databricks

Large Language Models

Application through Production

©2023 Databricks Inc. — All rights reserved

Course Outline

©2023 Databricks Inc. — All rights reserved

Course Outline

Course Introduction

Module 1 - Applications with LLMs

Module 2 - Embeddings, Vector Databases, and Search

Module 3 - Multi-stage Reasoning

Module 4 - Fine-tuning and Evaluating LLMs

Module 5 - Society and LLMs

Module 6 - LLMOps

©2023 Databricks Inc. — All rights reserved

Course Introduction

©2023 Databricks Inc. — All rights reserved

Before we begin

1. Introduction by Matei Zaharia: Why LLMs?
2. Primer on NLP
3. Setting up your Databricks lab environment

©2023 Databricks Inc. — All rights reserved

Why LLMs?

Matei Zaharia

Co-founder & CTO of Databricks

Associate Professor of Computer Science
at Stanford University

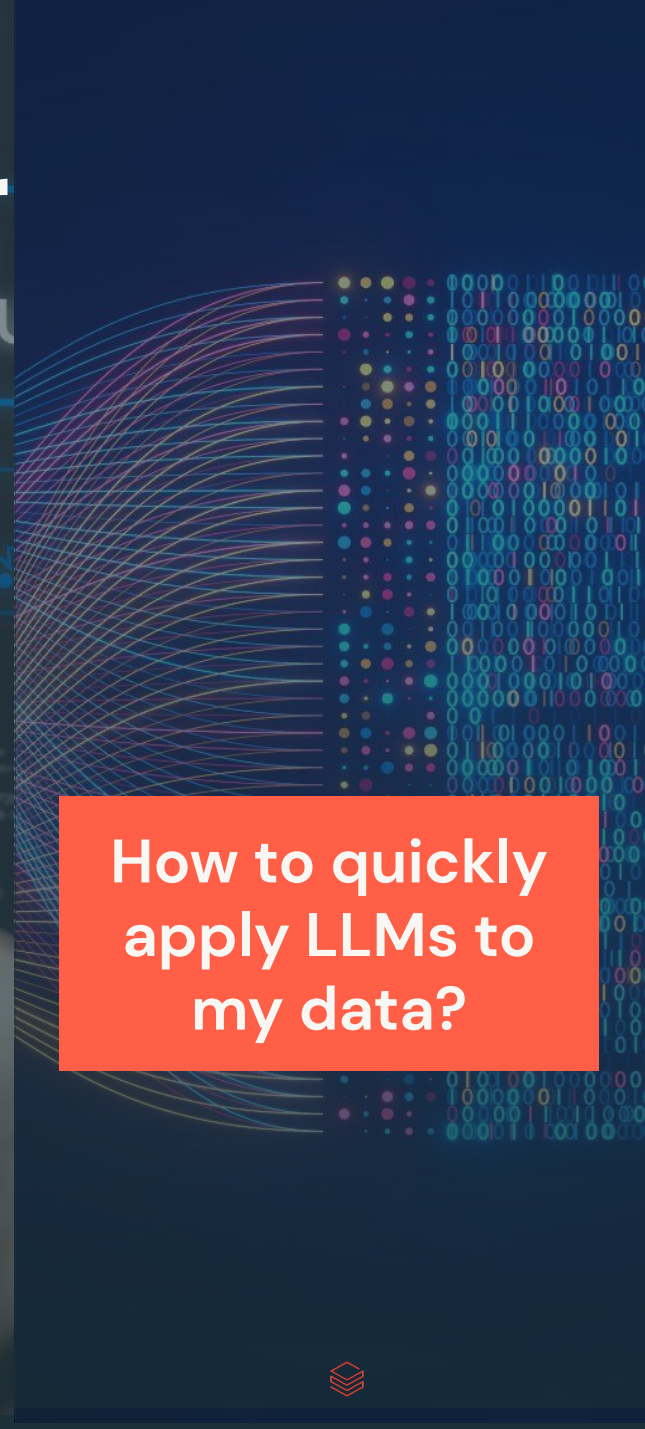
©2023 Databricks Inc. — All rights reserved

Questions we hear about LLMs

Is the LLM hype real? Is this an iPhone moment?

Are LLMs a threat or an opportunity?

How to leverage LLMs to gain a competitive advantage?



How to quickly apply LLMs to my data?

©2023 Databricks Inc. — All rights reserved | Confidential and proprietary

LLMs are more than hype

They are revolutionizing every industry

"Chegg shares drop more than 40% after company says ChatGPT is killing its business"

CNBC 05/02/2023
Link

"[...] ask GitHub Copilot to explain a piece of code. Bump into an error? Have GitHub Copilot fix it. It'll even generate unit tests so you can get back to building what's next."

 03/22/2023*
Link

"[YouChat is an] AI search assistant that you can talk to right in your search results. It stays up-to-date with the news and cites its sources so that you can feel confident in its answers."

