Classifying Data into Predefined Categories

Swetha Kolalapudi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Recognize Classification problems in different fields : from Spam Detection to Quant Trading

Set up all the elements of a classification problem : Problem statement, Features, Labels

Classifying Data into Predefined Categories

Is this e-mail **Spam** or **Ham**?

Is this tweet **positive** or **negative**?

Is this trading day an **up-day** or a **down-day**?

All Classification Problems have the same setup

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

“Train a model” using the training data

Test

“Test the model” using test data

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

“Train a model” using the training data

Test

“Test the model” using test data

Problem Statement

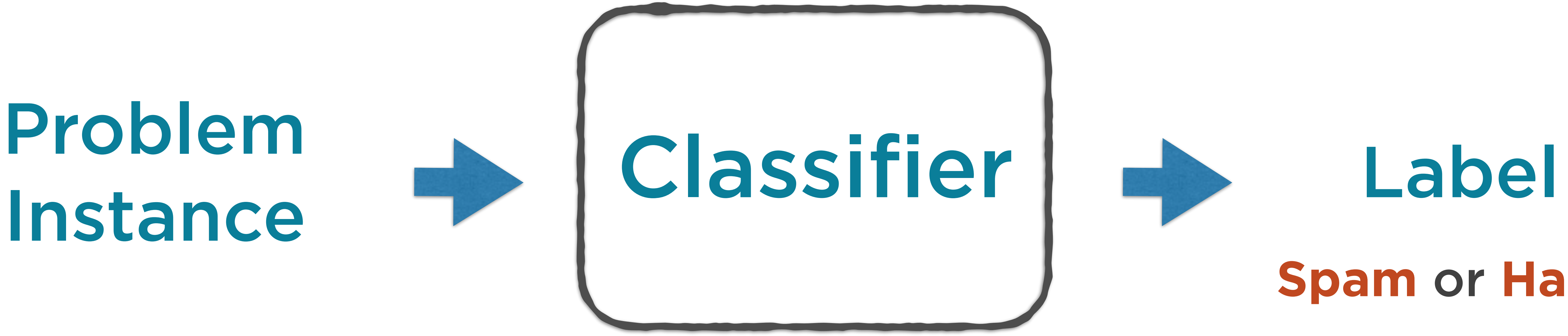
We are given a Problem Instance

An e-mail

A Tweet

A trading day

Problem Statement



Problem Instance

Classifier

Label

Spam or **Ham**?

positive or **negative**?

up-day or **down-day**?

The Classifier **assigns a label**

Classifier

This classifier is like a black box

Machine Learning Objective

Classifier

Build this black box

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

"Train a model" using the training data

Test

"Test the model" using test data

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

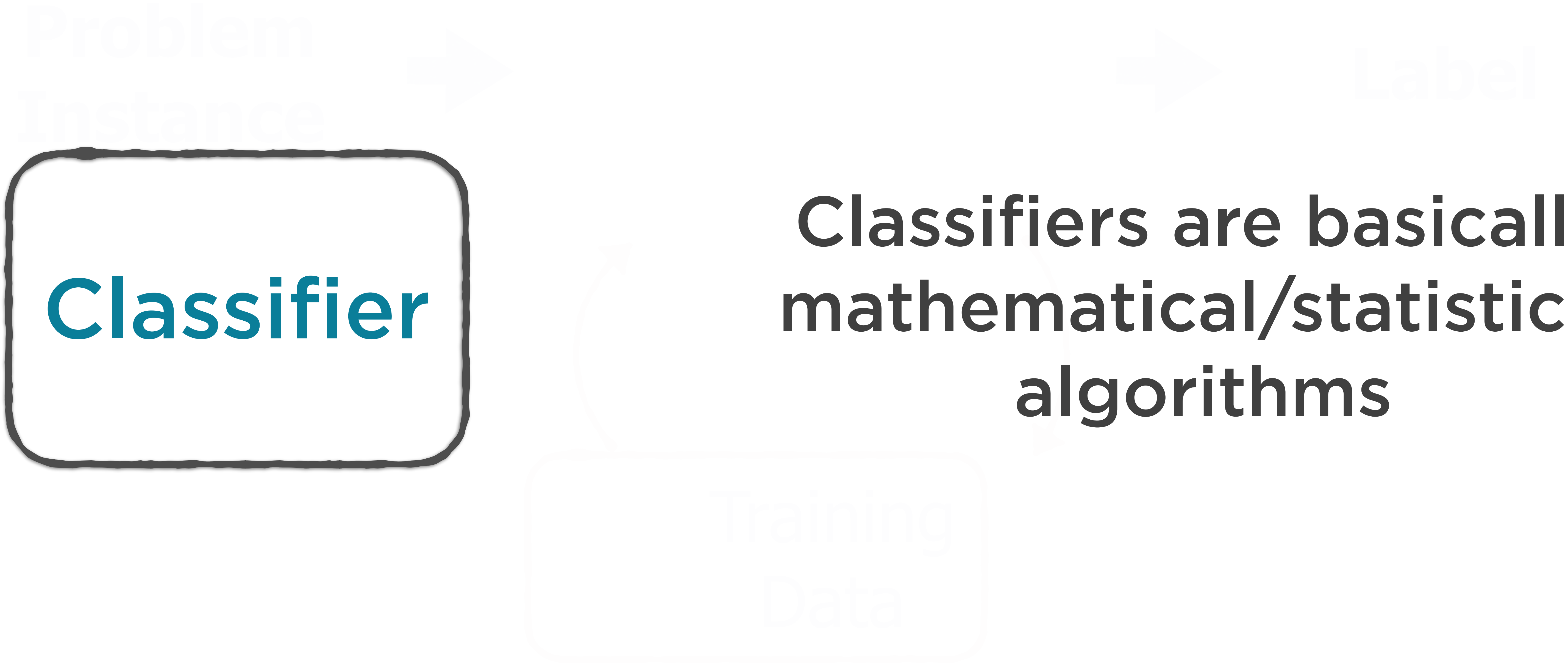
Training

“Train a model” using the training data

Test

“Test the model” using test data

Features



Problem Instance

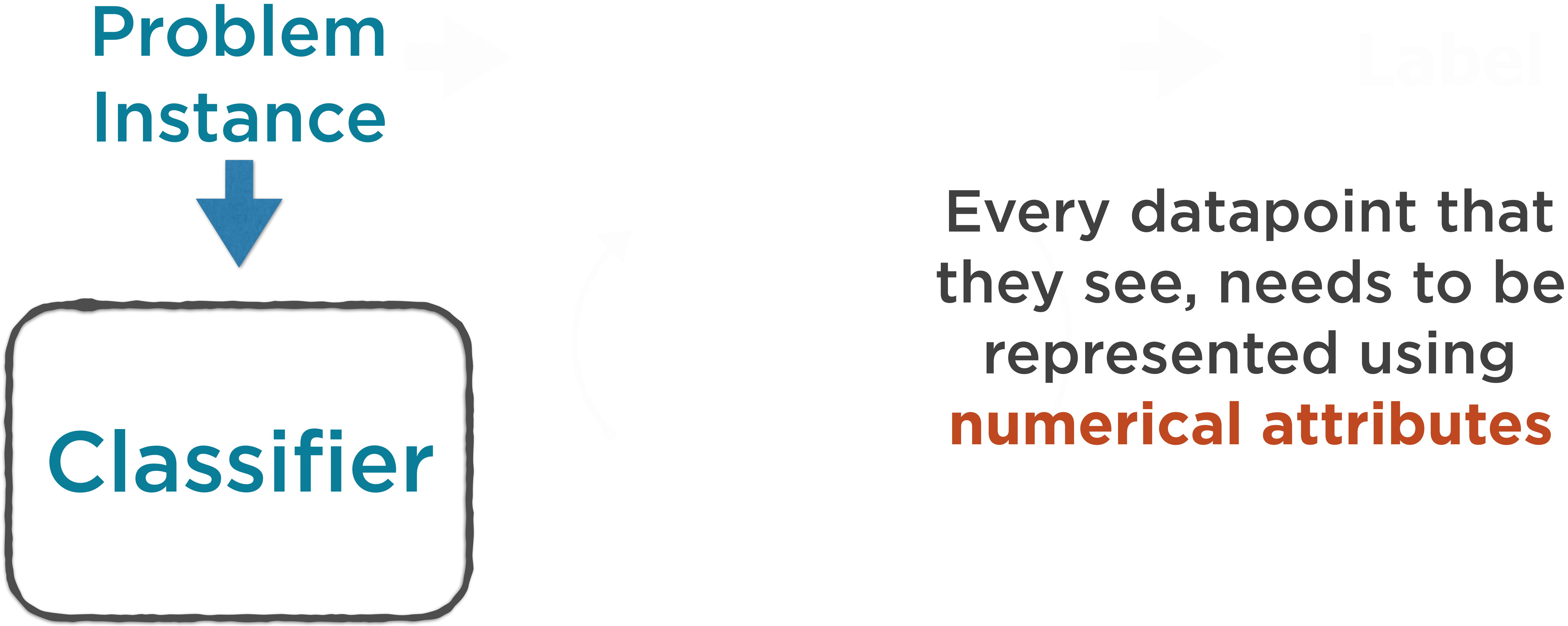
Classifier

Label

Training Data

Classifiers are basically mathematical/statistical algorithms

Features



Every datapoint that they see, needs to be represented using **numerical attributes**

