Analyses Using the Million Song Database

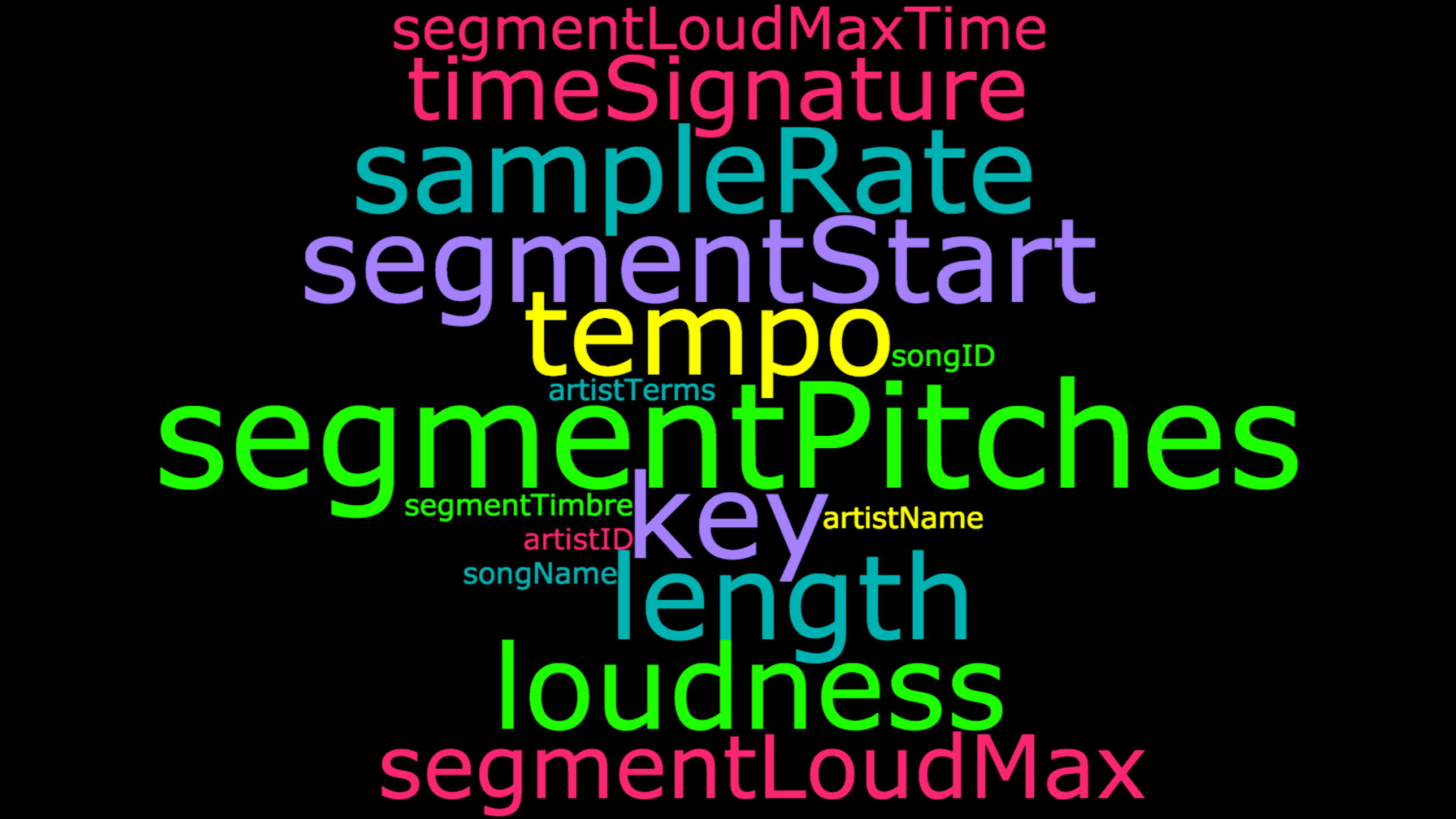
Wanlin Ji • Rodrigo Valdés • Erin M. Ochoa
CS 123 • 2017 Spring

Goals

- ☐ Implement big-data techniques
  - ☐ Multiprocessing
  - ☐ MPI
  - ☐ MapReduce

Goals

- ☐ Explore the dataset
  - ☐ Distribution of pairwise similarities
  - ☐ Ten most eclectic songs
  - ☐ Ten most formulaic songs
  - ☐ Two most similar songs
  - ☐ Two least similar songs


segmentLoudMaxTime
timeSignature
sampleRate
segmentStart
tempo
songID
artistTerms
segmentPitches
segmentTimbre
key
artistName
artistID
songName
length
loudness
segmentLoudMax

Pairwise Comparisons

Songs	Comparisons Needed
10	45
100	4,950
1,000	499,500
10,000	49,995,000
100,000	4,999,950,000
1,000,000	499,999,500,000

Methodological Overview

- ☐ Store selected songs in pickles
  - ☐ Administrative & musical pickles
- ☐ Pass pairs of pickles to worker nodes
  - ☐ Pairwise comparisons between songs
  - ☐ Report results to master node
- ☐ Compile results
- ☐ Run MapReduce scripts on results

Extracting Data

- HDF5 files: tables package (Python)
- Access administrative details (song name & ID and artist name, ID, & genre tags)
  - Store in dictionary
- Access musical details (sampleRate, length, key, loudness, tempo, timeSignature, segLoudMax, segLoudMaxTime, segPitches, segStart, segTimbre)
  - Store in second dictionary
- Return tuple: songIDX, adminDicto, musicDicto

Pairwise Comparison

- Compare lengths
- Pad arrays in shorter song
- Vectorize each song
- Compare vectors: cosine similarity
  - [-1,1]
    - →0←: Less similar
    - ←0→: More similar

Preliminary Results

- Sample of 100,000 pairwise comparisons
- Subset of musical features:
  - segLoudMax
  - segLoudMaxTime
  - segPitches
  - segStart
  - segTimbre

Preliminary Results

(using a sample of 100,000 comparisons and a subset of musical features)

Meta-distribution of pairwise distances	
Maximum	0.6847
Mean	0.2841
Median	0.2744
Minimum	-0.0565
Standard Deviation	0.1519
Skew	0.2505

Preliminary Results

(using a sample of 100,000 comparisons and a subset of musical features)

- The most eclectic song:
  - 9720: Hinge, “Pray The I Miss” (sic)
  - Mean pairwise distance of 0.01248307392
- The most formulaic song:
  - 9963: Owsley, “Class Clown/Good Old Days (Reprise)”
  - Mean pairwise distance of 0.41185859225

Preliminary Results

(using a sample of 100,000 comparisons and a subset of musical features)

- The two least similar songs: 1.546504982019e-05
  - 9734: Van Halen, “Why Can't This Be Love (Remastered Version)”
  - 9943: X-Ecutioners featuring Big Pun, “The Drama (intro)”
- The two most similar songs: 0.900037116659938
  - 9725: Livio Minafra, “Campane”
  - 9925: Semprini, “Rhapsody In Blue (2003 Digital Remaster)”

Challenges

- Time!
- Defining goals
- Accessing data from HDF5 files
- Distance metrics
  - Distance correlation or correlation distance?
    - Values > 1
    - Nebulous support pages
  - Increased computation time with cosine similarity
- Large number of comparisons necessary