O'REILLY®

Data Chunking

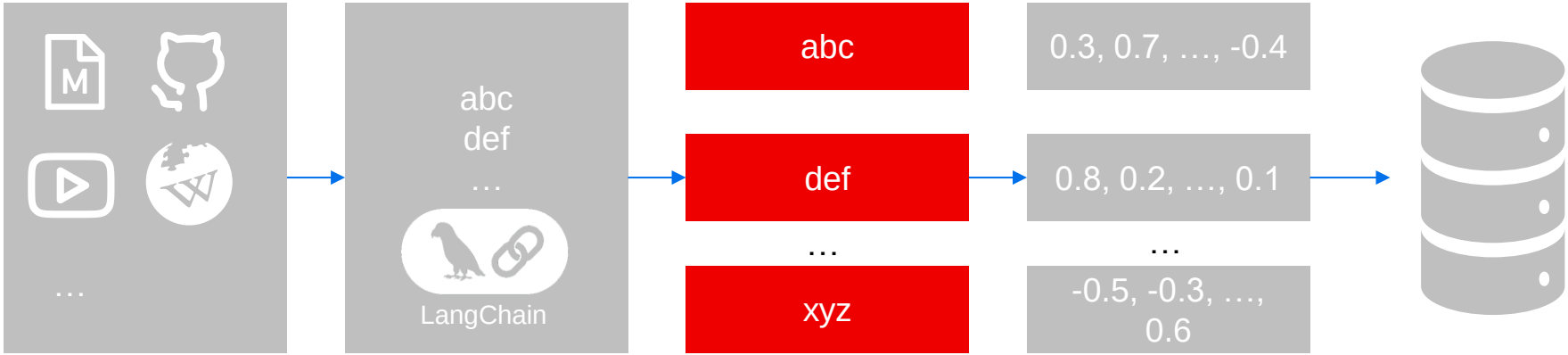
Bert Gollnick

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

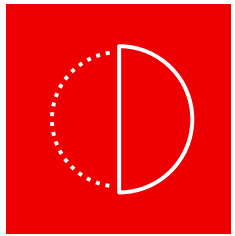
Data Ingestion Process



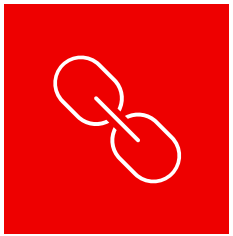
What is „Chunking“?

- 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

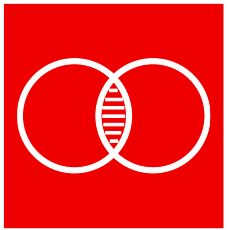
Process



Split

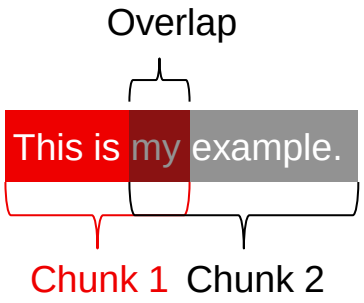


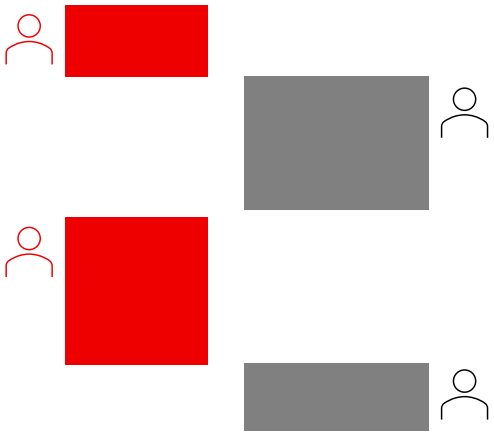
Combine

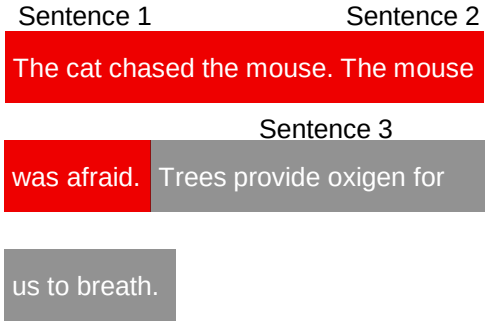


Overlap

Chunking Approaches







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

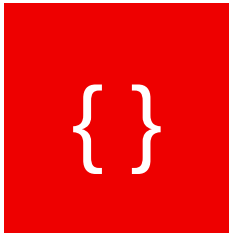
Fixed Chunk-Sizes
- Identical
- pre-defined

