Advanced Natural Language Processing and Information Retrieval

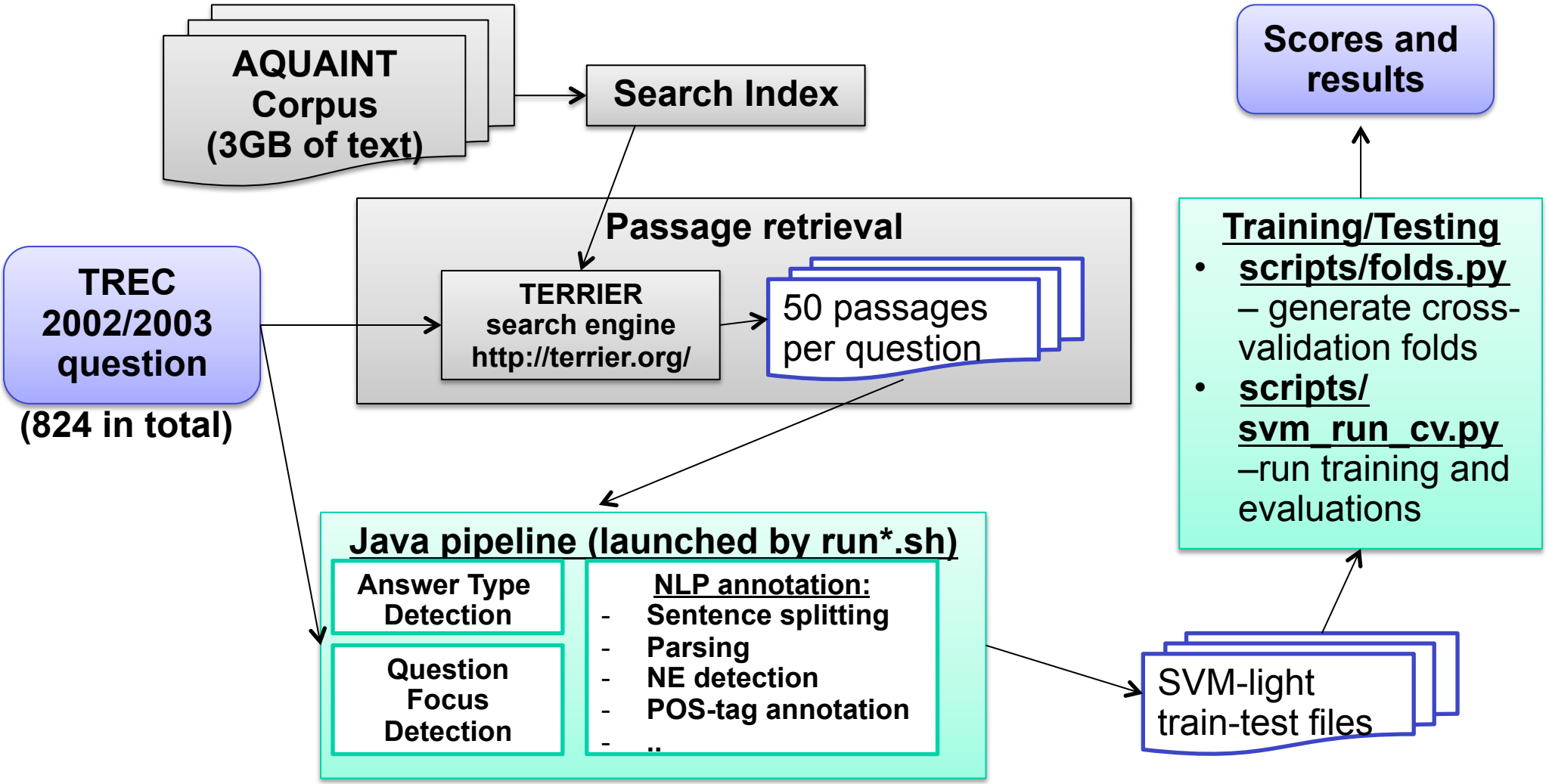
Lab 4: Answer sentence reranking pipeline

