SparseMAP: Differentiable Sparse Structured Inference

Presented by Vlad Niculae

Joint work with André FT Martins

Mathieu Blondel

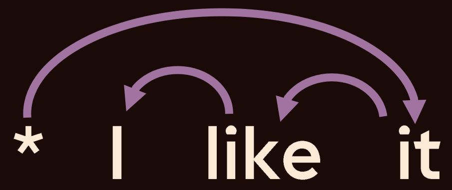
Claire Cardie

poster #66 tonight

github.com/vene/sparsemap

Structured Inference

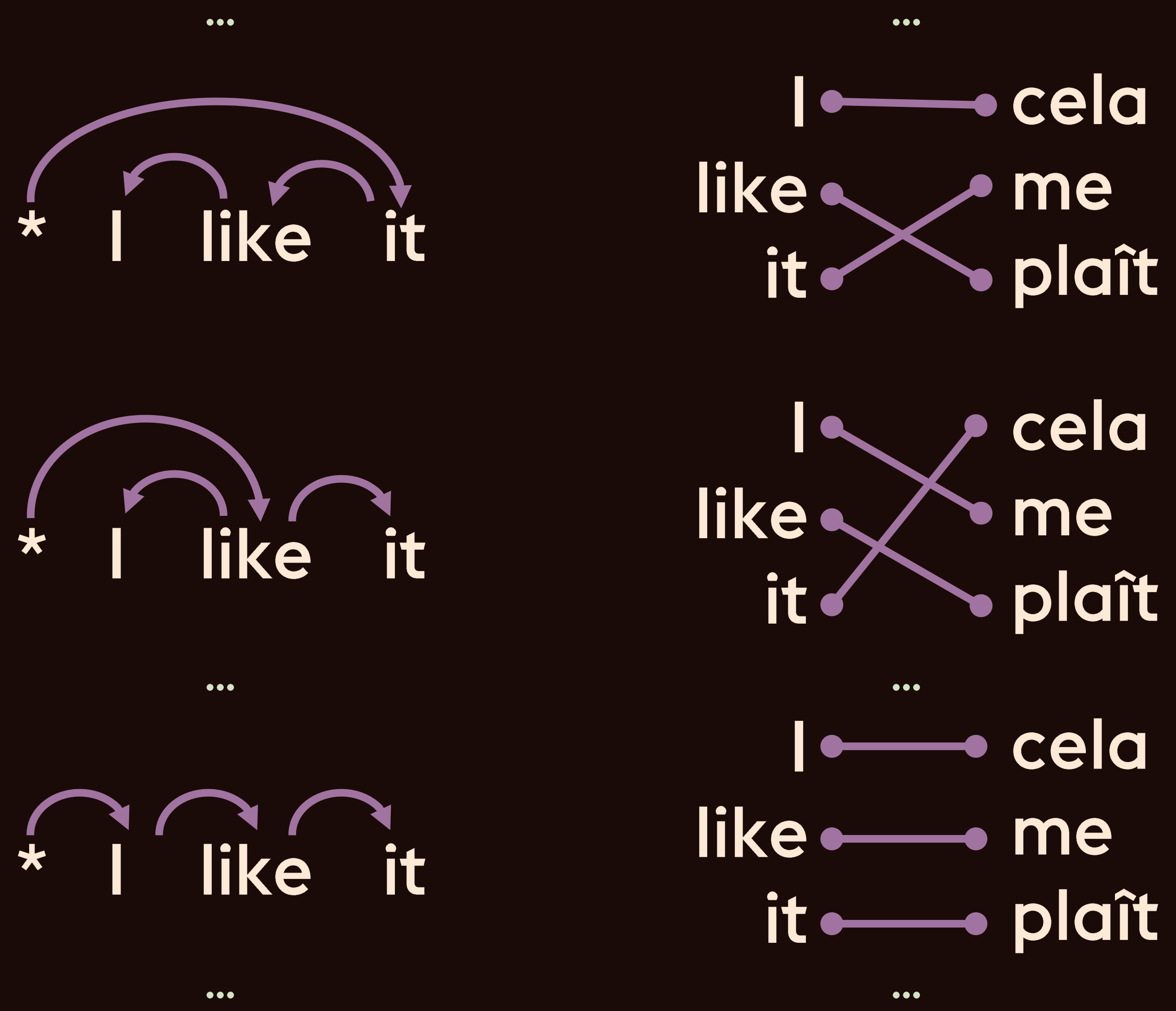
...



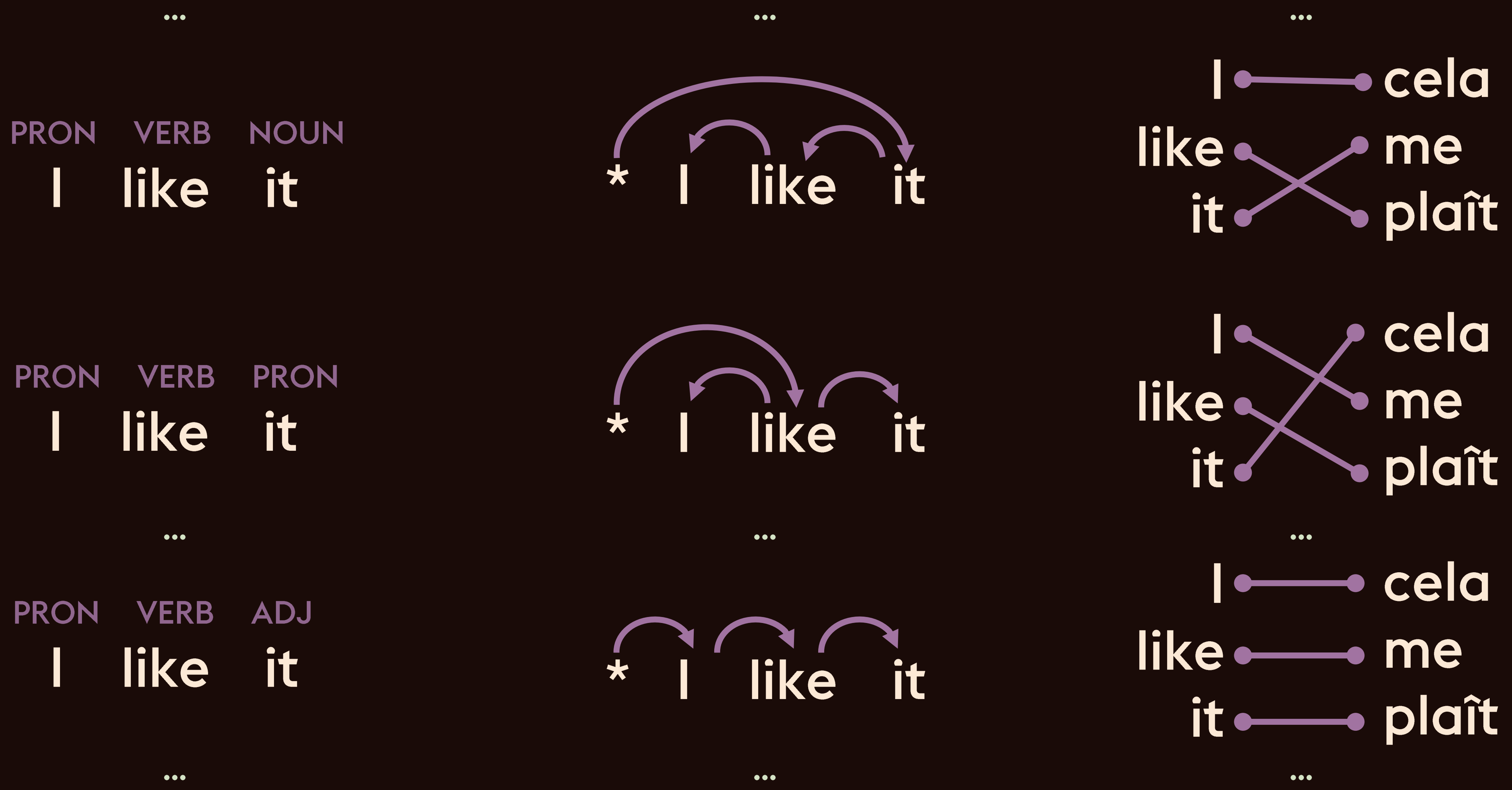
...

...

