CSC 390
Topics in Artificial Intelligence

"Unsupervised Machine Learning"

Fall 2016

Prof. Sara Mathieson
Smith College

Outline: 11/3

- Today:
  - **Lujun** 10:30am
  - **Farida** 10:45am
  - **Sharon** 11:00am
  - **Hera** 11:15am
  - **Jessica Tran** 11:30am

- Office Hours today: 4-5pm, Ford 355

- Presenters:
  - Speak loudly
  - I will give you a 2 min warning after 10 minutes

- Audience:
  - Give presenters your full attention and ask questions
  - I will ask questions too

Smart Reply: Automated Response

...

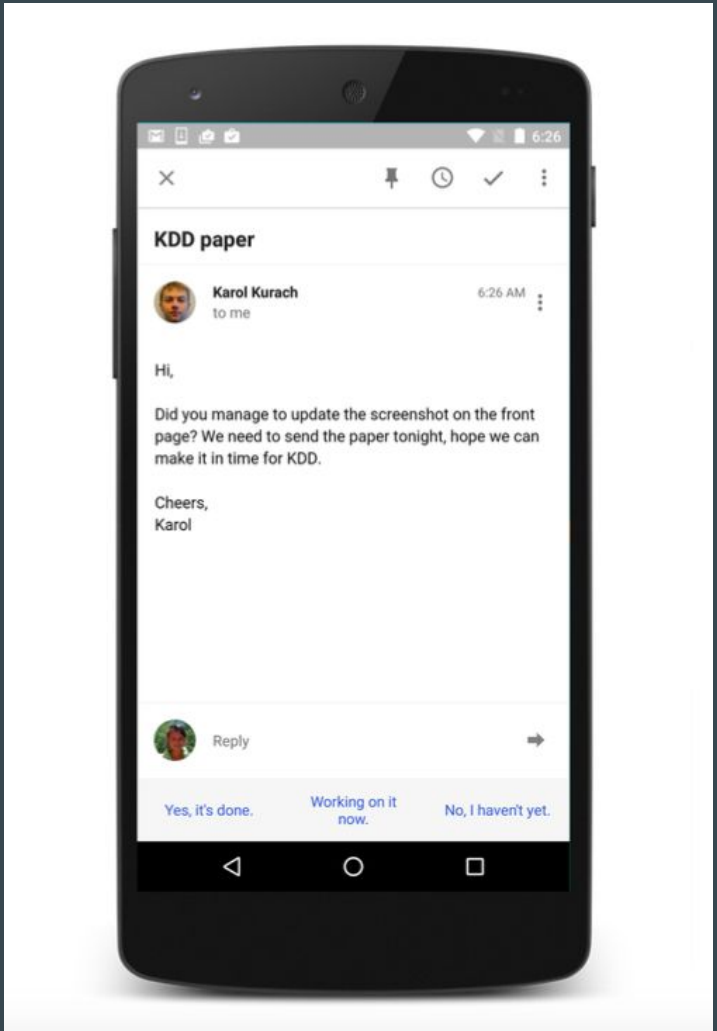
By Anjuli Kannan, ; Karol Kurach*, Google; Sujith Ravi, Google; Tobias Kaufmann, Google, Inc.; Andrew Tomkins, ; Balint Miklos, Google, Inc.; Greg Corrado, ; László Lukács, ; Marina Ganea, ; Peter Young, ; Vivek Ramavajjala

Today's presenter: Lujun Jian

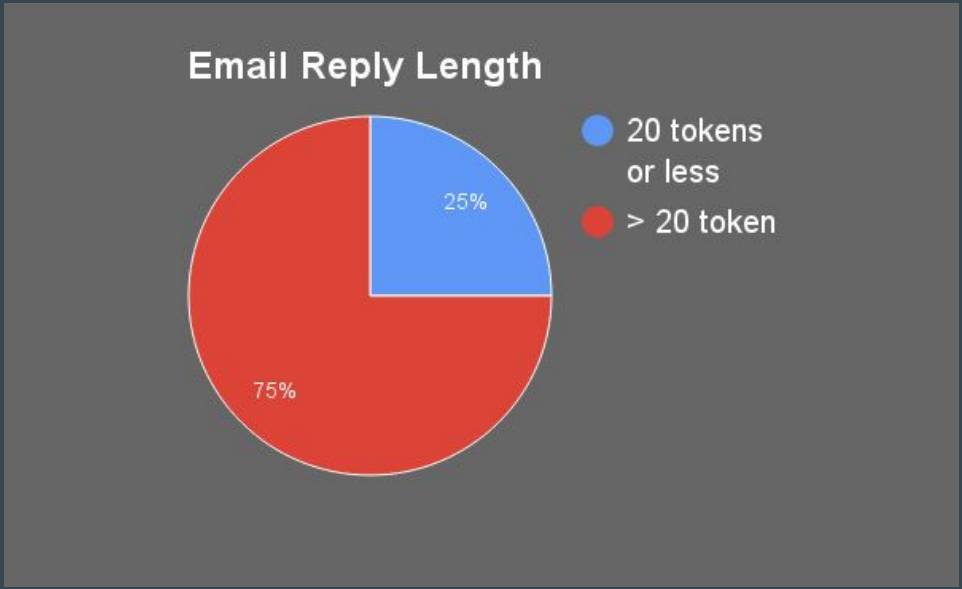
Motivation

Currently, more and more emails and messages are received and responded on the mobile device. To help user with the message overload, technologists want to assist user with automatically generated short message response.

This paper proposes and investigates into the mechanism/ methods behind Smart Reply which is currently used use Gmail mobile app.



Background



Several million email-reply pairs shows that around 25% of email replies have 20 tokens or less. It shows potential needs and possibility for creating auto-reply suggestions.

This Smart Reply system is used in Inbox by Gmail and 10% of mobile replies in Inbox are now composed with assistance from the Smart Reply system.

Key Terminology

Tokenization - Subject and message body are broken into words and punctuation marks.

Normalization - Infrequent words and entities like personal names, URLs, email addresses etc. are replaced by special tokens.

Long Short Term Memory (LSTM) - a special kind of Recurrent Neural Network, capable of learning long-term dependencies.

Methods

1. Triggering model - Is there is a need to generate responses?
2. Response Set Generation - only select responses from appropriate response space.
3. Response Selection (LSTM) - predict most likely responses.
4. Diversity Selection - make sure the diversity of the suggestions.

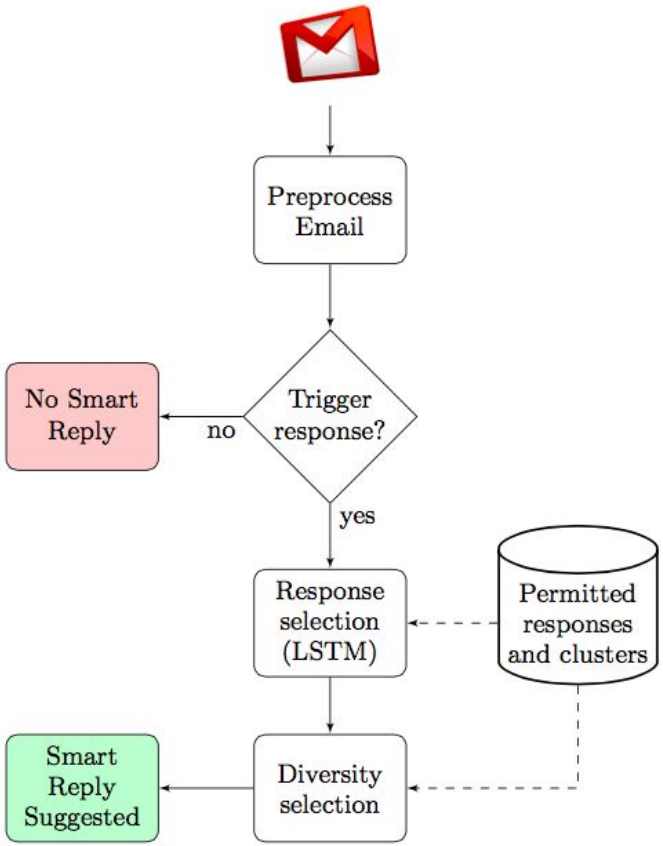


Figure 2: Life of a message. The figure presents the overview of inference.

Triggering model

Filter out input message for which short replies are not appropriate:

- Open-ended questions
- Contains sensitive topics
- Promotional message
- Auto-generated updates

Feedforward neural network produces a probability score for the input message.

Roughly 11% message will triggering next step in the current system.

Response Set Generation

To generate a structured response set that effectively captures various intents conveyed by people in natural language conversations.

1. Standardize Email Response
   a. “Thanks for your kind update” and “Thank you for your updating” are the same.
   b. Ignore modifiers, parse with a dependency parser, and generate a new standard representation.
2. Semantic Intent Clustering
   a. Ex. “lol”, “haha”, “that’s funny” → “funny” cluster | “thanks!”, “I’m grateful” → “thank you” cluster
   b. Graph Construction
      i. Frequent response messages (V_R nodes) ← lexical features (V_F nodes)
      ii. Inter-message relations
   c. Semi-Supervised Learning with Cluster Validation

I'm grateful.

Thanks.

Thank you.

"Thank" intent

Sounds good.

Works for me.

Sure.

"date" intent

How about today?

Tomorrow.

lol.

That's funny.

Haha.

How about Friday?

