Building Your First scikit-learn Solution

EXPLORING SCIKIT-LEARN FOR MACHINE LEARNING

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

scikit-learn for data and ML modeling

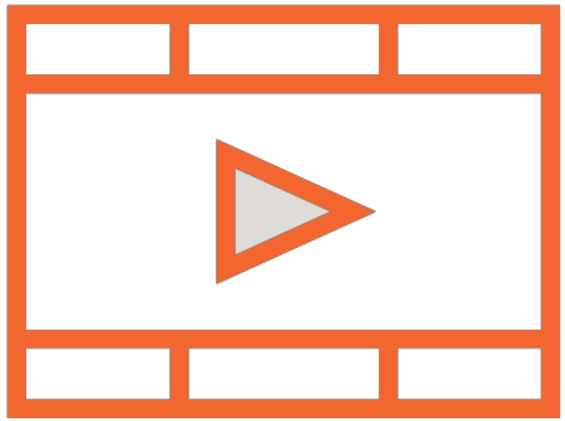
Relationship with NumPy, SciPy, Pandas, and Matplotlib

Algorithms for supervised and unsupervised learning

Contrast with TensorFlow, Keras, and other deep learning frameworks

Prerequisites and Course Outline

Prerequisites



Basic Python programming

Intended to be first ML course

No ML knowledge required

Course Outline

Introduction to ML and scikit-learn

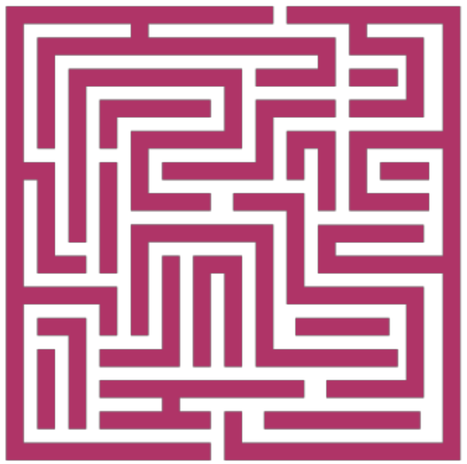
ML workflow with scikit-learn

Building simple ML models for regression and classification

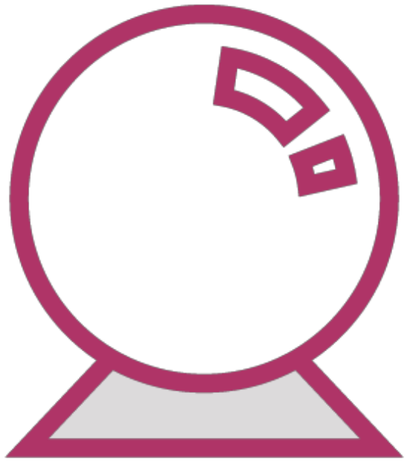
Introducing Machine Learning

A machine learning algorithm is an algorithm that is able to learn from data

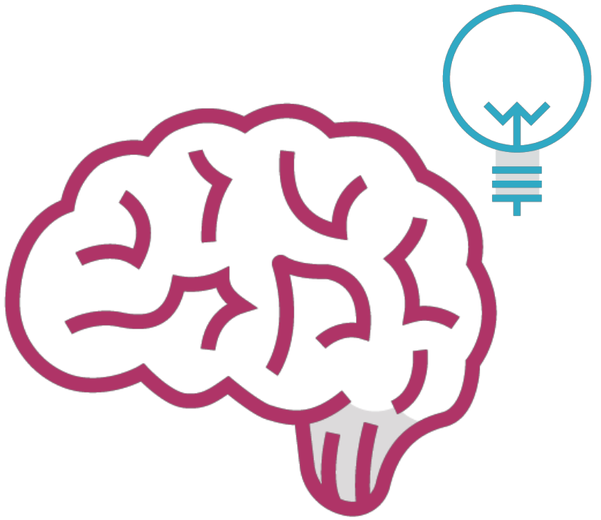
Machine Learning



Work with a huge maze of data



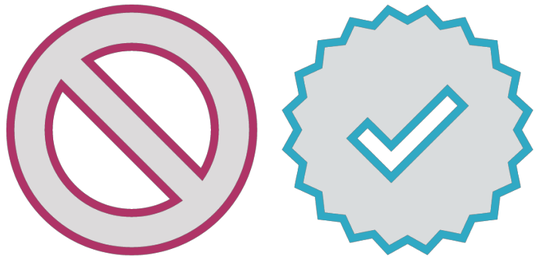
Find patterns



Make intelligent decisions

Machine Learning

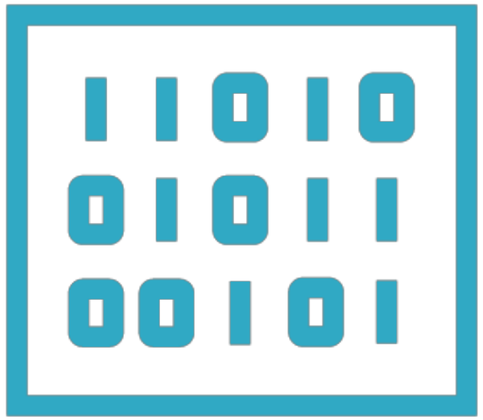
Emails on a server



Spam or Ham?

Trash or Inbox

Machine Learning



Images represented as pixels

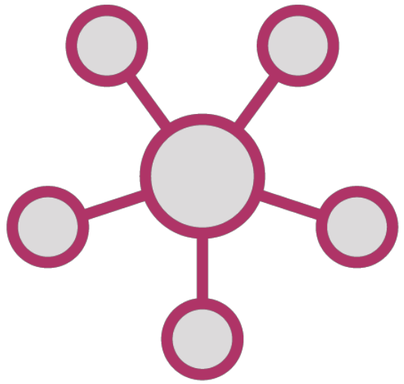
Identify edges, colors, and shapes

A photo of a little girl

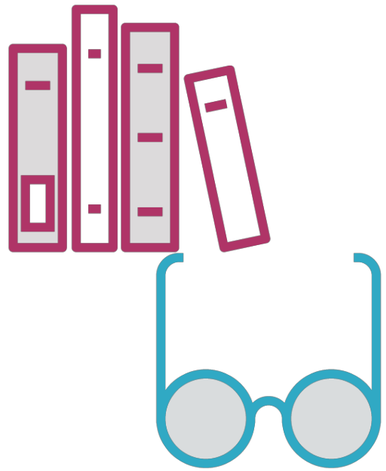
Types of Machine Learning Problems

Classification

Regression



Clustering



Dimensionality reduction

Types of Machine Learning Problems

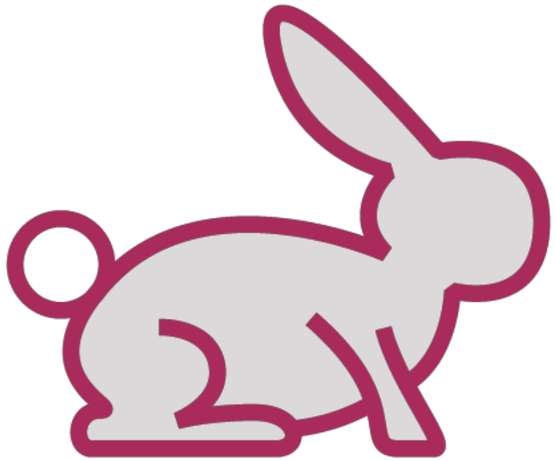
Classification

Regression

Clustering

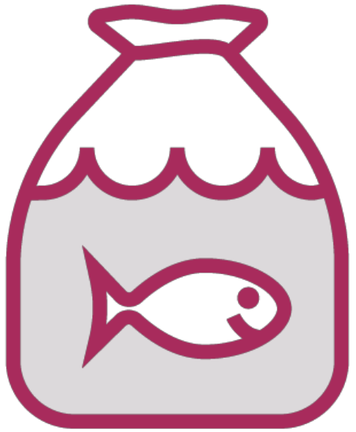
Dimensionality reduction

Whales: Fish or Mammals?



