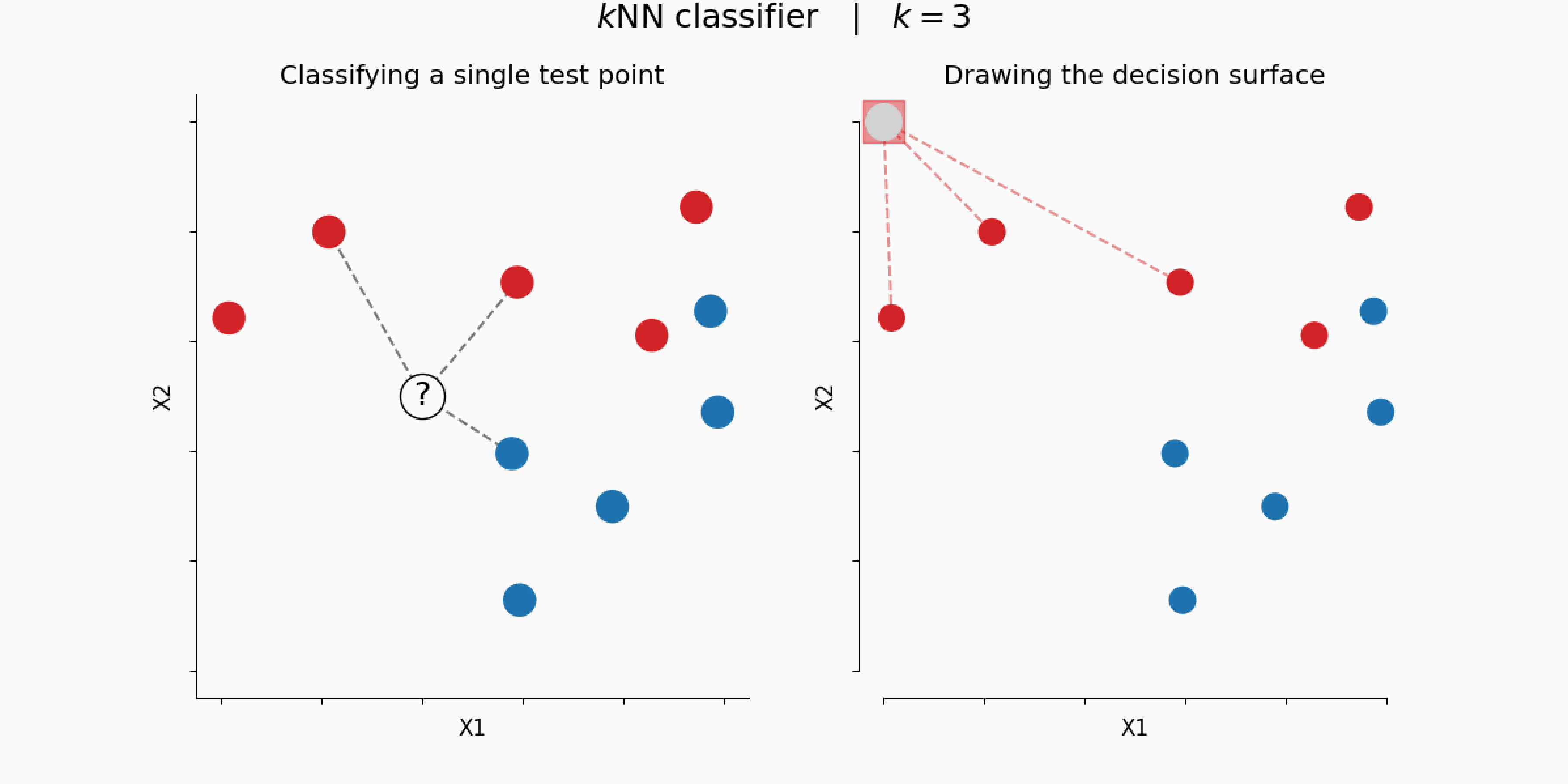
