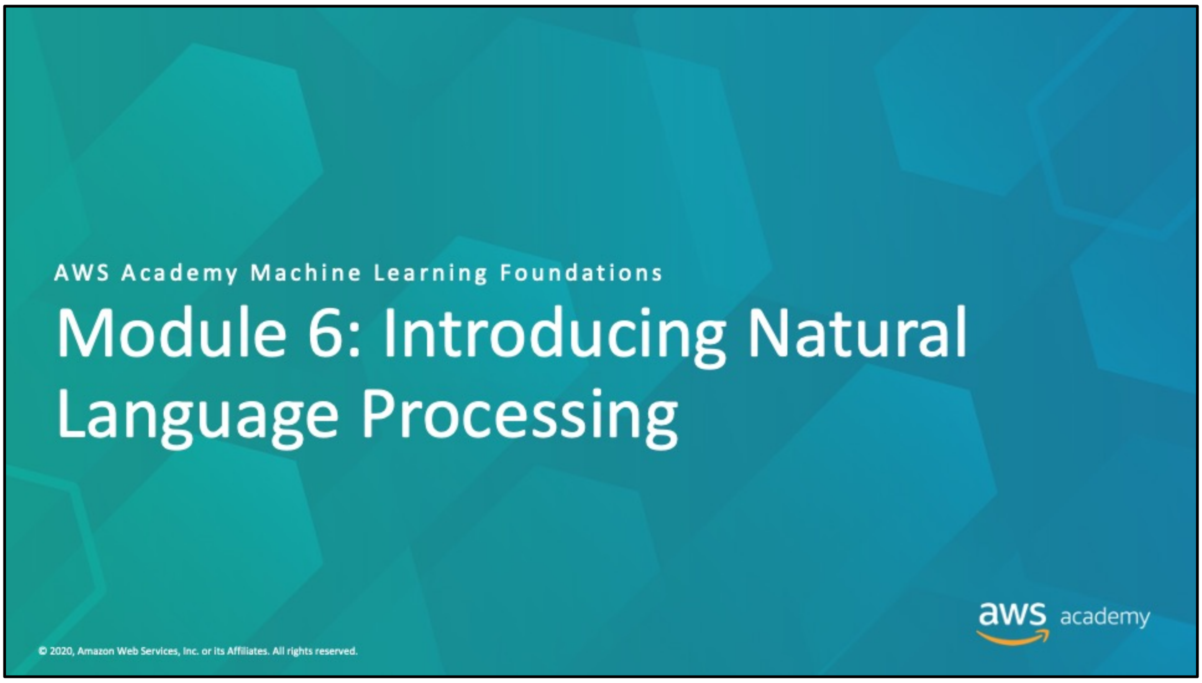
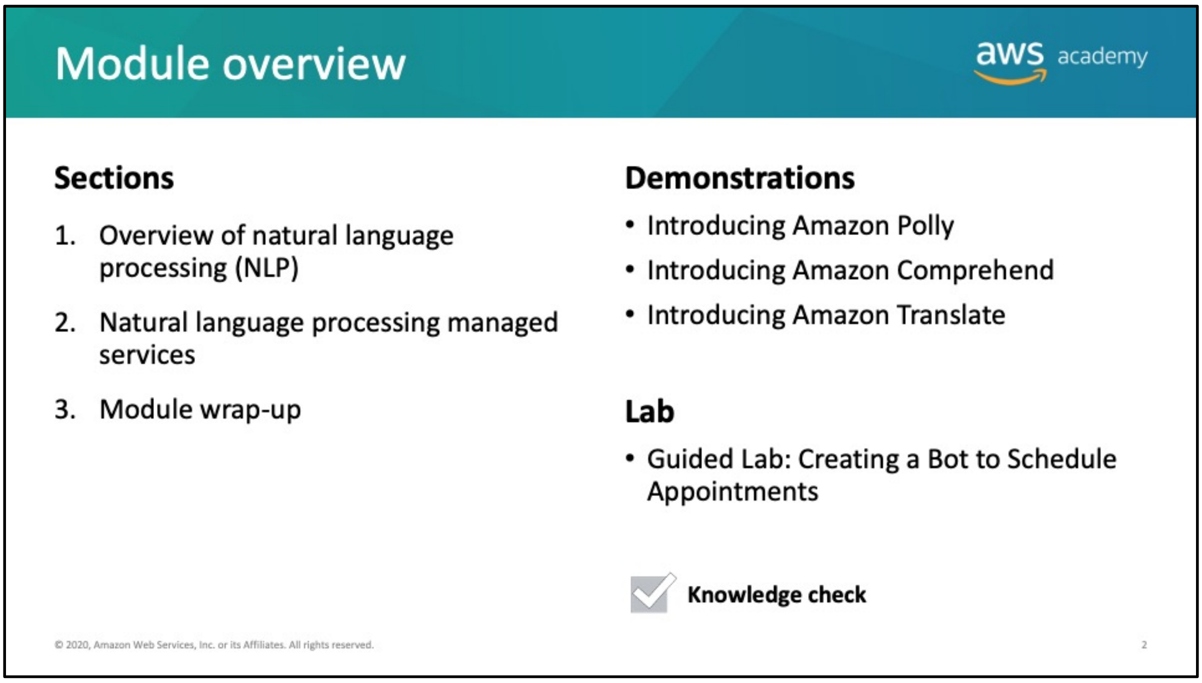


AWS Academy Machine Learning Foundations

Module 6: Introducing Natural Language Processing

Welcome to Module 6: Introducing Natural Language Processing.



Module overview

Sections

1. Overview of natural language processing (NLP)
2. Natural language processing managed services
3. Module wrap-up

Demonstrations

- Introducing Amazon Polly
- Introducing Amazon Comprehend
- Introducing Amazon Translate

Lab

- Guided Lab: Creating a Bot to Schedule Appointments

☑ **Knowledge check**

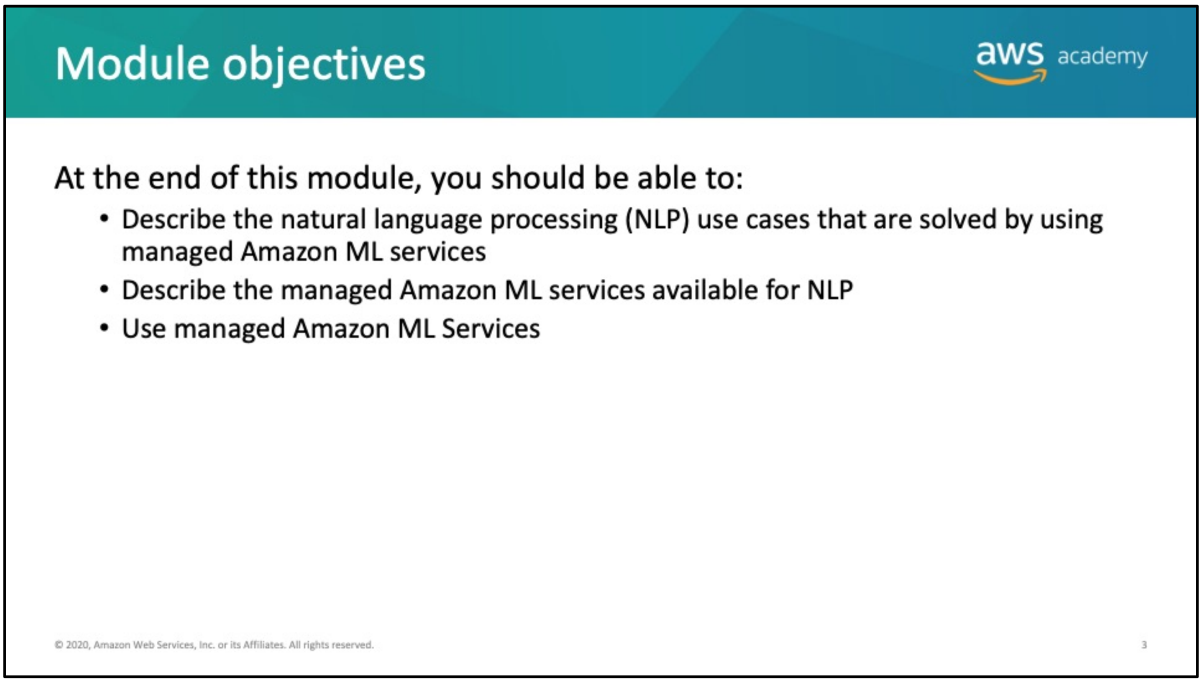
This module includes the following sections:

- Introduction to natural language processing (NLP). This introduction includes a description of the major challenges that are faced when you work with NLP, and the overall development process for NLP applications.
- A review of five AWS services that you can use to speed up the development of NLP-based applications.

This module also includes <recorded / educator-led> demonstrations that will show you how to use Amazon Comprehend, Amazon Polly and Amazon Translate by using the AWS Management Console.

The module also includes a hands-on guided lab where you will learn how to use Amazon Lex to develop a bot for scheduling appointments.

Finally, you will be asked to complete a knowledge check that will test your understanding of key concepts covered in this module.



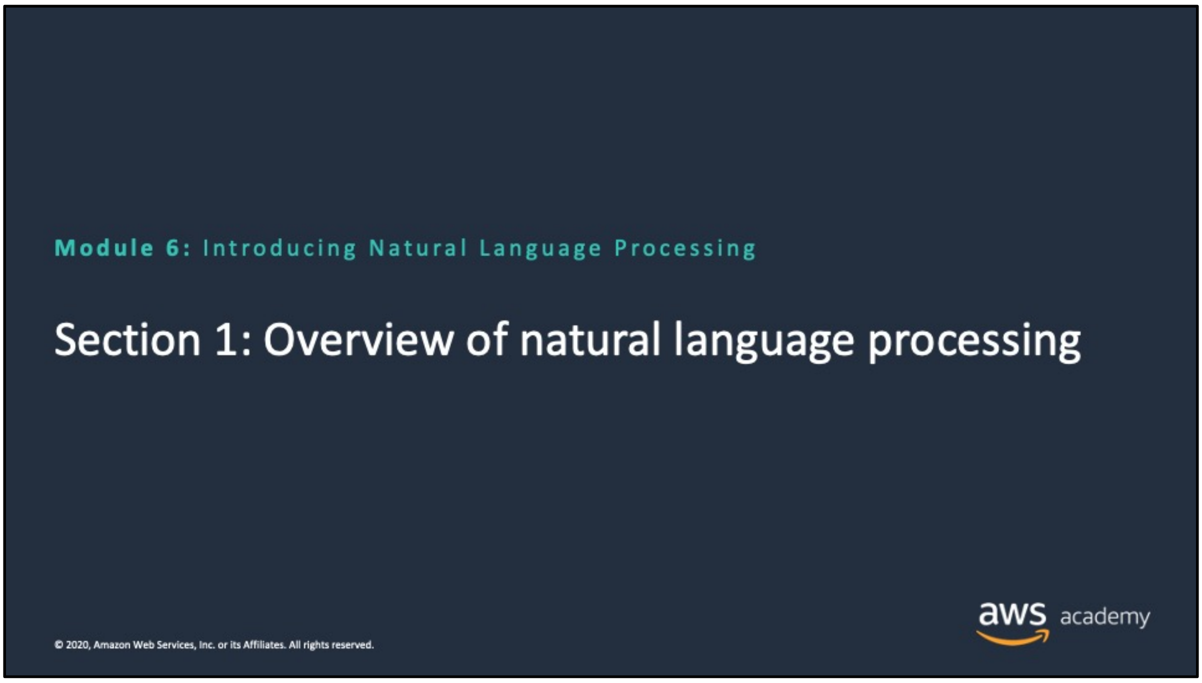
Module objectives

At the end of this module, you should be able to:

- Describe the natural language processing (NLP) use cases that are solved by using managed Amazon ML services
- Describe the managed Amazon ML services available for NLP
- Use managed Amazon ML Services

After completing this module, you should be able to:

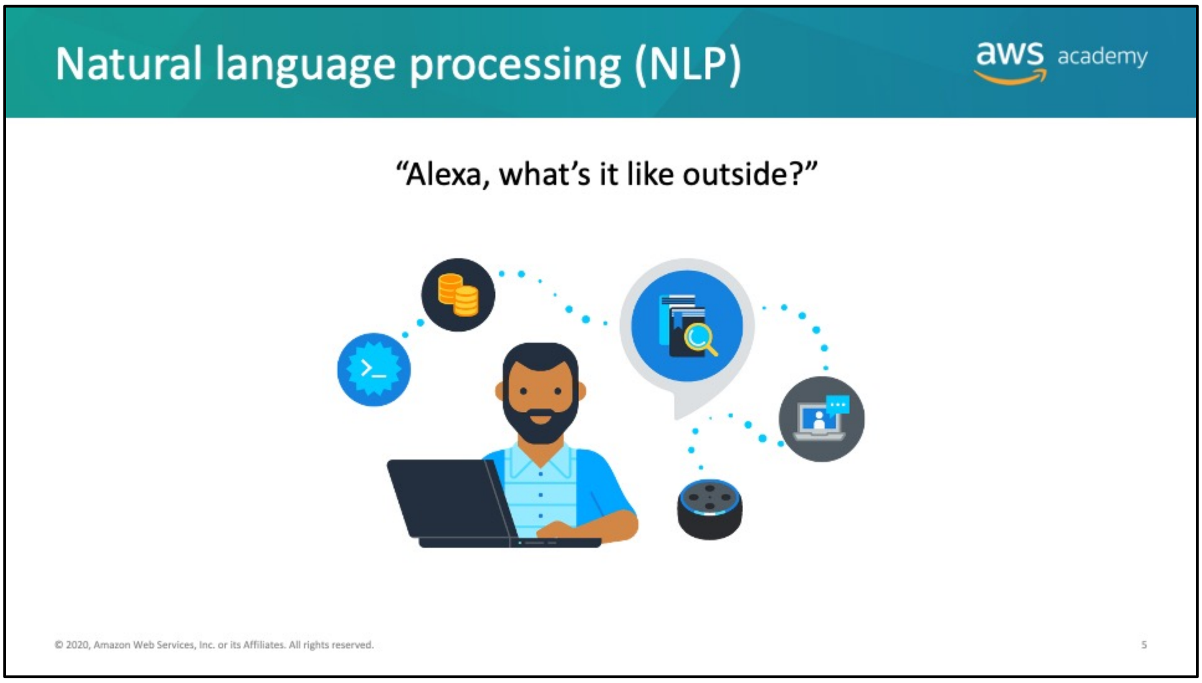
- Describe the natural language processing (NLP) use cases that are solved by using managed Amazon ML services
- Describe the managed Amazon ML services available for NLP
- Use managed Amazon ML Services



Module 6: Introducing Natural Language Processing

Section 1: Overview of natural language processing

Introducing Section 1: Overview of natural language processing. In this section, you will review the meaning of natural language processing.