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

"Train a model" using the training data

Test

"Test the model" using test data

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

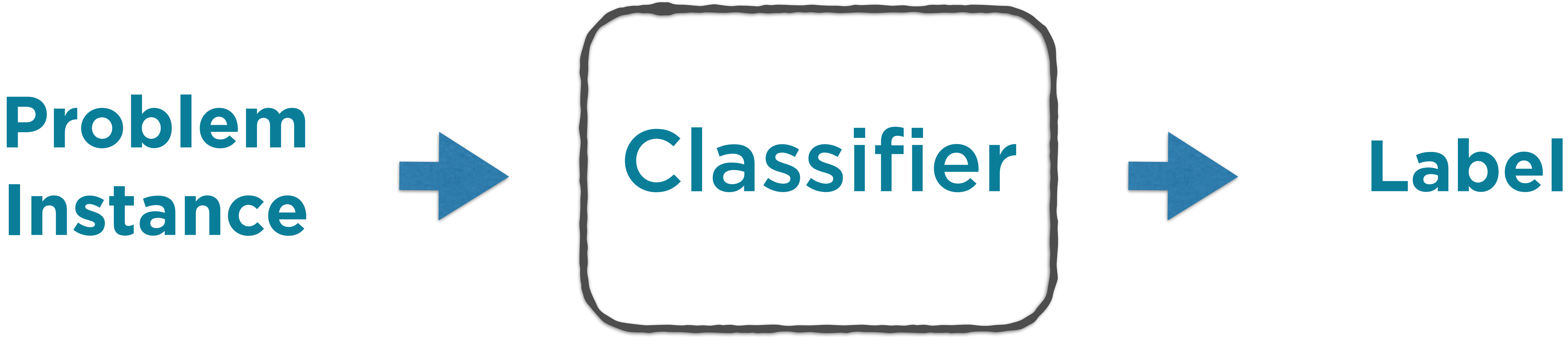
Training

“Train a model” using the training data

Test

“Test the model” using test data

Training Phase



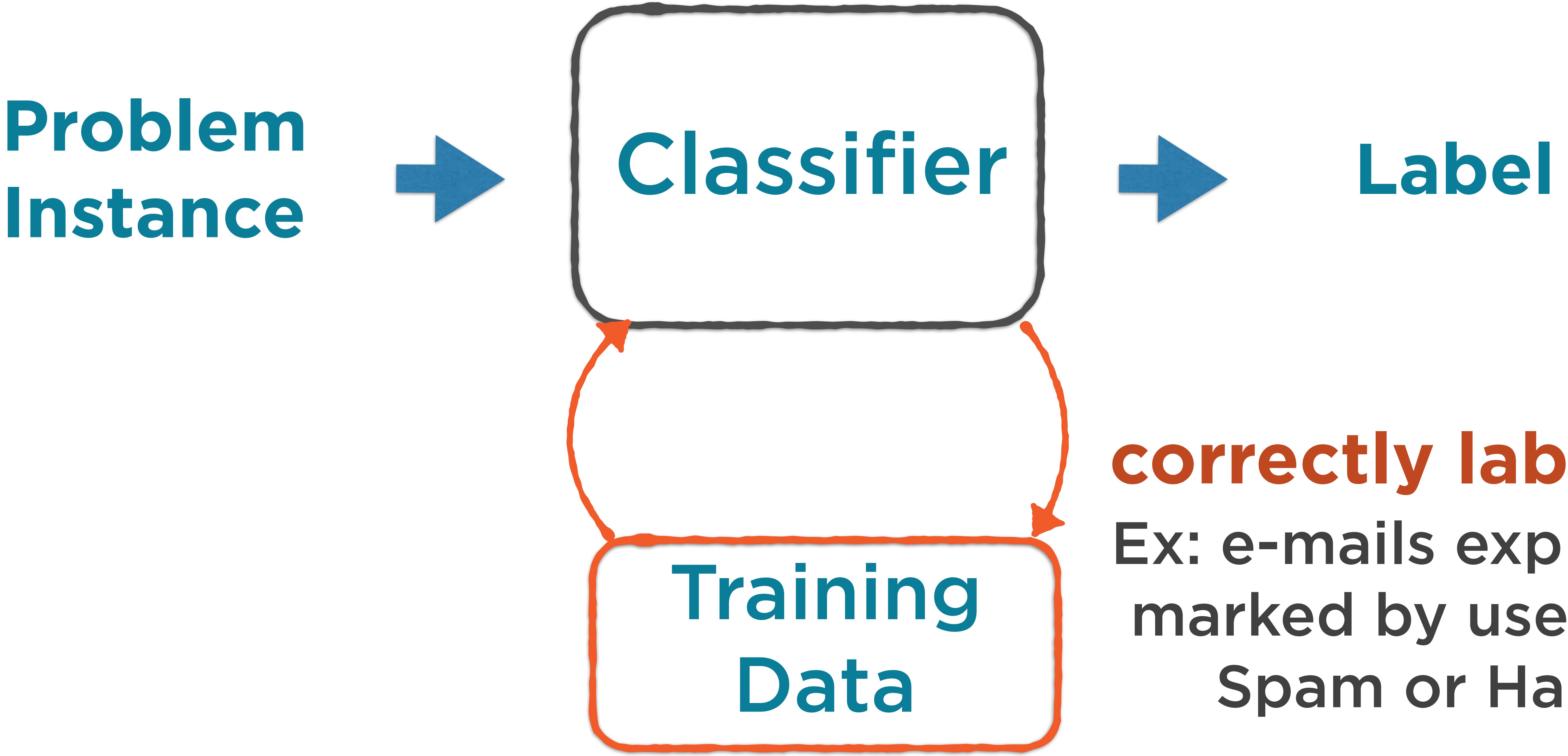
Problem Instance

Classifier

Label

The Classification algorithm will look at a set of instances which are **correctly labeled**

