COMPSCI 5100
ML & AI for Data Science

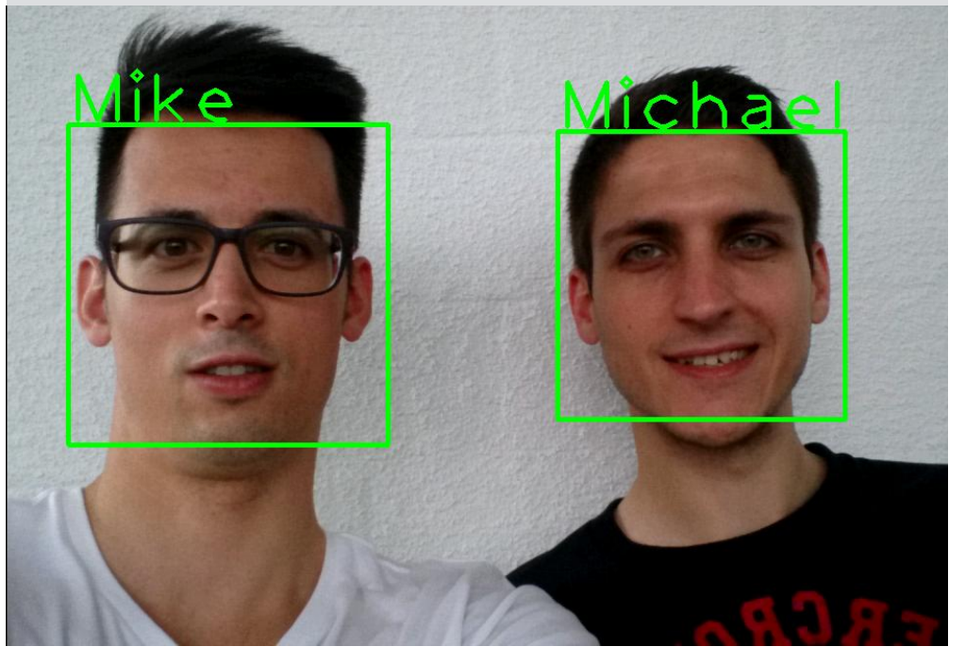
Ali Gooya

ali.gooya@glasgow.ac.uk

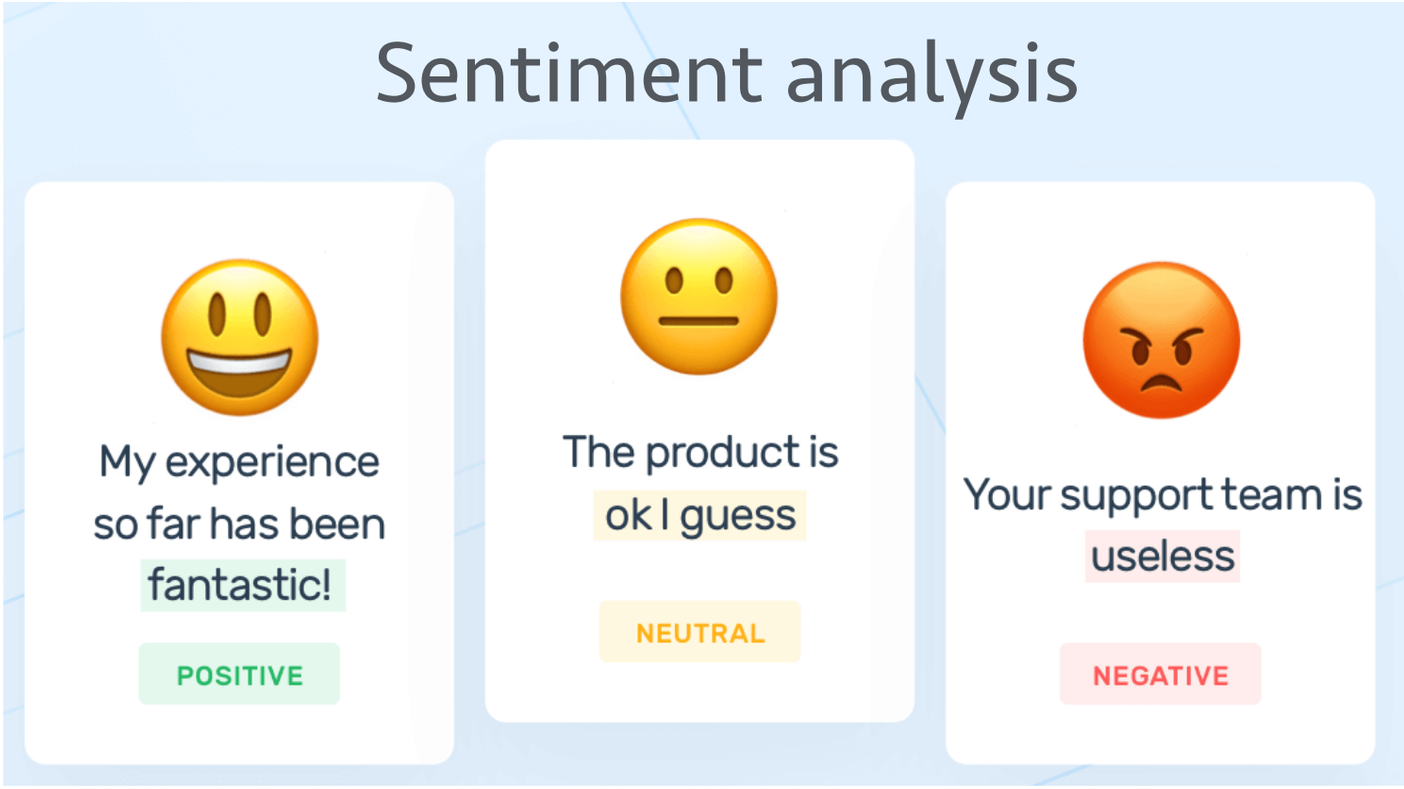
Unit 2: Classification

Classification = Automatically label data

Face recognition



Sentiment analysis



Learning Objectives

- To be able to formulate a problem to a classification task
- To understand the components of a classifier
- To be able to evaluate a classifier's performance

Classification: Part I

ML paradigms

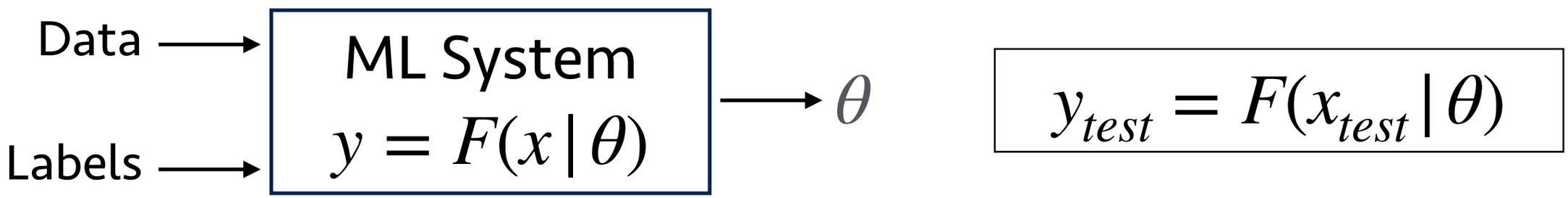
[AI ≠ ML, but in today's world ML and AI are often synonymous]

Machine Learning paradigms

- **Supervised**
  - need labelled training data
- **Semi-supervised**
  - some labelled training data
- **Unsupervised/ Self-supervised**
  - no labelled training data is needed

Supervised learning paradigm

- **Observe** a set of examples:
  - Training data (measurements of any kind)= x
  - Lables $= y$
- **Model** the relationship between data and labels
- **Predict** the label for new data (test data)



Classification vs. Regression

Classification

- Lables = $\mathbf{y}$ is discrete
- For example, y identifies an image as: *dog, cat, mouse* (class)
- As the number of classes increase, difficulty of a classification task increases

Regression

- Lables = $\mathbf{y}$ is continuous
- For example, y predicts the price of houses in Glasgow

Can we transform a regression task to a classification task?

