Caltech

Machine Learning & Data Mining

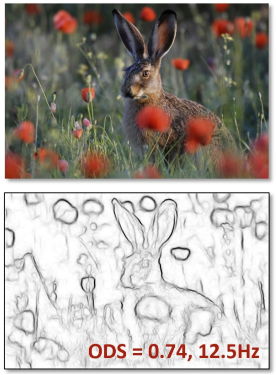
CS/CNS/EE 155

Lecture 12:

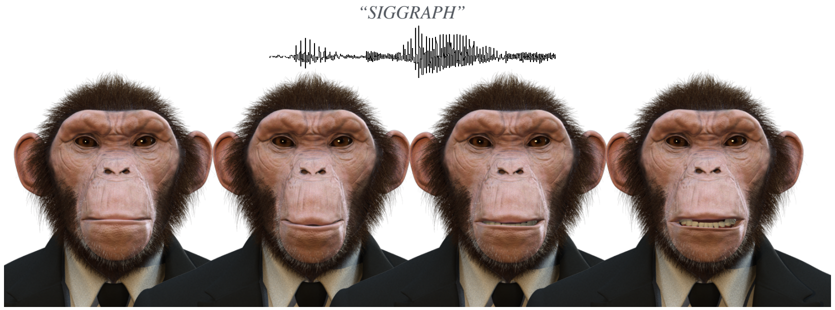
Recent Applications

Today: Recent Applications

Edge Detection



Speech Animation



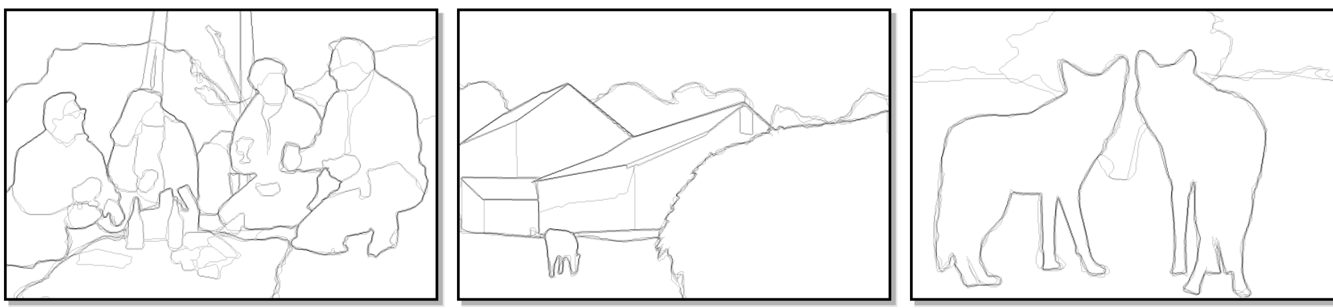
Embeddings of Visual Style

(Briefly) Generating Faces

Edge Detection

X:

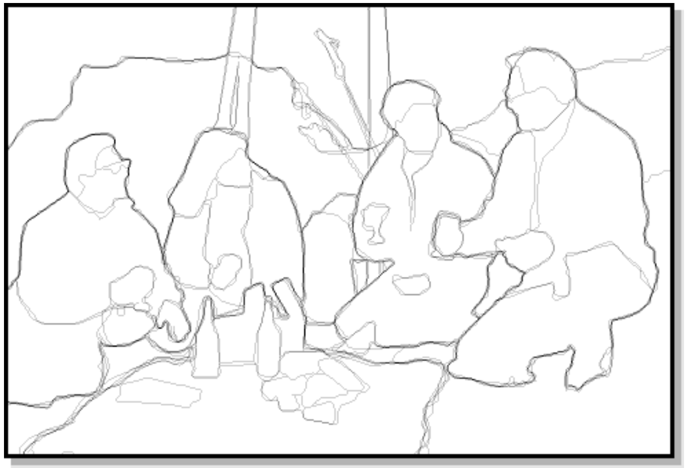
Y:



Challenges

- Output Space?

- 400x300 Image
  - 120000 Pixels
  - **2^{120000} Labels!**



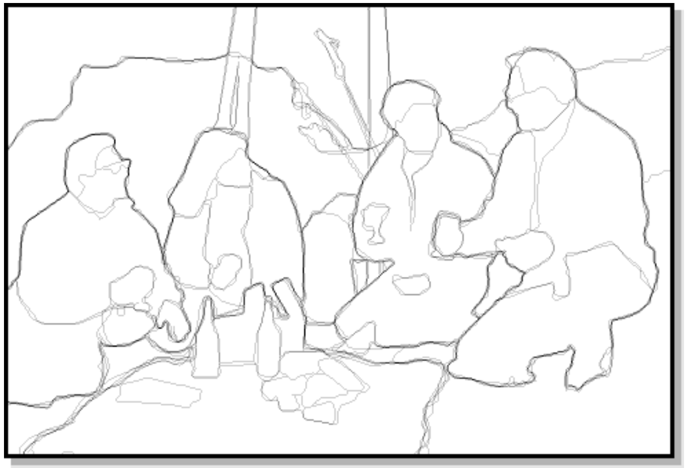
Today (first half): Learning Reductions

- Convert complicated problem into simpler ones
  - Use complex models for simpler problems
  - E.g., decision trees, neural nets

- Recompose predictions for complicated problem

Strong Local Properties

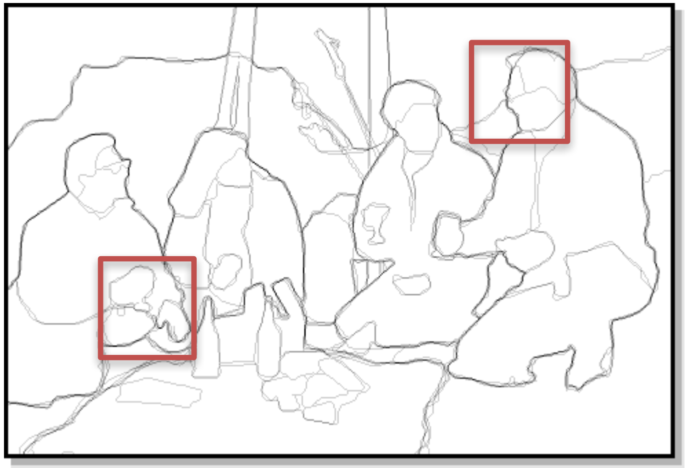
- Local patterns matter
  - E.g., image patches
- Complex relationship
  - Non-linear



Weak Global Properties

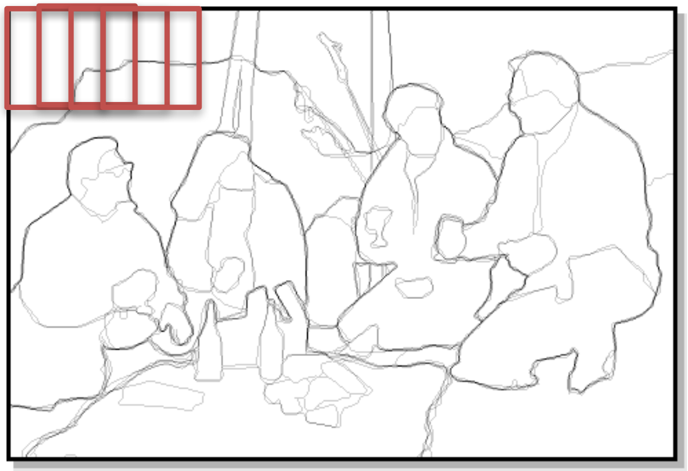
- Edge detections local

- Can ignore most of image

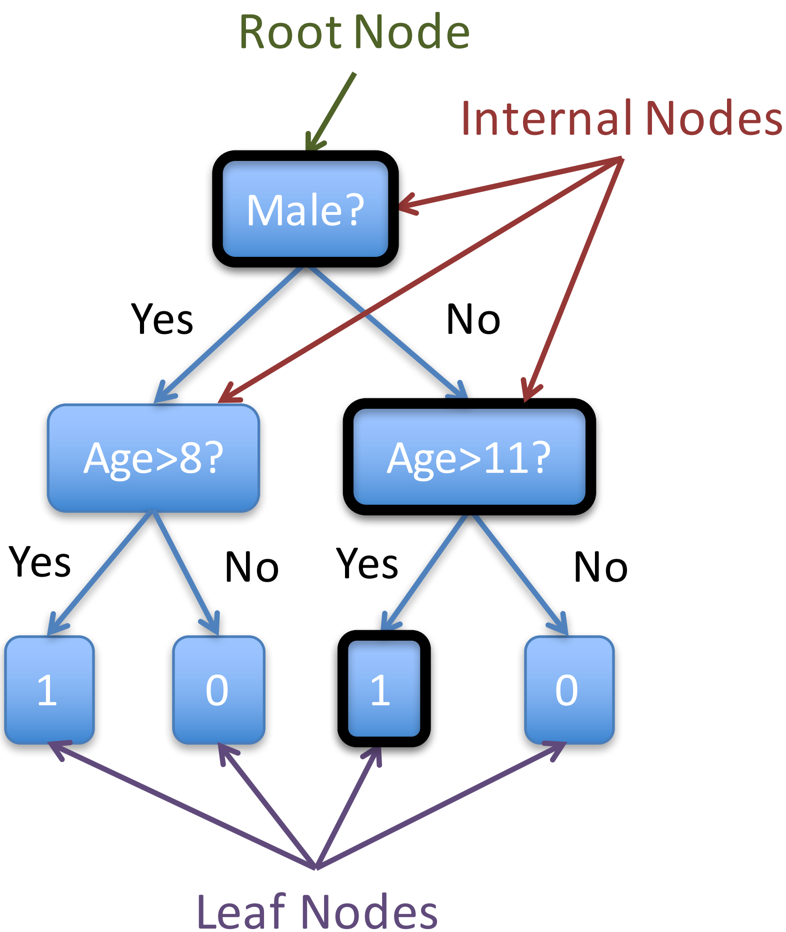


Sliding Window Approach
(Decomposition)

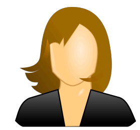
- Train model to predict patches
  - E.g., 16x16
- Slide across image
- **What model?**



Recall: Binary Decision Tree



Input:



Alice
Gender: Female
Age: 14

Prediction: Height > 55”

Every **internal node** has a **binary** query function q(x).

