

# one million songs



**Million Song Dataset**

Overview

**A Brief History**

Timeline of the dataset's development:

- 2009: Initial release
- 2010: Expansion
- 2011: Further growth
- 2012: Continued expansion
- 2013: Latest update

## The Million Song Dataset

WHAT YOU NEED TO REMEMBER FROM THIS TALK



**Future Improvements**

- More sources of data
- More metadata
- More audio quality
- More consistent metadata

**Some discussions we'd like to have**

- tasks
- other communities
- releasing data

**Some specifics / Getting started**

- resources
- original data - HDF5

**Work on the MSD at ISMIR**

- Large-scale music similarity search with spatial trees, McFee and Lanckriet, Tuesday 12pm - 1:30pm
- The natural language of playlists, McFee and Lanckriet, Wednesday 3:30pm - 4:30pm
- The Million Song Dataset, Bertin-Mahieux and Ellis, Wednesday 4:40pm-5pm
- Audio-based Music Classification with a pretrained convolutional network, Dileman, Breakel and Schirwen, Thursday 3:30pm-5pm

**ISSUES WITH THE MSD**

- DUPLICATES
- FORMAT

**BREAK! (15 min)**

Break quiz: If you have a recording A, which of the following are "the same song"?

- re-release of A
- radio edit, slightly different duration
- radio edit, some words changed
- re-mastering of A
- re-mastering of A, slightly diff. duration
- re-mastering of A, some more backvocals
- remix of A
- ...

<http://labrosa.ee.columbia.edu/millionsong/ismir2011>

# Million Song Dataset

ISMIR 2011 tutorial

by Thierry Bertin-Mahieux  
Matt Hoffman  
Dan Ellis

<http://labrosa.ee.columbia.edu/millionsong/ismir2011>

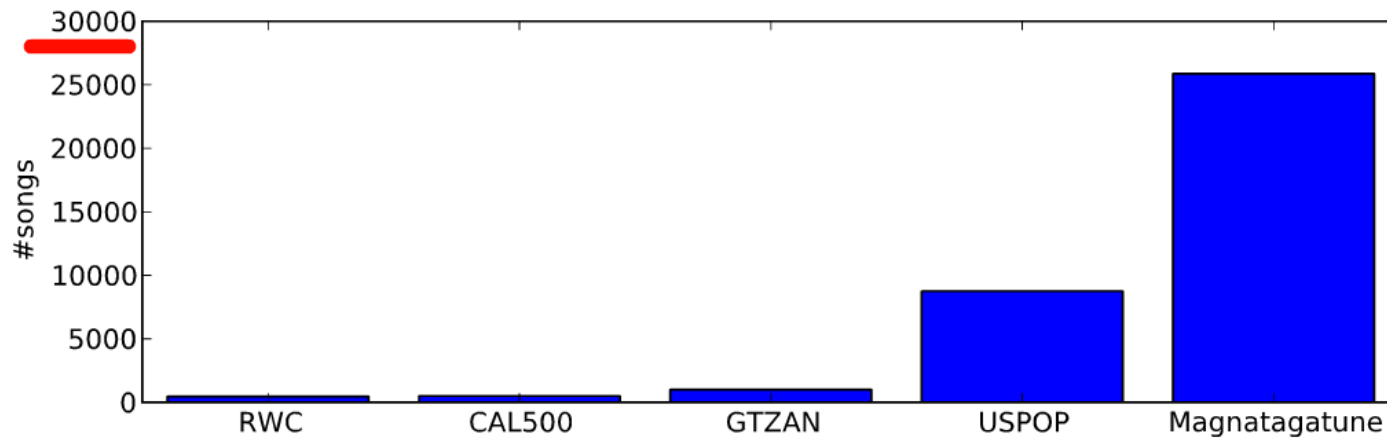
# Overview

- Origin and goals, brief history
- What data is available?
- Quick examples, some code, some issues

(break)

- Practical demonstrations

# Existing Datasets



Commercial

- Spotify: 15M
- iTunes: 14M
- Last.fm: 12M
- MOG: 12M
- RDIO: 9M



Magnatagatune

# one million songs



### Million Song Dataset

ISMRIR 2008 tutorial

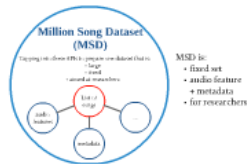
#### Overview

- High quality, big, free!
- Open access, public!
- Open research, new data, new ideas!

Music!

#### A Brief History

The Music Genome Project → The Million Song Dataset



### Some discussion tasks

and random labels

- ### Work on the M
- Large-scale music similarity trees, McFee and Lanckriet
  - The natural language of music, Lanckriet, Wednesday 2
  - The Million Song Dataset, Wednesday 4:40pm-5pm
  - Audio-based Music Classification, convolutional network,

WHAT YOU NEED TO REMEMBER FROM THIS TALK

Th

rele  
d

# Existing Web Resources - avoiding copyrights

compute your  
audio features

NEMA

MusiCLEF

access audio  
features

Echo Nest

Semantic  
Web

other kinds of data

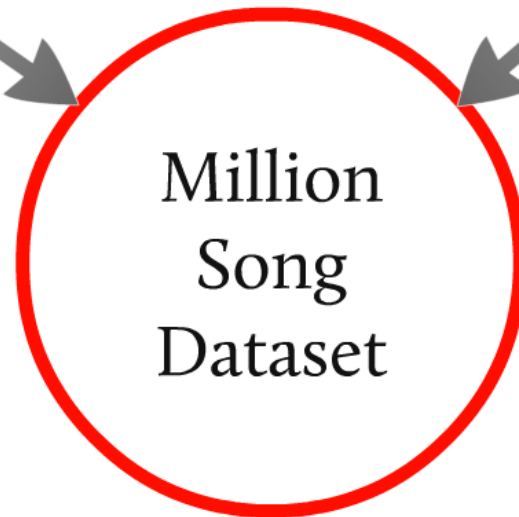
MusicBrainz

musiXmatch

Last.fm

Musicmetric





**WAIT!**  
**Aren't APIs better?**

Different goals!

- APIs give small amount of data, on demand, is maintained (may change)
- MSD is a frozen dataset, all data available (no latency), fixed so results can be reproduced

# **WAIT!**

## **Aren't APIs better?**

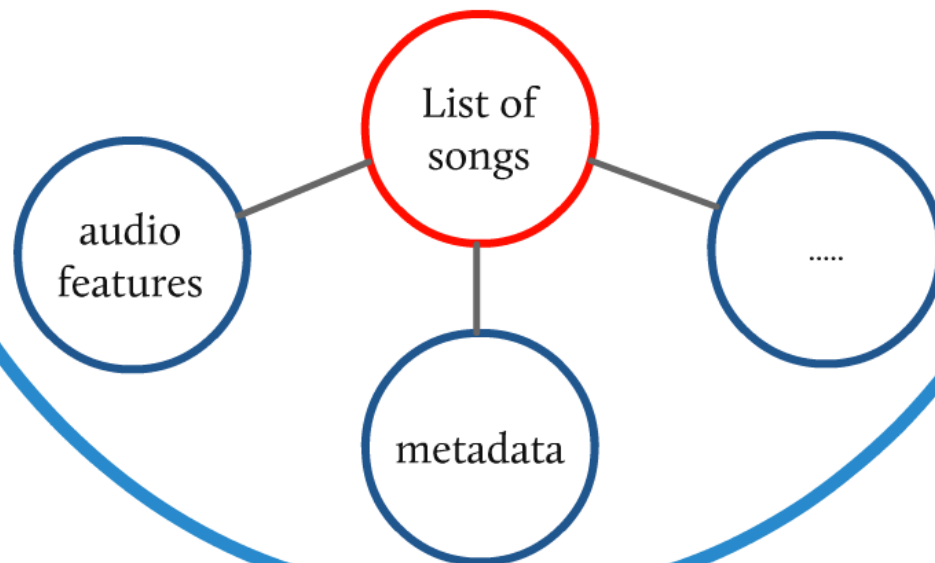
Different goals!

- APIs give small amount of data, on demand, is maintained (may change)
- MSD is a frozen dataset, all data available (no latency), fixed so results can be reproduced

# Million Song Dataset (MSD)

Tapping into these APIs to prepare one dataset that is:

