Until now, we have seen data that are structured, numerical, or categorical.

On the other hand, tweets are loosely structured.

They are often textual.

They have poor spelling, often contain non-traditional grammar, and they are multilingual.

In this example here, we see two examples of this aspect of tweets.

We have also discussed why people care about textual data.

A key question, however, is how to handle this information, including in the tweets.

Humans cannot, of course, keep up with internet-scale volumes of data as there are about half a billion tweets per day.

Even at the small scale, the cost and time required to process this is of course prohibitive.

How can computers help?

The field that addresses how computers understand text is called Natural Language Processing.

The goal is to understand and derive meaning from human language.

In 1950, Alan Turing, a major computer scientist of the era, proposed a test of machine intelligence.

That the computer program passes it if it can take part in a real-time conversation and cannot be distinguished from a human.

Let's discuss briefly the history of Natural Language Processing.

There has been some progress-- for example, the "chatterbot" ELIZA.

The initial focus has been on understanding grammar.

Later, the focus shifted towards statistical, machine learning techniques that learn from large bodies of text.

Today, there are modern versions of these Natural Language Processing.

Apple is using Siri, and Google is using Google Now.

Why is it hard?

Let us give you an example.

Suppose we say the phrase, I put my bag in the car.

Is it large and blue?

The question is, does the "it" refer to the bag or the "it" refers to car?

The context is often important.

Humans use homonyms, metaphors, often sarcasm.

In this lecture, we'll see how can build analytics models using text as our data.