Linguistic scaffolds for policy learning

Jacob Andreas

Berkeley → Microsoft Semantic Machines → MIT

~~Linguistic scaffolds for policy learning~~

Work on language!

Jacob Andreas

Berkeley → Microsoft Semantic Machines → MIT

What RL can do for language

What language can do for RL

```
replace the last letter of the word
drop head
change the final letter to t i
add a z if the last character is a
every vowel becomes y
change only the first consonant to
first & last 3 letters
delete every vowel
replace all n s with c
```

What RL can do for language

Daniel Fried

Ronghang Hu

Volkan Cirik

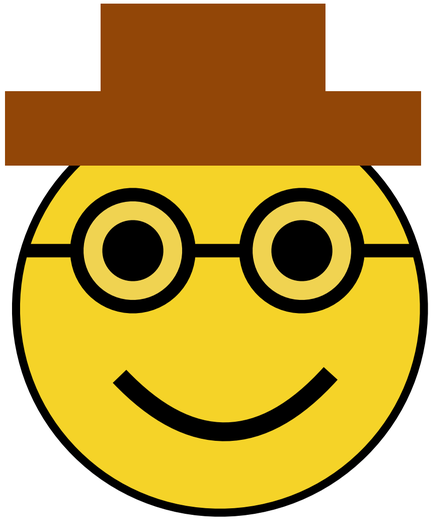
w/ Anja Rohrbach, L.P. Morency, Taylor Berg-Kirkpatrick, Trevor Darrell and Dan Klein

Generation & understanding

Turn right and walk through the kitchen. Go right into the living room and stop by the rug.

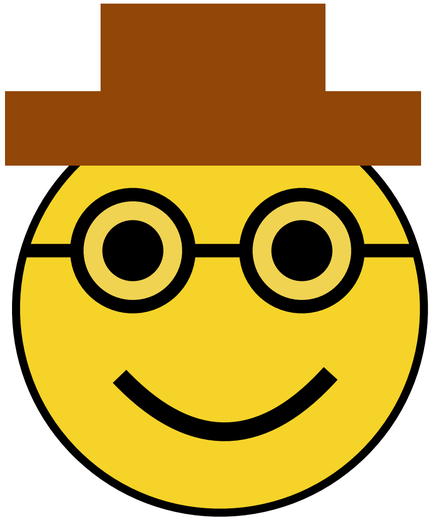
[Anderson et al. 18]

A reference game



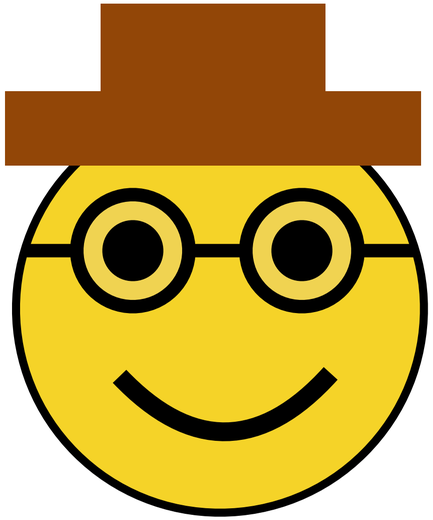
[Frank & Goodman 12]

“glasses"



[Frank & Goodman 12]

“*glasses*"

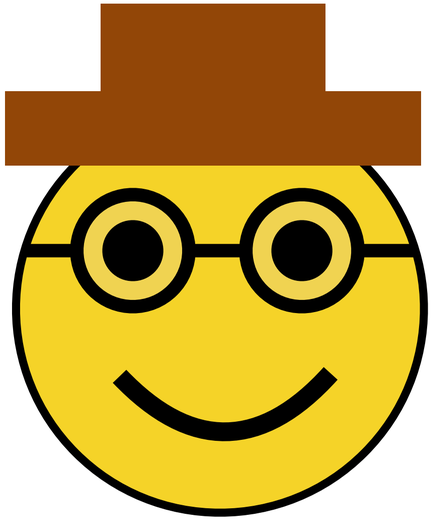


[Frank & Goodman 12]

“*glasses*"

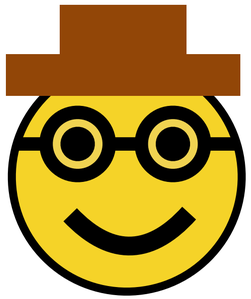
[Frank & Goodman 12]

“glasses"



[Frank & Goodman 12]

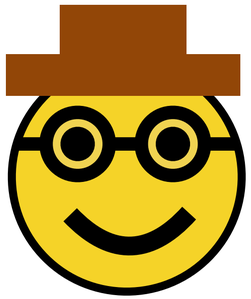
The rational speech acts model



$L_0(\ .\ \mid \textit{glasses})$	1/2	1/2
$L_0(\ .\ \mid \textit{hat})$	0	1

[Frank & Goodman 12]

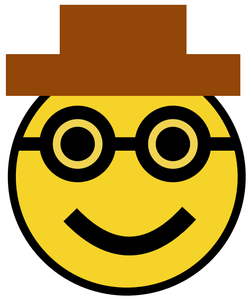
The rational speech acts model



$L_0(\ .\ \mid glasses)$	1/2	1/2
$L_0(\ .\ \mid hat)$	0	1
$S_1(\ glasses \mid .\ ) \propto L_0(\ .\ \mid glasses)$	1	1/3
$S_1(\ hat \mid .\ )$	0	2/3

[Frank & Goodman 12]

The rational speech acts model



$L_1(\,.\mid glasses\,) \propto S_1(\,glasses \mid .\,)$	3/4	1/4
$L_1(\,.\mid hat\,)$	0	1
$S_1(\,glasses \mid .\,) \propto L_0(\,.\mid glasses)$	1	1/3
$S_1(\,hat \mid .\,)$	0	2/3

[Frank & Goodman 12]

Pragmatics

Q: Do you know what time it is?

Pragmatics

Q: Do you know what time it is?

A: Yes

Pragmatics

Q: Do you know what time it is?

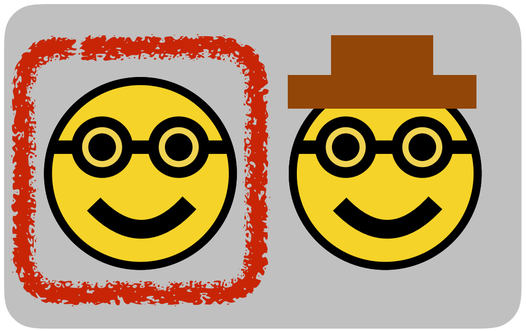
A: Yes

I find his cooking very interesting.

[Grice 70]

