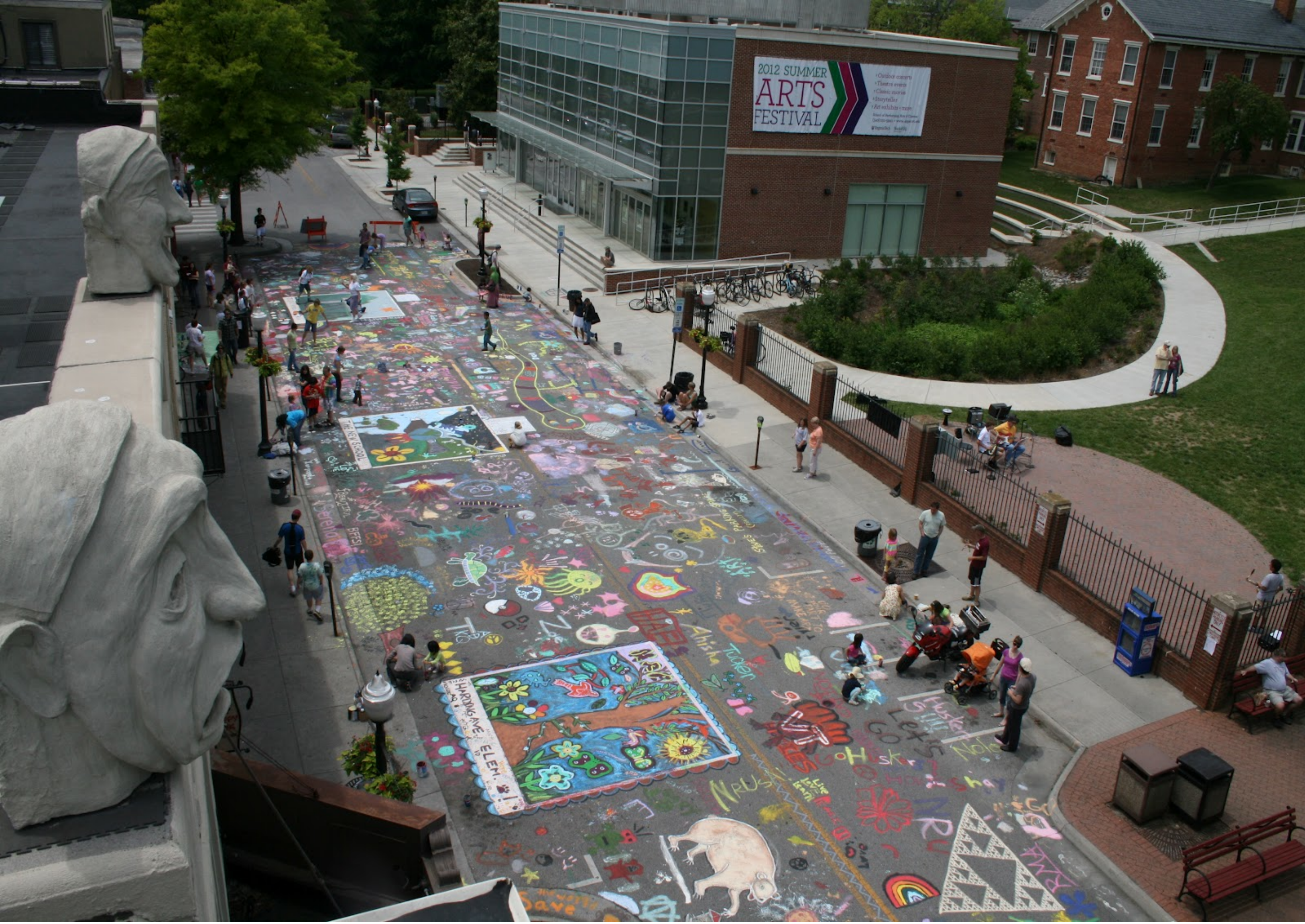
Learning by Playing

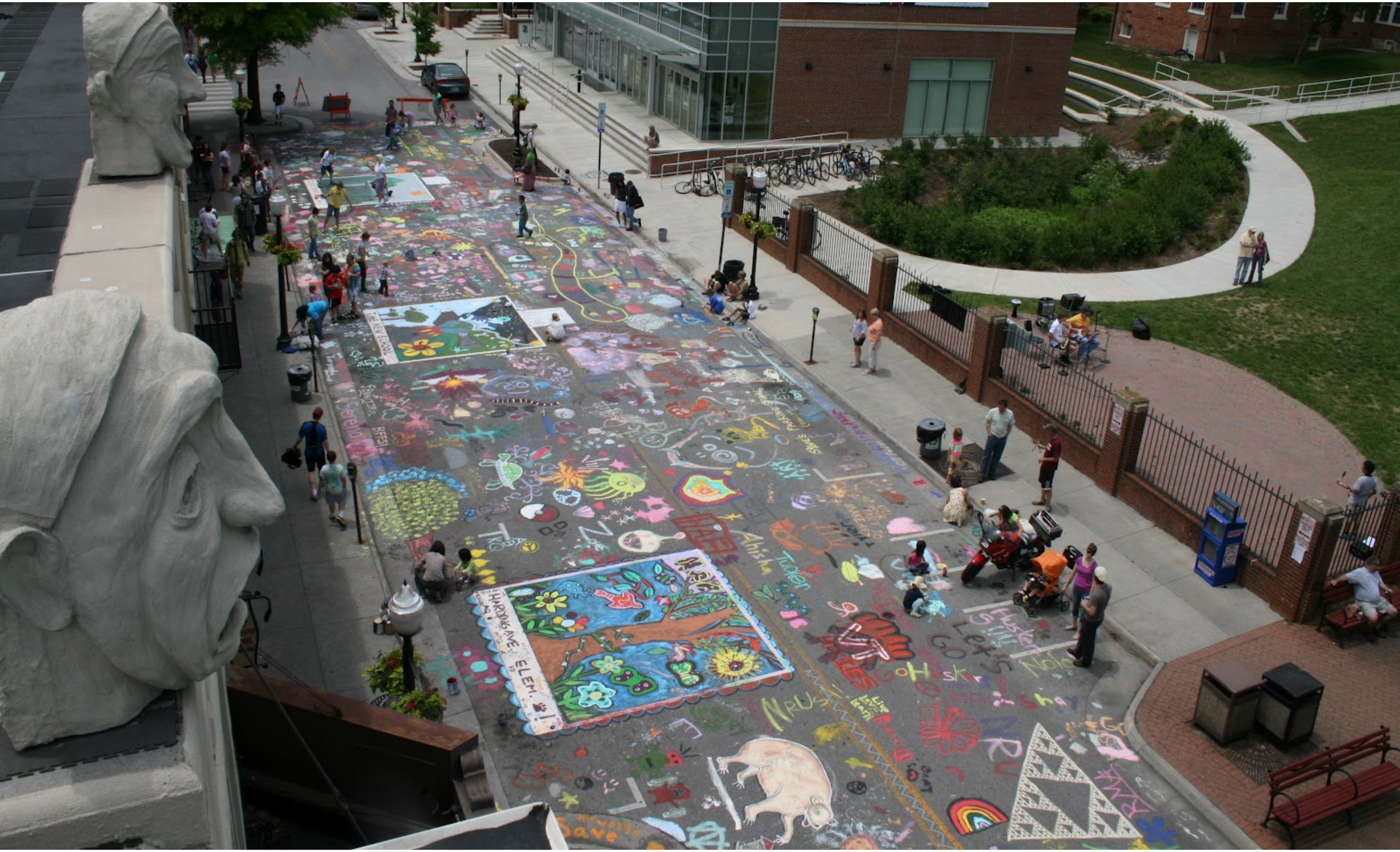
Devi Parikh

Georgia Tech



“Color College Avenue”, Blacksburg, VA, May 2012

People coloring a street in rural Virginia.



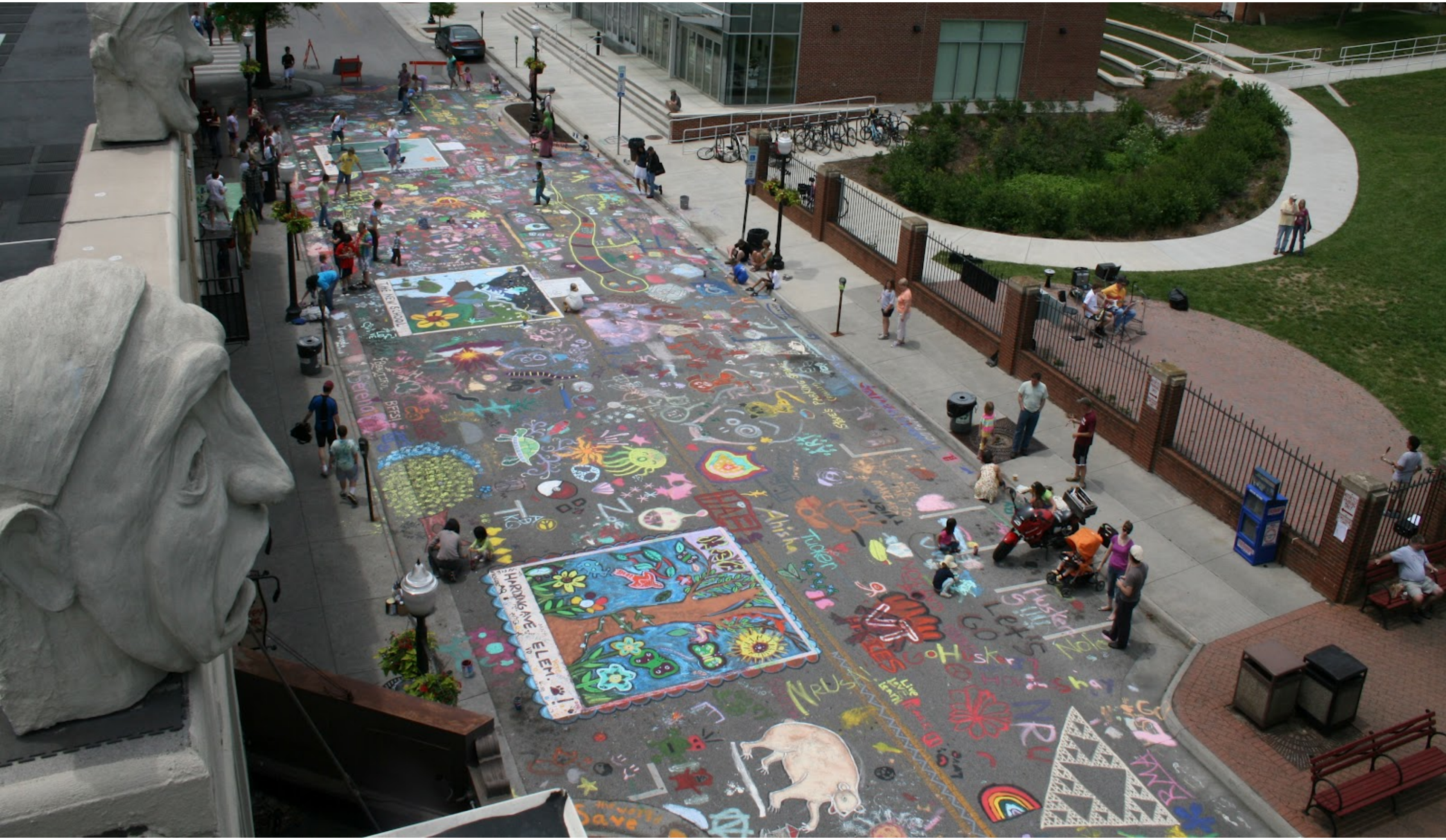
“Color College Avenue”, Blacksburg, VA, May 2012

It was a great event! It brought families out, and the whole community together.

“Color College Avenue”, Blacksburg, VA, May 2012

Q. What are they coloring the street with?

A. Chalk



“Color College Avenue”, Blacksburg, VA, May 2012

AI: What a nice picture! What event was this?

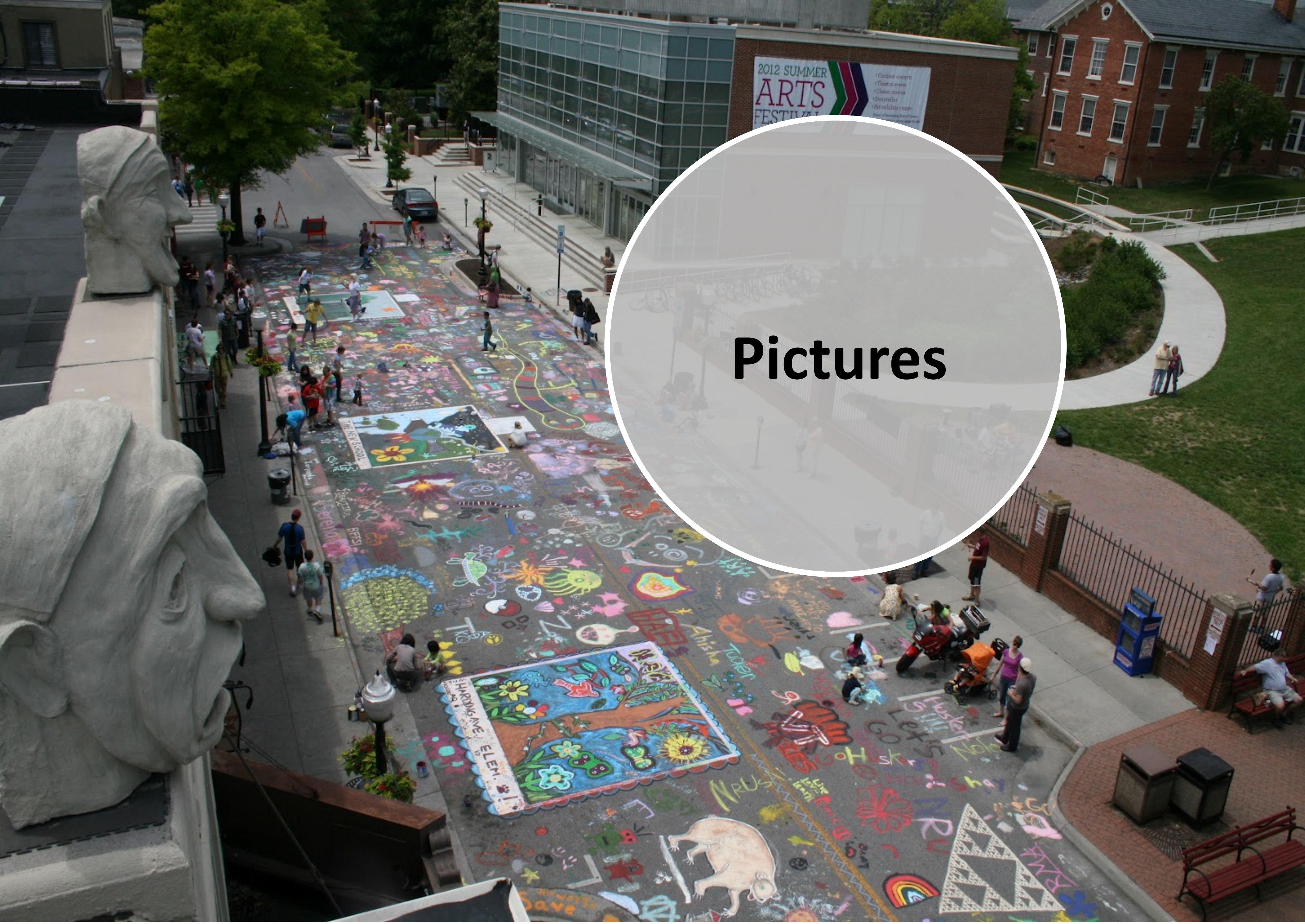
User: *“Color College Avenue”. It was a lot of fun!*

AI: I am sure it was! Do they do this every year?

User: *I wish they would. I don’t think they’ve organized it again since 2012.*

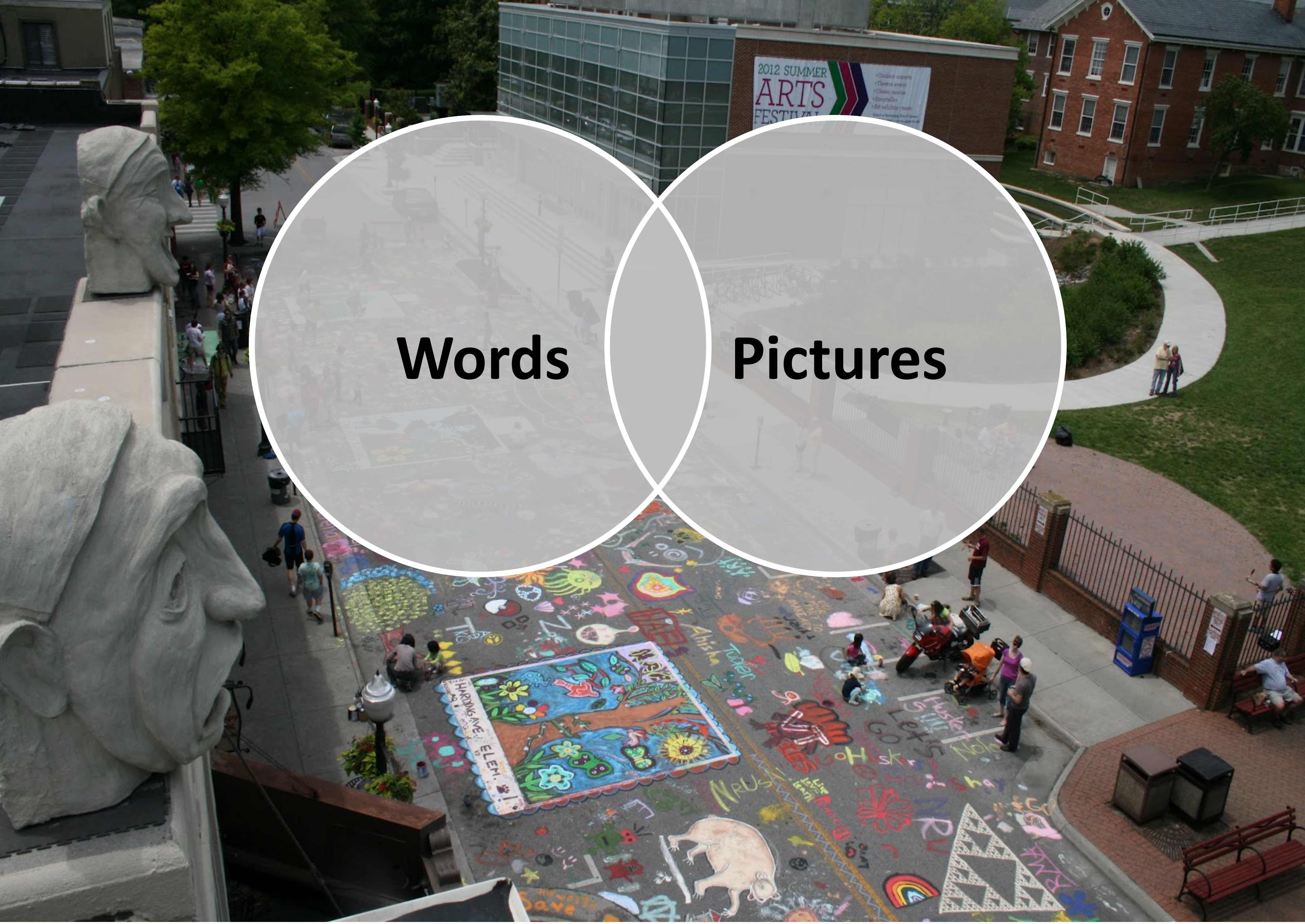
…

“Color College Avenue”, Blacksburg, VA, May 2012



Pictures

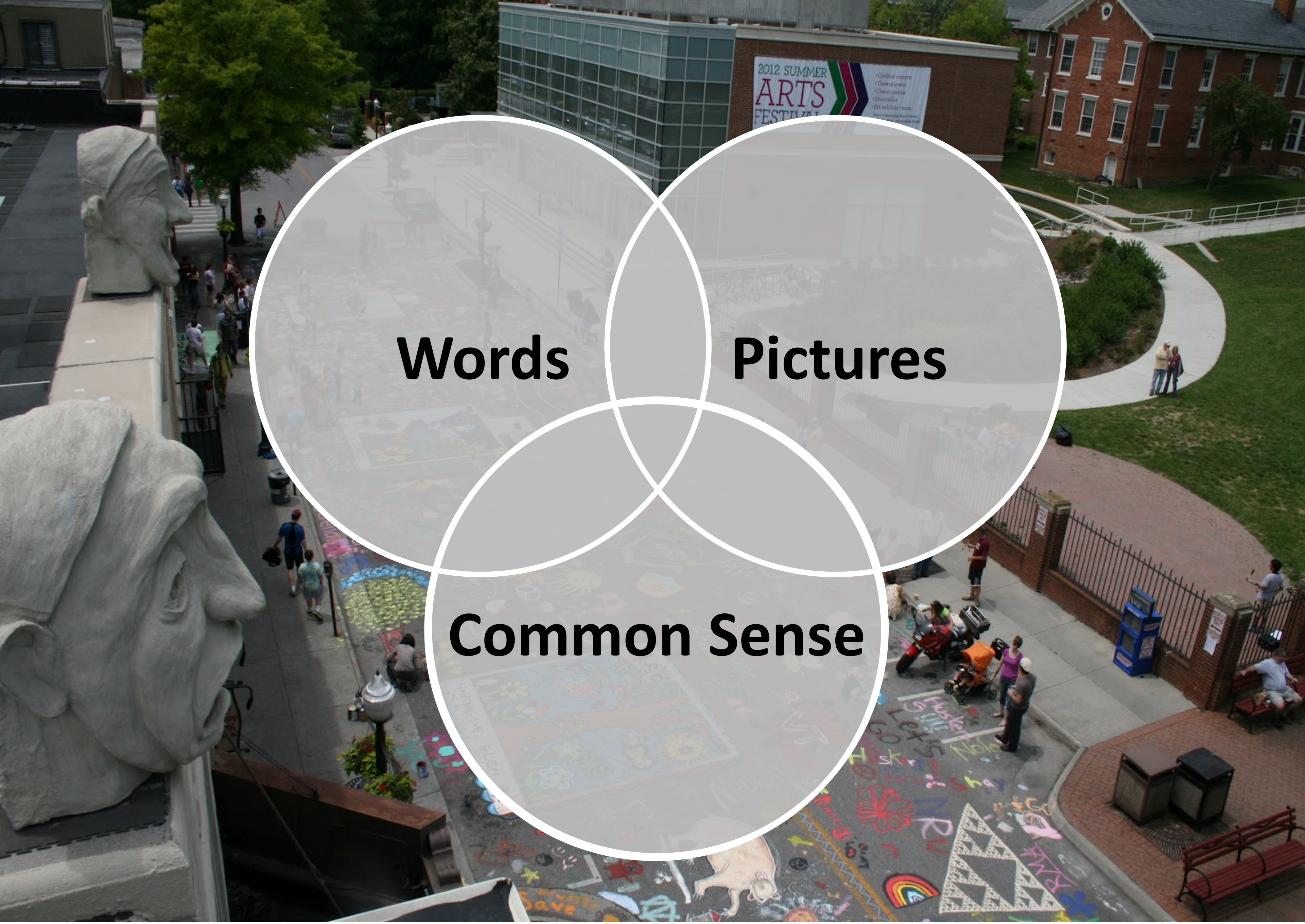
“Color College Avenue”, Blacksburg, VA, May 2012



Words

Pictures

“Color College Avenue”, Blacksburg, VA, May 2012



Words

Pictures

Common Sense

“Color College Avenue”, Blacksburg, VA, May 2012

Man in blue wetsuit is surfing on wave

Karpathy and Fei-Fei (Stanford) 2015

A group of young people playing a game of Frisbee

Vinyals et al. (Google) 2015

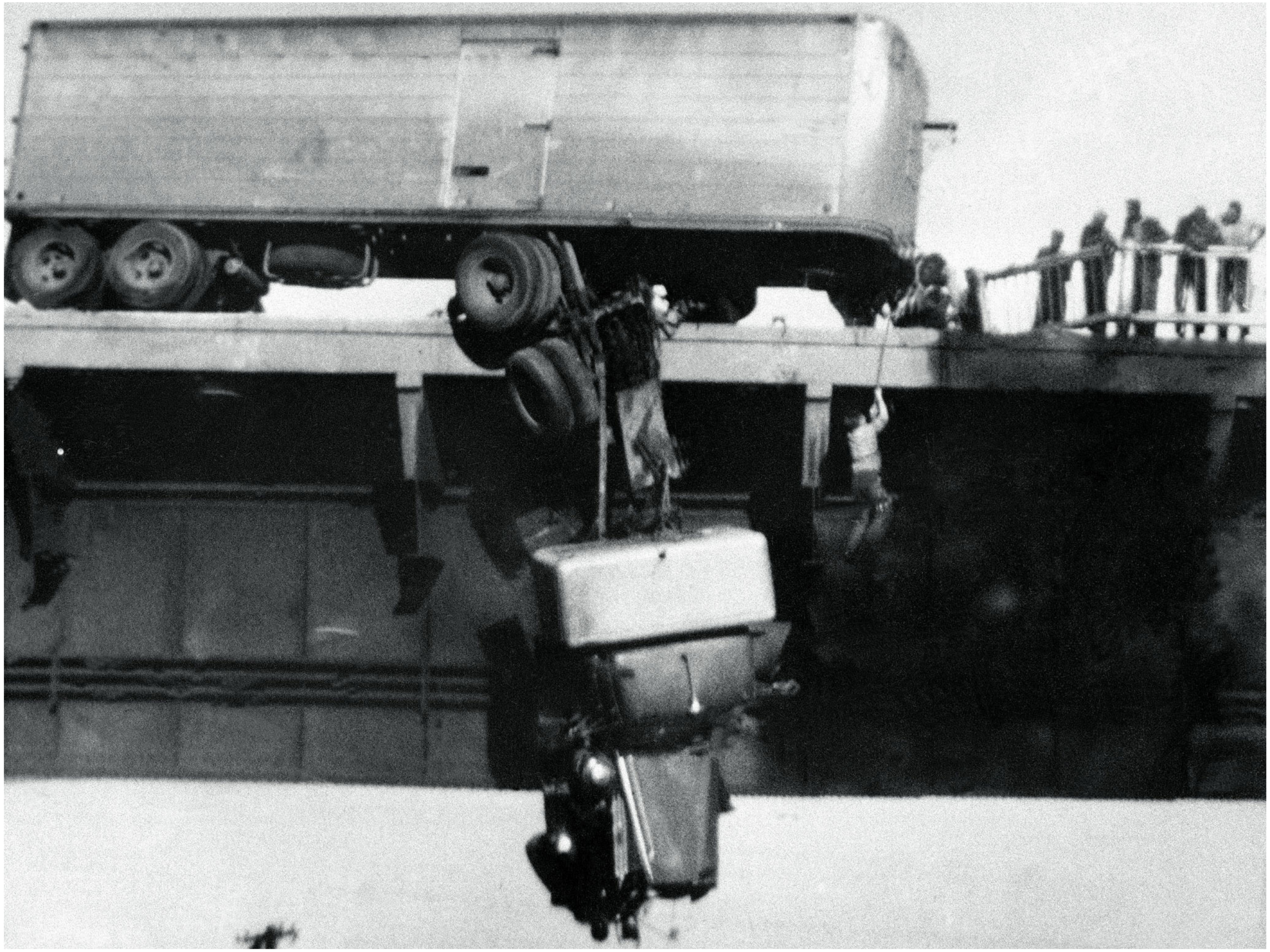
A car is parked in the middle of nowhere

Kiros et al. (University of Toronto) 2015

A pot of broccoli on a stove.

