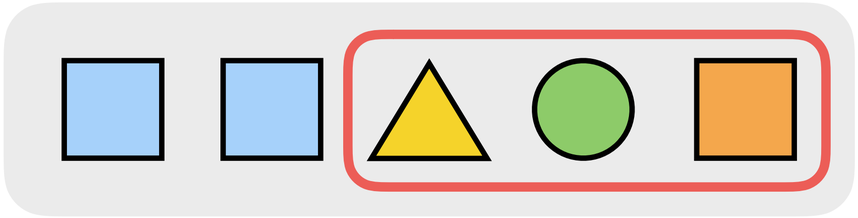
Analogs of Linguistic Structure in Deep Representations

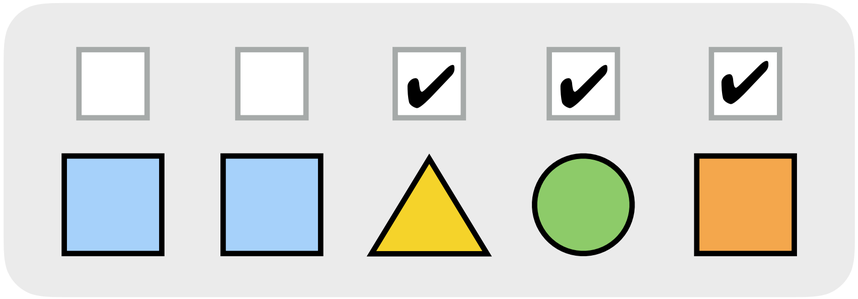
Jacob Andreas and Dan Klein

A game for humans



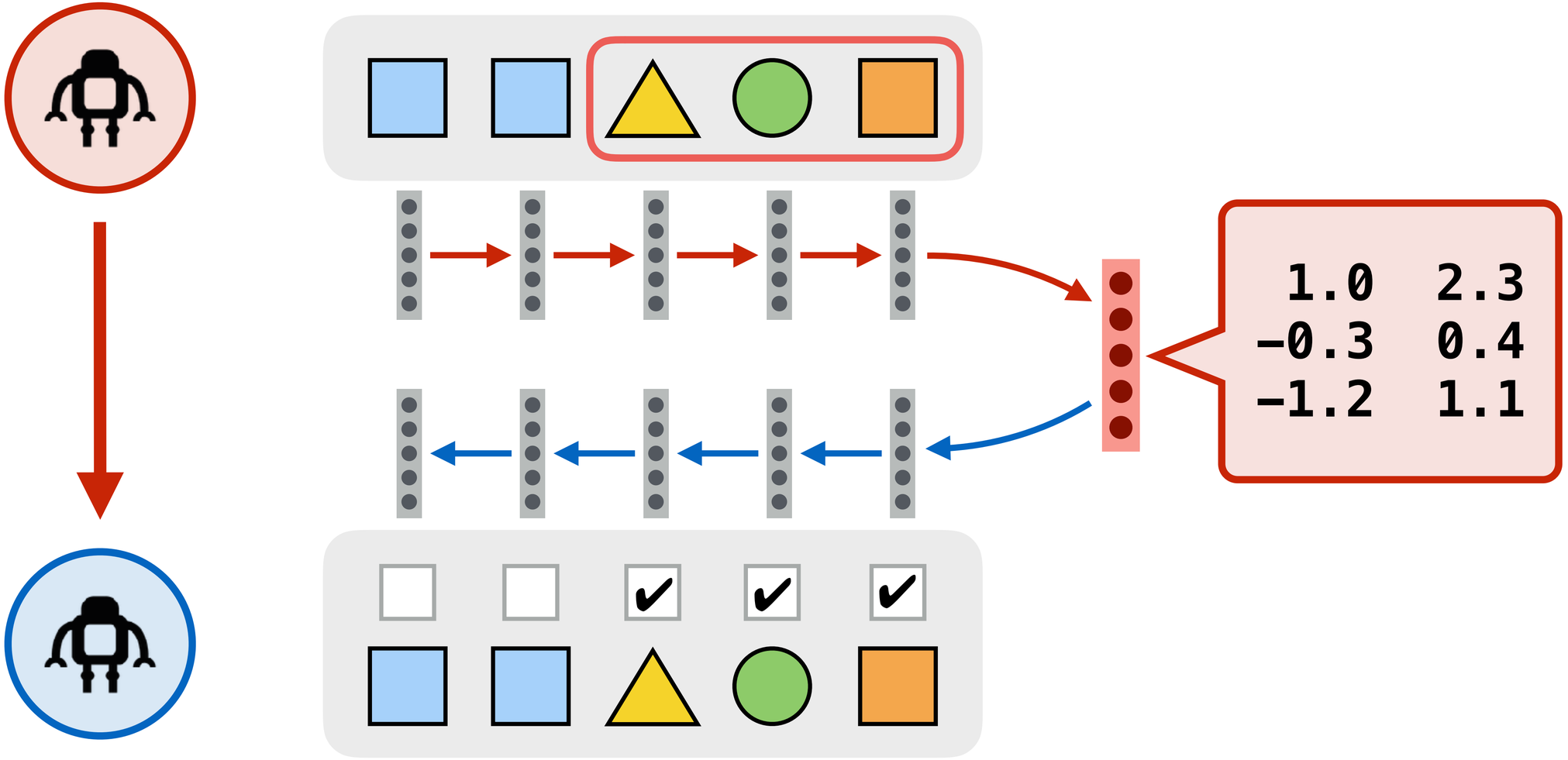
everything but the blue shapes

orange square and non-squares



[FitzGerald et al. 2013]

A game for RNNs



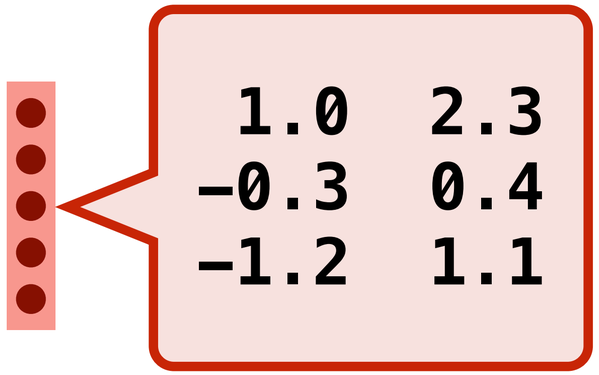
[e.g. Lazaridou et al. 2016]

Questions

1. Does the RNN employ a human-like communicative strategy?

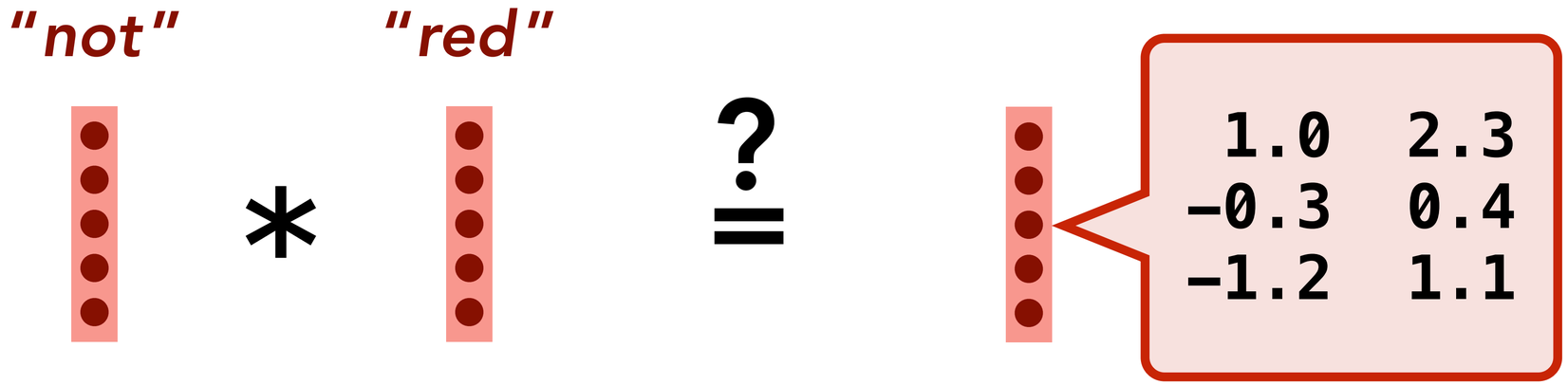
everything but squares

?