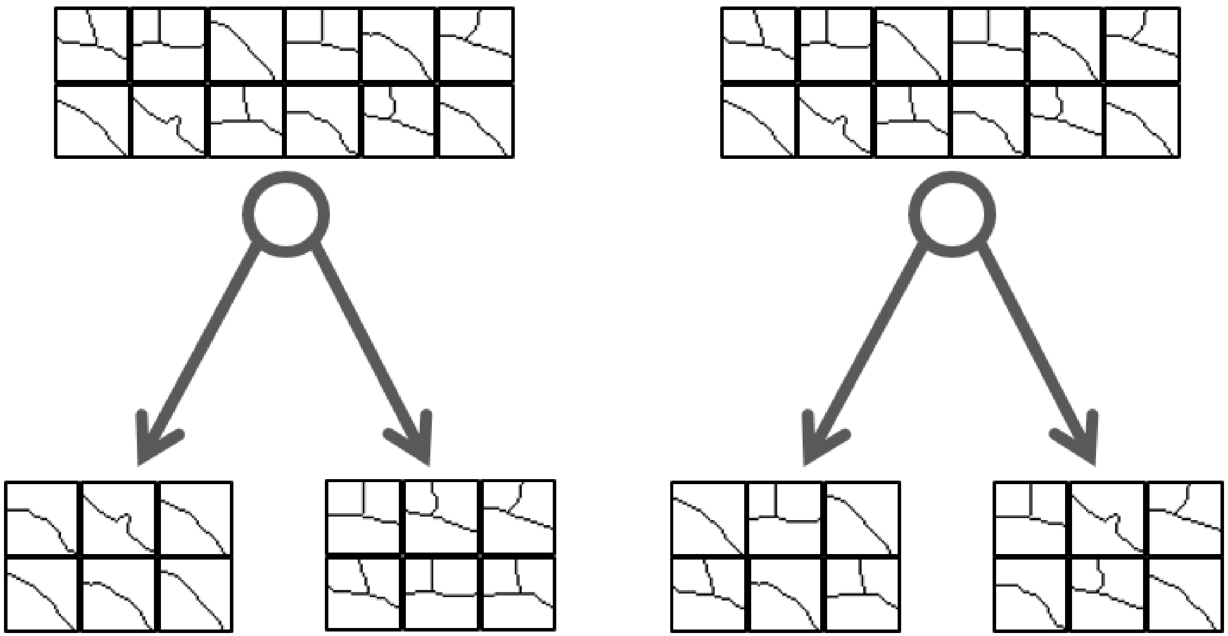
Every **leaf node** has a prediction, e.g., 0 or 1.

Prediction starts at **root node**.
Recursively calls query function.
Positive response → Left Child.
Negative response → Right Child.
Repeat until Leaf Node.

Structured Decision Tree

- Each leaf node predicts a 16x16 edge matrix
  - Average of all training patch labels
- Prediction is very fast!
  - Slide predictor across image, average results
  - No need for Viterbi-type algorithms
- What is splitting criterion?
- What is query set?

Structured Information Gain



Good!

Bad!

"Structured Random Forests for Fast Edge Detection"
Dollár & Zitnick, ICCV 2013

Structured Information Gain

1. First map labels to coordinate system
   A. For each coordinate, choose pair of pixels
   B. Set coordinate to 1 if in same segment, 0 o.w.
      - Coordinate 1 = 0

(Actual approach more complicated.)

“Structured Random Forests for Fast Edge Detection”
Dollár & Zitnick, ICCV 2013

Structured Information Gain

1. First map labels to coordinate system
   A. For each coordinate, choose pair of pixels
   B. Set coordinate to 1 if in same segment, 0 o.w.
      - Coordinate 1 = 0
      - Coordinate 2 = 1
      - Etc…

For each training example!

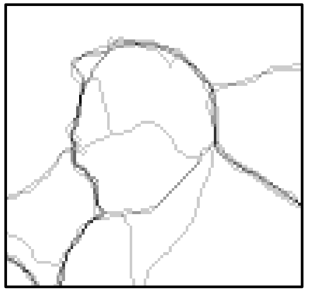
(Actual approach more complicated.)

"Structured Random Forests for Fast Edge Detection"
Dollár & Zitnick, ICCV 2013

Structured Information Gain

1. First map labels to coordinate system
   A. For each coordinate, choose pair of pixels
   B. Set coordinate to 1 if in same segment, 0 o.w.
      - Coordinate 1 = 0
      - Coordinate 2 = 1
      - Etc…
2. Cluster training labels

For each training example!

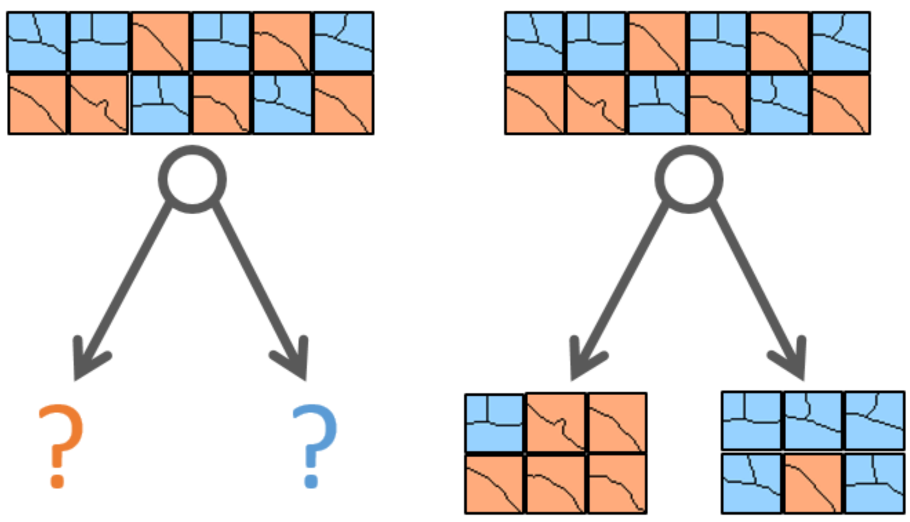


(Actual approach more complicated.)

“Structured Random Forests for Fast Edge Detection”
Dollár & Zitnick, ICCV 2013

Multiclass Entropy

- Reduced training labels to K clusters
  - Can treat as multiclass classification
- Impurity measure = multiclass entropy



Query Set

- Features about color gradients
  - Image gets darker from column 1 to column 5
  - Image gets more blue from row 7 to row 3
  - Etc…
  - 7228 features total

(Actual approach more complicated.)

“Structured Random Forests for Fast Edge Detection”
Dollár & Zitnick, ICCV 2013

Putting it Together

- Create new training set $\hat{S} = \{(x,\hat{y})\}$
  - x = 16x16 image patch
  - $\hat{y}$ = 16x16 ground truth edges

Decomposition

- Train structured DT on $\hat{S}$

- Predict by sliding DT over input image
  - Average predictions

Recomposition

(Actual approach more complicated.)

"Structured Random Forests for Fast Edge Detection"
Dollár & Zitnick, ICCV 2013


Input
Ground Truth
Four Versions of Method

Comparable accuracy vs state-of-the-art

Much faster!

	ODS	OIS	AP	FPS
Human	.80	.80	-	-
Canny	.60	.64	.58	15
Felz-Hutt [11]	.61	.64	.56	10
Hidayat-Green [16]	.62†	-	-	20
BEL [9]	.66†	-	-	1/10
gPb + GPU [6]	.70†	-	-	1/2‡
gPb [1]	.71	.74	.65	1/240
gPb-owt-ucm [1]	.73	**.76**	.73	1/240
Sketch tokens [21]	.73	.75	**.78**	1
SCG [31]	**.74**	**.76**	.77	1/280
SE-SS, T=1	.72	.74	.77	**60**
SE-SS, T=4	.73	.75	.77	30
SE-MS, T=4	**.74**	**.76**	**.78**	6

Accuracy Measures

Speed

"Structured Random Forests for Fast Edge Detection"
Dollár & Zitnick, ICCV 2013

Speech Animation

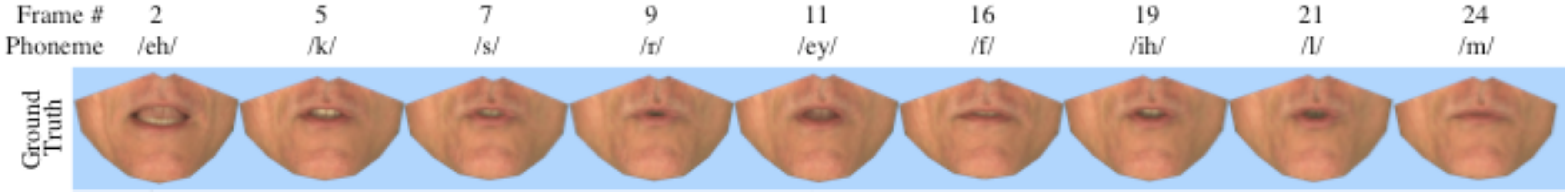
Automatically Animate to Input Audio?

