Meaning from Music: Automatically tagging audio files using supervised learning on acoustic features

Douglas Eck

University of Montreal Computer Science

Intl. Lab for Research in Brain Music and Sound (BRAMS)

Université de Montréal

BRAMS Brain Music and Sound

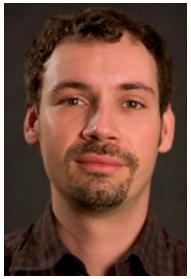
Fonds de recherche sur la nature et les technologies Québec

NSERC CRSNG

Acknowledgments

Research done with:

- Norman Casagrande (former grad. student; now researcher at Last.FM)
- James Bergstra (grad. student)
- Thierry Bertin-Mahieux (grad. student)
- Balazs Kegl (colleague)
- Paul Lamere (Sun Microsystems, Boston)



Structure of Talk

- Overview: machine music listening
- Audio feature extraction
- Feature selection using boosting
- Predicting artist and genre from audio features
- Predicting arbitrary tags from audio features

3 Current Commercial Approaches

- Collaborative Filtering (Amazon)
  - \+ Captures popularity and similarity among discs; unquestionably useful
  - \- More than books, songs are multipurpose (dinner party, jogging, close listening) but are purchased only once

- Social Recommendation (Last.FM)
  - \+ Better view of current musical tastes than Amazon
  - \+ Measures popularity
  - \- Cold-start problem
  - \- "All roads lead to Radiohead" (Popularity bias)

- Human Labeled Content-Based Recommendation (Pandora)

  Pandora: 40 experts label music on ~=400 params (7000 songs/month)
  - \+ Can capture multidimensional similarity
  - \+- No popularity bias
  - \- Not scalable (slow; what happens if they want the 401st parameter?)

Alternative: listen to music with machine learning

- Map acoustic features onto classes or distributions using labeled data
- Embed music, even new music, in a multidimensional space useful for
  - Automatic annotation (“autotagging”)
  - Similarity measure
  - Visualization
- Augment/regularize social tags & other web data

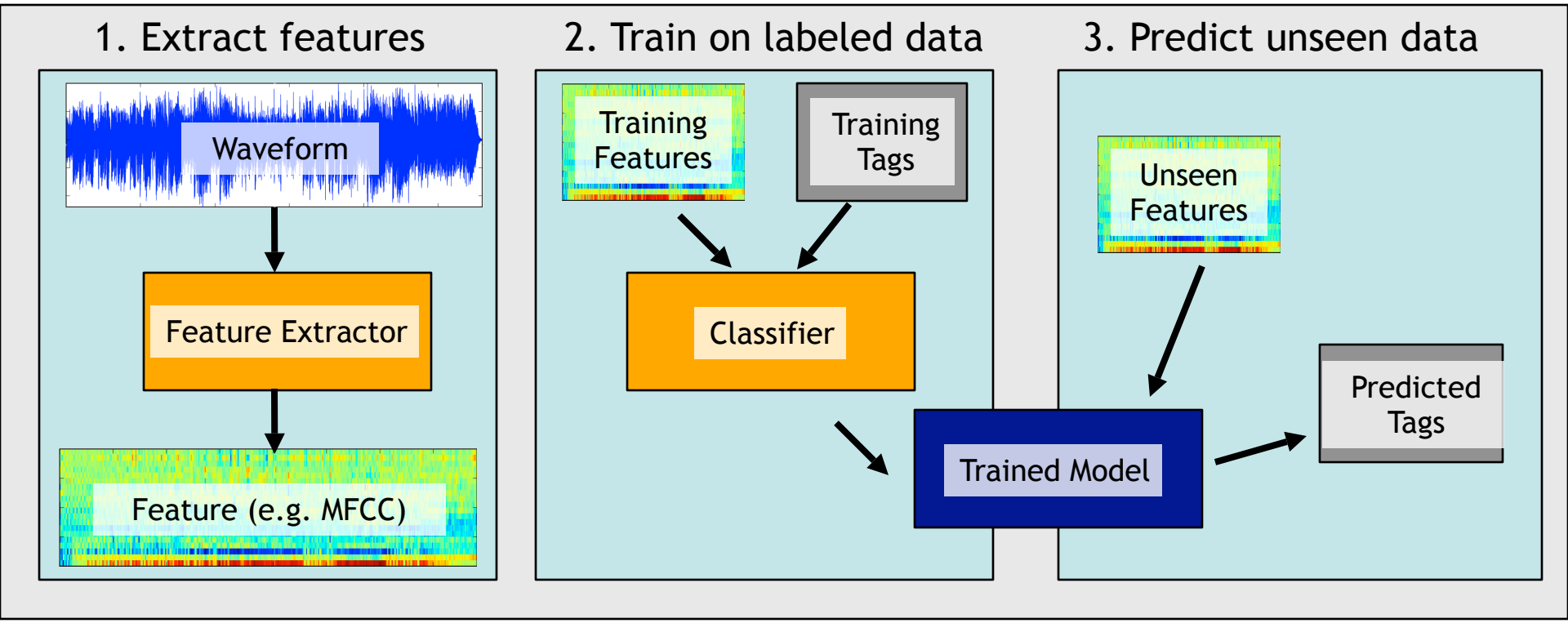
Pros

- More scalable than human expert annotation (Pandora)
- Can predict attributes not likely to be tagged (e.g. tempo in BPM; "highly compressed")
- **Tells us something about content of music audio on the web**

Cons

- Acoustic features not predictive of some attributes (e.g. "Protest music")
- Hard to measure quality (John Coltrane was just another bebop sax player?)
- Engineering is challenging (db construction, large-scale ML)

Automatic annotation ("Autotagging")



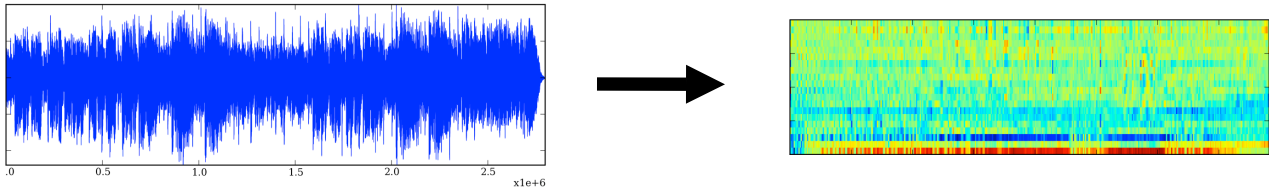
Challenges and previous work

- Challenges
  - What features to use?
  - What machine learning algorithm to use?
  - How to scale to huge datasets?

- Previous approaches (genre and artist prediction):
  - SVM (Ellis & Mandel 2006)
  - Decision Trees (West, 2005)
  - Nearest Neighbors (Palmpalk, 2005)
  - AdaBoost (Bergstra, Casagrande, Erhan, Eck & Kegl 2006)

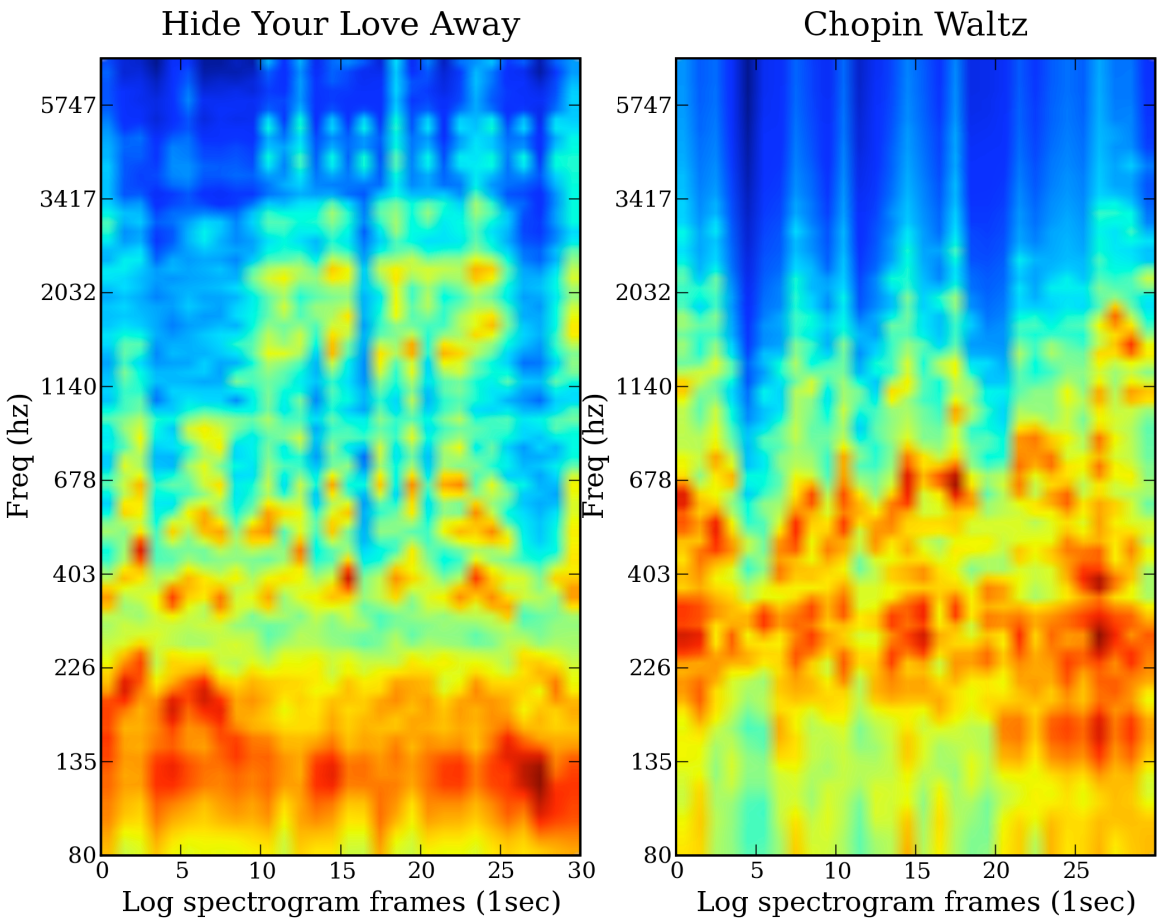
- Current approach (extension of our 2006 AdaBoost method):
  - Iteratively apply feature selection to build small feature set
  - Boost simple classifiers on individual features
  - Predict lots of independent classes (social tags)