Questions

2. Do RNN representations have interpretable compositional structure?



Computing meaning representations

not the red squares

Computing meaning representations

not the red squares

λx.¬(sqr(x)∧red(x))

Computing meaning representations

not the red squares

not red or not square

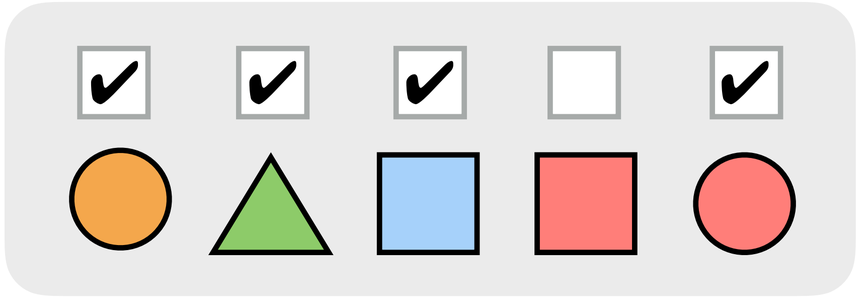
$\lambda x.\neg(sqr(x) \wedge red(x))$

$\lambda x.\neg red(x) \vee \neg sqr(x)$

Computing meaning representations

not the red squares

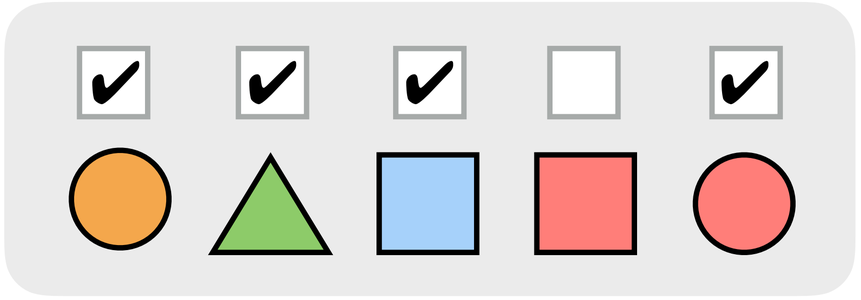
$\lambda x.\neg(sqr(x) \wedge red(x))$



=

not red or not square

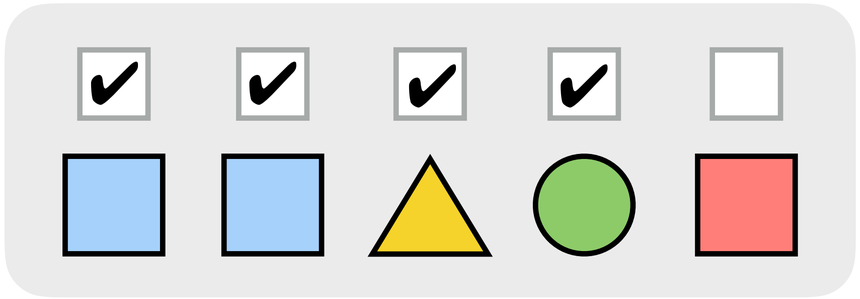
$\lambda x.\neg red(x) \vee \neg sqr(x)$



Computing meaning representations

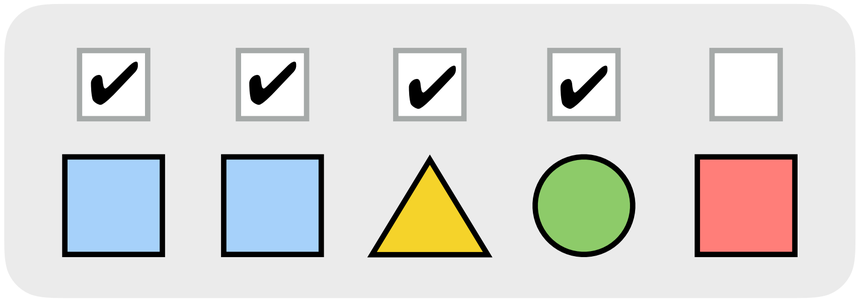
not the red squares

$\lambda x.\neg(sqr(x) \wedge red(x))$



not red or not square

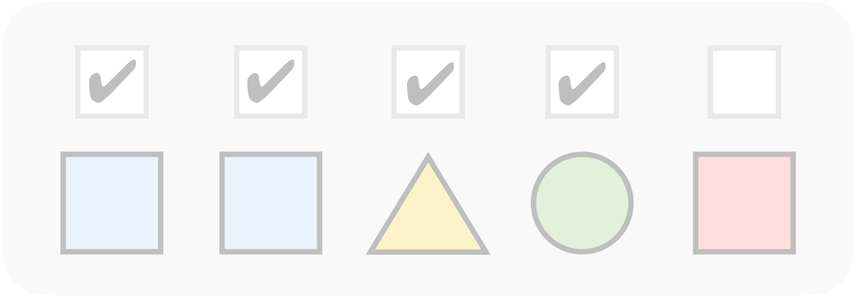
$\lambda x.\neg red(x) \vee \neg sqr(x)$



Computing meaning representations

not the red squares

λx.¬(sqr(x)∧red(x))

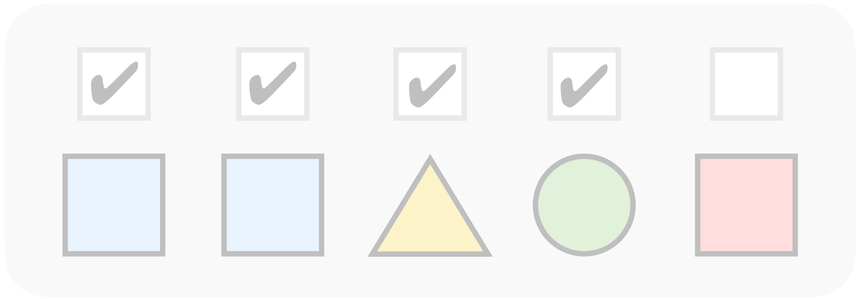


-0.1 1.3 0.5 -0.4

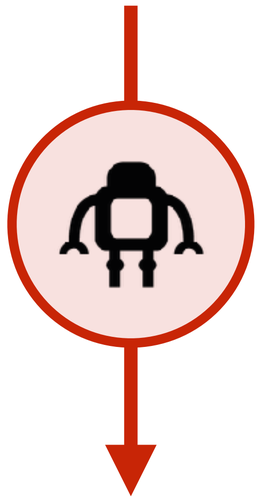
Computing meaning representations

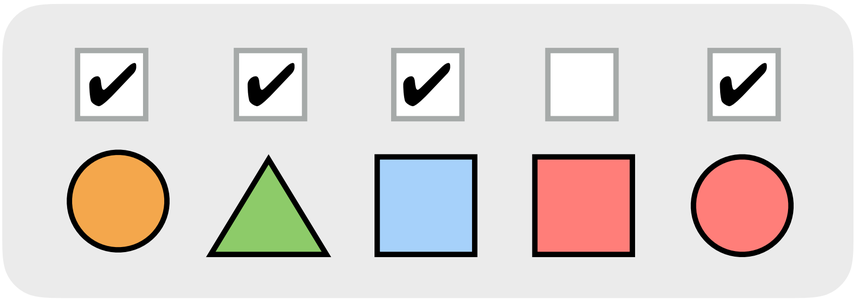
not the red squares

λx.¬(sqr(x)∧red(x))