(Given Training Data)

A Decision Tree Framework for Spatiotemporal Sequence Prediction
Taehwan Kim, Yisong Yue, Sarah Taylor, Iain Matthews. KDD 2015
A Deep Learning Approach for Generalized Speech Animation
Sarah Taylor, Taehwan Kim, Yisong Yue, et al. SIGGRAPH 2017

Training Data

- ~2500 Sentences
  - Recorded at 30 Hz
  - ~10 hours of recorded speech
- Active Appearance Model
  - Actor's lower face
  - 30 degrees of freedom (also 100+)



Data from [Taylor et al., 2012]

Prediction Task

Input sequence $X = < x_1, x_2, \ldots, x_{|x|} >$

Output sequence $Y = < y_1, y_2, \ldots, y_{|y|} > \ , y_t \in R^D$

Goal: learn predictor $h : X \to Y$

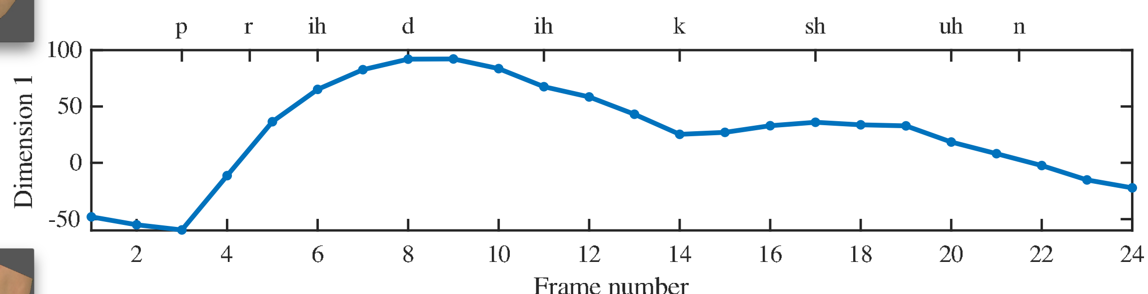
X

Frame	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Token	-	p	p	r	ih	ih	d	d	ih	ih	ih	ih	k	k	sh	sh	sh	sh	uh	uh	n	-

Phoneme sequence

h

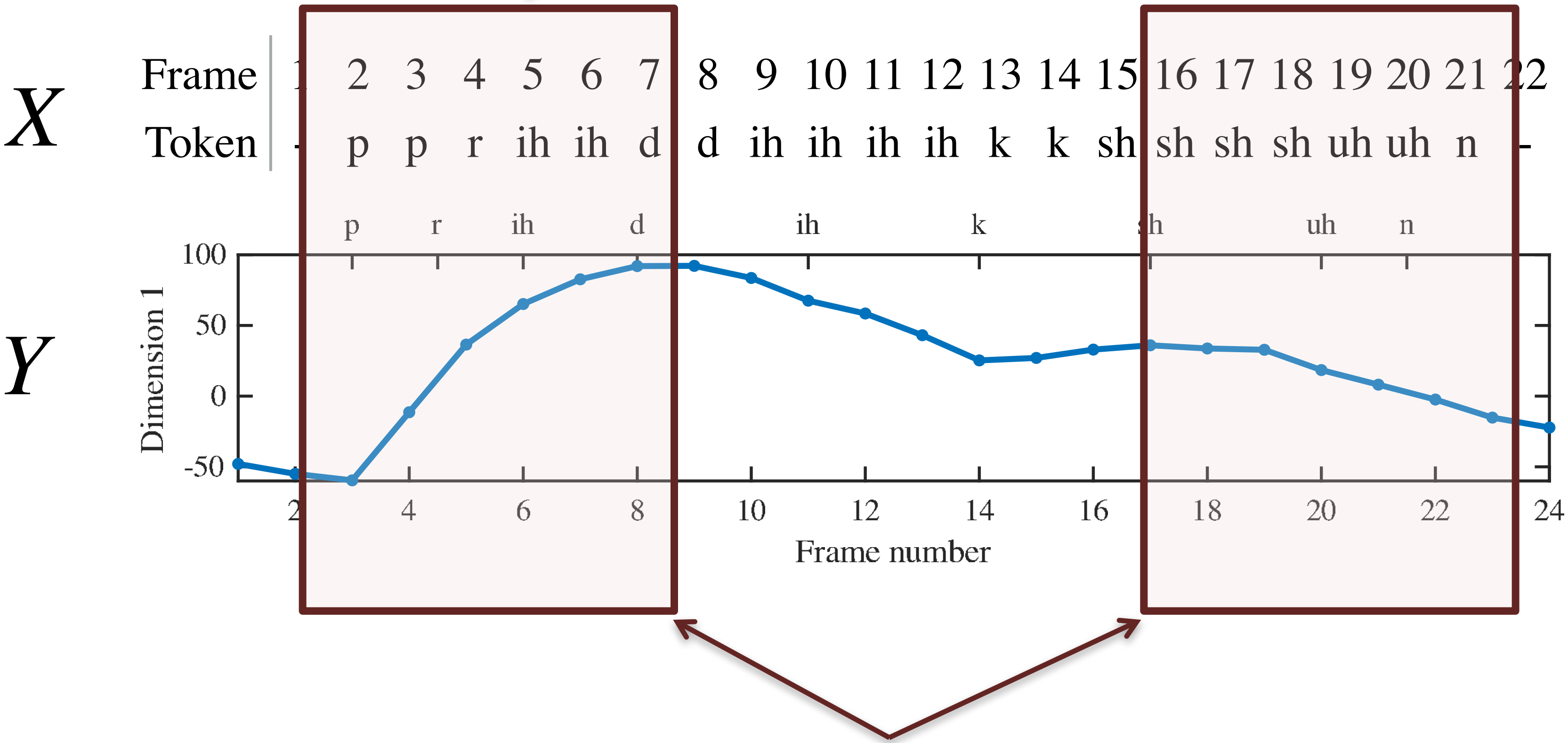
Y



Sequence of face configurations

Temporal curvature can vary smoothly or sharply

(Depends on context – this is the co-articulation problem)

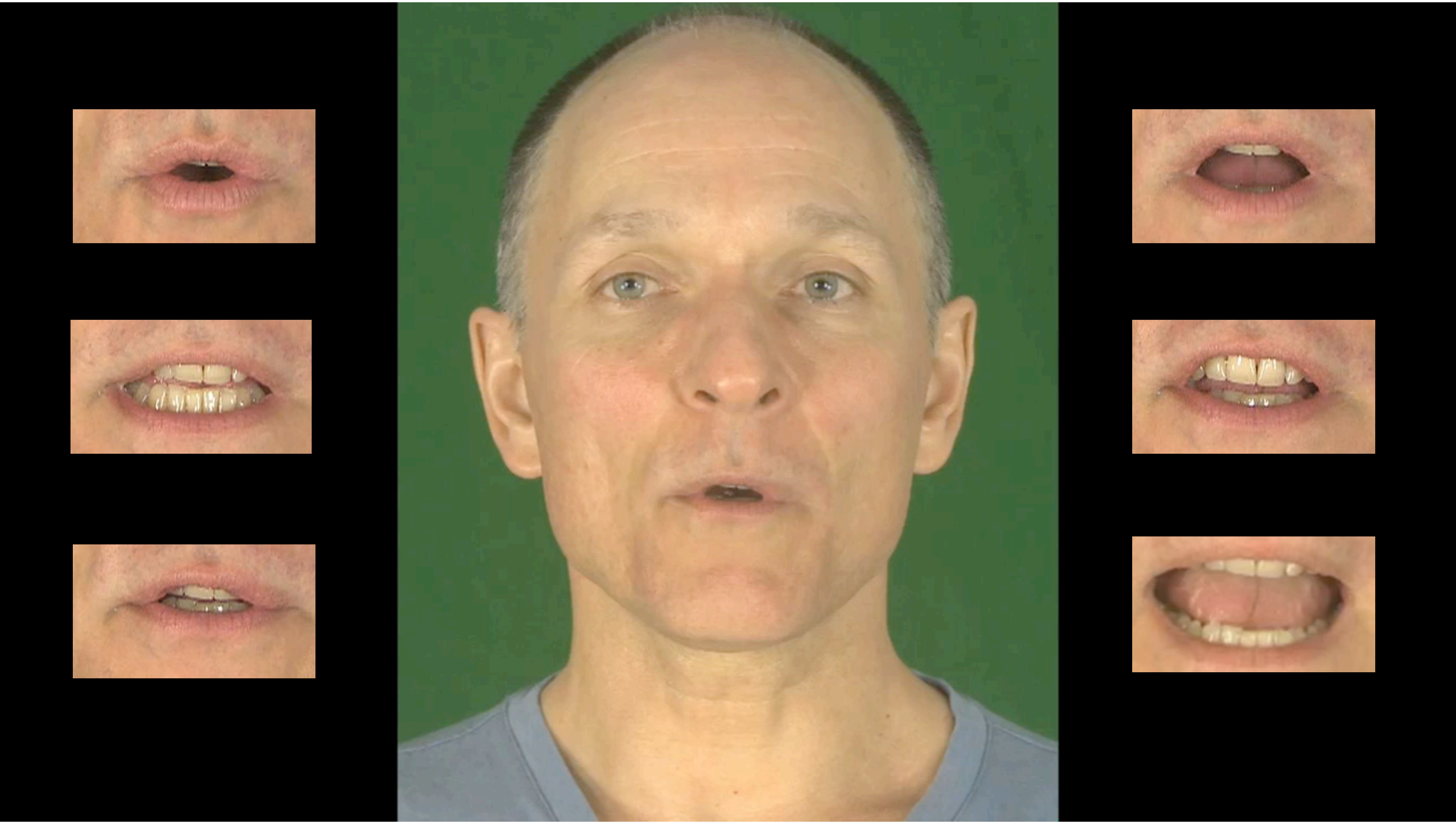


Minimal long-range dependencies

(prediction = construction = election…)

Co-Articulation is Hard to Get Right

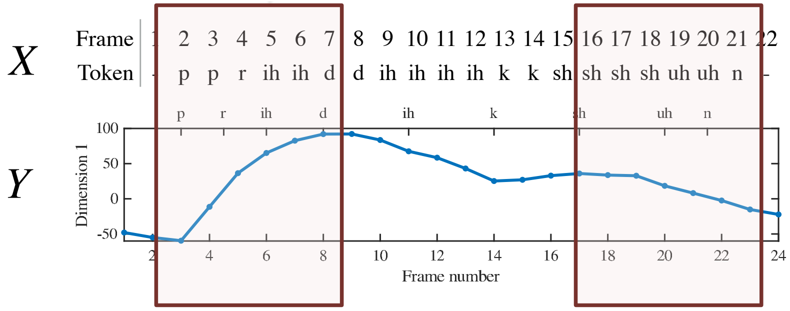
(Strong Local Properties)



/k/

Weak Global Properties

- No need to model entire chain directly



Minimal long-range dependencies
(predi**ction** = constru**ction** = ele**ction**…)

- Motivates sliding window approach!

Input speech: " P R E D I C T I O N "

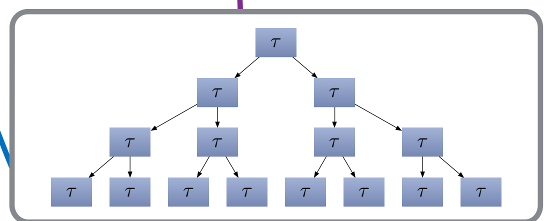
	Frame	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
x	Token	-	p	p	r	ih	ih	d	d	ih	ih	ih	ih	k	k	sh	sh	sh	sh	uh	uh	n	-

$\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots$

… r ih **ih** d d
ih ih **d** d ih
ih d **d** ih ih
d d **ih** ih ih
d ih **ih** ih ih …

Overlapping Sliding Window of Inputs

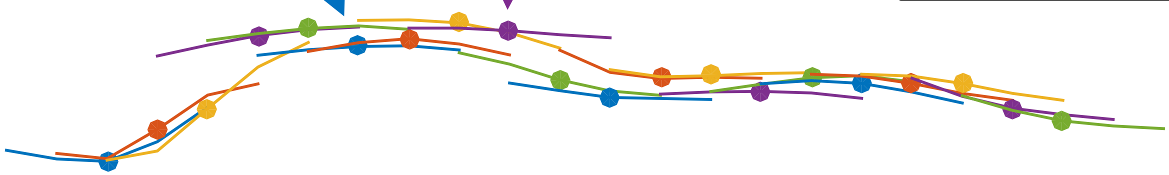
$h(\hat{\mathbf{x}})$



Decision Tree Model
150-variate regression

This is the only thing that requires machine learning!

