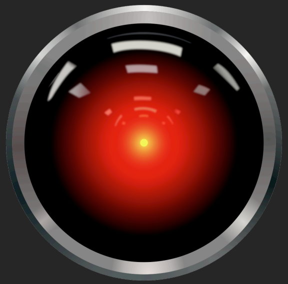
Language and Vision: Learning Knowledge about the World

Yejin Choi

Computer Science & Engineering

UNIVERSITY of WASHINGTON

Goal: Intelligent Communication



Intelligent Communication

Reading between the lines

Understanding

what is said

+

what is *not* said

language in physical context

Blueberry Muffins

Ingredients

1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

Need knowledge about the cooking world!

Bake what? where?

Procedure

1. Preheat oven to 400 degrees F. Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.
3. Bake for 20 minutes. Serve hot.

http://allrecipes.com/Recipe/Blueberry-Muffins-I/

Intelligent Communication

Reading between the lines

Understanding

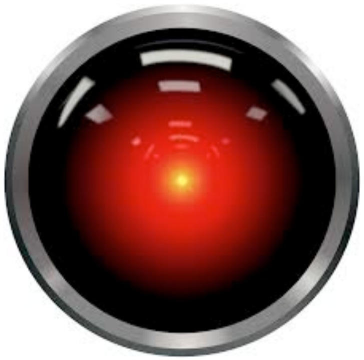
what is said

+

what is *not* said

Language is contextual:

- social / emotional context
- visual / physical context



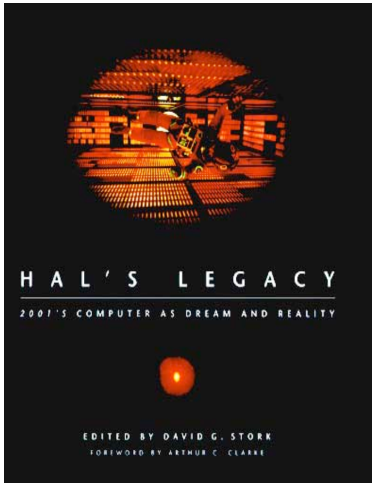
HAL (A space odyssey, 1968)

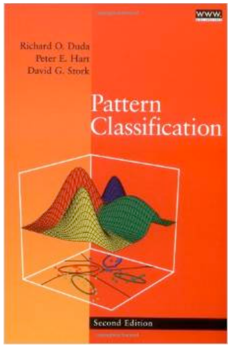
- David Stork (HAL's Legacy, 1998)

"Imagine, for example, a computer that could look at an arbitrary scene anything from a sunset over a fishing village to Grand Central Station at rush hour and produce a verbal description.

This is a problem of overwhelming difficulty, relying as it does on finding solutions to both vision and language and then integrating them.

I suspect that scene analysis will be one of the last cognitive tasks to be performed well by computers"





Web in 1995


MONEY & INVESTING UPDATE
from THE WALL STREET JOURNAL.
Front Page
STOCKS
U.S.
Small U.S.
Americas
Asia
Europe
Heard on the Street
Credit Markets
Foreign Exchange
Commodities
Mutual Funds
Wednesday, September 6, 1995
What's News —
Business and Finance
MARKETS DIARY 5 p.m. EDT
DJIA 4683.81 + 13.73
S&P 500 + 0.18%
Nasdaq Composite + 0.48%
Tokyo (Nikkei 225) – 0.98%
London (FT 100) + 0.72%
30-Yr Treasury Yield 6.59%
Japanese yen (per US$) 98.82
German mark (per US$) 1.4768
Computer Shares Lift Stocks Again; Bonds Are Weak
By DAVE PETTIT
Money & Investing Update


amazon.com
Earth's biggest bookstore
Welcome to Amazon.com Books!
One million titles, consistently low prices.
(If you explore just one thing, make it our personal notification service. We think it's very cool!)
SPOTLIGHT! -- AUGUST 16TH
These are the books we love, offered at Amazon.com low prices. The spotlight moves EVERY day so please come often.
ONE MILLION TITLES
Search Amazon.com's million title catalog by author, subject, title, keyword, and more... Or take a look at the books we recommend in over 20 categories... Check out our customer reviews and the award winners from the Hugo and Nebula to the Pulitzer and Nobel... and bestsellers are 30% off the publishers list...
EYES & EDITORS, A PERSONAL NOTIFICATION SERVICE
Like to know when that book you want comes out in paperback or when your favorite author

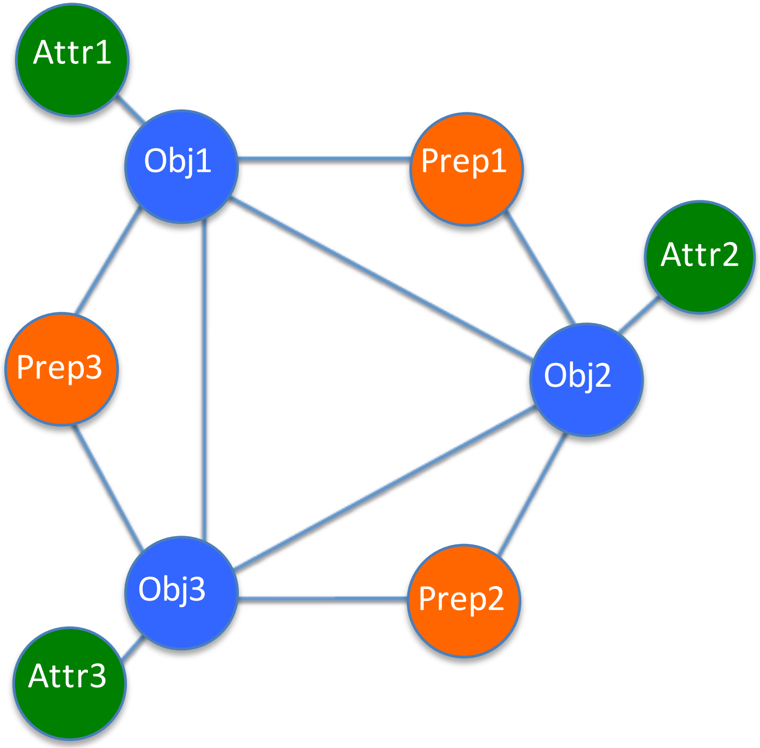
Web Today: Increasingly Visual

-- social media, news media, online shopping

- Facebook.com has over 250 billion images uploaded as of Jun 2013
- 1.15 billion users uploading 350 million images a day on average

Image Captioning - Take I - Baby Talk (CVPR 2011)

Conditional random fields (CRF) model to combine visual detection with language priors



"This picture shows one person, one grass, one chair, and one potted plant. The person is near the green grass, and in the chair. The green grass is by the chair, and near the potted plant."

Image Captioning - Take II – Tree Talk (TACL 2014)

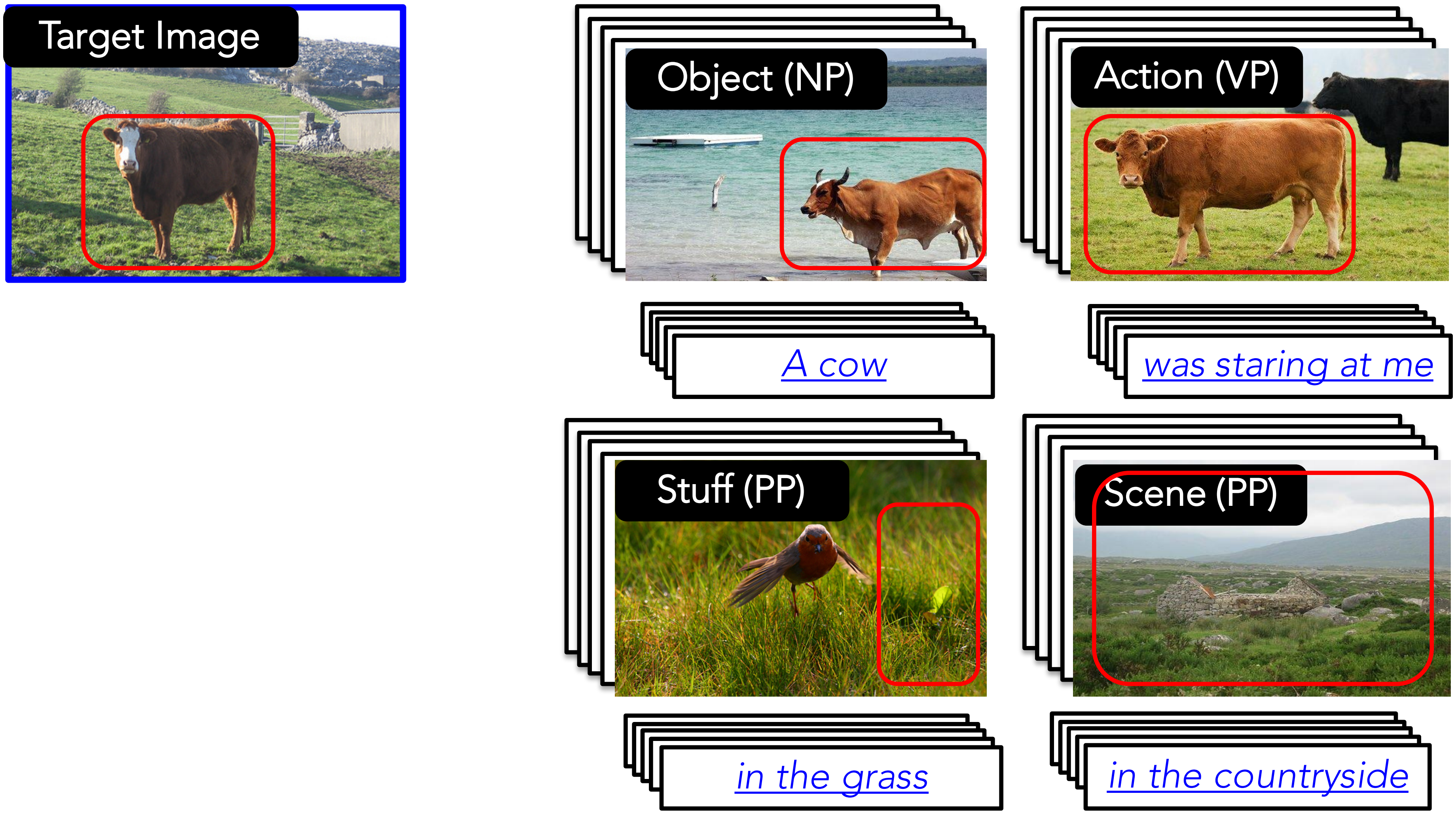
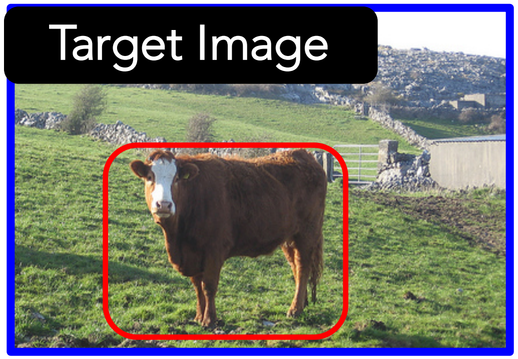


Image Captioning - Take II – Tree Talk (TACL 2014)



A cow
in the grass
was staring at me
in the countryside

A cow
was staring at me
~~in the grass~~
in the countryside

Tree Structure --- Probabilistic Context Free Grammars (PCFG)

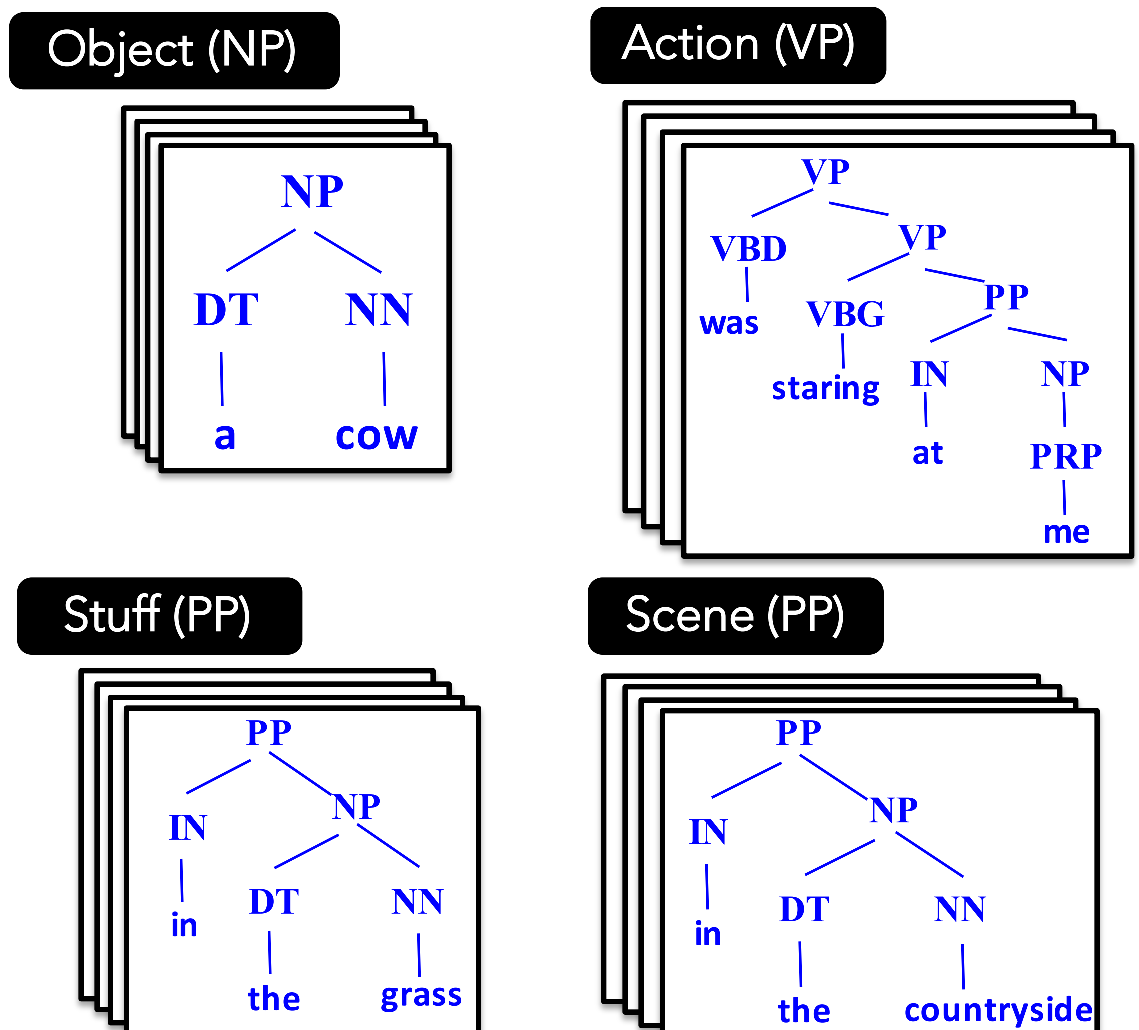


Image Captioning - Take II – Tree Talk (TACL 2014)

Blue flowers have no scent. Small white flowers have no idea what they are.

My cat laying in my duffel bag.

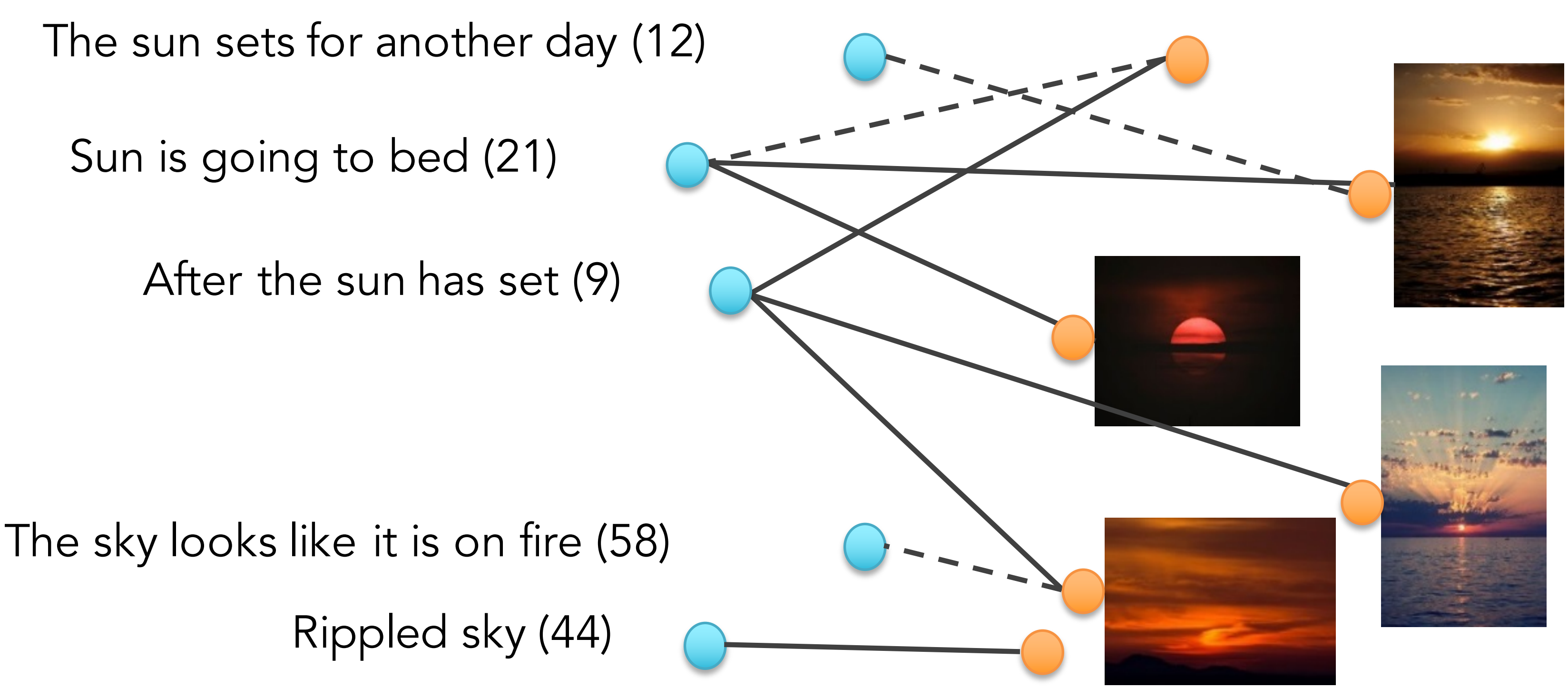
Blue flowers are running rampant in my garden.

Mini Turing Test: our system wins in ~ 24 % cases!

Image Captioning - Take III – Deja Captions (NAACL 2015)

Deja Image-caption corpus (NAACL 2015):

- Of 750 million pairs of image-caption pairs from Flickr
- Retain only those captions that are repeated verbatim by more than one user
- Yielding 4 million images with 180K unique captions



Related Work

- Donahue et al., 2015, Vinyals et al, 2015, Fang et al., 2015, Karpahty et al, 2015, Xu et al, 2015, Delvin et al., 2015, …

- MS CoCo Dataset
  - 120,000 images, 5 captions per image
  - 80 objects

Data problem?
Or Modeling problem?

