Building Sentiment Analysis Systems in Python

IDENTIFYING APPLICATIONS OF SENTIMENT ANALYSIS

Vitthal Srinivasan
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

Recognise applications of sentiment analysis

Frame sentiment analysis as a binary classification problem

Contrast ML-based and rule-based approaches to sentiment analysis

Sentiment Analysis Introduced

Changing Patterns of Online Behavior

“Surf/Browse”

c. 1990 - c. 2000

“Search-Find-Obtain”

c. 2000 - c. 2008

“Share-Discover”

c. 2008 - Present

“Share-Discover”

c. 2008 - Present

Always online

Share with network

Discover through network

Stream of online opinions

Opinions Contain Information

Reviews

Tweets and Posts

Messages

Swipes

Data Analyst

Collect opinions

Extract information from them

Act on that information

Changing Patterns of Online Behavior

Collect Opinions

Scrape/harvest comments, articles, tweets...

Extract Information

This is sentiment analysis

Act

Buy/sell stocks, target advertising spend,...

Collect Opinions

Researchers use public datasets

Companies use proprietary data

Scrapers use media signals

“Big Data”

Unstructured data

Extract Information

Tag data item with values for sentiments

One/more categorical data series created

Analyse categorical data

Extract Information

Data item to analyse

Tweet, email, message, review, …

Sentiment identified

“Positive”, “Negative”,”Neutral”

Categorical variable

+1, 0, -1

Analysing Categorical Sentiment Data

Logistic Regression

Relationships between variables

Quadrant Analysis

Clusters of data with similar characteristics

Act

Trade financial markets

Change or reallocate ad budgets

Tailor electoral strategy

Decide product recall strategies

Applications of Sentiment Analysis

Changing Patterns of Online Behavior

Collect Opinions

Scrape/harvest comments, articles, tweets...

Extract Information

This is sentiment analysis

Act

Buy/sell stocks, target advertising spend...

Sentiment Analysis in Event-Driven Trading

Analyst Sentiment Before Earnings

Company Earnings, versus Forecast

Negative	Positive	
		Exceeded Forecast
		Missed Forecast

Sentiment Analysis in Event-Driven Trading

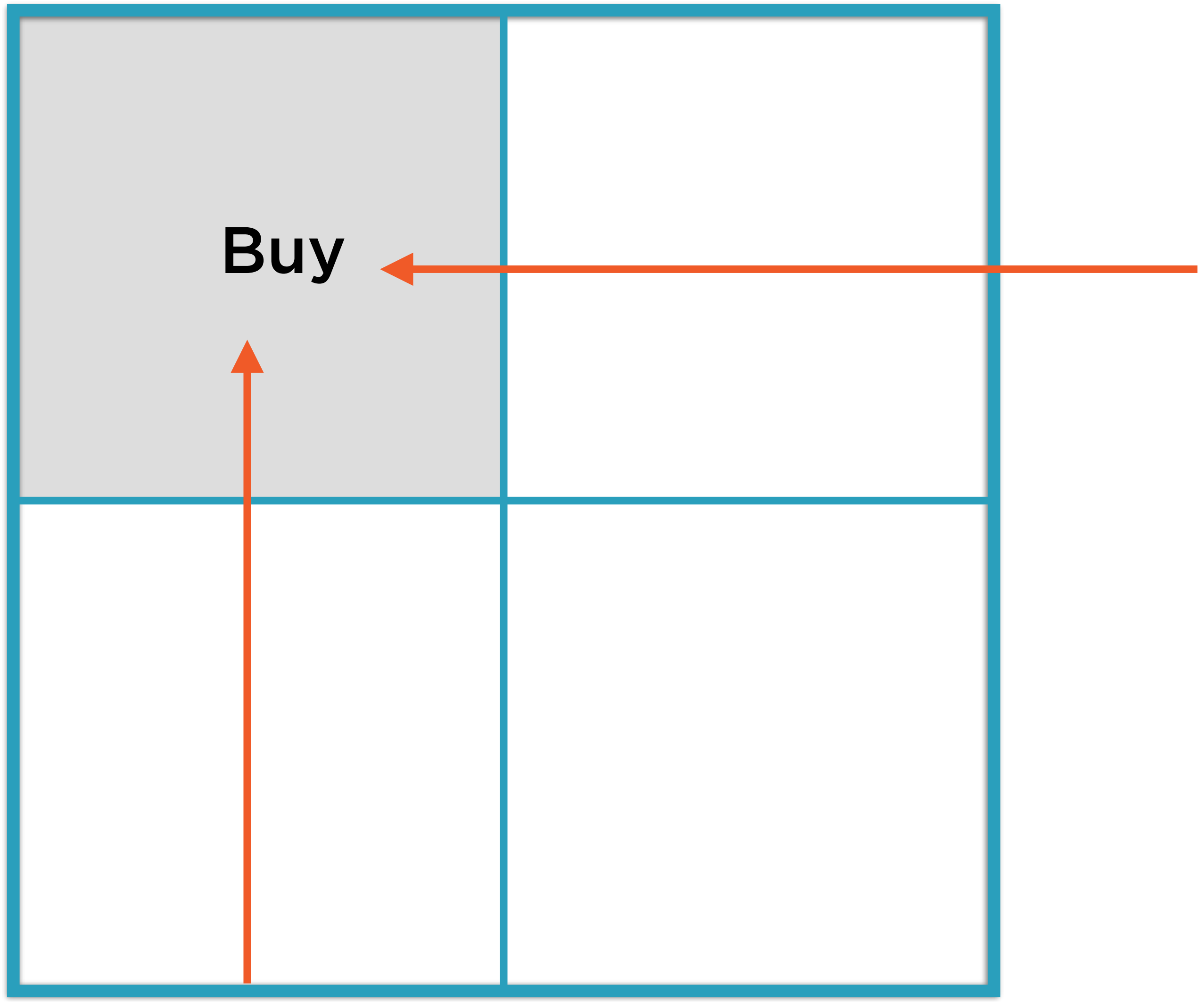
Company Earnings Releases

Better or worse than analyst expectations?

Financial Traders

Buy or sell?