"time" intent

Thursday evening.

This afternoon.

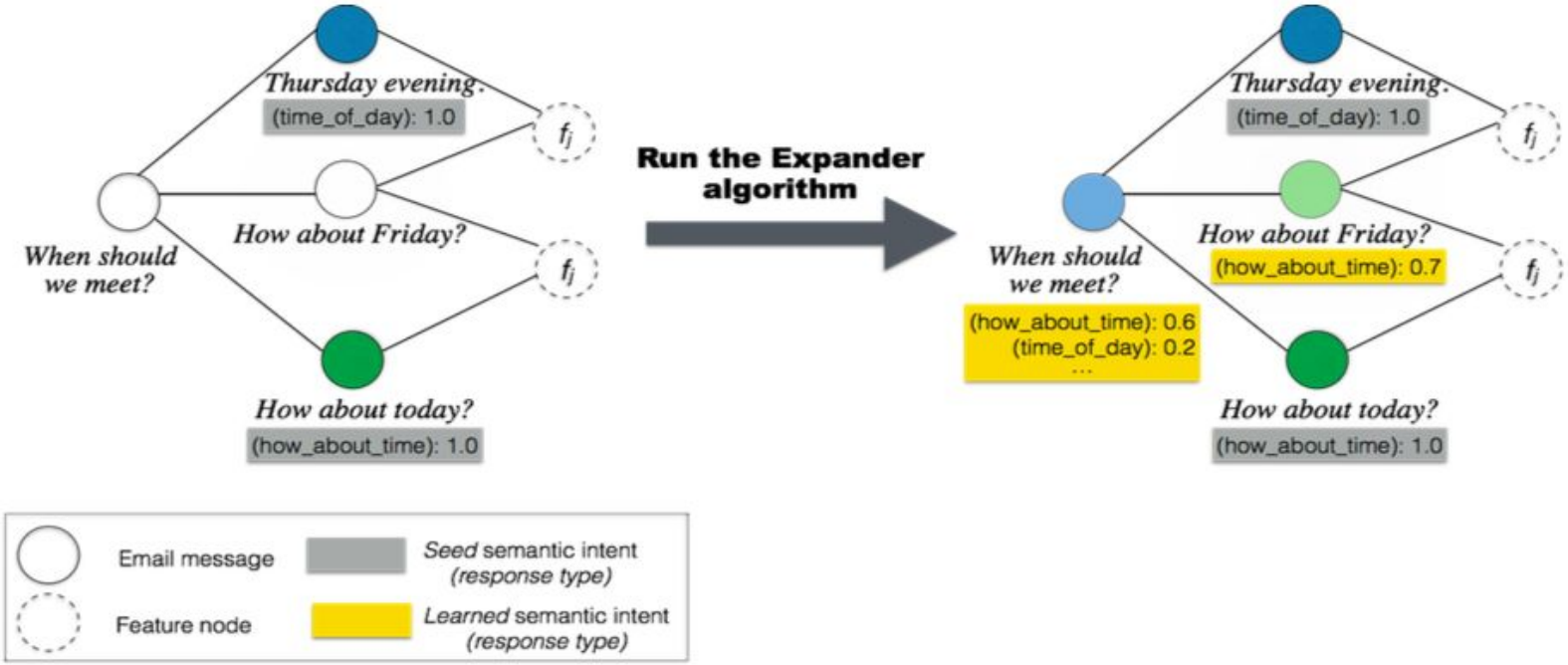


Figure 4: Semantic clustering of response messages.

I’m grateful.

Thanks.

Thank you.

“Thank” intent

Sounds good.

Works for me.

Sure.

“date” intent

lol.

How about today?

Tomorrow.

That’s funny.

How about Friday?

Haha.

“time” intent

Thursday evening.

This afternoon.

I'm grateful.

Thanks.

Thank you.

"Thank" intent

Sounds good.

Works for me.

"yes" intent

Sure.

"date" intent

How about today?

Tomorrow.

How about Friday?

"time" intent

Thursday evening.

This afternoon.

lol.

That's funny.

Haha.

"fun" intent

I'm grateful.

Thanks.

Thank you.

"Thank" intent

Sounds good.

Works for me.

Sure.

"yes" intent

"date" intent

How about today?

Tomorrow.

How about Friday?

"time" intent

Thursday evening.

This afternoon.

lol.

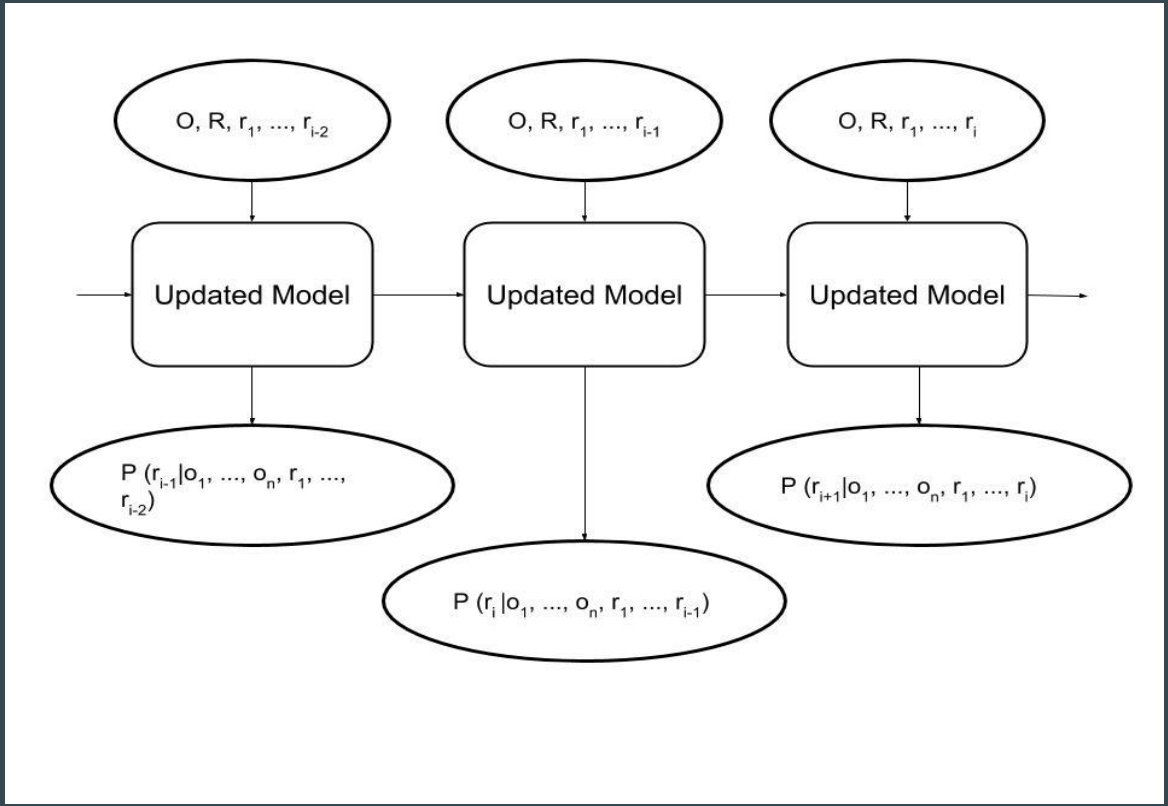
That's funny.

Haha.

"fun" intent

Response Selection / LSTM scoring

Give a score to each possible response based on the conditional probability distribution of the sequence of response tokens given the input.



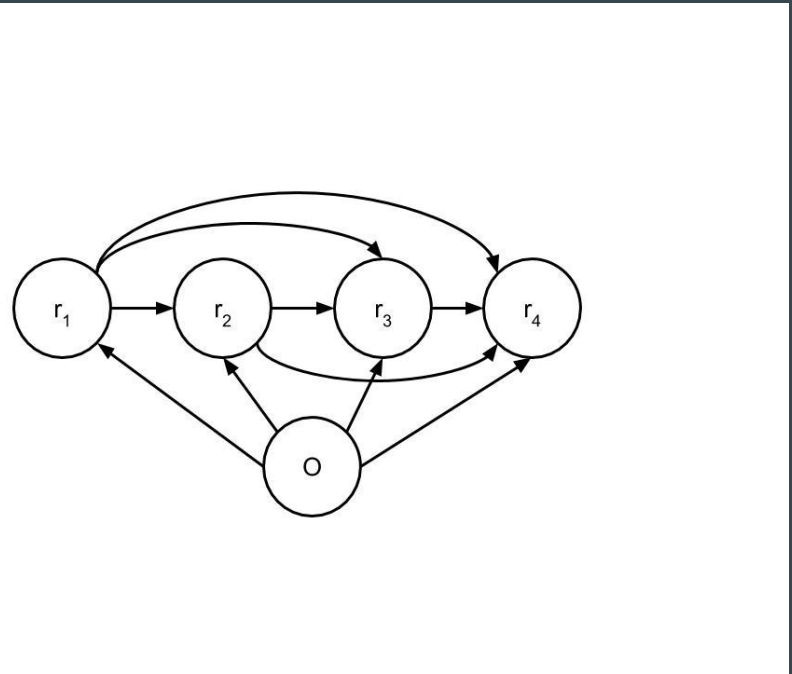
LSTM scoring - Model

$$P(r_1, ..., r_m | o_1, ..., o_n) = \prod_{i=1}^{m} P(r_i | o_1, ..., o_n, r_1, ..., r_{i-1})$$

R - the set of all possible responses;

r - response token

o - original message token



Diversity Selection

To make sure the delivered suggestions have sufficient variety.