RSA game tree

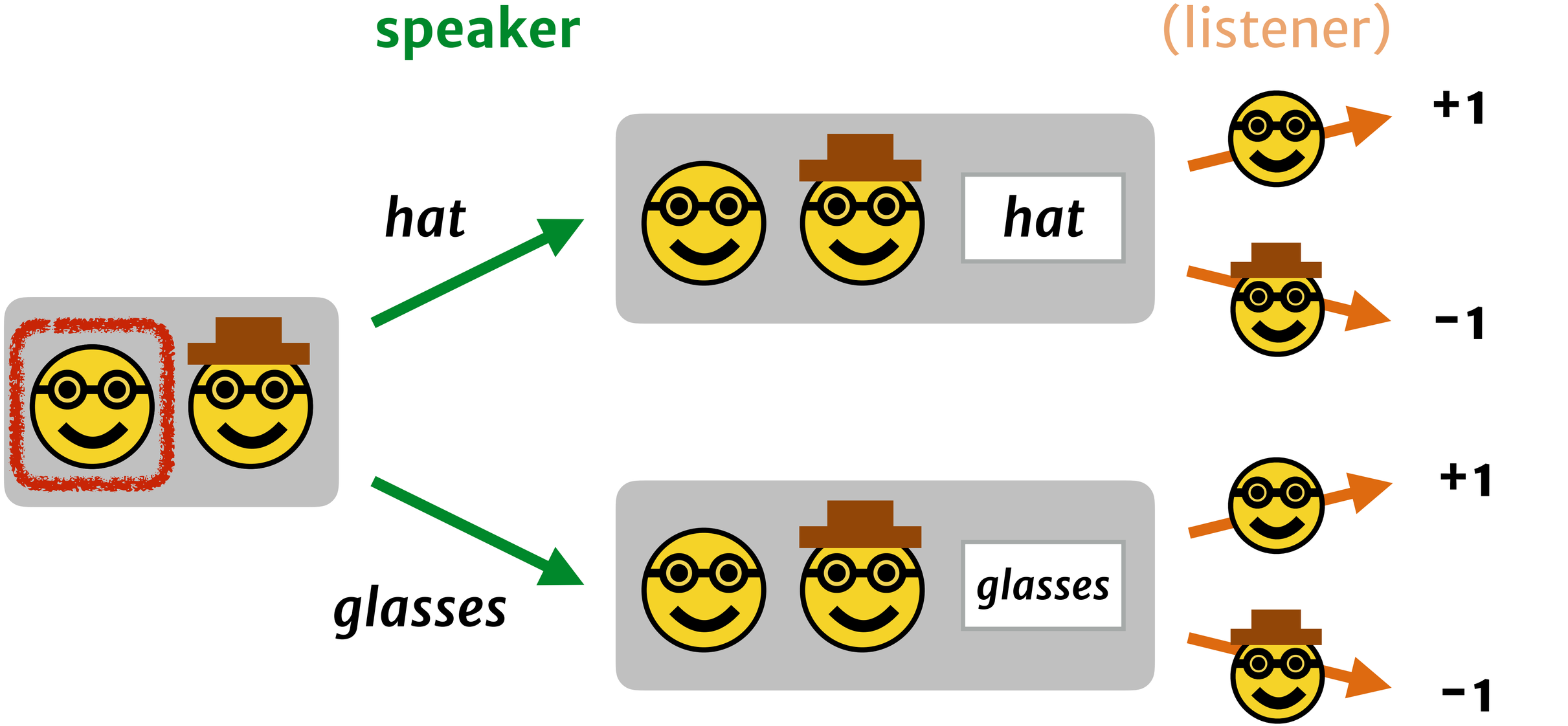
speaker



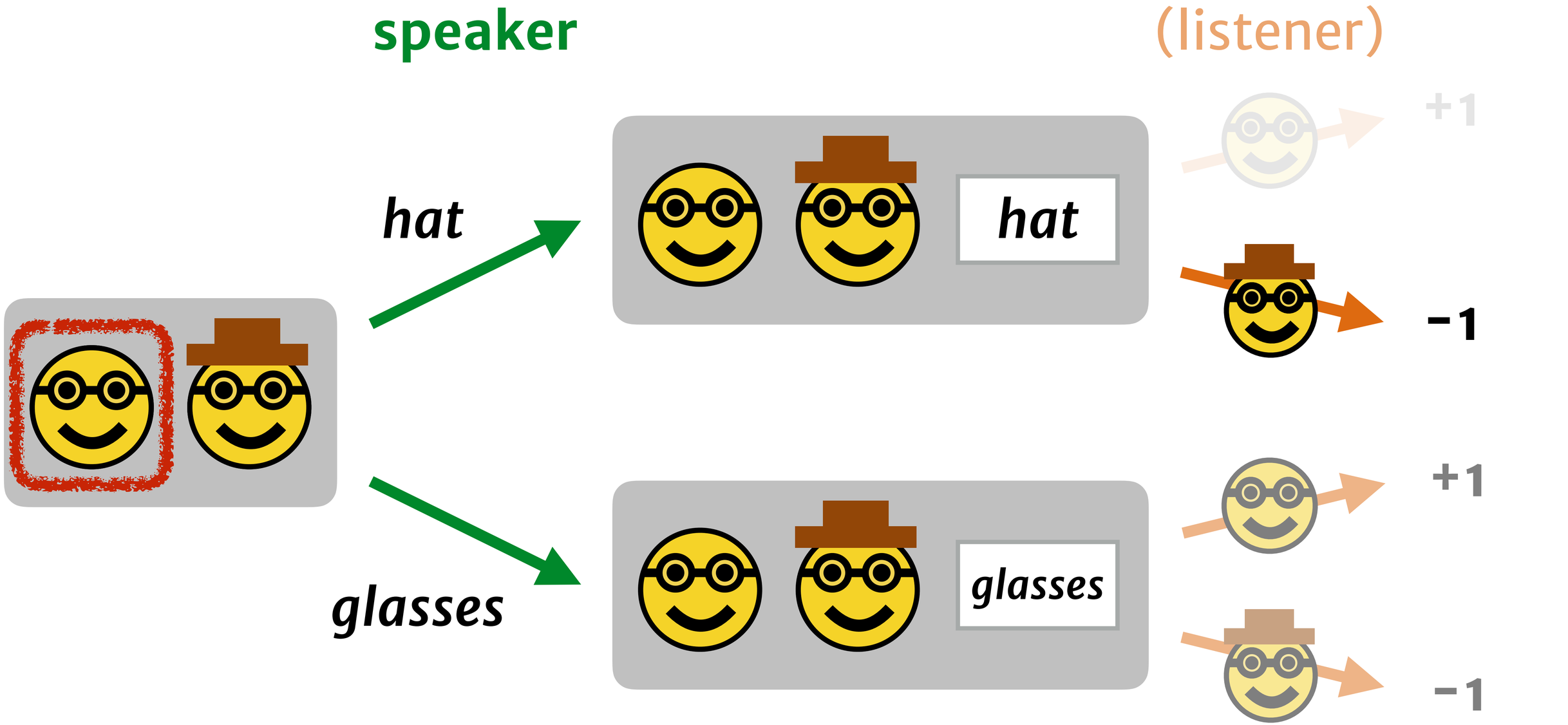
hat

glasses

RSA game tree: as speaker

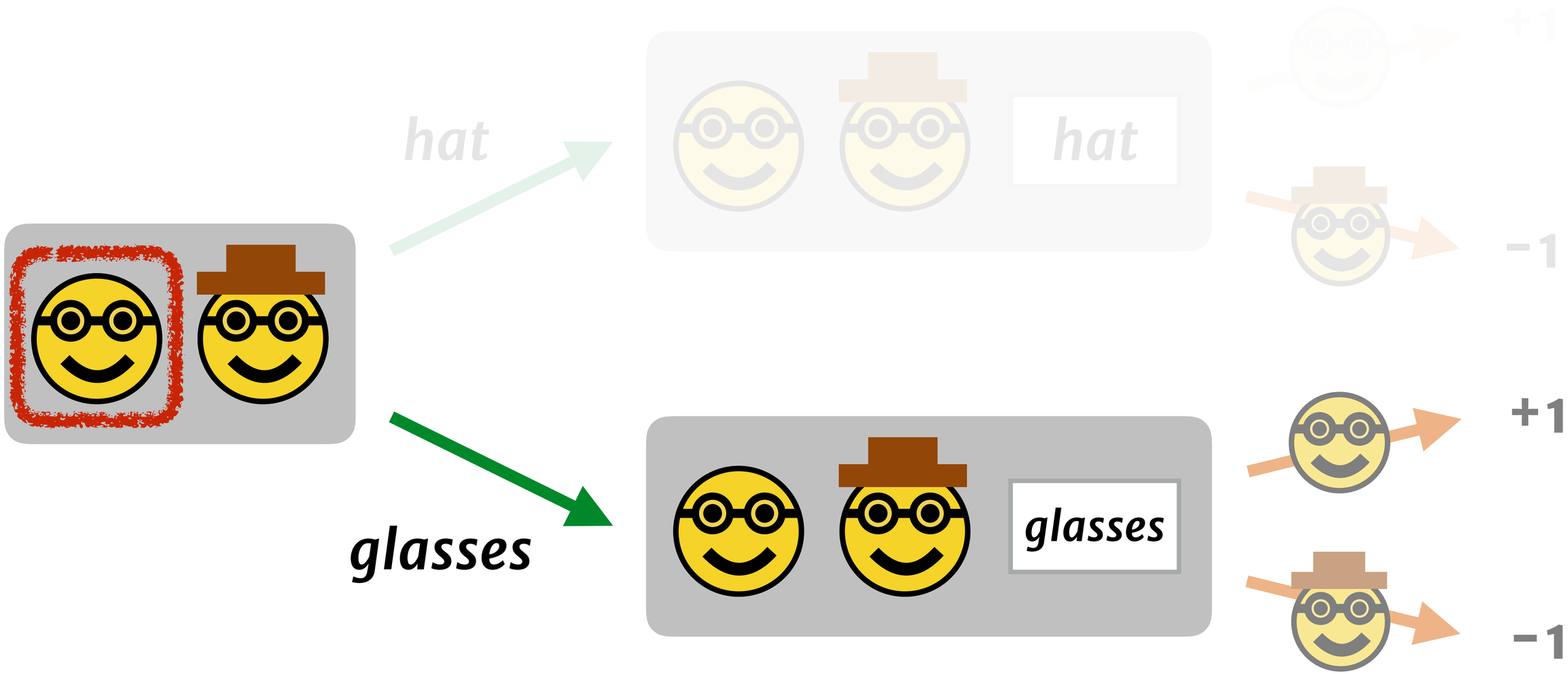


RSA game tree: as speaker



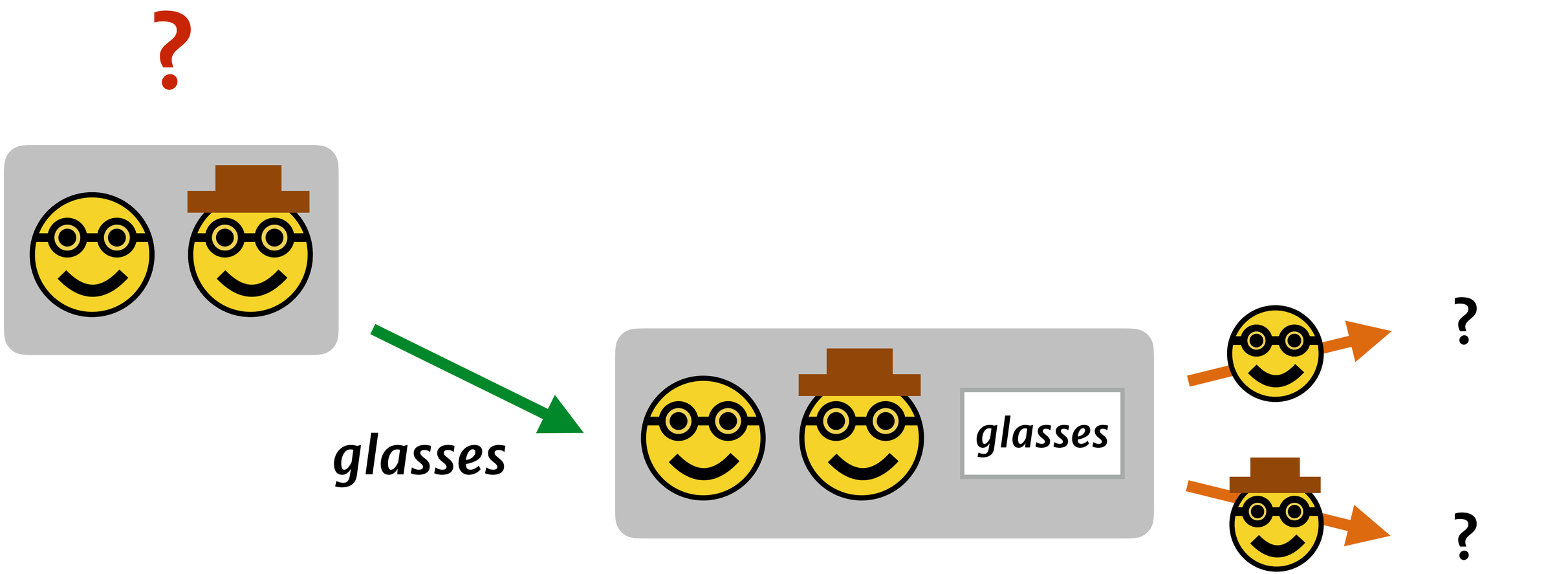
RSA game tree: as speaker

speaker (listener)



RSA game tree: as listener

(speaker) **listener**

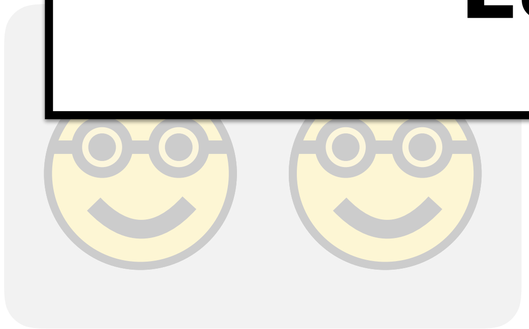


RSA game tree: as listener

(speaker) listener

?

Language use is gameplay!



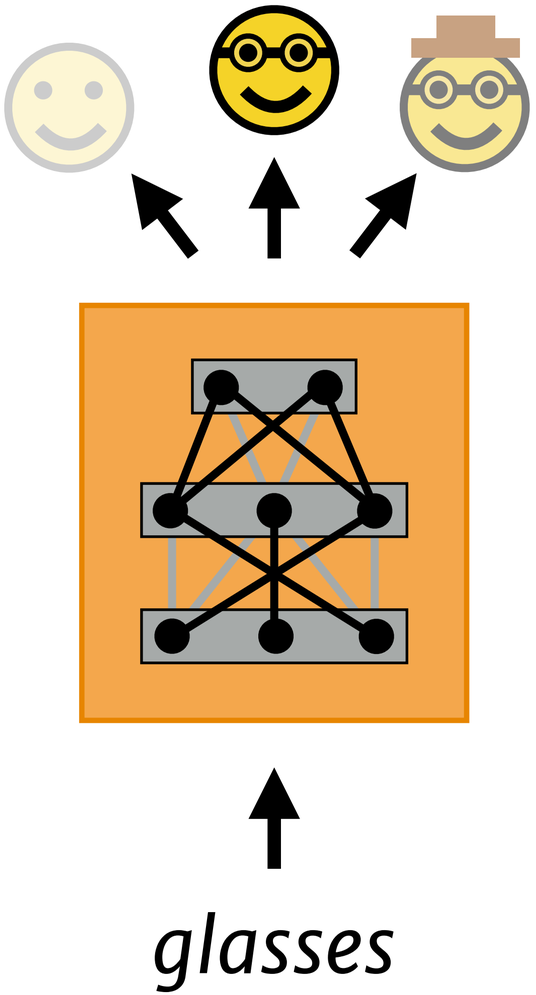
glasses

?

?

A recipe for pragmatic text generation

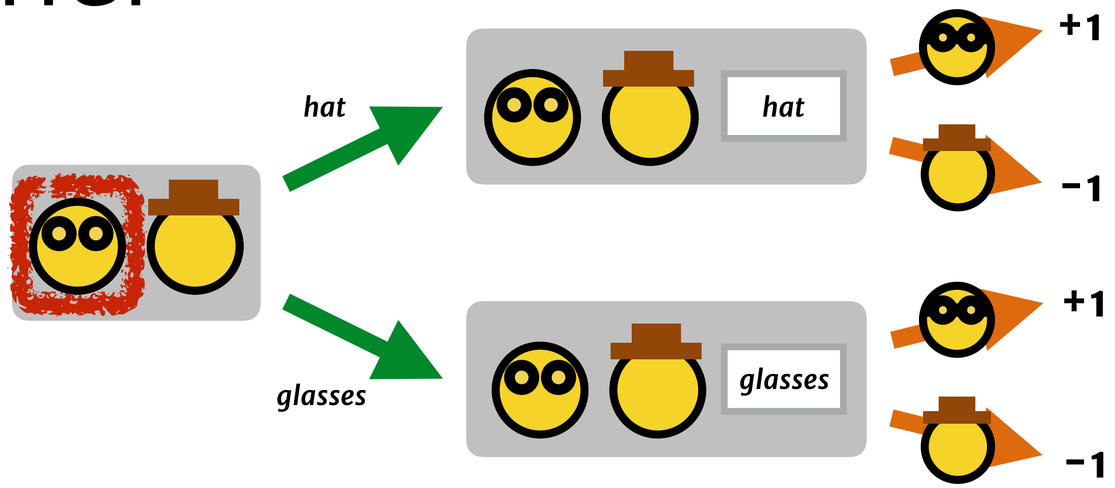
1. Train a base **listener** model



A recipe for pragmatic text generation

1. Train a base **listener** model

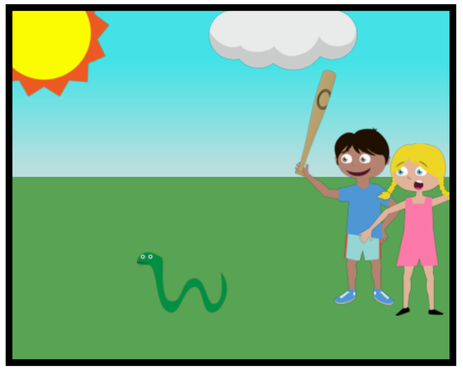
2. Train a reasoning **speaker** to win when playing with the listener

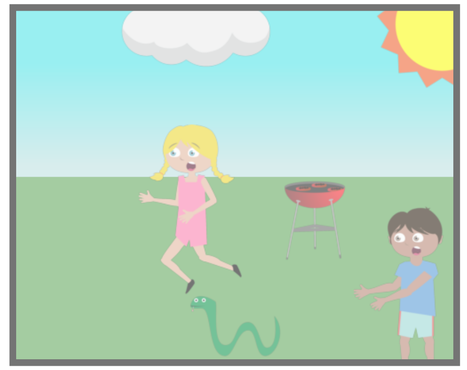


Application: image captioning

1. Train an image **retrieval** / gen model

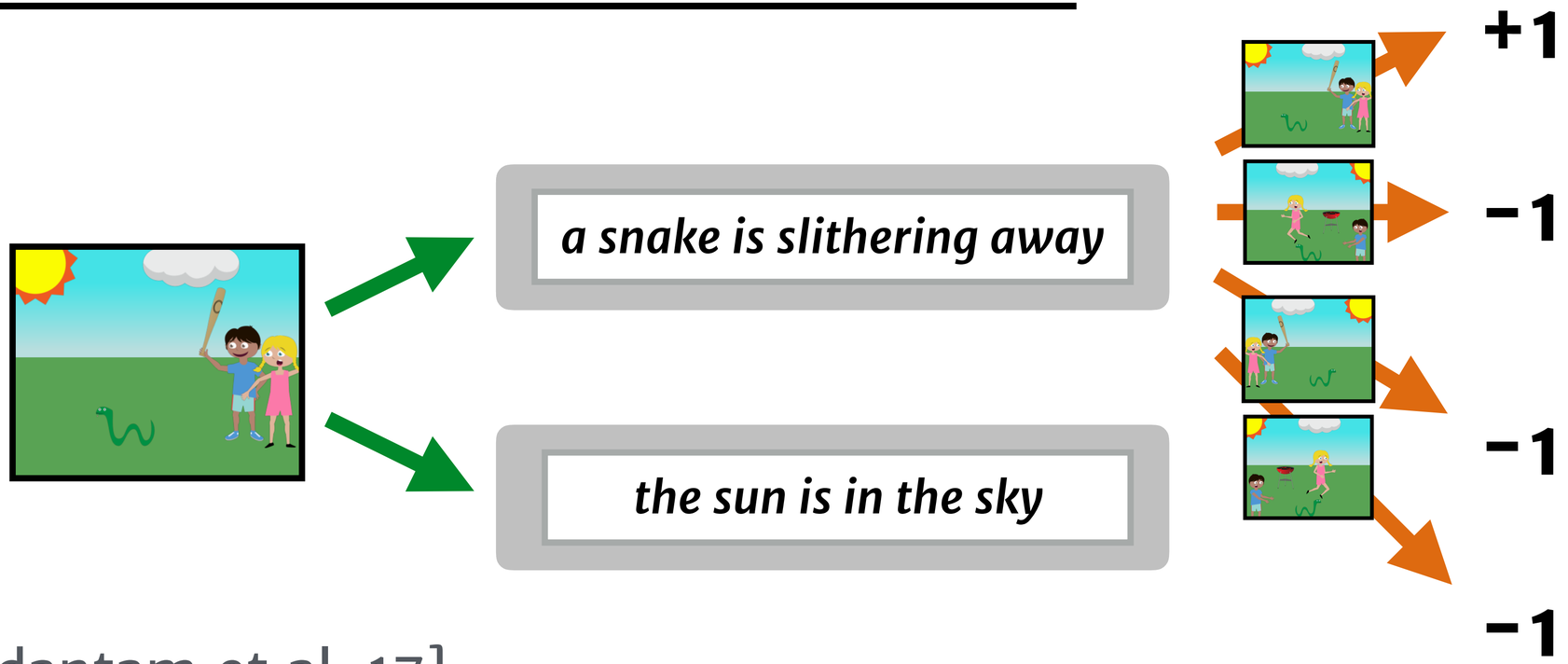
a snake is slithering away from Jenny





Application: image captioning

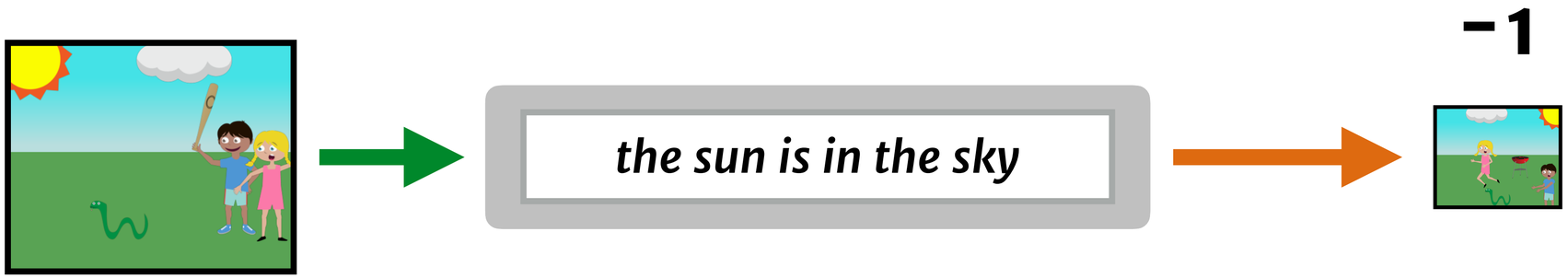
2. **Describe** images using the listener model for search at inference time



[A & Klein 16, Vedantam et al. 17]

Application: image captioning

2. **Describe** images using the listener model as a training-time reward ("self-play")



captioner model **retrieval loss**

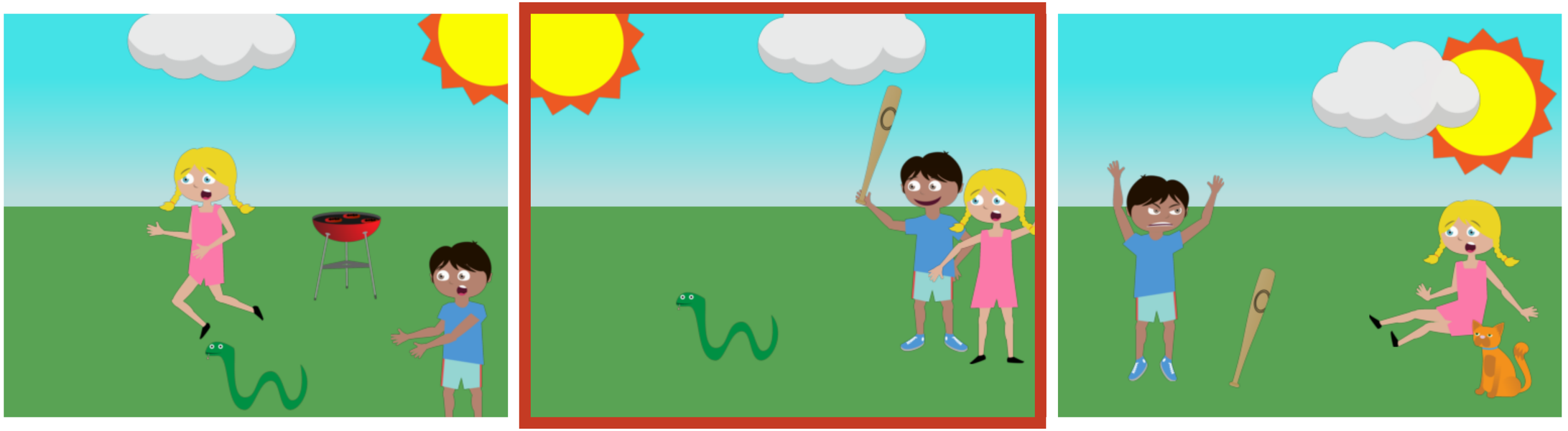
[Yu et al. 16, Mao et al. 16]

Descriptive captions [Vedantam et al. 17]

seq2seq captioner: *this bird has a yellow breast with a short pointy bill*

pragmatic captioner: *a small yellow bird with black stripes on its body and black stripe on the wings.*

Contrastive captions without contrastive data!



