Musical query-by-description as a multi-class learning problem

Brian Whitman

MIT Media Lab

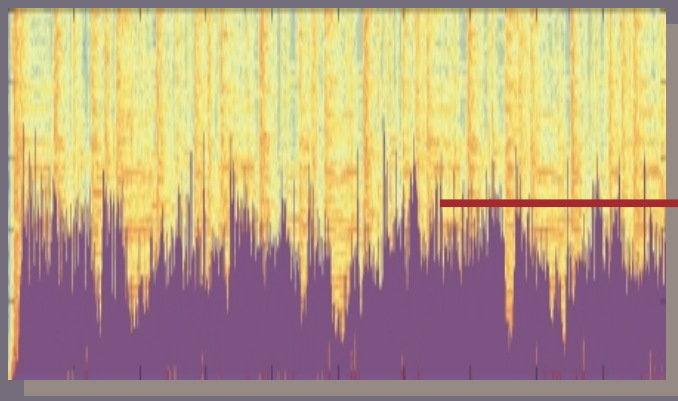
Music, Mind and Machine Group
(formerly Machine Listening)

Ryan Rifkin

(MIT AI Lab, CBCL) / Honda R&D Americas

MMSP -- December 10th, 2002

Music intelligence

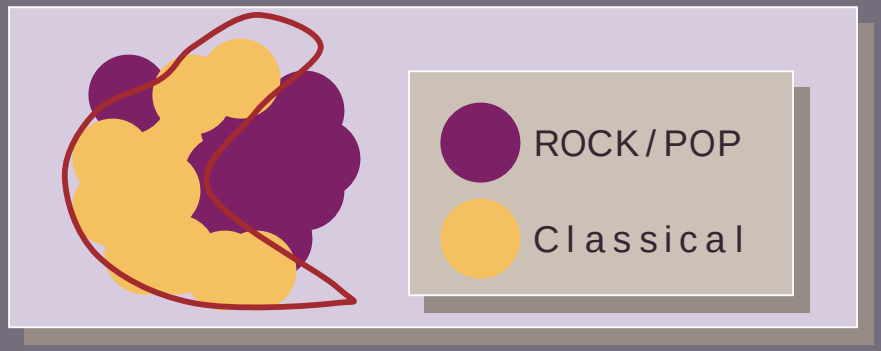


Structure | Recommendation

Genre / Style ID | Artist ID

Song similarity | Synthesis

- Extracting salience from a signal
- Learning is features and regression



Better understanding through semantics

"Loud college rock with electronics."

Structure

Recommendation

Genre / Style ID

Artist ID

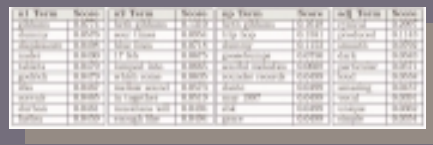
Song similarity

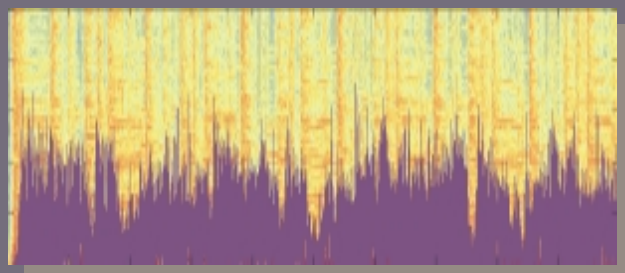
Synthesis

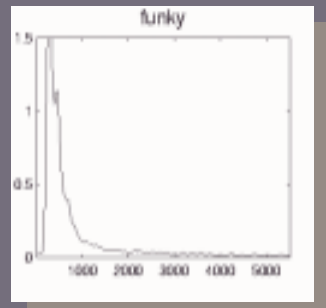
- What if a system learned the meaning of the underlying perception?
- How can we get context to computationally influence understanding?

