30 Questions in Neural Machine Translation

Jia Xu

Graduate Center & Hunter College, CUNY

MT Tutorial @ JSALT'19

human language translation: old problem

ἌΑνδρα μοι ἔννεπε, Μοῦσα, πολύτροπον…

Odyssey, Homer, 700 BC

human language translation: old problem

ἌΑνδρα μοι ἔννεπε, Μοῦσα, πολύτροπον…

Odyssey, Homer, 700 BC

ancient Greek

human language translation: old problem

ἌΑνδρα μοι ἔννεπε, Μοῦσα, πολύτροπον…

Odyssey, Homer, 700 BC

ancient Greek

human language translation: old problem

ἄνδρα μοι ἔννεπε, Μοῦσα, πολύτροπον…

Odyssey, Homer, 700 BC

ancient Greek

Virum mihi , Camena, insece versutum…

Latin

human language translation: old problem

ἄνδρα μοι ἔννεπε, Μοῦσα, πολύτροπον…

Odyssey, Homer, 700 BC

ancient Greek

Virum mihi , Camena, insece versutum…

Latin

2000 year old translation of
Sing in me, Muse, and through me tell the story of the man of twists and turns…

human language translation: old problem

ἄνδρα μοι ἔννεπε, Μοῦσα, πολύτροπον…

Odyssey, Homer, 700 BC

ancient Greek

Virum mihi , Camena, insece versutum…

Latin

2000 year old translation of
Sing in me, Muse, and through me tell the story of the man of twists and turns…

translating between human languages is an old problem

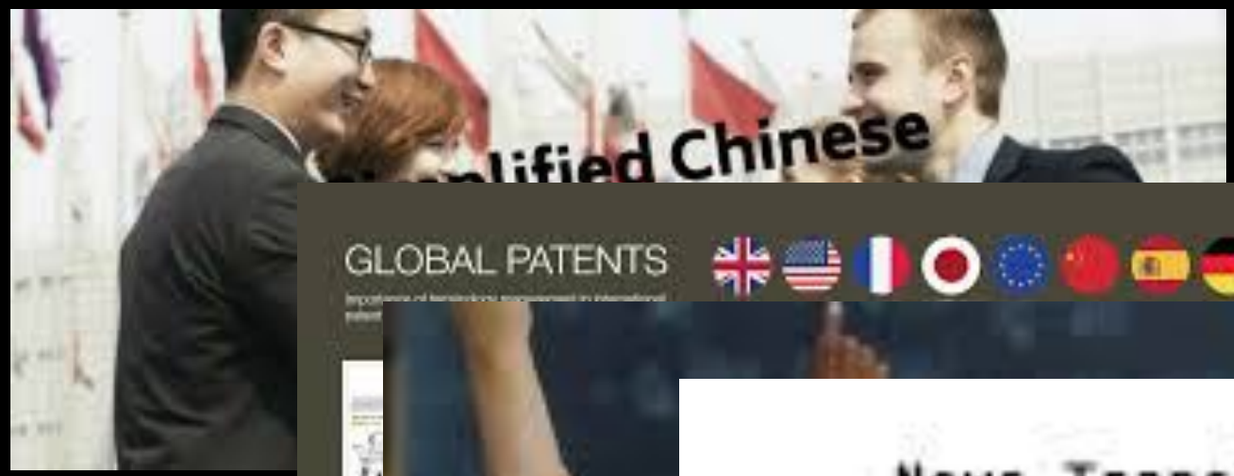
human language translation: nowadays

human language translation: nowadays

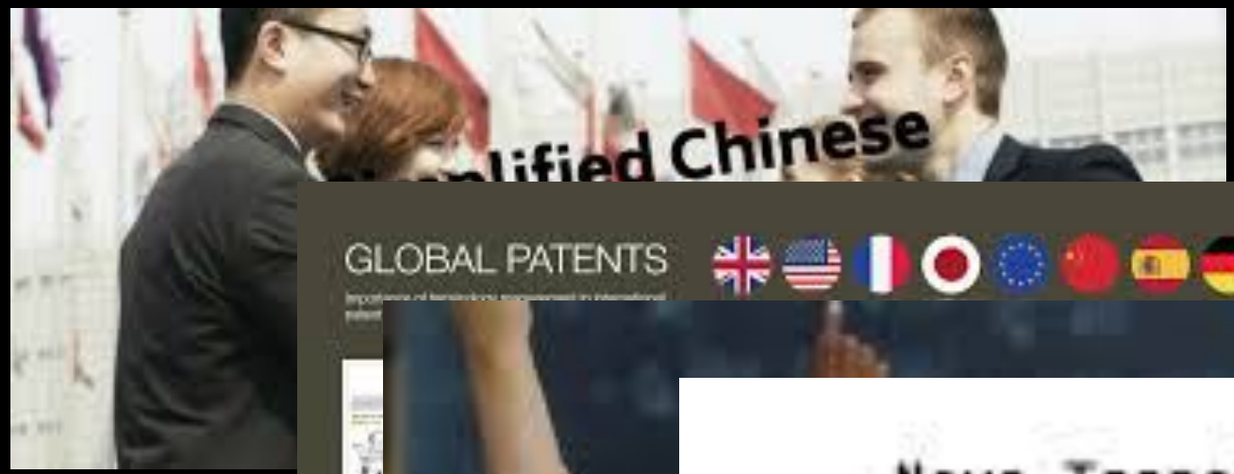
human language translation: nowadays

human language translation: nowadays

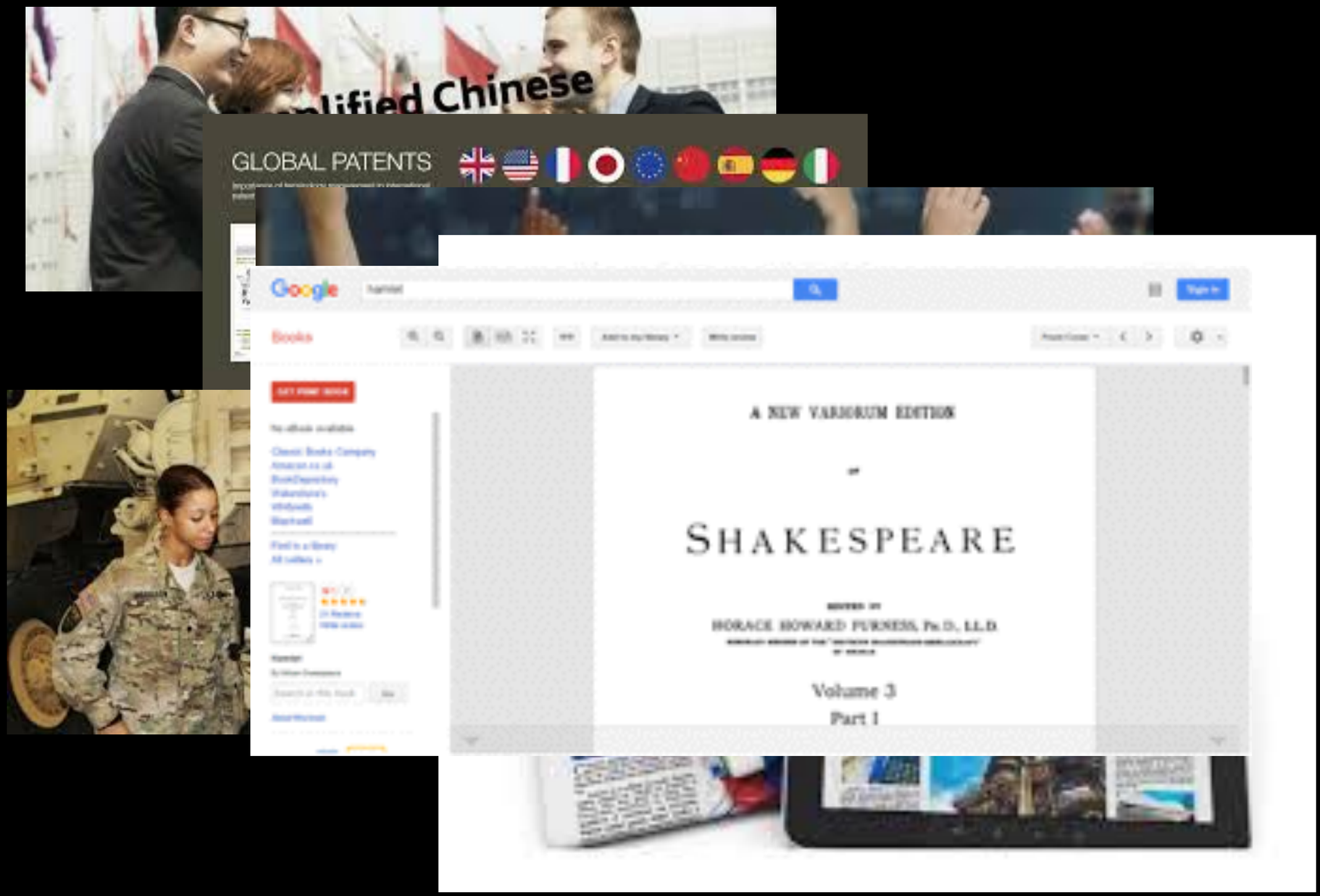
human language translation: nowadays



human language translation: nowadays



human language translation: nowadays



human language translation: nowadays

human language translation: nowadays

human language translation: nowadays

nowadays machine translation is applied in society, science, arts, commerce and finance, literature, military, …

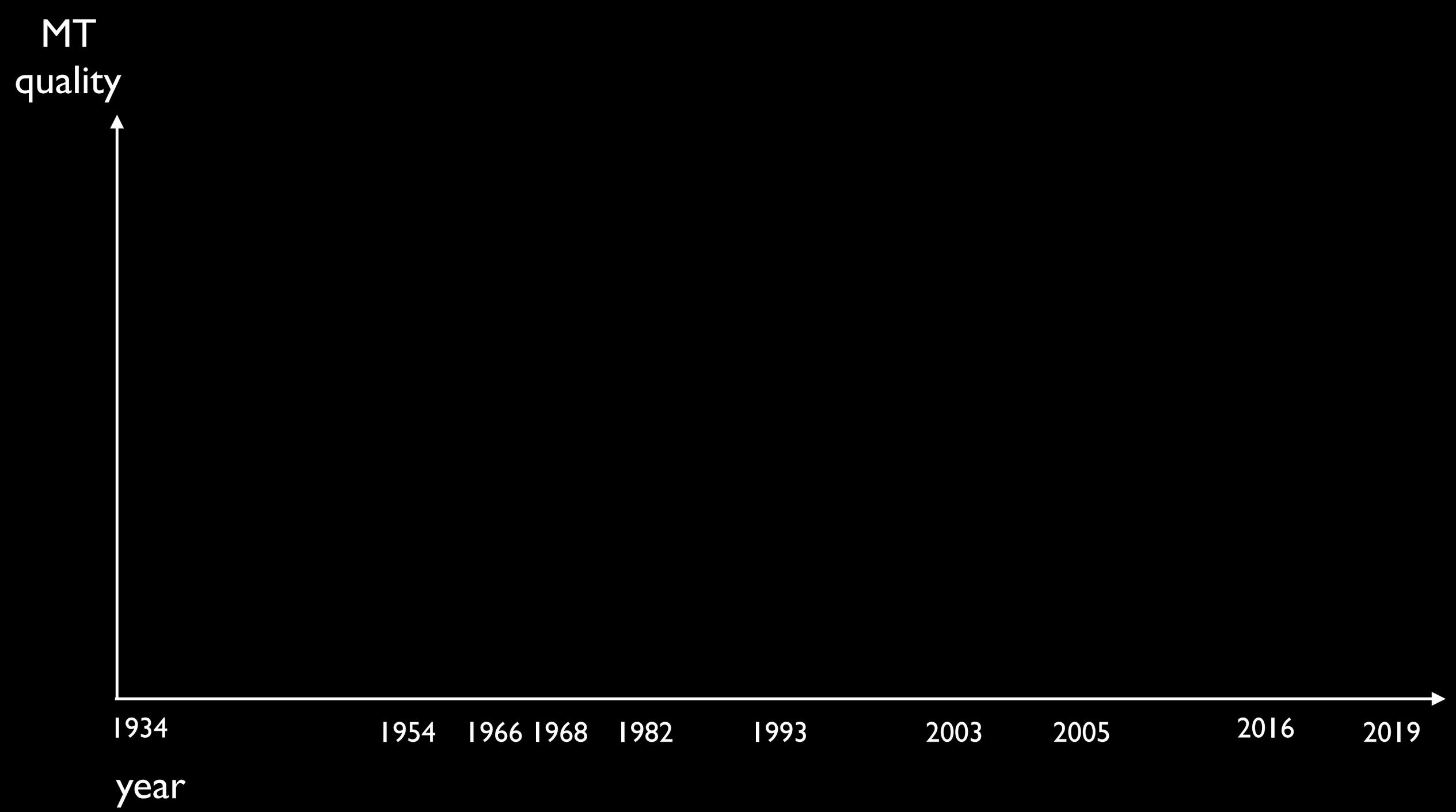
machine translation:

automatically translate from
one human language to another

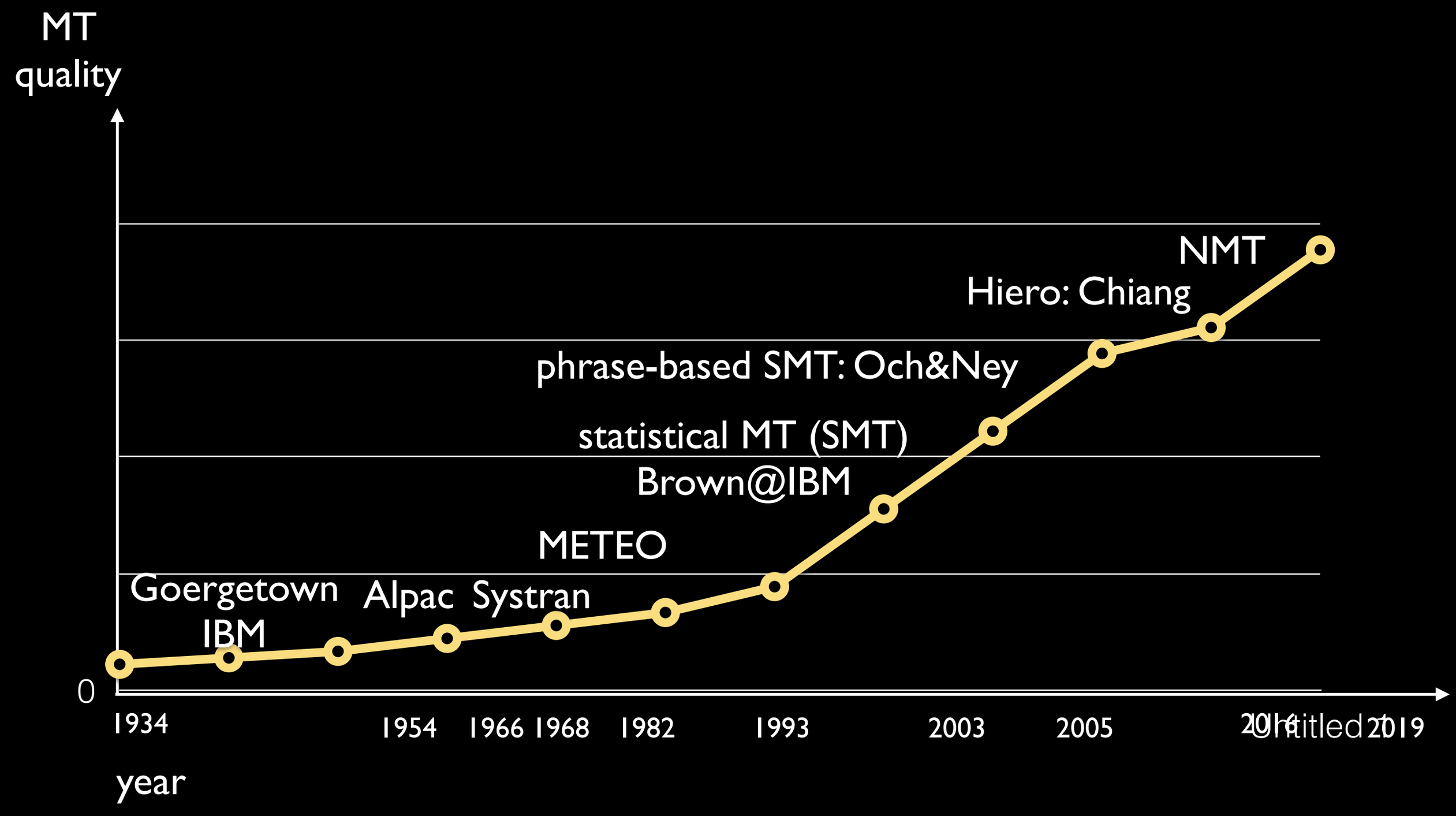
machine translation evaluations & resources

- competitions with data resources and baseline platforms, e.g.
  - NIST: 2002- [https://www.nist.gov/itl/iad/mig/openmt15-evaluation]
  - WMT: 2006- [http://www.statmt.org]
  - IWSLT: 2004- [https://workshop2019.iwslt.org]
- projects, e.g.
  - GALE, TC-Star, EuroMatrix, BOLT, and more and more
- datasets, e.g.
  - LDC: [https://www.ldc.upenn.edu]

machine translation development

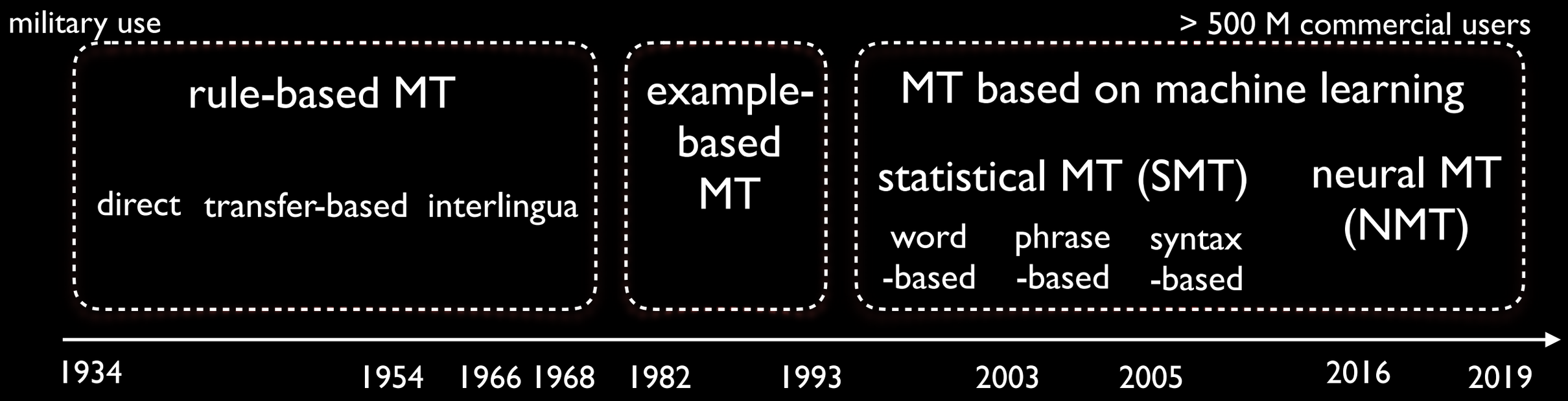


machine translation development



machine translation history & challenges

machine translation history & challenges



machine translation history & challenges

Google translates over 100 billion words a day

military use

> 500 M commercial users

rule-based MT

direct transfer-based interlingua

example-based MT

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)

1934 1954 1966 1968 1982 1993 2003 2005 2016 2019

machine translation history & challenges

Google translates over 100 billion words a day

military use

rule-based MT

direct transfer-based interlingua

example-based MT

> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)

problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

machine translation history & challenges

Google translates over 100 billion words a day

military use

> 500 M commercial users

rule-based MT

direct transfer-based interlingua

example-based MT

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)

problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

machine translation history & challenges

Google translates over 100 billion words a day

military use

rule-based MT

direct transfer-based interlingua

example-based MT

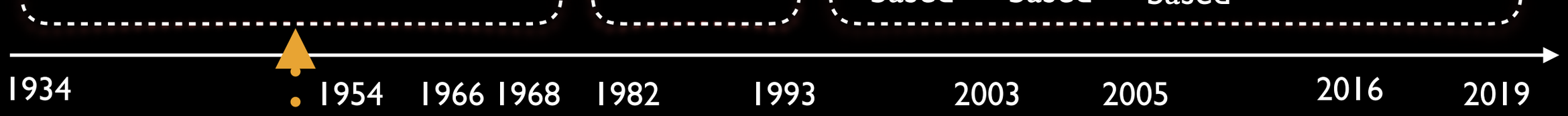
> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)



problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

machine translation history & challenges

Google translates over 100 billion words a day

military use

rule-based MT

direct transfer-based interlingua

example-based MT

> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)

1934 1954 1966 1968 1982 1993 2003 2005 2016 2019

problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

machine translation history & challenges

Google translates over 100 billion words a day

military use

rule-based MT

direct transfer-based interlingua

example-based MT

> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)

problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

machine translation history & challenges

Google translates over 100 billion words a day

military use

rule-based MT

direct transfer-based interlingua

example-based MT

> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)

problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

machine translation history & challenges

Google translates over 100 billion words a day

military use

rule-based MT

direct transfer-based interlingua

example-based MT

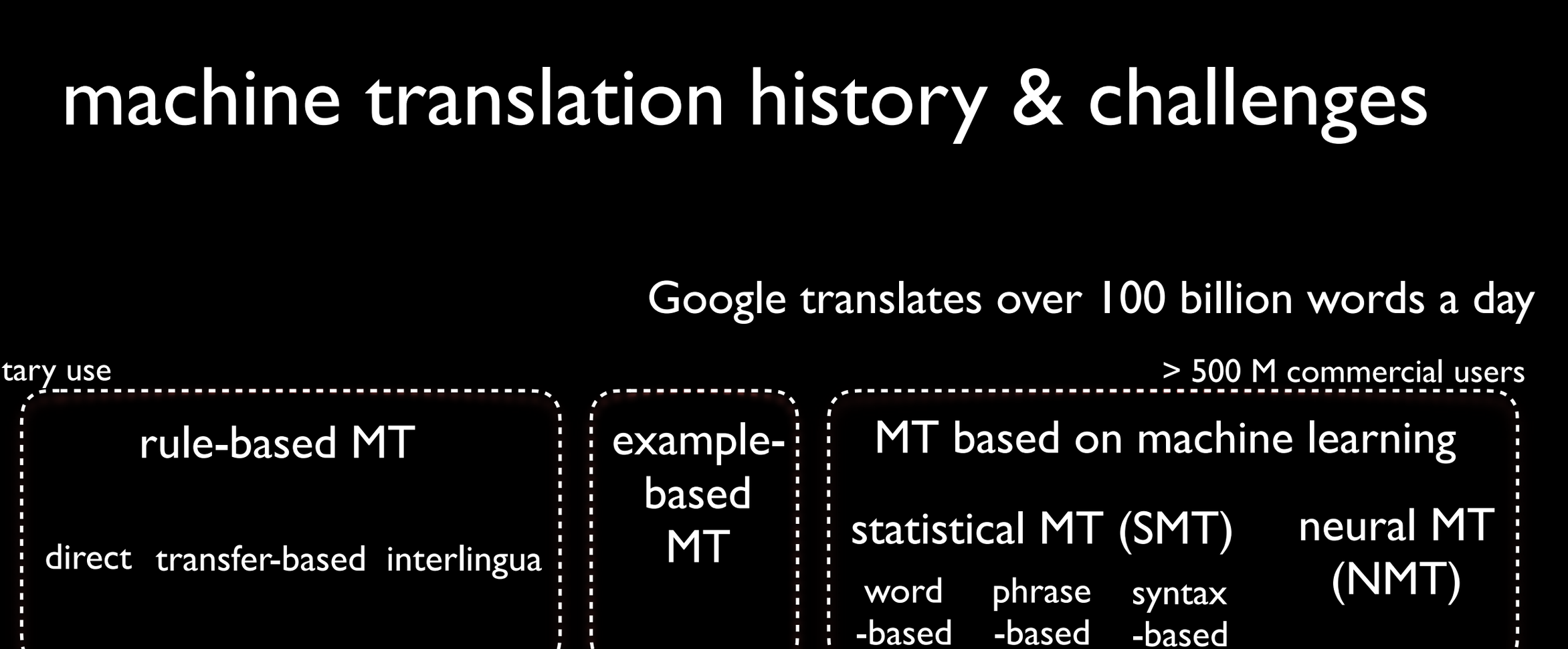
> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)



problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

machine translation history & challenges

Google translates over 100 billion words a day

military use

> 500 M commercial users

rule-based MT

direct transfer-based interlingua

example-based MT

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)

problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

machine translation history & challenges

Google translates over 100 billion words a day

military use

> 500 M commercial users

rule-based MT

direct transfer-based interlingua

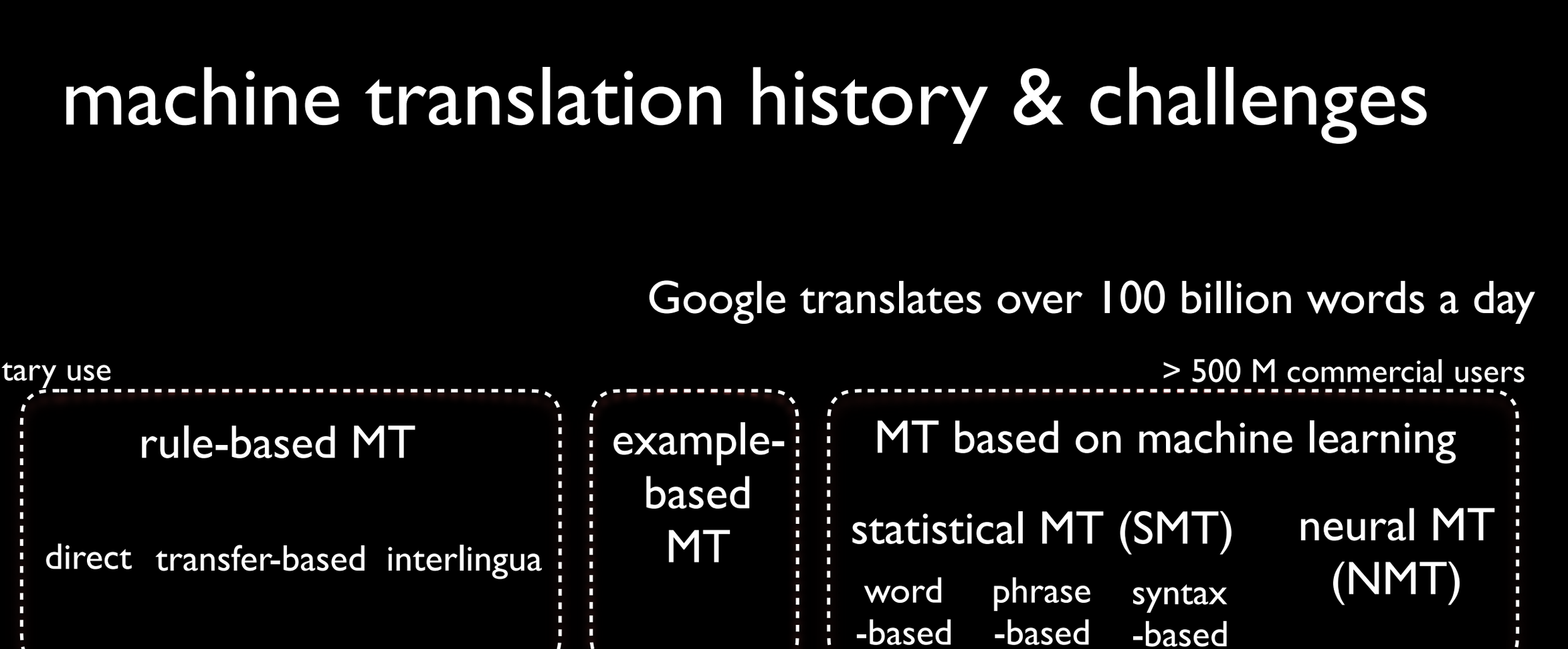
example-based MT

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)



problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

machine translation history & challenges

Google translates over 100 billion words a day

military use

rule-based MT

direct transfer-based interlingua

example-based MT

> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)

problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

machine translation history & challenges

Google translates over 100 billion words a day

military use

> 500 M commercial users

rule-based MT

direct transfer-based interlingua

example-based MT

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)

problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

Question #1: how to enhance NMT robustness?

machine translation history & challenges

Google translates over 100 billion words a day

military use

rule-based MT

direct transfer-based interlingua

example-based MT

> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)

problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

machine translation history & challenges

Google translates over 100 billion words a day

military use

rule-based MT

direct transfer-based interlingua

example-based MT

> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)

problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

machine translation history & challenges

Google translates over 100 billion words a day

military use

> 500 M commercial users

rule-based MT

direct transfer-based interlingua

example-based MT

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)

problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

Question #2: how to increase interpretability?

machine translation history & challenges

Google translates over 100 billion words a day

military use

rule-based MT

direct transfer-based interlingua

example-based MT

> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)

problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

machine translation history & challenges

Google translates over 100 billion words a day

military use

rule-based MT

direct transfer-based interlingua

example-based MT

> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based phrase -based syntax -based

neural MT (NMT)

problem of each approach:

lack generality	language dependent	difficult to define	lack generality	lack context	need syntax structure	human defined features	**robustness interpretability**

machine translation using machine learning

> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based

phrase -based

syntax -based

neural MT (NMT)

machine translation using machine learning

noisy channel model

> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based

phrase -based

syntax -based

neural MT (NMT)

machine translation using machine learning

noisy channel model

[Shannon, 1948] Information Theory

> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based

phrase -based

syntax -based

neural MT (NMT)

machine translation using machine learning

> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based

phrase -based

syntax -based

neural MT (NMT)

noisy channel model

[Shannon, 1948] Information Theory

- input: source sentence (observation) f
- output: target sentence (decision) e
- Bayes decision rule

machine translation using machine learning

> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based | phrase -based | syntax -based

neural MT (NMT)

noisy channel model

[Shannon, 1948] Information Theory

- input: source sentence (observation) f
- output: target sentence (decision) e
- Bayes decision rule

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

machine translation using machine learning

> 500 M commercial users

MT based on machine learning

statistical MT (SMT)

word -based | phrase -based | syntax -based

neural MT (NMT)

noisy channel model

[Shannon, 1948] Information Theory

- input: source sentence (observation) f
- output: target sentence (decision) e
- Bayes decision rule

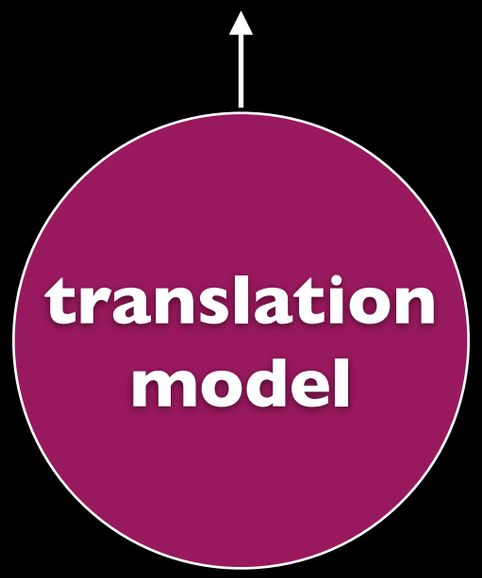
$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

machine translation using machine learning

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

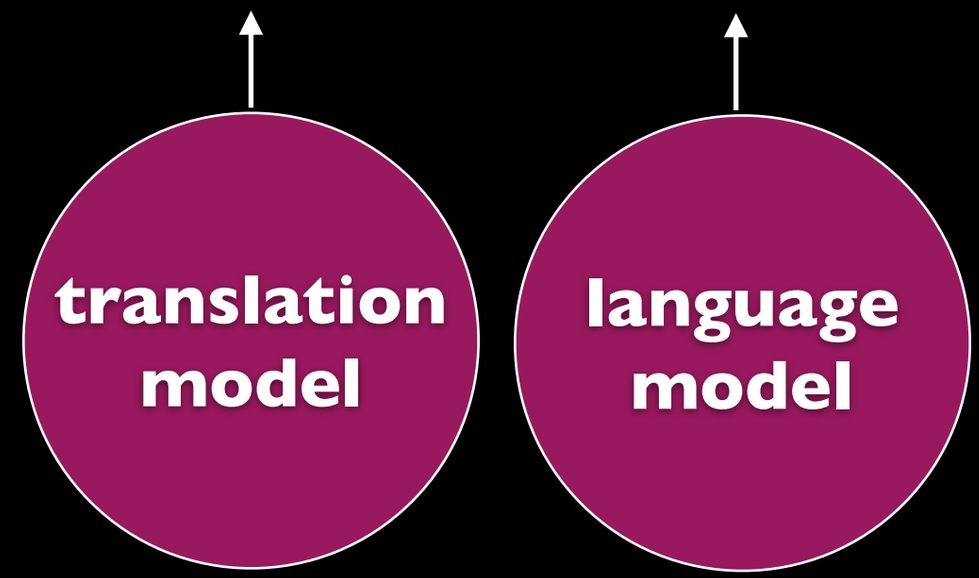
machine translation using machine learning

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$



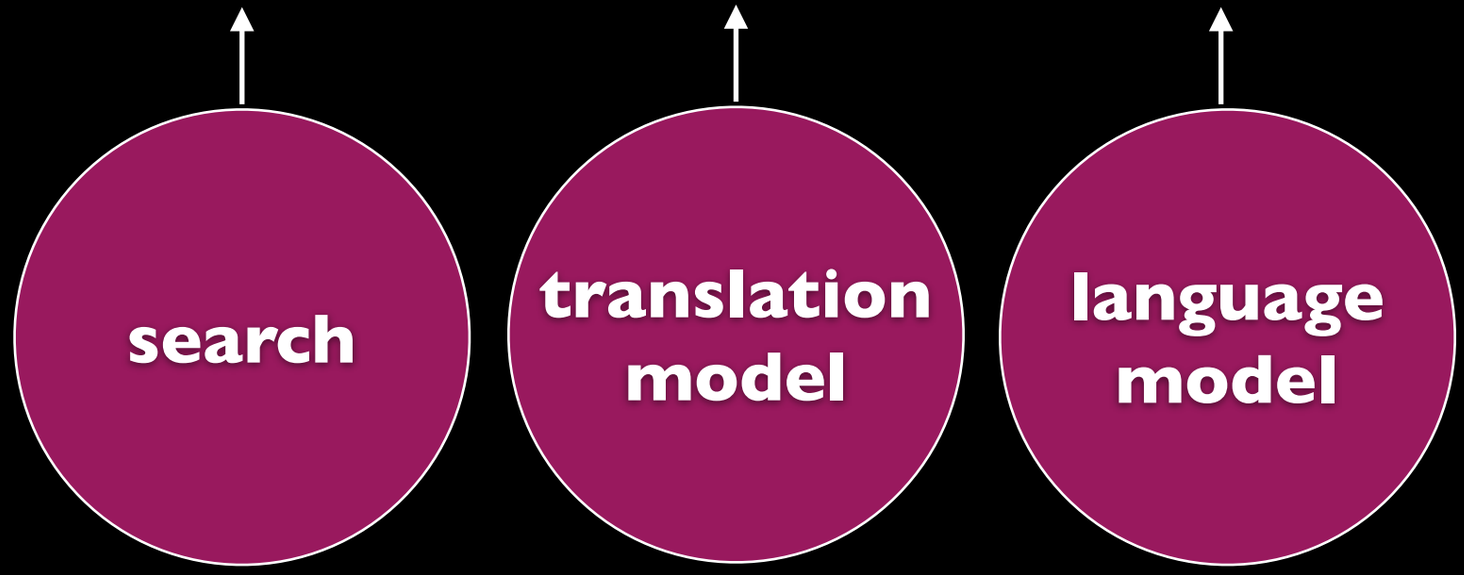
machine translation using machine learning

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$



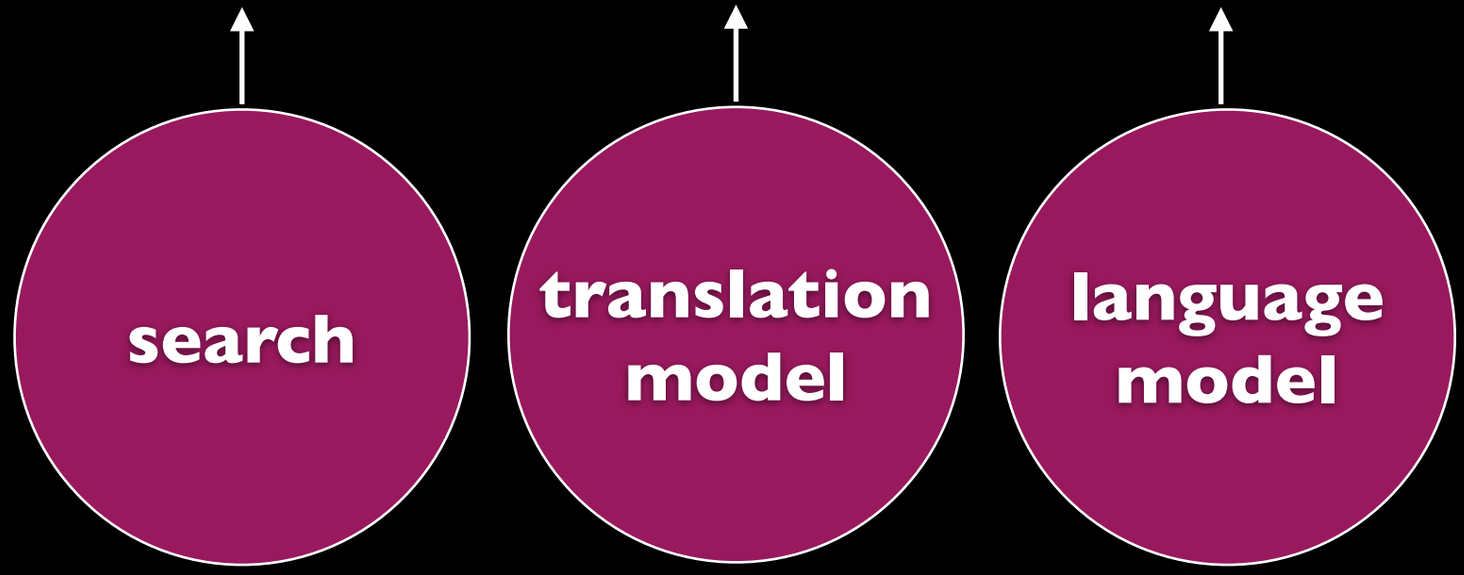
machine translation using machine learning

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$



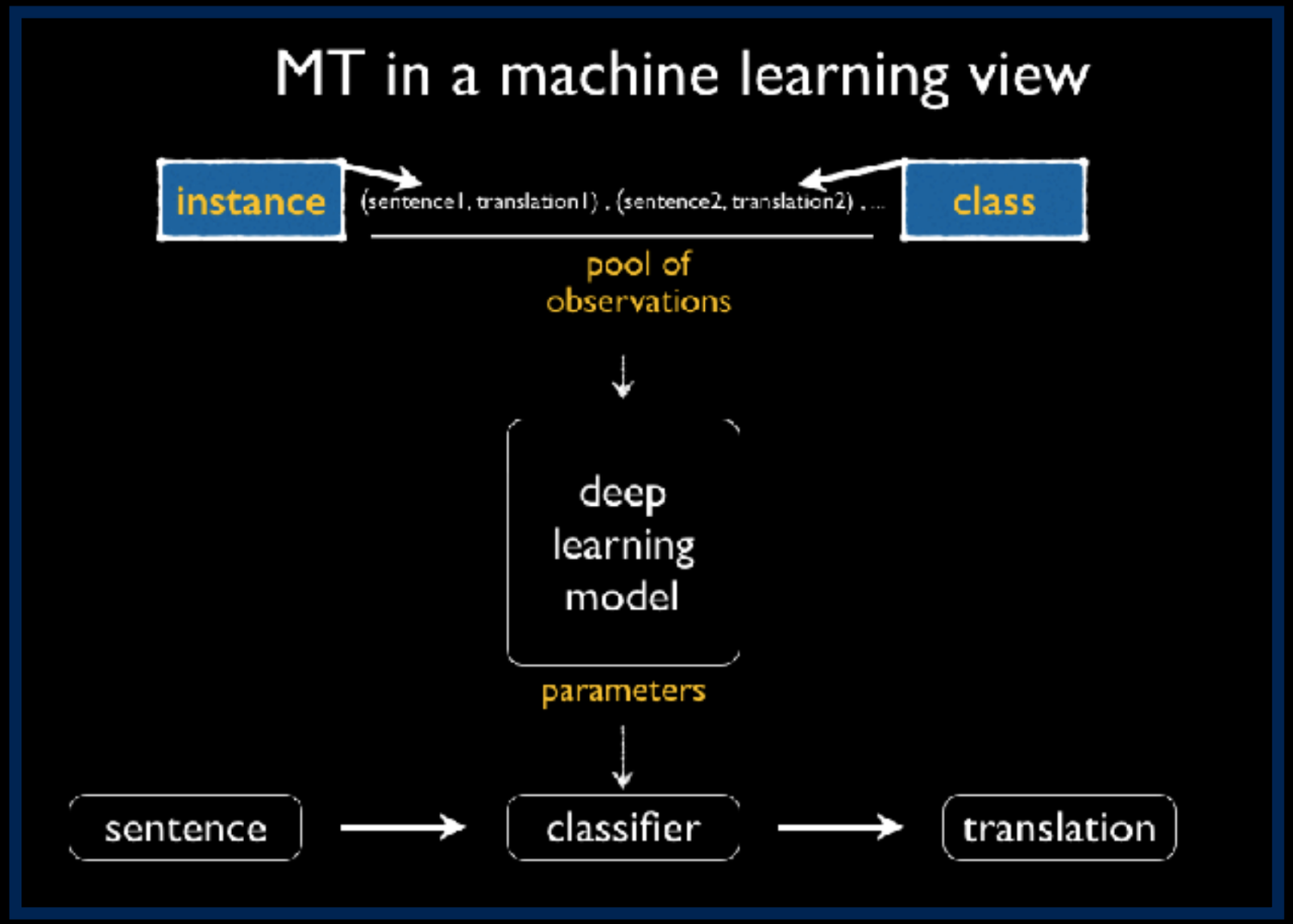
machine translation using machine learning

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$



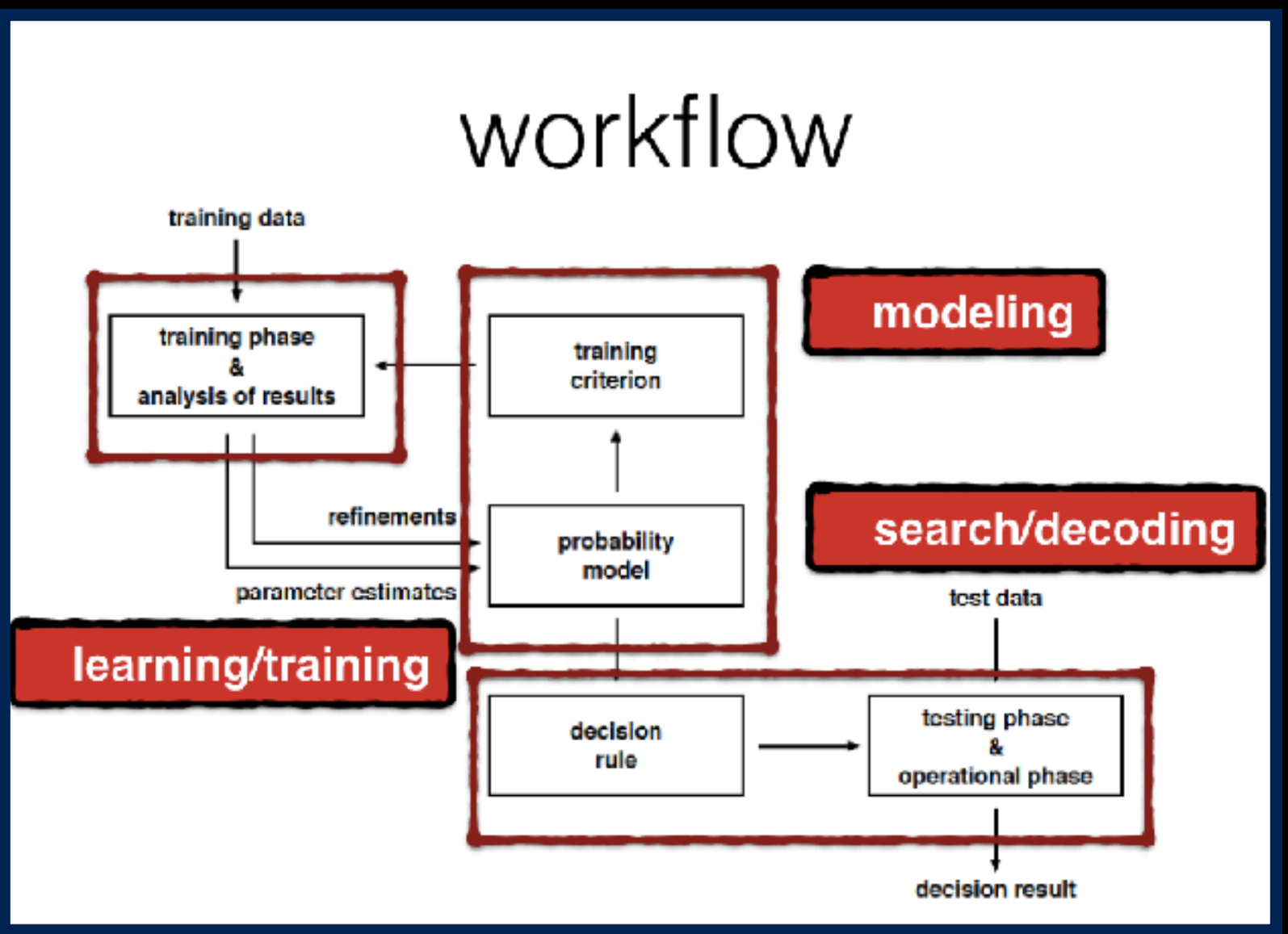
machine translation using machine learning

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$



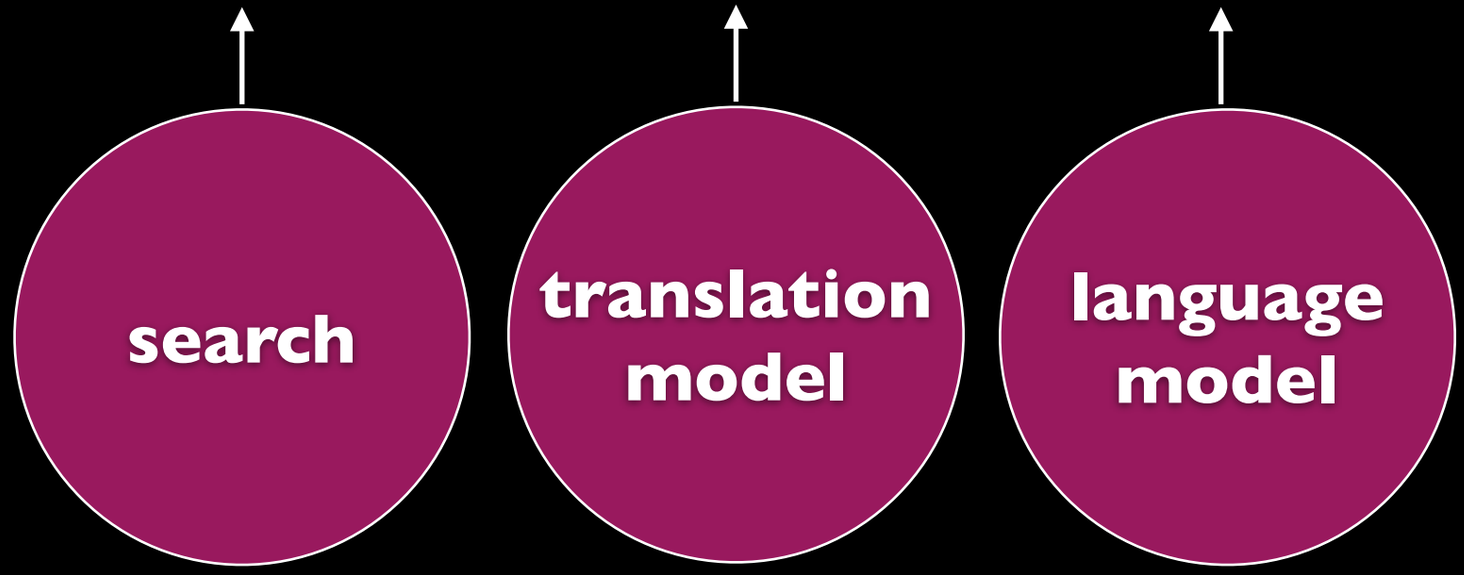
machine translation using machine learning

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

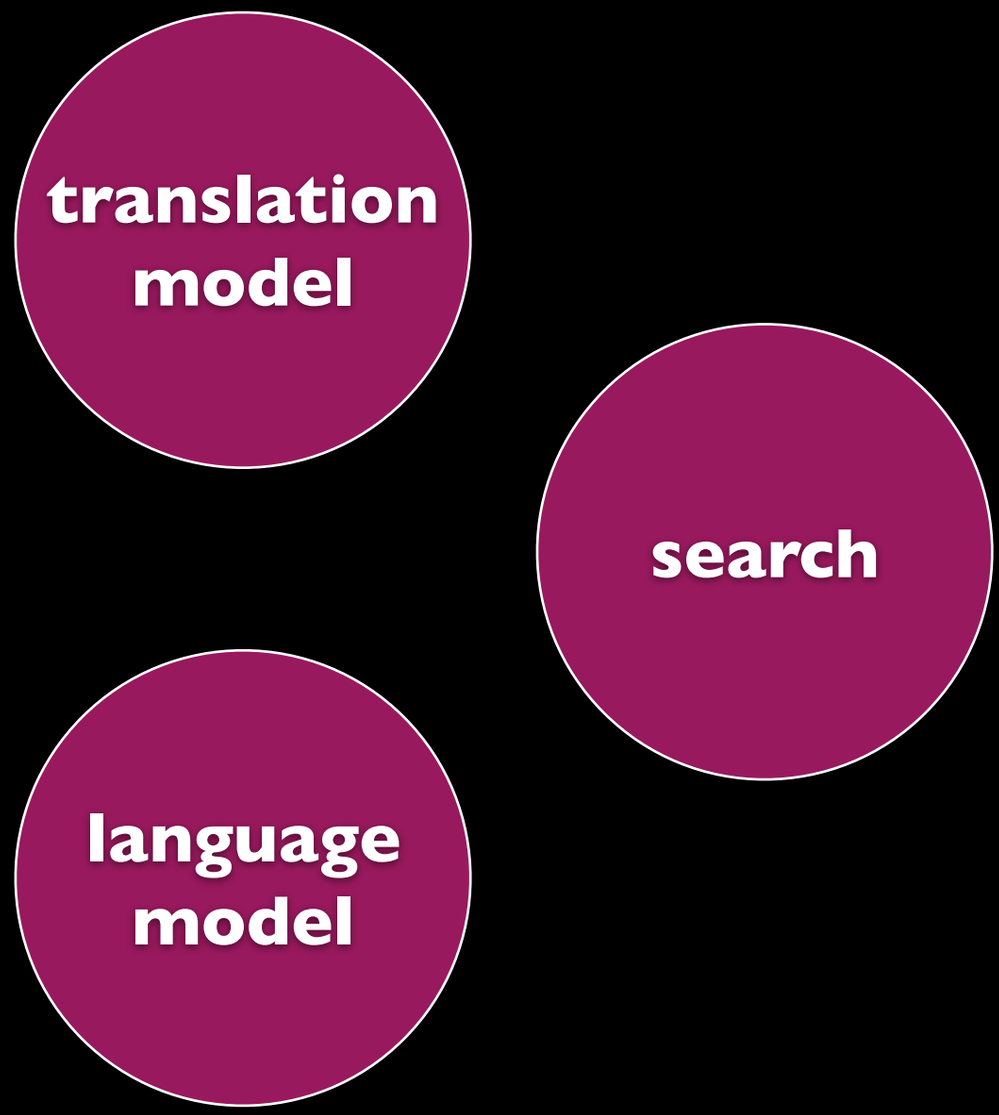


machine translation using machine learning

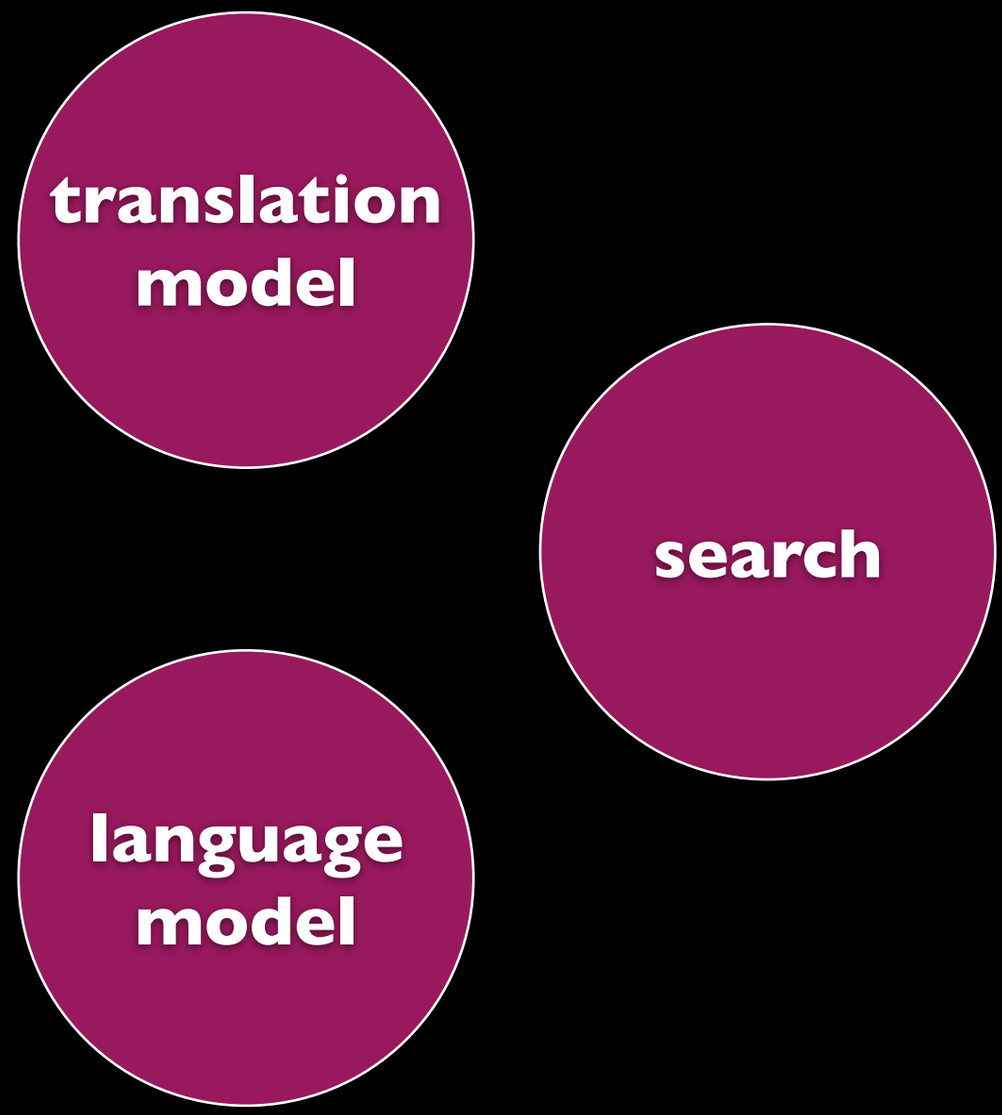
$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$



machine translation



neural machine translation

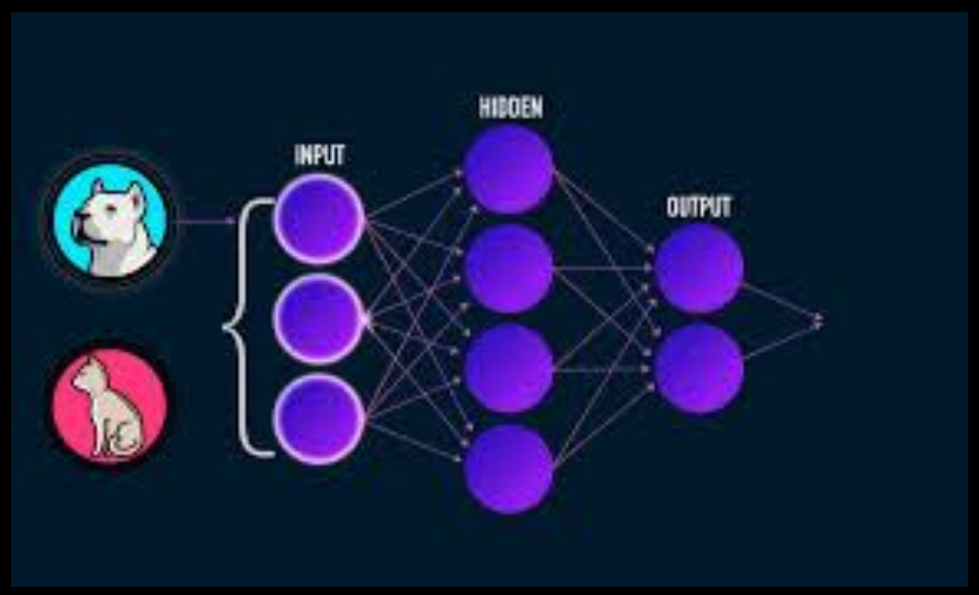


neural machine translation

translation model

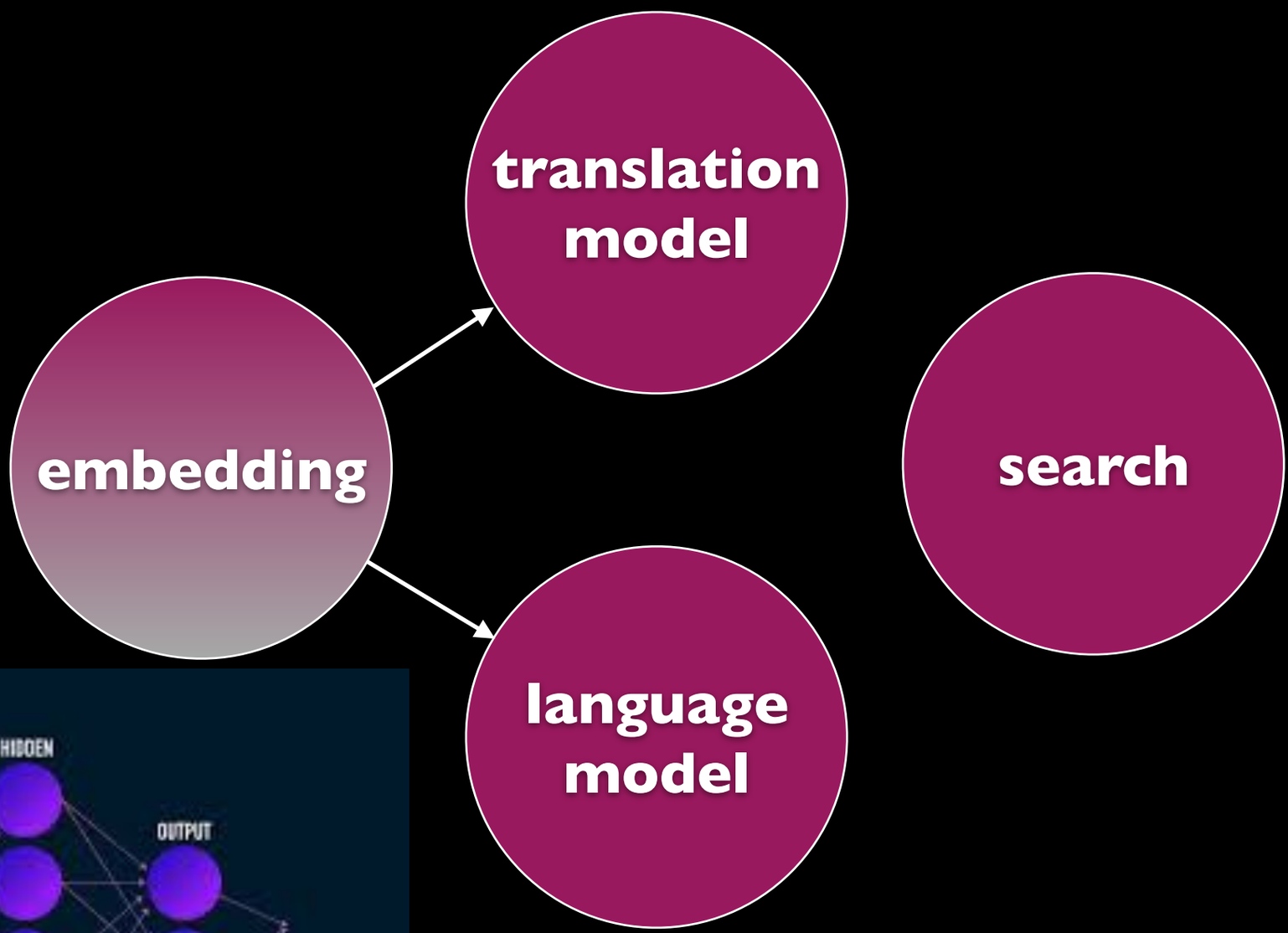
search

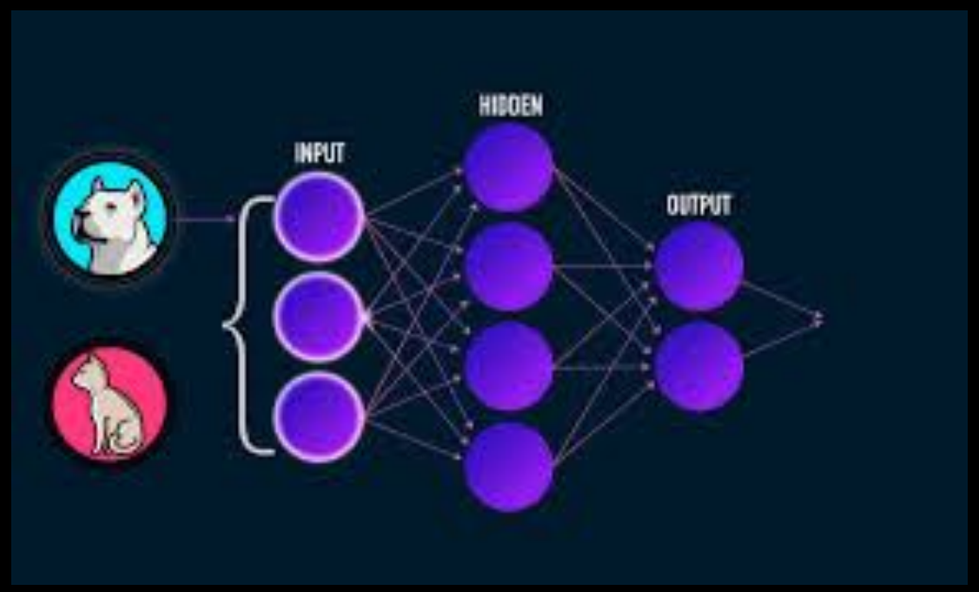
language model



MT with a single neural network

neural machine translation





MT with a single neural network

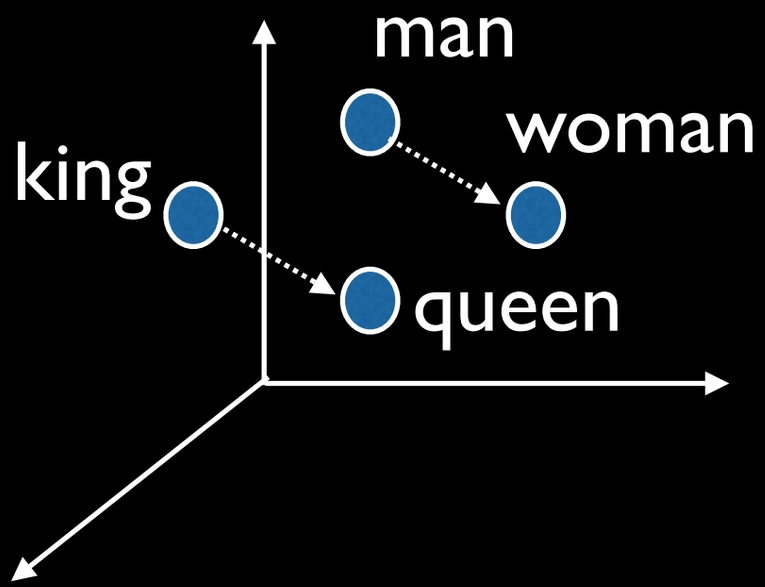
neural machine translation

neural machine translation

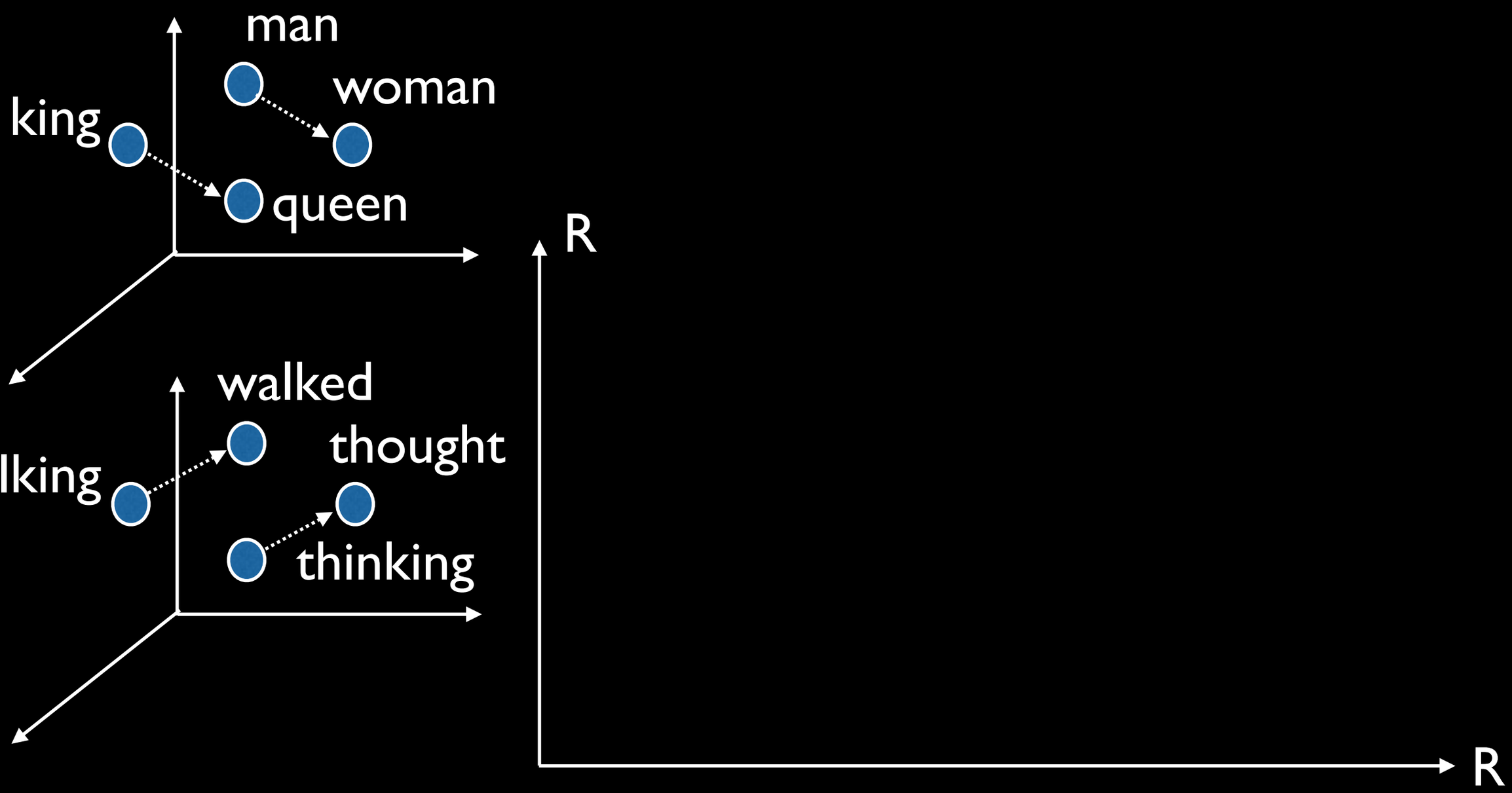
introduced in the previous tutorial

word embedding maps words to vectors

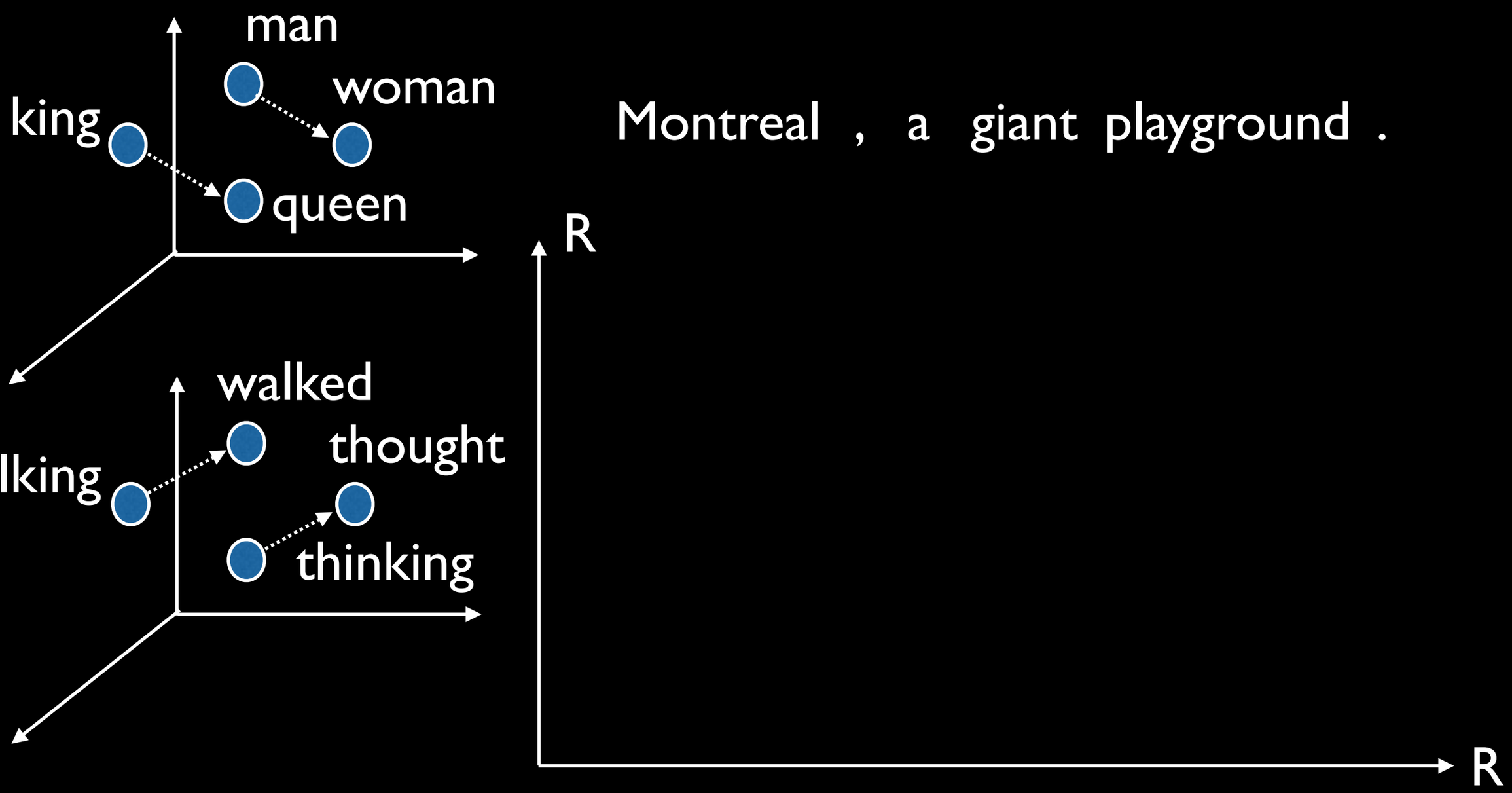
word embedding maps words to vectors



word embedding maps words to vectors

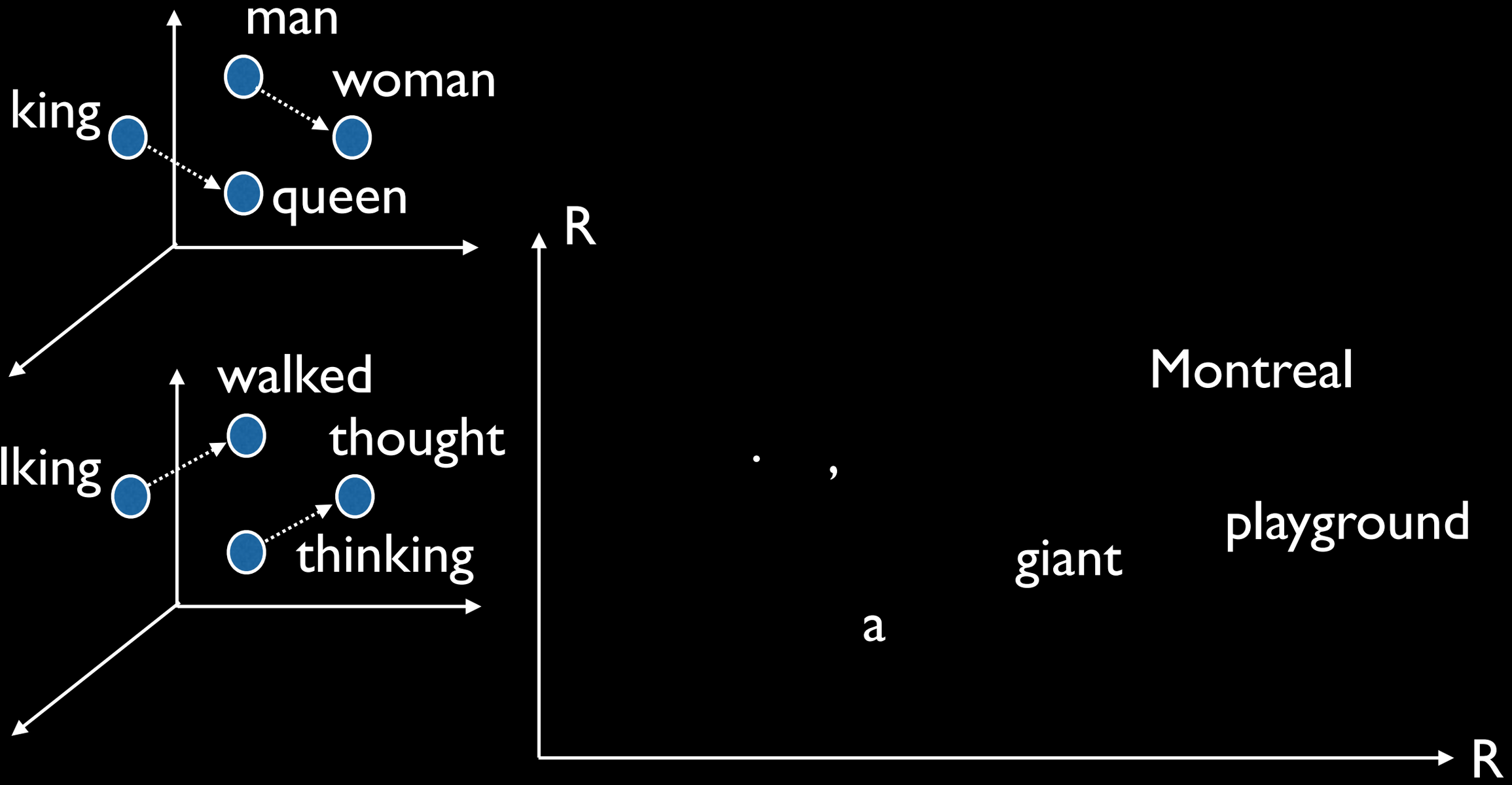


word embedding maps words to vectors

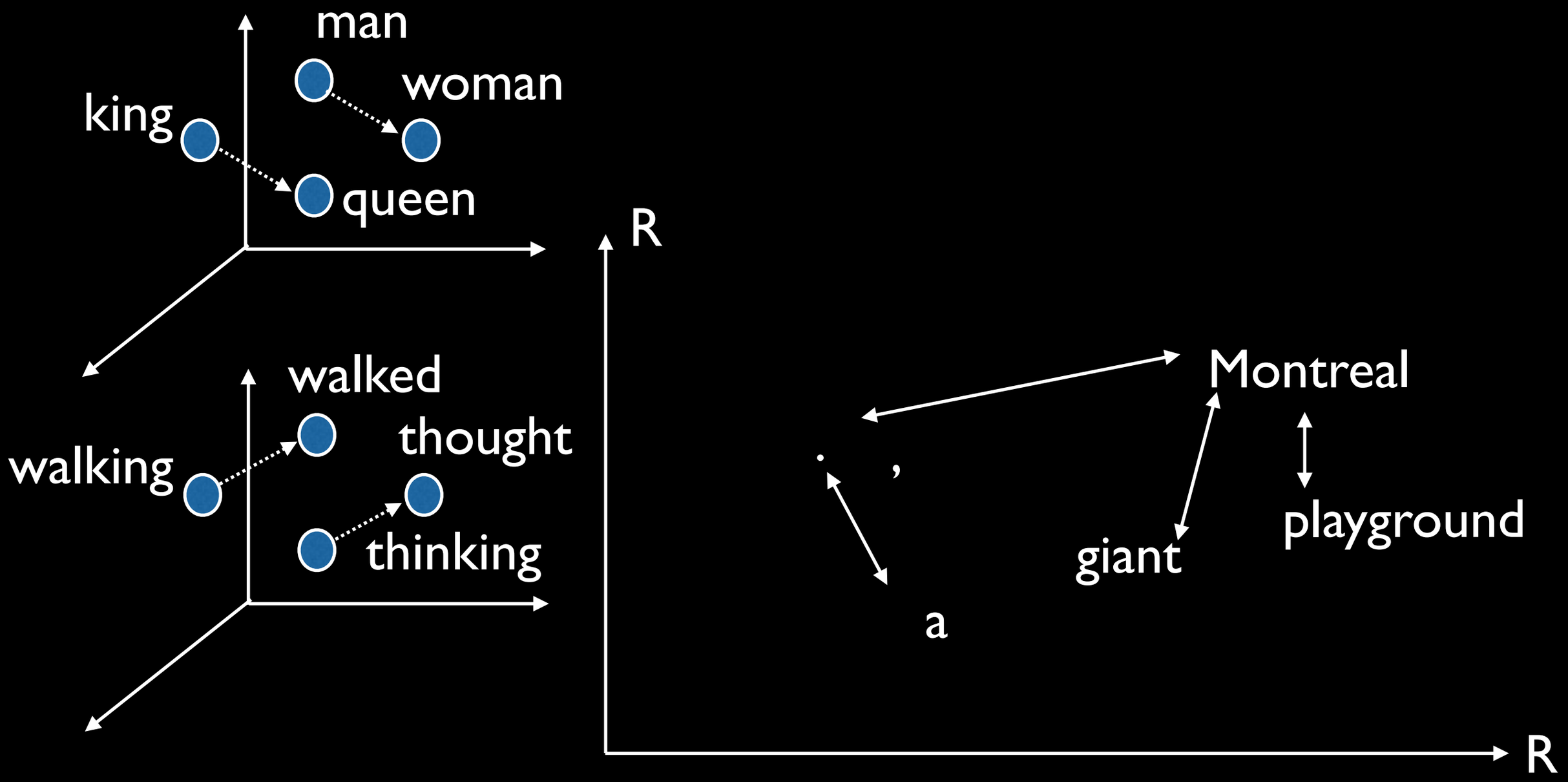


Montreal , a giant playground .

