Magnatagatune

-

one million songs

Million Song Dataset

ISMIR 2011 tutorial

by Thierry Bertin-Mahieux
Matt Hoffman
Dan Ellis

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

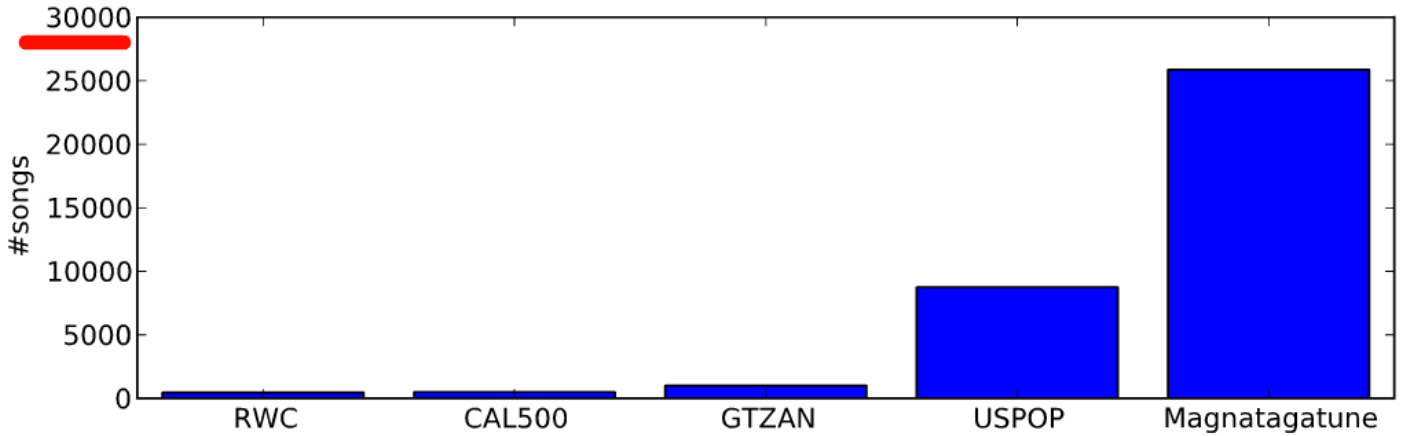
Overview

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

(break)

- Practical demonstrations

Existing Datasets



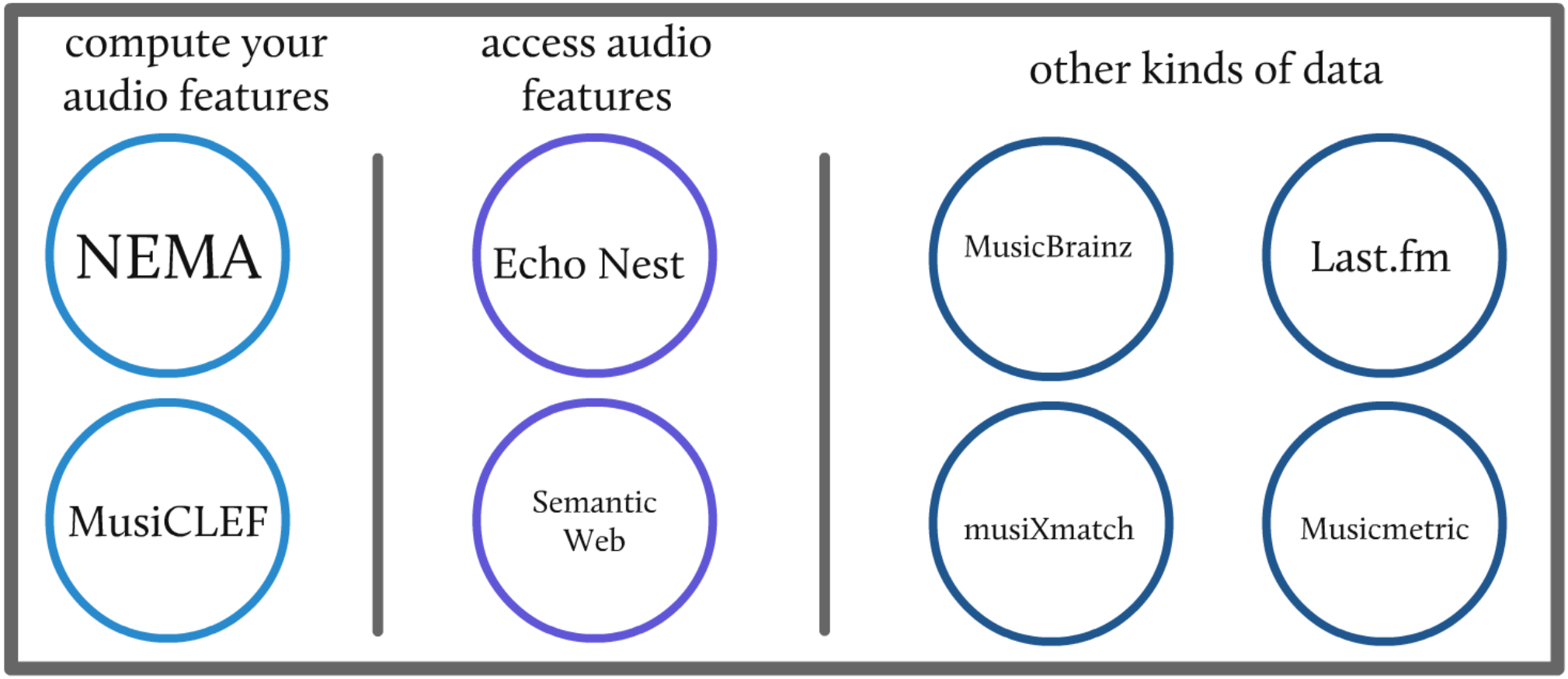
Commercial

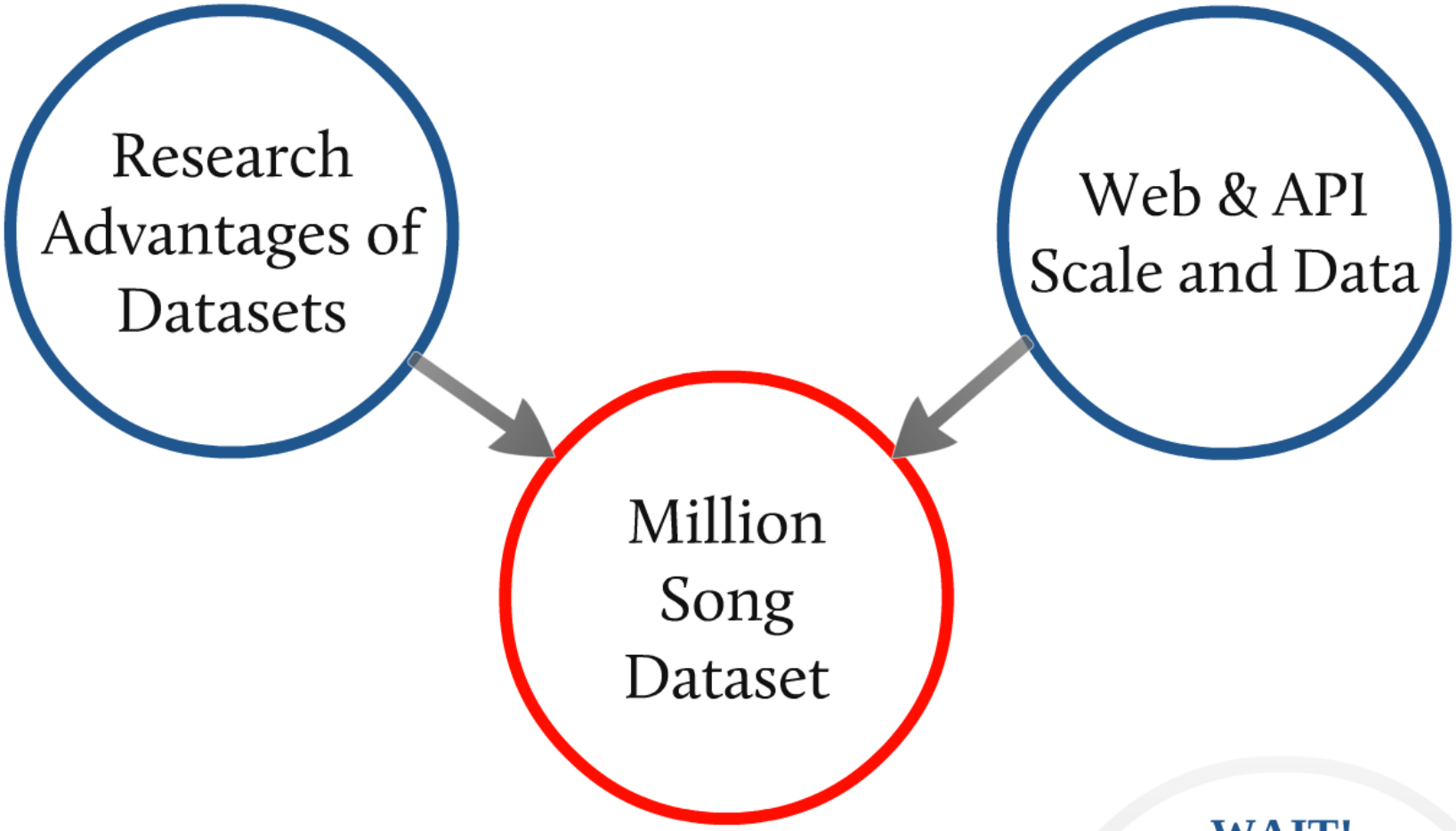
- Spotify: 15M
- iTunes: 14M
- Last.fm: 12M
- MOG: 12M
- RDIO: 9M

one million songs

Magnatagatune

Existing Web Resources - avoiding copyrights





WAIT! Aren't APIs better?

