Getting Started with Natural Language Processing with Python

GETTING STARTED

Swetha Kolalapudi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Recognize applications of Natural Language Processing

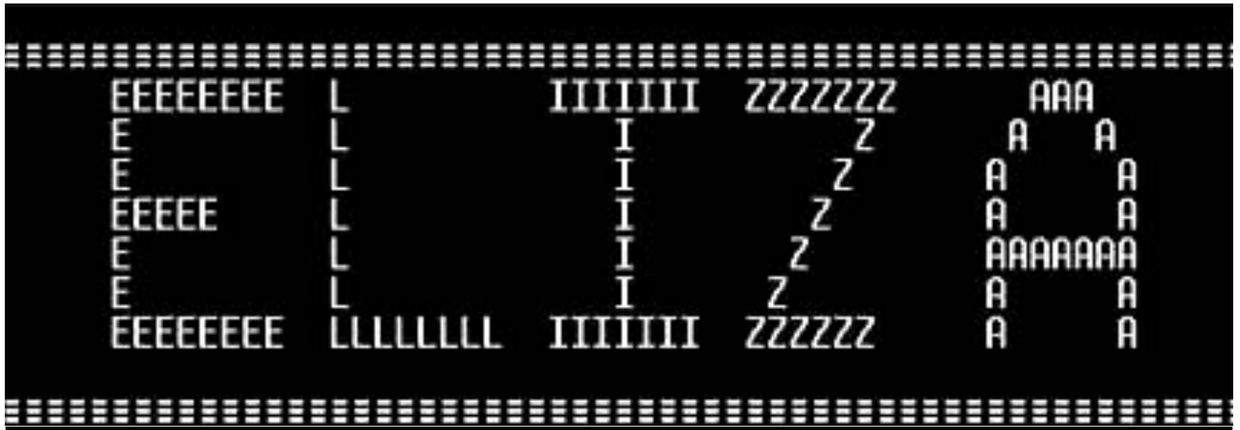
Understand the role of Machine learning in NLP

Introduce NLTK and it’s basic functionality

Eliza in 1964


```
    EEEEEEEE  L           IIIIIII  ZZZZZZZ      AAA
    E         L              I          Z      A   A
    E         L              I         Z      A     A
    EEEEE     L              I        Z       A     A
    E         L              I       Z        AAAAAAA
    E         L              I      Z         A     A
    EEEEEEEE  LLLLLLLL   IIIIIII  ZZZZZZ     A     A
```

ELIZA at MIT

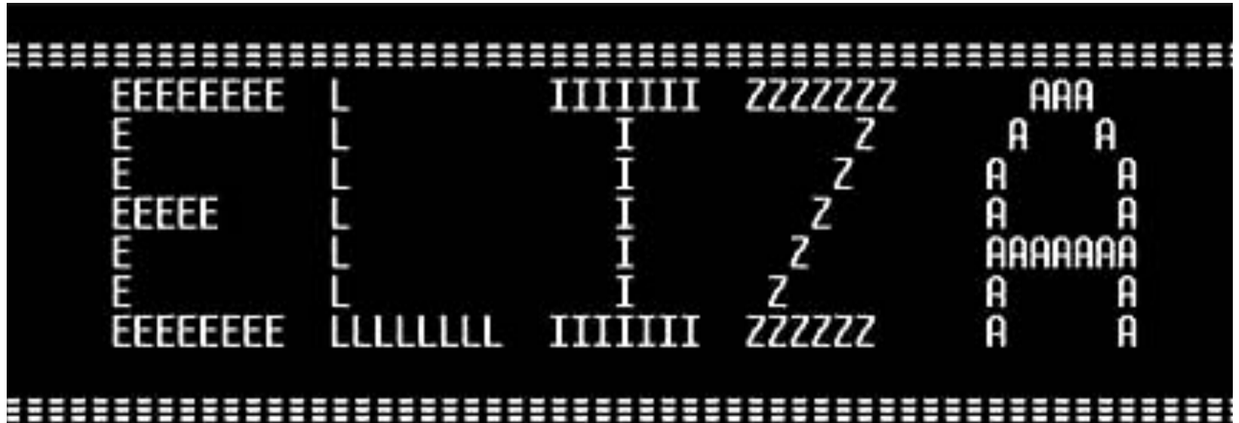


ELIZA> I am the psychotherapist. Please describe your problem

I have a terrible headache

ELIZA> Maybe you should consult a doctor of medicine. I am a psychiatrist.