Mike is holding a baseball bat.

The snake is slithering away from Mike & Jenny.

[A & Klein 16]

Application: instruction generation

1. Train a base **instruction following** model

2. Train an **instruction generation** model to get the follower to goal states

Application: instruction generation

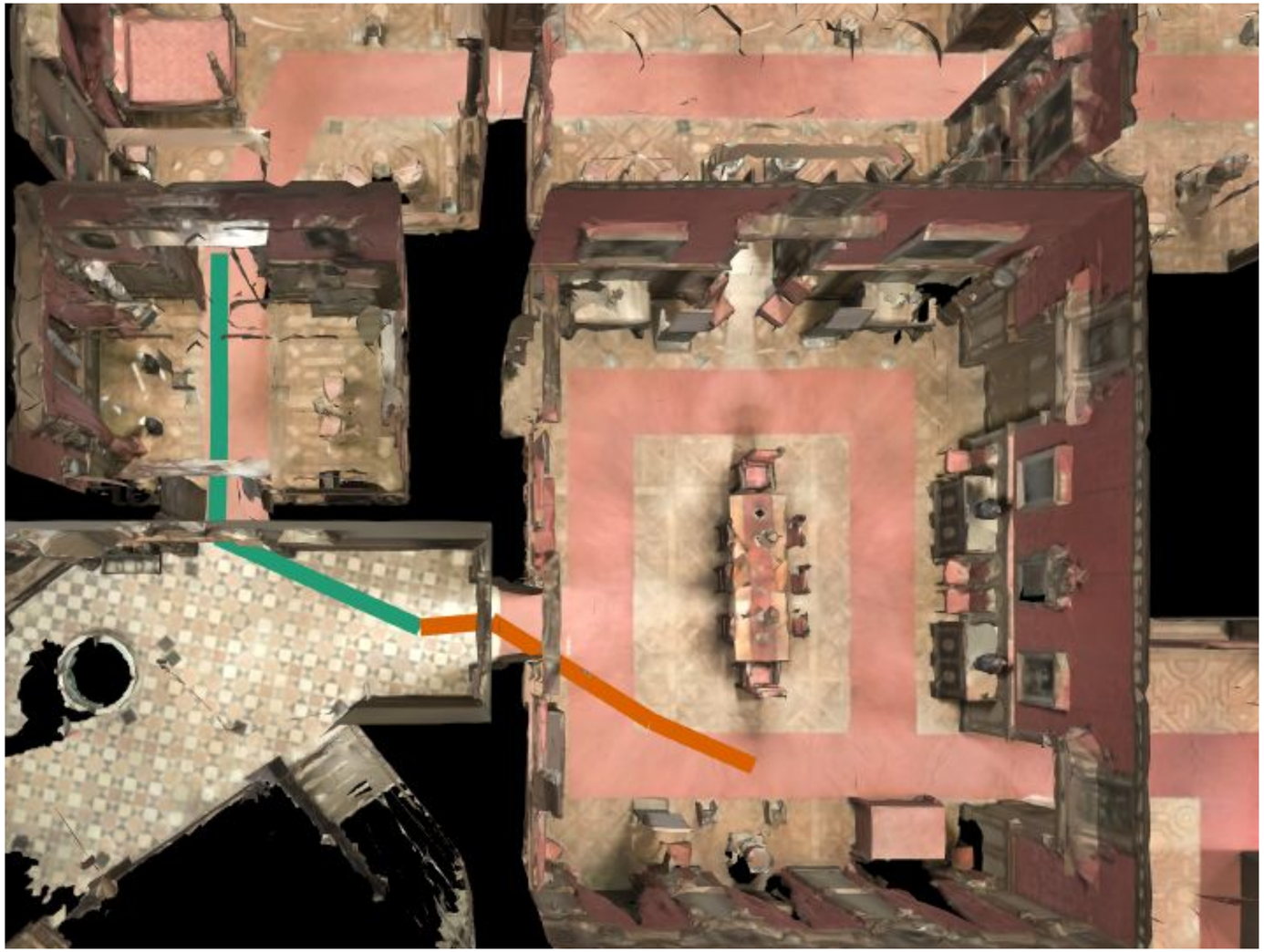
seq2seq: *Walk past the dining room table and chairs and wait there.*

speaker-listener: *Walk past the dining room table and chairs and take a right into the living room. Stop once you are on the rug.*

human: *Turn right and walk through the kitchen. Go right into the living room and stop by the rug.*

[Fried, Hu, Cirik et al. 18]

Listener mode



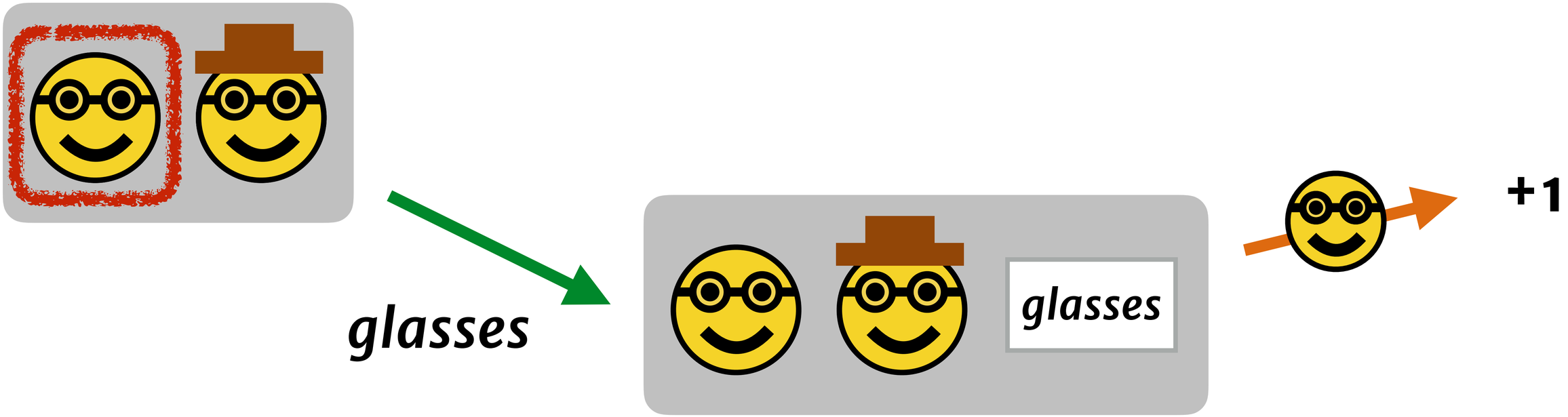
***human**: Go through the door on the right and continue straight. Stop in the next room in front of the bed.*

seq-to-seq

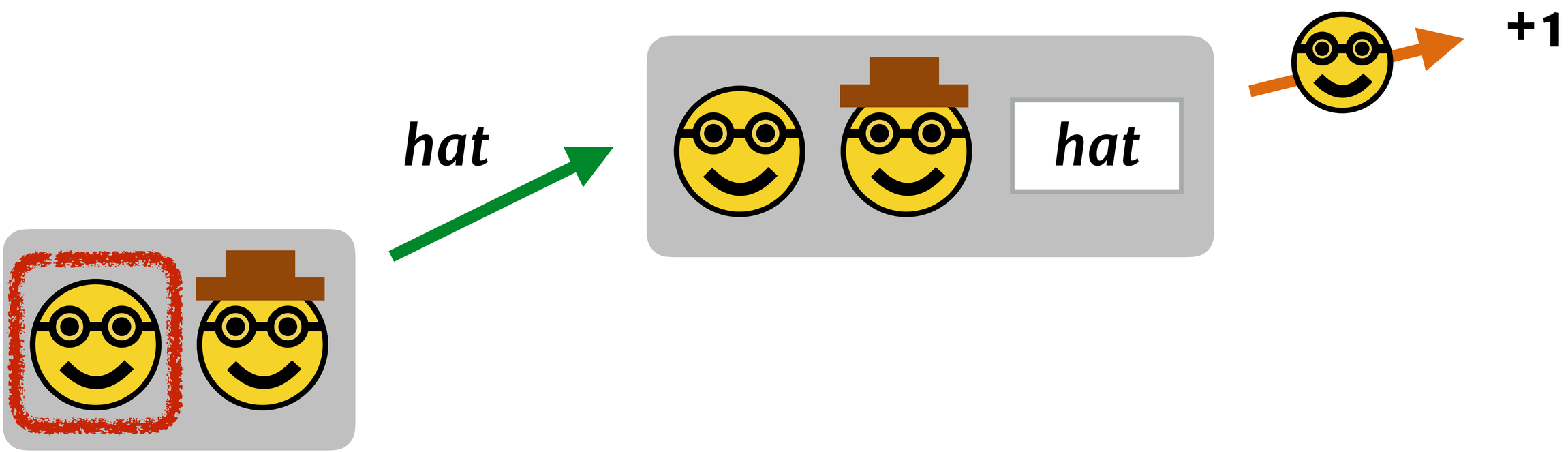
speaker-listener

[Fried, Hu, Cirik et al. 18]

The rules of the game



The rules of the game



Killer robots [Lewis et al. 17]

Bob: *i can i i everything else . . . . . .*

Alice: *balls have zero to me to me to me to me to me to me to me to me to*

Bob: *you i everything else . . . . . .*

Alice: *balls have a ball to me to me to me to me to me to me to me to me*

Killer robots [Lewis et al. 17]

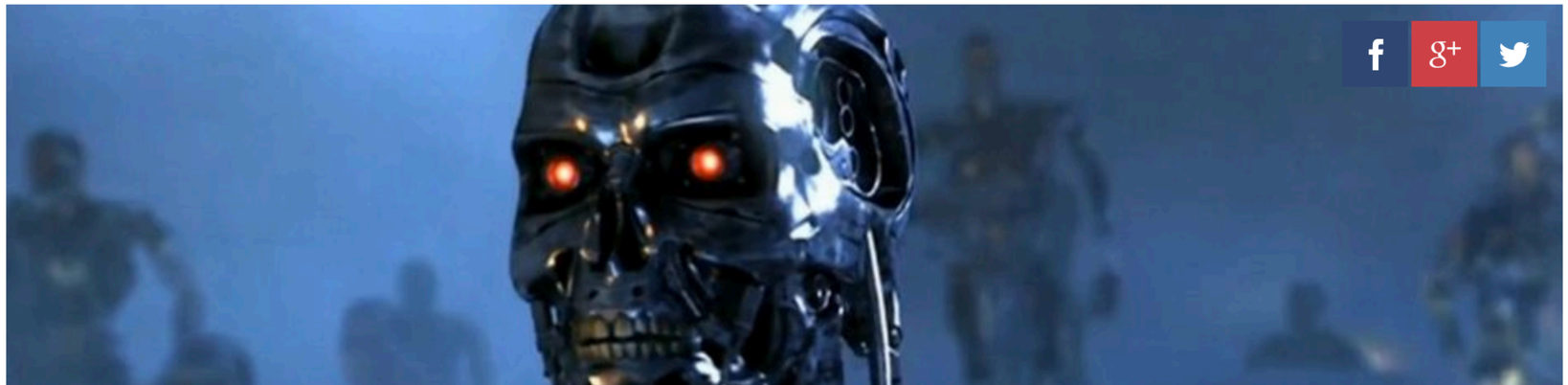
Bob: *i can i i everything else . . . . . .*

Alice: *balls have zero to me to me to me to me to me to me to me to me to*

Bob: *you i everything else . . . . . .*

Alice: *balls have a ball to me to me to me to me to me to me to me*

‘Terminator’ Come To Life? – Facebook Shuts Down Artificial Intelligence After It Developed Its Own Language



Problems to work on

How do we use tools like self-play and tree search while remaining within the rules of natural language?

How do we do efficient search in string-valued action spaces?

Problems to work on

How do we use tools like self-play and tree search while remaining within the rules of natural language?

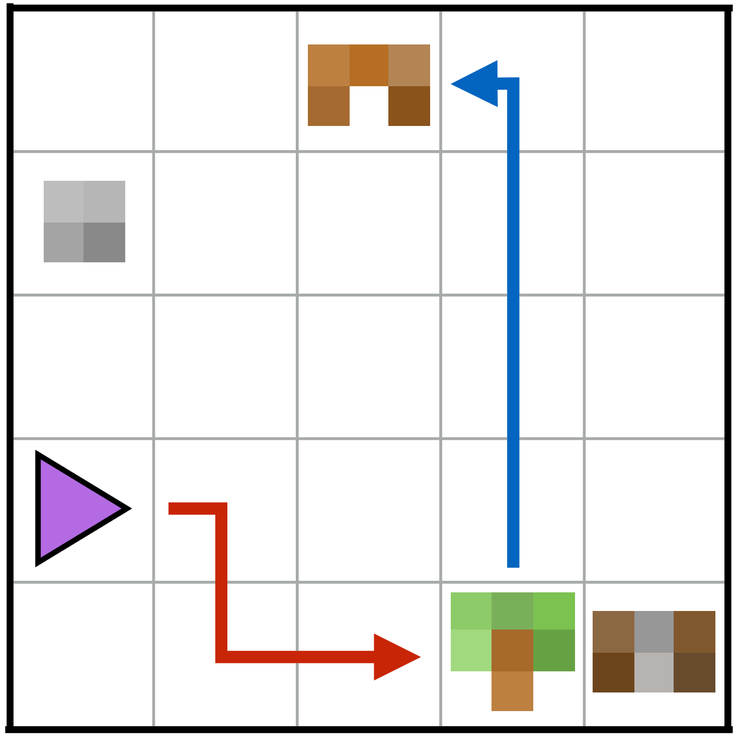
How do we do efficient search in string-valued action spaces?

