Large-scale Music Data

Alastair Porter & Dmitry Bogdanov
March 1 2016

Dataset sizes in MIR

Genre example

- GTZAN Genre Collection: the most used genre dataset in MIR
  - 1000 track, 10 genres
  - **cross-validation accuracy: 75%**
  - **real accuracy (on 261K annotated tracks): 8%**
- Music Audio Benchmark Data Set
  - 1886 tracks, 9 genres
  - cross-validation accuracy: 60%
  - real accuracy (on 261K annotated tracks): 28%
- ROS dataset
  - 400 tracks, 8 genres
  - cross-validation accuracy: 88%
  - real accuracy (on 261K annotated tracks): 41%

Dataset sizes in MIR

GTZAN confusion matrix

cross-evaluation on 1K tracks:

Accuracy: 75.528701%

Predicted (%)												
	blu	**cla**	**cou**	**dis**	**hip**	**jaz**	**met**	**pop**	**reg**	**roc**		**Proportion**
blu	78.00	1.00	8.00	3.00	1.00	1.00	3.00	0.00	2.00	3.00	**blu**	10.07 %
cla	2.15	92.47	1.08	2.15	0.00	0.00	0.00	1.08	0.00	1.08	**cla**	9.37 %
cou	1.00	1.00	78.00	7.00	0.00	2.00	0.00	4.00	3.00	4.00	**cou**	10.07 %
dis	0.00	1.00	5.00	71.00	3.00	1.00	2.00	7.00	5.00	5.00	**dis**	10.07 %
hip	2.00	1.00	0.00	6.00	73.00	0.00	3.00	3.00	11.00	1.00	**hip**	10.07 %
jaz	7.00	4.00	4.00	3.00	1.00	79.00	0.00	1.00	1.00	0.00	**jaz**	10.07 %
met	2.00	0.00	0.00	1.00	3.00	2.00	86.00	1.00	0.00	5.00	**met**	10.07 %
pop	0.00	1.00	6.00	6.00	5.00	0.00	0.00	75.00	4.00	3.00	**pop**	10.07 %
reg	3.00	2.00	4.00	4.00	11.00	2.00	0.00	5.00	64.00	5.00	**reg**	10.07 %
roc	7.00	2.00	6.00	10.00	3.00	2.00	4.00	2.00	4.00	60.00	**roc**	10.07 %

Actual (%)

on 261K tracks:

Ground-truth genre	size (%)	Estimated genre: blues	classical	country	hip hop	jazz	pop	rock	ska
blues	2.7	**1.70**	2.15	0.44	2.10	89.48	0.37	3.53	0.23
classical	2.0	4.16	**8.59**	0.07	2.23	73.48	2.07	8.75	0.67
country	5.0	0.94	0.64	**1.38**	1.61	92.63	0.30	2.20	0.31
hip hop	2.4	0.40	0.17	0.02	**1.78**	92.43	1.38	3.69	0.14
jazz	7.3	1.79	3.51	0.23	2.67	**84.95**	1.00	5.37	0.50
pop	5.7	0.38	0.43	0.14	1.47	93.95	**0.60**	2.91	0.13
rock	43.9	0.35	0.13	0.04	1.12	95.40	0.35	**2.60**	0.02
ska	0.6	0.44	0.05	0.00	3.13	85.78	3.13	6.70	**0.77**

Problems with datasets in MIR

- Legal issues with sharing audio data
  - No public audio collections larger than few thousand of tracks are currently available in MIR
- Expert annotations are costly. Companies won’t publish their data.
- Pandora Music Genome Project
  - 450 characteristics of music
  - manually annotated by experts
  - ~1M tracks
- Researcher can’t build good enough datasets on their own
- Using external data sources for annotations
- Inconsistencies in annotations, noisy annotations

Last.fm

- Editorial metadata (artist, release, label, year)
- License: Proprietary
- Scrobbling history and statistics
  - To "**scrobble**" a song means that when you listen to it, the name of the song is sent to a Web site (for example, Last.fm) and added to your music profile.
  - User profiles with a history and statistics of scrobbles
  - Music statistics
- Collaborative tags (“folksonomy”) for artists, albums, and tracks
  - Tags for genres, moods, uses of music, …, weighted by frequency
- Music similarity (similar artists, tracks)
- Frequently used in MIR research as a “dirty” source of semantic annotations
- API: http://www.last.fm/api

Discogs

- Collaborative editorial metadata
- License: Public Domain
- Releases, artists, labels, genres, styles, release country and year, relations between artists, artist credits
    - 6M releases by 3.9M artists, across over 743K labels, contributed from nearly 238K contributor user accounts
- User ratings, reviews, and lists
- Rarely used in MIR research
- API: https://www.discogs.com/developers/

AllMusic

- Expert editorial metadata
- License: Proprietary
- Genre, styles, album moods, song moods, song themes
- Expert reviews and biographies
- Sometimes used in MIR research as a source for genre and mood annotations
- No API

RateYourMusic

- Collaborative editorial metadata
- License: Proprietary
- Releases, artists, labels, artist credits, artist relations
- 1M artists, 3M releases, 38M ratings, 89K labels, 1934K reviews, 470K users
- Album annotations:
  - Genres (based on user votes) http://rateyourmusic.com/rgenre/
  - Album music qualities (moods, …) https://rateyourmusic.com/music_descriptor/
- User ratings, reviews, and lists
- Never (?) used in MIR before
- No API

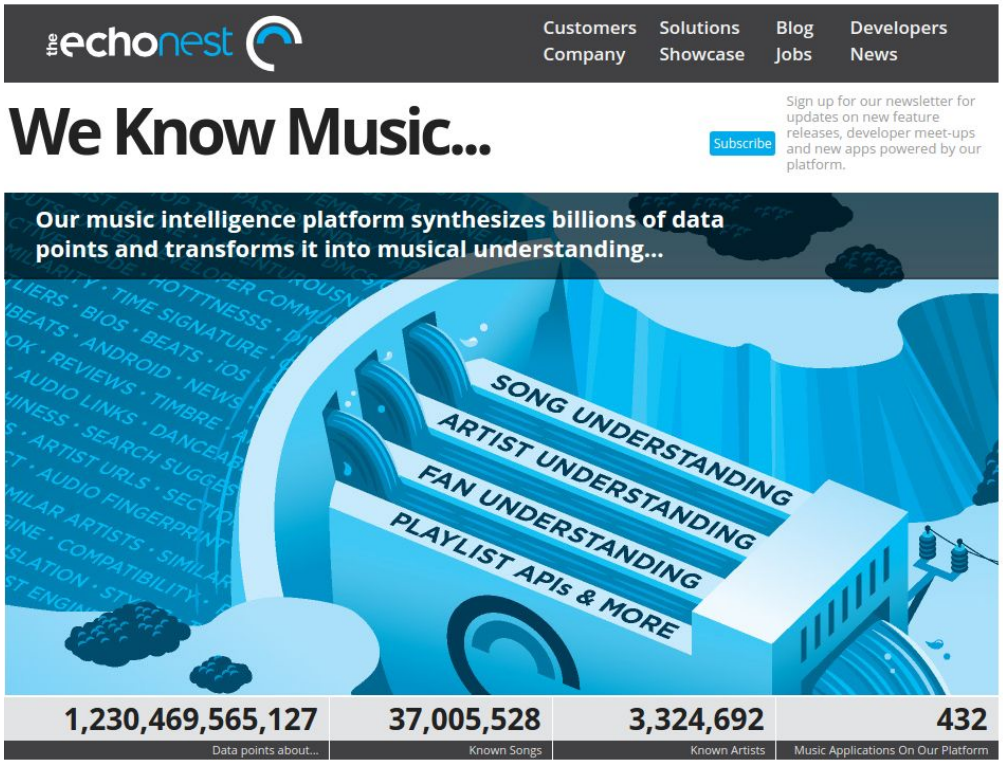
Wikipedia

- Wiki-based free general interest encyclopedia
- License: Public Domain
- Album pages
  - artist, year, label, artist credits, studio, tracklists
- Artist pages
  - biography, genres, instruments, labels, artist relations (associated acts)
- Loosely structured data
- Frequently used in MIR
- Has a structured data service called WikiData for storing information
- https://www.wikidata.org/wiki/Q11399, https://www.wikidata.org/wiki/Q2306

and more

https://musicbrainz.org/doc/Other_Databases

The Echo Nest (the.echonest.com)



Feature selection and knowledge

The Echo Nest contains both musical knowledge learned from how people talk about music, and also signal analysis to understand what music sounds like

Feature selection and knowledge

API Interface: http://developer.echonest.com/docs/v4/index.html

Million Song Dataset

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

Acoustic information for 1 million songs, computed with the Echo Nest's algorithms

Information about rhythm, tone, melody, chords

Million Song Dataset

Used by lots of researchers since it was published.

