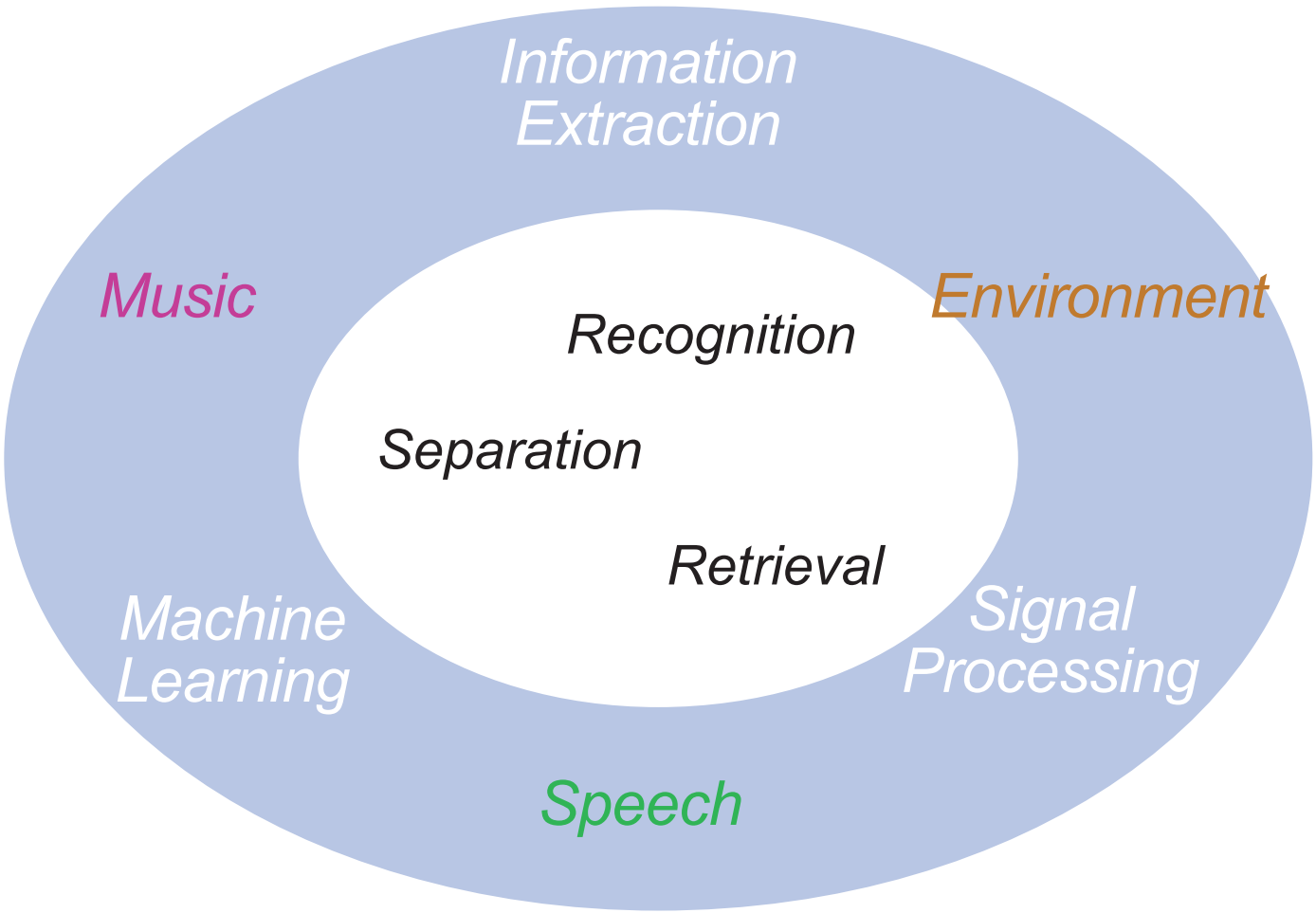
Mining for the Meaning of Music

Dan Ellis
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Engineering, Columbia University, NY USA
http://labrosa.ee.columbia.edu/

1. Motivation: Oodles of Music
2. Eigenrhythms & Eigenmelodies
3. Melodic-Harmonic Fragments
4. Other Projects

LabROSA Overview



1. Motivation: Oodles of Music

- The impact of the iPod
  - creates new research questions (music IR)
  - but also: provides new tools for old questions

- What can you do with 100k+ tracks?
  - around 9 months of listening..
  - unsupervised data

"The Meaning of Music"

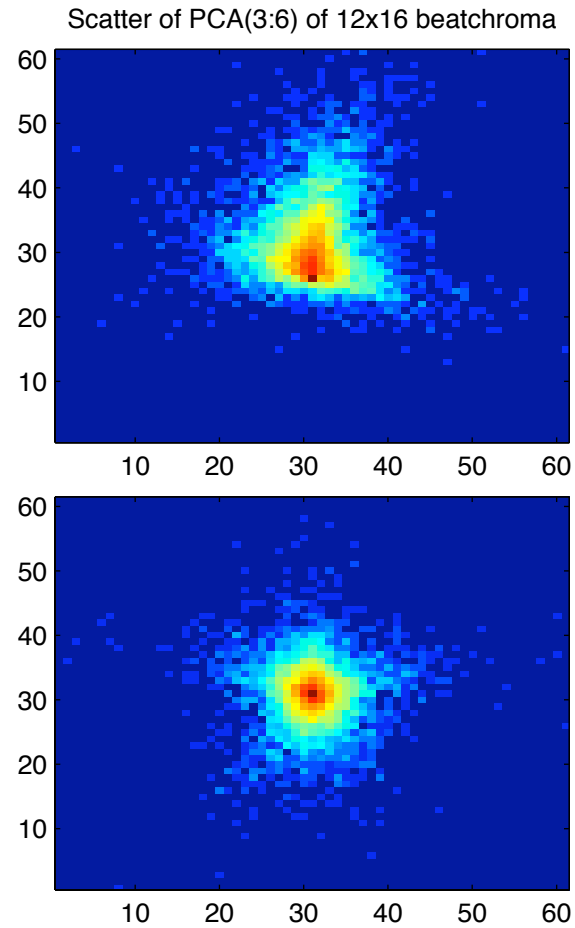
Two kinds of "meaning":

- What does music evoke in a listener's mind?
  - i.e. "what does it all mean?" (metaphysics?)
  - study with subjective experiments
  - (then build detectors for specific responses ...?)

- What phenomena are denoted by "music"?
  - i.e. delineate the "set of all music"
  - (the ultimate music/nonmusic classifier?)
  - .. this talk's topic

Re-used Musical Elements

- "What are the most popular chord progressions?"
  - a well-formed question...
  - music occupies a small subset of some space
  - look at massive audio archive?
- How can we distill a large collection of music audio into a compact description of what "music" means?
  - or at least a vocabulary...



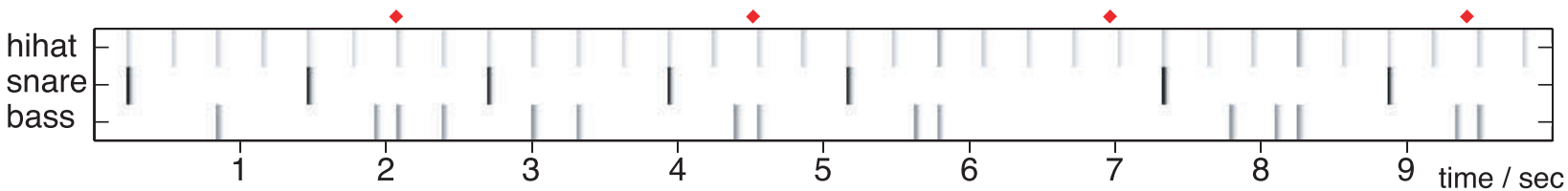
Potential Applications

- Given a description of the musically valid subspace...
  - compression: represent a given piece by its indices/ parameters in the subspace
  - classification: subspace representation reveals 'essence'; neighbors are interesting
  - manipulation: modifications within the space remain musically valid

2. Eigenrhythms: Drum Track Structure

Ellis & Arroyo ISMIR'04

- To first order,
  All pop music has the same drum track:



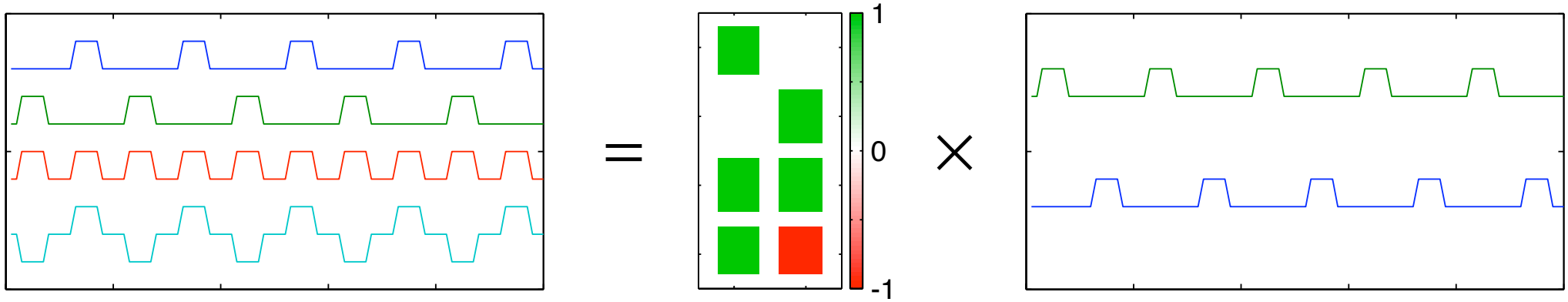
- Can we capture this from examples?
  - .. including the variations
- Can we exploit it?
  - .. for synthesis
  - .. for classification
  - .. for insight

Basis Sets

- Dataset reduced to linear combinations of a few basic patterns

weights

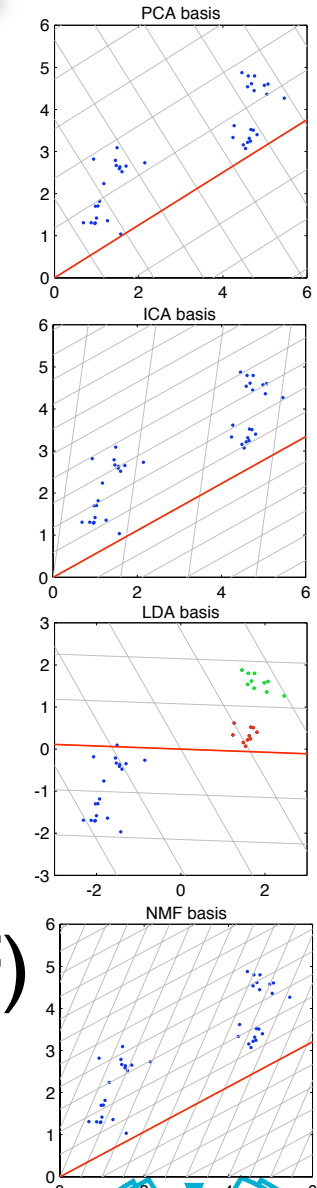
data $\mathbf{X} = \mathbf{W} \times \mathbf{H}$ *bases*



- bases $\mathbf{H}$: subspace that spans the data
- weights $\mathbf{W}$: dimension-reduced projection of data

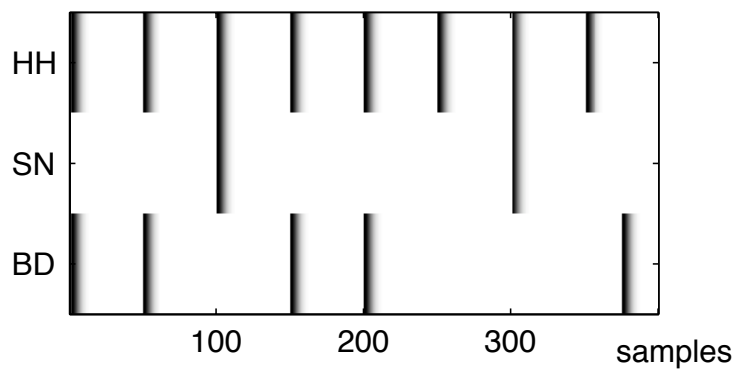
Different basis projections

- Principal Component Analysis (PCA)
  - optimizes MSE of low-D reconstruction
- Independent Component Analysis (ICA)
  - projections are independent (cf decorrelated)
- Linear Discriminant Analysis (LDA)
  - given class labels for data, find dimensions to separate them
- Nonnegative Matrix Factorization (NMF)
  - each basis function only adds bits in



Data

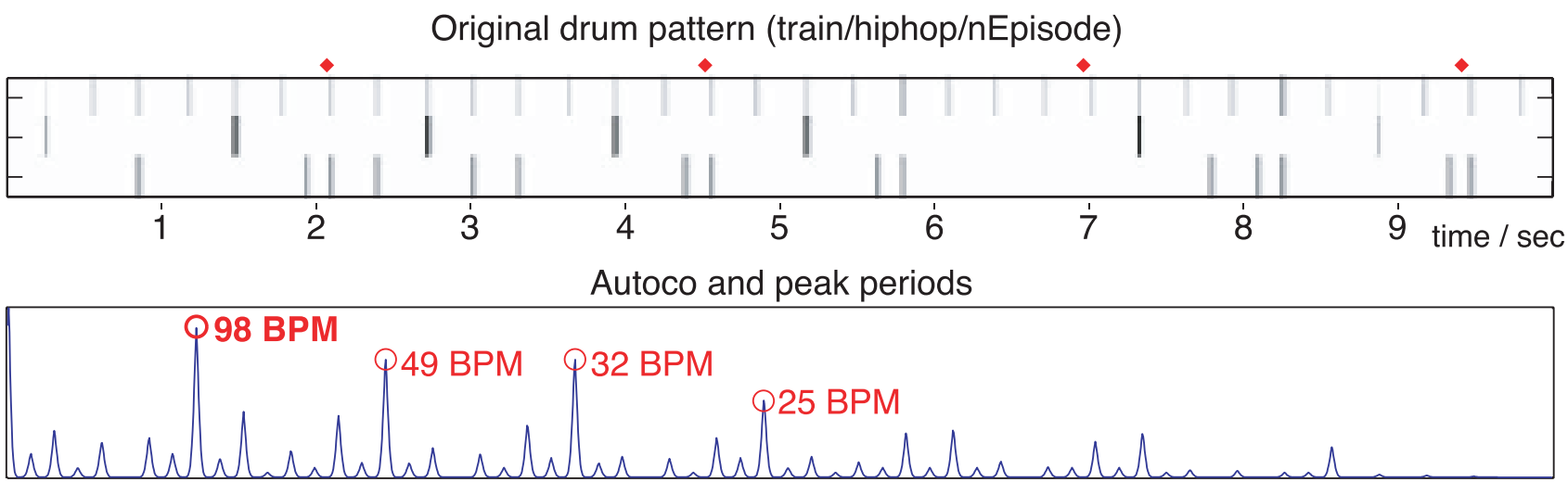
- **Drum tracks extracted from MIDI**
  - 100 examples (10 × 10 genre classes)
  - fixed mapping to 3 instruments: bass drum, snare, hi-hat
  - temporary proxy for audio transcription...
- **Pseudo-envelope representation**
  - 40ms half-Gauss window sampled at 200 Hz



