CS598PS – Machine Learning for Signal Processing

Machine Listening and Music Information Retrieval

16 November 2020

Today's lecture

- Like the title said …
- Music Information Retrieval
  - Making computers understand music
- Machine Listening
  - Making machines that can hear

Music Information Retrieval

- Making sense out of music data
  - Since we do signals we'll do music audio data
- Genre/Artist identification
- Song fingerprinting
  - Large scale correlations

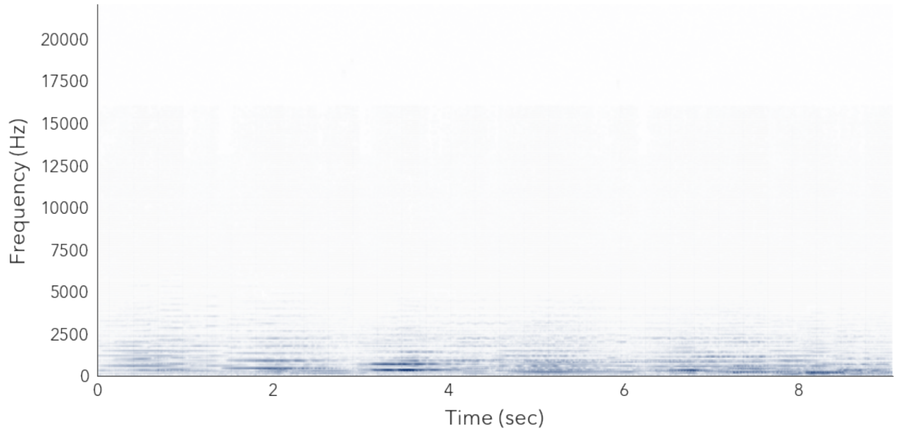
Musical genre identification

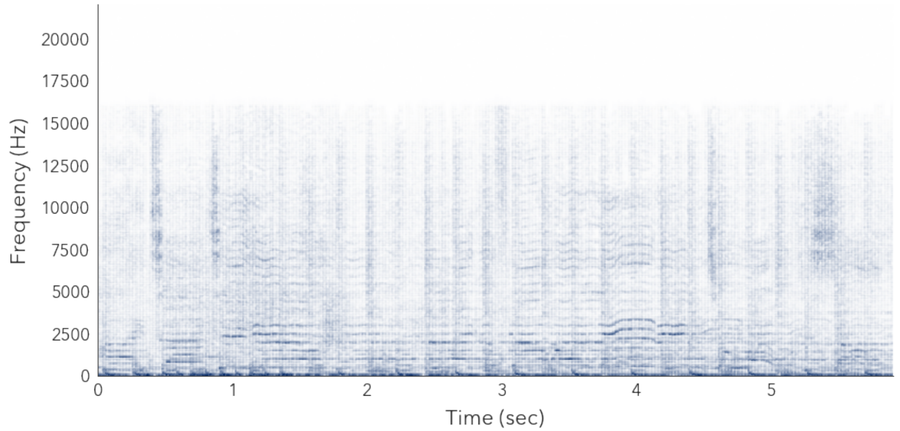
- Identify the style of a music recording
  - Jazz, rock, rap, classical, etc ...

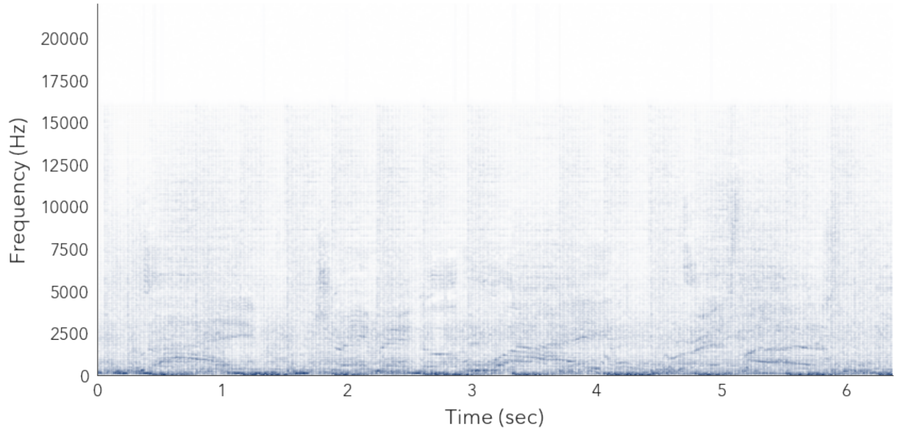
- You should know how to do that!
  - Get training data, find decent features, make a classifier, ...

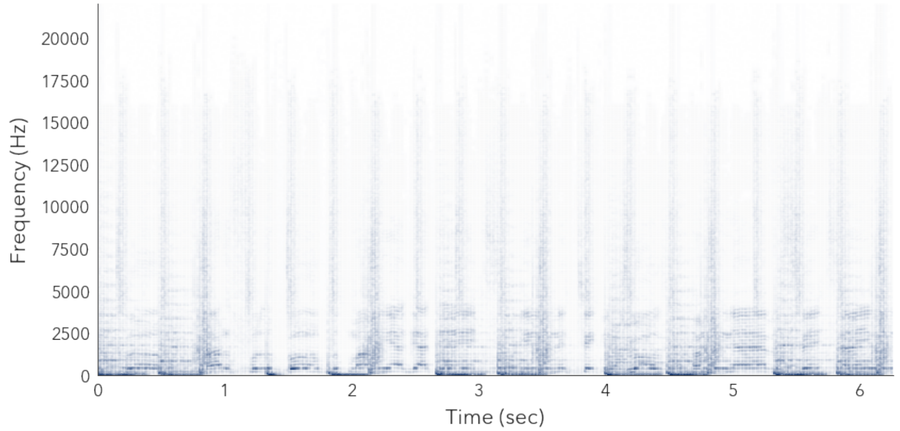
Basic premise

- Different genres sound different
  - i.e. their audio features are different classes









Features for music

- Spectrograms are a little too detailed

- We sometimes need to derive coarser features
  - Overall sound and rhythm quantities

Timbral features

- Timbre: “Quality of sound”
  - What is not pitch or amplitude
- A function of the spectral shape
  - “hissy”, “muted”, “nasal”, “boomy”, …

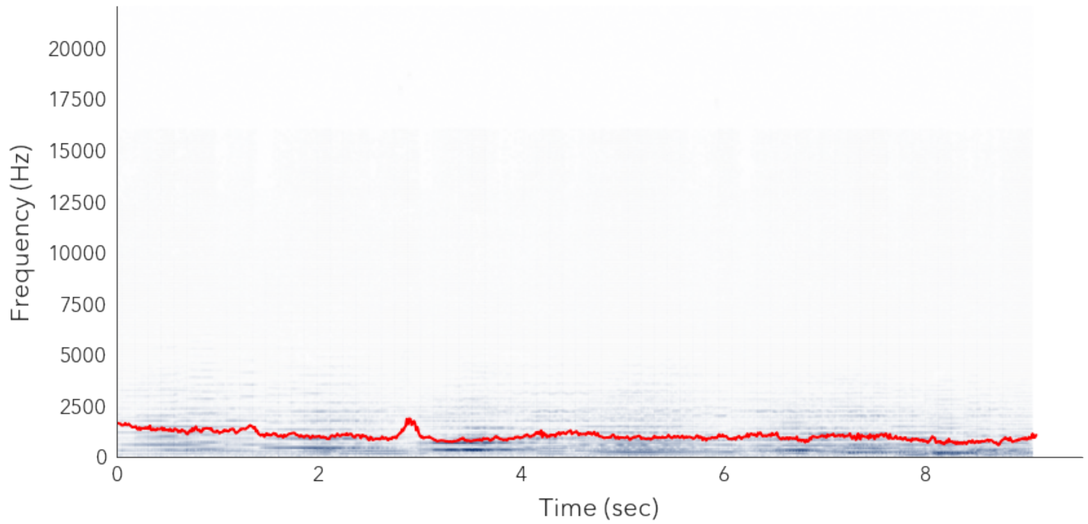
Spectral centroid

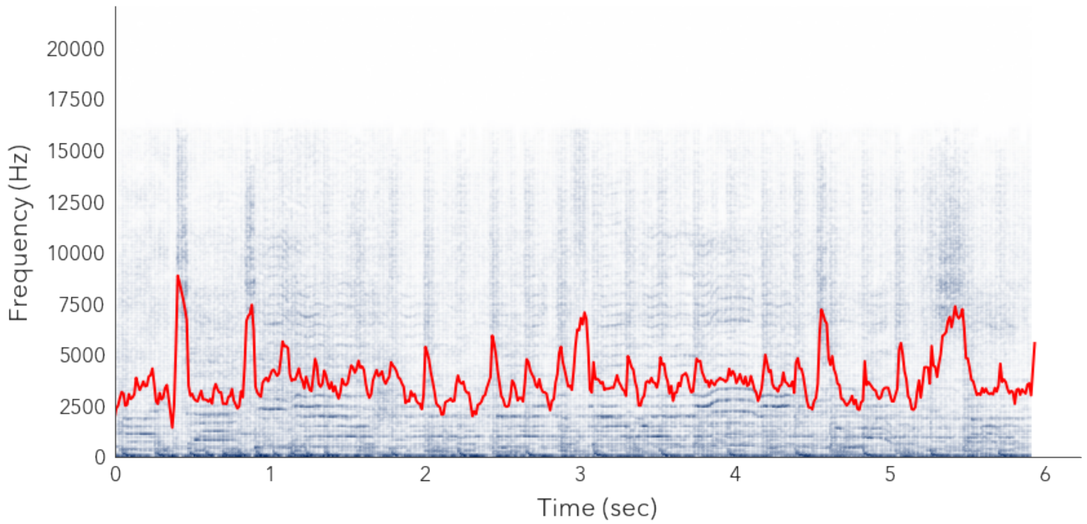
- "Center of mass" of the spectrum

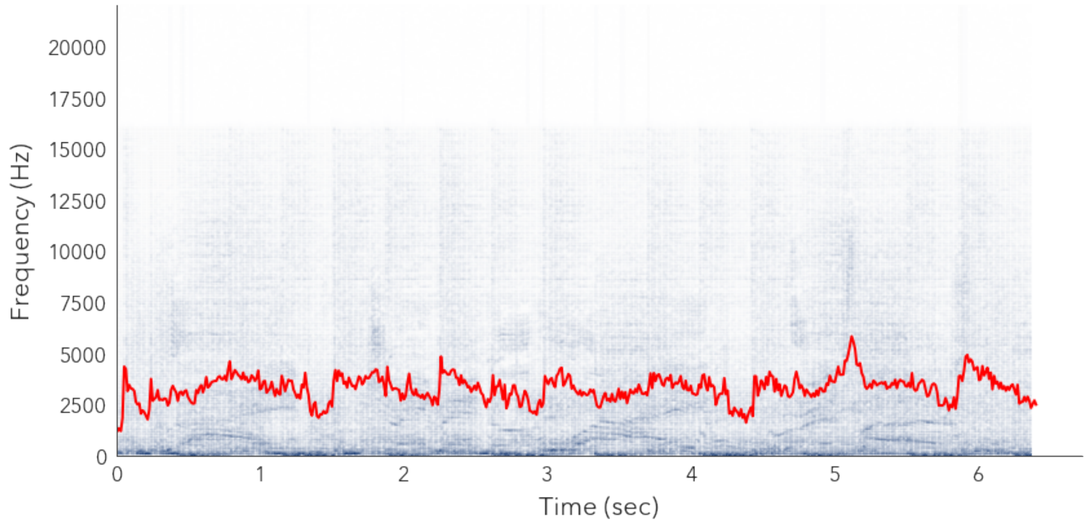
$$C_t = \frac{\sum \left\| F_t(f) \right\| f}{\sum \left\| F_t(f) \right\|}$$