ELIZA at MIT



ELIZA> I am the psychotherapist. Please describe your problem

I have a terrible headache

ELIZA> Maybe you should consult a doctor of medicine. I am a psychiatrist.

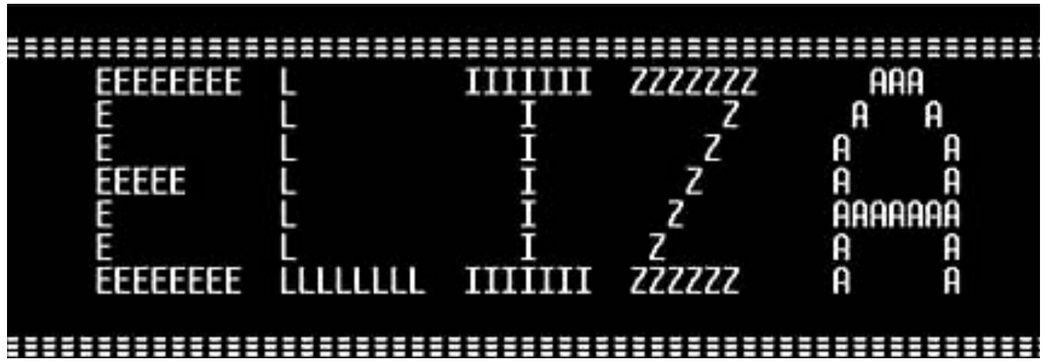
An early attempt at a program that could mimic human understanding of language

“A computer would deserve to be called intelligent if it could deceive a human into believing that it was human.”

