# Interpretable Neural Predictions with Differentiable Binary Variables

Joost Bastings    Wilker Aziz    Ivan Titov

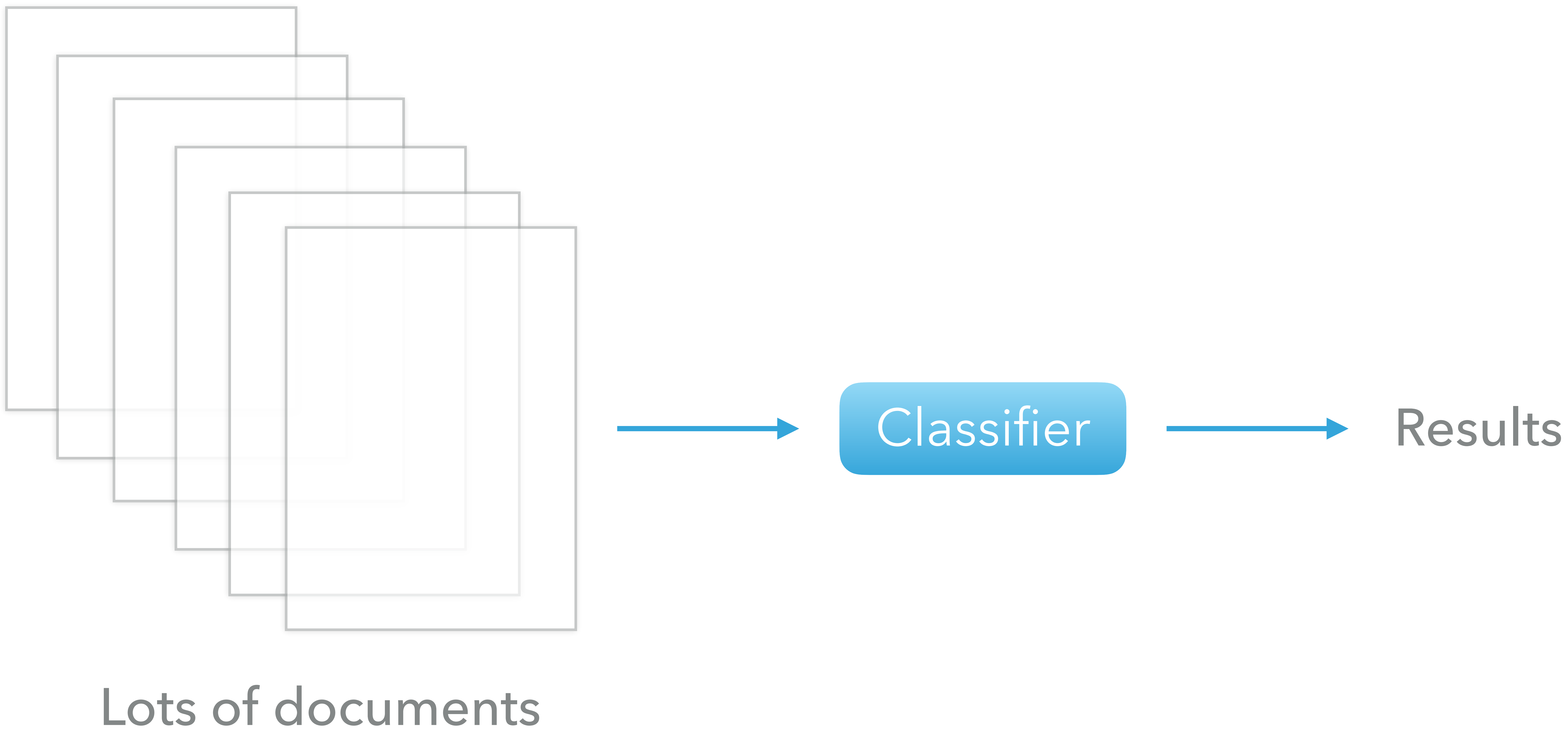
ILLC, University of Amsterdam
ILCC, University of Edinburgh

bastings.github.io

ACL, 30 July 2019

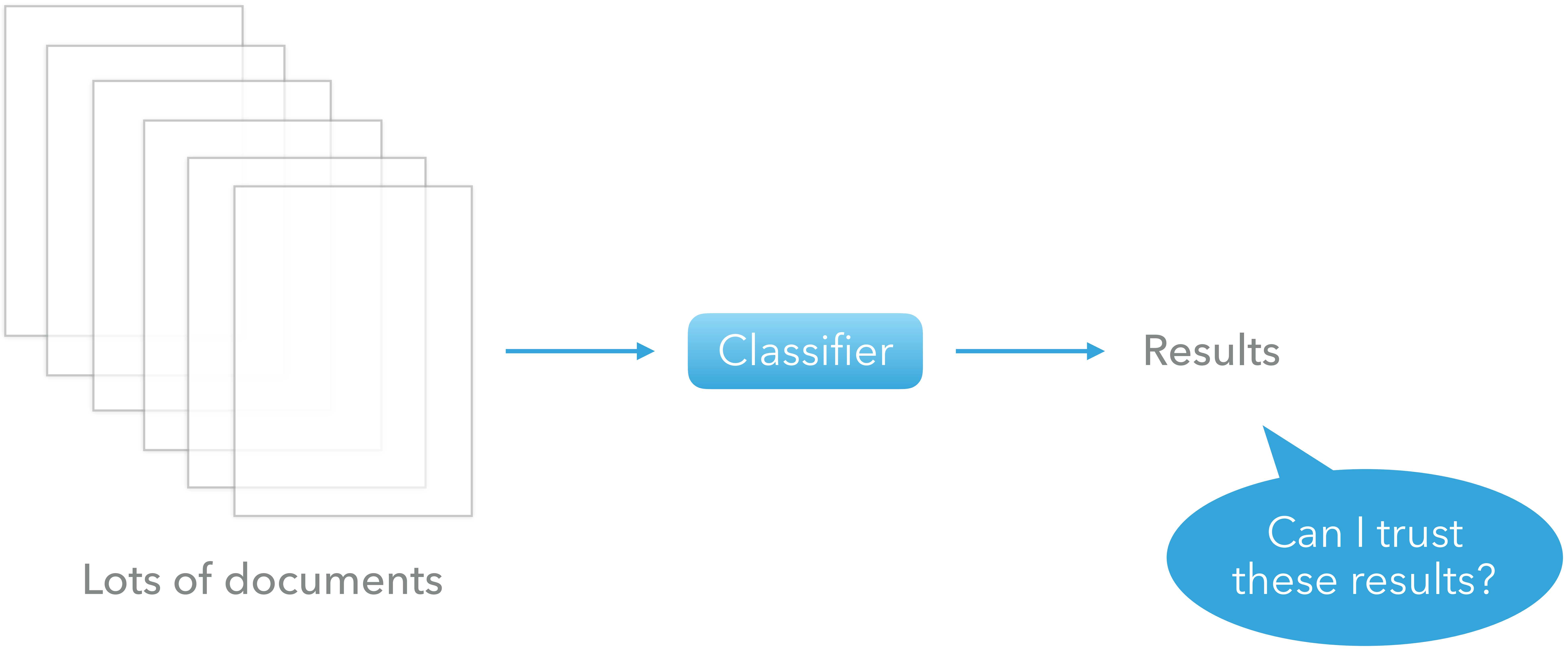
# Interpreting neural networks is difficult





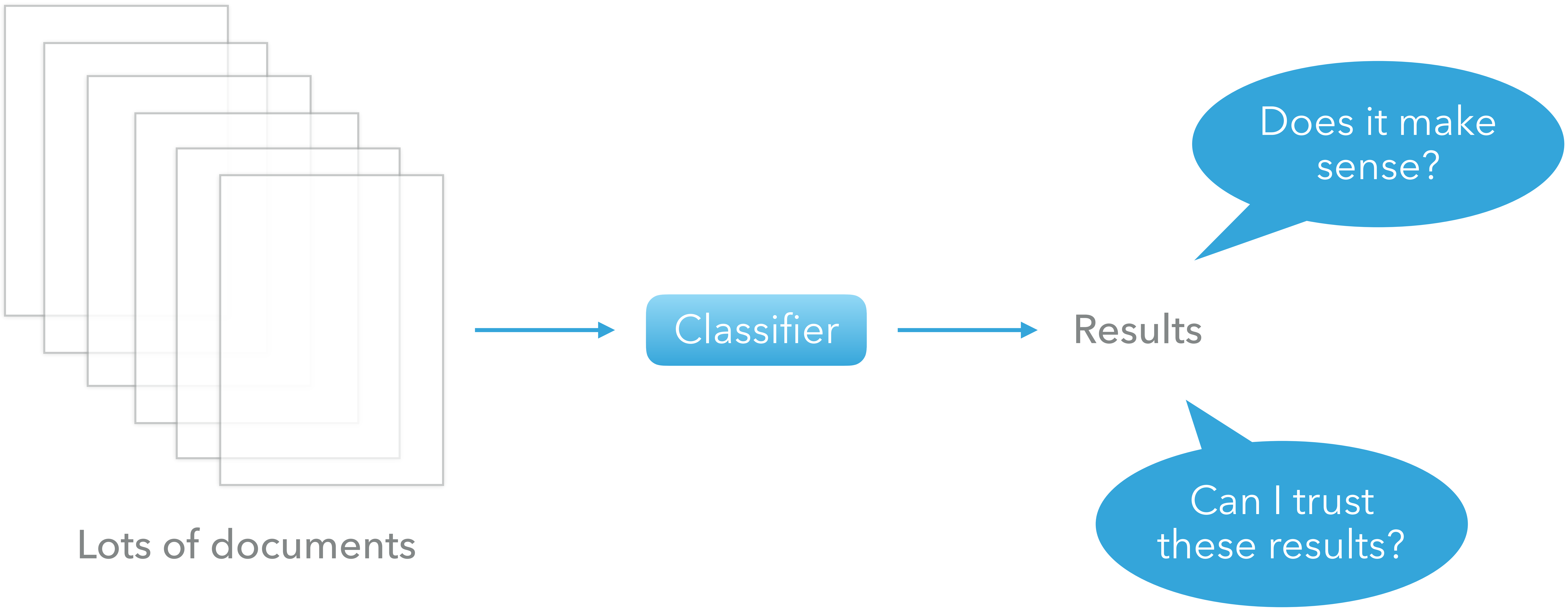
# Interpreting neural networks is difficult



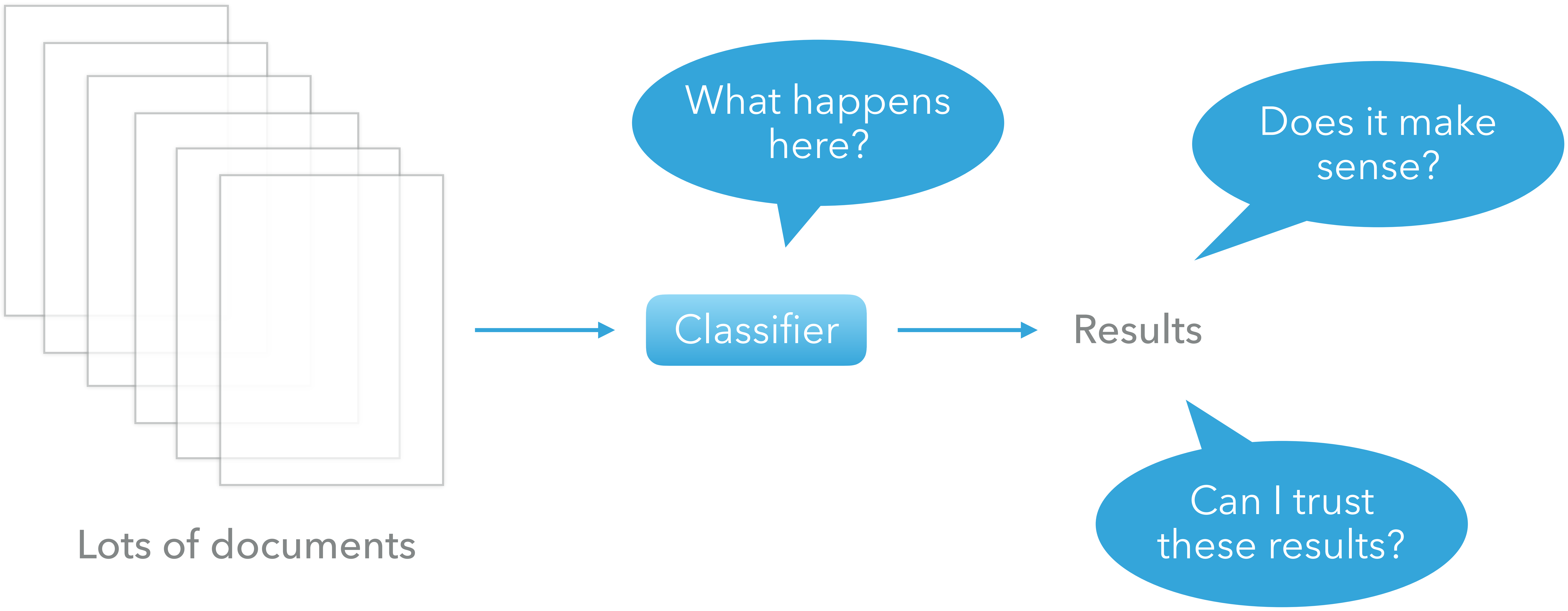


# Interpreting neural networks is difficult





# Interpreting neural networks is difficult



# Approach (Lei et al., 2016)



pours a dark amber color with decent head that does not recede much . it 's a tad too dark to see the carbonation , but fairs well . smells of roasted malts and mouthfeel is quite strong in the sense that you can get a good taste of it before you even swallow .

# Approach (Lei et al., 2016)



pours **a dark amber color with decent head** that does not recede much . it 's **a tad too dark to see the carbonation , but fairs well .** smells of roasted malts and mouthfeel is quite strong in the sense that you can get a good taste of it before you even swallow .

Rationale Extractor

pours a dark amber color with decent head that does not recede much . it 's a tad too dark to see the carbonation , but fairs well . smells of roasted malts and mouthfeel is quite strong in the sense that you can get a good taste of it before you even swallow .

# Approach (Lei et al., 2016)



look:
★★★★

Classifier

Rationale

pours **a dark amber color with decent head** that does not recede much . it 's **a tad too dark to see the carbonation , but fairs well .** smells of roasted malts and mouthfeel is quite strong in the sense that you can get a good taste of it before you even swallow .