Mammals

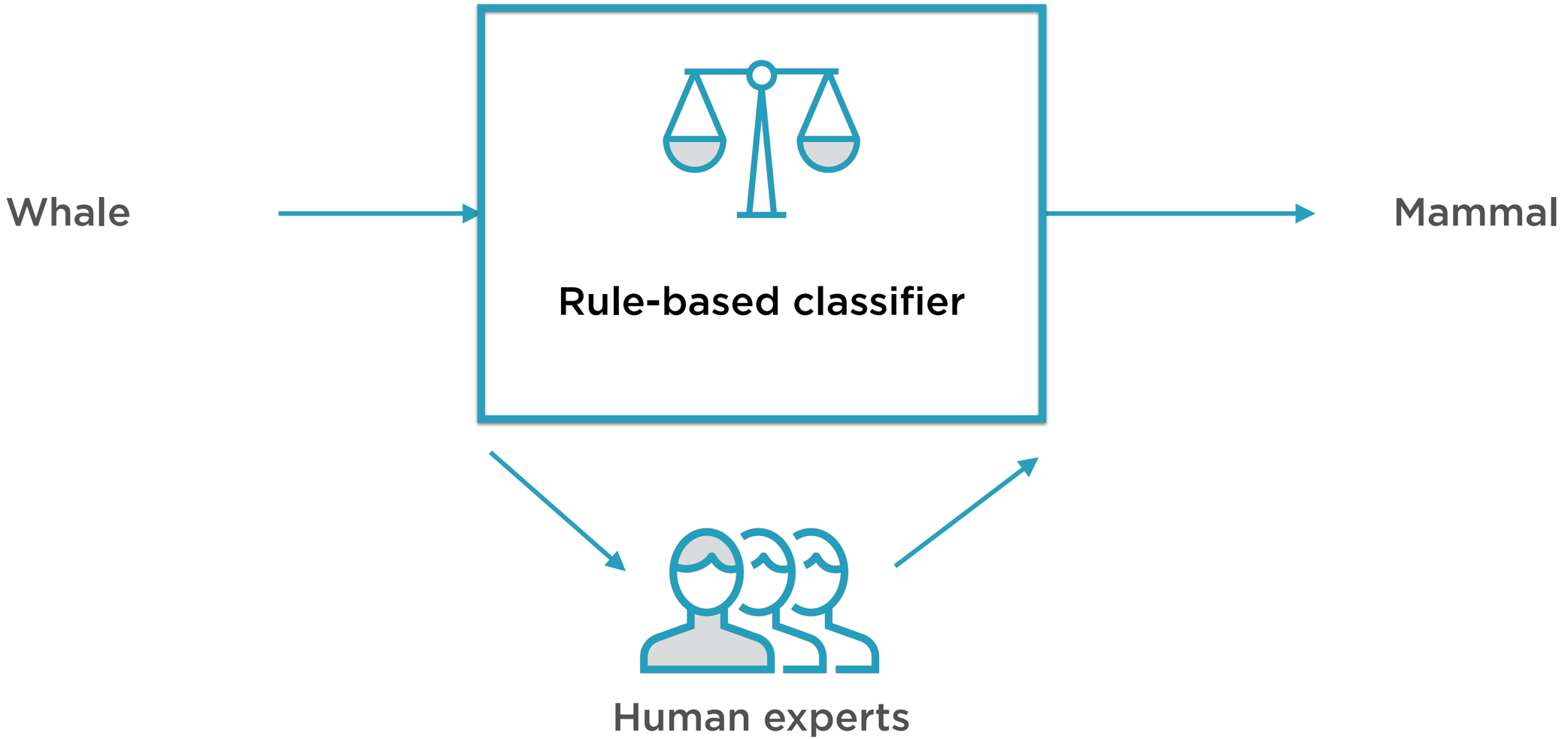
Members of the infraorder *Cetacea*



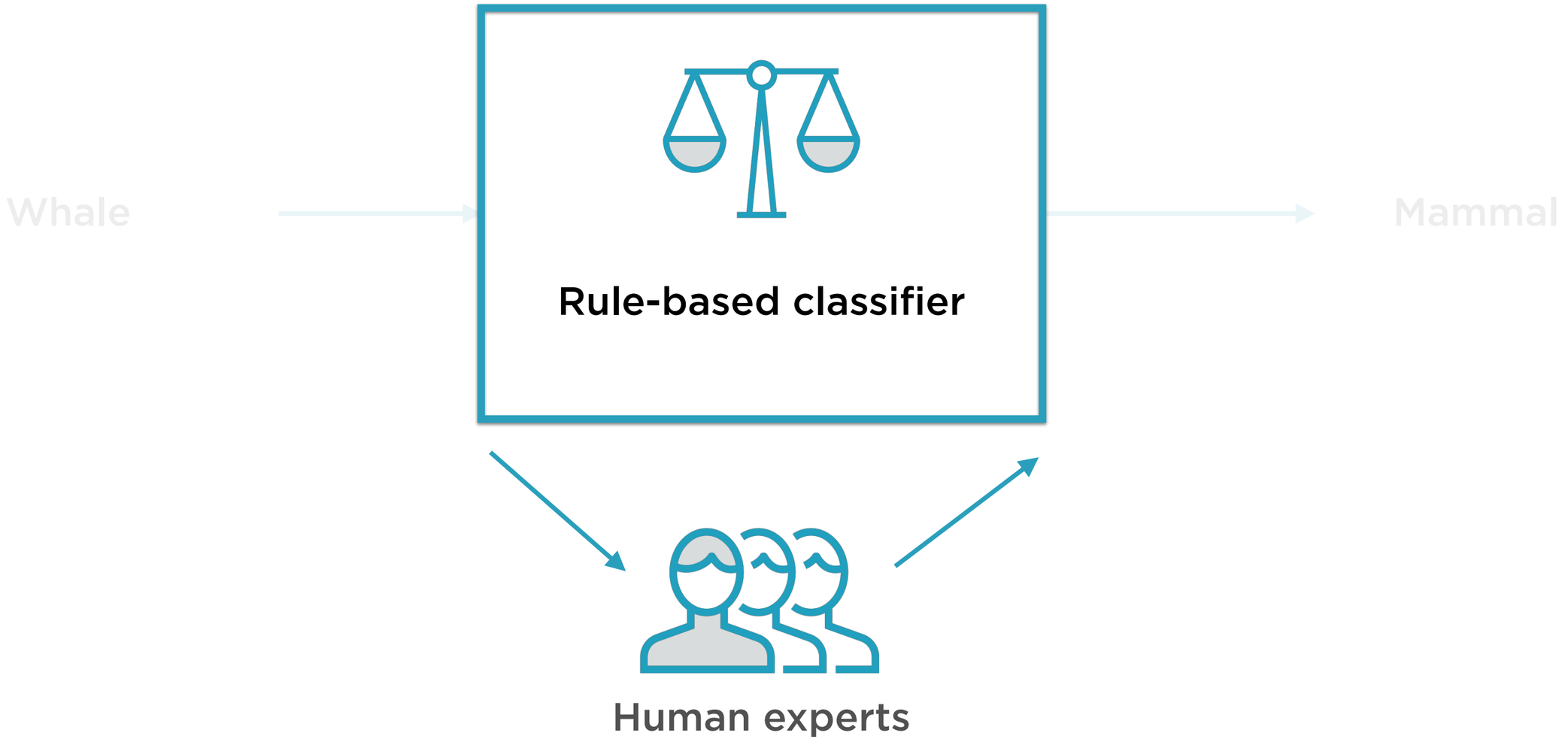
Fish

Look like fish, swim like fish, and move with fish

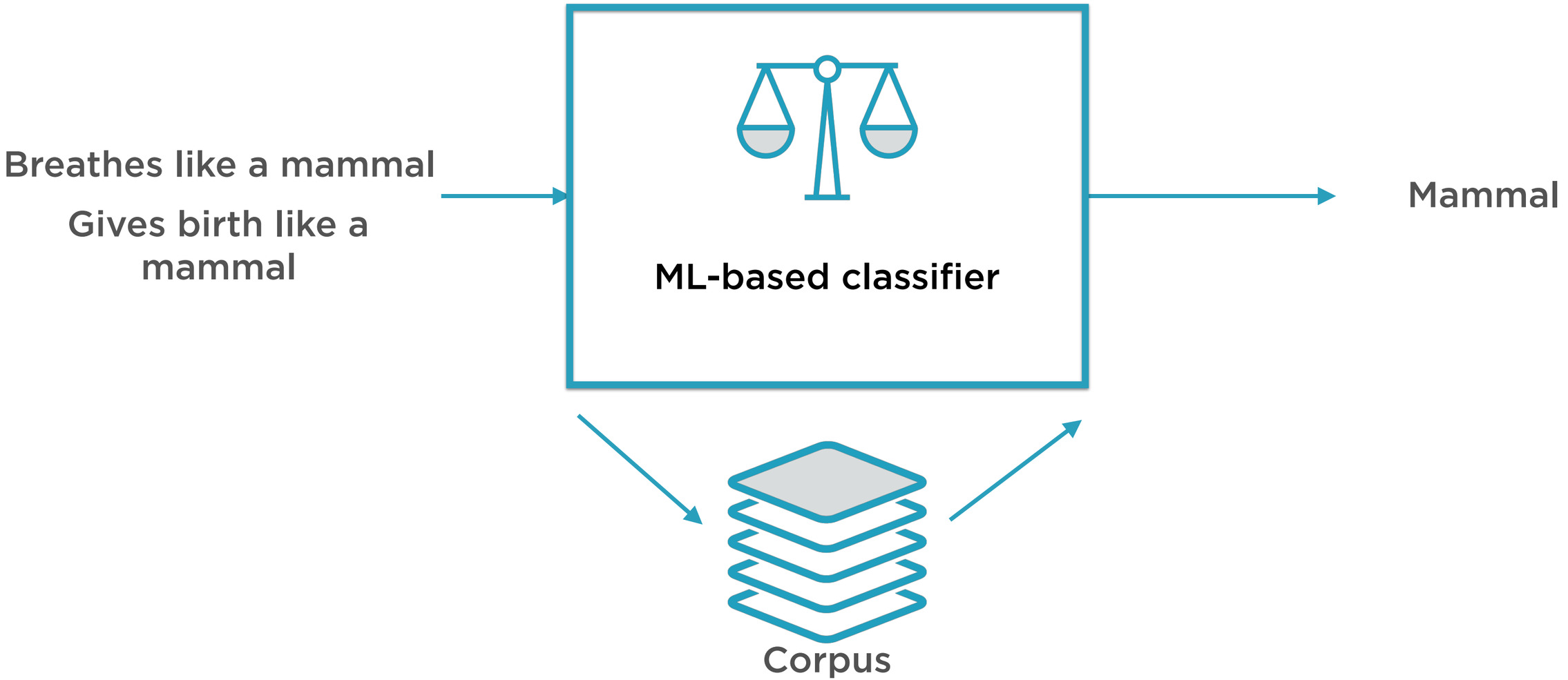
Rule-based Binary Classifier



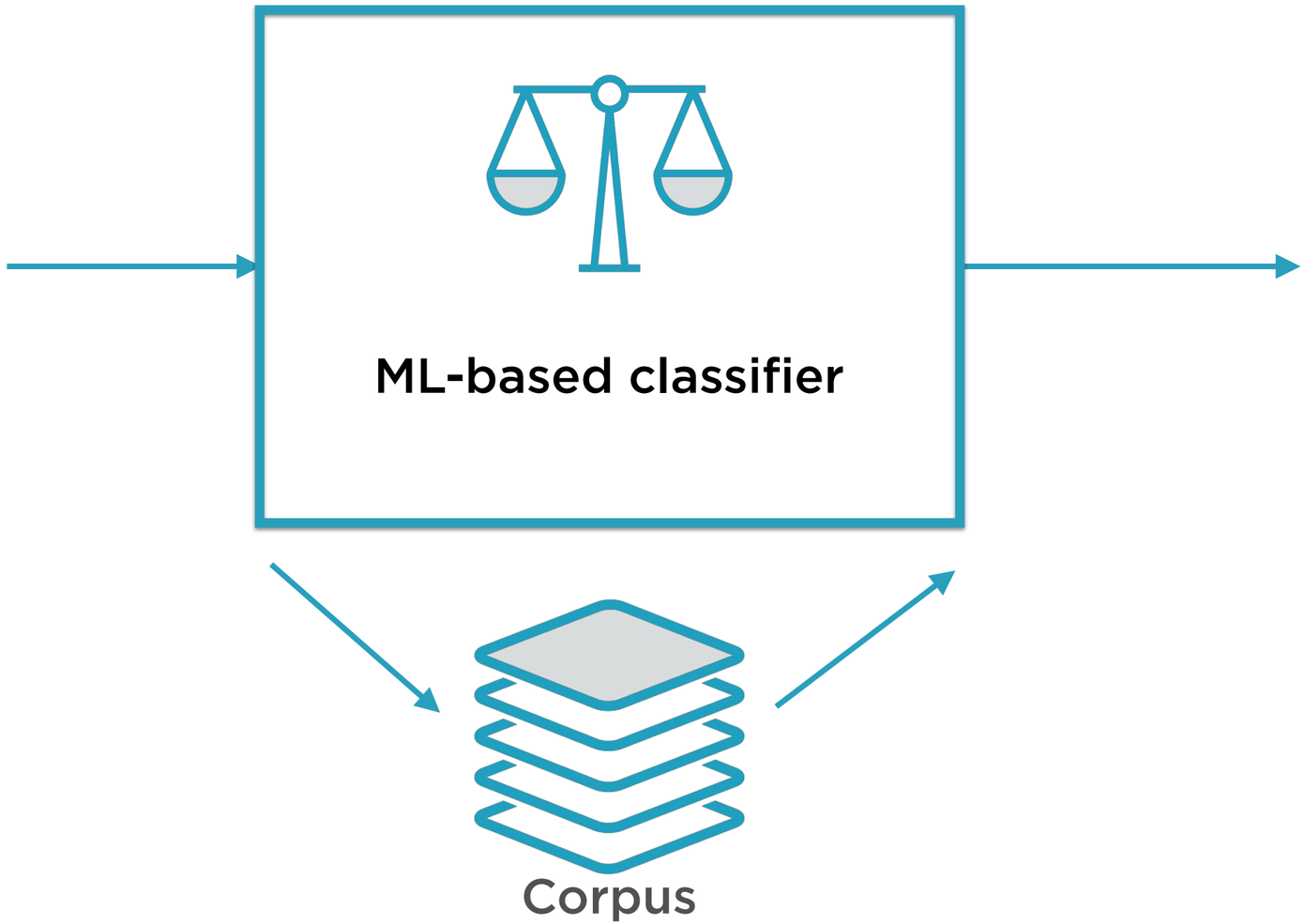
Experts Know What Rules to Apply



ML-based Binary Classifier



Data Used to Train Model Parameters



ML-based Classifier

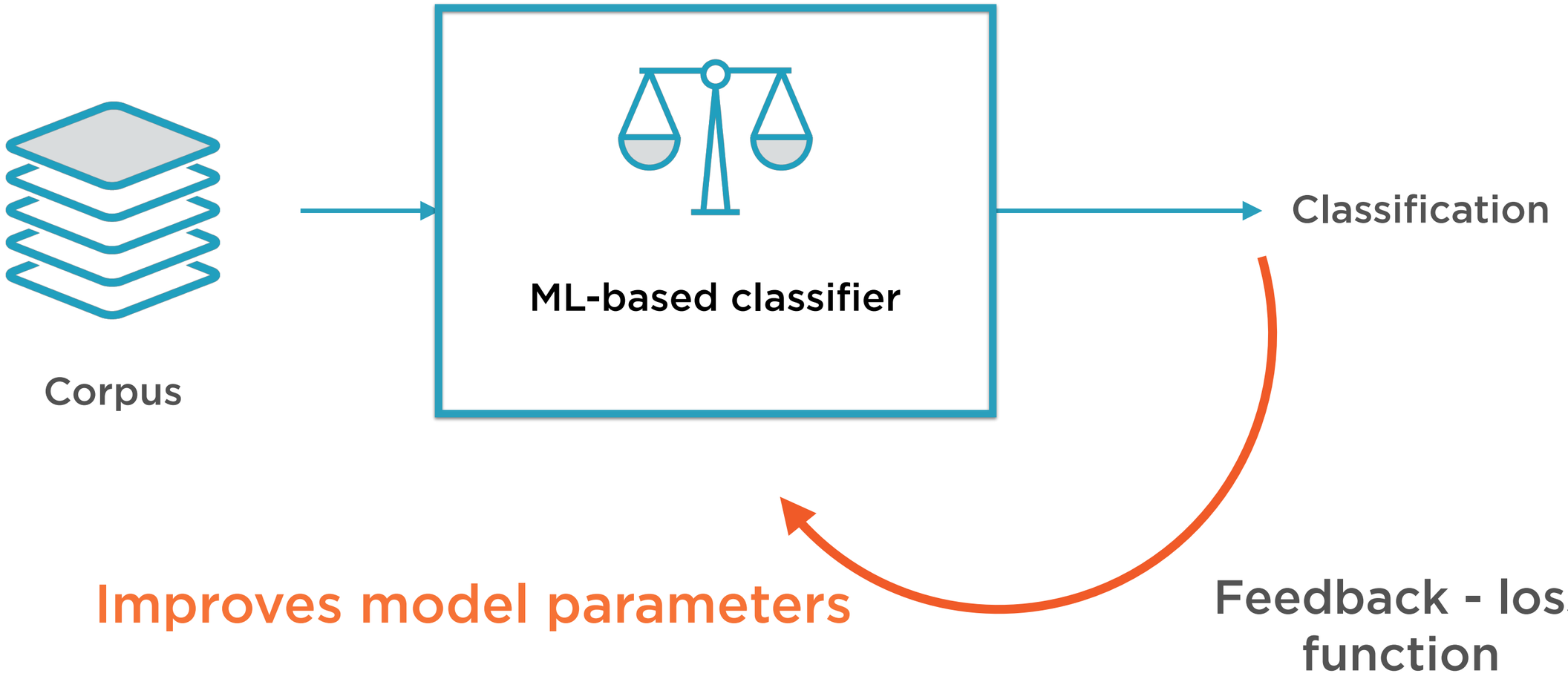
Training

Feed in a large corpus of data classified correctly

Prediction

Use it to classify new instances which it has not seen before

Training the ML-based Classifier



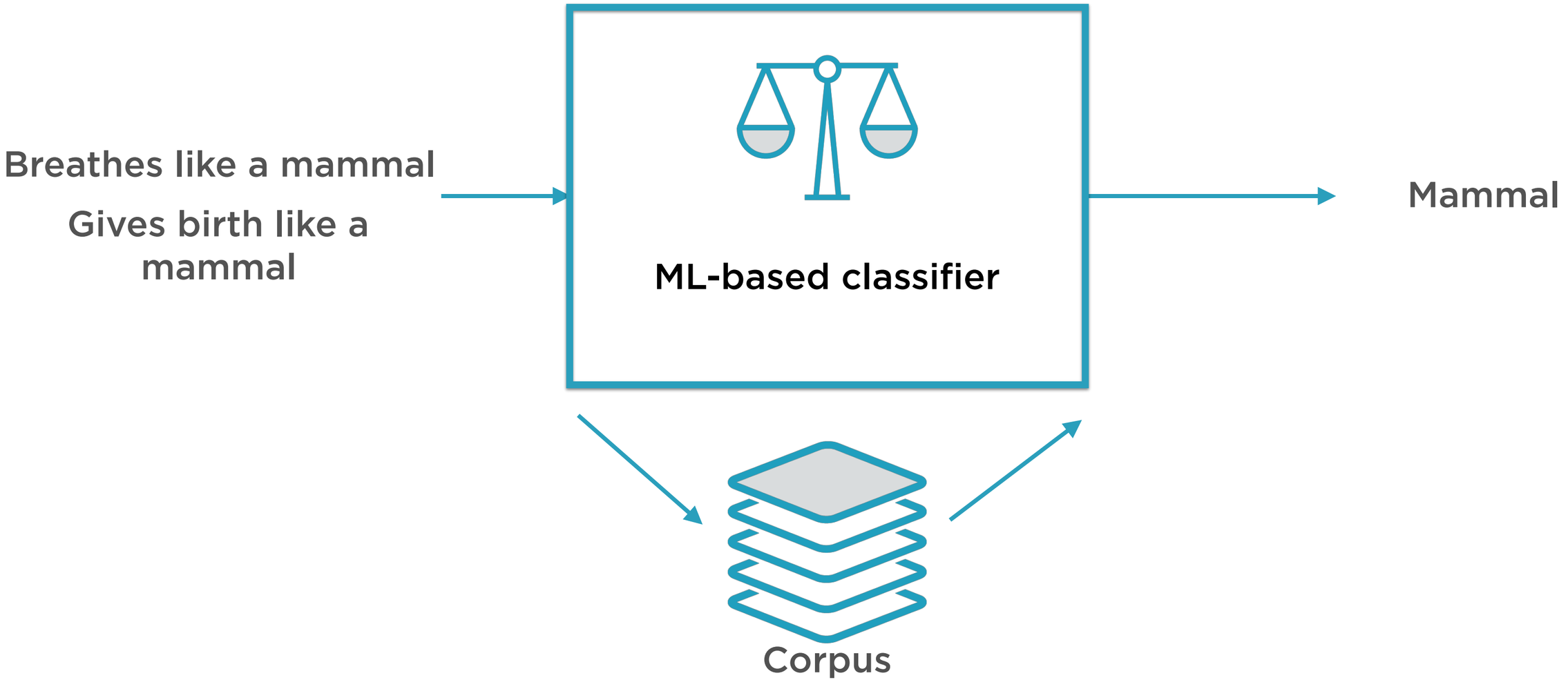