What language can do for RL

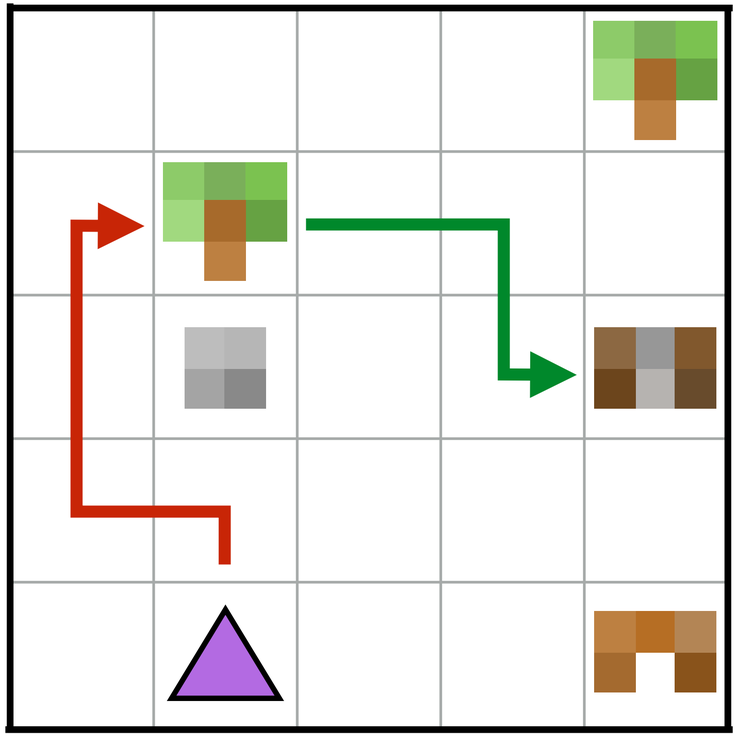
w/ Dan Klein and Sergey Levine

A crafting game

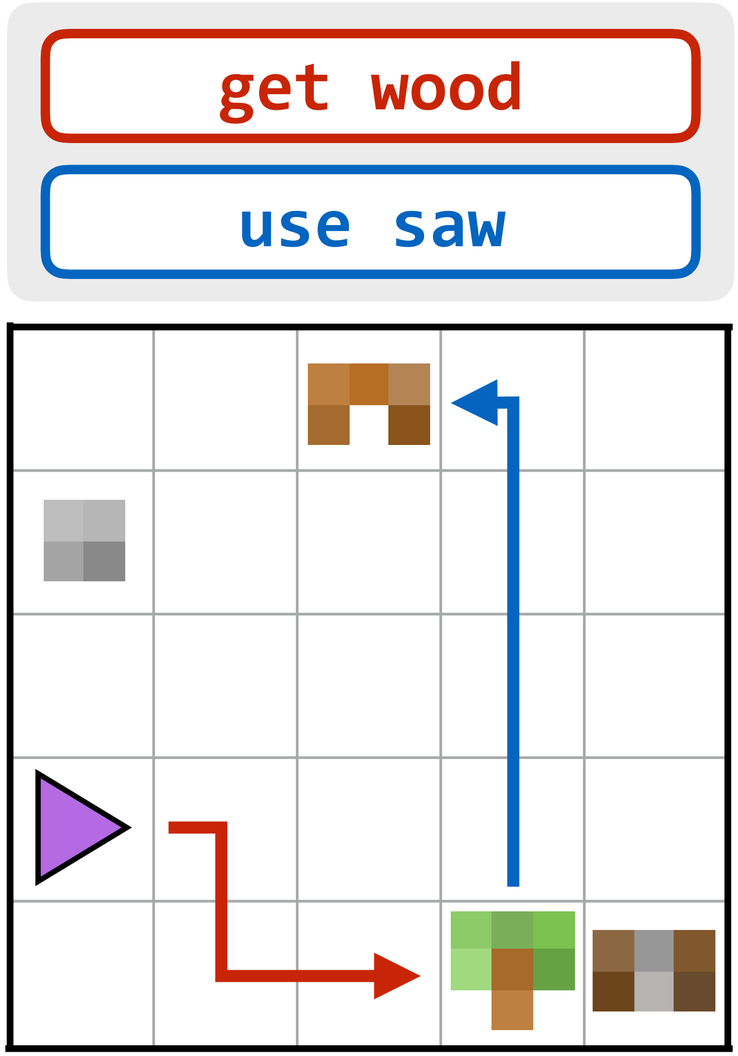
`make planks`

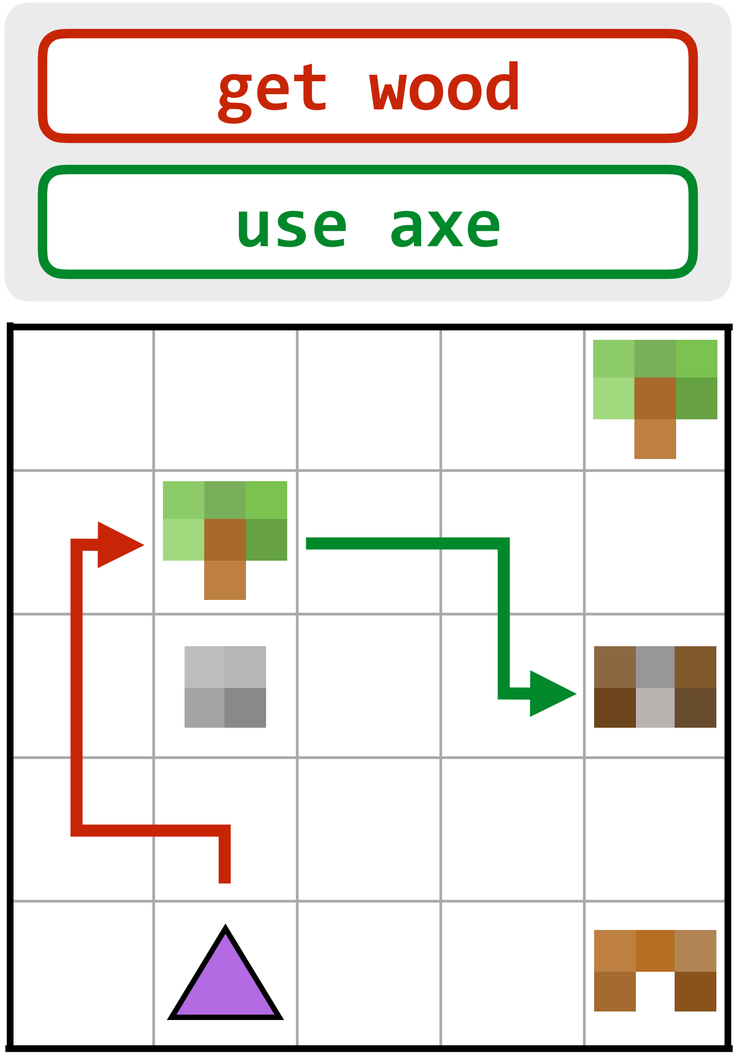


`make sticks`

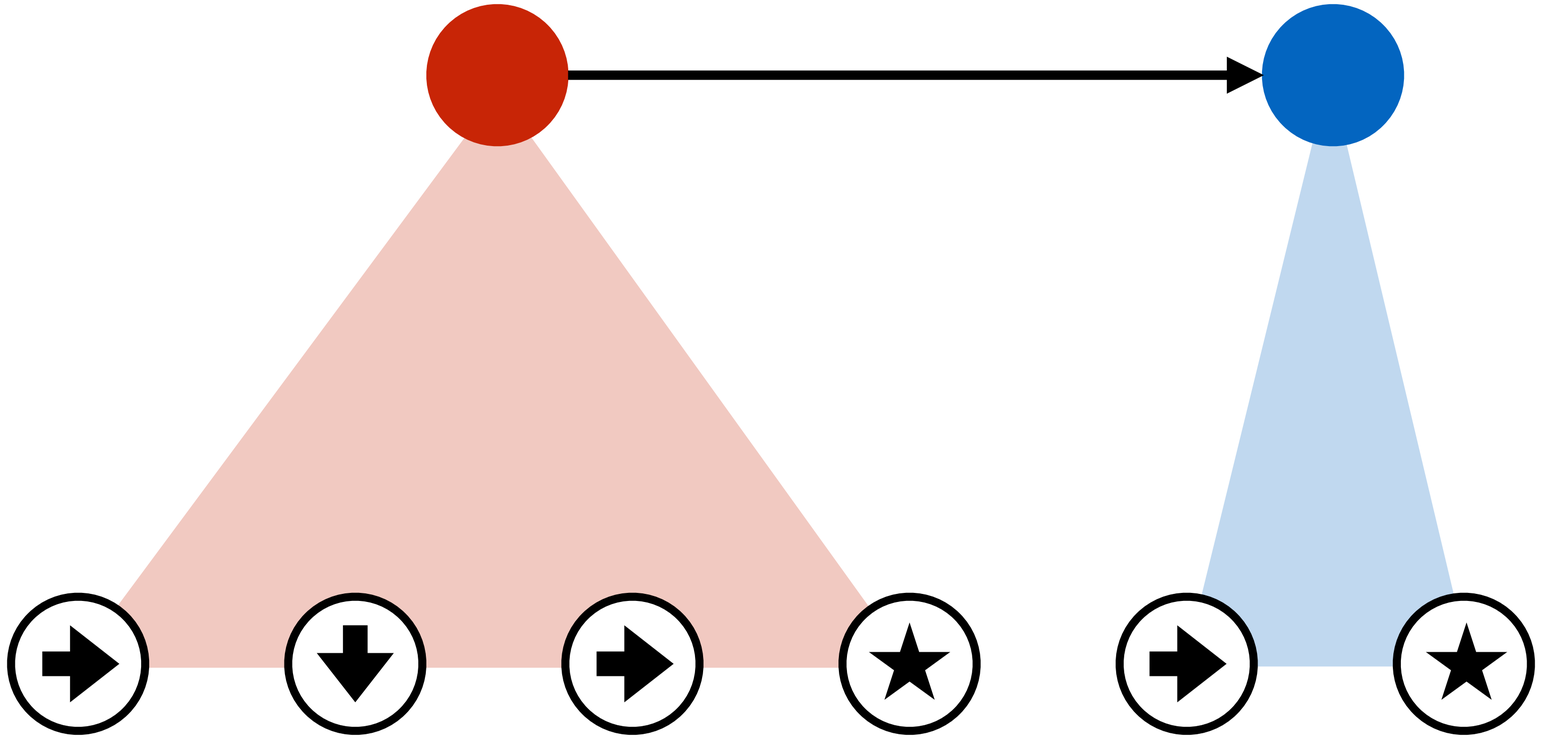


Learning with sketches



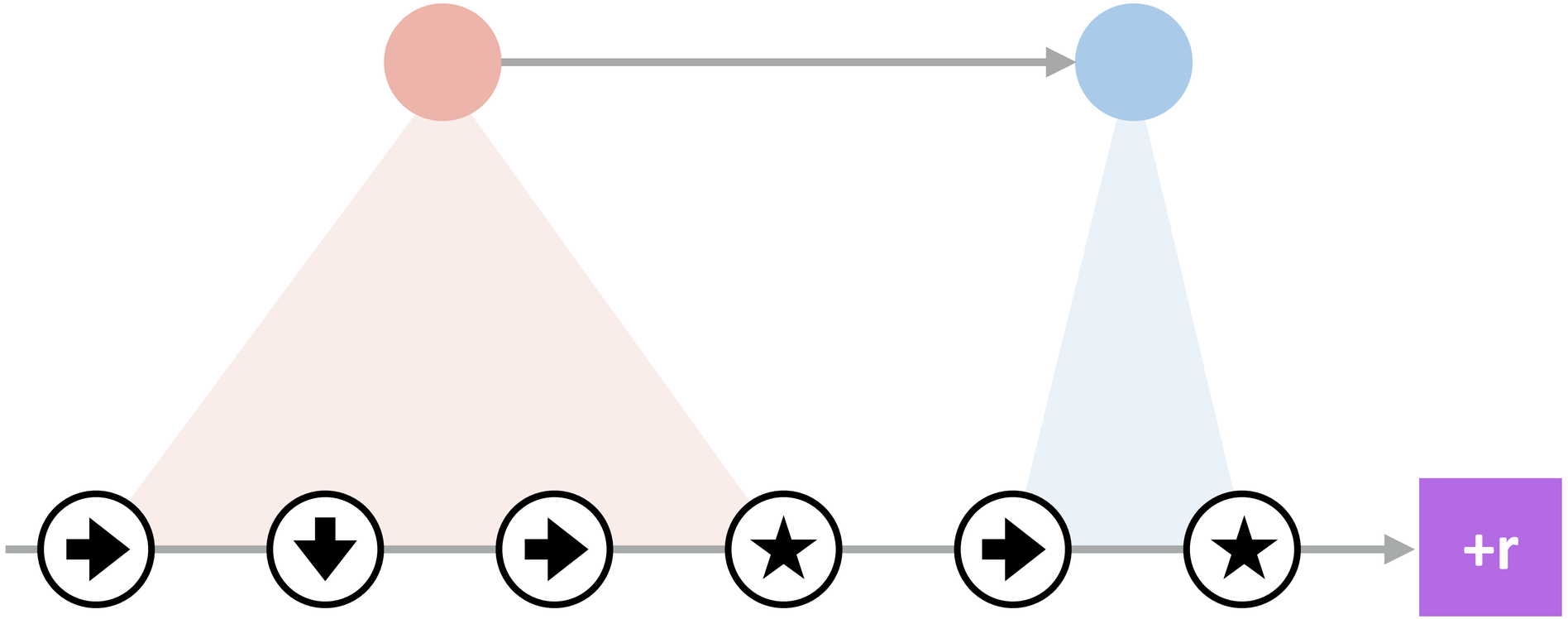


The options framework



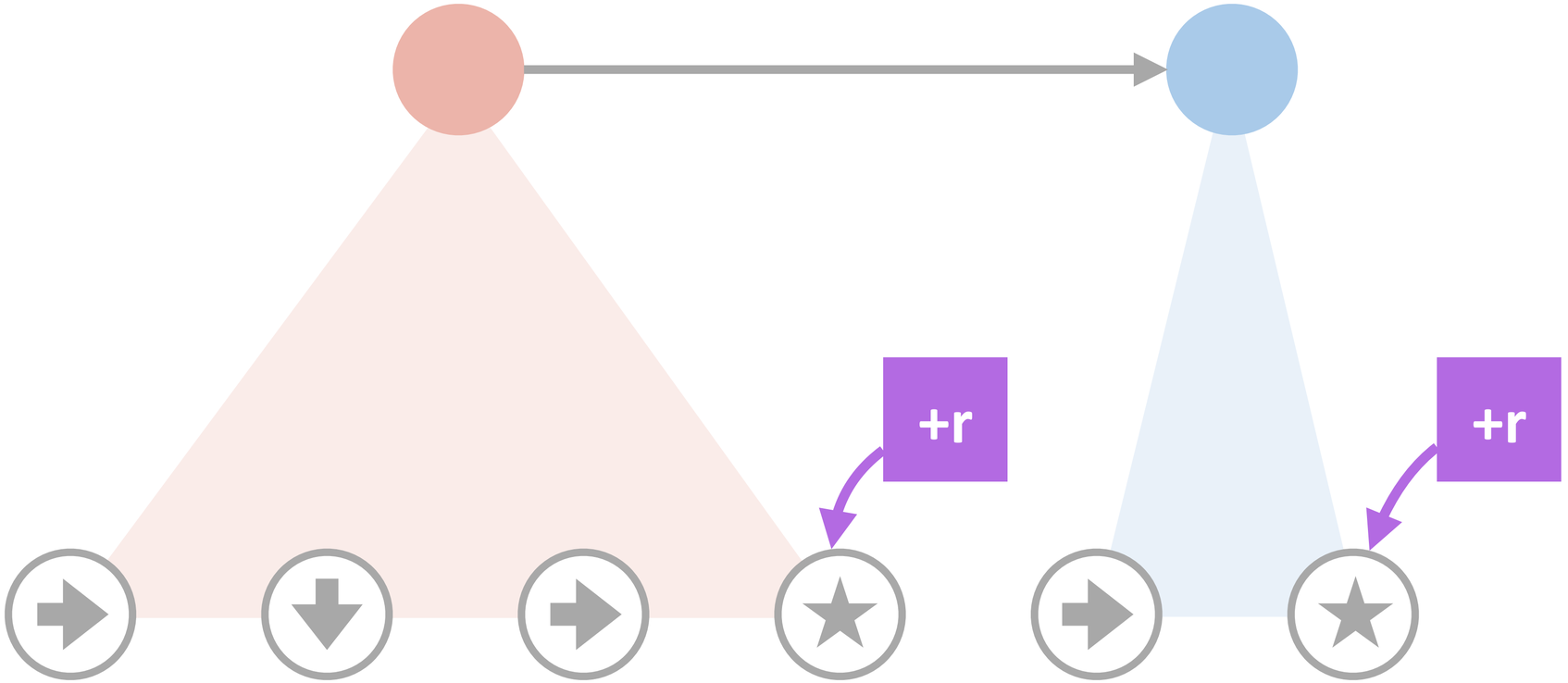
[Sutton et al. 99]