Alessandro Moschitti, Massimo Nicosia, Kateryna Tymoshenko
Email: moschitti@disi.unitn.it, m.nicosia@gmail.com, k.tymoshenko@gmail.com

Overview

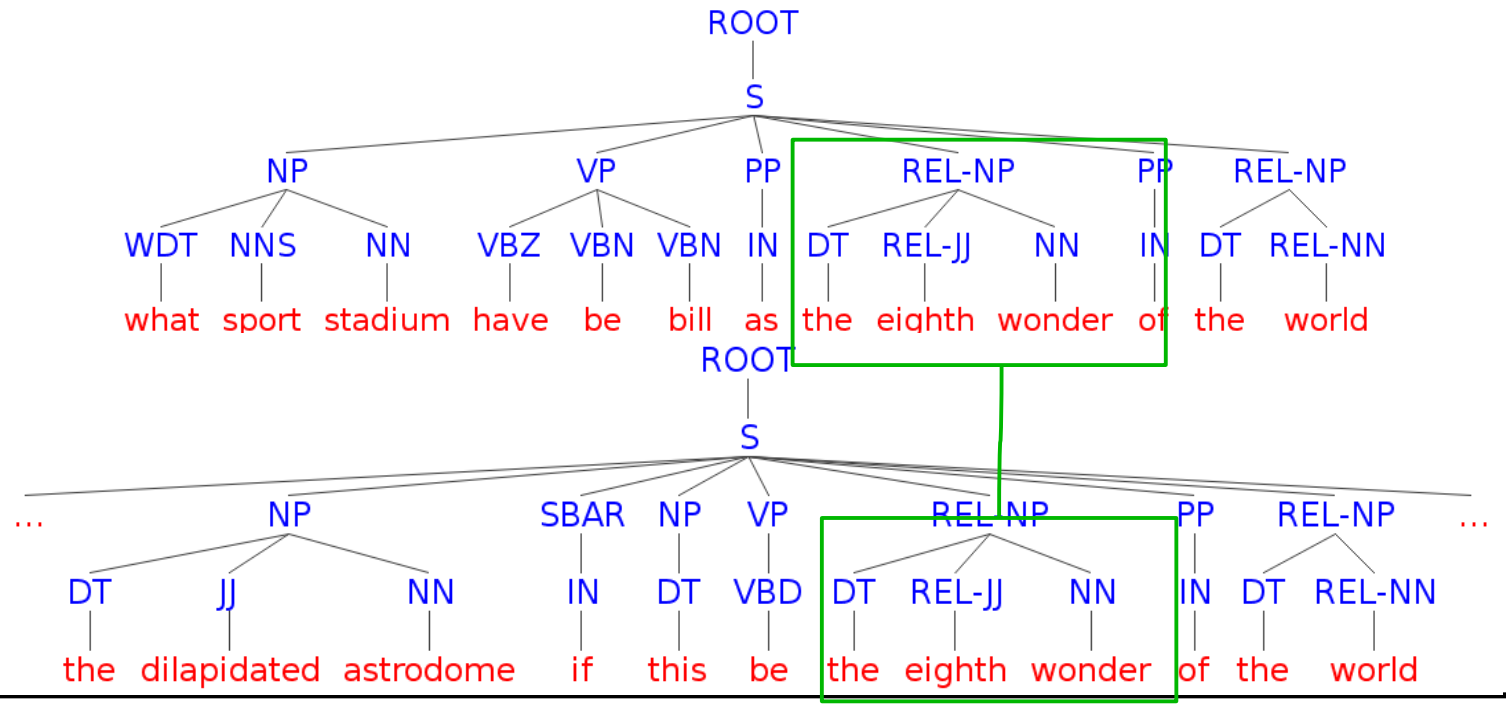
- Software
  - **https://github.com/ktymoshenko/minimalpipeline**
- Answer passage reranking on TREC 2002/2003
- 5 fold-cross validation
- We will build shallow structures with the relational links

