with labelled data (**supervised learning**)
  - classes —> classification
  - values —> regression
- A very simple model that resembles human reasoning when making predictions:
  - Answering a lot of yes/no questions based on feature values
- Problems:
  - Which questions to answer?
  - How many questions? (Tree depth)
  - In which order?



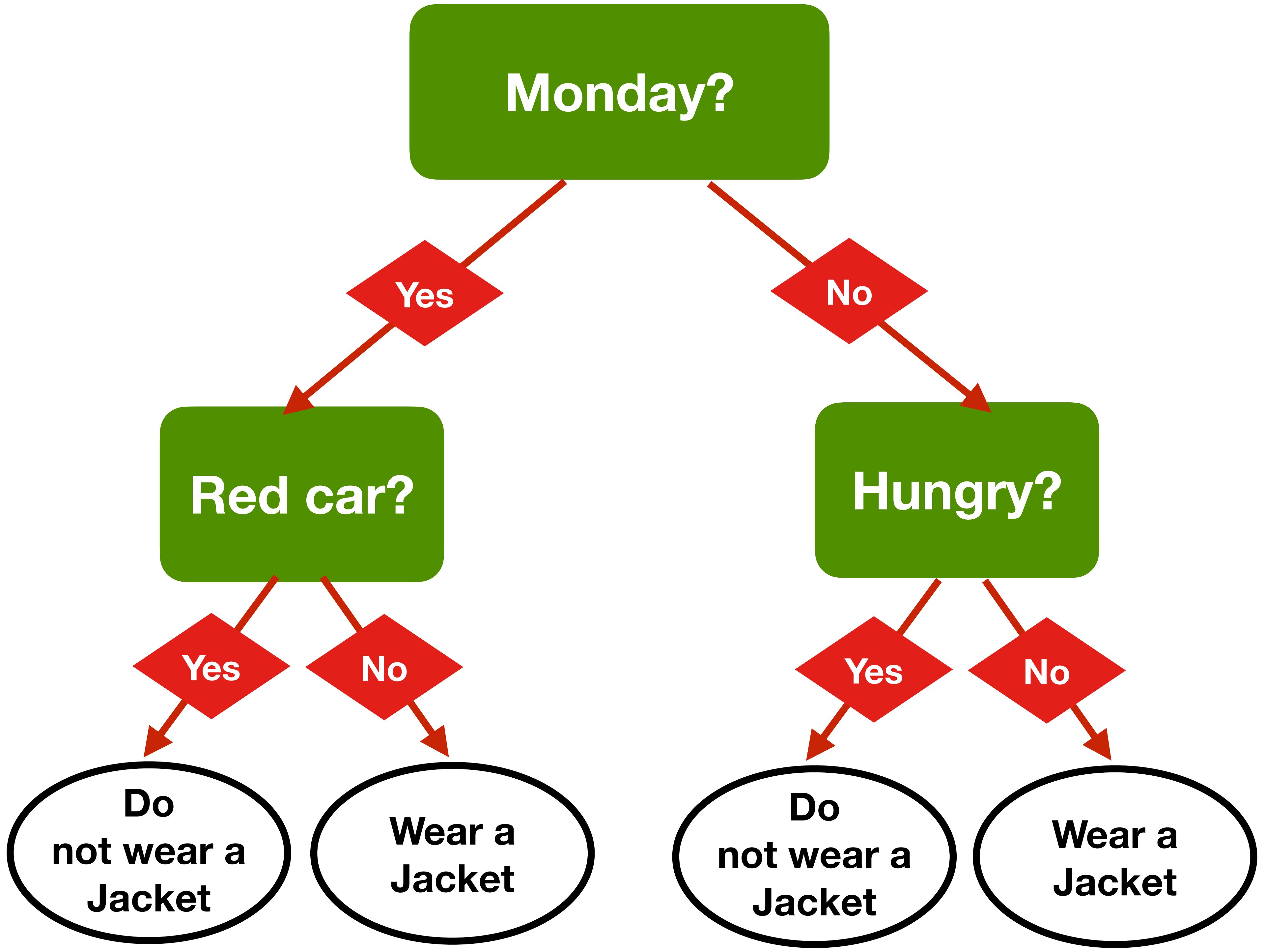
Same Problem, Multiple Trees

- Feature space
  - Am I hungry?
  - Is there a red car outside?
  - Is it Monday?
  - Is it raining?
  - Is it cold outside?



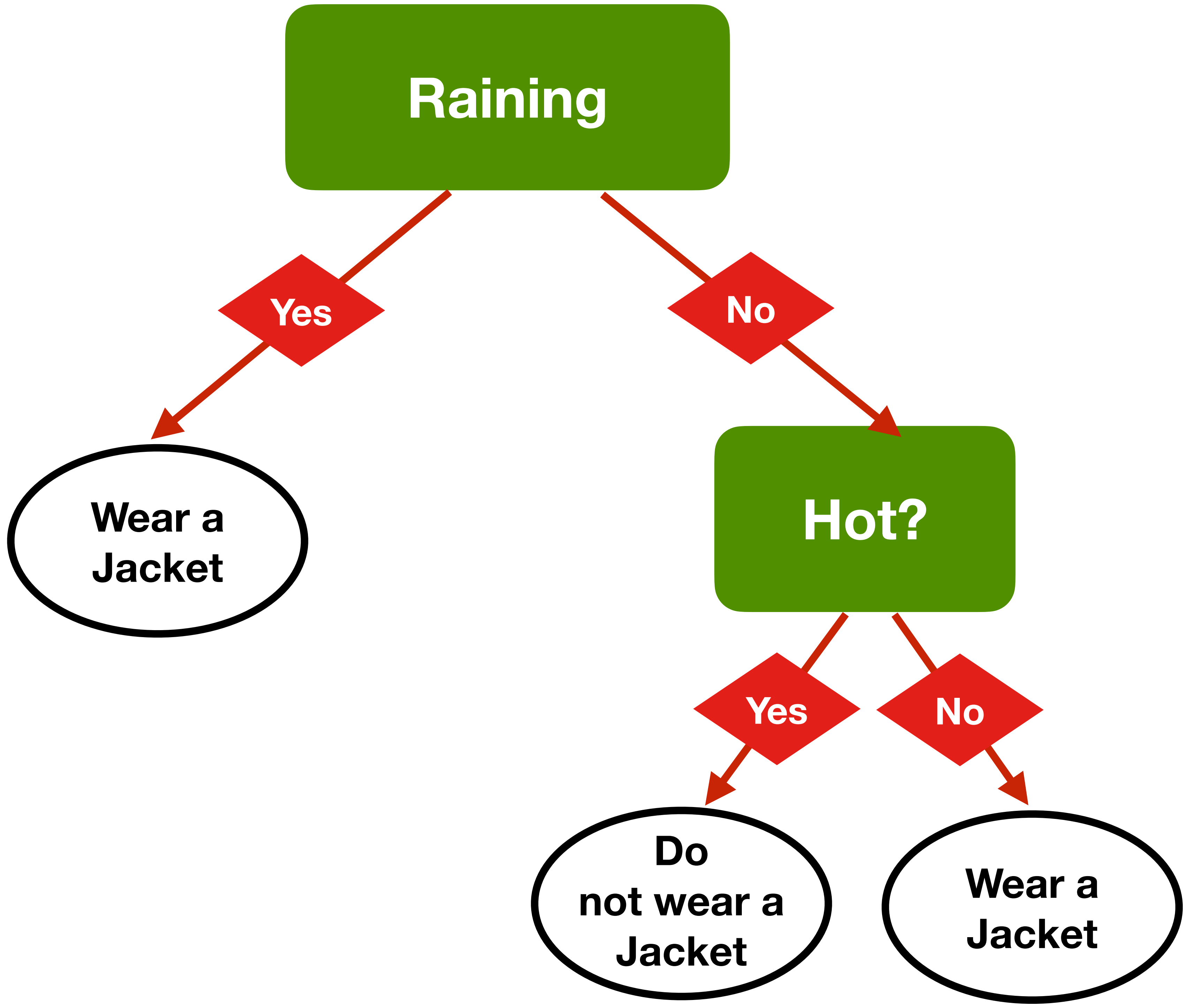
Same Problem, Multiple Trees

- Feature space
  - ~~Am I hungry?~~
  - ~~Is there a red car outside?~~
  - ~~Is it Monday?~~
  - Is it raining?
  - Is it cold outside?

