Structured Output Prediction

CS4780/5780 – Machine Learning
Fall 2013

Thorsten Joachims
Cornell University

Reading:
T. Joachims, T. Hofmann, Yisong Yue, Chun-Nam Yu,
Predicting Structured Objects with Support Vector Machines,
Communications of the ACM, Research Highlight, 52(11):97-104, 2009.
http://mags.acm.org/communications/200911/

Discriminative vs. Generative

Bayes Decision Rule

- $$h_{bayes}(x) = argmax_{y \in \mathrm{Y}} \left[P(Y = y|X = x)\right]$$
  $$= argmax_{y \in \mathrm{Y}} \left[P(X = x|Y = y)P(Y = y)\right]$$

Generative:

- Idea: Make assumptions about $P(X = x|Y = y), P(Y = y)$
- Method: Estimate parameters of the two distributions, then apply Bayes decision rule.

Discriminative:

- Idea: Define set of prediction rules (i.e. hypotheses) H, then search for $h \in H$ that best approximates

  $$h_{bayes}(x) = argmax_{y \in \mathrm{Y}} \left[P(Y = y|X = x)\right]$$

- Method: find $h \in H$ that minimizes training error.

Question: Can we train HMM's discriminately?

Idea for Discriminative Training of HMM

Start:

- $h_{bayes}(x) = argmax_{y \in \mathrm{Y}} \left[P(Y = y | X = x) \right]$
  $= argmax_{y \in \mathrm{Y}} \left[P(X = x | Y = y) P(Y = y) \right]$

Idea:

- Model $P(Y = y | X = x)$ with $\vec{w} \cdot \phi(x, y)$ so that
  $\left(argmax_{y \in \mathrm{Y}} \left[P(Y = y | X = x) \right] \right) = \left(argmax_{y \in \mathrm{Y}} \left[\vec{w} \cdot \phi(x, y) \right] \right)$

Intuition:

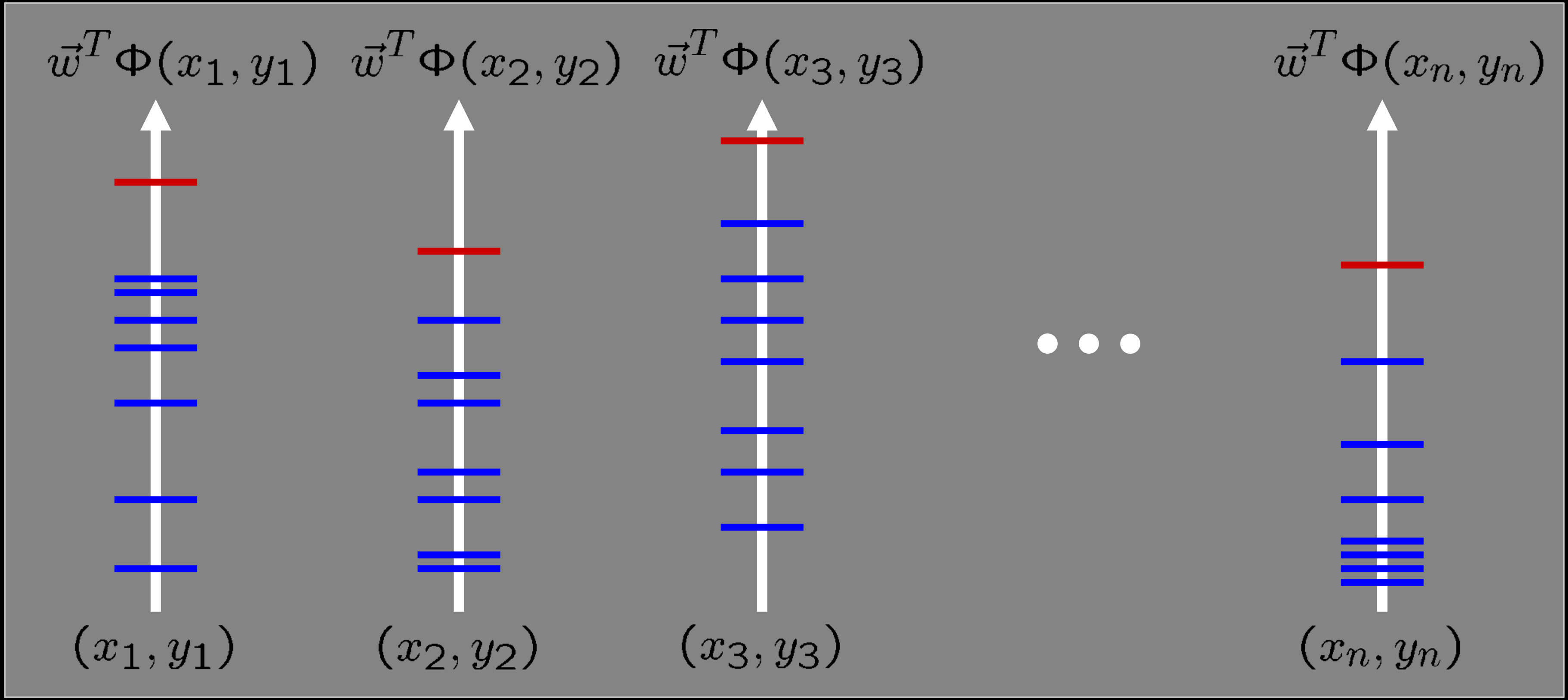
- Tune $\vec{w}$ so that correct y has the highest value of $\vec{w} \cdot \phi(x, y)$
- $\phi(x, y)$ is a feature vector that describes the match between x and y