ML-based Binary Classifier

Corpus

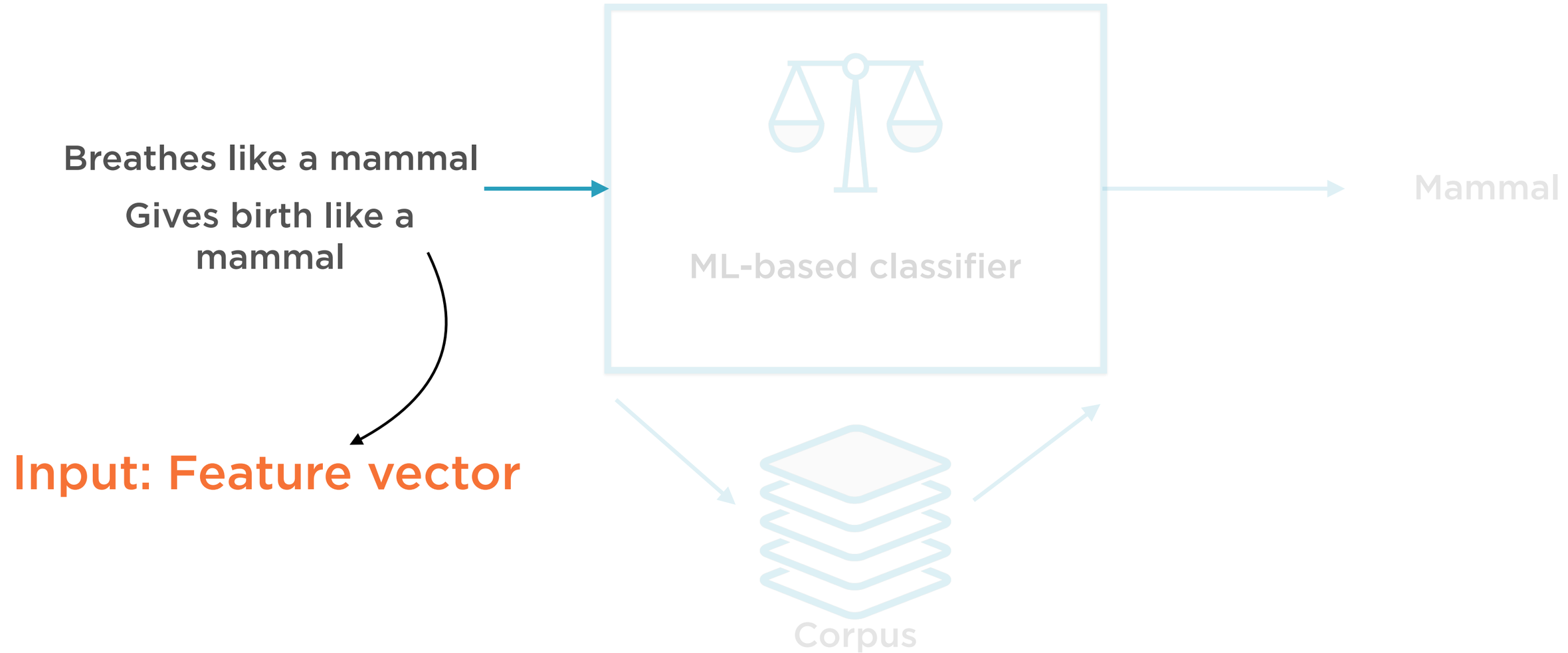
Classification algorithm

ML-based classifier

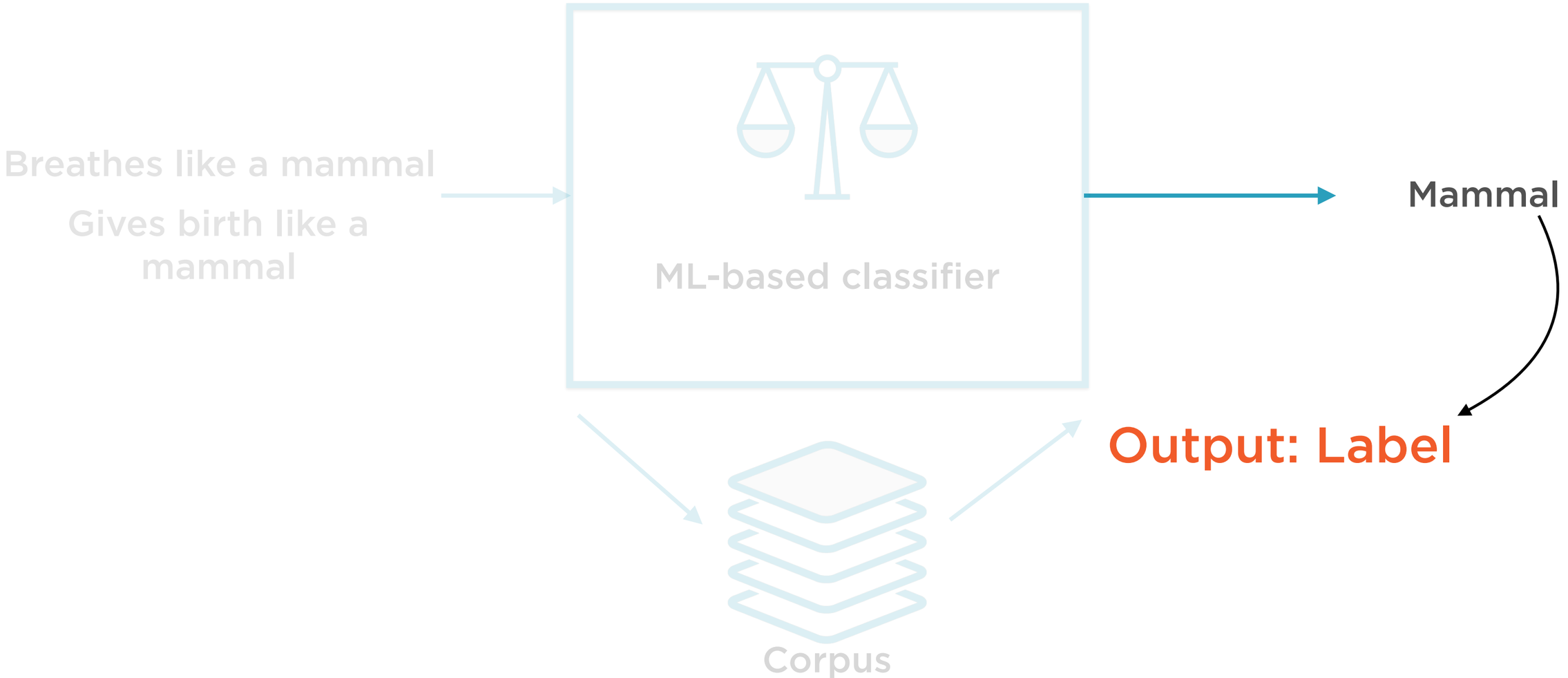
ML-based Binary Classifier



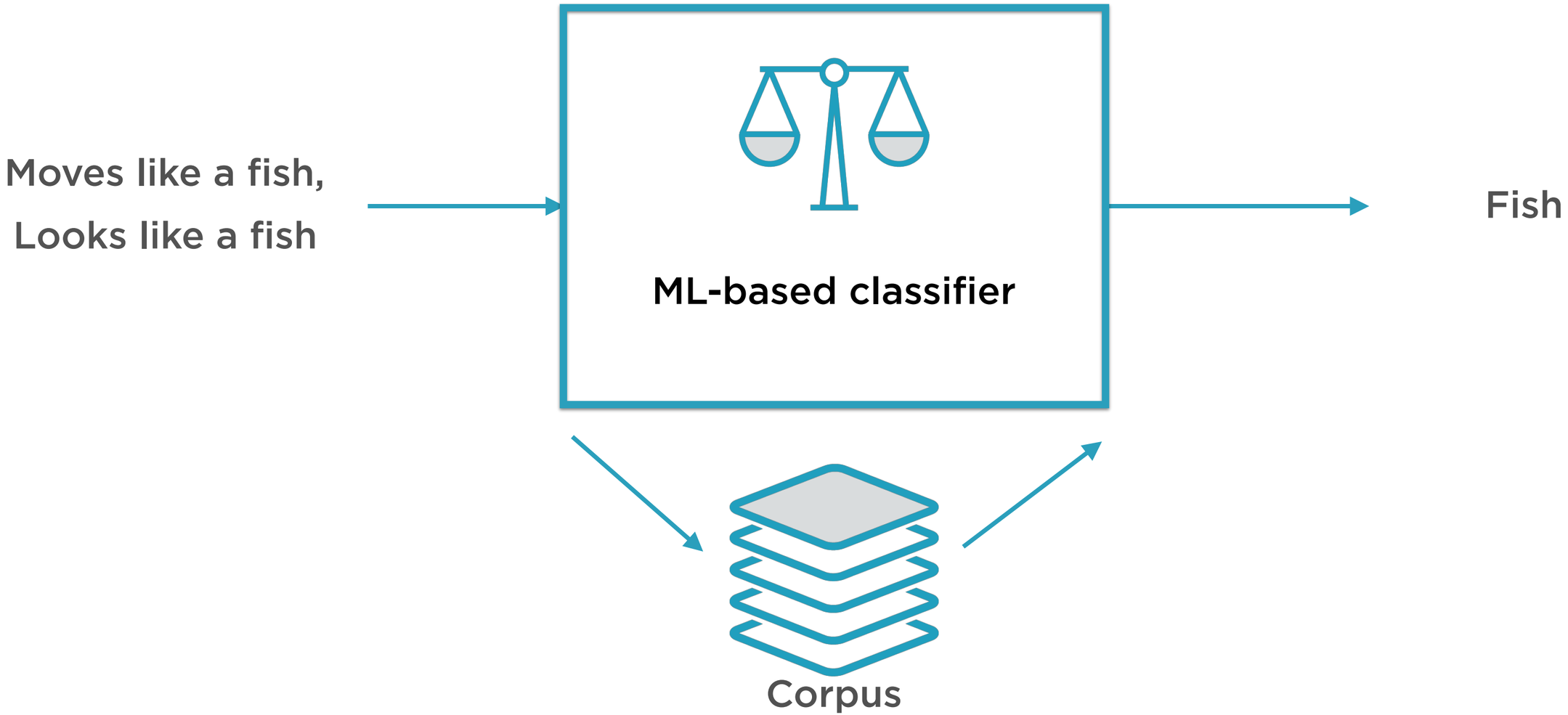
ML-based Binary Classifier



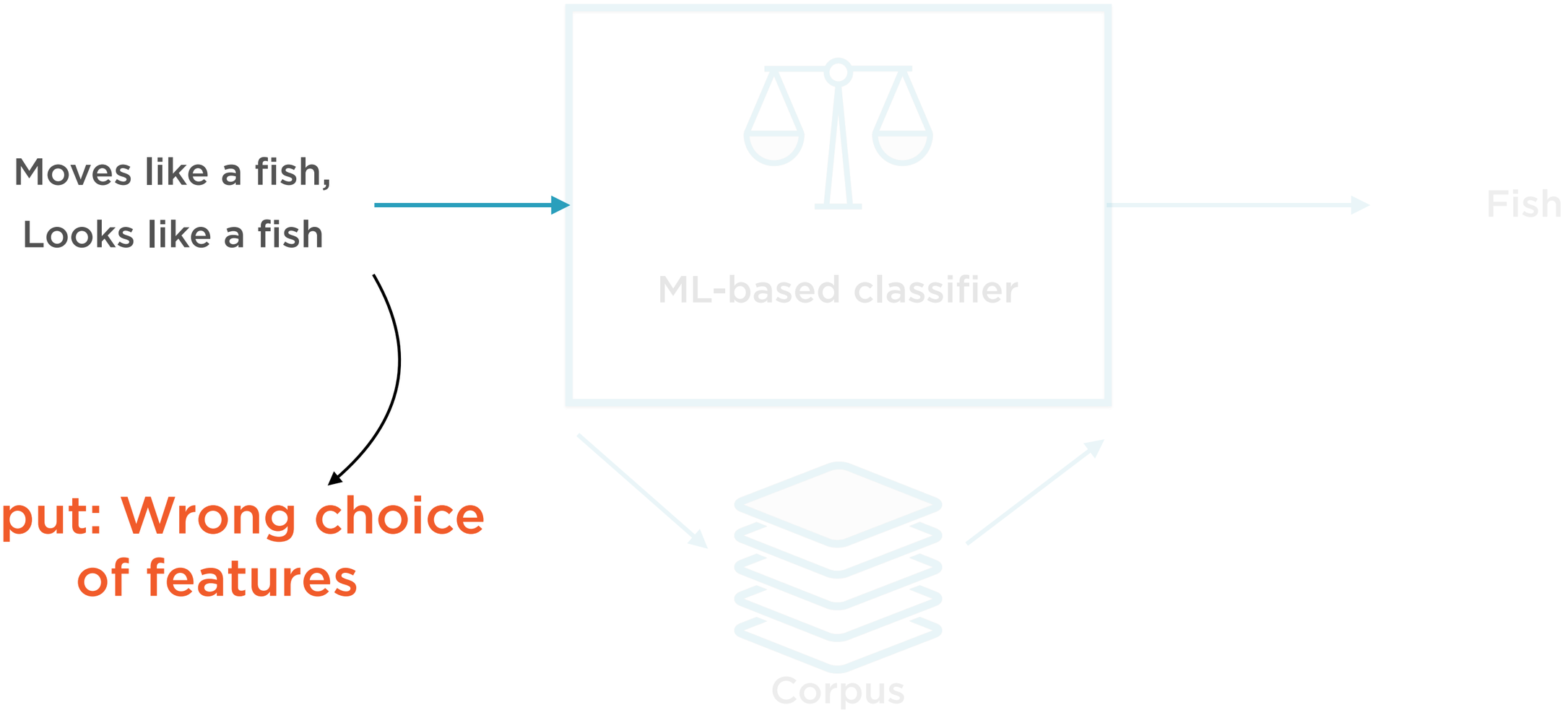
ML-based Binary Classifier



ML-based Binary Classifier

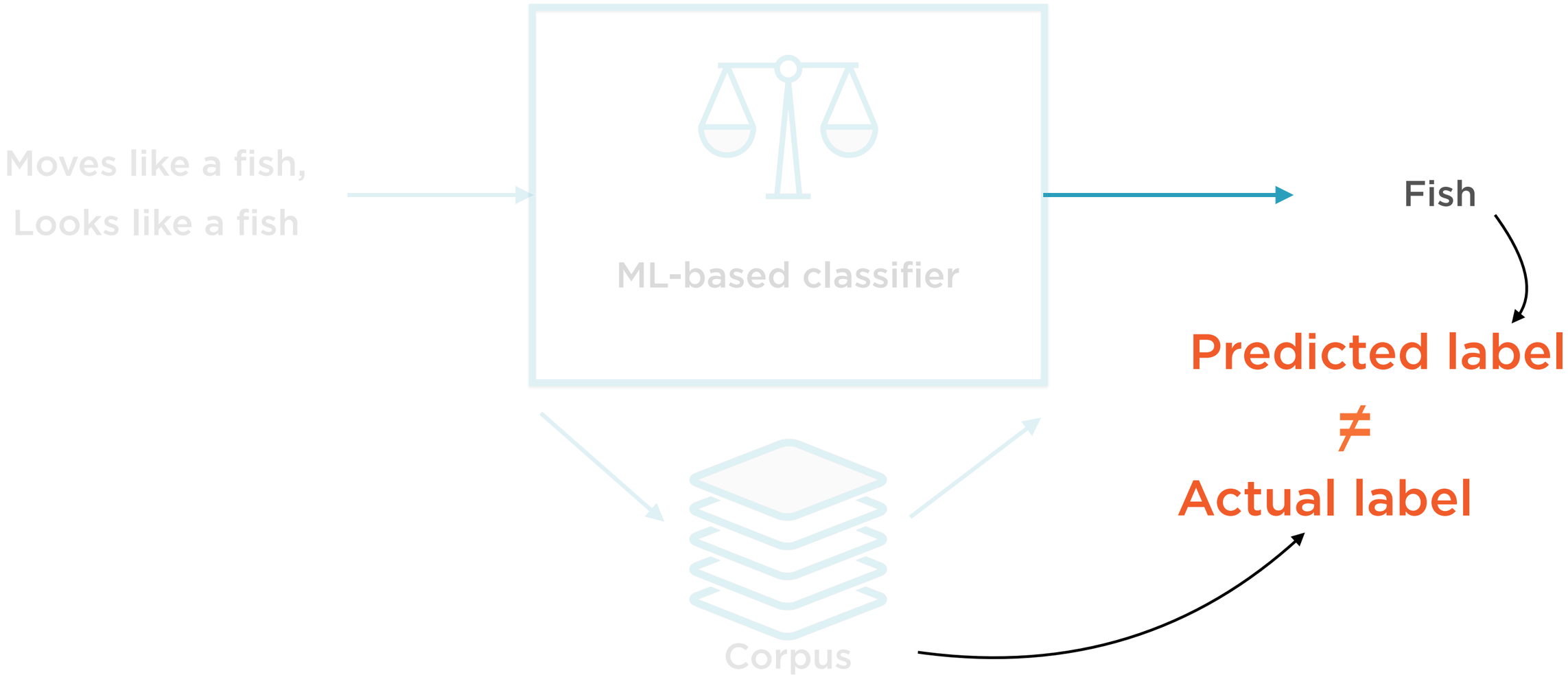


ML-based Binary Classifier



Input: Wrong choice of features

ML-based Binary Classifier



Traditional and Representational Machine Learning

ML-based Binary Classifier

Corpus

Classification algorithm

ML-based classifier

Specific Algorithm Which Learns From Data

