Natural Language Explanations of Deep Networks

Jacob Andreas

MIT CSAIL

LINGO.csail.mit.edu

Deep learning is hard



Deep learning is hard



What is this a picture of?

[Photo credit: Lucy Li (Wikipedia), dialogue: http://demo.visualdialog.org/]

Deep learning is hard



What is this a picture of?

```
it looks like a
<unknown_word> of some
sort
```

[Photo credit: Lucy Li (Wikipedia), dialogue: http://demo.visualdialog.org/]

Deep learning is hard



What is this a picture of?

```
it looks like a
<unknown_word> of some
sort
```

How many giraffes are in this picture?

[Photo credit: Lucy Li (Wikipedia), dialogue: http://demo.visualdialog.org/, giraffe question: Janelle Shane]

Deep learning is hard



What is this a picture of?

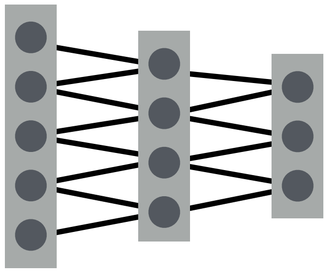
```
it looks like a
<unknown_word> of some
sort
```

How many giraffes are in this picture?

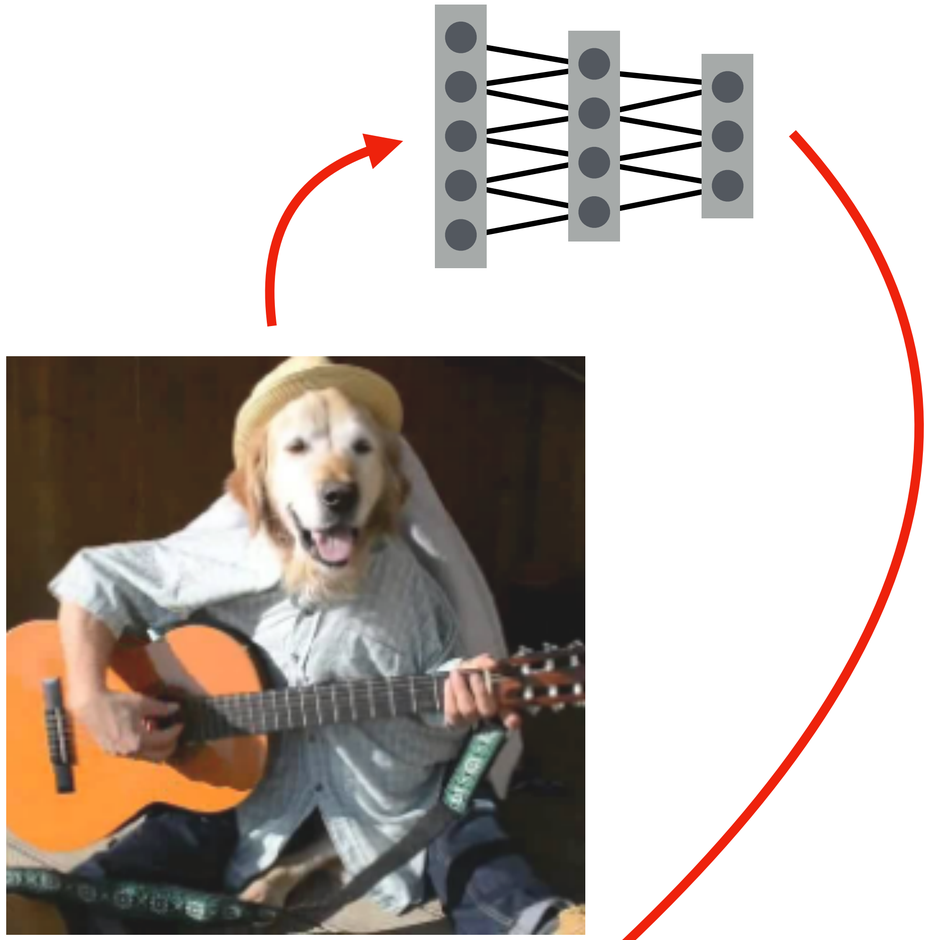
```
too many to count
```

[Photo credit: Lucy Li (Wikipedia), dialogue: http://demo.visualdialog.org/, giraffe question: Janelle Shane]

Explaining deep networks is hard



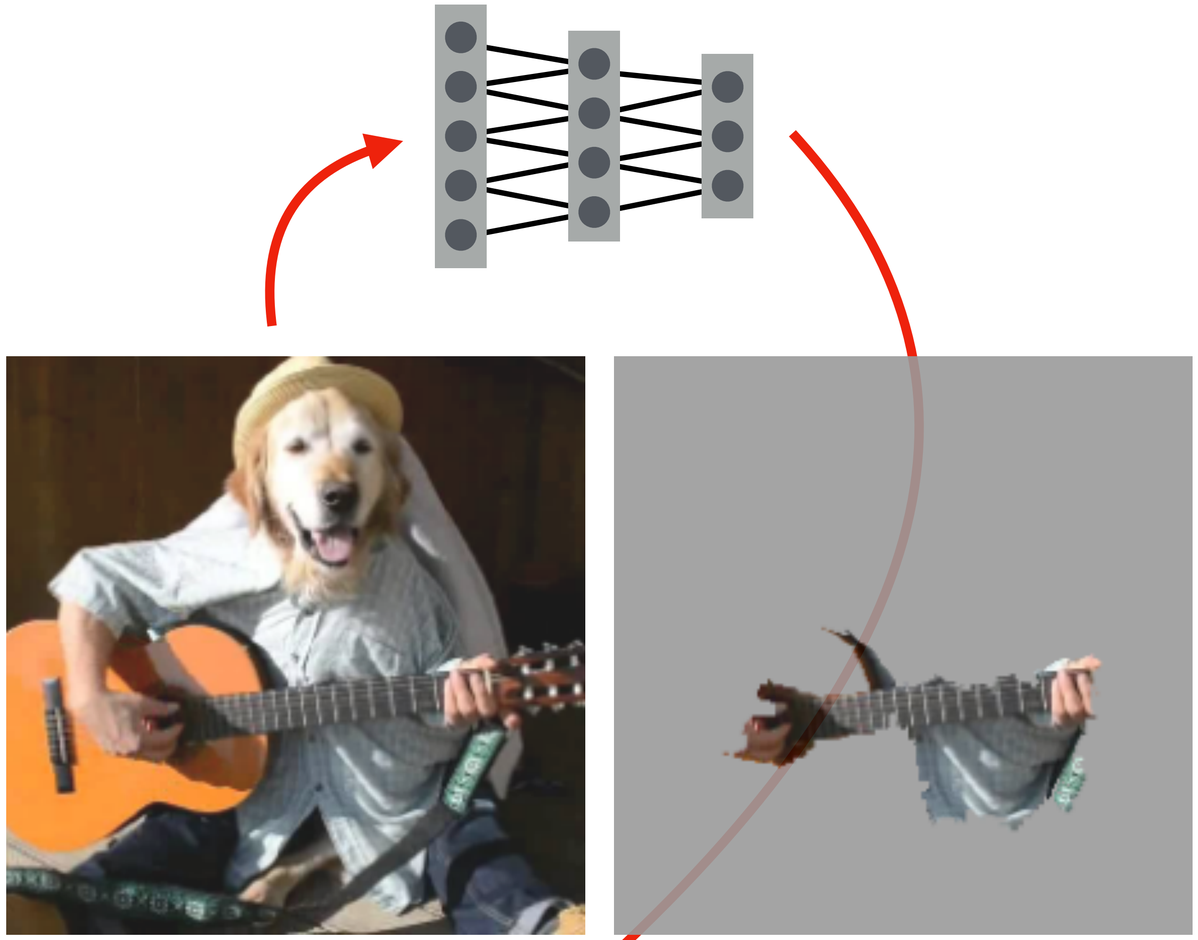
Explaining deep networks is hard



`electric guitar`

[Ribeiro et al. 2016. LIME]

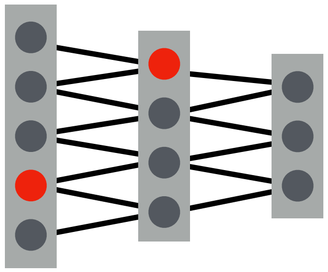
Explaining deep networks is hard



`electric guitar`

[Ribeiro et al. 2016. LIME]

Explaining deep networks is hard





electric guitar



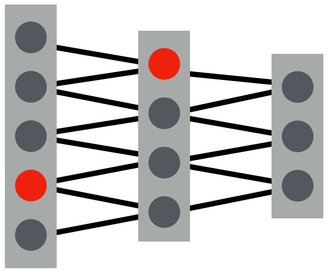
[Ribeiro et al. 2016. LIME]





[Olah et al. 2017. The building blocks of interpretability.]

Explaining deep networks is hard

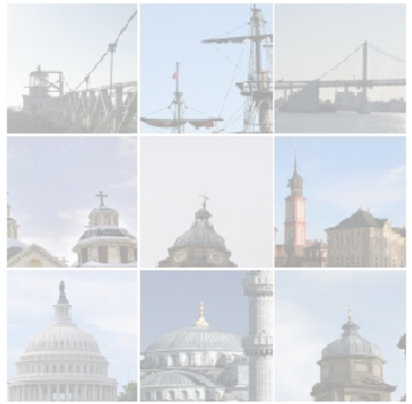


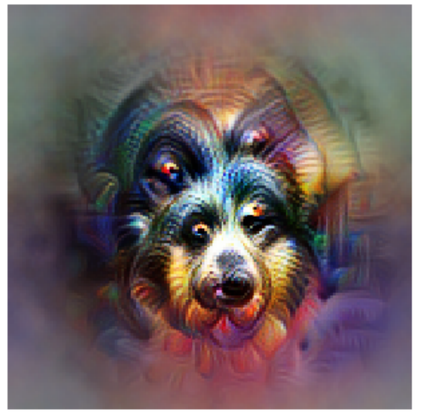




electric guitar [Ribeiro et al. 2016. LIME]



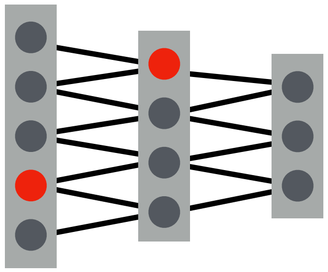


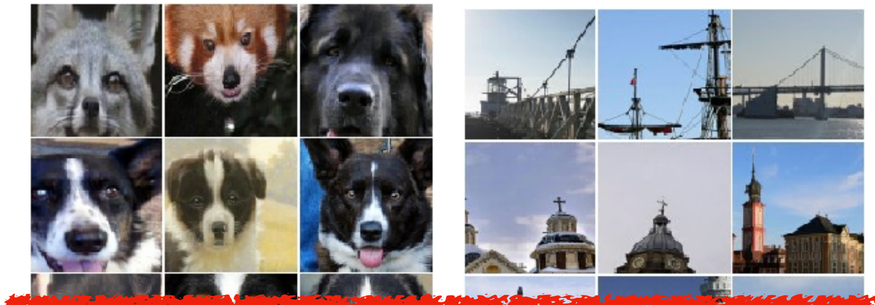




[Olah et al. 2017. The building blocks of interpretability.]

Explaining deep networks is hard





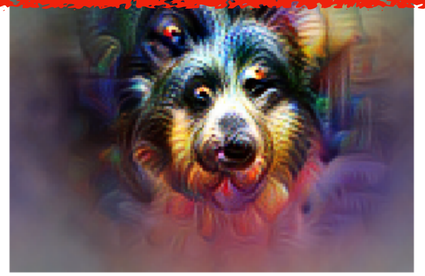
What aspect of this region is relevant?

What do these pictures have in common?





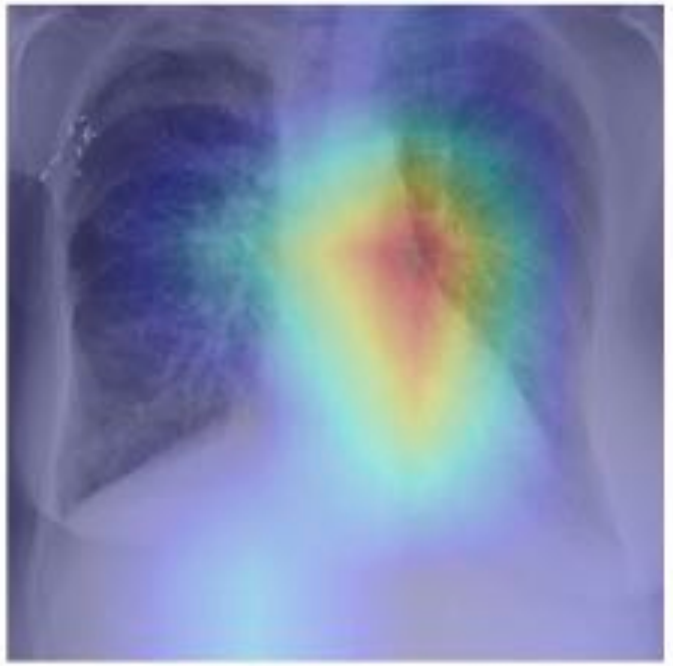
electric guitar [Ribeiro et al. 2016. LIME.]





[Olah et al. 2017. The building blocks of interpretability.]

Explaining deep networks is hard



`moderate pulmonary edema`

[Chauhan et al. 2020. Joint Modeling of Chest Radiographs and Radiology Reports for Pulmonary Edema Assessment]

What are explanations for?

Communicating:

- Model safety and potential vulnerabilities
- Changes needed to produce desired outputs
- Actions that *users* should take in collaborative settings

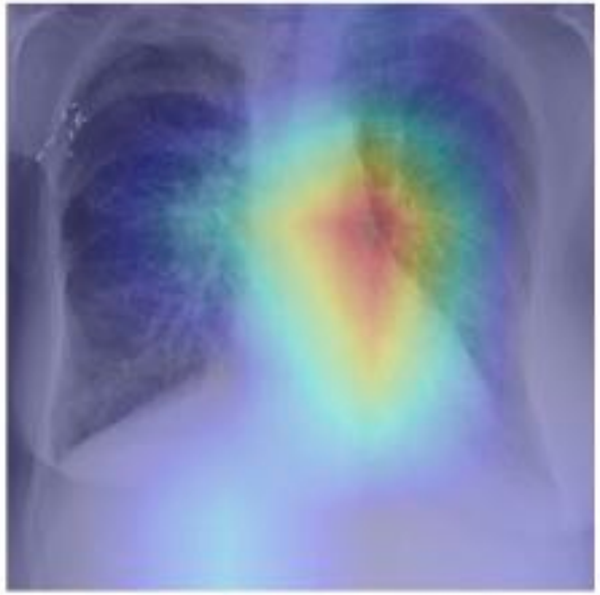
Communicating explanations



the fretboard is gray



mammals with black noses

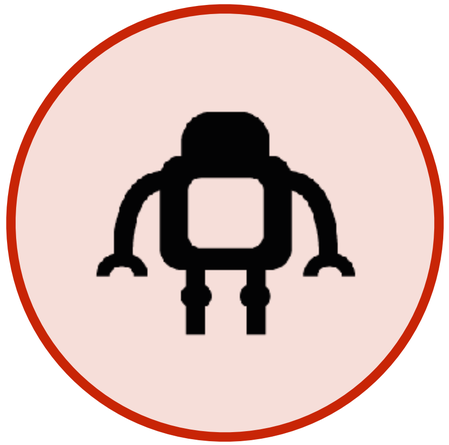


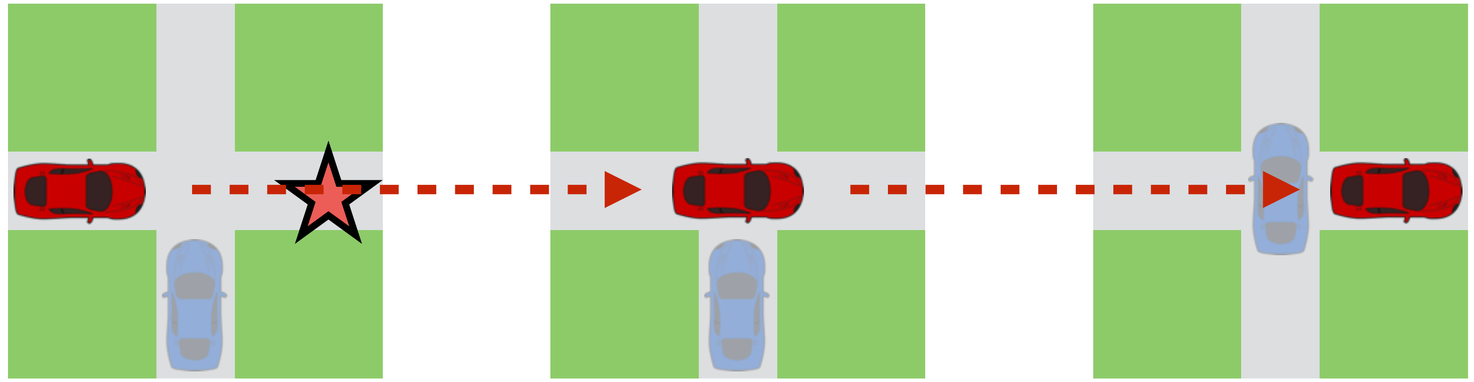
interstitial abnormality with peribronchal cuffing

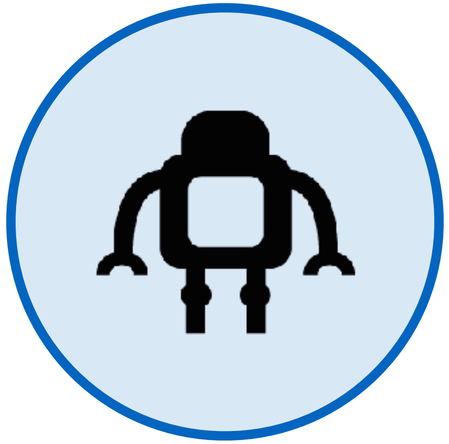
[c.f. Bansal et al. 2020. Does the Whole Exceed its Parts?
Zaidan et al. 2007. Using “Annotator Rationales” to Improve Machine Learning for Text Categorization.]