Training Phase



Problem Instance

Classifier

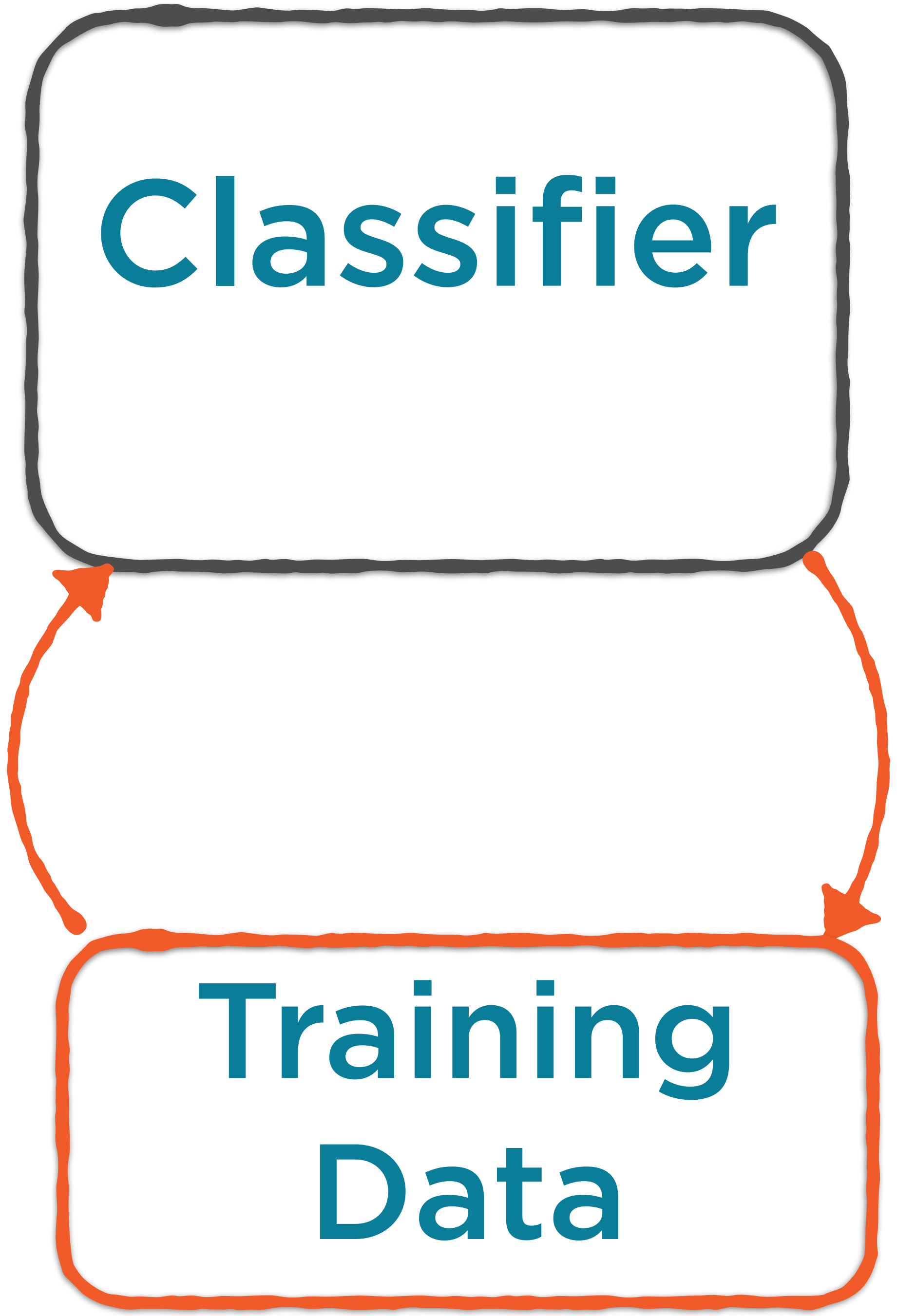
Label

Training Data

correctly labeled

Ex: e-mails explicitly marked by users as Spam or Ham

Training Phase



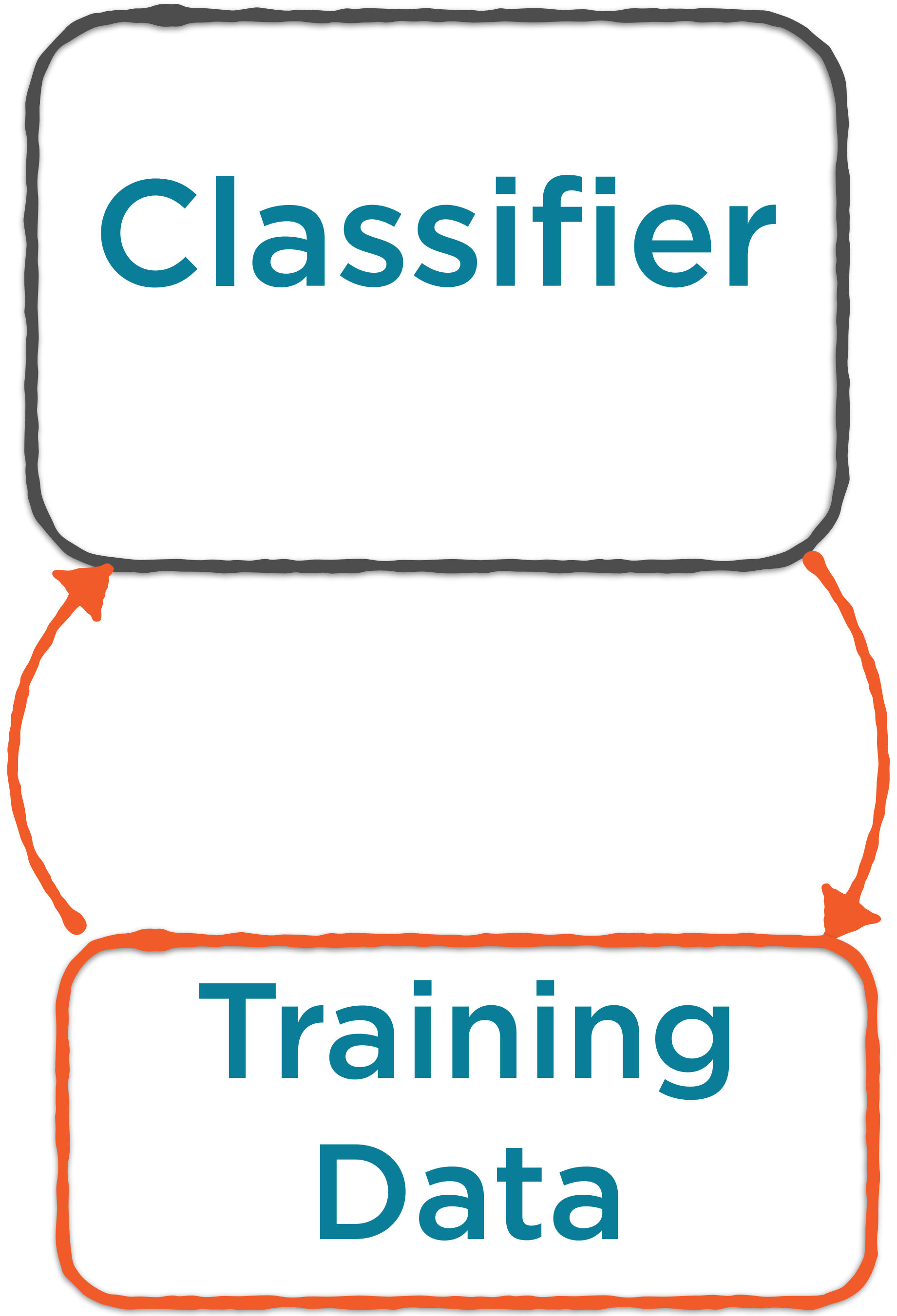
The classifier **"learns"** from the training data

Training Phase

Training Data

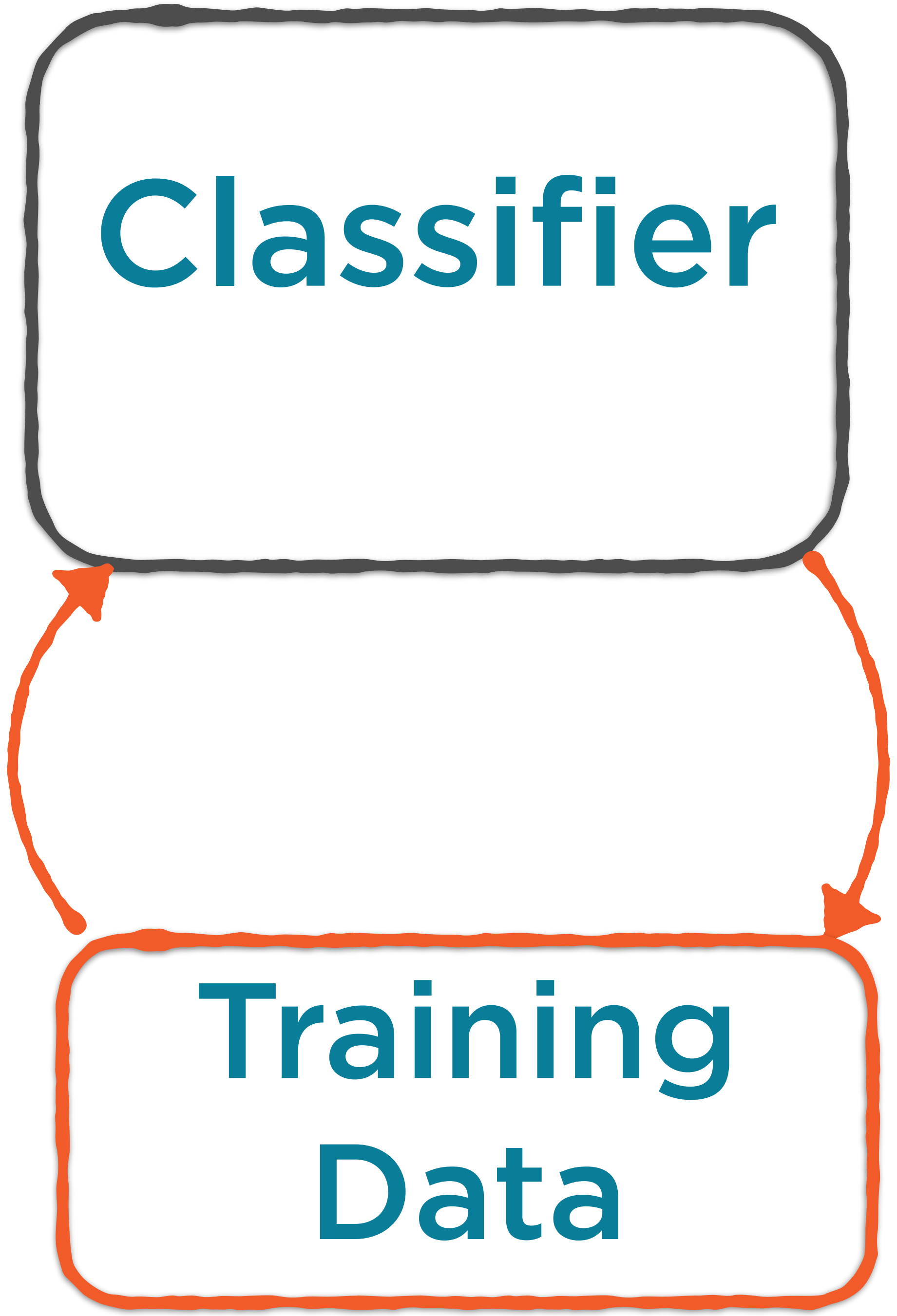
Tuples of (Features, Label)

Training Phase



The patterns that the Classifier “learns” in this phase, constitute the **Model**