-0.1 1.3 0.5 -0.4



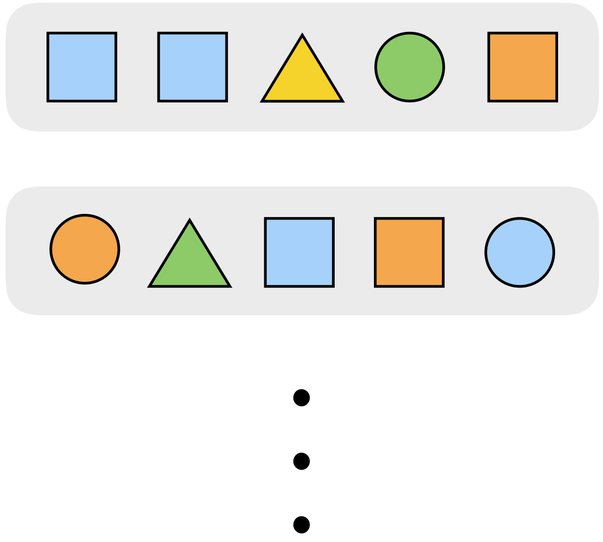


Computing meaning representations

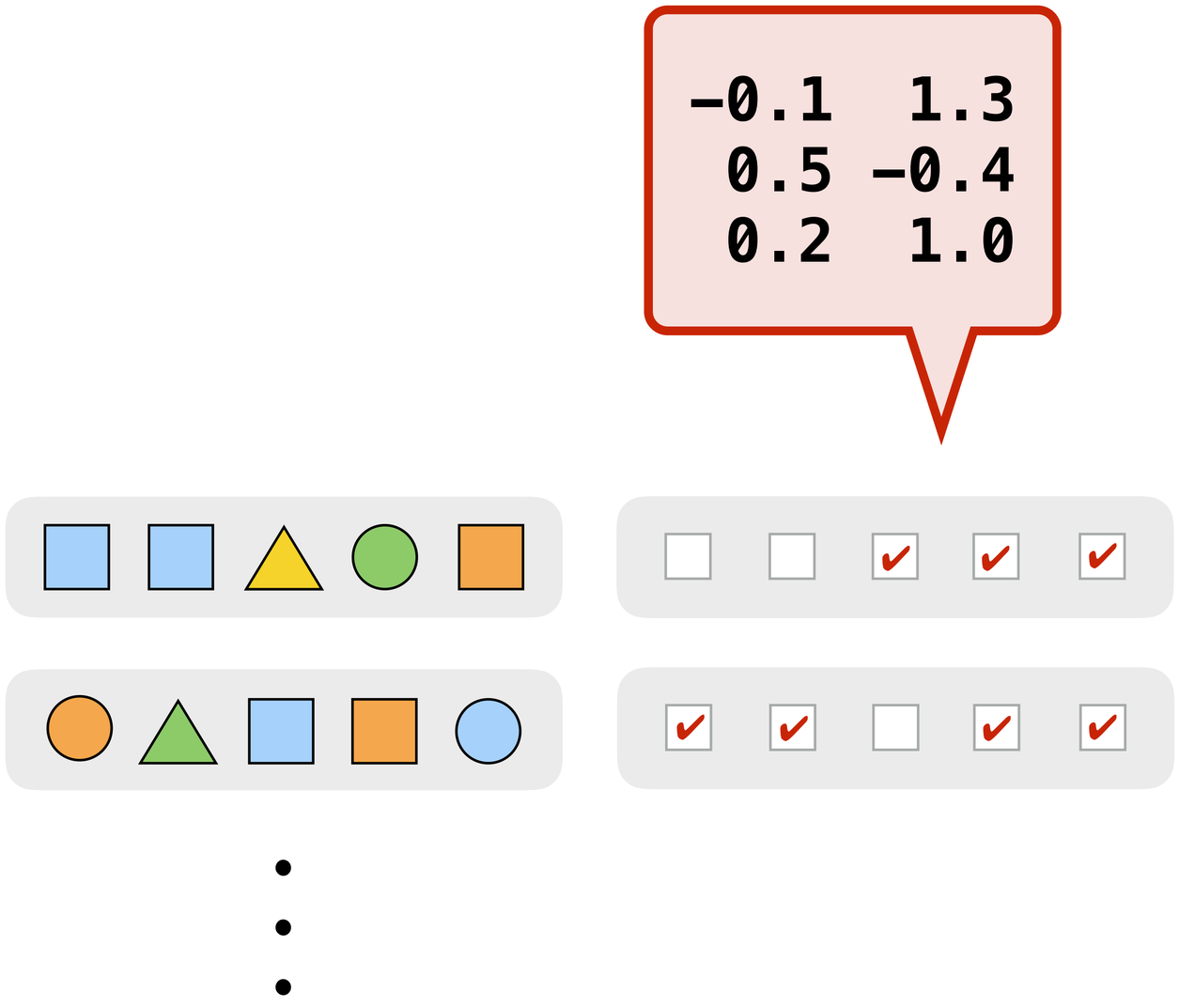
```
-0.1  1.3
 0.5 -0.4
 0.2  1.0
```

Computing meaning representations

```
-0.1  1.3
 0.5 -0.4
 0.2  1.0
```



Computing meaning representations

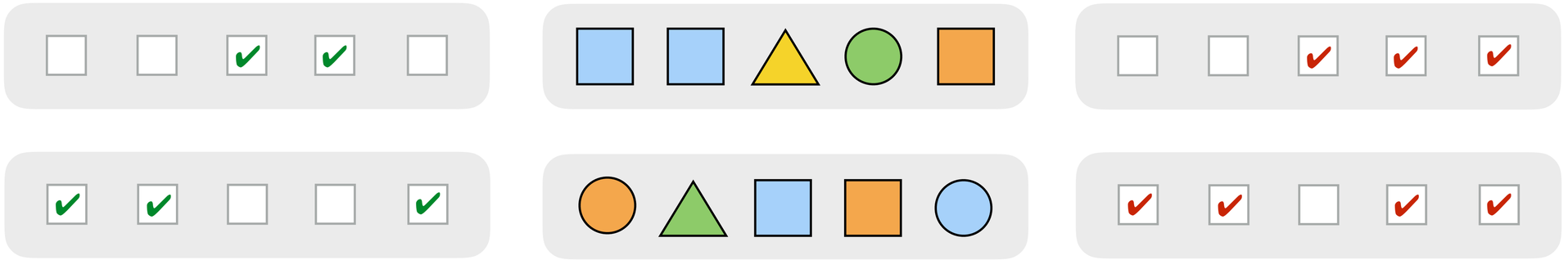


Computing meaning representations

everything but squares

≠

```
-0.1  1.3
 0.5 -0.4
 0.2  1.0
```

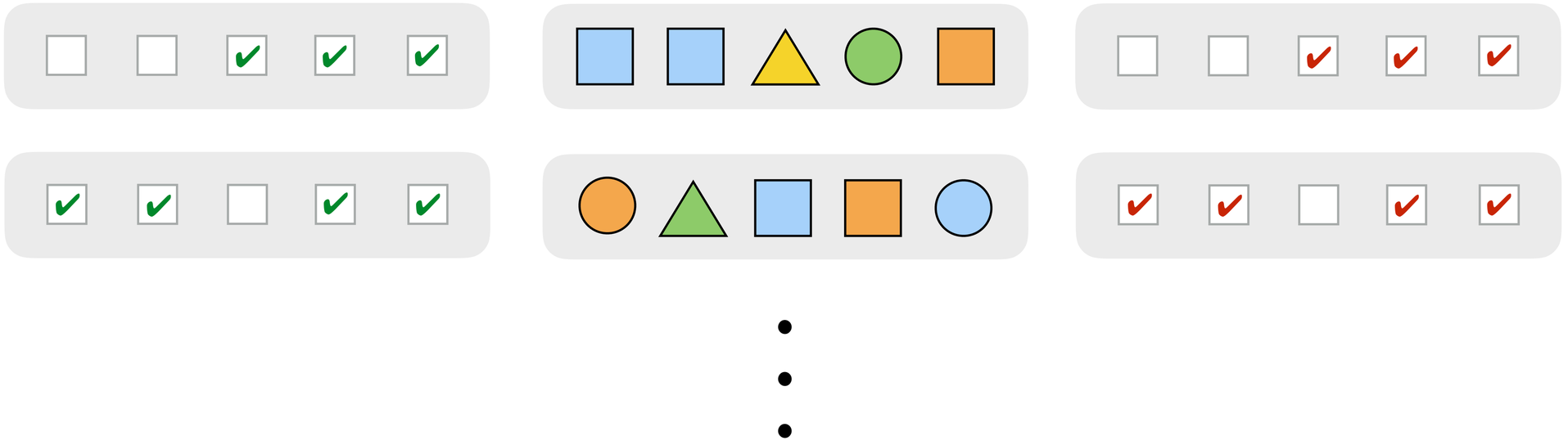


Computing meaning representations

not the blue squares

=

```
-0.1  1.3
 0.5 -0.4
 0.2  1.0
```



Translating

By comparing denotations from

logical forms and the **decoder model**,

we can find **utterances** and **vectors**

with the same meaning.