$\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots$



Aggregate Outputs

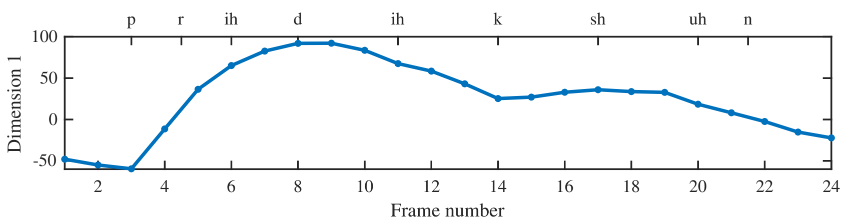
Very fast!

Training

Input speech: " P R E D I C T I O N "

	Frame	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
x	Token	-	p	p	r	ih	ih	d	d	ih	ih	ih	ih	k	k	sh	sh	sh	sh	uh	uh	n	-

y

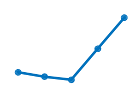
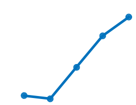


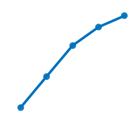
Original Training Data

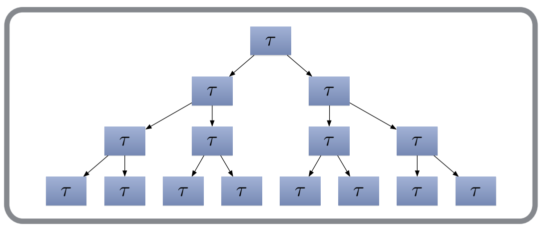
(Variable-Length Trajectory Prediction)

Modified Training Data

(Fixed-Length Multivariate Regression)

$\Big(\langle -, p, p, r, ih \rangle,$  $\Big), \Big(\langle p, p, r, ih, ih \rangle,$  $\Big)$

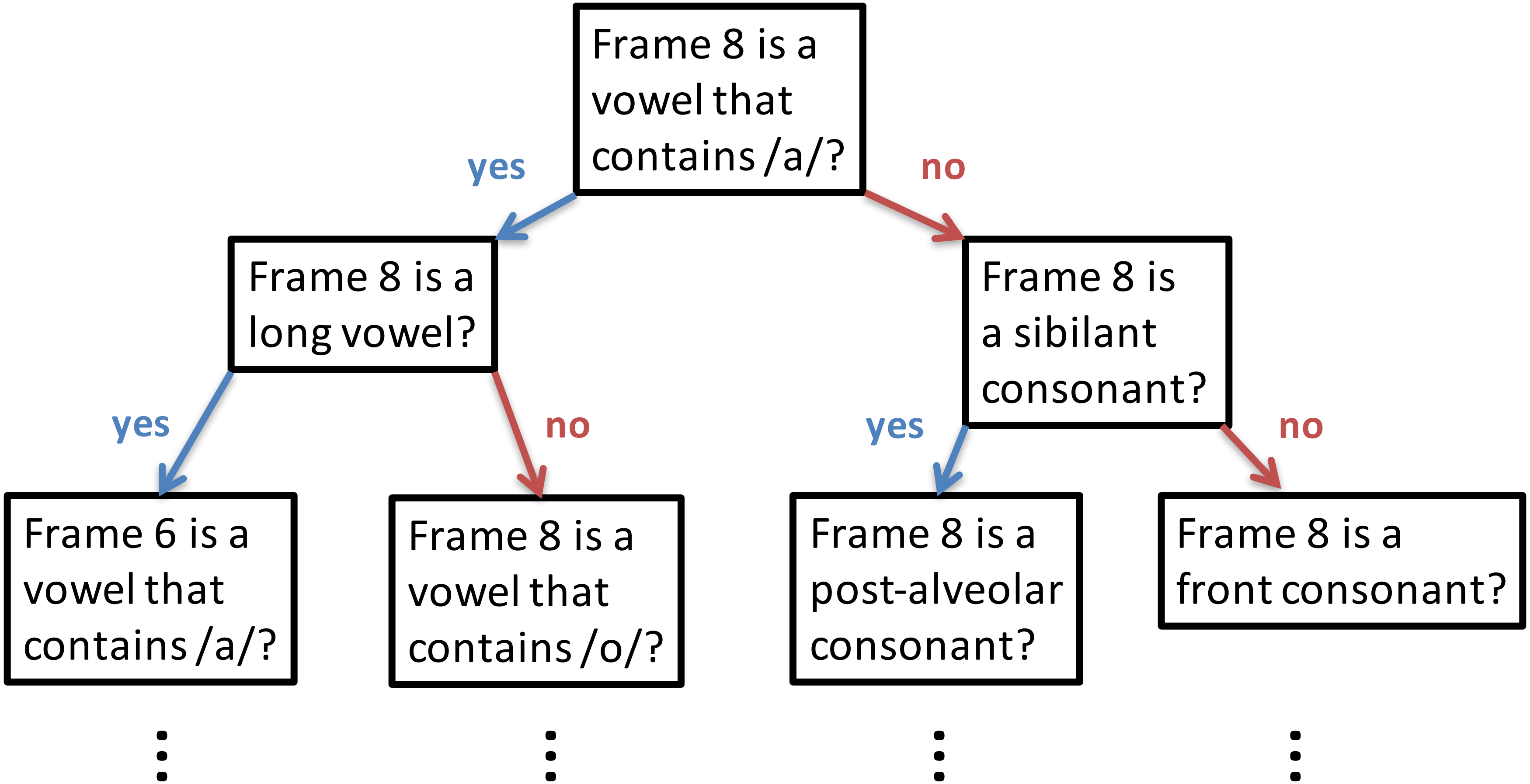
$\Big(\langle p, r, ih, ih, d \rangle,$  $\Big),$ ...



Train Decision Tree

(Or some other regression model)

Query Set for Speech Animation



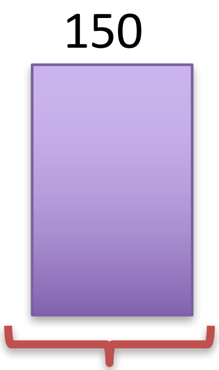
Frames indexed by 1-11 (center is frame 6)

Full tree has 5K+ leaf nodes

Multivariate Regression Tree

- **Prediction:**

Training Data
in Leaf Node:



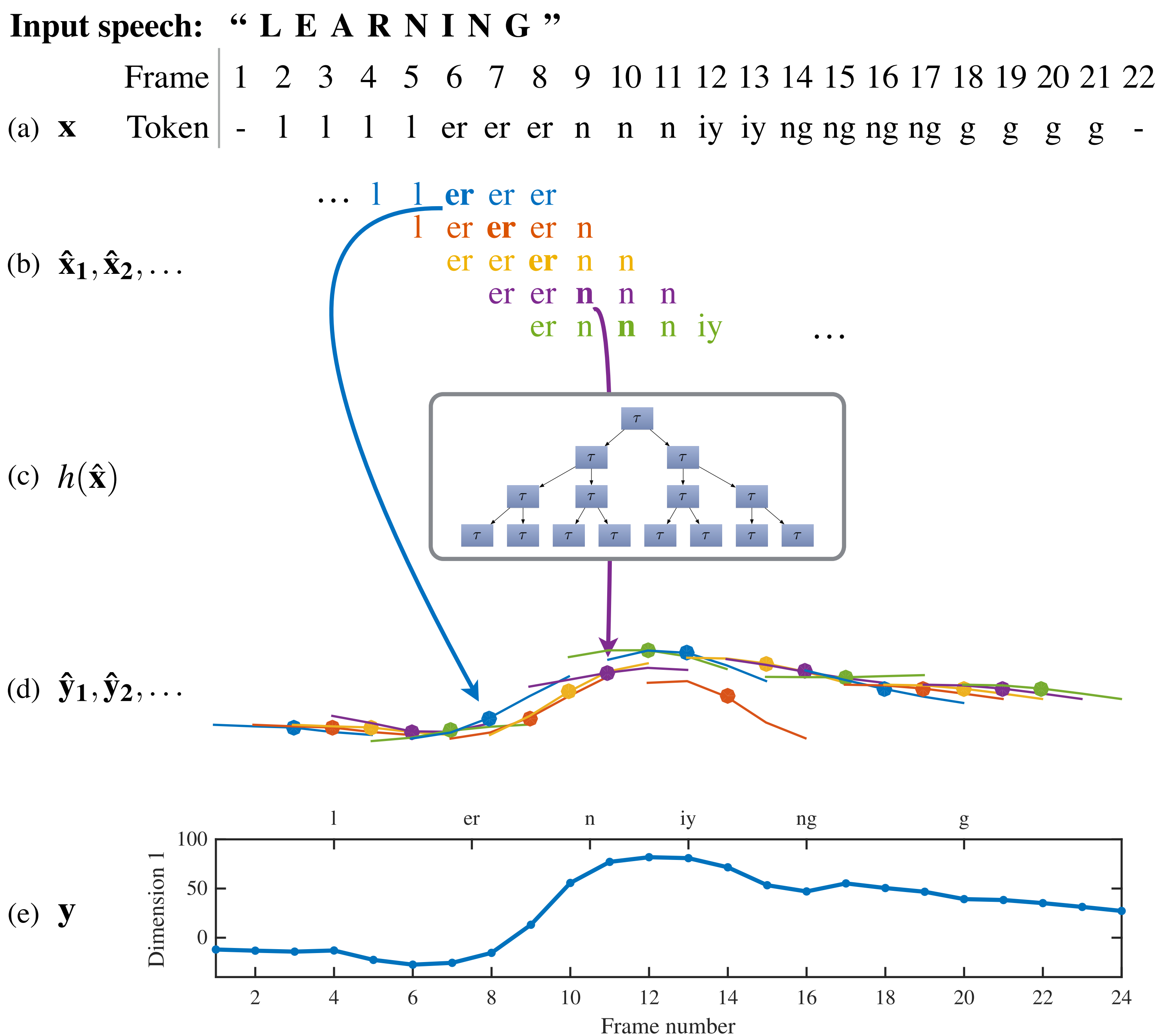
Prediction:  = Mean

- **Training loss:** multivariate squared loss:

$$\sum_{Leaf} \sum_{\hat{y} \in Leaf} \left\| \hat{y}_{Leaf} - \hat{y} \right\|^2$$

