

# SparseMAP: DIFFERENTIABLE SPARSE STRUCTURED INFERENCE

Presented by **Vlad Niculae**

Joint work with André FT Martins

Mathieu Blondel

Claire Cardie

poster #66 tonight



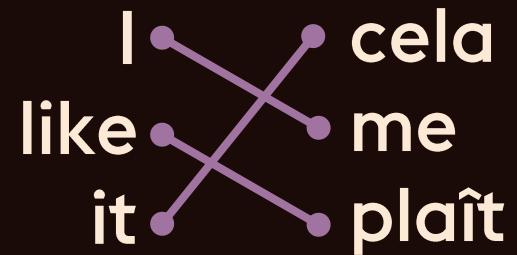
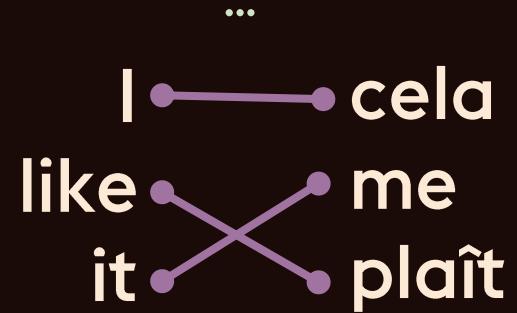
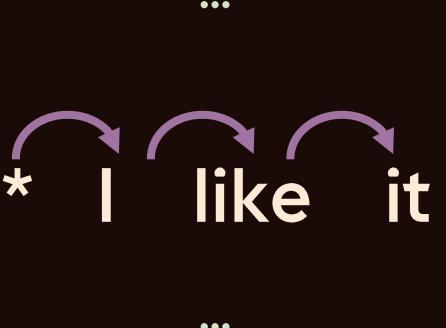
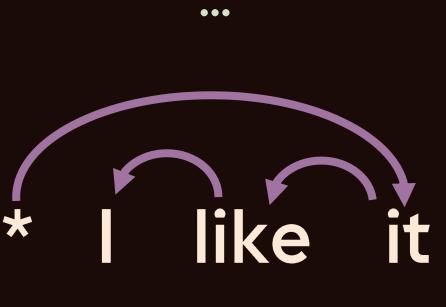
[github.com/vene/sparsemap](https://github.com/vene/sparsemap)

# Structured Inference



...

# Structured Inference



# Structured Inference

...  
PRON   VERB   NOUN  
I   like   it

\*   ...  
I   like   it

PRON   VERB   PRON  
I   like   it  
...

\*   ...  
I   like   it

PRON   VERB   ADJ  
I   like   it  
...  
...

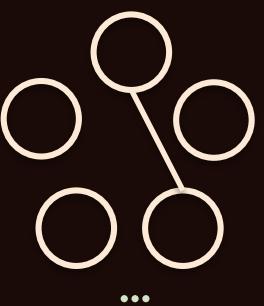
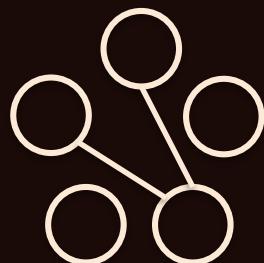
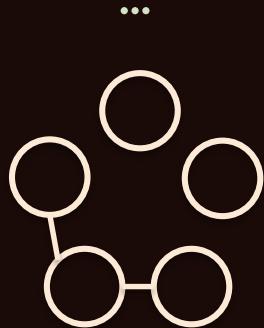
\*   ...  
I   like   it