Corpus

Classification algorithm

ML-based classifier

Choice of Algorithm Determined by Experts

Corpus

Classification algorithm

ML-based classifier

Features Determined by Experts

Corpus

Classification algorithm

ML-based classifier

Traditional ML Models

Regression models: Linear, Lasso, Ridge, SVR

Classification models: Naive Bayes, SVMs, Decision trees, Random forests

Dimensionality Reduction: Manifold learning, factor analysis

Clustering: K-means, DBSCAN, Spectral clustering

Traditional ML Models

Have a fundamental algorithmic structure to solve problems

The algorithm is fed data which trains the algorithms parameters

Called model parameters

Traditional ML Models

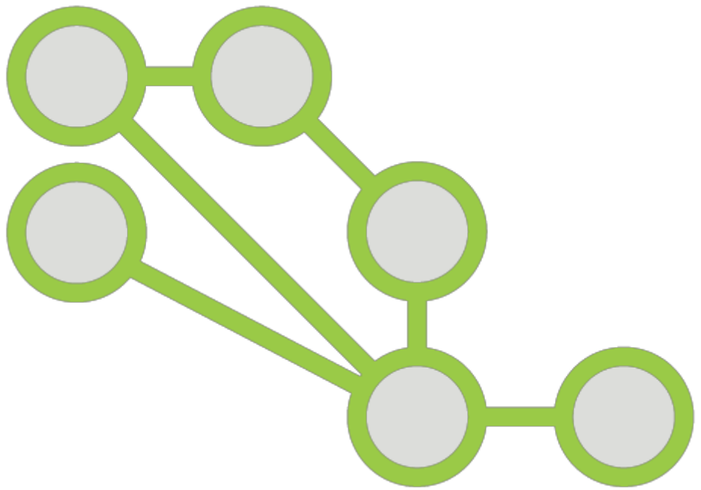
Build a tree structure to classify instances

Fit a line or a curve on data to make predictions

Apply probabilities on input data to get output probabilities

“Traditional” ML-based systems rely on experts to decide what features to pay attention to - and how

Representation ML Models



Also used to solve classification, regression, clustering, and dimensionality reduction

Learn significant features from the underlying data

Deep learning models such as neural networks

“Representation” ML-based systems figure out by themselves what features to pay attention to - and how

What Is a Neural Network?