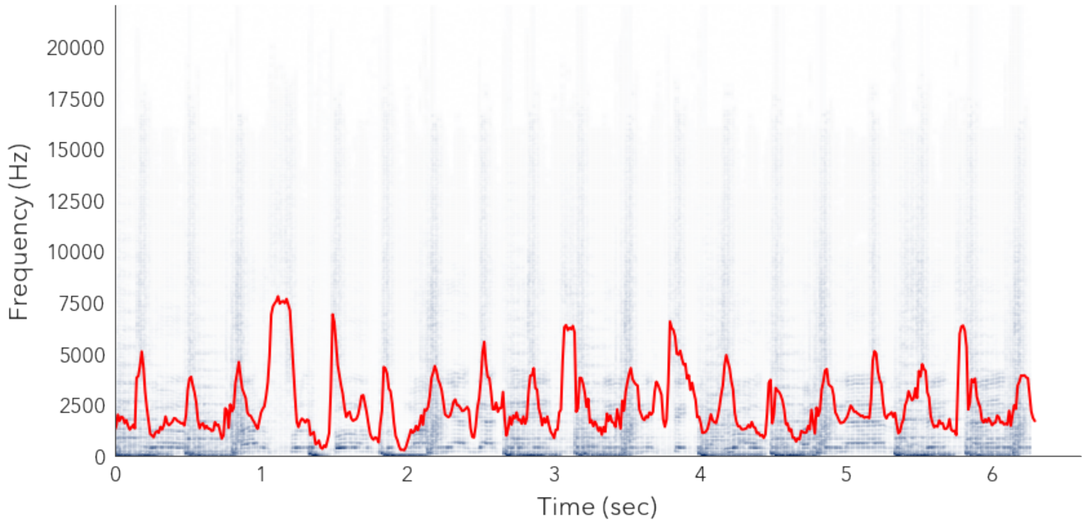
- Correlates to perceived "brightness"

Spectral centroid









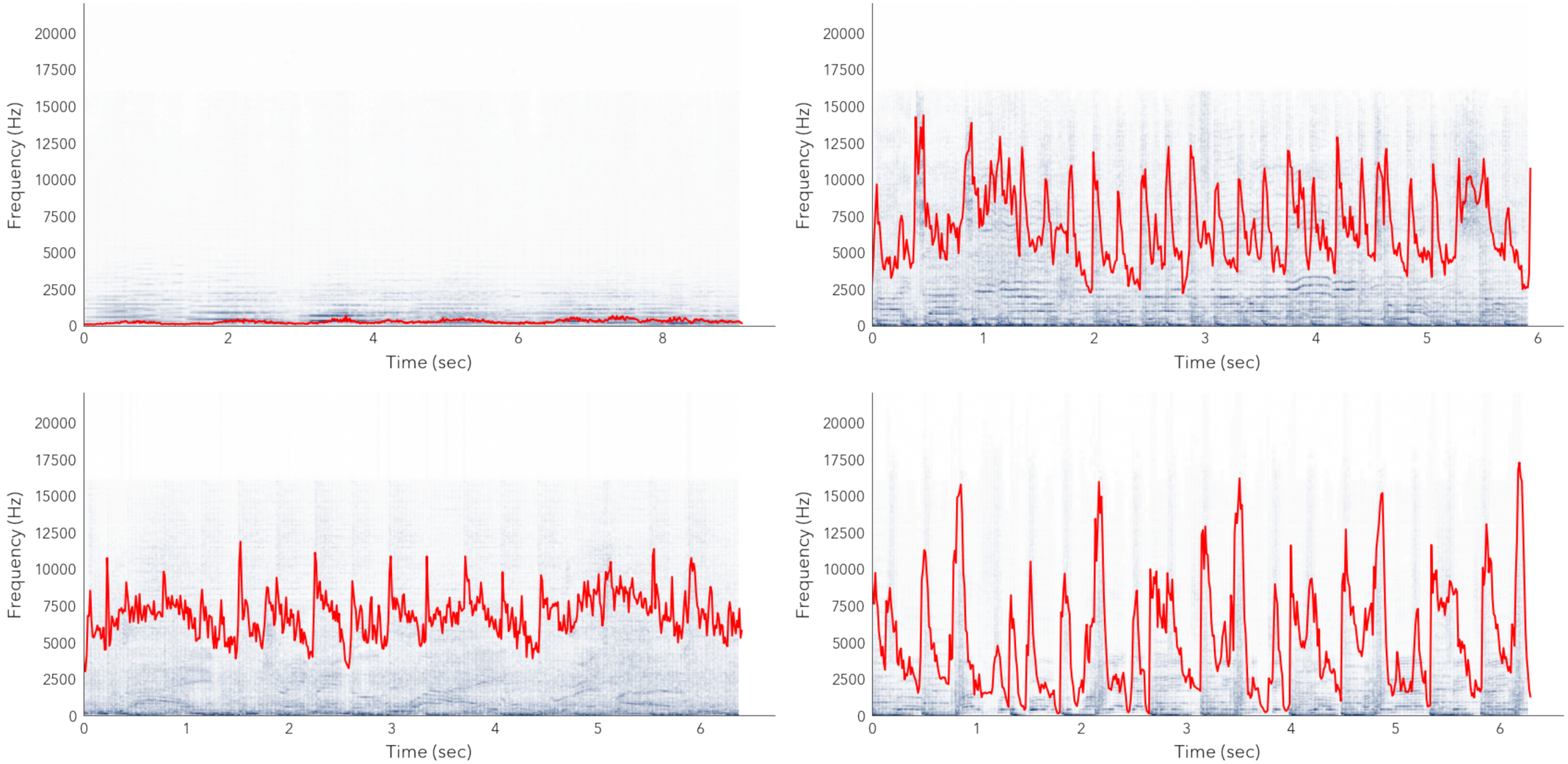
Spectral flux

- Difference between successive spectra

$$L_t = \sum \left(F_t(f) - F_{t-1}(f) \right)^2$$

- Provides a sense of local timbre change
  - Is the music static or dynamic?

Spectral flux



Many more …

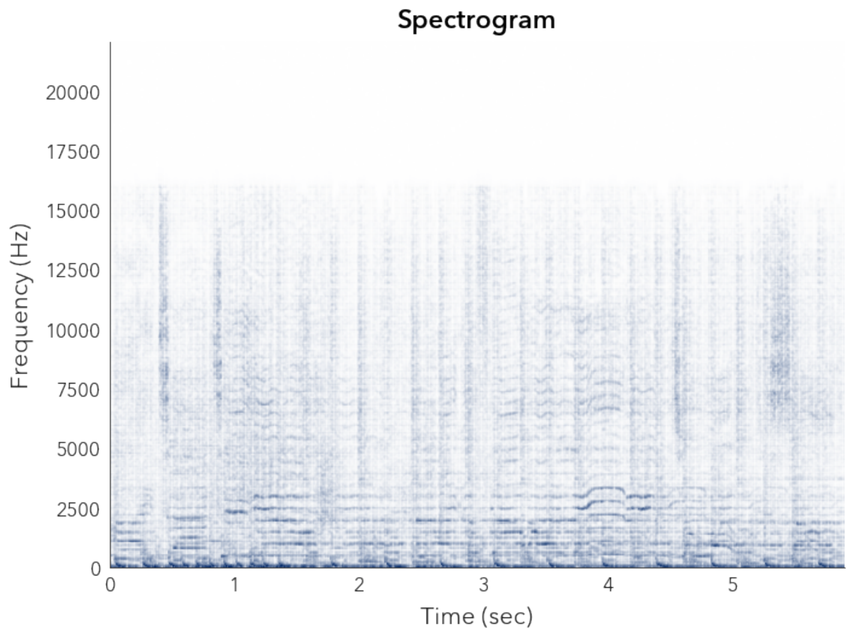
- Spectral slope
  - Fit a line to the spectrum
- Spectral statistics
  - Spread, skew, kurtosis
- Zero crossings per time window
  - Provides a sense of noisiness

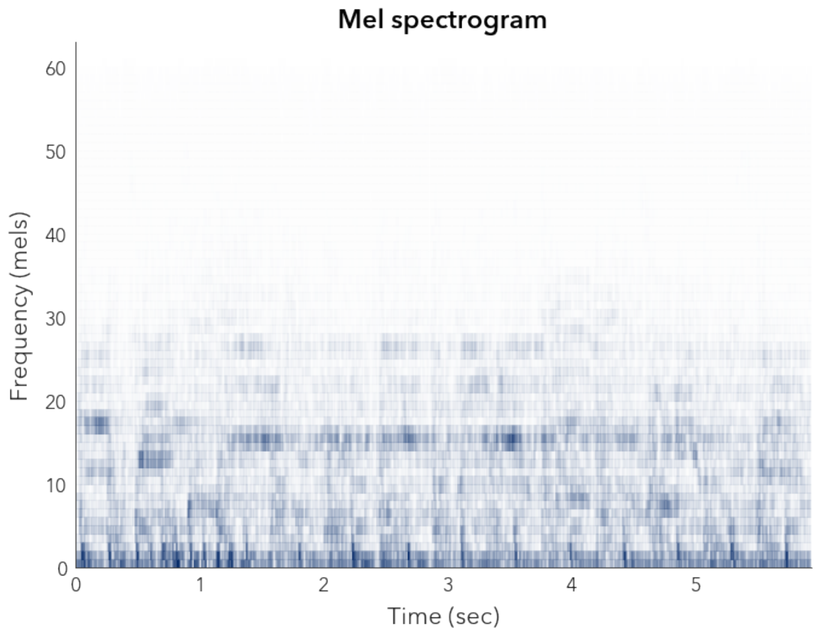
Mel Frequency Cepstral Coeffs (MFCC)