Rationale Extractor

pours a dark amber color with decent head that does not recede much . it 's a tad too dark to see the carbonation , but fairs well . smells of roasted malts and mouthfeel is quite strong in the sense that you can get a good taste of it before you even swallow .

# Rationale



Enough to make the right prediction

A rationale is a **short** and **sufficient** part of the input text.

To be a good explanation

# Text classification with rationales (Lei et al., 2016)



$$Z_i \mid x \sim \text{Bernoulli}\left(g_i(x;\phi)\right)$$

Rationale Extractor

NN that predicts a sequence *n* of Bernoulli parameters

# Text classification with rationales (Lei et al., 2016)



Classifier

$z_i \in \{0,1\}$ can erase $x_i$

$$Y \mid x, z \sim \mathrm{Cat}\big(f(x \odot z; \theta)\big)$$

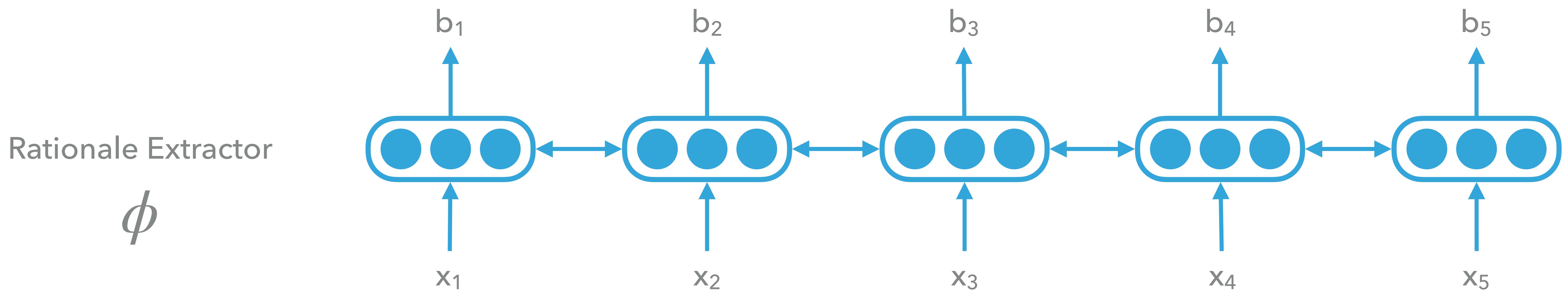
$$Z_i \mid x \sim \mathrm{Bernoulli}\big(g_i(x; \phi)\big)$$

Rationale Extractor

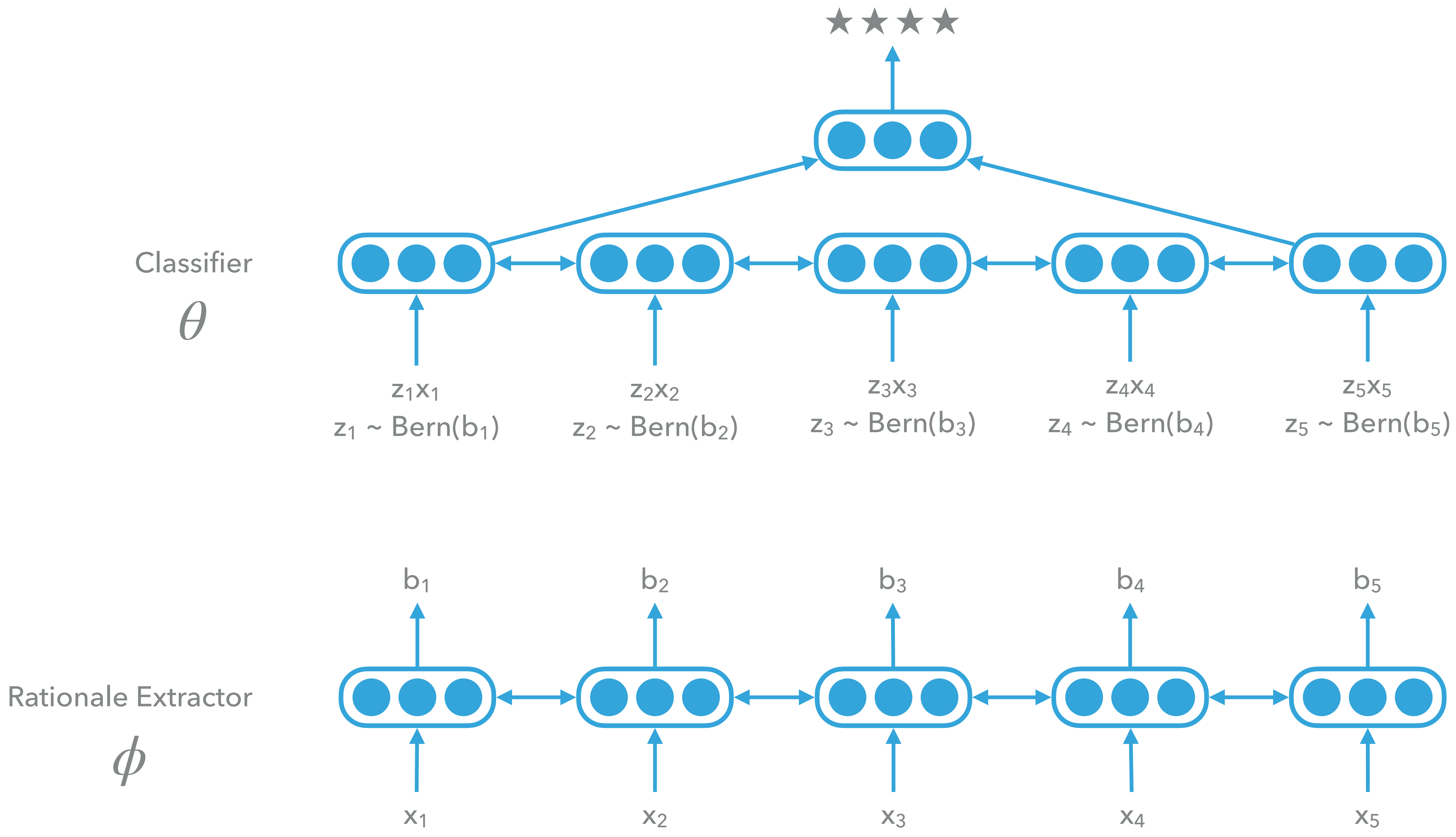
NN that predicts a sequence *n* of Bernoulli parameters