Unsupervised option learning



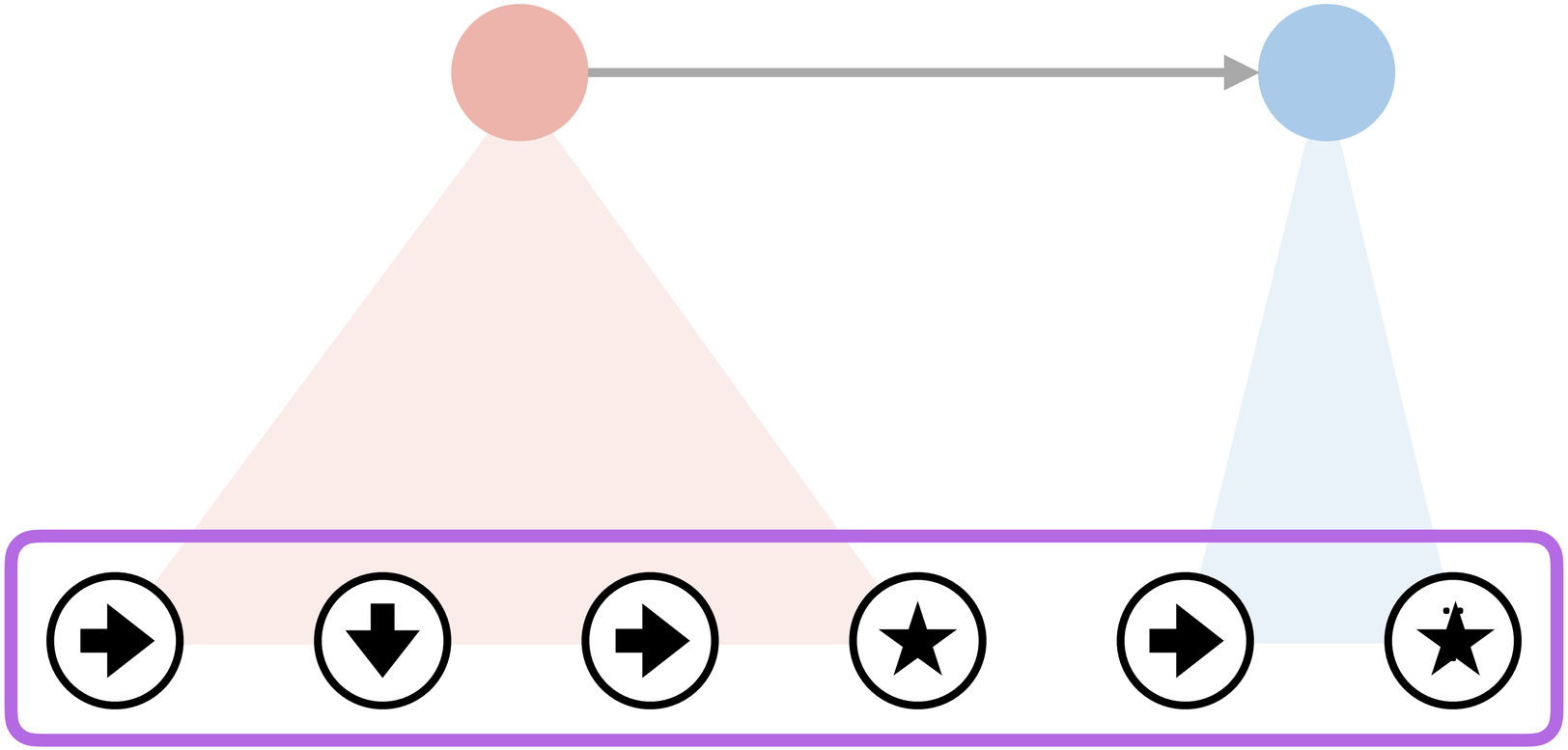
[Bacon & Precup 16]

Learning with intermediate rewards



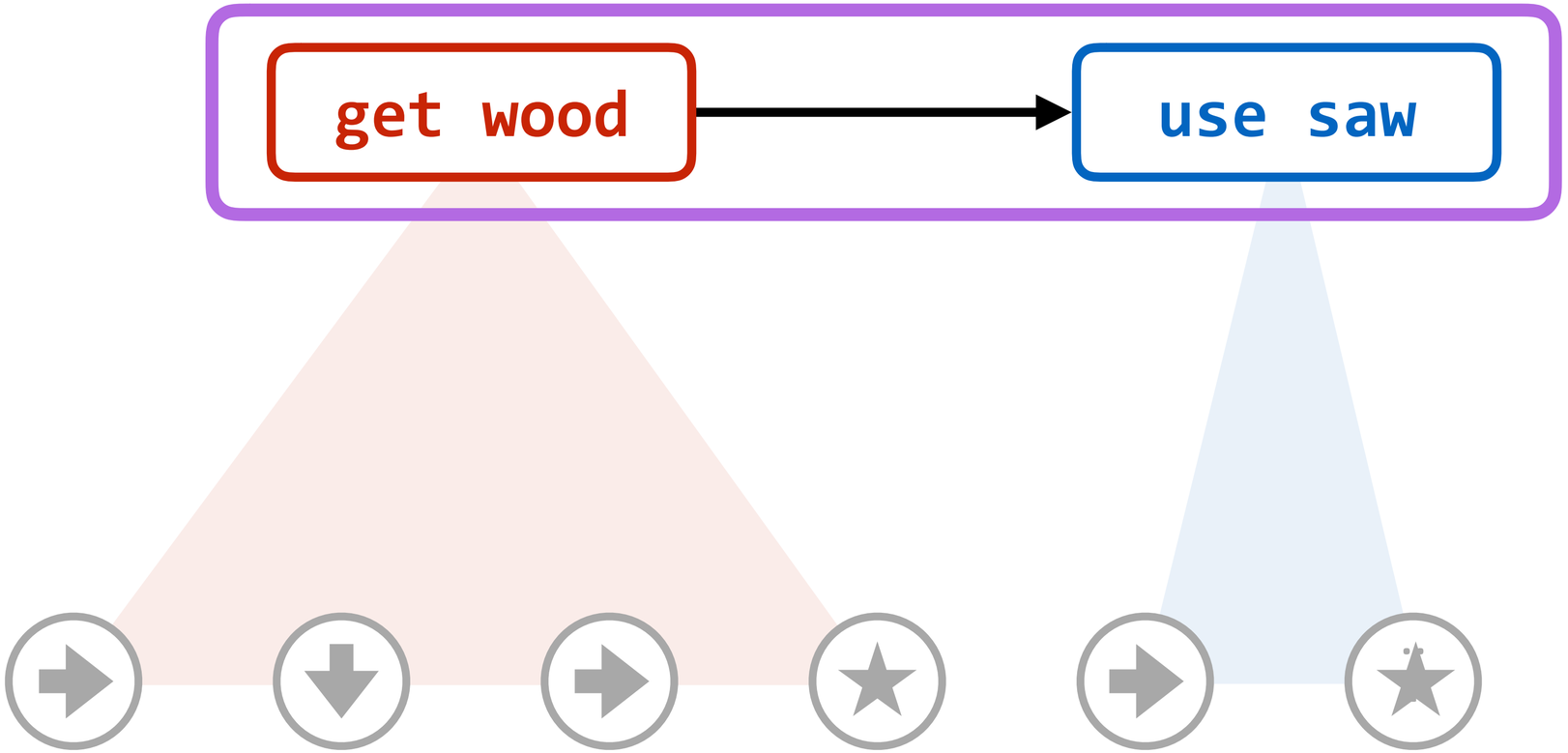
[Kearns & Singh 02, Kulkarni et al. 16]

Segmenting demonstrations



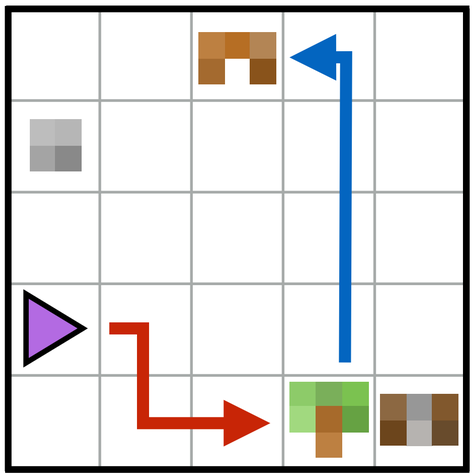
[Stolle & Precup 02, Fox & Krishnan et al. 16]

Learning from sketches



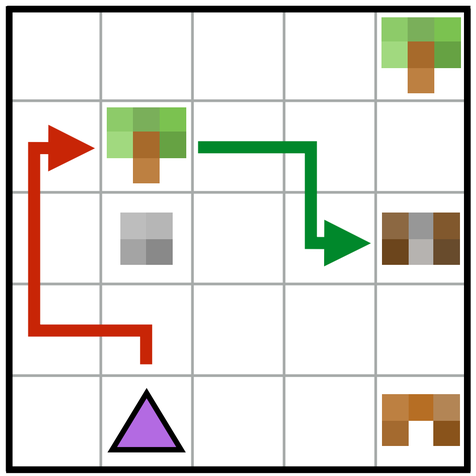
[A, Klein & Levine 17]

