Semi-Automated Music Catalog Correction Using Audio and Metadata

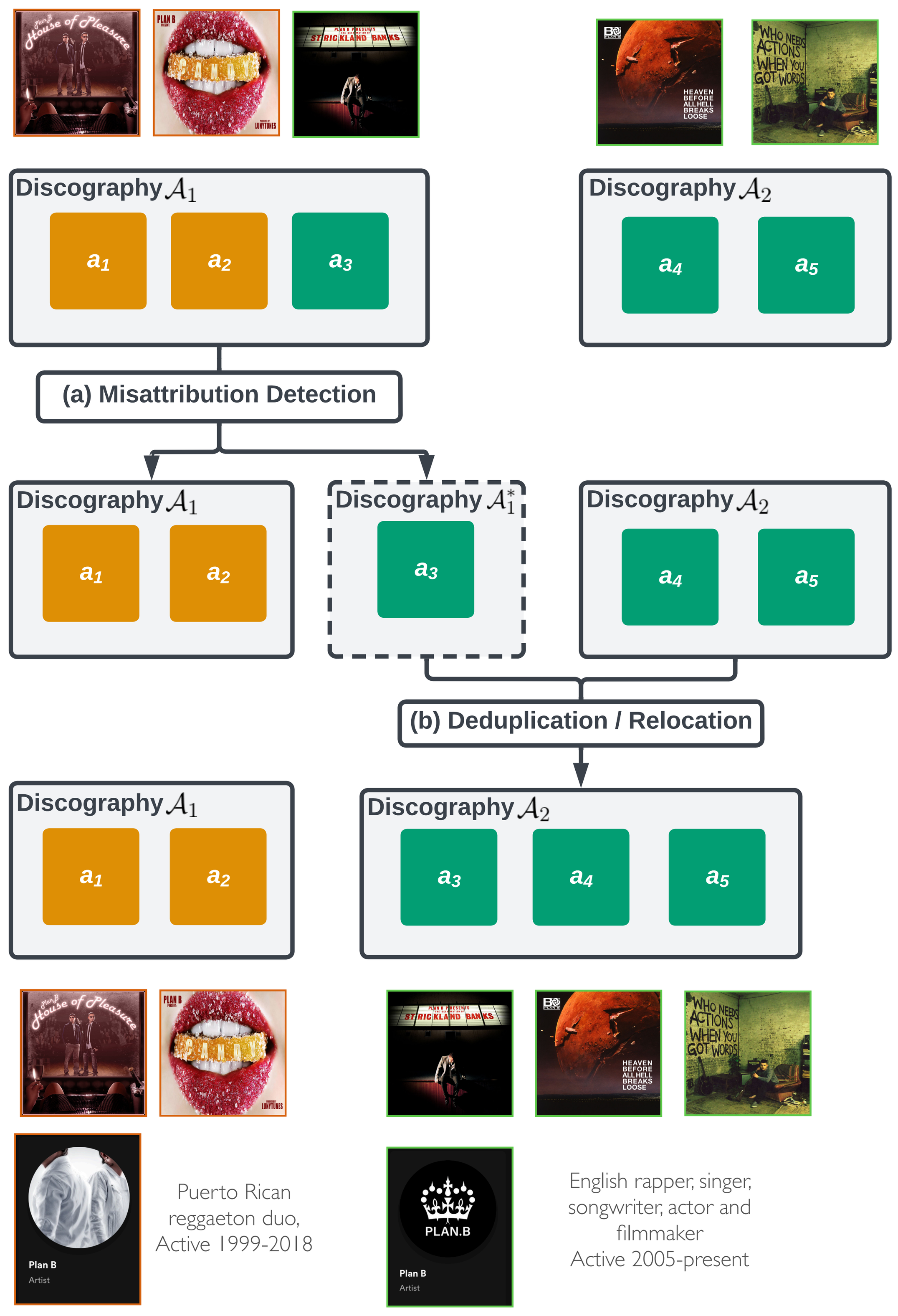
Brian Regan, Desislava Hristova, and Mariano Beguerisse-Díaz

Spotify Inc.