...  
I   cela  
like   me  
it   plaît

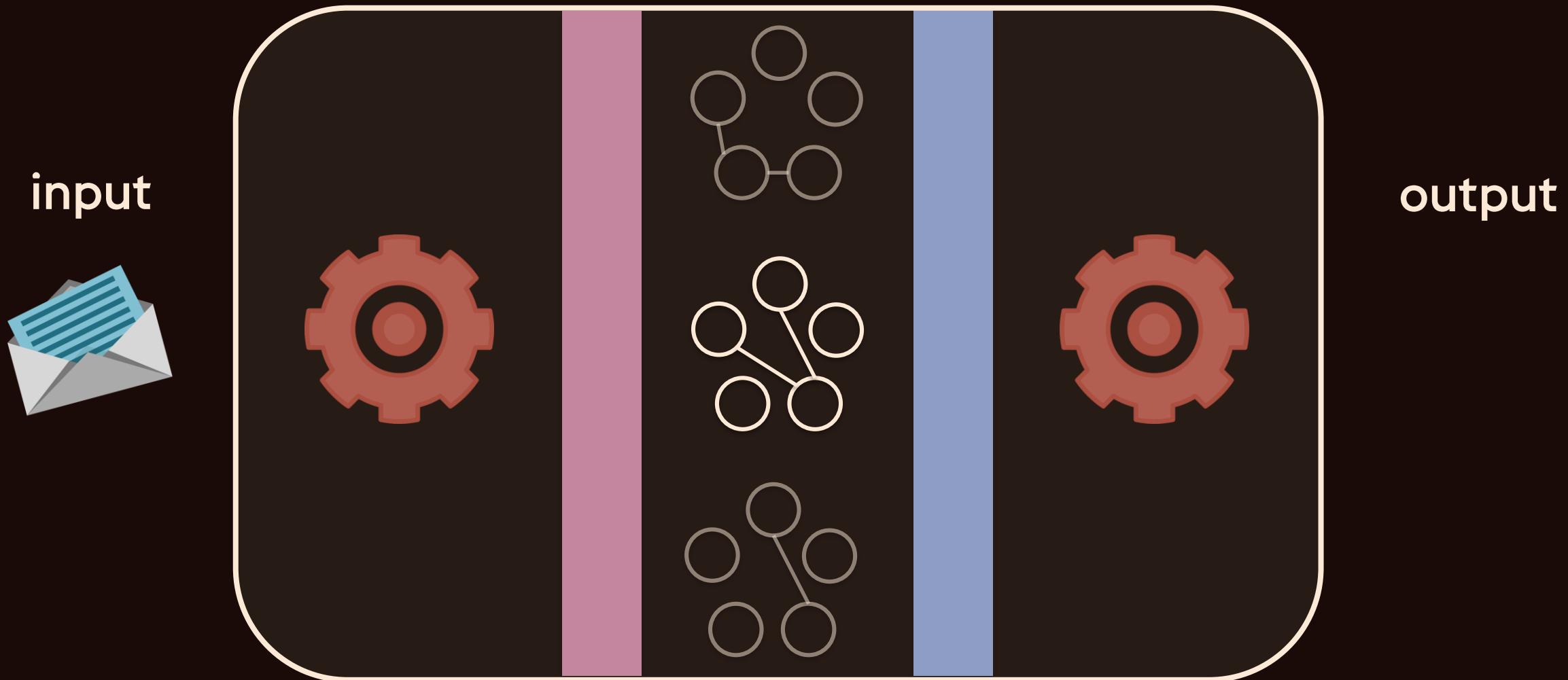
I   cela  
like   me  
it   plaît

...  
I   cela  
like   me  
it   plaît

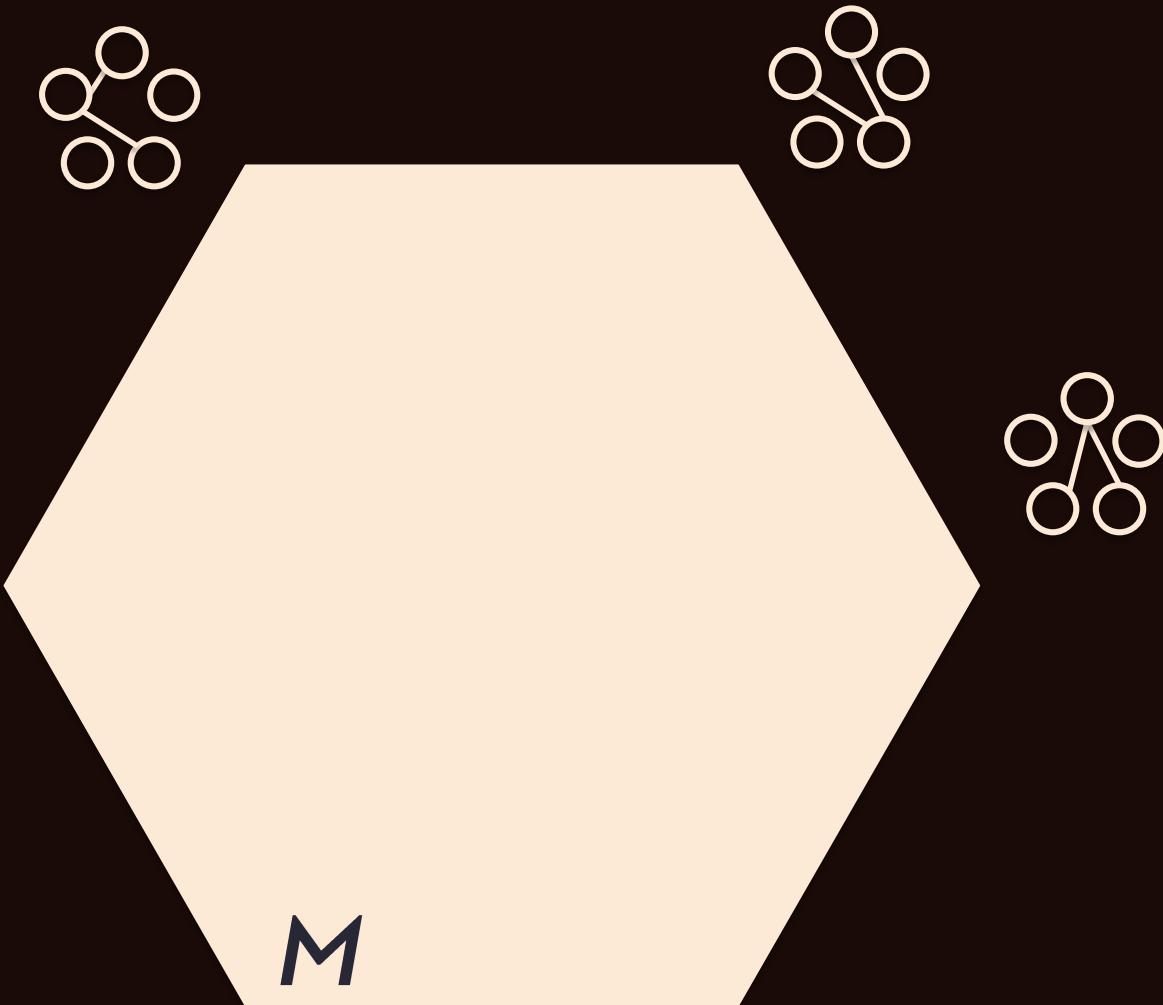
# Structured Inference



# (Latent) Structured Inference



$M = \text{conv}(Y)$  where  $Y = \{ \quad , \quad , \quad \}$



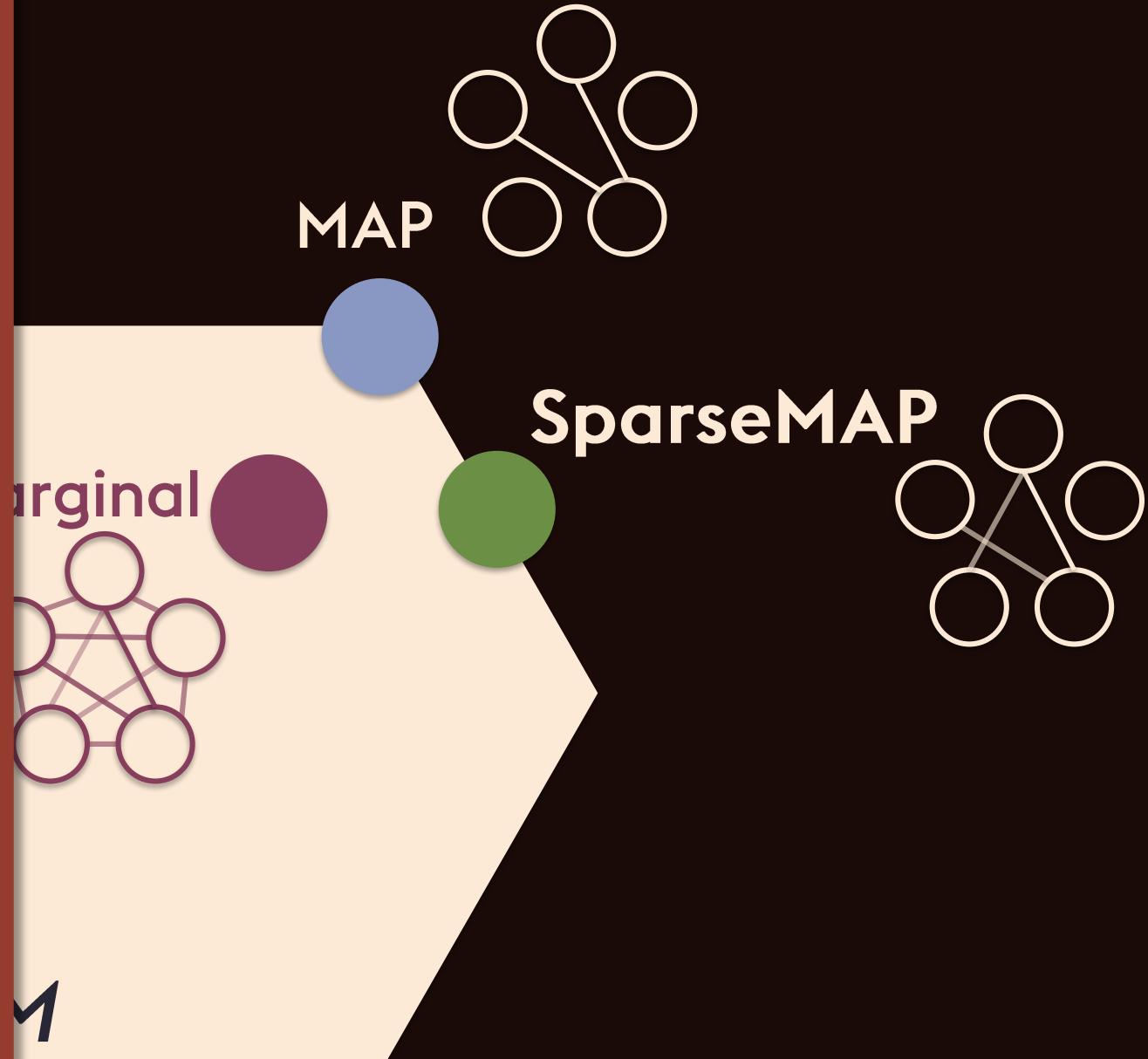
# SparseMAP

Efficient & simple to:

- compute
- back-propagate

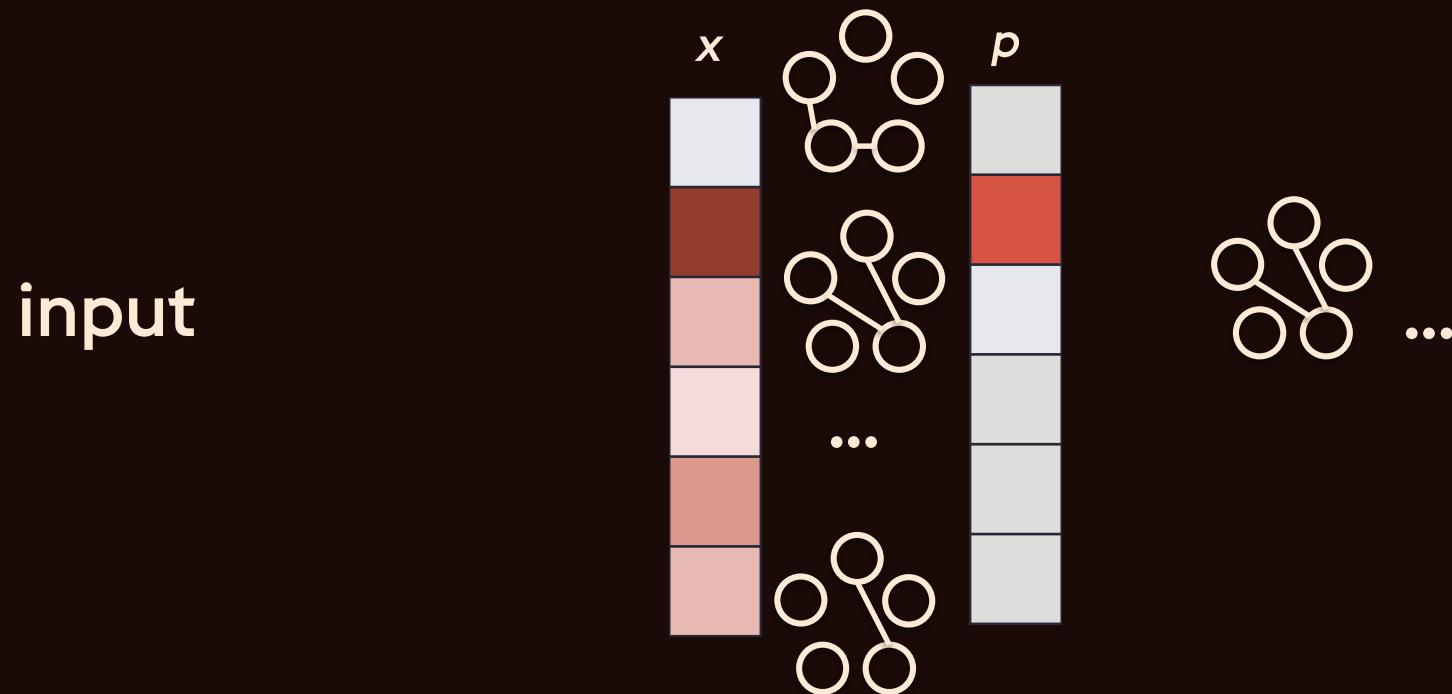
Useful as:

- hidden layer
- output layer



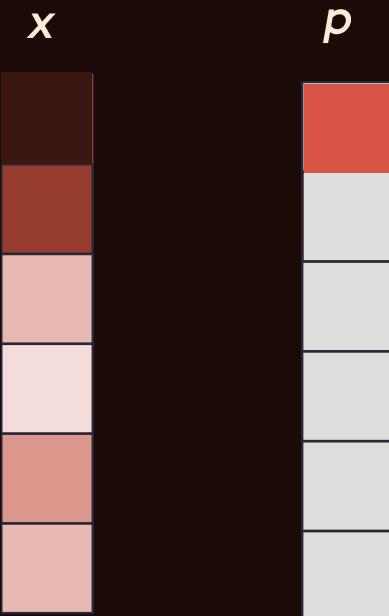
# Deriving SparseMAP

# Structured Inference as argmax



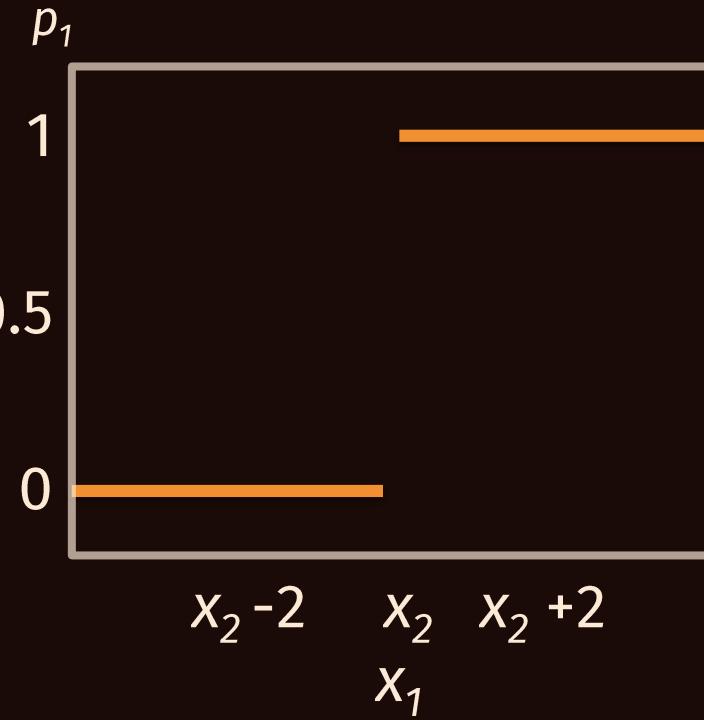
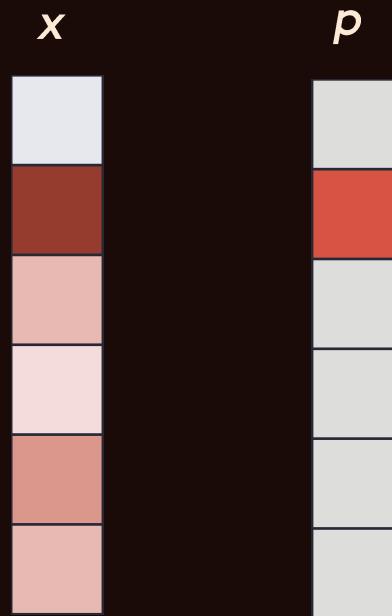
# argmax

