CMPT 983

Grounded Natural Language Understanding

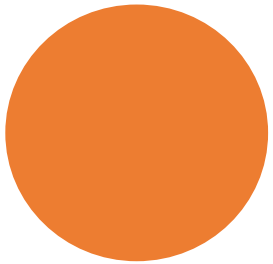
January 11, 2021

Introduction to grounding and course logistics

Today

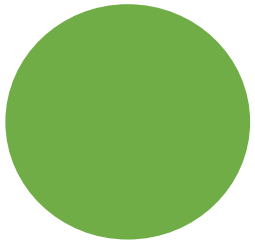
- Introductions
- What is grounding?
- Course overview and logistics
- Topics in grounded NLU

What is grounding?



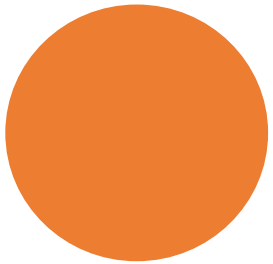
osk

vap



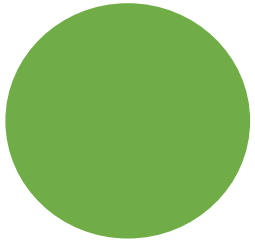
osk

vap



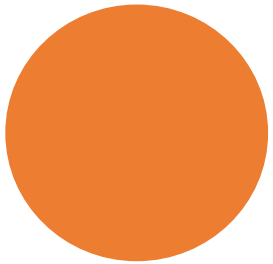
tod

be



bo

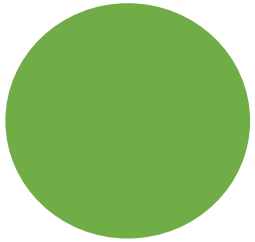
tod



osk tod

vap be

vap bo



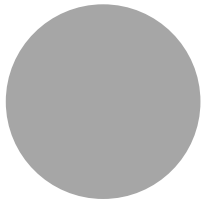
osk bo

vap tod

What can humans do?

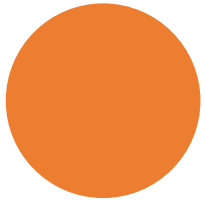
Grounding

osk



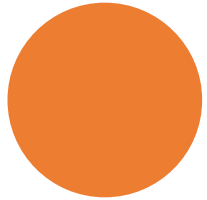
vap

tod



bo

Compositionality



osk tod

Generalization

vap bo

What is symbol grounding?

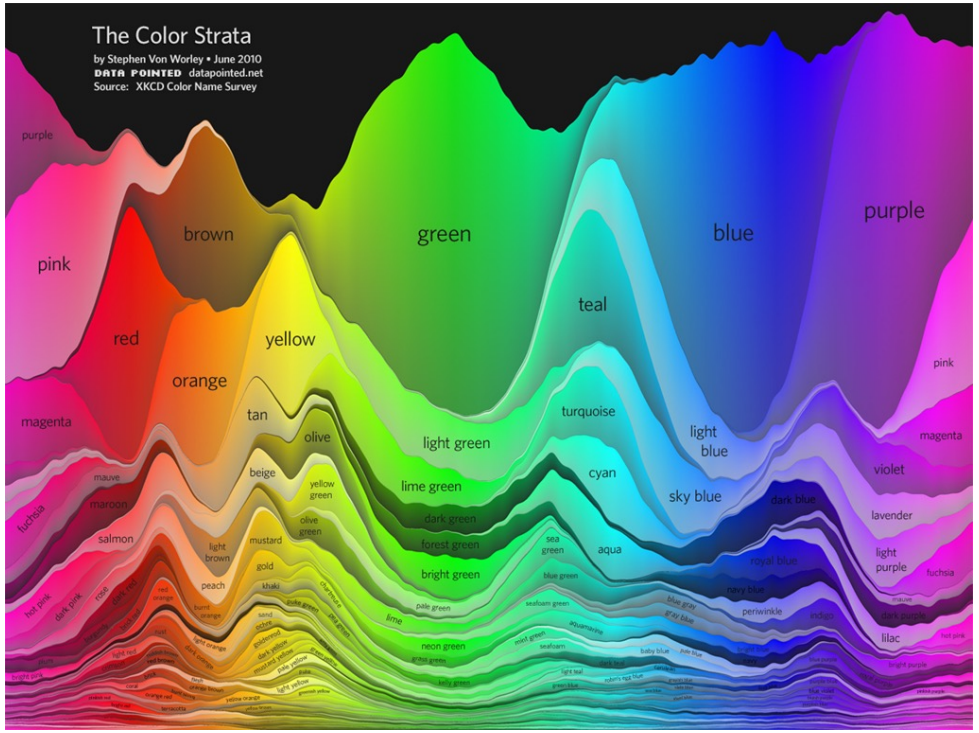
- Connecting **linguistic symbols** to **perceptual** experiences and **actions**

- **Connecting words and sentences to their meaning**

Types of grounding

Perceptual

- Visual: *green* = [0,1,0] in RGB
- Auditory: *loud* = >120 dB
- Taste: *sweet* = >some threshold level of sensation on taste buds
- Touch: *pain*, *cold*, *soft*



Types of grounding: high-level concepts

Things (objects)

cat

dog

Actions

running

eating

Types of grounding

Temporal

- *winter, summer*
- *late evening* = after 6pm
- *fast, slow* = describing rates of change

Spatial

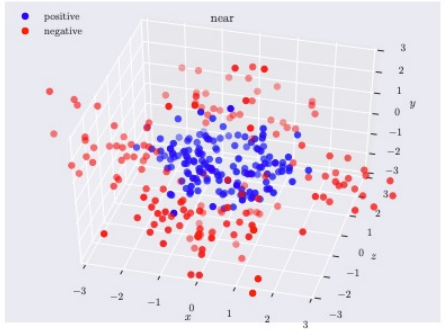
- *Vancouver*
- *north, south*
- *left, on top of, in front of*

relations

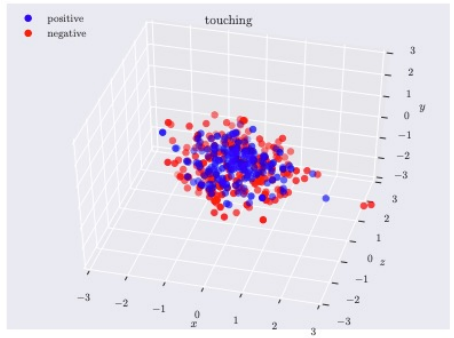
Types of grounding

Relations

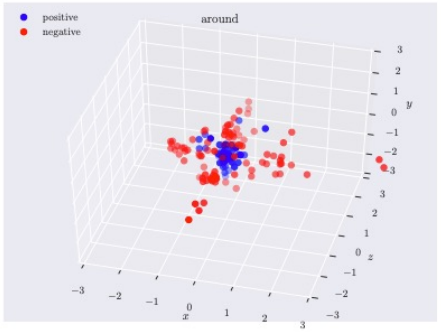
- Spatial
  - *left, on top of, in front of*
- Functional
  - Jacket *keeps* people warm
  - Mug *holds* water
- Size
  - Whales are *larger* than lions



(n) near



(ab) touching



(b) around

"Rel3D: A Minimally Contrastive Benchmark for Grounding Spatial Relations in 3D"
[Goyal et al, NeurIPS 2020]

Types of grounding

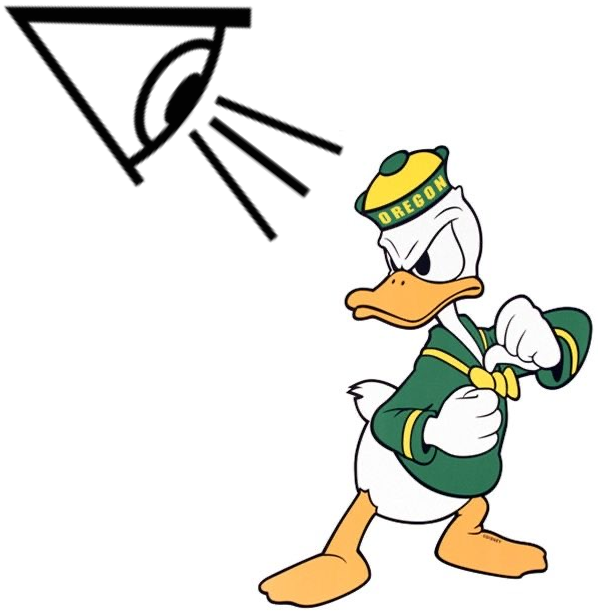
Compositional

- *Dog reading newspaper*

- *Climb on chair to turn on lamp* (VP)

Ambiguity in grounding

I saw her duck.



Slide credit: Stefan Lee

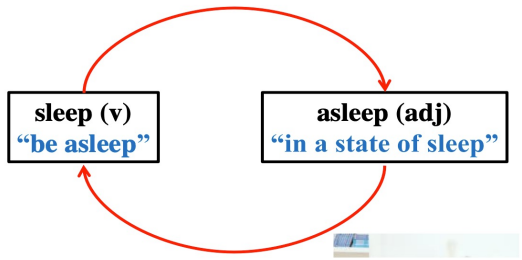
Choices in what to ground to

Connecting linguistic symbols to

- perceptual experiences and actions
- other symbols
- to executable programs

``Sleep" means ``be asleep"

Circular definitions



sleep(n): ``a natural and periodic state of rest during which consciousness of the world is suspended"

Create a key `key` if it does not exist in dict `dic` and append element `value` to value

```
dic.setdefault(key, []).append(value)
```