Alan Turing

A Long Way Since Then

1966



2016

Siri

Google Now

Human beings communicate using natural language

Natural Language Processing

Enable computers to derive meaning from natural language

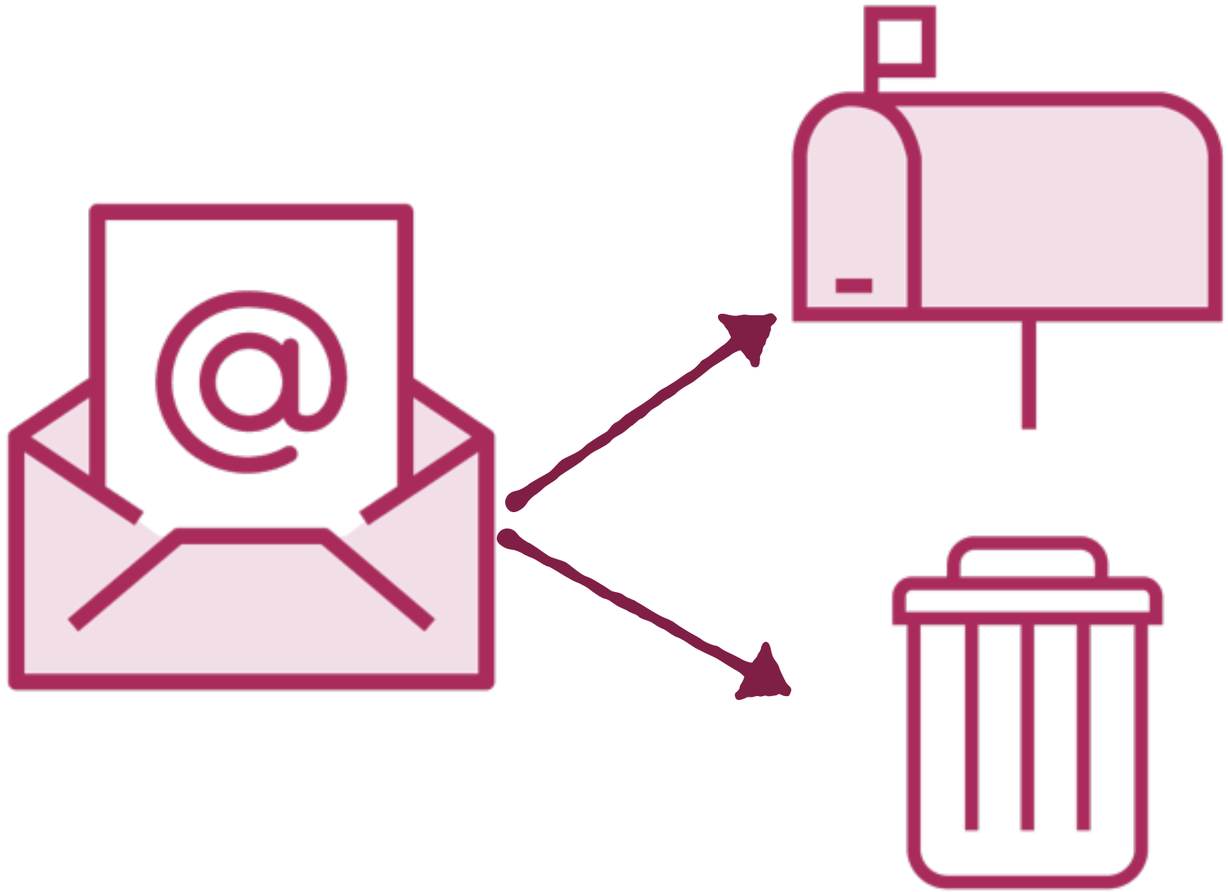
Natural Language Processing

denormalize the table data using sql

mysql sql

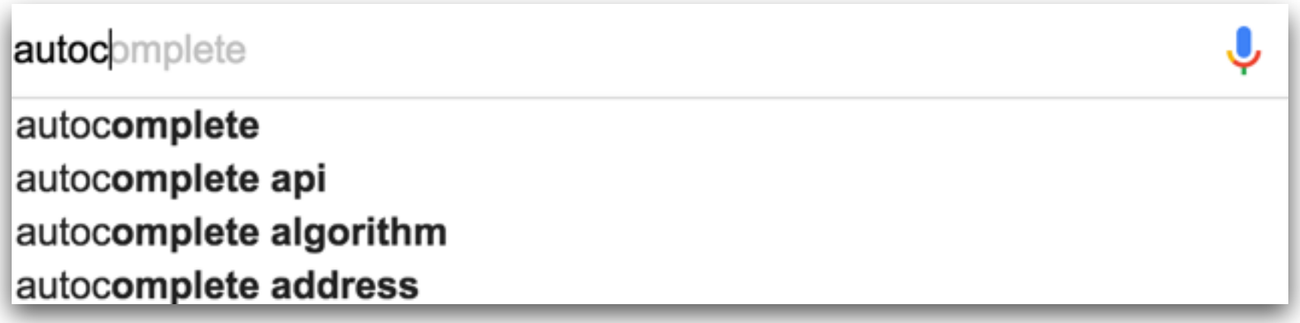
Auto-Tagging

Natural Language Processing



Spam Detection

Natural Language Processing



Autocomplete

Natural Language Processing

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?