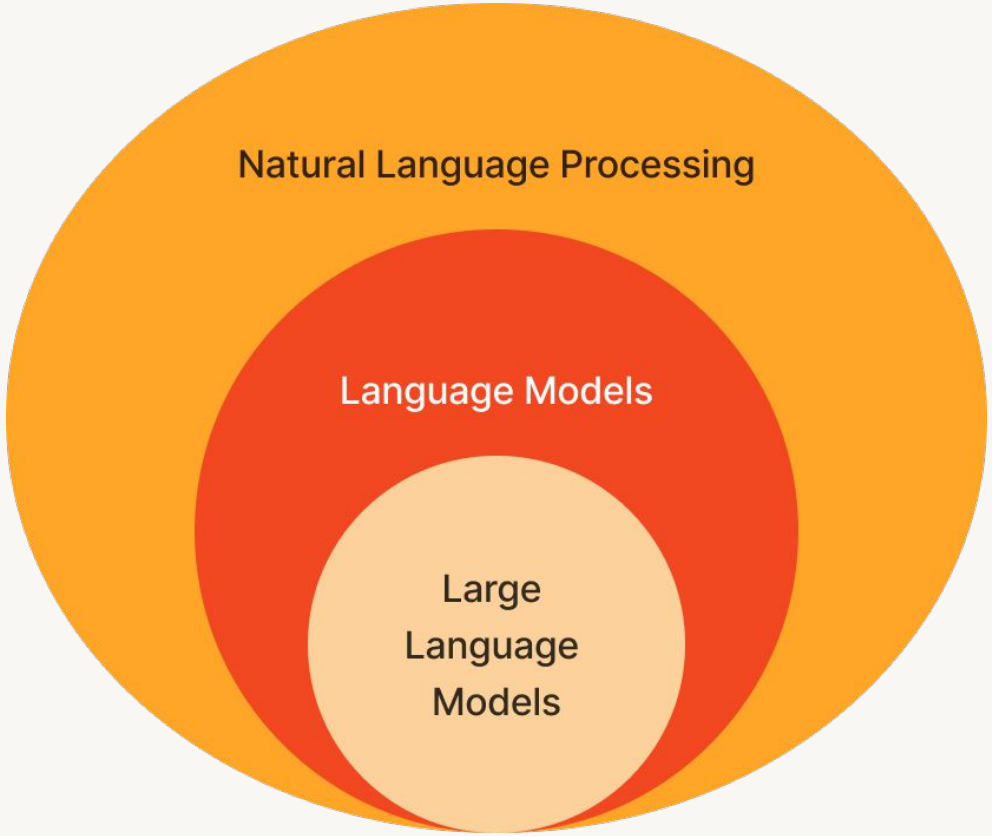
YOU 12/23/2022
Link

©2023 Databricks Inc. — All rights reserved

*Announcement date instead of article date

LLMs are not *that* new

Why should I care now?



Accuracy and effectiveness has hit a tipping point

- Many new use cases are unlocked!
- Accessible by all.

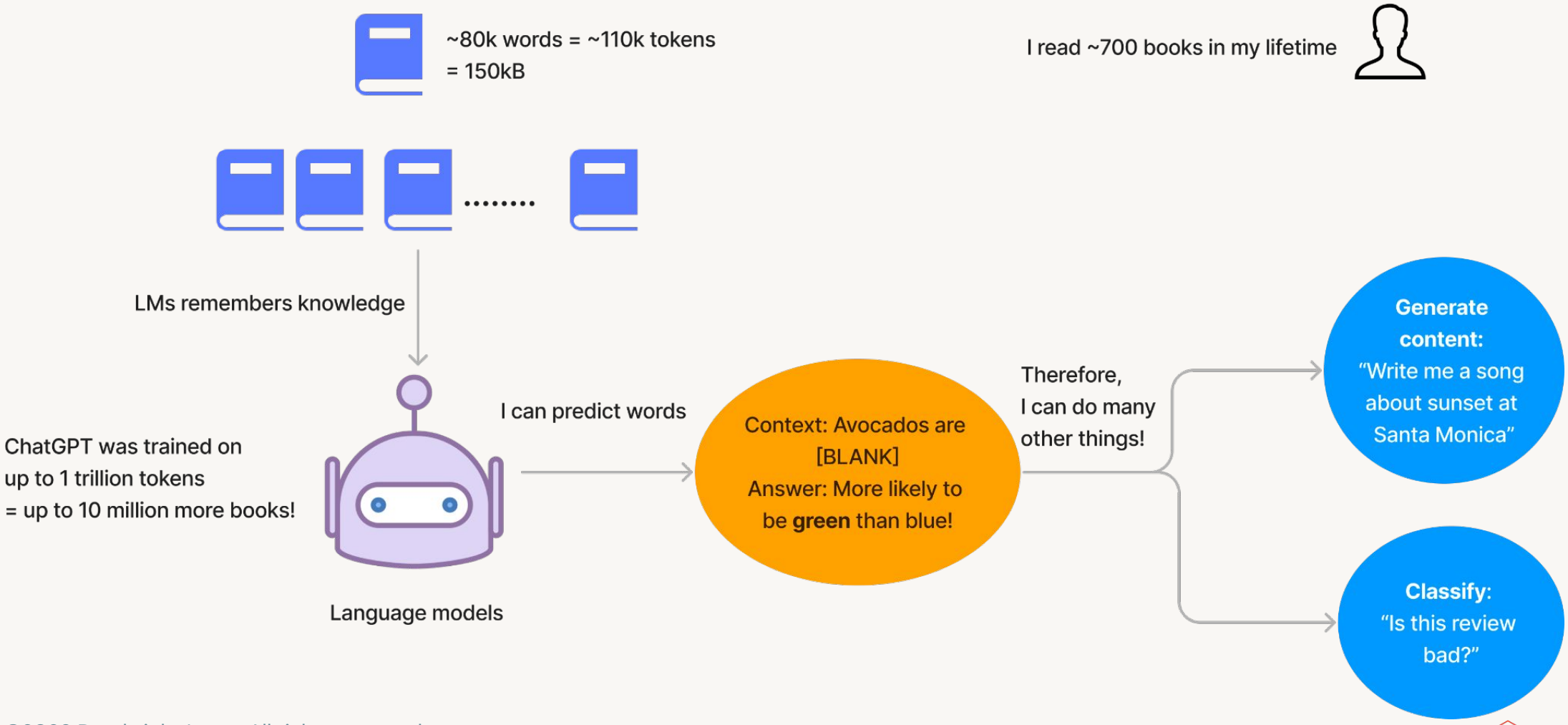
Readily available data and tooling

- Large datasets.
- Open-sourced model options.
- Requires powerful GPUs, but are available on the cloud.