Why? Largest dataset of its kind

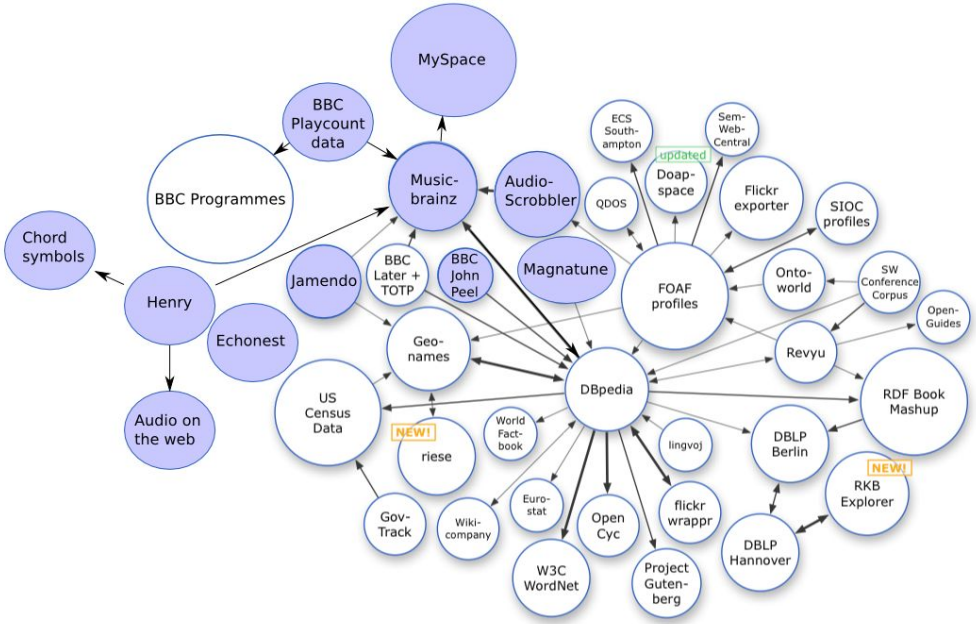
People create related datasets for matching to other data sources.

Cited by many other papers

https://scholar.google.es/scholar?q=related:PjIwLbOvMecJ:scholar.google.com/&hl=en&as_sdt=0,5

Music Ontology

The Music Ontology provides a model for publishing structured music-related data on your web site or through your API. (http://musicontology.com)



MusicBrainz http://musicbrainz.org

- Database of open information about music
  - People, Bands
  - Albums, Recordings
  - Compositions
  - Relationships (*Person* played *instrument* on *recording)*
  - Other interesting information - tags, external links, information about bootlegs, information about concerts, places, ...

MusicBrainz http://musicbrainz.org

- Data is added by *editors*, that is, anybody who wants to add the data
- Data is validated by other users and then approved
- Data is used all over the internet (Google, BBC, …)

MusicBrainz http://musicbrainz.org

Artists:	913,246
Releases:	1,397,658
Mediums:	1,560,427
Recordings:	14,411,759
Tracks:	17,659,503

MusicBrainz http://musicbrainz.org

All items in the database have a unique identifier

Other services on the internet know how to ask for MusicBrainz ids (this is important)

Getting data

- There is an API available to get all of the data from MusicBrainz
- Documented at http://musicbrainz.org/doc/Development/XML_Web_Service/Version_2
- Most people use a library to help them

Python client

https://github.com/alastair/python-musicbrainzngs

http://musicbrainz.org/recording/2bd0efdb-99e6-4d86-82b8-cdb098e7dd20

```
In [1]: import musicbrainzngs as mb

In [2]: mb.get_recording_by_id("2bd0efdb-
                               99e6-4d86-82b8-cdb098e7dd20")
Out[2]:
{'recording': {'id': '2bd0efdb-99e6-4d86-82b8-
cdb098e7dd20',
  'length': '233000',
  'title': 'Monsters in the Ballroom'}}
```

```
In [3]: result = mb.get_recording_by_id("2bd0efdb-99e6-
4d86-82b8-cdb098e7dd20", includes=["releases"])

In [4]: num_releases = len(result["recording"]["release-
list"])

In [5]: num_releases
Out[5]: 3

In [6]: print release["title"]
Siren Charms

In [7]: release = result["recording"]["release-list"][1]

In [8]: print release["release-event-list"][0]["date"]
2014-09-08
```

MetaBrainz

upf. Music Technology Group

AcousticBrainz

What is AcousticBrainz

An open database of audio features

Features are submitted by users running a feature

extractor on their music collections

A collaboration between MTG-UPF and the

MetaBrainz foundation (MusicBrainz)

http://acousticbrainz.org

Why do we need AcousticBrainz?

Much research requires access to features from audio, but

collecting audio can be hard

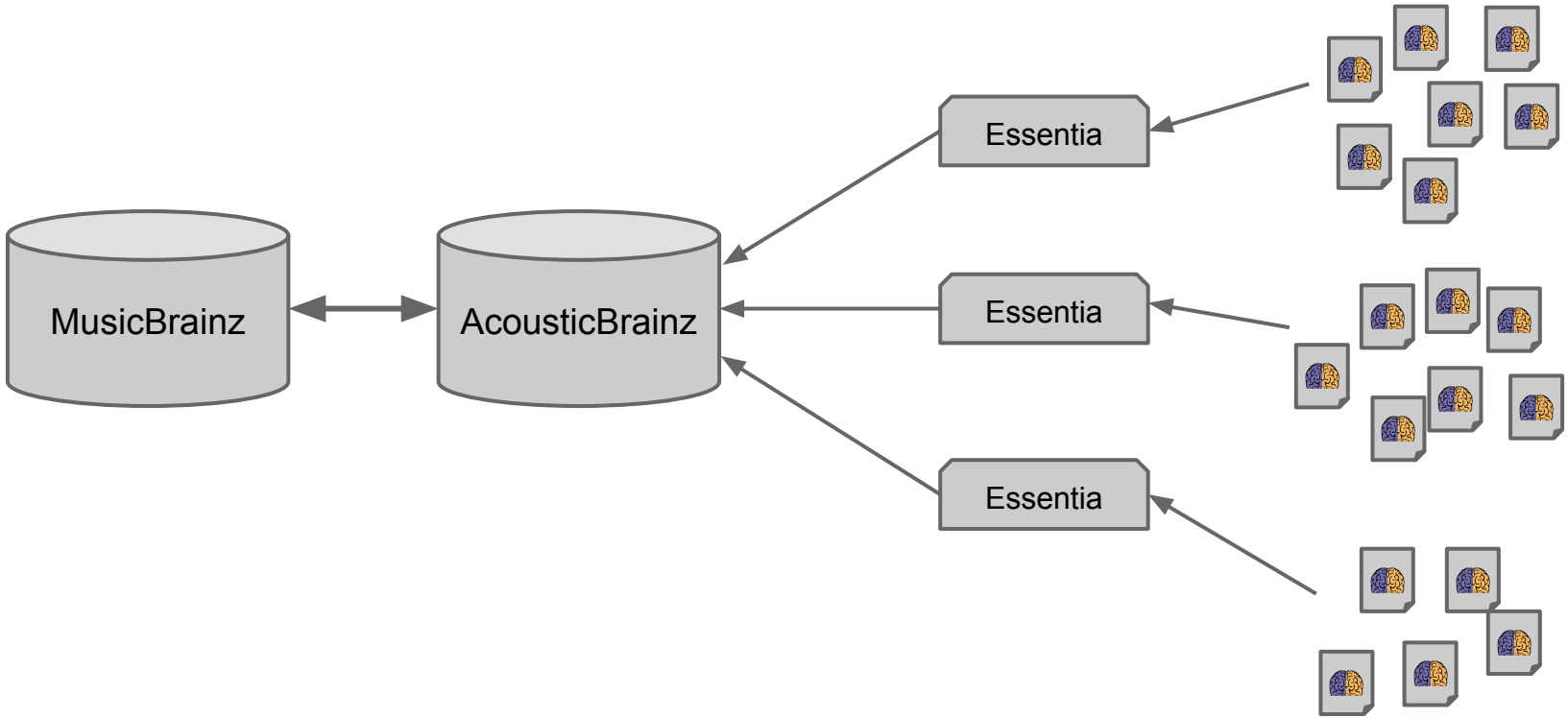
- Licensing
- Quantity
- Processing power

The need for another platform

The Million Song Dataset is great. It has contributed a lot to the MIR field

- Closed algorithms
- Has not been updated since it was released
- The EchoNest extractor has changed since the MSD was released

Crowd sourcing data



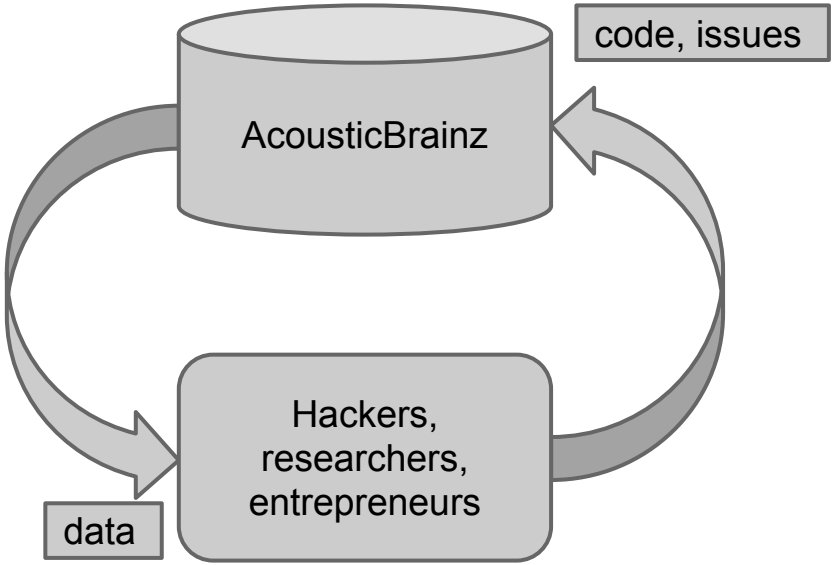
Feedback loop

Geeks use the AcousticBrainz data…
… and will find issues.