# Model of Lei et al. (2016)



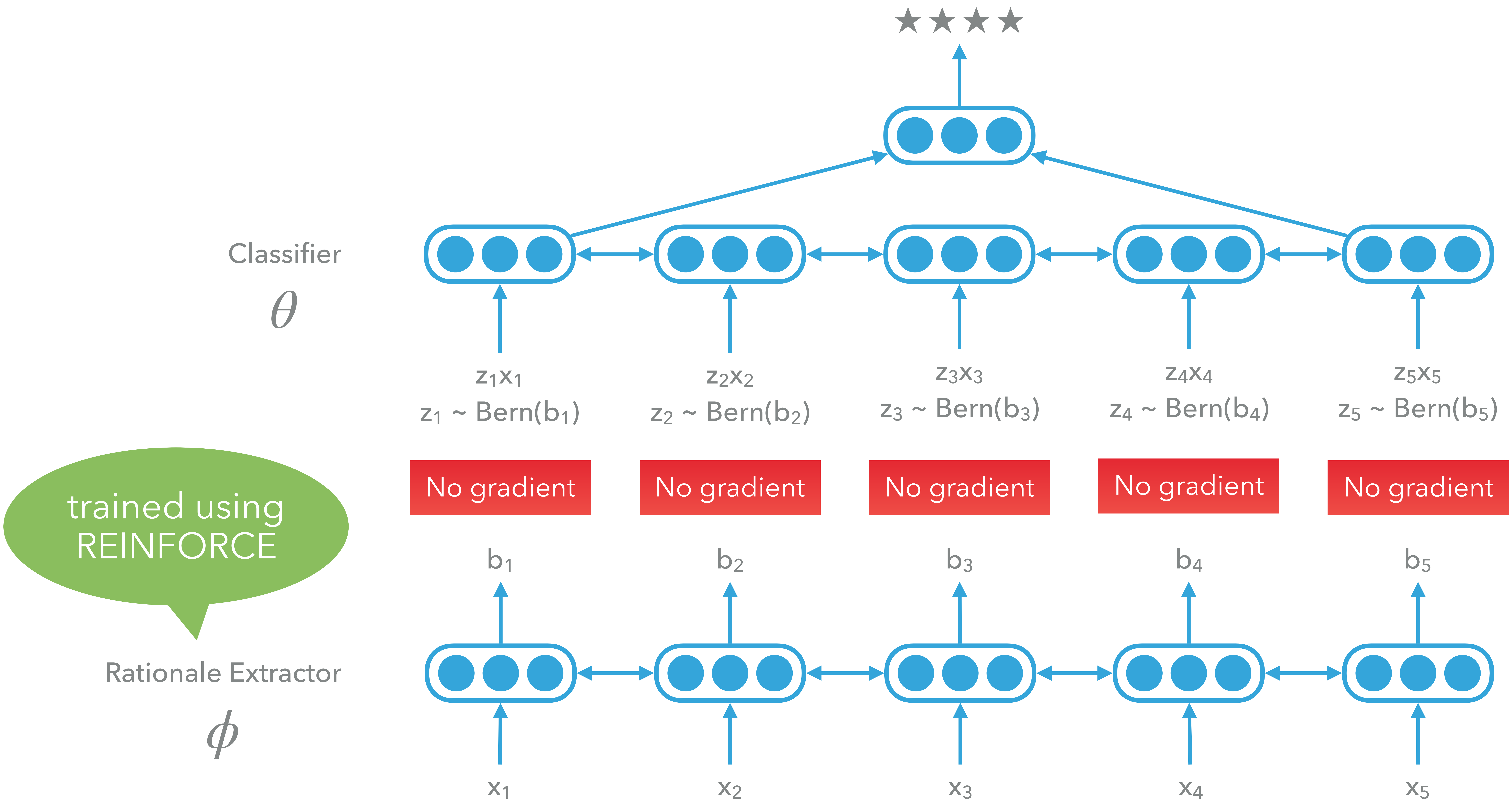


# Model of Lei et al. (2016)



# Model of Lei et al. (2016)



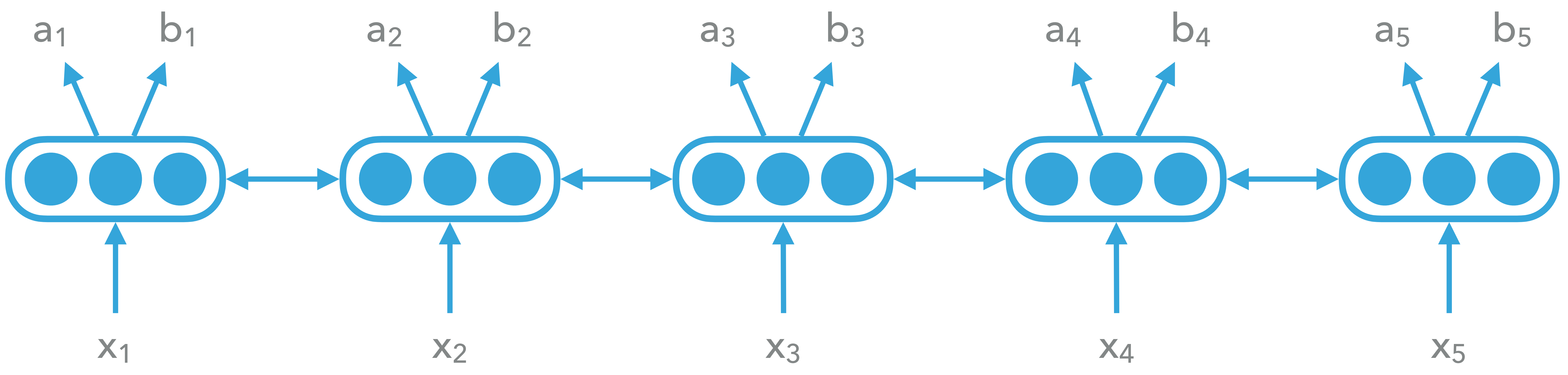


# Our Proposed Model



Rationale Extractor

$\phi$



# Our Proposed Model



# Our Proposed Model

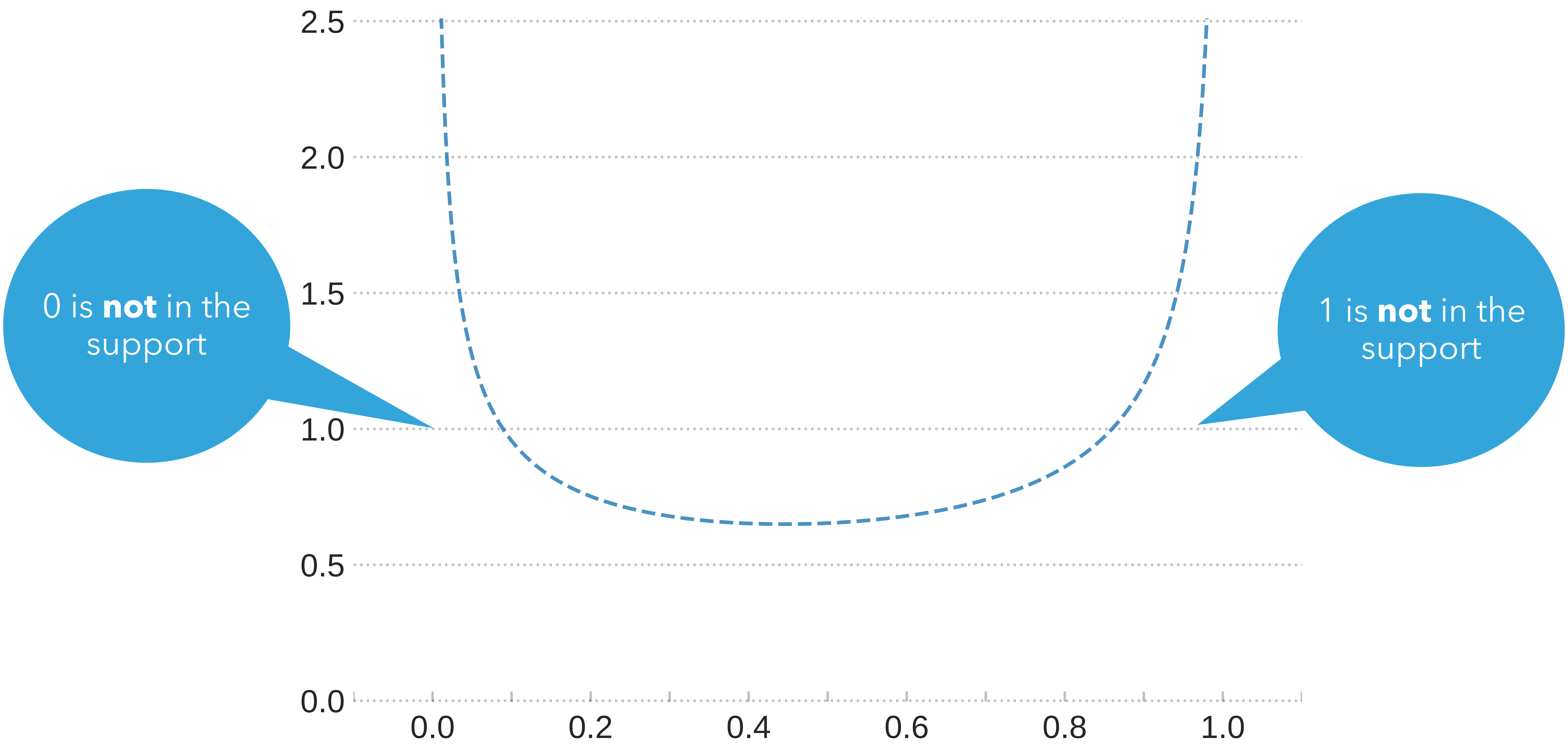


# Our Proposed Model



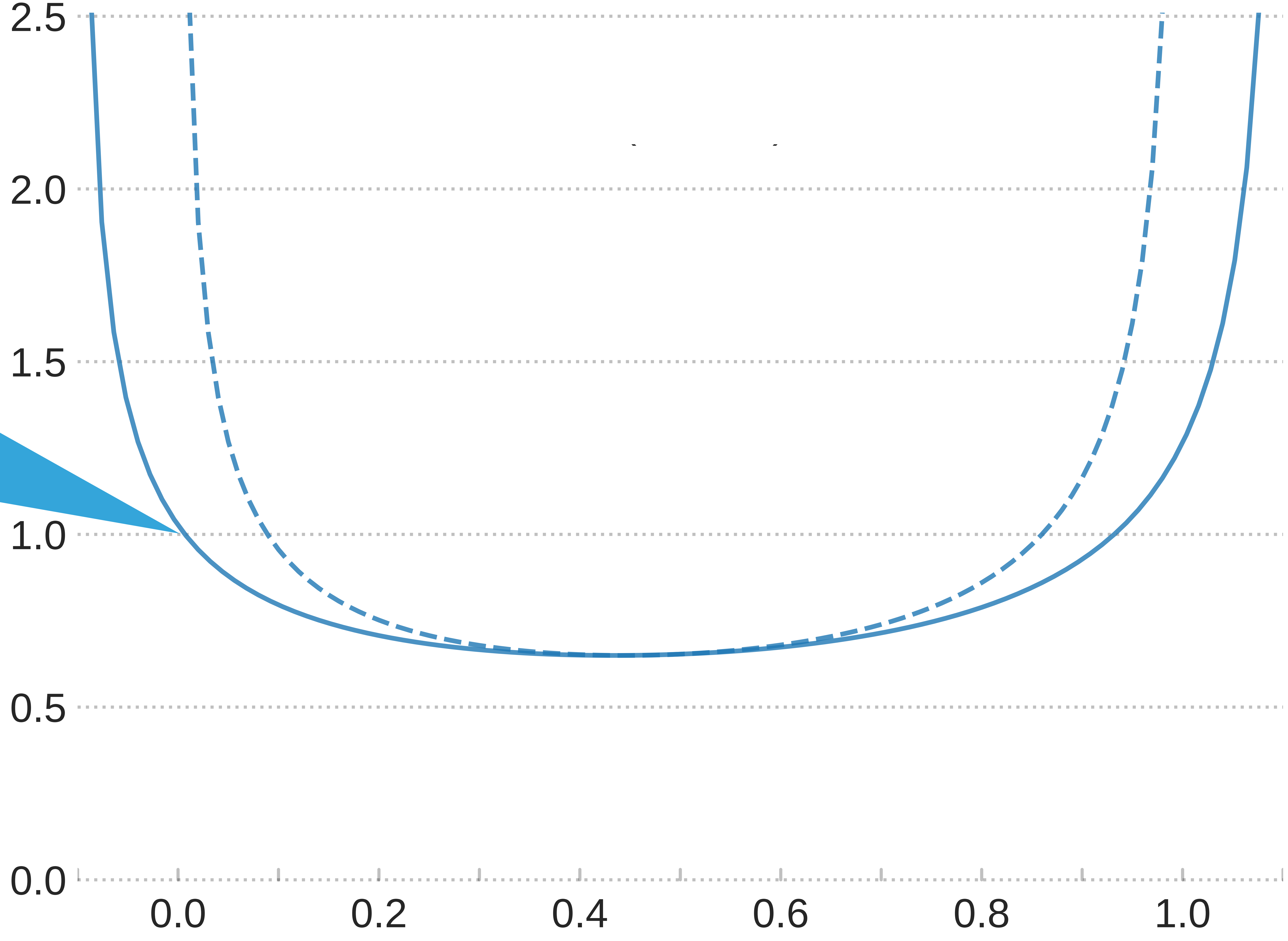
# Stretch-and-rectify (Louizos et al., 2018)





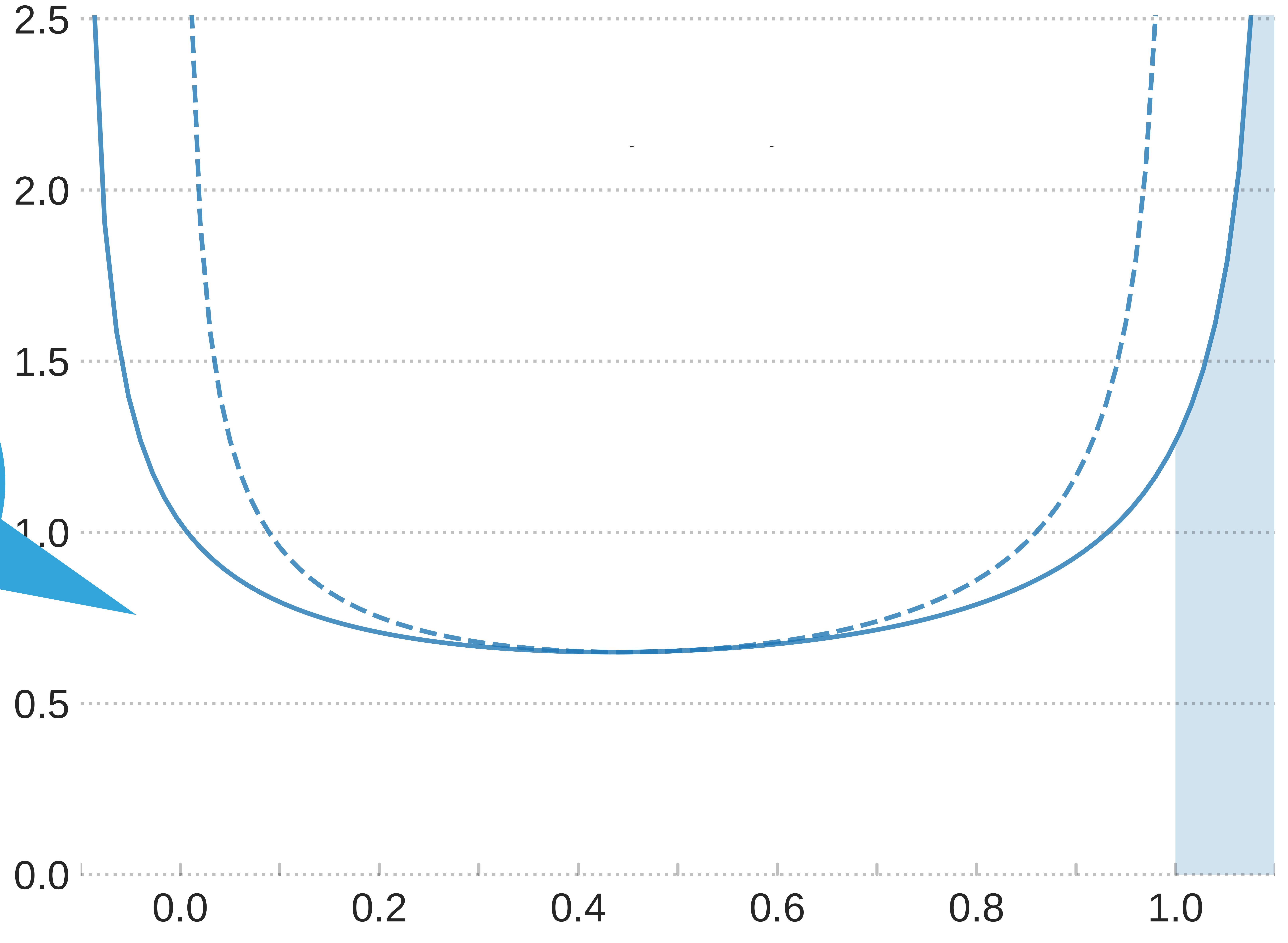
# Stretch-and-rectify (Louizos et al., 2018)






**0 is in the support now**, but we will never sample it

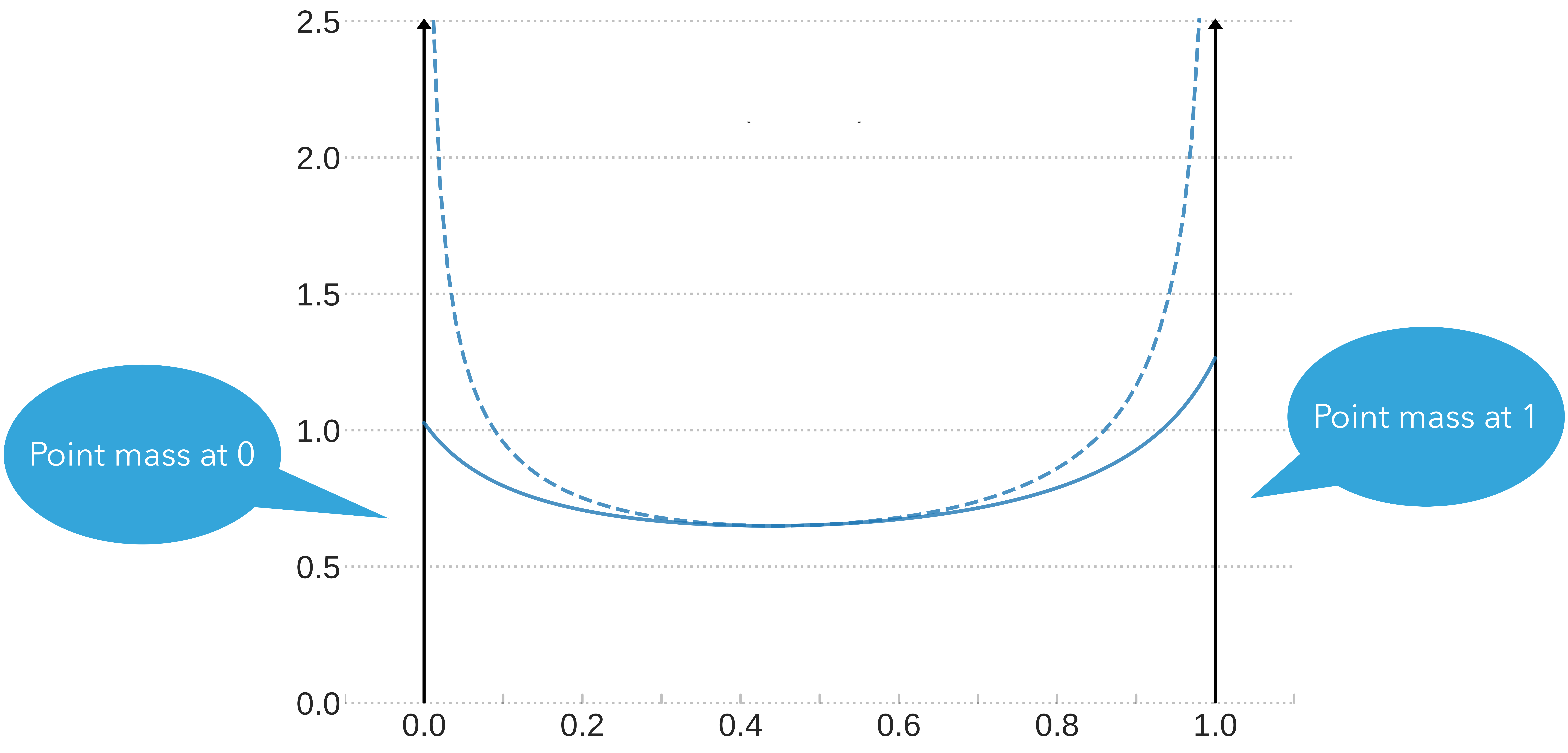
# Stretch-and-rectify (Louizos et al., 2018)