  - sports (10 categories):
    - tennis racket (3561 images), baseball bat, baseball gloves, snowboard, skateboard, surf board,…
  - street (5 categories)
    - traffic light (4330 images), fire hydrant (1797 images), stop sign (1803 images), parking meters (742 images), bench (5805 images)
  - person (6 categories)
    - tie (3955 images), umbrella (4142 images)

Moving Forward …

- Image captioning is an emblematic task, not the end goal
- Seeing beyond the literal content

- Why did this happen?
- How do they feel?
- Reasoning about the situation
- Need knowledge about the world

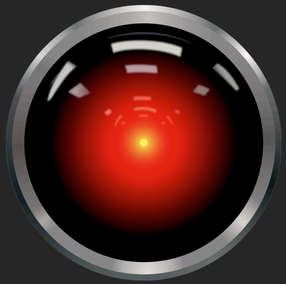
Learning Knowledge about the World

I: Size

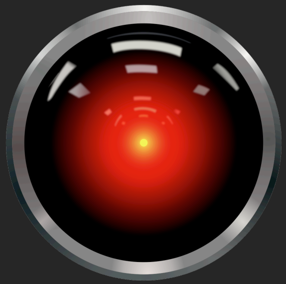
II: Entailment

III: Cooking

IV: Event



Learning Knowledge about the World
Take I: Size



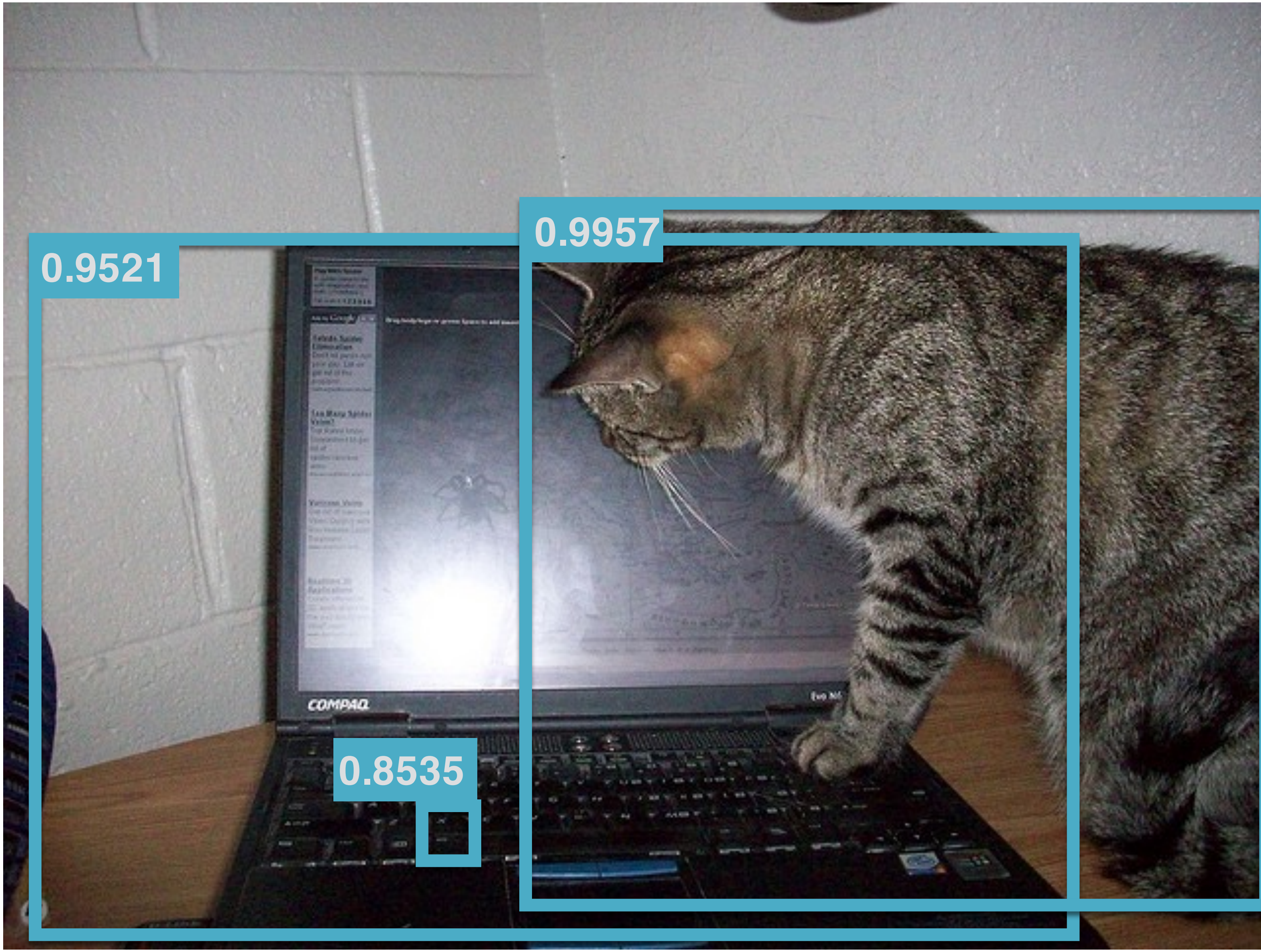
Bagherinezhad et al. @ AAAI 2016

Are Elephants Bigger than Butterflies?

Bagherinezhad et al. @ AAAI 2016

Knowledge on Size Useful for

- Vision:
  - Prune out implausible detections


0.9521
0.9957
0.8535
COMPAQ

Knowledge on Size Useful for

- Vision:
  - Prune out implausible detections
- Language:
  - The trophy would not fit in the brown suitcase because it was too **big**. What was too **big**?
    Answer 0: the trophy
    Answer 1: the suitcase

Related Work

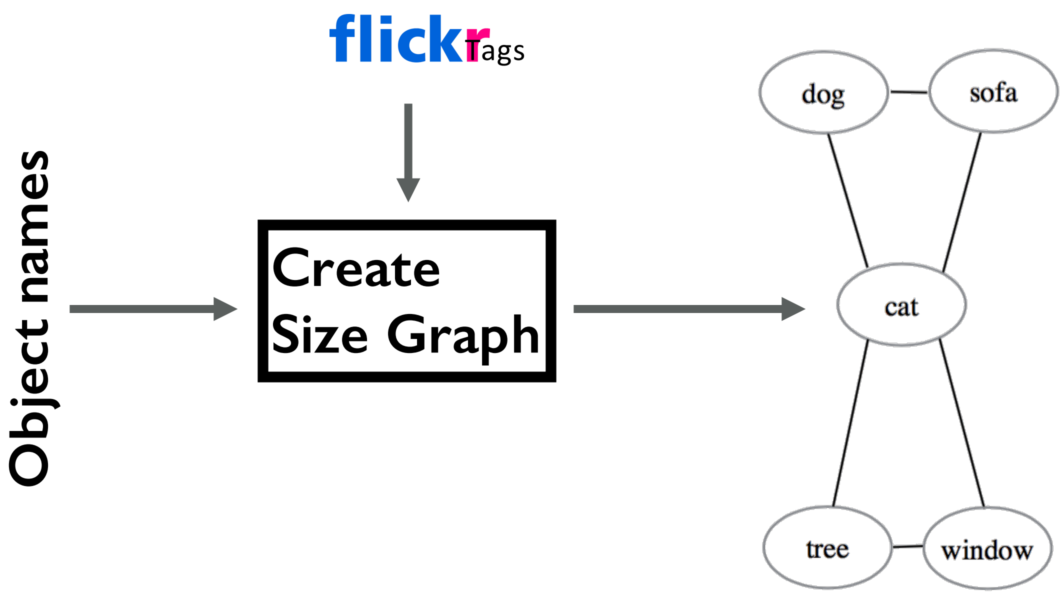
- Narisawa et al. 2013 -- Is a 204 cm Man Tall or Small?
- Tandon et al. 2014 -- WebChild
- Takamura et al. 2015

➔ Text only

Elephants Bigger than People?

- Reporting bias: do not state the obvious
- Use both language and images!
- Elephants bigger than butterflies?

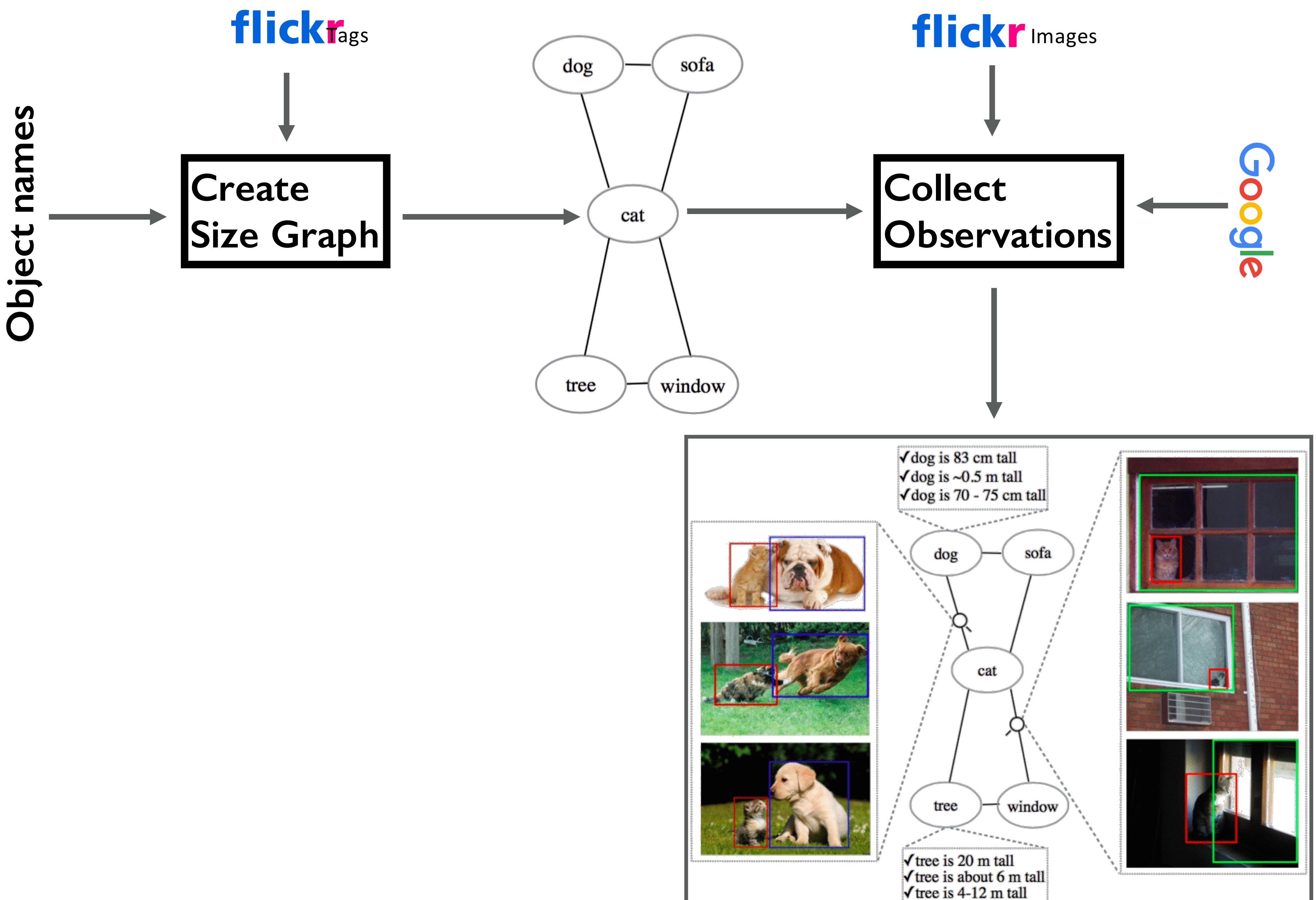
➔ Need multi-hop inference



Construction of size graph

- Not all object pairs co-occur in many images.
    - e.g. “airplane” and “watermelon”
- It is not scalable to see images for all object pairs.
- An edge (A,B) only if A and B co-occur in many images.
- 2 edge connected (2 disjoint edge paths between every pair)

Bagherinezhad et al. @ AAAI 2016


flickr Tags
Object names
Create Size Graph
dog
sofa
cat
tree
window
flickr Images
Google
Collect Observations
✓dog is 83 cm tall
✓dog is ~0.5 m tall
✓dog is 70 - 75 cm tall
dog
sofa
cat
tree
window
✓tree is 20 m tall
✓tree is about 6 m tall
✓tree is 4-12 m tall

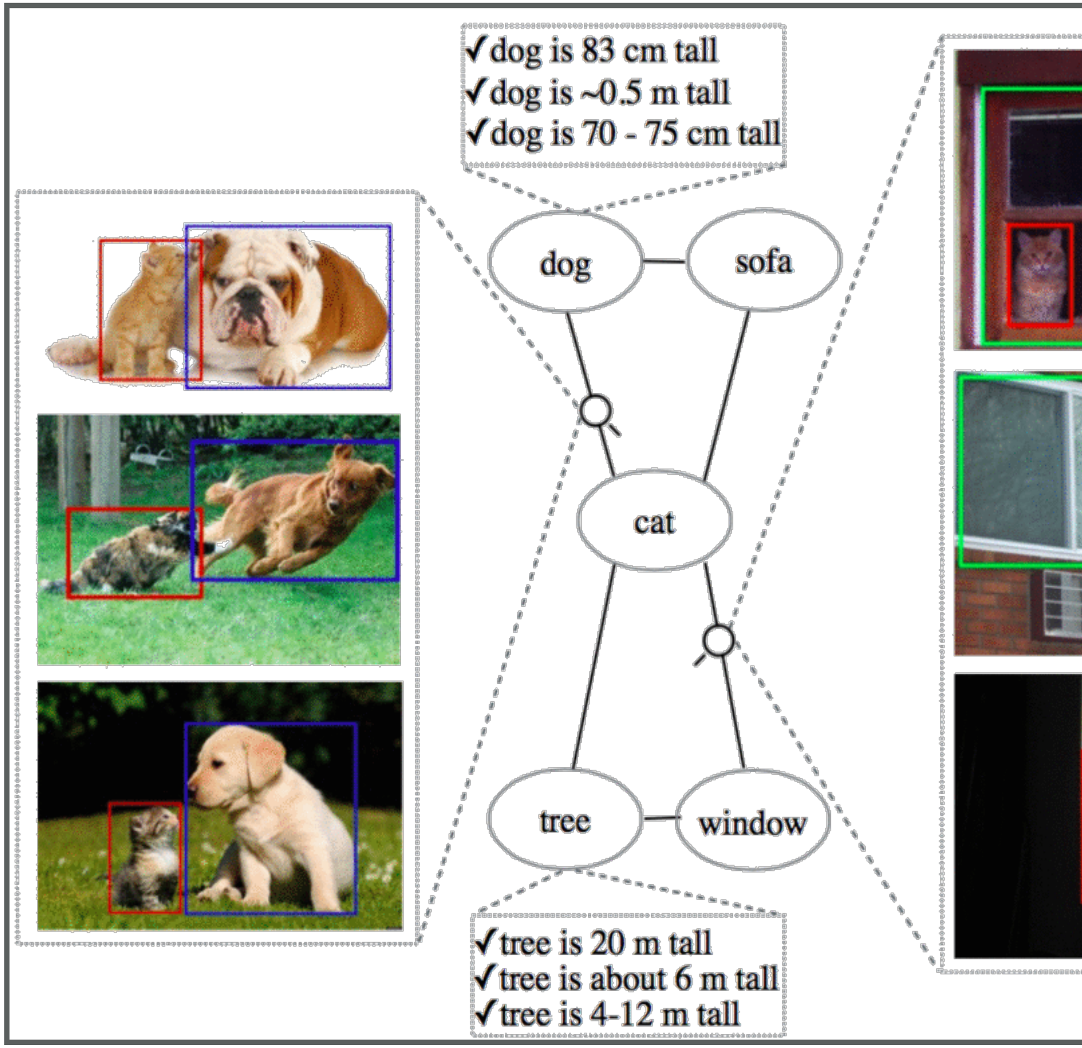
Language – absolute estimation

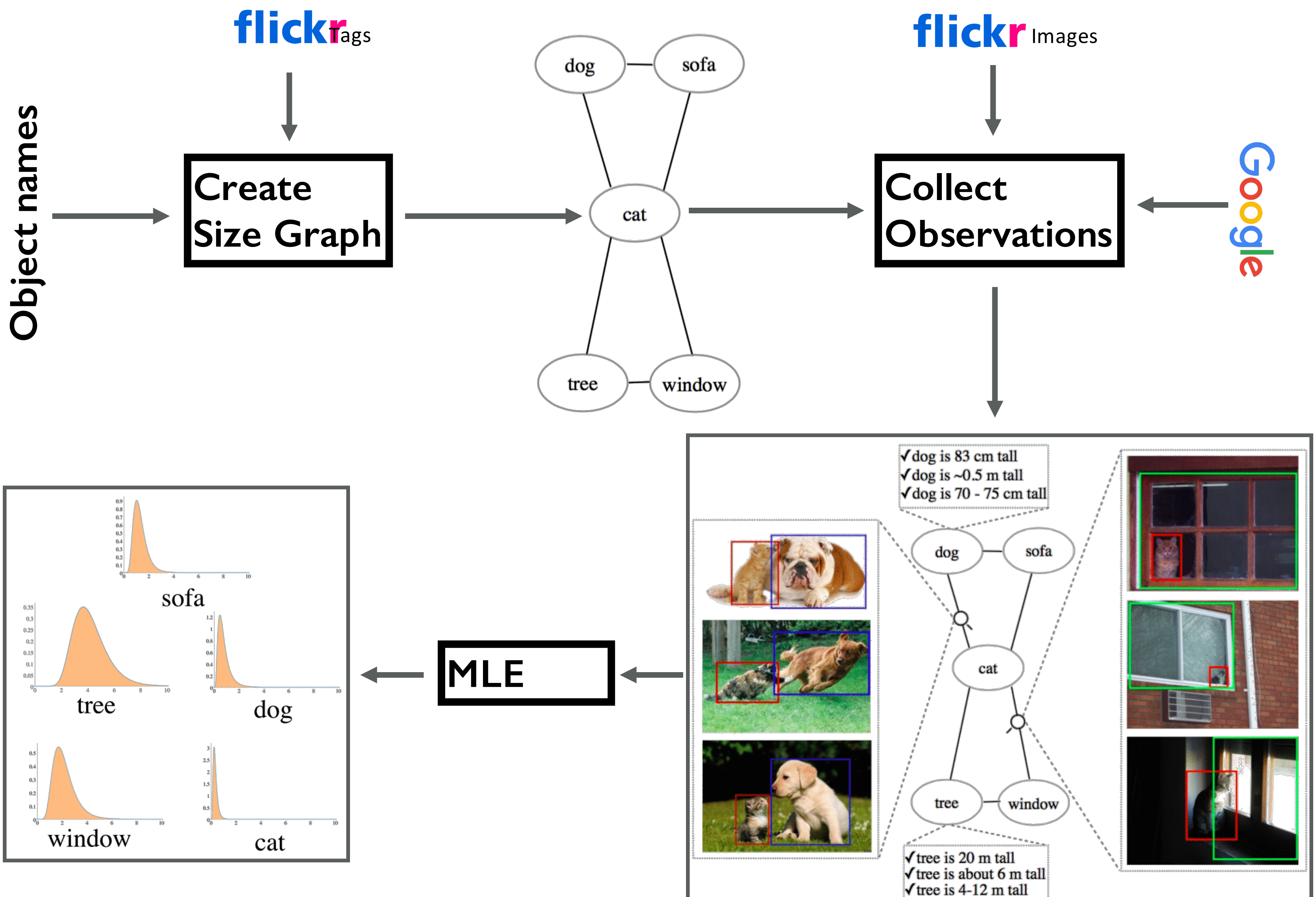
- "car is * x * m"
- "person is * m tall"