- **Extract just one pattern from each MIDI**
  - looking for variety, quantity not a problem

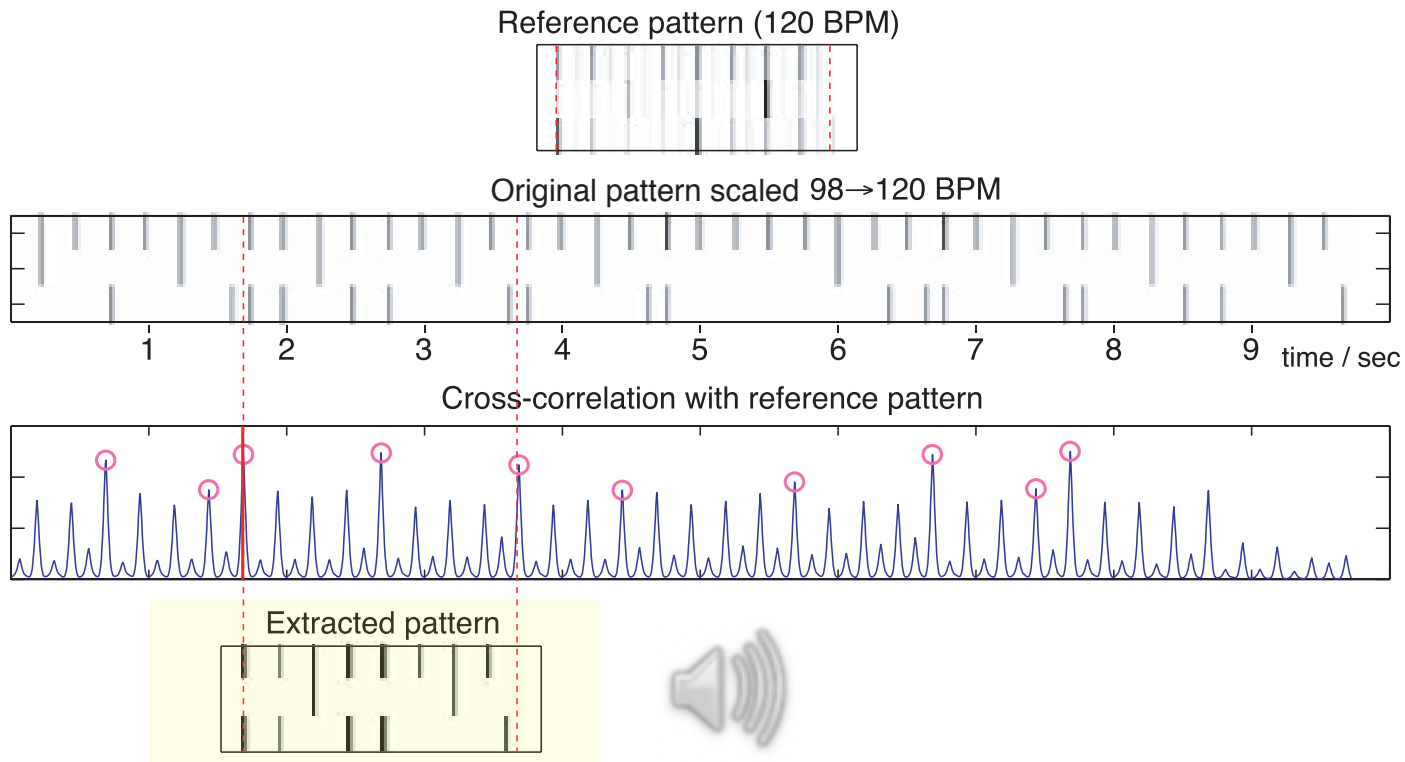
Aligning Data: Tempo

- Need to align patterns prior to PCA/...
- First, normalize tempo
  - autocorrelation gives BPM candidates
  - keep them all for now...



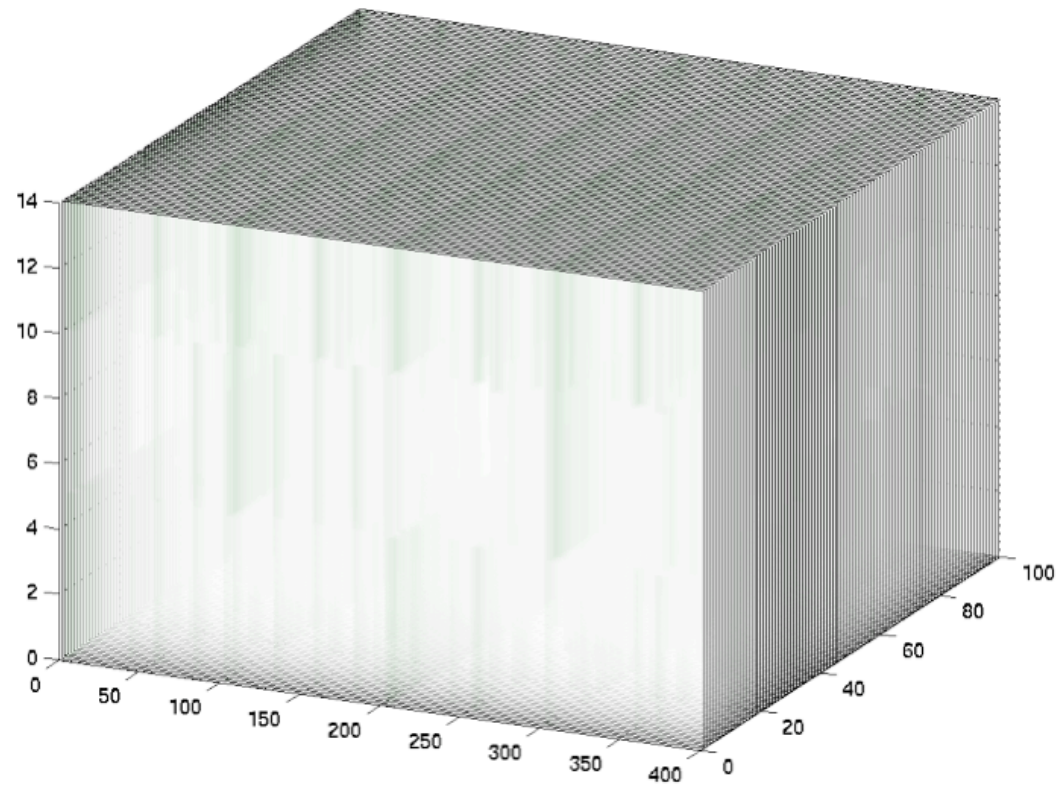
Aligning Data: Downbeat

- Downbeat from best match of tempo-normalized pattern to mean template
  - try every tempo hypotheses, choose best match



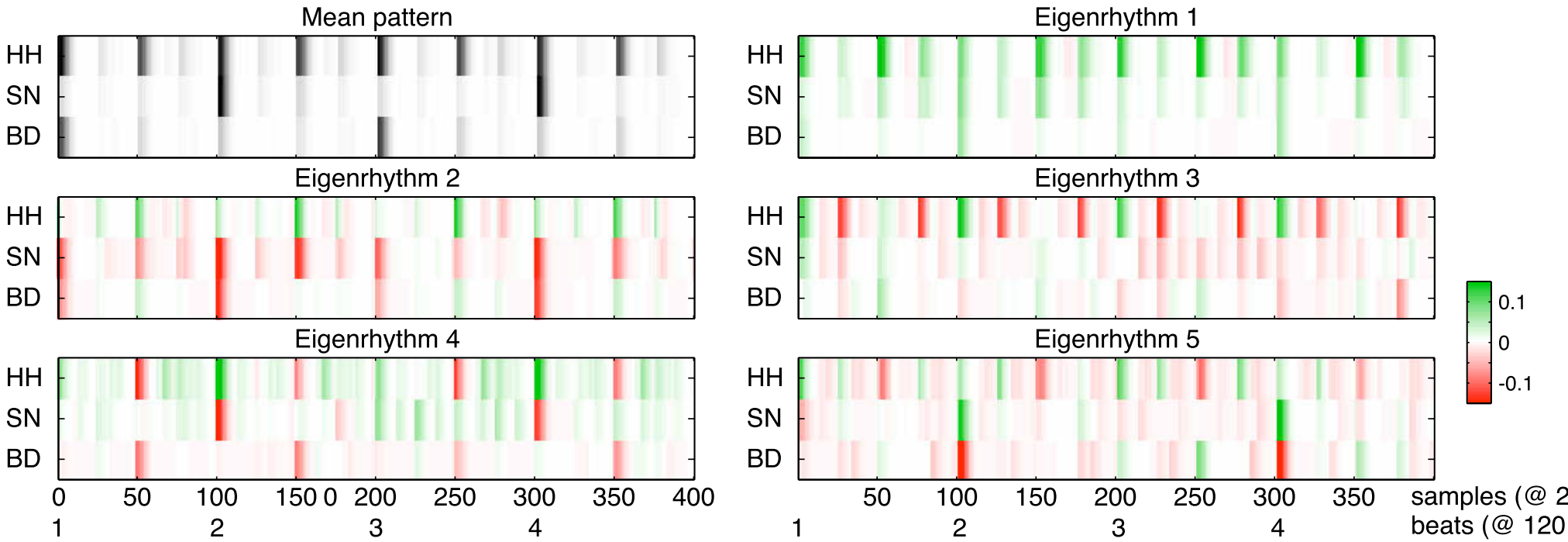
Aligned Data

- Tempo normalization + downbeat alignment
  → 100 excerpts (2 bars each):



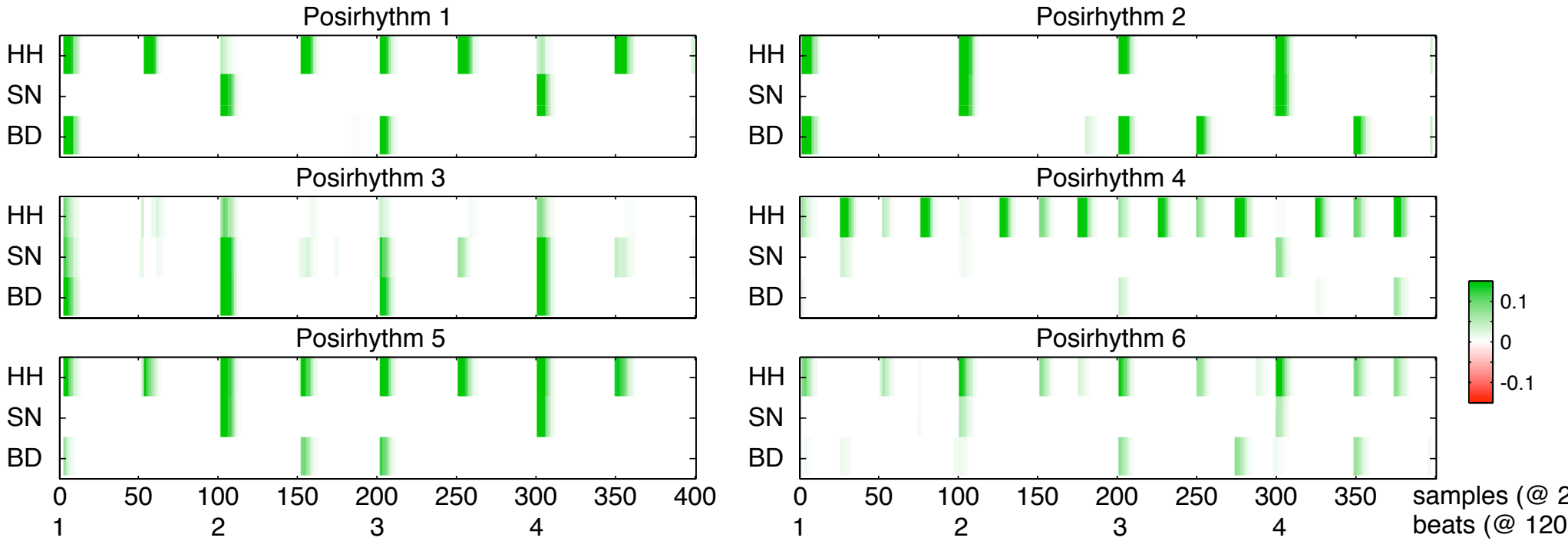
- Can now extract basis projection(s)

Eigenrhythms (PCA)



- Need 20+ Eigenvectors for good coverage of 100 training patterns (1200 dims)
- Eigenrhythms both add and subtract

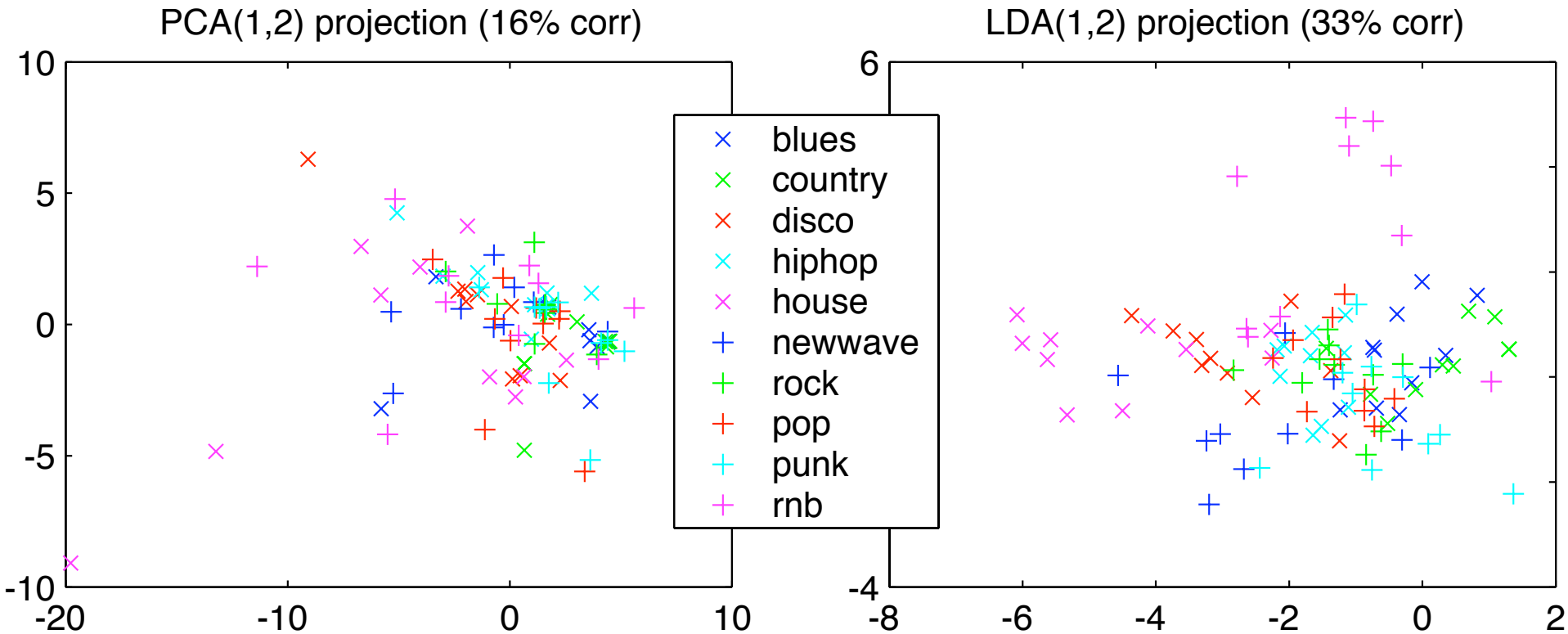
Posirhythms (NMF)



