The Million Song Dataset

Thierry Bertin-Mahieux[1] Daniel P.W. Ellis[1]
Brian Whitman[2] Paul Lamere[2]

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

[1] **LabROSA**
Columbia University
New York, USA

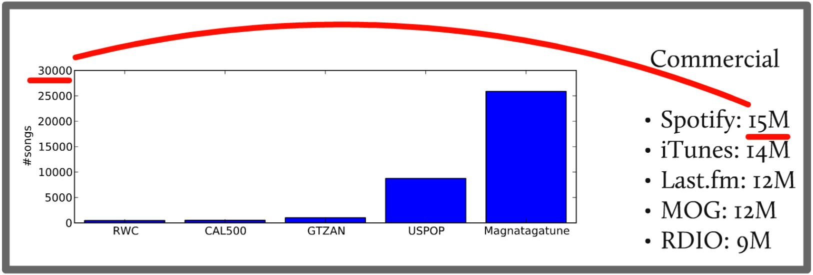
[2] **The Echo Nest**
Somerville, MA

the echonest

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

Academic vs. Commercial Resources

There is a gap to fill.

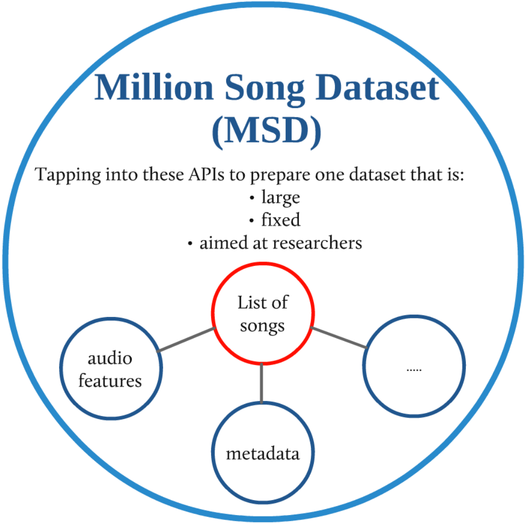


Commercial

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

- We still want datasets
- We want web data (scale and type), e.g. from APIs

Goals of the MSD



MSD is:

- fixed set
- audio feature
  + metadata
- for researchers

A very brief history

- collaboration LabROSA - The Echo Nest
- NSF grant for an academia - industry project
- Idea of a million song sounded **cool**!
- **Feasible** (API)
- **Useful** (larger MIR benchmark, easier industrial transfer)

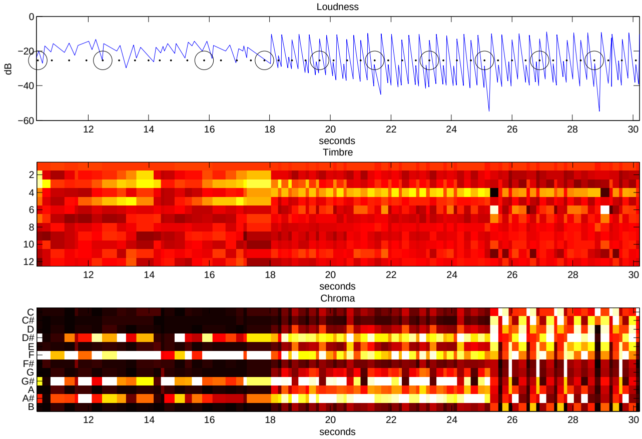
The Echo Nest data

Quick overview of the original MSD data

- **basic metadata**: artist name, title, IDs, ...
- **link** to other data: Musicbrainz, 7digital, ...
- **artist info**: name, origin, tags, similar artists, ...
- **audio features**: pitches and timbre per segment
- **segmentation data**: beats, bars, sections
- **evaluation** of: key, mode, tempo, ...

All this was contained in the original release (February 2011). Takes 280GB of space because of audio features.

Audio Features

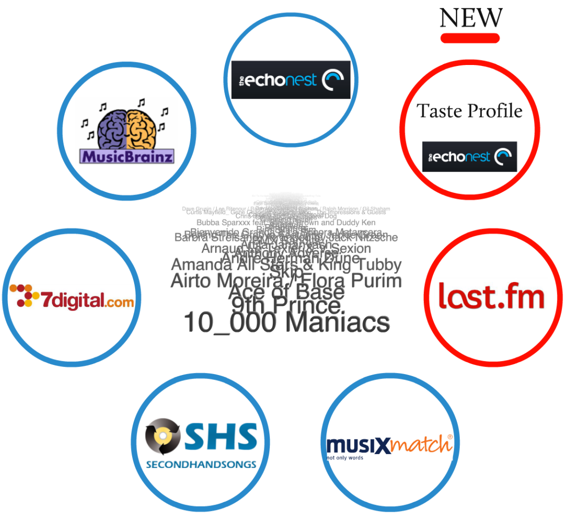


segment: between two note onsets

pitch: chroma

timbre: similar to MFCC

What can you get from the MSD?



