Audio Retrieval

David Kauchak

cs160

Fall 2009

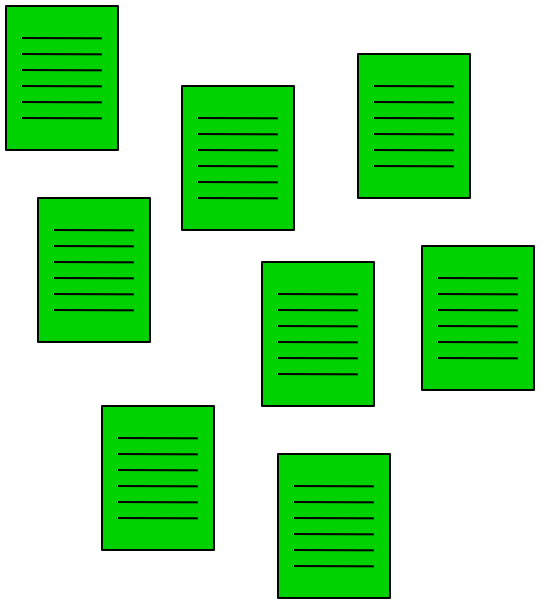
Administrative

- Assign 4 due Friday
- Previous scores

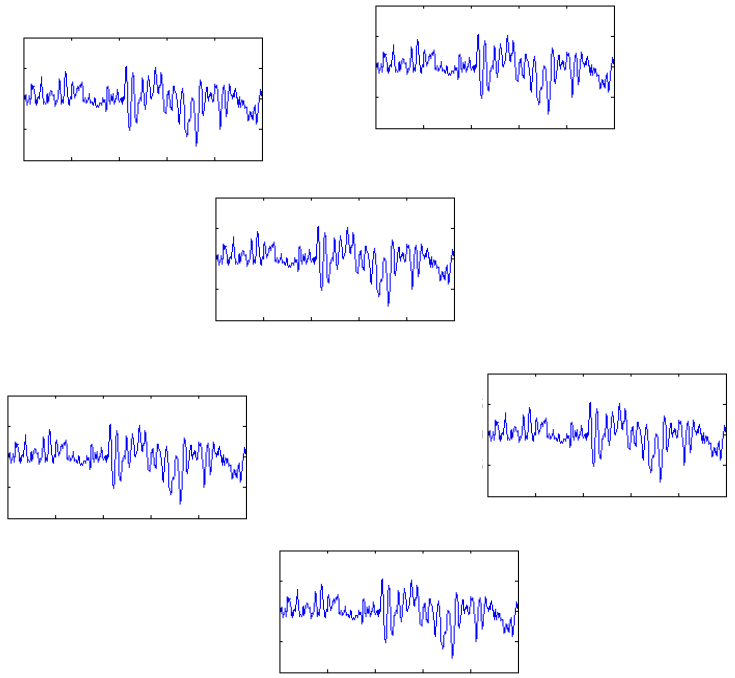
Final project

Audio retrieval

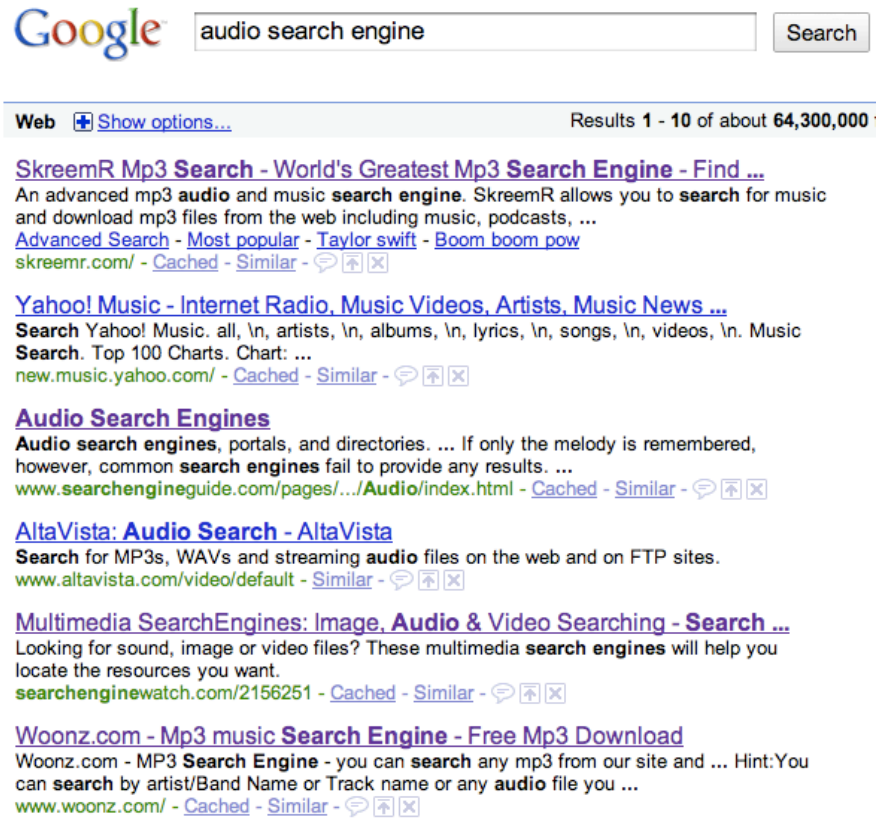
text retrieval corpus

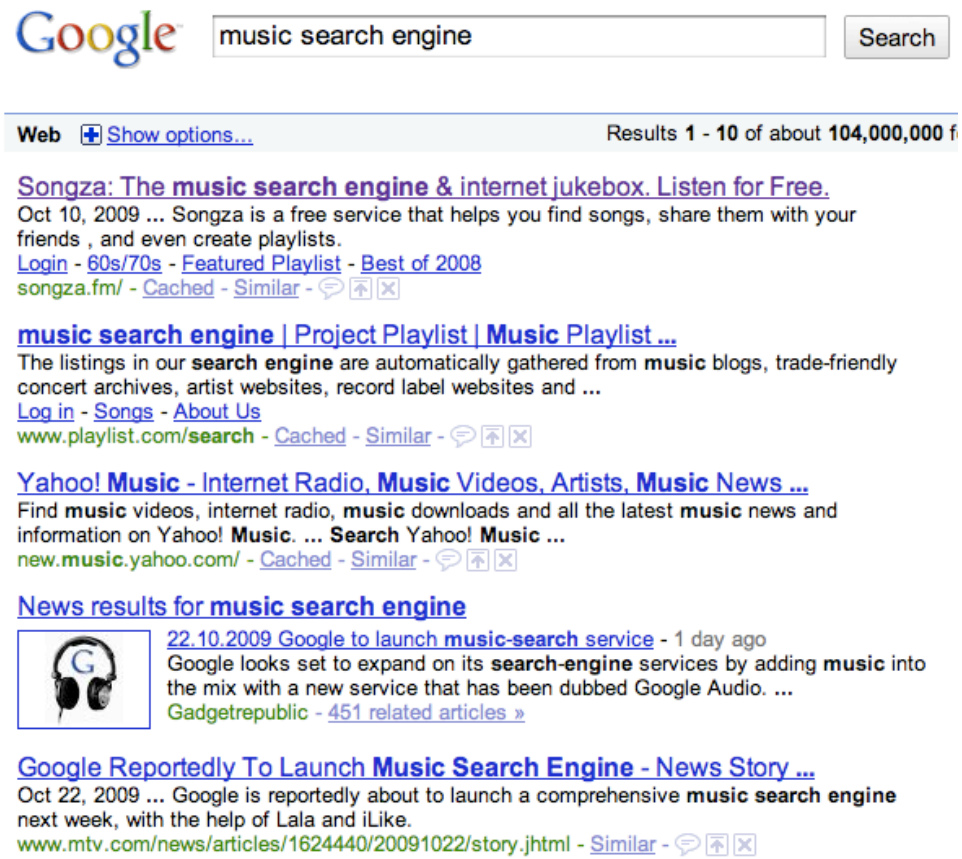


audio retrieval corpus



Current audio search engines





What do you want from an audio search engine?

- **Name**: You might know the name of the song or the artist
- **Genre**: You might try "Bebop," "Latin Jazz," or "Rock"
- **Instrumentation**: The tenor sax, guitar, and double bass are all featured in the song
- **Emotion**: The song has a "cool vibe" that is "upbeat" with an "electric texture"

- Some other approaches to search:
  - musicovery.com
  - pandora.com (song similarity)
  - Genius (collaborative filtering)

Text Index construction

Documents to be indexed

Friends, Romans, countrymen.

text preprocessing

friend , roman , countrymen .

indexer

Inverted index

friend → 2 → 4

roman → 1 → 2

countryman → 13 → 16

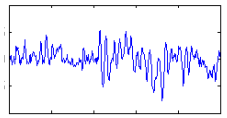
Audio Index construction

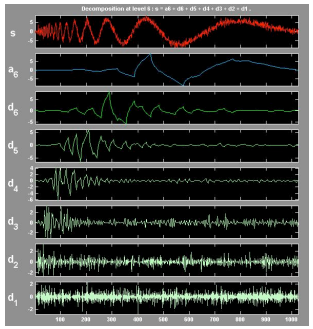
Audio files to be indexed

Today

wav **midi** **mp3**

audio preprocessing





slow, jazzy, punk

indexer

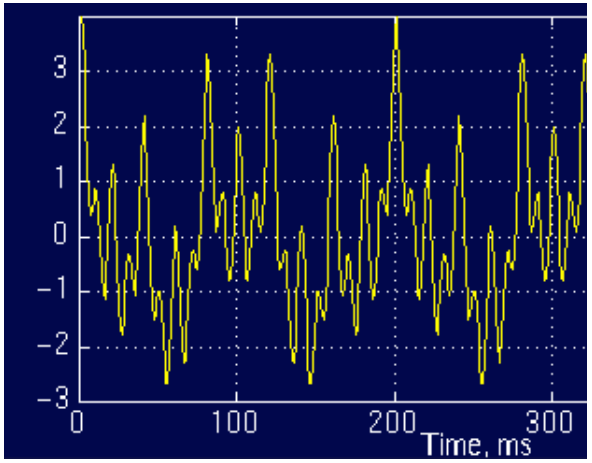
Index

may be keyed off of text

may be keyed off of audio features

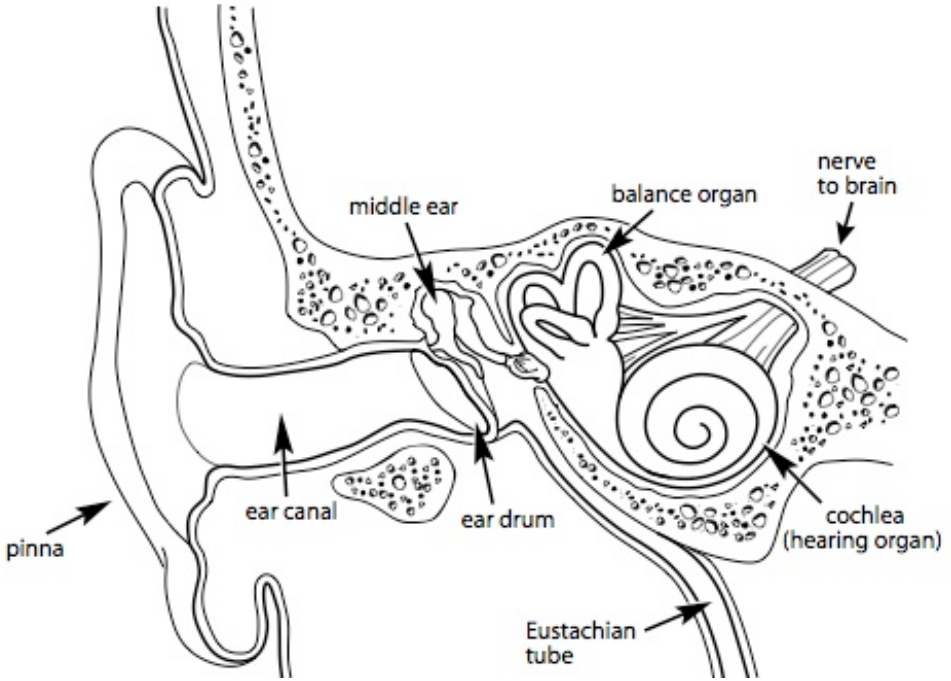
Sound

- What is sound?
  - A longitudinal compression wave traveling through some medium (often, air)
  - Rate of the wave is the frequency
  - You can think of sounds as a sum of sign waves