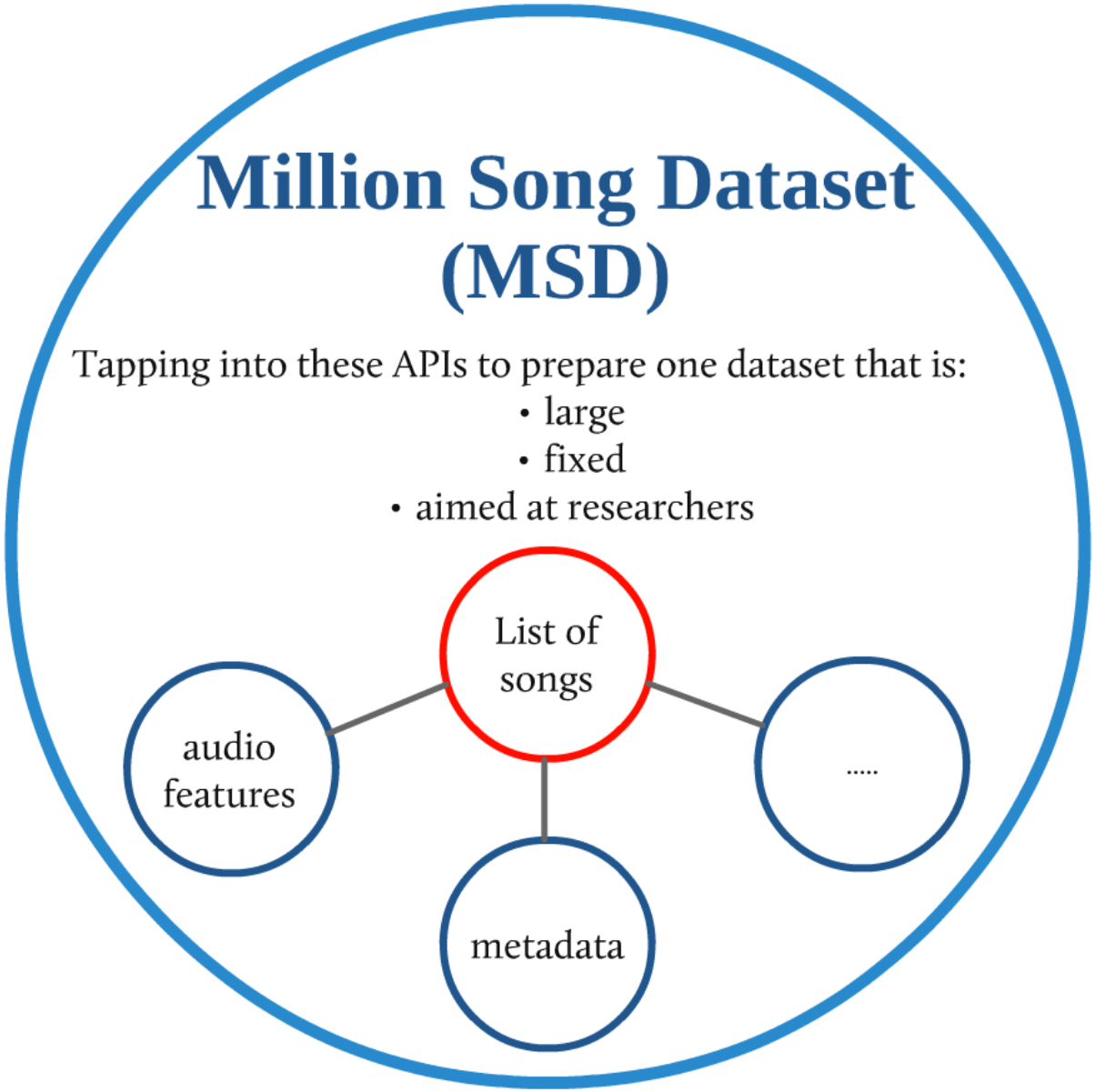
Different goals!

- 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



MSD is:

- fixed set
- audio feature + metadata
- for researchers

A Brief History

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

Feasability

- 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!

Origins



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

idea: One Million Songs

Selecting One Million Songs

surprisingly difficult...

how many songs have you heard in your life?

Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

echonest

Origins

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

idea: One Million Songs

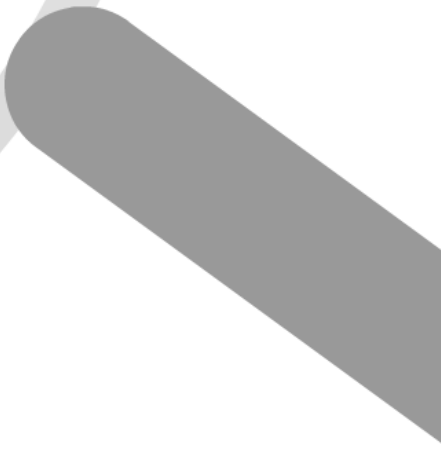
(audio features and metadata)

- 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!

Feasability

- Echo Nest API already exists
- threading enables downloading a lot
- storage is cheap
- CPUs are getting faster

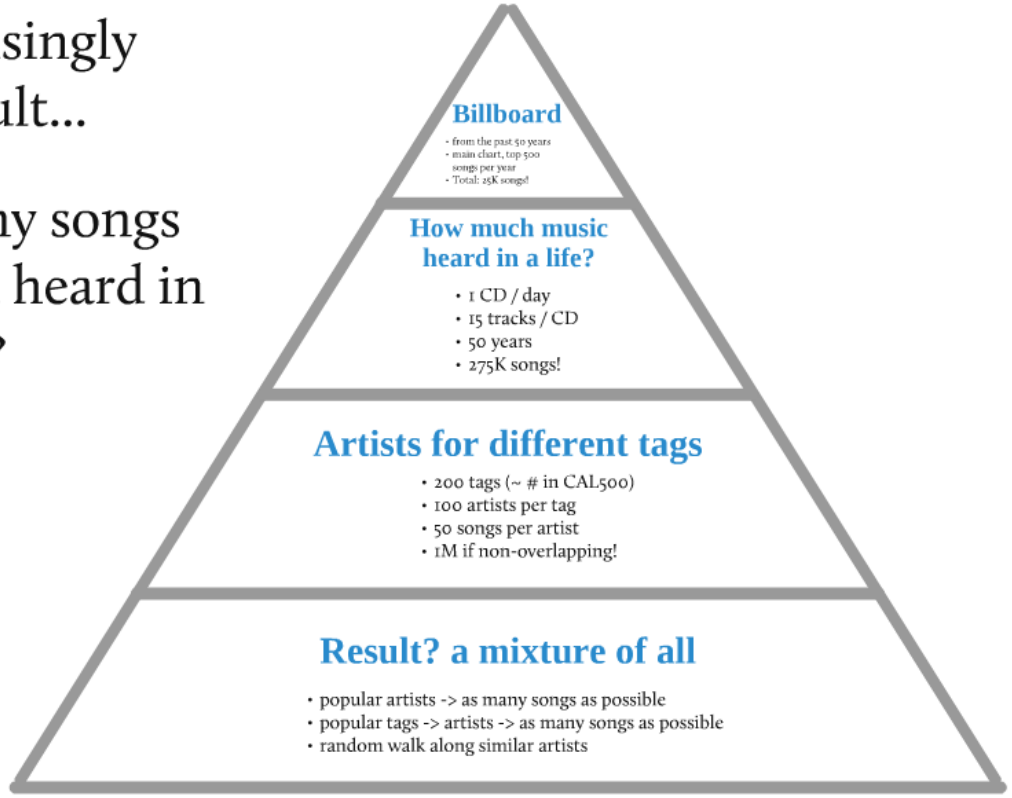
- MSD was first a big, organized API snapshot!