Training HMMs with Structural SVM

- Define $\phi(x, y)$ so that model is isomorphic to HMM
  - One feature for each possible start state
  - One feature for each possible transition
  - One feature for each possible output in each possible state
  - Feature values are counts

Structural Support Vector Machine

- Joint features $\phi(x, y)$ describe match between *x* and *y*
- Learn weights $\vec{w}$ so that $\vec{w} \cdot \phi(x, y)$ is max for correct *y*



Structural SVM Training Problem

Hard-margin optimization problem:

$$\min_{\vec{w}} \quad \frac{1}{2}\vec{w}^T\vec{w}$$

$$s.t. \quad \forall y \in Y \setminus y_1 : \vec{w}^T\Phi(x_1, y_1) \geq \vec{w}^T\Phi(x_1, y) + 1$$

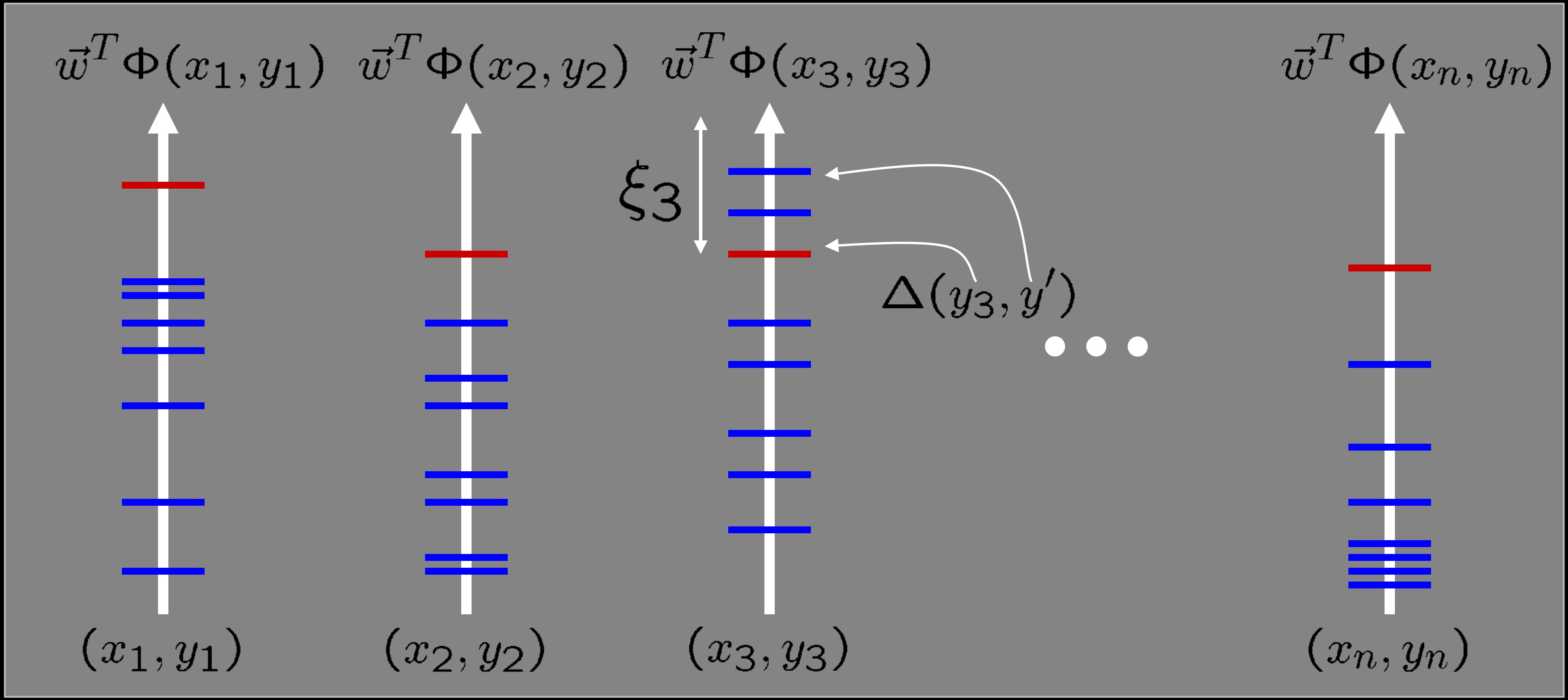
$$\ldots$$

$$\forall y \in Y \setminus y_n : \vec{w}^T\Phi(x_n, y_n) \geq \vec{w}^T\Phi(x_n, y) + 1$$

- Training Set: $(x_1, y_1), \ldots, (x_n, y_n)$
- Prediction Rule: $\mathrm{h}_{svm}(x) = argmax_{y \in \mathrm{Y}} \, [\vec{w} \cdot \phi(x, y)]$
- Optimization:
  - Correct label y_i must have higher value of $\vec{w} \cdot \phi(x, y)$ than any incorrect label y
  - Find weight vector with smallest norm

Soft-Margin Structural SVM

- Loss function $\Delta(y_i, y)$ measures match between target and prediction.



Soft-Margin Structural SVM

Soft-margin optimization problem:

$$\min_{\vec{w},\vec{\xi}} \quad \frac{1}{2}\vec{w}^T\vec{w} + C\sum_{i=1}^{n}\xi_i$$

$$s.t. \quad \forall y \in Y \backslash y_1 : \vec{w}^T\Phi(x_1, y_1) \geq \vec{w}^T\Phi(x_1, y) + \Delta(y_1, y) - \xi_1$$

$$\dots$$

$$\forall y \in Y \backslash y_n : \vec{w}^T\Phi(x_n, y_n) \geq \vec{w}^T\Phi(x_n, y) + \Delta(y_n, y) - \xi_n$$

