

O'REILLY®

Data Chunking

Bert Gollnick





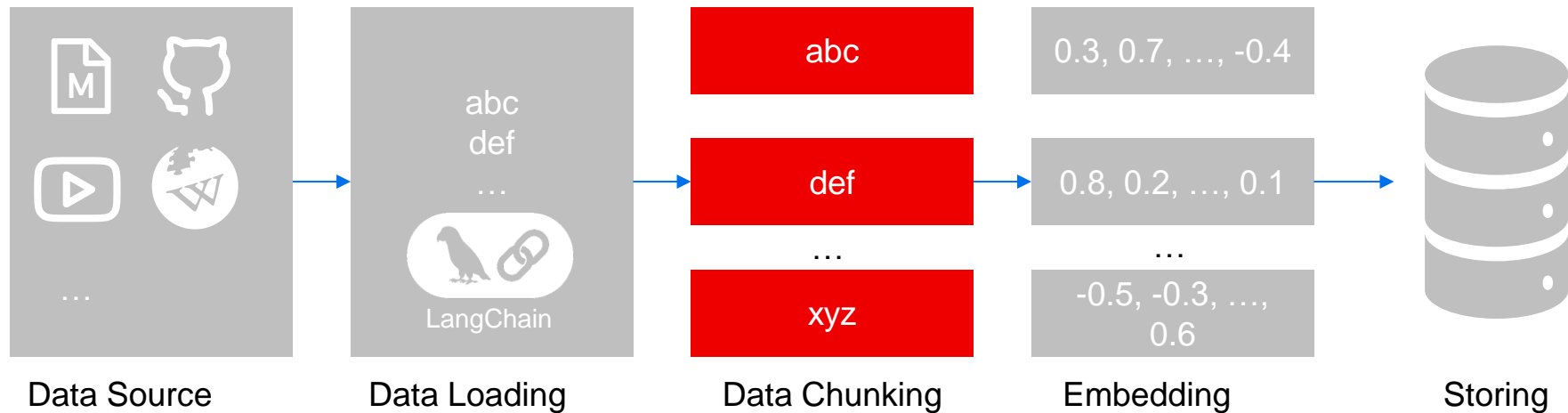
Learning Objectives

By the end of this module, you will:

- Know how to split the data
- Why it is important to split the data
- Know what tokens and sequence lengths are



Data Ingestion Process

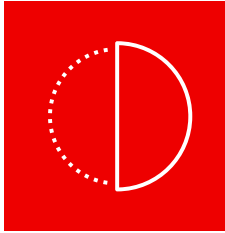




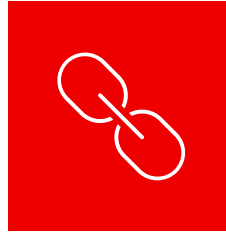
What is „Chunking“?

- Dividing larger pieces of information into smaller, manageable units
- These units called „chunks“
- Required to fit model context window
- Chunks should be:
 - Small
 - Semantically meaningful

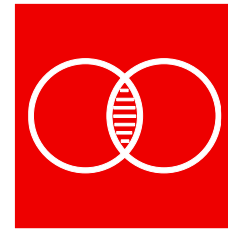
Process



Split



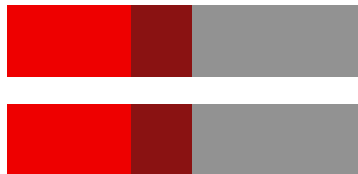
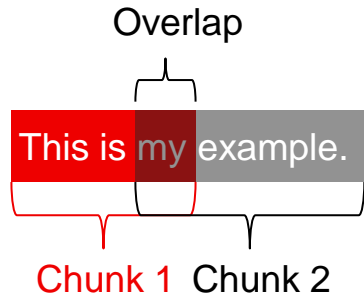
Combine



