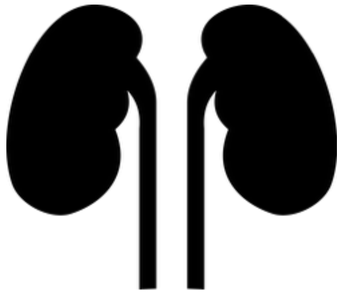
# Applied Text Mining in Python

*Text Classification*

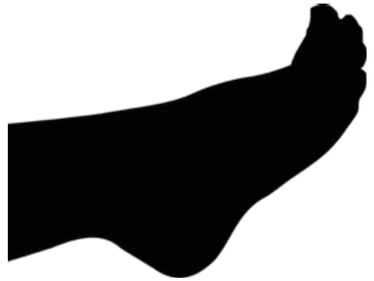
# Which medical speciality does this relate to?

TINEA PEDIS, or ATHLETE'S FOOT, is a very common fungal skin infection of the foot. It often first appears between the toes. It can be a one-time occurrence or it can be chronic. The fungus, known as Trichophyton, thrives under warm, damp conditions so people whose feet sweat a great deal are more susceptible. It is easily transmitted in showers and pool walkways. Those people with immunosuppressive conditions, such as diabetes mellitus, are also more susceptible to athlete's foot.



Nephrology

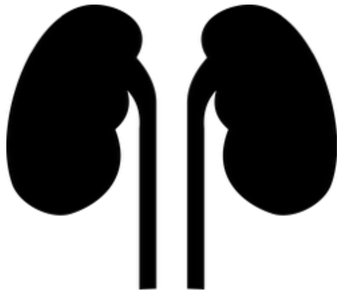
Neurology



Podiatry

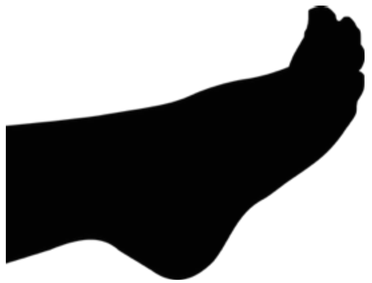
# Which medical speciality does this relate to?

KIDNEY FAILURE, also known as RENAL FAILURE or RENAL INSUFFICIENCY, is a medical condition of impaired kidney function in which the kidneys fail to adequately filter metabolic wastes from the blood.The two main forms are acute kidney injury, which is often reversible with adequate treatment, and chronic kidney disease, which is often not reversible. In both cases, there is usually an underlying cause.



Nephrology

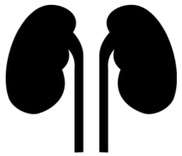
Neurology



Podiatry

# What is Classification?

Given a set of classes:



Classification: Assign the correct class label to the given input

# Examples of Text Classification

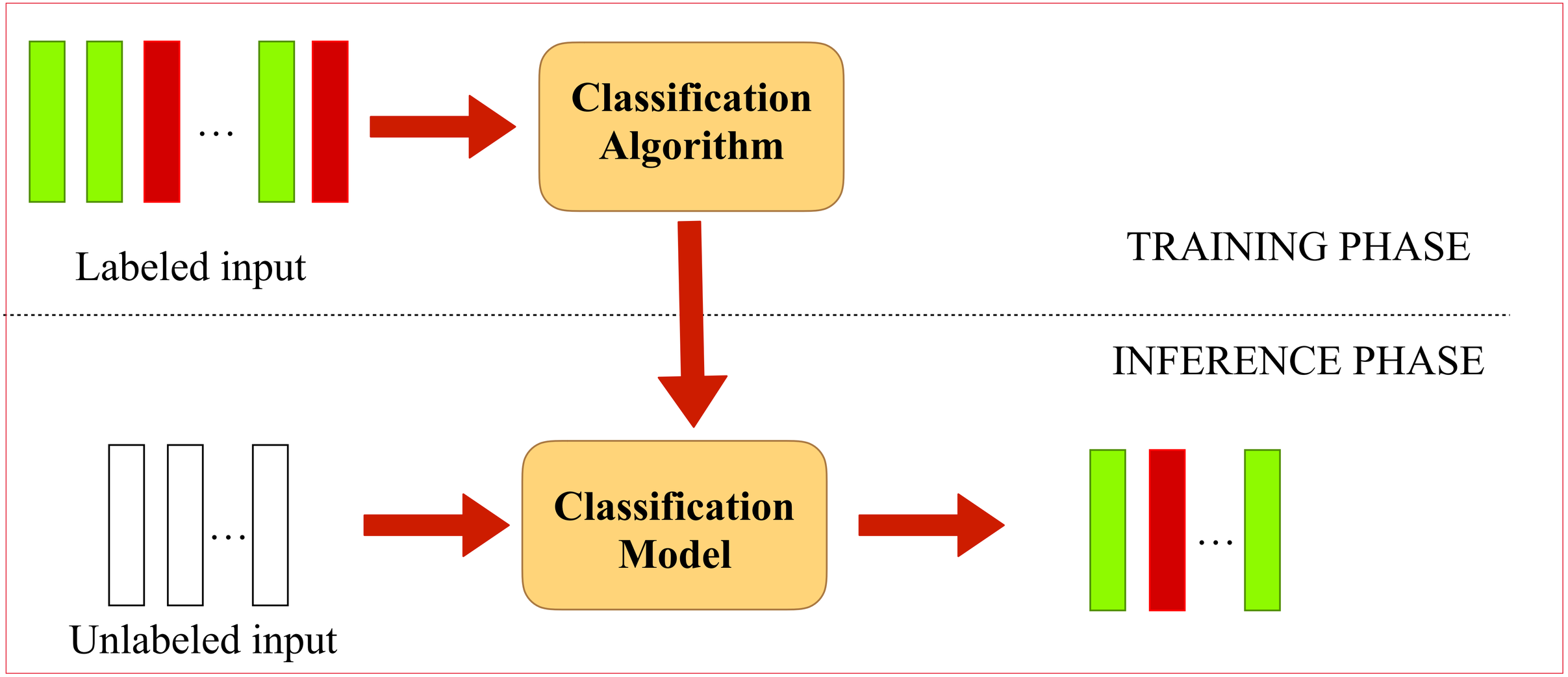
- **Topic identification**: Is this news article about Politics, Sports, or Technology?
- **Spam Detection**: Is this email a spam or not?
- **Sentiment analysis**: Is this movie review positive or negative?
- **Spelling correction**: weather or whether?
  color or colour?