Training Phase



ML techniques which have an explicit "training a model" phase are examples of

Supervised Learning

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

“Train a model” using the training data

Test

“Test the model” using test data

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

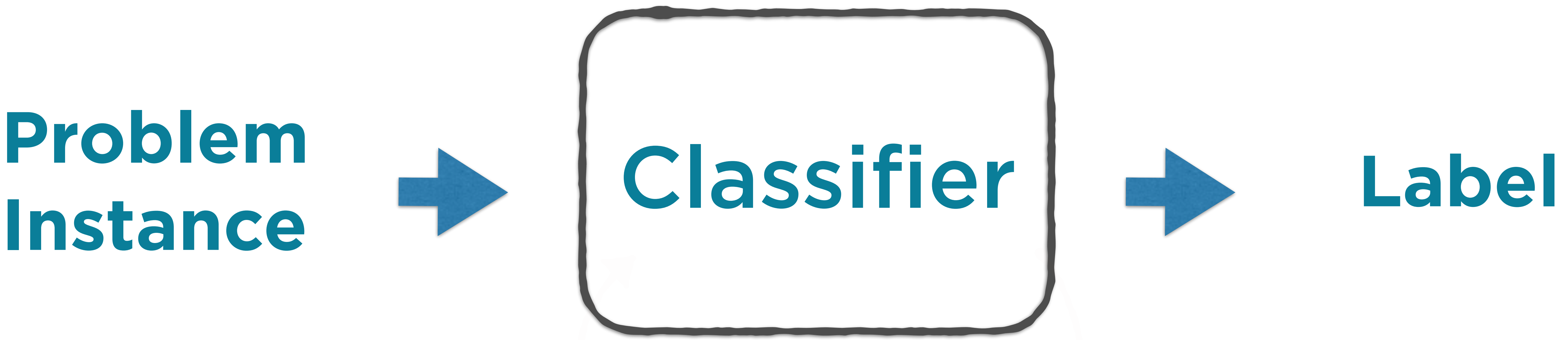
Training

“Train a model” using the training data

Test

“Test the model” using test data

Test Phase



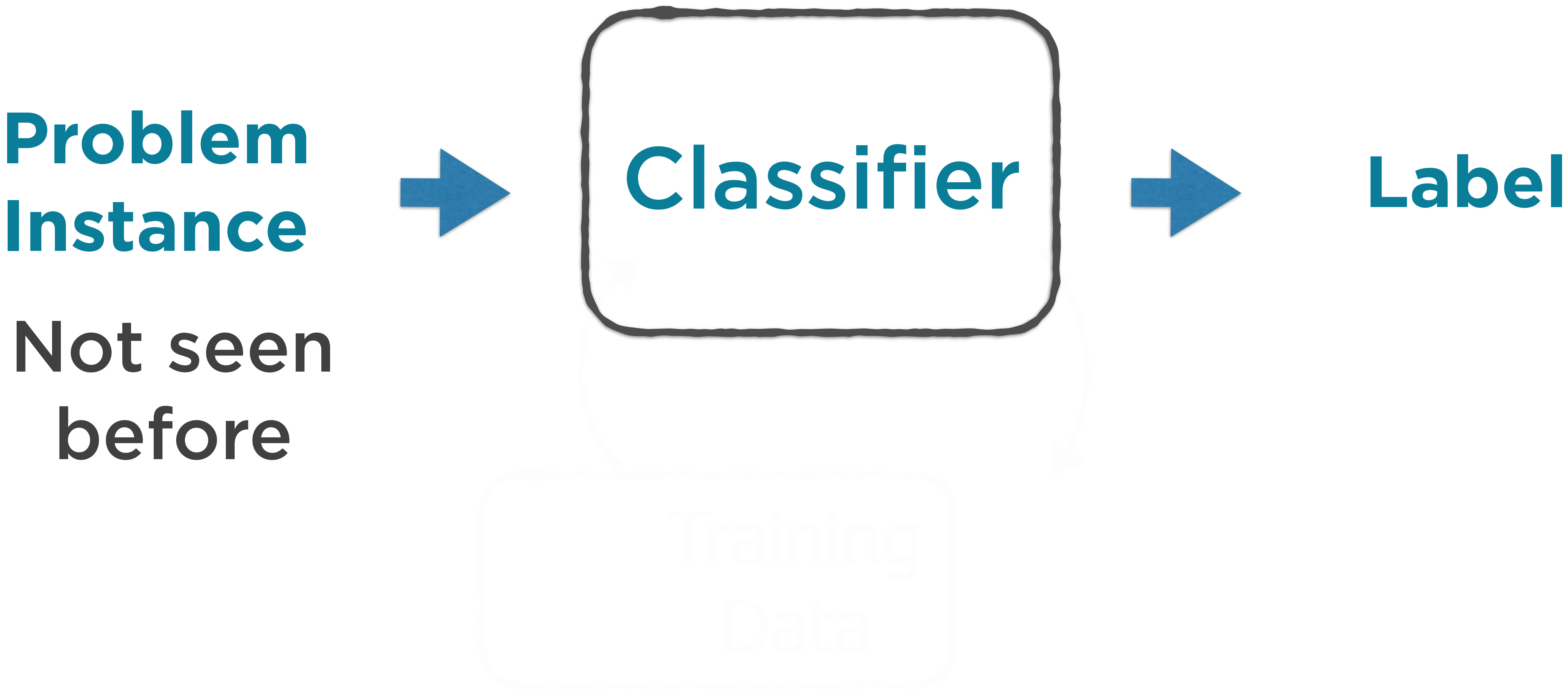
Problem Instance

Classifier

Label

The classifier **classifies** new instances

Test Phase



Problem Instance

Not seen before

Classifier

Label

Training Data

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

“Train a model” using the training data

Test

“Test the model” using test data

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

These 2 steps require careful consideration

Typical Classification Setup

Plug and play a standard algorithm using pre-built libraries

Training

"Train a model" using the training data

Test

"Test the model" using test data

Typical Classification Setup

There are several standard algorithms to choose from

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

“Train a model” using the training data

Test

“Test the model” using test data

Algorithms for Solving Classification Problems

Naive Bayes

Support Vector Machines

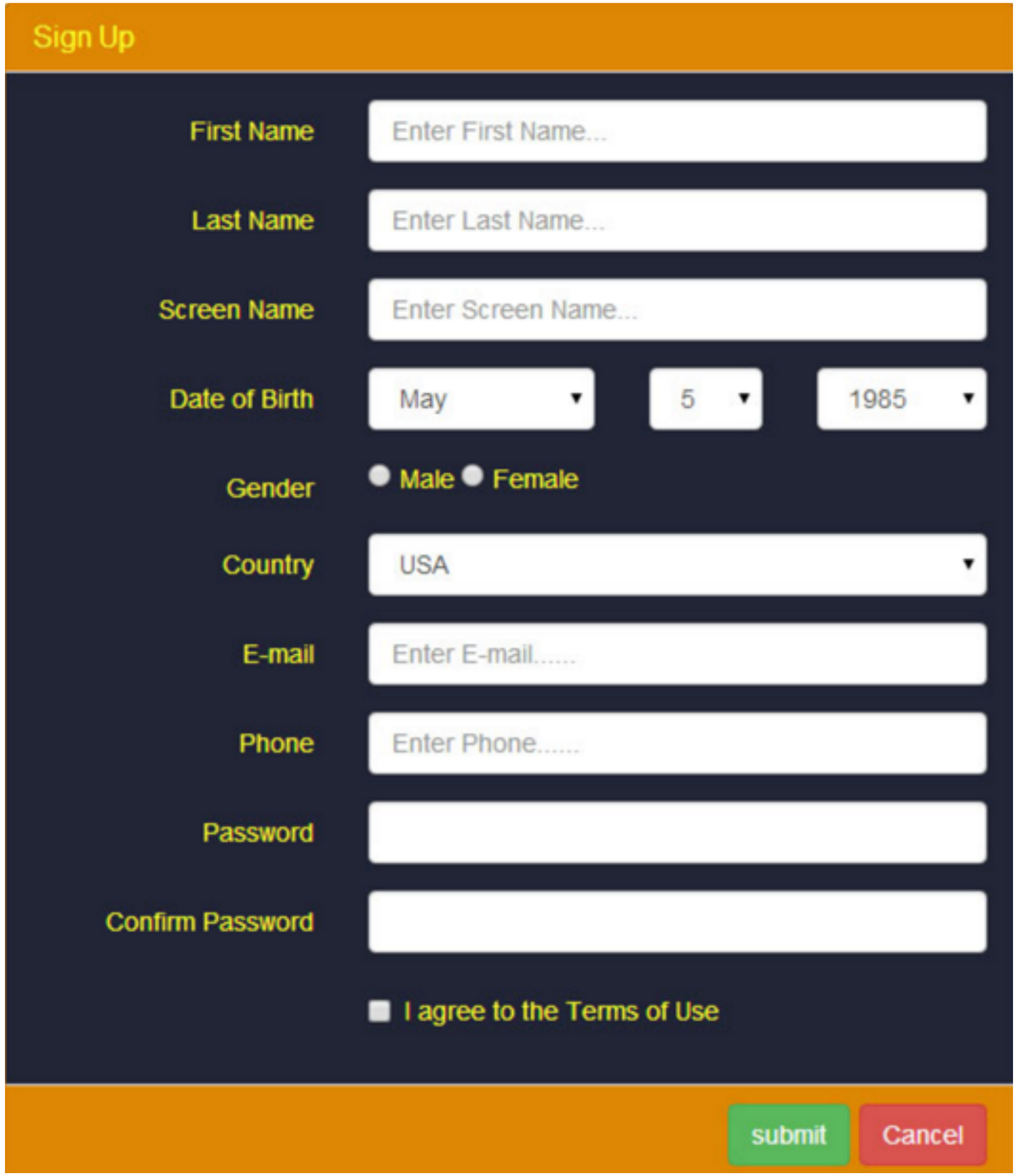
Decision Trees

K-Nearest Neighbors

Random Forests

Logistic Regression

Many online services collect customer information during the registration process



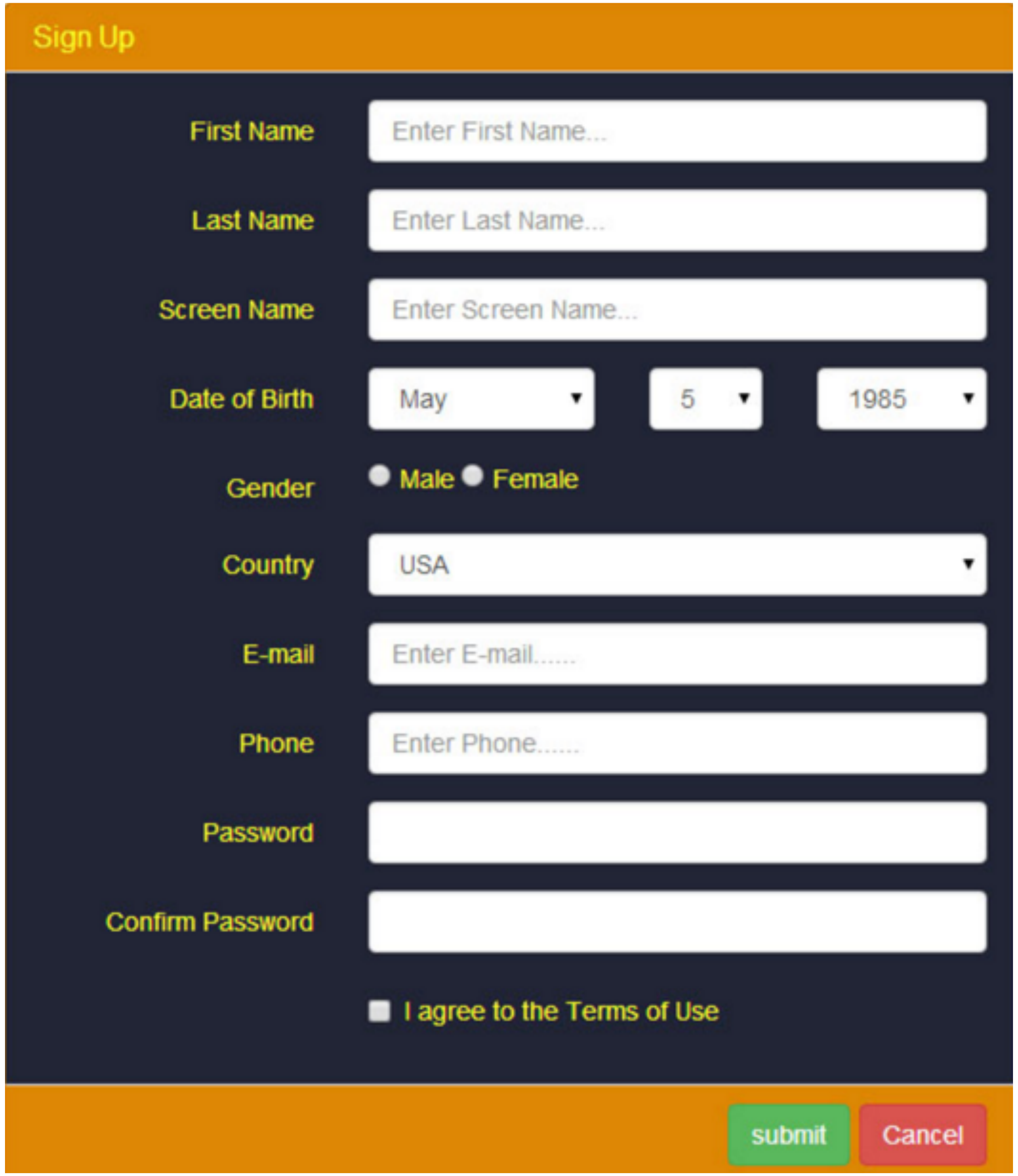
Sign Up

First Name	Enter First Name...
Last Name	Enter Last Name...
Screen Name	Enter Screen Name...
Date of Birth	May ▾ 5 ▾ 1985 ▾
Gender	○ Male ○ Female
Country	USA ▾
E-mail	Enter E-mail......
Phone	Enter Phone......
Password	
Confirm Password	

