Coding for Economists

Advanced Session 3

Jian Cao

15 May 2025

Module Files

Google Drive Folder

Outline

Outline

- Text Analysis Overview

Outline

- Text Analysis Overview
- Text Preprocessing

Outline

- Text Analysis Overview
- Text Preprocessing
- Word Embedding

Outline

- Text Analysis Overview
- Text Preprocessing
- Word Embedding
- Sentiment Analysis

Outline

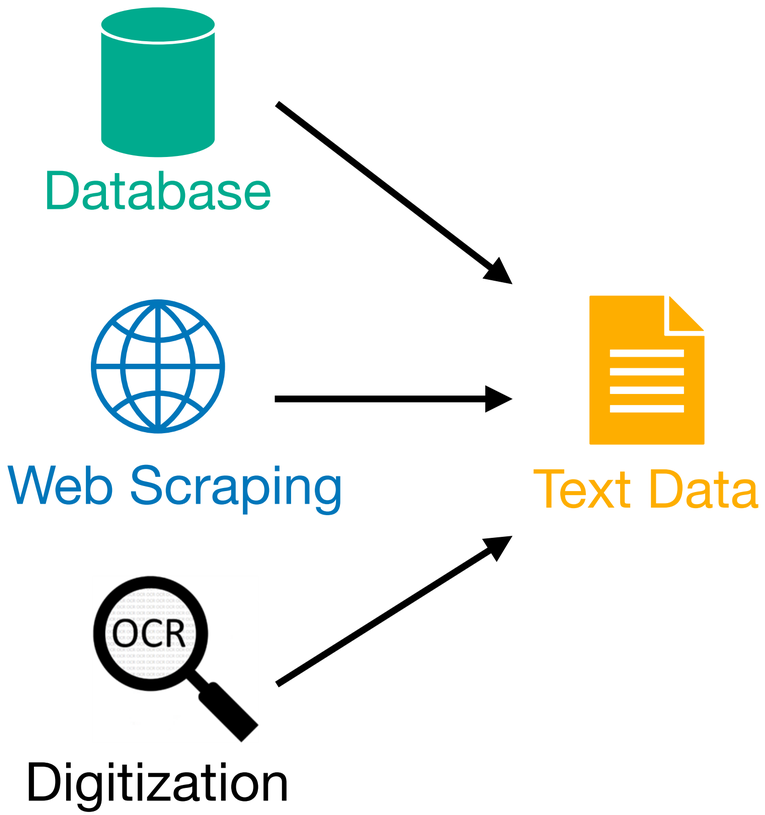
- Text Analysis Overview
- Text Preprocessing
- Word Embedding
- Sentiment Analysis
- Topic Analysis

Text Analysis

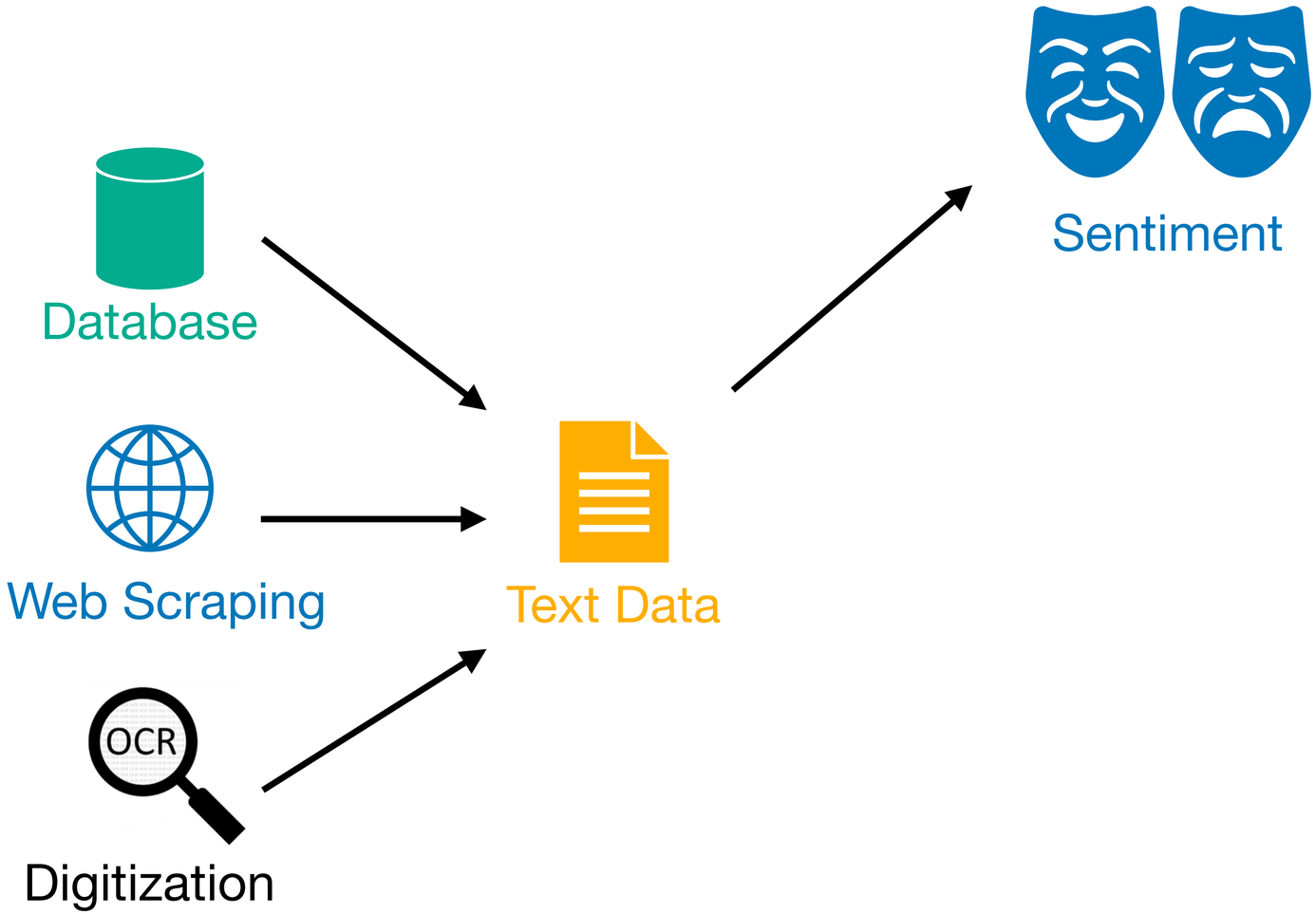
Text Analysis

Text Data

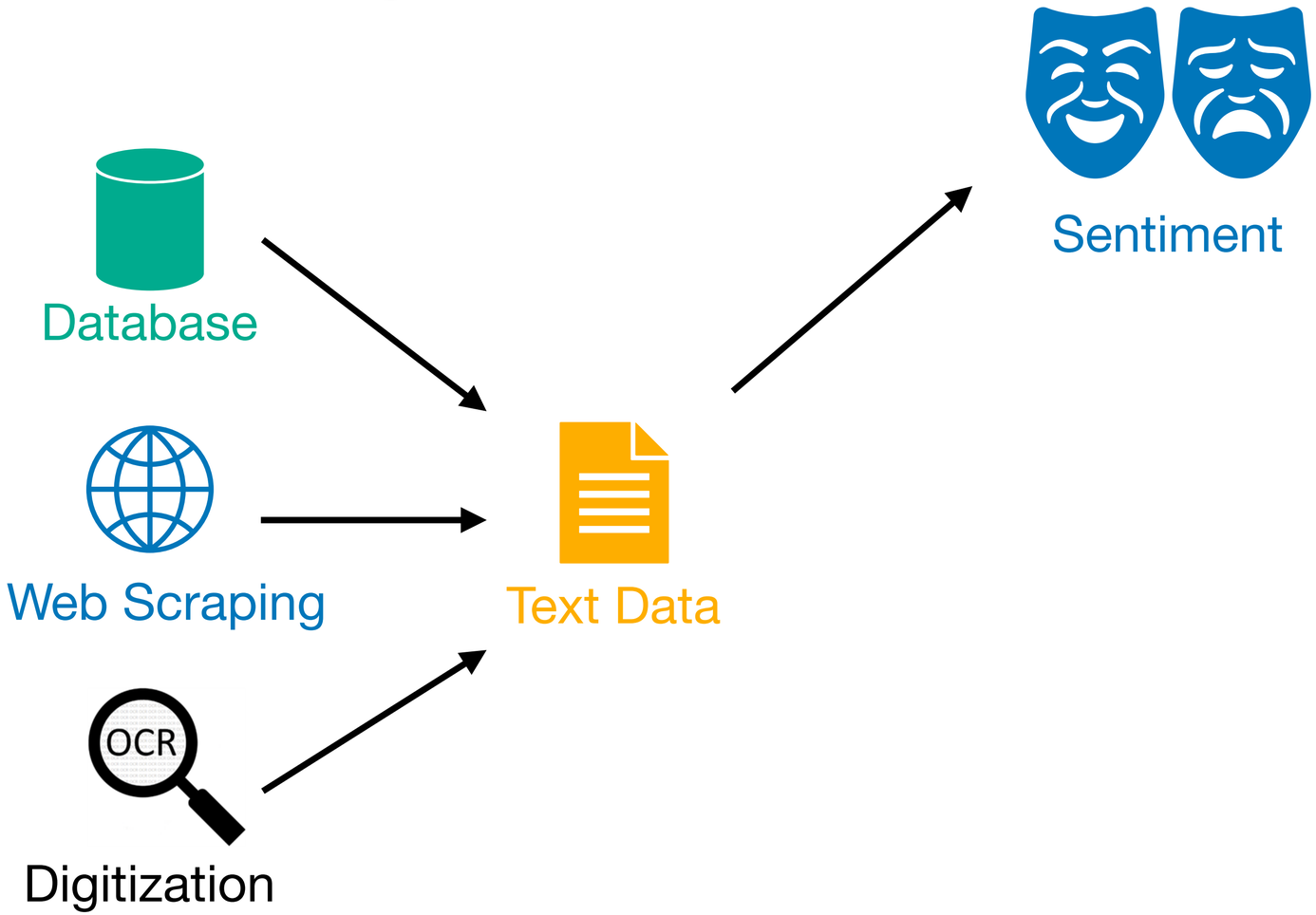
Text Analysis



Text Analysis

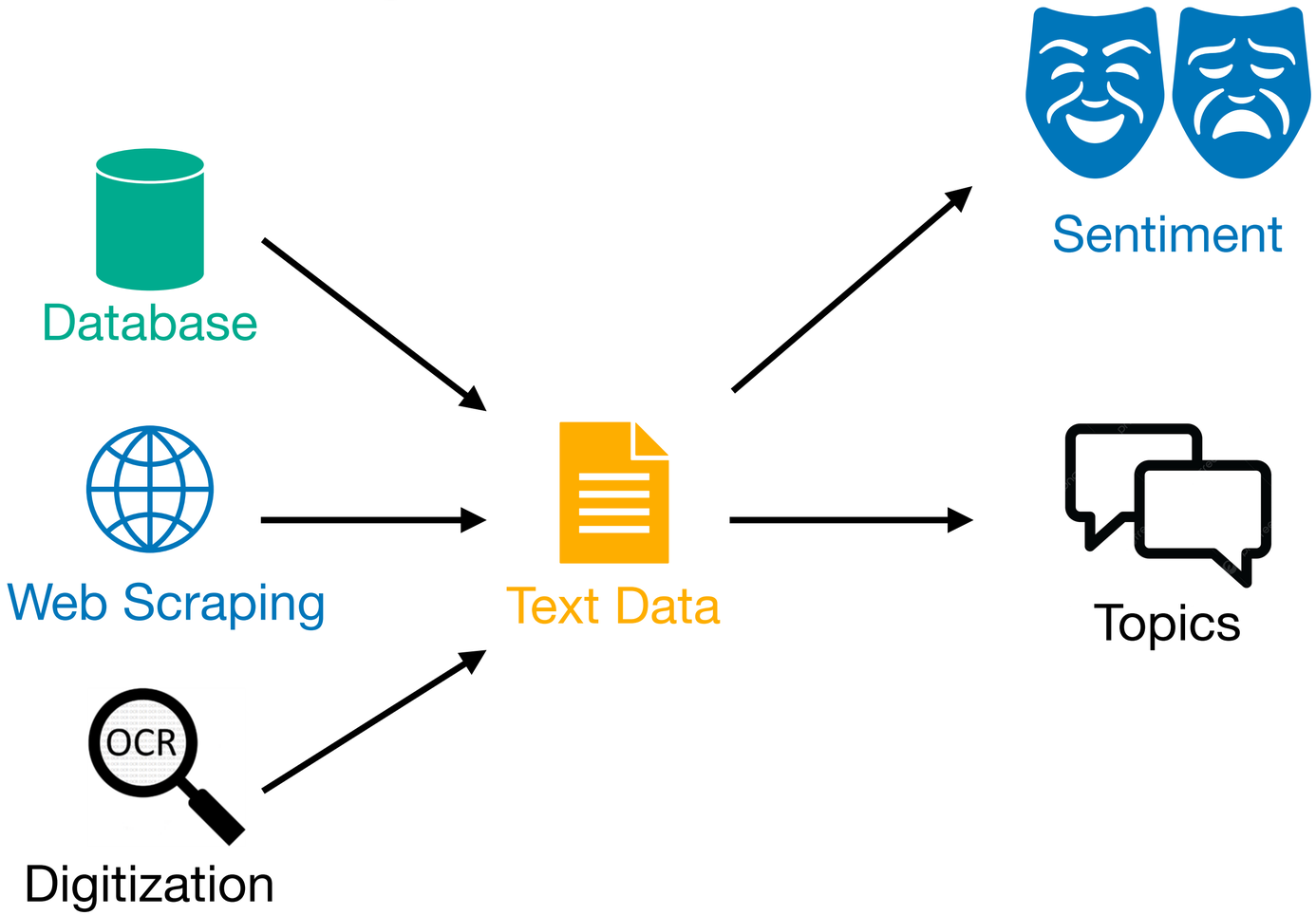


Text Analysis



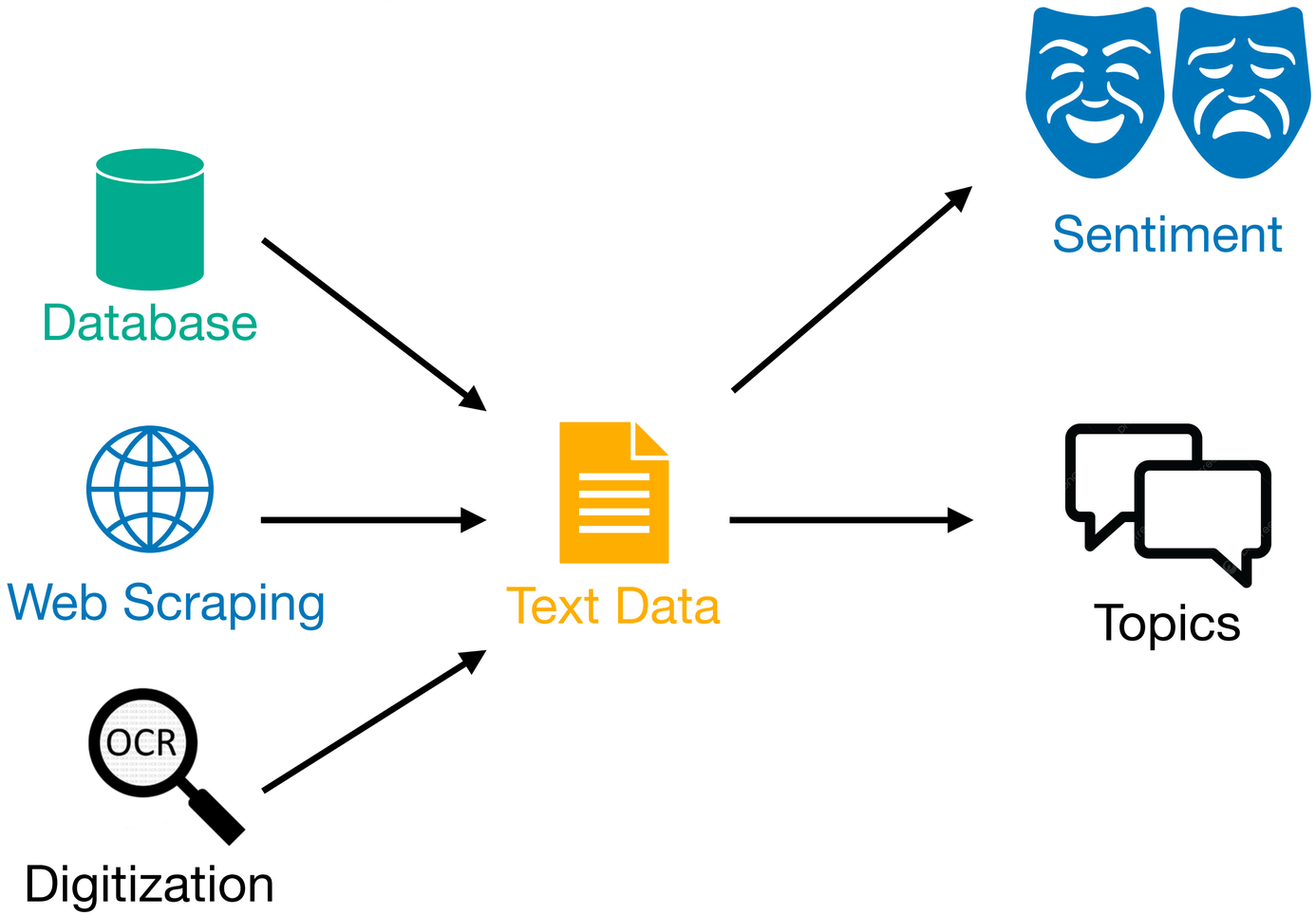
- Monetary Policy
- Consumer Confidence
- Earning Surprises
- Labor Market