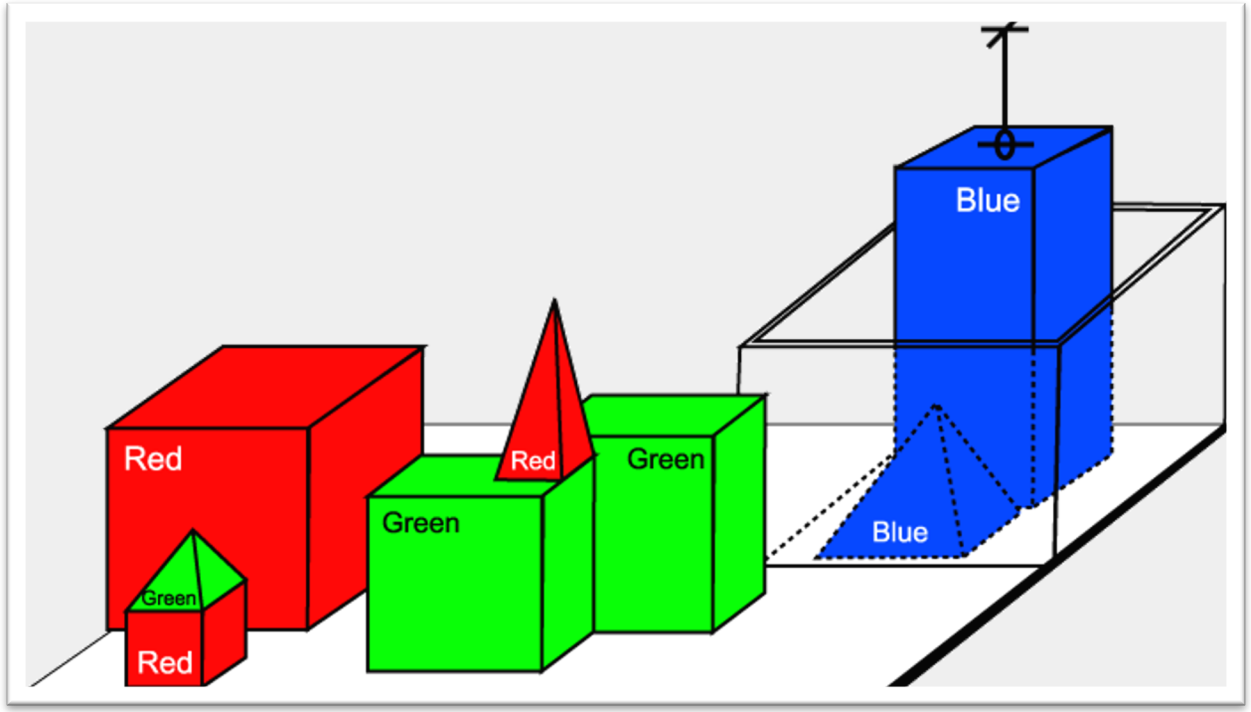
Topics in grounded NLU

SHRDLU (Winograd, 1968)

Video of actual system:

https://www.youtube.com/watch?v=bo4RvYJYOzI

Person: Pick up a big red block.
Computer: OK.
Person: Grasp the pyramid.
Computer: I don't understand which pyramid you mean.
Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.
Computer: By "it", I assume you mean the block which is taller than the one I am holding.
Computer: OK.
Person: What does the box contain?
Computer: The blue pyramid and the blue block.
Person: What is the pyramid supported by?
Computer: The box.



Topics

- Representation
  - Embeddings
  - Structured representations
- Concepts
  - Compositionality
  - Speaker-listener models
- Learning
  - Generalization
  - Fully supervised vs weakly supervised
  - Embodied setting
  - Interactive / Incremental learning

Tasks

- Translation: Captioning, text to X generation
- Alignment: Reference resolution
- Question Answering: VQA, EQA
- Instruction following
- Dialogue

Representations

Representations

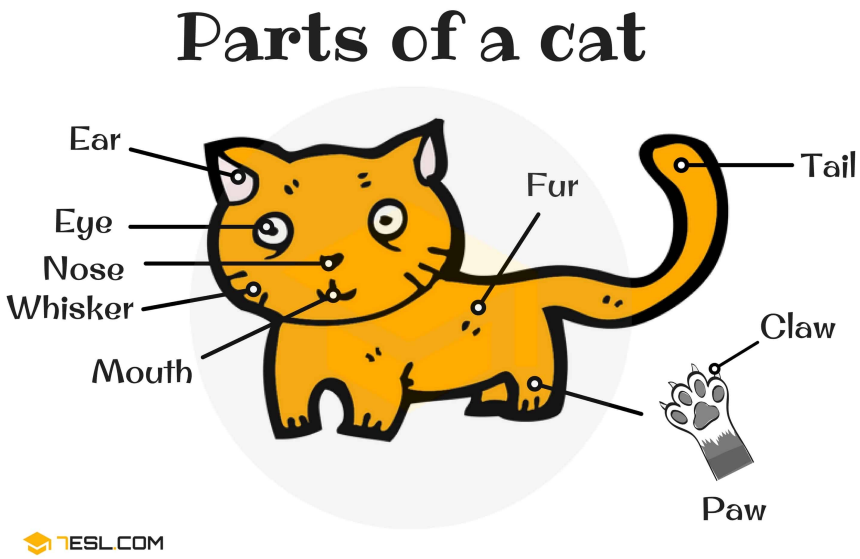
How to represent the meaning of something?

“cat”

cat: a small domesticated carnivore, *Felis domestica* or *F. catus,* bred in a number of varieties.

cat → {
  isMammal: true
  hasFur: true
  hasLegs: true
  meows: true
  barks: false
  height: 9.1 – 9.8 in
  weight: 7.9 – 9.9 lbs
  …
}

Attributed representation



Representations

“cat”

“dog”

Representing meaning as vectors

- common representation space
- enables information sharing
- can be learned from data

- One-hot

  cat = [0 0 0 1 0 0 0]

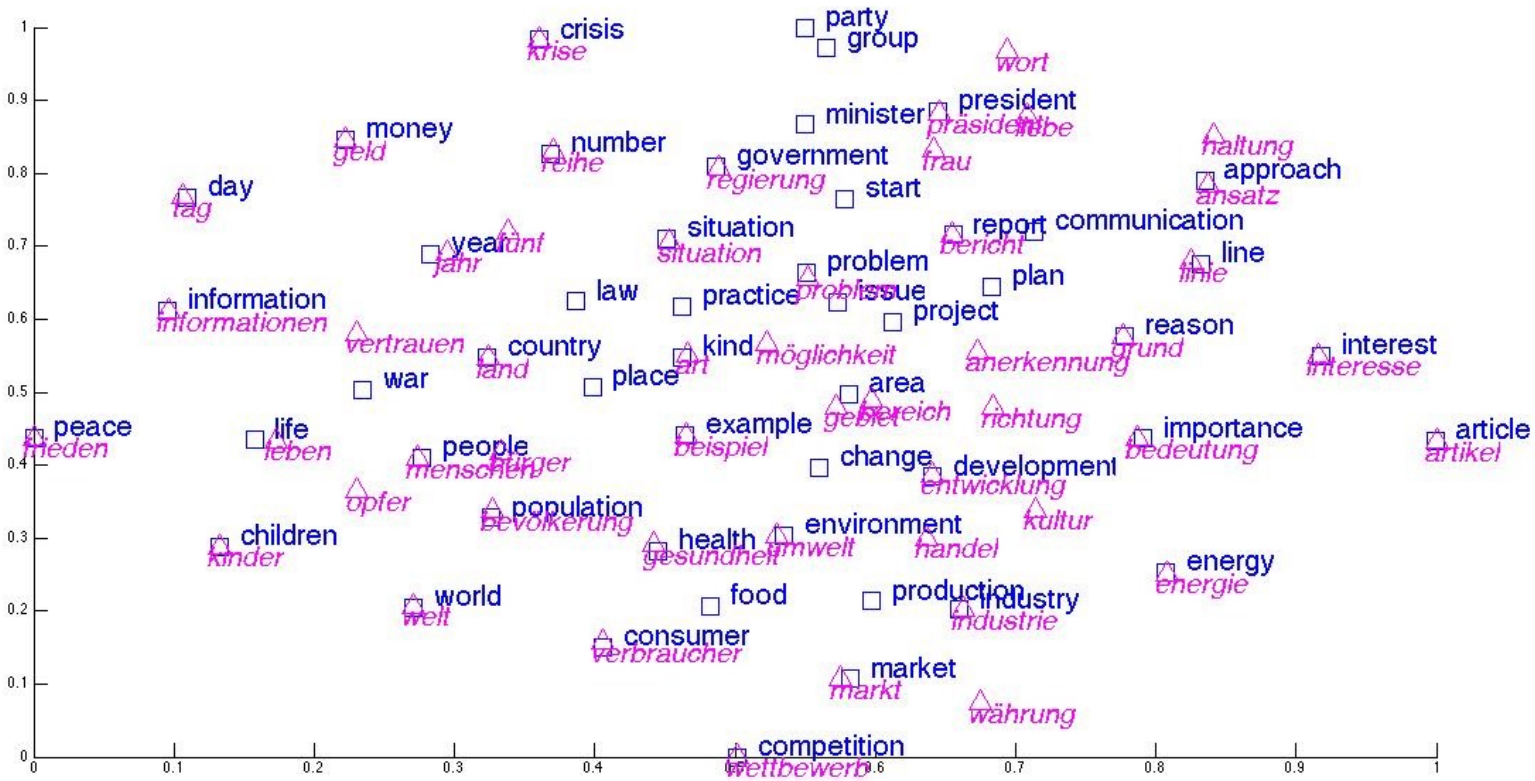
  dog = [0 0 0 0 0 1 0]

- Embeddings

  cat = [0.04 1.79 -1.79 1.07 0.48]