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

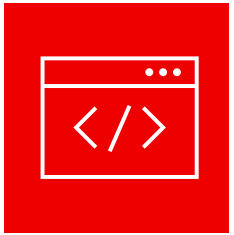
Semantic Chunking
- based on semantic similarity
- e.g. when semantic break is observed

Splitter-Types

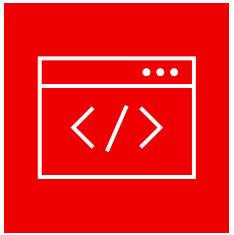
Text



JSON



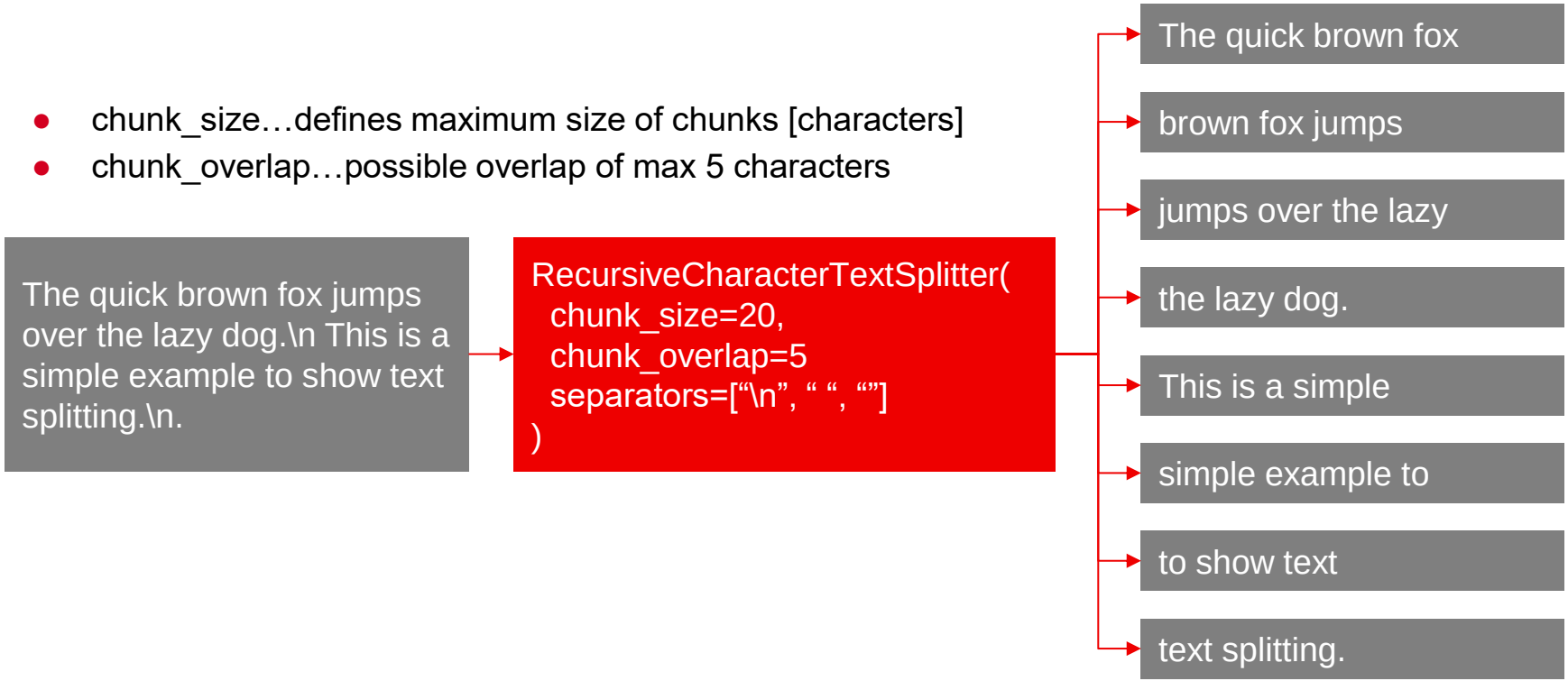
HTML



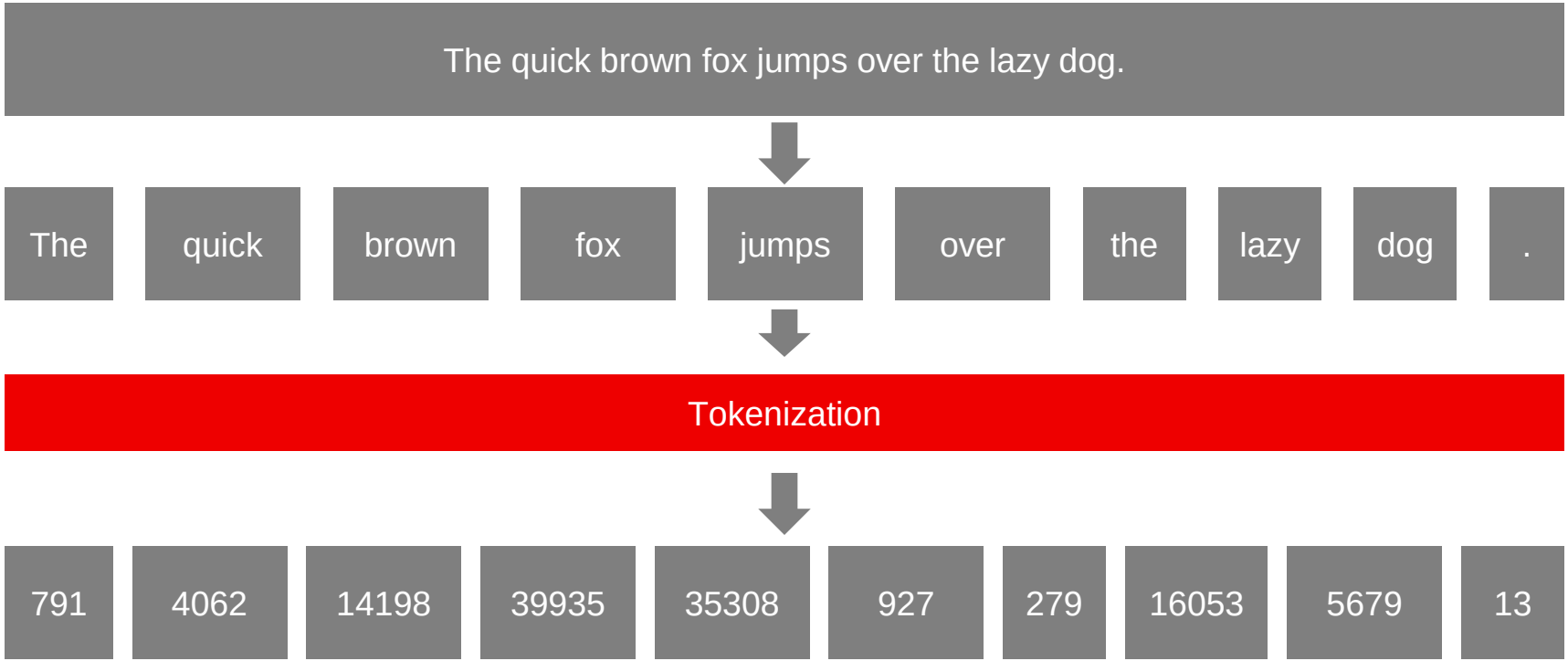
Code

Recursive Splitting

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

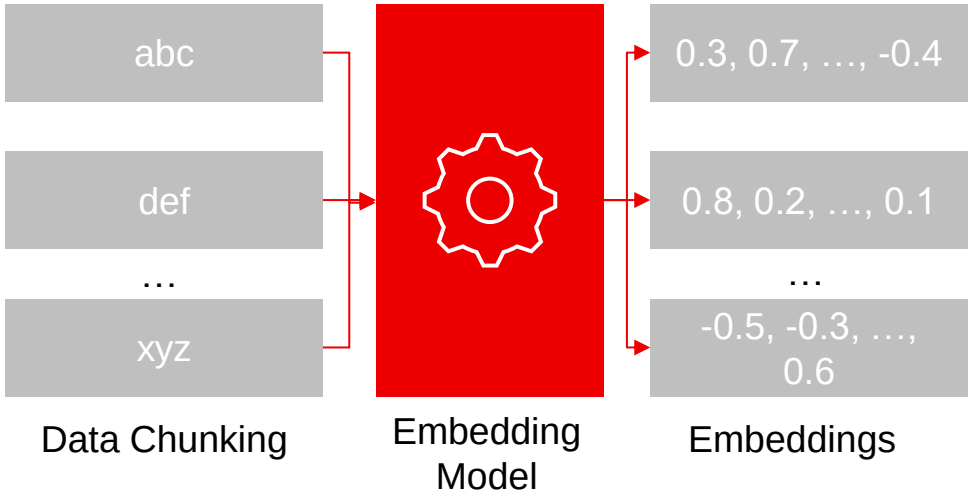


Tokenization



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




O'REILLY®