Using context to learn descriptions of perception

- “Grounding” meanings (Harnad 1990): defining terms by linking them to the ‘outside world’





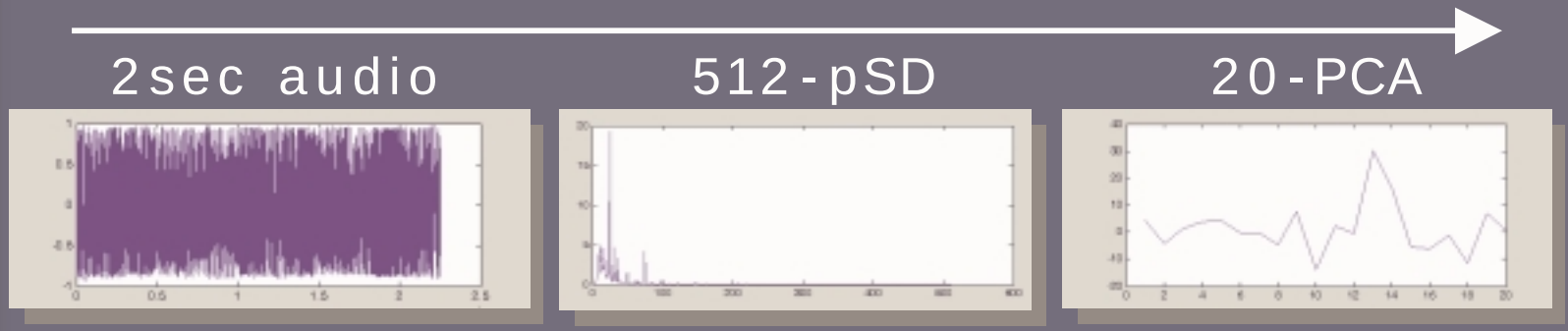


Query-by-description as evaluation case

- QBD: “Play me something loud with an electronic beat.”
- With what probability can we accurately describe music?
- Training: We play the computer songs by a bunch of artists, and have it read about the artists on the Internet.
- Testing: We play the computer more songs by different artists and see how well it can describe it.

The audio data

- Large set of music audio
  - Minnowmatch testbed (1000 albums)
  - Most popular on OpenNap August 2001
  - 51 artists randomly chosen, 5 songs each
- Each 2sec frame an observation:
  - TD→PSD→PCA to 20 dimensions



“Community metadata”

- Whitman / Lawrence (ICMC2002)
- Internet-mined description of music
- Embed description as kernel space
- Community-derived meaning
- Time-aware!

n1 Term	Score
gibbons	0.0774
dummy	0.0576
displeasure	0.0498
nader	0.0490
tablets	0.0479
godrich	0.0479
irks	0.0467
corvair	0.0465
durban	0.0461
farfisa	0.0459

n2 Term	Score
beth gibbons	0.1310
sour times	0.0954
blue lines	0.0718
17 feb	0.0675
lumped into	0.0665
which come	0.0635
mellow sound	0.0573
in together	0.0519
musicians will	0.0494
enough like	0.0494

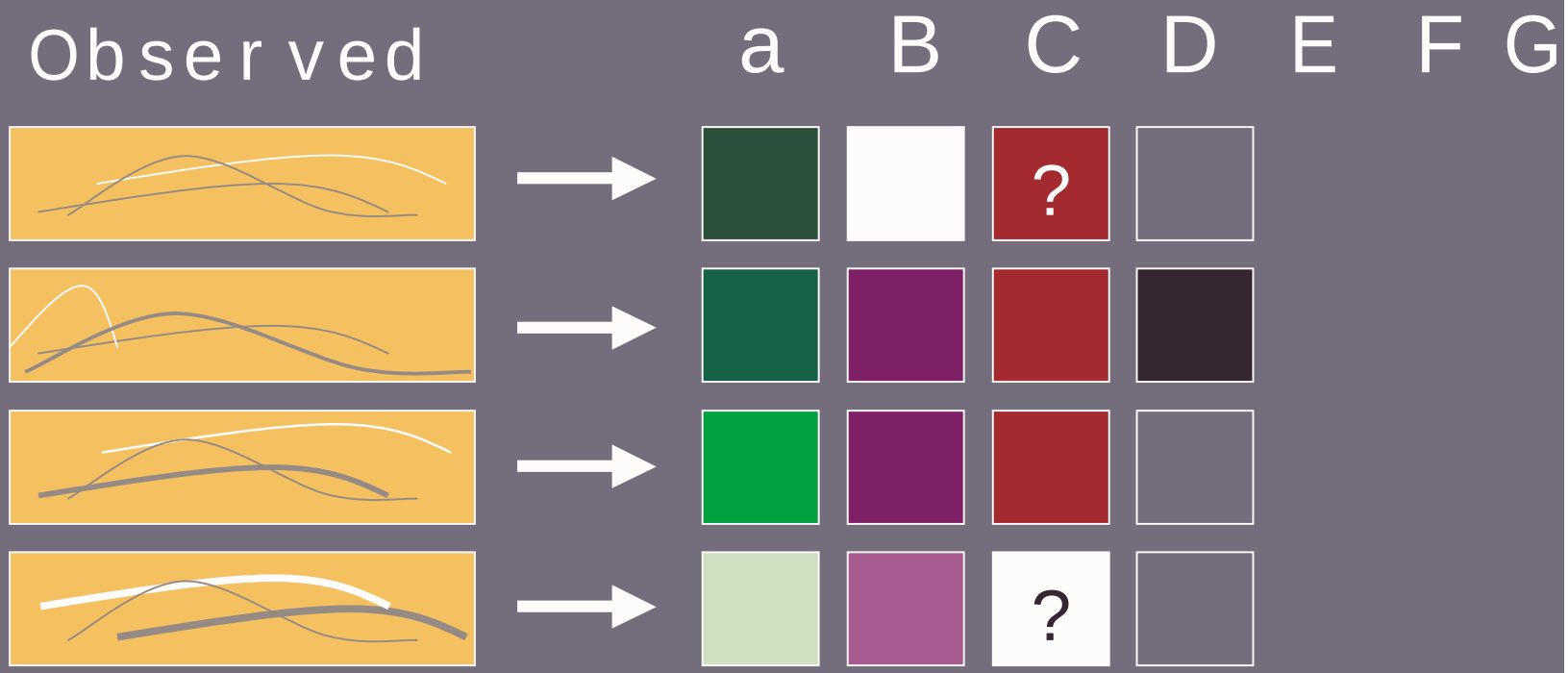
np Term	Score
beth gibbons	0.1648
trip hop	0.1581
dummy	0.1153
goosebumps	0.0756
soulful melodies	0.0608
rounder records	0.0499
dante	0.0499
may 1997	0.0499
sbk	0.0499
grace	0.0499