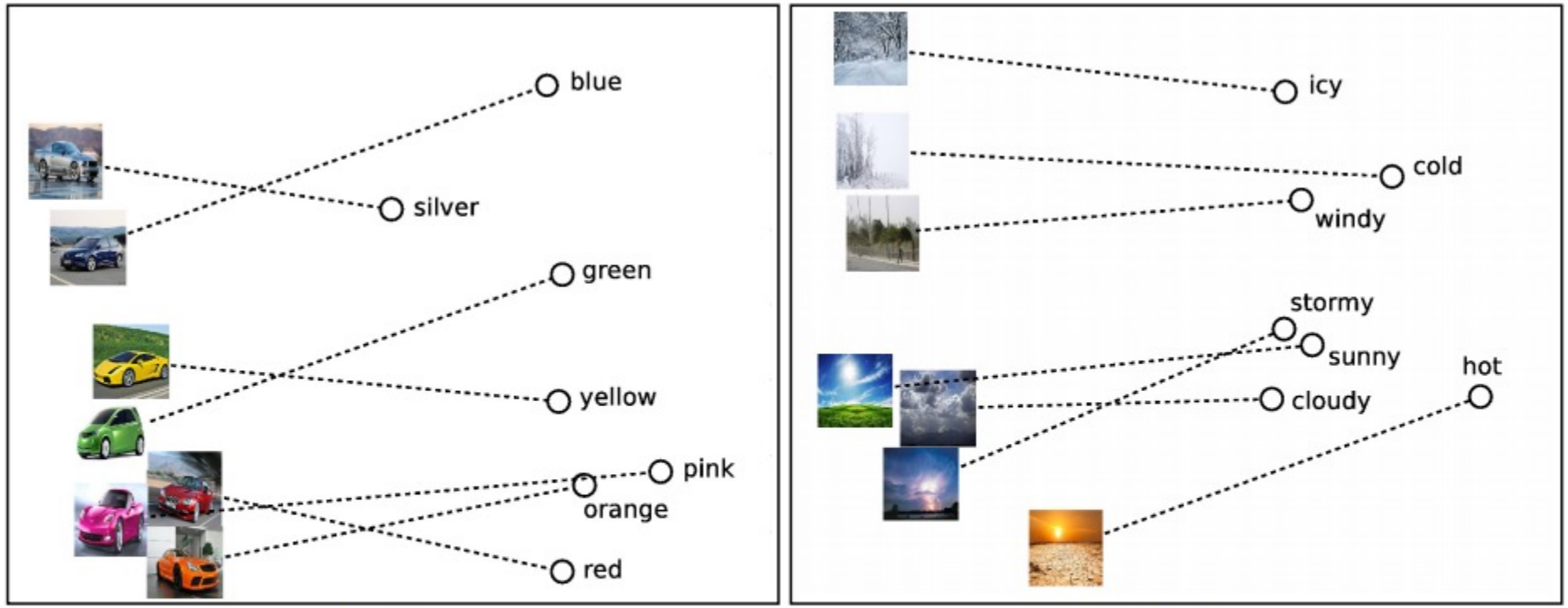
  dog = [0.61 1.84 -1.12 0.52 0.53]

Word Embeddings



"Bilingual Word Representations with Monolingual Quality in Mind"
[Minh-Thang Luong, Hieu Pham, and Christopher D. Manning NAACL 2015 VSM Workshop]

Multimodal Embeddings

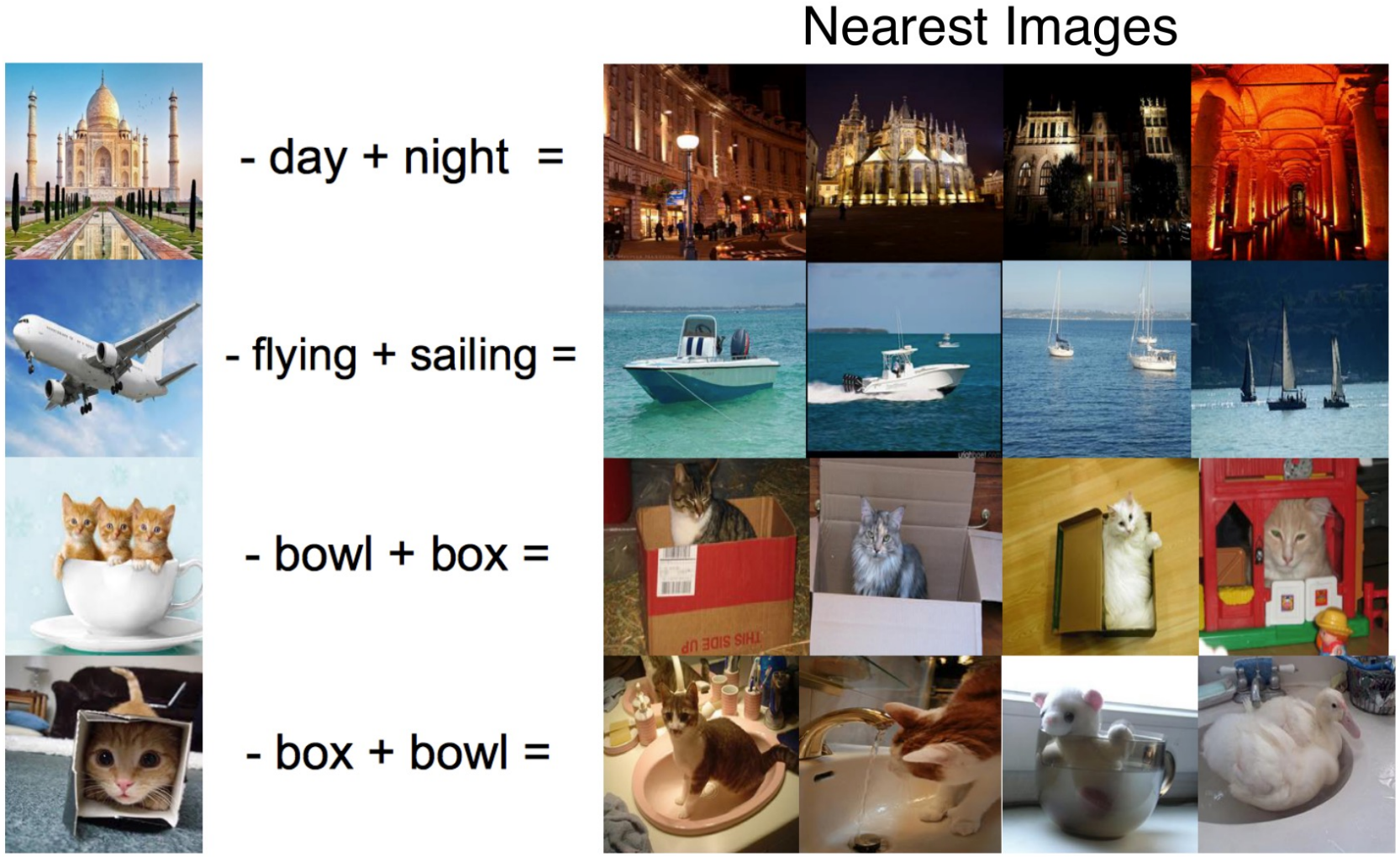


(a) Colors (b) Weather

Figure 5: PCA projection of the 300-dimensional word and image representations for (a) cars and colors and (b) weather and temperature.

"Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models"
[Kiros, Salakhutdinov, Zemel TACL 2015]

Multimodal Embeddings



(Kiros, Salakhutdinov, Zemel, TACL 2015)

Compositional Semantics

How do units of meaning combine?

“house” + “teapot” = “house teapot”

Compositional word embeddings

"house teapot"

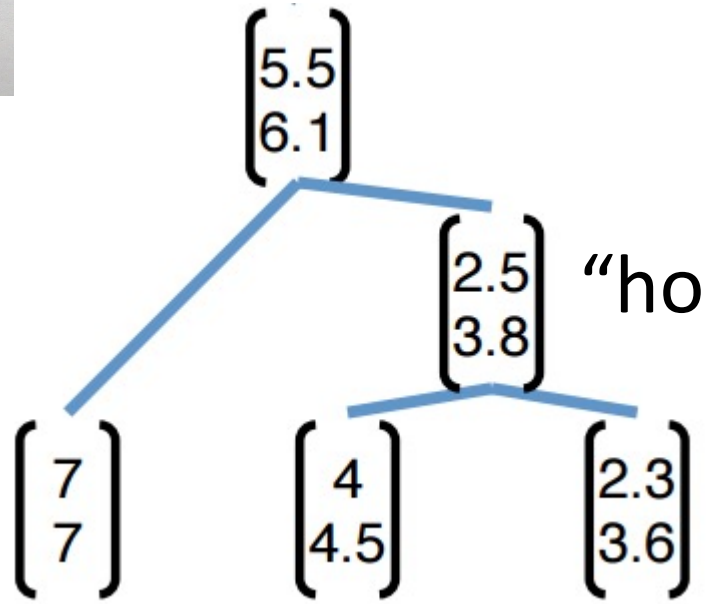
$$\begin{bmatrix} 2.5 \\ 3.8 \end{bmatrix}$$

"house" $\begin{bmatrix} 4 \\ 4.5 \end{bmatrix}$ $\begin{bmatrix} 2.3 \\ 3.6 \end{bmatrix}$ "teapot"

Compositional word embeddings

"mushroom house teapot"

"house teapot"



"mushroom" "house" "teapot"

Other representations

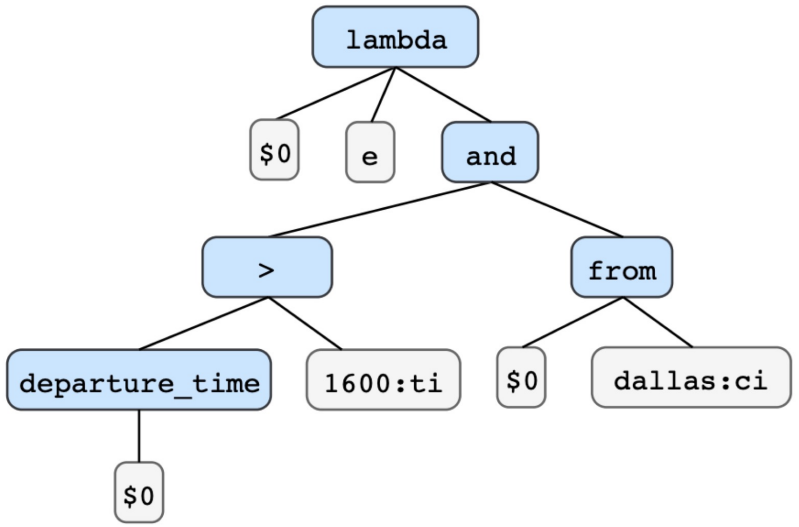
Logical forms

Show me flights from Pittsburgh to Seattle

```
lambda $0 e (and (flight $0)
     (from $0 pittsburgh:ci)
     (to $0 seattle:ci))
```