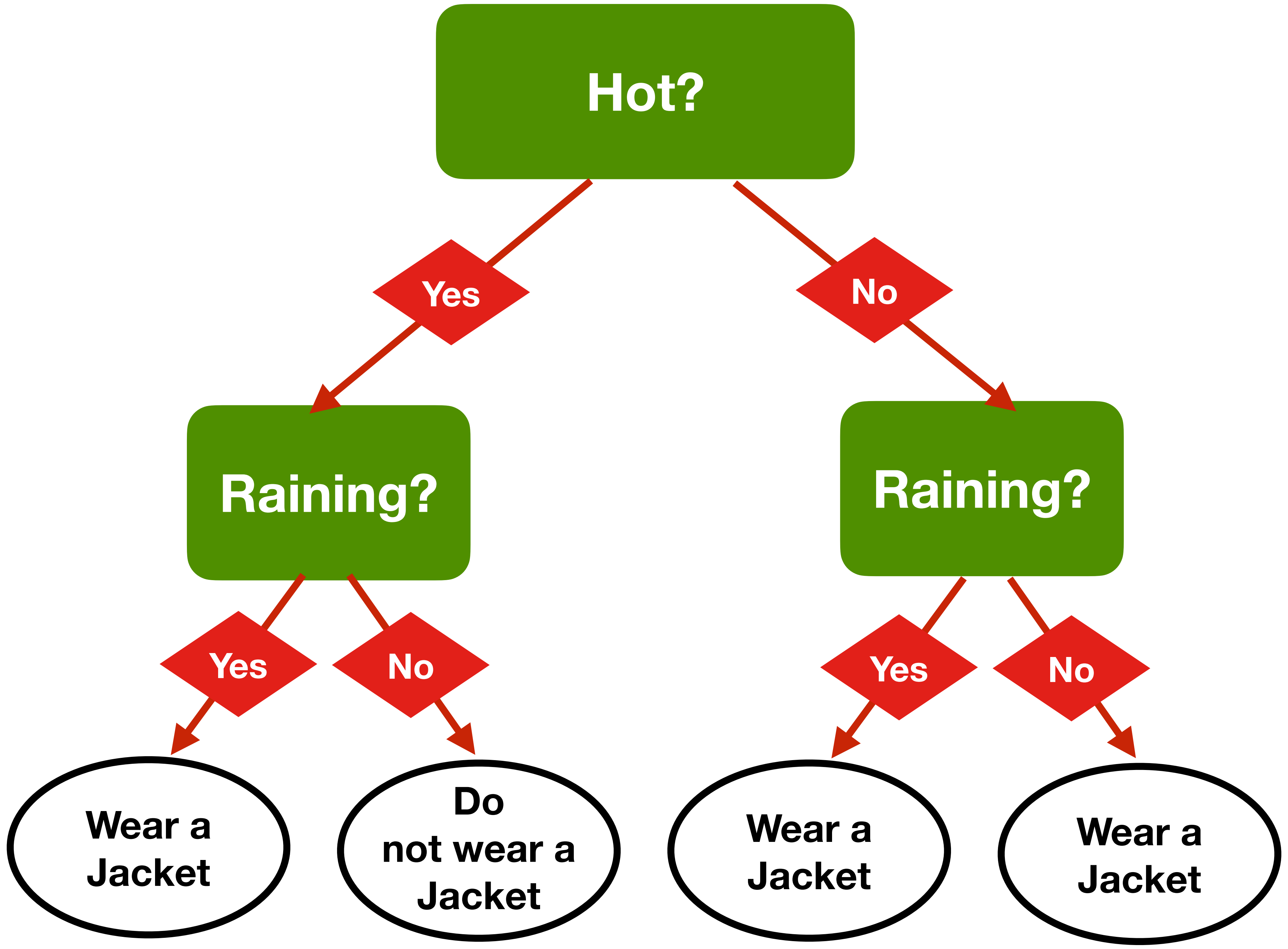


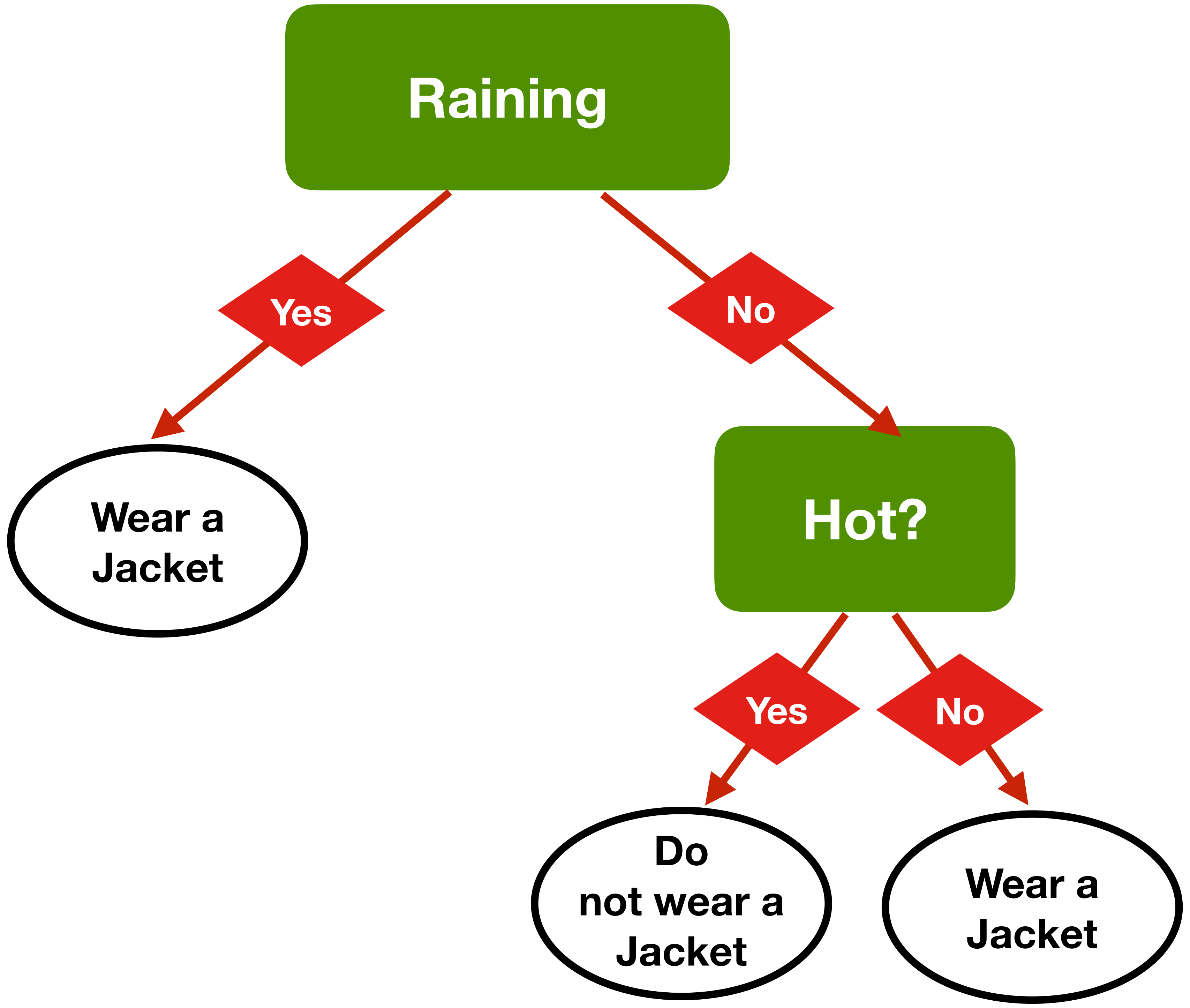
Same Problem, Multiple Trees

- Feature space
  - ~~Am I hungry?~~
  - ~~Is there a red car outside?~~
  - ~~Is it Monday?~~
  - Is it raining?
  - Is it cold outside?

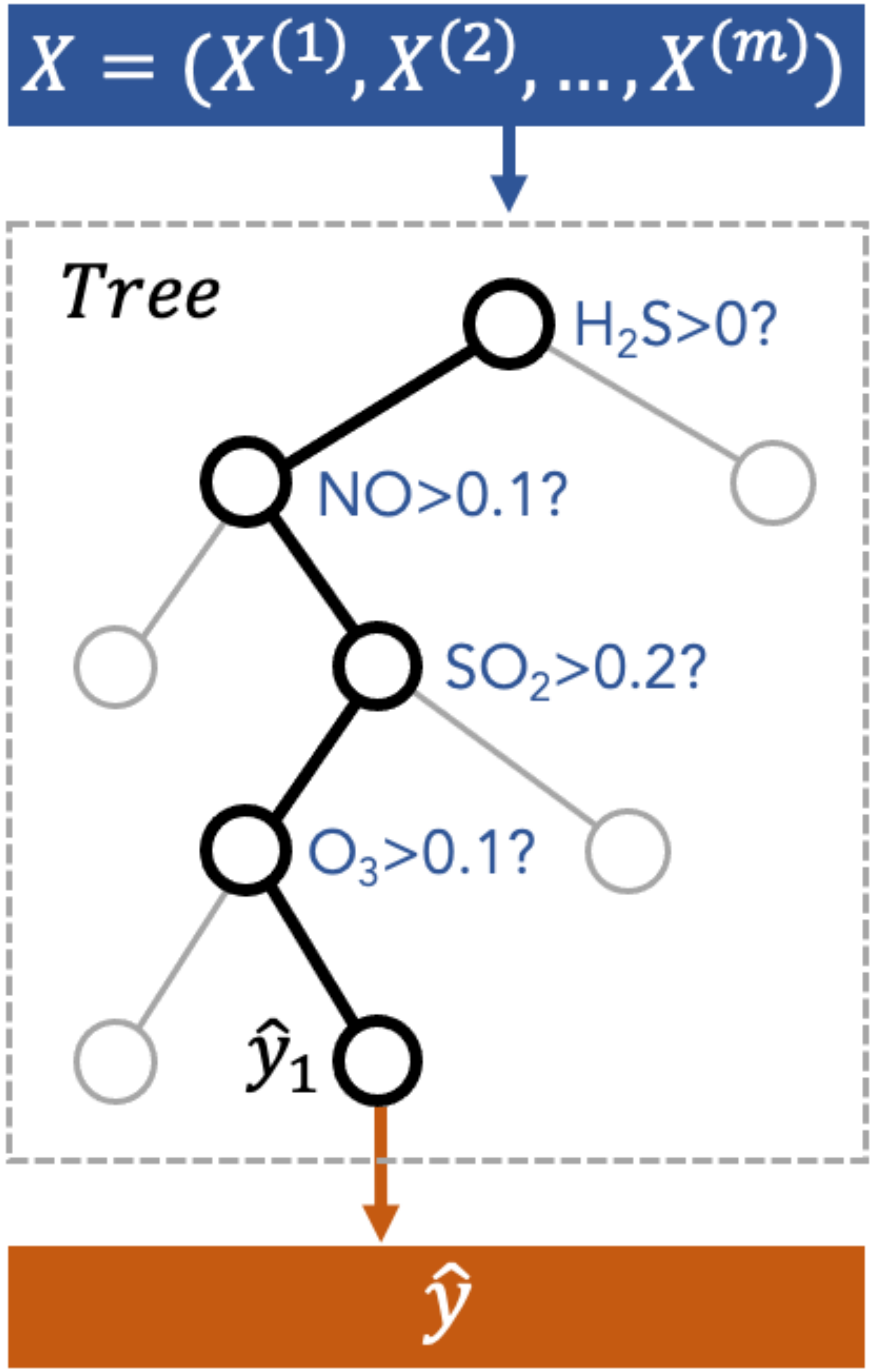


Same decision, different trees





Tutorial 3



PM, SO_2, CO, NO, NO_2, O_3, H_2S, and wind information

Prediction of bad smell (yes/no)

How to decide the best question to ask?

- **3 metrics**
  - **Accuracy**
    - Which question helps me be **correct** more often?
  - **Gini Impurity Index**
    - A measure of *diversity* in a dataset —> diversity of classes in a given leaf node
      - *index = 0* means that all the items in a leaf node have the same class
    - Which question helps me obtain the lowest average **Gini impurity Index**?
  - **Entropy**
    - Another measure of *diversity* linked to information theory
    - Which question helps me obtain the lowest average **entropy**?