Sentiment Analysis

Natural Language Processing

- **Auto-summarizing text**
- **Identifying the genre of a book**
- **Recognizing the themes/topics in an article**

Tasks in Natural Language Processing

Tokenization

Breaking down text into words and sentences

Stopword Removal

Filtering common words

N-Grams

Identifying commonly groups of words

Word Sense Disambiguation

Identifying the context in which the word occurs

Parts-of-Speech

Identifying Part-of-Speech

Stemming

Removing ends of the words

Tokenization

Stopword Removal

Stopword Removal

Mary little lamb

N-Grams

New York is a great city. Have you ever been to New York?

Bigrams

Word Sense Disambiguation

The movie had really cool effects.

I'd like a tall glass of cool water.

Parts of Speech Tagging

Noun Verb Adj. Noun

Mary had a little lamb.

Stemming

- Close
- Closed
- Closely
- Closer

Demo

Tokenize text into sentences and words

Demo

Remove stop words

Demo

Identify bigrams

Demo

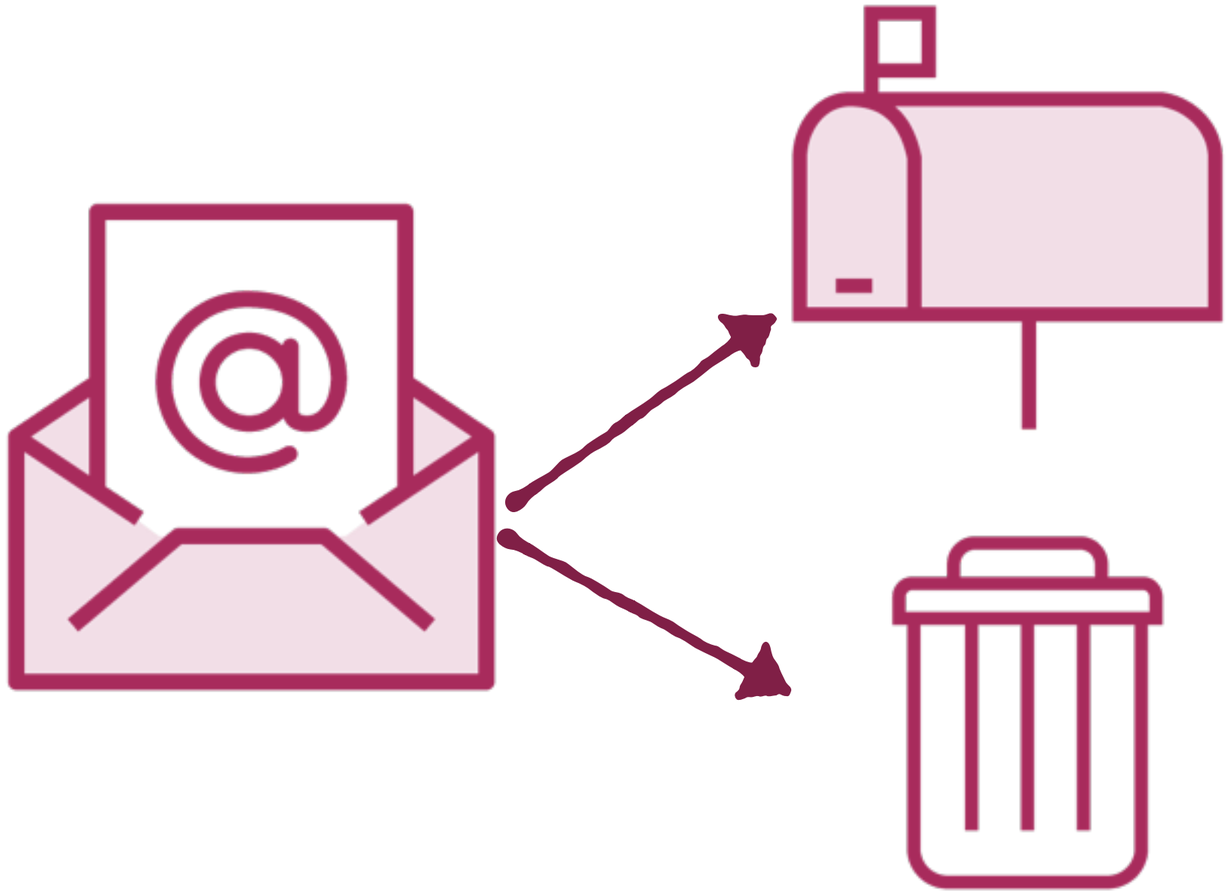
Stemming

Parts of Speech Tagging

Demo