©2023 Databricks Inc. — All rights reserved

What is an LLM?

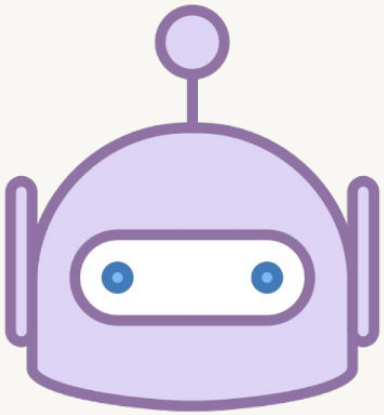
It's a *large* language model trained on *enormous* data



©2023 Databricks Inc. — All rights reserved

What does that mean for me?

LLMs *automate* many human-led tasks



I can accelerate innovation

I can increase ROI

- Faster software development
- More users can leverage AI
- More use cases
- Reduce development cost
- Reduce monotonous tasks

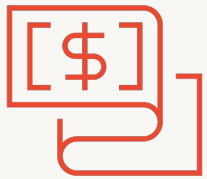
©2023 Databricks Inc. — All rights reserved

Choose the right LLM

There is no “perfect” model. Trade-offs are required.

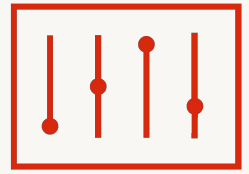
Decision criteria

Model Quality



Serving Cost

Serving Latency



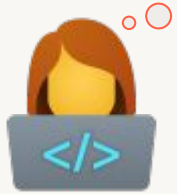
Customizability

©2023 Databricks Inc. — All rights reserved

Who is this course for?

Bridging the gap between black-box solutions and academia for practitioners

Exec:
We need to add LLMs



You:
"Where do I start?"

Academic Materials

This Course

SaaS API Materials

Base Theory/Algorithms — Build Your Own — Black-Box Solutions

©2023 Databricks Inc. — All rights reserved

Enjoy the course!

©2023 Databricks Inc. — All rights reserved

Before we begin

1. Introduction by Matei Zaharia: Why LLMs?
2. Primer on NLP
3. Setting up your Databricks lab environment

©2023 Databricks Inc. — All rights reserved

Primer on NLP

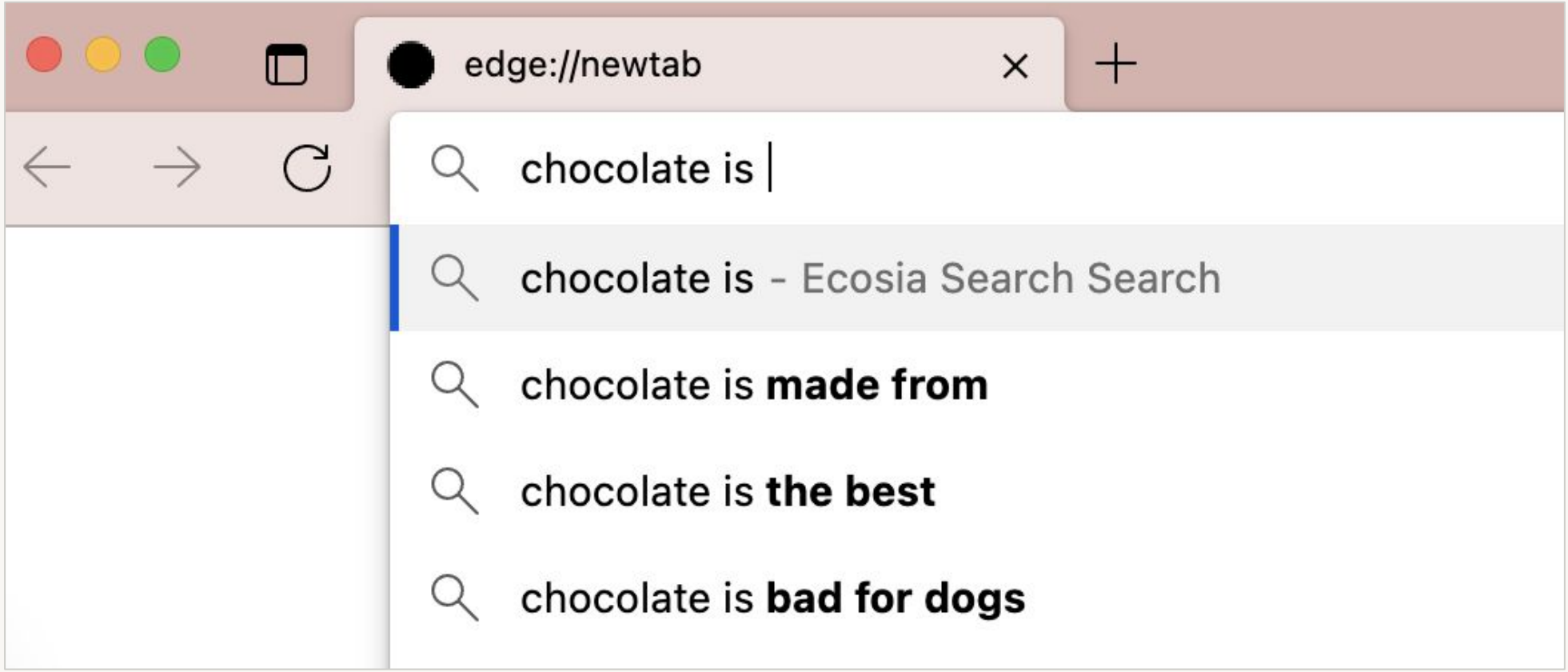
©2023 Databricks Inc. — All rights reserved

Natural Language Processing

What is NLP?