Sentiment Analysis in Event-Driven Trading



Analyst Sentiment Before Earnings

Company Earnings, versus Forecast

Buy

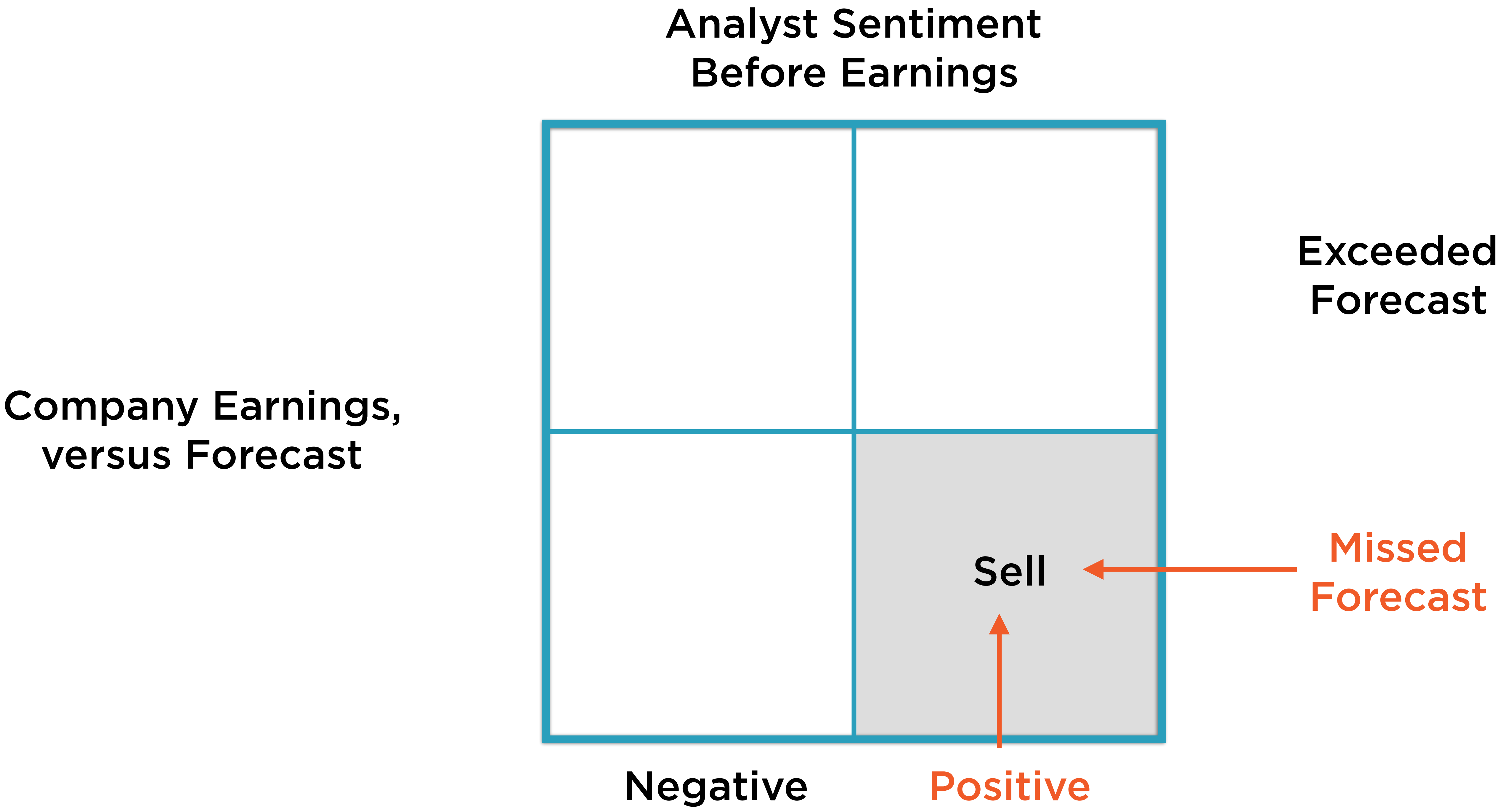
Exceeded Forecast

Missed Forecast

Negative

Positive

Sentiment Analysis in Event-Driven Trading



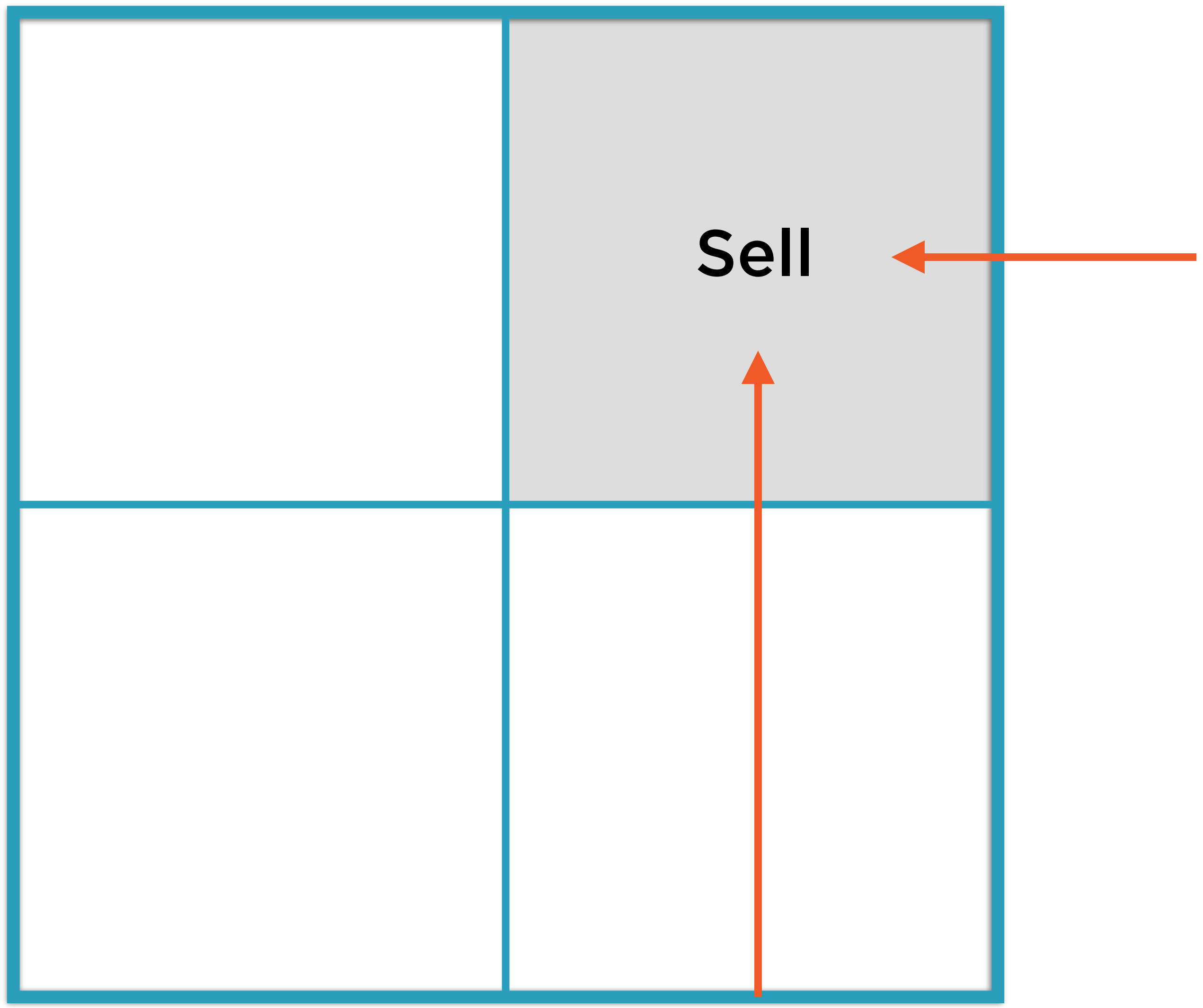
Analyst Sentiment Before Earnings

Company Earnings, versus Forecast

	Negative	Positive
Exceeded Forecast		
Missed Forecast		Sell

Sentiment Analysis in Event-Driven Trading

Analyst Sentiment
Before Earnings



Company Earnings,
versus Forecast

Sell

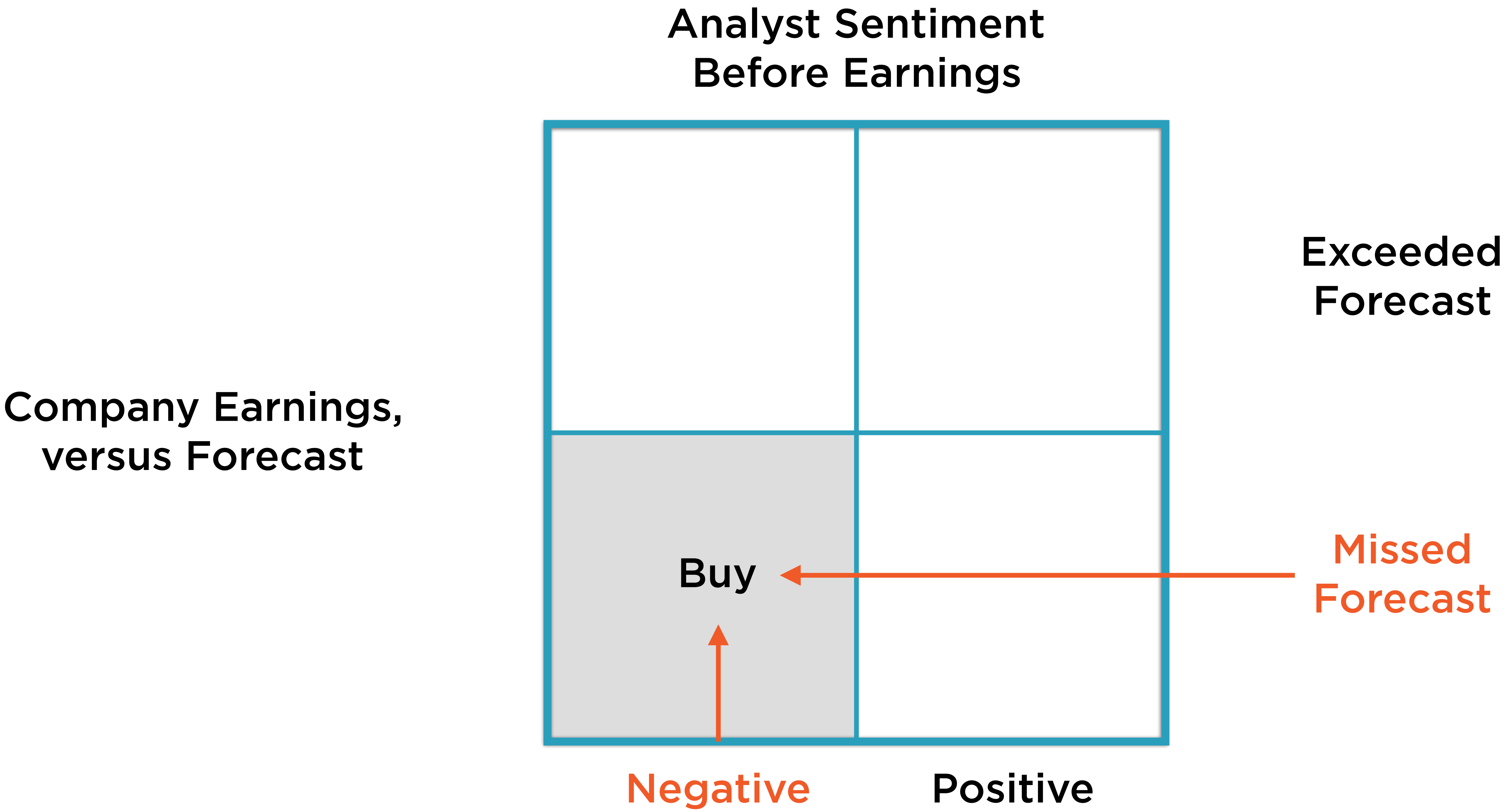