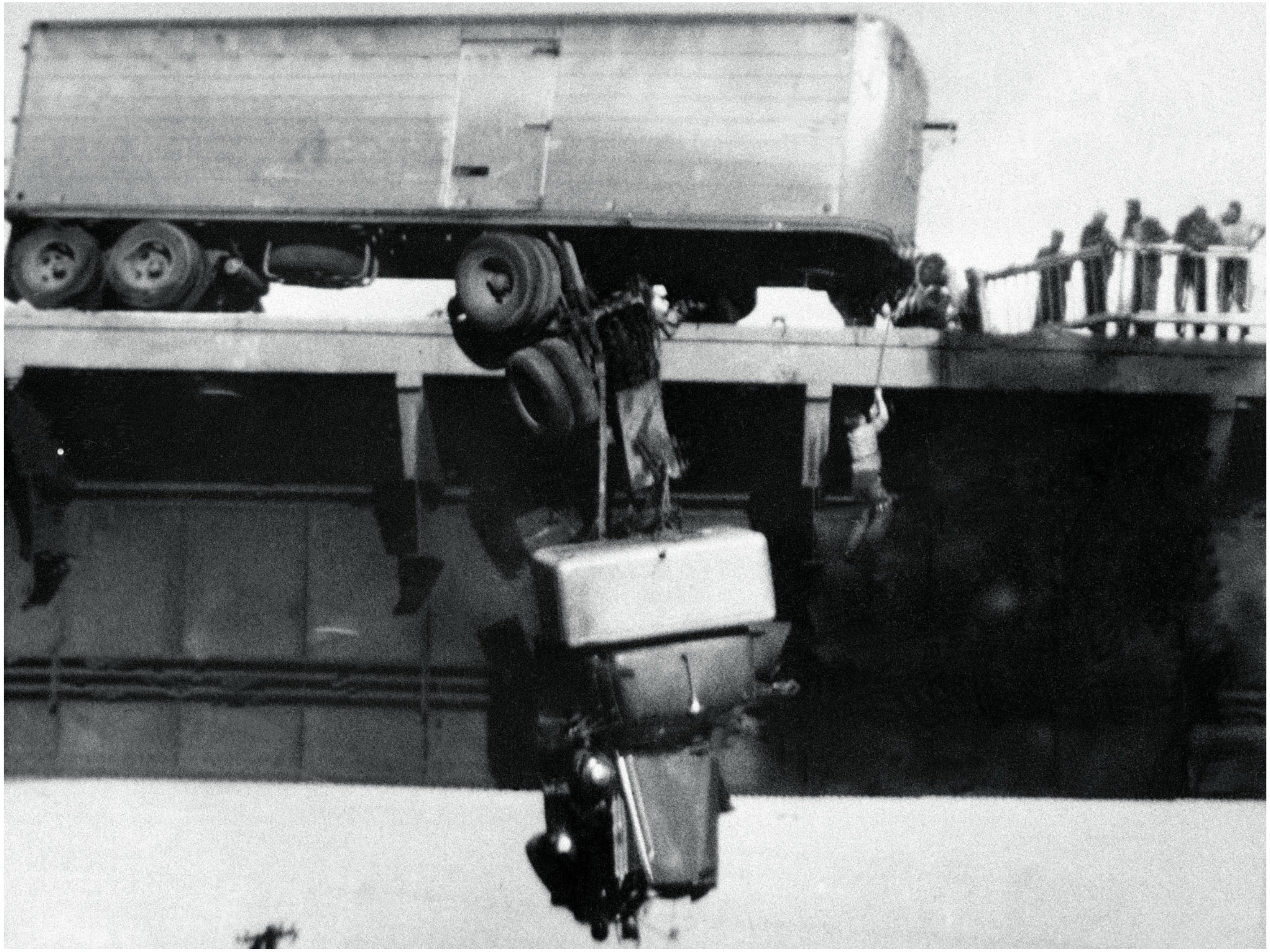
Fang et al. (Microsoft Research) 2015

A man is rescued from his truck that is hanging dangerously from a bridge.



Slide credit: Larry Zitnick

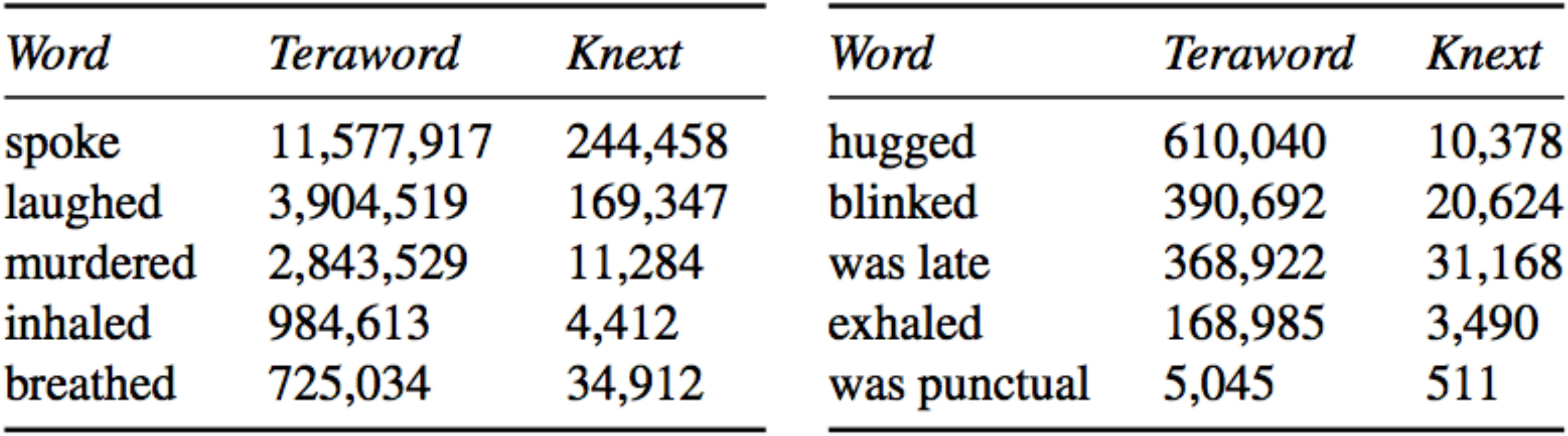
A man is *rescued* from his truck that is hanging *dangerously* from a bridge.



Learning Common Sense

- Text
  - Reporting bias

Reporting bias in text

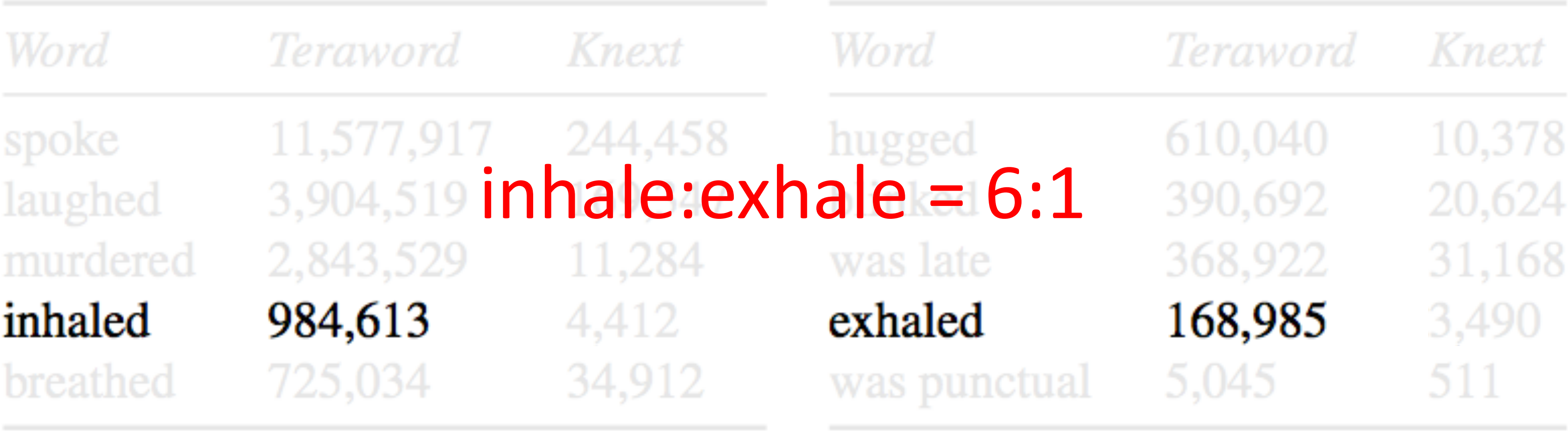


Word	*Teraword*	*Knext*
spoke	11,577,917	244,458
laughed	3,904,519	169,347
murdered	2,843,529	11,284
inhaled	984,613	4,412
breathed	725,034	34,912

Word	*Teraword*	*Knext*
hugged	610,040	10,378
blinked	390,692	20,624
was late	368,922	31,168
exhaled	168,985	3,490
was punctual	5,045	511

[Gordon et al. 2013]

Reporting bias in text

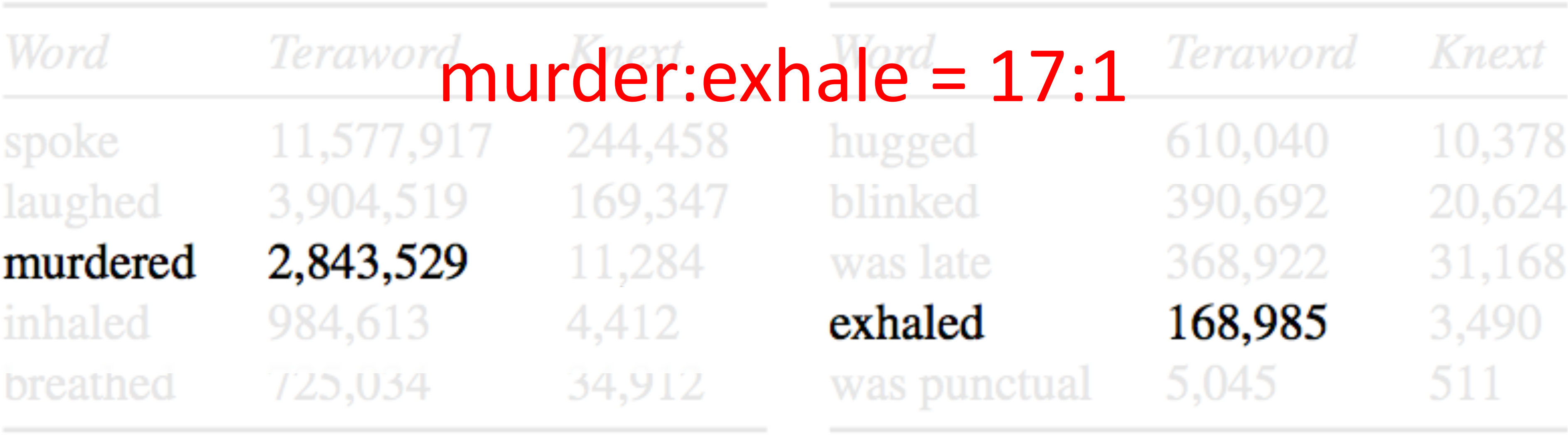


Word	Teraword	Knext	Word	Teraword	Knext
spoke	11,577,917	244,458	hugged	610,040	10,378
laughed	3,904,519	169,347	blinked	390,692	20,624
murdered	2,843,529	11,284	was late	368,922	31,168
inhaled	**984,613**	4,412	**exhaled**	**168,985**	3,490
breathed	725,034	34,912	was punctual	5,045	511

inhale:exhale = 6:1

[Gordon et al. 2013]

Reporting bias in text



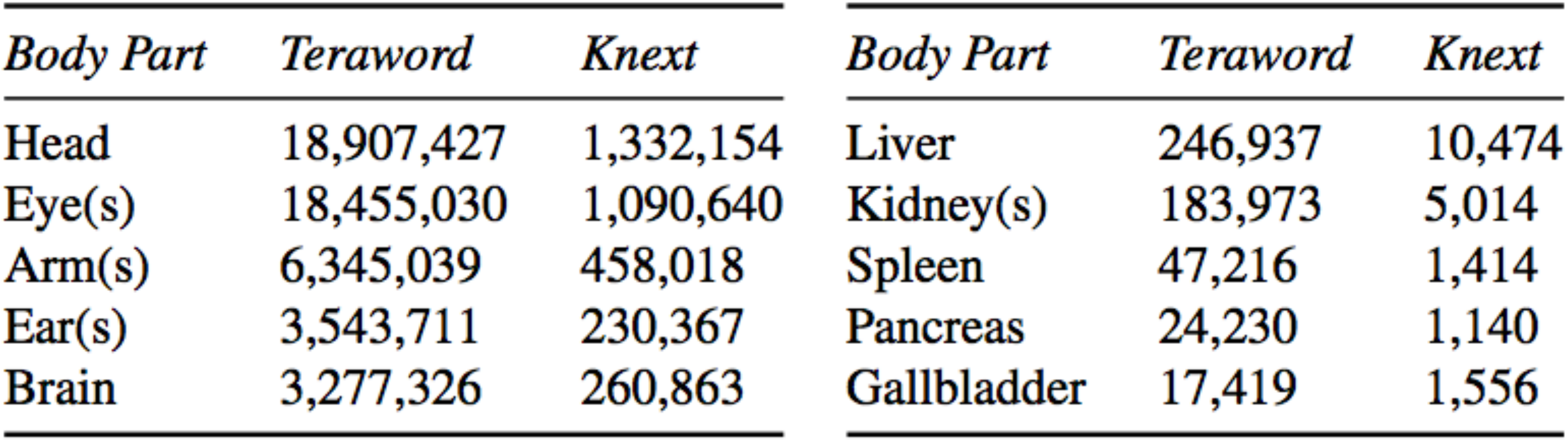
murder:exhale = 17:1

Word	Teraword	Knext
spoke	11,577,917	244,458
laughed	3,904,519	169,347
murdered	**2,843,529**	11,284
inhaled	984,613	4,412
breathed	725,034	34,912

Word	Teraword	Knext
hugged	610,040	10,378
blinked	390,692	20,624
was late	368,922	31,168
exhaled	**168,985**	3,490
was punctual	5,045	511

[Gordon et al. 2013]

Reporting bias in text

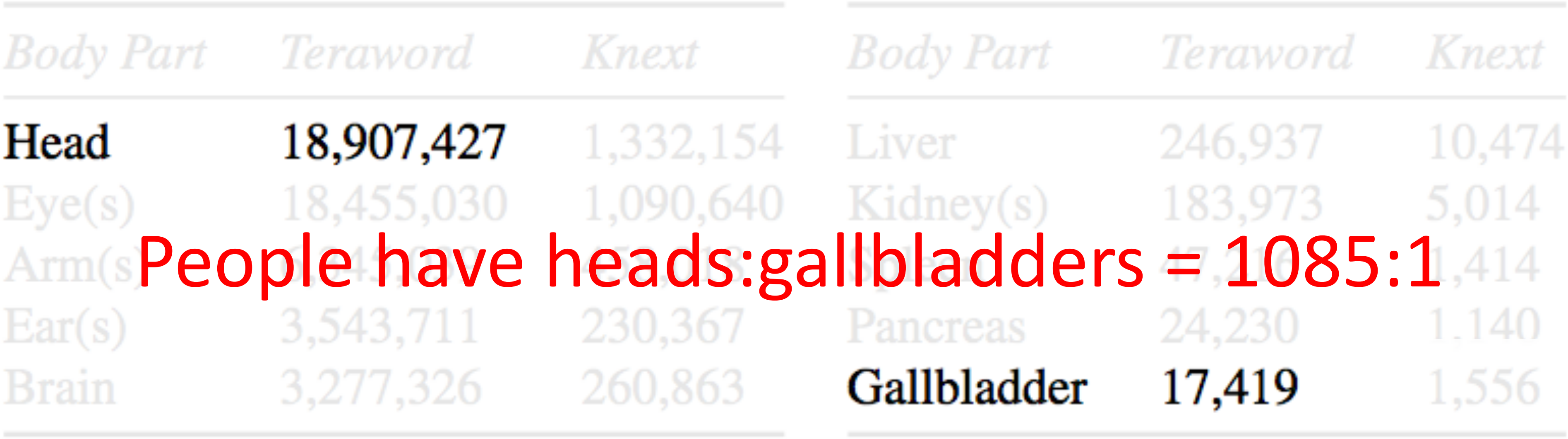


Body Part	*Teraword*	*Knext*
Head	18,907,427	1,332,154
Eye(s)	18,455,030	1,090,640
Arm(s)	6,345,039	458,018
Ear(s)	3,543,711	230,367
Brain	3,277,326	260,863

Body Part	*Teraword*	*Knext*
Liver	246,937	10,474
Kidney(s)	183,973	5,014
Spleen	47,216	1,414
Pancreas	24,230	1,140
Gallbladder	17,419	1,556

[Gordon et al. 2013]

Reporting bias in text



Body Part	*Teraword*	*Knext*	*Body Part*	*Teraword*	*Knext*
Head	**18,907,427**	1,332,154	Liver	246,937	10,474
Eye(s)	18,455,030	1,090,640	Kidney(s)	183,973	5,014
Arm(s)	[illegible]	[illegible]	[illegible]	[illegible]	1,414
Ear(s)	3,543,711	230,367	Pancreas	24,230	1,140
Brain	3,277,326	260,863	**Gallbladder**	**17,419**	1,556

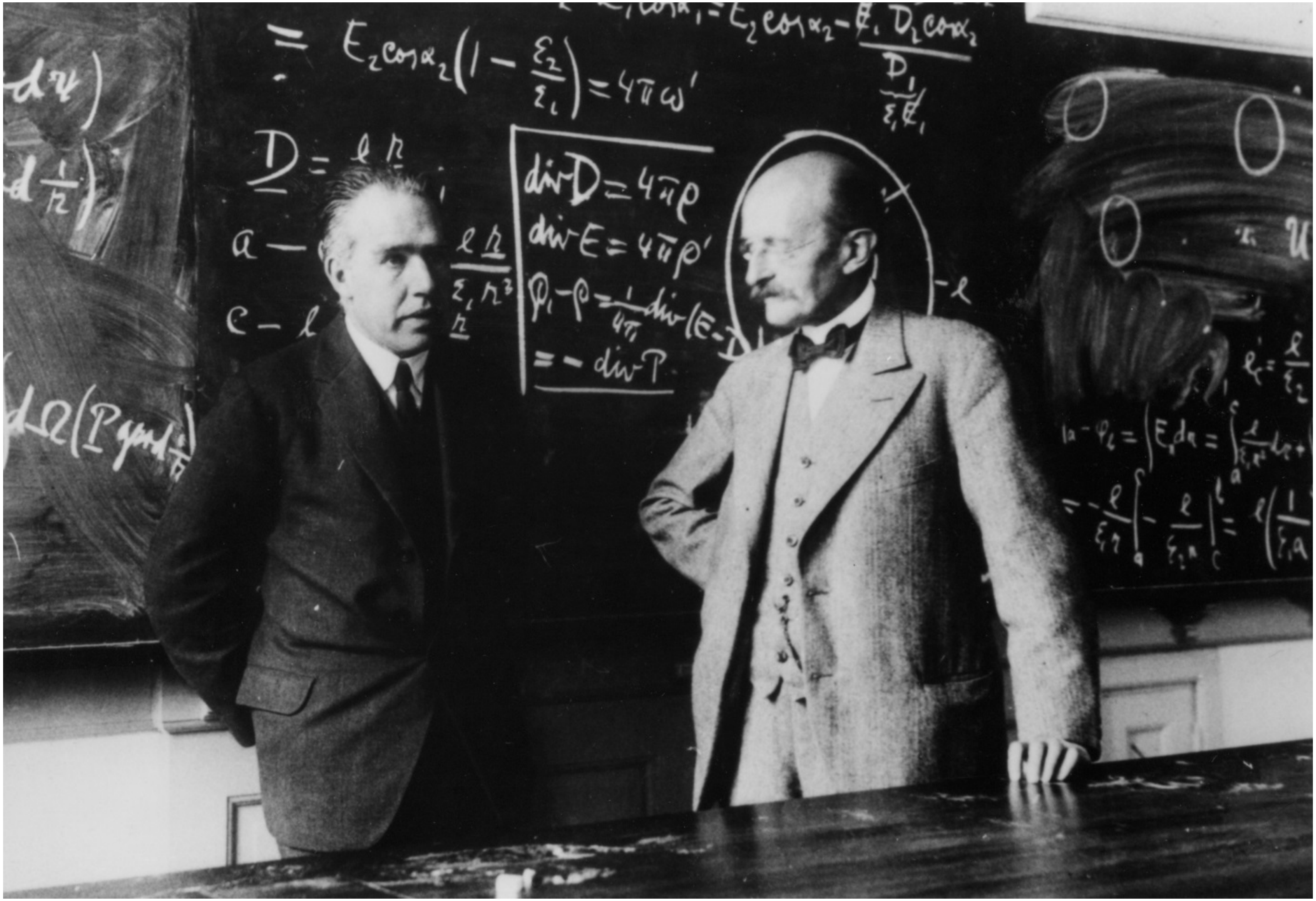
People have heads:gallbladders = 1085:1

[Gordon et al. 2013]

Learning Common Sense

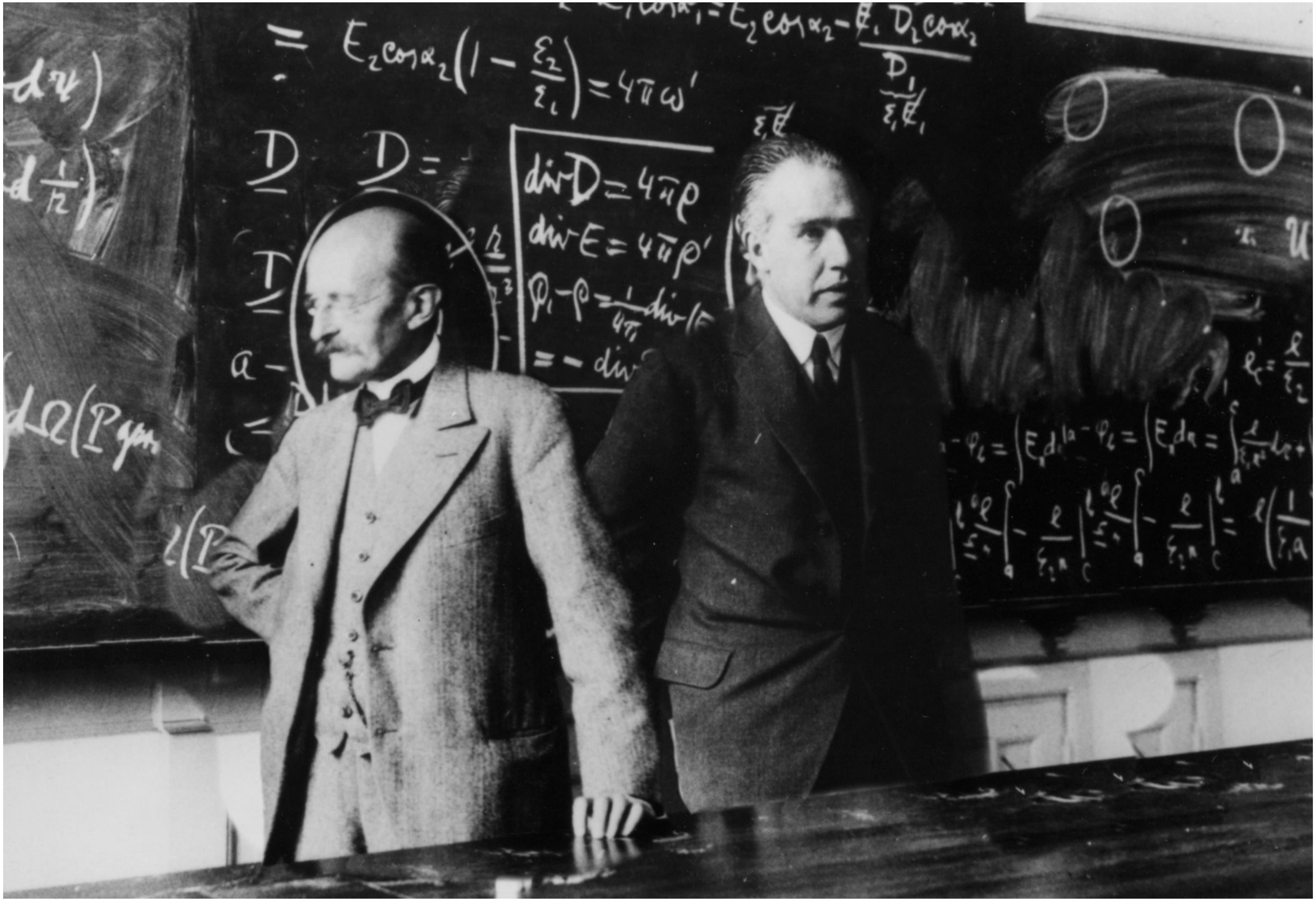
- Text
  - Reporting bias
- From structure in our visual world?