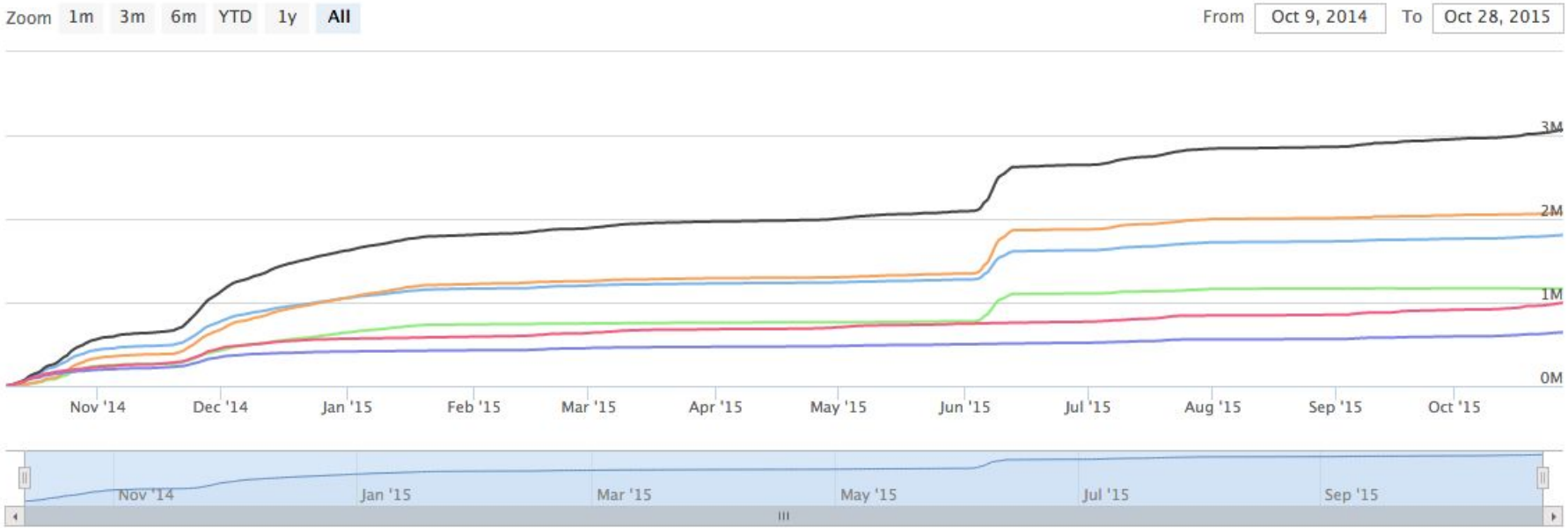
Geeks will come up with better ways of doing things…
… and will share code with AcousticBrainz.

AcousticBrainz will incorporate the best changes …
… and run the improved code over the data.

Geeks use the new and improved data…
… and will find more issues.



Statistics


Zoom
1m
3m
6m
YTD
1y
All
From
Oct 9, 2014
To
Oct 28, 2015
3M
2M
1M
0M
Nov '14
Dec '14
Jan '15
Feb '15
Mar '15
Apr '15
May '15
Jun '15
Jul '15
Aug '15
Sep '15
Oct '15
Nov '14
Jan '15
Mar '15
May '15
Jul '15
Sep '15

Statistics

In about 1.5 years:

- 3.5 million tracks of data
- almost 2 million unique tracks of data
- Low-level data: ~~33GB~~ >70GB
- High-level data: ~~3GB~~ ~10GB

Most important: **CC0 license**. (Free)

Statistics

Much self-reported metadata from MusicBrainz—tags, year, artists, instrumentation, genre

2,000,000 submitted files have genre reported

Top 20 genres represent 52.9% of these files, including

- Rock: 200,000
- Pop: 100,000
- Classical: 90,000
- Jazz: 89,000
- Soundtrack: 80,000

Top submitted tracks

The Beatles–I’m Down (129)
The Beatles–Slow Down (114)
The Beatles–Something (105)
The Beatles–Can’t Buy Me Love (103)
The Beatles–A Hard Day’s Night (103)
The Beatles–Octopus’s Garden (102)
The Beatles–Come Together (98)
The Beatles–And I Love Her (97)
The Beatles–Ticket to Ride (94)

Top submitted artists

The Beatles (26,616)

Johann Sebastian Bach (16,604)

Antonio Vivaldi (13,101)

Ludwig van Beethoven (13,090)

Wolfgang Amadeus Mozart (12,588)

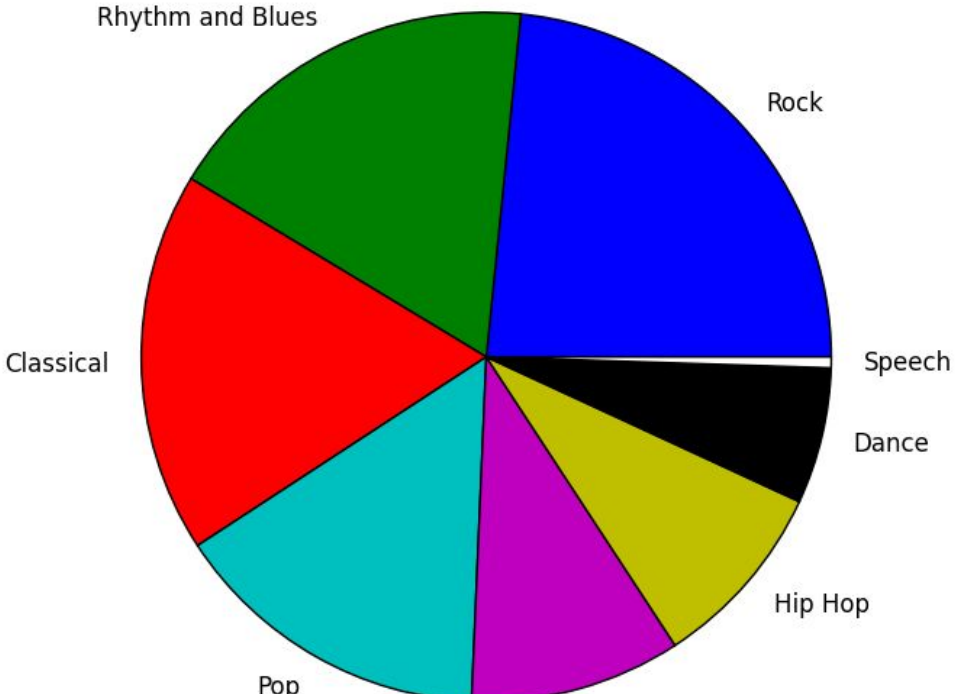
Jean Sibelius (10,667)

John Williams (9,027)

Bob Dylan (7,928)

