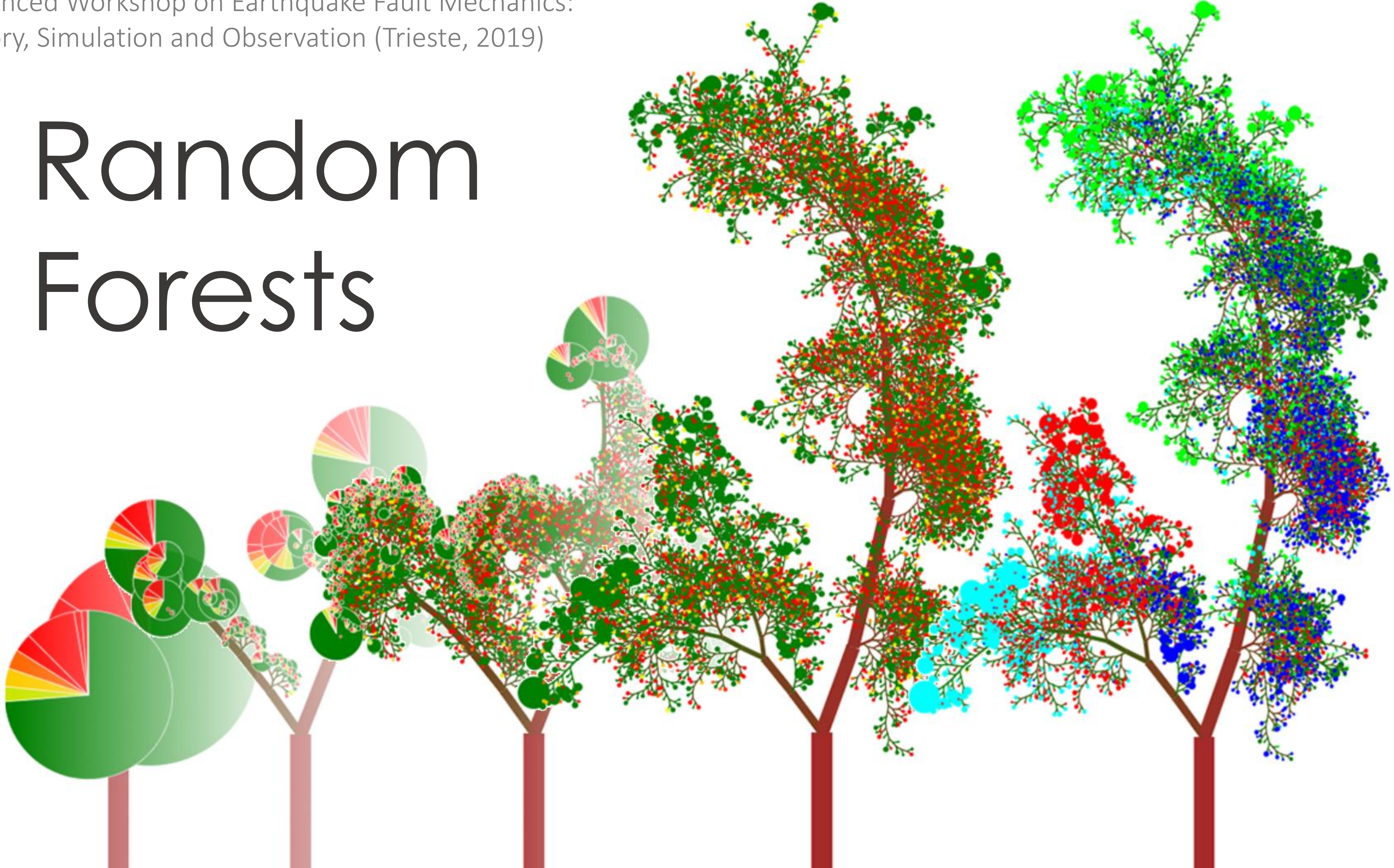
# Random Forests



http://www.rhaensch.de/rfvis.html

# AI vs. ML vs. DL

## Artificial Intelligence (AI)

- Chess computers
- Computer games
- Robotics
- Decision policies

### Machine Learning (ML)

- Random Forests
- Support Vector Machines

#### Deep Learning (DL)

Neural Networks with many (up to hundreds) of “layers”

# What's the difference?

- Neural Networks make decisions based on… well… *something*
- Random Forests (RF) make decisions based on well-defined rules
- RFs are easier to interpret, decision process can be visualised
- … but RFs require a particular type of input