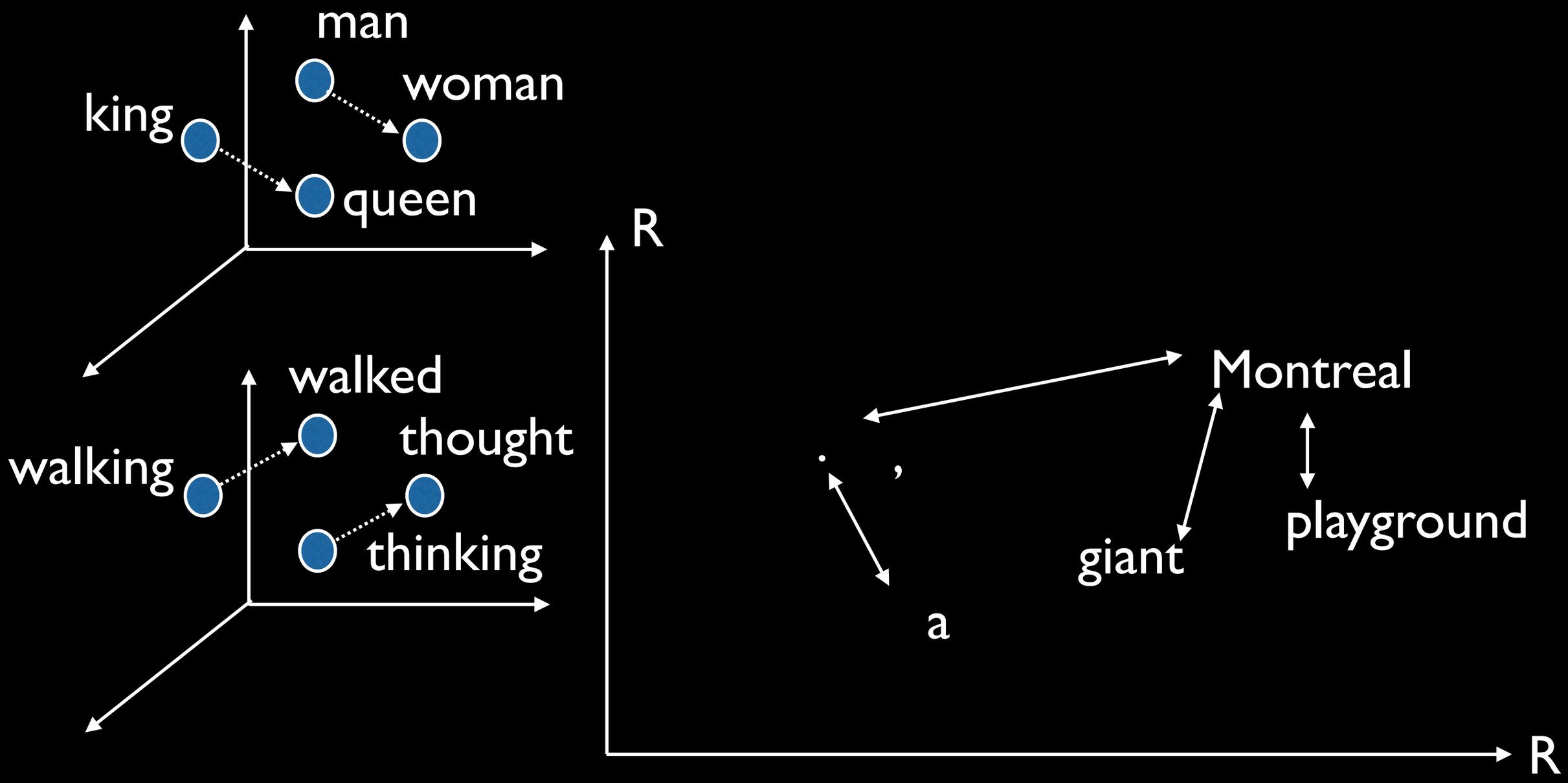
word embedding maps words to vectors



word embedding maps words to vectors

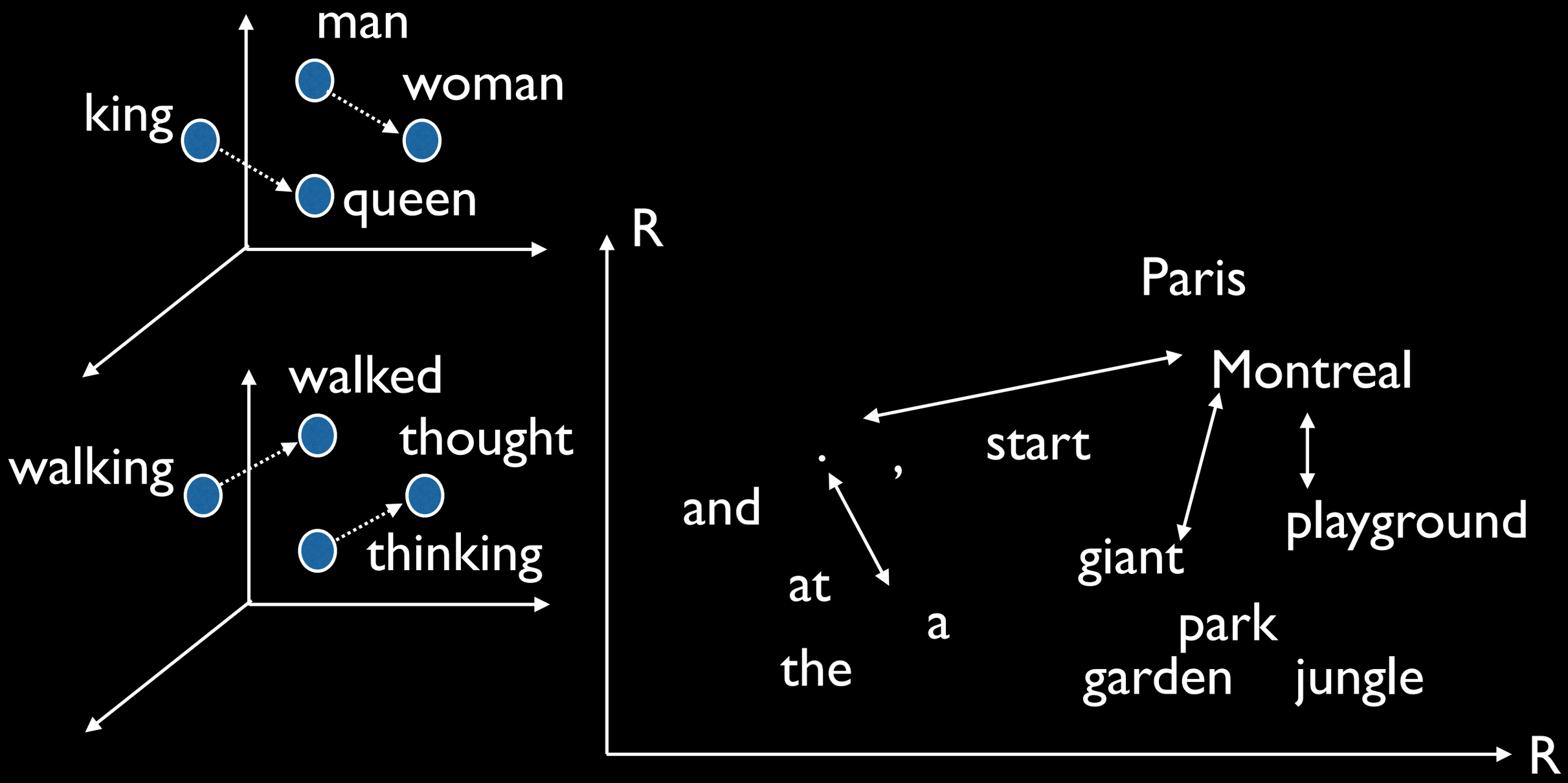


word embedding maps words to vectors



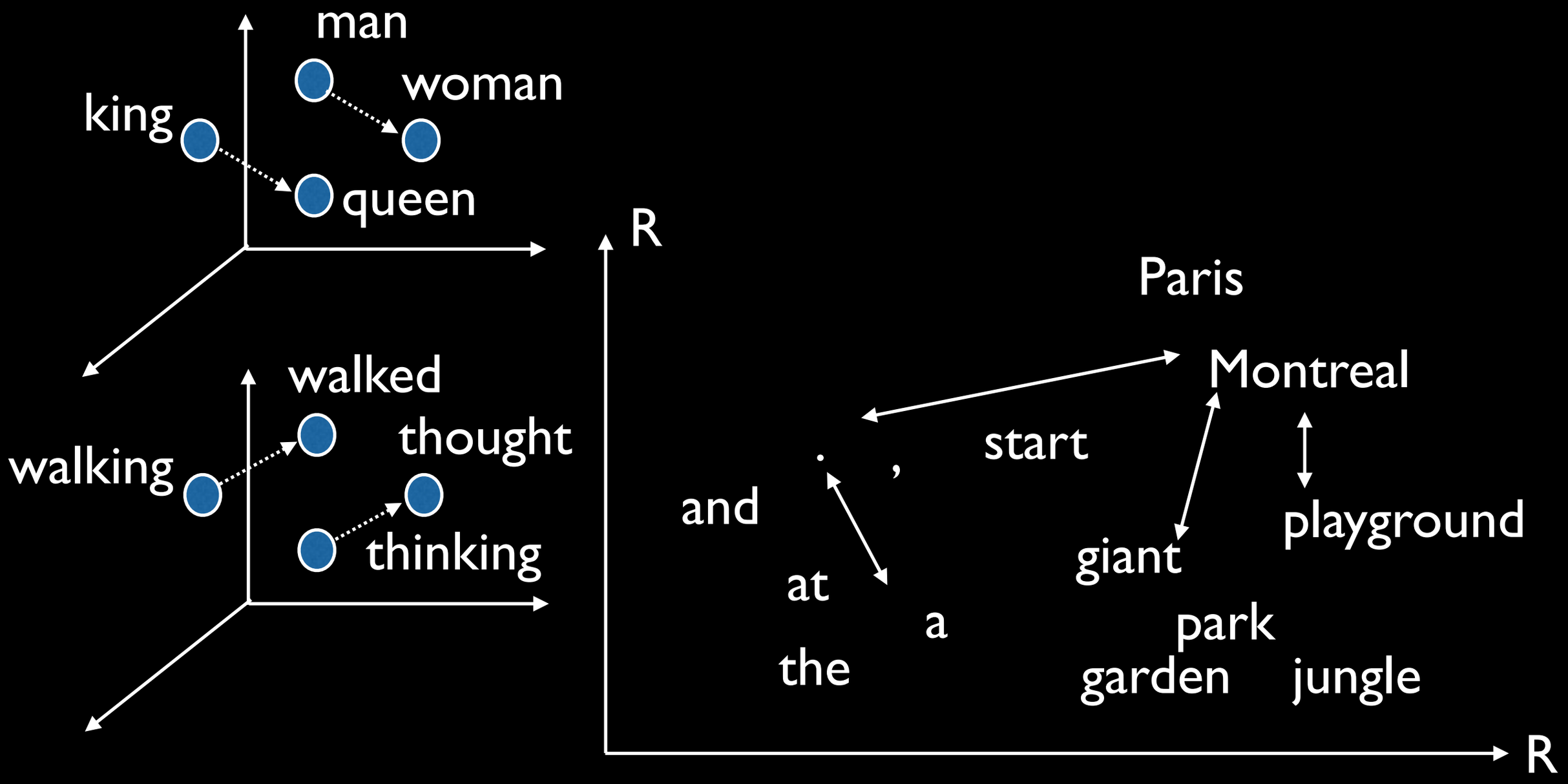
word pair semantic distance preserved

word embedding maps words to vectors



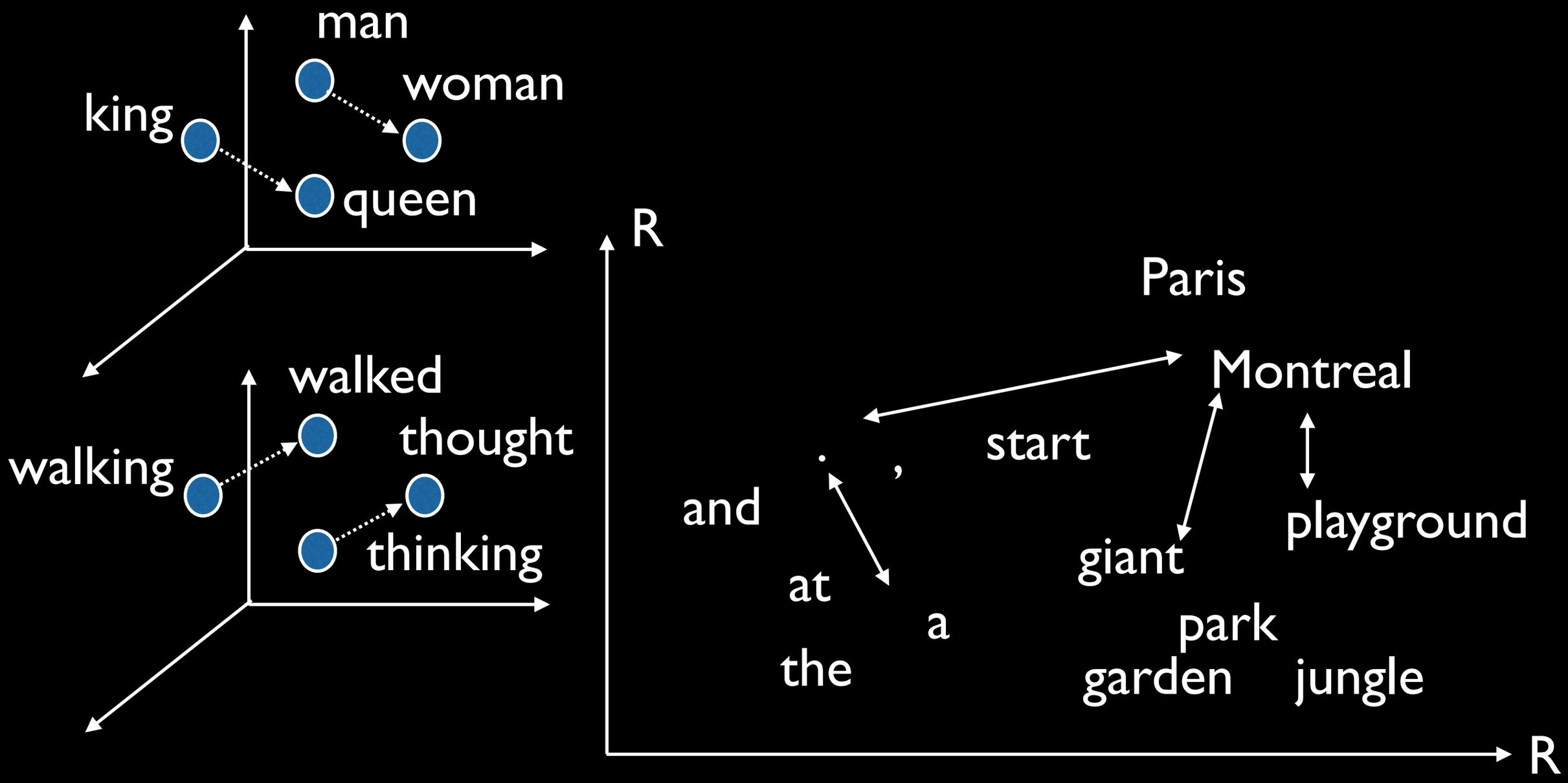
word pair semantic distance preserved

word embedding maps words to vectors



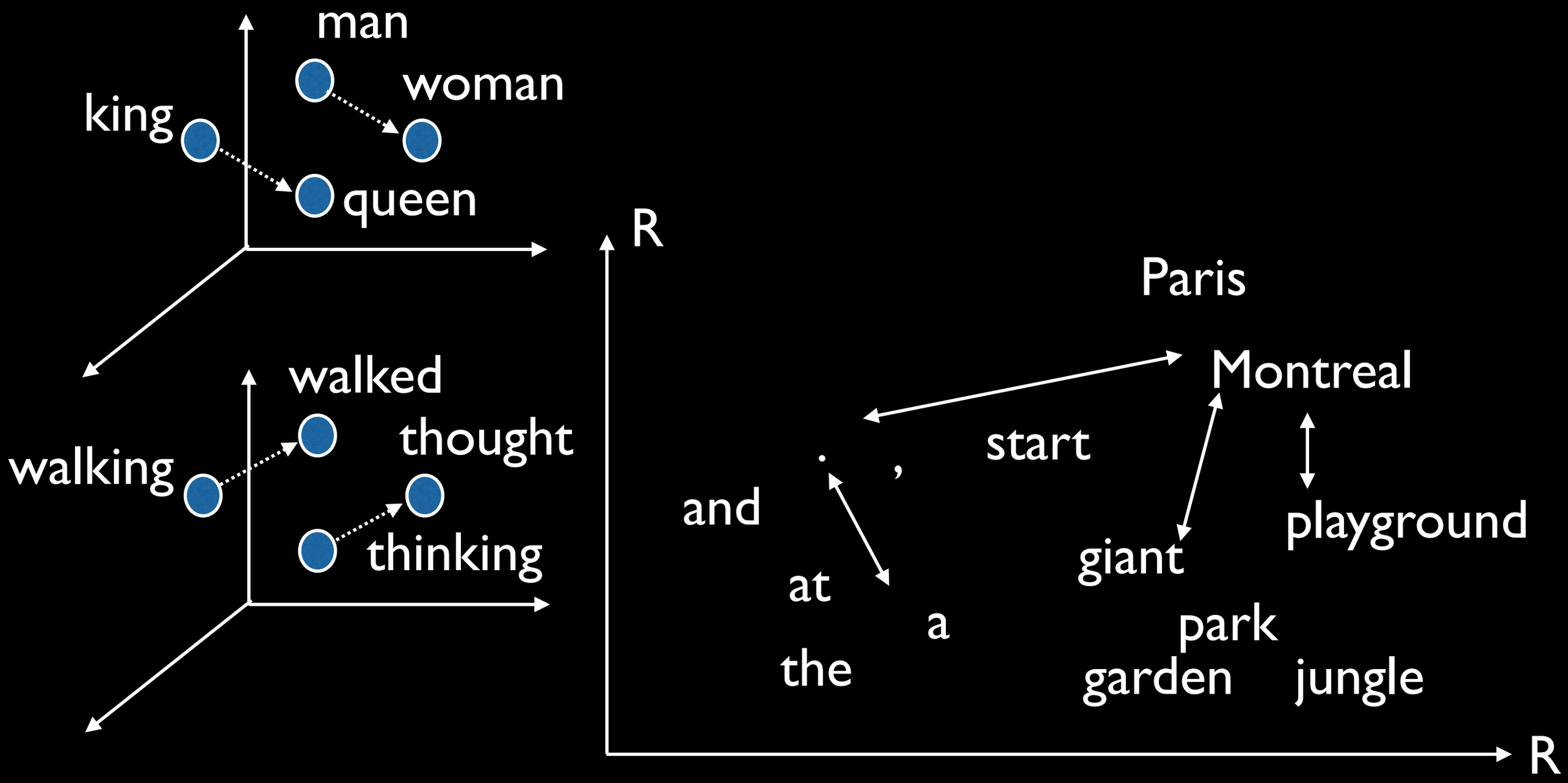
Question #3: better text embedding/representation?
BERT, ELMO, GloVec, FastText, …

word embedding maps words to vectors



word pair semantic distance preserved

word embedding maps words to vectors

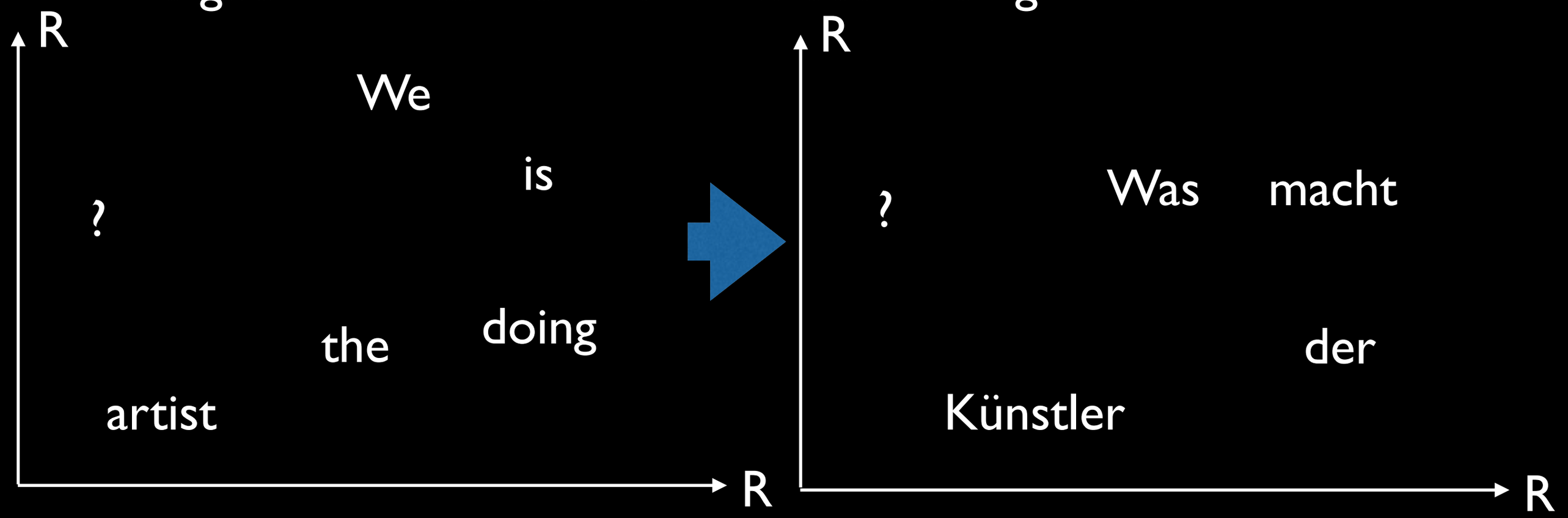


word pair semantic distance preserved

source to target language embedding

Source: Was macht der Künstler ?

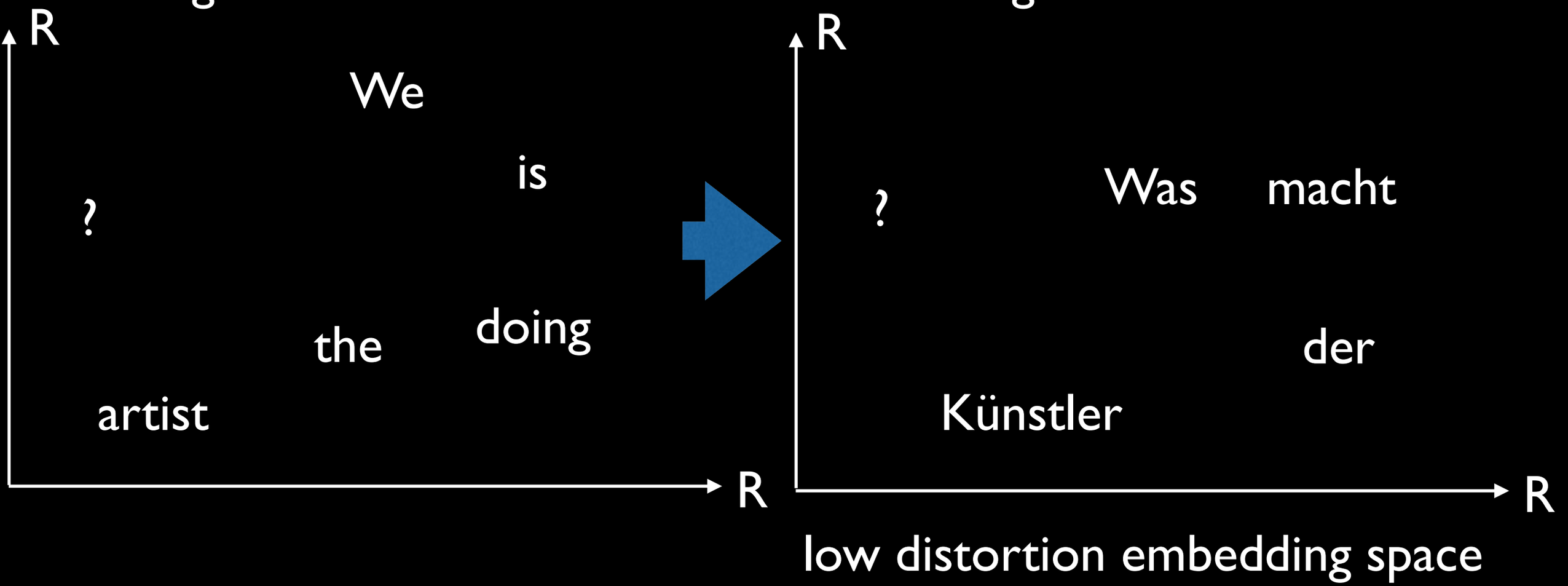
Target: What is the artist doing ?



source to target language embedding

Source: Was macht der Künstler ?

Target: What is the artist doing ?

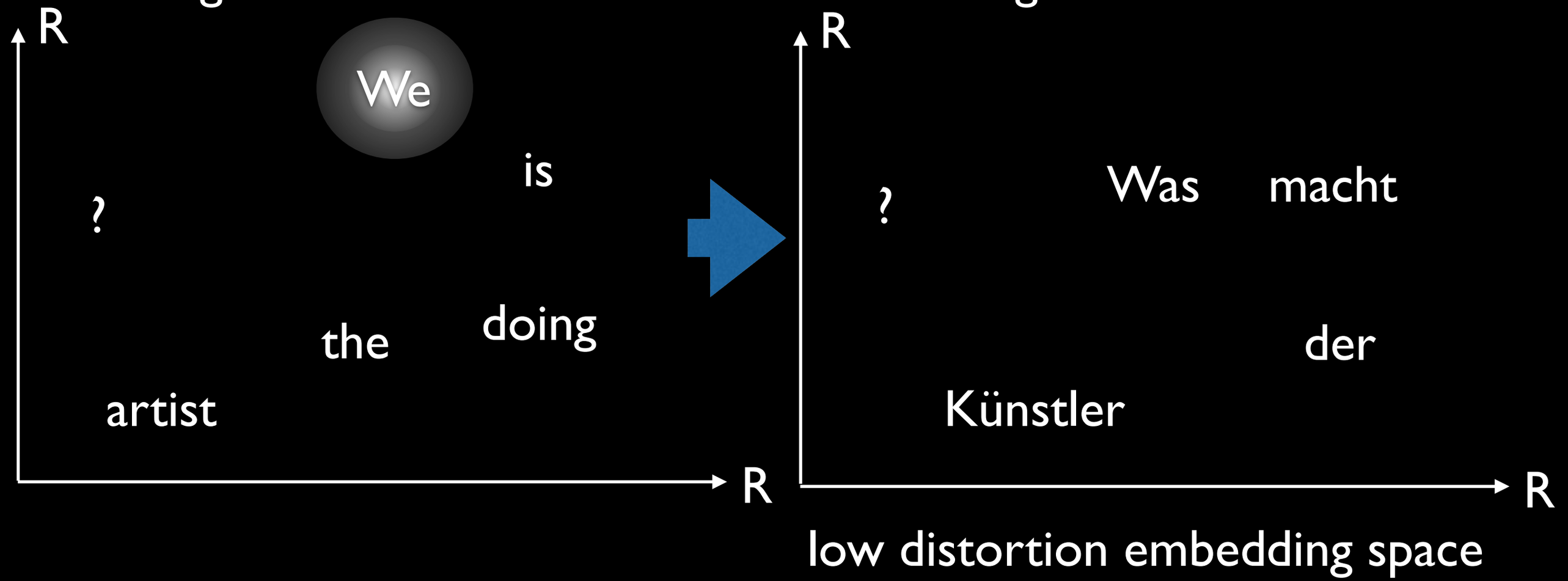


low distortion embedding space

source to target language embedding

Source: Was macht der Künstler ?

Target: What is the artist doing ?

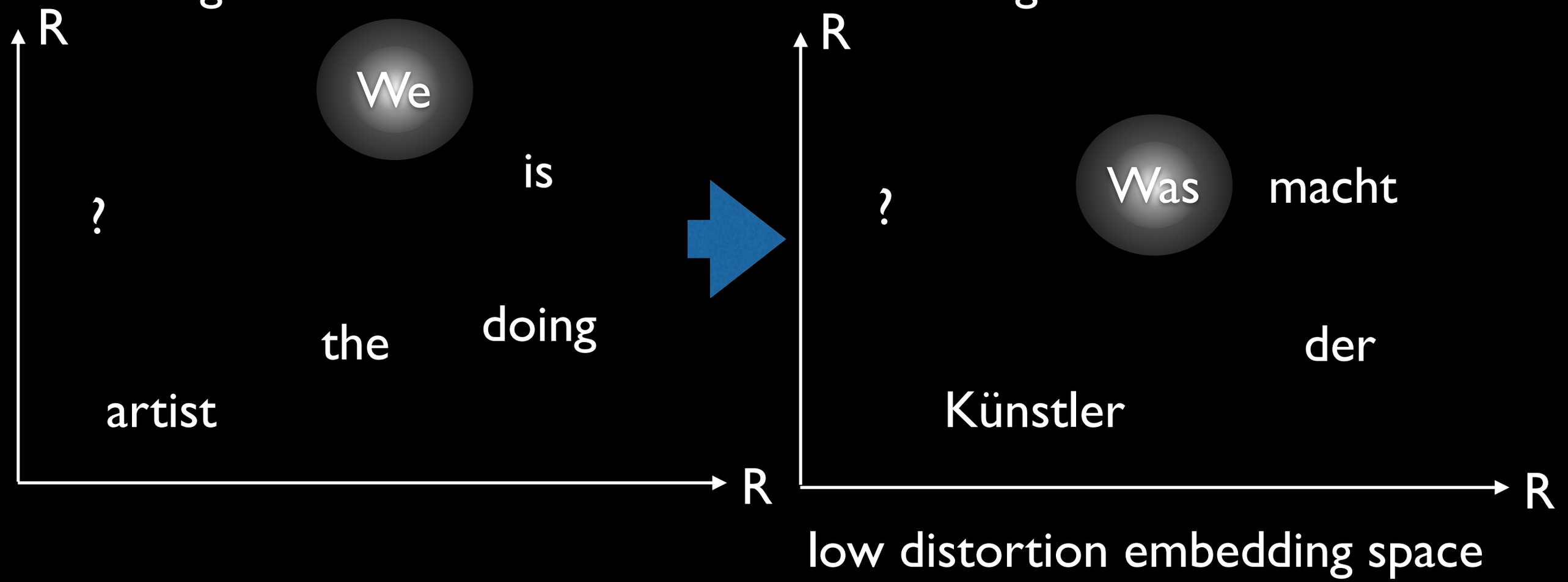


low distortion embedding space

source to target language embedding

Source: Was macht der Künstler ?

Target: What is the artist doing ?

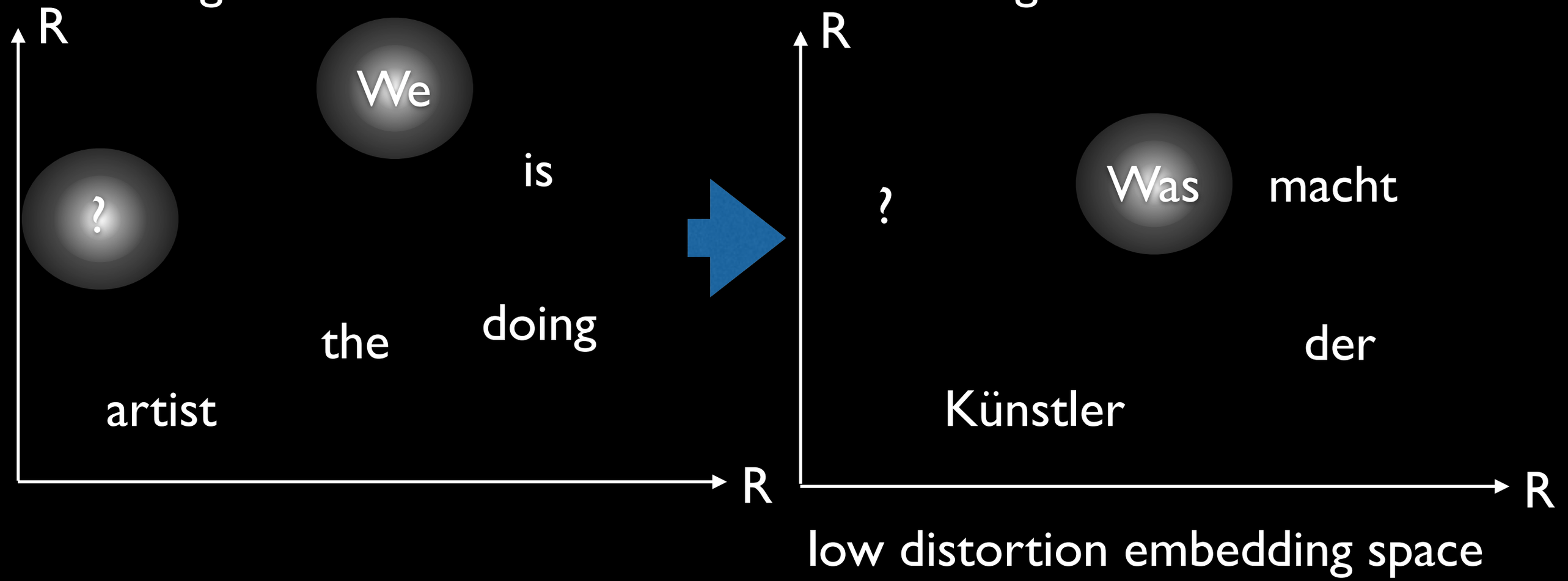


low distortion embedding space

source to target language embedding

Source: Was macht der Künstler ?

Target: What is the artist doing ?

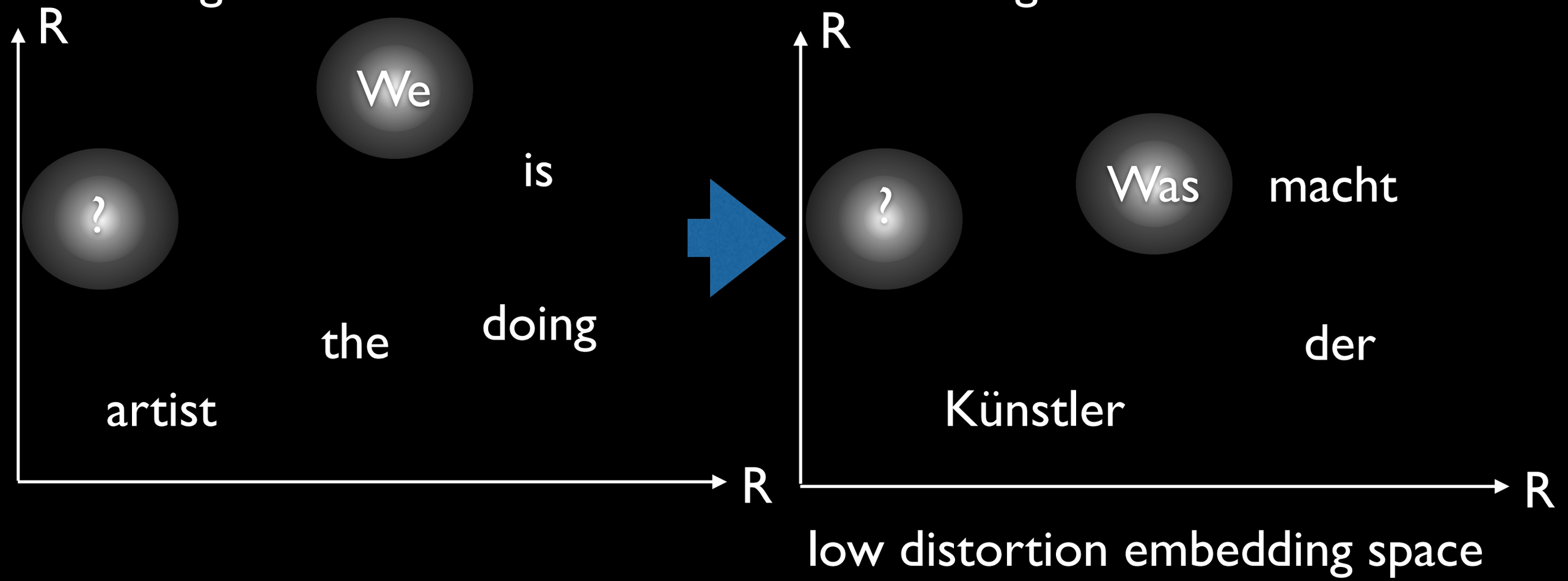


low distortion embedding space

source to target language embedding

Source: Was macht der Künstler ?

Target: What is the artist doing ?

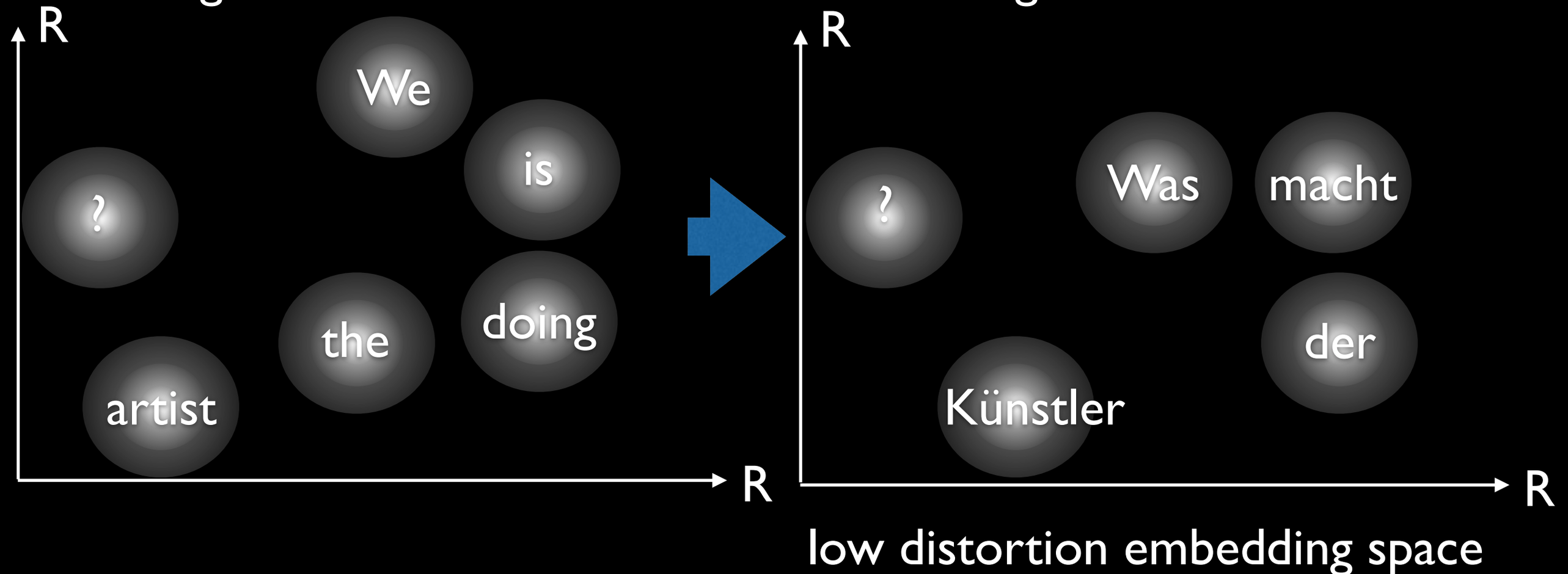


low distortion embedding space

source to target language embedding

Source: Was macht der Künstler ?

Target: What is the artist doing ?

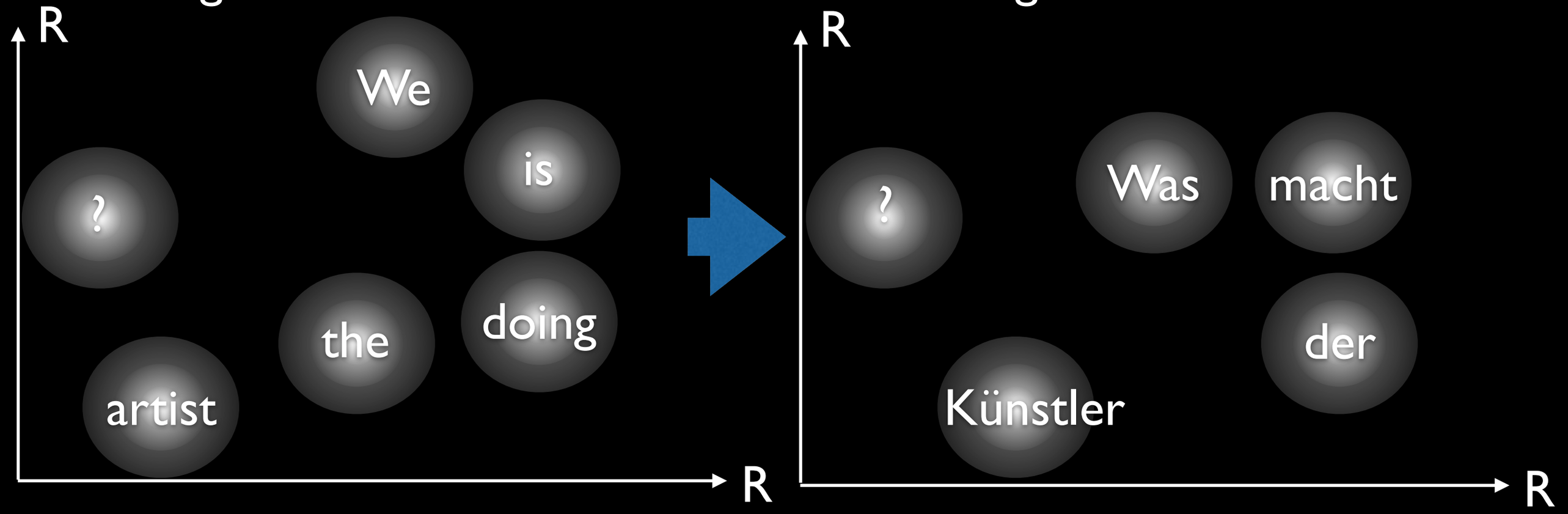


low distortion embedding space

source to target language embedding

Source: Was macht der Künstler ?

Target: What is the artist doing ?

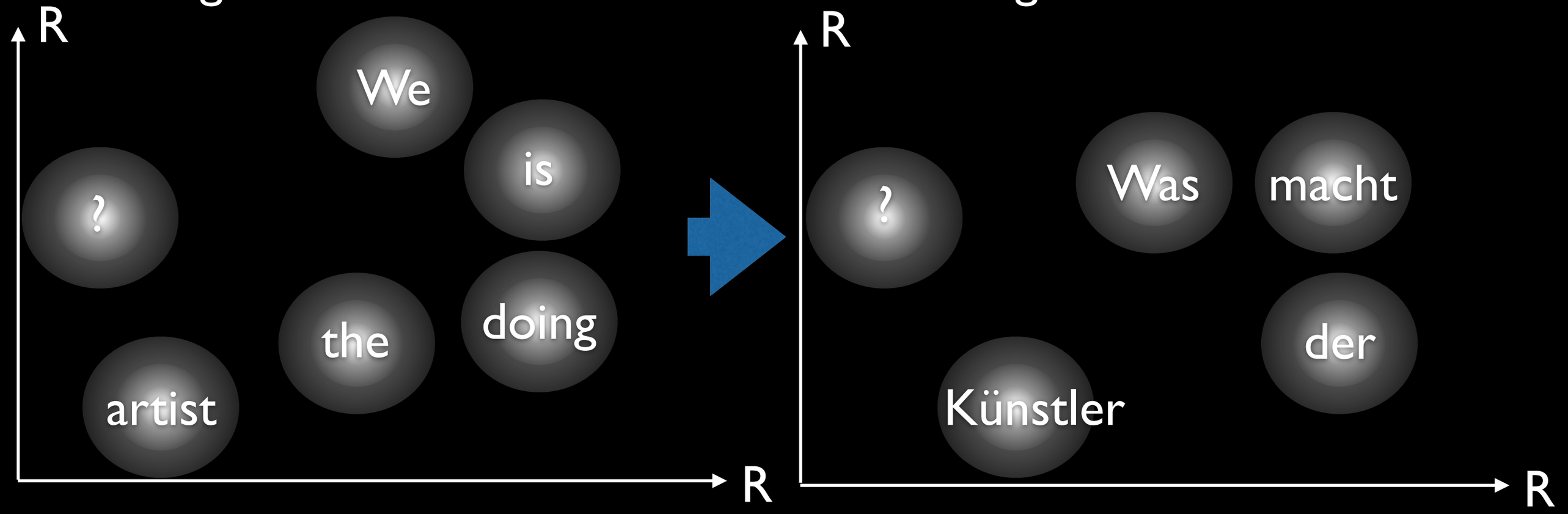


Can we learn only from distance in low distortion embedding space ?

source to target language embedding

Source: Was macht der Künstler ?

Target: What is the artist doing ?



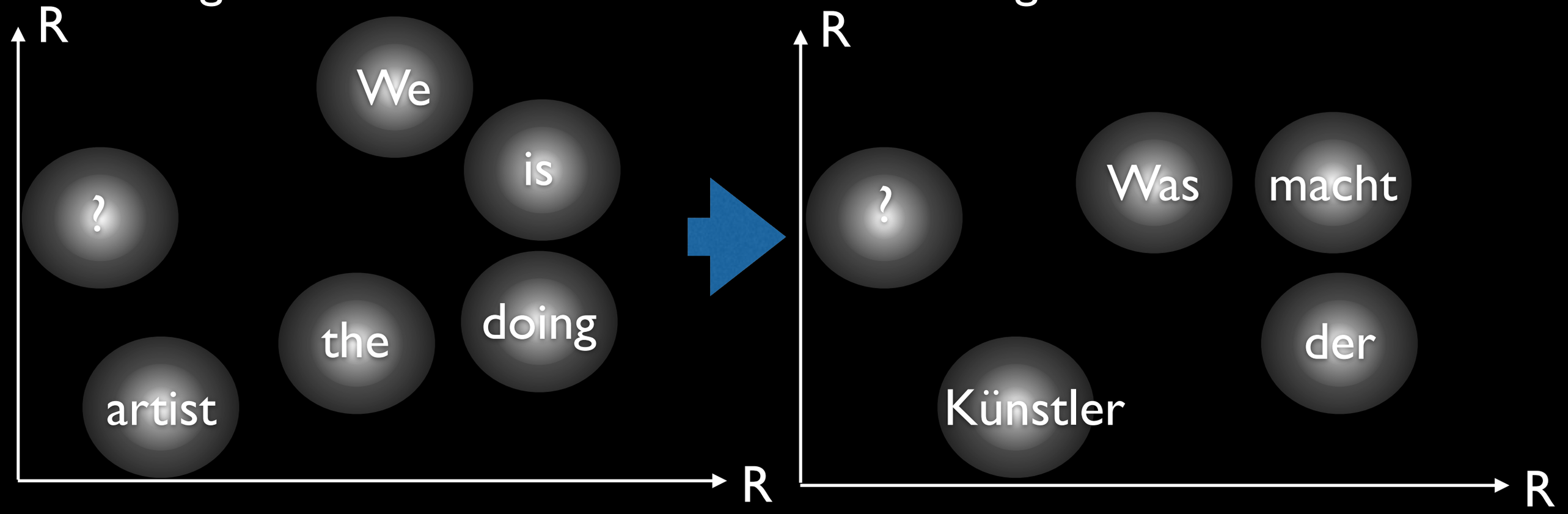
Can we learn only from distance in low distortion embedding space ?

No, in NLP we need language models

source to target language embedding

Source: Was macht der Künstler ?

Target: What is the artist doing ?



Can we learn only from distance in low distortion embedding space ?

No, in NLP we need language models

setting: multi-class classification + metric structure + 2 experts
we cannot combine them to a better one by querying them [PYX, 16]

machine translation components

what are

language models

machine translation components

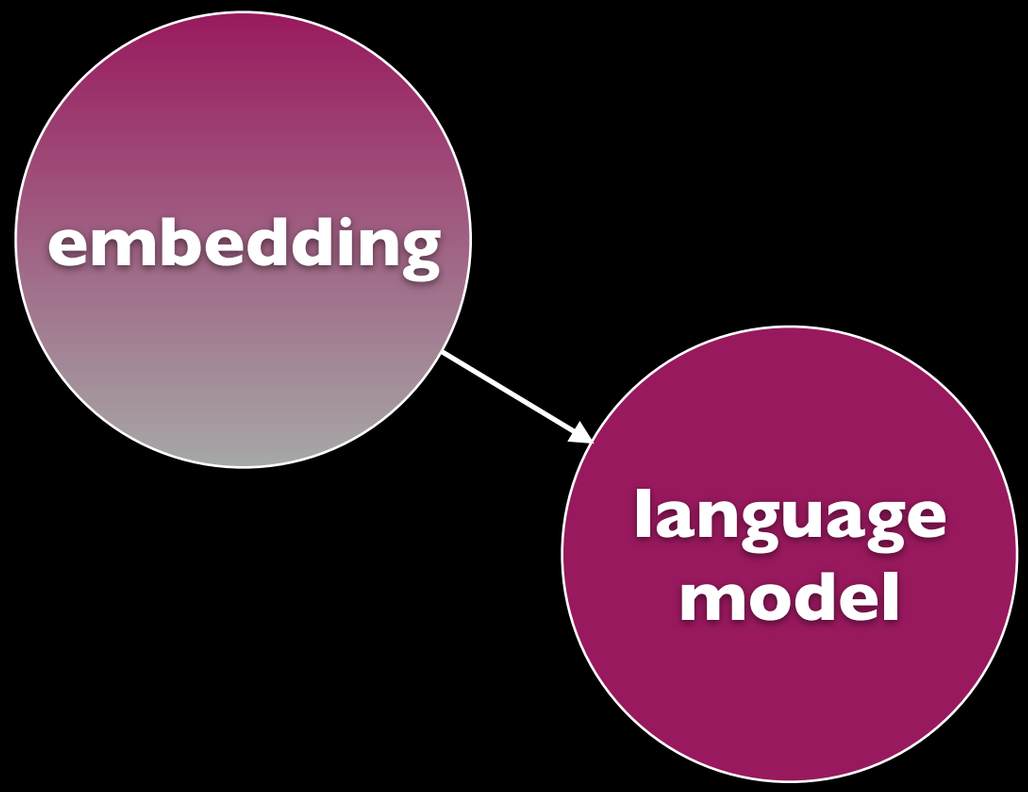
what are

language models

machine translation components

what are

language models



n-gram language model

n-gram language model

previously

n-gram language model

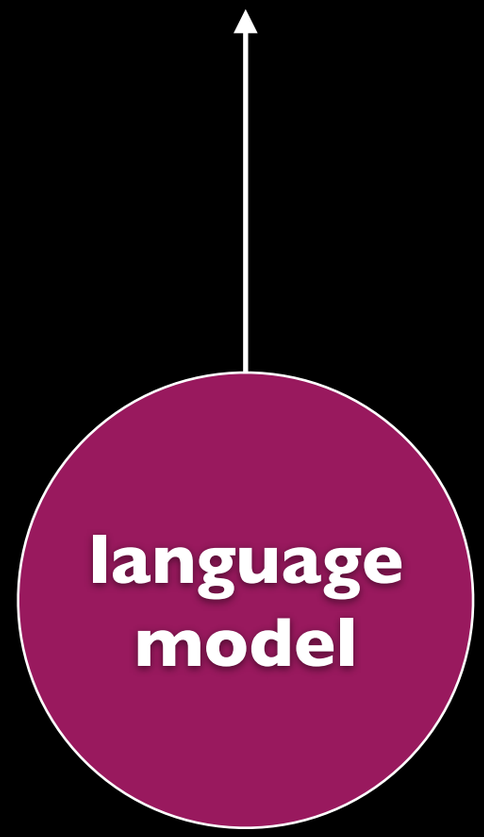
previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

n-gram language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$



evaluating language models: perplexity

let e_i be a word in the document that contains N words

PPL (perplexity) is measured as

$$\log \text{PPL} = -\frac{1}{N}\sum_{i=1}^{N} \log P(e_i|h_i)$$

h_i is the history of word e_i

language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

let e_1^I be a sentence of a sequence of words $e_1, e_2 \cdots e_I$

language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

let e_1^I be a sentence of a sequence of words $e_1, e_2 \cdots e_I$

$$Pr(e_1^I) = Pr(e_1, e_2, \cdots, e_I)$$
$$= \prod_i Pr(e_i|e_1, \cdots, e_{i-1}) \quad \text{chain rule}$$
$$\approx P(e_i|e_1, \cdots, e_{i-n+1}) \quad \text{estimated as relative frequency}$$

language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

let e_1^I be a sentence of a sequence of words $e_1, e_2 \cdots e_I$

$$Pr(e_1^I) = Pr(e_1, e_2, \cdots, e_I)$$
$$= \prod_i Pr(e_i|e_1, \cdots, e_{i-1}) \quad \text{chain rule}$$
$$\approx P(e_i|e_1, \cdots, e_{i-n+1}) \quad \text{estimated as relative frequency}$$

n-gram based; if n-gram unseen, then back-off

language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

let e_1^I be a sentence of a sequence of words $e_1, e_2 \cdots e_I$

$$Pr(e_1^I) = Pr(e_1, e_2, \cdots, e_I)$$
$$= \prod_i Pr(e_i|e_1, \cdots, e_{i-1}) \quad \text{chain rule}$$

statistical → $\approx P(e_i|e_1, \cdots, e_{i-n+1})$ estimated as relative frequency

→ n-gram based; if n-gram unseen, then back-off

language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

let e_1^I be a sentence of a sequence of words $e_1, e_2 \cdots e_I$

$$Pr(e_1^I) = Pr(e_1, e_2, \cdots, e_I)$$
$$= \prod_i Pr(e_i|e_1, \cdots, e_{i-1}) \quad \text{chain rule}$$

statistical → $\approx P(e_i|e_1, \cdots, e_{i-n+1})$ estimated as relative frequency

→ n-gram based; if n-gram unseen, then back-off

language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

let e_1^I be a sentence of a sequence of words $e_1, e_2 \cdots e_I$

$$Pr(e_1^I) = Pr(e_1, e_2, \cdots, e_I)$$
$$= \prod_i Pr(e_i|e_1, \cdots, e_{i-1}) \quad \text{chain rule}$$

statistical → $\approx P(e_i|e_1, \cdots, e_{i-n+1})$ estimated as relative frequency

→ n-gram based; if n-gram unseen, then back-off

language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

let e_1^I be a sentence of a sequence of words $e_1, e_2 \cdots e_I$

$$Pr(e_1^I) = Pr(e_1, e_2, \cdots, e_I)$$
$$= \prod_i Pr(e_i|e_1, \cdots, e_{i-1}) \quad \text{chain rule}$$

statistical → $\approx P(e_i|e_1, \cdots, e_{i-n+1})$ estimated as relative frequency

→ n-gram based; if n-gram unseen, then back-off

language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

let e_1^I be a sentence of a sequence of words $e_1, e_2 \cdots e_I$

$$Pr(e_1^I) = Pr(e_1, e_2, \cdots, e_I)$$
$$= \prod_i Pr(e_i|e_1, \cdots, e_{i-1}) \quad \text{chain rule}$$

statistical → $\approx P(e_i|e_1, \cdots, e_{i-n+1})$ estimated as relative frequency

→ n-gram based; if n-gram unseen, then back-off

language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

let e_1^I be a sentence of a sequence of words $e_1, e_2 \cdots e_I$

$$Pr(e_1^I) = Pr(e_1, e_2, \cdots, e_I)$$
$$= \prod_i Pr(e_i|e_1, \cdots, e_{i-1}) \quad \text{chain rule}$$

statistical → $\approx P(e_i|\cdot, e_{i-n+1})$ estimated as relative frequency

→ n-gram based; if n-gram unseen, then back-off

language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

let e_1^I be a sentence of a sequence of words $e_1, e_2 \cdots e_I$

$$Pr(e_1^I) = Pr(e_1, e_2, \cdots, e_I)$$
$$= \prod_i Pr(e_i|e_1, \cdots, e_{i-1}) \quad \text{chain rule}$$

statistical → $\approx P(e_i|e_1, \cdots, e_{i-n+1})$ estimated as relative frequency

→ n-gram based; if n-gram unseen, then back-off

language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

let e_1^I be a sentence of a sequence of words $e_1, e_2 \cdots e_I$

$$Pr(e_1^I) = Pr(e_1, e_2, \cdots, e_I)$$
$$= \prod_i Pr(e_i|e_1, \cdots, e_{i-1}) \quad \text{chain rule}$$
$$\approx P(e_i|e_1, \cdots, e_{i-n+1})$$

statistical → estimated as relative frequency

→ n-gram based; if n-gram unseen, then back-off

language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

let e_1^I be a sentence of a sequence of words $e_1, e_2 \cdots e_I$

$$Pr(e_1^I) = Pr(e_1, e_2, \cdots, e_I)$$
$$= \prod_i Pr(e_i|e_1, \cdots, e_{i-1}) \quad \text{chain rule}$$

statistical → $\approx P(e_i|e_1, \cdots, e_{i-n+1})$ estimated as relative frequency

→ n-gram based; if n-gram unseen, then back-off

language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

let e_1^I be a sentence of a sequence of words $e_1, e_2 \cdots e_I$

$$Pr(e_1^I) = Pr(e_1, e_2, \cdots, e_I)$$
$$= \prod_i Pr(e_i|e_1, \cdots, e_{i-1}) \quad \text{chain rule}$$

statistical → $\approx P(e_i|e_1, \cdots, e_{i-n+1})$ estimated as relative frequency

→ n-gram based; if n-gram unseen, then back-off

language model

previously

$$\hat{e} = \text{argmax}_e\{Pr(e|f)\}$$
$$= \text{argmax}_e\{Pr(f|e) \cdot Pr(e)\}$$

let e_1^I be a sentence of a sequence of words $e_1, e_2 \cdots e_I$

$$Pr(e_1^I) = Pr(e_1, e_2, \cdots, e_I)$$
$$= \prod_i Pr(e_i|e_1, \cdots, e_{i-1}) \quad \text{chain rule}$$

statistical → $\approx P(e_i|e_1, \cdots, e_{i-n+1})$ estimated as relative frequency

→ n-gram based; if n-gram unseen, then back-off

long-term memory for language models

long-term memory for language models

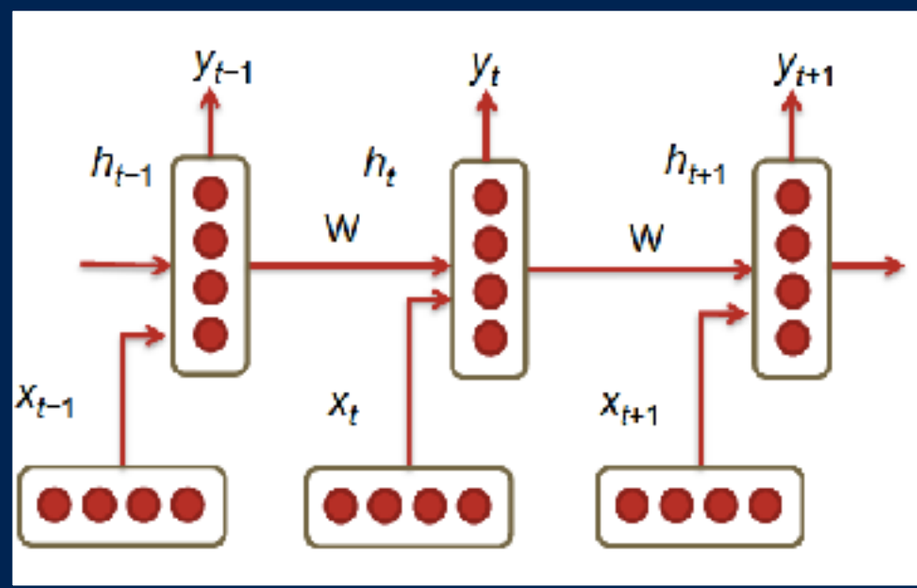
- n-gram LM making prediction on fixed windows
- past n words my not be sufficient to capture the context

long-term memory for language models

- n-gram LM making prediction on fixed windows
- past n words my not be sufficient to capture the context