Sound

- How do people hear sound?
  - The cochlea in the inner ear has hair cells that "wiggle" when certain frequency are encountered



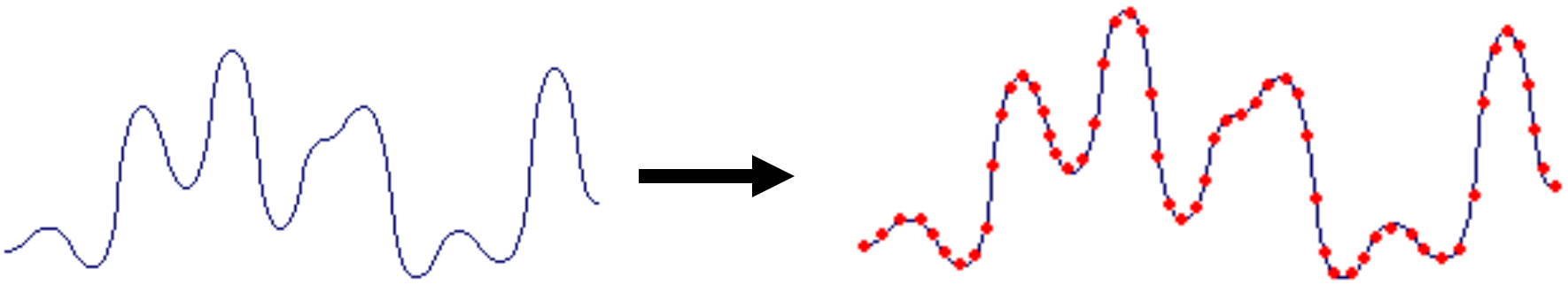
http://www.bcchildrens.ca/NR/rdonlyres/8A4BAD04-A01F-4469-8CCF-EA2B58617C98/16128/theear.jpg

Digital Encoding

- Like everything else for computers, we must represent audio signals digitally
- Encoding formats:
  - WAV
  - MIDI
  - MP3
  - Others…

WAV

- Simple encoding
- Sample sound at some interval (e.g. 44 KHz).
- High sound quality
- Large file sizes



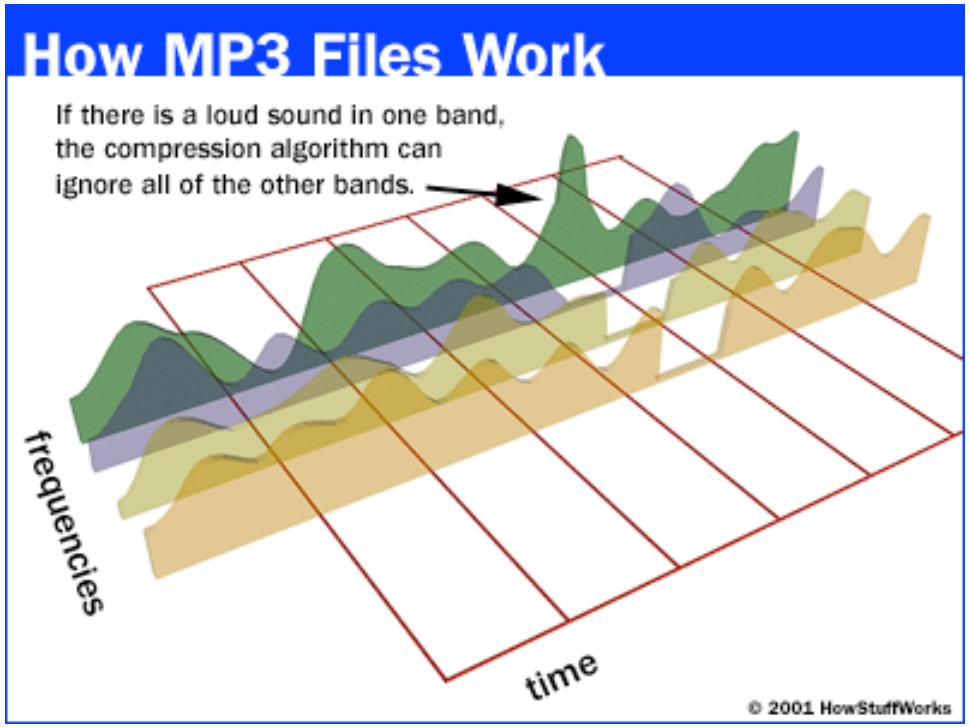
MIDI

- Musical Instrument Digital Interface
- MIDI is a language
- Sentences describe the channel, note, loudness, etc.
- 16 channels (each can be thought of and recorded as a separate instrument)
- Common for audio retrieval and classification applications

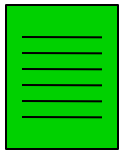
MP3

- Common compression format
- 3-4 MB vs. 30-40 MB for uncompressed
- Perceptual noise shaping
  - The human ear cannot hear certain sounds
  - Some sounds are heard better than others
  - The louder of two sounds will be heard
- Lossy or lossless?
  - Lossy compression
  - quality depends on the amount of compression
  - like many compression algorithms, can have issues with randomness (e.g. clapping)

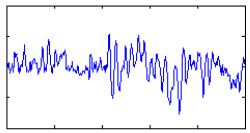
MP3 Example


How MP3 Files Work
If there is a loud sound in one band, the compression algorithm can ignore all of the other bands.
frequencies
time
© 2001 HowStuffWorks

Features



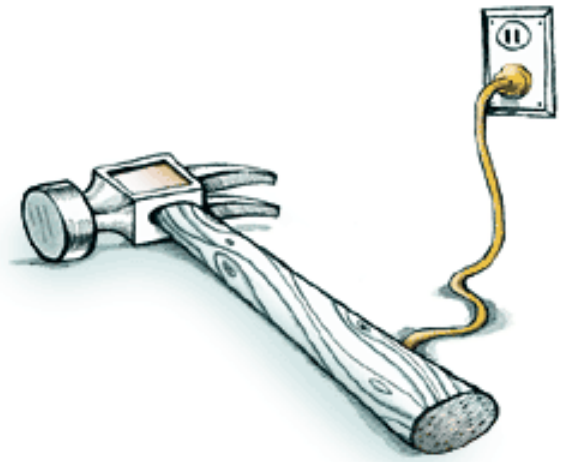
Weight vectors
- word frequency
- count normalization
- idf weighting
- length normalization



?

Tools for Feature Extraction

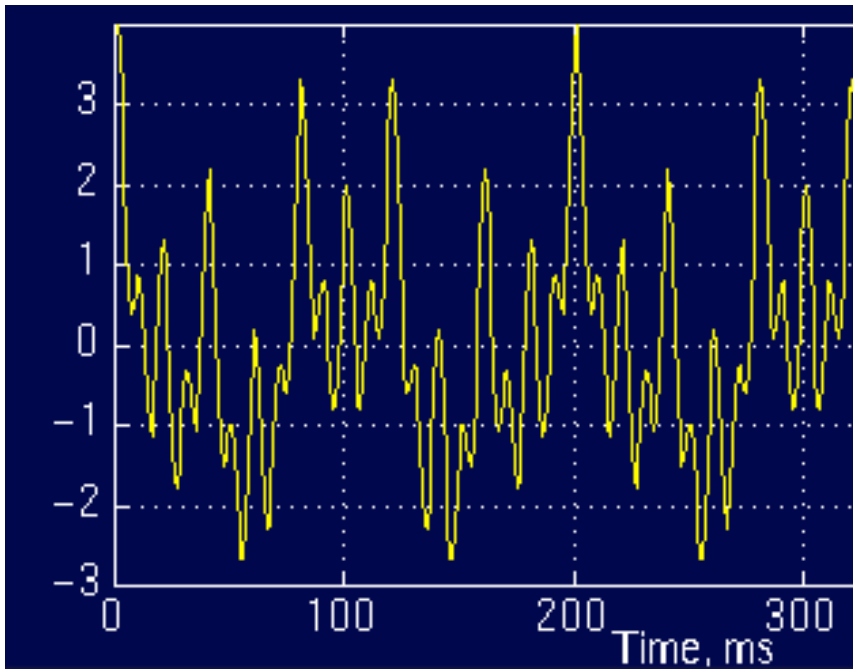
- Fourier Transform (FT)
- Short Term Fourier Transform (STFT)
- Wavelets

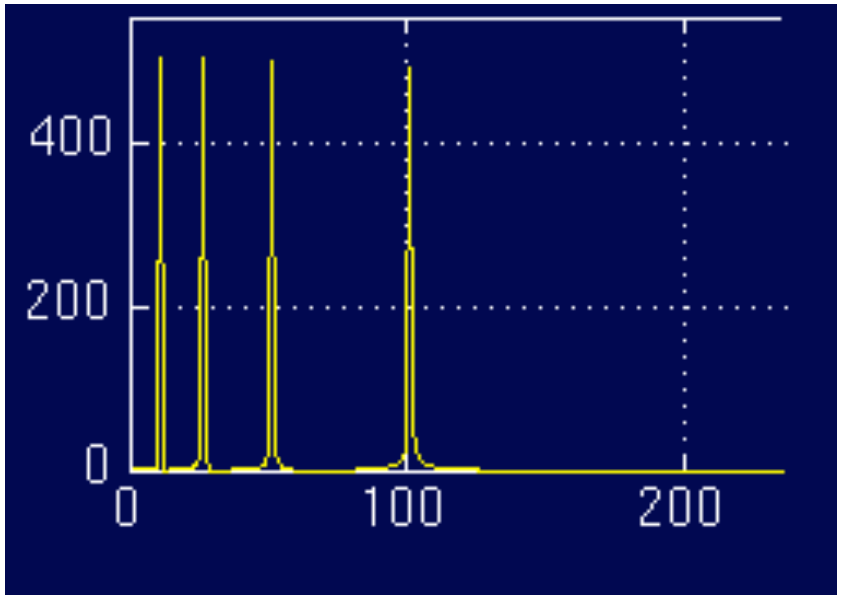


Paul Antonson

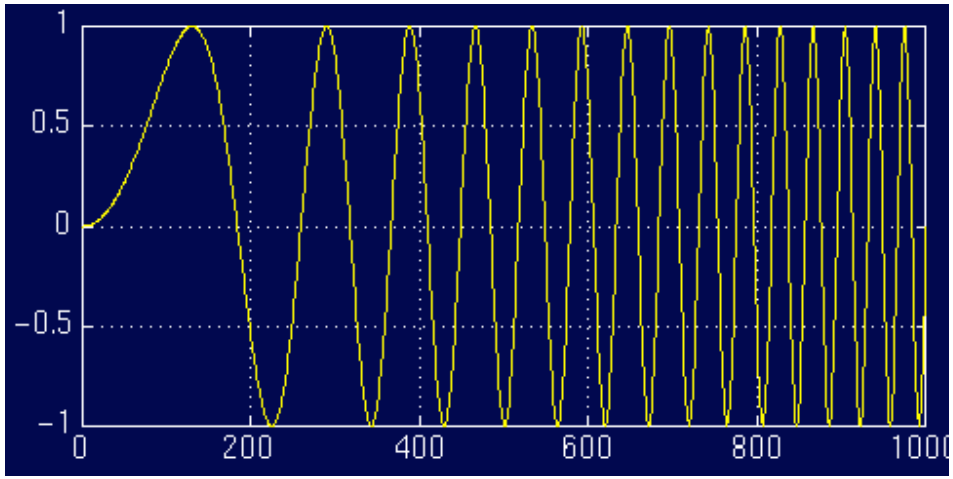
Fourier Transform (FT)