input



$\partial p / \partial x ?$

input

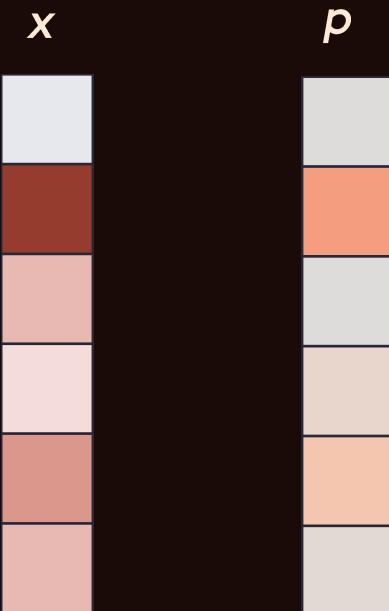


$\partial p / \partial x ?$

# **argmax → softmax**

$$p_i = \exp x_i / Z$$

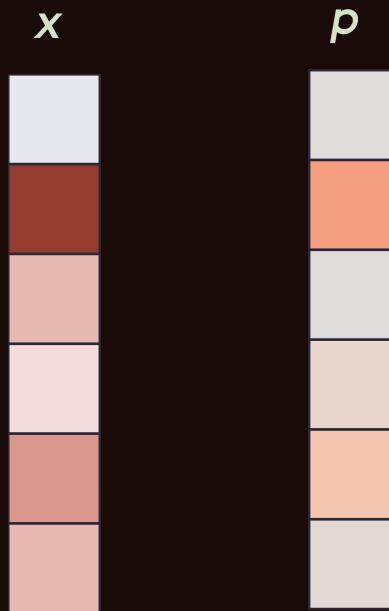
input



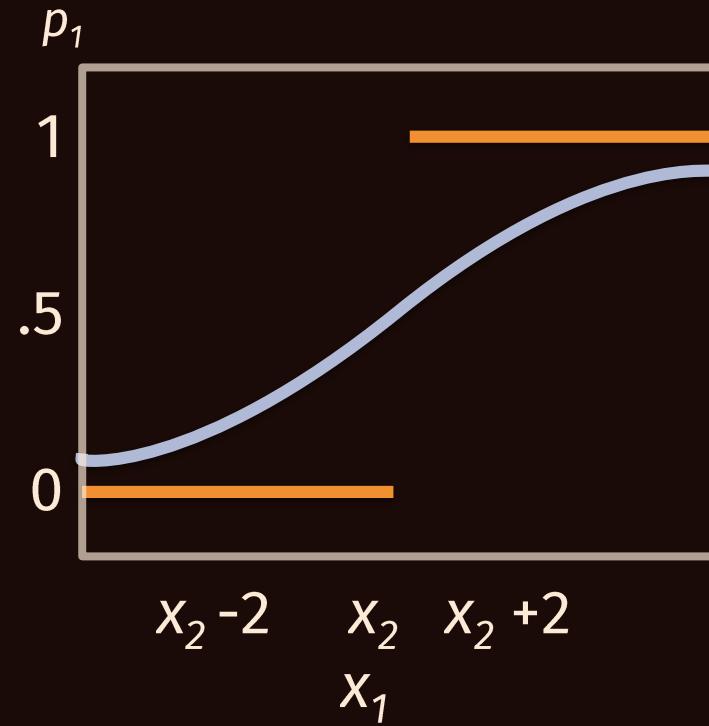
$$\partial p / \partial x ?$$

# argmax → softmax

input



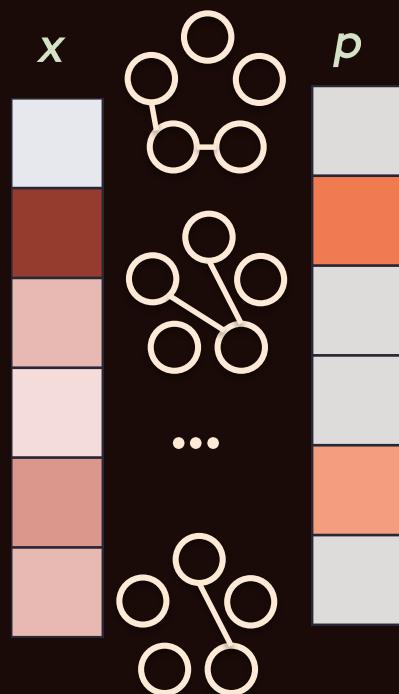
$\partial p / \partial x ?$



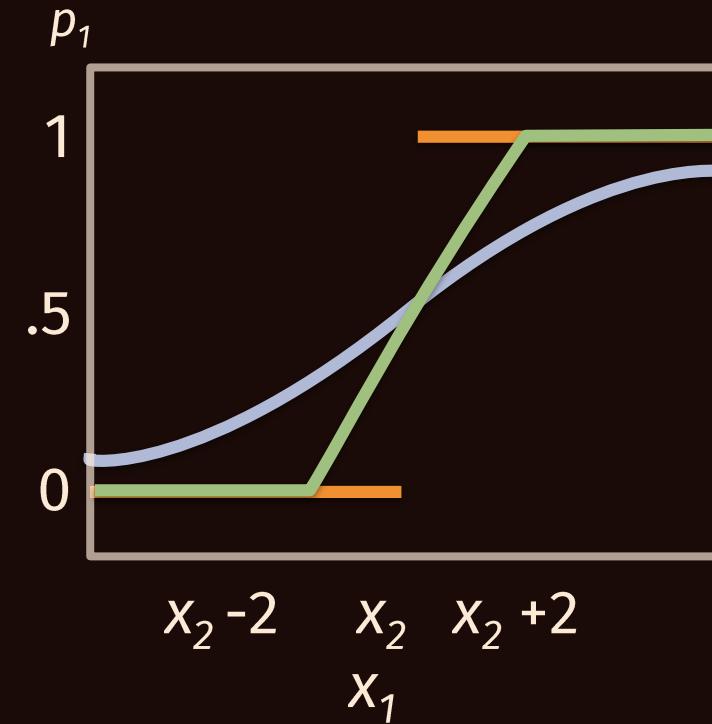
# **argmax → softmax → sparsemax**

**$\dim(x)$  = number of possible structures!**

(exponentially large)



$\partial p / \partial x ?$



[Martins and Astudillo, 2016]  
[Niculae and Blondel, 2017]

$$\textbf{\textit{x}} \in \mathbb{R}^k = \textbf{\textit{A}}^\top \boldsymbol{\eta} \in \mathbb{R}^d \qquad k \gg d$$

$$\mathbf{x} = \mathbf{A}^T \boldsymbol{\eta}$$

\* I like it



The diagram illustrates a sequence of words: an asterisk (\*), followed by the letters I, like, and it. Purple curved arrows connect the word I to the word like, and the word like to the word it, creating a visual link between these three words.

$$\mathbf{x} = \mathbf{A}^T \boldsymbol{\eta}$$

\* I like it

$$\mathbf{A} = \left[ \begin{array}{cccc|cc} 1 & 0 & 0 & 0 & * & \rightarrow I \\ 0 & 1 & 1 & 0 & like & \rightarrow I \\ 0 & 0 & 0 & 0 & it & \rightarrow I \\ \hline 0 & 1 & 1 & 0 & * & \rightarrow like \\ 1 & \dots & 0 & 0 & I & \rightarrow like \\ 0 & 0 & 0 & 0 & it & \rightarrow like \\ \hline 0 & 0 & 0 & 0 & * & \rightarrow it \\ 0 & 1 & 0 & 0 & I & \rightarrow it \\ 1 & 0 & 1 & 0 & like & \rightarrow it \end{array} \right]$$

$\boldsymbol{\eta} = \begin{bmatrix} .3 \\ .8 \\ -.5 \\ .2 \\ -.1 \\ -.2 \\ .7 \\ .6 \\ .1 \end{bmatrix}$

$$\mathbf{x} = \mathbf{A}^T \boldsymbol{\eta}$$

\* I like it

I cela  
like me  
it plait

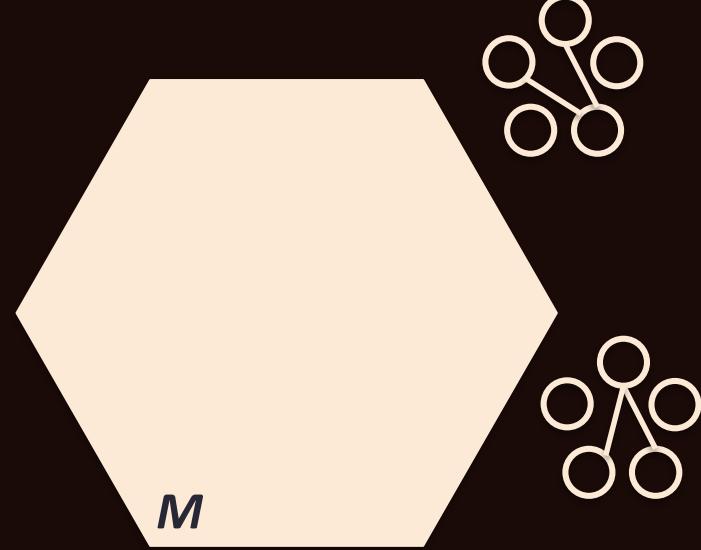
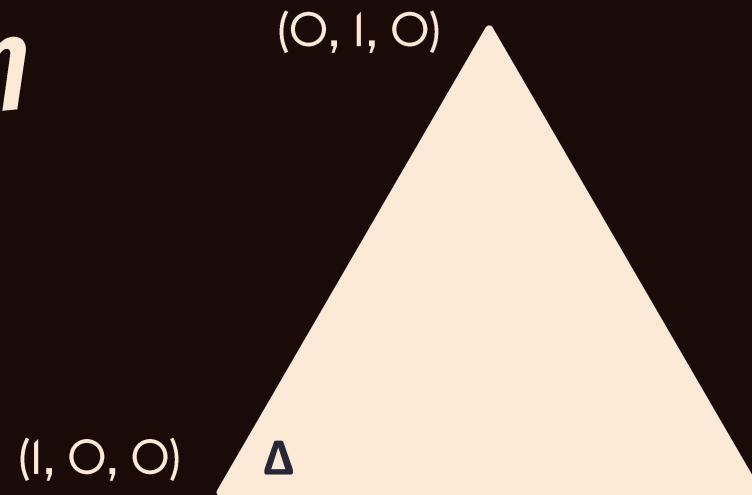
$$\mathbf{A} = \left[ \begin{array}{cccc|cc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & \dots & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{array} \right] \quad \begin{array}{l} * \rightarrow I \\ \text{like} \rightarrow I \\ \text{it} \rightarrow I \\ \hline * \rightarrow \text{like} \\ I \rightarrow \text{like} \\ \text{it} \rightarrow \text{like} \\ \hline * \rightarrow \text{it} \\ I \rightarrow \text{it} \\ \text{like} \rightarrow \text{it} \end{array}$$

$$\boldsymbol{\eta} = \begin{pmatrix} .3 \\ .8 \\ -.5 \\ .2 \\ -.1 \\ -.2 \\ .7 \\ .6 \\ .1 \end{pmatrix}$$

$$\mathbf{A} = \left[ \begin{array}{cccc|cc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & \dots & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{array} \right] \quad \begin{array}{l} I - cela \\ I - me \\ I - plait \\ \hline \text{like} - cela \\ \text{like} - me \\ \text{like} - plait \\ \hline \text{it} - cela \\ \text{it} - me \\ \text{it} - plait \end{array}$$

$$\boldsymbol{\eta} = \begin{pmatrix} .3 \\ .8 \\ -.5 \\ .2 \\ -.1 \\ -.2 \\ .7 \\ .6 \\ .1 \end{pmatrix}$$

$$x = A^\top \eta$$



$$\mu = Ap$$

~

$$\begin{aligned} & \operatorname{argmax} \langle x, p \rangle \\ \text{s.t. } & p \in \Delta \end{aligned}$$

