

What is a song, and how many do I have?

*lessons learned from the
Million Song Dataset*

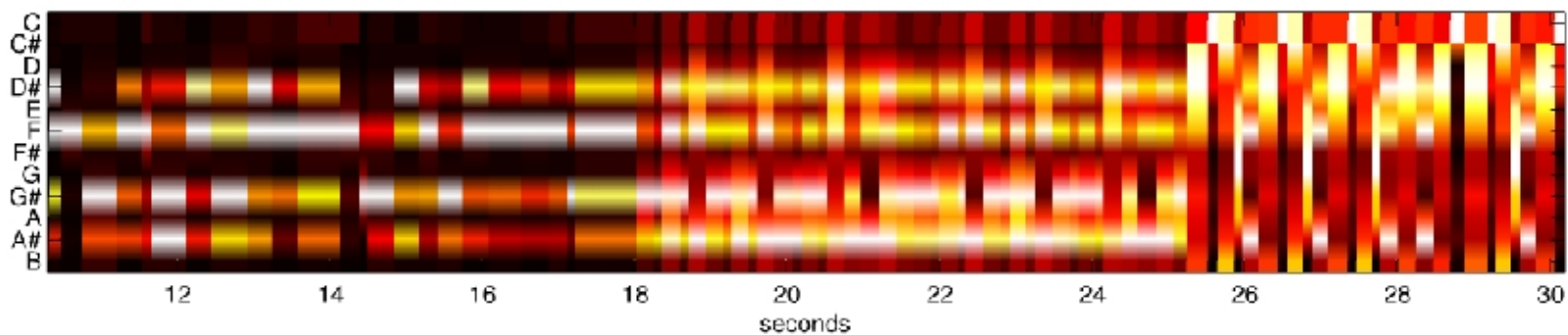
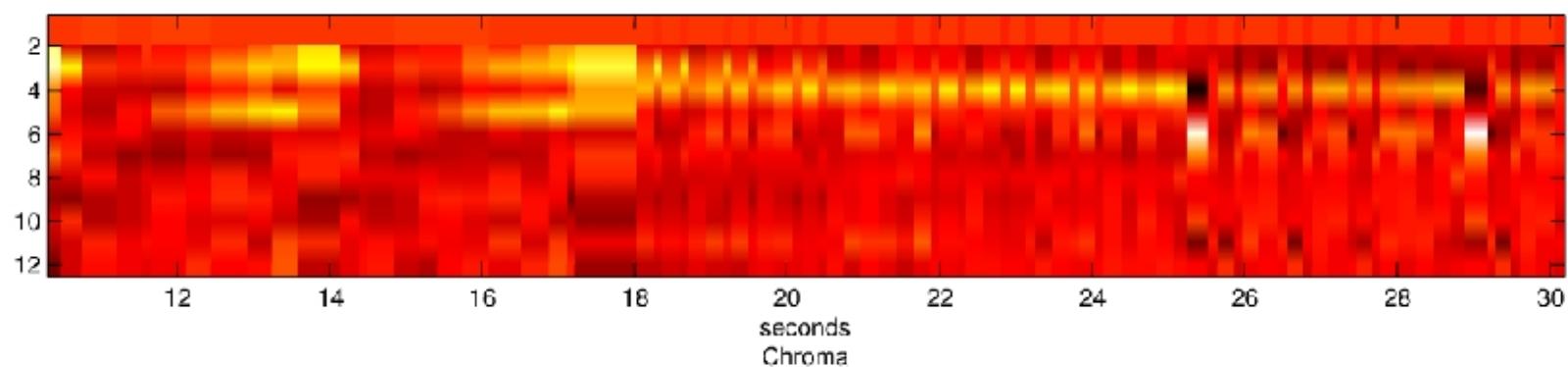
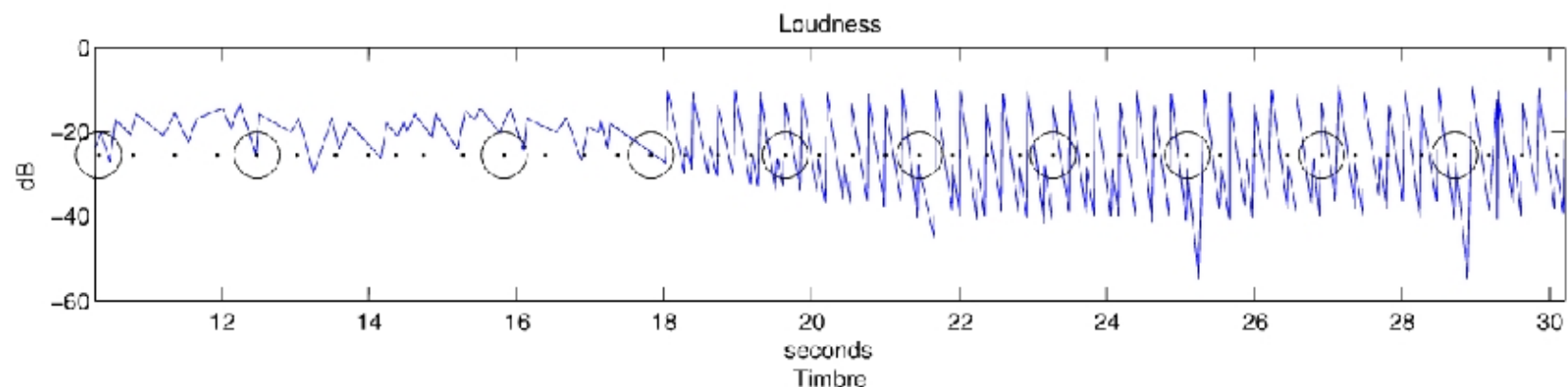