We **collapse** the shaded area into a **point mass**

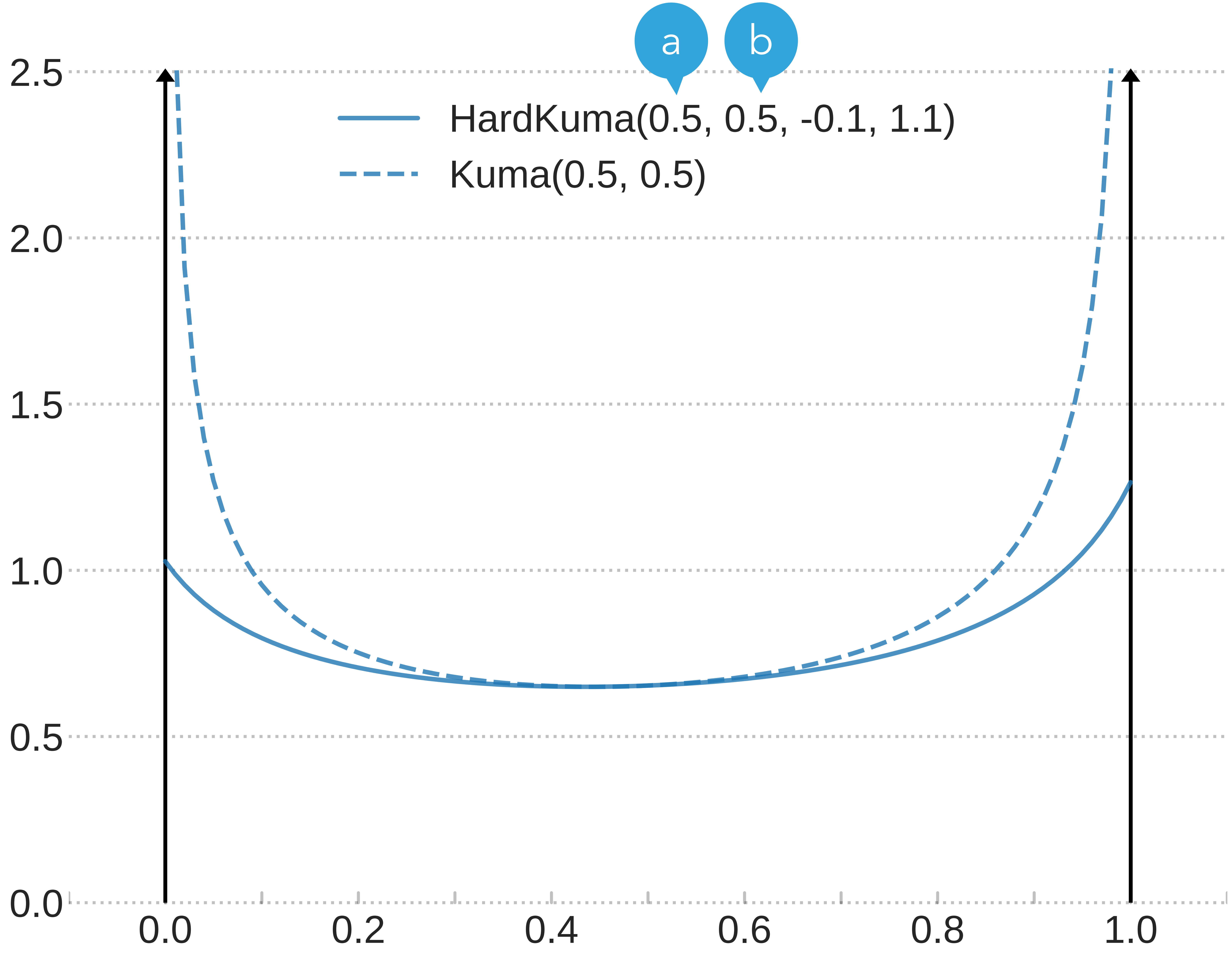
# Stretch-and-rectify (Louizos et al., 2018)





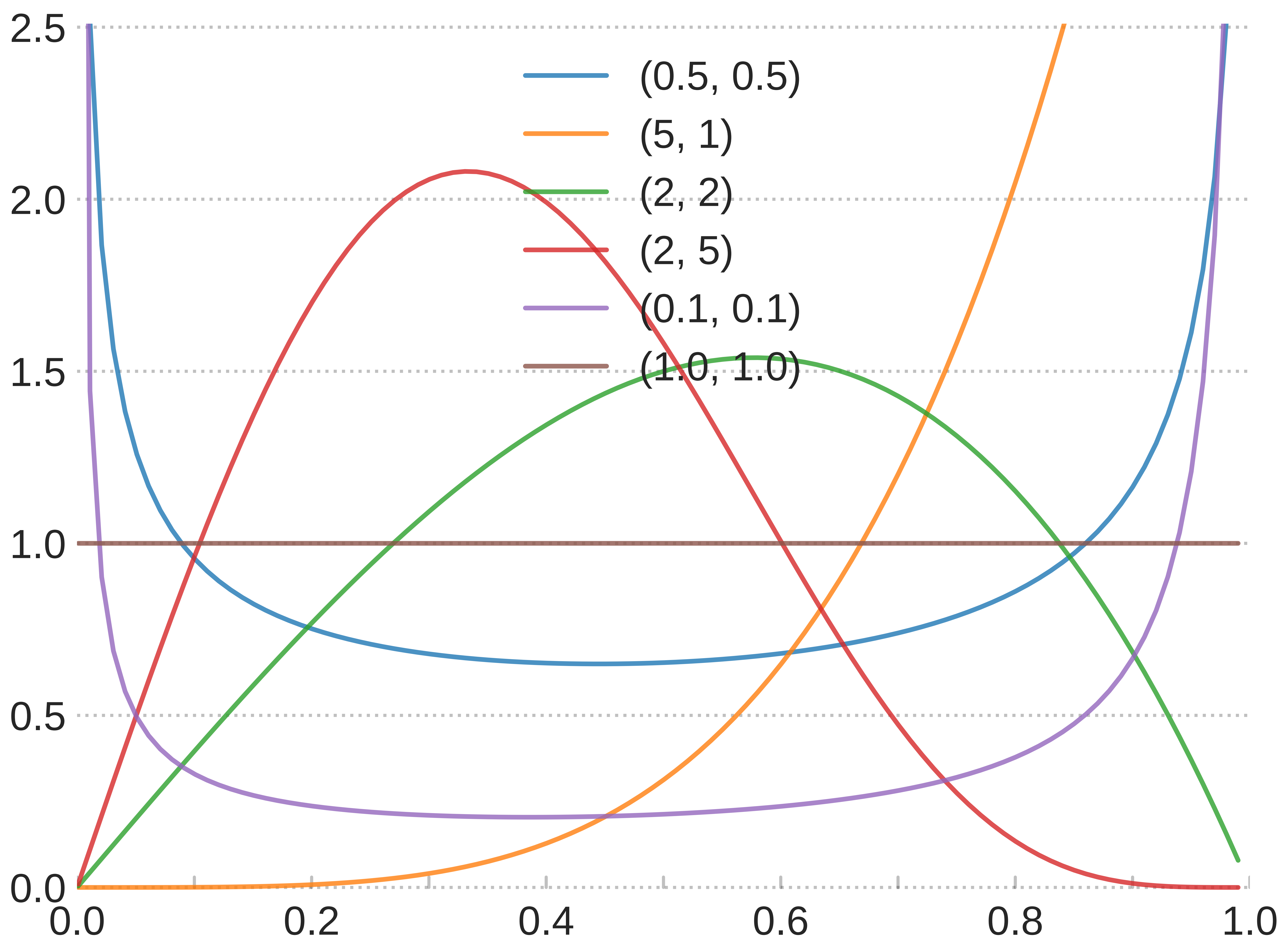
# In this work: Hard Kumaraswamy Distribution





# Kumaraswamy with various (a, b)





# Getting short rationales

- We want **short** rationales without breaking backpropagation
- Solution: relax $L_0$ (Louizos et al., 2018)

$$L_0(z) \qquad \mathbb{E}_{p(z|x)}\left[L_0(z)\right]$$

**Lei et al.**
Compute $L_0$ for one specific assignment of z

**Proposed model:**
$L_0$ computed for all assignments **on expectation**

# Getting coherent rationales



- Baseline: **penalty for transitions** using fused lasso

$$\sum_{i=1}^{n-1} | z_i - z_{i+1} |$$

- Proposed model: compute a **relaxation** of fused lasso by computing the expected number of **zero-to-nonzero** and **nonzero-to-zero** changes:

$$\mathbb{E}_{p(z|x)}\left[\sum_{i=1}^{n-1} \mathbb{I}[z_i = 0, z_{i+1} \neq 0]\right] + \mathbb{E}_{p(z|x)}\left[\sum_{i=1}^{n-1} \mathbb{I}[z_i \neq 0, z_{i+1} = 0]\right]$$

# Specify target selection rate



- We want a maximum selection rate e.g. 10% of the text
- We propose a **constrained optimization** problem:

$$\min_{\phi,\theta} L(\phi, \theta) \quad \text{s.t. } \mathbb{E}[L_0] < r$$

- We use Lagrangian relaxation

# Experiments



1. Multi-aspect sentiment analysis (BeerAdvocate, Lei et al. 2016)
   - Regression, sentiment score in [0,1]
2. Stanford Sentiment (SST)
   - Classification {very negative, …, very positive}
3. Stanford Natural Language Inference (SNLI)
   - Classification {entailment, contradiction, neutral}

# Beer Precision per Aspect



Multiple aspects

| | | Look | | Smell | | Taste | |
|---|---|---|---|---|---|---|---|
| | | Precision | % Selected | Precision | % Selected | Precision | % Selected |
| Attention (Lei et al.) | Threshold | 80.6 | 13 | 88.4 | 7 | 65.3 | 7 |

# Beer Precision per Aspect



Multiple aspects

| | | Look | | Smell | | Taste | |
|---|---|---|---|---|---|---|---|
| | | Precision | % Selected | Precision | % Selected | Precision | % Selected |
| Attention (Lei et al.) | Threshold | 80.6 | 13 | 88.4 | 7 | 65.3 | 7 |
| Bernoulli / REINFORCE (Lei et al.) | Tuned $\lambda$ | 96.3 | 14 | 95.1 | 7 | 80.2 | 7 |

# Beer Precision per Aspect

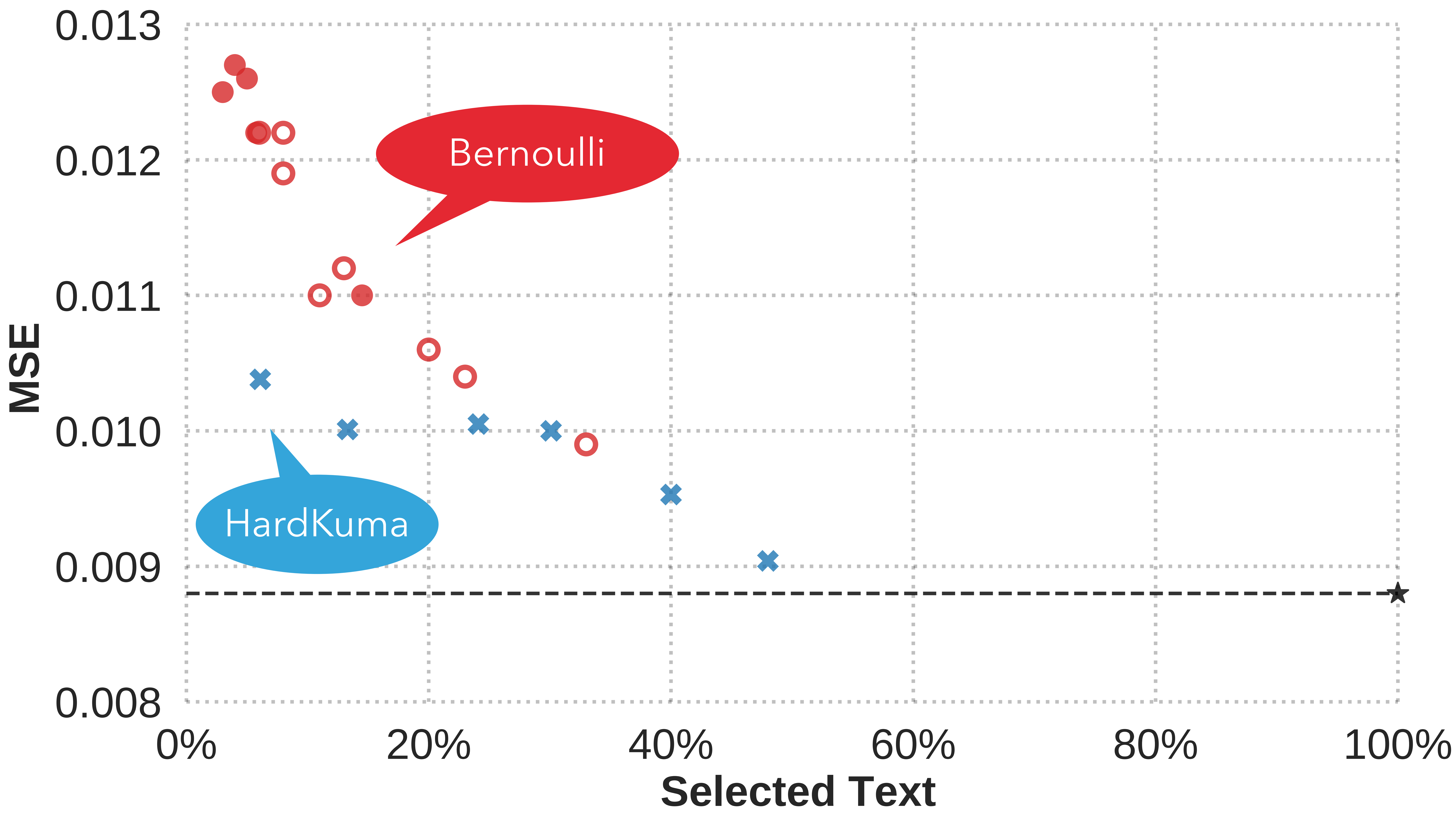


Multiple aspects

| | | Look | | Smell | | Taste | |
|---|---|---|---|---|---|---|---|
| | | Precision | % Selected | Precision | % Selected | Precision | % Selected |
| Attention (Lei et al.) | Threshold | 80.6 | 13 | 88.4 | 7 | 65.3 | 7 |
| Bernoulli / REINFORCE (Lei et al.) | Tuned λ | 96.3 | 14 | 95.1 | 7 | 80.2 | 7 |
| **HardKuma** | Lagrange | **98.1** | **13** | **96.8** | **7** | **89.8** | **7** |