Exceeded
Forecast

Missed
Forecast

Negative

Positive

Sentiment Analysis in Event-Driven Trading



Analyst Sentiment Before Earnings

Company Earnings, versus Forecast

	Negative	Positive
Exceeded Forecast		
Missed Forecast	Buy	

Insight: “Buy the Rumor, Sell the News”

Buy the rumor

If market sentiment was negative, buy even if earnings are poor

Sell the news

If market sentiment was positive, sell even if earnings are great

Sentiment Analysis in Customer Satisfaction

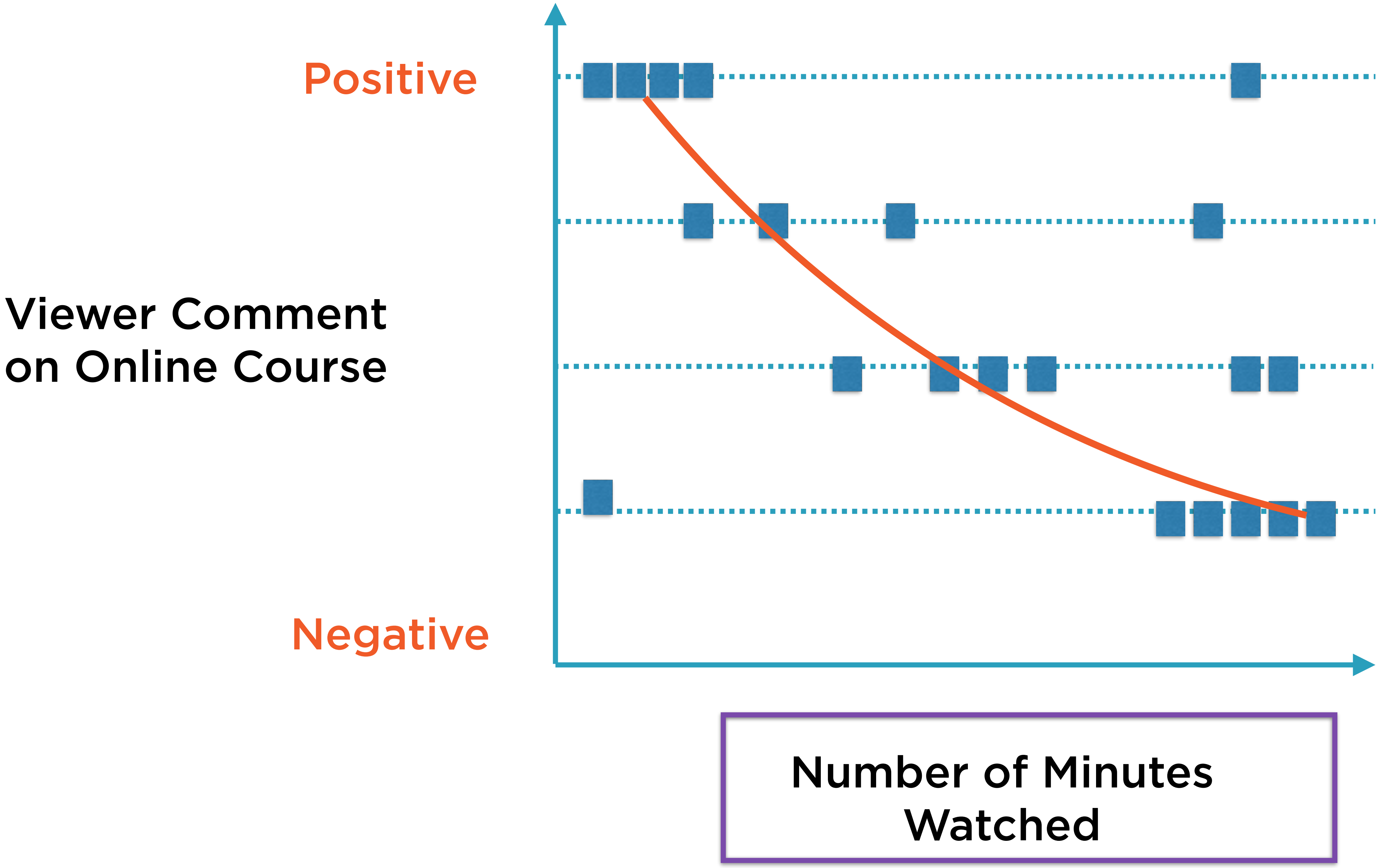
User Messages on Learning Platform

Irritated or satisfied?