# Example: Anderson's Irises

*Iris setosa*

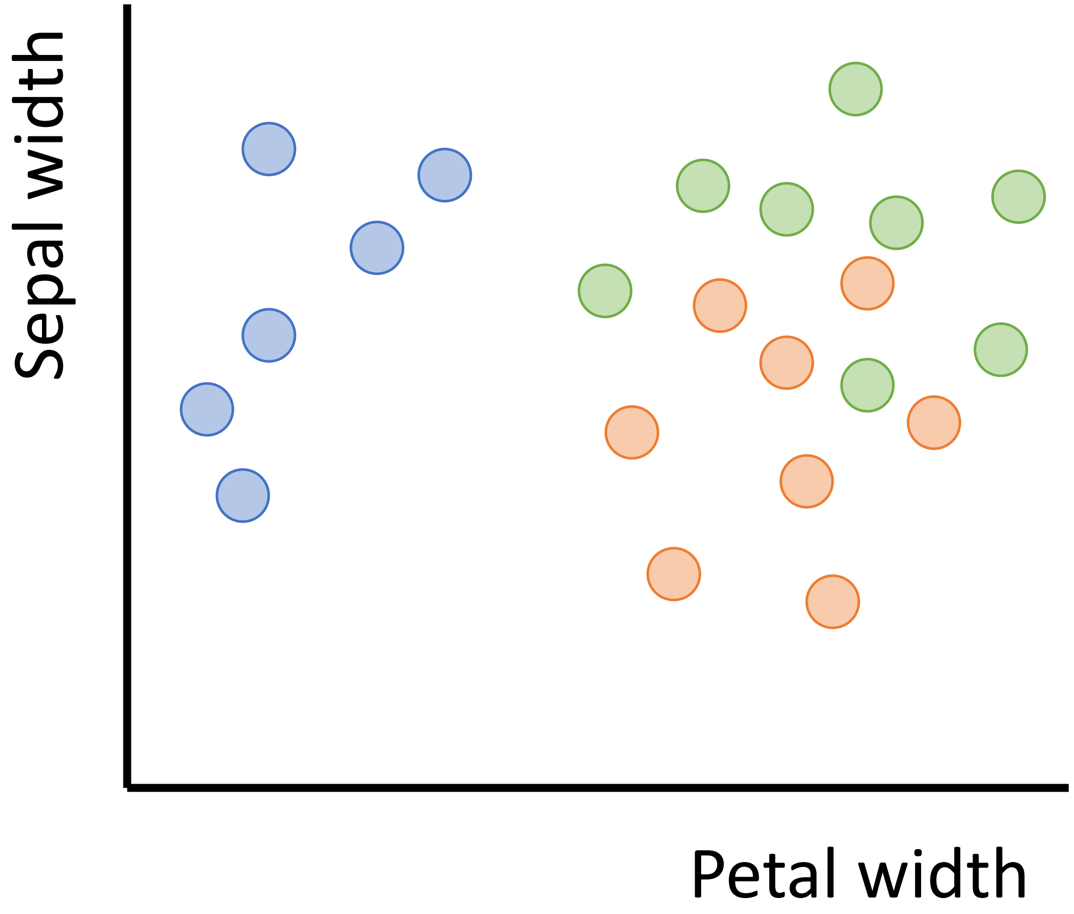
*Iris virginica*

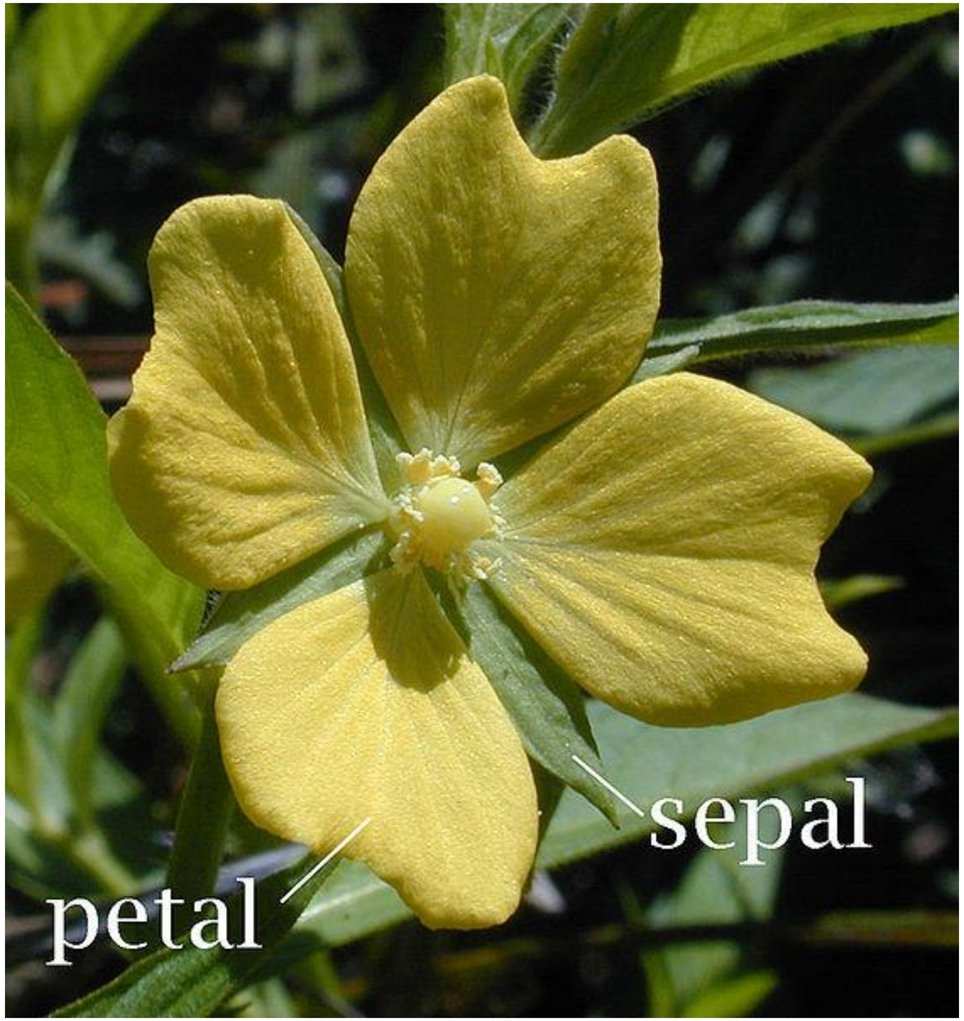
*Iris versicolor*

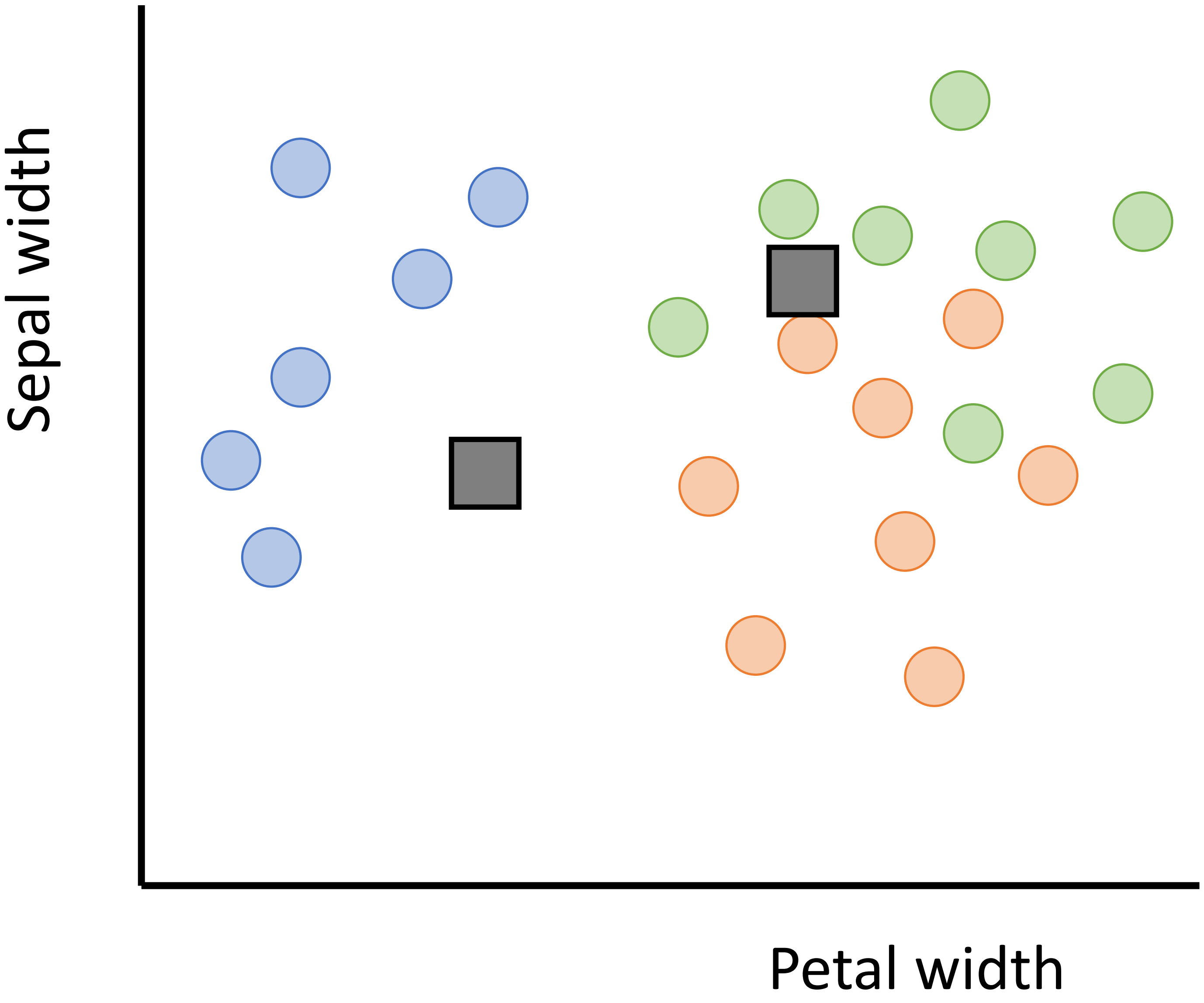
Wikipedia

# Example: Anderson's Irises

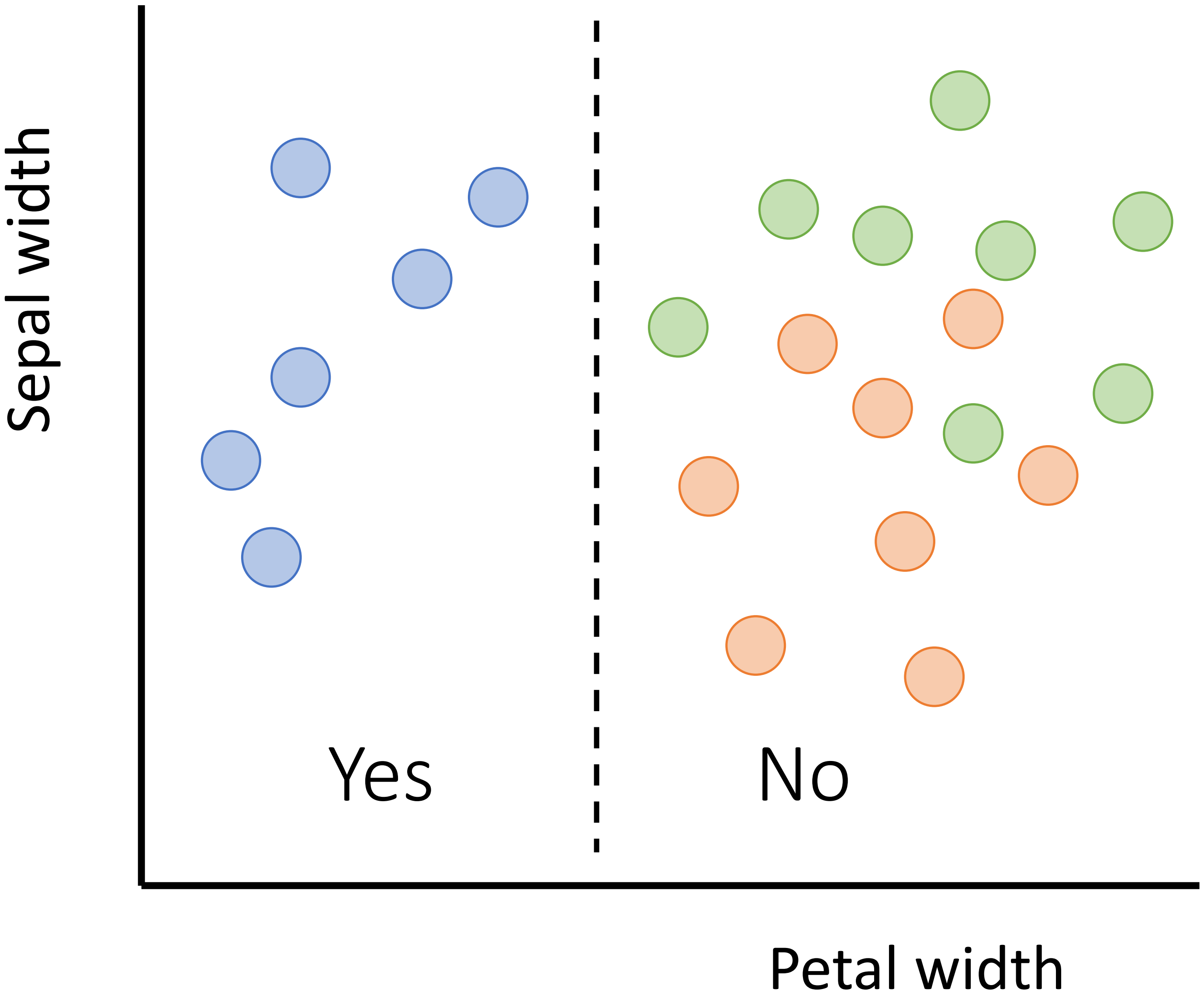
