Introduction

Lecture notes by Ethem Alpaydın
Introduction to Machine Learning (Boğaziçi Üniversitesi)

Lecture notes by Kevyn Collins-Thompson
Applied Machine Learning (Coursera)

Lecture notes by Andrew NG
Machine Learning by Stanford University (Coursera)

What is Machine Learning?

- The study of computer programs (algorithms) that can learn by example
- ML algorithms can generalize from existing examples of a task

Machine Learning Definition

Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

Machine Learning Definition

Tom Mitchell (1998) Well-posed Learning Problem: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?

A. Classifying emails as spam or not spam.

B. Watching you label emails as spam or not spam.

C. The number (or fraction) of emails correctly classified as spam/not spam.

D. None of the above—this is not a machine learning problem.

Machine Learning brings together statistics, computer science, and more..

Statistical methods

– *Infer conclusions from data*

– *Estimate reliability of predictions*

Computer science

– *Large-scale computing architectures*

– *Algorithms for capturing, manipulating, indexing, combining, retrieving and performing predictions on data*

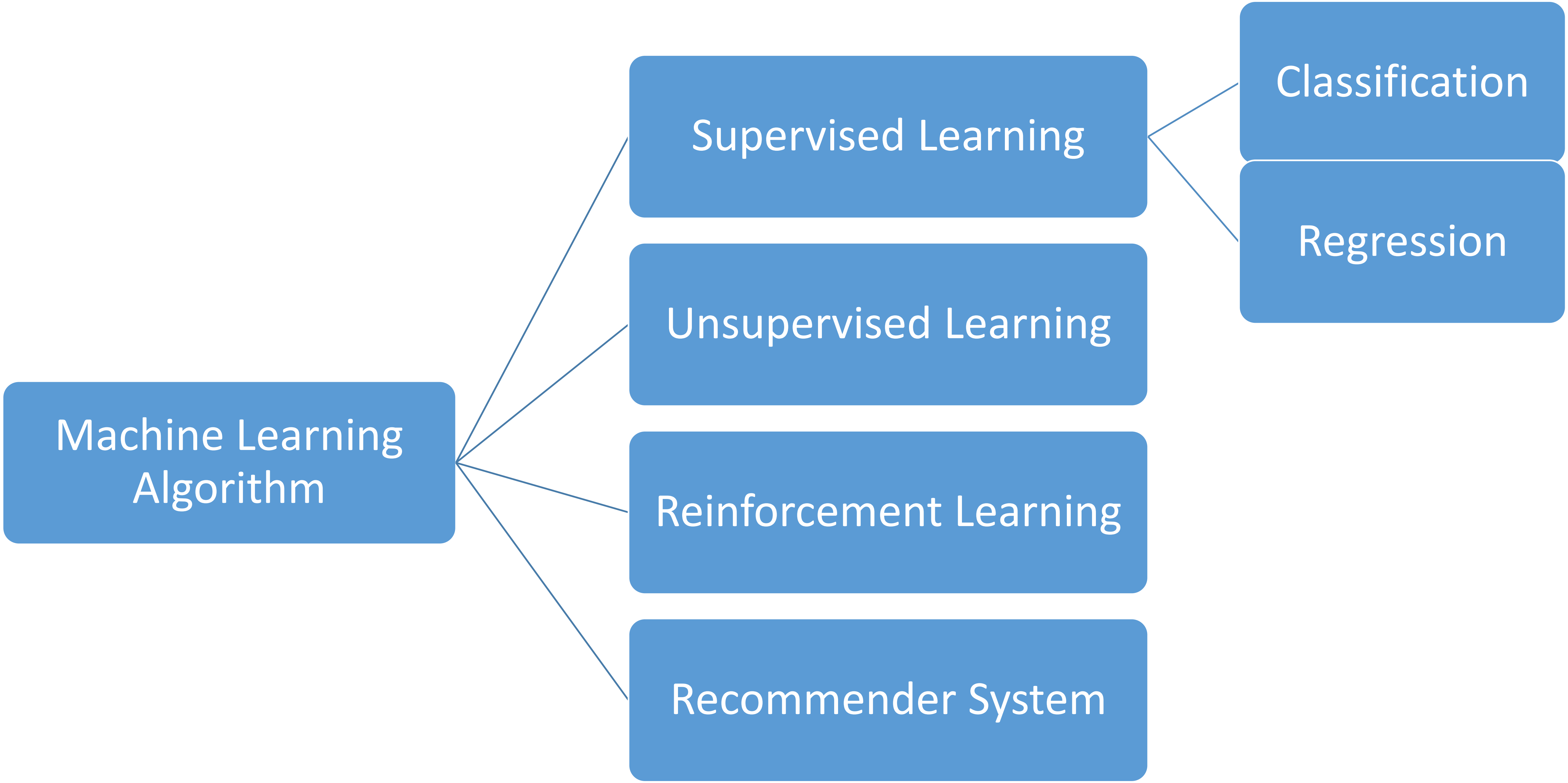
– *Software pipelines that manage the complexity of multiple subtasks*

Economics, biology, psychology

– *How can an individual or system efficiently improve their performance in a given environment?*

– *What is learning and how can it be optimized?*

Key Concepts in Machine Learning



Machine Learning Algorithm

Supervised Learning

Classification

Regression

Unsupervised Learning

Reinforcement Learning

Recommender System

Supervised Learning

Learn to predict target values from labelled data.