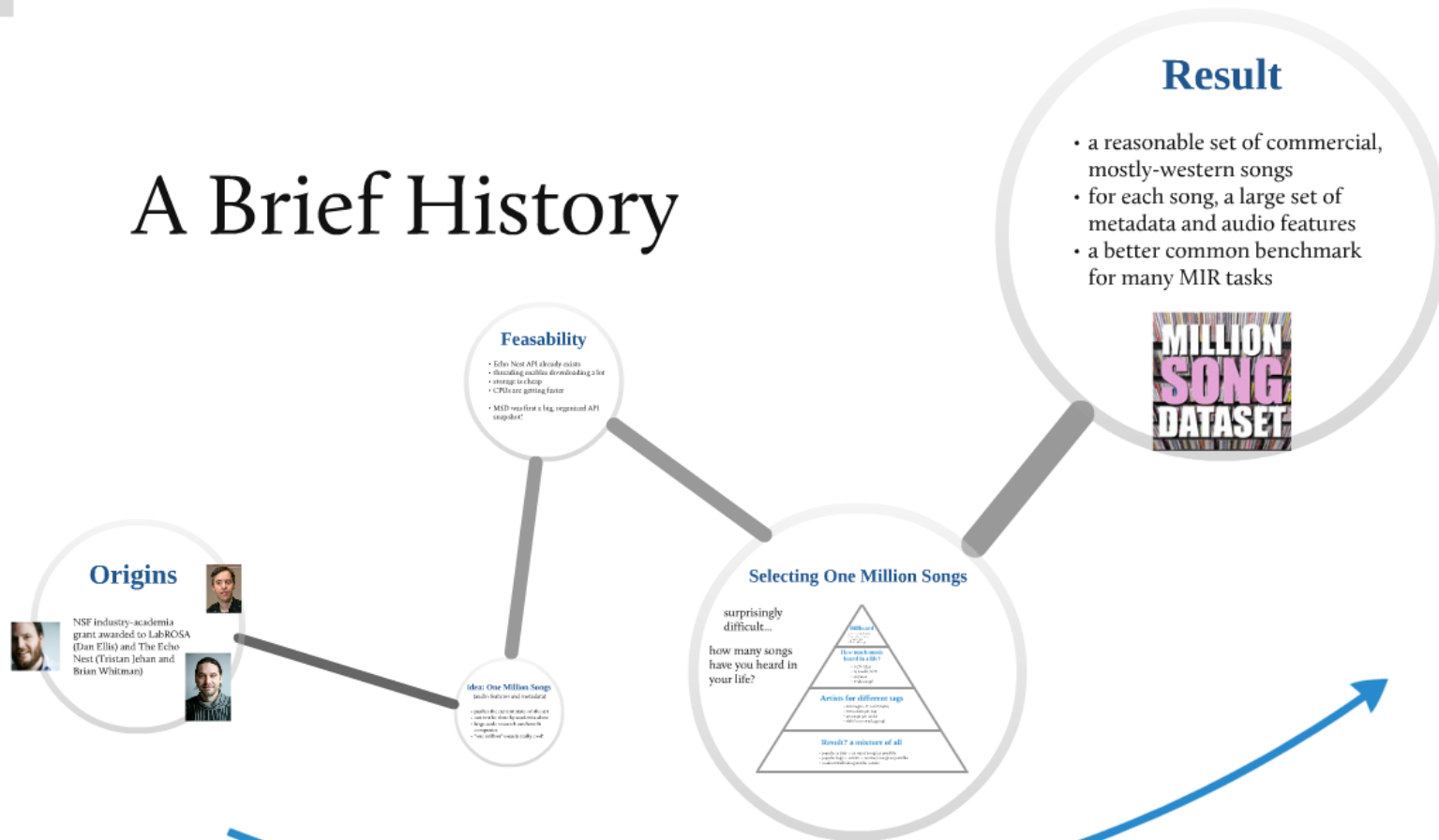
- large
- fixed
- aimed at researchers



MSD is:

- fixed set
- audio feature + metadata
- for researchers

# A Brief History



# Origins



NSF industry-academia  
grant awarded to LabROSA  
(Dan Ellis) and The Echo  
Nest (Tristan Jehan and  
Brian Whitman)



# **idea: One Million Songs**

(audio features and metadata)

- pushes the current state-of-the-art
- can not be done by academia alone
- large-scale research can benefit companies
- "one million" sounds really cool!



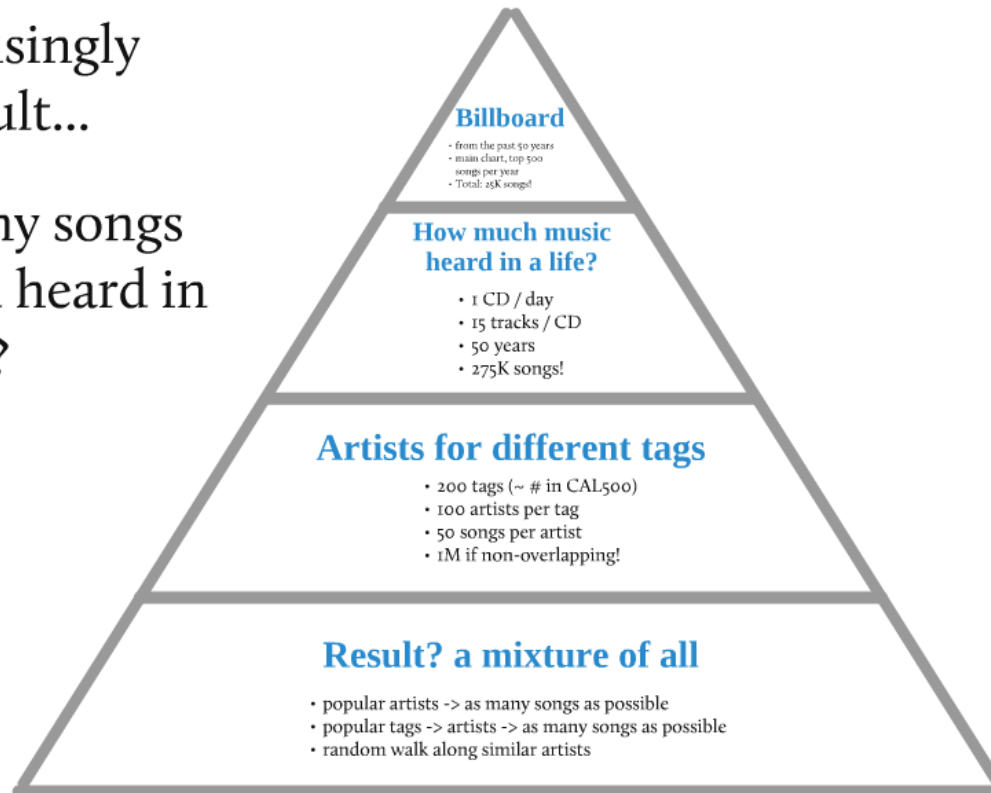
# Feasibility

- Echo Nest API already exists
- threading enables downloading a lot
- storage is cheap
- CPUs are getting faster
  
- MSD was first a big, organized API snapshot!

# Selecting One Million Songs

surprisingly  
difficult...

how many songs  
have you heard in  
your life?





# Billboard

- from the past 50 years
- main chart, top 500 songs per year
- Total: 25K songs!

## Billboard

- from the past 50 years
- main chart, top 500 songs per year
- Total: 25K songs!

## How much music heard in a life?

- 1 CD / day
- 15 tracks / CD
- 50 years
- 275K songs!

## Artists for different tags

- 200 tags (~ # in CAI 500)

gs  
d in

## How much music heard in a life?

- 1 CD / day
- 15 tracks / CD
- 50 years
- 275K songs!

## Artists for different tags

- 200 tags (~ # in CAL500)
- 100 artists per tag
- 50 songs per artist
- 1M if non-overlapping!

## Result? a mixture of all

- popular artists -> as many songs as possible
- popular tags -> artists -> as many songs as possible
- random walk along similar artists

- 15 tracks / CD
- 50 years
- 275K songs!

## Artists for different tags

- 200 tags (~ # in CAL500)
- 100 artists per tag
- 50 songs per artist
- 1M if non-overlapping!

## Result? a mixture of all

- popular artists -> as many songs as possible
- popular tags -> artists -> as many songs as possible
- random walk along similar artists

# Result

- a reasonable set of commercial, mostly-western songs
- for each song, a large set of metadata and audio features
- a better common benchmark for many MIR tasks



# The Million Song Dataset

WHAT YOU NEED TO REMEMBER FROM THIS TALK





# The Echo Nest



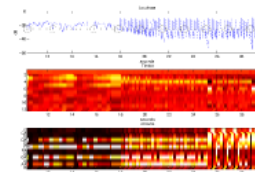
- First piece of the MSD
- Data taken from their API

artist data

```
• analysis_sample_rate: 22050
• artist_digitalid: 2200
• artist_familiarity: 0.784704224711
• artist_hottness: 0.623782610977
• artist_id: ARULZ74118789ADzEF
• artist_latitude: nan
• artist_location: Tupelo, MS
• artist_longitude: nan
• artist_mbidi: 01809552-4f87-45b0-aff-2c6f0730a2be
• artist_mbtags_count: shape = 10 x 1
• artist_name: Elvis Presley
• artist_playmeid: 542
• artist_terms: shape = 35 x 1
• artist_terms_freq: shape = 35 x 1
• artist_terms_weight: shape = 35 x 1
• audio_md5: 0642e5988f9c252609a4b44a0a12e1a
• bars_confidence: shape = 61 x 1
• bars_start: shape = 61 x 1
• beats_confidence: shape = 226 x 1
• beats_start: shape = 226 x 1
• danceability: 0.0
• duration: 200.09751
• end_of_fade_in: 0.287
• energy: 0.0
• key: 0
• key_confidence: 0.782
• loudness: -16.34
• mode: 1
• mode_confidence: 0.642
• release: Blue Christmas With Elvis
• release_digitalid: 443482
• sections_confidence: shape = 10 x 1
• sections_start: shape = 10 x 1
• segments_confidence: shape = 392 x 1
• segments_loudness_max: shape = 392 x 1
• segments_loudness_max_time: shape = 392 x 1
• segments_loudness_start: shape = 392 x 1
• segments_pitches: shape = 392 x 12
• segments_start: shape = 392 x 1
• segments_timbre: shape = 392 x 12
• similar_artists: shape = 100 x 1
• song_hottness: nan
• song_id: SOLDEL1A2AB0180E83
• start_of_fade_out: 197.677
• tatums_confidence: shape = 449 x 1
• tatums_start: shape = 449 x 1
• tempo: 67.219
• time_signature: 3
• time_signature_confidence: 0.743
• title: (There'll Be) Peace On The Valley (For Me)
• track_digitalid: 4924766
• track_id: TRABXW128F92FB9DF
• year: 0
```

