

# More data means less inference: A pseudo-max approach to structured learning

W10

David Sontag  
Microsoft Research

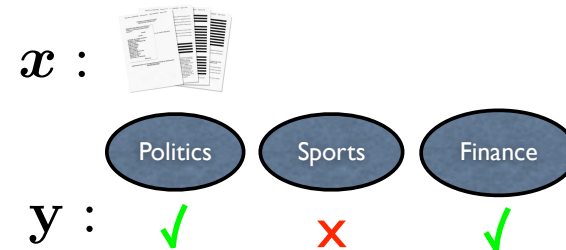
Ofer Meshi  
Hebrew University

Tommi Jaakkola  
MIT

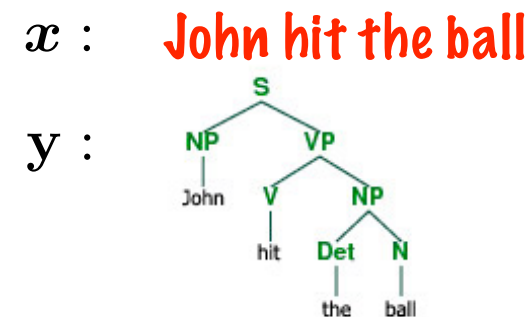
Amir Globerson  
Hebrew University

## Structured prediction

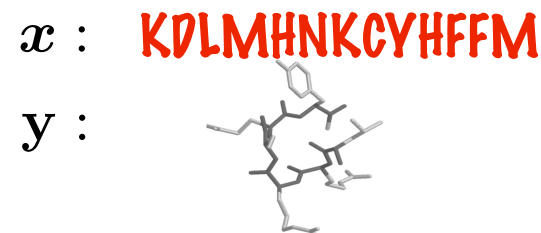
- Multi-label prediction:



- Parsing of natural language:



- Protein side-chain placement:



# More data means less inference: A pseudo-max approach to structured learning

W10

David Sontag  
Microsoft Research

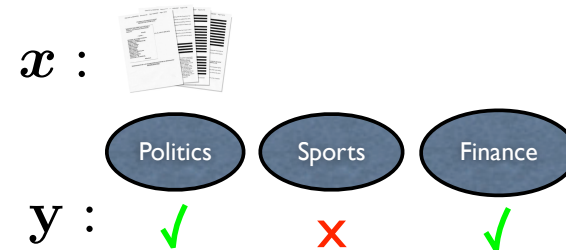
Ofer Meshi  
Hebrew University

Tommi Jaakkola  
MIT

Amir Globerson  
Hebrew University

## Structured prediction

- Multi-label prediction:



- Each prediction task is specified by a feature function  $f(x, y)$  and a weight vector  $w$ .
- Prediction is given by 
$$y \leftarrow \operatorname{argmax}_{\hat{y} \in Z} w \cdot f(x, \hat{y})$$
- Typically decomposes as  $f(x, y) = \sum_c f_c(x, y_c)$ , where  $c$  is a small set of variables

# More data means less inference: A pseudo-max approach to structured learning

W10

David Sontag  
Microsoft Research

Ofer Meshi  
Hebrew University

Tommi Jaakkola  
MIT

Amir Globerson  
Hebrew University

---

## Learning problem (separable setting)

- Given training data  $\{\mathbf{x}_m, \mathbf{y}_m\}_{m=1}^M$
- Assume that there exist “true” parameters  $\mathbf{w}$  such that
$$\mathbf{y}_m \leftarrow \operatorname{argmax}_{\hat{\mathbf{y}} \in Z} \mathbf{w} \cdot f(\mathbf{x}_m, \hat{\mathbf{y}}) \quad \text{for all } m$$
- Structured perceptron, stochastic subgradient, cutting-plane, ...  
**All repeatedly do prediction during learning - very slow!**
- Is there some way to circumvent prediction during learning?
- We give an efficient learning algorithm which, when distribution of training examples is sufficiently “nice”, is asymptotically consistent

# More data means less inference: A pseudo-max approach to structured learning

W10

David Sontag  
Microsoft Research

Ofer Meshi  
Hebrew University

Tommi Jaakkola  
MIT

Amir Globerson  
Hebrew University

---

## The Pseudo-max Method

- **Exact:**  $\{\mathbf{w} \cdot f(\mathbf{x}^m, \mathbf{y}^m) > \mathbf{w} \cdot f(\mathbf{x}^m, \mathbf{y}), \forall m \text{ and } \mathbf{y} \neq \mathbf{y}^m\}$
- **Pseudo-max:**  $\{\mathbf{w} \cdot f(\mathbf{x}^m, \mathbf{y}^m) > \mathbf{w} \cdot f(\mathbf{x}^m, \mathbf{y}_{-i}^m, y_i), \forall m \text{ and } i, y_i \neq y_i^m\}$
- Very small number of constraints:  $M \times \# \text{Vars} \times \# \text{Values}$
- Does this ever work?
  - Yes, under some conditions on  $p(\mathbf{x})$ .
  - When  $f$  corresponds to a pairwise Markov random field, these constraints suffice to identify  $\mathbf{w}^*$ .
- We also show how to apply to non-separable setting
- **Very fast**, and gives good results for multi-label prediction and protein side-chain placement