- Omitting Redundant Responses
  - Iterated over the responses in the order of decreasing score
  - Add to the list if the intent has not yet been covered
  - Resulting list contains only the highest scored representatives of each intent
- Enforcing Negatives and Positives
  - If the top two response contain at least one positive response and none of the top three responses are negative, the third response will be replaced with a negative one.

Results

- Response Selection Results
  - Perplexity - a perplexity equal to k means that when the model predicts the next word, there on average k like candidates. (1 is the optimal).
    - Smart Reply LSTM is 17.0 v.s. an n-grams language model with Katz backoff which is 31.4
  - Precision@K - for a given value of K it is computed as the number of cases for which target response r was within the topK responses that were ranked by the model.

Model	Precision@10	Precision@20	MRR
Random	$5.58e-4$	$1.12e-3$	$3.64e-4$
Frequency	0.321	0.368	0.155
Multiclass-BOW	0.345	0.425	0.197
Smart Reply	0.483	0.579	0.267

Table 3: Response ranking

Results

- Usage

	Daily Count	Seen	Used
Unique Clusters	376	97.1%	83.2%
Unique Suggestions	12.9k	78%	31.9%

Table 4: Unique cluster/suggestions usage per day

Conclusion

- This paper addresses the method that
  - Ensure the quality of the response options by selecting them from a constructed response space.
  - Increase the total utility of the combination of suggestion by enforcing diversity.
- Inspiration to my final project
- KDD 2016

Thank you!

Q & A

KDD 2016 paper 1069

Brain Tumor Detection

using Unsupervised Learning based Neural Network

—by Ms. Suchita Goswami & Mr. Lalit Kumar Bhaiya

Motivation

- **Goal:** automate brain image classification (ie. normal or abnormal)

- MRI Images present the alternative of tumor detection using an non-invasive procedure.

- Classification of brain MRI can be extremely useful to weed out normal patients, especially since manual readings can be quite labor intensive

Key Terminology

- **MRI:** Magnetic Resonance Imaging
- **ICA:** Independent Component Analysis --> a powerful technique used mainly on signals to separate independent sources that are linearly mixed with other signals.
- **SOM:** Self-Organizing Map --> artificial neural network based on competitive unsupervised learning

Methodology

A. Image Pre-processing
    1. Edge Detection
    2. Histogram Equalization
    3. Thresholding

B. Feature Recognition
  --> when data is too large to be processed, we transform it into a reduced representative set of features using ICA

C. Classification (SOM + K-means)

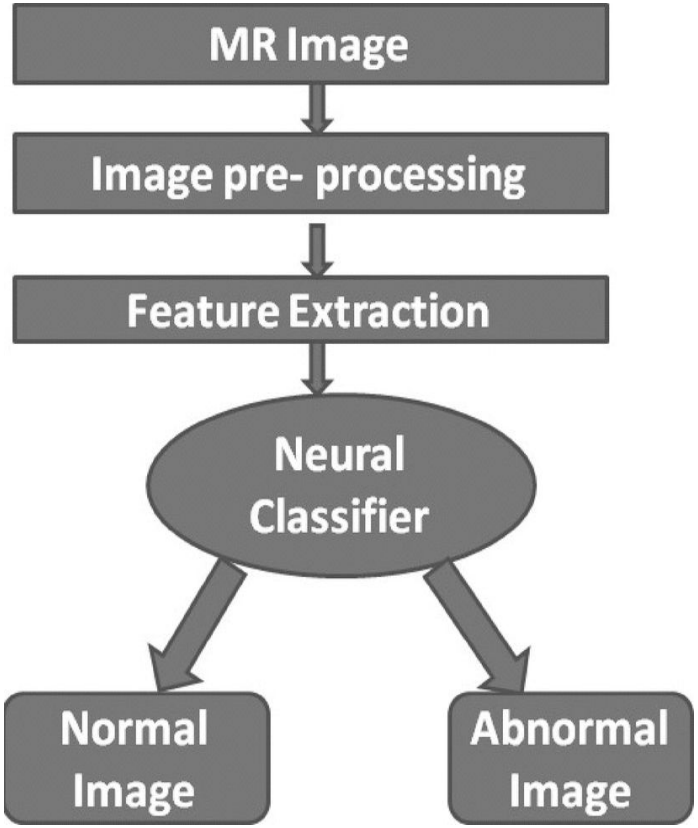


Fig 3.1. Block Diagram of proposed methodology

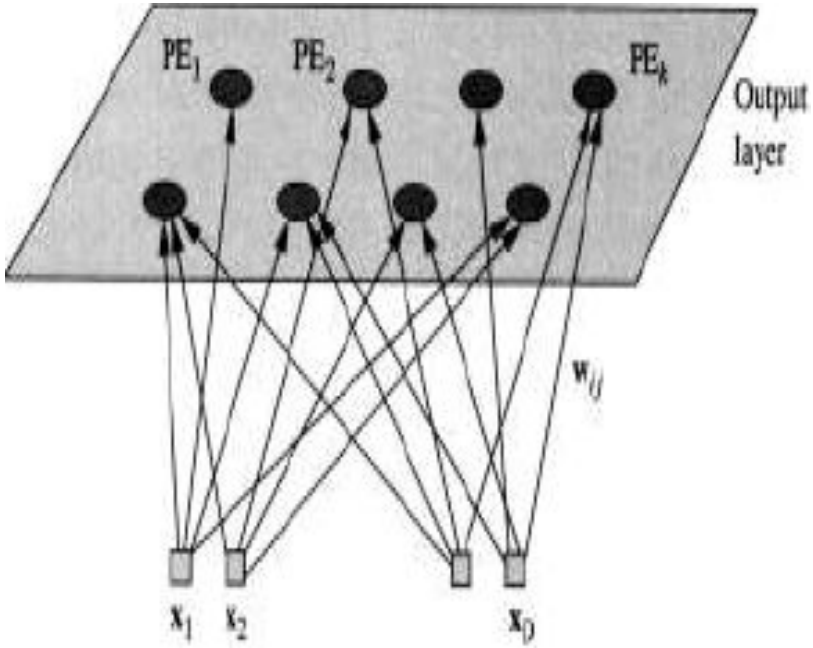


Fig 3.2: Architecture of a SOM with a 2-D output

Results

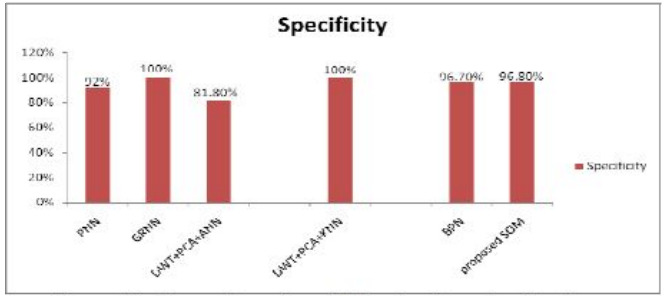


Figure 4.3: Comparison of specificityof various classification techniques

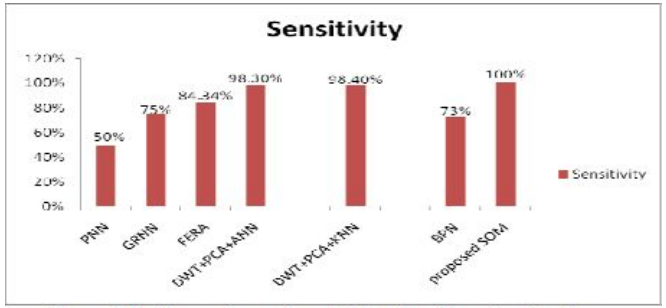


Figure 4.4: Comparison of sensitivity of various classification techniques

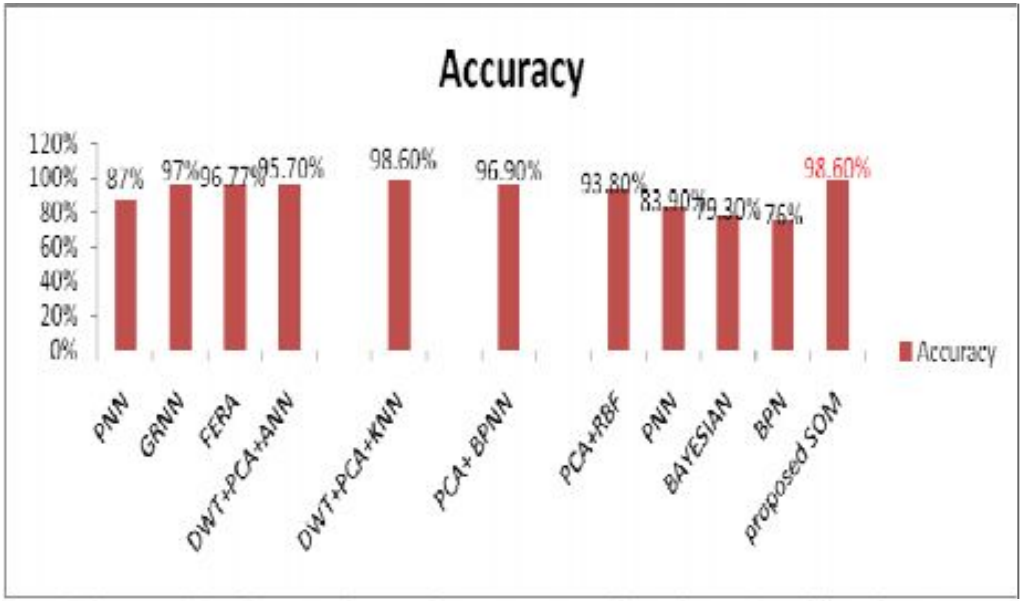


Figure 4.5: Comparison of s accuracy of various classification techniques

Conclusions

MRI Image Classification

--> 69/70 images were classified correctly

--> model was more accurate than other techniques

--> I'm interested in possibly pursuing a final project related to precision medicine

THANK YOU FOR LISTENING!