audio features

audio features



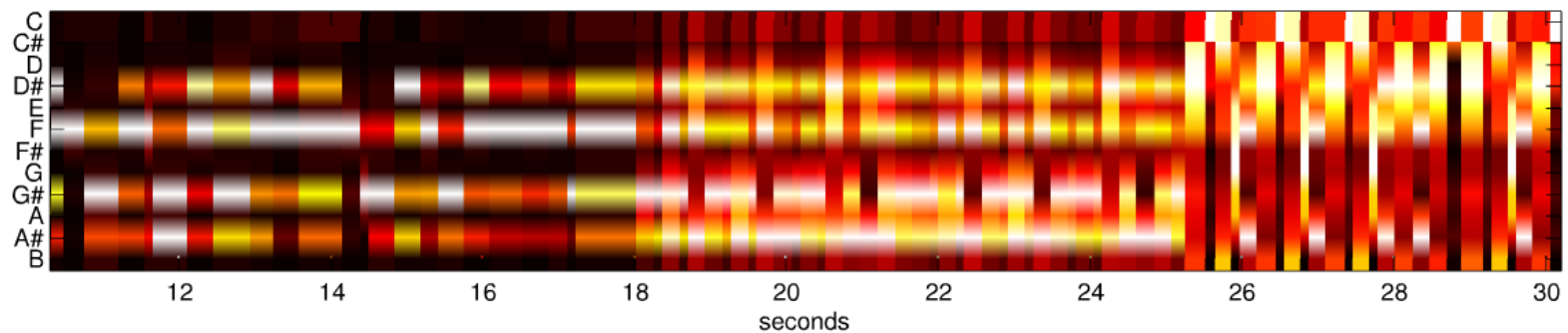
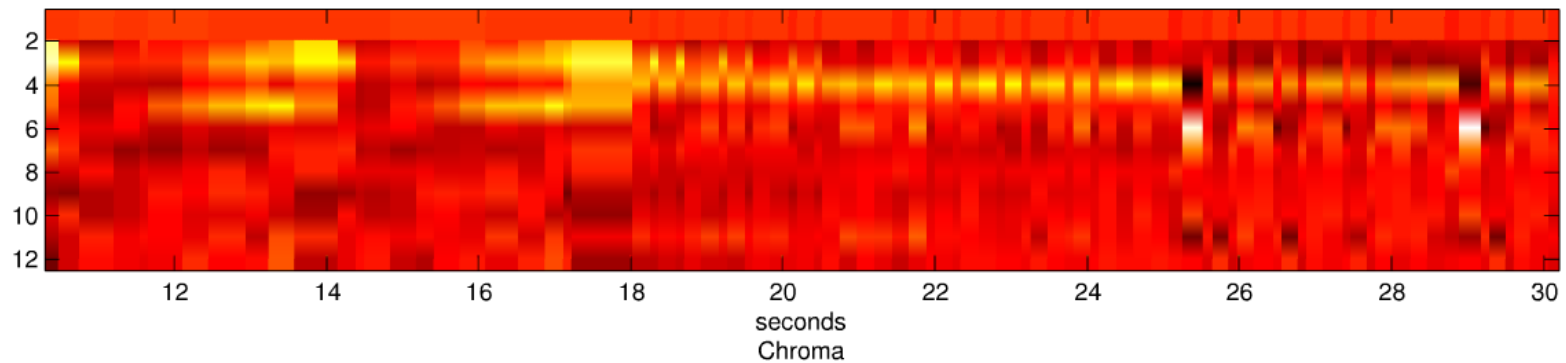
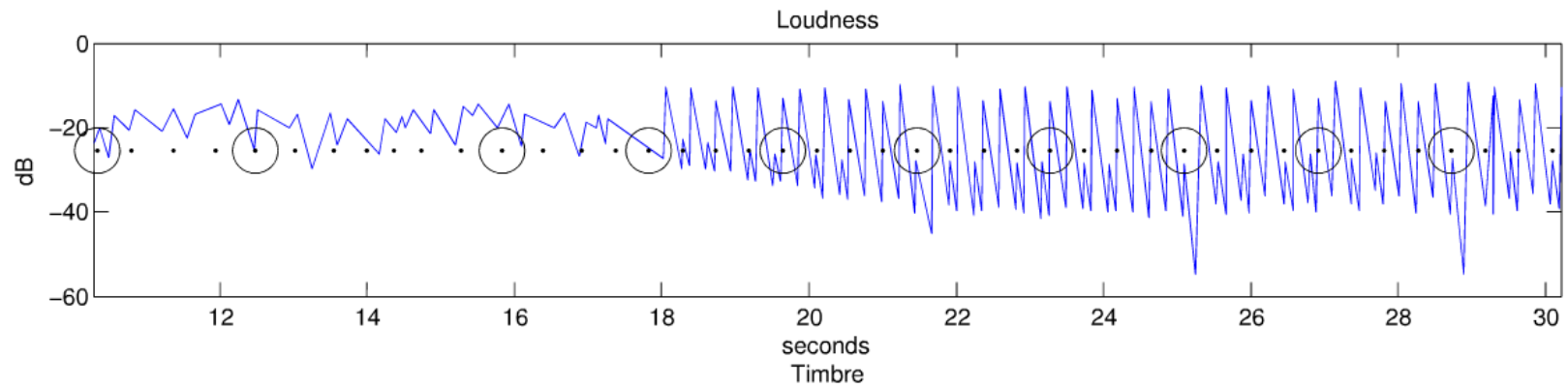
## artist data

- analysis\_sample\_rate: 22050
- artist\_7digitalid: 2200
- artist\_familiarity: 0.784704224711
- artist\_hotttnesss: 0.623782610977
- artist\_id: ARULZ741187B9AD2EF
- artist\_latitude: nan
- artist\_location: Tupelo, MS
- artist\_longitude: nan
- artist\_mbid: 01809552-4f87-45b0-afff-2c6fo730a3be
- artist\_mbtags: shape = 10 x 1
- artist\_mbtags\_count: shape = 10 x 1
- artist\_name: Elvis Presley
- artist\_playmeid: 542
- artist\_terms: shape = 25 x 1
- artist\_terms\_freq: shape = 25 x 1
- artist\_terms\_weight: shape = 25 x 1
- audio\_md5: 6642e8598869c025609a4b94a0a82e8a
- bars\_confidence: shape = 61 x 1
- bars\_start: shape = 61 x 1
- beats\_confidence: shape = 226 x 1
- beats\_start: shape = 226 x 1
- danceability: 0.0
- duration: 200.09751
- end\_of\_fade\_in: 0.287
- energy: 0.0

- key: 0
- key\_confidence: 0.782
- loudness: -16.34
- mode: 1
- mode\_confidence: 0.642
- release: Blue Christmas With Elvis
- release\_7digitalid: 443482
- sections\_confidence: shape = 10 x 1
- sections\_start: shape = 10 x 1
- segments\_confidence: shape = 392 x 1
- segments\_loudness\_max: shape = 392 x 1
- segments\_loudness\_max\_time: shape = 392 x 1
- segments\_loudness\_start: shape = 392 x 1
- segments\_pitches: shape = 392 x 12
- segments\_start: shape = 392 x 1
- segments\_timbre: shape = 392 x 12
- similar\_artists: shape = 100 x 1
- song\_hotttnesss: nan
- song\_id: SOLDELA12AB0180E83
- start\_of\_fade\_out: 197.677
- tatums\_confidence: shape = 449 x 1
- tatums\_start: shape = 449 x 1
- tempo: 67.219
- time\_signature: 3
- time\_signature\_confidence: 0.743
- title: (There'll Be) Peace On The Valley (For Me)
- track\_7digitalid: 4924766
- track\_id: TRABXWX128F92FB9DF
- year: 0

## audio features

# audio features



# MusicBrainz



<http://musicbrainz.org/>

- Reminder: you can download a local musicbrainz copy
- The Echo Nest provide artist level musicbrainz ID
- We matched ~ half of the songs to get year information
- More data to be exploited...!

# 7digital



- Online MP3 store
- Echo Nest provides 7digital IDs
- API gives you free 30 seconds sample

## GREAT FOR:

- quick demo / testing
- user experiments

**just in!**

Brian McFee has the  
Radio IDs for ~ half  
the million songs,  
more on this soon...

<http://>

# just in!

Brian McFee has the  
Radio IDs for ~ half  
the million songs,  
more on this soon...

# SHS



- Database of cover songs
- 18,196 matches in the MSD
- 5,854 "cover cliques"

# musiXmatch

**MUSIX**match<sup>®</sup>

not only words

- API of lyrics
- Provided lyrics in bag-of-words format for 237K tracks

i [28], babi [25], me [20], you [14], oh [12], not [10], my [8], still [8], believ [8], is [7], to [6], and [6], that [6], know [6], a [5], now [5], time [5], one [5], more [5], give [5], must [5], kill [5], hit [5], sign [5], confess [5], loneli [5], it [4], be [4], how [4], do [3], am [3], have [3], with [3], when [3], was [3], here [3], mind [3], lose [3], the [2], will [2], go [2], want [2], let [2], yeah [2], need [2], caus [2], tell [2], show [2], should [2], becaus [2], pretti [2], suppos [2], of [1], are [1], there [1], out [1], got [1], way [1], would [1], die [1], right [1], noth [1], e [1], someth [1], und [1], boy [1], breath [1], reason [1], blind [1], plan [1], sight [1]

Britney Spears - ... Baby One More Time!

---



i [28], babi [25], me [20], you [14], oh [12], not [10], my [8], still [8], believ [8], is [7], to [6], and [6], that [6], know [6], a [5], now [5], time [5], one [5], more [5], give [5], must [5], kill [5], hit [5], sign [5], confess [5], loneli [5], it [4], be [4], how [4], do [3], am [3], have [3], with [3], when [3], was [3], here [3], mind [3], lose [3], the [2], will [2], go [2], want [2], let [2], yeah [2], need [2], caus [2], tell [2], show [2], should [2], becaus [2], pretti [2], suppos [2], of [1], are [1], there [1], out [1], got [1], way [1], would [1], die [1], right [1], noth [1], e [1], someth [1], und [1], boy [1], breath [1], reason [1], blind [1], plan [1], sight [1]



# Last.fm

## last.fm

song-level tags and song to song similarity!

- \* 943,347 matched tracks MSD <-> Last.fm
- \* 505,216 tracks with at least one tag
- \* 584,897 tracks with at least one similar track
- \* 522,366 unique tags
- \* 8,598,630 (track - tag) pairs
- \* 56,506,688 (track - similar track) pairs

Comparison: CAL500 contains  
~1,700 (track - tag) pairs

# Taste Profile



subset of a larger user data dataset provided from The Echo Nest

- contains >40M (user - song - playcount) triplets
- more soon!

\*8,3  
\*56,50

Co  
~1

NEW



the  
Echo Nest

- First piece of the MSD
- Data taken from their API

audio features

audio features

echo nest

Taste Profile

subset of a larger user data dataset provided from The Echo Nest

- contains >40M (user - song - playcount) triplets
- more soon!

echo nest

Morrison / Gil Shaham  
Sessions & Guests  
by Ken  
Mera  
Zsche

# Future Improvements

## More sources of data

- Ongoing discussions with other API providers, e.g. Musicmetric



- We would love to have image/video features, from album covers or videoclip?

If you have such data (as a research lab, a company, ...) please let us know :)

## the MSD is an "open" community resource

- Columbia started the MSD project
- we are interested in improving it
  - we work with it
  - we will maintain the website

But...

- not much active development planned
- we used most of our contacts

Anyone can contribute!

- releasing additional datasets
- releasing code
- reformatting the whole MSD
- ...



# More sources of data

- Ongoing discussions with other API providers, e.g. Musicmetric



- We would love to have image/video features, from album covers or videoclip?

If you have such data (as a research lab, a company, ...) please let us know :)

# the MSD is an "open" community resource

Columbia started the MSD project

- we are interested in improving it
- we work with it
- we will maintain the website

But...