RNNs are capable of conditioning the model on all previous words

neural language model with RNN



$$h_t = \sigma(W^{(hh)}h_{t-1} + W^{(hx)}x_{[t]})$$

$$\hat{y}_t = softmax(W^{(S)}h_t)$$

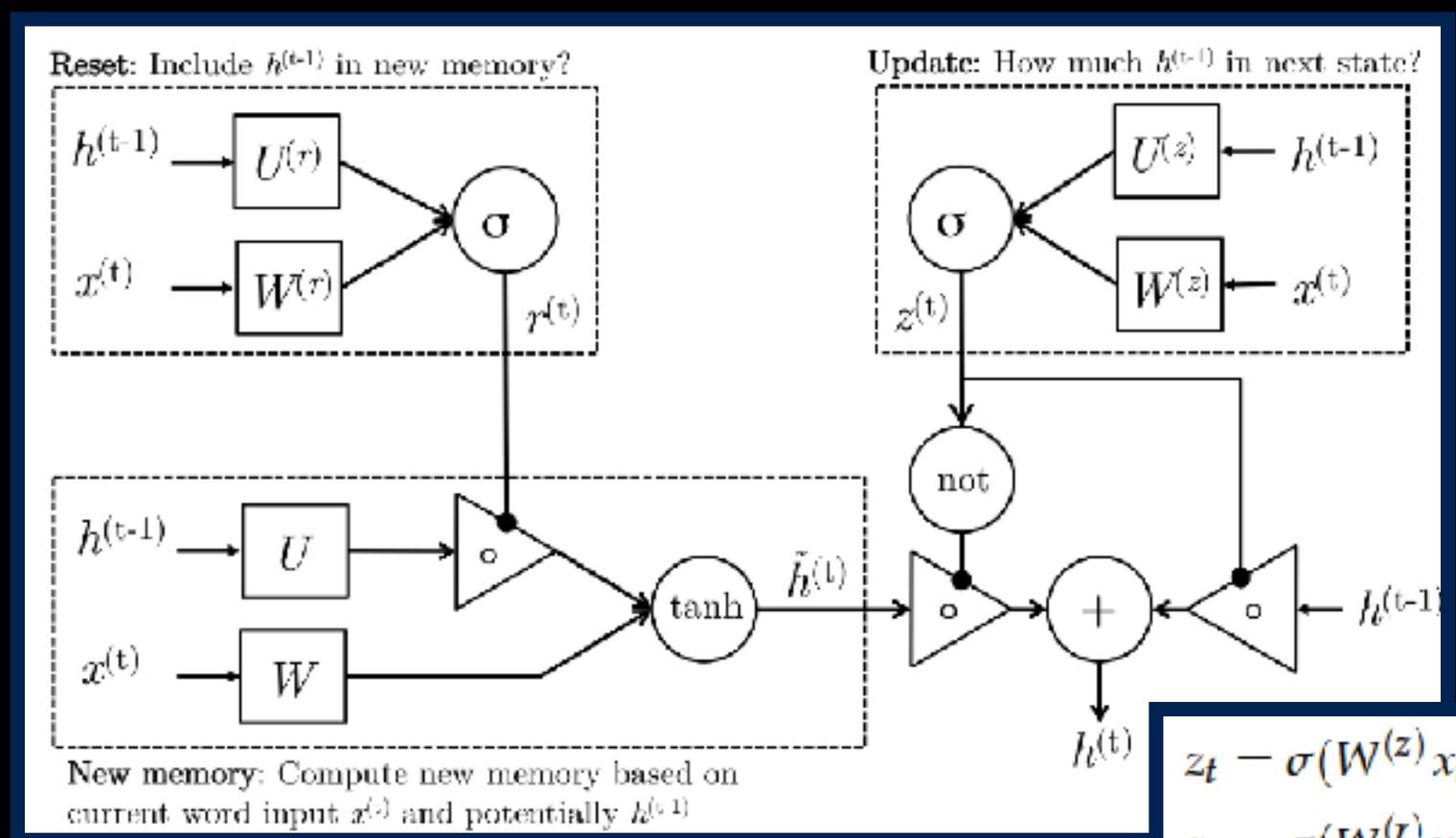
$$J = -\frac{1}{T}\sum_{t=1}^{T} J^{(t)}(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{|V|} y_{t,j} \times log(\hat{y}_{t,j})$$

- x_t : input word vector at time t
- W: weights matrix to condition t
- h_{t-1}: output of the non-linear function at the previous time step
- σ : the non-linearity function

Gated Recurrent Units (GRU)

- problem of vanishing gradients makes RNNs hard to train for long-term dependency

- use more complex units for activation



$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1}) \quad \text{(Update gate)}$$

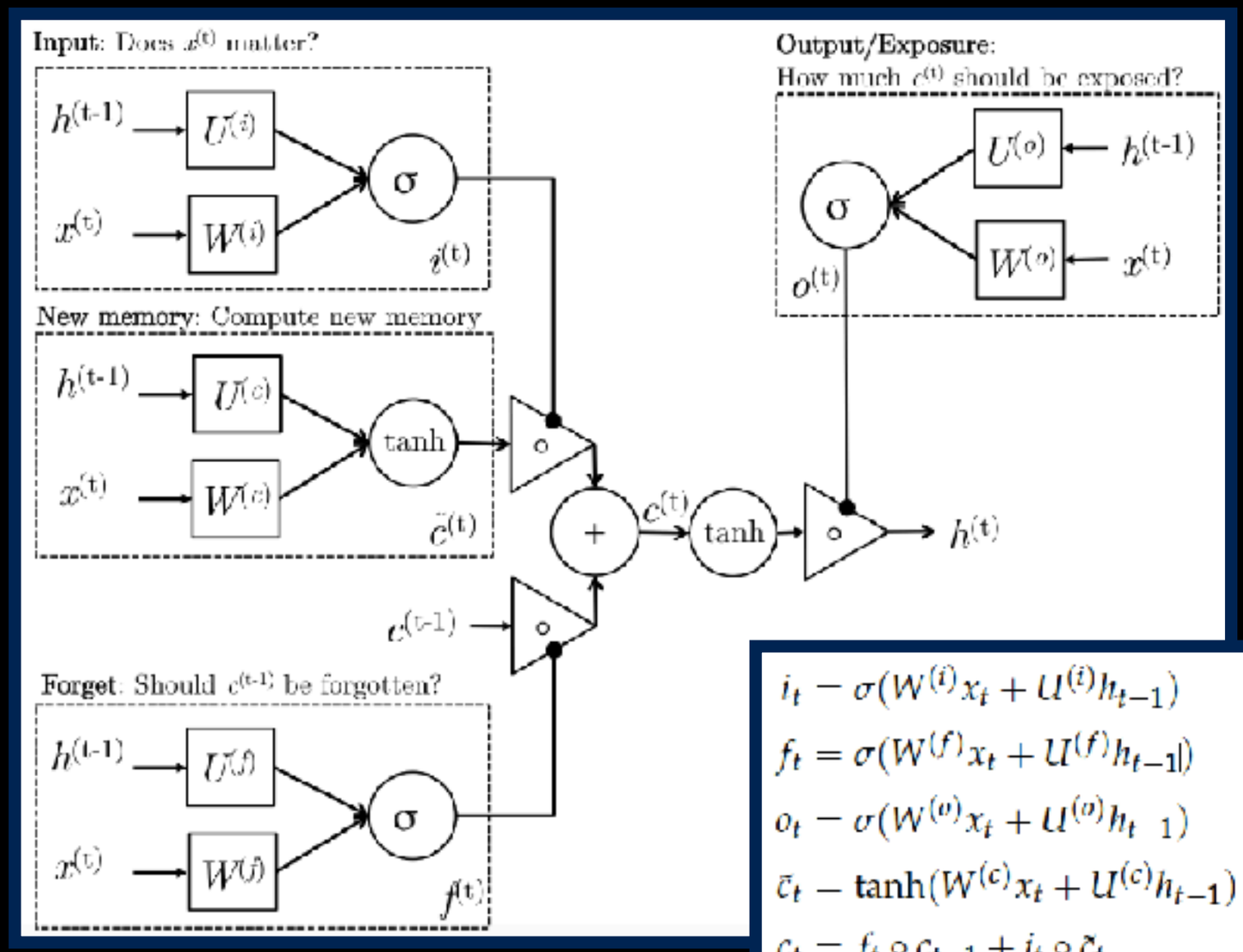
$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1}) \quad \text{(Reset gate)}$$

$$\tilde{h}_t = \tanh(r_t \circ Uh_{t-1} + Wx_t) \quad \text{(New memory)}$$

$$h_t = (1 - z_t) \circ \tilde{h}_t + z_t \circ h_{t-1} \quad \text{(Hidden state)}$$

Long-Short-Term-Memories

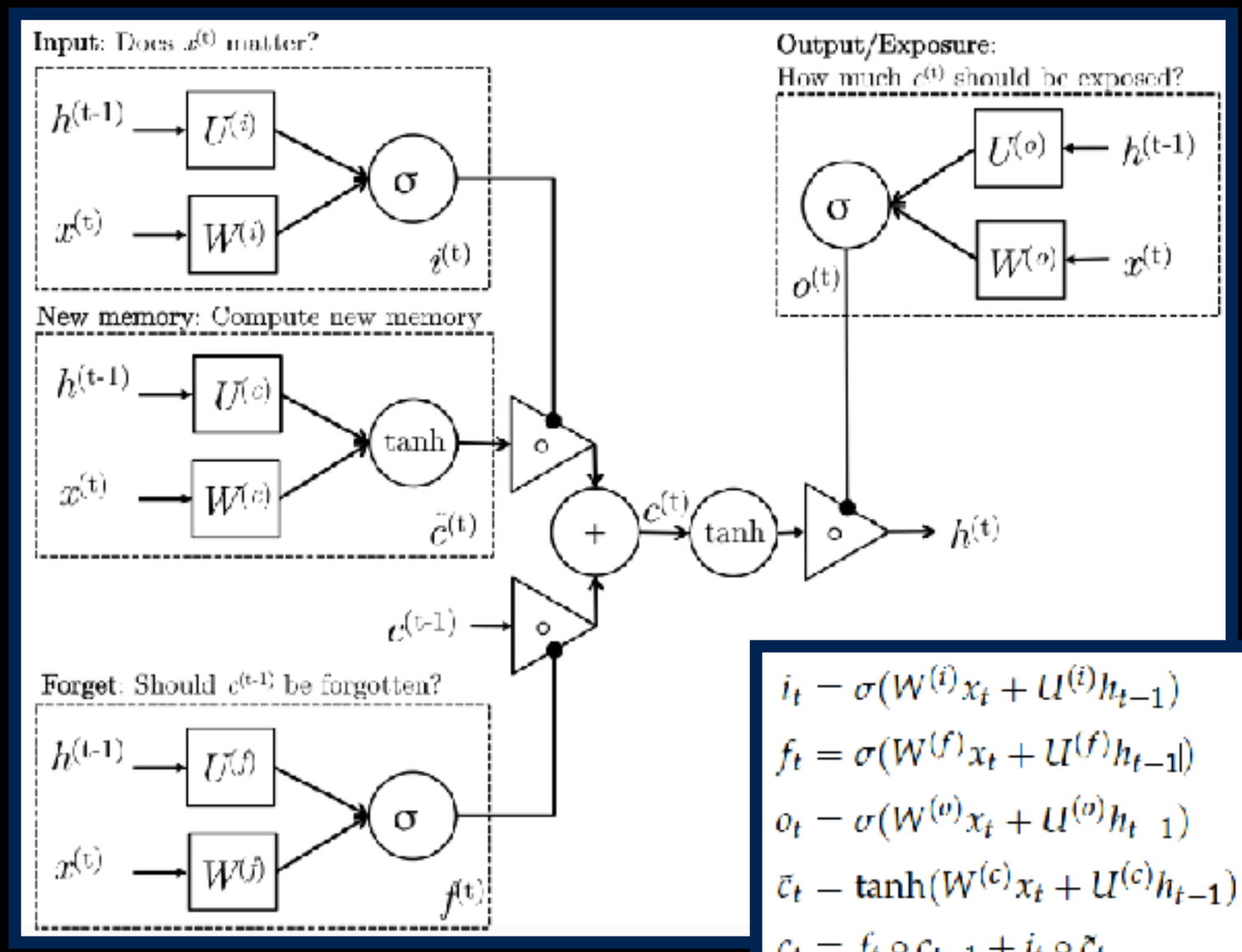
another type of complex activation unit



$$
\begin{aligned}
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1}) && \text{(Input gate)} \\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1}) && \text{(Forget gate)} \\
o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{t-1}) && \text{(Output/Exposure gate)} \\
\tilde{c}_t &= \tanh(W^{(c)}x_t + U^{(c)}h_{t-1}) && \text{(New memory cell)} \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t && \text{(Final memory cell)} \\
h_t &= o_t \circ \tanh(c_t)
\end{aligned}
$$

Long-Short-Term-Memories

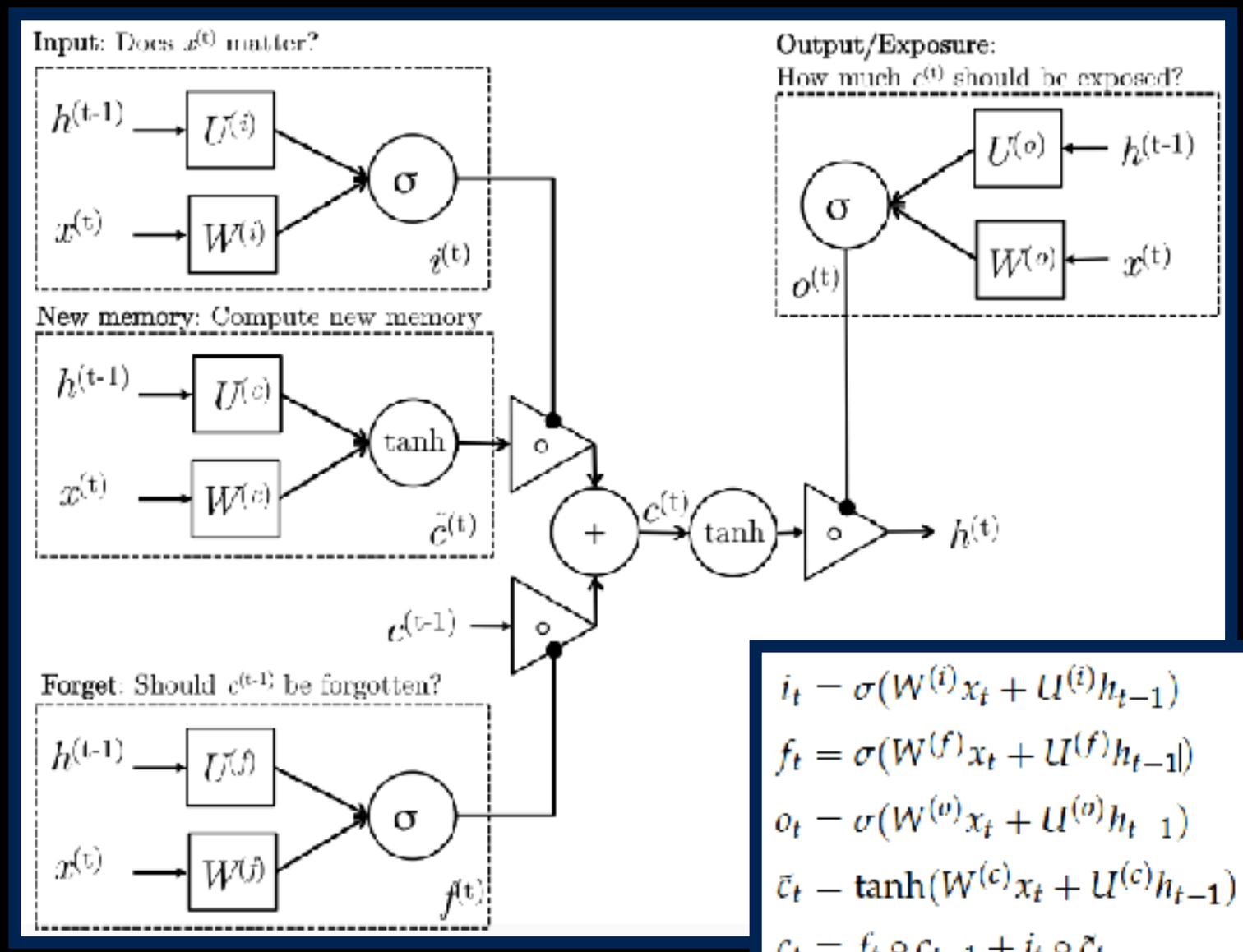
Question #4: contextual memory in language model



$$
\begin{aligned}
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1}) && \text{(Input gate)} \\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1}) && \text{(Forget gate)} \\
o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{t-1}) && \text{(Output/Exposure gate)} \\
\tilde{c}_t &= \tanh(W^{(c)}x_t + U^{(c)}h_{t-1}) && \text{(New memory cell)} \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t && \text{(Final memory cell)} \\
h_t &= o_t \circ \tanh(c_t)
\end{aligned}
$$

Long-Short-Term-Memories

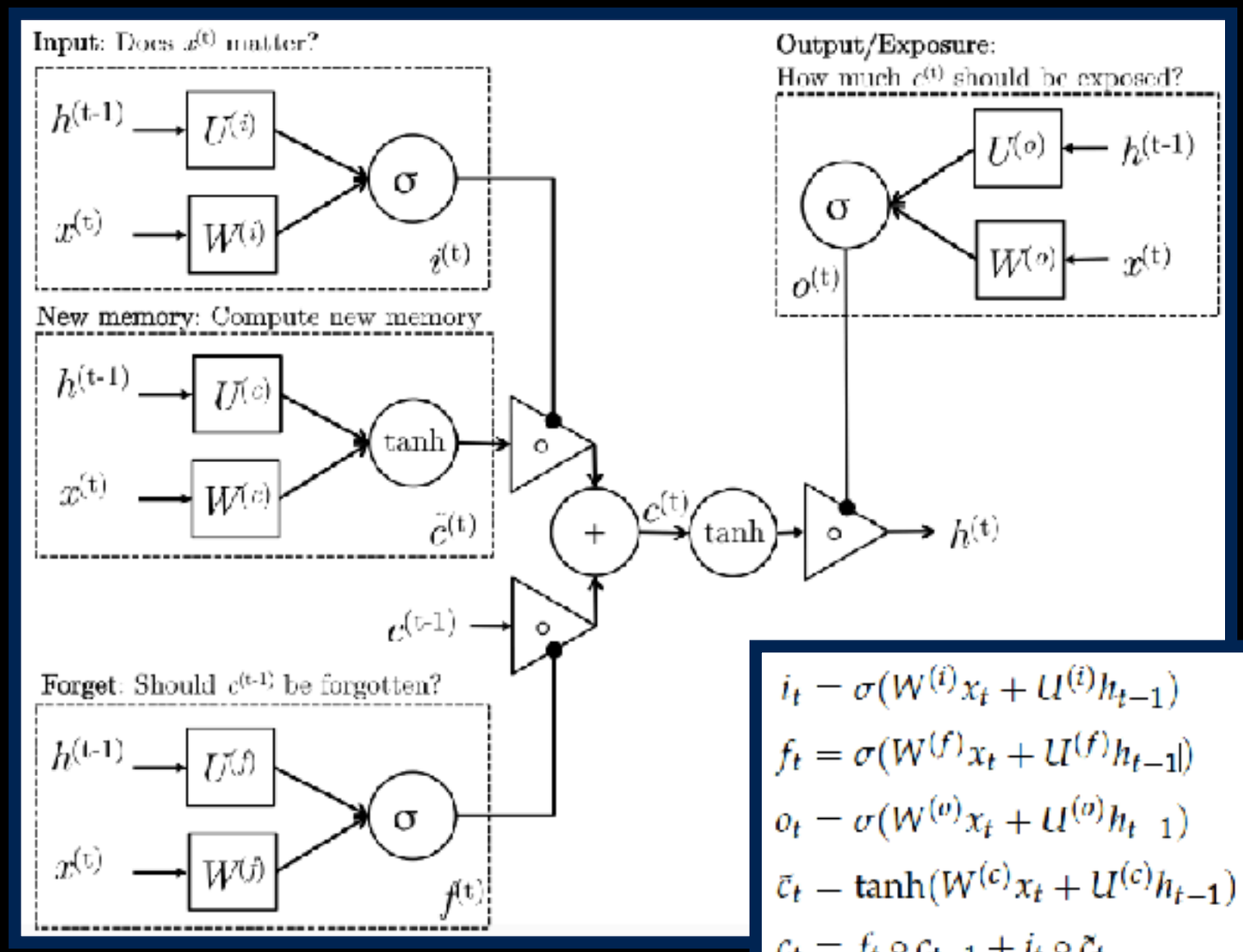
another type of complex activation unit



$$
\begin{aligned}
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1}) && \text{(Input gate)} \\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1}) && \text{(Forget gate)} \\
o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{t-1}) && \text{(Output/Exposure gate)} \\
\tilde{c}_t &= \tanh(W^{(c)}x_t + U^{(c)}h_{t-1}) && \text{(New memory cell)} \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t && \text{(Final memory cell)} \\
h_t &= o_t \circ \tanh(c_t)
\end{aligned}
$$

Long-Short-Term-Memories

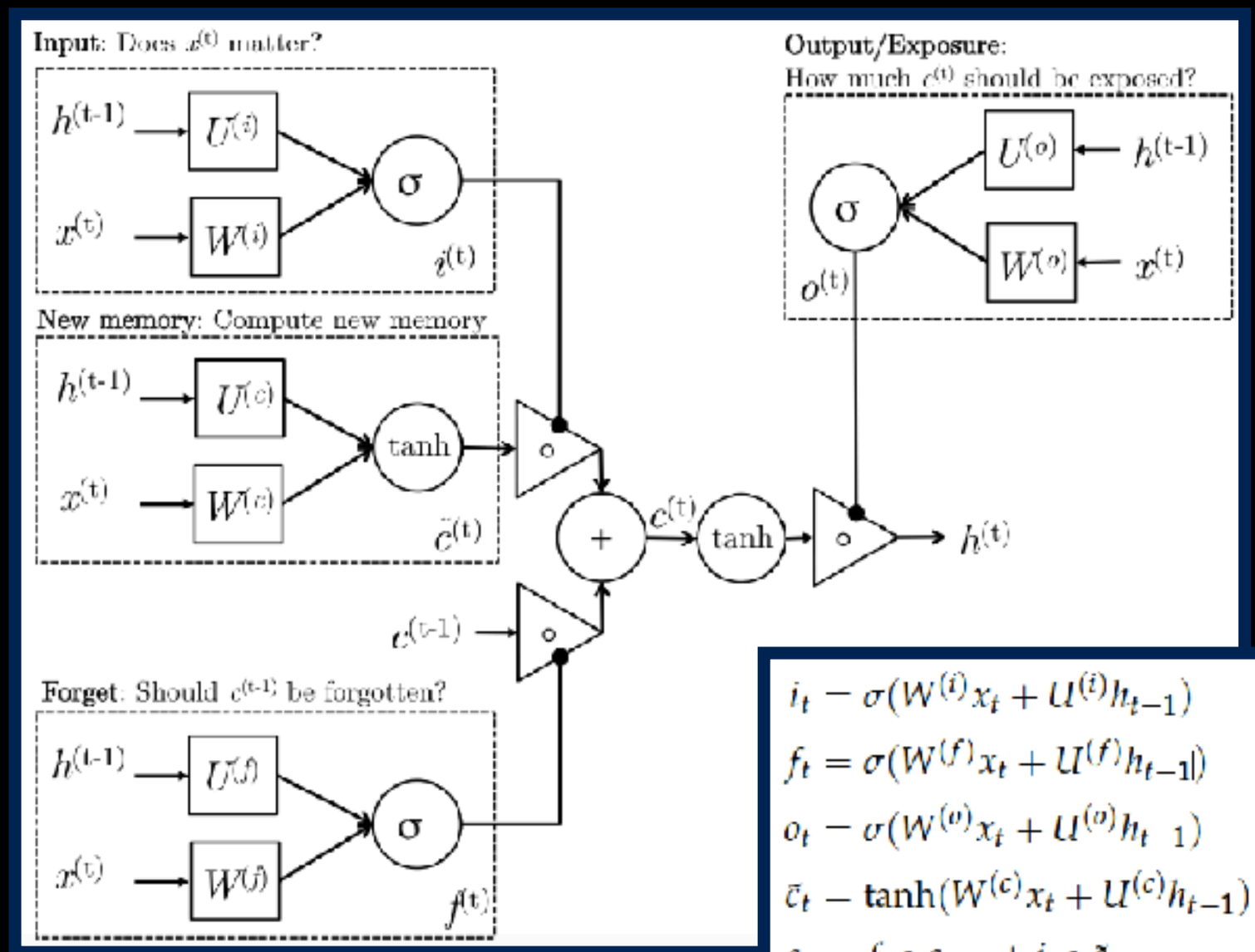
another type of complex activation unit



$$
\begin{aligned}
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1}) && \text{(Input gate)} \\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1}) && \text{(Forget gate)} \\
o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{t-1}) && \text{(Output/Exposure gate)} \\
\tilde{c}_t &= \tanh(W^{(c)}x_t + U^{(c)}h_{t-1}) && \text{(New memory cell)} \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t && \text{(Final memory cell)} \\
h_t &= o_t \circ \tanh(c_t)
\end{aligned}
$$

Long-Short-Term-Memories

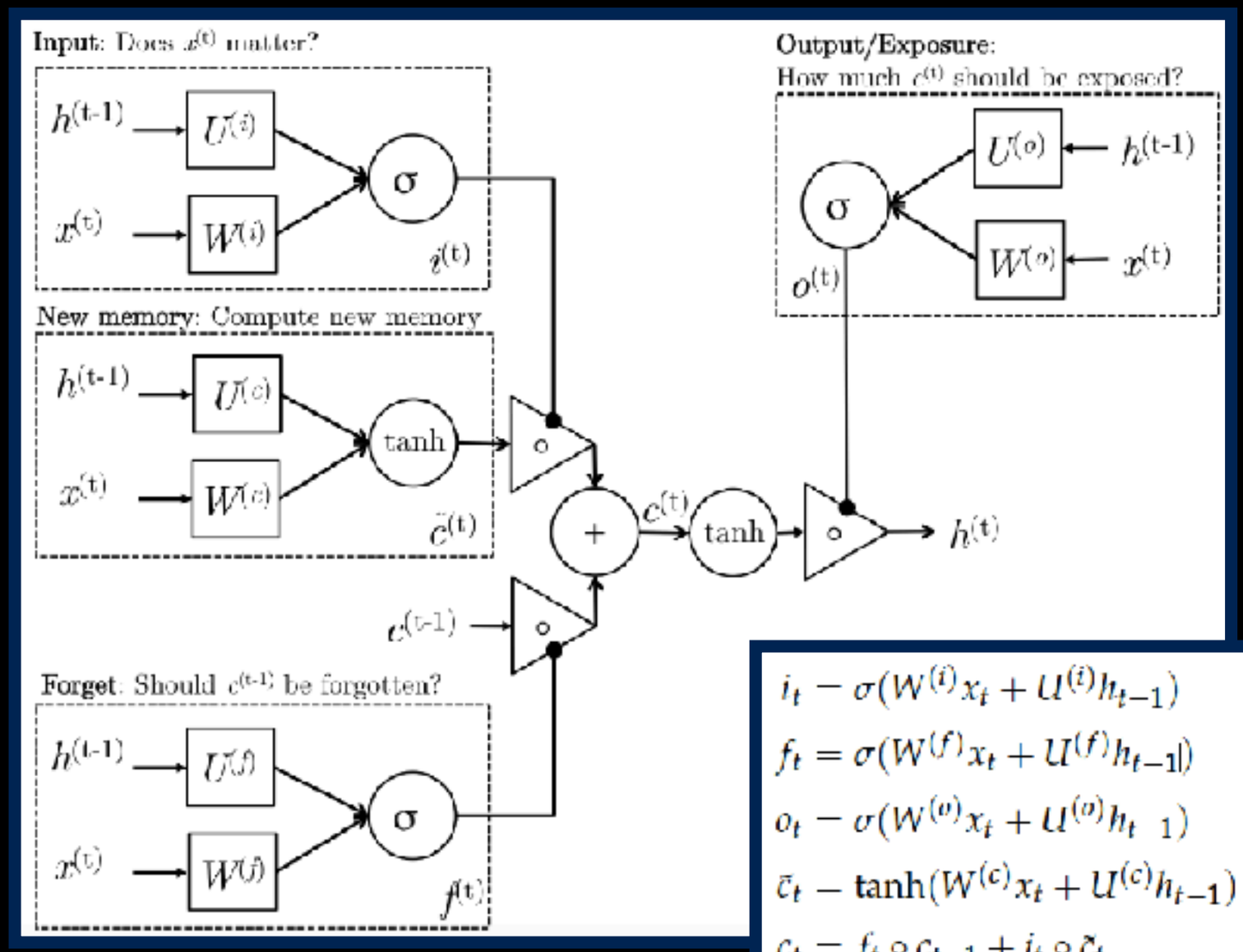
Question #5: affective neuron activation function



$$
\begin{aligned}
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1}) && \text{(Input gate)} \\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1}) && \text{(Forget gate)} \\
o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{t-1}) && \text{(Output/Exposure gate)} \\
\tilde{c}_t &= \tanh(W^{(c)}x_t + U^{(c)}h_{t-1}) && \text{(New memory cell)} \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t && \text{(Final memory cell)} \\
h_t &= o_t \circ \tanh(c_t)
\end{aligned}
$$

Long-Short-Term-Memories

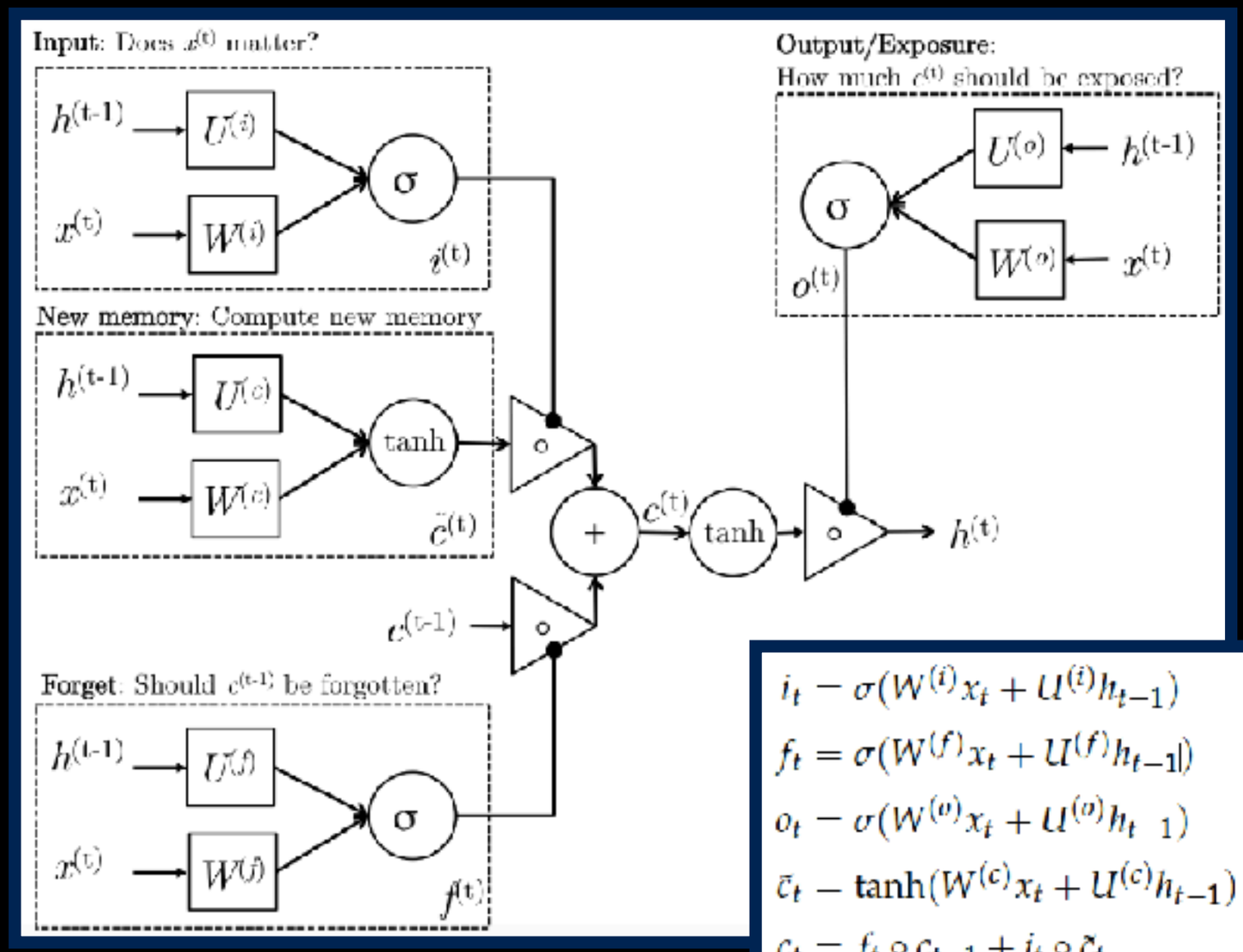
another type of complex activation unit



$$
\begin{aligned}
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1}) && \text{(Input gate)} \\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1}) && \text{(Forget gate)} \\
o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{t-1}) && \text{(Output/Exposure gate)} \\
\tilde{c}_t &= \tanh(W^{(c)}x_t + U^{(c)}h_{t-1}) && \text{(New memory cell)} \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t && \text{(Final memory cell)} \\
h_t &= o_t \circ \tanh(c_t)
\end{aligned}
$$

Long-Short-Term-Memories

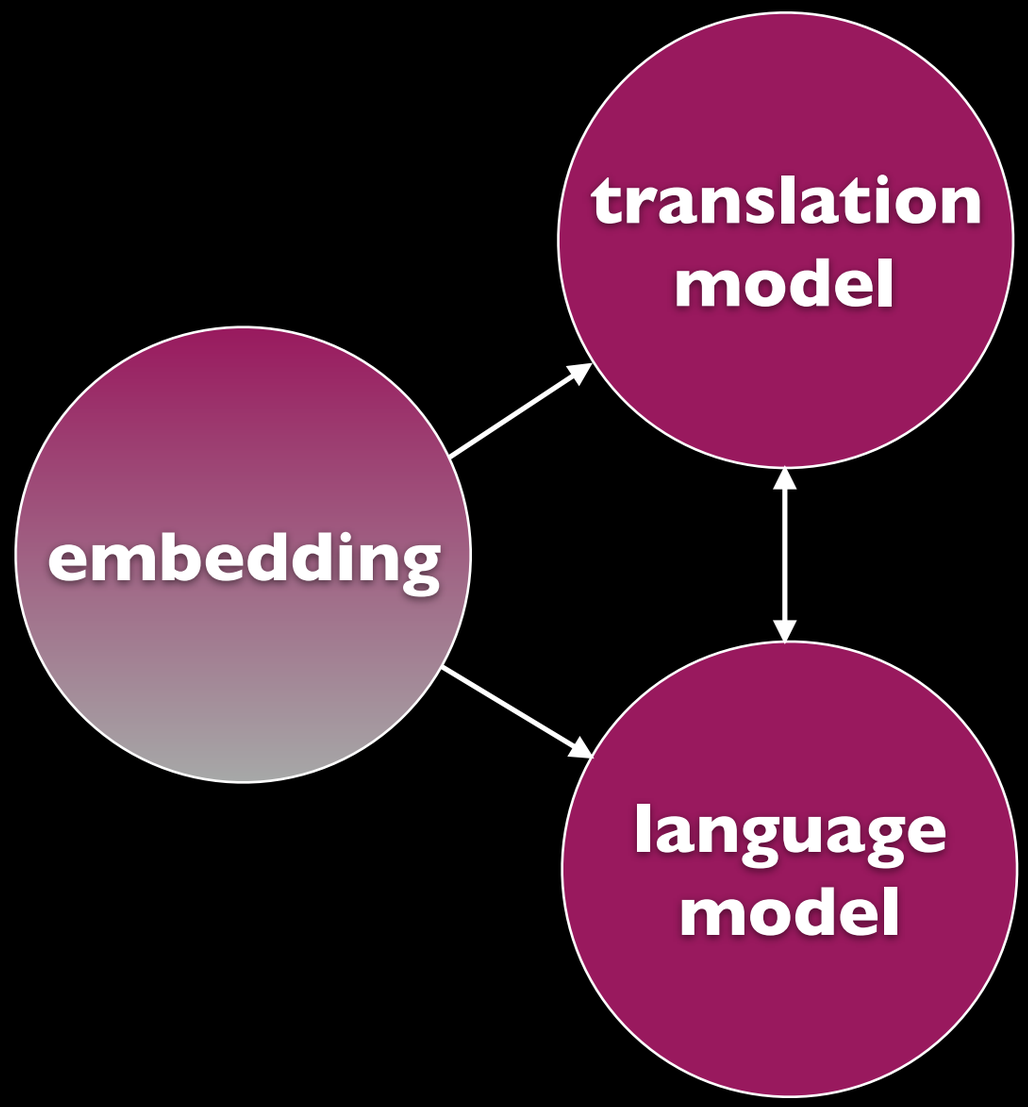
another type of complex activation unit



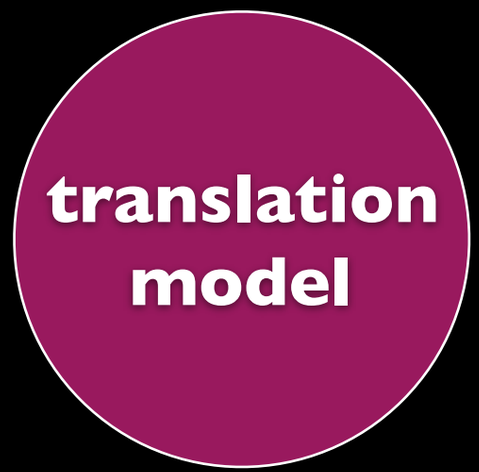
$$
\begin{aligned}
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1}) && \text{(Input gate)} \\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1}) && \text{(Forget gate)} \\
o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{t-1}) && \text{(Output/Exposure gate)} \\
\tilde{c}_t &= \tanh(W^{(c)}x_t + U^{(c)}h_{t-1}) && \text{(New memory cell)} \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t && \text{(Final memory cell)} \\
h_t &= o_t \circ \tanh(c_t)
\end{aligned}
$$

machine translation components

machine translation components

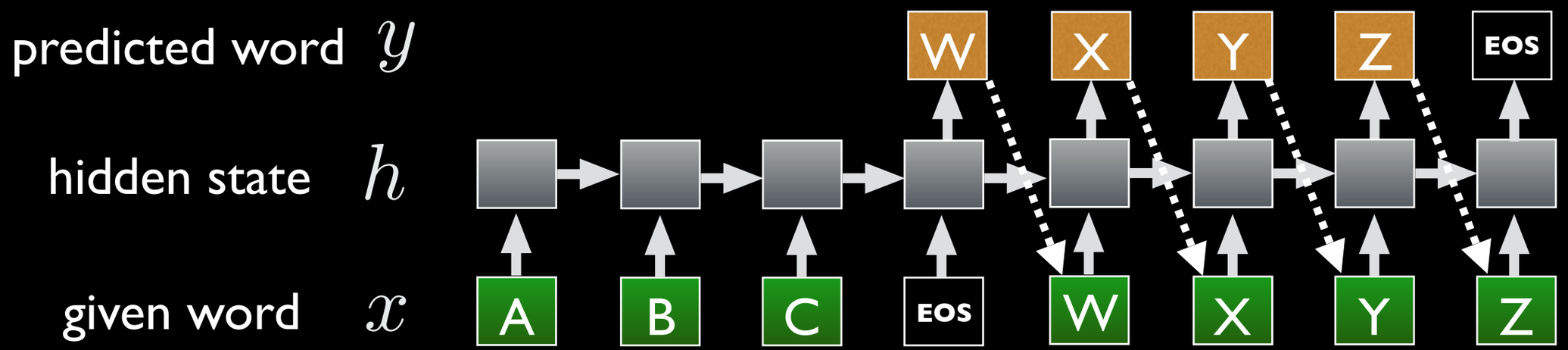


machine translation components



NMT I: sequence-to-sequence with RNN

[Sutskever, 93]



$$P(y_1, \cdots, y_{T'} | x_1, \cdots, x_T) = \prod_{t=1}^{T'} P(y_t | x, y_1 \cdots, y_{t-1})$$

$$h_t = \mathrm{sigm}(W^{hx} x_t + W^{hh} h_{t-1})$$

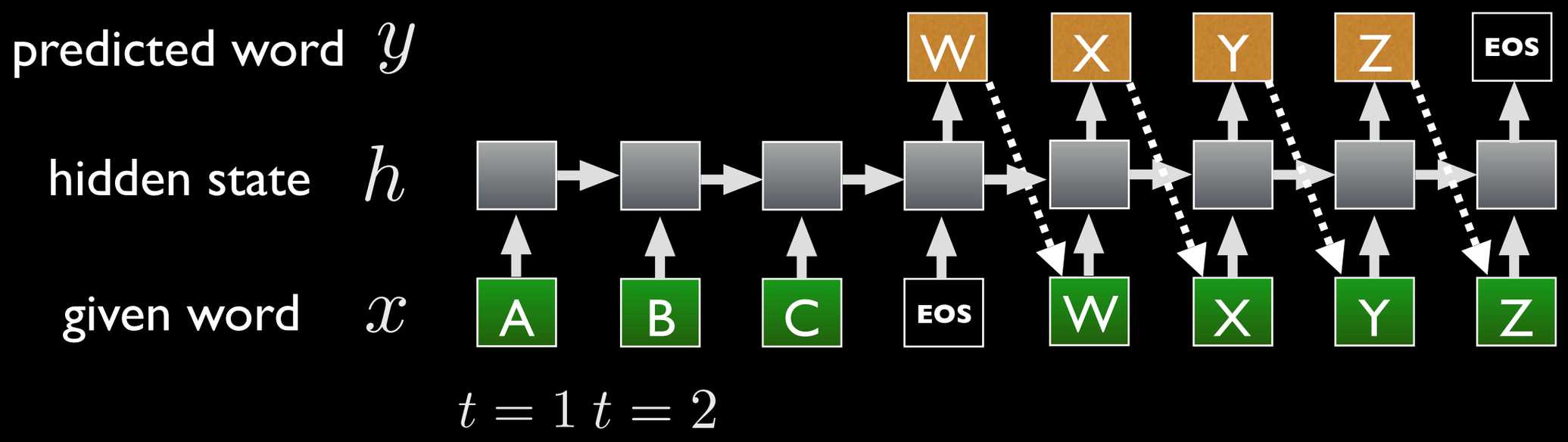
$$y_t = W^{yh} h_t$$

words are in embedded representation

NMT I: training sequence-to-sequence

[Sutskever, 93]

$$\frac{1}{4}\left[-P(\text{“W”})-P(\text{“X”})-P(\text{“Y”})-P(\text{“Z”})\right]$$

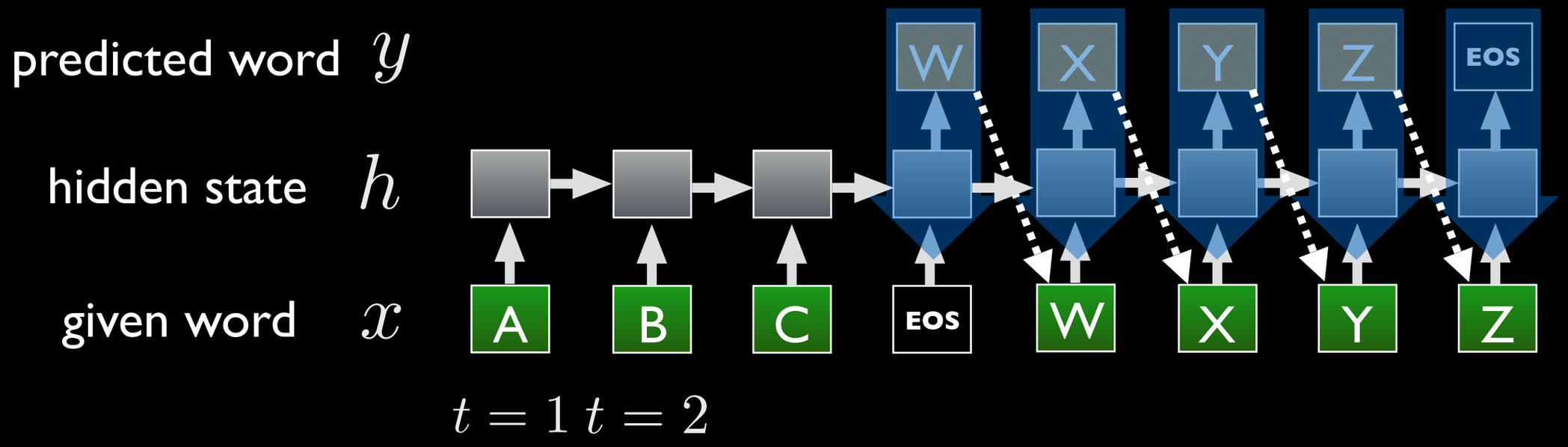


maximize the log probability of a correct translation given the source sentence

NMT I: training sequence-to-sequence

[Sutskever, 93]

$$\frac{1}{4}\left[-P(\text{“W”})-P(\text{“X”})-P(\text{“Y”})-P(\text{“Z”})\right]$$

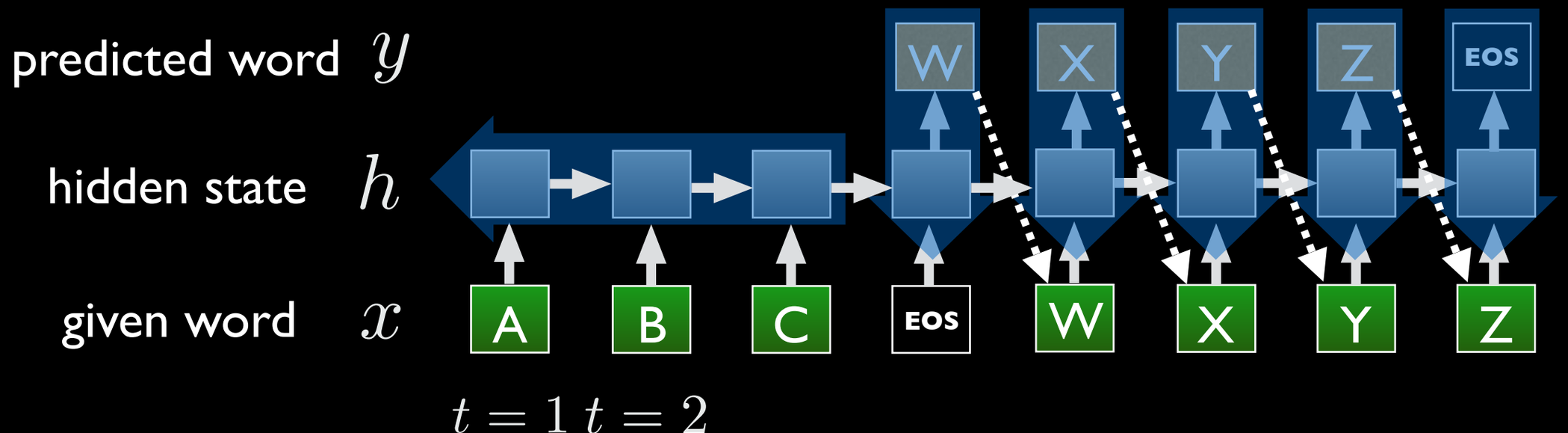


maximize the log probability of a correct translation given the source sentence

NMT I: training sequence-to-sequence

[Sutskever, 93]

$$\frac{1}{4}\left[-P(\text{“W”})-P(\text{“X”})-P(\text{“Y”})-P(\text{“Z”})\right]$$

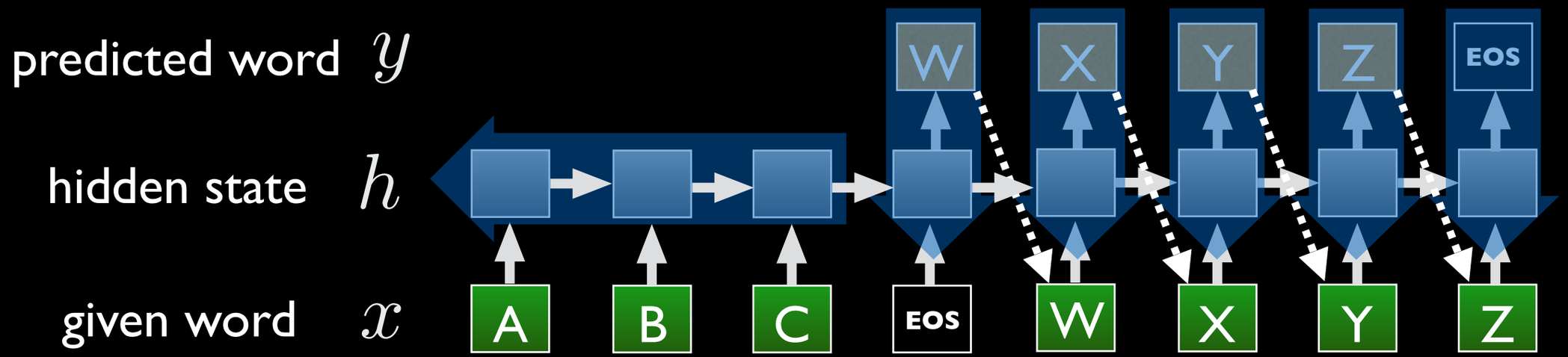


maximize the log probability of a correct translation given the source sentence

NMT I: training sequence-to-sequence

[Sutskever, 93]

$\frac{1}{4}$ [$-P(\text{“W”})-P(\text{“X”})-P(\text{“Y”})-P(\text{“Z”})$]



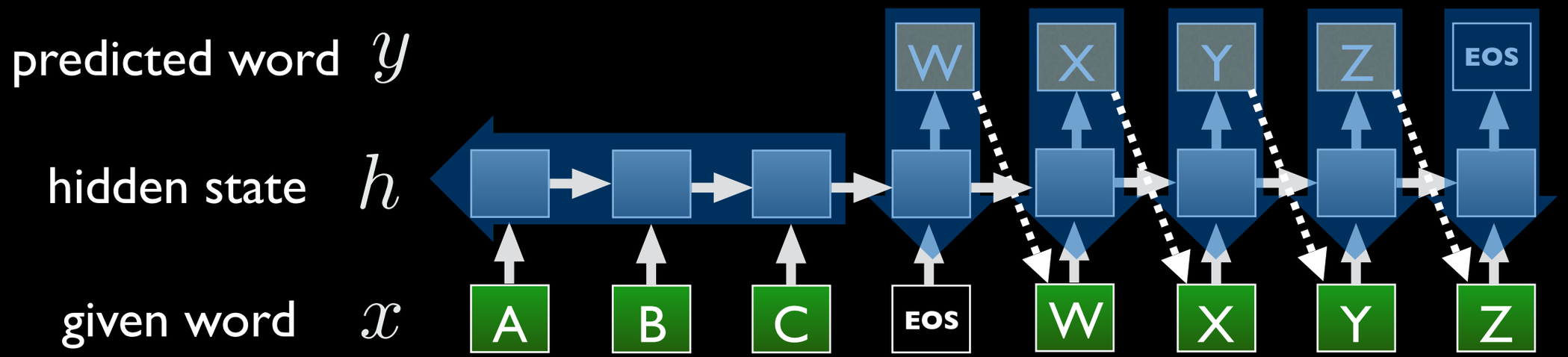
back propagation operates “end-to-end”

maximize the log probability of a correct translation given the source sentence

NMT I: training sequence-to-sequence

[Sutskever, 93]

$\frac{1}{4}$ [$-P(\text{“W”})-P(\text{“X”})-P(\text{“Y”})-P(\text{“Z”})$]



back propagation operates “end-to-end”

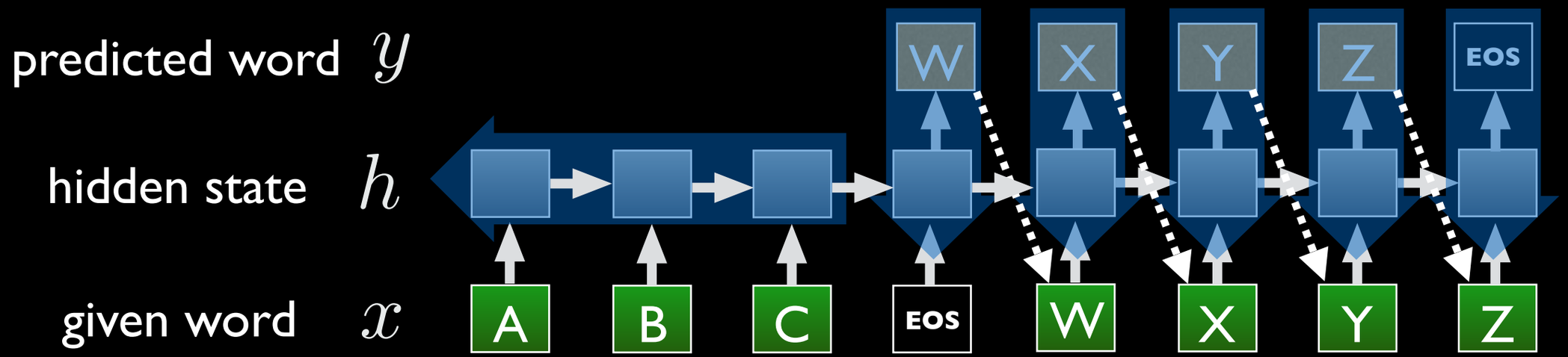
maximize the log probability of a correct translation given the source sentence

Question #6: better training criterion?
Maximum Likelihood, squared error, MAP, cross-entropy, minimum risk, ..

NMT I: training sequence-to-sequence

[Sutskever, 93]

$\frac{1}{4}$ [$-P(\text{“W”})-P(\text{“X”})-P(\text{“Y”})-P(\text{“Z”})$]



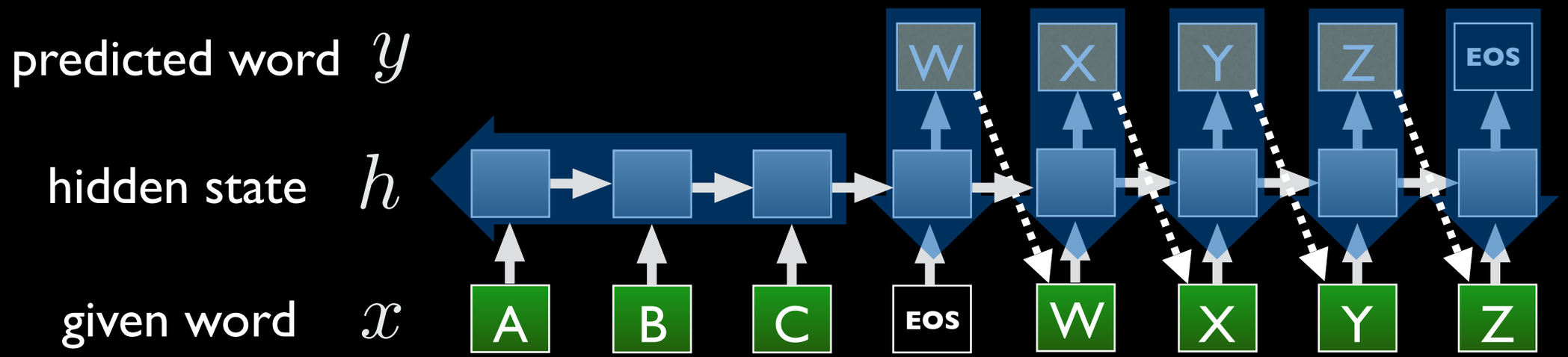
back propagation operates “end-to-end”

maximize the log probability of a correct translation given the source sentence

NMT I: training sequence-to-sequence

[Sutskever, 93]

$\frac{1}{4}$ [$-P(\text{“W”})-P(\text{“X”})-P(\text{“Y”})-P(\text{“Z”})$]



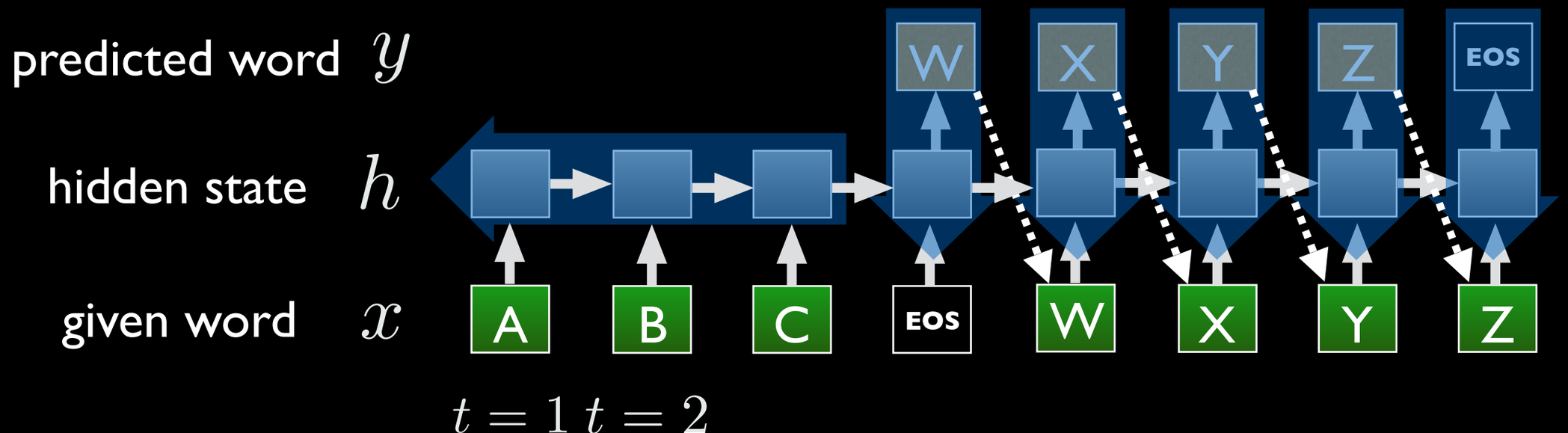
back propagation operates “end-to-end”

maximize the log probability of a correct translation given the source sentence

NMT I: training sequence-to-sequence

[Sutskever, 93]

$\frac{1}{4}$ [$-P(\text{“W”})-P(\text{“X”})-P(\text{“Y”})-P(\text{“Z”})$]



back propagation operates “end-to-end”

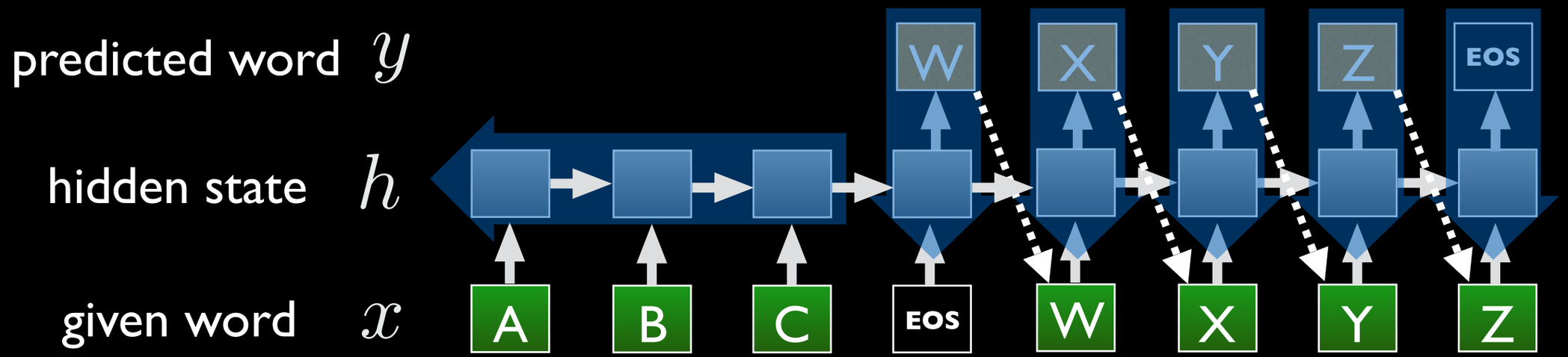
maximize the log probability of a correct translation given the source sentence

Question #7: better training algorithm?
error back propagation, contrastive estimation, …

NMT I: training sequence-to-sequence

[Sutskever, 93]

$\frac{1}{4}$ [$-P(\text{“W”})-P(\text{“X”})-P(\text{“Y”})-P(\text{“Z”})$]



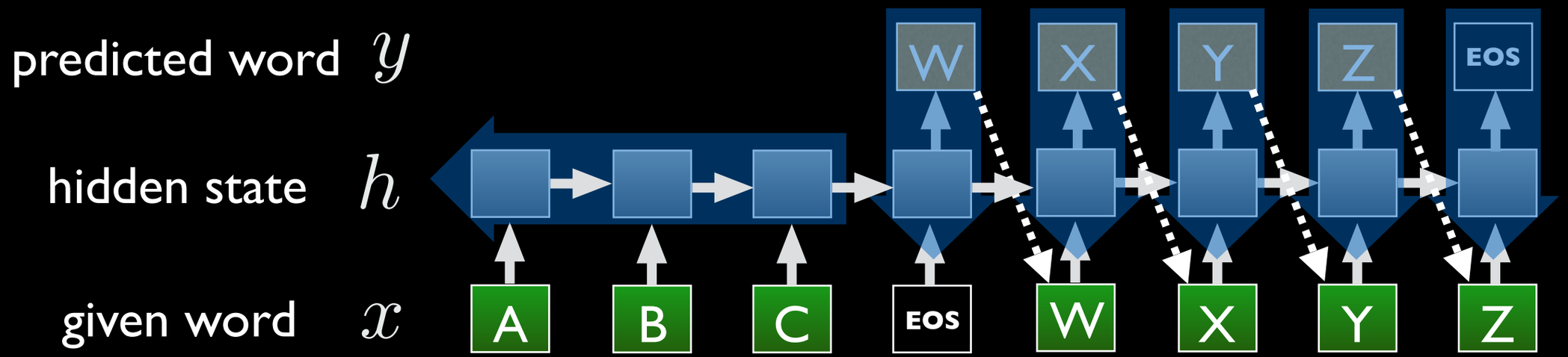
back propagation operates “end-to-end”

maximize the log probability of a correct translation given the source sentence

NMT I: training sequence-to-sequence

[Sutskever, 93]

$\frac{1}{4}$ [$-P(\text{“W”})-P(\text{“X”})-P(\text{“Y”})-P(\text{“Z”})$]

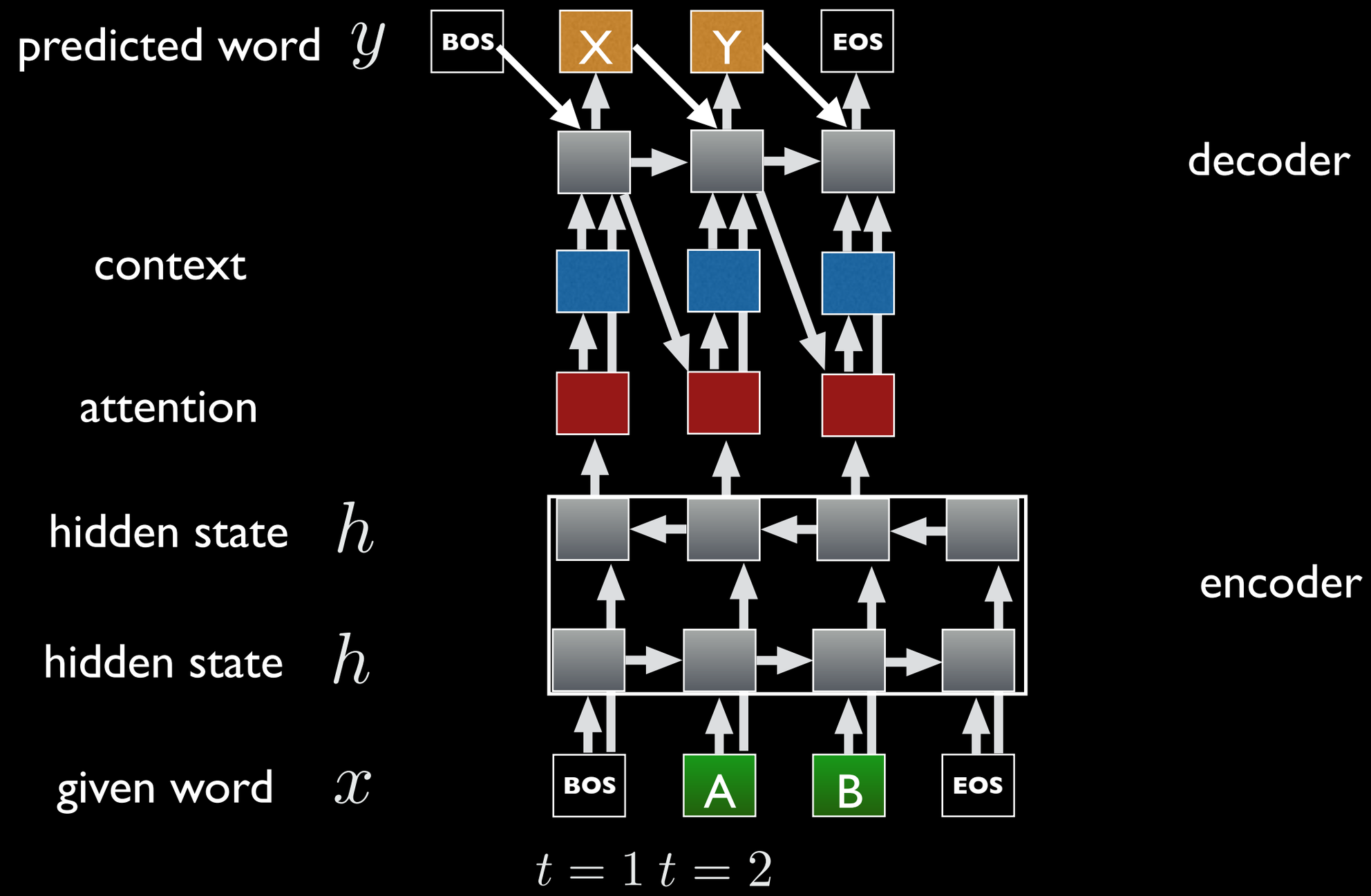


back propagation operates “end-to-end”

maximize the log probability of a correct translation given the source sentence

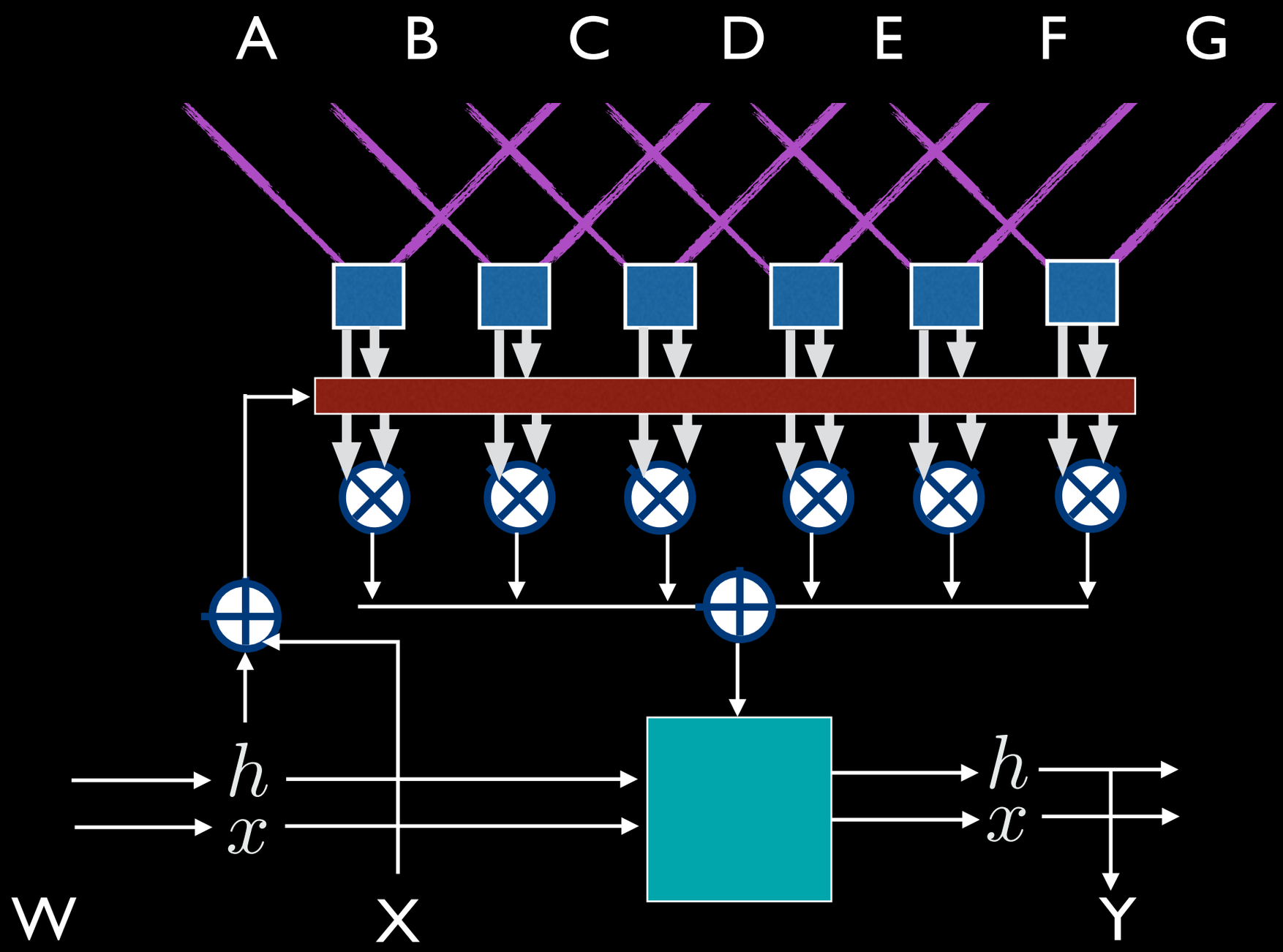
NMT II: encoder & decoder with attention

[Luong et.al., 15]



