

Jun Takagi† Yasunori Ohishi‡ Akisato Kimura‡ Masashi Sugiyama† Makoto Yamada† Hirokazu Kameoka‡

†Graduate School of Information Science and Engineering, Tokyo Institute of Technology ‡NTT Communication Science Laboratories, NTT Corporation

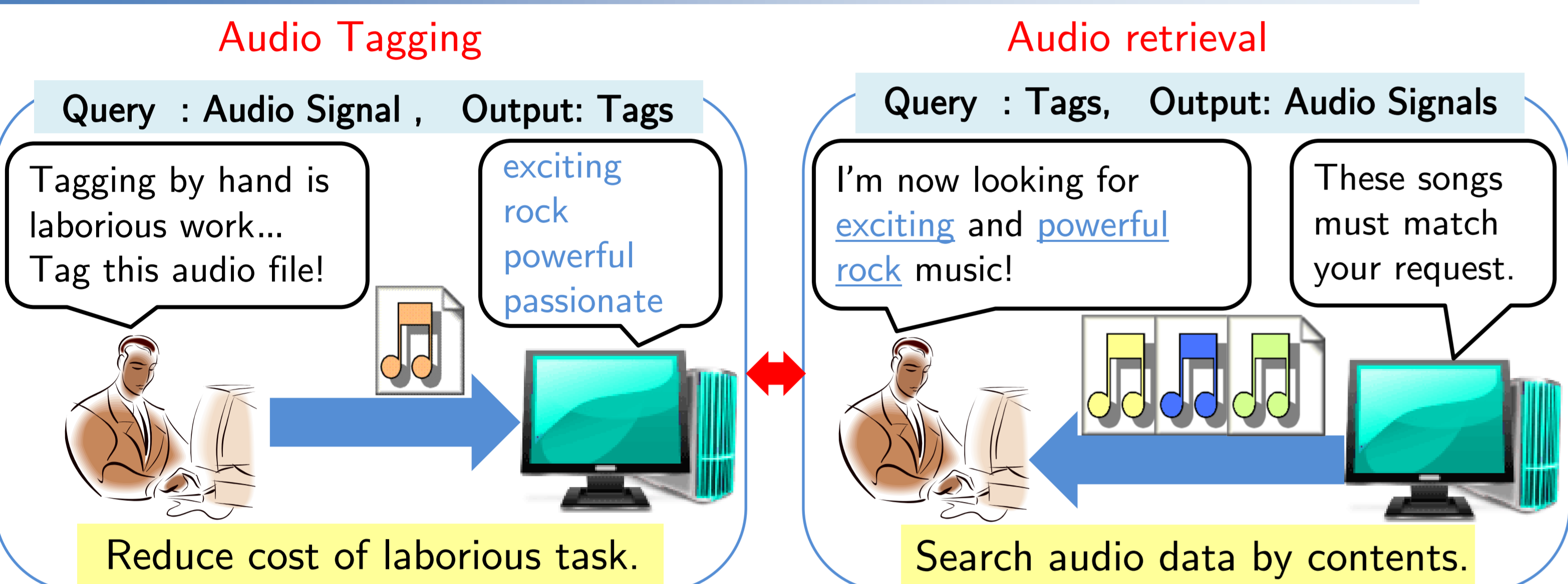
Contact information
E-mail: takagi@sg.os.titech.ac.jp

Overview

Apply semi-supervised method to audio tagging/retrieval.

Utilize untagged samples and reduce the needed number of expensive tagged samples.

What is Audio Tagging/Retrieval?