Parse trees

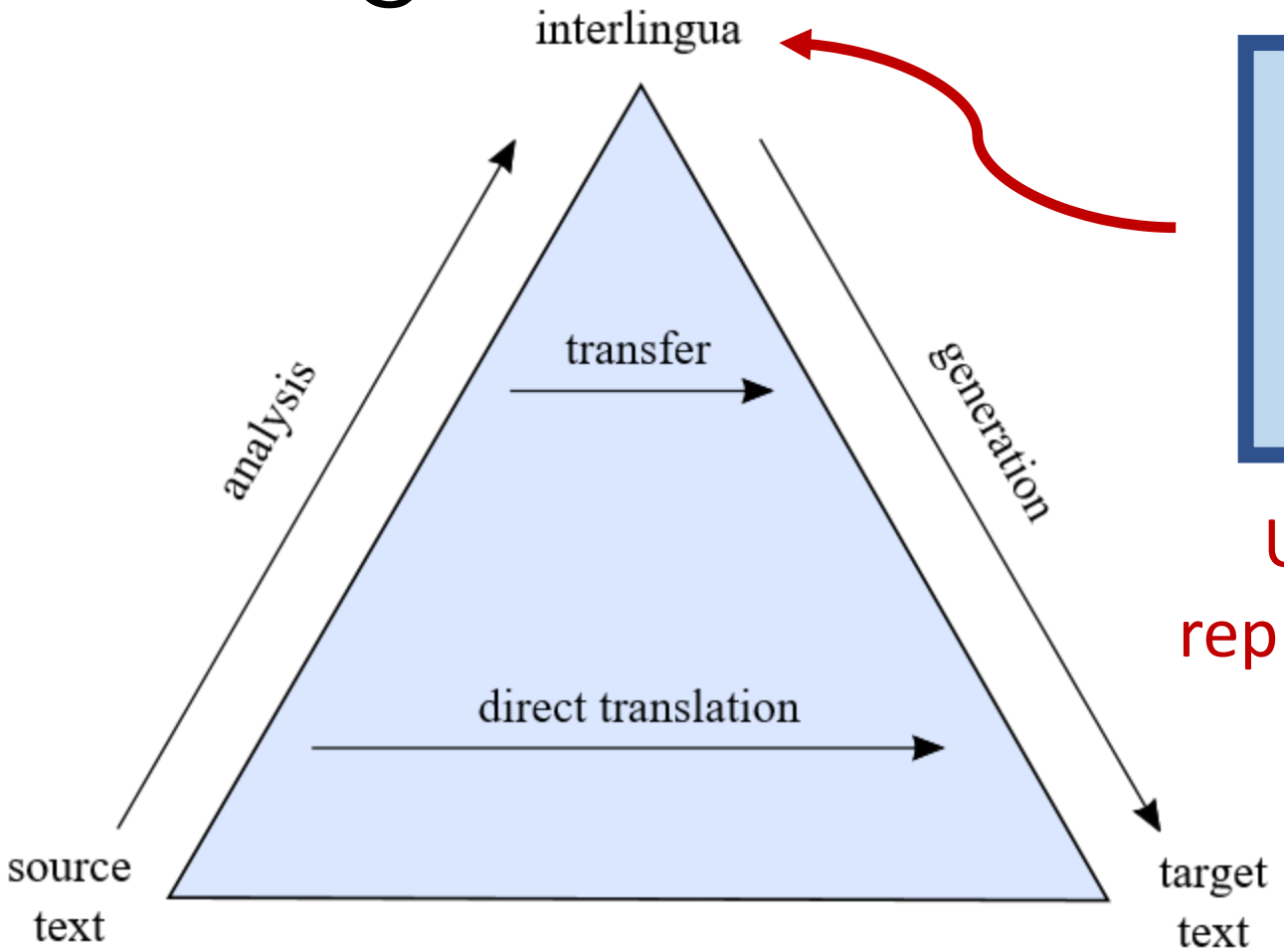
Show me flight from Dallas departing after 16:00