Prediction on New Speaker

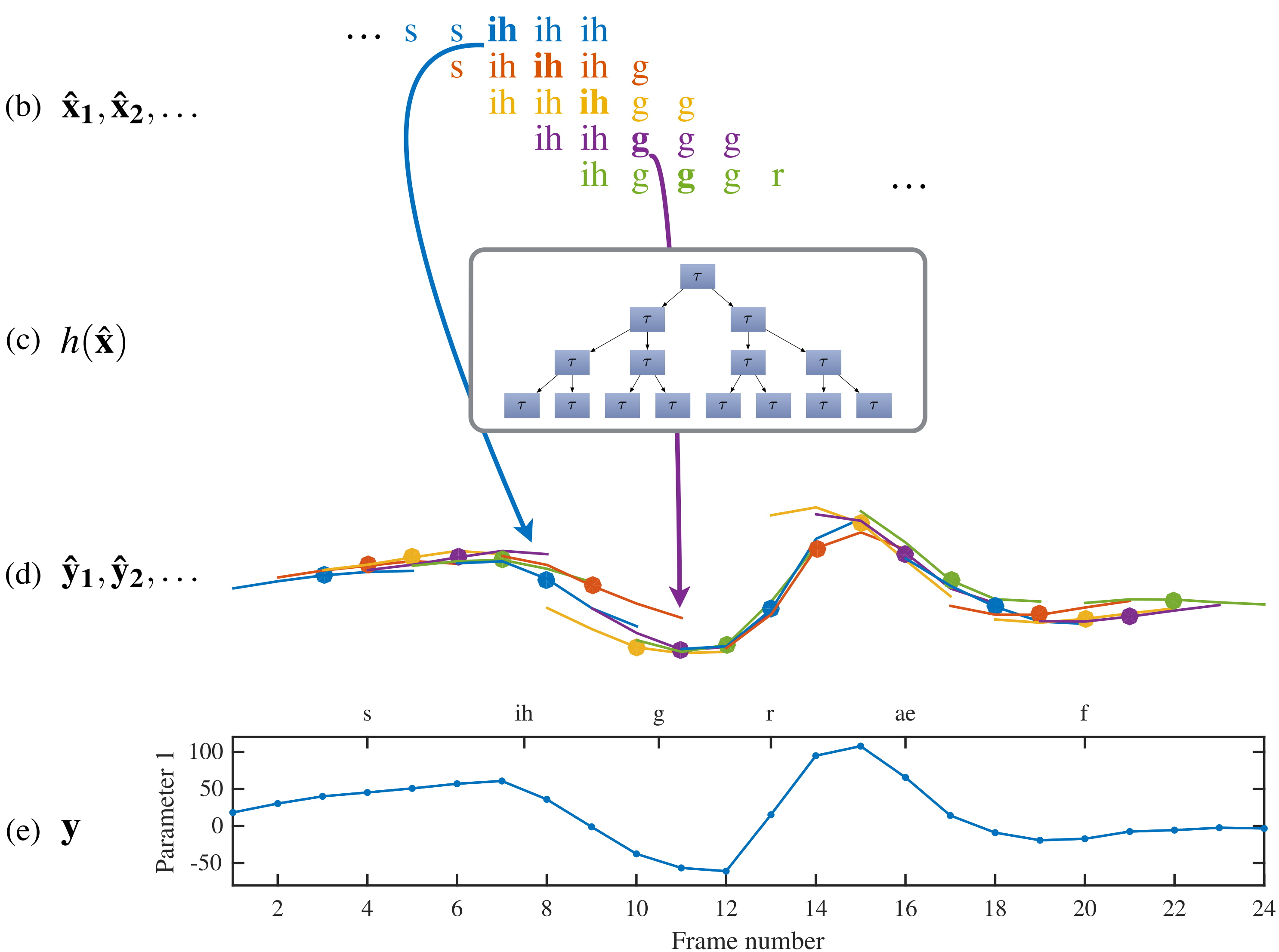
A Decision Tree Framework for Spatiotemporal Sequence Prediction
Taehwan Kim, Yisong Yue, Sarah Taylor, Iain Matthews. KDD 2015
A Deep Learning Approach for Generalized Speech Animation
Sarah Taylor, Taehwan Kim, Yisong Yue, et al. SIGGRAPH 2017



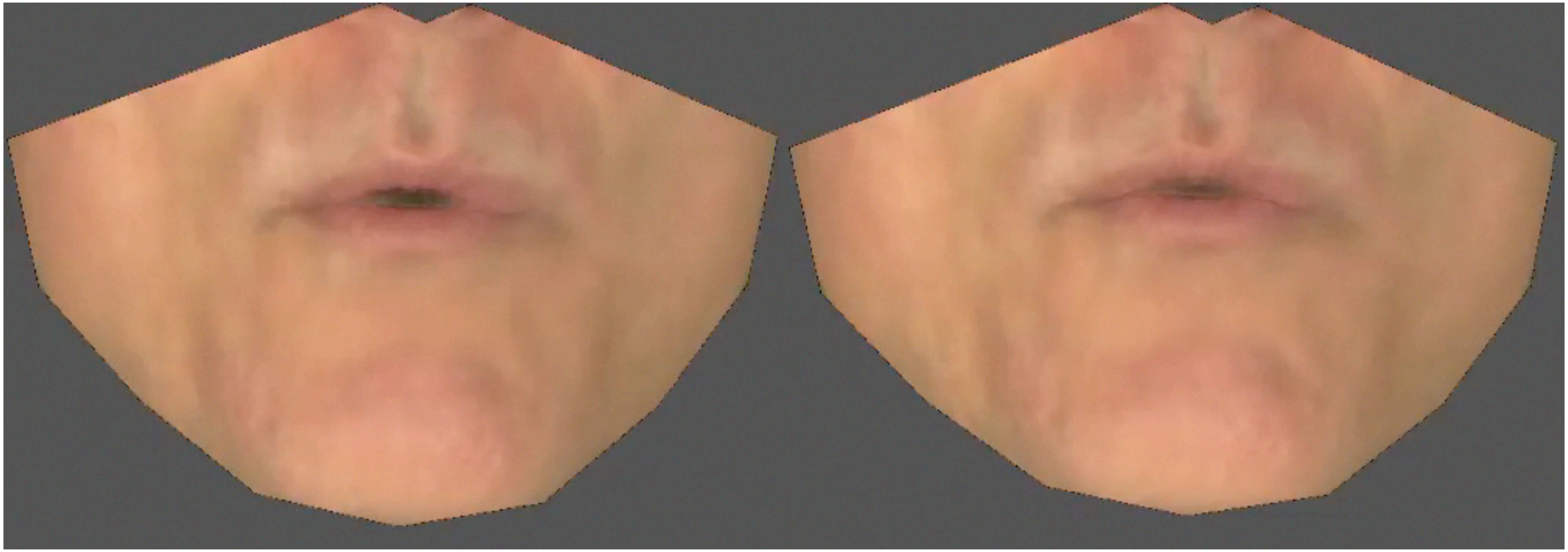
Input speech: “ L E A R N I N G ”
Frame 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
(a) x Token - l l l l er er er n n n iy iy ng ng ng ng g g g g -
(b) $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots$
… l l er er er
l er er er n
er er er n n
er er n n n
er n n n iy
…
(c) $h(\hat{\mathbf{x}})$
τ
(d) $\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots$
(e) y
l er n iy ng g
Dimension 1
100
50
0
2 4 6 8 10 12 14 16 18 20 22 24
Frame number

Input speech: “ S I G G R A P H ”

(a) $\mathbf{x}$																						
Frame	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Label	-	s	s	s	s	ih	ih	ih	g	g	g	r	r	ae	ae	ae	ae	f	f	f	f	-