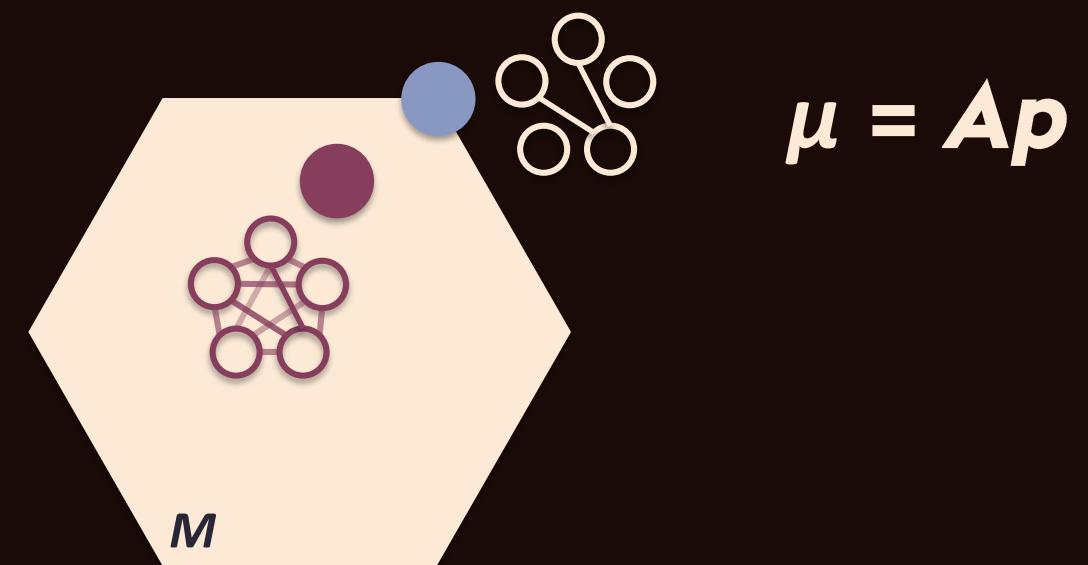
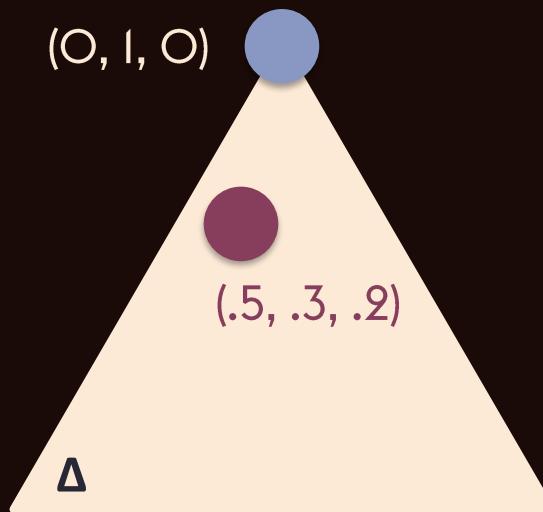
MAP

$$\begin{aligned} & \operatorname{argmax} \langle \eta, \mu \rangle \\ \text{s.t. } & \mu \in M \end{aligned}$$

$$p^* = e_i \text{ where } i = \operatorname{argmax}(x)$$

~

$$x = A^T \eta$$



$$\mu = Ap$$

# MAP inference:

Maximum spanning tree  
(Chu-Liu/Edmonds)

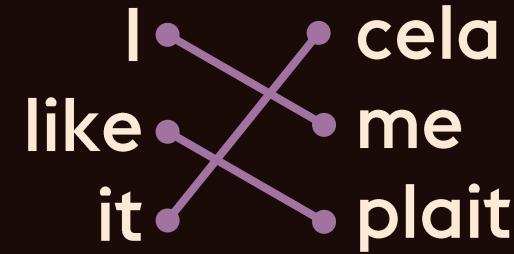


$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ \hline 0 & 1 & 1 \\ 1 & \dots & 0 & 0 & \dots \\ 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$* \rightarrow I$   
 $\text{like} \rightarrow I$   
 $\text{it} \rightarrow I$   
 $* \rightarrow \text{like}$   
 $I \rightarrow \text{like}$   
 $\text{it} \rightarrow \text{like}$   
 $* \rightarrow \text{it}$   
 $I \rightarrow \text{it}$   
 $\text{like} \rightarrow \text{it}$

$$\eta = \begin{bmatrix} .3 \\ .8 \\ -.5 \\ .2 \\ -.1 \\ -.2 \\ .7 \\ .6 \\ .1 \end{bmatrix}$$

# Hungarian algorithm



$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 0 & 0 & 1 \\ 1 & \dots & 0 & 0 & \dots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

$I - \text{cela}$   
 $I - \text{me}$   
 $I - \text{plait}$   
 $\text{like} - \text{cela}$   
 $\text{like} - \text{me}$   
 $\text{like} - \text{plait}$   
 $\text{it} - \text{cela}$   
 $\text{it} - \text{me}$   
 $\text{it} - \text{plait}$

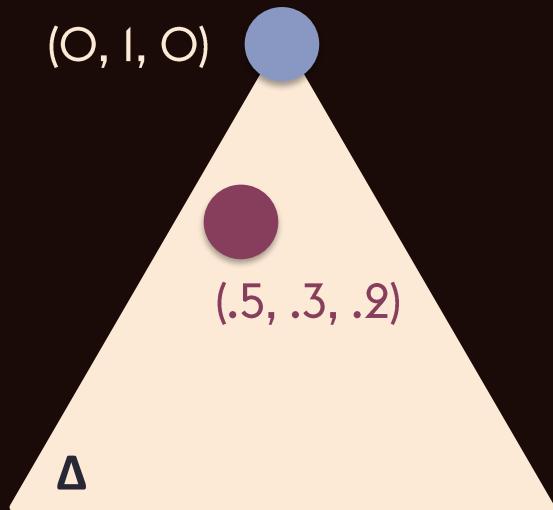
$$\eta = \begin{bmatrix} .3 \\ .8 \\ -.5 \\ .2 \\ -.1 \\ -.2 \\ .7 \\ .6 \\ .1 \end{bmatrix}$$

$$\operatorname{argmax} \langle x, p \rangle \\ \text{s.t. } p \in \Delta$$

$$\operatorname{argmax} \langle x, p \rangle + H(p) \\ \text{s.t. } p \in \Delta$$

**softmax, closed-form  
solution:**  $p^* = \exp(x) / Z$

$$x = A^T \eta$$



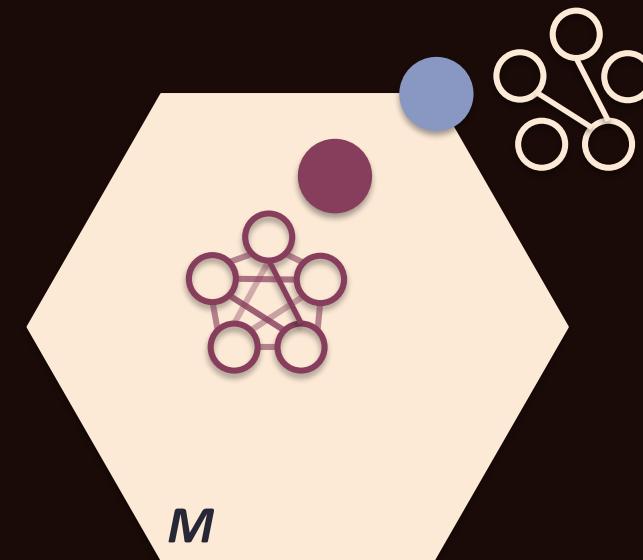
MAP

$$\operatorname{argmax} \langle \eta, \mu \rangle \\ \text{s.t. } \mu \in M$$

Marginal

$$\operatorname{argmax} \langle \eta, \mu \rangle + \tilde{H}(\mu) \\ \text{s.t. } \mu \in M$$

*structured attention networks*  
[Kim et al, 2017],  
[Liu et al, 2017]



$$\mu = Ap$$

MAP inference:  
Maximum spanning tree

Marginal inference:  
Matrix-Tree theorem



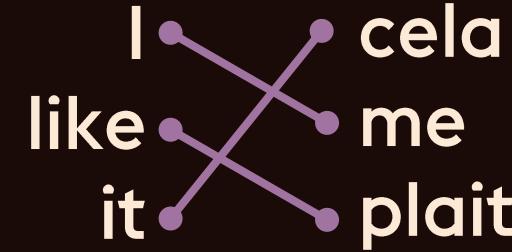
$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ \hline 0 & 1 & 1 \\ 1 & \dots & 0 & 0 & \dots \\ 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$* \rightarrow |$   
 $\text{like} \rightarrow |$   
 $| \rightarrow |$   
 $* \rightarrow \text{like}$   
 $| \rightarrow \text{like}$   
 $\text{it} \rightarrow \text{like}$   
 $* \rightarrow \text{it}$   
 $| \rightarrow \text{it}$   
 $\text{like} \rightarrow \text{it}$

$$\eta = \begin{bmatrix} .3 \\ .8 \\ -.5 \\ .2 \\ -.1 \\ -.2 \\ .7 \\ .6 \\ .1 \end{bmatrix}$$

Hungarian algorithm

#P complete



$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 0 & 0 & 0 \\ 1 & \dots & 0 & 0 & \dots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

$| - \text{cela}$   
 $| - \text{me}$   
 $| - \text{plait}$   
 $\text{like} - \text{cela}$   
 $\text{like} - \text{me}$   
 $\text{like} - \text{plait}$   
 $\text{it} - \text{cela}$   
 $\text{it} - \text{me}$   
 $\text{it} - \text{plait}$

$$\eta = \begin{bmatrix} .3 \\ .8 \\ -.5 \\ .2 \\ -.1 \\ -.2 \\ .7 \\ .6 \\ .1 \end{bmatrix}$$

$$\begin{aligned} & \operatorname{argmax} \langle x, p \rangle \\ \text{s.t. } & p \in \Delta \end{aligned}$$