Audio snippets

7digital.com

- online mp3 / streaming store
- free snippets for each of the MSD song through API
- enough for user testing or experiment

Cover songs

SHS SECONDHANDSONGS

The SecondHandSongs dataset

- online community-maintained list of cover songs
- **18,196** covers in **5,854** “cover cliques”
- finding a cover out of 1M song is a challenge
- preliminary results at WASPAA
- we provide split for train / test

Lyrics

musiXmatch®
not only words

- lyrics for **237,662** songs
- bag-of-word format
- stemmed words

For *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],
```

Tags & Similarity

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

How much can you do with a large similarity groundtruth?

Taste Profile subset

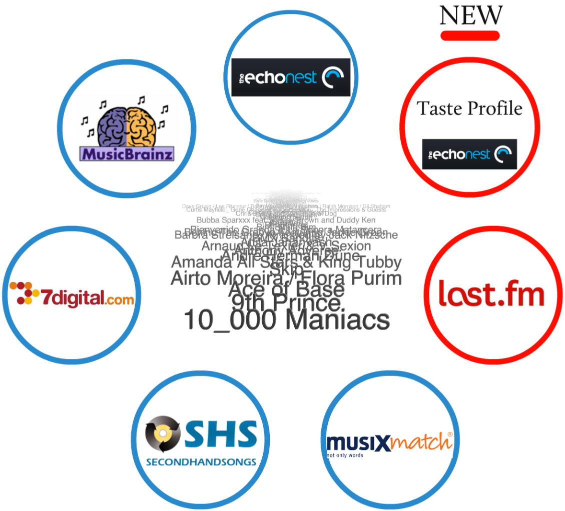
User data -> collaborative filtering
tons of (user - track - playcount) triplets!

Still in progress, but:

- subset available through The Echo Nest API
- already large! 120K users with at least 10 songs each

```
{u'id': u'CACNYVZ1332EB0BA9D',
         u'artist_name': u'M83',
         u'date_added': u'2011-10-23T15:59:59',
         u'foreign_id': u'CACNYVZ1332EB0BA9D:song:10286694_usercat',
         u'play_count': 1,
         u'song_id': u'SOFMYVK12A58A7A675',
         u'song_name': u'Skin Of The Night'},
         ...
```

What can you get from the MSD?



Year Prediction

Task definition

Predict the release year of a song
solely based on audio features.

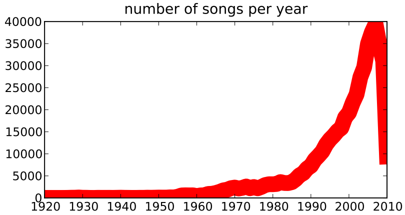
Why did we choose this task?

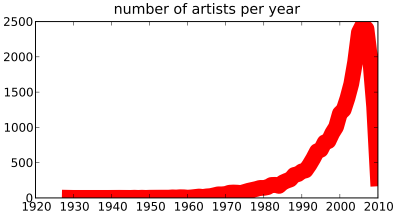
- Almost no mention of it in the literature
- Could not be done without a proper large dataset

Data

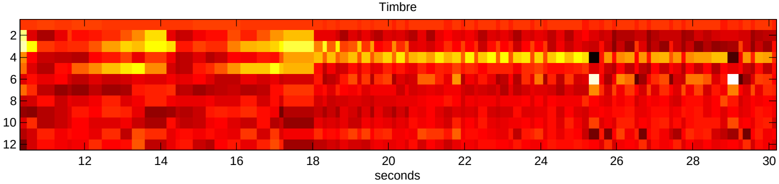
Year information from Musicbrainz

- **515,576** tracks
- **28,223** artists with at least one dated song





Features



- timbre features
- we take average and covariance
- one feature vector of dimension 90 per track
  (12 -> average, 78 -> upper triangle of covariance matrix)

Methods

k-NN

- euclidean distance
- we present results with $k = 1$ and $k = 50$

Vowpal Wabbit (VW)

- very fast linear predictor from J. Langford (Yahoo!)
- the magic is in the error function / gradient descent
- `http://hunch.net/~vw/`

Results

We measure *predicted year - real year*

method	avg. abs. difference	avg. sq. difference
constant pred.	8.13	10.80
1-NN	9.81	13.99
50-NN	7.58	10.20
vw	**6.14**	**8.76**

- VW better than k-NN
- 6 years error not that bad...
- but data is very short-tailed

Conclusion

What is the Million Song Dataset?

- Large collection of audio features and metadata
- Frozen set, for research
- Data sources are linked -> new possibilities
- Open project, anyone can contribute

... and year prediction is fun.

Thanks!

Any question?

And we're here all week, if you have questions come talk to me, or Brian (The Echo Nest), or Mark (Last.fm), or ...

And vist the website!

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