Building the tree (pseudo-code)

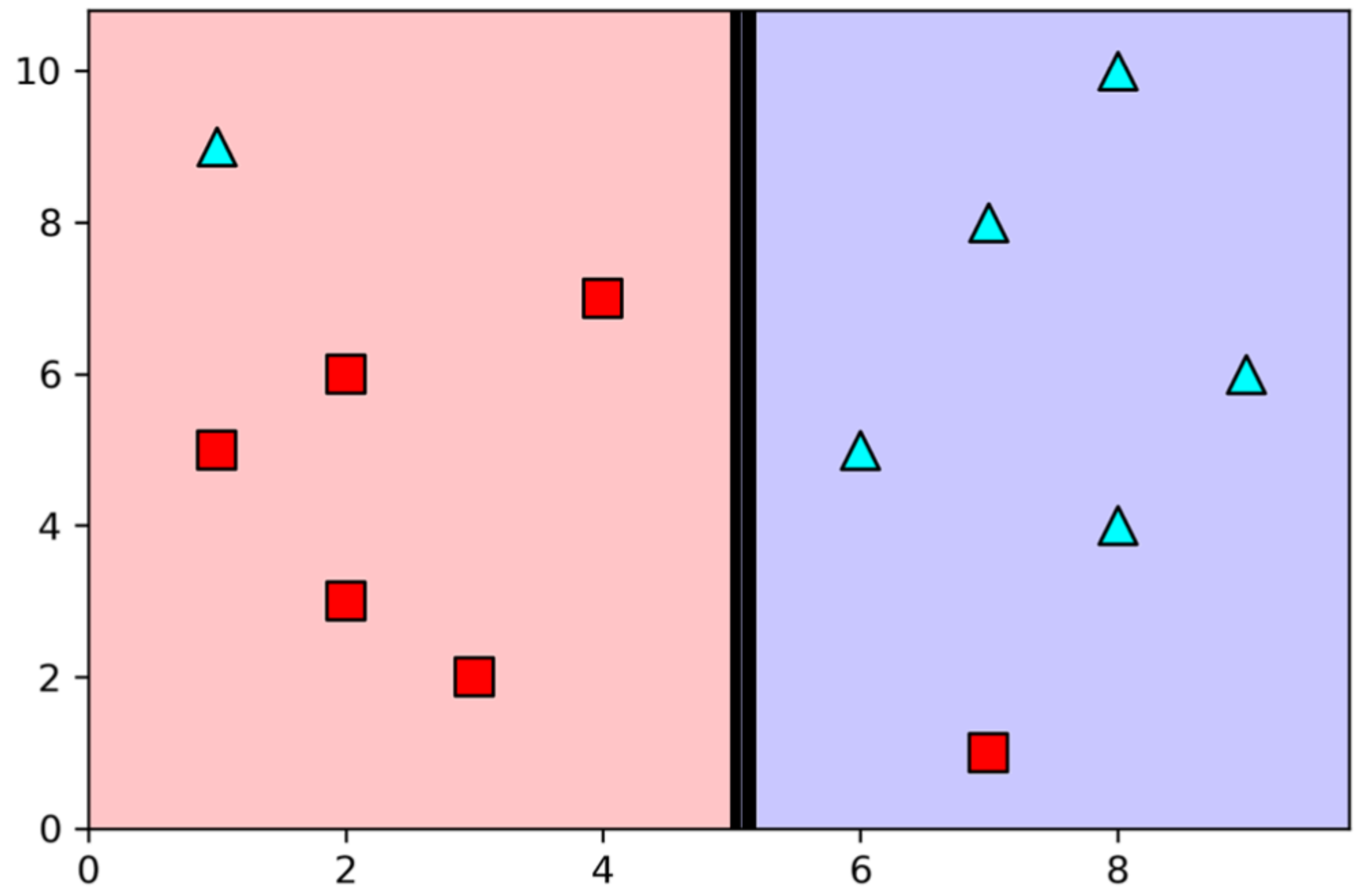
Hyperparameter: tree depth

Stopping condition

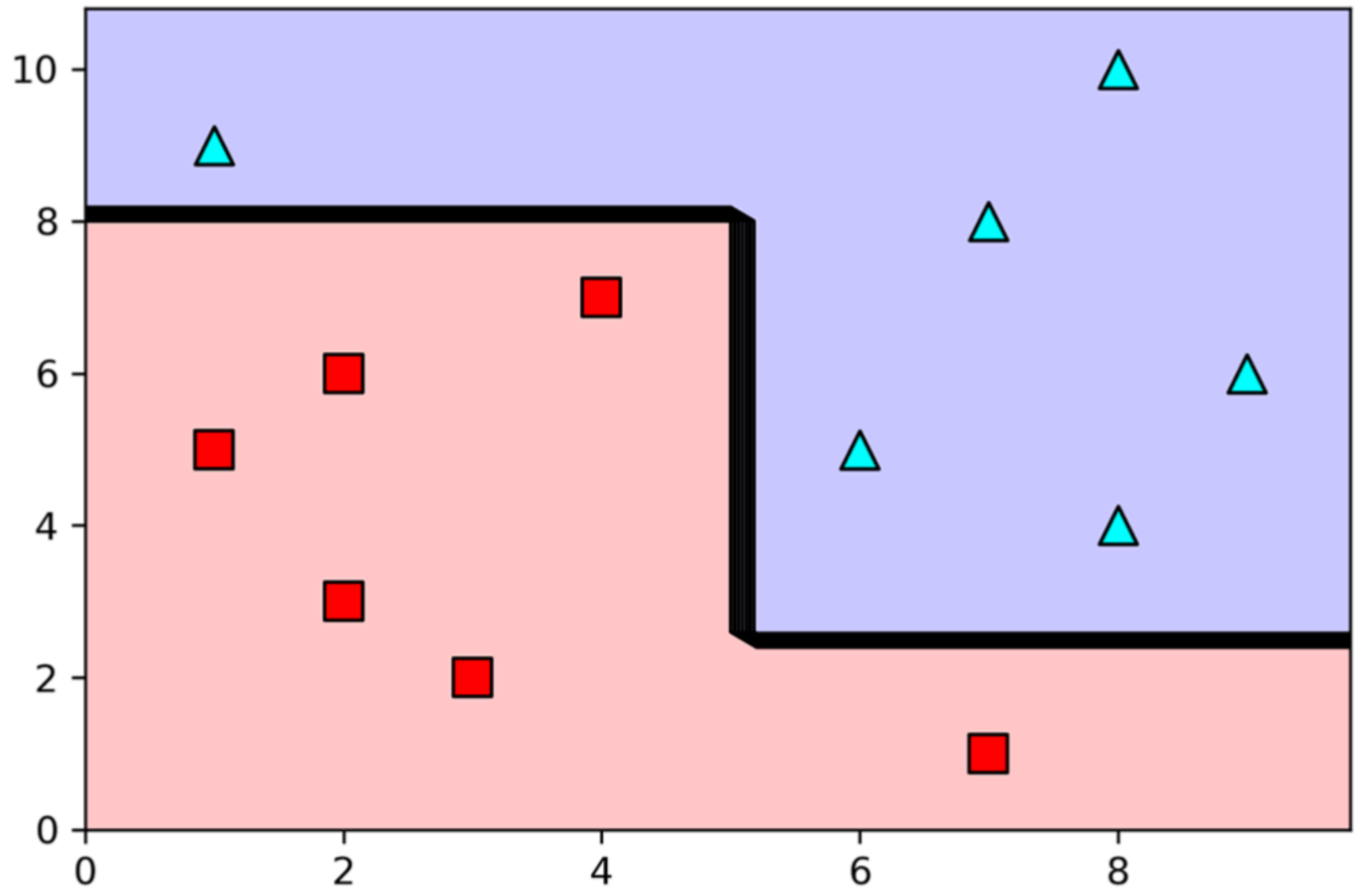
- **Add a root node, and associate it with the entire dataset**
  - This node has level 0. Call it a leaf node

- **Repeat until the stopping conditions are met at every leaf node**
  - Pick one of the leaf nodes at the highest level
  - Go through all the features, and select the one that splits the samples corresponding to that node in an optimal way, according to the selected metric.
    - Associate that feature to the node
  - This feature splits the dataset into two branches
    - Create two new leaf nodes, one for each branch
    - Associate the corresponding samples to each of the nodes
  - If the stopping conditions allow a split, turn the node into a decision node, and add two new leaf nodes underneath it
    - If the level of the node is i, the two new leaf nodes are at level i + 1
  - If the stopping conditions don't allow a split, the node becomes a leaf node
    - Associate the most common label among its samples
    - That label is the prediction at the leaf

A geometrical perspective



- Step 1 - Select the first question
- **X >= 5**
  - Best possible prediction accuracy with one feature



- Step 2 - Iterate
- **x < 5 & y< 8; x >=5 & y>=2**
  - Perfect split of the feature space

Decision Trees: Pros and Cons

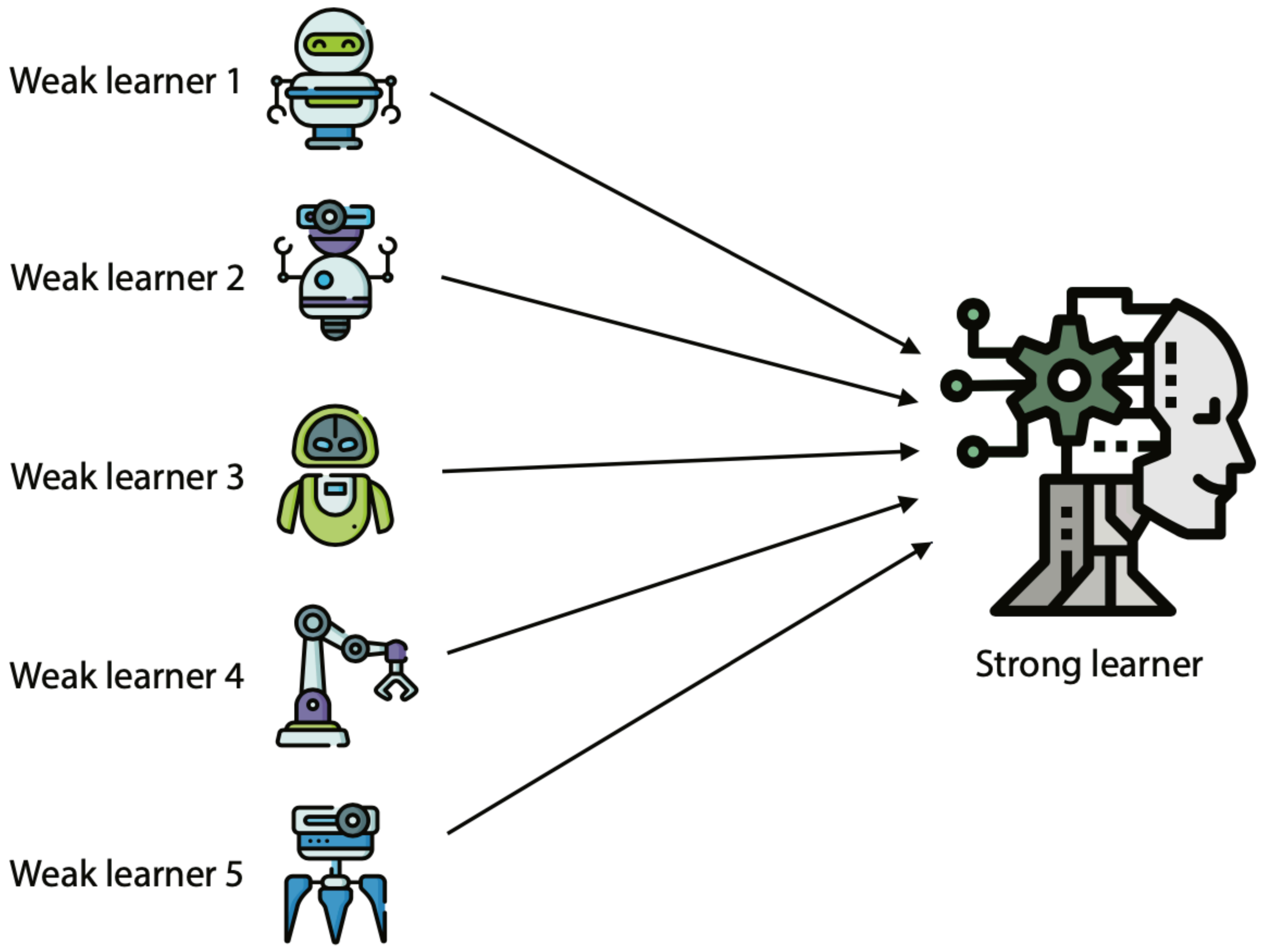
- **PROs**
  - Simple to understand and to interpret. Trees can be visualised
  - Requires little data preparation. Other techniques often require data normalisation, dummy variables need to be created and blank values to be removed
  - Able to handle both numerical and categorical data