Feature Extraction



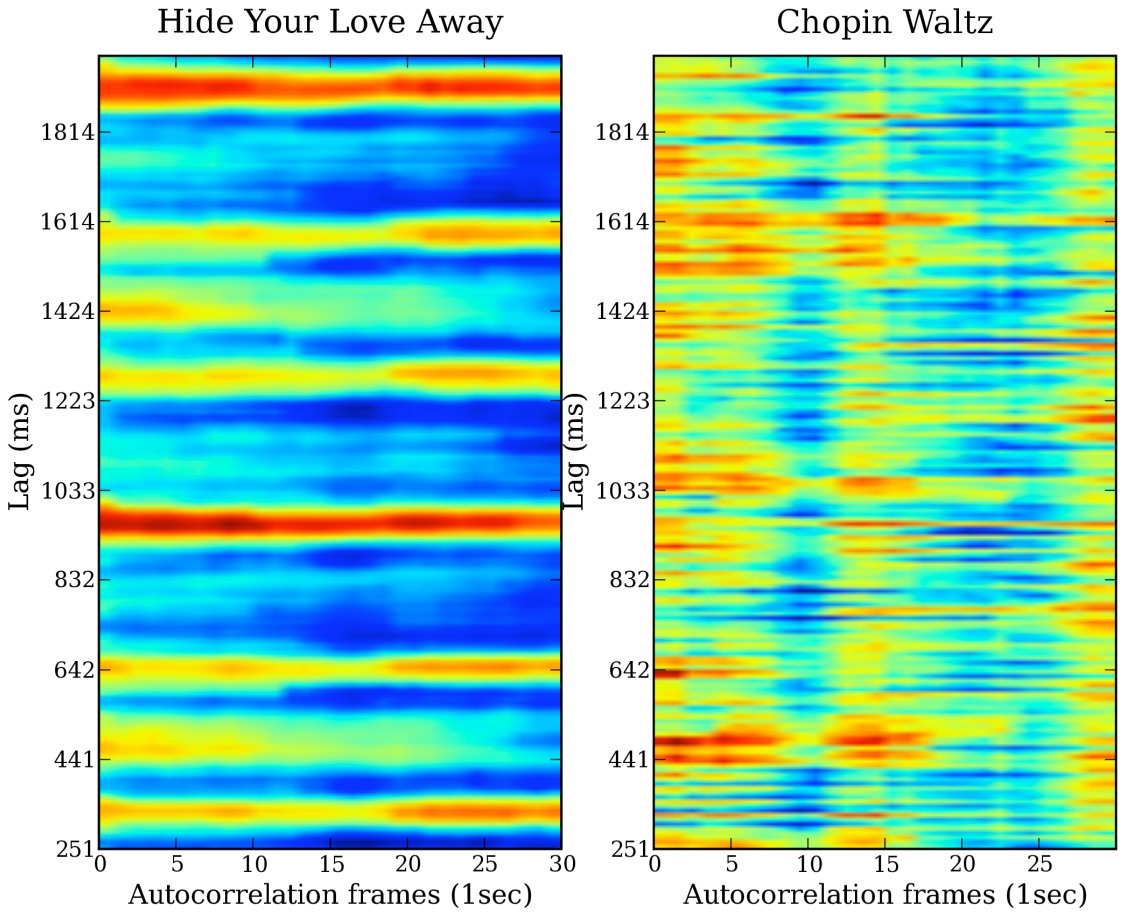
- Extract features from audio which reveal musical content
- Many features come from speech recognition
- Three major categories:
  - Spectral features (Fourier Transform; time-->frequency)
    Example: Spectrogram
  - Cepstral features (Fourier transform of spectral features; time-->frequency-->time)
    Example: Cepstral coefficients; Mel-Frequency cepstral coefficients; Autocorrelation
  - Statistical features
    Example: zero crossing rate over time, LPCs

Spectral feature (Log spectrogram ~= Constant-Q transform)



Short timescale Fourier transform (STFT) with 100ms hops; frequencies sampled logarithmically

Cepstral feature (Autocorrelation)



Compare:

Cepstrum = ifft (log(fft(x)))

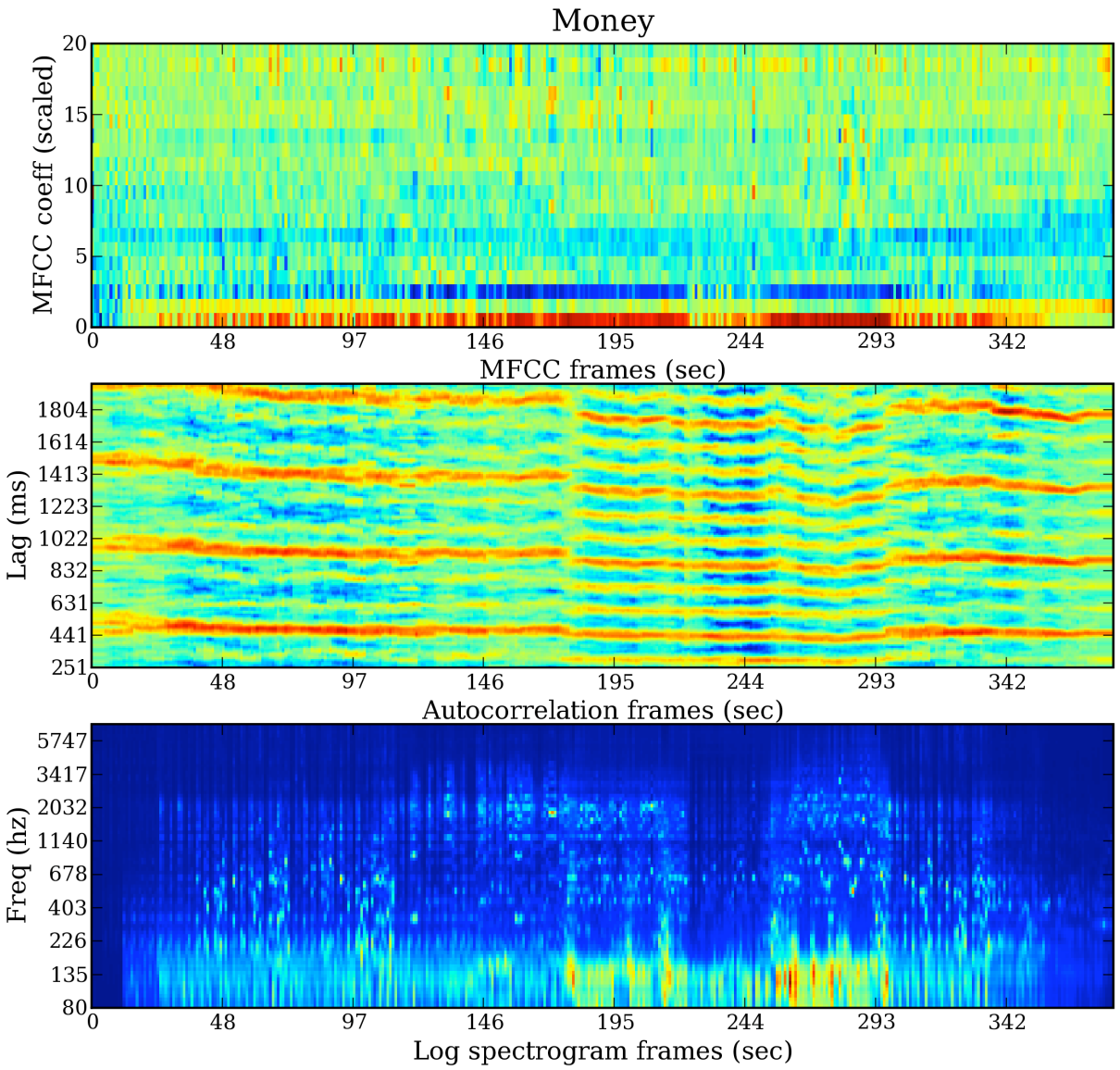
Autocorr = ifft (|(fft(x))|^2)

Example: Pink Floyd "Money"

MFCC
Timbre compressed
Into a few coeffs.

