# A Multi-World Approach to Question Answering about Real-World Scenes

Mateusz Malinowski, Mario Fritz

# Outline

1. **Goal**
2. **Dataset**
3. **Performance Measure**

4. Technical Approach

# Motivation

- **“full scene understanding”**
  - semantic segmentation
  - image captioning

- Q & A is the most complete

# Full Scene Understanding?

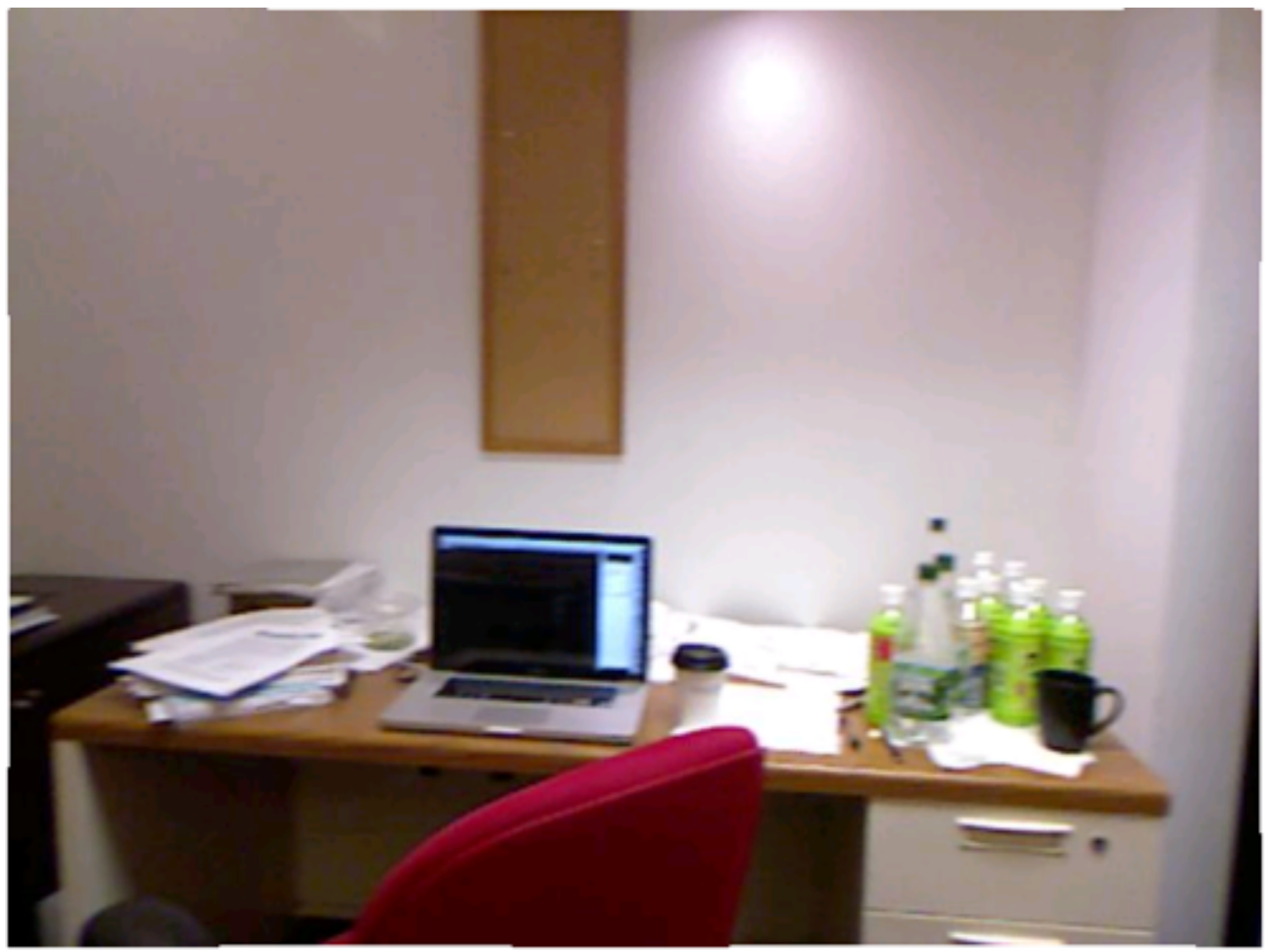
**Semantic Segmentation**


**Image Captioning**

*a car parked outside of a grassy field*

# Goal

To answer **natural-language** queries about images



**Question:** what is on the desk and behind the black cup?
**Answer:** bottle

# Dataset

# Dataset: Images

- 1449 RGB-D Images and pixel labels from NYU-Depth v2
- 894 object categories (!)
  - restricted to 37 for most evaluations