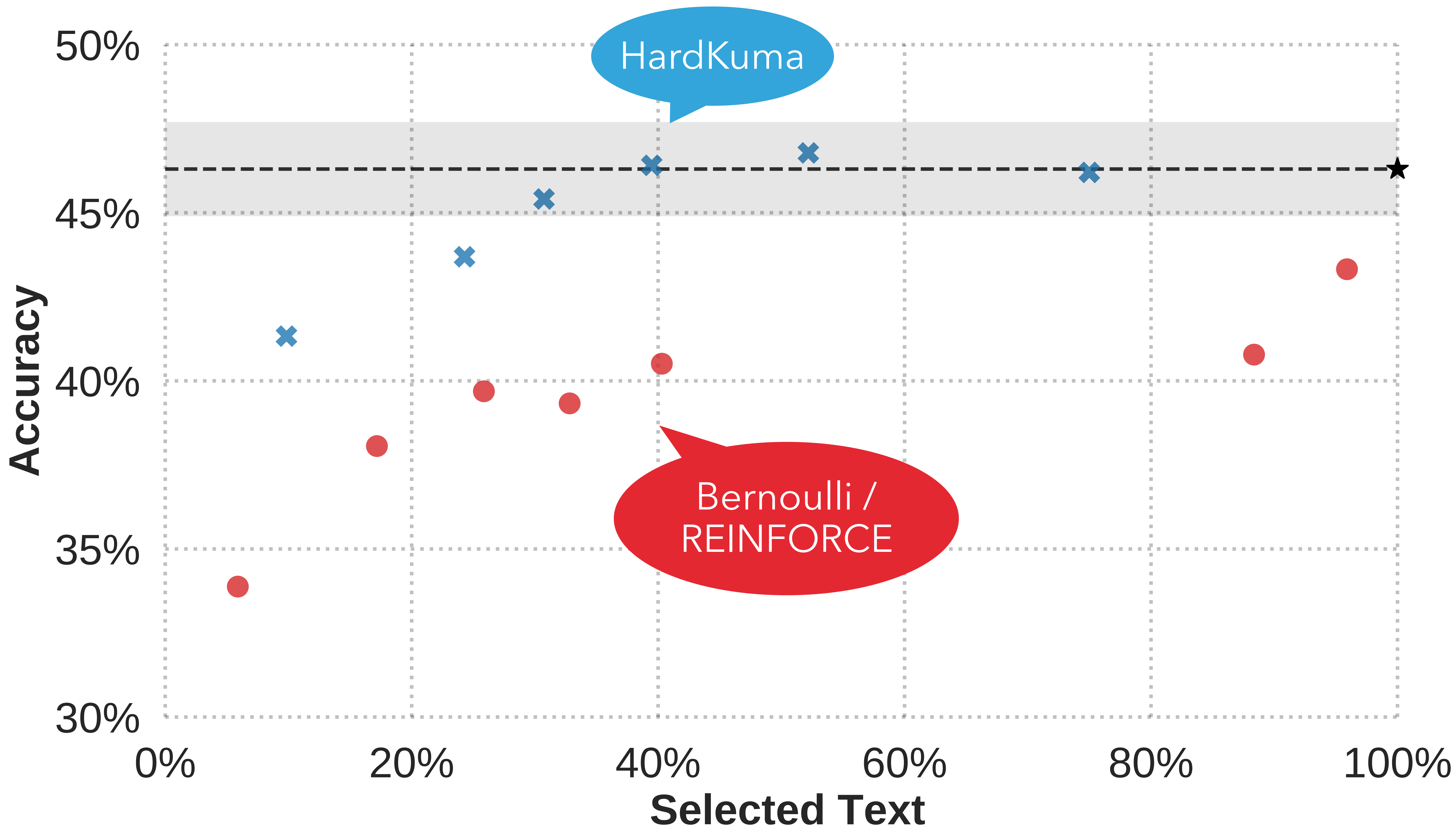
# Beer MSE/selection tradeoff (all aspects)





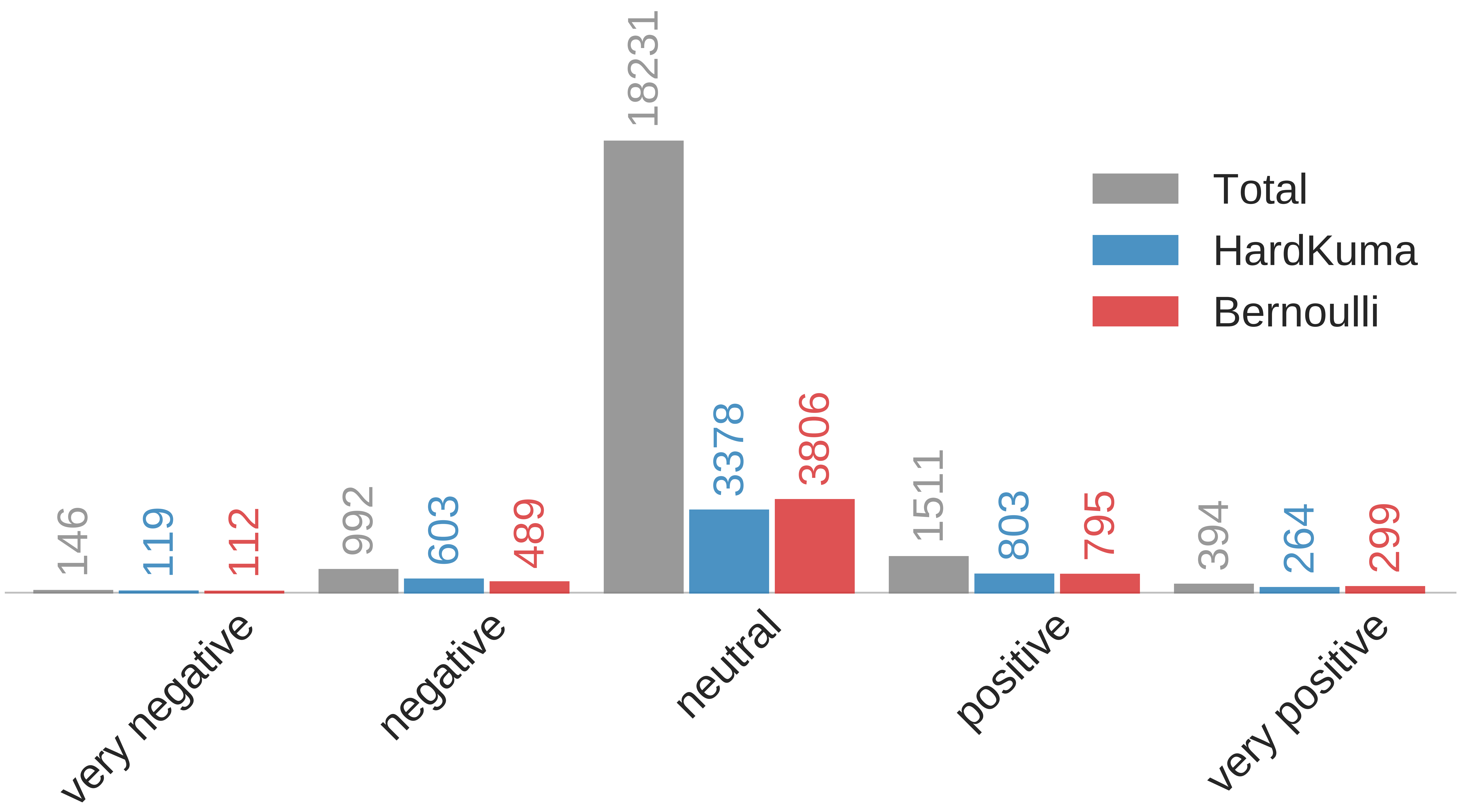
# SST: accuracy/selection tradeoff





# Analysis: Word Count per Sentiment

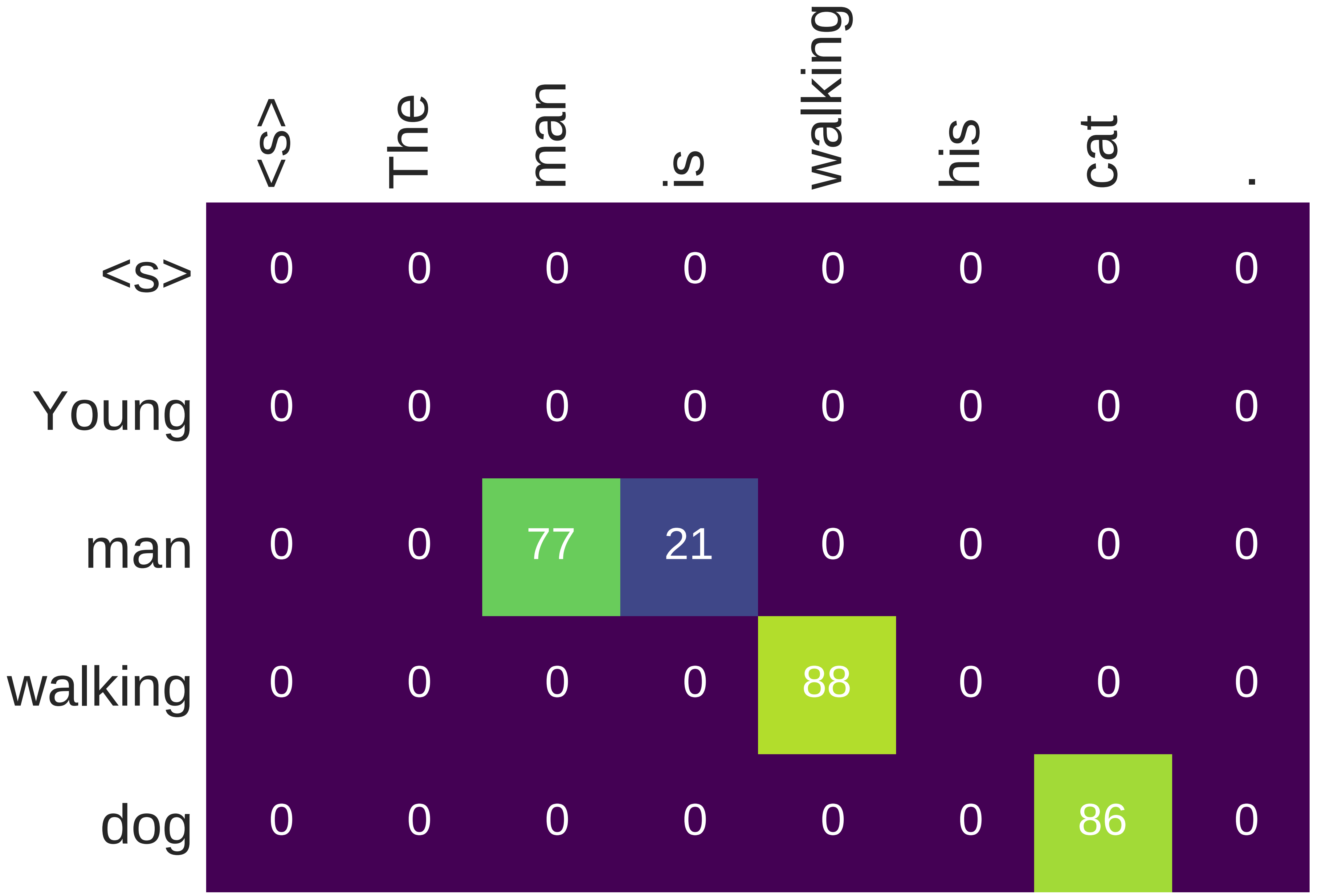




# SNLI: HardKuma attention



- NLI: predict {entailment, contradiction, neutral} given premise & hypothesis
- Baseline: Decomposable Attention model (Parikh et al., 2016)
- We replace Hypothesis-Premise attention with HardKuma attention



# SNLI: HardKuma attention

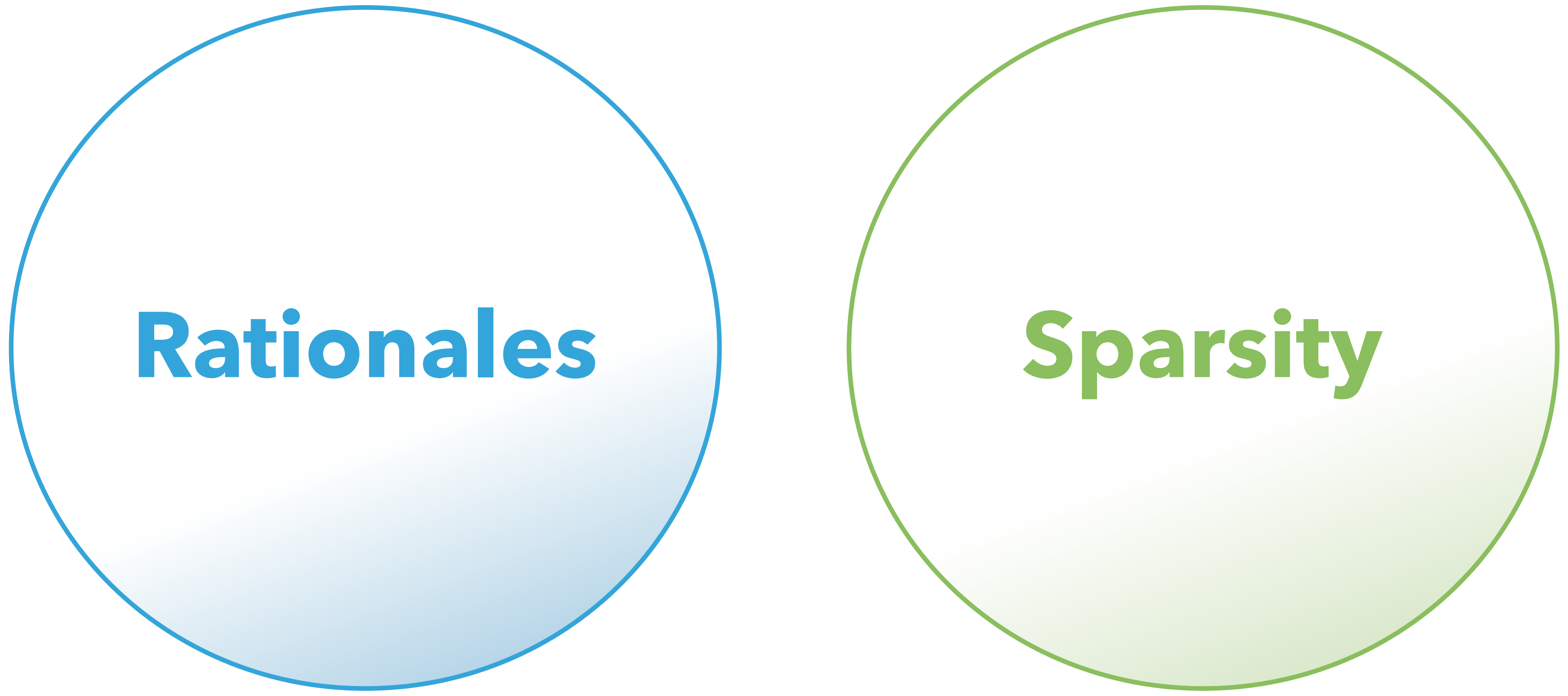


| | SNLI Accuracy | |
|---|---|---|
| | Dev | Test |
| LSTM (Bowman et al.) | - | 80.6 |
| DA (Parikh et al.) | - | 86.3 |
| **DA (reimpl.)** | **86.9** | **86.5** |
| **DA HardKuma** | **86.0** | **85.5** |