Two professors converse in front of a blackboard.


$= E_2 \cos\alpha_2 \left(1 - \frac{\varepsilon_2}{\varepsilon_1}\right) = 4\pi\omega'$
$\operatorname{div} D = 4\pi\rho$
$\operatorname{div} E = 4\pi\rho'$
$\rho_1 - \rho = \frac{1}{4\pi} \operatorname{div}(E-$
$= - \operatorname{div} P$

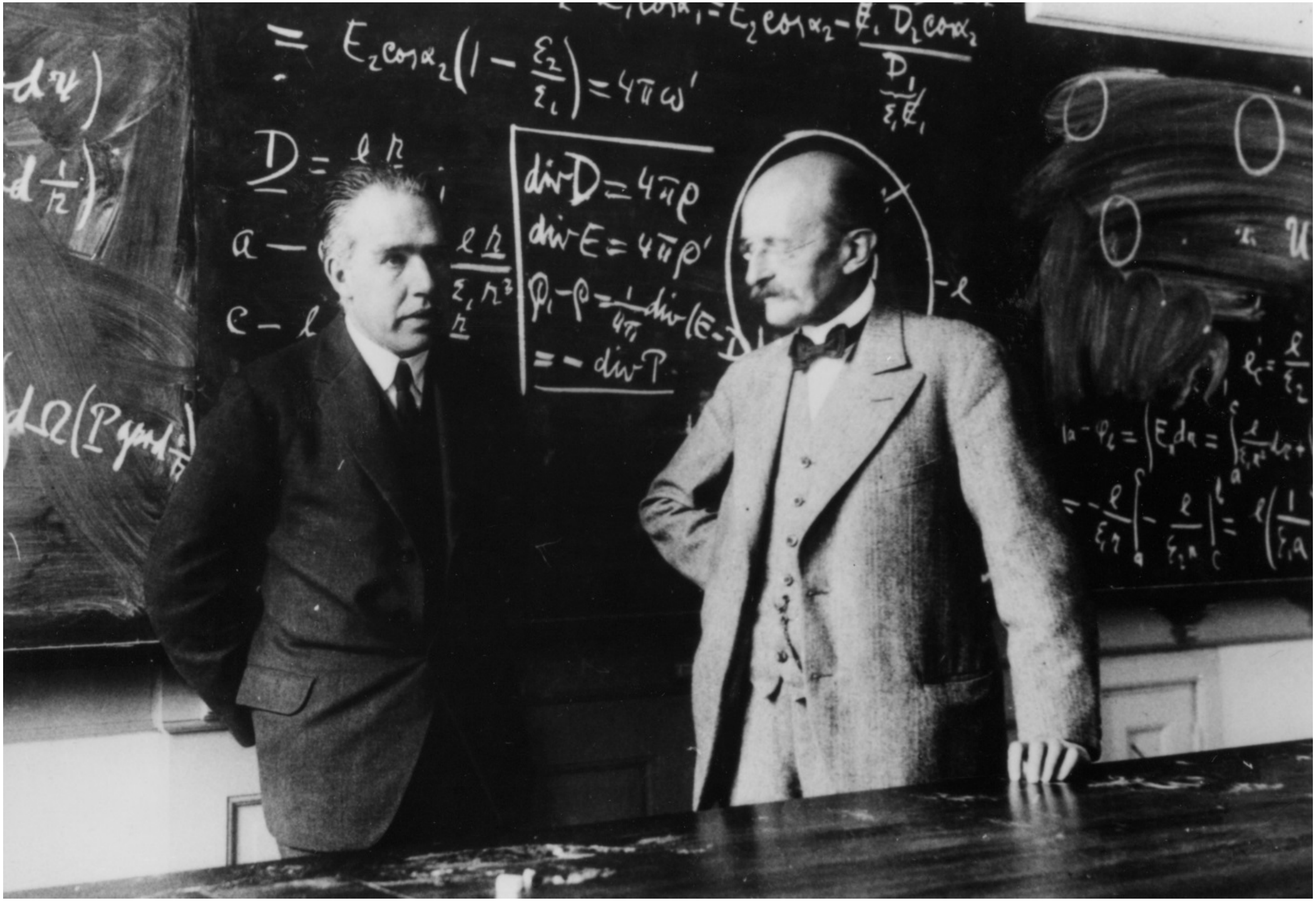
Slide credit: Larry Zitnick

Two professors stand in front of a blackboard.



Slide credit: Larry Zitnick

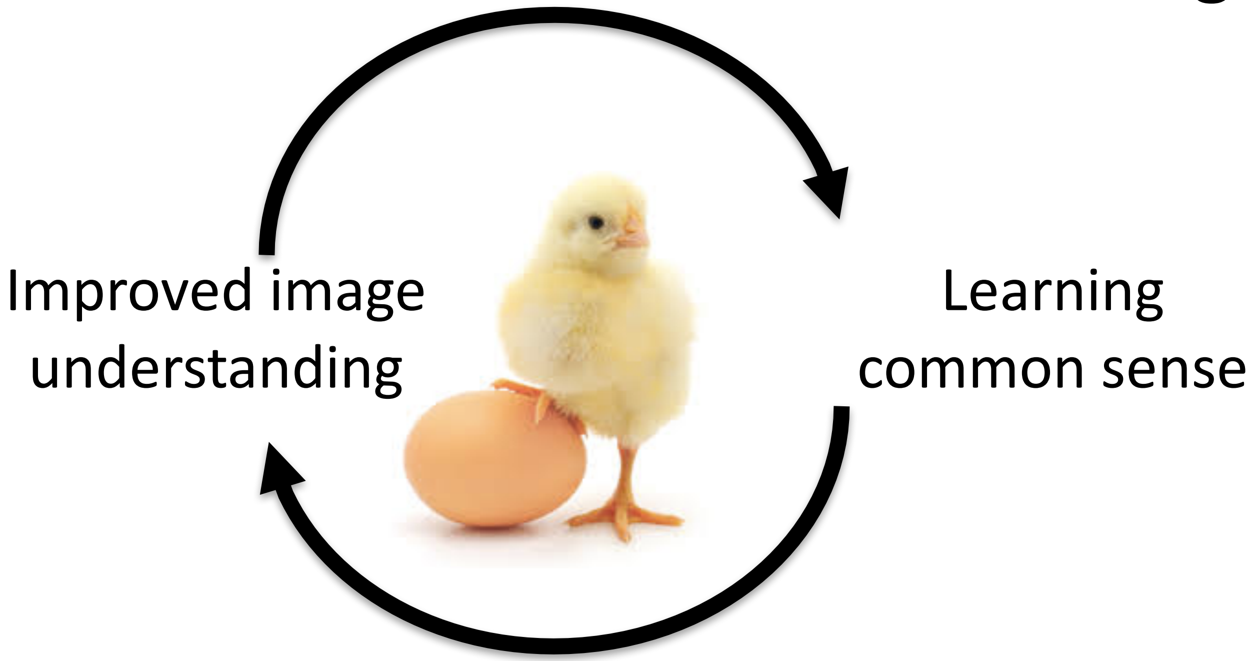
Two professors converse in front of a blackboard.



Slide credit: Larry Zitnick

Challenges

- Lacking visual density
- Annotations are expensive
- Computer vision doesn't work well enough

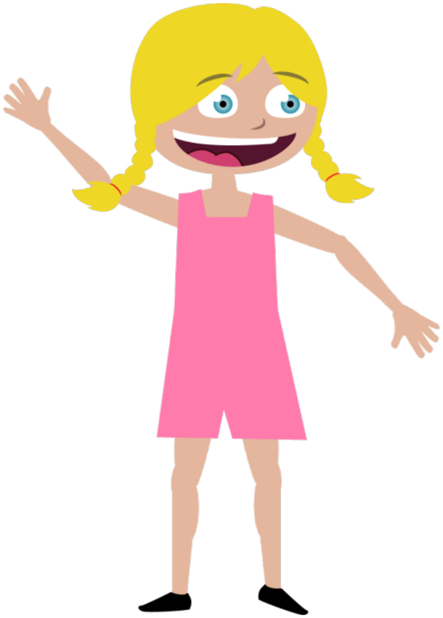


Improved image understanding

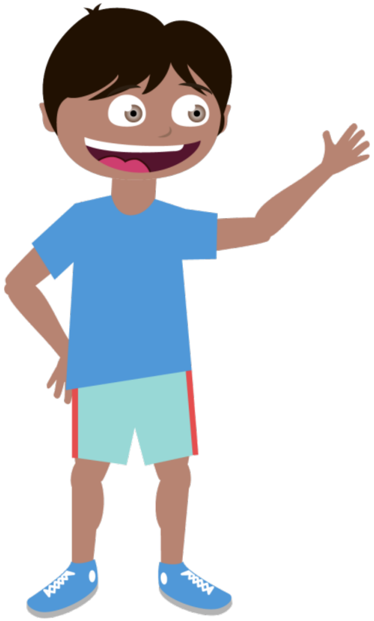
Learning common sense

Is photorealism necessary?

Slide credit: Larry Zitnick



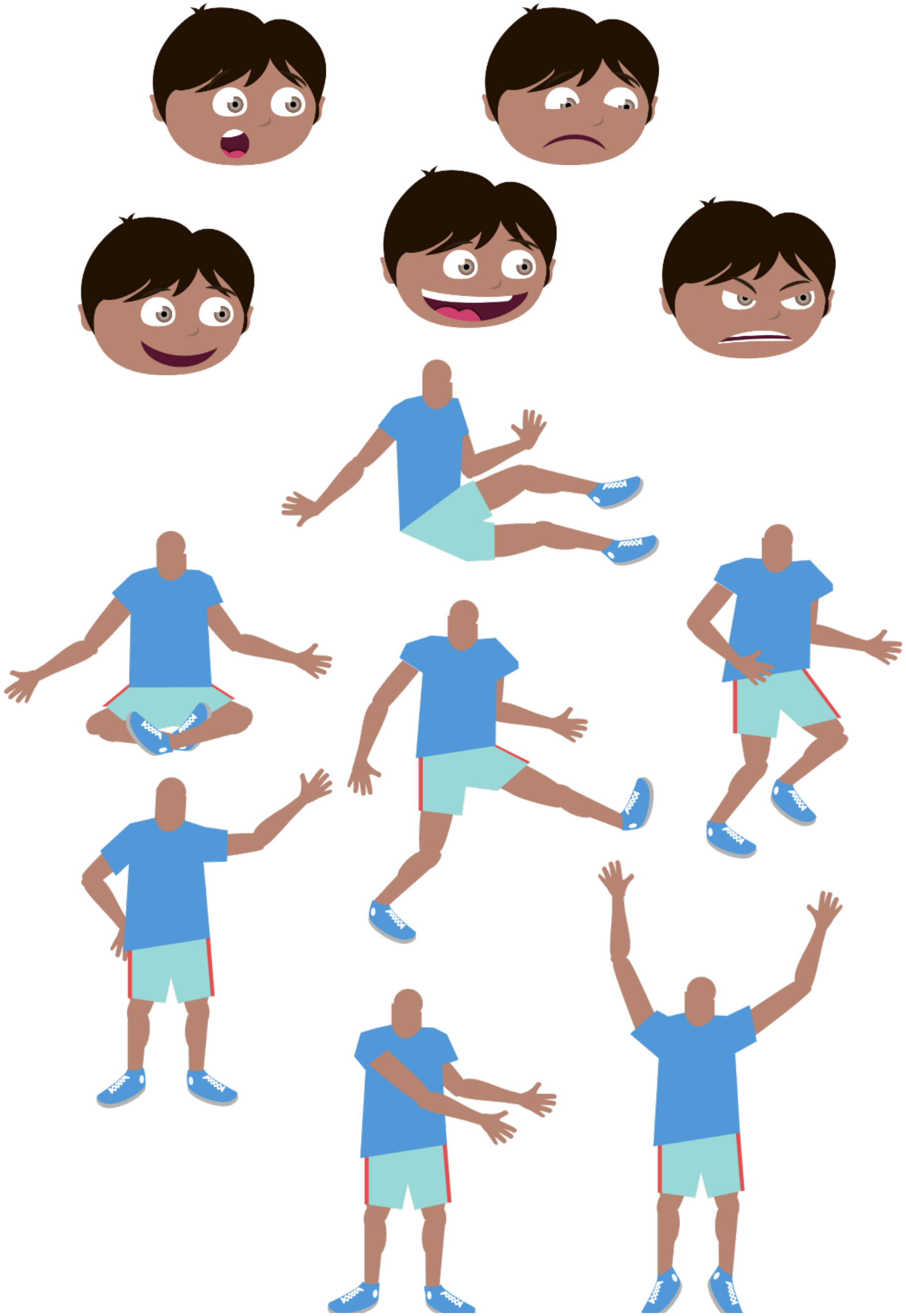
Jenny

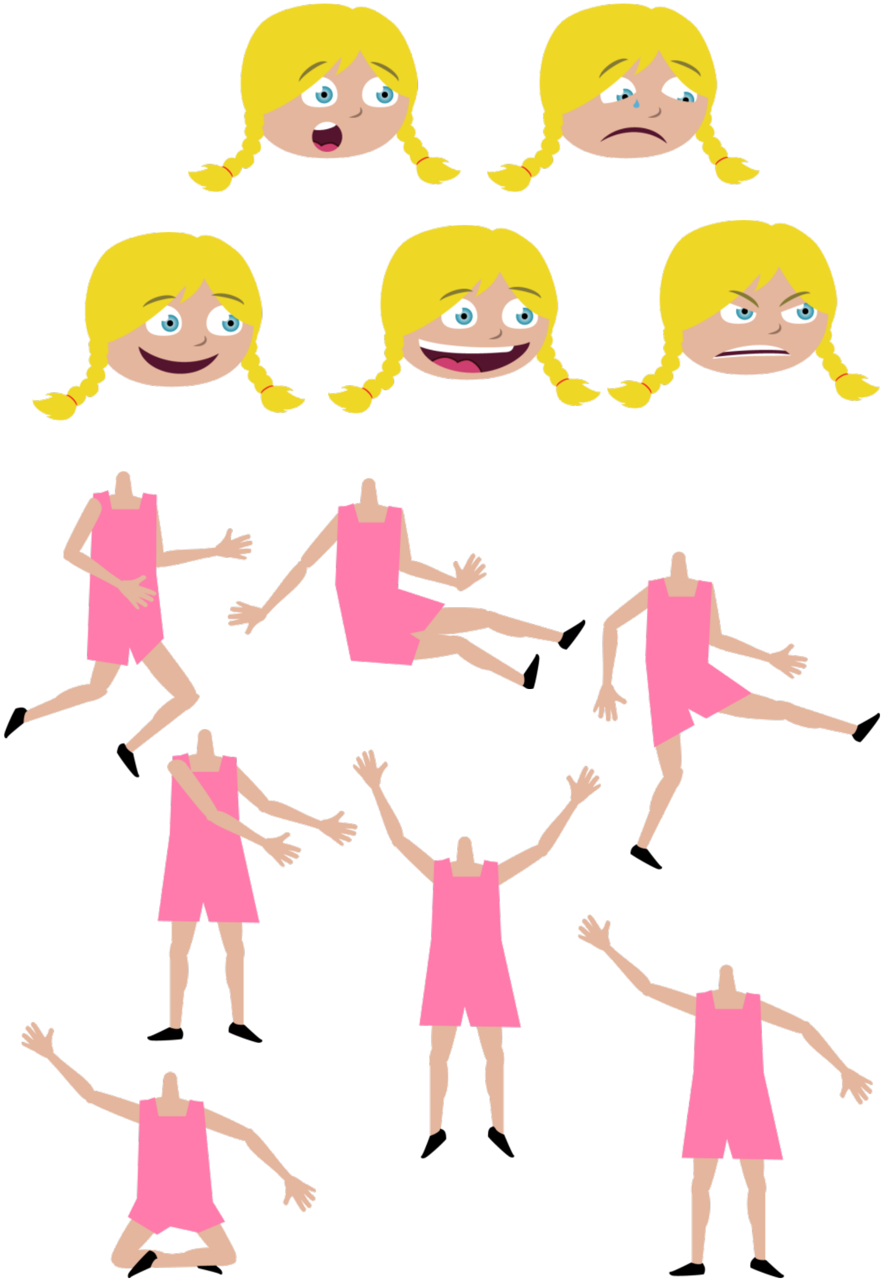


Mike

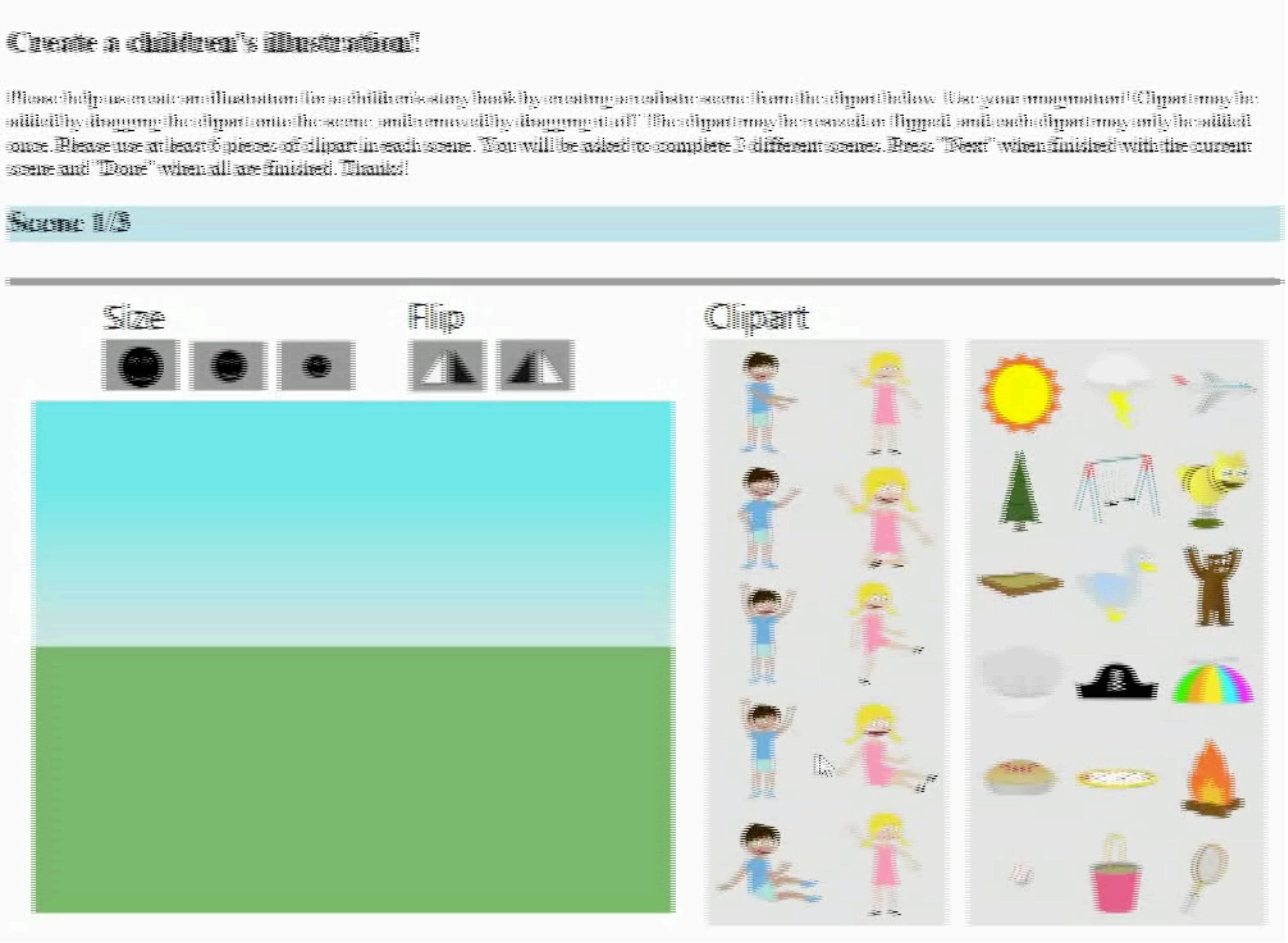
Slide credit: Larry Zitnick

Slide credit: Larry Zitnick





Slide credit: Larry Zitnick

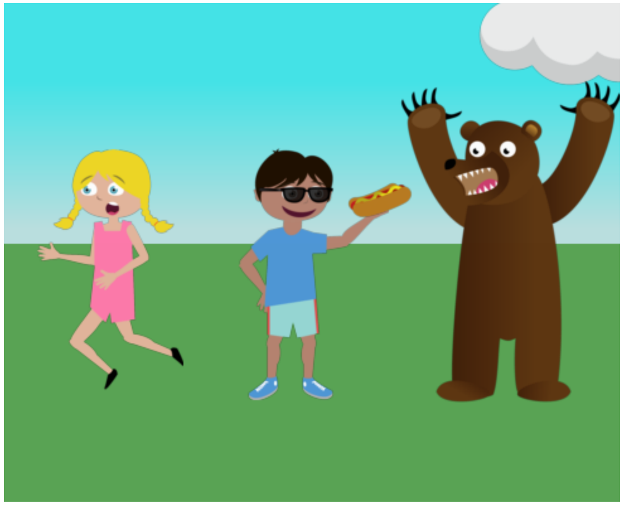


Create a children's illustration!

Please help us create an illustration for a children's story book by creating a realistic scene from the clipart below. Use your imagination! Clipart may be added by dragging the clipart onto the scene, and removed by dragging it off. The clipart may be resized or flipped, and each clipart may only be added once. Please use at least 6 pieces of clipart in each scene. You will be asked to complete 3 different scenes. Press "Next" when finished with the current scene and "Done" when all are finished. Thanks!

Scene 1/3

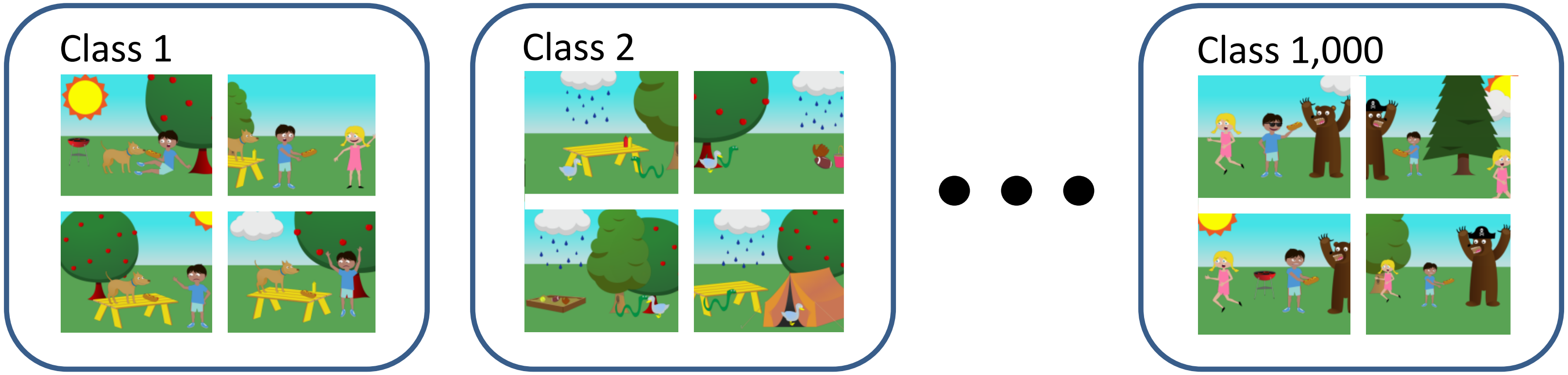
Mike fights off a bear by giving him a hotdog while Jenny runs away.