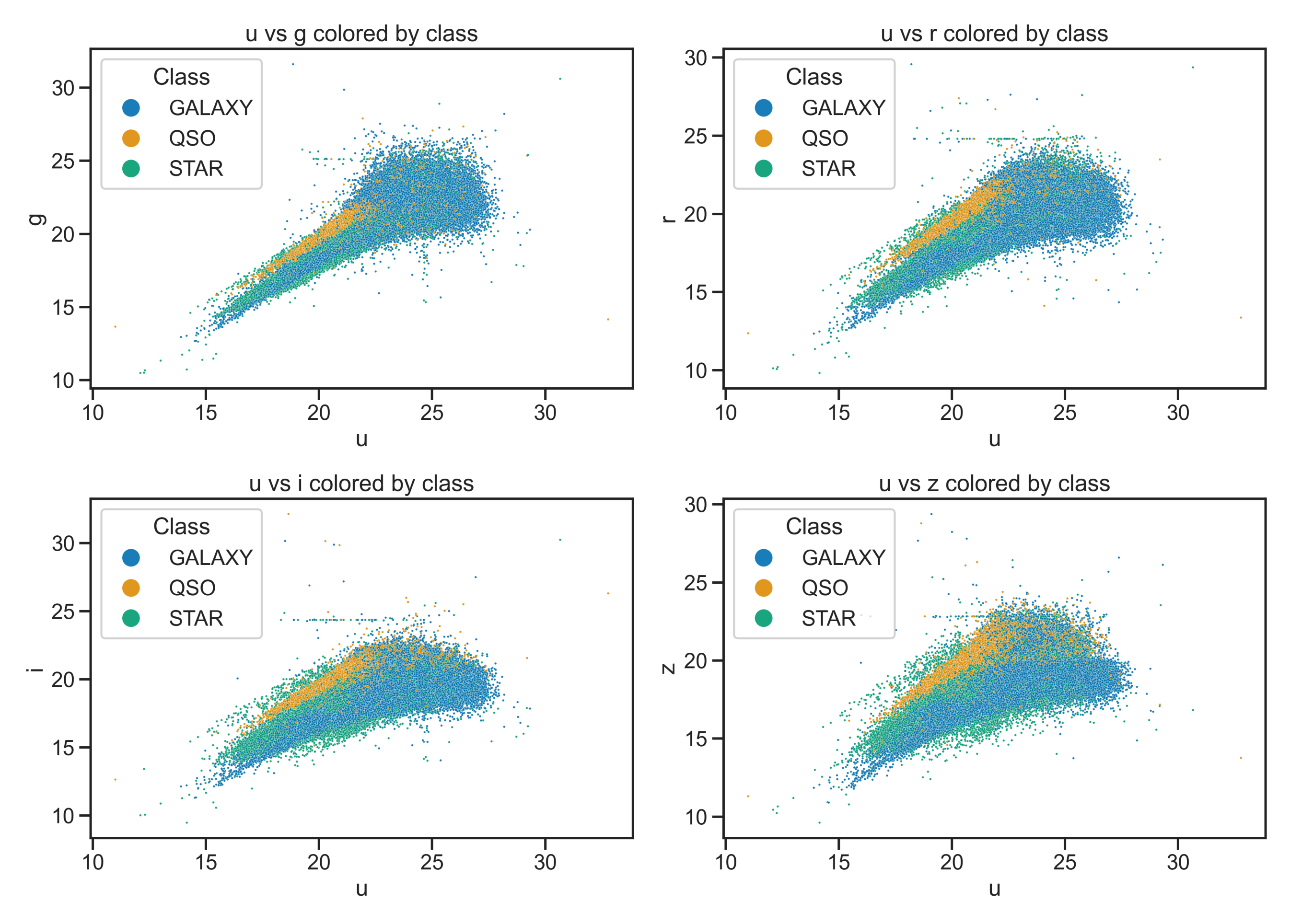