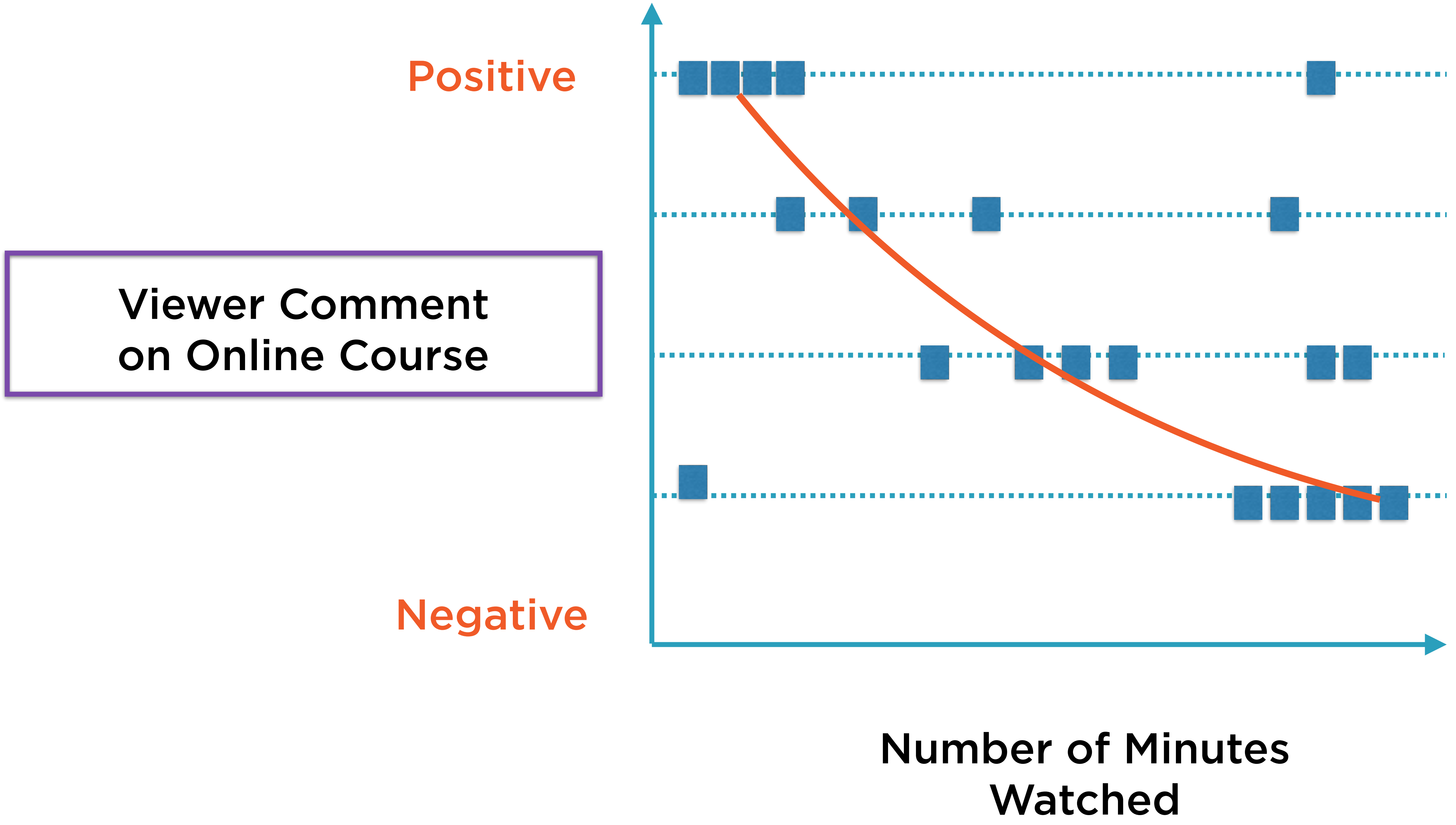
Minutes Watched of Online Content

Engaged or checked out?

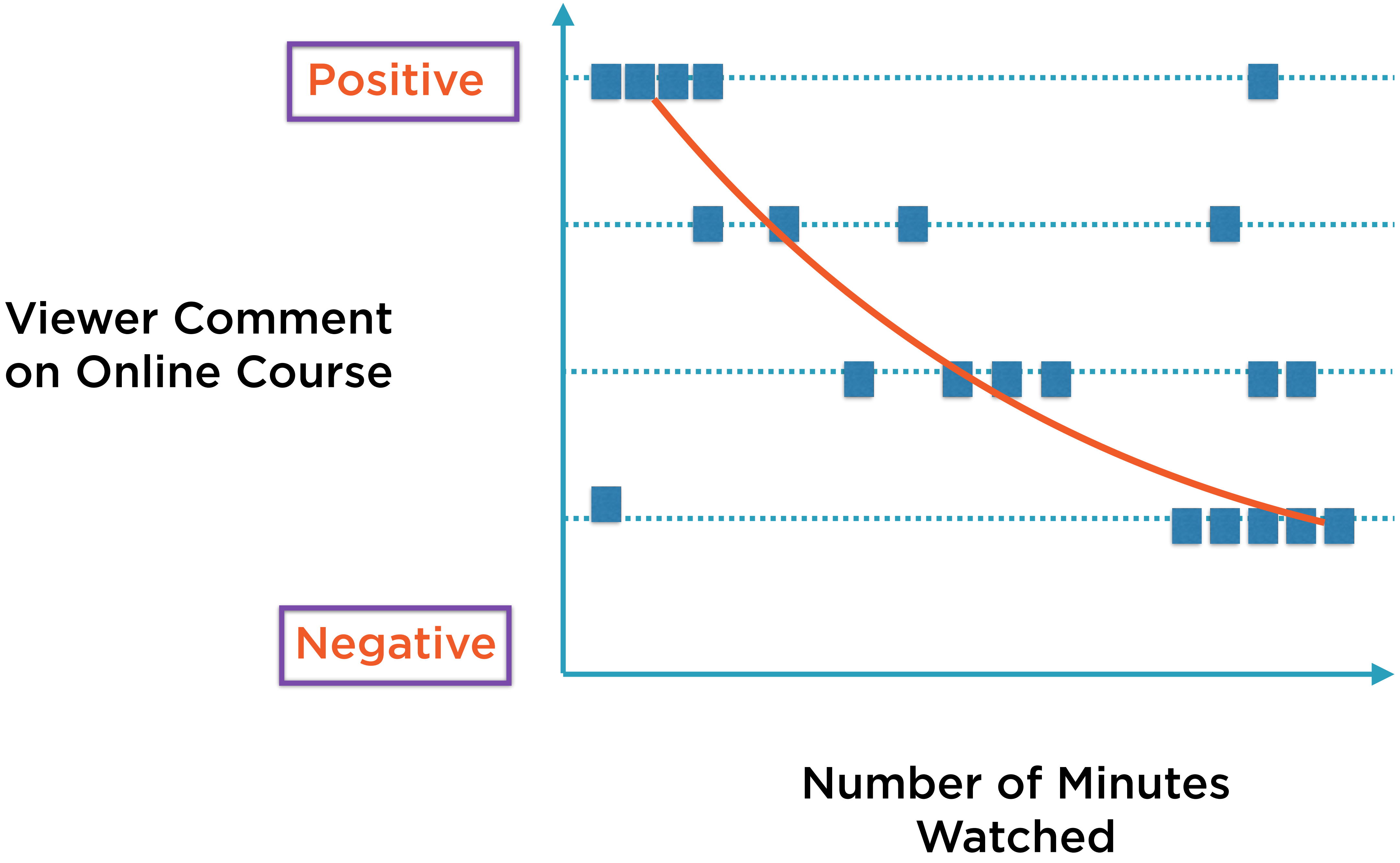
Sentiment Analysis in Customer Satisfaction



Sentiment Analysis in Customer Satisfaction



Sentiment Analysis in Customer Satisfaction



Insight: Fix the Finish

Strong start

The module starts well and makes a strong first impression

Weak finish

The latter part of the course fails to hold viewer interest

Polarity Detection for Sentiment Analysis

Sentiment Analysis Systems

Polarity

Positive or negative?