- Nonnegative: only adds beat-weight
- Capturing some structure

Eigenrhythms for Classification

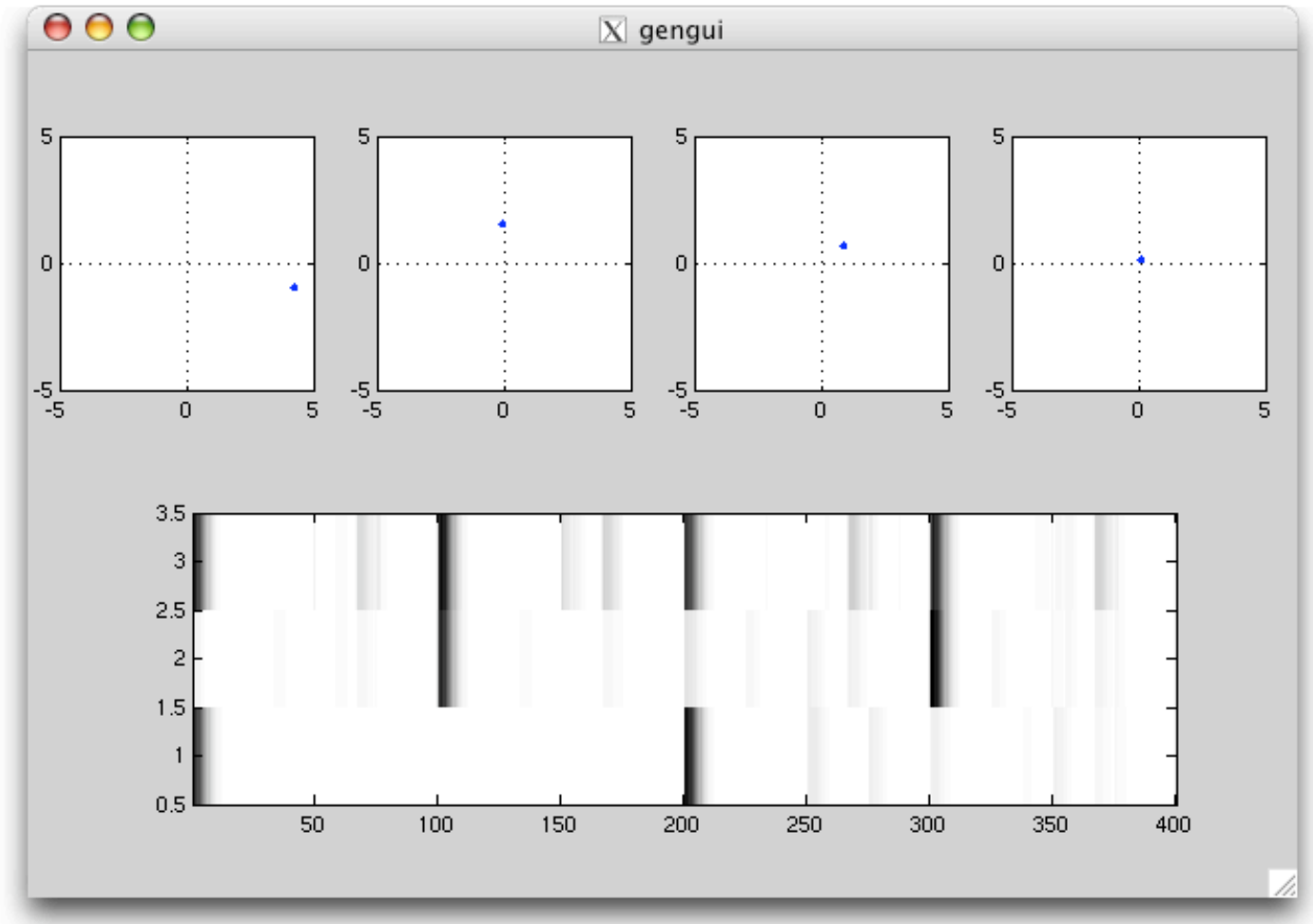
- Projections in Eigenspace / LDA space



- 10-way Genre classification (nearest nbr):
  - PCA3: 20% correct
  - LDA4: 36% correct

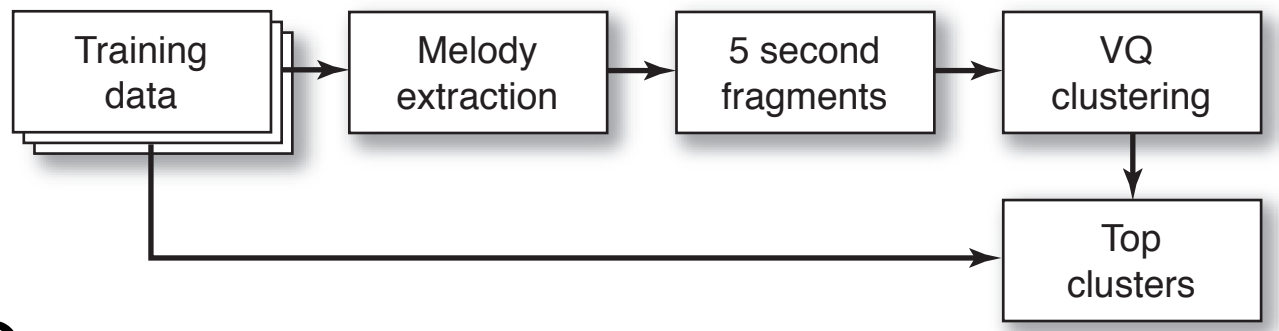
Eigenrhythm BeatBox

- Resynthesize rhythms from eigen-space



Eigenmelodies?

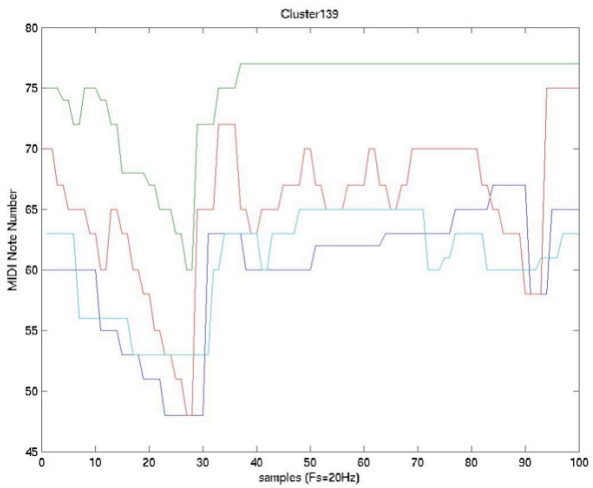
- Can we do a similar thing with melodies?
- Cluster 'fragments' that recur in melodies
  - .. across large music database
  - .. one way to get fragment alignment?
  - .. trade data for model sophistication

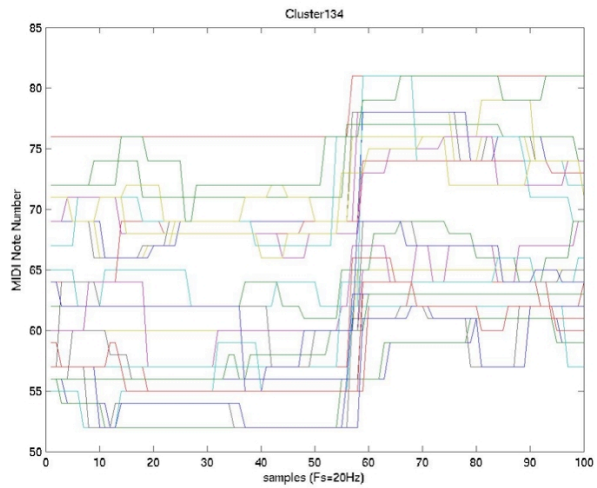


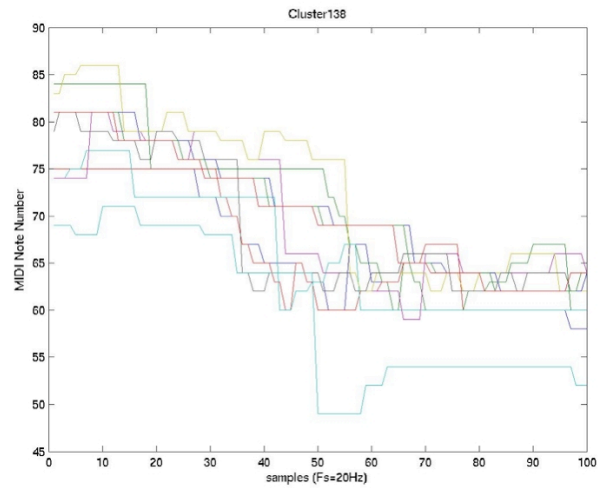
- Data sources
  - pitch tracker, or MIDI training data
- Melody fragment representation
  - DCT(1:20) - removes average, smoothes detail

Melody Clustering

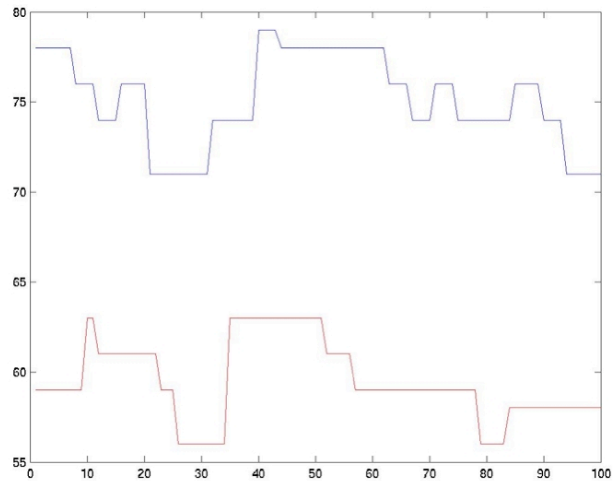
- Clusters match underlying contour:







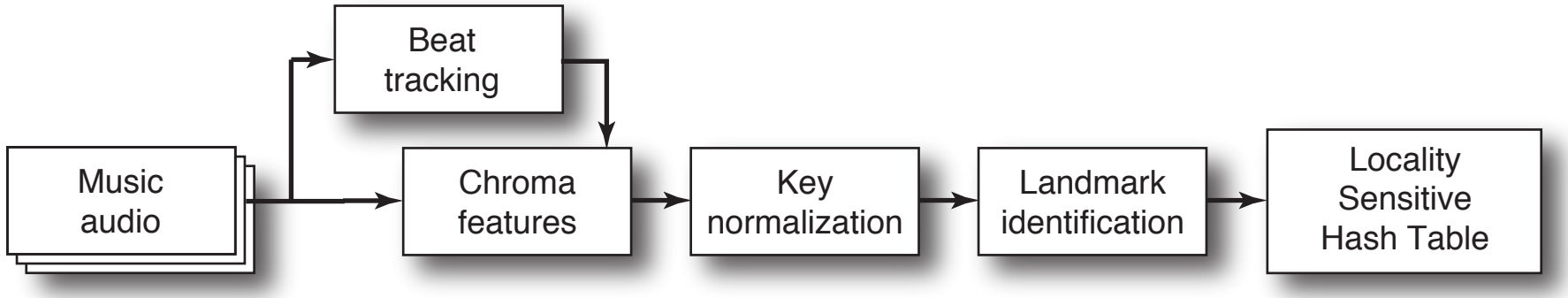
- Some interesting matches:
  - e.g. Pink + Nsync



3. Melodic-Harmonic Fragments

- Can we use the subspace and clustering ideas with our oodles of music?
  - use lots of real music audio
  - capture both melodic and harmonic context...
- Goal: Dictionary of common motifs (clichés)
  - build up into longer sequences
  - reveal quotes & inspirations, genre/style idioms
- Questions
  - what representation and similarity measure?
  - what clustering scheme?
  - tractability: how large can we go?

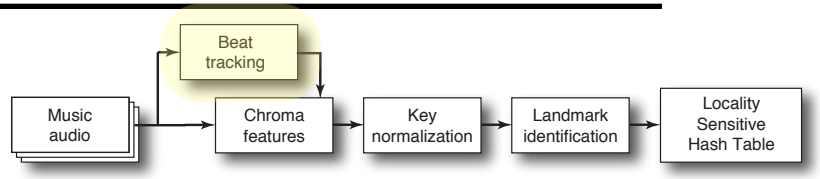
Finding Common Fragments



- Chop up music into short descriptions of musical content
  - 24-beat beat-chroma matrices
- Choose a few at "starts" (landmarks)
- Put into LSH table
  - similar items fall in same bin
- Find the bins with most entries

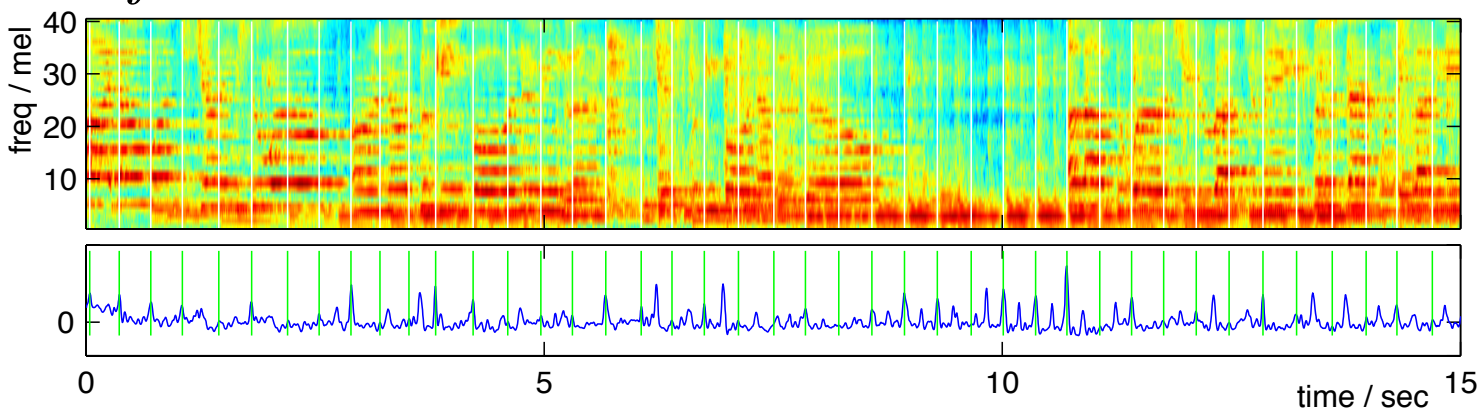
= most commonly reused motifs

Beat Tracking

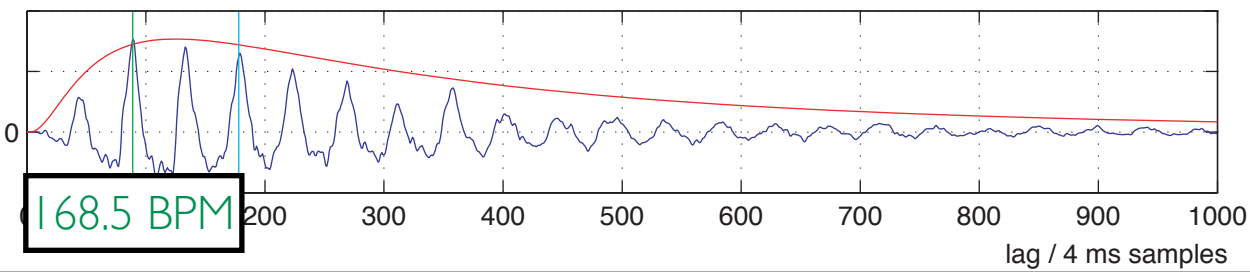


Ellis '06,'07

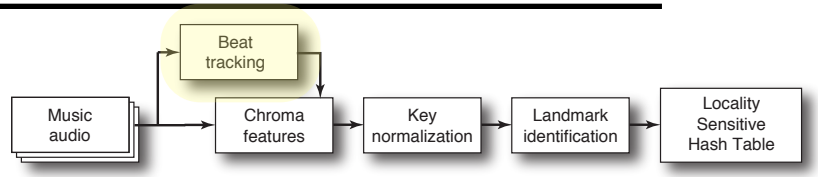
- Goal: Per-'beat' (tatum) feature vector
  - for tempo normalization, efficiency
- "Onset Strength Envelope"
  - $\text{sum}_f(\max(0, \text{diff}_t(\log |X(t,f)|)))$