Components of a classification system

1. Labeled data for training

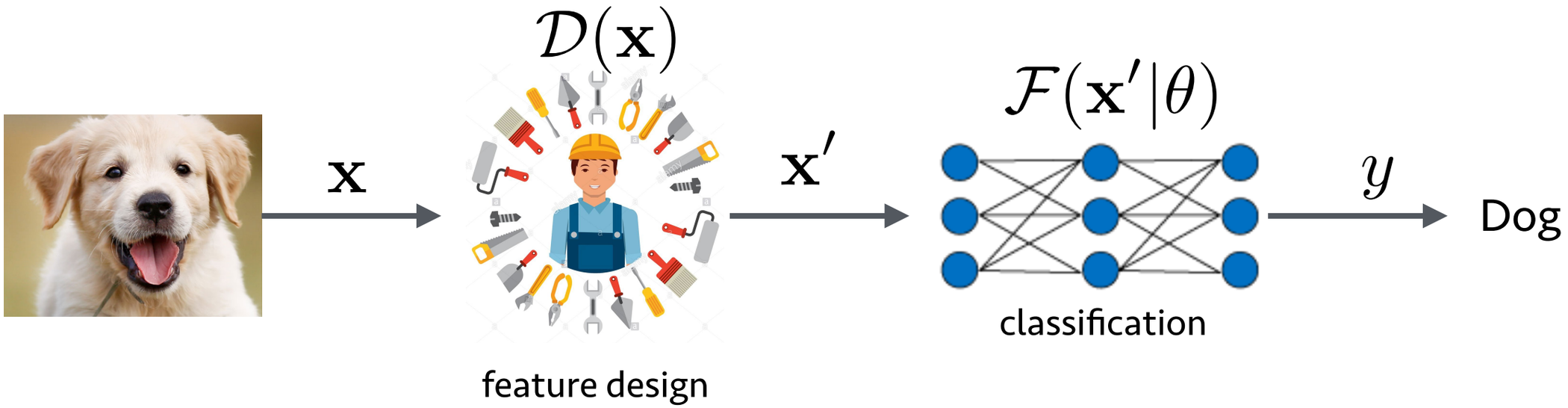
2. Features (hand crafted for traditional ML, learned in DL)