Deep Learning

Algorithms that learn what features matter

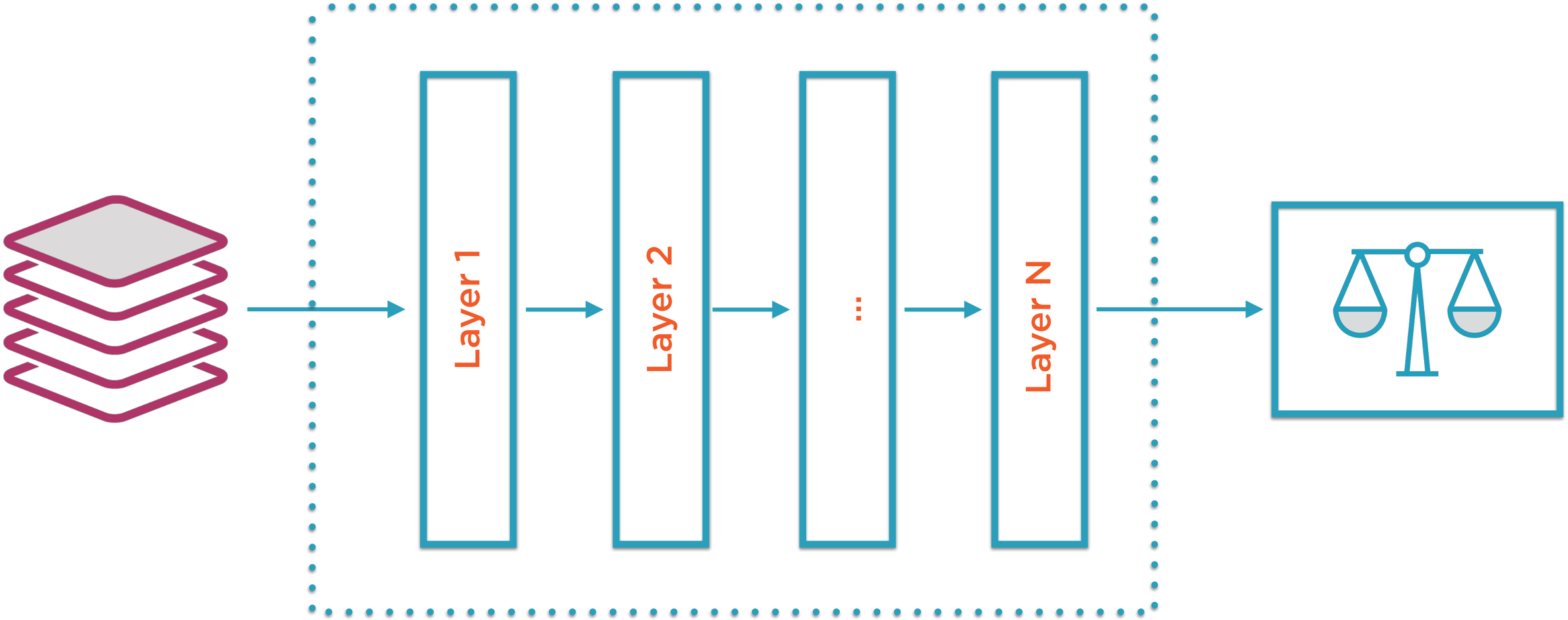
Neural Networks

The most common class of deep learning algorithms

Neurons

Simple building blocks that actually “learn”

Neural Networks

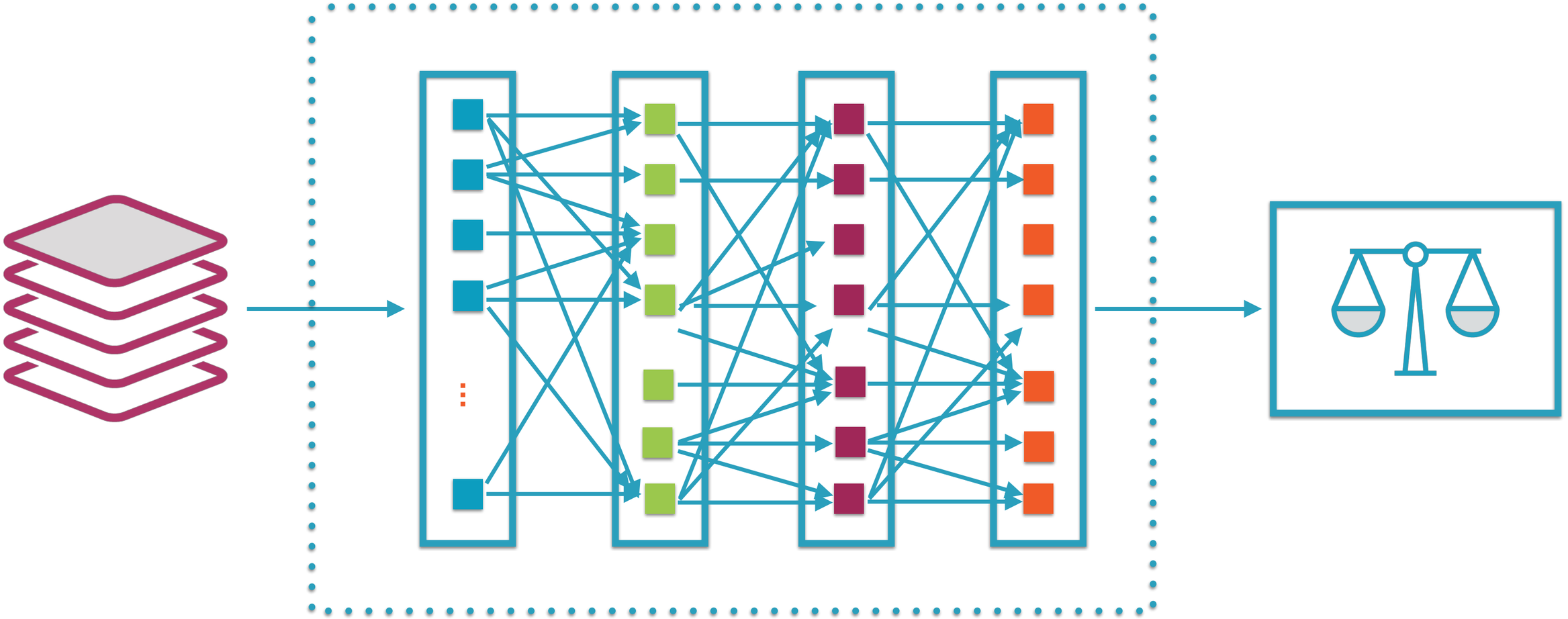


Corpus

Layers in a neural network

ML-based classifier

Neural Networks

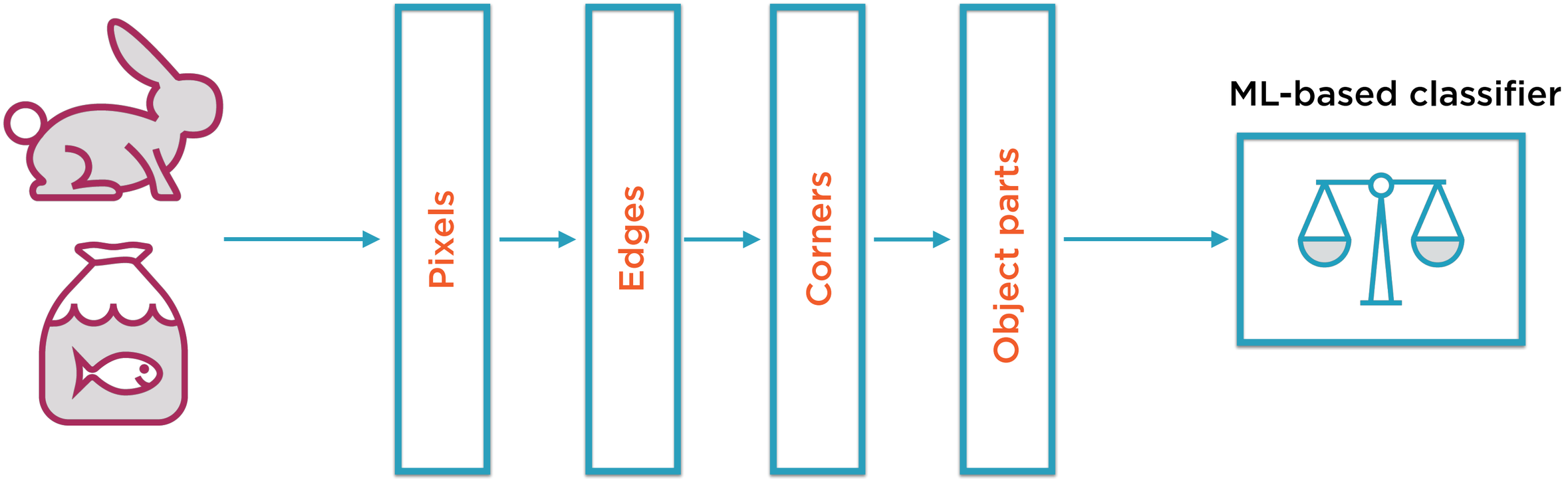


Corpus

Each layer consists of individual interconnected neurons

ML-based classifier

Each Layer Extracts Information from Data



Traditional vs. Deep Learning Models

Traditional ML Models

Features used in models explicitly chosen by domain experts

Structured data such as numbers and probabilities

Classification, regression, clustering, and dimensionality reduction

Deep Learning ML Models

Features used in models implicitly chosen by model itself

Unstructured data such as images and movies

Classification, regression, clustering, and dimensionality reduction

Traditional vs. Deep Learning Models

Traditional ML Models

Wide range of problem-specific solution techniques

Each solution technique adopts characteristic approach

User has more insight into mechanics and internals of models

scikit-learn

Deep Learning ML Models

Neural networks by far the most common solution technique

All solution techniques rely on neurons and interconnections between them

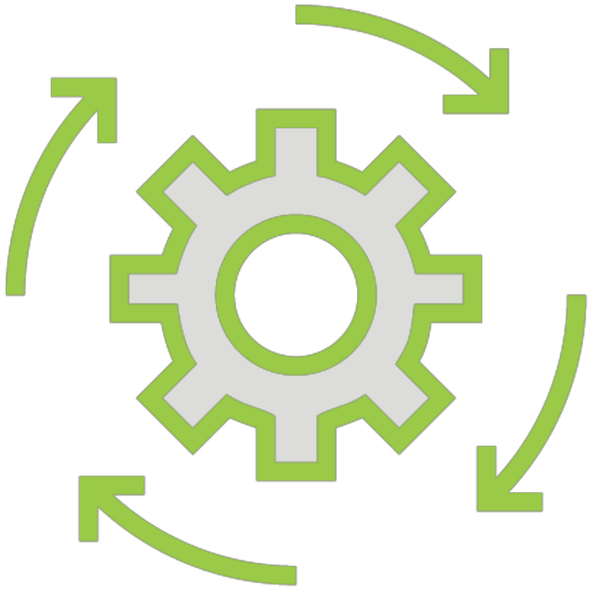
Black-box models that are hard to question or reverse-engineer

TensorFlow, Keras, PyTorch

The Niche of scikit-learn in Machine Learning

scikit-learn - easy-to-use, very comprehensive and efficient Python library for **traditional** ML models

scikit-learn



Developed as a Google summer of code project in 2007

Currently has 30+ active contributors

Sponsored by INRIA, Google, Tinyclues, and the Python Software Foundation

Attractions of scikit-learn

Easy-to-use

Comprehensive

Efficient

Attractions of scikit-learn

Easy-to-use

Comprehensive

Efficient

Ease of Use

- Estimator API for consistent interface
- Estimators for all kinds of models
- Create a model object
- Fit to training data
- Predict for new data
- Pipelines for complex operations