Text Analysis



- Monetary Policy
- Consumer Confidence
- Earning Surprises
- Labor Market

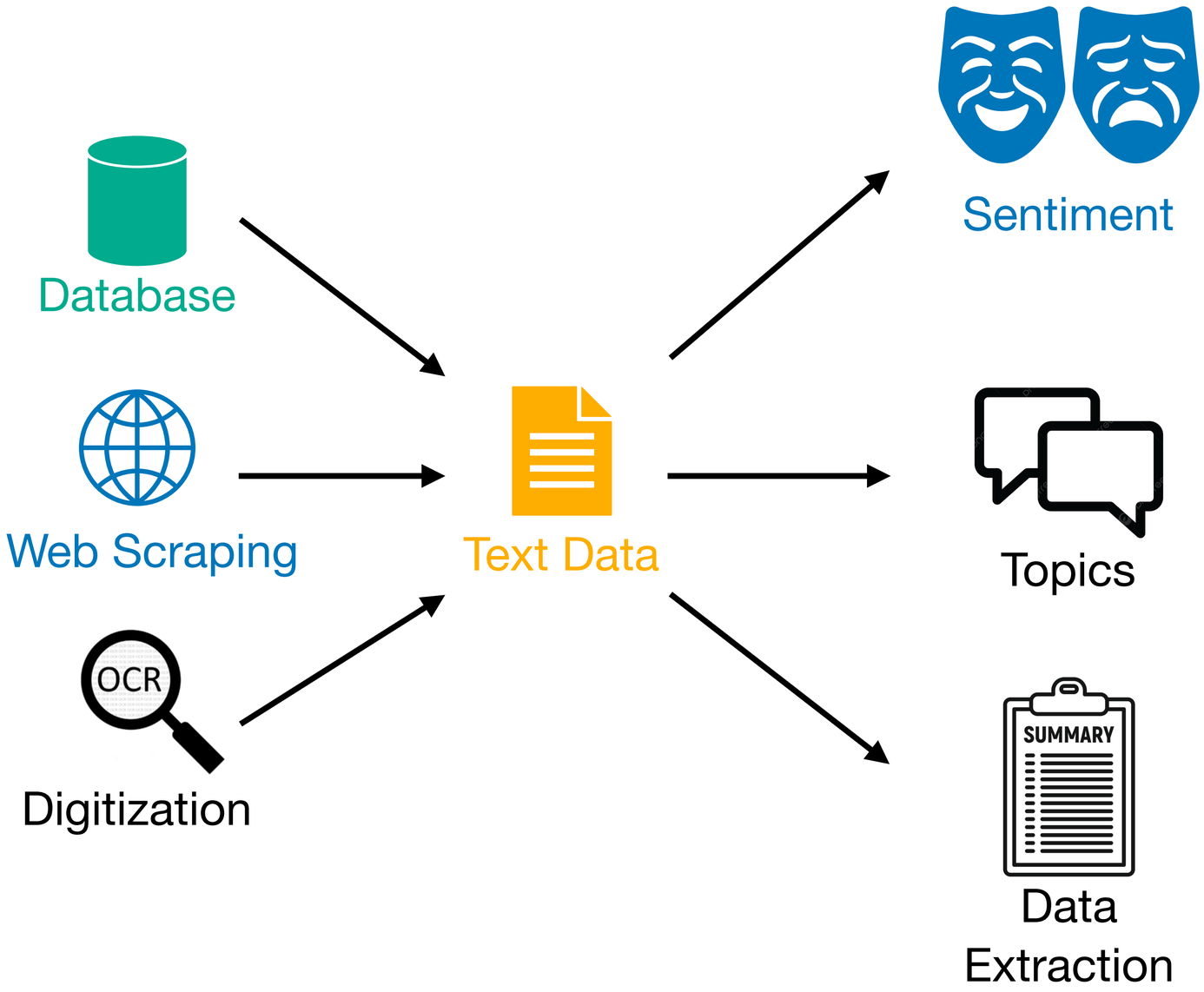
Text Analysis



- Monetary Policy
- Consumer Confidence
- Earning Surprises
- Labor Market

- Legislative Speech
- Qualitative Survey

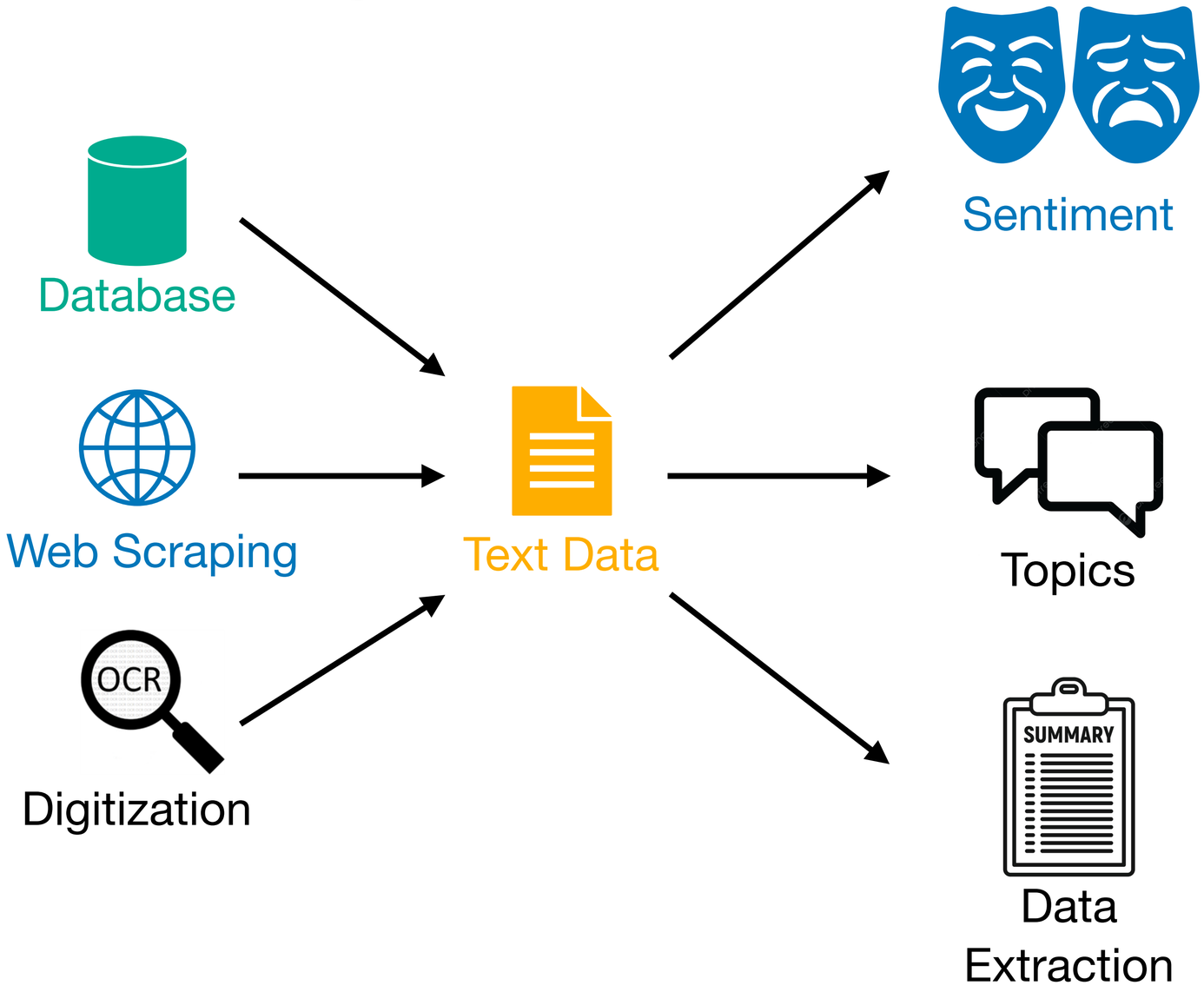
Text Analysis



- Monetary Policy
- Consumer Confidence
- Earning Surprises
- Labor Market

- Legislative Speech
- Qualitative Survey

Text Analysis



- Monetary Policy
- Consumer Confidence
- Earning Surprises
- Labor Market

- Legislative Speech
- Qualitative Survey

- Summarize Document
- Key Info Extraction
- Quantitative Modeling

Text Preprocessing

Text Preprocessing

- Noise Removal (e.g. url, email, emoji)

Text Preprocessing

- Noise Removal (e.g. url, email, emoji)
- Tokenization (sentence -> list of tokens)

Text Preprocessing

- Noise Removal (e.g. url, email, emoji)
- Tokenization (sentence -> list of tokens)
- Remove Stopwords (e.g. the, a, it)

Text Preprocessing

- Noise Removal (e.g. url, email, emoji)
- Tokenization (sentence -> list of tokens)
- Remove Stopwords (e.g. the, a, it)
- Stemming (doing -> do)

Text Preprocessing

- Noise Removal (e.g. url, email, emoji)
- Tokenization (sentence -> list of tokens)
- Remove Stopwords (e.g. the, a, it)
- Stemming (doing -> do)
- Lemmatizing (teeth -> tooth)

Text Preprocessing

- Noise Removal (e.g. url, email, emoji)
- Tokenization (sentence -> list of tokens)
- Remove Stopwords (e.g. the, a, it)
- Stemming (doing -> do)
- Lemmatizing (teeth -> tooth)

Advanced 3

Word Embedding

Word Embedding

Understand the Meaning of Words

Word Embedding

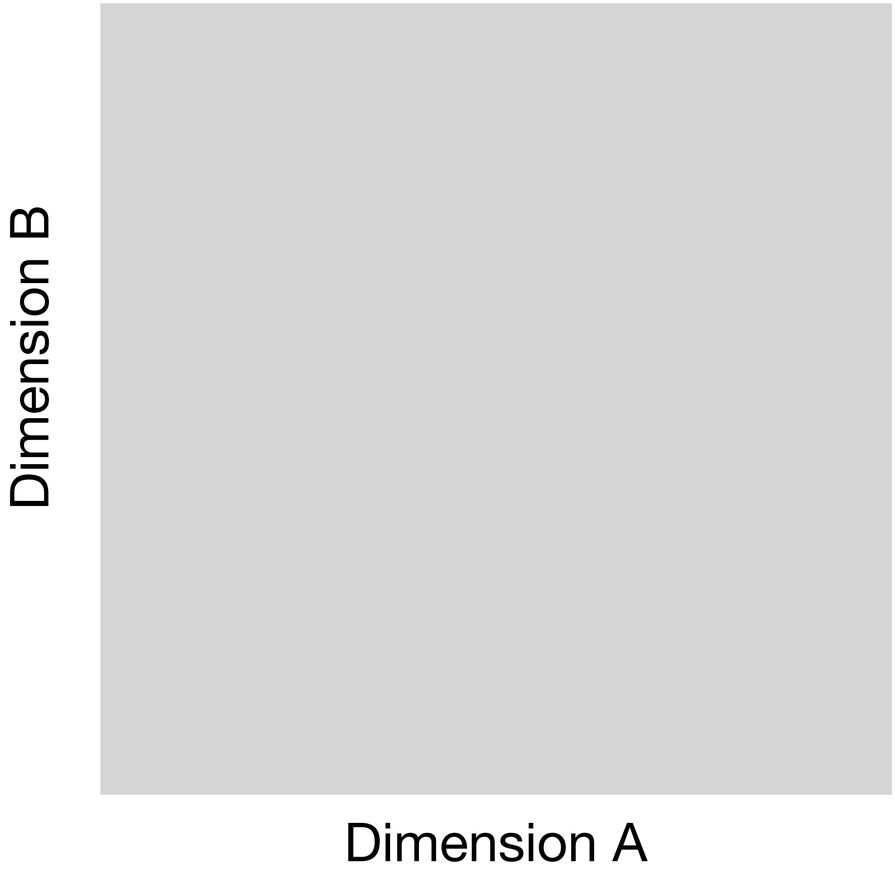
Understand the Meaning of Words

“A playful **Cat** stalked the curious **Dog** in the yard, as the sleek **Car** drove by and the **Jet** airplane thundered overhead.”

Word Embedding

Understand the Meaning of Words

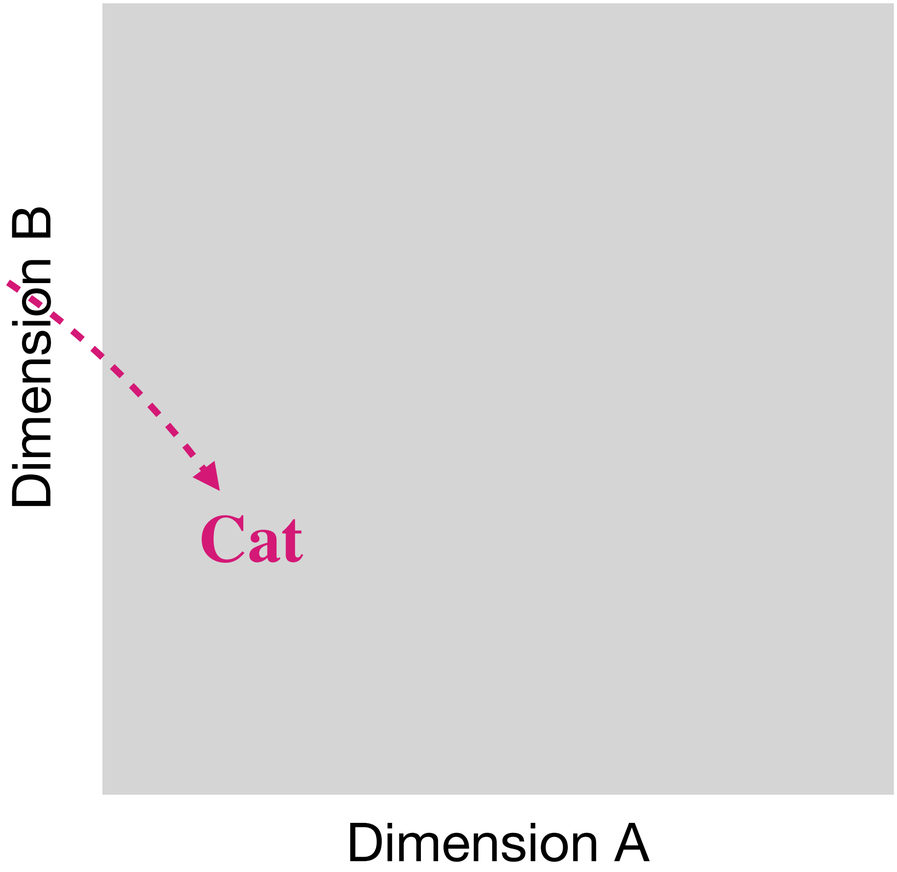
"A playful **Cat** stalked the curious **Dog** in the yard, as the sleek **Car** drove by and the **Jet** airplane thundered overhead."



Word Embedding

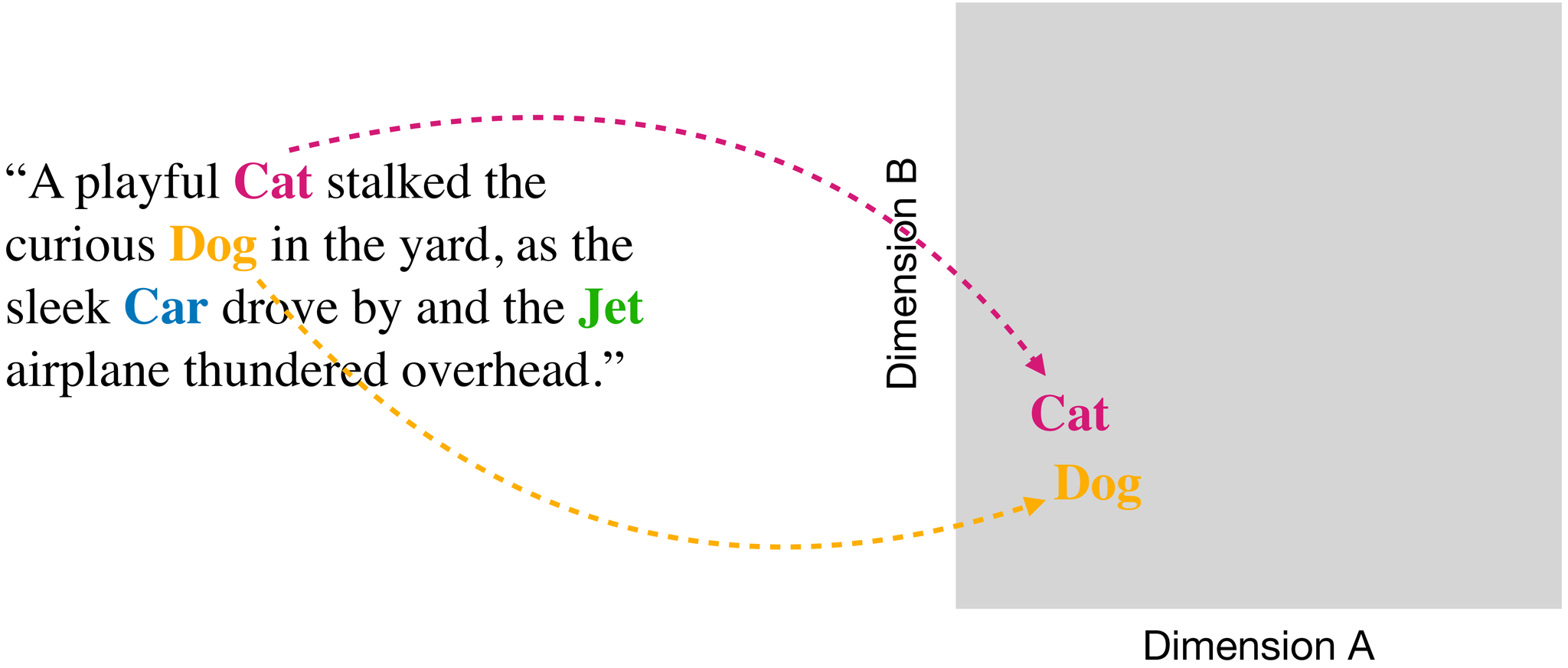
Understand the Meaning of Words

"A playful **Cat** stalked the curious **Dog** in the yard, as the sleek **Car** drove by and the **Jet** airplane thundered overhead."