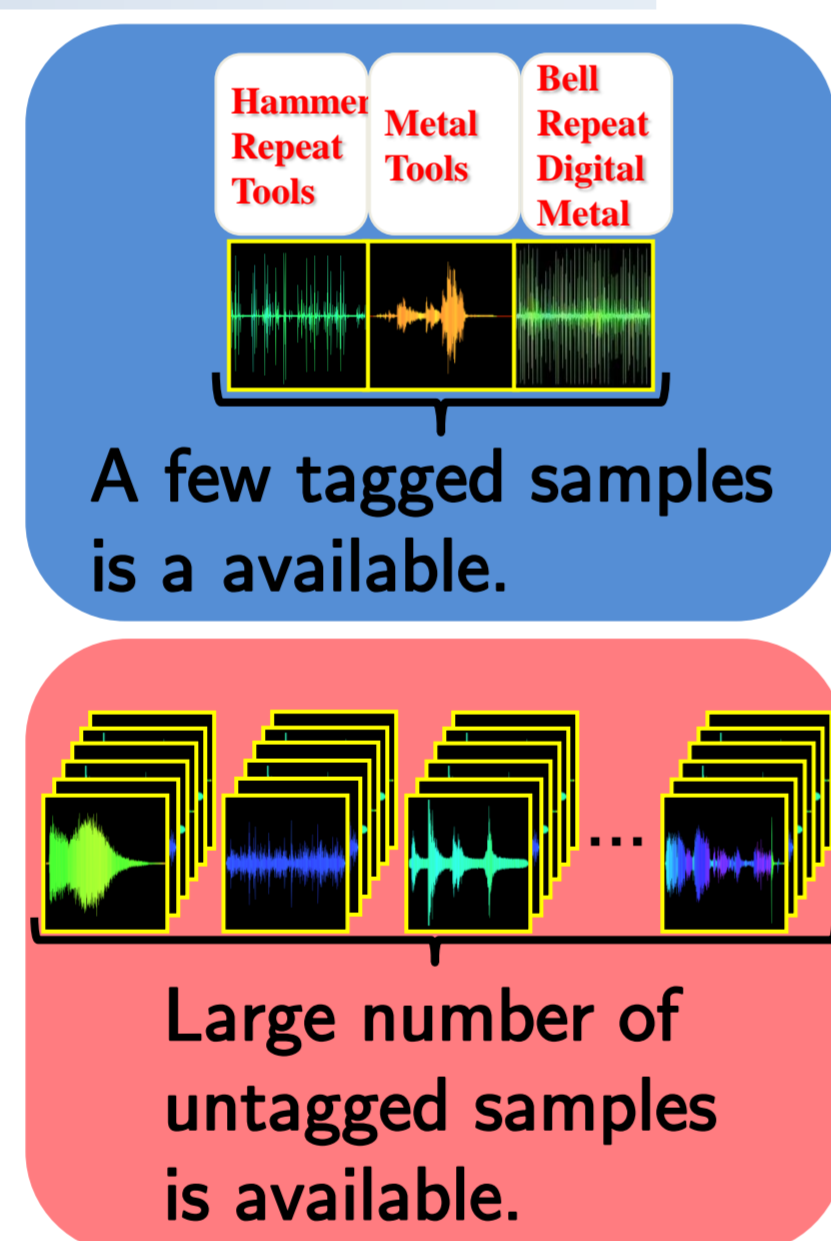
SSCDE can cope with both tasks in the same framework!

Technical Challenges : Use Inexpensive Untagged Audios

Audio samples having high-quality tags are **very expensive!**

Semi-supervised method utilizes inexpensive untagged audio samples!