MAP

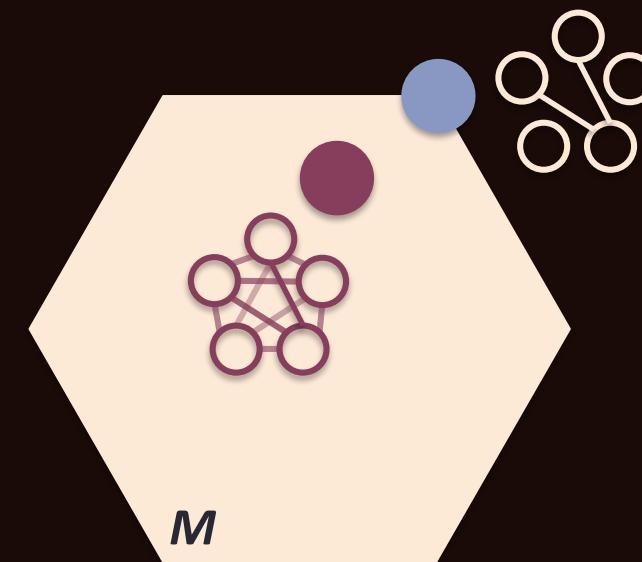
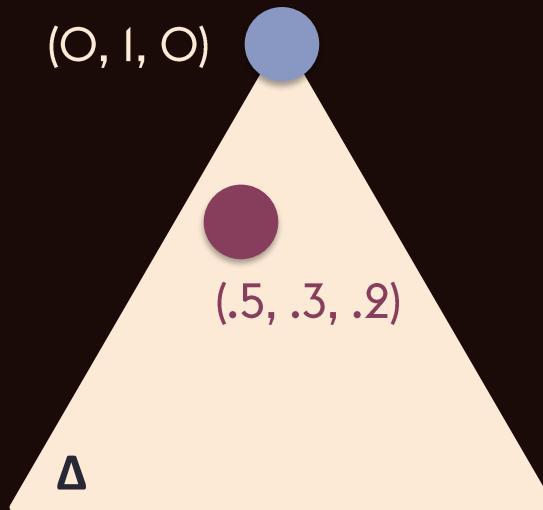
$$\begin{aligned} & \operatorname{argmax} \langle \eta, \mu \rangle \\ \text{s.t. } & \mu \in M \end{aligned}$$

$$\begin{aligned} & \operatorname{argmax} \langle x, p \rangle + H(p) \\ \text{s.t. } & p \in \Delta \end{aligned}$$

Marginal

$$\begin{aligned} & \operatorname{argmax} \langle \eta, \mu \rangle + \tilde{H}(\mu) \\ \text{s.t. } & \mu \in M \end{aligned}$$

$$x = A^T \eta$$



$$\mu = Ap$$

$$\begin{aligned} & \operatorname{argmax} \langle x, p \rangle \\ \text{s.t. } & p \in \Delta \end{aligned}$$

**MAP**

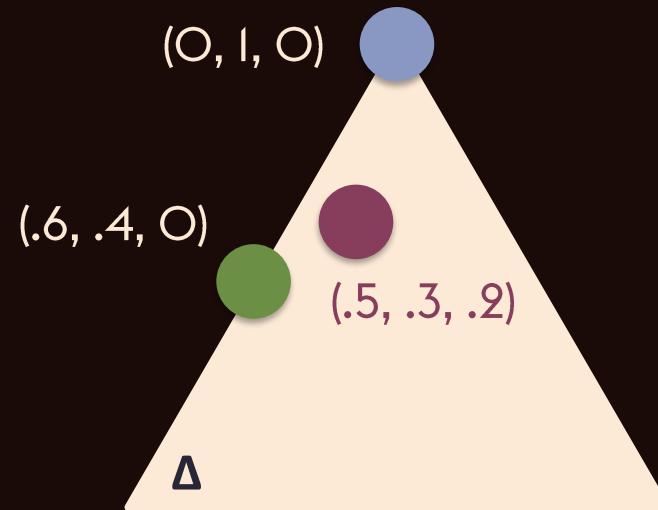
$$\begin{aligned} & \operatorname{argmax} \langle x, p \rangle + H(p) \\ \text{s.t. } & p \in \Delta \end{aligned}$$

**Marginal**

$$\begin{aligned} & \operatorname{argmax} \langle x, p \rangle - \frac{1}{2} \|Ap\|^2 \\ \text{s.t. } & p \in \Delta \end{aligned}$$

**SparseMAP**

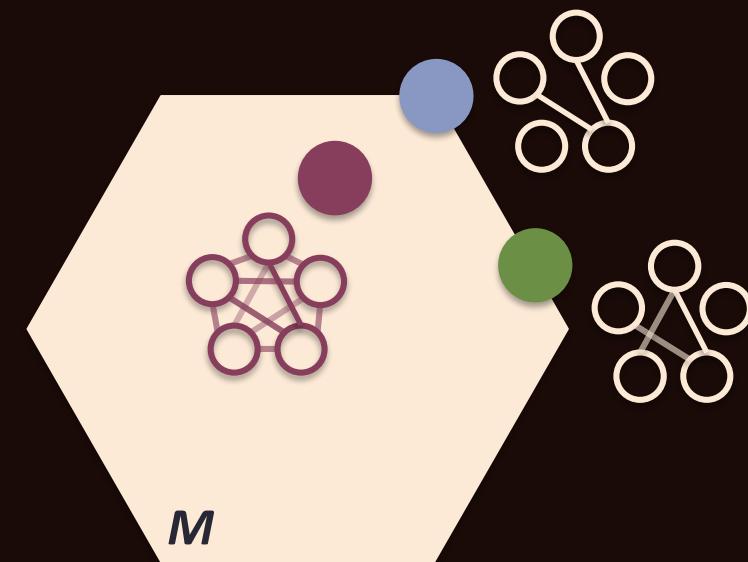
$$x = A^T \eta$$



$$\begin{aligned} & \operatorname{argmax} \langle \eta, \mu \rangle \\ \text{s.t. } & \mu \in M \end{aligned}$$

$$\begin{aligned} & \operatorname{argmax} \langle \eta, \mu \rangle + \tilde{H}(\mu) \\ \text{s.t. } & \mu \in M \end{aligned}$$

$$\begin{aligned} & \operatorname{argmax} \langle \eta, \mu \rangle - \frac{1}{2} \|\mu\|^2 \\ \text{s.t. } & \mu \in M \end{aligned}$$



$$\mu = Ap$$

# Efficiently Computing SparseMAP

$$\begin{aligned} & \operatorname{argmax} \langle \eta, \mu \rangle - \frac{1}{2} \|\mu\|^2 \\ \text{s.t. } & \mu \in M \end{aligned}$$

QP with exponentially many vertices!

## Forward Pass:

Active Set algorithm

only accesses  $M$  through MAP calls

linear & finite convergence

## Backward Pass:

$$\frac{\partial \mu^*}{\partial \eta}$$

Linear in  $\dim(M)$  and in # selected structures

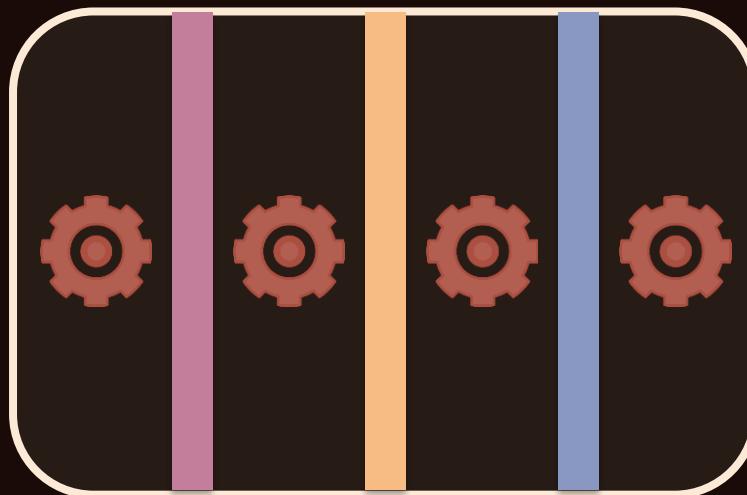
# Sparse Latent Structure

# Natural Language Inference

Prem: A gentleman overlooking a neighborhood situation.

Hypo: A police officer watches a situation closely.

(P, H)



{ entailment  
contradiction  
neither

# Natural Language Inference

Prem: A gentleman overlooking a neighborhood situation.

Hypo: A police officer watches a situation closely.



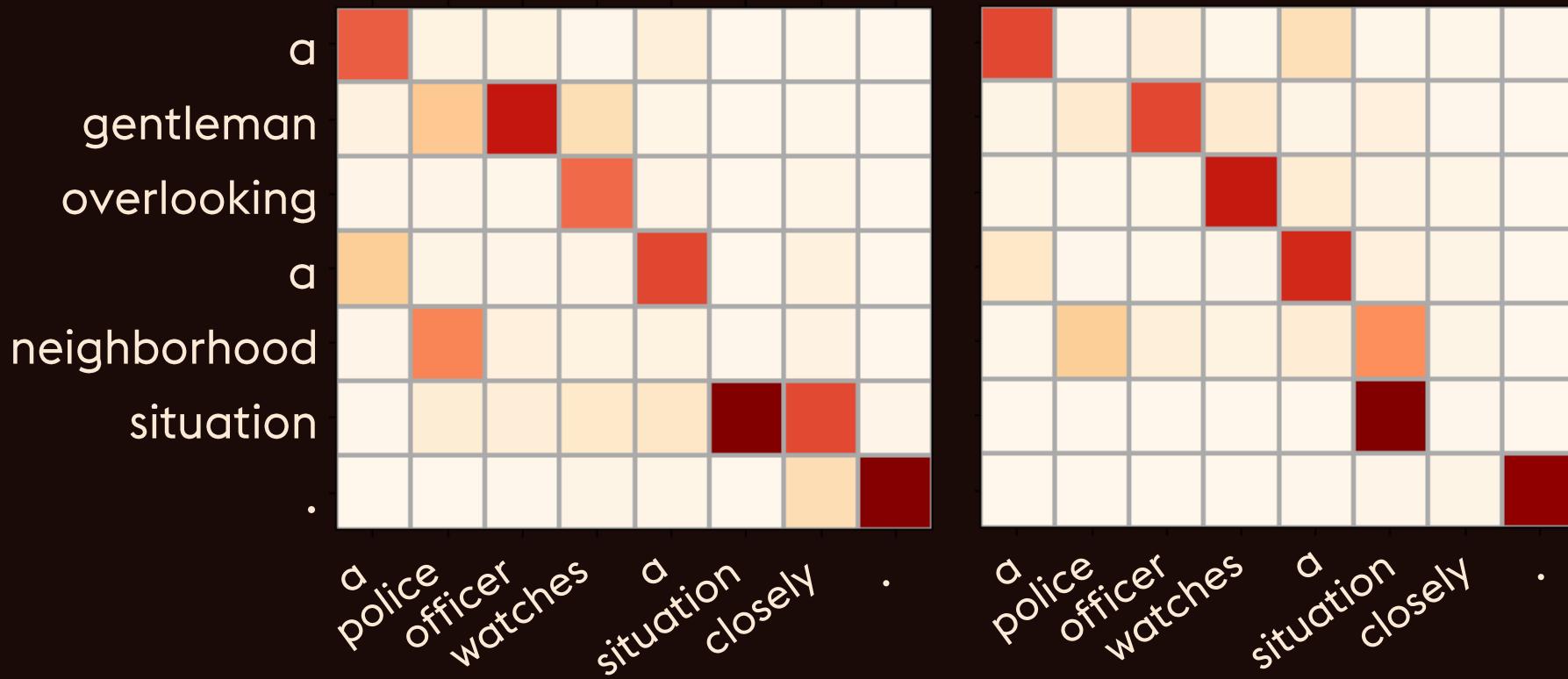
# Natural Language Inference

Prem: A gentleman overlooking a neighborhood situation.