Slide credit: Larry Zitnick

Dataset

1,000 classes of semantically similar scenes:



Class 1

Class 2

Class 1,000

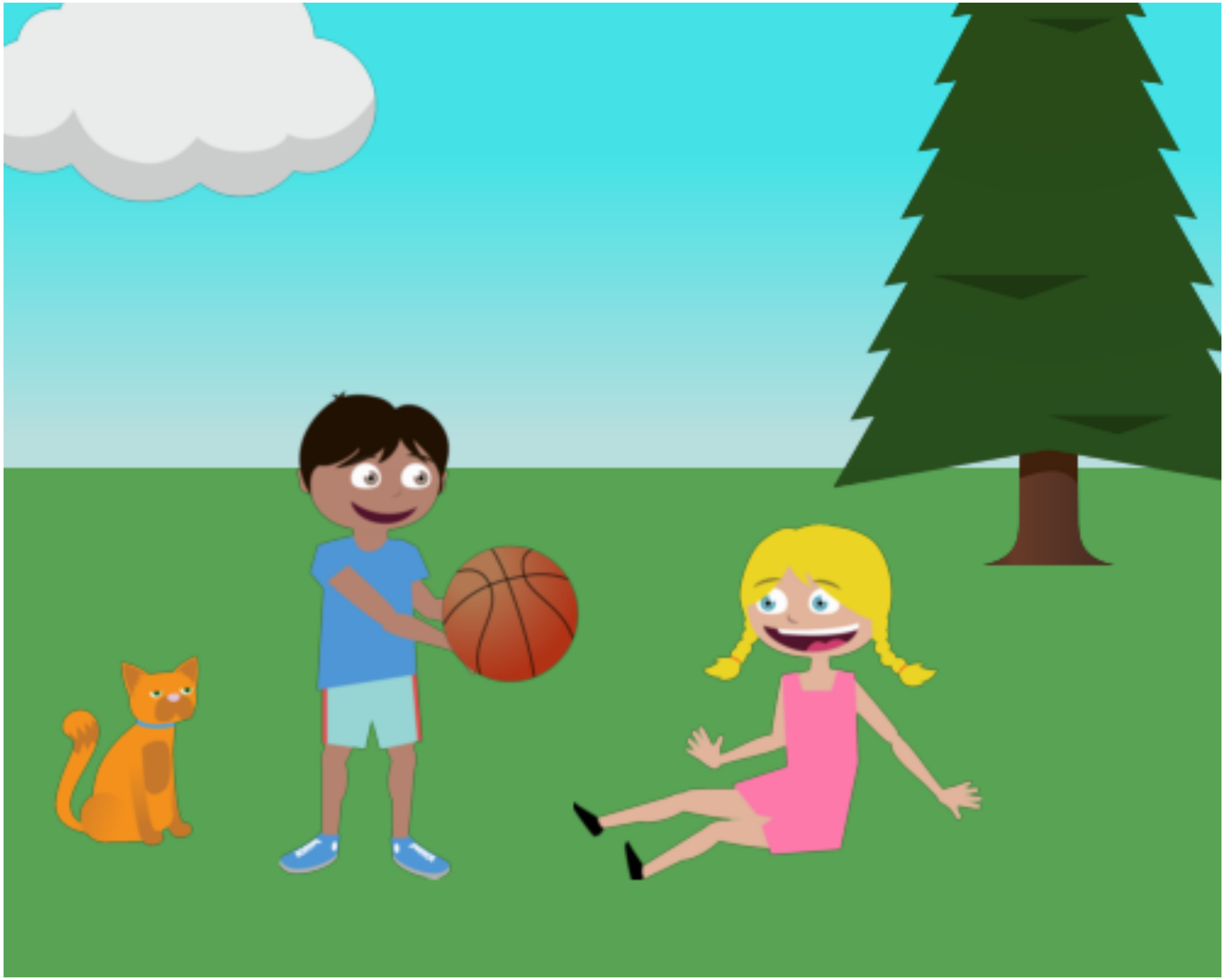
1,000 classes x 10 scenes per class = 10,000 scenes

[Zitnick and Parikh, CVPR 2013, Oral]

Dataset online

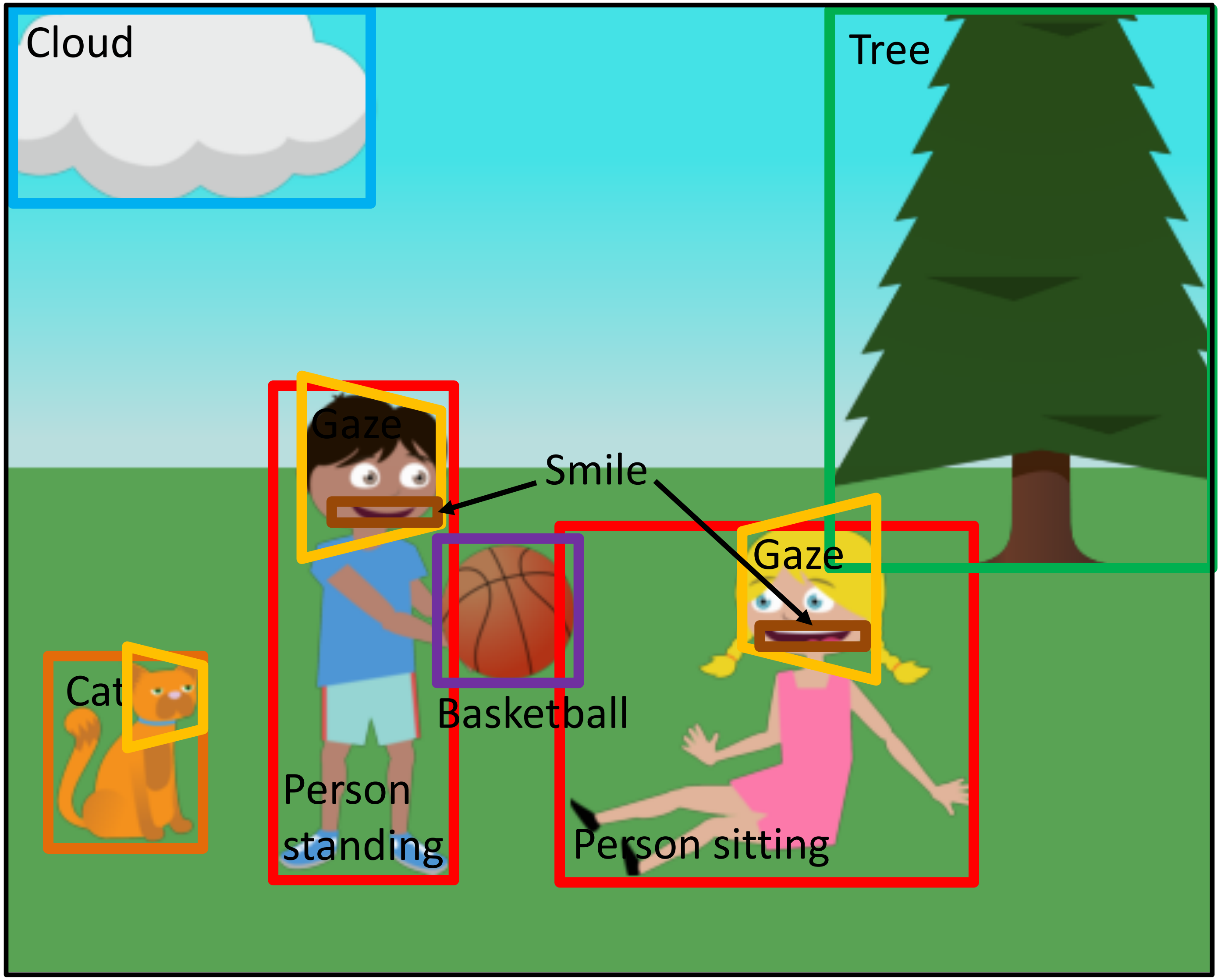
Slide credit: Larry Zitnick

Visual Features



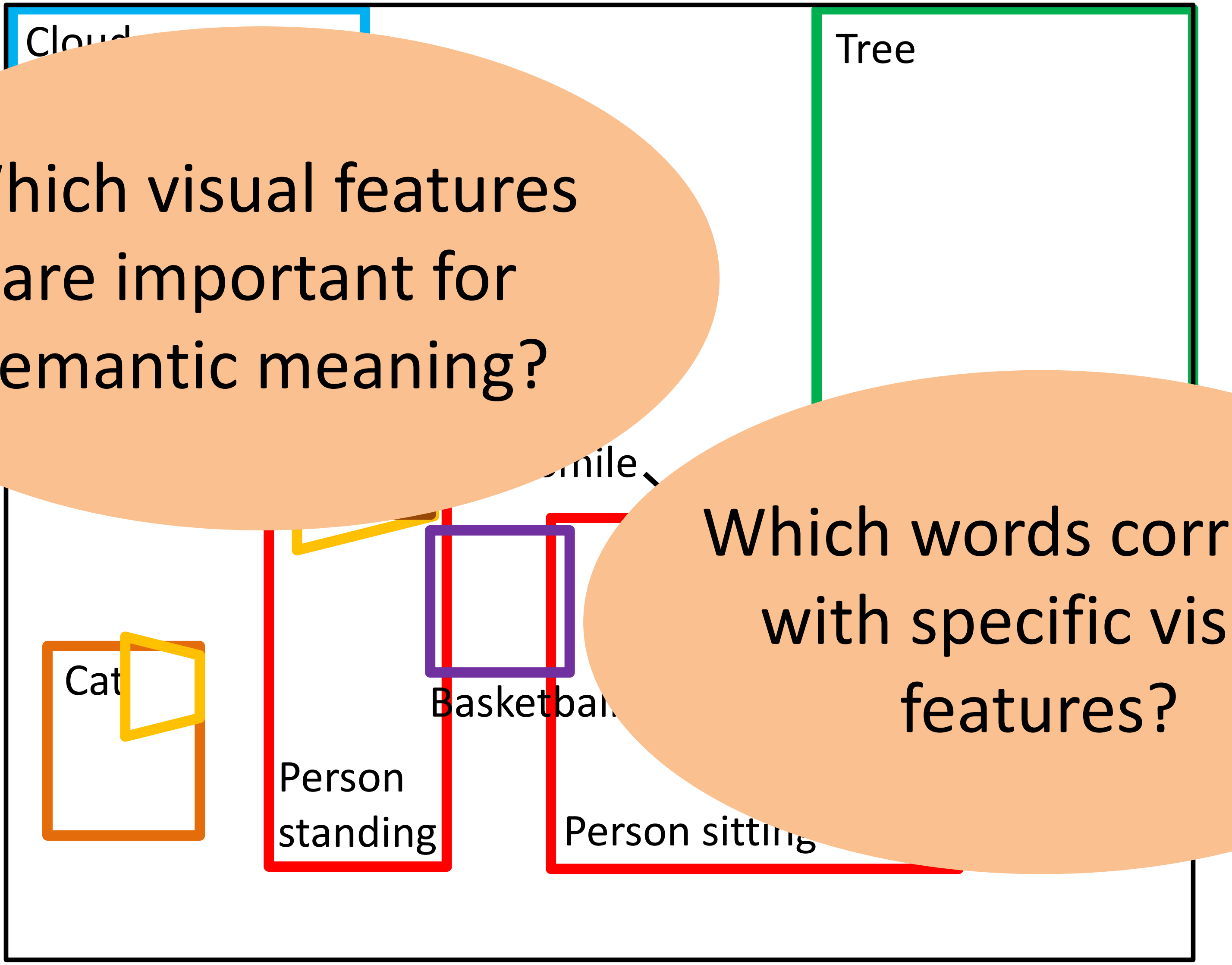
Slide credit: Larry Zitnick

Visual Features



Slide credit: Larry Zitnick

Visual Features



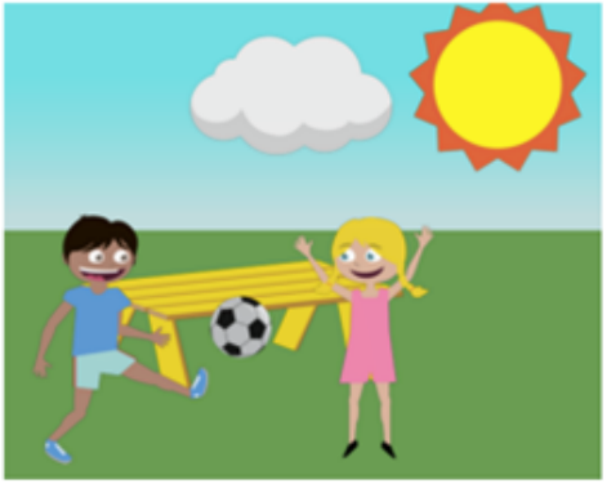
Which visual features are important for semantic meaning?

Which words correlate with specific visual features?

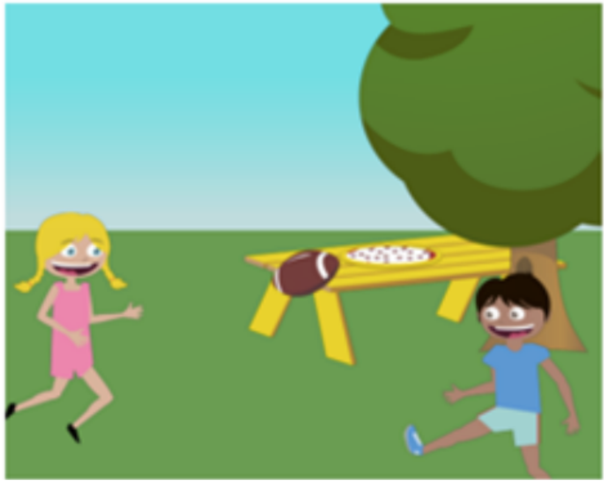
Generate Scenes

Input: Jenny is catching the ball. Mike is kicking the ball. The table is next to the tree.

Tuples: <<Jenny>,<catch>,<ball>> <<Mike>,<kick>,<ball>> <<table>,<be>,<>>



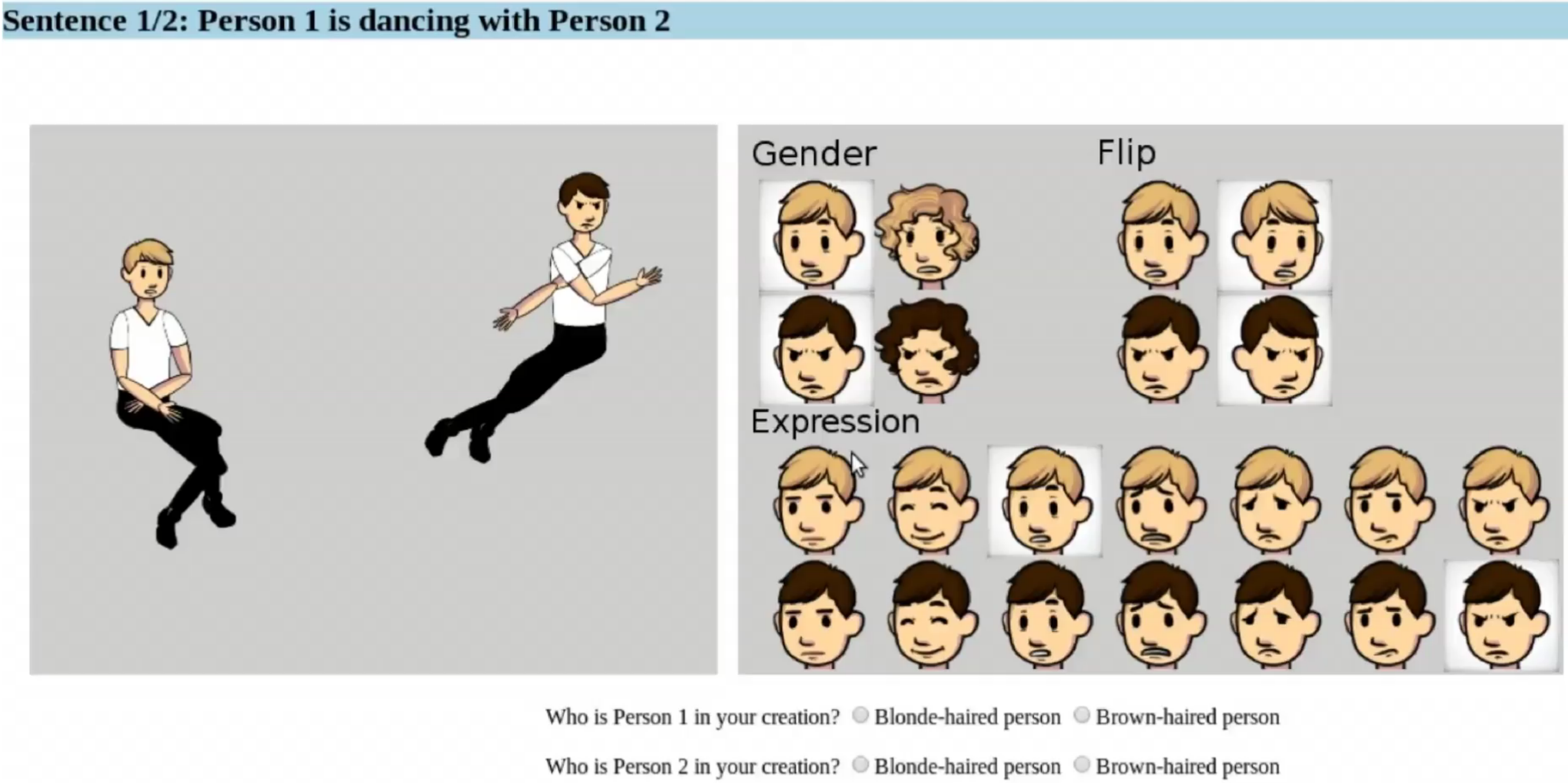
Automatically Generated



Human Generated

[Zitnick, Parikh and Vanderwende, ICCV 2013]

Learning Fine-grained Interactions



Sentence 1/2: Person 1 is dancing with Person 2

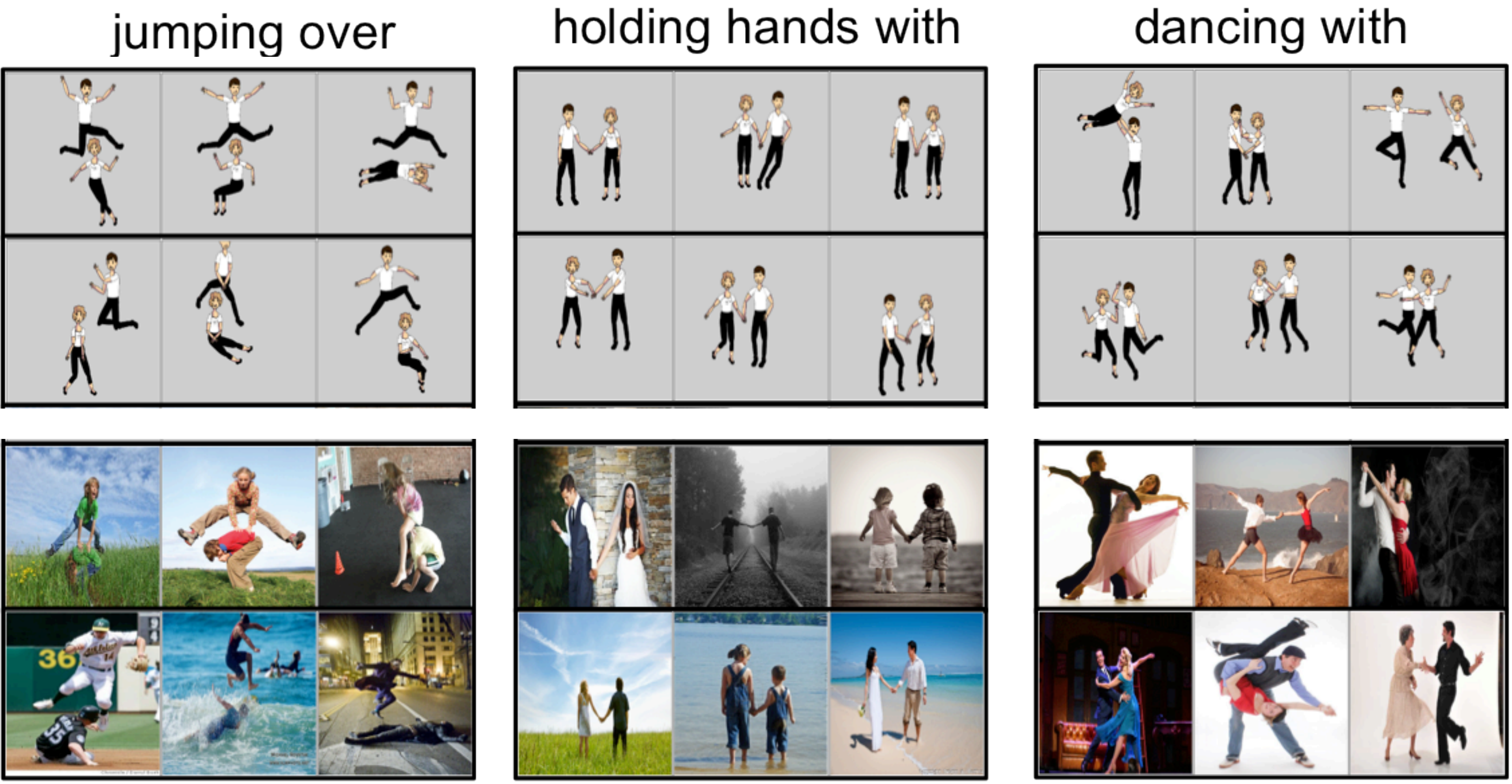
Who is Person 1 in your creation? ○ Blonde-haired person ○ Brown-haired person

Who is Person 2 in your creation? ○ Blonde-haired person ○ Brown-haired person

3x

[Antol, Zitnick and Parikh, ECCV 2014]

Learning Fine-grained Interactions



jumping over

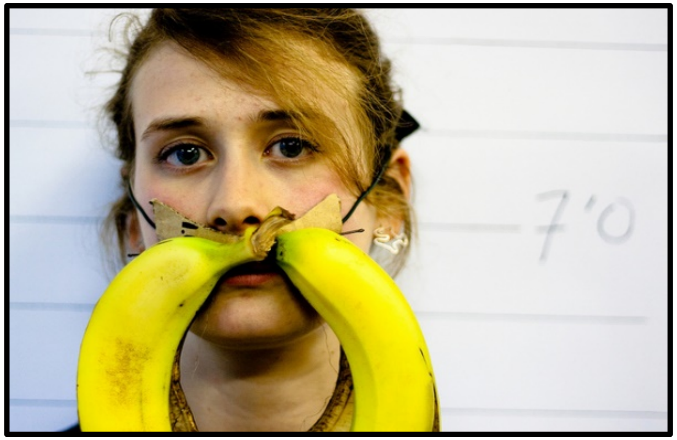
holding hands with

dancing with

Train on clipart, test on real

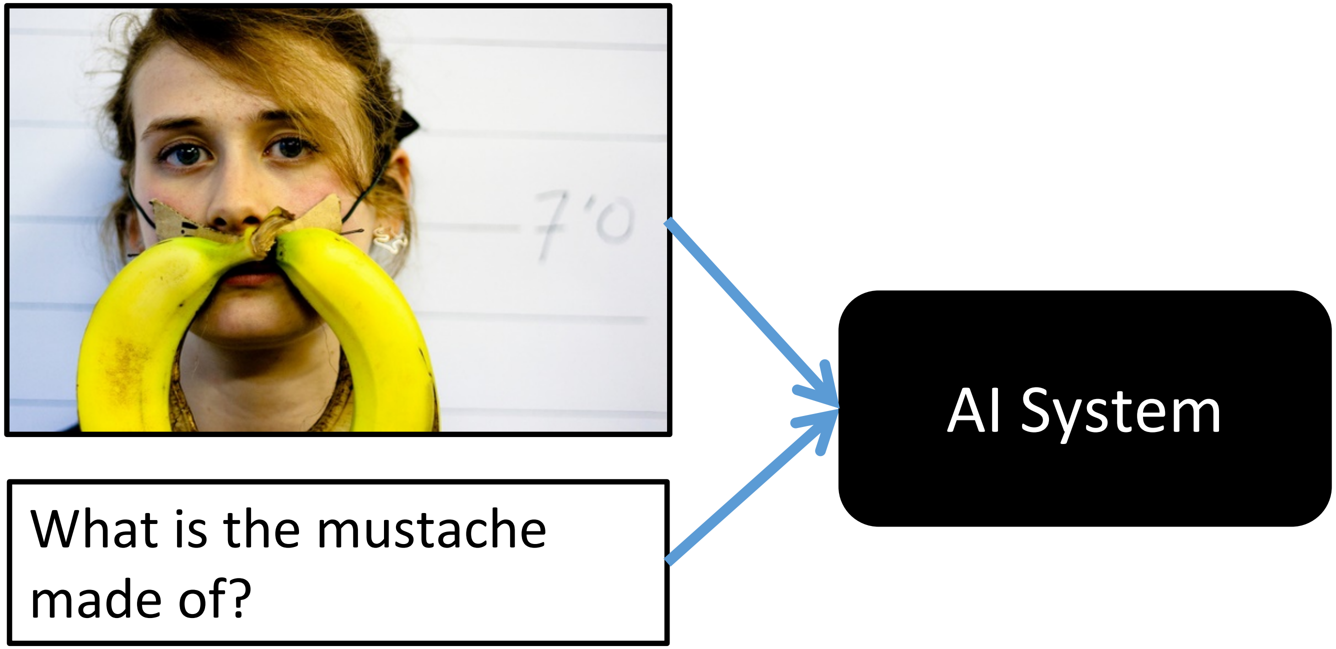
Visual Question Answering (VQA)

Visual Question Answering (VQA)



What is the mustache made of?