- Classification (target values are discrete classes)
- Regression (target values are continuous values)

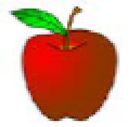
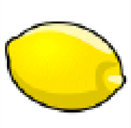
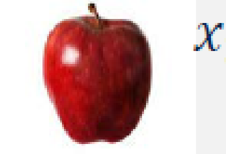
Right answers are given

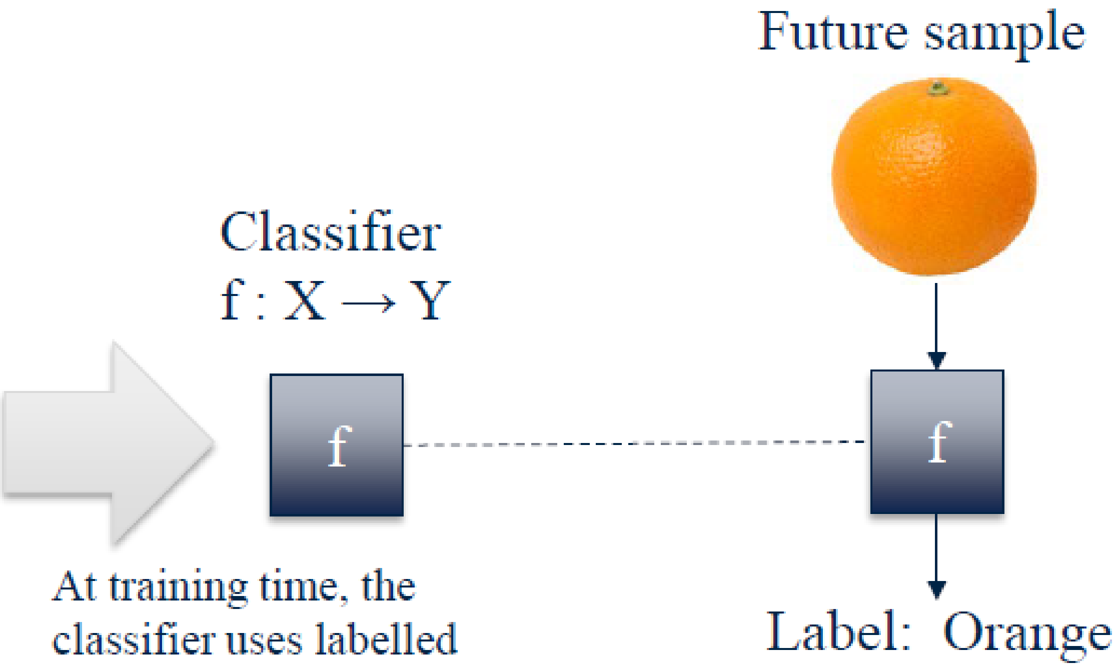
Supervised Learning - Classification

- Discrete valued output
- The function that we learn is called the ***classifier***.

Supervised Learning - Classification

Training set

X Sample	Y Target Value (Label)
 x_1	Apple y_1
 x_2	Lemon y_2
 x_3	Apple y_3
x_4	Orange y_4



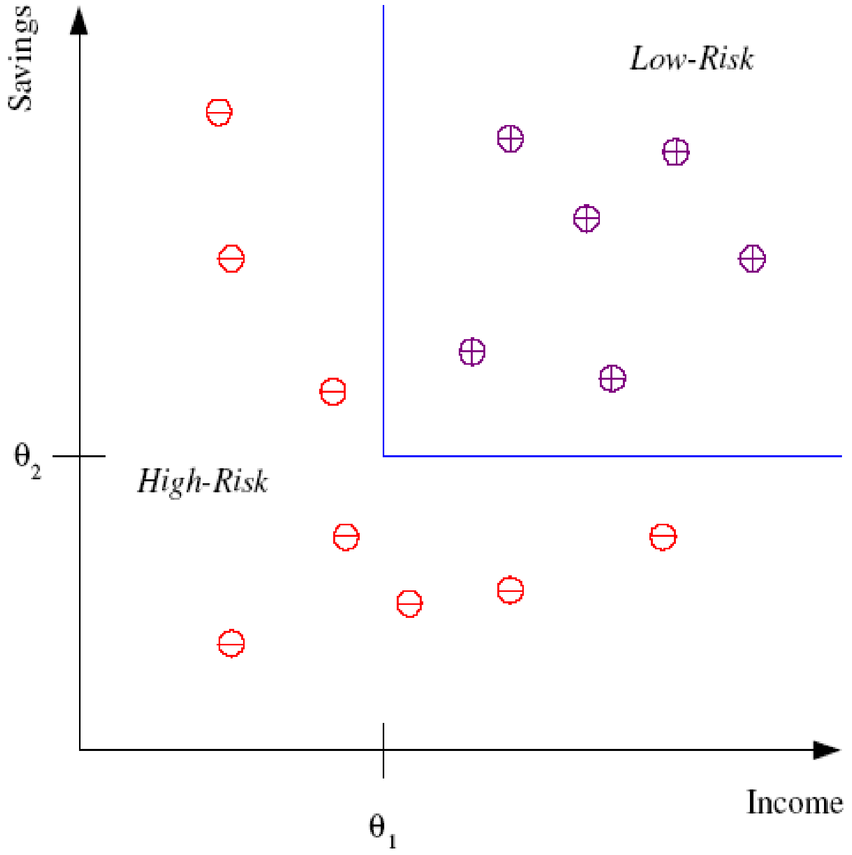
At training time, the classifier uses labelled examples to learn rules for recognizing each fruit type.