Modular policies



`get wood` `use saw`

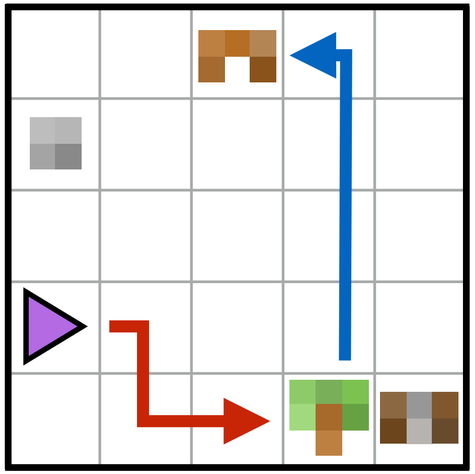
π_1 π_2



`get wood` `use axe`

π_1 π_3

Modular policies

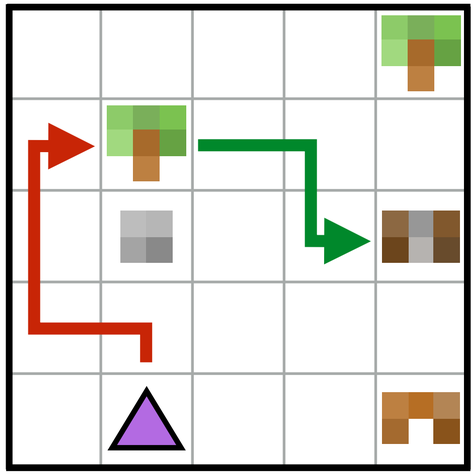


get wood

use saw

π_1

π_2



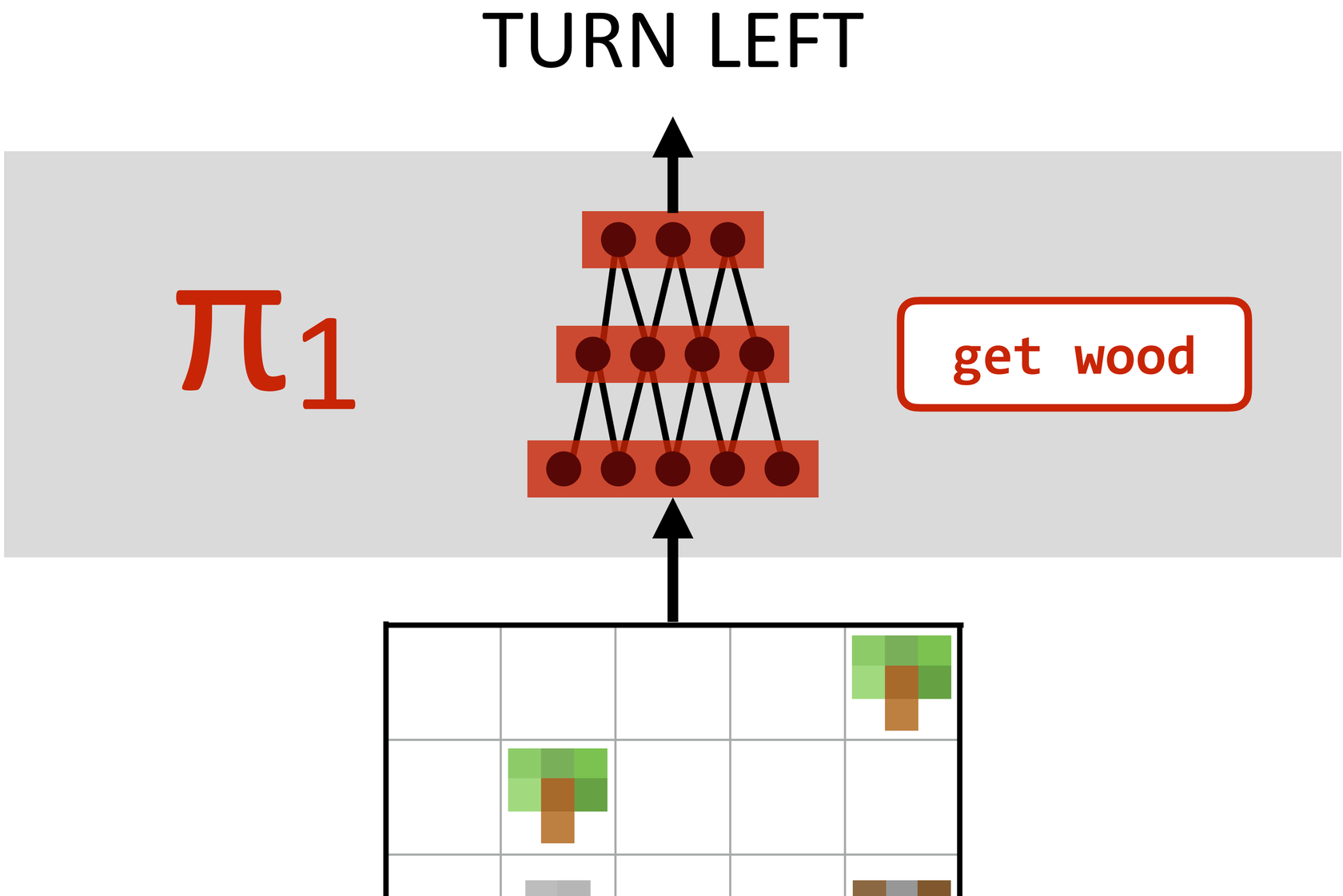
get wood

use axe

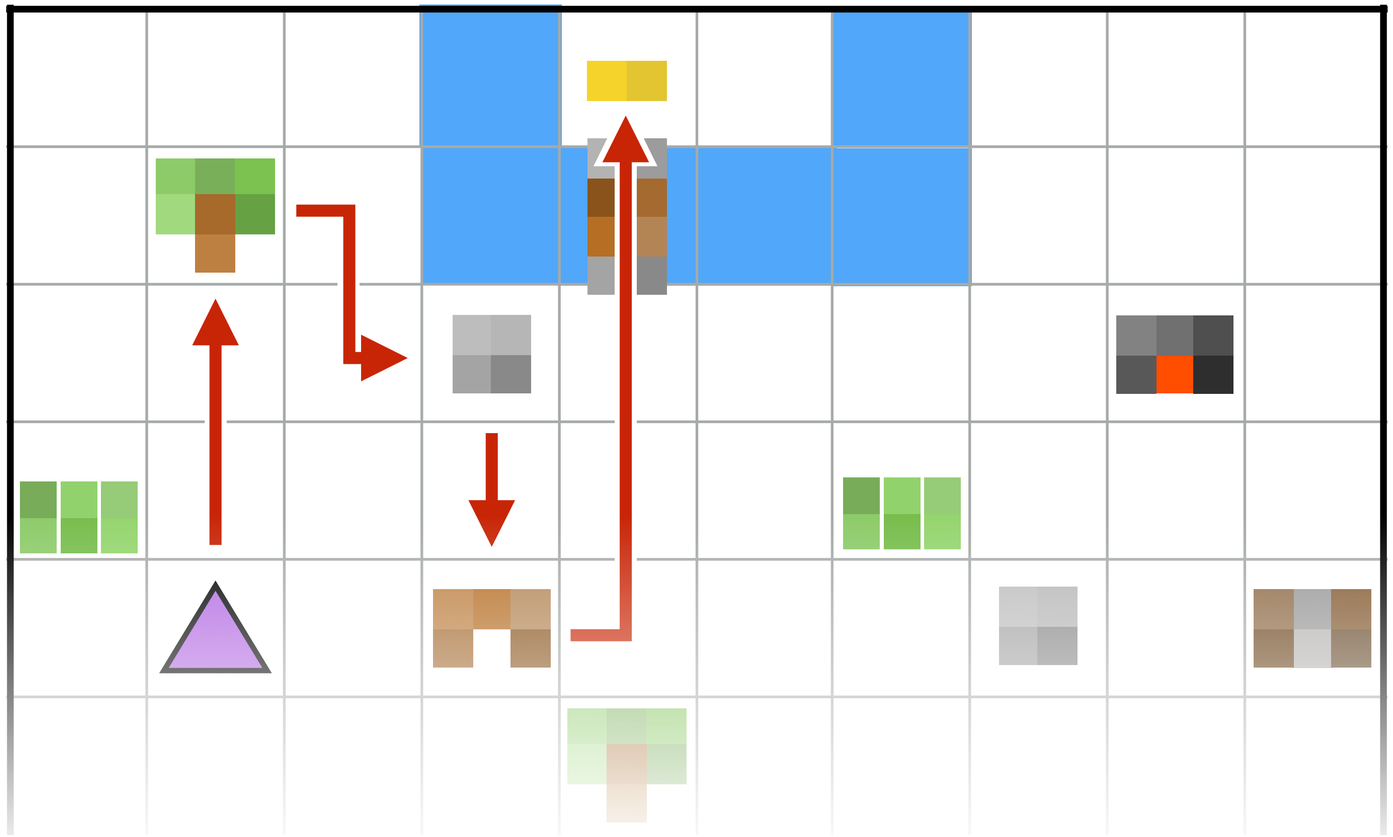
π_1

π_3

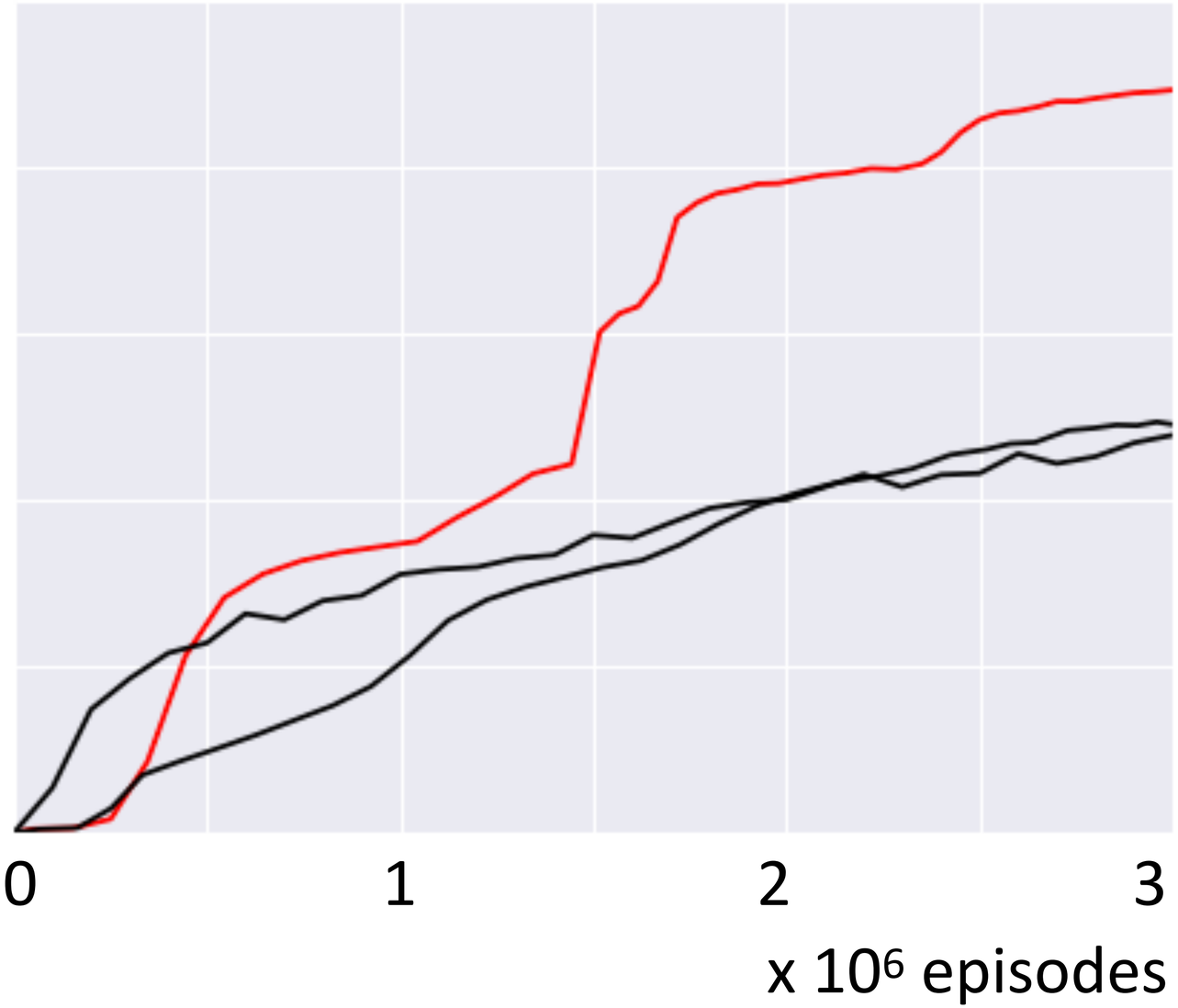
Modular policies



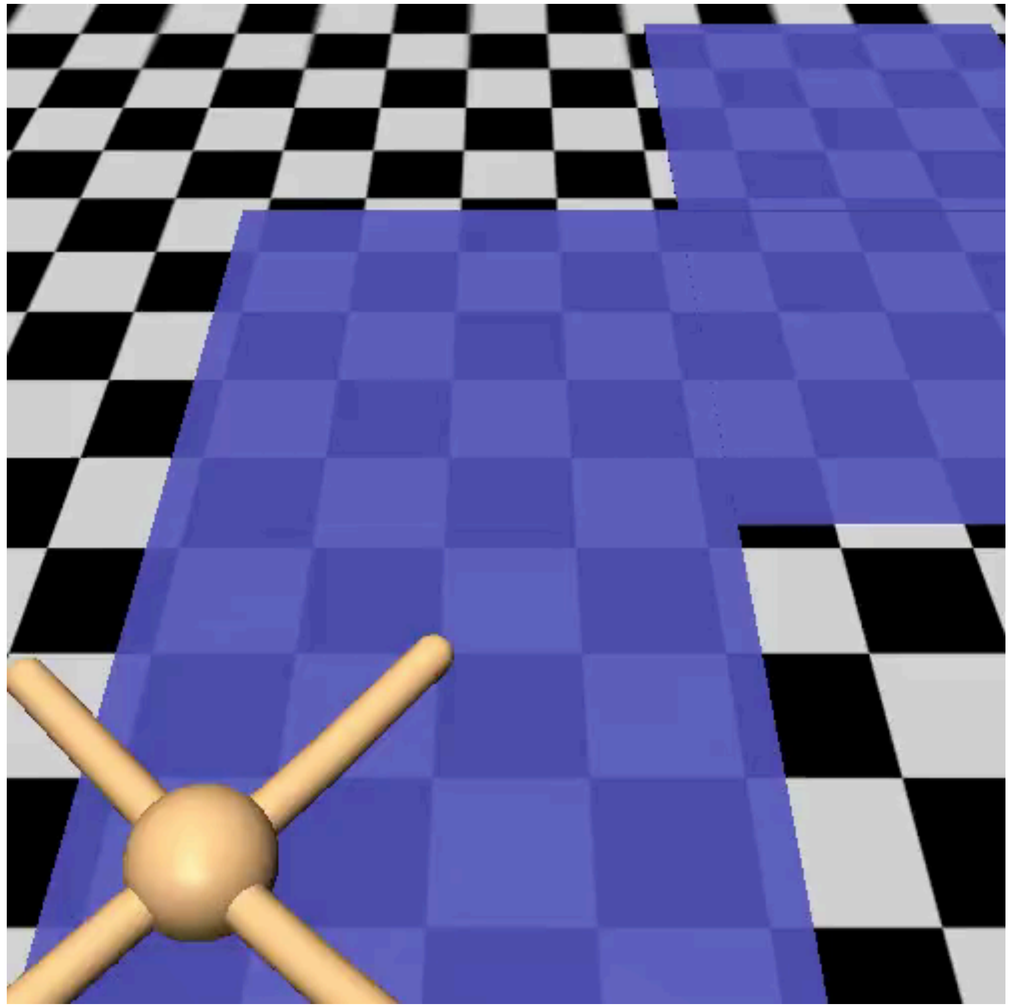
Results: crafting game



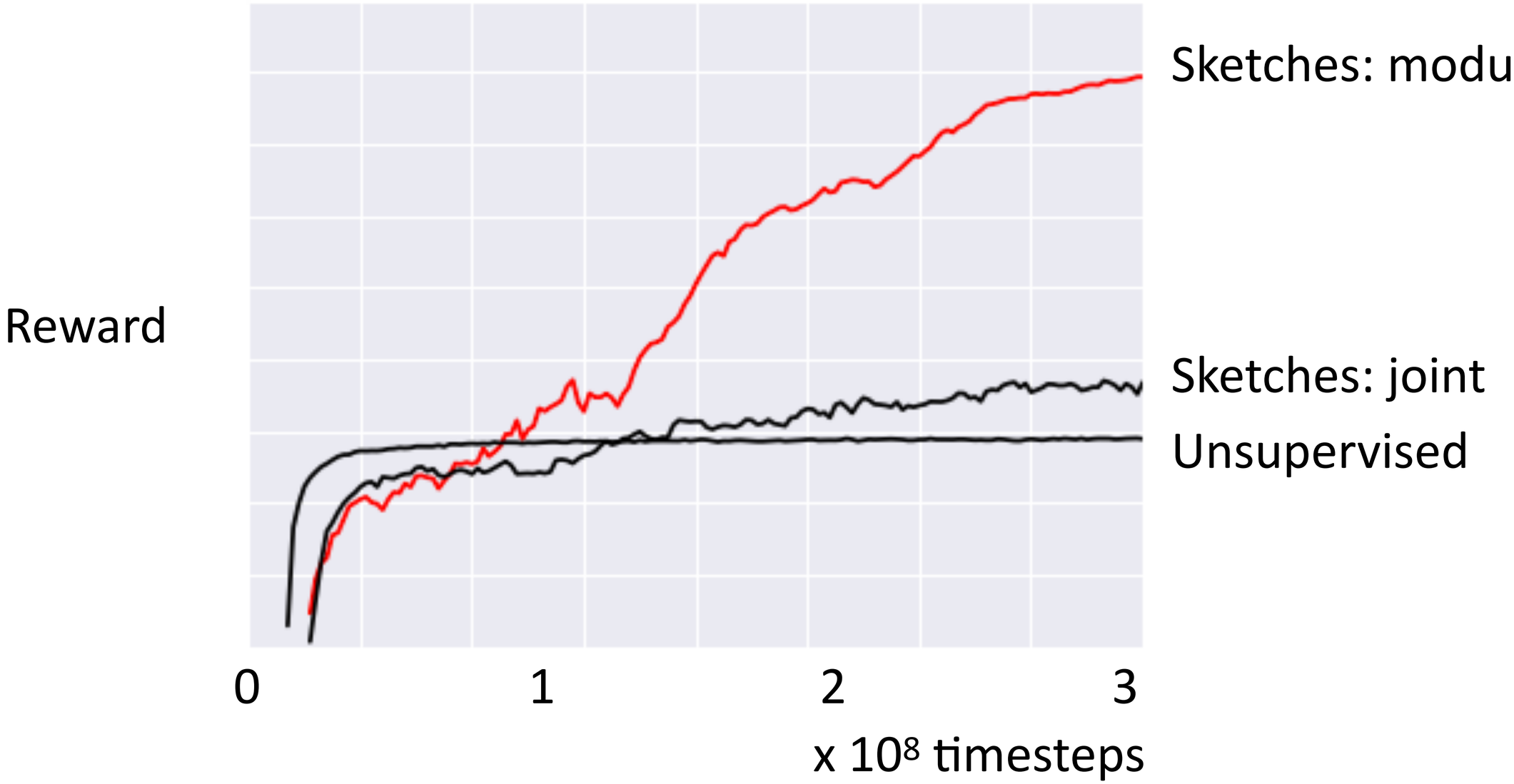
Results: crafting game



Results: locomotion

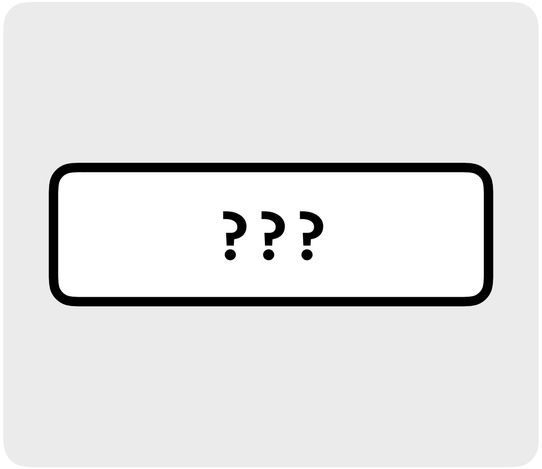


Results: locomotion



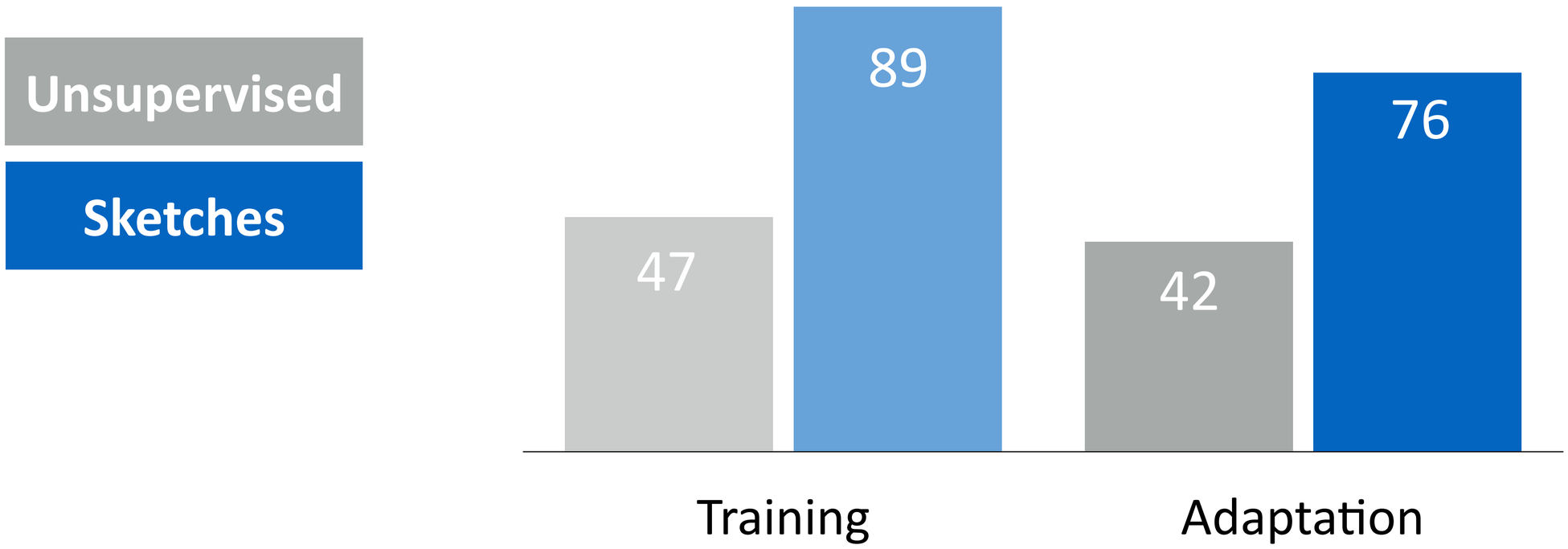
Generalization

What if I don’t get a sketch at test time?

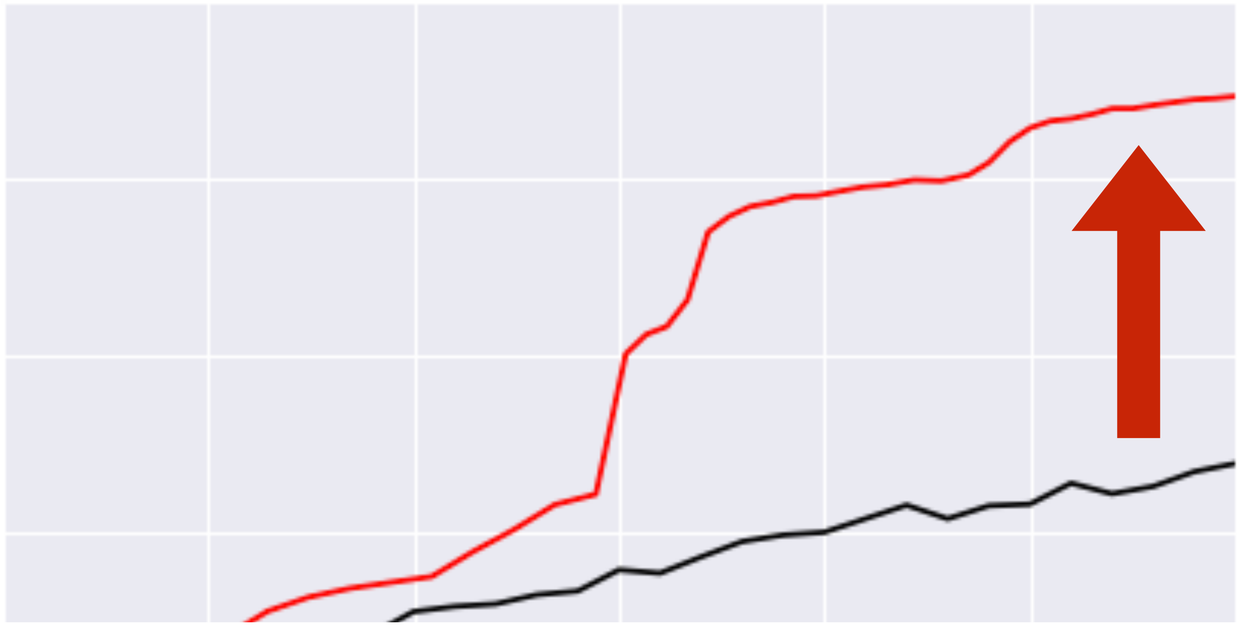


Generalization

What if I don’t get a sketch at test time?