[A, Dragan & Klein 2013]

Questions

1. Does the RNN employ a human-like communicative strategy?

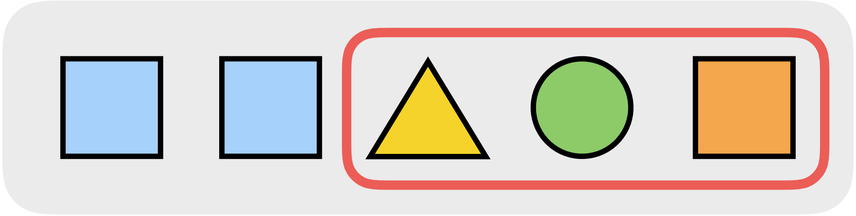
2. Do RNN representations have interpretable compositional structure?

Questions

1. Does the RNN employ a human-like communicative strategy?

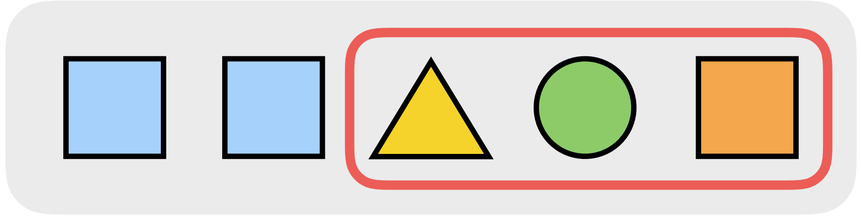
2. Do RNN representations have interpretable compositional structure?

Comparing strategies



Comparing strategies

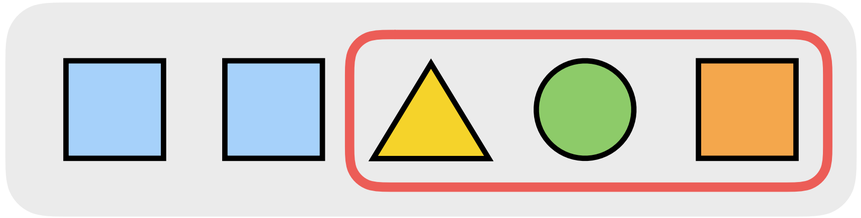
```
-0.1  1.3
 0.5 -0.4
 0.2  1.0
```



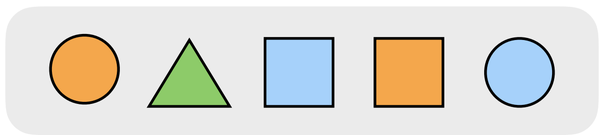
everything but squares

Comparing strategies

```
-0.1  1.3
 0.5 -0.4
 0.2  1.0
```



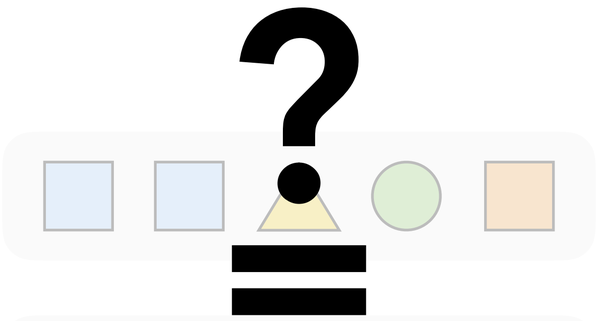
everything but squares



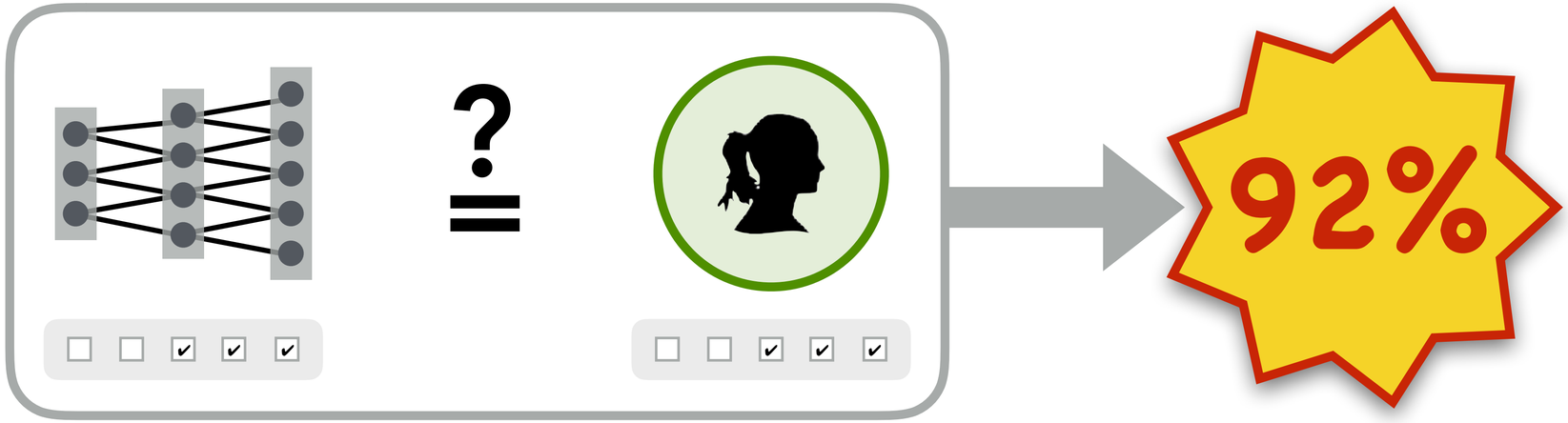
Comparing strategies

```
-0.1  1.3
 0.5 -0.4
 0.2  1.0
```