Lemma: The training loss is upper bounded by

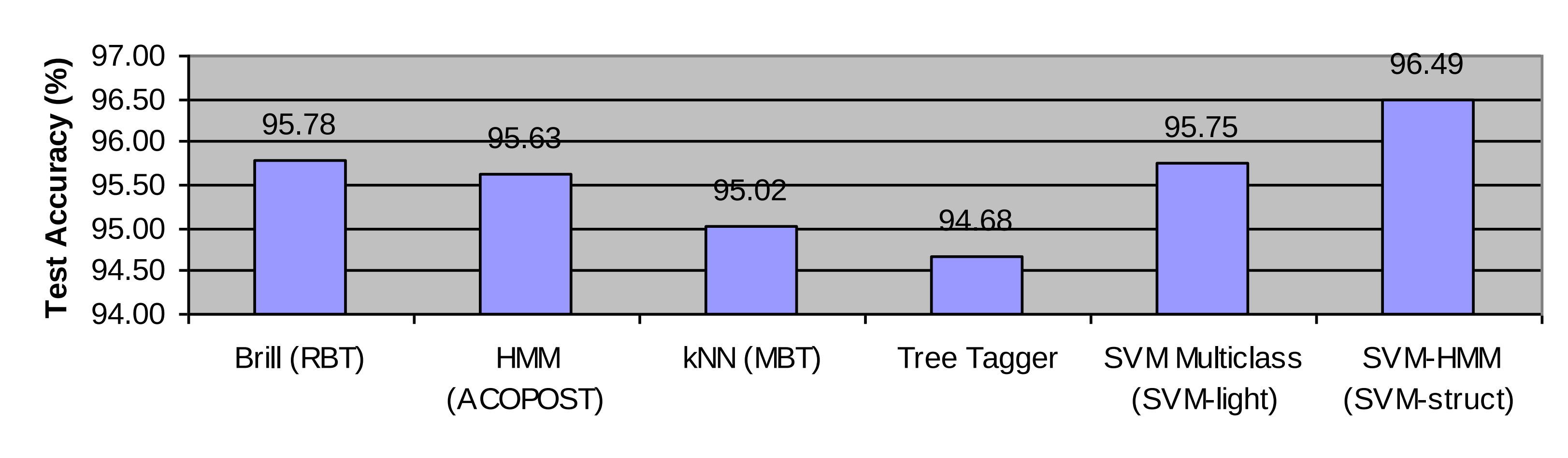
$$Err_S(h) = \frac{1}{n}\sum_{i=1}^{n}\Delta(y_i, h(\vec{x}_i)) \leq \frac{1}{n}\sum_{i=1}^{n}\xi_i$$

Experiment: Part-of-Speech Tagging

- **Task**
  - Given a sequence of words *x*, predict sequence of tags *y*.

    x The dog chased the cat → **y** Det → N → V → Det → N

  - Dependencies from tag-tag transitions in Markov model.
- **Model**
  - Markov model with one state per tag and words as emissions
  - Each word described by ~250,000 dimensional feature vector (all word suffixes/prefixes, word length, capitalization …)
- **Experiment (by Dan Fleisher)**
  - Train/test on 7966/1700 sentences from Penn Treebank



NE Identification

- Identify all named locations, named persons, named organizations, dates, times, monetary

The delegation, which included the commander of the U.N. troops in Bosnia, Lt. Gen. Sir Michael Rose, went to the Serb stronghold of Pale, near Sarajevo, for talks with Bosnian Serb leader Radovan Karadzic.

Este ha sido el primer comentario publico del presidente Clinton respecto a la crisis de Oriente Medio desde que el secretario de Estado, Warren Christopher, decidiera regresar precipitadamente a Washington para impedir la ruptura del proceso de paz tras la violencia desatada en el sur de Libano.

1. Locations
2. Persons
3. Organizations

Figure 1.1 Examples. Examples of correct labels for English text and for Spanish text.