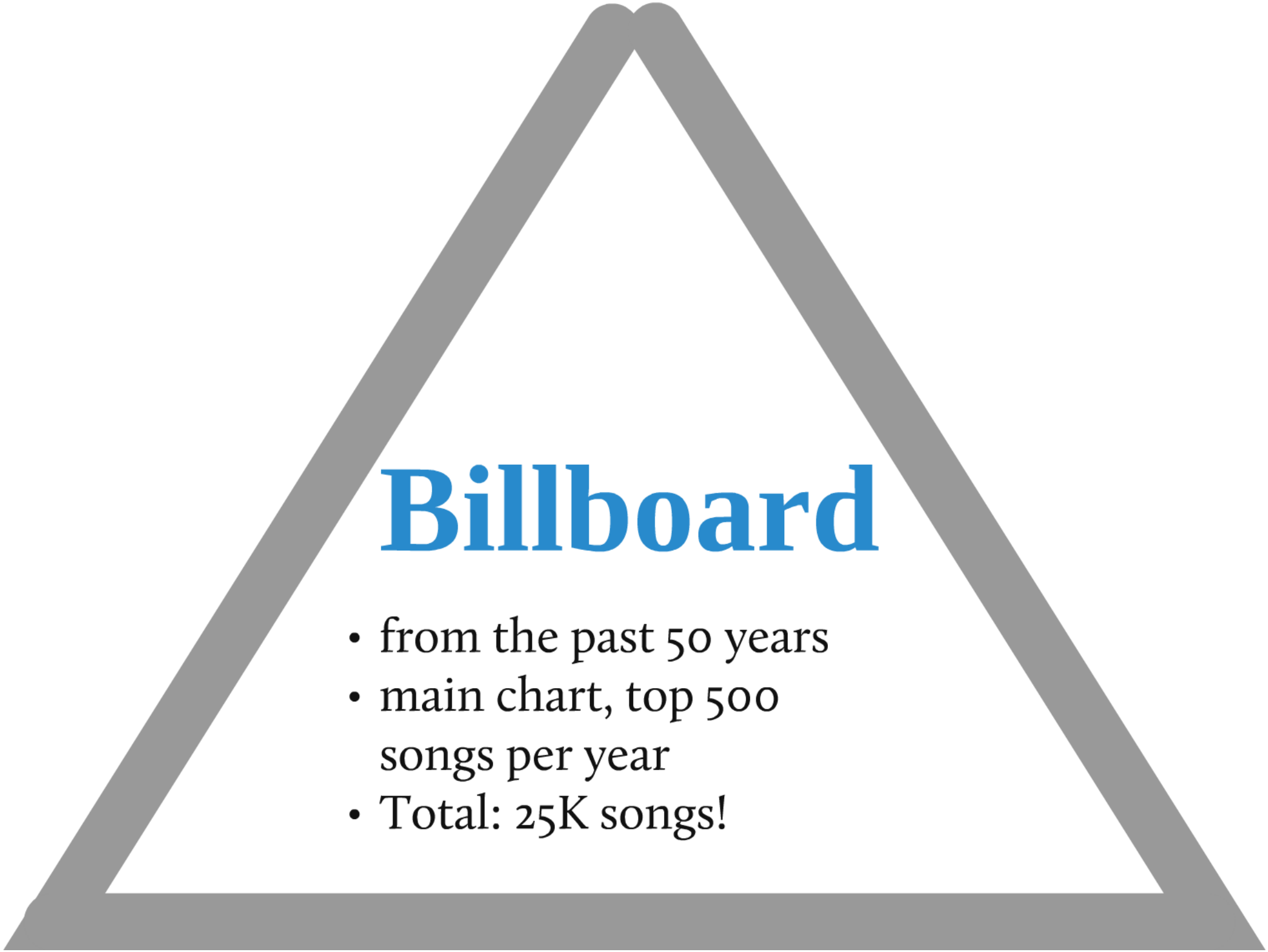


Selecting One Million Songs

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 CAL500)

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

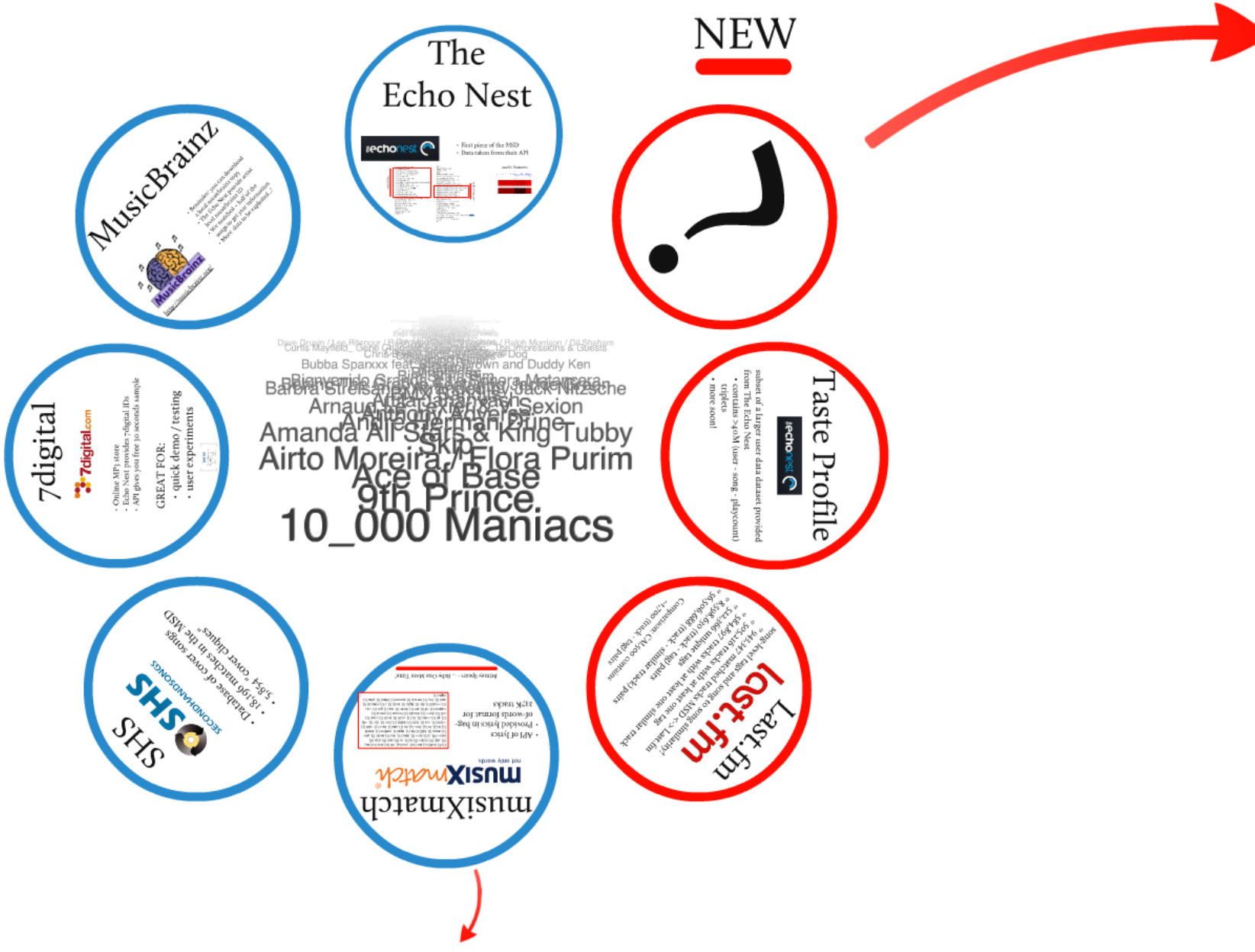
- popular artists -> as many songs as possible
- popular tags -> artists -> as many songs as possible
- random walk along similar artists

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

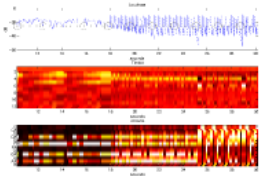
artist data

- analysis_sample_rate: 22050
- artist_7digitalid: 2200
- artist_familiarity: 0.7847042247II
- artist_hotttnesss: 0.623782610977
- artist_id: ARULZ741187B9AD2EF
- artist_latitude: nan
- artist_location: Tupelo, MS
- artist_longitude: nan
- artist_mbid: 01809552-4f87-45b0-afff-2c6f073023be
- 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: 6642e859886c025b0944b9420a82e8a
- 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

audio features

- 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

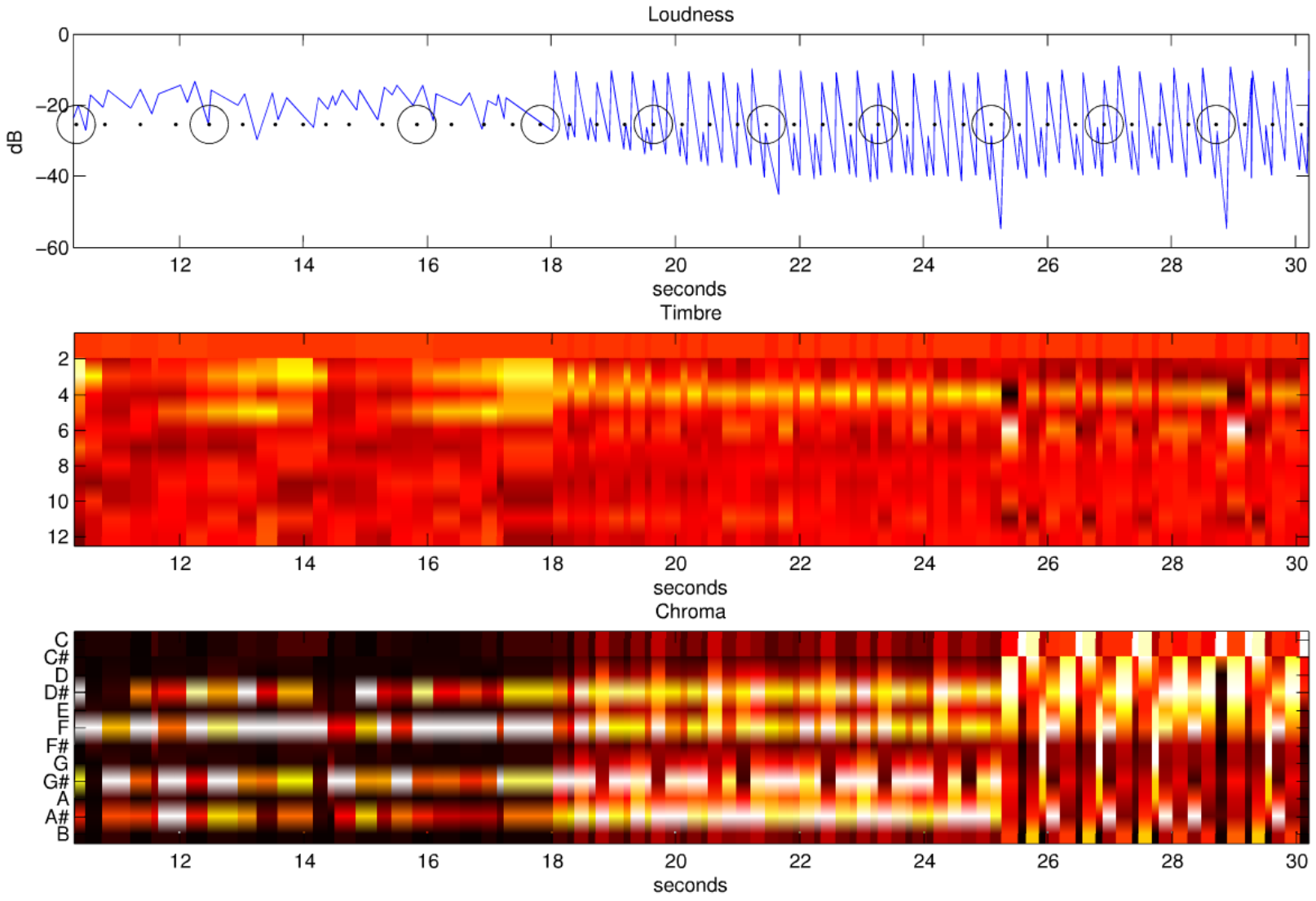


artist data

- analysis_sample_rate: 22050
- artist_7digitalid: 2200
- artist_familiarity: 0.847042247 11
- artist_hotttnesss: 0.623782610977
- artist_id: ARULZ741187B9AD2EF
- artist_latitude: nan
- artist_location: Tupelo, MS
- artist_longitude: nan
- artist_mbid: 01809552-4f87-45b0-afff-2c6f0730a3be
- 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