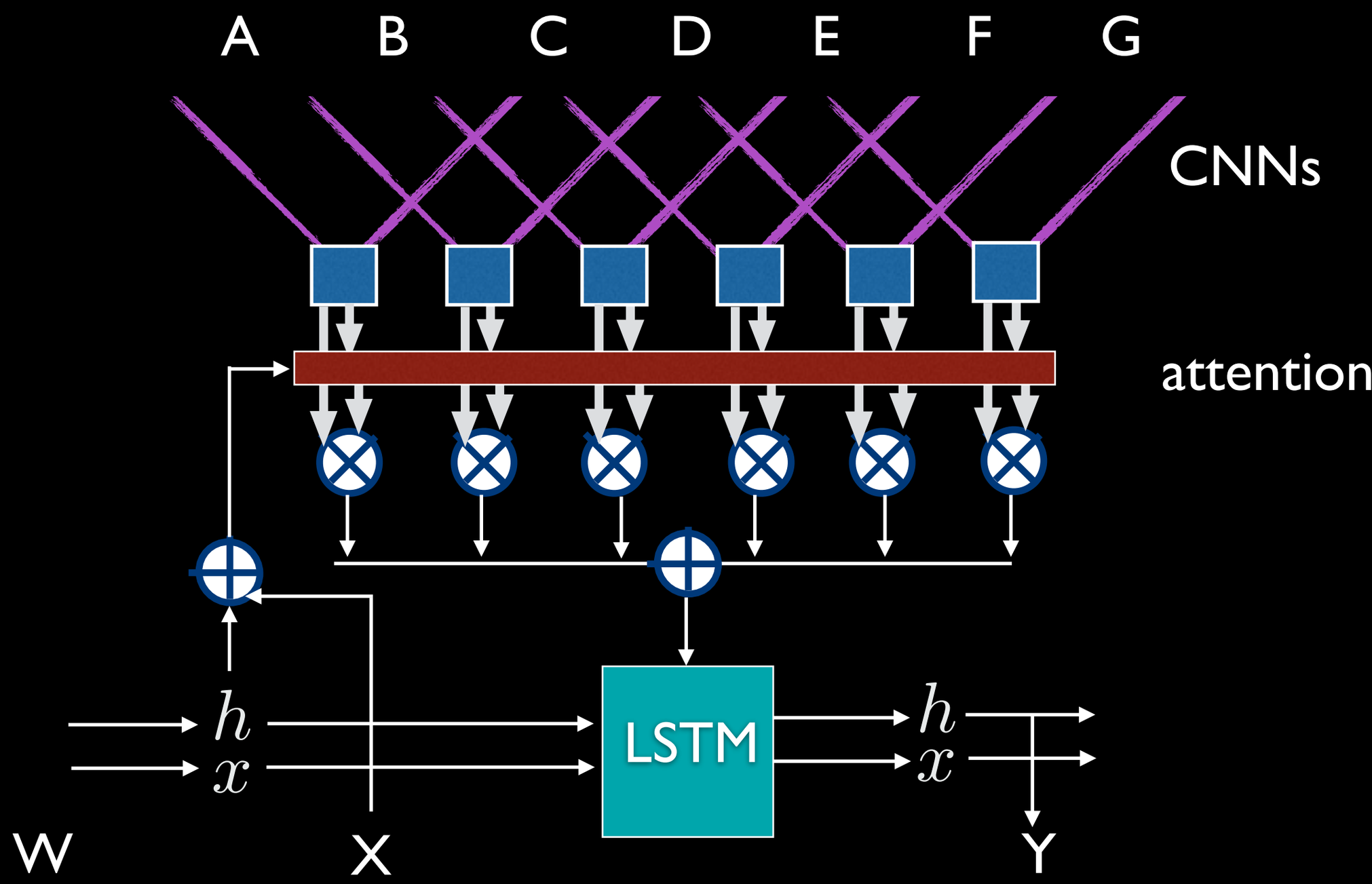
NMT III: multiple models with CNN

[Gehring, 16]

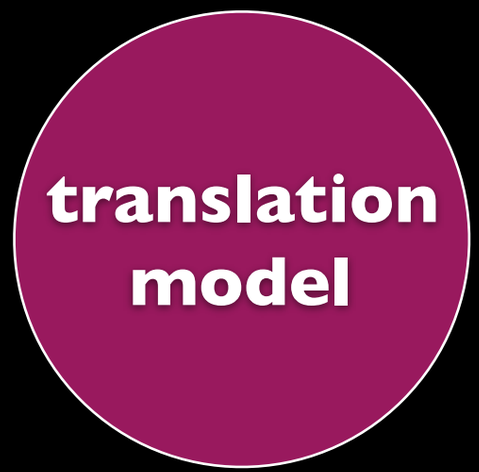


NMT III: multiple models with CNN

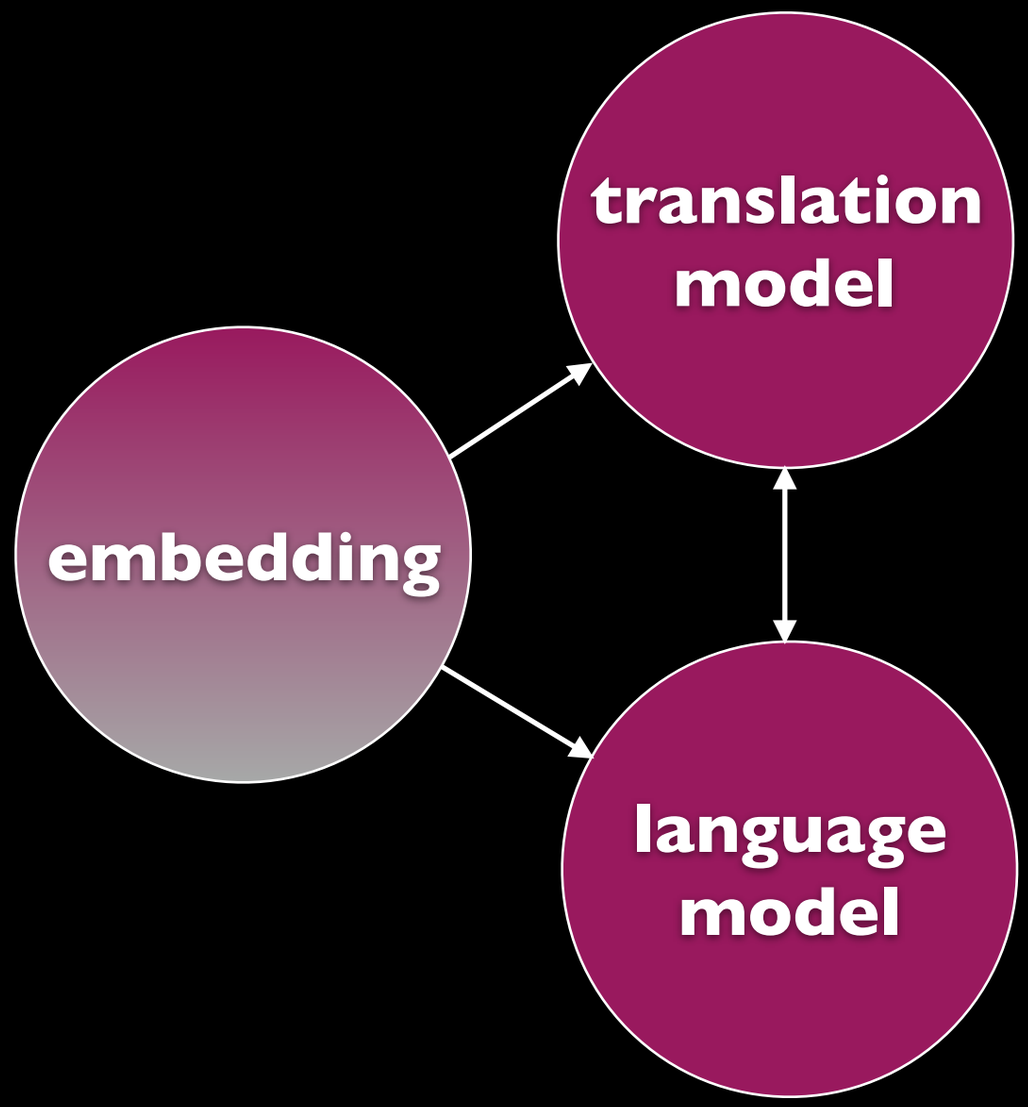
[Gehring, 16]



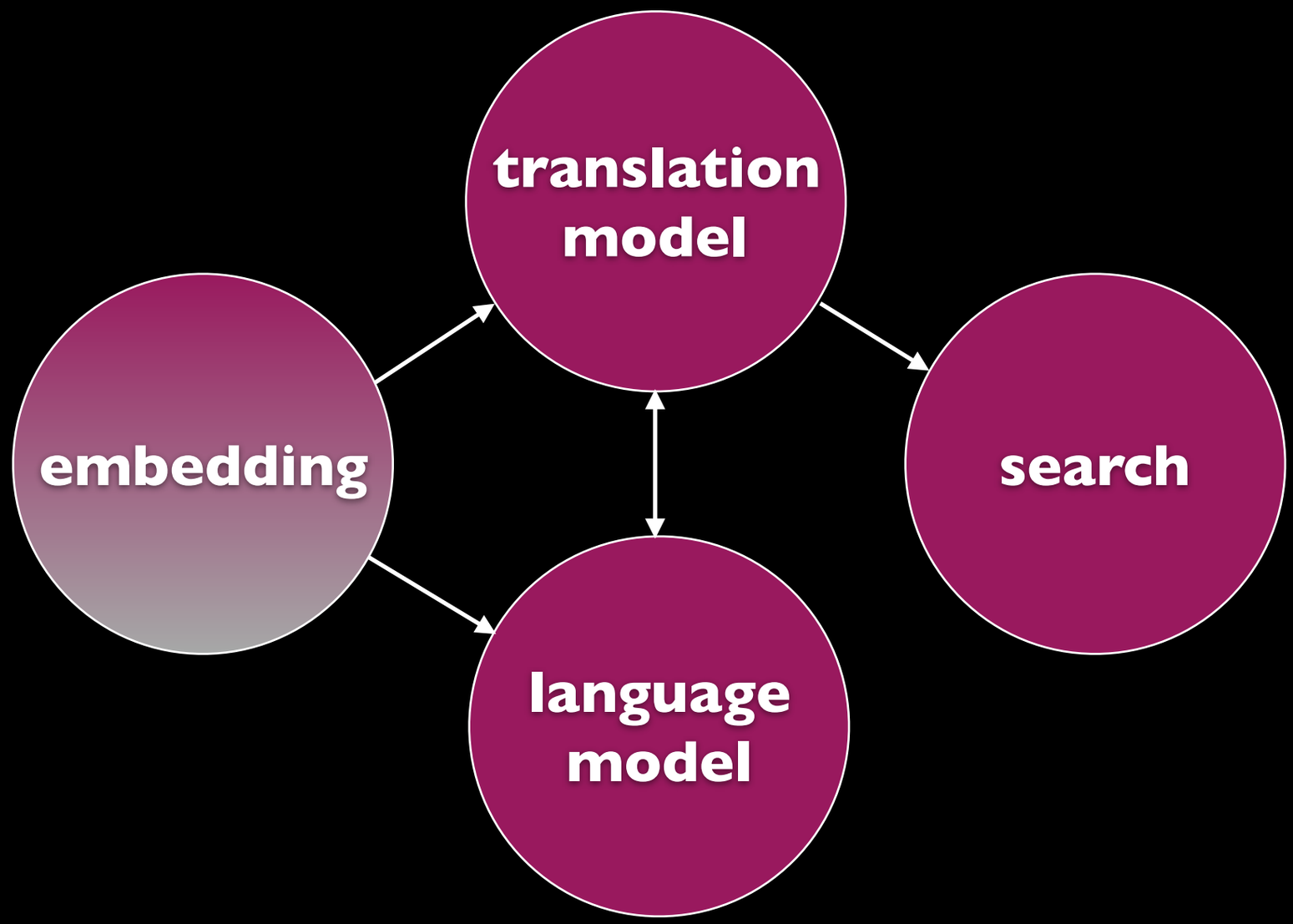
machine translation components



machine translation components

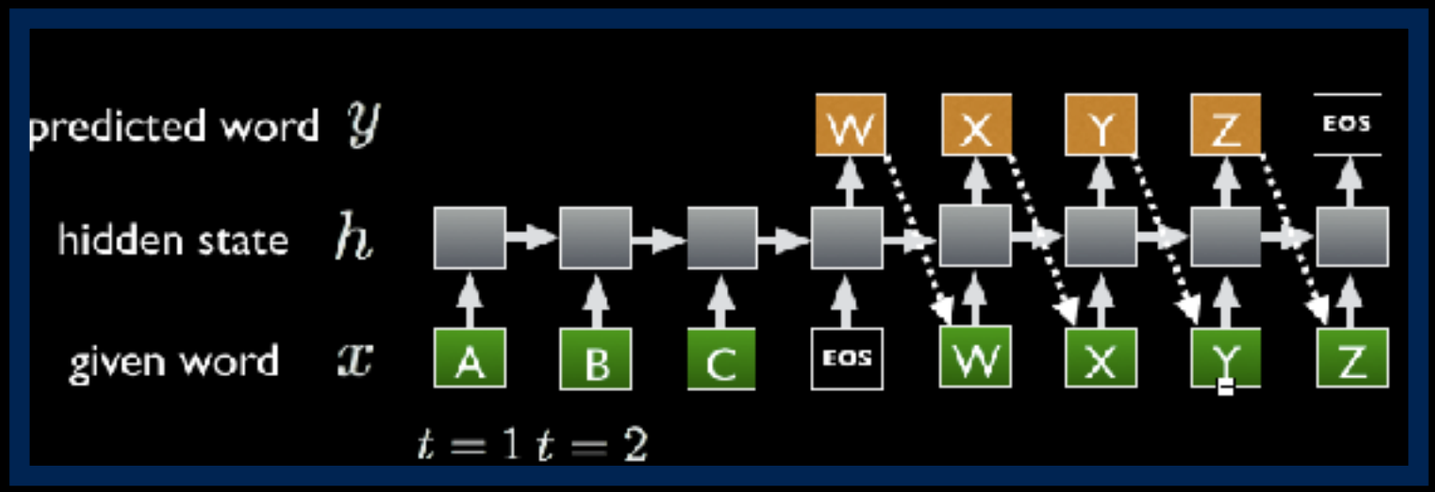


machine translation components



machine translation components

greedy search

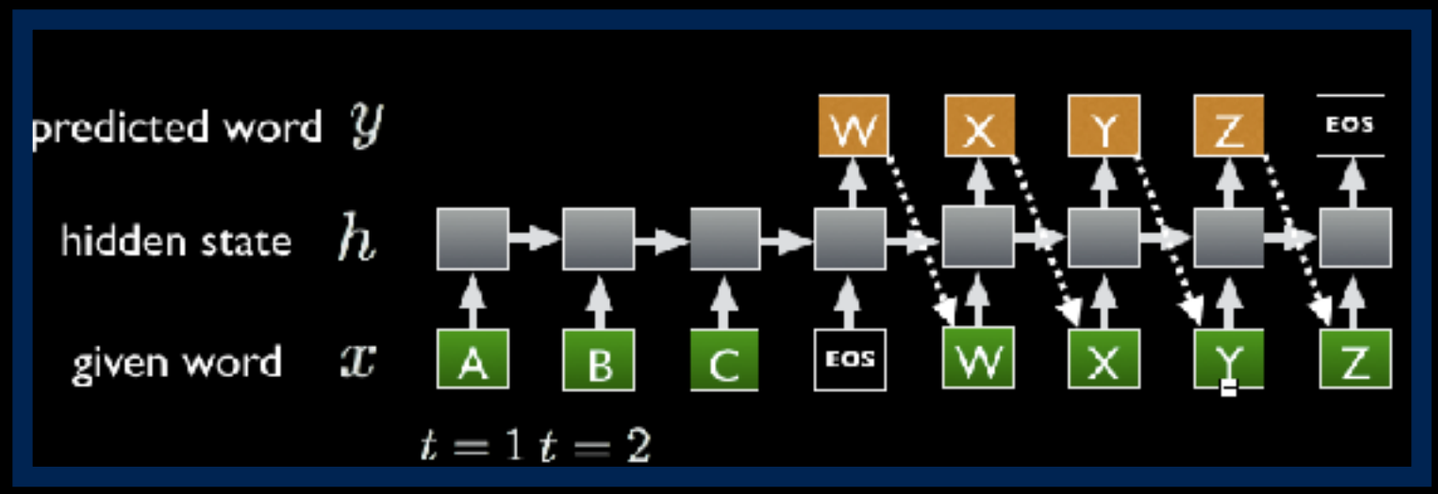


method: take most probable word in each step

problem: no way to undo decisions

- W ___
- W X ___
- W X Z ___ (no way back!)

exhaustive search



ideally: find a translation that maximize

$$P(y_1, \cdots, y_{T'} | x_1, \cdots, x_T) = \prod_{t=1}^{T'} P(y_t | x, y_1 \cdots, y_{t-1})$$

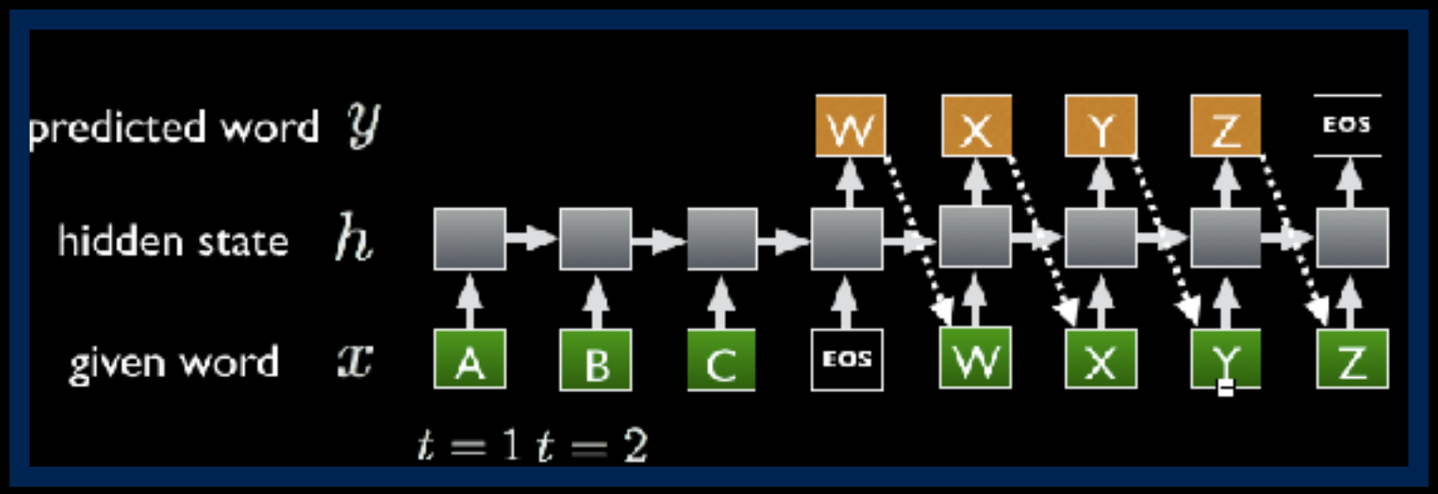
method: compute all possible sequences y

problem: expensive

each step tracking *V (vocabulary)* words

complexity $O(V^T)$

beam search



$$\text{score}(y_1, \cdots, y_t) = \log P_{LM}(y_1, \cdots, y_t | x)$$

$$= \sum_{i=1}^{t} \log P_{LM}(y_i | y_1 \cdots, y_{i-1}, x)$$

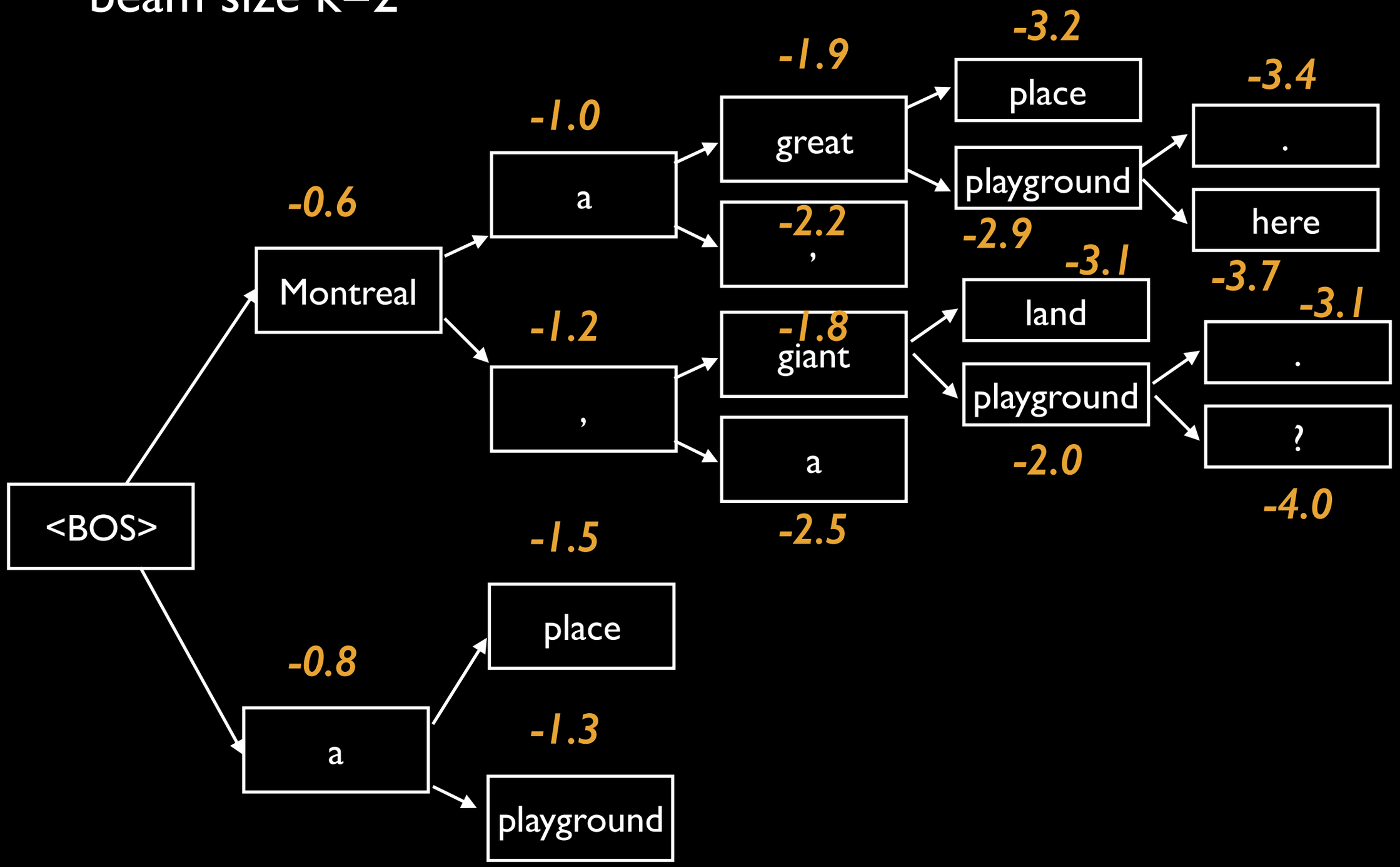
method: on each search step, keep track of the k most probable (higher score) partial translations

problem: no guarantee for optimal solution

efficient!

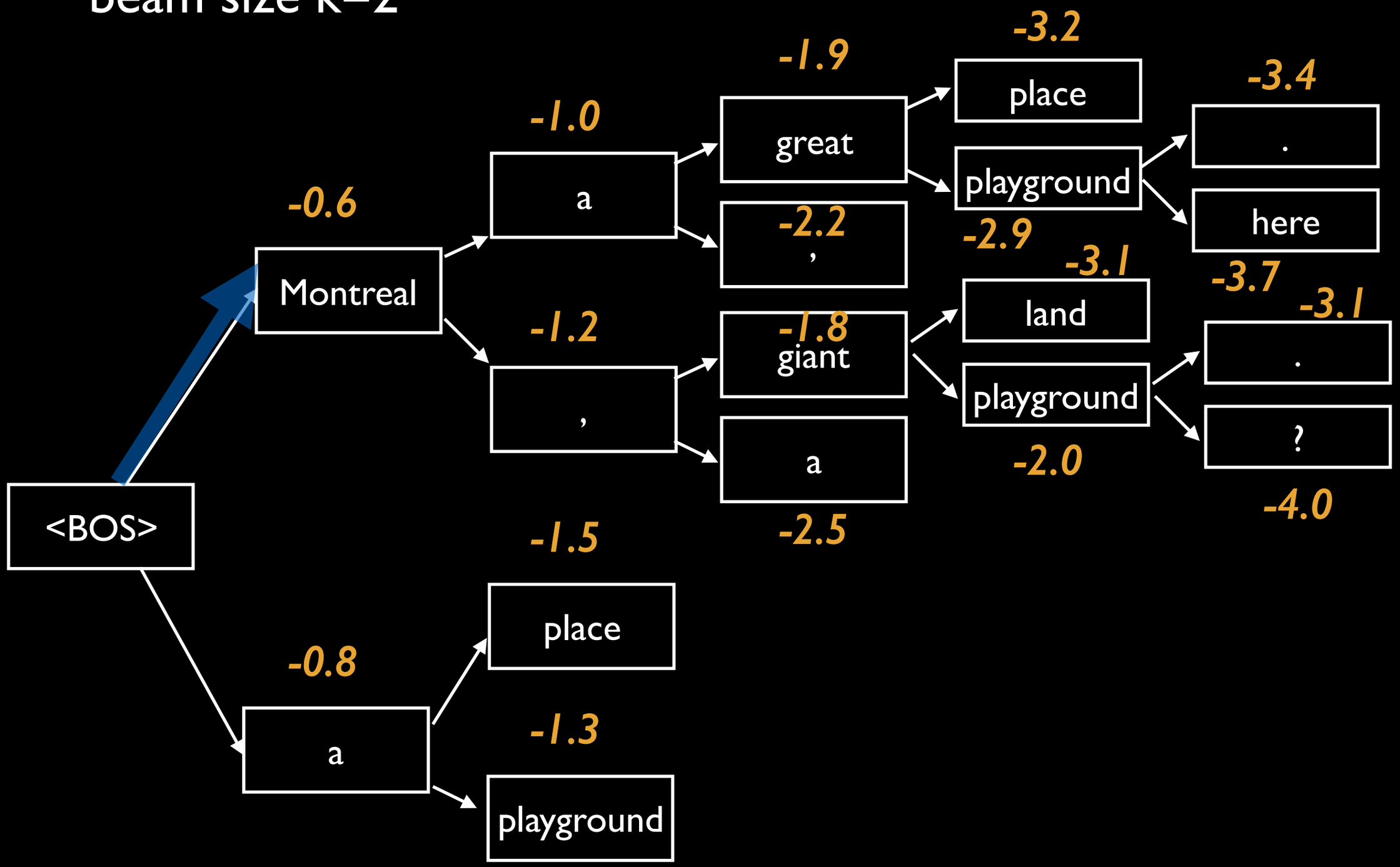
beam search example

beam size k=2



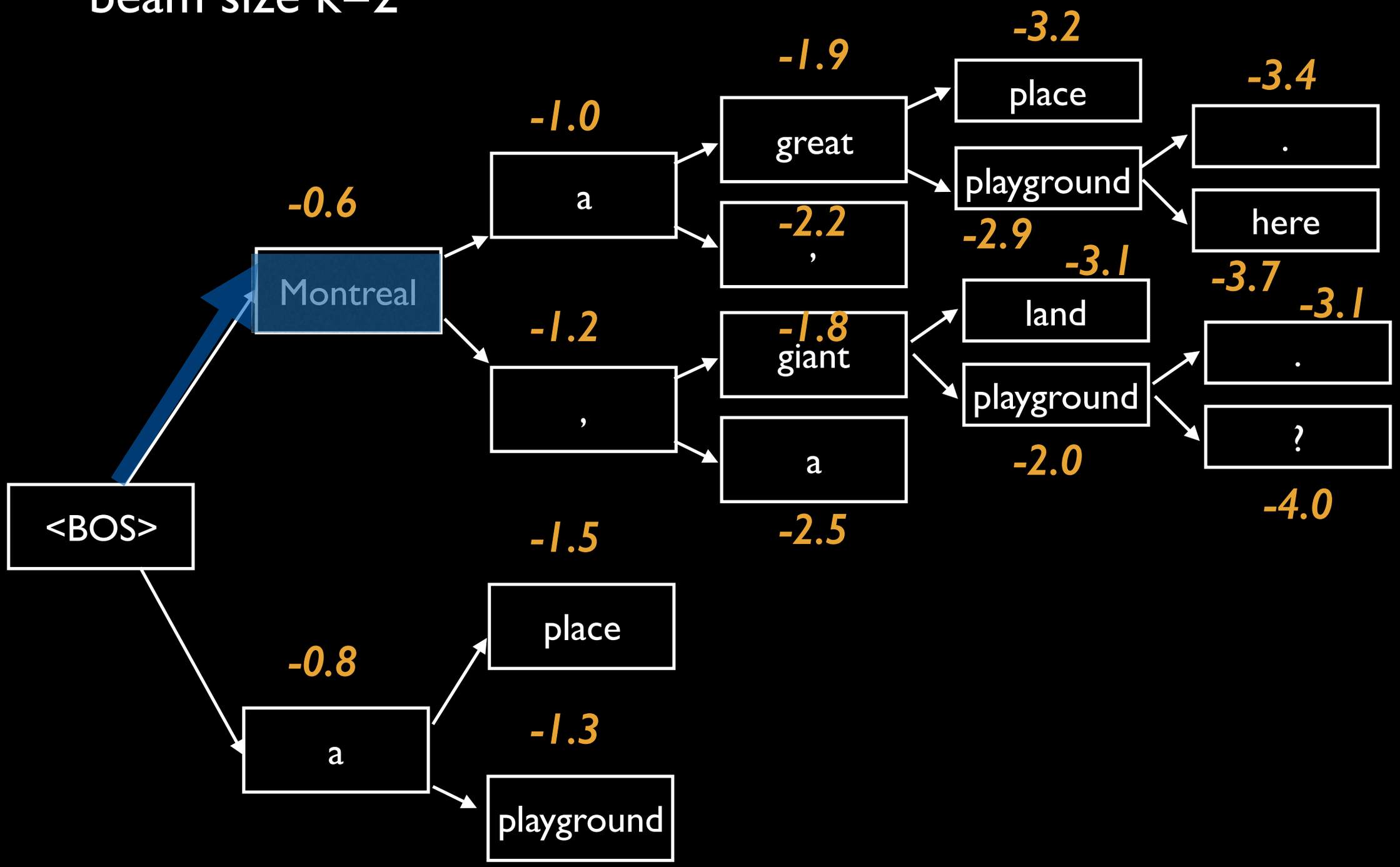
beam search example

beam size k=2



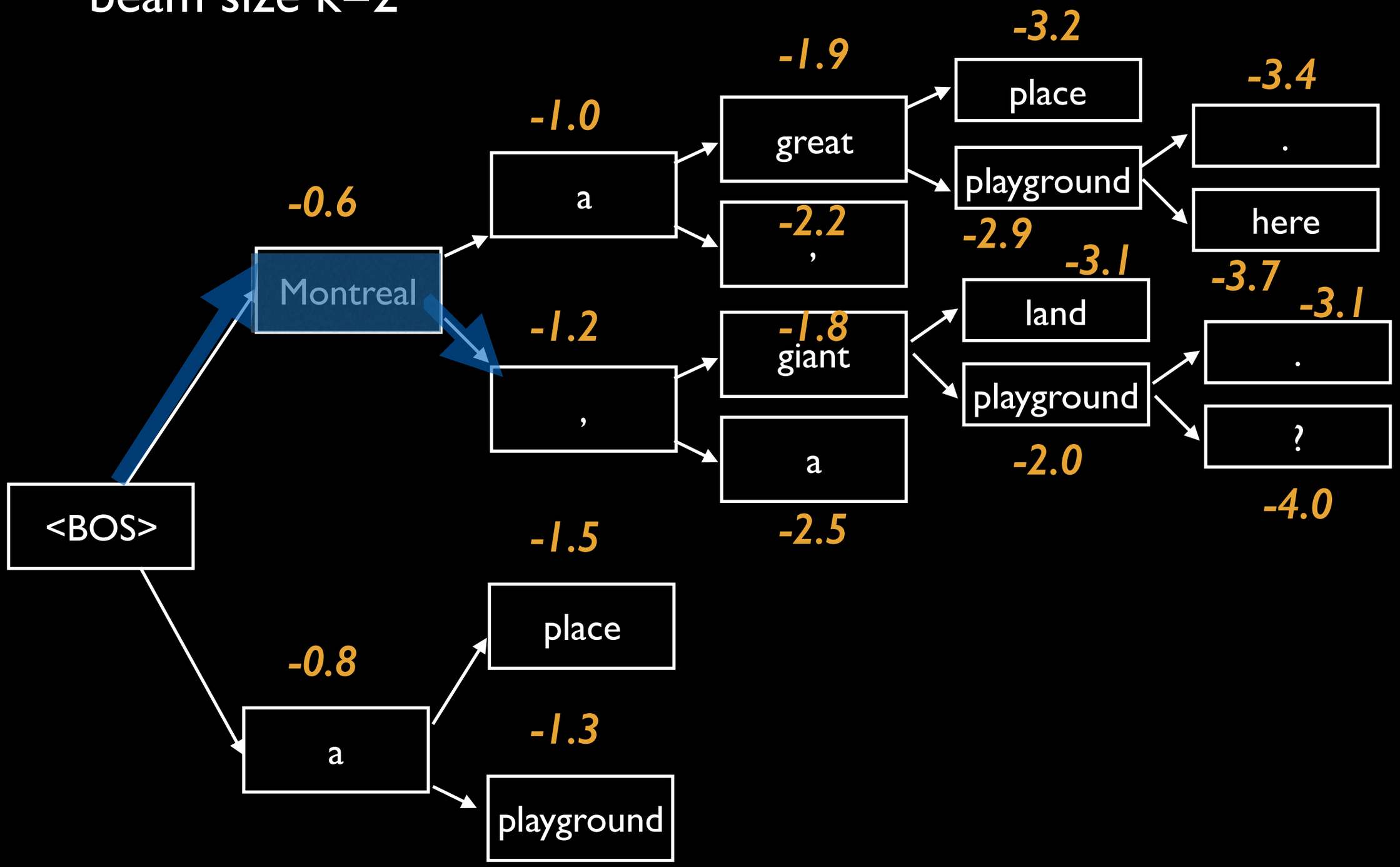
beam search example

beam size k=2



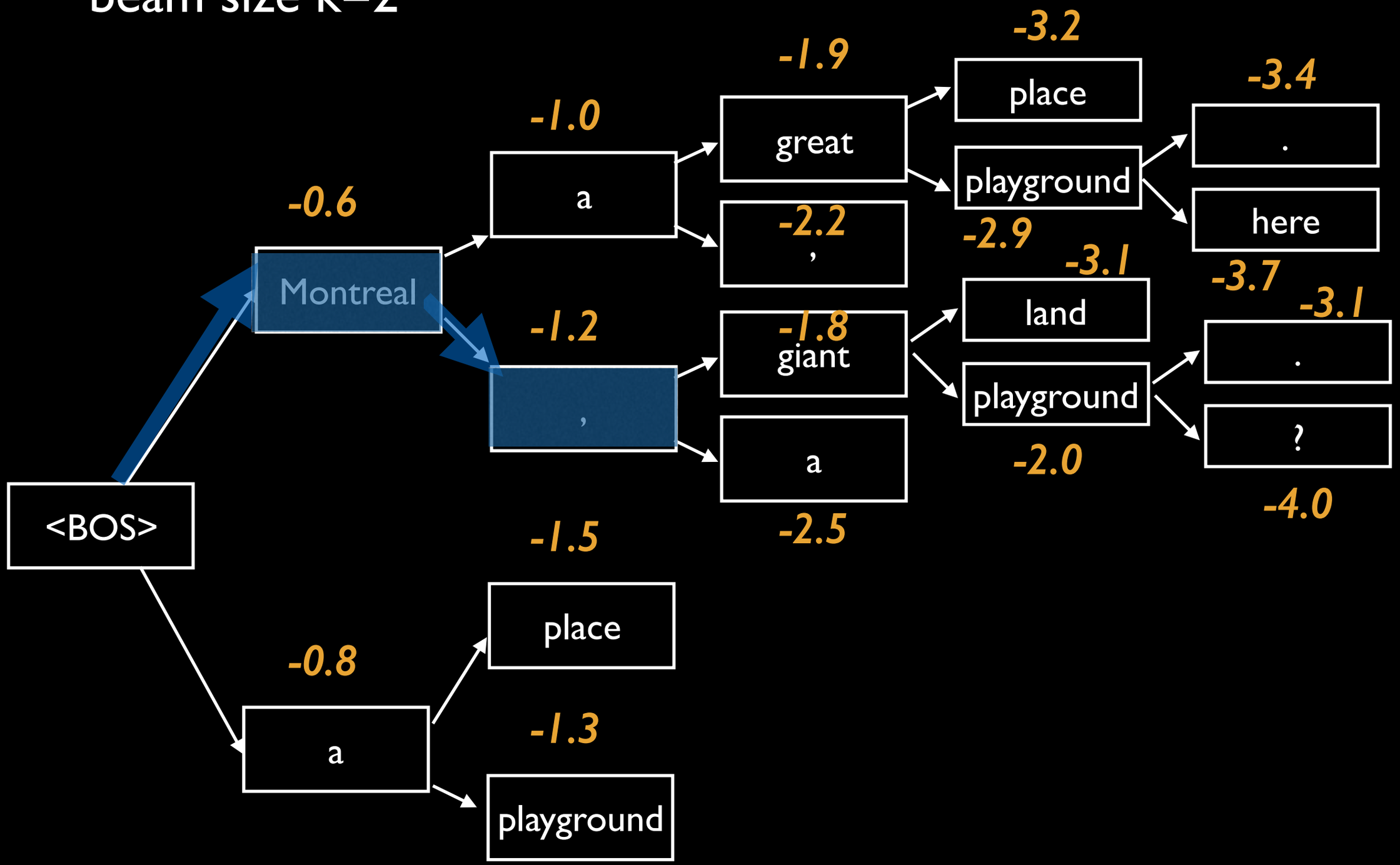
beam search example

beam size k=2



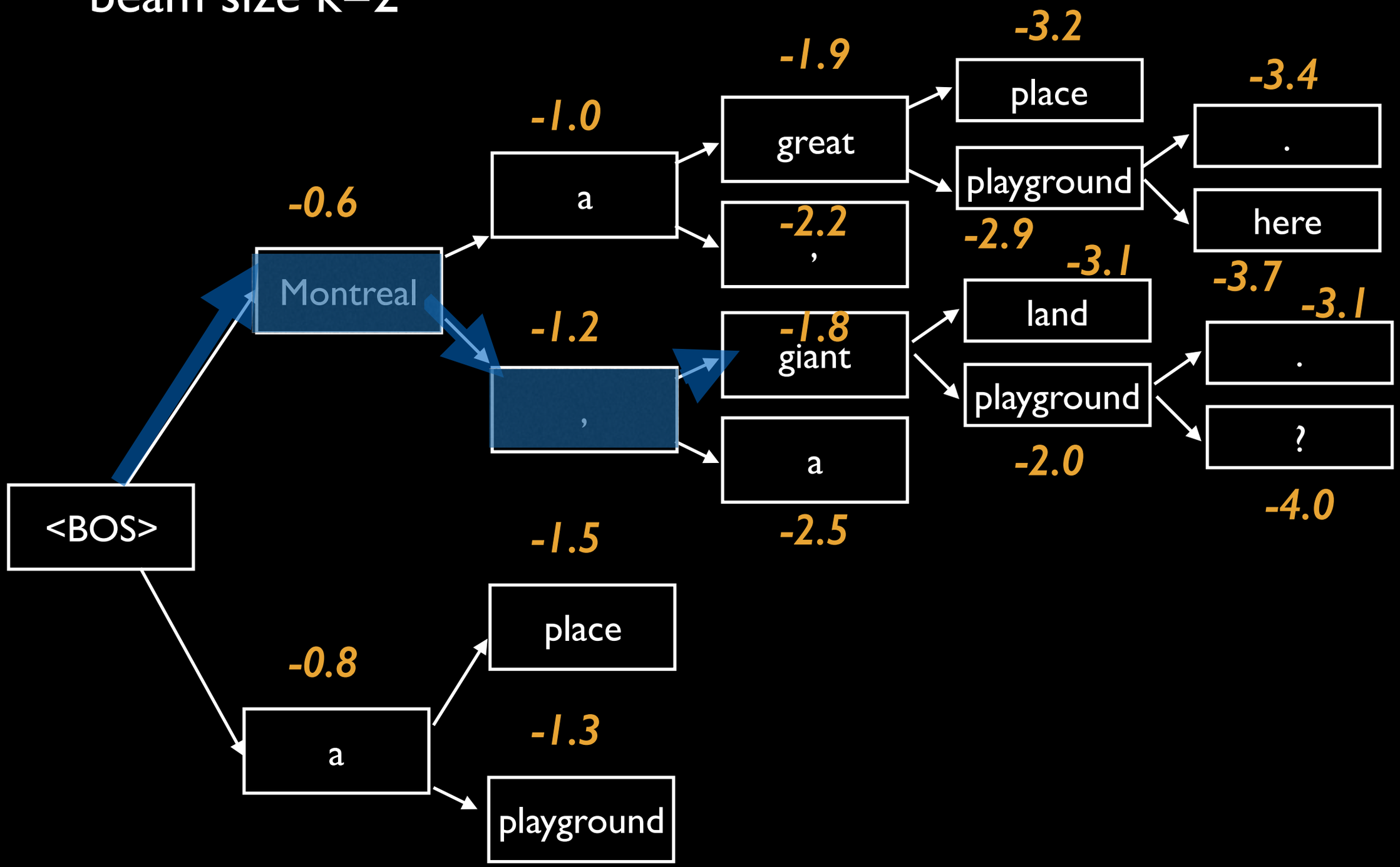
beam search example

beam size k=2



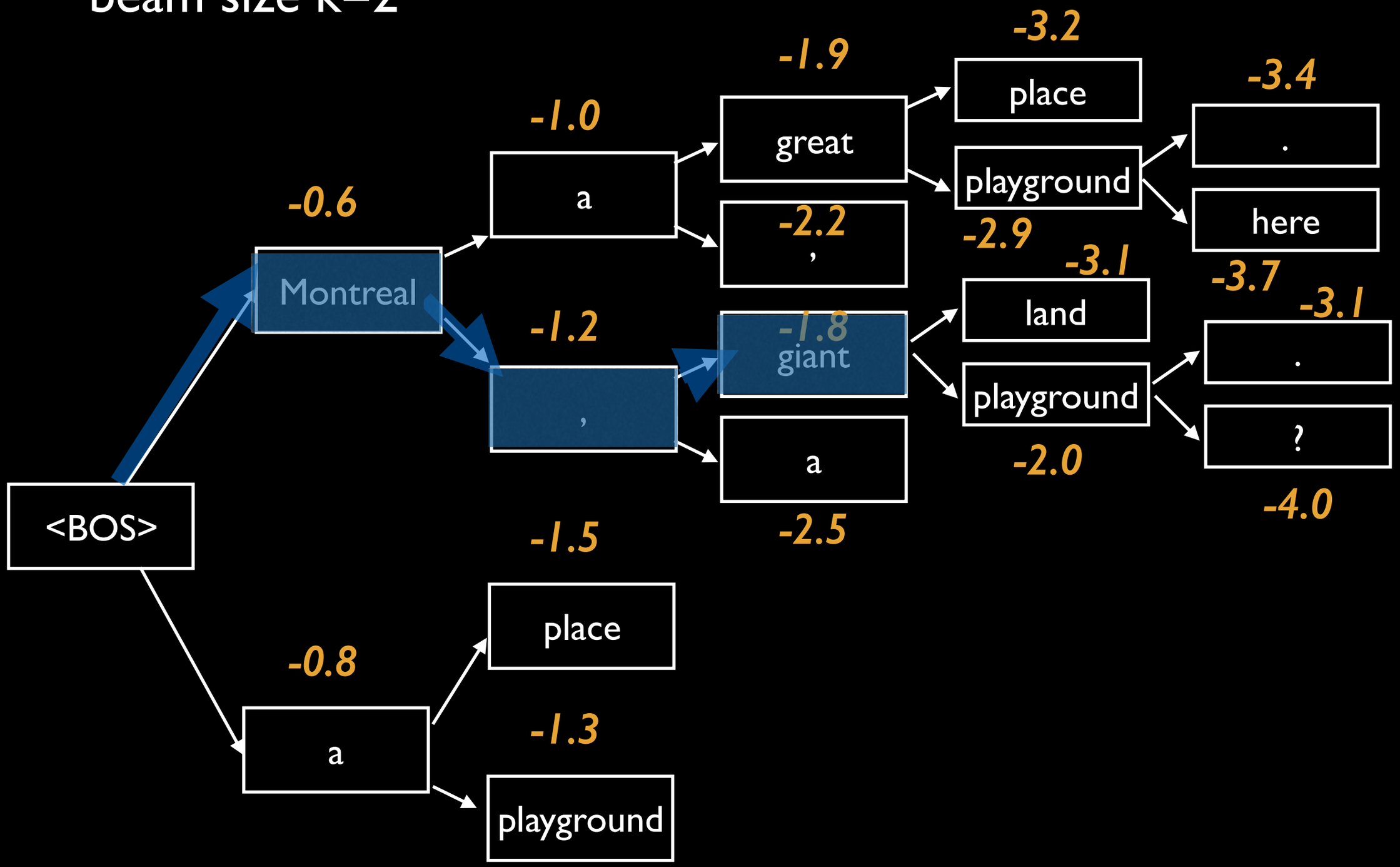
beam search example

beam size k=2



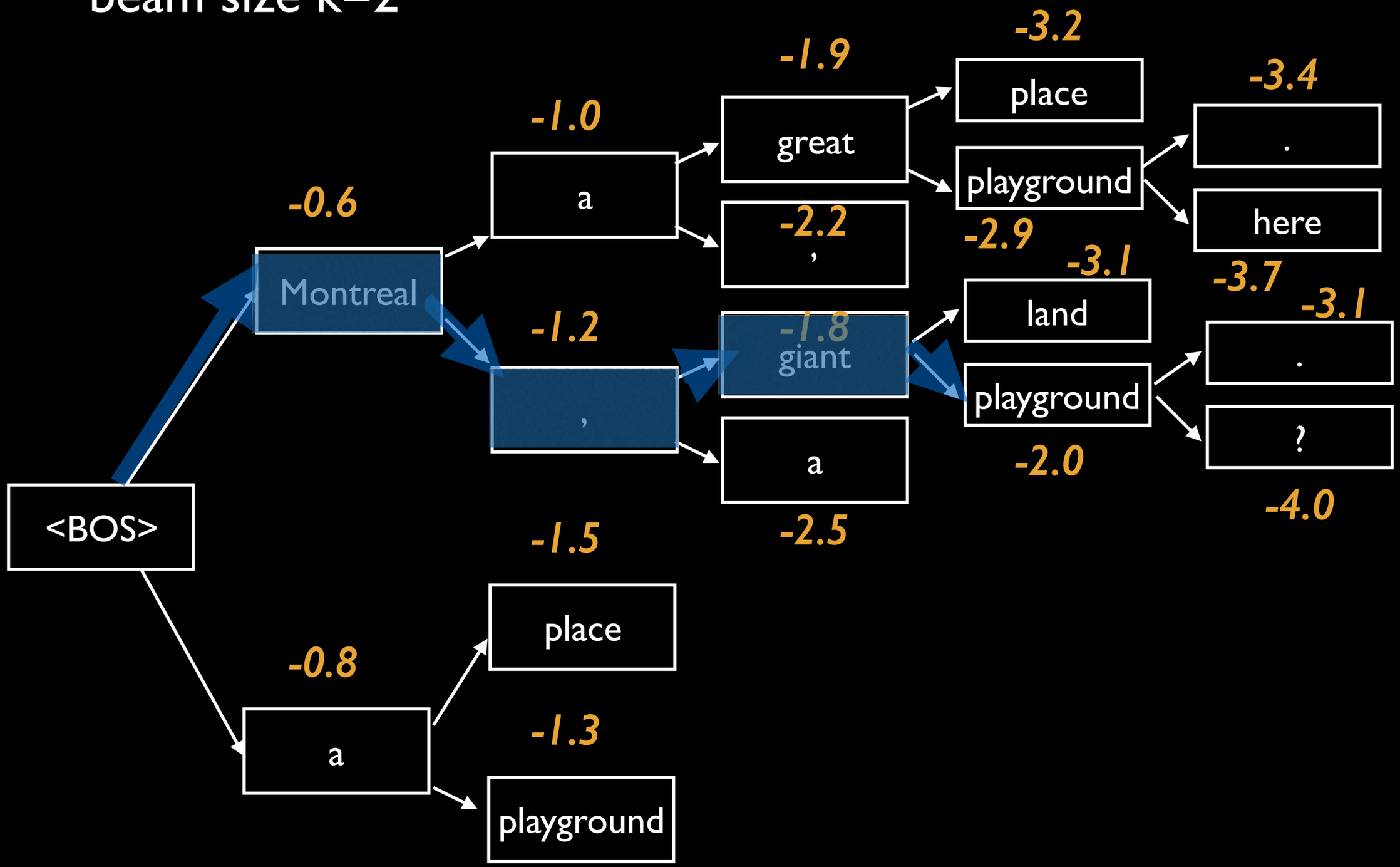
beam search example

beam size k=2



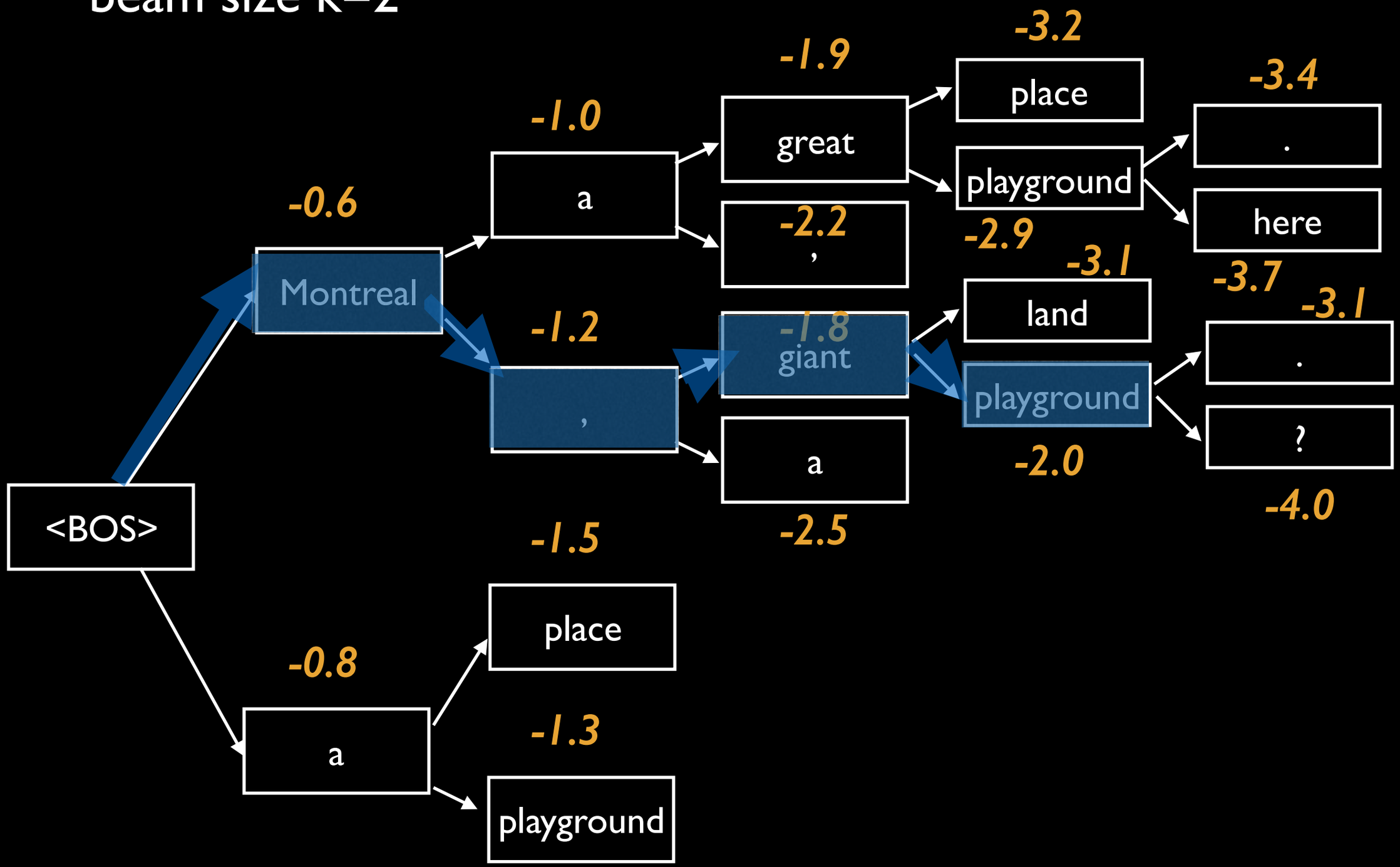
beam search example

beam size k=2



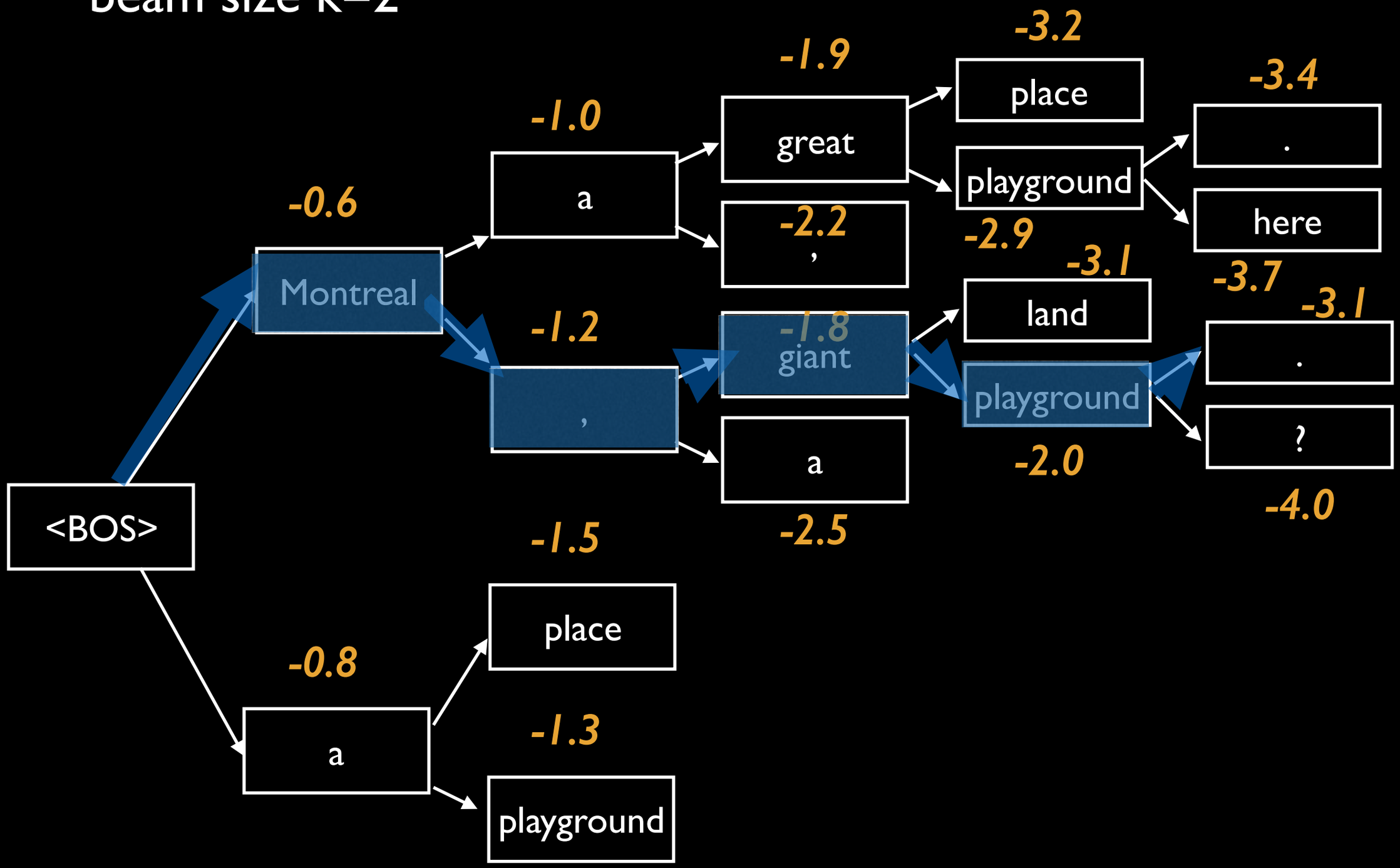
beam search example

beam size k=2



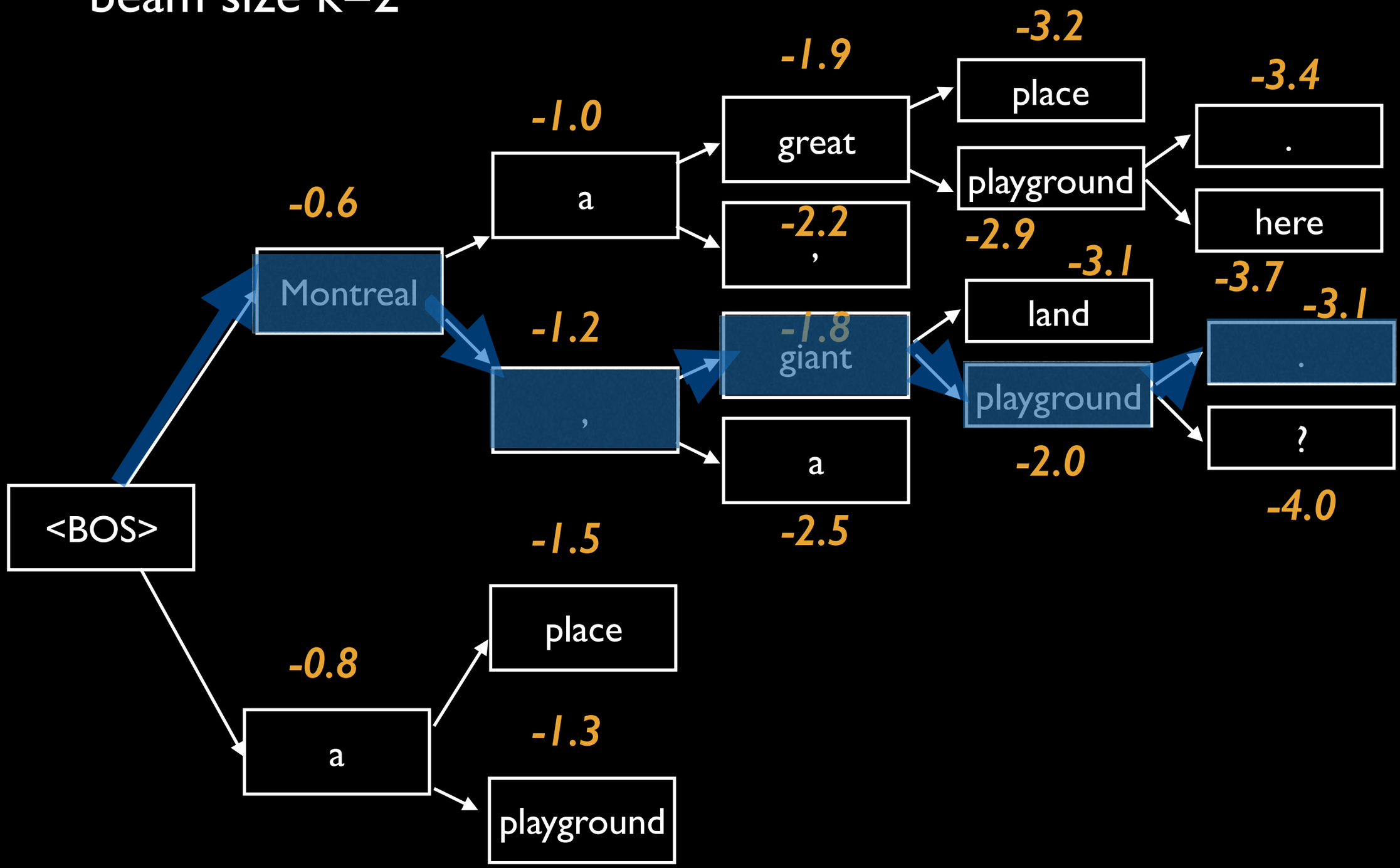
beam search example

beam size k=2



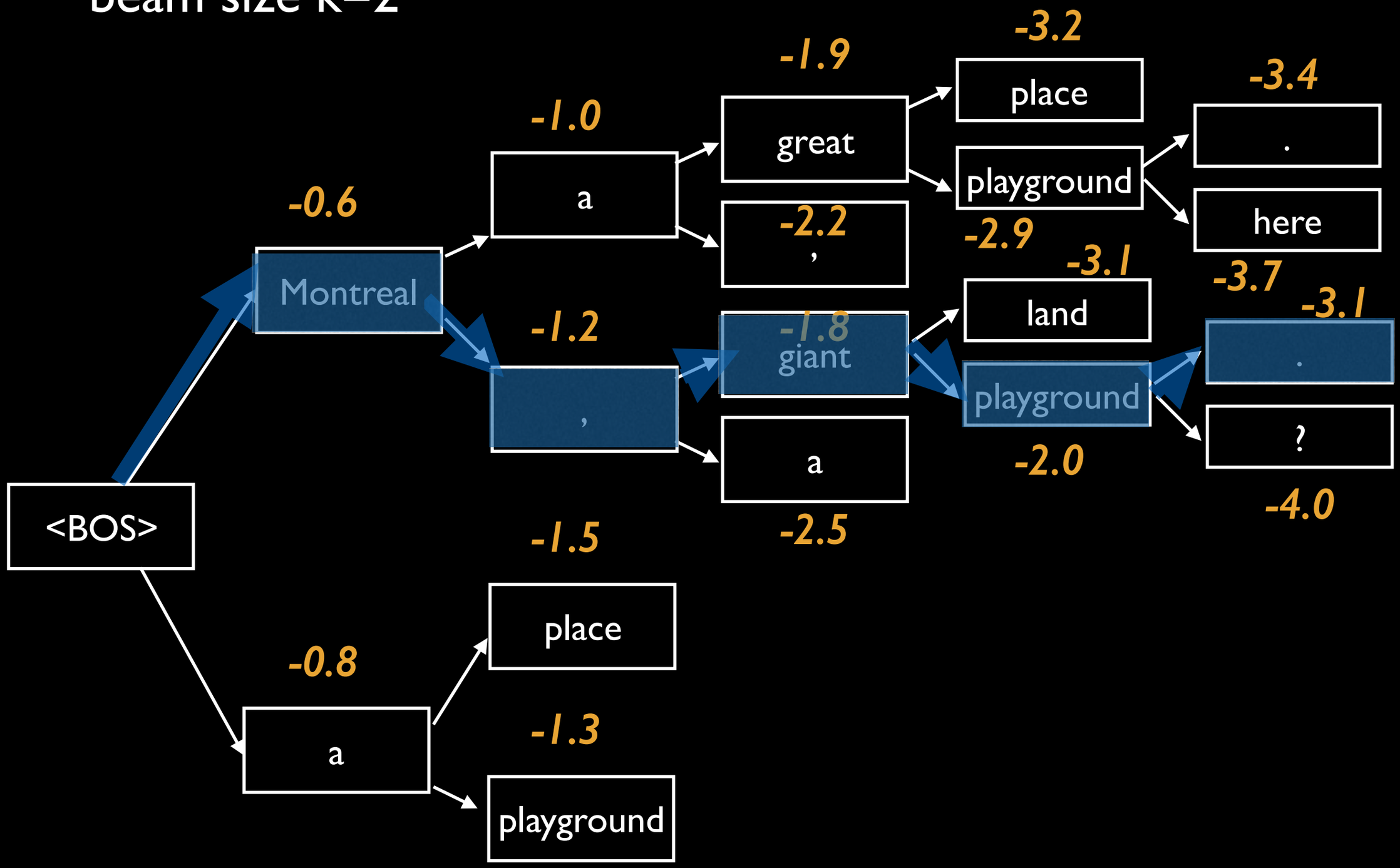
beam search example

beam size k=2



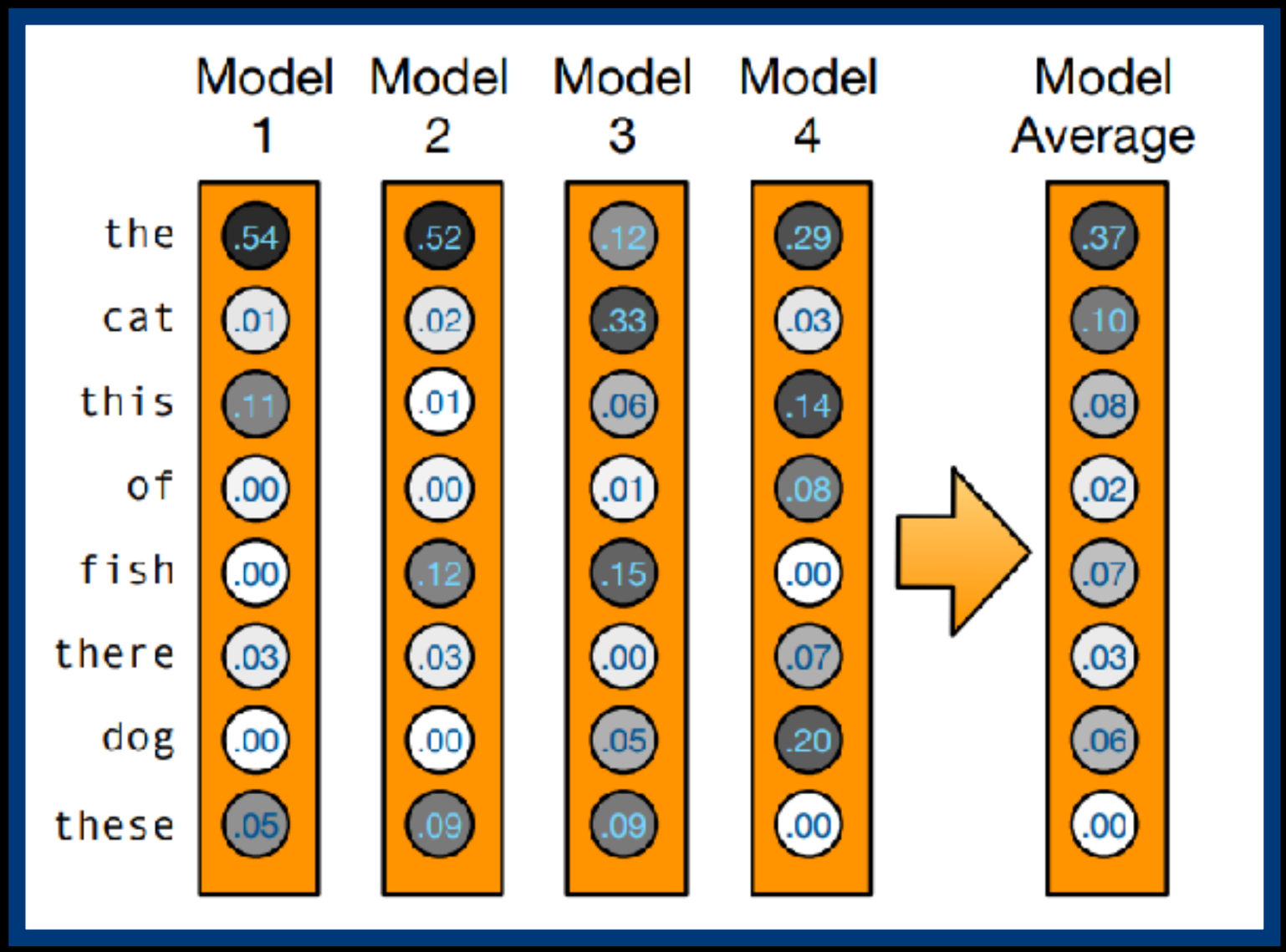
beam search example

beam size k=2



Question #8: more efficient or controlled search? binary NMT, constraint

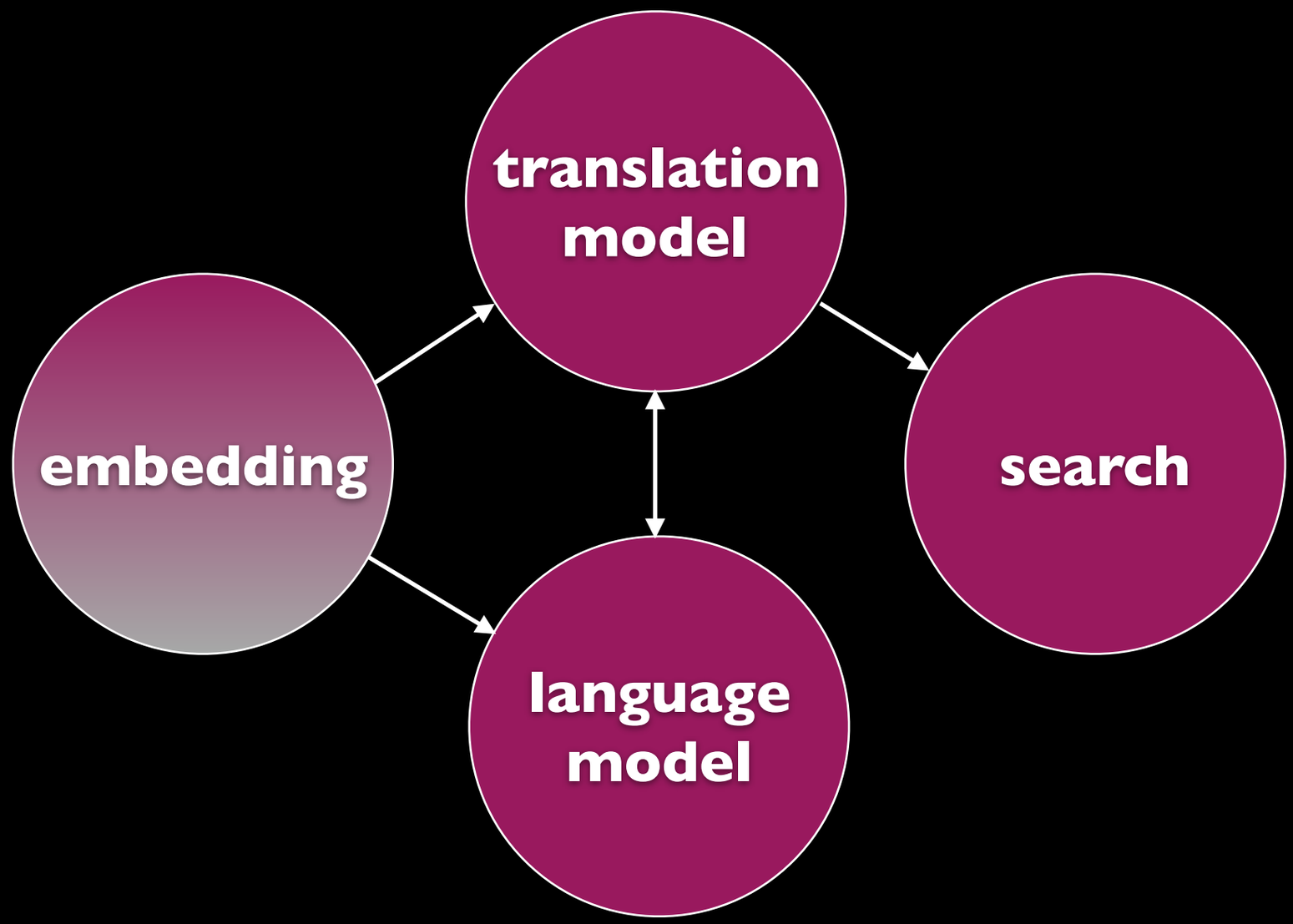
ensemble



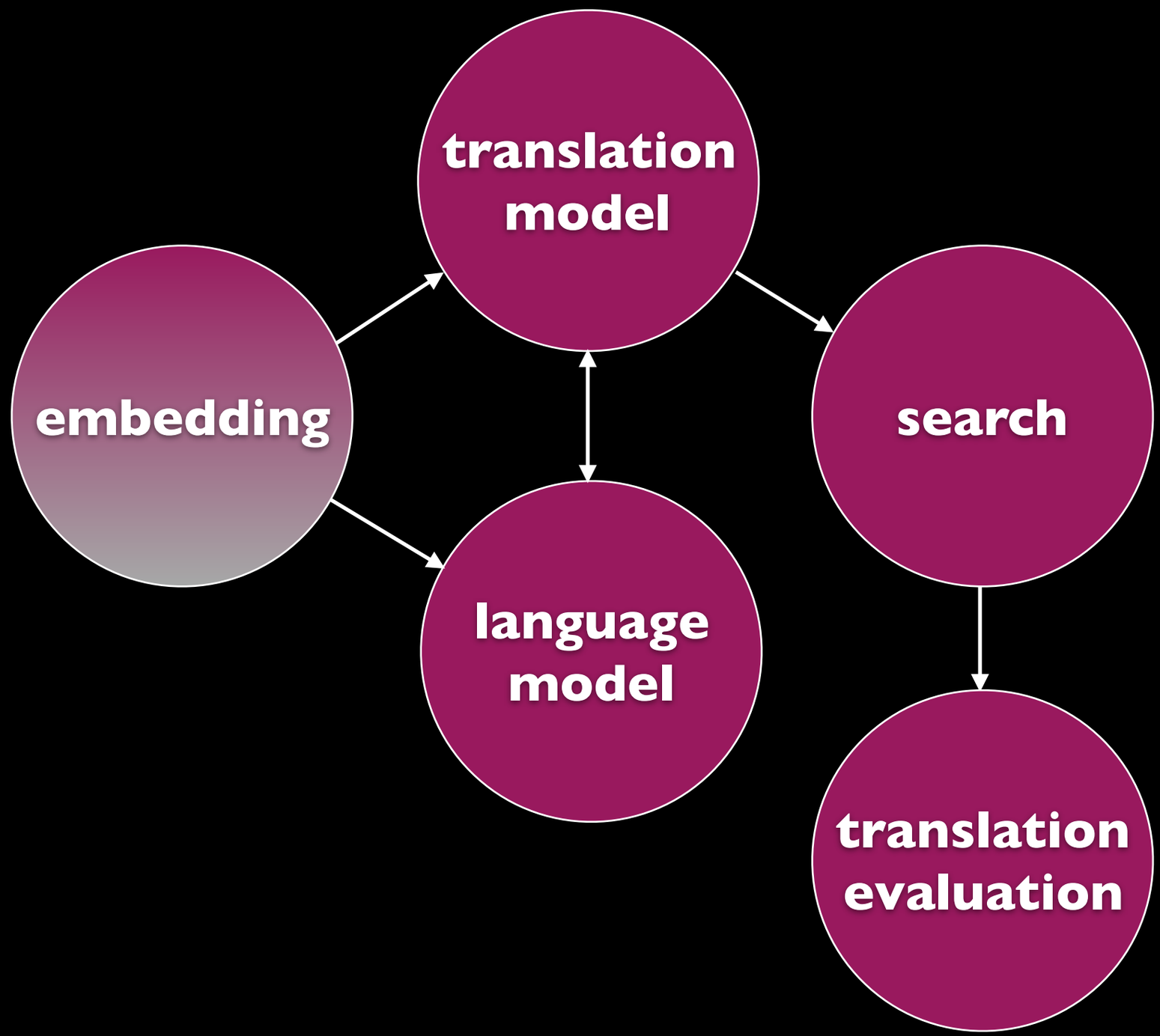
random initialization or
outputs from different iterations

machine translation components

machine translation components



machine translation components



machine translation components

machine translation evaluation

human evaluation is expensive, develop automatic evaluation criteria

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

- method to measure error rates:
- edit distance: insertion, deletion, substitution
- word error rate: normalized edit distance
- HTER: Human Targeted Translation Error Rate

machine translation evaluation

human evaluation is expensive, develop automatic evaluation criteria

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

substitution#= 1

- method to measure error rates:
- edit distance: insertion, deletion, substitution
- word error rate: normalized edit distance
- HTER: Human Targeted Translation Error Rate

machine translation evaluation

human evaluation is expensive, develop automatic evaluation criteria

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

substitution#=1+1

- method to measure error rates:
- edit distance: insertion, deletion, substitution
- word error rate: normalized edit distance
- HTER: Human Targeted Translation Error Rate

machine translation evaluation

human evaluation is expensive, develop automatic evaluation criteria

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

substitution#=1+1 deletion#=1

- method to measure error rates:
- edit distance: insertion, deletion, substitution
- word error rate: normalized edit distance
- HTER: Human Targeted Translation Error Rate

machine translation evaluation

human evaluation is expensive, develop automatic evaluation criteria

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

substitution#=1+1 deletion#=1 insertion#=0

- method to measure error rates:
  - edit distance: insertion, deletion, substitution
  - word error rate: normalized edit distance
  - HTER: Human Targeted Translation Error Rate

machine translation evaluation

human evaluation is expensive, develop automatic evaluation criteria

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

substitution#=1+1 deletion#=1 insertion#=0

edit distance#=1+1+1=3

- method to measure error rates:
  - edit distance: insertion, deletion, substitution
  - word error rate: normalized edit distance
  - HTER: Human Targeted Translation Error Rate

machine translation evaluation

human evaluation is expensive, develop automatic evaluation criteria

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

substitution#=1+1 deletion#=1 insertion#=0

edit distance#=1+1+1=3

- method to measure error rates:
- edit distance: insertion, deletion, substitution
- word error rate: normalized edit distance 3/6=0.5
- HTER: Human Targeted Translation Error Rate

machine translation evaluation: BLEU

BLEU (Bilingual Evaluation Understudy)