Thierry Bertin-Mahieux
PhD Candidate, Columbia U.
CIRMMT Workshop 2012
tb2322@columbia.edu

The Million Song Dataset



audio features



Awesome things to do with the MSD!

(buzzword alert!)

- co-clustering on tags, lyrics, features, ...
- 5-dimensional tensor factorization for recommendation
- describing the manifold of pop music
- year prediction / music evolution from features and years
- mood prediction using lyrics, tags, features
- large-scale tag prediction
- large-scale cover song detection
- management system for multimedia libraries
- hashing methods for music data
- ...

...but I won't talk about any of them.

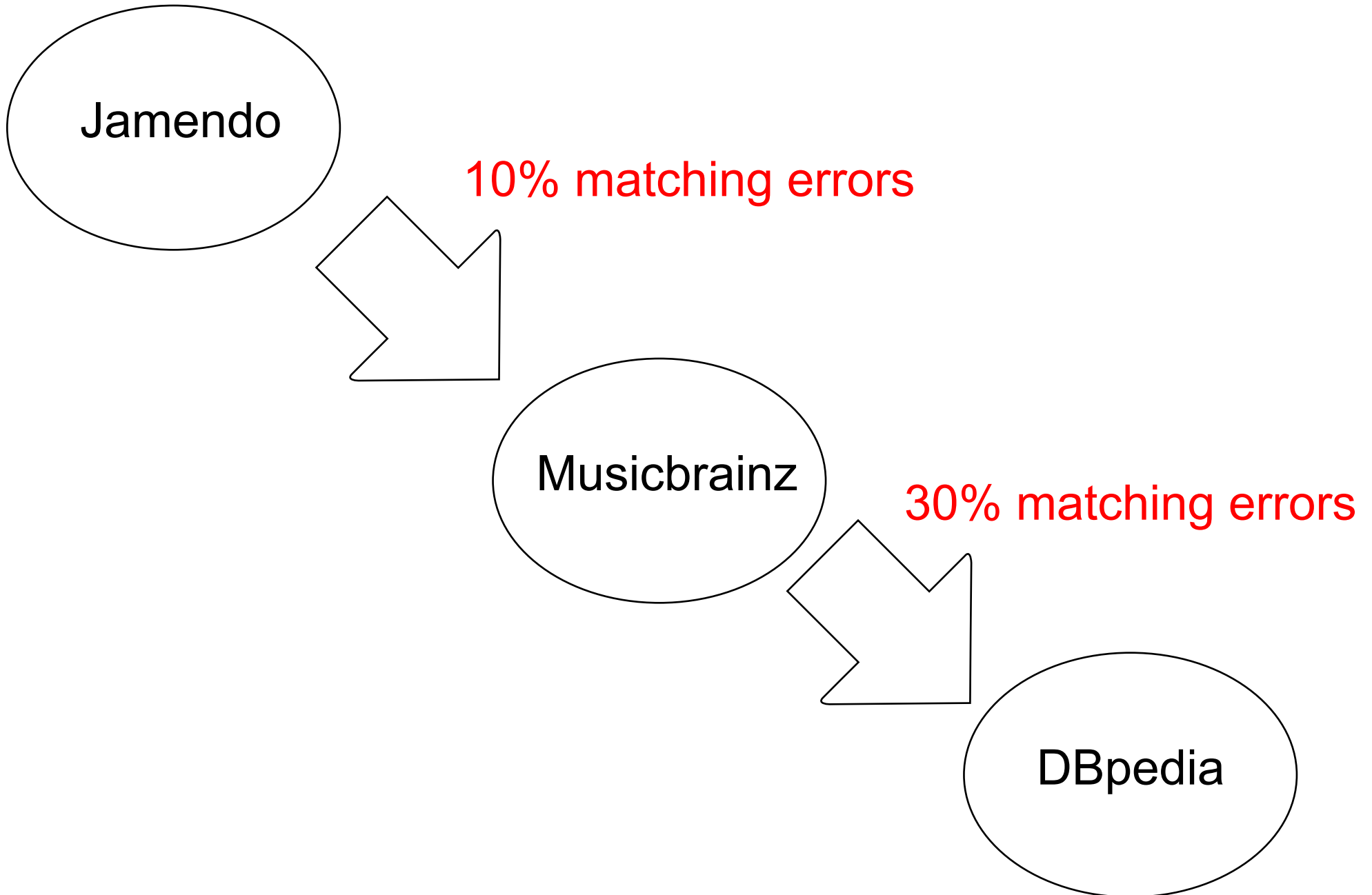
Let's talk about song matching

a MIR problem as sexy as this



Bloodhound Gang, *Hefty Fine*, 2005

Motivation: RDF

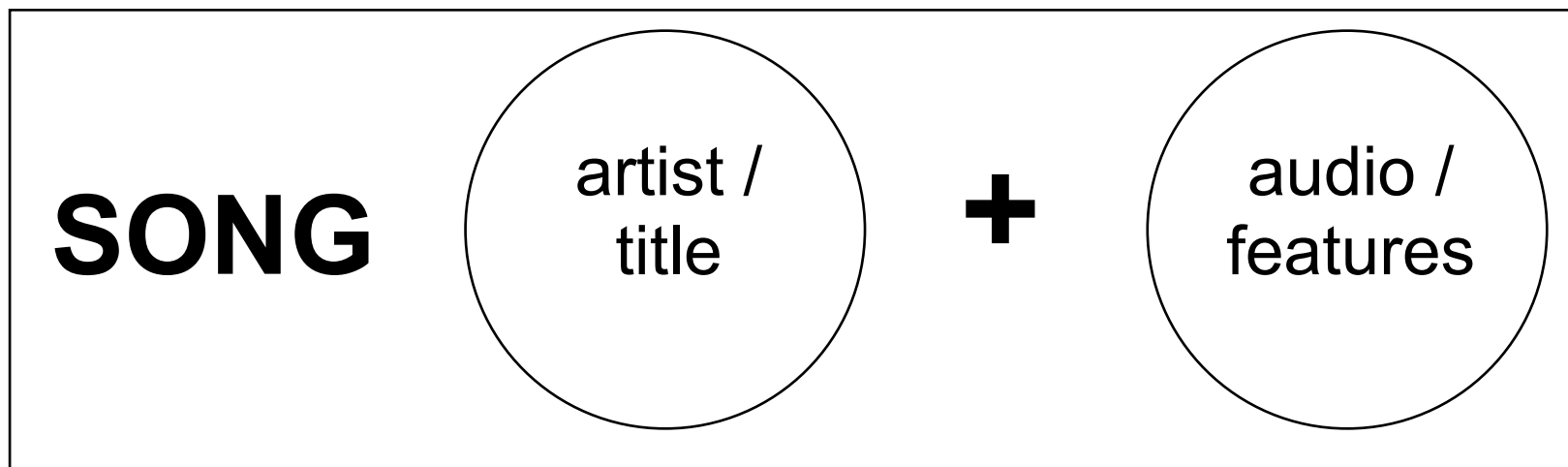


What I want to see at ISMIR / MCMC:

*Here is the definition of a song,
and here is some code to decide if
song A is the same as song B*