Vision – relative estimation

- From Flickr images that are tagged with both objects
- LEVAN [CVPR14], a webly supervised object detector.
- Run a depth estimator to infer the object distances

$$\frac{size(O_i)}{size(O_j)} = \frac{area(box_1)}{area(box_2)} \times \frac{depth(box_1)^2}{depth(box_2)^2}$$

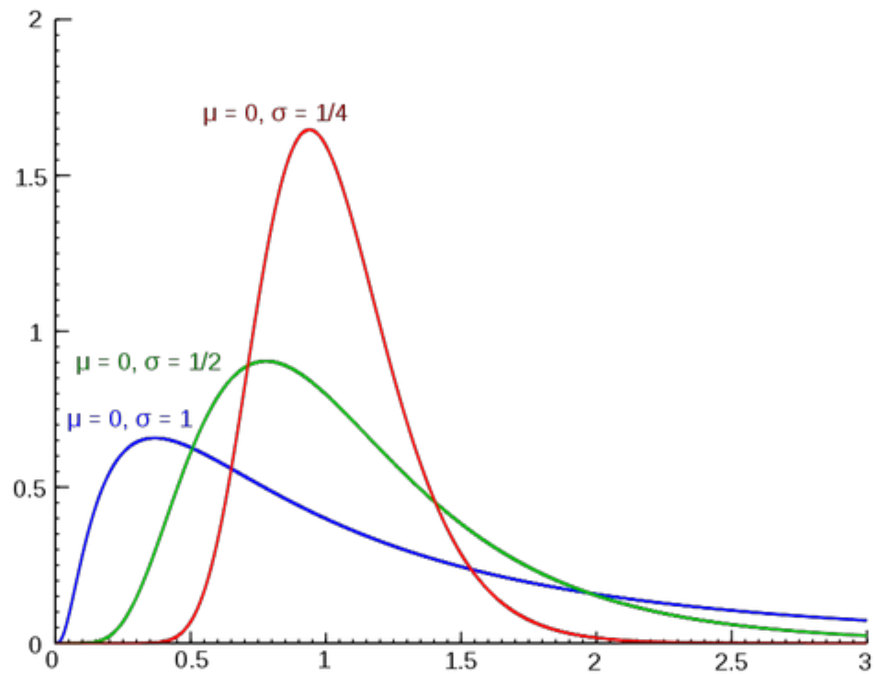




Object names
flickr Tags
Create Size Graph
dog
sofa
cat
tree
window
flickr Images
Collect Observations
Google
✓dog is 83 cm tall
✓dog is ~0.5 m tall
✓dog is 70 - 75 cm tall
✓tree is 20 m tall
✓tree is about 6 m tall
✓tree is 4-12 m tall
MLE
sofa
tree
dog
window
cat

Collective Inference

- Resolving potential inconsistencies across different language and vision estimates
- Assumption: size follows log-normal distribution
- Size is always positive, thus log-normal instead of normal
- Also motivated by a psychology study (Konkle and Olivia 2011)



Collective Inference

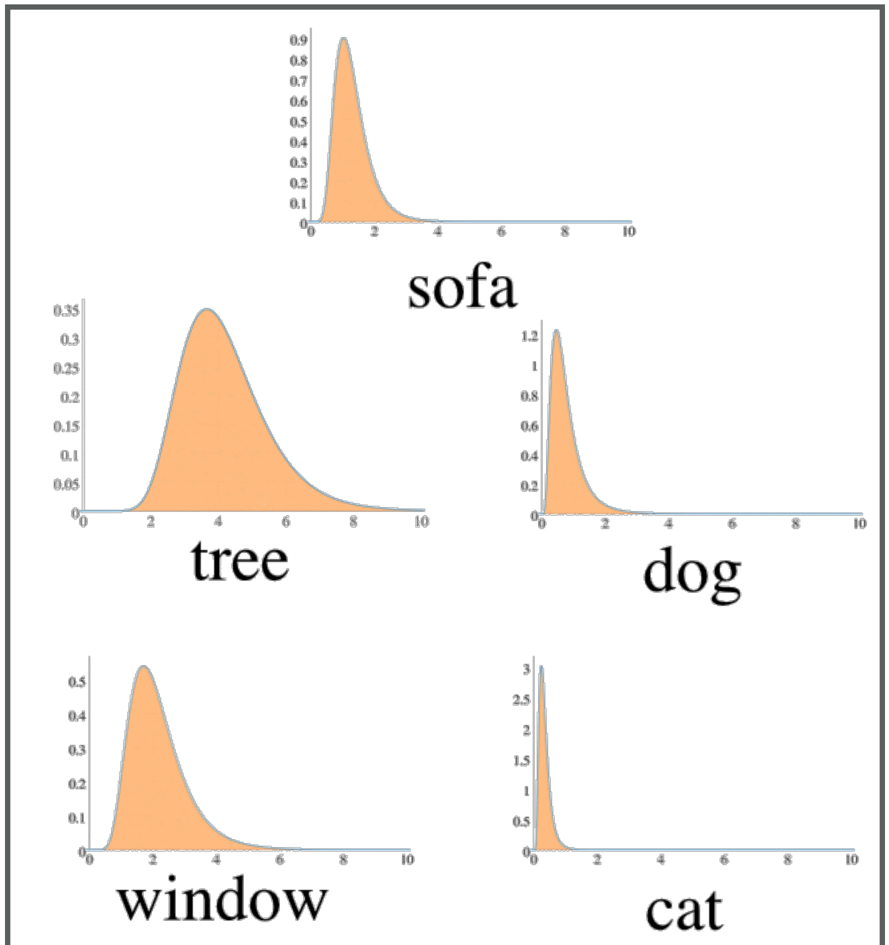
- By optimizing LL over the entire graph (MLE)

$$\sum_{(i,j)\in E} \sum_{r=1}^{n_{ij}} \log f(g_i - g_j = y_{ij}^{(r)} | g_i \sim N(\mu_i, \sigma_i^2), g_j \sim N(\mu_j, \sigma_j^2))$$

$$+ \sum_{i\in V} \sum_{r=1}^{n_i} \log f(g_i = y_i^{(r)} | g_i \sim N(\mu_i, \sigma_i^2))$$

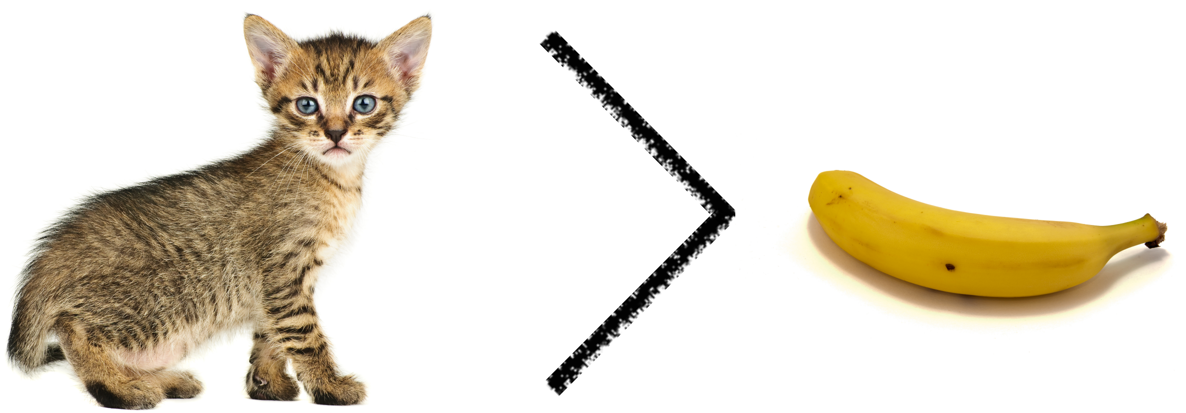
- Coordinate ascent (not convex)

Final output: log-normal dist of sizes



Evaluation

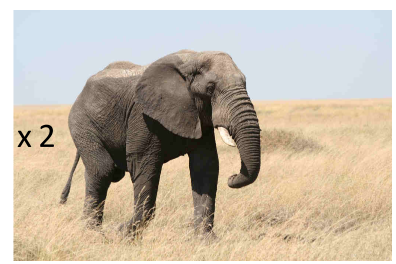
Dataset: annotated labels for 41 physical objects with 486 comparisons.



Baselines

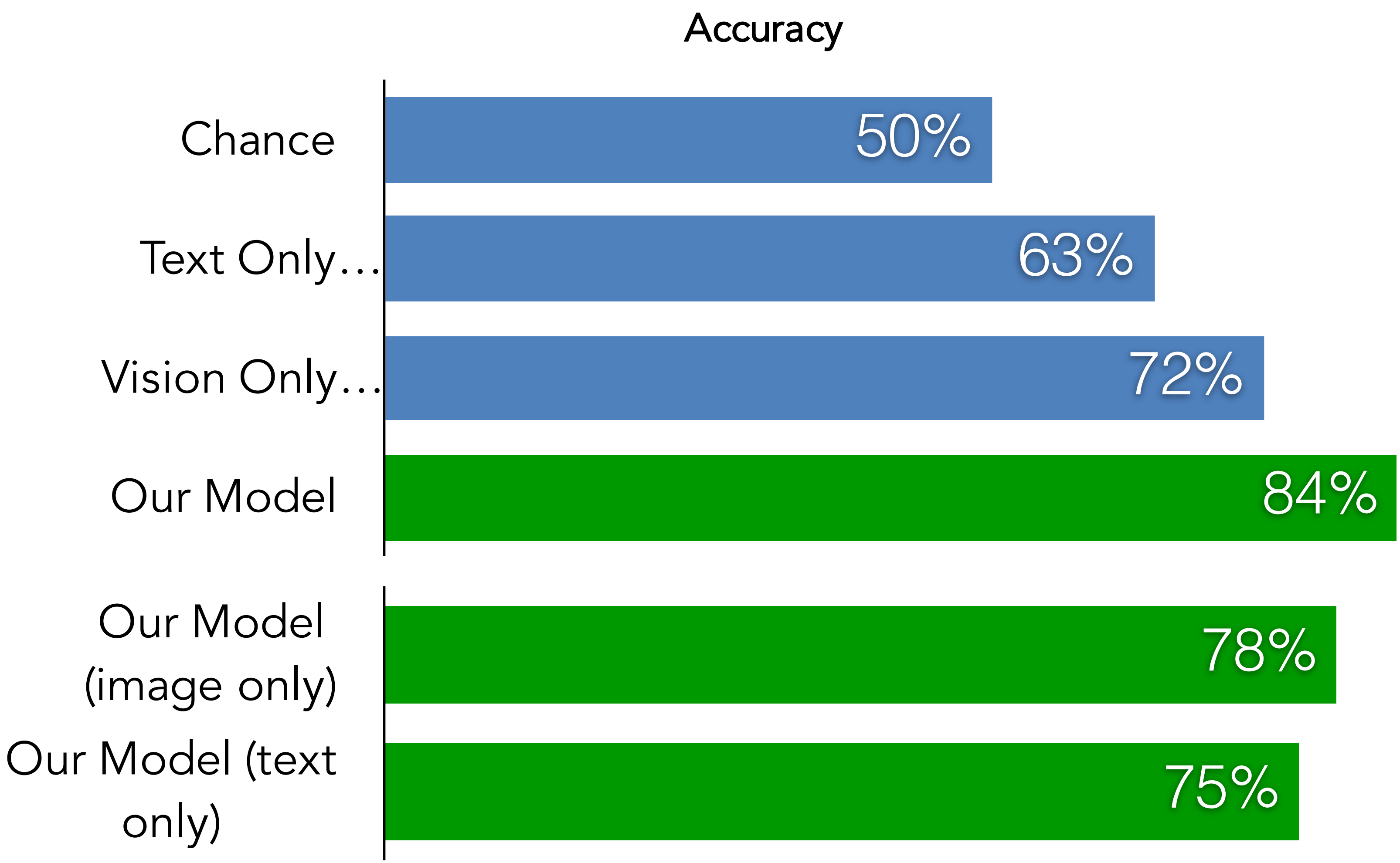
- Text Baseline (inspired by Davidov et al. ACL 2010): Search for some fixed templates and get the mean for each object.
  - e.g. “object is * x * m” and “object’s width is * m”

- Vision Baseline: To answer query (A < B) find a *reliable* path between A and B in in the complete graph and multiply ratios.

$>$

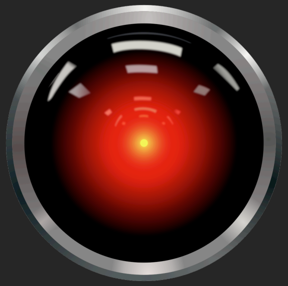
Which of objects A or B is bigger?



To Conclude

- Learning size of objects
- Integrating language and vision
  - to overcome the reporting bias
- Future work: learning physical knowledge

Learning Knowledge about the World
Take II: Entailment



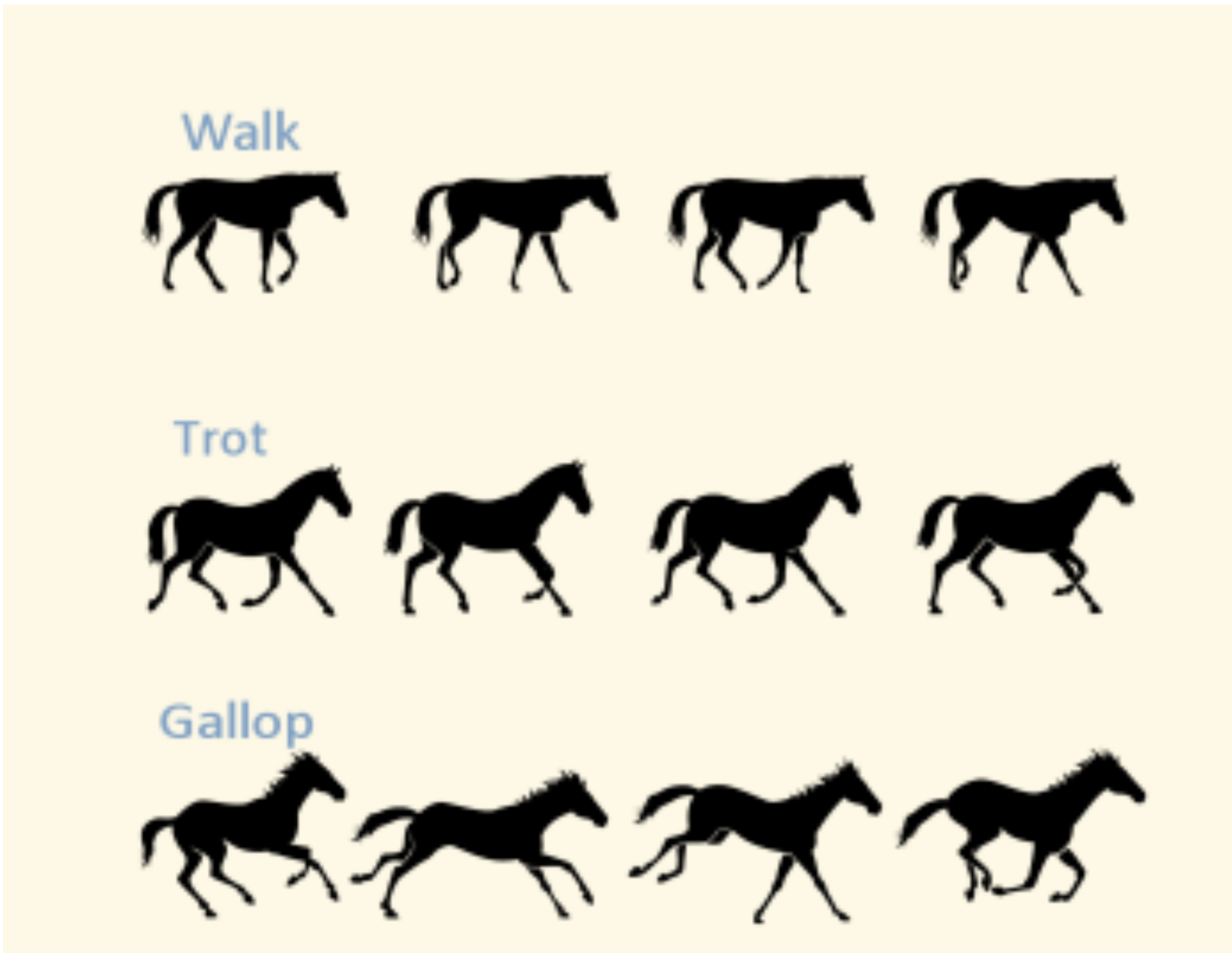
Izadinia et al. @ ICCV 2015

A horse is eating.
Is that horse standing or sitting?