- Time-domain → Frequency-domain



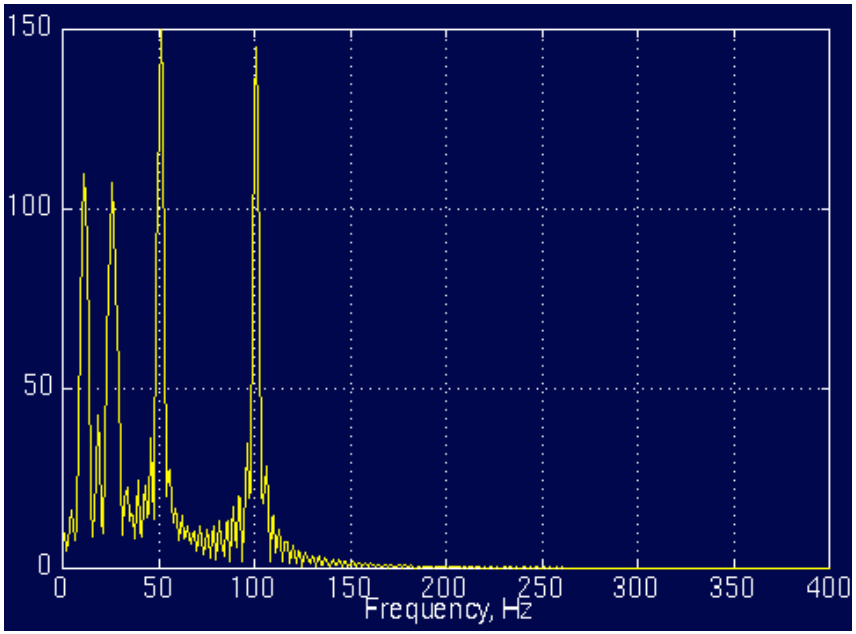


Another FT Example

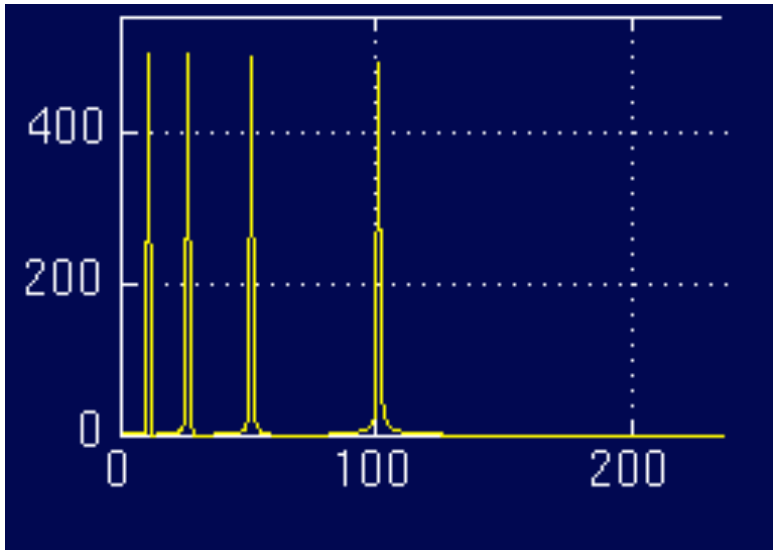


Time

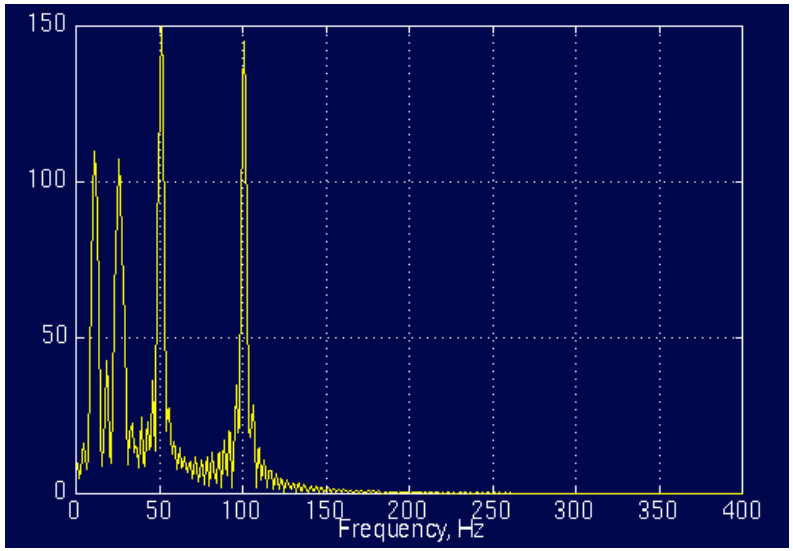
Frequency



Problem?


400
200
0
0
100
200

≈


150
100
50
0
0
50
100
150
200
250
300
350
400
Frequency, Hz

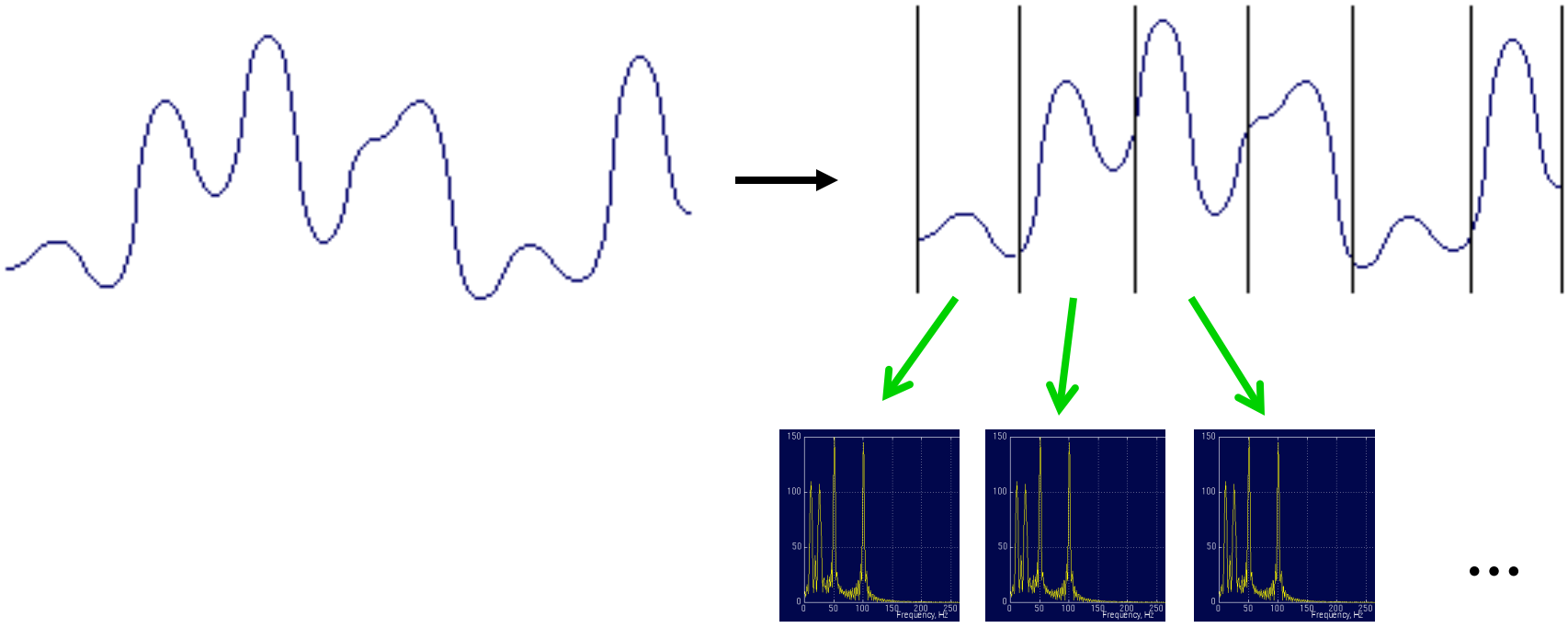
Problem with FT

- FT contains only frequency information
- No **time** information is retained
- Works fine for stationary signals
- Non-stationary or changing signals cause problems
  - FT shows frequencies occurring at all times instead of specific times

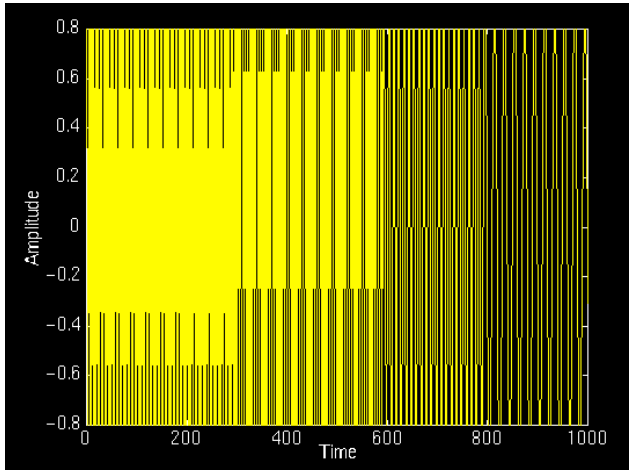
Ideas?

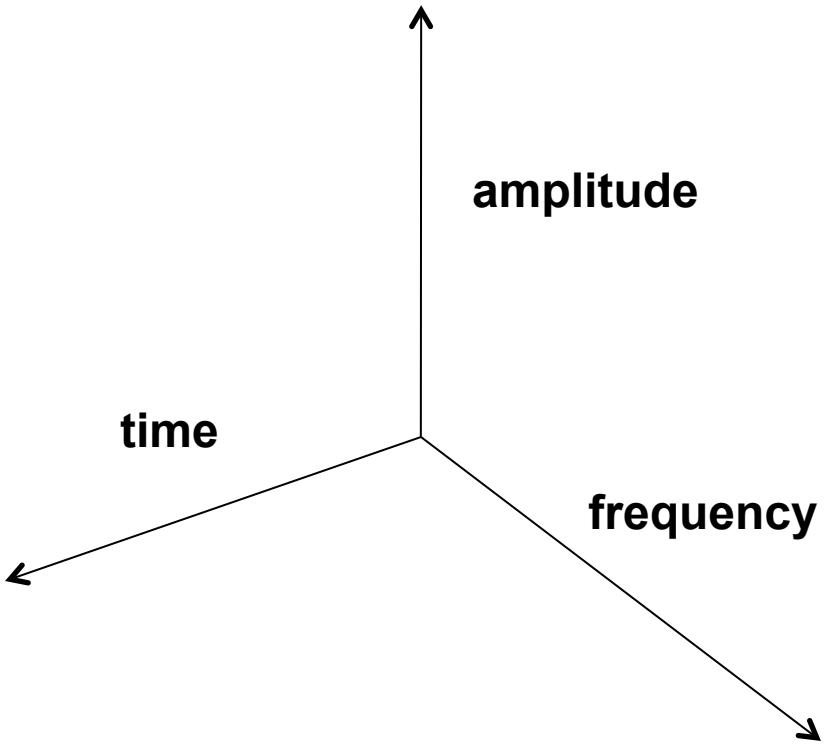
Short-Time Fourier Transform (STFT)

- Idea: Break up the signal into discrete windows
- Treat each signal within a window as a stationary signal
- Take FT over each part