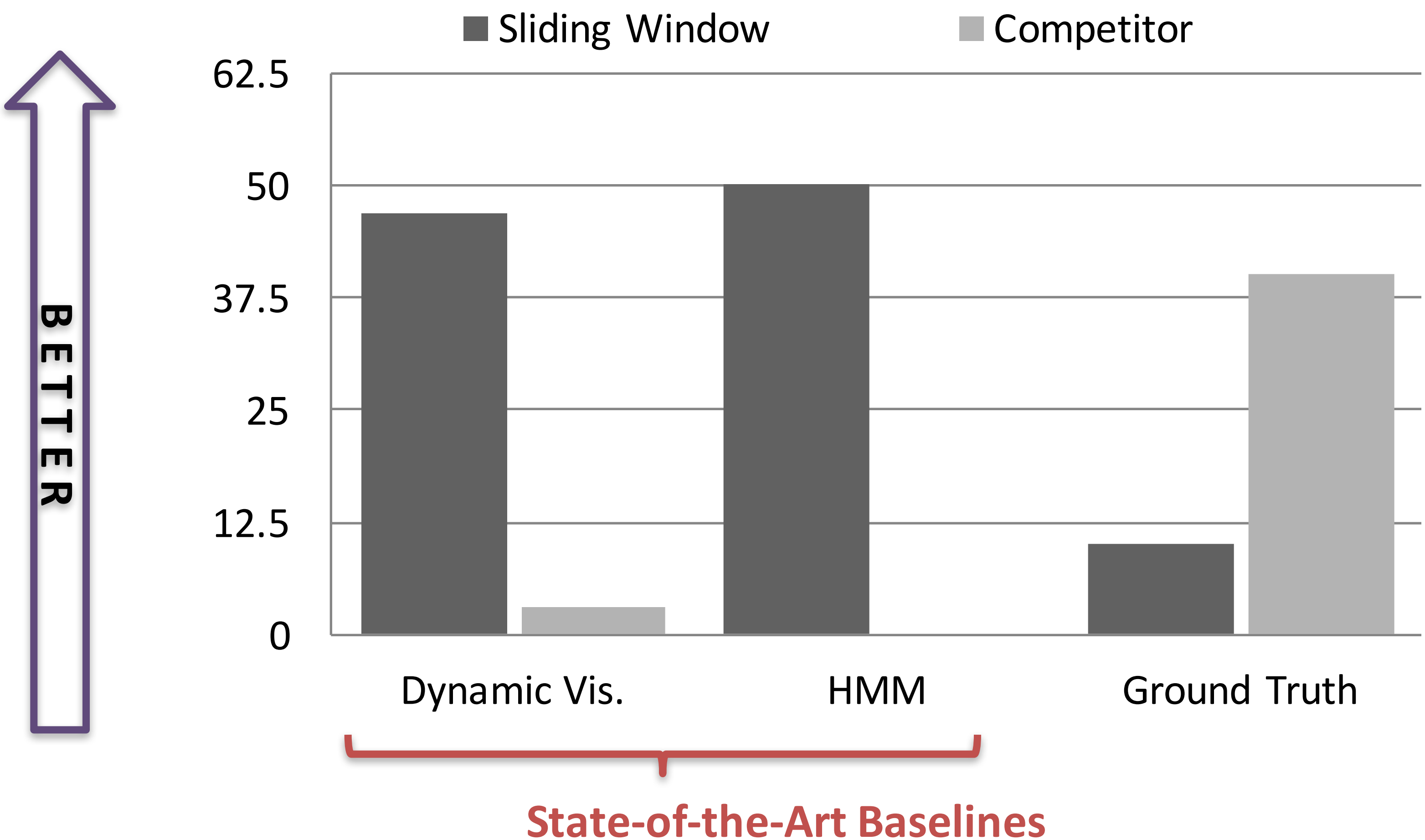
Side-by-Side User Study



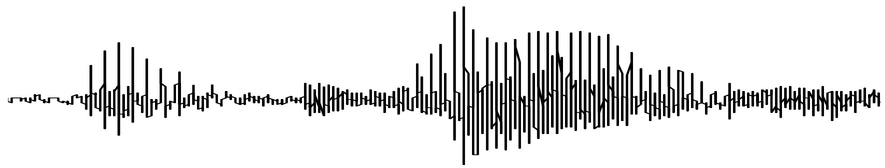
Comparing our approach versus competitor on 50 held-out test sentences.

"A Decision Tree Framework for Spatiotemporal Sequence Prediction"
Kim, Yue, Taylor, Matthews, KDD 2015, http://projects.yisongyue.com/visual_speech

Side-by-Side User Study



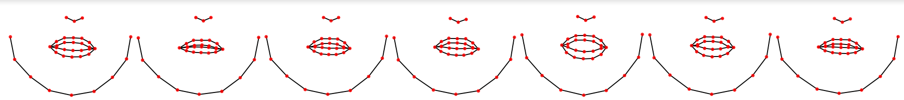
Comparing our approach versus competitor on 50 held-out test sentences.



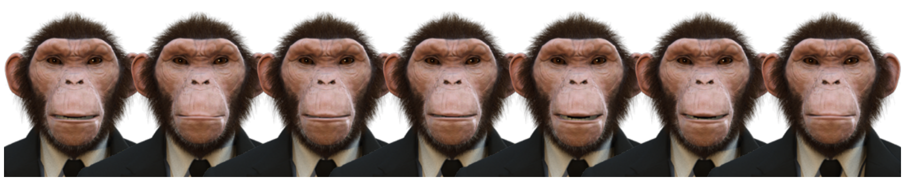
Input Audio

s s s s s ih ih ih g g r r ae ae ae ae f f f

Speech Recognition



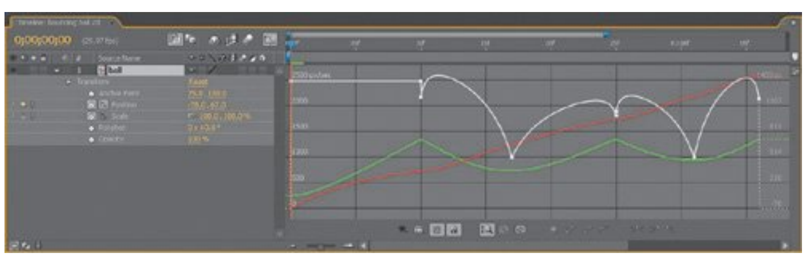
Speech Animation



(chimp rig courtesy of Hao Li)

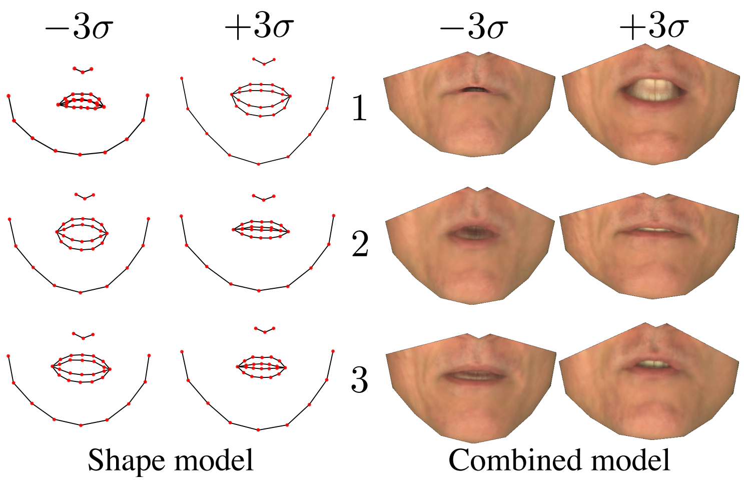
Retargeting

E.g., [Sumner & Popovic 2004]



Editing

Aside: Retargeting



Reference face ➔ target face

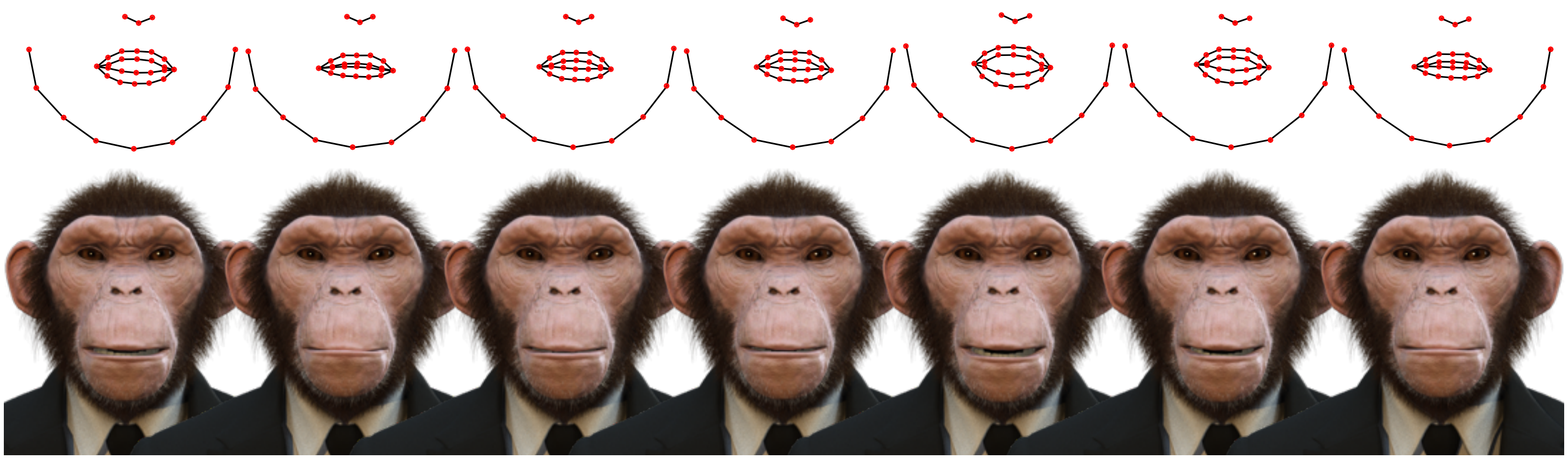
(Semi-)Automatic:

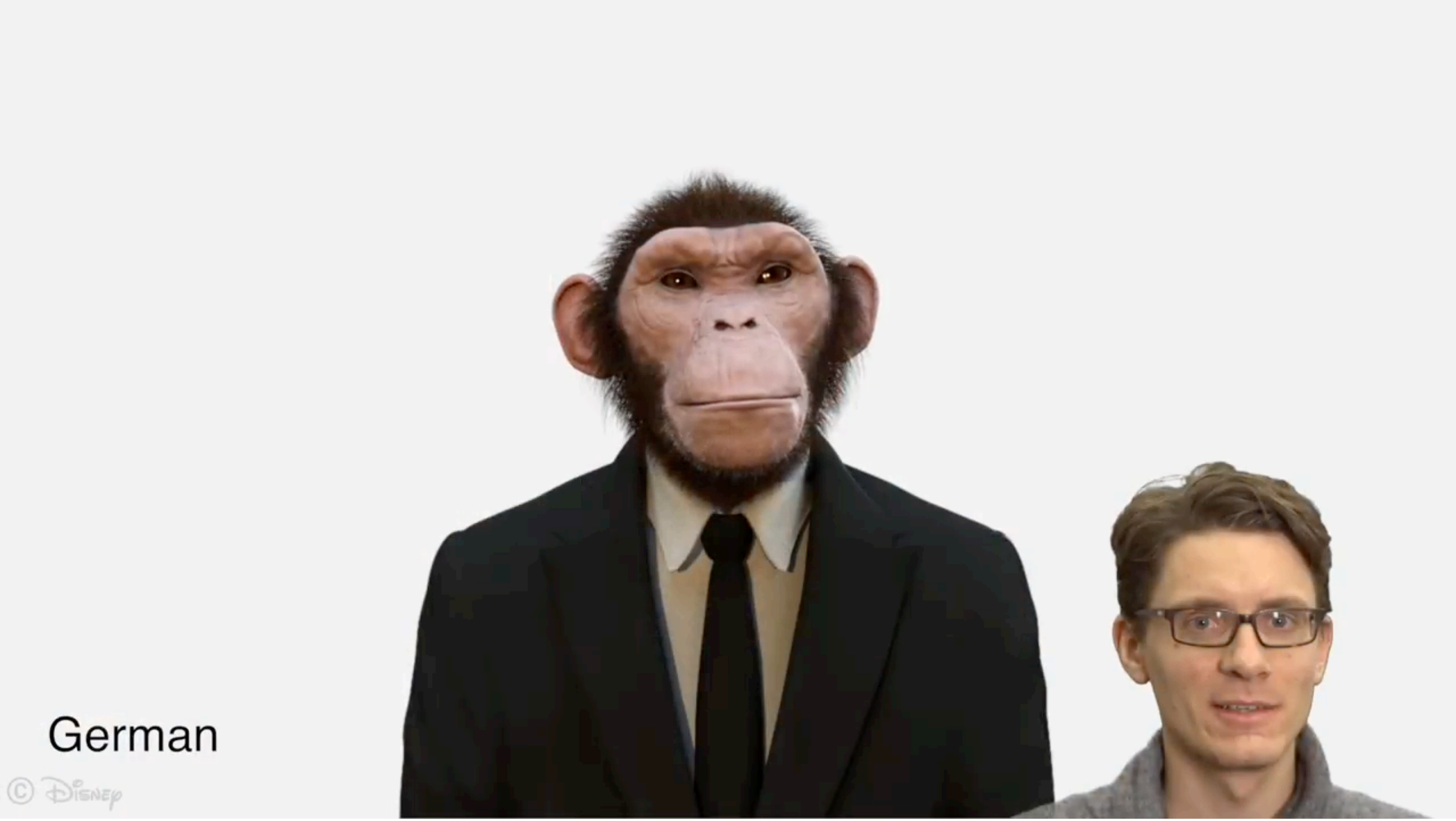
Deformation Transfer [Sumner & Popovic 2004]