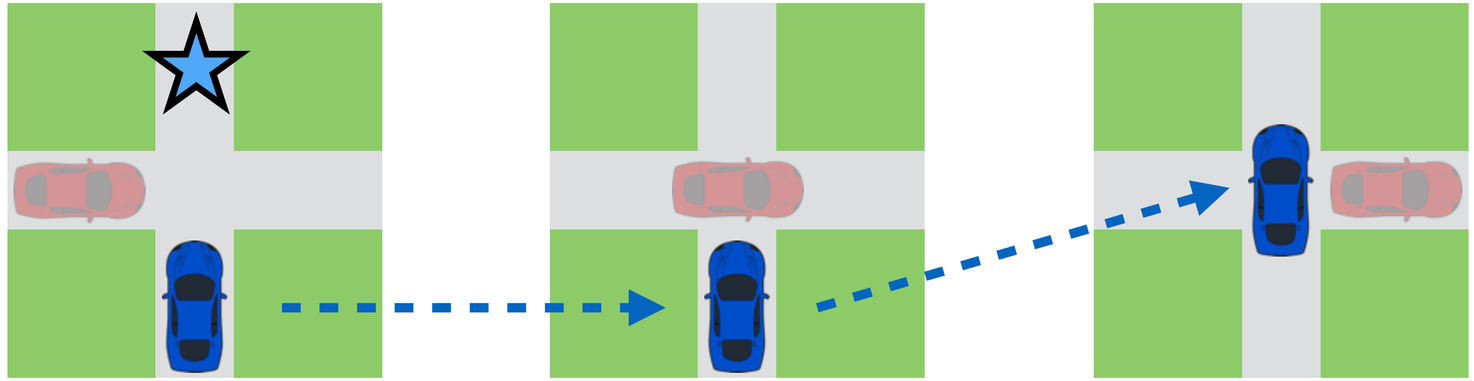
Explanation as translation

Multi-agent communication

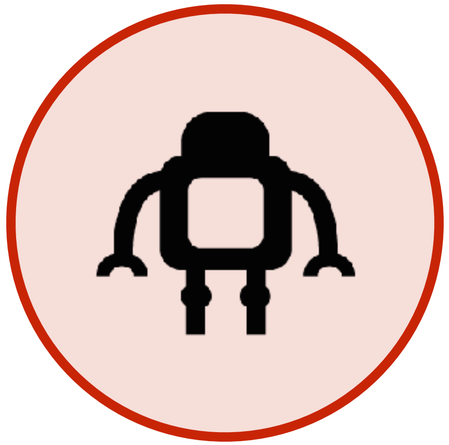


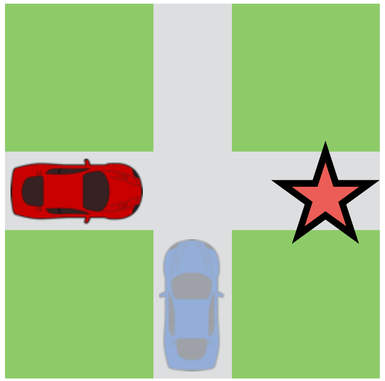


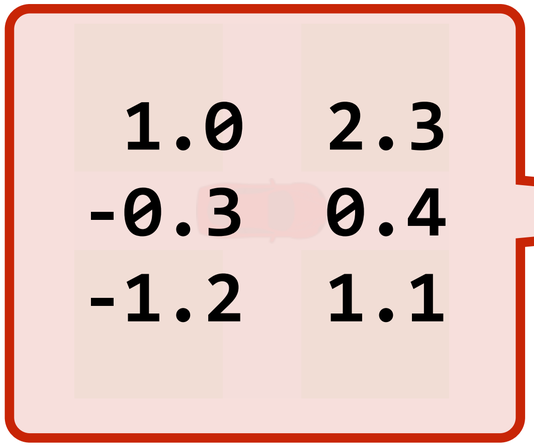


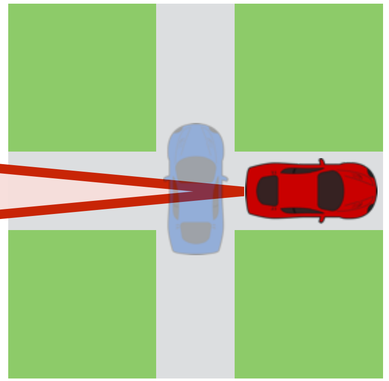


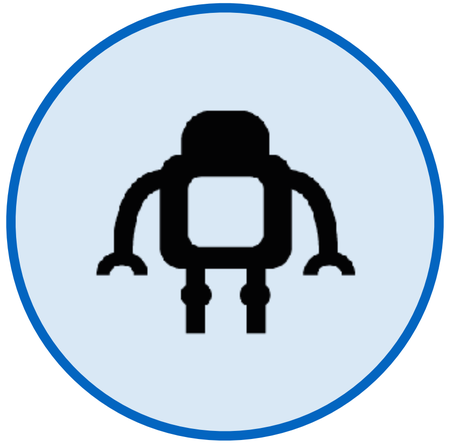
Neuralese

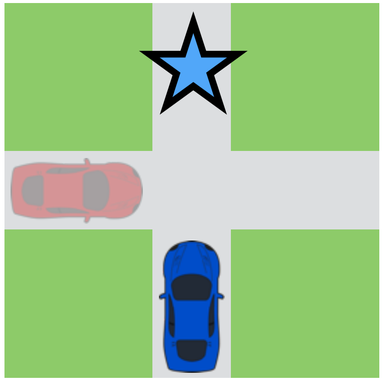


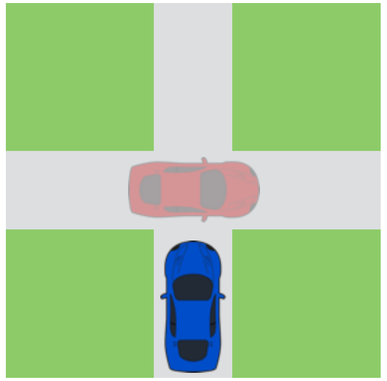


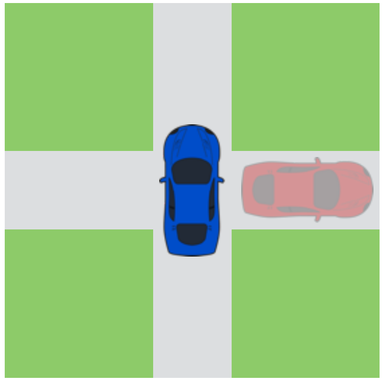






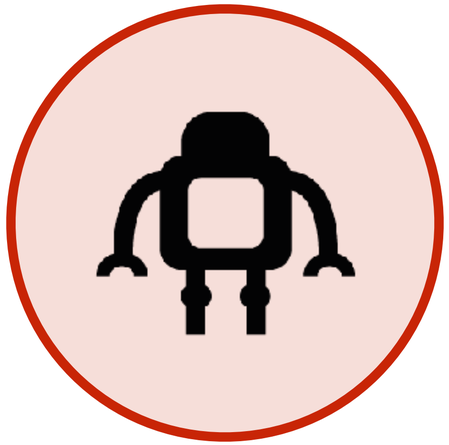






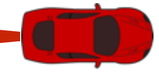
[Sukhbataar et al. 16. Learning multiagent communication with backpropagation.]

Translating neuralese



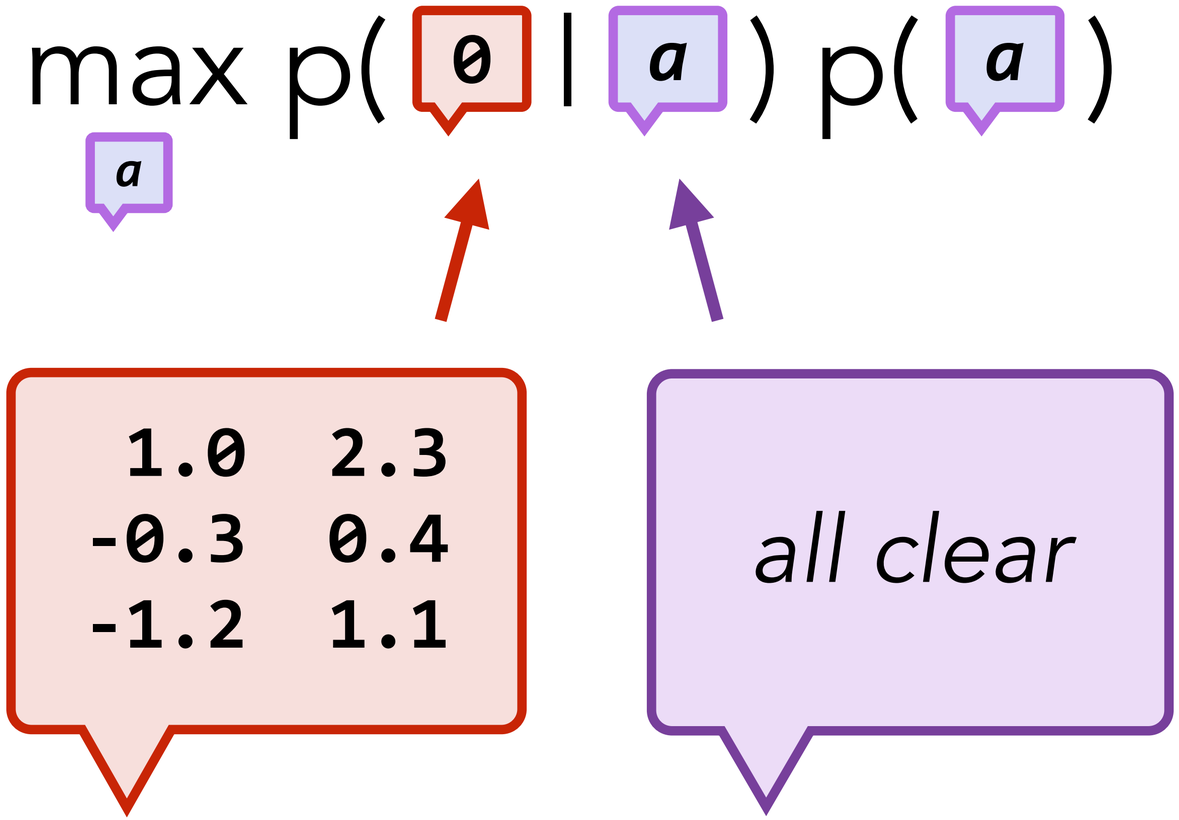


```
 1.0  2.3
-0.3  0.4
-1.2  1.1
```



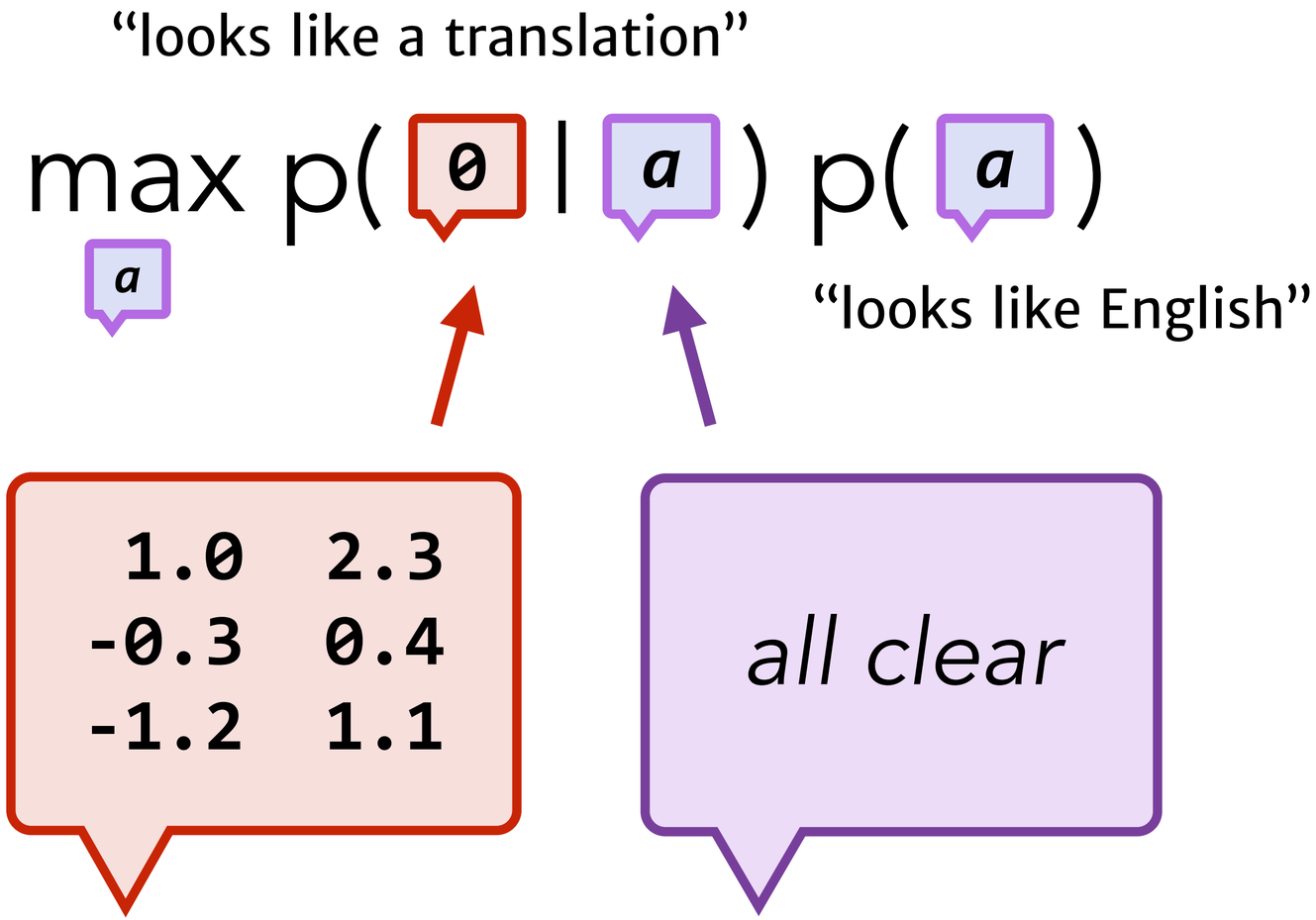
all clear

A statistical MT problem



[e.g. Koehn 10. Phrase-based statistical machine translation.]

A statistical MT problem



[e.g. Koehn 10. Phrase-based statistical machine translation.]

A statistical MT problem

p(a)

p(ñ | a)

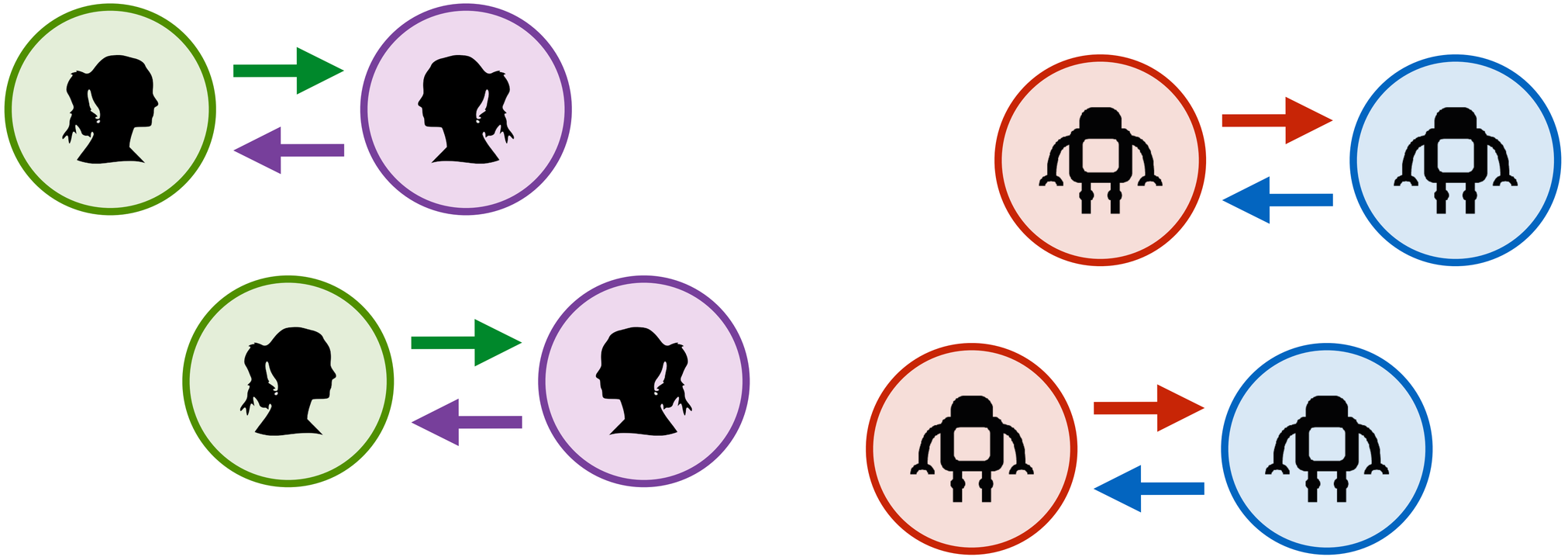
all clear

I'm going north

listo

voy al norte

A statistical MT problem

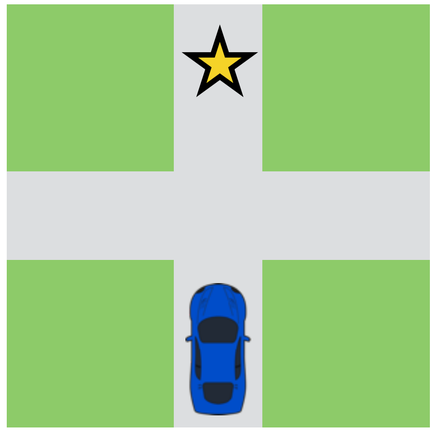


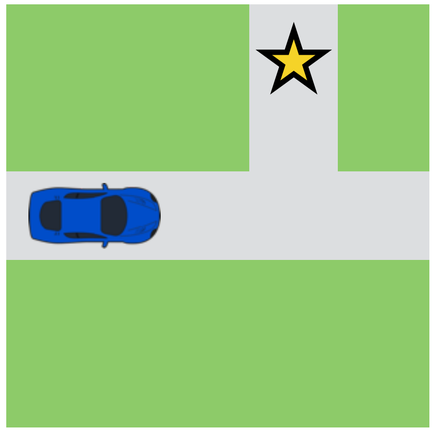
How do we induce a translation model?

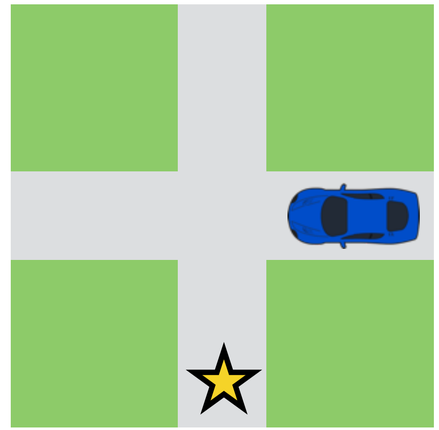
A “semantic MT” problem

The meaning of an utterance is given by its **truth conditions**

I’m going north





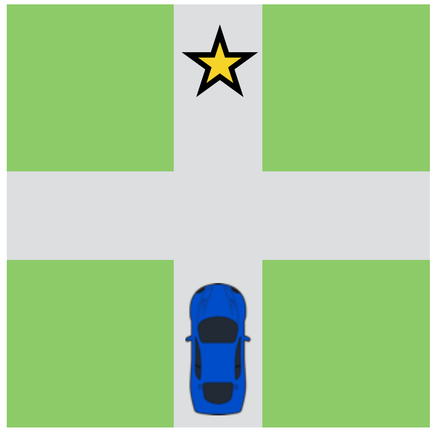


[Davidson 67. Truth and meaning.]