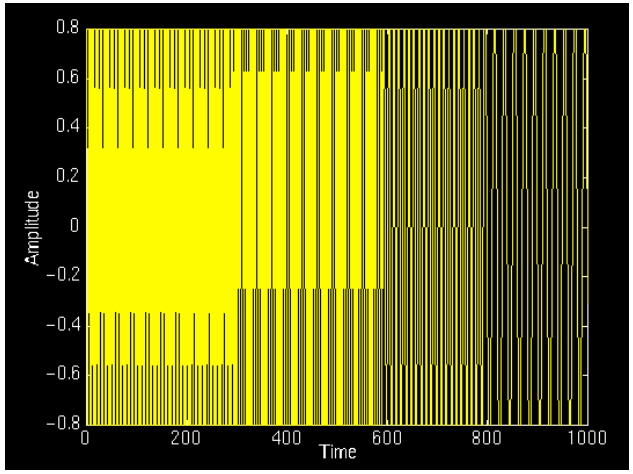


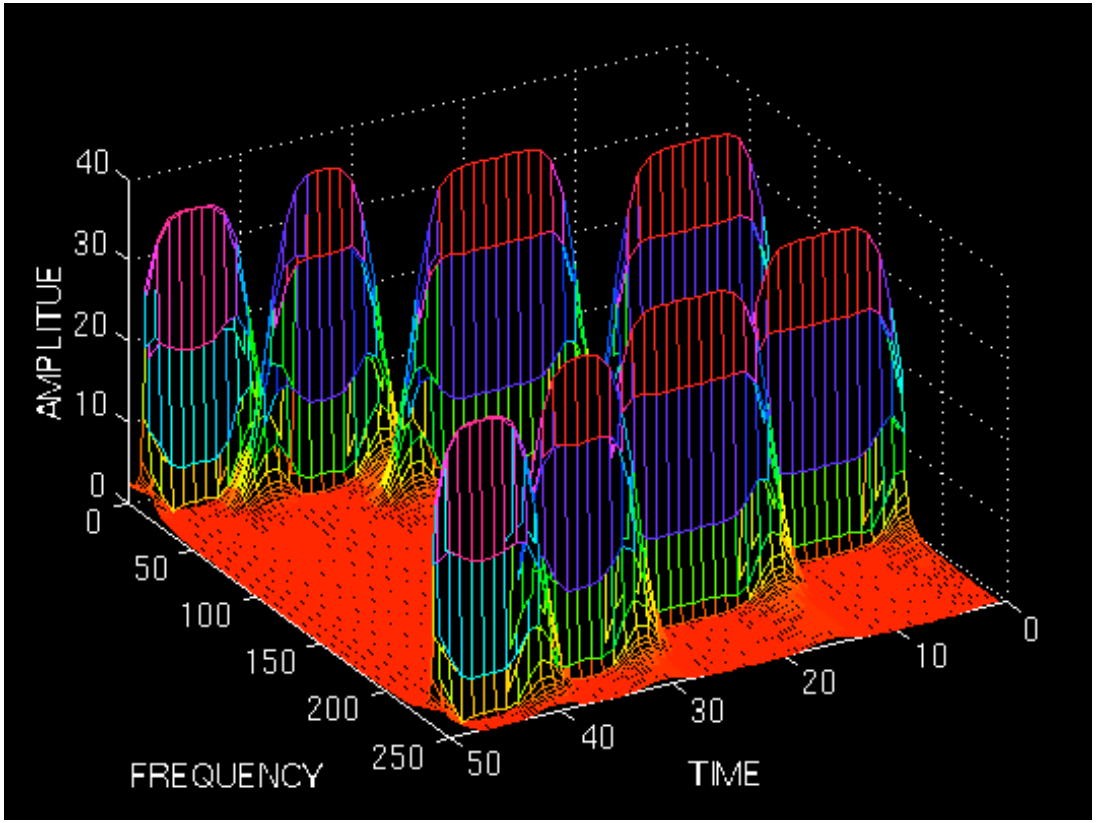
STFT Example





STFT Example

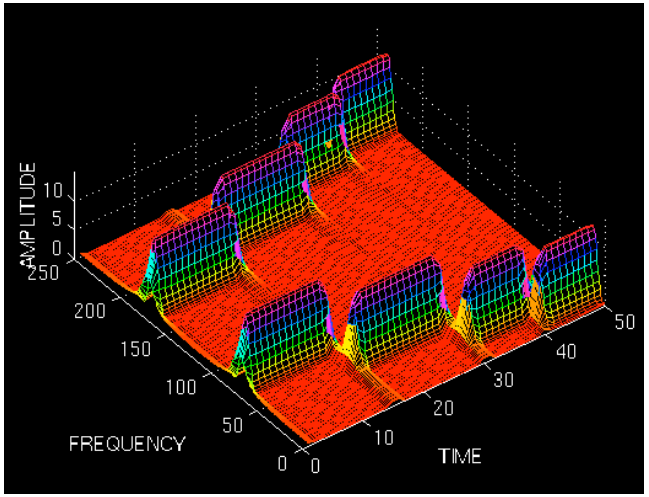


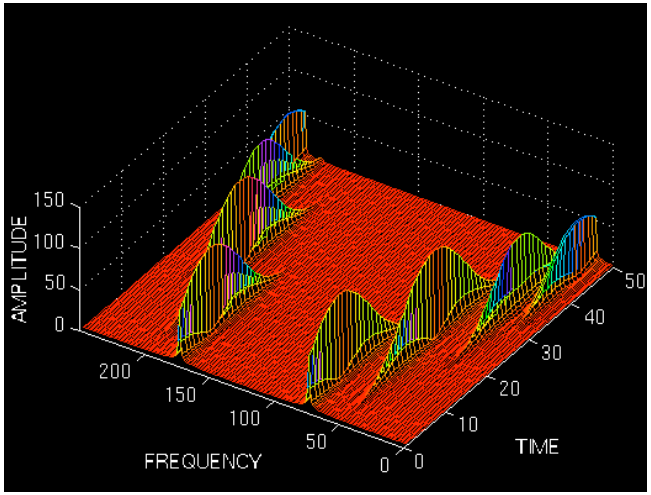


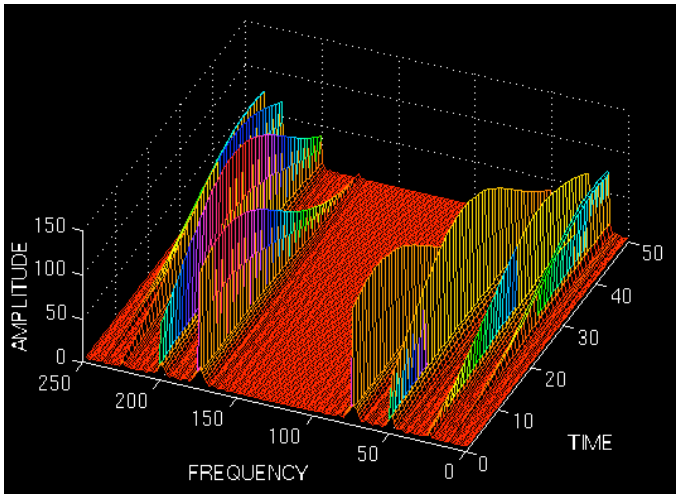
Problem: Resolution

- How do we pick the window size?
- We can vary time and frequency accuracy
  - Narrow window: good time resolution, poor frequency resolution
  - Wide window: good frequency resolution, poor time resolution

Varying the resolution







Ideas?