Only drop **1%** with **8.6%** non-zero attention cells

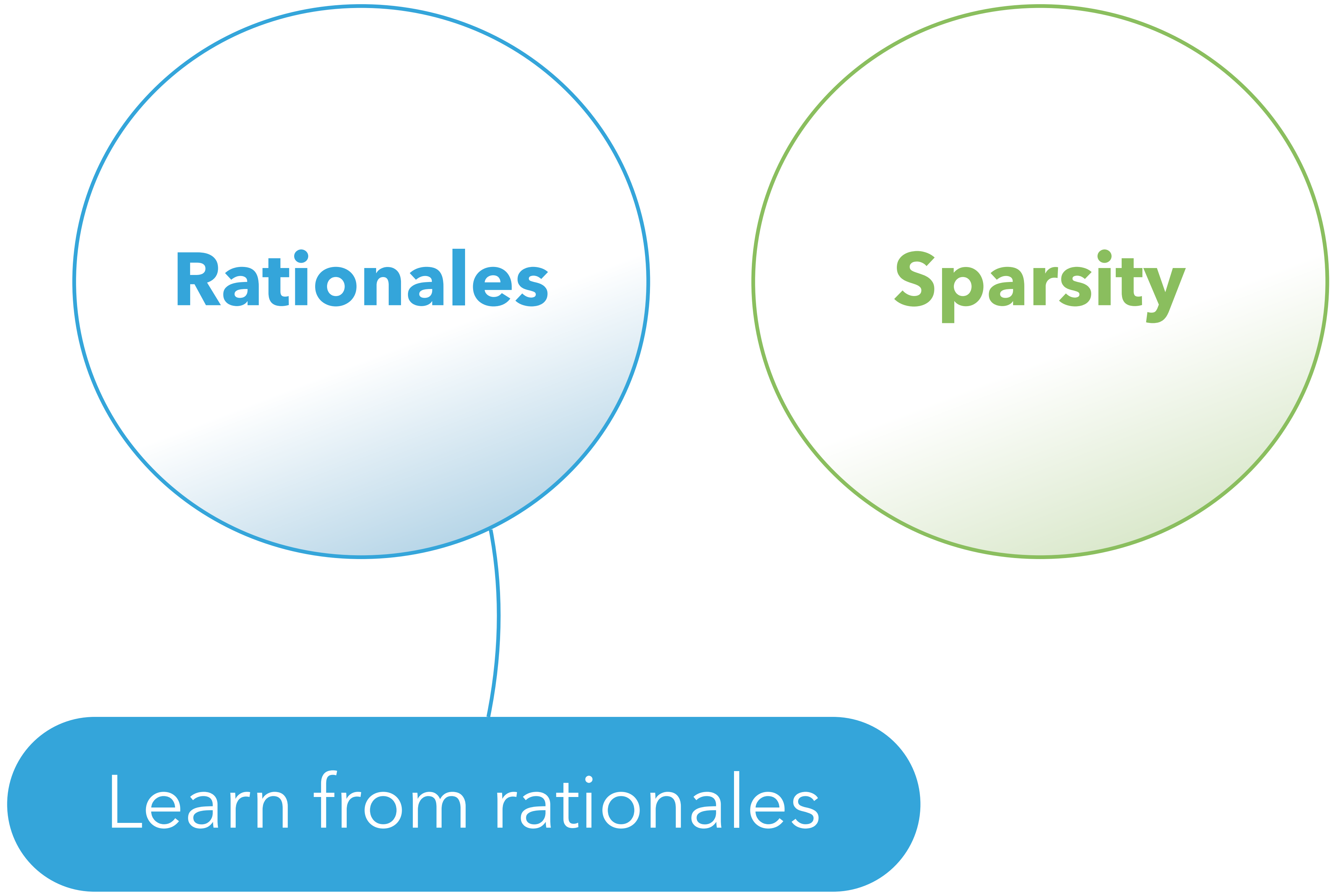
# Related work





# Related work

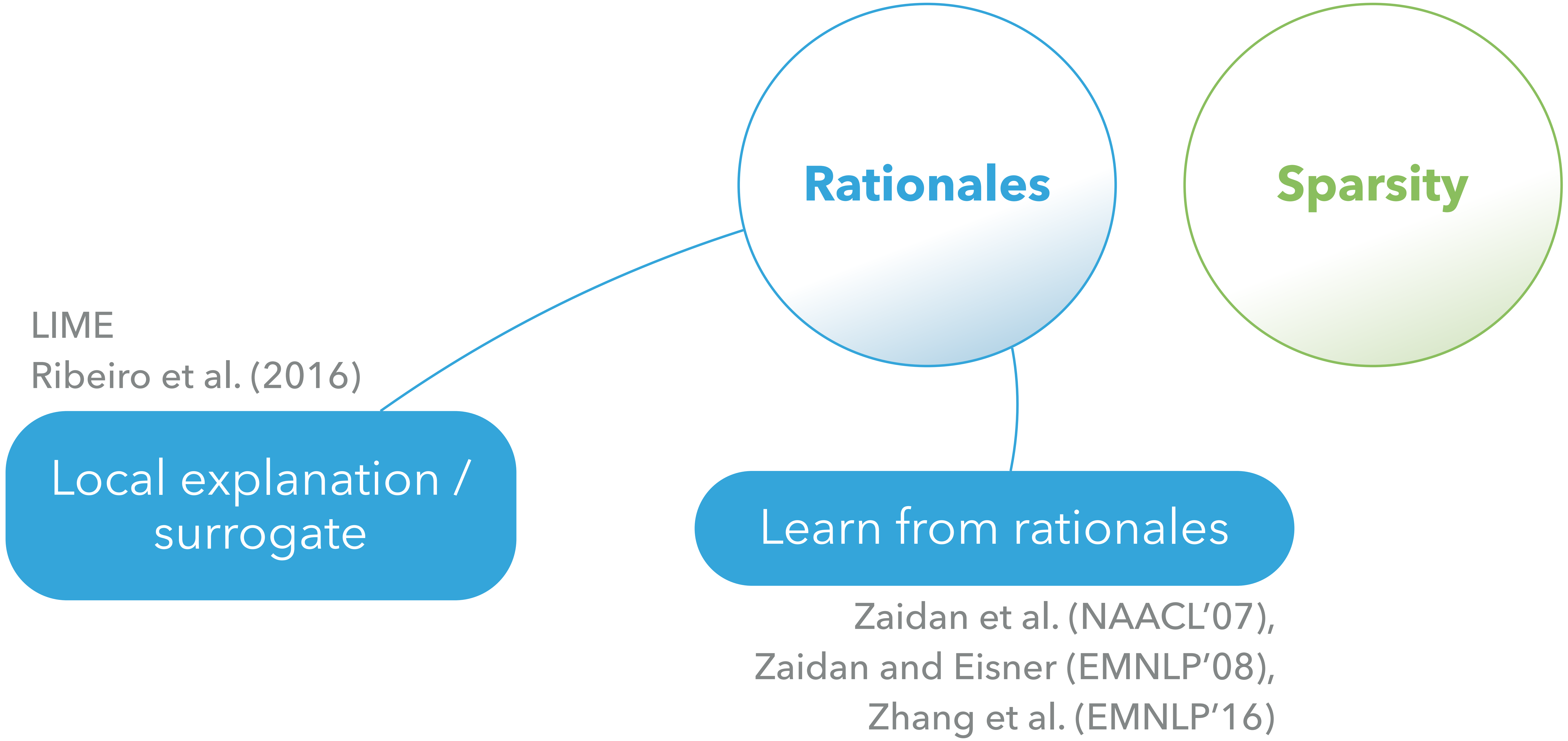





Zaidan et al. (NAACL'07),
Zaidan and Eisner (EMNLP'08),
Zhang et al. (EMNLP'16)

# Related work

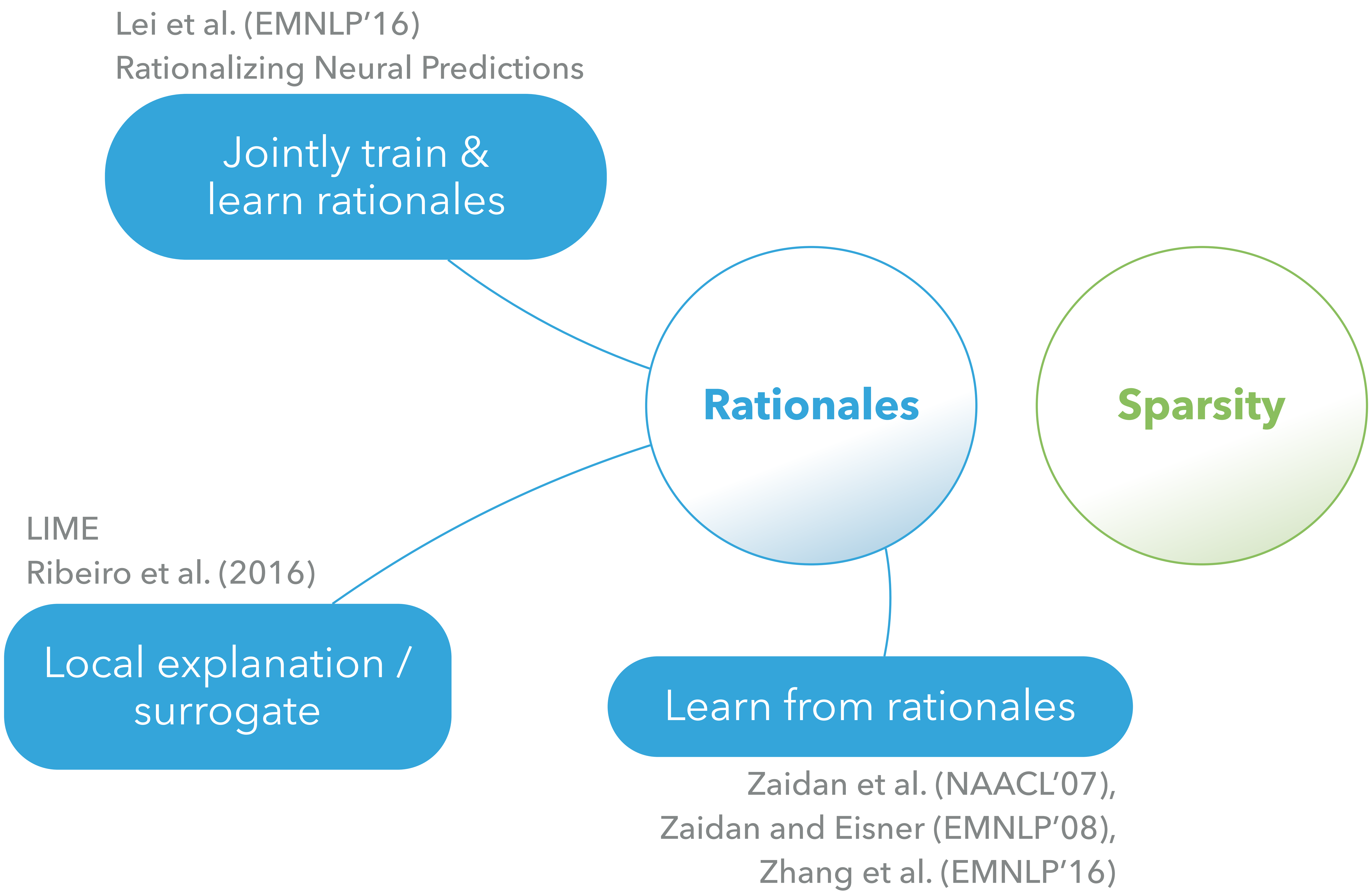




# Related work



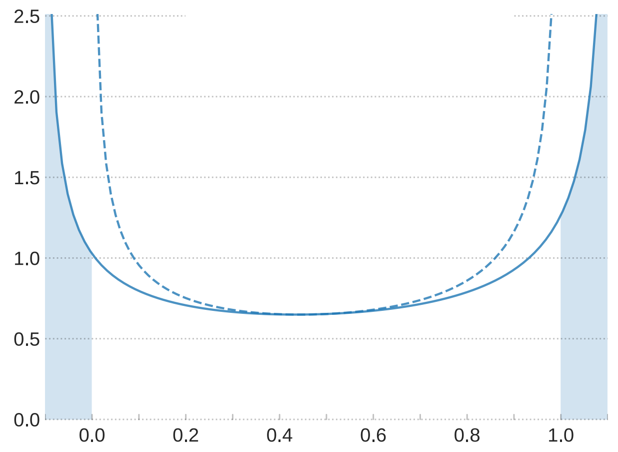
this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here ... first, the aroma is kind of bubblegum-like and grainy. next, the taste is sweet and grainy with an unpleasant bitterness in the finish. ... ... overall, the fat weasel is good for a fairly cheap buzz, but only if you like your beer grainy and bitter .



# Related work



this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here … first, the aroma is kind of bubblegum-like and grainy. next, the taste is sweet and grainy with an unpleasant bitterness in the finish. … … overall, the fat weasel is good for a fairly cheap buzz, but only if you like your beer grainy and bitter .




Lei et al. (EMNLP'16)
Rationalizing Neural Predictions

Jointly train & learn rationales

Louizos et al. (ICLR'18)
Sparsifying parameters

Stretch & rectify

**Rationales**

**Sparsity**

LIME
Ribeiro et al. (2016)

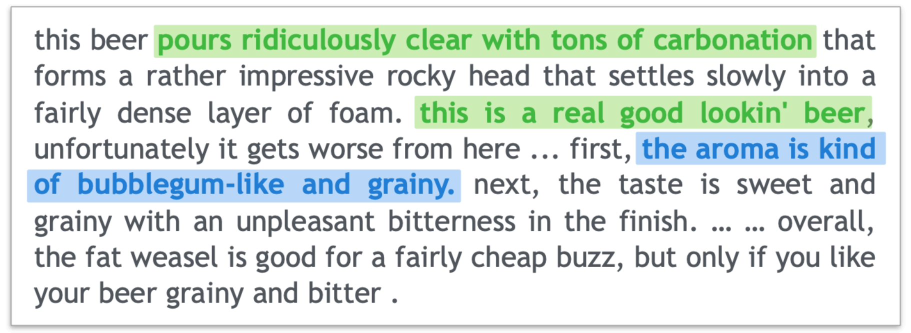
Local explanation / surrogate

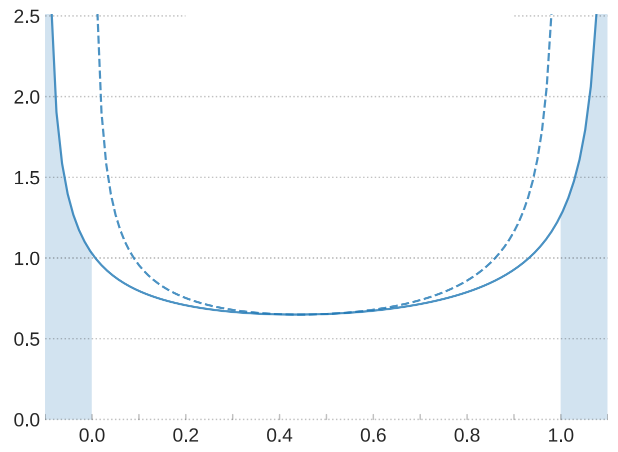
Learn from rationales

Zaidan et al. (NAACL'07),
Zaidan and Eisner (EMNLP'08),
Zhang et al. (EMNLP'16)

# Related work




this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here ... first, the aroma is kind of bubblegum-like and grainy. next, the taste is sweet and grainy with an unpleasant bitterness in the finish. ... ... overall, the fat weasel is good for a fairly cheap buzz, but only if you like your beer grainy and bitter .