©2023 Databricks Inc. — All rights reserved

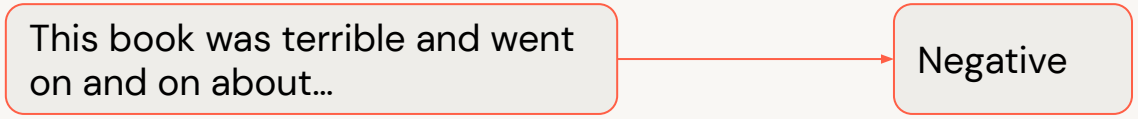
We use NLP everyday



©2023 Databricks Inc. — All rights reserved

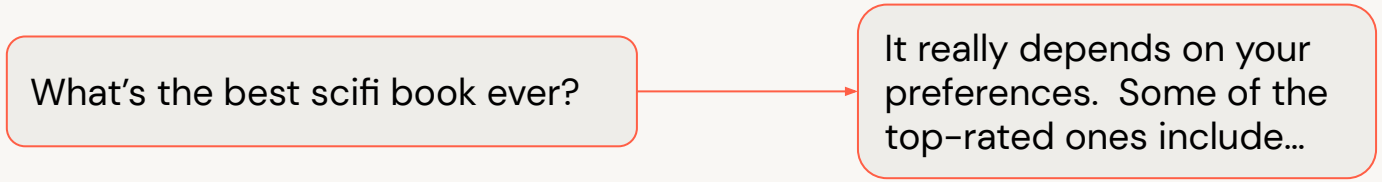
NLP is useful for a variety of domains

Sentiment analysis: product reviews



Translation

Question answering: chatbots



Other use cases

Semantic similarity

- Literature search.
- Database querying.
- Question-Answer matching.

Summarization

- Clinical decision support.
- News article sentiments.
- Legal proceeding summary.

Text classification

- Customer review sentiments.
- Genre/topic classification.

©2023 Databricks Inc. — All rights reserved

Some useful NLP definitions

The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years.

Token

Basic building block

- The
- Moon
- ,
- Earth’s
- Only
- …..
- years

Sequence

Sequential list of tokens

- The moon,
- Earth’s only natural satellite
- Has been a subject of
- ….
- Thousands of years

Vocabulary

Complete list of tokens

```
{
1:"The",
569:"moon",
122: ",",
430:"Earth",
50:"**’s",
…}
```

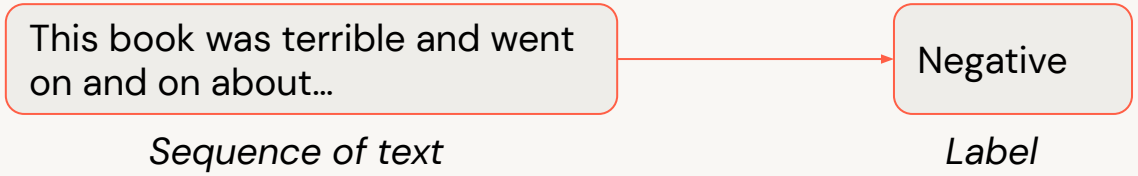
©2023 Databricks Inc. — All rights reserved

Types of sequence tasks

Translation

Sequence to sequence prediction

Sentiment analysis (product reviews)



*Sequence to **non sequence** prediction*

Question answering (chatbots)

*Sequence to sequence **generation***

©2023 Databricks Inc. — All rights reserved

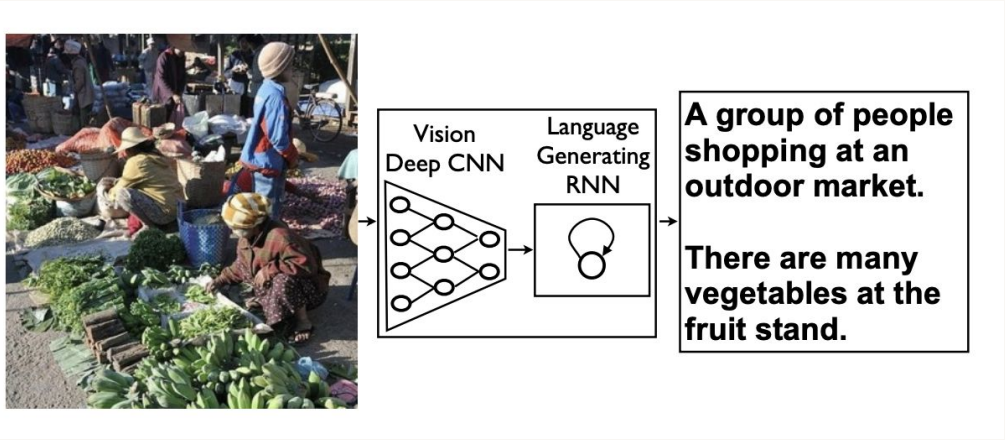
NLP goes beyond text

Speech recognition

Image caption generation

Image generation from text

...



©2023 Databricks Inc. — All rights reserved

Source: Show and Tell: A Neural Image Caption Generator

Text interpretation is challenging

"The ball hit the table and it broke."

Language is ambiguous.

Context can change the meaning.

"What's the best sci-fi book ever?"

There can be multiple good answers.

Input data format matters.
Lots of work has gone into text representation for NLP.

Model size matters.
Big models help to capture the diversity and complexity of human language.