Experiment: Named Entity Recognition

- Data
  - Spanish Newswire articles
  - 300 training sentences
  - 9 tags
    - no-name,
    - beginning and continuation of person name, organization, location, misc name
  - Output words are described by features (e.g. starts with capital letter, contains number, etc.)
- Error on test set (% mislabeled tags):
  - Generative HMM: 9.36%
  - Support Vector Machine HMM: 5.08%

Cutting-Plane Algorithm for Structural SVM

- Input: $(x_1, y_1), \ldots, (x_n, y_n), C, \epsilon$
- $S \leftarrow \emptyset, \vec{w} \leftarrow 0, \vec{\xi} \leftarrow 0$
- REPEAT
  - FOR $i = 1, \ldots, n$
    - compute $\hat{y} = argmax_{y \in Y} \{\Delta(y_i, y) + \vec{w}^T \Phi(x_i, y)\}$
    - IF $(\Delta(y_i, \hat{y}) - \vec{w}^T[\Phi(x_i, y_i) - \Phi(x_i, \hat{y})]) > \xi_i + \epsilon$
      - $S \leftarrow S \cup \{\vec{w}^T[\Phi(x_i, y_i) - \Phi(x_i, \hat{y})] \geq \Delta(y_i, \hat{y}) - \xi_i\}$
      - $[\vec{w}, \vec{\xi}] \leftarrow$ optimize StructSVM over S
    - ENDIF
  - ENDFOR
- UNTIL S has not changed during iteration

Find most violated constraint

Violated by more than ε ?

Add constraint to working set

→ Polynomial Time Algorithm (SVM-struct)

General Problem: Predict Complex Outputs

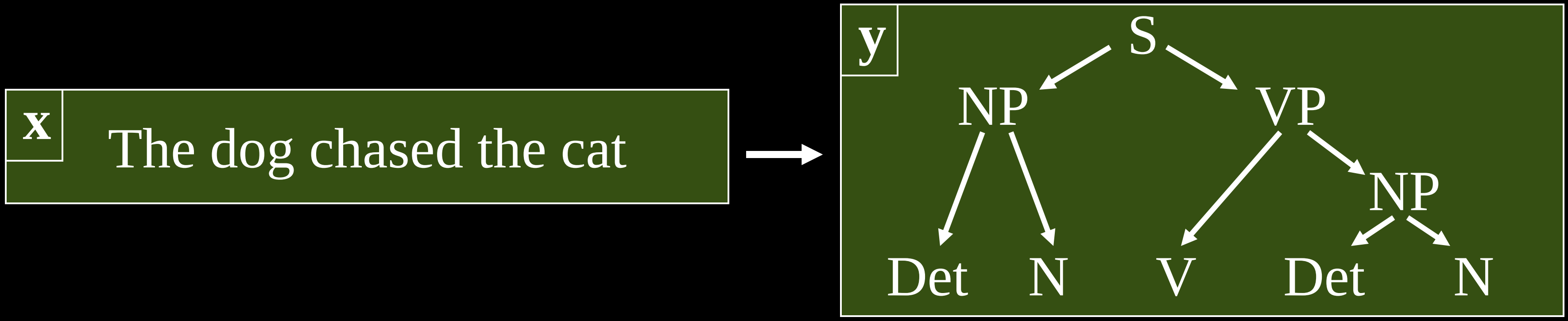
- Supervised Learning from Examples
  - Find function from input space X to output space Y

$$h: X \to Y$$

  such that the prediction error is low.
- Typical
  - Output space is just a single number
    - Classification: -1,+1
    - Regression: some real number
- General
  - Predict outputs that are complex objects

Examples of Complex Output Spaces

- Natural Language Parsing
  - Given a sequence of words *x*, predict the parse tree *y*.
  - Dependencies from structural constraints, since *y* has to be a tree.



Examples of Complex Output Spaces

- Multi-Label Classification
  - Given a (bag-of-words) document x, predict a set of labels *y*.
  - Dependencies between labels from correlations between labels (“iraq” and “oil” in newswire corpus)

