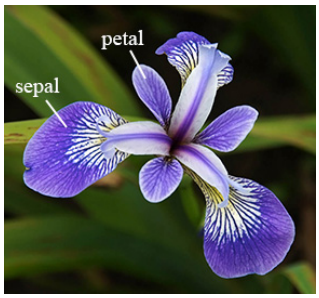
Iris Flower Dataset

To practice using the k-nearest neighbor algorithm in python, we'll be using a very famous dataset, the *Iris flower data set* from statistician and biologist Ronald Fisher in 1936. This dataset consists of measurements taken from 150 iris flowers from three different iris species, Iris setosa, Iris virginica, and Iris versicolor. Four features were measured for each flower sample: sepal length, sepal width, petal length, and petal width.



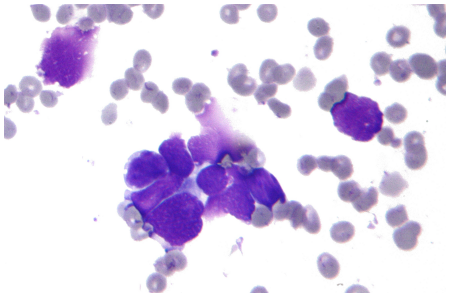
The *petals* are the colorful leaves of a flower, which surround the reproductive parts of the flower.

The *sepals* are outside of the petals, and they protect the flower while it's in bud. When the flower blooms, the sepals support the petals. In most flowers, the sepals are green, but for these irises, the sepals are purple.

When Fisher first introduced the iris data set, he showed how a machine learning method called *linear discriminant analysis* can be used for classification of these iris species. Since then, the Iris dataset has become one of the most commonly used datasets for practicing, testing, and demonstrating machine learning classification algorithms. As you continue to study machine learning, you're likely to come across the iris dataset again and again!

Breast Cancer Dataset

Another famous dataset is the Wisconsin breast cancer dataset. This dataset consists of measurements taken from an image of a breast mass, and these measurements describe the shape of the cell nuclei visible in the image. Each sample is classified as malignant or benign, telling us if it is a cancerous cell. The dataset contains measurements from 357 benign samples, and 212 malignant samples.



The following measurements were taken for each nucleus in each image: radius, texture, perimeter, area, smoothness, compactness, concavity, number of concave points, symmetry, and fractal dimension.

For each image, the mean, standard error, and average of the three largest values were computed over all nuclei in the image. These measurements are the features in the dataset.

The goal of studying this dataset is to answer the question: can we predict if a sample is cancerous or not, based on measurements of the shape of the cells? This would be useful as an initial screening procedure; if the cells appear to be cancerous, further tests could then confirm this preliminary diagnosis.