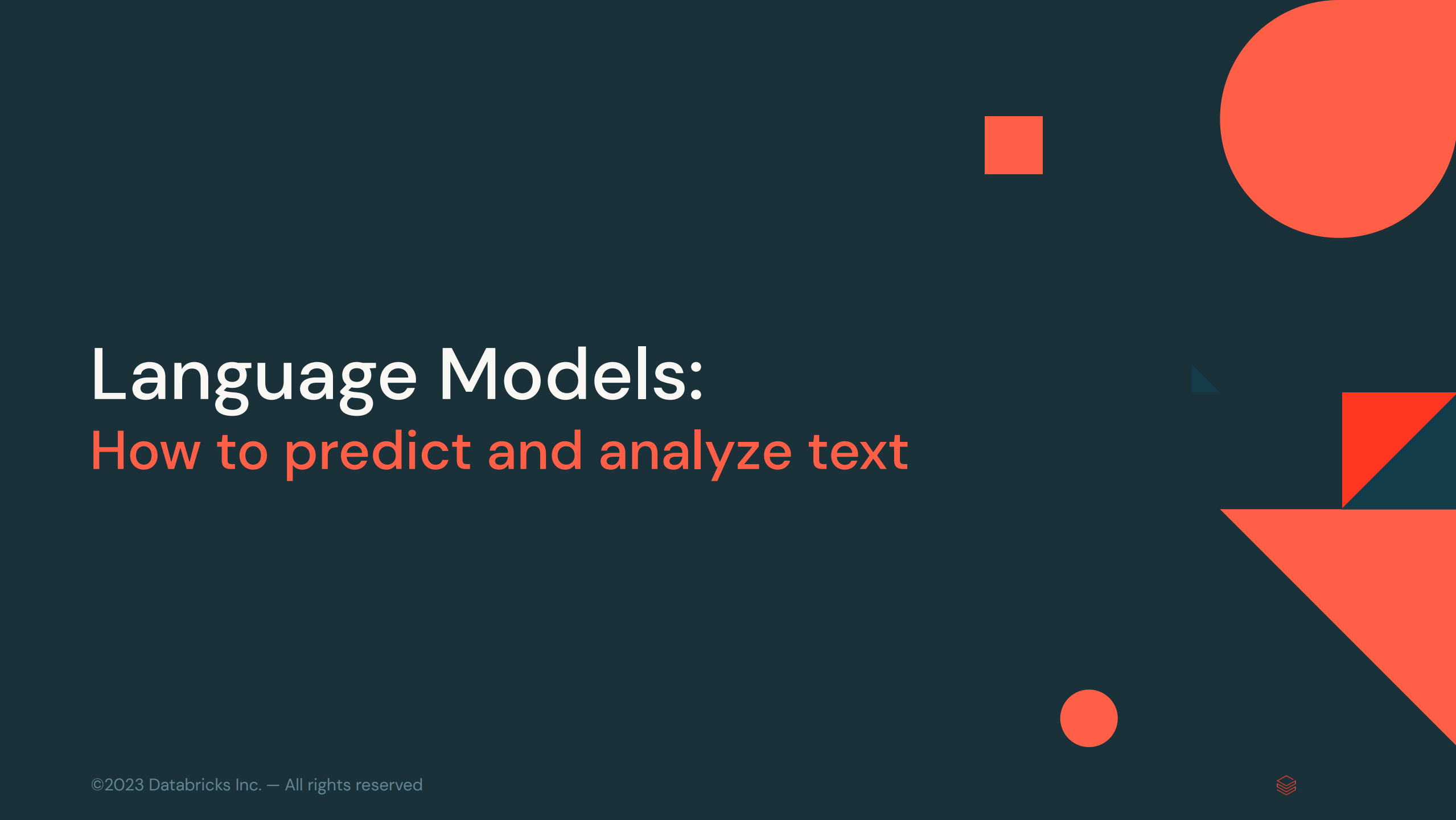
Training data matters.
It helps to have high-quality data and lots of it.

©2023 Databricks Inc. — All rights reserved

Language Models:

How to predict and analyze text

©2023 Databricks Inc. — All rights reserved



What is a Language Model?

The term **Large Language Models** is everywhere these days.
But let's take a closer look at that term:

Large **Language Model**—What is a Language Model?

Large Language Model—What about these makes them "larger" than other language models?

©2023 Databricks Inc. — All rights reserved

Source: txt.cohere.com

What is a Language Model?

LMs assign probabilities to word sequences: find the most likely word

Categories:

- **Generative**: find the most likely next word
- Classification: find the most likely classification/answer

©2023 Databricks Inc. — All rights reserved

What is a Large Language Model?