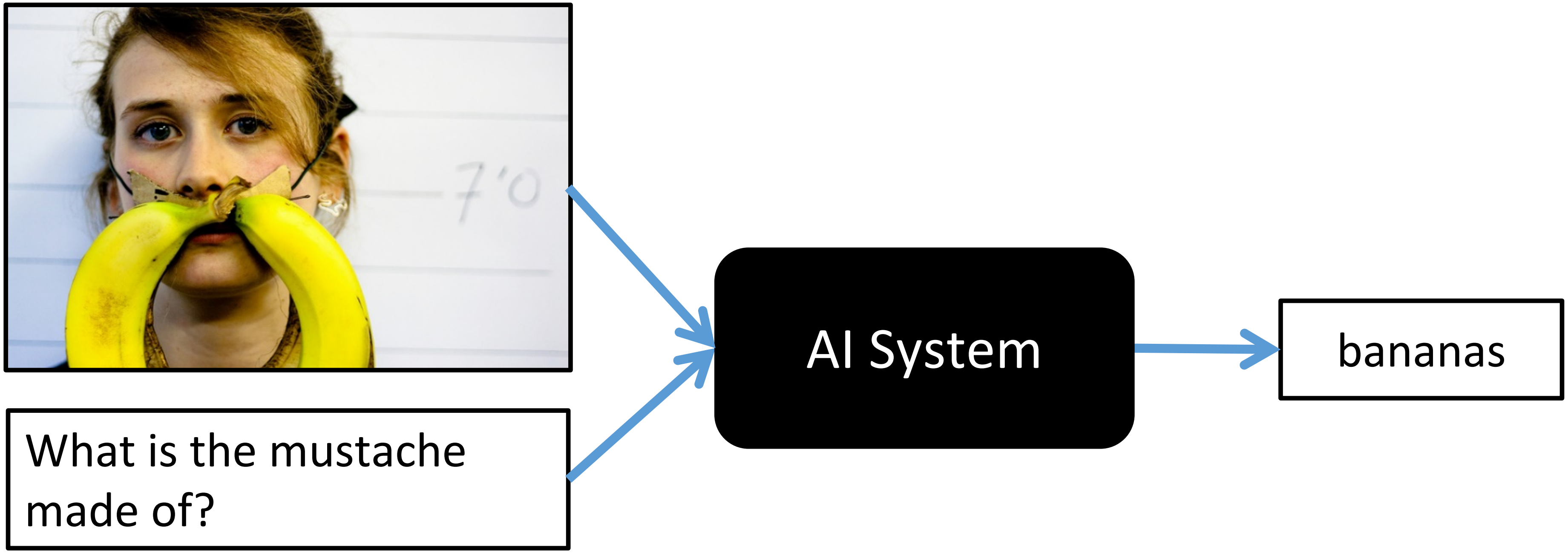
Visual Question Answering (VQA)



What is the mustache made of?

AI System

Visual Question Answering (VQA)

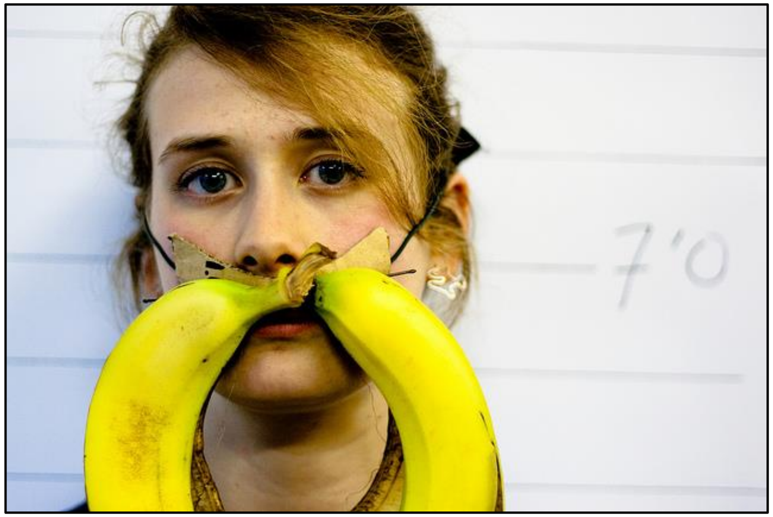


What is the mustache made of?

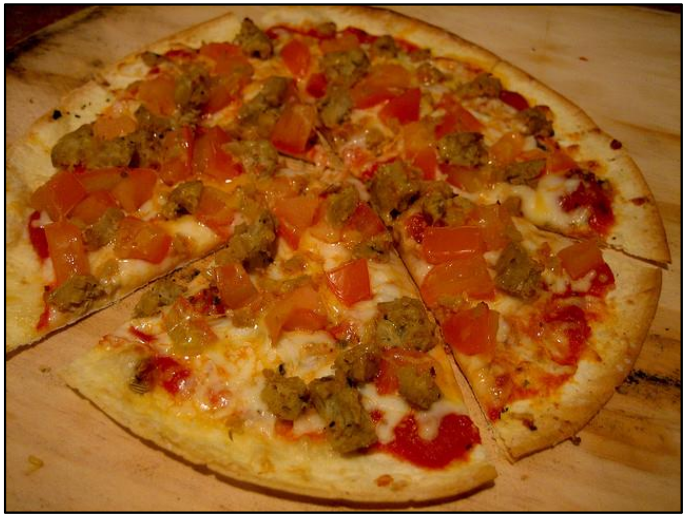
AI System

bananas

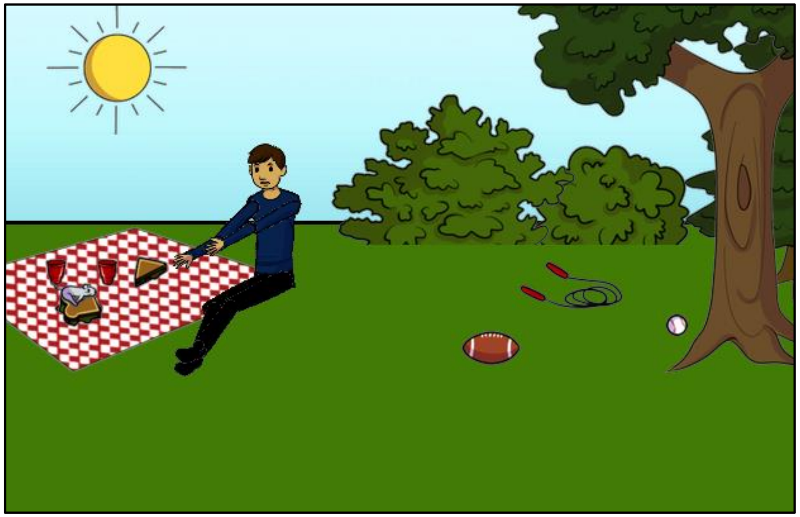
Visual Question Answering (VQA)



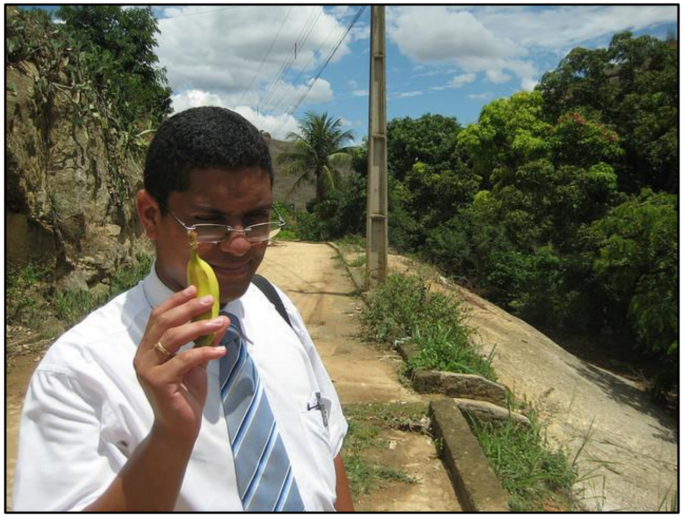
What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

Language Bias

Is there a clock … ?

‘yes’ 98%

……

Is the man wearing glasses … ?

‘yes’ 94%

……

Are the lights on … ?

‘yes’ 85%

……

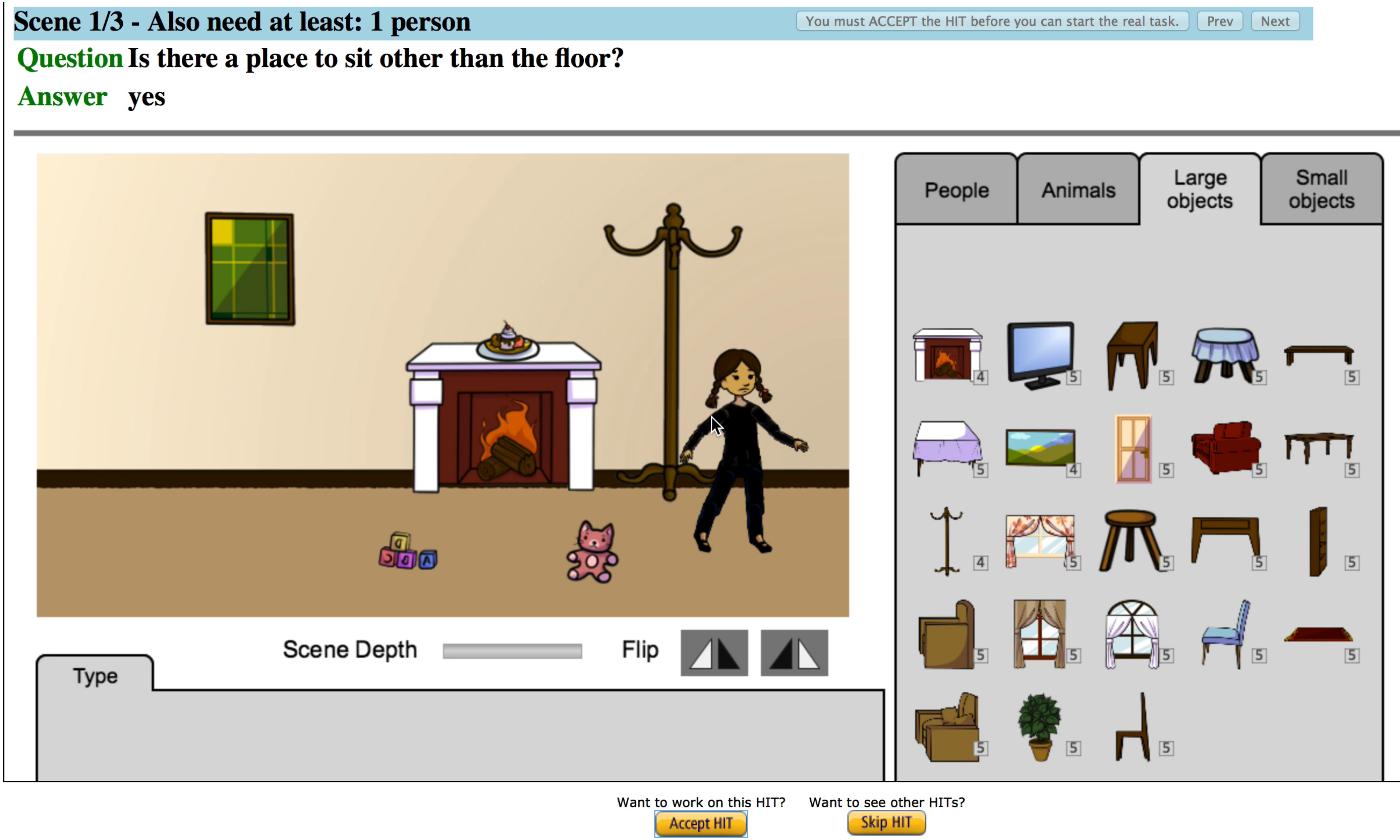
Do you see a … ?

‘yes’ 87%

……

Slide credit: Yash Goyal and Peng Zhang

Removing Language Priors



Scene 1/3 - Also need at least: 1 person

You must ACCEPT the HIT before you can start the real task.

Question **Is there a place to sit other than the floor?**

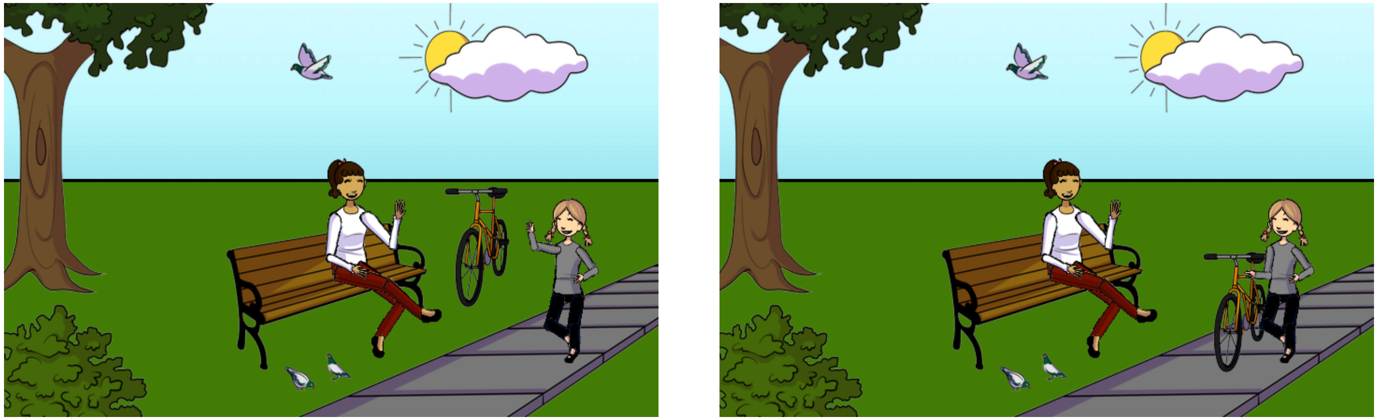
Answer yes

Want to work on this HIT?

Want to see other HITs?

Slide credit: Yash Goyal and Peng Zhang

Removing Language Priors



Answer: No

Answer: Yes

complementary scenes

Question: Is the girl walking the bike?

[Zhang, Goyal, Summers-Stay, Batra, Parikh, CVPR 2016]

Slide credit: Yash Goyal and Peng Zhang

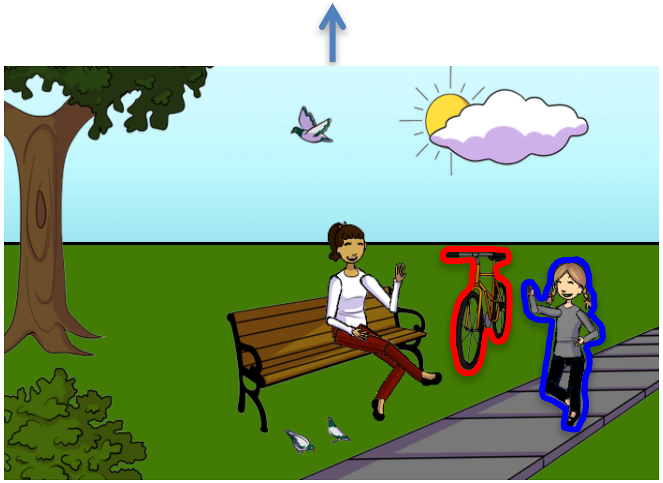
Classifying a pair of complementary scenes

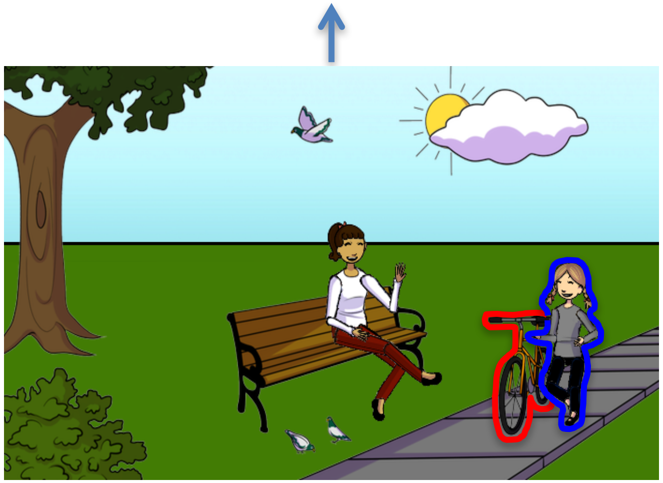
	Training set	
	Unbalanced	Balanced
Blind (no image features)		
Holistic image features		

Answering Binary Questions

Answer: No

Answer: Yes





Tuple: <girl, walking, bike>

Question: Is the girl walking the bike?

[Zhang, Goyal, Summers-Stay, Batra, Parikh, CVPR 2016]

Slide credit: Yash Goyal and Peng Zhang

Classifying a pair of complementary scenes

	Training set	
	Unbalanced	Balanced
Blind (no image features)	0	0
Holistic image features	03.20	23.13
Attention-based image features		

Abstract Scenes

- Learning by playing

- Fully annotated visual data

- Allow full control over the distribution and density of data
  - to learn from
  - to evaluate on

Commonsense Tasks

- Text-based tasks

Key idea

- Imagine the scene behind the text
- Reason about the visual interpretation of the text, not just the text alone

Commonsense Tasks

- Assess plausibility of relations
  - `man holds meal`
  - `tree grows in table`

[Vedantam, Lin, Batra, Zitnick, and Parikh, ICCV 2015]

Fill-in-the-blank:

Mike is having lunch when he sees a bear.
______________.

A. Mike orders a pizza.
B. Mike hugs the bear.
C. Bears are mammals.
D. Mike tries to hide.

[Lin and Parikh, CVPR 2015]

Approach: Imagination

________________.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

A. There is a tree near a table.

B. The brown dog is standing next to Mike.

C. The sun is in the sky.

D. Jenny is standing dangerously on the swing.

Slide credit: Xiao Lin

Approach: Imagination

There is a tree near a table.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

A

A. There is a tree near a table.

B. The brown dog is standing next to Mike.

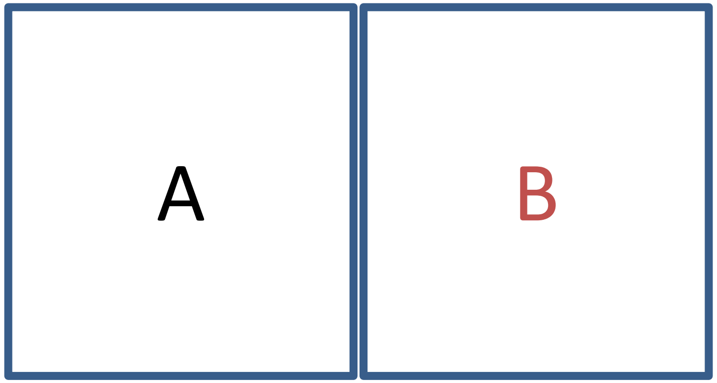
C. The sun is in the sky.

D. Jenny is standing dangerously on the swing.

Slide credit: Xiao Lin

Approach: Imagination

The brown dog is standing next to Mike.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.



A. There is a tree near a table.

B. The brown dog is standing next to Mike.

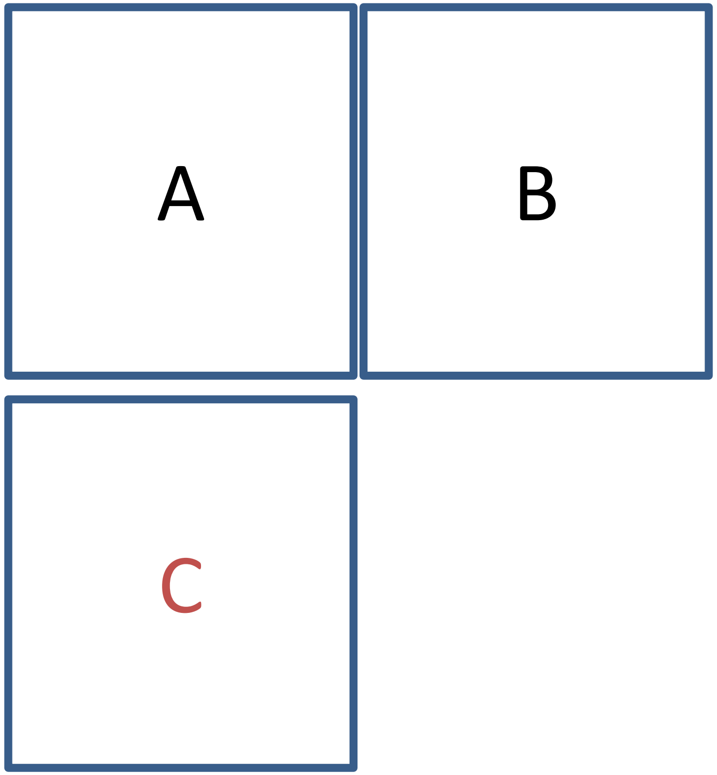
C. The sun is in the sky.

D. Jenny is standing dangerously on the swing.

Slide credit: Xiao Lin

Approach: Imagination

The sun is in the sky.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.