- not much active development planned
- we used most of our contacts

Anyone can contribute!

- releasing additional datasets
- releasing code
- reformatting the whole MSD
- ...



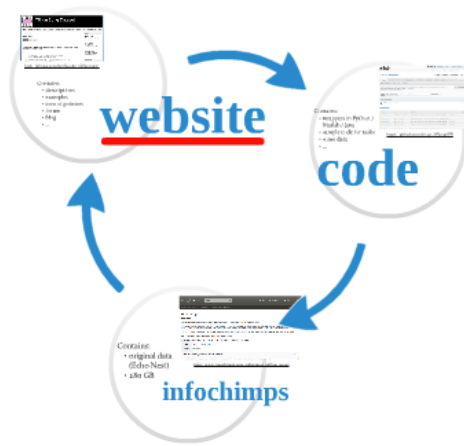
# The Million Song Dataset

WHAT YOU NEED TO REMEMBER FROM THIS TALK



# Some specifics / Getting started

## resources



## original data - HDF5

### HDF5

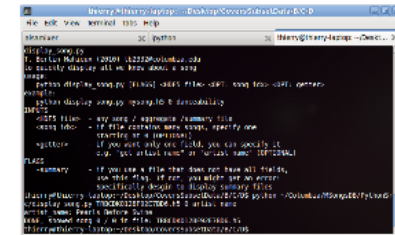
one file per track

HDF5 format  
contains all Echo Nest data +  
7 digital ID +  
artist musicbrainz ID

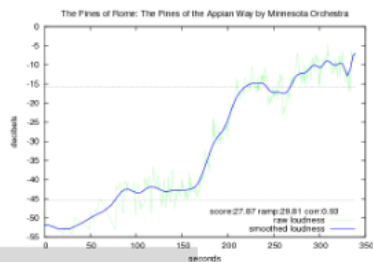
HDF5 is a format to contain and search large heterogeneous data developed by NASA... cool stuff!

1 Million Files!

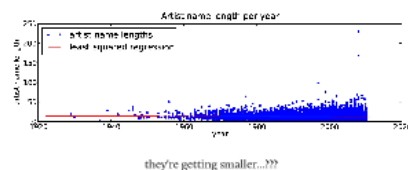
Files organized by Echo Nest track ID:  
track "TRBCDKO128F92E7BD6"  
in "/B/C/D/TRBCDKO128F92E7BD6.h5"



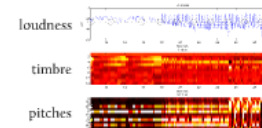
Another example: finding songs with slow build (from Paul Lamere - musicmachinery.com)



A quick MSD example: artist name length per year



## Echo Nest Audio Features



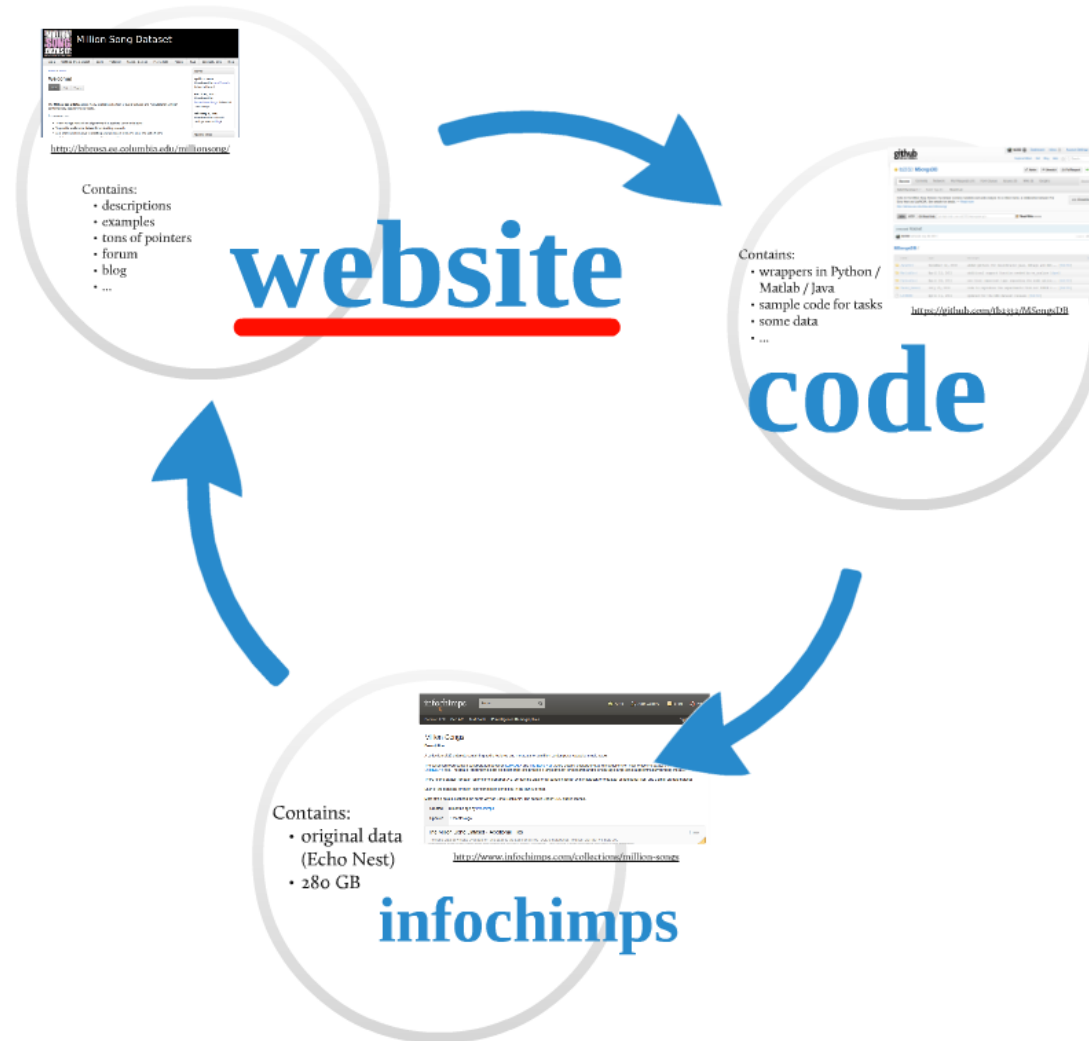
timbre and pitch information on the segment level (= between 2 note onsets), plus approximation of the beats, bars, sections

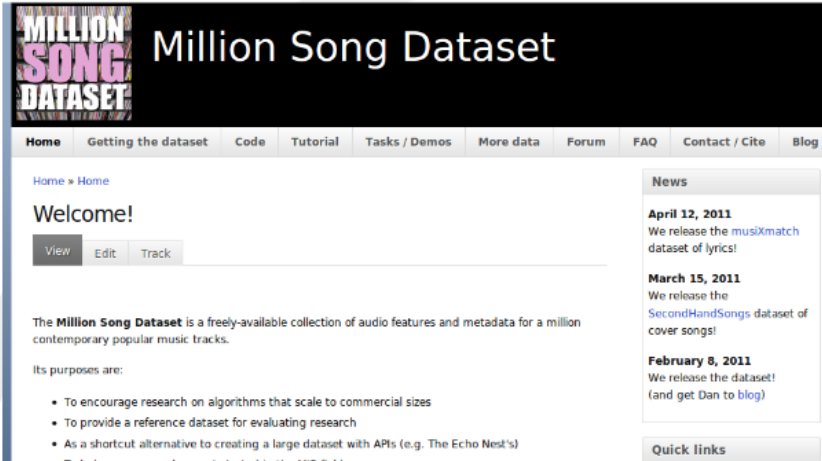
Start with the subset!

10K Songs

- Complete dataset on its own
- 10K songs / HDF5 file
  - small to download & play
  - all from tagging training set

# resources





<http://labrosa.ee.columbia.edu/millionsong/>

Contains:

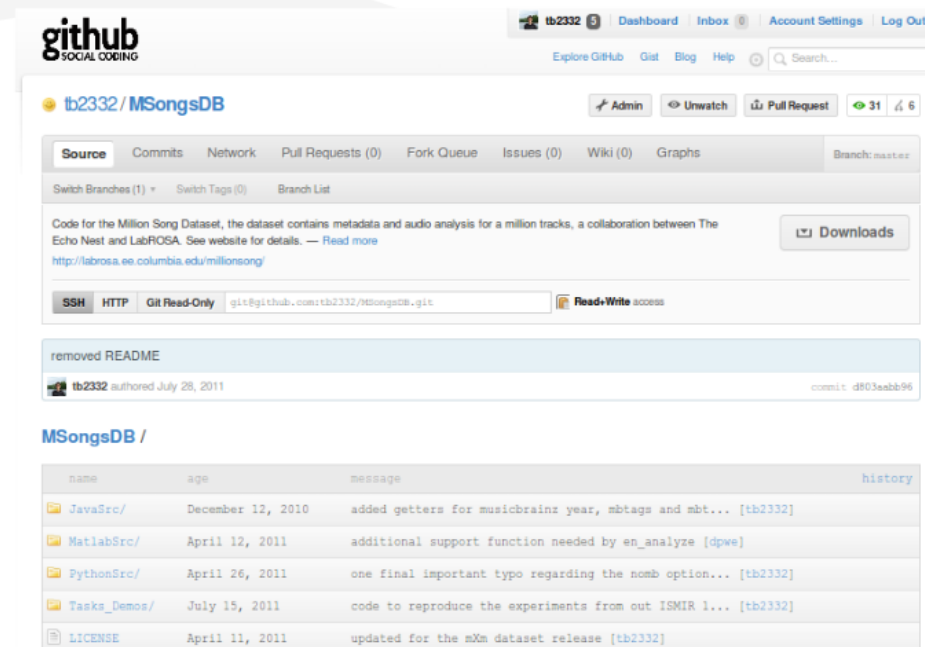
- descriptions
- examples
- tons of pointers
- forum
- blog
- ...

webs

Contains:

- wrappers in Python / Matlab / Java
- sample code for tasks
- some data
- ...