# Supervised Learning

- Humans learn from past experiences, machines learn from past instances!

# Supervised Classification

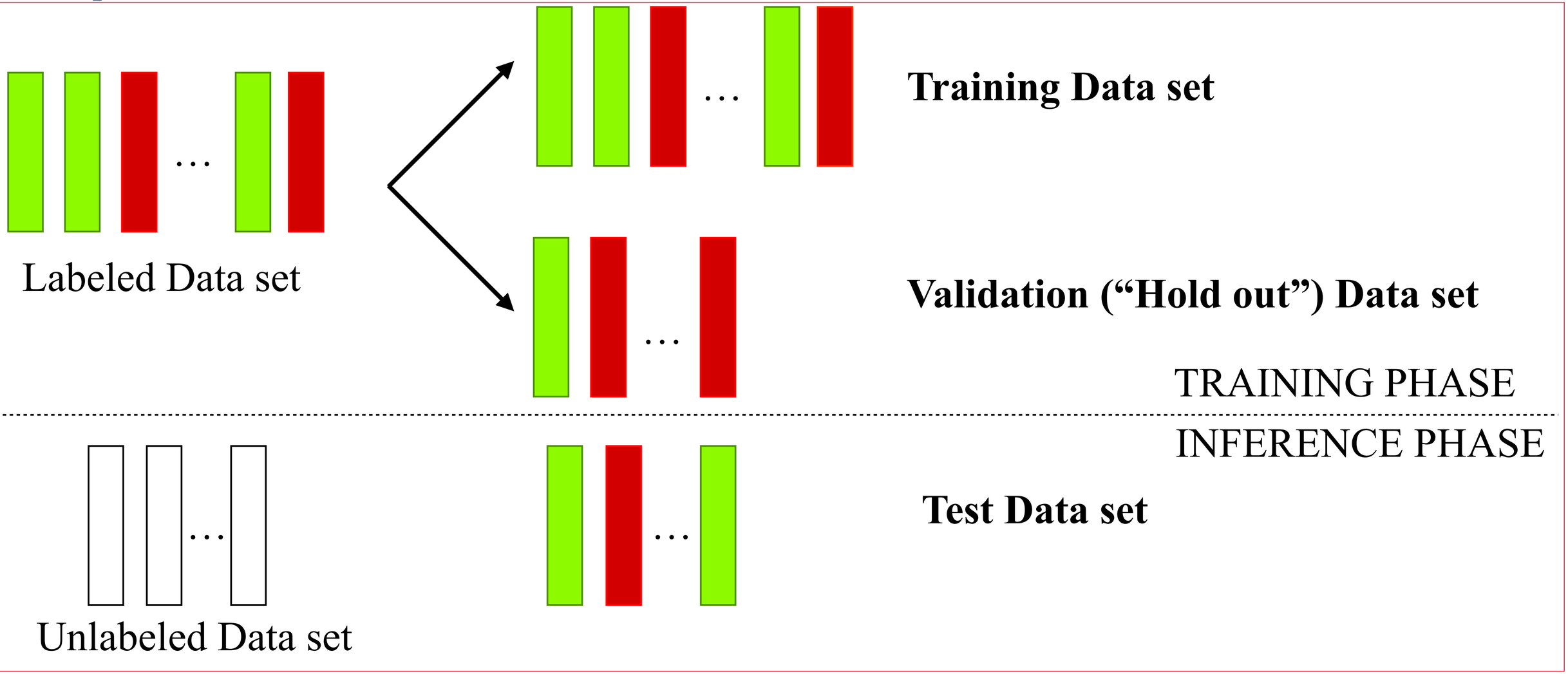


# Supervised Classification

- Learn a classification model on properties ("features") and their importance ("weights") from labeled instances
  - X: Set of attributes or features $\{x_1, x_2, \ldots, x_n\}$
  - y: A "class" label from the label set $Y = \{y_1, y_2, \ldots, y_k\}$

- Apply the model on new instances to predict the label

# Supervised Classification: Phases and Datasets



# Classification Paradigms

- When there are only two possible classes; |Y| = 2 :

  **Binary Classification**

- When there are more than two possible classes; |Y| > 2 :

  **Multi-class Classification**

- When data instances can have two or more labels :

  **Multi-label Classification**

# Questions to ask in Supervised Learning

- Training phase:
  - What are the features? How do you represent them?
  - What is the classification model / algorithm?
  - What are the model parameters?
- Inference phase:
  - What is the expected performance? What is a good measure?