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

- method to measure accuracy
- most well-cited evaluation criterion so far
  - precision of 1-gram through 4-gram
  - brevity penalty
- mainly relies on n-gram coverage

machine translation evaluation: BLEU

BLEU (Bilingual Evaluation Understudy)

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

- method to measure accuracy
- most well-cited evaluation criterion so far
  - precision of 1-gram through 4-gram
  - brevity penalty
- mainly relies on n-gram coverage

machine translation evaluation: BLEU

BLEU (Bilingual Evaluation Understudy)

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

- method to measure accuracy
- most well-cited evaluation criterion so far
  - precision of 1-gram through 4-gram
  - brevity penalty
- mainly relies on n-gram coverage

machine translation evaluation: BLEU

BLEU (Bilingual Evaluation Understudy)

hypothesis: ``Montreal is a city . ”

reference: ``Montreal , a giant playground . ”

- method to measure accuracy
- most well-cited evaluation criterion so far
  - precision of 1-gram through 4-gram
  - brevity penalty
- mainly relies on n-gram coverage

machine translation evaluation: BLEU

BLEU (Bilingual Evaluation Understudy)

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

- method to measure accuracy
- most well-cited evaluation criterion so far
  - precision of 1-gram through 4-gram
  - brevity penalty
- mainly relies on n-gram coverage

1-gram#=3
2-gram#=0
3-gram#=0
4-gram#=0

machine translation evaluation: BLEU

BLEU (Bilingual Evaluation Understudy)

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

- method to measure accuracy
  - most well-cited evaluation criterion so far
    - precision of 1-gram through 4-gram
    - brevity penalty
  - mainly relies on n-gram coverage

1-gram#=3
2-gram#=0
3-gram#=0
4-gram#=0

Question #9: higher correlation with human judgement? rich literature

machine translation evaluation: BLEU

BLEU (Bilingual Evaluation Understudy)

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

- method to measure accuracy
  - most well-cited evaluation criterion so far
    - precision of 1-gram through 4-gram
    - brevity penalty

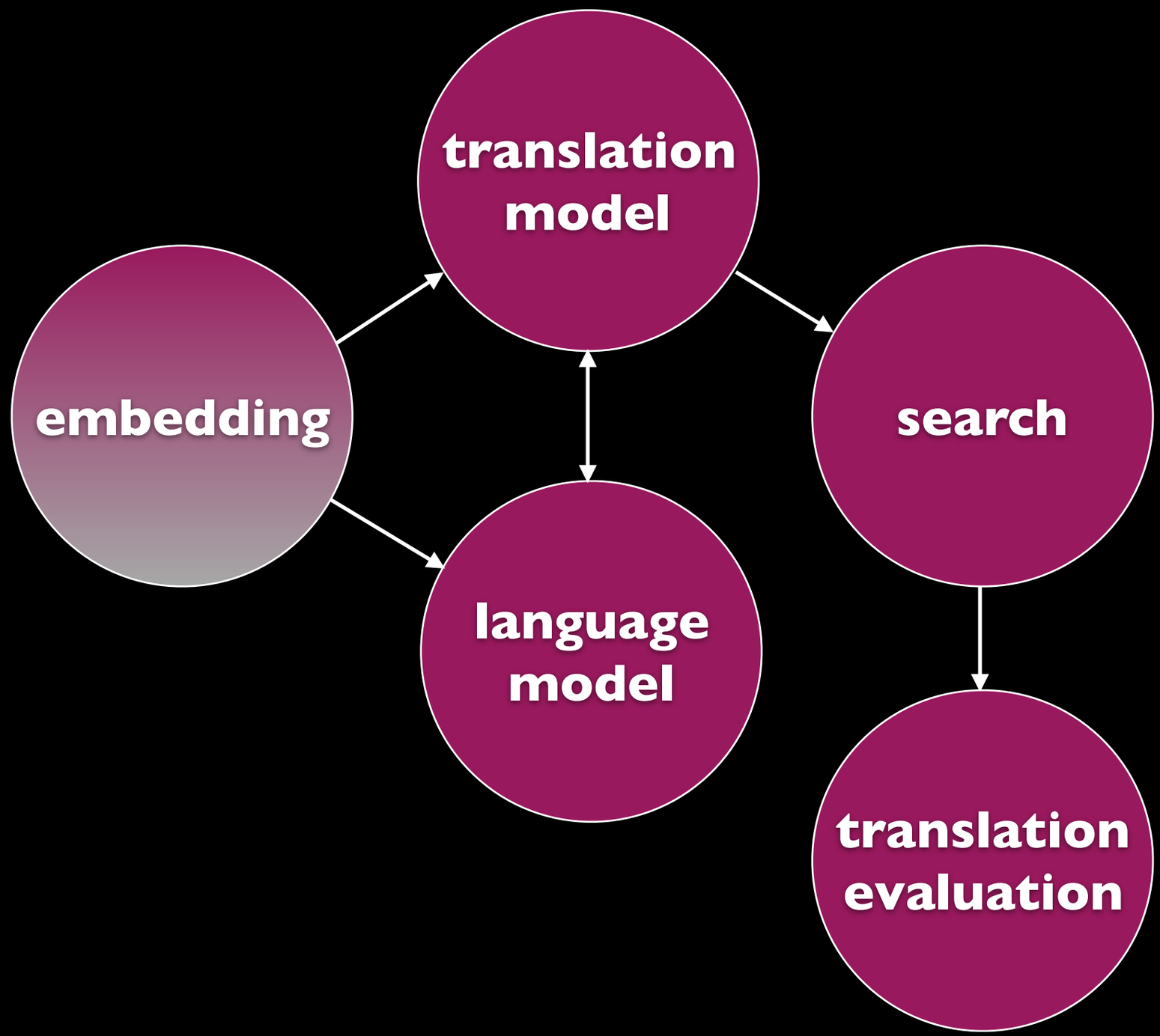
1-gram#=3
2-gram#=0
3-gram#=0
4-gram#=0

Question #10: better quality estimation?

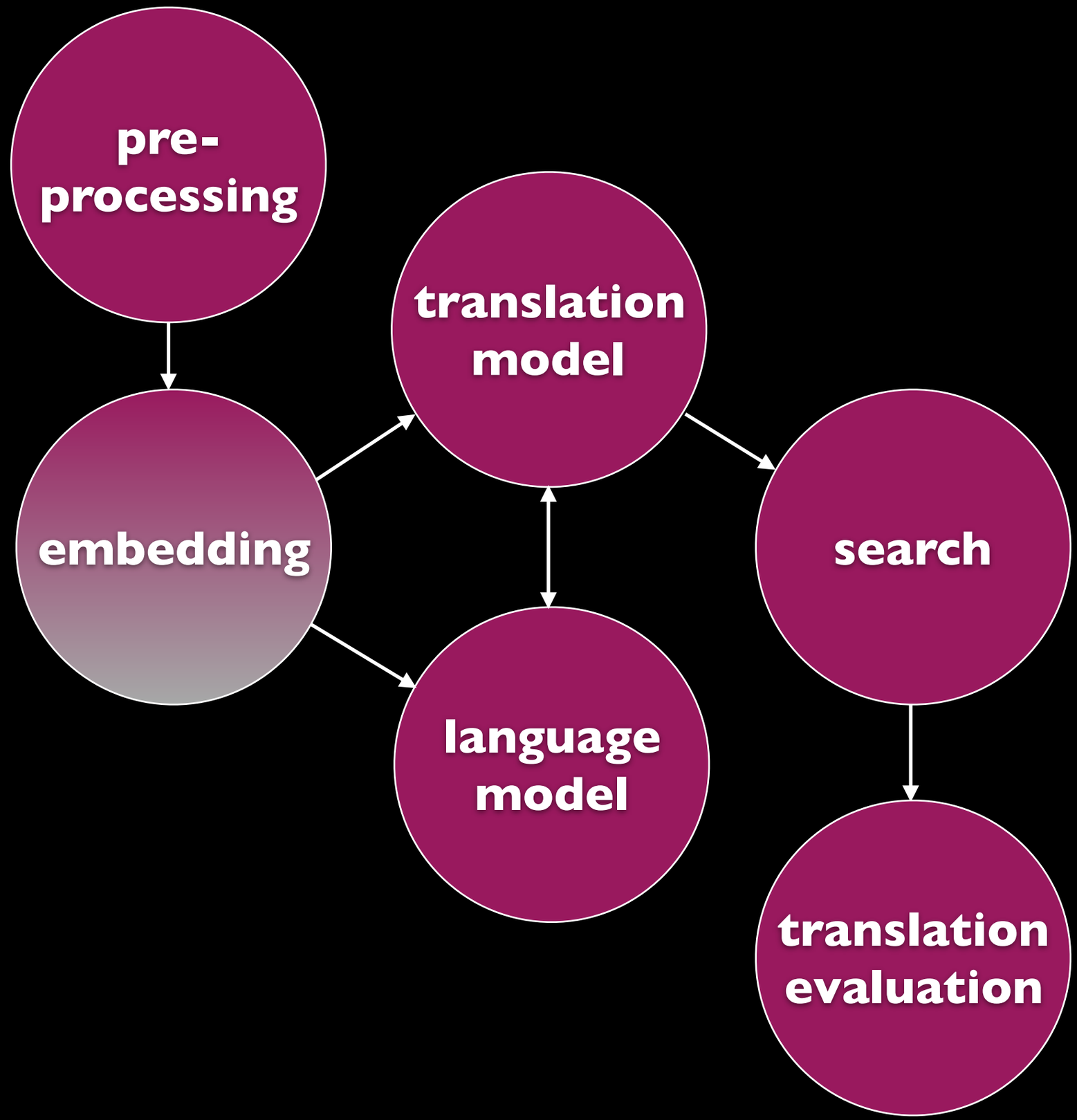
Question #9: higher correlation with human judgement? rich literature

machine translation components

machine translation components



machine translation components



machine translation components

pre-processing

- tokenization: separate words from punctuation marks, typically based on rules
- text normalization
- word segmentation
- sentence segmentation
- domain classification

pre-processing

- tokenization: separate words from punctuation marks, typically based on rules
- text normalization
- word segmentation
- sentence segmentation
- domain classification

Question #11: text normalization

pre-processing

- tokenization: separate words from punctuation marks, typically based on rules
- text normalization
- word segmentation
- sentence segmentation
- domain classification

Question #11: text normalization

Question #12: better subword?

pre-processing

- tokenization: separate words from punctuation marks, typically based on rules
- text normalization
- word segmentation
- sentence segmentation
- domain classification

Question #11: text normalization

Question #12: better subword?

Question #13: monolingual and bilingual sentence segmentation

pre-processing

- tokenization: separate words from punctuation marks, typically based on rules
- text normalization
- word segmentation
- sentence segmentation
- domain classification

Question #11: text normalization

Question #12: better subword?

Question #13: monolingual and bilingual sentence segmentation

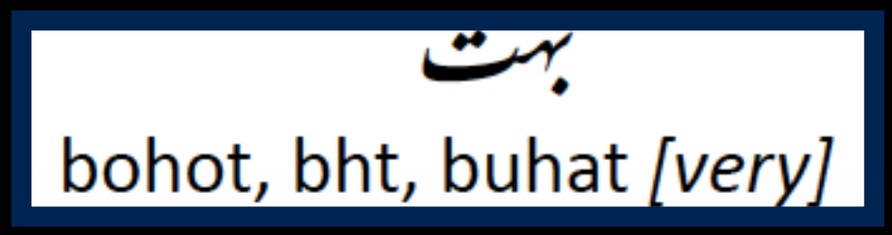
Question #14: domain adaptation

text normalization

formal text: ``are you coming to the class tomorrow?''

informal text: ``r u cuming 2 class tomr?''

- bad translation: style, domain change, noise e.g. mis-spelling
- goal: translate different lexical variations
  - add noise to training: [Michell, et.al., 19]
  - word clustering [Khan et. al., 19]



word segmentation

明 天 来 上 课 吗?

- possible segmentation boundaries: 2^{k-1}
  - k: number of characters

- n-gram approach:

n-gram approach

$$\hat{f}_1^J(c_1^K) = \underset{k_1^J, J}{\operatorname{argmax}} \prod_{j-1}^{J} Pr(c_{k_{j-1}}^{k_j} | c_{k_{j-1-n+2}}^{k_{j-1}}, \ldots, c_{k_{j-1}}^{k_{j-1}})$$

unigram

$$= \underset{k_1^J, J}{\operatorname{argmax}} \prod_{j-1}^{J} Pr(c_{k_{j-1}+1}^{k_j})$$

word segmentation

明天 来 上 课 吗?

- possible segmentation boundaries: 2^{k-1}
  - k: number of characters

- n-gram approach:

n-gram approach

$$\hat{f}_1^J(c_1^K) = \underset{k_1^J, J}{\operatorname{argmax}} \prod_{j-1}^{J} Pr(c_{k_{J-1}}^{k_J} | c_{k_{J-2}}^{k_{J-1-n}}, \ldots, c_{k_{J-2}}^{k_{J-1}})$$

unigram

$$= \underset{k_1^J, J}{\operatorname{argmax}} \prod_{j-1}^{J} Pr(c_{k_{J-1}+1}^{k_J})$$

word segmentation

明|天|来 上 课 吗?

- possible segmentation boundaries: 2^{k-1}
  - k: number of characters

- n-gram approach:

n-gram approach

$$\hat{f}_1^J(c_1^K) = \underset{k_1^J, J}{\operatorname{argmax}} \prod_{j-1}^{J} Pr(c_{k_{J-1}}^{k_J} | c_{k_{J-1-n+1}}^{k_{J-1}}, \ldots, c_{k_{J-1}}^{k_{J-1}})$$

unigram

$$= \underset{k_1^J, J}{\operatorname{argmax}} \prod_{j-1}^{J} Pr(c_{k_{J-1}+1}^{k_J})$$

word segmentation

明|天|来|上 课 吗?

- possible segmentation boundaries: 2^{k-1}
  - k: number of characters

- n-gram approach:

n-gram approach

$$\hat{f}_1^J(c_1^K) = \underset{k_1^J, J}{\operatorname{argmax}} \prod_{j-1}^{J} Pr(c_{k_{j-1}}^{k_j} | c_{k_{j-1-n+1}}^{k_{j-1}}, \ldots, c_{k_{j-1}}^{k_{j-1}})$$

unigram

$$= \underset{k_1^J, J}{\operatorname{argmax}} \prod_{j-1}^{J} Pr(c_{k_{j-1}+1}^{k_j})$$

word segmentation

明|天|来|上|课 吗?

- possible segmentation boundaries: 2^{k-1}
  - k: number of characters

- n-gram approach:

n-gram approach

$$\hat{f}_1^J(c_1^K) = \underset{k_1^J, J}{\operatorname{argmax}} \prod_{j-1}^{J} Pr(c_{k_{J-1}}^{k_J} | c_{k_{J-1-n}}^{k_{J-1}})$$

unigram

$$= \underset{k_1^J, J}{\operatorname{argmax}} \prod_{j-1}^{J} Pr(c_{k_{J-1}}^{k_J})$$

word segmentation

明|天|来|上|课|吗?

- possible segmentation boundaries: 2^{k-1}
  - k: number of characters

- n-gram approach:

n-gram approach

$$\hat{f}_1^J(c_1^K) = \underset{k_1^J, J}{\operatorname{argmax}} \prod_{j-1}^{J} Pr(c_{k_{J-1}}^{k_J} | c_{k_{J-1-n+1}}^{k_{J-1}}, \ldots, c_{k_{J-1}}^{k_{J-1}})$$

unigram

$$= \underset{k_1^J, J}{\operatorname{argmax}} \prod_{j-1}^{J} Pr(c_{k_{J-1}+1}^{k_J})$$

word segmentation

明|天|来|上|课|吗?

- possible segmentation boundaries: 2^{k-1}
  - k: number of characters

- n-gram approach:

statistical

n-gram approach

$$\hat{f}_1^J(c_1^K) = \underset{k_1^J, J}{\operatorname{argmax}} \prod_{j-1}^{J} Pr(c_{k_{j-1}}^{k_j} | c_{k_{j-1-n+2}}^{k_{j-1}}, \ldots, c_{k_{j-1}}^{k_{j-1}})$$

unigram

$$= \underset{k_1^J, J}{\operatorname{argmax}} \prod_{j-1}^{J} Pr(c_{k_{j-1}+1}^{k_j})$$

character based statistical MT

明|天|来|上 课|吗?

- Gibbs sampling: joint model word alignment and word segmentation

$$\mathrm{argmax}_{f_1^J}\{P(f_1^J{}^{\lambda_1} P(e_1^J, b_1 I|f_1^J)^{\lambda_2} P(f_1^J, a_1^J|e_1^I)^{\lambda_3}$$

Transition from a boundary(+) to a no-boundary (−)

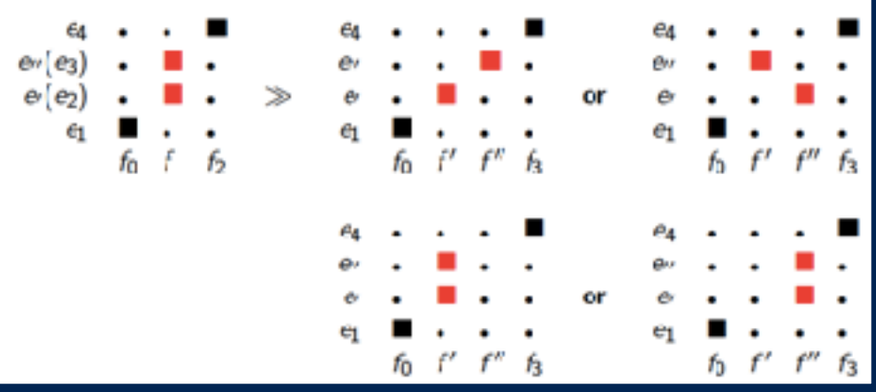
- **C-to-E:** $P_{fe}(cba^-{}_k|dh_{sk}{}^+) : \frac{1}{I+1} P_G(f|e'_*)$ **or** $\frac{1}{I+1} P_G(f|e''_*)$, $I-3$

Transition from boundary(+) to no-boundary (−)

- **E-to-C:** $P_{ef}(cba_k|dh_{sk}{}^-) : \frac{1}{J}{}^{|e|} P_G(e'|f) P_G(e''|f)$, **here** $|e|-2, J-3$

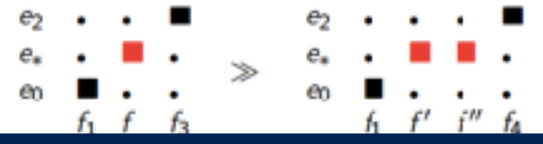
Transition from no-boundary (−) to boundary (+)

- **E-to-C:** $P_{ef}(cba_k^+|dh_{sk}{}^-) : \frac{1}{J+2}{}^{|e|} P(e_1|f') P(e_2|f'')$ **or** ..., $J=2$



Transition from no-boundary(−) to boundary(+)

- **C-to-E:** $P_{fe}(cba_k^-|dh_{sk}{}^-) : \frac{1}{I+1}{}^2 P_G(f'|e_*) P_G(f''|e_*)$, $I=2$



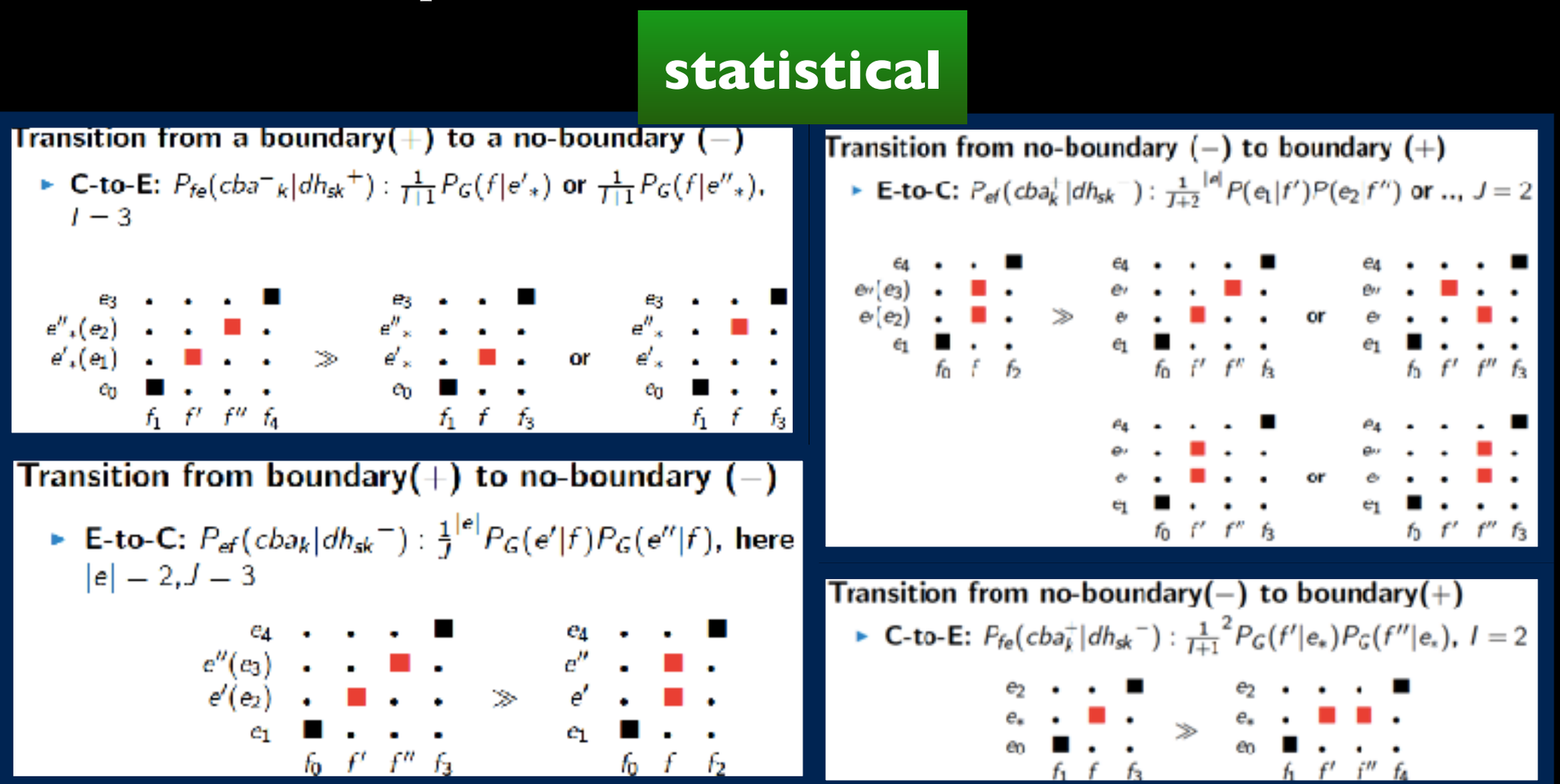
character based statistical MT

明|天|来|上 课|吗?

- Gibbs sampling: joint model word alignment and word segmentation

$$\mathrm{argmax}_{f_1^J}\{P(f_1^J{}^{\lambda_1} P(e_1^J, b_1 I|f_1^J)^{\lambda_2} P(f_1^J, a_1^J|e_1^I)^{\lambda_3}$$

statistical



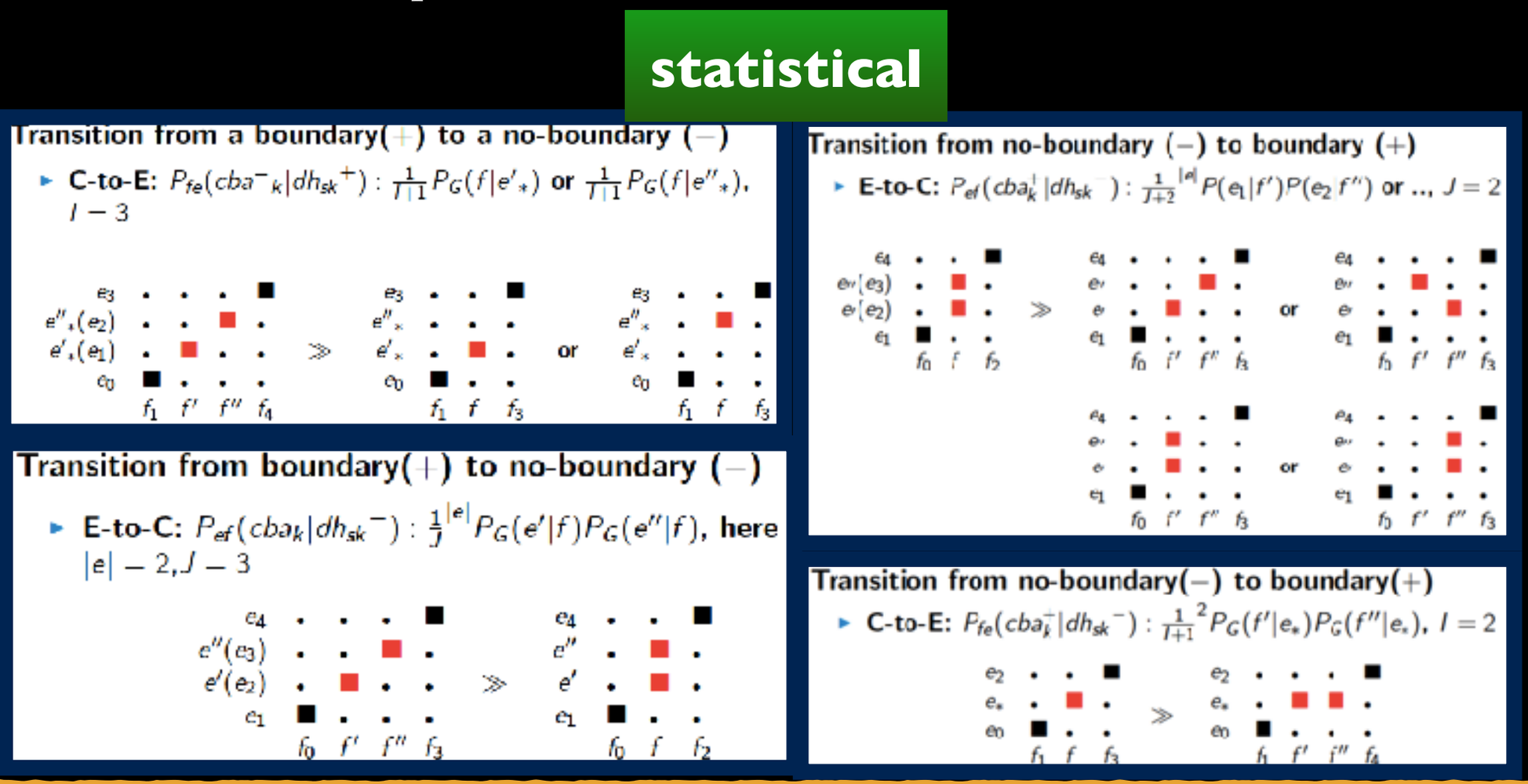
character based statistical MT

明|天|来|上 课|吗?

- Gibbs sampling: joint model word alignment and word segmentation

$$\text{argmax}_{f_1^J}\{P(f_1^J{}^{\lambda_1} P(e_1^J, b_1 I|f_1^J)^{\lambda_2} P(f_1^J, a_1^J|e_1^I)^{\lambda_3}$$

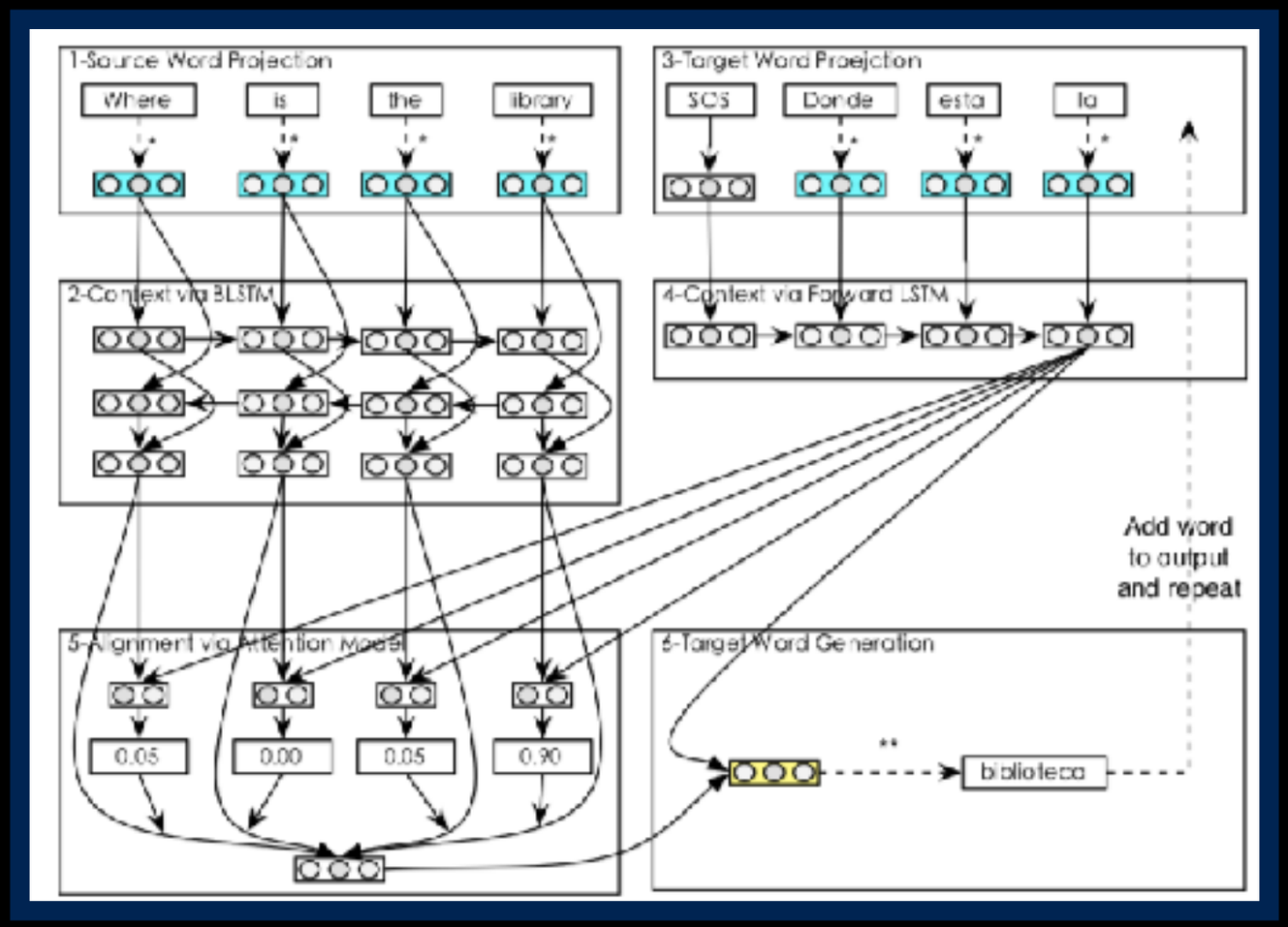
statistical



Question #15: what can we borrow from statistical MT?

character based neural MT

- integrate with neural network framework
  - [Ling, et.al., 15], [Cherry, et.al. 18], [Lee, et.al., 18]…



Byte-Pair-Encoding (BPE)

- efficient and affective, NMT independent
- [Gage, 1994], [Sennrich, et.al., 16]
  - aaabdaaabac
  - ZabdZabac (Z=aa)
  - ZYdZYac (Y=ab; Z=aa)
  - XdXac (X=ZY;Y=ab; Z=aa)

Byte-Pair-Encoding (BPE)

- efficient and affective, NMT independent
- [Gage, 1994], [Sennrich, et.al., 16]
  - aaabdaaabac
  - ZabdZabac (Z=aa)
  - ZYdZYac (Y=ab; Z=aa)
  - XdXac (X=ZY; Y=ab; Z=aa)

Question #16: better subword e.g. with morphological knowledge?

Byte-Pair-Encoding (BPE)

- efficient and affective, NMT independent
- [Gage, 1994], [Sennrich, et.al., 16]
  - aaabdaaabac
  - ZabdZabac (Z=aa)
  - ZYdZYac (Y=ab; Z=aa)
  - XdXac (X=ZY; Y=ab; Z=aa)

Question #16: better subword e.g. with morphological knowledge?

Question #17: unseen words?

Byte-Pair-Encoding (BPE)

- efficient and affective, NMT independent
- [Gage, 1994], [Sennrich, et.al., 16]
  - aaabdaaabac
  - ZabdZabac (Z=aa)
  - ZYdZYac (Y=ab; Z=aa)
  - XdXac (X=ZY; Y=ab; Z=aa)

Question #16: better subword e.g. with morphological knowledge?

Question #17: unseen words?

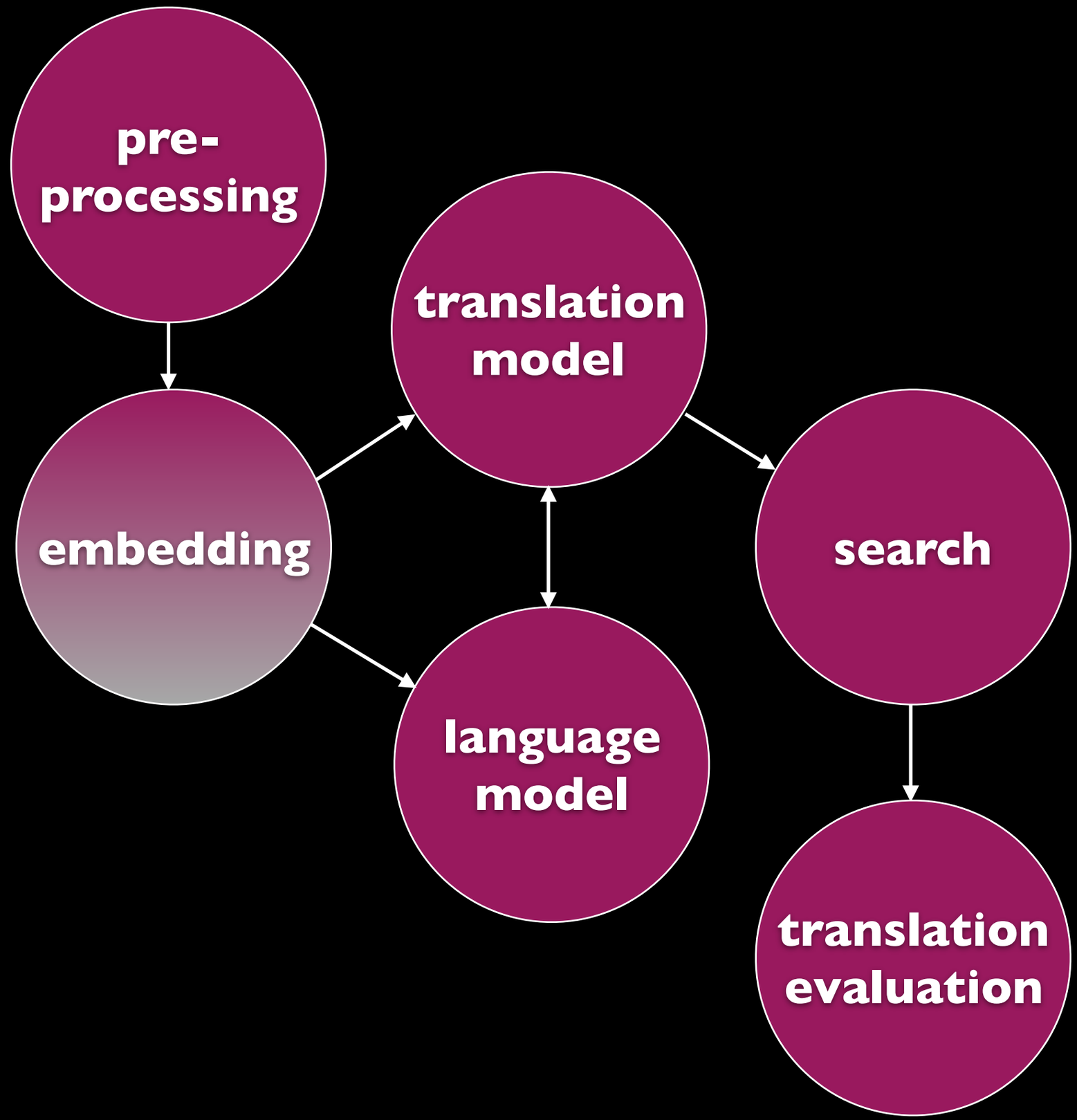
Question #18: named entities?

Byte-Pair-Encoding (BPE)

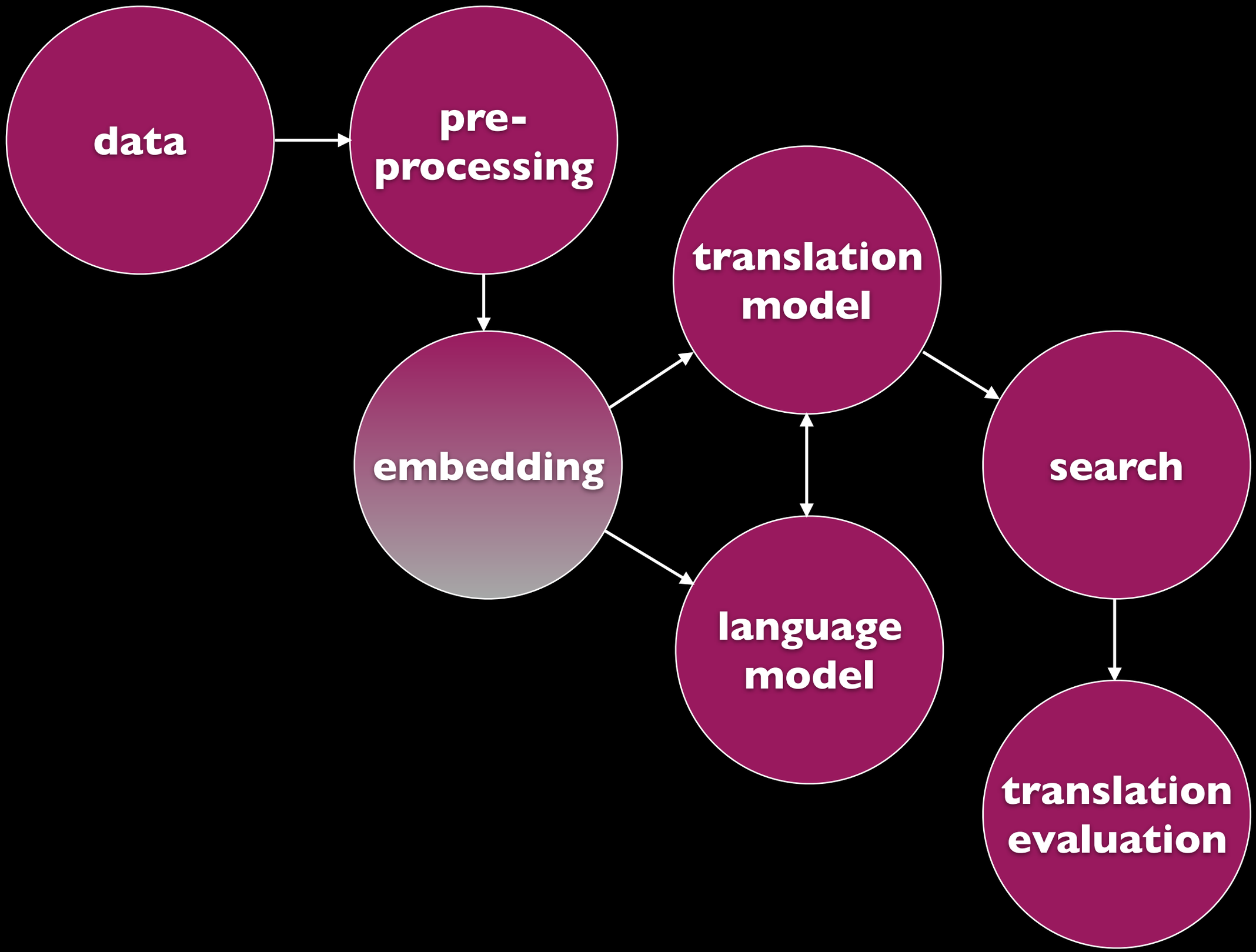
- efficient and affective, NMT independent
- [Gage, 1994], [Sennrich, et.al., 16]
  - aaabdaaabac
  - ZabdZabac (Z=aa)
  - ZYdZYac (Y=ab; Z=aa)
  - XdXac (X=ZY;Y=ab; Z=aa)

machine translation components

machine translation components



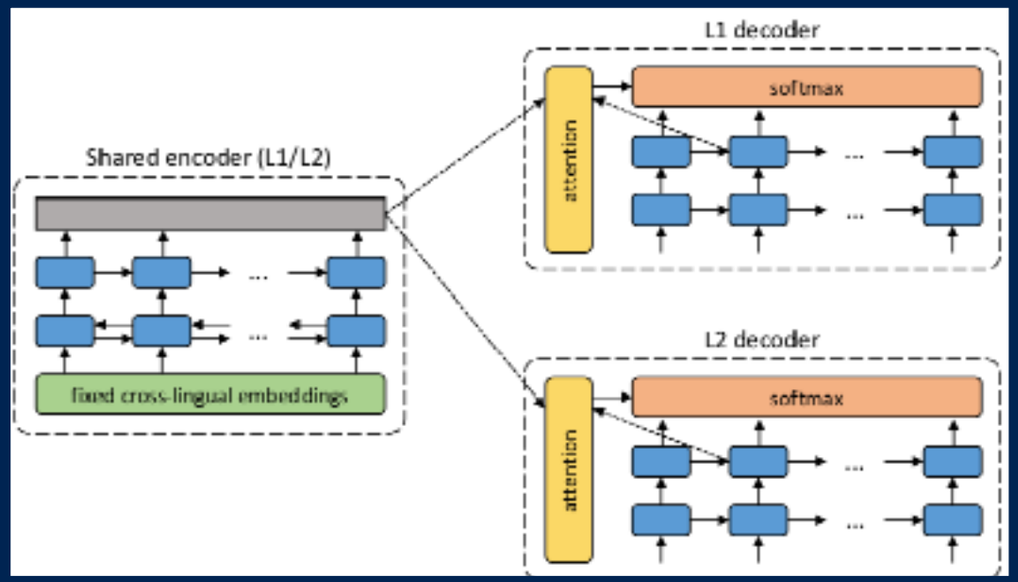
machine translation components



machine translation components

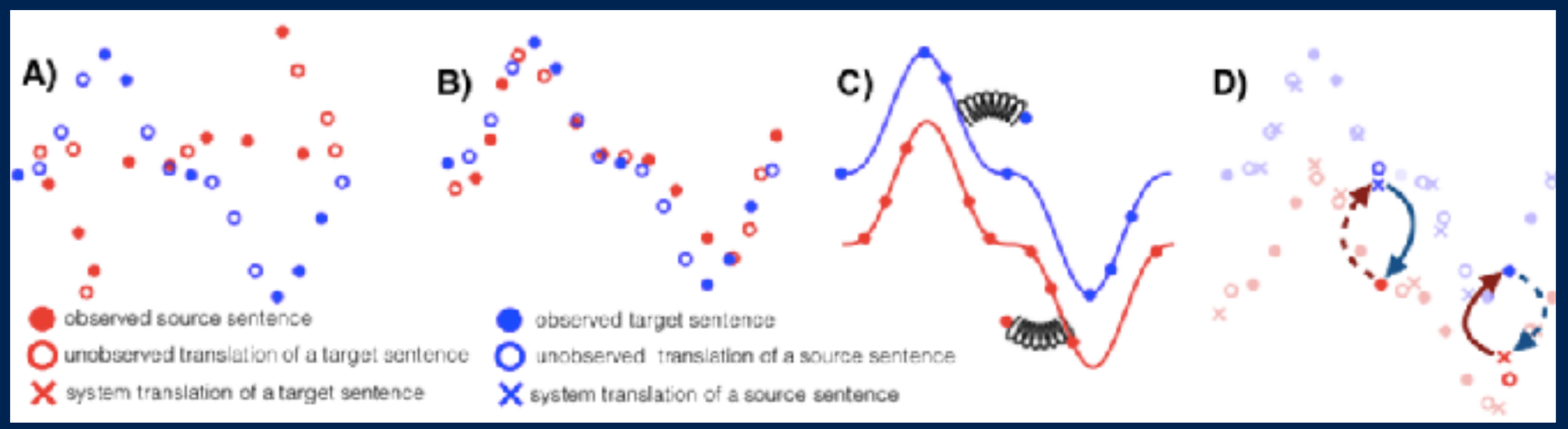
unsupervised NMT

- NO parallel training data possible?
  - [Artetxe, et.al., 17],[Lample, et.al., 16]



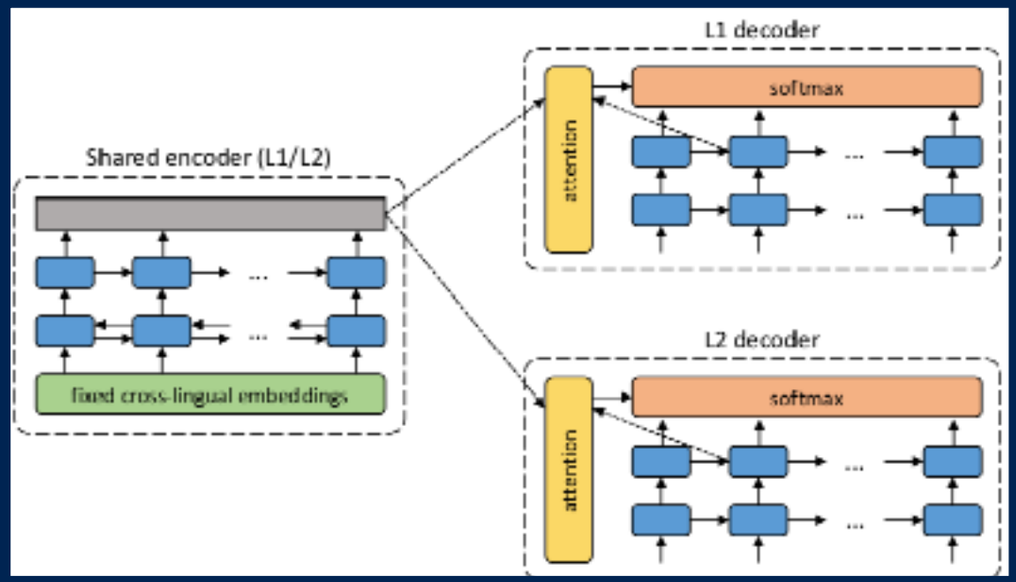
Algorithm 1: Unsupervised MT

1 **Language models:** Learn language models P_s and P_t over source and target languages;
2 **Initial translation models:** Leveraging P_s and P_t, learn two initial translation models, one in each direction: $P_{s\to t}^{(0)}$ and $P_{t\to s}^{(0)}$;
3 **for** k=1 **to** N **do**
4 **Back-translation:** Generate source and target sentences using the current translation models, $P_{t\to s}^{(k-1)}$ and $P_{s\to t}^{(k-1)}$, factoring in language models, P_s and P_t;
5 Train new translation models $P_{s\to t}^{(k)}$ and $P_{t\to s}^{(k)}$ using the generated sentences and leveraging P_s and P_t;
6 **end**



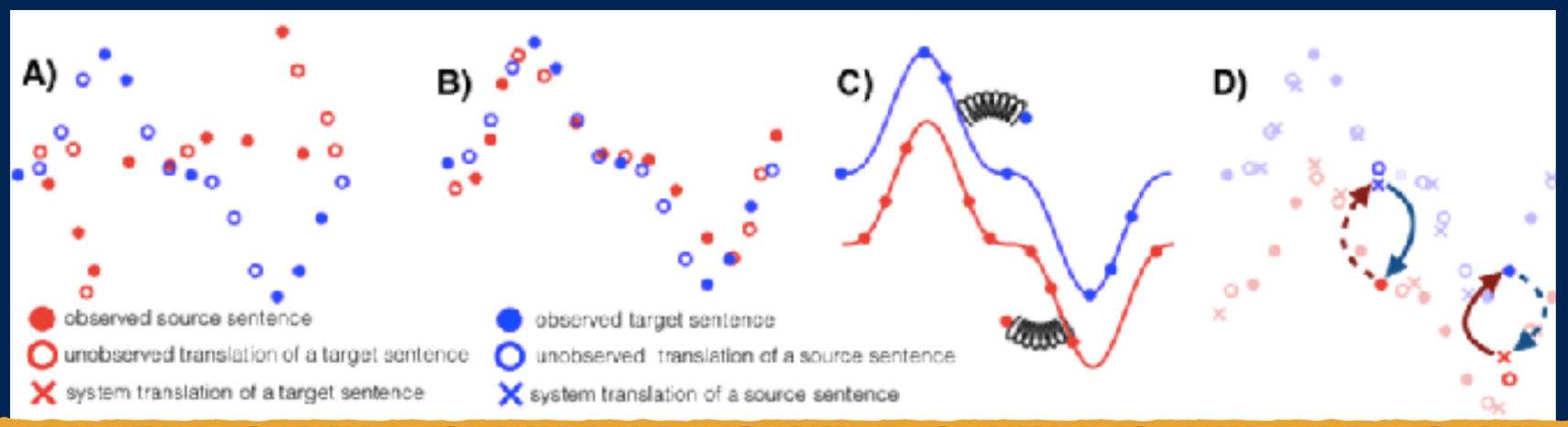
unsupervised NMT

- NO parallel training data possible?
  - [Artetxe, et.al., 17],[Lample, et.al., 16]



Algorithm 1: Unsupervised MT

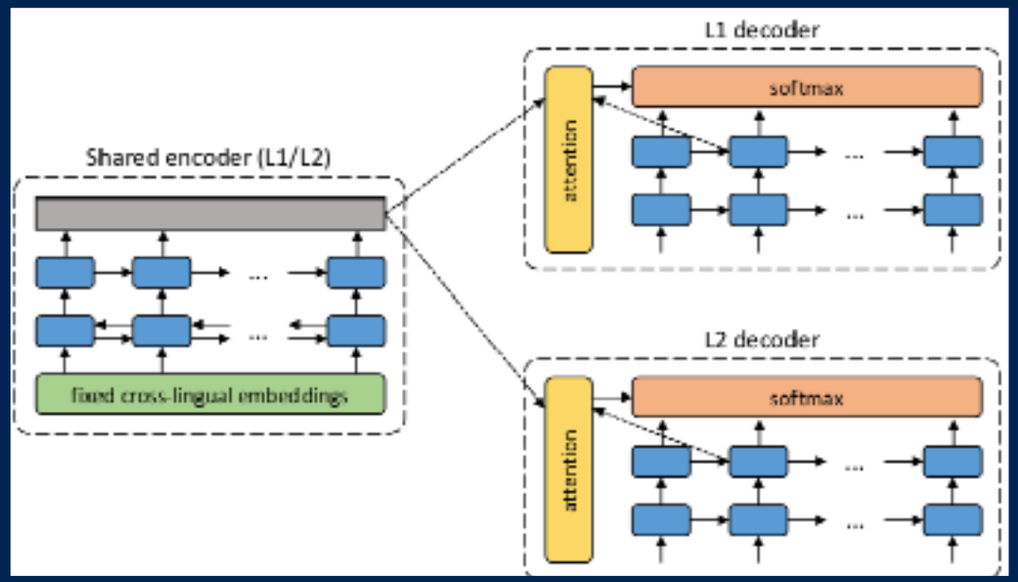
1 **Language models:** Learn language models P_s and P_t over source and target languages;
2 **Initial translation models:** Leveraging P_s and P_t, learn two initial translation models, one in each direction: $P_{s\to t}^{(0)}$ and $P_{t\to s}^{(0)}$;
3 **for** k=1 **to** N **do**
4 **Back-translation:** Generate source and target sentences using the current translation models, $P_{t\to s}^{(k-1)}$ and $P_{s\to t}^{(k-1)}$, factoring in language models, P_s and P_t;
5 Train new translation models $P_{s\to t}^{(k)}$ and $P_{t\to s}^{(k)}$ using the generated sentences and leveraging P_s and P_t;
6 **end**



Question #19: higher quality in unsupervised MT?

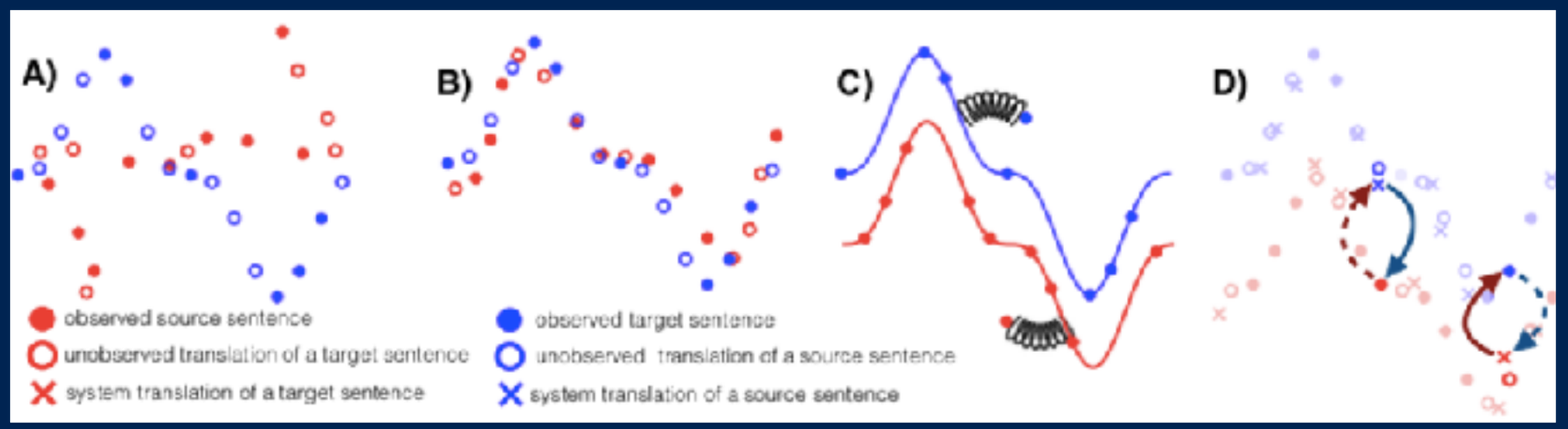
unsupervised NMT

- NO parallel training data possible?
  - [Artetxe, et.al., 17],[Lample, et.al., 16]



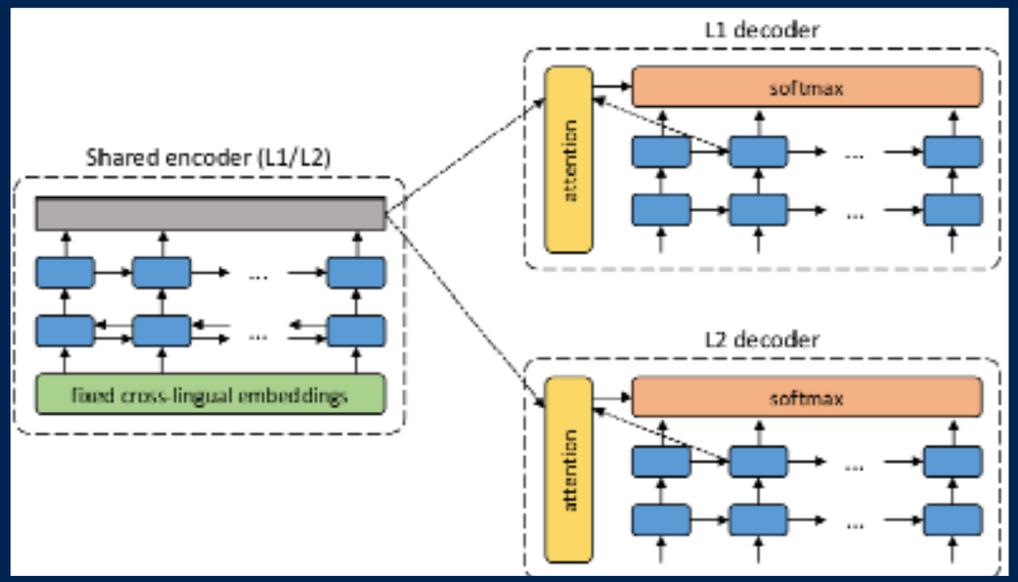
Algorithm 1: Unsupervised MT

1 **Language models:** Learn language models P_s and P_t over source and target languages;
2 **Initial translation models:** Leveraging P_s and P_t, learn two initial translation models, one in each direction: $P_{s\to t}^{(0)}$ and $P_{t\to s}^{(0)}$;
3 **for** k=1 **to** N **do**
4 **Back-translation:** Generate source and target sentences using the current translation models, $P_{t\to s}^{(k-1)}$ and $P_{s\to t}^{(k-1)}$, factoring in language models, P_s and P_t;
5 Train new translation models $P_{s\to t}^{(k)}$ and $P_{t\to s}^{(k)}$ using the generated sentences and leveraging P_s and P_t;
6 **end**



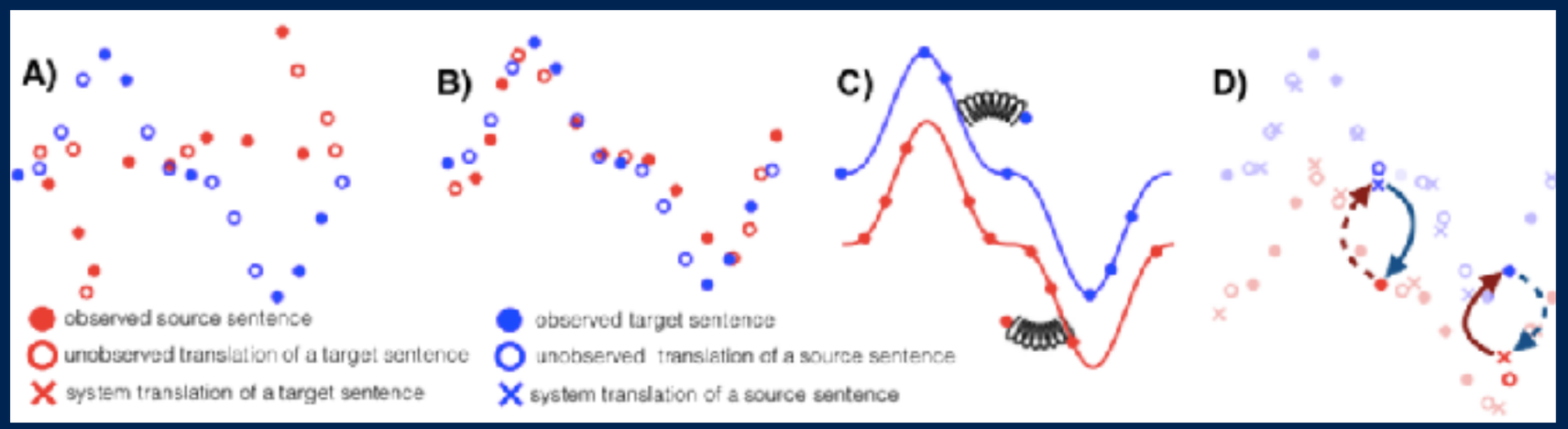
unsupervised NMT

- NO parallel training data possible?
  - [Artetxe, et.al., 17],[Lample, et.al., 16]



Algorithm 1: Unsupervised MT

1 **Language models:** Learn language models P_s and P_t over source and target languages;
2 **Initial translation models:** Leveraging P_s and P_t, learn two initial translation models, one in each direction: $P_{s\to t}^{(0)}$ and $P_{t\to s}^{(0)}$;
3 **for** k=1 **to** N **do**
4 **Back-translation:** Generate source and target sentences using the current translation models, $P_{t\to s}^{(k-1)}$ and $P_{s\to t}^{(k-1)}$, factoring in language models, P_s and P_t;
5 Train new translation models $P_{s\to t}^{(k)}$ and $P_{t\to s}^{(k)}$ using the generated sentences and leveraging P_s and P_t;
6 **end**



back translation

- German to French translation is good
- French to German translation is bad
- use German to French MT system to translate German monolingual data e.g. [Sennrich, et.al.,16]
- add the synthetic data into French to German MT training
  - usually greatly improves translation quality due to data augmentation

back translation

- German to French translation is good
- French to German translation is bad
- use German to French MT system to translate German monolingual data e.g. [Sennrich, et.al.,16]
- add the synthetic data into French to German MT training
  - usually greatly improves translation quality due to data augmentation

Question #20: back translation

back translation

- German to French translation is good
- French to German translation is bad
- use German to French MT system to translate German monolingual data e.g. [Sennrich, et.al.,16]
- add the synthetic data into French to German MT training
  - usually greatly improves translation quality due to data augmentation

back translation

- German to French translation is good
- French to German translation is bad
- use German to French MT system to translate German monolingual data e.g. [Sennrich, et.al.,16]
- add the synthetic data into French to German MT training
  - usually greatly improves translation quality due to data augmentation

pivot translation

- lack of German - French parallel training data
- rich data of German - English, and French - English
- generate German - French parallel data using German - English and French - English MT systems
- add the synthetic data into French to German MT training
  - usually greatly improves translation quality

pivot translation

- lack of German - French parallel training data
- rich data of German - English, and French - English
- generate German - French parallel data using German - English and French - English MT systems
- add the synthetic data into French to German MT training
  - usually greatly improves translation quality