- A coarse version of the spectrogram
  - Blurring frequencies, removing redundancy
- Standard tool for speech, very popular for non-speech sounds

Frequency warping/rescaling

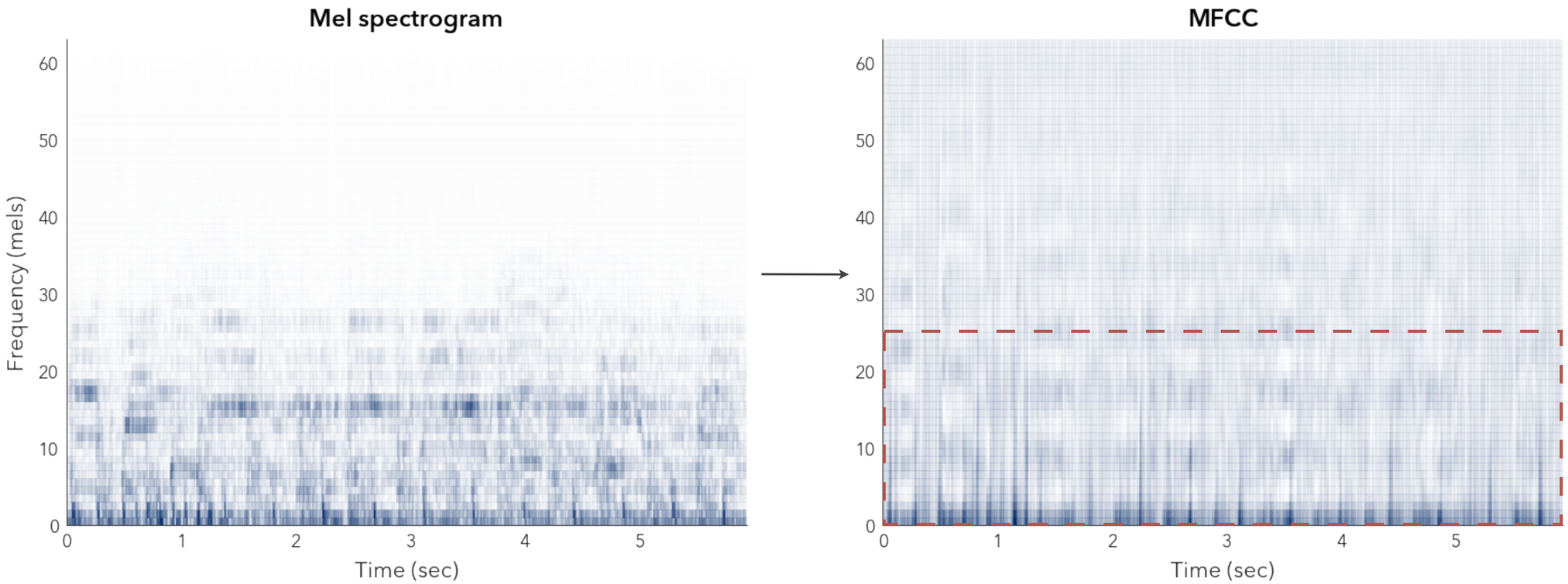
- Spectrograms are too detailed
  - We first blur the frequencies
    - Use Mel warping and reduce their number
  - Also take their log to improve contrast





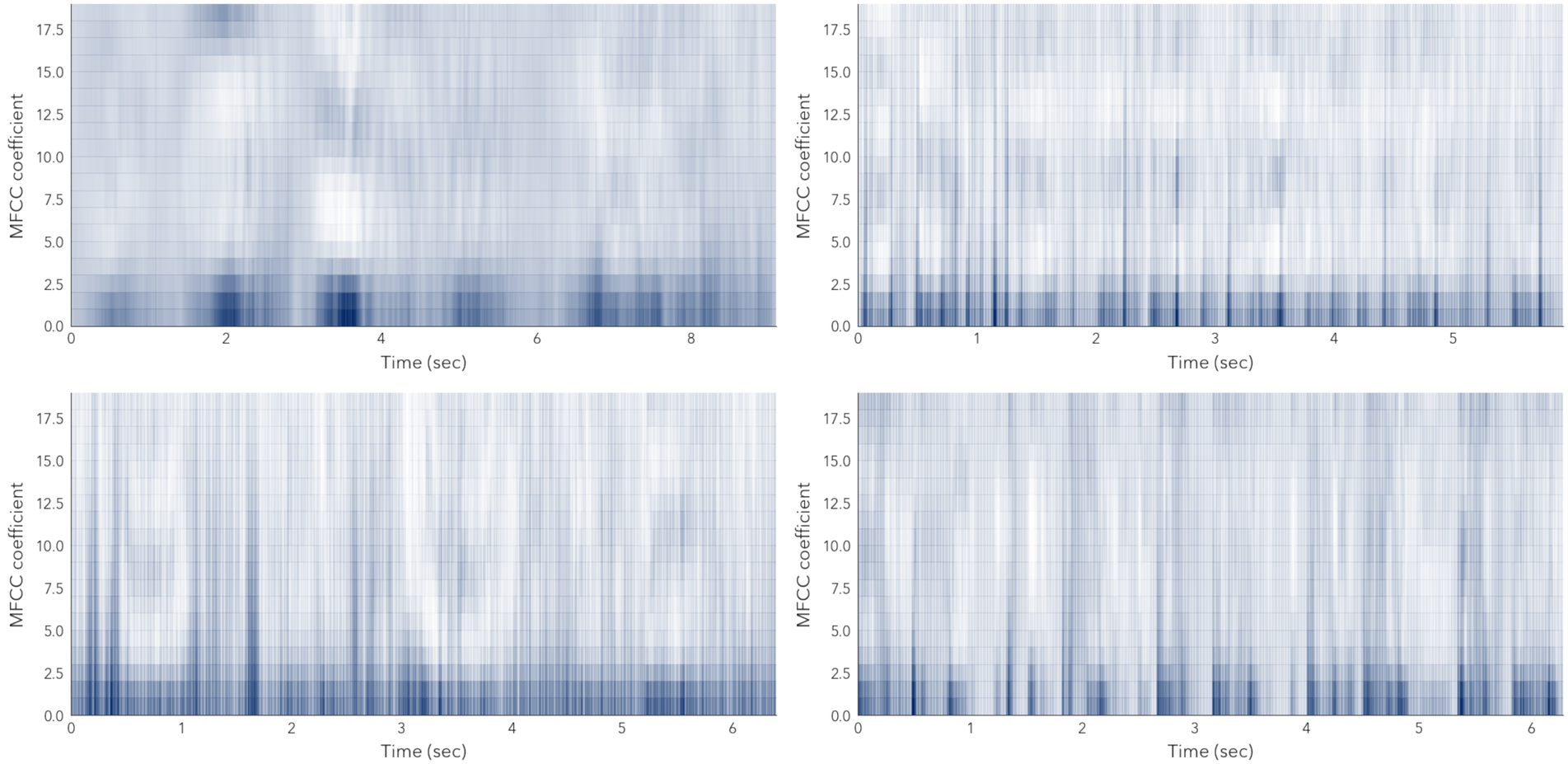
Dimensionality reduction

- Do a DCT (Discrete Cosine Transform)
  - Keep lowest frequencies



Resulting transform

- Not as intuitive, but informative



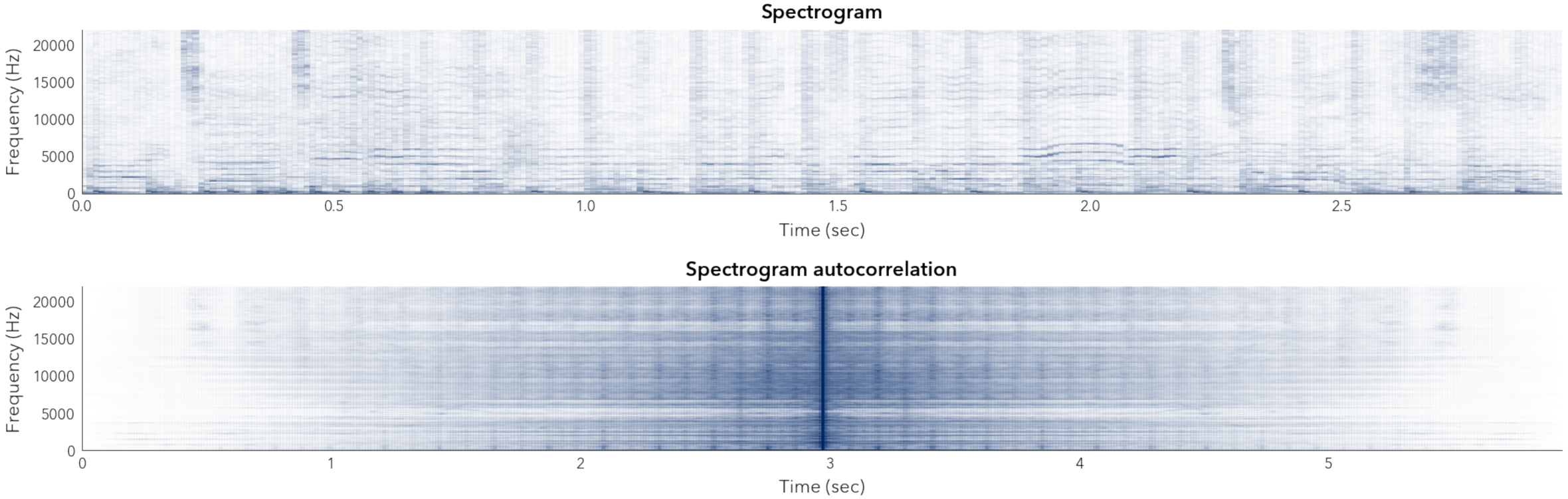
The statistical interpretation

- Why the log magnitudes?
  - Magnitude spectra are exponentially distributed
  - Their log makes them more Gaussian distributed
    - We can then use Gaussian models!