- **Cons**
  - Possible to create over-complex trees that do not generalise the data well
    - overfitting
  - Unstable —> small variations in the data might result in a completely different tree being generated
  - Biased trees if some classes dominate

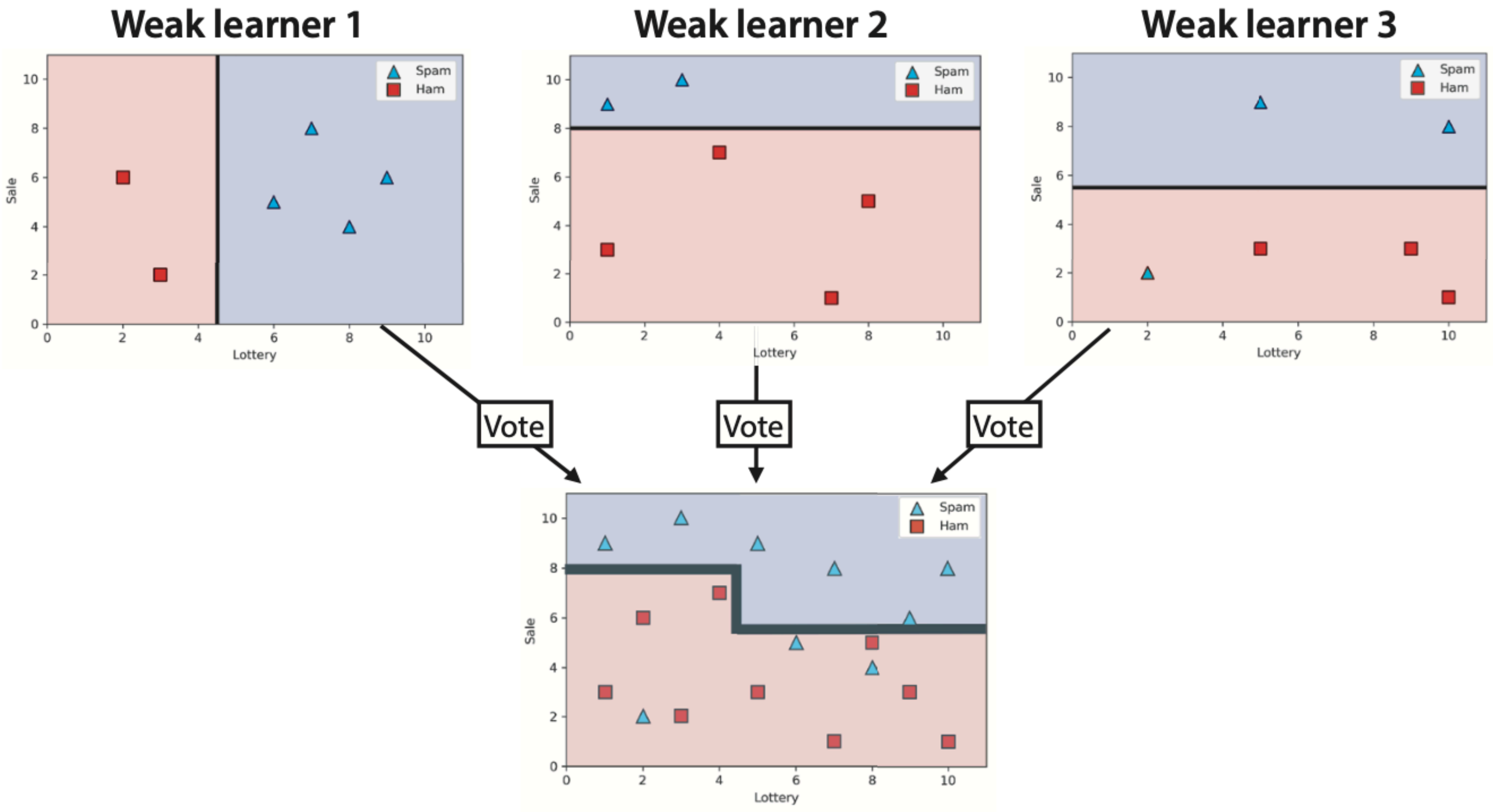
https://scikit-learn.org/stable/modules/tree.html

Ensemble learning: Random Forest

- Idea: combine several “weak” learners to build a strong learner
  - Build random training sets from the dataset
  - Train a different model on each of the sets
    - weak learners
  - Combination the weak models by voting (if it is a classification model) or averaging the predictions (if it is a regression model)
    - For any input, each of the weak learners predicts a value
    - The most common output (or the average) is the output of the strong learner



- Random Forest
  - Weak learners are **decision trees**



Weak learner 1

Weak learner 2

Weak learner 3

Vote

Vote

Vote

Strong learner (random forest)

Clustering

What is clustering?

- Grouping items that “belong together” (i.e. have similar features)
- **Unsupervised learning**: we only use data features, not the labels
- We can detect patterns
  - Group emails or search results
  - Customer shopping patterns
  - Regions of images
- Useful when don’t know what you’re looking for
  - But: can get gibberish
- If the goal is classification, we can later ask a human to label each group (cluster)

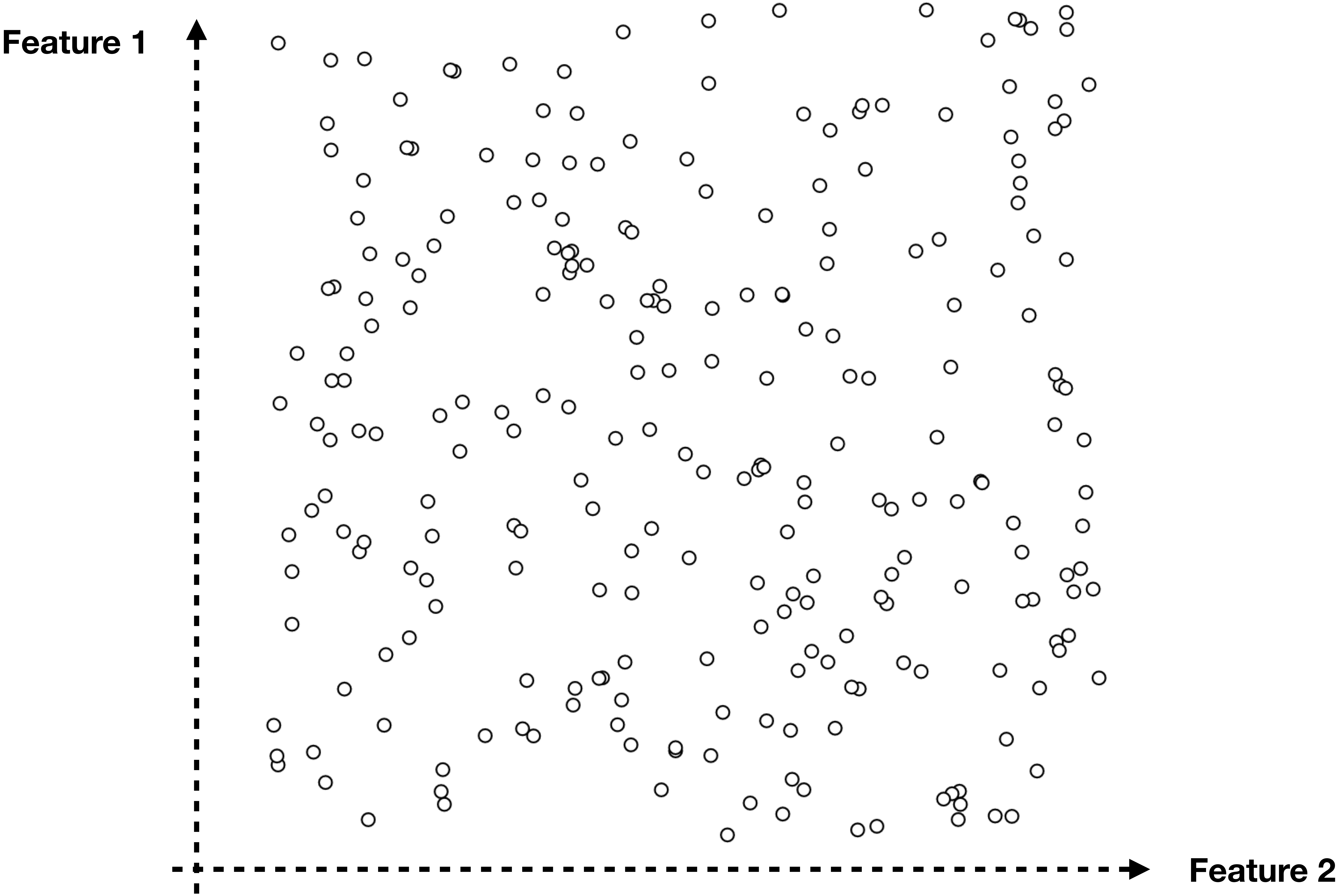
Why do we cluster?