Autocorrelation
Temporal structure
(rhythm, meter)

Spectrogram
Pitch, timbre distributed
over many coeffs.

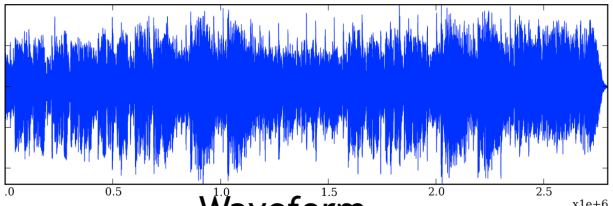


High-dimensional features

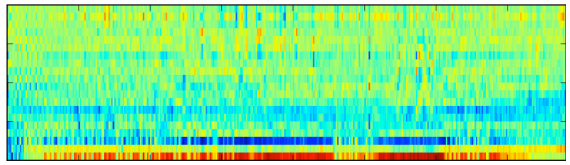
- 1 minute of CD-quality PCM audio
  44025 * 2 *60 ~= 5.3M values per min

- 512-point spectrogram computed with 50ms frames
  512* 20 * 60 = 614,440 values per min (8.6x compression)

- 13-point Mel Frequency Cepstral Coeffs with 50ms frames
  13 * 20 * 60 = 15600 values per min (340x compression)



Waveform



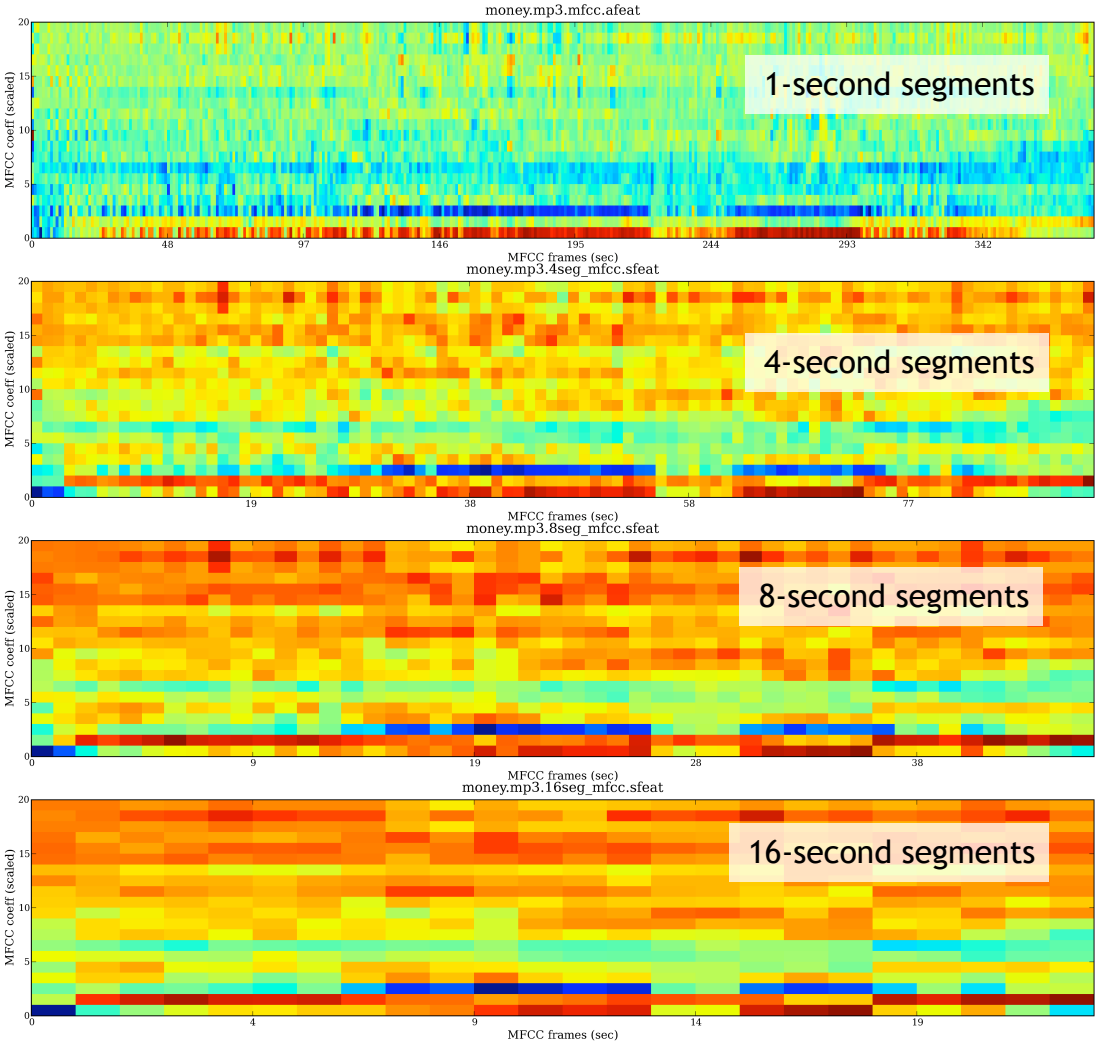
MFCC=340x compression

?

Aggregate Features

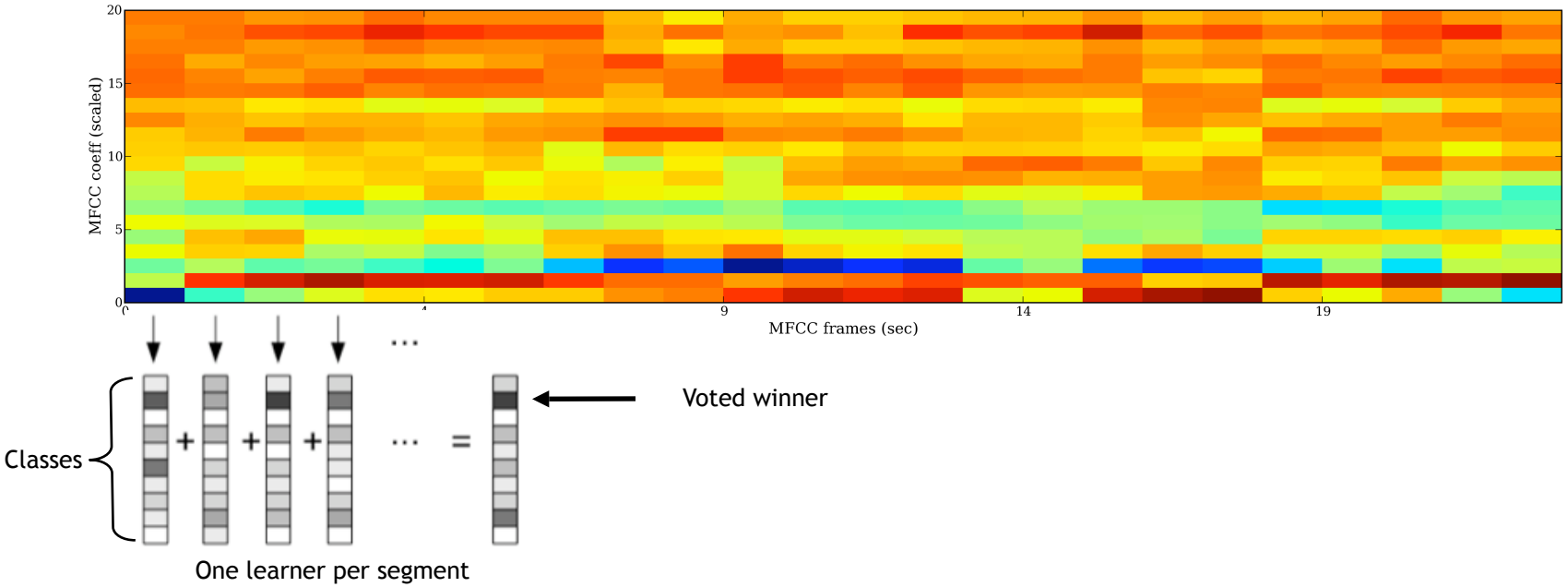
- Aggregate chunks of feature frames into longer-timescale segments
- Method: independent Gaussians
- More complex approaches possible (e.g. mixture of Gaussians, virtually any dim. reduction algorithm)
- Question: What is the best segment size?