After training, at prediction time, the trained model is used to predict the fruit type for new instances using the learned rules.

Supervised Learning - Classification

Example: Credit scoring

Differentiating between low-risk and high-risk customers from their *income* and *savings*



Discriminant: IF *income* > θ_1 AND *savings* > θ_2

THEN low-risk ELSE high-risk

Supervised Learning - Regression

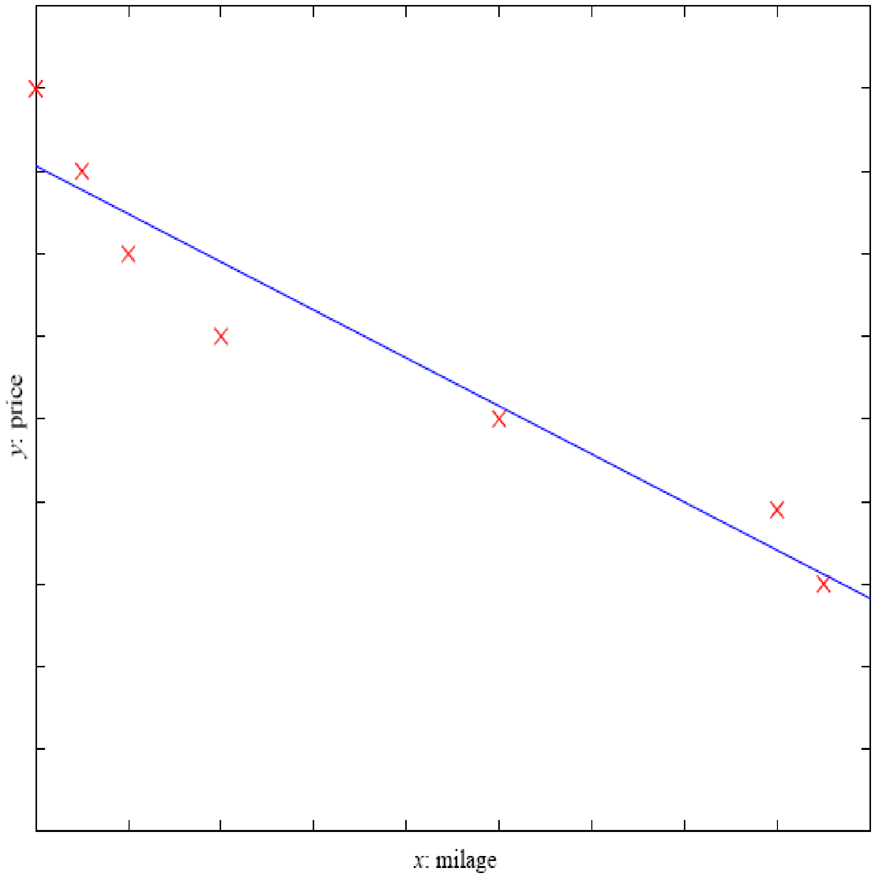
- Continuous valued output
- Regression function

Supervised Learning - Regression

Example: Price of a used car

x : car attributes

y : price



Supervised Learning: Uses

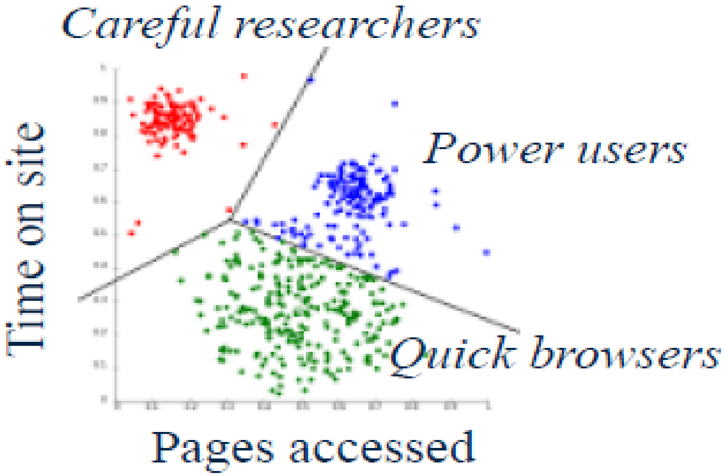
- Prediction of future cases: Use the rule to predict the output for future inputs
- Knowledge extraction: The rule is easy to understand
- Compression: The rule is simpler than the data it explains
- Outlier detection: Exceptions that are not covered by the rule, e.g., fraud