x

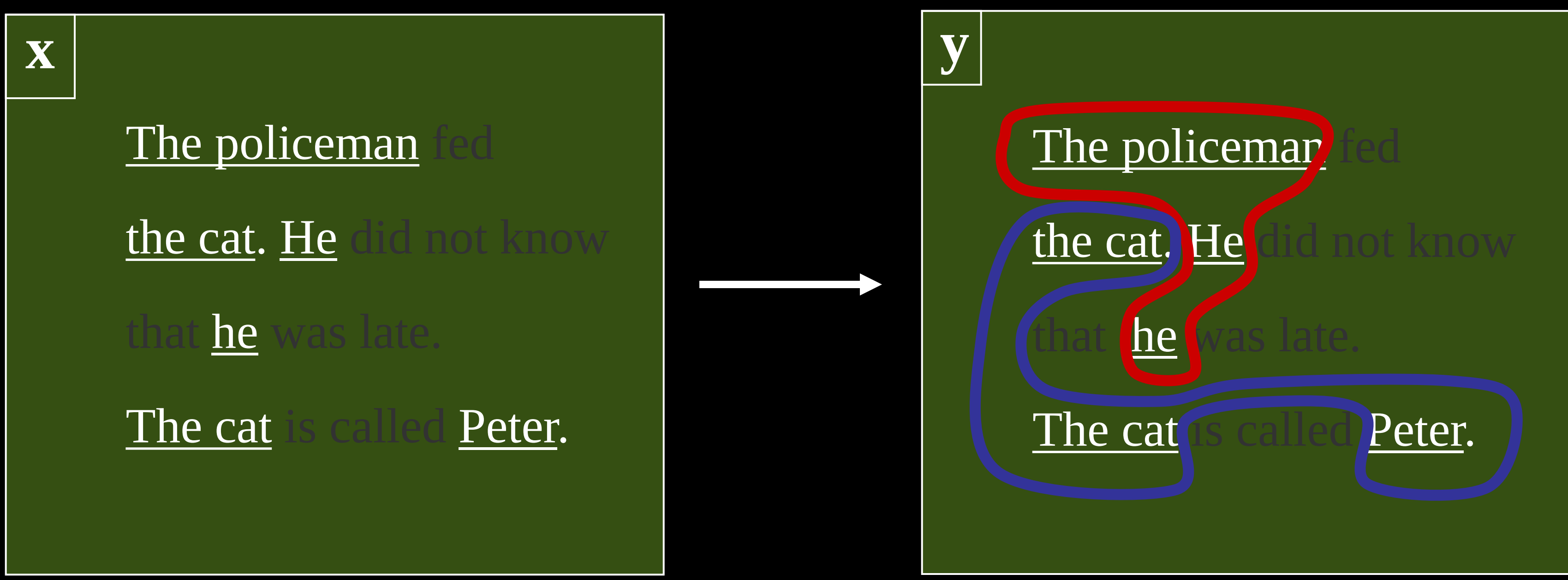
Due to the continued violence in Baghdad, the oil price is expected to further increase. OPEC officials met with …

→

y	
-1	antarctica
-1	benelux
-1	germany
+1	iraq
+1	oil
-1	coal
-1	trade
-1	acquisitions

