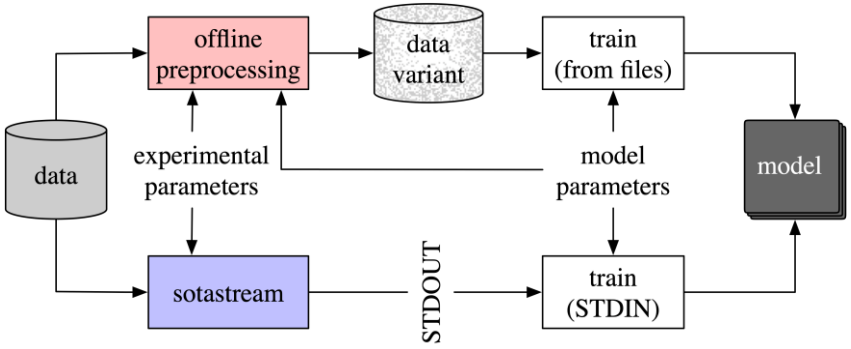
SOTASTREAM: A Streaming Approach to Machine Translation Training

Matt Post, Thamme Gowda, Roman Grundkiewicz, Huda Khayrallah, Rohit Jain, Marcin Junczys-Dowmunt

- Problem: standard off-line data preparation is expensive
  - data tensorized ahead of time
  - takes up time and disk space
  - ties the prepared dataset to a model configuration (e.g., vocabulary)
- Solution: generate data dynamically, on-the-fly!
- SOTASTREAM
  - ✓ Just as accurate
  - ✓ Just as fast
  - ✓ Saves disk space
  - ✓ More flexible

</>

- **pip install sotastream**
- https://github.com/marian-nmt/sotastream
- MIT License



Use Cases

Mixing multiple streams of data	Data augmentation for robustness	Filtering bad data examples
Subword tokenization sampling	Training document-context models	Alignments and other data types
Data collection tools: e.g., mtdata	Generating datasets for offline use	...