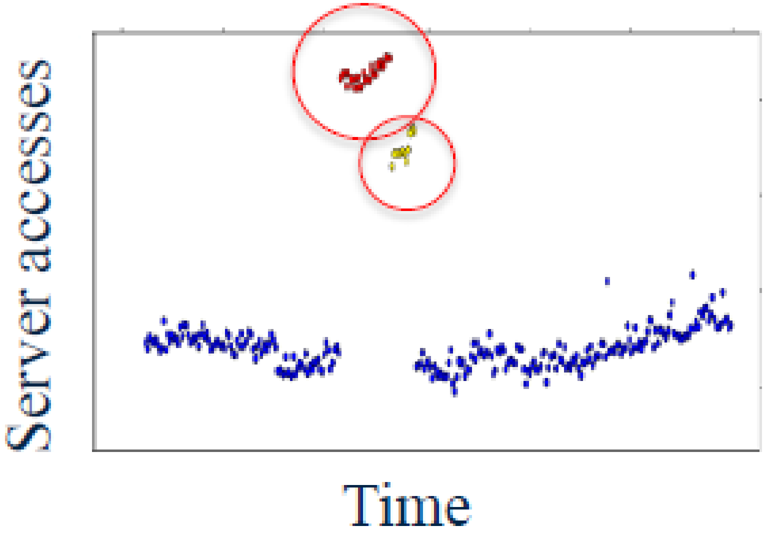
Unsupervised Learning

Find structure in *unlabeled data*

- Find groups of similar instances in the data (clustering)
- Finding unusual patterns (outlier detection)

Unsupervised Learning

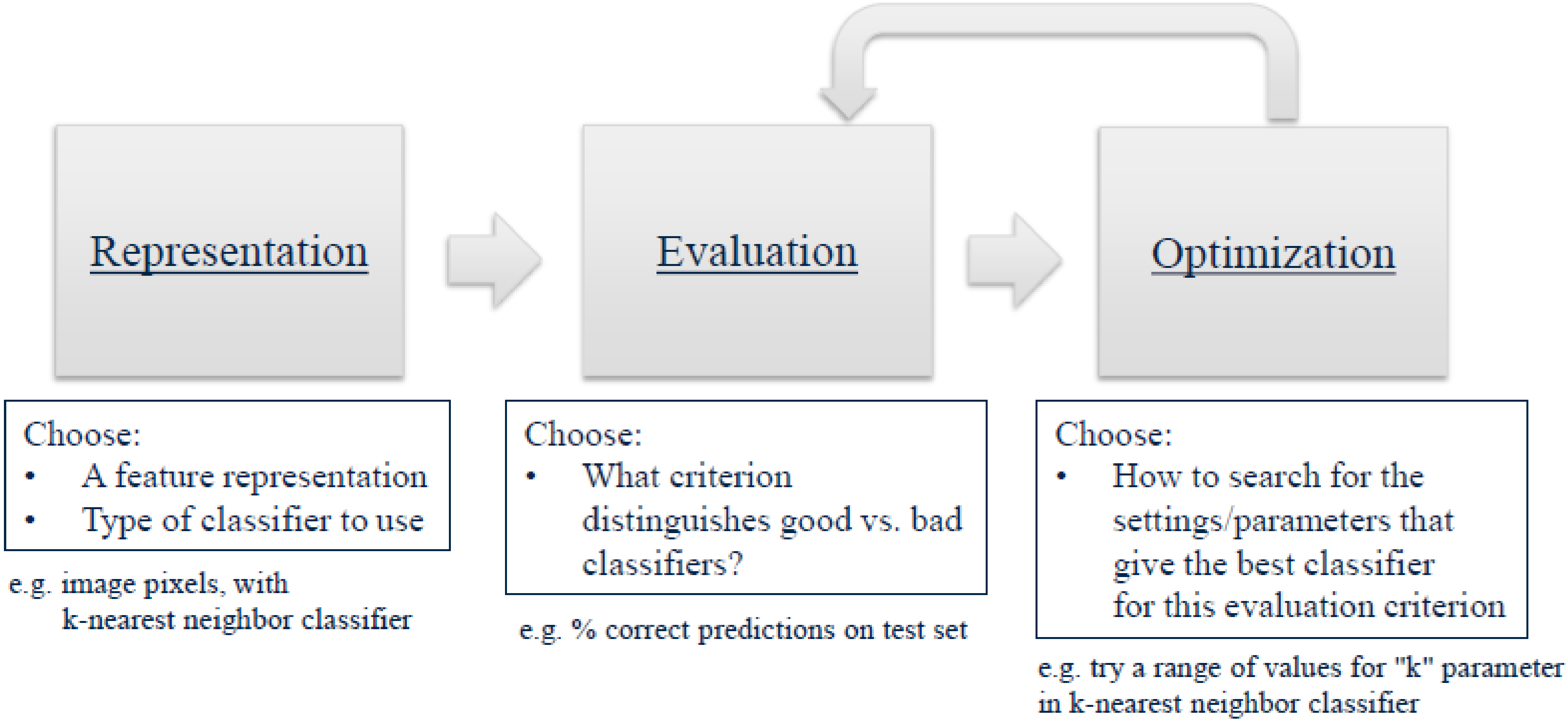




Finding clusters of similar users (clustering)