Finds linear transform (requires reference pose)

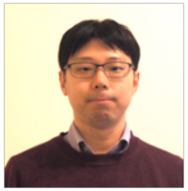
Manual:

Pose basis shapes & linear blending





Sarah Taylor

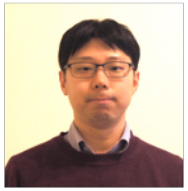


Taehwan Kim

A Decision Tree Framework for Spatiotemporal Sequence Prediction
Taehwan Kim, Yisong Yue, Sarah Taylor, Iain Matthews. KDD 2015

A Deep Learning Approach for Generalized Speech Animation
Sarah Taylor, Taehwan Kim, Yisong Yue, et al. SIGGRAPH 2017

Sarah Taylor



Taehwan Kim

A Decision Tree Framework for Spatiotemporal Sequence Prediction

Taehwan Kim, Yisong Yue, Sarah Taylor, Iain Matthews. KDD 2015

A Deep Learning Approach for Generalized Speech Animation

Sarah Taylor, Taehwan Kim, Yisong Yue, et al. SIGGRAPH 2017

Learning Reductions Recap

- Know how to solve "standard" ML problems
  - Classification, regression, etc.
  - SVMs, logistic regression, decision trees, neural nets, etc.

Many toolkits available!

- "Reduce" complex problems to simple ones?
  - Variable-length trajectories ➔ multivariate regression

Still non-trivial!

- Similar to other reduction problems
  - E.g., NP-complete reductions
  - Some learning reductions have provable guarantees

Other Learning Reductions

- Multiclass → Binary
- Cost-weighted → Unweighted
- Ranking → Binary
- Sequential → Multiclass
- And many more…

http://hunch.net/~jl/projects/reductions/reductions.html

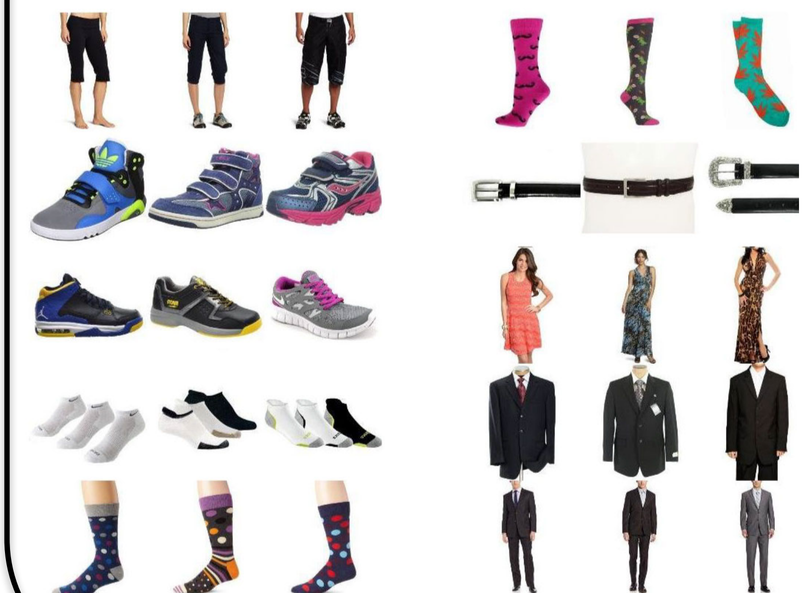
Learning Visual Style

Learning Visual Clothing Style with Heterogeneous Dyadic Co-occurrences

Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, Serge Belongie, ICCV 2015

Visually Compatible

Visually Incompatible



http://vision.cornell.edu/se3/projects/clothing-style/

Training Data

- Ground set of items
  - ~1M items
  - Image of item x
  - Category of item c
    - Coat, belt, pants, socks, etc.

- Pairwise relationships
  - "frequently bought together"
  - Interpret as visually compatible

Training Goal

(ignoring regularization)

Penalizes too far

Penalizes too close

Embedding of image

Embedding of image

$$\underset{\Theta}{\operatorname{argmin}} \sum_{(i,j)\in D} L^{+}\Big(\Phi(x_i),\Phi(x_j)\Big) + \sum_{(i,j)\in \tilde{D}} L^{-}\Big(\Phi(x_i),\Phi(x_j)\Big)$$