- Why the DCT?
  - The DCT acts like a PCA decomposition
    - But it is easier/cheaper to implement as opposed to PCA

Rhythm features

- Analysis of temporal correlations
  - Autocorrelate varying frequency areas
  - Find “frequencies of frequencies”



Putting it all together

- Append all/some of the above features
  - large vector [MFCC , centroid , flux , ...]
- Do standard classification on it
  - Dimensionality reduction: PCA, MDA, ...
  - Classifiers: GMMs, SVMs, Neural Nets, ...
- Sort of works ...
  - 60%-70% accuracy in semi-realistic sets

Reuse of this framework

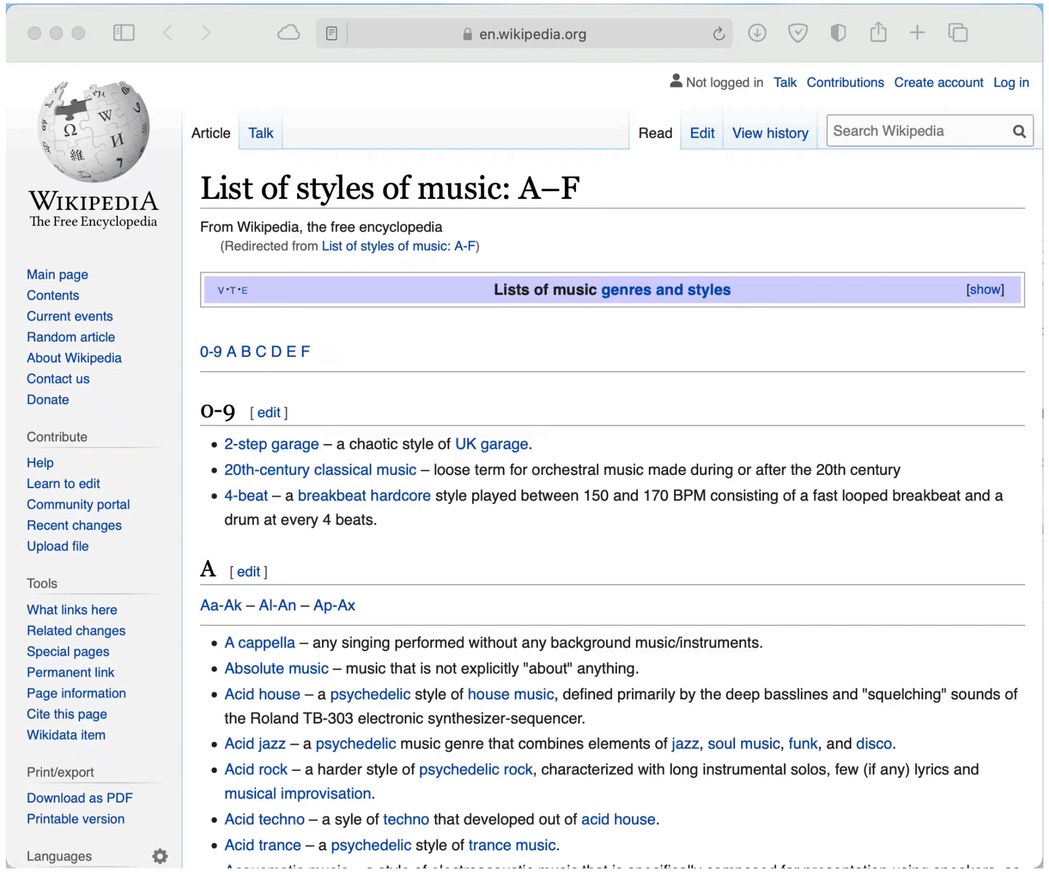
- Artist recognition
  - Is it Madonna? Is it Metallica?
- Music similarity
  - Measure similarity between two recordings
- Music clustering
  - Make sense of your huge music collection

Problems

- What are some of the problems here?

Music classification issues

- Too many music classes
  - http://en.wikipedia.org/wiki/List_of_styles_of_music:_A-F



- Audio by itself doesn't convey all the info we want

What style is this?

Is this the same artist?

Looking beyond the signal

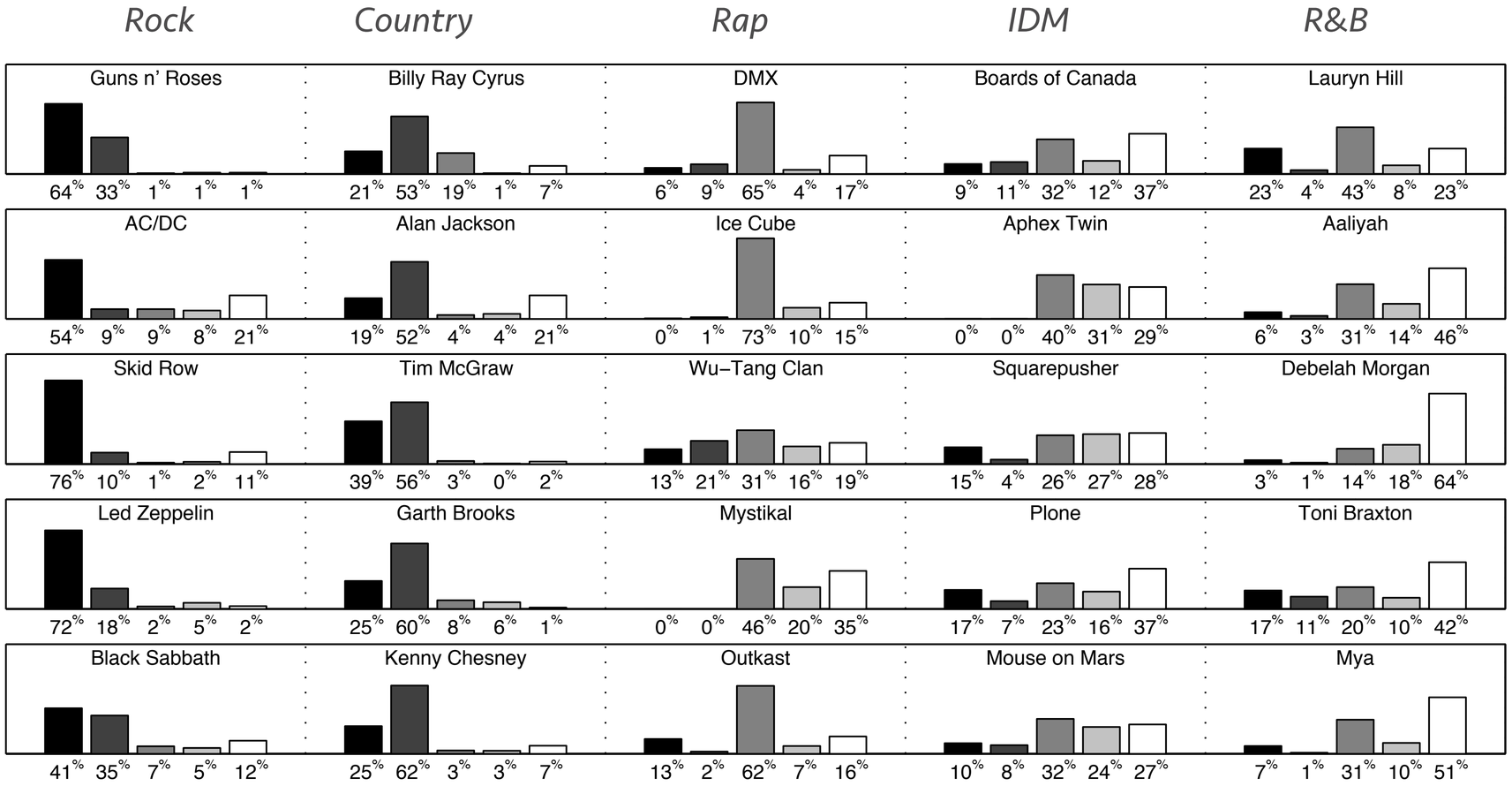
- Music is all about context
  - marketing, posing, image, sex appeal, etc …
- Audio doesn’t capture context
  - But text does!

Supplementing with text

- Text about an artist provides context
  - Can we harness that information?
- Combining acoustic and text classifiers
  - Signal + semantics approach

Example case

- Audio-only classification can fail

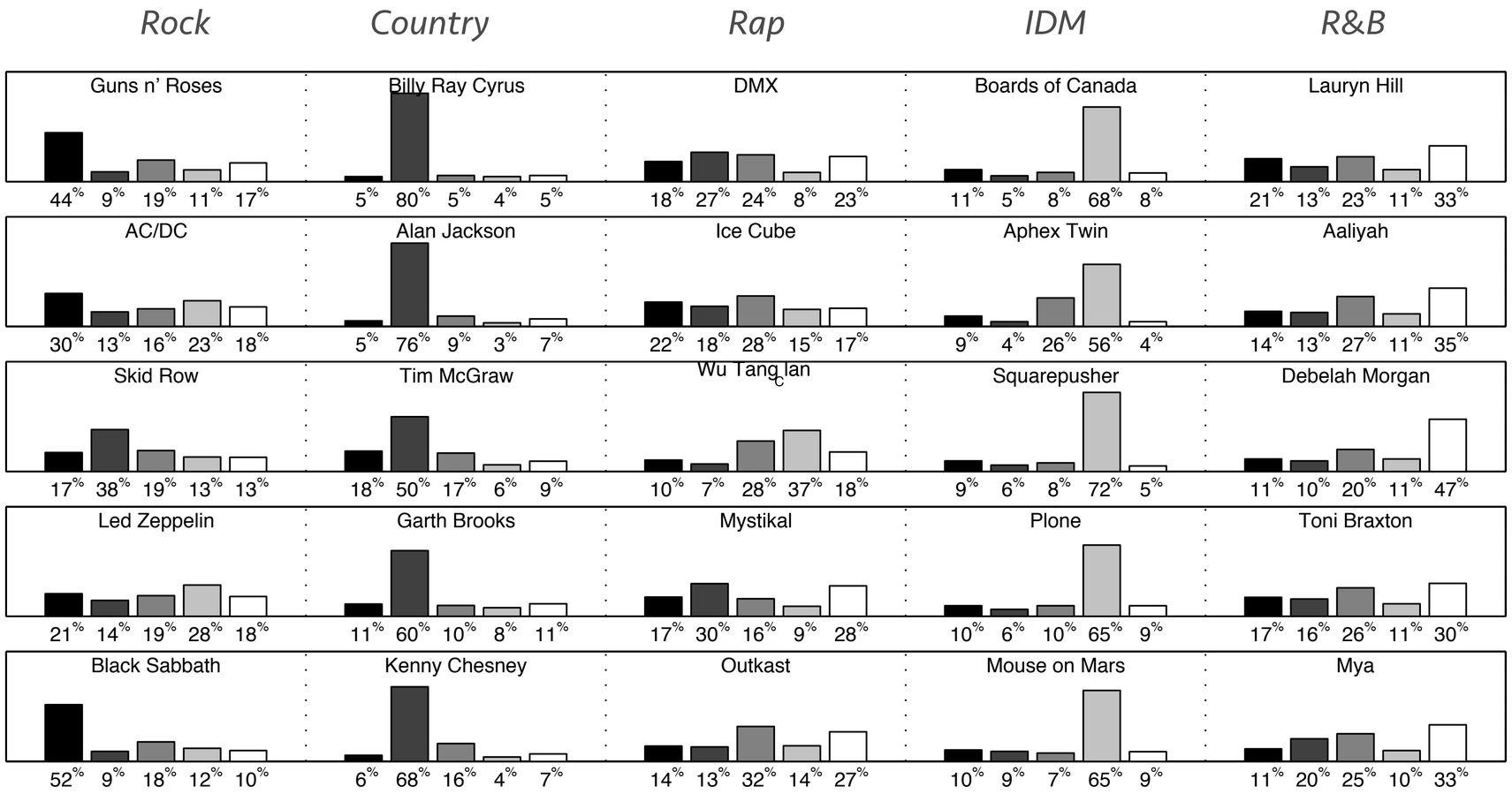


Obtaining some text info

- Scour the web to get text about these data
  - Lyrics, descriptions, social site comments, etc.
- Try to do the classification using text

Text-based classification

- Also fails!