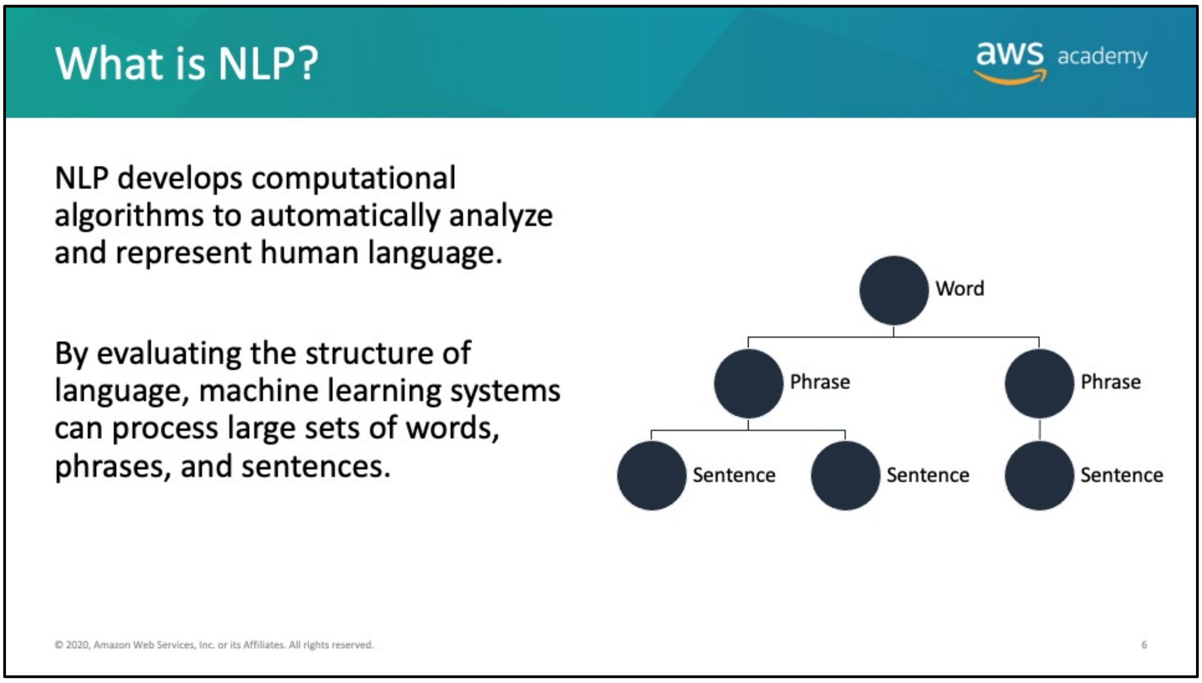


Natural language processing (NLP)

"Alexa, what's it like outside?"

Before you see an explanation of natural language processing (NLP), look at an NLP example with Amazon Alexa.

1. An Amazon device, such as an Echo, records your words. The recording of your speech is sent to Amazon servers to be analyzed more efficiently.
2. Amazon breaks down your phrase into individual sounds. Then, it connects to a database that contains the pronunciations of various words to find which words most closely correspond to the combination of individual sounds.
3. It identifies important words to make sense of the tasks and to carry out corresponding functions. For instance, if Alexa notices words like *outside* or *temperature*, it opens the weather app.
4. Amazon servers send the information back to your device, and Alexa might speak.



What is NLP?

NLP develops computational algorithms to automatically analyze and represent human language.

By evaluating the structure of language, machine learning systems can process large sets of words, phrases, and sentences.

Word

Phrase

Phrase

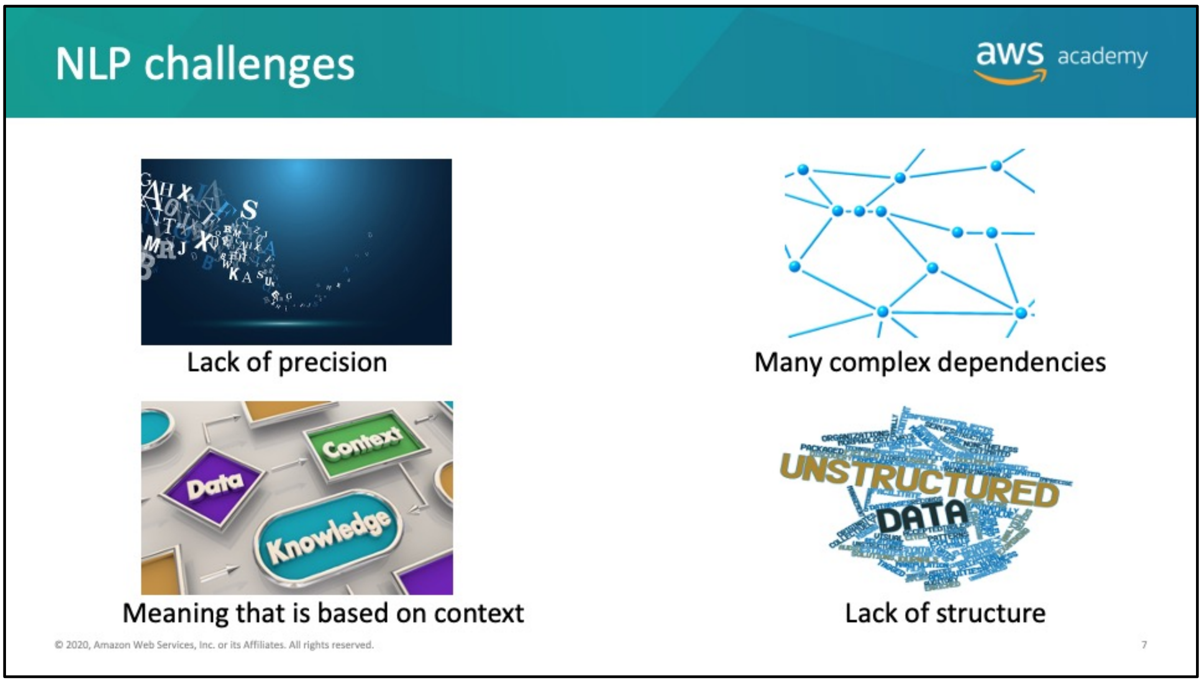
Sentence

Sentence

Sentence

NLP is a broad term for a general set of business or computational problems that you can solve with machine learning (ML). NLP systems predate ML. Two examples are speech-to-text on your old cell phone and screen readers. Many NLP systems now use some form of machine learning. NLP considers the hierarchical structure of language. Words are at the lowest layer of the hierarchy. A group of words make a phrase. The next level up consists of phrases, which make a sentence, and ultimately, sentences convey ideas.

NLP systems face several significant challenges, which you will learn about next.



NLP challenges

Lack of precision

Many complex dependencies

Meaning that is based on context

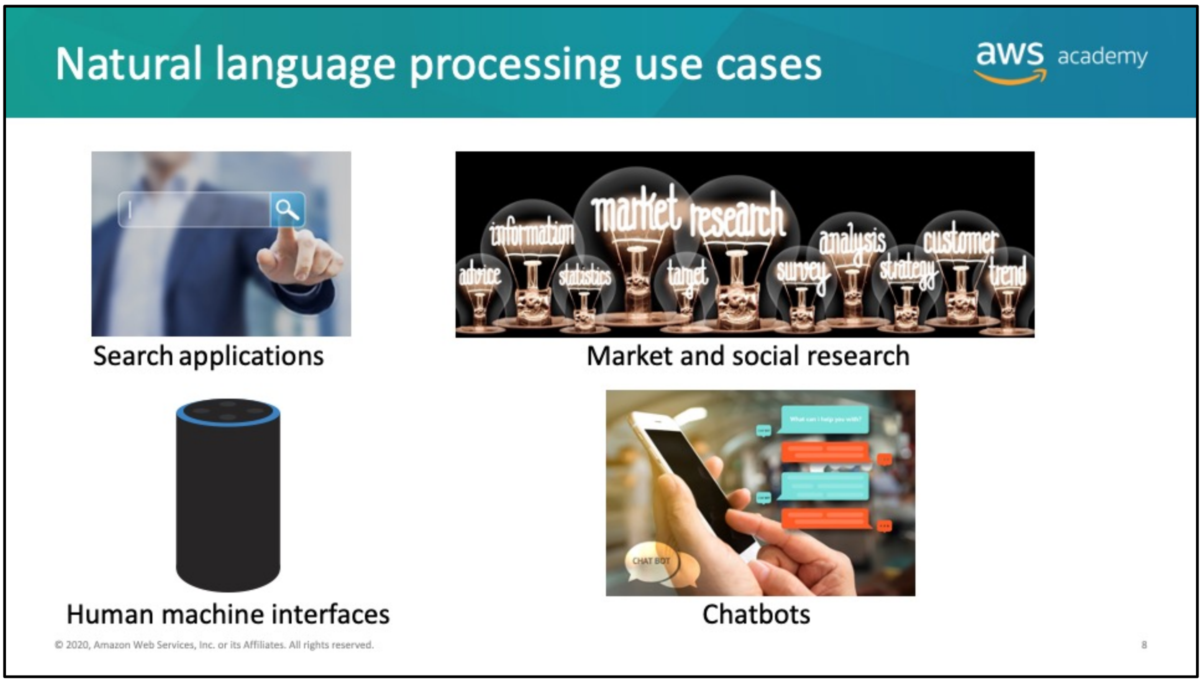
Lack of structure

Language is not precise. Words can have different meanings, which are based on the other words that surround them (context). Often, the same words or phrases can have multiple meanings. For example, consider the term *weather*. You might be *under the weather*, which means that you are sick. However, *there is wonderful weather outside* means that the weather conditions outside are good. The phrase *Oh, really?* might convey surprise, disagreement, or many other meanings, depending on context and inflection.

Some of the main challenges for NLP include:

- Discovering the structure of the text – One of the first tasks of any NLP application is to break the text into meaningful units, such as words, phrases, and sentences.
- Labeling data – After the system converts the text to data, the next challenge is to apply labels that represent the various parts of speech. Every language requires a different labeling scheme to match the language’s grammar.
- Representing context – Because word meaning depends on context, any NLP system needs a way to represent context. It is a big challenge because of the large number of contexts. Converting context into a form that computers can understand is difficult.
- Applying grammar – Although grammar defines a structure for language, the application of grammar is nearly infinite. Dealing with the variation in how humans use language is a major challenge for NLP systems. Addressing this challenge is where machine learning can

have a big impact.



Natural language processing use cases

Search applications

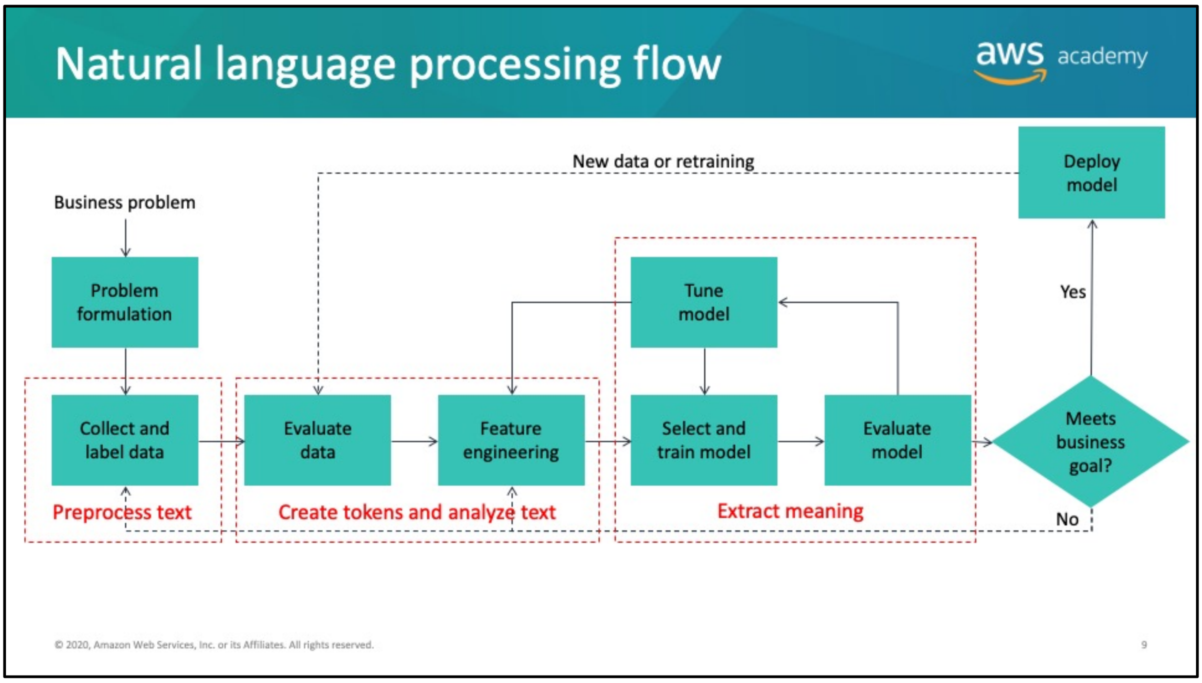
Market and social research

Human machine interfaces

Chatbots

You can apply NLP to a wide range of problems. Some of the more common applications include:

- Search applications (such as Google and Bing)
- Human machine interfaces (such as Alexa)
- Sentiment analysis for marketing or political campaigns
- Social research that is based on media analysis
- Chatbots to mimic human speech in applications



Natural language processing flow

aws academy

Business problem

Problem formulation

Collect and label data

Preprocess text

Evaluate data

Feature engineering

Create tokens and analyze text

Tune model

Select and train model

Evaluate model

Extract meaning

Meets business goal?