Question #21: pivot translation

pivot translation

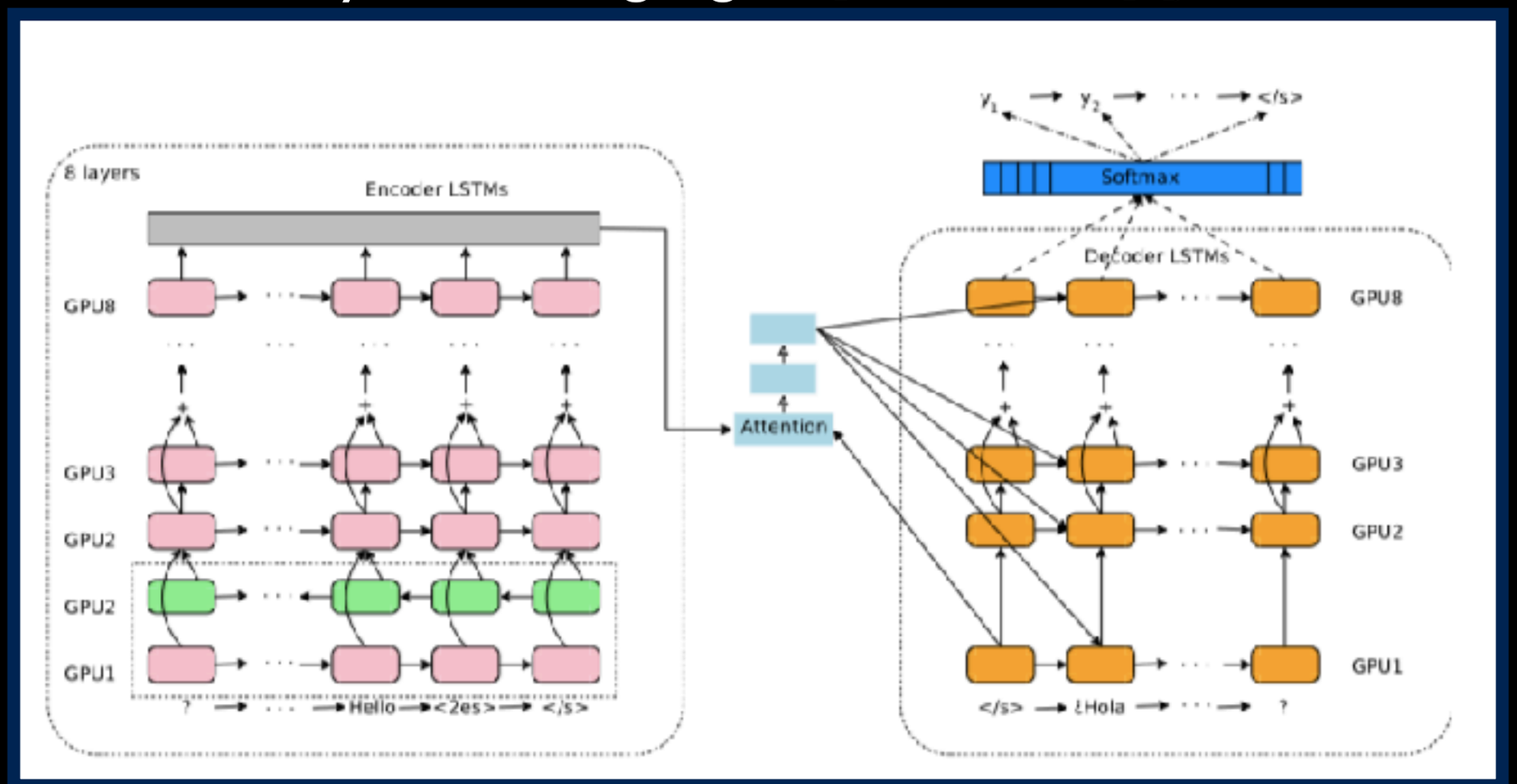
- lack of German - French parallel training data
- rich data of German - English, and French - English
- generate German - French parallel data using German - English and French - English MT systems
- add the synthetic data into French to German MT training
  - usually greatly improves translation quality

pivot translation

- lack of German - French parallel training data
- rich data of German - English, and French - English
- generate German - French parallel data using German - English and French - English MT systems
- add the synthetic data into French to German MT training
  - usually greatly improves translation quality

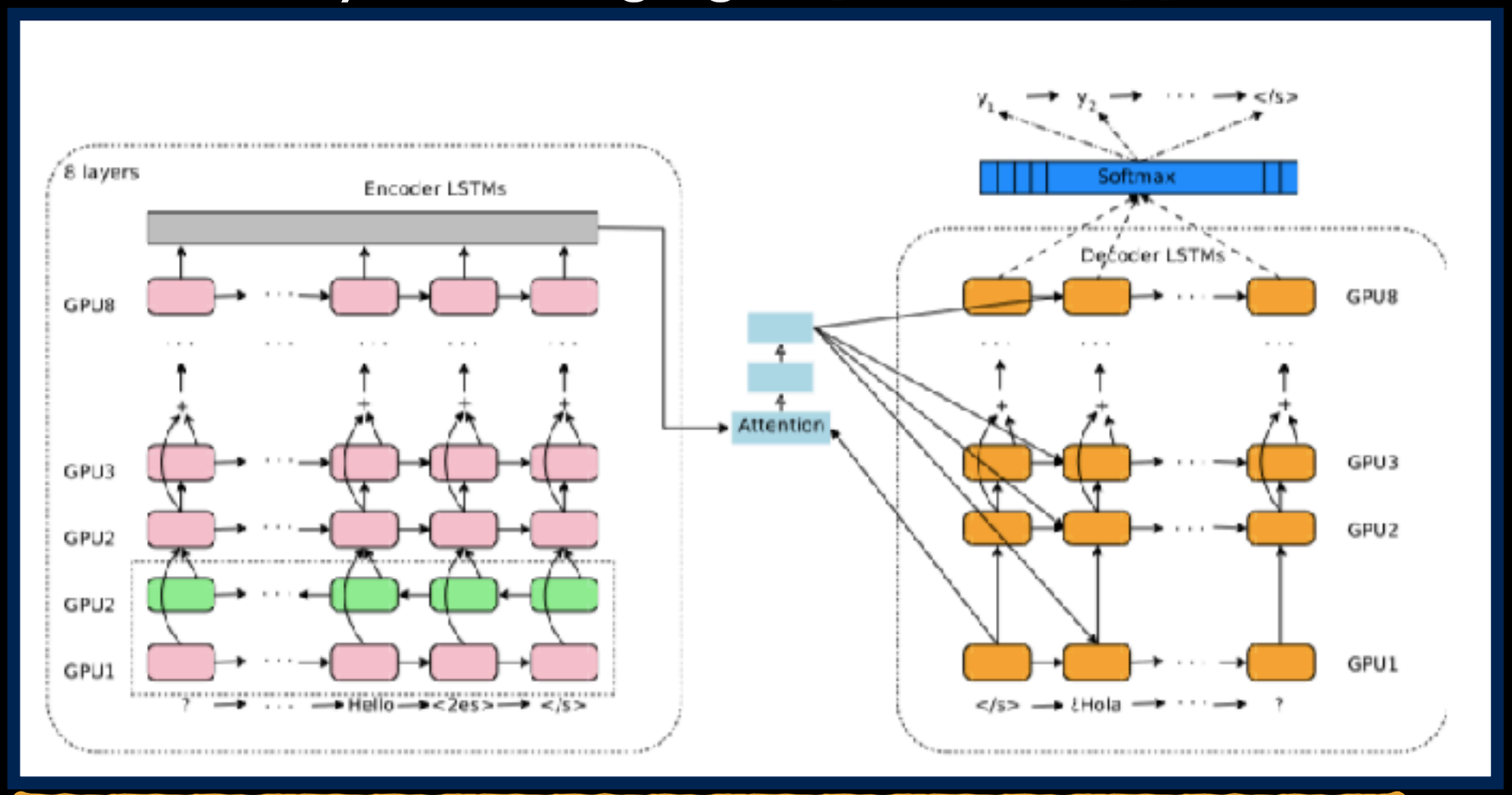
multiple languages

- observation in WMT'19: adding Hindi - English parallel data improves Gujarati - English translation
- how about many other languages? [Firat, et.al., 16], [Johnson, et.al., 17]



multiple languages

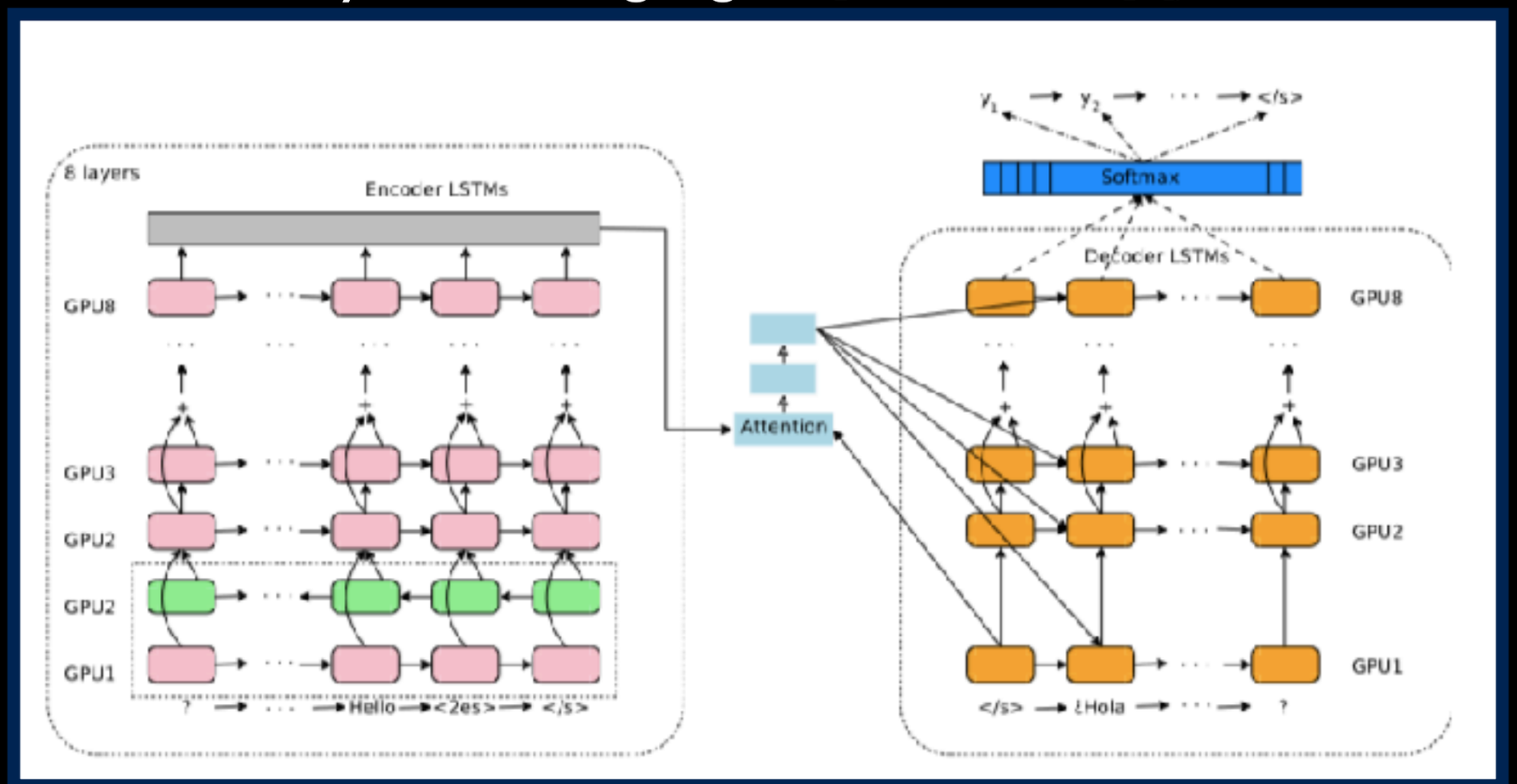
- observation in WMT'19: adding Hindi - English parallel data improves Gujarati - English translation
- how about many other languages? [Firat, et.al., 16], [Johnson, et.al., 17]



Question #22: multi-lingual and zero resource

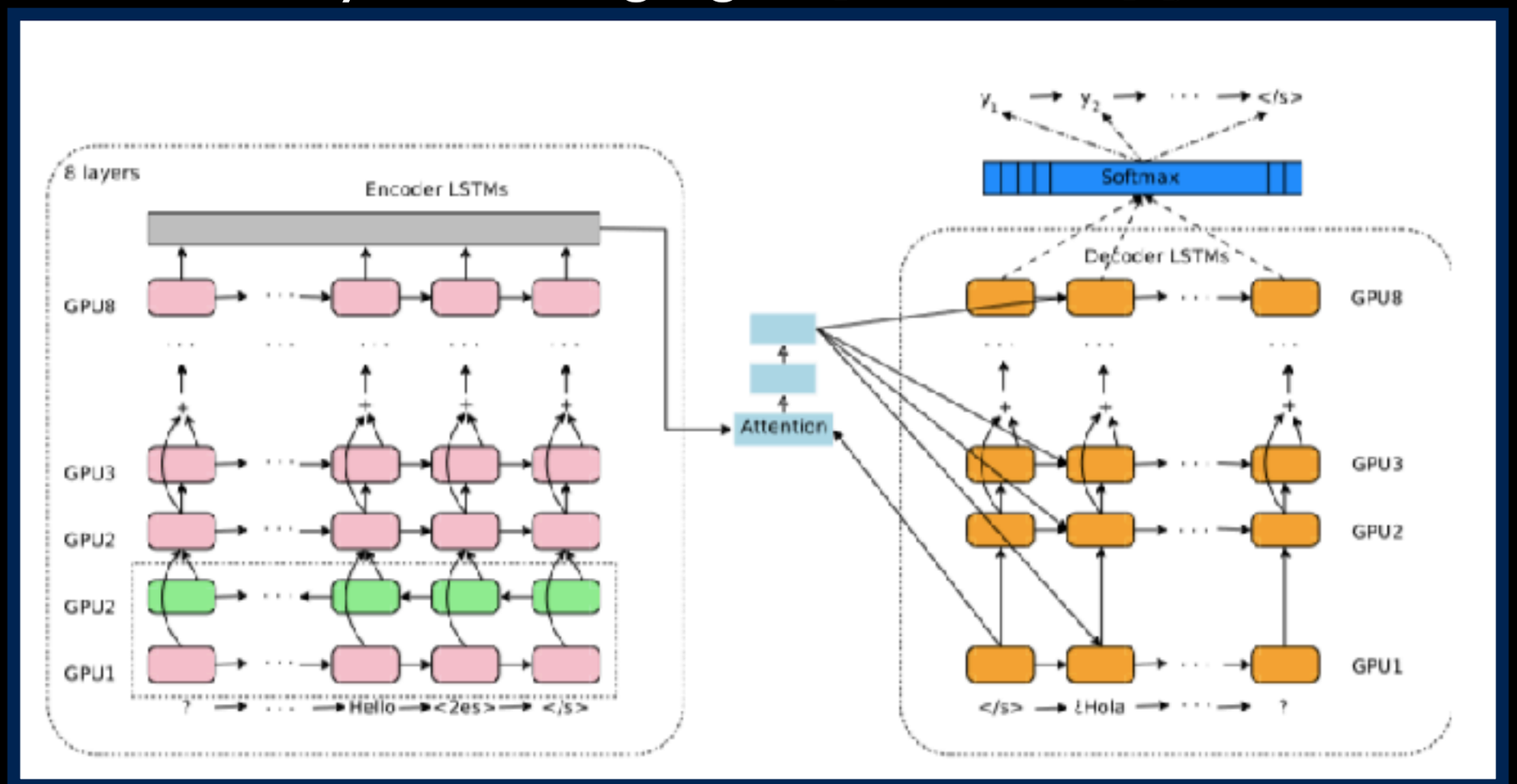
multiple languages

- observation in WMT'19: adding Hindi - English parallel data improves Gujarati - English translation
- how about many other languages? [Firat, et.al., 16], [Johnson, et.al., 17]

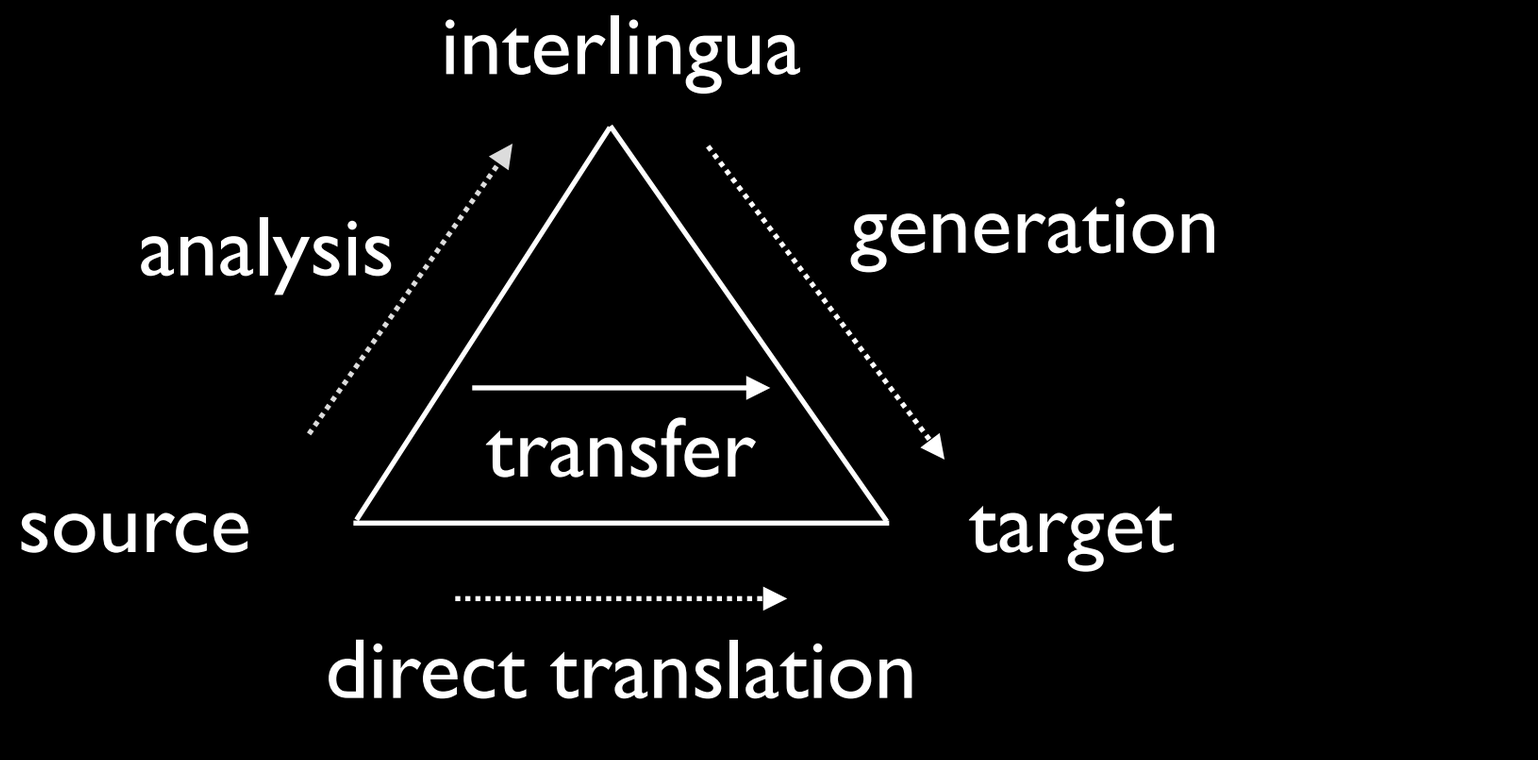


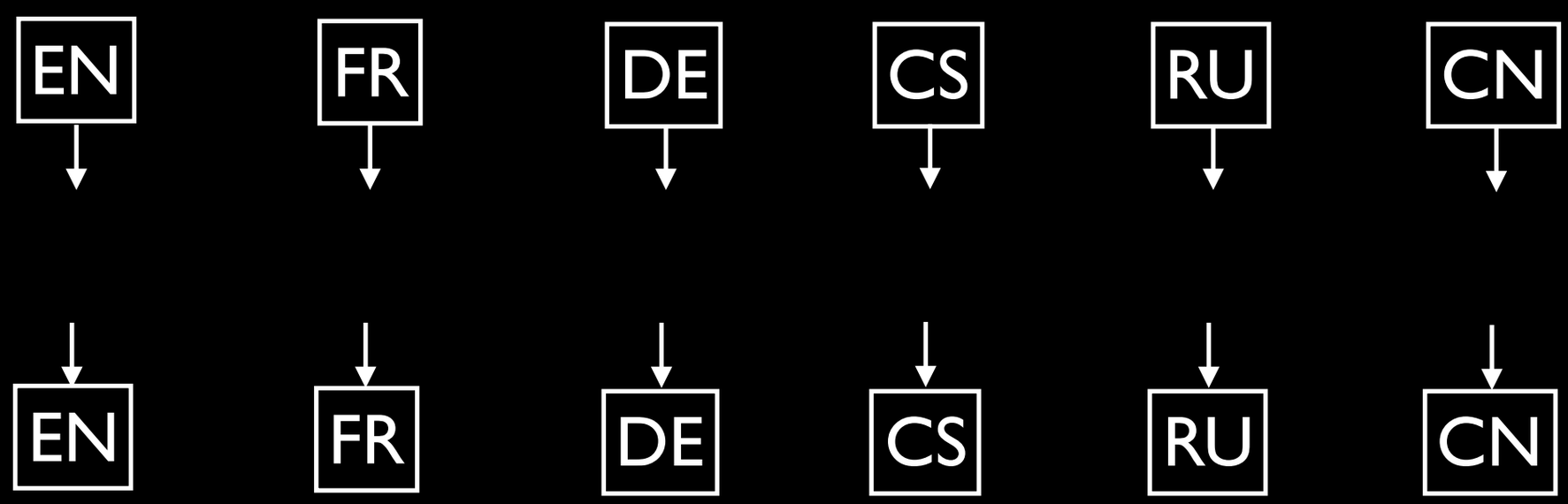
multiple languages

- observation in WMT'19: adding Hindi - English parallel data improves Gujarati - English translation
- how about many other languages? [Firat, et.al., 16], [Johnson, et.al., 17]

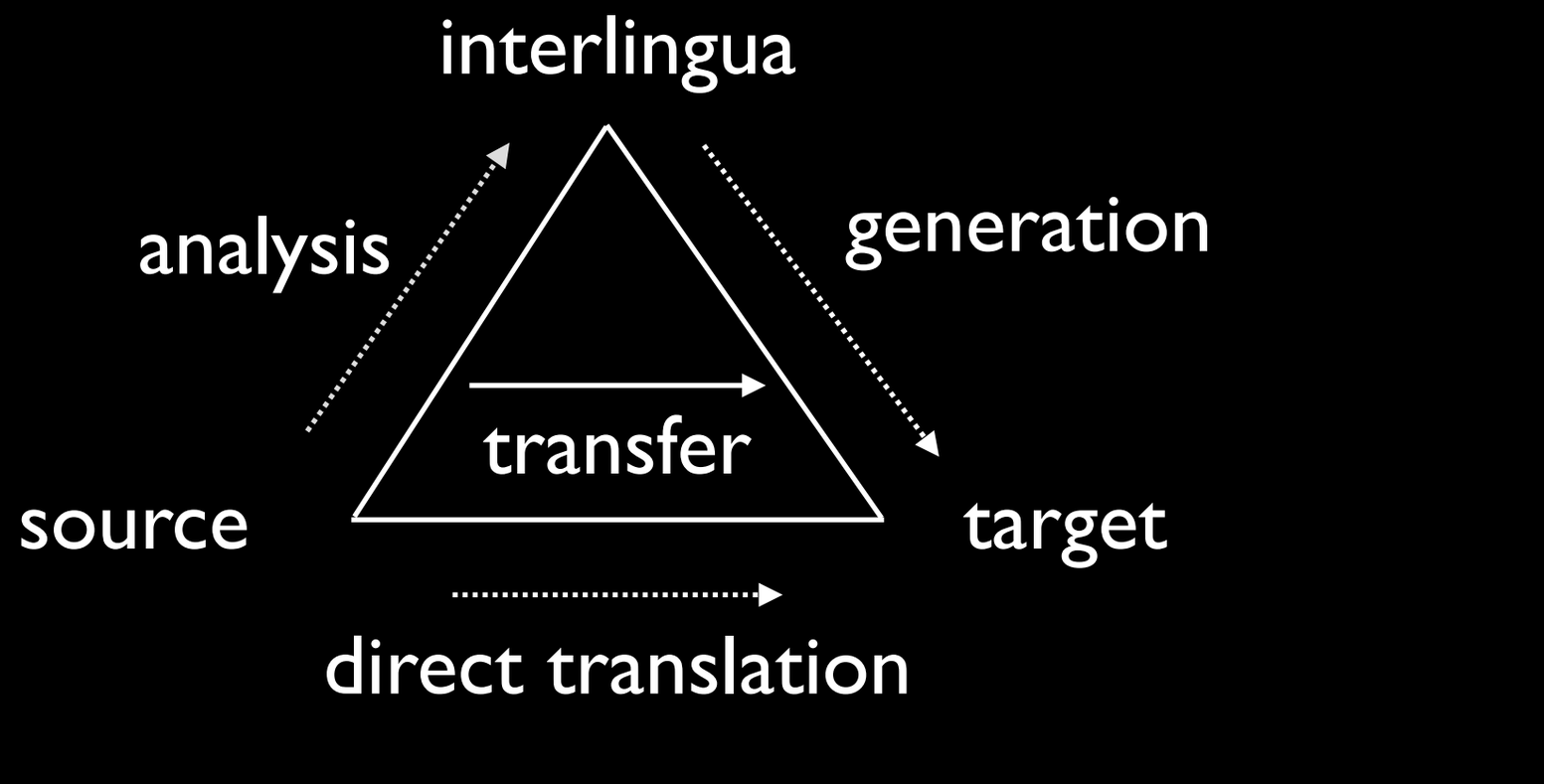


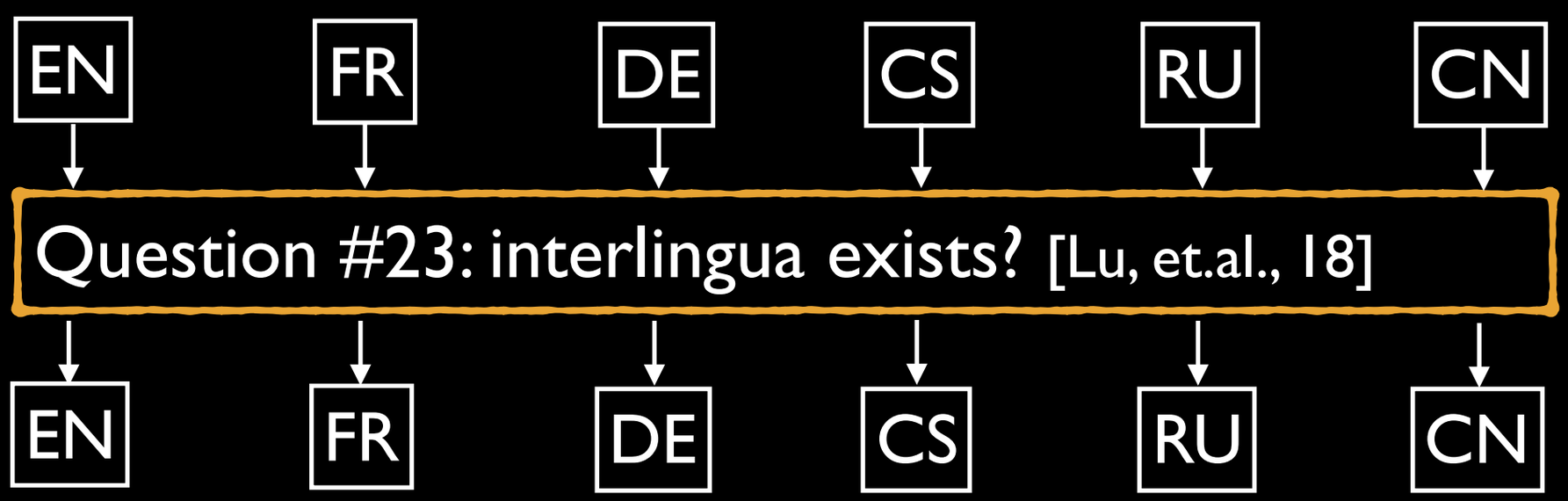
the concept of interlingua



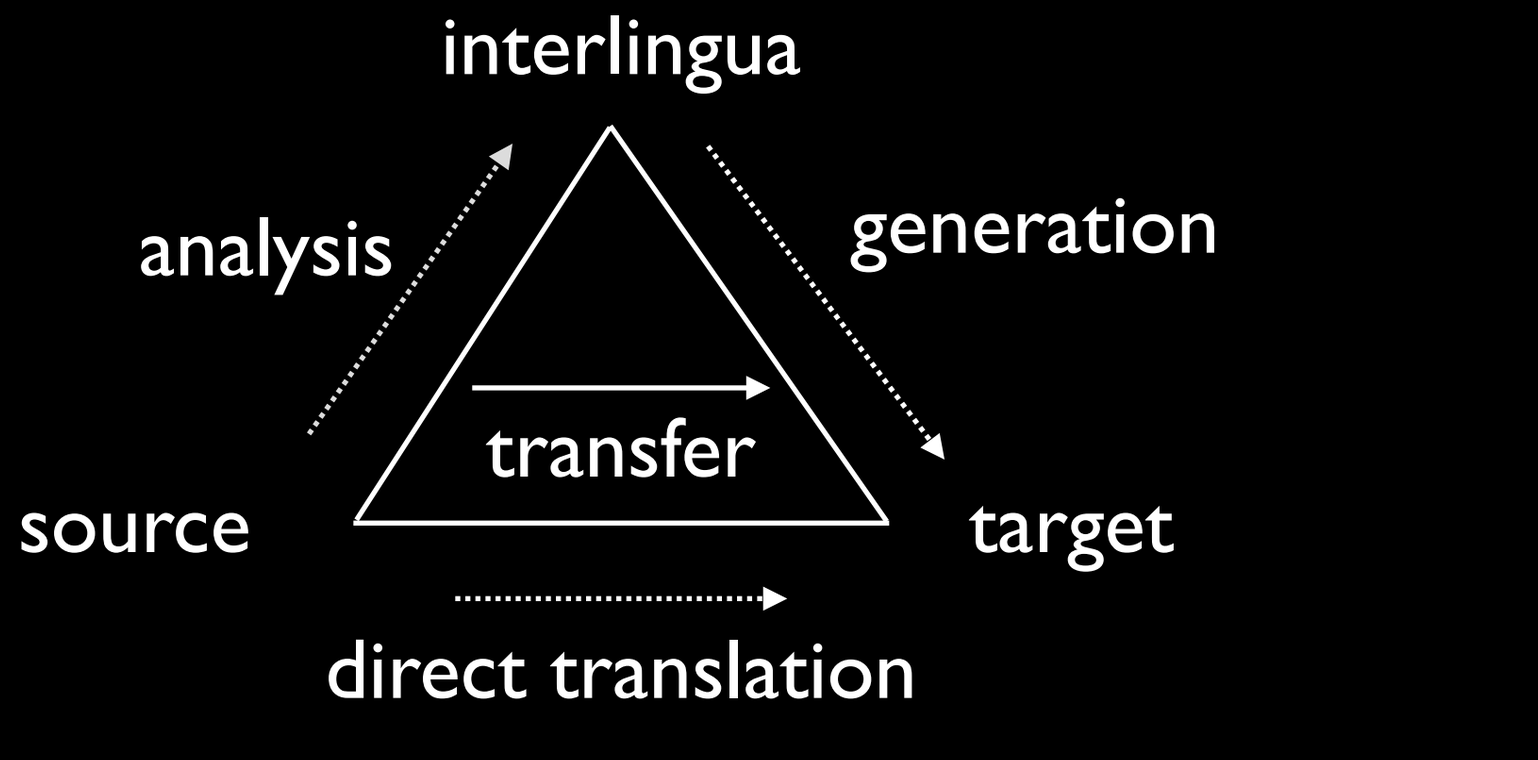


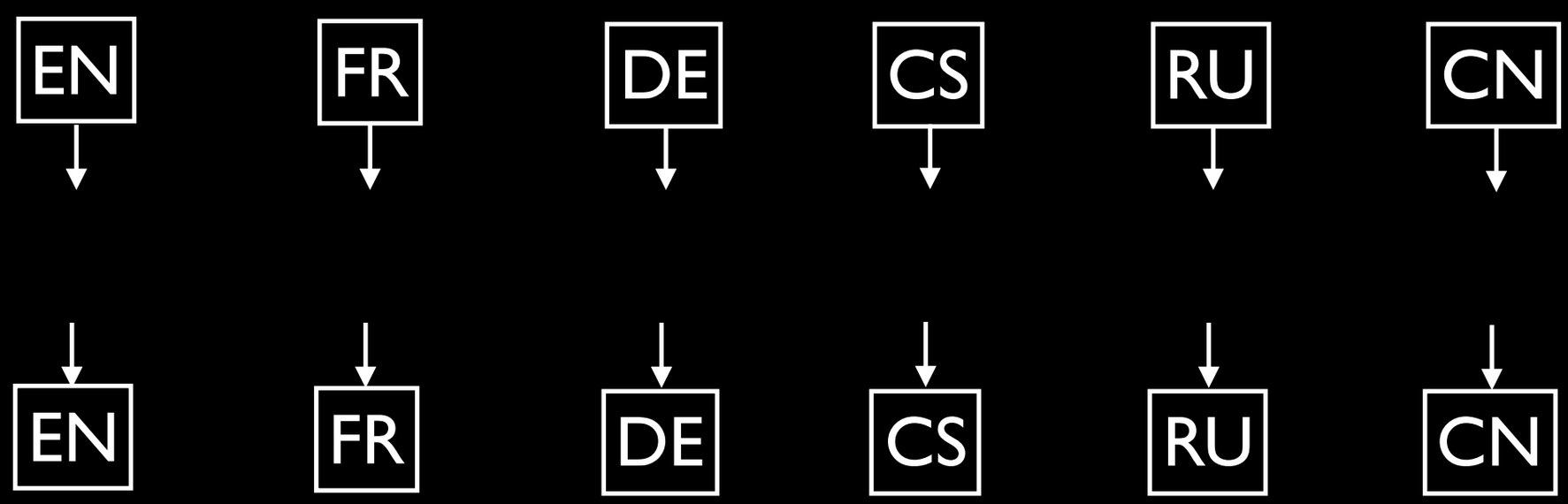
the concept of interlingua



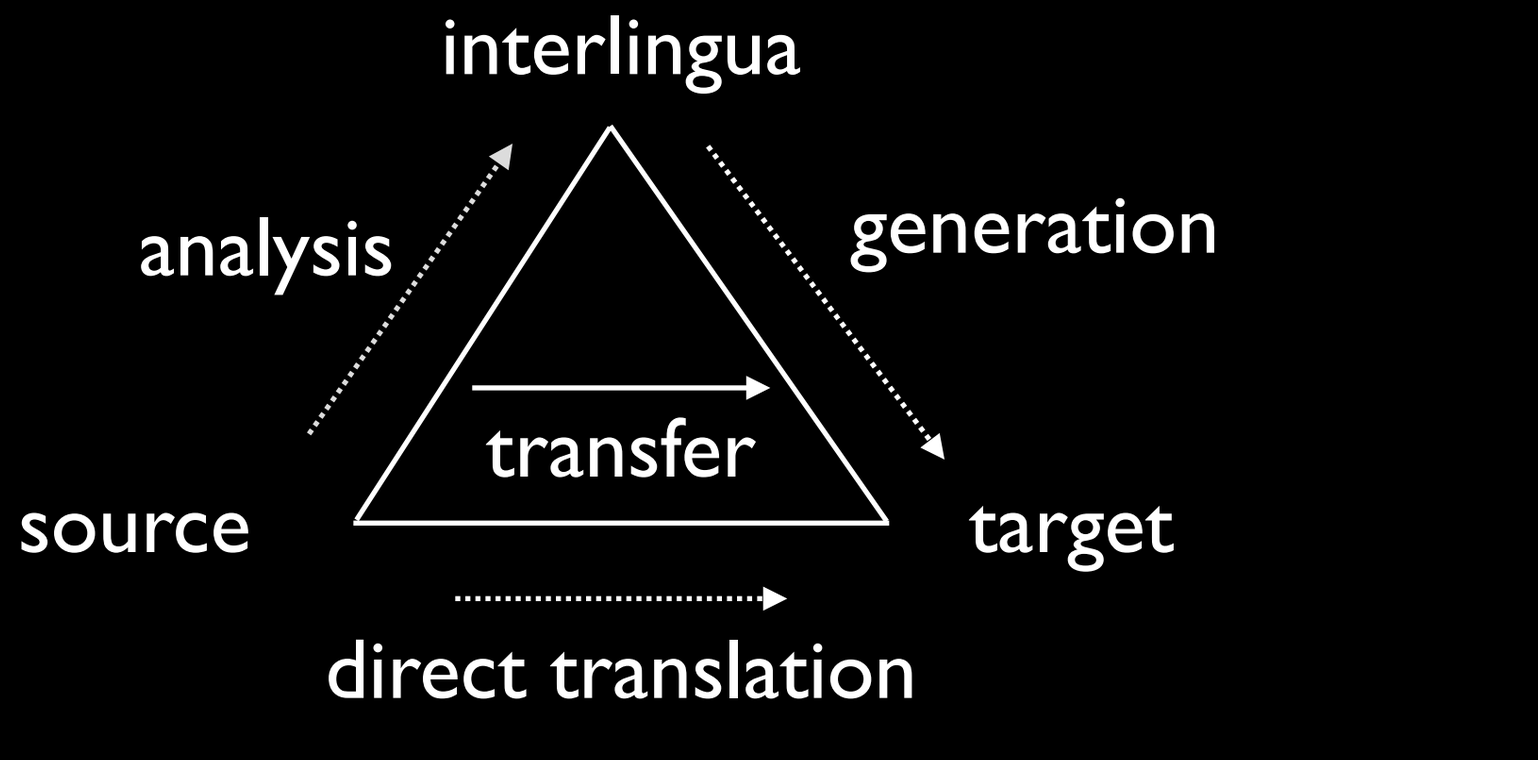


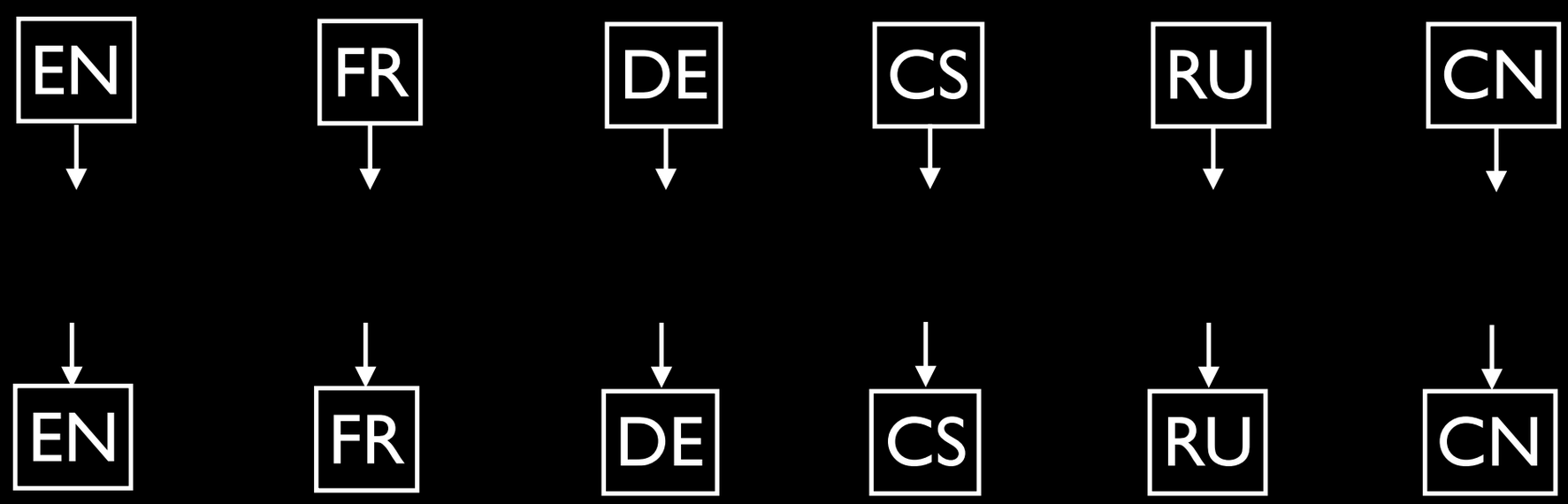
the concept of interlingua





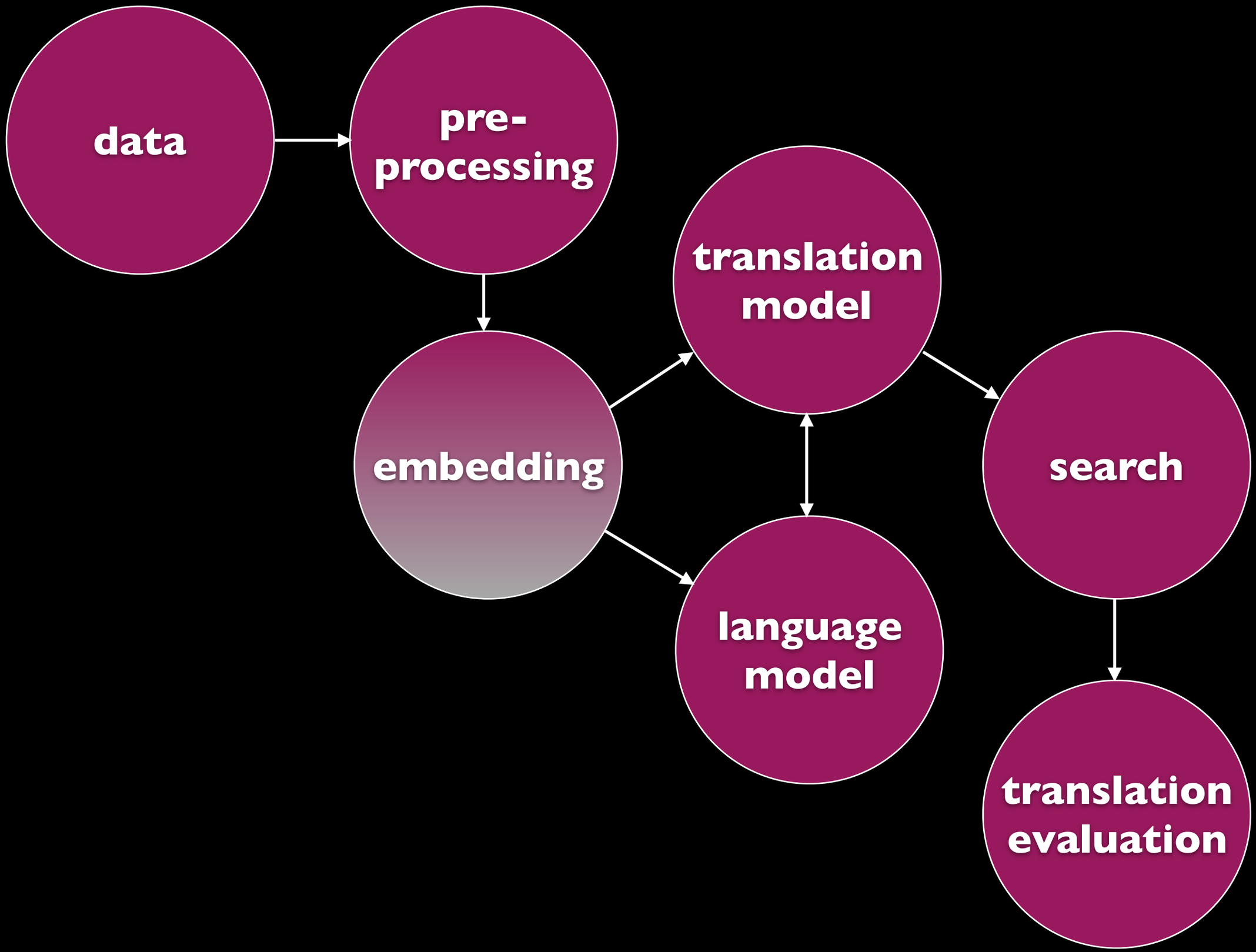
the concept of interlingua



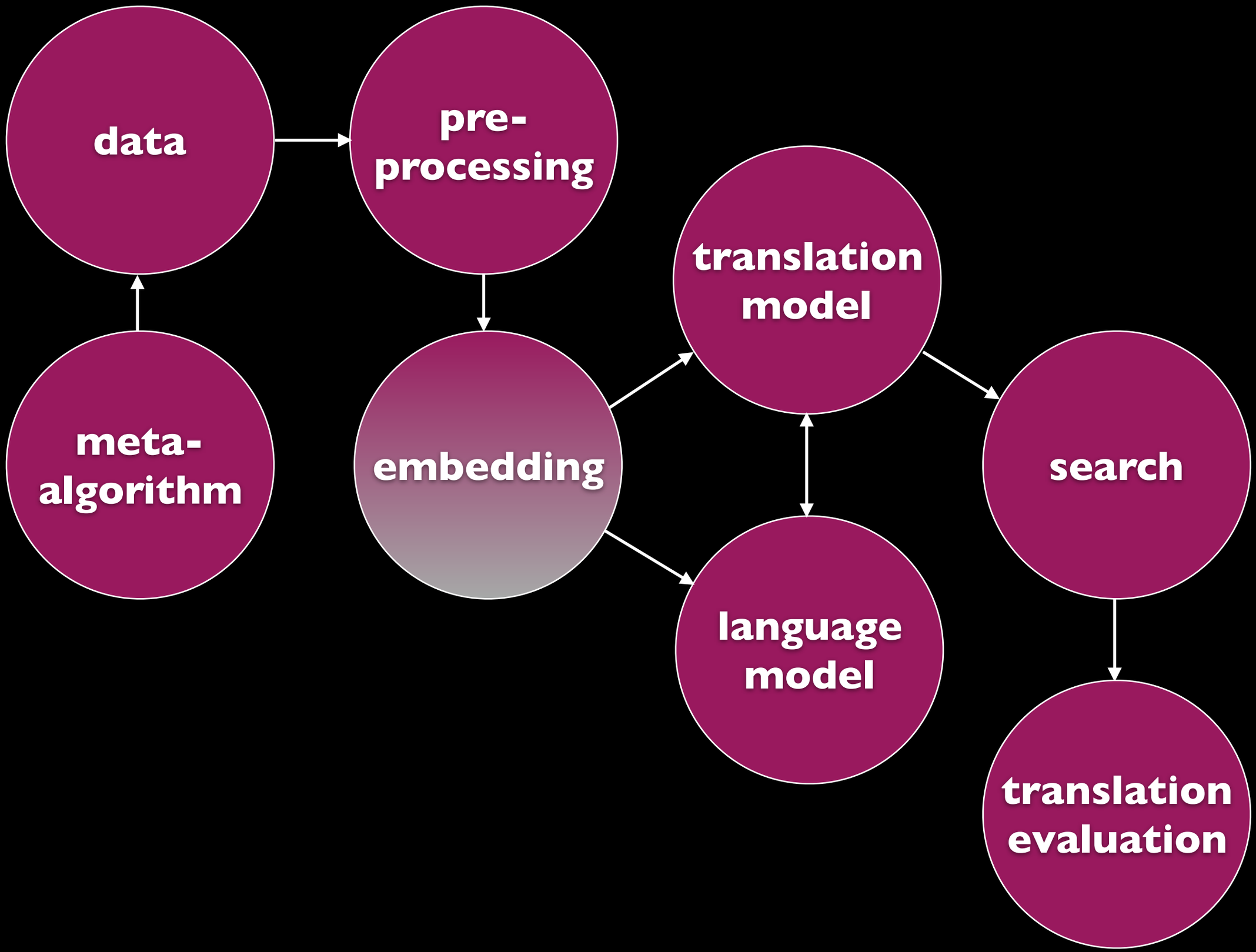


machine translation components

machine translation components



machine translation components

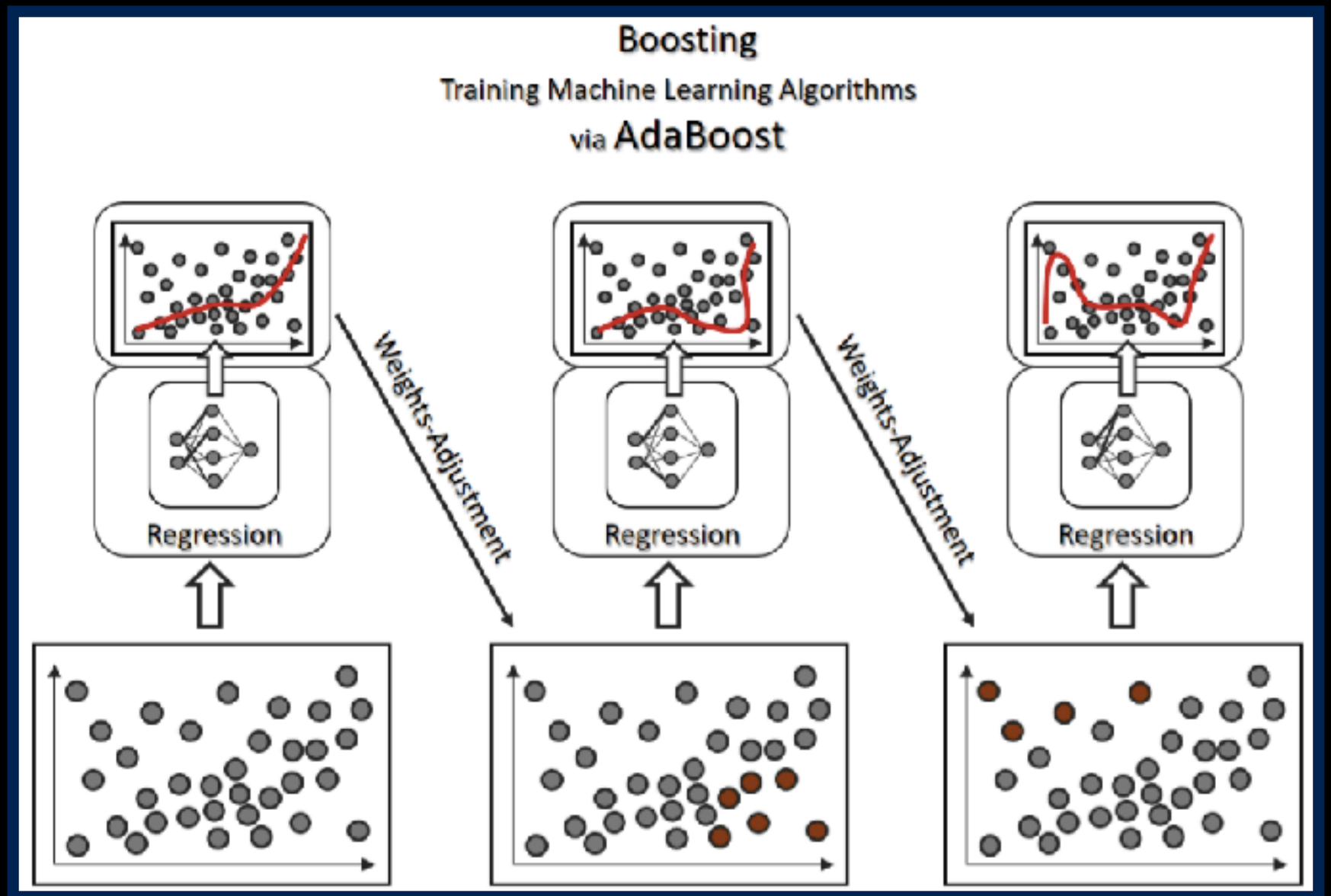


machine translation components

meta-algorithm

boosting

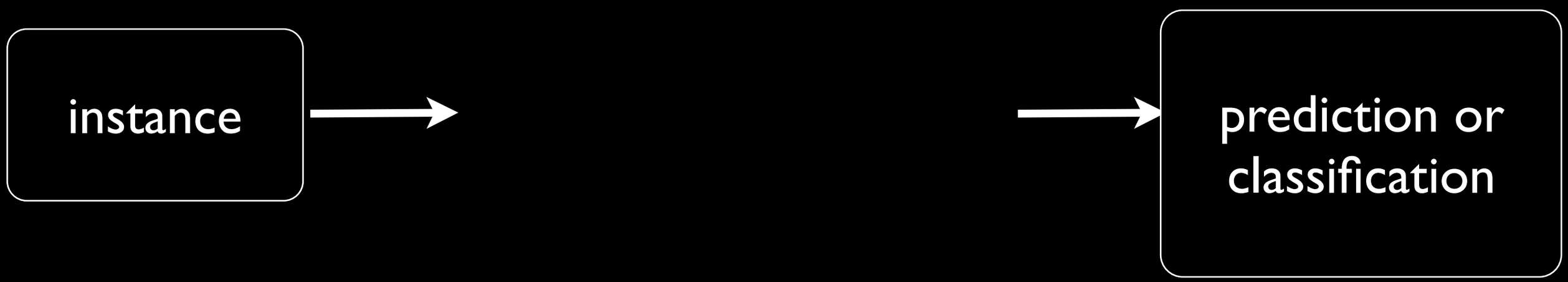
- improves prediction accuracy; non-parallelizable



subsamples generate multiple classifiers

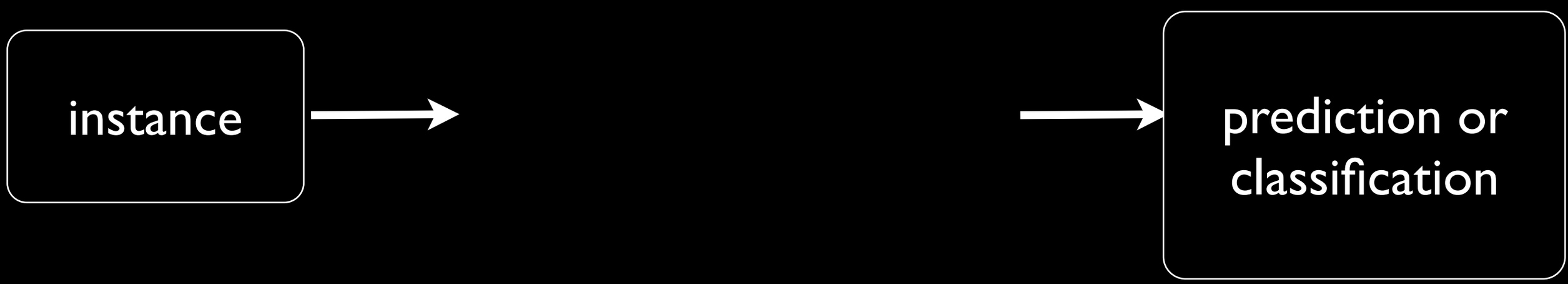
Natürliche Lebensräume wurden zerstört.
Dies ist eine ihrer Hauptaufgaben.
Das kann so nicht weitergehen.
Die Aussprache ist geschlossen.

Natural habitats were destroyed.
This is a major task.
That cannot continue.
That concludes the debate.



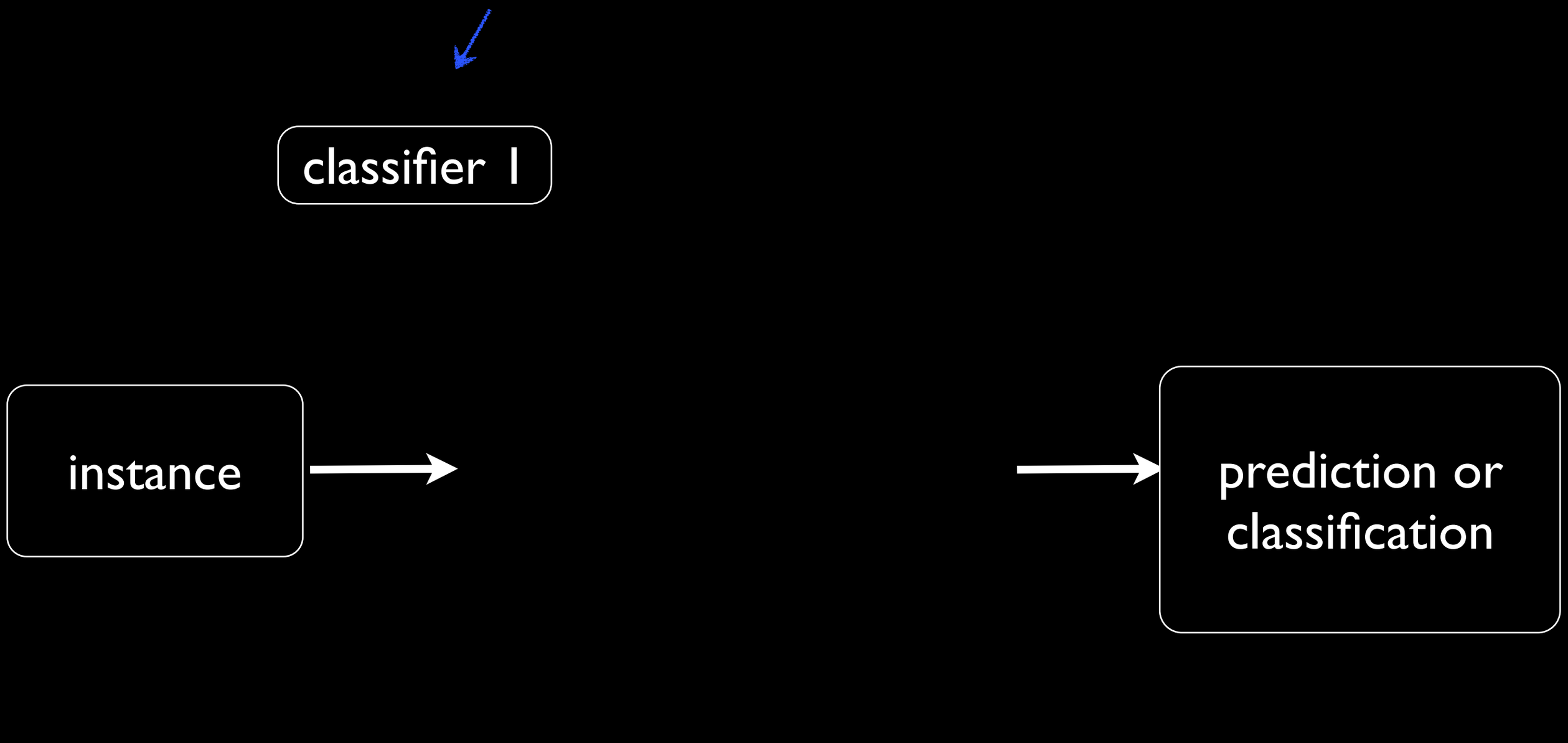
subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.	Natural habitats were destroyed.
Dies ist eine ihrer Hauptaufgaben.	This is a major task.
Das kann so nicht weitergehen.	That cannot continue.
Die Aussprache ist geschlossen.	That concludes the debate.



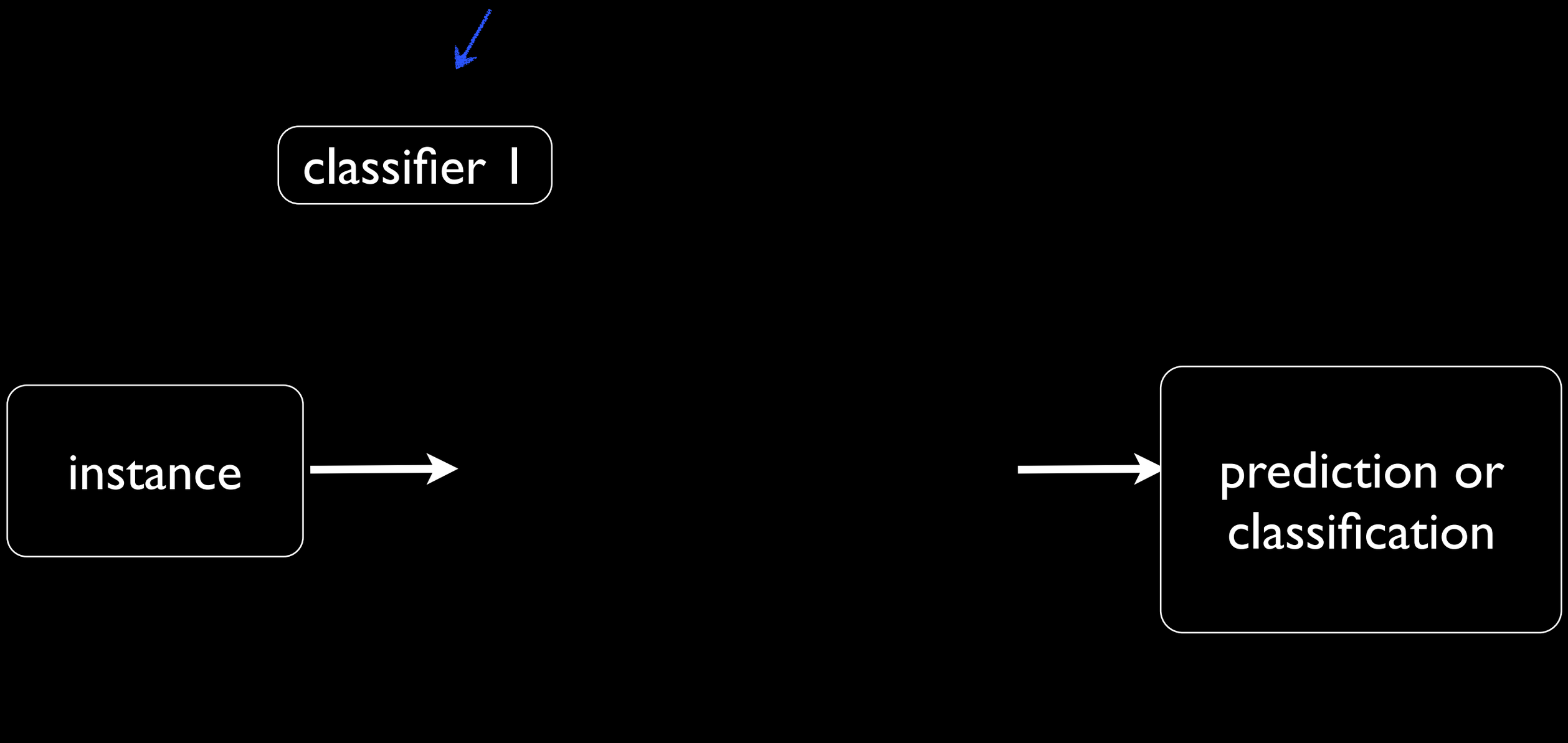
subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.	Natural habitats were destroyed.
Dies ist eine ihrer Hauptaufgaben.	This is a major task.
Das kann so nicht weitergehen.	That cannot continue.
Die Aussprache ist geschlossen.	That concludes the debate.



subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.	Natural habitats were destroyed.
Dies ist eine ihrer Hauptaufgaben.	This is a major task.
Das kann so nicht weitergehen.	That cannot continue.
Die Aussprache ist geschlossen.	That concludes the debate.

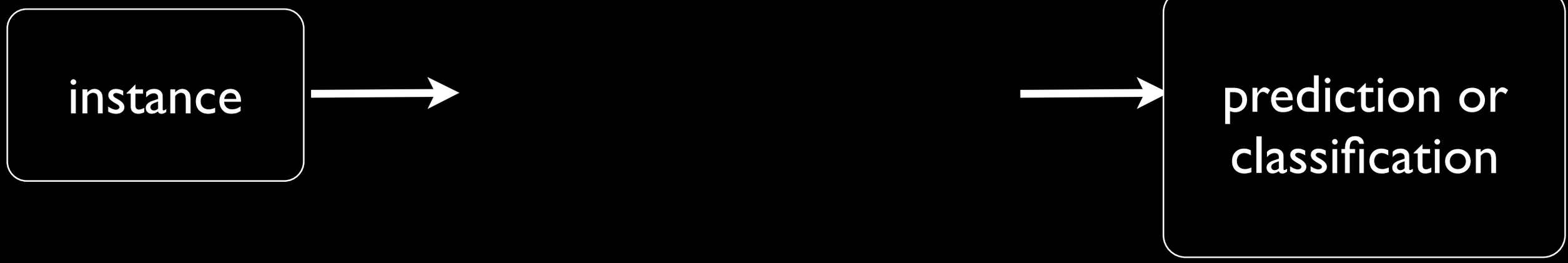


subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.	Natural habitats were destroyed.
Dies ist eine ihrer Hauptaufgaben.	This is a major task.
Das kann so nicht weitergehen.	That cannot continue.
Die Aussprache ist geschlossen.	That concludes the debate.

classifier 1

classifier 2

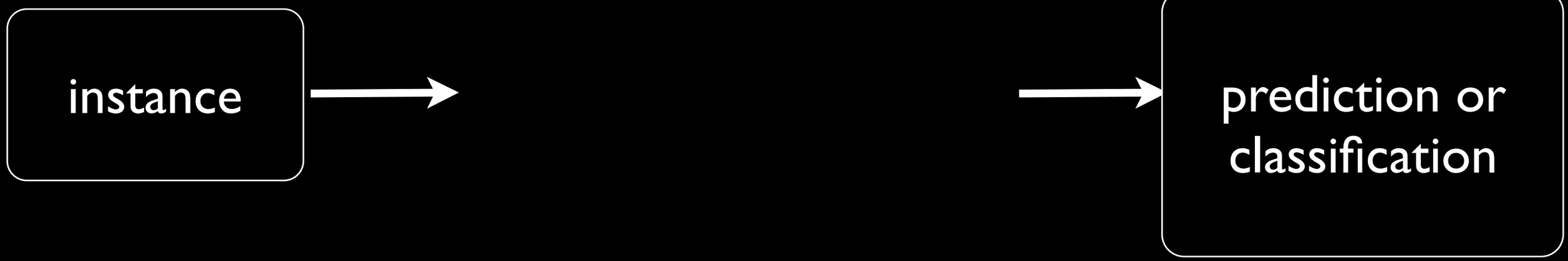


subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.	Natural habitats were destroyed.
Dies ist eine ihrer Hauptaufgaben.	This is a major task.
Das kann so nicht weitergehen.	That cannot continue.
Die Aussprache ist geschlossen.	That concludes the debate.

classifier 1

classifier 2



subsamples generate multiple classifiers

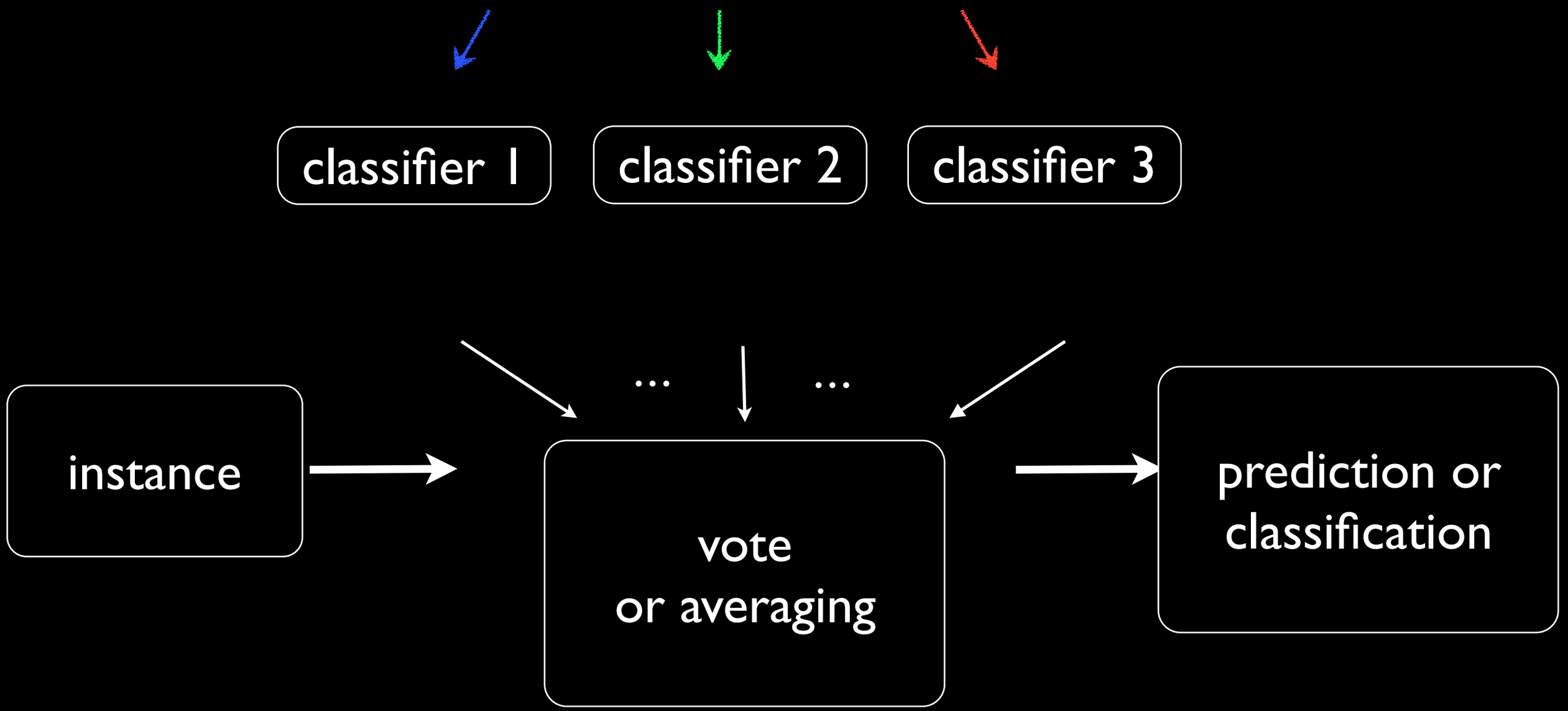
Natürliche Lebensräume wurden zerstört.	Natural habitats were destroyed.
Dies ist eine ihrer Hauptaufgaben.	This is a major task.
Das kann so nicht weitergehen.	That cannot continue.
Die Aussprache ist geschlossen.	That concludes the debate.

classifier 1 | classifier 2 | classifier 3

instance → prediction or classification

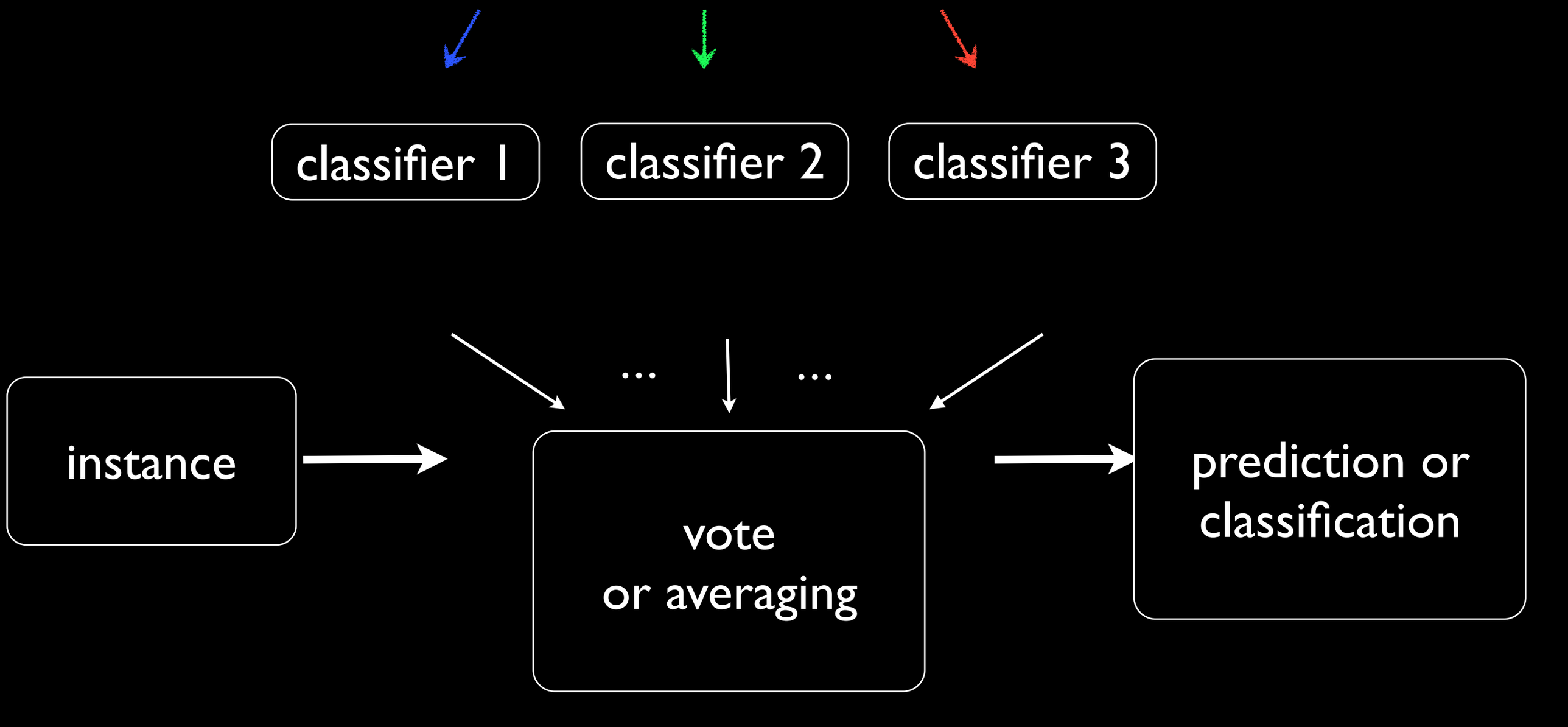
subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.	Natural habitats were destroyed.
Dies ist eine ihrer Hauptaufgaben.	This is a major task.
Das kann so nicht weitergehen.	That cannot continue.
Die Aussprache ist geschlossen.	That concludes the debate.