Pink Floyd "Money"



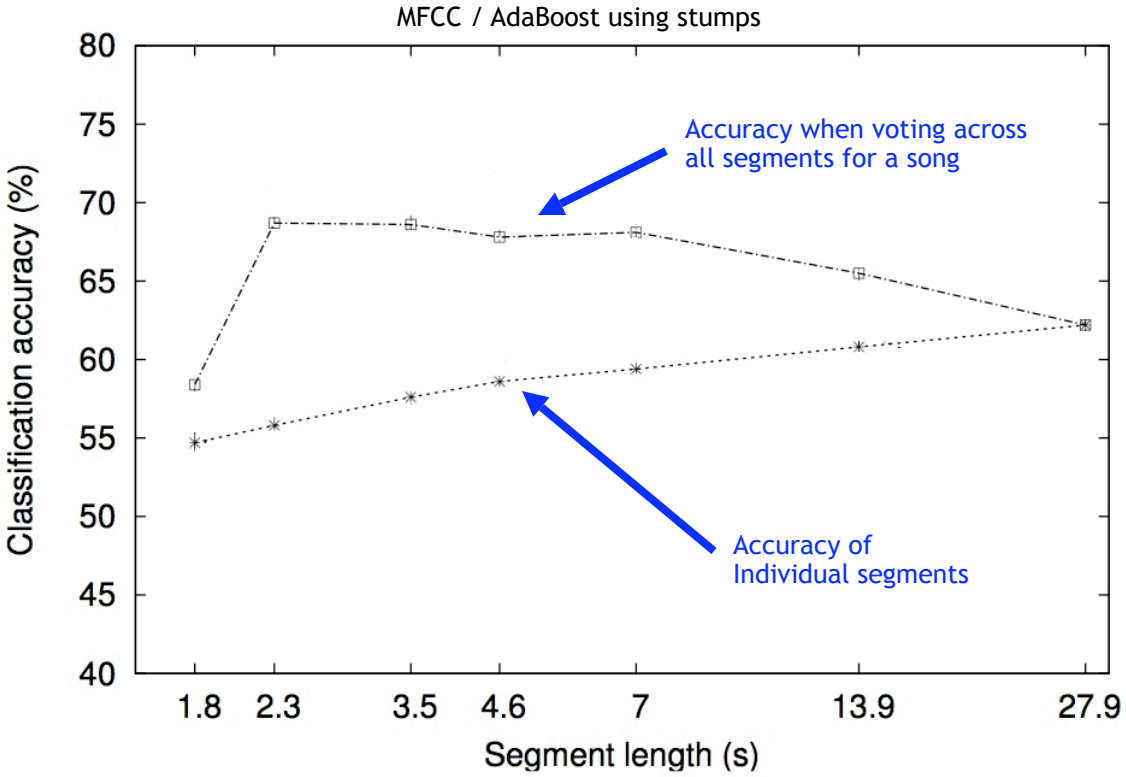
Voting

- Spectral features more predictive over short timescales
- Train segments individually using song-level label
- Vote to choose winner
- Segment size is important



What is the best segment size?

- Tested range of segment sizes for four features: FFT, RCEPS, MFCC, MiscStat
- Tested 4 different learners (2 Boosters, ANN, SVM) with
- 1000-song 10-class dataset
- Segments trained individually and voted for prediction
- **Result: segment size between 3 and 8 sec is optimal**



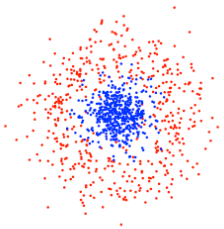
AdaBoost (Freund & Schapire 1995)

Build an initial model using a single weak learner

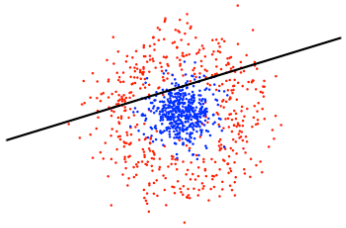
While (error criterion not met) :

1. ID wrongly-classified points in the dataset
2. Boost these points so that they will receive more attention
3. Add best weak learner over boosted dataset to model

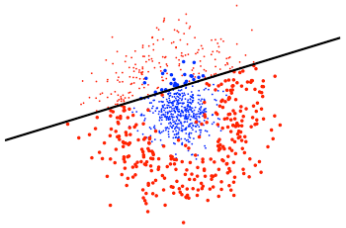
Ex: pos. points $= N(0,1)$, neg. points $= \frac{1}{r\sqrt{8\pi^3}} e^{-1/2(r-4)^2}$



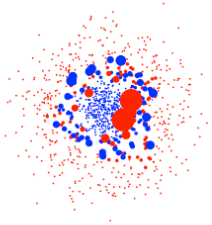
Training set



First cut



After first cut



After many cuts

Algorithm AdaBoost

Input: sequence of N labeled examples $\langle (x_1, y_1), \ldots, (x_N, y_N) \rangle$
distribution D over the N examples
weak learning algorithm **WeakLearn**
integer T specifying number of iterations

Initialize the weight vector: $w_i^1 = D(i)$ for $i = 1, \ldots, N$.

Do for $t = 1, 2, \ldots, T$

1. Set
$$\mathbf{p}^t = \frac{\mathbf{w}^t}{\sum_{i=1}^N w_i^t}$$

2. Call **WeakLearn**, providing it with the distribution $\mathbf{p}^t$; get back a hypothesis $h_t : X \to [0, 1]$.

3. Calculate the error of h_t: $\epsilon_t = \sum_{i=1}^N p_i^t |h_t(x_i) - y_i|$.

4. Set $\beta_t = \epsilon_t / (1 - \epsilon_t)$.

5. Set the new weights vector to be
$$w_i^{t+1} = w_i^t \beta_t^{1-|h_t(x_i)-y_i|}$$

Output the hypothesis

$$h_f(x) = \begin{cases} 1 & \text{if } \sum_{t=1}^T \left(\log \frac{1}{\beta_t}\right) h_t(x) \geq \frac{1}{2} \sum_{t=1}^T \log \frac{1}{\beta_t} \\ 0 & \text{otherwise} \end{cases} .$$

Observations

- In our case the $h_t(x)$ weak learners are decision stumps on individual features
- We perform feature selection based on minimization of empirical error
- Slow to converge relative to Sims but linear complexity w.r.t. dataset size
- Multi-class learning done using AdaBoost.MH (weak learners are 1-versus-all)
- Many improvements address weaknesses of AdaBoost (e.g. not good with noisy data)

Task 1: Genre & Artist Prediction

1. 1000-song 10-class dataset from Tzanetakis
2. Genre prediction contest from 2005 MIREX Contest at ISMIR
3. Artist prediction contest from 2005 MIREX

Used spectral and cepstral features + several additional features:

- 256 RCEPS
- 64 MFCC
- 32 Fourier coefficients
- 32 LPCs (linear predictive coefficients)
- 16 Rolloff coefficients
- 1 linear prediction error
- 1 zero crossing rate

Tzanetakis Database

- 1000 30sec audio segments
- 10 classes: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, rock

Model	Correct rate
G. Tzanetakis (2002)	61%
T. Li (2003)	71%
Our approach	**83%**

MIREX Artist Identification (2005)

- "MAGNATUNE" database (www.magnatune.com)
  - 1005 training and 510 testing files (1515 total)
  - 77 artists

- "USPOP" database (Dan Ellis, Columbia)
  - 940 training and 474 testing files (1414 total)
  - 77 artists

Rank	Participant	Overall	Magnatune	USPOP
1	Mandel & Elllis (SVM)	**72.45%**	76.60%	**68.30%**
2	Our approach A (AdaBoost w/ stumps)	68.57%	**77.26%**	59.88%
3	Our approach B (AdaBoost w/ trees)	66.71%	74.45%	58.96%
4	Pampalk (Nearest neighbors)	61.28%	66.36%	56.20%
5	West & Lamere (Decision Trees)	47.24%	53.43%	41.04%

MIREX Genre Prediction (2005)

- "MAGNATUNE" database (www.magnatune.com)
  - 1005 training and 510 testing files (1515 total)
  - 10 genre (hierarchical)

- "USPOP" database (Dan Ellis, Columbia)
  - 940 training and 474 testing files (1414 total)
  - 6 genre (flat)

Rank	Participant	Overall	Magnatune	USPOP
1	Our approach B (AdaBoost w/ trees)	**82.34%**	**75.10%**	**86.92%**
2	Our approach A (AdaBoost w/ stumps)	81.77%	74.71%	86.29%
3	Mandel & Elllis (SVM)	78.81%	67.65%	85.65%
4	West (Decision Trees)	75.29%	68.43%	78.90%
5	Lidy & Rauber (SVM)	75.27%	67.65%	79.75%

Task 2: Automatic Annotation of Social Tags

- Collected tags and tag frequencies for over 50k artist from Last.FM
- Genre, mood, instrumentation account for 77% of tags

Tag Type	Frequency	Examples
Genre	68%	heavy metal, punk
Locale	12%	French, Seattle, NYC
Mood	5%	chill, party
Opinion	4%	love, favorite
Instrumentation	4%	piano, female vocal
Style	3%	political, humor
Misc	3%	Coldplay, composers
Personal	1%	seen live, I own it