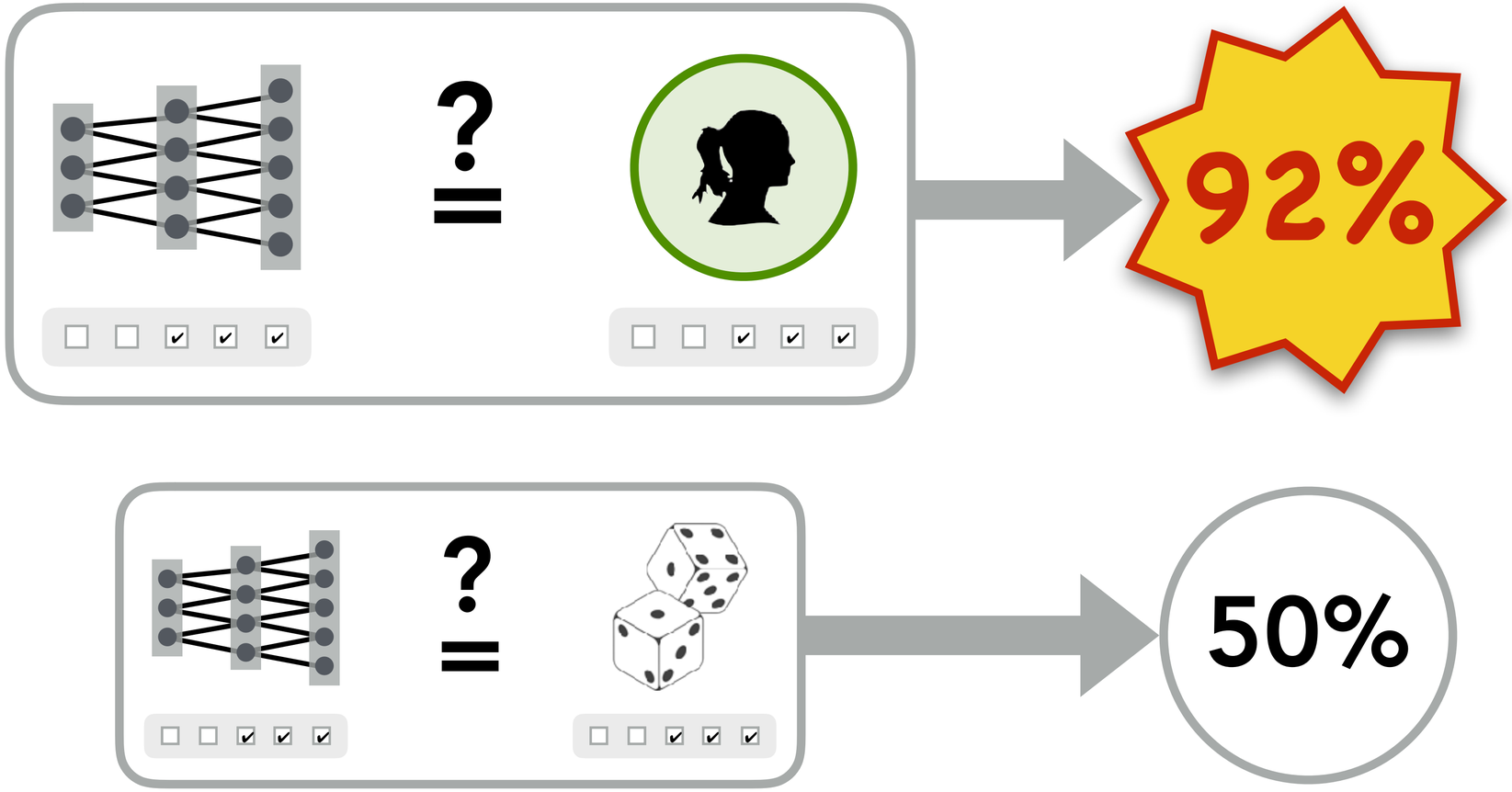
everything but squares



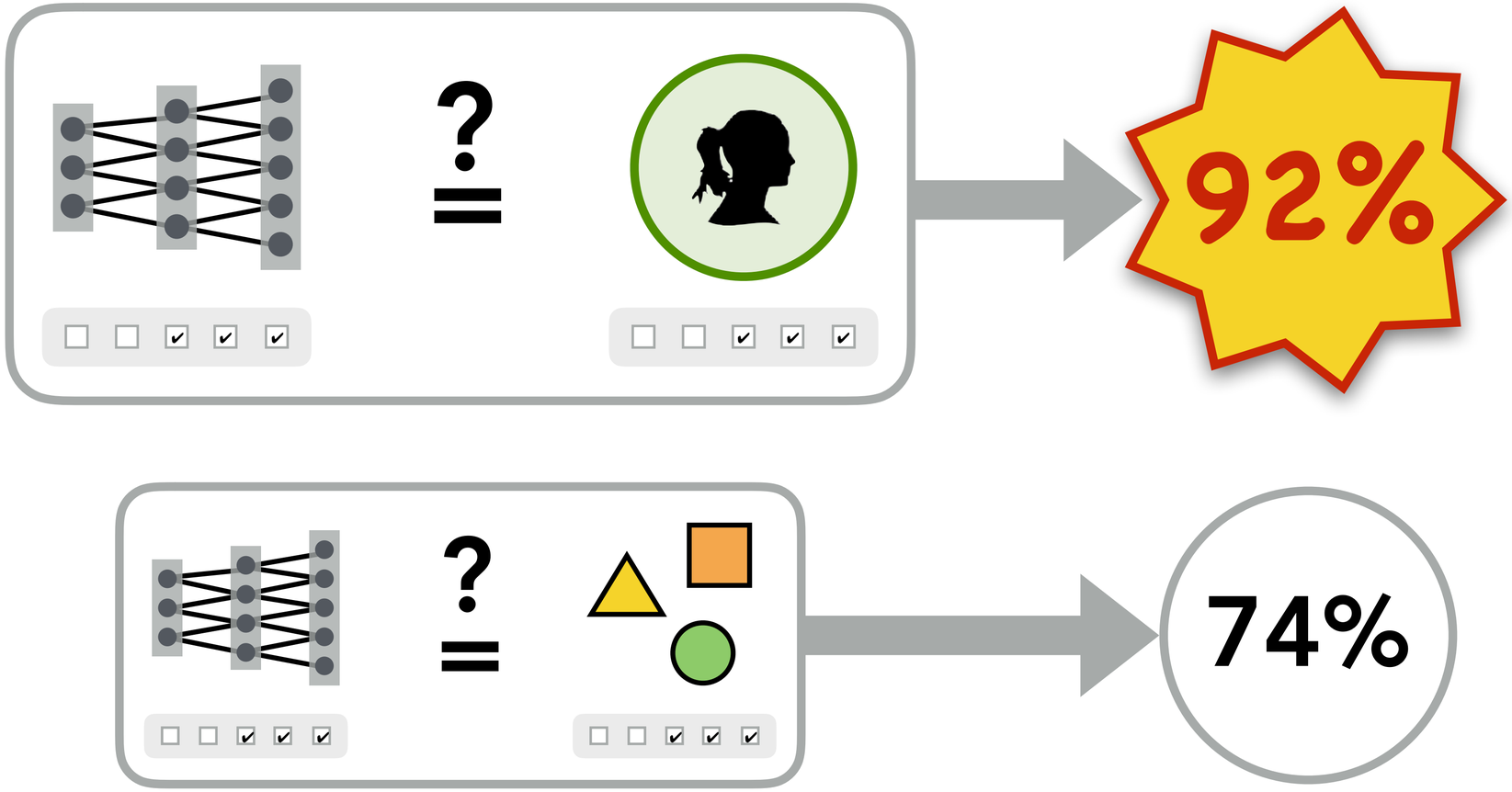
Evaluation: strategies



Evaluation: strategies



Evaluation: strategies



Experiments

1. Does the RNN employ a human-like communicative strategy?

2. Do RNN representations have interpretable compositional structure?

Collecting translation data

all the red shapes

blue objects

everything but red

green squares

not green squares

Collecting translation data

λx.red(x)

λx.blu(x)

λx.¬red(x)

λx.grn(x)∧sqr(x)

λx.¬(grn(x)∧sqr(x))

Collecting translation data

λx.red(x)	→	0.1 −0.3 0.5 1.1
λx.blu(x)	→	−0.3 0.2 0.1 0.1
λx.¬red(x)	→	1.4 −0.3 −0.5 0.8
λx.grn(x)∧sqr(x)	→	0.2 −0.2 0.5 −0.1
λx.¬(grn(x)∧sqr(x))	→	0.3 −1.3 −1.5 0.1