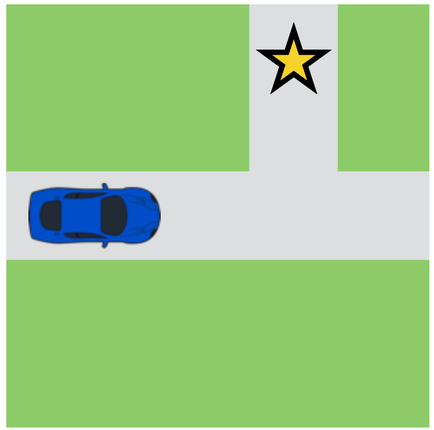
A “semantic MT” problem

The meaning of an utterance is given by its **truth conditions**

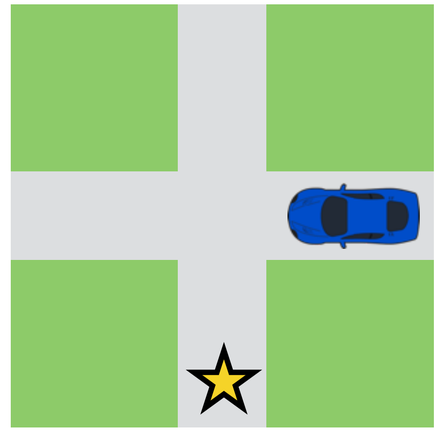
I’m going north



✔



✔



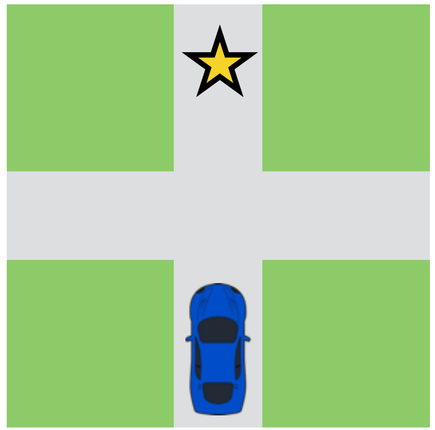
✘

A “semantic MT” problem

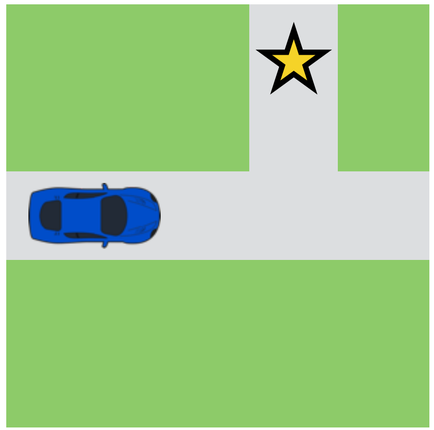
The meaning of an utterance is given by its truth conditions

→ the **distribution over states** in which it is uttered

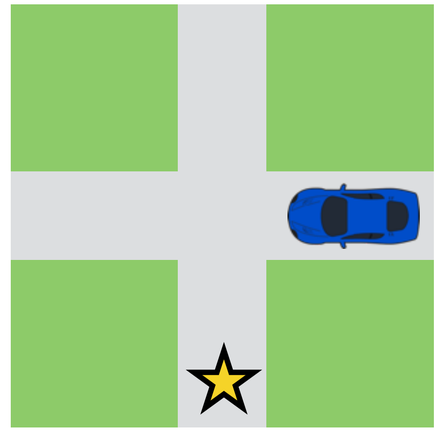
I’m going north



0.4



0.2



0.001

[Beltagy et al. 14. Efficient Markov Logic Inference for Natural Language Semantics]

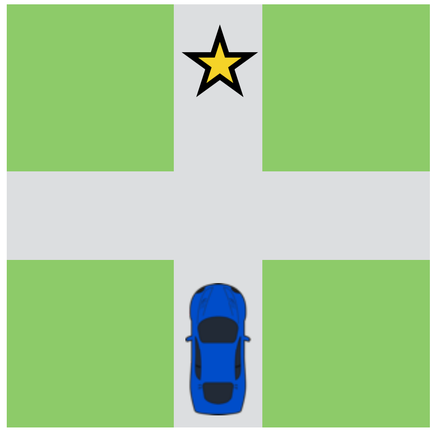
A “semantic MT” problem

The meaning of an utterance is given by its **truth conditions**

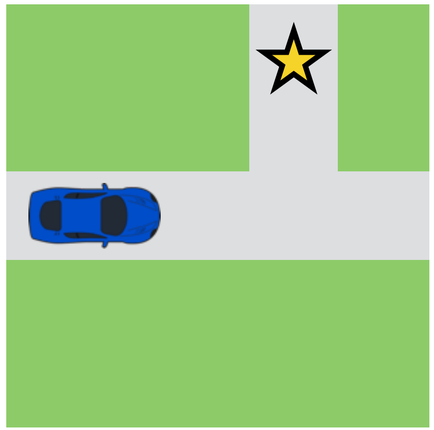
→ the **distribution over states** in which it is uttered

→ the **belief** it induces in listeners

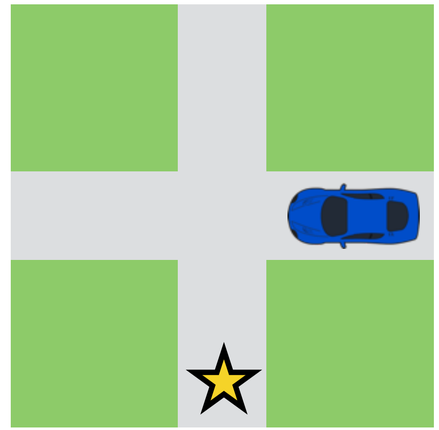
I’m going north



0.4



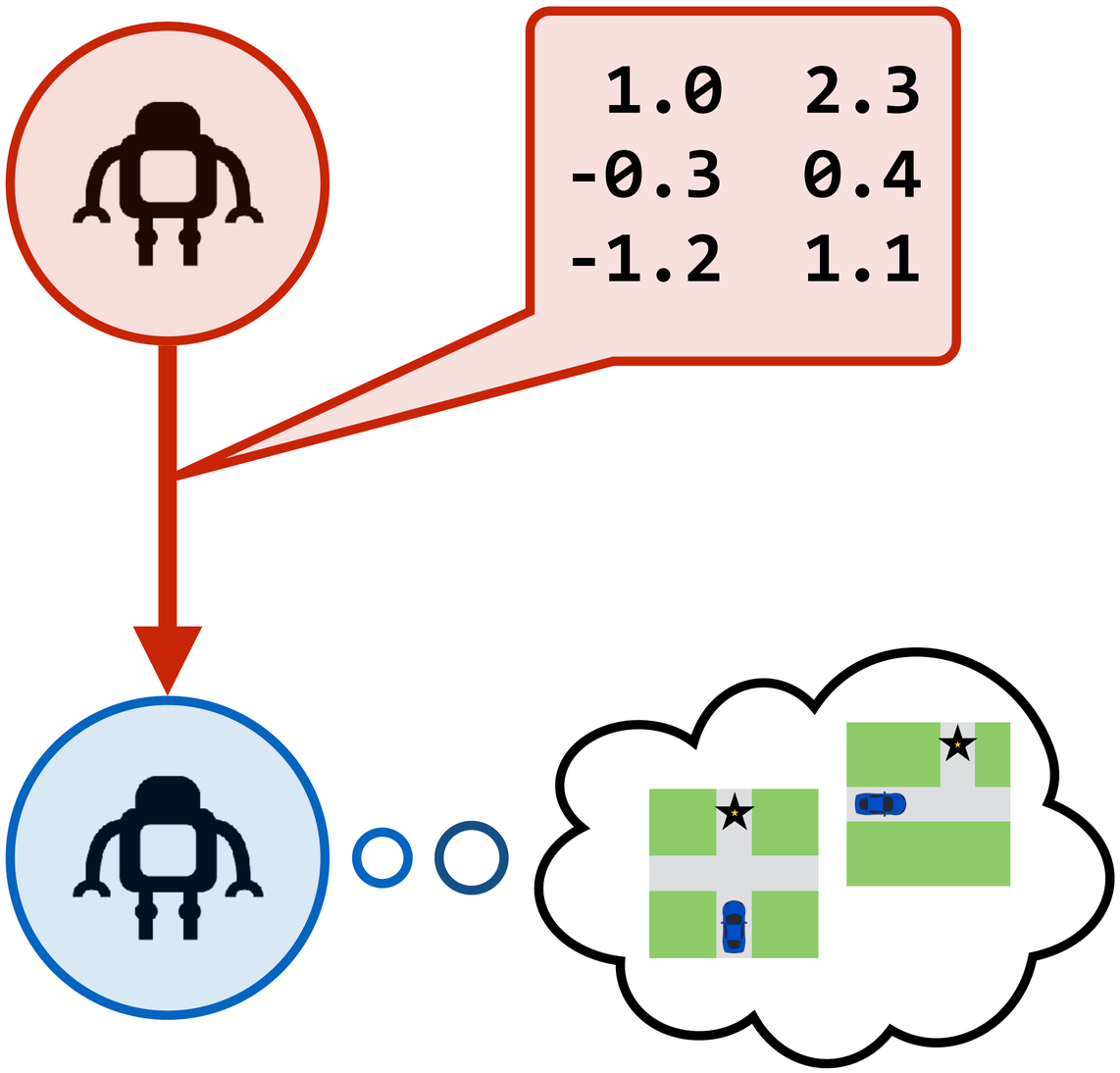
0.2



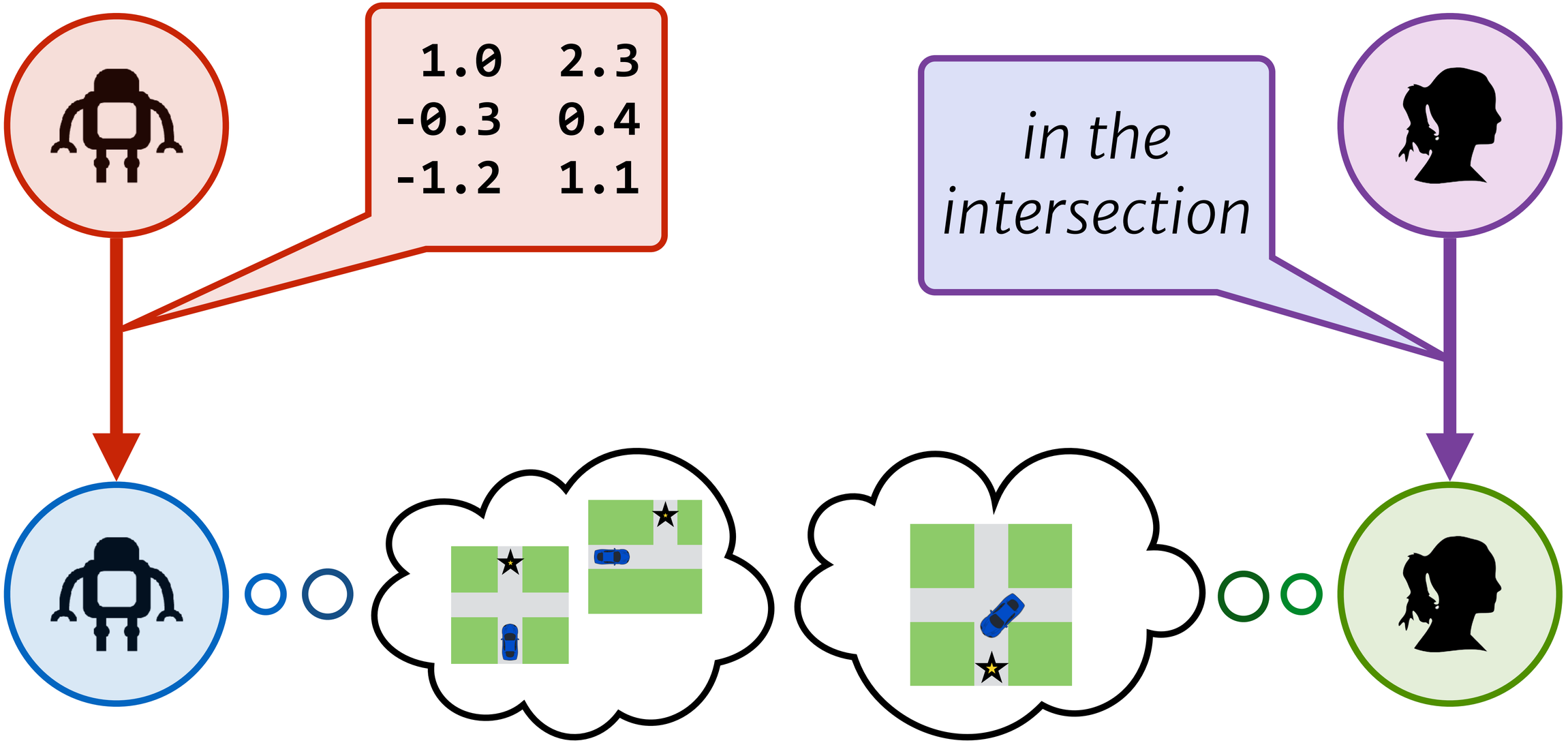
0.001

[Frank et al. 09. Informative communication in word production and word learning.]

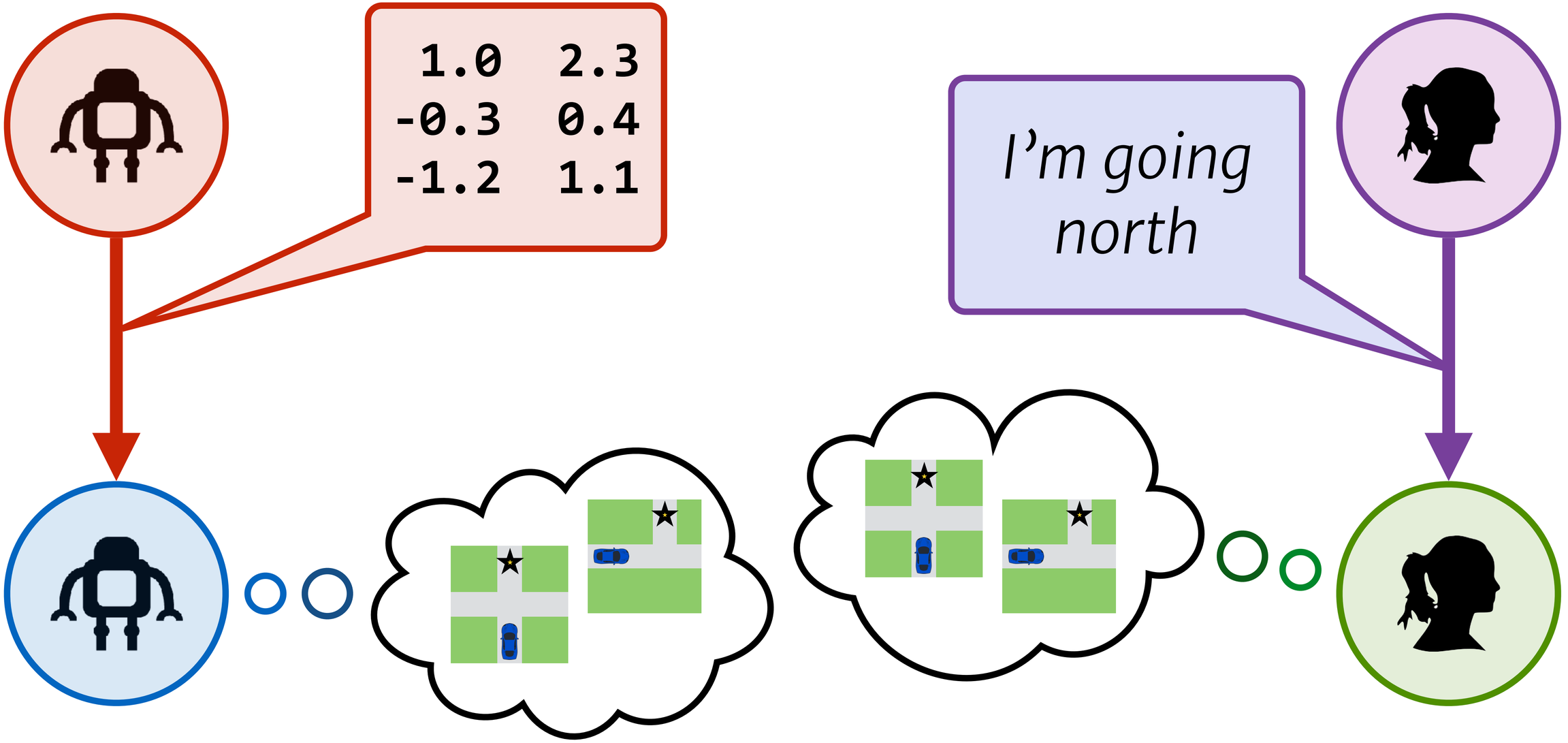
Translating via belief



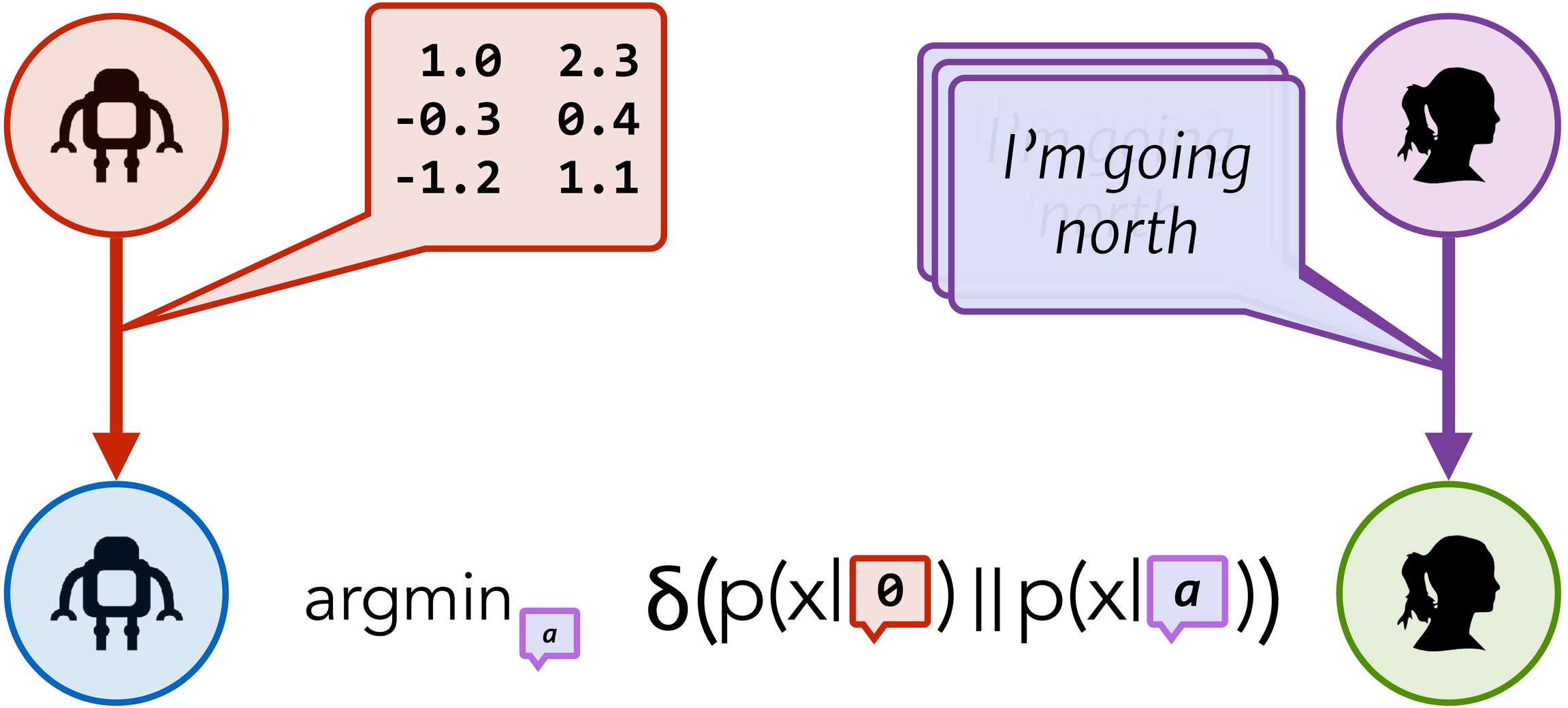
Translating via belief



Translating via belief



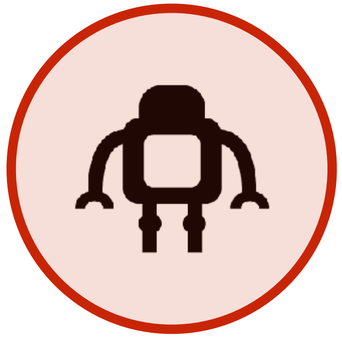
Translating via belief

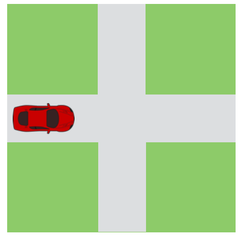


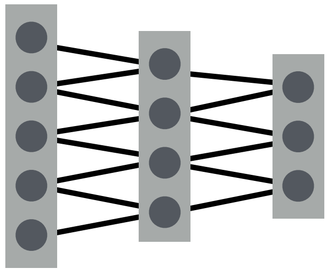
[A, Dragan and Klein 17. Translating neuralese.]