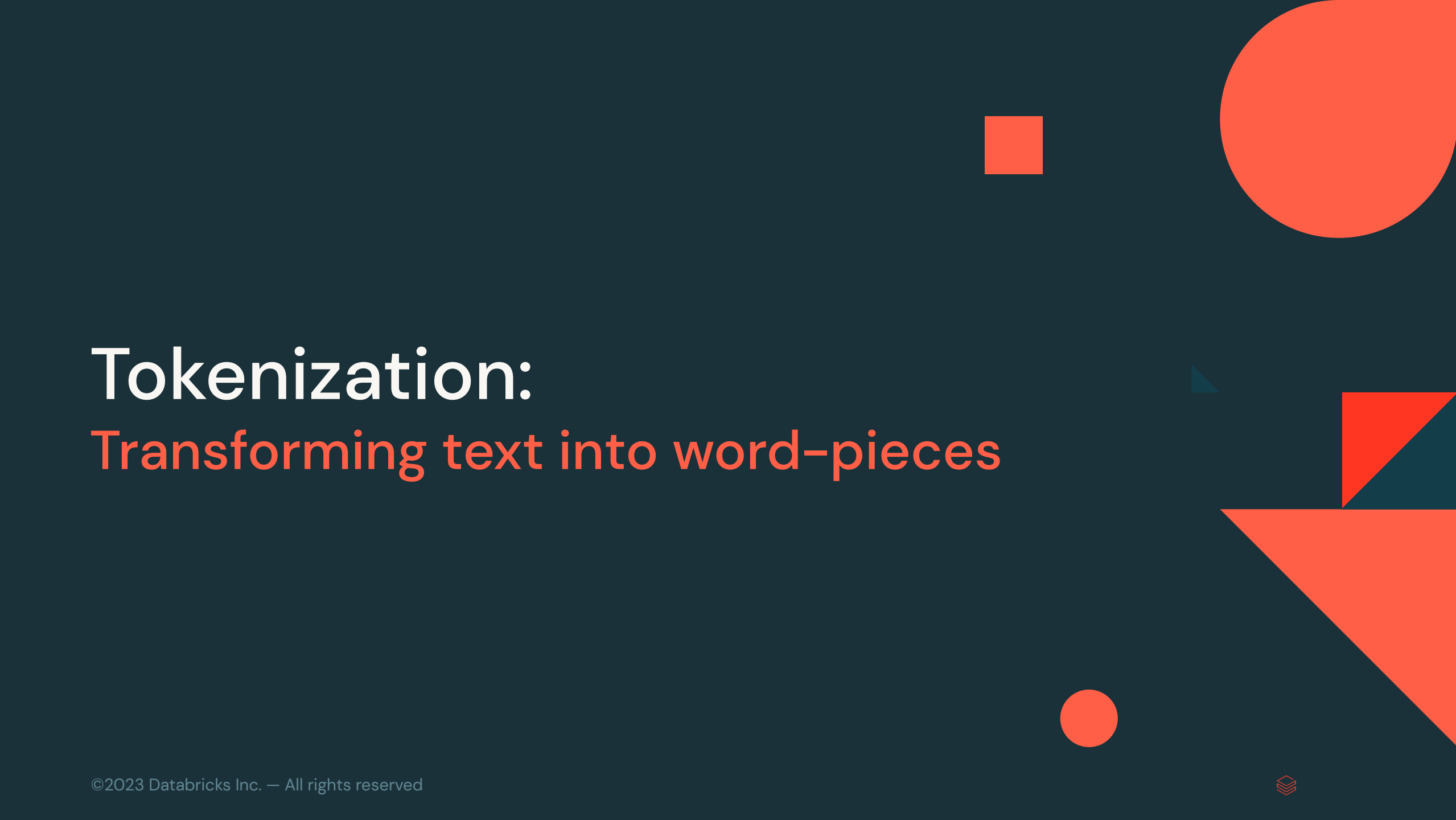
Language Model	Description	"Large"?	Emergence
Bag-of-Words Model	Represents text as a set of unordered words, without considering sequence or context	No	1950s-1960s
N-gram Model	Considers groups of N consecutive words to capture sequence	No	1950s-1960s
Hidden Markov Models (HMMs)	Represents language as a sequence of hidden states and observable outputs	No	1980s-1990s
Recurrent Neural Networks (RNNs)	Processes sequential data by maintaining an internal state, capturing context of previous inputs	No	1990s-2010s
Long Short-Term Memory (LSTM) Networks	Extension of RNNs that captures longer-term dependencies	No	2010s
Transformers	Neural network architecture that processes sequences of variable length using a self-attention mechanism	**Yes**	**2017-Present**

©2023 Databricks Inc. — All rights reserved

Tokenization:

Transforming text into word-pieces

©2023 Databricks Inc. — All rights reserved



Tokenization - Words

This vocab is too big!

The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years.

Corpus of training data used to build our vocabulary.

Building Vocabulary

Build index (dictionary of tokens = words)

```
a: 0
The: 1
is: 2
what: 3
I: 4
and: 5
...
```

Tokenization

Map tokens to indices

```
{The moon, Earth's only natural satellite ... }
{ [1], [45600], [8097], [43], [1323], [754] ... }
```

Pros

Intuitive.

Cons

Big vocabularies.

Complications such as handling misspellings and other out-of-vocabulary words.

©2023 Databricks Inc. — All rights reserved

Tokenization - Characters

This vocab is too small!

The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years.

Corpus of training data used to build our vocabulary.

Build index (dictionary of tokens = letters/characters)

```
a: 0
b: 1
c: 2
d: 3
e: 4
f: 5
...
```

Map tokens to indices

```
t   → 19
h   →  7
e   →  4
m   → 12
o   → 14
o   → 14
n   → 13
... → ...
```

<u>Pros</u>

Small vocabulary.

No out-of-vocabulary words.

<u>Cons</u>

Loss of context within words.

Much longer sequences for a given input.

©2023 Databricks Inc. — All rights reserved

Tokenization – Sub-words

This vocab is just right!

The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years.

Corpus of training data used to build our vocabulary.

Build index (dictionary of tokens = mix of words and sub-words)

```
a: 0
as: 1
ask: 2
be: 3
ca: 4
cd: 5
...
```

Map tokens to indices

The	→	319
moon	→	12
**,	→	391
Earth	→	178
**'s	→	198
on	→	79
ly	→	281
...	→	...

Byte Pair Encoding (BPE) a popular encoding.

Start with a small vocab of characters.

Iteratively merge frequent pairs into new bytes in the vocab (such as "b","e" → "be").

<u>Compromise</u>

"Smart" vocabulary built from characters which co-occur frequently.

More robust to novel words.