Vector representations

Tasks

Vauquois Triangle for translation



1.2
3.4
0.5
1.6
9.2
2.1

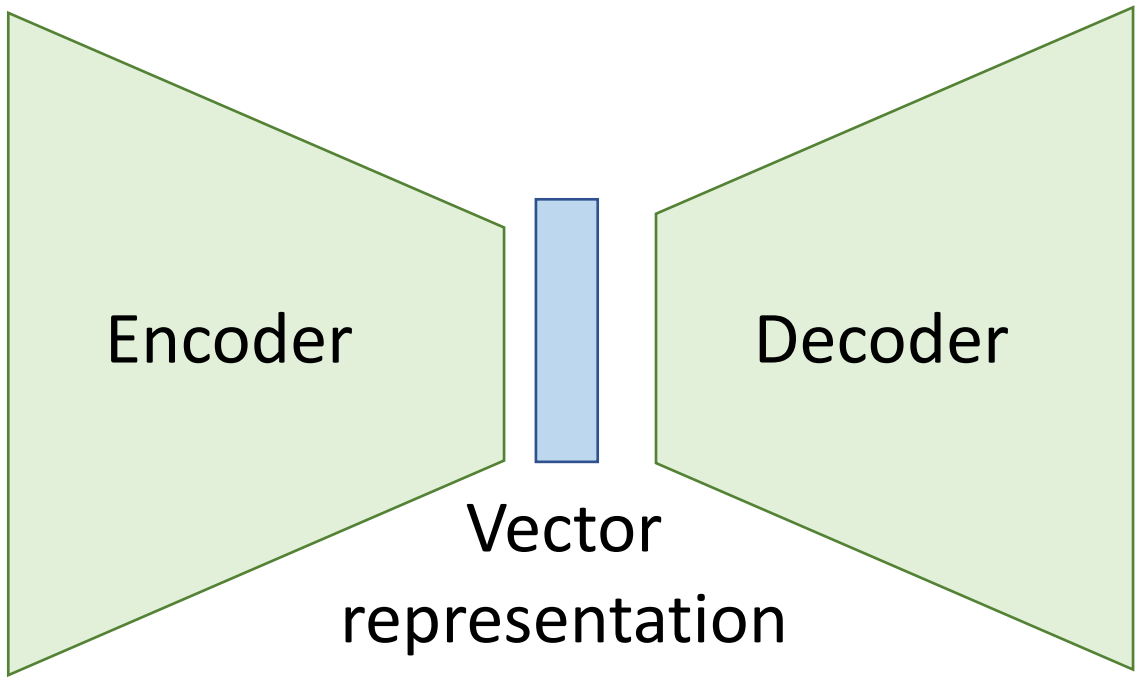
Use vector to represent meaning

the cat sat on the table

η γάτα κάθισε στο τραπέζι

Translating between languages

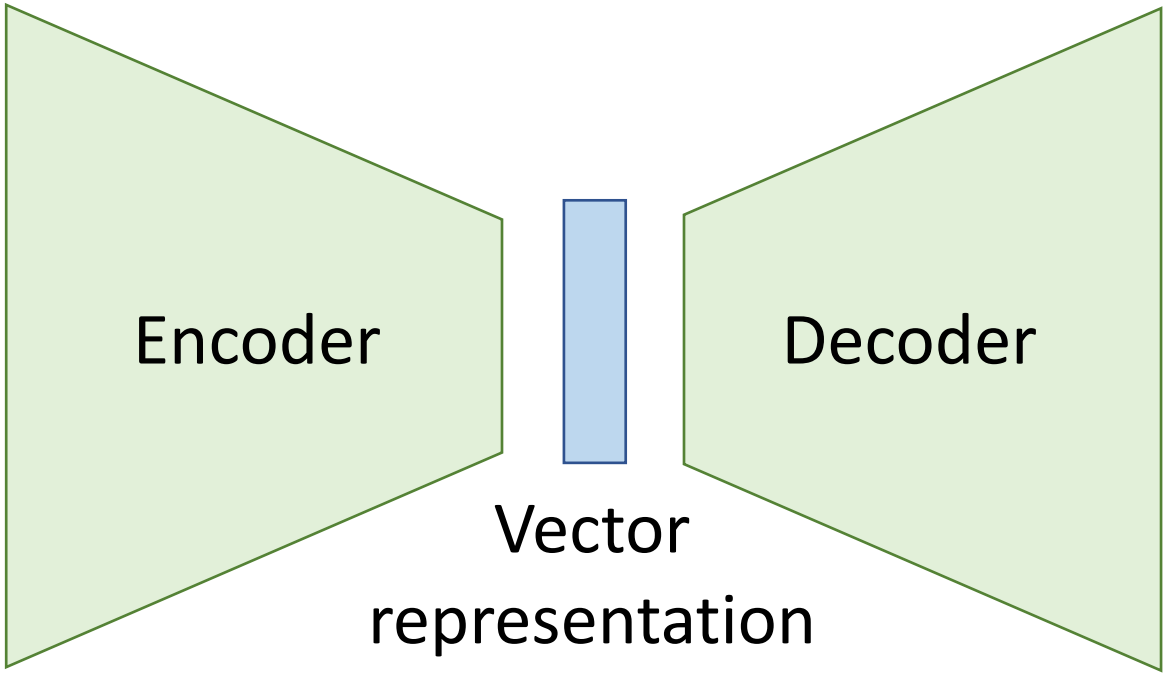
the cat sat
on the table



η γάτα κάθισε
στο τραπέζι

Translating across modalities

Image



Text

man in black shirt
is playing guitar

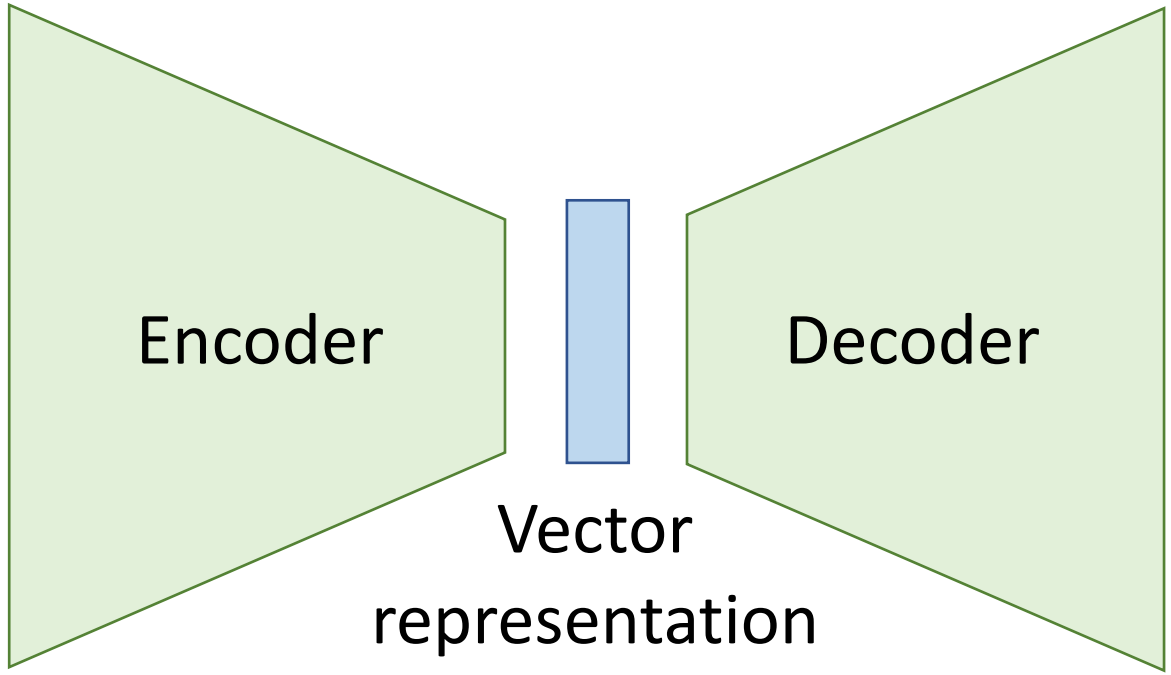
Image captioning

"Deep Visual-Semantic Alignments for Generating Image Descriptions"
[Karpathy and Fei-Fei CVPR 2015]

Translating across modalities

Text

A white bird with
a black crown
and yellow beak



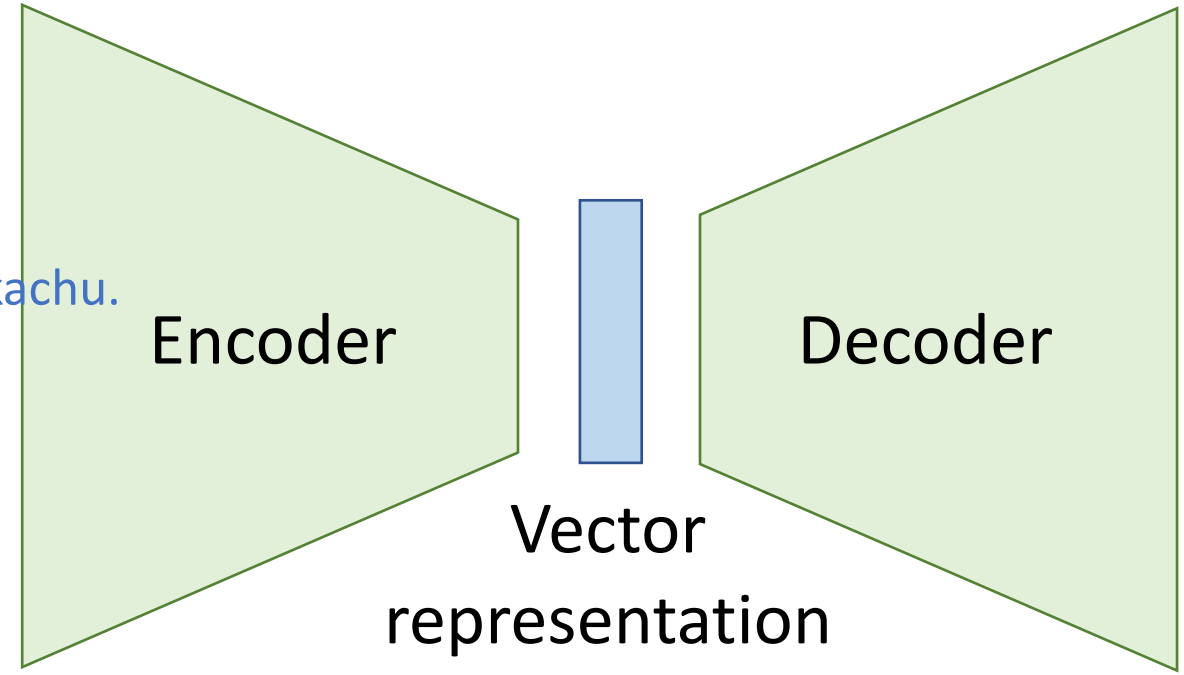
Image

"StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks"
[Zhang et al, ICCV 2017]

Translating across modalities

Text

a teapot in the shape of a pikachu.
a teapot imitating a pikachu



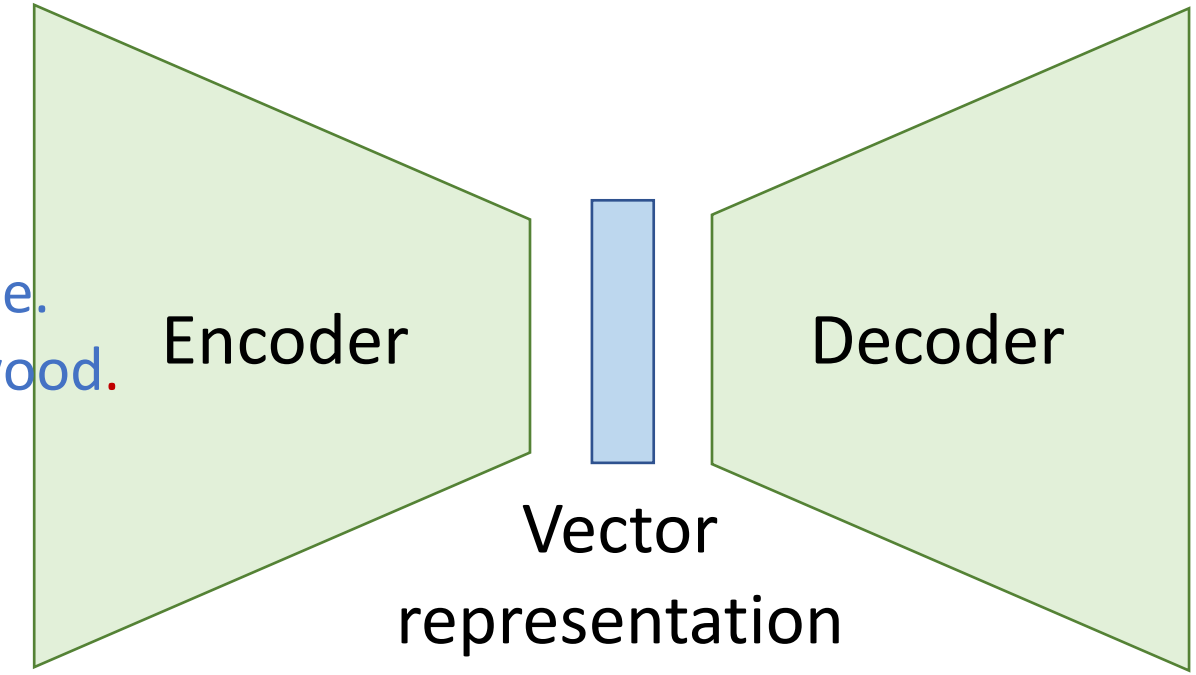
Image

"Dall-e"
[Ramesh et al, https://openai.com/blog/dall-e/]

Translating across modalities

Text

Brown colored dining table.
It has four legs made of wood.



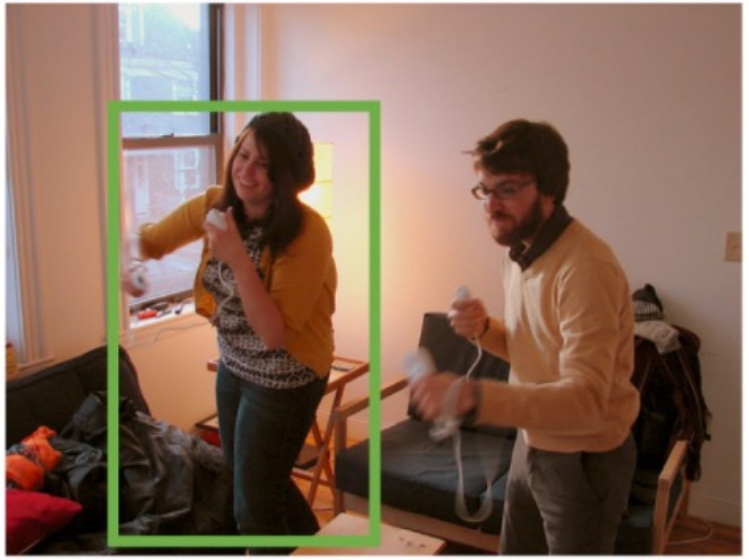
3D Shape

"Text2Shape: Generating Shapes from Natural Language by Learning Joint Embeddings"
[Chen et al, ACCV 2018]

Referring Expressions

Task 1: Expression Generation

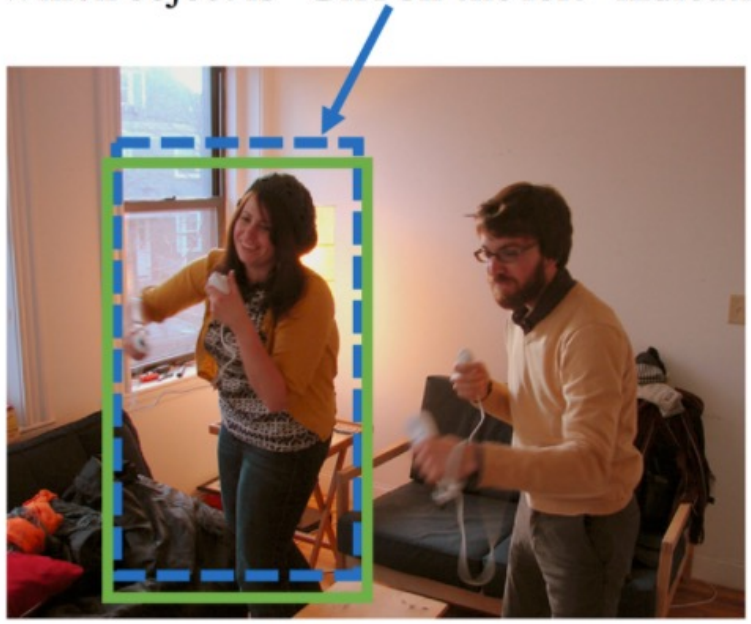
Generate referring expression for this target person.