Computing beliefs

agent policy

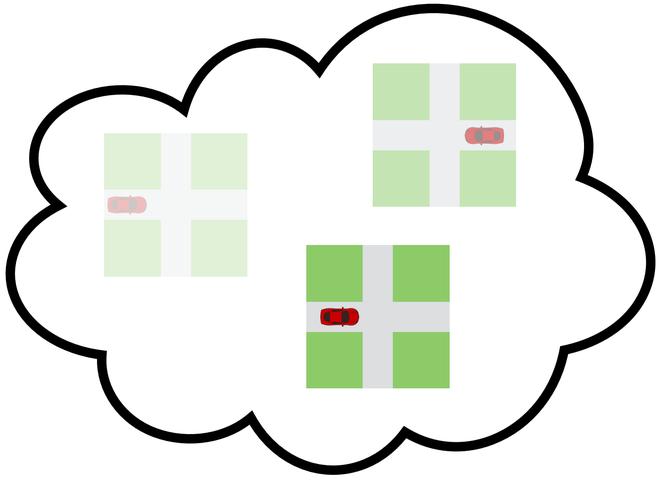


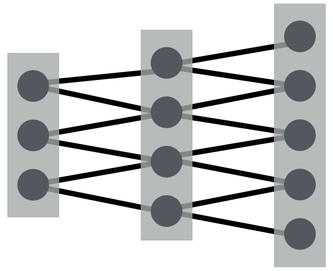




actions & messages

agent model

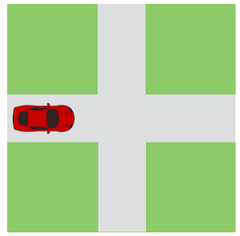




Computing beliefs

human policy

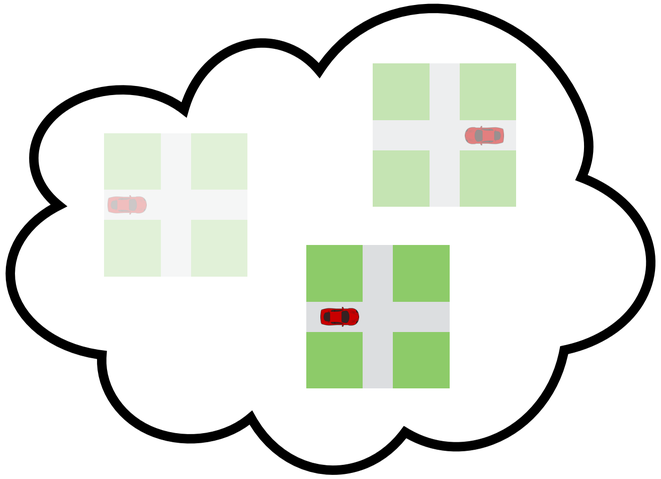


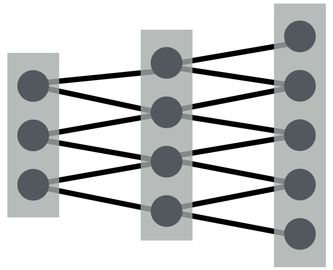




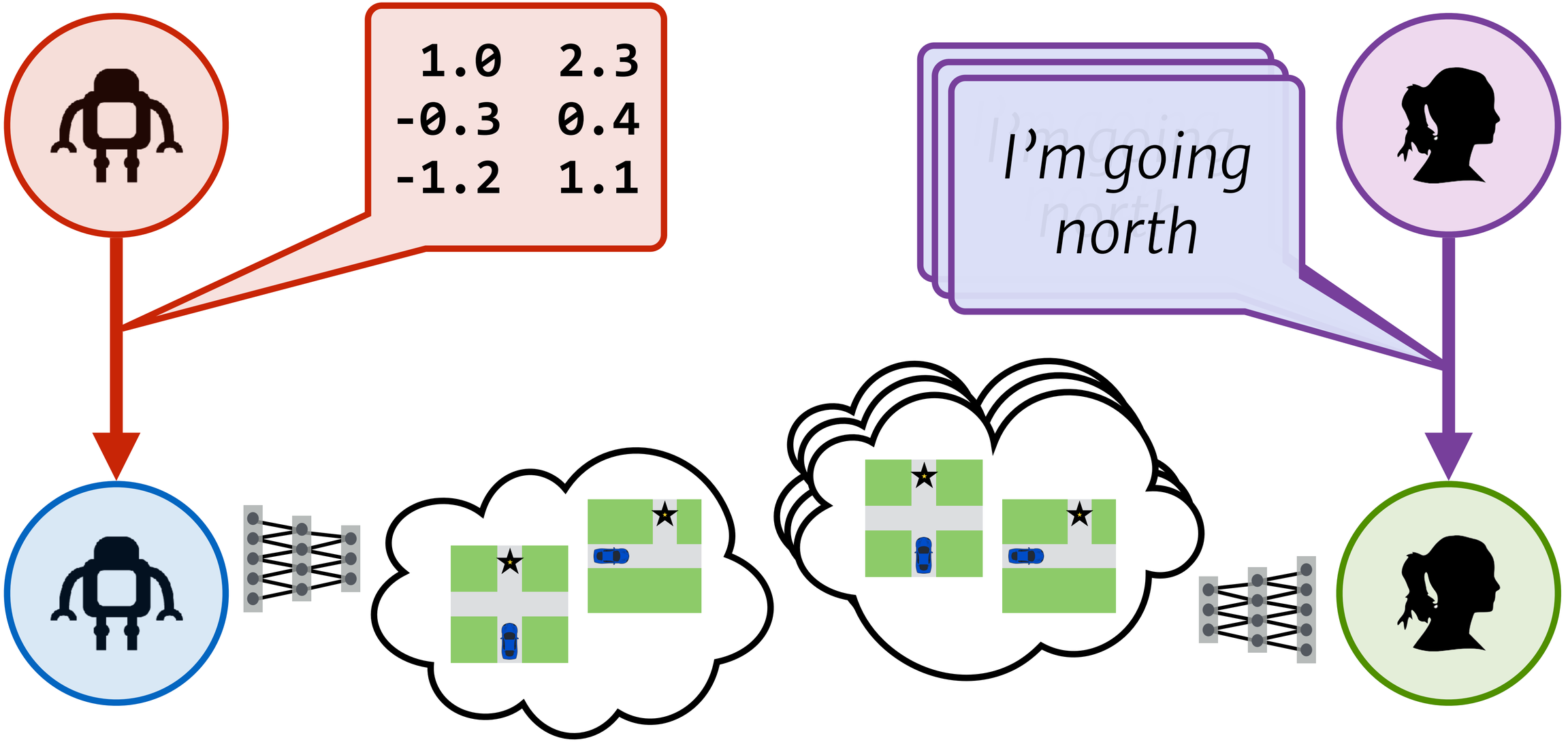
actions & messages

human model

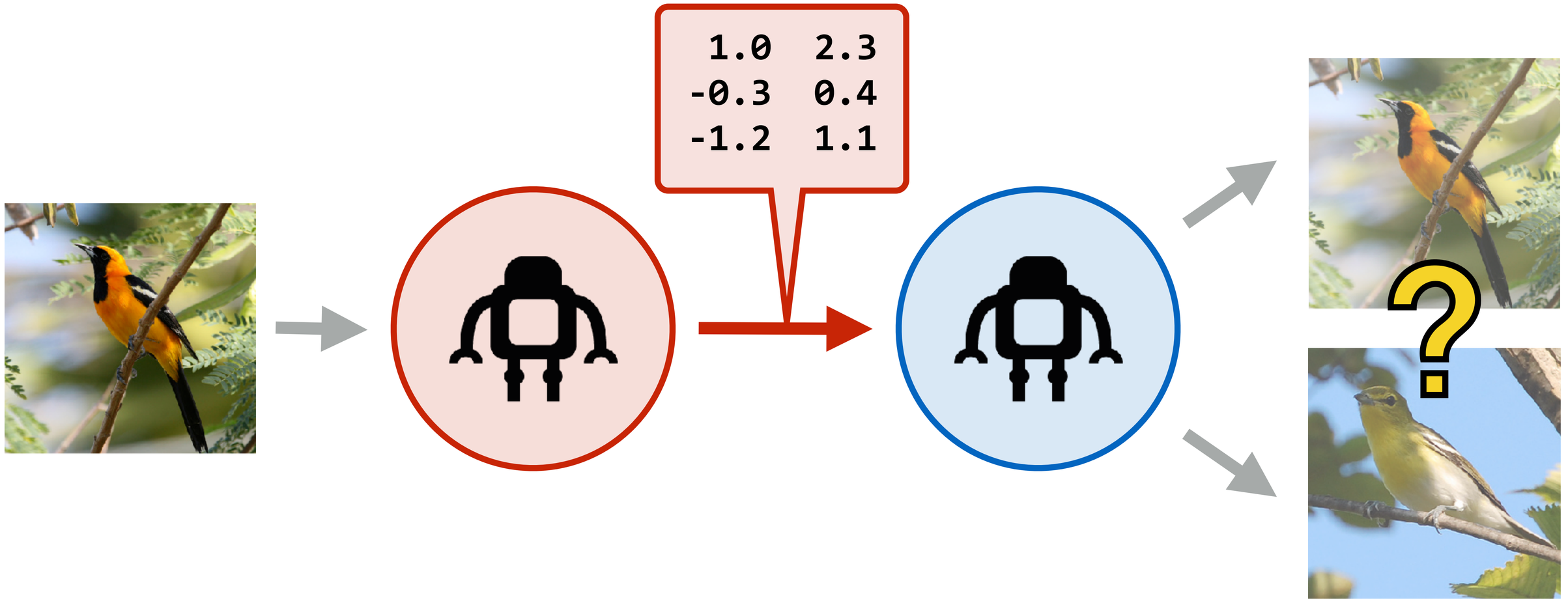




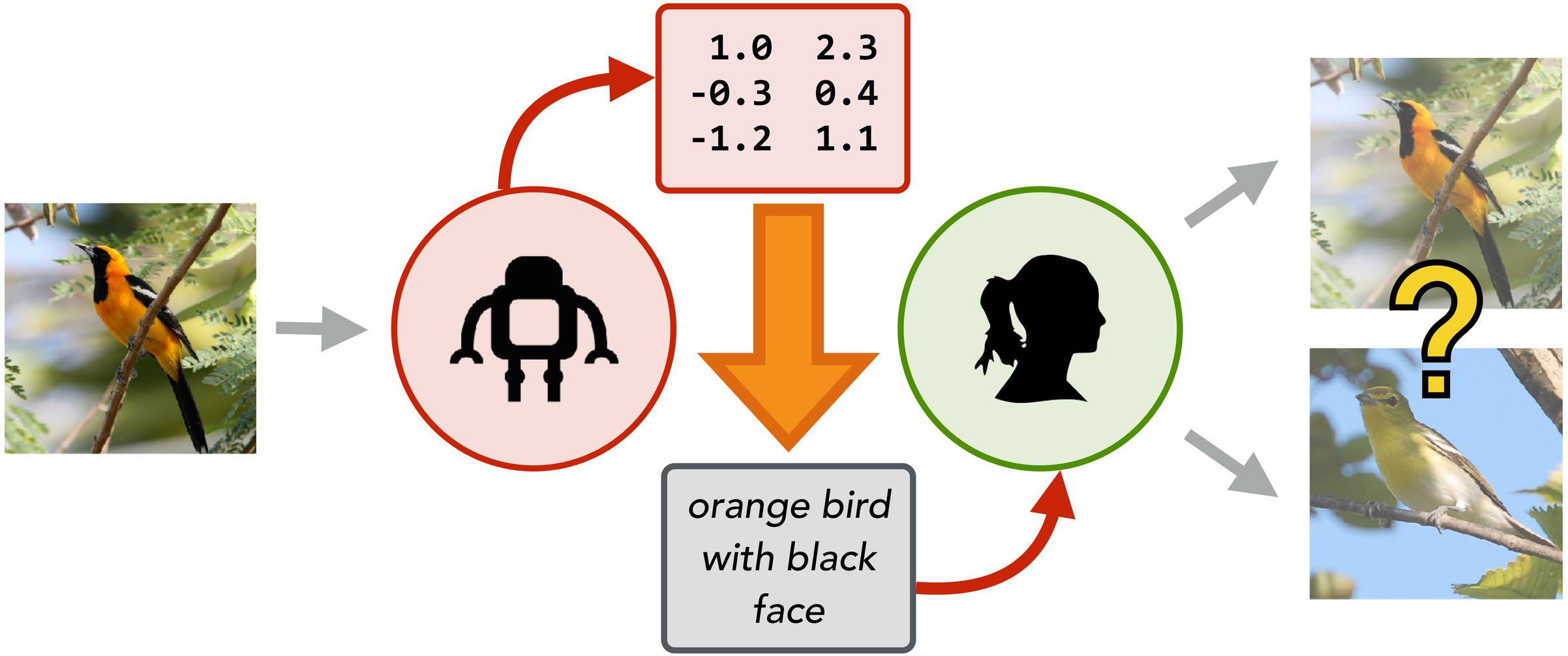
Translating via belief



Reference games



Evaluation: translator-in-the-loop



Example translations: image classification

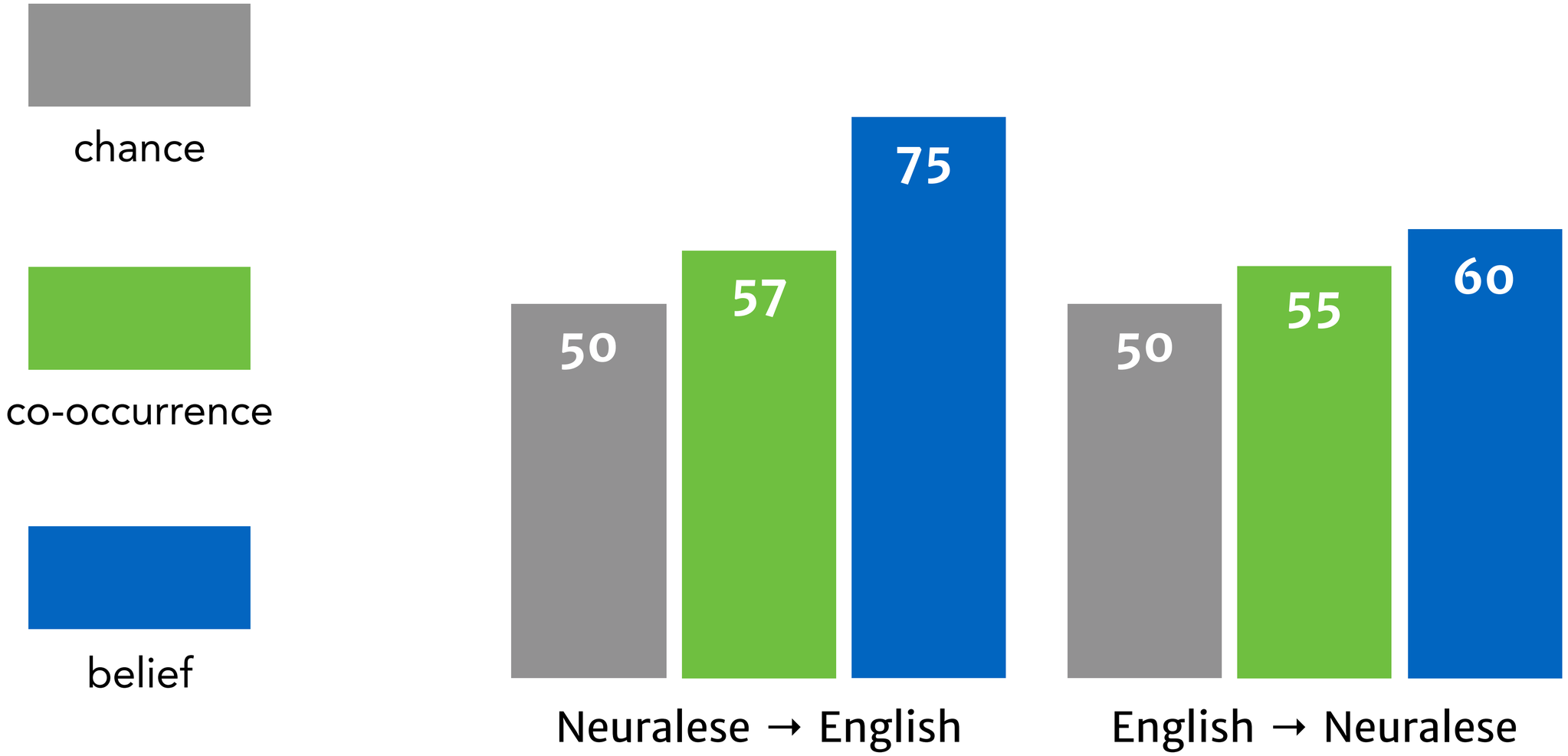


large bird, black wings, black crown

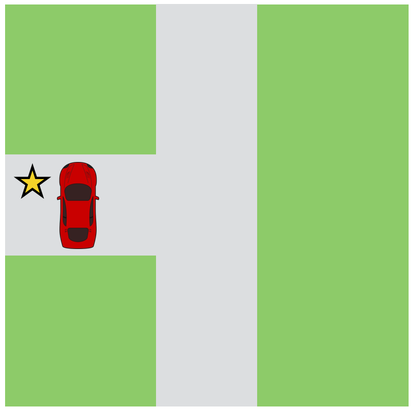


small brown, light brown, dark brown

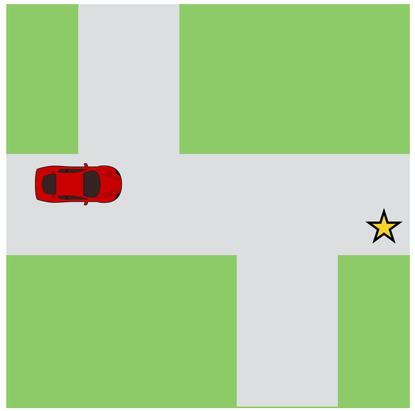
Experiment: image references



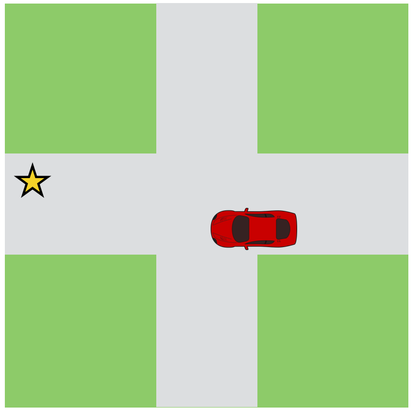
Example translations: driving



at goal
done
left to top



you first
following
going down

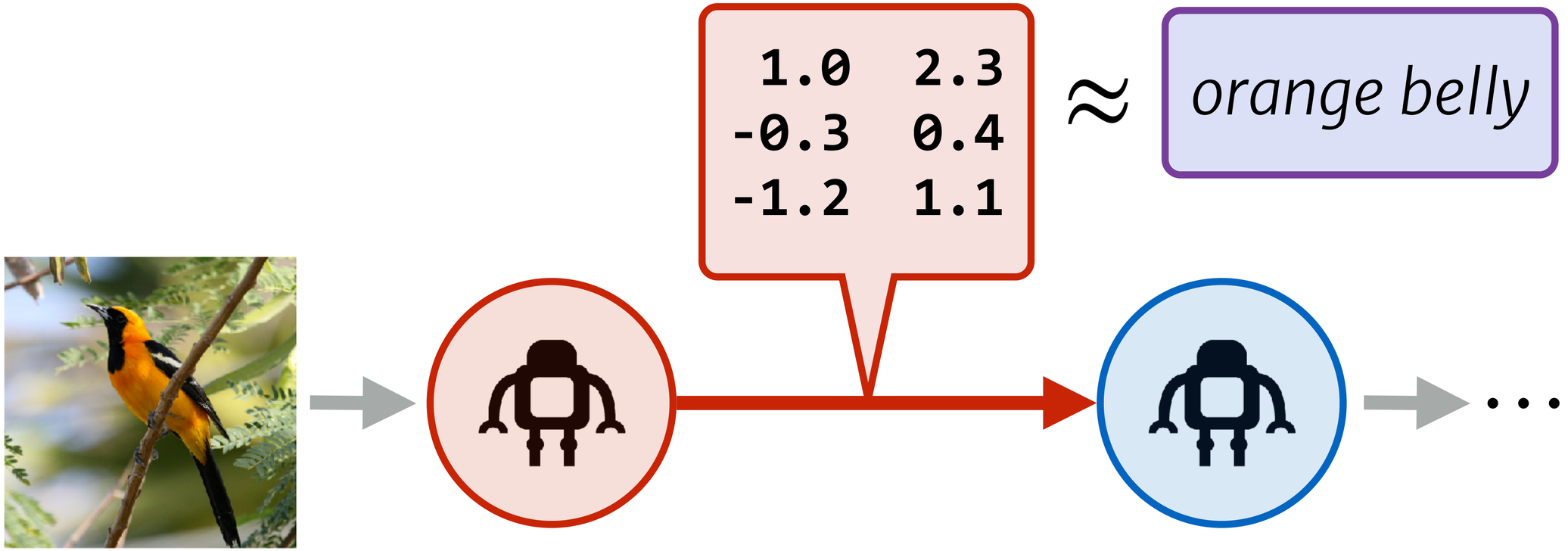


going in intersection
proceed
going

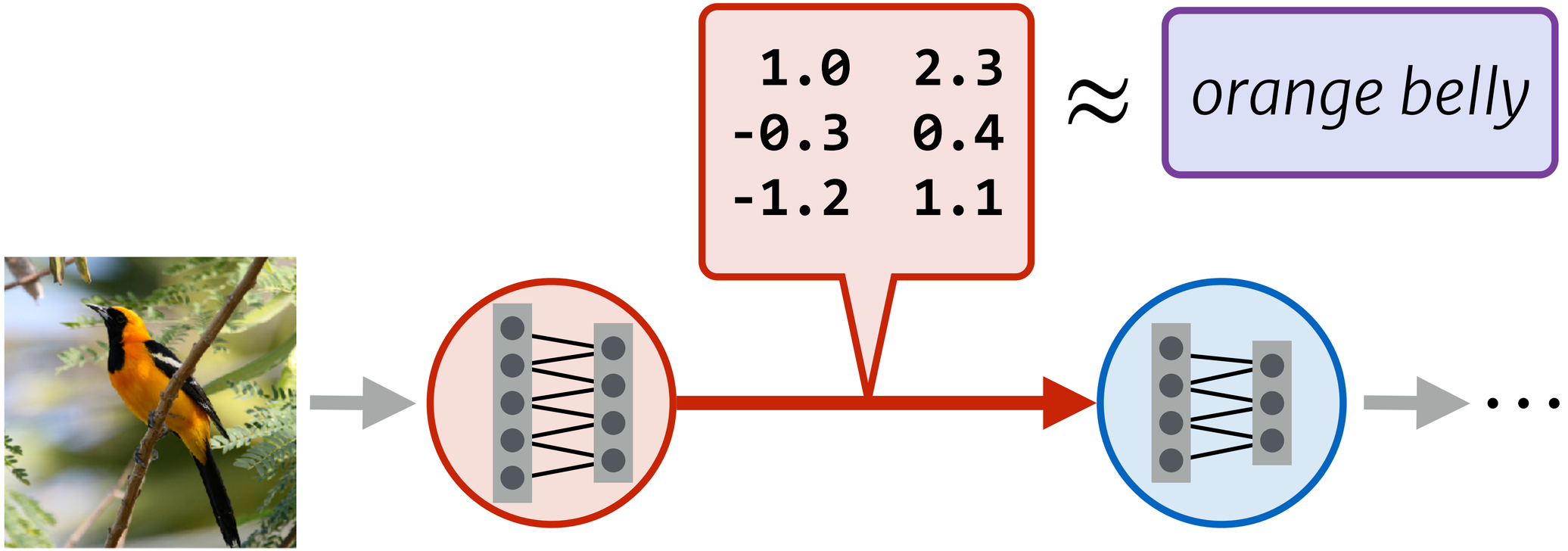
Experiment: driving game



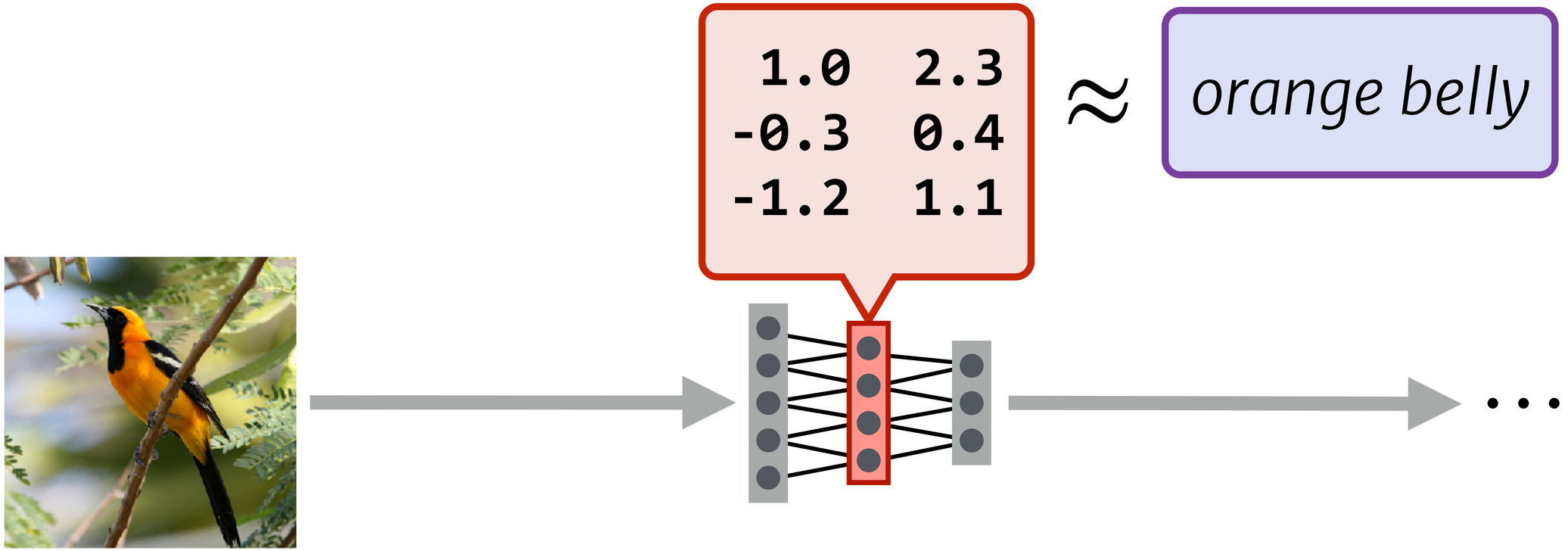
Beyond communication