# code



github SOCIAL CODING

tb2332 Dashboard Inbox Account Settings Log Out

Explore GitHub Git Blog Help Search...

tb2332 / MSongsDB Admin Unwatch Pull Request 31 6

Source Commits Network Pull Requests (0) Fork Queue Issues (0) Wiki (0) Graphs Branch: master

Switch Branches (1) Switch Tags (0) Branch List

Code for the Million Song Dataset, the dataset contains metadata and audio analysis for a million tracks, a collaboration between The Echo Nest and LabROSA. See website for details. — Read more  
<http://labrosa.ee.columbia.edu/millionsong> Downloads

SSH HTTP Git Read-Only git@github.com:tb2332/MSongsDB.git Read+Write access

removed README

tb2332 authored July 28, 2011 commit d803aabb96

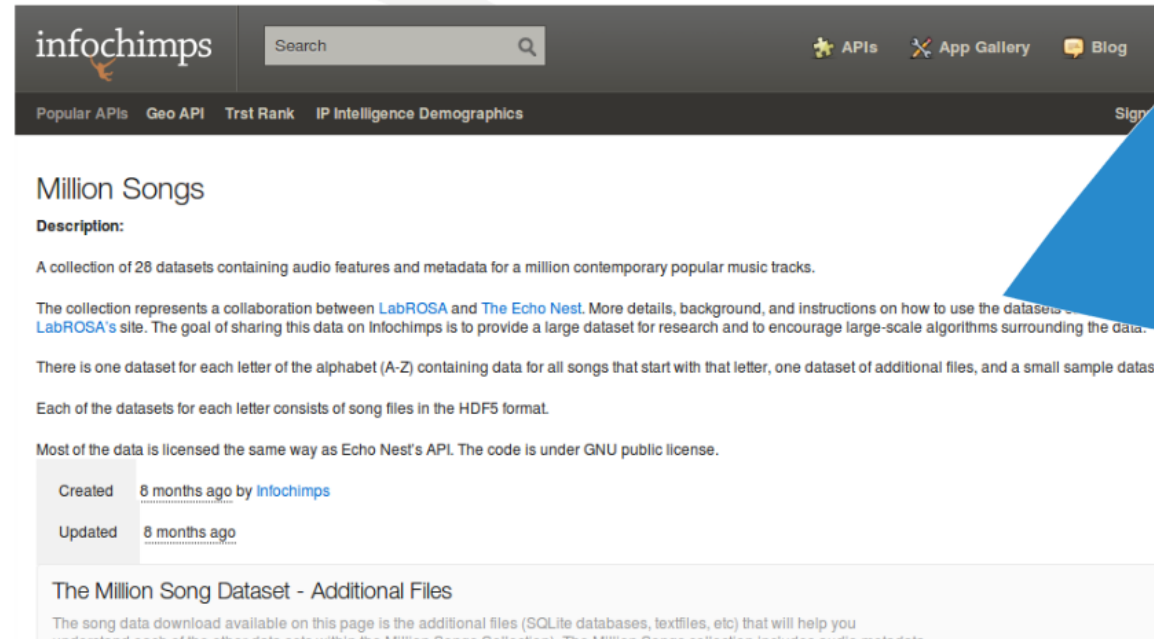
MSongsDB /

name	age	message	history
JavaSrc/	December 12, 2010	added getters for musicbrainz year, mbtags and mbt... [tb2332]	
MatlabSrc/	April 12, 2011	additional support function needed by en_analyze [dpwe]	
PythonSrc/	April 26, 2011	one final important typo regarding the nomb option... [tb2332]	
Tasks_Demos/	July 15, 2011	code to reproduce the experiments from out ISMIR 1... [tb2332]	
LICENSE	April 11, 2011	updated for the mXm dataset release [tb2332]	

<https://github.com/tb2332/MSongsDB>

## Contains:

- original data (Echo Nest)
- 280 GB

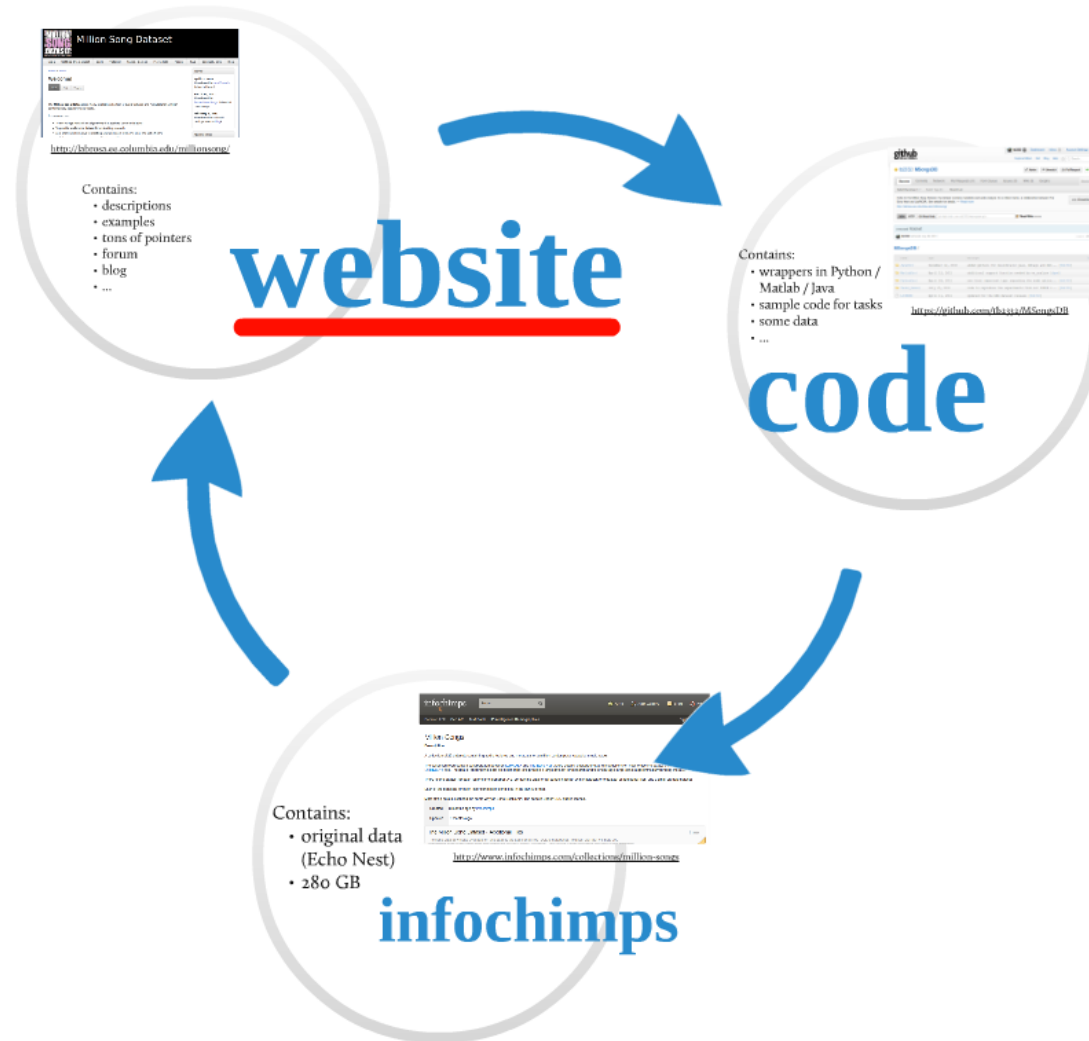


The screenshot shows the 'Million Songs' dataset page on the infochimps website. The page header includes the 'infochimps' logo, a search bar, and navigation links for 'APIs', 'App Gallery', and 'Blog'. Below the header, there are links for 'Popular APIs', 'Geo API', 'Trust Rank', and 'IP Intelligence Demographics'. The main content area is titled 'Million Songs' and includes a 'Description' section. The description states that the collection consists of 28 datasets containing audio features and metadata for a million contemporary popular music tracks. It mentions a collaboration between LabROSA and The Echo Nest and provides instructions on how to use the datasets. The page also includes a table with 'Created' and 'Updated' dates, both listed as '8 months ago by infochimps'. Below the table, there is a section titled 'The Million Song Dataset - Additional Files' which describes the additional files available for download.

<http://www.infochimps.com/collections/million-songs>

# infochimps

# resources



# original data - HDF5

## HDF5

one file per track

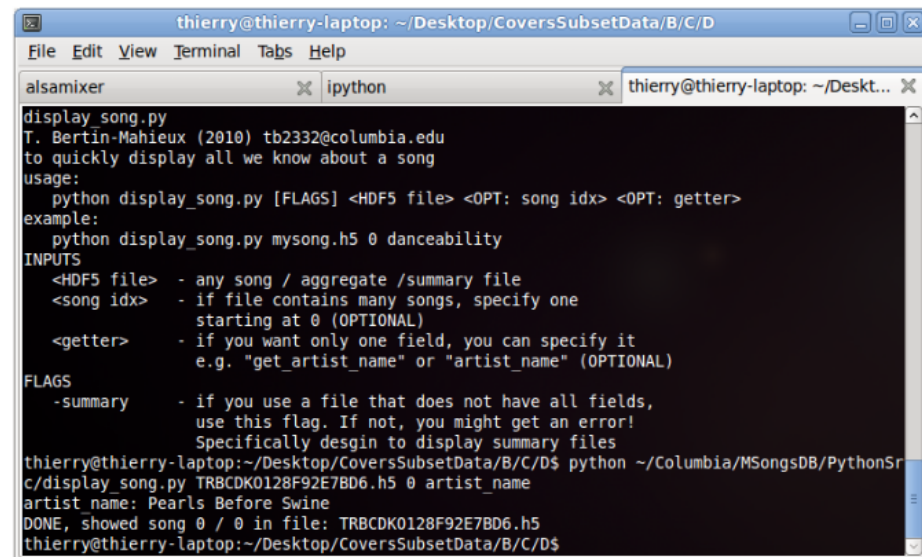
HDF5 format

contains all Echo  
Nest data +  
7digital ID +  
artist musicbrainz  
ID

HDF5 is a format to  
contain and search large  
heterogeneous data  
developed by NASA...  
cool stuff!