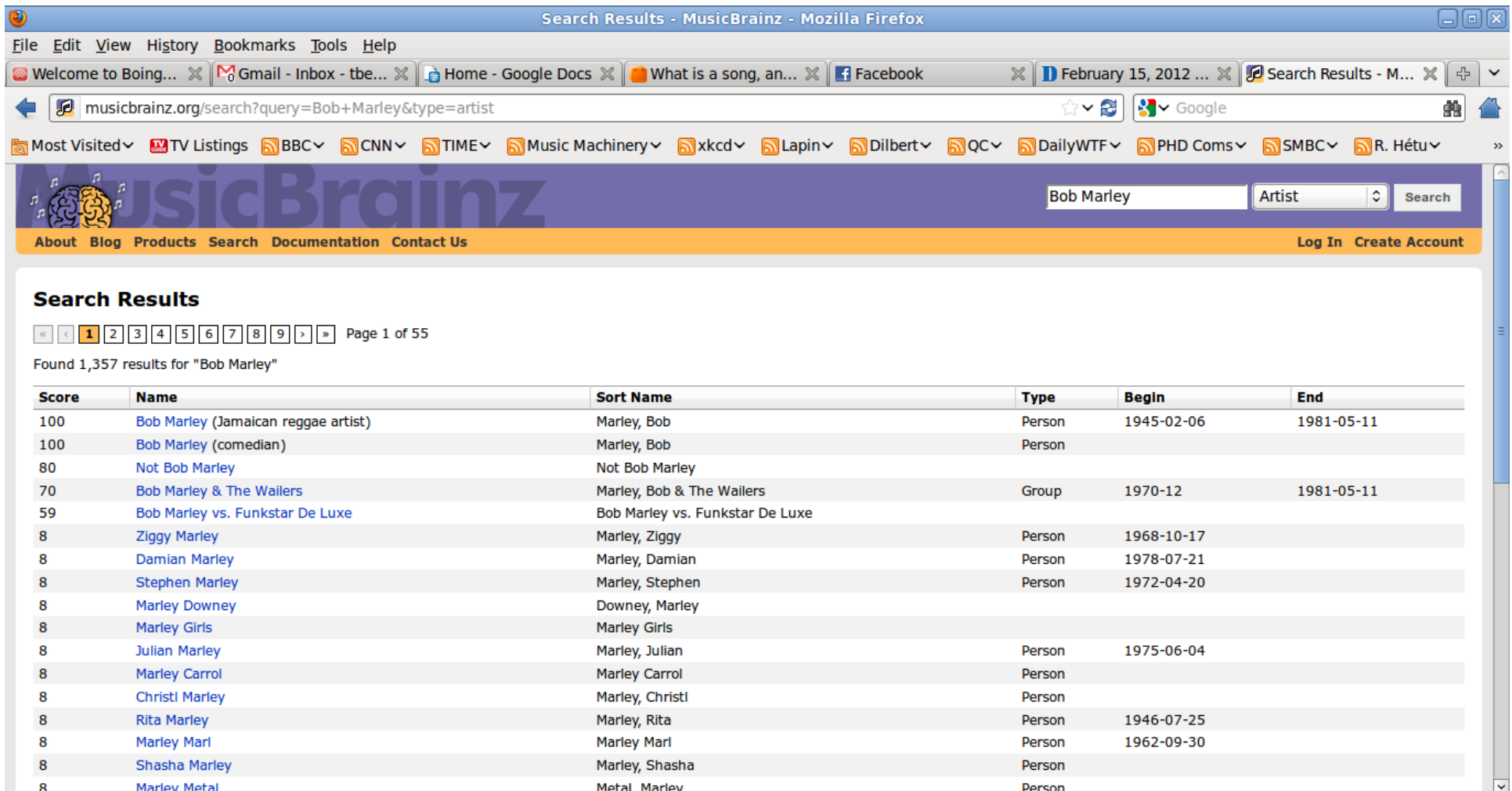
Comment 1: A song is a song. Identification is solved by Shazaam + Musicbrainz. Done.

Comment 2: what are we dealing with?



Specific Example

Matching the MSD with Musicbrainz



The screenshot shows a Mozilla Firefox browser window with the title "Search Results - MusicBrainz - Mozilla Firefox". The address bar contains the URL "musicbrainz.org/search?query=Bob+Marley&type=artist". The search results page displays a table of results for "Bob Marley".