Attractions of scikit-learn

Easy-to-use

Comprehensive

Efficient

Support for Complete ML Workflow

All common families of models supported

Data pre-processing, cleaning, feature selection, and extraction

Model validation and evaluation

Completeness

Regression, classification, clustering, dimensionality reduction

Feature extraction and selection using statistical and dimensionality reduction

Data pre-processing

Data generation

- Swiss rolls, S-curves

Cross-validation to evaluate models

Hyperparameter tuning

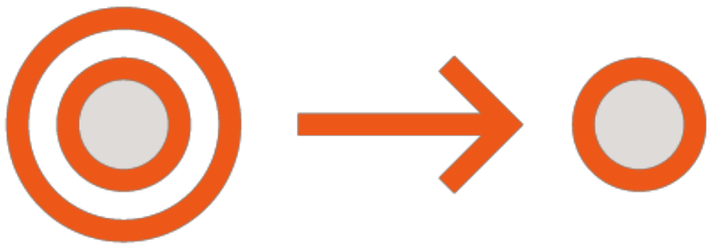
Attractions of scikit-learn

Easy-to-use

Comprehensive

Efficient

Efficiency

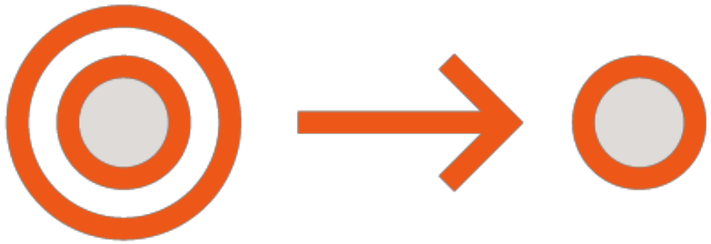


Highly optimized implementations

Built on SciPy, hence scikit prefix

Interoperates with all common Python libraries for data science

Efficiency



NumPy: Base n-dimensional array package

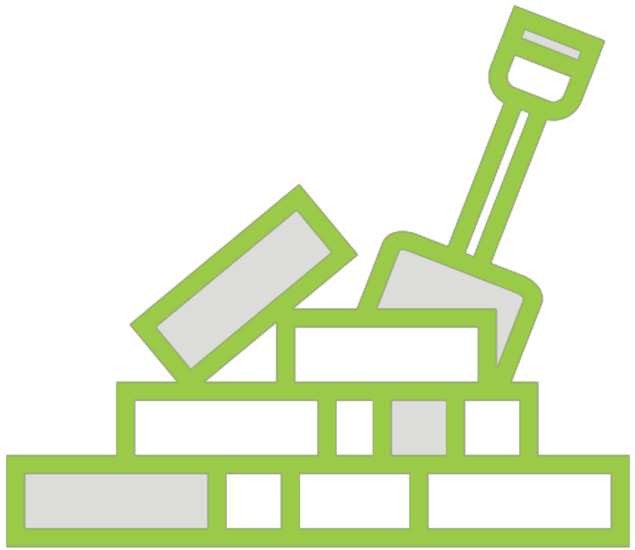
SciPy: Fundamental library for scientific computing

Matplotlib: Comprehensive 2D/3D plotting

Sympy: Symbolic mathematics

Pandas: Data structures and analysis

Foundational Libraries for scikit-learn



NumPy: Base n-dimensional array package

SciPy: Fundamental library for scientific computing

Matplotlib: Comprehensive 2D/3D plotting

Sympy: Symbolic mathematics

Pandas: Data structures and analysis

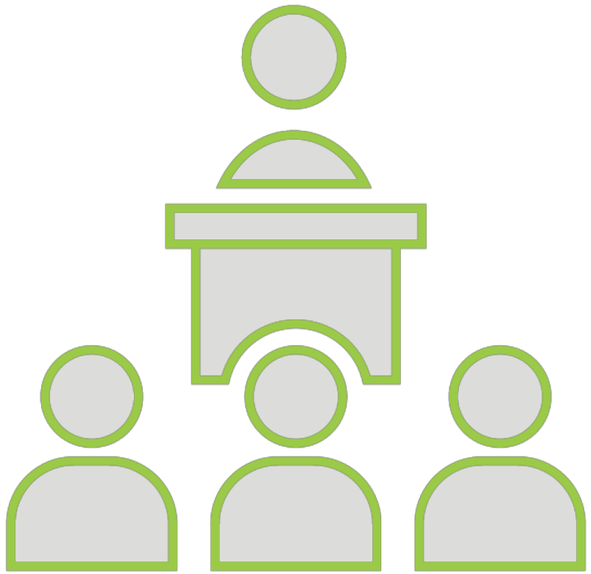
Demo

Exploring the scikit-learn webpage

Exploring documentation, packages, and libraries

Supervised and Unsupervised Learning

Types of ML Algorithms



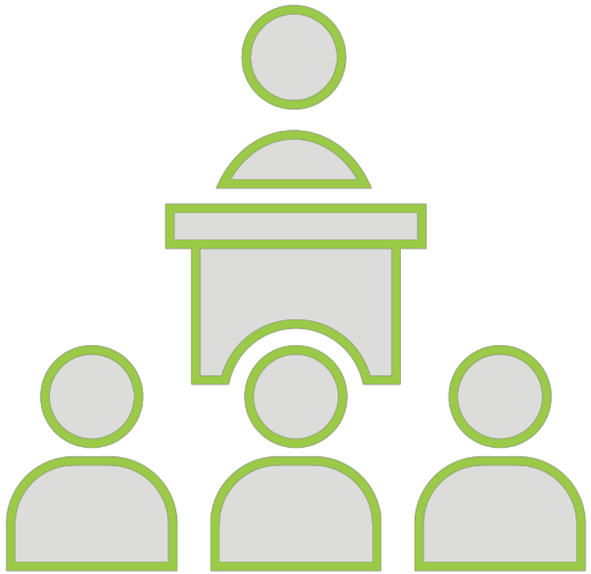
Supervised

Labels associated with the training data is used to correct the algorithm

Unsupervised

The model has to be set up right to learn structure in the data

Types of ML Algorithms



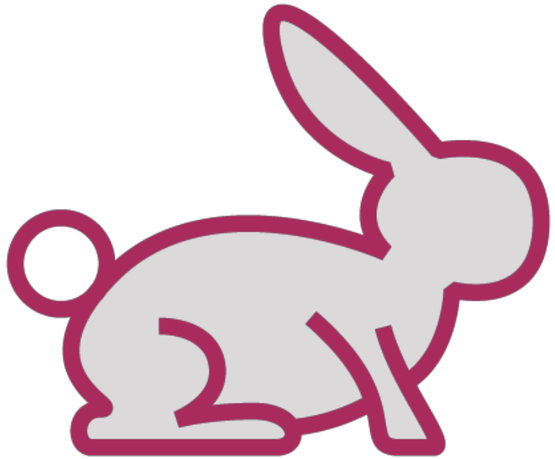
Supervised

Labels associated with the training data is used to correct the algorithm

Unsupervised

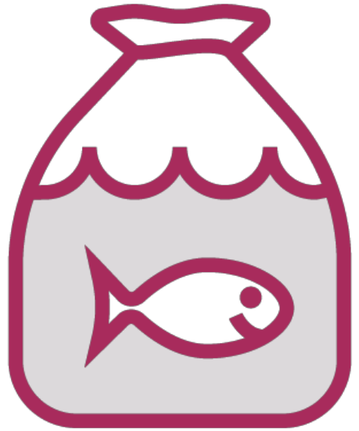
The model has to be set up right to learn structure in the data

Whales: Fish or Mammals?



Mammals

Members of the infraorder *Cetacea*



Fish

Look like fish, swim like fish, move with fish

Whales: Fish or Mammals?