1 Million Files!

Files organized by Echo Nest track ID:  
track "TRBCDKO128F92E7BD6"  
in "/B/C/D/TRBCDKO128F92E7BD6.h5"



```
thierry@thierry-laptop: ~/Desktop/CoversSubsetData/B/C/D
File Edit View Terminal Tabs Help
alsamixer ipython thierry@thierry-laptop: ~/Desk...
display_song.py
T. Bertin-Mahieux (2010) tb2332@columbia.edu
to quickly display all we know about a song
usage:
  python display_song.py [FLAGS] <HDF5 file> <OPT: song idx> <OPT: getter>
example:
  python display_song.py mysong.h5 0 danceability
INPUTS
  <HDF5 file> - any song / aggregate /summary file
  <song idx>  - if file contains many songs, specify one
               starting at 0 (OPTIONAL)
  <getter>   - if you want only one field, you can specify it
               e.g. "get_artist_name" or "artist_name" (OPTIONAL)
FLAGS
  -summary   - if you use a file that does not have all fields,
               use this flag. If not, you might get an error!
               Specifically desgin to display summary files
thierry@thierry-laptop:~/Desktop/CoversSubsetData/B/C/D$ python ~/Columbia/MSongsDB/PythonSrc
c/display_song.py TRBCDKO128F92E7BD6.h5 0 artist_name
artist name: Pearls Before Swine
DONE, showed song 0 / 0 in file: TRBCDKO128F92E7BD6.h5
thierry@thierry-laptop:~/Desktop/CoversSubsetData/B/C/D$
```



thierry@thierry-laptop: ~/Desktop/CoversSubsetData/B/C/D

File Edit View Terminal Tabs Help

alsamixer

ipython

thierry@thierry-laptop: ~/Deskt... X

display\_song.py

T. Bertin-Mahieux (2010) tb2332@columbia.edu

to quickly display all we know about a song

usage:

```
python display_song.py [FLAGS] <HDF5 file> <OPT: song idx> <OPT: getter>
```

example:

```
python display_song.py mysong.h5 0 danceability
```

INPUTS

<HDF5 file> - any song / aggregate /summary file

<song idx> - if file contains many songs, specify one starting at 0 (OPTIONAL)

<getter> - if you want only one field, you can specify it e.g. "get\_artist\_name" or "artist\_name" (OPTIONAL)

FLAGS

-summary - if you use a file that does not have all fields, use this flag. If not, you might get an error! Specifically desgin to display summary files

```
thierry@thierry-laptop:~/Desktop/CoversSubsetData/B/C/D$ python ~/Columbia/MSongsDB/PythonSrc/display_song.py TRBCDK0128F92E7BD6.h5 0 artist_name
```

```
artist_name: Pearls Before Swine
```

```
DONE, showed song 0 / 0 in file: TRBCDK0128F92E7BD6.h5
```

```
thierry@thierry-laptop:~/Desktop/CoversSubsetData/B/C/D$
```

Start with the subset!



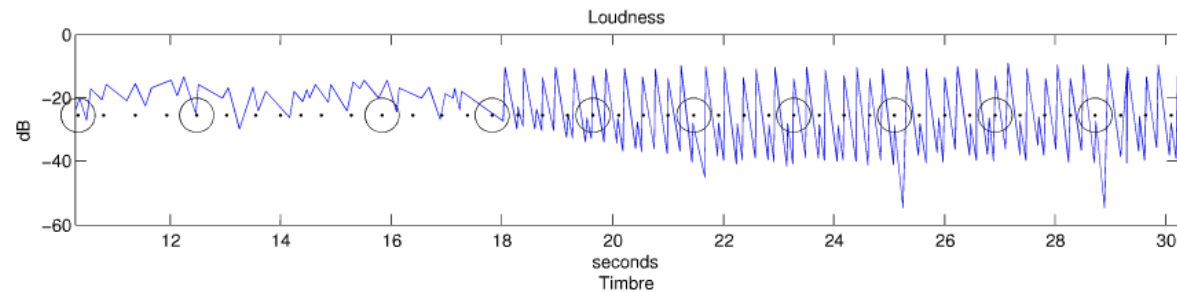
10K Songs

Complete dataset on its own

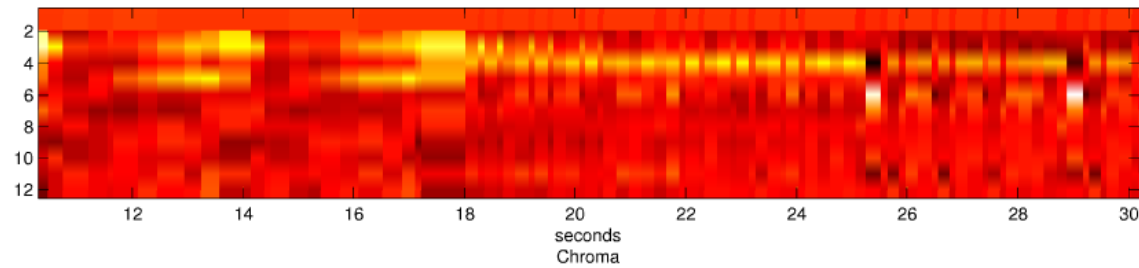
- 10K songs / HDF5 file
- small to download & play
- all from tagging training set

# Echo Nest Audio Features

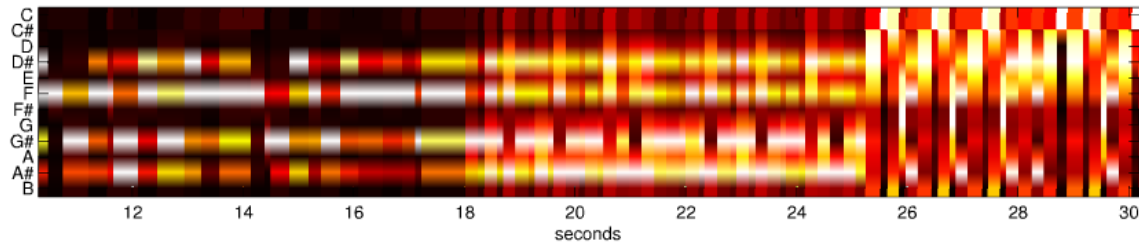
loudness



timbre

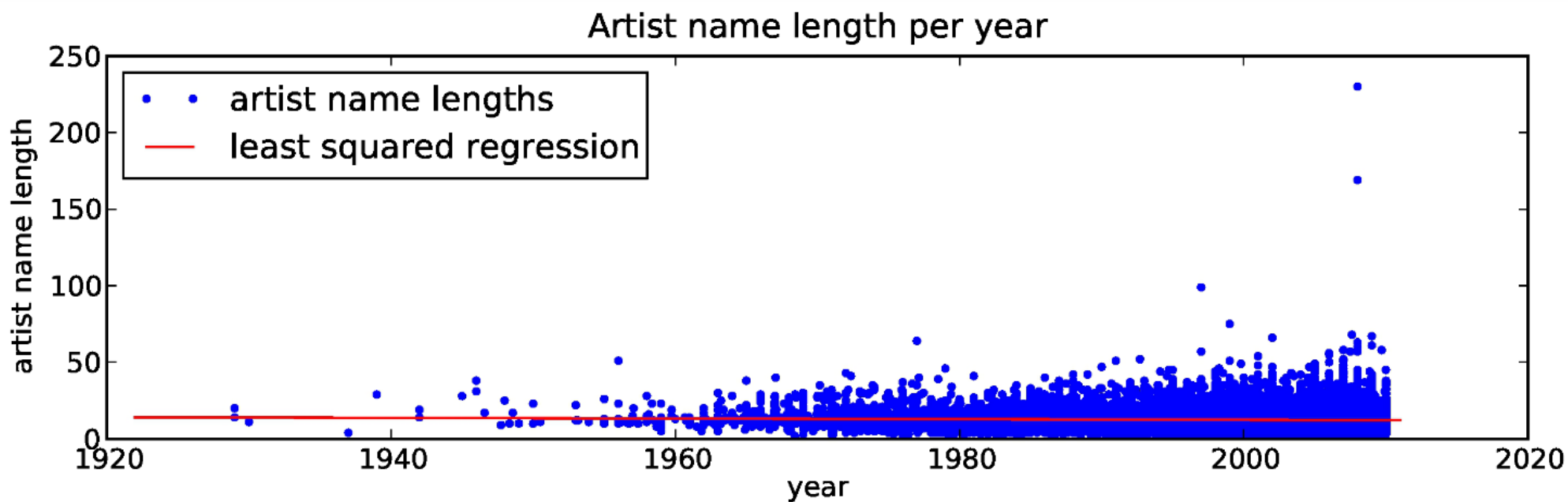


itches



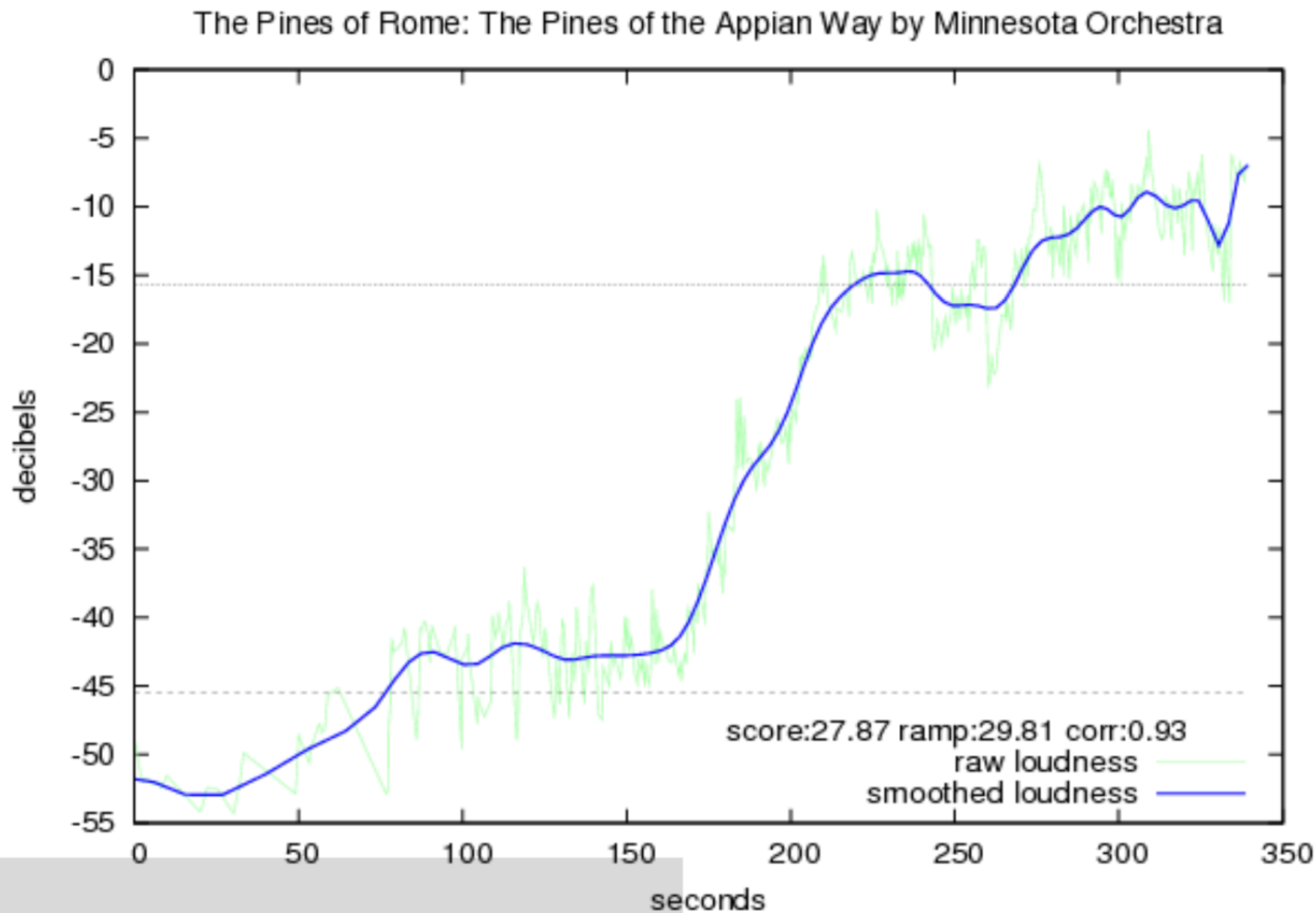
timbre and pitch information on the segment level (~ between 2 note onsets), plus approximation of the beats, bars, sections