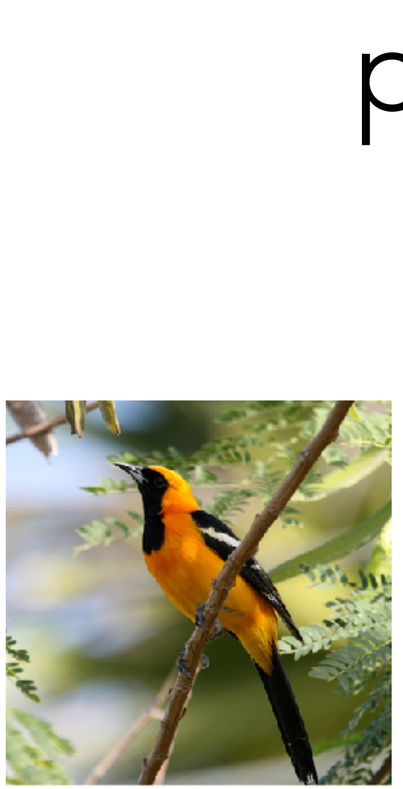
Beyond communication



Beyond communication



Beyond communication



x

$p(x|\texttt{0})$

1.0 2.3
-0.3 0.4
-1.2 1.1

$\approx$ *orange belly*

…

$\text{argmin}_{a}\ \delta(p(x|\texttt{0}) \,||\, p(x|a))$

Summary

1. Represent “meanings” of learned representations and features as distributions over model inputs, and generate explanations by summarizing these distributions with language.

Explanation as (program) induction



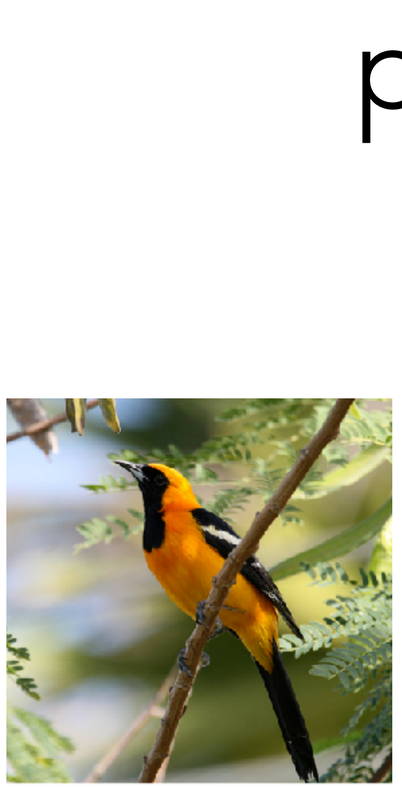
Evan Hernandez



Jesse Mu

(thanks to Sarah Schwettman, David Bau, Antonio Torralba)

Explaining representations



x

$p(x|\,\mathbf{0})$

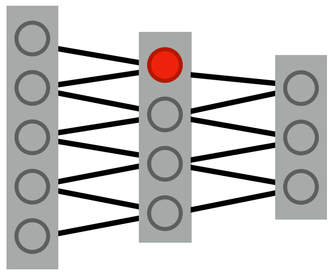
1.0 2.3
-0.3 0.4
-1.2 1.1

$\approx$ *orange belly*

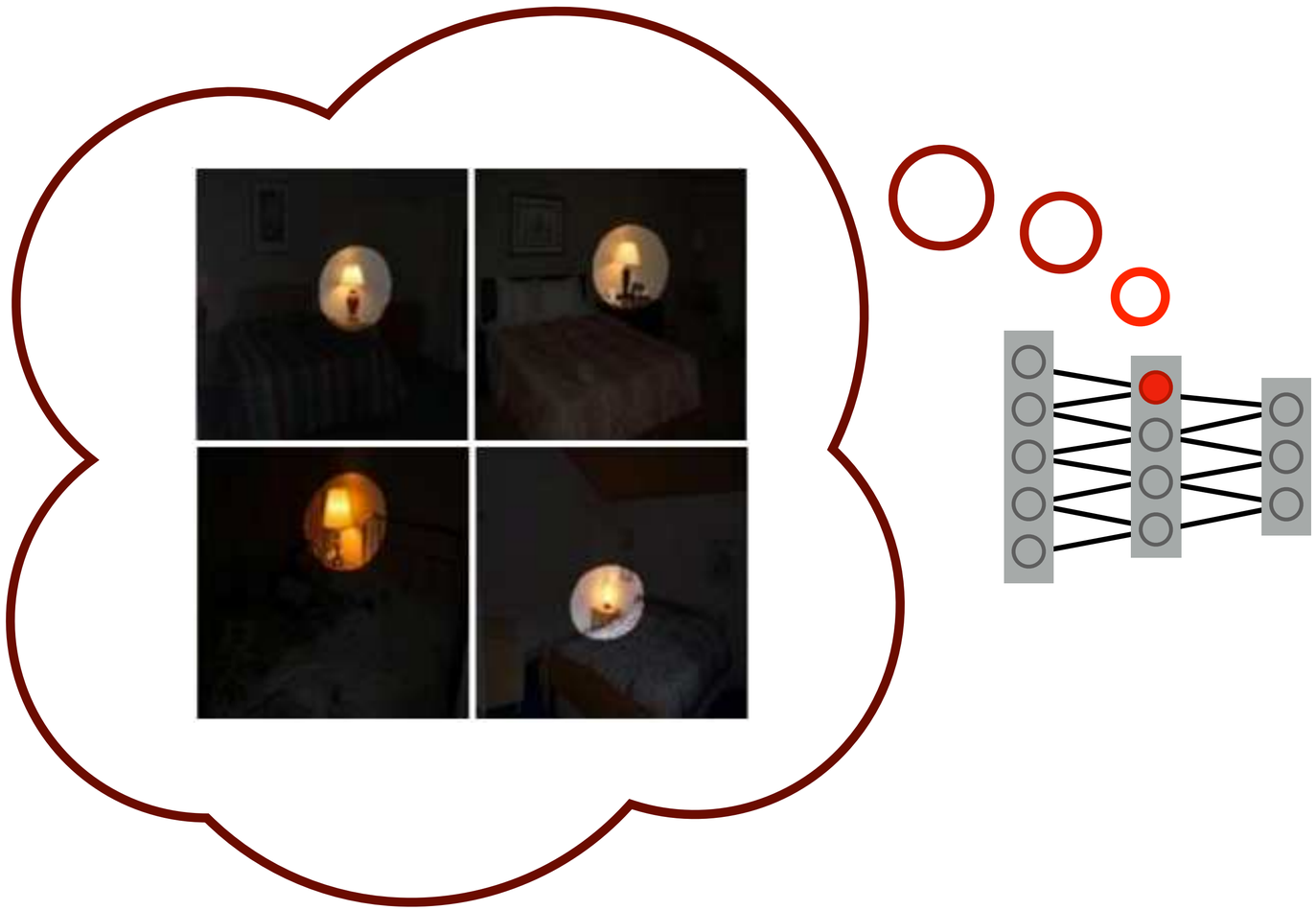
…

$\text{argmin}_{a}\ \delta(p(x|\,\mathbf{0}) \,||\, p(x|\,a))$

Explaining individual neurons

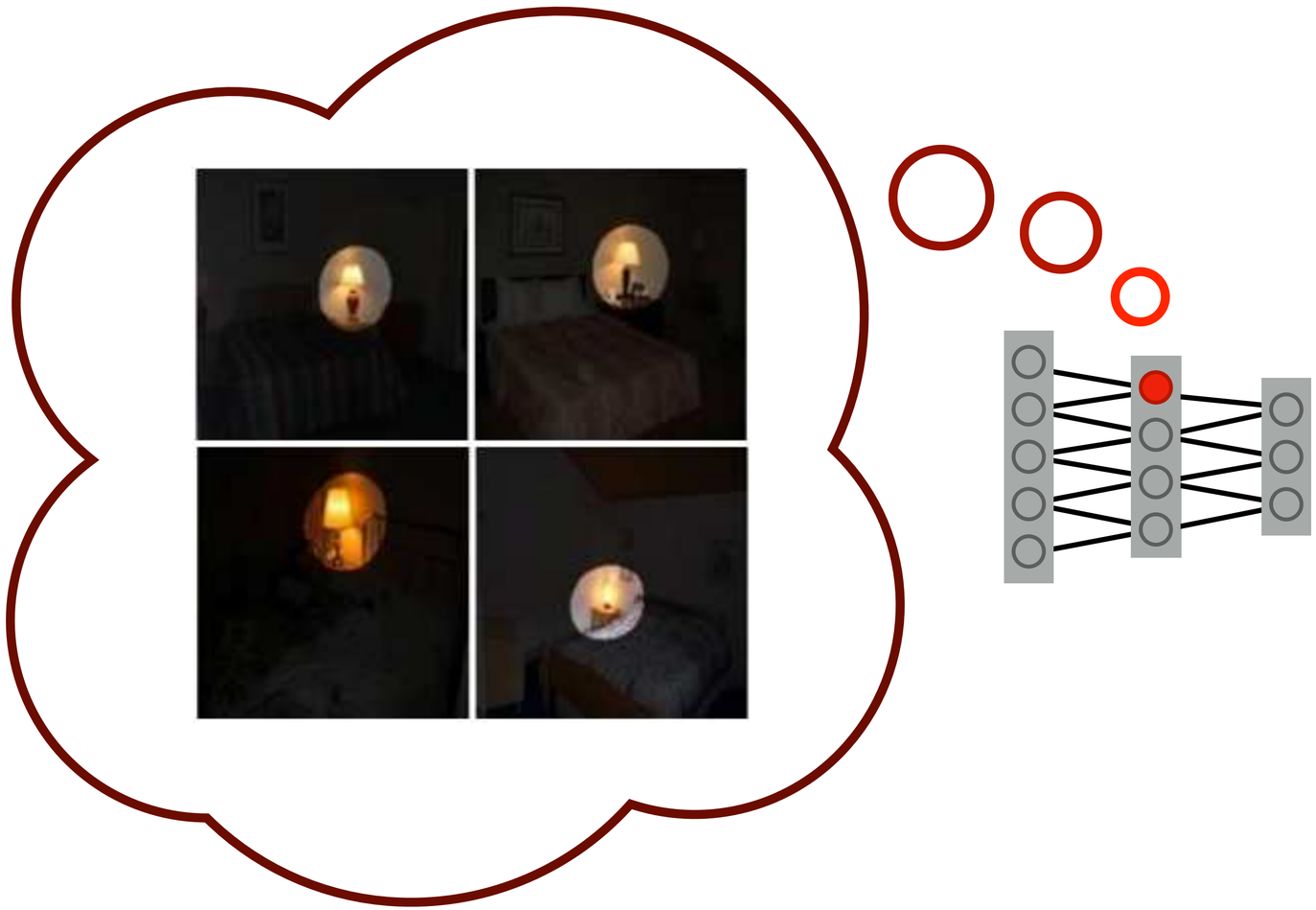


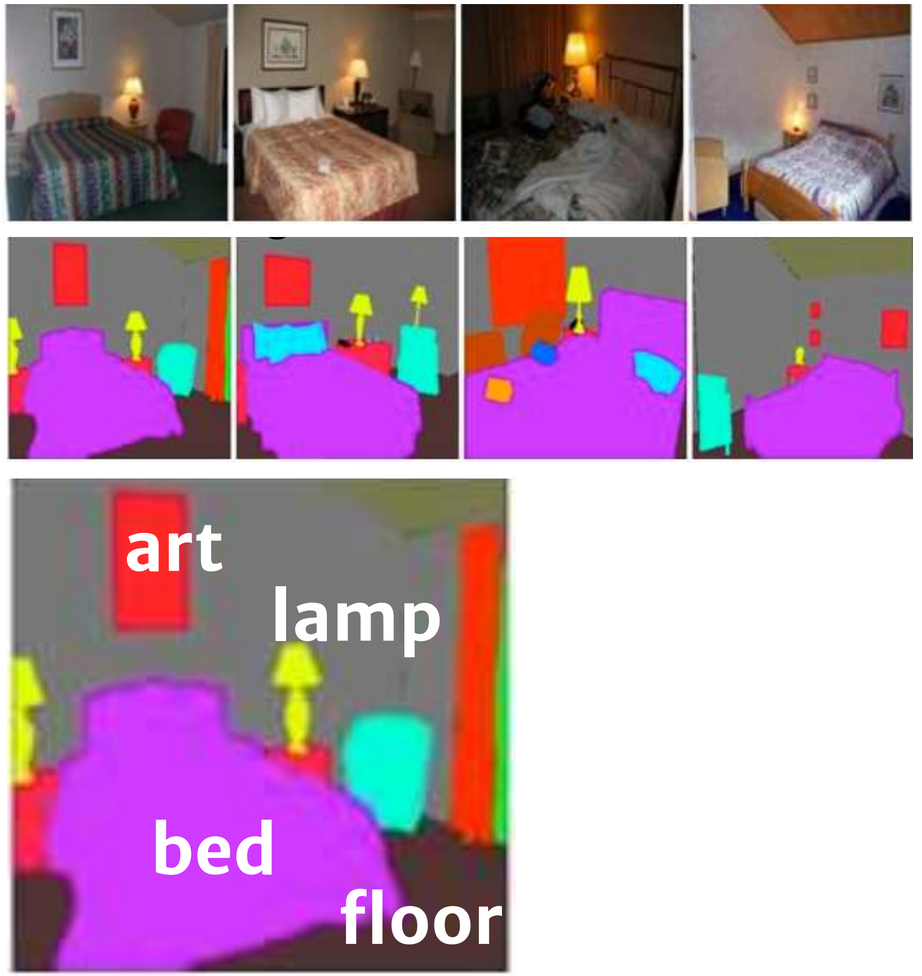
Explaining individual neurons



[Bau et al. 16. Network Dissection.]

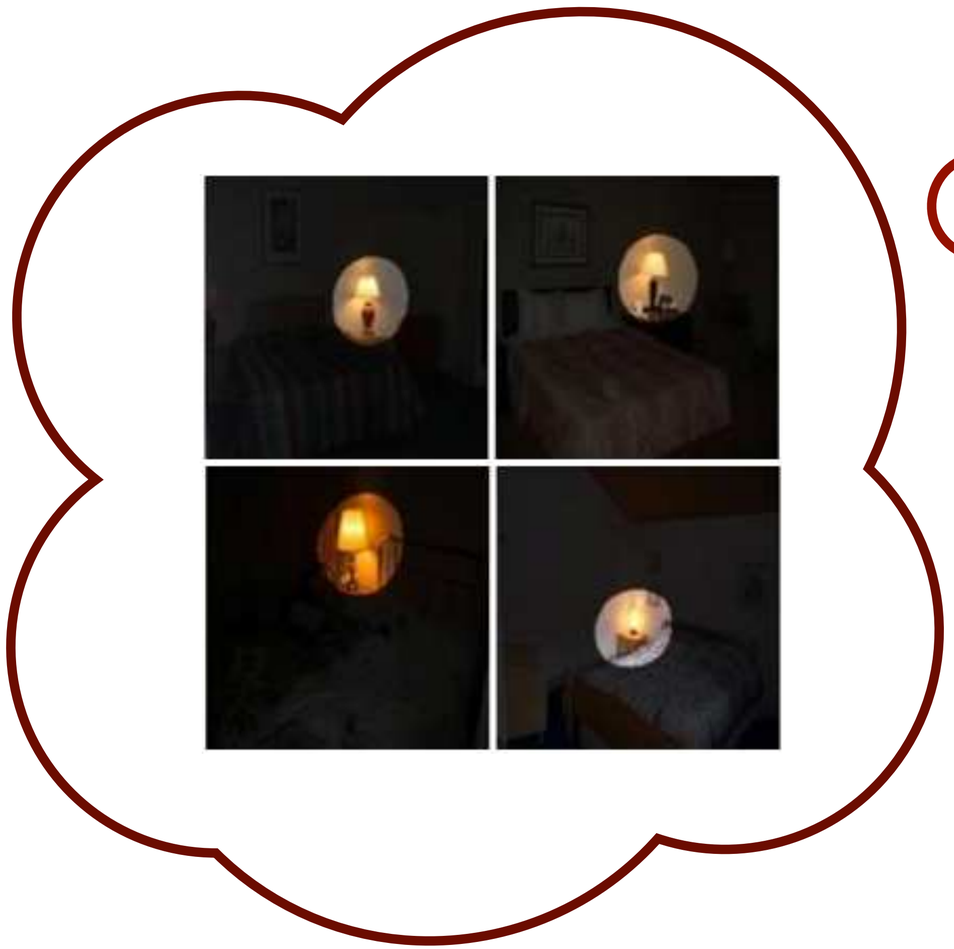
Labeling neurons with categories

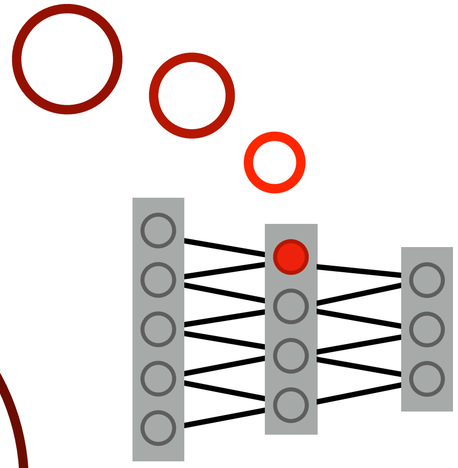




[Bau et al. 16. Network Dissection.]