- Summarizing data
    - Look at large amounts of data
    - Represent a large continuous vector with the cluster number
- Counting
    - Computing feature histograms
- Prediction
    - Images in the same cluster may have the same labels
- Segmentation
    - Separate the image into different regions

K-Means

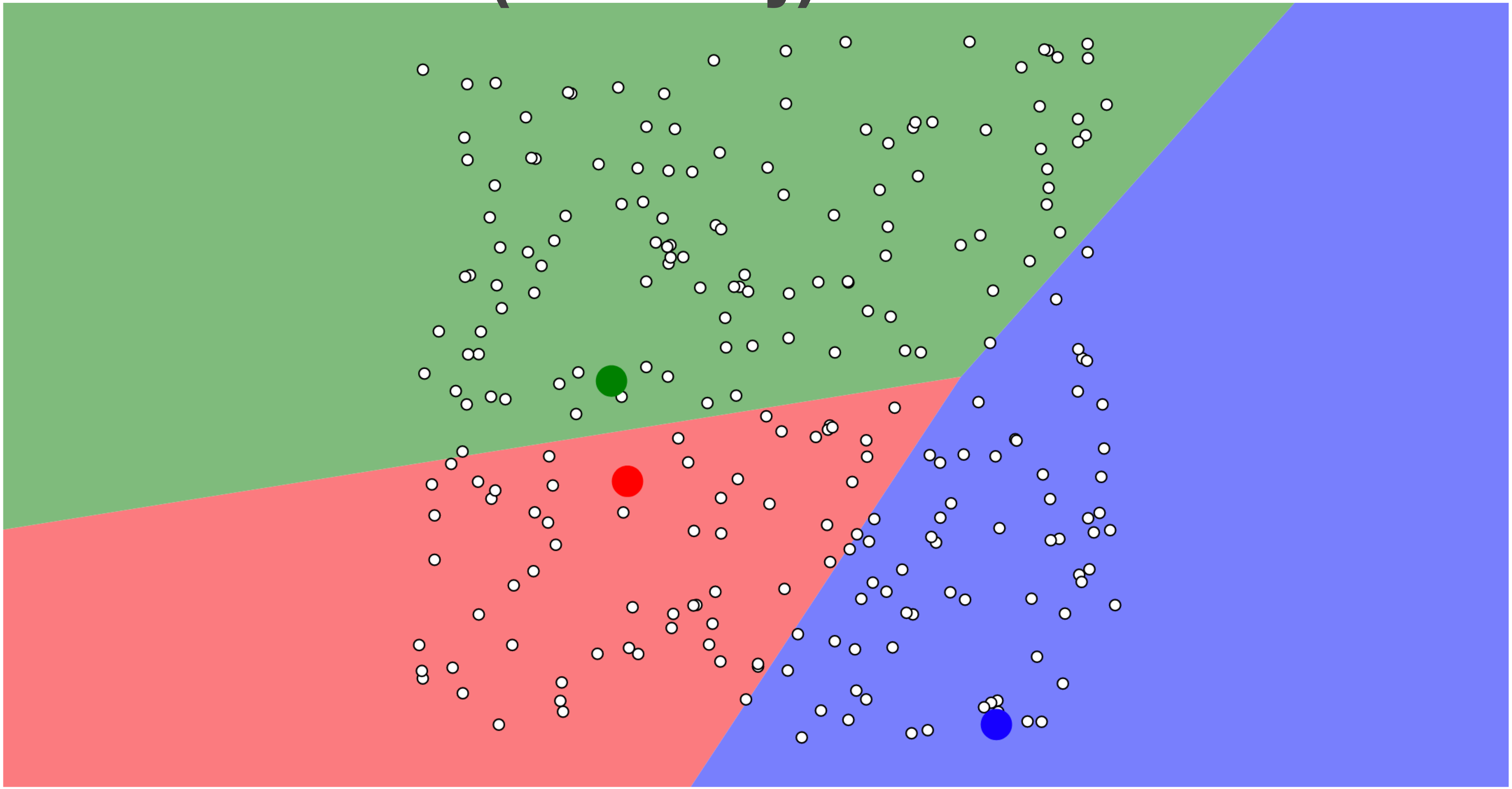
- An iterative clustering algorithm
  - **Initialize**: Pick K random points as cluster centres
  - **Alternate**:
    - Assign data points to the closest cluster centre
    - Change the cluster centre to the average of its assigned points
  - **Stop** when no points' assignments change

Data items distribution

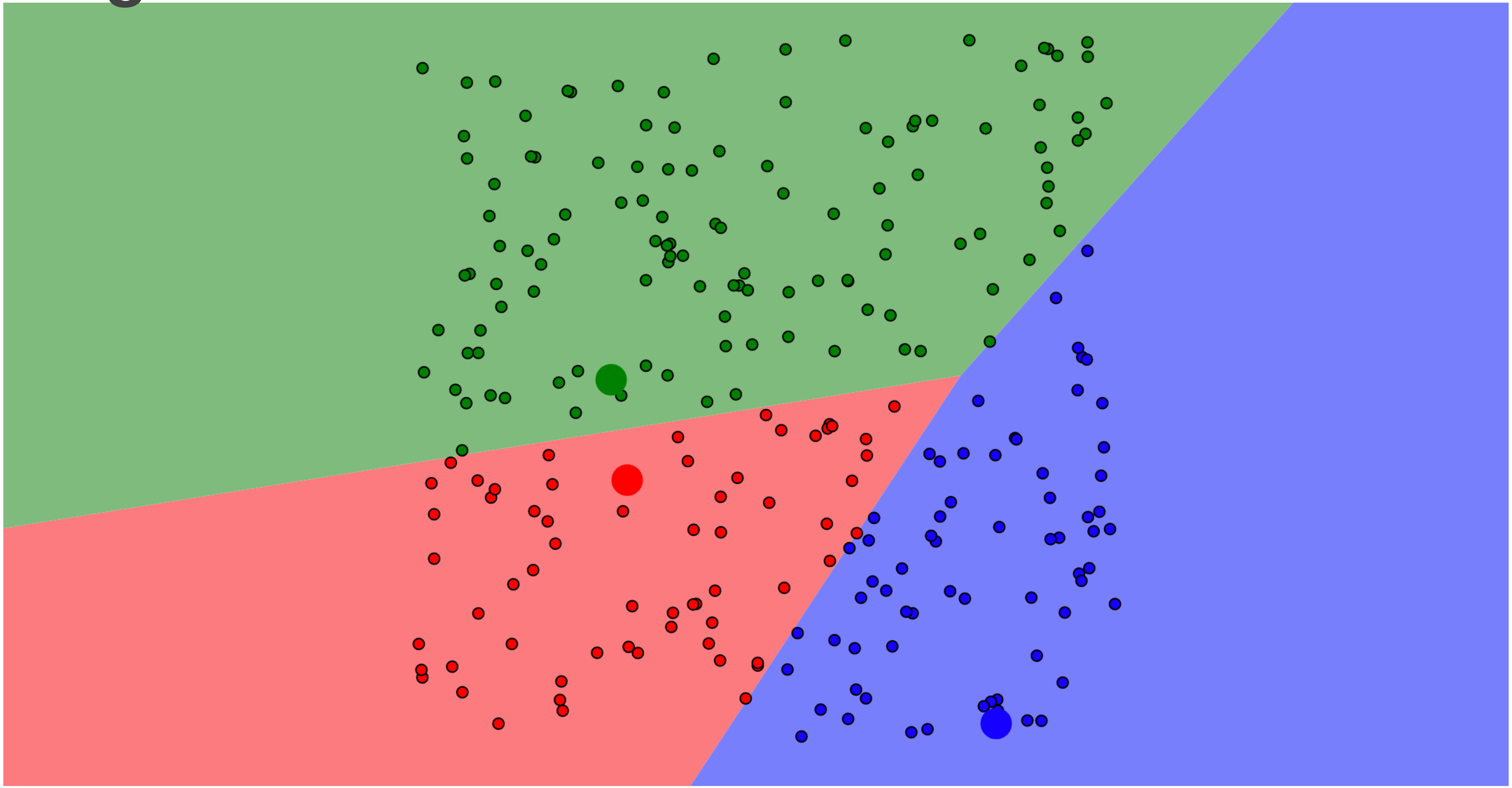


https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

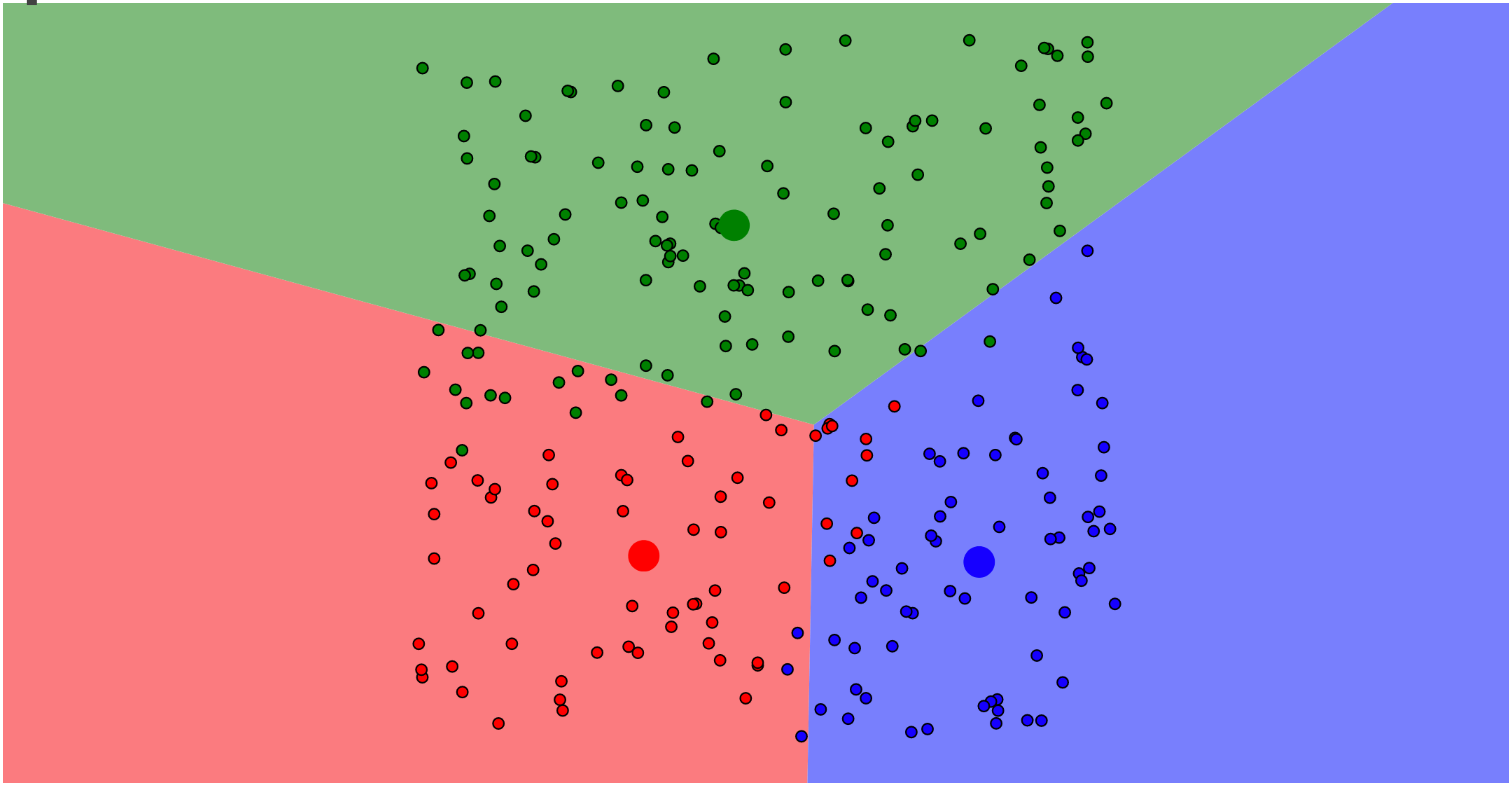
Add 3 Centroids (randomly)