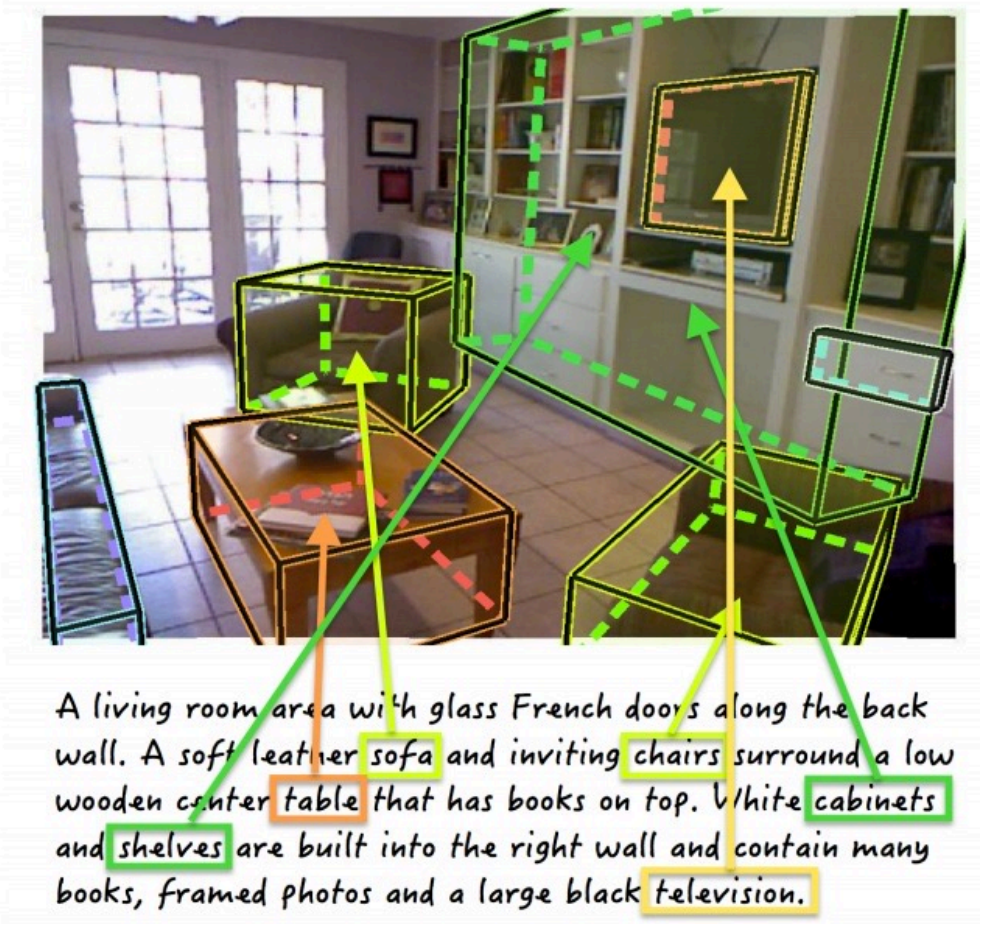
Algorithm: The girl playing wii

Task 2: Expression Comprehension

Which object is "**Girl on the left**" indicating?



Alignment



"What are you talking about? Text-to-Image Coreference"
[Kong et al, CVPR 2014]

Visual Question Answering

Who is wearing glasses?

man

woman

Where is the child sitting?

fridge

arms

Is the umbrella upside down?

yes

no

How many children are in the bed?

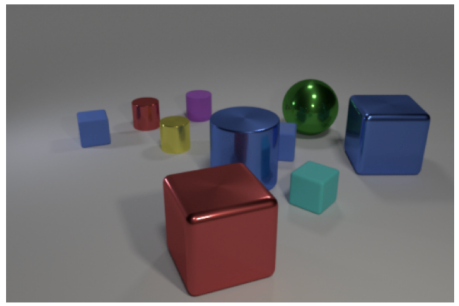
2

1

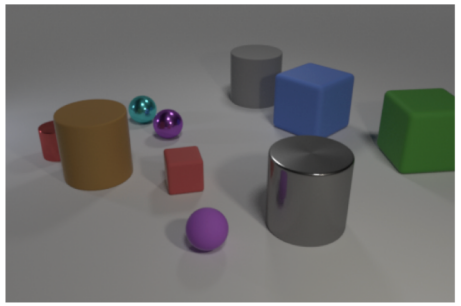
[Antol et al, ICCV 2015]

Visual Question Answering

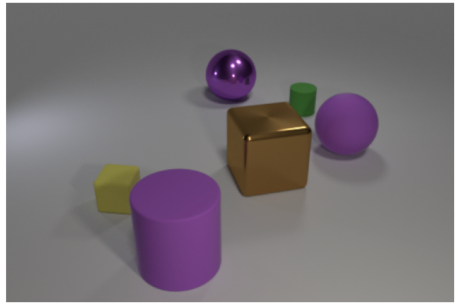
Compositionality and reasoning (CLEVR dataset, Johnson et al, 2017)



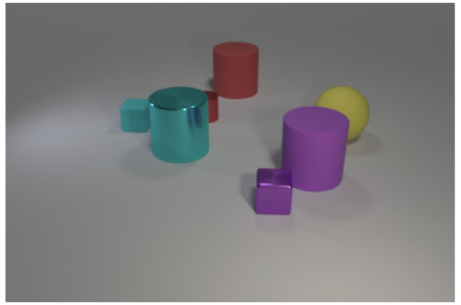
Q: What shape is the object reflected in the blue cylinder?
A: cube



Q: What number of cylinders share the same color?
A: 2

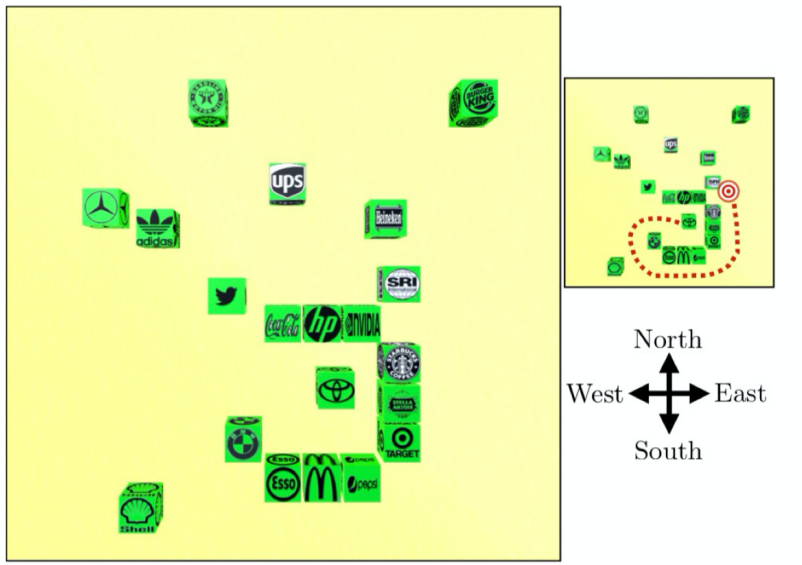


Q: How many objects are not purple and not metallic?
A: 2



Q: What color is the object partially blocked by the purple cylinder?
A: yellow

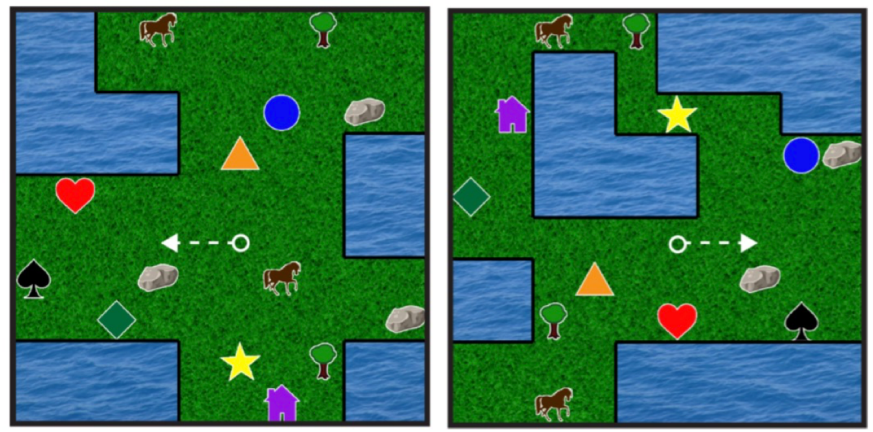
Spatial Reasoning



Put the Toyota block in the same row as the SRI block, in the first open space to the right of the SRI block
Move Toyota to the immediate right of SRI, evenly aligned and slightly separated
Move the Toyota block around the pile and place it just to the right of the SRI block
Place Toyota block just to the right of The SRI Block
Toyota, right side of SRI

Robotic Manipulation

(Bisk et al., 2016, Misra et al., 2017)