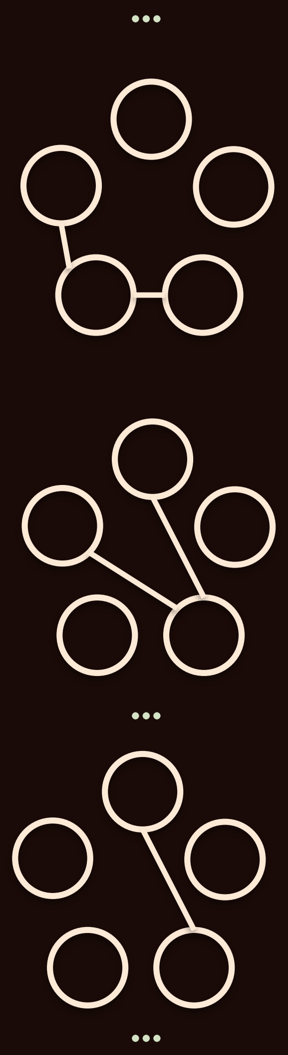
Structured Inference



Structured Inference

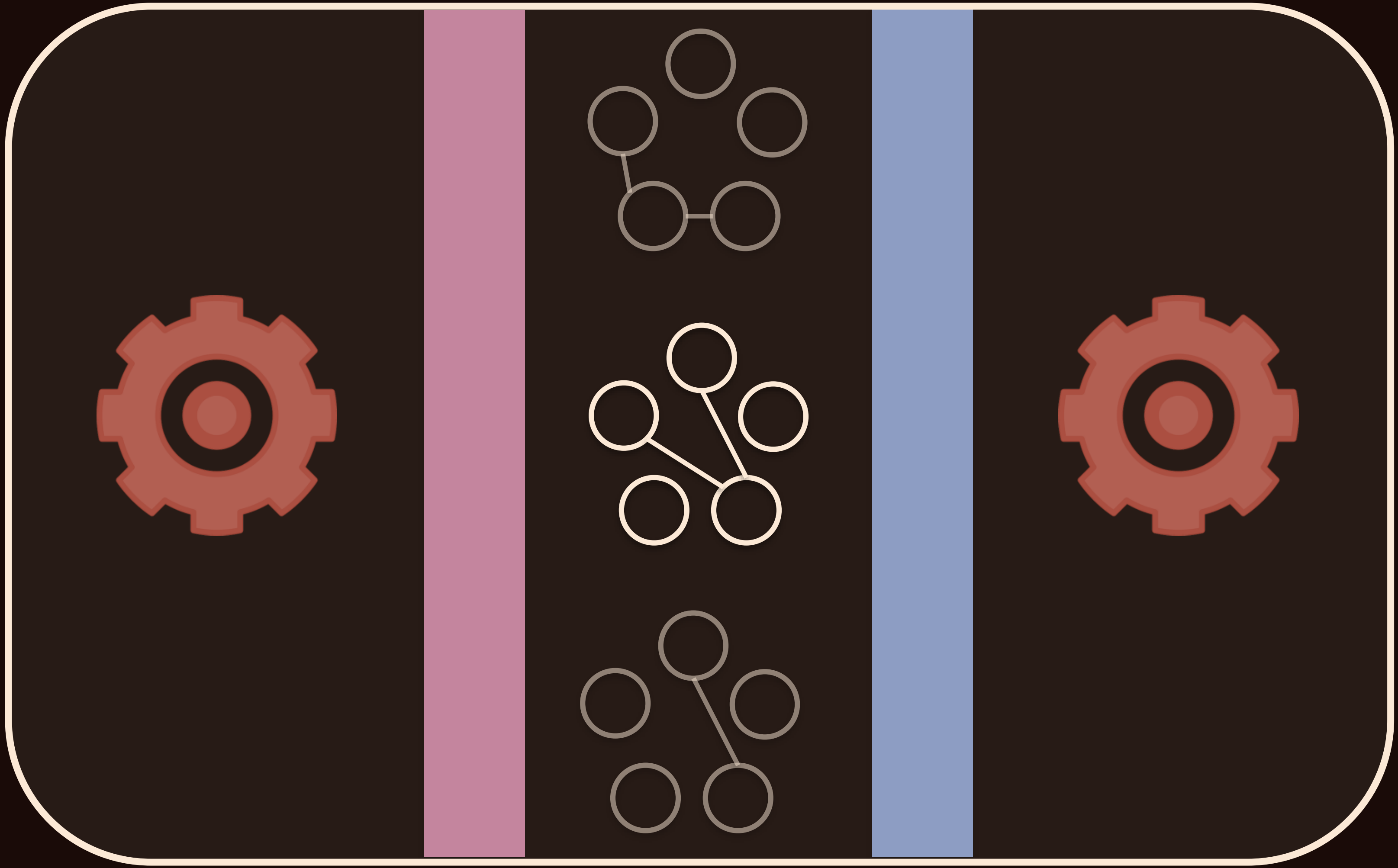


Structured Inference



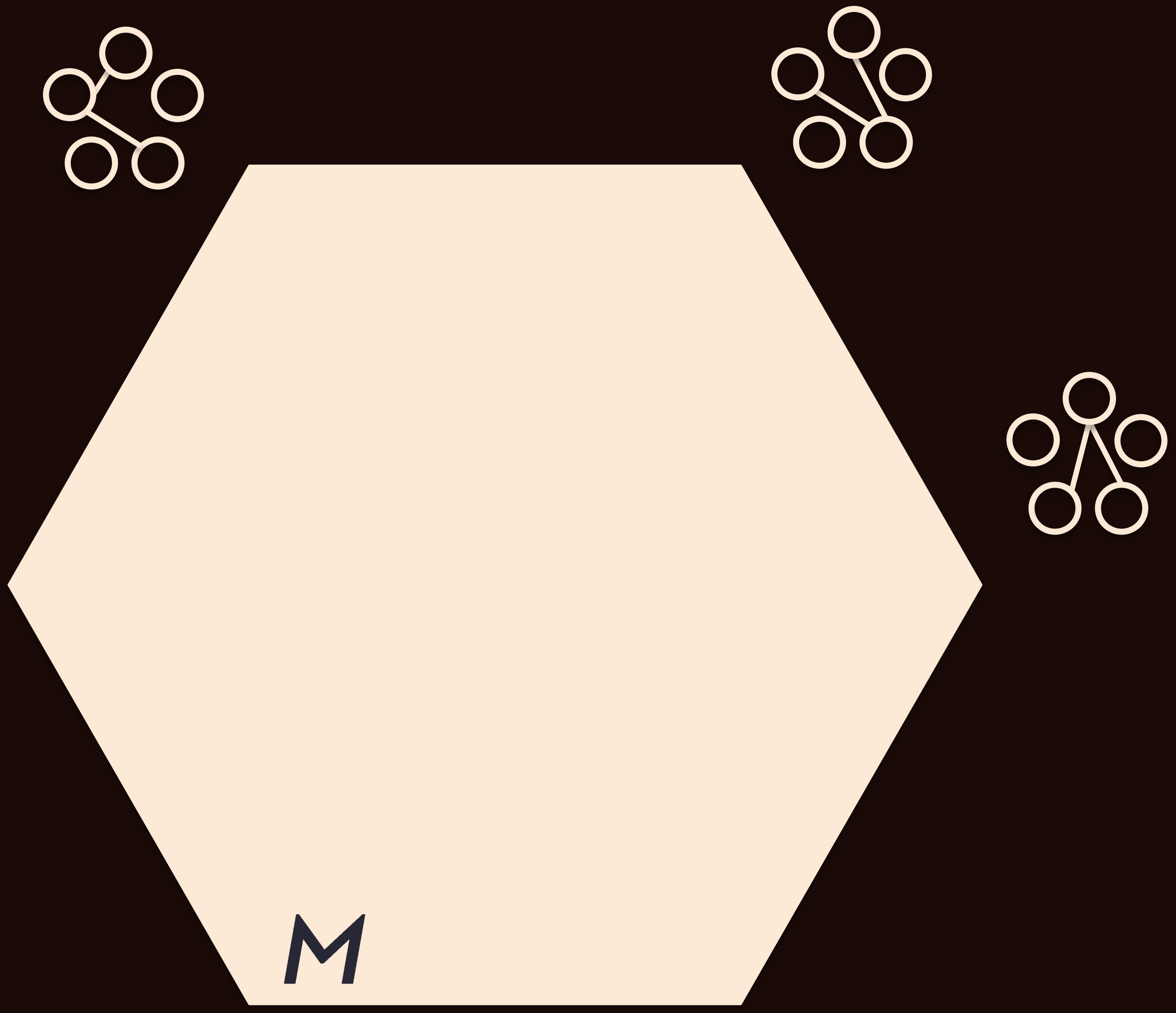
(Latent) Structured Inference

input



output

$\boldsymbol{M} = \mathbf{conv}(\mathcal{Y})$ where $\mathcal{Y} = \{$, , $\}$



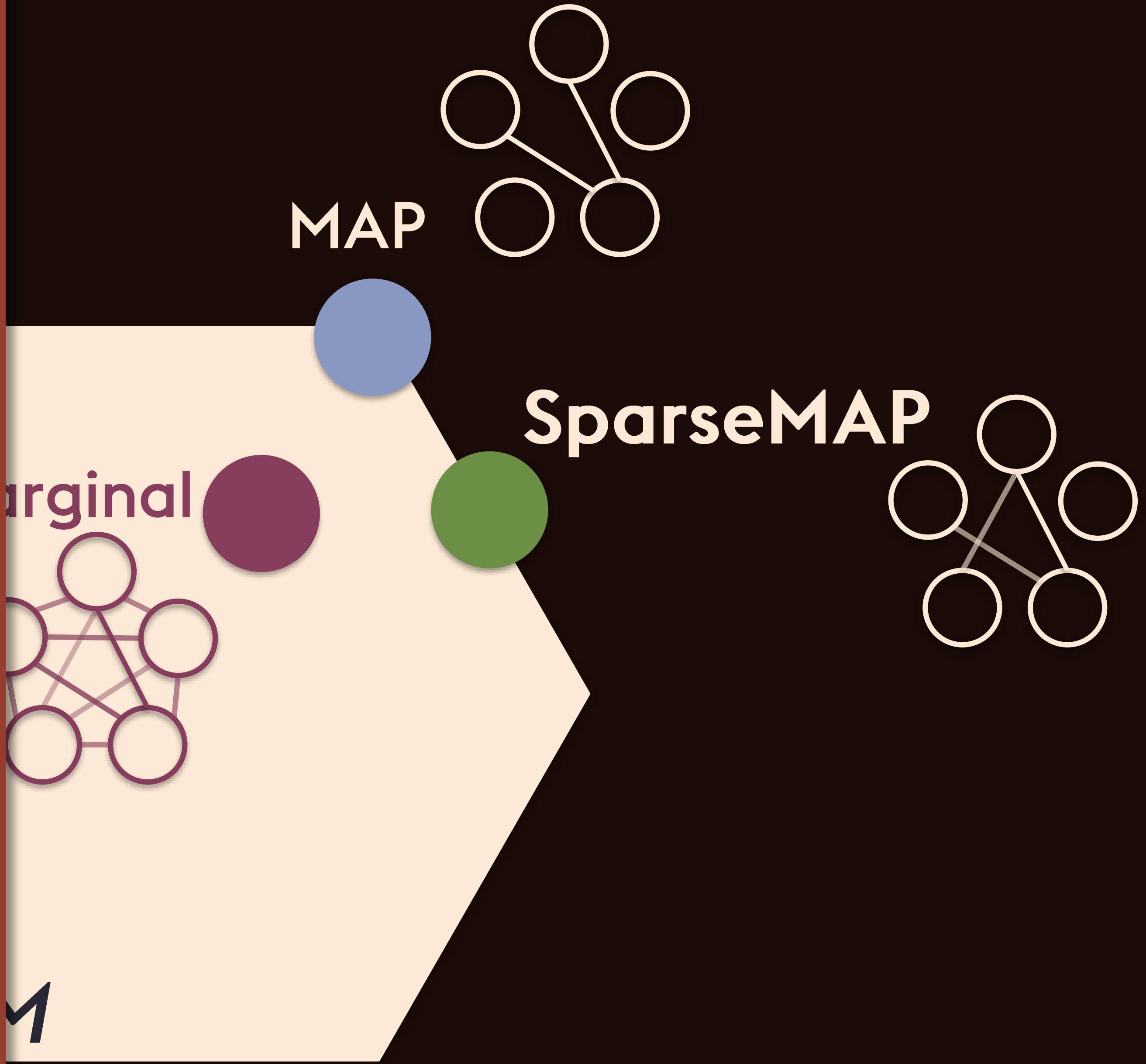
SparseMAP

Efficient & simple to:

- compute
- back-propagate

Useful as:

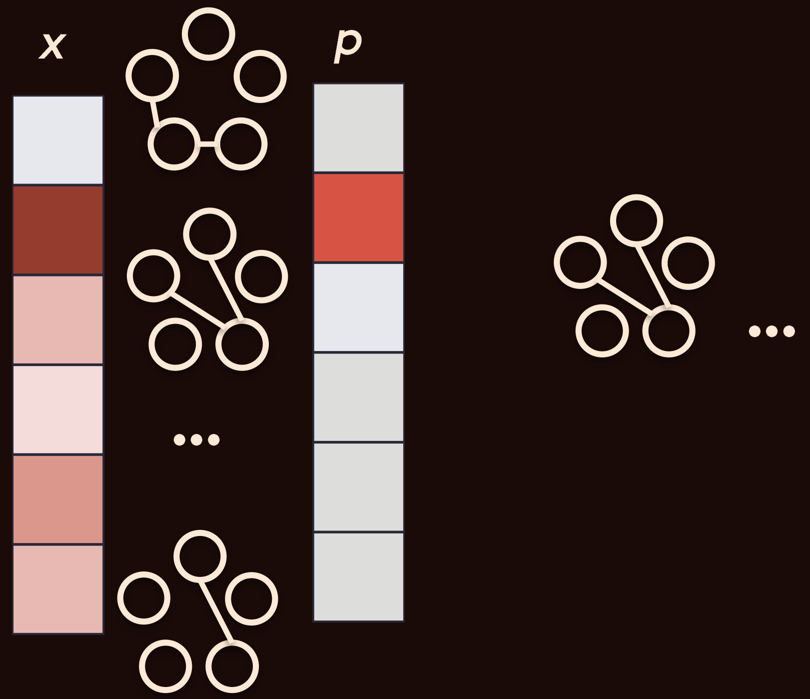
- hidden layer
- output layer



Deriving SparseMAP

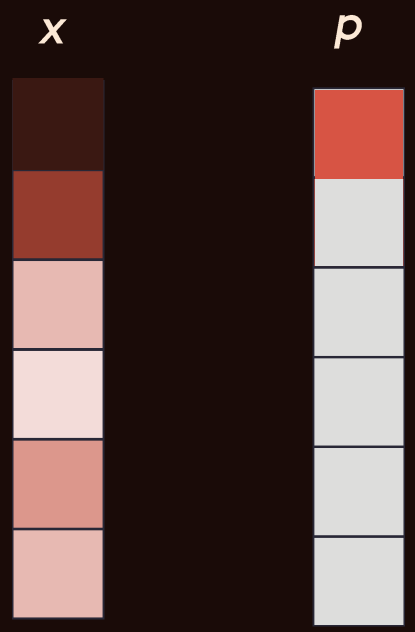
Structured Inference as argmax

input



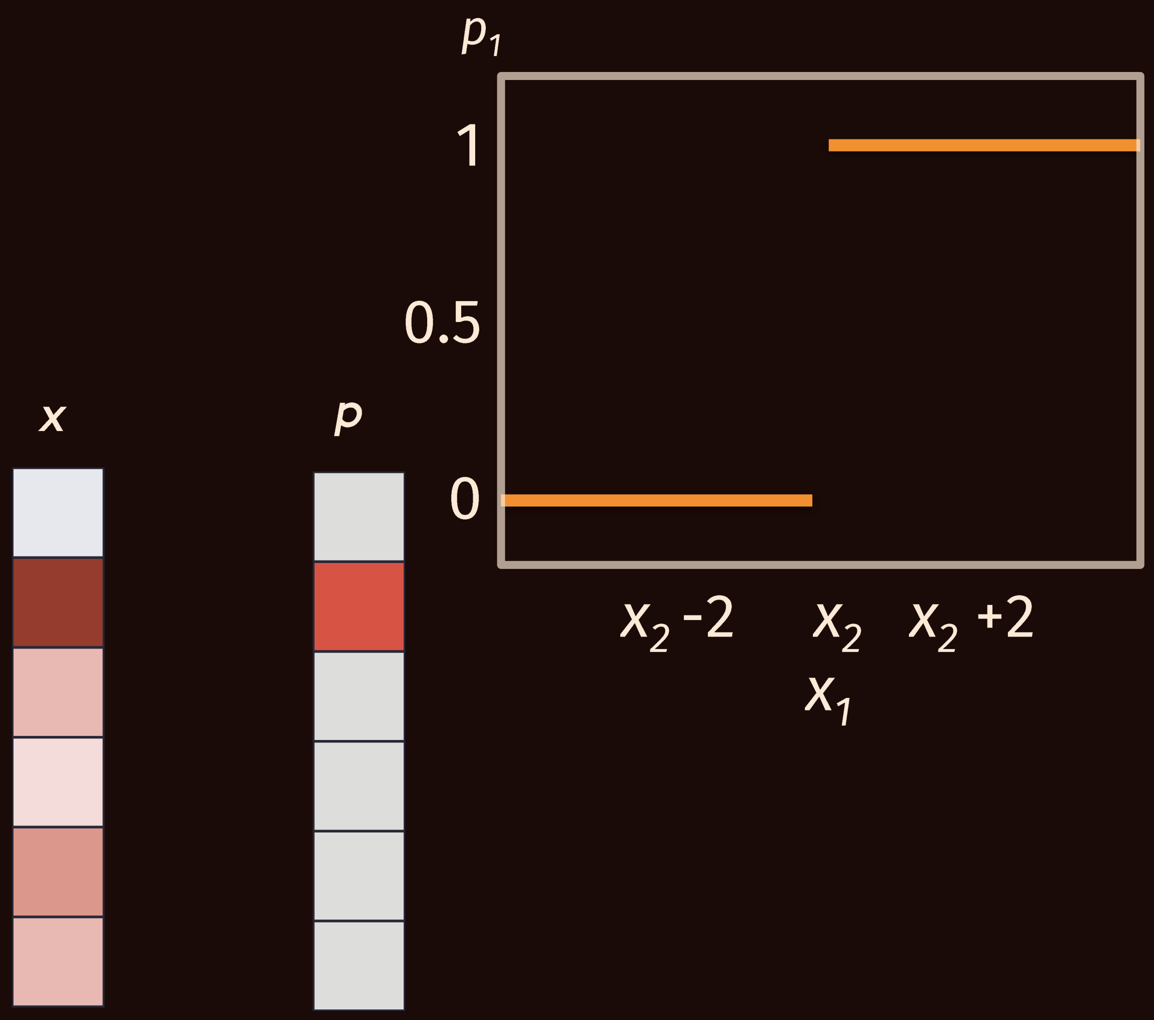
argmax

input



$\partial p / \partial x$?

input

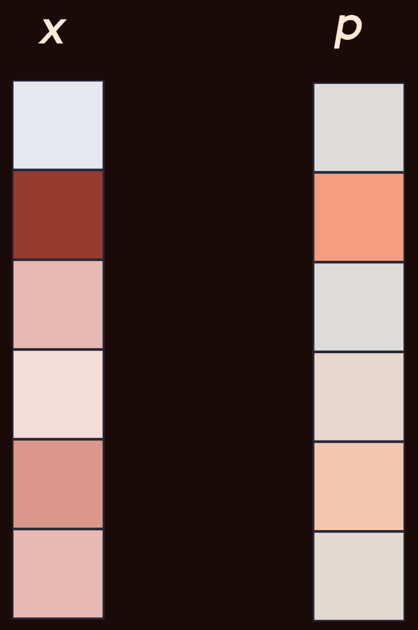


$\partial p / \partial x$?

argmax → softmax

$$p_i = \exp x_i \ / Z$$

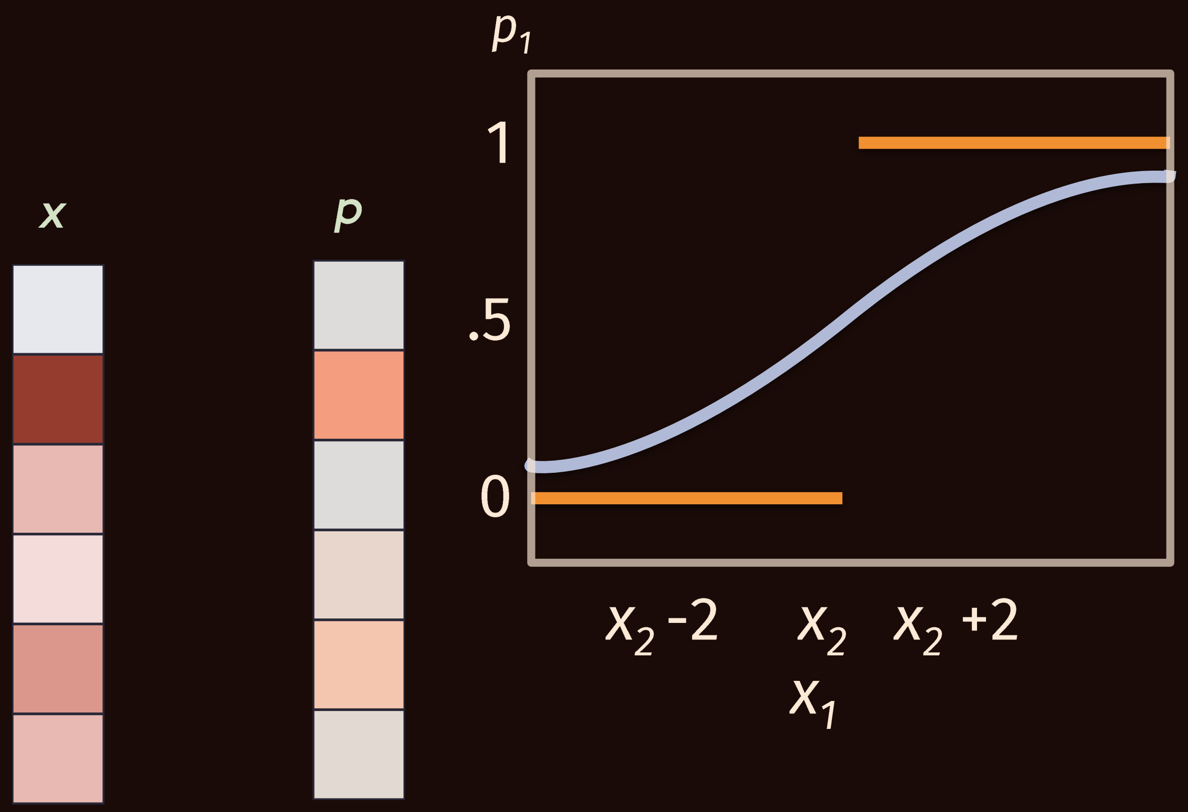
input



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

argmax → softmax

input

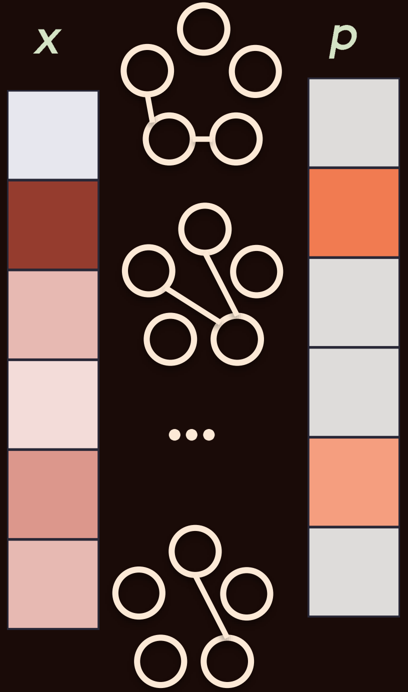


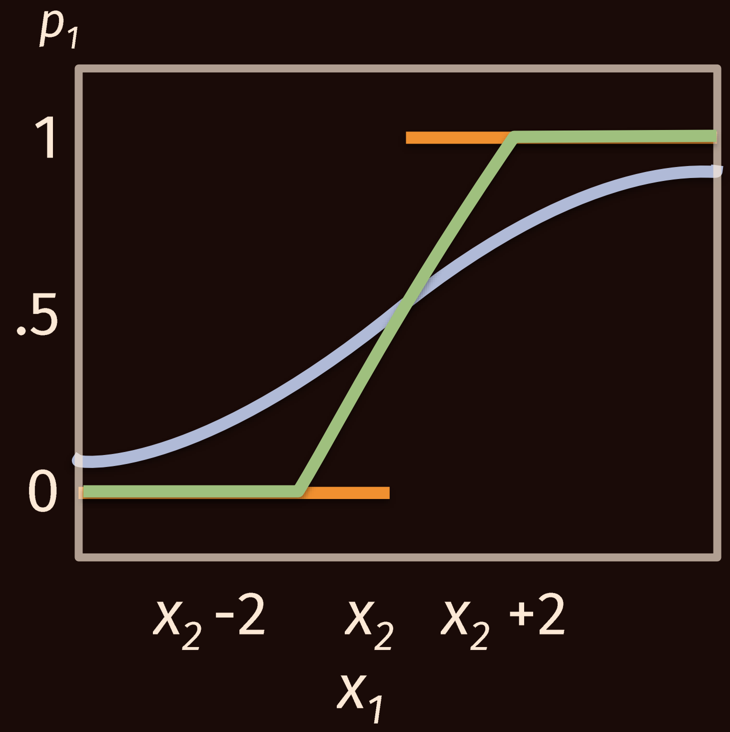
$$\partial p \,/\, \partial x\,?$$

argmax → softmax → sparsemax

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

(exponentially large)





$\partial p / \partial x$?

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

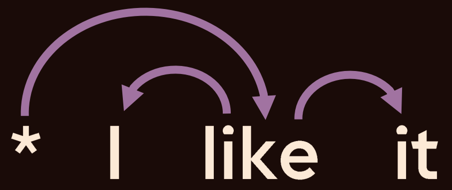
$$\boldsymbol{x} = \boldsymbol{A}^\top \boldsymbol{\eta}$$

$\boldsymbol{x} \in \mathbf{R}^k$, $\boldsymbol{\eta} \in \mathbf{R}^d$

$k \gg d$

$$\boldsymbol{x} = \boldsymbol{A}^{\top}\boldsymbol{\eta}$$

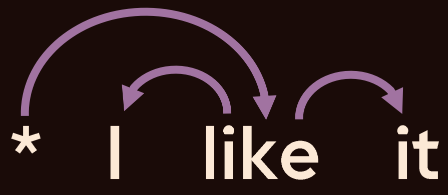
$$x = A^{\top}\eta$$

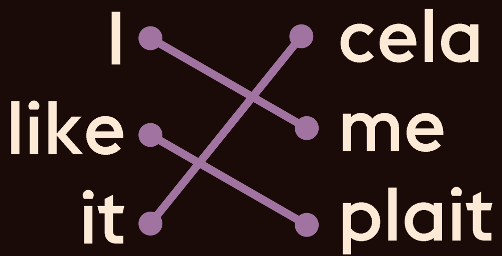


$$A = \begin{bmatrix} 1 & & 0 & 0 & \\ 0 & & 1 & 1 & \\ 0 & & 0 & 0 & \\ \hdashline 0 & & 1 & 1 & \\ 1 & \cdots & 0 & 0 & \cdots \\ 0 & & 0 & 0 & \\ \hdashline 0 & & 0 & 0 & \\ 0 & & 1 & 0 & \\ 1 & & 0 & 1 & \end{bmatrix} \quad \begin{matrix} * \to I \\ \text{like} \to I \\ \text{it} \to I \\ \hdashline * \to \text{like} \\ I \to \text{like} \\ \text{it} \to \text{like} \\ \hdashline * \to \text{it} \\ I \to \text{it} \\ \text{like} \to \text{it} \end{matrix}$$

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

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





A=

1		0	0		*	→ I
0		1	1		like	→ I
0		0	0		it	→ I
0		1	1		*	→ like
1	…	0	0	…	I	→ like
0		0	0		it	→ like
0		0	0		*	→ it
0		1	0		I	→ it
1		0	1		like	→ it

