BERT Pre-training: Two-Task Learning
BERT Training: Masked Language Modeling (MLM)

Input:

The	[MASK]	sat	on	[MASK]	mat

Model:

BERT
(Bidirectional Transformer)

Prediction:

cat   the
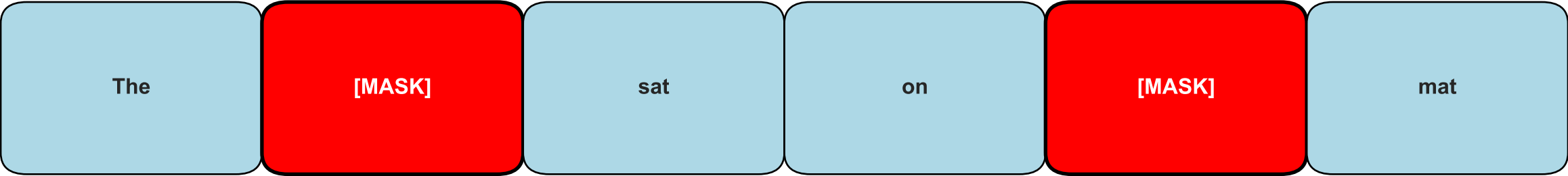
15% of words masked randomly during training

Next Sentence Prediction (NSP)

Sentence A: "The man went to the store."

Sentence B: "He bought milk and bread."

Label: IsNext

Sentence A: "The cat slept all day."

Sentence B: "Quantum physics is complex."

Label: NotNext

Training Data

• BookCorpus: 800M words

• Wikipedia: 2.5B words

• **Total: 3.3B words**

• Training time: 4 days

• Hardware: 4-16 TPUs

• **Parameters: 110M-340M**