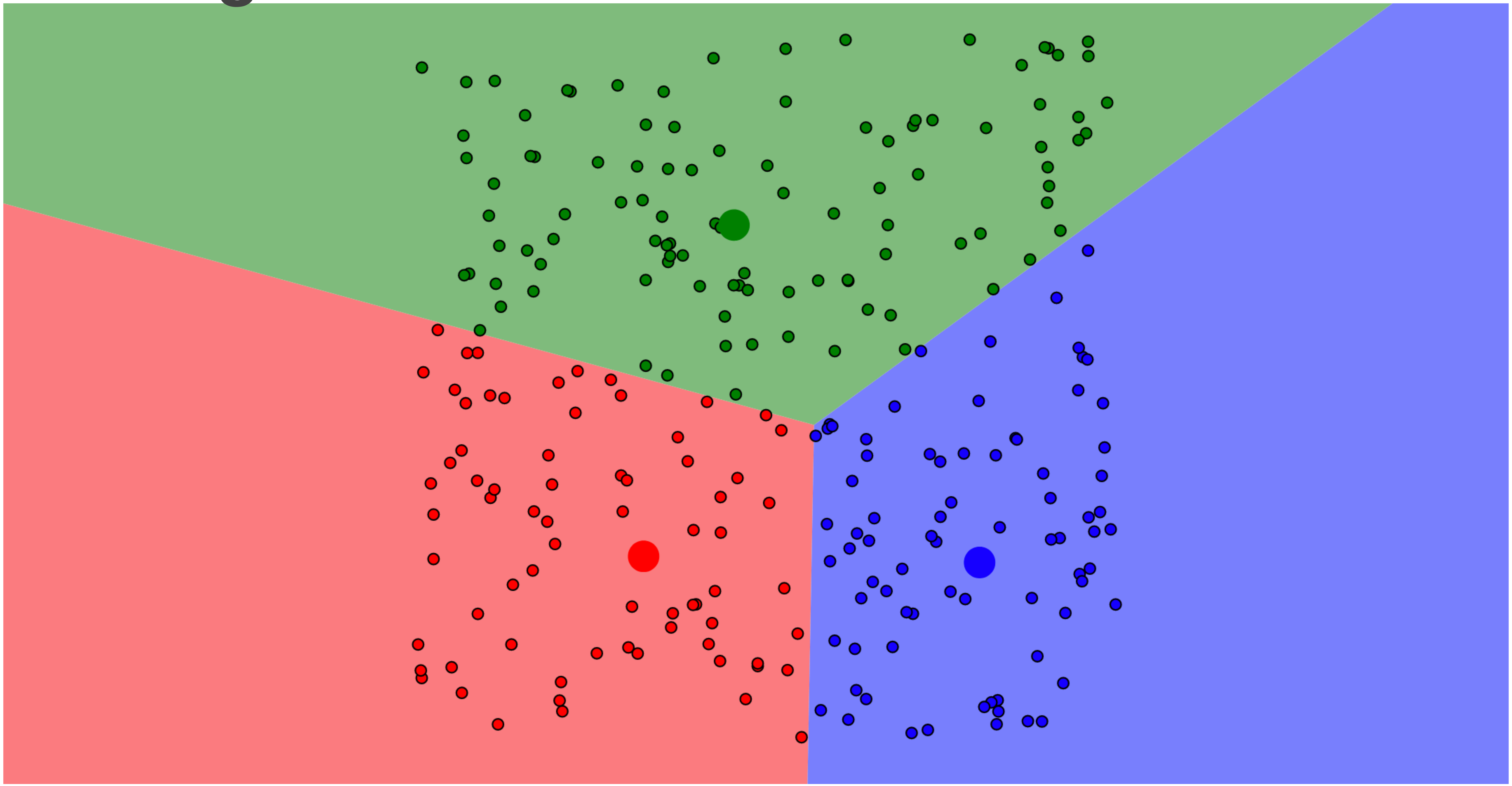
Assign Data Points



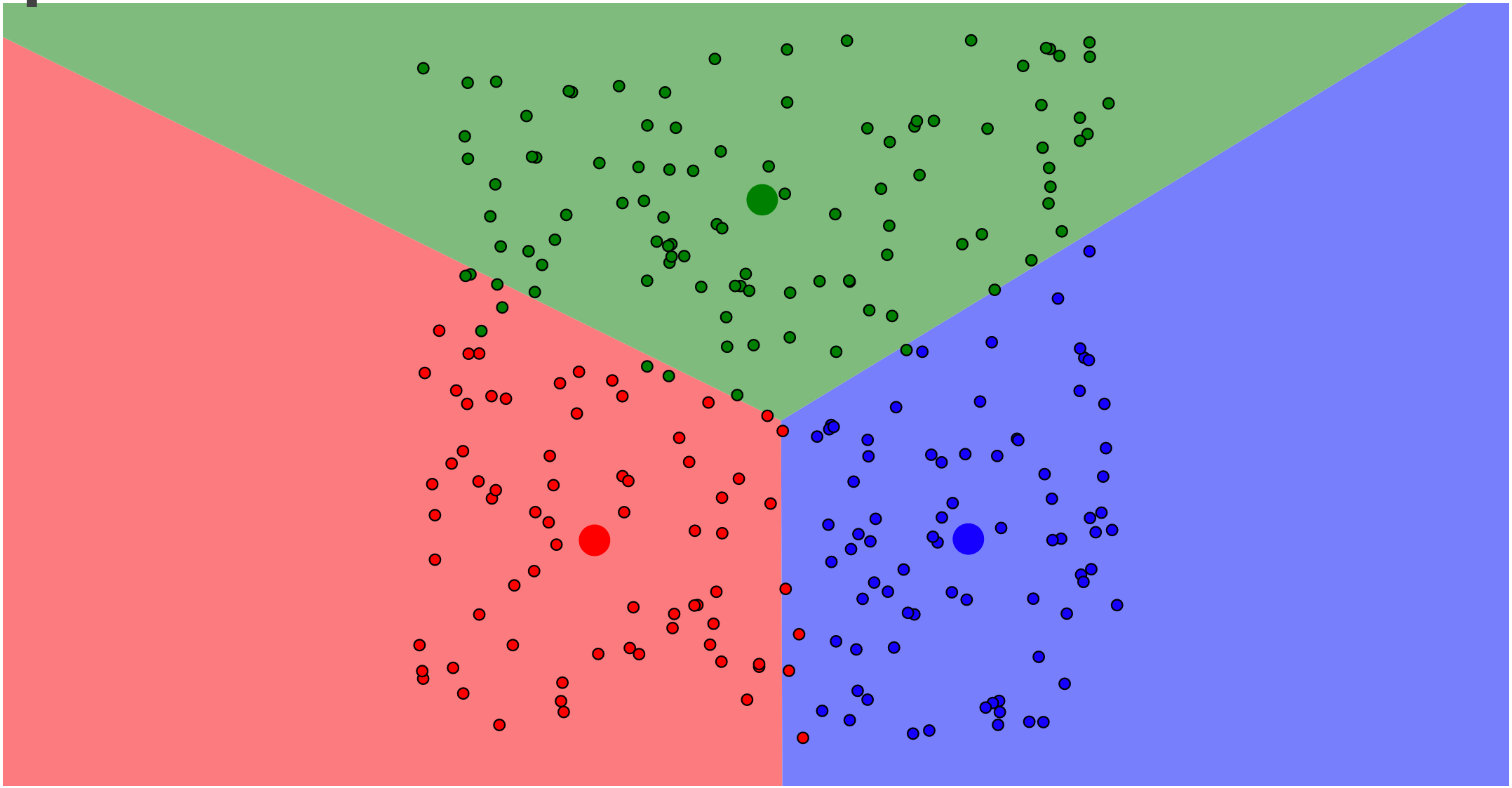
Update Centroids



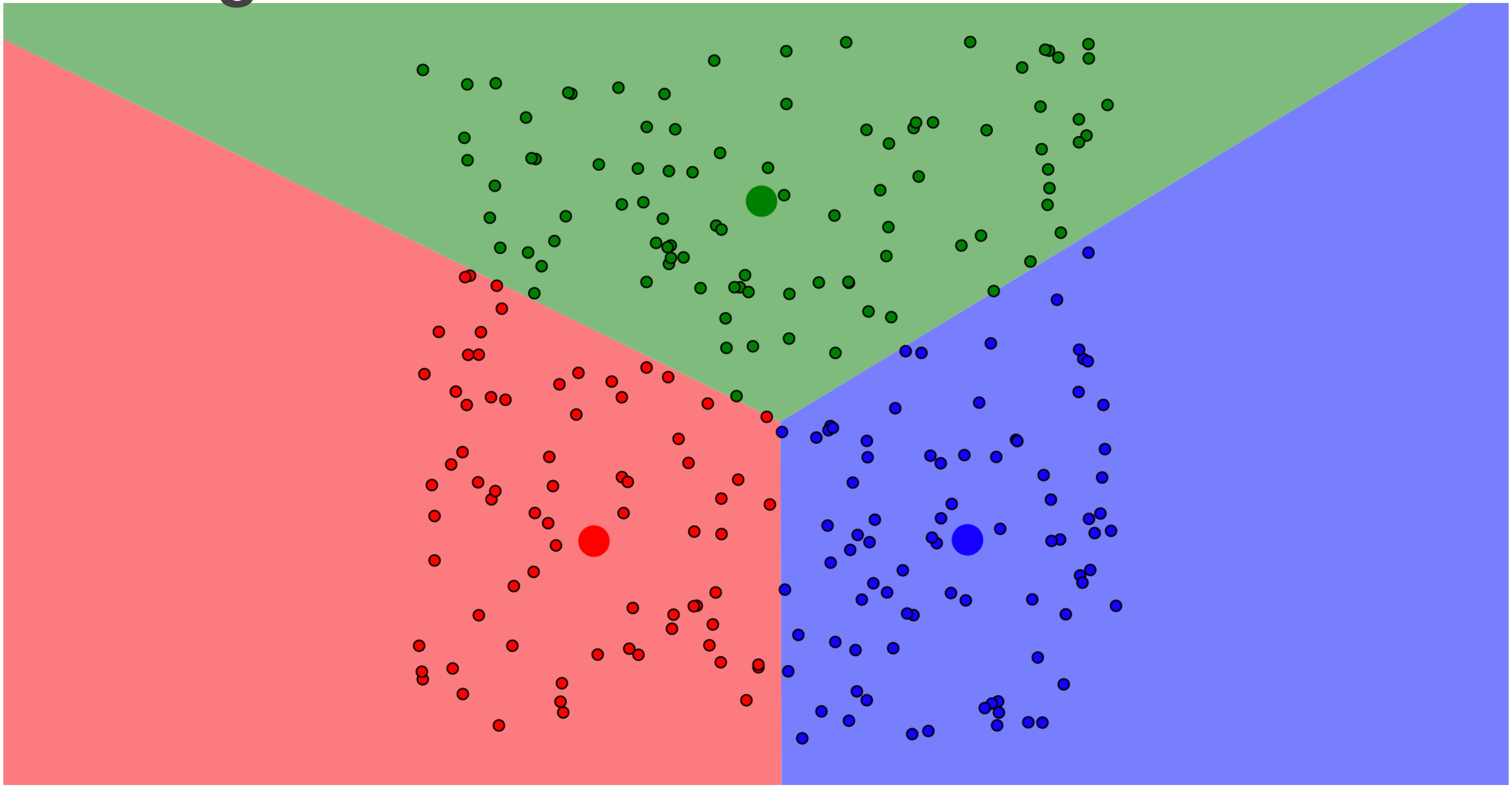
Re-Assign Data Points



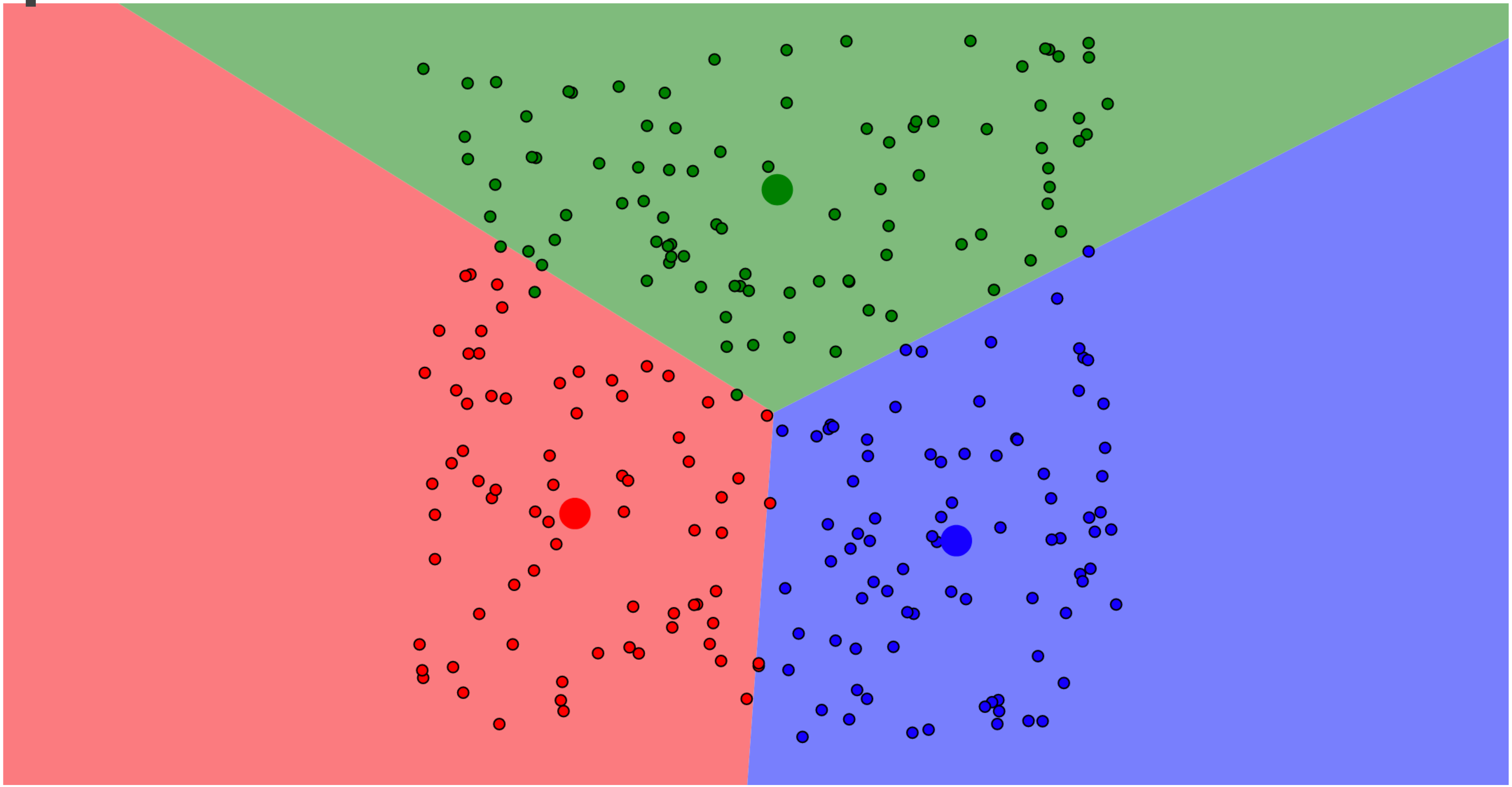
Update Centroids



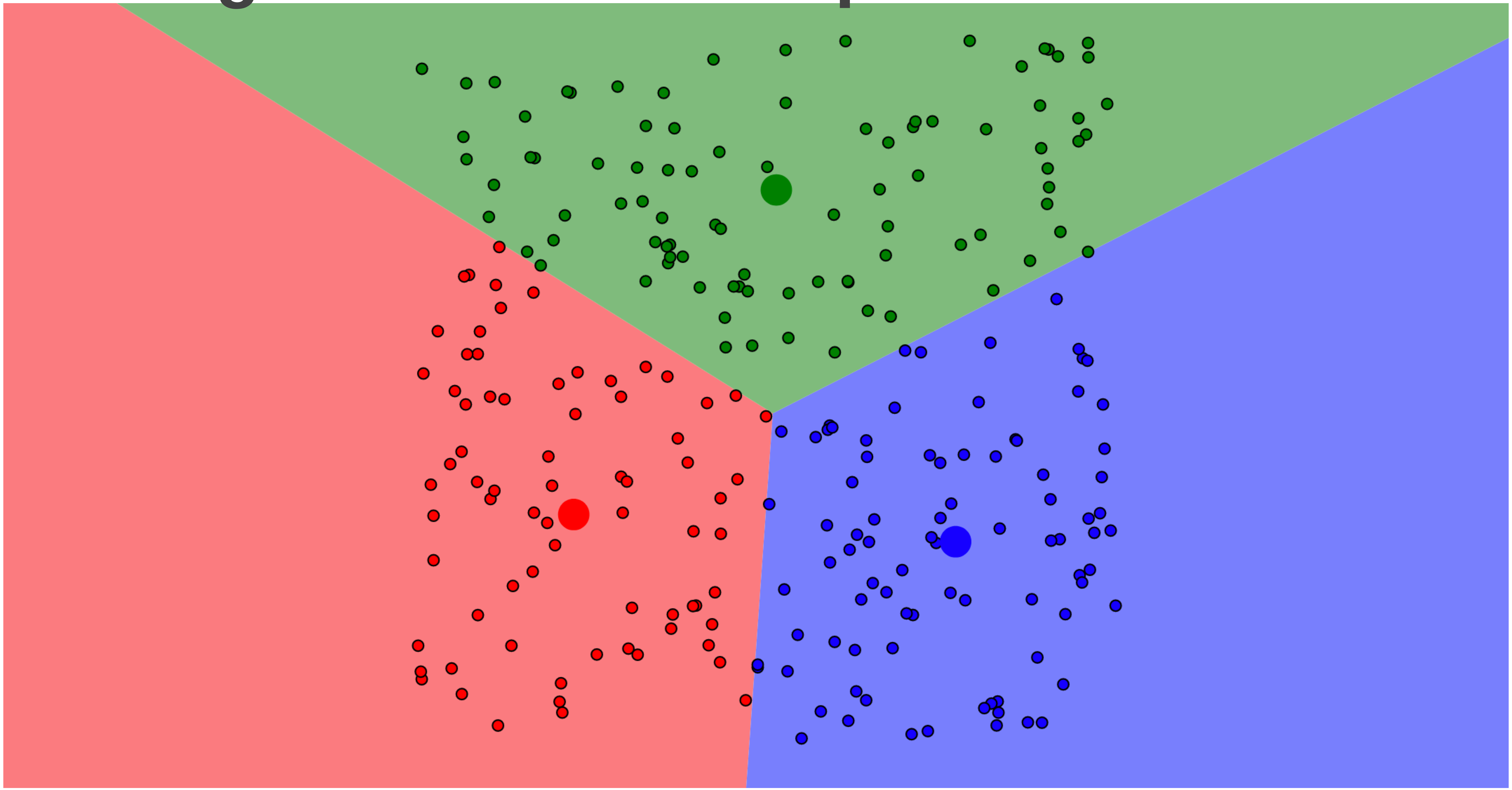
Re-Assign Data Points



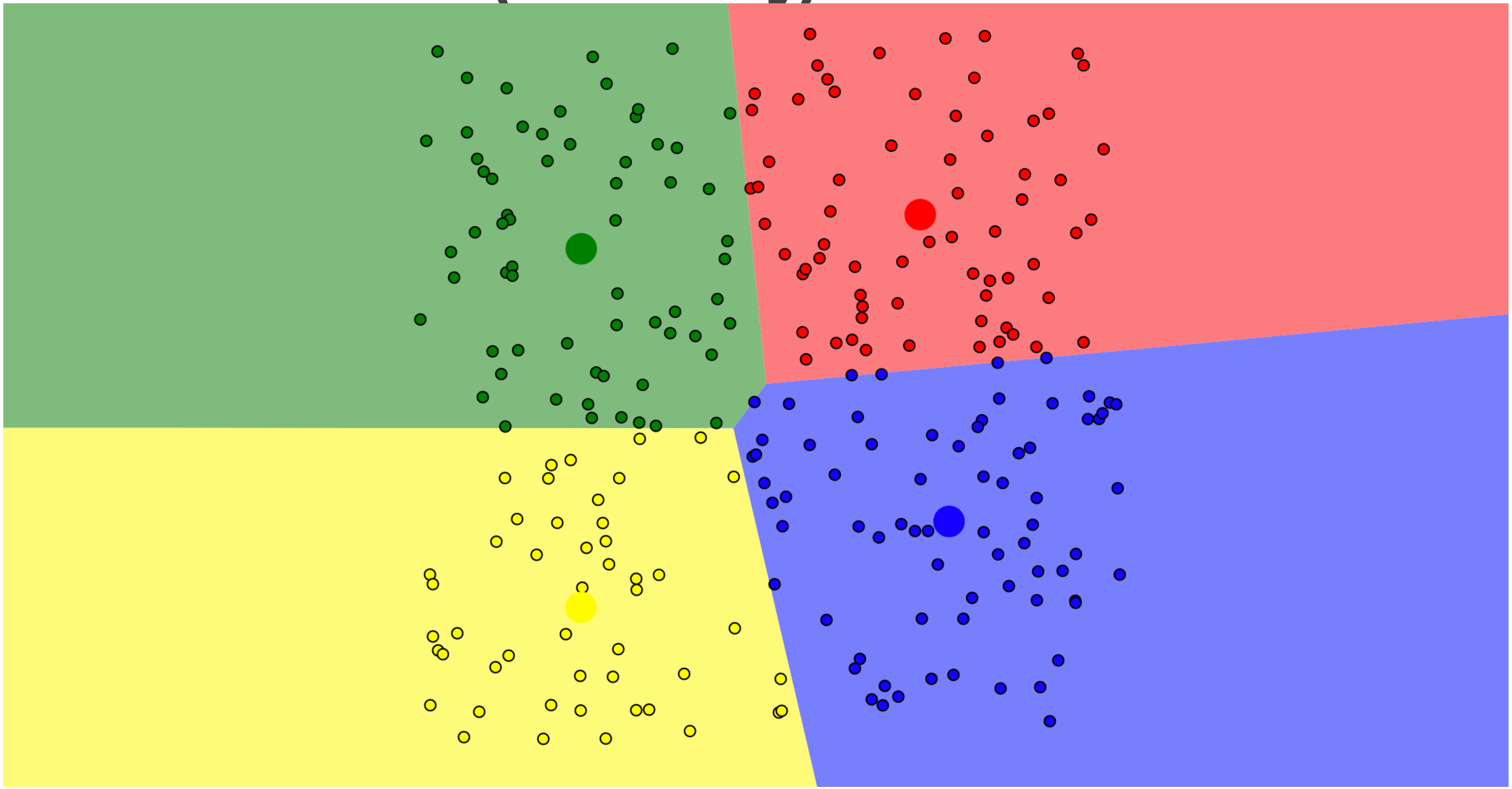
Update Centroids



Re-Assign Data Points - Stop

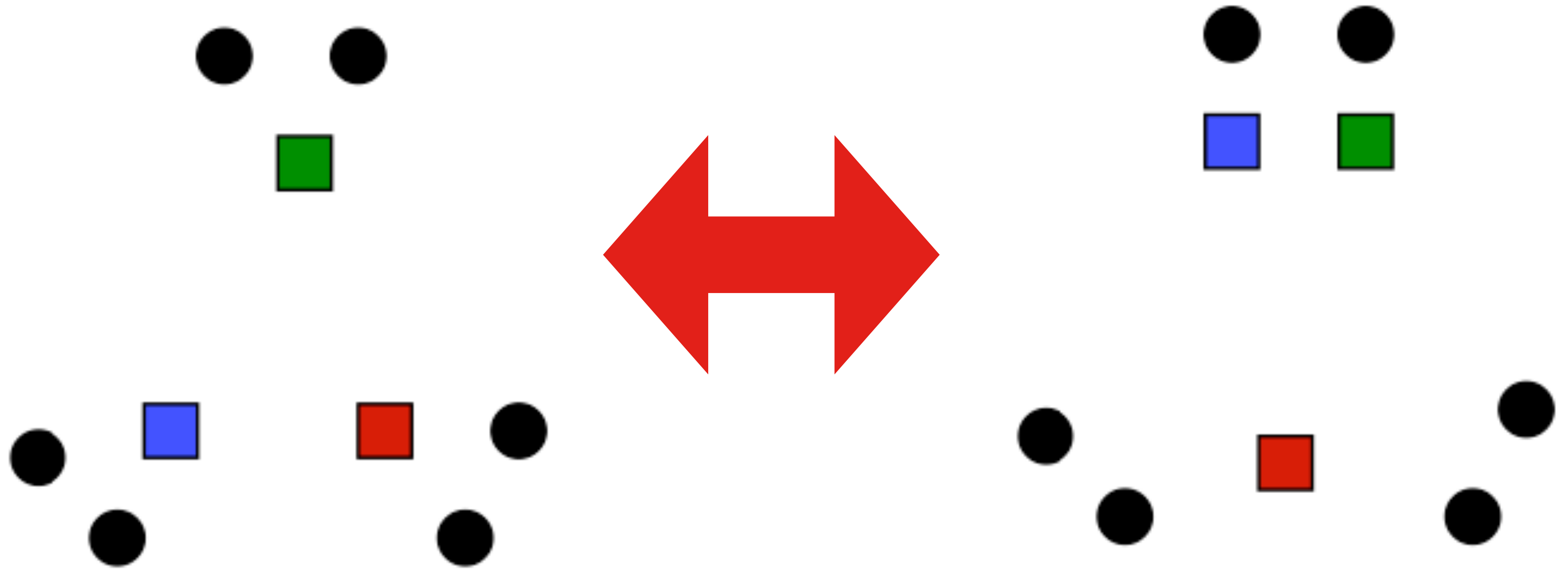