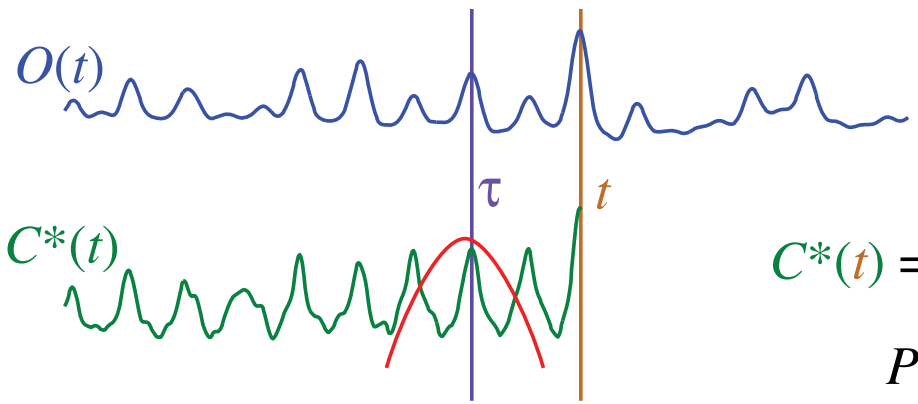
- Autocorr. + window → global tempo estimate



Beat Tracking (2)



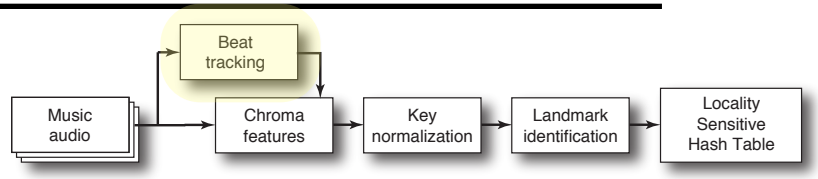
- Dynamic Programming finds beat times $\{t_i\}$
  - optimizes $\Sigma_i\, O(t_i) + \alpha\, \Sigma_i\, F(t_{i+1} - t_i\ ,\ \tau_p)$
  - where $O(t)$ is onset strength envelope (local score)
    $W(t)$ is a log-Gaussian window (transition cost)
    τ_p is the default beat period per measured tempo
  - incrementally find best predecessor at every time
  - backtrace from largest final score to get beats



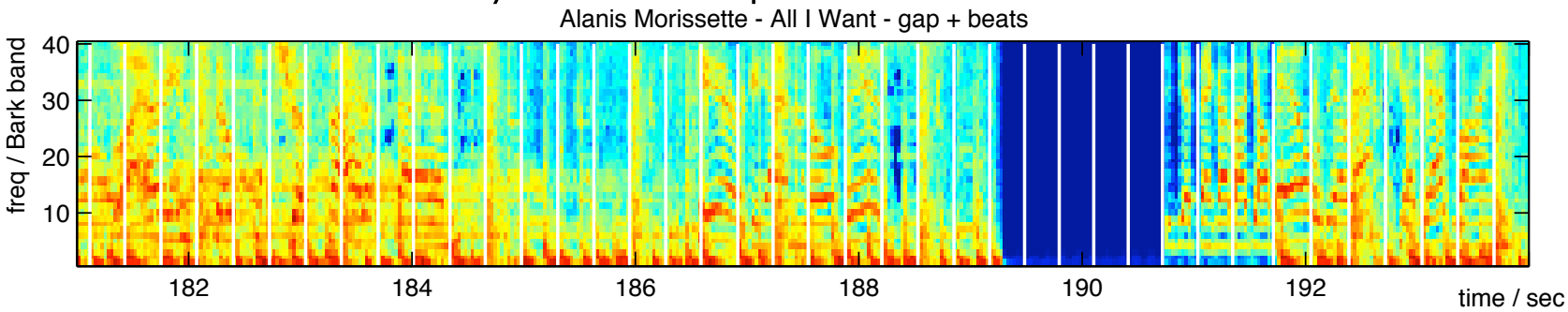
$$C^*(t) = O(t) + \max_{\tau}\{\alpha F(t-\tau, \tau_p) + C^*(\tau)\}$$

$$P(t) = \operatorname{argmax}_{\tau}\{\alpha F(t-\tau, \tau_p) + C^*(\tau)\}$$

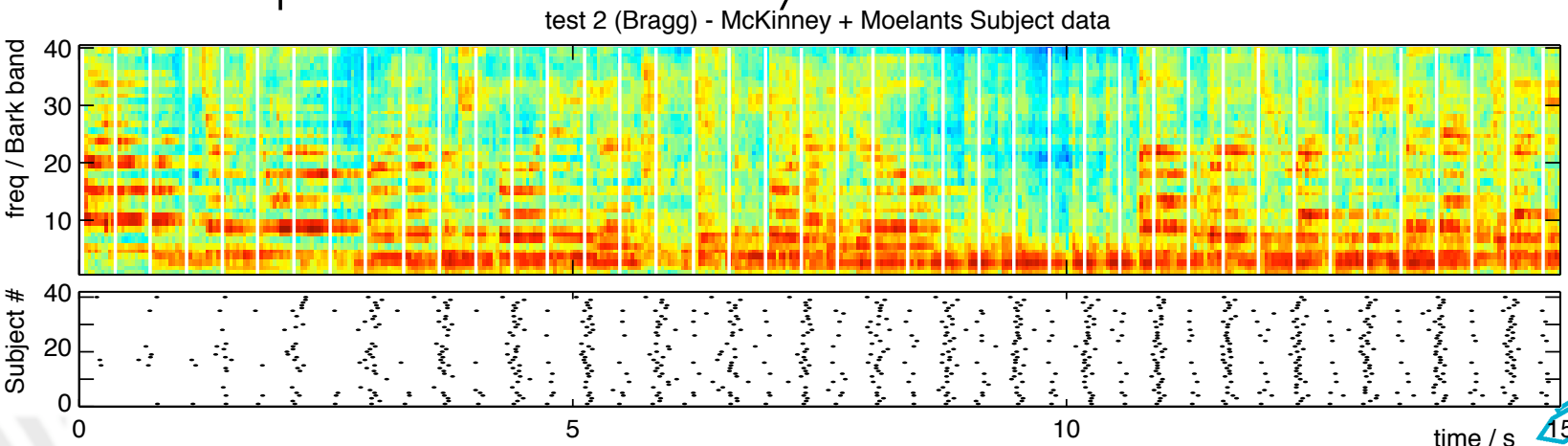
Beat Tracking (3)



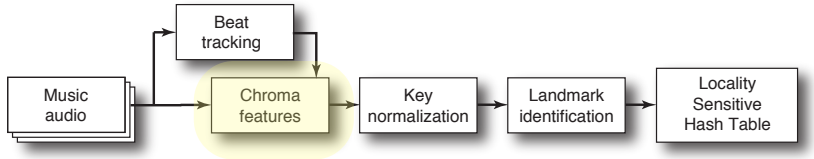
- DP will bridge gaps (non-causal)
  - there is always a best path ...



- 2nd place in MIREX 2006 Beat Tracking
  - compared to McKinney & Moelants human data

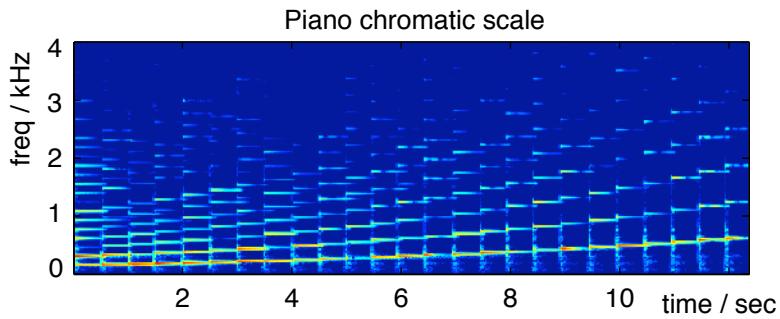


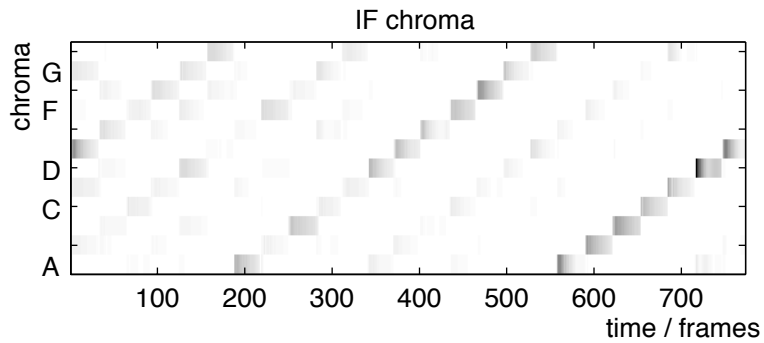
Chroma Features



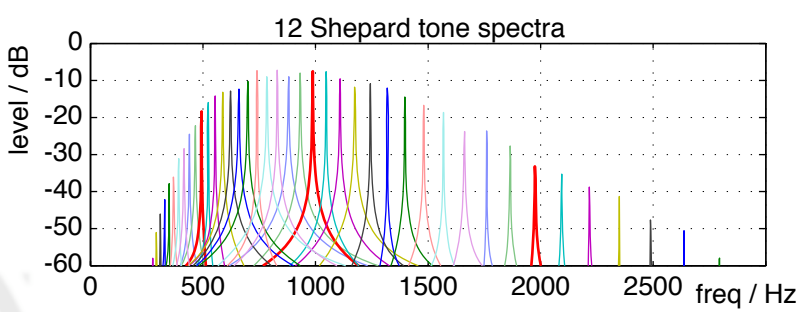
- Chroma features convert spectral energy into musical weights in a canonical octave
  - i.e. 12 semitone bins

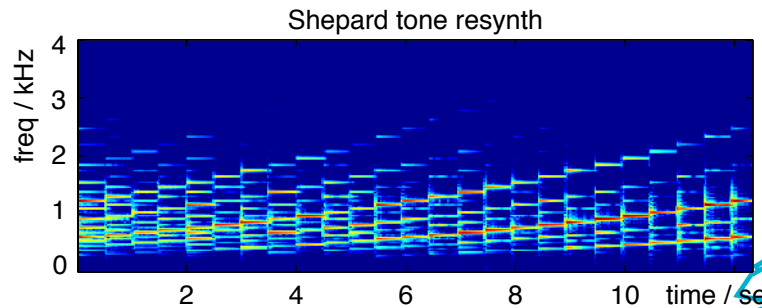
Piano scale



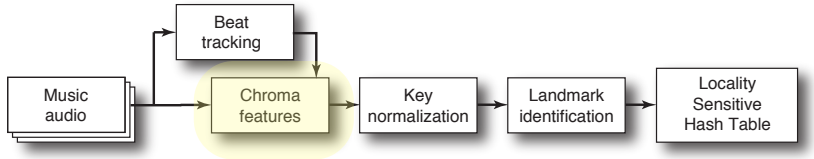


- Can resynthesize as "Shepard Tones"
  - all octaves at once

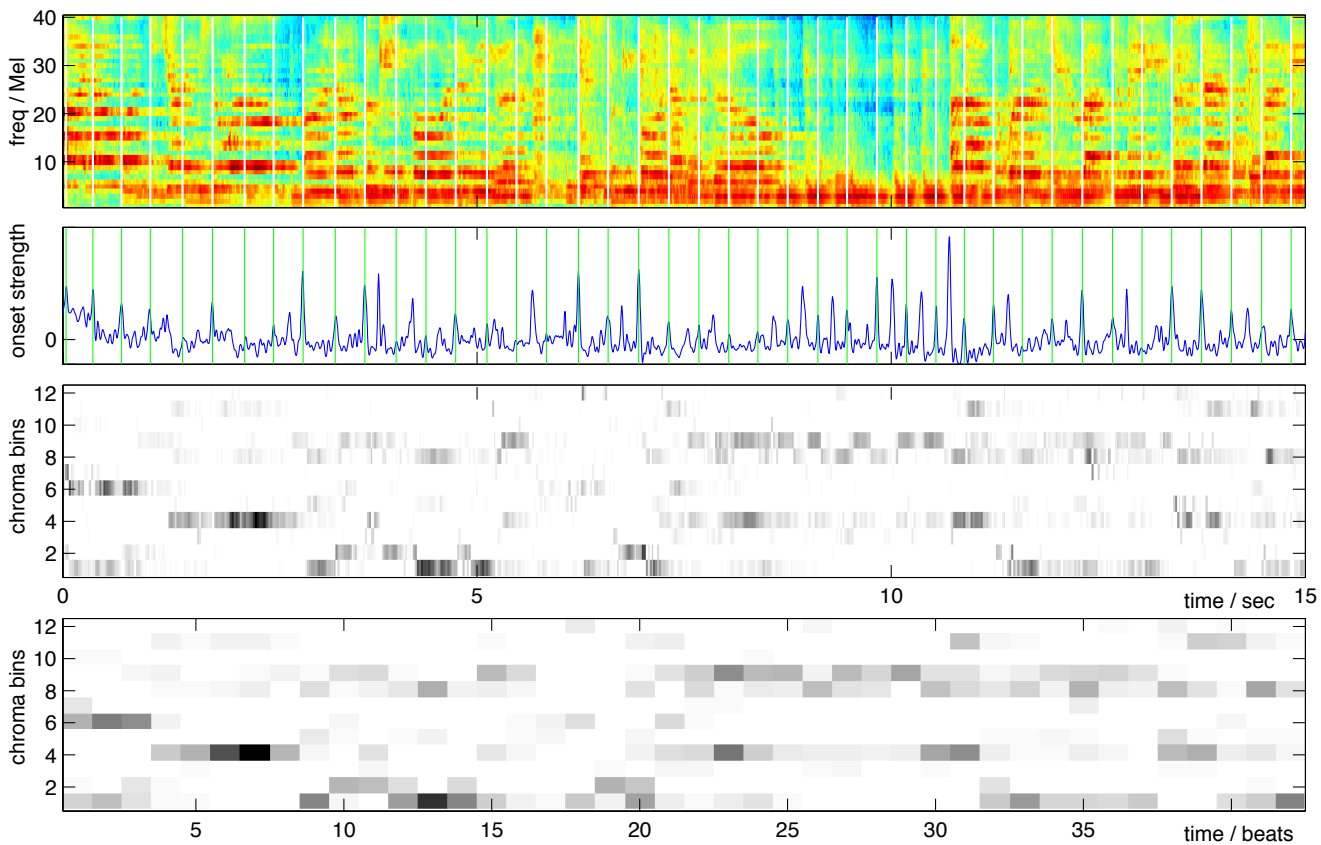




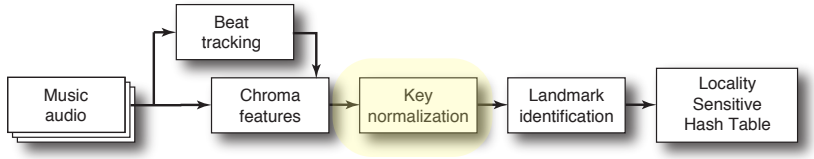
Beat-Chroma Features



- Beat + chroma features / 30ms frames
  → average chroma within each beat
  - compact; sufficient?