briann@spotify.com
desih@spotify.com
marianob@spotify.com

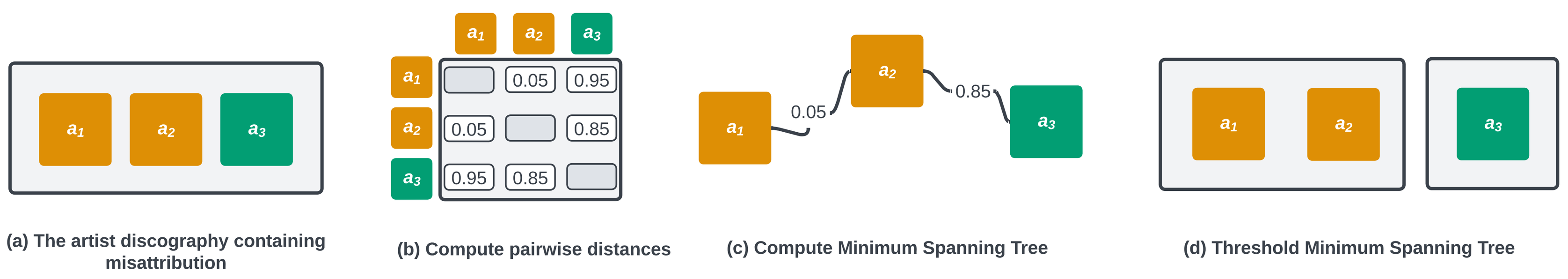
Introduction

- We present a system to assist Subject Matter Experts (SMEs) in the curation of large music catalogs.
- Online music catalogs, such as Spotify's, contain millions of releases; it is common for multiple artists to share the same name.
- In the absence of unique identifiers, it is inevitable that on rare occasions a release is incorrectly attributed (e.g. due to incomplete or incorrect metadata, extreme ambiguity, or human error).
- These errors can manifest in two different ways:
  - ***Misattribution***: when a release is incorrectly attributed to an artist, so their discography now contains releases from two separate real-world artists.
  - ***Duplication***: when a release is not attributed to the correct existing discography but to a new one, so that a single artist's work is split across two discographies.



Method

1. Misattribution detection



Pairwise model

	Attribute	Functions
Metadata	Music Label*	Exact Match*, Dice Score [2]
	Music Licensor*	Exact Match
	Music Source*	Exact Match
	Release Name	Exact Match, Dice Score
	Release Group* [3]	Exact Match
	Release Artists	Overlap, Dice Overlap [4]
	Release Track Names*	At Least 1 Exact Match, Min Dice Score
	Release Track Artists	Max Overlap, Max Dice Overlap
	Release Track Language*	At Least One Exact Match
	Release Type†*	Categorical
	Release Is Remix†	Categorical
	At Least One Track Is Remix†*	Categorical
Audio	Track Audio Vectors*	Min/Max/Mean Cosine Similarity
	Track Speechiness†	Min/Max/Mean