3. Model

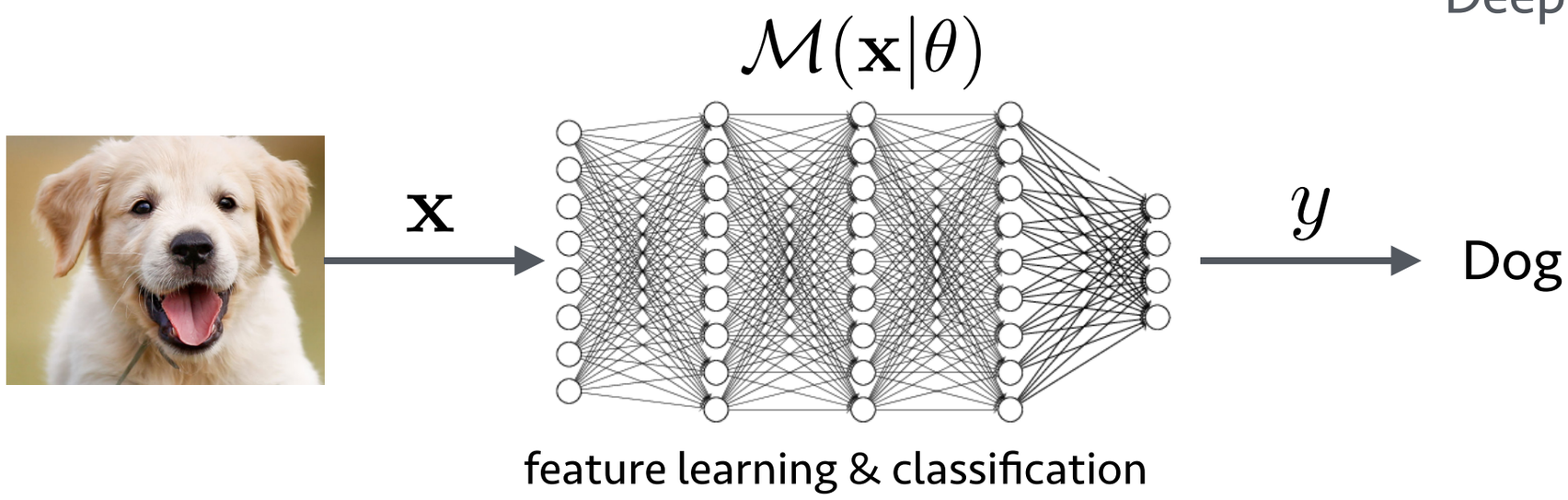
4. Evaluation metrics

Components of a classification system

Traditional machine learning



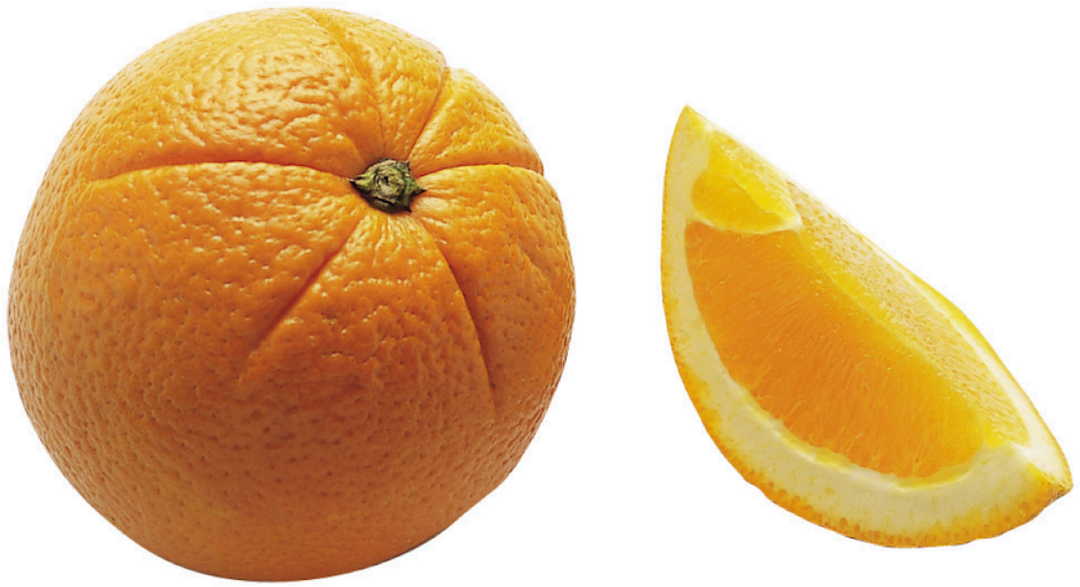
Deep learning



Example

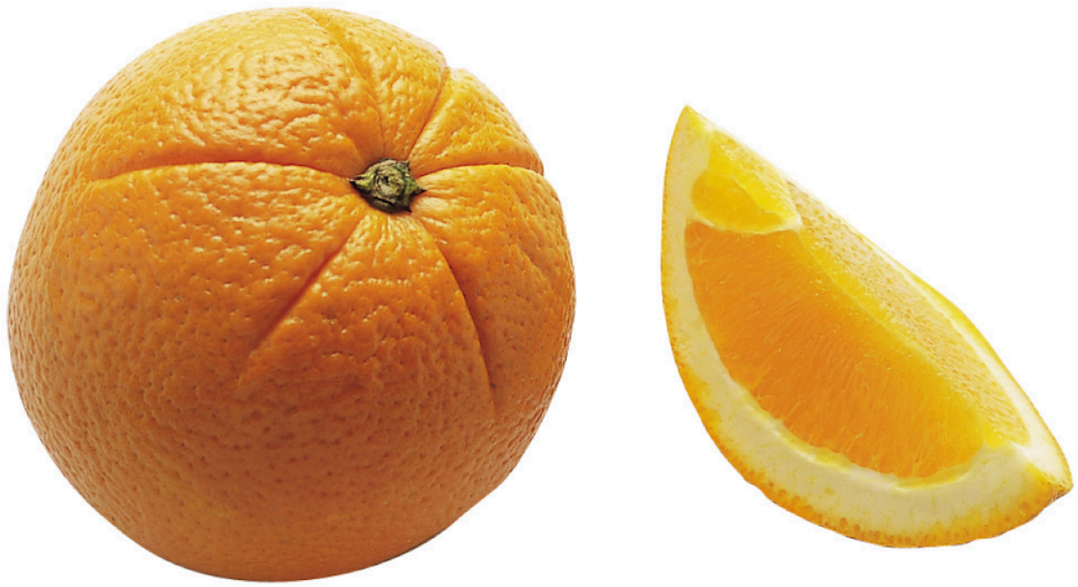
Task: Build a classification system to identify oranges and lemons

(not necessarily using images)