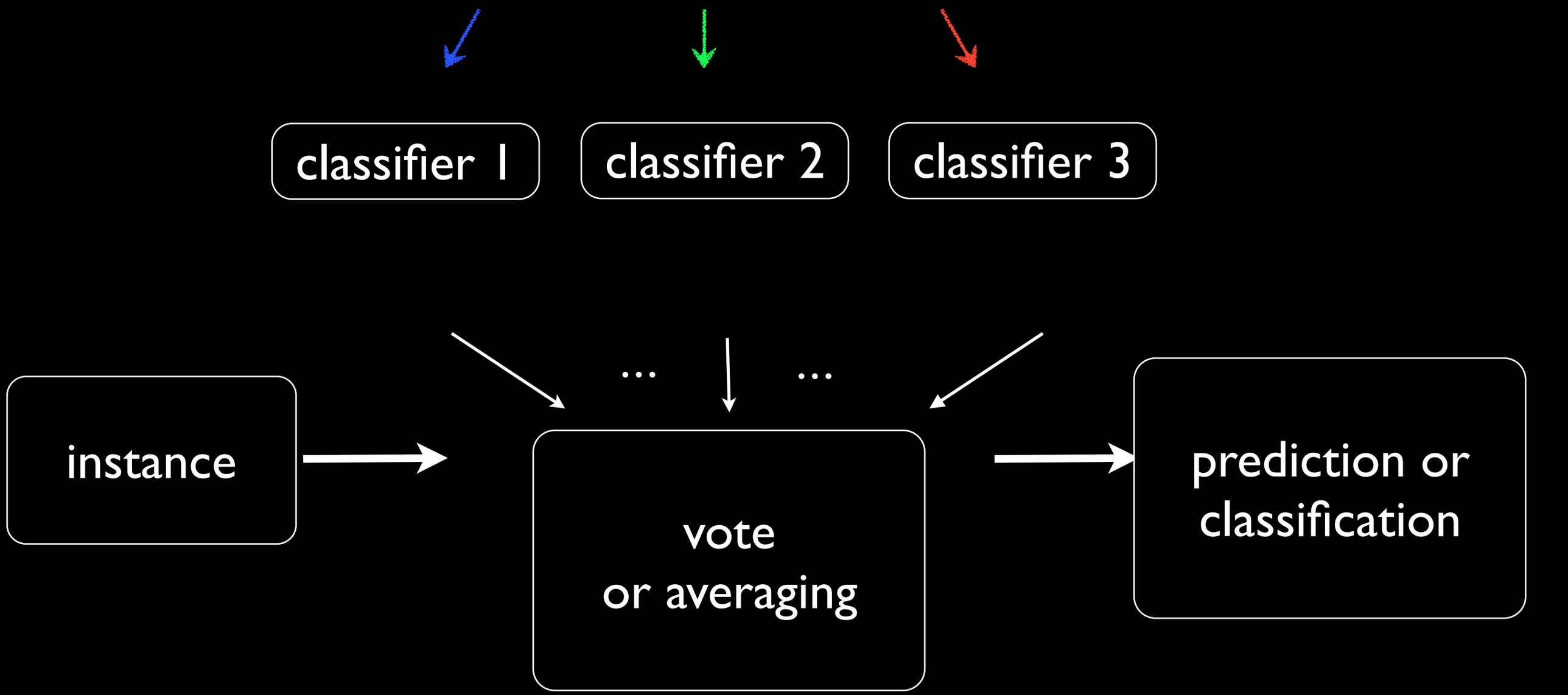
subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.	Natural habitats were destroyed.
Dies ist eine ihrer Hauptaufgaben.	This is a major task.
Das kann so nicht weitergehen.	That cannot continue.
Die Aussprache ist geschlossen.	That concludes the debate.



subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.	Natural habitats were destroyed.
Dies ist eine ihrer Hauptaufgaben.	This is a major task.
Das kann so nicht weitergehen.	That cannot continue.
Die Aussprache ist geschlossen.	That concludes the debate.



subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.	Natural habitats were destroyed.
Dies ist eine ihrer Hauptaufgaben.	This is a major task.
Das kann so nicht weitergehen.	That cannot continue.
Die Aussprache ist geschlossen.	That concludes the debate.

different ways to subsample

classifier 1 | classifier 2 | classifier 3

instance

...

vote
or averaging

prediction or classification

subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört. — Natural habitats were destroyed.
Dies ist eine ihrer Hauptaufgaben. — This is a major task.
Das kann so nicht weitergehen. — That cannot continue.
Die Aussprache ist geschlossen. — That concludes the debate.

different ways to subsample

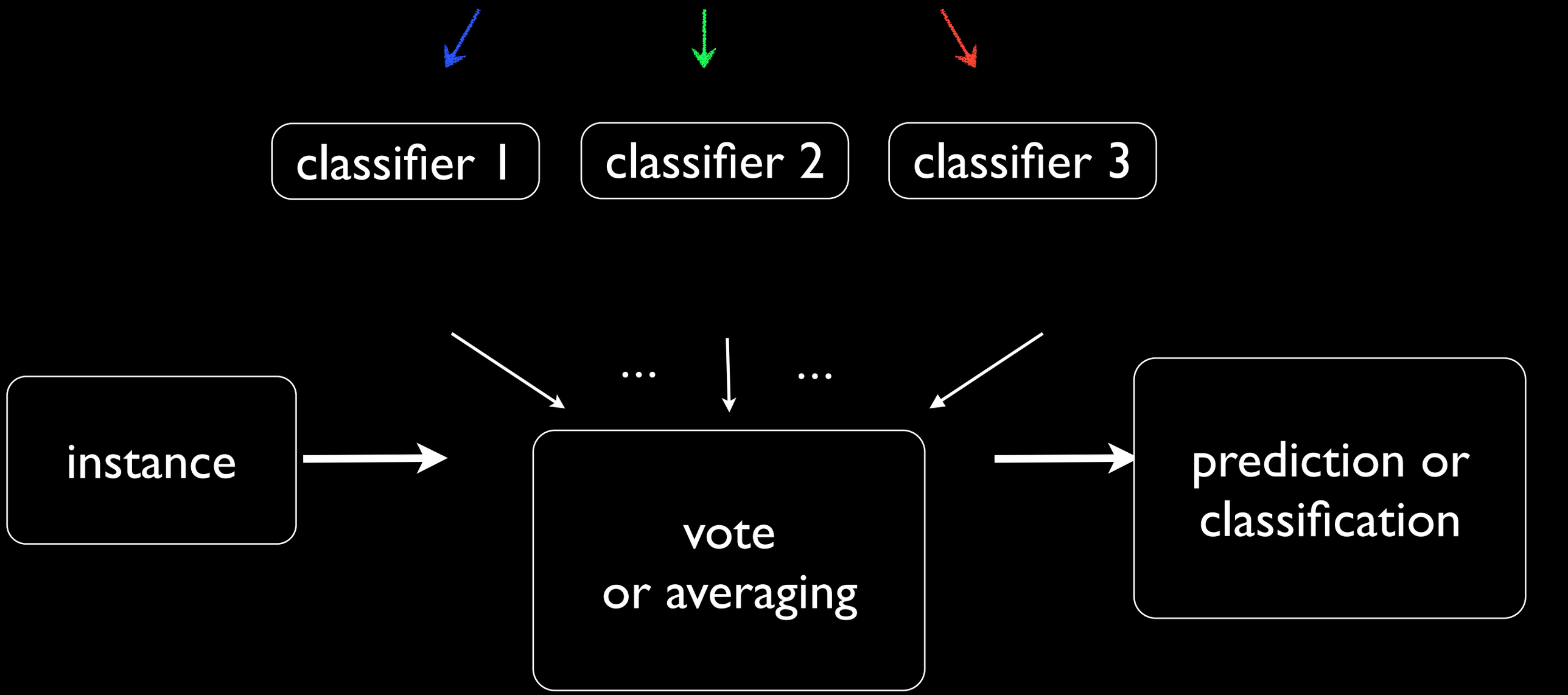
classifier 1 | classifier 2 | classifier 3

better subsampling, better translation?
analyze with abstract formulation for all applications, algorithms, datasets and domains

instance → vote or averaging → prediction or classification

subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.	Natural habitats were destroyed.
Dies ist eine ihrer Hauptaufgaben.	This is a major task.
Das kann so nicht weitergehen.	That cannot continue.
Die Aussprache ist geschlossen.	That concludes the debate.



subsamples generate multiple classifiers

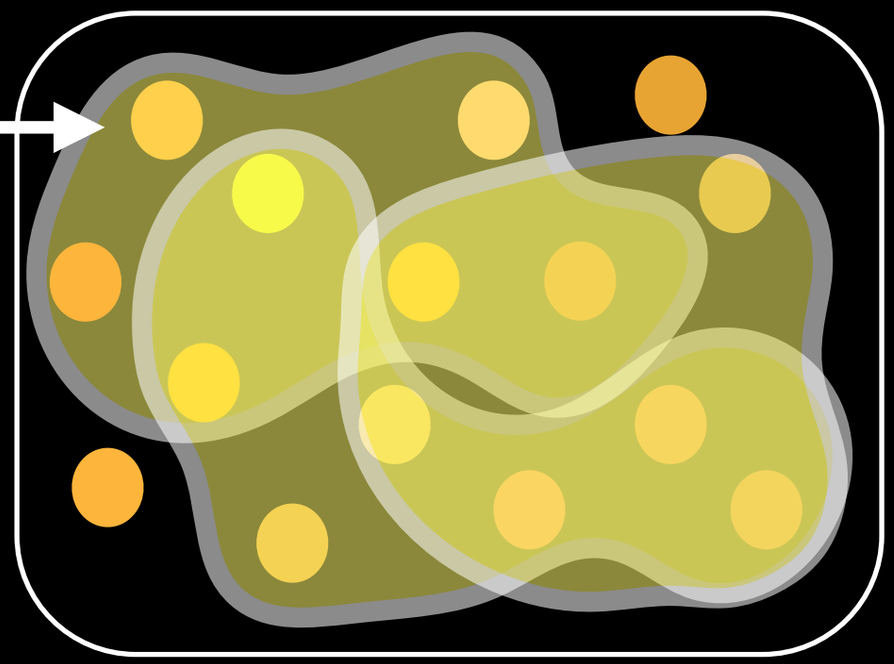
each sample ● is
a *parallel sentence*

Natürliche Lebensräume wurden zerstört.	Natural habitats were destroyed.
Dies ist eine ihrer Hauptaufgaben.	This is a major task.
Das kann so nicht weitergehen.	That cannot continue.
Die Aussprache ist geschlossen.	That concludes the debate.

classifier 1 | classifier 2 | classifier 3

instance → … … → vote or averaging → prediction or classification

a sample:
instance+class



classifier 1

classifier 2

classifier 3

instance

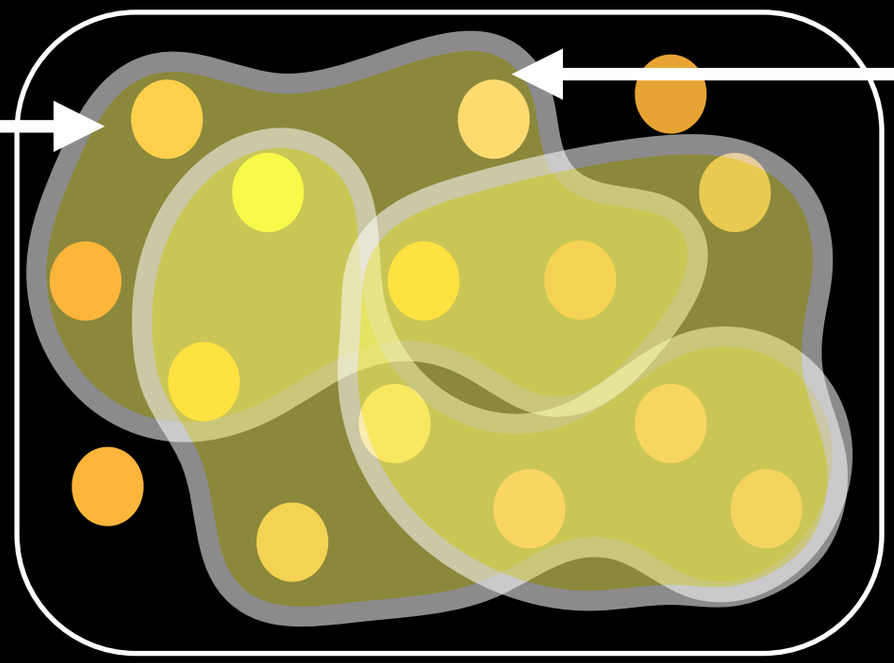
...

...

vote
or averaging

prediction or
classification

a sample:
instance+class



bootstap 1

classifier 1

classifier 2

classifier 3

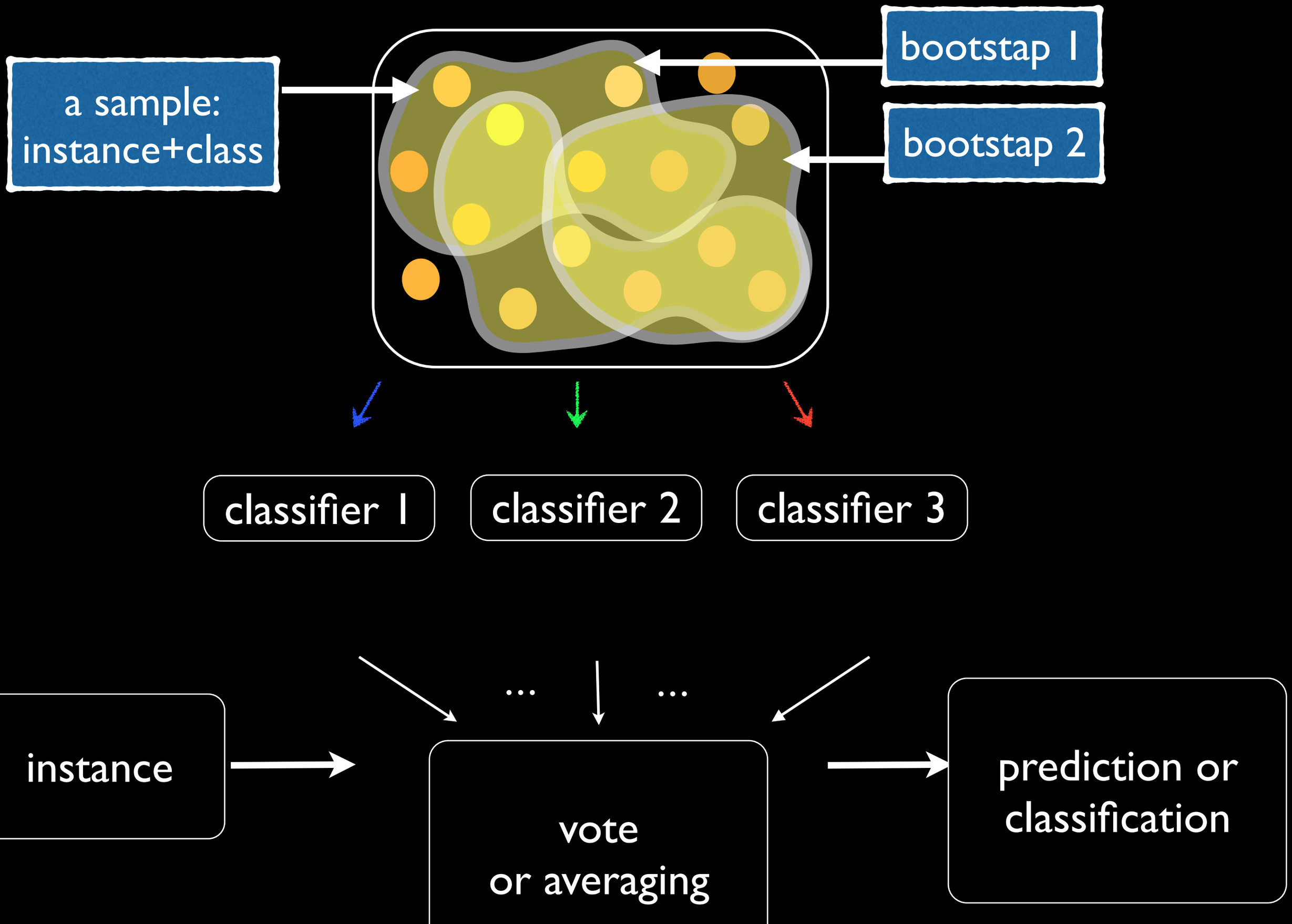
instance

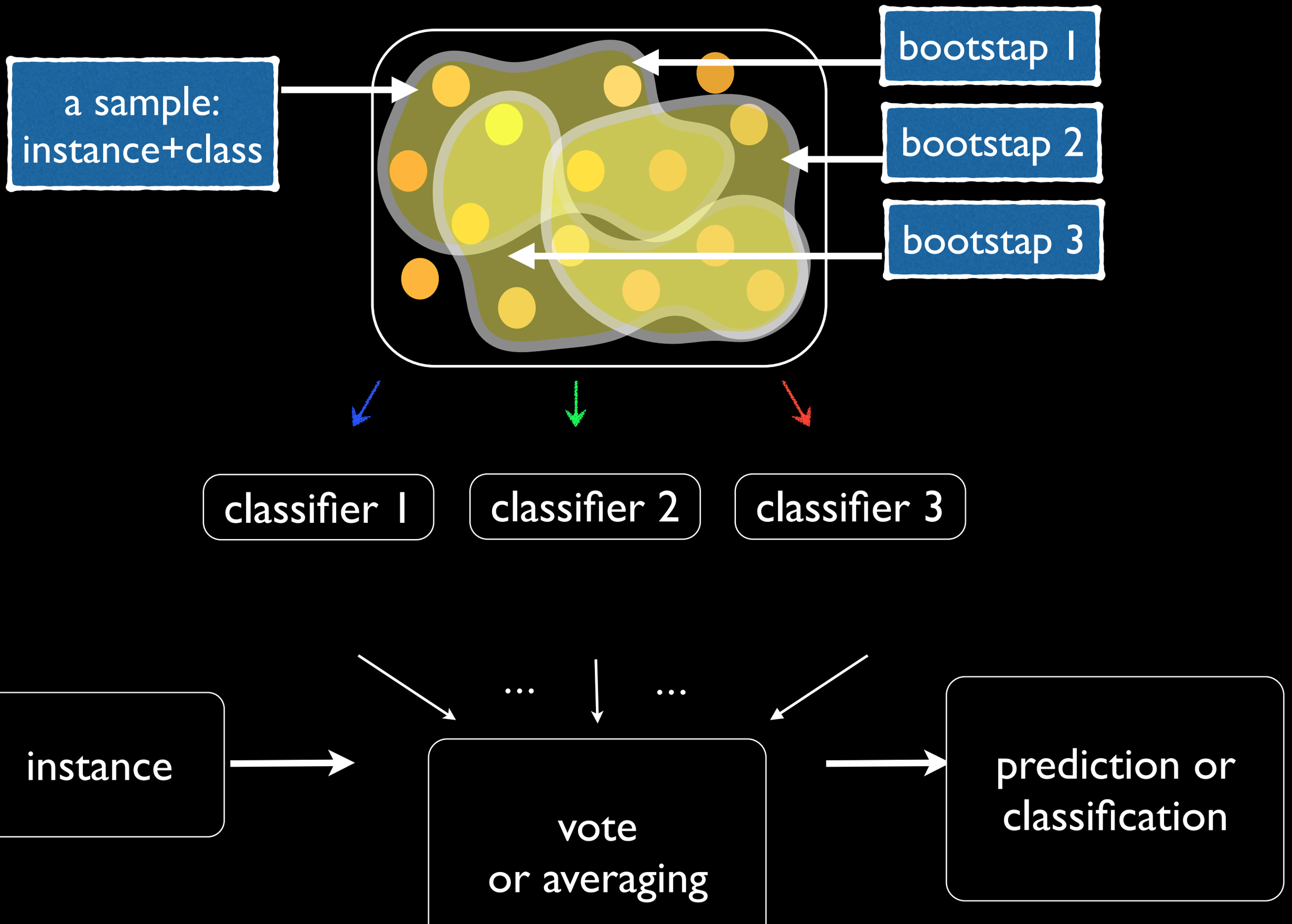
...

...

vote
or averaging

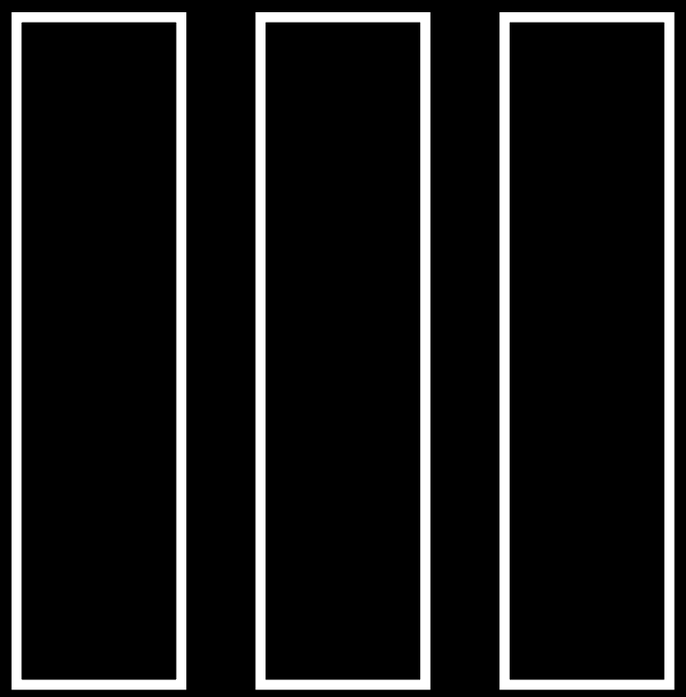
prediction or
classification


a sample:
instance+class
bootstap 1
bootstap 2
classifier 1
classifier 2
classifier 3
instance
...
...
vote
or averaging
prediction or
classification


a sample:
instance+class
bootstap 1
bootstap 2
bootstap 3
classifier 1
classifier 2
classifier 3
instance
...
...
vote
or averaging
prediction or
classification

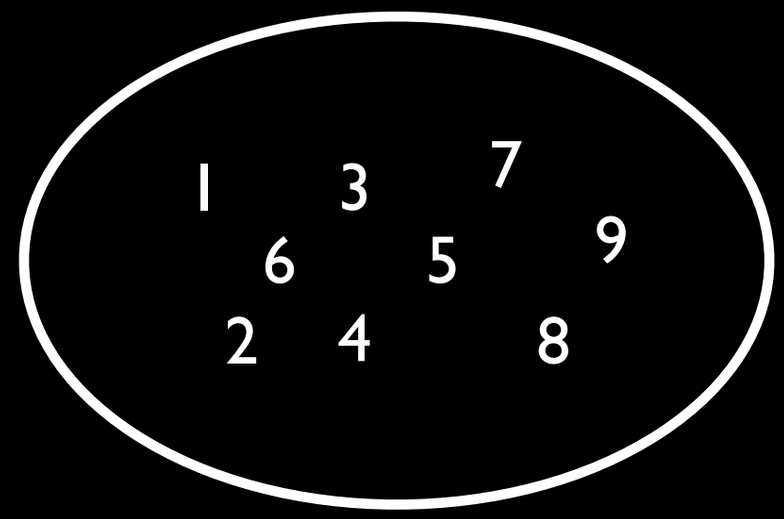
bootstrapping with combinatorial design

N=9, m=3, b=6



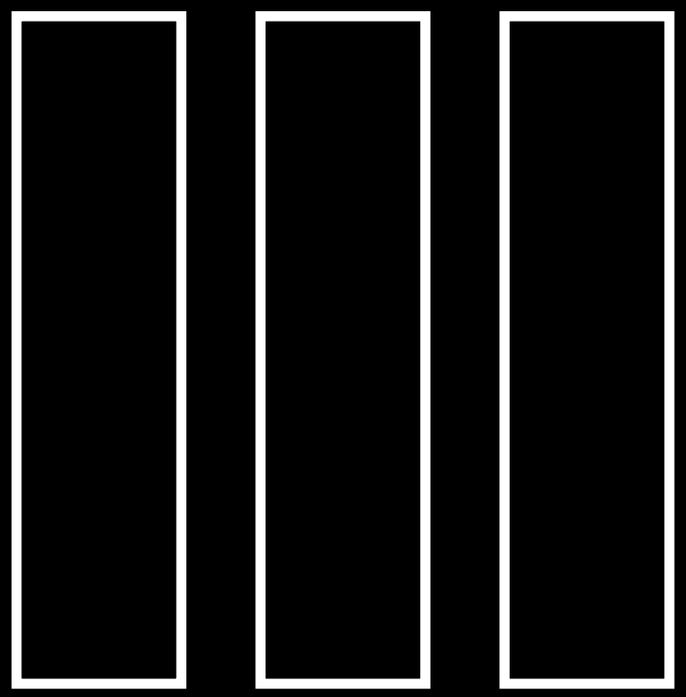
bootstrap: 1 2 3

universe

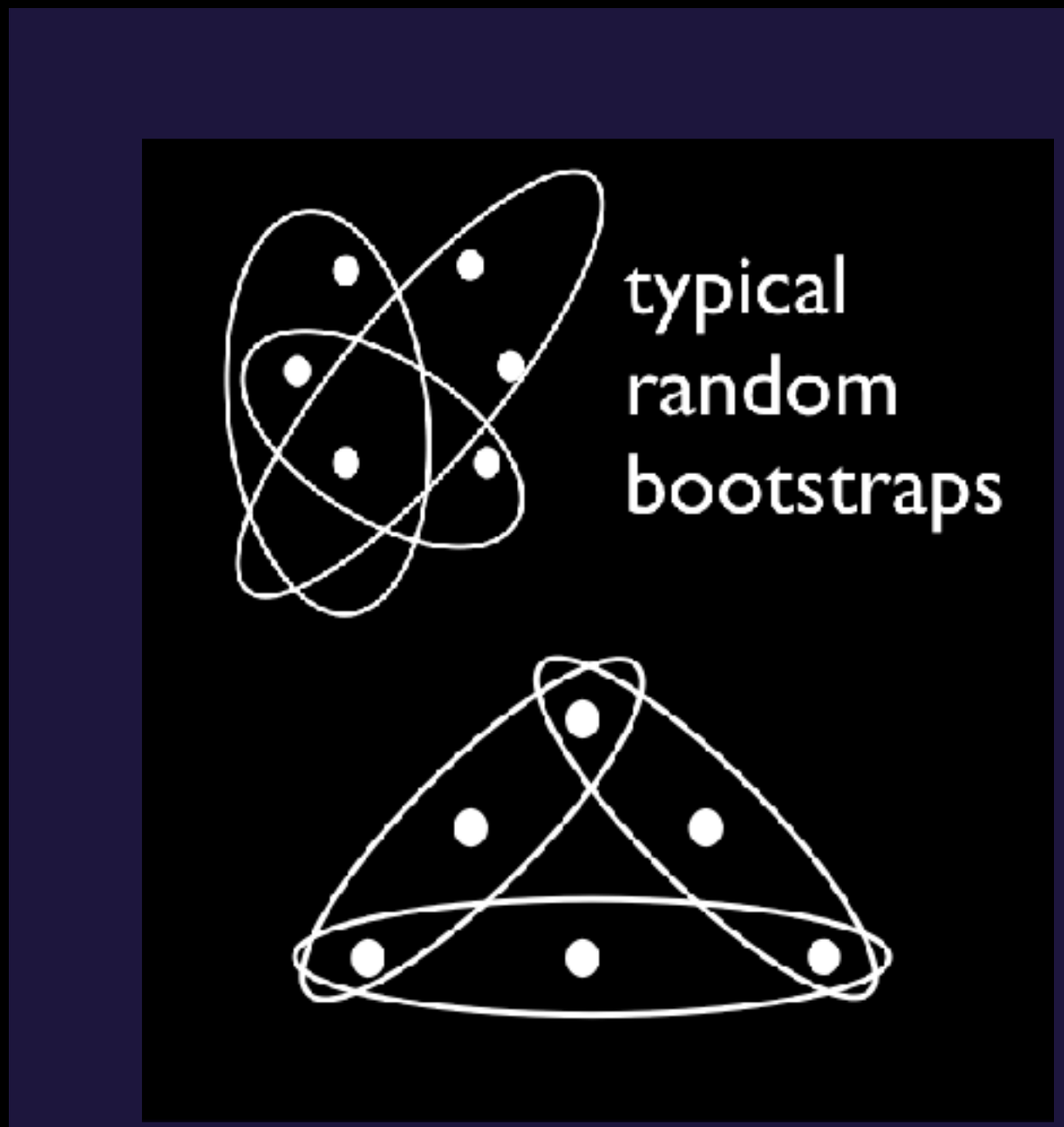


bootstrapping with combinatorial design

N=9, m=3, b=6

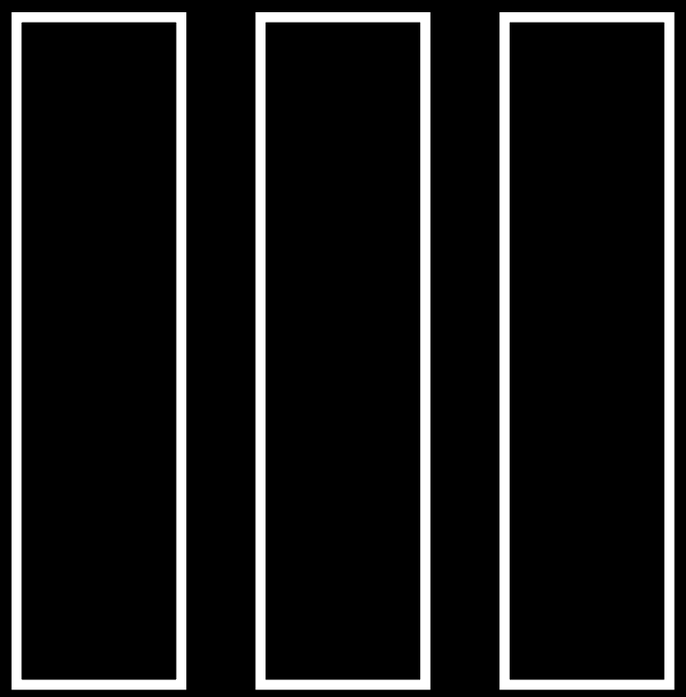


bootstrap: 1 2 3



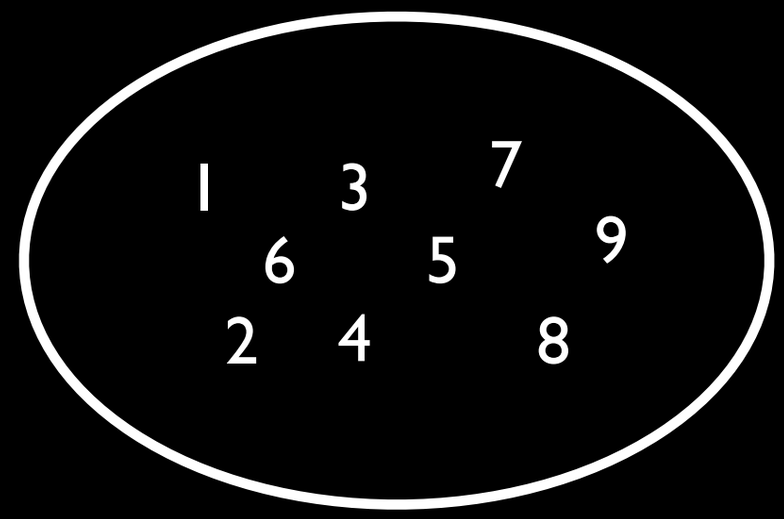
bootstrapping with combinatorial design

N=9, m=3, b=6



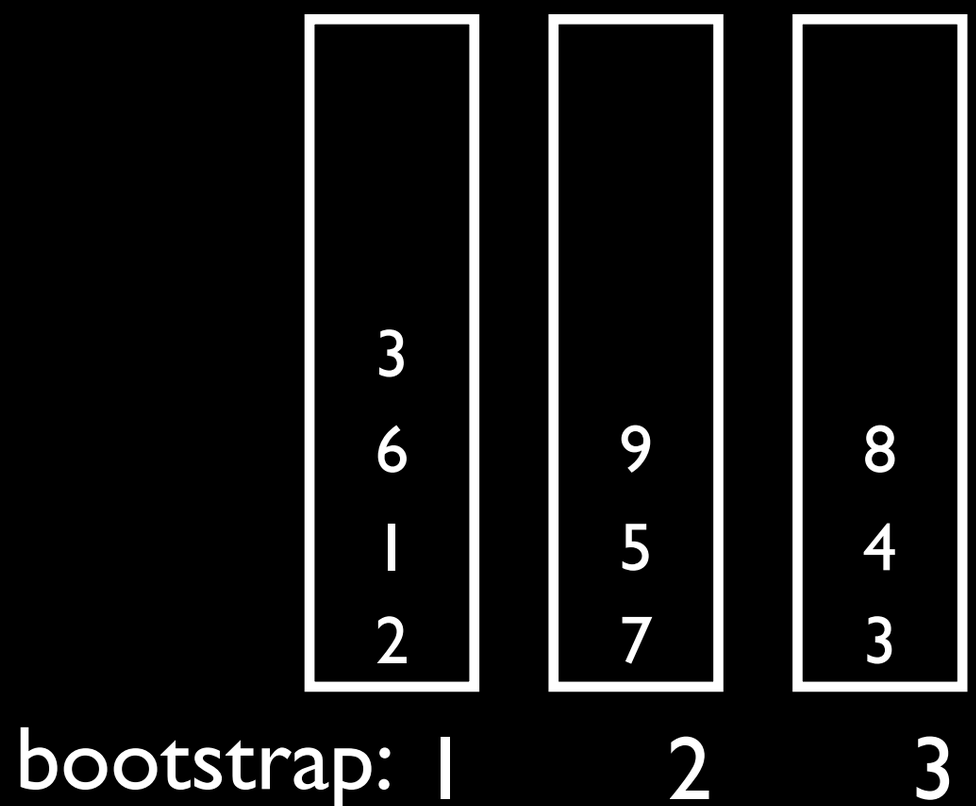
bootstrap: 1 2 3

universe



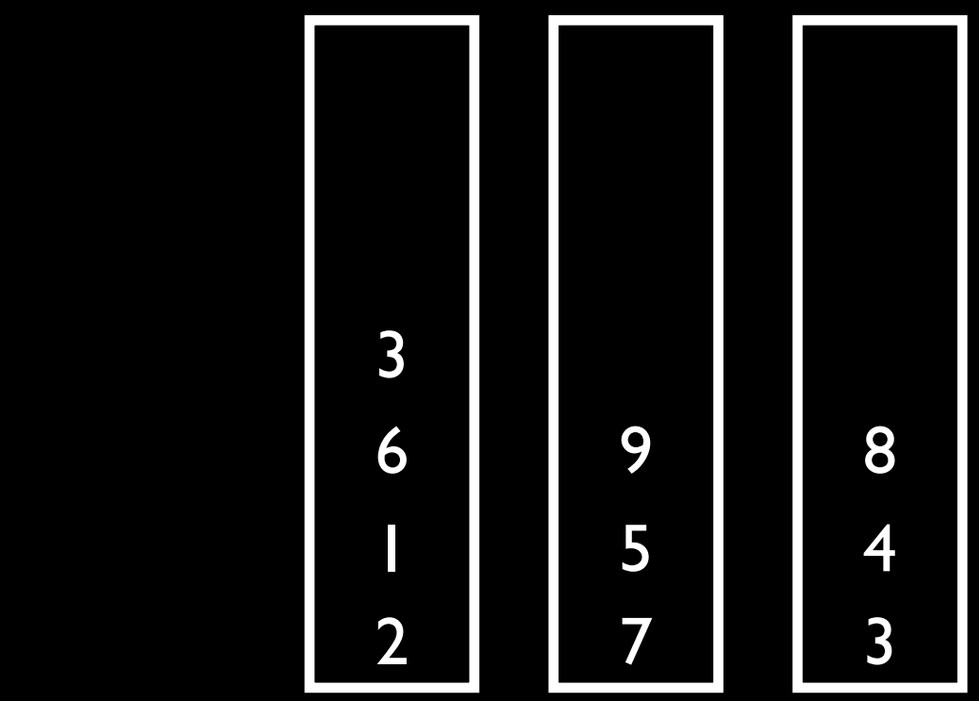
bootstrapping with combinatorial design

N=9, m=3, b=6

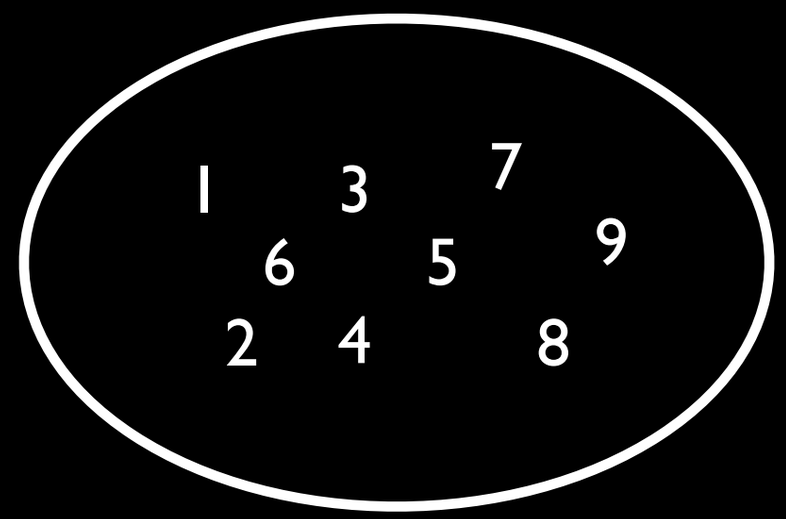


bootstrapping with combinatorial design

N=9, m=3, b=6

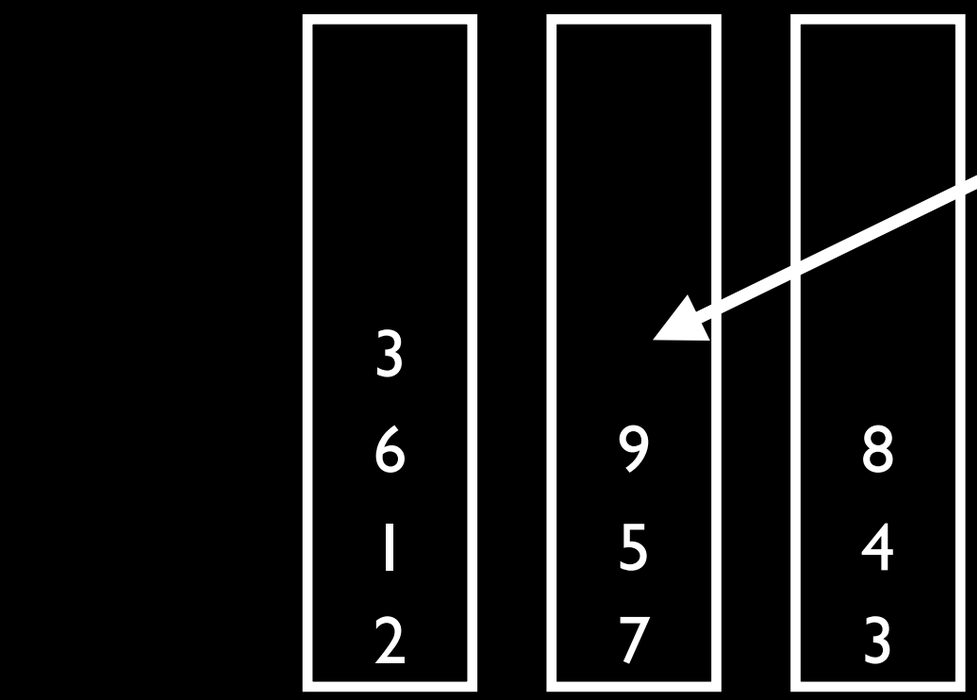


now we are at this point of the execution of Design-Bagging



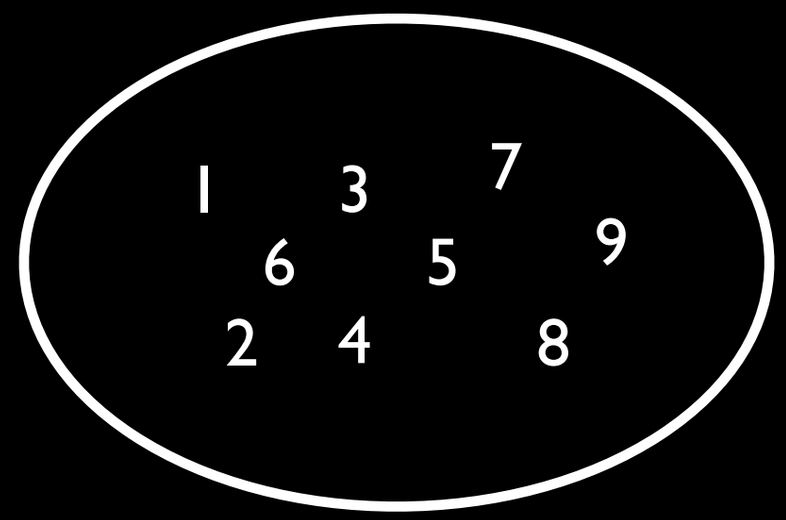
bootstrapping with combinatorial design

N=9, m=3, b=6



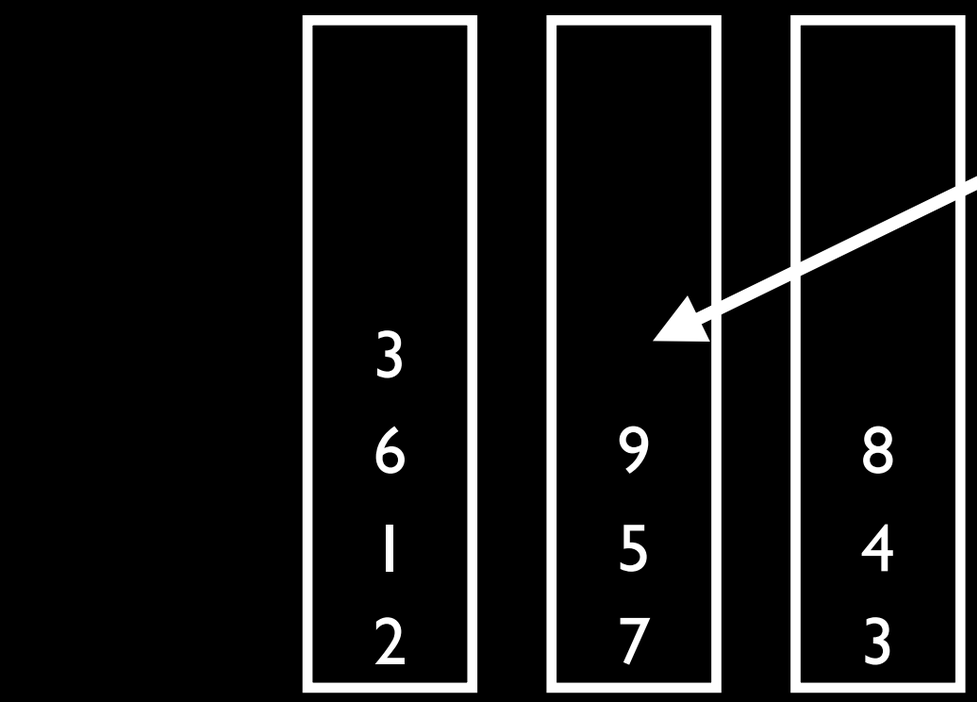
choose at random block #2 to add new element

now we are at this point of the execution of Design-Bagging



bootstrapping with combinatorial design

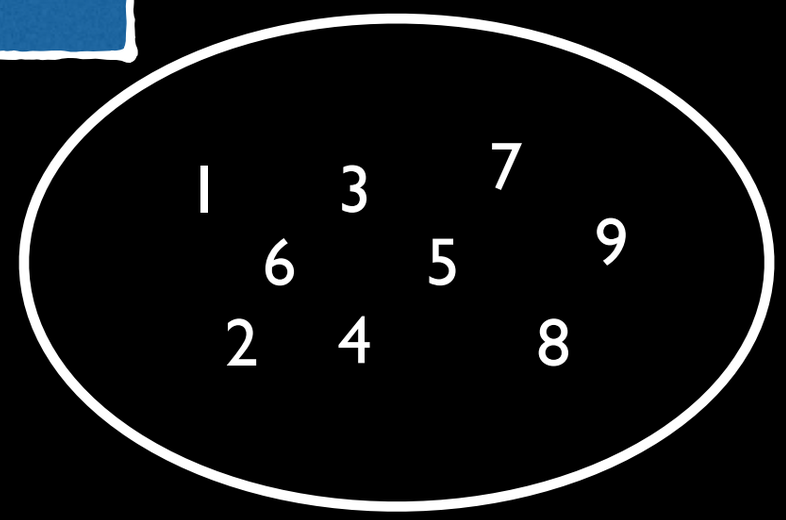
N=9, m=3, b=6



choose at random block #2 to add new element

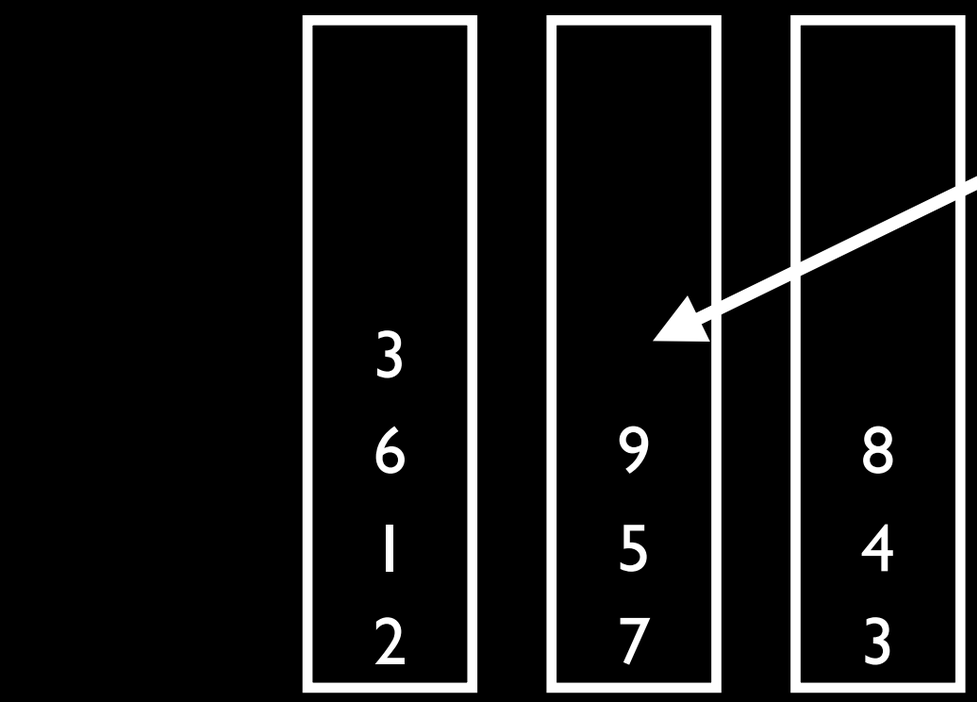
now we are at this point of the execution of Design-Bagging

the elements in Universe - Block#2 are {1,2,3,4,6,8}



bootstrapping with combinatorial design

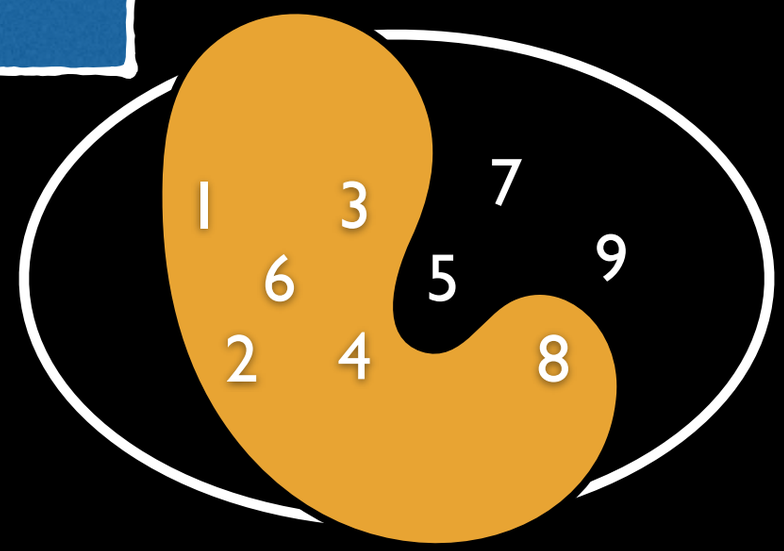
N=9, m=3, b=6



choose at random block #2 to add new element

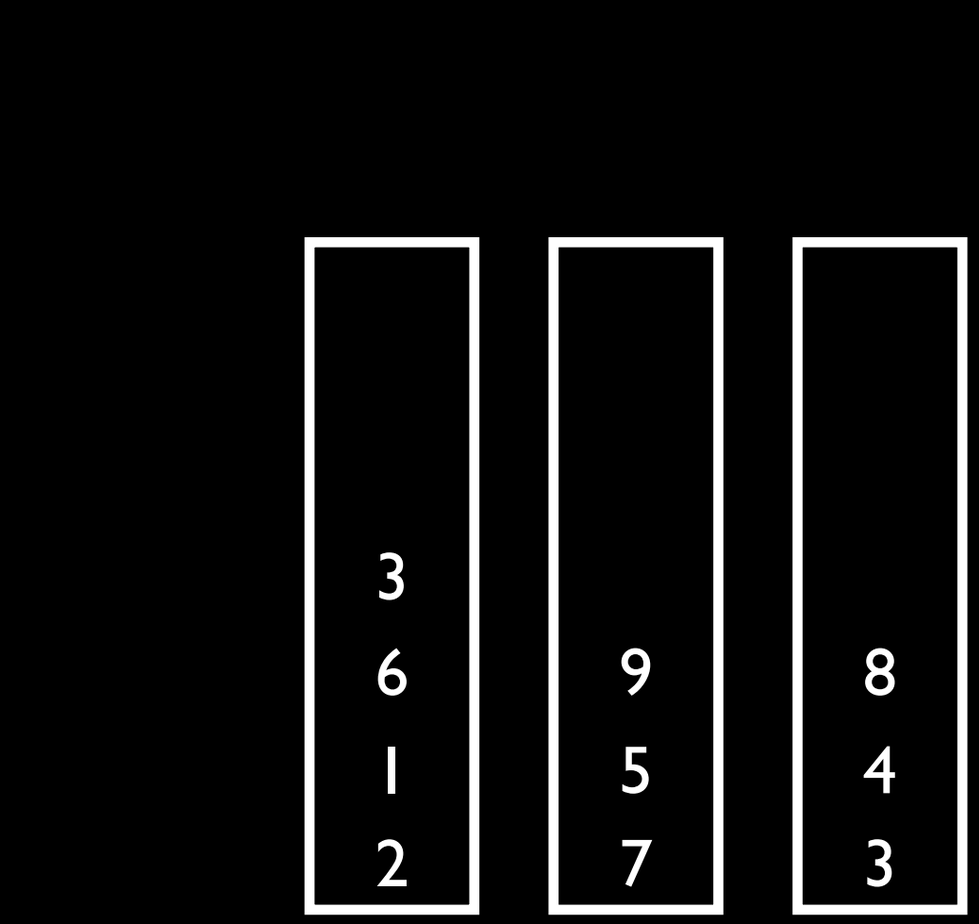
now we are at this point of the execution of Design-Bagging

the elements in Universe - Block#2 are {1,2,3,4,6,8}



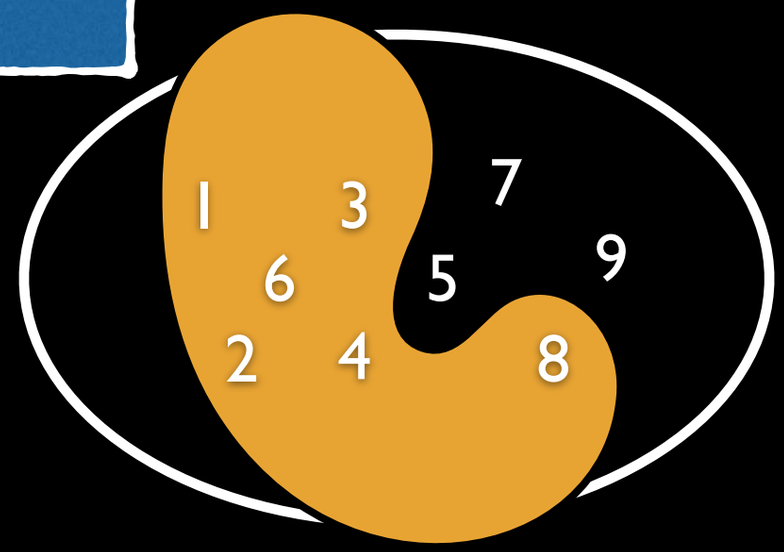
bootstrapping with combinatorial design

N=9, m=3, b=6



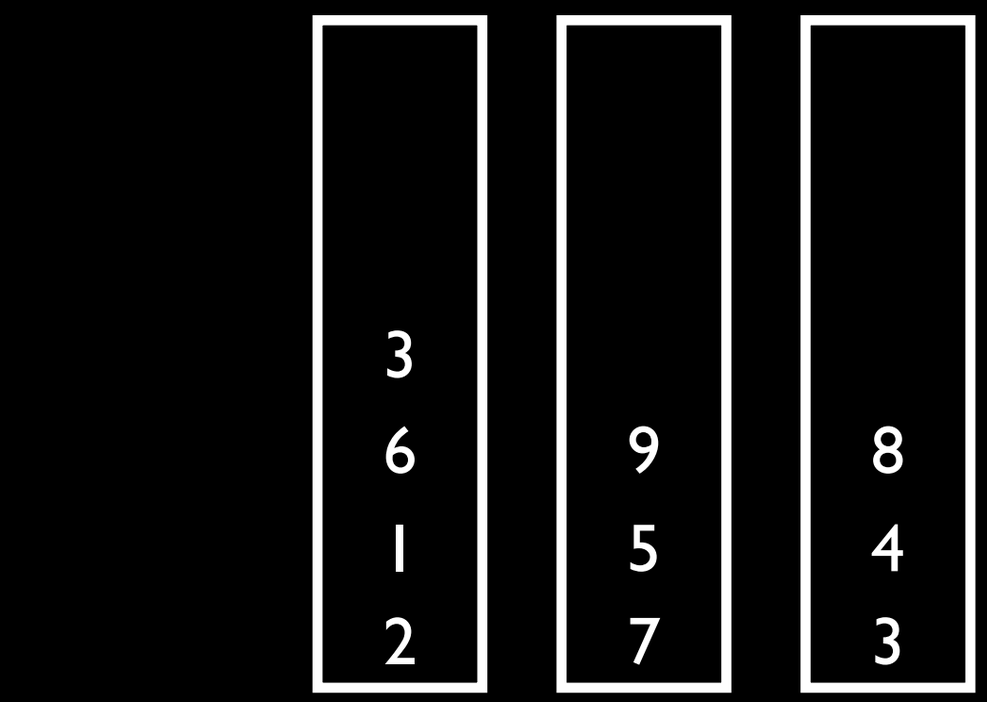
among {1,2,3,4,6,8} those that appear the least are {1,2,4,6,8}

the elements in Universe - Block#2 are {1,2,3,4,6,8}



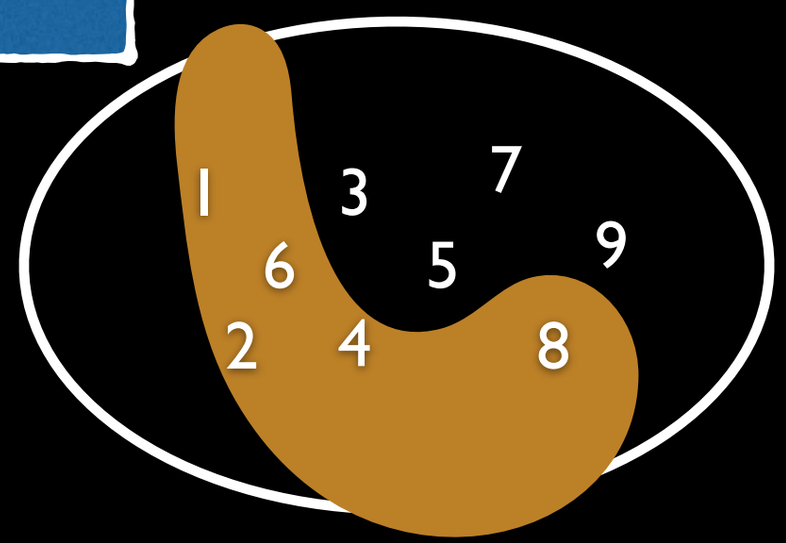
bootstrapping with combinatorial design

N=9, m=3, b=6



among {1,2,3,4,6,8} those that appear the least are {1,2,4,6,8}

the elements in Universe - Block#2 are {1,2,3,4,6,8}



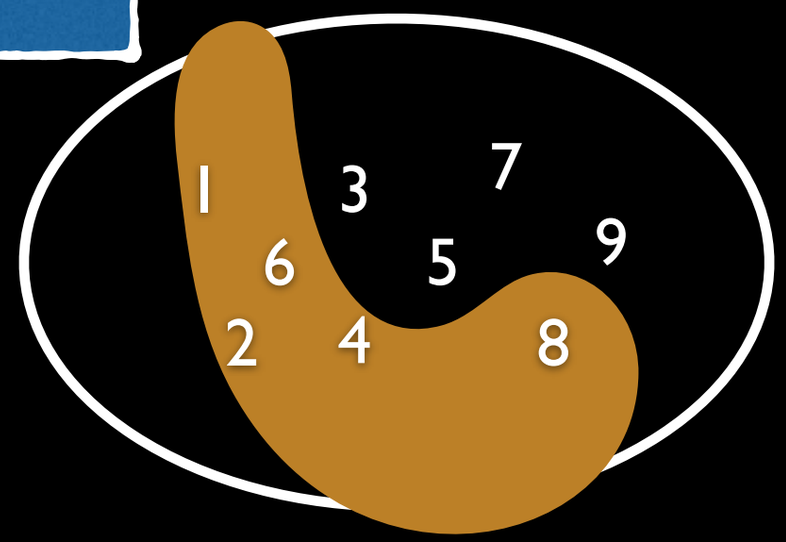
bootstrapping with combinatorial design

N=9, m=3, b=6

among {1,2,3,4,6,8} those that appear the least are {1,2,4,6,8}

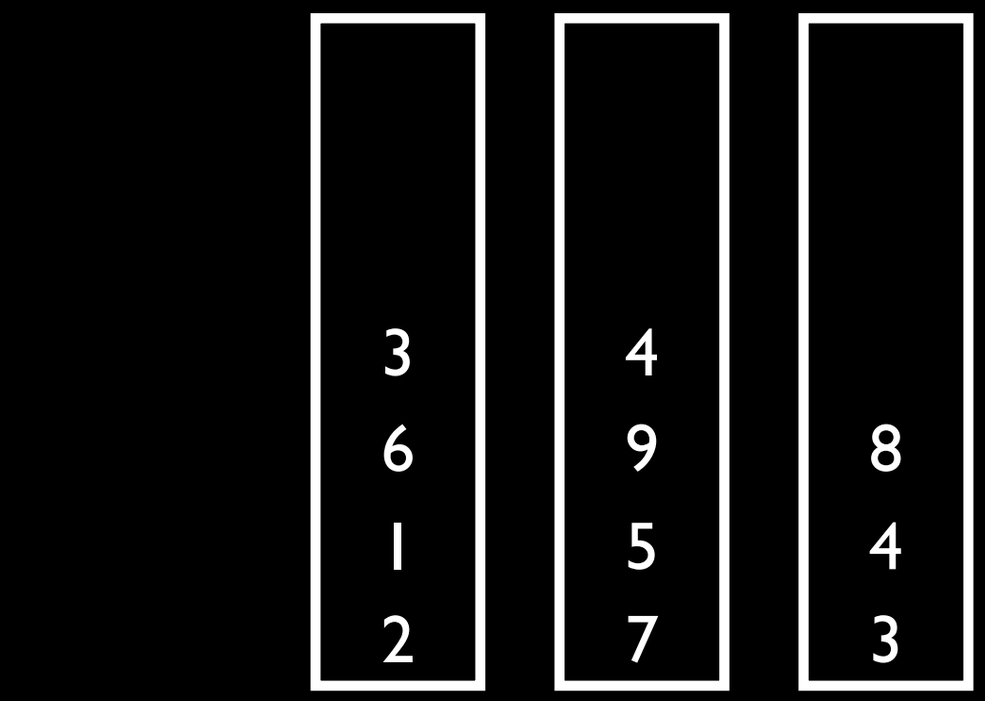
sampled: 4
— chosen uniformly at random from {1,2,4,6,8} —

bootstrap: 1	2	3
3	4	
6	9	8
1	5	4
2	7	3



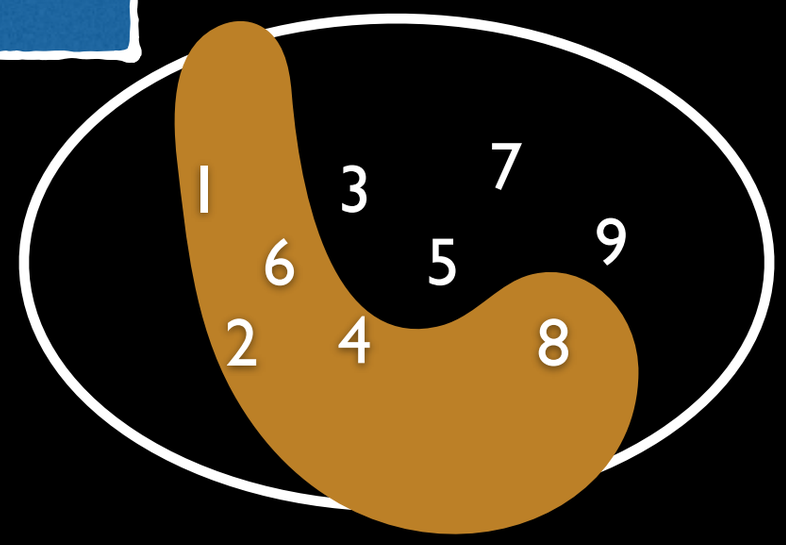
bootstrapping with combinatorial design

N=9, m=3, b=6



among {1,2,3,4,6,8} those that appear the least are {1,2,4,6,8}

sampled: 4
— chosen uniformly at random from {1,2,4,6,8} —



this is a biased sampling process

bootstrapping with combinatorial design

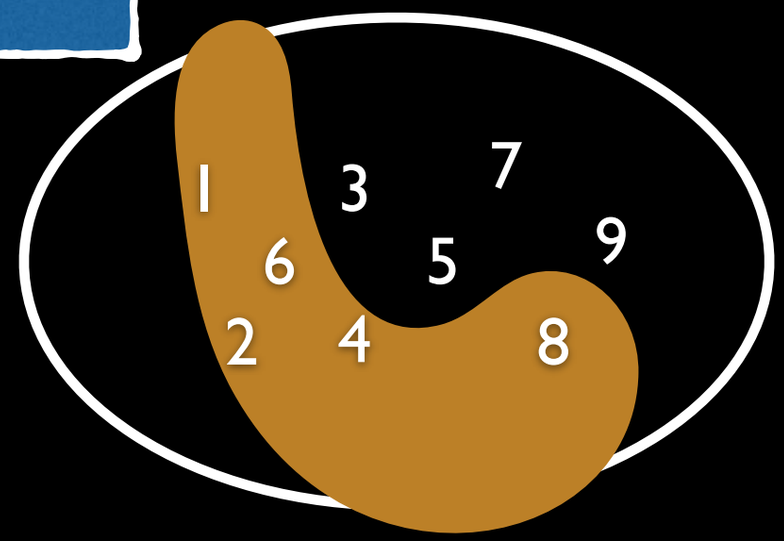
N=9, m=3, b=6

Question #24: enhance bootstrapping

among {1,2,3,4,6,8} those that appear the least are {1,2,4,6,8}

sampled: 4
— chosen uniformly at random from {1,2,4,6,8} —

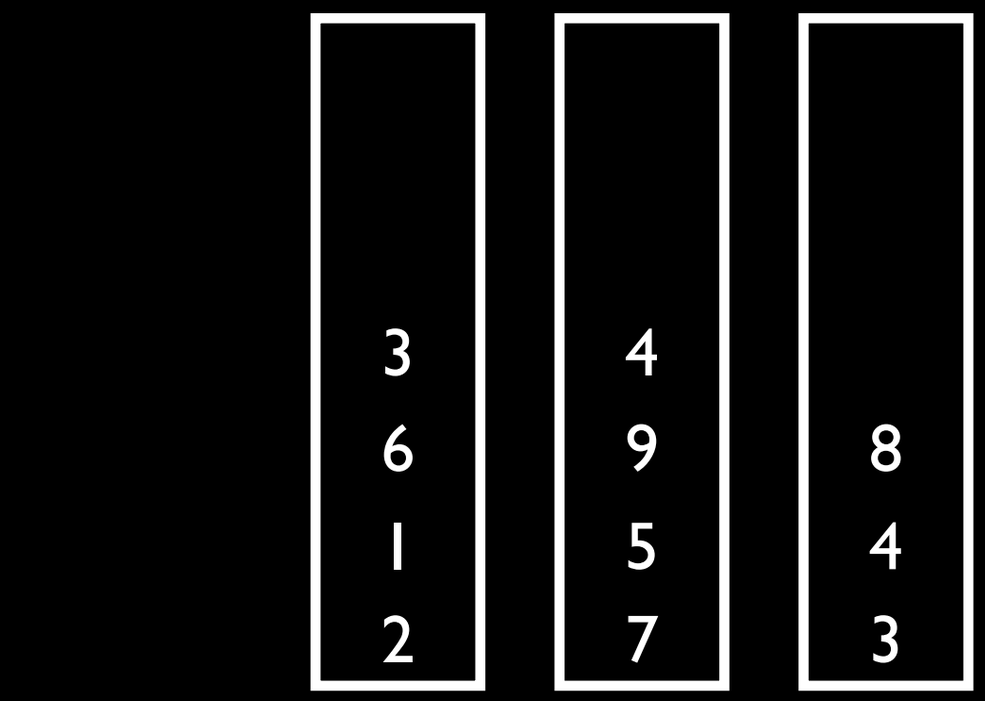
bootstrap: 1	2	3
3	4	
6	9	8
1	5	4
2	7	3



this is a biased sampling process

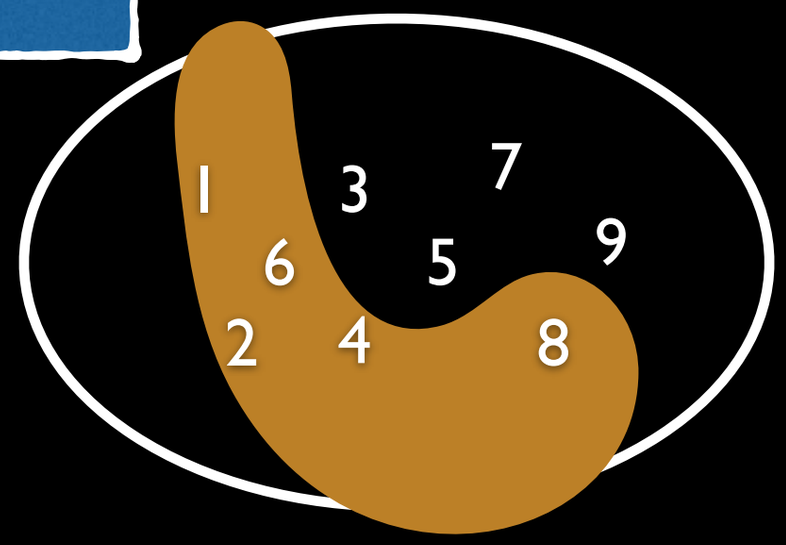
bootstrapping with combinatorial design

N=9, m=3, b=6



among {1,2,3,4,6,8} those that appear the least are {1,2,4,6,8}

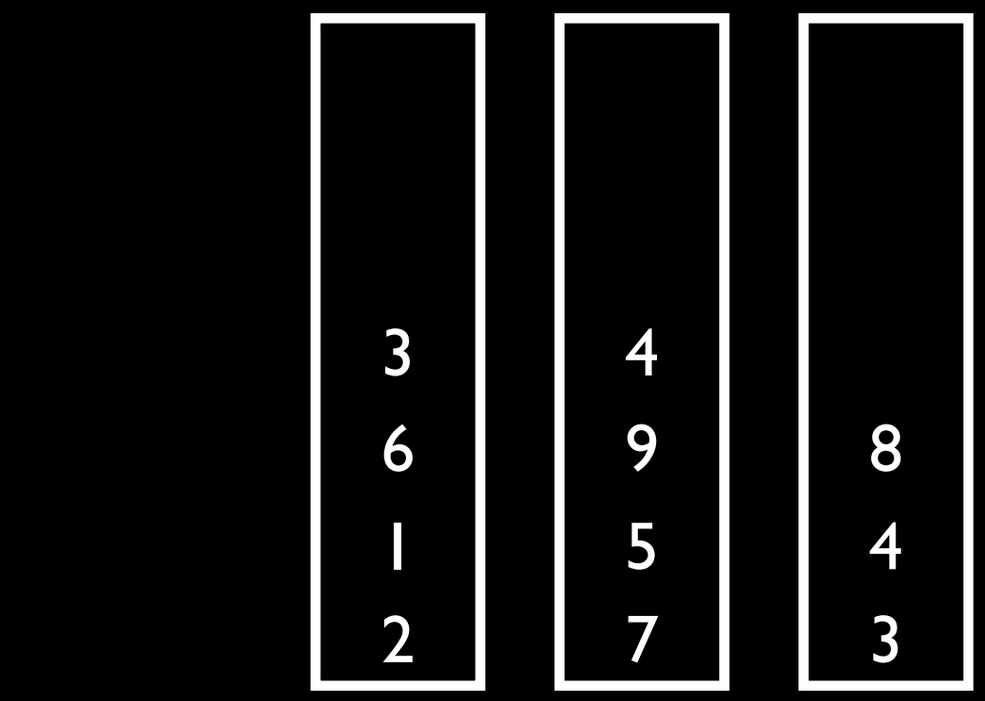
sampled: 4
— chosen uniformly at random from {1,2,4,6,8} —



this is a biased sampling process

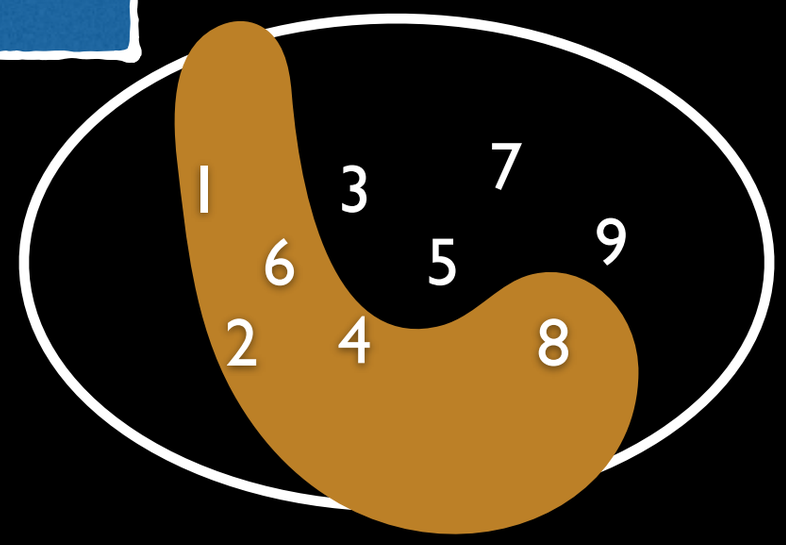
bootstrapping with combinatorial design

N=9, m=3, b=6



among {1,2,3,4,6,8} those that appear the least are {1,2,4,6,8}

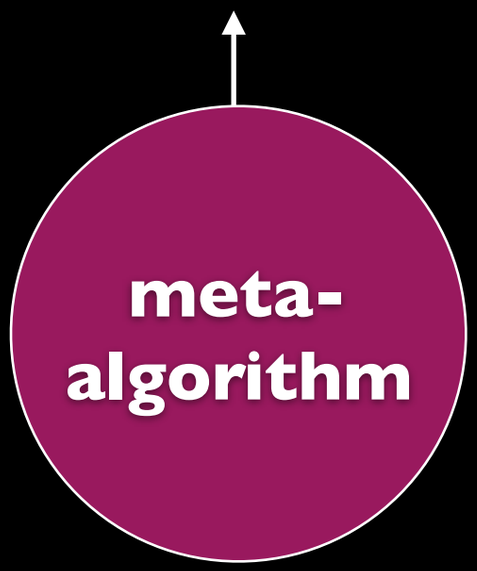
sampled: 4
— chosen uniformly at random from {1,2,4,6,8} —