Any questions?

UNSUPERVISED DISCOVERY OF TEMPORAL STRUCTURE IN MUSIC

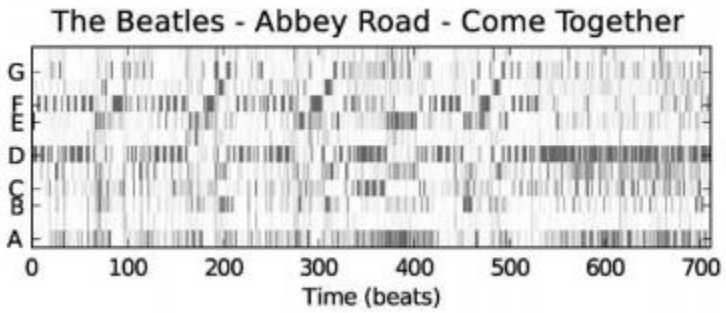
SHARON VIZCAINO

INTRODUCTION

- Repetition plays a fundamental role in music, and patterns in rhythm are necessary to understand and analyze music
- Discovering repetitive patterns is easy using symbolic representations of music (sheet music)
- Becomes a complicated problem when trying to derive those patterns from audio signals
- Novel approach for extraction and localization of repeated patterns
- Based on SI-PLCA, and uses sparse prior distributions to minimize the number of parameters specified in advance

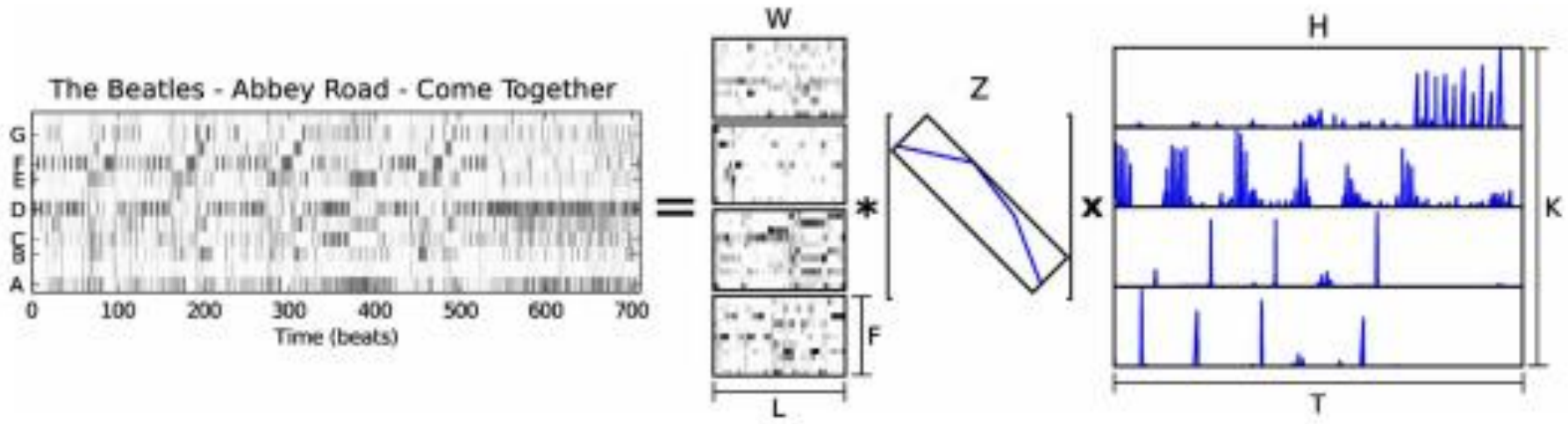
TERMS

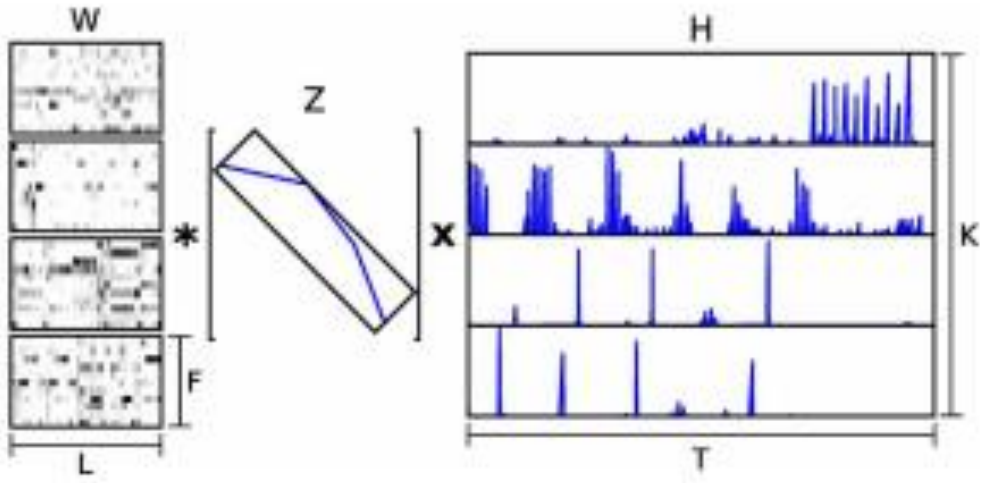
- **Chroma feature:** representation of music audio in which spectrum of notes is projected into 12 bins, each representing one of the distinct semitones in an octave
- **Time signature:** an indication of rhythm that shows the number of beats per measure
- **Transposition:** process of moving a collection of notes up or down in pitch by a constant interval



SI-PLCA

- Algorithm treats a musical recording as a set of short, repeated patterns
- Able to simultaneously estimate both the patterns and their repetitions throughout the song





$$V \approx WH$$

$$V \approx WZH = \sum_{k=0}^{K-1} z_k \mathbf{w}_k \mathbf{h}_k^T$$

$$V = P(f,t) \approx \sum_k P(k)P(f|k)P(t|k)$$

In which W represents basis vectors used repeatedly throughout V, and H represents a time-frequency decomposition of the audio signal, telling us when each basis vector is being activated over time

PLCA adds Z, an additional distribution over the set of bases (mixing weights) and normalizes parameters

K is the rank of the decomposition (the number of bases in W), 4 in example above

$$P(k) = z_k,\ P(f|k) = w_{kf},\ P(t|k) = h_{kt},\ f \in [0,F), t \in [0,T)$$

F = 12 pitch classes

T = beats

SPARSE PRIOR DISTRIBUTIONS

$$P(\mathbf{z}|\alpha_z) \propto \prod_k z_k^{\alpha_z - 1}, \quad \alpha_z \geq 0$$

- To learn the number of patterns K, it is set to a large value
- The sparse prior on z prunes out bases that don't contribute to the reconstruction of V
- Removes need to know K in advance
- To learn the pattern length L, a similar approach is used, which also sets L to an upper bound
- Adds c, which is the beat at which the prior becomes active, and m, the minimum pattern length

$$P\left(\mathbf{h}_k^T|\beta_h\right) \propto \exp\left(\beta_h \sum_t h_{kt} \log h_{kt}\right), \quad \beta_h \geq 0.$$

RIFF IDENTIFICATION

- Tested on songs with a single chord progression repeated throughout
- Also considers transposition, which, due to the chroma features, is a simple rotation of Wk and Hk
- K = 1 on these examples
- R = 12 to allow for all key modulations
- Basis length L = 40
- C = 10 and m = -0.0003 to identify underlying pattern length
- βh = 0.1 to encourage identification of correct patterns
- βr = 1 to enforce that only one key is active at a time