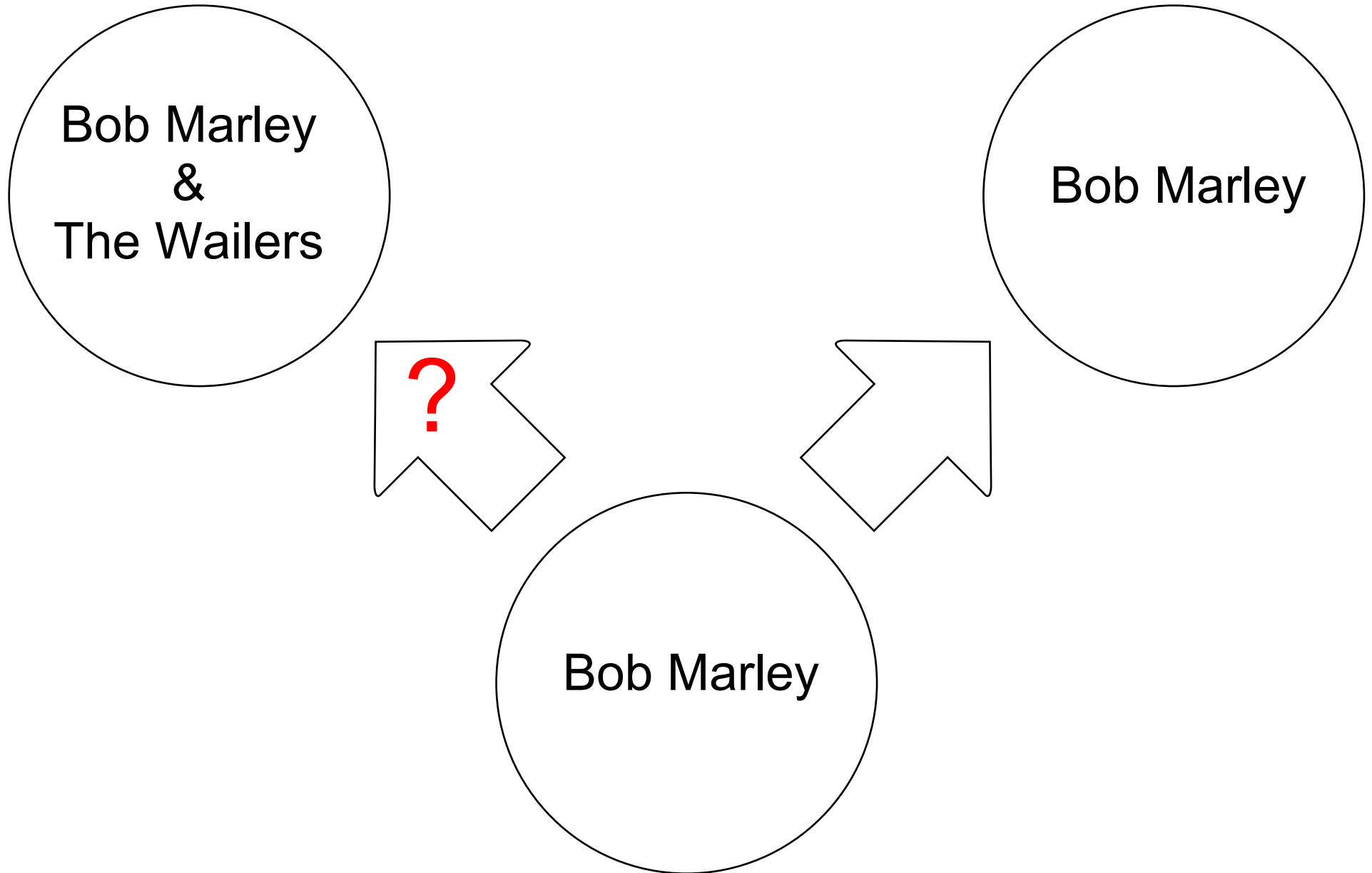
Search Results

Page 1 of 55

Found 1,357 results for "Bob Marley"