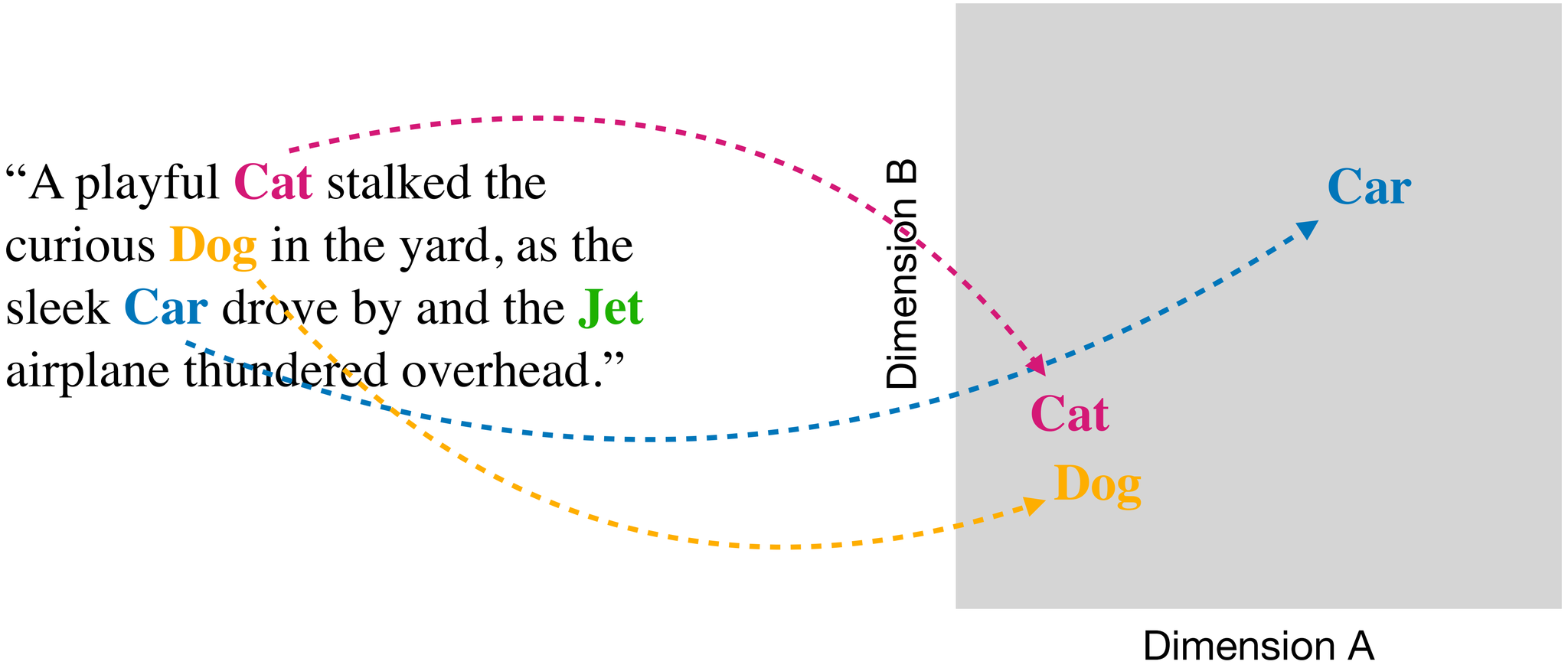
Word Embedding

Understand the Meaning of Words



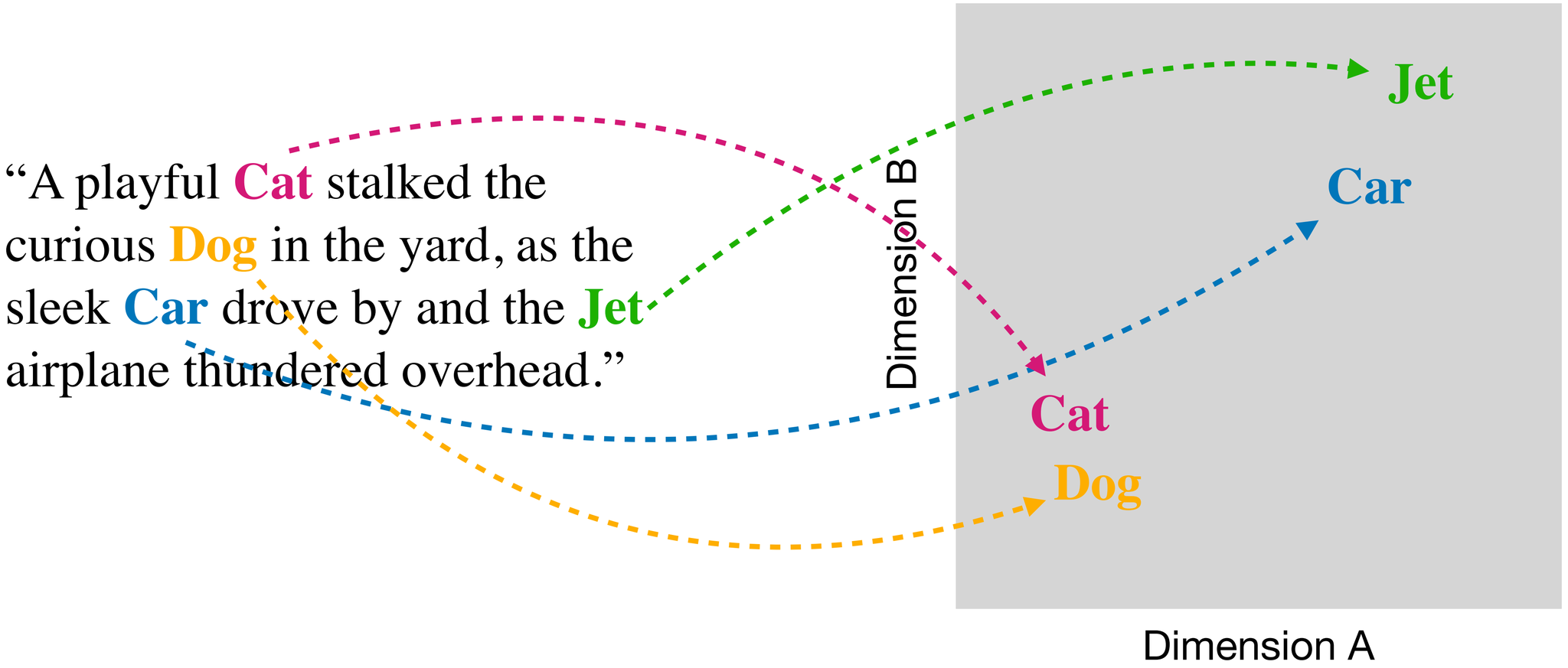
Word Embedding

Understand the Meaning of Words



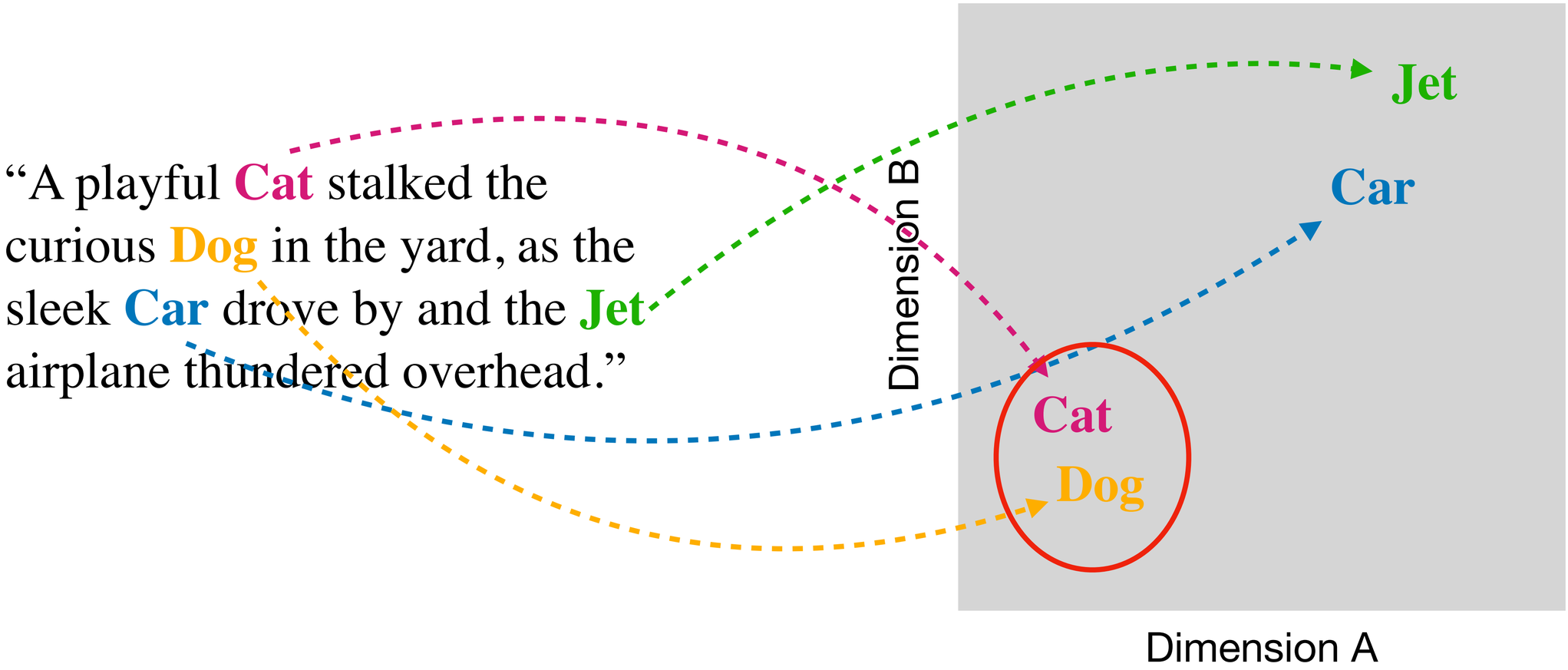
Word Embedding

Understand the Meaning of Words



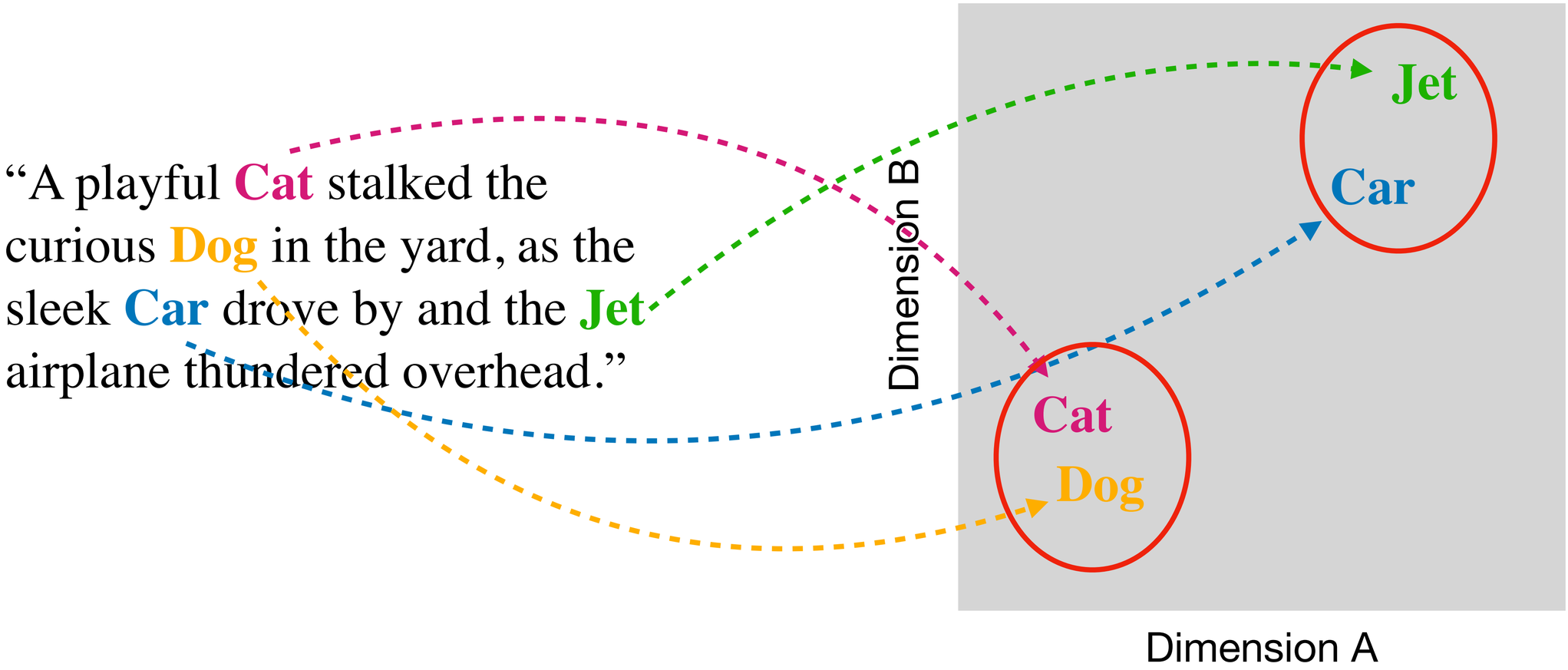
Word Embedding

Understand the Meaning of Words



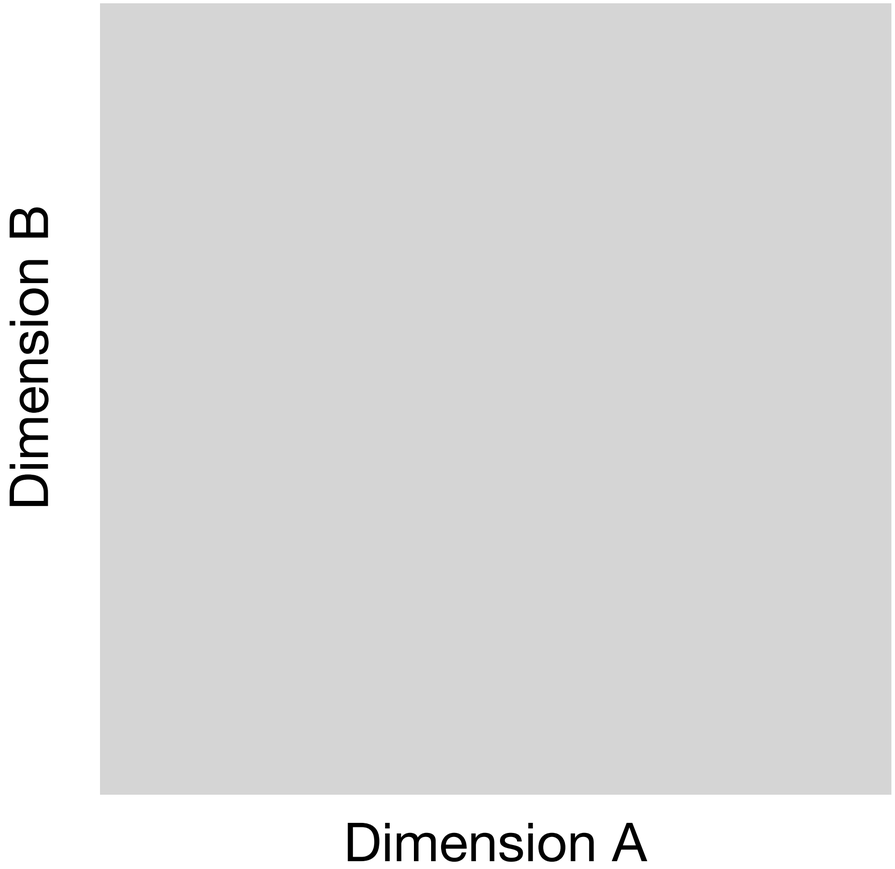
Word Embedding

Understand the Meaning of Words



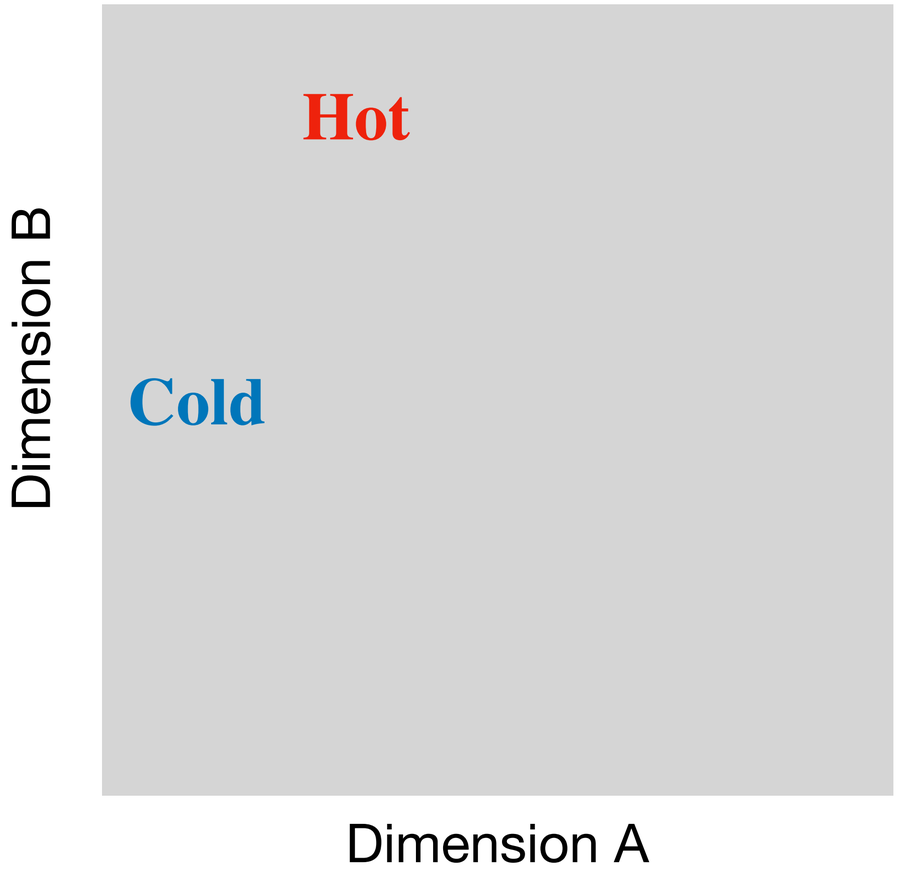
Word Embedding

Understand the Meaning of Words



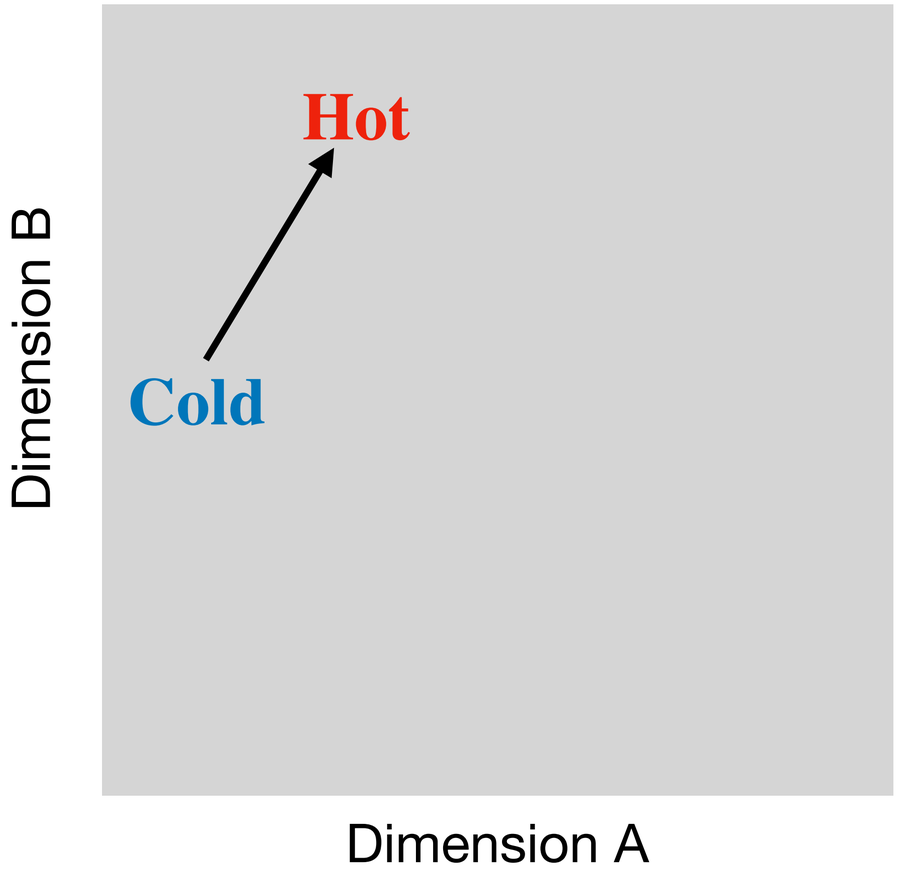
Word Embedding

Understand the Meaning of Words