# A quick MSD example: artist name length per year



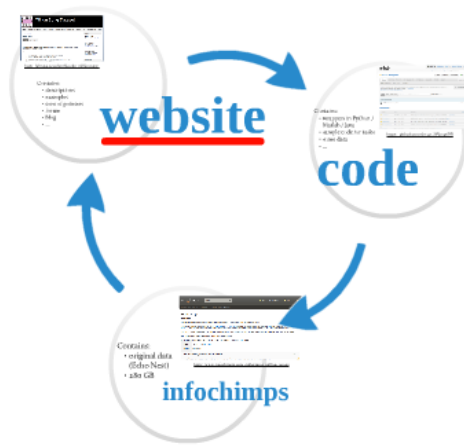
they're getting smaller...???

# Another example: finding songs with slow build (from Paul Lamere - musicmachinery.com)



# Some specifics / Getting started

## resources



## original data - HDF5

### HDF5

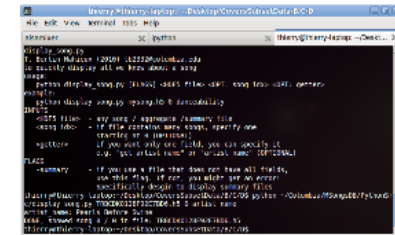
one file per track

HDF5 format  
contains all Echo Nest data +  
7 digital ID +  
artist musicbrainz ID

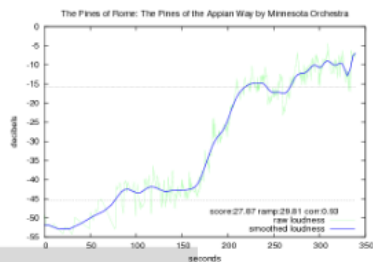
HDF5 is a format to contain and search large heterogeneous data developed by NASA... cool stuff!

1 Million Files!

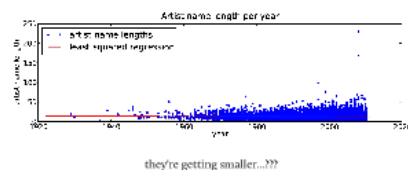
Files organized by Echo Nest track ID:  
track "TRBCDKO128F92E7BD6"  
in "/B/C/D/TRBCDKO128F92E7BD6.h5"



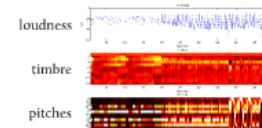
Another example: finding songs with slow build (from Paul Lamere - musicmachinery.com)



A quick MSD example: artist name length per year



## Echo Nest Audio Features



timbre and pitch information on the segment level (= between 2 note onsets), plus approximation of the beats, bars, sections

Start with the subset!

10K Songs

- Complete dataset on its own
- 10K songs / HDF5 file
  - small to download & play
  - all from tagging training set

# ISSUES WITH THE MSD

## DUPLICATES

Many duplicates!  
some are really different versions  
some are clear mistakes  
Dataset contains ~850K unique songs  
(we provide a list of many duplicates)

```
thierry@thierry-laptop:~/Columbia/MSongsDB/Tasks_Demos/SQLite
File Edit View Terminal Help
In [17]: sql = "SELECT track_id, artist_name, title, duration FROM songs WHERE
title='s' AND artist_name='s' % ('Bad Romance', 'Lady Gaga')
In [18]: res = conn.execute(sql); res.fetchall()
Out[18]:
[[('TRACXCT128F934A4AF', u'Lady Gaga', u'Bad Romance', 294.47791000000001),
 ('TRWGTZK128F934A4A81', u'Lady Gaga', u'Bad Romance', 433.37097999999997),
 ('TRWQWNG128F934A4A8B', u'Lady Gaga', u'Bad Romance', 311.58612),
 ('TRKJOUH128F934A4A86', u'Lady Gaga', u'Bad Romance', 295.79567),
 ('TRACFSA12903CC089B', u'Lady Gaga', u'Bad Romance', 262.68689000000001)]
```

and  
more

...



sorry.

## FORMAT

IM files? come on!

- requires this weird HDF5 library
- difficult to iterate over IM songs
- 26 folders / tar files to download?!
- Duplicate information, e.g. artist data

yes, yes, yes and yes... but...



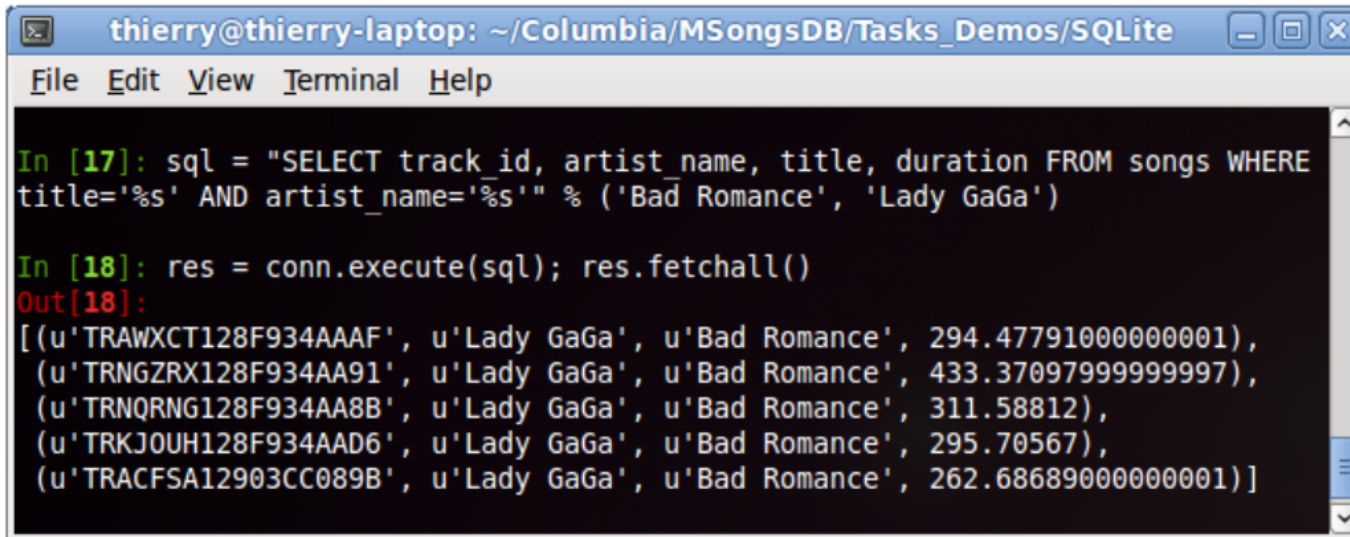
# DUPLICATES

Many duplicates!

some are really different versions

some are clear mistakes

Dataset contains ~850K unique songs  
(we provide a list of many duplicates)



```
thierry@thierry-laptop: ~/Columbia/MSongsDB/Tasks_Demos/SQLite
File Edit View Terminal Help
In [17]: sql = "SELECT track_id, artist_name, title, duration FROM songs WHERE
title='%s' AND artist_name='%s'" % ('Bad Romance', 'Lady GaGa')
In [18]: res = conn.execute(sql); res.fetchall()
Out[18]:
[(u'TRAWXCT128F934AAAF', u'Lady GaGa', u'Bad Romance', 294.47791000000001),
(u'TRNGZRX128F934AA91', u'Lady GaGa', u'Bad Romance', 433.37097999999997),
(u'TRNQRNG128F934AA8B', u'Lady GaGa', u'Bad Romance', 311.58812),
(u'TRKJOUH128F934AAD6', u'Lady GaGa', u'Bad Romance', 295.70567),
(u'TRACFSA12903CC089B', u'Lady GaGa', u'Bad Romance', 262.68689000000001)]
```



# FORMAT

IM files? come on!

- requires this weird HDF5 library
- difficult to iterate over IM songs
- 26 folders / tar files to download?!
- Duplicate information, e.g. artist data

yes, yes, yes and yes... but...

# ISSUES WITH THE MSD

## DUPLICATES

Many duplicates!  
some are really different versions  
some are clear mistakes  
Dataset contains ~850K unique songs  
(we provide a list of many duplicates)

```
thierry@thierry-laptop:~/Columbia/MSongsDB/Tasks_Demos/SQLite
File Edit View Terminal Help
In [17]: sql = "SELECT track_id, artist_name, title, duration FROM songs WHERE
title='s' AND artist_name='s' % ('Bad Romance', 'Lady Gaga')
In [18]: res = conn.execute(sql); res.fetchall()
Out[18]:
[[('TRACXCT128F934A4AF', u'Lady Gaga', u'Bad Romance', 294.47791000000001),
 ('TRWGTZK128F934A4A81', u'Lady Gaga', u'Bad Romance', 433.37097999999997),
 ('TRWQWNG128F934A4A8B', u'Lady Gaga', u'Bad Romance', 311.58612),
 ('TRKJOUH128F934A4A86', u'Lady Gaga', u'Bad Romance', 295.79567),
 ('TRACFSA12903CC089B', u'Lady Gaga', u'Bad Romance', 262.68689000000001)]
```

and  
more

...



sorry.