Izadinia et al. @ ICCV 2015

Inspiration: Visual Dictionary

Inspiration: Visual Dictionary



Segment-Phrase Table:
Webly supervised over 50000 instances

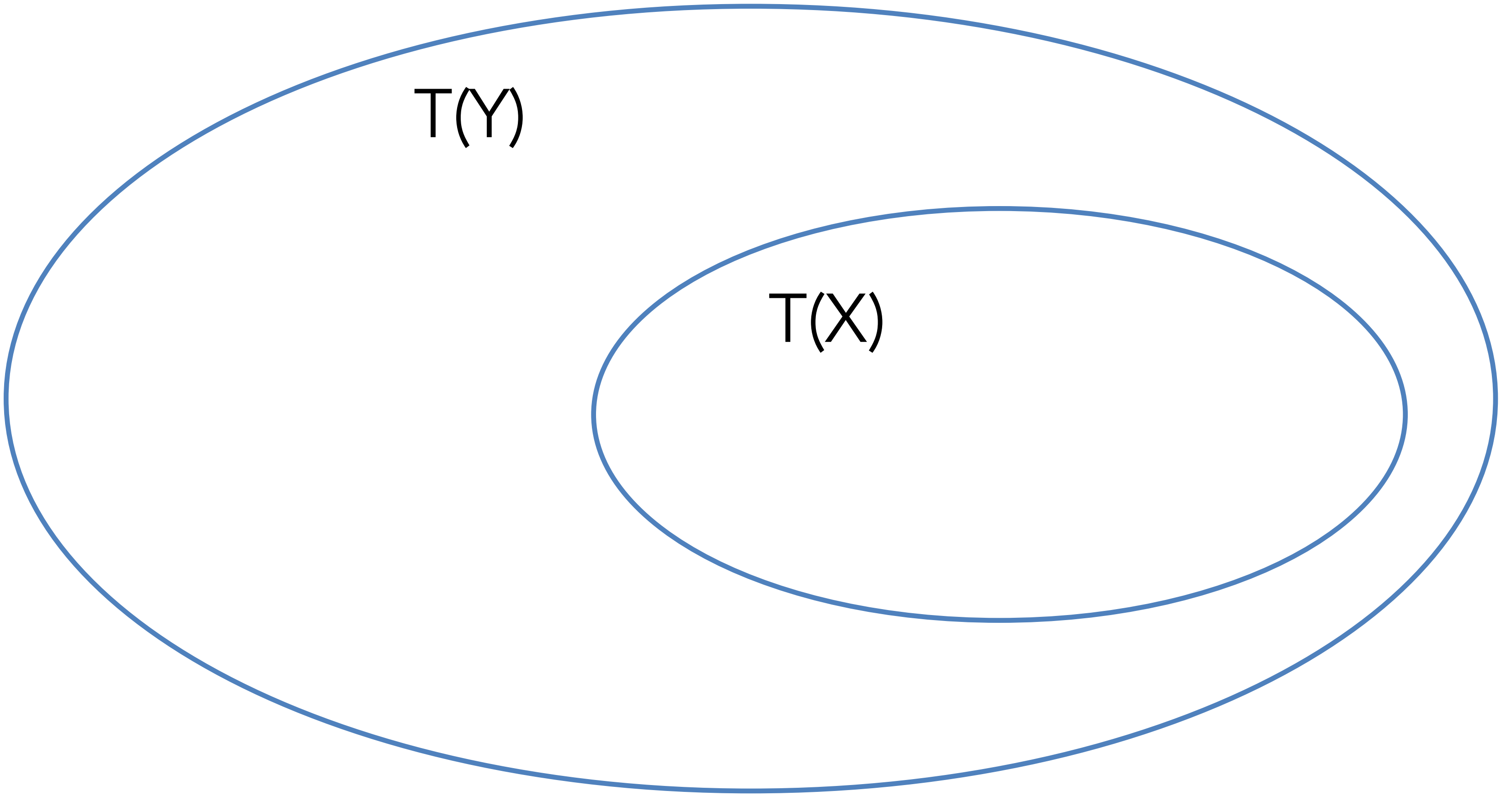
A horse is eating.
Is that horse standing or sitting?

Izadinia et al. @ ICCV 2015

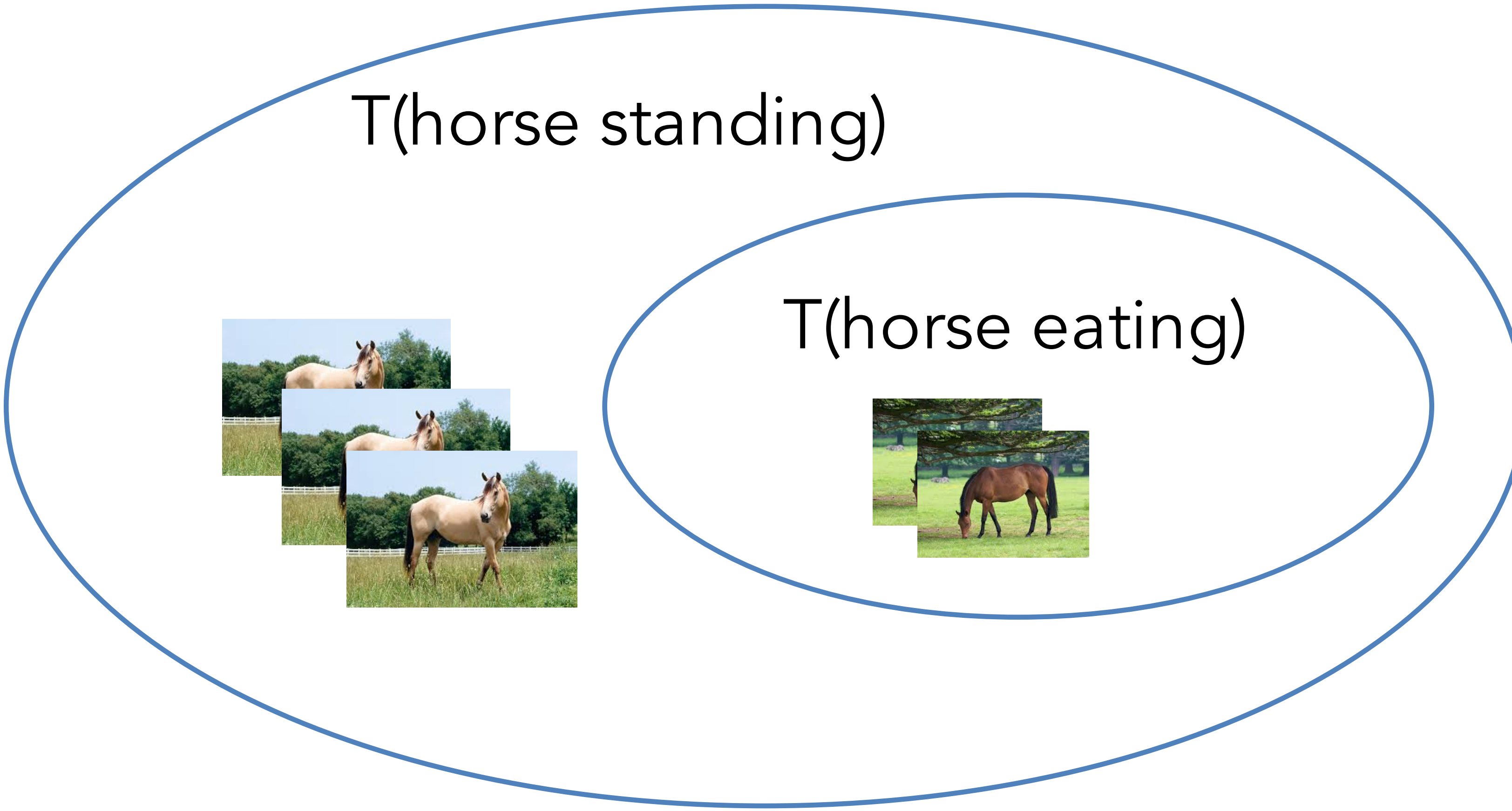
a horse eating => a horse standing

- Reporting bias: do not state the obvious
- Another case where language + vision can help!

Entailment X=>Y



Entailment X=>Y



Entailment X=>Y

$$\text{entail}(X \vDash Y) := Sim_{\overrightarrow{R2I}}(X, Y) - Sim_{\overrightarrow{R2I}}(Y, X)$$

$Sim_{\overrightarrow{R2I}}(X, Y)$ = average asymmetric region-to-image similarity measure (Kim and Grauman 2010) using top K segmentation masks

Global Inference

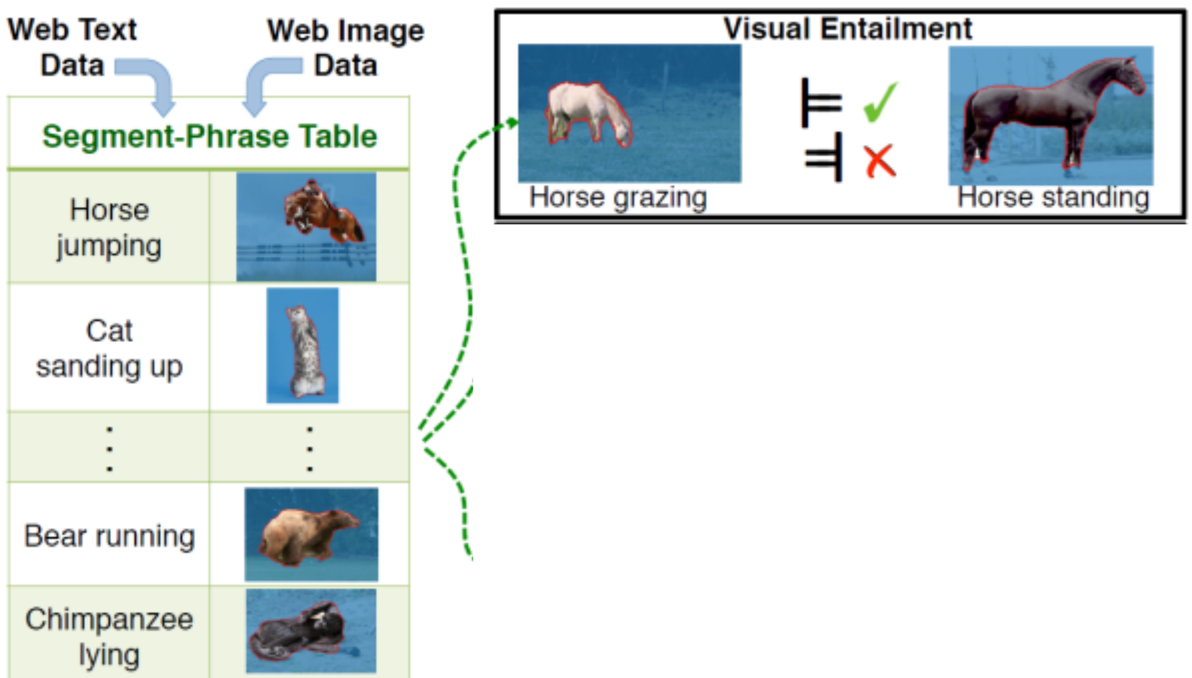
- Transitivity of entailment relations

$$\max \sum_{x \neq y} \text{entail}_{xy} W_{xy} - \lambda |W| \quad s.t. \quad W_{xy} \in \{0, 1\}$$

$$\forall x, y, z \in \mathcal{V}, W_{xy} + W_{yz} - W_{xz} \leq 1$$

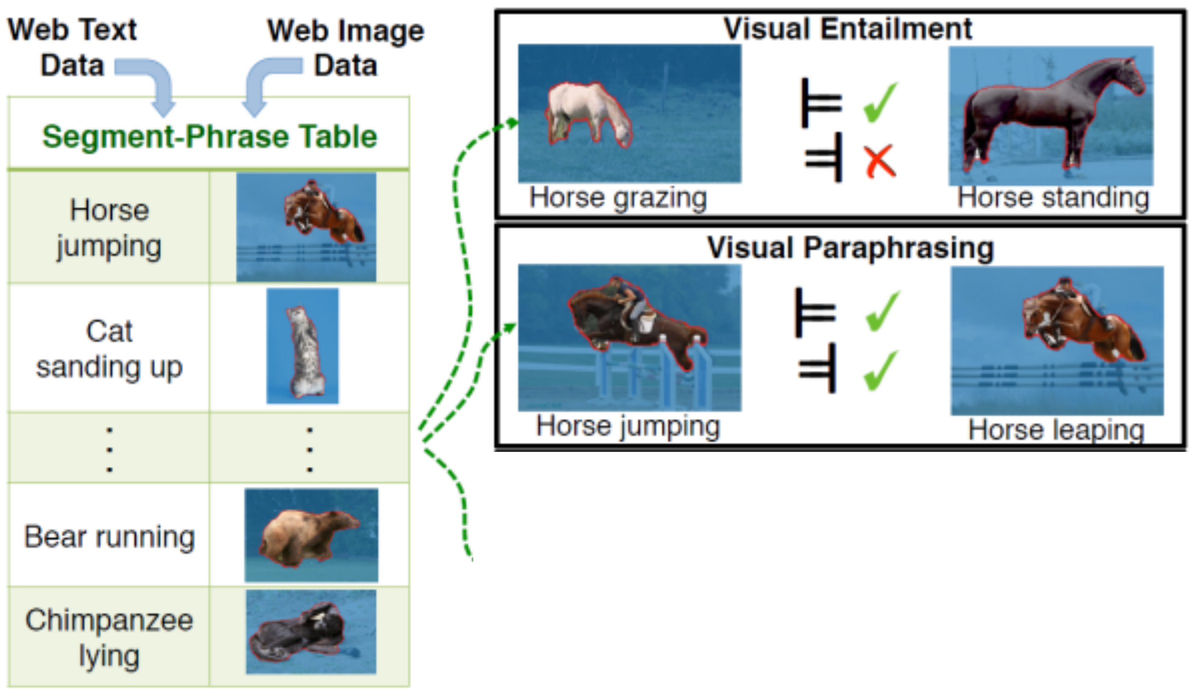
Visual Semantic Tasks

1. Visual Entailment



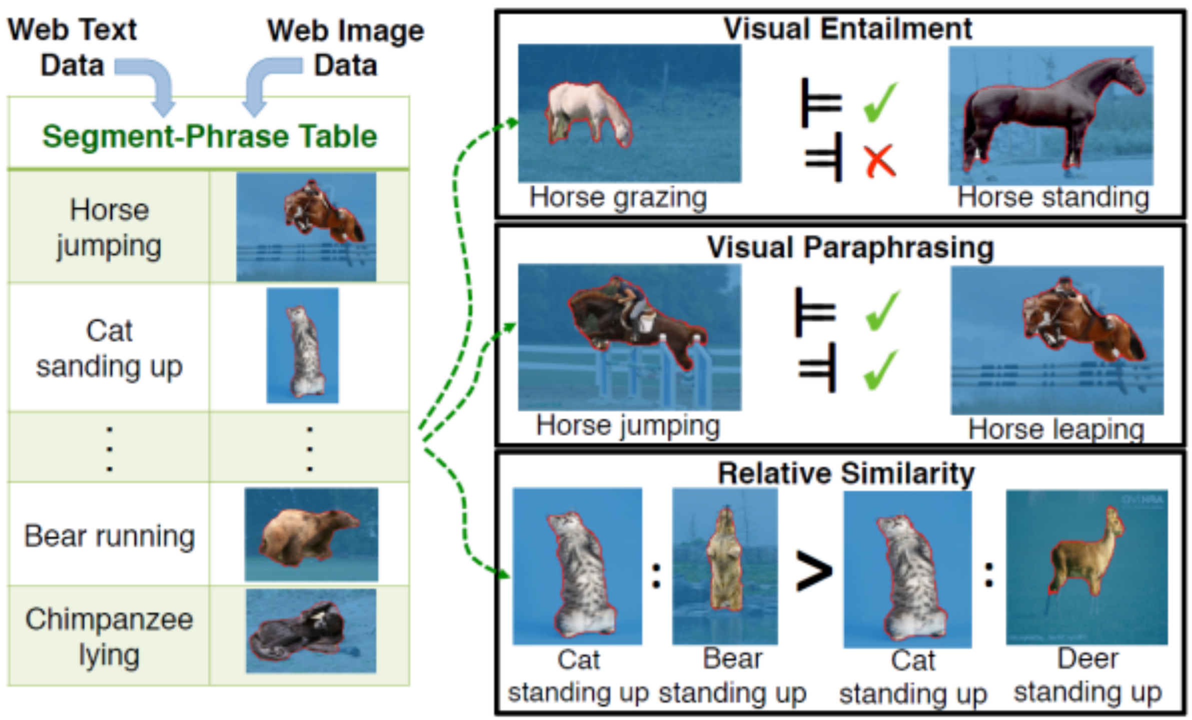
Visual Semantic Tasks

1. Visual Entailment
2. Visual Paraphrasing



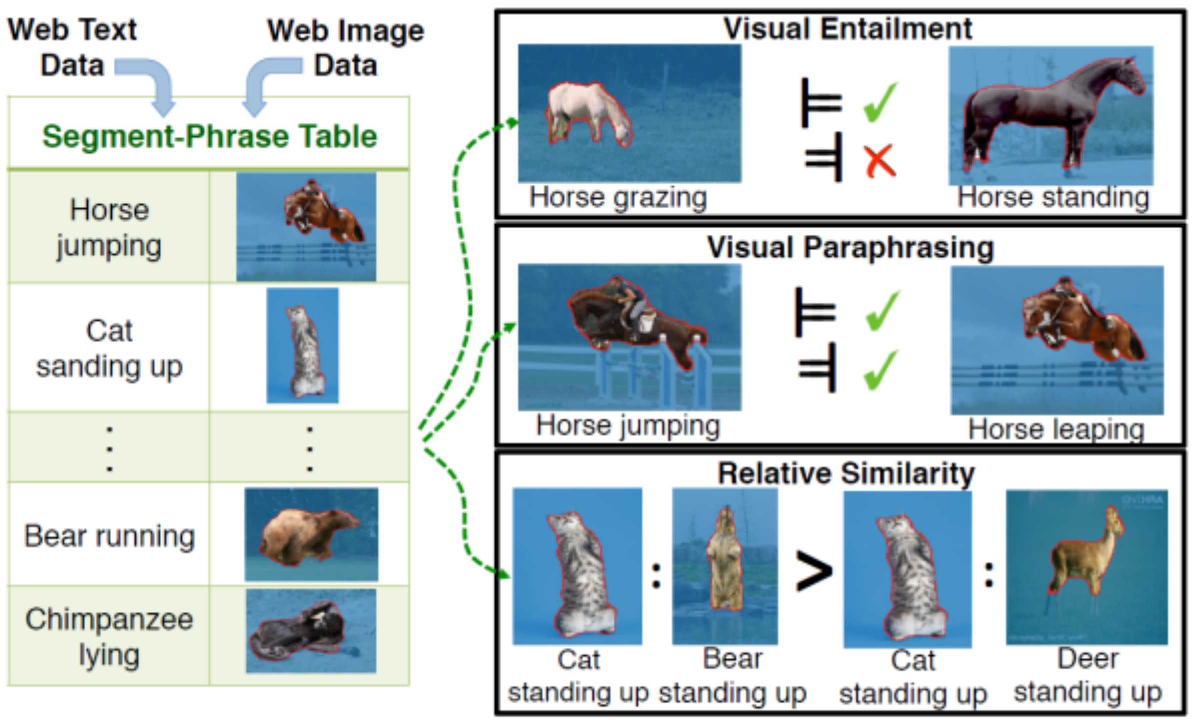
Visual Semantic Tasks

1. Visual Entailment
2. Visual Paraphrasing
3. Semantic Similarity

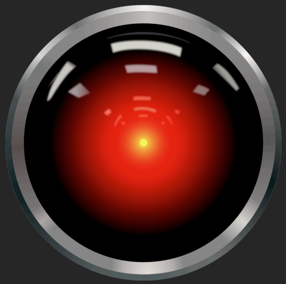


To Conclude

- Segment-Phrase Table
  - Translation dictionary between images and text
- Can learn visual entailment and paraphrases



Learning Knowledge about the World
Take III: Cooking with Action Diagrams



Kiddon et al. @ EMNLP 2015

Interpreting Natural Language Instructions as Action Diagrams

Smart devices and personal robots

- executing commands in natural language instructions
- not just one line command, but a sequence of commands

