INTRODUCTION TO MACHINE LEARNING

LECTURE 1
SECTION 1
JUNE 2ND

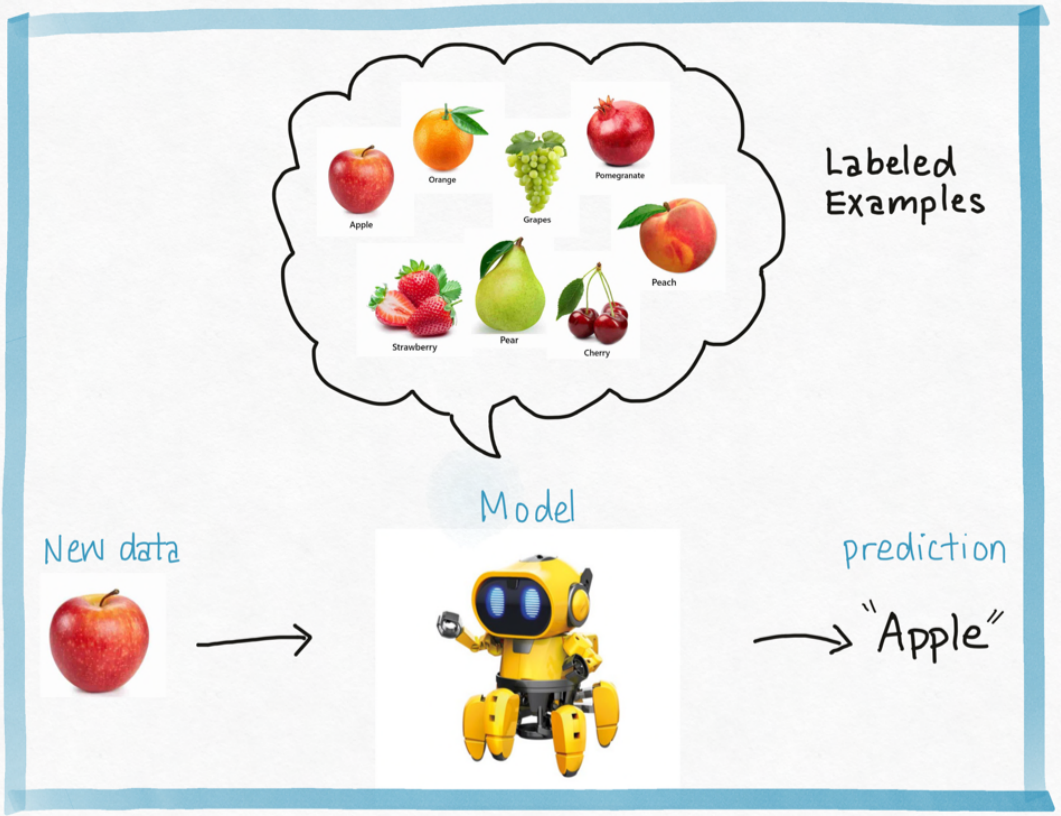
INSTITUTE FOR APPLIED
COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

UNIVERSITY *of*
RWANDA

DEFINING MODELS

WHAT IS A MACHINE LEARNING MODEL?

A machine learning model extracts patterns from data.



If data given to the model is labeled, and the task is to predict labels for new labels then this is a:

prediction problem

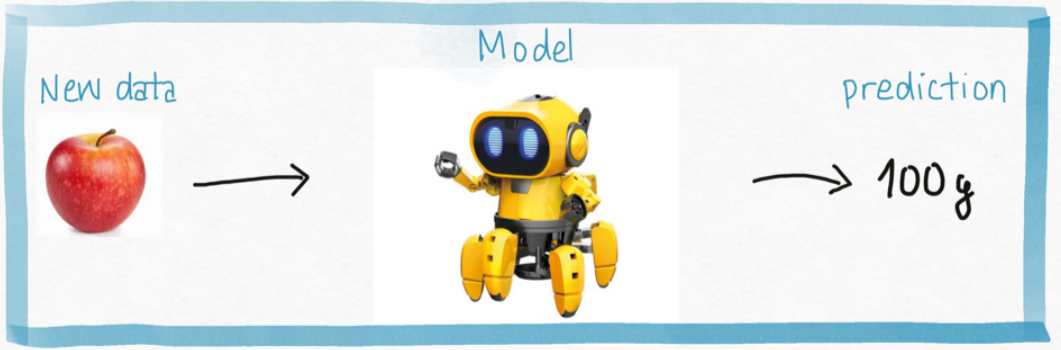
we also call this a:

supervised learning problem

CLASSIFICATION V.S. REGRESSION:

If the label is a category then this is a:

Classification problem

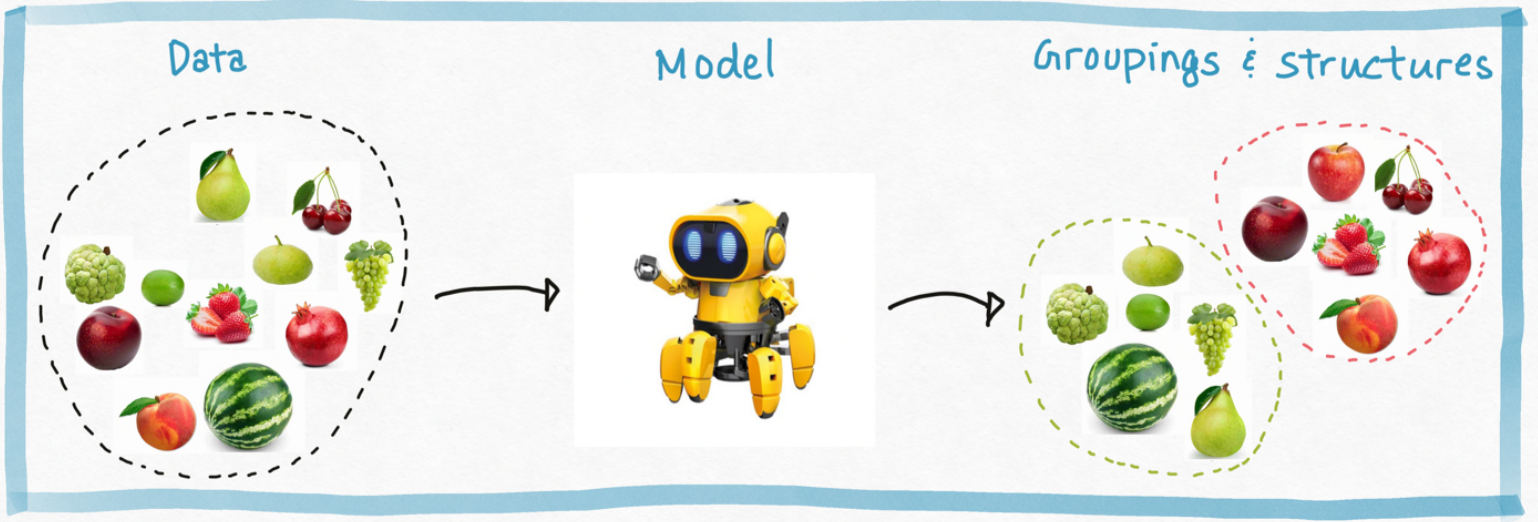


If the label is a real number then this is a:

regression problem

WHAT IS A MACHINE LEARNING MODEL?

A machine learning model extracts patterns from data.



If the data given to the model is unlabeled and the task is to find groupings and structures in the data, we call this a:

unsupervised learning problem

MODELS AS FUNCTIONS:

A model for supervised learning is a mathematical function f_w that takes in an input x, of D number of measurements

$$x = (x_1, \dots x_D) \in \mathbb{R}^D$$

and outputs a prediction

$$\hat{y} = f_w(x)$$

w are the parameters of the function f_w.

We call the measurements $x_1, \dots, x_D$ predictors, or covariates.

Example: Let x_1 be size of house (m^2) and x_2 be distance to city center (m); let $\hat{y}$ be selling price for the house, then

$$\hat{y} = w_1 x_1 - w_2 x_2 + w_3$$

is a model to predict the selling price of a house.

MODEL TRAINING

MODEL TRAINING:

In a supervised learning problem, we provide data with lables to the model, f_w,

$$D = \{(x^{(1)}, y^{(1)}), \dots (x^{(N)}, y^{(N)})\}$$

each $y^{(n)}$ is the true label for the covariates $x^{(n)} = (x_1^{(n)}, \dots, x_D^{(n)})$.
We call y the target, and we call D the training set.

We learn the best parameters w for f_w using the training data.
We call this process model training, model fitting or inference.

Example: For the housing price model $f_w(x) = w_1x_1 + w_2x_2 + w_3$
To train the model, we need to provide a training set of houses including the size, the distance to city center and the actual selling price:

x_1	x_2	y
$70\,m^2$	2,000m	\$500k
$90m^2$	10k m	\$800k
$65\,m^2$	60km	\$300k
⋮	⋮	⋮

MODEL EVALUATION

MODEL EVALUATION:

After fitting the model, f_w, we evaluate our model by comparing its predictions $\hat{y} = f_w(x)$ to the target or true label y on the training data D.

But we also need to test our model on new data that was not used in the training!

Training evaluation

Testing yourself on materials you studied from. You expect to do well if studied hard!

Test evaluation

Testing yourself on questions you haven't seen before. Your performance maybe worse!

TRAIN TEST SPLIT:

So we know to evaluate our models on new, test data.

But new data maybe hard or impossible to get:

- clinical trials
- expensive experiments
- data was historical

We must simulate getting new data by holding out a portion of the data we are given, and not use it for model training.

After fitting the model, we evaluate it on this hold-out data.

This hold-out procedure is called the train-test split.