Subjectivity

Subjective or objective?

Aspects

Part or whole?

Opinions Are Very Complex

But sentiment analysis need not be (if we set up the problem right)

Either-or Decisions Are Simple

Human brains are very efficient at making binary decisions

Binary Decisions

Hot or not?

Buy or sell?

Fight or flight?

For or against?

Opinions Are Very Complex

Model sentiment analysis as a Binary Classification problem

Binary Classification

Positive

Not Positive

Model sentiment analysis as a Binary Classification problem

Binary Classification

Positive

Not Positive

Binary classification is a well-studied, well-understood problem

Binary Decisions

Comment: Positive or negative?

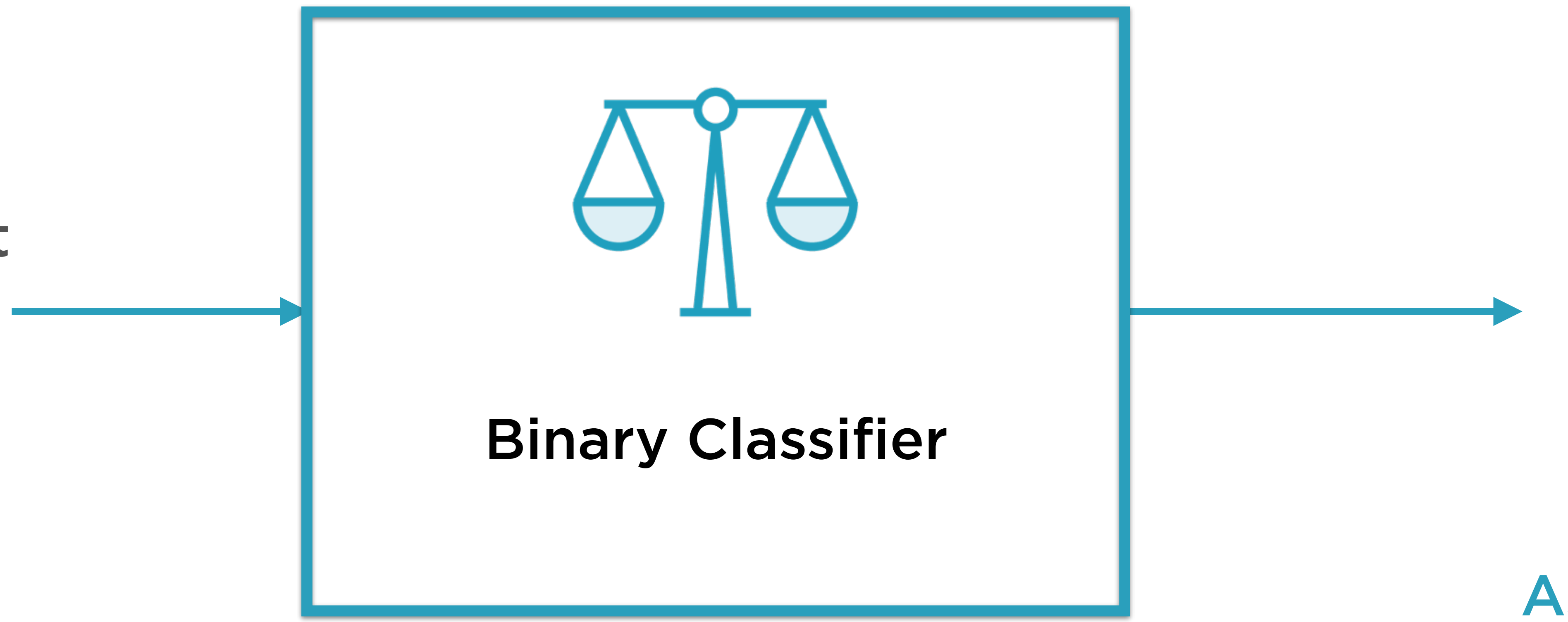
Email: Spam or ham?

Transactions: Fraud or legit?

Sentiment Analysis as Binary Classification

"This is not the worst restaurant in the metropolis, not by a long way"

Problem Instance



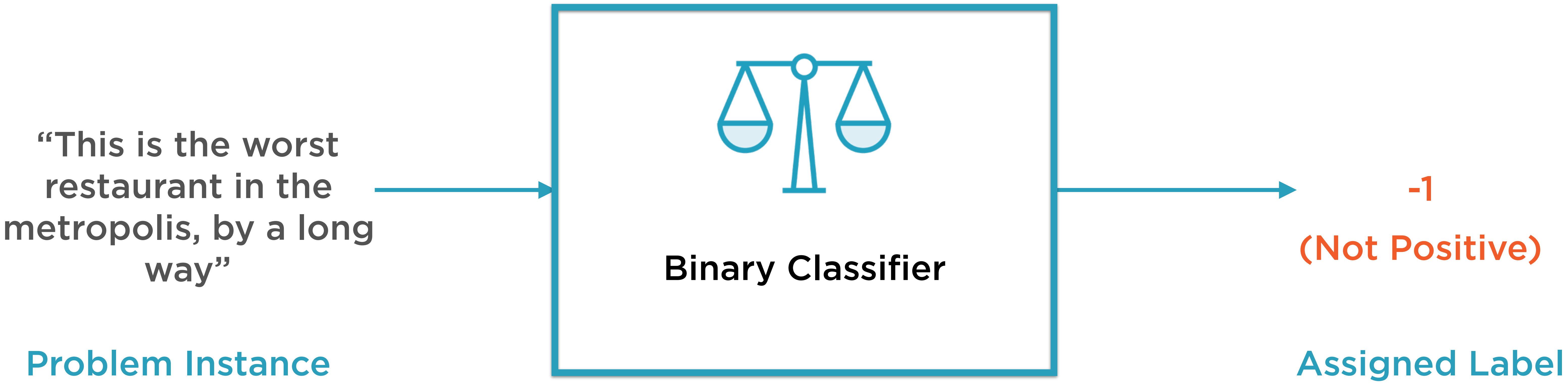
Binary Classifier

+1

(Positive)

Assigned Label

Sentiment Analysis as Binary Classification



"This is the worst restaurant in the metropolis, by a long way"