A. There is a tree near a table.

B. The brown dog is standing next to Mike.

C. The sun is in the sky.

D. Jenny is standing dangerously on the swing.

Slide credit: Xiao Lin

Approach: Imagination

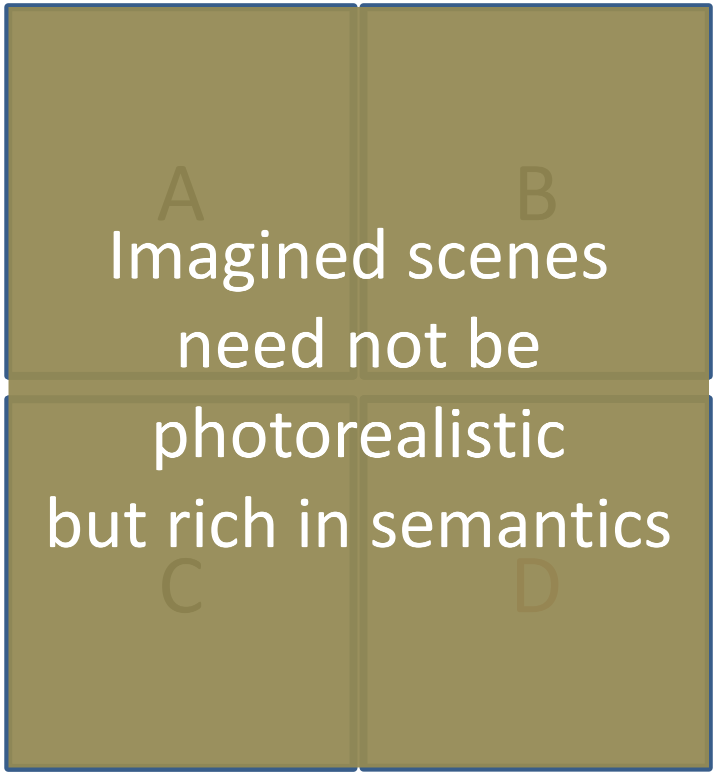
Jenny is standing dangerously on the swing.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

A. There is a tree near a table.

B. The brown dog is standing next to Mike.

C. The sun is in the sky.

D. Jenny is standing dangerously on the swing.



Imagined scenes need not be photorealistic but rich in semantics

Slide credit: Xiao Lin

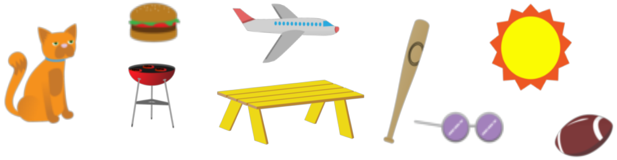
Approach: Imagination

- Clipart Visual World

  [CVPR 2013]

  – Two children playing in the park

  – 58 objects



  – 7 poses and 5 expressions

Slide credit: Xiao Lin

Approach: Imagination

- Scene generation given description [ICCV 2013]

There is a tree near a table.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

Slide credit: Xiao Lin

Approach: Imagination

- Scene generation given description [ICCV 2013]
- Semantic parsing into tuples

<Tree, near table>
<Mike, wear, cap>
<Mike, tell, get> <Jenny, get off, swing>

Slide credit: Xiao Lin

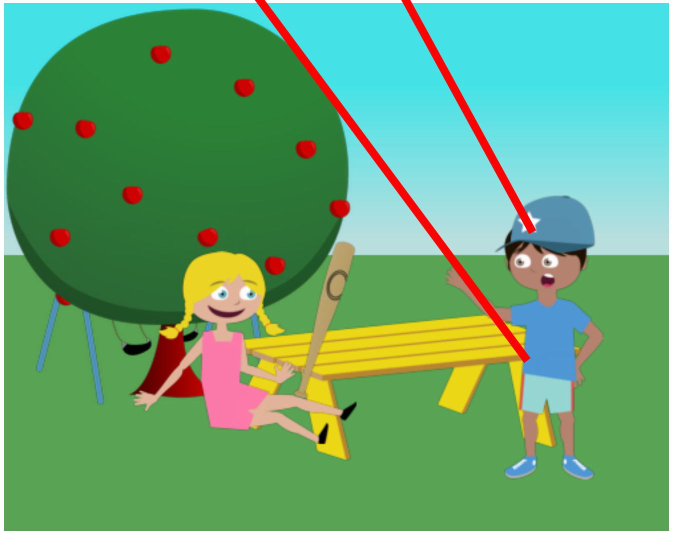
Approach: Imagination

- Scene generation given description [ICCV 2013]
- Semantic parsing into tuples
- Scene generation

Conditional Random Field (CRF)

$$p(\text{objects}|\text{tuples})$$

<Tree, near table>
<Mike, wear, cap>
<Mike, tell, get> <Jenny, get off, swing>



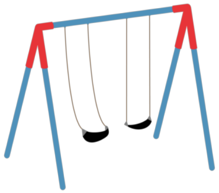
Slide credit: Xiao Lin

Approach: Imagination

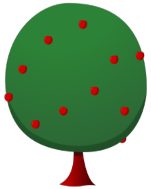
- Scene generation given description [ICCV 2013]
- Semantic parsing into tuples
- Scene generation CRF

Which objects are present

mike
he
ground
shirt

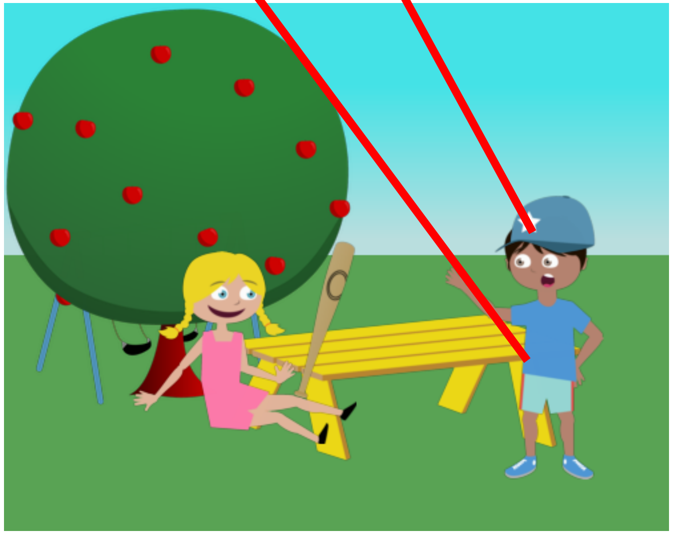


swing
set
swingset
playground



tree
apple
fruit
grab

<Tree, near table>
<Mike, wear, cap>
<Mike, tell, get> <Jenny, get off, swing>



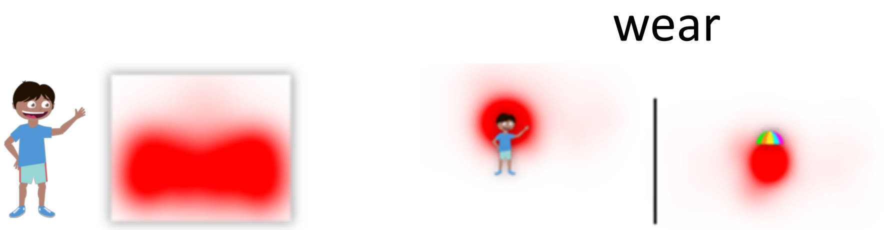
Slide credit: Xiao Lin

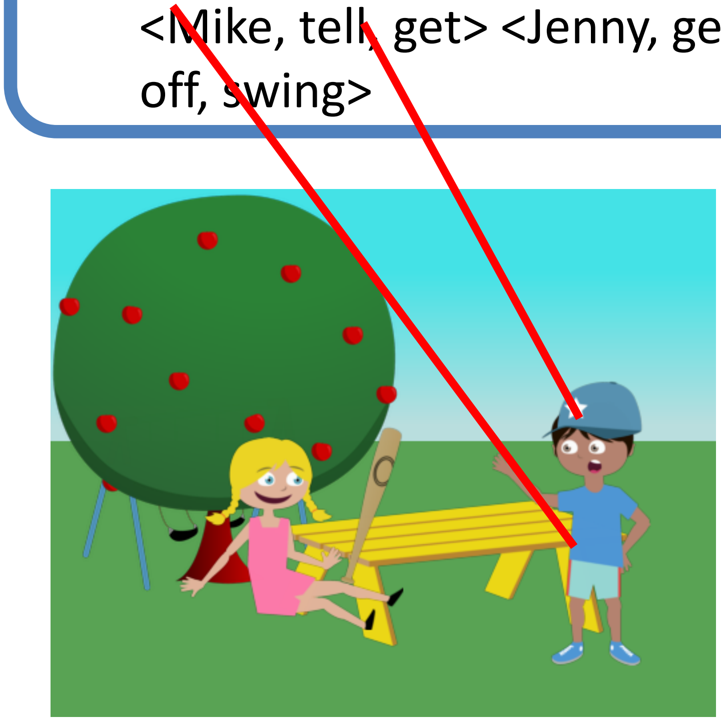
Approach: Imagination

- Scene generation given description [ICCV 2013]
- Semantic parsing into tuples
- Scene generation CRF

Where objects are

<Tree, near table>
<Mike, wear, cap>
<Mike, tell, get> <Jenny, get off, swing>



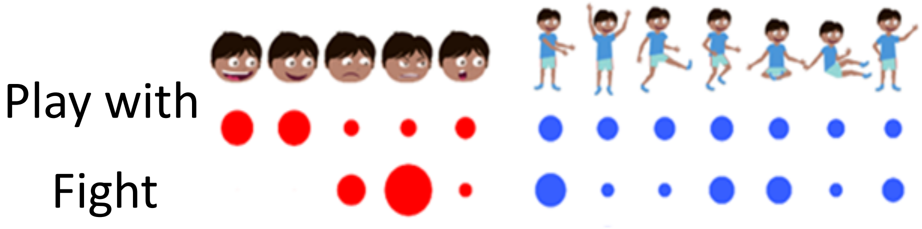


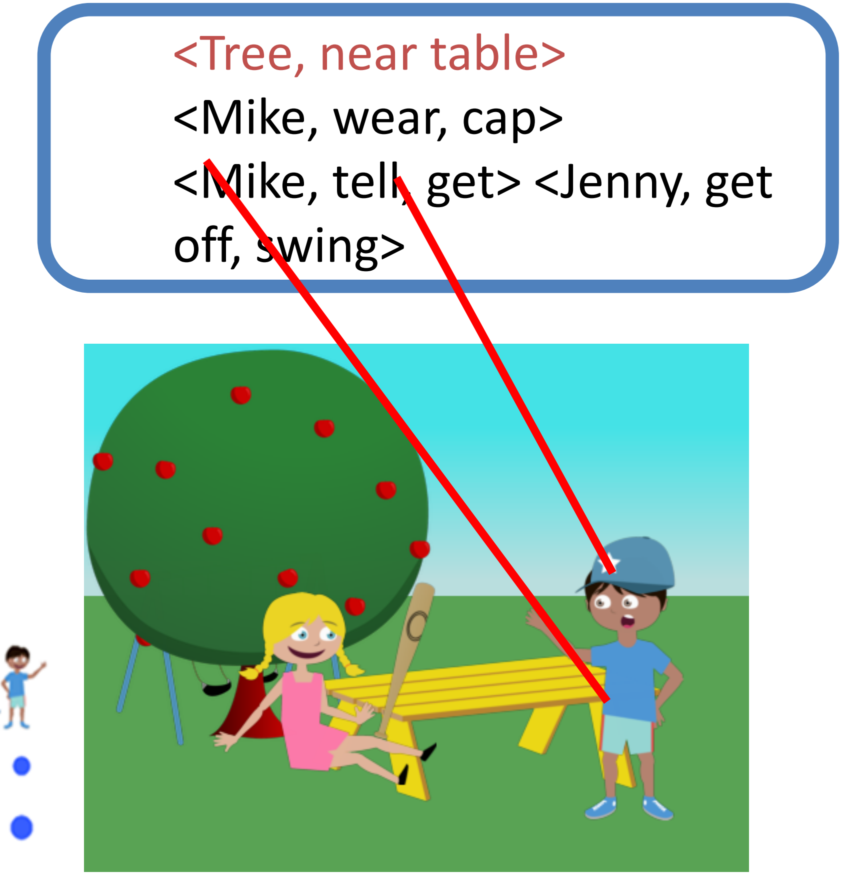
Slide credit: Xiao Lin

Approach: Imagination

- Scene generation given description [ICCV 2013]
- Semantic parsing into tuples
- Scene generation CRF

  What are the poses and expressions





<Tree, near table>
<Mike, wear, cap>
<Mike, tell, get> <Jenny, get off, swing>

Slide credit: Xiao Lin

Approach: Imagination

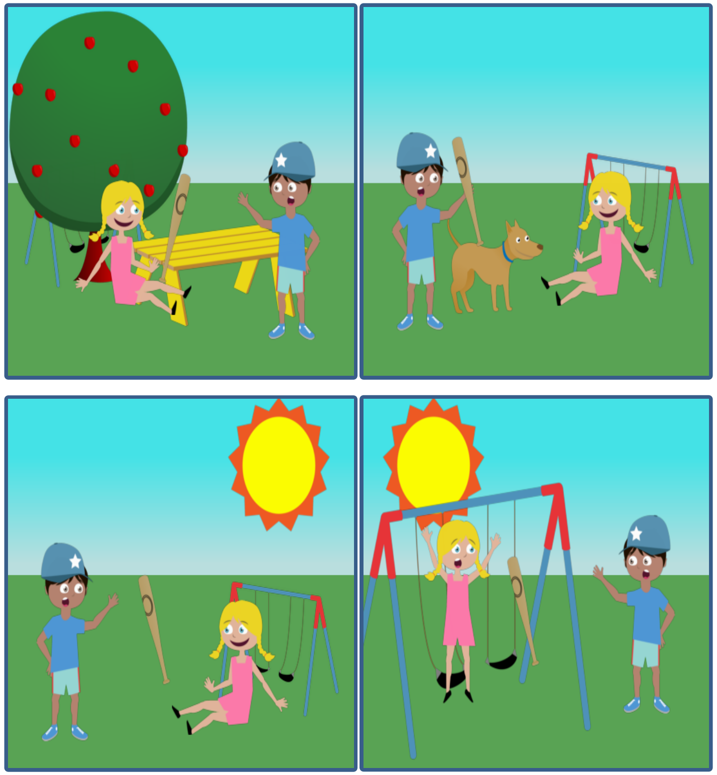
________________.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

A. There is a tree near a table.

B. The brown dog is standing next to Mike.

C. The sun is in the sky.

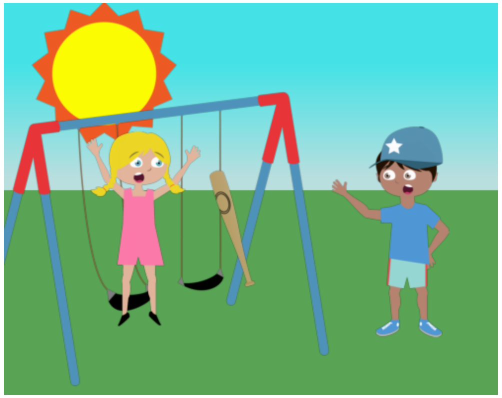
D. Jenny is standing dangerously on the swing.



Slide credit: Xiao Lin

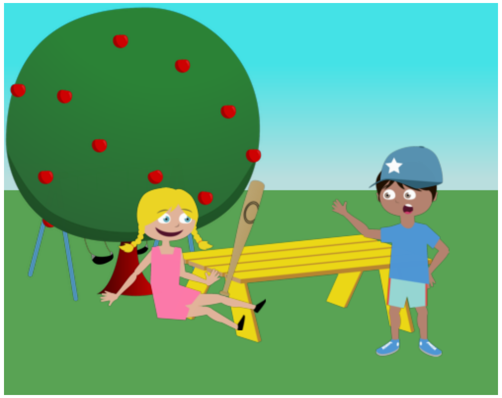
Approach: Joint Text + Visual Reasoning

Jenny is standing dangerously on the swing. Mike is wearing a blue cap. Mike is telling Jenny to get off the



$\geq$

There is a tree near a table. Mike is wearing a blue cap. Mike is telling Jenny to get off the swing.



$$w^T \phi_i^{\text{gt}} \geq w^T \phi_i^j + 1$$

Ranking Support Vector Machine (Ranking SVM)

Slide credit: Xiao Lin

Results

	Fill-in-the-blanks (FITB) Accuracy (+/- ~0.15)	Visual Paraphrasing (VP) AP (+/- ~0.02)
Random	25.00	33.33

[Lin and Parikh, CVPR 2015]

Results

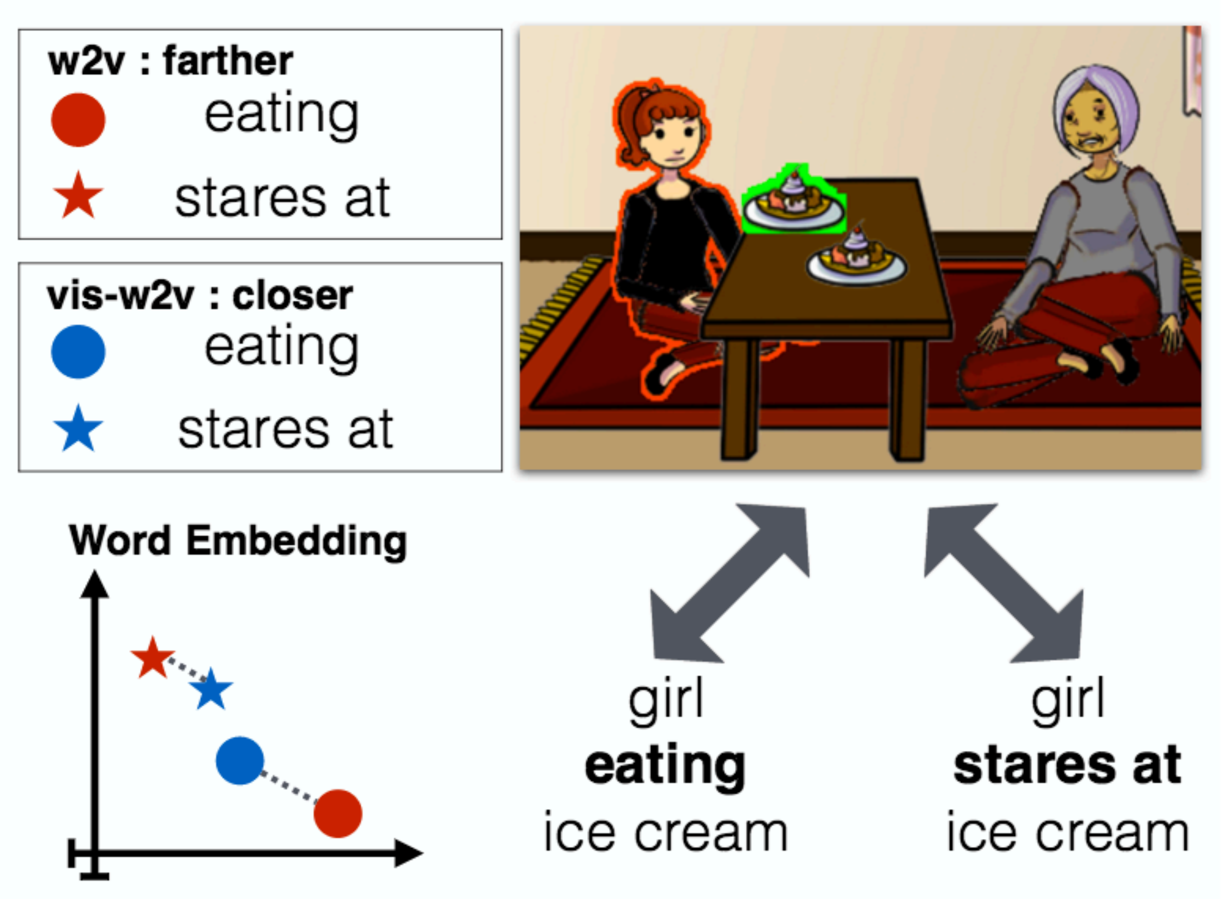
Given *any* tuple, can assess its plausibility

	Average Precision	Rank Correlation
Text alone		
Visual alone		
Text + visual		

[Vedantam, Lin, Batra, Zitnick, and Parikh, ICCV 2015]

Visual word2vec

- Learn word embeddings that respect visual (as well as textual) similarity



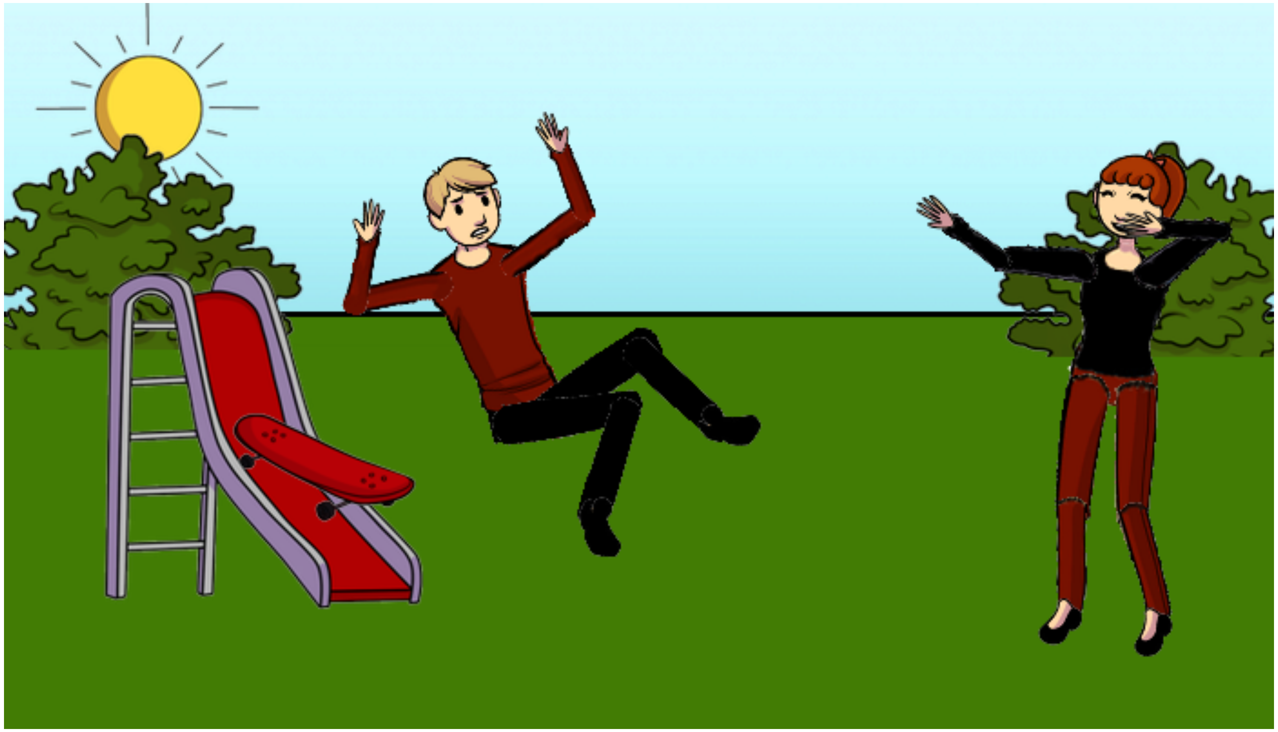
[Kottur, Vedantam and Parikh, CVPR 2016]

Understanding Visual Humor

[Chandrasekaran, Kalyan, Antol, Bansal, Batra, Zitnick, and Parikh, CVPR 2016]

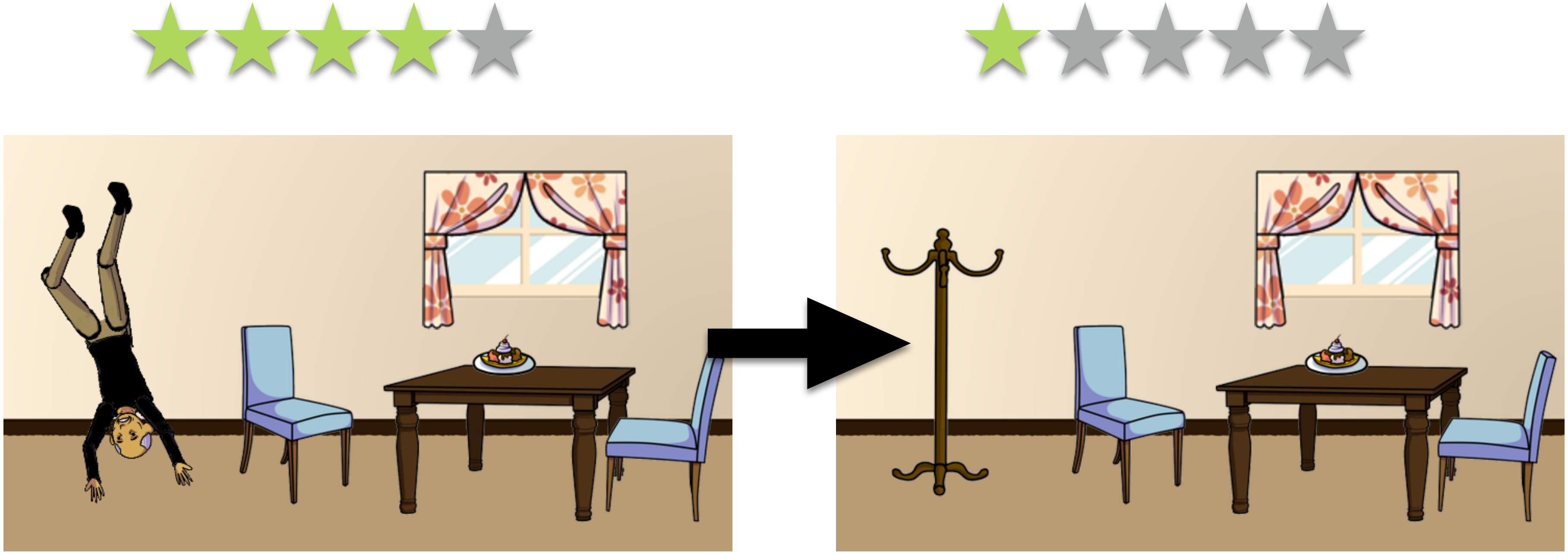
Task 1: Rating humor

★★★★☆



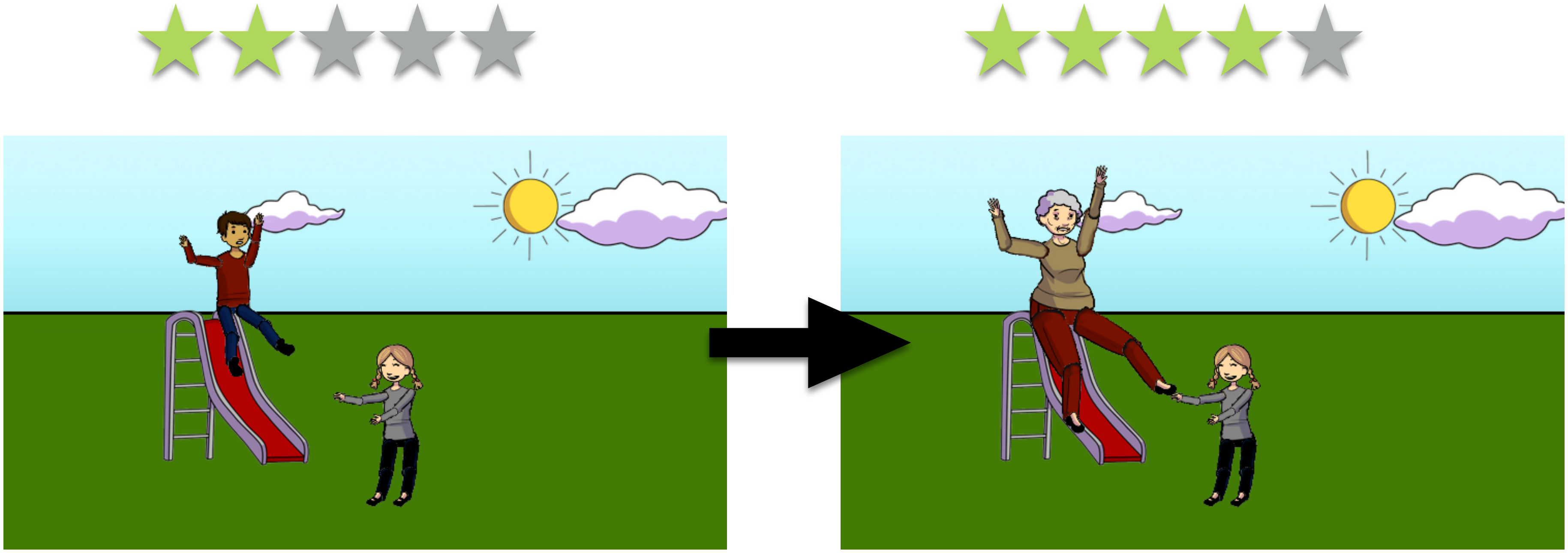
Slide credit: Arjun Chandrasekaran

Task 2: Remove humor



Slide credit: Arjun Chandrasekaran

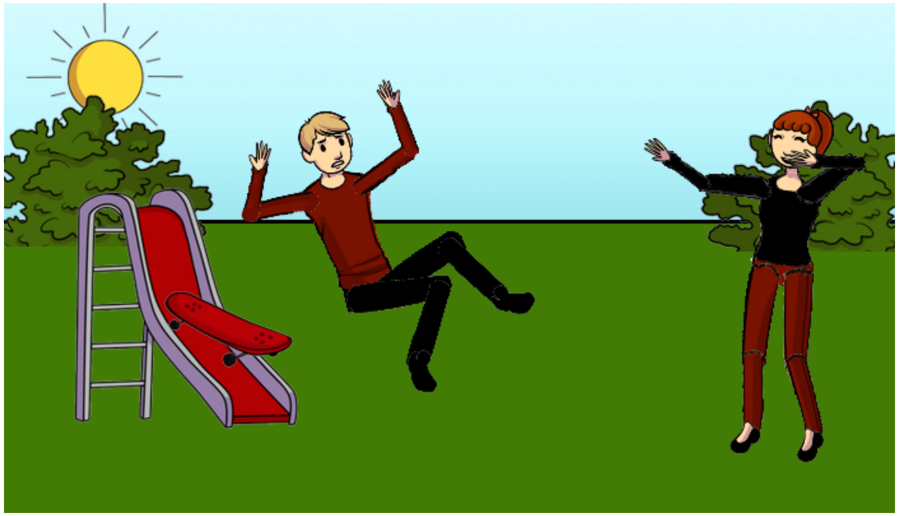
Task 2: Add humor



Slide credit: Arjun Chandrasekaran

Dataset: Abstract Visual Humor (AVH)