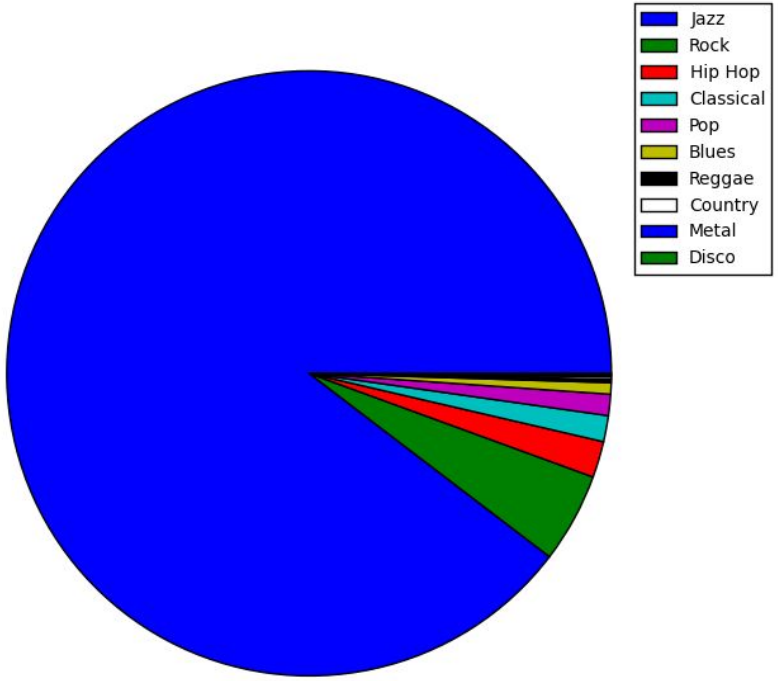
Extracted features

Genre, in-house dataset. 87% cross-validation accuracy

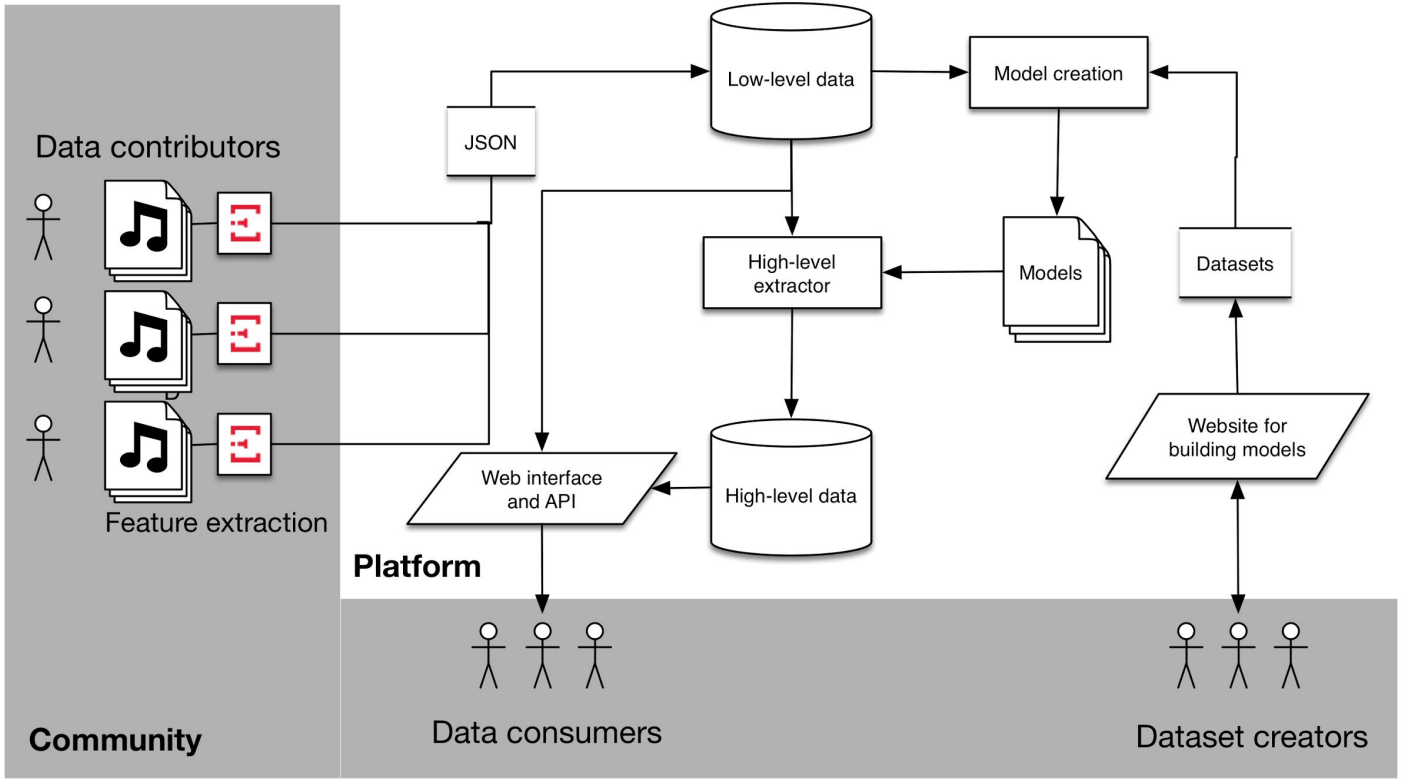


Extracted features

GTZAN. 76% cross-validation accuracy



Architecture



Feature extractor

Feature extractor based on Essentia

Users download the extractor and run it on their collection

runs at about 20x realtime

Audio must have MusicBrainz IDs (MBIDs)

Shared IDs

Remember how I said other things on the Internet know how to speak MusicBrainz ids?

So does AcousticBrainz

Data access

REST API to submit or download features by MBID

Duplicates (by MBID) are accepted

Feature data is tagged with the version number of the extractor

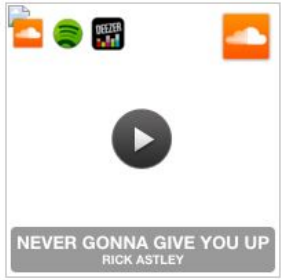
When we have new features, we release a new extractor, data slowly becomes up to date

Data dumps available in JSON or PostgreSQL

http://acousticbrainz.org/data

Recording "Never Gonna Give You Up" by Rick Astley

Metadata	value
MBID	98255a8c-017a-4bc7-8dd6-1fa36124572b
title	Never Gonna Give You Up
artist	Rick Astley
release	Absolute 80's, Volume 4: 86/87
track number	17 / 19
track length	03:32



Submission #1 out of 4 →

Low-level information Summary

Tonal & Rhythm	value
key	G#
scale	major
danceability	1.49614965916
bpm	113.307624817
beat count	399

Download data

For any MusicBrainz ID that you want to get data for:

http://acousticbrainz.org/2bd0efdb-99e6-4d86-82b8-cdb098e7dd20/low-level
http://acousticbrainz.org/2bd0efdb-99e6-4d86-82b8-cdb098e7dd20/high-level

How to get a MusicBrainz ID? Search the MusicBrainz API

Data

We generate two kinds of data

low level

- Submitted by users. Features computed from audio

high level

- Features inferred from low-level data using machine learning techniques.
- Can be recomputed without audio

http://essentia.upf.edu/documentation/streaming_extractor_music.html#music-descriptors

Low level features

Spectral features

MFCC, Bark/Mel/ERB bands, spectral energy, flux, dissonance, complexity

Time-domain / rhythmic

loudness, dynamics, onsets, beats, BPM, BPM histogram, danceability, beats loudness

Tonal

tuning frequency, chroma, key, scale, chords

High level features

Using Gaia, we generate SVM models for a number of music classification tasks, including genre, mood, instrumentation, timbre, rhythm

The data contains descriptors computed with these models, returning probability values for each class

Crowdsourcing datasets

Models are typically built in-house, or are small

We've seen that their accuracy isn't always that good

We are experimenting with letting other people generate models

Dataset "LastFm Dataset"

Evaluate | Edit | Delete

← Back to dataset list

A dataset built using tags from lastfm for ground truth

Author: alastairp
Creation time: Oct 24, 2015, 16:00 +02

View | Evaluation

Classes

avant-garde *753 recordings*	**classical** *4345 recordings*	**easylistening** *1158 recordings*	**rnb** *9190 recordings*
clatinamer *4118 recordings*	**jazz** *20524 recordings*	**rock** *104414 recordings*	**pop** *15855 recordings*
country *16449 recordings*	**hiphop** *5717 recordings*	**other** *224 recordings*	**ska** *1815 recordings*
asian *2338 recordings*	**african** *829 recordings*	**blues** *8106 recordings*	**comedy** *1300 recordings*
electronic *56098 recordings*	**folk** *8203 recordings*		

Dataset "LastFm Dataset"

Evaluate | Edit | Delete

← Back to dataset list

A dataset built using tags from lastfm for ground truth

Author: alastairp
Creation time: Oct 24, 2015, 16:00 +02

View | Evaluation

LastFm Dataset / classical

← Back to class list

MusicBrainz ID	Recording
424be4a4-4d7f-4d3a-af51-38f6af7eb580	Genloc - Markus Schneider
e87d087f-fce6-4343-947a-ea9d836dea23	Eden Roc - Ludovico Einaudi
315579fa-fd2a-42dd-b5f3-5e6ef8a0f04d	Hooked on Baroque - Royal Philharmonic Orchestra
d99a97cc-5c91-4197-a9fc-8080f418f6d7	Sabre Dance - Արամ Խաչատրյան
e07e8726-1036-4d5e-8dfe-509d9d084094	Air - 平野義久
f3a3facd-08e0-42d5-aa46-dc57b6d6cc12	Bilder einer Ausstellung: 2. Il Vecchio Castello - Chicago Symphony Orchestra, Carlo Maria Giulini
bfce96bf-8ac5-4038-b428-b82ef403ba6e	Summertime From Porgy and Bess - George Gershwin
91f19007-682c-4bb0-934a-2cbae7138d1e	Symphony No. 1 in C minor, Op. 68: III. Un poco allegretto e grazioso - Johannes Brahms
4c97e44b-5340-49d5-87de-1beb8b08a5dd	Für Sarah - Jim Steinman
8b51d893-b8ec-49bc-ac98-4f29000d3c04	Fantaisie in F minor, op. 49 - İdil Biret
e5db6ab7-513c-450c-a54f-c804b711f755	Carmen Suite No. 1: Intermezzo - Georges Bizet
77ac0592-2b1f-4141-a395-0bcfe19368fb	Symphony no. 2 in D major, op. 36: IV. Allegro molto - Wiener Philharmoniker, Christian Thielemann
6e656196-0b4b-4dde-80f9-318654336972	Romance no. 2 in F major, op. 50 - Gil Shaham, Orpheus Chamber Orchestra
effd36b9-c599-4cfa-bb95-4591a517bdf8	Farewell - Danny Elfman
2f629117-3d12-41da-badc-0b70ab2799f6	Cello Concerto in G minor, RV 417: I. Allegro - Jonathan Cohen, The King's Consort, Robert King

Job e0f3a2b9-a72a-44d4-89e6-68535d7cefe1

← Back to job list

Creation time: Sun, 25 Oct 2015 15:38:22 GMT

This evaluation job has been completed on Tue, 27 Oct 2015 09:01:36 GMT. You can find results below.