Problem Instance

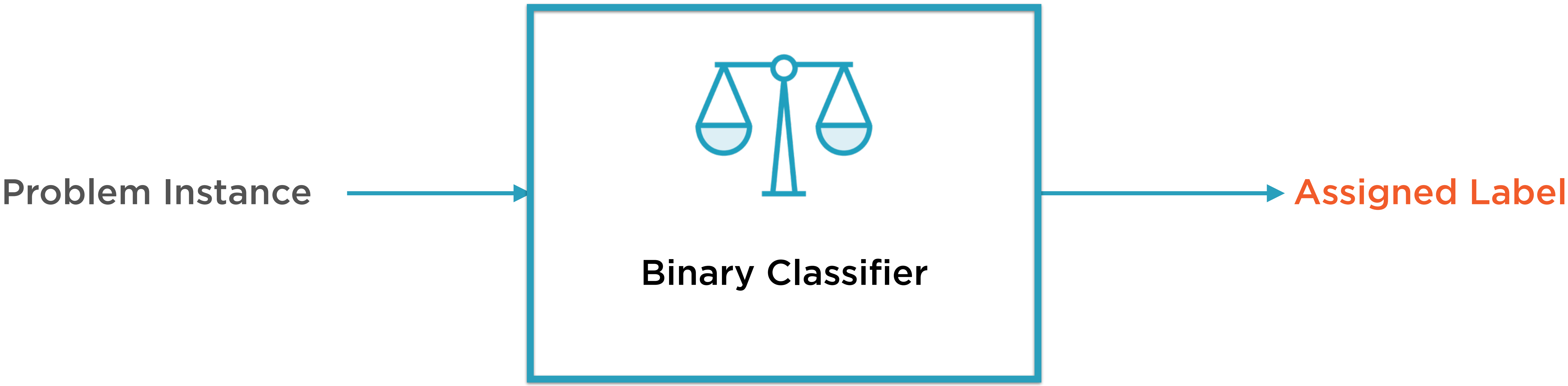
Binary Classifier

-1

(Not Positive)

Assigned Label

Sentiment Analysis as Binary Classification



Setting up a Binary Classification Problem

```
“This is not the worst restaurant in the metropolis, not
by a long way”
```

Problem Instance

The data item to be classified - usually unstructured text

“This is ~~not~~ the worst restaurant in the metropolis, ~~not~~ by a long way”

Problem Instance

The data item to be classified - usually unstructured text

“This is the worst restaurant in the metropolis, by a long way”

Another Problem Instance

The data item to be classified - usually unstructured text

“This is not the worst restaurant in the metropolis, not by a long way”

```
(“This”, “is”,”not”,”the”,”worst”,”restaurant”,”in”,”the”,
“metropolis”, “not”,”by”,”a”,”long”,”way”)
```

Feature Vector: Word Tuple

Any representation of the attributes of the problem instance is called a feature vector

```
“This is not the worst restaurant in the metropolis, not
by a long way”
```

```
{“This”:1, “is”:1,”not”:2,”the”:2,”worst”:1,”restaurant”:
1,”in”:1, “metropolis”:1, “by”:1,”a”:1,”long”:1,”way”:1}
```

Feature Vector: Word Frequency Set

A different representation - setting up the feature vector correctly is quite a skill

“This is not the worst restaurant in the metropolis, not by a long way”

{~~“This”:1, “is”:1,~~”not”:2,~~”the”:2,~~”worst”:1,”restaurant”:1,~~”in”:1,~~ “metropolis”:1, ~~“by”:1,”a”:1,~~”long”:1,”way”:1}

Feature Vector: Word Frequency Set

A different representation - setting up the feature vector correctly is quite a skill

“This is not the worst restaurant in the metropolis, not by a long way”

```
{”not”:2,”worst”:1,”restaurant”:1,“metropolis”:1,”long”:
1,”way”:1}
```

Feature Vector: Stop Words Eliminated

Yet another version, this one eliminates common words called stop words

`{Positive, Not Positive}`

Category Set

The set with the two values - need to find which value applies to the problem instance

{+1, -1}

Categorical Variable Values

Numeric values are often assigned to each category label - handy for use in logistic regression

“This is not the worst restaurant in the metropolis, not by a long way”

Positive

Assigned Label

The category that the problem instance belongs to - as decided by the classifier

“This is the worst restaurant in the metropolis, by a long way”

Negative

Assigned Label

The category that the problem instance belongs to - as decided by the classifier

…

“This is not the worst restaurant in the metropolis, not by a long way”

“This is the worst restaurant in the metropolis, by a long way”

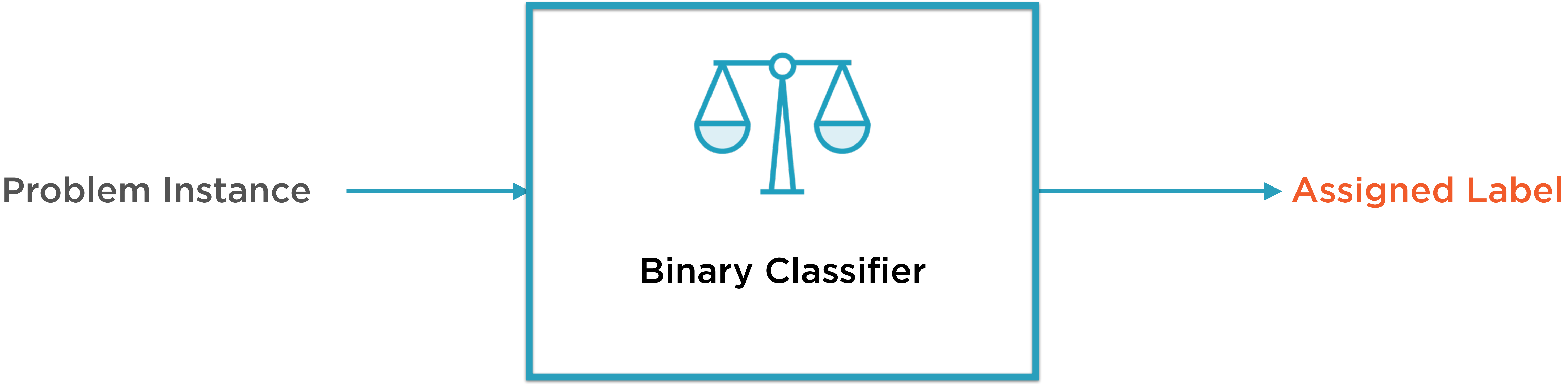
…

Corpus

A large number of data items, collectively available to the classifier

Rule-based and ML-based Binary Classifiers

Sentiment Analysis as Binary Classification



The binary classifier is a function that takes in a problem instance, and assigns a label