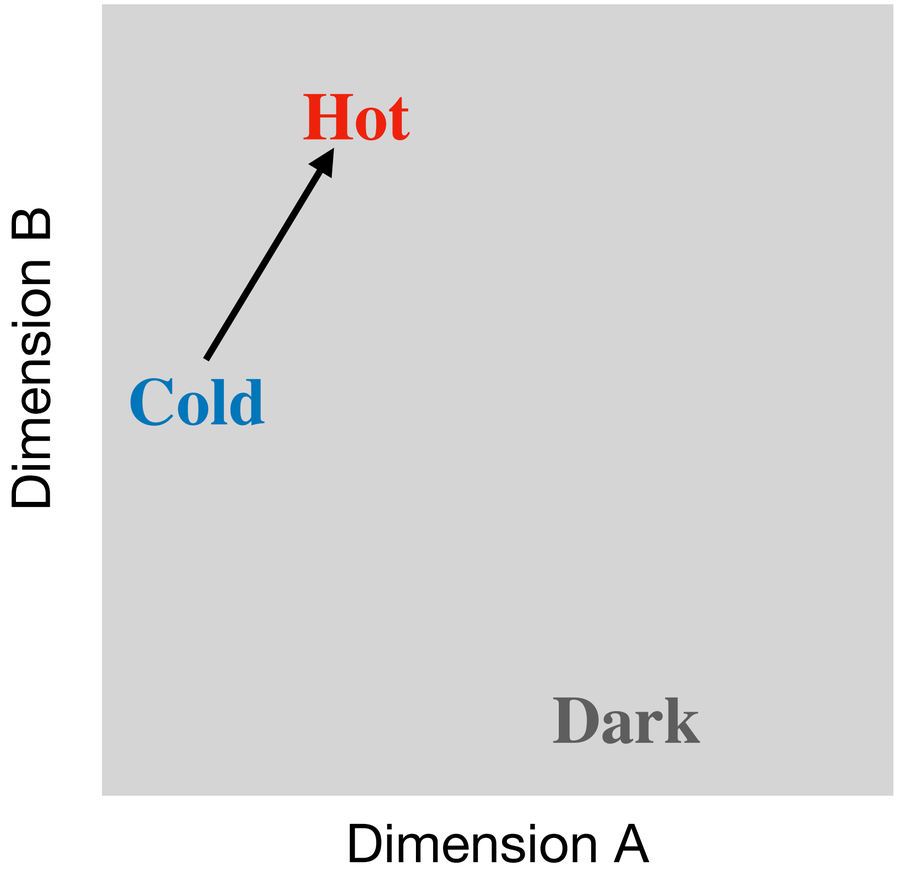
Word Embedding

Understand the Meaning of Words



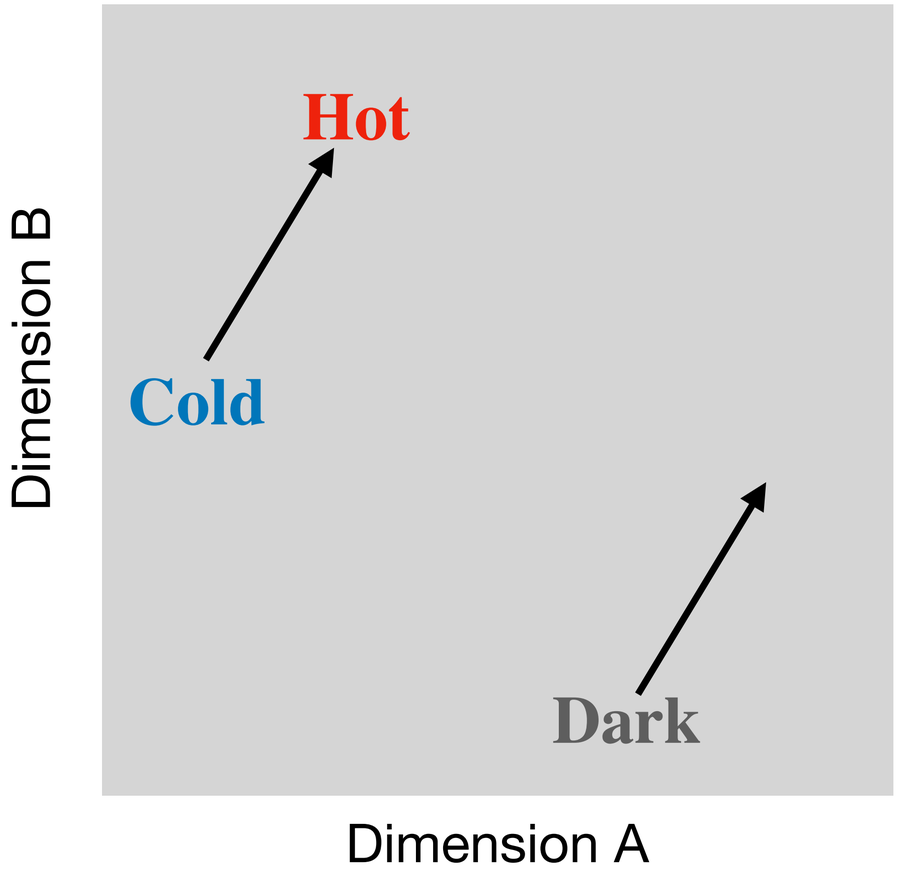
Word Embedding

Understand the Meaning of Words



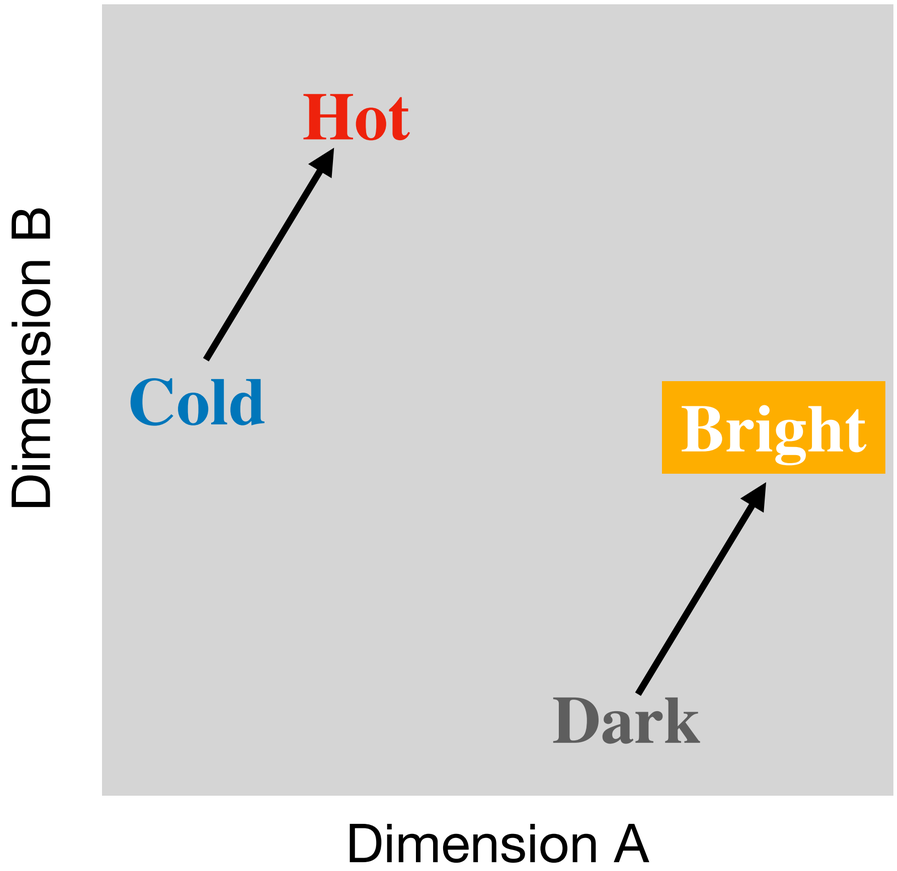
Word Embedding

Understand the Meaning of Words



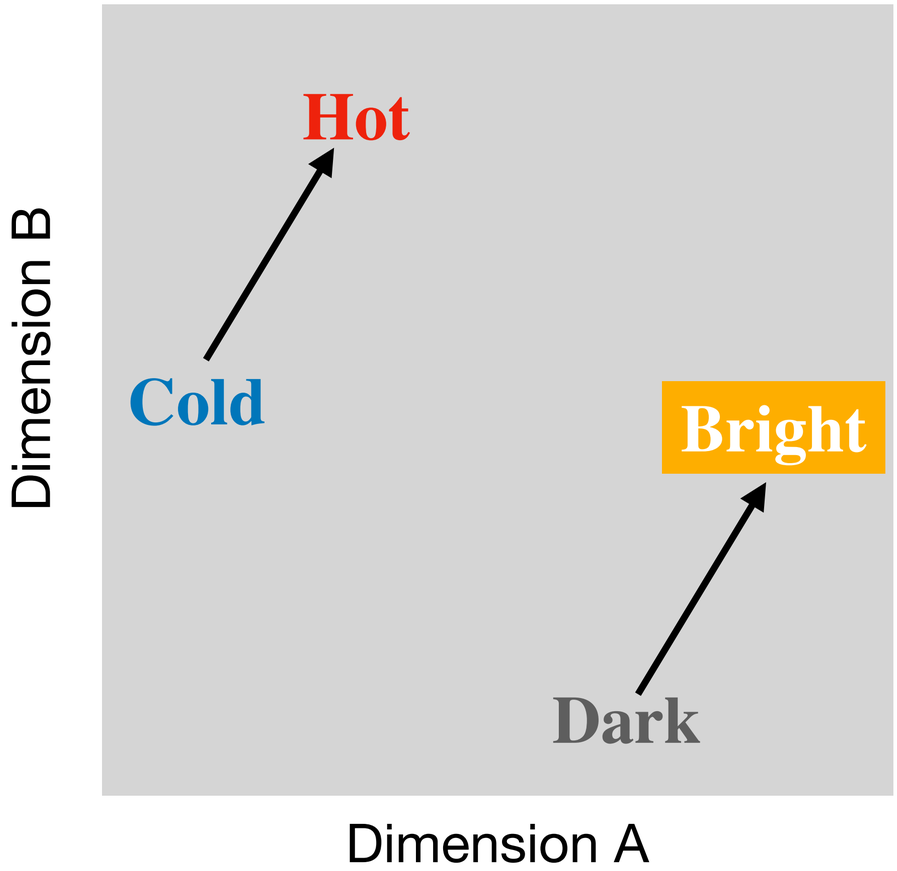
Word Embedding

Understand the Meaning of Words



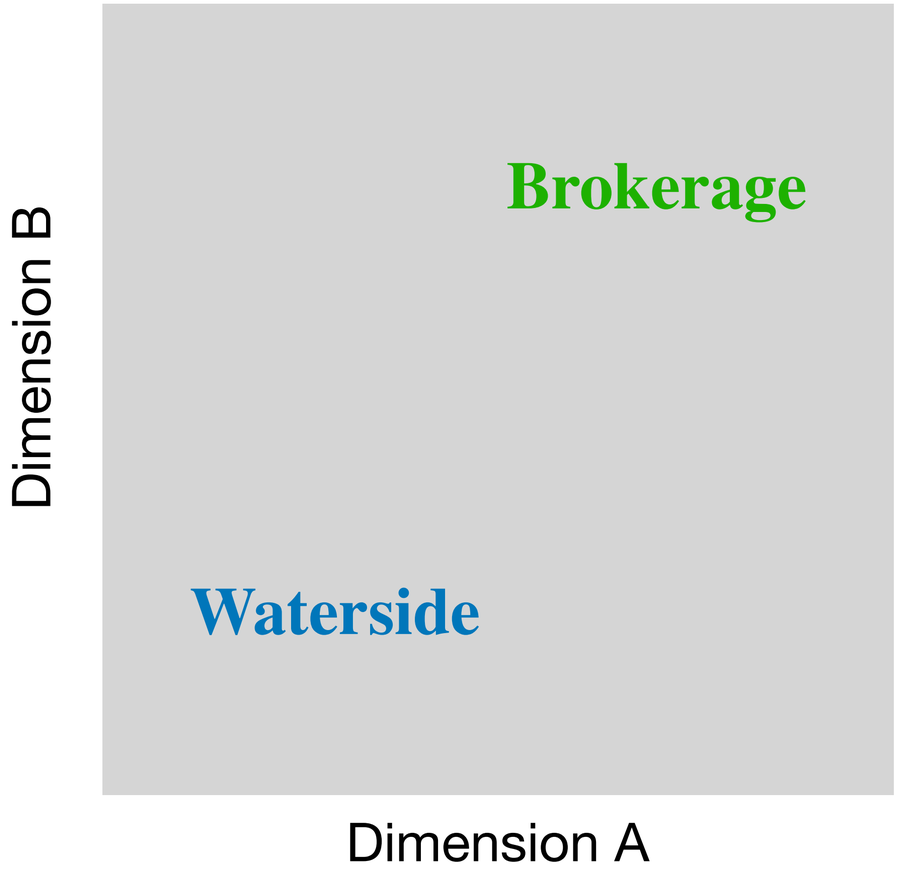
Word Embedding

Understand the Meaning of Words



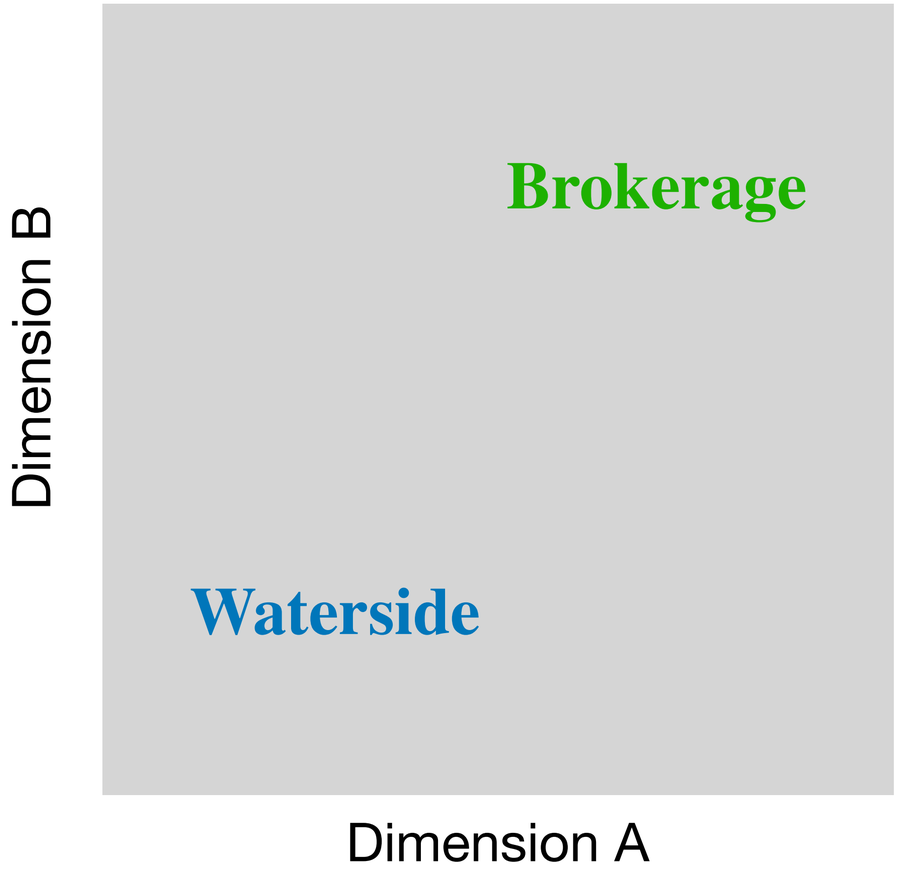
Advanced 3

Contextual Word Embedding

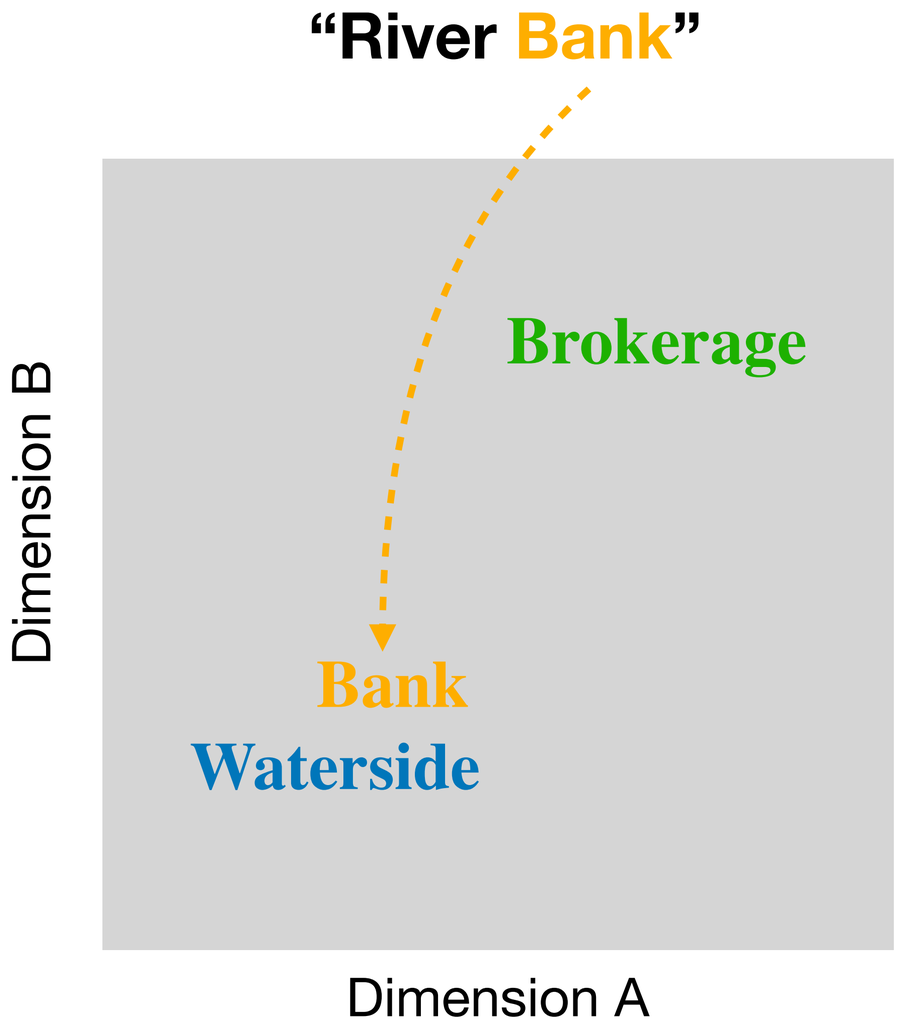


Contextual Word Embedding

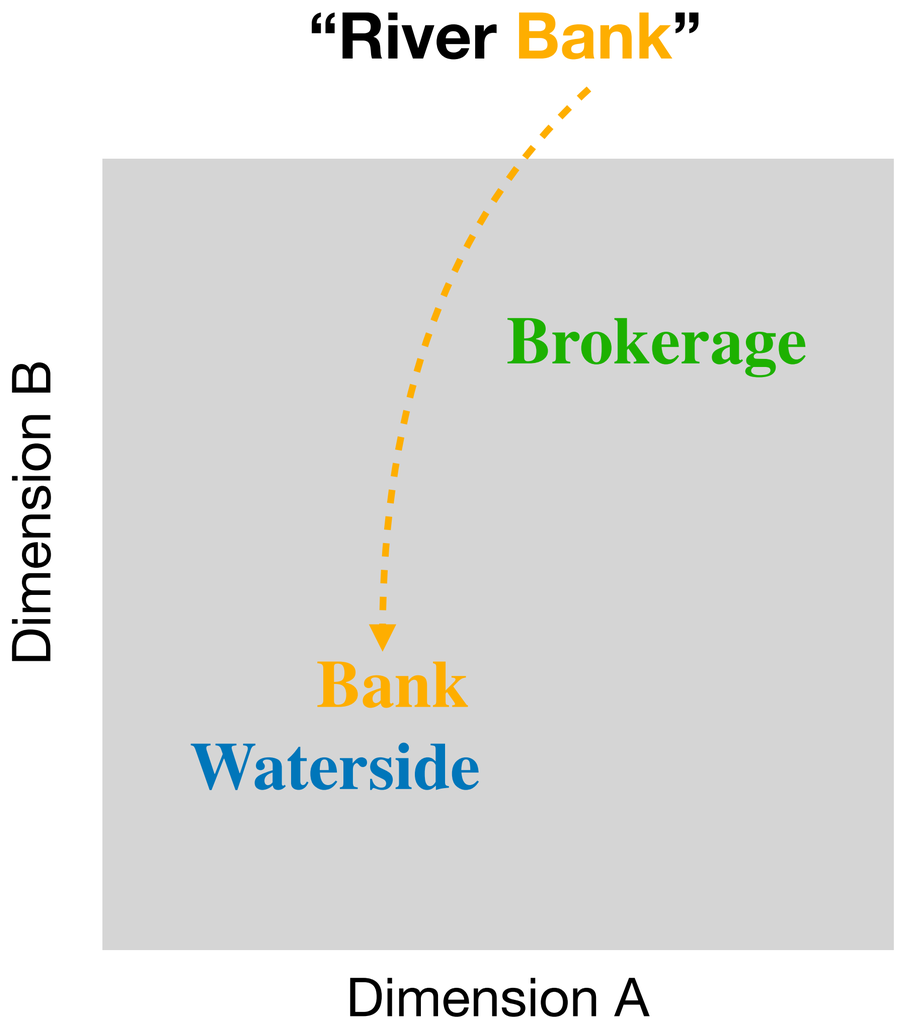
“River Bank”