## FORMAT

IM files? come on!

- requires this weird HDF5 library
- difficult to iterate over IM songs
- 26 folders / tar files to download?!
- Duplicate information, e.g. artist data

yes, yes, yes and yes... but...

# Some discussions we'd like to have

## tasks

- What tasks can evolve thanks to more data?
- What new task can appear?  
(year prediction)
- How many tasks can be merged into one framework?  
(similarity with lyrics AND features AND tags...)

## other communities

- We have this large heterogeneous multimedial dataset!
- how can we share our research to other fields (AI, vision, speech, ...)?
  - how can we convince other fields to use our data?
  - how can we make ISMIR a stronger idea worldwide?
  - can we link this data to even more data?  
(images, web search, geo-data, cultural data, ...)

## releasing data

- After 5 datasets, we still have no clue what people want!
- ok, HDF5 is weird, but what are the alternatives?
  - what would we win from using RDE?
  - how easy is it to share databases, like MySQL?
  - are text files tab-delimited better?
  - who in the field is active on AWS?

and a random baby goat...

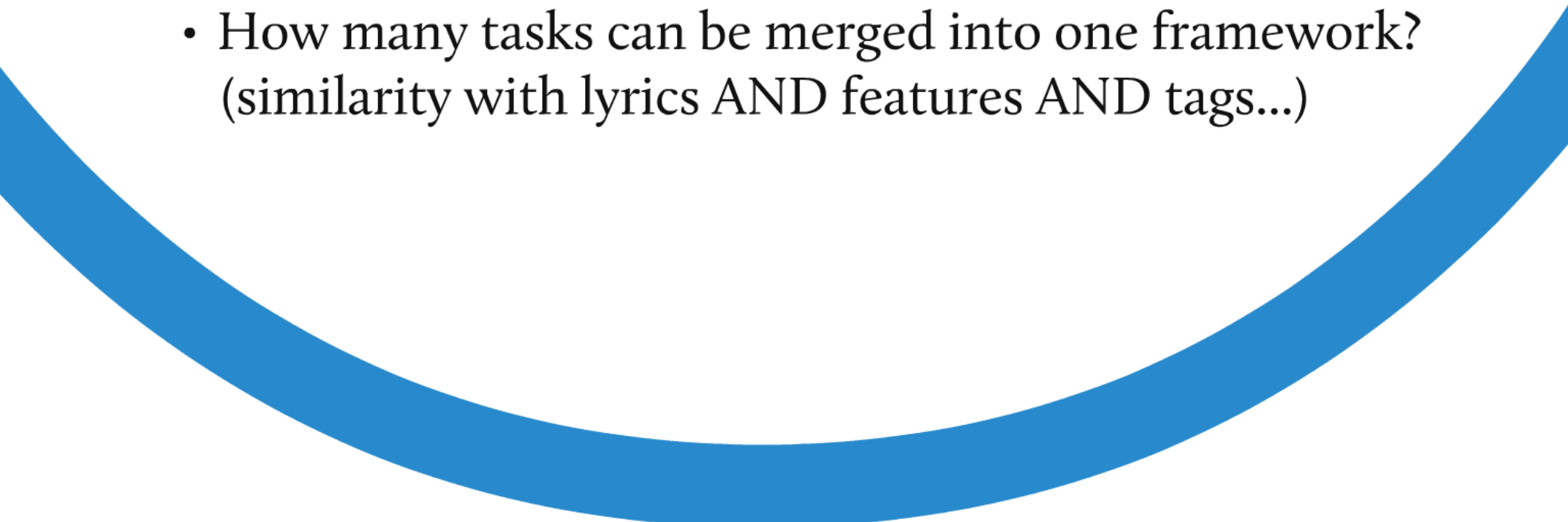




# tasks

- What tasks can evolve thanks to more data?
- What new task can appear?  
(year prediction)
- How many tasks can be merged into one framework?  
(similarity with lyrics AND features AND tags...)

# tasks

- What tasks can evolve thanks to more data?
  - What new task can appear?  
(year prediction)
  - How many tasks can be merged into one framework?  
(similarity with lyrics AND features AND tags...)
- 



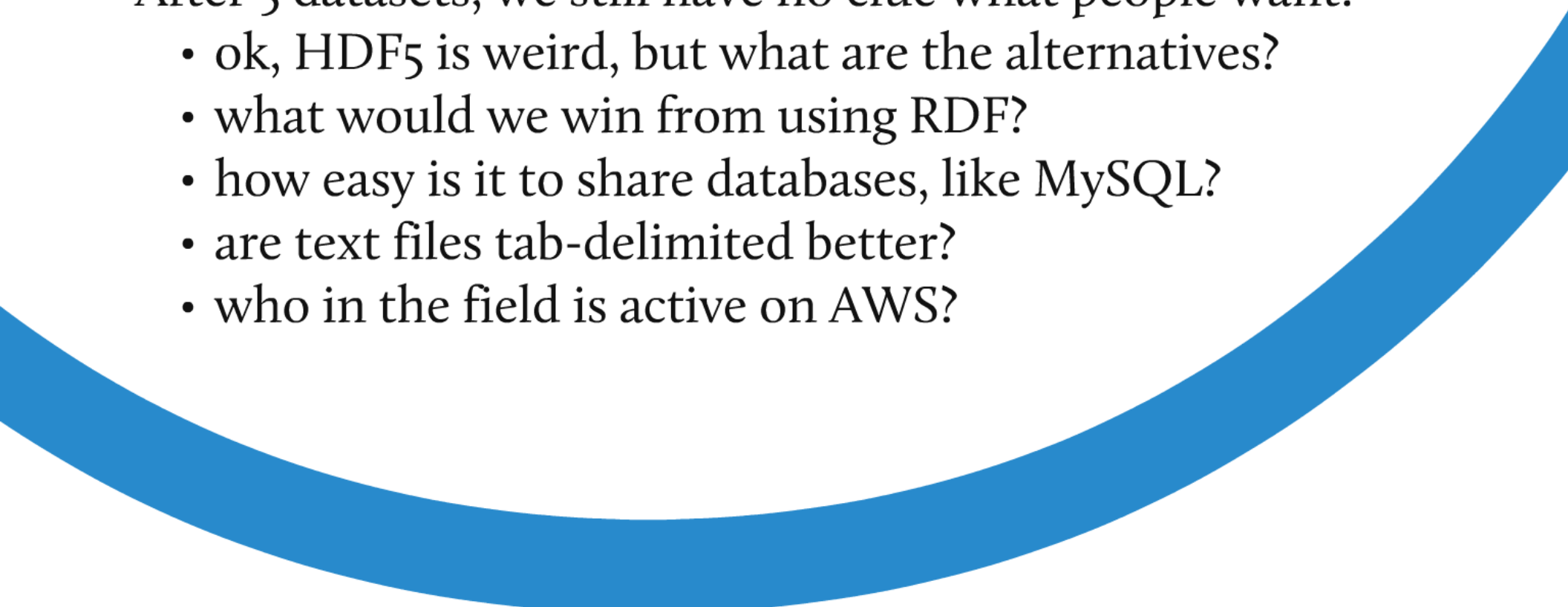
# releasing data

After 5 datasets, we still have no clue what people want!

- ok, HDF5 is weird, but what are the alternatives?
- what would we win from using RDF?
- how easy is it to share databases, like MySQL?
- are text files tab-delimited better?
- who in the field is active on AWS?

# data

After 5 datasets, we still have no clue what people want!

- ok, HDF5 is weird, but what are the alternatives?
  - what would we win from using RDF?
  - how easy is it to share databases, like MySQL?
  - are text files tab-delimited better?
  - who in the field is active on AWS?
- 



# other commu- nities

We have this large heterogeneous multimedia dataset!

- how can we show off our research to other fields (AI, vision, speech, ...)?
- how can we convince other fields to use our data?
- how can we make ISMIR a stronger idea incubator?
- can we link this data to even more data?  
(images, web search, geo data, cultural data, ...)



# minuties

We have this large heterogeneous multimedia dataset!

- how can we show off our research to other fields (AI, vision, speech, ...)?
- how can we convince other fields to use our data?
- how can we make ISMIR a stronger idea incubator?
- can we link this data to even more data?  
(images, web search, geo data, cultural data, ...)

# Some discussions we'd like to have

## tasks

- What tasks can evolve thanks to more data?
- What new task can appear?  
(year prediction)
- How many tasks can be merged into one framework?  
(similarity with lyrics AND features AND tags...)

## other communities

- We have this large heterogeneous multimedial dataset!
- how can we share our research to other fields (AI, vision, speech, ...)?
  - how can we convince other fields to use our data?
  - how can we make ISMIR a stronger idea worldwide?
  - can we link this data to even more data?  
(images, web search, geo-data, cultural data, ...)

## releasing data

- After 5 datasets, we still have no clue what people want!
- ok, HDF5 is weird, but what are the alternatives?
  - what would we win from using RDE?
  - how easy is it to share databases, like MySQL?
  - are text files tab-delimited better?
  - who in the field is active on AWS?

and a random baby goat...



## Work on the MSD at ISMIR

- Large-scale music similarity search with spatial trees, McFee and Lanckriet, Tuesday 12pm - 1:50pm
- The natural language of playlists, McFee and Lanckriet, Wednesday 2:50pm - 4:30pm
- The Million Song Dataset, Bertin-Mahieux and Ellis, Wednesday 4:40pm-5pm
- Audio-based Music Classification with a pretrained convolutional network, Dieleman, Brakel and Schrauwen, Thursday 3:10pm-5pm

# BREAK!

## (15 min)

Break quizz: if you have a recording A, which of the following are "the same song"?

- re-release of A
- radio edit, slightly different duration
- radio edit, some words changed
- re-mastering of A
- re-mastering of A, slightly diff. duration
- re-mastering of A, some more backvocals
- remix of A
- ...

<http://labrosa.ee.columbia.edu/millionsong/ismir2011>