Moral

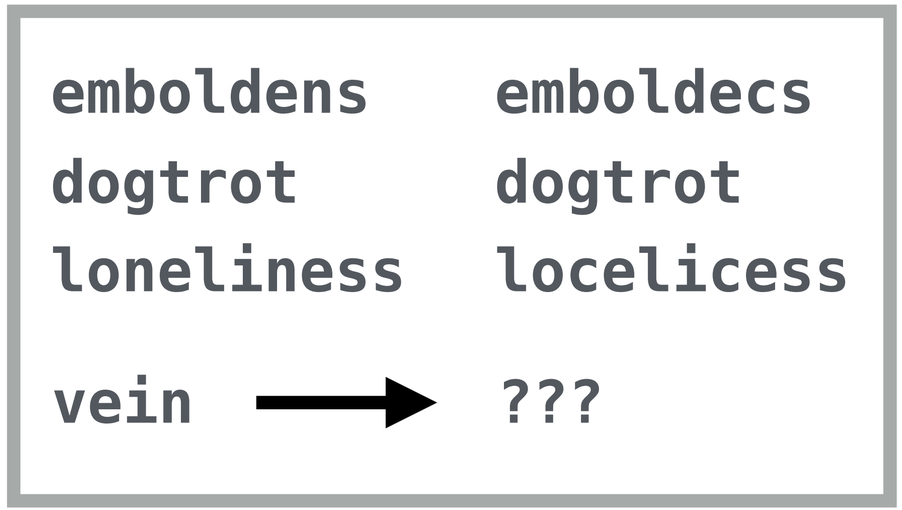


A little bit of (structured) language goes a long way!

Beyond structured sketches

Language learning

Learning from demonstrations

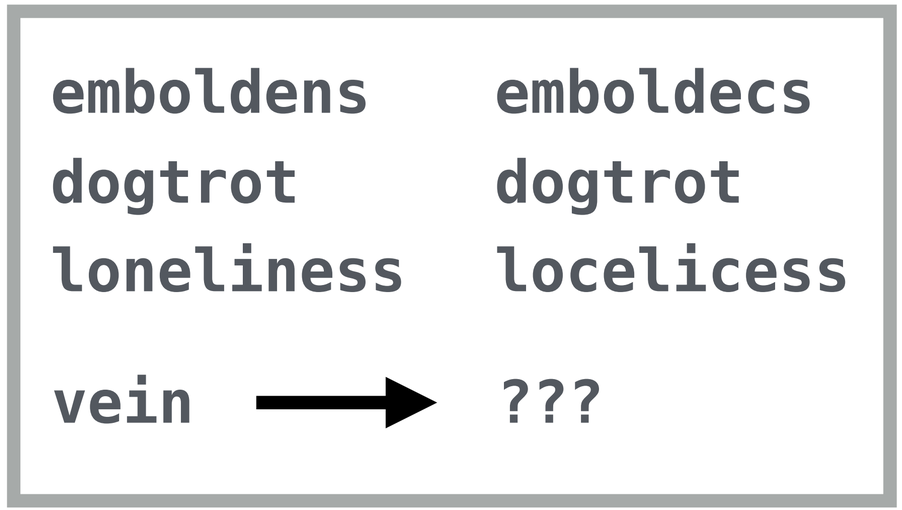


[A, Klein & Levine 17]

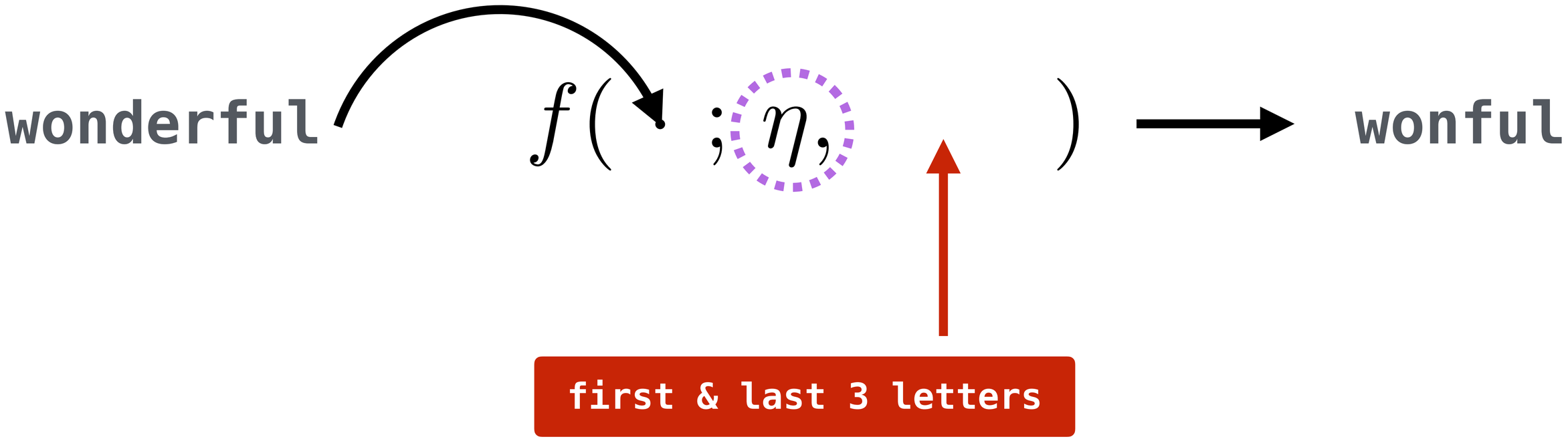
Beyond structured sketches

Language learning

Learning from demonstrations

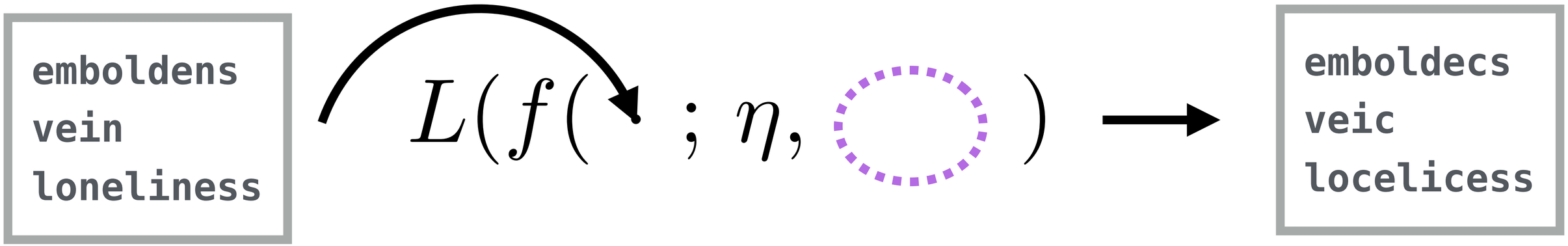


Pretraining via language learning

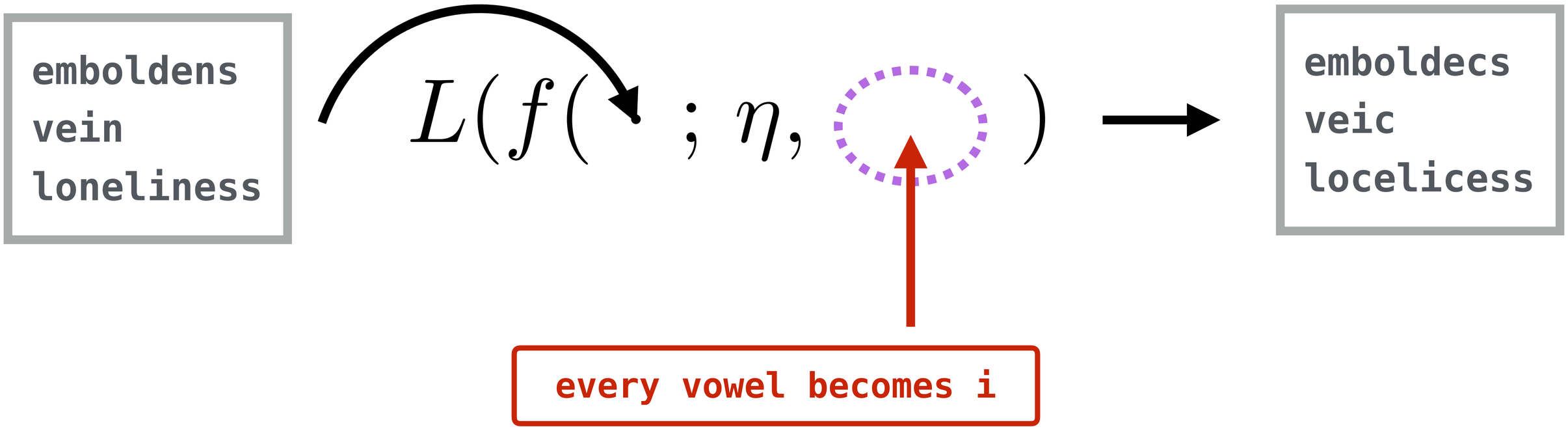


[Branavan et al., 09]