https://en.wikipedia.org/wiki/Sepal



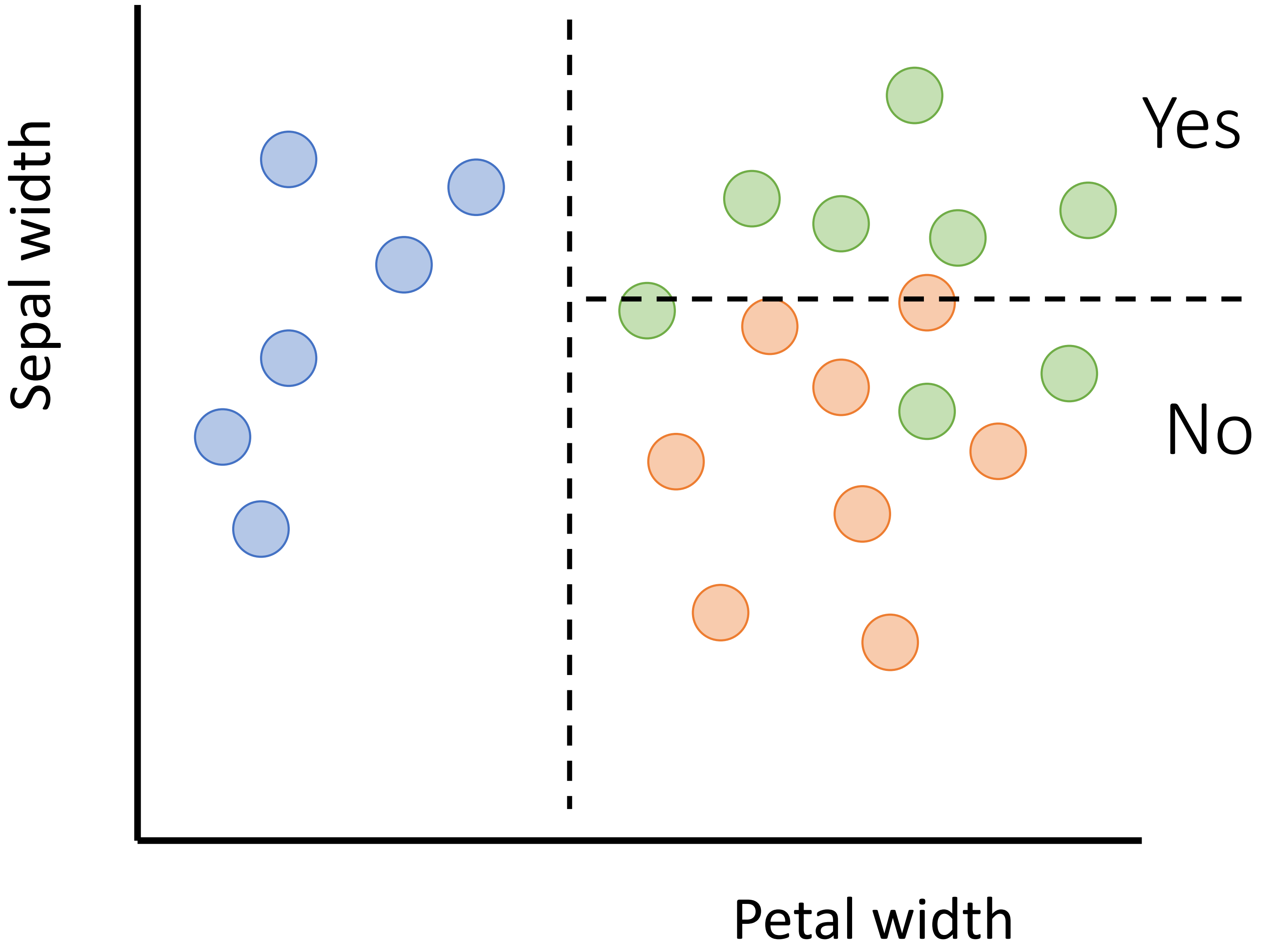




Sepal width
Petal width

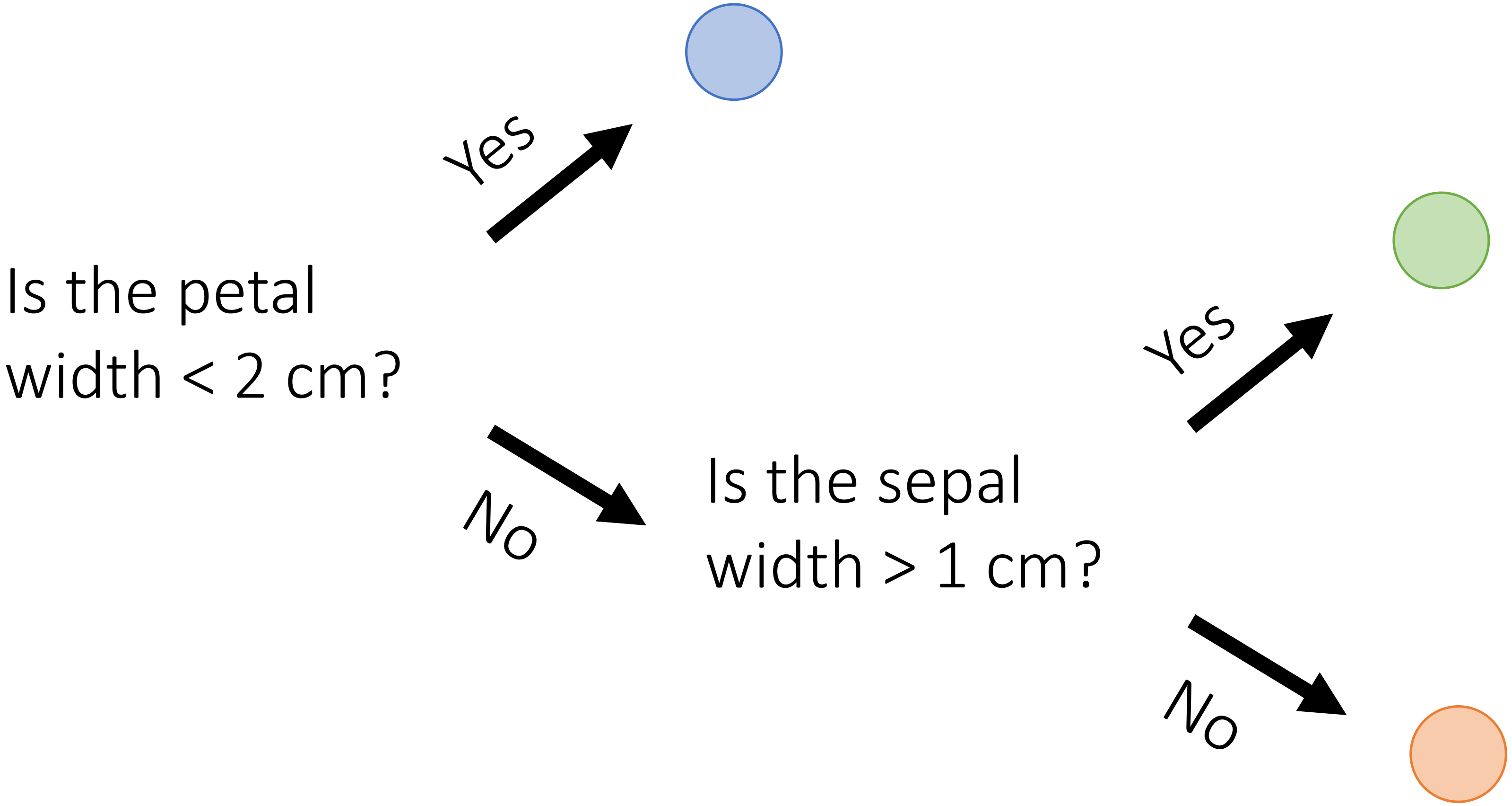
*Is the petal width < 2 cm?*

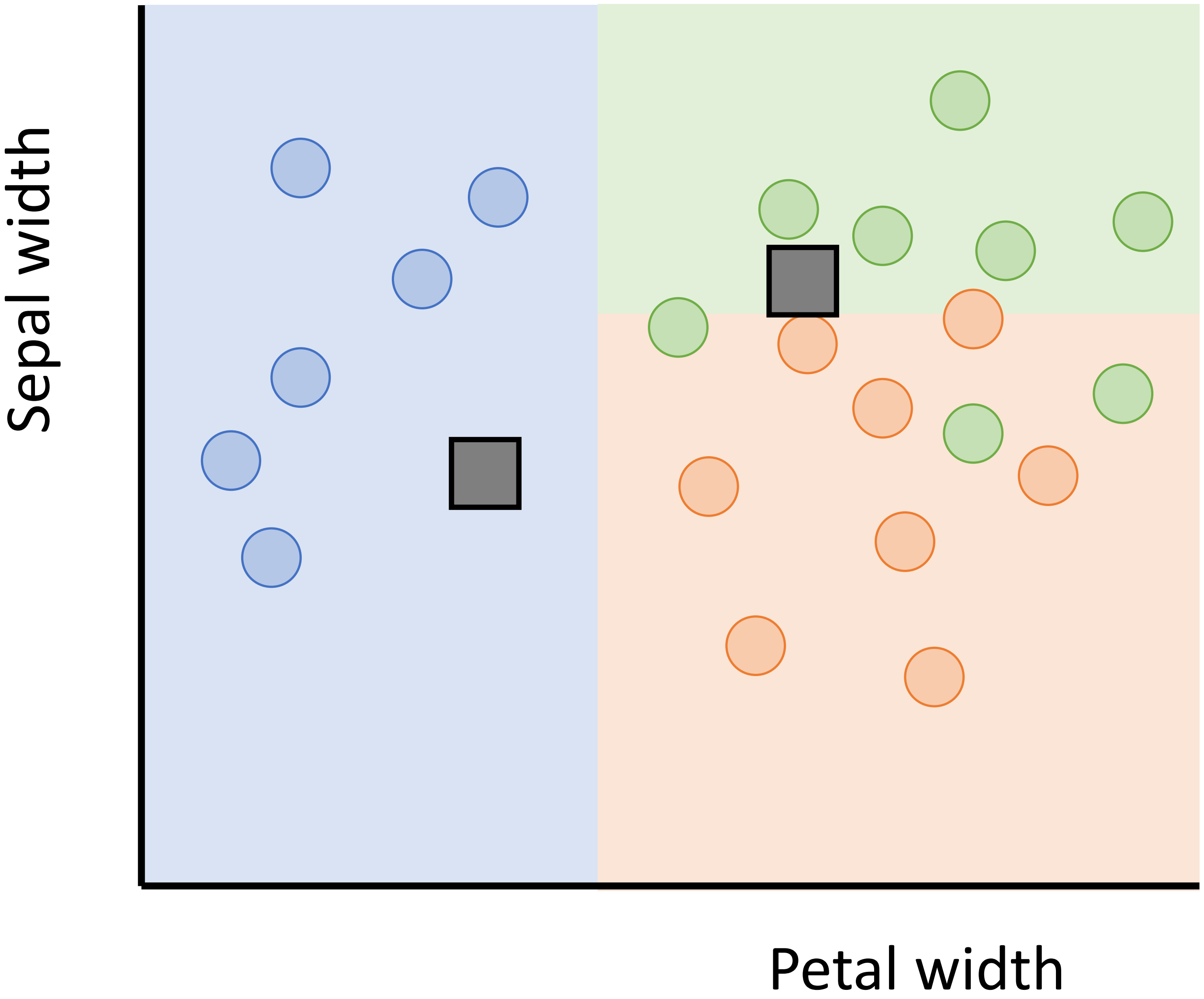


*Is the sepal width > 1 cm?*