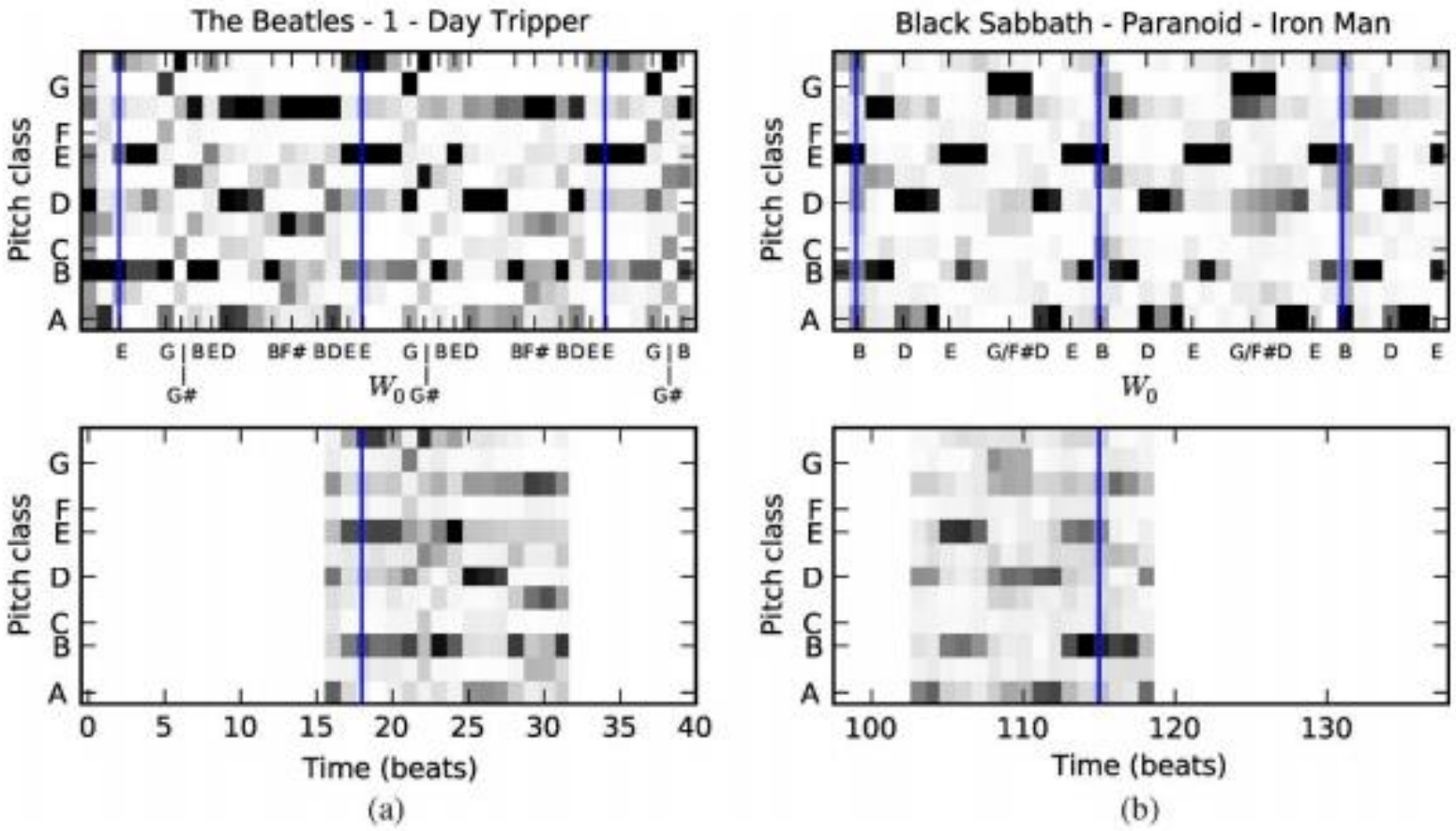
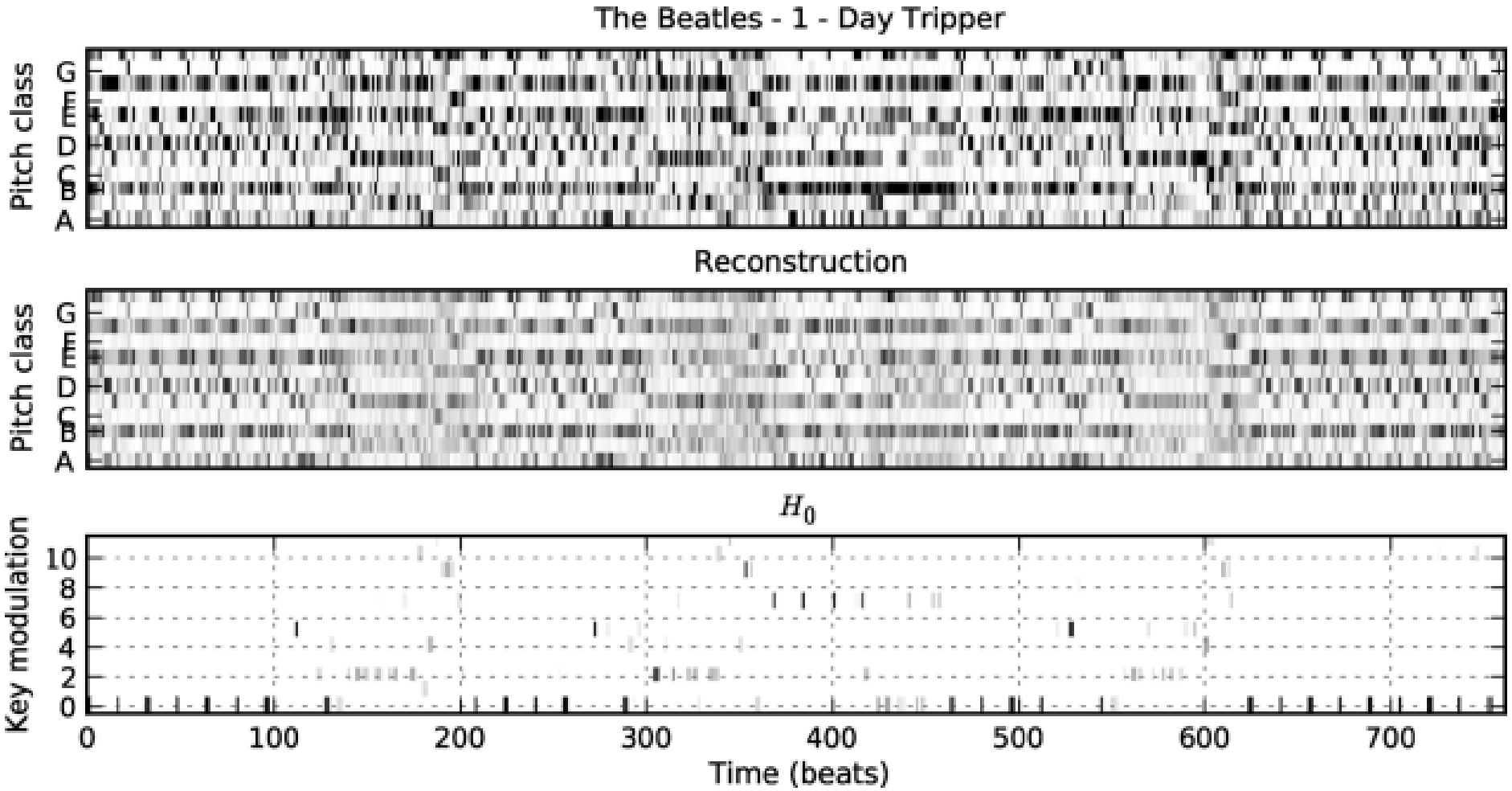
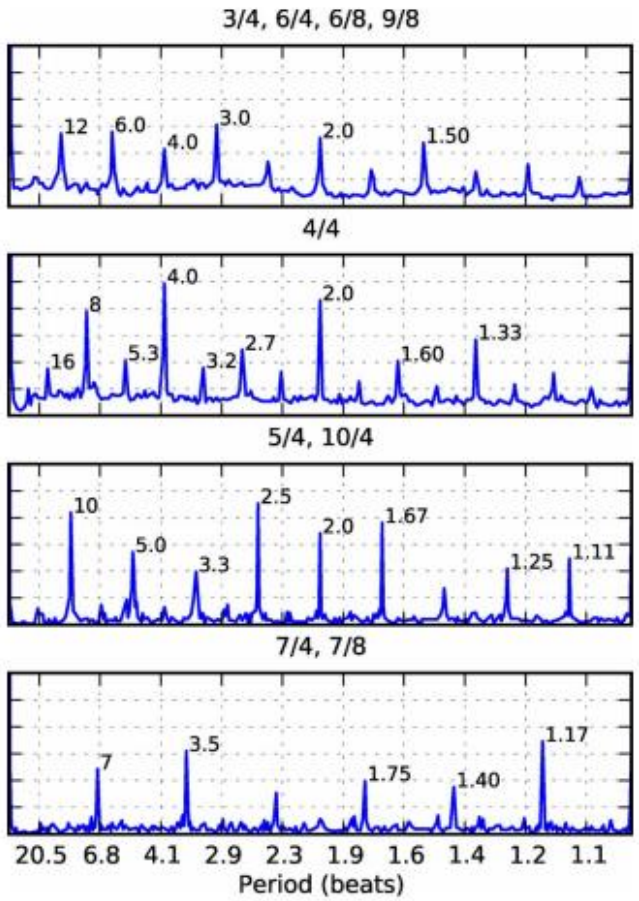


Fig. 4. Main riffs identified in (a) *Day Tripper* by The Beatles and (b) *Iron Man* by Black Sabbath. The top panels show chromagram excerpts from each song including two repetitions of the main riff. The bottom panels show the identified riff W_0 aligned against the top panels. Blue vertical lines indicate the beginning of the riff.