Overlap

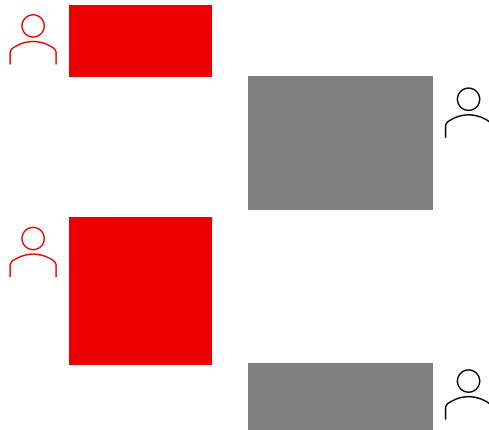


Chunking Approaches



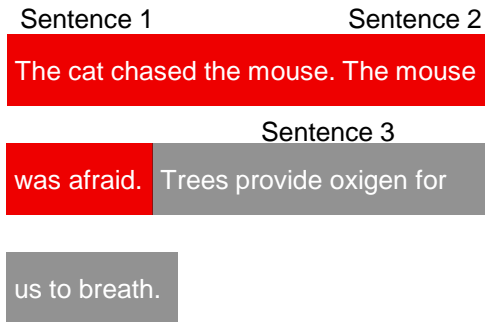
Fixed Chunk-Sizes

- Identical
- pre-defined



Structure-Based Chunk-Sizes

- e.g. chat messages should be consistent, no mix of users and chunks



- Sentence 1 and 2 are very similar → same chunk
- Sentence 3 different → new chunk

Semantic Chunking

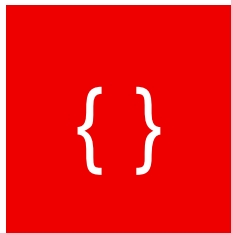
- based on semantic similarity
- e.g. when semantic break is observed



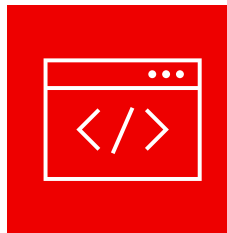
Splitter-Types



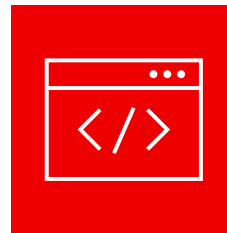
Text



JSON



HTML



Code



Recursive Splitting

- `chunk_size`...defines maximum size of chunks [characters]
- `chunk_overlap`...possible overlap of max 5 characters

The quick brown fox jumps
over the lazy dog.\n This is a
simple example to show text
splitting.\n.

```
RecursiveCharacterTextSplitter(  
    chunk_size=20,  
    chunk_overlap=5  
    separators=["\n", " ", "\""]  
)
```

The quick brown fox

brown fox jumps

jumps over the lazy

the lazy dog.

This is a simple

simple example to

to show text

text splitting.



Tokenization

The quick brown fox jumps over the lazy dog.

The

quick

brown

fox

jumps

over

the

lazy

dog

.

Tokenization

791

4062

14198

39935

35308

927

279

16053

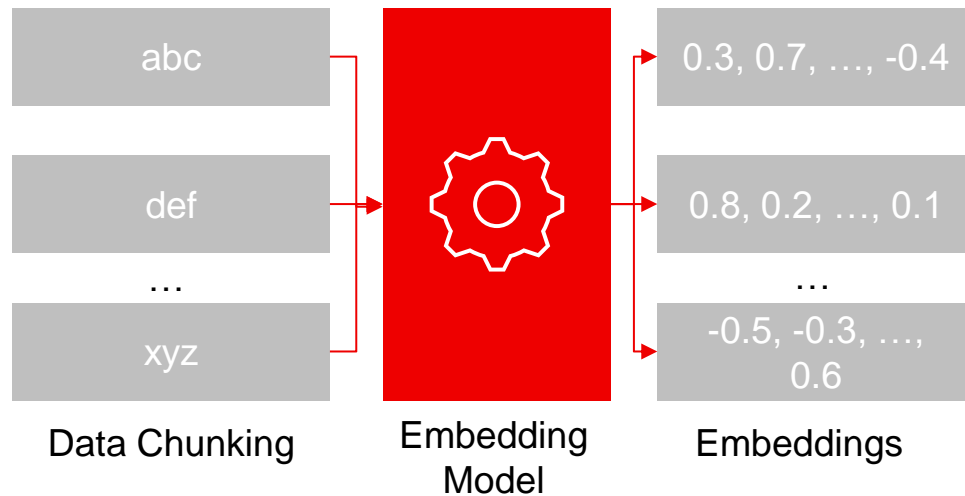
5679

13



Context Window

- Embedding model works with tokens, NOT words
- Model can cover only specific sequence lengths
- Too long text (longer than context window) will be truncated



The image features the O'Reilly logo in white, bold, sans-serif capital letters, with a registered trademark symbol (®) at the end. The logo is centered horizontally and positioned in the middle of the frame. The background is a smooth gradient from deep orange on the left to bright yellow on the right. Overlaid on this background are several large, semi-transparent circles in various shades of orange and red, creating a layered, abstract effect.

O'REILLY®