Unsupervised outlier detection

A Basic Machine Learning Workflow - Classification



Representation

Evaluation

Optimization

Choose:
- A feature representation
- Type of classifier to use

e.g. image pixels, with
k-nearest neighbor classifier

Choose:
- What criterion distinguishes good vs. bad classifiers?

e.g. % correct predictions on test set

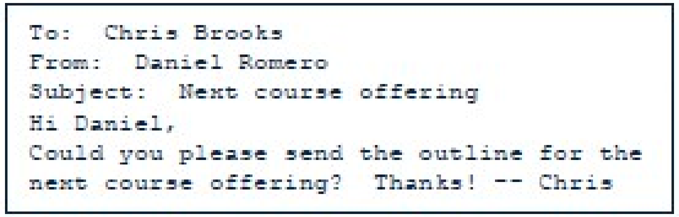
Choose:
- How to search for the settings/parameters that give the best classifier for this evaluation criterion

e.g. try a range of values for "k" parameter
in k-nearest neighbor classifier

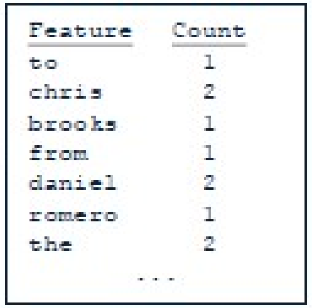
Feature Representations

Feature representation

Email



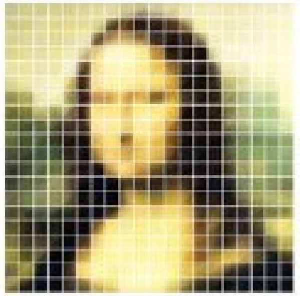
```
To:  Chris Brooks
From:  Daniel Romero
Subject:  Next course offering
Hi Daniel,
Could you please send the outline for the
next course offering?  Thanks! -- Chris
```



Feature	Count
to	1
chris	2
brooks	1
from	1
daniel	2
romero	1
the	2
...	

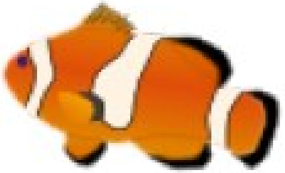
A list of words with their frequency counts

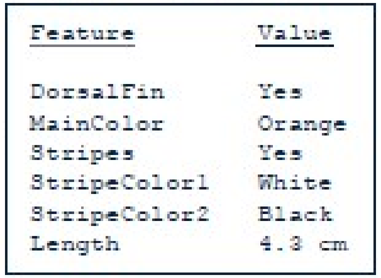
Picture



A matrix of color values (pixels)