The Beatles - 1 - Day Tripper
Pitch class
G
F
D
C
B
A
Reconstruction
Pitch class
G
F
D
C
B
A
H_0
Key modulation
10
8
6
4
2
0
0
100
200
300
400
500
600
700
Time (beats)

METER ANALYSIS

- Simple analysis of *H* used to discriminate between metrical patterns
- In western music, chord changes are likely to happen at downbeat
- SI-PLCA assumes that the time signature stays constant
- Data set of 342 pop songs in different time signatures, broken into four classes
- Class 3: 144 songs in triple meter (3/4, 6/4, 6/8, 9/8)
- Class 4: 155 songs in 4/4
- Class 5: 25 songs in 5/4 and 10/4
- Class 7: 18 songs in 7/4 and 7/8
- Split into training and testing, generated meter templates

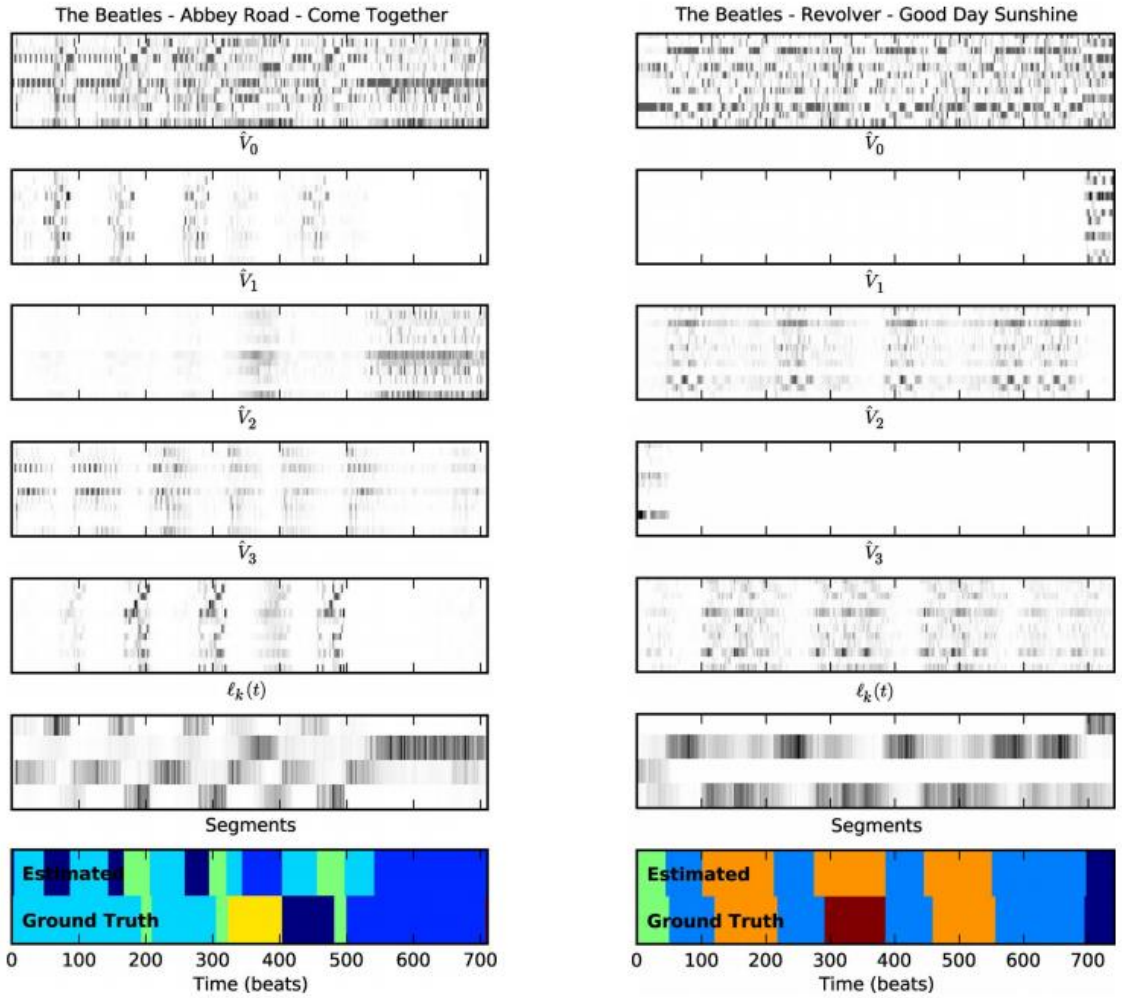


RESULTS

- Grid search over SI-PLCA parameters was used to find best performing features
- Achieved using L = 60, c = 10, m = 5 x 10^-6, and βh = 0.1
- Using these features, accuracy was 61%
- Songs in triple meter confused with 4/4 due to peaks at 2 and 4 beats
- Classes 5 and 7 suffer from overfitting because of small amounts of training data
- When restricted to duple vs triple meter, performance increases to 78%

		Predicted			
	Class	3	4	5	7
True	3	35	15	8	14
	4	13	59	1	4
	5	0	5	3	4
	7	2	0	1	6

STRUCTURE SEGMENTATION RESULTS



COMPARISON TO STATE-OF-THE-ART

TABLE II
SEGMENTATION PERFORMANCE ON THE BEATLES DATA SET. THE NUMBER OF LABELS PER SONG WAS FIXED TO 4 FOR SI-PLCA, QMUL, AND RANDOM. THE AVERAGE EFFECTIVE RANKS FOR SI-PLCA-α_z AND MAUCH *ET AL* WERE 3.9 AND 5.5, RESPECTIVELY

System	PFM	PPR	PRR	S_o	S_u
Mauch et al [46]	0.66	0.61	0.77	0.76	0.64
SI-PLCA-α_z	0.60	0.57	0.69	0.62	0.56
SI-PLCA	0.59	0.60	0.61	0.57	0.57
QMUL [44]	0.54	0.58	0.53	0.50	0.57
Random	0.47	0.43	0.56	0.39	0.41

- Mauch et al outperforms SI-PLCA by 8% and SI-PLCA outperforms QMUL by 16%

CONCLUSIONS

- Approach is successful at finding motif, meter analysis, and structure segmentation of popular music
- Potential for improvement by using more sophisticated post-processing
- Future directions include improving the algorithm to work better with other styles of music, such as classical

Hidden Topic Markov Models

Gruber, Rosen-Zvi and Weiss

—

Hera Liu, 2016 Fall

Motivation & Background

- Modeling word document relationships

- Existing model: LDA (Latent Dirichlet Allocation)
  - Assumption: topics of all words in the same document are independent
  - “an unrealistic oversimplification”

- Proposed model: HTMM (Hidden Topic Markov Model)
  - Assumption: all words in the same sentence have the same topic, and successive sentences are more likely to have the same topics

HTMM

1. for z=1...K,
   Draw $\beta_z \sim \text{Dirichlet}(\eta)$

2. for d=1...D,
   Document d is generated as follows:

   (a) Draw $\theta \sim \text{Dirichlet}(\alpha)$

   (b) Set $\psi_1 = 1$

   (c) for n=2 $\ldots N_d$

      i. If (begin_sentence) draw $\psi_n \sim Binom(\epsilon)$
         else $\psi_n = 0$

   (d) for n=1$\ldots N_d$

      i. if $\psi_n == 0$ then $z_n = z_{n-1}$
         else $z_n \sim multinomial(\theta)$

      ii. Draw $w_n \sim multinomial(\beta_{z_n})$

K: number of topics
z: topic
w: word
βz: prior parameter
Θ: vector that defines distribution of topics
ψn: topic transition variable
Nd: length of document d

Example - LDA vs. HTMM

Abstract We give necessary and sufficient conditions for uniqueness of the **support** vector solution for the problems of pattern recognition and regression estimation, for a general class of cost functions. We show that if the solution is not unique, all **support** vectors are necessarily at bound, and we give some simple examples of non-unique solu- tions. We note that uniqueness of the primal (dual) solution does not necessarily imply uniqueness of the dual (primal) solution. We show how to compute the threshold b when the solution is unique, but when all **support** vectors are at bound, in which case the usual method for determining b does not work.

recognition and regression estimation algorithms [12], with arbitrary convex costs, the value of the normal w will always be unique. Acknowledgments C. Burges wishes to thank W. Keasler, V. Lawrence and C. Nohl of Lucent Tech- nologies for their **support**. References [1] R. Fletcher. Practical Methods of Optimization. John Wiley and Sons, Inc., 2 nd edition, 1987.

LDA

Example - LDA vs. HTMM

Abstract We give necessary and sufficient conditions for uniqueness of the **support** vector solution for the problems of pattern recognition and regression estimation, for a general class of cost functions. We show that if the solution is not unique, all **support** vectors are necessarily at bound, and we give some simple examples of non-unique solu- tions. We note that uniqueness of the primal (dual) solution does not necessarily imply uniqueness of the dual (primal) solution. We show how to compute the threshold b when the solution is unique, but when all **support** vectors are at bound, in which case the usual method for determining b does not work.

recognition and regression estimation algorithms [12], with arbitrary convex costs, the value of the normal w will always be unique. Acknowledgments C. Burges wishes to thank W. Keasler, V. Lawrence and C. Nohl of Lucent Tech- nologies for their **support**. References [1] R. Fletcher. Practical Methods of Optimization. John Wiley and Sons, Inc., 2 nd edition, 1987.