Yes

No

Deploy model

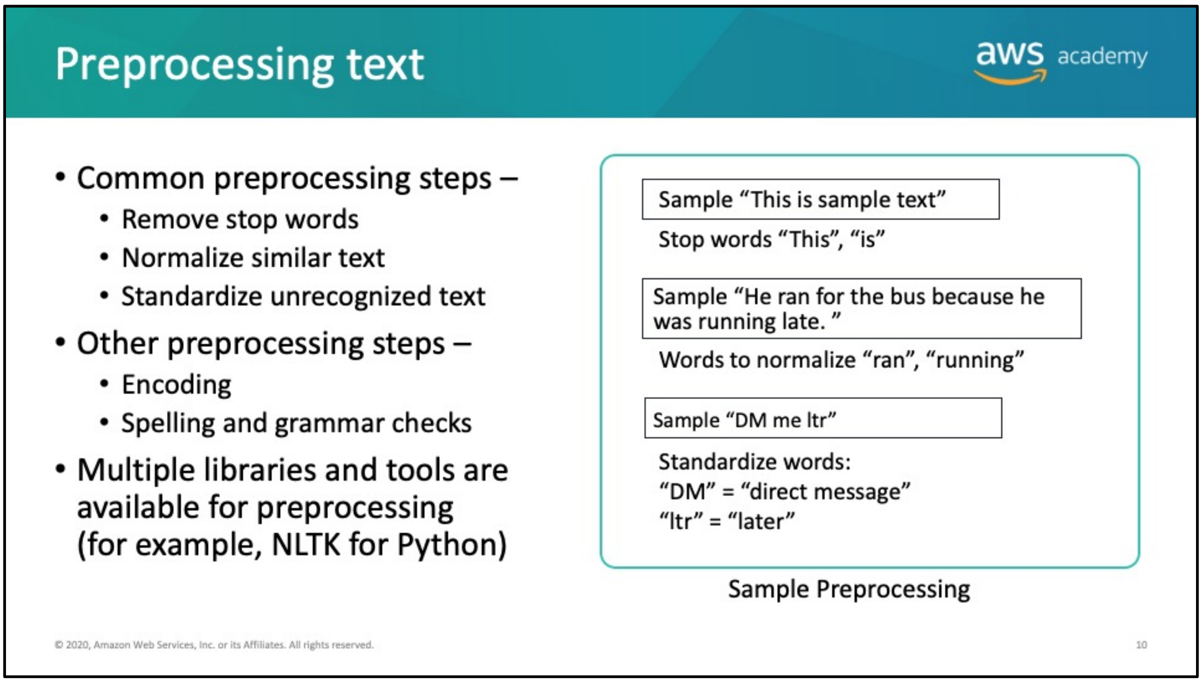
New data or retraining

9

You can apply the ML development pipeline that you have seen throughout this course when you develop an NLP solution. The first task is to formulate a problem, and then collect and label data.

For NLP, collecting data consists of breaking the text into meaningful subsets and labeling the sets. Feature engineering is a large part of NLP applications. This process gets more complicated when you have irregular or unstructured text. For example, if you build an application to classify documents, you must be able to distinguish between words with common terms but different meanings.

Labeling data in the NLP domain is sometimes also called *tagging*. In the labeling process, you must assign individual text strings to different part of speech. You can use specialized tools to help with NLP labeling.



Preprocessing text

aws academy

- Common preprocessing steps –
  - Remove stop words
  - Normalize similar text
  - Standardize unrecognized text
- Other preprocessing steps –
  - Encoding
  - Spelling and grammar checks
- Multiple libraries and tools are available for preprocessing (for example, NLTK for Python)

Sample “This is sample text”

Stop words “This”, “is”

Sample “He ran for the bus because he was running late. ”

Words to normalize “ran”, “running”

Sample “DM me ltr”

Standardize words:
“DM” = “direct message”
“ltr” = “later”

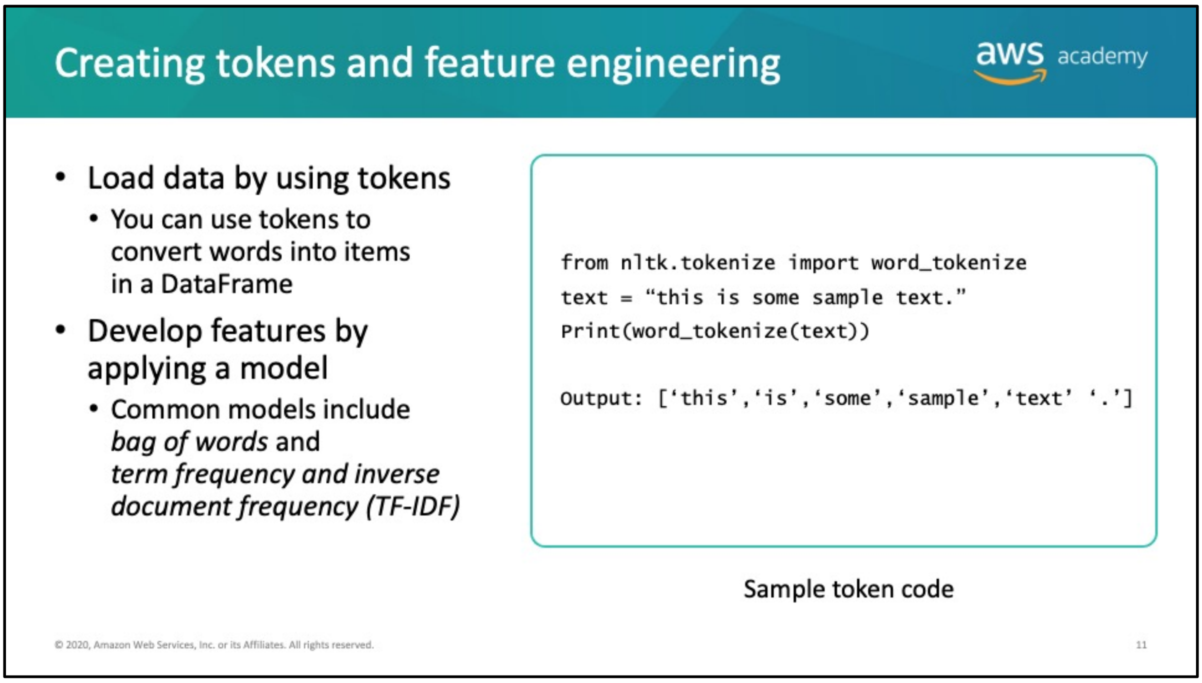
Sample Preprocessing

The first task for an NLP application is to convert the text to data so that it can be analyzed. You convert text by removing words that are not needed for the analysis from the input text. In the example, the words *This* and *is* are removed to leave the phrase *sample text*.

After you remove these stop words, you can normalize text by converting similar words into a common form. For example, the words *run*, *runner*, *ran*, and *running* are all different forms of the word *run*. You can normalize all instances of these words within a block of text by using processes of stemming and lemmatization.

After you normalize the text, you can standardize it by removing words that are not in the dictionary that you use for analysis. Examples include acronyms, slang, and special characters.

The Natural Language Toolkit (NLTK) Python library provides functions that you can use to remove stop words, normalize, and standardize text.



Creating tokens and feature engineering

- Load data by using tokens
  - You can use tokens to convert words into items in a DataFrame
- Develop features by applying a model
  - Common models include *bag of words* and *term frequency and inverse document frequency (TF-IDF)*

```
from nltk.tokenize import word_tokenize
text = “this is some sample text.”
Print(word_tokenize(text))

Output: [‘this’,‘is’,‘some’,‘sample’,‘text’ ‘.’]
```

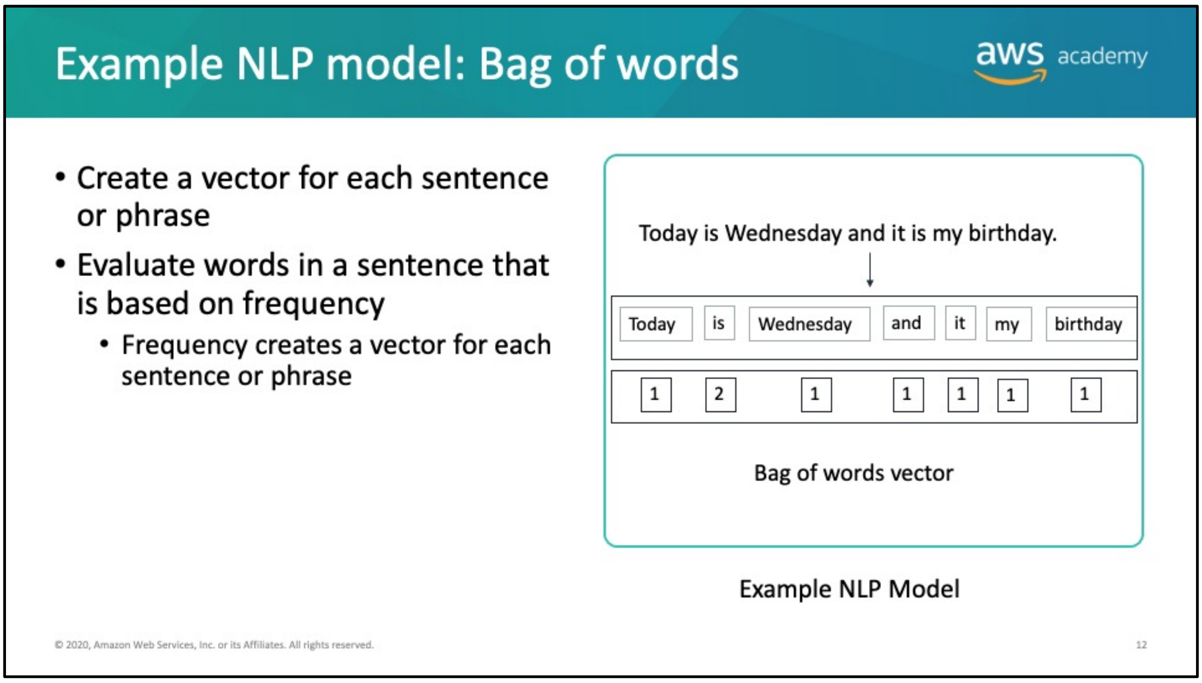
Sample token code

One of the first steps for creating an NLP system is to convert the text into a data collection, such as a DataFrame. All the NLP libraries provide functions to assist with this type of conversion. This example shows how to use the word_tokenize function that’s provided in the NLTK library.

After you clean up your text and load it into a DataFrame, you can apply one of the NLP models to create features.

Common models include:

- Bag of words – This simple model captures the frequency of words in a document. For each word in the document, a key is created, with a value that is the number of occurrences within that document.
- Term frequency and inverse document frequency (TF-IDF) – *Term frequency* is a count of how many times a word appears in a document. *Inverse document frequency* is the number of times a word occurs in a group of documents. These two values are used together to calculate a weight for the words. Words that frequently appear in many documents have a lower weight.



Example NLP model: Bag of words

- Create a vector for each sentence or phrase
- Evaluate words in a sentence that is based on frequency
  - Frequency creates a vector for each sentence or phrase

Today is Wednesday and it is my birthday.

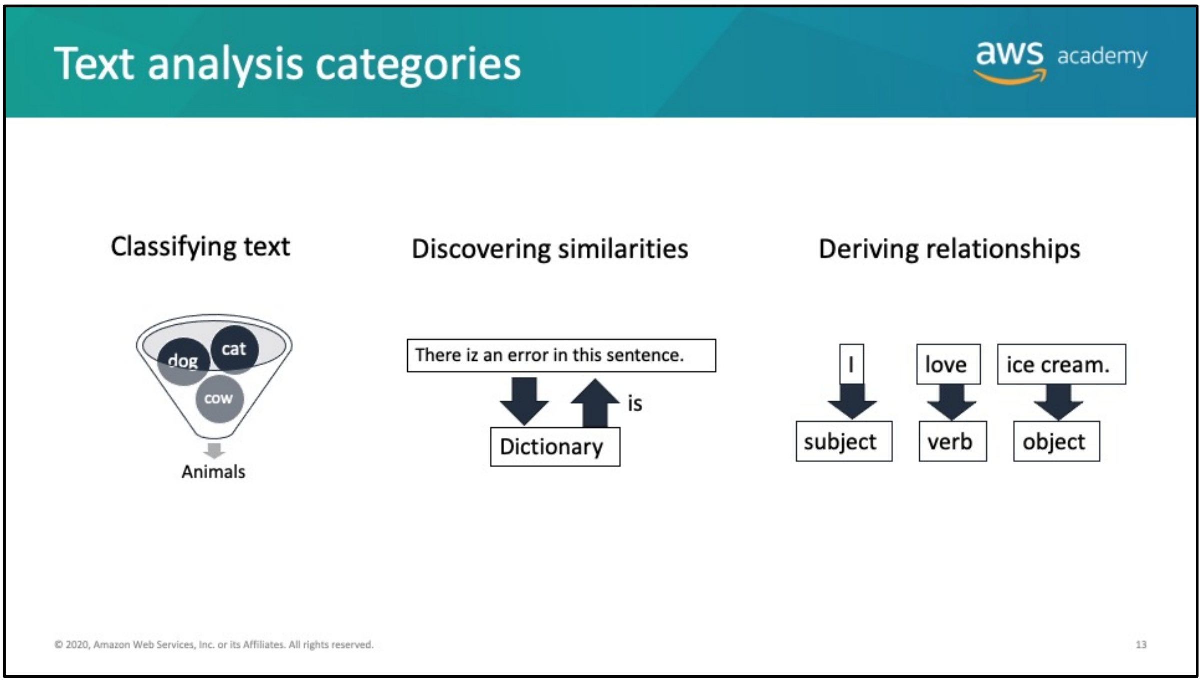
Today	is	Wednesday	and	it	my	birthday
1	2	1	1	1	1	1

Bag of words vector

Example NLP Model

Many models have been established in the NLP field. This example shows the bag of words model. Bag of words is a vector model. Vector models convert each sentence or phrase into a vector, which is a mathematical object that records both directionality and magnitude. In the example, a simple sentence is converted into a vector where each word is recorded in terms of frequency. The word *is* has a value of *2* because it appears twice in the sentence.

Bag of words is often used to classify documents into different categories. Bag of words is also used to derive attributes that feed into NLP applications, such as sentiment analysis.