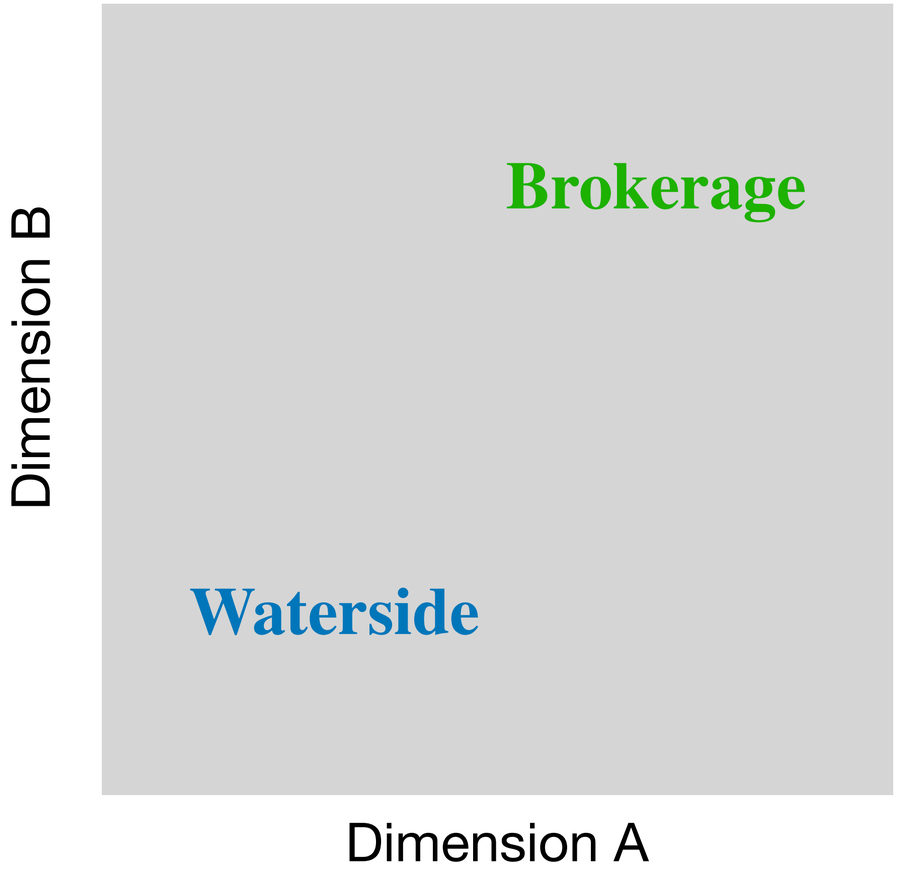


Contextual Word Embedding

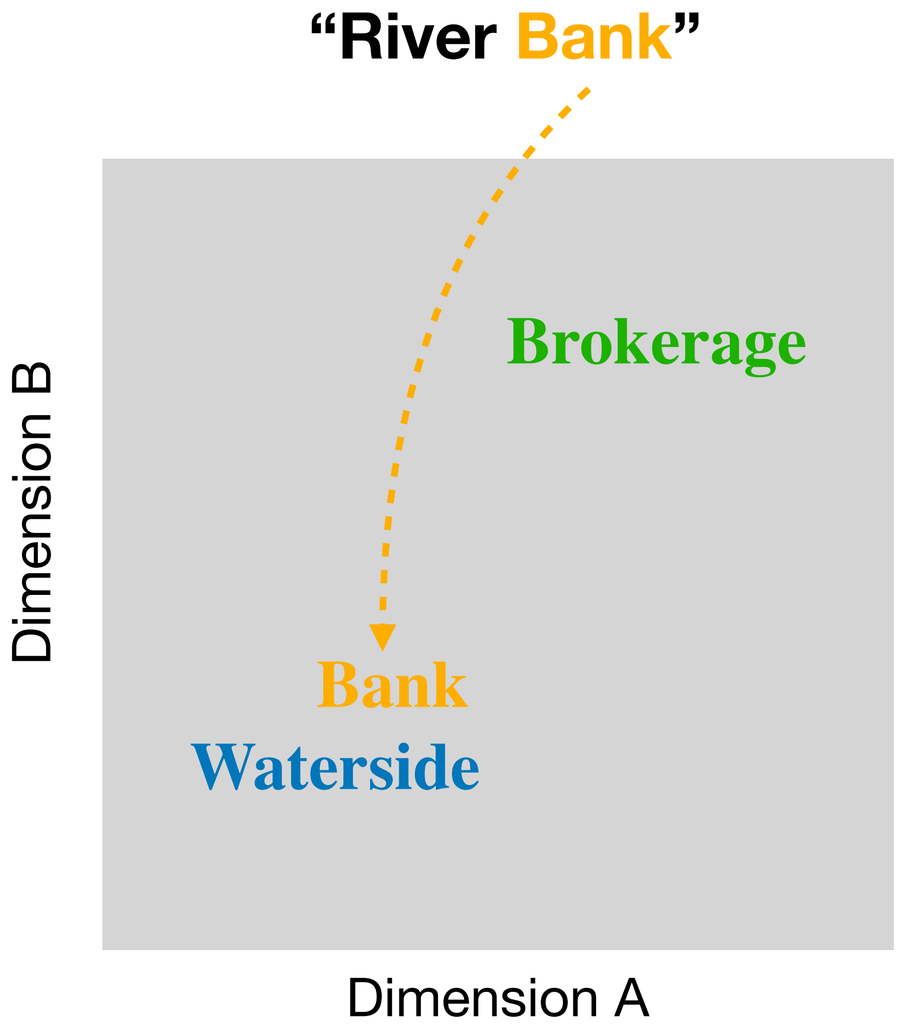


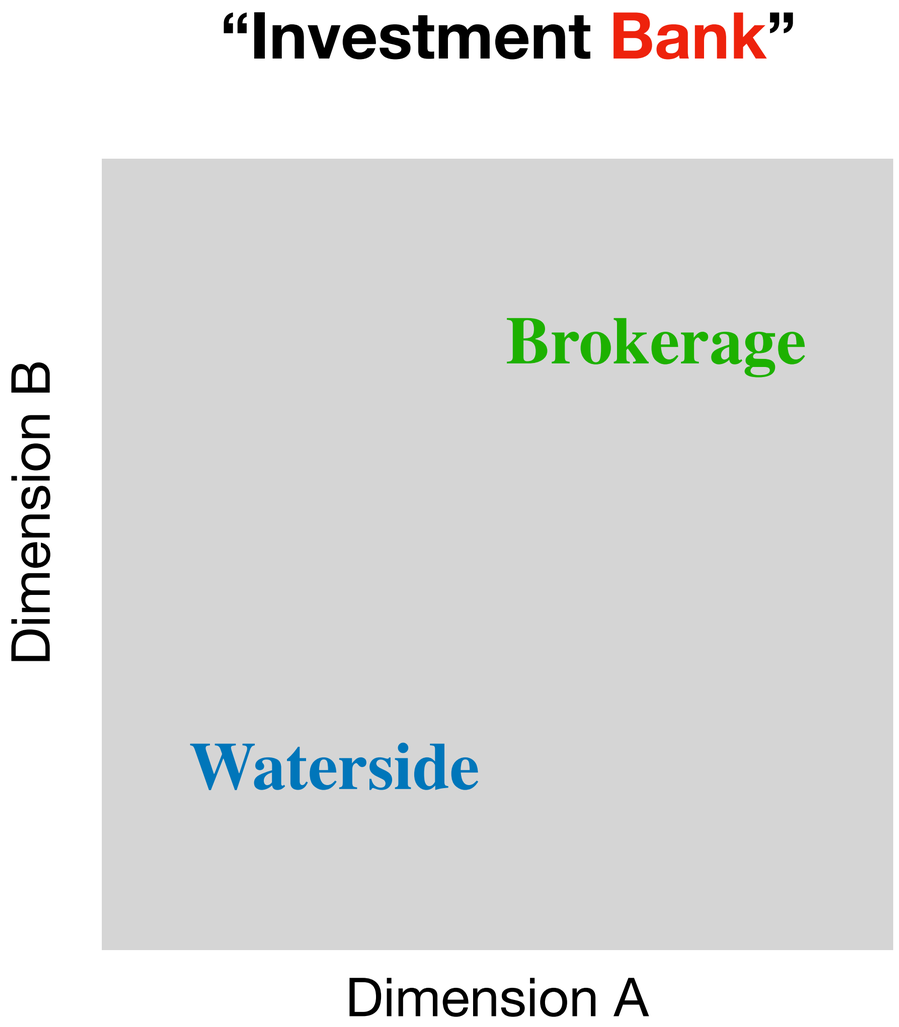
Contextual Word Embedding



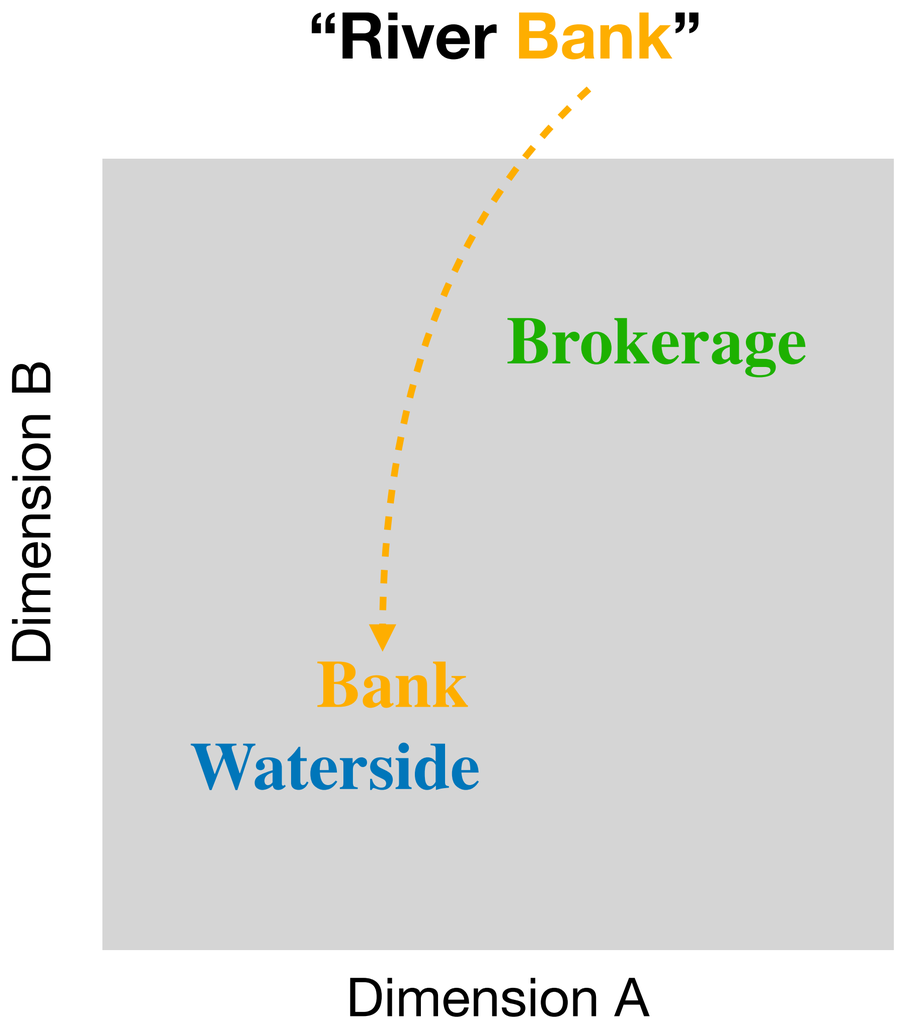


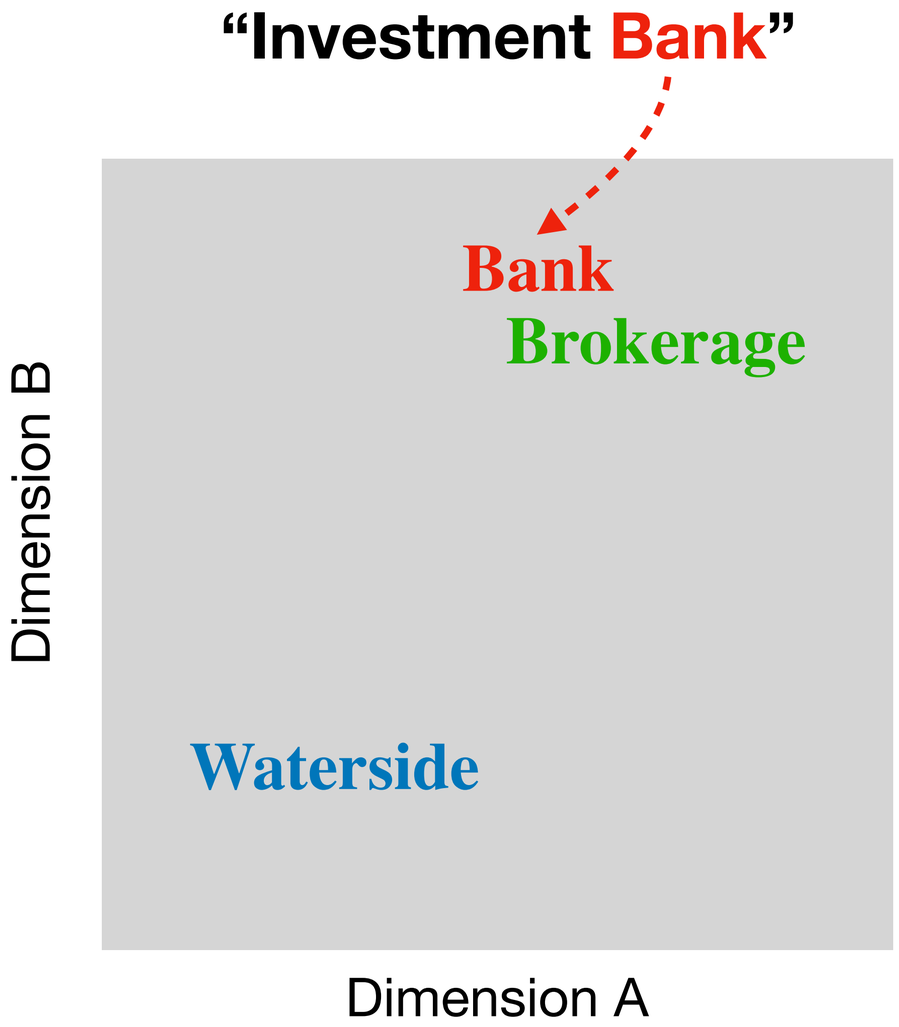
Contextual Word Embedding



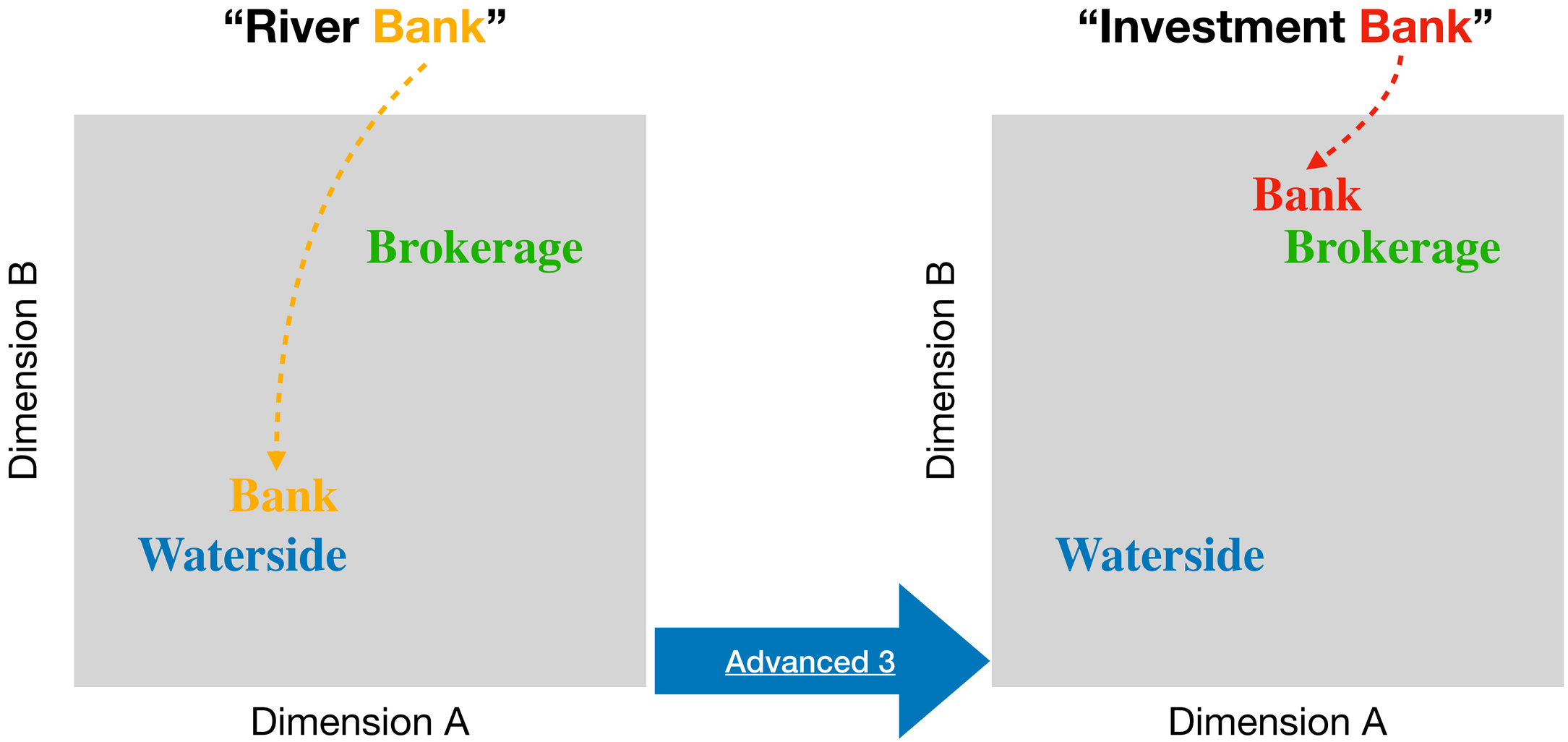


Contextual Word Embedding





Contextual Word Embedding



Pre-Trained Model

Pre-Trained Model

- Pretrained on large data set

Pre-Trained Model

- Pretrained on large data set
- Broad linguistic/semantic knowledge

Pre-Trained Model

- Pretrained on large data set
- Broad linguistic/semantic knowledge
- Specialize in the area

Pre-Trained Model

- Pretrained on large data set
- Broad linguistic/semantic knowledge
- Specialize in the area
- Stable

Pre-Trained Model

- Pretrained on large data set
- Broad linguistic/semantic knowledge
- Specialize in the area
- Stable

Fine-tuning > Training from scratch

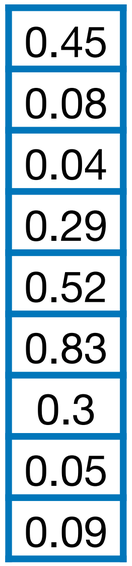
Document Embedding

Document Embedding

“A **playful** **Cat** stalked …”

Document Embedding

“A **playful** **Cat** stalked …”



Document Embedding

“A **playful** **Cat** stalked …”

Document Embedding

“A **playful** **Cat** stalked …”

Document Embedding

“A **playful** **Cat** stalked …”

Contextual
Word Embedding

Document Embedding

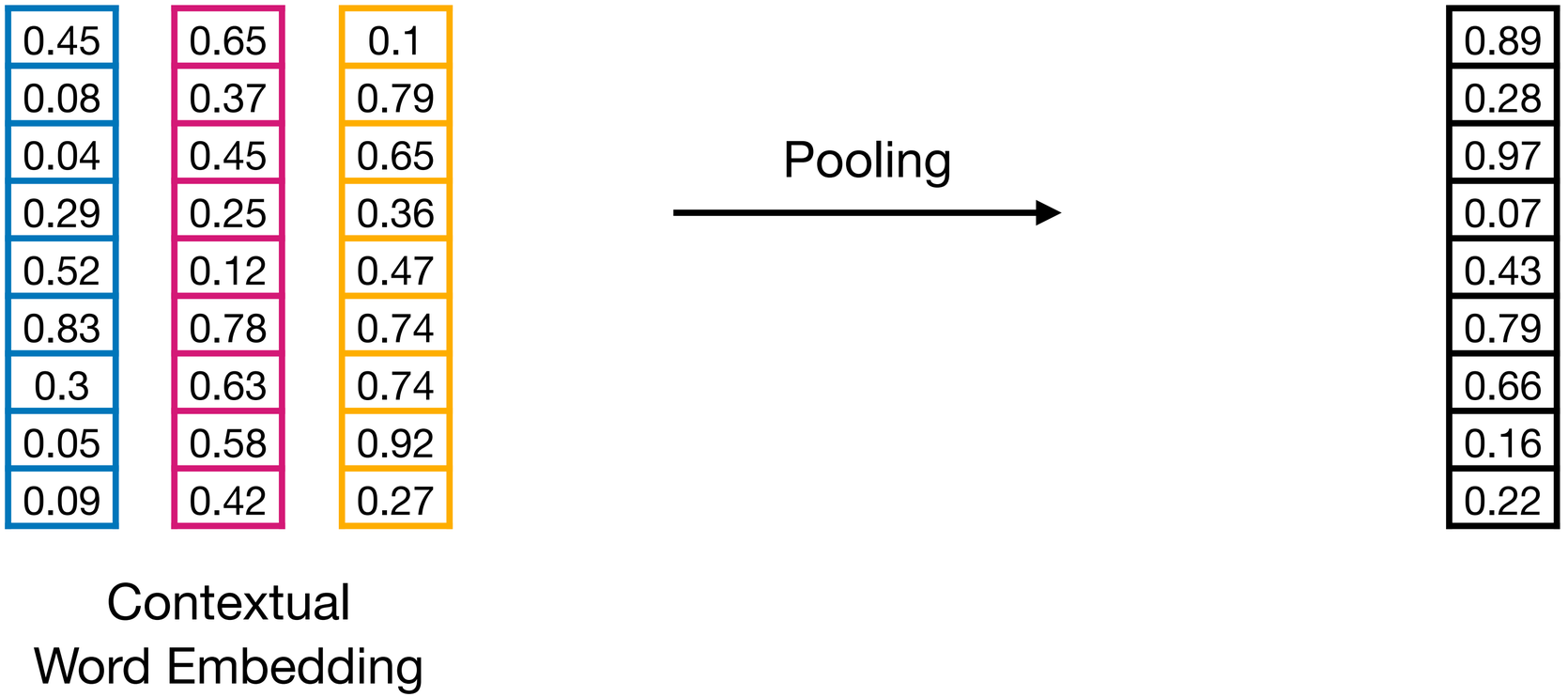
“A **playful** **Cat** stalked …”

Contextual
Word Embedding

Pooling

Document Embedding

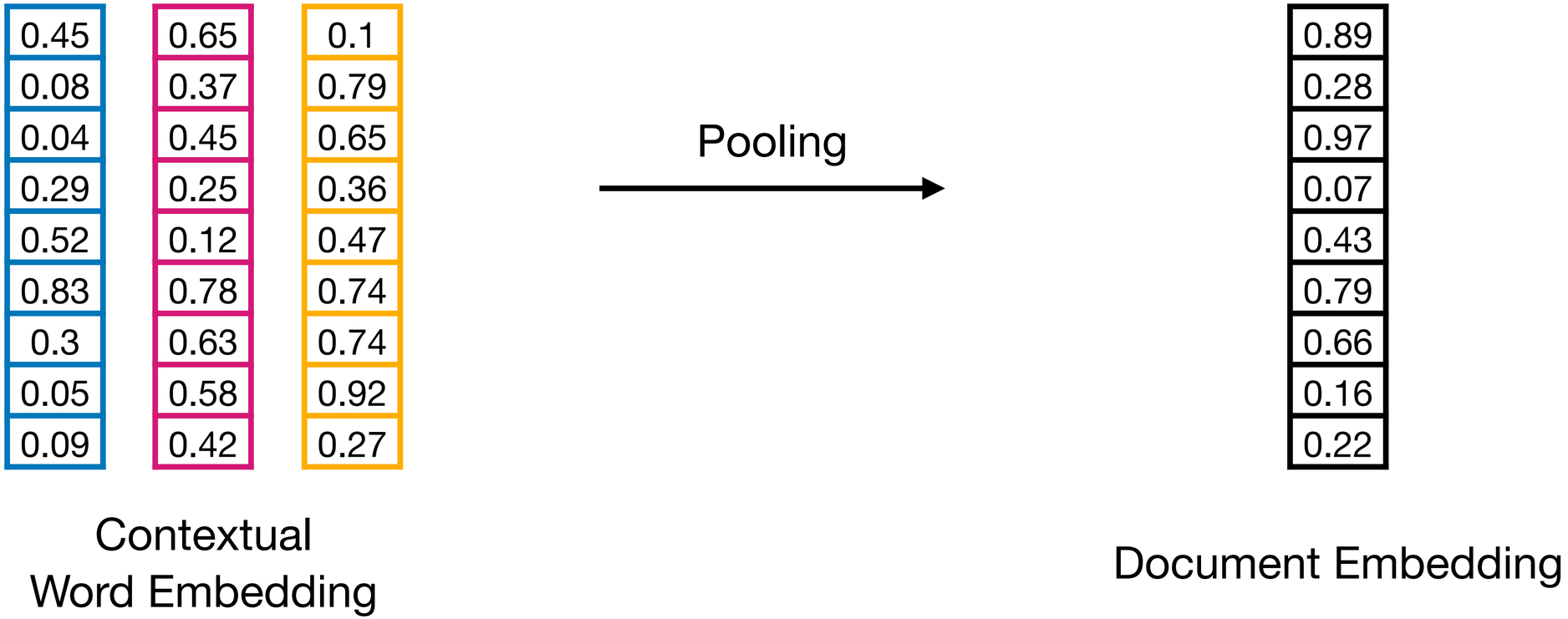
“A **playful** **Cat** stalked …”



Contextual
Word Embedding

Document Embedding

“A **playful** **Cat** stalked …”



Applications of Word Embedding

Applications of Word Embedding

- Sentiment analysis

Applications of Word Embedding

- Sentiment analysis
- Topic analysis

Applications of Word Embedding

- Sentiment analysis
- Topic analysis
- Entity recognition

Applications of Word Embedding

- Sentiment analysis
- Topic analysis
- Entity recognition
- Question answering

Applications of Word Embedding

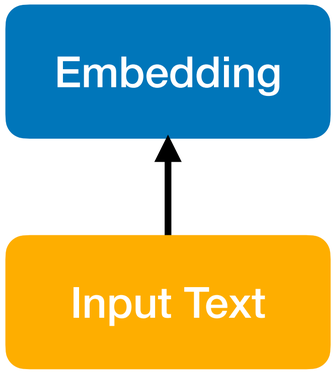
- Sentiment analysis
- Topic analysis
- Entity recognition
- Question answering
- Translation