Binary Classifiers

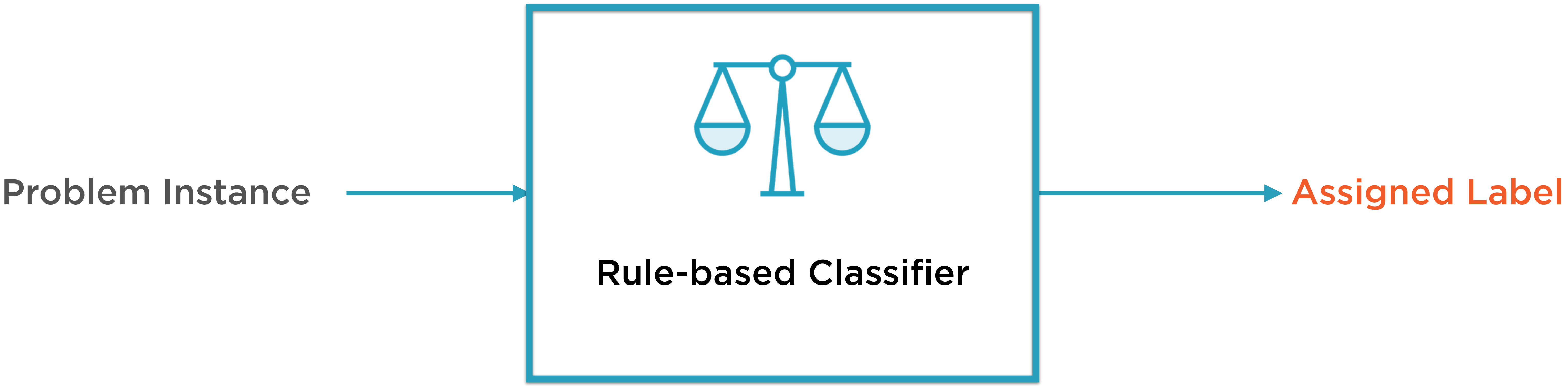
Rule-based Classifiers

Rules drawn up by experts are used to assign a label to problem instance

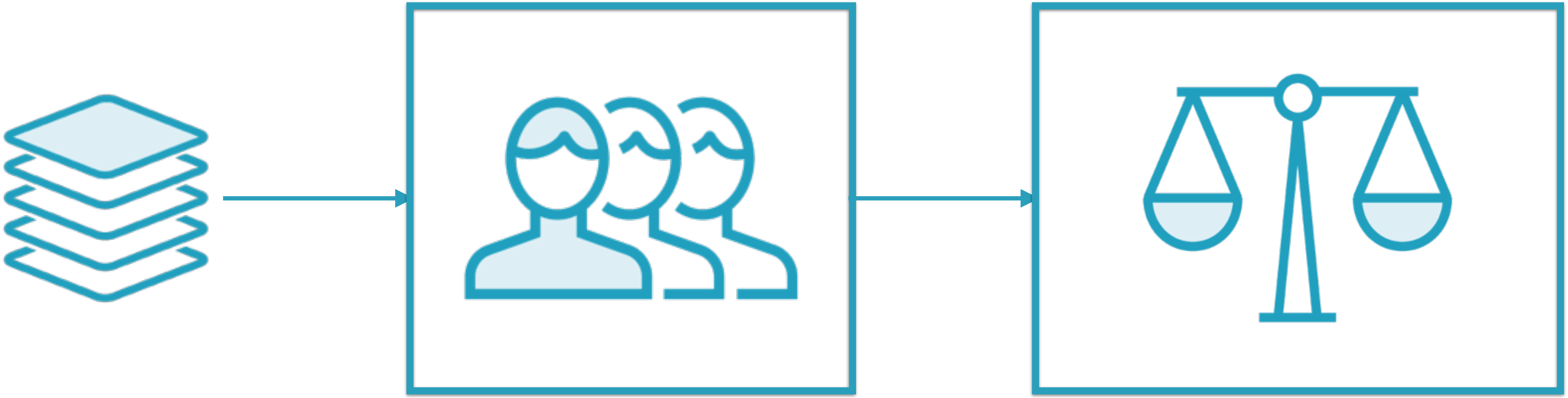
ML-based Classifiers

Label is assigned based on patterns displayed in aggregate data

Rule-based Binary Classifier



Rule-based Binary Classifier

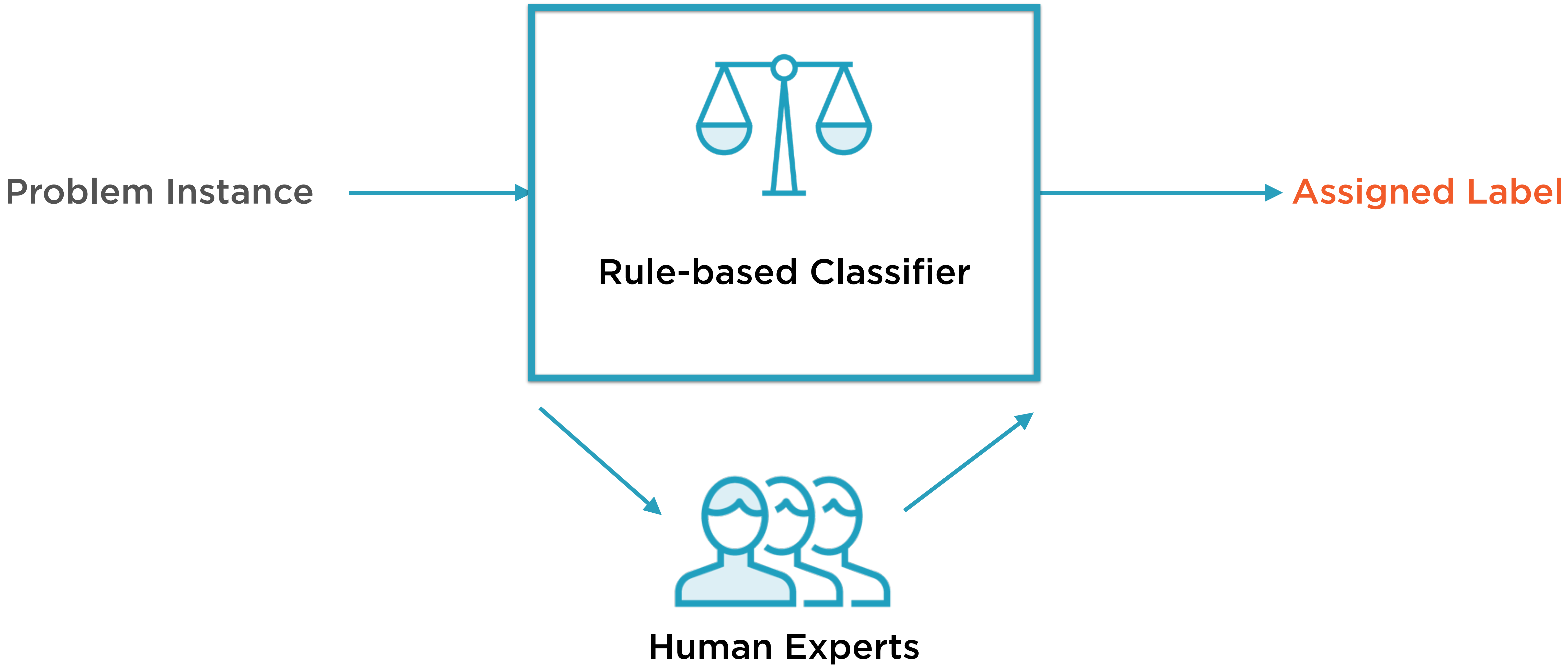


Corpus

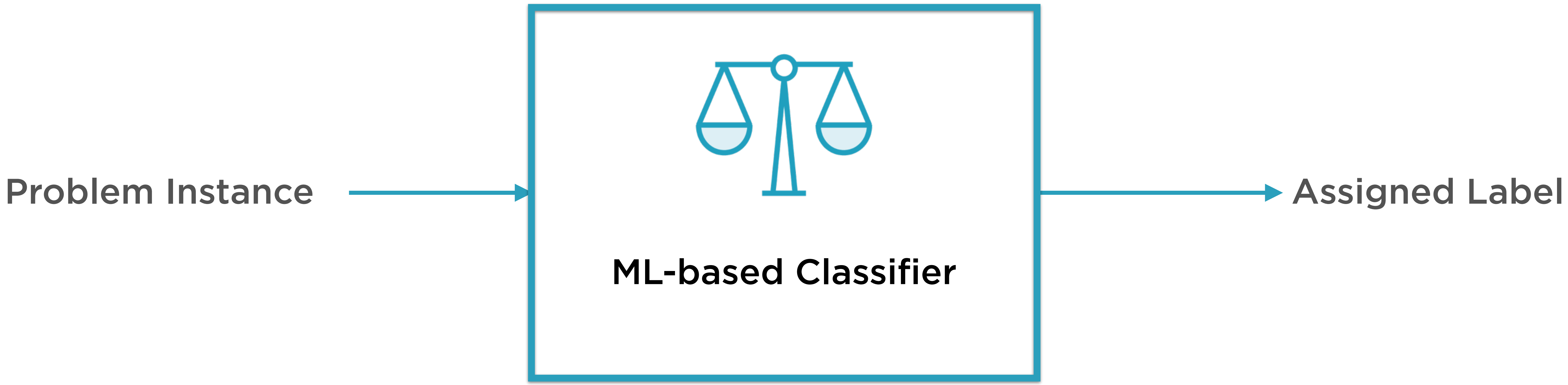
Human Experts

Rule-based Classifier

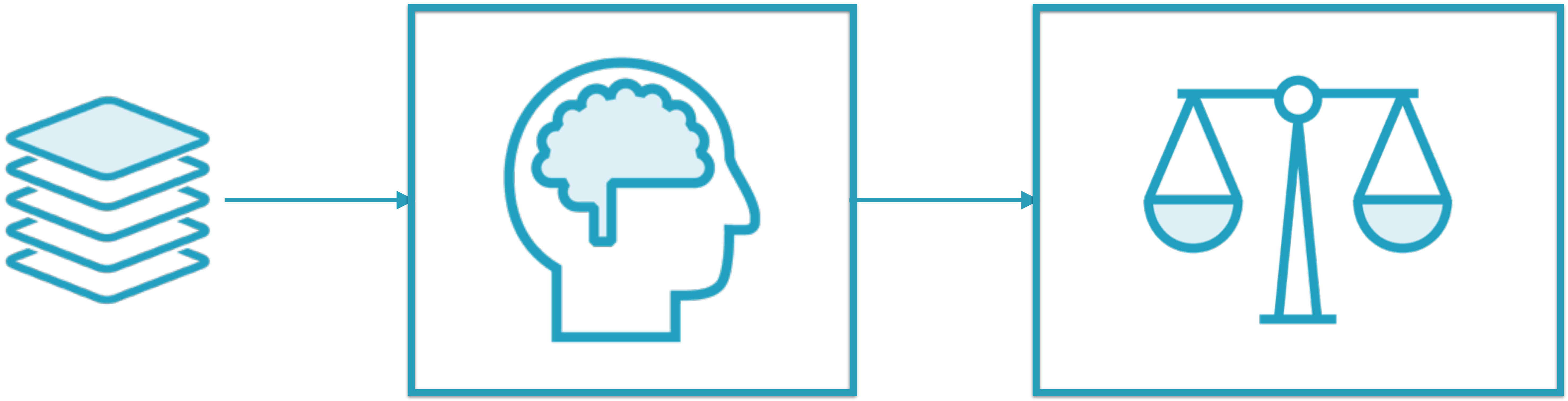
Rule-based Binary Classifier



ML-based Binary Classifier



ML-based Binary Classifier

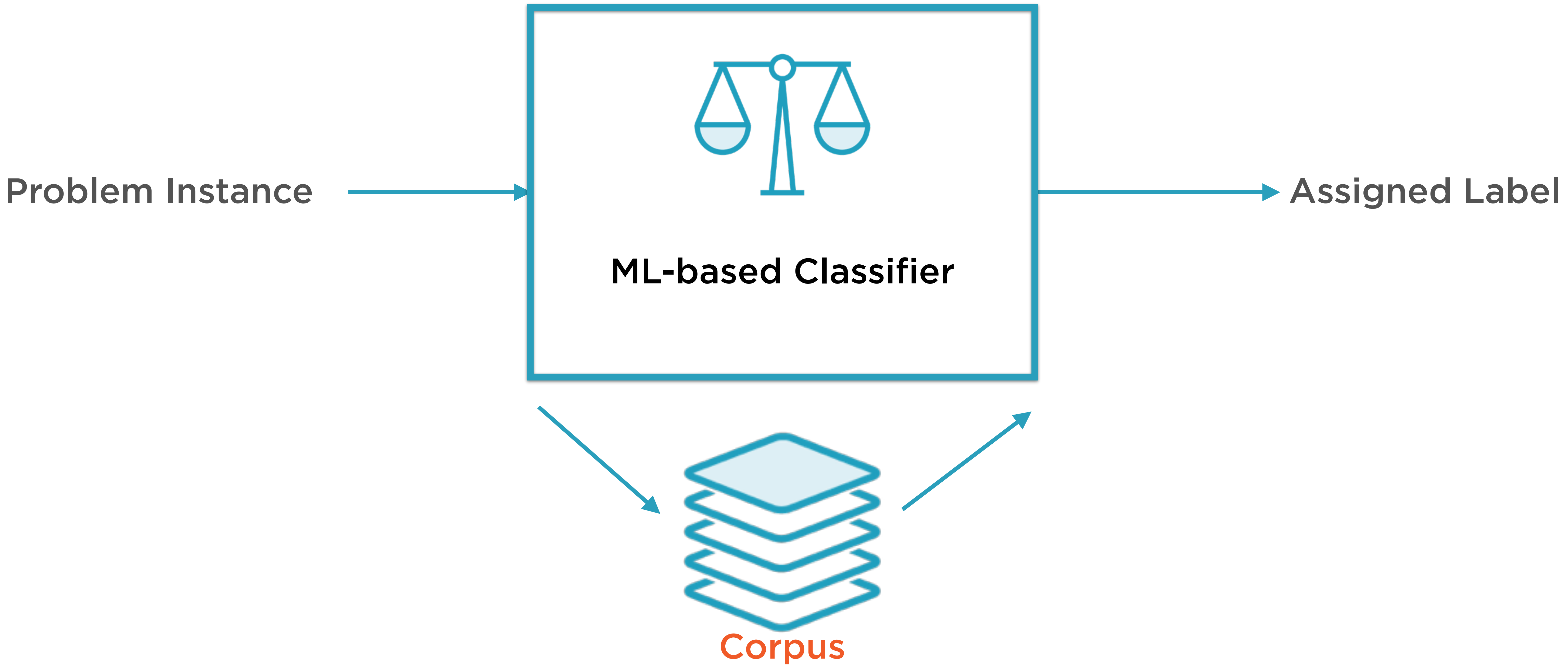


Corpus

Classification Algorithm

ML-based Classifier

ML-based Binary Classifier



ML-based and Rule-based Classifiers

ML-based	Rule-based
Dynamic - alter output based on patterns in data	Static - rules are applied independent of data being analysed
No need for expert skill	Experts needed to formulate rules
Corpus of data needed, cannot operate on isolated problem instance	Can operate on isolated problem instances
To update classifier, update corpus	To update classifier, update rules
Might require an explicit ‘training’ step (depends on the ML technique employed)	No training step required

Rule-based classifiers can be just as complex and effective as ML-based ones

Summary

Sentiment analysis extracts information from opinions

Polarity detection is the commonest form of sentiment analysis

ML-based classifiers alter their working based on the data

Rule-based classifiers don’t