Text analysis categories

aws academy

Classifying text

Discovering similarities

There iz an error in this sentence.

is

Dictionary

Deriving relationships

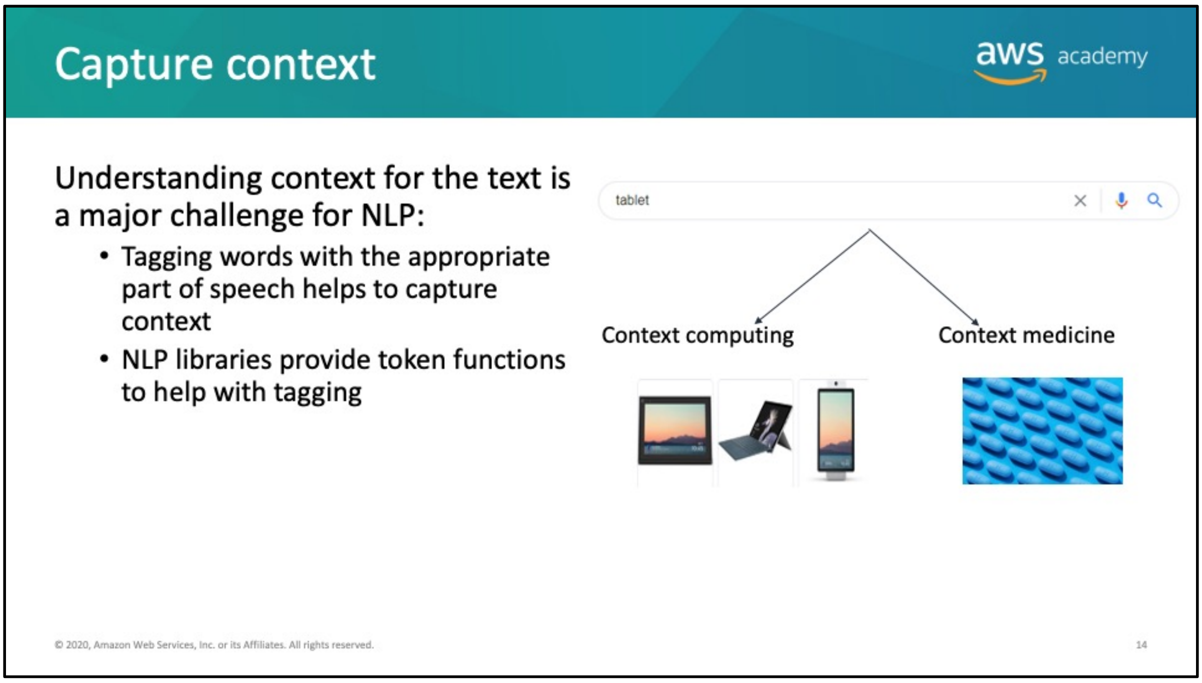
I → subject

love → verb

ice cream. → object

Text analysis has three broad categories:

- Classifying text – This category of analysis is similar to other classification systems that you have seen in this course. Text provides the input to a process that extracts features, which are then sent though an ML algorithm. This algorithm interacts with a classifier model to infer the classification. You can use the NLTK Python library to create a classification system.
- Discovering similarities – Text matching has many applications. For example, auto-correct, spell check, and grammar check are all based on text matching. The edit distance (also known as the Levenshtein distance) algorithm is frequently used.
- Deriving relationships – You can derive relationships between different words or phrases in the text by using a process called coreference resolution. Several NLP systems provide Python libraries for deriving relationships.



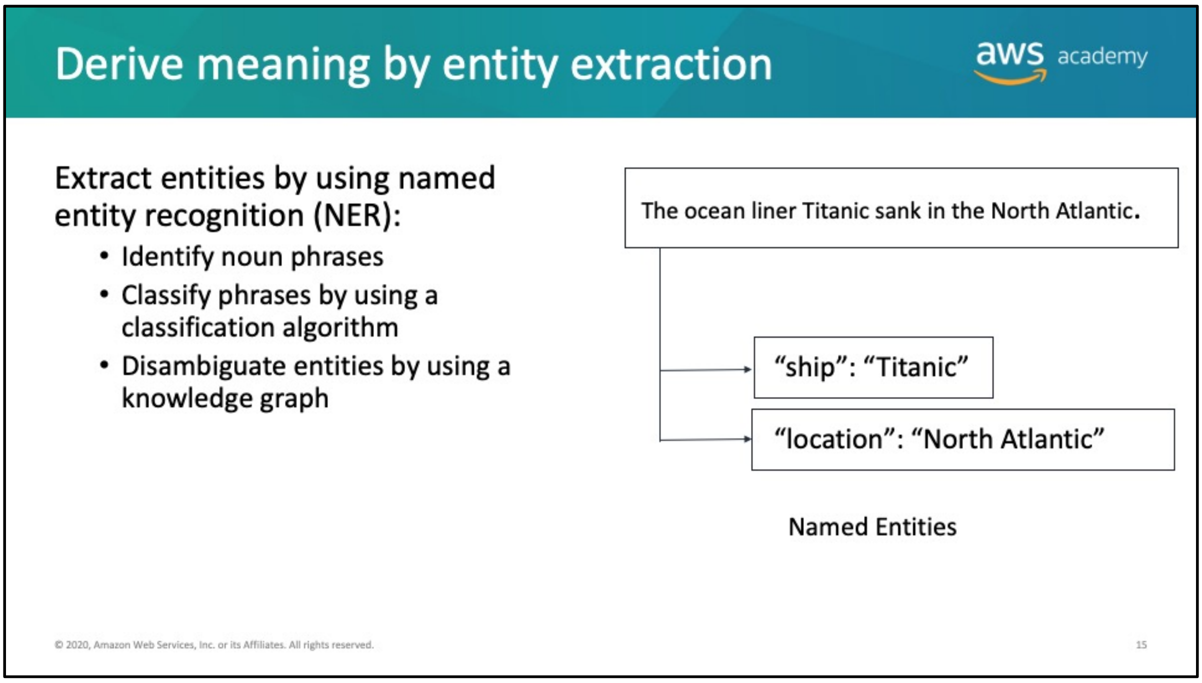
Capture context

Understanding context for the text is a major challenge for NLP:

- Tagging words with the appropriate part of speech helps to capture context
- NLP libraries provide token functions to help with tagging

tablet
Context computing
Context medicine

One of the largest challenges for NLP is how to describe the context for the text. Consider the example where a user is searching for the term *tablet*. Because the word *tablet* has at least two distinct meanings, the search engine must know which meaning the user has in mind. The term might be qualified further (for example, by adding another term like *medicine* or *computing*). If it's not qualified by other terms, then most search engines rely on the most commonly used context.



Derive meaning by entity extraction

aws academy

Extract entities by using named entity recognition (NER):

- Identify noun phrases
- Classify phrases by using a classification algorithm
- Disambiguate entities by using a knowledge graph

The ocean liner Titanic sank in the North Atlantic.

“ship”: “Titanic”

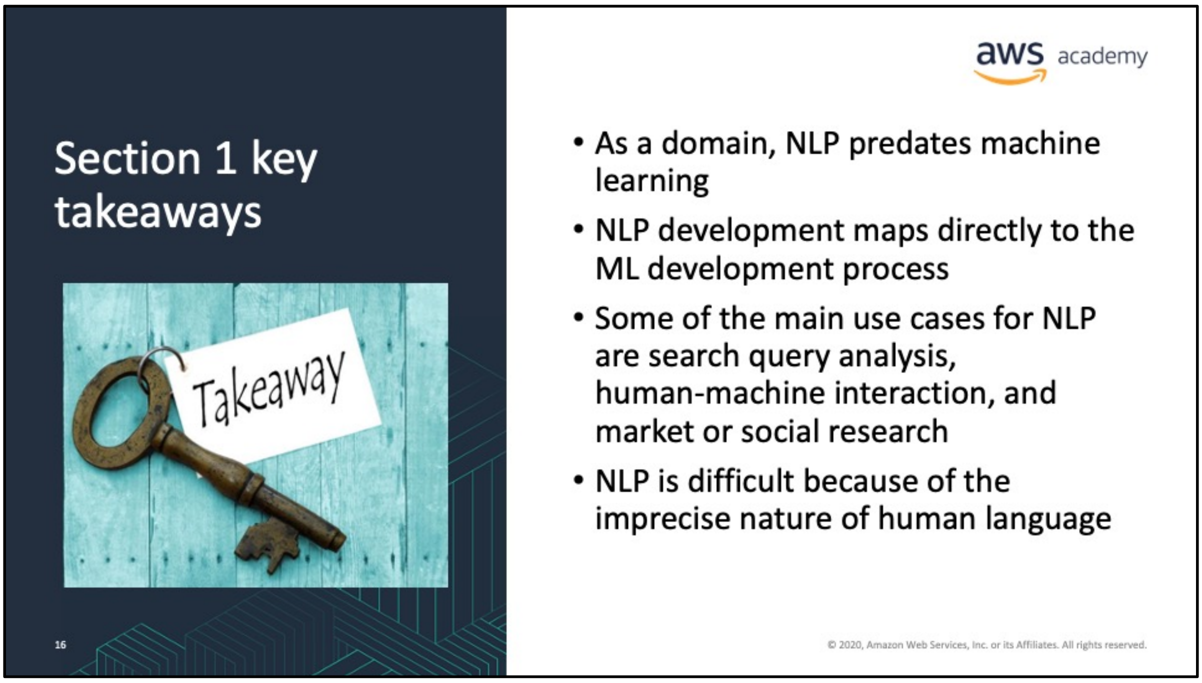
“location”: “North Atlantic”

Named Entities

The process of extracting entities is known as named entity recognition (NER). A NER model has the following components:

- Identify noun phrases by using dependency charts and part of speech tagging
- Classify phrases by using a classification algorithm, such as Word2Vec
- Disambiguate entities by using a knowledge graph

This example shows how to use NER to extract the entities *Titanic* and *North Atlantic* from the text. After the named entities are extracted, you can use a knowledge graph to extract meaning. A knowledge graph combines subject matter expertise with machine learning to derive meaning. The Amazon recommendations engine is an example of a knowledge graph.

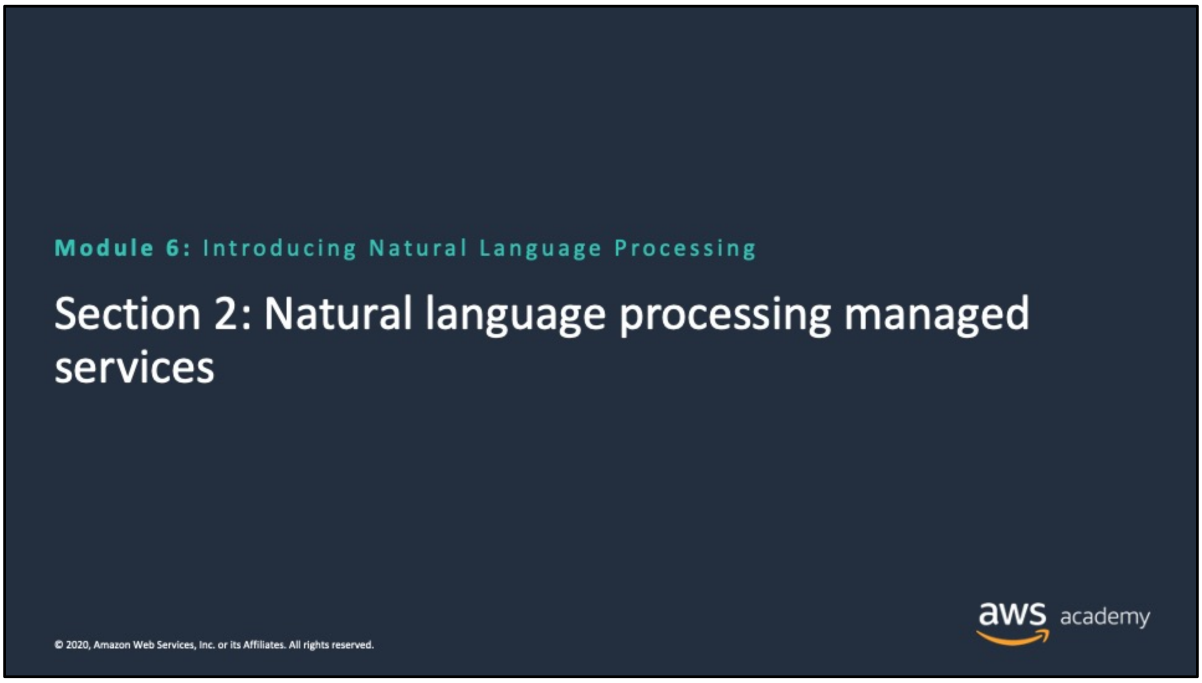


Section 1 key takeaways

- As a domain, NLP predates machine learning
- NLP development maps directly to the ML development process
- Some of the main use cases for NLP are search query analysis, human-machine interaction, and market or social research
- NLP is difficult because of the imprecise nature of human language

Some key takeaways from this section of the module include:

- As a domain, NLP predates machine learning.
- NLP development maps directly to the ML development process.
- Some of the main use cases for NLP are search query analysis, human-machine interaction, and market or social research.
- NLP is difficult because of the imprecise nature of human language.

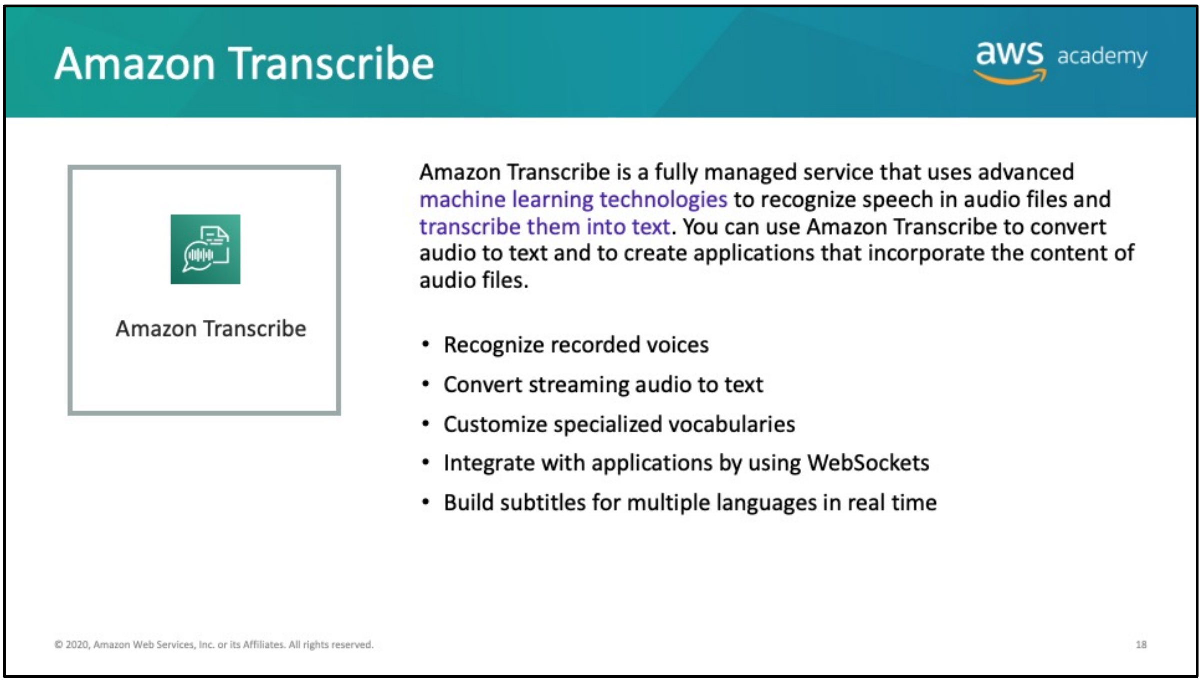


Module 6: Introducing Natural Language Processing

Section 2: Natural language processing managed services

aws academy

Introducing Section 2: Natural language processing managed services. In this section, you will review five managed Amazon ML services that you can use for various NLP use cases. These services simplify the process of creating a NLP application.