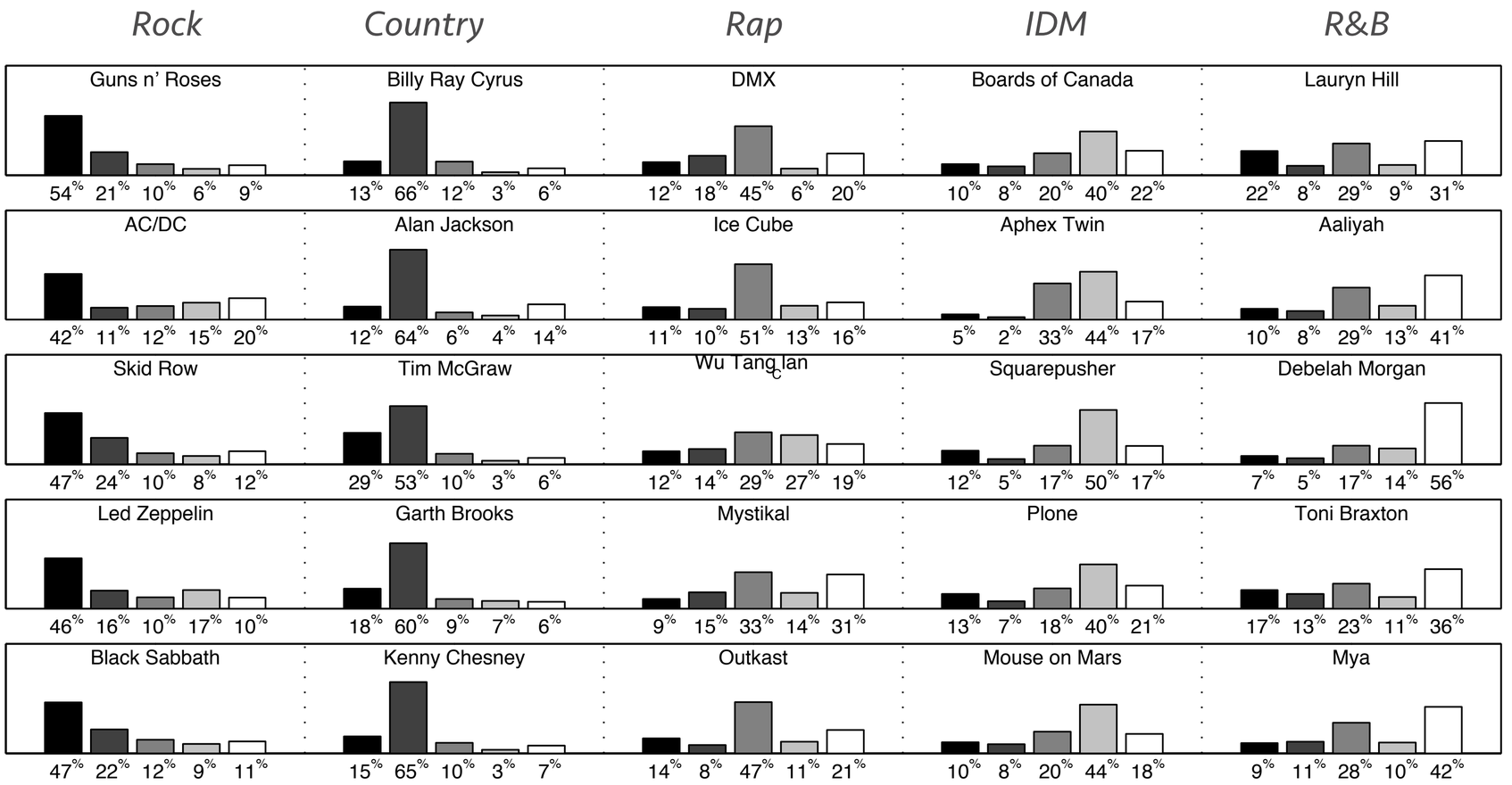


Combining the two modalities

- The failure points are not overlapping
- We can combine the two classifiers
  - Hopefully will cancel each other's faults

Combined system

- Higher accuracy, no contextual mistakes



Audio fingerprinting

- So far we did an approximate match
  - Similar style/artist
- Fingerprinting is about finding *exact* matches
- Is this song "x" or not?
  - Many commercial tools available

How to match a recording

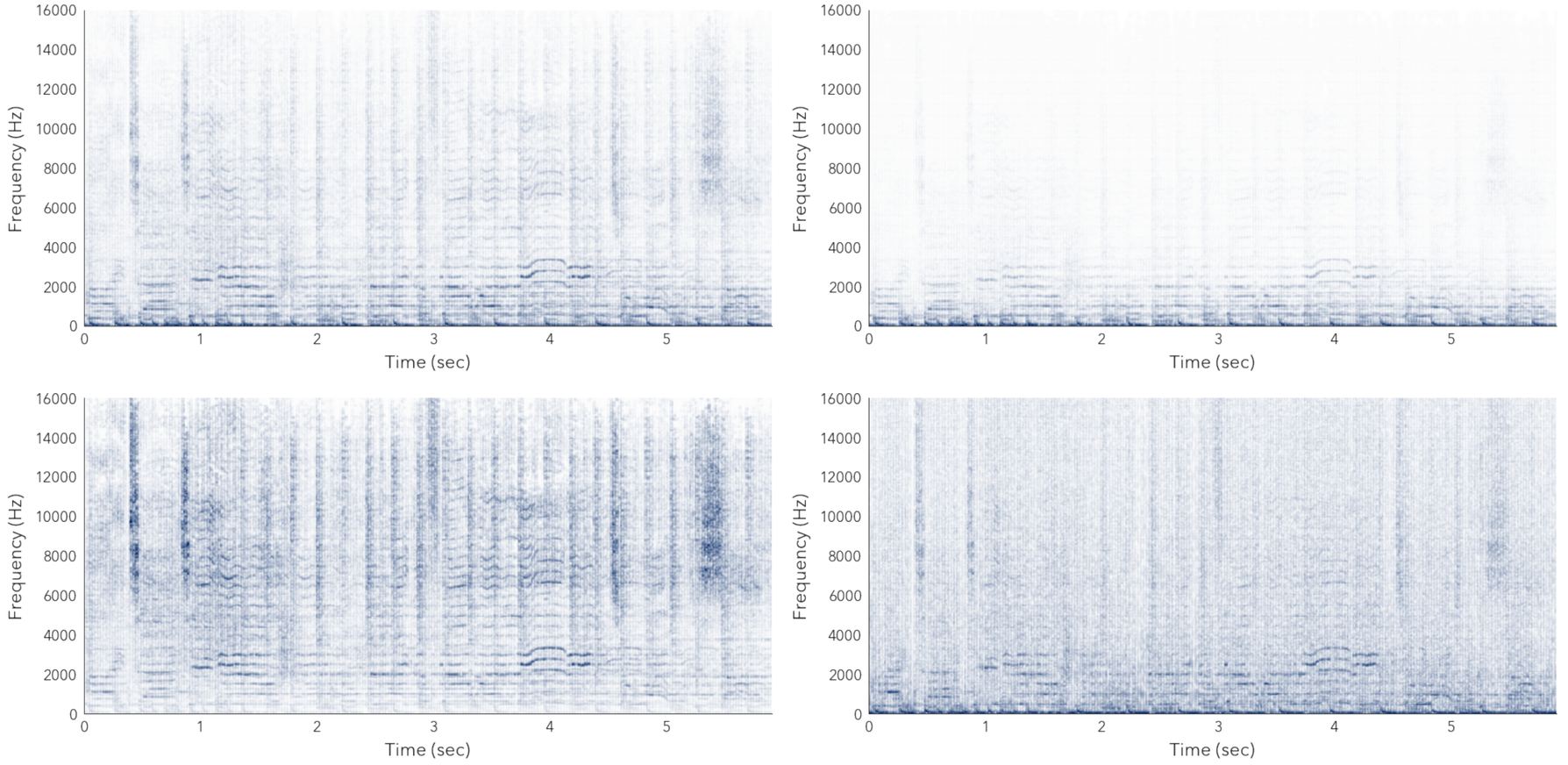
- Obvious approach
  - The matched filter!

- Problems
  - What if the query is noisy?
  - Can you efficiently correlate with your entire database?