Extracting related pairs

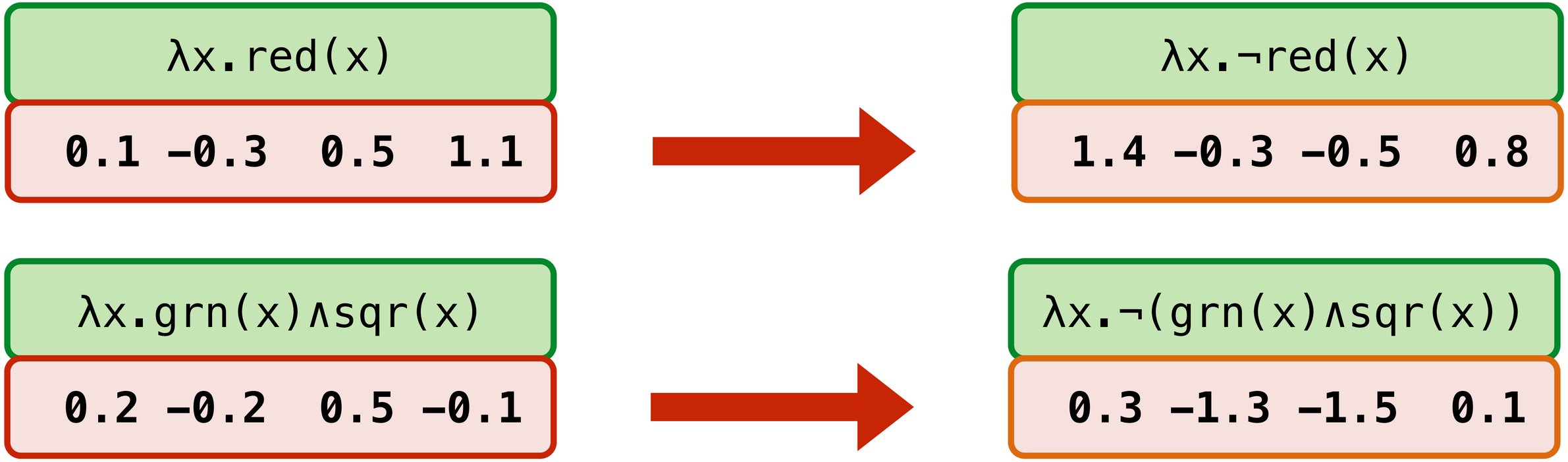
λx.red(x)
0.1 −0.3 0.5 1.1

λx.¬red(x)
1.4 −0.3 −0.5 0.8

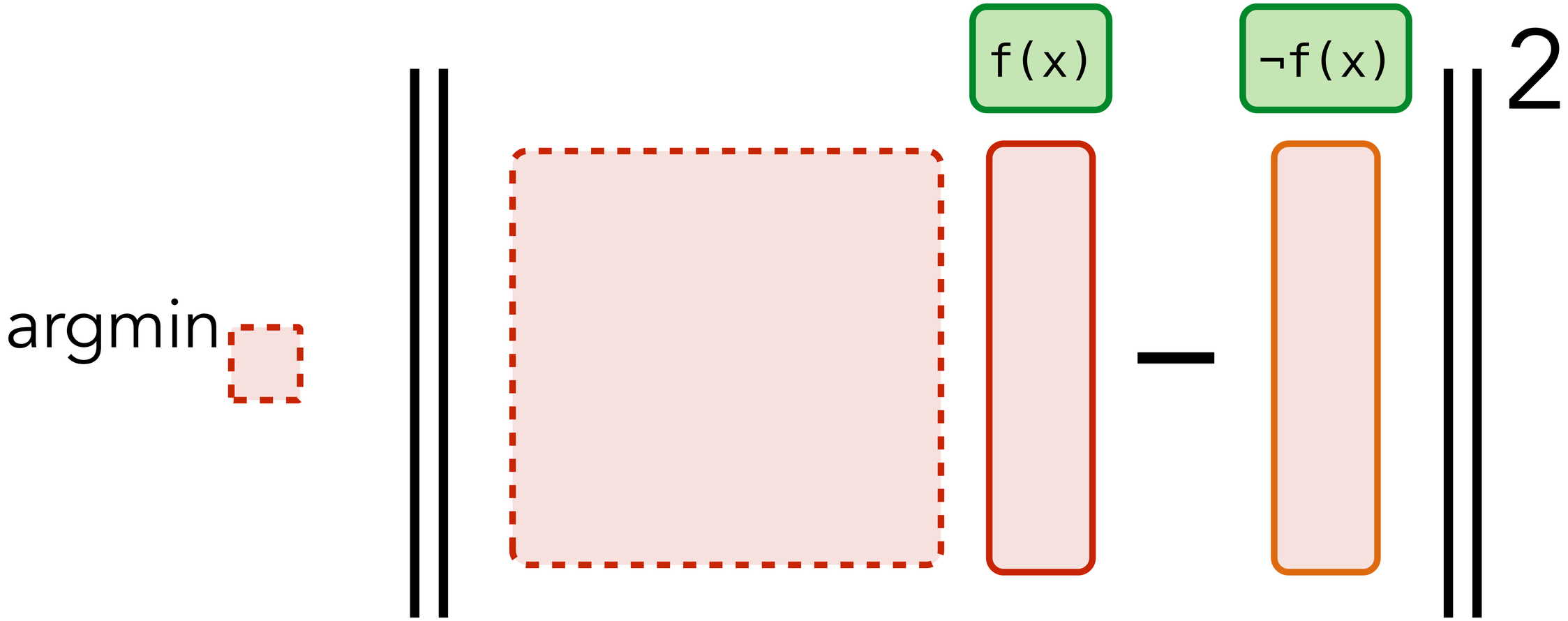
λx.grn(x)∧sqr(x)
0.2 −0.2 0.5 −0.1

λx.¬(grn(x)∧sqr(x))
0.3 −1.3 −1.5 0.1

Extracting related pairs



Learning compositional operators



Evaluating learned operators

λx.red(x)

0.1 −0.3 0.5 1.1

λx.¬red(x)

1.4 −0.3 −0.5 0.8

λx.grn(x)∧sqr(x)

0.2 −0.2 0.5 −0.1

λx.¬(grn(x)∧sqr(x))

0.3 −1.3 −1.5 0.1

λx.f(x)

0.2 −0.2 0.5 −0.1

Evaluating learned operators

λx.red(x)

0.1 −0.3 0.5 1.1

λx.¬red(x)

1.4 −0.3 −0.5 0.8

λx.grn(x)∧sqr(x)

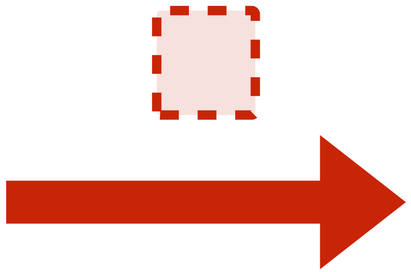
0.2 −0.2 0.5 −0.1

λx.¬(grn(x)∧sqr(x))

0.3 −1.3 −1.5 0.1

λx.f(x)

0.2 −0.2 0.5 −0.1



−0.2 0.4 −0.3 0.0

Evaluating learned operators

λx.red(x)

0.1 −0.3 0.5 1.1

λx.¬red(x)

1.4 −0.3 −0.5 0.8

λx.grn(x)∧sqr(x)

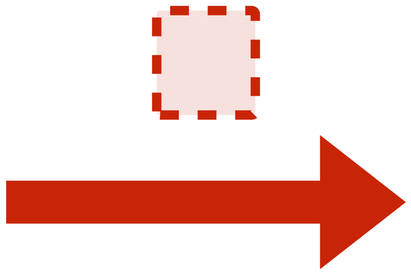
0.2 −0.2 0.5 −0.1

λx.¬(grn(x)∧sqr(x))

0.3 −1.3 −1.5 0.1

λx.f(x)