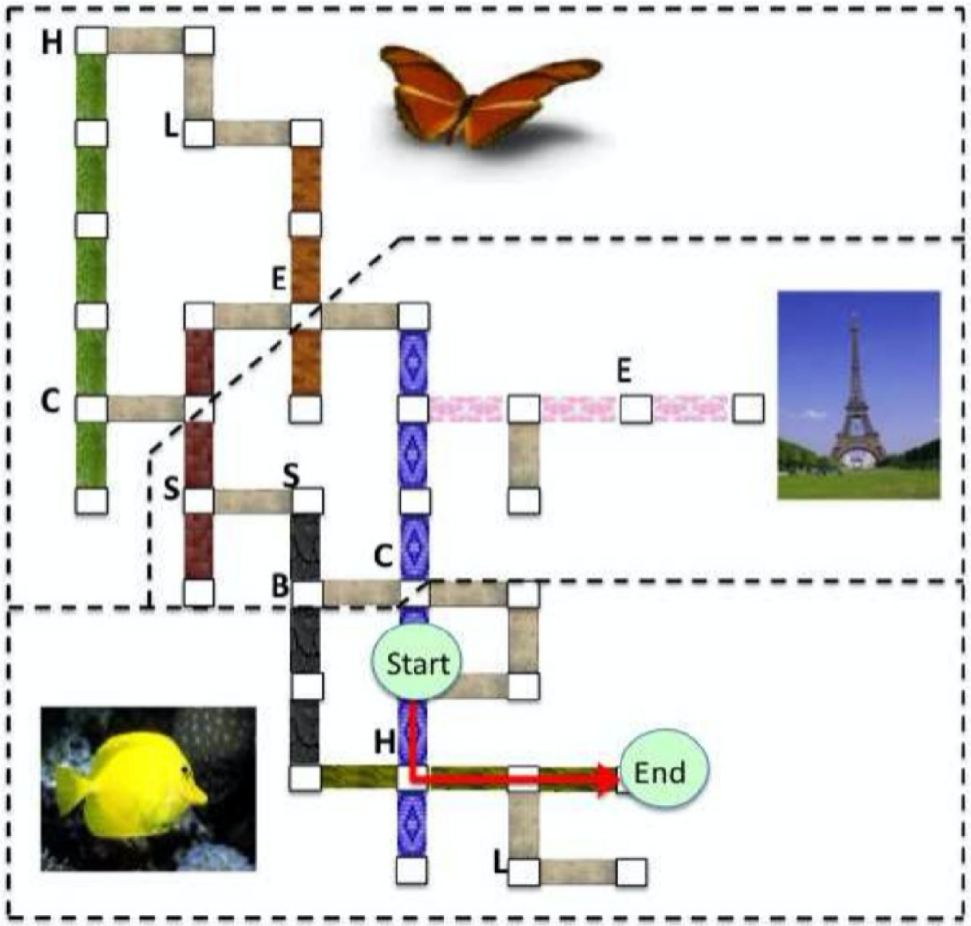


Reach the cell above the westernmost rock

Autonomous navigation

(Janner et al., 2017)

Instruction Following



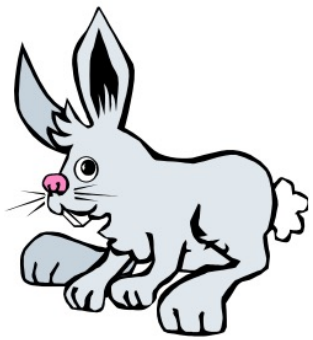
- Want to be able to follow instructions in a virtual environment
- "Go along the blue hall, then turn left away from the fish painting and walk to the end of the hallway"

Learning

There's a WOCKET in my POCKET!

Dr. Seuss

What does “gavagai” mean?

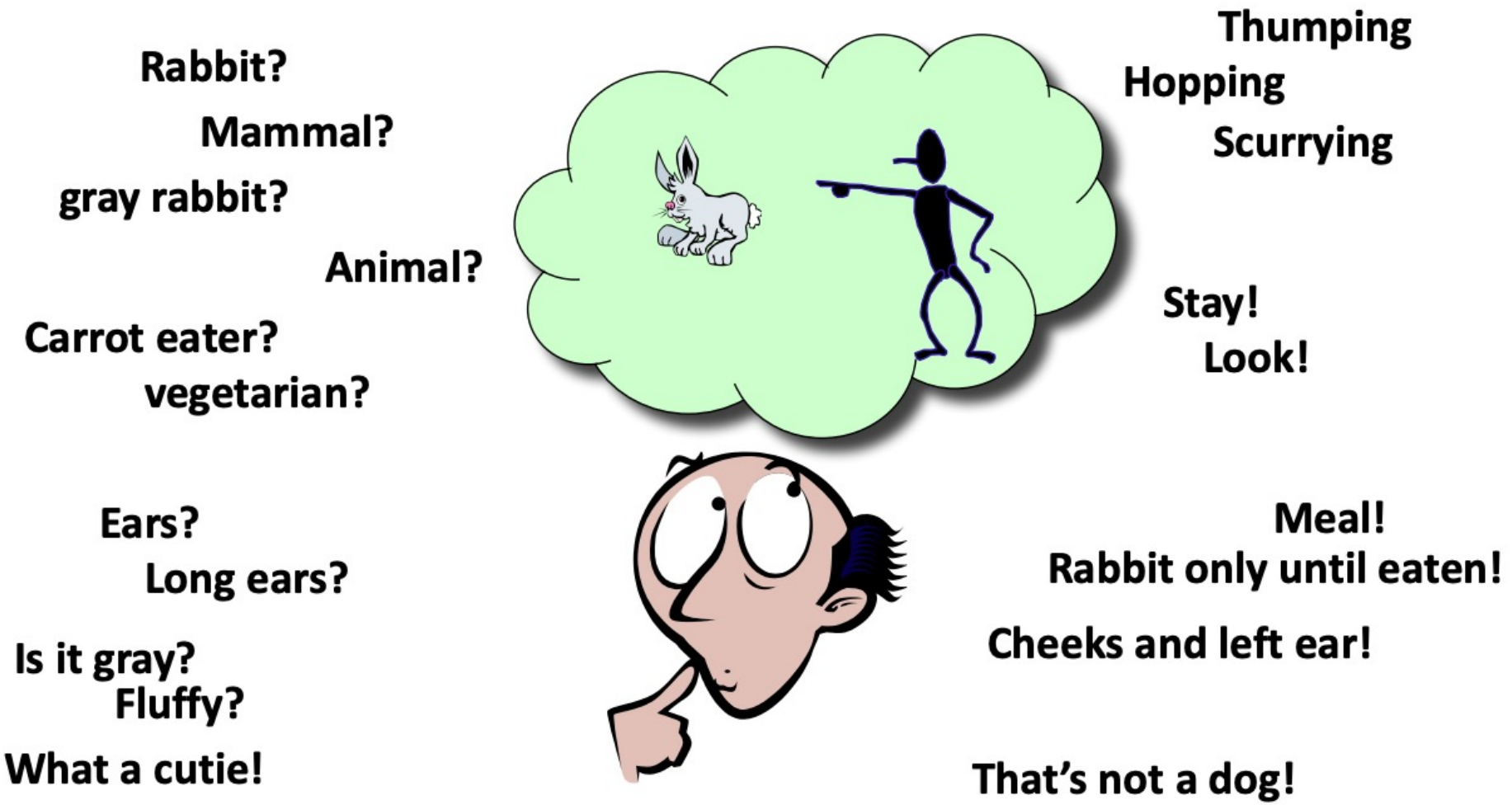


Slide credit: Lisa Pearl

What does “gavagai” mean?

How can we learn the correct association?

Do computer models learn the correct association?

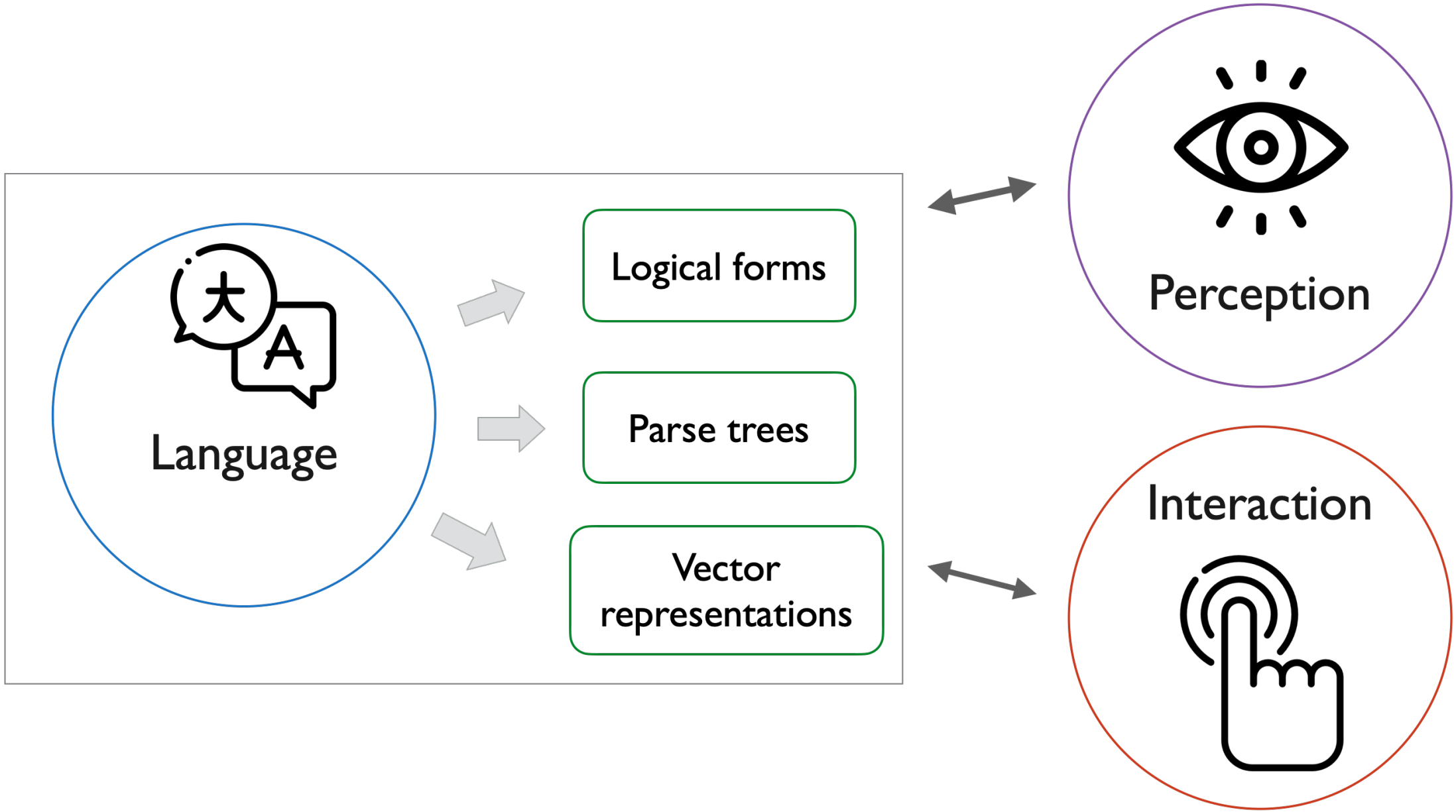


Slide credit: Lisa Pearl

- Children do not learn language from raw text or passively watching TV

- Natural way to learn language in the context of its use in the physical and social world