□ I agree to the Terms of Use

Knowing demographic information like gender

Can help the business create targeted offers for specific customers



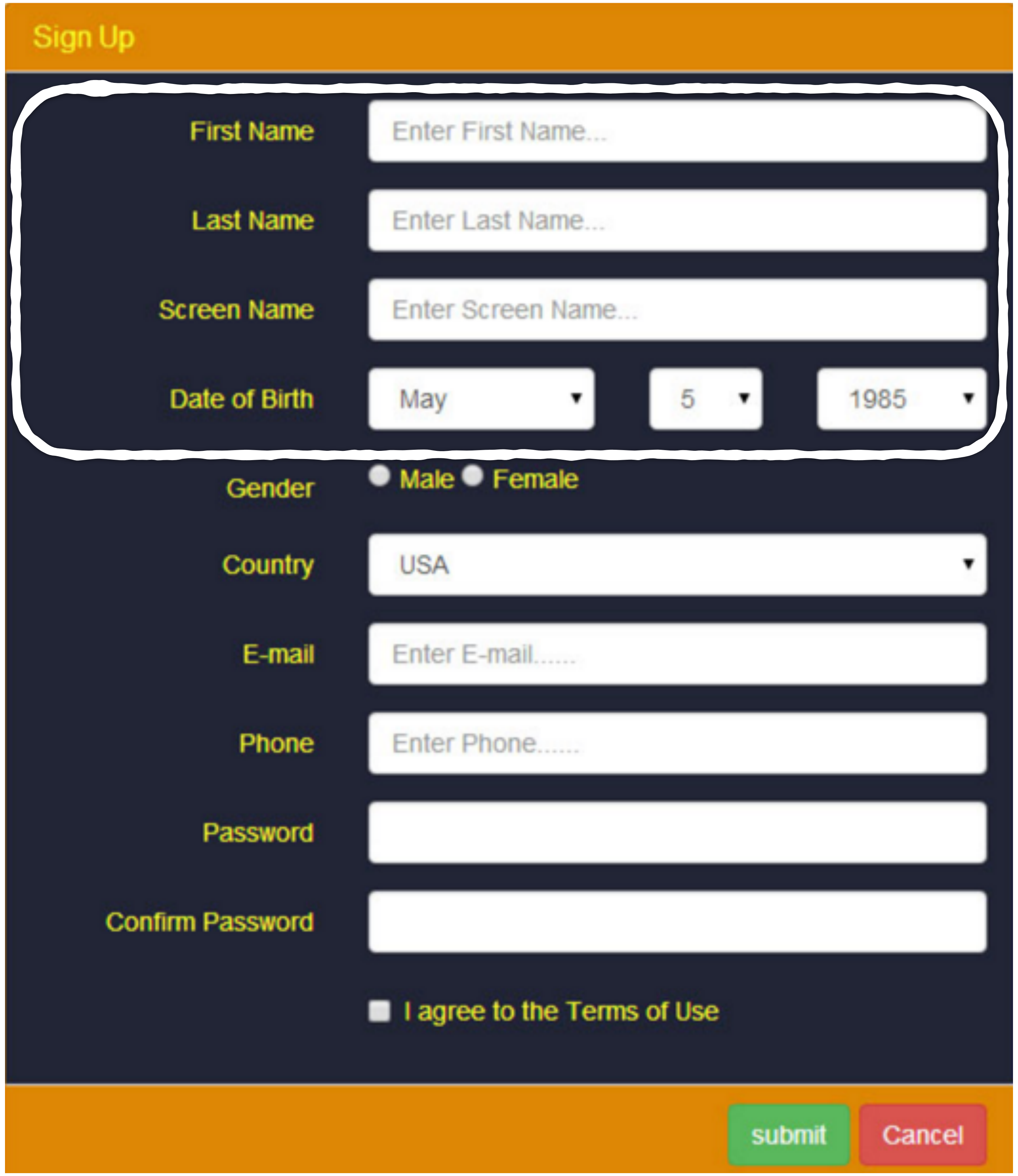
Sign Up

First Name	Enter First Name...
Last Name	Enter Last Name...
Screen Name	Enter Screen Name...
Date of Birth	May ▾ 5 ▾ 1985 ▾
Gender	○ Male ○ Female
Country	USA ▾
E-mail	Enter E-mail......
Phone	Enter Phone......
Password	
Confirm Password	

□ I agree to the Terms of Use

The problem

Folks will often fill in just the mandatory fields

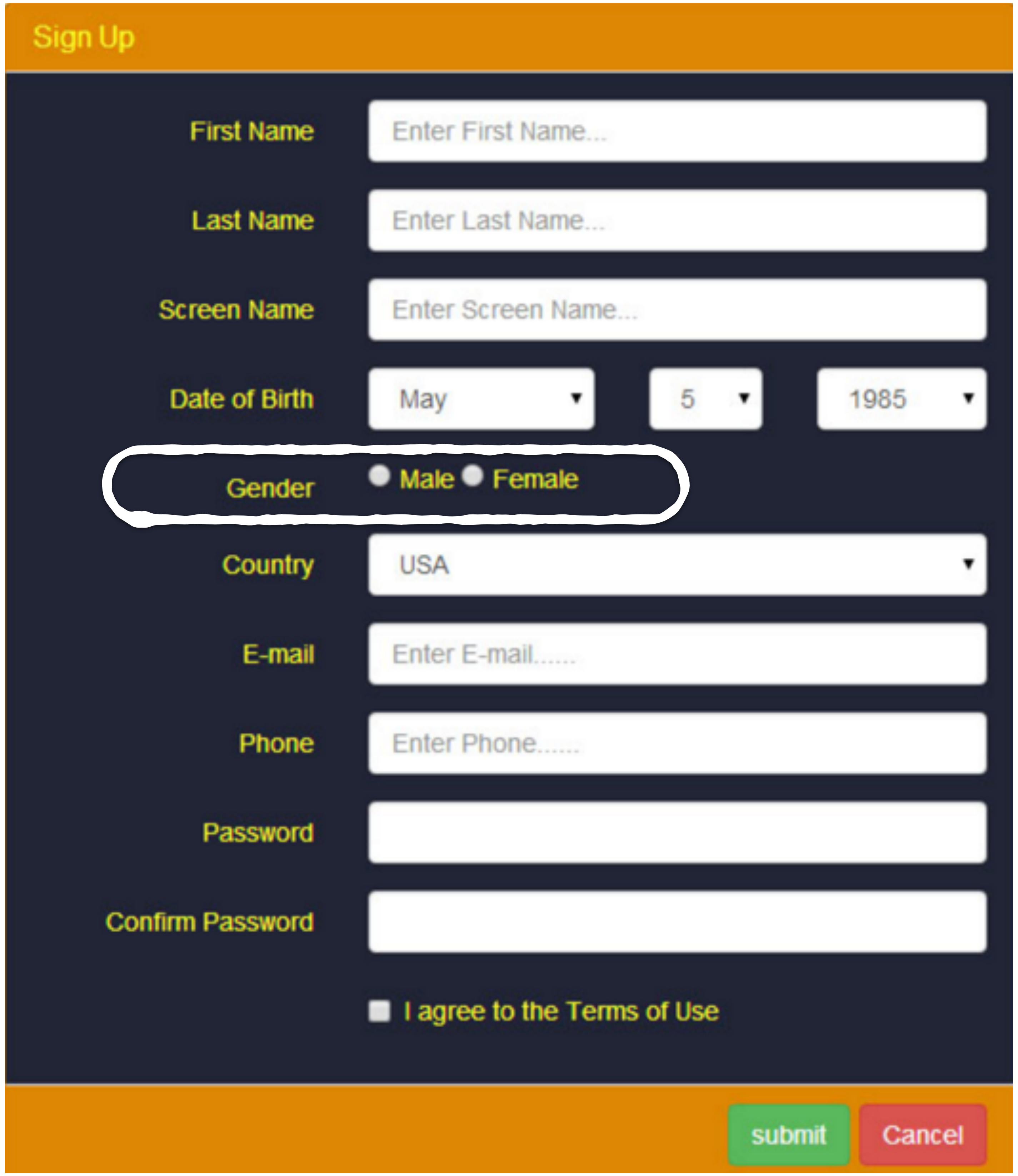


Sign Up

First Name	Enter First Name...
Last Name	Enter Last Name...
Screen Name	Enter Screen Name...
Date of Birth	May ▾ 5 ▾ 1985 ▾
Gender	○ Male ○ Female
Country	USA ▾
E-mail	Enter E-mail......
Phone	Enter Phone......
Password	
Confirm Password	

□ I agree to the Terms of Use

Only a fraction fill out all the fields



Sign Up

First Name: Enter First Name...

Last Name: Enter Last Name...

Screen Name: Enter Screen Name...

Date of Birth: May ▾ 5 ▾ 1985 ▾

Gender: ○ Male ○ Female

Country: USA ▾

E-mail: Enter E-mail......