Step 1: interpret instructions as action diagrams

Instructional Recipes

Blueberry Muffins

Ingredients

1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

Procedure

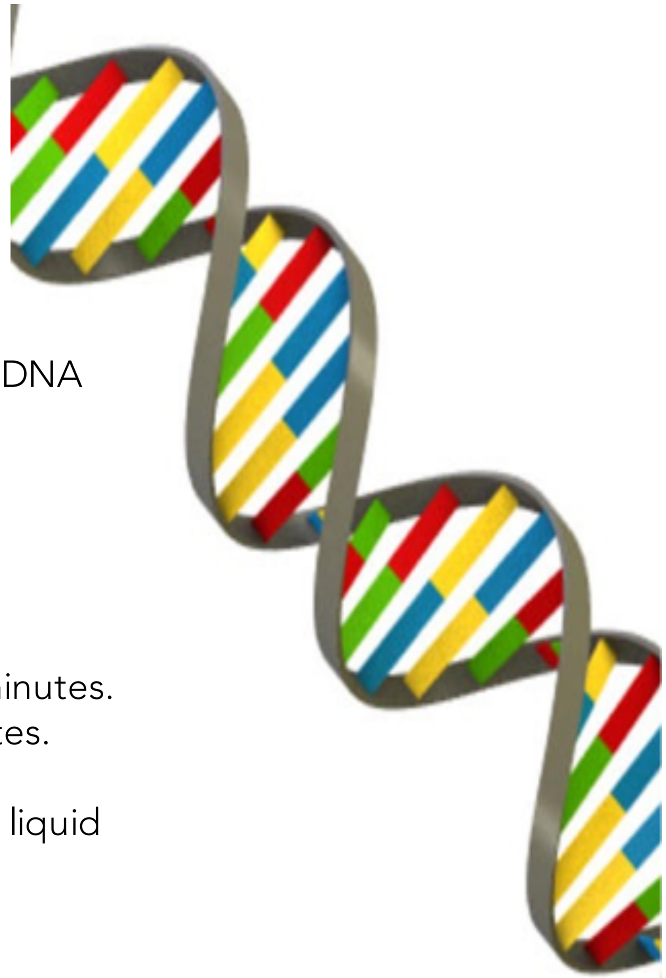
1. Preheat oven to 400 degrees F. Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.
3. Bake for 20 minutes. Serve hot.

http://allrecipes.com/Recipe/Blueberry-Muffins-I/

From Kitchen to Biology Labs

DNA Precipitation

Materials
3M NaOAc pH 5.2
EtOH 95%
Glycogen (optional)



Procedure
1. Add 0.1 volumes of 3M Sodium Acetate solution to 1 volume of DNA sample.
2. Add 1ul Glycogen to the DNA sample.
3. Add 2 volumes of 95% EtOH to the DNA Sample.
4. Store the solution overnight at -20°C or for 30 minutes at -80°C.
5. Centrifuge the solution at maximum speed for least 15 minutes.
6. Decant and discard the supernatant.
7. (Optional) Add 1 ml of 70% EtOH to the pellet and let sit for 5 minutes.
8. (Optional) Centrifuge the sample at maximum speed for 5 minutes.
9. (Optional) Decant and Discard the supernatant.
10. Air-dry the pellet for 10-15 minutes at room temperature until all liquid is gone.
11. Resuspend in desired volume of water or buffer

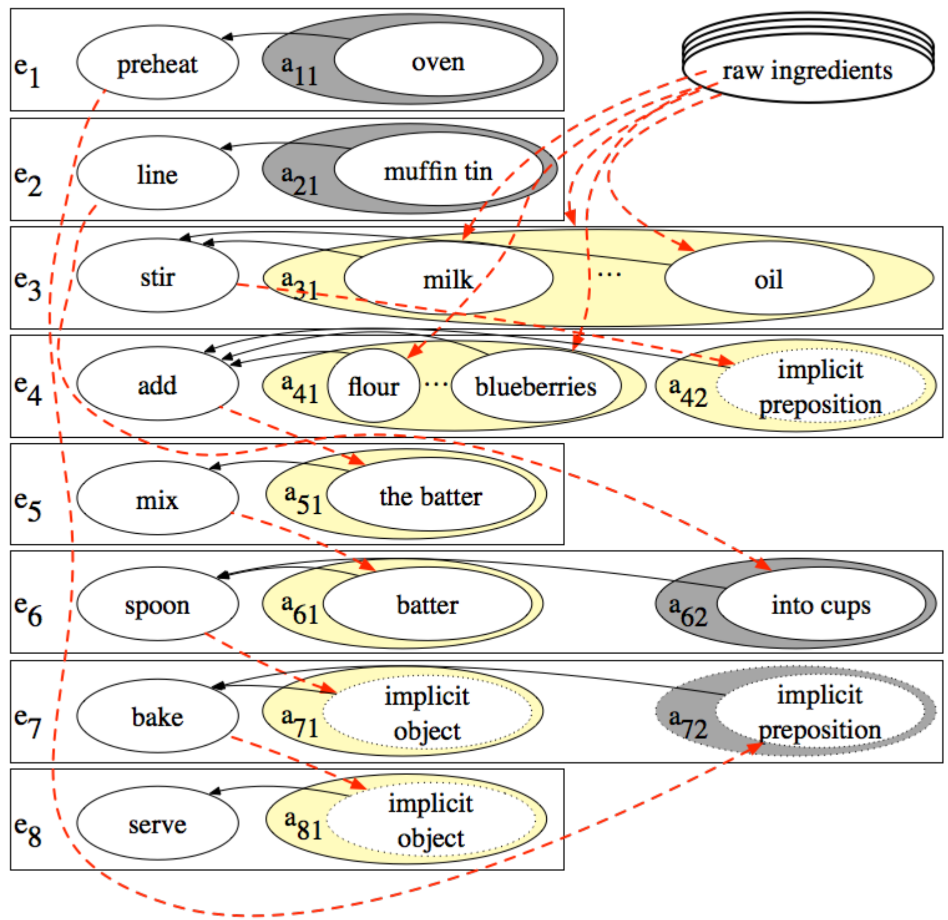
http://openwetware.org/wiki/DNA_Precipitation

Action graph for blueberry muffins

Blueberry Muffins

Ingredients
1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

Procedure
1. Preheat oven to 400 degrees F (205 degrees C). Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.
3. Bake for 20 minutes. Serve hot.



Finding best action graph

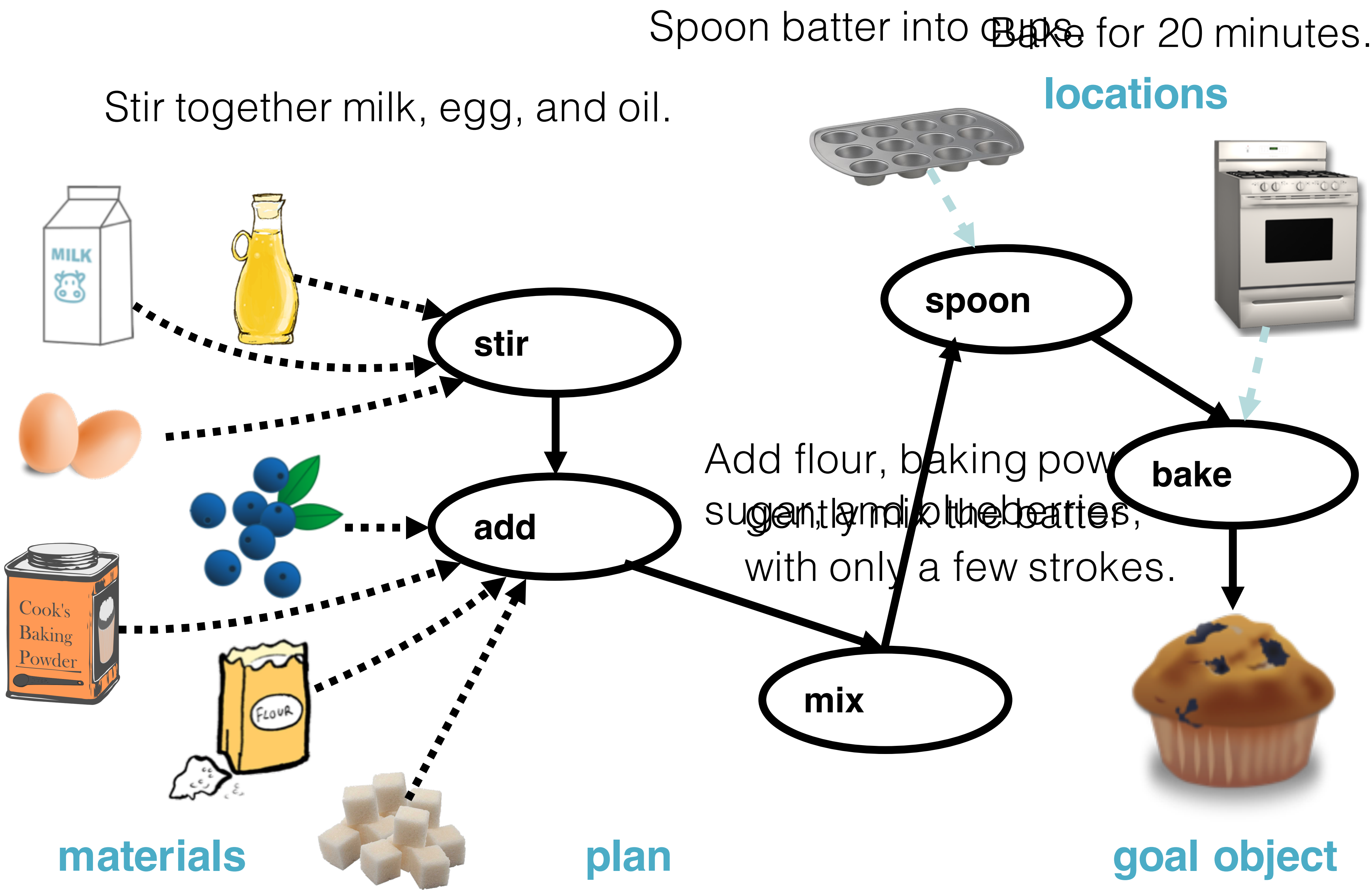
Stir together milk, egg, and oil.

Add flour, baking powder, sugar, and blueberries;

Gently mix the batter with only a few strokes.

Spoon batter into cups.

Bake for 20 minutes.



Spoon batter into cups.
Bake for 20 minutes.
locations
Stir together milk, egg, and oil.
MILK
stir
spoon
bake
Add flour, baking powder, sugar, and blueberries,
gently mix the batter with only a few strokes.
add
Cook's Baking Powder
FLOUR
mix
materials
plan
goal object

Semantic challenges

- Traditional parsers have trouble with imperatives
  - Grease with butter. **Grease = noun?**
- Elided arguments are common.
  - Bake for 30 minutes. **Bake what? Bake where?**
- Referring expressions use physical properties
  - Whisk eggs. Add flour. Fold sugar into **the wet mixture.**

Action graph for blueberry muffins

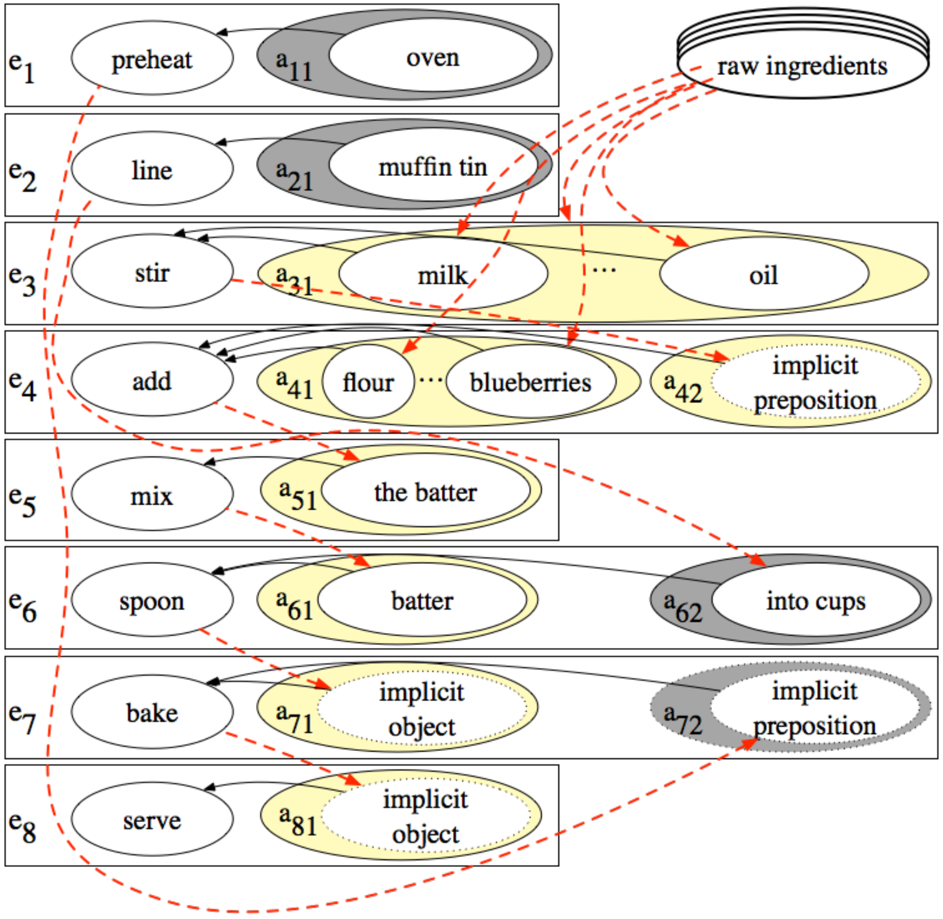
Blueberry Muffins

Ingredients

1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

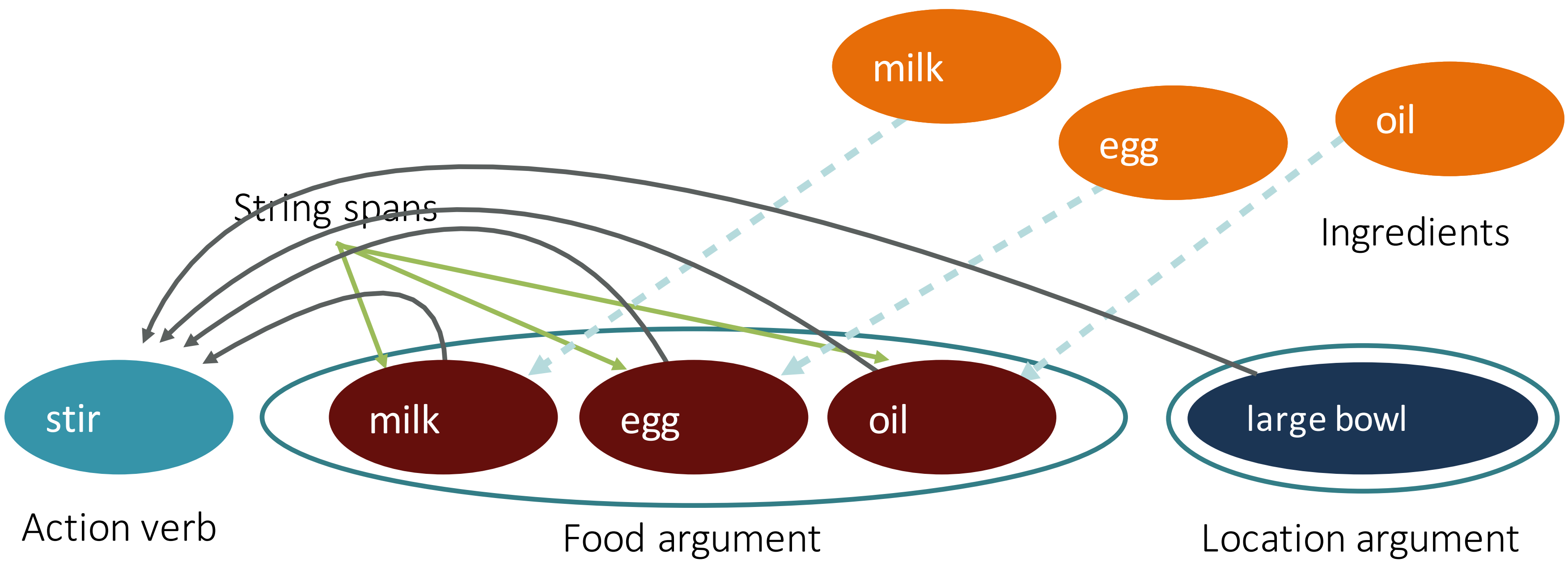
Procedure

1. Preheat oven to 400 degrees F (205 degrees C). Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.
3. Bake for 20 minutes. Serve hot.



Action graphs

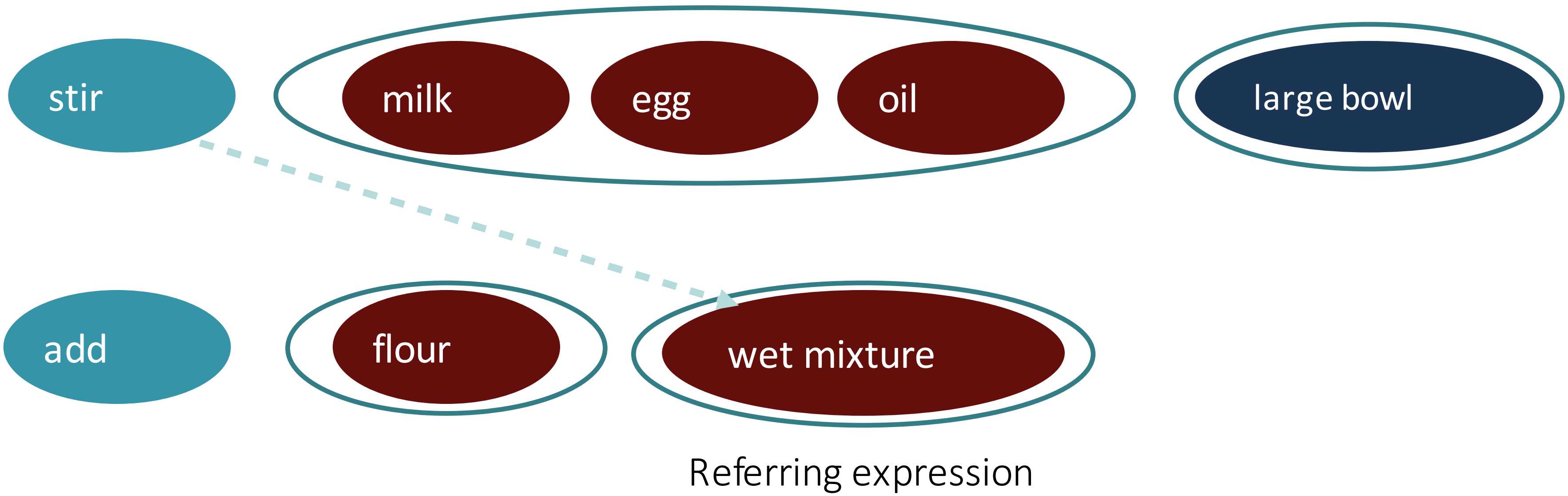
Model the **flow** of ingredients as a DAG



"In a large bowl, stir together milk, egg, and oil."