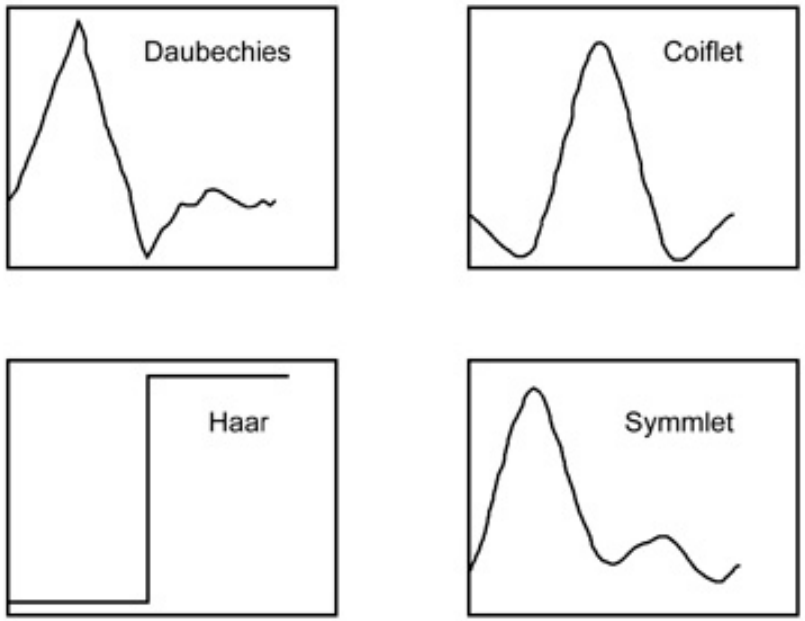
Wavelets

Wave

Wavelets ☺

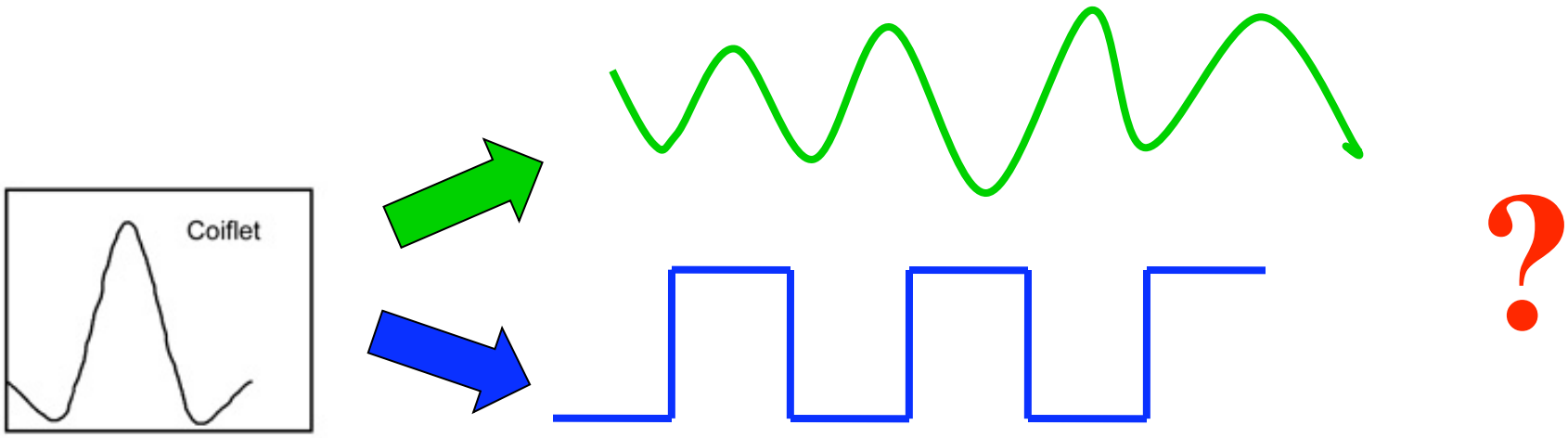
Wavelets

- Wavelets ***respond*** to signals that are similar



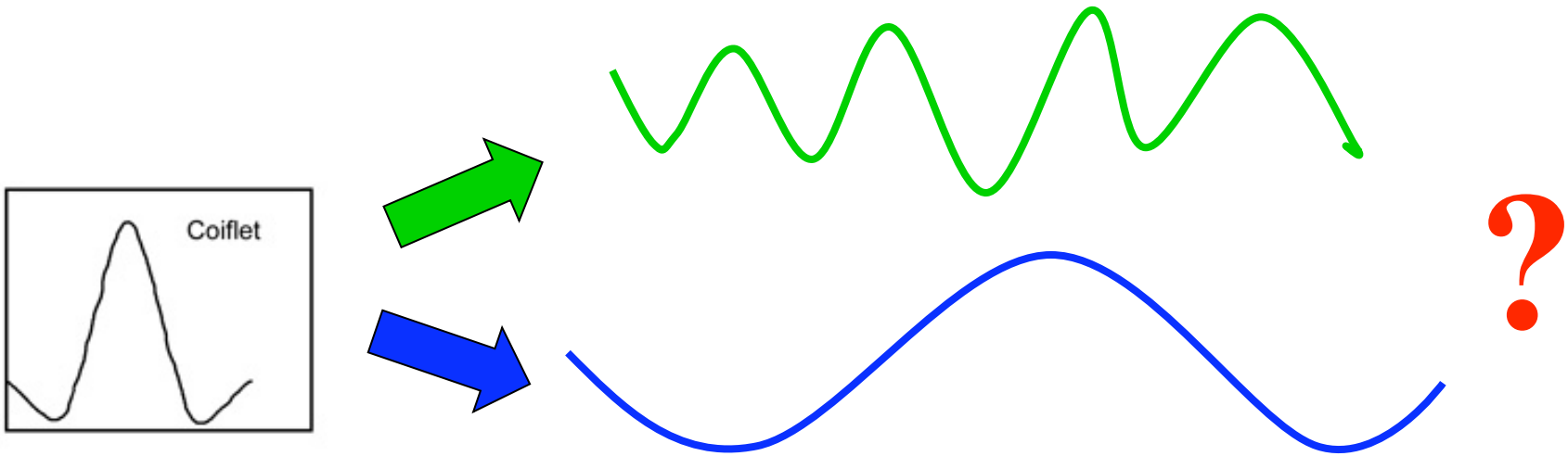
Wavelet response

A wavelet responds to signals that are similar to the wavelet



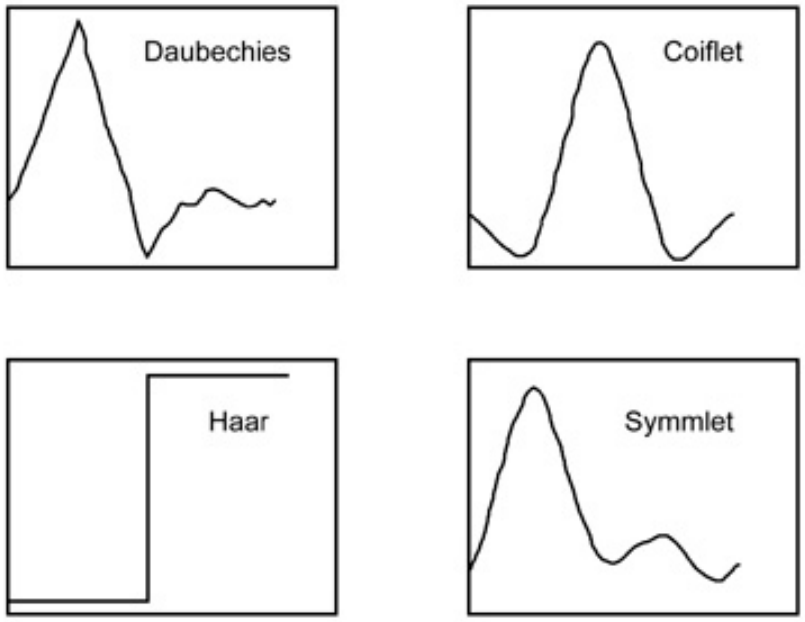
Wavelet response

Scale matters!

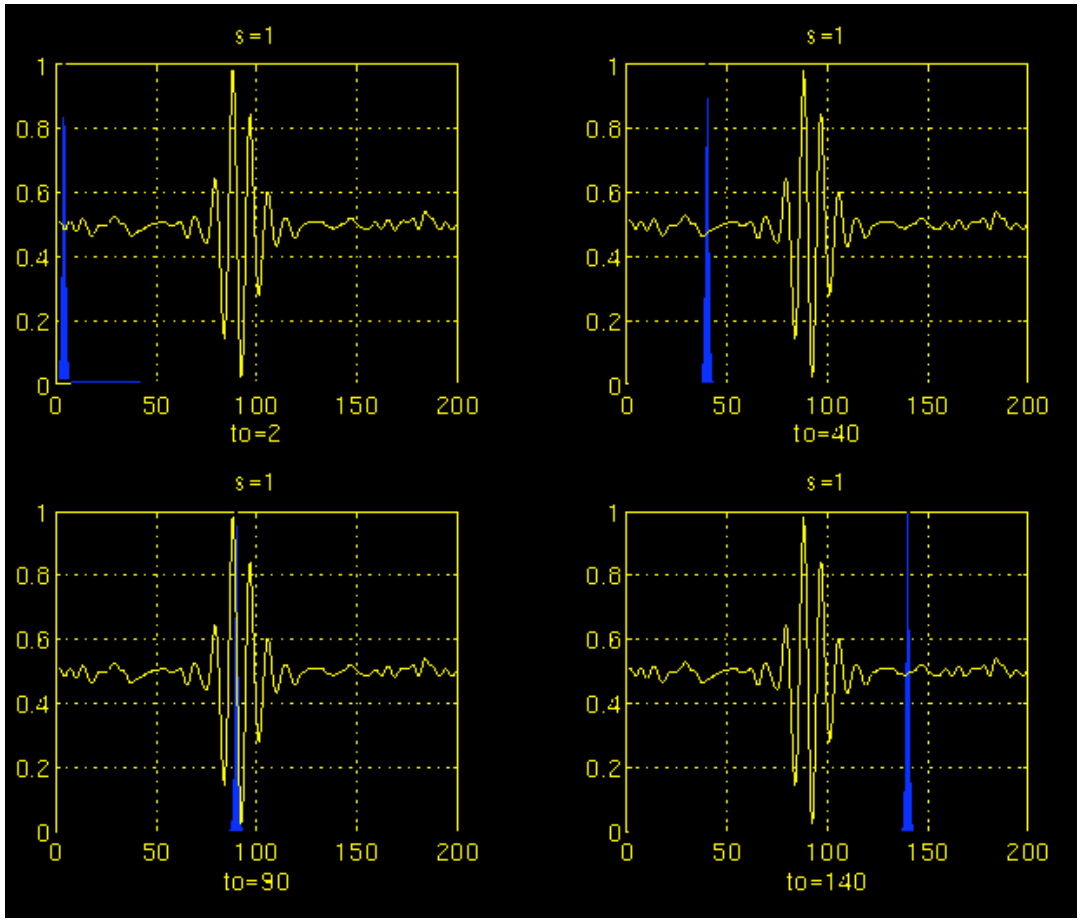


Wavelet Transform

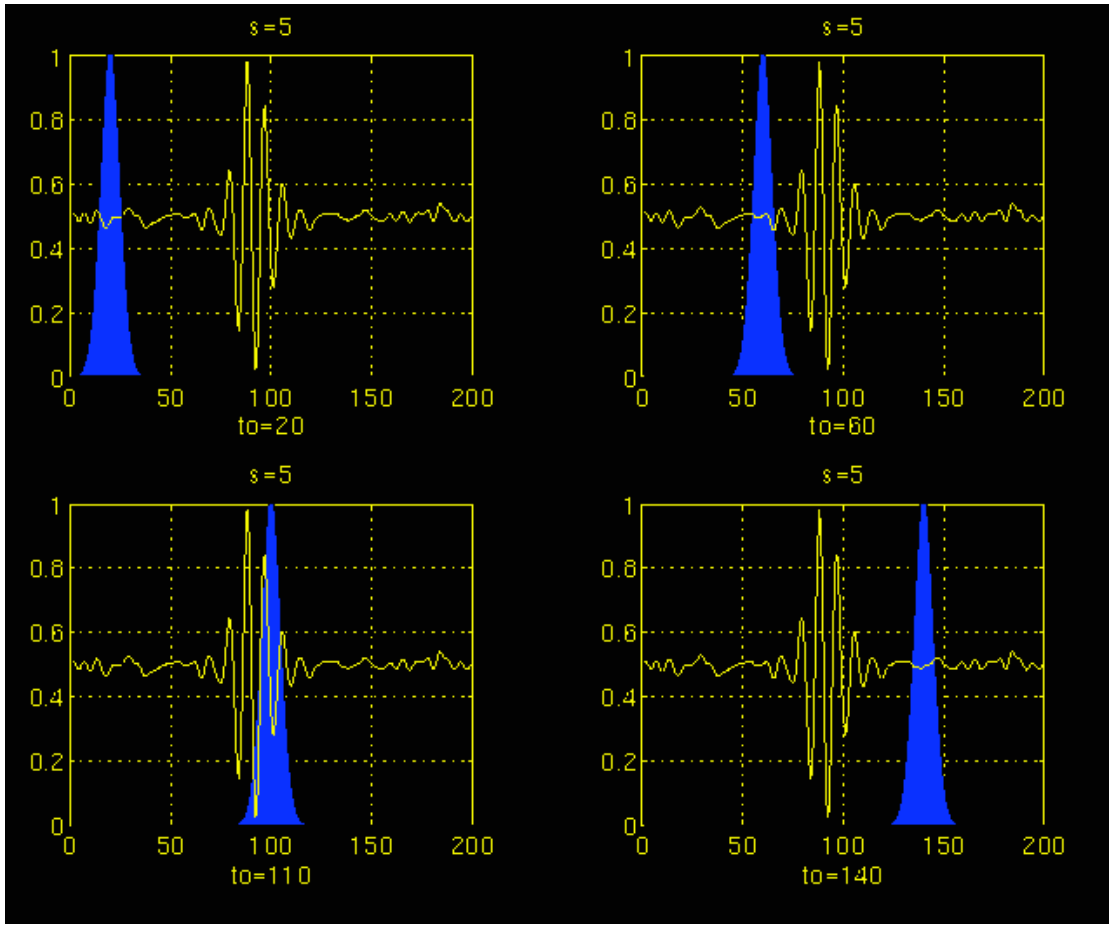
- Idea: Take a wavelet and vary scale
- Check response of varying scales on signal