Top 20 tags applied to "The Shins"

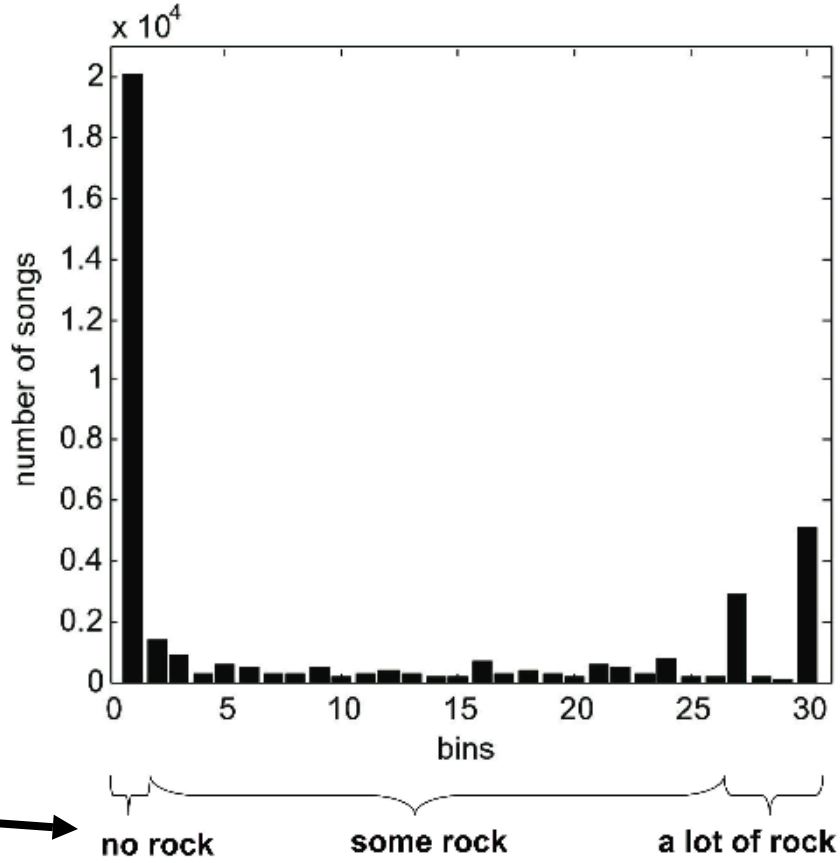
Tag	Freq	Tag	Freq
Indie	2375	Mellow	85
Indie rock	1138	Folk	85
Indie pop	841	Alternative rock	83
Alternative	653	Acoustic	54
Rock	512	Punk	49
Seen Live	298	Chill	45
Pop	231	Singer-songwriter	41
The Shins	190	Garden State	39
Favorites	138	Favorite	37
Emo	113	Electronic	36

last·fm
the social music revolution

What can we learn from tags?

- Focus on tag types that are learnable from audio features (e.g. genre, mood)
- Regression or ranking would be difficult
  - Uneven coverage
  - Many untagged songs
  - Difficult to normalize
  - Distribution constantly in flux
- Classification of *binned* tag data (e.g. no rock, some rock,a lot of rock)
- K bins with same # songs per bin
  - K=3 except classical where K=2

Histogram of tag "Rock" frequency / song



Autotagging experiments

- Simplified feature set
  - MFCCs (20)
  - Log spectrogram coefficients (85)
  - Autocorrelation coefficients (88)
- Learn tags independently (not 1-vs-all)
  - Currently: 13 hand-chosen tags related to genre
- Use magnitudes of weak learner predictions to form graded prediction (regression)

13 selected tags from Last.fm			
Tag	# Artists	Tag	# Artists
jazz	225	soul	44
rock	185	alternative	41
electronic	115	country	15
classical	108	punk	15
folk	59	reggae	14
indie	55	britpop	10
classic rock	51		

Number of artists with at least one tag	937
Number of artists with no tag	605
Number of songs	43597

Summary of results

	Songs	Segments
alternative	55.4%	59.3%
britpop	59.1%	62.3%
classic rock	56.7%	60.4%
classical	86.8%	90.0%
country	58.9%	64.9%
electronic	58.0%	61.1%
folk	59.9%	62.6%
indie	53.7%	57.5%
jazz	58.0%	62.5%
punk	59.4%	63.4%
reggae	61.3%	64.8%
rock	54.9%	58.7%
soul	55.5%	58.8%

2 bins; all others had 3 bins

Test error % (1 fold of 5-fold cross-validation) with 2000 single-stump learners.

Classic rock

1 INXS
2 Creedence Clearwater Revival
3 Steppenwolf
4 The Cars
5 The Psychedelic Furs
6 The Zombies
7 Eric Burdon and the Animals
8 The Lovin' Spoonful
9 Crowded House
10 Ramones
11 Meat Loaf
12 Jimmy Buffett
13 Tom Petty and the Heartbreakers
14 ZZ Top
15 The Mamas & The Papas
16 The Byrds
17 Tina Turner
18 X
19 Guns N' Roses
20 Elvis Costello & The Attractions
21 The Rolling Stones
22 The Housemartins
23 The Beatles
24 Al Green
25 Darlene Love
26 Fugazi
27 Bob Dylan
28 Arlo Guthrie
29 Elvis Costello
30 Jeff Wayne
31 Violent Femmes
32 The B-52's
33 Gin Blossoms
34 Joe Jackson
35 The Commitments
36 Lloyd Cole and the Commotions
37 Talking Heads
38 Cream
39 Bryan Ferry
40 The Band

Electronic

1 Sasha & John Digweed
2 Paul van Dyk
3 Aqua
4 Paul Oakenfold
5 Sasha
6 John Digweed
7 BT
8 Juno Reactor
9 Ministry of Sound
10 Fluke
11 The Crystal Method
12 Les Rythmes Digitales
13 808 State
14 Orbital
15 Nortec Collective
16 Hybrid
17 ATB
18 Leftfield
19 Tangerine Dream
20 Daft Punk
21 Olive
22 Laurent Garnier
23 Eiffel 65
24 The Shamen
25 The Chemical Brothers
26 Basement Jaxx
27 Chicane
28 bis
29 Aaliyah
30 Brazilian Girls
31 Kraftwerk
32 Underworld
33 John Lydon
34 Sneaker Pimps
35 Electronic
36 Boom Boom Satellites
37 Massive Attack vs. Mad Professor
38 Kid Loco
39 Fatboy Slim
40 The Other Two

Reggae

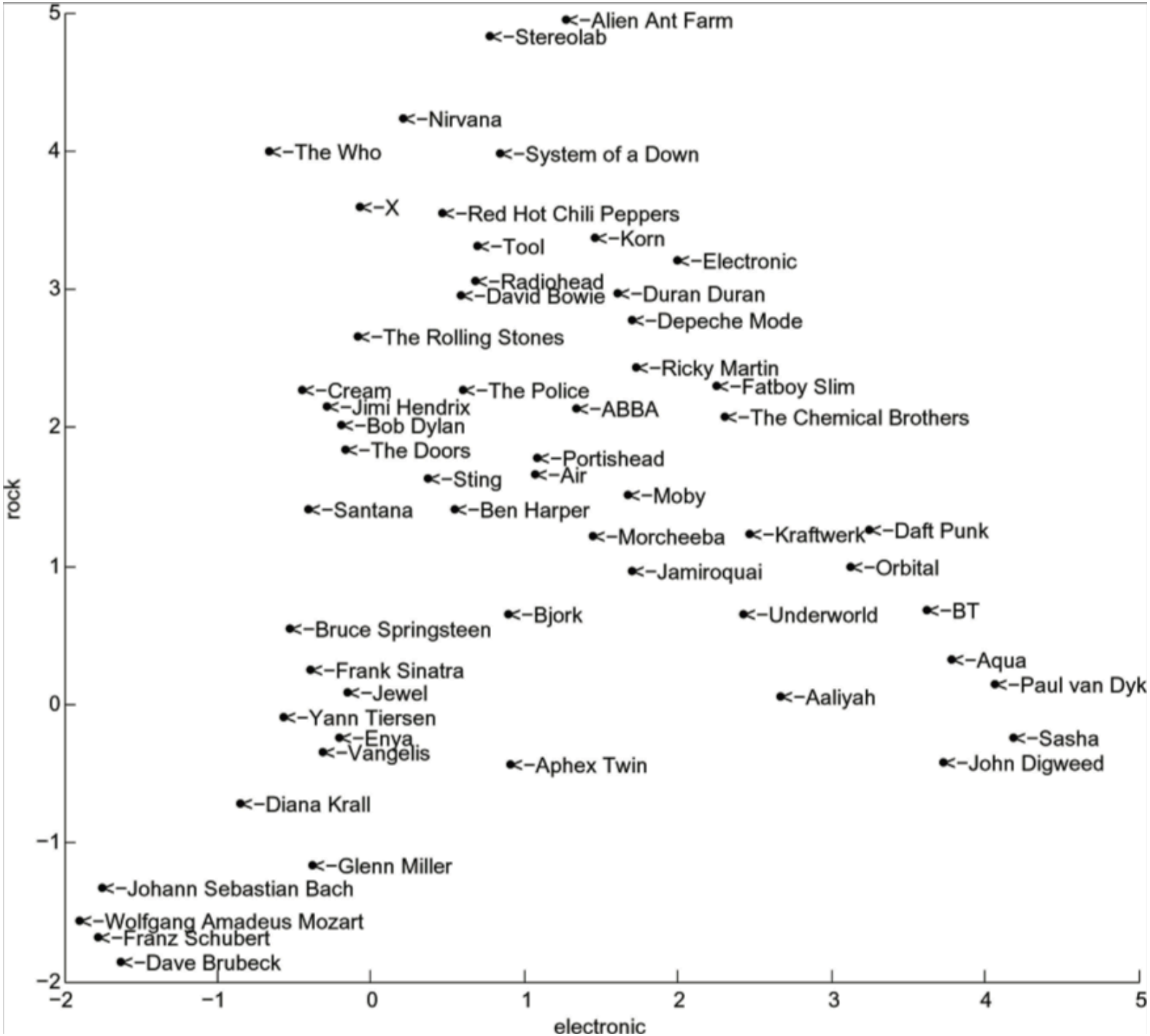
1 Bunny Wailer
2 Burning Spear
3 The Abyssinians
4 Dennis Brown
5 Jimmy Cliff
6 Fugees
7 Peter Tosh
8 Steel Pulse
9 The Melodians
10 Culture
11 D'Angelo
12 Third World
13 OutKast
14 Sublime
15 Aaliyah
16 Jill Scott
17 George Clinton
18 Missy Elliott
19 Fela Kuti
20 Dispatch
21 Parliament
22 Ben Harper
23 Johnny Nash
24 John Lennon & Yoko Ono
25 Big Audio Dynamite
26 Gilberto Gil
27 The Police
28 Joss Stone
29 Ernest Ranglin
30 Sneaker Pimps
31 Ursula 1000
32 Erykah Badu
33 Bebel Gilberto
34 Los Amigos Invisibles
35 Us3
36 Mandalay
37 _Weird Al_ Yankovic
38 Prince
39 Al Green
40 Soul Coughing