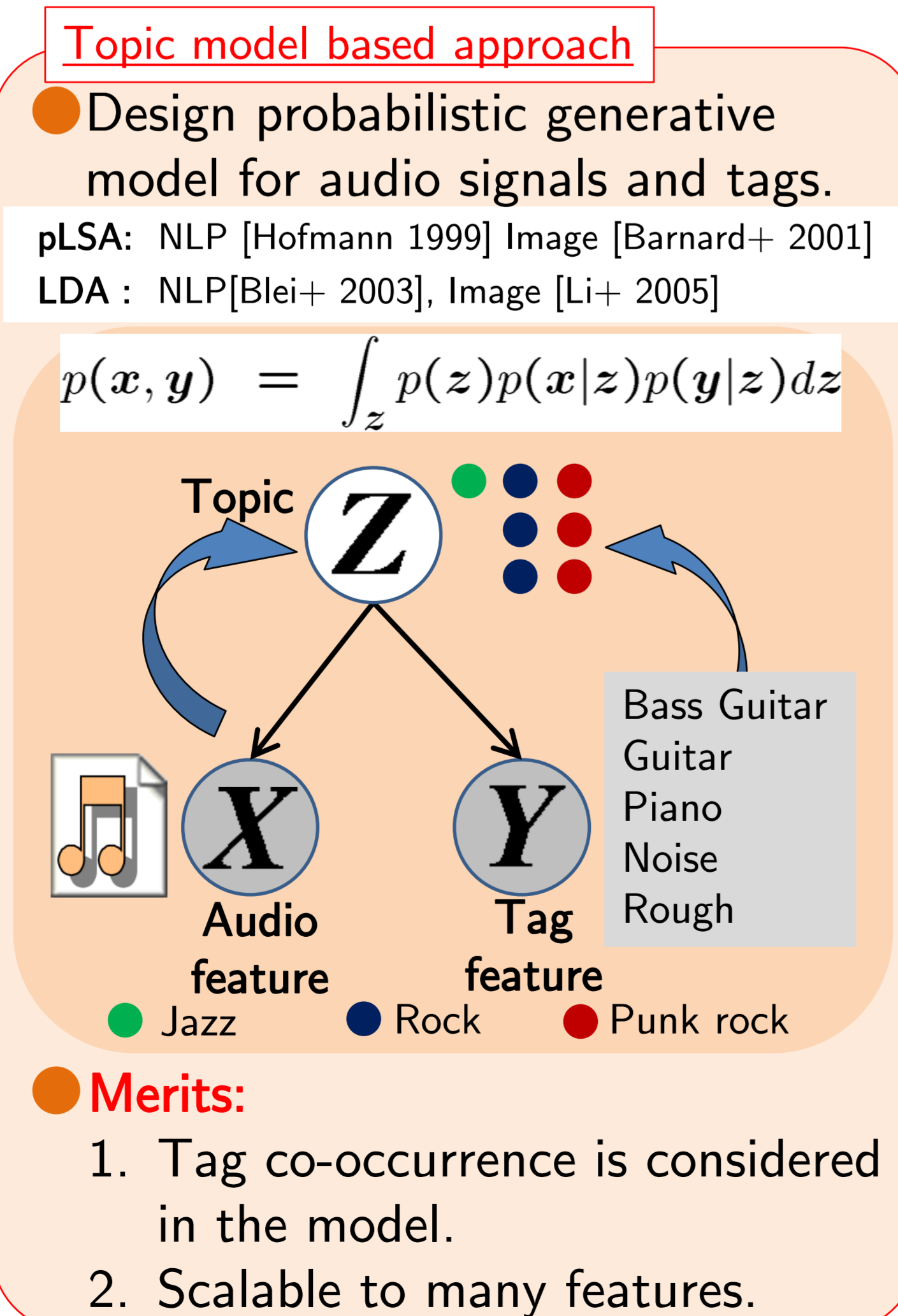