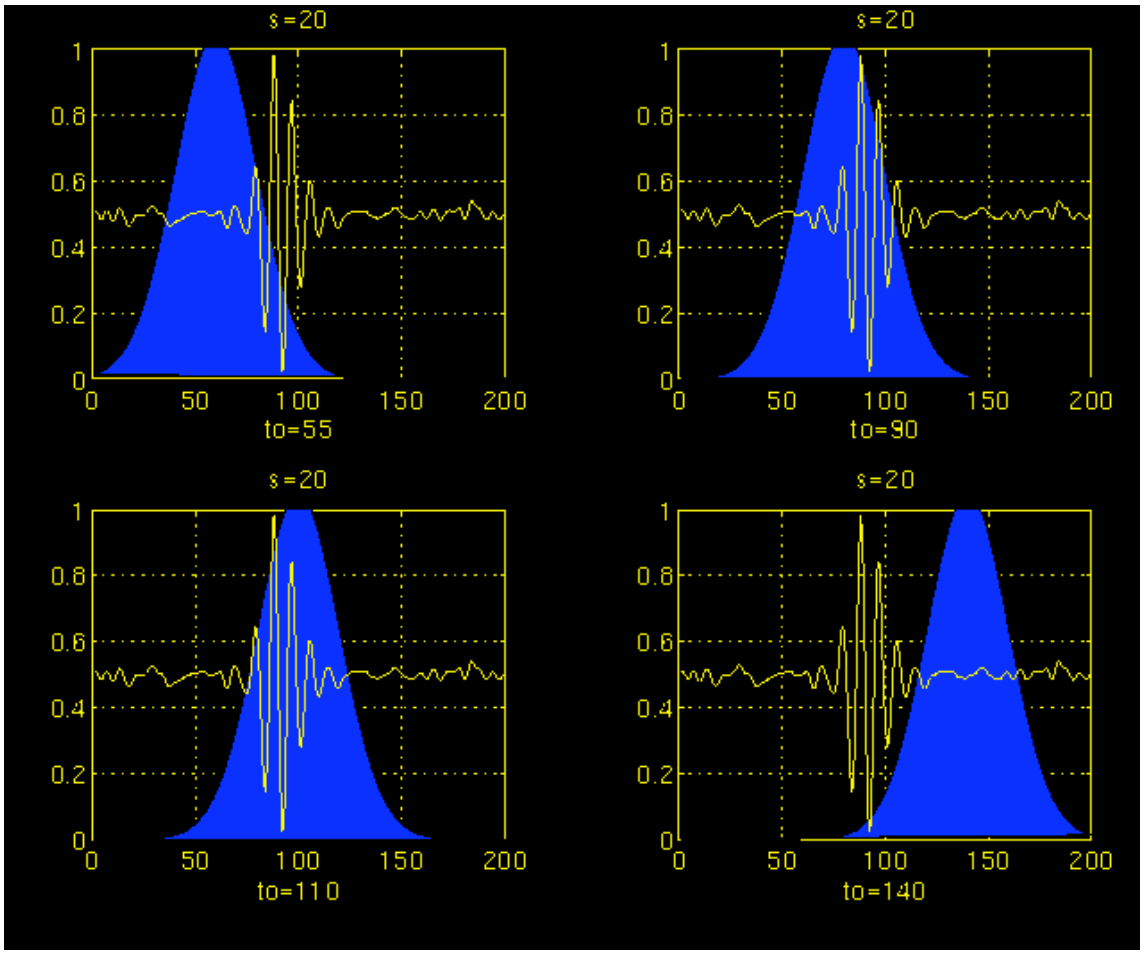
Wavelet Example: Scale 1



Wavelet Example: Scale 2



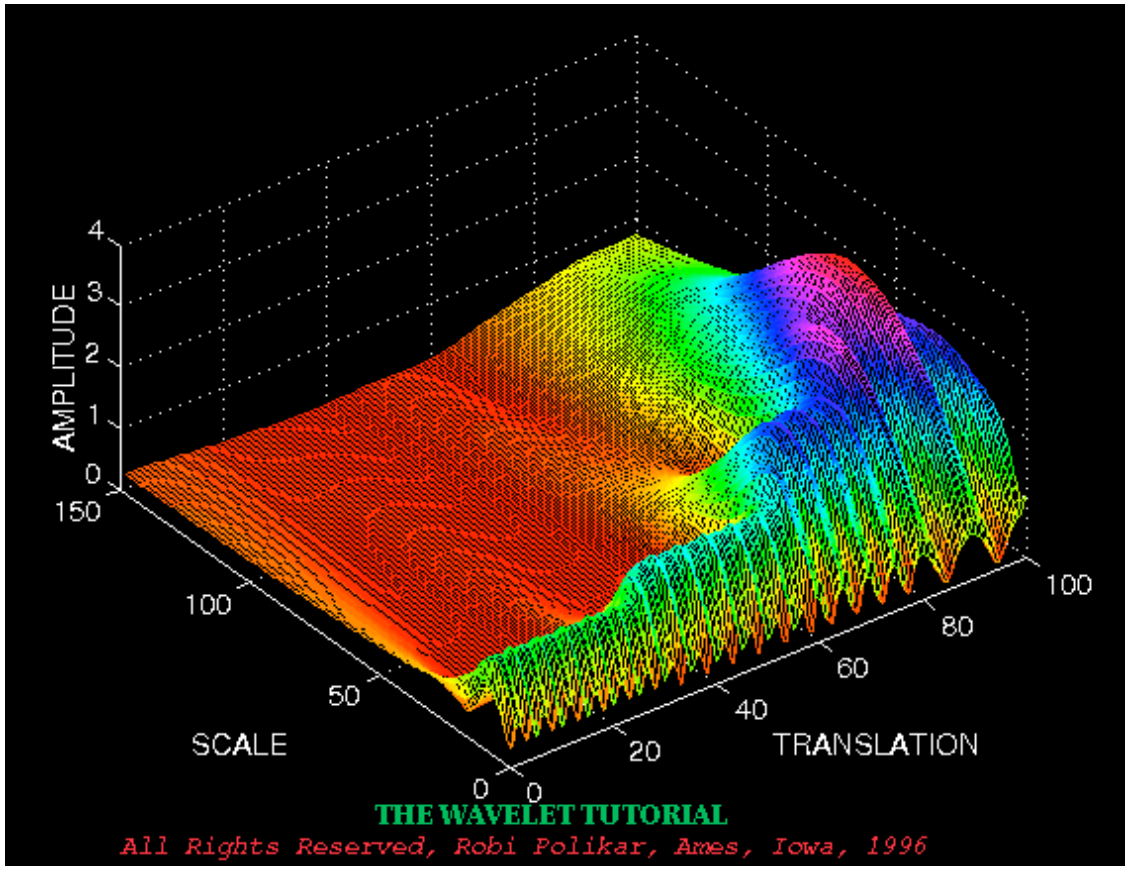
Wavelet Example: Scale 3



Wavelet Example

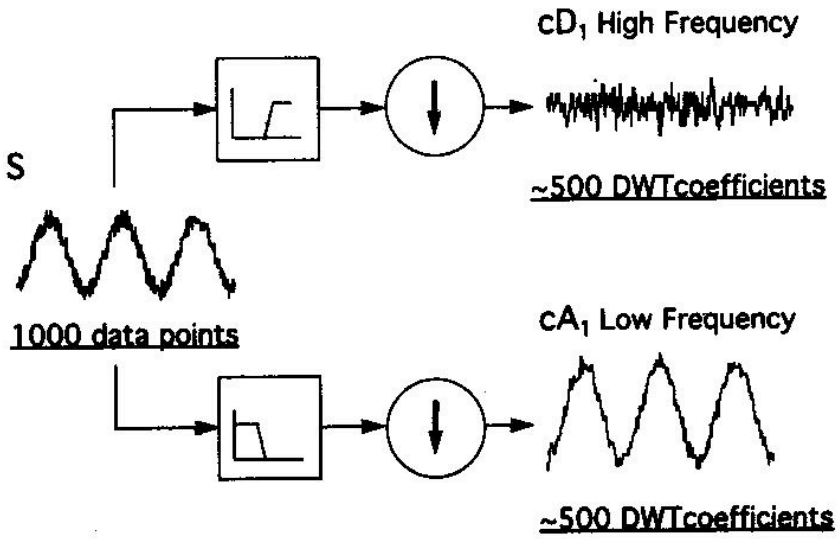
Scale = 1/frequency

Translation ≈ Time



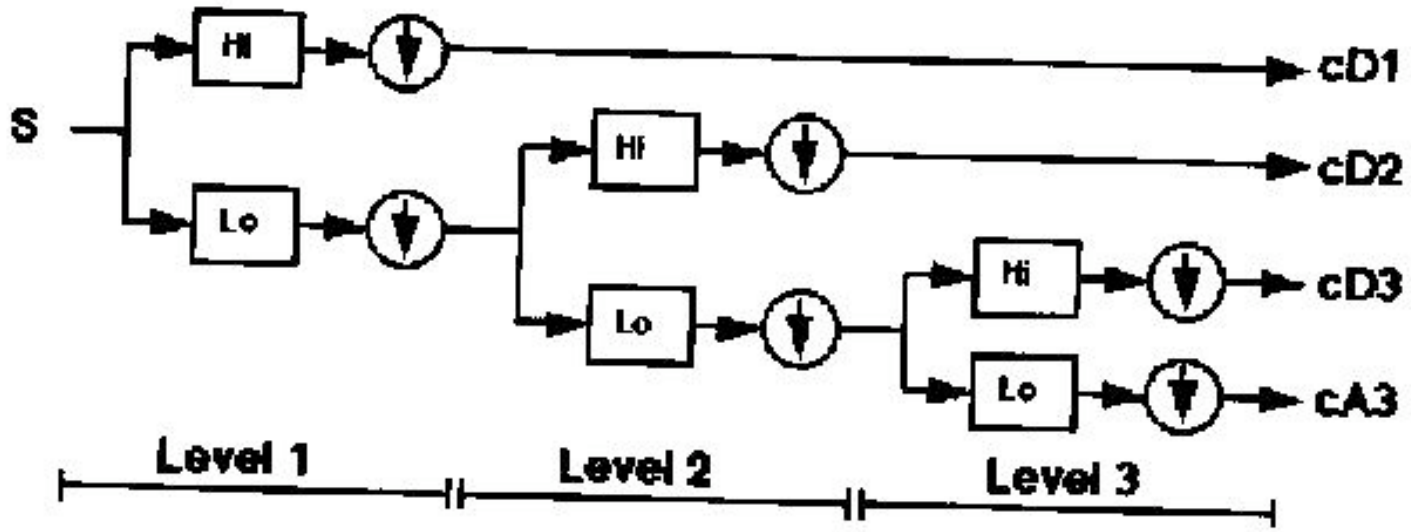
Discrete Wavelet Transform (DWT)

- Wavelets come in pairs (high pass and low pass filter)
- Split signal with filter and downsample

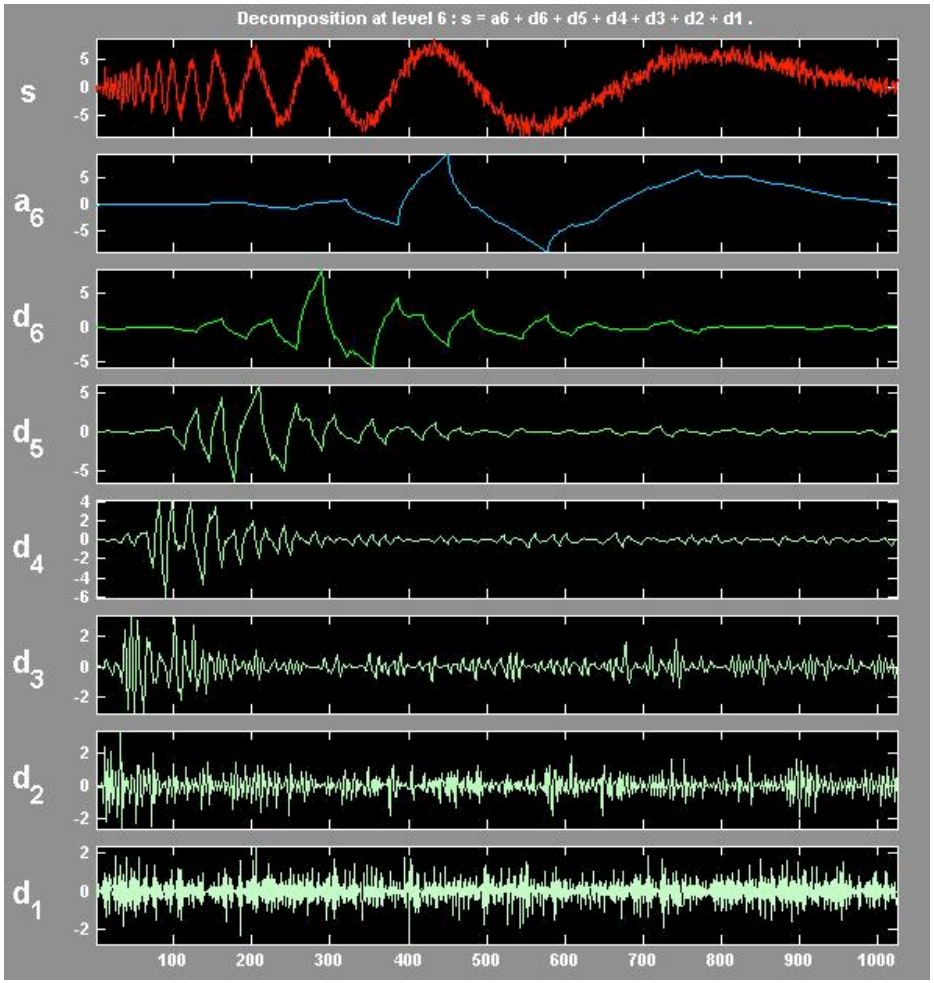


DWT cont.