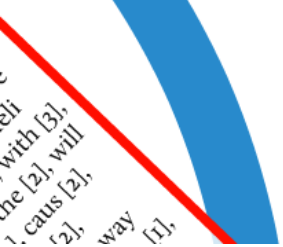
just in!

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

SHS

SECONDHANDSONGS

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



musiXmatch

musiXmatch®

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]

word cloud of "electronic" lyrics
(by Andrew Clegg, Last.fm)

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!

NEW

?

Taste Profile

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

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

Echo Nest

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

audio features

audio features

Future Improvements

More sources of data

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

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
- ...

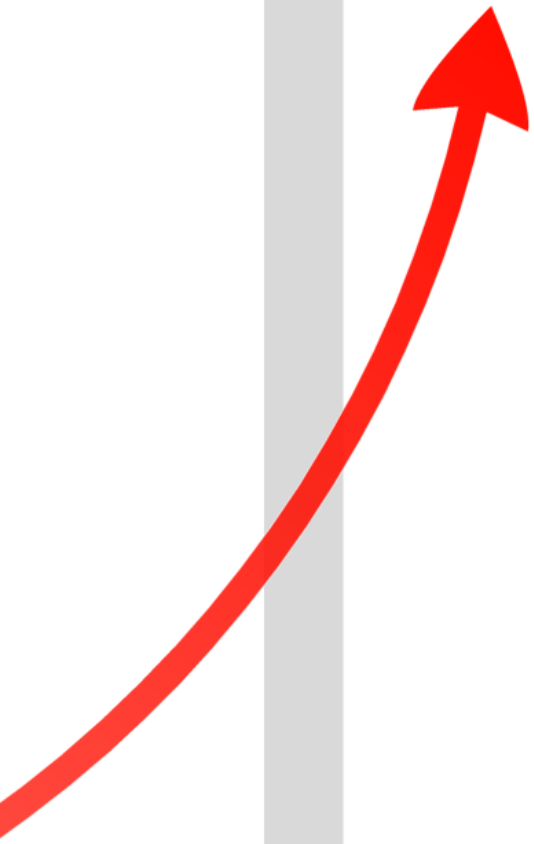
More sources of data

- 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...

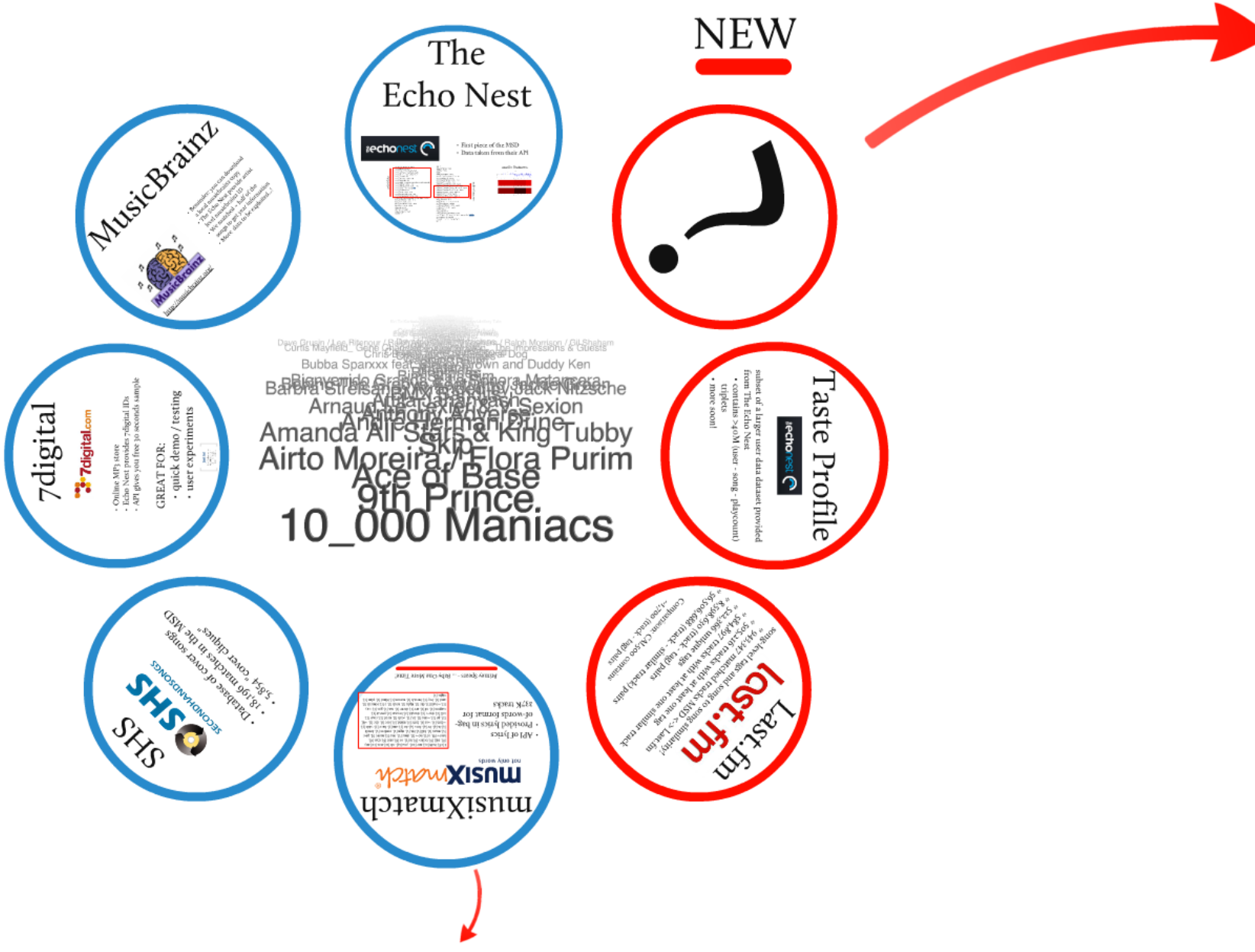
- not much active development planned
- we used most of our contacts

Anyone can contribute!

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

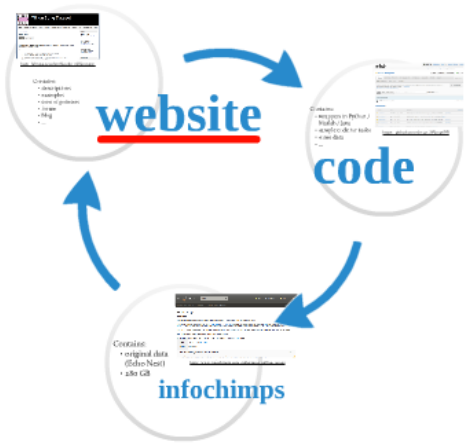
The Million Song Dataset

WHAT YOU NEED TO REMEMBER FROM THIS TALK



Some specifics / Getting started

resources



original data - HDF5

HDF5

one file per track

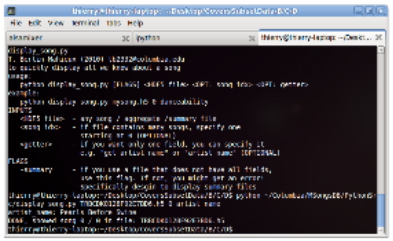
HDF5 format

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

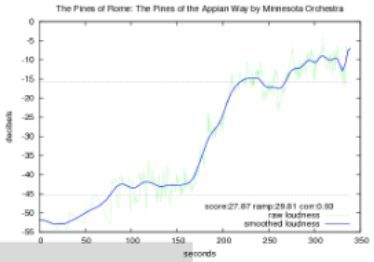
HDF5 is a format to contain and search large heterogeneous data developed by NASA... cool stuff!

1 Million Files!

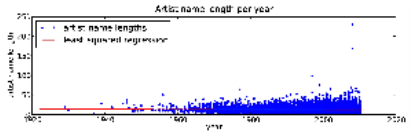
Files organized by Echo Nest track ID:
track "TRBCDKO128F92E7BD6"
in "/B/C/D/TRBCDKO128F92E7BD6.h5"



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

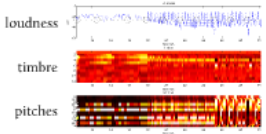


A quick MSD example: artist name length per year



they're getting smaller...???