2. Discography deduplication

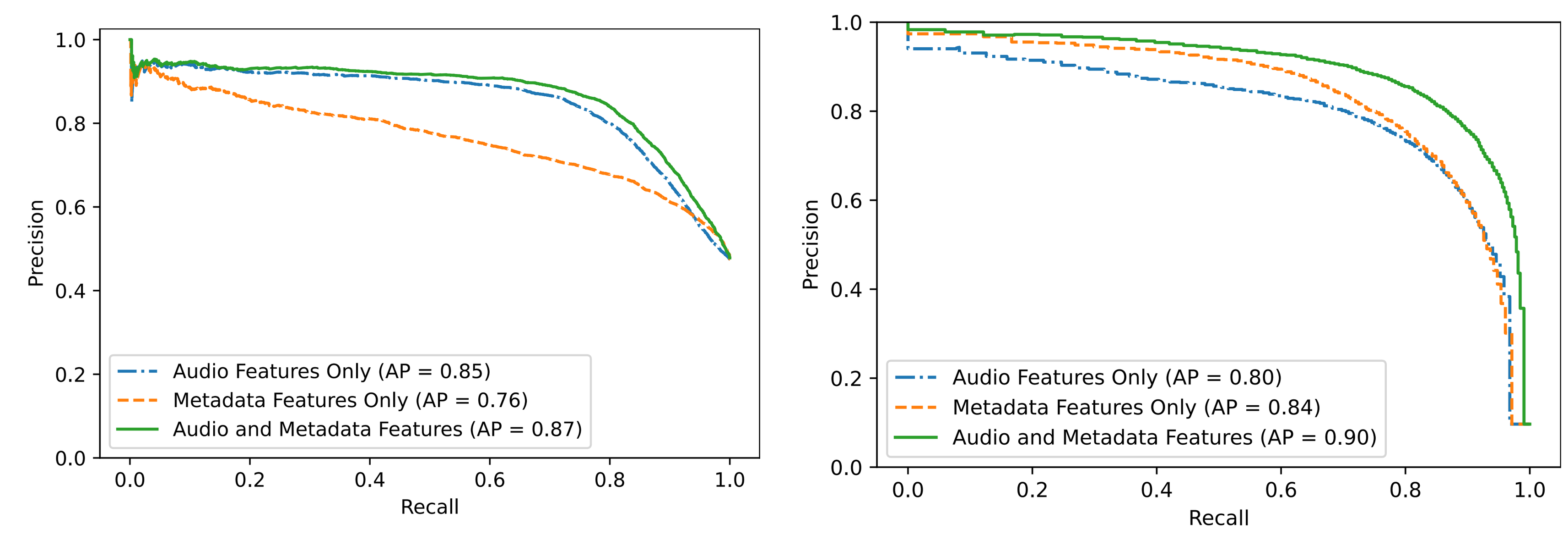
Duplicate detection model

	Attribute	Functions
Metadata	Elasticsearch relevance score	See [24]
	Artist name similarity	2-gram Dice coefficient
	Release Names	Jaccard similarity
	Release Track Names	Jaccard similarity
	Release Artists	Overlap between artist names of collaborators on releases
	Release Track Artists	Overlap between artist names of collaborators on release tracks
	Number of releases	$\lvert\mathcal{A}_i \bigcup \mathcal{A}_j\rvert$
Audio	Track Audio Vectors	Mean Cosine Similarity

- Our system consists of two machine learning sub-systems:
  - a pairwise distance model combined with a Minimum Spanning Tree that splits discographies with releases from multiple artists into their constituent sub-discographies;
  - a duplicate detection model that takes pairs of discographies or sub-discographies and decides if they should be combined.
- Both models are random forest ensemble classifiers and use a combination of features from audio and metadata.