Hypo: A police officer watches a situation closely.



# Natural Language Inference



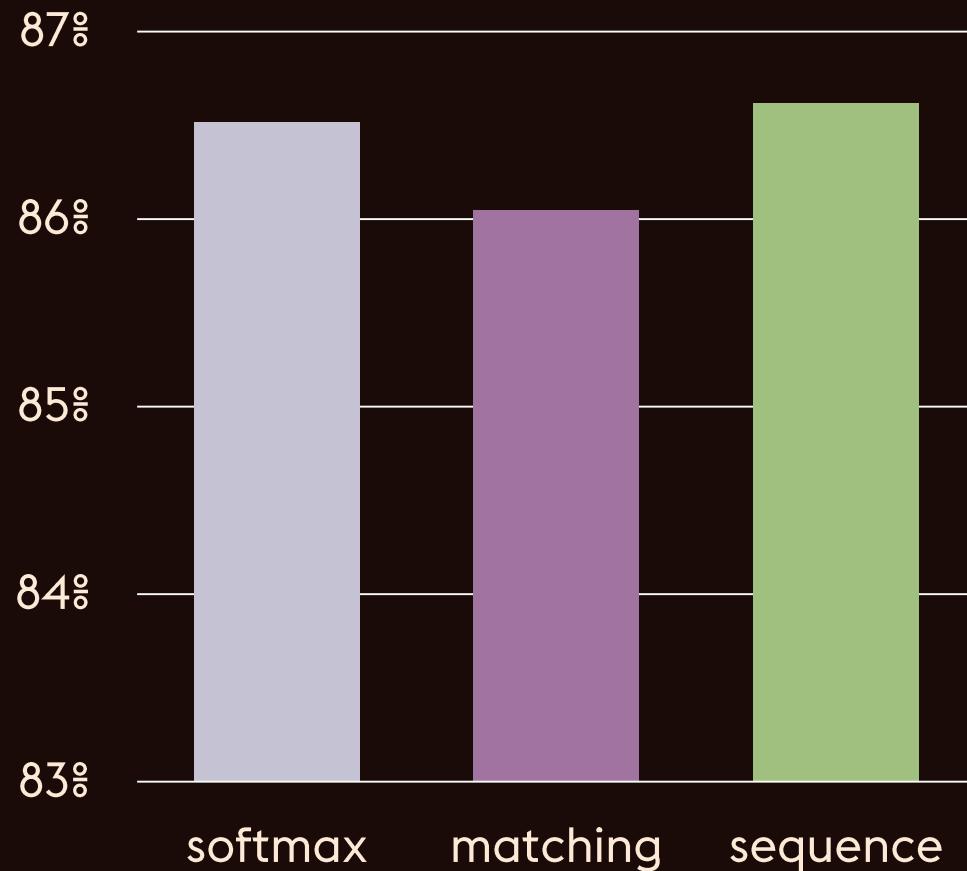
# Natural Language Inference with Linear Assignment

Prem: A gentleman overlooking a neighborhood situation.

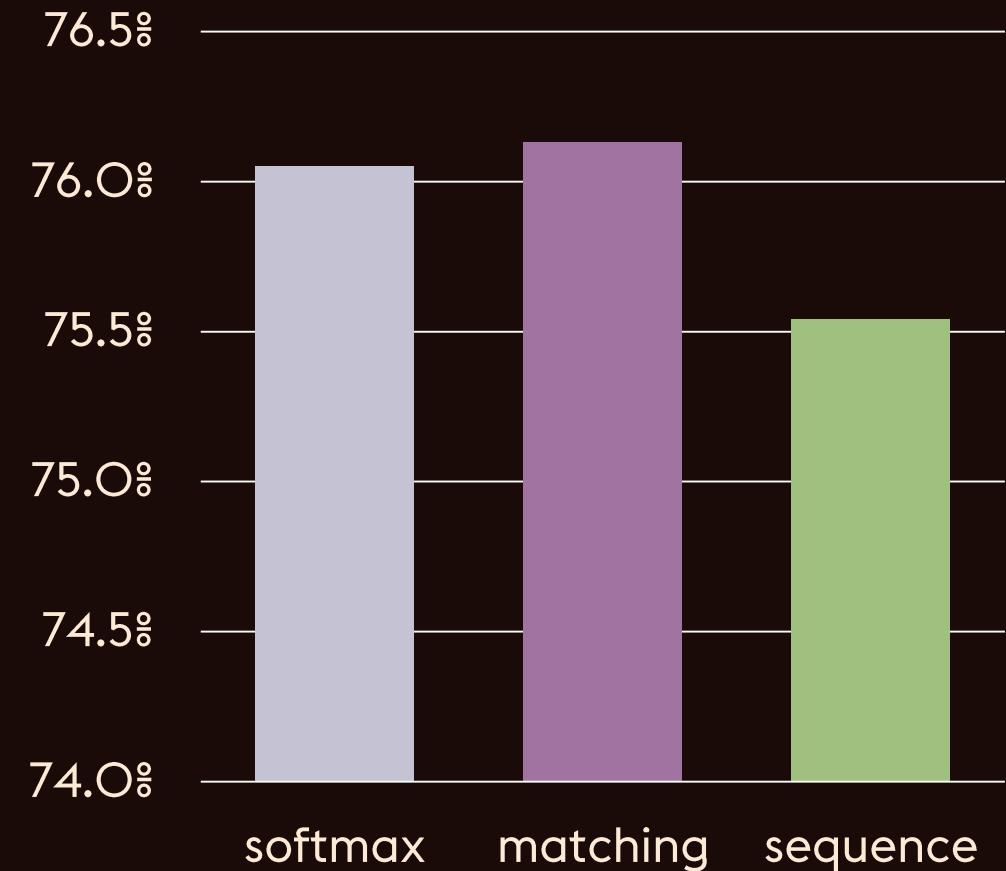
Hypo: A police officer watches a situation closely.



## SNLI

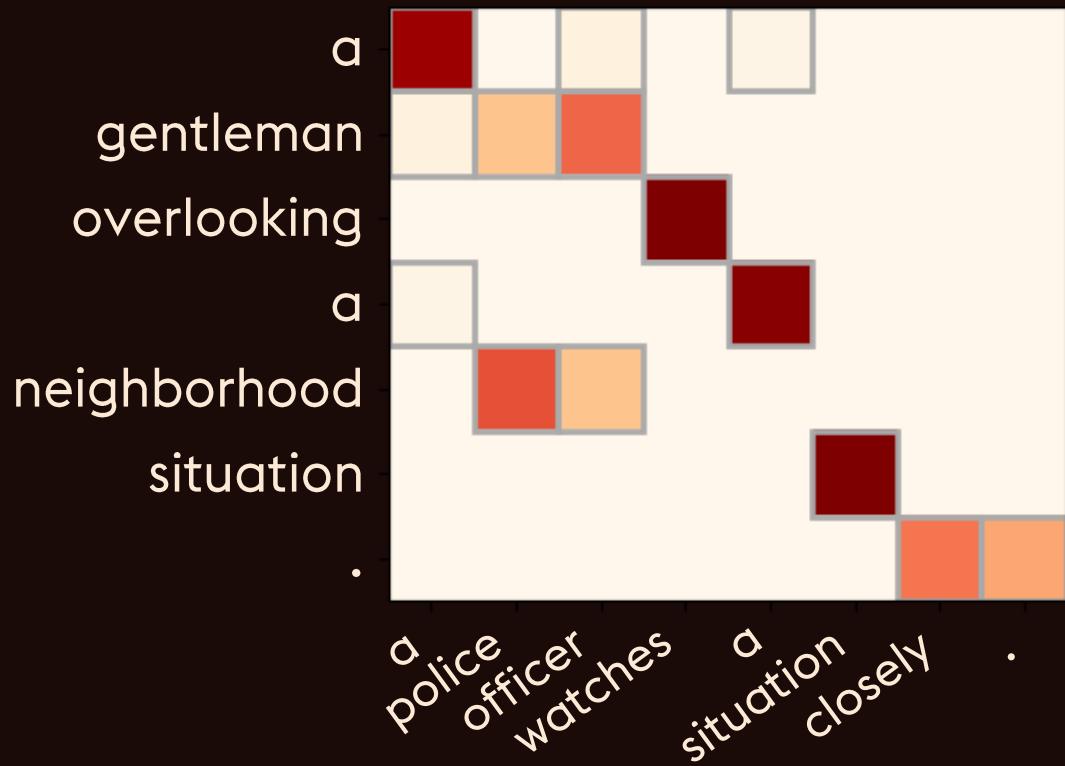


## MultiNLI

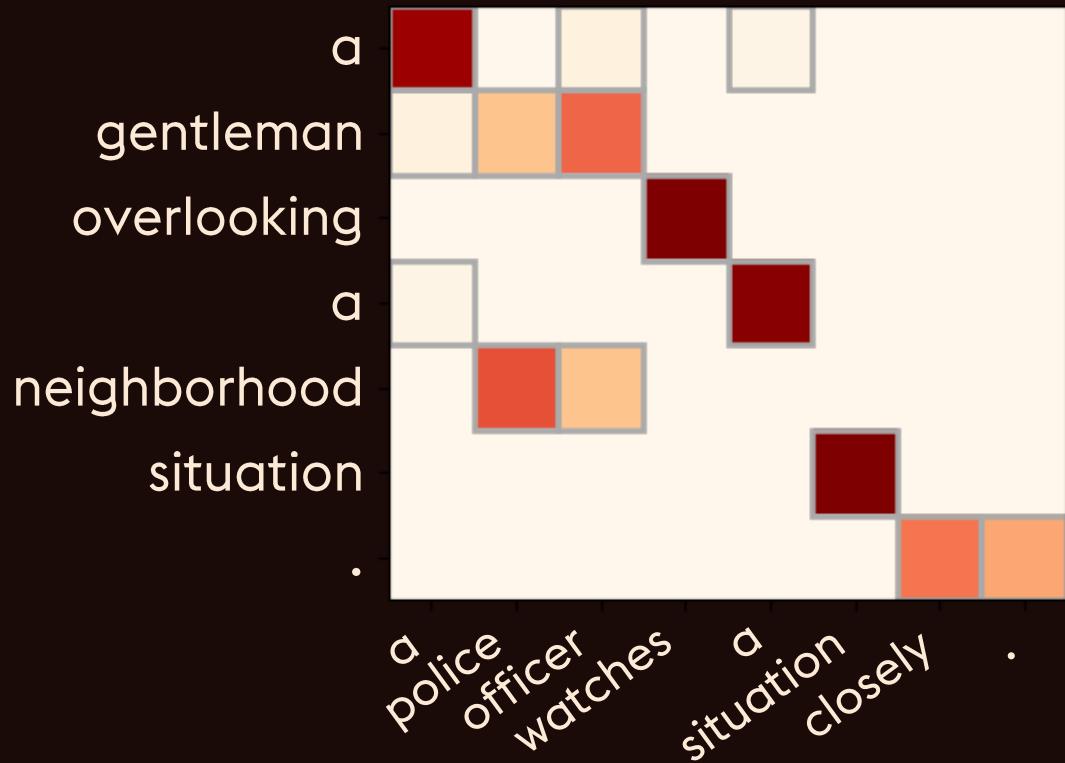


(3-class accuracy)