- Continue this process on the low frequency portion of the signal



DWT Example



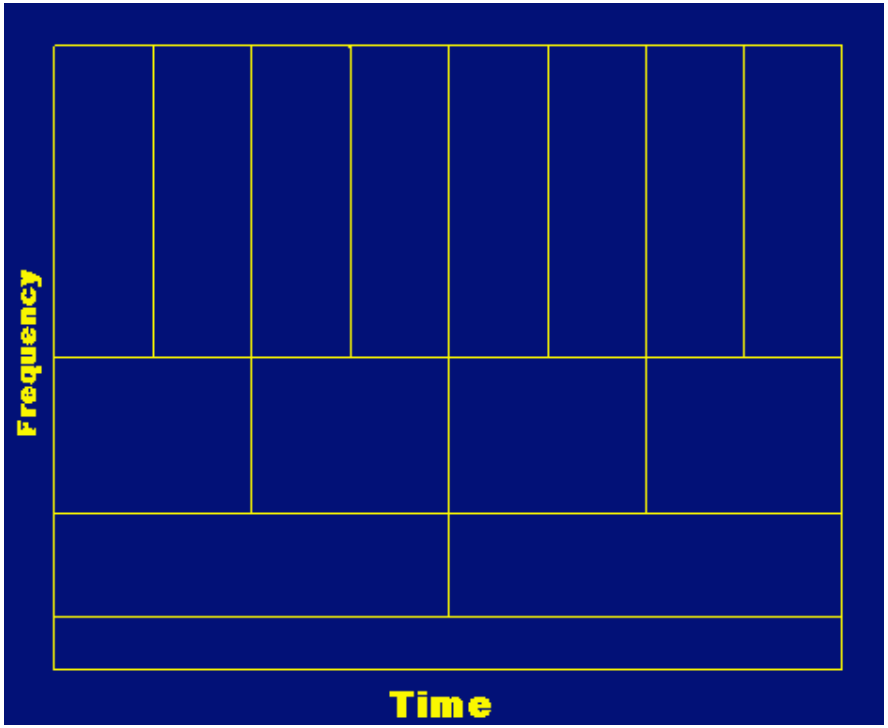
signal

low frequency

high frequency

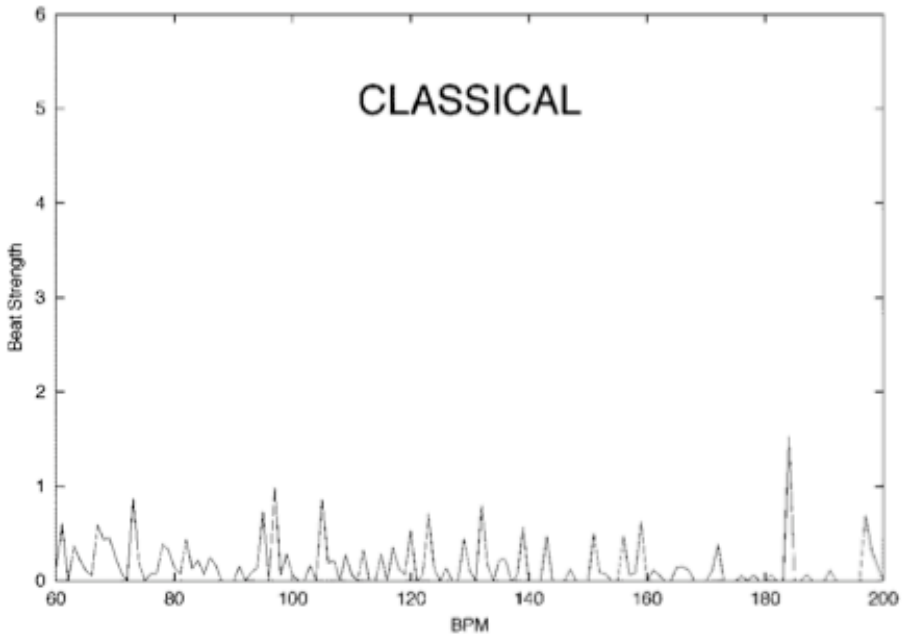
How did this solve the resolution problem?

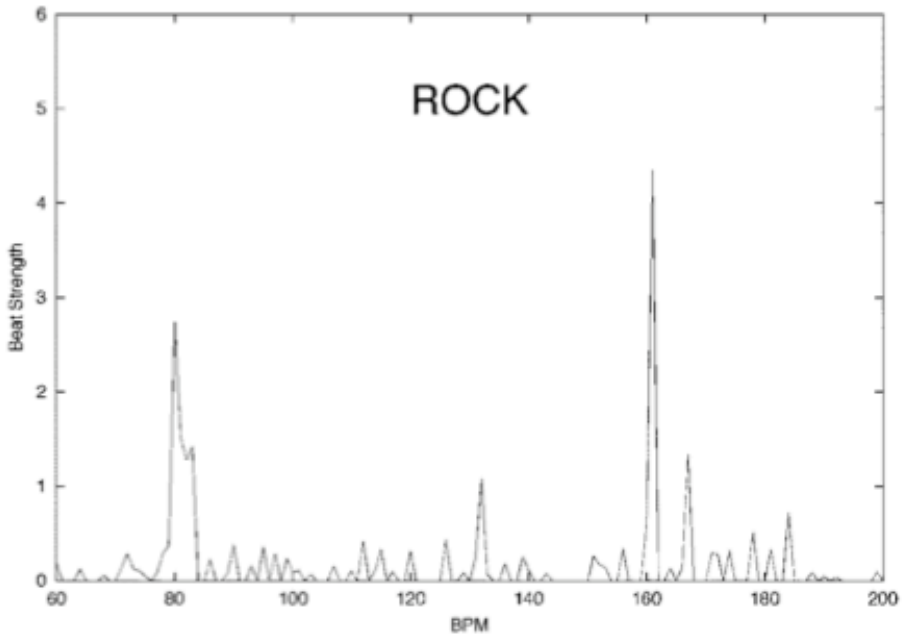
- Higher frequency resolution at high frequencies
- Higher time frequency at low frequencies

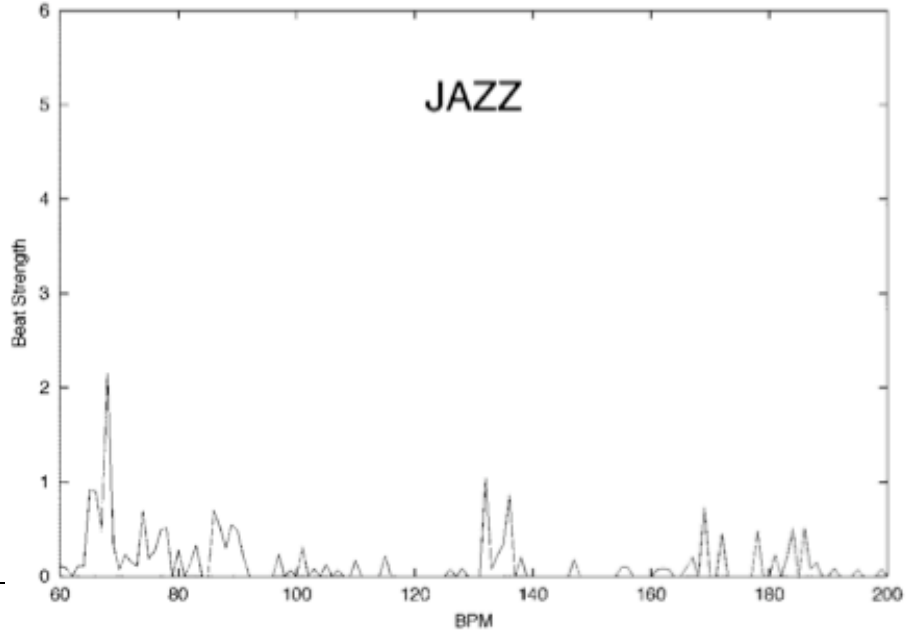


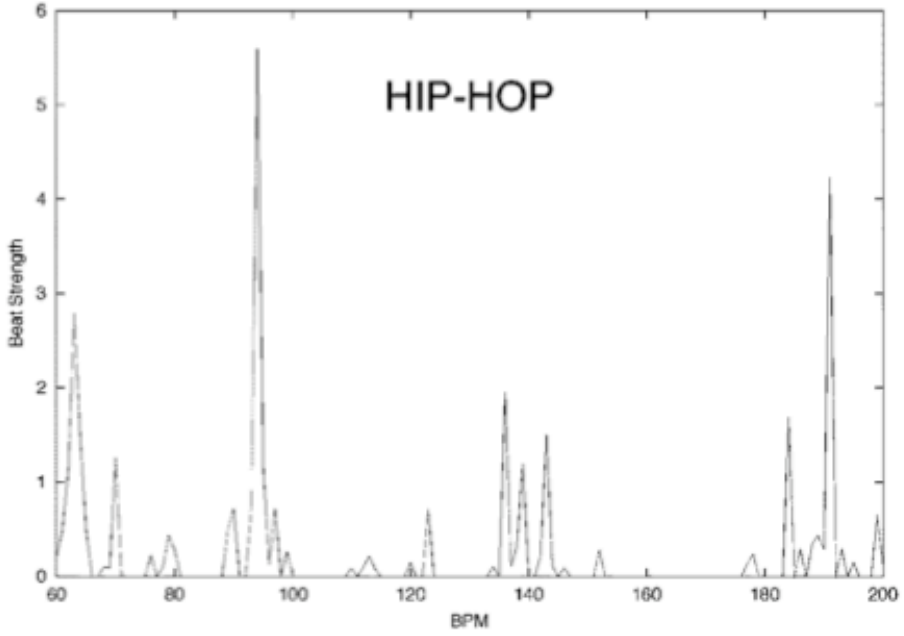
Feature Extraction

- All these transforms help us understand how the frequencies changes over time
- Features extraction:
  - Mel-frequency cepstral coefficients (MFCCs)
    - Attempt to mimic human ear
  - Surface features (texture, timbre, instrumentation)
    - Capture frequency statistics of STFT
  - Rhythm features (i.e the “beat”)
    - Characteristics of low-frequency wavelets


CLASSICAL
Beat Strength
BPM
0
1
2
3
4
5
6
60
80
100
120
140
160
180
200


ROCK
Beat Strength
BPM
0
1
2
3
4
5
6
60
80
100
120
140
160
180
200


JAZZ
Beat Strength
BPM
0
1
2
3
4
5
6
60
80
100
120
140
160
180
200


HIP-HOP
Beat Strength
BPM
0
1
2
3
4
5
6
60
80
100
120
140
160
180
200

Music Classification

- Data
  - Audio collected from radio, CDs and Web
    - Genres: classic, country, hiphop, jazz, rock
    - Speech vs. music
    - 4-types of classical music
  - 50 samples for each class, 30 sec. long
  - Task is to predict the genre of the clip
- Approach
  - Extract features
  - Learn genre classifier

General Results

	Music vs. Speech	Genres	Classical
Random	50%	16%	25%
Classifier	86%	62%	76%

Results: Musical Genres

	Classic	Country	Disco	Hiphop	Jazz	Rock
Classic	86	2	0	4	18	1
Country	1	57	5	1	12	13
Disco	0	6	55	4	0	5
Hiphop	0	15	28	90	4	18
Jazz	7	1	0	0	37	12
Rock	6	19	11	0	27	48