Example

Task: Build a classification system to identify oranges and lemons

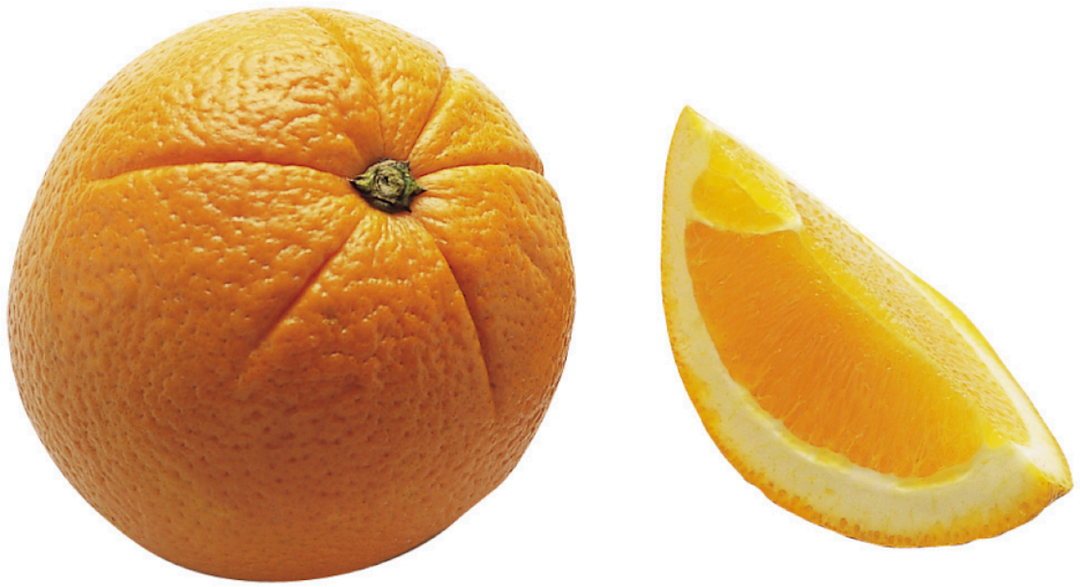


Step 1: Decide which attributes can you measure

- weight
- shape
- colour (images)

Example

Task: Build a classification system to identify oranges and lemons



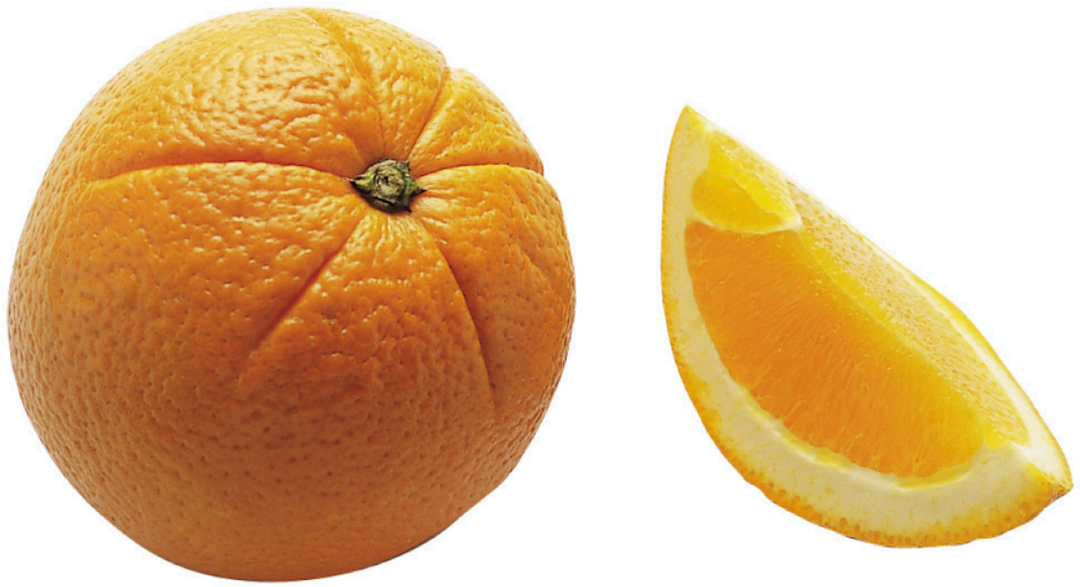
Step 2: Collect data to create training set

(Ideal) Training data

- **ML models can not recognise data they do not see**
- Need a (large) set of training data
- Balanced samples across classes
- Not noisy (accurate measurement)
- Varied examples for each class