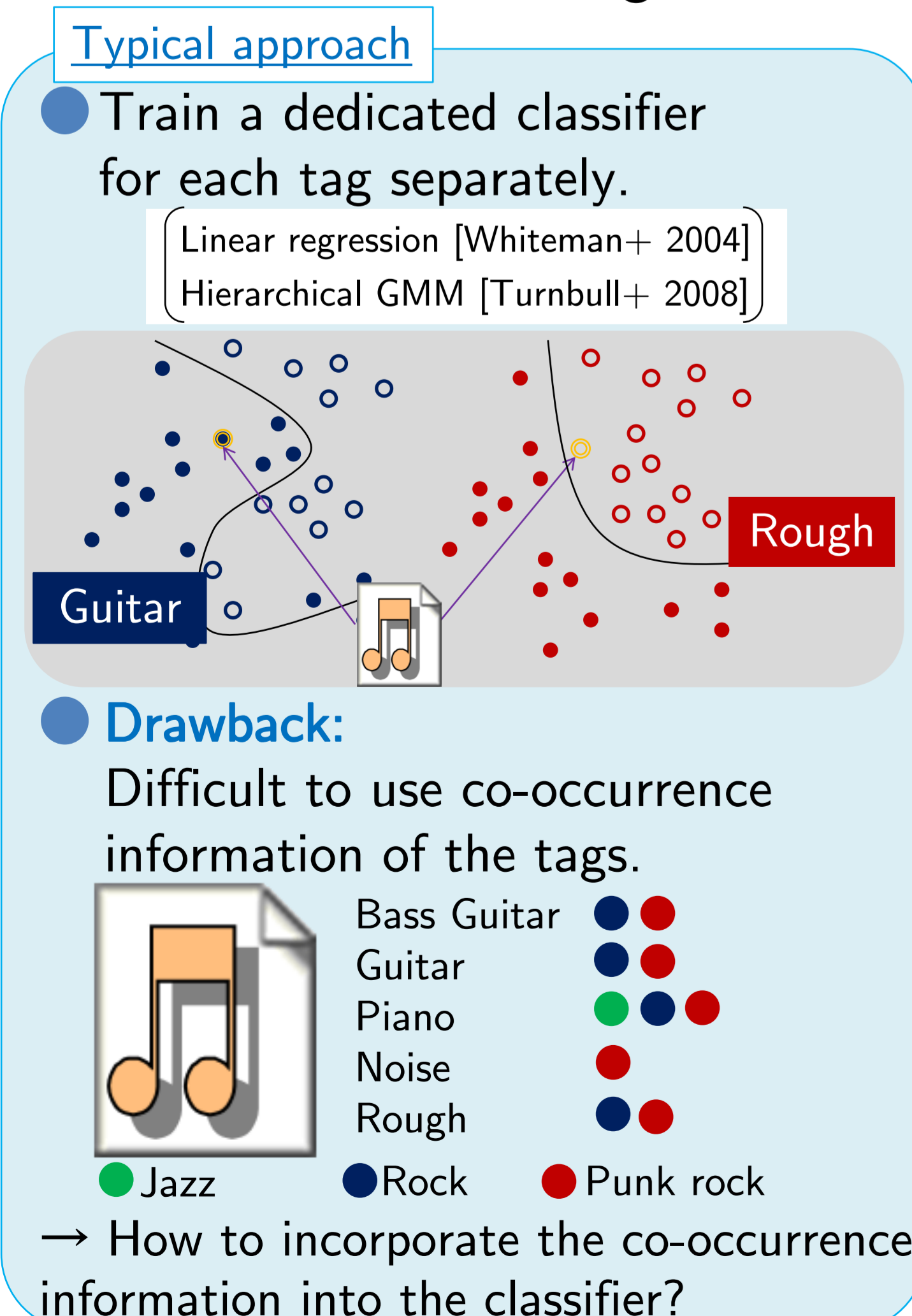
Easy to collect at low cost!