The need for robust features

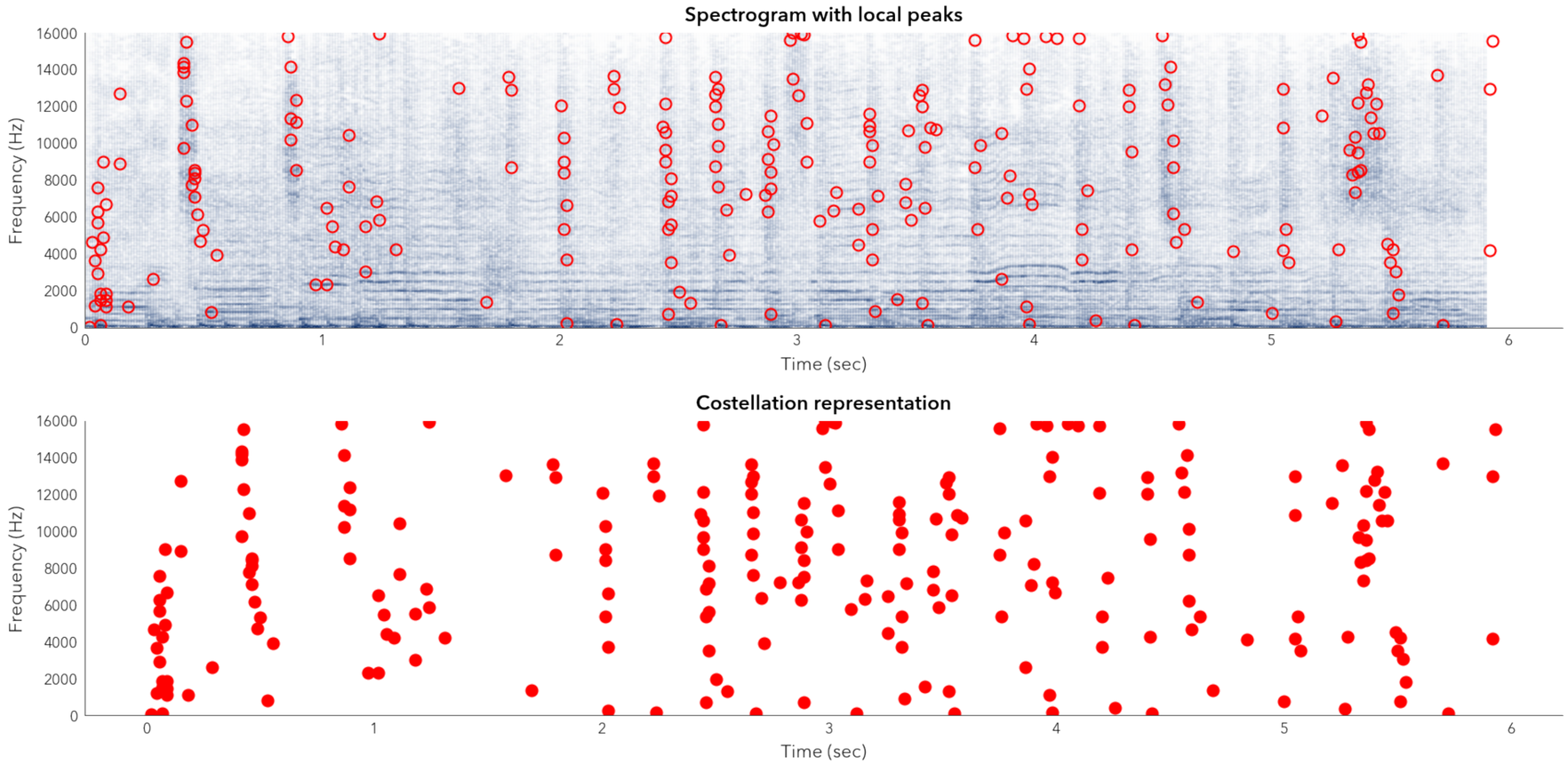
- Query can be contaminated
  - We must find features that look past that!

Same recording, different conditions



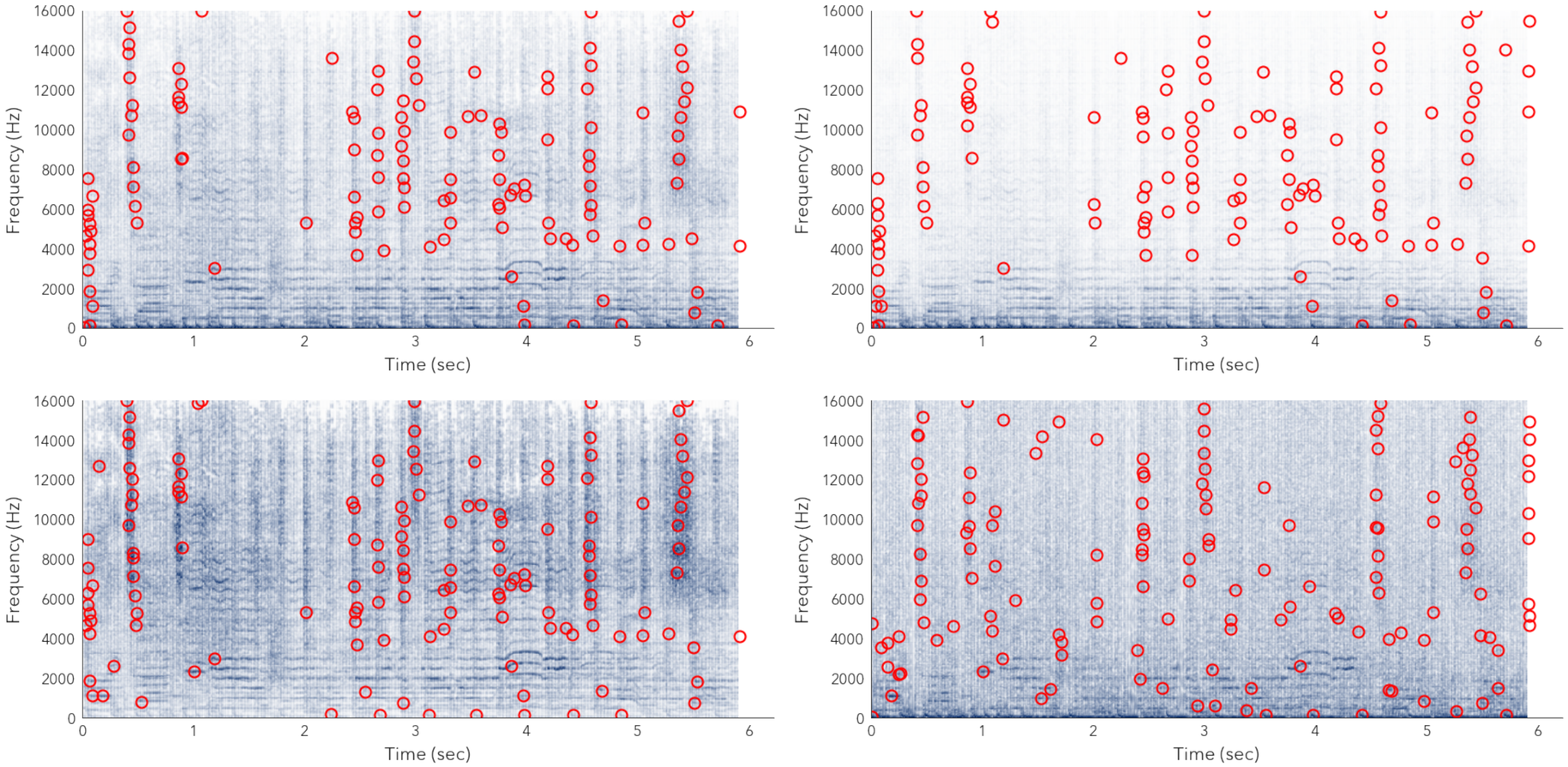
Peak constellations

- Find local peaks and keep their positions
  - Spectrogram is reduced to prominent time-frequency locations



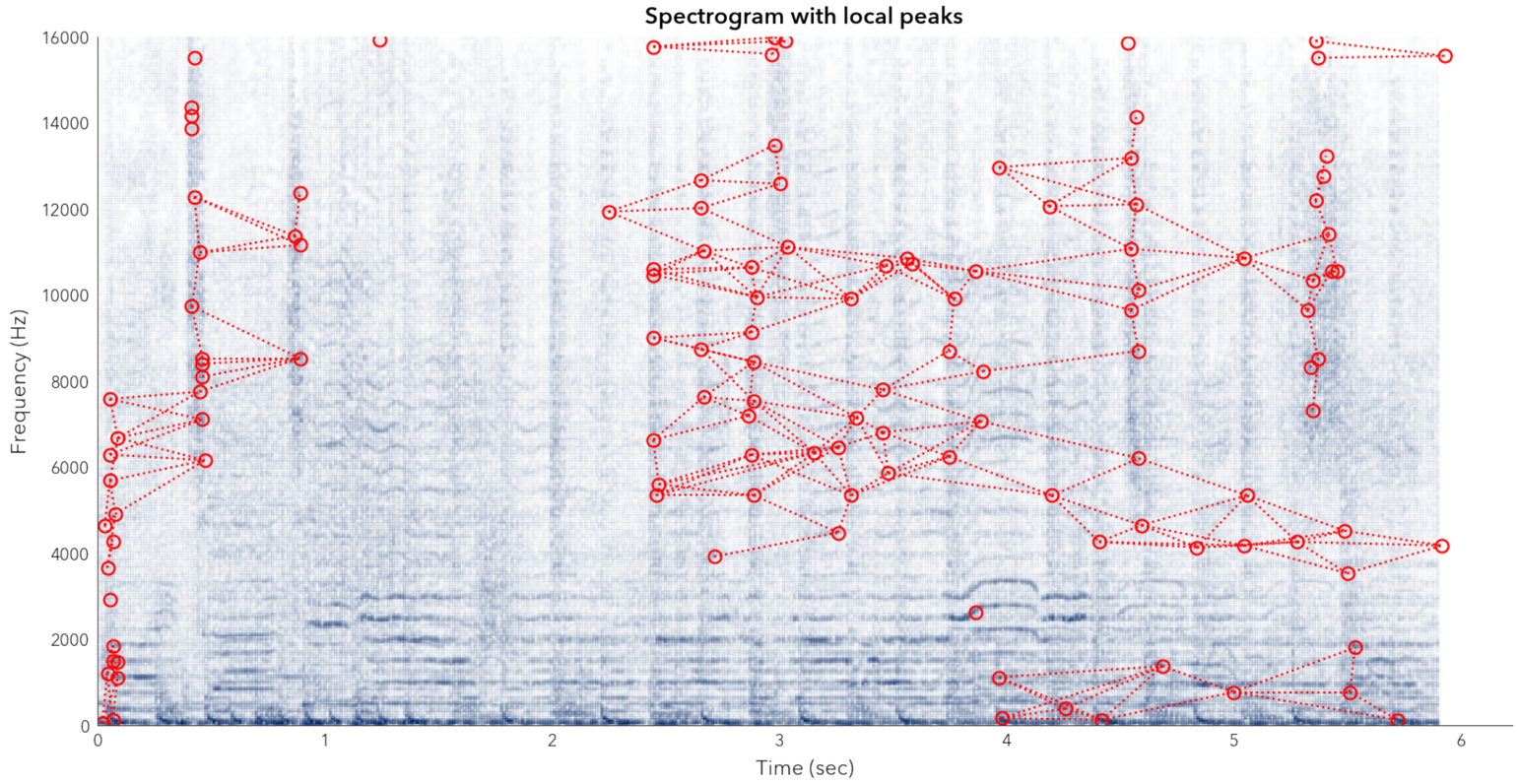
The advantage of constellations

- Constellations are rather robust
  - Peaks don't change much under transforms



Adding a bit more info

- Connect neighboring points to form pairs
  - Makes it very unlikely to get false matches
  - New representation: [time, Δtime, $freq_1$, $freq_2$]