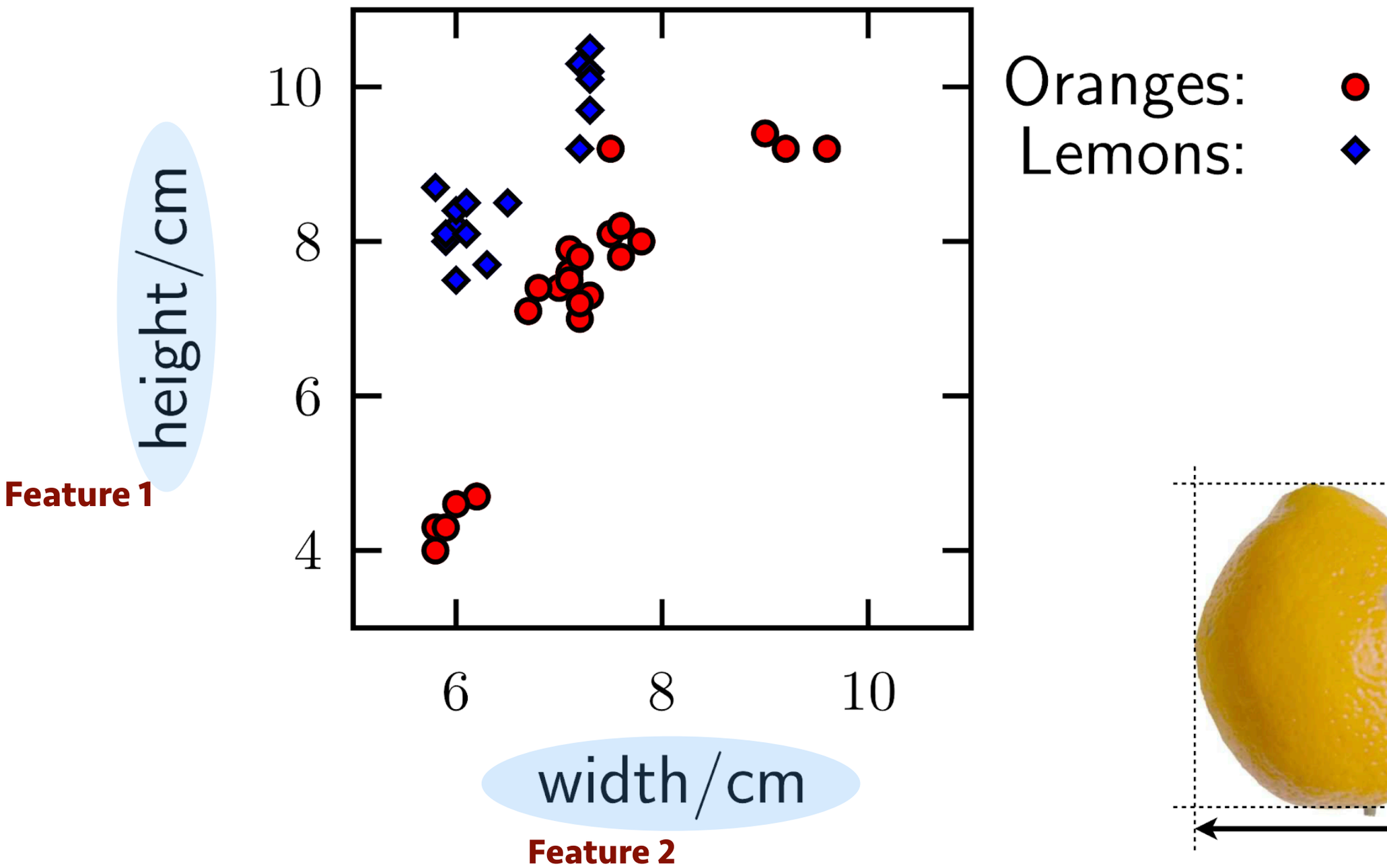
Example

Task: Build a classification system to identify oranges and lemons



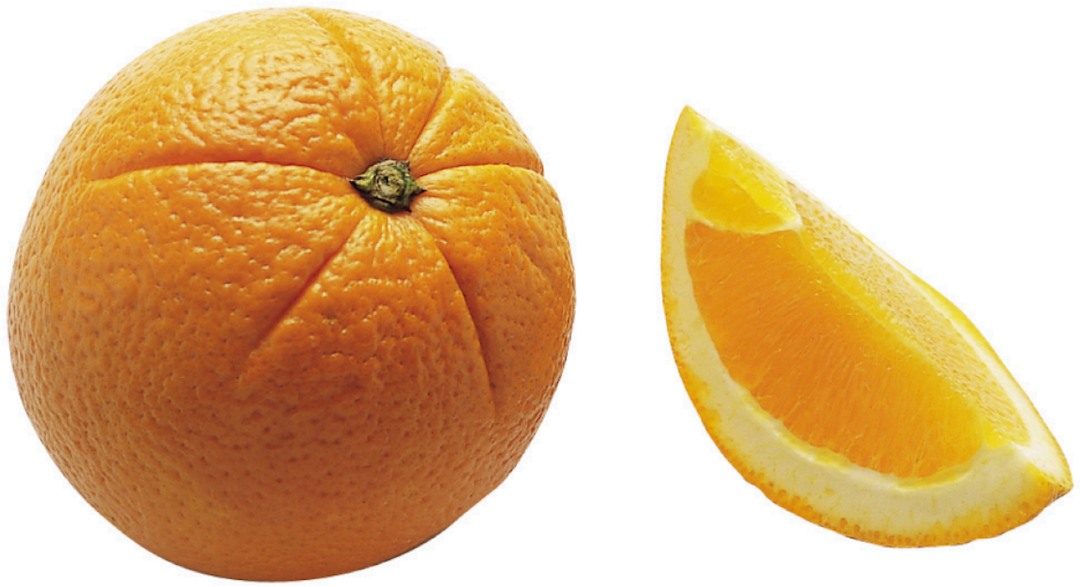
Step 3: Extract features

Features


10
8
6
4
6
8
10
height/cm
width/cm
Oranges:
Lemons:
Feature 1
Feature 2

Example

Task: Build a classification system to identify oranges and lemons



Step 4: Build ML Model (classifier)

Features

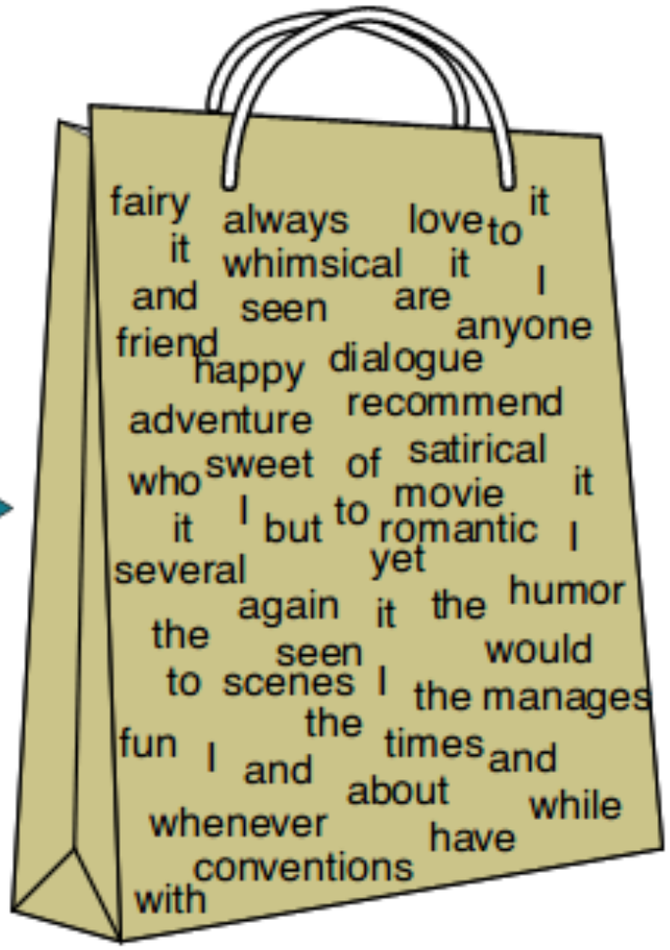
Features = Representation of Data

- Features have high impact on model performance
- For classification, features should be discriminative
- Depends on data type (modality): image, text …
    - Needs domain knowledge
- Depends on application
- Usually difficult to identify the ‘right’ set of features
    - Rely on prior work, Trial and error
    - Systematic ‘selection’ of features

Features example: Bag of Words (BoW)

Features depend on data type

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Summary: Part I

- Learning paradigms
    - supervised, unsupervised, semi-supervised

- Components of a Classification System
    - Data, features, model, evaluation metrics

- Model and Evaluation metrics coming up