Phone: Enter Phone......

Password:

Confirm Password:

□ I agree to the Terms of Use

Given the first name of a user

Can the system make a good guess ?

or

or

This can be set up as a **Classification problem**

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

“Train a model” using the training data

Test

“Test the model” using test data

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

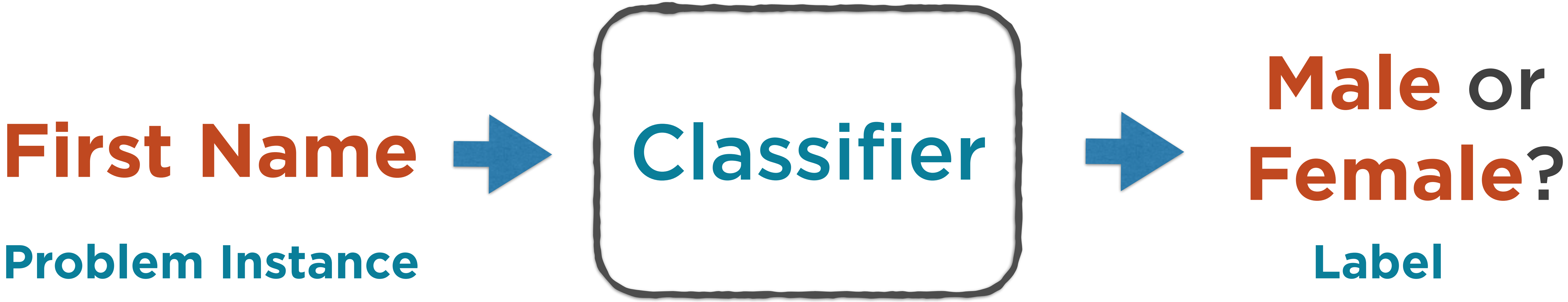
"Train a model" using the training data

Test

"Test the model" using test data

Let's set these up for Gender detection

Problem Statement



First Name

Problem Instance

Classifier

Male or Female?

Label

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

“Train a model” using the training data

Test

“Test the model” using test data

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

"Train a model" using the training data

Test

"Test the model" using test data

We need to represent Names using numeric attributes

Name

Use characteristics that usually differentiate male and female names

Name

Last letter a vowel? (1/0)

Number of characters

Presence of prefixes/suffixes common to a specific gender

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

"Train a model" using the training data

Test

"Test the model" using test data

Typical Classification Setup

Use the data of folks who did fill in their gender

Training

“Train a model” using the training data

Test

“Test the model” using test data

Typical Classification Setup

Feed the data to any standard Classification algorithm

Training

"Train a model" using the training data

Test

"Test the model" using test data

What is the market sentiment around Apple’s latest product launch?

How are voters feeling towards a particular candidate?

What do customers think about a particular brand?

The answers to all of these questions involve analyzing how people feel about something

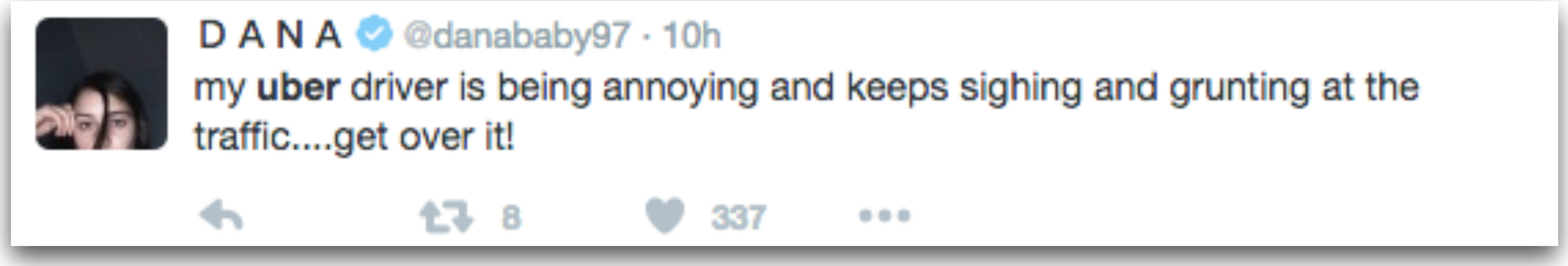
How people feel about something

🙂 or 🙁

can be measured using a technique known as

Sentiment Analysis

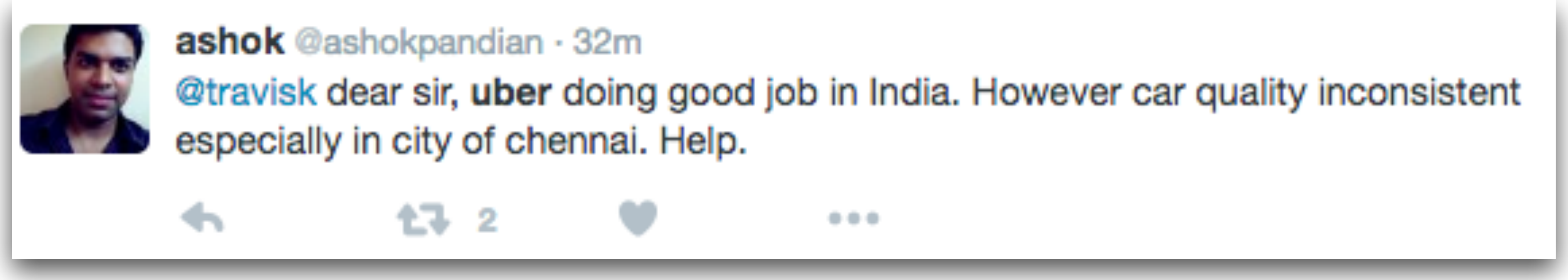
These days, folks express - all too freely and in public online forums - how they feel



DANA @danababy97 · 10h

my **uber** driver is being annoying and keeps sighing and grunting at the traffic....get over it!

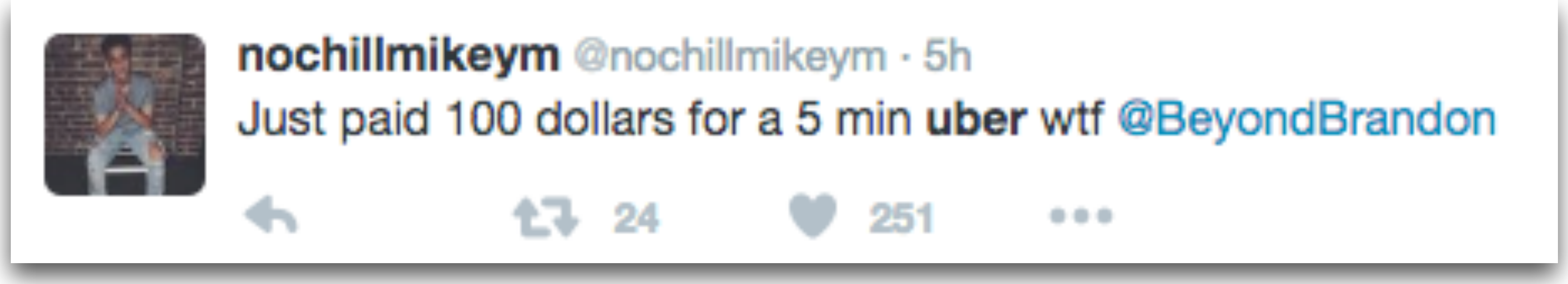
8 337



ashok @ashokpandian · 32m

@travisk dear sir, **uber** doing good job in India. However car quality inconsistent especially in city of chennai. Help.

2



nochillmikeym @nochillmikeym · 5h

Just paid 100 dollars for a 5 min **uber** wtf @BeyondBrandon

24 251

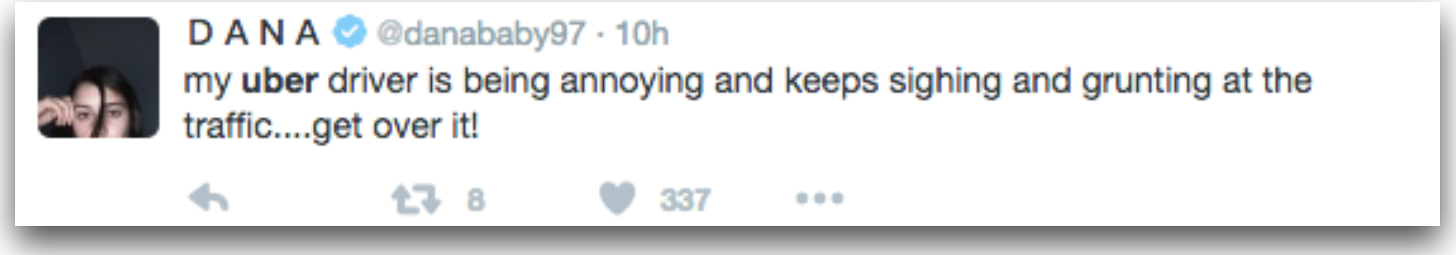
This data is

Huge (100s/1000s of tweets, reviews)

Unstructured

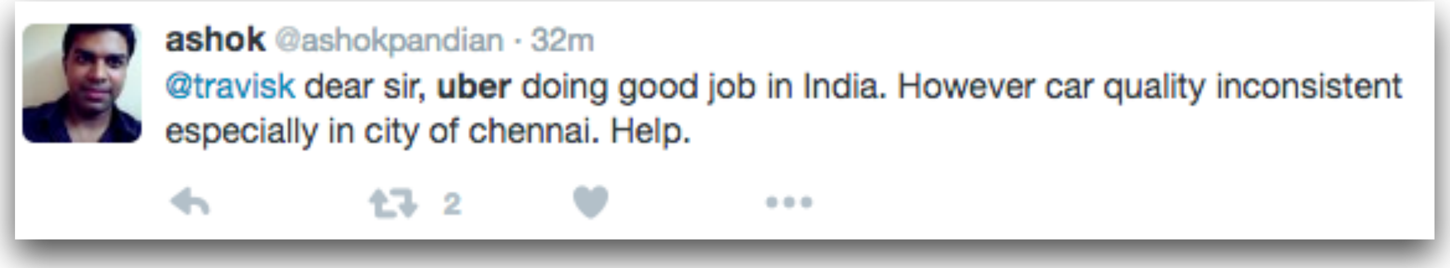
Semantically complicated

Freely and publicly available for anyone to analyse!