Hashing for fast matching

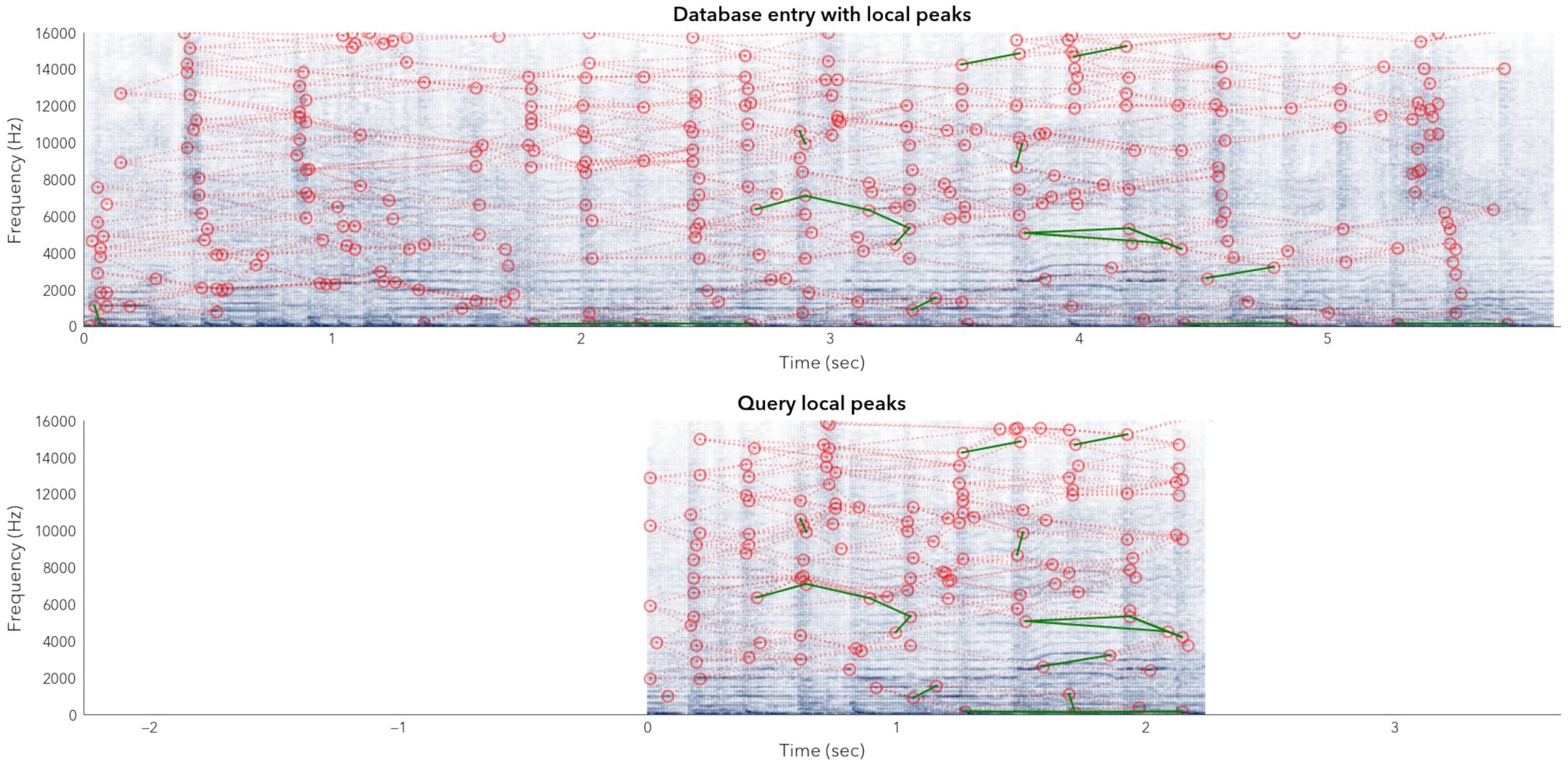
- For each recording make a table of constellations
  - Take all constellation features and hash them
    - e.g. pack in 16 bits int: [0000 $\underbrace{\text{xxxx}}_{dtime}$ $\underbrace{\text{xxxx}}_{freq_1}$ $\underbrace{\text{xxxx}}_{freq_2}$]

- Find database entries sharing hashes with query
  - Common hashes imply similar spectral features
    - Lots of these imply a very similar spectrogram

- Very efficient since we can precompute database hashes

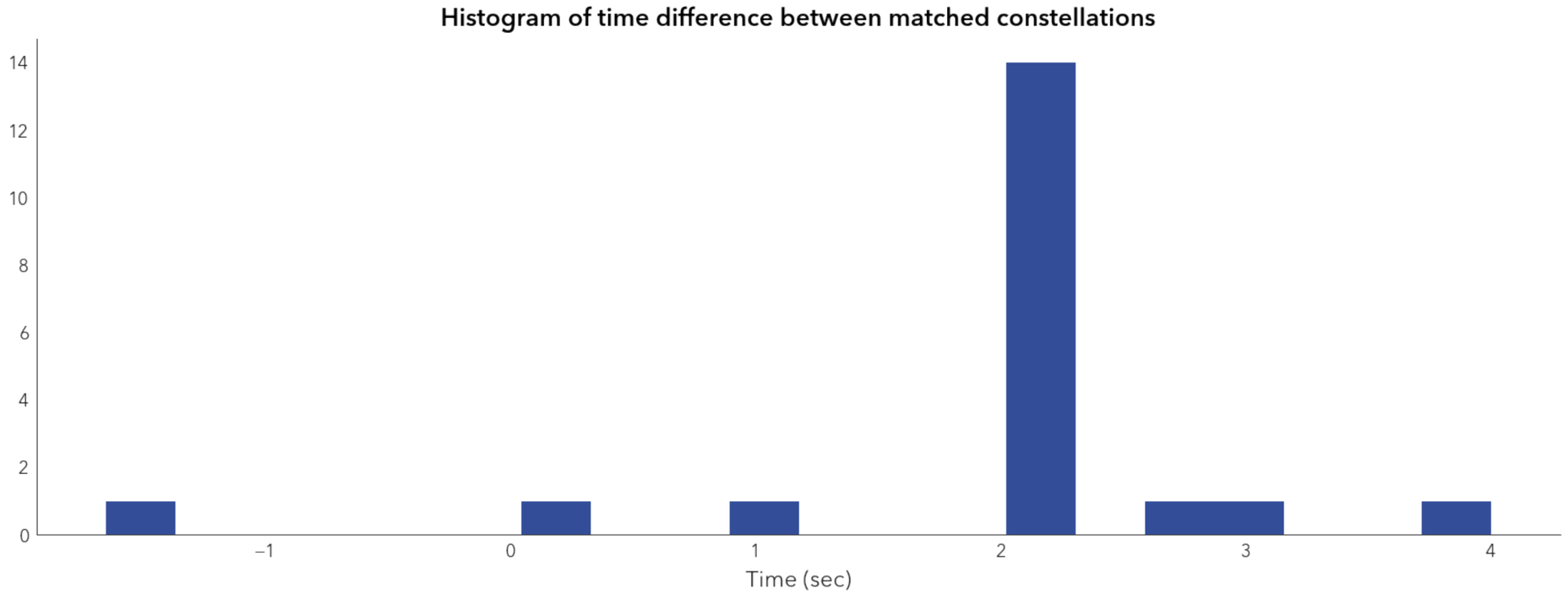
Matching a song snippet

- The query is usually offset and smaller



Looking at matches that matter

- Not all matches will be correct
  - But these that have the same time offset between the files will suggest that the two files are probably the same



Fast and robust performance

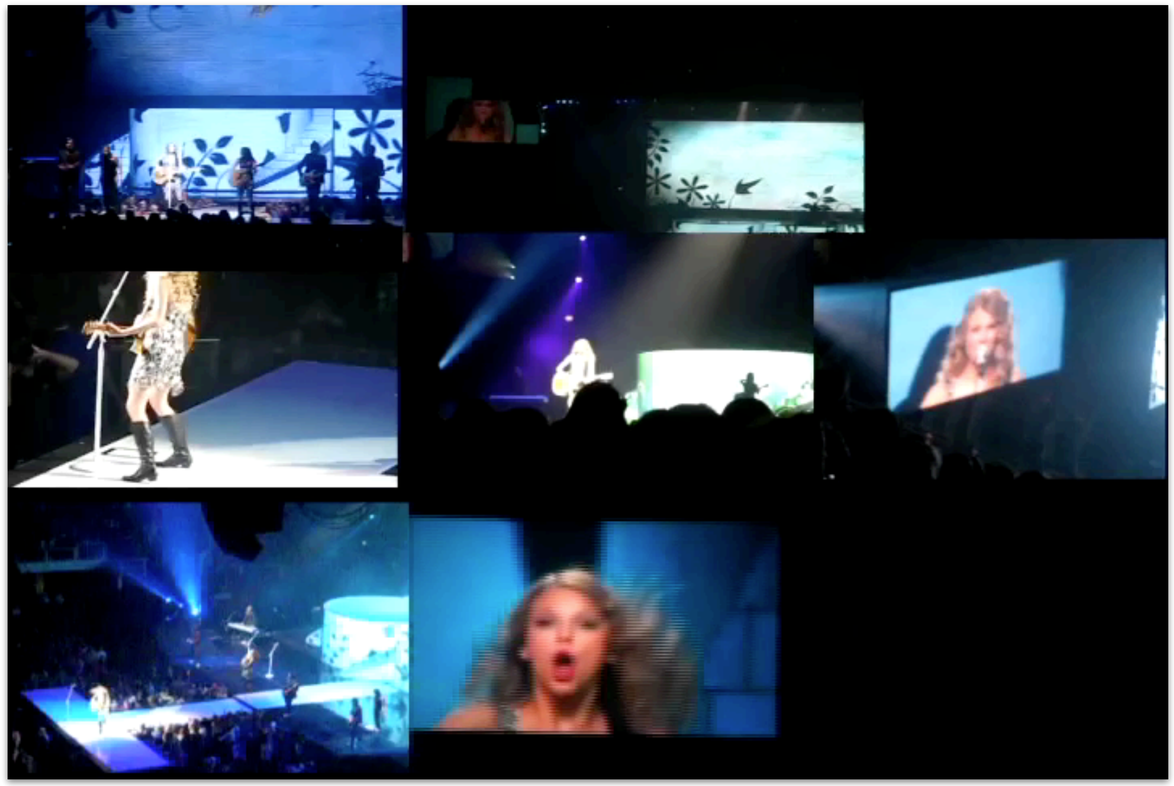
- Robust to noise contaminated queries
  - 90% for 3dB SNR, 80% for 0dB SNR
- Rapid searching due to features/hashing
  - 20,000 song database, 5-500ms per query!
    - How fast would correlation be?