Accuracy: 43.54%

Predicted (%)

	asian	blues	classical	clatinamer	country	easylistening	electronic	folk	hiphop	jazz	pop	rnb	rock	ska		Proportion
asian	64.44	1.11	2.89	2.22	2.67	2.00	2.67	3.33	1.33	1.11	6.00	1.56	5.11	3.56	**asian**	7.14
blues	1.33	35.56	2.89	2.89	12.00	5.78	1.11	8.67	1.11	7.11	4.00	6.67	7.11	3.78	**blues**	7.14
classical	2.00	2.22	65.78	0.67	2.44	4.67	4.22	6.44	1.11	4.44	1.11	1.33	2.44	1.11	**classical**	7.14
clatinamer	3.33	3.56	0.89	32.22	5.33	2.00	4.22	4.89	8.89	6.67	4.22	4.44	2.89	16.44	**clatinamer**	7.14
country	2.00	6.00	2.00	2.44	53.11	4.44	0.44	9.78	0.22	2.89	6.67	5.11	3.56	1.33	**country**	7.14
easylistening	3.11	2.89	9.33	4.44	11.11	23.78	3.56	10.00	2.44	9.33	9.78	5.33	2.44	2.44	**easylistening**	7.14
electronic	3.11	0.89	5.78	2.22	0.22	1.78	52.00	3.33	8.44	2.22	2.89	1.11	8.22	7.78	**electronic**	7.14
folk	4.67	6.67	7.33	4.22	9.11	8.44	3.11	36.22	0.67	4.44	6.67	1.56	5.33	1.56	**folk**	7.14
hiphop	1.56	0.67	0.22	6.89	0.44	0.44	6.67	0.44	63.78	0.44	1.56	3.11	2.44	11.33	**hiphop**	7.14
jazz	3.56	7.56	9.11	6.00	3.11	8.22	3.11	4.67	2.89	37.33	3.33	6.67	1.78	2.67	**jazz**	7.14
pop	9.11	1.78	2.89	5.11	10.00	5.56	6.44	6.89	5.56	3.11	25.56	7.11	6.67	4.22	**pop**	7.14
rnb	5.56	5.78	0.89	5.56	8.22	8.22	4.00	4.22	6.67	4.22	10.89	27.56	3.11	5.11	**rnb**	7.14
rock	5.33	4.22	2.89	0.44	4.44	2.44	4.44	5.33	0.67	1.78	7.11	2.22	52.00	6.67	**rock**	7.14
ska	1.78	3.56	0.22	13.11	2.44	1.33	8.89	1.11	12.89	1.33	3.11	3.11	6.89	40.22	**ska**	7.14

Actual (%)

Contributions

People who want to contribute their audio can download our submission tools at http://acousticbrainz.org/download

The tool runs the AcousticBrainz essentia extractor on audio, creates a file with all of the data, and uploads it to the AcousticBrainz site

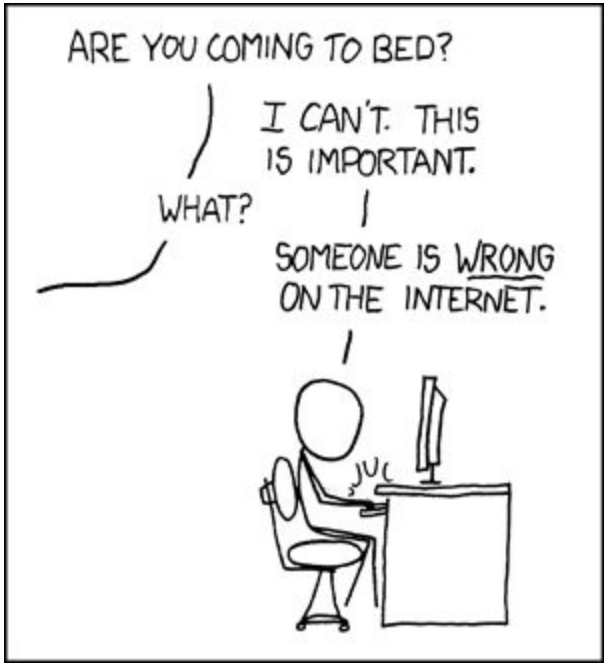
Data quality

A lot of the data needs work

We’re very early on in this project, we can’t expect it to be perfect.

Cunningham’s law—we’re going to use it!

Data quality



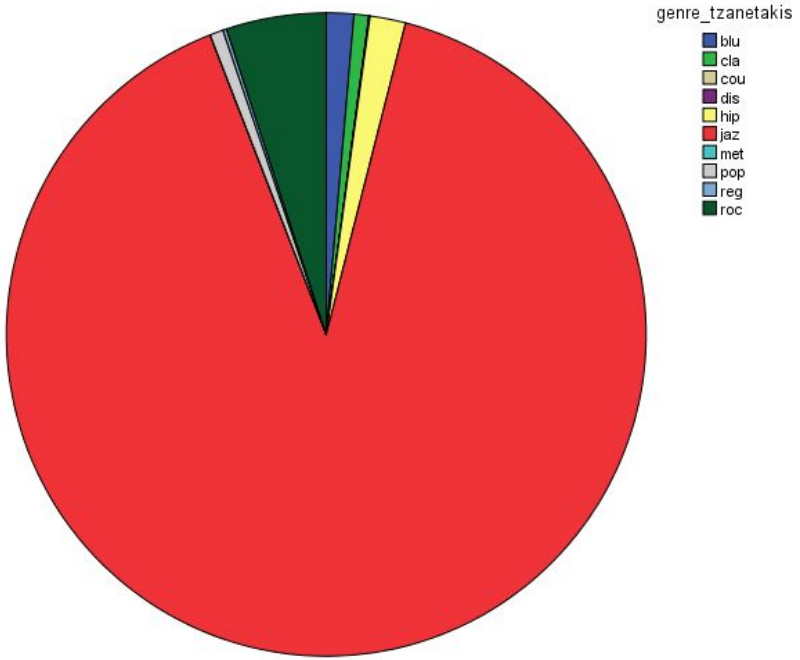
Preliminary results

Some interesting results from our initial 600k files

Emilia did this analysis a few months after we released the data

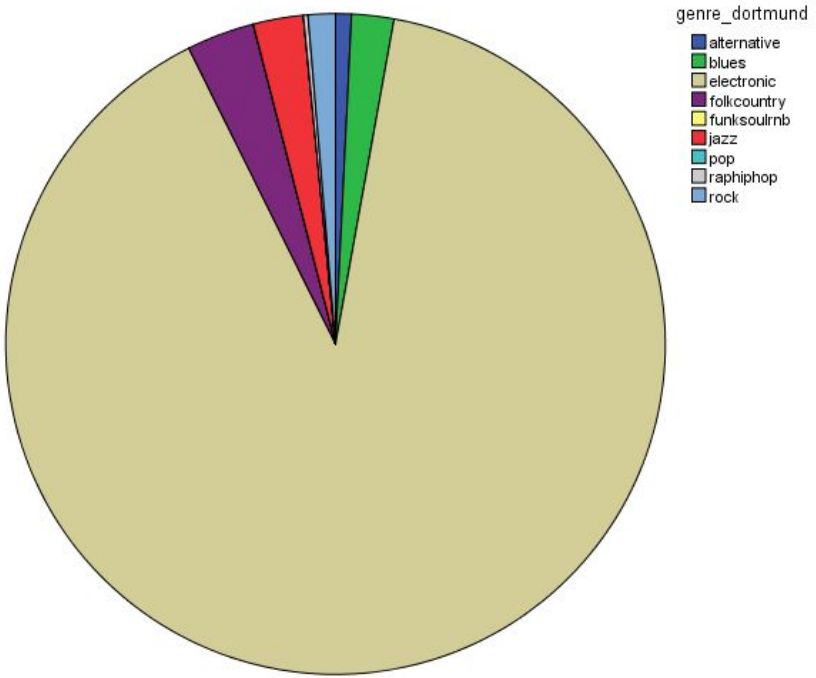
Genre distribution

The Tzanetakis model shows that almost all of the music in the collection is Jazz