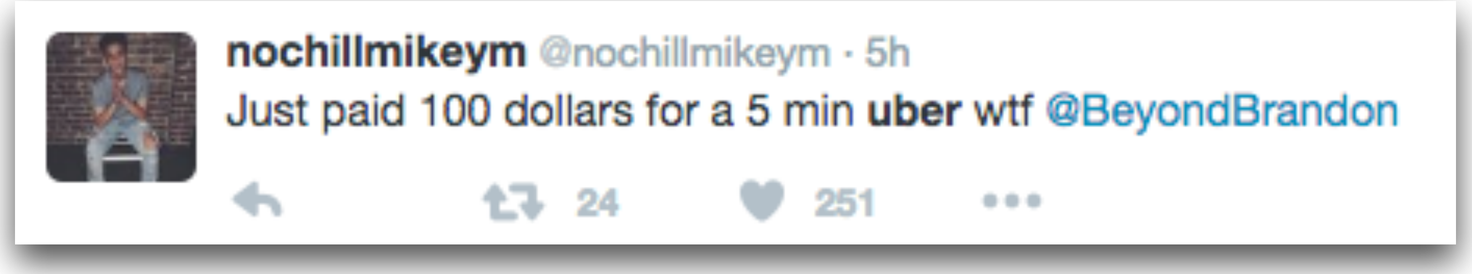
DANA @danababy97 · 10h

my **uber** driver is being annoying and keeps sighing and grunting at the traffic....get over it!



ashok @ashokpandian · 32m

@travisk dear sir, **uber** doing good job in India. However car quality inconsistent especially in city of chennai. Help.



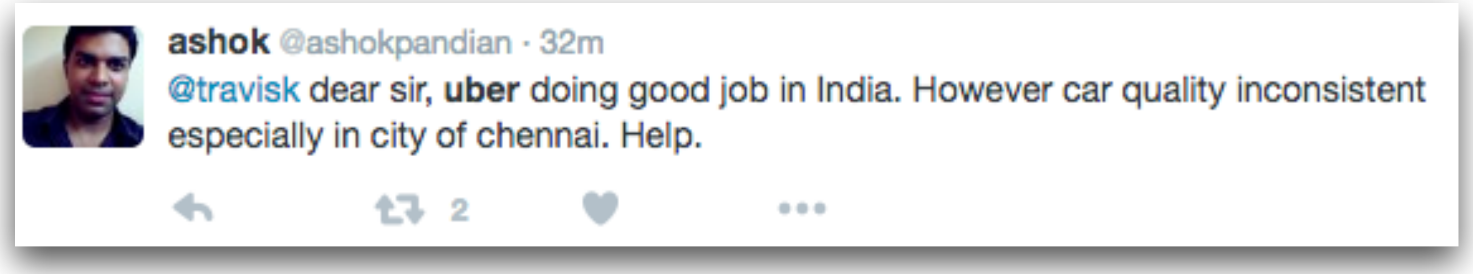
nochillmikeym @nochillmikeym · 5h

Just paid 100 dollars for a 5 min **uber** wtf @BeyondBrandon

To paraphrase Bill Gates, any big dataset is a learning opportunity - use Sentiment Analysis to seize it!

Sentiment Analysis

Positive

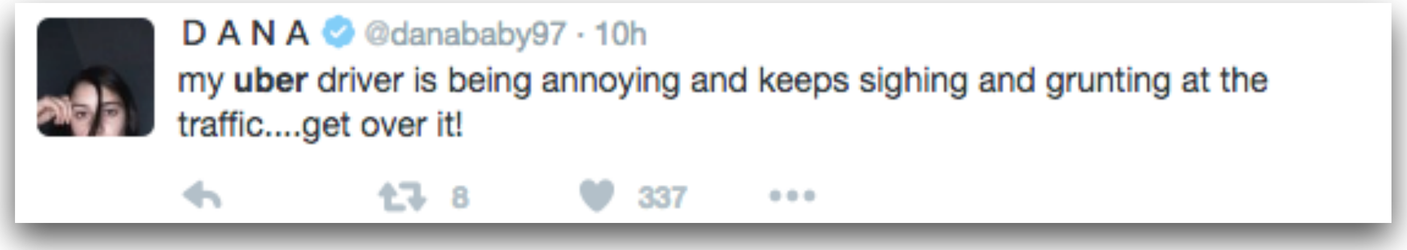


ashok @ashokpandian · 32m

@travisk dear sir, **uber** doing good job in India. However car quality inconsistent especially in city of chennai. Help.

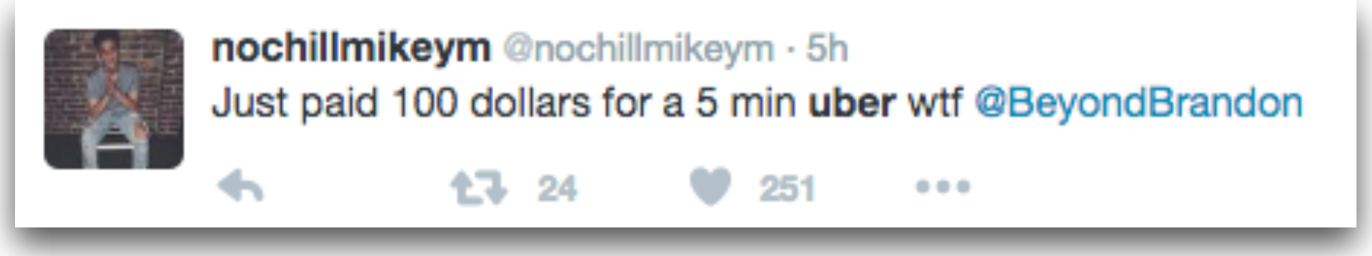
This comment is positive

Negative



D A N A @danababy97 · 10h

my **uber** driver is being annoying and keeps sighing and grunting at the traffic....get over it!



nochillmikeym @nochillmikeym · 5h

Just paid 100 dollars for a 5 min **uber** wtf @BeyondBrandon

These comments are negative

The Key Challenge

Positive or **Negative**

This is called

Identifying the Polarity

of a comment

Identifying the Polarity

Positive or **Negative**

This is a classic example of a Classification problem

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

“Train a model” using the training data

Test

“Test the model” using test data

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

"Train a model" using the training data

Test

"Test the model" using test data

Let's set these up for Sentiment Analysis

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

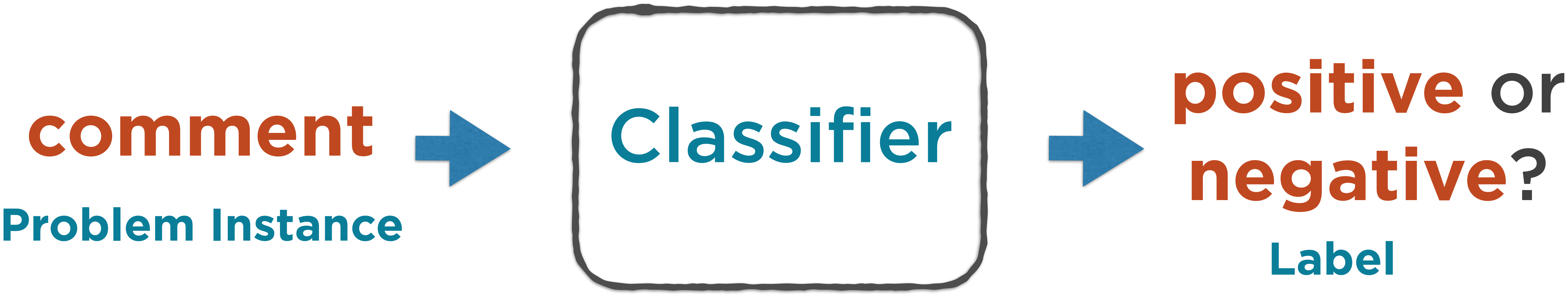
Training

"Train a model" using the training data

Test

"Test the model" using test data

Problem Statement



comment

Problem Instance

Classifier

positive or **negative**?

Label

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

"Train a model" using the training data

Test

"Test the model" using test data

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

“Train a model” using the training data

Test

“Test the model” [illegible] test data

We need to represent text data using numeric attributes

Features

Create a list representing the universe of all words that can appear in any text

$(W_1, W_2, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots W_N)$

(hello, this, is, the, universe, of, all, words, in, any, text, a, an, test, goodbye)

Any text can then be represented using the frequencies of these words

Features

Hello, this is a test

(hello, this, is, the, universe, of, all, words, in, any, text, a, an, test, goodbye)

(1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0)