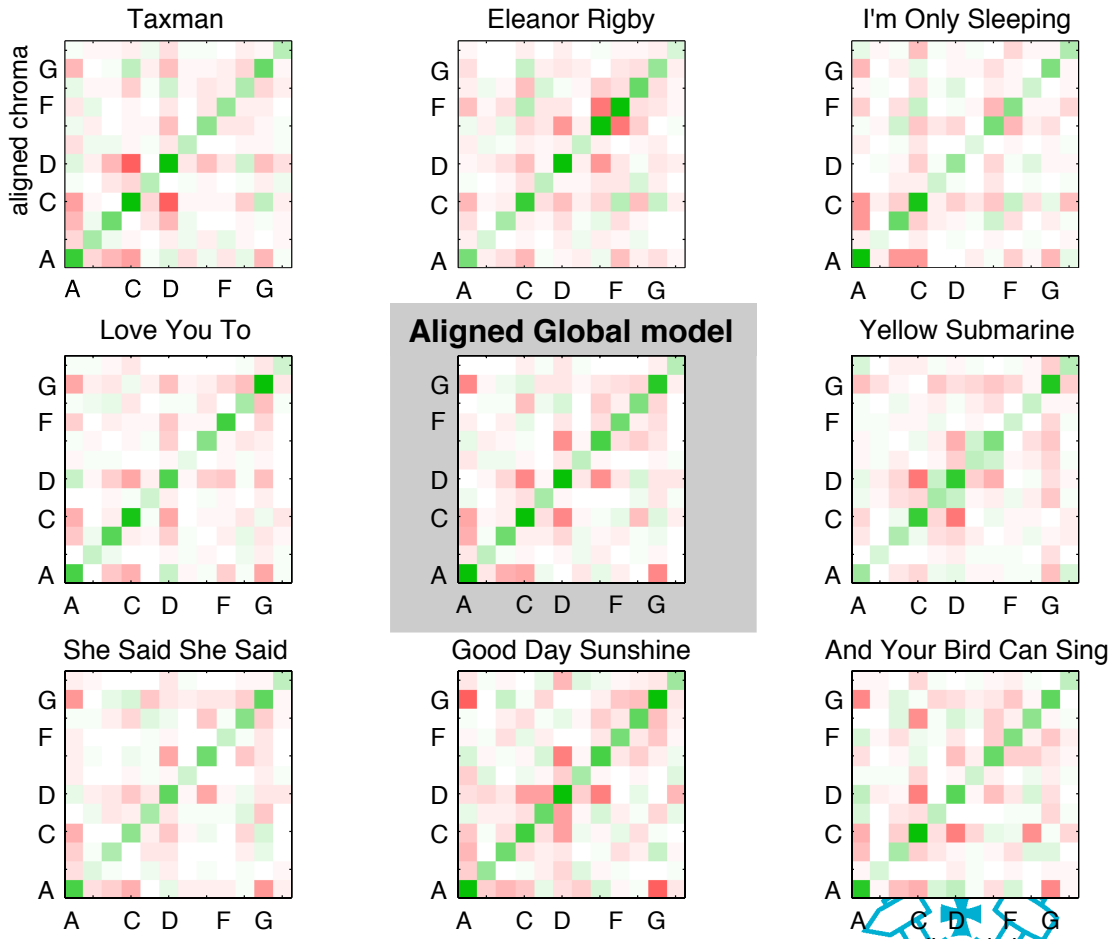


Key Estimation

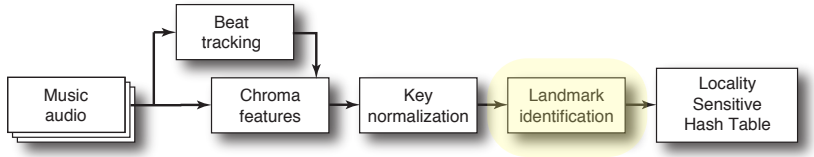


Ellis ICASSP '07

- Covariance of chroma reflects key
- Normalize by transposing for best fit
  - single Gaussian model of one piece
  - find ML rotation of other pieces
  - model all transposed pieces
  - iterate until convergence

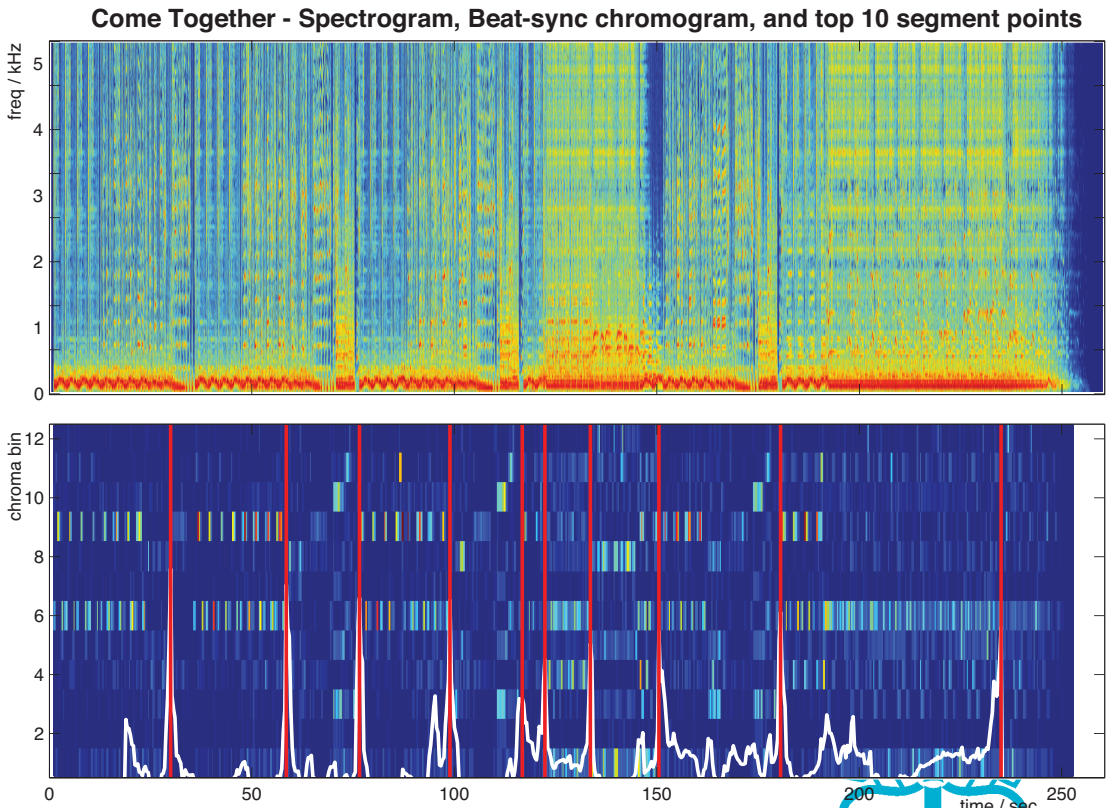


Landmark Location

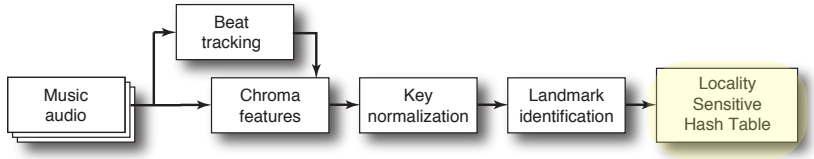


- Looking for "beginnings" of phrases
  - e.g. abrupt change in harmony, instruments, etc.
  - use likelihood ratio test: data following under model up to boundary

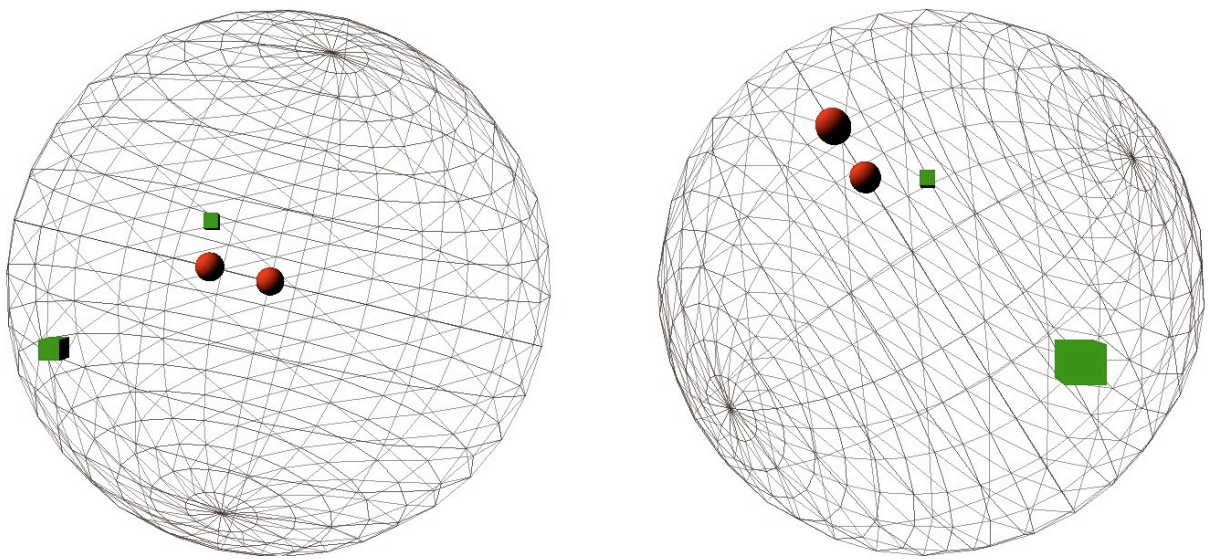
- Choose top 10 locally-normalized peaks
  - .. to control data size
  - ? include ± 2 beats to catch errors



Locality Sensitive Hash



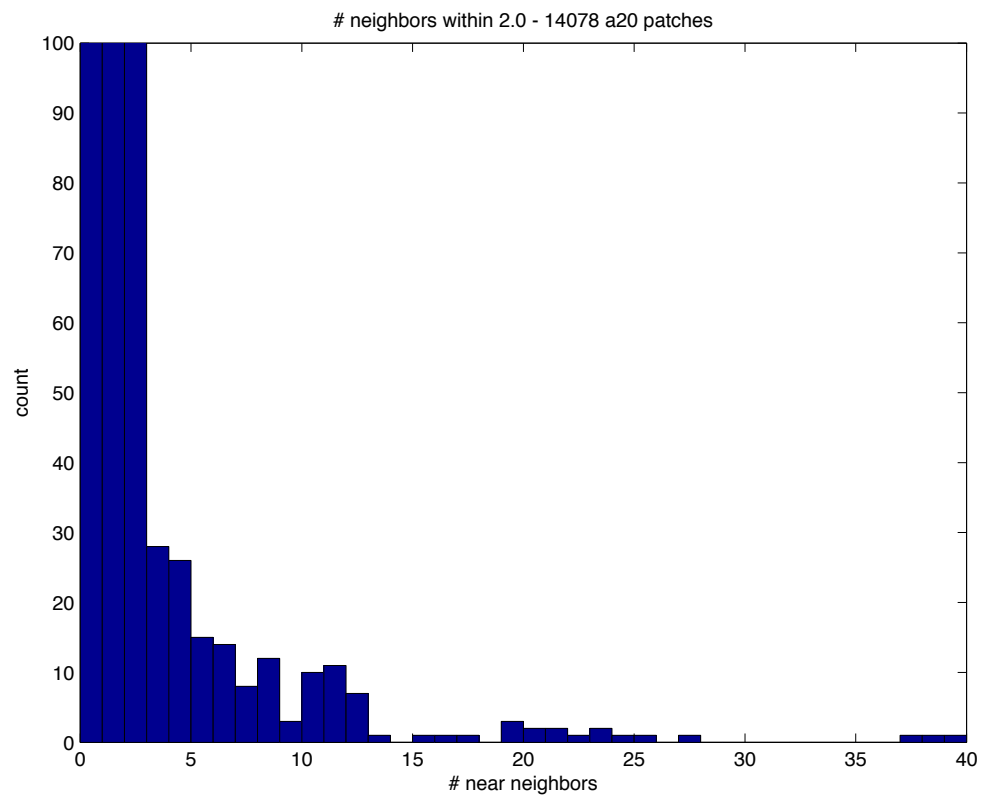
- Goal: Quantize high-dimensional data so 'similar' items fall into same bin
  - .. for fast and scalable nearest-neighbor search
- Idea: Multiple random scalar projections
  - each one will tend to keep neighbors nearby
  - items close together in all projections are probably neighbors