| Score | Name | Sort Name | Type | Begin | End |
|-------|---|---------------------------------|--------|------------|------------|
| 100 | Bob Marley (Jamaican reggae artist) | Marley, Bob | Person | 1945-02-06 | 1981-05-11 |
| 100 | Bob Marley (comedian) | Marley, Bob | Person | | |
| 80 | Not Bob Marley | Not Bob Marley | | | |
| 70 | Bob Marley & The Wailers | Marley, Bob & The Wailers | Group | 1970-12 | 1981-05-11 |
| 59 | Bob Marley vs. Funkstar De Luxe | Bob Marley vs. Funkstar De Luxe | | | |
| 8 | Ziggy Marley | Marley, Ziggy | Person | 1968-10-17 | |
| 8 | Damian Marley | Marley, Damian | Person | 1978-07-21 | |
| 8 | Stephen Marley | Marley, Stephen | Person | 1972-04-20 | |
| 8 | Marley Downey | Downey, Marley | | | |
| 8 | Marley Girls | Marley Girls | | | |
| 8 | Julian Marley | Marley, Julian | Person | 1975-06-04 | |
| 8 | Marley Carrol | Marley Carrol | Person | | |
| 8 | Christl Marley | Marley, Christl | Person | | |
| 8 | Rita Marley | Marley, Rita | Person | 1946-07-25 | |
| 8 | Marley Marl | Marley Marl | Person | 1962-09-30 | |
| 8 | Shasha Marley | Marley, Shasha | Person | | |
| 8 | Marley Metal | Metal, Marley | Person | | |

Specific Example



First Stab at an Implementation

Run-D.M.C. and Aerosmith - "Walk This Way"

- Aerosmith and Run-D.M.C.
- Run-D.M.C.
- Aerosmith
- Run-D.M.C. vs. Aerosmith
- Run DMC / Aerosmith
- Run DMC
- run dmc
- Ron Dmc and Areosmit
- Run-DMC and Aerosmith ~**!AWESOME!**~ (LUV)
- Run-D.M.C feat. Steven Tyler
- Simons/McDaniels/Mizell/Tyler/Perry

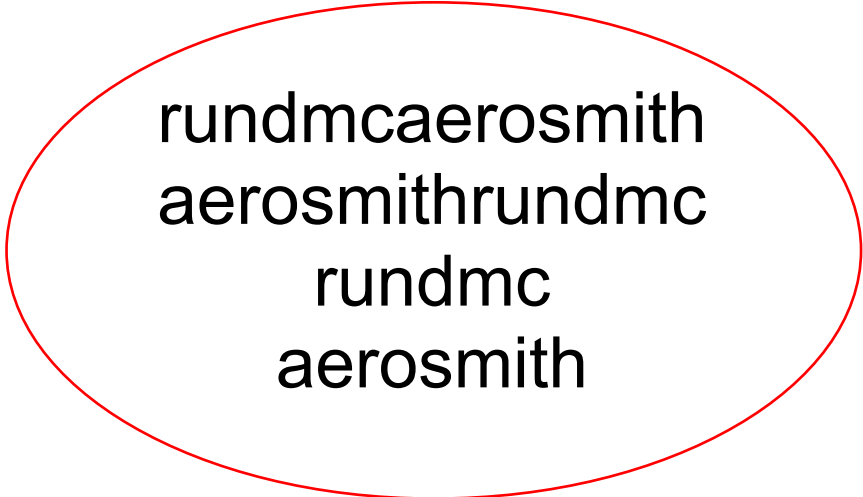
(need aliases)

First Stab at an Implementation

Possible method: generate set of possible "standard names":

Run-D.M.C. and Aerosmith

=>



rundmcaerosmith
aerosmithrundmc
rundmc
aerosmith

Weakness: Classical Music

*Les Choeurs de l'Armée Rouge -
Les Rossignols
The Red Army Choir -
Nightingales*

*Pablo Casals -
Sinfonia Concertante for Violin, Viola and Orchestra in E-flat Major, KV. 364/III. Presto
Isaac Stern_ William Primrose -
Sinfonia Cencertante in E Major_ K. 364: III. Presto*

Matching is Imperfect, Period.

- Assuming name matching is 90% correct
- Assuming audio fingerprinting is 95% correct
- Assuming 10M songs