Sea Creatures

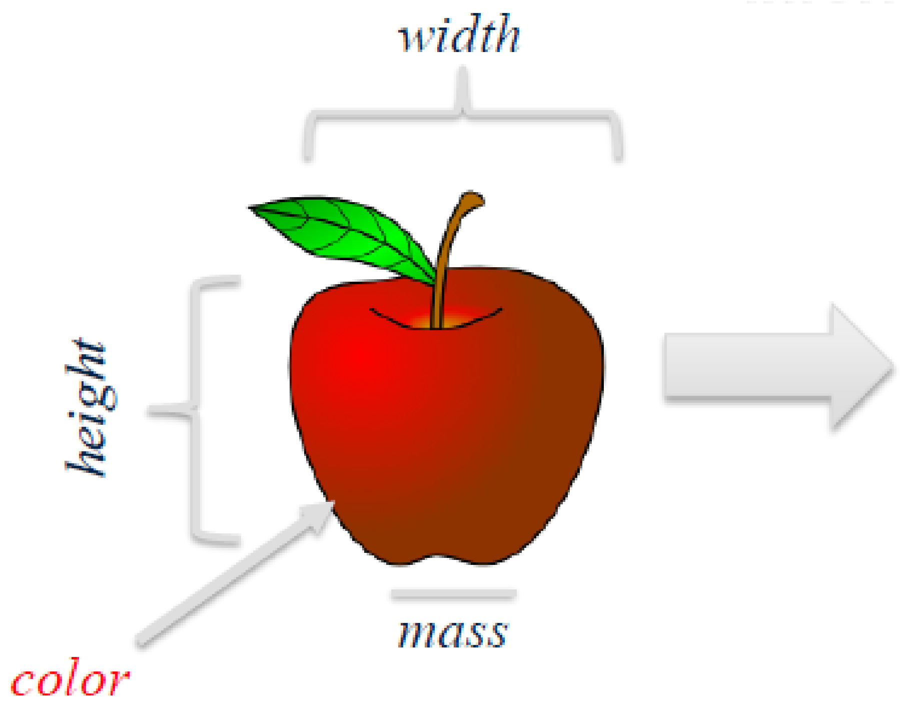




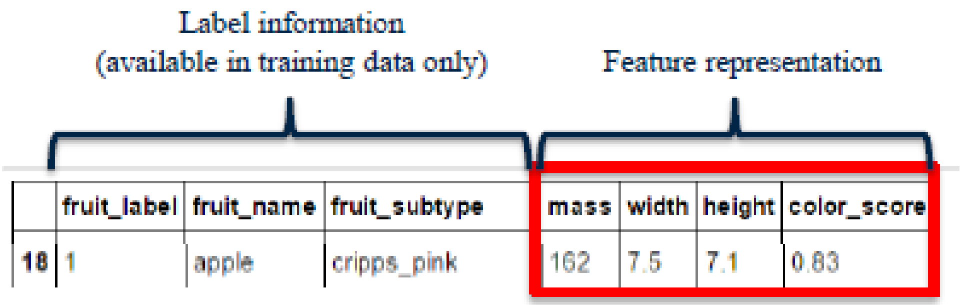
Feature	Value
DorsalFin	Yes
MainColor	Orange
Stripes	Yes
StripeColor1	White
StripeColor2	Black
Length	4.3 cm

A set of attribute values

Feature Representations



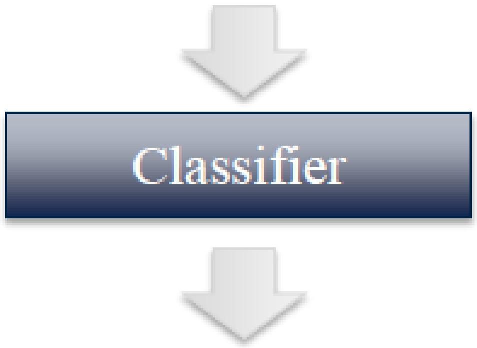
1. Feature representation



Label information (available in training data only)

Feature representation

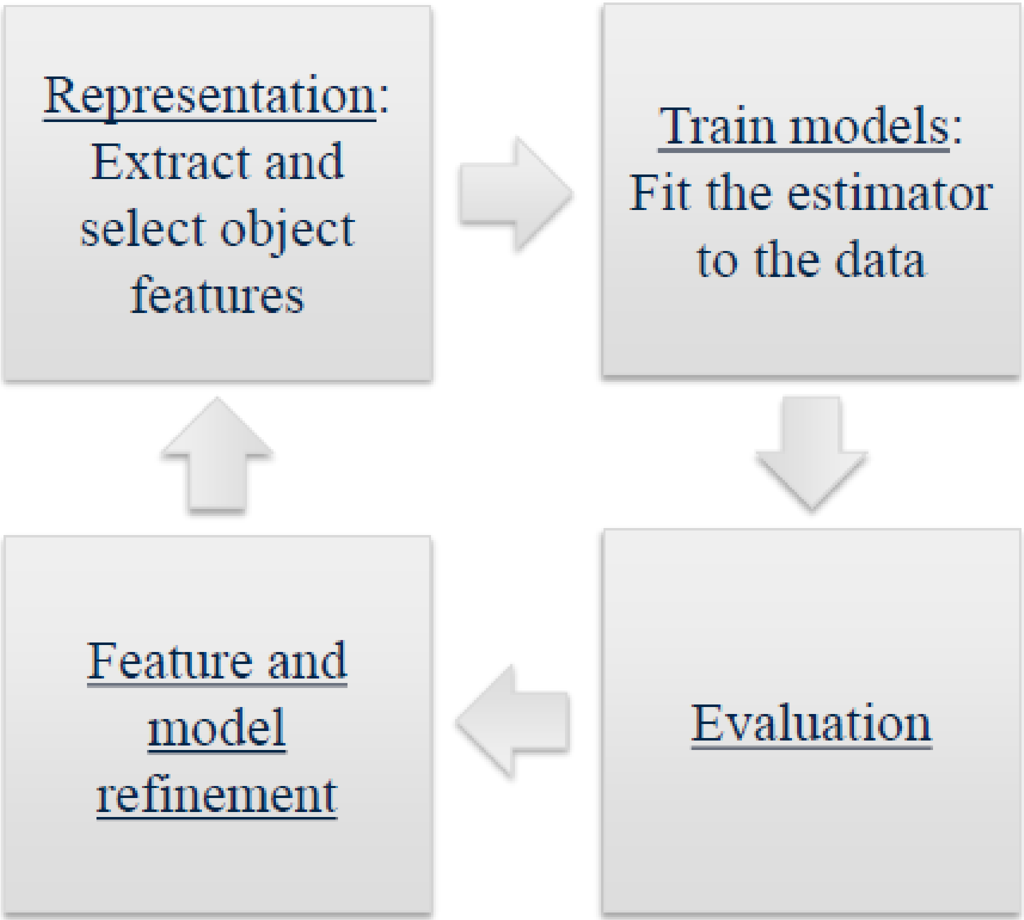
	fruit_label	fruit_name	fruit_subtype	mass	width	height	color_score
18	1	apple	cripps_pink	162	7.5	7.1	0.83



2. Learning model

Predicted class (apple)