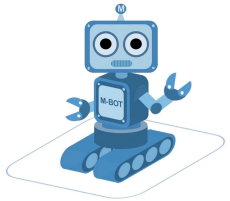
- This requires inferring the meaning of utterances from their perceptual context


Language
Logical forms
Parse trees
Vector representations
Perception
Interaction

Embodied AI

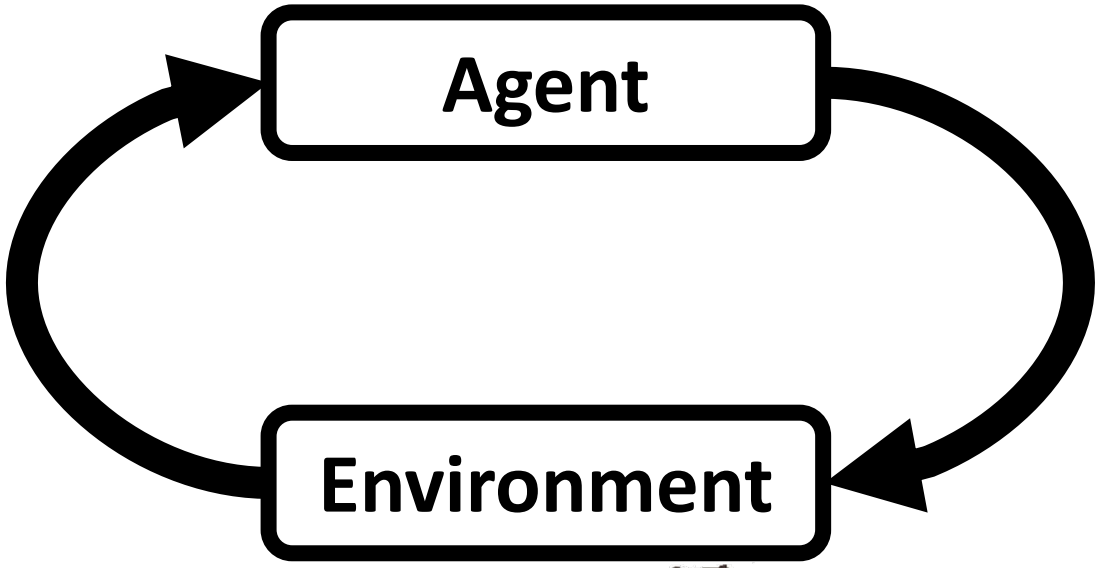
Learning to perceive + act + communicate with physical embodiment

Exit the bedroom. Turn left down the hall and stop in the kitchen.

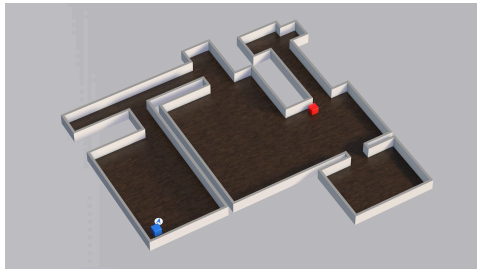


- Trained using reinforcement learning
- Agent can be purely reactive, or use memory or map representations

Observations



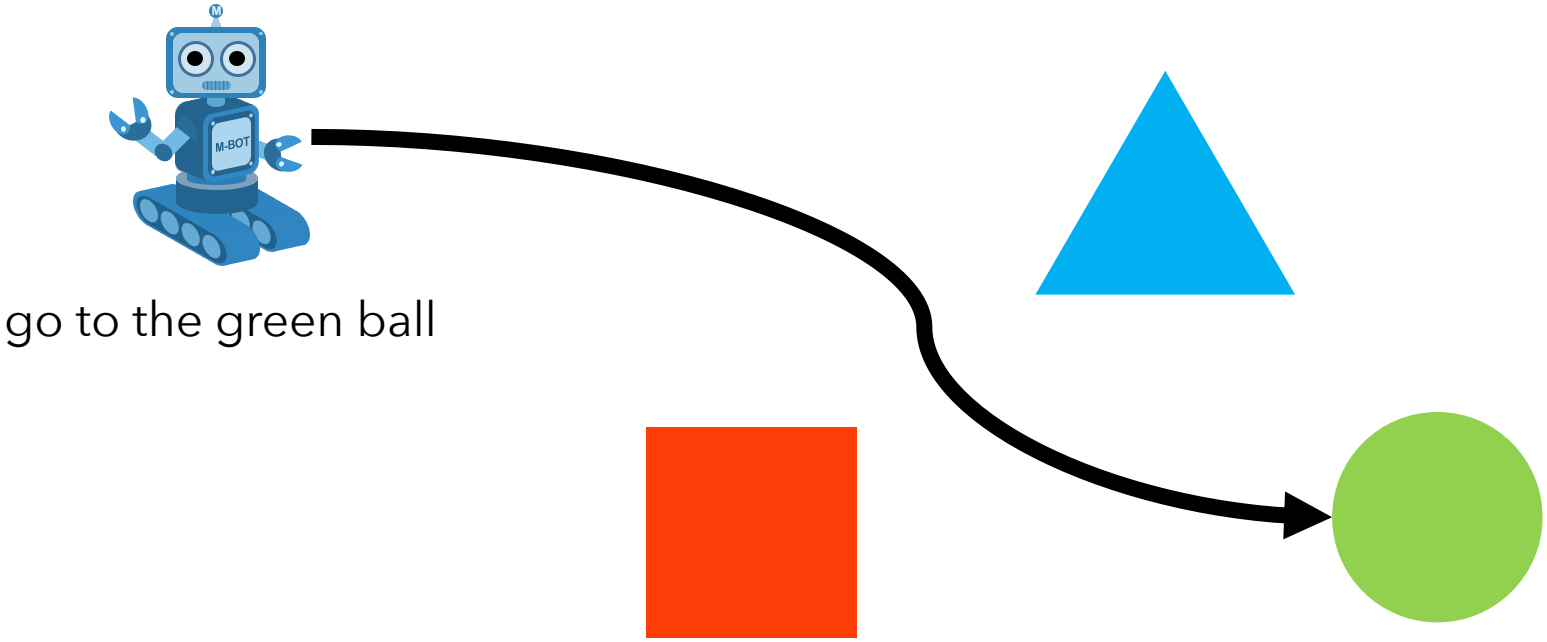
Actions



Embodied language learning

Grounded language learning for embodied agents

Learning natural language by interacting with an environment



Slide credit: Stefan Lee

Grounded Language Learning

Goal specified as an attributed object

- Focus is on language learning – often study generalization to compositionally novel instances

go to the small red object

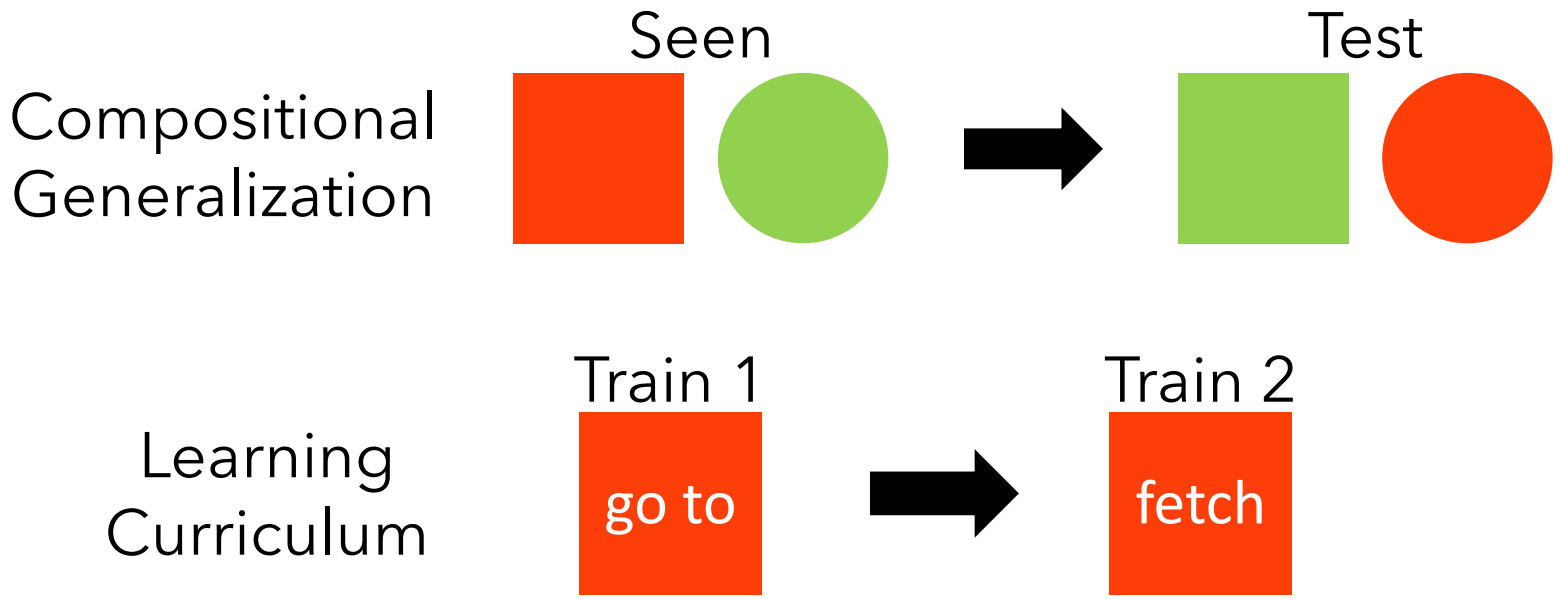
the target is left of the hair dryer

go to any green object

Slide credit: Stefan Lee

Grounded Language Learning

Controlled settings to study specific aspects of language learning:



Slide credit: Stefan Lee

Grounded Language Learning

- Grounded Language Learning in a Simulated 3D World arxiv.org/abs/1706.06551
- Understanding Grounded Language Learning Agents arxiv.org/abs/1710.09867

Slide credit: Stefan Lee

Upcoming

- Next time: Reading papers and project overview

- Next week:
  - Review of deep learning building blocks
    - MLPs
    - CNNs
    - RNNs
  - Multimodal representations