Action graphs

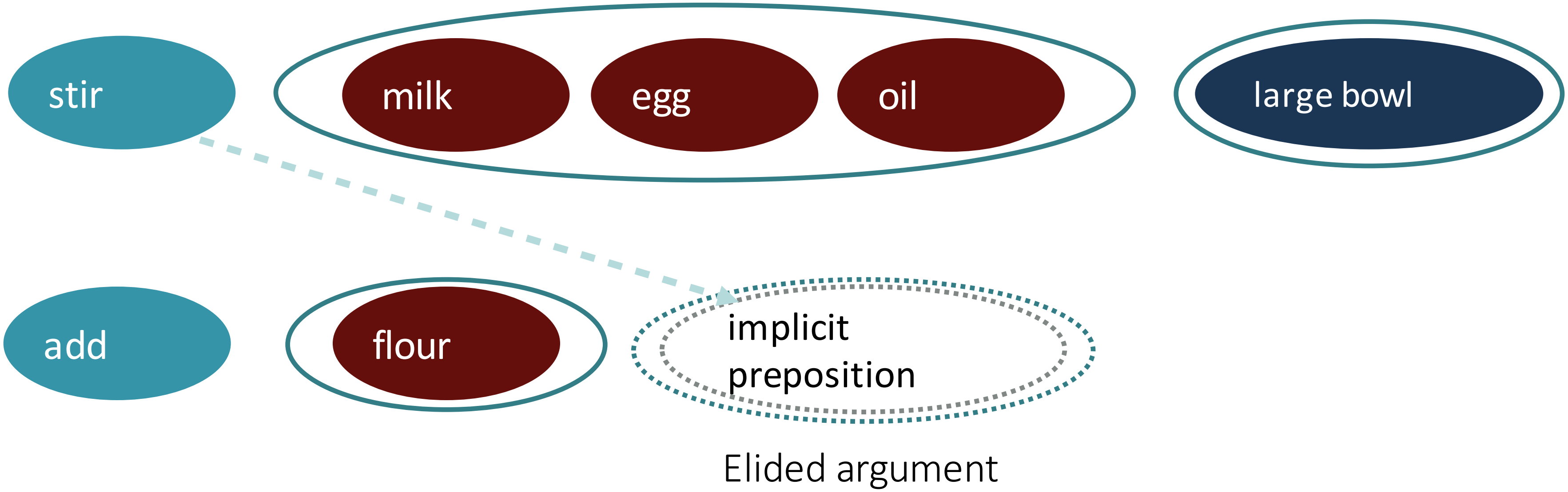
Model the **flow** of ingredients as a DAG



"In a large bowl, stir together milk, egg, and oil.
Add flour to the wet mixture."

Action graphs

Model the **flow** of ingredients as a DAG



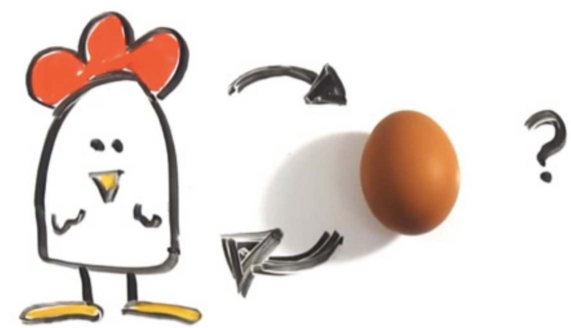
"In a large bowl, stir together milk, egg, and oil.
Add flour."

Related Work

- Maeta et al. 2015,
- Mori et al. 2014
- Tasse and Smith 2008

Unsupervised Learning (Kiddon et al. 2015)

- Chicken and Egg
  - Parsing (unstructured text → action graph) requires knowledge
  - Knowledge requires parsing



- Model:
  - Probabilistic Model
- Learning:
  - Expectation-Maximization

Probability model P(C,R) (Kiddon et al. 2015)

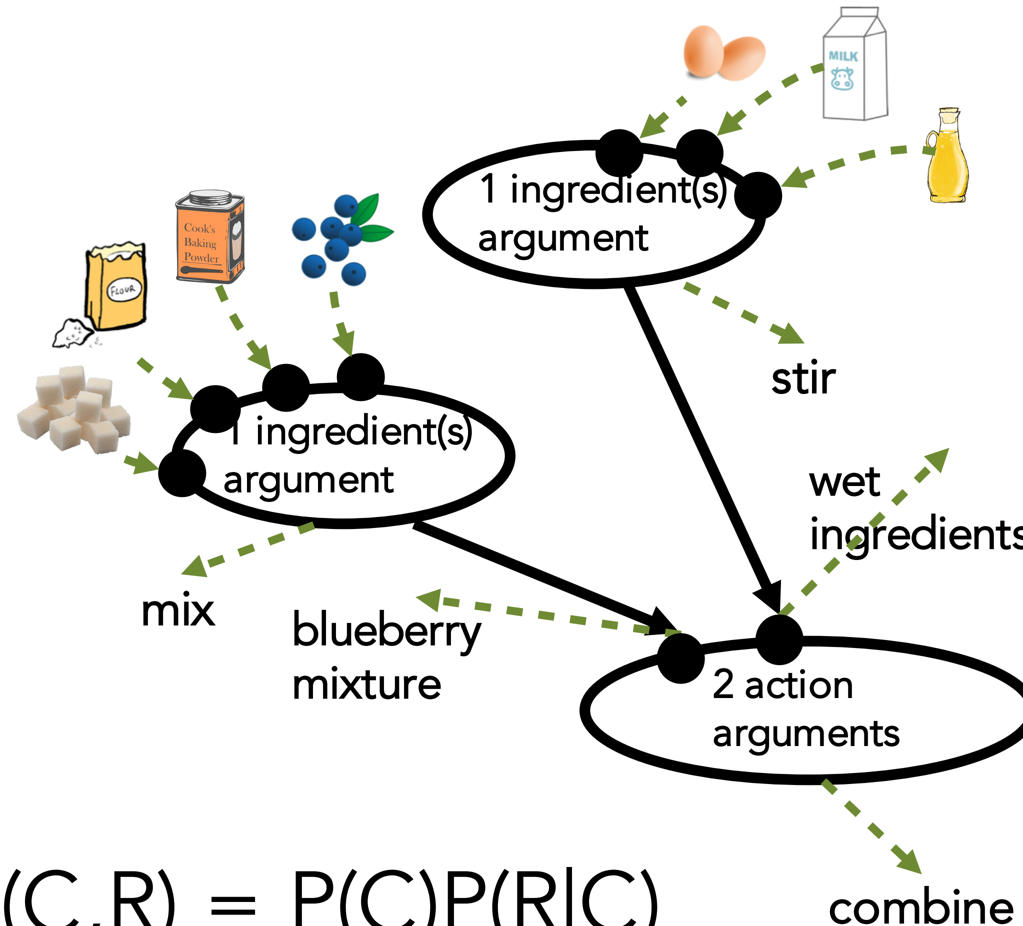
- **Input:** A set of connections C and a recipe R segmented (**Sec. 6**) into its actions $\{e_1 = (v_1, \mathbf{a}_1), \ldots, e_n = (v_n, \mathbf{a}_n)\}$
- The joint probability of C and R is $P(C, R) = P(C)P(R|C)$, each defined below:

1. **Connections Prior (Sec. 3.1)**: $P(C) = \prod_i P(\mathbf{d}_i|\mathbf{d}_1, \ldots, \mathbf{d}_{i-1})$
   Define $\mathbf{d}_i$ as the list of connections with destination index i. Let $c_p = (o, i, j, k, t^{syn}, t^{sem}) \in \mathbf{d}_i$. Then,
   - $P(\mathbf{d}_i|\mathbf{d}_1, \ldots, \mathbf{d}_{i-1}) = P(vs(\mathbf{d_i})) \prod_{c_p \in \mathbf{d_i}} P(\mathbb{1}(o \to s_{ij}^k)|vs(\mathbf{d_i}), \mathbf{d}_1, \ldots, \mathbf{d}_{i-1}, c_1, \ldots, c_{p-1})$
     (a) $P(vs(\mathbf{d_i}))$: multinomial verb signature model (**Sec. 3.1.1**)
     (b) $P(\mathbb{1}(o \to s_{ij}^k)|vs(\mathbf{d_i}), \mathbf{d}_1, \ldots, \mathbf{d}_{i-1}, c_1, \ldots, c_{p-1})$: multinomial connection origin model, conditioned on the verb signature of $\mathbf{d}_i$ and all previous connections (**Sec. 3.1.2**)
2. **Recipe Model (Sec. 3.2)**: $P(R|C) = \prod_i P(e_i|C, e_1, \ldots, e_{i-1})$
   For brevity, define $\mathbf{h}_i = (e_1, \ldots, e_{i-1})$.
   - $P(e_i|C, \mathbf{h}_i) = P(v_i|C, \mathbf{h}_i)P(a_{ij}|C, \mathbf{h}_i)$ (**Sec. 3.2**)
     Define argument a_{ij} by its types and spans, $a_{ij} = (t_{ij}^{syn}, t_{ij}^{sem}, S_{ij})$.
     (a) $P(v_i|C, \mathbf{h}_i) = P(v_i|g_i)$: multinomial verb distribution conditioned on verb signature (**Sec. 3.2**)
     (b) $P(a_{ij}|C, \mathbf{h}_i) = P(t_{ij}^{syn}, t_{ij}^{sem}|C, \mathbf{h}_i) \prod_{s_{ij}^k \in S_{ij}} P(s_{ij}^k|t_{ij}^{syn}, t_{ij}^{sem}, C, \mathbf{h}_i)$
       i. $P(t_{ij}^{syn}, t_{ij}^{sem}|C, \mathbf{h}_i)$: deterministic argument types model given connections (**Sec. 3.2.1**)
       ii. $P(s_{ij}^k|t_{ij}^{syn}, t_{ij}^{sem}, C, \mathbf{h}_i)$: string span model computed by case (**Sec. 3.2.2**):
         A. $t_{ij}^{sem} = food$ and $origin(s_{ij}^k) \neq 0$: IBM Model 1 generating composites (**Part-composite model**)
         B. $t_{ij}^{sem} = food$ and $origin(s_{ij}^k) = 0$: naïve Bayes model generating raw food references (**Raw food model**)
         C. $t_{ij}^{sem} = location$: model for generating location referring expressions (**Location model**)

Figure 2: Summary of the joint probabilistic model $P(C, R)$ over connection set C and recipe R.

Probabilistic model

- Assume we are given a preprocessed recipe text R that has been segmented into actions
- Probabilistic model over action graphs to determine most likely connections C for the recipe



$$P(C,R) = P(C)P(R|C)$$

prior over connections

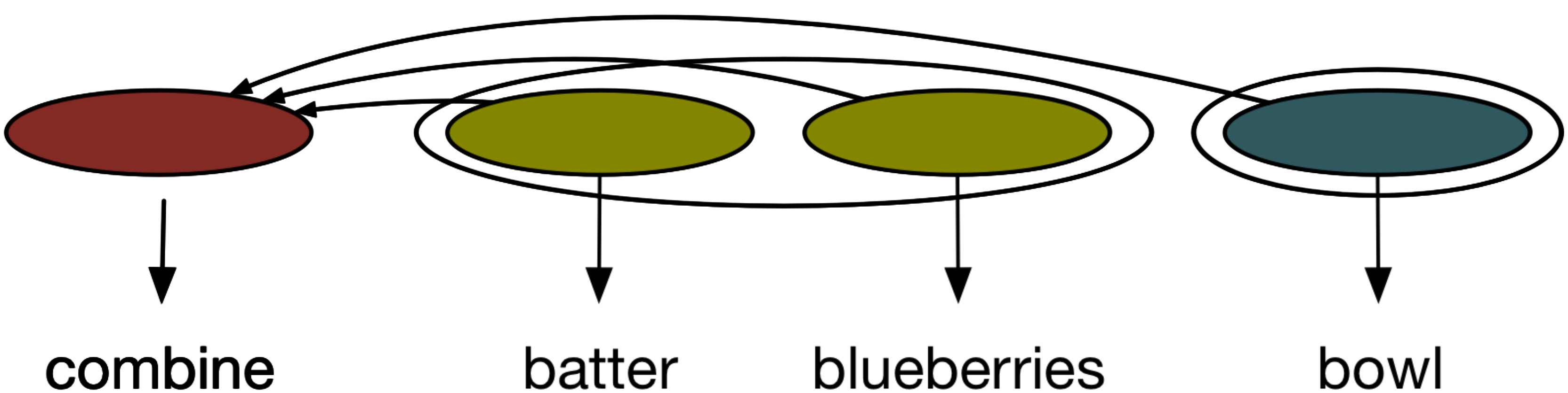
probability of recipe text given connections

Recipe distribution: P(R|C)

- R is a sequence of actions $e_1, \ldots, e_n$

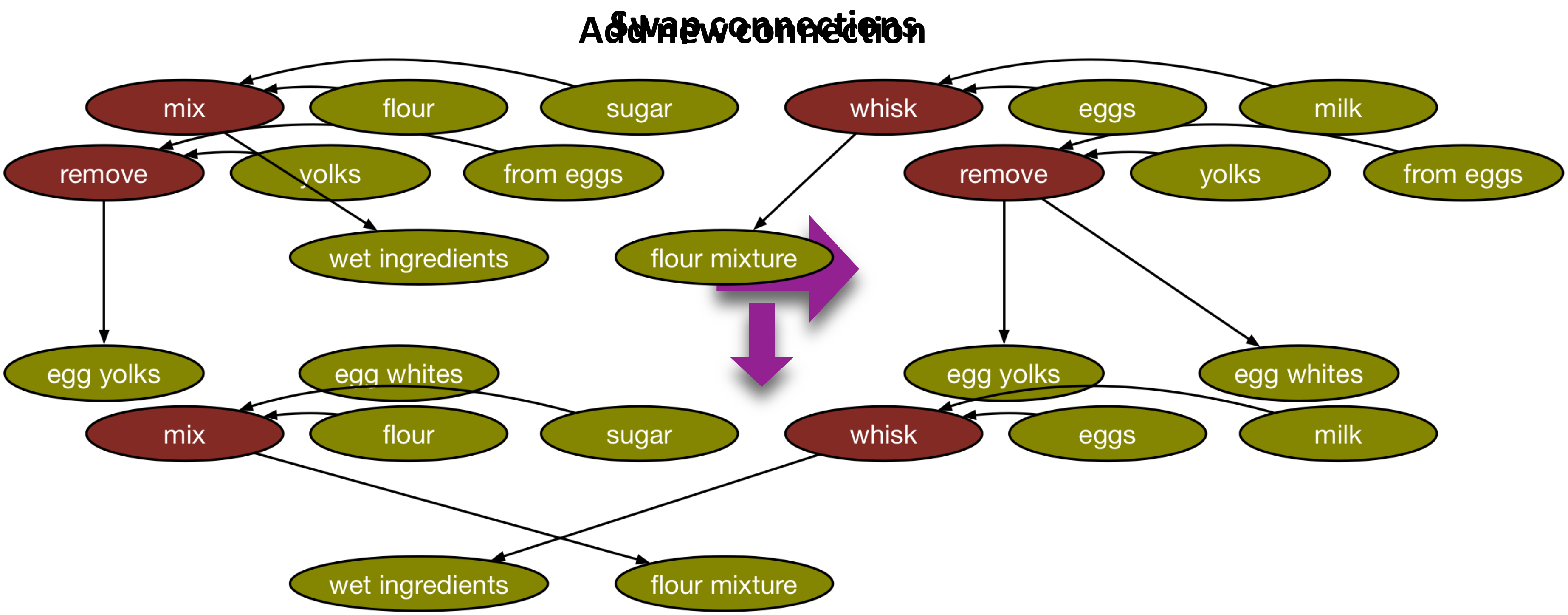
$$P(R|C) = \prod_i P(e_i|C, e_1, \ldots, e_{i-1})$$

- Actions decompose into the probability of the verbs, arguments, and spans



Local search

- Initialize with sequential connections
- Score local search operators and greedily apply



Model learning

- Unsupervised hard EM method
- First, initialize models. Then:

Recurse:

- **E-step:** Update $C \leftarrow \text{argmax}_C\, P(C,R)$ for each R in dataset using local search

- **M-step**: Update parameters of $P(C,R)$ using action graphs generated in E-step

Knowledge in the Model

- **Part-composite model**: how likely it is to generate a composite word given the incoming ingredients/raw materials
  - P("dressing" | "oil" "vinegar") > P("batter" | "oil" "vinegar")
- **Raw materials model**: how likely a word is to be a initial reference
  - P("batter" | initial reference) < P("flour" | initial reference)
- **Location model**: how likely a location is given the action verb

Learned cooking knowledge

Learned good composite words for different ingredients

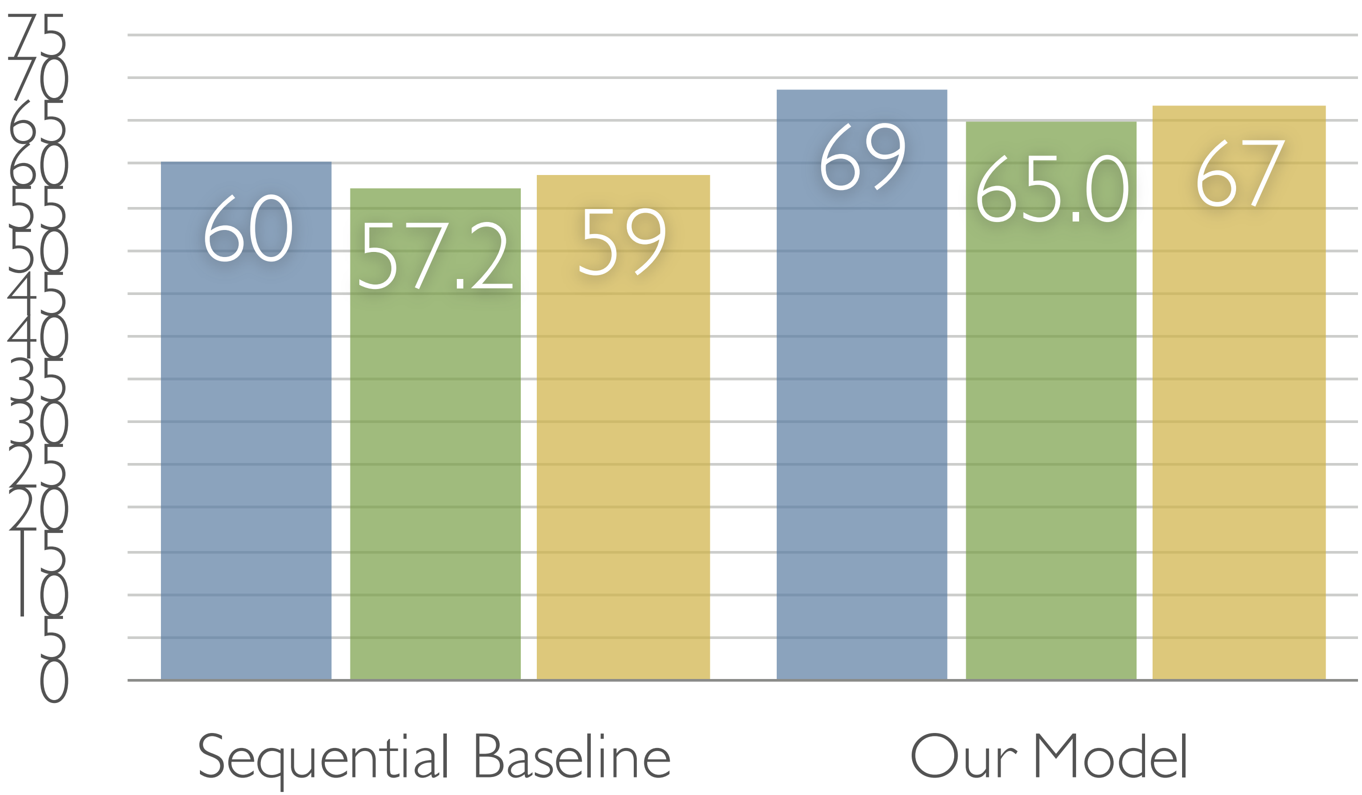
eggs	egg, yolk, mixture, noodles, whites
beef	beef, mixture, grease, meat, excess
flour	flour, mixture, dough, batter, top, crust