Echo Nest Audio Features



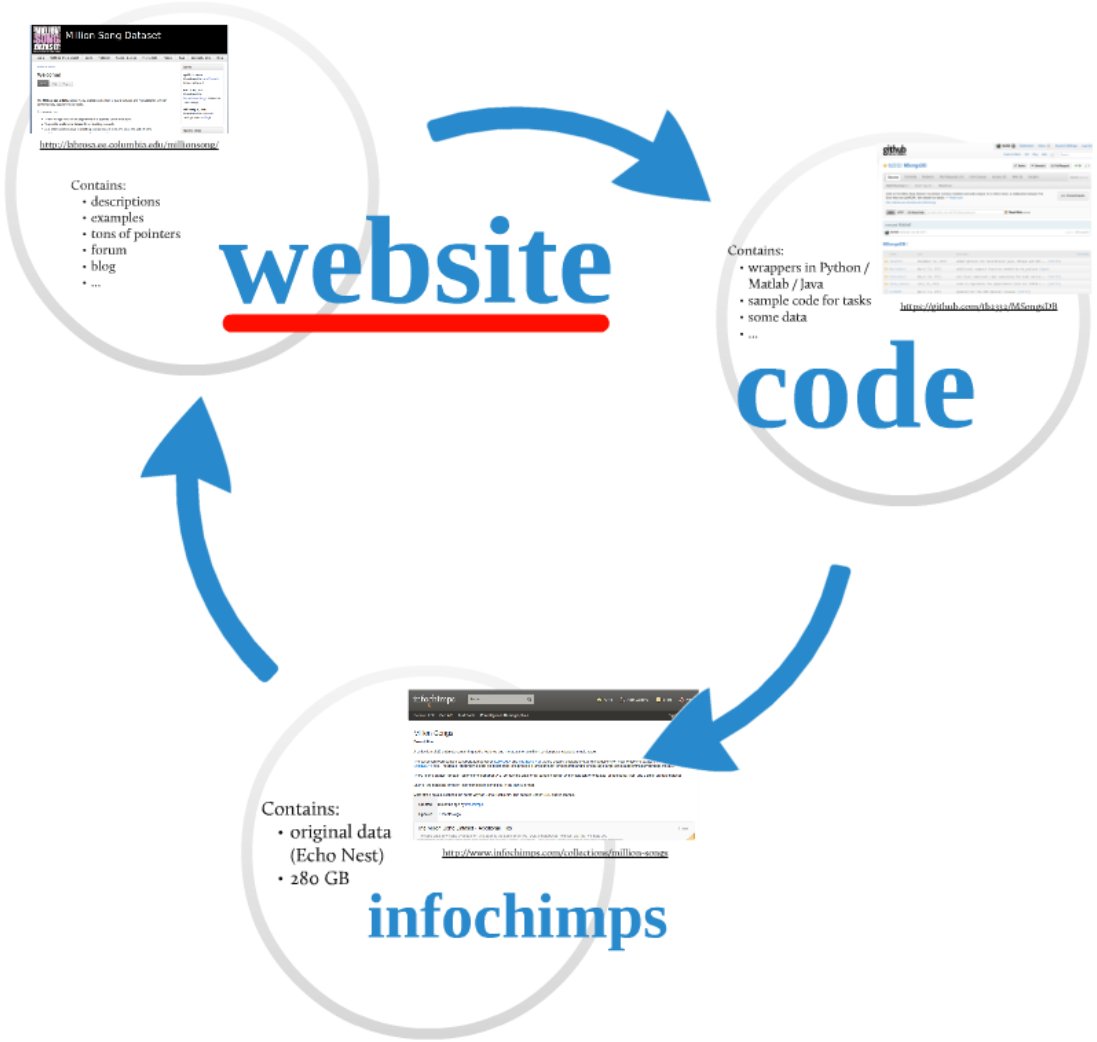
timbre and pitch information on the segment level (~ between 2 note onsets), plus approximation of the beats, bars, sections

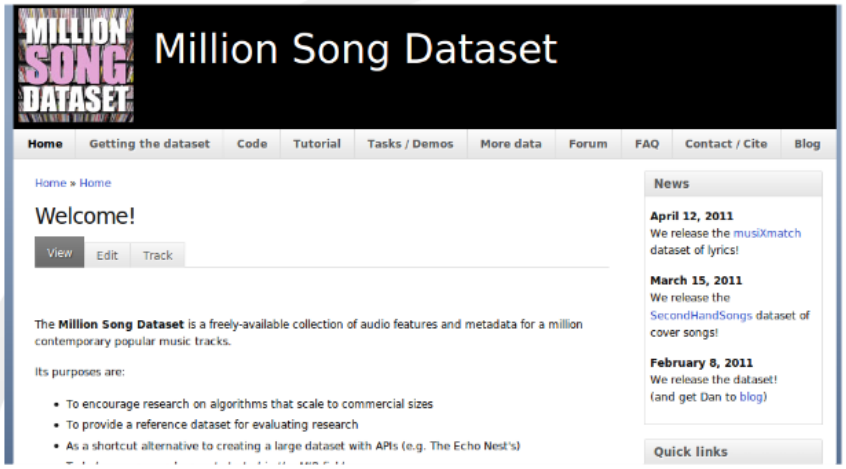
Start with the subset!

10K Songs

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

resources


Million Song Dataset
http://labrosa.ee.columbia.edu/millionsong/
Contains:
• descriptions
• examples
• tons of pointers
• forum
• blog
• ...
website
Contains:
• wrappers in Python / Matlab / Java
• sample code for tasks
• some data
• ...
https://github.com/tb1337/MSongsDB
code
Contains:
• original data (Echo Nest)
• 280 GB
http://www.infochimps.com/collections/million-songs
infochimps



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

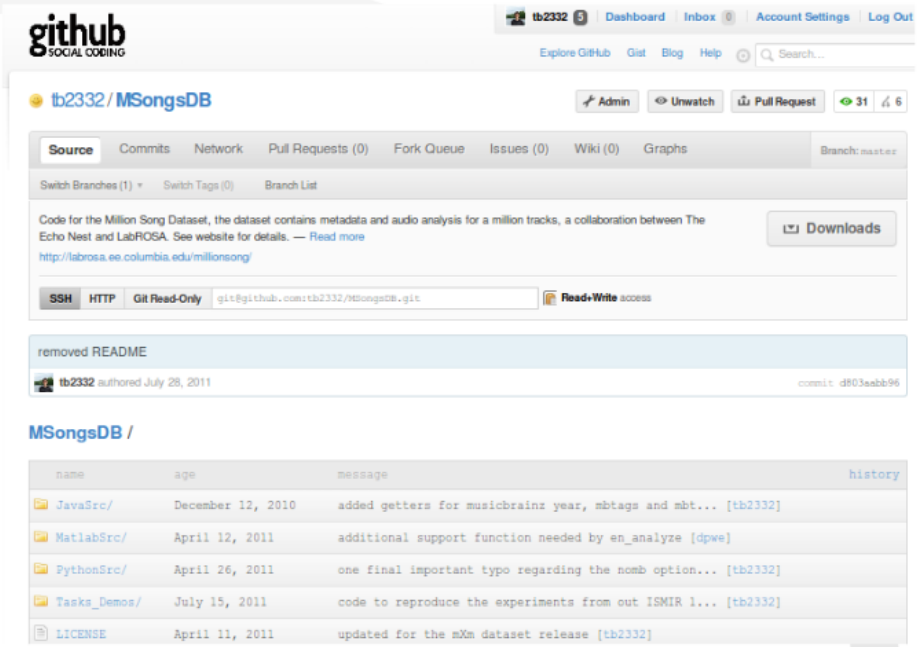
Contains:

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

web

Contains:

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

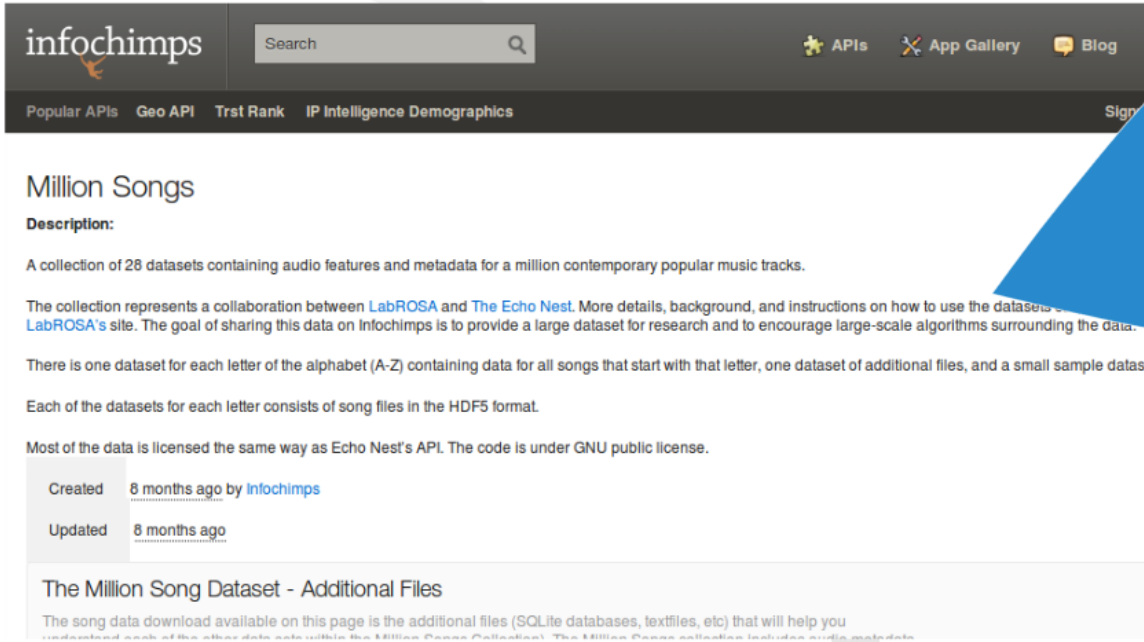


https://github.com/tb2332/MSongsDB

code

Contains:

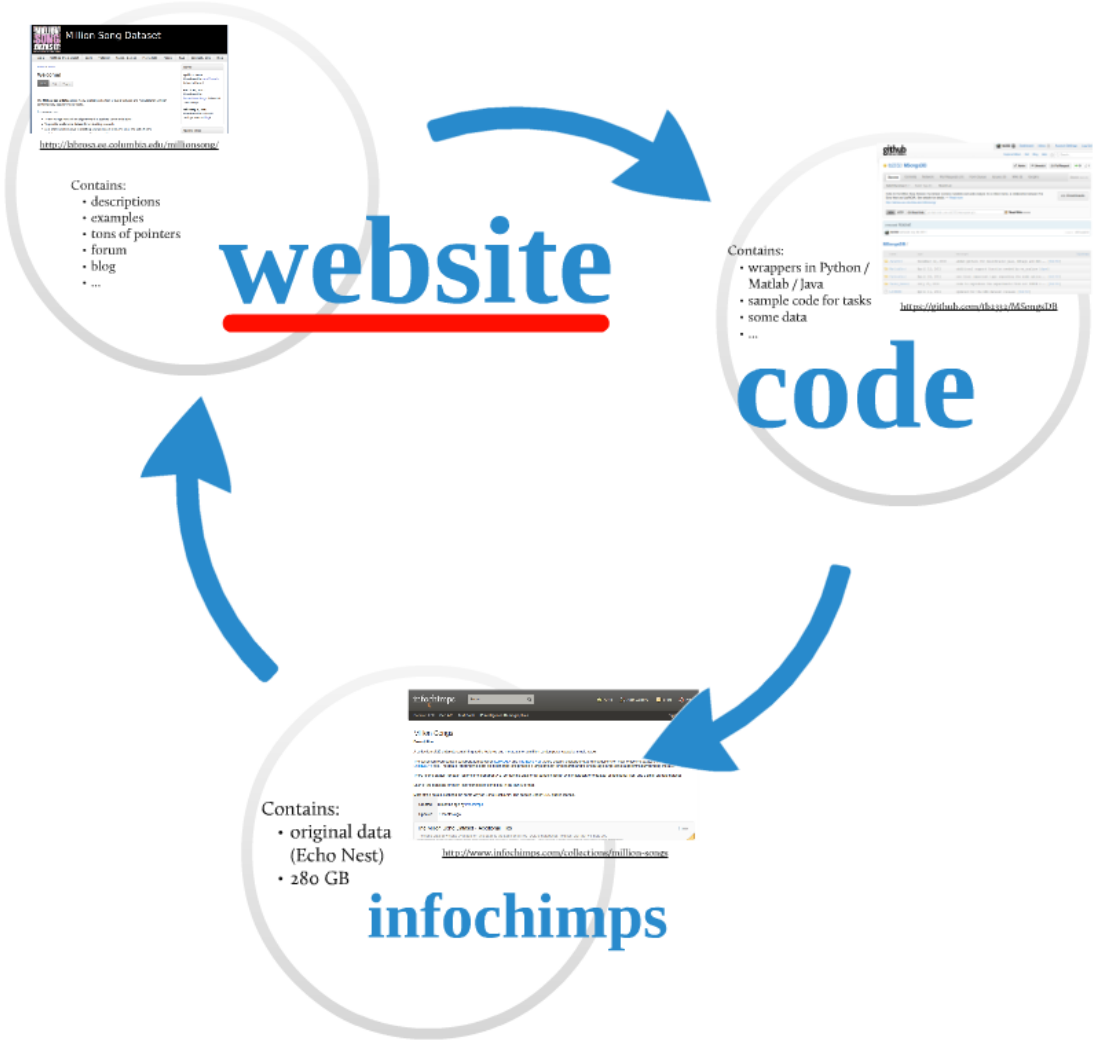
- original data (Echo Nest)
- 280 GB



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

infochimps

resources


Million Song Dataset
http://labrosa.ee.columbia.edu/millionsong/
Contains:
• descriptions
• examples
• tons of pointers
• forum
• blog
• ...
website
Contains:
• wrappers in Python / Matlab / Java
• sample code for tasks
• some data
• ...
https://github.com/tb1337/MSongsDB
code
Contains:
• original data (Echo Nest)
• 280 GB
http://www.infochimps.com/collections/million-songs
infochimps

original data - HDF5

HDF5

one file per track

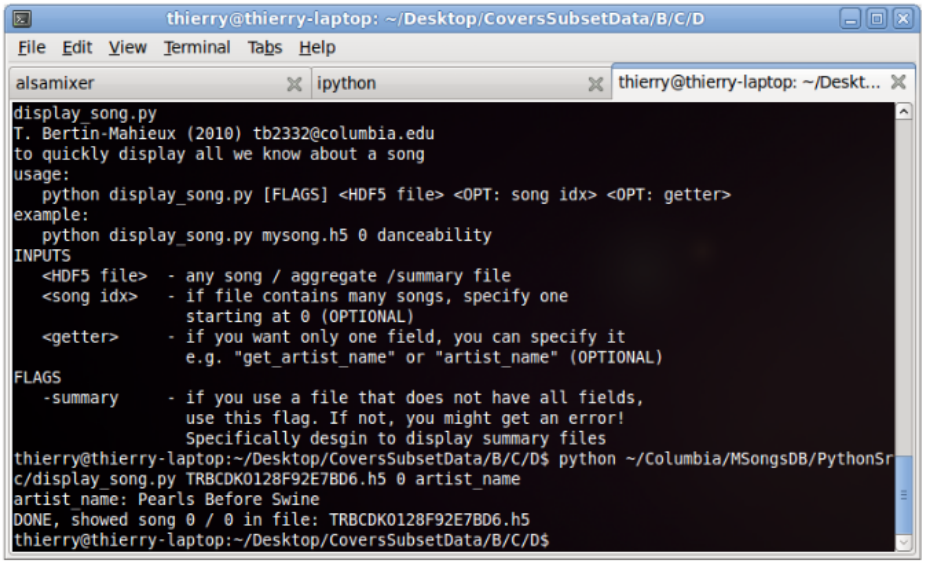
HDF5 format

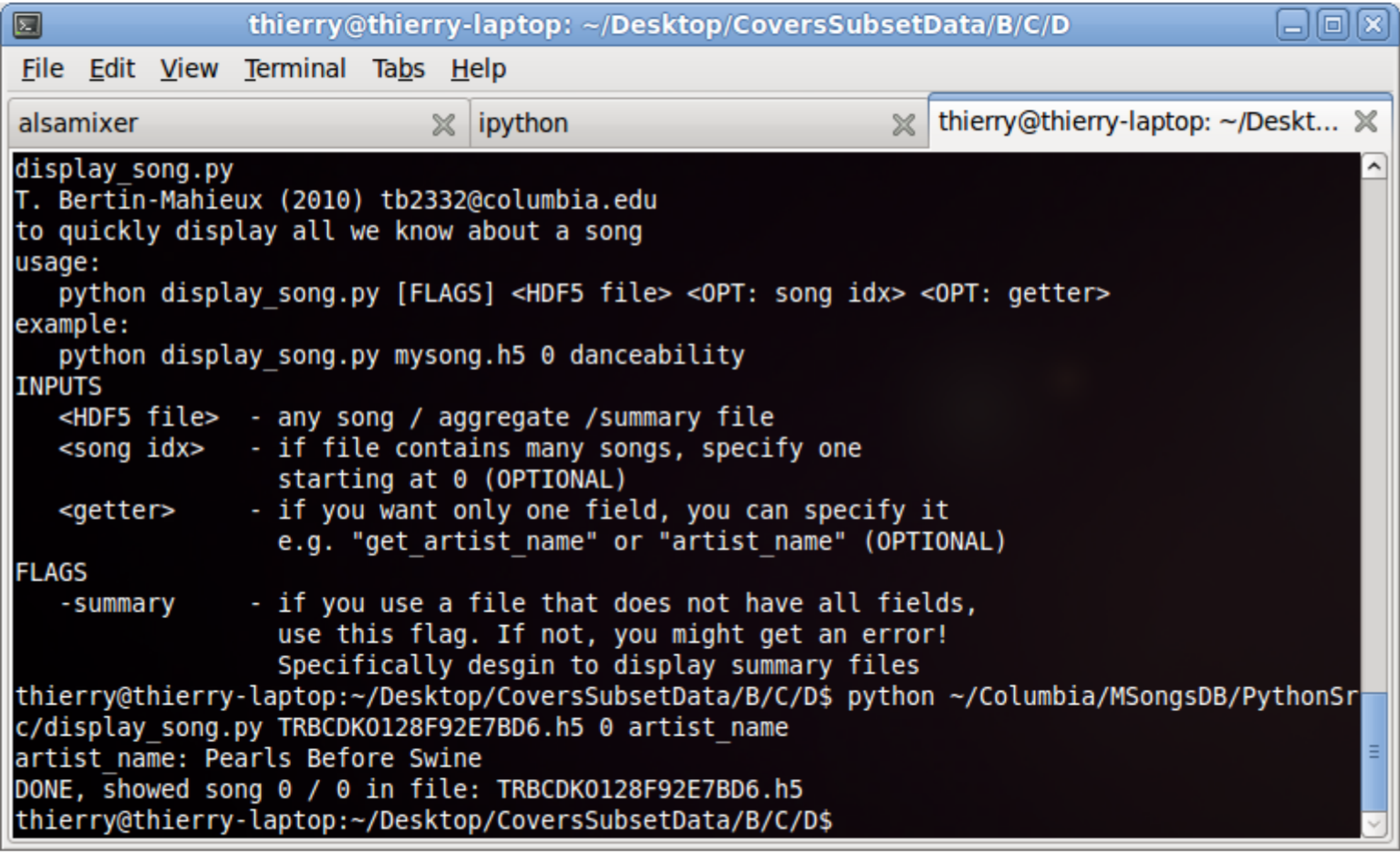
contains all Echo Nest data + 7digital ID + artist musicbrainz ID

1 Million Files!

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

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




thierry@thierry-laptop: ~/Desktop/CoversSubsetData/B/C/D
File Edit View Terminal Tabs Help
alsamixer
ipython
thierry@thierry-laptop: ~/Deskt...
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 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$

Start with the subset!