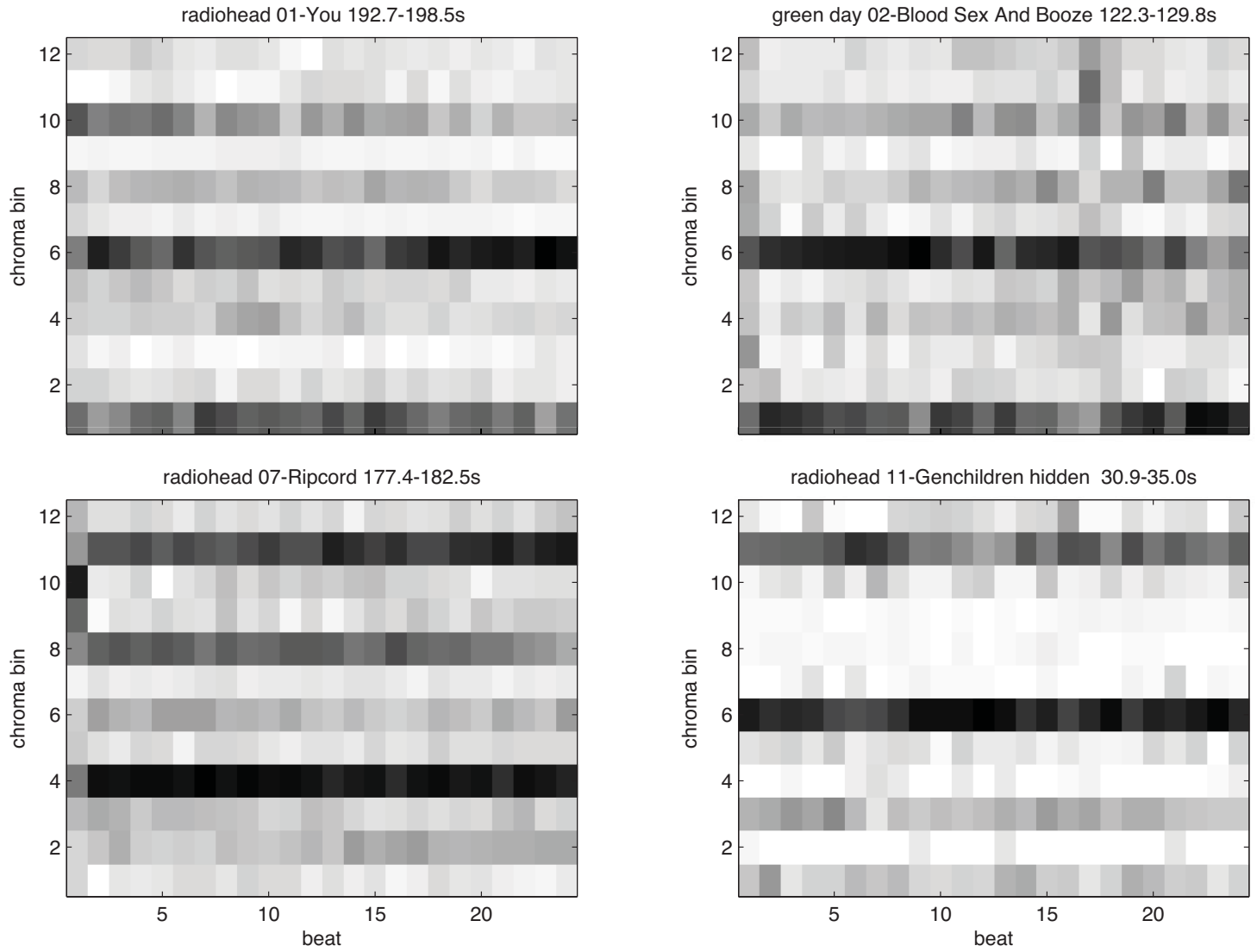
from Slaney & Casey '08

Experiments

- Data
  - "artist20" - 20 artist x 6 albums = 1413 tracks
  - (up to) 10 landmarks/track = 14,078 patches
  - each patch = 12 chroma bins x 24 beats (288 dims)
- Performance
  - feature calculation: ~ 60 min
  - LSH 14k NNs: ~ 30 sec
  - 51 patches have >10 NNs within r = 2.0



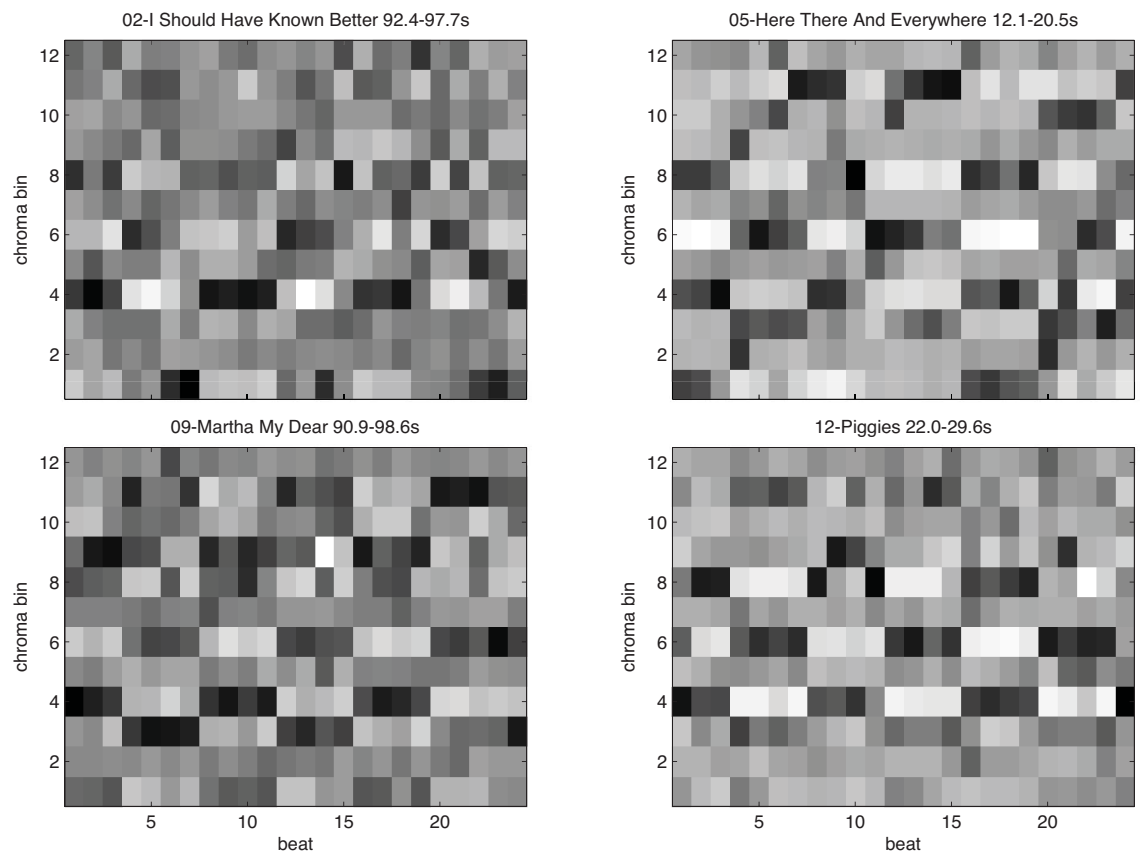
Results - artist20



- mainly sustained notes

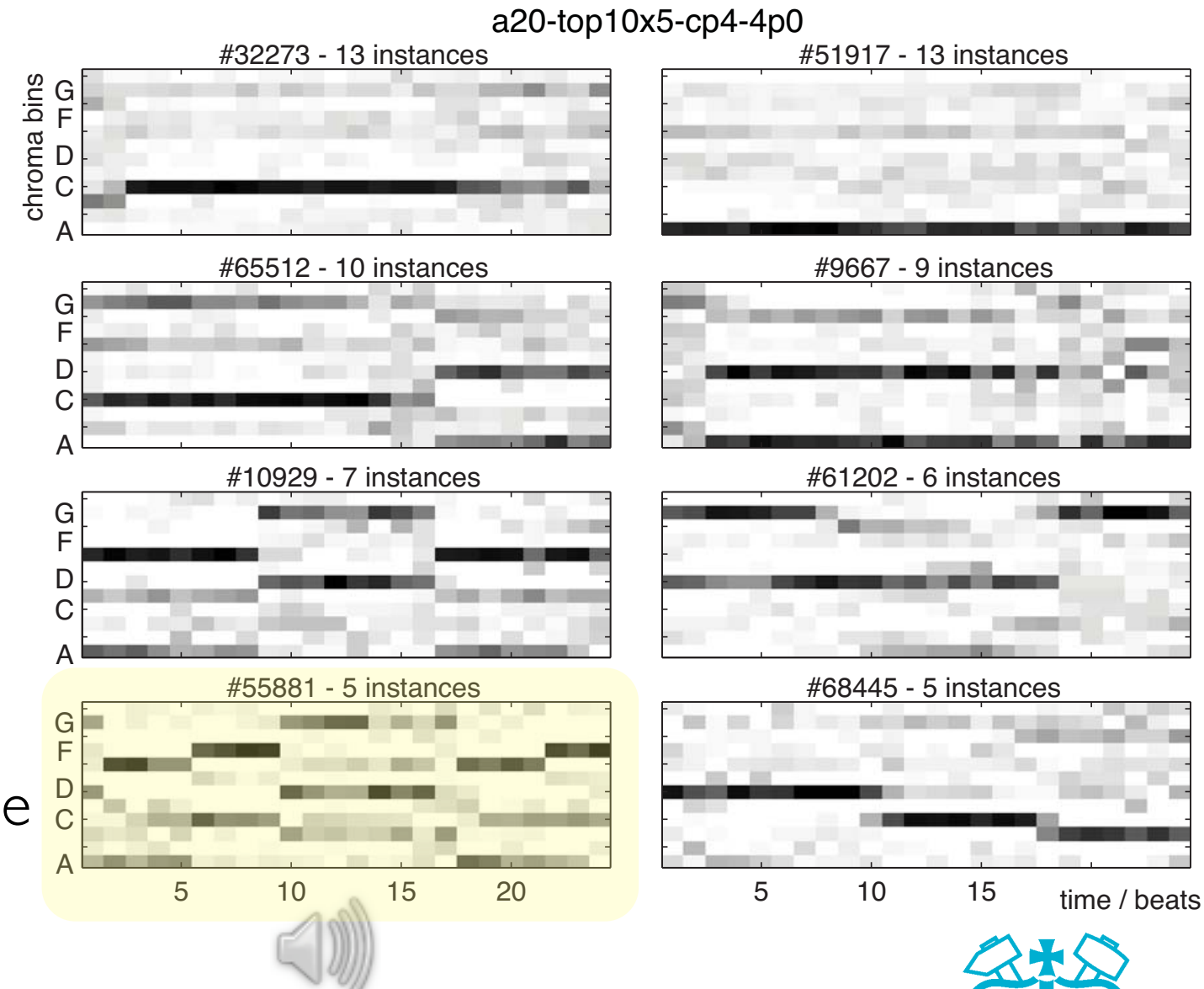
Results - Beatles

- Only the 86 Beatles tracks
- All beat offsets = 41,705 patches
  - LSH takes 300 sec - approx NlogN in patches?
- High-pass along time
  - to avoid sustained notes
- Song filter
  - remove hits in same track



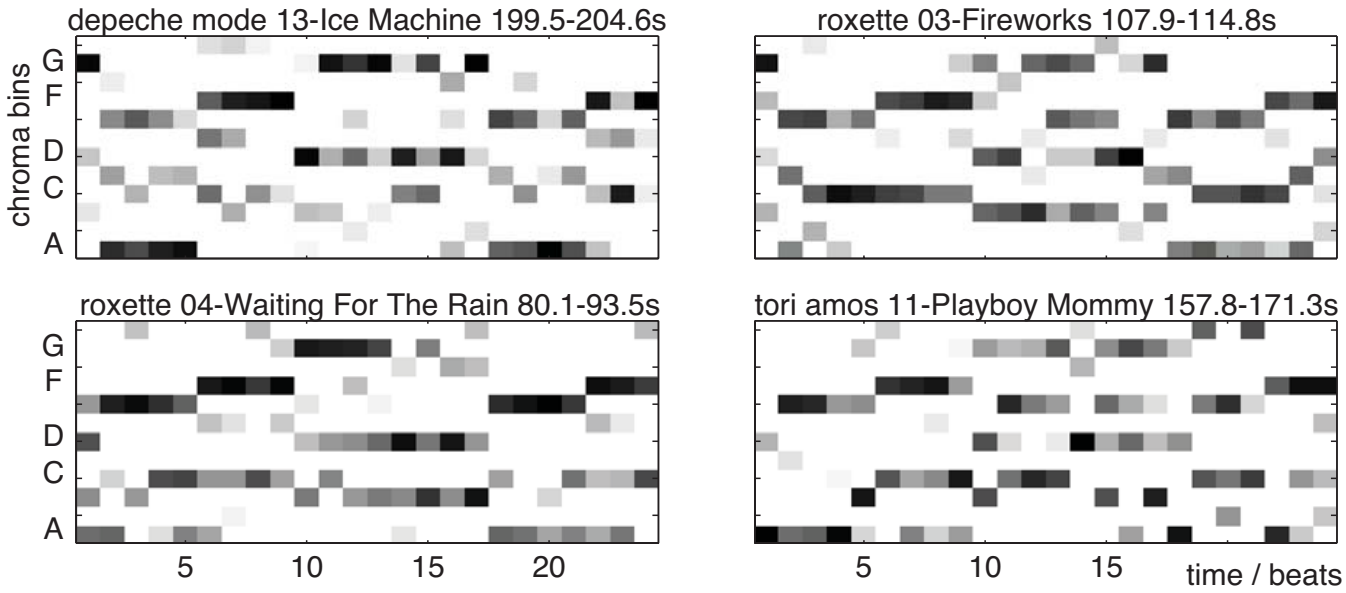
Results - Chroma Peaks

- Beat-chroma
  too diverse
  - reduce variation by keeping only 4 chroma/frame
- Landmarks
  off-by-1 →
  use $t_r - 2 \ldots t_r + 2$
  - 70,606 fragments (all beats would be 1.3M fragments)



Results - Detail

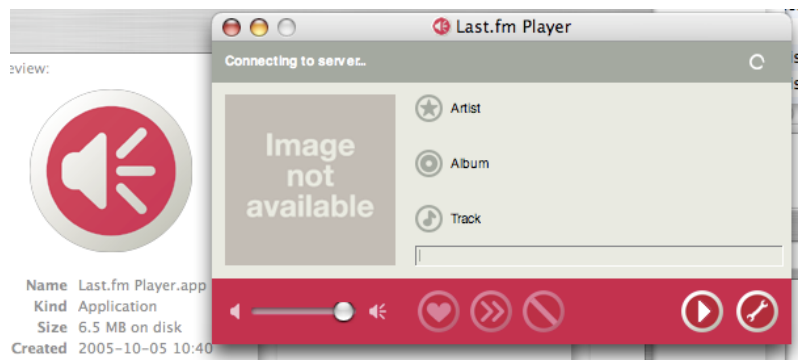
- Interesting fragment cluster...



- Not that interesting...
  - further simplification of fragments?
  - larger dataset?