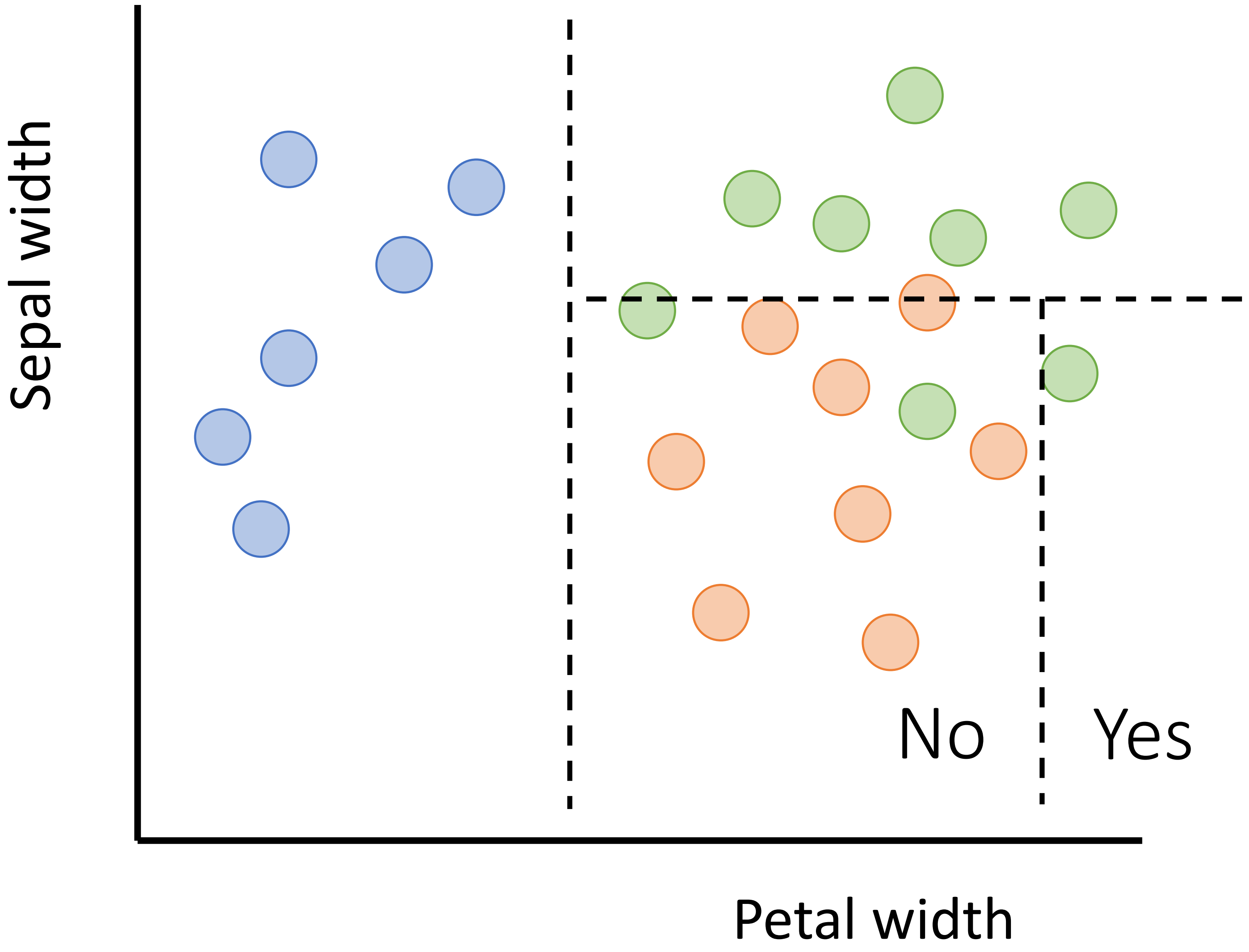


# Decision Trees

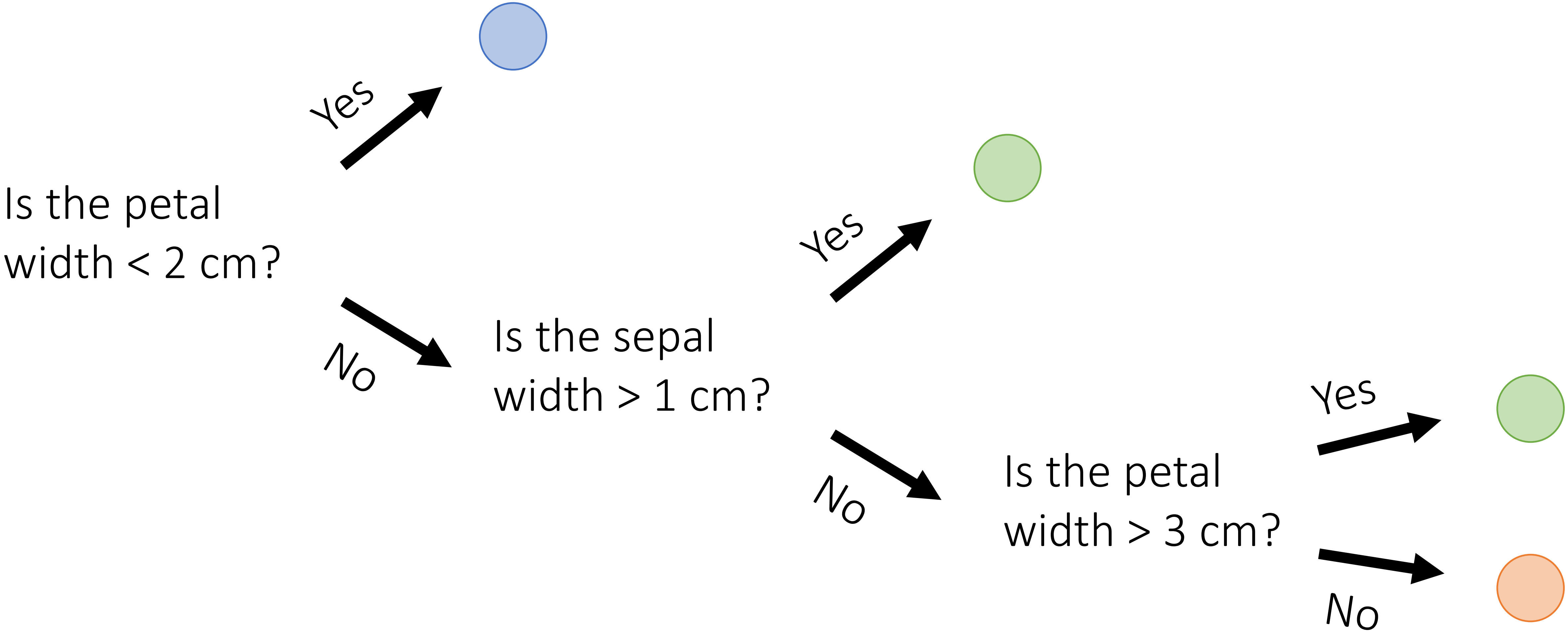


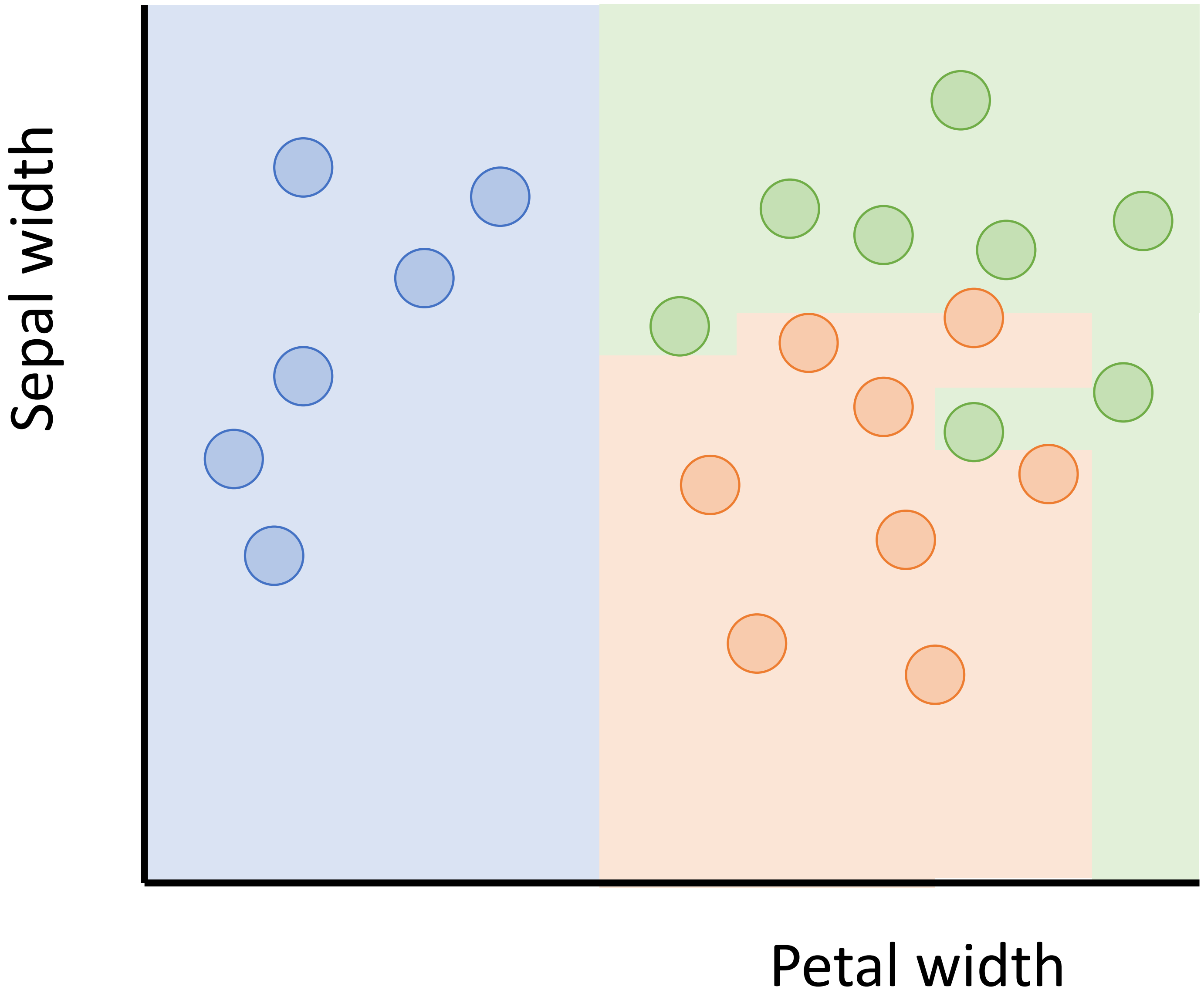

Sepal width
Petal width

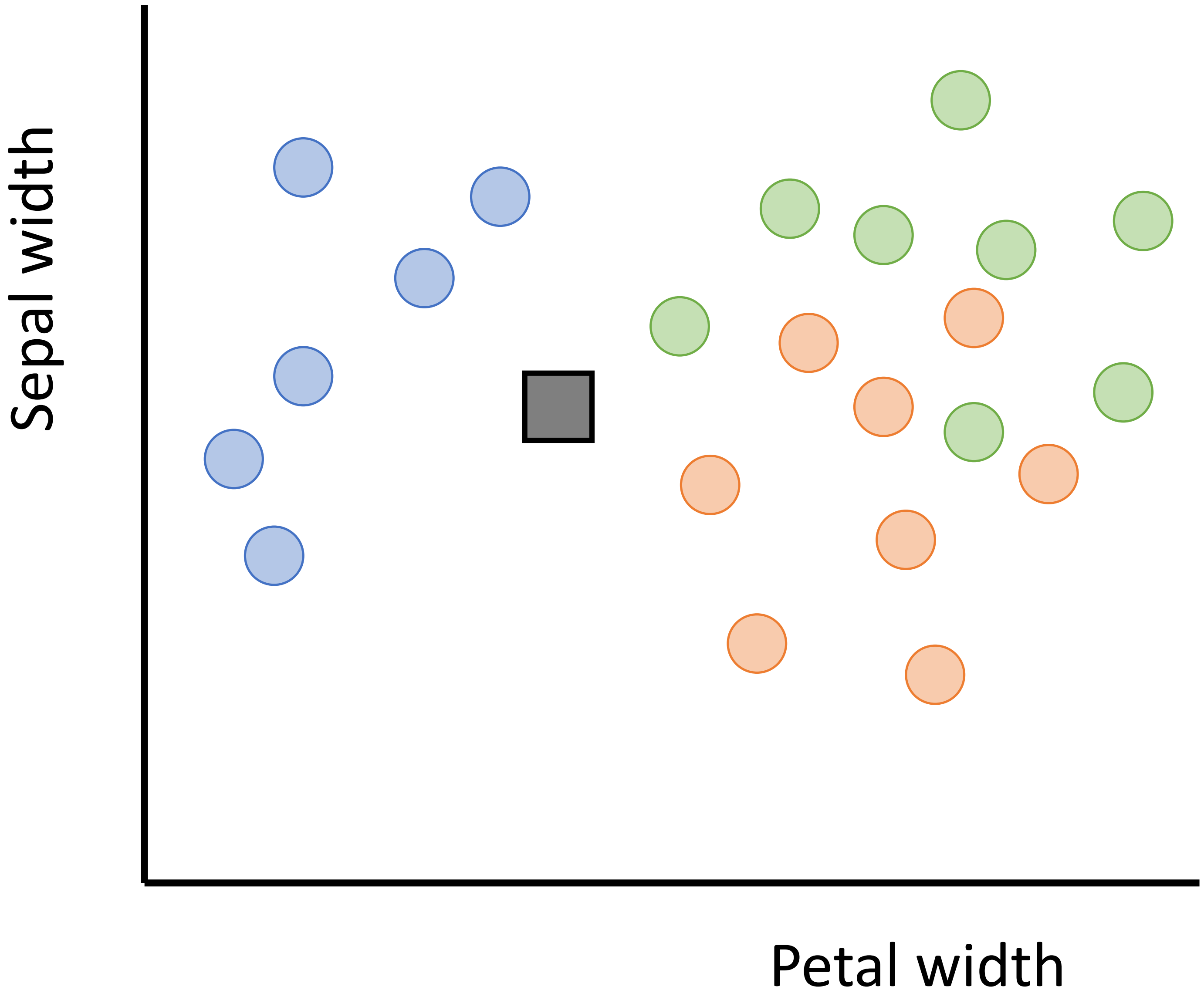
*Is the petal width > 3 cm?*



# Decision Trees




Sepal width
Petal width