Complete dataset on its own

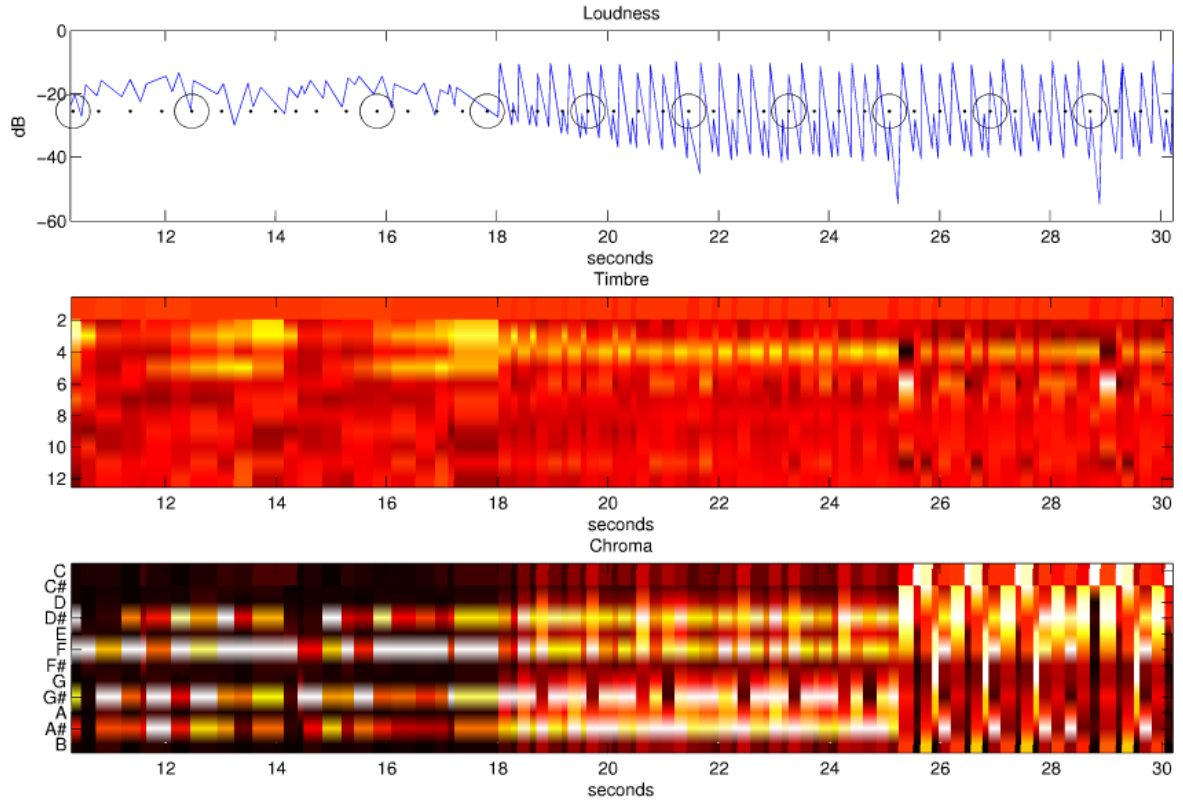
- 10K songs / HDF5 file
- small to download & play
- all from tagging training set

Echo Nest Audio Features

loudness

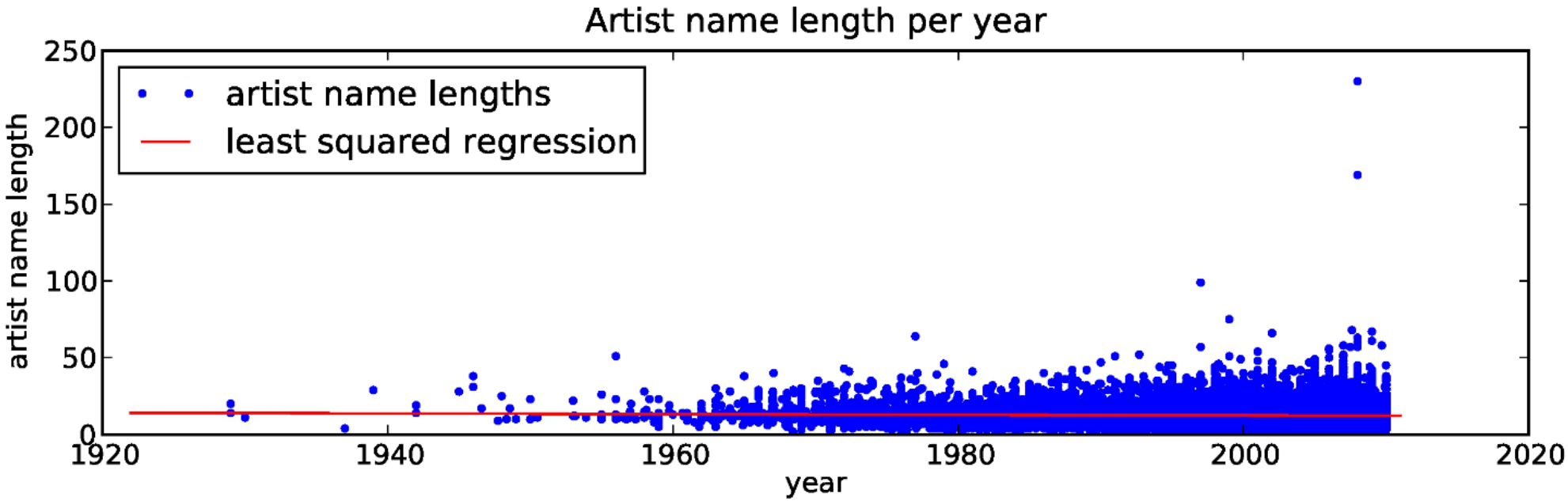
timbre

pitches



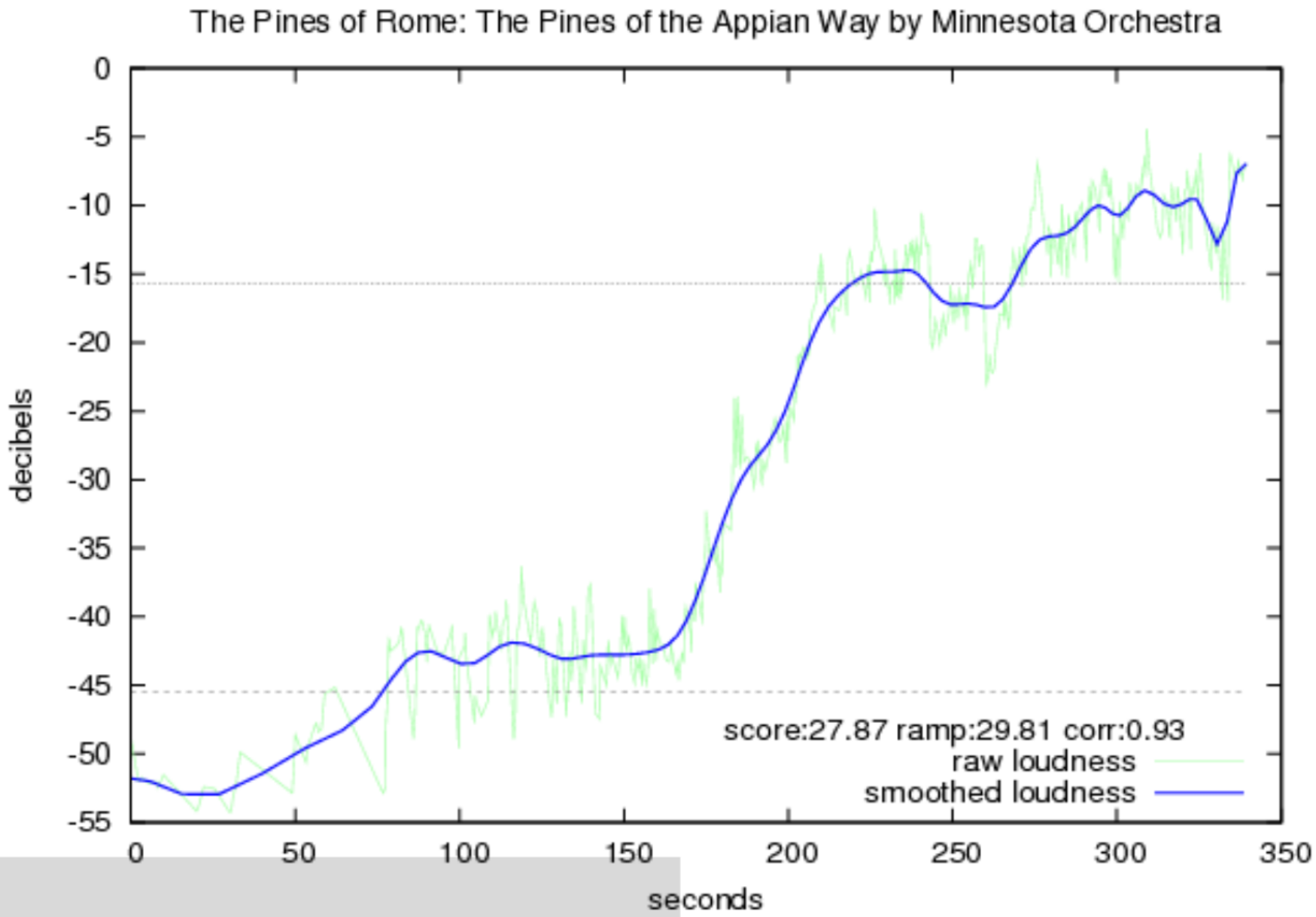
timbre and pitch information on the segment level (~ between 2 note onsets), plus approximation of the beats, bars, sections