# Natural Language Inference with Linear Assignment



# Natural Language Inference with Linear Assignment



# Sparse Structured Output Prediction

# Sparse Structured Output Prediction

SparseMAP  
loss

scores      gold structure

$$L_A(\eta, \mu) = \max_{\mu \in M} \{ \langle \eta, \mu \rangle - \frac{1}{2} \|\mu\|^2 \}$$
$$- \langle \eta, \mu \rangle + \frac{1}{2} \|\mu\|^2$$

margin-SparseMAP  
loss

cost (as in structured SVM)

$$L_A^\rho(\eta, \mu) = \max_{\mu \in M} \{ \langle \eta, \mu \rangle - \frac{1}{2} \|\mu\|^2 + \rho(\mu, \bar{\mu}) \}$$
$$- \langle \eta, \mu \rangle + \frac{1}{2} \|\mu\|^2$$

Instance of a structured Fenchel-Young loss, like CRF, SVM, etc.

[Blondel, Martins, Niculae '18]

# Dependency Parsing with bi-LSTM features

[Kiperwasser & Goldberg, 2016]

90

85

80

75

70

65

60

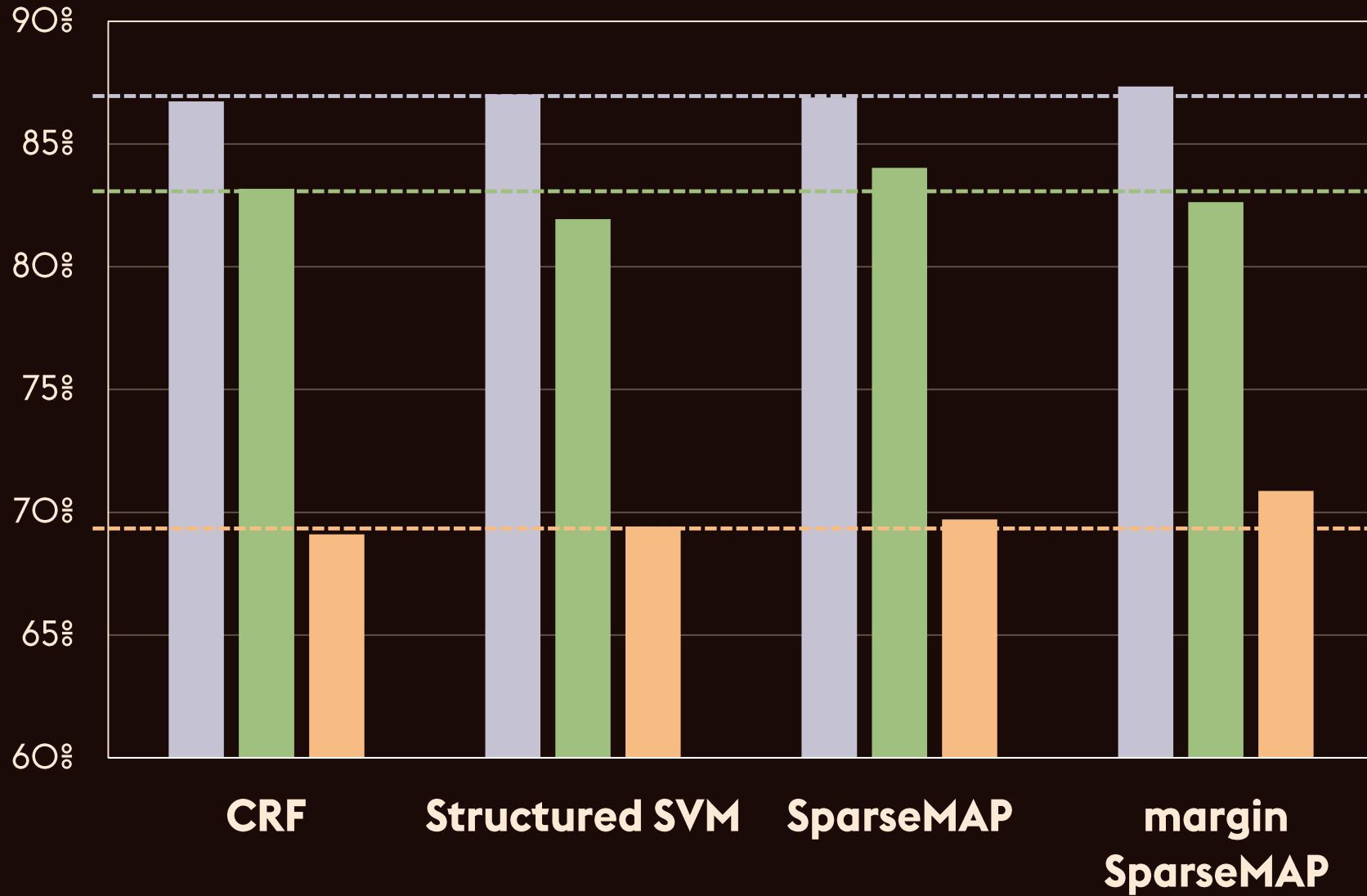
**CRF**

**Structured SVM**

**SparseMAP**

**margin  
SparseMAP**

■ English ■ Chinese ■ Vietnamese



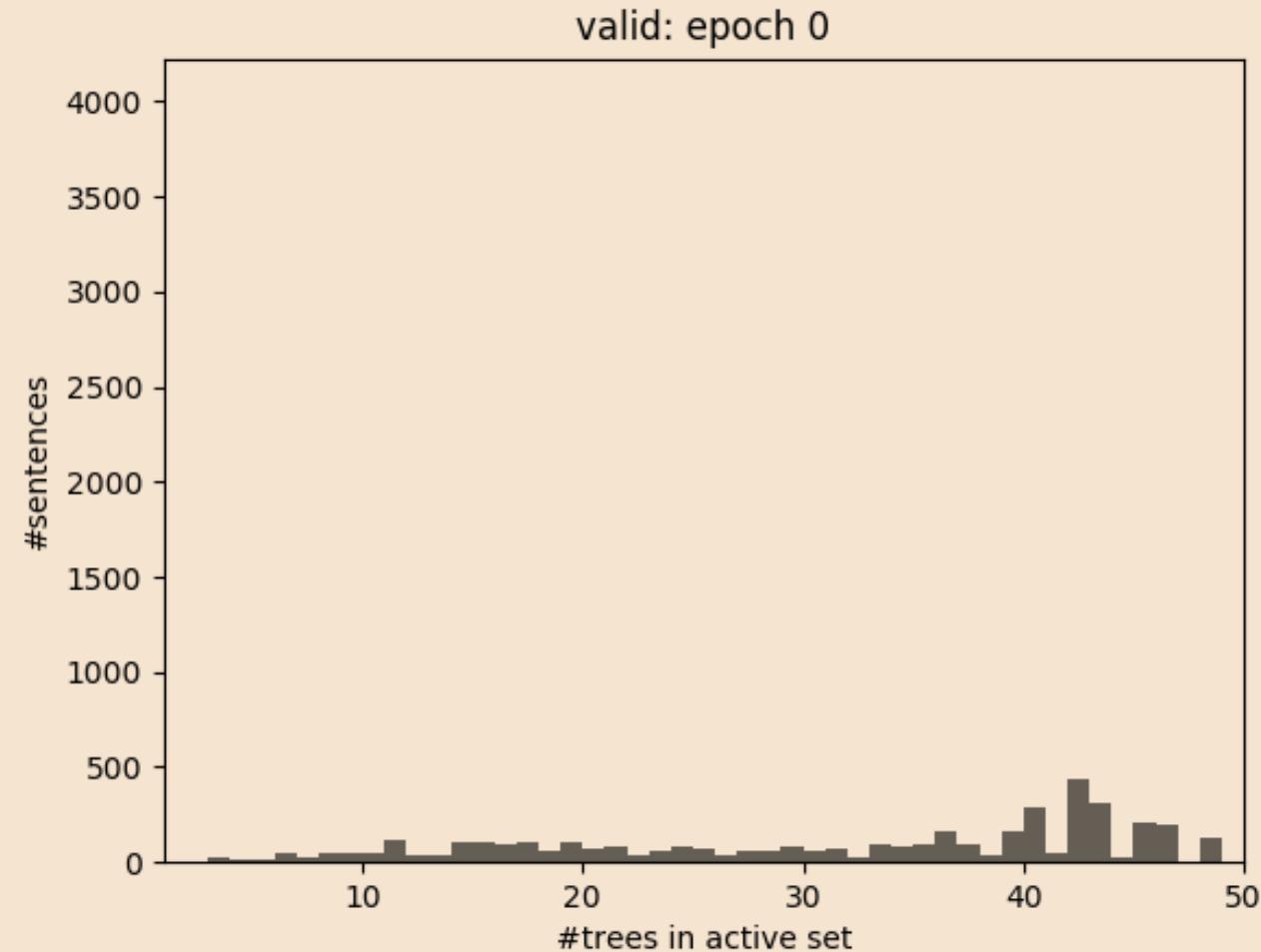
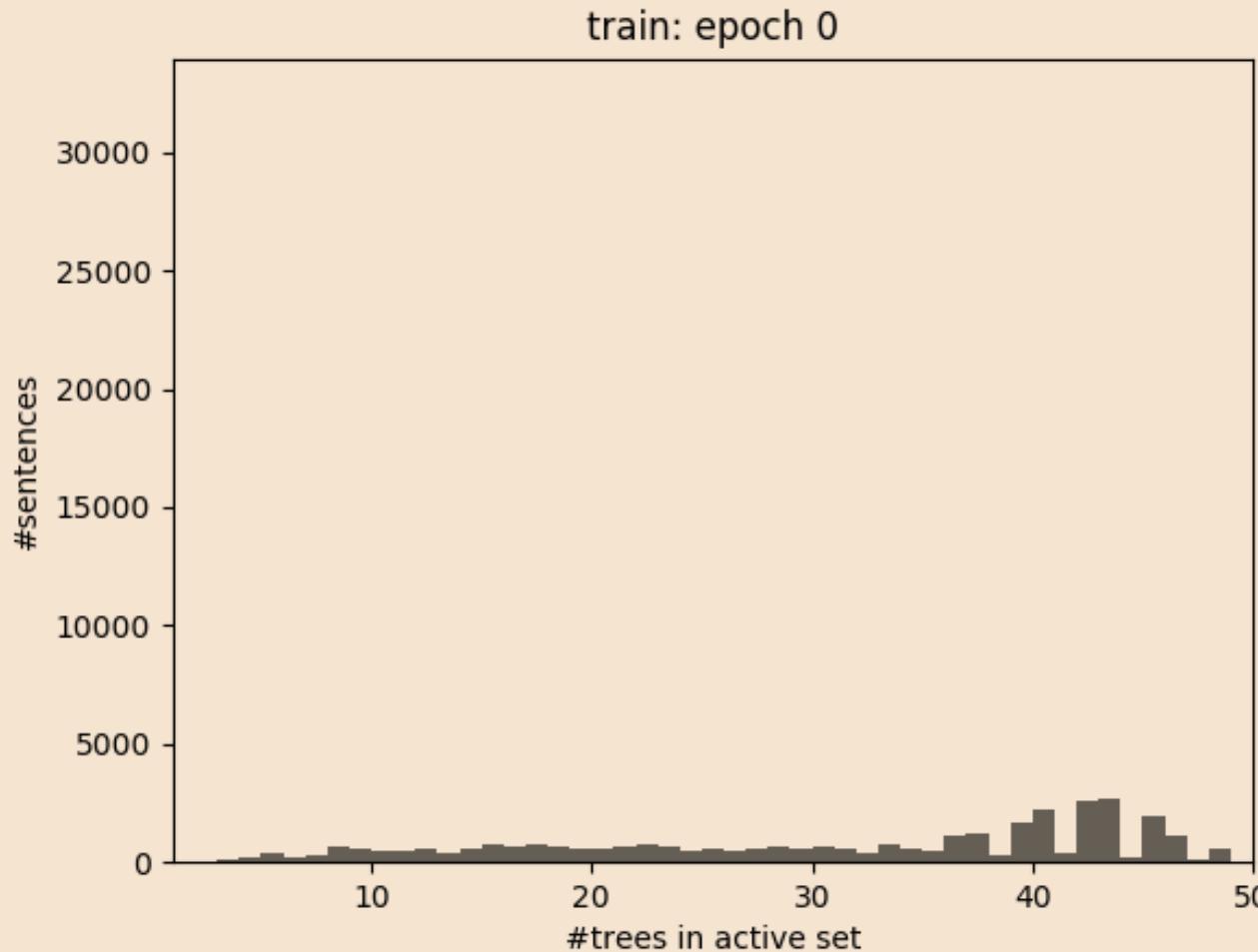
Unlabeled Accuracy (UAS)