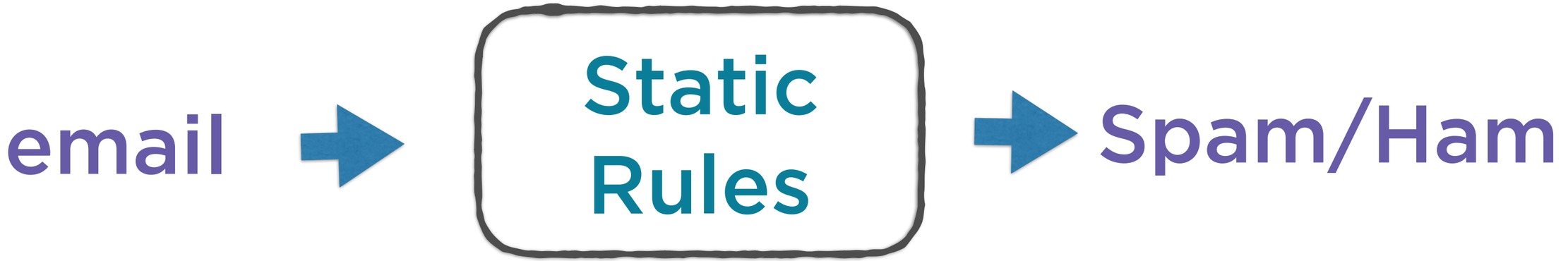
Word Sense Disambiguation

Spam Detection



Rule Based Approach

- **Write rules by hand**

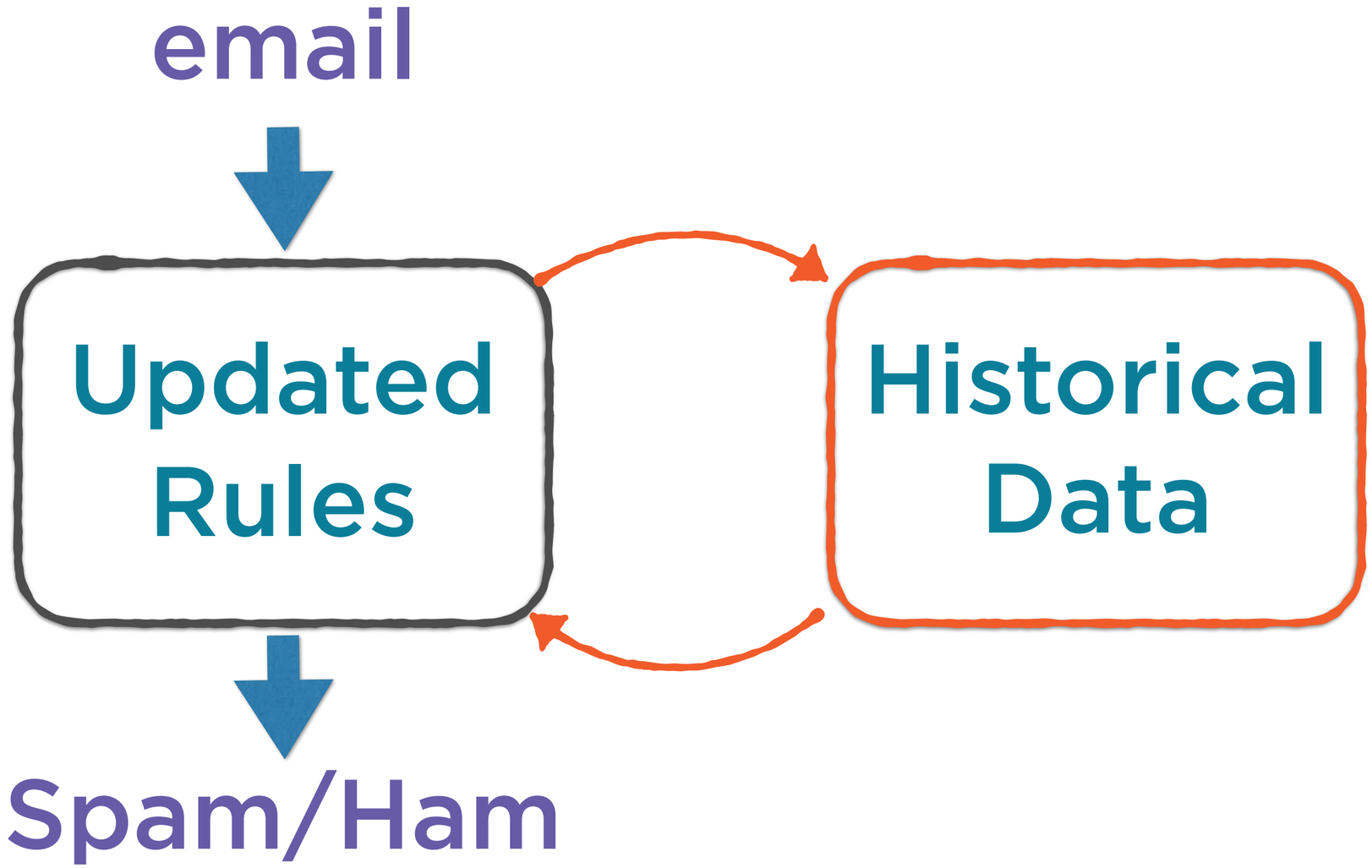


- Contains specific keywords

Use Machine Learning

- Difficult for humans to express rules
- Patterns/Relationships are dynamic
- A large amount of historical data is available

Machine Learning Approach



Two Approaches

Rule Based Approach

Machine Learning Approach

Two Approaches

Rule Based Approach

Machine Learning Approach

Typical ML Workflow

Pick your Problem

Identify which type of problem we need to solve

Represent Data

Represent data using numeric attributes

Apply an Algorithm

Use a standard algorithm to find a model

Typical ML Workflow

Pick your Problem

Identify which type of problem we need to solve

Represent Data

Represent data using numeric attributes

Apply an Algorithm

Use a standard algorithm to find a model

Pick your Problem

ML problems generally fall under a broad set of categories

Classification

Clustering

Pick your Problem

ML problems generally fall under a broad set of categories

Classification

Clustering

Classification

Spam Detection

Is this email **Spam** or **Ham**?

Sentiment Analysis

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

Classification

We are given a **problem instance**

An e-mail

A Tweet

Classification

We need to assign a category to the problem instance

Spam or **Ham**?

positive or **negative**?

Classification

Algorithms which perform classification are known as **Classifiers**

Classification

A Classifier

uses a set of instances for which the correct category membership is known

Training Data

Ex: Tweets which are correctly classified as positive or negative

Pick your Problem

ML problems generally fall under a broad set of categories

Classification

Clustering

Pick your Problem

ML problems generally fall under a broad set of categories

Classification

Clustering

Clustering

Say you have a large group of articles

Divide the articles into **groups** based on some **common attributes**

Clustering

The key thing here is that..