Labeling neurons with categories



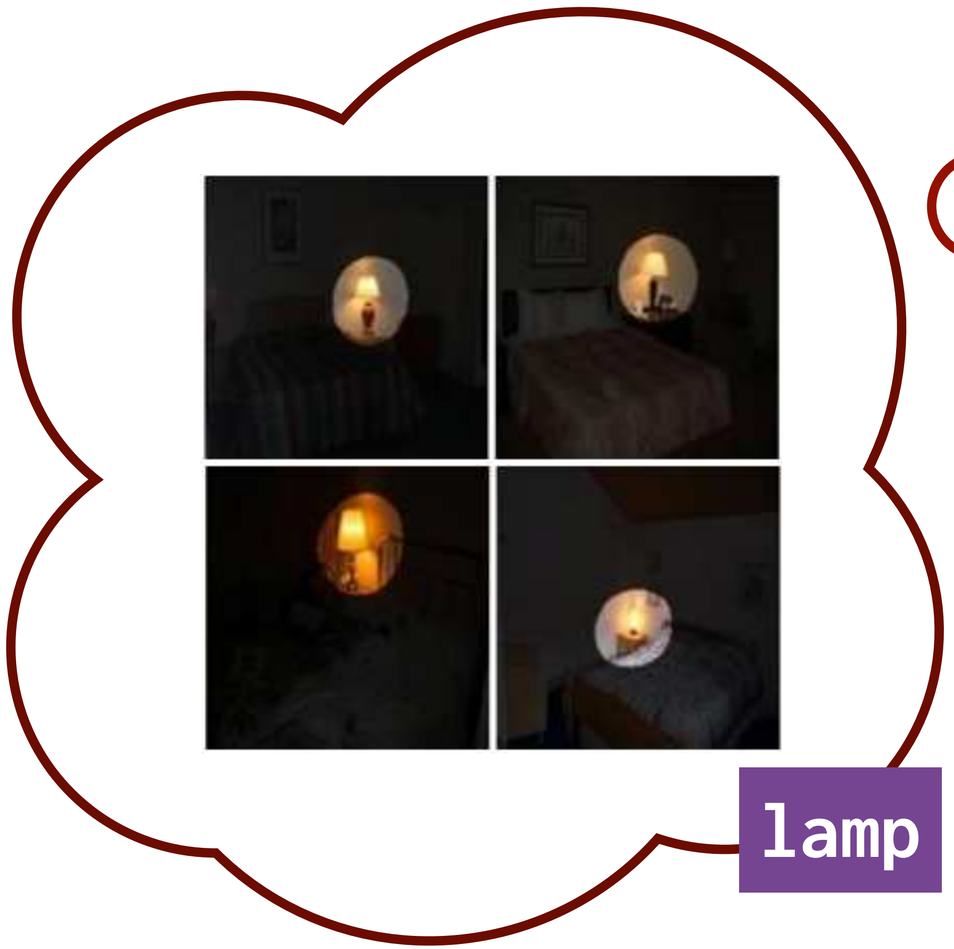


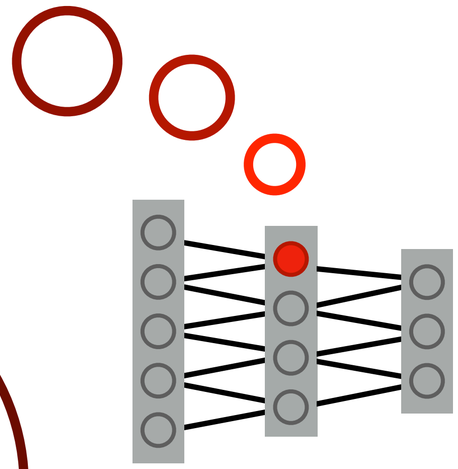


$$\text{argmax}_{a} \; \text{sim}(\cdot \rightarrow \, , \; \cdot \rightarrow a \;)$$

[Bau et al. 16. Network Dissection.]

Labeling neurons with categories



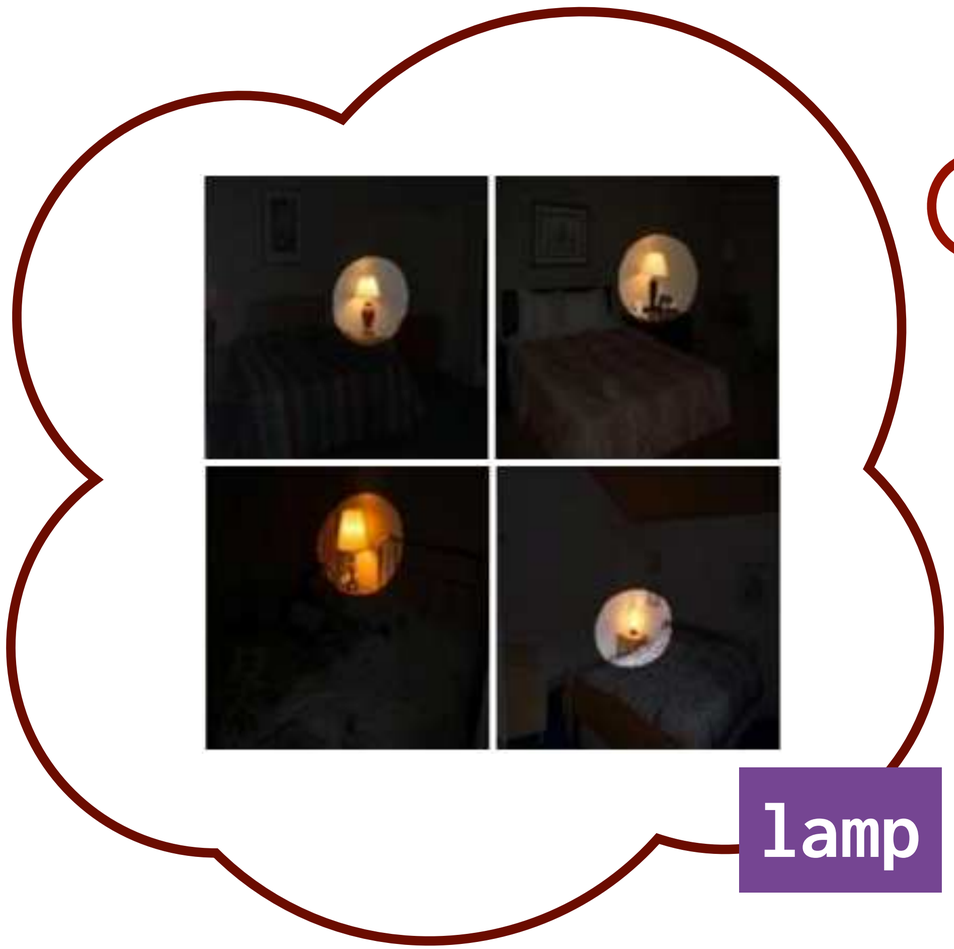


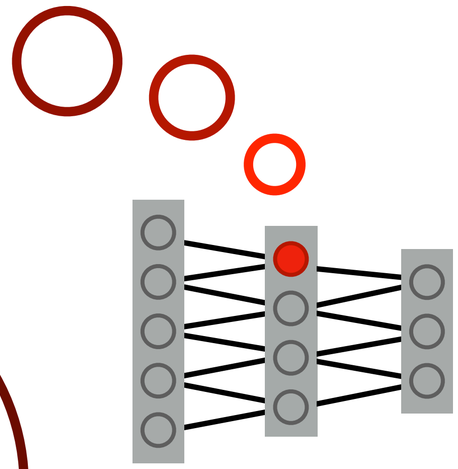


argmax$_a$ sim(· → , · → lamp)

[Bau et al. 16. Network Dissection.]

Neurons as “words” of neuralese



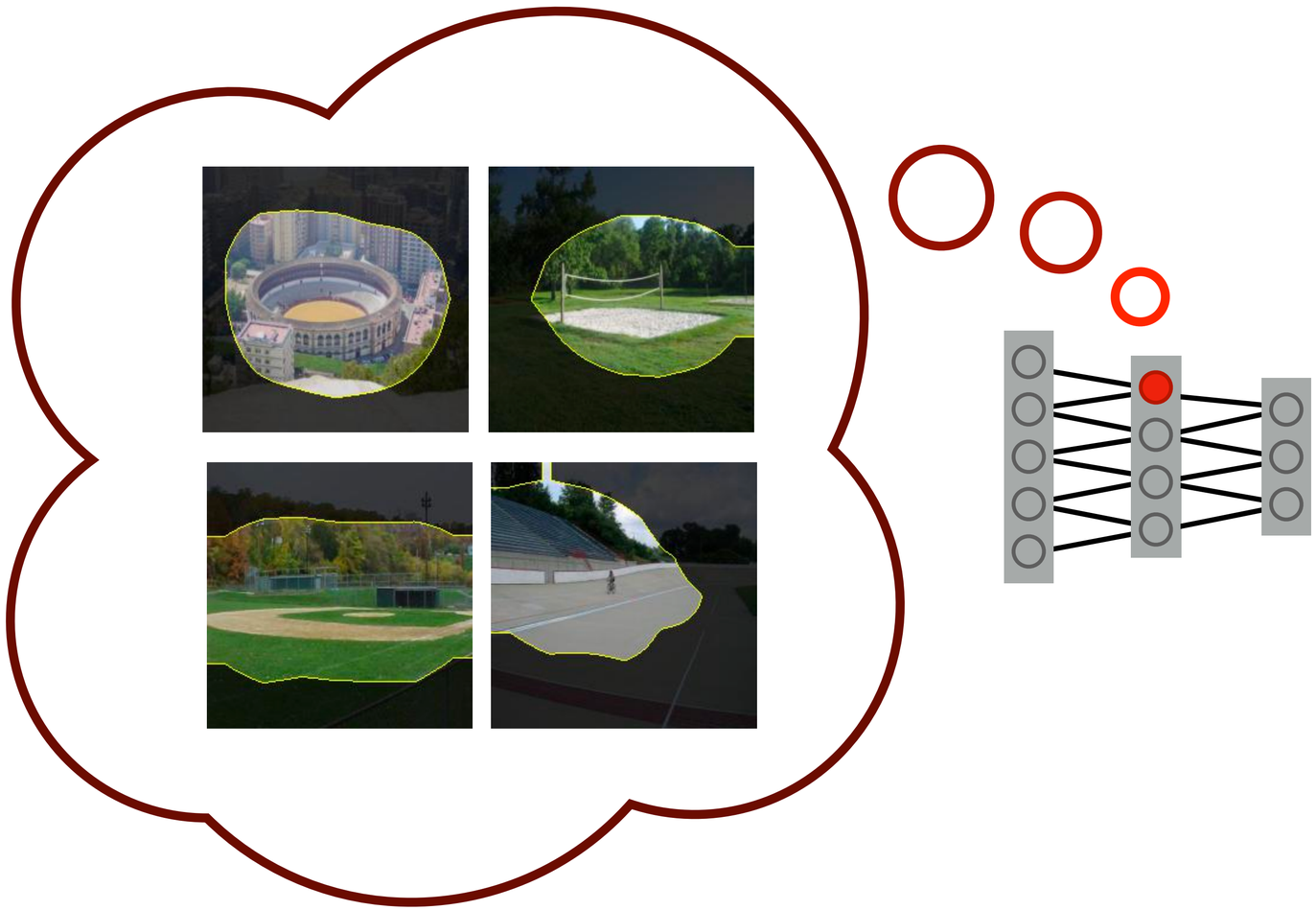




$$\text{argmin}_{a} \; \delta\big(p(\cdot \mid \text{[neuron]}),\, p(\cdot \mid \texttt{lamp})\big)$$

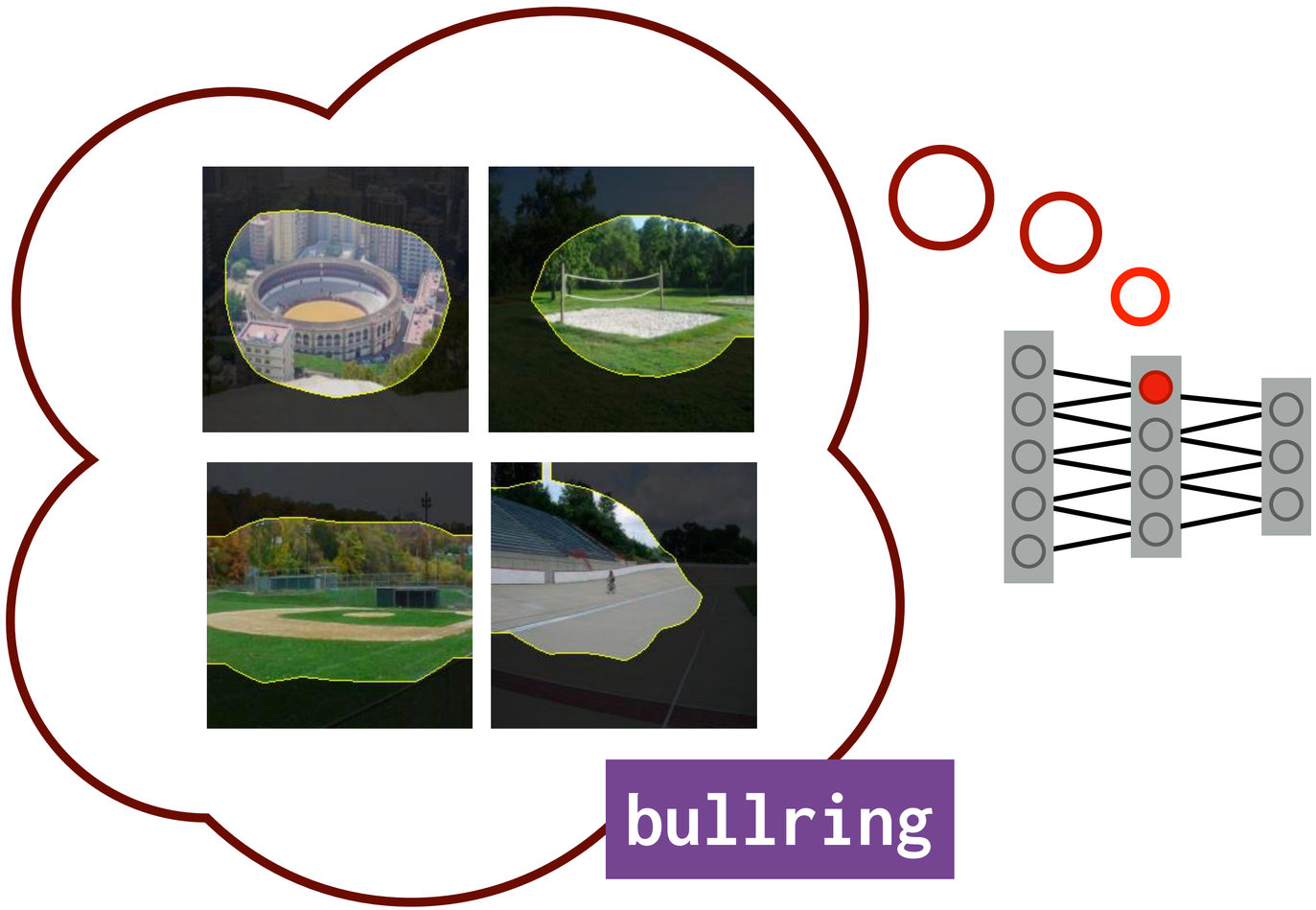
[Bau et al. 16. Network Dissection.]

Are categories enough?



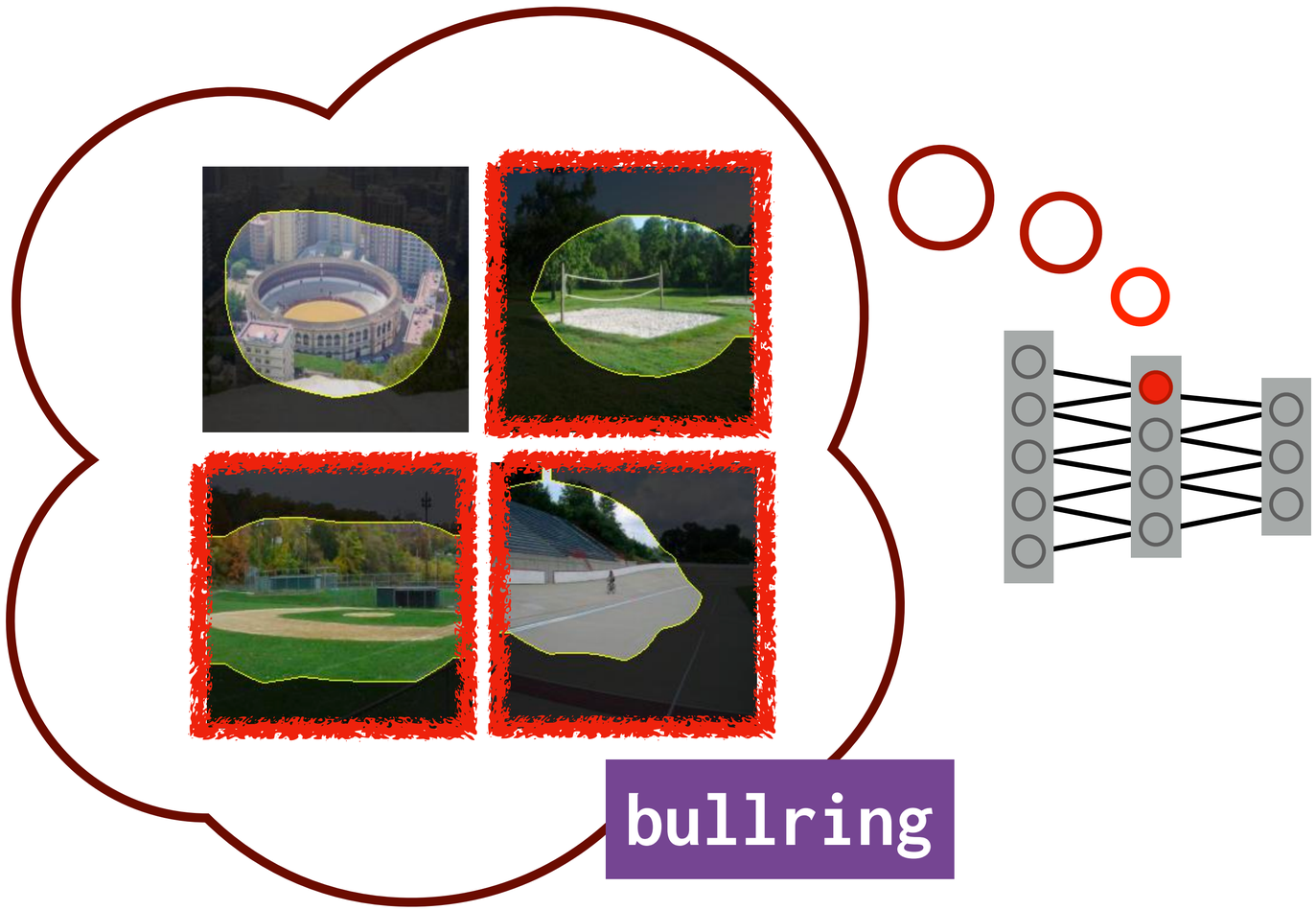
$$\operatorname{argmin}_{a \in \{\text{arena, bank, } \ldots\}} \delta(p(\cdot \mid \text{image}), p(\cdot \mid a))$$

Are categories enough?