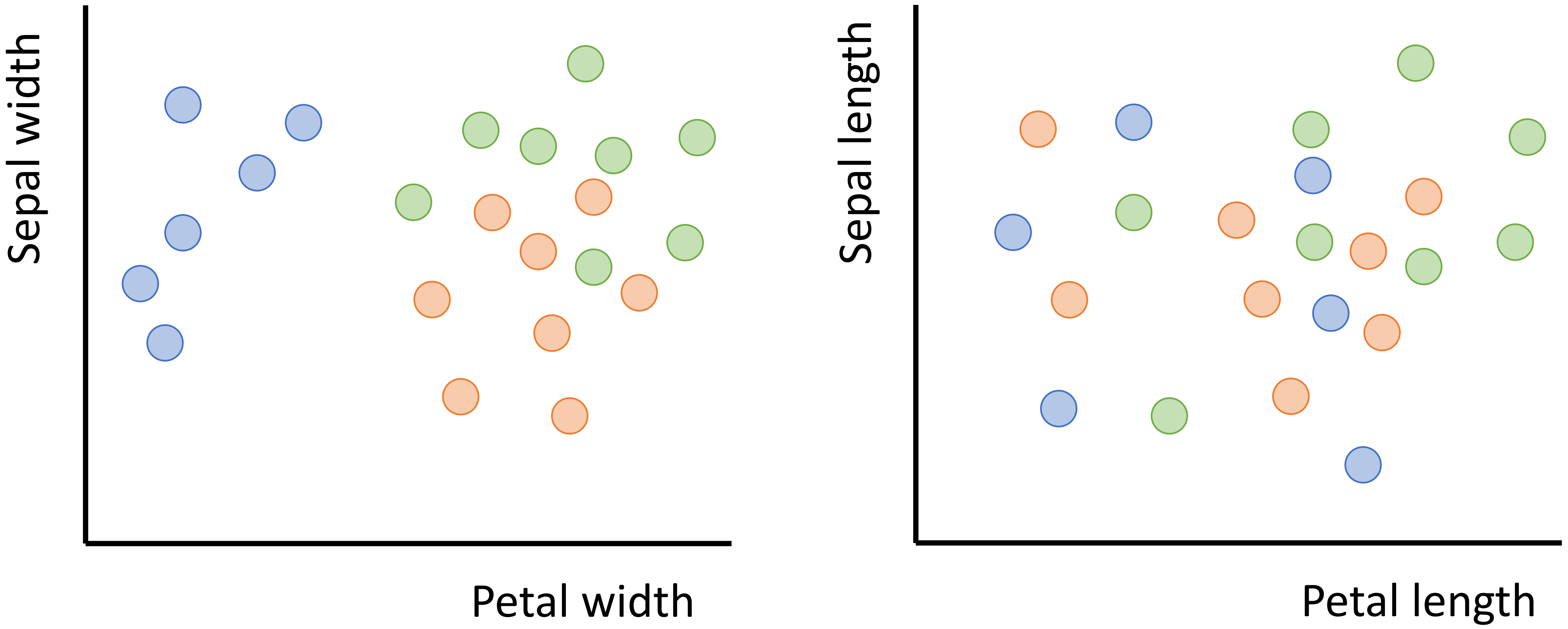

Sepal width
Petal width

# RF: Democracy of Decision Trees

- Decision Trees make decisions that split the data most efficiently
- Two trees with different data will make different decisions
- Random Forests:
  - Create $N$ Decision Trees
  - Give each tree a different subset of the data (randomly)
  - Average the predictions of all the trees in the "forest"

# Visualise feature importance

- Input data has “features” (sepal width/length, petal width/length)
- Which of these features is most important?


Sepal width
Petal width
Sepal length
Petal length

# Visualise feature importance

- Input data has “features” (sepal width/length, petal width/length)
- Which of these features is most important?
- With RFs it is possible to “calculate” relative importance of features