adj Term	Score
cynical	0.2997
produced	0.1143
smooth	0.0792
dark	0.0583
particular	0.0571
loud	0.0558
amazing	0.0457
vocal	0.0391
unique	0.0362
simple	0.0354

Learning formalization

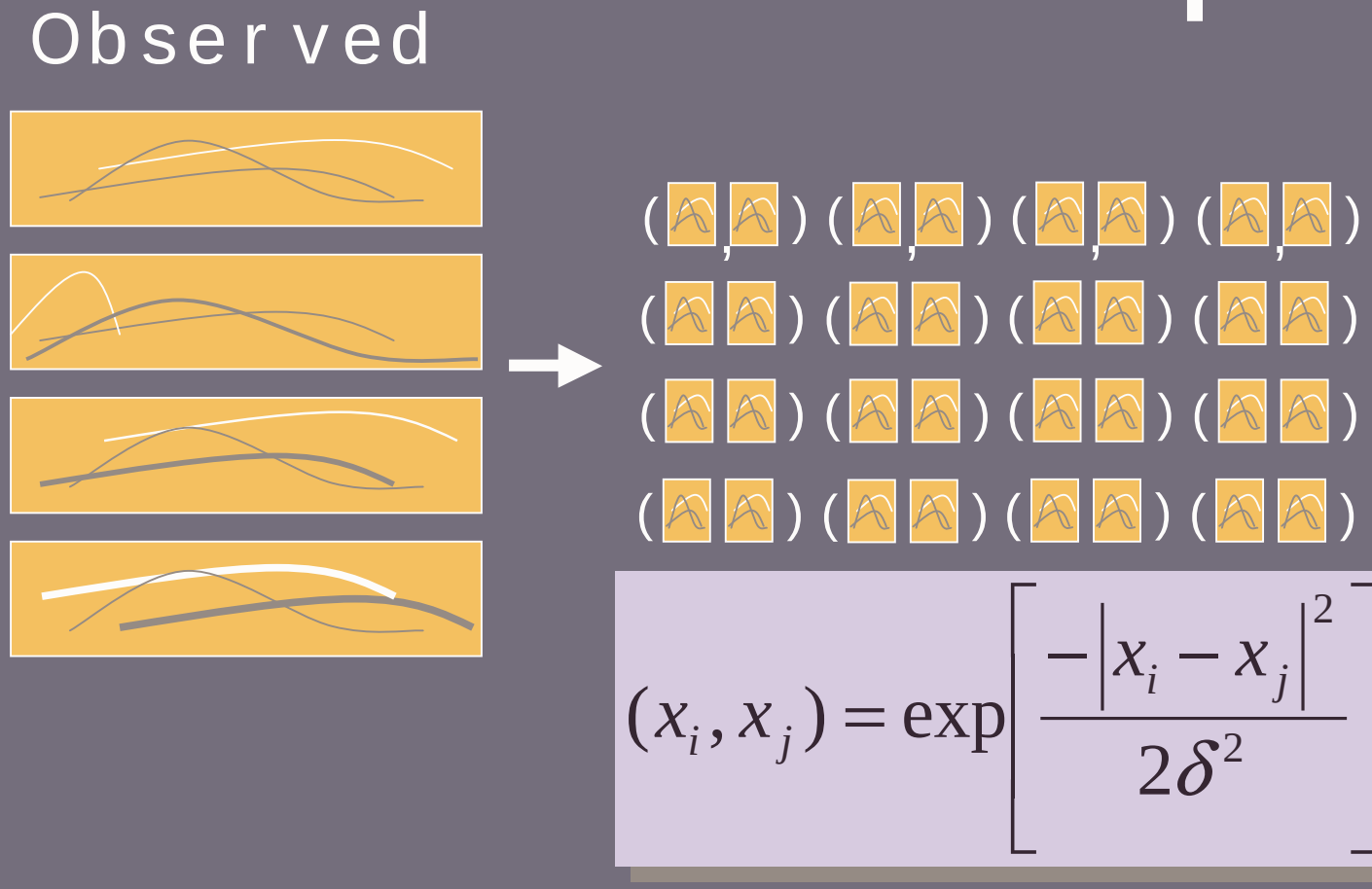
- Learn relation between audio and naturally encountered description
- Can’t trust target class!
  - Opinion
  - Counterfactuals
  - Wrong artist
  - Not musical
- 200,000 possible terms (output classes!)
  - (For this experiment we limit it to adjectives)