A quick MSD example: artist name length per year



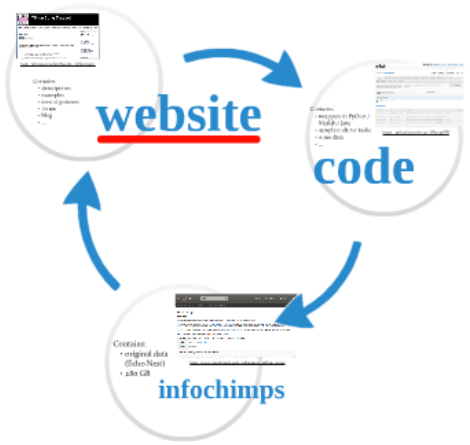
they're getting smaller...???

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

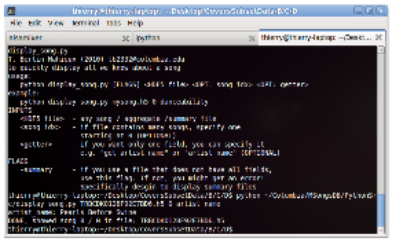
HDF5 format

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

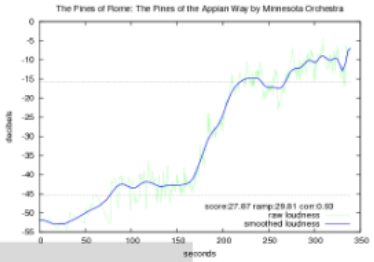
HDF5 is a format to contain and search large heterogeneous data developed by NASA... cool stuff!

1 Million Files!

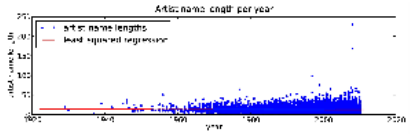
Files organized by Echo Nest track ID:
track "TRBCDKO128F92E7BD6"
in "/B/C/D/TRBCDKO128F92E7BD6.h5"



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

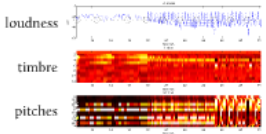


A quick MSD example: artist name length per year



they're getting smaller...???

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!

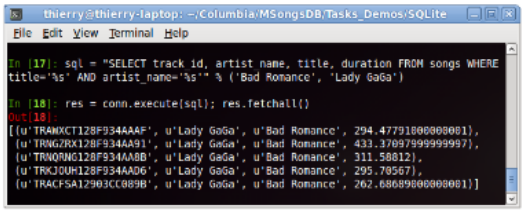
10K Songs

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)



and more ...

FORMAT

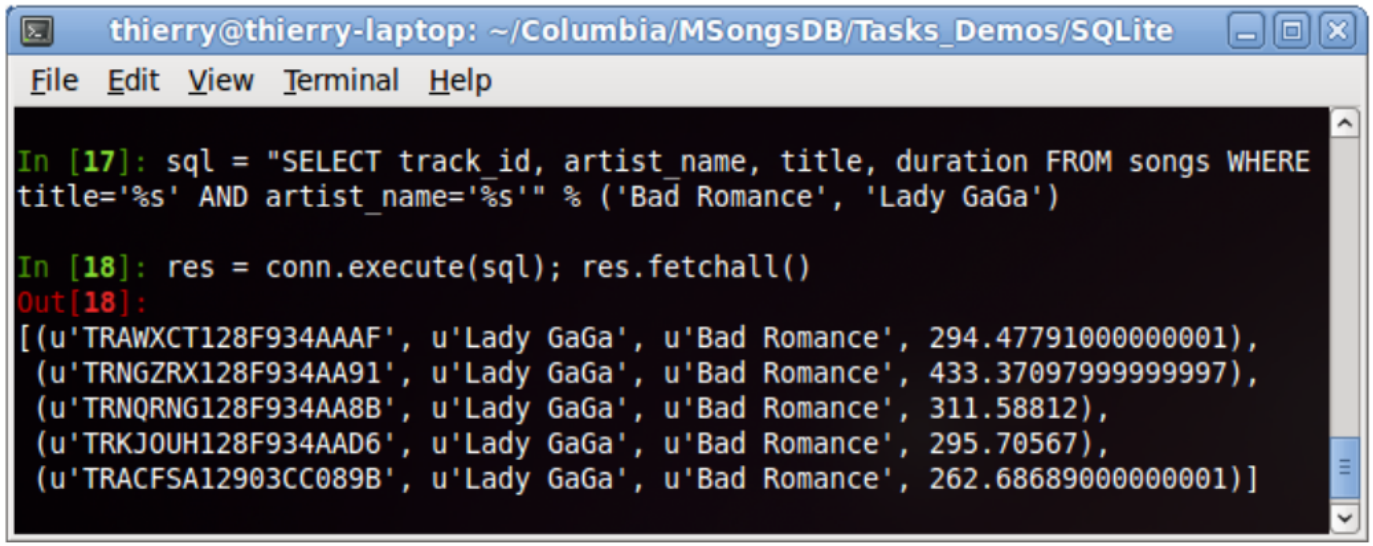
1M files? come on!

- requires this weird HDF5 library
- difficult to iterate over 1M 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)



FORMAT

1M files? come on!

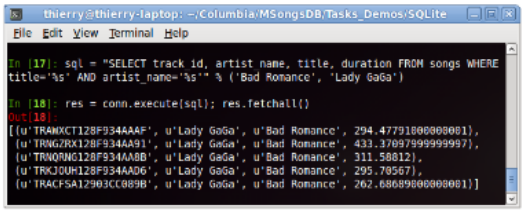
- requires this weird HDF5 library
- difficult to iterate over 1M songs
- 26 folders / tar files to download?!
- Duplicate information, e.g. artist data

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)



and more ...

FORMAT

1M files? come on!

- requires this weird HDF5 library
- difficult to iterate over 1M songs
- 26 folders / tar files to download?!
- Duplicate information, e.g. artist data

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

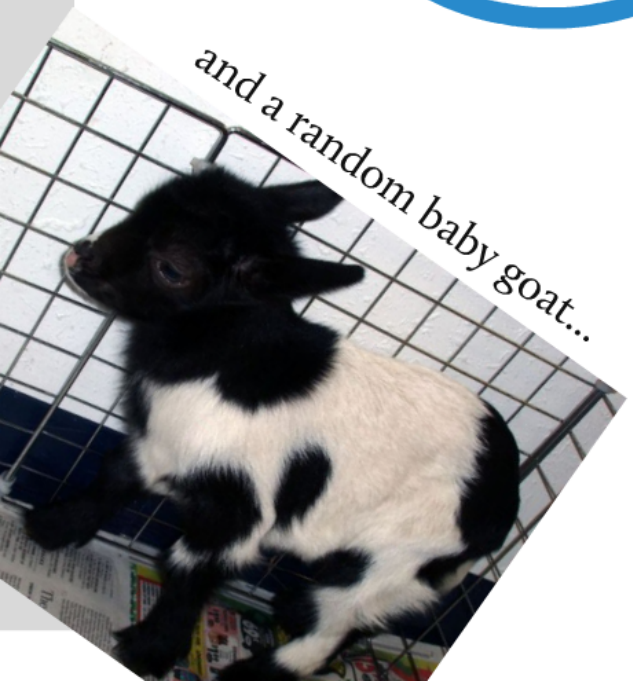
Some discussions we'd like to have

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 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, ...)



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?

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 communities

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, ...)

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, ...)

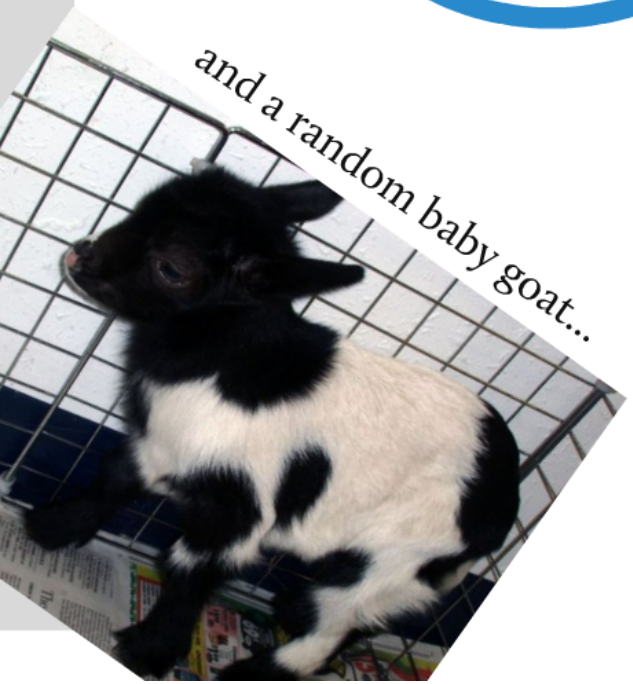
Some discussions we'd like to have

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 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, ...)



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?

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