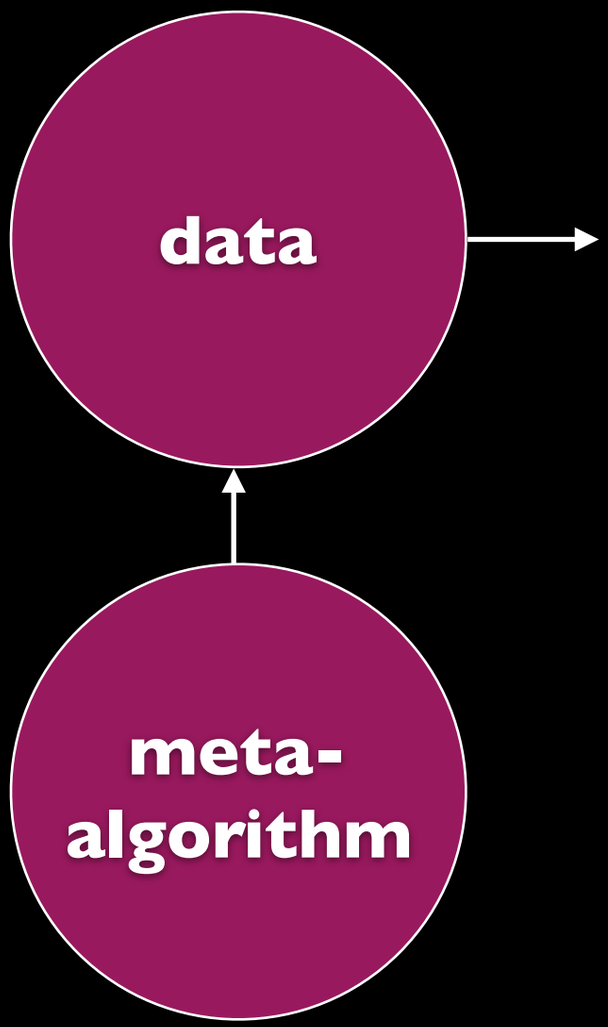
this is a biased sampling process

machine translation components

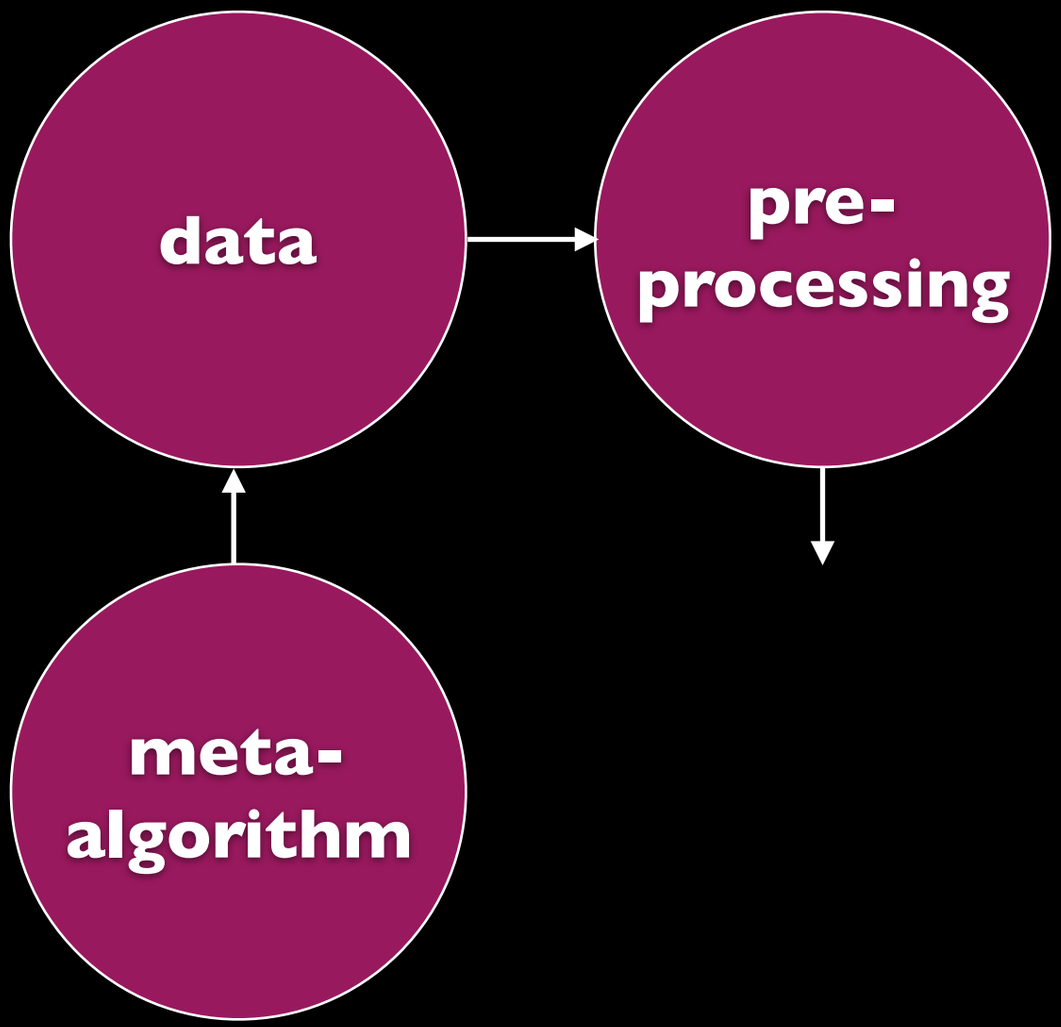
machine translation components



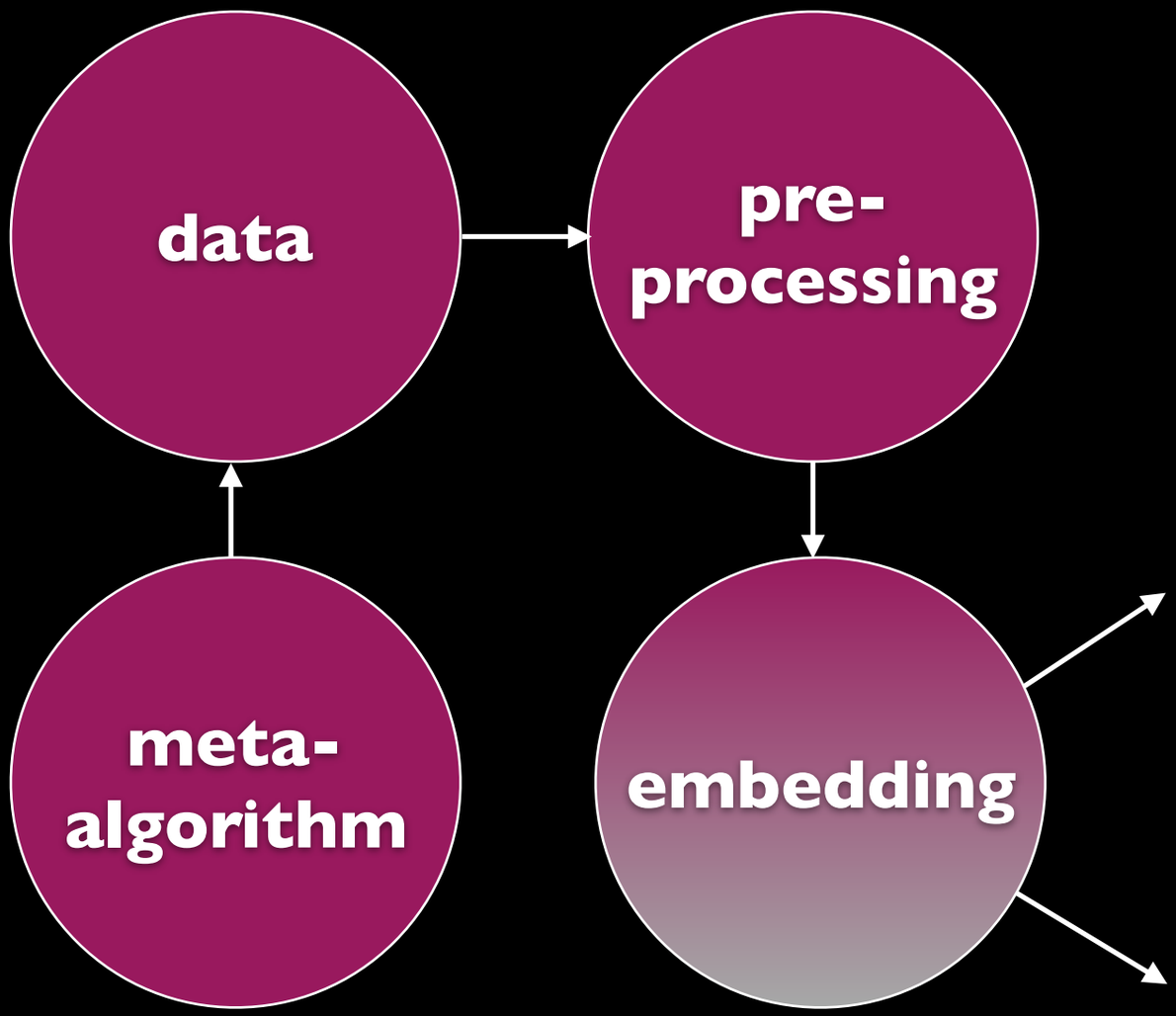
machine translation components



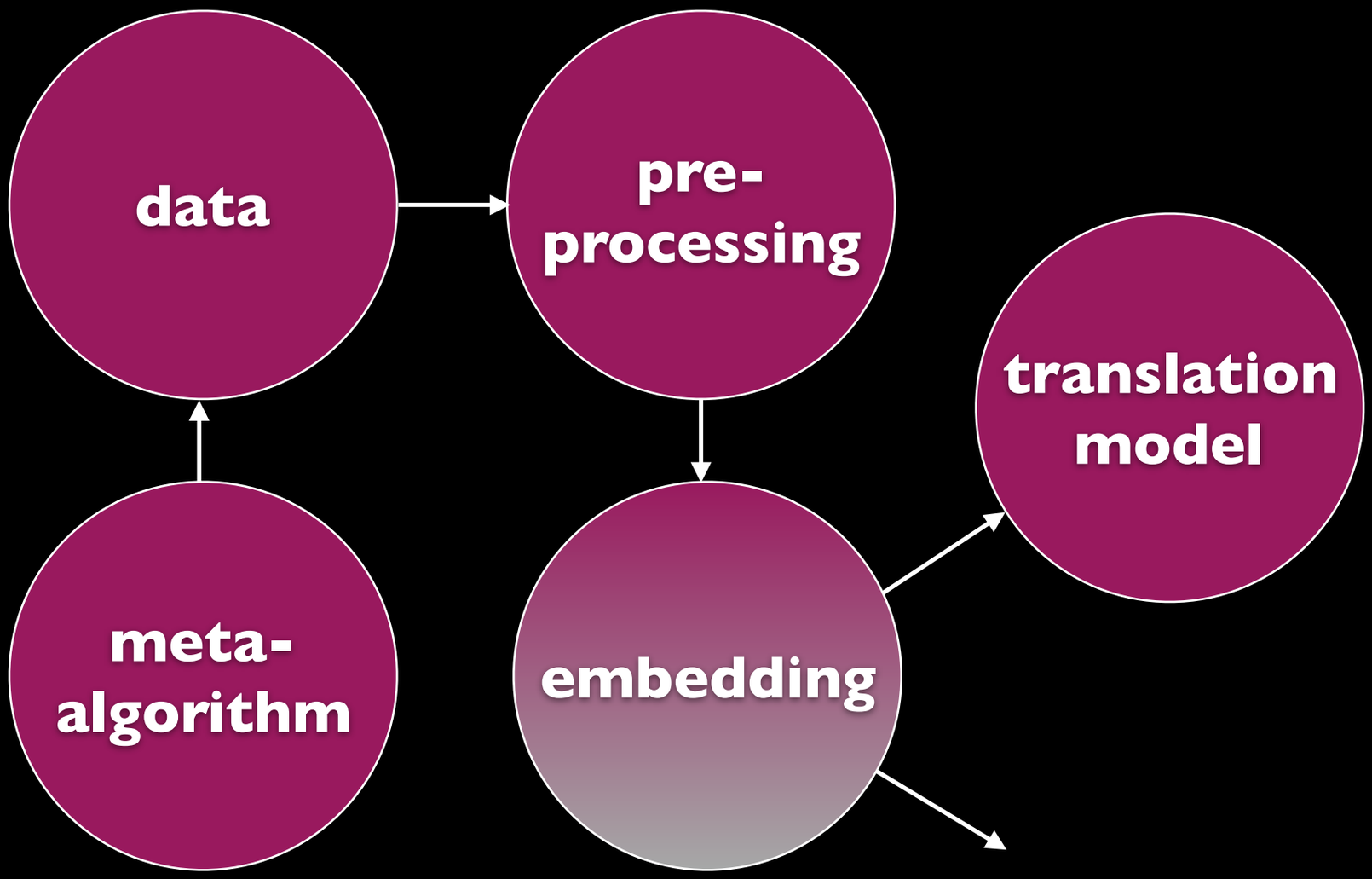
machine translation components



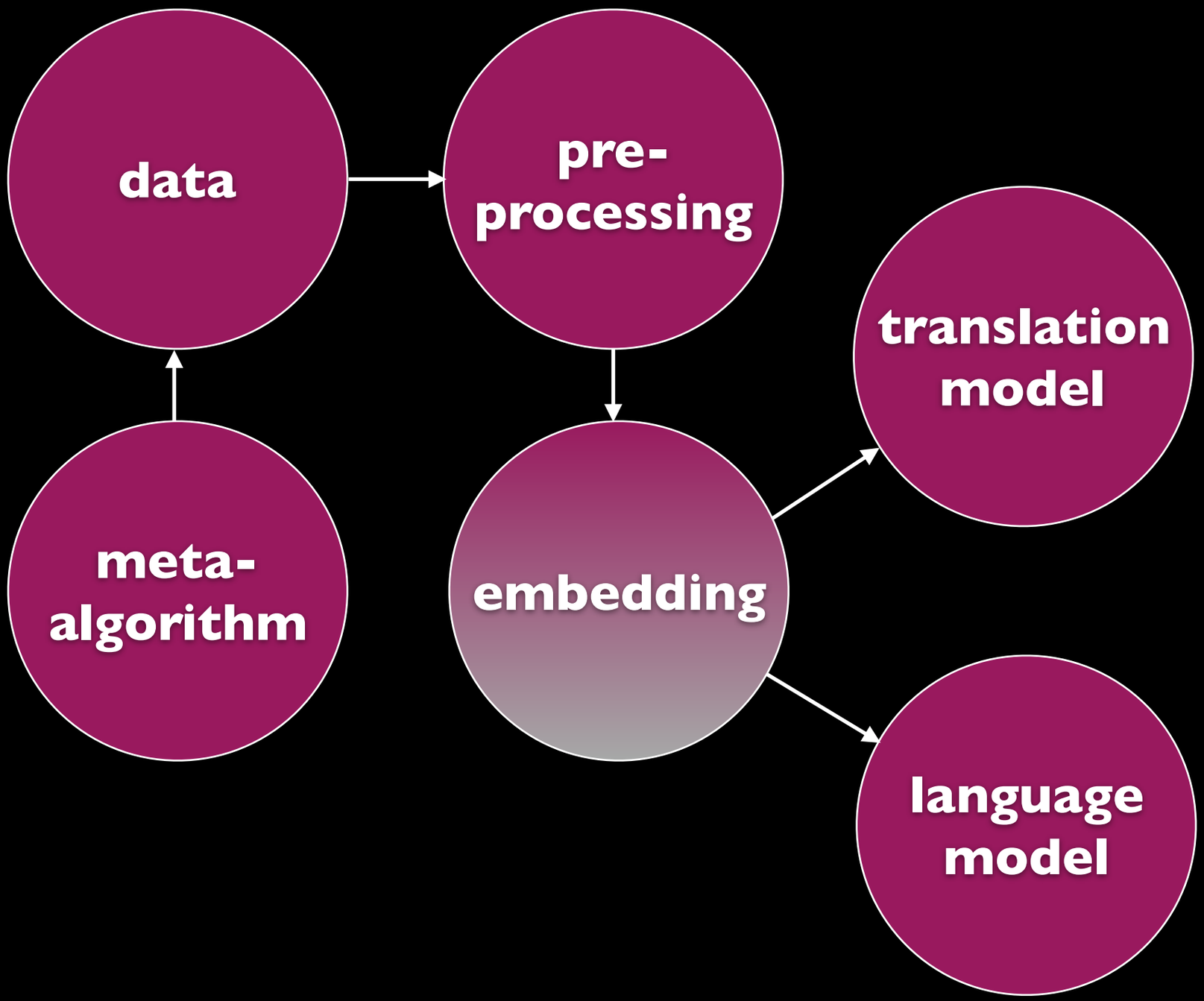
machine translation components



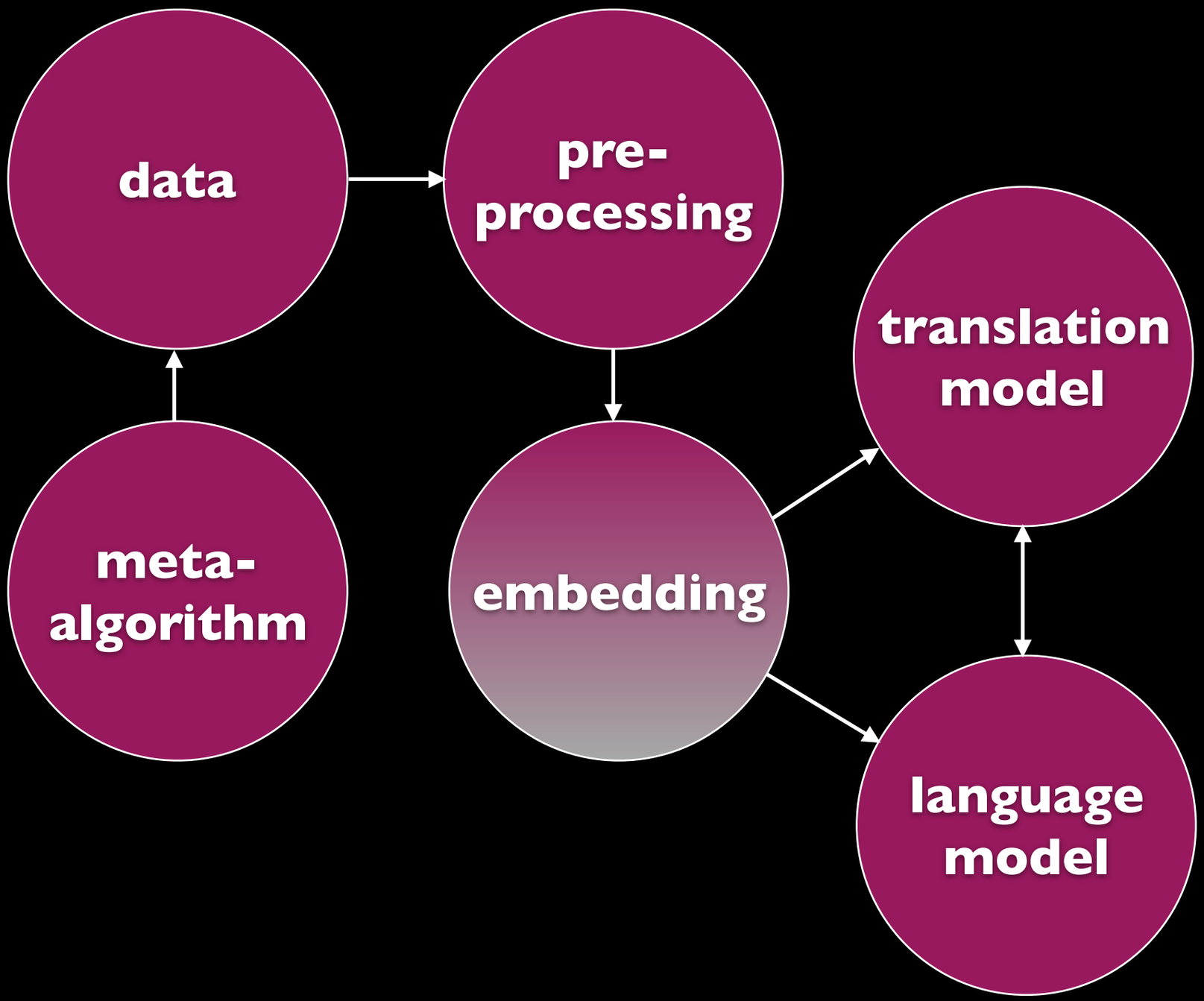
machine translation components



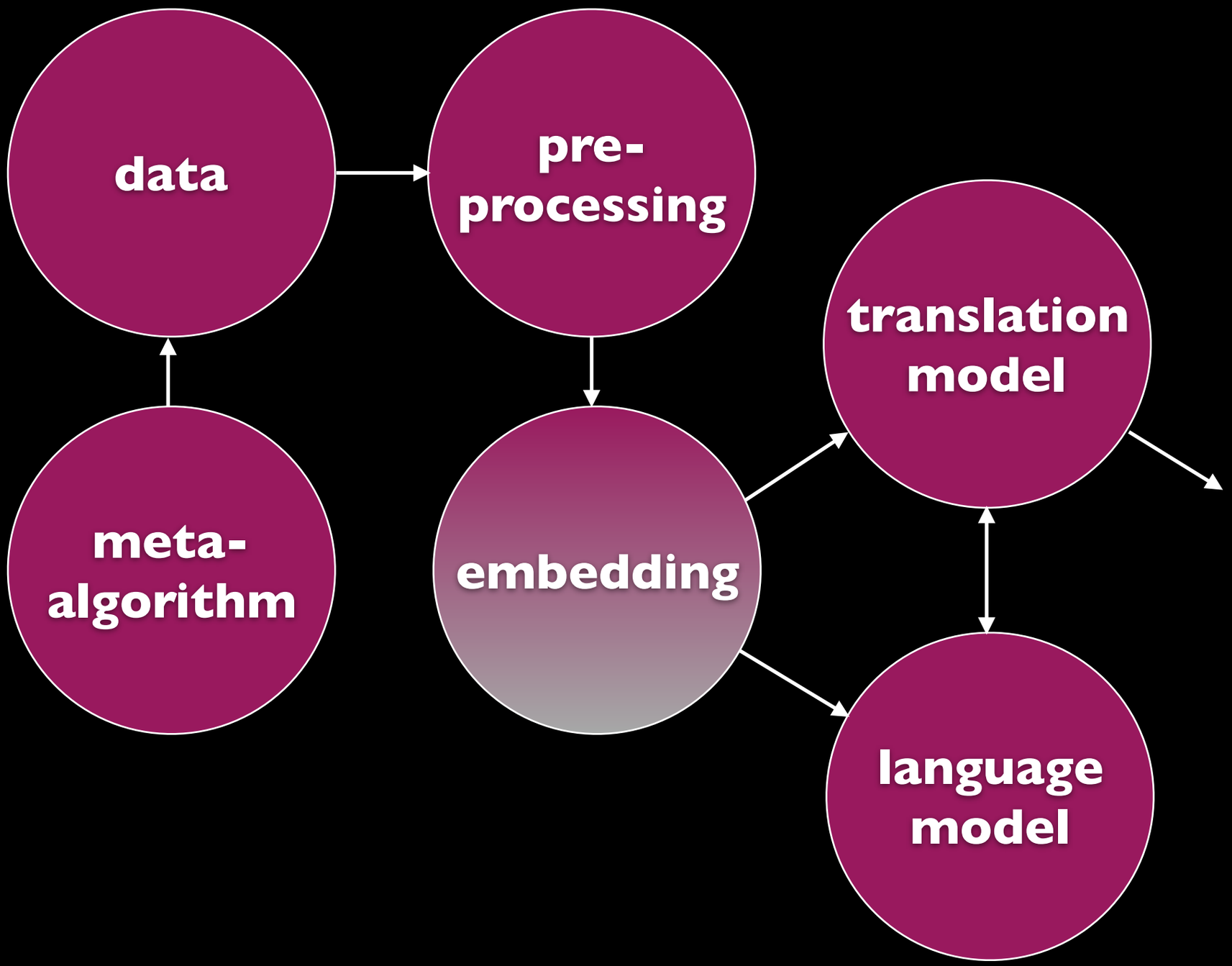
machine translation components



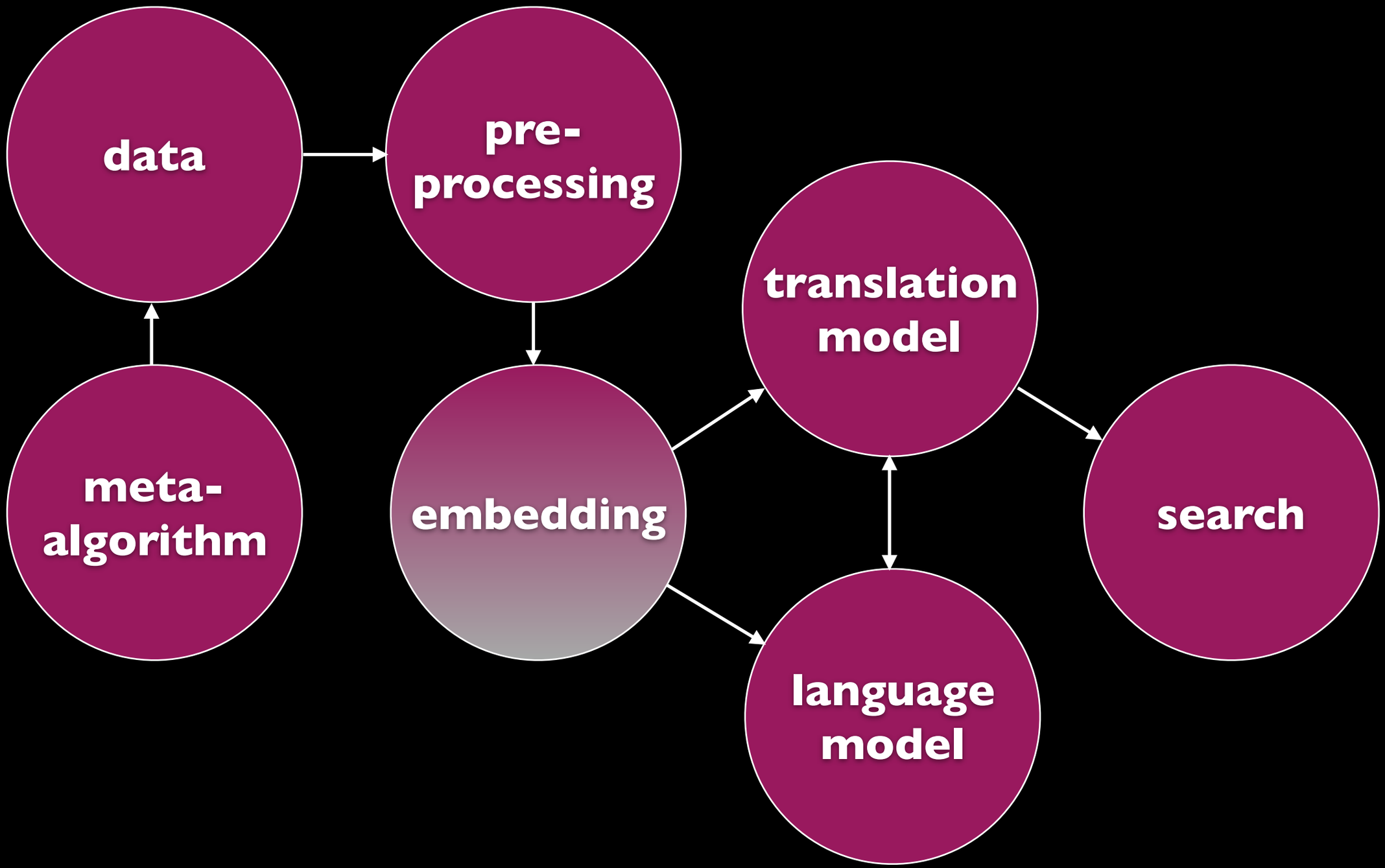
machine translation components



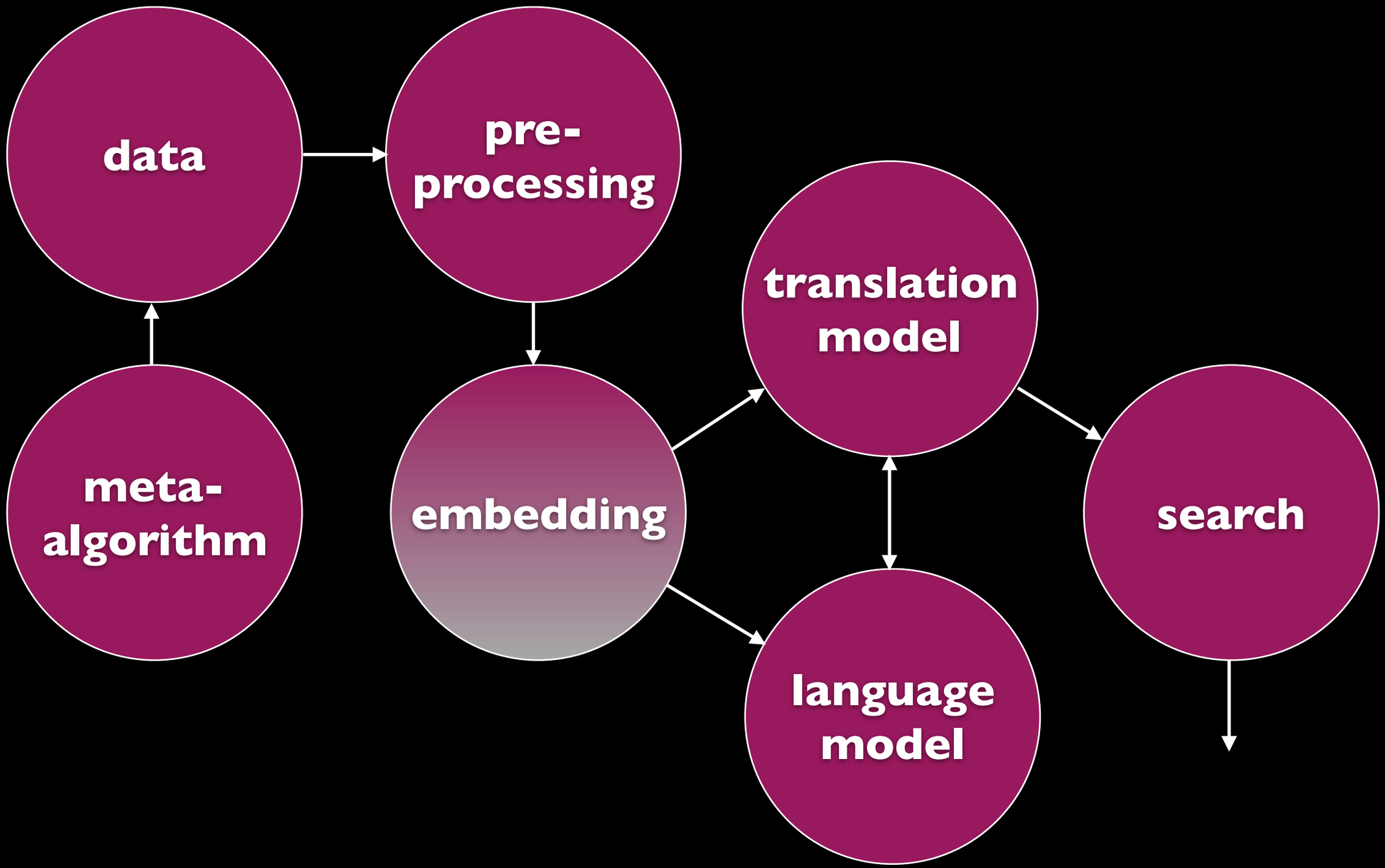
machine translation components



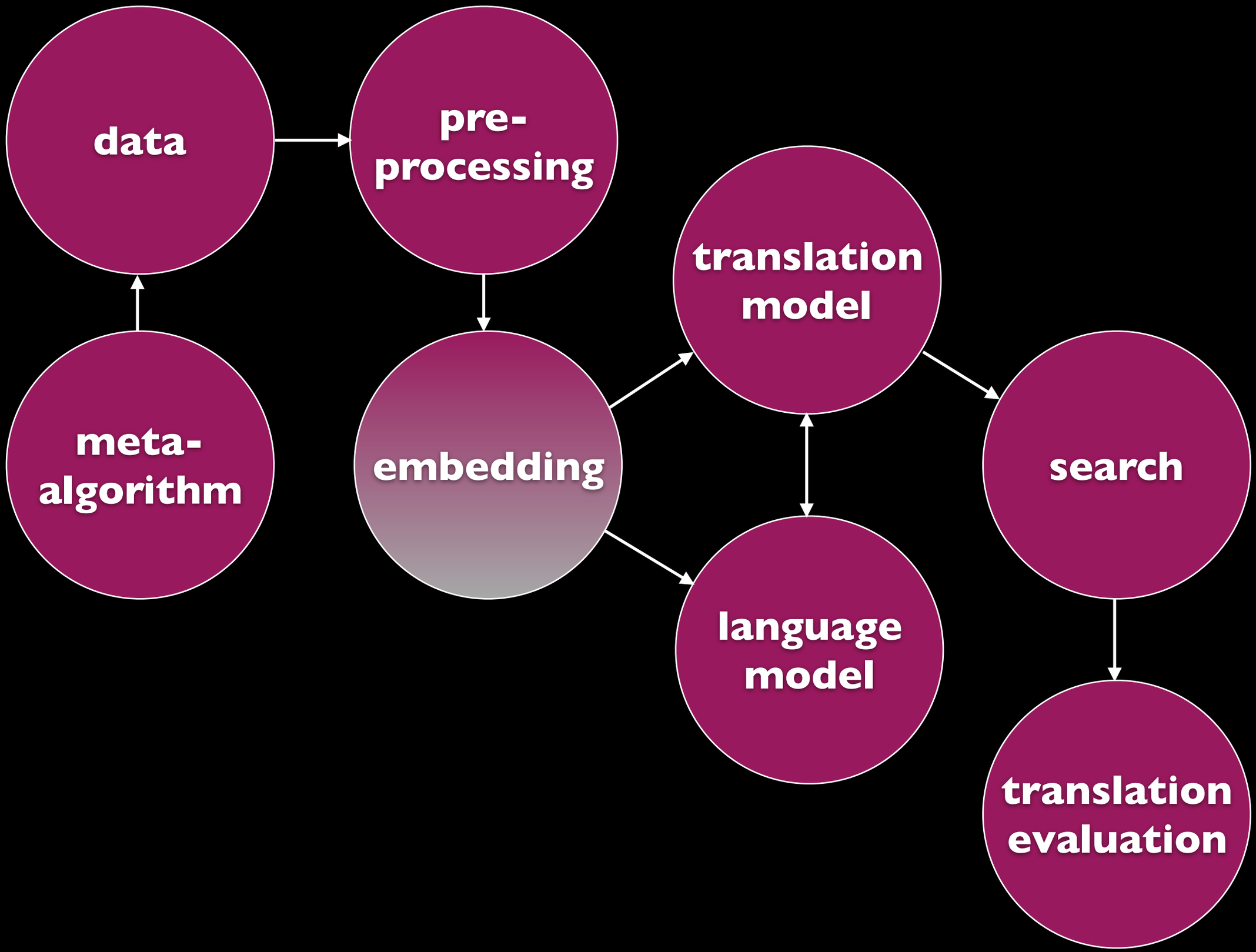
machine translation components



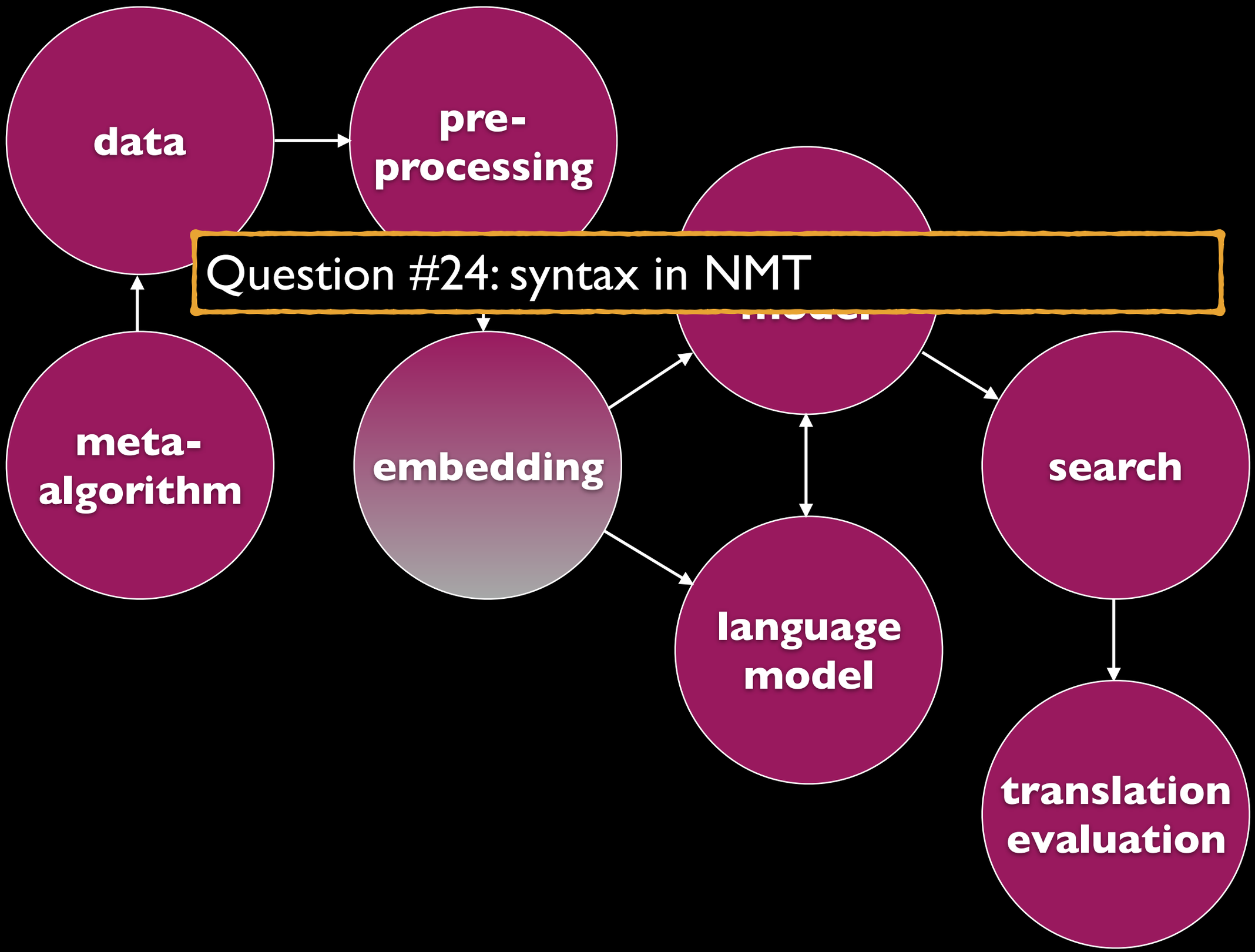
machine translation components



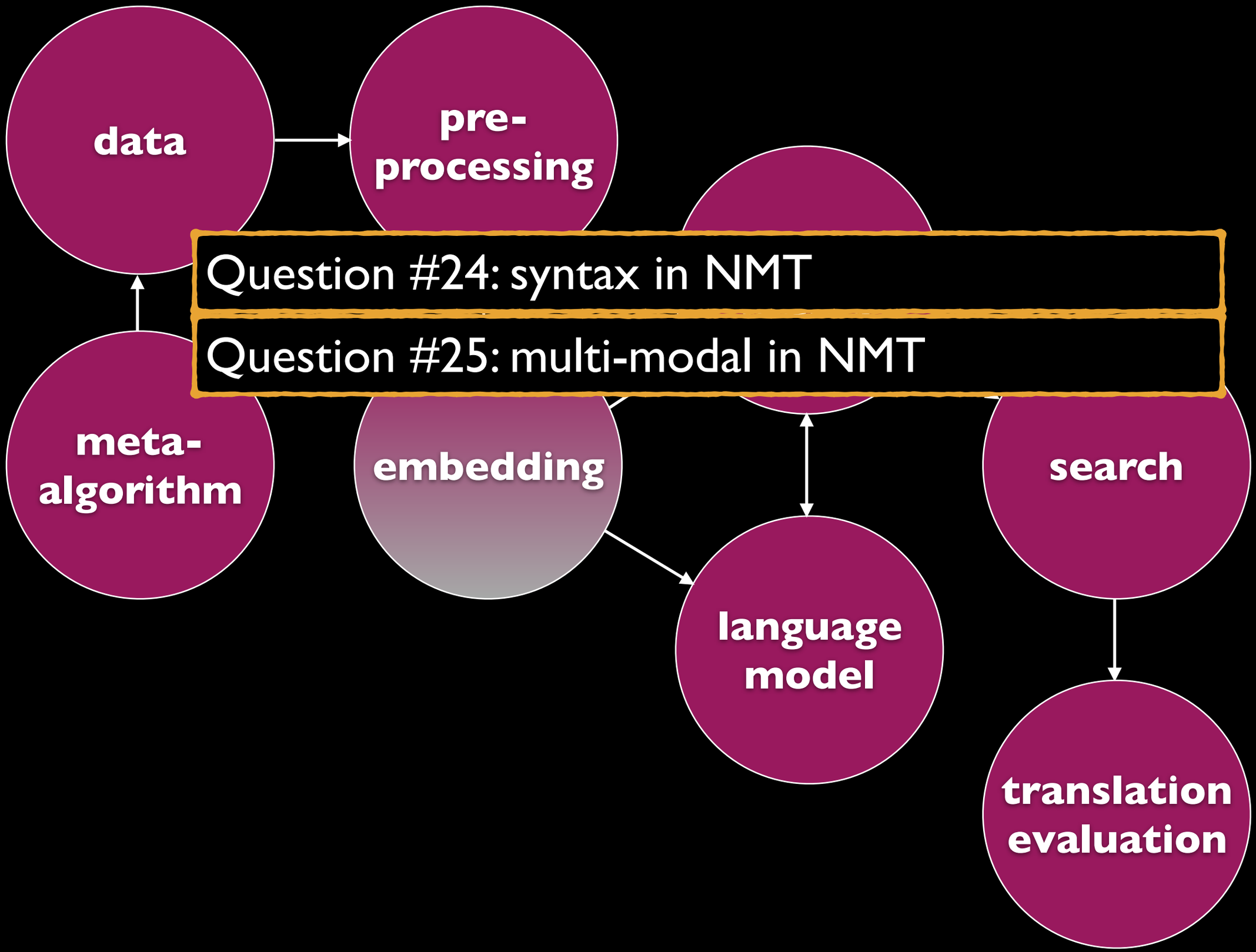
machine translation components



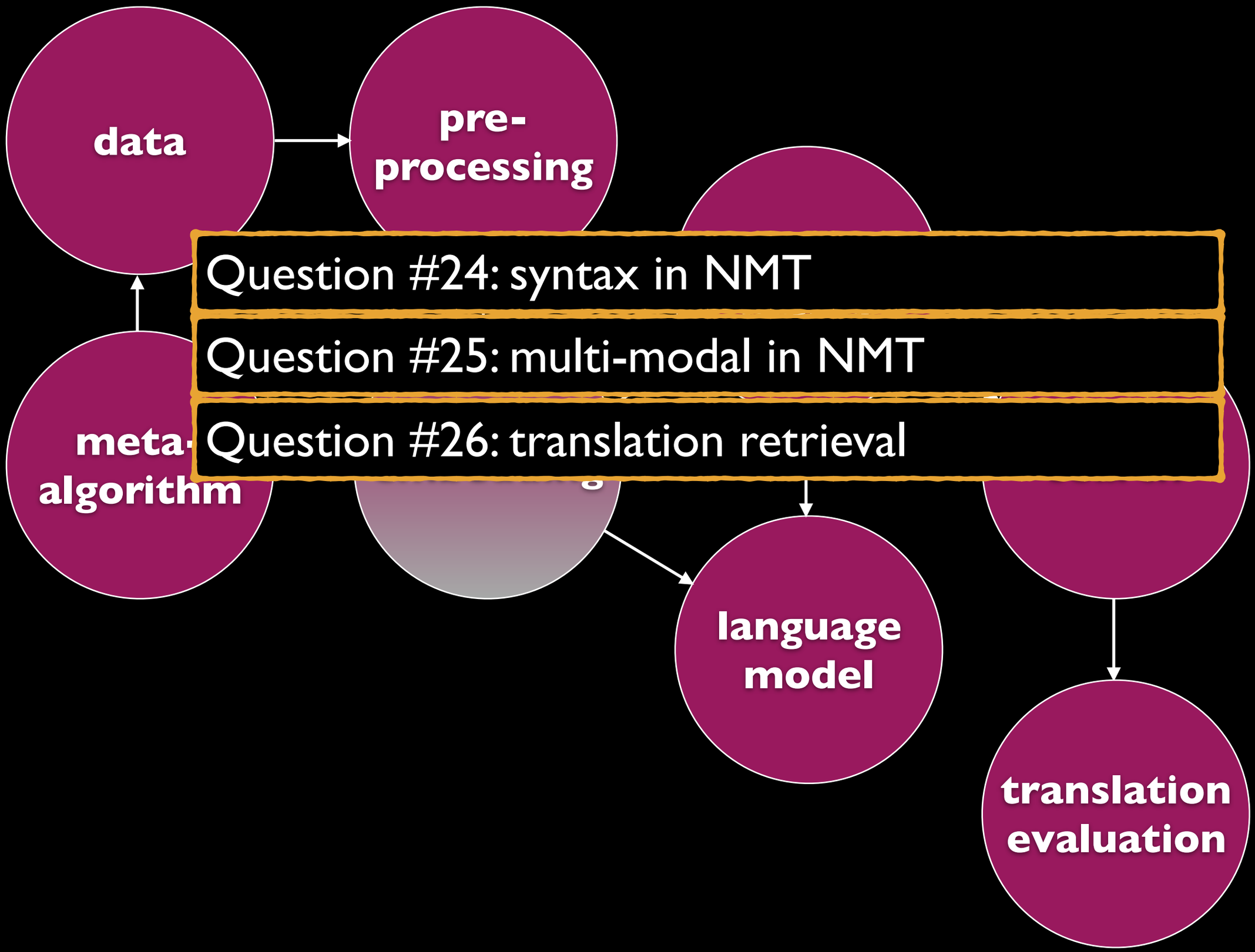
machine translation components



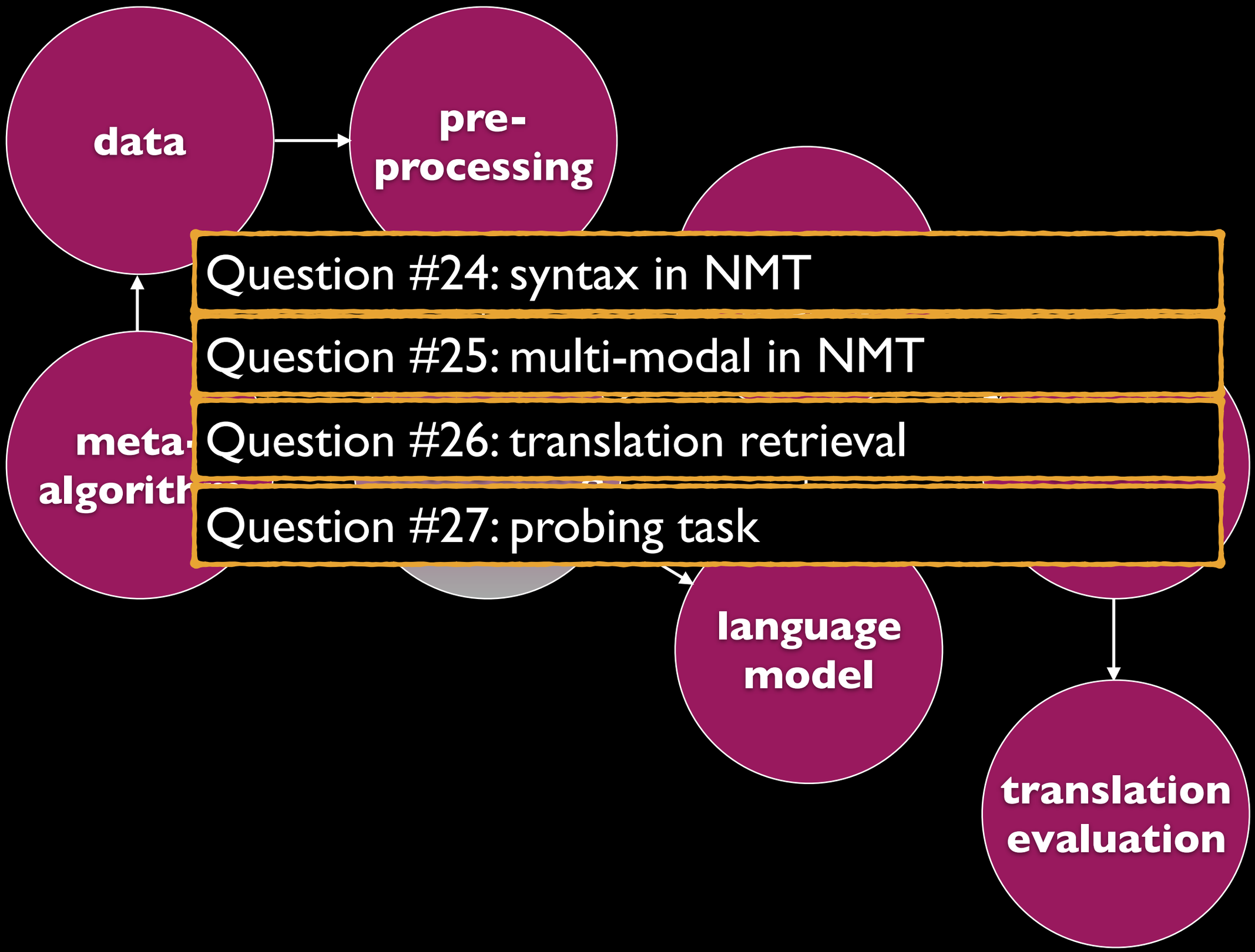
machine translation components



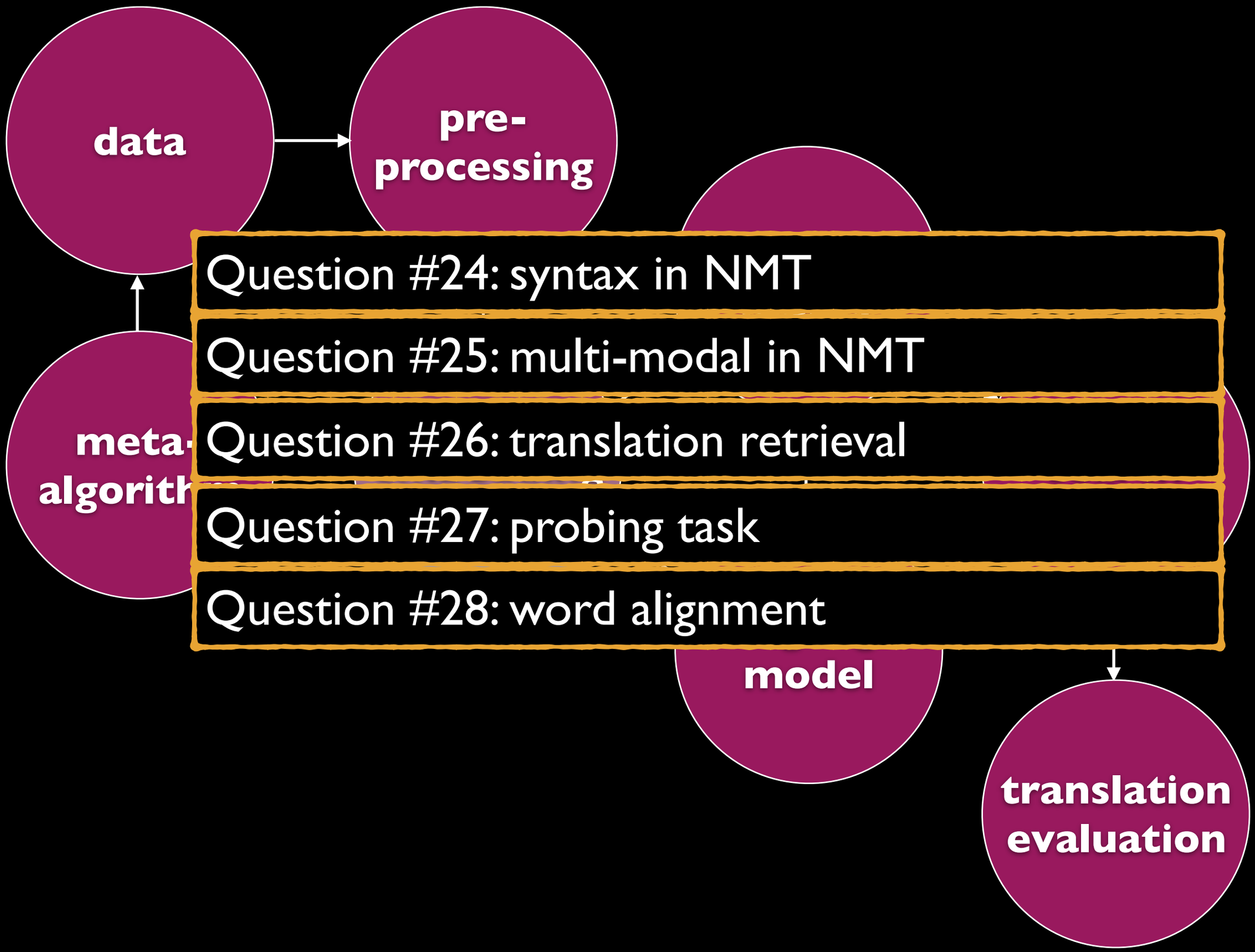
machine translation components



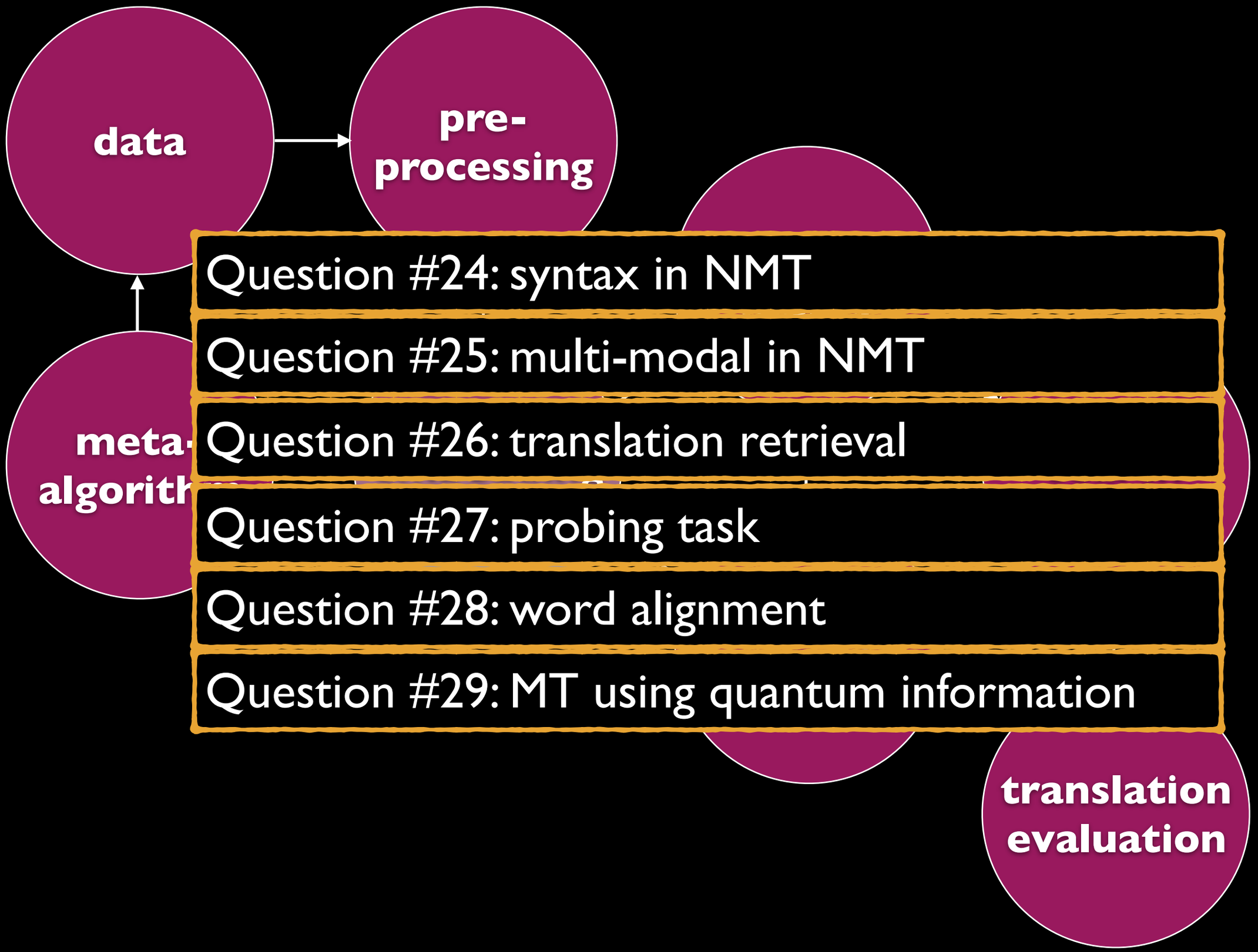
machine translation components



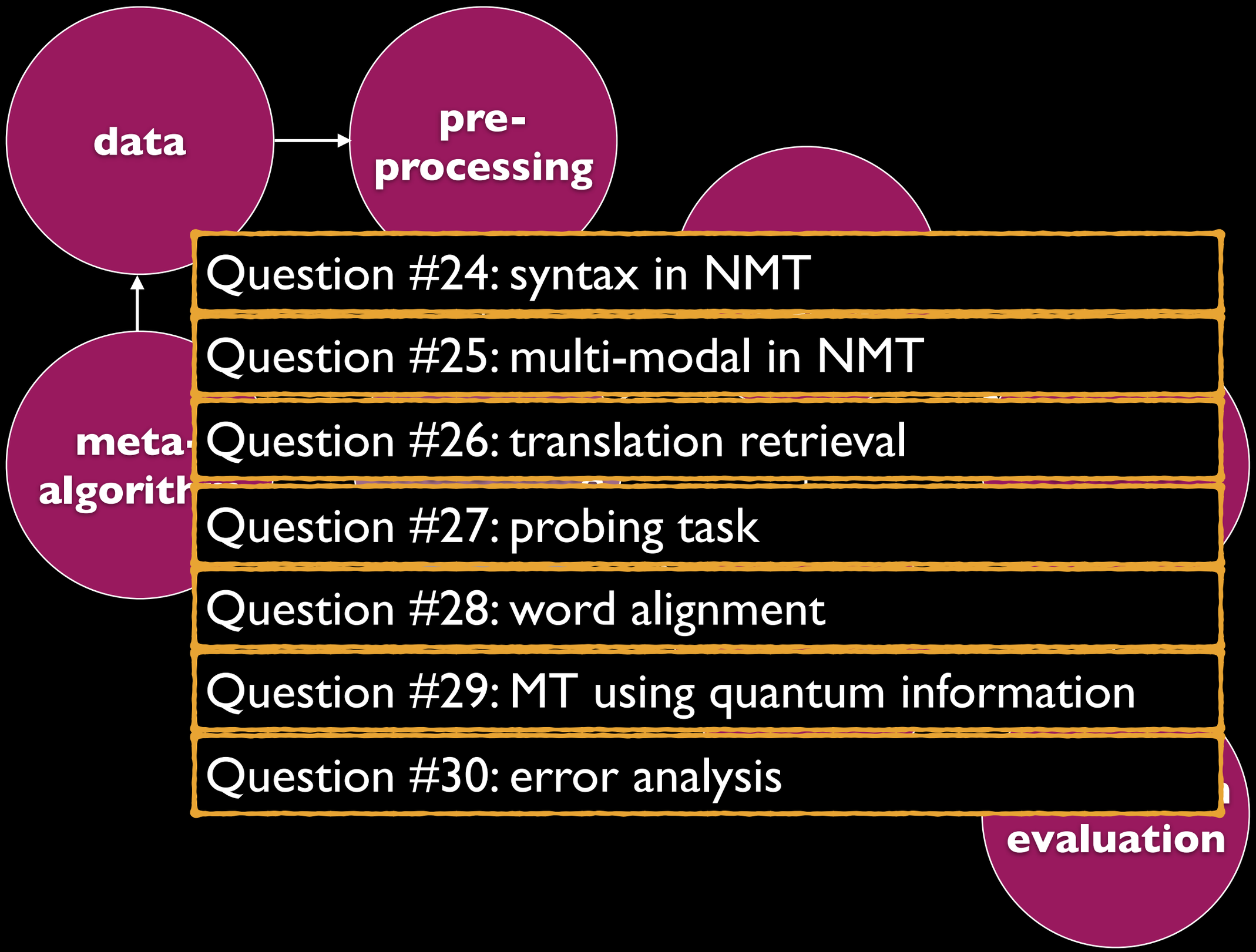
machine translation components



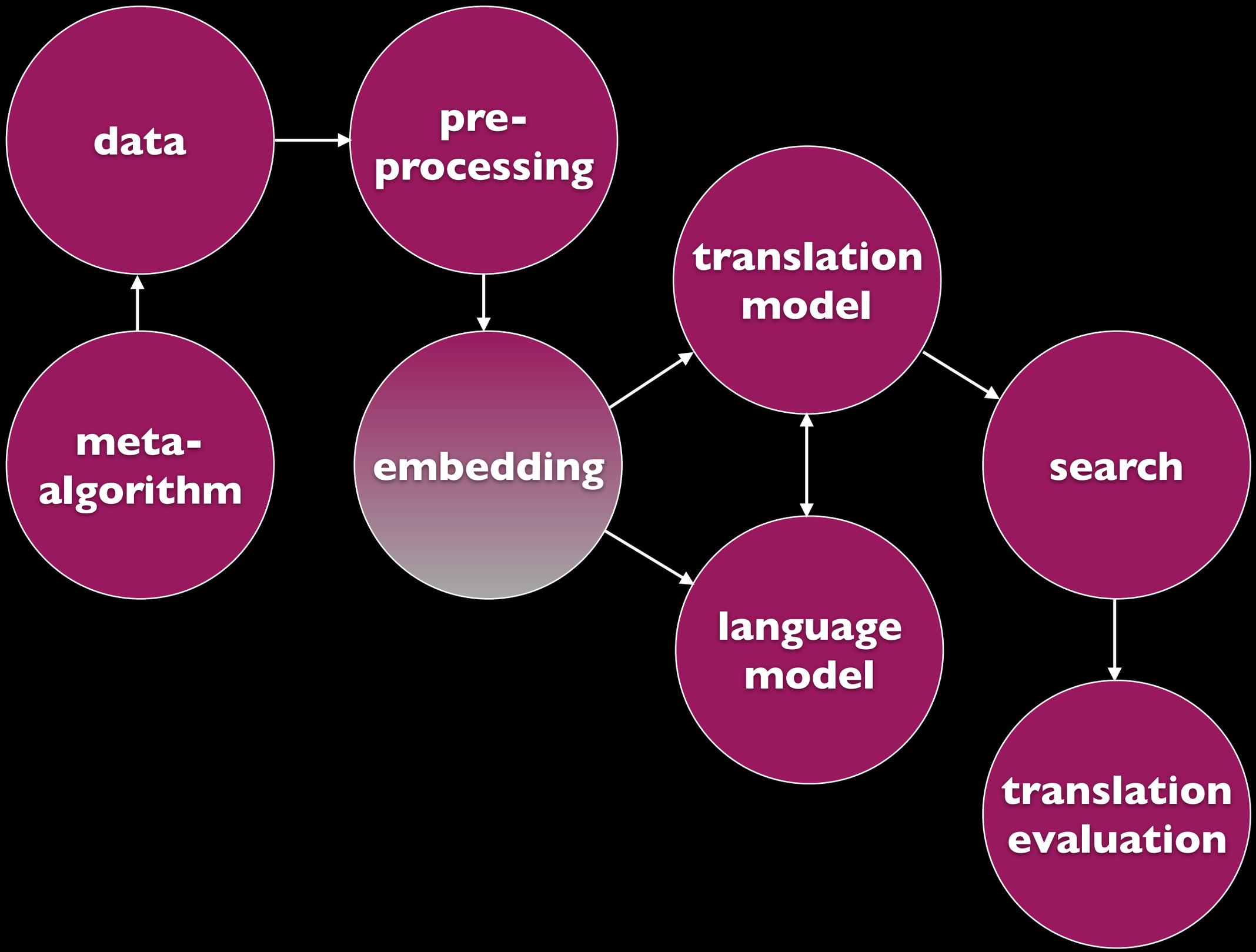
machine translation components



machine translation components



machine translation components



30 Questions

Question #1: how to enhance NMT robustness?

Question #2: how to increase interpretability?

Question #3: better text embedding/representation?
BERT, ELMO, GloVec, FastText, …

Question #4: contextual memory in language model

Question #5: affective neuron activation function

Question #6: better training criterion?
Maximum Likelihood, squared error, MAP, cross-entropy, minimum risk, ..

Question #7: better training algorithm?
error back propagation, contrastive estimation, …

Question #8: more efficient or controlled search? binary NMT, constraint

Question #9: higher correlation with human judgement? rich literature

Question #10: better quality estimation?

Question #11: text normalization

Question #12: better subword?

Question #13: monolingual and bilingual sentence segmentation

Question #14: domain adaptation

Question #15: what can we borrow from statistical MT?

Question #16: better subword e.g. with morphology?

Question #17: unseen words?

Question #18: named entities?

Question #19: higher quality in unsupervised MT?

Question #20: back translation

Question #21: pivot translation

Question #22: multi-lingual and zero resource

Question #23: interlingua exists? [Lu, et.al., 18]

Question #24: syntax in NMT

Question #25: multi-modal in NMT

Question #26: translation retrieval

Question #27: probing task

Question #28: word alignment

Question #29: MT using quantum information

Question #30: error analysis

questions?