©2023 Databricks Inc. — All rights reserved

Tokenization

Tokenization method	Tokens	Token count	Vocab size
Sentence	‘The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years.’	1	# sentences in doc
Word	'The', 'moon,', "Earth's", 'only', 'natural', 'satellite,', 'has', 'been', 'a', 'subject', 'of', 'fascination', 'and', 'wonder', 'for', 'thousands', 'of', 'years.'	18	171K (English[1])
Sub-word	'The', 'moon', ',', 'Earth', "'", 's', 'on', 'ly', 'n', 'atur', 'al', 's', 'ate', 'll', 'it', 'e', ',', 'has', 'been', 'a', 'subject', 'of', 'fascinat', 'ion', 'and', 'w', 'on', 'd', 'er', 'for', 'th', 'ous', 'and', 's', 'of', 'y', 'ears', '.'	37	(varies)
Character	'T', 'h', 'e', ' ', 'm', 'o', 'o', 'n', ',', ' ', 'E', 'a', 'r', 't', 'h', "'", 's', ' ', 'o', 'n', 'l', 'y', ' ', 'n', 'a', 't', 'u', 'r', 'a', 'l', ' ', 's', 'a', 't', 'e', 'l', 'l', 'i', 't', 'e', ',', ' ', 'h', 'a', 's', ' ', 'b', 'e', 'e', 'n', ' ', 'a', ' ', 's', 'u', 'b', 'j', 'e', 'c', 't', ' ', 'o', 'f', ' ', 'f', 'a', 's', 'c', 'i', 'n', 'a', 't', 'i', 'o', 'n', ' ', 'a', 'n', 'd', ' ', 'w', 'o', 'n', 'd', 'e', 'r', ' ', 'f', 'o', 'r', ' ', 't', 'h', 'o', 'u', 's', 'a', 'n', 'd', 's', ' ', 'o', 'f', ' ', 'y', 'e', 'a', 'r', 's', '.'	110	52 + punctuation (English)

©2023 Databricks Inc. — All rights reserved

[1]Source: BBC.com

Word Embeddings:

The surprising power of similar context

©2023 Databricks Inc. — All rights reserved

Represent words with vectors

Words with similar meaning tend to occur in similar contexts:

The cat meowed at me for food.

The kitten meowed at me for treats.

The words cat and kitten share context here, as do food and treats.

If we use vectors to encode tokens we can attempt to store this meaning.

- Vectors are the basic inputs for many ML methods.
- Tokens that are similar in meaning can be positioned as neighbors in the vector space using the right mapping functions.

©2023 Databricks Inc. — All rights reserved

How to convert words into vectors?

Initial idea: Let's count the frequency of the words!

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

We now have length-6 vectors for each document:

- 'the cat sat' → [1 1 1 0 0 0]
- 'the cat sat in the hat' → [2 1 1 1 1 0]
- 'the cat with the hat' → [2 1 0 0 1 1]

BIG limitation: **SPARSITY**

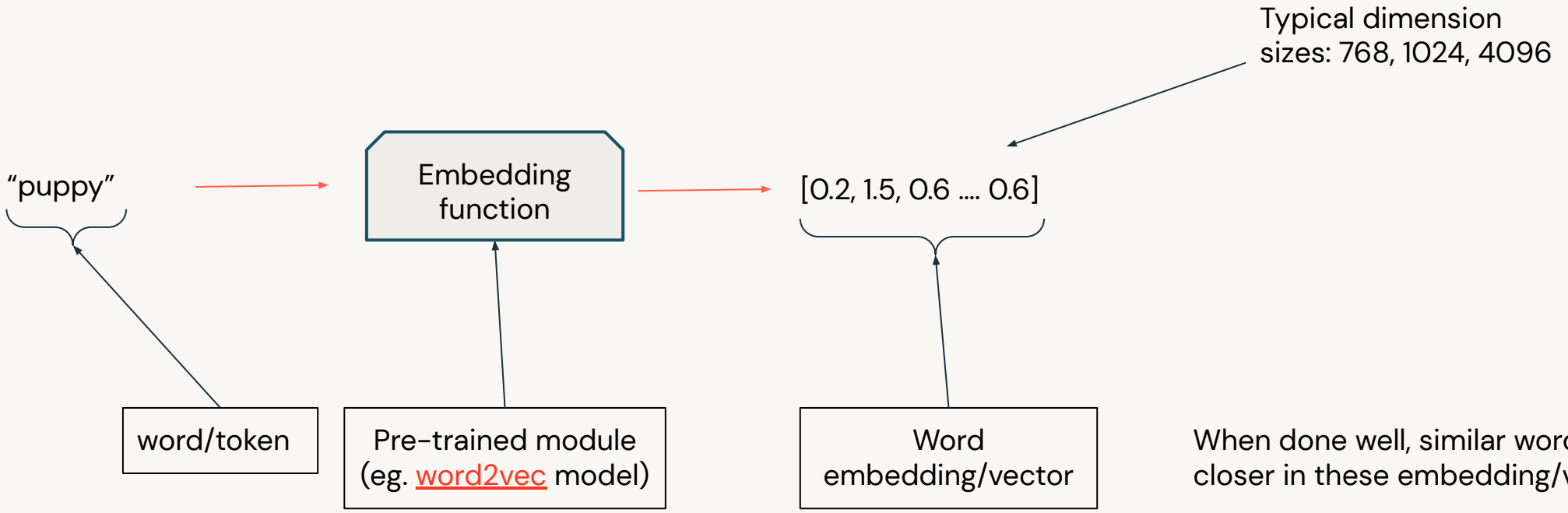
©2023 Databricks Inc. — All rights reserved

Source: victorzhou.com

Creating dense vector representation

Sparse vectors lose meaningful notion of similarity

New idea: Let's give **each word** a vector representation and use data to build our embedding space.