Severe multi-class problem



1. Incorrect ground truth
2. Bias
3. Large number of output classes

Kernel space



$$(x_i, x_j) = \exp\left[\frac{-\left|x_i - x_j\right|^2}{2\delta^2}\right]$$

- Distance function represents data
  - (gaussian works well for audio)

Regularized least-squares classification (RLSC)

- (Rifkin 2002)

$$(K + \frac{I}{C})\mathbf{c_t} = \mathbf{y_t} \quad \rightarrow \quad \mathbf{c_t} = (K + \frac{I}{C})^{-1}\mathbf{y_t}$$

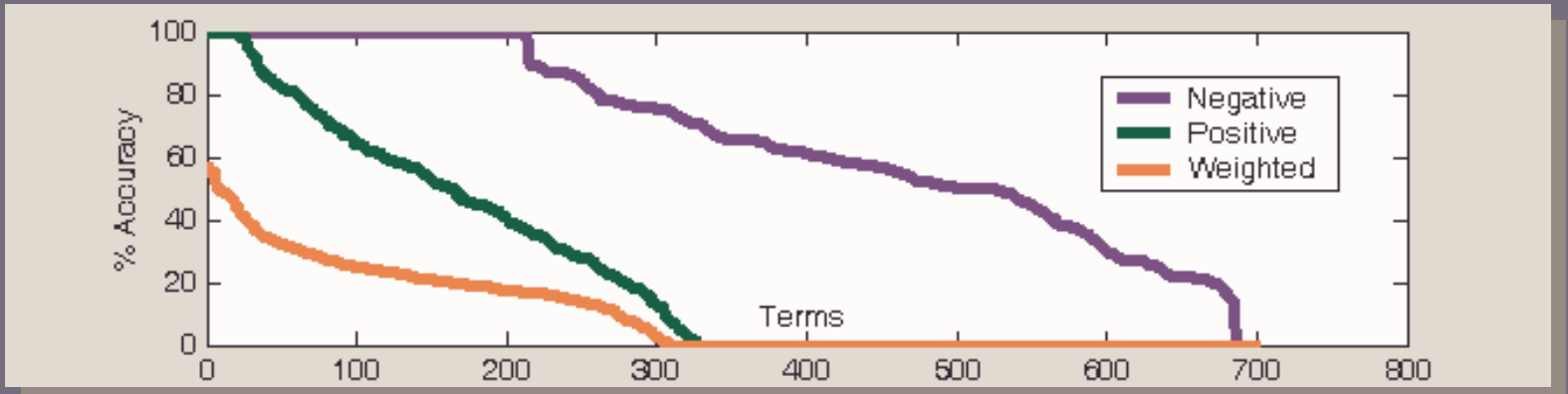
$\mathbf{c_t}$ = machine for class t
$\mathbf{y_t}$ = truth vector for class t
C = regularization constant (10)

Computational benefit

- (*n* classes, *d* dimensions of input, *l* examples)
- Store 2 *l*x*l* gram matrices in memory or train *n* SVMs?
- SVM *(d*,*n*,*l* dependent to time, *d*,*l* dependent to memory)*:*
  - 250hrs for training 1/10th of the SVMs
  - Max 16MB cache needed
- RLSC *(d*,*l* dependent to time, *l* dependent to memory*):*
  - 1.5hrs for precomputing & inverting
  - 256MB of space for both gram matrices

QBD evaluation results

- Compute ‘weighted precision’ P(*p*)P(*n*)



- Usual IR evals worthless because of incredibly low baseline, mistrust of data, bias
- What is important are deltas!