η=

.3
.8
-.5
.2
-.1
-.2
.7
.6
.1

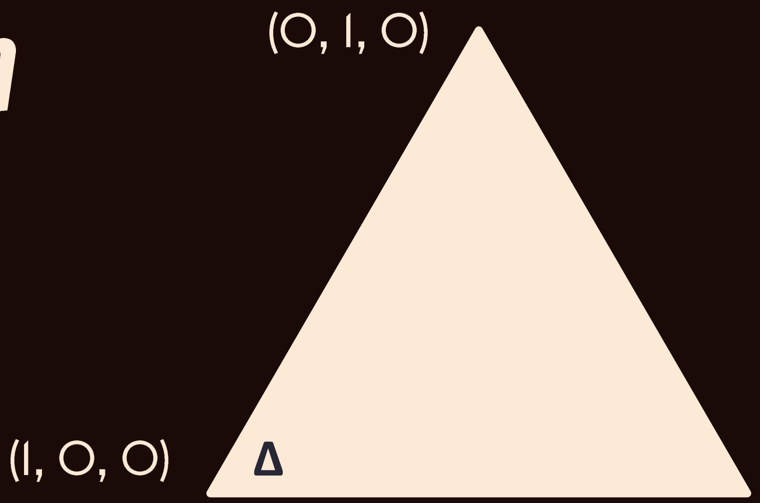
A=

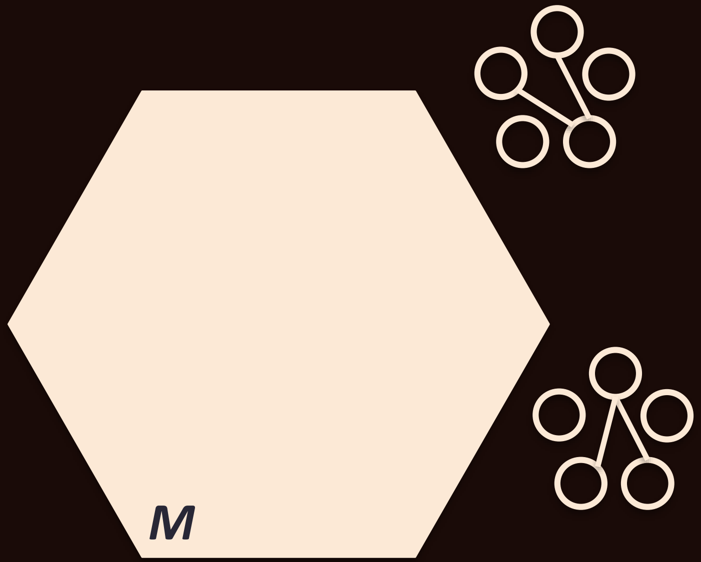
1		0	0		I	- cela
0		1	0		I	- me
0		0	1		I	- plait
0		0	1		like	- cela
1	…	0	0	…	like	- me
0		1	0		like	- plait
0		1	0		it	- cela
0		0	1		it	- me
1		0	0		it	- plait

η=

.3
.8
-.5
.2
-.1
-.2
.7
.6
.1

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





$$\mu = Ap$$

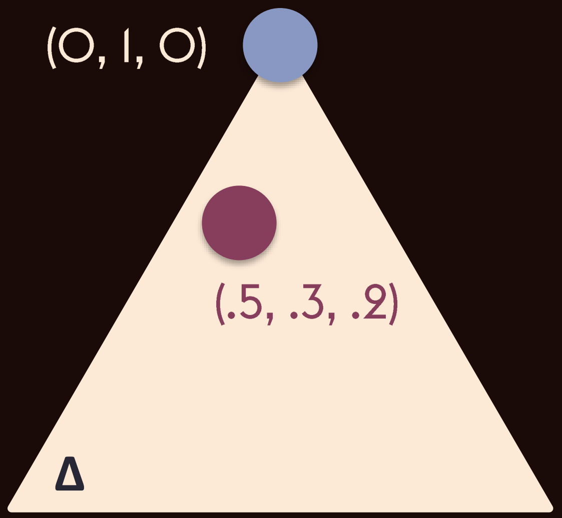
$\operatorname{argmax} \langle \boldsymbol{x}, \boldsymbol{p} \rangle$
s.t. $\boldsymbol{p} \in \Delta$

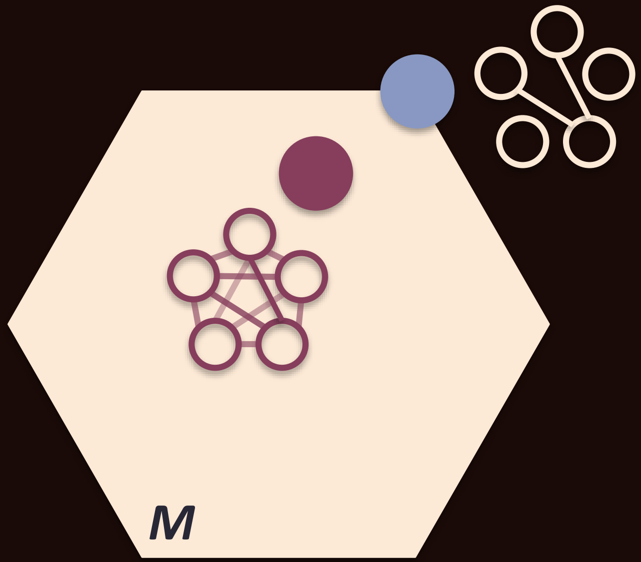
MAP

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

$\boldsymbol{p}^* = \boldsymbol{e}_i$ where $i = \operatorname{argmax}(\boldsymbol{x})$

$\boldsymbol{x} = \boldsymbol{A}^\top \boldsymbol{\eta}$

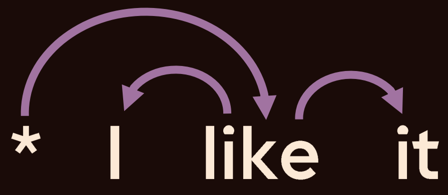




$\boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{p}$

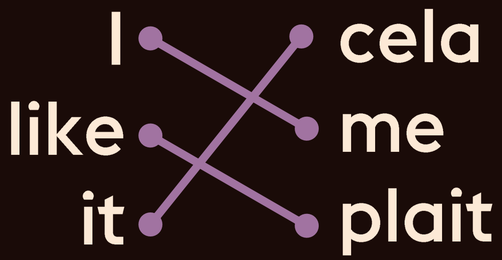
MAP inference:

Maximum spanning tree (Chu-Liu/Edmonds)



$$A=\begin{bmatrix} 1 & & 0 & 0 & \\ 0 & & 1 & 1 & \\ 0 & & 0 & 0 & \\ \hdashline 0 & & 1 & 1 & \\ 1 & \dots & 0 & 0 & \dots \\ 0 & & 0 & 0 & \\ \hdashline 0 & & 0 & 0 & \\ 0 & & 1 & 0 & \\ 1 & & 0 & 1 & \end{bmatrix} \quad \begin{matrix} * \to I \\ \text{like} \to I \\ \text{it} \to I \\ * \to \text{like} \\ I \to \text{like} \\ \text{it} \to \text{like} \\ * \to \text{it} \\ I \to \text{it} \\ \text{like} \to \text{it} \end{matrix} \qquad \eta=\begin{bmatrix} .3 \\ .8 \\ -.5 \\ \hdashline .2 \\ -.1 \\ -.2 \\ \hdashline .7 \\ .6 \\ .1 \end{bmatrix}$$

Hungarian algorithm



$$A=\begin{bmatrix} 1 & & 0 & 0 & \\ 0 & & 1 & 0 & \\ 0 & & 0 & 1 & \\ \hdashline 0 & & 0 & 1 & \\ 1 & \dots & 0 & 0 & \dots \\ 0 & & 1 & 0 & \\ \hdashline 0 & & 1 & 0 & \\ 0 & & 0 & 1 & \\ 1 & & 0 & 0 & \end{bmatrix} \quad \begin{matrix} I - \text{cela} \\ I - \text{me} \\ I - \text{plait} \\ \text{like} - \text{cela} \\ \text{like} - \text{me} \\ \text{like} - \text{plait} \\ \text{it} - \text{cela} \\ \text{it} - \text{me} \\ \text{it} - \text{plait} \end{matrix} \qquad \eta=\begin{bmatrix} .3 \\ .8 \\ -.5 \\ \hdashline .2 \\ -.1 \\ -.2 \\ \hdashline .7 \\ .6 \\ .1 \end{bmatrix}$$

MAP

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

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

Marginal

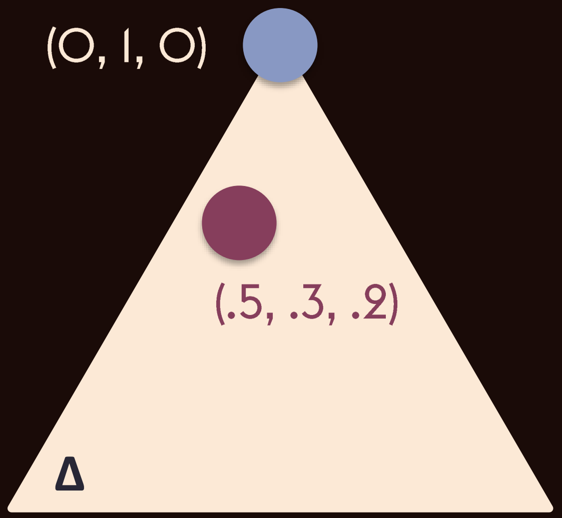
$$\operatorname{argmax} \langle \boldsymbol{x}, \boldsymbol{p} \rangle + \mathrm{H}(\boldsymbol{p}) \quad \text{s.t.} \quad \boldsymbol{p} \in \Delta$$

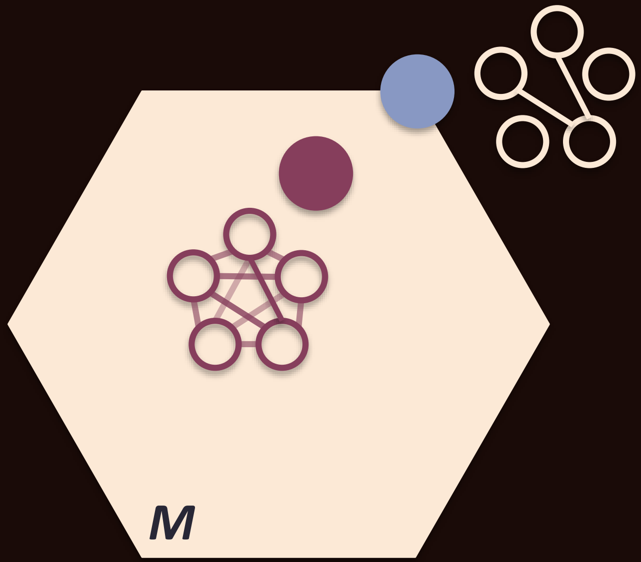
softmax, closed-form solution: $\boldsymbol{p}^* = \exp(\boldsymbol{x}) / Z$

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

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

$$\boldsymbol{x} = \boldsymbol{A}^{\top} \boldsymbol{\eta}$$





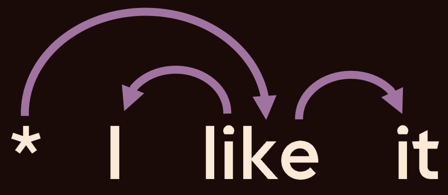
$$\boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{p}$$

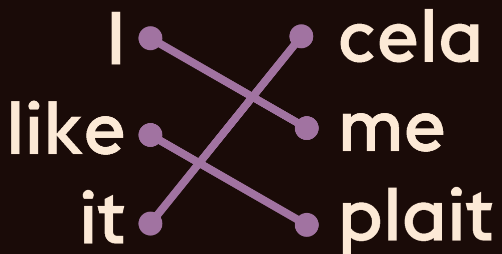
MAP inference:
Maximum spanning tree

Hungarian algorithm

Marginal inference:
Matrix-Tree theorem

#P complete





A =

1		0	0		*	→ I
0		1	1		like	→ I
0		0	0		it	→ I
0		1	1		*	→ like
1	…	0	0	…	I	→ like
0		0	0		it	→ like
0		0	0		*	→ it
0		1	0		I	→ it
1		0	1		like	→ it

η = [.3, .8, -.5, .2, -.1, -.2, .7, .6, .1]