Amazon Transcribe

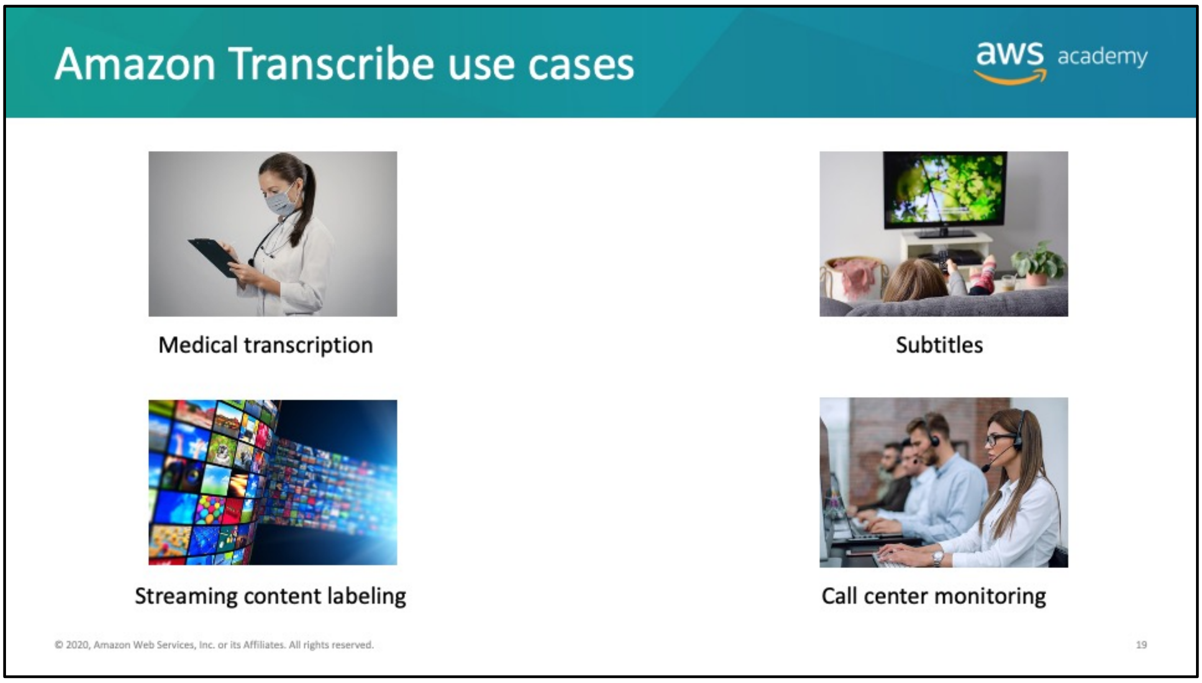
aws academy

Amazon Transcribe

Amazon Transcribe is a fully managed service that uses advanced machine learning technologies to recognize speech in audio files and transcribe them into text. You can use Amazon Transcribe to convert audio to text and to create applications that incorporate the content of audio files.

- Recognize recorded voices
- Convert streaming audio to text
- Customize specialized vocabularies
- Integrate with applications by using WebSockets
- Build subtitles for multiple languages in real time

Amazon Transcribe is the first managed machine learning service that you will learn about. You can use Amazon Transcribe to recognize speech in audio files and produce a transcription. Amazon Transcribe can recognize specific voices in an audio file, and you can create customized vocabulary for terms that are specialized for a particular domain. You can also add transcription service to your applications by integrating with WebSockets. WebSockets provide an internet-facing interface that enables two-way communication between an application and Amazon Transcribe.



Amazon Transcribe use cases

Medical transcription

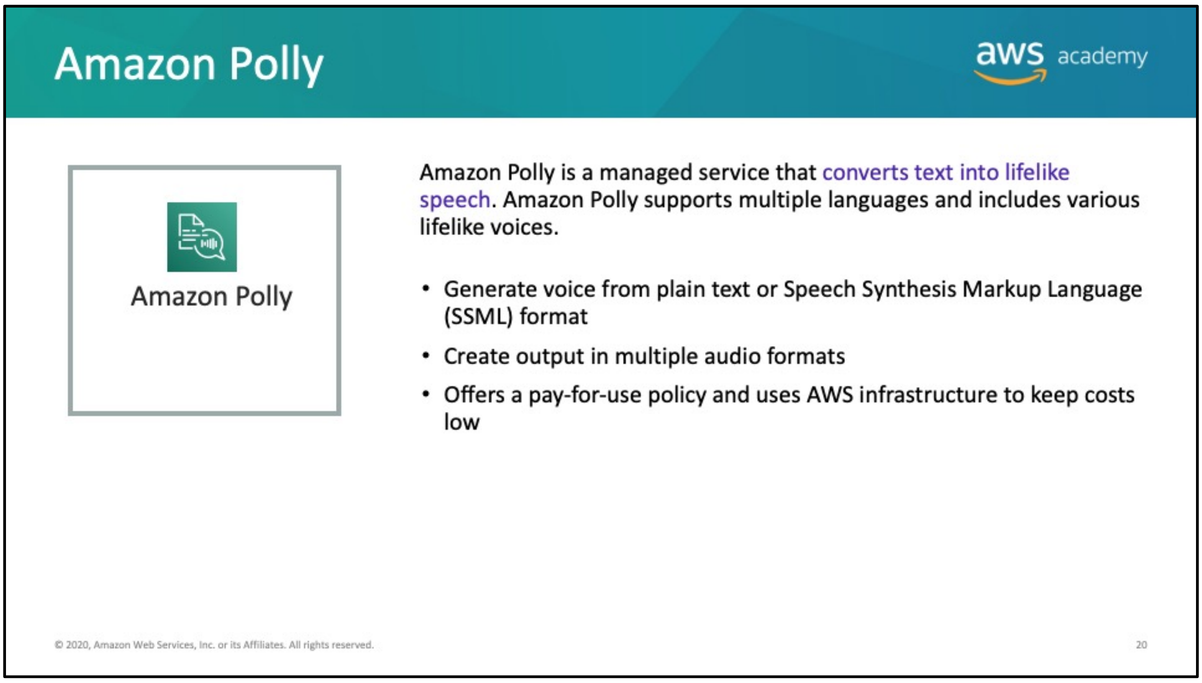
Subtitles

Streaming content labeling

Call center monitoring

Some of the more common use cases for Amazon Transcribe include:

- Medical transcription – Medical professionals can record their notes, and Amazon Transcribe can capture their spoken notes as text.
- Video subtitles – Video production organizations can generate subtitles automatically from video. It can also be done in real time for a live feed to add closed captioning (CC).
- Streaming content labeling – Media companies can capture and label content, and then feed the content into Amazon Comprehend for further analysis.
- Customer call center monitoring – Companies can record customer service or sales calls, and then analyze the results for training or strategic opportunities.



Amazon Polly

Amazon Polly

Amazon Polly is a managed service that converts text into lifelike speech. Amazon Polly supports multiple languages and includes various lifelike voices.

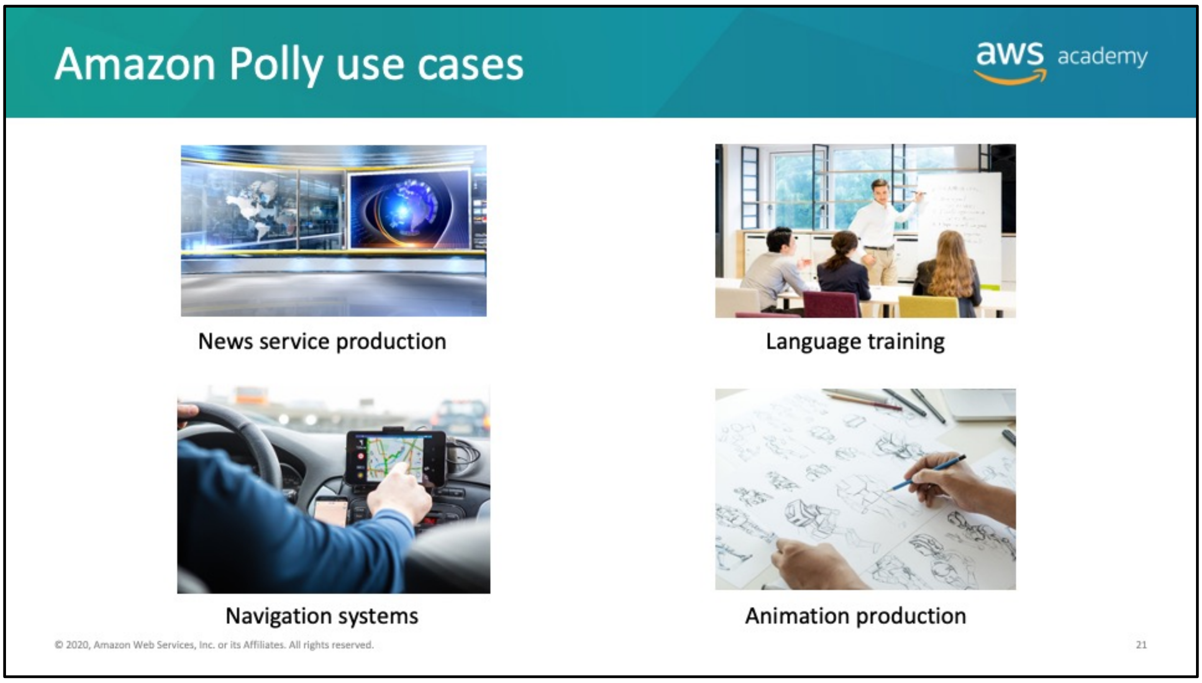
- Generate voice from plain text or Speech Synthesis Markup Language (SSML) format
- Create output in multiple audio formats
- Offers a pay-for-use policy and uses AWS infrastructure to keep costs low

Amazon Polly can convert text into lifelike speech. You can input either plaintext files or a file in Speech Synthesis Markup Language (SSML) format. SSML is a markup language that you can use to provide special instructions for how speech should sound. For example, you might want to introduce a pause in the flow of speech. You can add an SSML tag to instruct Amazon Polly to pause between two words.

You can also output speech from Amazon Polly to MP3, Vorbis, and pulse-code modulation (PCM) audio stream formats.

Amazon Polly has various applications. Common use cases for Amazon Polly include mobile apps (such as newsreaders), games, eLearning platforms, and accessibility applications for visually impaired people.

Amazon Polly is eligible for use with regulated workloads for the U.S. Health Insurance Portability and Accountability Act of 1996 (HIPAA), and the Payment Card Industry Data Security Standard (PCI DSS).



Amazon Polly use cases

News service production

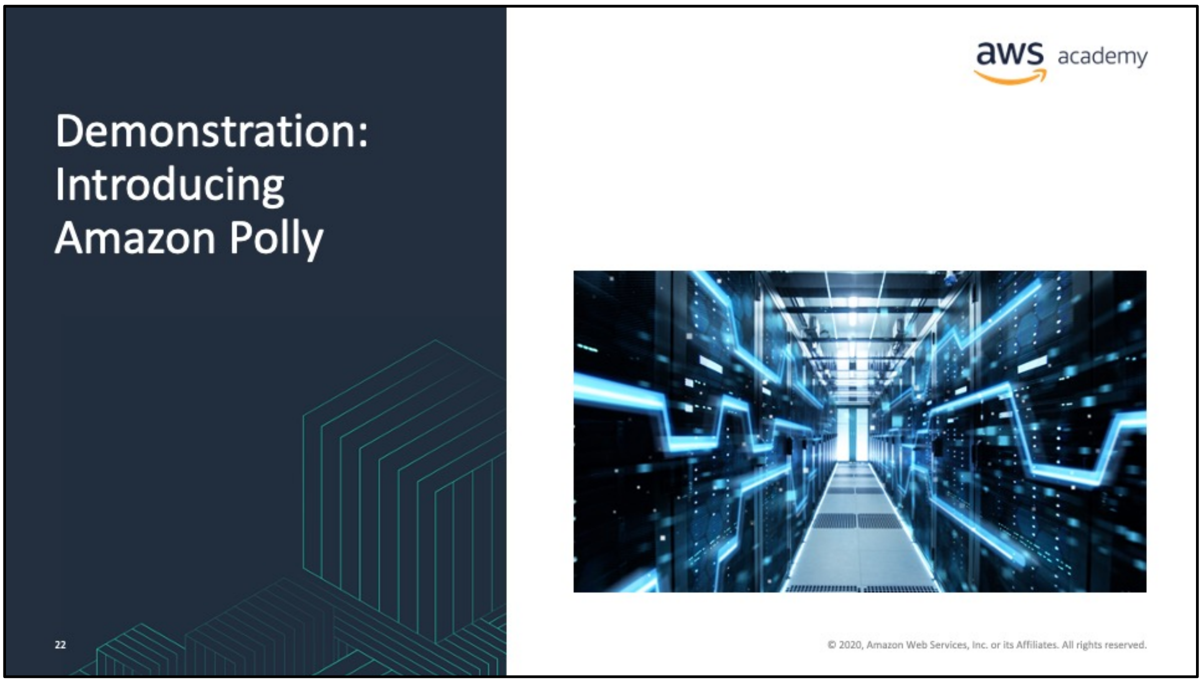
Language training

Navigation systems

Animation production

Some of the more common use cases for Amazon Polly include:

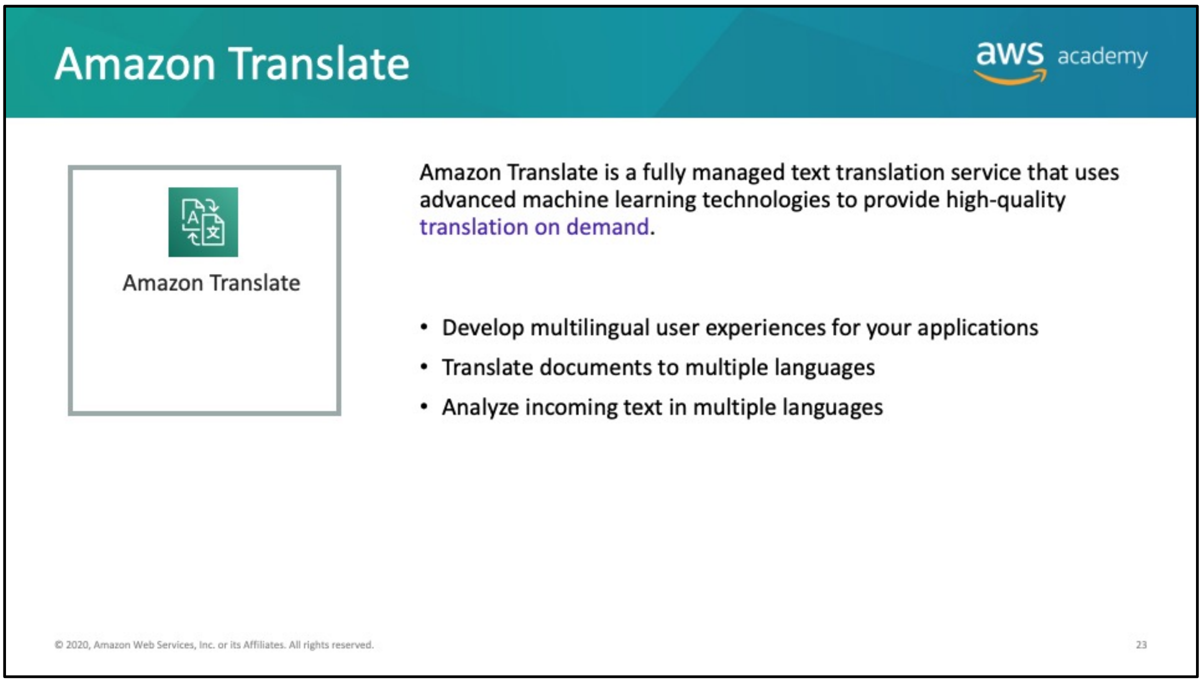
- News service production – Major news companies use Amazon Polly to generate vocal content directly from their written stories.
- Language training systems – Language training companies use Amazon Polly to create systems for learning a new language.
- Navigation systems – Amazon Polly is embedded in mapping application programming interfaces (APIs) so that developers can add voice to their geo-based applications.
- Animation production – Animators use Amazon Polly to add voices to their characters.



Demonstration: Introducing Amazon Polly

22

Your instructor will now either demonstrate Amazon Polly or provide you with access to a recorded demonstration.



Amazon Translate

aws academy

Amazon Translate

Amazon Translate is a fully managed text translation service that uses advanced machine learning technologies to provide high-quality translation on demand.

- Develop multilingual user experiences for your applications
- Translate documents to multiple languages
- Analyze incoming text in multiple languages

With Amazon Translate, you can create multilanguage experiences in your applications. You can create systems for reading documents in one language, and then render or storing it in another language. You can also use Amazon Translate as part of a document analysis system.

Amazon Translate is fully integrated with other Amazon ML services, such as Amazon Comprehend, Amazon Transcribe, and Amazon Polly. With this integration, you can:

- Extract named entities, sentiment, and key phrases by integrating with Amazon Comprehend
- Create multilingual subtitles with Amazon Transcribe
- Speak translated content with Amazon Polly

Amazon Translate use cases

International websites

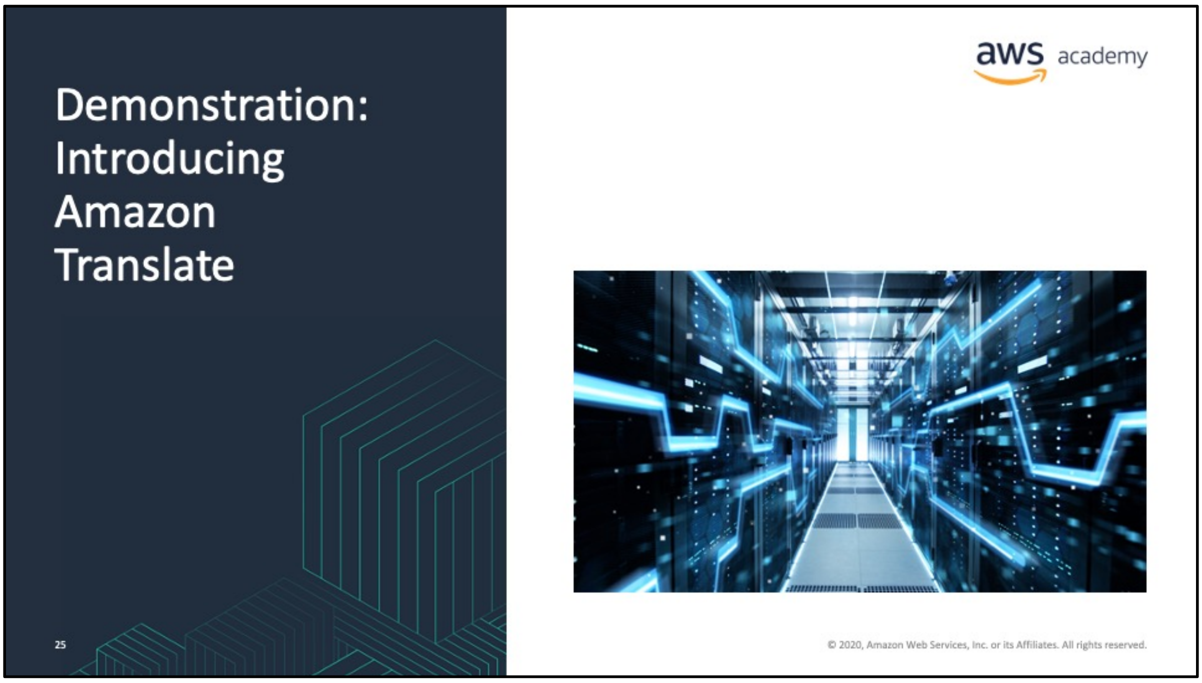
Software localization

Multilingual chatbots

International media

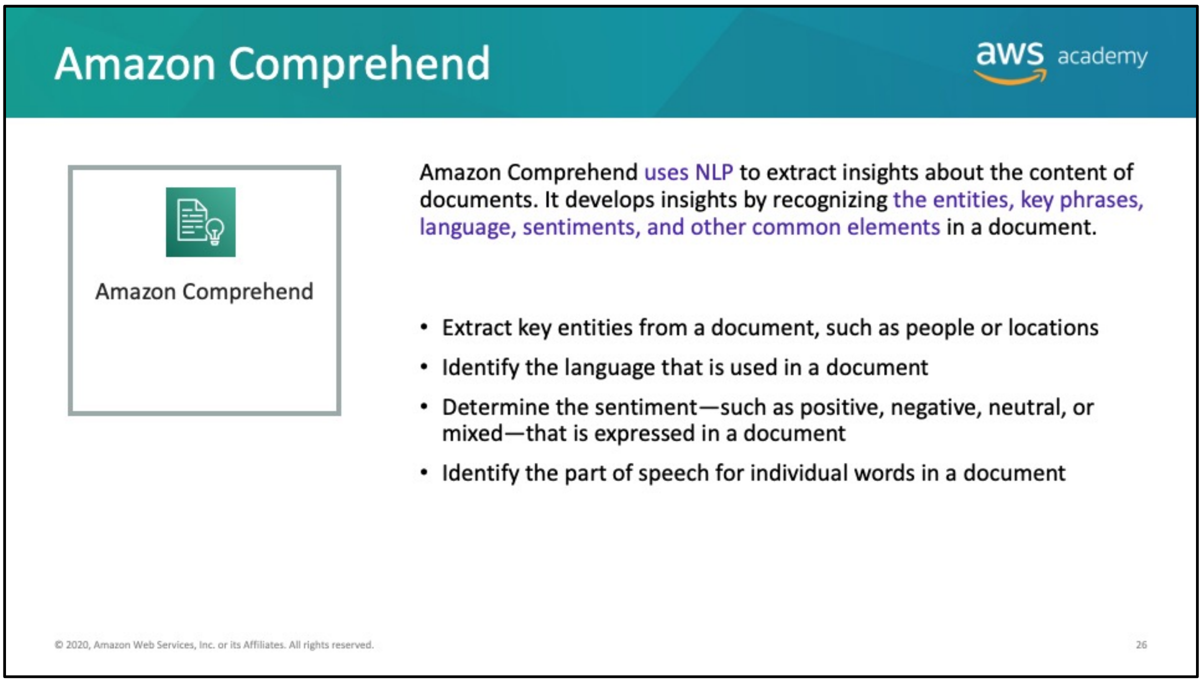
Some of the more common use cases for Amazon Translate include:

- International websites – You can use Amazon Translate to quickly globalize your websites.
- Software localization – Localization is a major cost for all software that is aimed at a global audience. Amazon Translate can decrease software development time and significantly reduce costs for localizing software.
- Multilingual chatbots – Chatbots are used to create a more human-like interface to applications. With Amazon Translate, you can create a chatbot that speaks multiple languages.
- International media management – Companies that manage media for a global audience use Amazon Translate to reduce their costs for localization.



Demonstration: Introducing Amazon Translate

Your instructor will now either demonstrate Amazon Translate or provide you with access to a recorded demonstration.

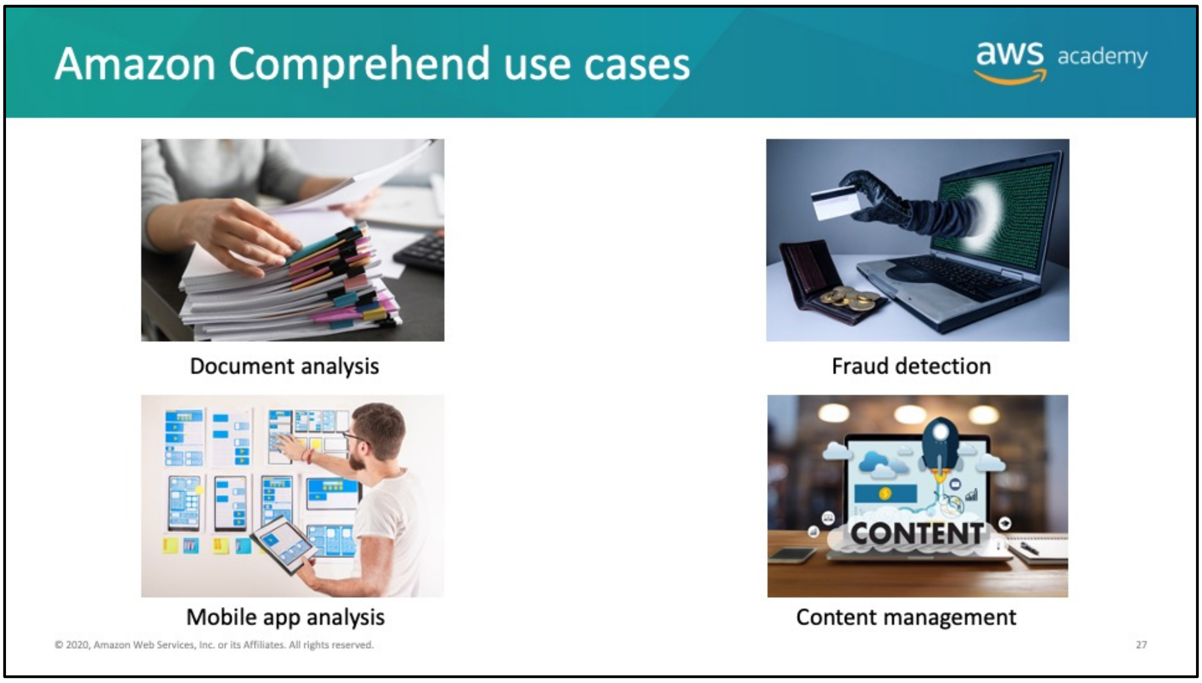


Amazon Comprehend

Amazon Comprehend uses NLP to extract insights about the content of documents. It develops insights by recognizing the entities, key phrases, language, sentiments, and other common elements in a document.

- Extract key entities from a document, such as people or locations
- Identify the language that is used in a document
- Determine the sentiment—such as positive, negative, neutral, or mixed—that is expressed in a document
- Identify the part of speech for individual words in a document

Amazon Comprehend implements many of the NLP techniques that you reviewed earlier in this module. You can extract key entities, perform sentiment analysis, and tag words with parts of speech.