Per-term accuracy

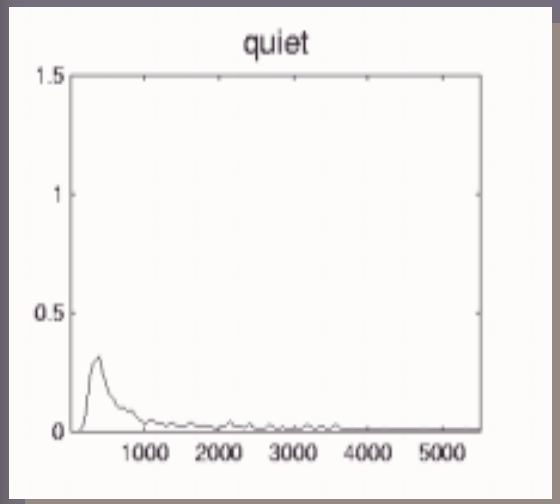
Good terms		Bad terms	
Electronic	33%	Annoying	0%
Digital	29%	Dangerous	0%
Gloomy	29%	Fictional	0%
Unplugged	30%	Magnetic	0%
Acoustic	23%	Pretentious	1%
Dark	17%	Gator	0%
Female	32%	Breaky	0%
Romantic	23%	Sexy	1%
Vocal	18%	Wicked	0%
Happy	13%	Lyrical	0%
Classical	27%	Worldwide	2%

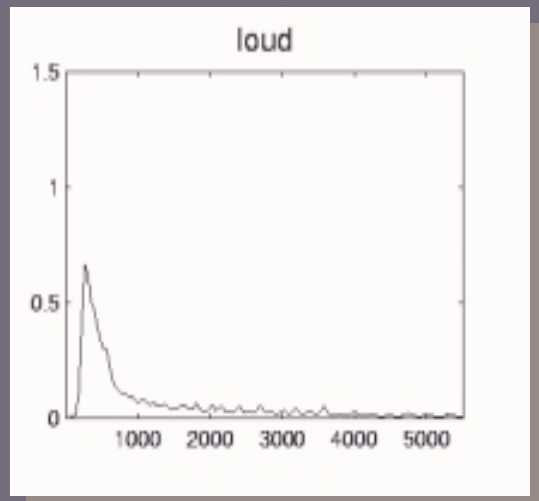
“Baseline” = 0.14%

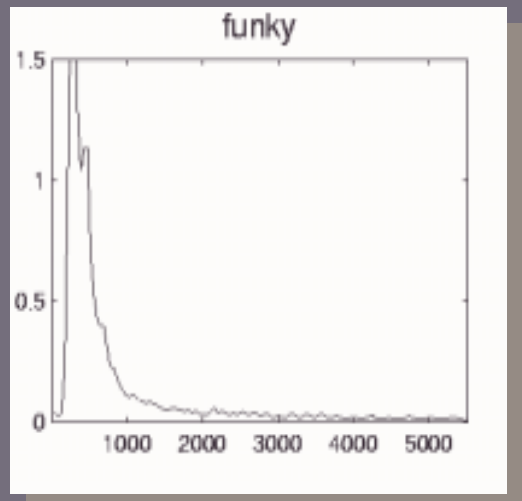
- Good term set as restricted grammar?

Synthesizing opinion

- “What does loud mean?”







- Weighted mean of labeled observations
- (Smarter: eigenfilters, etc.)

What's next

- Human evaluation
  - cf. Reiger/Carlson
  - "can we trust the internet for community meaning?"
- Time-aware features
- Learning parameter spaces
  - "fast .. slow" "loud .. soft"
  - Knobs for retrieval/synthesis
- Bootstrapping terms from expert
- Hierarchy learning

Wrap-up

- On its own: QBD as multimedia interface
- Thinking ahead: multimedia understanding from semantic attachment
- Thanks: Steve Lawrence, Dan Ellis (Columbia), MMM group!