Evaluation

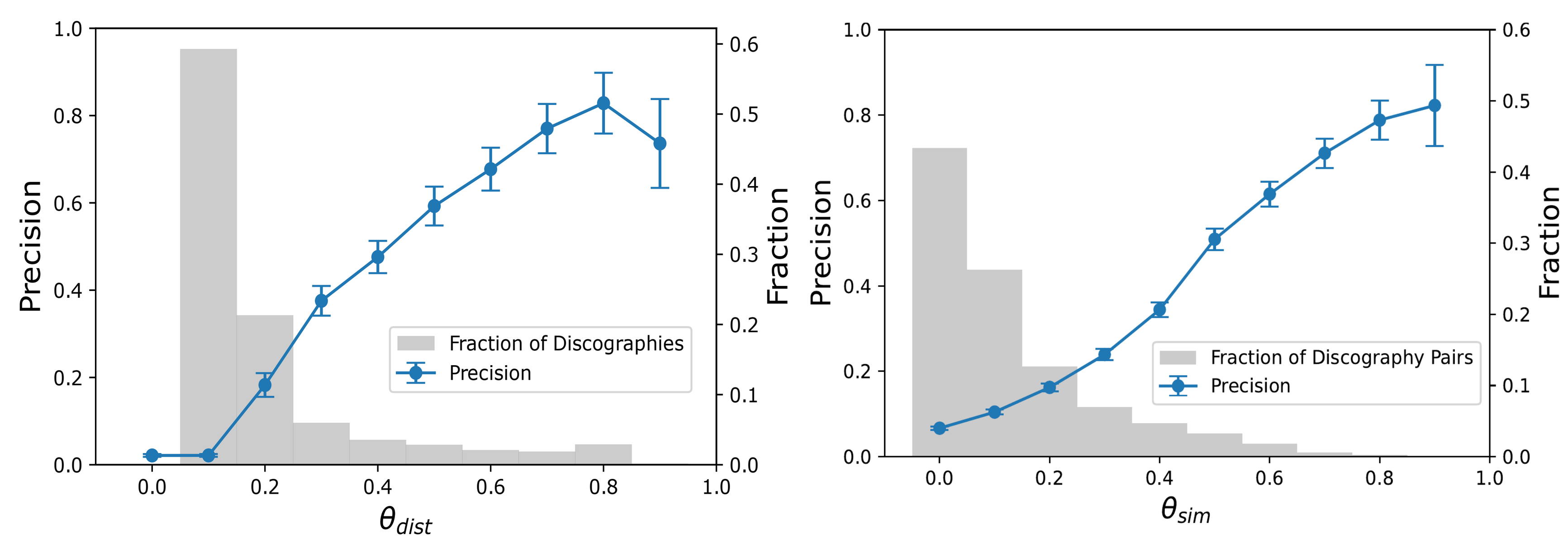
1. Audio and Metadata Feature Ablations



Misattribution Detection Pairwise Model — Duplicate detection model

- The pairwise model, a model using the audio features alone, performs well in the task of classifying albums from distinct artists. Adding metadata signals improves its average precision by 2%.
- The duplicate detection method, a model using metadata features alone, performs well in the task of identifying duplicate discographies, but adding audio features improves the average precision by 6%.

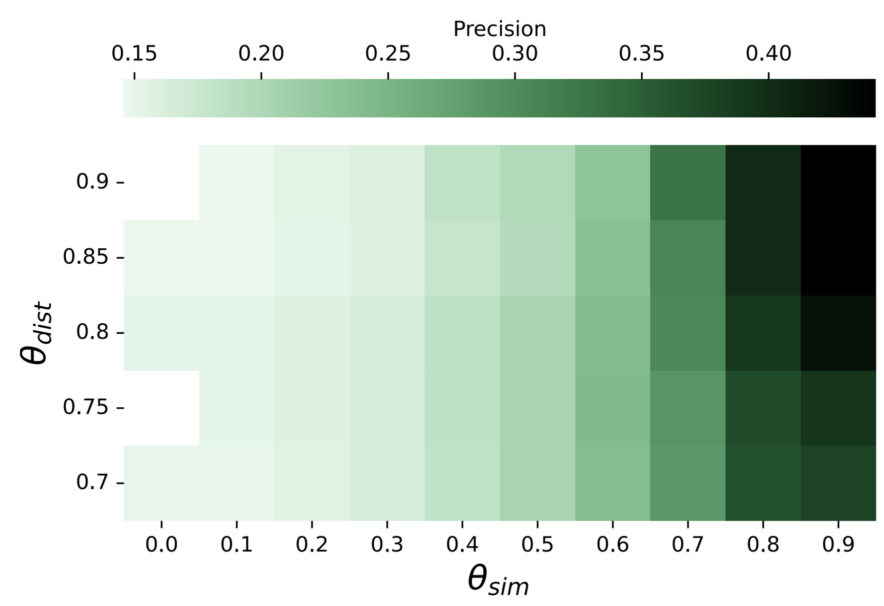
2. Experiments with Subject Matter Experts (SMEs)



Misattribution Detection — Duplicate detection model

- We sample ~1,000 pairs of releases / pairs of discographies for each respective task and asked SMEs to review the predicted misattributions and duplicates:
  - Misattribution detection: *are the two releases by the same artist or by different artists?*
  - Duplicate detection: do *the two discographies belong to the same real world artist?*
- We report the precision of detecting these misattributions/duplicates as well as the fraction of discographies in population at each score interval.

Predicted relocation



- We use the duplicate detection model to predict relocations for the misattributed releases detected by the misattribution model
- We evaluate performance on ~1,000 release-discography pairs, asking SMEs: *Does the release belong with the discography?*

- The highest precision is 45%, which is achieved when both the misattribution step and deduplication (relocation) step have a high threshold.
- The relocation task is more difficult because it inherits the performance (and uncertainties) of the misattribution and duplicate detection models. Sometimes a relocation is not possible, and creating a new discography is the correct solution.