Examples…

	britpop	classical	classicrock	electronic	indie	jazz	reggae	rock
Gustav Mahler	-3.145	2.604	-1.313	-1.258	-1.022	-1.715	-4.076	-1.080
Wilco	0.187	-3.011	0.142	0.139	0.960	-1.322	-1.603	1.234
The Beatles	0.110	-2.887	0.972	-0.146	0.512	-0.614	-0.907	1.239
New Order	0.954	-2.678	0.355	1.214	0.874	-1.475	-1.249	1.173
Jamiroquai	-0.302	-2.017	-0.475	0.805	0.033	-0.579	0.164	0.254
Thelonious Monk	-3.200	-1.451	-1.159	-1.221	-1.417	1.440	-3.590	-1.250
Tom Waits	-1.661	-1.485	-0.312	-0.498	-0.295	0.441	-1.227	-0.212

Near-neighbor artists		
Seed Artist	Last.fm Tags	Our Prediction
Fatboy Slim	The Prodigy Basement Jaxx Apollo 440	Chemical Brothers Apollo 440 Beck
The Beatles	John Lennon The Beach Boys The Doors	Eric Clapton Marvin Gaye The Rolling Stones
Mozart	Bach Beethoven John Williams	Schubert Haydn Brahms

Near-neighbor artists using Euclidian distance across all 13 genre


Alien Ant Farm
Stereolab
Nirvana
The Who
System of a Down
X
Red Hot Chili Peppers
Tool
Korn
Electronic
Radiohead
David Bowie
Duran Duran
Depeche Mode
The Rolling Stones
Ricky Martin
Cream
The Police
Fatboy Slim
Jimi Hendrix
ABBA
The Chemical Brothers
Bob Dylan
The Doors
Portishead
Sting
Air
Moby
Santana
Ben Harper
Morcheeba
Kraftwerk
Daft Punk
Jamiroquai
Orbital
Bjork
Underworld
BT
Bruce Springsteen
Aqua
Frank Sinatra
Paul van Dyk
Jewel
Aaliyah
Yann Tiersen
Enya
Sasha
Vangelis
Aphex Twin
John Digweed
Diana Krall
Glenn Miller
Johann Sebastian Bach
Wolfgang Amadeus Mozart
Franz Schubert
Dave Brubeck
rock
electronic
5
4
3
2
1
0
−1
−2
−2
−1
0
1
2
3
4
5

Results 2

- In later work (NIPS 2007) trained 60 tags
- One day of processing per tag

	Mean	Median	Min	Max
Segment	40.93	43.1	21.3	49.6
Song	37.61	39.69	17.8	46.6

Table 2: Summary of test error (%) on predicting bins for songs and segments.

- Currently working on new FilterBoost model with ~2hrs per tag

Similarity Measures

- How to measure fit to some known list of similar artists?
- *TopN* measures how well we fit the top N artists. Let k_j be the position in list B of the *j*th element from list A.

$$s_i = \frac{\sum_{j=1}^{N} \alpha_r^j \alpha_c^{k_j}}{\sum_{l=1}^{N} (\alpha_r * \alpha_c)^l}$$

where $\alpha_r = 0.5^{1/3}$, and $\alpha_c = 0.5^{2/3}$

- *TopBucket* is percentage of commen elements in top N positions of two ranked lists

Similarity Measures continued

A third measure is Kendall's Tau. Here is the text from the NIPS paper:

Our second measure is Kendall's Tau, a classic measure in collaborative filtering which measures the number of discordant pairs in 2 lists. Let $R_A(i)$ be the rank of the element i in list A, if i is not explicitly present, $R_A(i) = length(A) + 1$. Let C be the number of concordant pairs of elements (i, j), e.g. $R_A(i) > R_A(j)$ and $R_B(i) < R_B(j)$. In a similar way, D is the number of discordant pairs. We use τ's approximation in [8]. We also define T_A and T_B the number of ties in list A and B. In our case, it's the number of pairs of artists that are in A but not in B, because they end up having the same position $R_B = length(B) + 1$, and reciprocally. Kendall's tau value is defined as:

$$\tau = \frac{C - D}{sqrt((C + D + T_A)(C + D + T_B))} \tag{3}$$

Ground truth

- Used Last.fm social tags for popular artists as ground truth.
- Correlations from listening habits
- If significant number of listeners all listen to artist A and B, we treat A and B as similar
- TF/IDF adjustment used to normalize
- Treat artists as documents, users as words

Similarity results

	TopN 10	Kendall (N=5)	TopBucket (N=5)
autotags	0.636	-0.099	**61.0%**
random	0.111	-0.645	8.1%

Table 3: Results for all three measures on tag order for 100 out-of-sample artists.

Groundtruth	Model	TopN 10	Kendall 50	TopBucket 20
Last.FM	social tags	0.26	-0.23	34.6%
	autotags	0.118	-0.406	22.5%
	random	0.005	-0.635	3.9%
MusicSeer	social tags	0.237	-0.182	29.7%
	autotags	0.184	-0.161	28.2%
	random	0.051	-0.224	21.5%

Table 4: Performance against Last.Fm (top) and MusicSeer (bottom) ground truth.