# Application of RF

AGU PUBLICATIONS

Geophysical Research Letters

RESEARCH LETTER

10.1002/2017GL076708

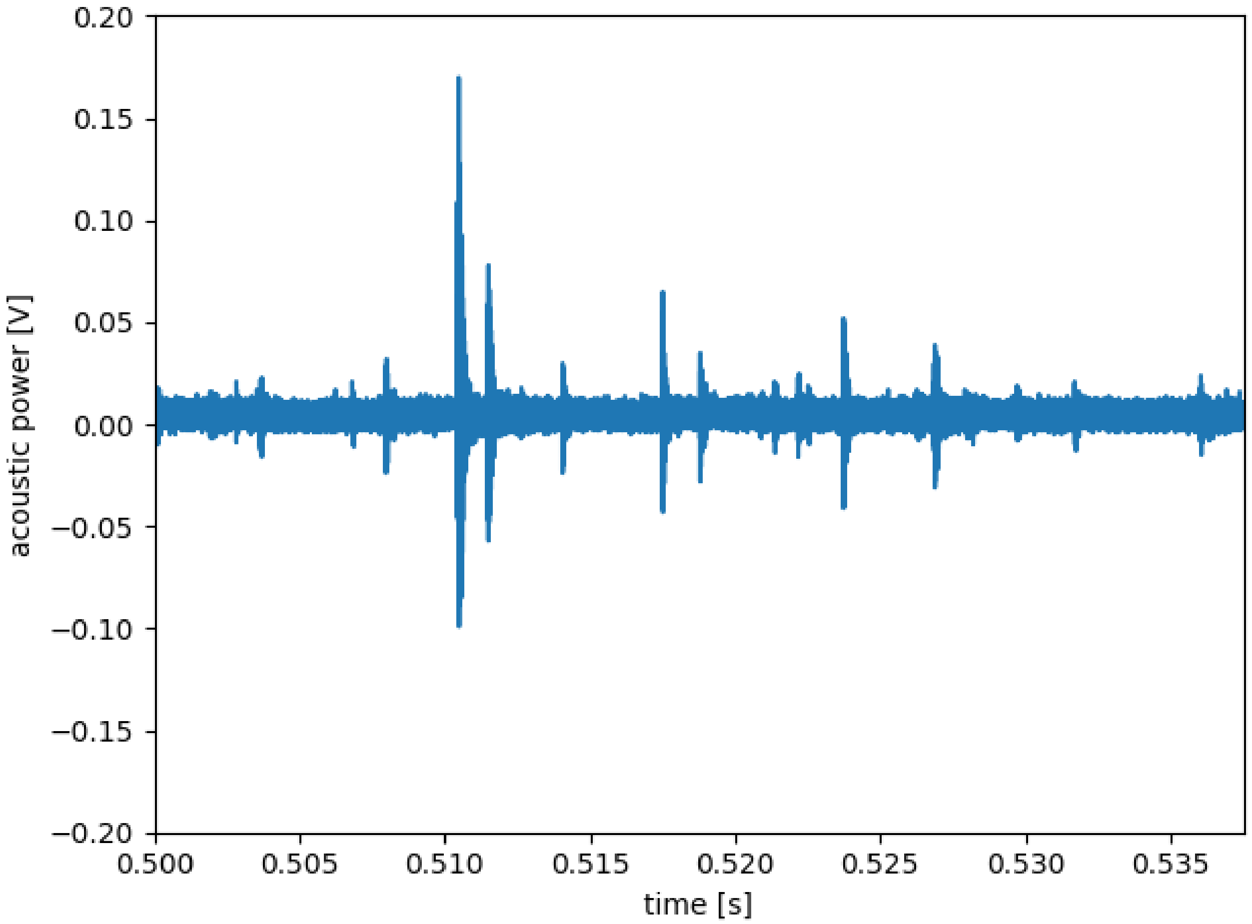
## Estimating Fault Friction From Seismic Signals in the Laboratory

Bertrand Rouet-Leduc[1], Claudia Hulbert[1], David C. Bolton[2], Christopher X. Ren[3], Jacques Riviere[2,4], Chris Marone[2], Robert A. Guyer[1], and Paul A. Johnson[1]

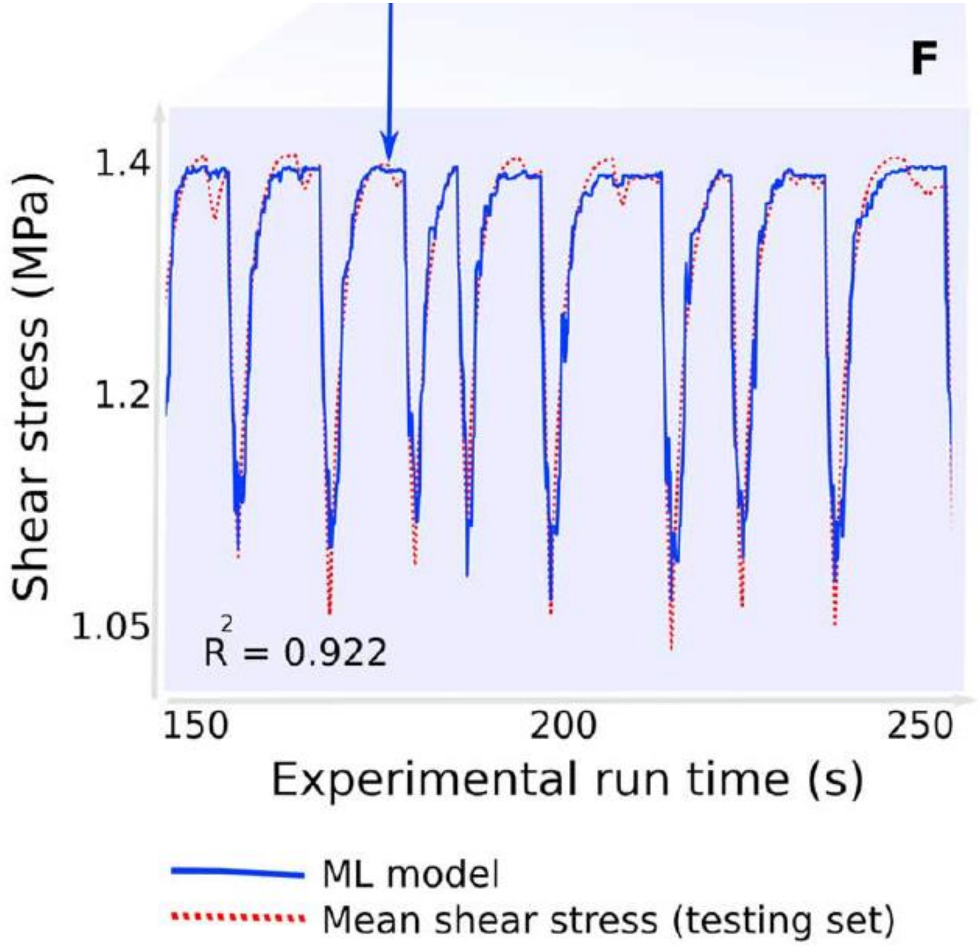
**Key Points:**

- Machine learning models can discern the frictional state of a laboratory fault from the statistical characteristics

# Application of RF



# Application of RF




Rouet-Leduc et al. (2018)

# RFs only accept “features”

- RFs are not suitable to analyse time series data (seismograms, GPS) or higher-dimensional data (spectrograms, images)
- Quality of predictions depends on selected features (“feature engineering”)
- Interpretation of certain features not always obvious
    - What is the meaning of the kurtosis of the signal squared?

# RF vs DL

- Random Forests are more interpretable, and are usually easier/faster to train (+ require less data)
- DL facilitates a wide range of architectures to handle different types of data, and are more flexible
- Pick the right tool for the job!

# Tutorial: Estimating EQ Damage

- After the 2015 Gorkha earthquake ($M_w$ 7.8) the Nepalese government initiated a large survey of the structural damage across the country
- For each building, the damage was classified as
    1. No/little damage
    2. Moderately damaged
    3. Severely damaged

DrivenData.org

# Tutorial: Estimating EQ Damage

- In addition, various socio-economical factors were recorded:
  - Building’s surface area, height, number of floors
  - Construction materials, foundation type
  - Primary use (residential, governmental, educational)
  - Number of families
  - Etc.

# Tutorial: Estimating EQ Damage

DrivenData Challenge:

Given the socio-economical factors (= features), predict the damage class of the building (1, 2, 3)