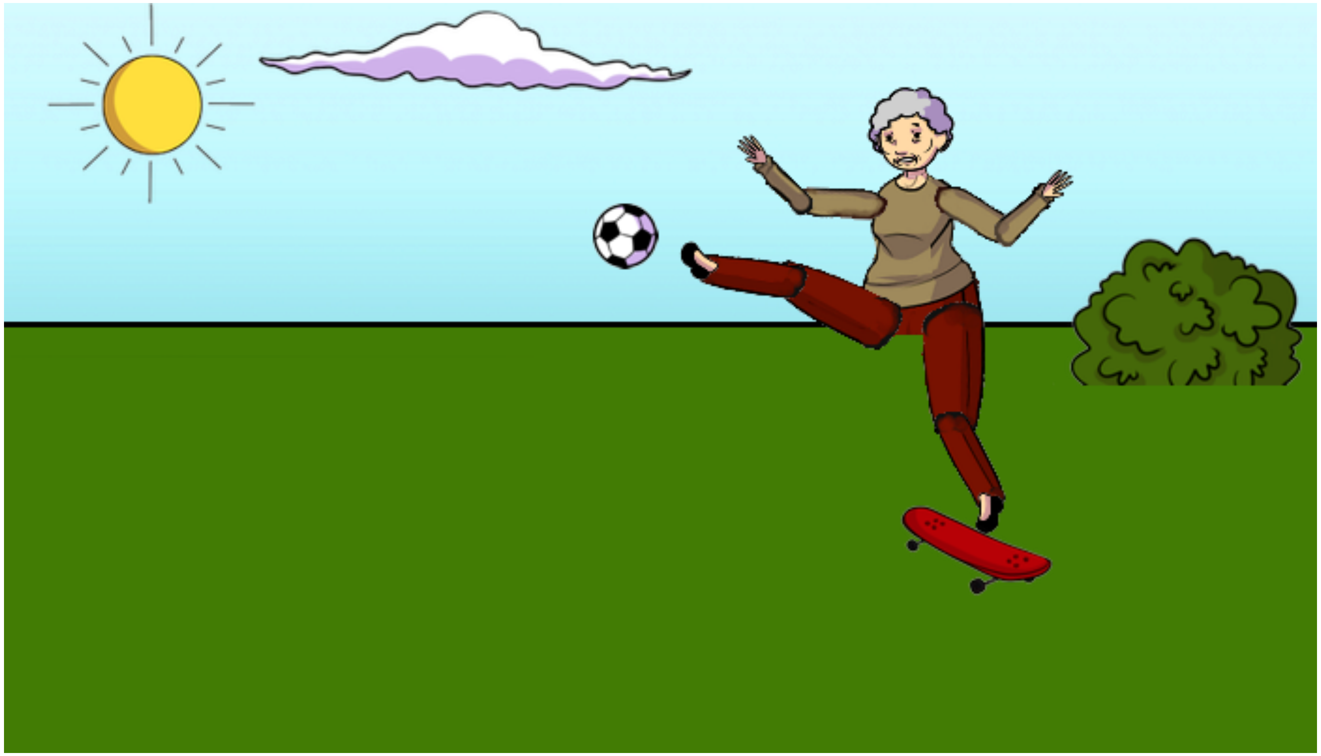
Funny



Not funny

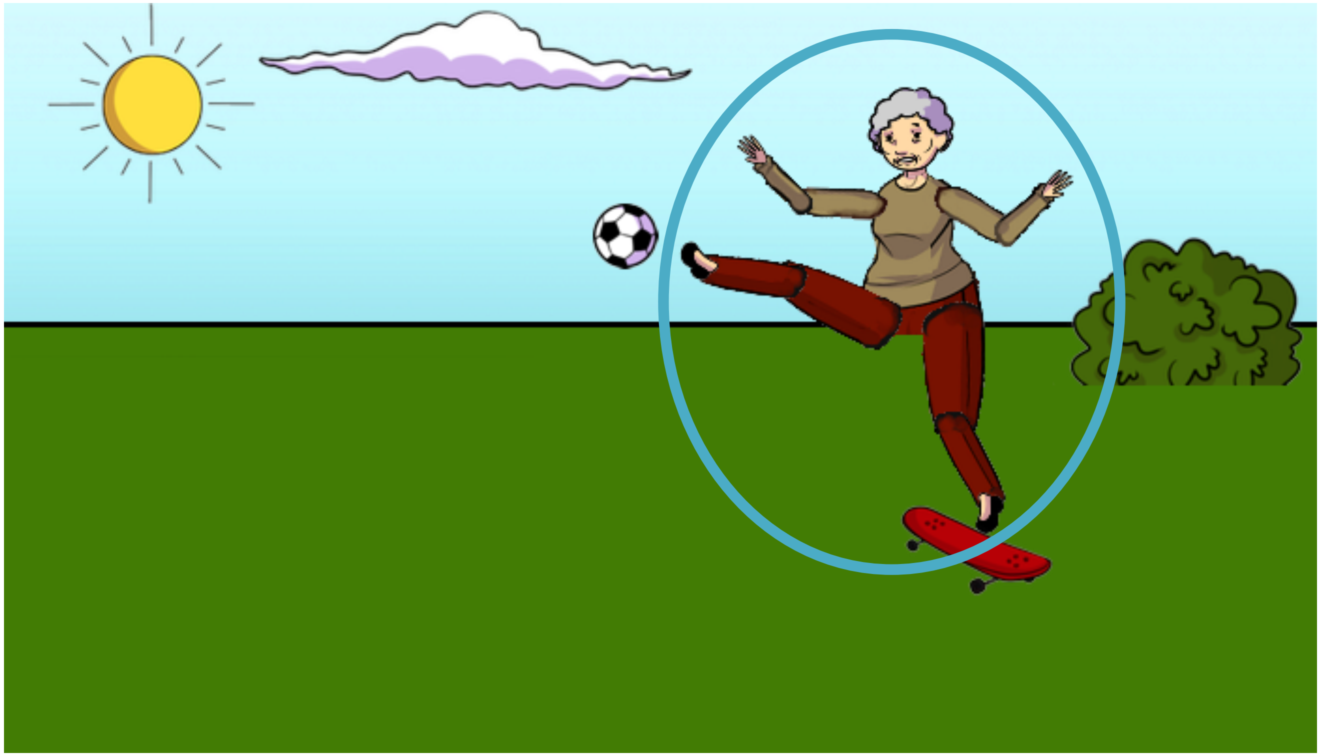
Slide credit: Arjun Chandrasekaran

Dataset: Funny Object Replaced (FOR)



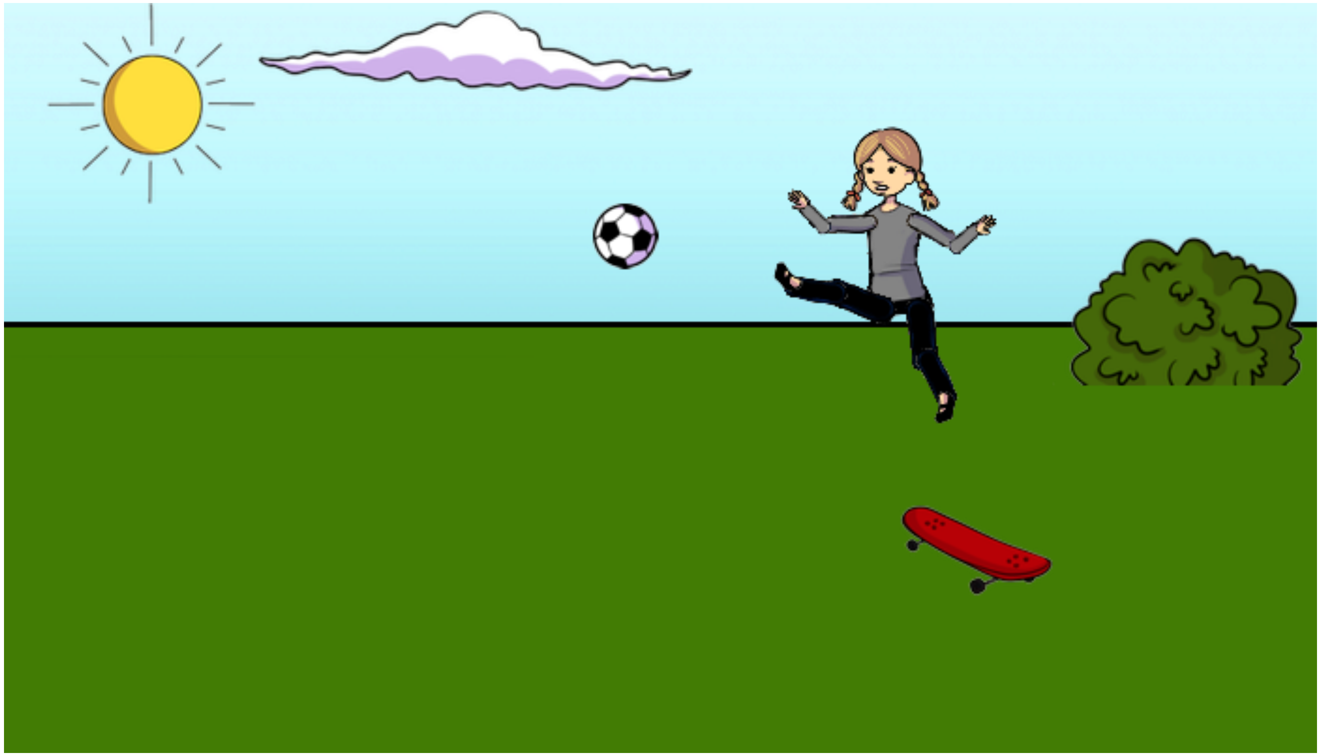
Slide credit: Arjun Chandrasekaran

Dataset: Funny Object Replaced (FOR)



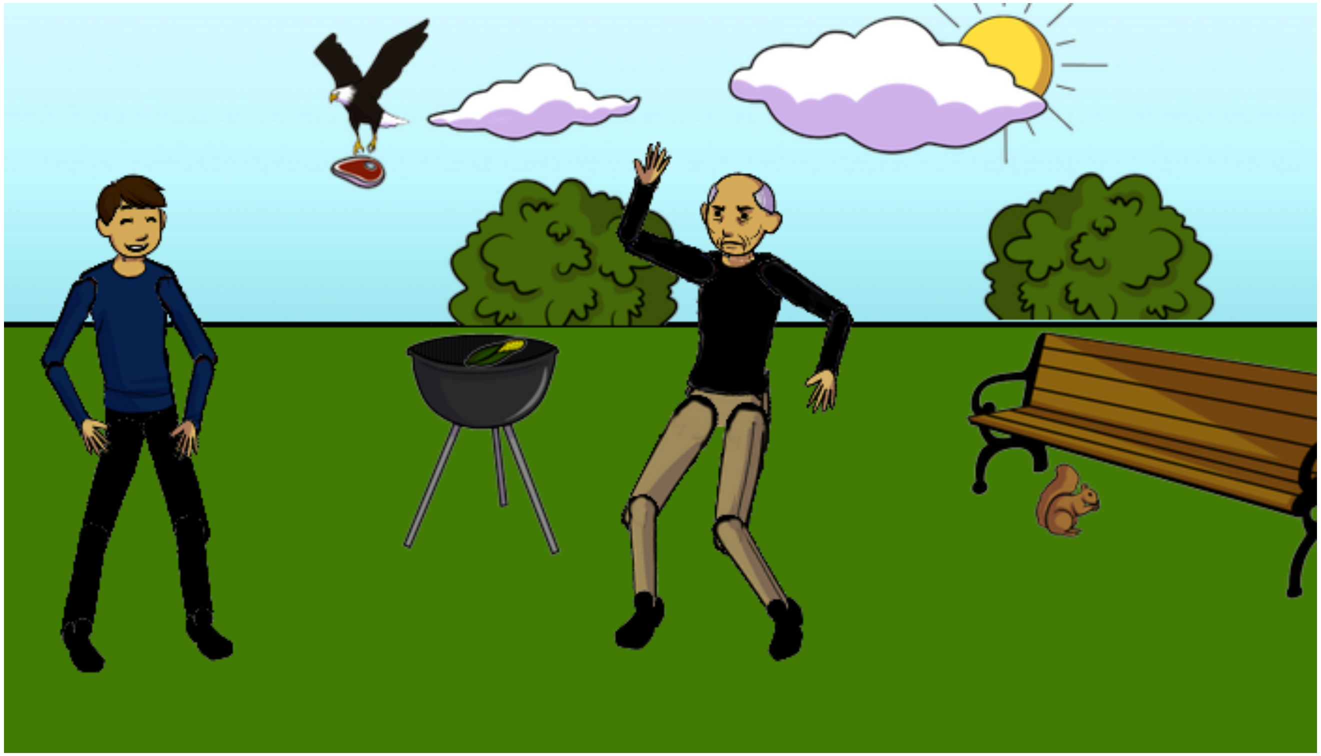
Slide credit: Arjun Chandrasekaran

Dataset: Funny Object Replaced (FOR)