Pseudo-confusion matrix

Results: Classical

	Choral	Orchestral	Piano	String
Choral	99	10	16	12
Orchestral	0	53	2	5
Piano	1	20	75	3
String	0	17	7	80

Confusion matrix

Google Books

Thanks

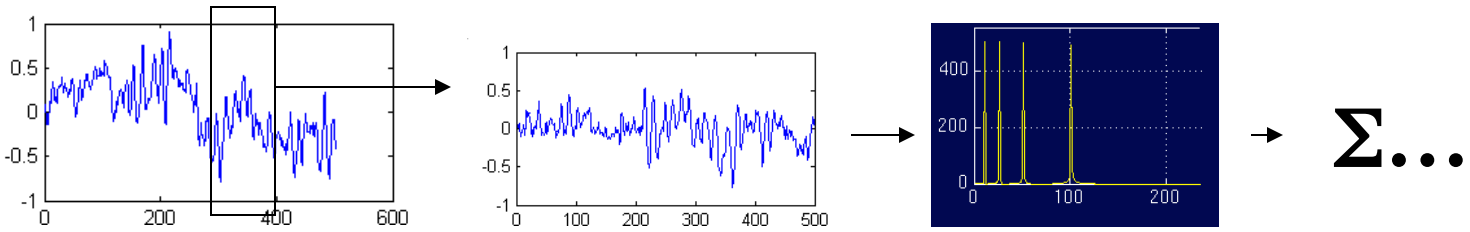
- Robi Polikar for his old tutorial (http://www.public.iastate.edu/~rpolikar/WAVELETS/WTtutorial.html)

Musical surface features

- What we'd like to do:
  - Represents characteristics of music
    - Texture
    - Pitch
    - Timbre
    - Instrumentation
- We need to quantify these things
  - Statistics that describe frequency distribution
    - Average frequency
    - Shape of the distribution
    - Number zero Crossings
  - Rhythm features

Calculating Surface Features

Signal → **Divide into windows** → **FFT over window** → **Calculate feature for window** → **Calculate mean and std. dev. over windows**



Surface Features

- Centroid: Measures spectral brightness

$$C = \frac{\sum_{f=1}^{N} f * M[f]}{\sum_{f=1}^{N} M[f]}$$

- Rolloff: Spectral Shape

R such that: $\sum_{f=1}^{R} M[f] = 0.85 * \sum_{f=1}^{N} M[f]$

M[f] = magnitude of FFT at frequency bin f over N bins

More surface features

- Flux: Spectral change

$$F = \left\| M[f] - M_p[f] \right\|$$

Where, $M_p[f]$ is M[f] of the previous window

- Zero Crossings: Noise in signal
- Low Energy: Percentage of windows that have energy less than average

Rhythm Features

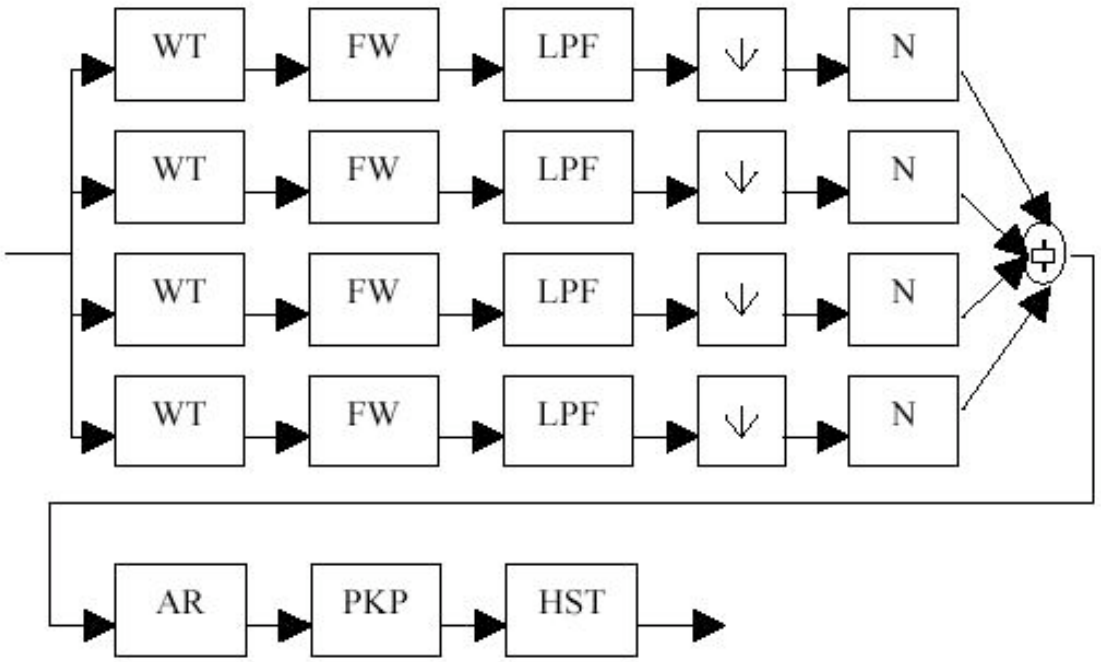
Wavelet Transform

Full Wave Rectification

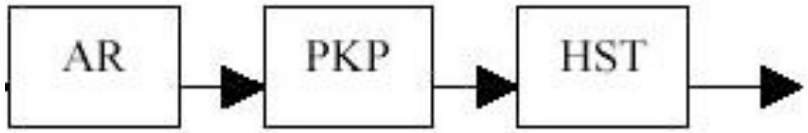
Low Pass Filtering

Downsampling

Normalize



Rhythm Features cont.



Autocorrelation – The cross-correlation of a signal with itself (i.e. portions of a signal with it's neighbors)

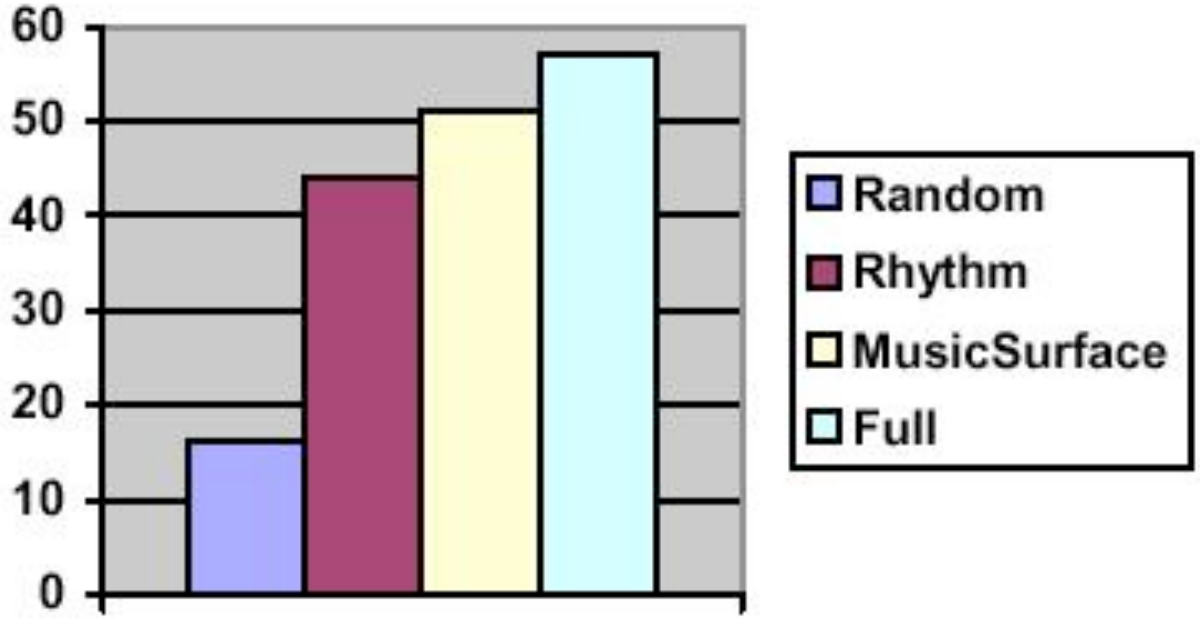
Take first 5 peaks

Histogram over windows of the signal

Actual Rhythm Features

- Using the "beat" histogram…
  - Period0 - Period in bpm of first peak
  - Amplitude0 - First peak divided by sum of amplitude
  - RatioPeriod1 - Ratio of periodicity of first peak to second peak
  - Amplitude1- Second peak divided by sum of amplitudes
  - RatioPeriod2, Amplitude2, RatioPeriod3, Amplitude3

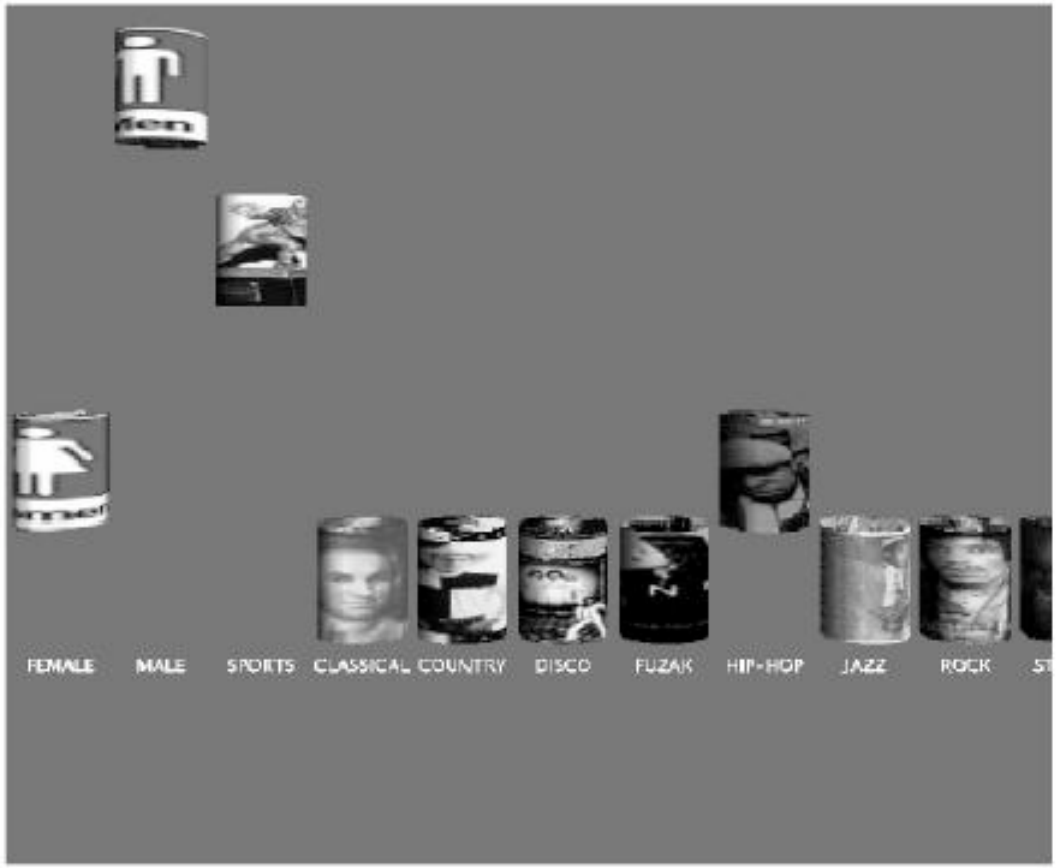
Analysis of Features



GUI for Audio Classification

- Genre Gram
  - Graphically present classification results
  - Results change in real time based on confidence
  - Texture mapped based on category
- Genre Space
  - Plots sound collections in 3-D space
  - PCA to reduce dimensionality
  - Rotate and interact with space

Genre Gram


FEMALE
MALE
SPORTS
CLASSICAL
COUNTRY
DISCO
FUZAK
HIP-HOP
JAZZ
ROCK

Genre Space