IR-based QA overall schema



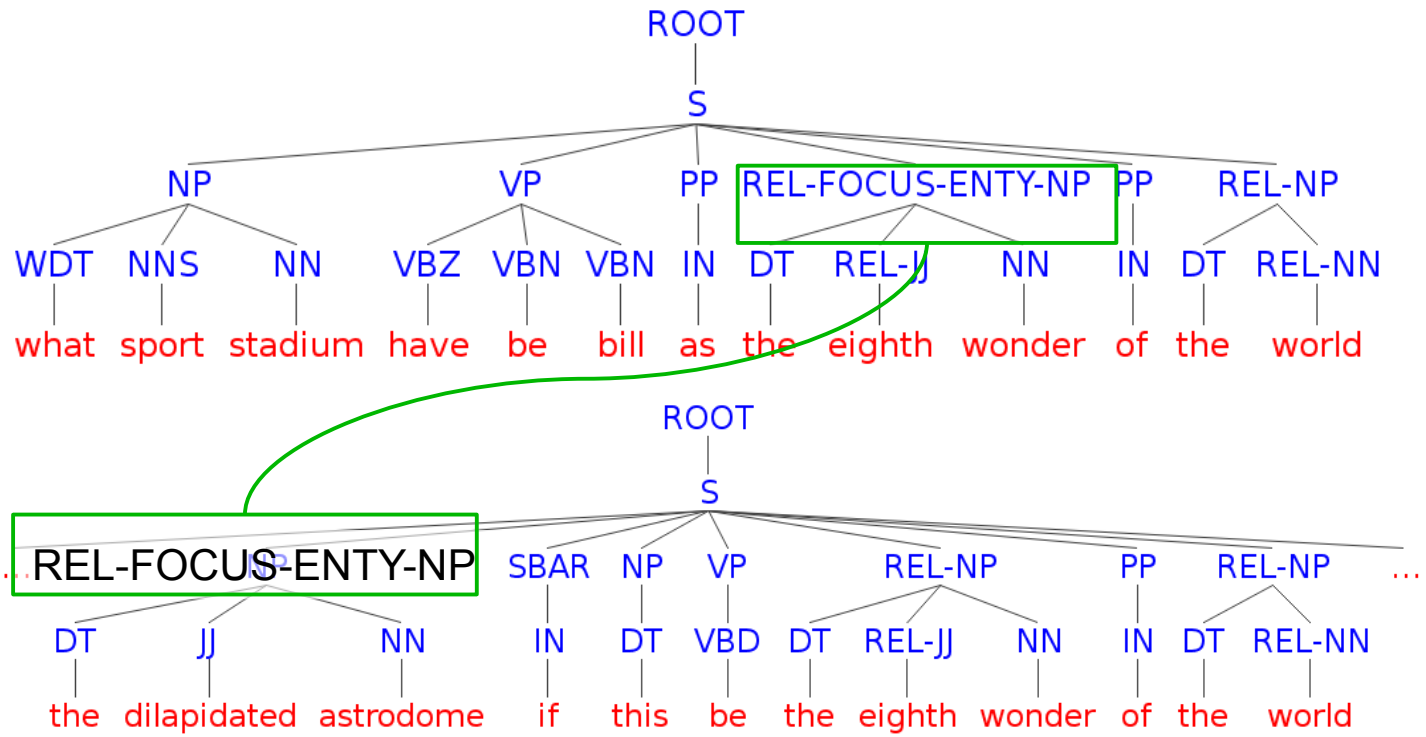
Java pipeline: no focus/question class

- ./run.sh outputs to data/trec-en, all questions
- ./run_toy.sh -> data/trec-en-toy-simple
  - only 200 questions, performance is lower but is faster