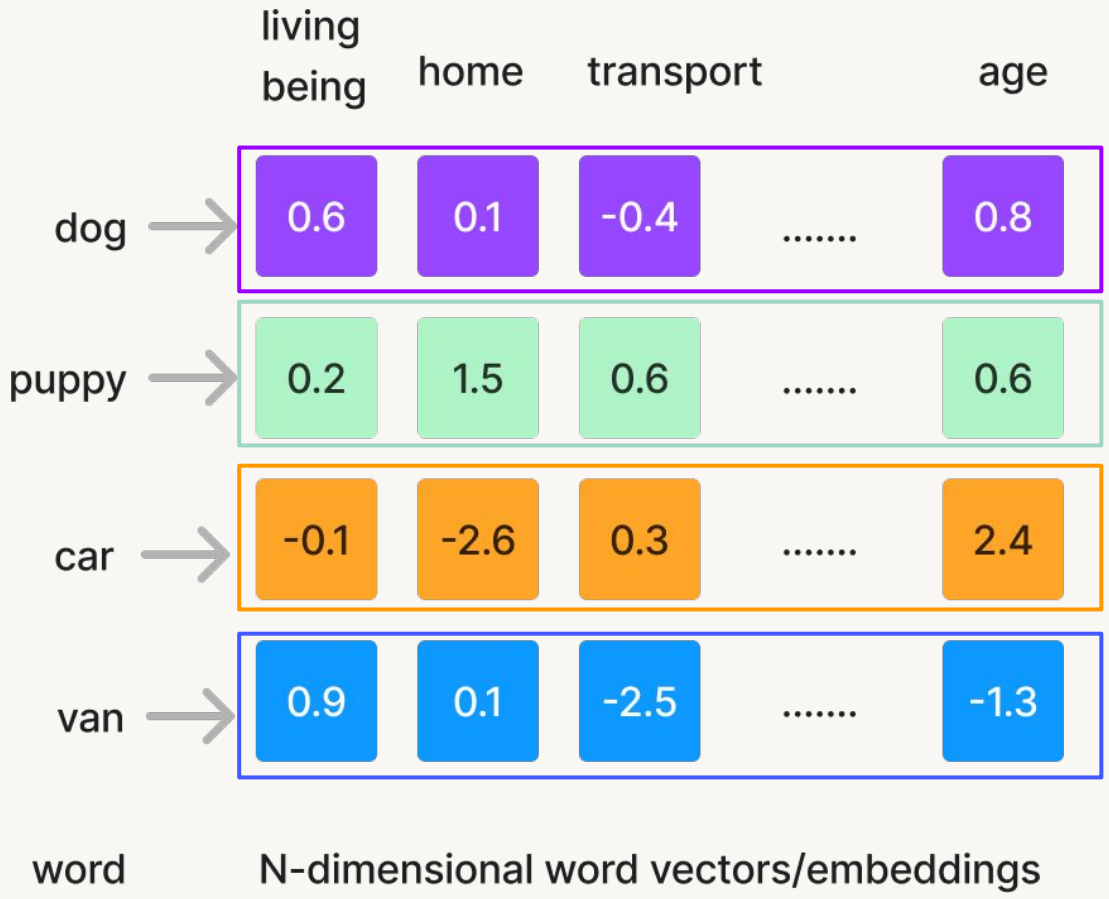


When done well, similar words will be closer in these embedding/vector spaces.

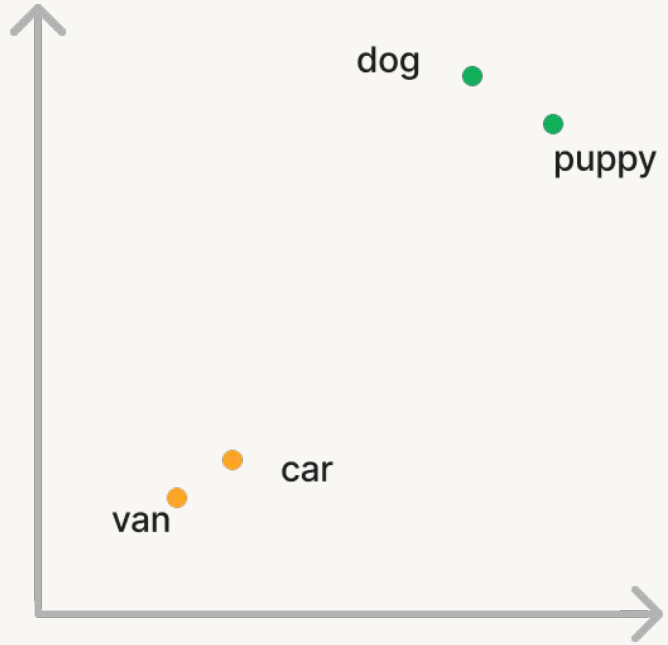
©2023 Databricks Inc. — All rights reserved

Dense vector representations

Visualizing common words using word vectors.



We can project these vectors onto 2D to see how they relate graphically



©2023 Databricks Inc. — All rights reserved

Natural Language Processing (NLP)

Let's review

- NLP is a field of methods to process text.
- NLP is useful: summarization, translation, classification, etc.
- Language models (LMs) predict words by looking at word probabilities.
- Large LMs are just LMs with transformer architectures, but bigger.
- Tokens are the smallest building blocks to convert text to numerical vectors, aka N-dimensional embeddings.

©2023 Databricks Inc. — All rights reserved

Before we begin

1. Introduction by Matei Zaharia: Why LLMs?
2. Primer on NLP
3. Setting up your Databricks lab environment

©2023 Databricks Inc. — All rights reserved

Databricks 101

A quick walkthrough of the platform

©2023 Databricks Inc. — All rights reserved