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

*lessons learned from the Million Song Dataset*
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

- Jamendo

  10% matching errors

- Musicbrainz

  30% matching errors

- DBpedia
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?

![Diagram](attachment://song_diagram.png)
Specific Example

Matching the MSD with Musicbrainz

<table>
<thead>
<tr>
<th>Score</th>
<th>Name</th>
<th>Sort Name</th>
<th>Type</th>
<th>Begin</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>Bob Marley (Jamaican reggae artist)</td>
<td>Marley, Bob</td>
<td>Person</td>
<td>1945-02-06</td>
<td>1981-05-11</td>
</tr>
<tr>
<td>100</td>
<td>Bob Marley (comedian)</td>
<td>Marley, Bob</td>
<td>Person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>Bob Marley</td>
<td>Marley, Bob</td>
<td>Person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Not Bob Marley</td>
<td>Not Bob Marley</td>
<td>Person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>Bob Marley vs. Funkstar De Luxe</td>
<td>Bob Marley vs. Funkstar De Luxe</td>
<td>Person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Ziggy Marley</td>
<td>Morley, Ziggy</td>
<td>Person</td>
<td>1968-10-17</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Damian Marley</td>
<td>Morley, Damian</td>
<td>Person</td>
<td>1978-10-21</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Stephen Marley</td>
<td>Morley, Stephen</td>
<td>Person</td>
<td>1972-04-20</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Marley Downey</td>
<td>Downey, Marley</td>
<td>Person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Merley Girls</td>
<td>Merley Girls</td>
<td>Person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Julian Marley</td>
<td>Morley, Julian</td>
<td>Person</td>
<td>1975-06-04</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Marley Carrol</td>
<td>Marley Carrol</td>
<td>Person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Christi Marley</td>
<td>Marley, Christi</td>
<td>Person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Rite Marley</td>
<td>Morley, Rite</td>
<td>Person</td>
<td>1946-07-25</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Merley Mot</td>
<td>Morley Mot</td>
<td>Person</td>
<td>1962-09-30</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Shaasha Marley</td>
<td>Marley, Shaasha</td>
<td>Person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Merley Metal</td>
<td>Metal, Marley</td>
<td>Person</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Specific Example

Bob Marley & The Wailers

Bob Marley

Bob Marley

?
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 Concertante 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!
Matching is a trade-off!

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

Bob Marley
*radio edit*

Bob Marley
*original*

Bob Marley
*remastered*

Bob Marley
*live*
Matching is a trade-off!

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

- 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