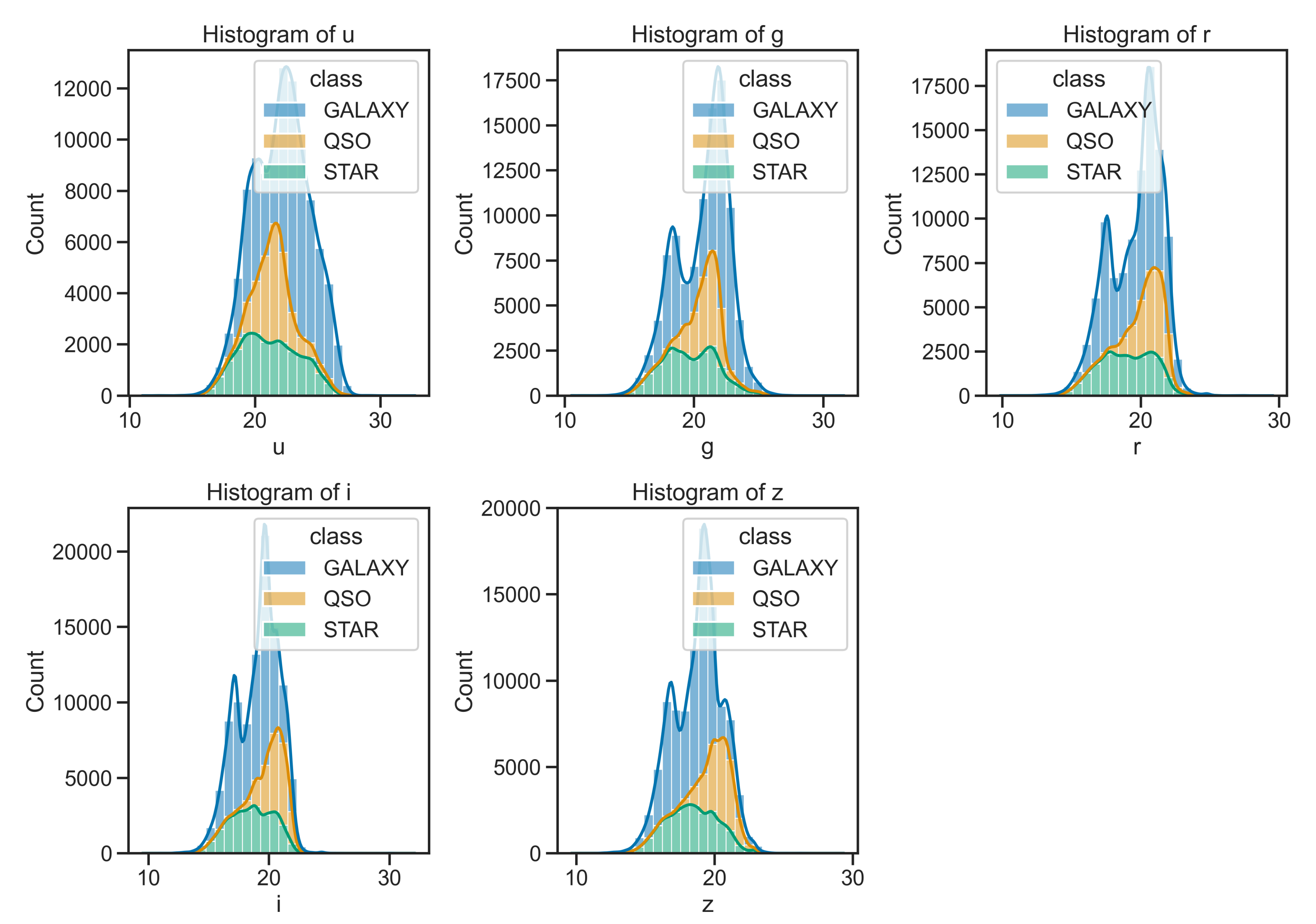
Day 02 - Classification