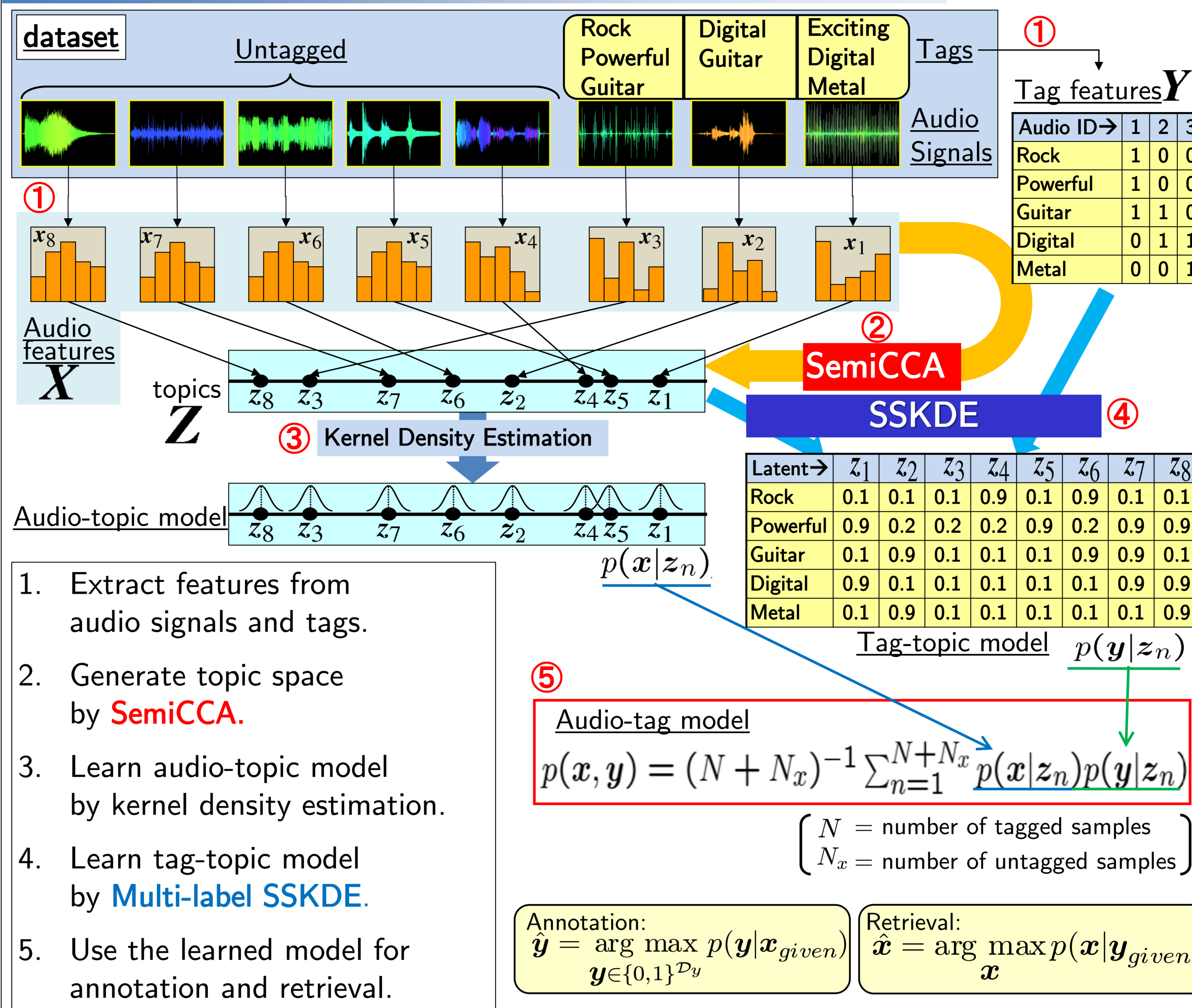
Technical Challenges : Use Tag Co-occurrence Information