Represent / Train / Evaluate / Refine Cycle



Representation: Extract and select object features

Train models: Fit the estimator to the data

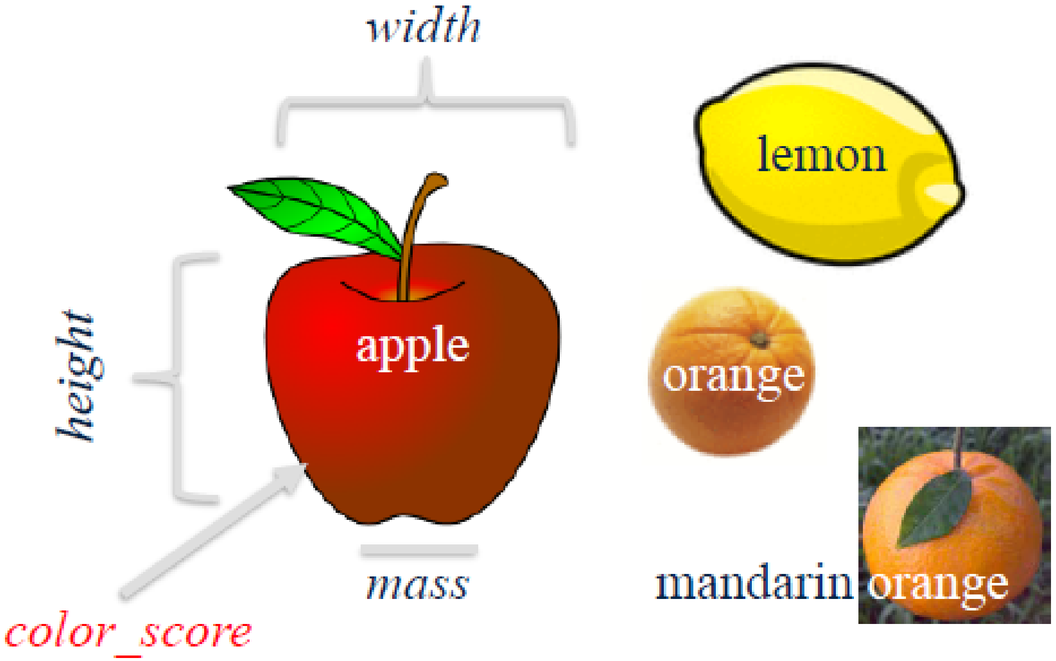
Evaluation

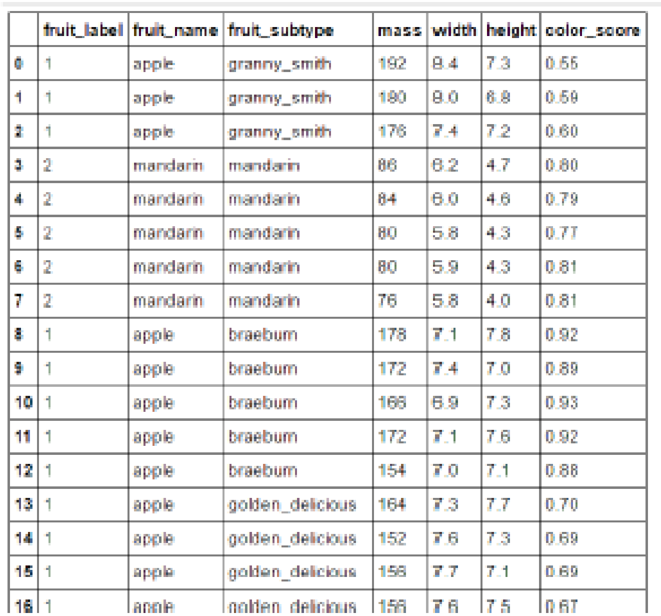
Feature and model refinement

Python Libraries

- scikit-learn : Python Machine Learning Library
- SciPy Library : Scientific Computing Tools
- NumPy : Scientific Computing Library
- Pandas : Data Manipulation and Analysis
- matplotlib and other plotting libraries