HTMM

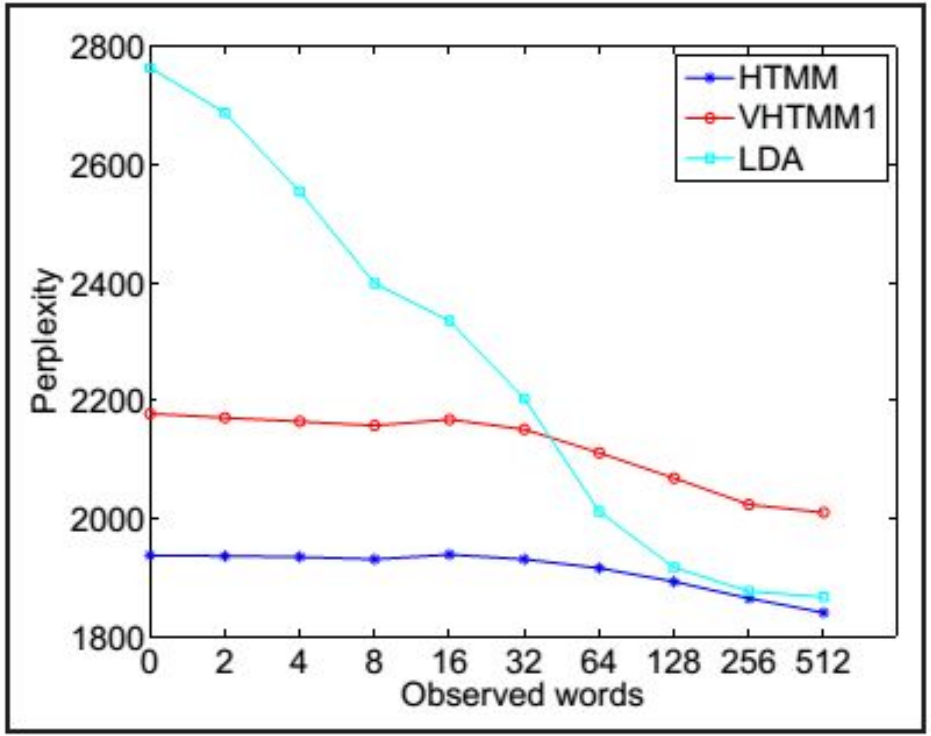
Experiment

- NIPS dataset
- extract words and preserve their orders
- divide the text to sentences according to the delimiters **.?!;**
  - omitted “e.g.” and “i.e.”
- average length about 1300 words per document
- set topic number = 100
- compare perplexities of models
  - perplexity: the difficulty of predicting a new unseen document after observing the first N words of the document

$$\text{Perplexity} = \exp\left(-\frac{\log \Pr(w_{N+1} \ldots w_{N_{test}} | w_1 \ldots w_N)}{N_{test} - N}\right)$$

  - the lower, the better

Result



- Comparison between the average perplexity of HTMM,LDA and VHTMM1.
- HTMM outperforms LDA and VHTMM1 models.

Discussion

- explicitly model the topic dynamics with a Markov chain

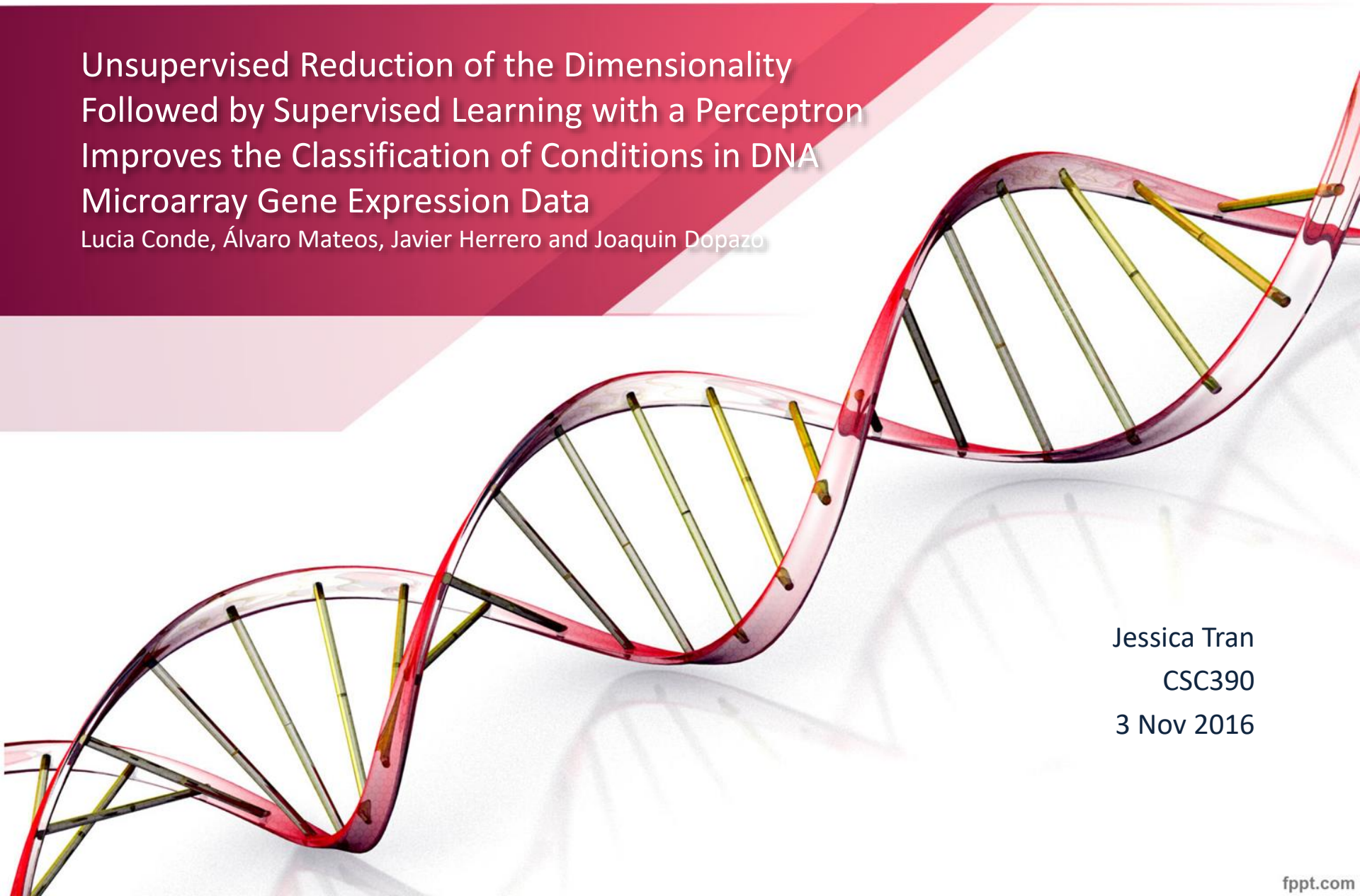
- learn more coherent topics and disambiguate the topics of ambiguous words

- combine different extensions of LDA together to form a better document analysis system
  - automatic translation of texts

Questions?

Unsupervised Reduction of the Dimensionality Followed by Supervised Learning with a Perceptron Improves the Classification of Conditions in DNA Microarray Gene Expression Data

Lucia Conde, Álvaro Mateos, Javier Herrero and Joaquin Dopazo



Jessica Tran

CSC390

3 Nov 2016

Motivation

- Classify conditions in DNA array gene expressions
  - Rows: genes in study
  - Columns: different experimental conditions
    - Usually cancer types
- Number of patterns of gene expression
  - Neural networks
    - Success with unsupervised and supervised methods
- Efficiency
  - Supervised methods
    - More efficient, takes metadata into consideration

Background Information

- Processing data to reduce dimensionality
  - Principal Component Analysis (PCA)
    - Used on all gene expression profiles (rows)
    - Principal components are used for training
    - Loss of biological meaning for statistical support
  - New proposal by Mateos:
    - Pre-clustering gene expression patterns
    - Average expression values of clusters are used with supervised neural network for classifying experimental conditions

Goals

- Clusters with optimal level of information:
  - Redundancies removed
  - Noise reduction
  - Number of elements (clusters of co-expressing genes) is as small as possible
  - Retains necessary information to produce an accurate classification by using a perceptron

Key Terminology

- Gene Ontology (GO) (*www.geneontology.org*)
- Perceptron – algorithm for supervised learning of binary classifiers (belongs to a specific class or not)
- Self Organizing Tree Algorithm (SOTA)
  - Unsupervised hierarchical neural network
  - Divisive method
  - Starts as a 2-neuron tree with a mother neuron
  - Recursively split neurons based on heterogeneity
  - End result: binary tree structure that accounts for the variability of the data set

Methods

1. SOTA - Reduced data via clustering
2. Train perceptron with average values from (1)
3. Find optimal resolution level using perceptron
   - Determine appropriate dimensionality/level of clustering for optimal classifications
     - 100% of true positives is obtained or maximum value followed by a decrease
4. Analyze magnitudes of interconnected weights of perceptron
   - Importance of different clusters

Using 2 datasets:

(A) – Khan's dataset (2285 genes)

(B) – Alizadeh's dataset (2414 genes)