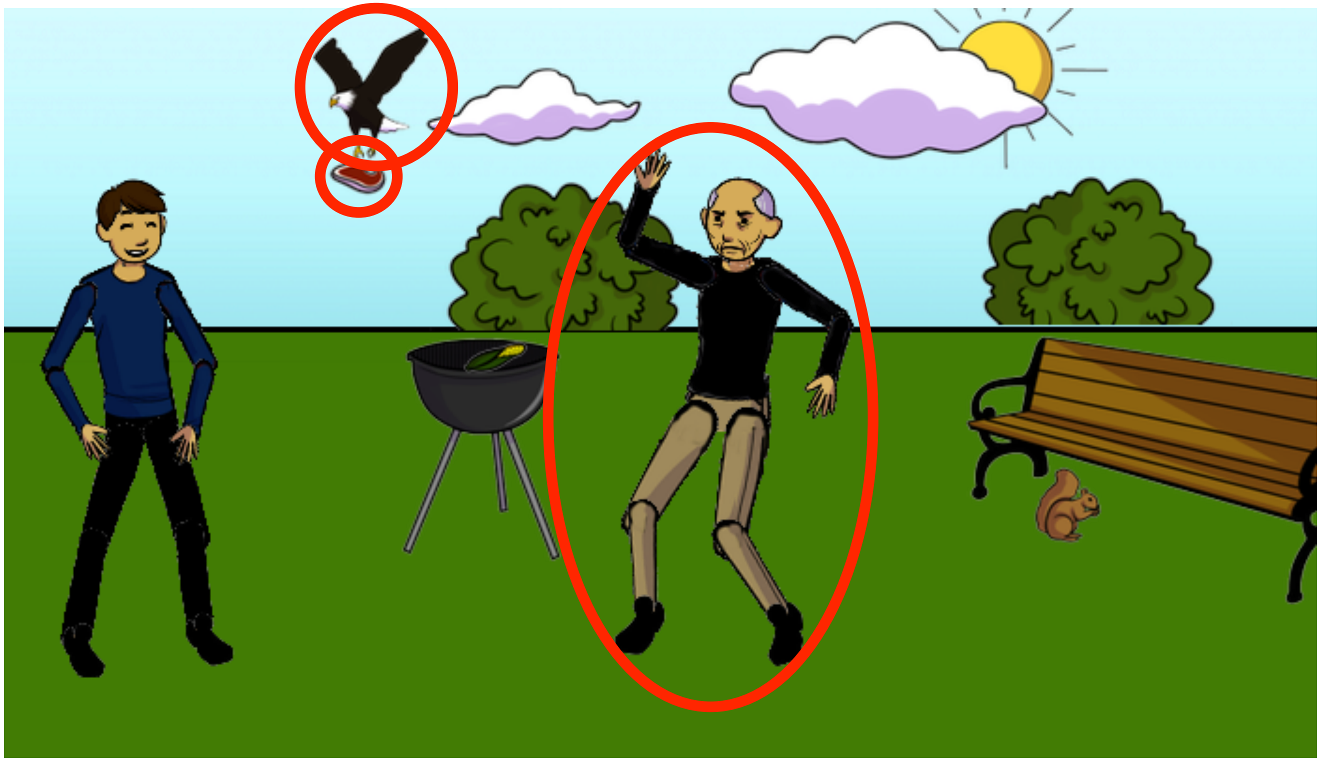
Slide credit: Arjun Chandrasekaran

Funny to unfunny



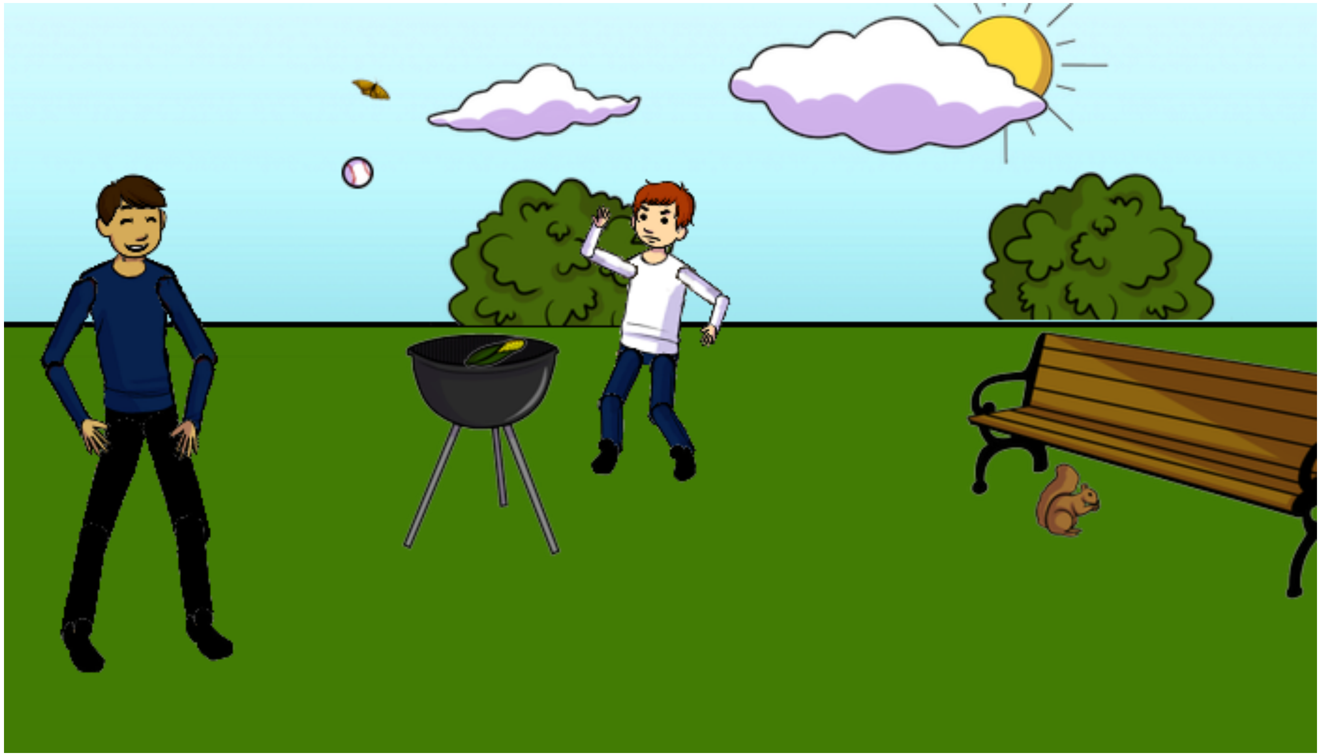
Slide credit: Arjun Chandrasekaran

Funny to unfunny



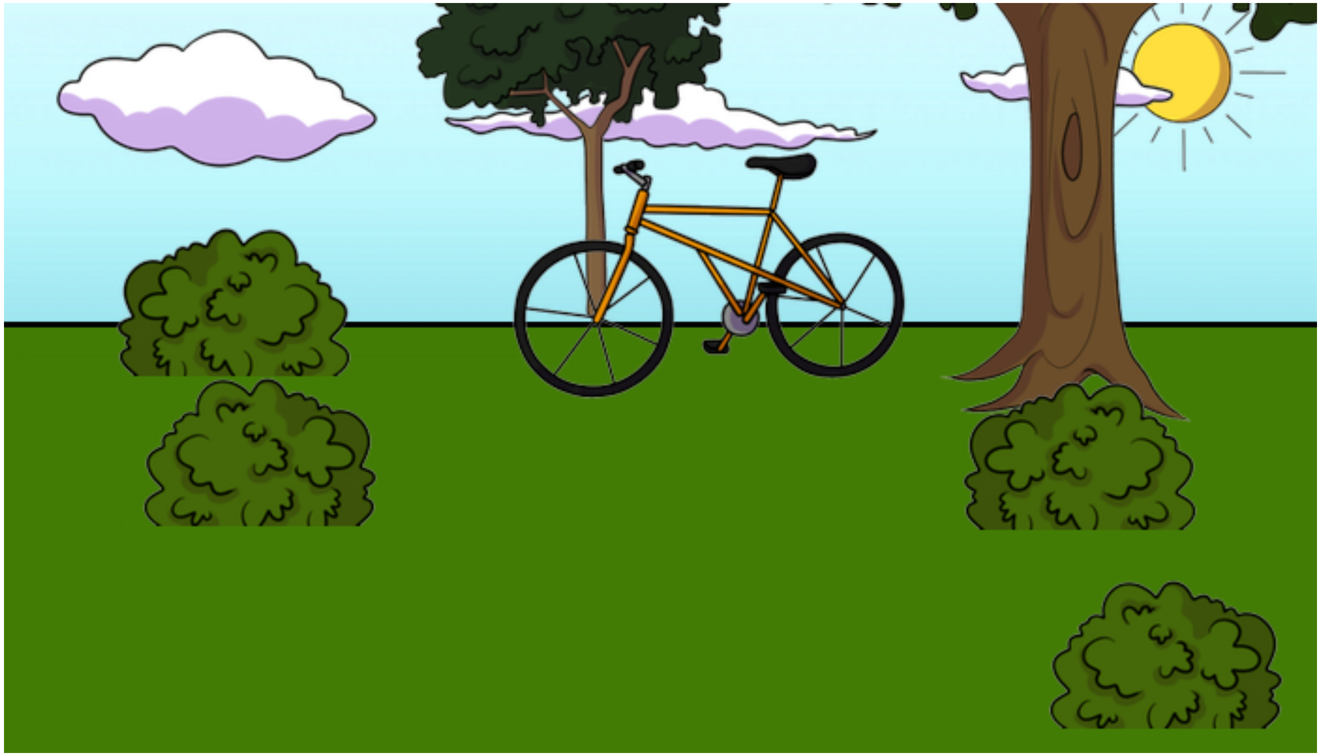
Slide credit: Arjun Chandrasekaran

Funny to unfunny



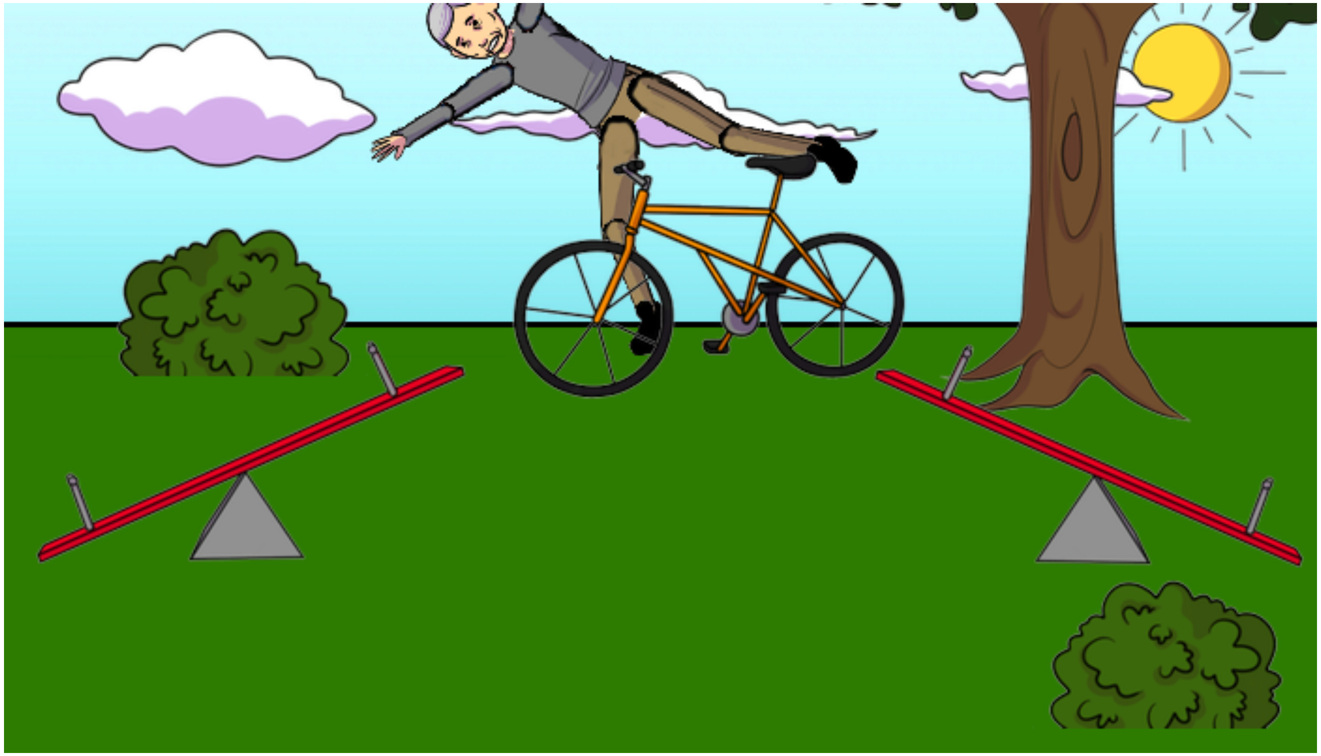
Slide credit: Arjun Chandrasekaran

Unfunny to funny



Slide credit: Arjun Chandrasekaran

Unfunny to funny

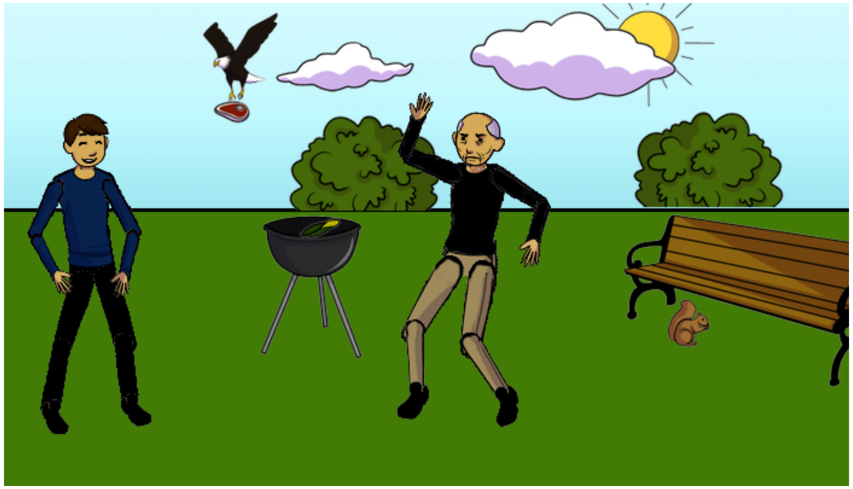


Slide credit: Arjun Chandrasekaran

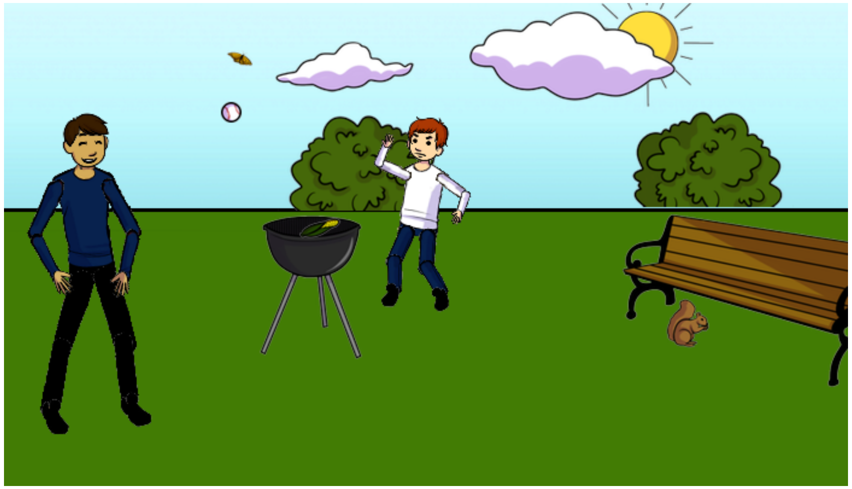
Human evaluation

Humor suppressor

Which scene is **LESS** funny?



5%



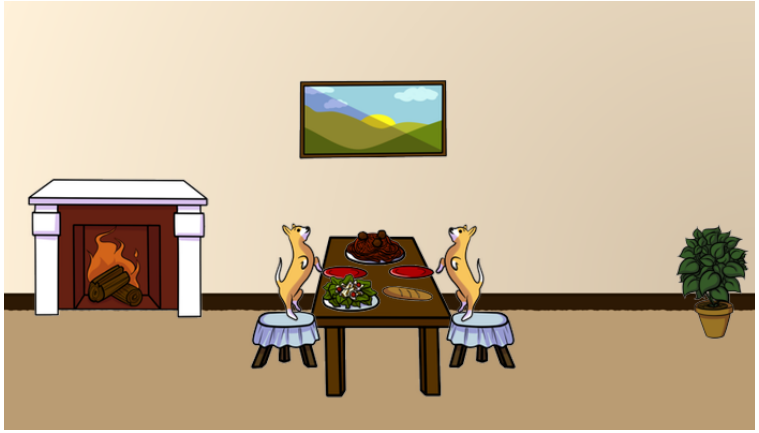
95%

Slide credit: Arjun Chandrasekaran

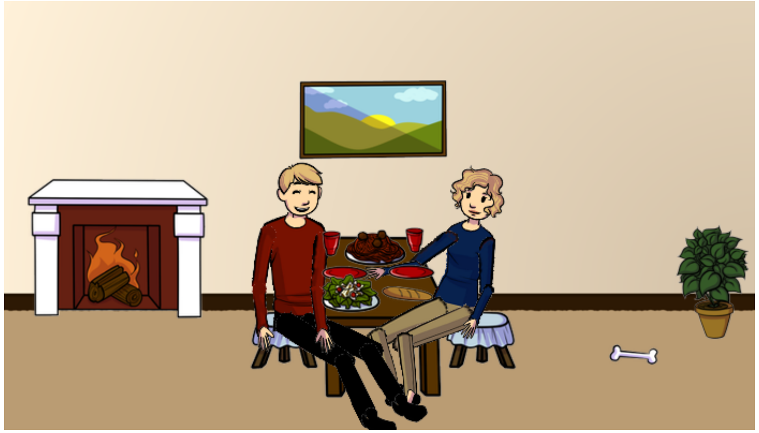
Human evaluation

Algorithm

Humor inducer



Not funny



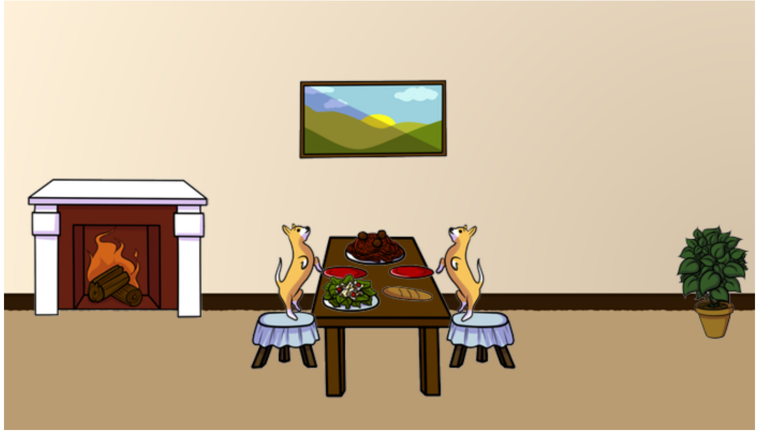
Slide credit: Arjun Chandrasekaran

Human evaluation

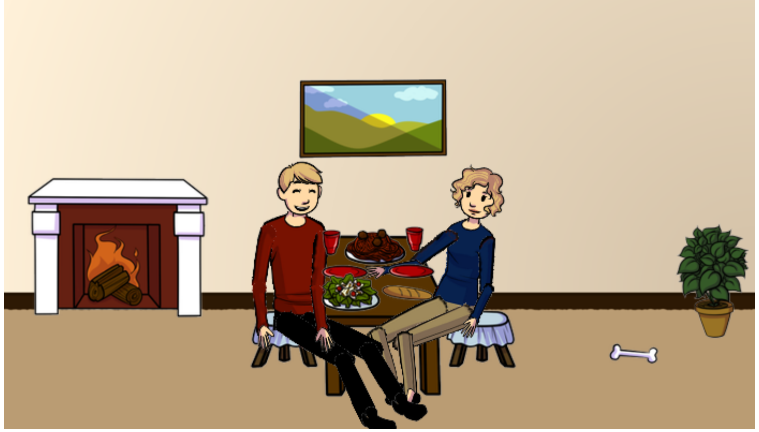
Humor inducer

Algorithm

28%



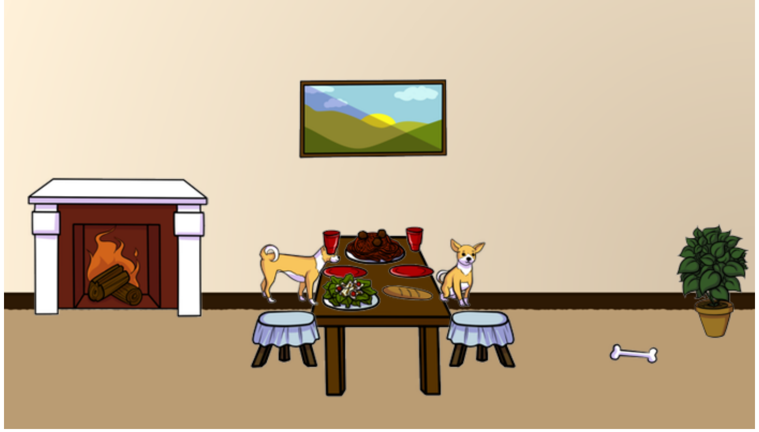
Not funny



Which scene is **MORE** funny?

Human

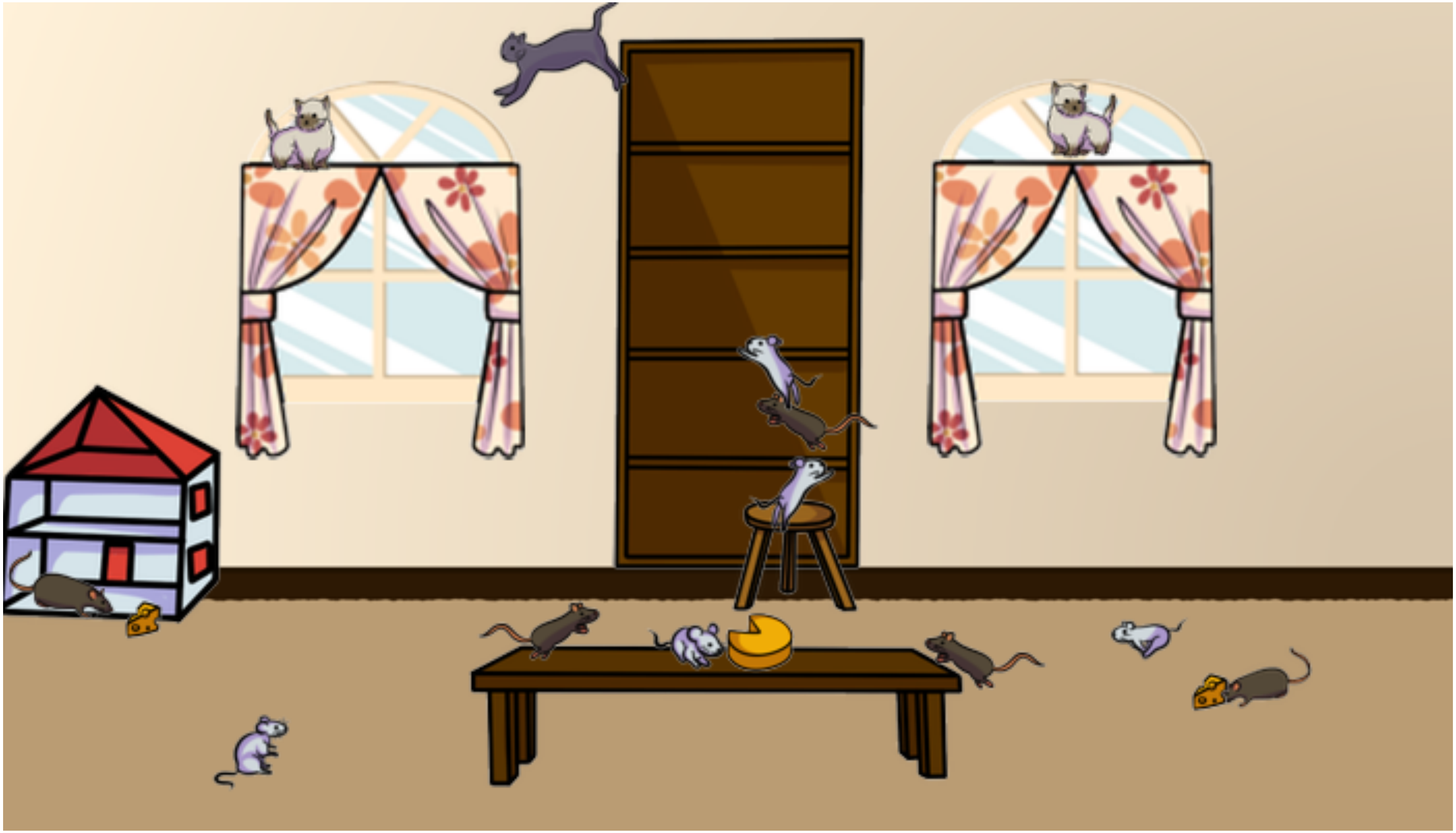
72%



Slide credit: Arjun Chandrasekaran

Funniest scene as per our algorithm

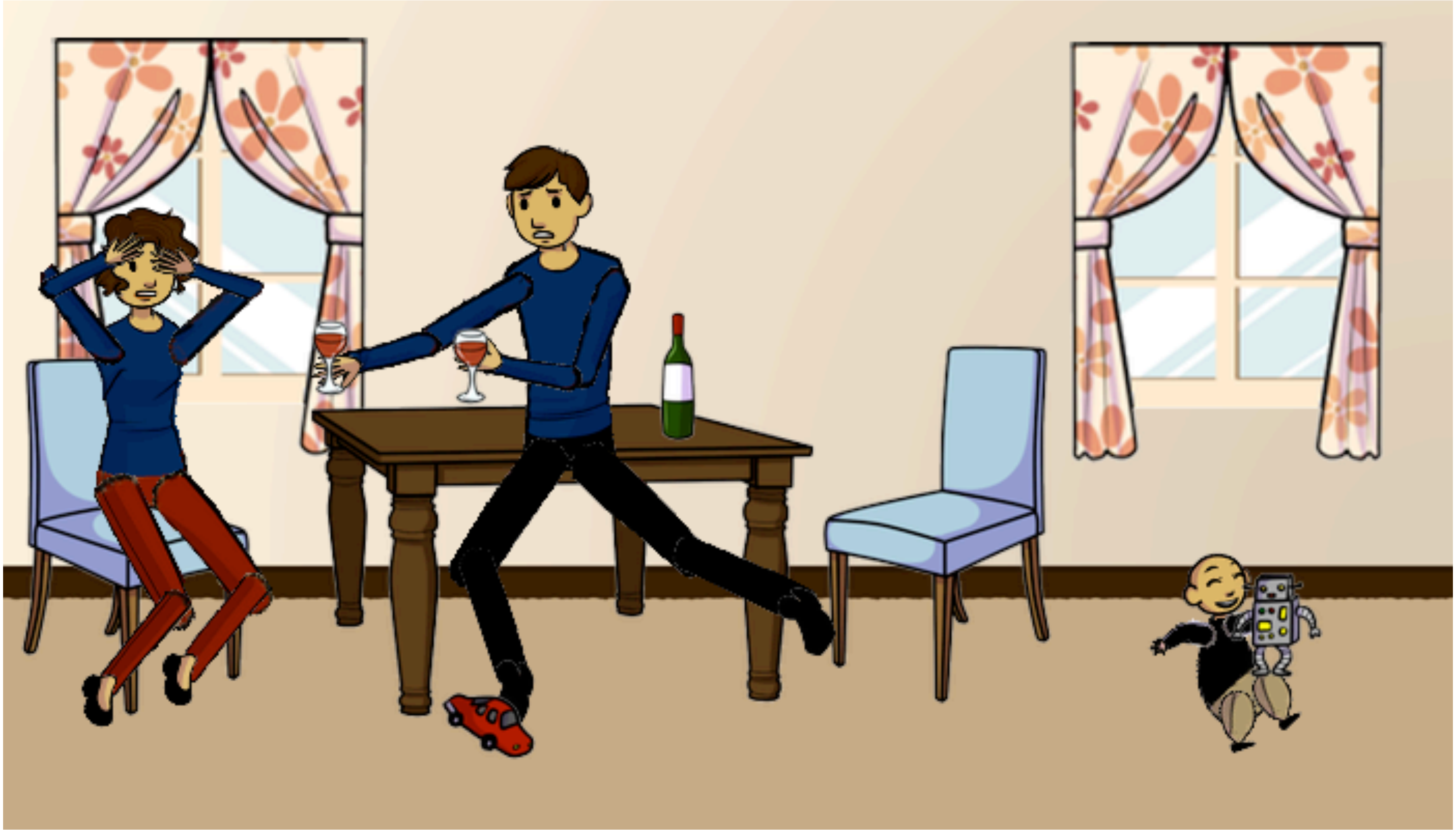
★★★★½



Slide credit: Arjun Chandrasekaran

“This terrified woman's home is being invaded by mice as the cat sleeps.”

“The man is about to trip on his child's car and spill wine on his wife.”



Visual Abstraction For…

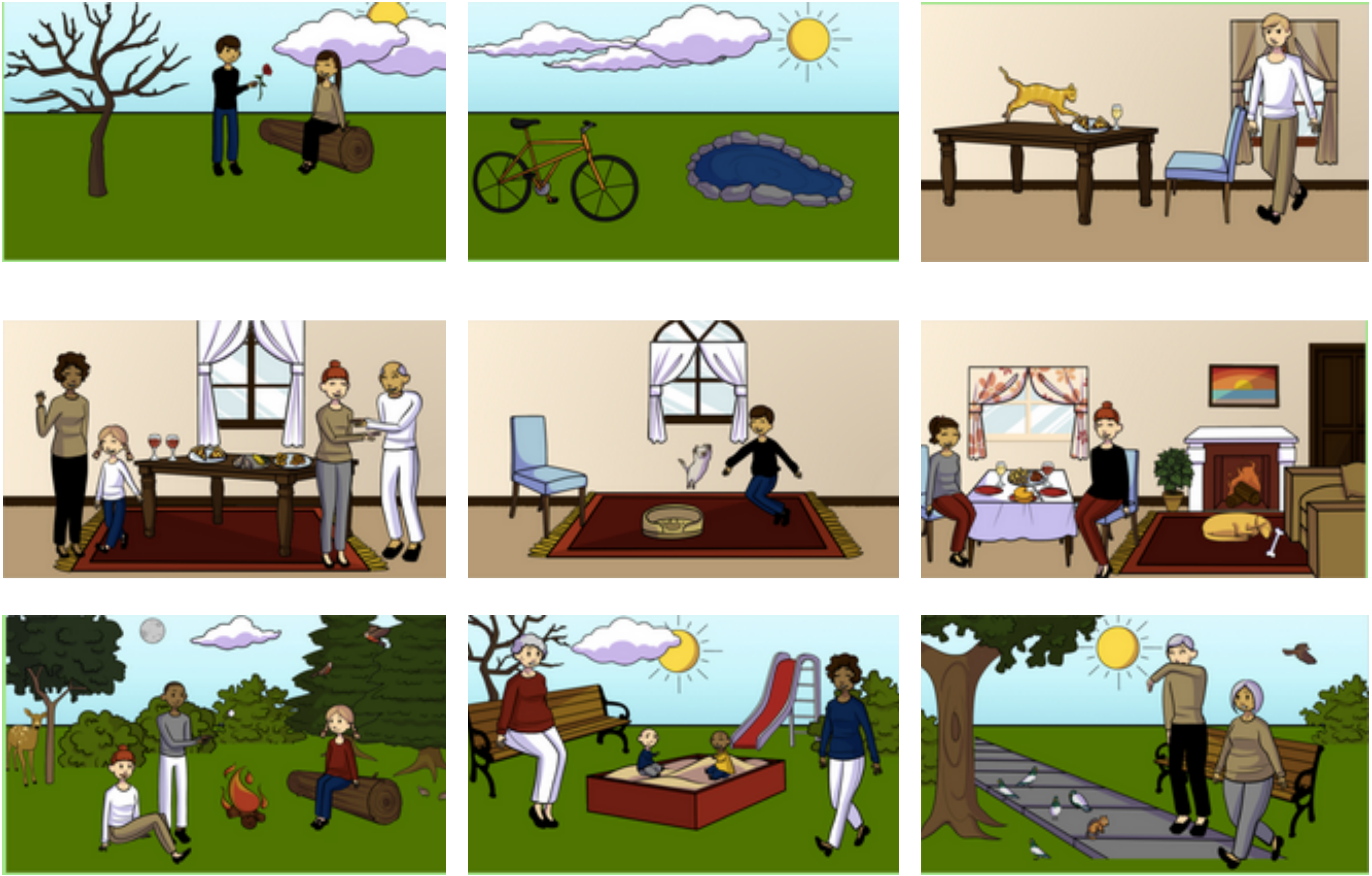
- Studying mappings between images and text [CVPR 2013, ICCV 2013]
- Zero-shot learning [ECCV 2014]
- Studying
  - Image memorability [PAMI 2016]
  - Image specificity [CVPR 2015]
  - Visual humor [CVPR 2016]

Visual Abstraction For…

- Studying mappings between images and text [CVPR 2013, ICCV 2013]
- Zero-shot learning [ECCV 2014]
- Studying
  - Image memor
  - Image specific
  - Visual humor [2016]
- Learning common sense knowledge [CVPR 2015, ICCV 2015, CVPR 2016]
- Rich annotation modality
  - Ask for descriptions
  - Ask for scene
  - Show scene a tion
  - Perturb a scene and ask for descriptions
  - …

Study high-level image understanding tasks without waiting for lower-level vision tasks to be solved

Future work:
Learning by "playing"



50k scenes, captions, QAs: available online!

Thank you.