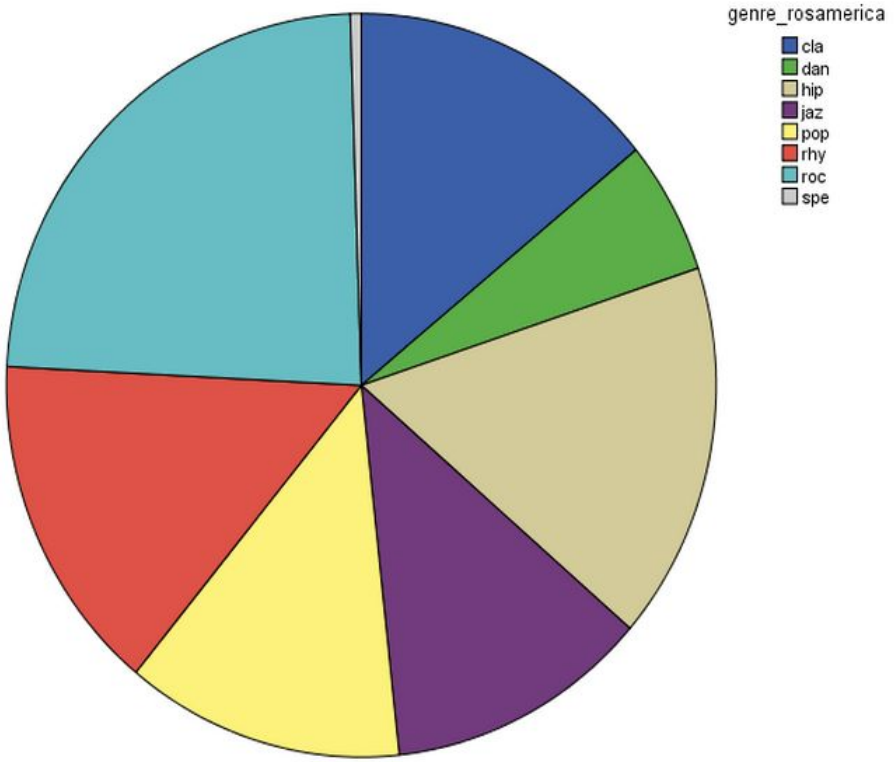
Genre distribution

Also, the Dortmund model shows that almost all the music is rap/hip hop (wut?)



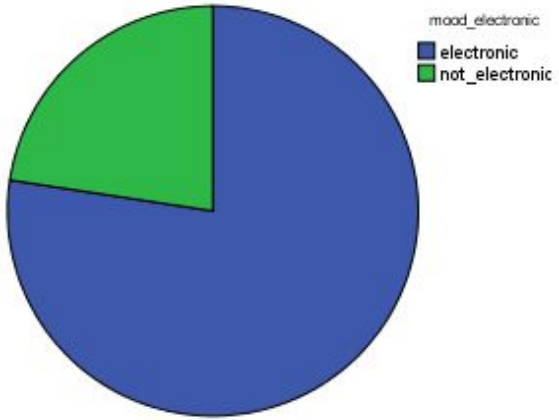
Genre distribution

The rosamerica genre training set seems to show a “nice” distribution, but...



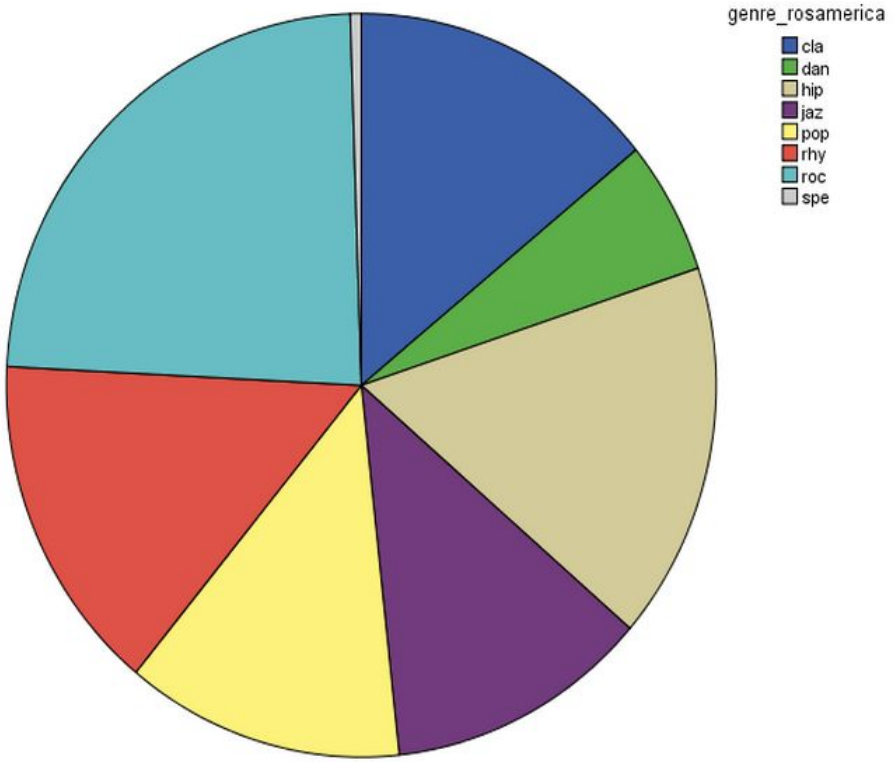
Genre distribution

We have a electronic mood model, which tells us if a song may be electronic (house, techno, trance, dnb, etc)