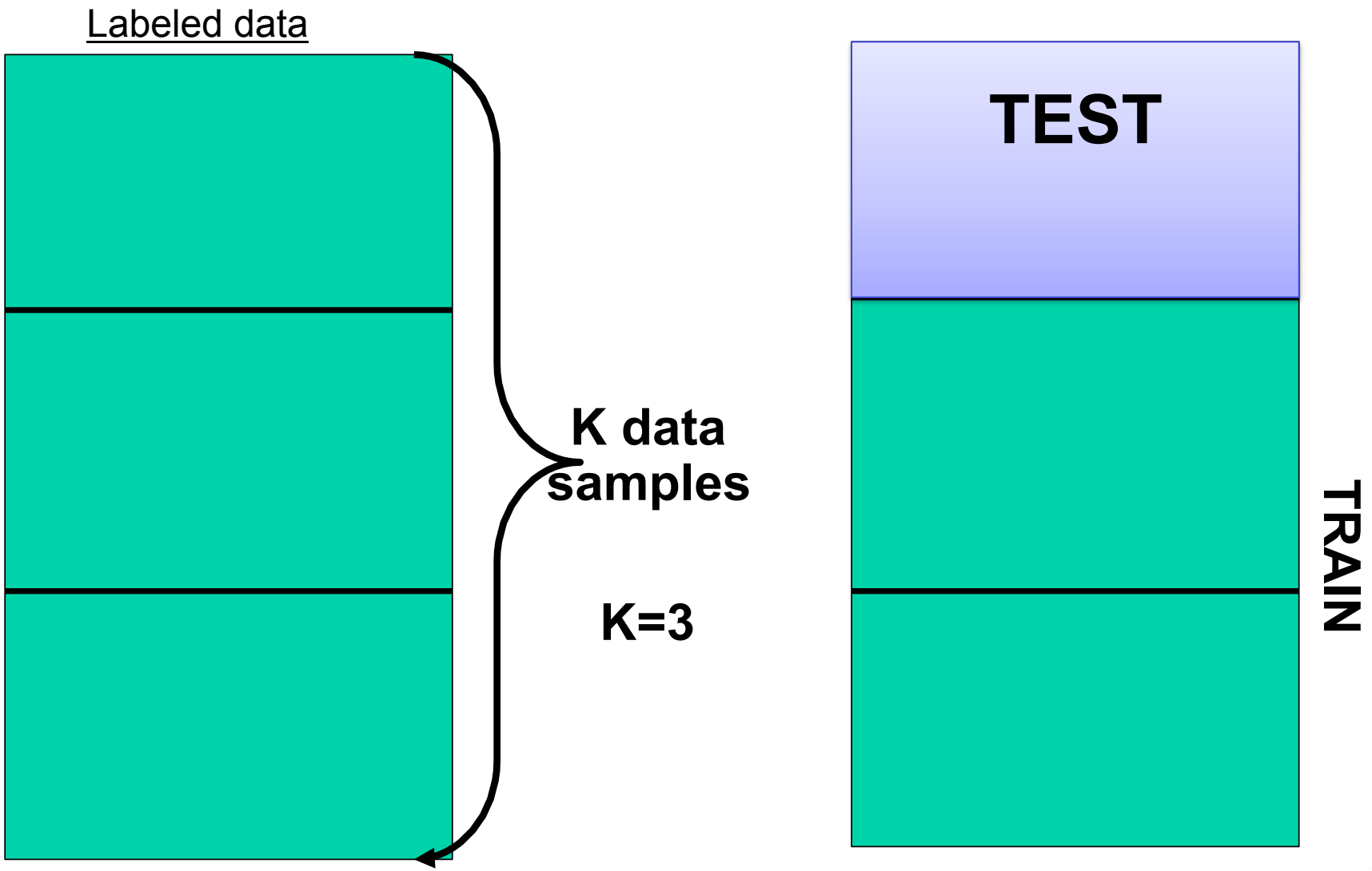


Java pipeline: with question focus/class

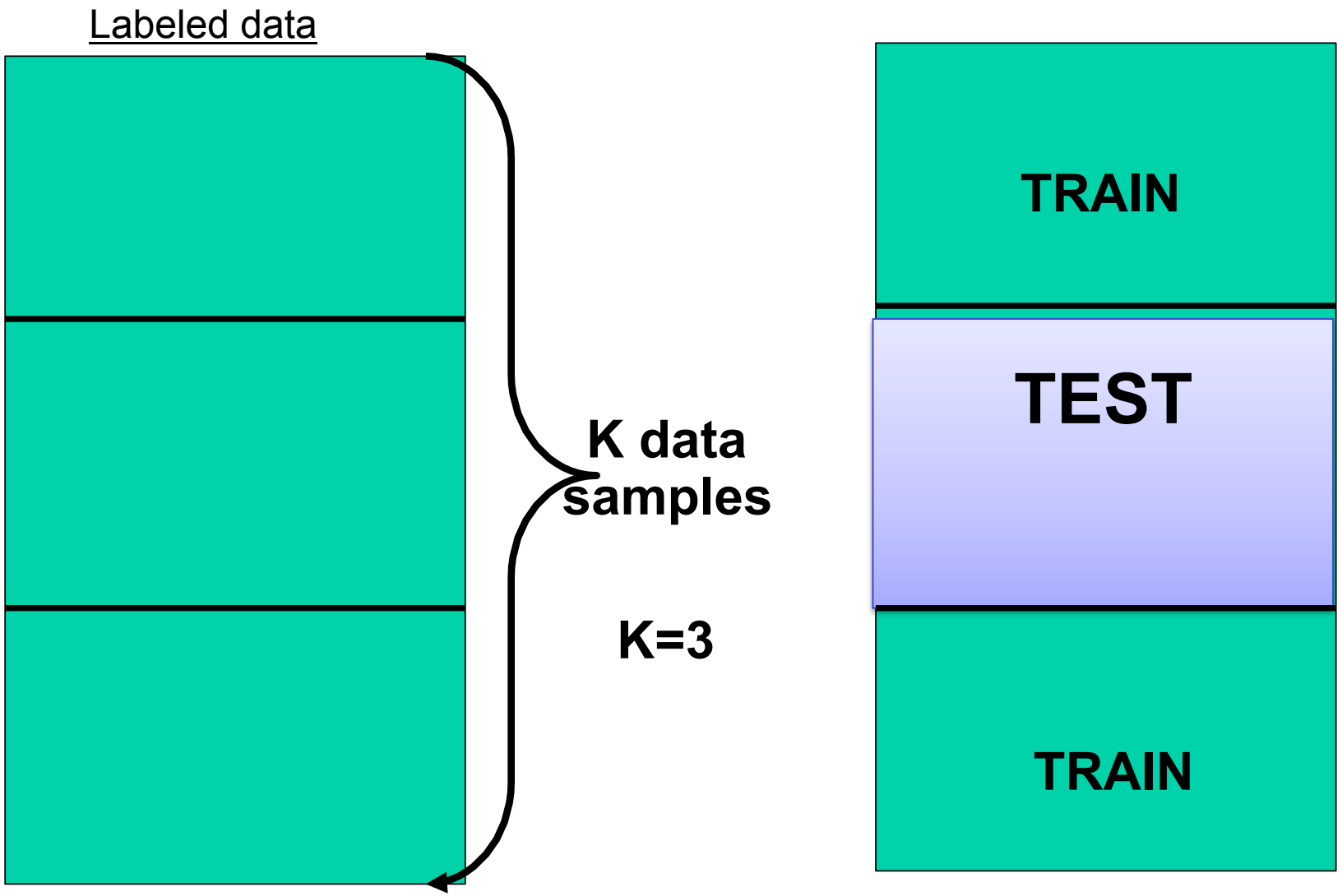
- ./run_qc_qf.sh outputs to data/trec-en, 824 questions
- ./run_qc_qf_toy.sh -> data/trec-en-toy
  - only 200 questions, performance is lower but is faster



N-fold cross-validation



N-fold cross-validation



Generate folds/run experiments

- Generating folds
  - python scripts/folds.py data/**trec-en-qc**/ 5
- Running training and testing on the folds
  - python scripts/svm_*run_cv.py --params="-t 5 -F 3* ***-C +*** *-W R -V R -m 400" --ncpus 2 data/trec-en-qc/ folds/*
  - "-C +" – use structures+similarity feature vectors
  - "-C V" – only feature vectors (fast, less accurate)
  - "-C T" – use only structures, no feature vectors