Tag co-occurrence information seems to be useful for tagging/retrieval task.

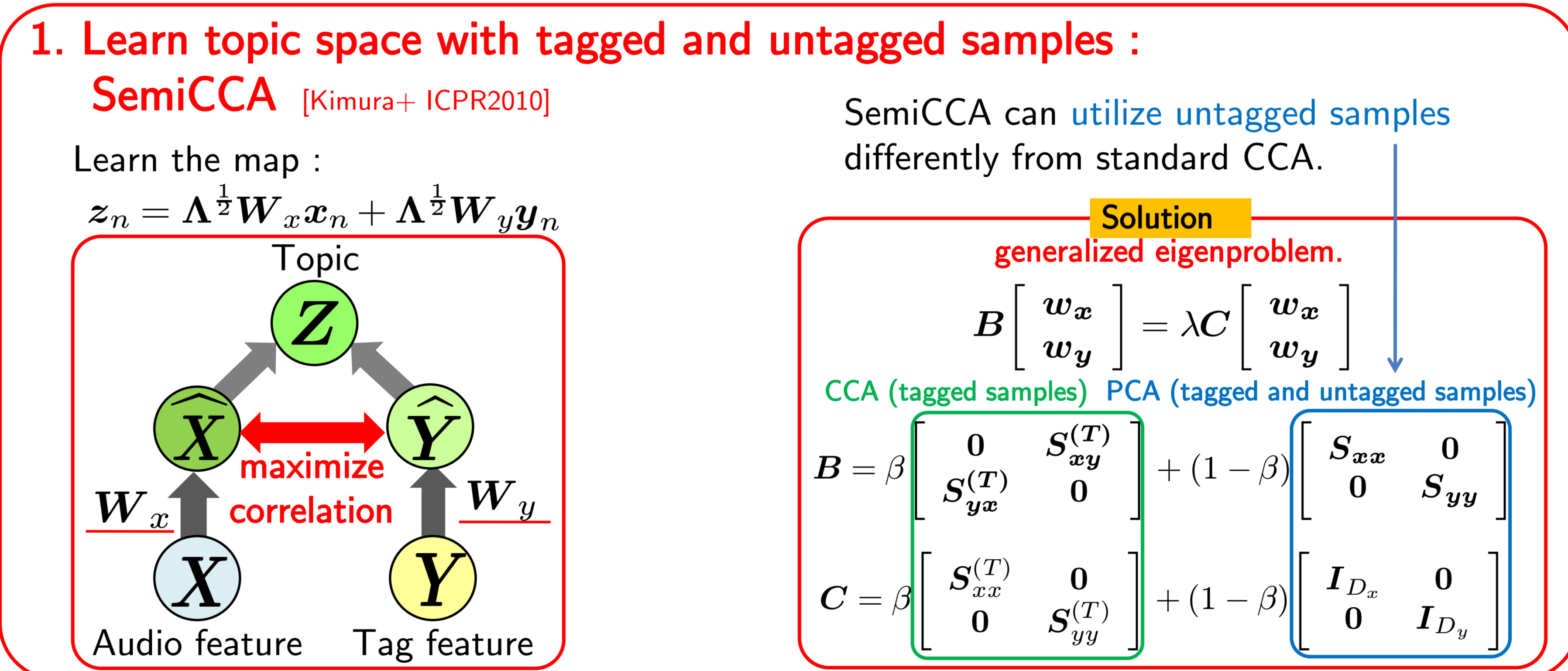
But almost all of existing method cannot utilize this information.