More similarity results

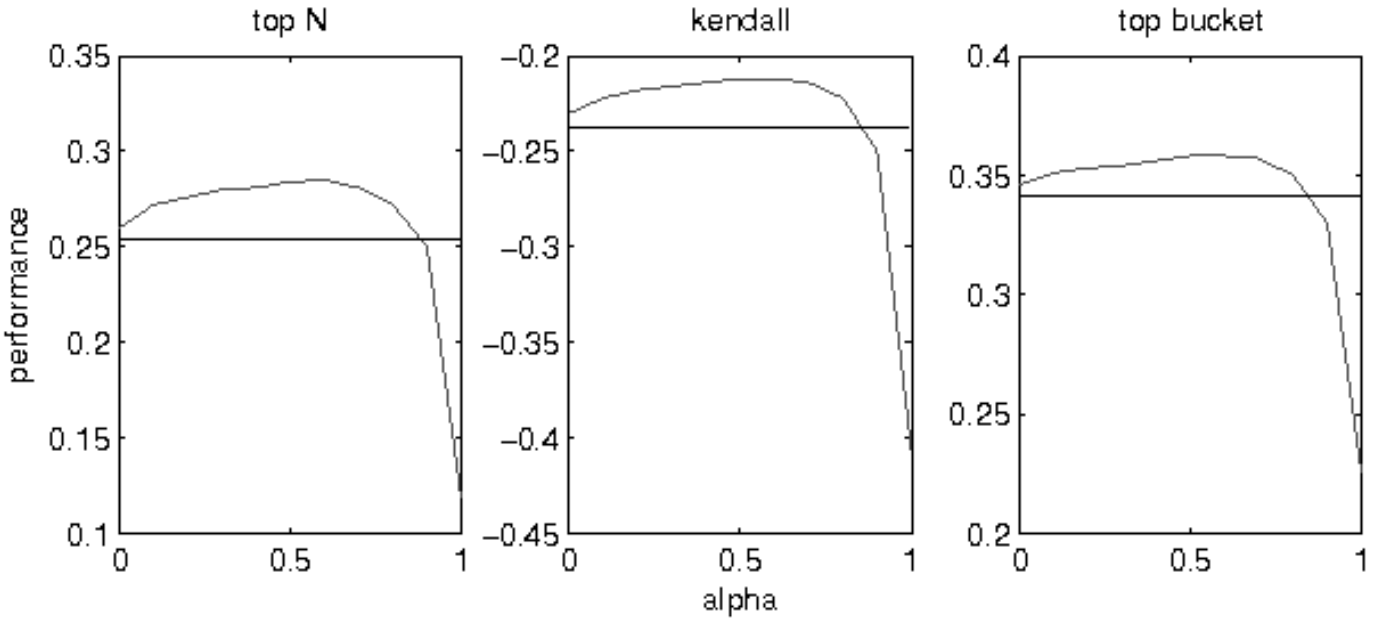
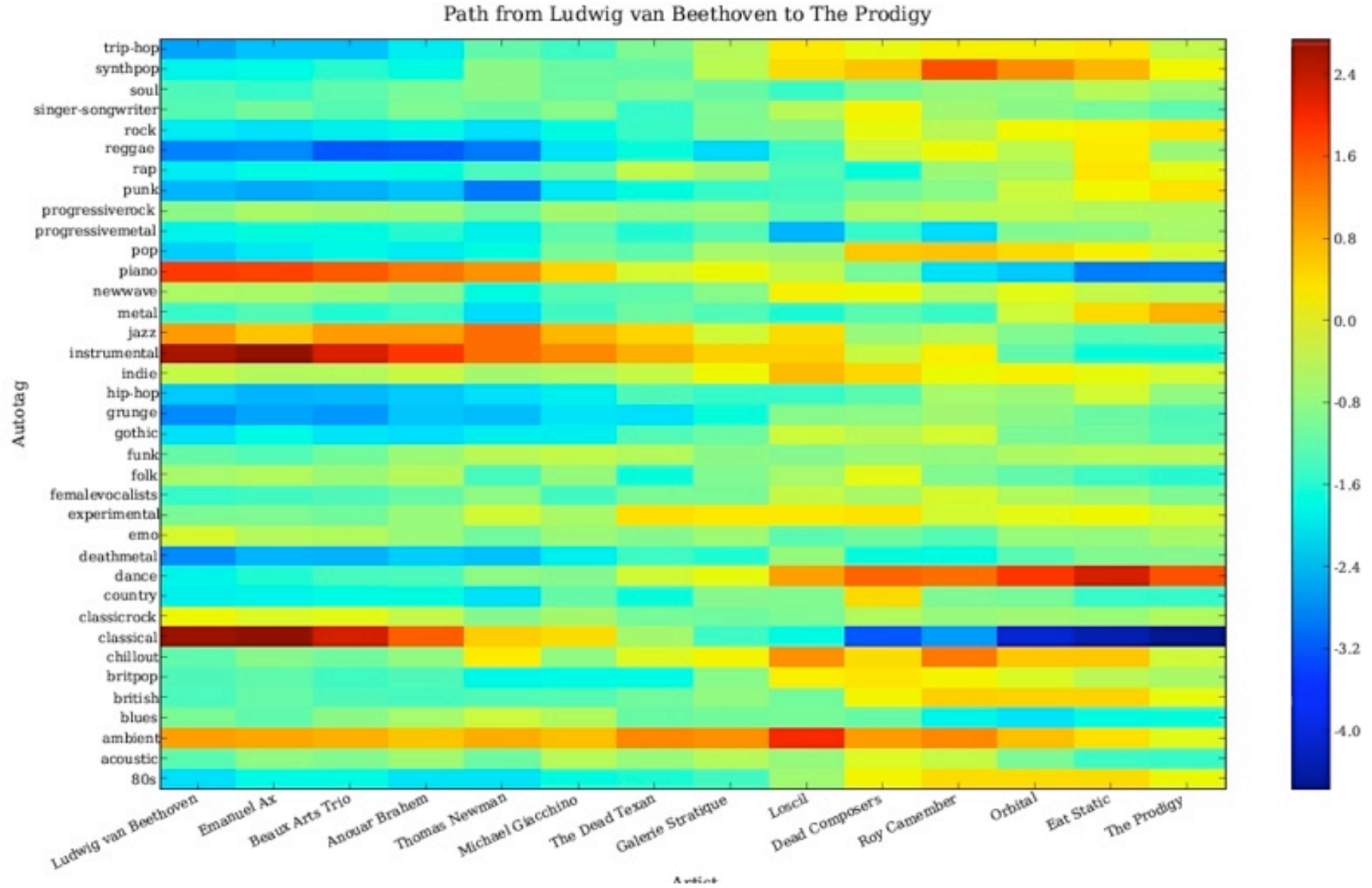
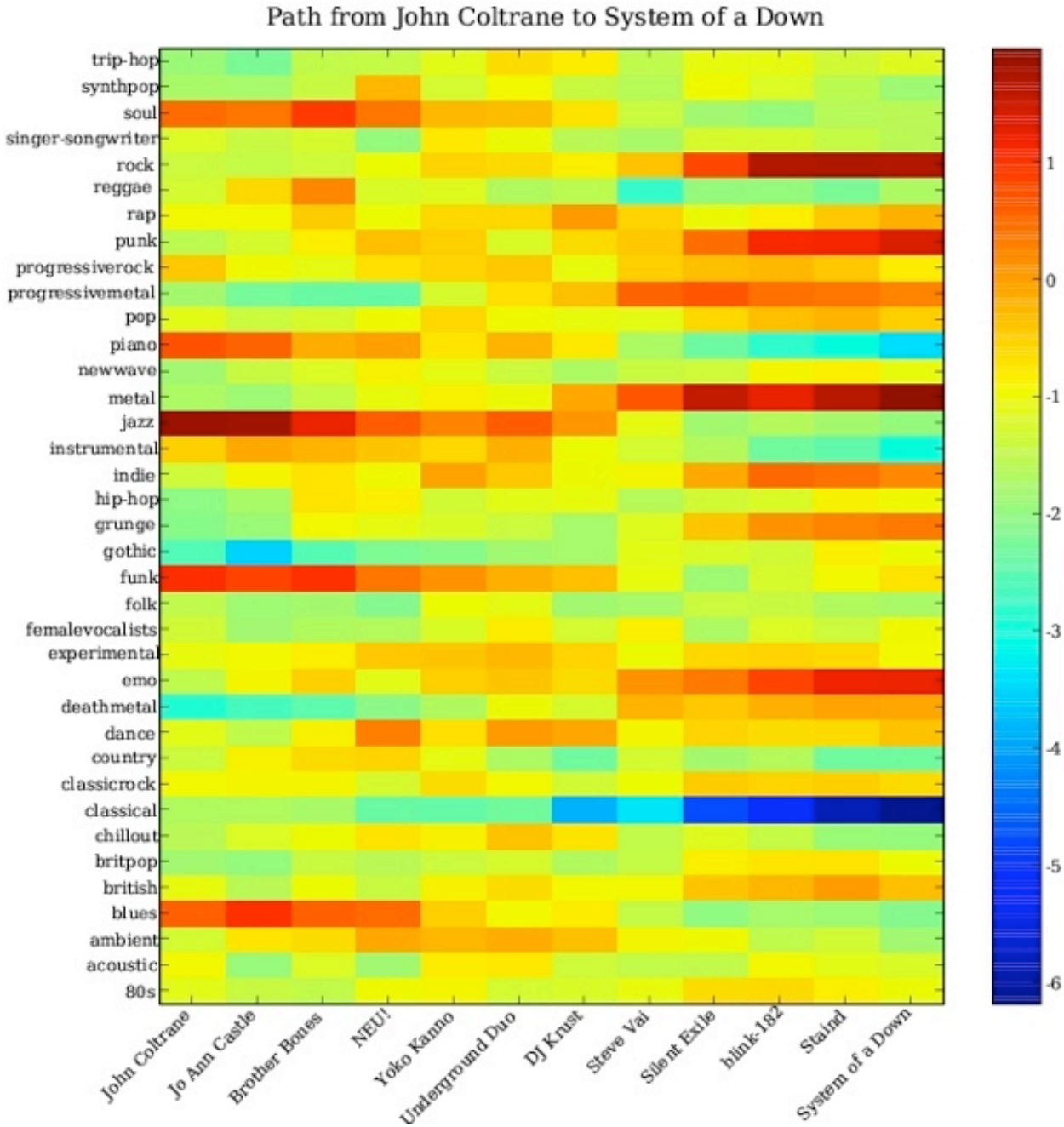