Add 4 Centroids (randomly)

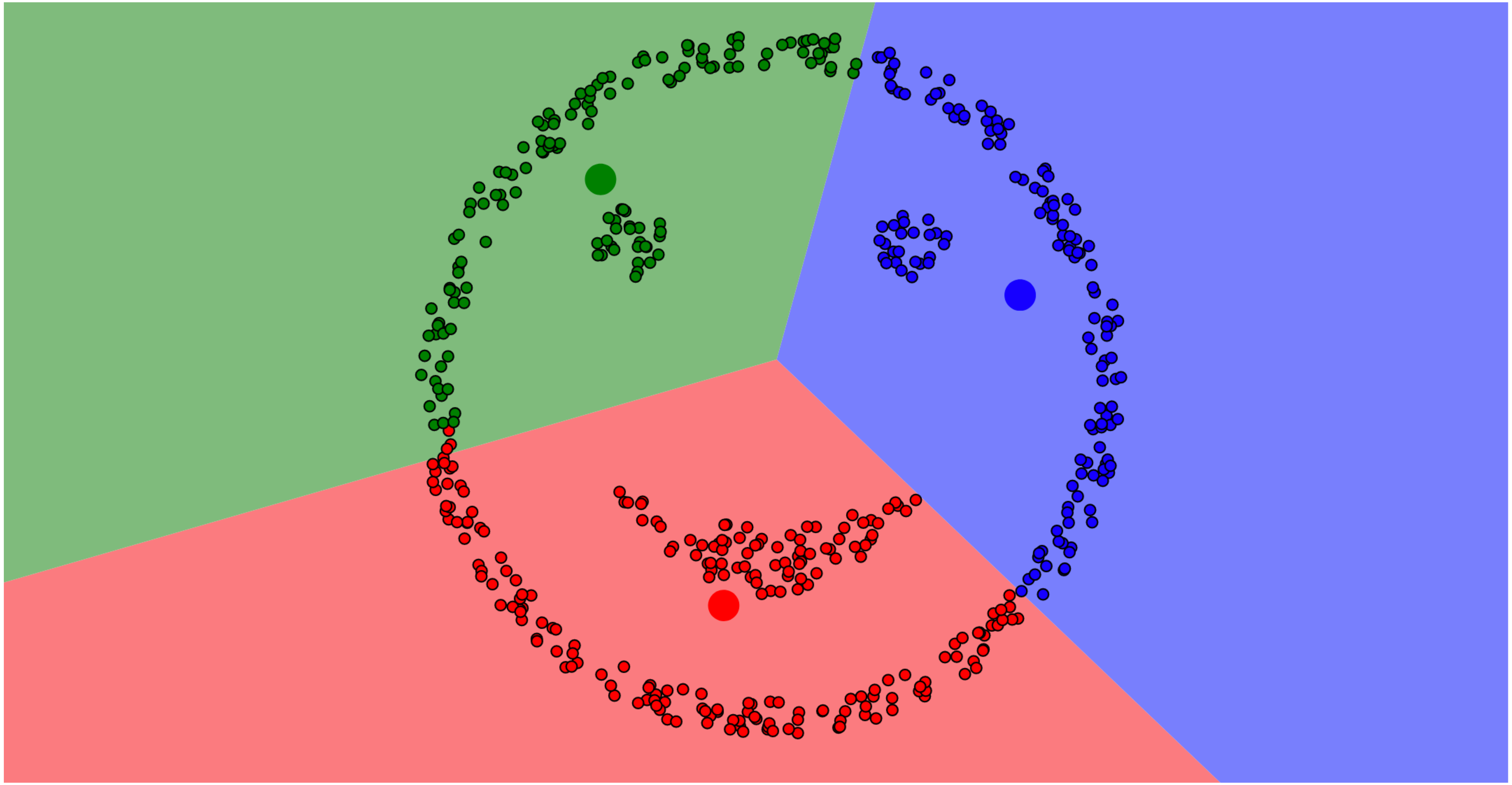


K-Means Pros and Cons

- **Pros**
  - Simple, fast to compute
  - Guaranteed to converge in a finite number of iterations
- **Cons/issues**
  - Setting *k*?
    - One way: silhouette coefficient
  - K-means algorithm is a heuristic
    - It does matter what random points you pick!
  - Sensitive to outliers
  - Detects spherical clusters



K-means not able to properly cluster



Machine Learning For Design

Lecture 8bis - Designing And Develop Machine Learning Models / Part 3

Alessandro Bozzon
30/03/2022

mlfd-io@tudelft.nl
www.ml4design.com

Credits

- Grokking Machine Learning. Luis G. Serrano. Manning, 2021
- https://scikit-learn.org/stable/modules/tree.html
- CIS 419/519 Applied Machine Learning. Eric Eaton, Dinesh Jayaraman. https://www.seas.upenn.edu/~cis519/spring2020/