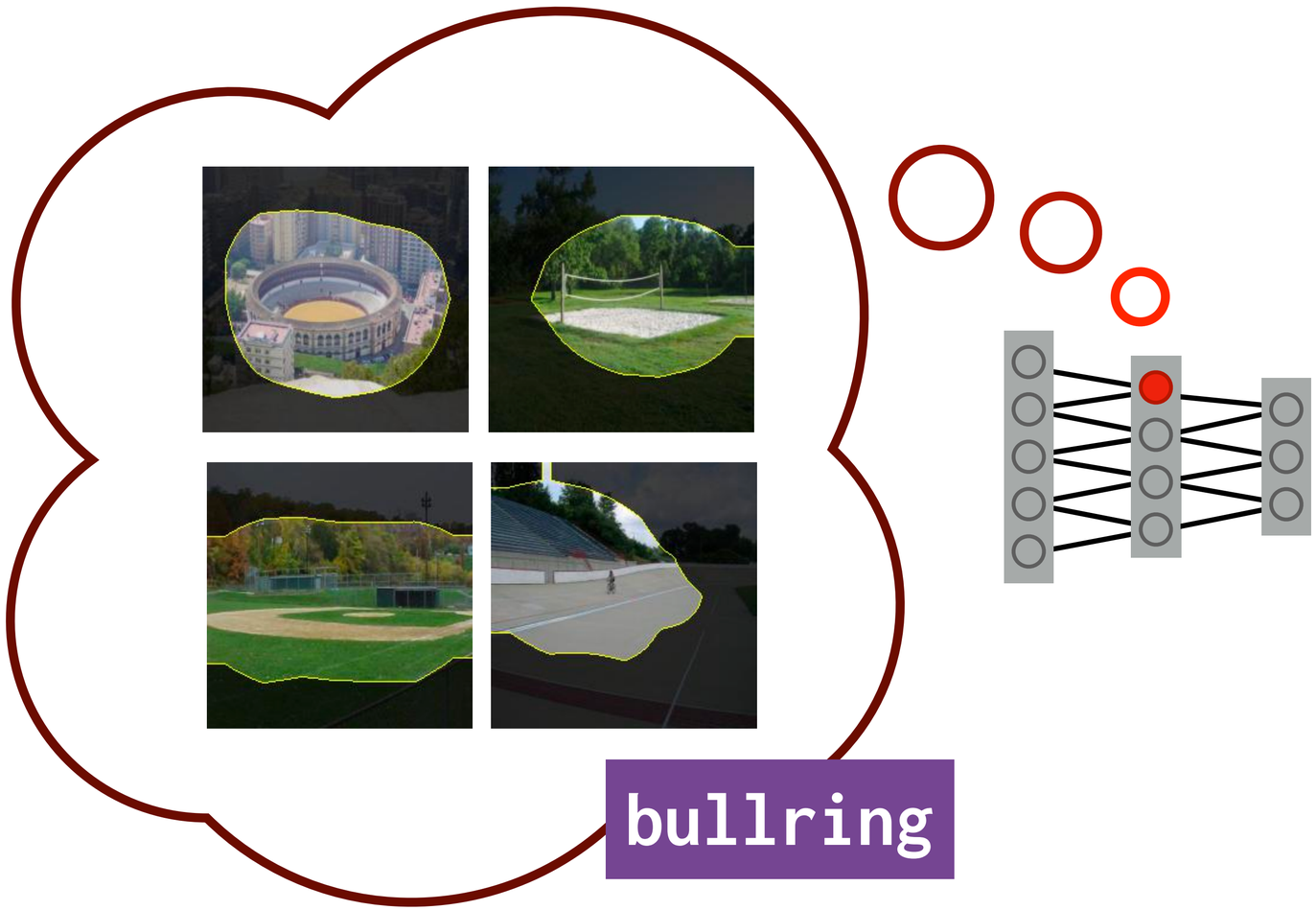
$$\text{argmin}_{a \in \{\text{arena, bank, } \ldots\}} \delta(p(\cdot \mid \text{image}), p(\cdot \mid a))$$

Are categories enough?



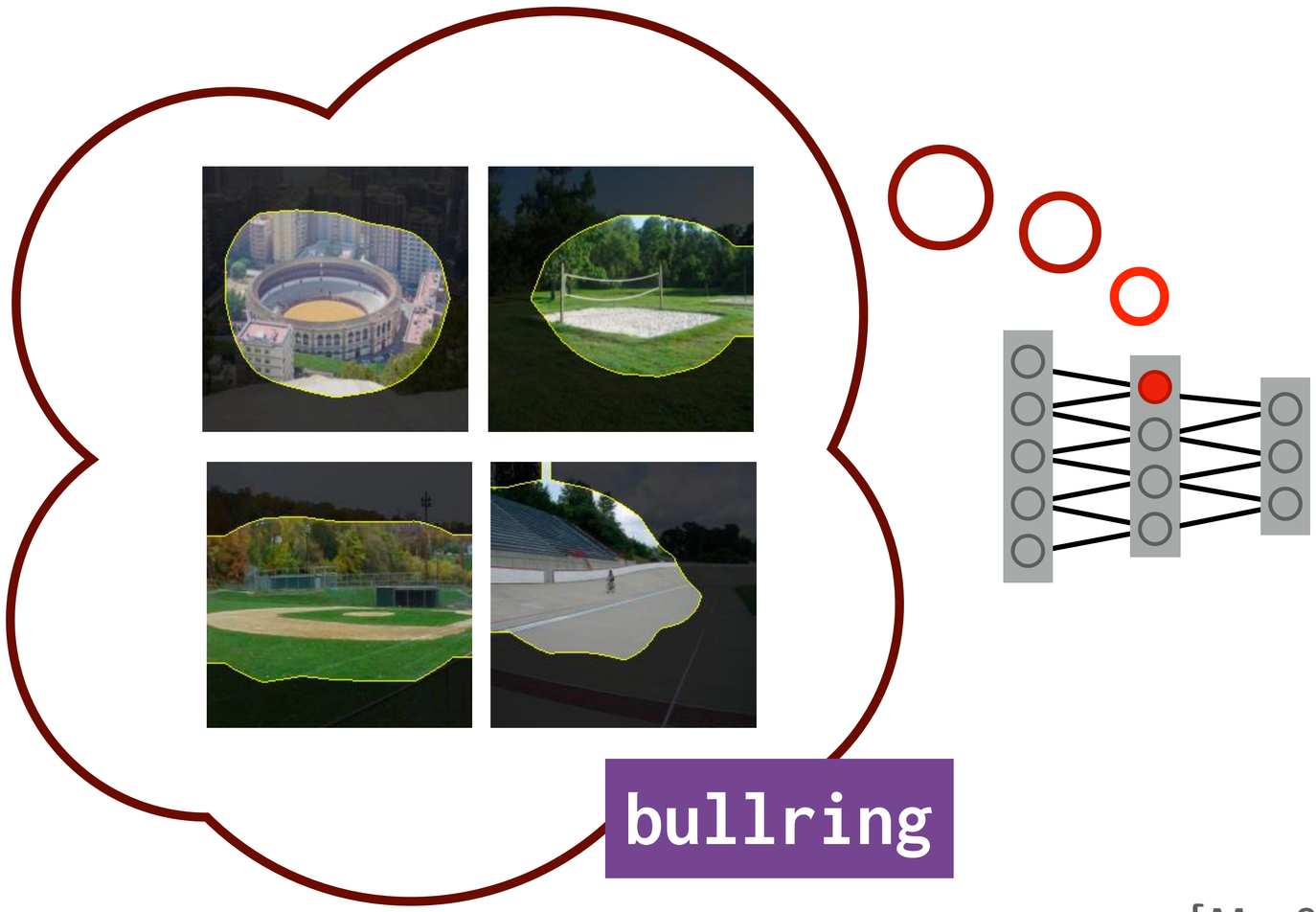
$$\text{argmin}_{a \in \{\text{arena, bank, } \ldots\}} \delta(p(\cdot \mid \text{image}), p(\cdot \mid a))$$

Are categories enough?



argmin $_{a \in \{\text{arena, bank, } \ldots\}}$ $\delta(p(\cdot \mid \text{traffic light}), p(\cdot \mid a))$

From categories to logical forms



argmin $_{a \in}$ {arena,
bank,
arena or bank,
arena and bank,
not bank,
arena and not…}

$\delta(p(\cdot \mid$ $), p(\cdot \mid a))$

[Mu & A 20. Compositional Explanations of Neurons.]

From categories to logical forms



(a) inputs $\mathbf{x}$



(b) neuron $f_{483}(\mathbf{x})$



(c) neuron masks $M_{483}(\mathbf{x})$

[Mu & A 20. Compositional Explanations of Neurons.]

From categories to logical forms



(a) inputs $\mathbf{x}$



(b) neuron $f_{483}(\mathbf{x})$



(c) neuron masks $M_{483}(\mathbf{x})$

`water` IoU .14



`river` IoU .08



`blue` IoU .006



(d) concepts $C(\mathbf{x})$

[Mu & A 20. Compositional Explanations of Neurons.]

From categories to logical forms



(a) inputs $\mathbf{x}$



(b) neuron $f_{483}(\mathbf{x})$



(c) neuron masks $M_{483}(\mathbf{x})$







(d) concepts $C(\mathbf{x})$







(e) logical forms $L(\mathbf{x})$

[Mu & A 20. Compositional Explanations of Neurons.]

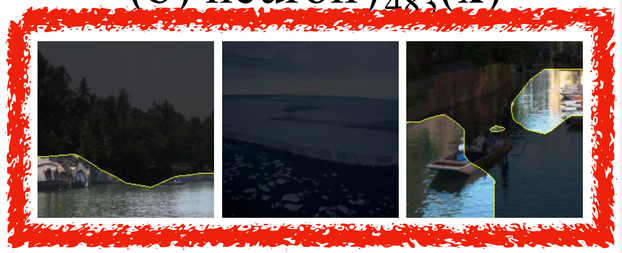
From categories to logical forms



(a) inputs $\mathbf{x}$



(b) neuron $f_{483}(\mathbf{x})$



(c) neuron masks $M_{483}(\mathbf{x})$

`water` IoU .14



`river` IoU .08



`blue` IoU .006



(d) concepts $C(\mathbf{x})$

`water` **OR** `river` IoU .15

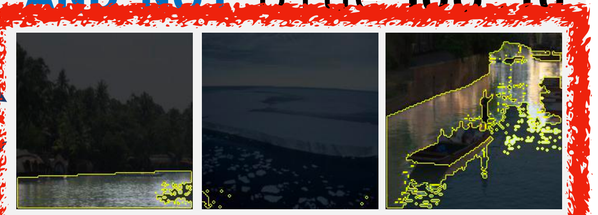


NOT `blue` IoU .004



(e) logical forms $L(\mathbf{x})$

(`water` **OR** `river`)
AND NOT `blue` **IoU .16**



[Mu & A 20. Compositional Explanations of Neurons.]

Compositional explanations

Unit 106 `bullring`



Compositional explanations

Unit 106 `bullring OR pitch OR volleyball court`

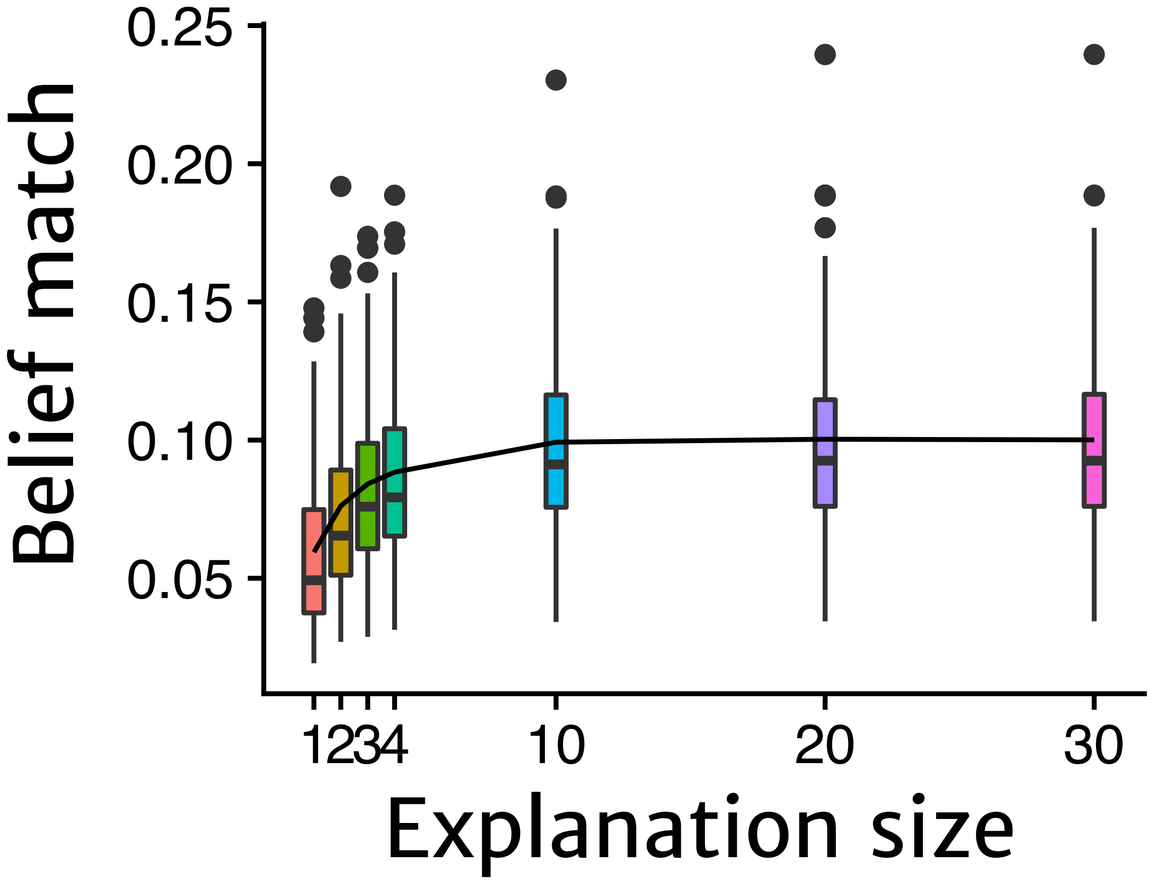


Compositional explanations

Unit 106 `bullring OR pitch OR volleyball court AND (NOT football field) AND (NOT railing)`