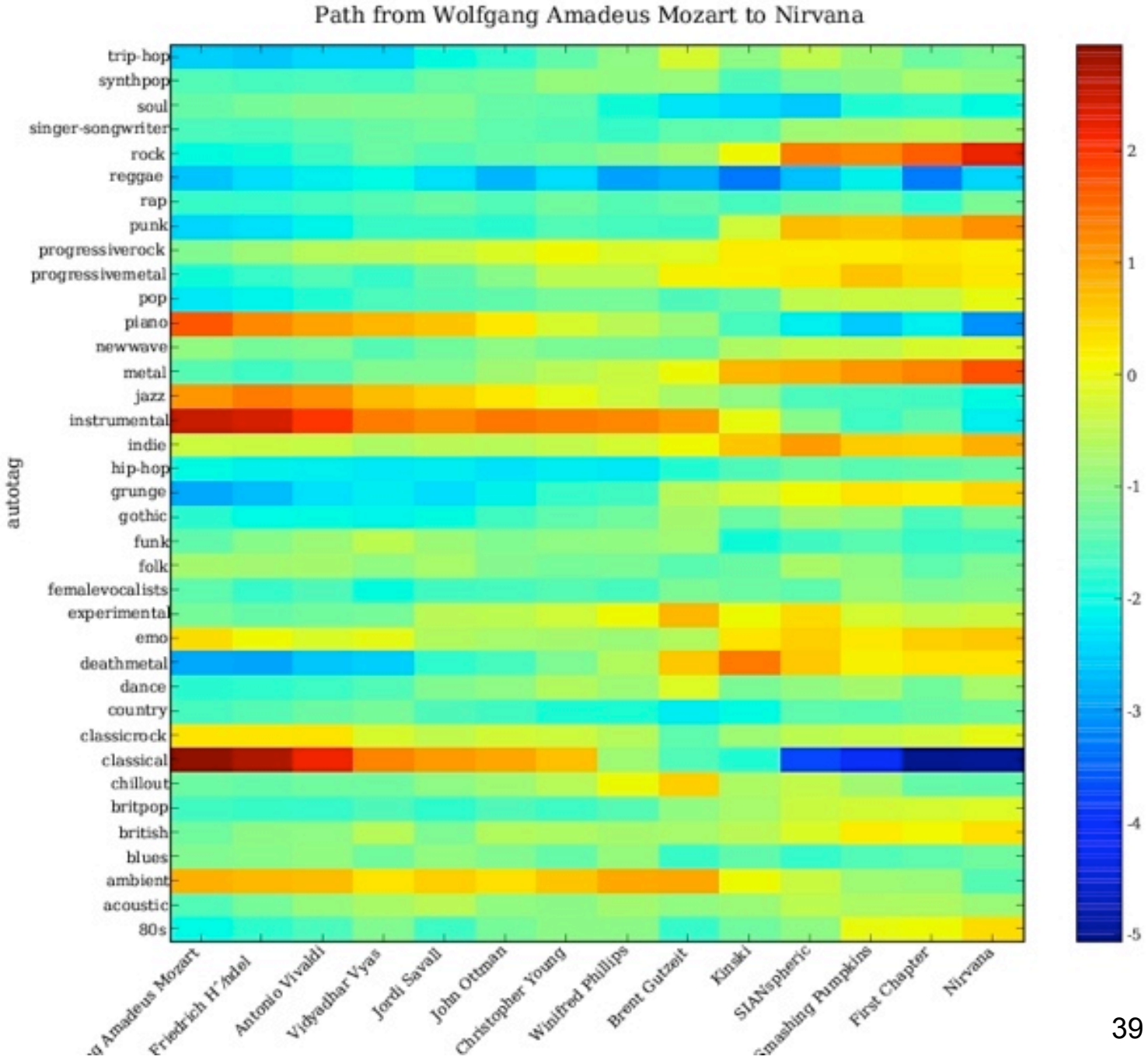
Figure 3: Similarity performance results when autotag similarities are blended with social tag similarities. The horizontal line is the performance of the social tags against ground truth.

Low dimensional music space

- Create low-dimensional space from autotags
- Find shortest path using Isomap

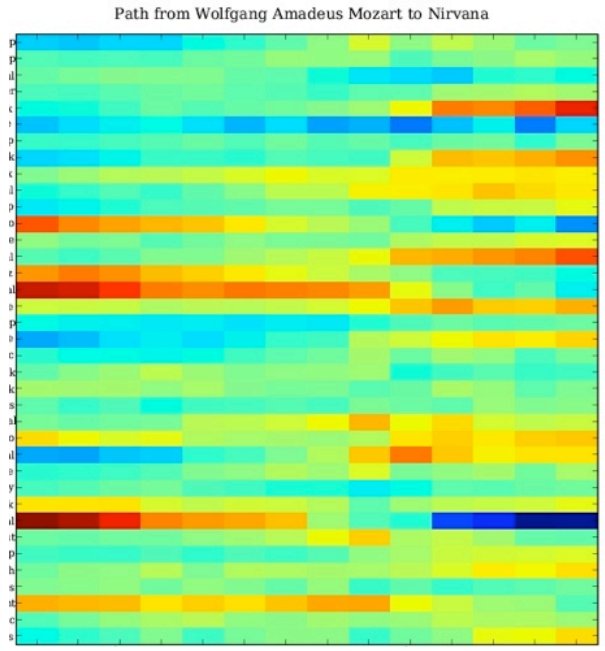

Path from Ludwig van Beethoven to The Prodigy
trip-hop
synthpop
soul
singer-songwriter
rock
reggae
rap
punk
progressiverock
progressivemetal
pop
piano
newwave
metal
jazz
instrumental
indie
hip-hop
grunge
gothic
funk
folk
femalevocalists
experimental
emo
deathmetal
dance
country
classicrock
classical
chillout
britpop
british
blues
ambient
acoustic
80s
Autotag
Ludwig van Beethoven
Emanuel Ax
Beaux Arts Trio
Anouar Brahem
Thomas Newman
Michael Giacchino
The Dead Texan
Galerie Stratique
Loscil
Dead Composers
Roy Camember
Orbital
Eat Static
The Prodigy
Artist
2.4
1.6
0.8
0.0
-0.8
-1.6
-2.4
-3.2
-4.0


Path from John Coltrane to System of a Down
trip-hop
synthpop
soul
singer-songwriter
rock
reggae
rap
punk
progressiverock
progressivemetal
pop
piano
newwave
metal
jazz
instrumental
indie
hip-hop
grunge
gothic
funk
folk
femalevocalists
experimental
emo
deathmetal
dance
country
classicrock
classical
chillout
britpop
british
blues
ambient
acoustic
80s
John Coltrane
Jo Ann Castle
Brother Bones
NEU!
Yoko Kanno
Underground Duo
DJ Krust
Steve Vai
Silent Exile
blink-182
Staind
System of a Down
1
0
-1
-2
-3
-4
-5
-6


Path from Wolfgang Amadeus Mozart to Nirvana
autotag
trip-hop
synthpop
soul
singer-songwriter
rock
reggae
rap
punk
progressiverock
progressivemetal
pop
piano
newwave
metal
jazz
instrumental
indie
hip-hop
grunge
gothic
funk
folk
femalevocalists
experimental
emo
deathmetal
dance
country
classicrock
classical
chillout
britpop
british
blues
ambient
acoustic
80s
Friedrich H˙ndel
Antonio Vivaldi
Vidyadhar Vyas
Jordi Savall
John Ottman
Christopher Young
Winifred Phillips
Brent Gutzeit
Kinski
SIANspheric
Smashing Pumpkins
First Chapter
Nirvana
2
1
0
-1
-2
-3
-4
-5

Compare two “classical” to “heavy”

Beethoven to The Prodigy



Mozart to Nirvana

Conclusions

- A simple framework for machine music listening:
  - Audio feature extraction + segmentation
  - Ensemble learning (boosting)
  - Classification of social tags using binning strategy

- Can be improved in many ways:
  - Ranking or regression
  - Regularization via weight sharing among song segments
  - Features derived from human audition
  - Many, many more features (source identification e.g. is there a female voice?)

- With relatively-simple machine learning, one can:
  - "Listen to" audio files to know their *musical* content
  - Embed songs and artists in a rich music space

- Important for companies like Last.FM and Pandora (music recommenders)

- Crucial for any company wanting to know what is going on with music on the web

Bibliography

- D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green. Automatic generation of social tags for music recommendation. In Neural Information Processing Systems Conference (NIPS) 20, 2007.

- D. Eck, T. Bertin-Mahieux, and P. Lamere. Autotagging music using supervised learning. In Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007), 2007. Submitted.

- J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl. Aggregate features and AdaBoost for music classification. Machine Learning, 65(2-3):473-484, 2006.

- Bergstra, A. Lacoste, and D. Eck. Predicting genre labels for artists using freedb. In Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006), 2006.

Questions to douglas.eck@umontreal.ca / douglas.eck@gmail.com