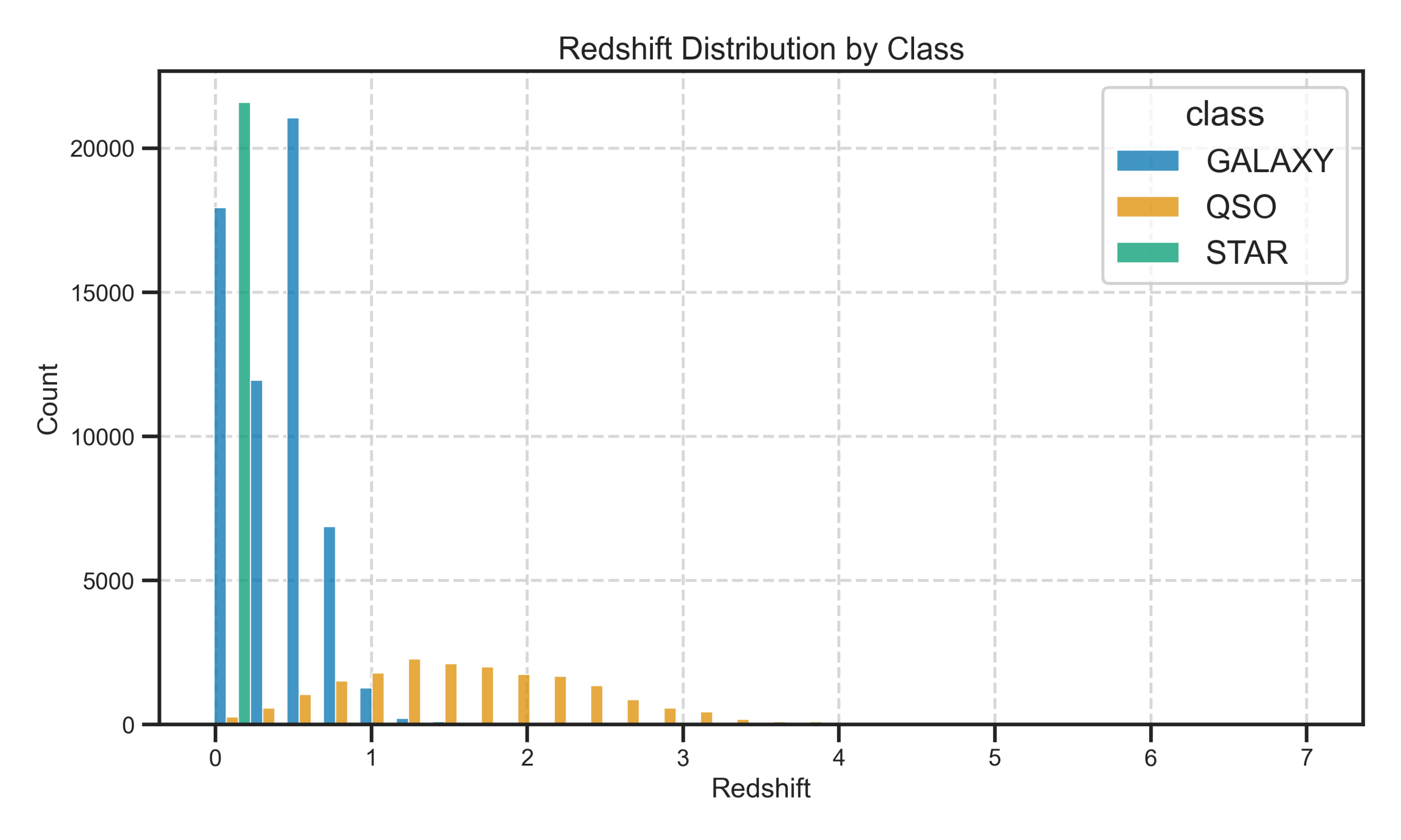
SDSS Data Set



SDSS Data Set



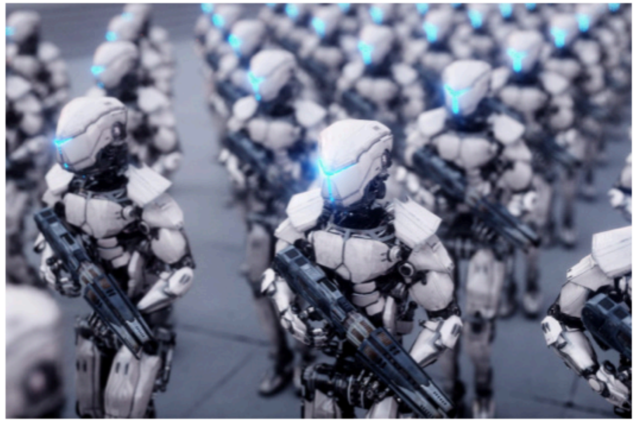
SDSS Data Set



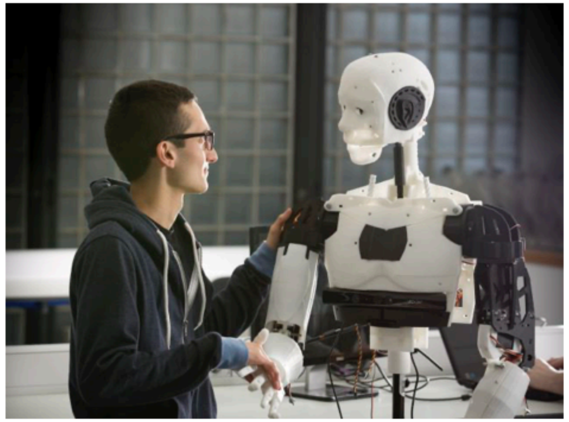
Classification Task

- Using the SDSS data set, we will classify objects as stars or quasars.
    - At first, we will only use the color information (u-g, g-r, r-i) to classify objects.
    - Later, we will add the redshift information (z) to improve our classification.
- Then, we will perform a 3-class classification to distinguish between stars, quasars, and galaxies; here we will use all available features including redshift.