Universal Dependencies dataset

■ English ■ Chinese ■ Vietnamese

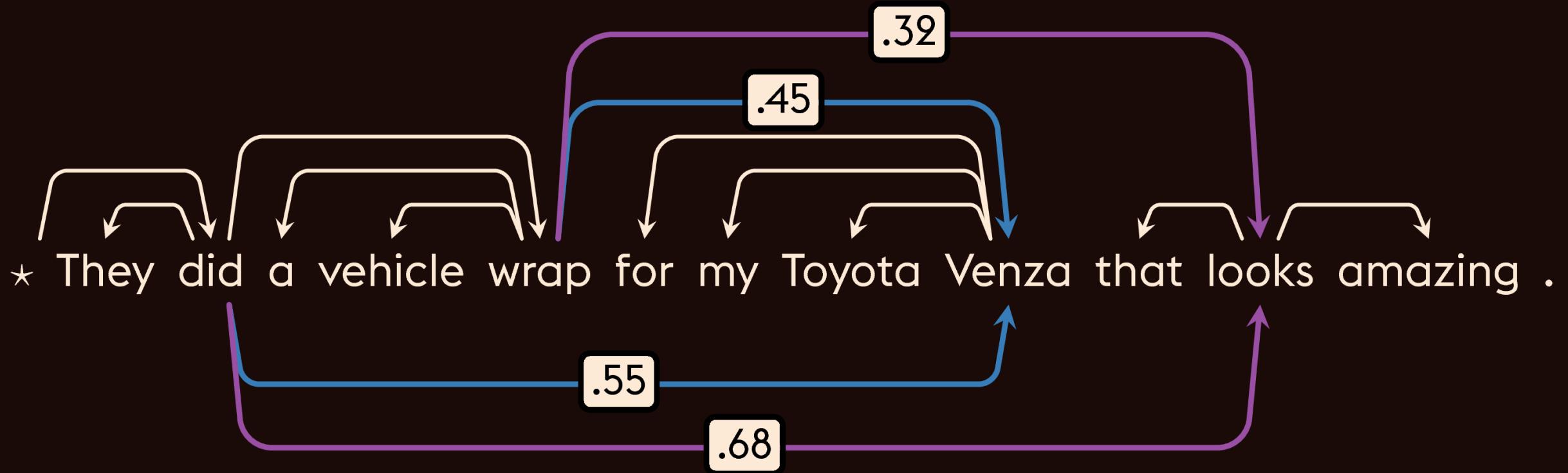
# Sparse Structured Output Prediction

As models train, inference gets sparser!



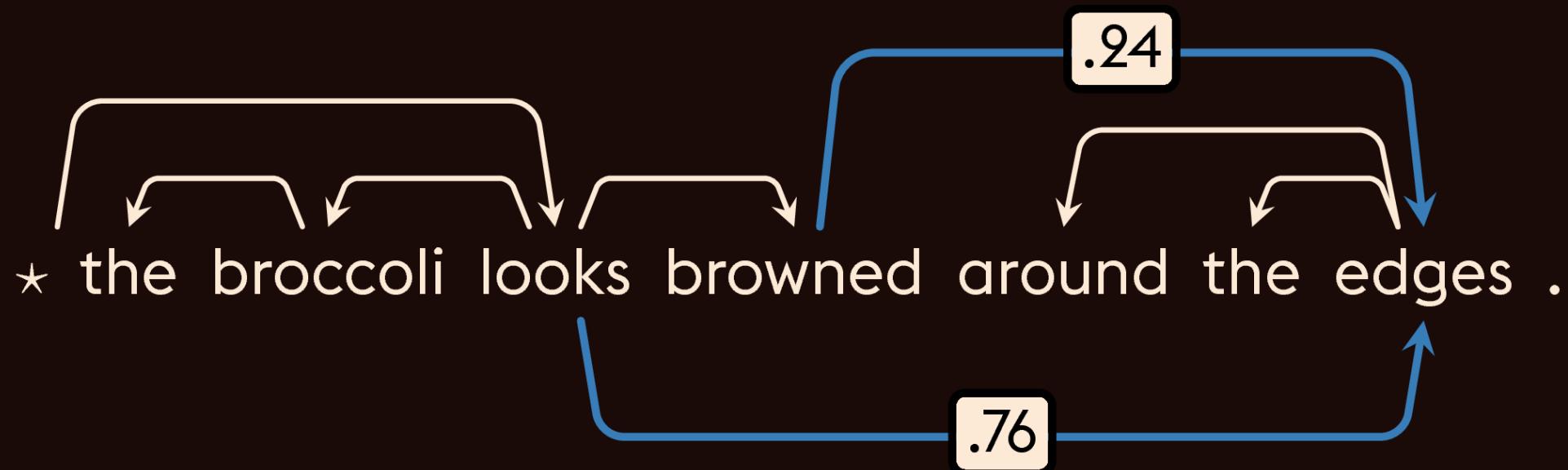
# Sparse Structured Output Prediction

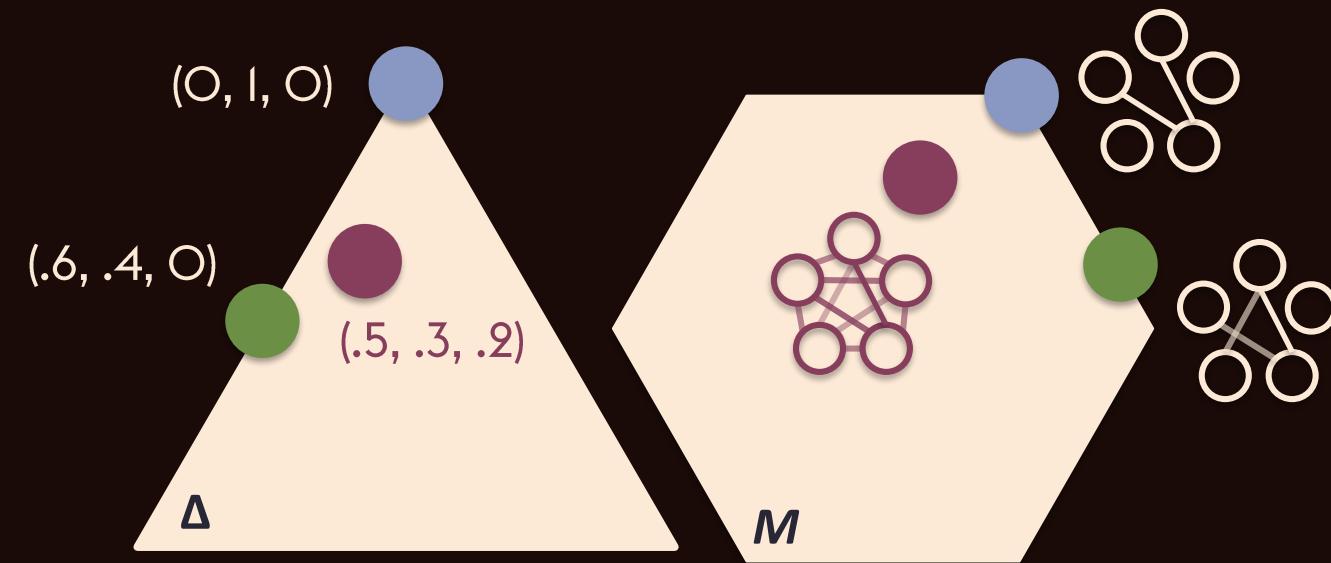
Inference captures linguistic ambiguity!



# Sparse Structured Output Prediction

Inference captures linguistic ambiguity!





**poster #66 tonight @6:15**



[github.com/vene/sparsemap](https://github.com/vene/sparsemap)

<https://vene.ro>



# vnfrombucharest

# Sparse Latent Structure

gentleman

overlooking

a

neighborhood

situation

.

a

police

officer

watches

a

situation

closely

# Sparse Structured Output Prediction

\* the broccoli looks browned around the edges.

.24

.76