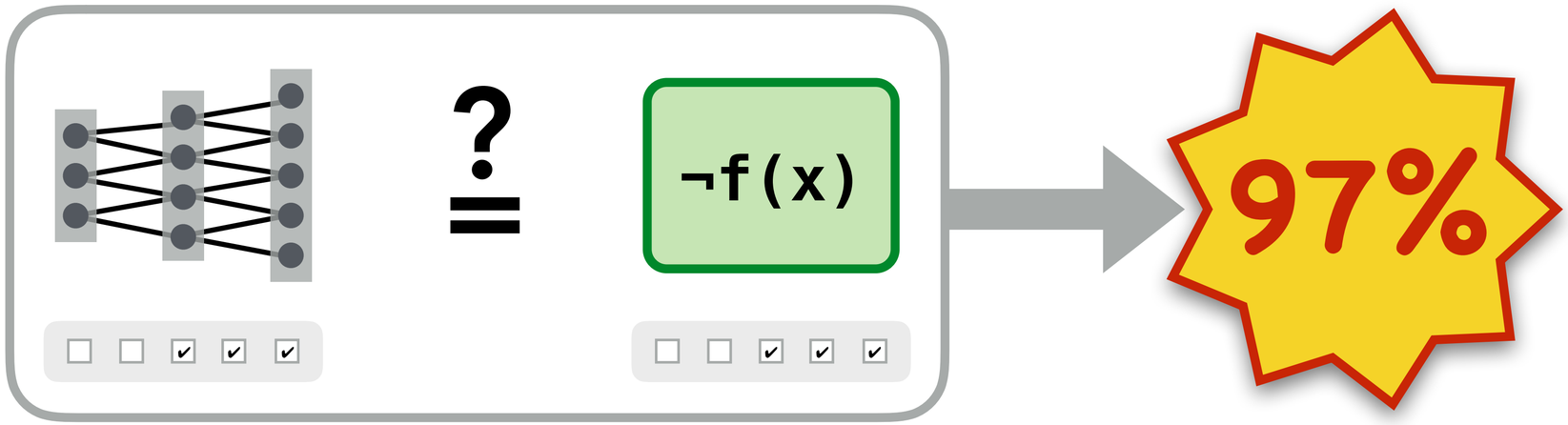
0.2 −0.2 0.5 −0.1



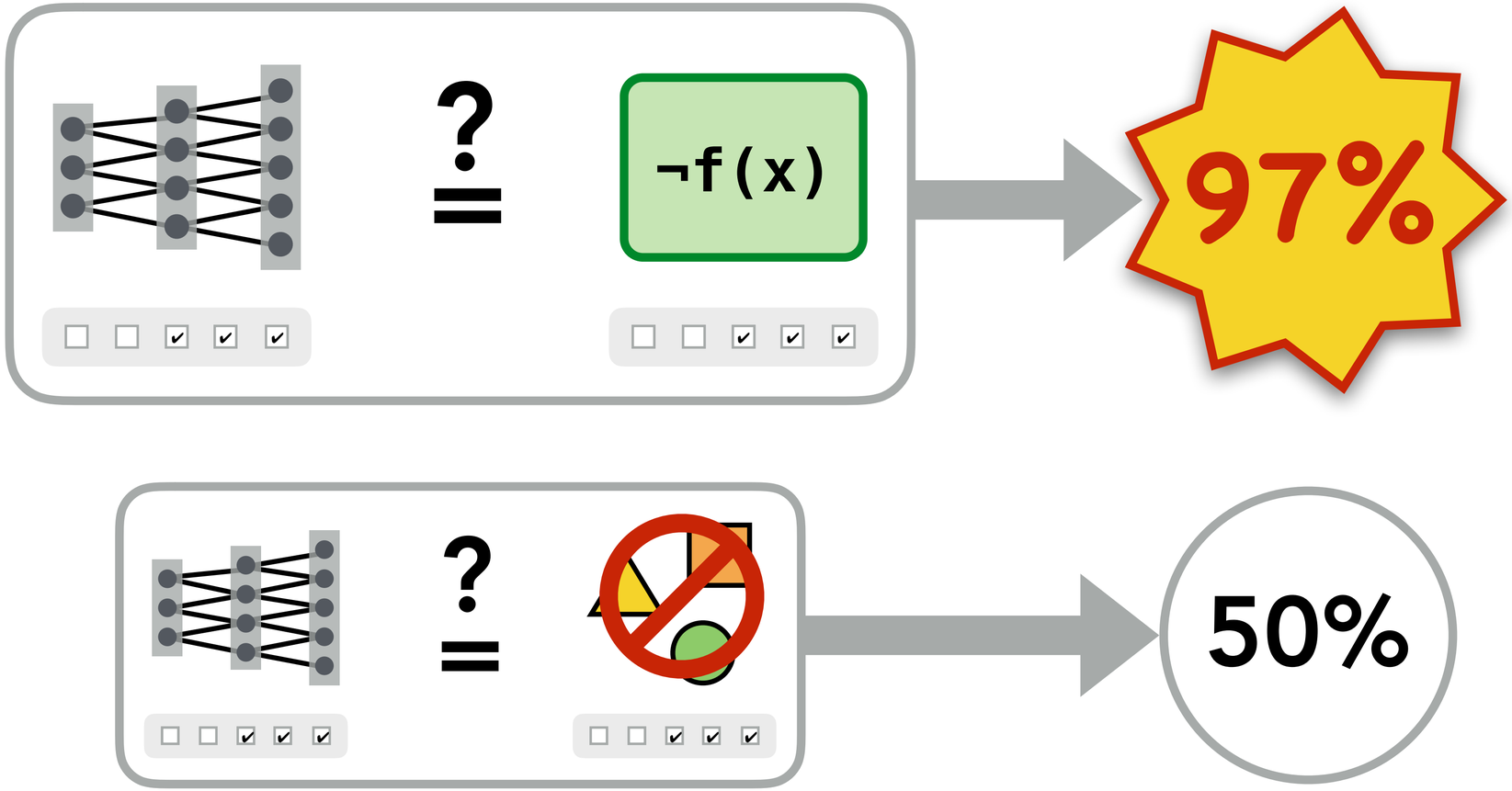
???