Machine Learning



What society thinks we do

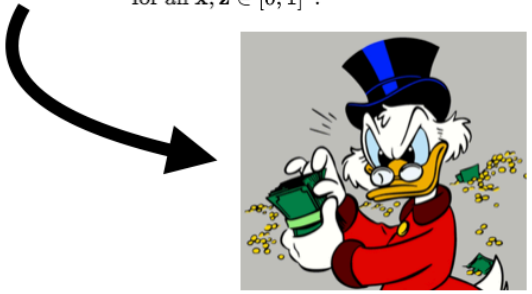


What our friends think we do

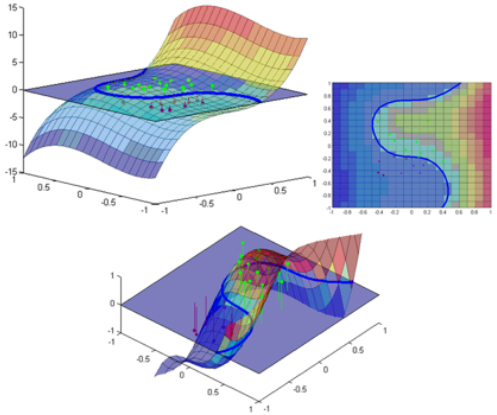
What our families thinks we do

$$K_k^{cc}(\mathbf{x},\mathbf{z}) = \sum_{k_1,\ldots,k_d,\sum_{j=1}^{d} k_j = k} \frac{k!}{k_1!\cdots k_d!}\left(\frac{1}{d}\right)^k \prod_{j=1}^{d} \mathbf{1}_{\lceil 2^{k_j} x_j \rceil = \lceil 2^{k_j} z_j \rceil},$$

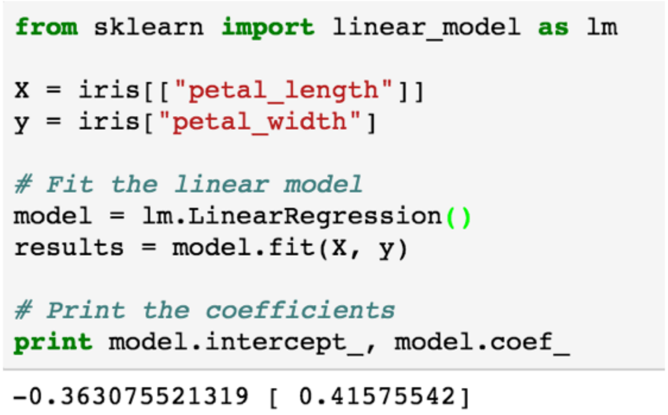
$$\text{for all } \mathbf{x},\mathbf{z} \in [0,1]^d.$$



What my boss thinks we do



What we think we do



```
from sklearn import linear_model as lm

X = iris[["petal_length"]]
y = iris["petal_width"]

# Fit the linear model
model = lm.LinearRegression()
results = model.fit(X, y)

# Print the coefficients
print model.intercept_, model.coef_
```