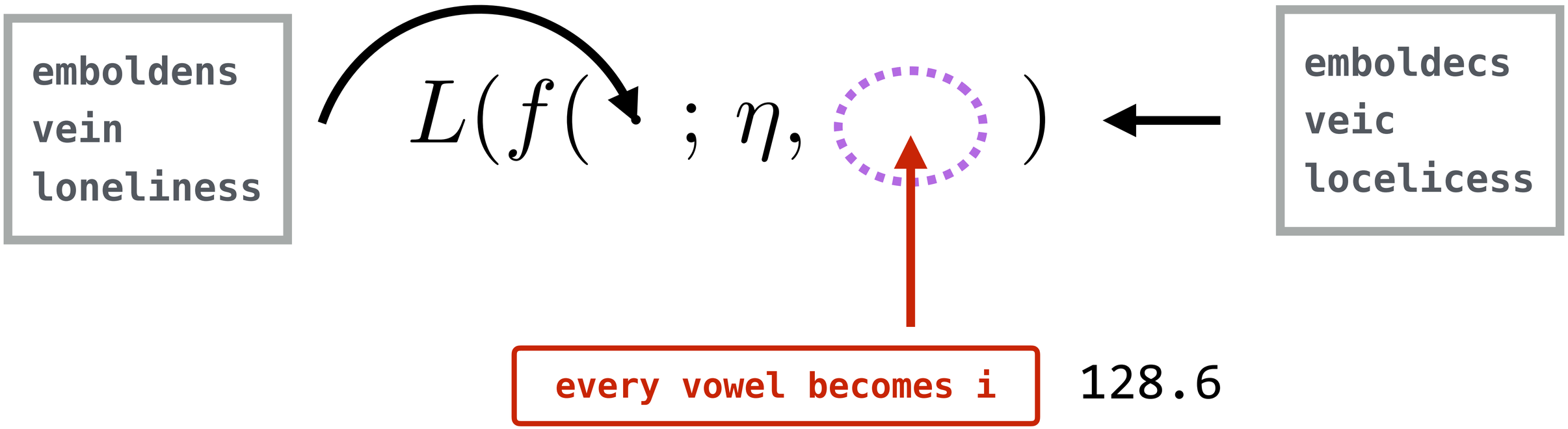
Concept learning



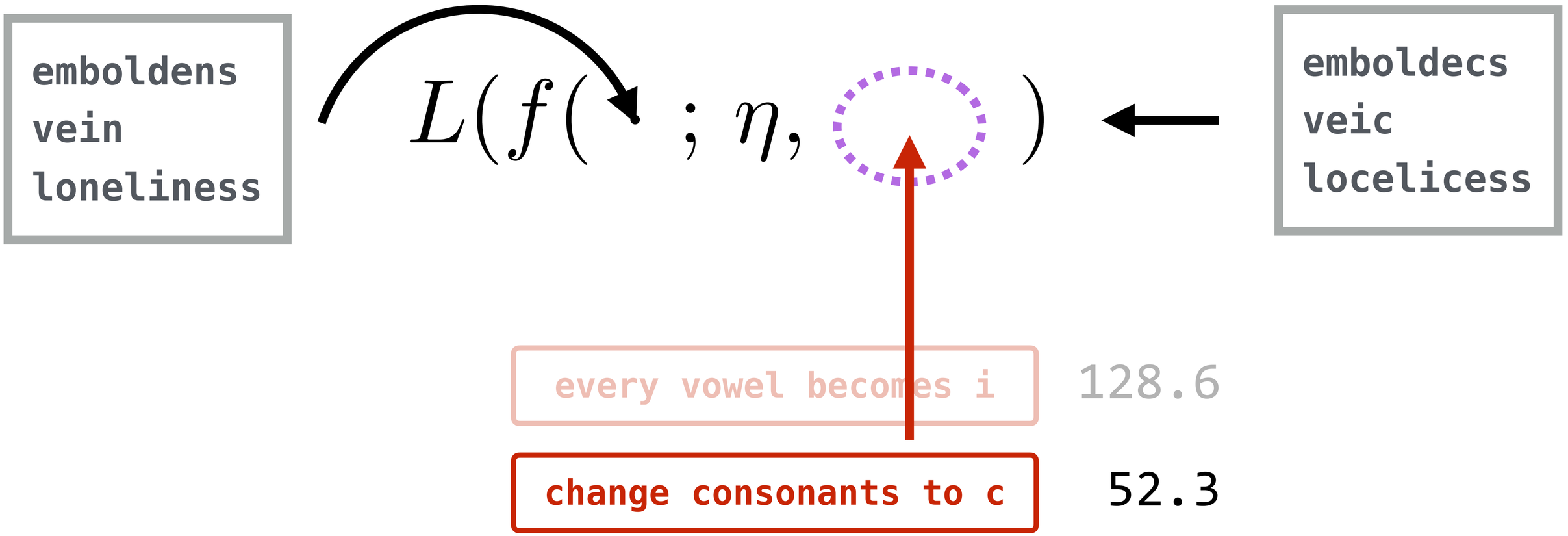
Concept learning



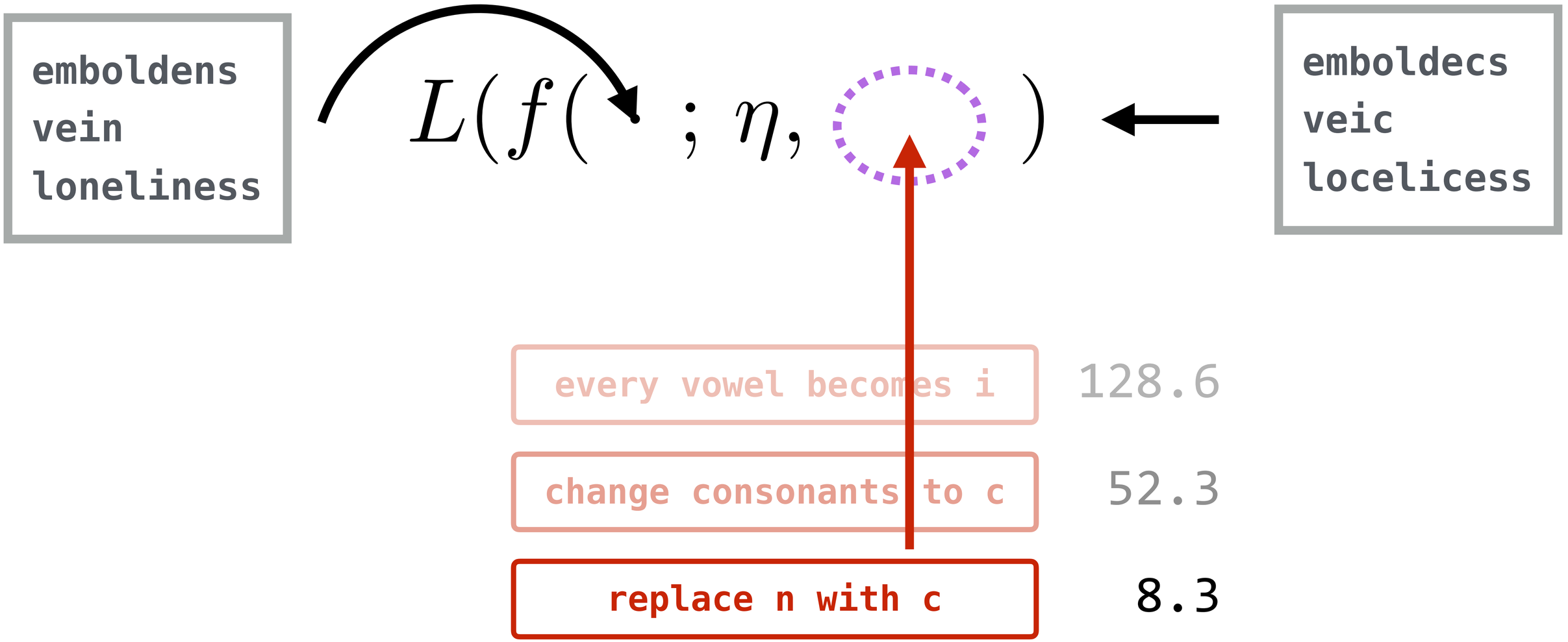
Concept learning



Concept learning

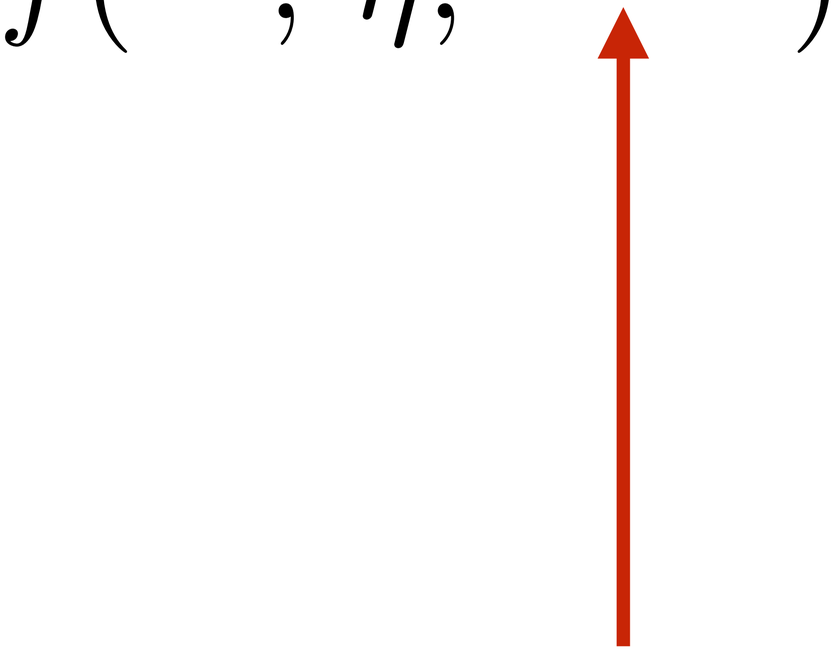


Concept learning

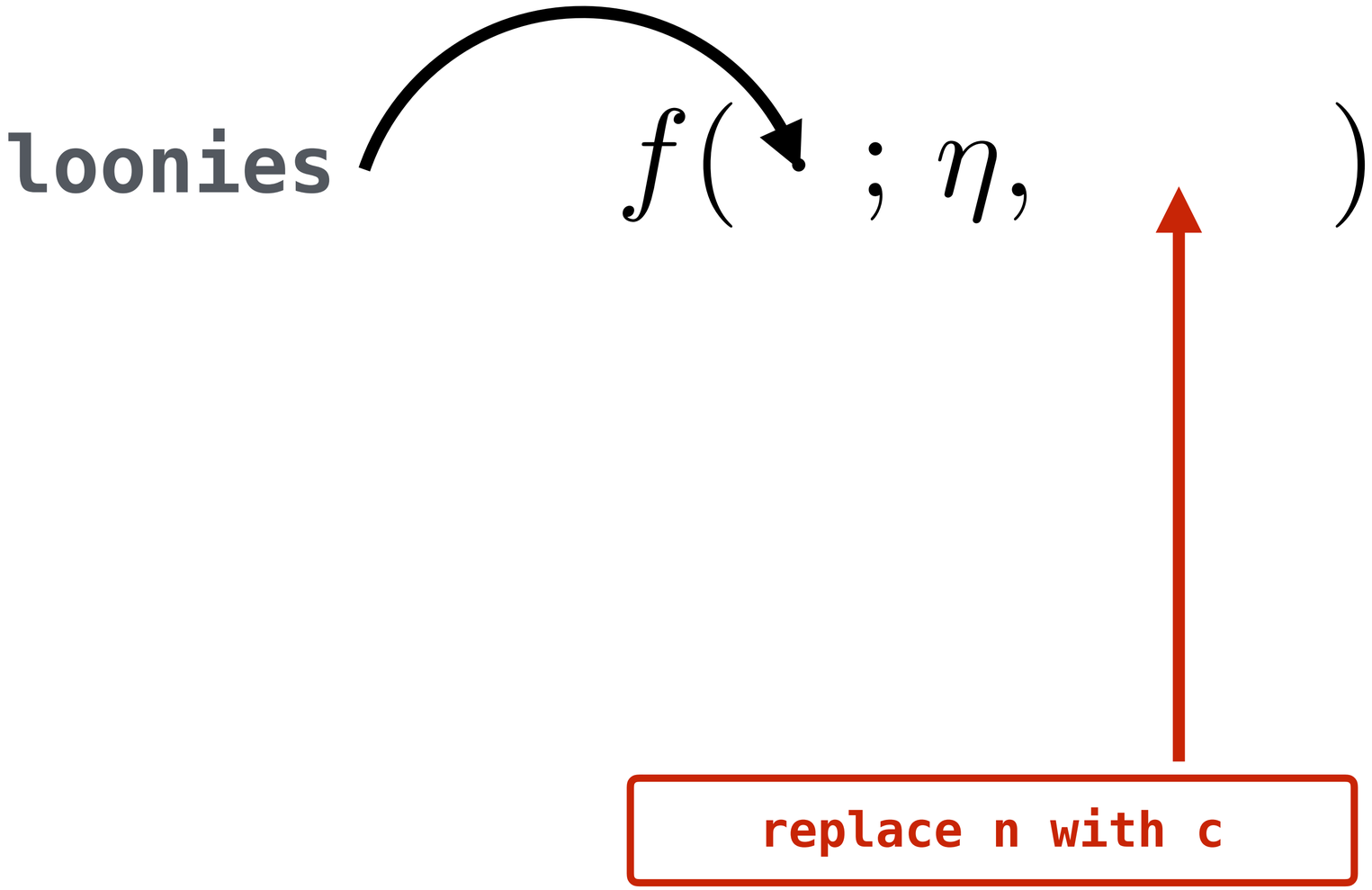


Prediction

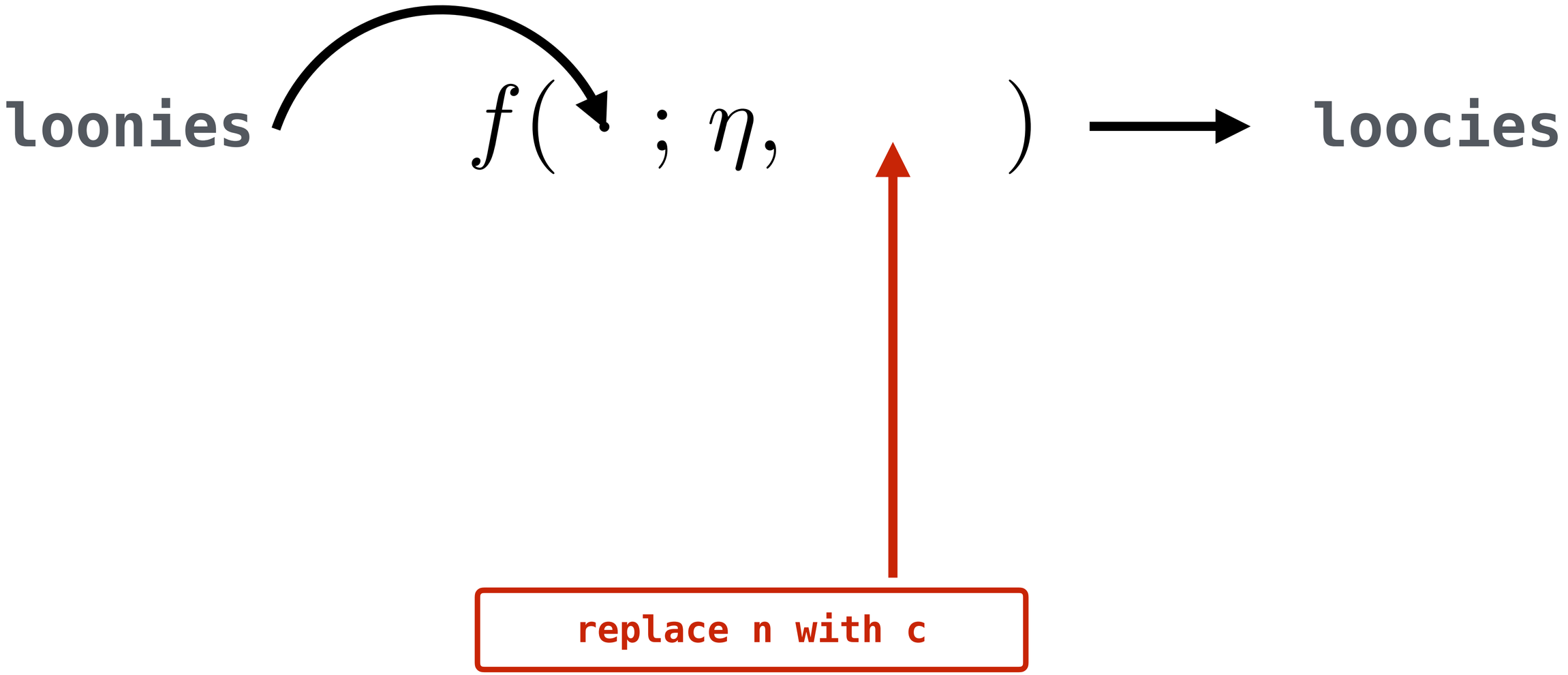
$$f(\ \cdot\ ;\ \eta,\qquad\qquad\ )$$



Evaluation



Evaluation



As multitask learning

Pretraining data

wonderful	wonful
glabrous	glaous
itch	itctch

Training data

emboldens	emboldecs
vein	veic
loneliness	locelicess

$$\arg\min_{\eta} L(f(\texttt{itctch} \mid \texttt{itch}; \eta, \blacktriangle))$$

first & last 3 letters

$$\arg\min_{\eta} L(f(\texttt{veic} \mid \texttt{vein}; \eta, \blacktriangle))$$

??? replace n with c

[Caruana 97]

As inverse reinforcement learning

cost function

$$\arg\min \hat{\mathbb{E}}_{\tau\sim\pi}[L(f(\tau|\,\texttt{vein}\,;\eta,\quad))]$$

???

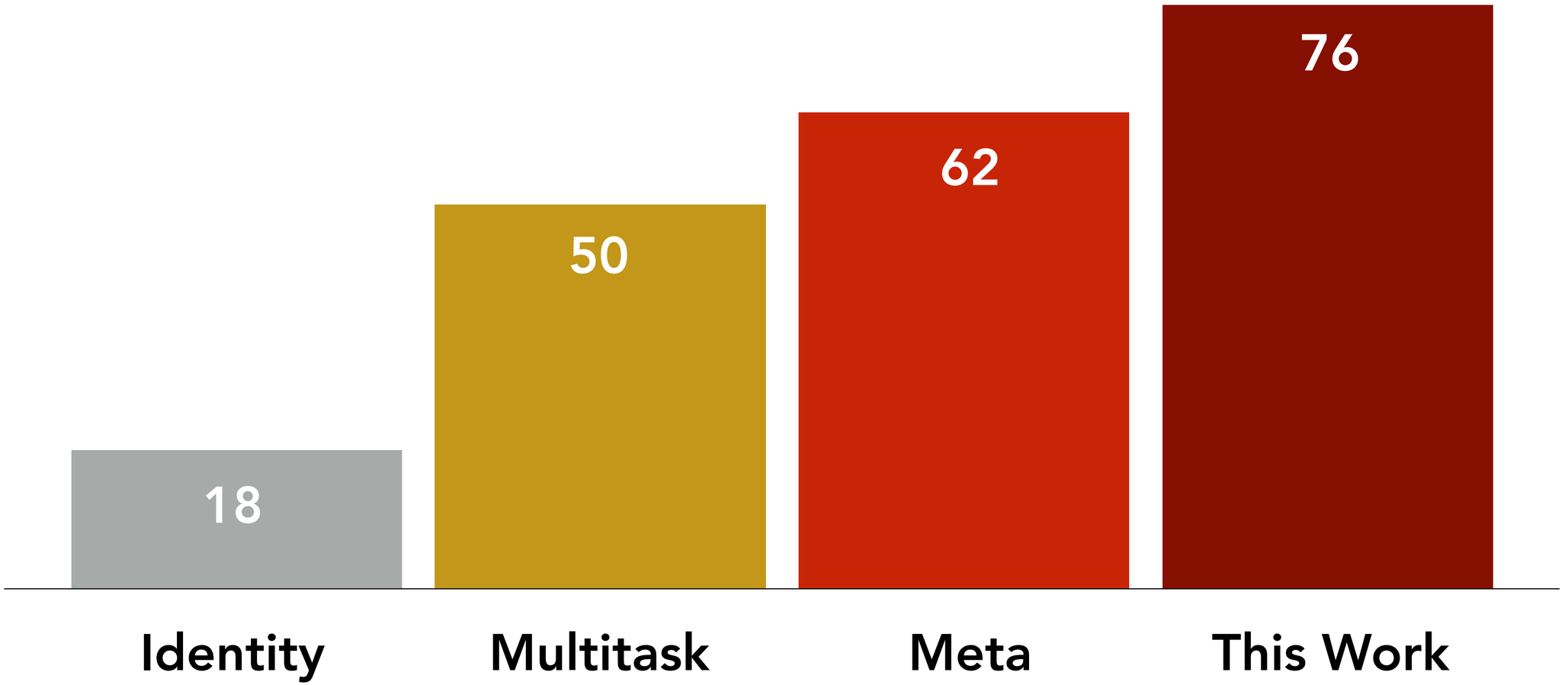
replace n with c

As a language game...

$$\arg\min_{???} f \circ L$$

speaker model **listener loss**

Results: string editing accuracy



Results

examples		
emboldens		emboldecs
kisses		kisses
loneliness	➡	locelicess
vein		veic
dogtrot		dogtrot

true description

replace all n s with c

change any n to a c

pred. description

true output

loocies

loonies

loocies

pred. output

Problems

How do we bootstrap from (unannotated) exploration of the environment alone?

How good are inferred descriptions as explanations?

Problems

How do we bootstrap from (unannotated) exploration of the environment alone

How good are inferred descriptions as explanations?

Conclusions

Conclusions

Use RL in NLP by formulating language generation / understanding as reward maximization rather than supervised learning.

Conclusions

Use language in (I)RL as a scaffold for learning options, goal representations, cost functions.

Languages encode 100k years of accumulated knowledge about which abstractions are useful —take advantage of it!

Thanks!

also...

Looking for NLP jobs? Ask me about Microsoft!

Applying to PhD programs? Ask me about MIT!

jaandrea@microsoft.com