Lei et al. (EMNLP'16)
Rationalizing Neural Predictions

Jointly train & learn rationales

This work

Louizos et al. (ICLR'18)
Sparsifying parameters

Stretch & rectify

Rationales

Sparsity

LIME
Ribeiro et al. (2016)

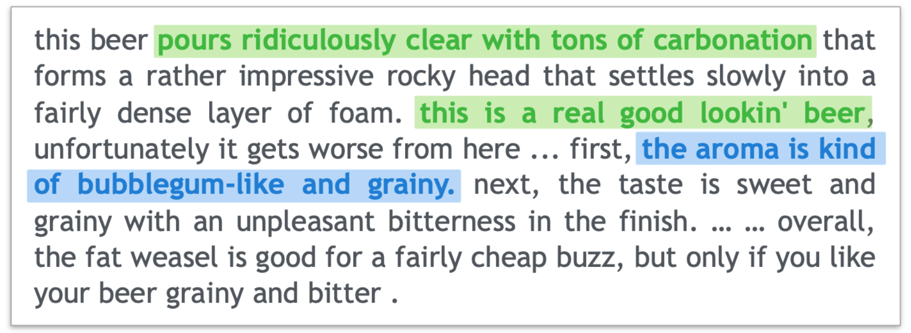
Local explanation / surrogate

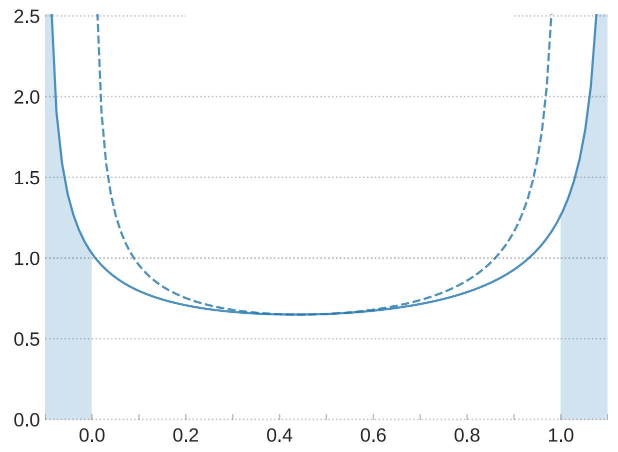
Learn from rationales

Zaidan et al. (NAACL'07),
Zaidan and Eisner (EMNLP'08),
Zhang et al. (EMNLP'16)

# Related work




this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here ... first, the aroma is kind of bubblegum-like and grainy. next, the taste is sweet and grainy with an unpleasant bitterness in the finish. ... ... overall, the fat weasel is good for a fairly cheap buzz, but only if you like your beer grainy and bitter .




**This work**

Lei et al. (EMNLP'16)
Rationalizing Neural Predictions

Jointly train & learn rationales

Louizos et al. (ICLR'18)
Sparsifying parameters

Stretch & rectify

**Rationales**

**Sparsity**

Analyzing Multi-Head Self-Attention

Voita et al. (ACL'19)

Tomorrow 16:40
Machine Translation 4

LIME
Ribeiro et al. (2016)

Local explanation / surrogate

Learn from rationales

Zaidan et al. (NAACL'07),
Zaidan and Eisner (EMNLP'08),
Zhang et al. (EMNLP'16)

# Summary



- Differentiable approach to extractive rationales
  - Stretch and rectify using HardKuma
  - Support for binary outcomes
- Objective to specify the percentage of selected text
- Future work: interpretable QA / fact checking
- Code online: github.com/bastings
  - DIY: add a HardKuma layer to your classifier!

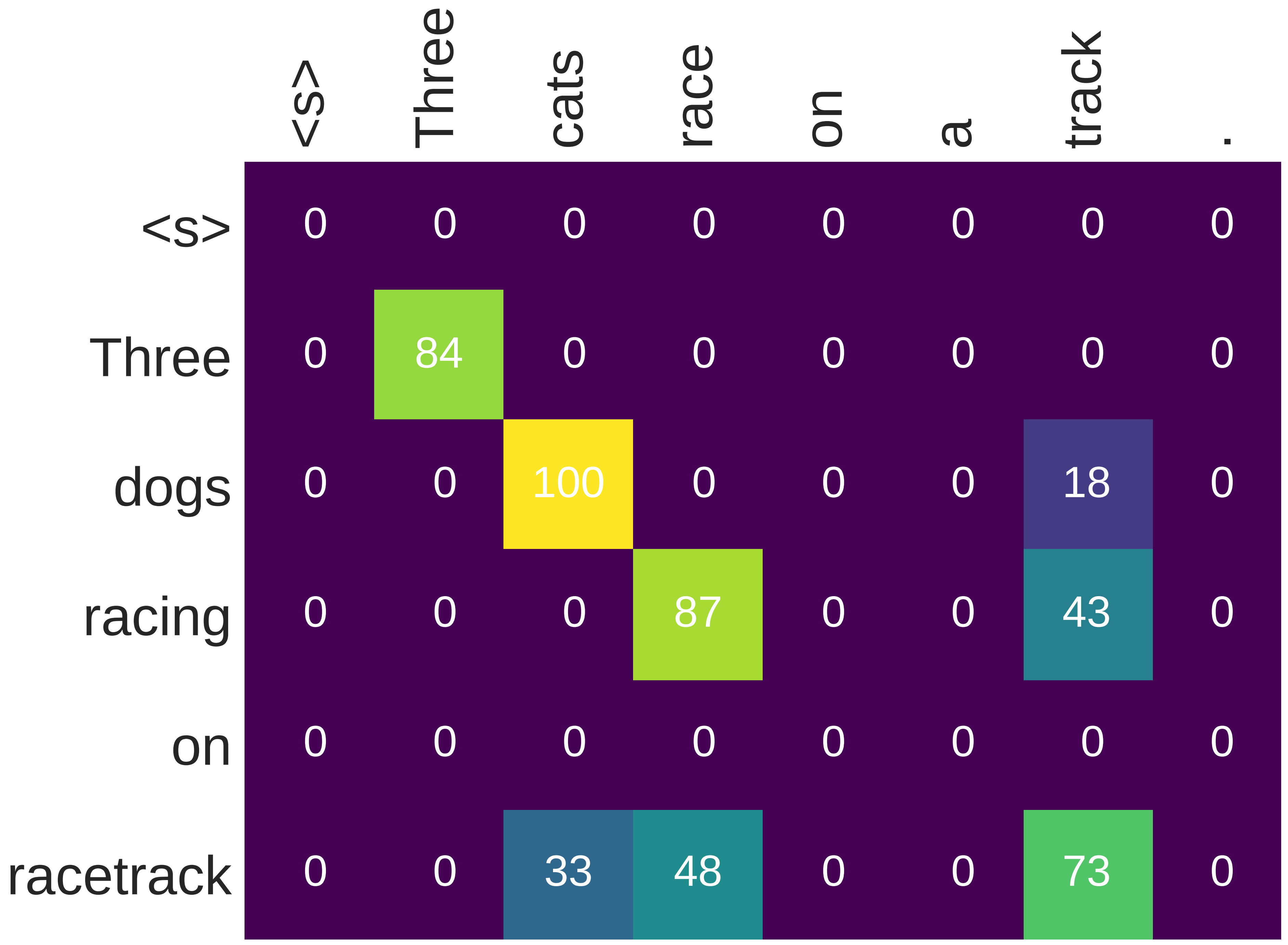
# Thank you!

Code online @ github.com/bastings

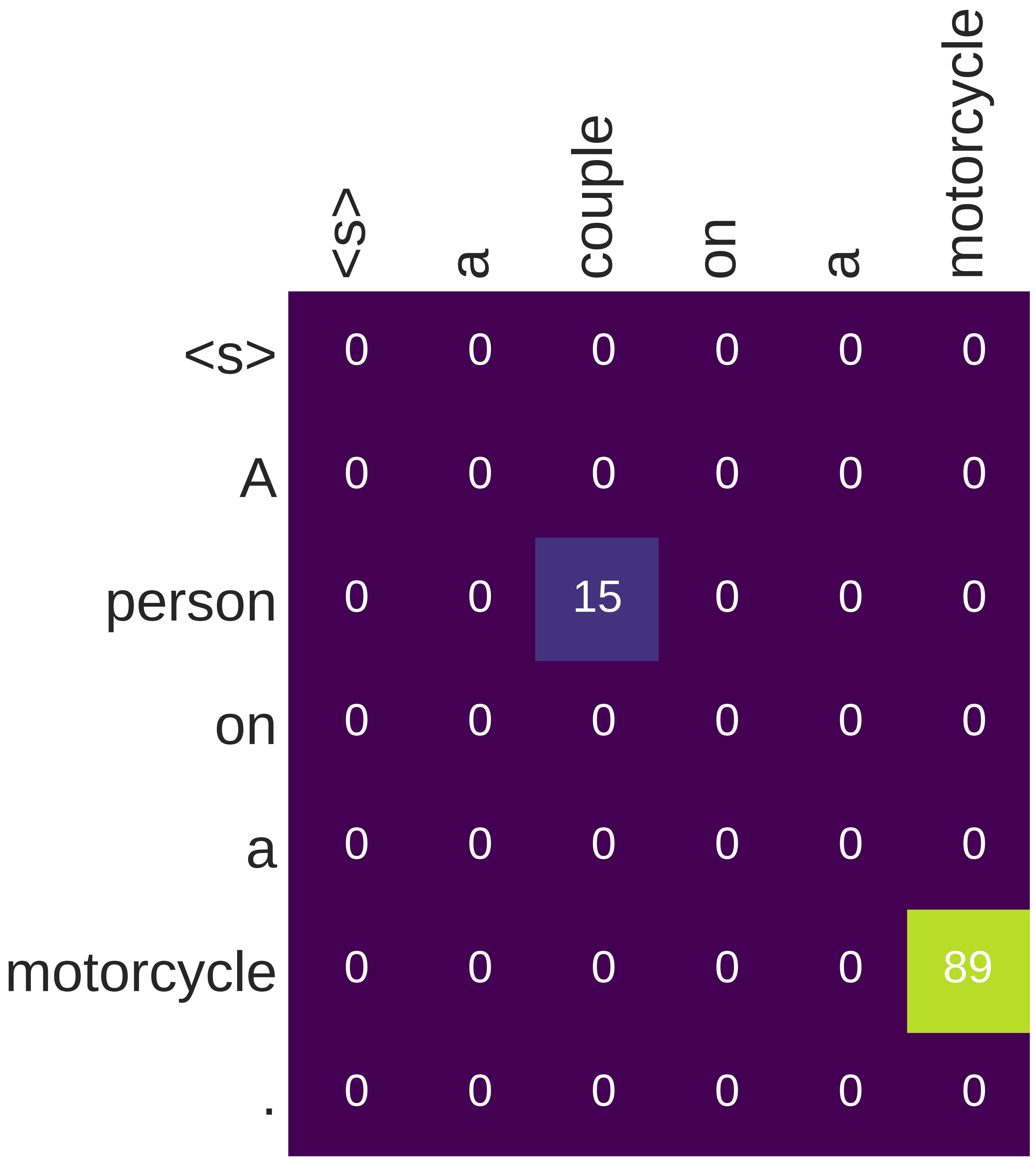
Check out our new NMT toolkit for novices Joey NMT at github.com/joeynmt

# Example: Contradiction






Correct




Incorrect

Prediction: entailment

# Example: Entailment



| | <s> | The | two | dogs | are | black | . |
|---|---|---|---|---|---|---|---|
| <s> | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Two | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| black | 0 | 0 | 0 | 0 | 0 | 100 | 0 |
| dogs | 0 | 0 | 0 | 90 | 0 | 0 | 0 |
| running | 0 | 0 | 0 | 23 | 0 | 0 | 0 |