..the groups to be divided into are **unknown beforehand**

Clustering

The algorithm divides articles into groups

Later, we might realize that these groups represent meaningful divisions

Themes, Topics

Pick your Problem

ML problems generally fall under a broad set of categories

Classification

Clustering

Pick your Problem

ML problems generally fall under a broad set of categories

Classification

Clustering

Typical ML Workflow

Pick your Problem

Identify which type of problem we need to solve

Represent Data

Represent data using numeric attributes

Apply an Algorithm

Use a standard algorithm to find a model

Typical ML Workflow

Pick your Problem

Identify which type of problem we need to solve

Represent Data

Represent data using numeric attributes

Apply an Algorithm

Use a standard algorithm to find a model

Represent Data

Use meaningful numeric attributes to represent text

Term Frequency

TF-IDF

Typical ML Workflow

Pick your Problem

Identify which type of problem we need to solve

Represent Data

Represent data using numeric attributes

Apply an Algorithm

Use a standard algorithm to find a model

Typical ML Workflow

Pick your Problem

Identify which type of problem we need to solve

Represent Data

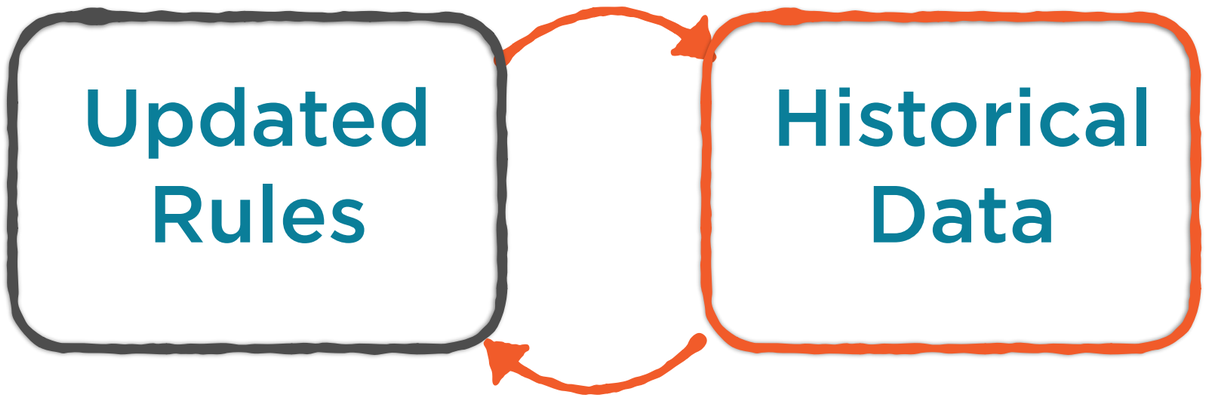
Represent data using numeric attributes

Apply an Algorithm

Use a standard algorithm to find a model

Apply an Algorithm

Use an algorithm to find patterns from the historical data



Apply an Algorithm

Updated Rules

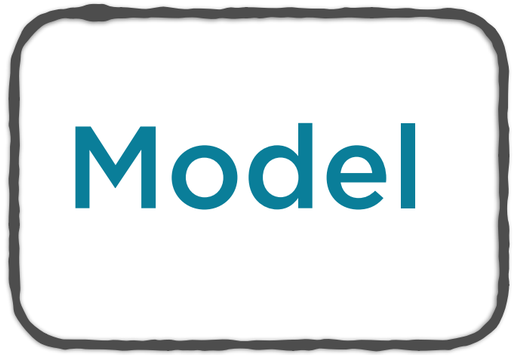
Rules are meant to quantify relationships between variables

Apply an Algorithm

Updated Rules

The rules together form something called a Model

Apply an Algorithm



A Model can be

- a mathematical equation
- a set of rules (if-then-else statements)

Apply an Algorithm

The choice of algorithm depends mainly on the type of problem

- Classification
- Naive Bayes
- Support Vector Machines

Apply an Algorithm

The choice of algorithm depends mainly on the type of problem

- Clustering
- K-Means
- Hierarchical Clustering

Summary

Understand Natural Language Processing and it's role in tasks like auto-tagging, spam detection, Siri, autocomplete

Performing common NLP tasks such as tokenization, stopword removal, word sense disambiguation etc

Understand the role of Machine learning in NLP and an overview of the process