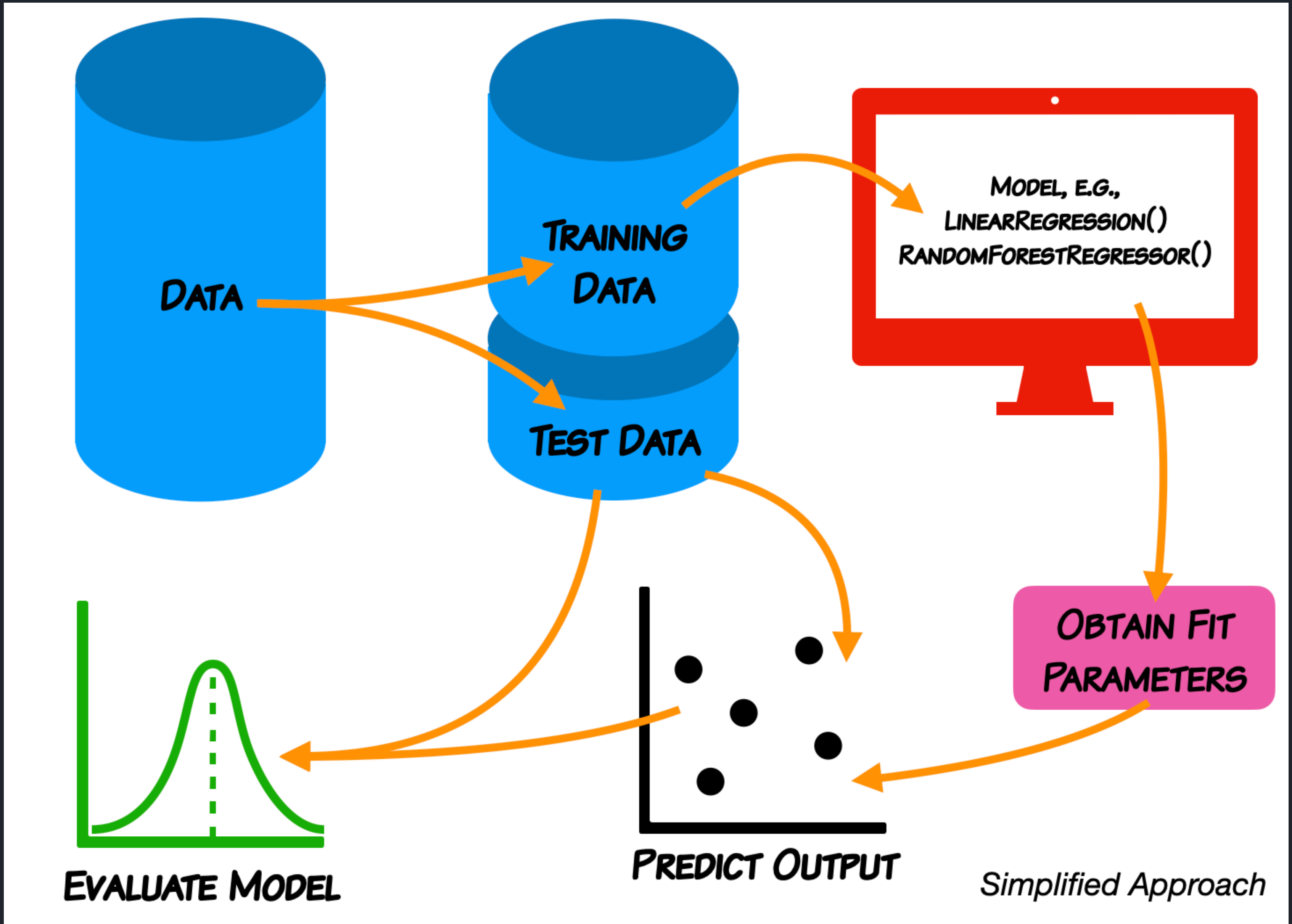
```
-0.363075521319 [ 0.41575542]
```

What we actually do

Sci-Kit-Learn Classification

- Sci-Kit-Learn is a powerful Python library for machine learning.
- It provides a wide range of classification algorithms, including:
    - k-Nearest Neighbors (kNN) & Logistic Regression
    - Decision Trees & Random Forests
    - Support Vector Machines (SVM)
- It also includes tools for model evaluation, such as cross-validation and confusion matrices.

https://scikit-learn.org/stable/index.html



Data

Training Data

Test Data

Model, e.g.,
LinearRegression()
RandomForestRegressor()