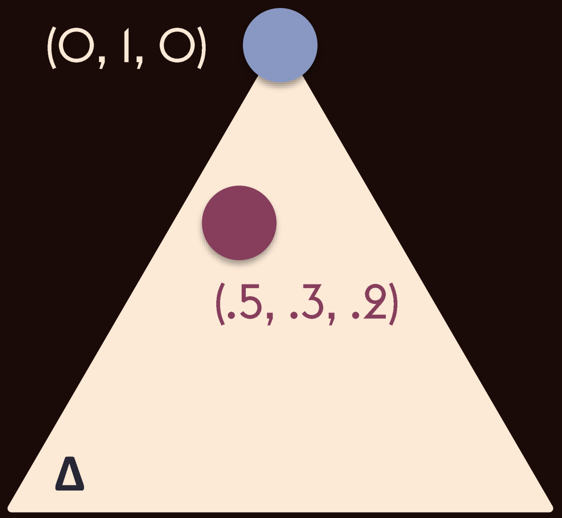
A =

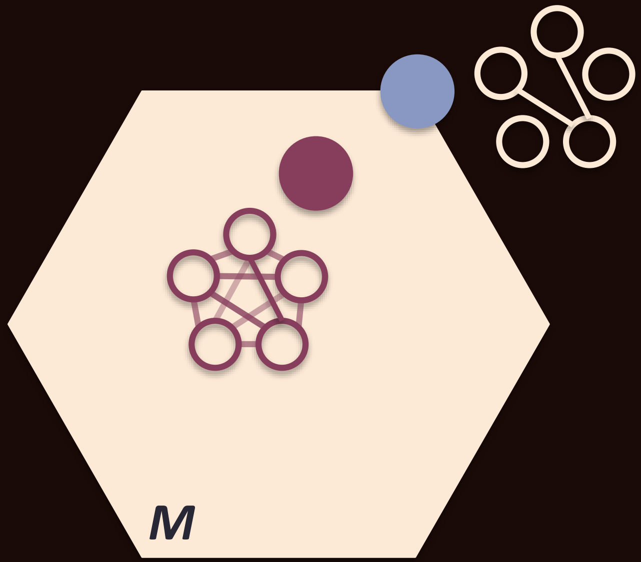
1		0	0		I	- cela
0		1	0		I	- me
0		0	1		I	- plait
0		0	1		like	- cela
1	…	0	0	…	like	- me
0		1	0		like	- plait
0		1	0		it	- cela
0		0	1		it	- me
1		0	0		it	- plait

η = [.3, .8, -.5, .2, -.1, -.2, .7, .6, .1]

$\operatorname{argmax} \langle \boldsymbol{x}, \boldsymbol{p} \rangle$ s.t. $\boldsymbol{p} \in \Delta$	**MAP**	$\operatorname{argmax} \langle \boldsymbol{\eta}, \boldsymbol{\mu} \rangle$ s.t. $\boldsymbol{\mu} \in M$
$\operatorname{argmax} \langle \boldsymbol{x}, \boldsymbol{p} \rangle + H(\boldsymbol{p})$ s.t. $\boldsymbol{p} \in \Delta$	**Marginal**	$\operatorname{argmax} \langle \boldsymbol{\eta}, \boldsymbol{\mu} \rangle + \tilde{H}(\boldsymbol{\mu})$ s.t. $\boldsymbol{\mu} \in M$

$$\boldsymbol{x} = \boldsymbol{A}^\top \boldsymbol{\eta}$$





$$\boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{p}$$

MAP

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

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

Marginal

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

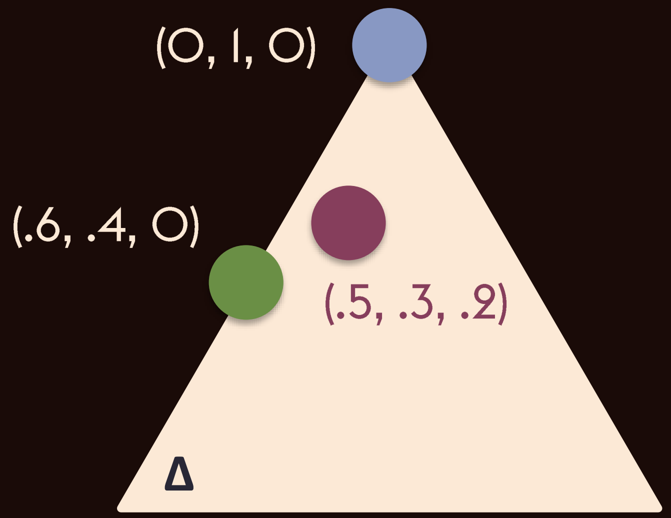
$$\operatorname{argmax} \langle \boldsymbol{\eta}, \boldsymbol{\mu} \rangle + \tilde{\mathrm{H}}(\boldsymbol{\mu}) \quad \text{s.t.} \quad \boldsymbol{\mu} \in \boldsymbol{M}$$

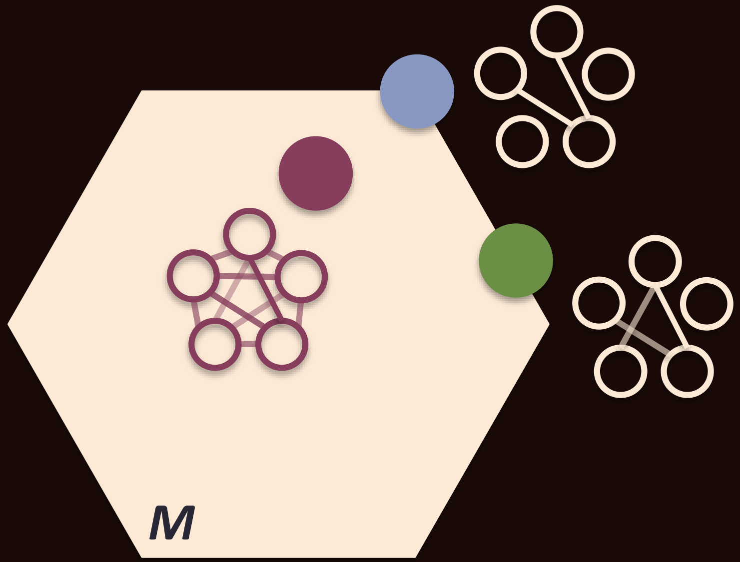
SparseMAP

$$\operatorname{argmax} \langle \boldsymbol{x}, \boldsymbol{p} \rangle - \tfrac{1}{2} \|\boldsymbol{A}\boldsymbol{p}\|^2 \quad \text{s.t.} \quad \boldsymbol{p} \in \boldsymbol{\Delta}$$

$$\operatorname{argmax} \langle \boldsymbol{\eta}, \boldsymbol{\mu} \rangle - \tfrac{1}{2} \|\boldsymbol{\mu}\|^2 \quad \text{s.t.} \quad \boldsymbol{\mu} \in \boldsymbol{M}$$

$$\boldsymbol{x} = \boldsymbol{A}^\top \boldsymbol{\eta}$$





$$\boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{p}$$

Efficiently Computing SparseMAP

$$\operatorname{argmax} \langle \eta, \mu \rangle - \tfrac{1}{2} \|\mu\|^2 \quad \text{s.t.} \quad \mu \in M$$

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

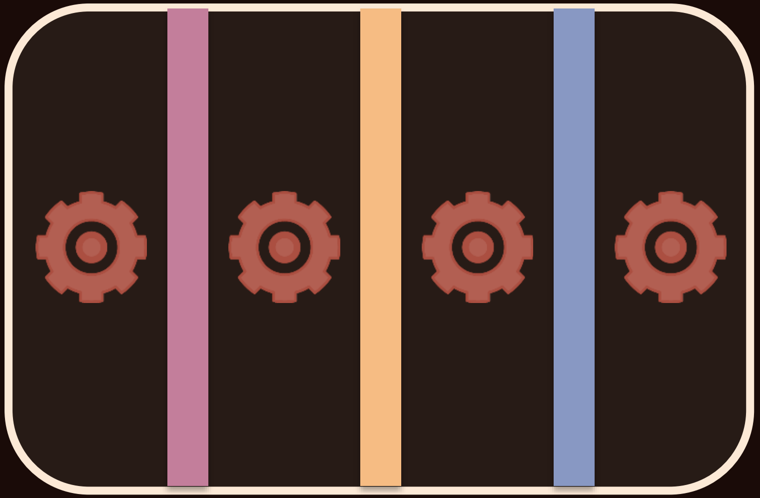
Sparse Latent Structure

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.

Model: ESIM [Chen & al, 2017]

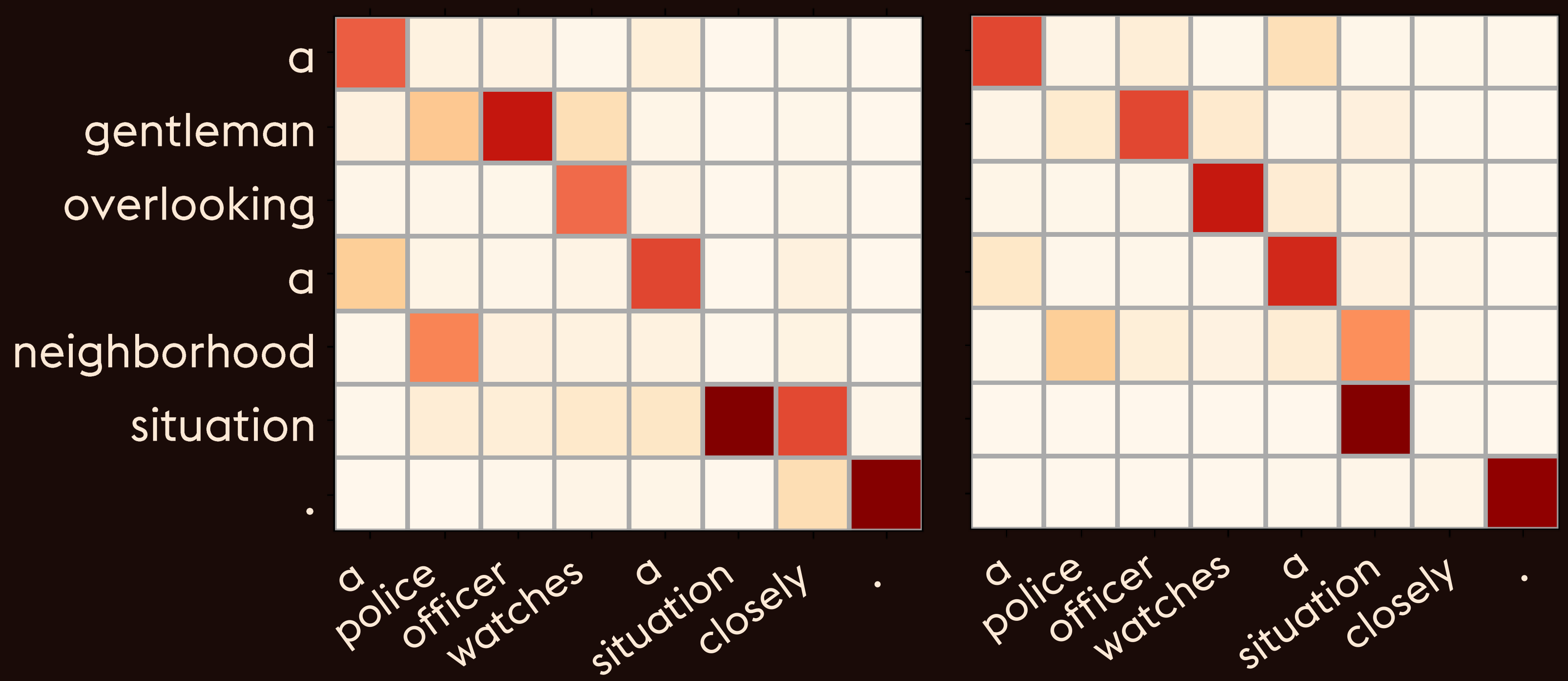
Natural Language Inference

Prem: A gentleman overlooking a neighborhood situation.

Hypo: A police officer watches a situation closely.