4. Other Projects: Music Similarity

- The most central problem...
  - motivates extracting musical information
  - supports real applications (playlists, discovery)
- But do we need content-based similarity?
  - compete with collaborative filtering
  - compete with fingerprinting + metadata

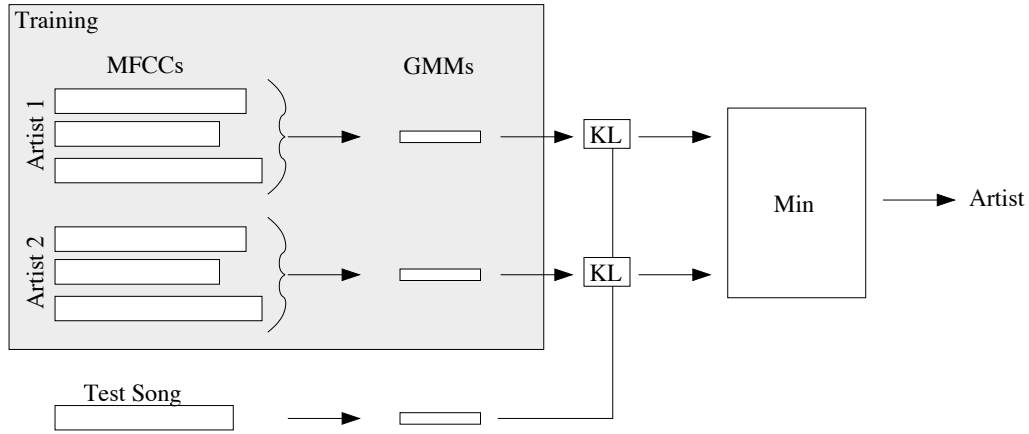


- Maybe ... for the Future of Music
  - connect listeners directly to musicians

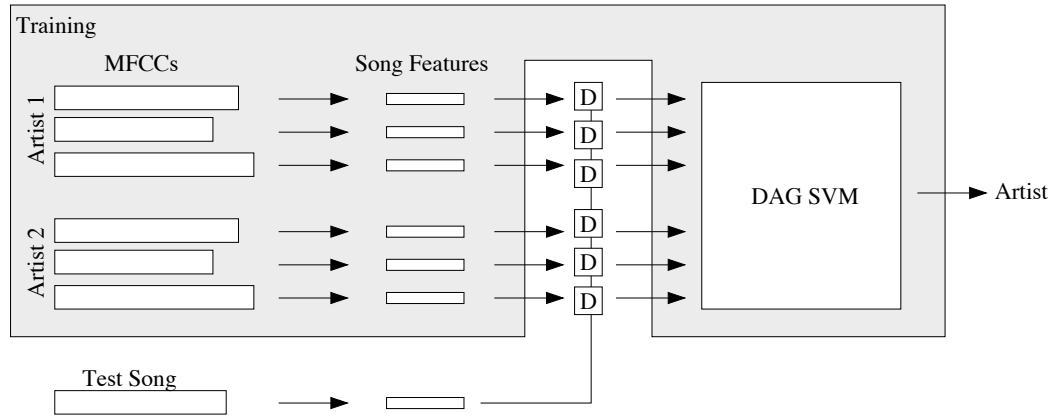
Discriminative Classification

Mandel & Ellis '05

- Classification as a proxy for similarity
- Distribution models...



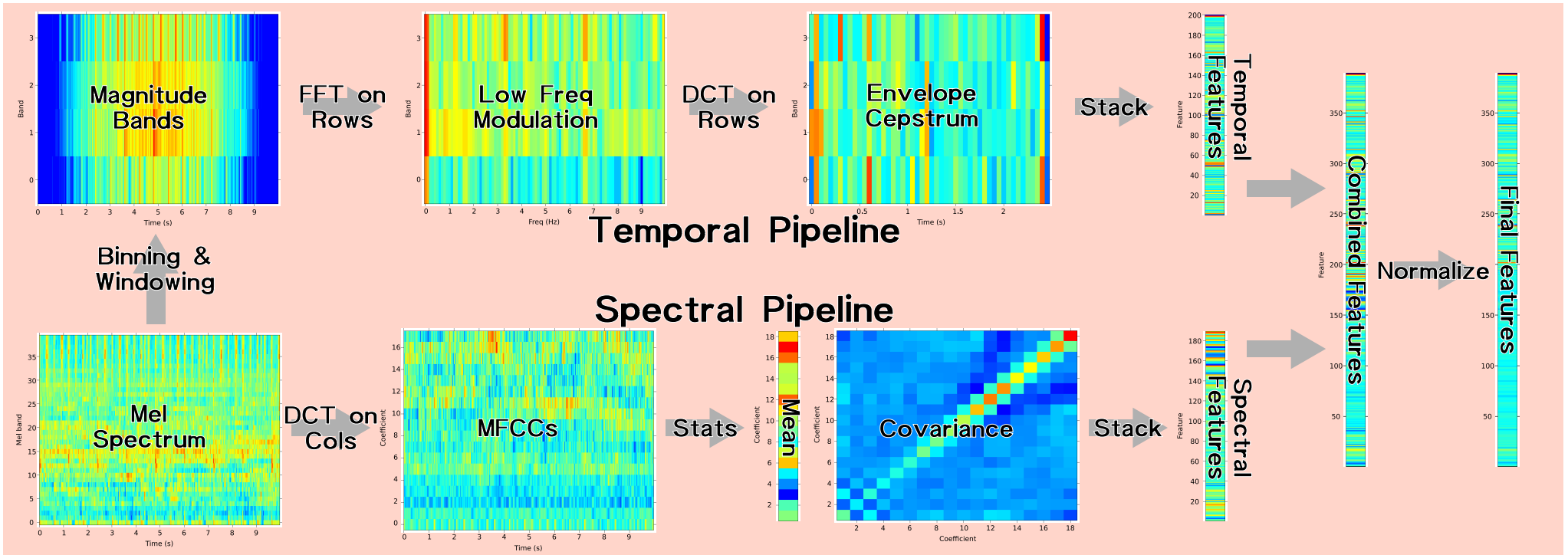
- vs. SVM



Segment-Level Features

Mandel & Ellis '07

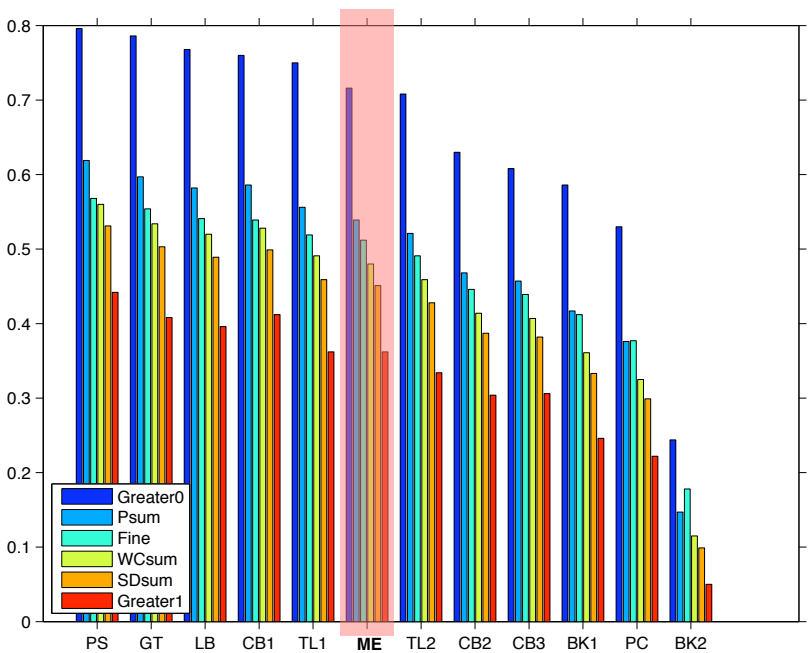
- Statistics of spectra and envelope define a point in feature space
  - for SVM classification, or Euclidean similarity...



MIREX'07 Results

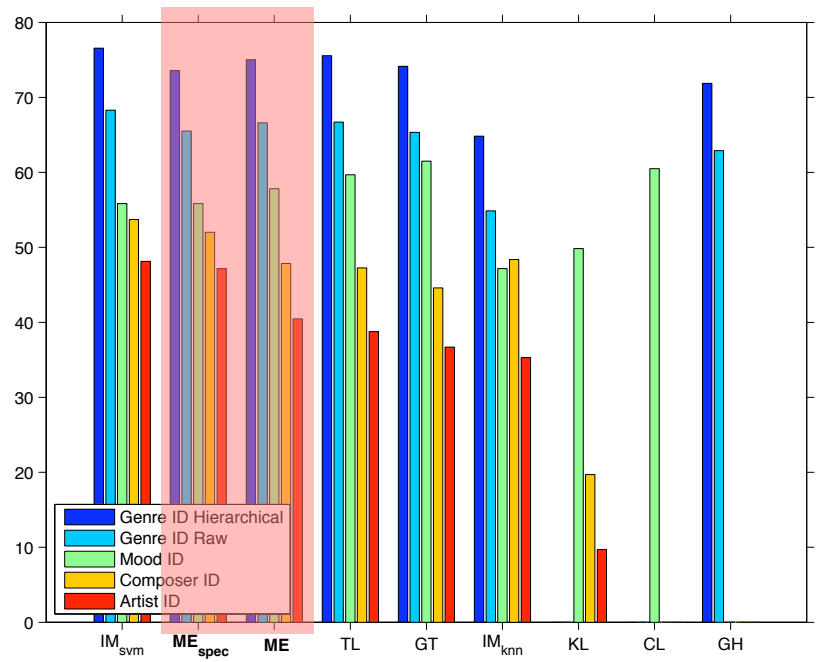
- One system for similarity and classification

Audio Music Similarity



PS = Pohle, Schnitzer; GT = George Tzanetakis; LB = Barrington, Turnbull, Torres, Lanckriet; CB = Christoph Bastuck; TL = Lidy, Rauber, Pertusa, Iñesta; ME = Mandel, Ellis; BK = Bosteels, Kerre; PC = Paradzinets, Chen

Audio Classification



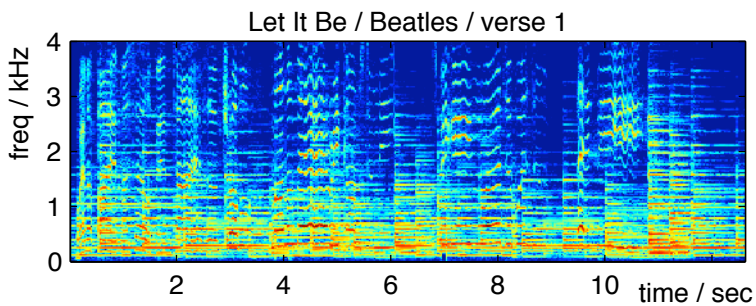
IM = IMIRSEL M2K; ME = Mandel, Ellis; TL = Lidy, Rauber, Pertusa, Iñesta; GT = George Tzanetakis; KL = Kyogu Lee; CL = Laurier, Herrera; GH = Guaus, Herrera

Cover Song Detection

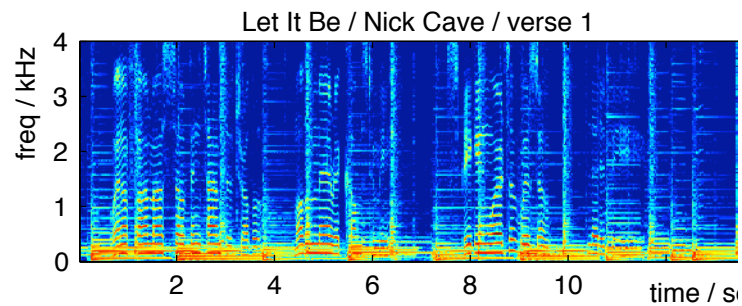
Ellis & Poliner '07

- "Cover Songs" = reinterpretation of a piece
  - different instrumentation, character
  - no match with "timbral" features

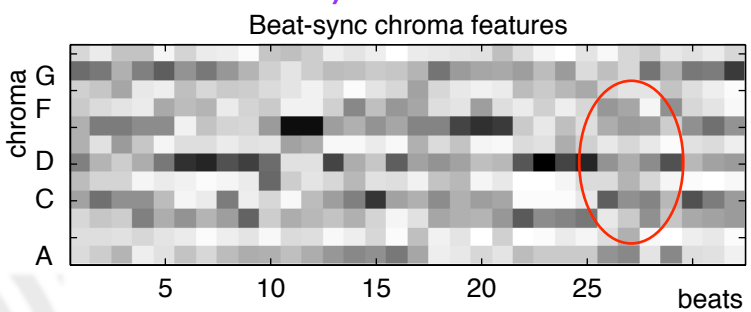
Let It Be - The Beatles

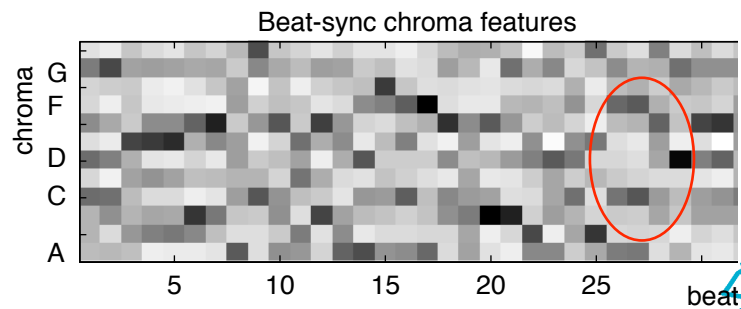


Let It Be - Nick Cave



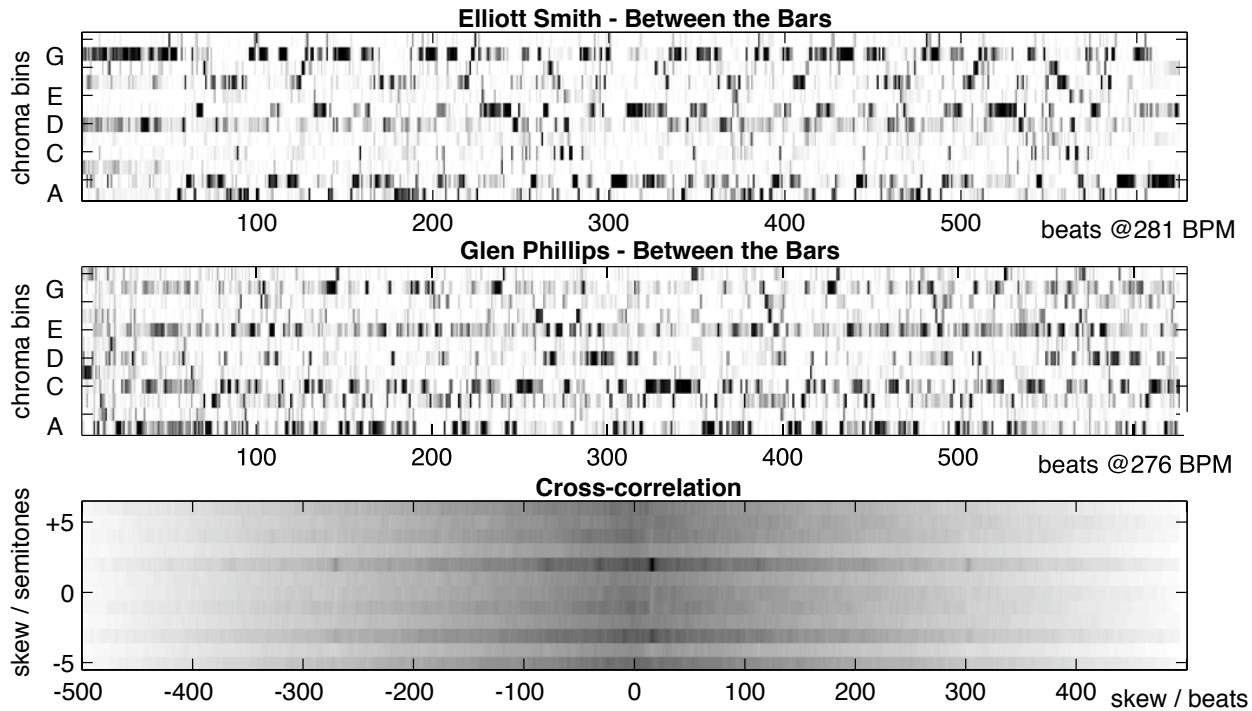
- Need a different representation!
  - beat-synchronous chroma features





Matching: Global Correlation

- Cross-correlate *entire* beat-chroma matrices
  - ... at all possible transpositions
  - implicit combination of match quality and duration