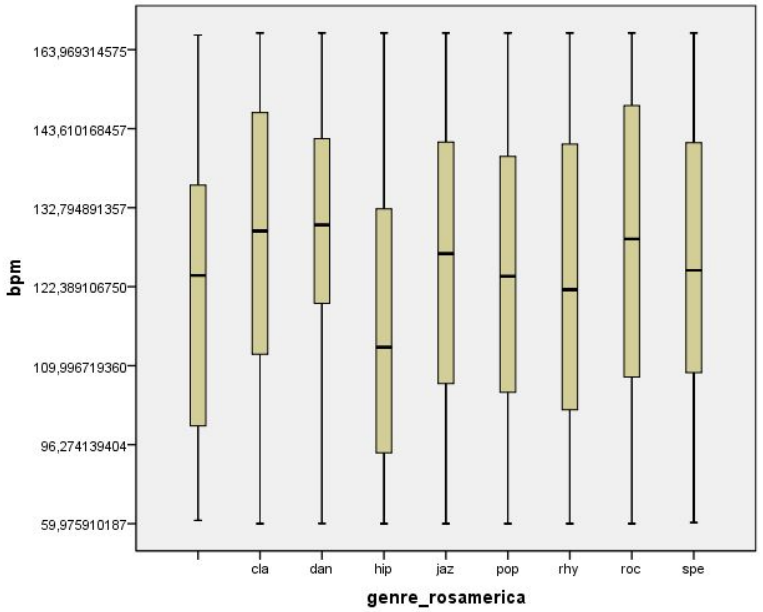
Genre distribution

And the rosamerica dataset doesn't have an "electronic" class



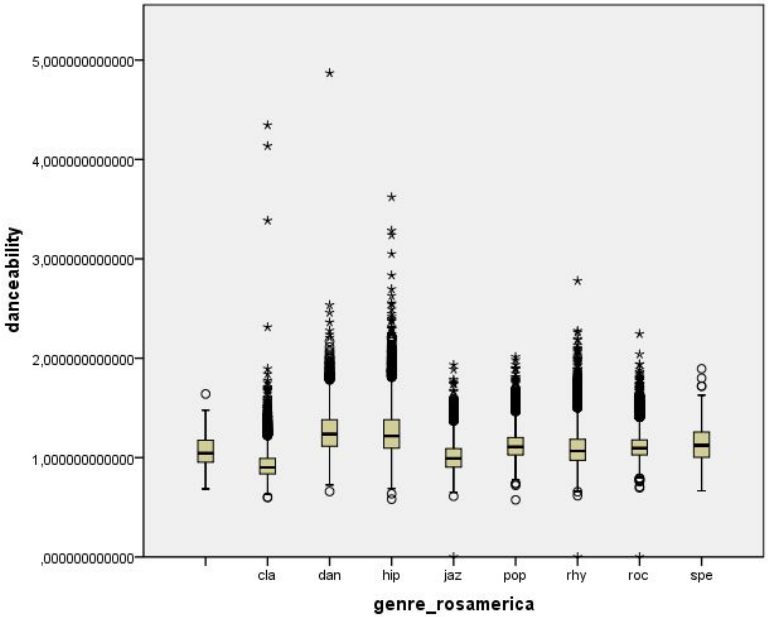
Genre-related information

Dance music tends to be faster, hip hop slower



Genre-related information

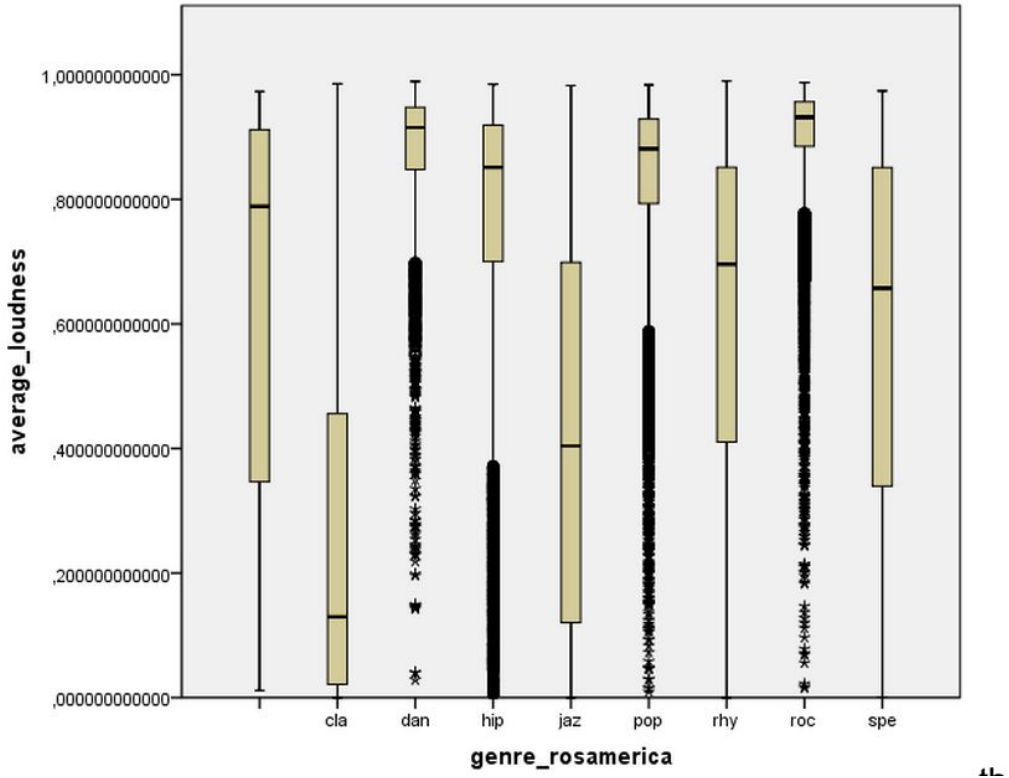
Hiphop and Dance music are the most “danceable” (have strong repetitive beats)



Loudness vs genre

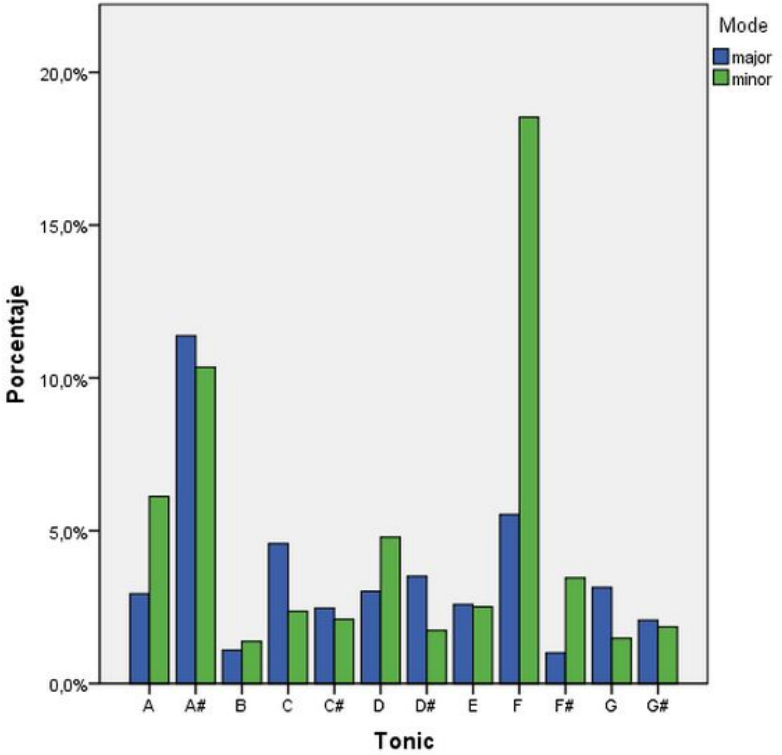
Classical & Jazz is quiet.

Rock, hip-hop, dance and pop are loud.



Key distribution

We wanted to automatically compute the musical key of files, but there's some problems with our detection

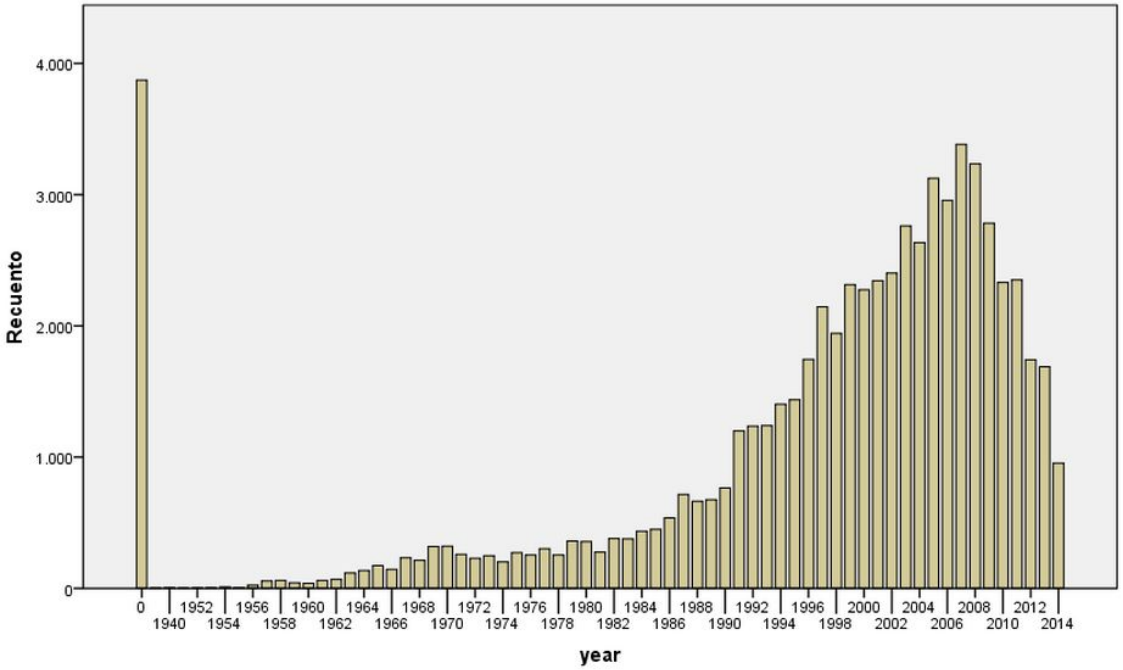


Year distribution

Clearly the music in AcousticBrainz is mostly from the last 30 years.

Is this a bias of our community or that there is much more music in the last 30 years?

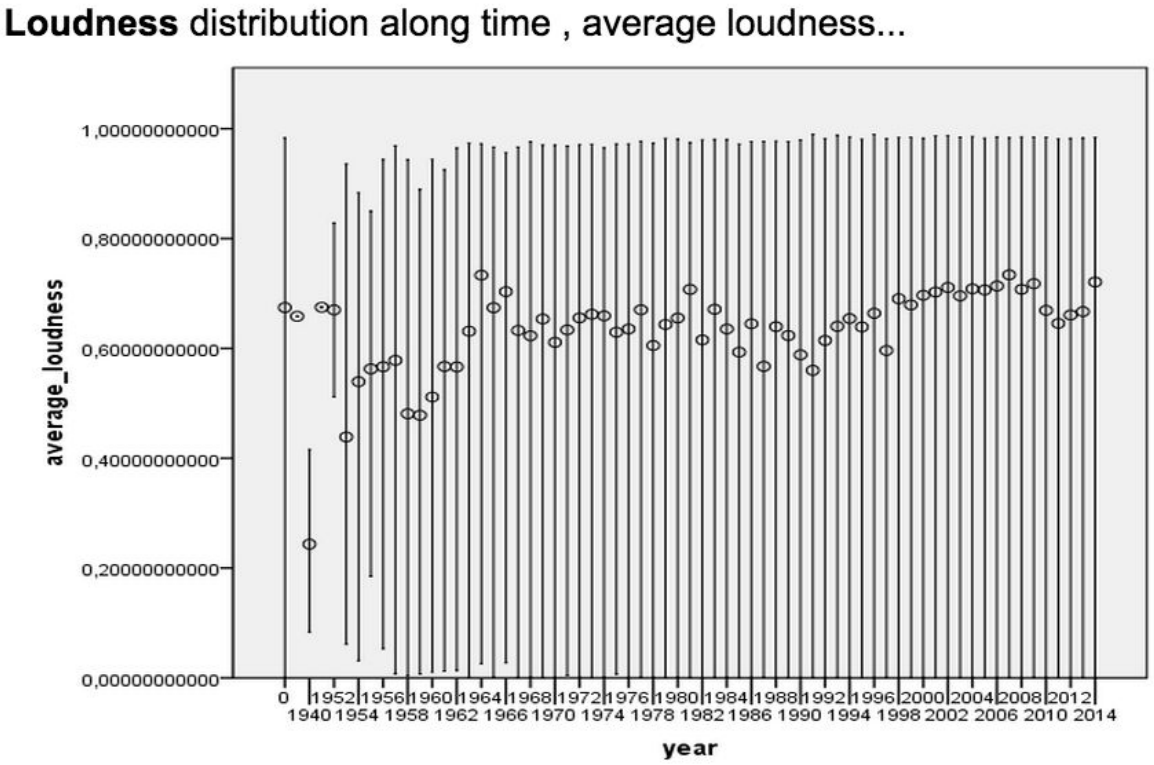
Probably both. Needs more data/analysis.



Loudness wars?