Learned selectional preferences for verb

- **add** is 58% likely to have two arguments that are not both raw materials
- **bake** is 95% likely to have one non-raw material argument

Evaluation

- Cooking recipe domain, 2456 recipes, 20 dish types
- 100 manually-annotated gold-standard recipes



To Conclude

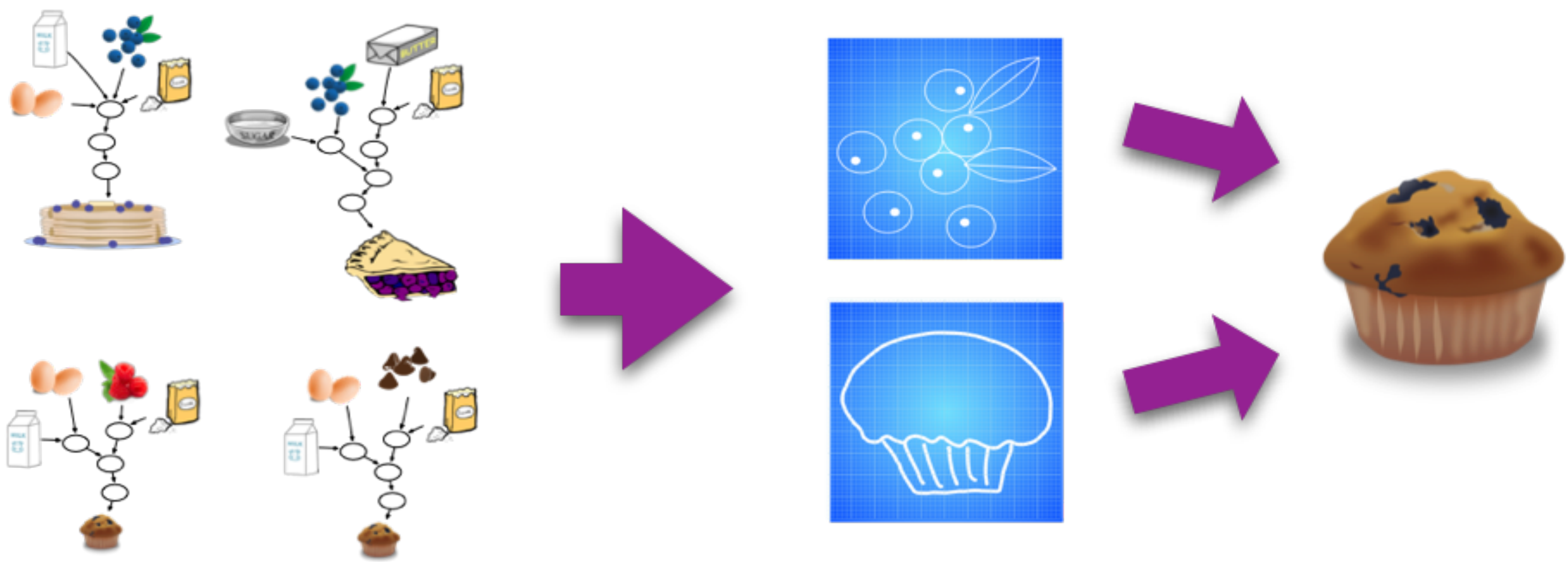
- Unsupervised parsing of instructional recipes to action diagrams
- Possible due to repeated patterns in naturally existing data
- Knowledge is a recurring theme.

What's Next: Composing a New Recipe

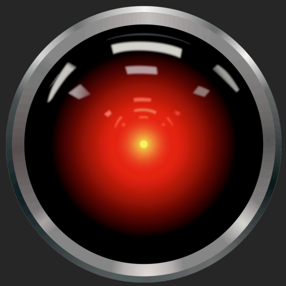
Compose new recipes given a recipe title (or what's in the fridge)!

- With or without explicit meaning representation
- New challenge: generating a **cohesive discourse**
- **zero-shot learning** for recipes

Grounding instructions with multimodal perception



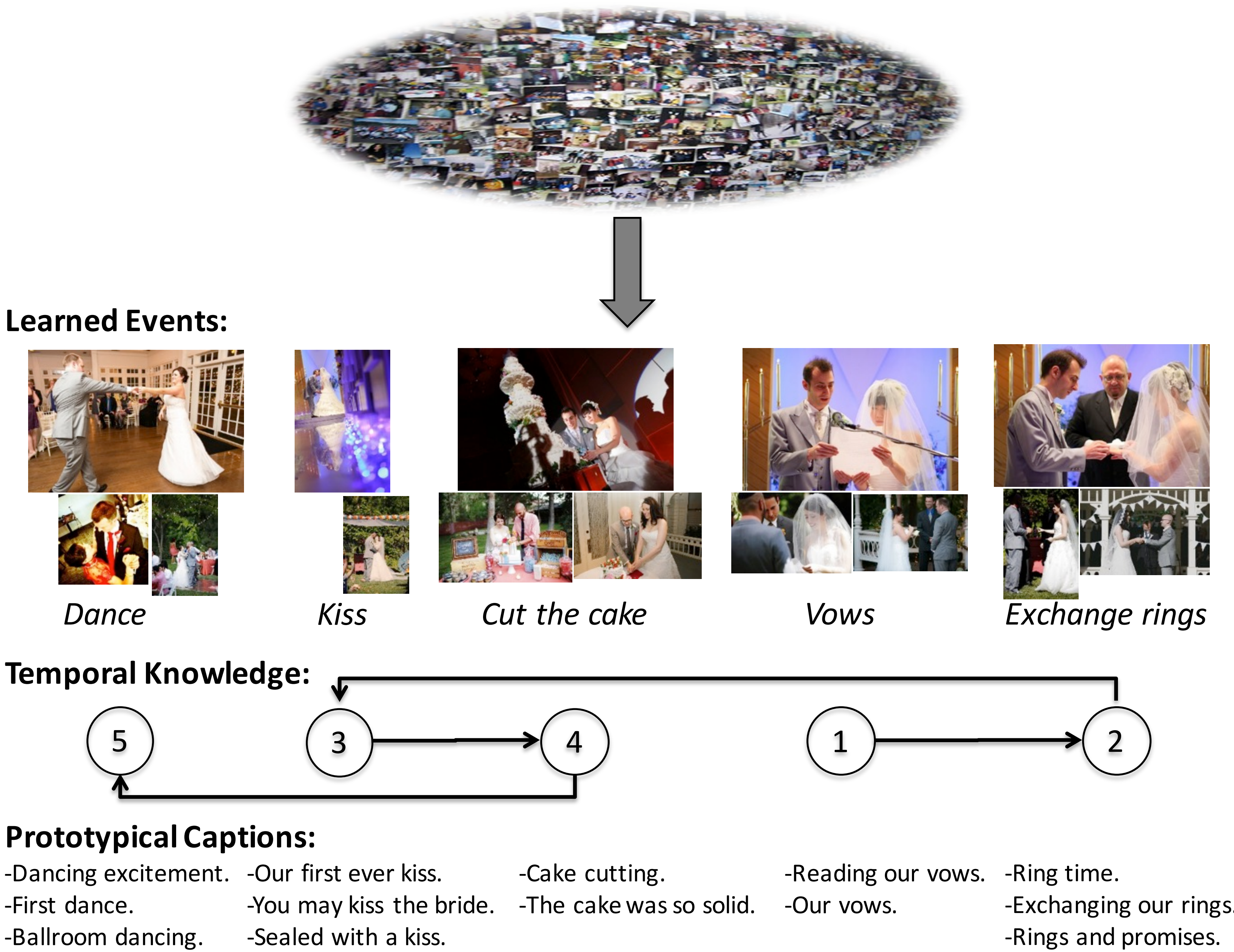
Learning Knowledge about the World
Take IV: Prototypical Events



Bosselut et al. @ ACL 2016

What makes a wedding a wedding?

Bosselut et al. @ ACL 2016



Learned Events:
Dance
Kiss
Cut the cake
Vows
Exchange rings
Temporal Knowledge:
5
3
4
1
2
Prototypical Captions:
-Dancing excitement.
-First dance.
-Ballroom dancing.
-Our first ever kiss.
-You may kiss the bride.
-Sealed with a kiss.
-Cake cutting.
-The cake was so solid.
-Reading our vows.
-Our vows.
-Ring time.
-Exchanging our rings.
-Rings and promises.

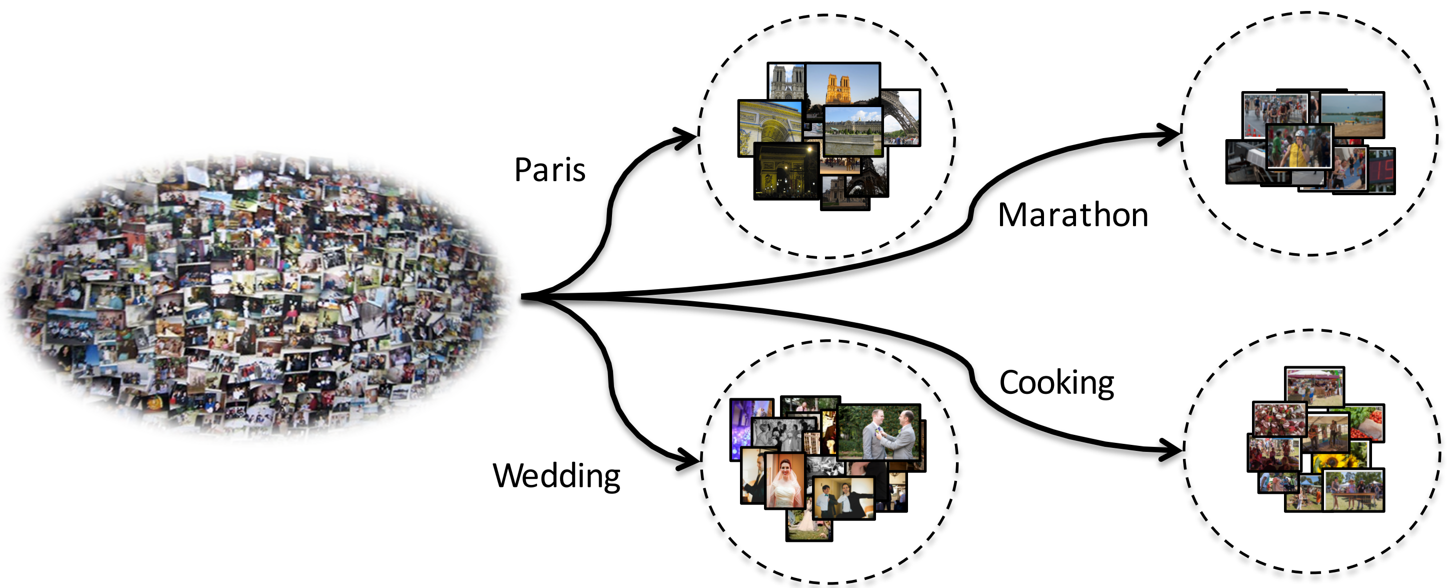
Circular Dependency

Better knowledge about stereotypical event structure

Better understanding of a new photo album

Data Compilation

- 12 common life scenarios
  - Wedding, Paris Trip, New York Trip, Barbecue, Funeral, Independence Day, Cooking, Camping, Marathon, Baby Birth, Christmas, Thanksgiving



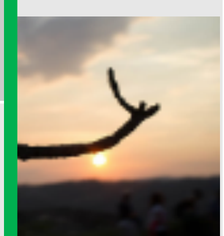
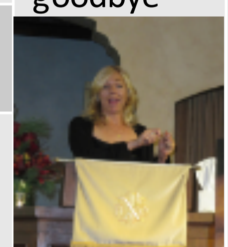
Learning Prototypical Events

- k-means clustering (on language only)

- Multimodal cluster representation
  - Weighted unigram features of content words
  - Visual Features from VGGNet

- Name each cluster with the most common word

Sample Events and Prototypical Captions

Wedding		Camping		Funeral	
aisle 	Walking down the aisle	tent 	Inside out tent	service 	Graveside service
	Bride walking down the aisle		Setting up the tent		The service
vow 	Exchanging vows	fire 	Building the fire	pay 	Paying Respects
	Reading the vows		Around the fire		Respect
	Reciting vows to each other		Getting the fire going		
dance 	First dance	sunset 	Sunset from camp	goodbye 	Saying goodbye
	Everybody dancing		Watching the sunset		
	Dancing the night away		Sunset on the first night		

Learn Temporal Knowledge

- Local transition probabilities – Probability that a photo assigned event A being followed by a photo assigned to event B.

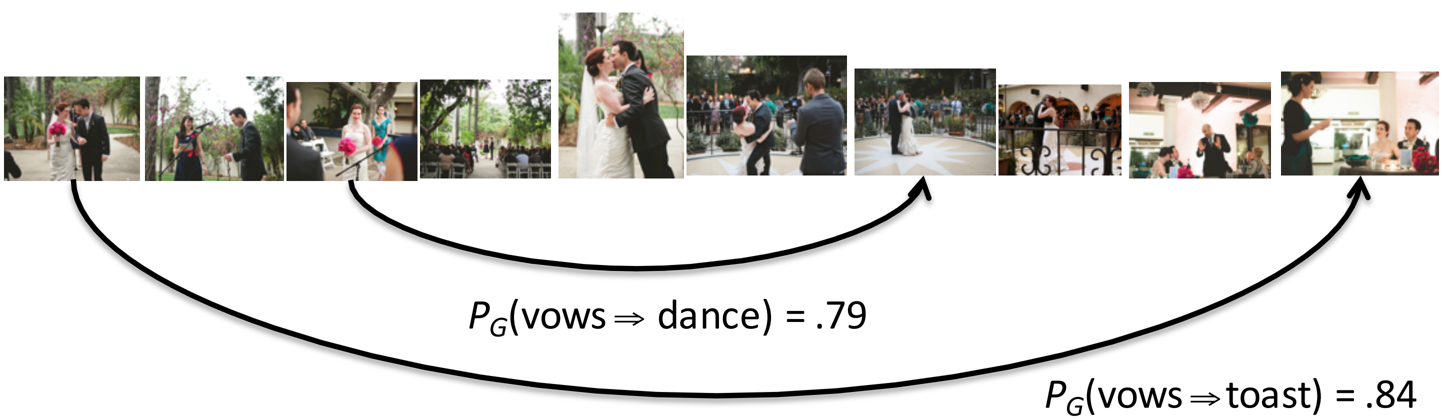
$$P_L(e_k \rightarrow e_l) = \frac{C(e_k \rightarrow e_l)}{\sum_{m=1}^{N} C(e_k \rightarrow e_m)}$$

$P_L(\text{kiss} \rightarrow \text{dance}) = .04$ $P_L(\text{dance} \rightarrow \text{toast}) = .17$

Learn Temporal Knowledge

- Global pairwise ordering probabilities – Probability that a photo assigned event A precedes a photo assigned event B anywhere in the album

$$P_G(e_k \Rightarrow e_l) = \frac{C(e_k \Rightarrow e_l)}{C(e_k \Rightarrow e_l) + C(e_l \Rightarrow e_k)}$$



Circular Dependency

Better knowledge about stereotypical event structure

Better understanding of a new photo album

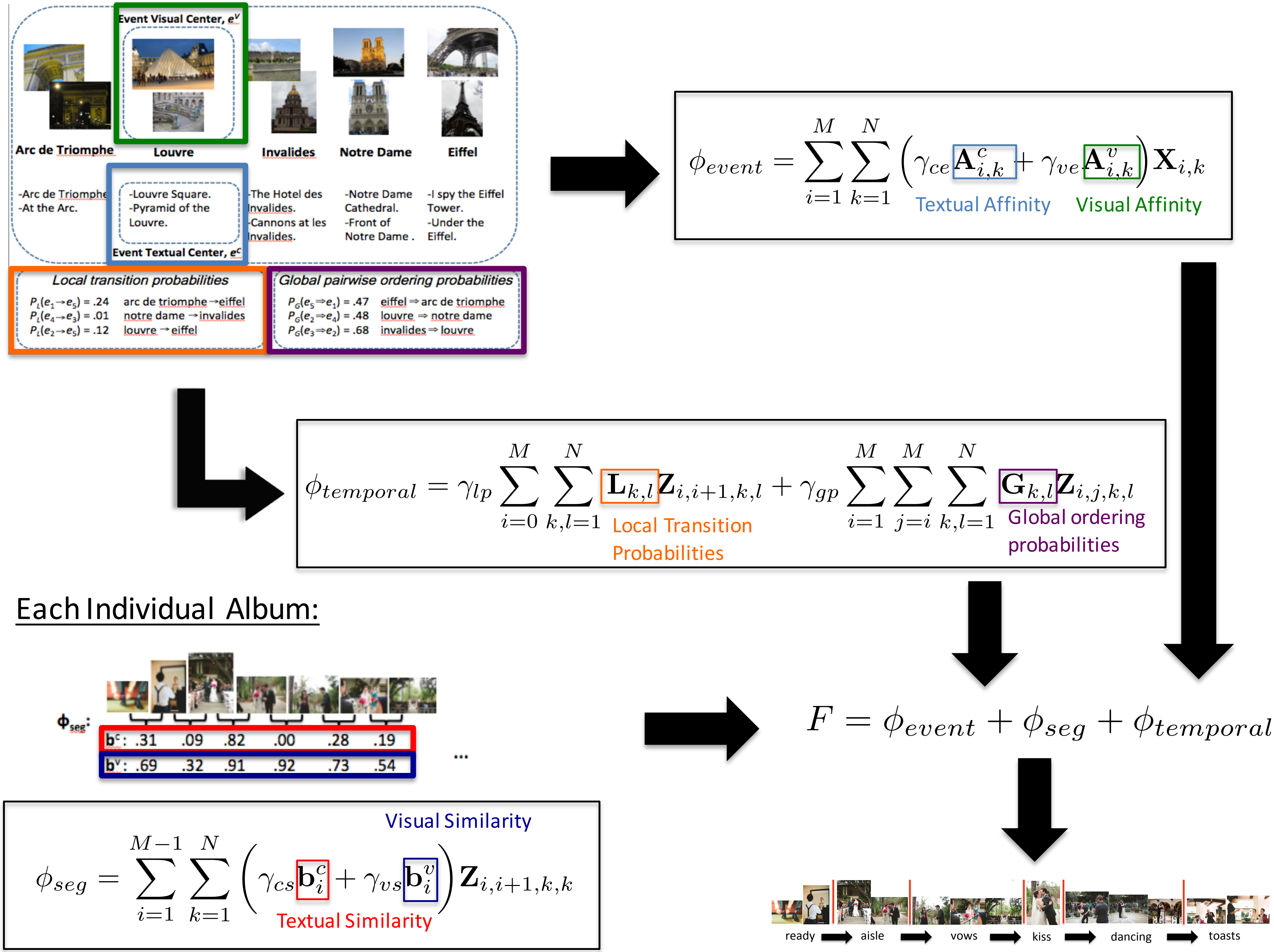
Individual Photo Album Analysis

- Input: An album of photos
- Output: An album partitioned by the scenario's compositional events