Model: ESIM [Chen & al, 2017]

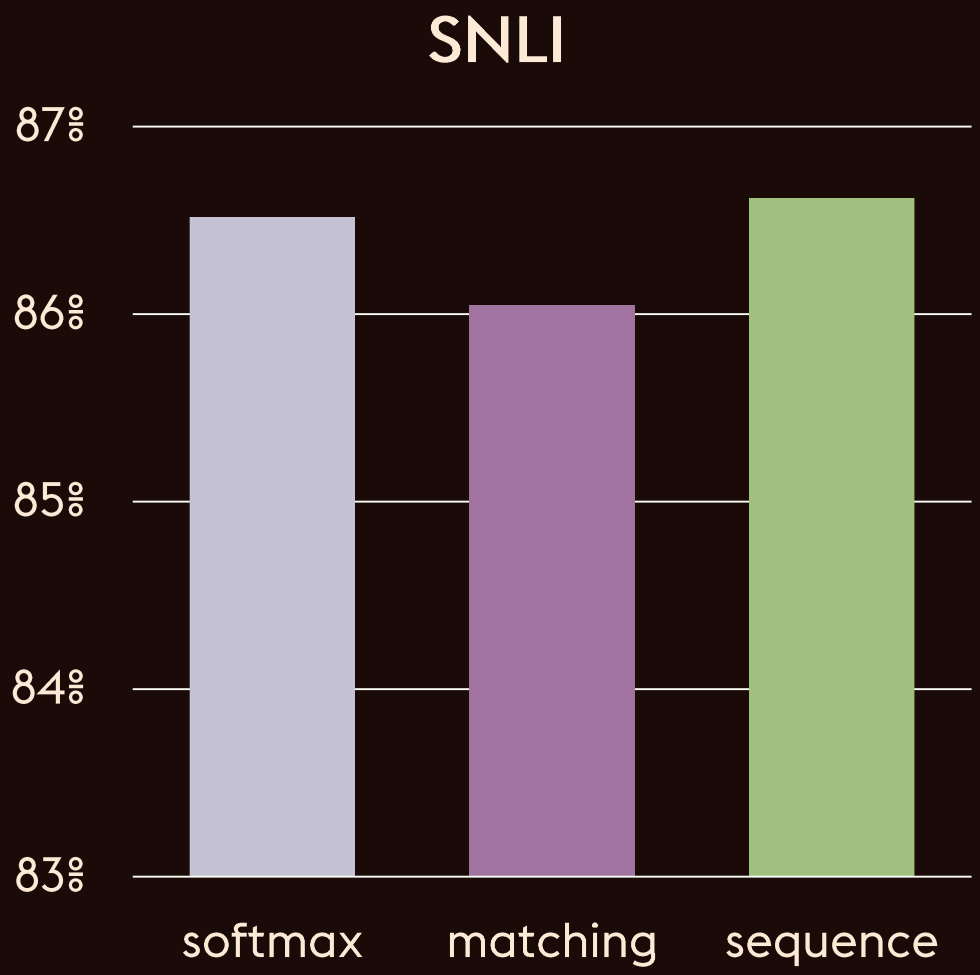
Natural Language Inference

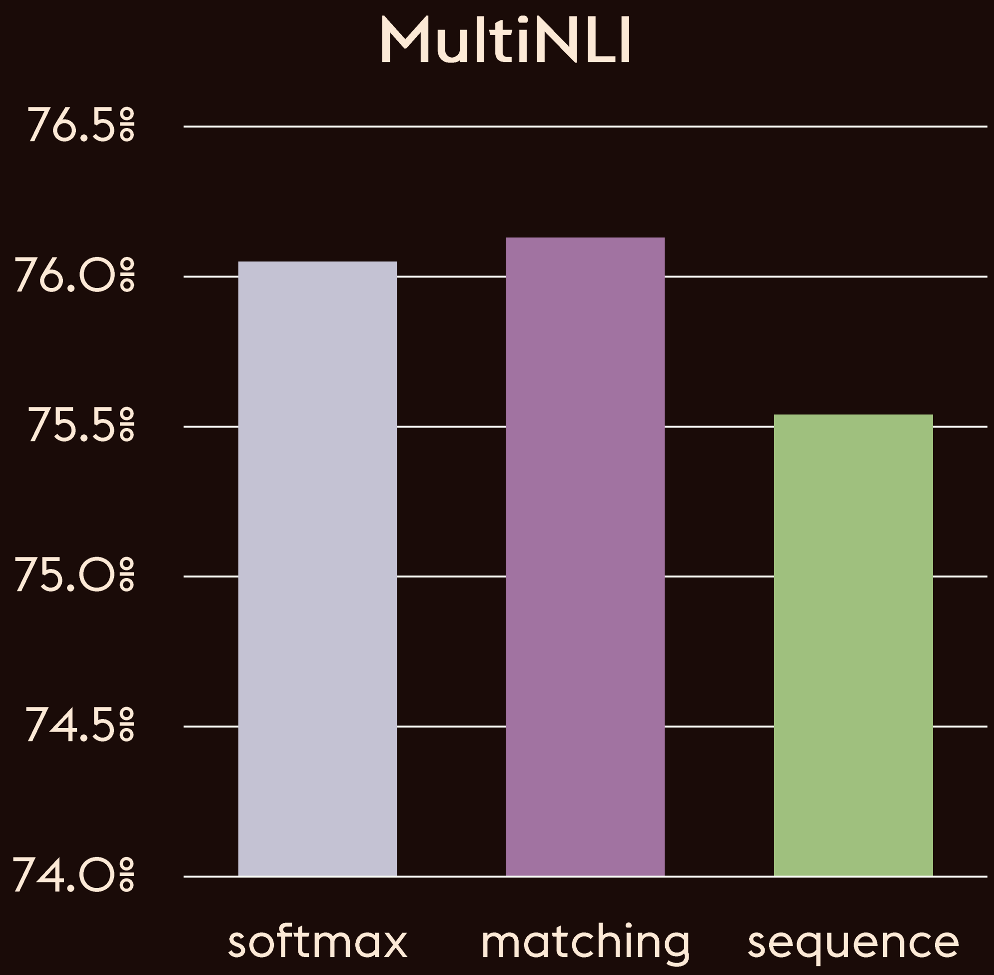


Natural Language Inference with Linear Assignment

Prem: A gentleman overlooking a neighborhood situation.

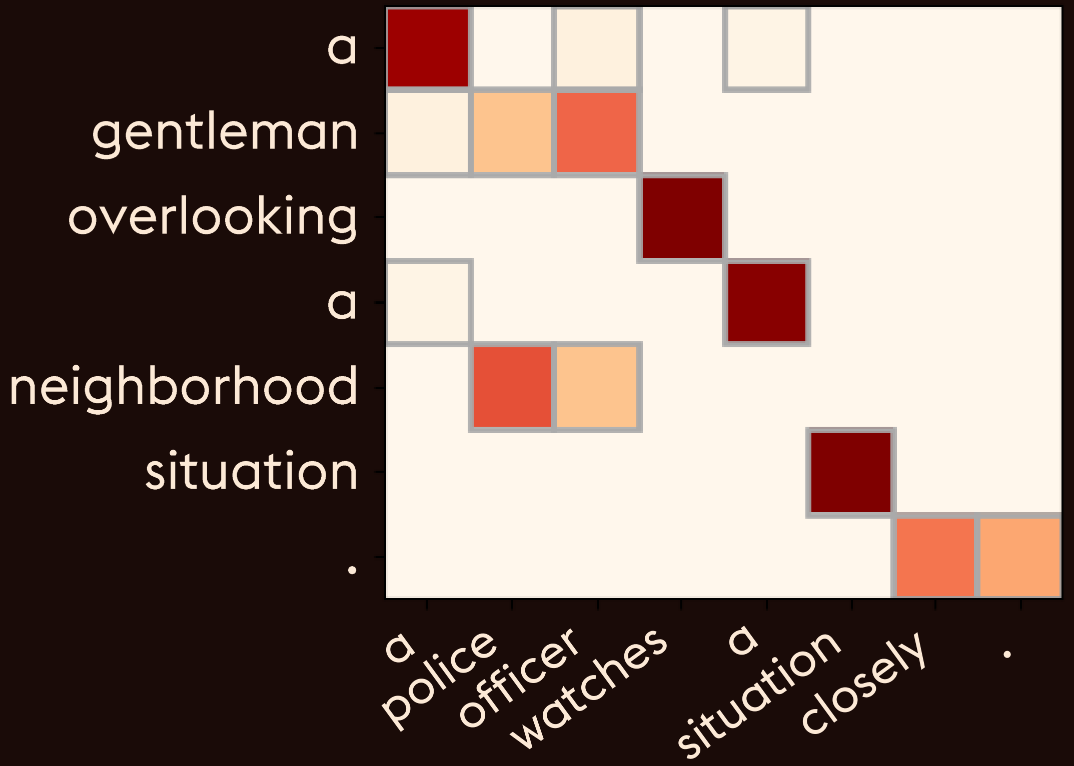
Hypo: A police officer watches a situation closely.



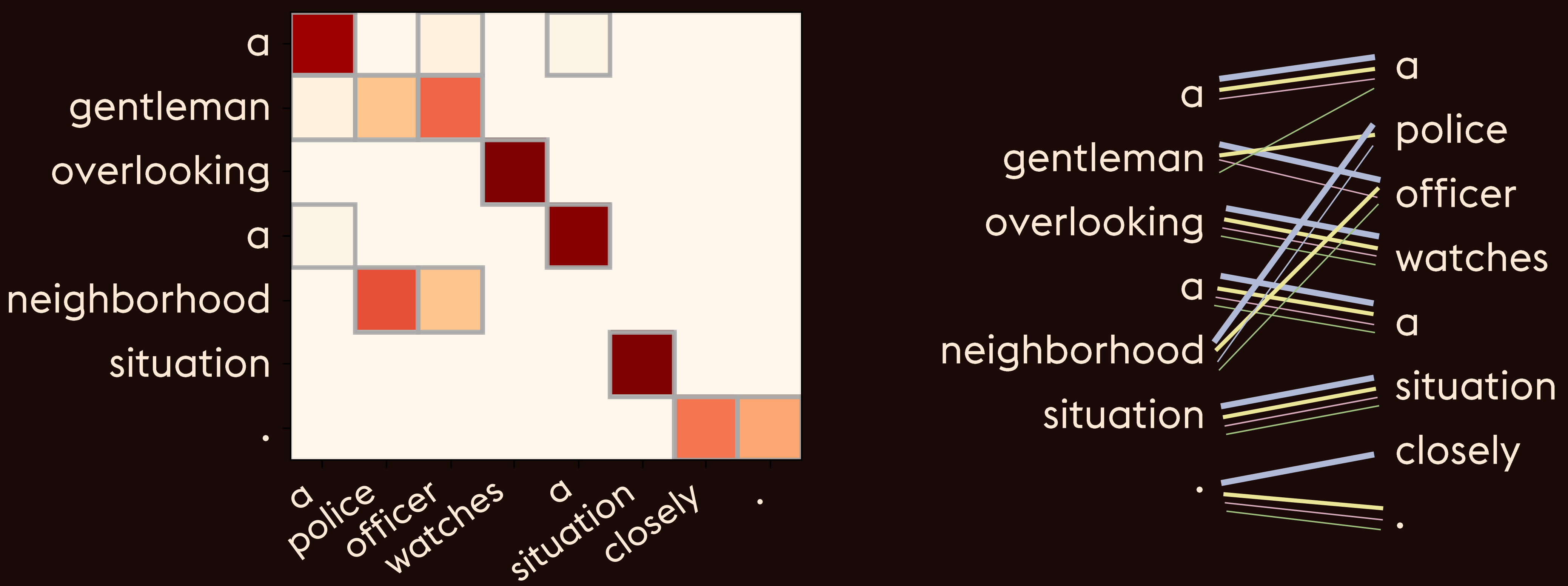


(3-class accuracy)