BEST CASE: ~ 50K errors

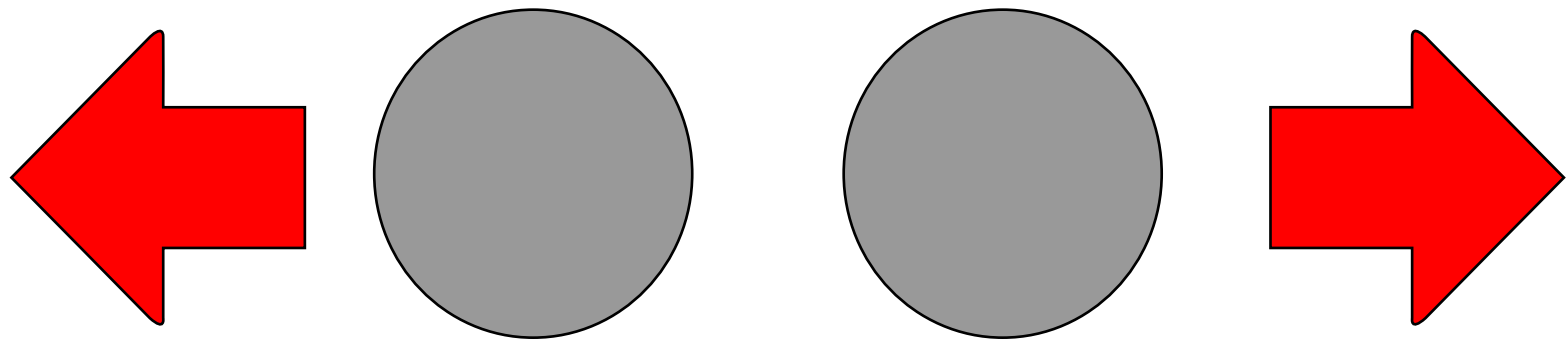
WORST CASE: ~ 500K errors

So we need to:

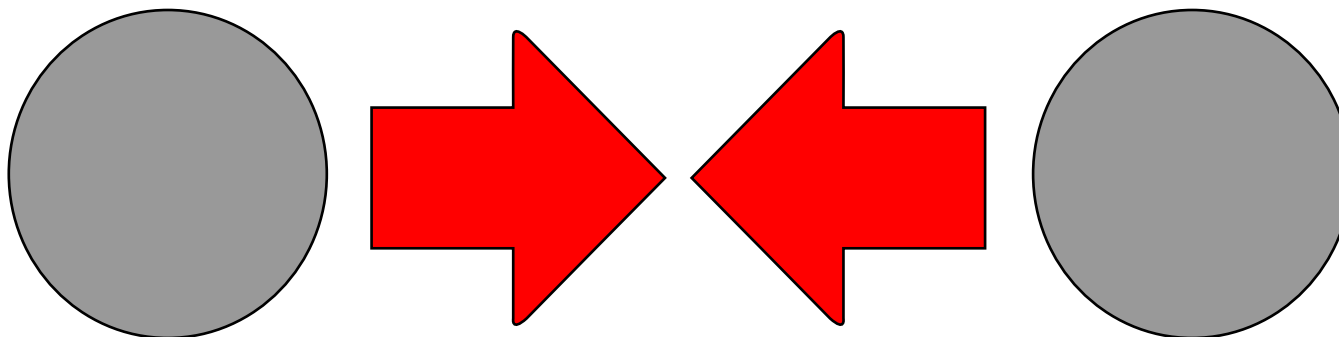
- improve both matching algorithms
- **deal with the noise**

Matching is a trade-off!

under-merge



over-merge



Matching is a trade-off!

If you're an online radio, start merging!

The screenshot shows a web browser window with the title "Air - Playground Love". The interface includes a menu bar (File, View, Tools, Controls, Account, Help) and a toolbar with icons for My Profile, Share, Tag, Playlist, Love, Ban, Stop, and Skip. A progress bar at the top right shows the track "Air - Playground Love" at 2:52, with a station name "Station: Air Radio" and a duration of 0:39.

The main content area displays the album cover for "Playground Love by Air" (The Virgin Suicides soundtrack). Below the cover, it lists the album title, release date (28 Feb 2000), and total tracks (16). A red circle highlights the artist name "Air" and its associated information:

Air
66,812,480 plays scrobbled on Last.fm

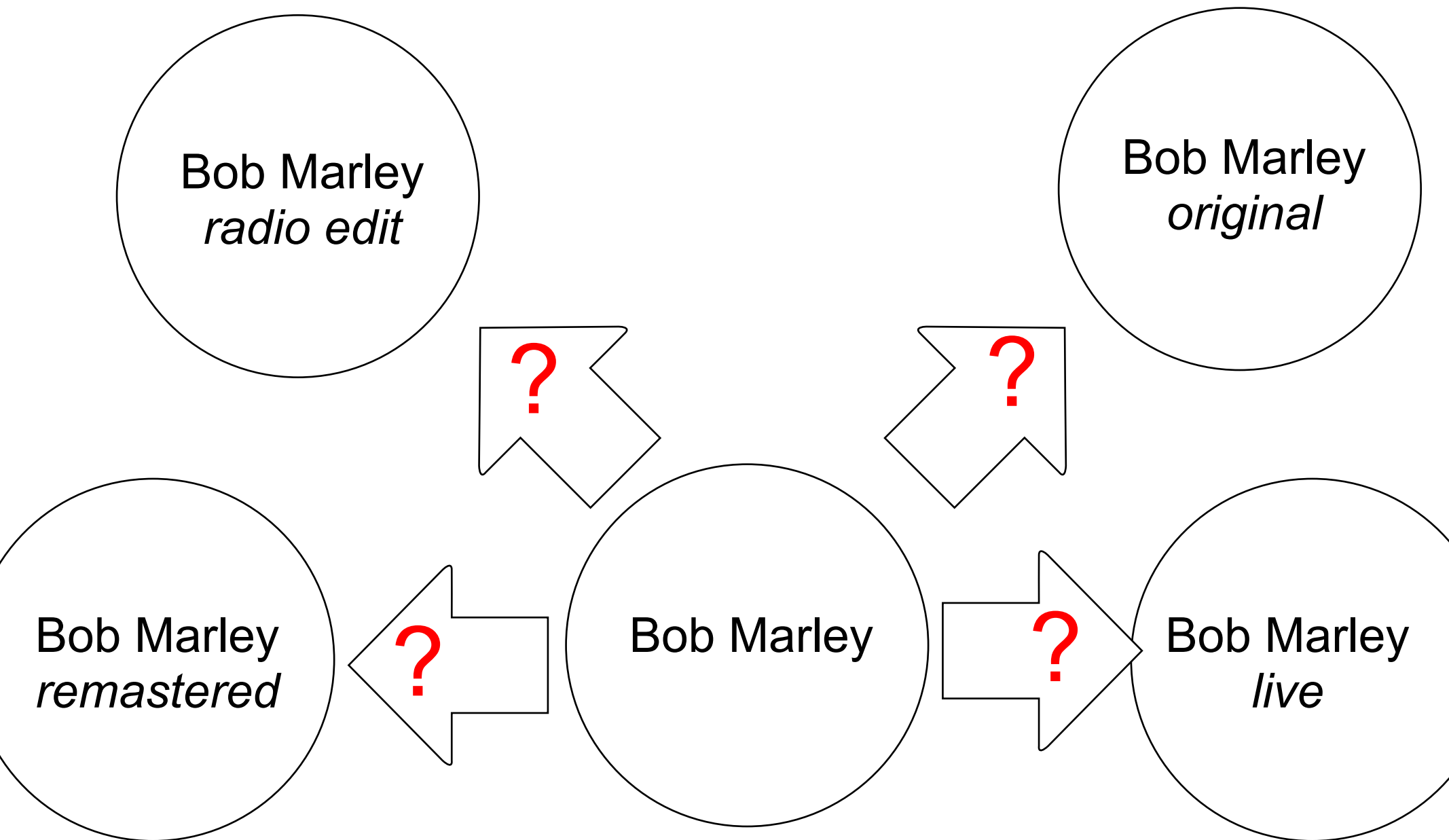
There are at least six artists with this name:

- 1) The French band Air is a duo consisting of Nicolas Godin and Jean-Benoit Dunckel. They went to school in Versailles, (Lycée Jules Ferry) before forming the band in 1995. Their critically acclaimed first album *Premiers Symptômes* was followed by the release of *Moon Safari*, *The Virgin Suicides* (soundtrack), *10,000Hz Legend*, and *Talkie Walkie*. In 2007, Air released the album *Pocket Symphony* and *Love 2* in 2009. [Read more...](#)

The left sidebar contains navigation links for "My Music Profile", "Start a Station", "Now Playing", "My Stations", "My Profile", and "History".

Matching is a trade-off!

If you're Shazaam, go ahead and have duplicates!



Matching is a trade-off!

If we match the MSD with Musicbrainz... take a deep breath?

- if we under-merge, we might miss connections with other resources
- if we over-merge, we don't know which artist is which anymore



Finally,
how many songs do I have?

(in the MSD)

between 500K and 850K

(and someone should publish tighter bounds)

Take-Home Message

- matching is under-studied
 - especially metadata + fingerprinting
- many trade-offs to analyze
- someone should implement something to get started!

Acknowledgements

- *Dan Ellis (LabROSA)*
- *The Echo Nest*
- *musiXmatch, SecondHandSongs, Last.fm*