Obtain Fit Parameters

Predict Output

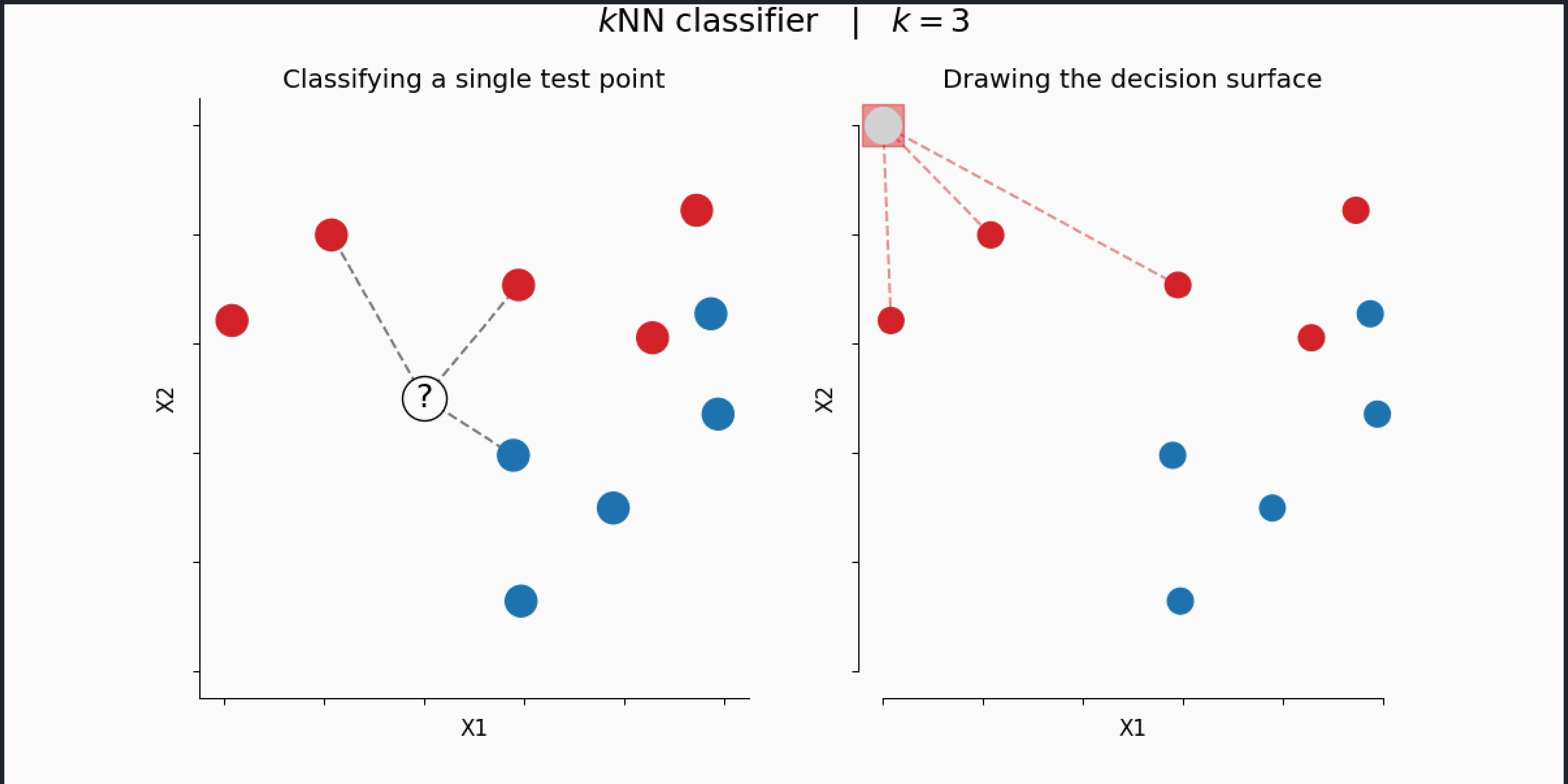
Evaluate Model

Simplified Approach

K-Nearest Neighbors (kNN)

- kNN is a simple and intuitive classification algorithm.
- It classifies a data point based on the majority class of its k nearest neighbors in the feature space.
- The distance metric (e.g., Euclidean distance) is used to determine the nearest neighbors.
- kNN is a non-parametric method, meaning it makes no assumptions about the underlying data distribution.
- It is sensitive to the choice of k and the distance metric.

K-Nearest Neighbors (kNN)



Today's Activity

- We will implement a kNN classifier using Sci-Kit-Learn to classify stars and quasars from the SDSS data set.
- We will:
    i. Load the SDSS data set and preprocess it.
    ii. Split the data into training and testing sets.
    iii. Train a kNN classifier on the training set.
    iv. Evaluate the classifier's performance on the testing set.
- We will focus on the evaluation metrics and visualizations to understand the classifier's performance.