Natural Language Inference with Linear Assignment



Natural Language Inference with Linear Assignment



Sparse Structured Output Prediction

Sparse Structured Output Prediction

scores — gold structure

SparseMAP loss

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

cost (as in structured SVM)

margin-SparseMAP loss

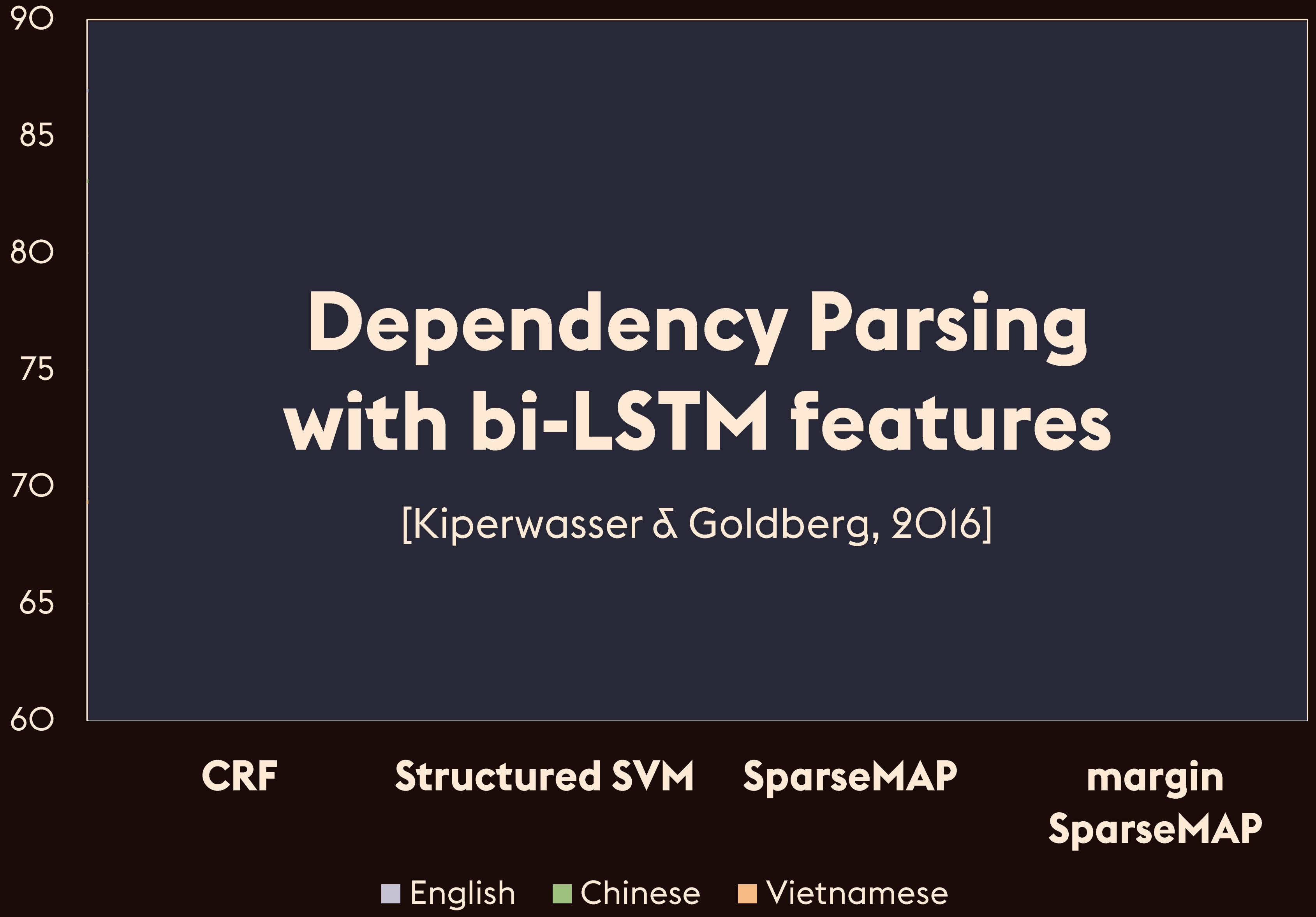
$$L_A^{\rho}(\eta, \bar{\mu}) = \max_{\mu \in M} \left\{ \langle \eta, \mu \rangle - \tfrac{1}{2} \|\mu\|^2 + \rho(\mu, \bar{\mu}) \right\} - \langle \eta, \bar{\mu} \rangle + \tfrac{1}{2} \|\bar{\mu}\|^2$$

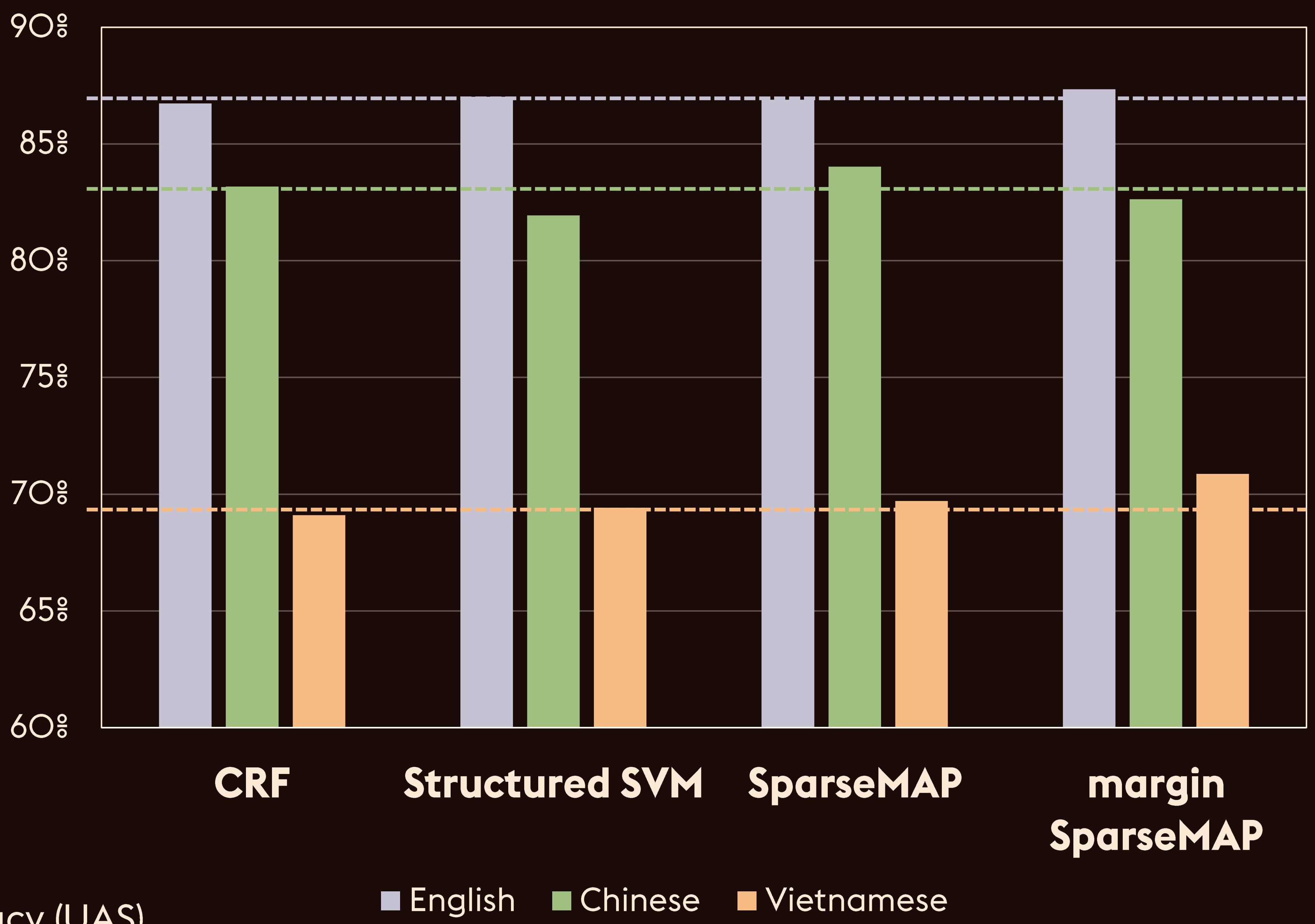
Instance of a structured Fenchel-Young loss, like CRF, SVM, etc.

[Blondel, Martins, Niculae '18]

Dependency Parsing with bi-LSTM features

[Kiperwasser & Goldberg, 2016]