Amazon Comprehend use cases

aws academy

Document analysis

Fraud detection

Mobile app analysis

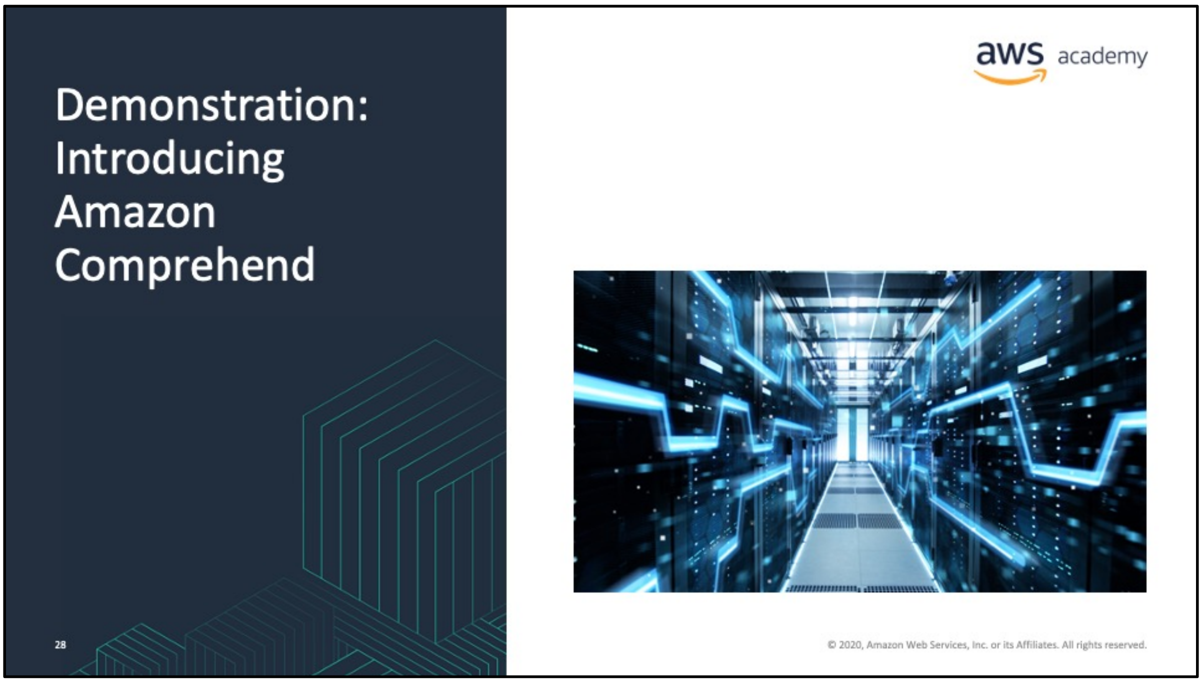
Content management

CONTENT

27

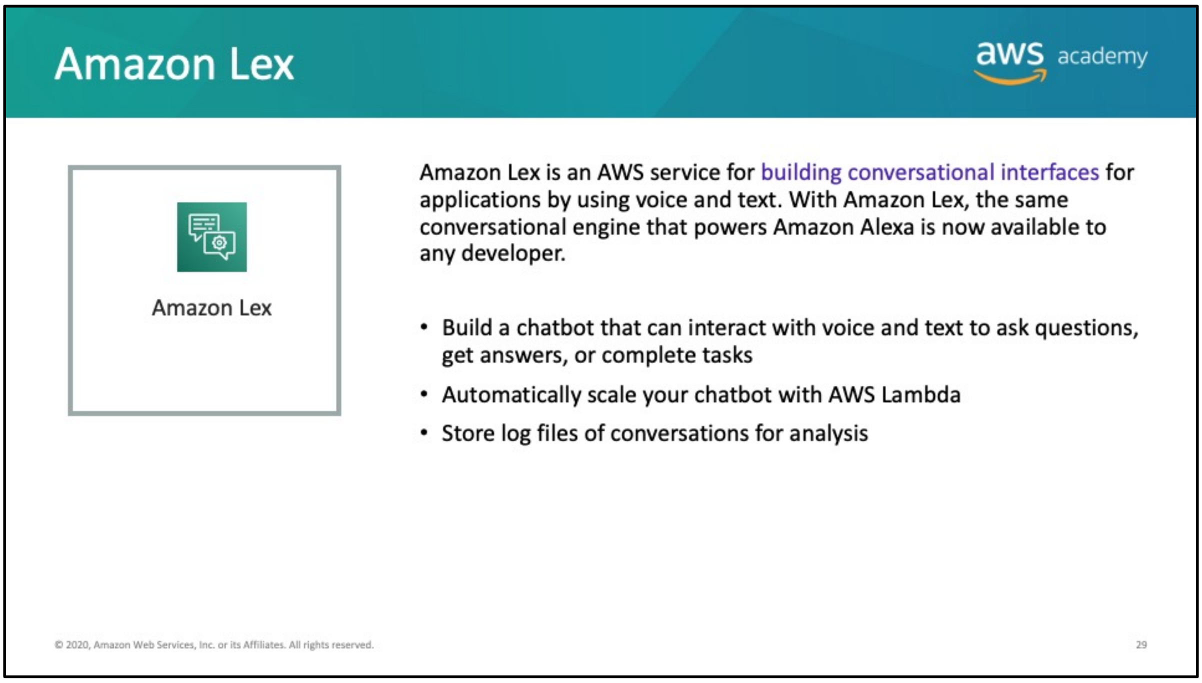
Some of the more common use cases for Amazon Comprehend include:

- Analysis of legal and medical documents – Legal, insurance, and medical organizations have used Amazon Comprehend to perform many of the NLP functions that you learned about in this module.
- Financial fraud detection – Banking, financial, and other institutions have used Amazon Comprehend to examine very large datasets of financial transactions to uncover fraud and look for patterns of illegal transactions.
- Large-scale mobile app analysis – Developers of mobile apps can use Amazon Comprehend to look for patterns in how their apps are used so they can design improvements.
- Content management – Media and other content companies can use Amazon Comprehend to tag content for analysis and management purposes.



Demonstration: Introducing Amazon Comprehend

Your instructor will now either demonstrate Amazon Comprehend or provide you with access to a recorded demonstration.



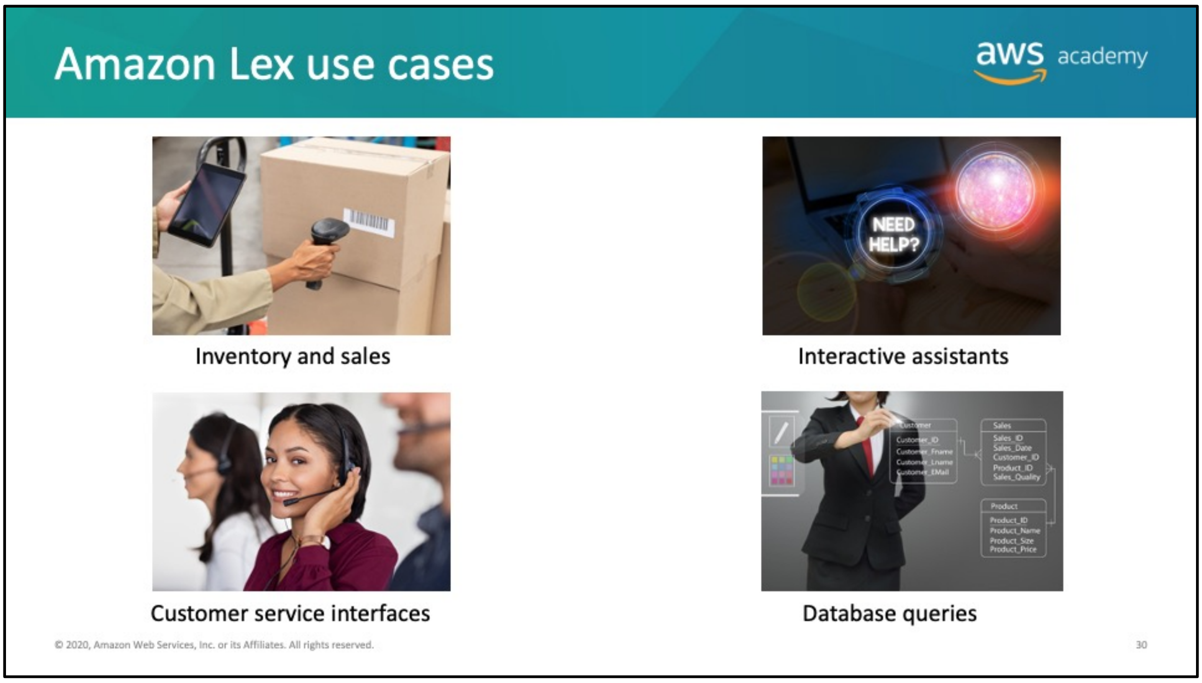
Amazon Lex

aws academy
Amazon Lex

Amazon Lex is an AWS service for building conversational interfaces for applications by using voice and text. With Amazon Lex, the same conversational engine that powers Amazon Alexa is now available to any developer.

- Build a chatbot that can interact with voice and text to ask questions, get answers, or complete tasks
- Automatically scale your chatbot with AWS Lambda
- Store log files of conversations for analysis

With Amazon Lex, you can add a human language frontend to your applications. Amazon Lex enables you to use the same conversational engine that powers Amazon Alexa. You can automatically increase capacity by creating AWS Lambda functions to scale on demand. In addition, you can store log files of the conversations for further analysis.



Amazon Lex use cases

Inventory and sales

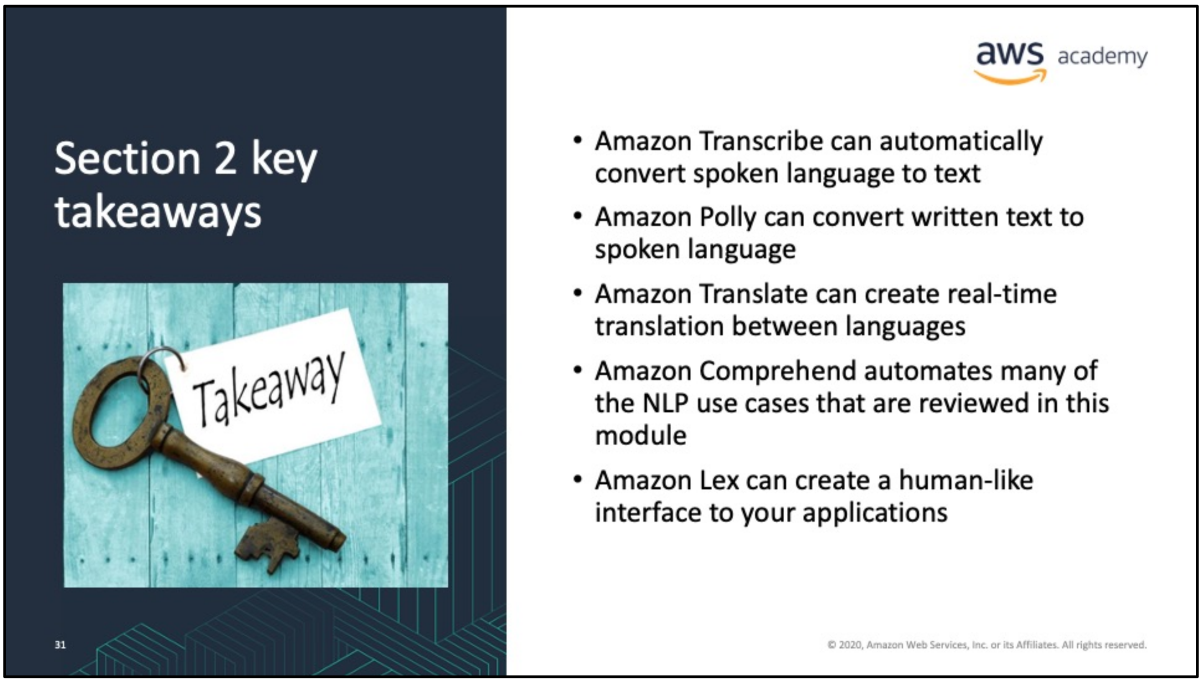
Interactive assistants

Customer service interfaces

Database queries

Some of the more common use cases for Amazon Lex include:

- Building frontend interfaces for inventory management and sales – Voice interfaces are becoming more common. Companies use Amazon Lex to add chatbots to their inventory and sales applications.
- Developing interactive assistants – By combining Amazon Lex with other ML services, customers create more sophisticated assistants for many different industries.
- Creating customer service interfaces – Human-like voice applications are quickly becoming the norm for customer service applications. Amazon Lex can reduce the time and increase the quality of these chatbots.
- Query databases with a human-like language – Amazon Lex is combined with other AWS database services to create sophisticated data analysis applications with a human-like language interface.

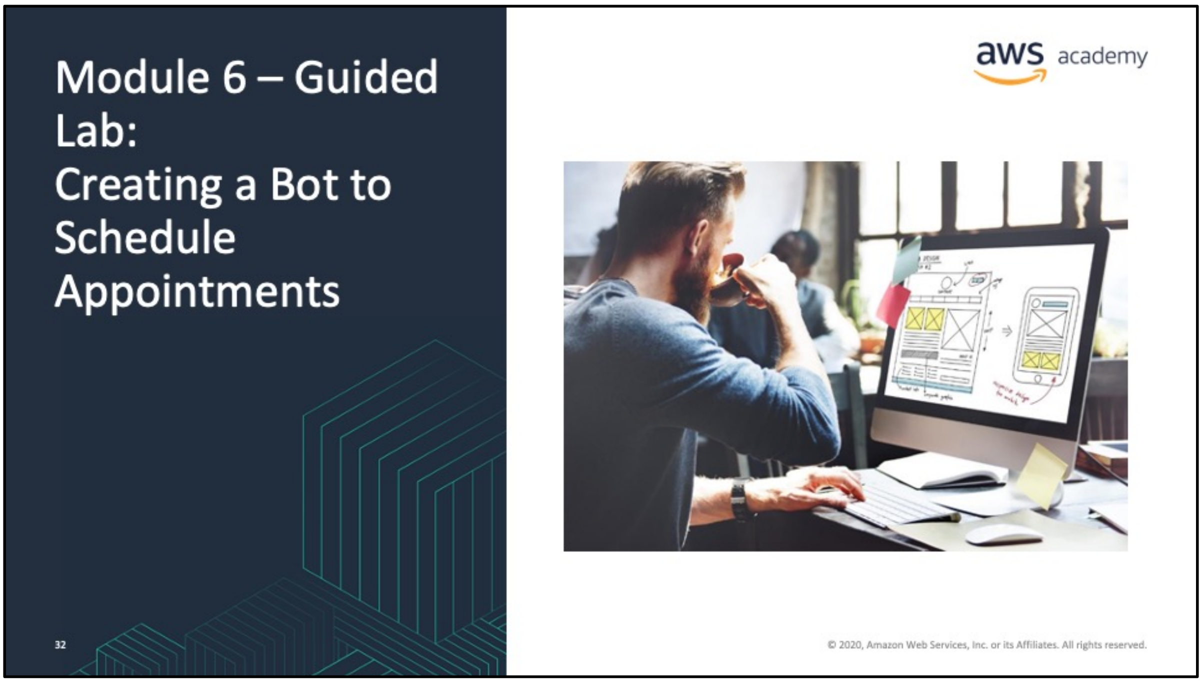


Section 2 key takeaways

- Amazon Transcribe can automatically convert spoken language to text
- Amazon Polly can convert written text to spoken language
- Amazon Translate can create real-time translation between languages
- Amazon Comprehend automates many of the NLP use cases that are reviewed in this module
- Amazon Lex can create a human-like interface to your applications

Some key takeaways from this section of the module include:

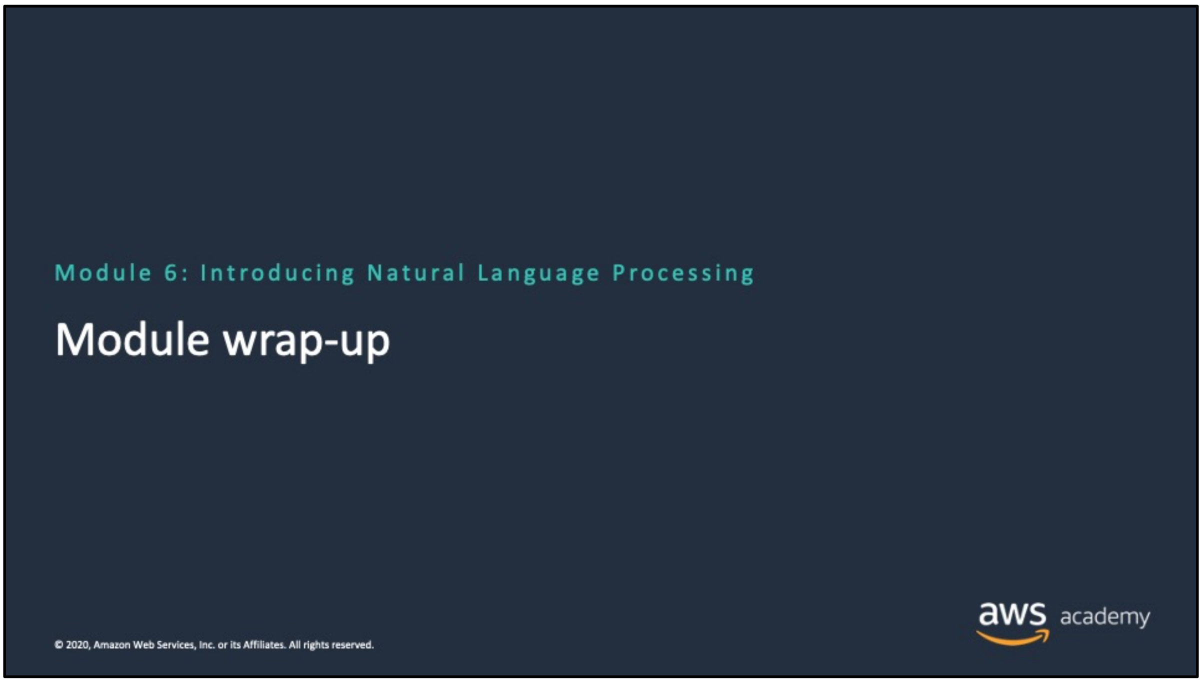
- Amazon Transcribe can automatically convert spoken language to text.
- Amazon Polly can convert written text to spoken language.
- Amazon Translate can create real-time translation between languages.
- Amazon Comprehend automates many of the NLP use cases that are reviewed in this module.
- Amazon Lex can create a human-like interface to your applications.



Module 6 – Guided Lab: Creating a Bot to Schedule Appointments

aws academy

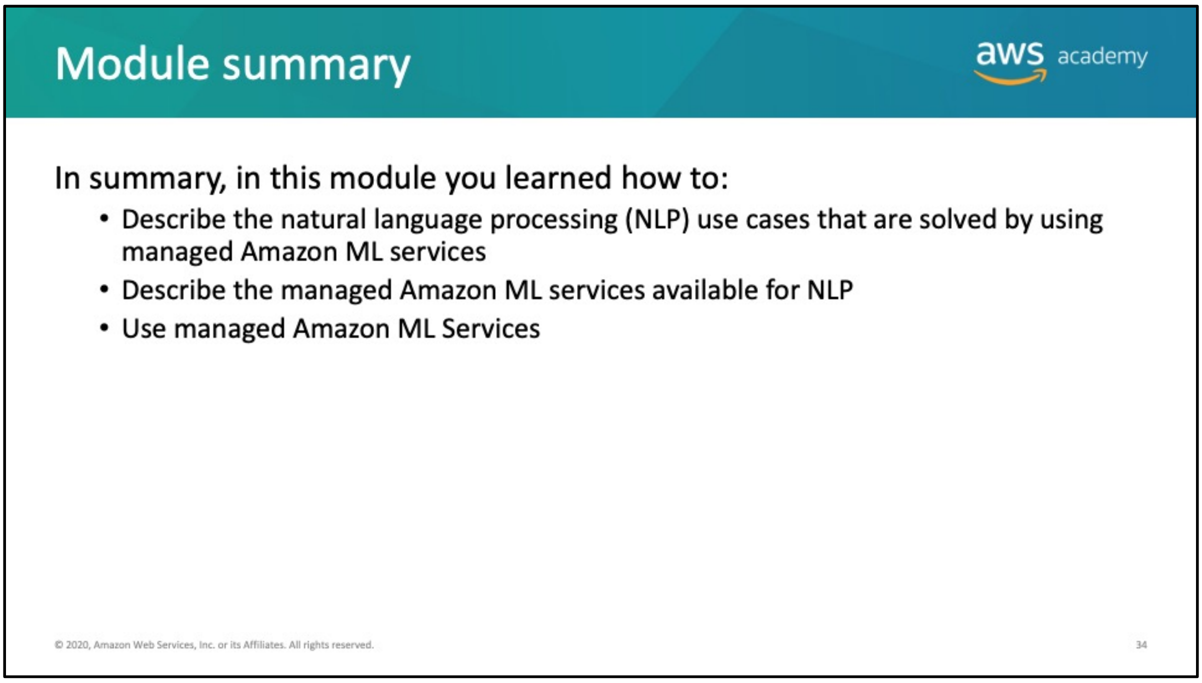
You will now complete Module 6 – Guided Lab: Creating a Bot to Schedule Appointments.



Module 6: Introducing Natural Language Processing

Module wrap-up

It’s now time to review the module and wrap up with a knowledge check.



Module summary

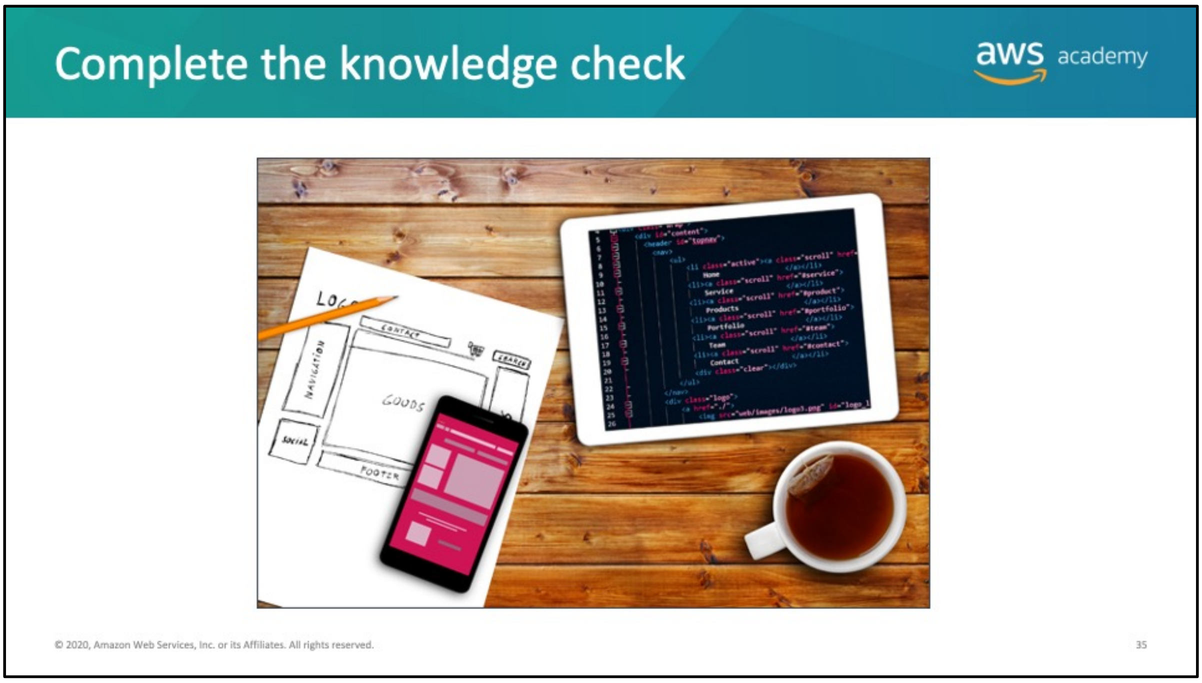
aws academy

In summary, in this module you learned how to:

- Describe the natural language processing (NLP) use cases that are solved by using managed Amazon ML services
- Describe the managed Amazon ML services available for NLP
- Use managed Amazon ML Services

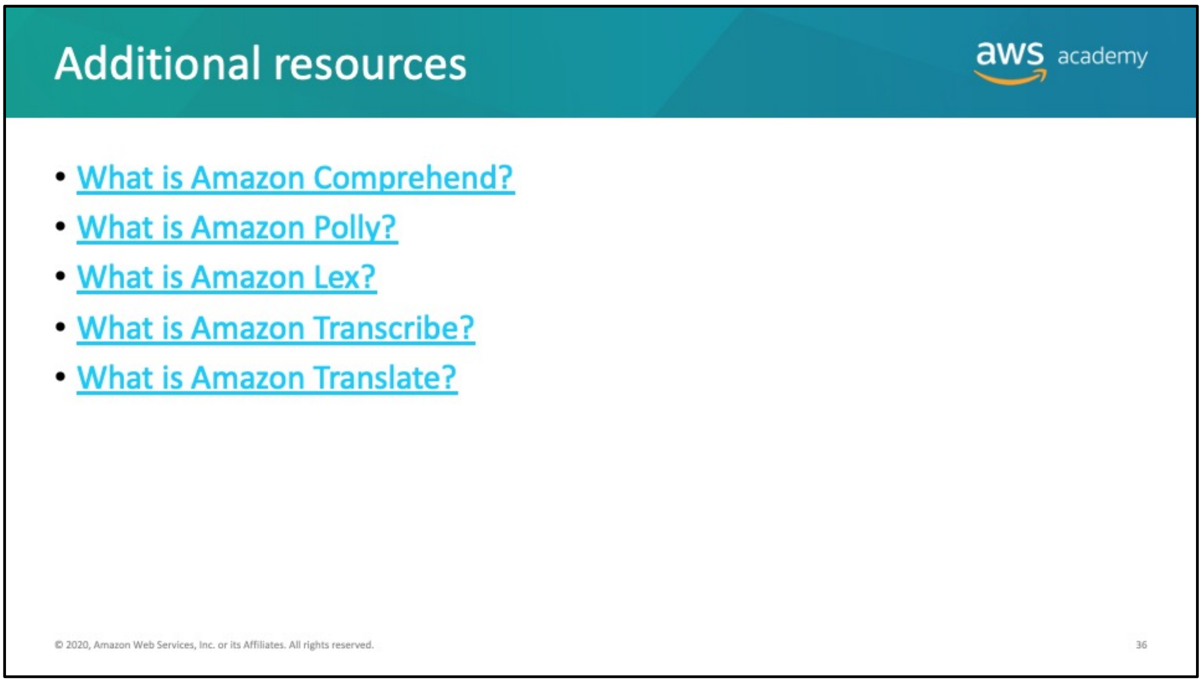
In summary, in this module, you learned how to:

- Describe the natural language processing (NLP) use cases that are solved by using managed Amazon ML services
- Describe the managed Amazon ML services available for NLP
- Use managed Amazon ML Services



Complete the knowledge check

It is now time to complete the knowledge check for this module.

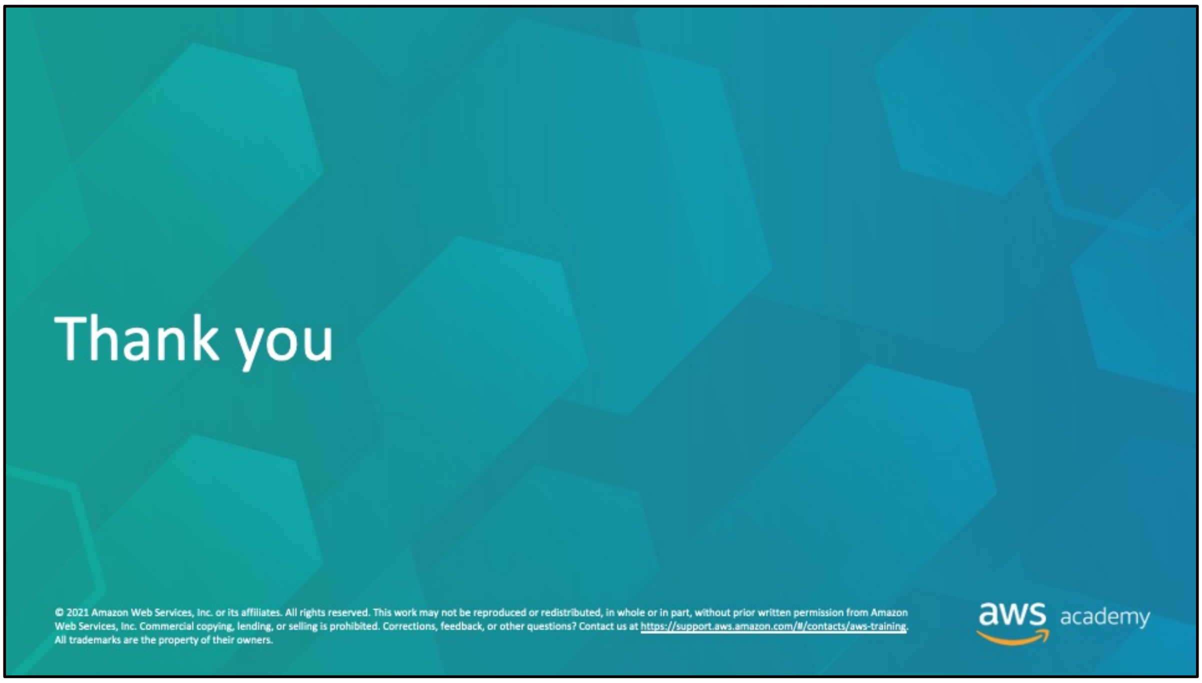


Additional resources

- What is Amazon Comprehend?
- What is Amazon Polly?
- What is Amazon Lex?
- What is Amazon Transcribe?
- What is Amazon Translate?

If you want to learn more about the topics that are covered in this module, you might find the following resources helpful:

- What is Amazon Comprehend?
- What is Amazon Polly?
- What is Amazon Lex?
- What is Amazon Transcribe?
- What is Amazon Translate?



Thank you

Thank you for completing this module.