Transformer

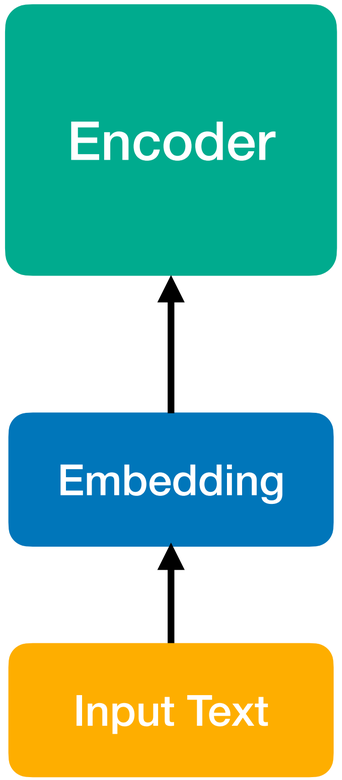
Transformer

Input Text

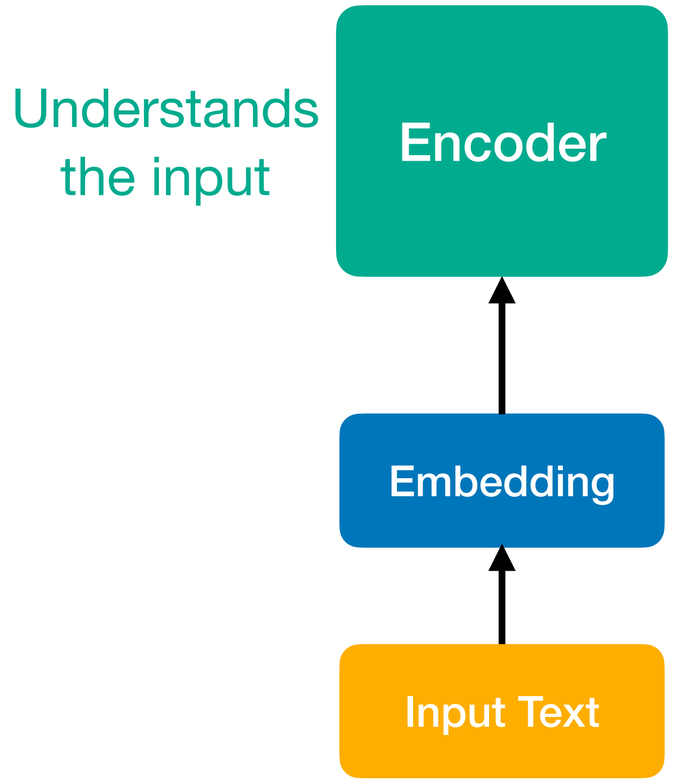
Transformer



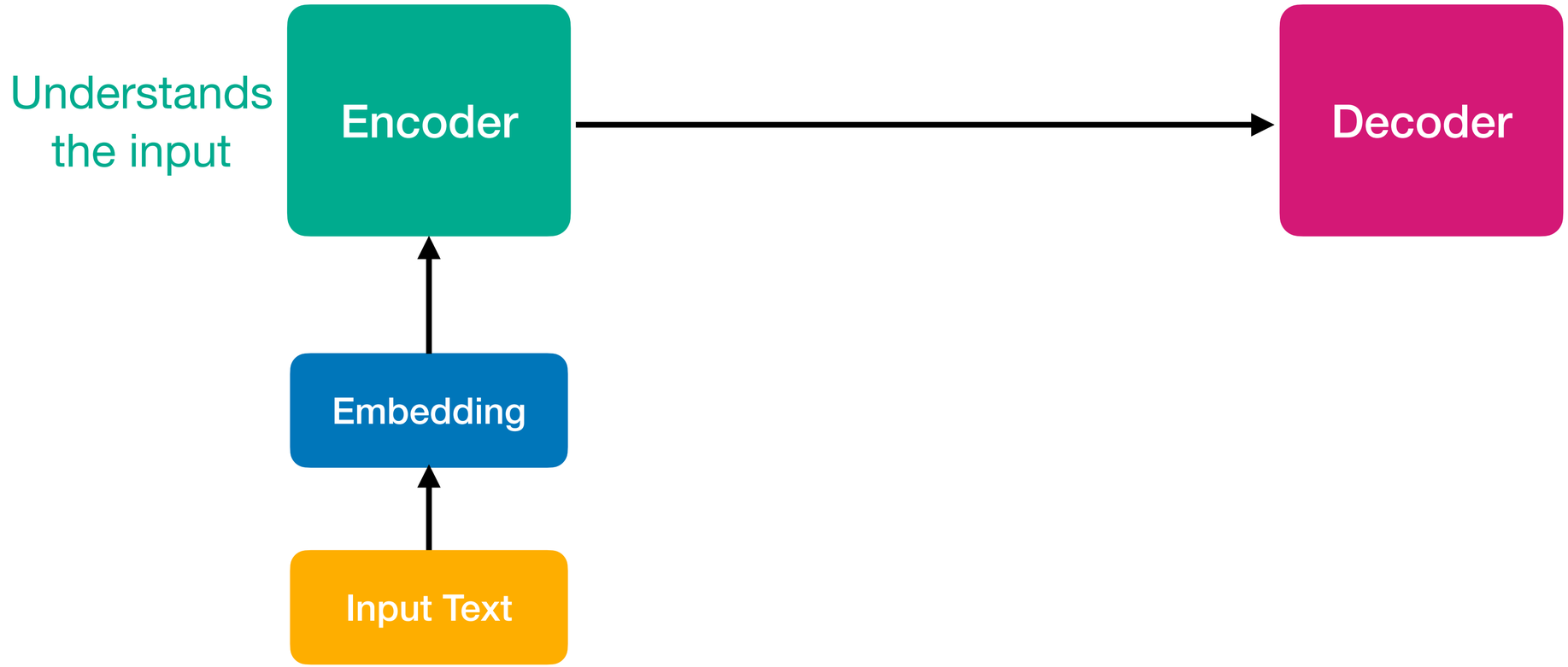
Transformer



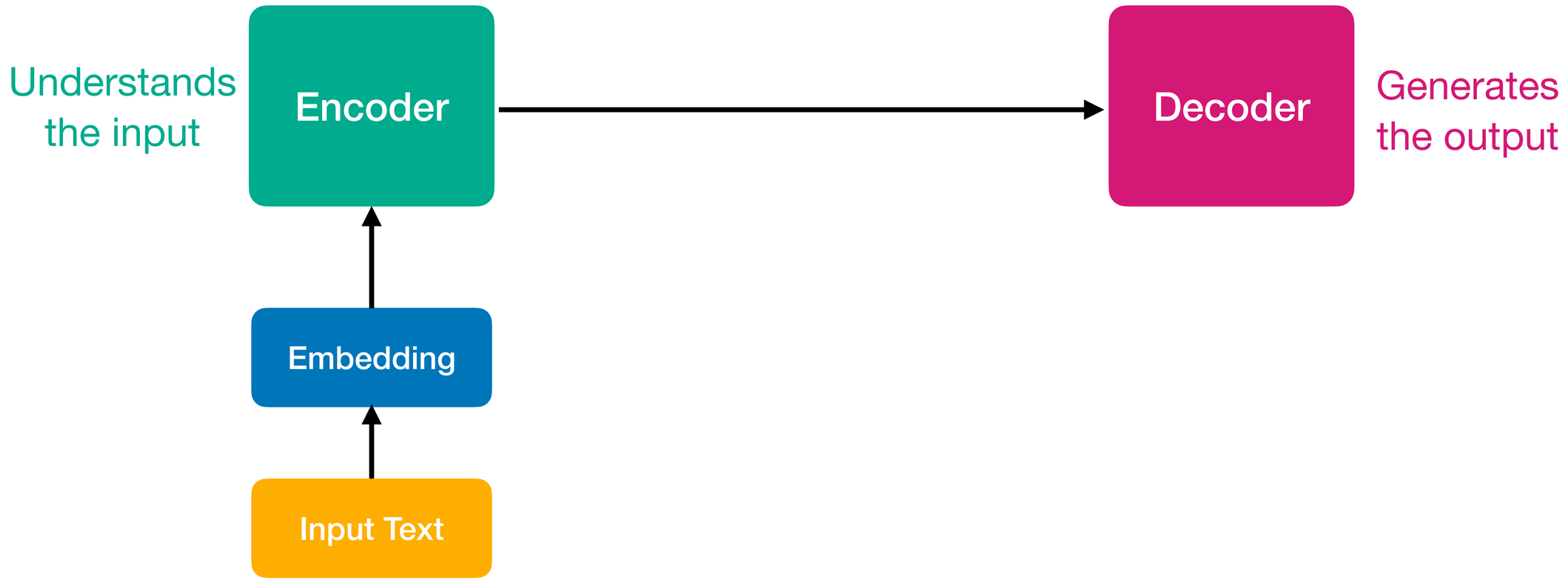
Transformer



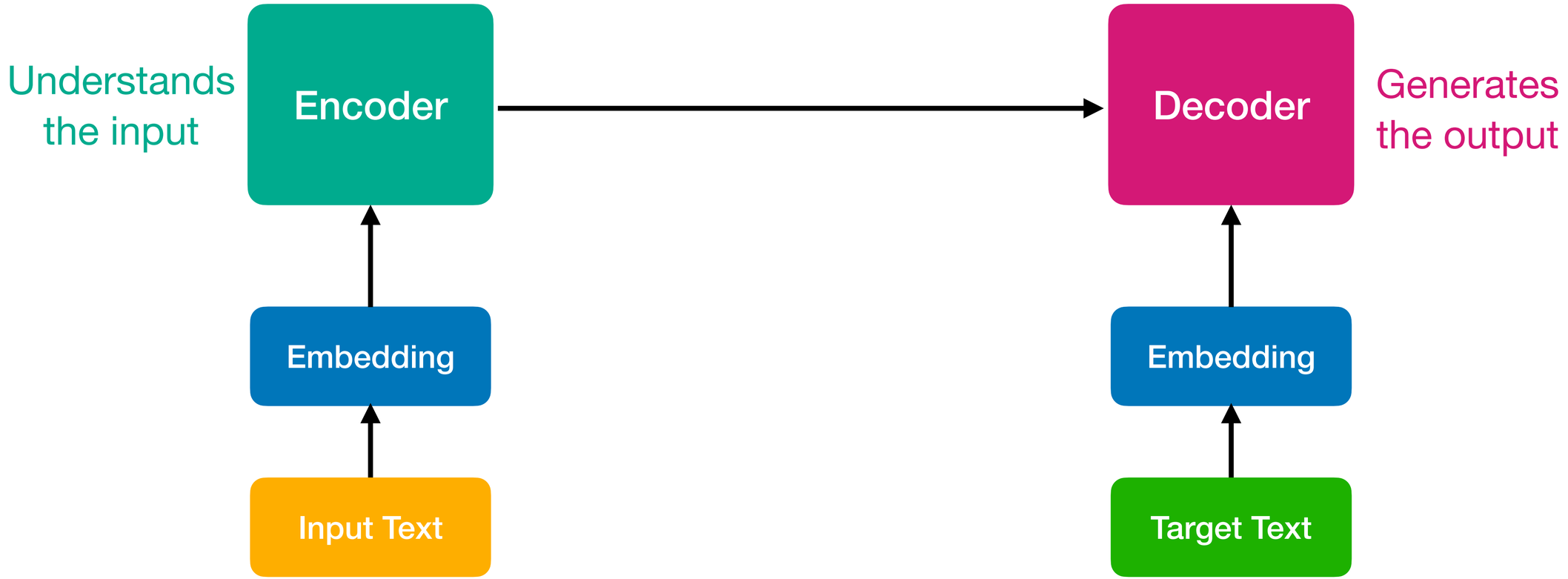
Transformer



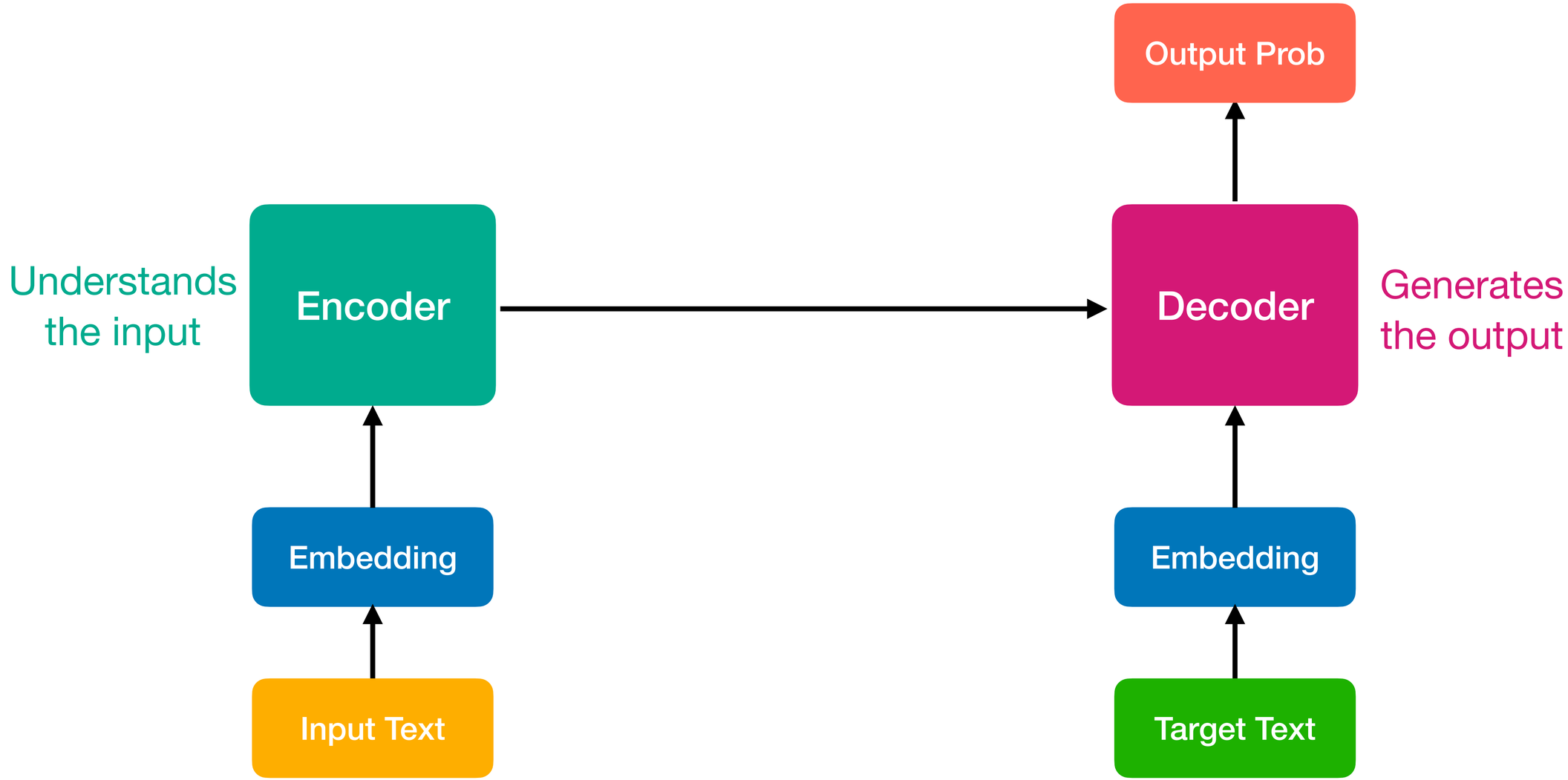
Transformer



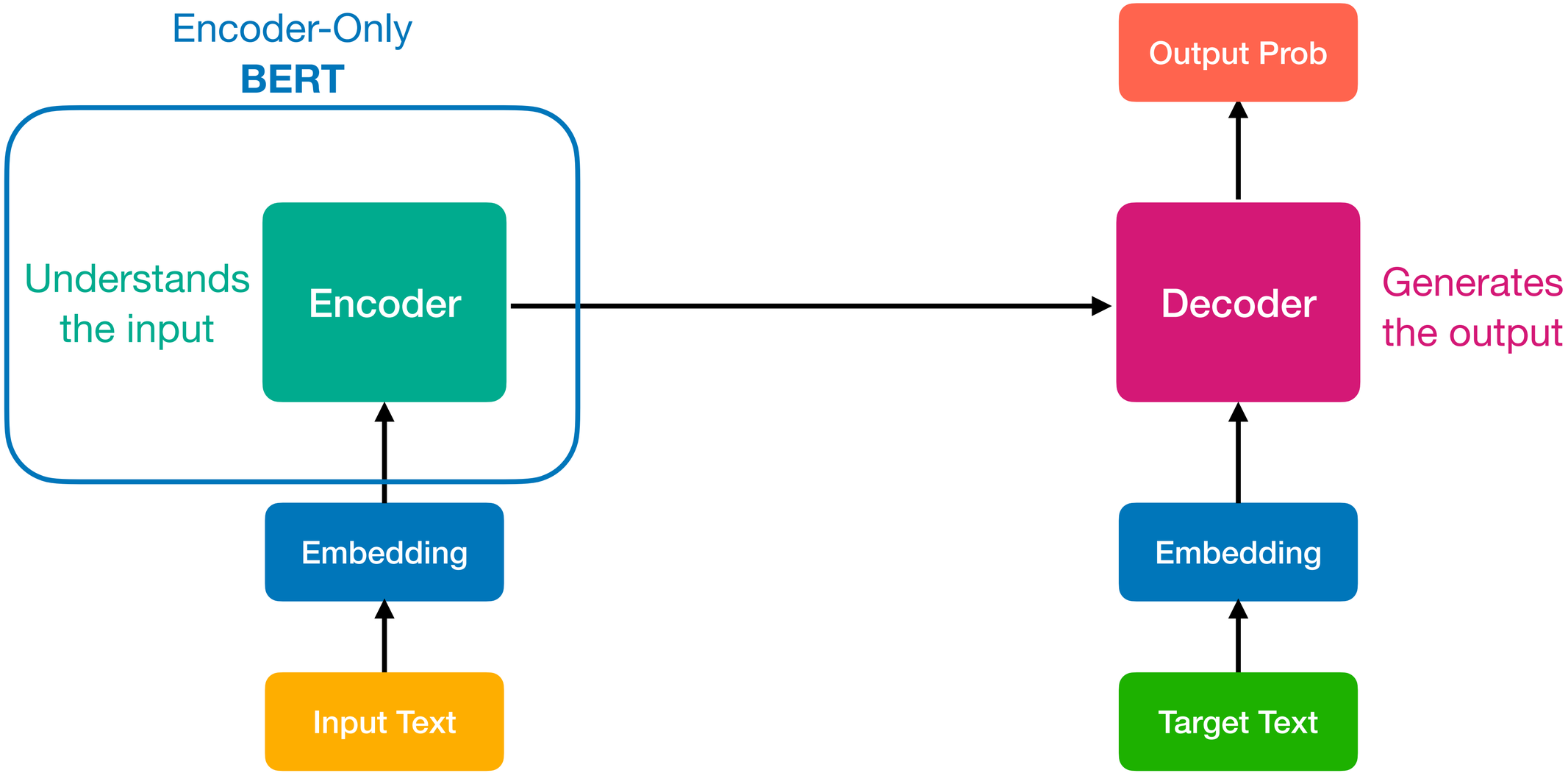
Transformer



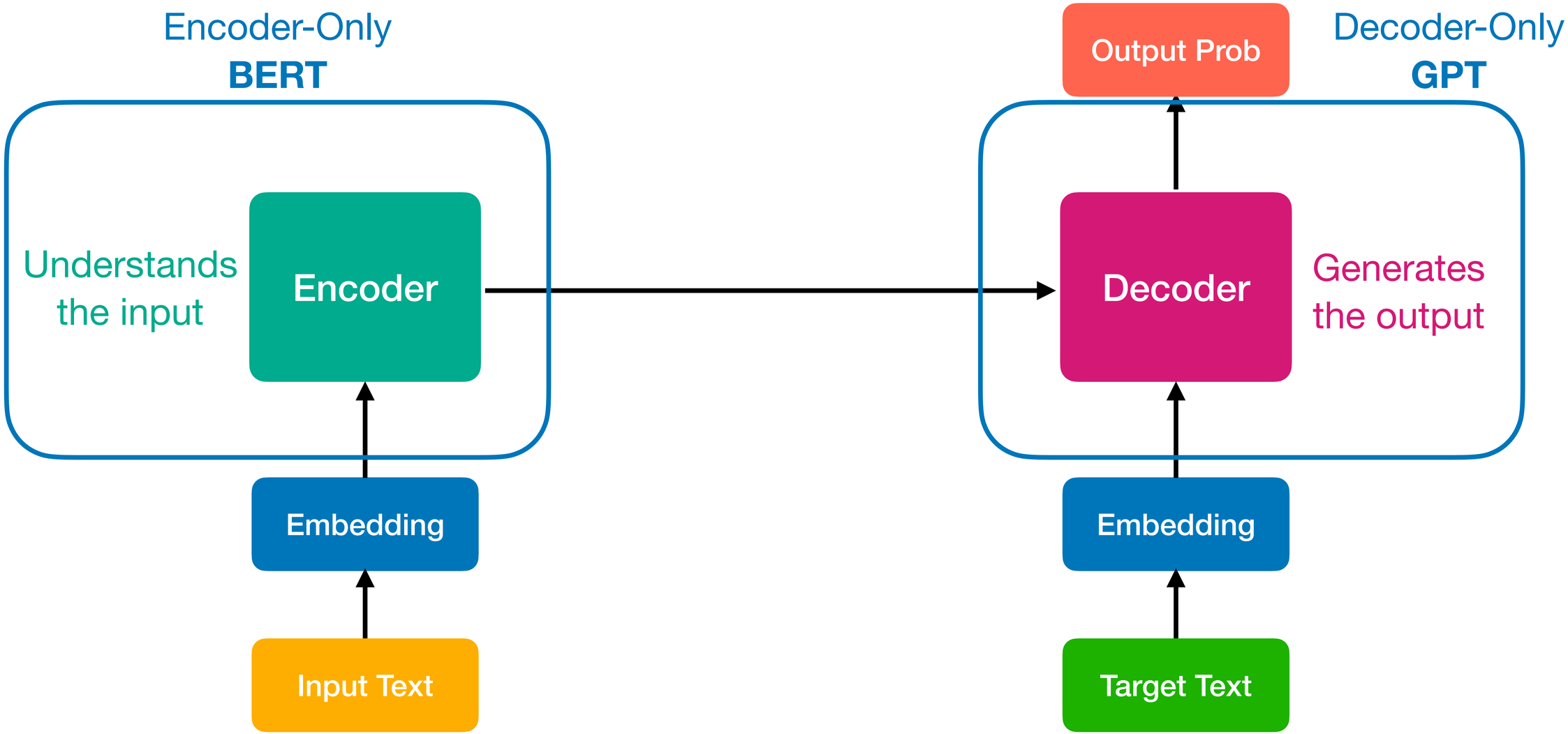
Transformer



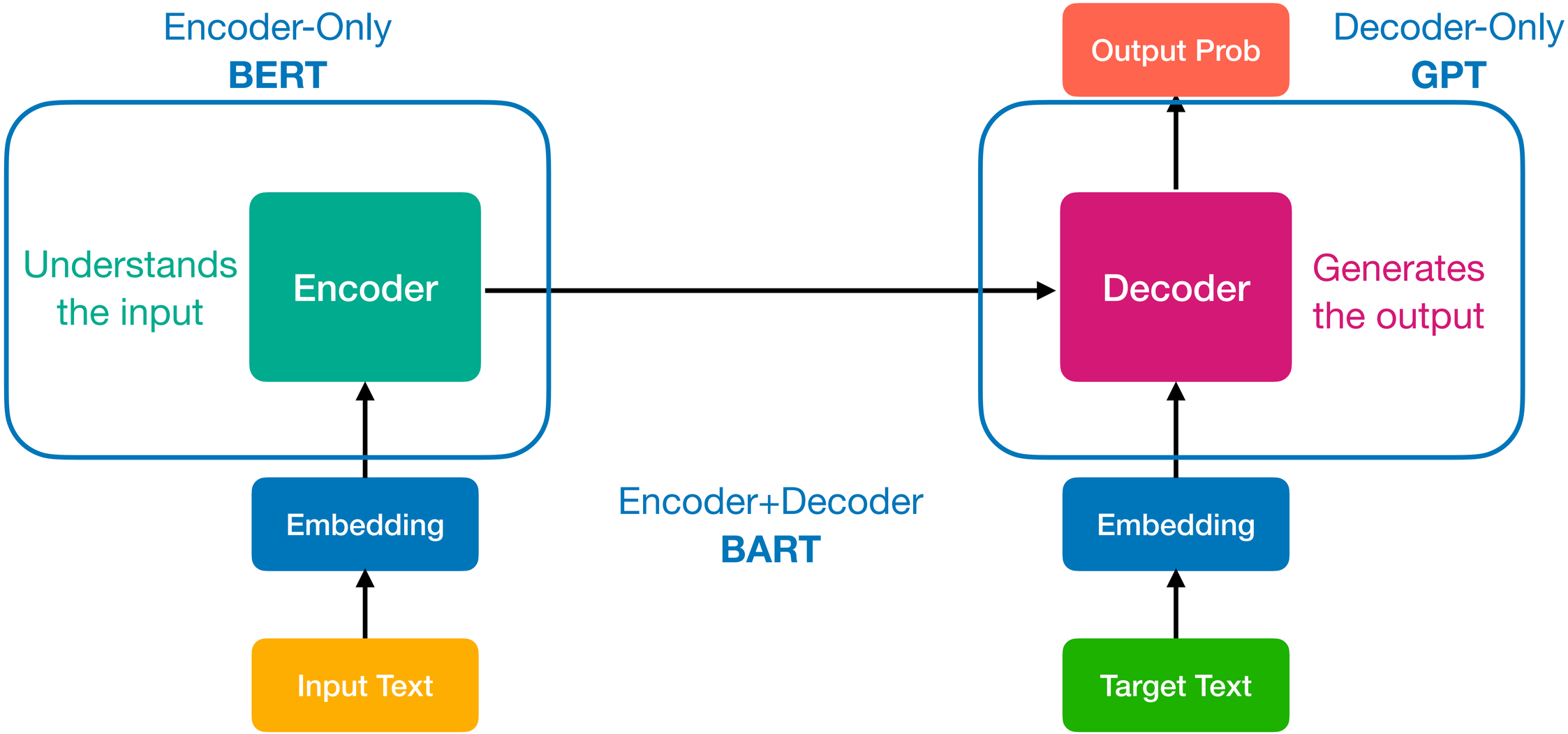
Transformer



Transformer



Transformer



Transformer

Transformer

- BERT: sentiment, topic, feature extraction

Transformer

- BERT: sentiment, topic, feature extraction
- GPT: freeform text generation, chatbots, creative writing

Transformer

- BERT: sentiment, topic, feature extraction
- GPT: freeform text generation, chatbots, creative writing
- BART: translation, summarization

Sentiment Analysis

Sentiment Analysis

“The plot was somewhat predictable,
but the acting was superb.”