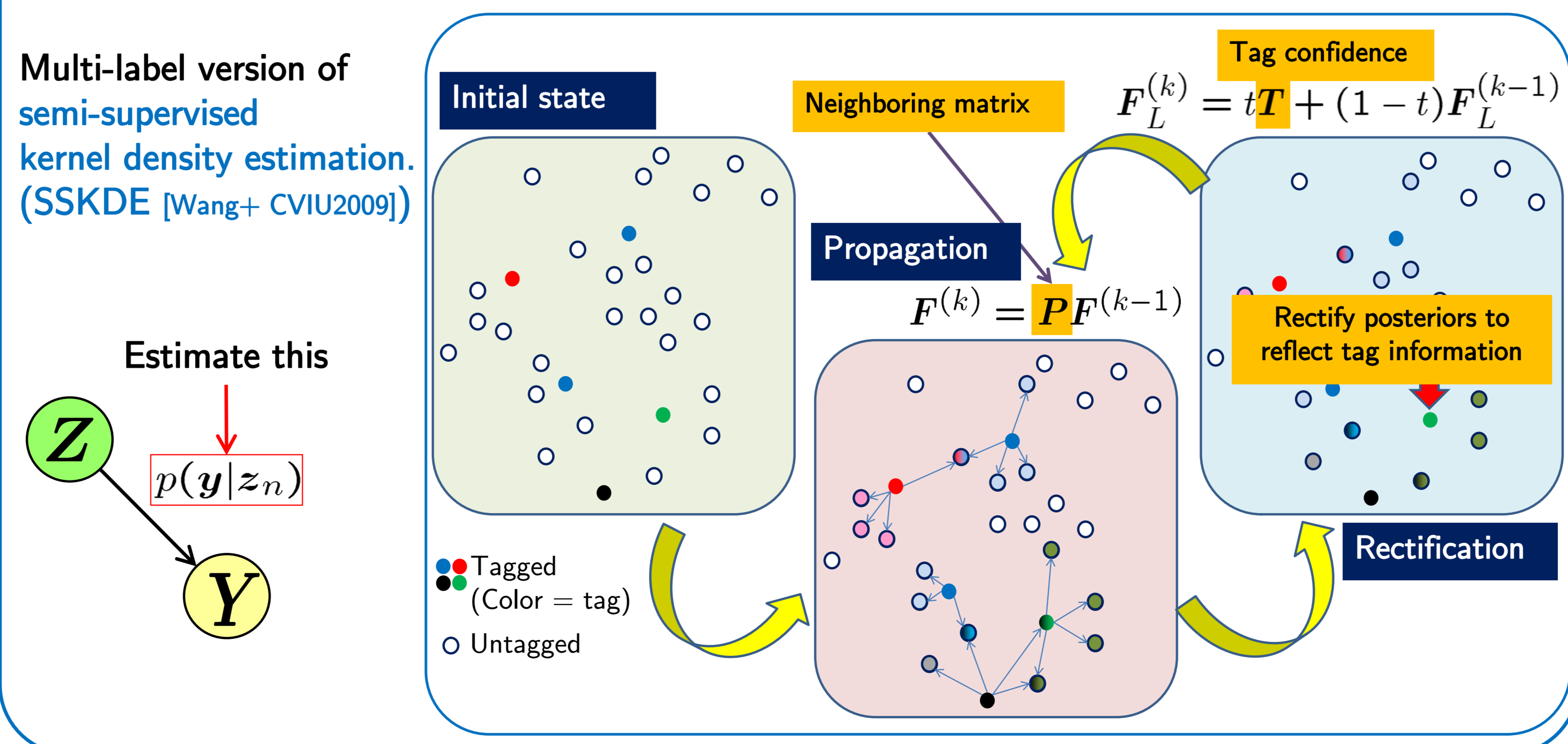
SSCDE : Model Learning Framework



Technical Points of SSCDE



2. Propagate tag information : Multi-label SSKDE



Experiment

Annotation performance of SSCDE is evaluated under following condition.

- Dataset:** 2012 audio files taken from "Freesound" (<http://www.freesound.org/>).
 - Database of Creative Commons licensed sounds.
 - Annotated with vocabulary.
- Evaluation condition:**
 - 2012 audio clips with WAV format.
 - Each clip has multiple tags selected from 230 words.