An alternative usage

- Hash matching speeds up correlations
- So we can speed up other correlation-based methods (e.g. work with arrays)

A social experiment

- Here's Taylor Swift at the HP Pavilion at San Jose
  - As the fans recorded her (hundreds of times)

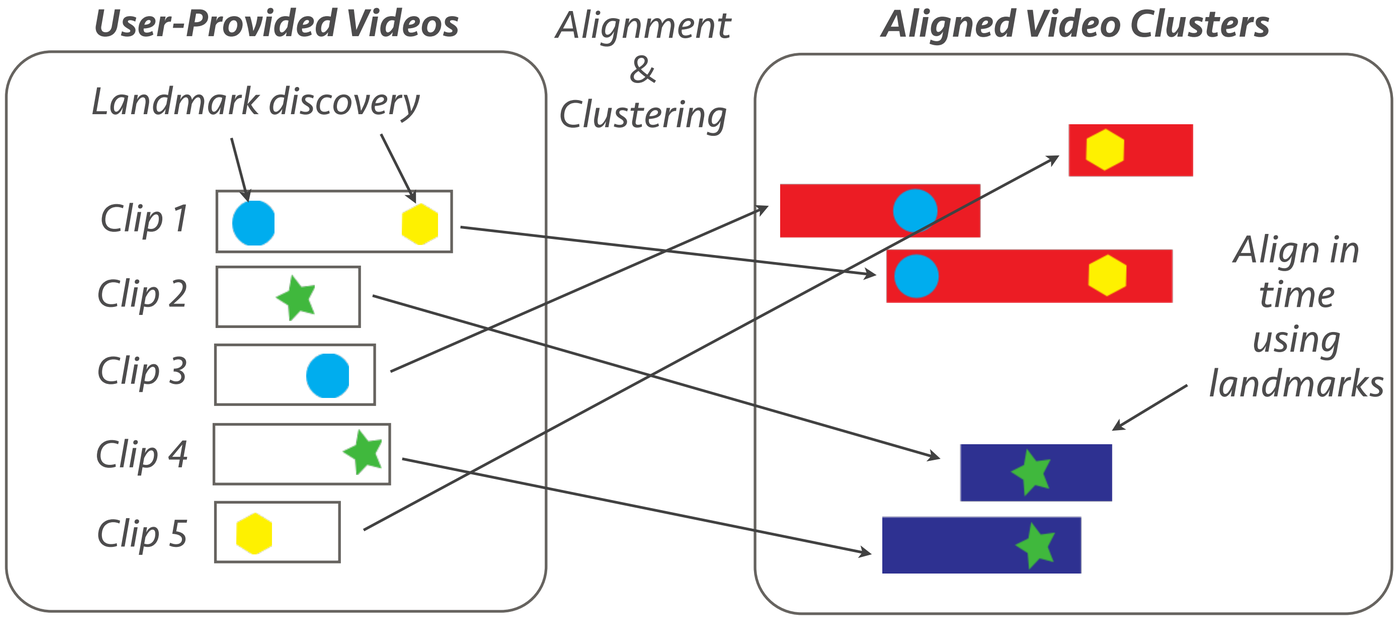


A problem

- How do we relate 100's or 1000's of videos?
- We don't want to directly correlate the audio tracks!
  - E.g. 200 people out of 5,000
  - 4 videos per person = 800 videos
  - 800^2 = 640,000 correlations
  - 2 min per average video = 1,600 minutes = 26 hours of footage
  - 44,100 samples per sec = 5,292,000 samples per clip
  - 1 correlation = 28 Trillion = 28 TeraFLOPS
  - Total cost = 17 Quintillion FLOPS = 17 ExaFLOPS

Can instead use fingerprinting

- Find spectrogram constellations
  - Match them using fast hashing to avoid cross-correlating



Outcome

- All the fan videos are now time aligned
  - We can do interesting followup processing now

Machine listening

- Broadening up to the domain of all sounds
  - Listen to anything that makes a sound!

- Tasks to do:
  - Recognition, scene analysis, localization, interpretation, object formation, ...

Some well-known fields of use

- Speech technology
  - Speech recognition, speaker verification, language identification, meeting diarization, …
- Music Information Retrieval
  - Just talked about it!
- But there’s more

Video Content Analysis

- Audio for event detection
  - Classify sounds to perform semantic analysis
    - Specific to broadcast type
- Use classification for content analysis
  - Clustering videos
  - Finding sections
  - Scene searching
  - Commercial Detection
    - Ads sound different
- Already built many consumer products

Hard computer vision problems

Was there a goal?

Sad or funny clip?

Detecting highlights

Traffic Monitoring

- Methodology
  - To optimize an intersection via monitoring
  - Record only the “interesting” parts
- “Interesting” event discovery
  - Difficult visual task
  - Easier in audio domain
- Sound classifier
  - 5%/4% false negatives/positives
- Outlier detection to find accidents

Examples of actual detected “interesting” parts

Normal crash

Hard-to-see crash

Near crash

Notable (?) event

Security Surveillance

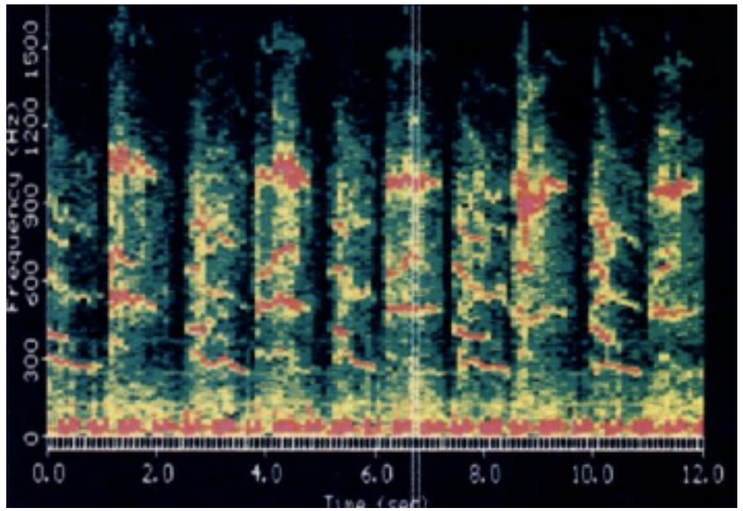
- Use 6 relevant sound classes
  - Normal speech, excited speech, footsteps, thumps, door open/close, screams
- When detecting suspicious sounds raise an alert
  - 96% accuracy in elevator test recordings with actors

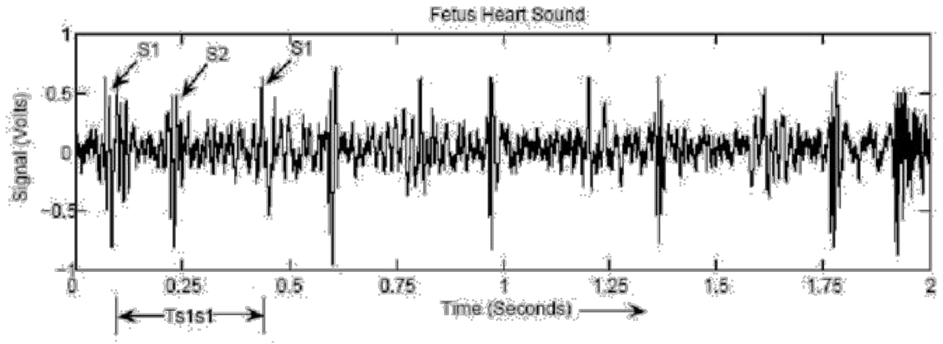
Elevators are a dark environment with poor visual analysis prospects

Audio analysis can provide optimal detection of distress sounds

Medical applications

- Medical acoustics and diagnoses
  - Pulmonary sounds
  - Cardiac sounds
  - Snoring analysis
  - Speech pathology





Mechanical monitoring

- Heavy machinery makes sounds
  - That convey information!
- Diagnostic applications
  - Cars, elevators, ACs, generators, etc …

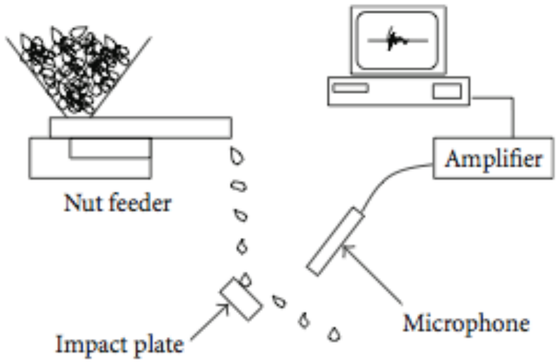
Many more …

- The list goes on …

Sound based location tracking

Snoring statistics

Gunshot detection and localization



Nuts quality control!

Recap

- Music information retrieval
  - Genre/Artist classification
  - Audio fingerprinting

- Machine listening
  - Video mining, surveillance, medical stuff, etc ...

Reading

- Genre classification
  - http://webhome.cs.uvic.ca/~gtzan/work/pubs/tsap02gtzan.pdf
  - http://paris.cs.illinois.edu/pubs/smaragdis-ismir02.pdf
- Audio fingerprinting
  - http://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf
- Audio features
  - http://recherche.ircam.fr/equipes/analyse-synthese/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf

Next time

- Secure Multiparty Computations

- By request:
  - Generative Neural Net Models
  - Attention and Transformers
  - Reinforcement training
  - A bit on regression/regularization
  - Current trends

Project presentations

- We are having presentations after the break
  - Add your project slides here:
    - https://bit.ly/32U1qn4
  - Do so in order of presentations as shown here:
    - https://bit.ly/3pEpF2r

- Keep it short - you only have 7 mins
  - Describe the problem
  - Outline your approach
  - Show any preliminary results