Inference

- Constrained Optimization to decode assignment and ordering of events

$$F = \phi_{event} + \phi_{seg} + \phi_{temporal}$$



Event Visual Center, e^V
Arc de Triomphe
Louvre
Invalides
Notre Dame
Eiffel
-Arc de Triomphe
-At the Arc.
-Louvre Square.
-Pyramid of the Louvre.
-The Hotel des Invalides.
-Cannons at les Invalides.
-Notre Dame Cathedral.
-Front of Notre Dame .
-I spy the Eiffel Tower.
-Under the Eiffel.
Event Textual Center, e^C
Local transition probabilities
$P_L(e_1 \rightarrow e_5) = .24$ arc de triomphe → eiffel
$P_L(e_4 \rightarrow e_3) = .01$ notre dame → invalides
$P_L(e_2 \rightarrow e_5) = .12$ louvre → eiffel
Global pairwise ordering probabilities
$P_G(e_5 \Rightarrow e_1) = .47$ eiffel ⇒ arc de triomphe
$P_G(e_2 \Rightarrow e_4) = .48$ louvre ⇒ notre dame
$P_G(e_3 \Rightarrow e_2) = .68$ invalides ⇒ louvre
$\phi_{event} = \sum_{i=1}^{M}\sum_{k=1}^{N}\left(\gamma_{ce}\boxed{\mathbf{A}^c_{i,k}} + \gamma_{ve}\boxed{\mathbf{A}^v_{i,k}}\right)\mathbf{X}_{i,k}$
Textual Affinity
Visual Affinity
$\phi_{temporal} = \gamma_{lp}\sum_{i=0}^{M}\sum_{k,l=1}^{N}\boxed{\mathbf{L}_{k,l}}\mathbf{Z}_{i,i+1,k,l} + \gamma_{gp}\sum_{i=1}^{M}\sum_{j=i}^{M}\sum_{k,l=1}^{N}\boxed{\mathbf{G}_{k,l}}\mathbf{Z}_{i,j,k,l}$
Local Transition Probabilities
Global ordering probabilities
Each Individual Album:
Φ_{seg}:
b^c: .31 .09 .82 .00 .28 .19
b^v: .69 .32 .91 .92 .73 .54
...
$F = \phi_{event} + \phi_{seg} + \phi_{temporal}$
Visual Similarity
$\phi_{seg} = \sum_{i=1}^{M-1}\sum_{k=1}^{N}\left(\gamma_{cs}\boxed{\mathbf{b}^c_i} + \gamma_{vs}\boxed{\mathbf{b}^v_i}\right)\mathbf{Z}_{i,i+1,k,k}$
Textual Similarity
ready
aisle
vows
kiss
dancing
toasts

Experiments

- Temporal Ordering
- Album Segmentation
- Learned Knowledge
  - Summarization
  - Captioning

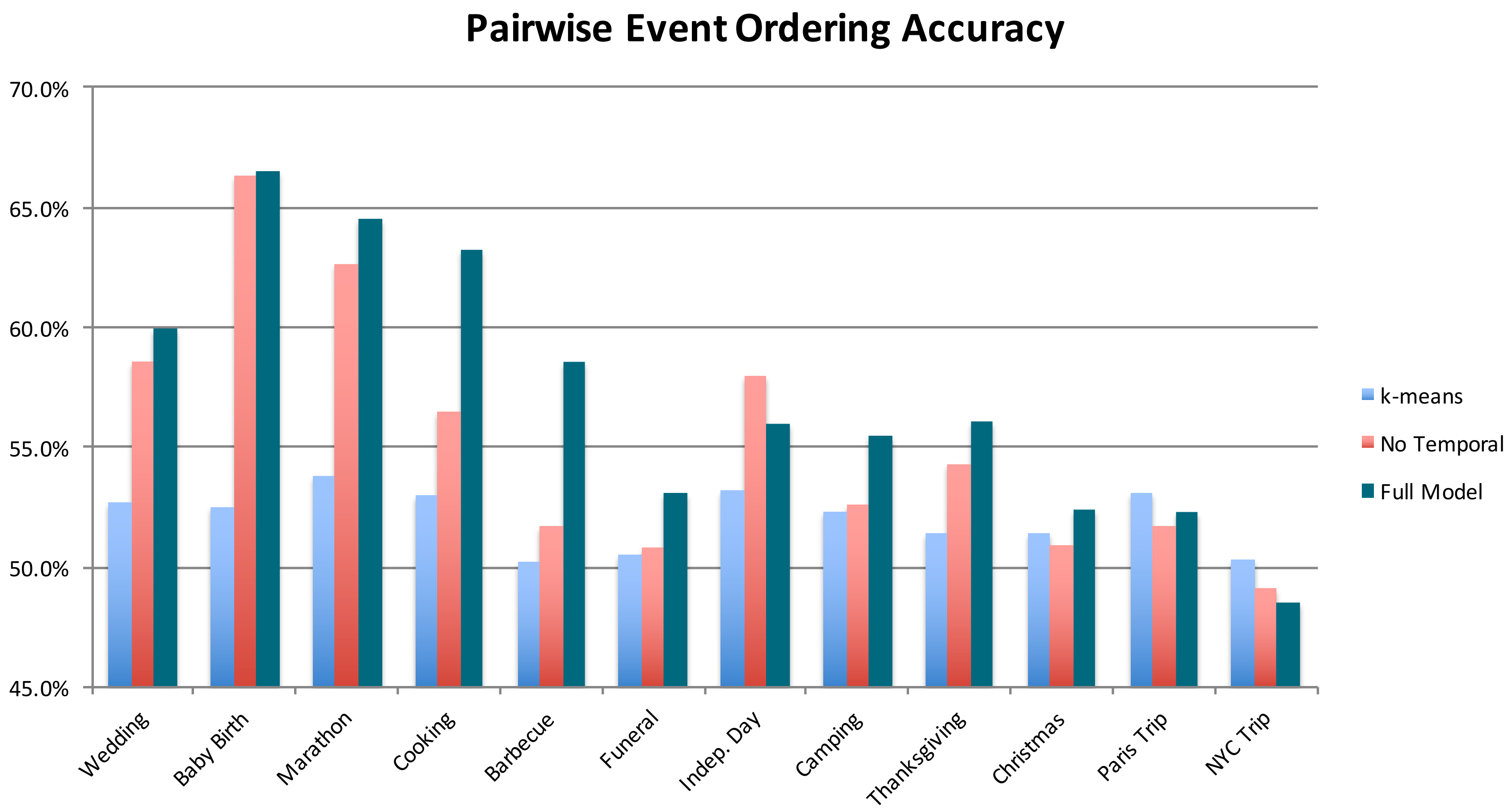
Temporal Ordering

- Compile pairwise event training set ordering statistics between all events
- In every album of the test set, pick two photos
- Based on the events assigned to those photos, predict which photo was taken before the other

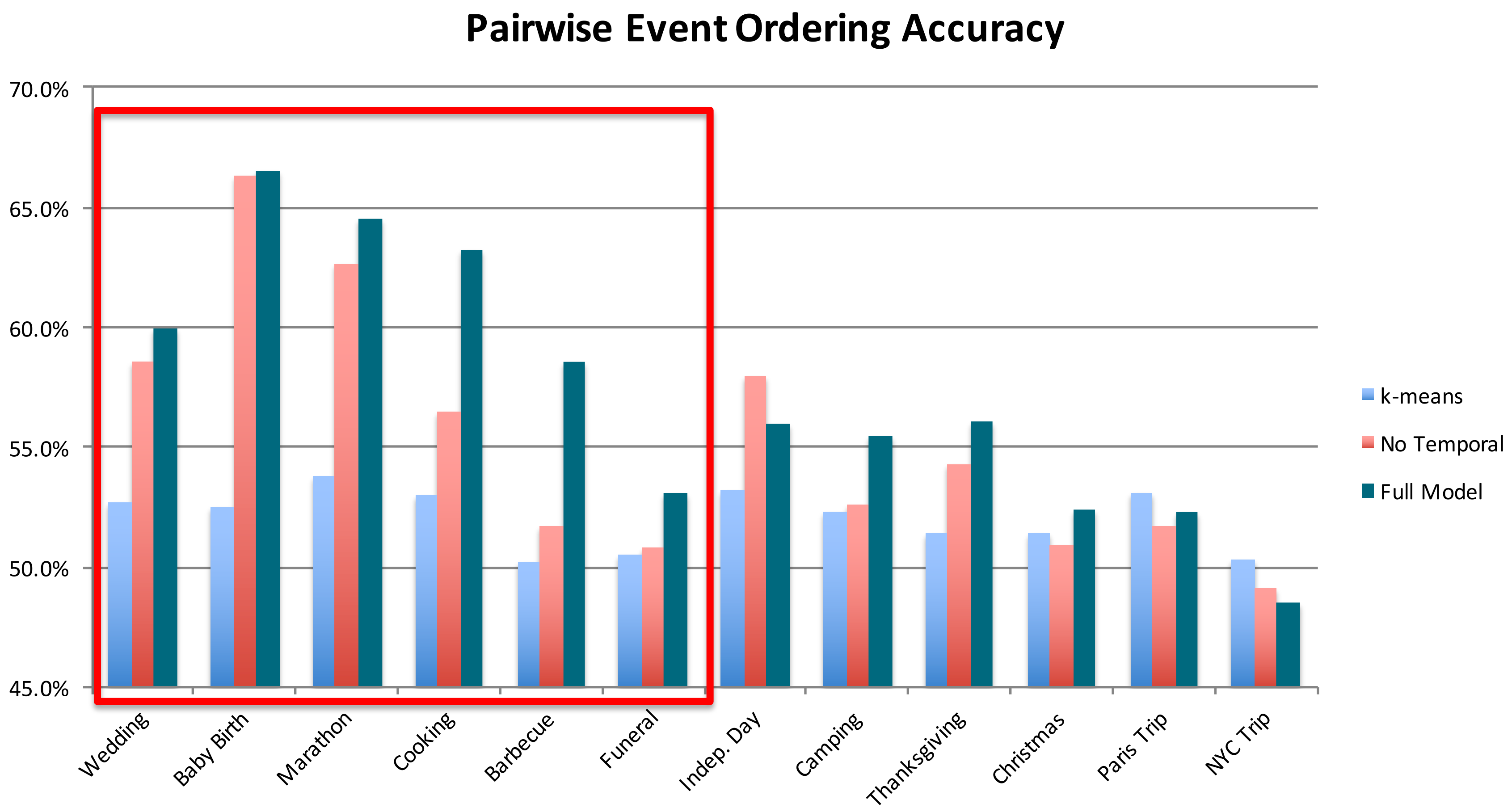
Aisle

Dance

Temporal Ordering



Temporal Ordering

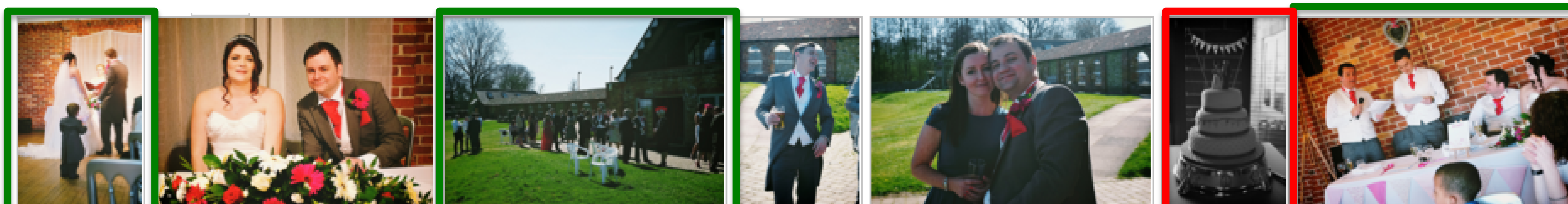


Learned Knowledge: Summarization

- Pick a set of *b* photos from an album as a summary

- Choose photos from *b* different events
- Choose photo with highest affinity for event
- Replace caption with a prototypical caption

Wedding Summaries

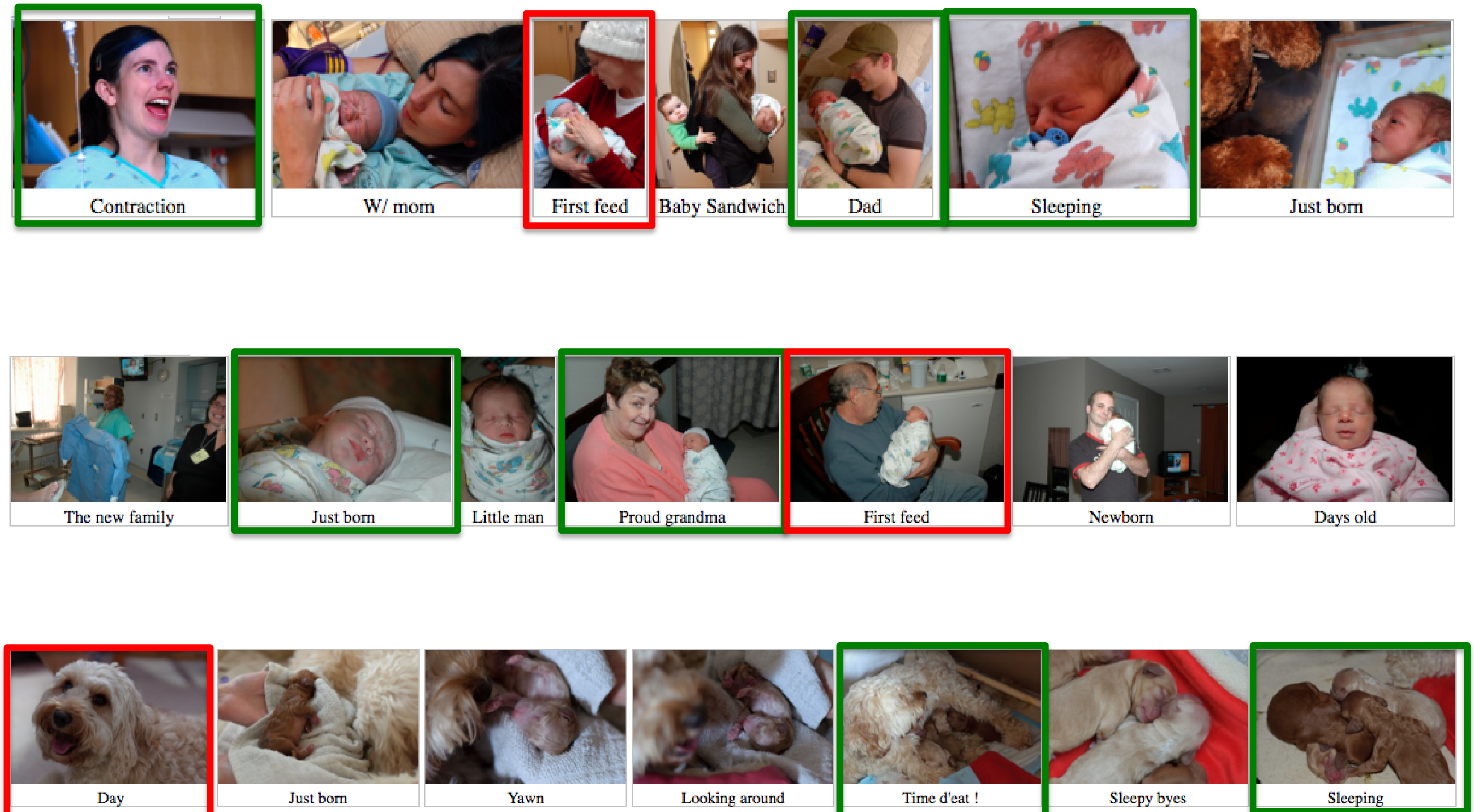


Watching the ceremony | Getting married | Greeting guests | Laughing | Mother of the bride | Eating cake | Listening to speeches

Watching the ceremony | Bridal party | Listening to the toasts | Father daughter dance | Cutting the cake | The Library | Last dance

The Arch | Watching the ceremony | ... heading for the reception | The kiss | Reciting vow 's to each other | First dance | Listening to the toasts

Baby Birth Summaries



To Conclude

- Multimodal script learning from photo albums
- Prototypical event structure of 12 common scenarios
- Future work: integration of videos, and scaling up the knowledge

In this talk

- Toward Intelligent Communication

- Learning knowledge about the world
  - Physical Knowledge (size)
  - Visual Entailment
  - Recipe Parsing with Cooking Knowledge
  - Prototypical Event Knowledge

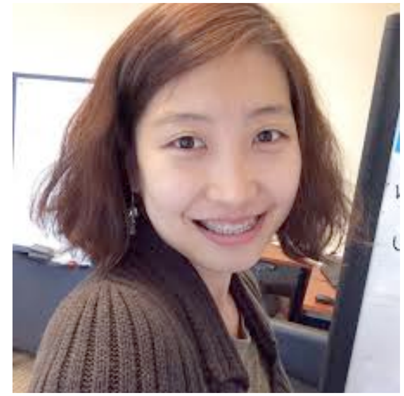
- From naturally existing data
  - No manually curated data for training

Acknowledgements

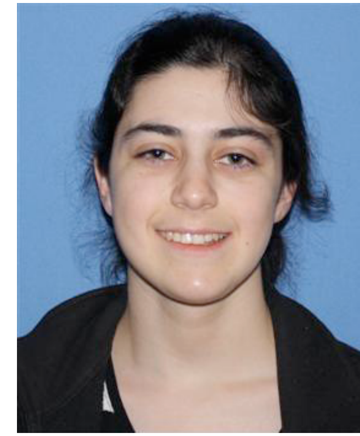
My PhD	Chloe Kiddon, Antoine Bosselut, Song Feng, Polina Kuznetsova,
Other PhD	Jianfu Chen, Vicente Ordonez, Karl Stratos, Siming Li, Jesse Dodge, Hamid Izadinia, Fereshteh Sadeghi
MS	Girish Kulkarni, Sagnik Dhar, Visruth Premraj
Professor	Ali Farhadi, Hannaneh, Hajishirzi, Hal Daumé III, Jia Deng, Alex Berg, Tamara Berg, David Warren, Luke Zettlemoyer
Industry	Margaret Mitchell, Santosh Divvala, Sujith Ravi, Ravi Kumar, Amit Goyal

Thanks!

Thanks!



Eunsol Choi



Hannah Rashkin

Maarten Sap

Max Forbes

Chloe Kiddon

Li Zilles

Antoine Bosselut