Explanatory power



Semantic neurons

Unit 192 `skyscraper OR lighthouse OR water tower`

IoU 0.06



Unit 310 `sink OR bathtub OR toilet`

IoU 0.16



Perceptual neurons

Unit 321 `ball pit OR orchard OR bounce game`

IoU 0.12



Unit 102 `cradle OR autobus OR fire escape`

IoU 0.12



Specialized neurons

Unit 483 `(water OR river) AND NOT blue`

IoU 0.13



Unit 432 `attic AND (NOT floor) AND (NOT bed)`

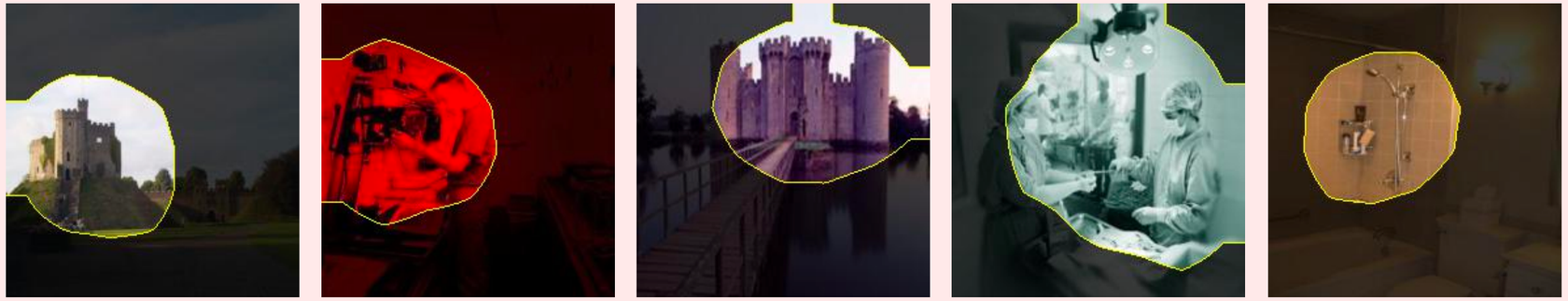
IoU 0.15



Overloaded neurons

Unit 314 `operating room OR castle OR bathroom`

IoU 0.05

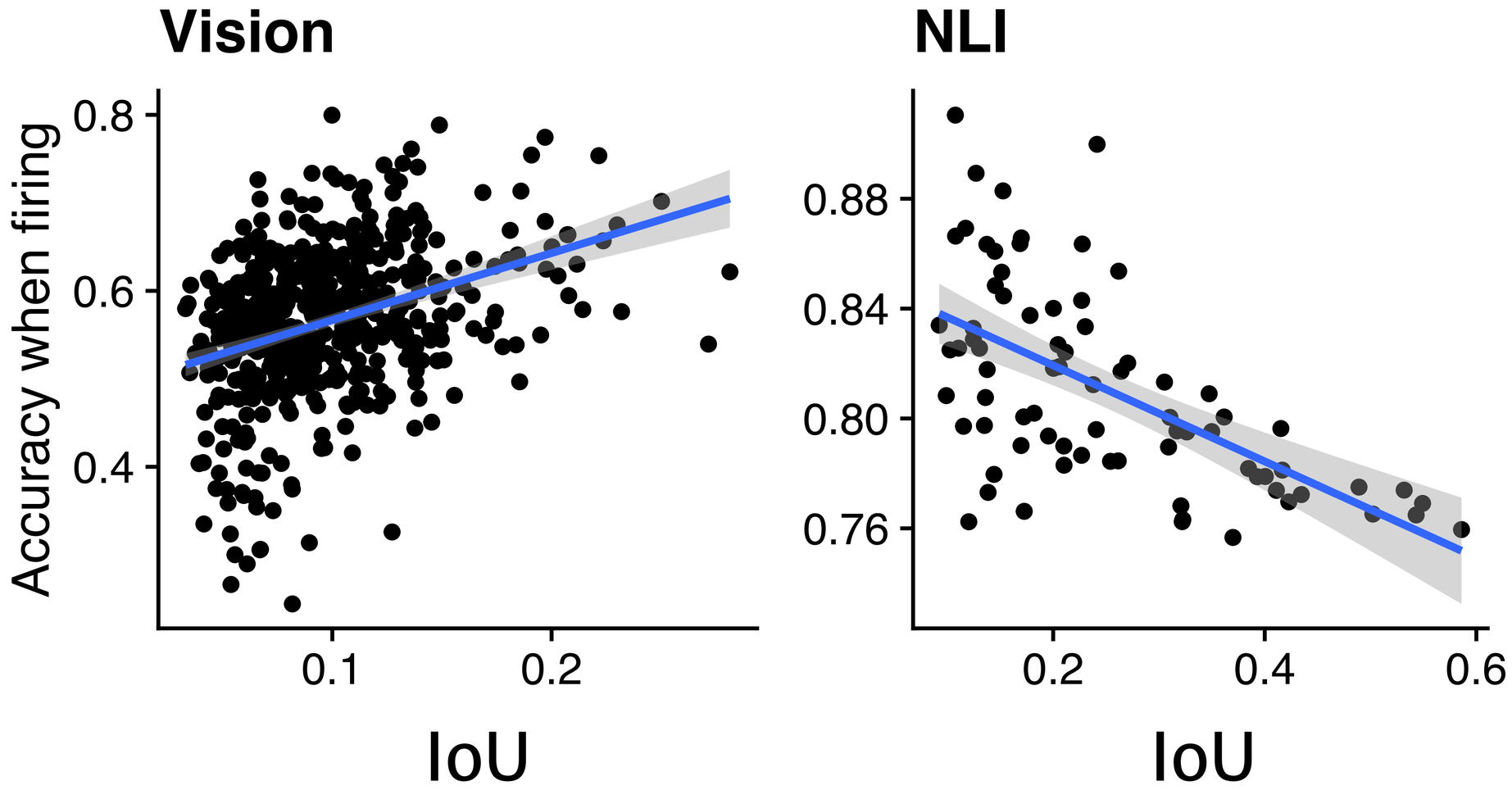


Unit 439 `bakery OR bank vault OR shopfront`

IoU 0.08



Explainability and model accuracy



Natural language inference

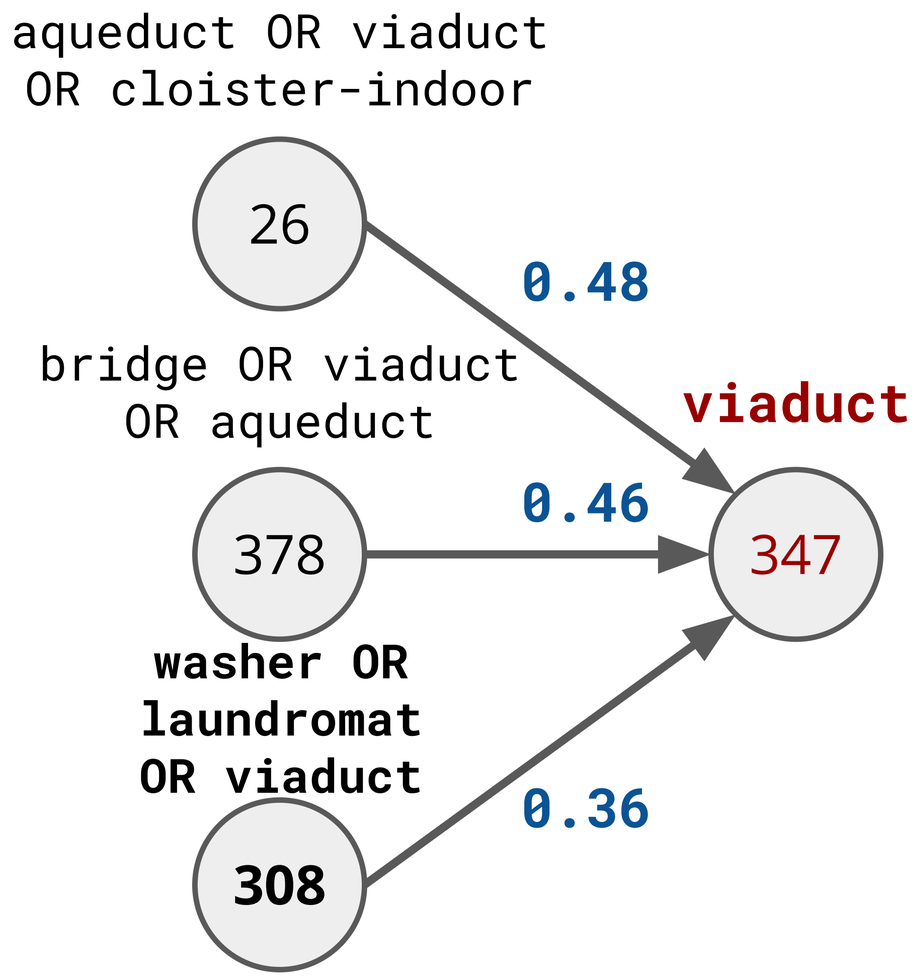
`Hypothesis DOES NOT contain man AND premise contains man`

IoU **0.123** w_{entail} -0.046 $w_{neutral}$ -0.021 w_{contra} 0.040

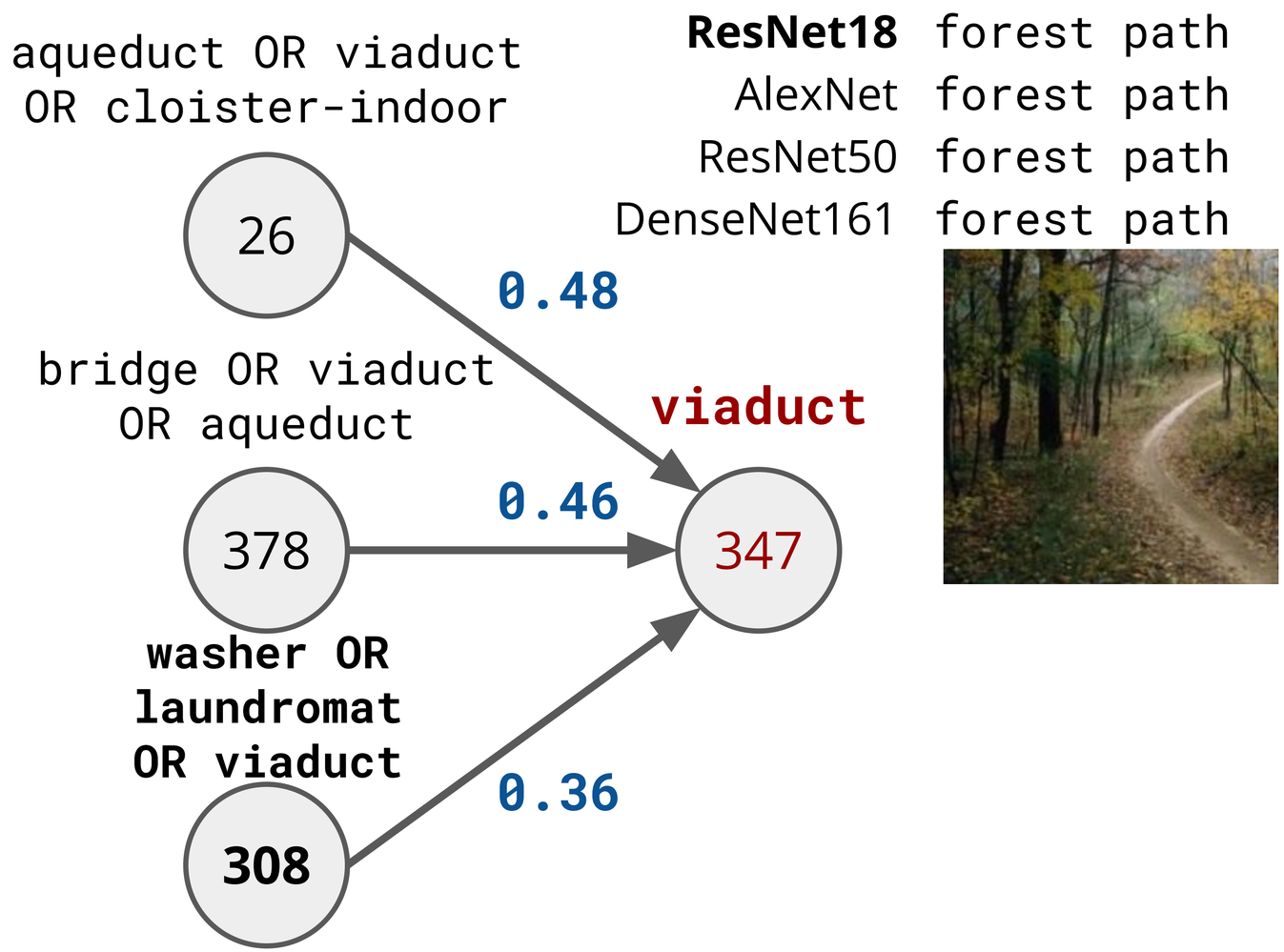
Pre A guy pointing at a giant blackberry.
Hyp A woman tearing down a giant display.
Act **29.31** True **contra** Pred **contra**

Pre A man in a hat is working with…flowers.
Hyp Women are working with flowers.
Act **27.64** True **contra** Pred **contra**

Adversarial examples

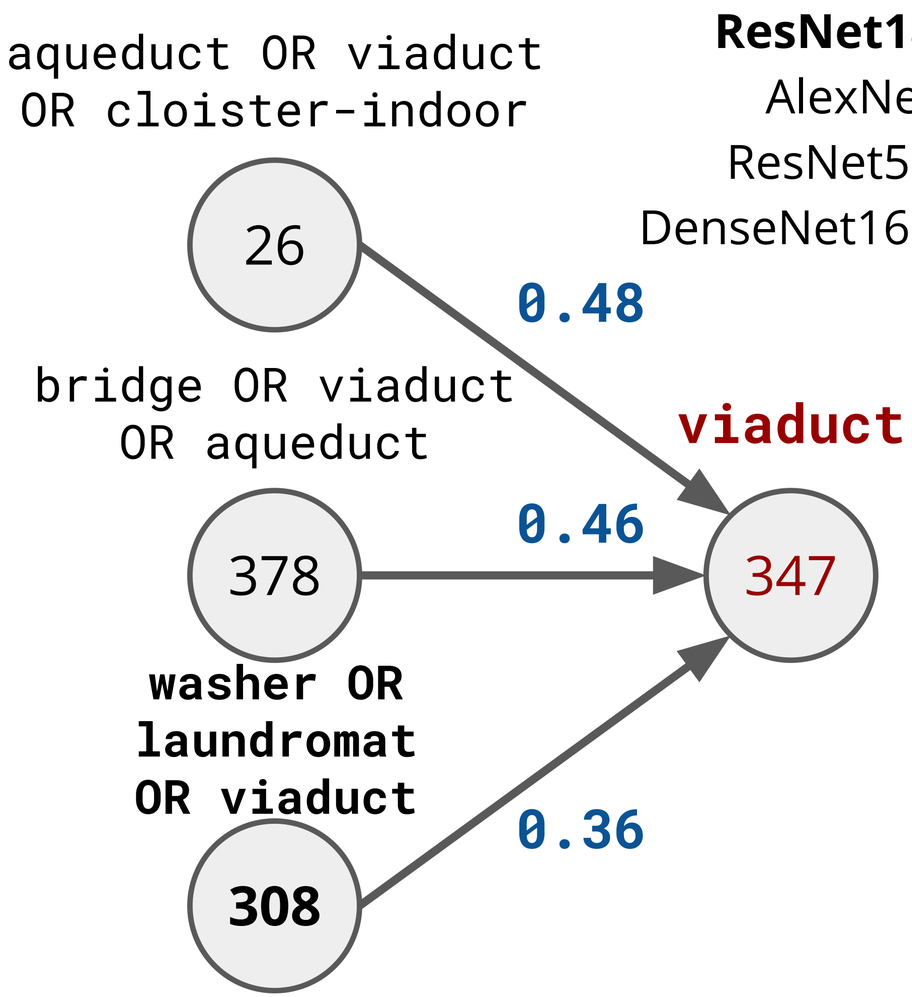


Adversarial examples



ResNet18 forest path
AlexNet forest path
ResNet50 forest path
DenseNet161 forest path

Adversarial examples



ResNet18	forest path	**viaduct**
AlexNet	forest path	**viaduct**
ResNet50	forest path	**viaduct**
DenseNet161	forest path	laundromat



Summary

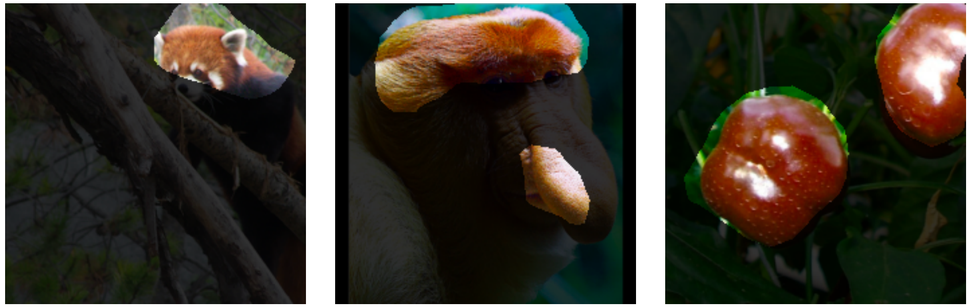
1. Represent “meanings” of learned representations and features as distributions over model inputs, and generate explanations by summarizing these distributions with language.

2. Compositionality matters! Lots of interpretable behaviors can’t be explained with a fixed vocabulary of primitives.

Are logical forms enough?



sides of cylinders



brown animals and foods



bird legs and beaks



high-contrast diagonals

Natural language neuron descriptions



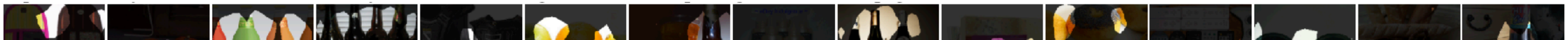
1. This is the side of containers. [A1WS884SI0SLO4]
2. The space on the sides of cylinders [A2CFSXJDV50DVF]
3. outer edges of cylindrical shapes [A9V0V2HY30WNG]



1. Human and animal faces and bodies, and vegetables. [ACJ33ZICZI18Q]
2. about animals eye [AH6W2VYAEZNM8]
3. These are animal heads. [A1WS884SI0SLO4]
4. They contain pointed features like noses or fruit stems. [A1DP4BJKMEF6LS]



1. Grass and a flower [ACJ33ZICZI18Q]
2. areas of vegetation in a background [A2U1NN8QDH1YVC]
3. These are spots of grass. [A1WS884SI0SLO4]
4. about animal [AH6W2VYAEZNM8]



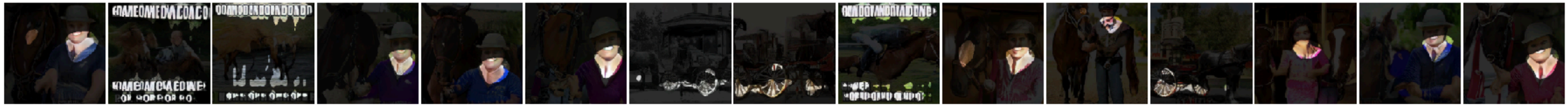
Natural language neuron descriptions



1. Human and horse faces are shown. [A1QAQ72HQ921UG]
2. Grass, the side of a horse head, and the human face beside a horse. [A2XVDB1OXWGTN9]
3. Faces from horses and humans, also leather belts from halters [AX7K5UODLEK72]



1. blue, red and green parts in clothing and objects [AX7K5UODLEK72]
2. This is a saddle. [A1WS884SI0SLO4]
3. They are colored basic shapes. [A1UOGYZFJF4BKW]



1. V shaped objects and parts of clothes [AX7K5UODLEK72]
2. These are chins. [A1WS884SI0SLO4]
3. Chins texts and wheels are shown. [A1QAQ72HQ921UG]



Language-based control of GANs



increase activation of neurons
whose description contains *jockey*

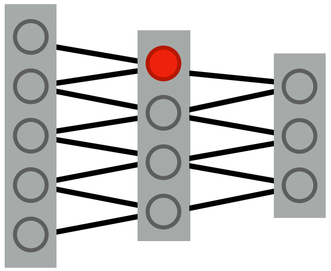
Language-based control of GANs

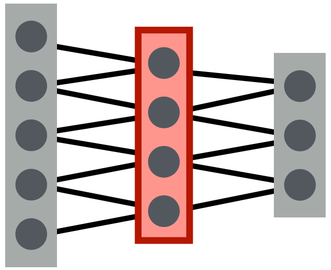


increase activation of neurons
whose description contains *brown*

Conclusion

Explain everything!





$$\operatorname{argmin}_{a \in \text{English}} \delta(p(x|\,\text{[neurons]}), p(x|\,a))$$

Explain everything?

Green and yellow animals, a yellow smiley face, and a firefighter's head and jacket.
The heads of animals
about animals
foods in the packet
Dog and fox eyes.
This is a animal head.
Body part of the birds
The bottom portions of faces of animals.
Dog heads with black and white.
Anything that has the color blue in it.
These are animal heads.
These are animal heads.
about animals
This is showing both parts of animals and parts of wheels.
The color white
These are diagonal lines.
Shifting contrast colors, either light-dark-light or dark-light-dark.
Doors, windows, and see-through objects.
Turtle shells and regular overlapping patterns are shown.

Red clothing, vehicles, plants, and objects.
Human skin
The black areas are highlighted in the images.
The images show body coverings of animals including fur, feathers, hair, and claws.
It shows an image that has a bit of white in it.
These are flowers and animal fur.
This regions is that is being highlighted are spots.
eyes and mouth
face of all animals
This is fruit and other circles.
Face of dogs
Dog faces and bodies.
The face area is highlighted in the images.
eyes and beak
People's faces and other body parts.
Green grass, plants, and objects.
This is black and white grids.
yellow spots surrounded by uniform colors
body of the dogs
arches
all the above are in green color
Honeycomb, grid, and keyboard button patterns.

This is text.
The regions depict lin
from a center.
They are the west or
objects that contain
This is the area above
They are brownish fu
Texts and blue or yell
This is very natural a
side angle part of all
letters of the all imag
Objects with curved e
Core hours are set th
be outside in the offi
hours is flexible.
All images are made
They are circular obje
Blue colored objects.
Animal skins are in th
The shiny white part
They are the midsect
Eyes of various anima
various objects and a

The next challenge: *pragmatic* explanation

Green and yellow animals, a yellow smiley face, and a firefighter's head and jacket.
The heads of animals
about animals
foods in the packet
Dog and fox eyes.
This is a animal head.
Body part of the birds
The bottom portions of faces of animals.
Dog heads with black and white.
Anything that has the color blue in it.
These are animal heads.
These are animal heads.
about animals
This is showing both parts of animals and parts of wheels.
The color white
These are diagonal lines.
Shifting contrast colors, either light-dark-light or dark-light-dark.
Doors, windows, and see-through objects.
Turtle shells and regular overlapping patterns are shown.

Red clothing, vehicles, plants, and objects.
Human skin
The black areas are highlighted in the images.
The images show body coverings of animals including fur, feathers, hair, and claws.
It shows an image that has a bit of white in it.
These are flowers and animal fur.
This regions is that is being highlighted are spots.
eyes and mouth
face of all animals
This is fruit and other circles.
Face of dogs
Dog faces and bodies.
The face area is highlighted in the images.
eyes and beak
People's faces and other body parts.
Green grass, plants, and objects.
This is black and white grids.
yellow spots surrounded by uniform colors
body of the dogs
arches
all the above are in green color
Honeycomb, grid, and keyboard button patterns.

This is text.
The regions depict lin
from a center.
They are the west or
objects that contain
This is the area above
They are brownish fu
Texts and blue or yell
This is very natural a
side angle part of all
letters of the all imag
Objects with curved e
Core hours are set th
be outside in the offi
hours is flexible.
All images are made
They are circular obje
Blue colored objects.
Animal skins are in th
The shiny white part
They are the midsect
Eyes of various anima
various objects and a

The next challenge: *pragmatic* explanation

Green and yellow animals, a yellow smiley face, and a firefighter's head and jacket.
The heads of animals
about animals
foods in the packet
Dog and fox eyes.
This is a animal head.
Body part of the birds
The bottom portions of faces of animals.
Dog heads with black and white.
Anything that has the color blue in it.
These are animal heads.
These are animal heads.
about animals
This is showing both parts of animals and parts of wheels.
The color white
These are diagonal lines.
Shifting contrast colors, either light-dark-light or dark-light-dark.
Doors, windows, and see-through objects.
Shells and irregular overlapping patterns are shown.

Red clothing, vehicles, plants, and objects.
Human skin
The black areas are highlighted in the images.
The images show body coverings of animals including fur, feathers, hair, and claws.
It shows an image that has a bit of white in it.
These are flowers and animal fur
eyes and mouth
face of all animals
Face of dogs
Dog faces and bodies
eyes and beak
People's faces and other body parts.
Green grass, plants, and objects
yellow spots surrounded by uniform colors
body of the dogs
arches
all the above are in green color
Honeycomb, grid, and keyboard button patterns.

This is text.
The regions depict lin
from a center.
They are the west or
objects that contain
This is the area above
They are brownish fu
This is very natural a
side angle part of all
letters of the all imag
Objects with curved e
Core hours are set the
hours is flexible.
All images are made
They are circular obje
Blue colored objects.
Animal skins are in th
The shiny white part
They are the midsect
Eyes of various anima
various objects and a

Grice's conversational maxims:

1. Say what is true
2. Say no more than is required
3. Be relevant
4. Avoid ambiguity

[Grice 1975]

The next challenge: *pragmatic* explanation

What can we learn from language users about how to explain models effectively?

Grice's conversational maxims:

1. Say what is true
2. Say no more than is required
3. Be relevant
4. Avoid ambiguity

[Grice 1975]

Summary

1. Represent “meanings” of learned representations and features as distributions over model inputs, and generate explanations by summarizing these distributions with language.

2. Compositionality matters! Lots of interpretable behaviors can’t be explained with a fixed vocabulary of primitives.

3. “Natural” language matters: real networks are probably too complex for logical explanations alone.

Thank you!

Researchers



Evan Hernandez



Jesse Mu

Sponsors

NVIDIA