ML-based Classifier

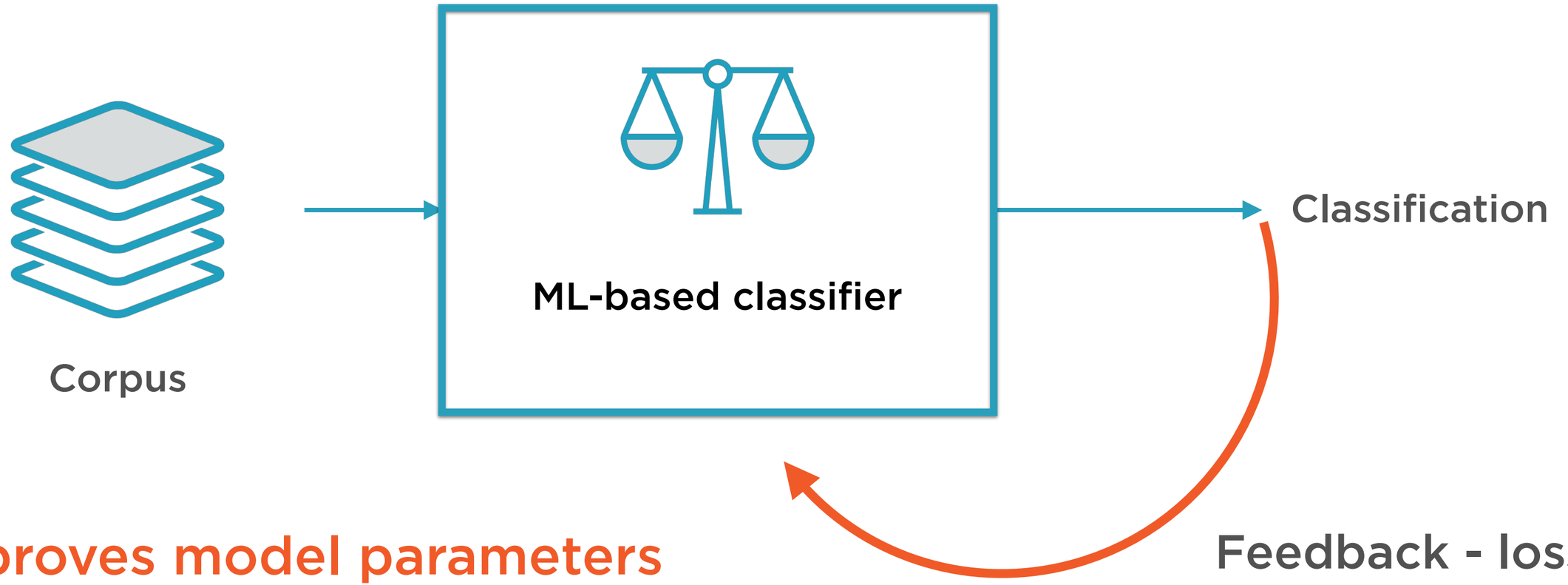
Training

Feed in a large corpus of data classified correctly

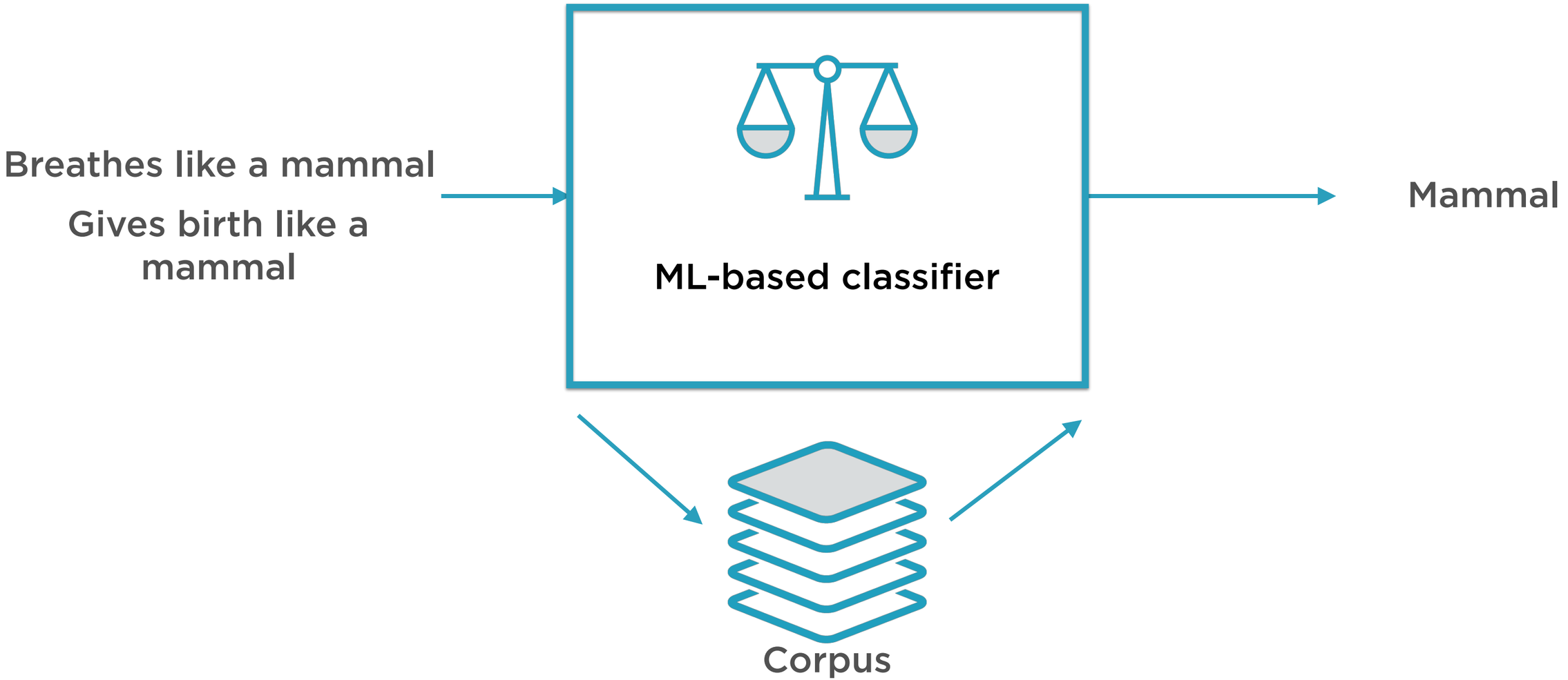
Prediction

Use it to classify new instances which it has not seen before

Training the ML-based Classifier



ML-based Binary Classifier



x Variables

The attributes that the ML algorithm focuses on are called features

Each data point is a list or vector of such features

Thus, the input into an ML algorithm is a feature vector

Feature vectors are usually called the x variables

y Variables

The attributes that the ML algorithm tries to predict are called labels

Types of labels

- **Categorical (classification)**
- **Continuous (regression)**

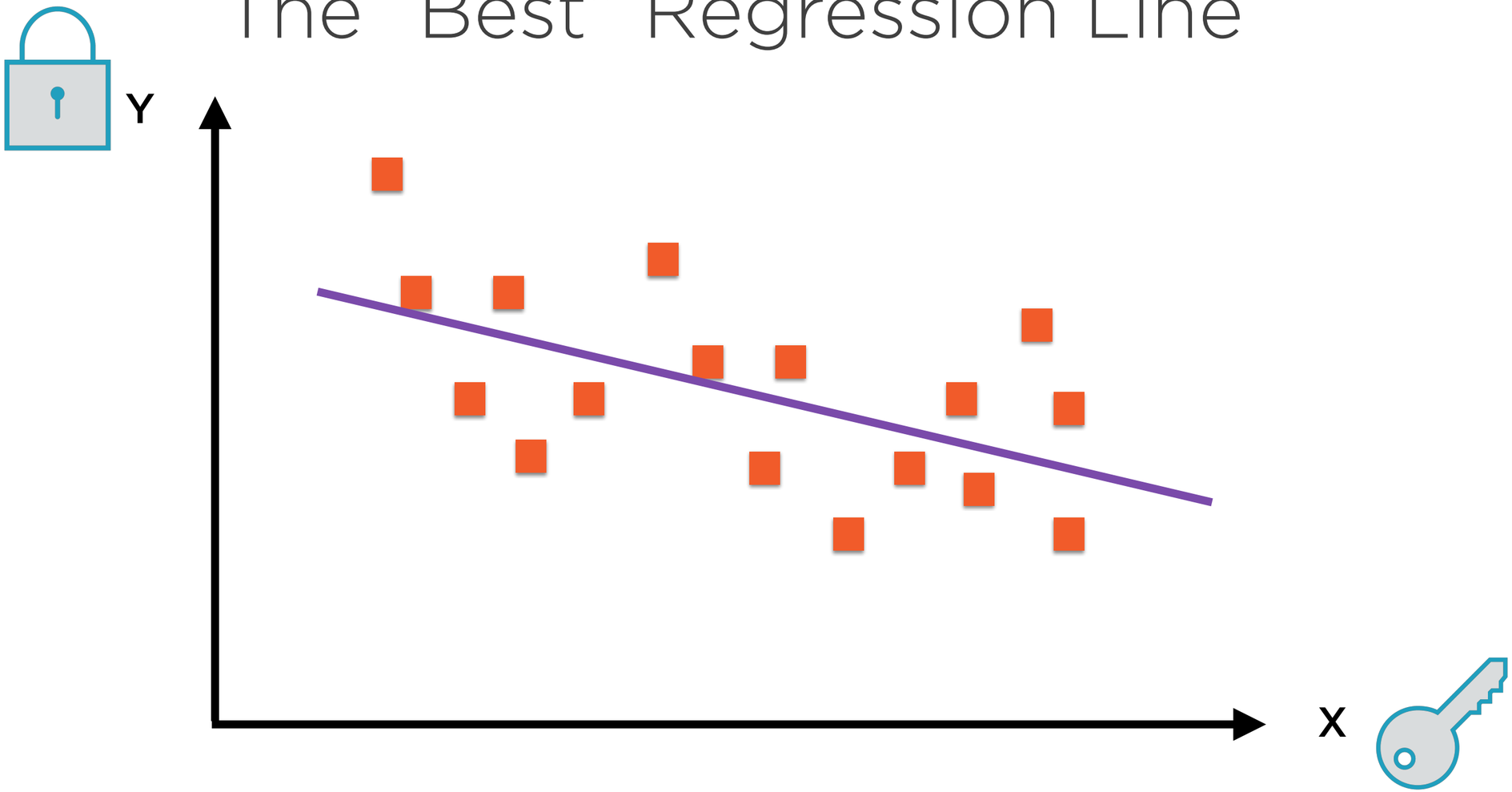
Labels are usually called the y variables

```
y = f(x)
```

Supervised Machine Learning

Most machine learning algorithms seek to “learn” the function f that links the features and the labels

The “Best” Regression Line



Linear regression involves finding the “best fit” line via a training process

```
y = Wx + b
```

$$f(x) = Wx + b$$

Linear regression specifies, up-front, that the function f is linear

```
def doSomethingReallyComplicated(x1,x2…):
   …
   …
   …
   return complicatedResult
```

---

f(x) = doSomethingReallyComplicated(x)

ML algorithms such as neural network can “learn” (reverse-engineer) pretty much anything given the right training data

Types of ML Algorithms

Supervised

Labels associated with the training data is used to correct the algorithm

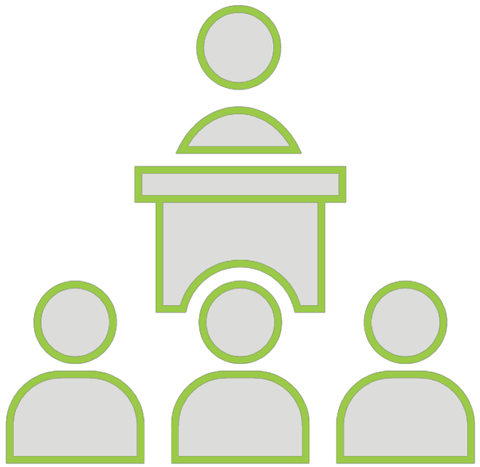
Unsupervised

The model has to be set up right to learn structure in the data

Unsupervised learning does **not** have:

- **y** variables
- **A labeled** corpus

Supervised Learning



Input variable x and output variable y

Learn the mapping function y = f(x)

Approximate the mapping function so for new values of x we can predict y

Use existing dataset to correct our mapping function approximation

Unsupervised Learning

Only have input data x – no output data

Model the underlying structure to learn more about data

Algorithms self-discover the patterns and structure in the data

Unsupervised ML Algorithms

Clustering

Identify patterns in data items e.g. K-means clustering

Dimensionality Reduction

Identify significant factors that drive data e.g. PCA

Demo

Installing scikit-learn and its dependencies on your local machine

Summary

scikit-learn for data and ML modeling

Relationship with NumPy, SciPy, Pandas, and Matplotlib

Algorithms for supervised and unsupervised learning

Contrast with TensorFlow, Keras, and other deep learning frameworks