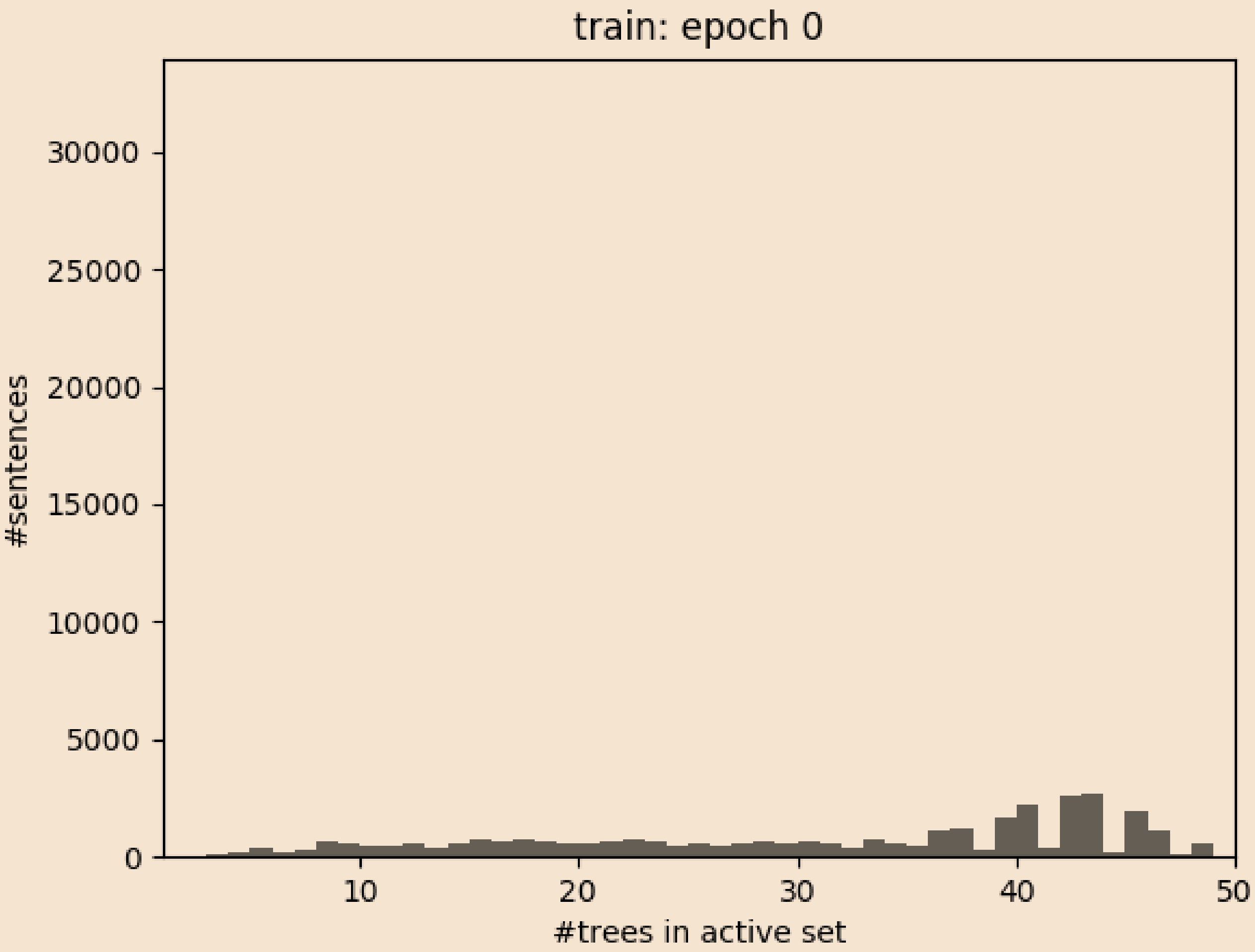


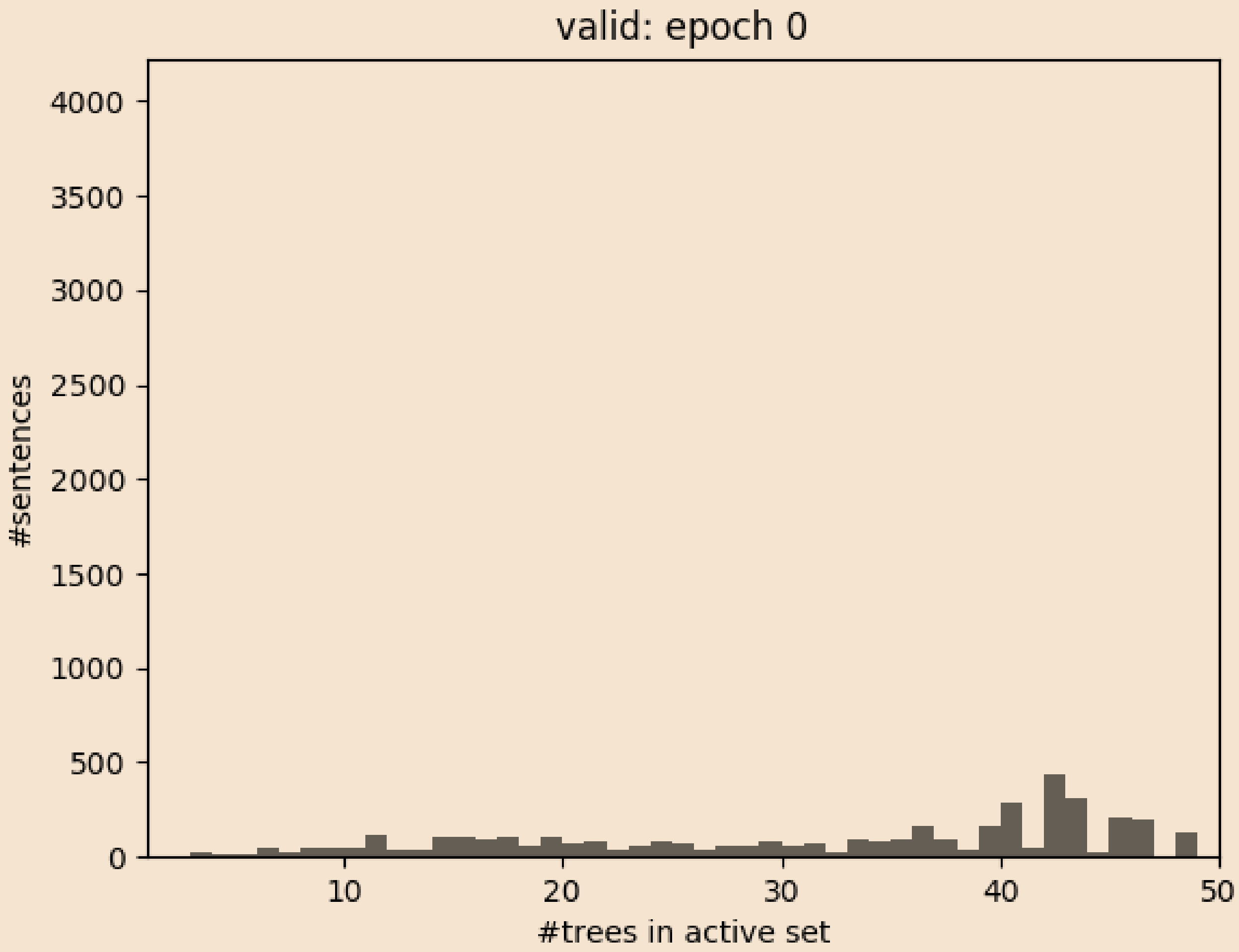


Unlabeled Accuracy (UAS)
Universal Dependencies dataset

Sparse Structured Output Prediction

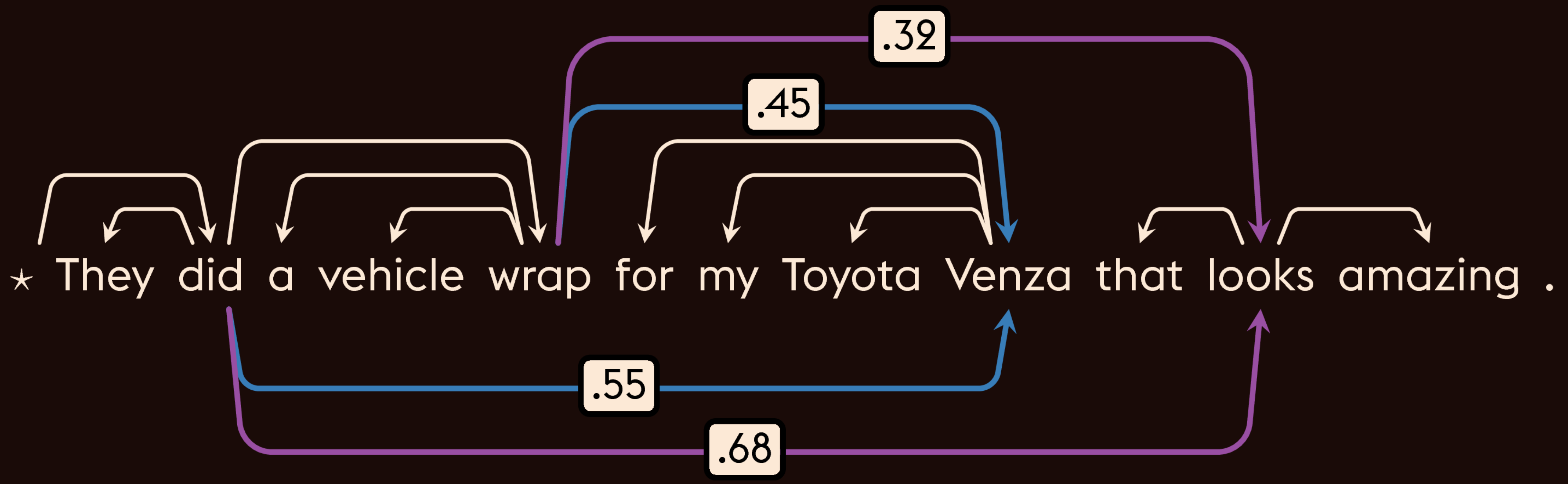
As models train, inference gets sparser!





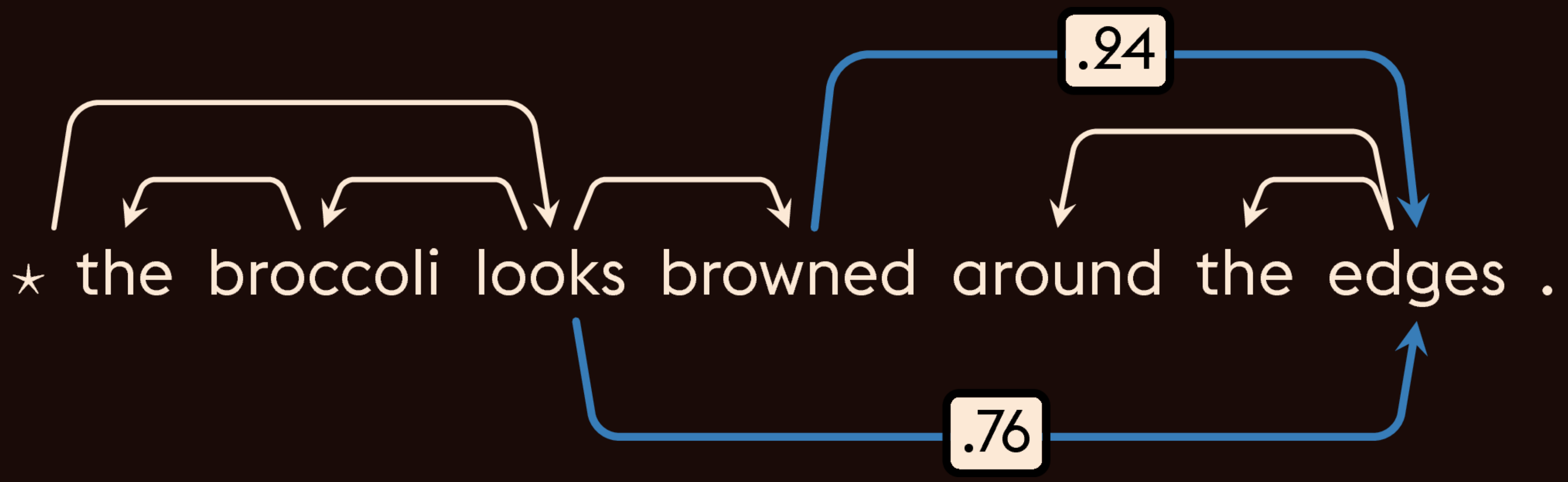
Sparse Structured Output Prediction

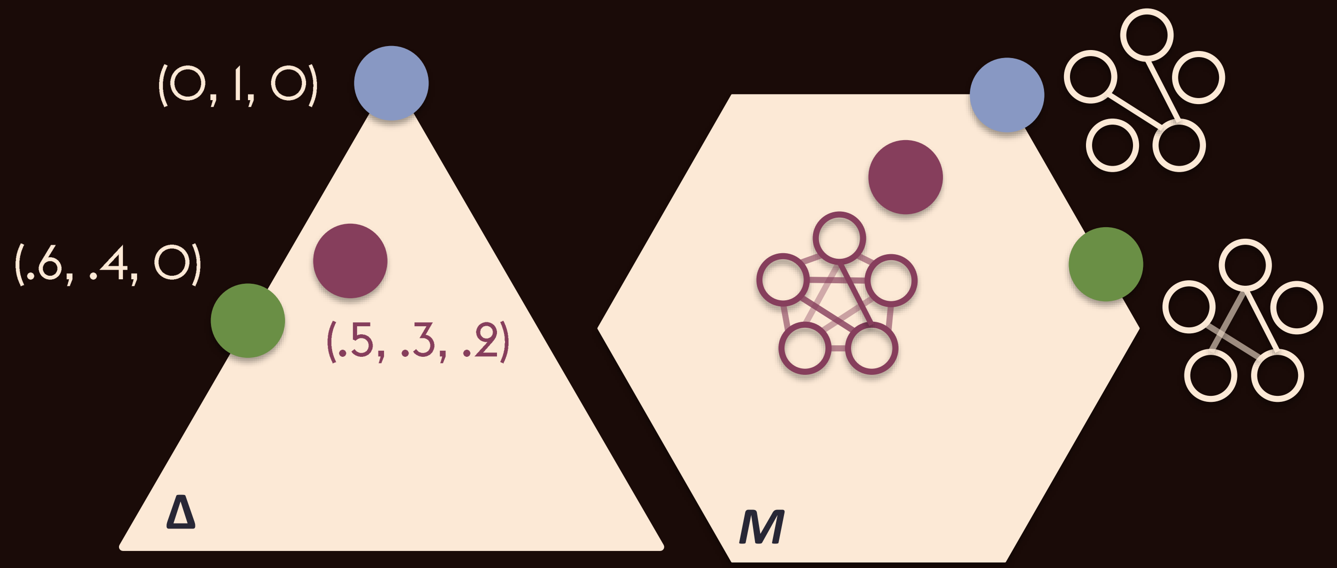
Inference captures linguistic ambiguity!



Sparse Structured Output Prediction

Inference captures linguistic ambiguity!





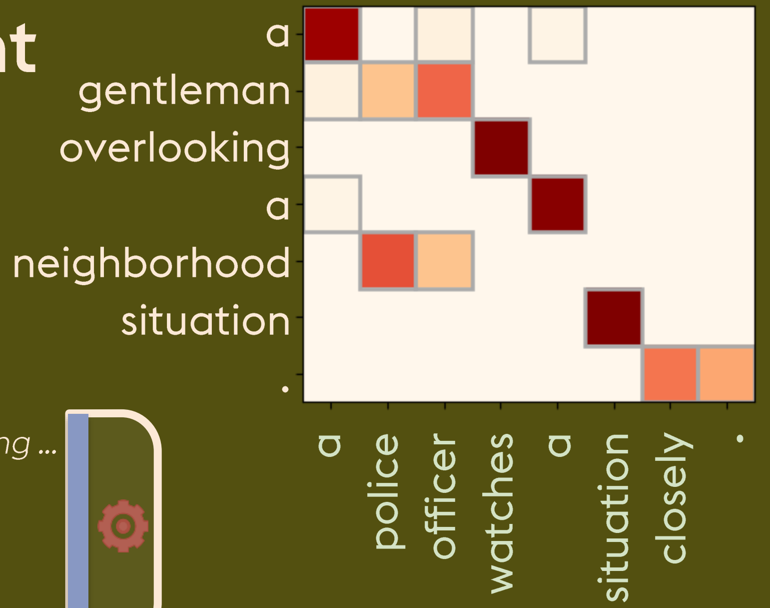
poster #66 tonight @6:15

github.com/vene/sparsemap

https://vene.ro

vnfrombucharest

Sparse Latent Structure



Sparse Structured Output Prediction