Term Frequency Representation

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

“Train a model” using the training data

Test

“Test the model” using test data

Typical Classification Setup

Use a comments dataset where comments are already labelled as positive/negative

Training

“Train a model” using the training data

Test

“Test the model” using test data

Typical Classification Setup

Feed the data to any standard Classification algorithm

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

“Train a model” using the training data

Test

“Test the model” using test data

Quant Trading

Let's say you work for a hedge fund

You trade stocks on a Stock Exchange

Quant Trading

Buy

Sell

Quant Trading

Every morning, you need to decide

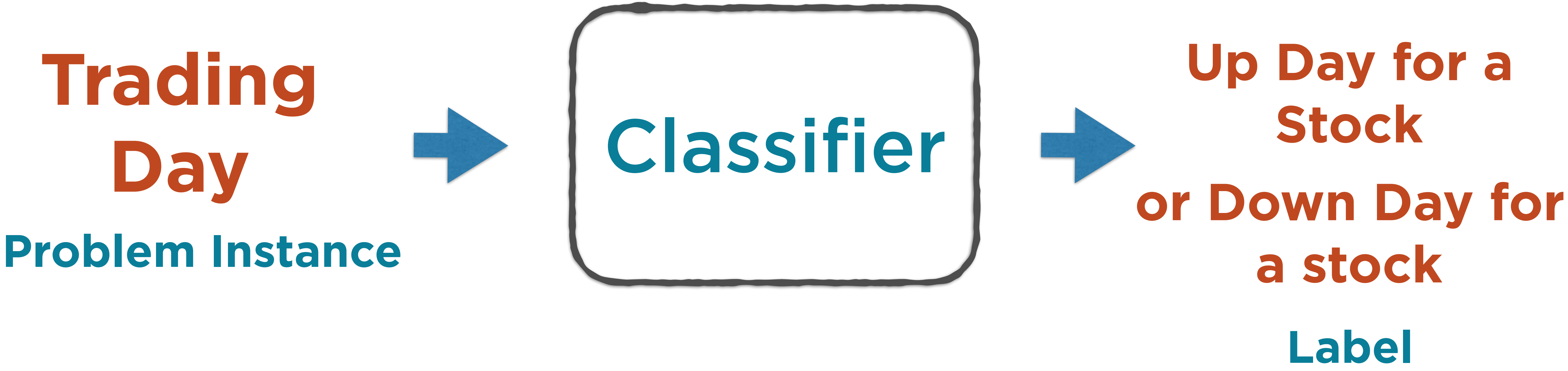
Buy Or Sell

Quant Trading

Buy Or Sell

This can be set up as a Classification problem

Problem Statement



Trading Day

Problem Instance

Classifier

Up Day for a Stock or Down Day for a stock

Label

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

"Train a model" using the training data

Test

"Test the model" using test data

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

"Train a model" using the training data

Test

"Test the model" using test data

We need to represent a Trading Day for a stock using numeric attributes

Trading Day Features

Day of the week

Month of the year

Price of the Stock on previous days

Price of related Stocks on previous days

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

"Train a model" using the training data

Test

"Test the model" using test data

Typical Classification Setup

Financial Data for the last 10 years

Represent each trading day as an up day or a down day for a stock

Training

“Train a model” using the training data

Test

“Test the model” using test data

Typical Classification Setup

Feed the data to any standard Classification algorithm

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

"Train a model" using the training data

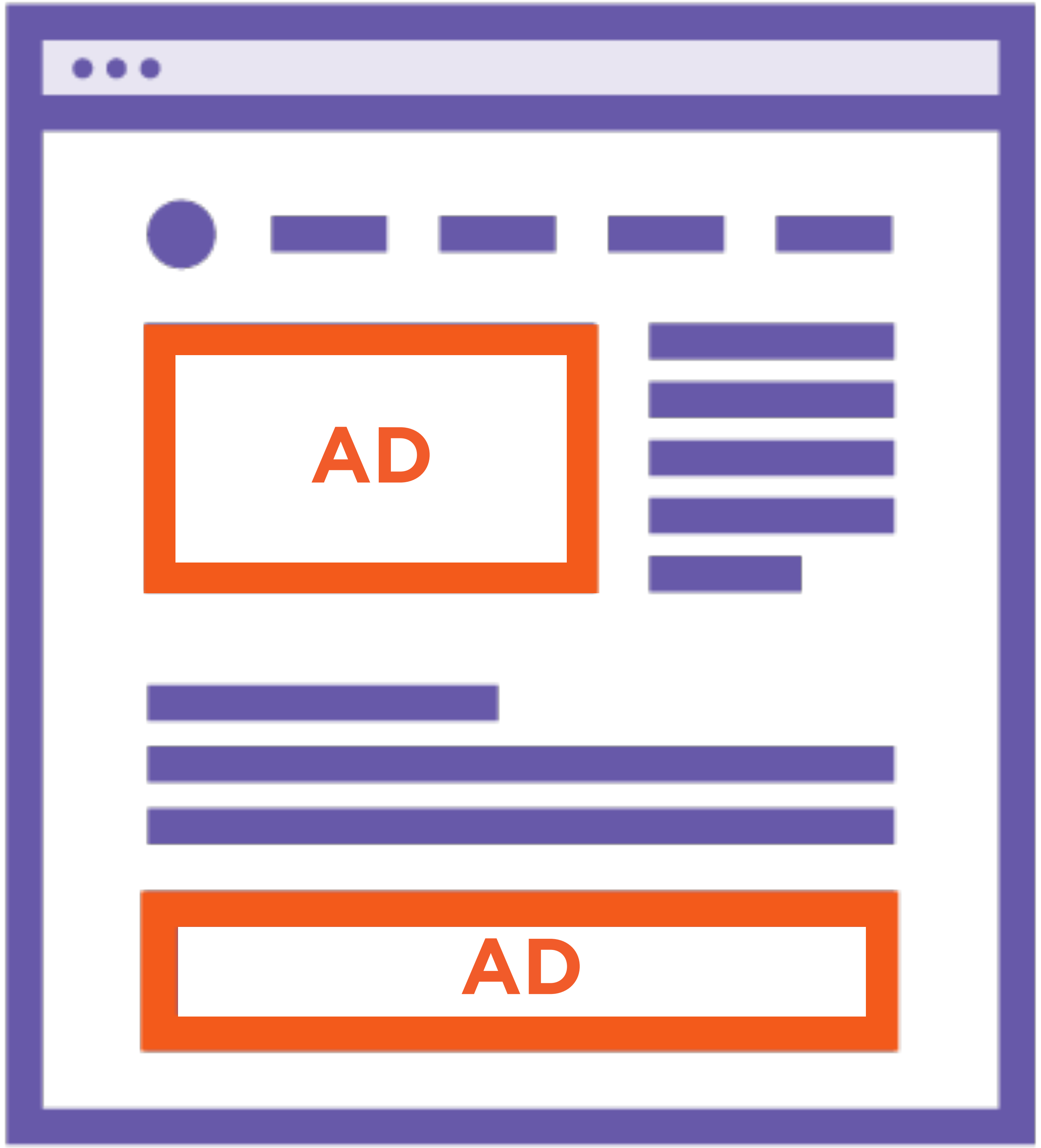
Test

"Test the model" using test data

Let’s say you want to build an ad-block extension for a browser

The browser has to render a number of images

Your extension should **block out** any ad images



This can be set up as a **Classification problem**

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

“Train a model” using the training data

Test

“Test the model” using test data

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

“Train a model” using the training data

Test

“Test the model” using test data

Let’s set these up for Ad Detection

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

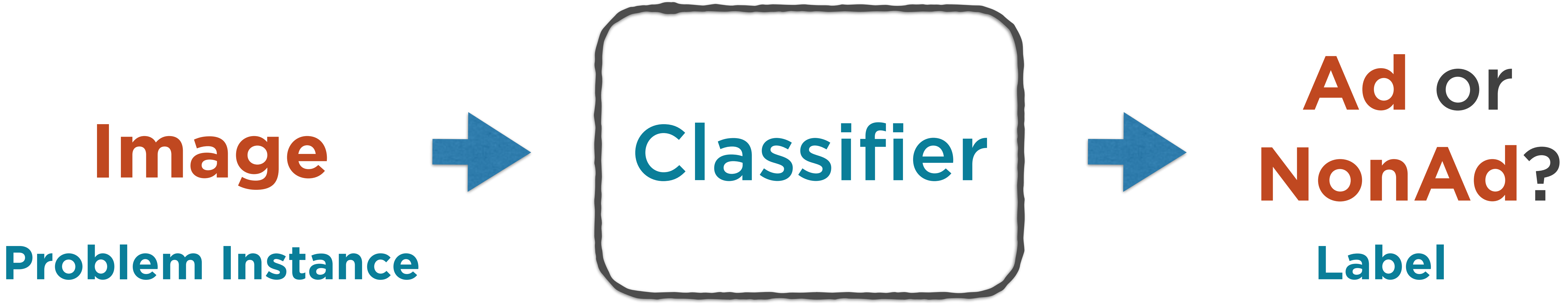
Training

"Train a model" using the training data

Test

"Test the model" using test data

Problem Statement



Image

Problem Instance

Classifier

Ad or **NonAd?**

Label

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

"Train a model" using the training data

Test

"Test the model" using test data

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Train

"Train a model" using the training data

Test

"Test the model" using test data

We need to represent an Image using numeric attributes

Image Features

Height, Width

Page URL

Image URL

Page text

Image Caption text

Image Features

Height, Width

Text attributes : Use a method like Term Frequency

Page URL

Image URL

Page text

Image Caption text

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

"Train a model" using the training data

Test

"Test the model" using test data

Typical Classification Setup

Use an image dataset where images are already labelled as Ad/NonAd

Training

“Train a model” using the training data

Test

“Test the model” using test data

Typical Classification Setup

Feed the data to any standard Classification algorithm

Training

"Train a model" using the training data

Test

"Test the model" using test data

Customer Behavior

Businesses often study customer activity to draw insights

Customer Churn

Does a customer’s behavior indicate that they will stop using our service in the future?

Fraud Detection

Does a customer’s behavior indicate that they are committing payment fraud?

Credit Risk

Does a customer’s behavior indicate that they are at risk of defaulting on their loan/payment?

Customer Churn

Fraud Detection

Credit Risk

Each of these can be set up as a **Classification problem**

Example 1: Customer Churn

Problem Instance	A Customer
Labels	Will repurchase, will not repurchase
Features	Purchases, demographics, days since last purchase
Training Data	A large number of customers categorized as repurchased, did not repurchase

Example 2: Fraud Detection

Problem Instance	A Payment
Labels	Fraud or Not Fraud
Features	Payment type, Frequency of use, Failed attempts in the last hour
Training Data	A large number of historical transactions

Example 3: Credit Risk

Problem Instance	A Customer
Labels	Will default payment, will not default payment
Features	Income, education, employment sector, history of defaults
Training Data	Past customers labelled as Defaulted/Did not Default

Summary

Recognize Classification problems in different fields : from Spam Detection to Quant Trading

Set up all the elements of a classification problem : Problem statement, Features, Labels