Results

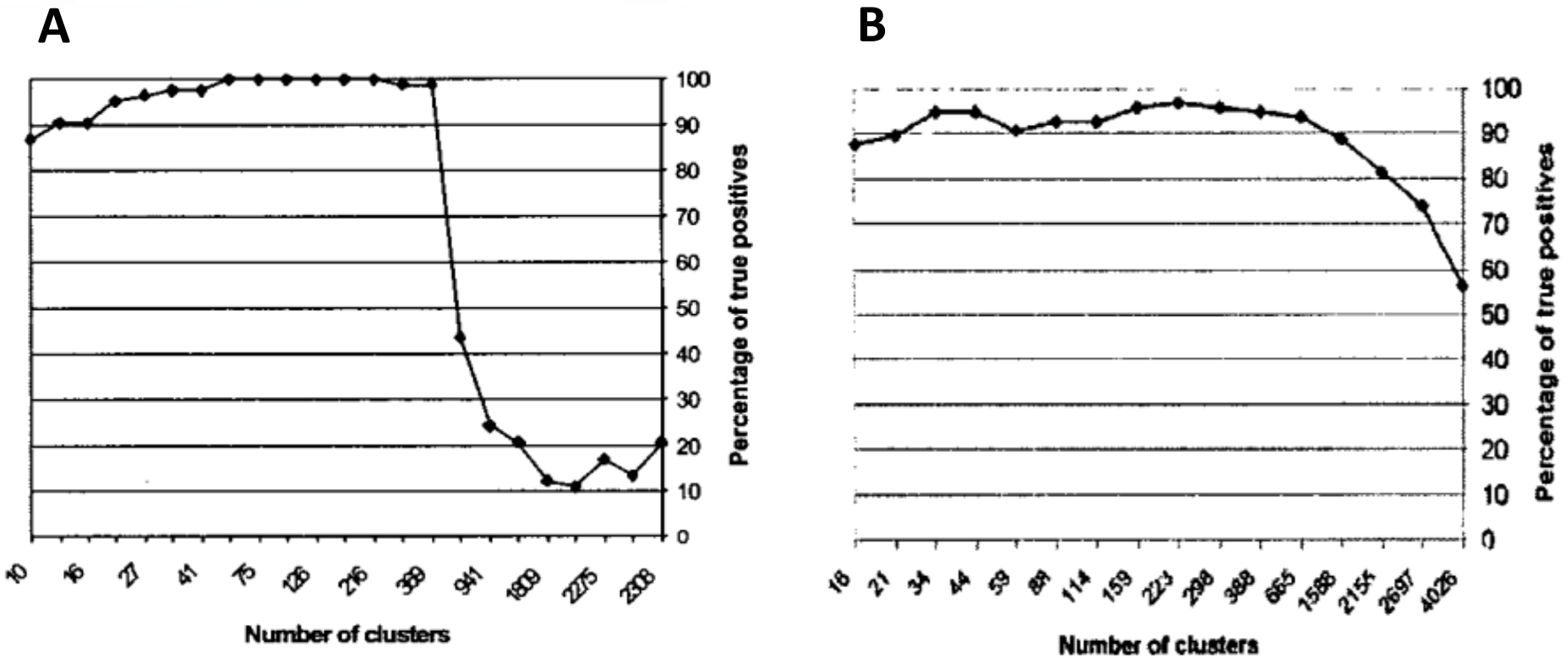


Figure 1. Number of true positives (cell lines properly classified) for different number of clusters taken as input nodes for the perceptron. The clusters are obtained by applying SOTA with different thresholds, which produces a hierarchical classification of the gene expression profiles at different levels of resolution, that is, a tree with different number of clusters. **A** corresponds to Khan's dataset [7] and **B** corresponds to Alizadeh's dataset [1].

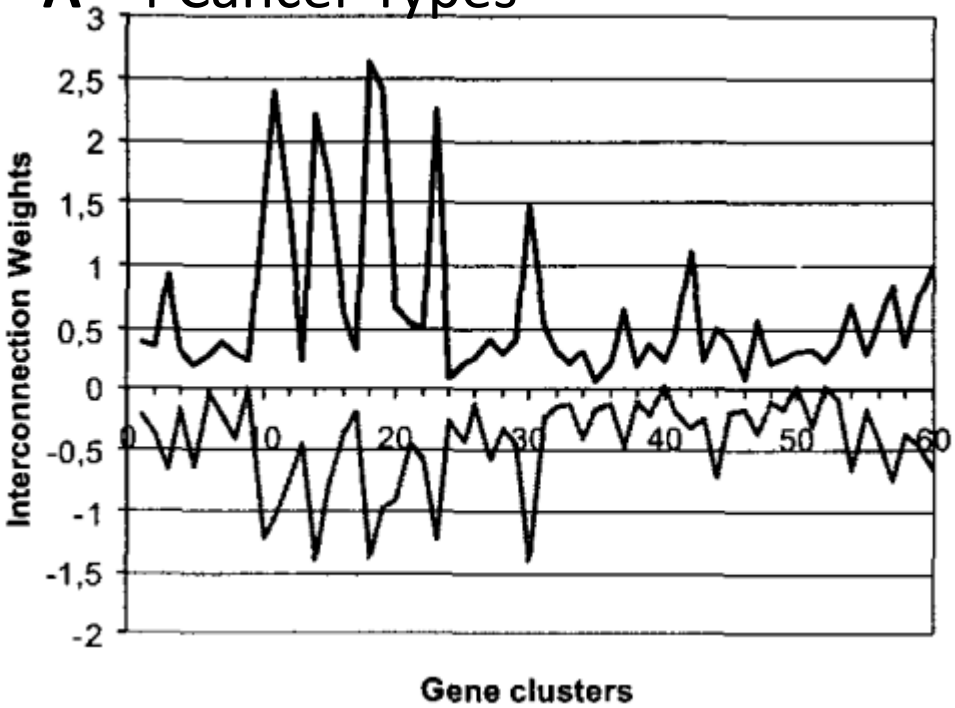
Results

Cell line	Total	Supervised pre-clustered (44 clusters)	Supervised pre-clustered (223 clusters)	Supervised on PCA	Unsuper-vised
DLBCL	46	43	45	45	43
Germinal centre B	2	2	2	0	2*
Nl. Lymph node/tonsil	2	1	1	0	1*
Activated blood B	10	10	10	10	10
Resting/acti -vated T	6	6	6	6	6
Transforme d cell lines	6	6	6	5	6
Follicular lymphoma	9	9	9	9	7**
Resting blood B	4	3	3	4	2***
Chronic lymphocyti c leukemia	11	11	11	11	11
Total	96	91	93	90	88

Table 1. Classification of cell lines. Supervised and unsupervised columns show the number of cell types properly classified out of the total number, shown in the column labeled as Total. Supervised pre-clustered results were obtained using a perceptron with 44 and 223 input nodes and nine output nodes (see text). Supervised on PCA results were obtained using a perceptron with 14 input nodes corresponding to the 14 principal components that explain 65% of the variance of the data. Unsupervised clustering was taken directly from the Alizadeth et al. [1] paper. The number of properly classified cases was obtained from the cluster most populated by a given class, no mater other classes are present in this cluster.

Results

A - 4 Cancer Types



B – 9 cell line classes

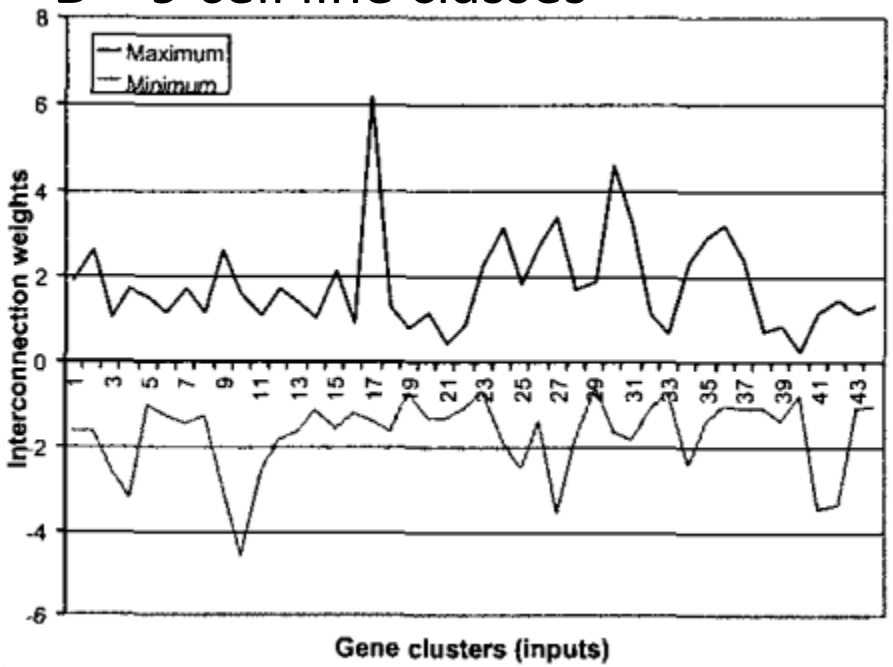


Figure 2 Maximum and minimum of the weight values connecting each of the clusters of co-expressing genes (input layer of the perceptron) to the different types of cell lines (output layer of the perceptron). **A**, resolution level corresponding to 60 clusters in Khan's dataset (Khan et al., 2001). **B**, resolution level corresponding to 44 clusters in Alizadeh's dataset [1].

Results

Alizadeh's Dataset

- Strongest weight connections in Alizadeh's dataset separates cancer and non-cancer cell types

Khan's Dataset

- Distinction between different cancer types
- Strong weights:
  - Nucleic acid binding
  - Protein binding
- Genes must be involved in processes that make these cancer types different
- Further research

Conclusions

- Mateos' clustering approach seems to perform better than PCA
- Study and analysis of weight connections is extremely valuable
  - Extract informative genes in dataset
  - Can suggest biological explanation for differences observed the different experimental conditions