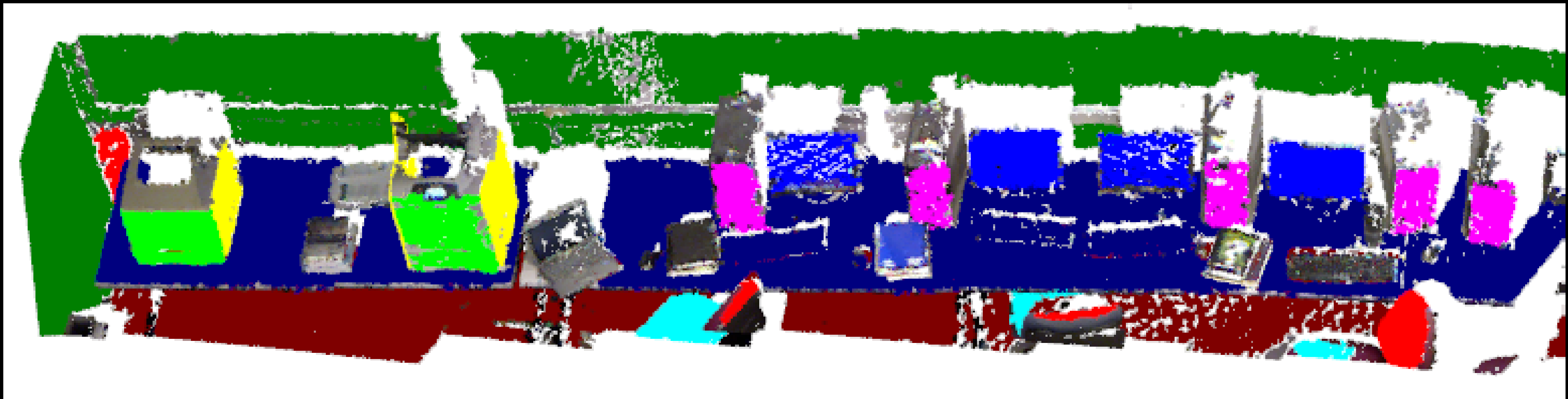
Examples of Complex Output Spaces

- Noun-Phrase Co-reference
  - Given a set of noun phrases *x*, predict a clustering *y*.
  - Structural dependencies, since prediction has to be an equivalence relation.
  - Correlation dependencies from interactions.



Examples of Complex Output Spaces

- Scene Recognition
  - Given a 3D point cloud with RGB from Kinect camera
  - Segment into volumes
  - Geometric dependencies between segments (e.g. monitor usually close to keyboard)



Wrap-Up

Classification

- Discriminative
  - Decision Trees
  - Perceptron
  - Linear SVMs
  - Kernel SVMs
- Generative
  - Multinomial Naïve Bayes
  - Multivariate Naïve Bayes
  - Less Naïve Bayes
  - Linear Discriminant
  - Nearest Neighbor

→ Methods + Theory + Practice

- Other Methods
  - Logical rule learning
  - Online Learning
  - Logistic Regression
  - Neural Networks
  - RBF Networks
  - Boosting
  - Bagging
  - Parametric (Graphical) Models
  - Non-Parametric Models

  - *-Regression
  - *-Multiclass

Structured Prediction

- Discriminative
  - Structural SVMs
- Generative
  - Hidden Markov Model
- Other Methods
  - Maximum Margin Markov Networks
  - Conditional Random Fields
  - Markov Random Fields
  - Bayesian Networks
  - Statistical Relational Learning

→ CS4782 Prob Graphical Models

Unsupervised Learning

- Clustering
  - Hierarchical Agglomerative Clustering
  - K-Means
  - Mixture of Gaussians and EM-Algorithm
- Other Methods
  - Spectral Clustering
  - Latent Dirichlet Allocation
  - Latent Semantic Analysis
  - Multi-Dimensional Scaling
- Other Tasks
  - Outlier Detection
  - Novelty Detection
  - Dimensionality Reduction
  - Non-Linear Manifold Detection

→ CS4850 Math Found for the Information Age

Other Learning Problems and Applications

- Recommender Systems, Search Ranking, etc.
- Reinforcement Learning and Markov Decision Processes
  - CS4758 Robot Learning
- Computer Vision
  - CS4670 Intro Computer Vision
- Natural Language Processing
  - CS4740 Intro Natural Language Processing

Other Machine Learning Courses at Cornell

- INFO 3300 - New course by David Mimno
- CS 4700 - Introduction to Artificial Intelligence
- CS 4780/5780 - Machine Learning
- CS 4758 - Robot Learning
- CS 4782 - Probabilistic Graphical Models
- OR 4740 - Statistical Data Mining
- CS 6756 - Advanced Topics in Robot Learning: 3D Perception
- CS 6780 - Advanced Machine Learning
- CS 6784 - Advanced Topics in Machine Learning
- ORIE 6740 - Statistical Learning Theory for Data Mining
- ORIE 6750 - Optimal learning
- ORIE 6780 - Bayesian Statistics and Data Analysis
- ORIE 6127 - Computational Issues in Large Scale Data-Driven Models
- BTRY 6502 - Computationally Intensive Statistical Inference
- MATH 7740 - Statistical Learning Theory