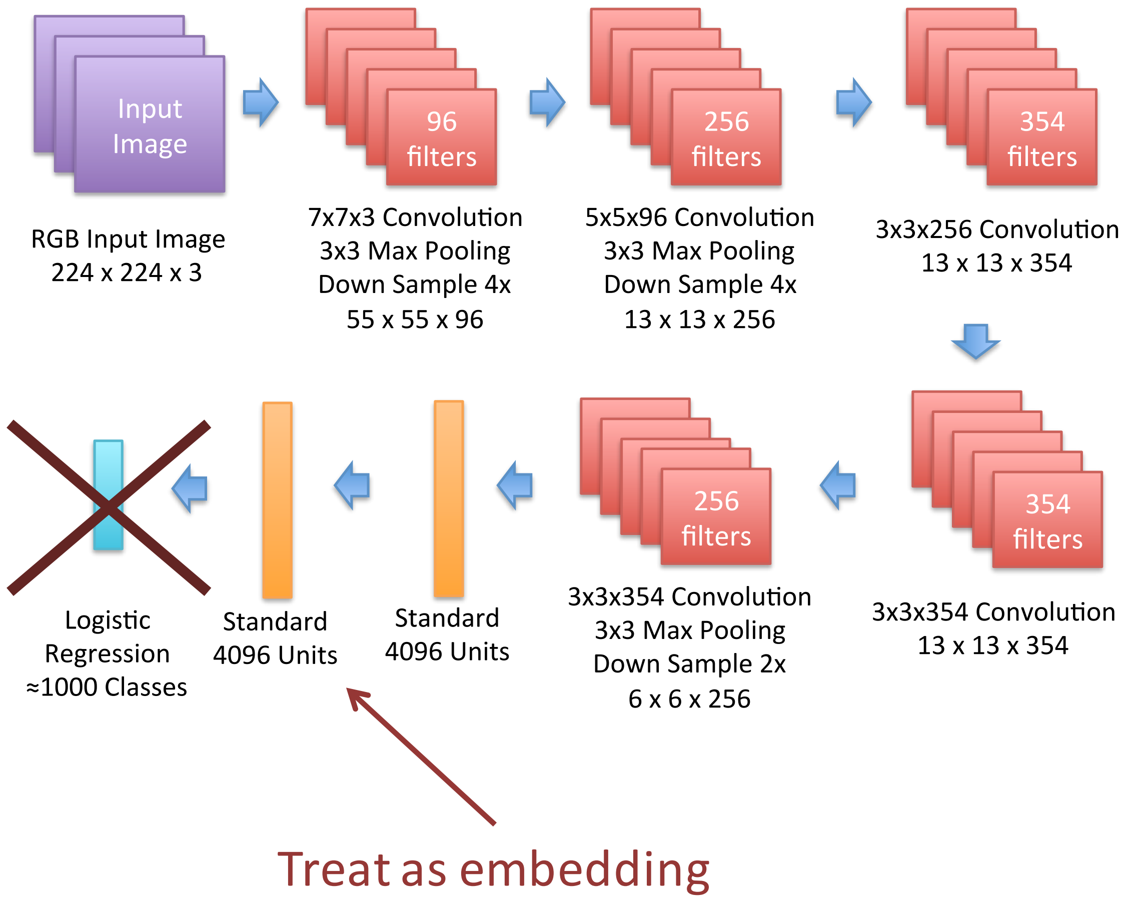
All Model Parameters

Compatible Pairs

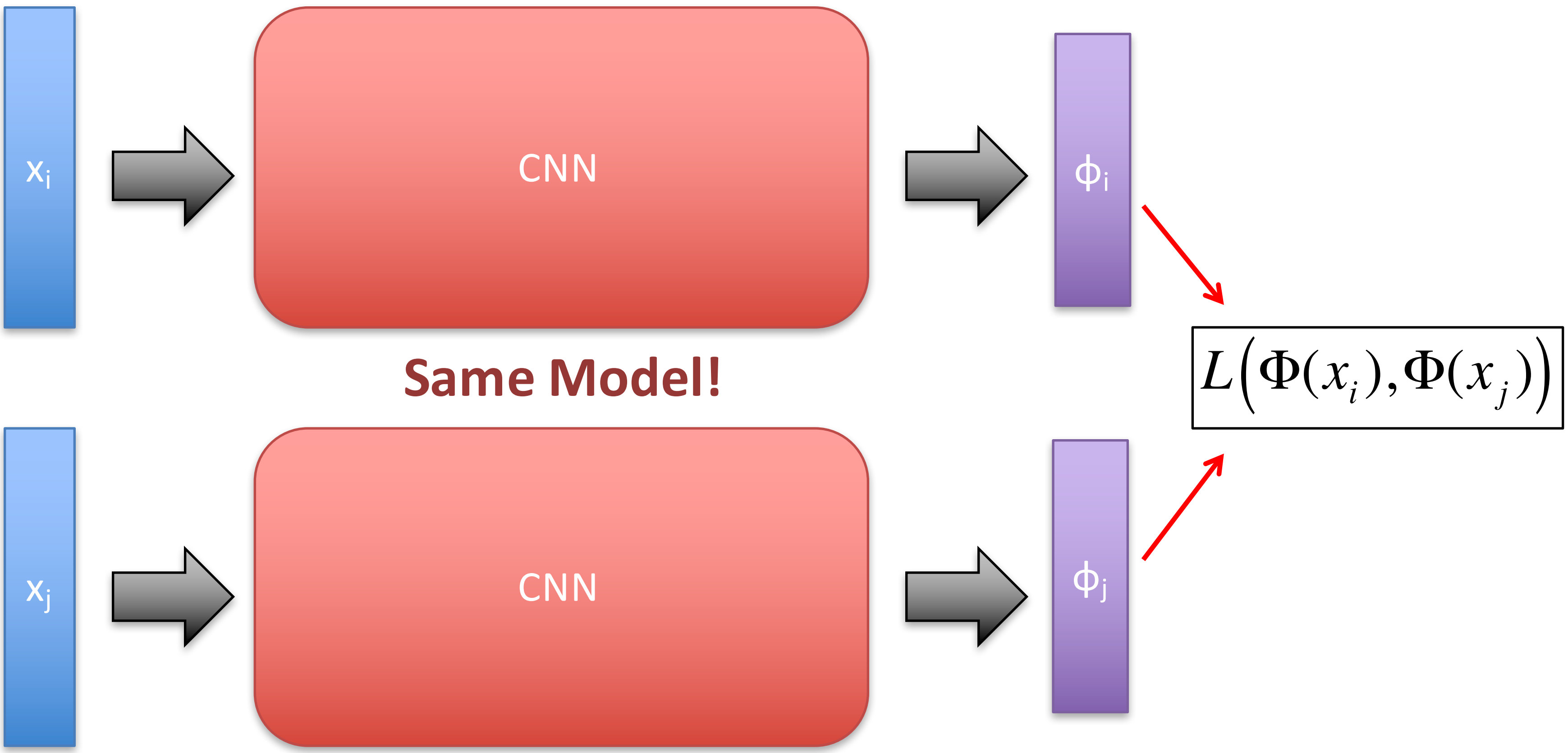
Incompatible Pairs

Only pairs in different categories.

Recall: Convolutional Neural Networks



Siamese Convolutional Neural Networks



More details: http://www.cs.cornell.edu/~kb/publications/SIG15ProductNet.pdf

Recap: Training Goal

Penalizes too far

Penalizes too close

Embedding of image

Embedding of image

$$\underset{\Theta}{\operatorname{argmin}} \sum_{(i,j)\in D} L^{+}\left(\Phi(x_i),\Phi(x_j)\right) + \sum_{(i,j)\in \tilde{D}} L^{-}\left(\Phi(x_i),\Phi(x_j)\right)$$

All Model Parameters

Compatible Pairs

Incompatible Pairs

Only pairs in different categories.

Model Embedding via Siamese Convolutional Neural Network!

Training Details

- Want embedding dimension smaller
  - E.g., 128 rather than 4096
- Need to subsample negative pairs
  - Most items are not frequently bought together
  - Negative component can overwhelm objective

http://www.cs.cornell.edu/~andreas/iccv15.pdf

Suggesting Outfits

Upper Garment

Lower Garment

Footware

http://www.cs.cornell.edu/~andreas/iccv15.pdf

Suggesting Outfits

- Given query item i
  - Embedding $\varphi_i=\Phi(x_i|\Theta)$
  - Category c_i
- For other categories
  - Recommend item with closest embedding φ
- **Not robust to label noise!**

http://www.cs.cornell.edu/~andreas/iccv15.pdf

Label Noise

- Amazon category labels are noisy
  - Eg., some pants mis-categorized as shoes

- Pants are visually very similar

$\Phi($  $)$ $\approx$ $\Phi($  $)$

Pants

Shoes

Mis-categorized!

Making Robust Suggestions

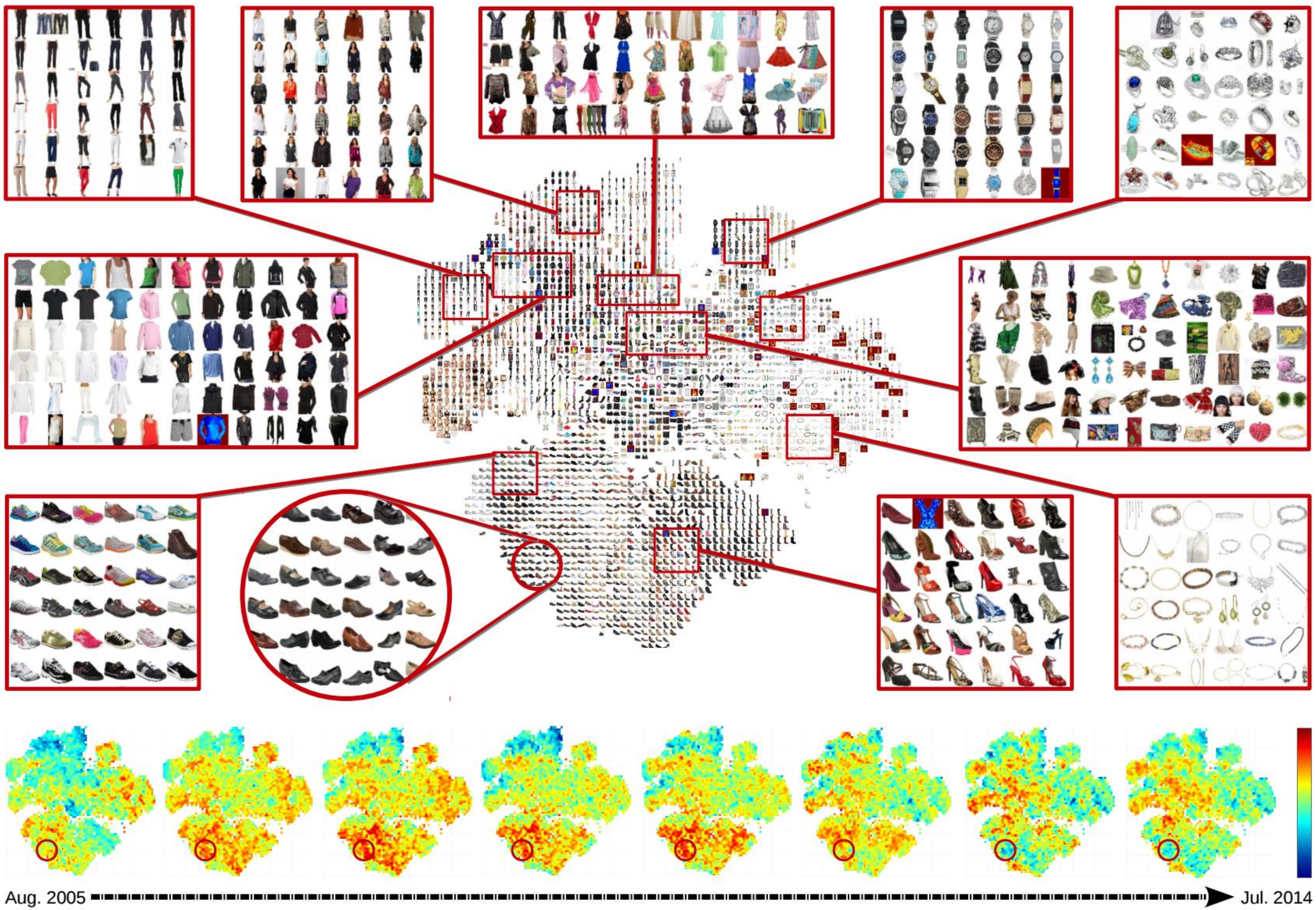
- Mis-categorizations are rare
  - Instead of predicting closest shoe…
  - Predict closest cluster of shoes!
- Preprocessing: cluster every category
- Given input query (category=pants)
  - Find closest cluster center (category=shoes)
  - Output shoes item close to cluster center

http://www.cs.cornell.edu/~andreas/iccv15.pdf

Compute Coherence of Outfit

Least coordinated

Most coordinated

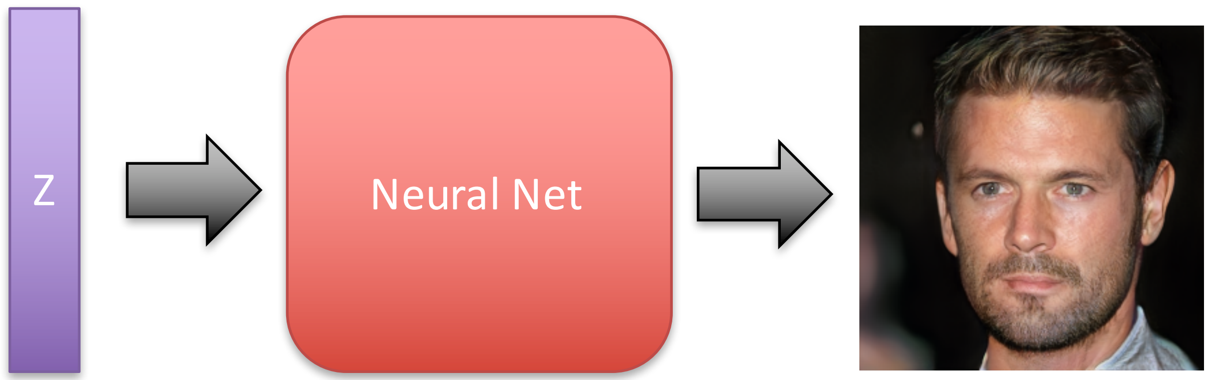


http://cseweb.ucsd.edu/~jmcauley/pdfs/www16a.pdf

(Briefly) Generating Faces

https://arxiv.org/abs/1710.10196 https://www.youtube.com/watch?v=XOxxPcy5Gr4

Latent-Variable Generative Models



Analogous to V matrix in SVD

Analogous to U matrix in SVD

Generative Adversarial Network (GAN)

(discussed further in Deep Generative Models lecture)

https://arxiv.org/abs/1710.10196

Next Few Lectures

- Probabilistic Modeling, HMMs, etc.

- Thursday Recitation: Probability