Correct

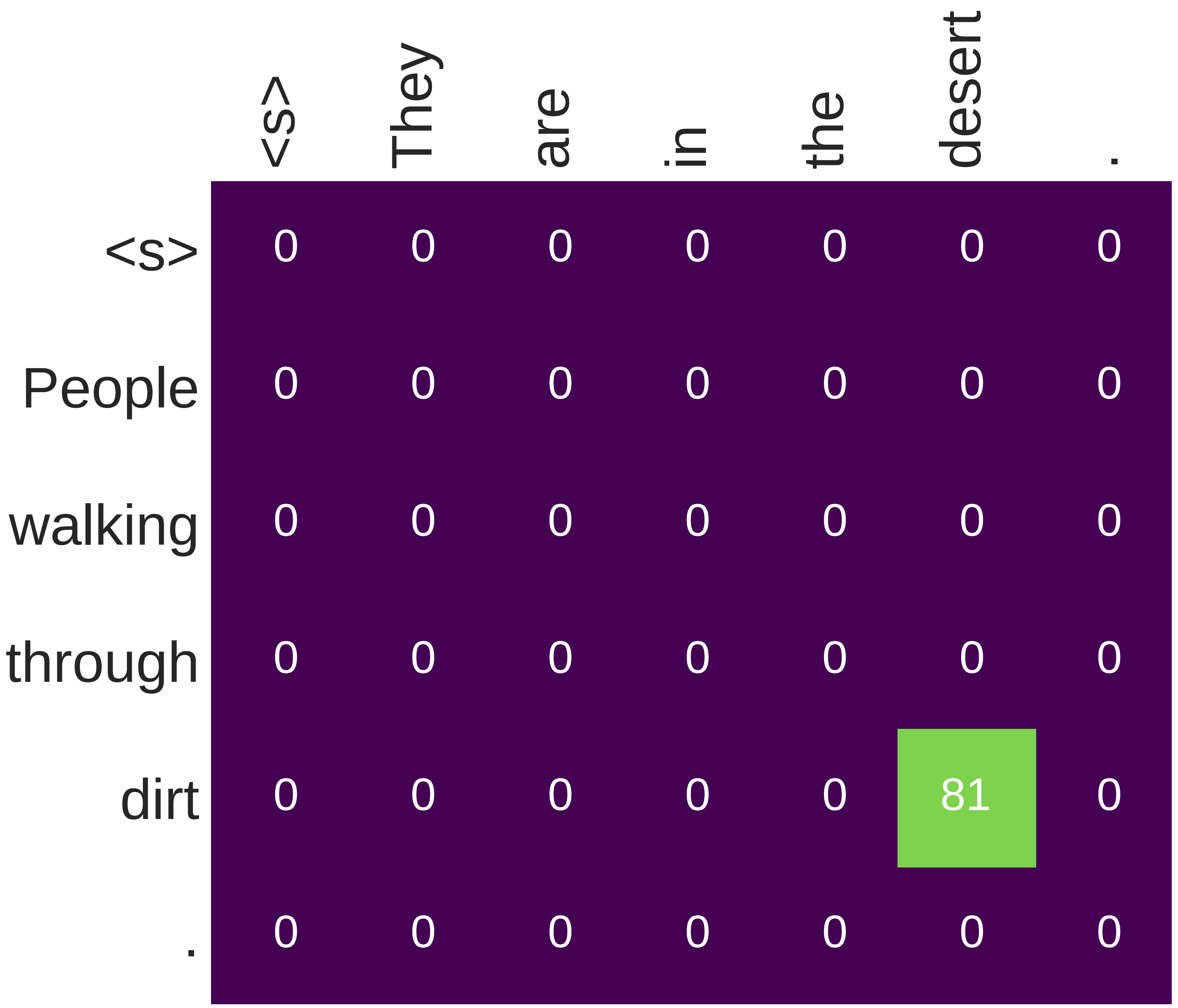
| | <s> | Four | people | in | a | kitchen | cooking | . |
|---|---|---|---|---|---|---|---|---|
| <s> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Four | 0 | 89 | 0 | 0 | 0 | 0 | 0 | 0 |
| people | 0 | 0 | 53 | 0 | 0 | 0 | 0 | 0 |
| in | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| a | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| kitchen | 0 | 0 | 0 | 0 | 0 | 100 | 74 | 0 |

Incorrect

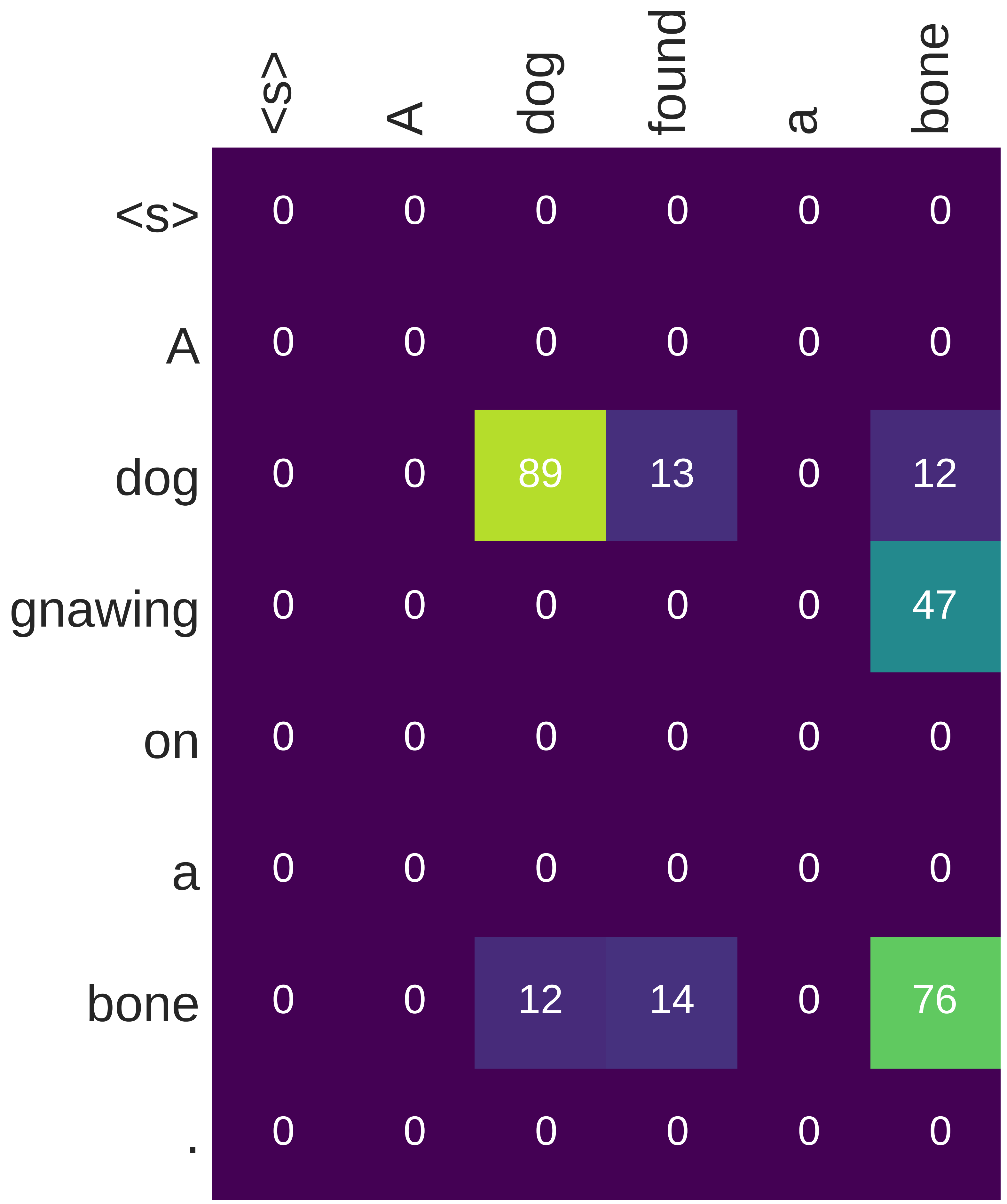
Prediction: neutral

# Example: Neutral






Correct




Incorrect

Prediction: entailment

# Parameter estimation

$$\begin{aligned}
\log P(y \mid x) &= \log \mathbb{E}_{P(z|x,\phi)}\left[P(y \mid x, z, \theta)\right] \\
&\geq \mathbb{E}_{P(z|x,\phi)}\left[\log P(y \mid x, z, \theta)\right] \\
&= \mathscr{E}(\phi, \theta)
\end{aligned}$$

We maximise this lower bound on the log-likelihood

# Baseline Regularizers



Loss for **sufficient** rationales

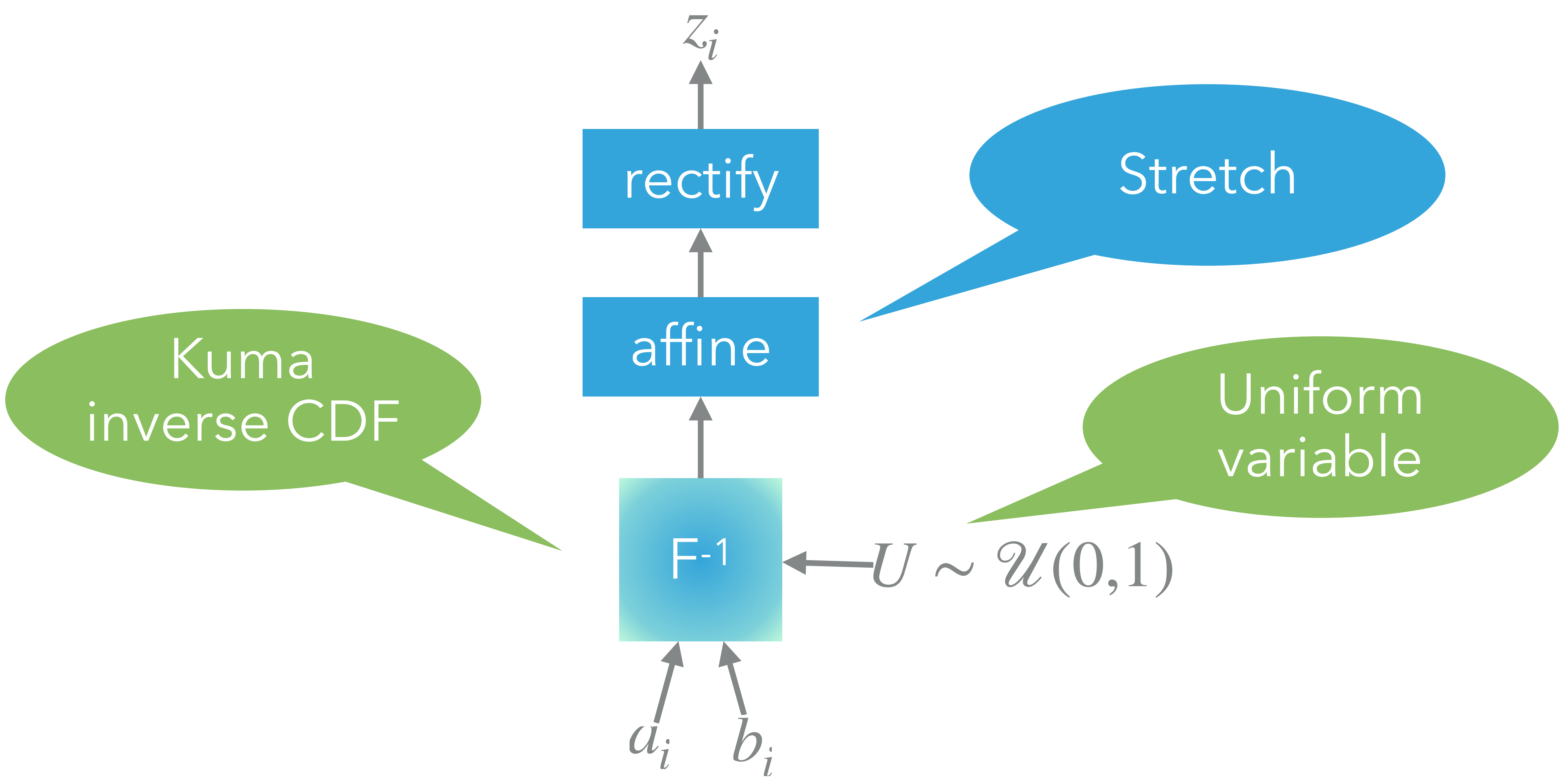
$$\min_{\phi,\theta} L(\phi, \theta) + \lambda_0 \sum_{i=1}^{n} z_i + \lambda_1 \sum_{i=1}^{n-1} | z_i - z_{i+1} |$$

$L_0$ for **short** rationales

Fused lasso for **coherent** rationales

# How to get a HardKuma sample from a uniform variable





# From uniform source to Kuma samples



Kuma inverse CDF

$$F_K^{-1}(u; a, b) = \left(1 - (1-u)^{\frac{1}{b}}\right)^{\frac{1}{a}} \qquad u \in [0,1]$$

Sample using uniform random source U

$$F_Z^{-1}(U; a, b) \sim \text{Kuma}(a, b) \qquad U \sim \mathcal{U}(0,1)$$

# Rectified Kumaraswamy (Formal)



- We **stretch** the support of the Kuma to *(l, r)*:

$$F_T(t; a, b, l, r) = F_K\left(\frac{(t-l)}{(r-l)}; a, b\right)$$

- And define a **rectified** random variable:

$$H \sim \text{HardKuma}(a, b, l, r)$$

- by passing a Kuma sample *t* through a hard sigmoid:

Support
in **closed** interval
[0, 1]

$$T \sim \text{Kuma}(a, b, l, r) \qquad h = \min(1, \max(0, t))$$

# Rectified Kumaraswamy (Formal) (2)



- Sampling $h$=0 means sampling any $t \in (l,0]$
- with mass under Kuma:

$$\mathbb{P}(H = 0) = F_K\left(\frac{-l}{r-l}; a, b\right)$$

- Sampling $h$=1 means sampling any $t \in [1,r)$
- with mass under Kuma:

$$\mathbb{P}(H = 1) = 1 - F_K\left(\frac{1-l}{r-l}; a, b\right)$$

# Detailed HardKuma relaxed $L_0$



$$\mathbb{E}_{p(z|x)}\left[L_0(z)\right] \overset{\text{ind}}{=} \sum_{i=1}^{n} \mathbb{E}_{p(z_i|x)}\left[\mathbb{I}[z_i \neq 0]\right]$$

$$= \sum_{i=1}^{n} 1 - \mathbb{P}(Z_i = 0) \ ,$$

# Detailed HardKuma fused lasso



- We can also compute a relaxation of fused lasso by computing the expected number of **zero-to-nonzero** and **nonzero-to-zero** changes:

$$\mathbb{E}_{p(z|x)}\left[\sum_{i=1}^{n-1} \mathbb{I}[z_i = 0, z_{i+1} \neq 0]\right] + \mathbb{E}_{p(z|x)}\left[\sum_{i=1}^{n-1} \mathbb{I}[z_i \neq 0, z_{i+1} = 0]\right]$$

$$\stackrel{\text{ind}}{=} \sum_{i=1}^{n-1} \mathbb{P}(Z_i = 0)(1 - \mathbb{P}(Z_{i+1} = 0)) + (1 - \mathbb{P}(Z_i = 0))\mathbb{P}(Z_{i+1} = 0)$$

# Reparameterization Trick



$$U \sim \mathcal{U}(0,1)$$

$$F_X^{-1}(u) \sim X$$

# Why are gradients possible?



- We consider the case where we need derivatives of a function L(u) of the underlying uniform variable *u*, as when we compute reparameterized gradients in variational inference.  By chain rule:

$$\frac{\partial \mathscr{L}}{\partial u} = \frac{\partial \mathscr{L}}{\partial h} \times \frac{\partial h}{\partial t} \times \frac{\partial t}{\partial k} \times \frac{\partial k}{\partial u}$$

Depends on differentiable observation model

Derivative for hardsigmoid
0 for `t<0`,
1 for `0<t<1`,
0 for `t>1`,
undef for `t={0,1}`

r-l

Depends on Kuma inverse CDF, no challenge