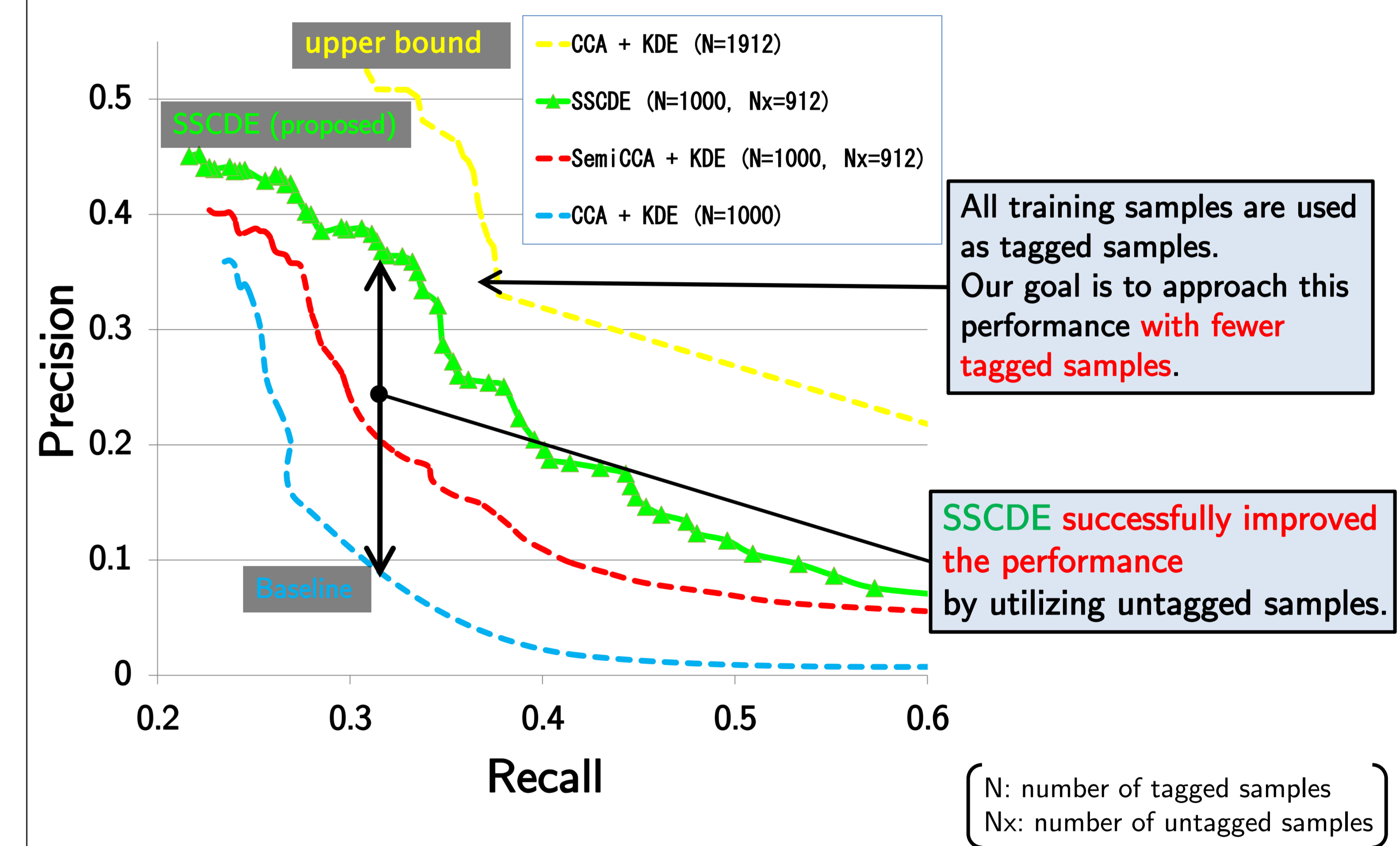
Tagged audio signal	Untagged audio signal	Evaluation sample
1000	912	100

- Audio feature:** bag-of-feature-vectors extracted by the following process.

- Audio signals are splitted into half-overlapping 23ms windows, and 39-dimensional vector (including first 13 MFCC, MFCC-Δ, MFCC-ΔΔ) is extracted from each window.
- 500 vectors are sampled from each audio signal (about 1,000,000 vectors in total). LBG algorithm (Linde-Buzo-Gray algorithm, algorithm for vector quantization : VQ) is applied to them, and VQ codebook (size :1024) is obtained.
- 1024-dimensional normalized vector representing each audio signal is created by VQ.

- Tag feature:** 230-dimensional binary vector. (Each element of the vector corresponds to specific tag.)

Result



Use Sparse Matrix as Neighboring Matrix

- Fix the number of non-zero elements in each row, then required **memory size** to hold the neighboring matrix **decreases**. ($\mathcal{O}(N^2) \rightarrow \mathcal{O}(N)$)
- In this case, SSKDE is equivalent to Graph spectral method. [Joachims 2003]
- Apply the same idea to audio tagging and retrieval, then **calculation needs only a few samples nearby query!**