Sentiment Analysis

“The plot was somewhat predictable,
but the acting was superb.”

Lexicon lookup:

- predicate: -1
- superb: +2

Sentiment Analysis

“The plot was somewhat predictable,
but the acting was superb.”

Lexicon lookup:

- predicate: -1
- superb: +2

Multipliers:

- somewhat: 0.5

Sentiment Analysis

"The plot was somewhat predictable,
but the acting was superb."

Lexicon lookup:

- predicate: -1
- superb: +2

Multipliers:

- somewhat: 0.5

Sentiment Score:

$0.5 \times (-1) + 2 = +1.5 > 0$

Sentiment Analysis

"The plot was somewhat predictable, but the acting was superb."

Lexicon lookup:

- predicate: -1
- superb: +2

Rule-Based Method

Multipliers:

- somewhat: 0.5

Sentiment Score:

$0.5 \times (-1) + 2 = +1.5 > 0$

Sentiment Analysis

"The plot was somewhat predictable, but the acting was superb."

Lexicon lookup:

- predicate: -1
- superb: +2

Rule-Based Method

Multipliers:

- somewhat: 0.5

Sentiment Score:

$0.5 \times (-1) + 2 = +1.5 > 0$

"The plot was somewhat predictable, but the acting was superb."

Sentiment Analysis

"The plot was somewhat predictable, but the acting was superb."

Lexicon lookup:

- predicate: -1
- superb: +2

Rule-Based Method

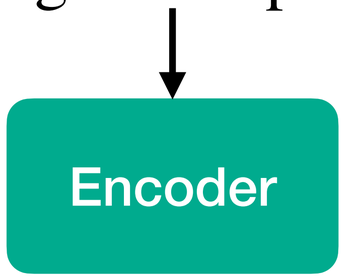
Multipliers:

- somewhat: 0.5

Sentiment Score:

$0.5 \times (-1) + 2 = +1.5 > 0$

"The plot was somewhat predictable, but the acting was superb."



Sentiment Analysis

"The plot was somewhat predictable, but the acting was superb."

Lexicon lookup:

- predicate: -1
- superb: +2

Rule-Based Method

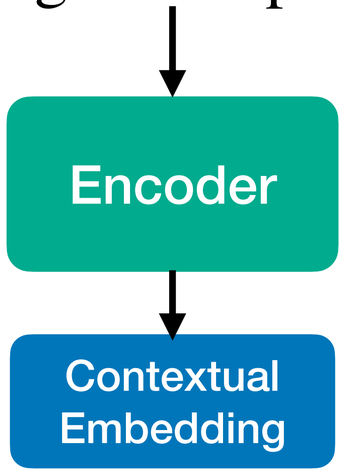
Multipliers:

- somewhat: 0.5

Sentiment Score:

$0.5 \times (-1) + 2 = +1.5 > 0$

"The plot was somewhat predictable, but the acting was superb."



Sentiment Analysis

"The plot was somewhat predictable, but the acting was superb."

Lexicon lookup:

- predicate: -1
- superb: +2

Rule-Based Method

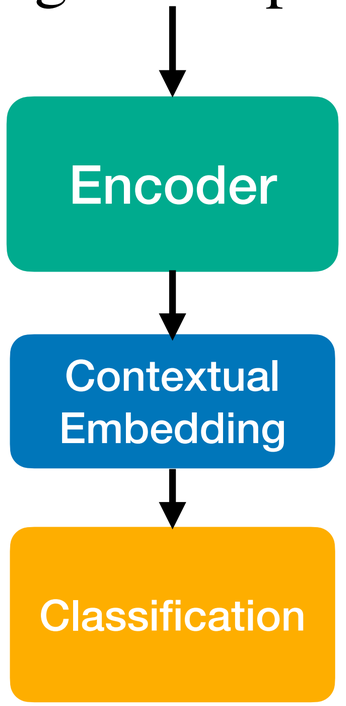
Multipliers:

- somewhat: 0.5

Sentiment Score:

$0.5 \times (-1) + 2 = +1.5 > 0$

"The plot was somewhat predictable, but the acting was superb."



Sentiment Analysis

"The plot was somewhat predictable, but the acting was superb."

Lexicon lookup:

- predicate: -1
- superb: +2

Rule-Based Method

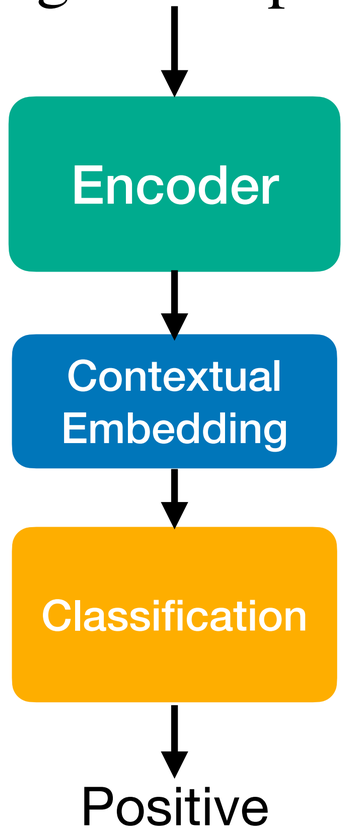
Multipliers:

- somewhat: 0.5

Sentiment Score:

$0.5 \times (-1) + 2 = +1.5 > 0$

"The plot was somewhat predictable, but the acting was superb."



Sentiment Analysis

"The plot was somewhat predictable, but the acting was superb."

Lexicon lookup:

- predicate: -1
- superb: +2

Rule-Based Method

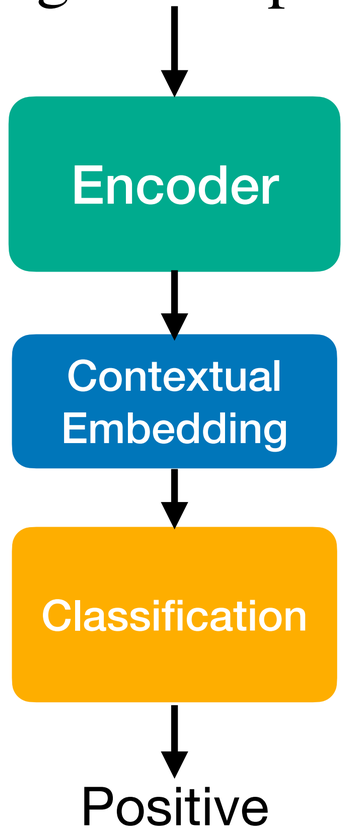
Multipliers:

- somewhat: 0.5

Sentiment Score:

$0.5 \times (-1) + 2 = +1.5 > 0$

"The plot was somewhat predictable, but the acting was superb."



BERT

Sentiment Analysis

"The plot was somewhat predictable, but the acting was superb."

Lexicon lookup:

- predicate: -1
- superb: +2

Multipliers:

- somewhat: 0.5

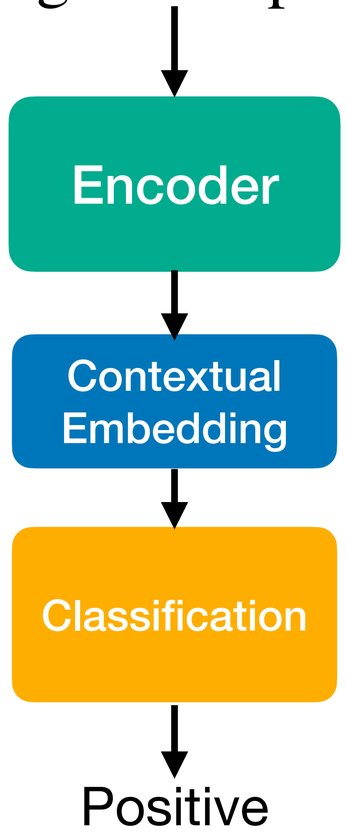
Sentiment Score:

$0.5 \times (-1) + 2 = +1.5 > 0$

Rule-Based Method

Advanced 3

"The plot was somewhat predictable, but the acting was superb."



BERT

Topic Analysis

Topic Analysis

LDA

Topic Analysis

LDA

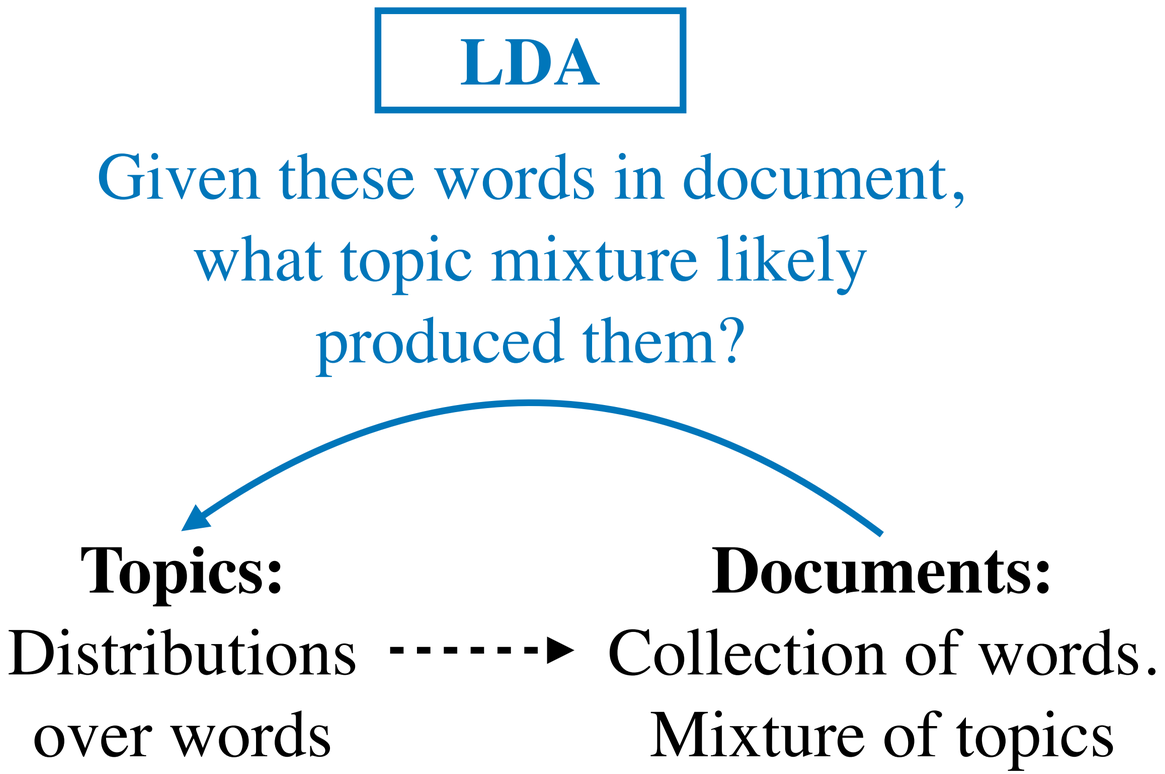
Topics:
Distributions
over words

Topic Analysis

LDA

Topics:		**Documents:**
Distributions over words	------►	Collection of words. Mixture of topics

Topic Analysis



Topic Analysis

LDA

Given these words in document,
what topic mixture likely
produced them?

Topics:
Distributions
over words

Documents:
Collection of words.
Mixture of topics

Raw Documents

Topic Analysis

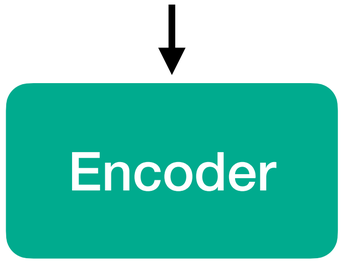
LDA

Given these words in document,
what topic mixture likely
produced them?

Topics:
Distributions
over words

Documents:
Collection of words.
Mixture of topics

Raw Documents



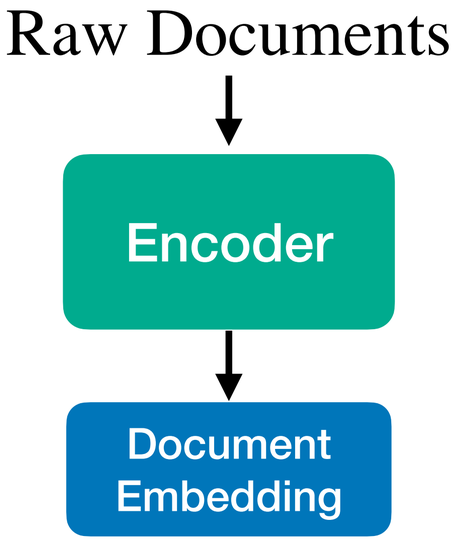
Topic Analysis

LDA

Given these words in document,
what topic mixture likely
produced them?

Topics:
Distributions
over words

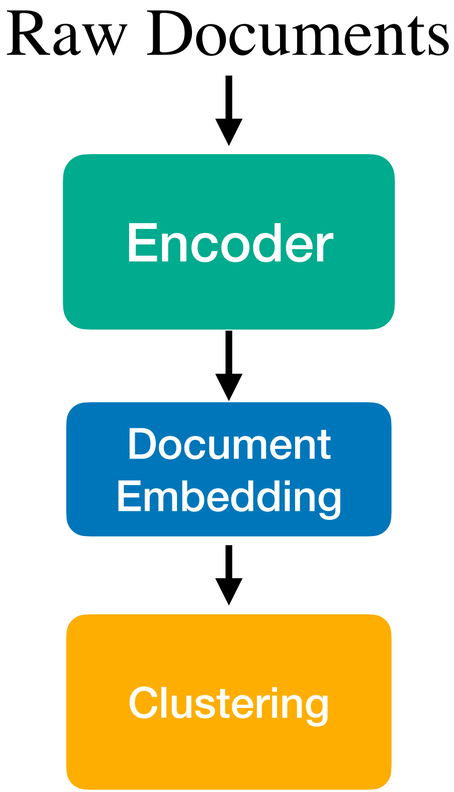
Documents:
Collection of words.
Mixture of topics



Topic Analysis

LDA

Given these words in document,
what topic mixture likely
produced them?



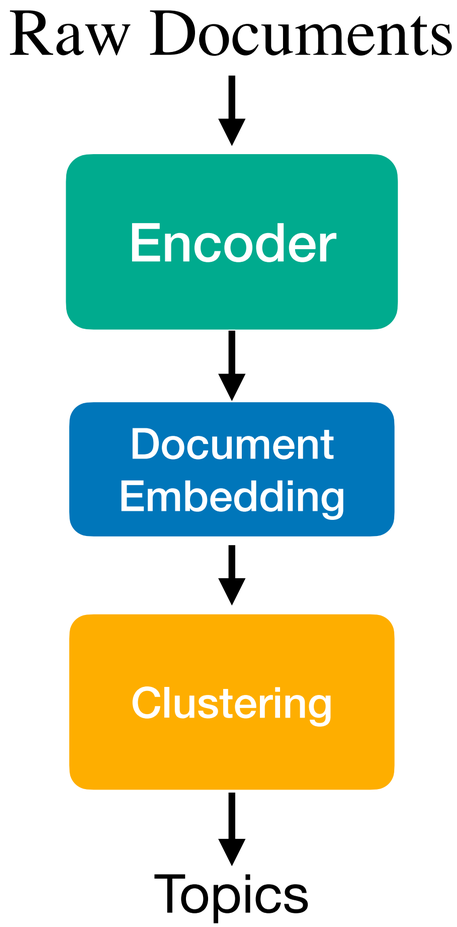
Topic Analysis

LDA

Given these words in document,
what topic mixture likely
produced them?

Topics:
Distributions
over words

Documents:
Collection of words.
Mixture of topics



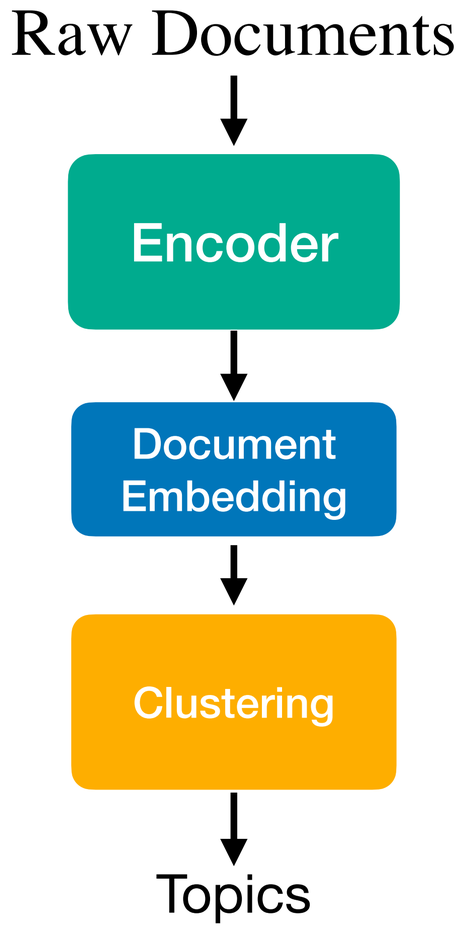
Topic Analysis

LDA

Given these words in document,
what topic mixture likely
produced them?

Topics:
Distributions
over words

Documents:
Collection of words.
Mixture of topics



BERTopic

Topic Analysis

LDA

Given these words in document,
what topic mixture likely
produced them?

Topics:
Distributions
over words

Documents:
Collection of words.
Mixture of topics

Advanced 3

Raw Documents

Encoder

Document Embedding

Clustering

Topics

BERTopic