The Fruit Dataset



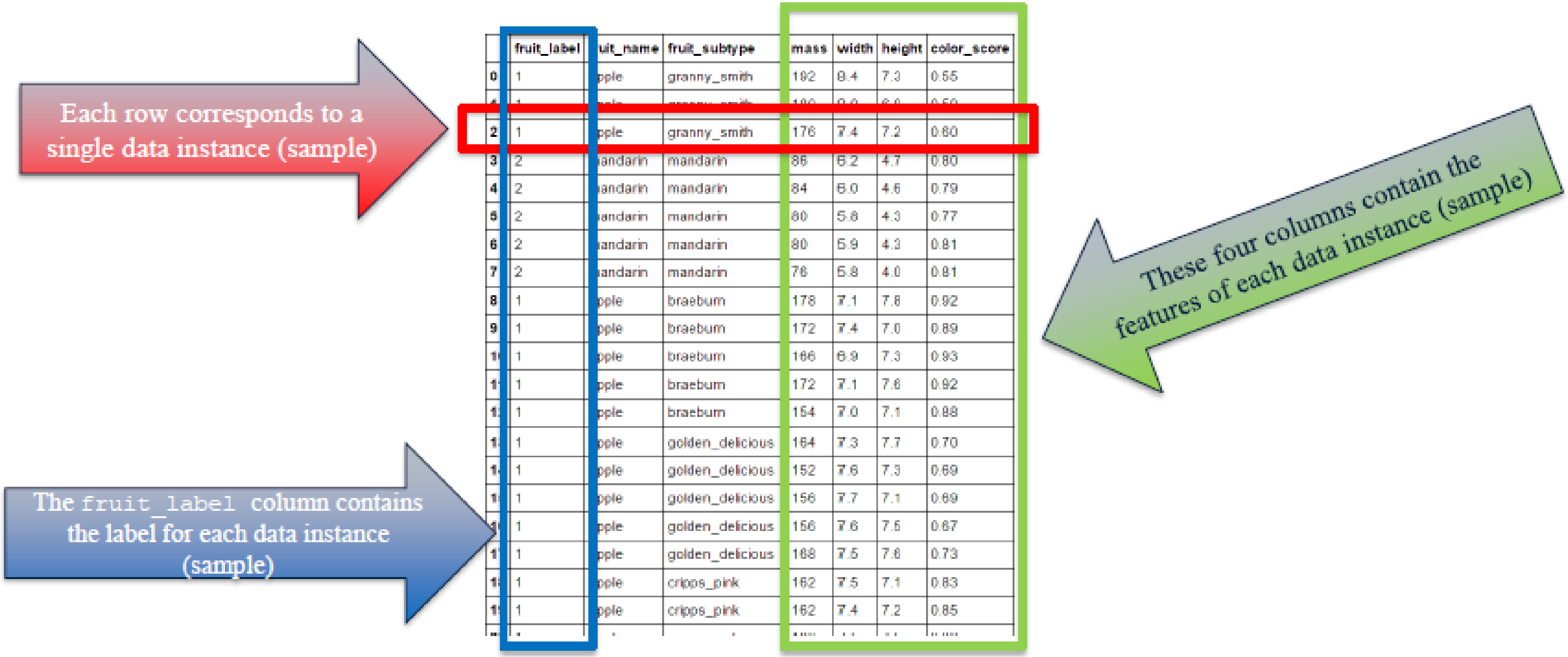


	fruit_label	fruit_name	fruit_subtype	mass	width	height	color_score
0	1	apple	granny_smith	192	8.4	7.3	0.55
1	1	apple	granny_smith	180	8.0	6.8	0.59
2	1	apple	granny_smith	176	7.4	7.2	0.60
3	2	mandarin	mandarin	86	6.2	4.7	0.80
4	2	mandarin	mandarin	84	6.0	4.6	0.79
5	2	mandarin	mandarin	80	5.8	4.3	0.77
6	2	mandarin	mandarin	80	5.9	4.3	0.81
7	2	mandarin	mandarin	76	5.8	4.0	0.81
8	1	apple	braeburn	178	7.1	7.8	0.92
9	1	apple	braeburn	172	7.4	7.0	0.89
10	1	apple	braeburn	166	6.9	7.3	0.93
11	1	apple	braeburn	172	7.1	7.6	0.92
12	1	apple	braeburn	154	7.0	7.1	0.88
13	1	apple	golden_delicious	164	7.3	7.7	0.70
14	1	apple	golden_delicious	152	7.6	7.3	0.69
15	1	apple	golden_delicious	156	7.7	7.1	0.69
16	1	apple	golden_delicious	156	7.6	7.5	0.67

`fruit_data_with_colors.txt`

Credit: Original version of the fruit dataset created by Dr. Iain Murray, Univ. of Edinburgh

The input data as a table



Each row corresponds to a single data instance (sample)

The `fruit_label` column contains the label for each data instance (sample)

These four columns contain the features of each data instance (sample)

	fruit_label	uit_name	fruit_subtype	mass	width	height	color_score
0	1	pple	granny_smith	192	8.4	7.3	0.55
1	1	pple	granny_smith	[illegible]	[illegible]	[illegible]	[illegible]
2	1	pple	granny_smith	176	7.4	7.2	0.60
3	2	andarin	mandarin	86	6.2	4.7	0.80
4	2	andarin	mandarin	84	6.0	4.6	0.79
5	2	andarin	mandarin	80	5.8	4.3	0.77
6	2	andarin	mandarin	80	5.9	4.3	0.81
7	2	andarin	mandarin	76	5.8	4.0	0.81
8	1	pple	braeburn	178	7.1	7.8	0.92
9	1	pple	braeburn	172	7.4	7.0	0.89
1	1	pple	braeburn	166	6.9	7.3	0.93
1	1	pple	braeburn	172	7.1	7.6	0.92
1	1	pple	braeburn	154	7.0	7.1	0.88
1	1	pple	golden_delicious	164	7.3	7.7	0.70
1	1	pple	golden_delicious	152	7.6	7.3	0.69
1	1	pple	golden_delicious	156	7.7	7.1	0.69
1	1	pple	golden_delicious	156	7.6	7.5	0.67
1	1	pple	golden_delicious	168	7.5	7.6	0.73
1	1	pple	cripps_pink	162	7.5	7.1	0.83
1	1	pple	cripps_pink	162	7.4	7.2	0.85

Some reasons why looking at the data initially is important

- Inspecting feature values may help identify what cleaning or preprocessing still needs to be done once you can see the range or distribution of values that is typical for each attribute.
- You might notice missing or noisy data, or inconsistencies such as the wrong data type being used for a column, incorrect units of measurements for a particular column, or that there aren’t enough examples of a particular class