−0.2 0.4 −0.3 0.0

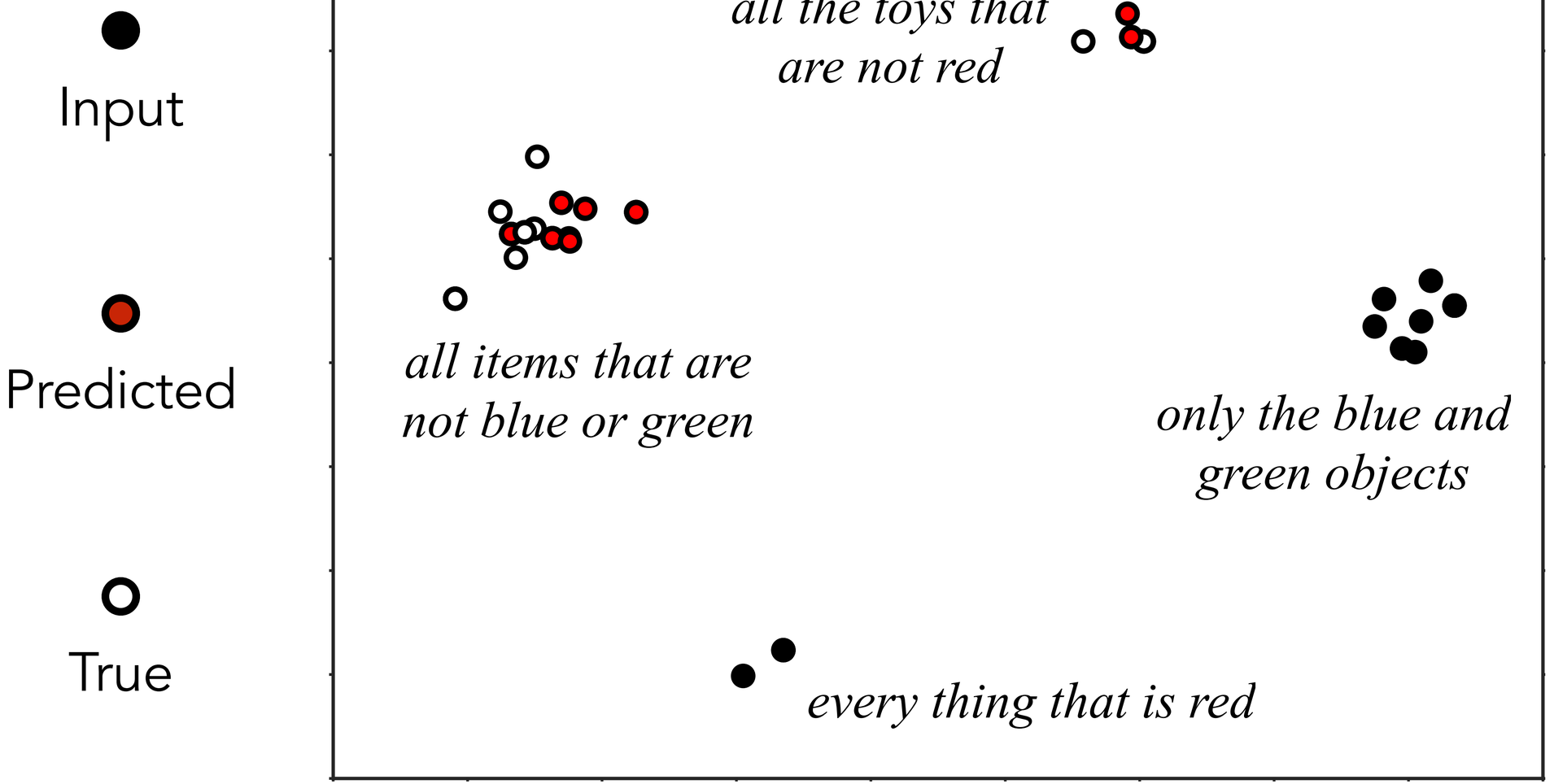
Evaluation: negation



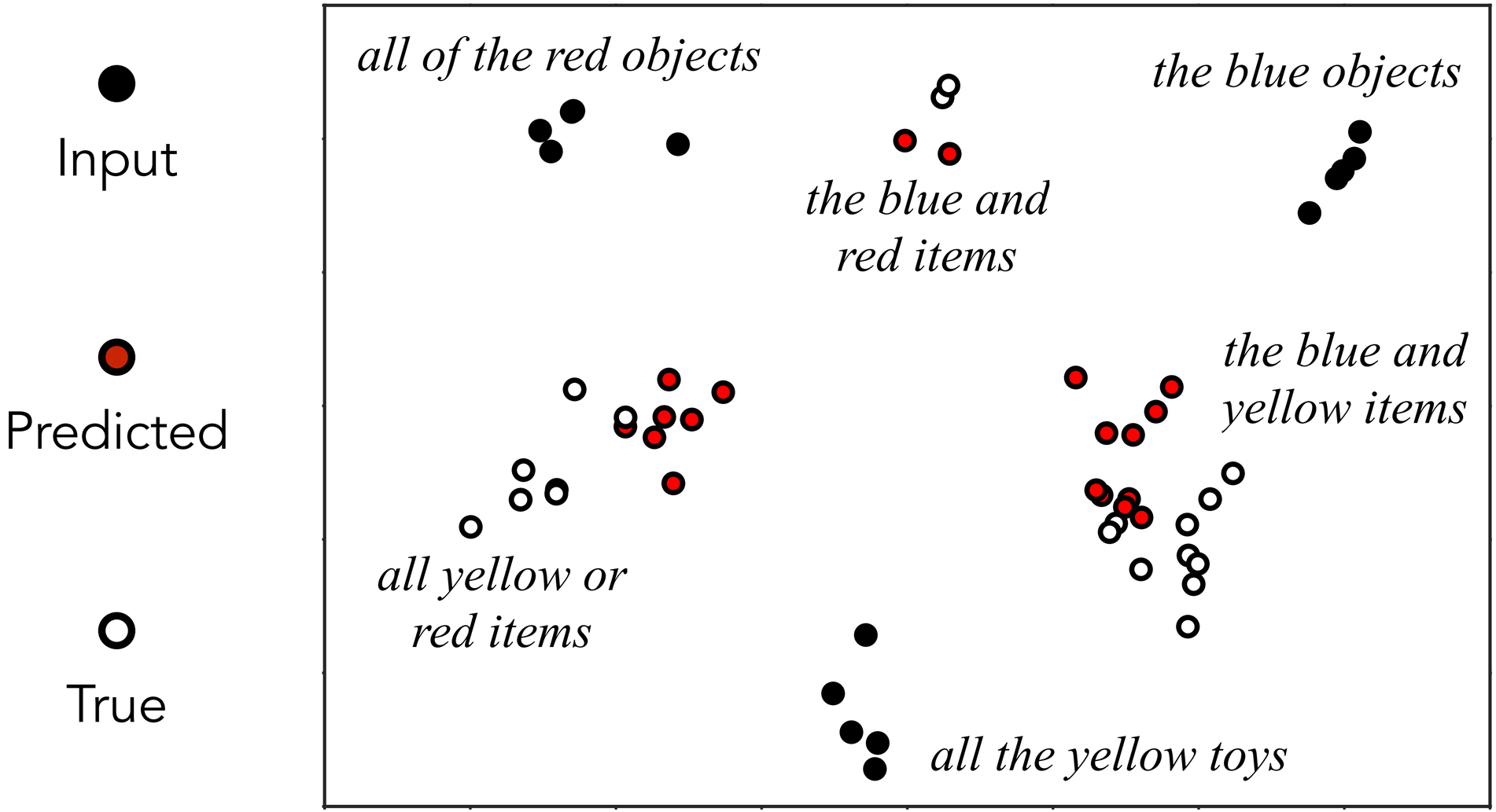
Evaluation: negation



Visualizing negation



Visualizing disjunction



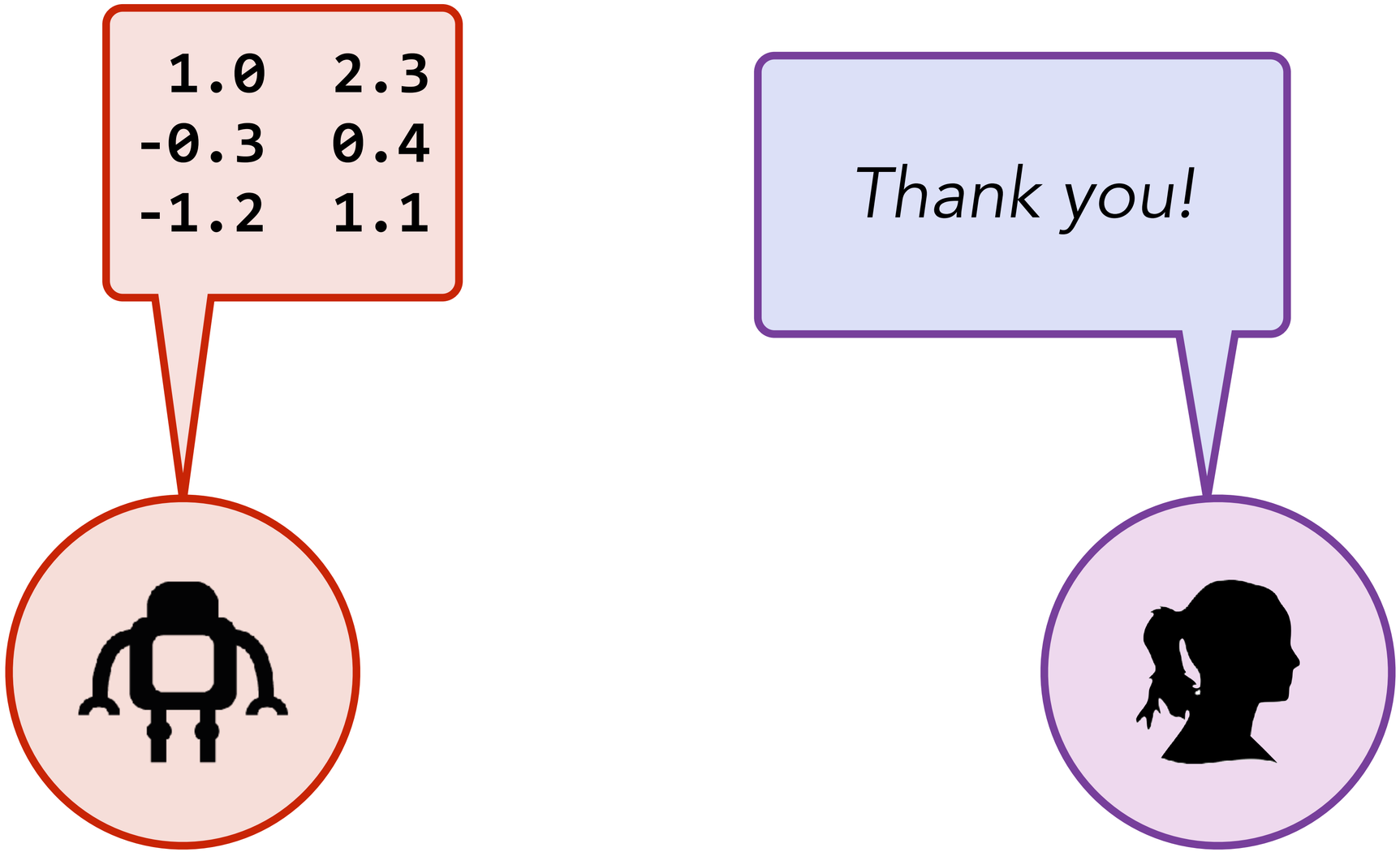
Conclusions

- Under the right conditions, RNN reprs exhibit interpretable pragmatics & compositional structure
- Not just communication games—language might be a good general-purpose tool for interpreting deep reprs.

Conclusions

- Under the right conditions, RNN reprs exhibit interpretable pragmatics & compositional structure

- Not just communication games—language might be a good general-purpose tool for interpreting deep reprs.



http://github.com/jacobandreas/rnn-syn