Is recorded music getting louder over time?

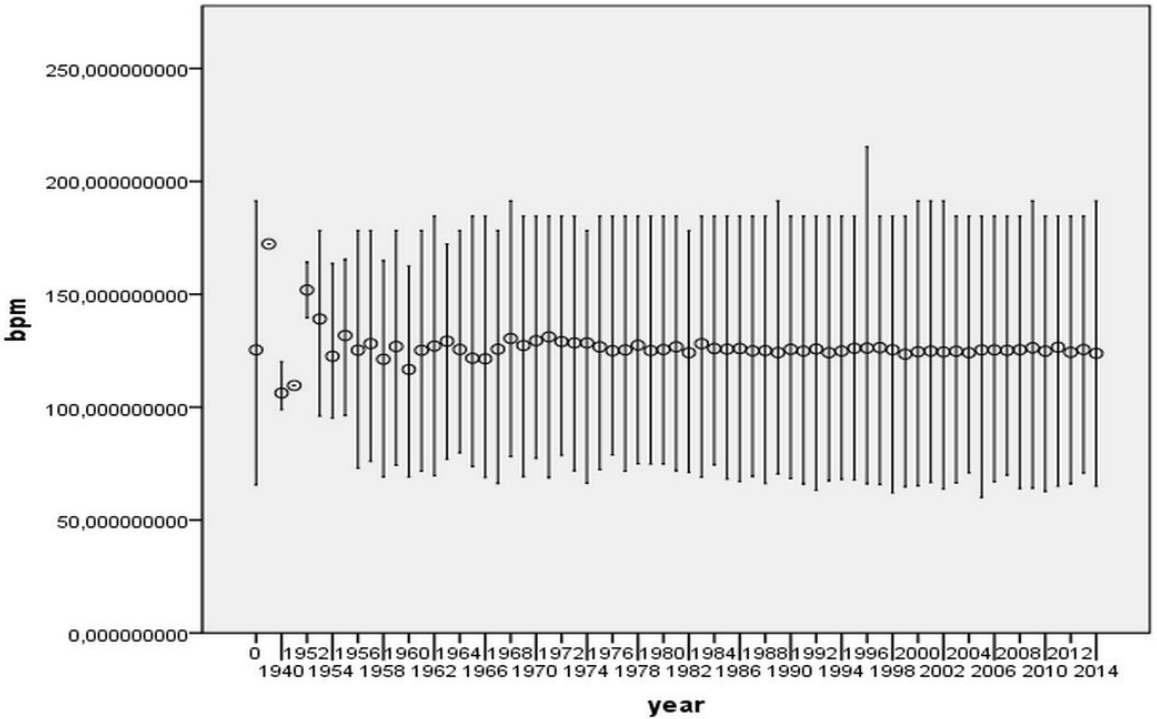
A little bit, but really, the data is not conclusive.



Is music getting faster?

No.

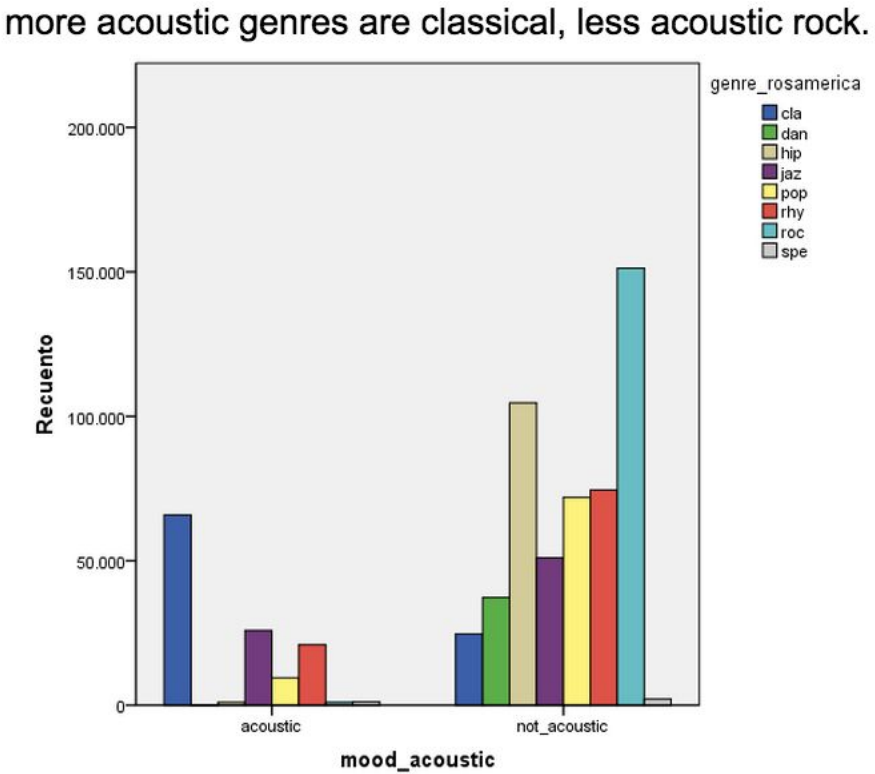
(At least we're not seeing any signs of it yet.)



Acoustic Moods

Acoustic moods are much more likely to be classical.

There isn’t much acoustic rock.



Follow the project on Twitter:

http://acousticbrainz.org

@AcousticBrainz

Your turn

You are to choose a mini-project to investigate an aspect of the AcousticBrainz data

Because of the scale of the data, we're providing some cut-down datasets

(Only a few gb, not 120gb)

We don’t know the answers

We’re investigating the unknown.

(This also means that there are no right or wrong answers. We will grade on process and discussion)

Do you need help?

Discuss with your classmates, but do the work and report independently

If you are stuck, come and see us! We're interested in the results as well, so are happy to help you if you're stuck.

Data science! (like real science except it won't explode)

The small dataset is still quite large. You will have to work out how to chop it up.

There are some tools available in Linux / OSX

head, tail, sort

a set of tools for reading and modifying csv files is csvkit `http://csvkit.readthedocs.org`

Data science!

Some data items are submitted in **JSON** format

This interchange format can be read by many languages (python, ruby, javascript, Matlab)

Data science!

You may not need to use all of the data that you get in the dataset

This is normal for data research projects. You need to look at the data and find the best subset for your task

Deliverables

Tuesday 8th: Give a 5 minute presentation on your task, what you have done so far, and discuss the results with the class.
Tuesday 15th: Submit a minimum 2 page report on the task (more if you need figures, etc) covering an outline of the problem, your approach, results, and a discussion of the results.

We also require any code that you wrote for the project. This **must** be submitted as a git repository, not a zip file. We should be able to recreate the results with the code.
You can create a free account on github. If you don't know how to use git, I'll give a small introduction to it on Thursday during the research methods class.

Task: dataset creation (from last.fm tags)

We are providing a list of tags from lastfm for a subset of recordings in AcousticBrainz
You have seen that we made a dataset by selecting genre tags and trained a model
Do the same for another aspect that you can determine from the tags. For example, mood, instrumentation, male/female
upload the dataset to acousticbrainz, and create the model.
This process will take about 2 hours per 1,000 instances in your dataset. If other students are doing the same task you **will not** have time in the last day before the task is due. Start early!

Task: data collection

For a previous project we have gathered annotations (tags) from last.fm

Choose another website to scrape annotations from.

Some examples:

- discogs (genre, style) [https://www.discogs.com/developers/]
- musicbrainz (tags, artist/release metadata) [http://wiki.musicbrainz.org/Development/XML_Web_Service/Version_2
- rate your music (genre) [http://rateyourmusic.com/, https://github.com/jcazevedo/beets-rymgenre]
- wikipedia [either search for wikipedia urls on musicbrainz, or search wikipedia or wikidata [https://www.wikidata.org/wiki/Property:P136]
- all music guide (mood, theme, genre, style) [more difficult - web scraping, https://github.com/daveisadork/picard-allmusic]

Task: data collection

Contributions should be made to the metadb project:

https://github.com/MTG/metadb

Set up the server, and write a plugin which scrapes and parses the data.
We will only allow one person per annotation source.

Open a ticket on github to reserve your spot.

Open a pull request to contribute your changes to the main project.

Task: duplicate analysis

We have many duplicate submissions. We are providing a dataset of **only** songs with more than 1 submission.

Analyse if these duplicates have similar descriptor values. Some examples of values that you could compare include:

`bpm, average loudness, onset rate, beat positions, chords histogram, length, hpcp->mean, key_key & key_scale, metadata->replaygain, tuning frequency`

If you can, see if any duplicate submission appears to be a different song, and explain why.

Task: Content analysis for genre classification

We have genre ground truth,

we built a dataset from this ground truth, and a model.

Non-linear SVMs are difficult to analyze, so we don’t know which descriptors contribute to the separation between classes (genres).

Using the data which we provide, try and find descriptors which differ between classes.

SVM only uses the mean and variance values of each feature, so you should only consider these

Next week

Student presentations of these projects. Come prepared, you will be called upon to present your work.

5 minutes (5-10 slides) about your task, and results so far.

We will also show how we decided to do some of the tasks

Any questions about the tasks?

Come and see Alastair (55.306) or

Dmitry (55.308)