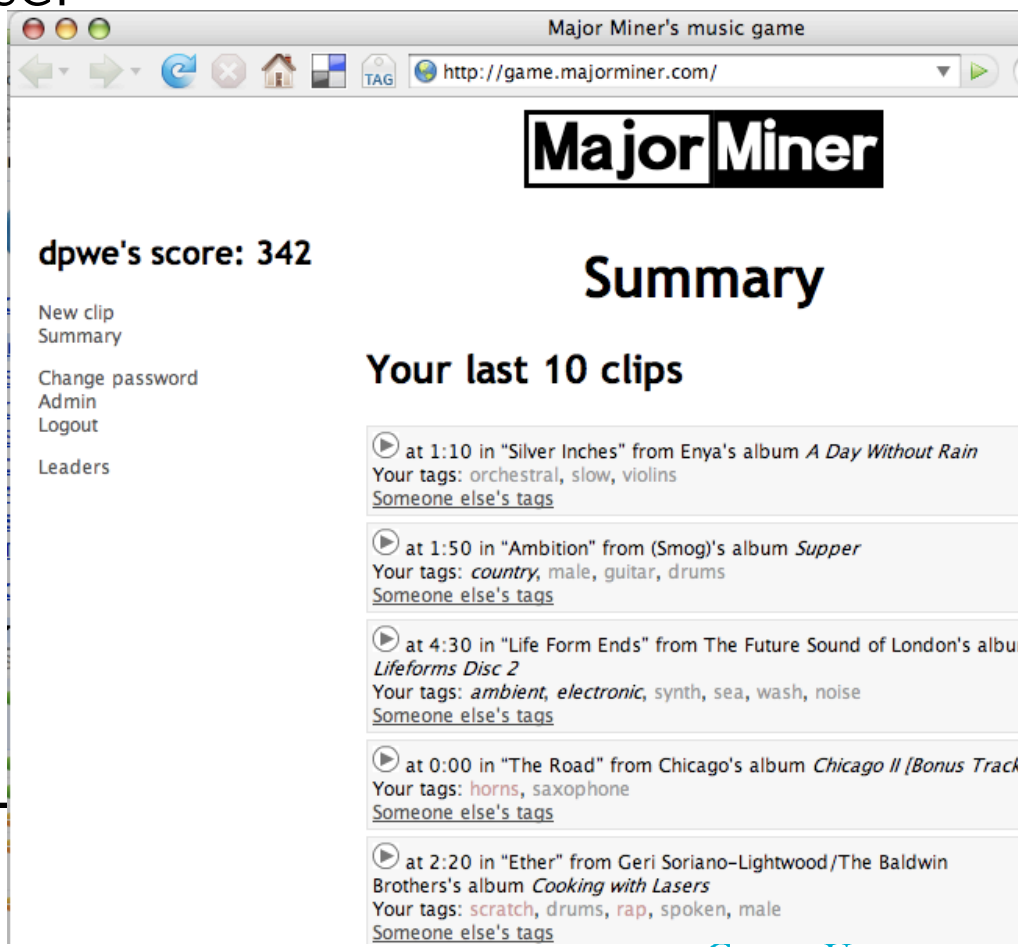
- One good matching fragment is sufficient...?

"Semantic Bases": MajorMiner

Mandel & Ellis '07,'08

- Describe segment in human-relevant terms
  - e.g. anchor space, but more so
- Need ground truth...
  - what words to people use?
- MajorMiner

  game:
  - 400 users
  - 7500 unique tags
  - 70,000 taggings
  - 2200 10-sec clips used
- Train classifiers...

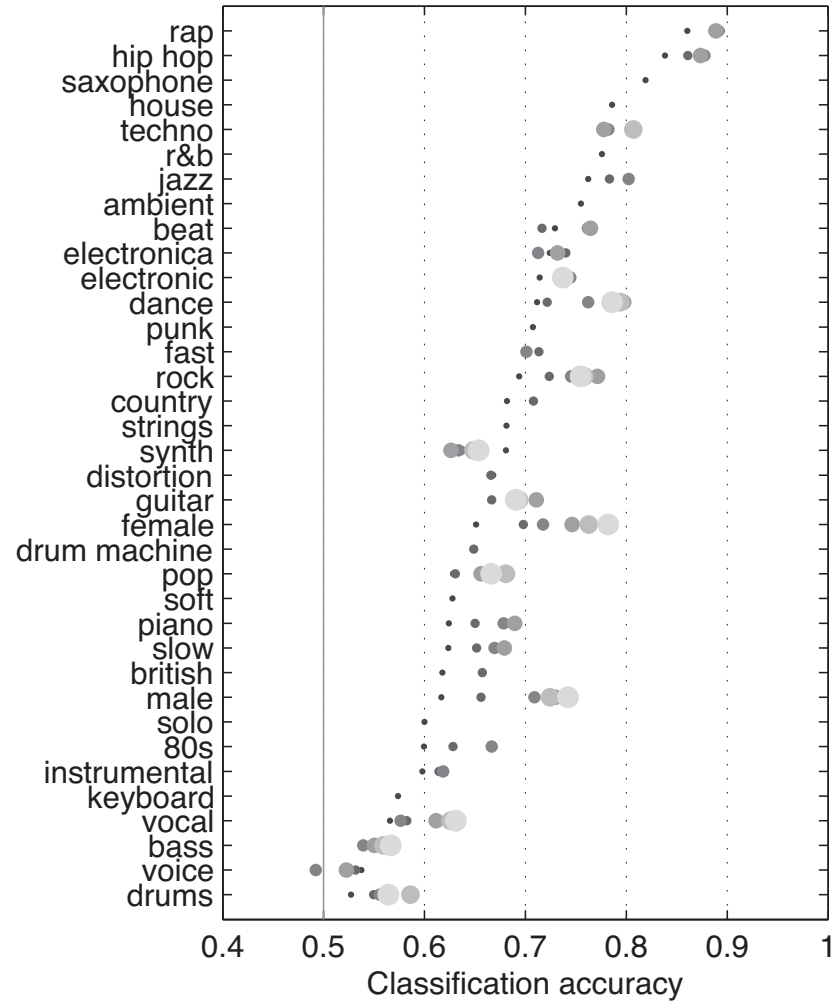


MajorMiner Autotagging Results

- Tags with enough verified clips → train SVM

- **Some good results**
  - test has 50% baseline; 7% better is significant
  - 50-300 training patterns

- **Next step: Propagate labels**
  - semi-supervised
  - "multi-instance" learning



Transcription as Classification

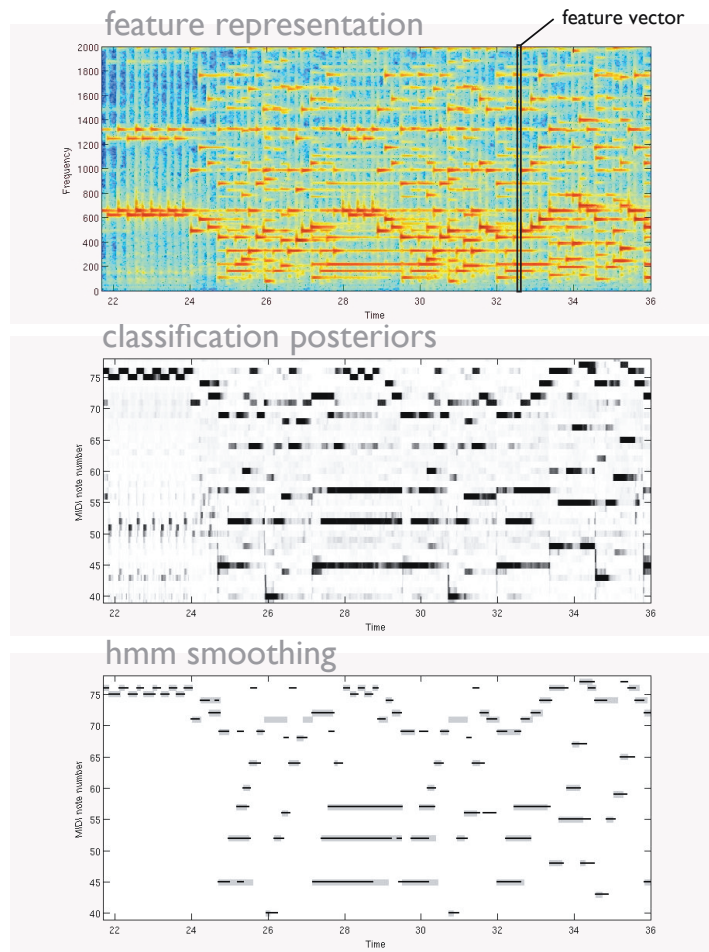
Poliner & Ellis '05,'06,'07

- Exchange signal models for data
  - transcription as pure classification problem:

Training data and features:
- MIDI, multi-track recordings, playback piano, & resampled audio (less than 28 mins of train audio).
- Normalized magnitude STFT.

Classification:
- N-binary SVMs (one for ea. note).
- Independent frame-level classification on 10 ms grid.
- Dist. to class bndy as posterior.

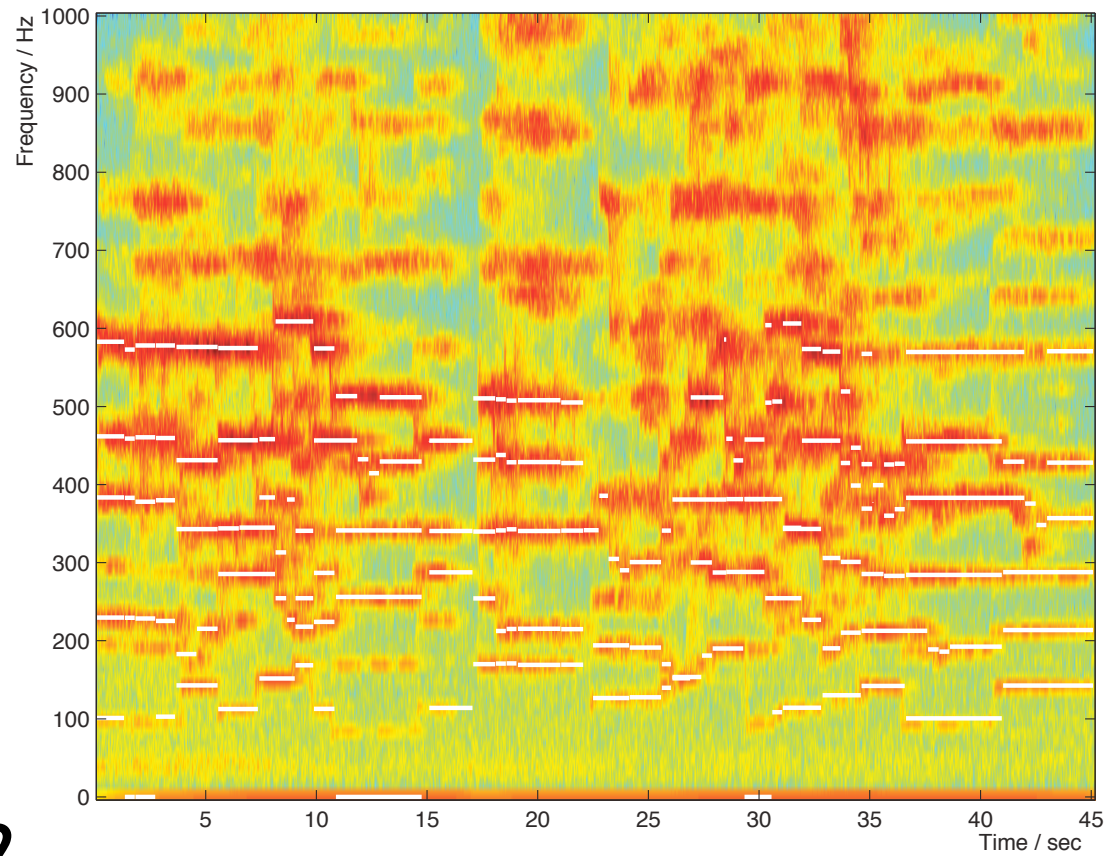
Temporal Smoothing:
- Two state (on/off) independent HMM for ea. note. Parameters learned from training data.
- Find Viterbi sequence for ea. note.



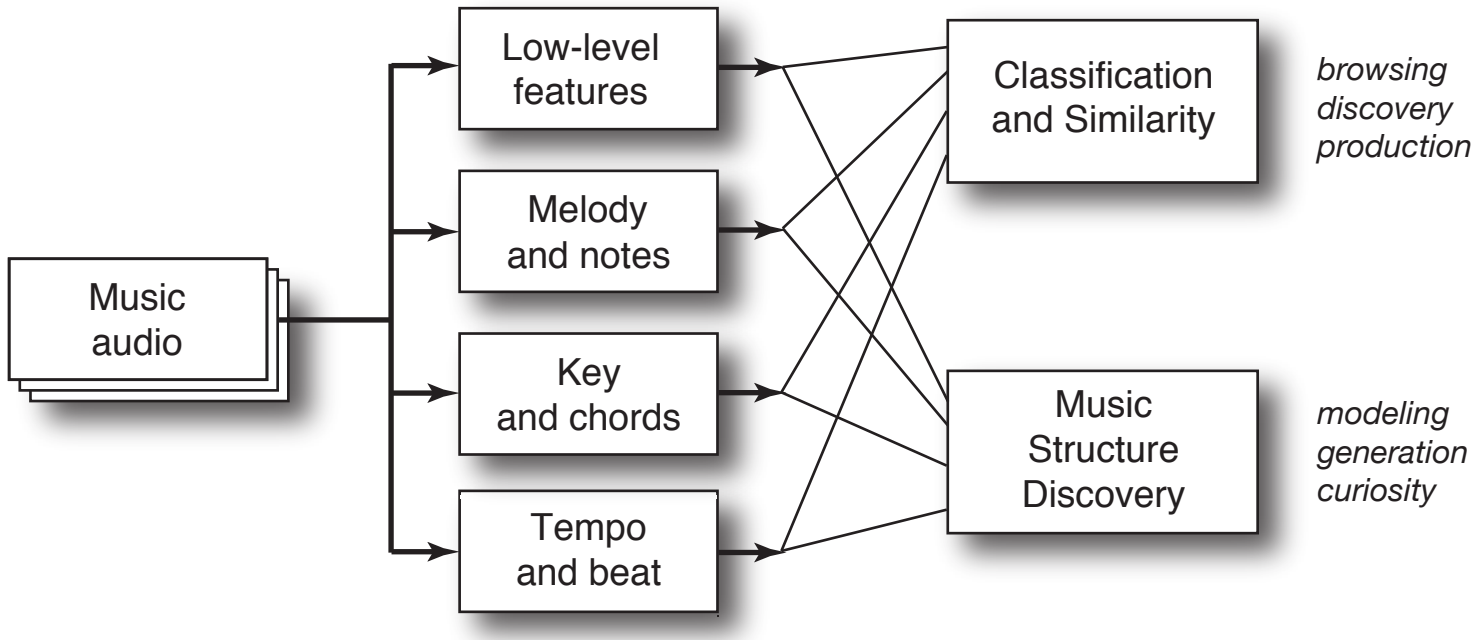
Singing Voice Modeling & Alignment

Christine Smit
Johanna Devaney

- How do singers sing?
  - e.g. "vowel modification" in classical voice
  - tuning variation...
- Collect the data
  - .. by aligning libretto to recordings
  - e.g. align Karaoke MIDI files to original recordings
  - detail at alignments
- Lyric Transcription?



Conclusions



- Lots of data
  + noisy transcription
  + weak clustering
  ⇒ musical insights?