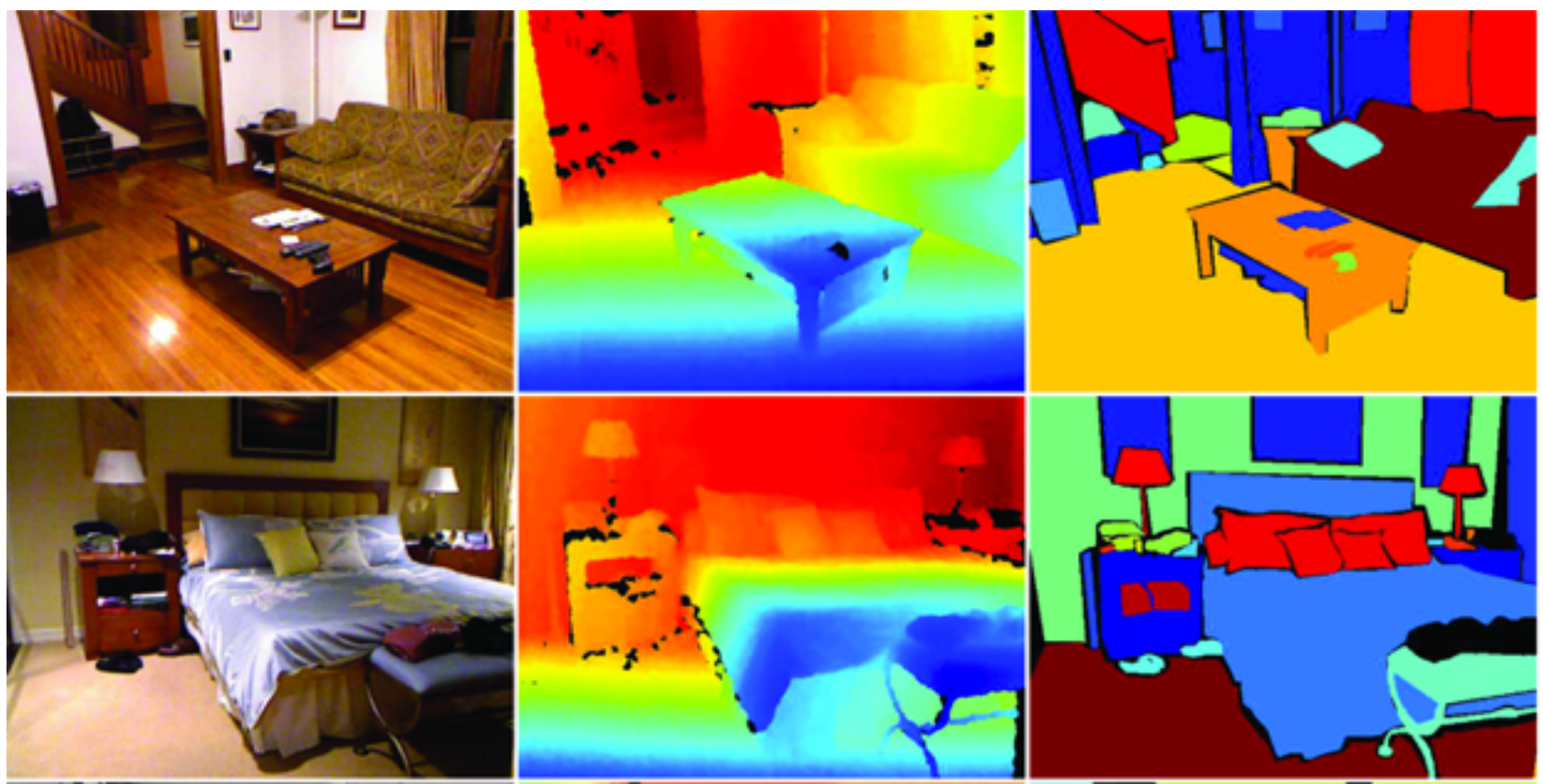


*Figure from Silberman et al, 2012*

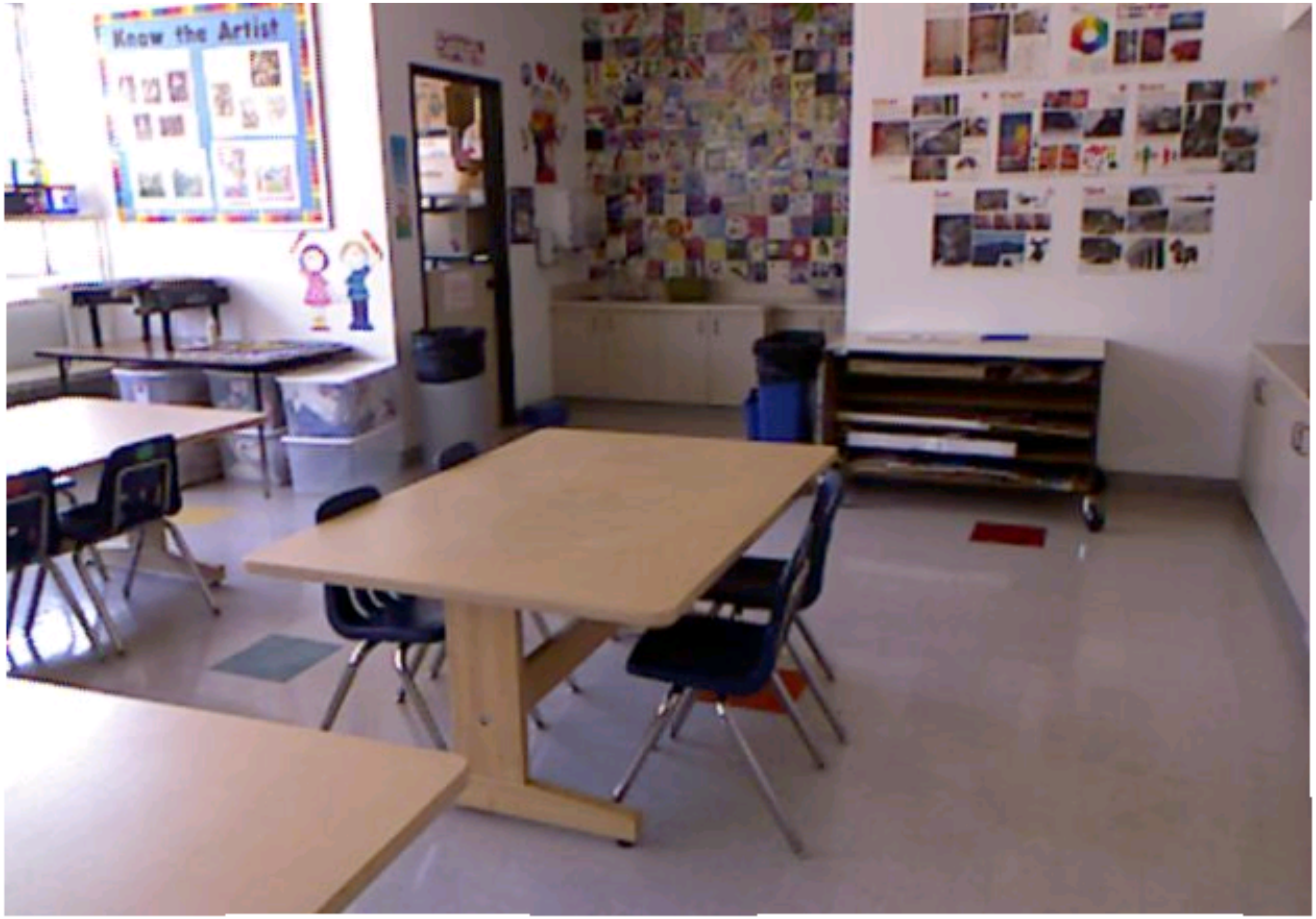
# Dataset: Q & A

**Human Dataset:**

- 12,000 Q&A pairs (~9 per image)
- questions unconstrained
- Each answer must be one of
  - a color
  - a number
  - a set of object categories - e.g. *{bed, couch}*

**Synthetic Dataset:**

- 420 Q&A pairs
- generated from templates
- answers can also be
  - scene types - e.g. *bedroom*
  - sets of images




**Question:** how many plastic toy containers are below the table in front of the wall?

**Answer:** 6

**Question:** what is on the desk?

**Answer:** {desk_mat, paper, book, napkin_dispenser}

**Question:** what color are the paper trays in the bookshelf on the left side of the wall divider not on the desk in front of the computer chair?

**Answer:** black

# Performance Measure

# Performance Measure

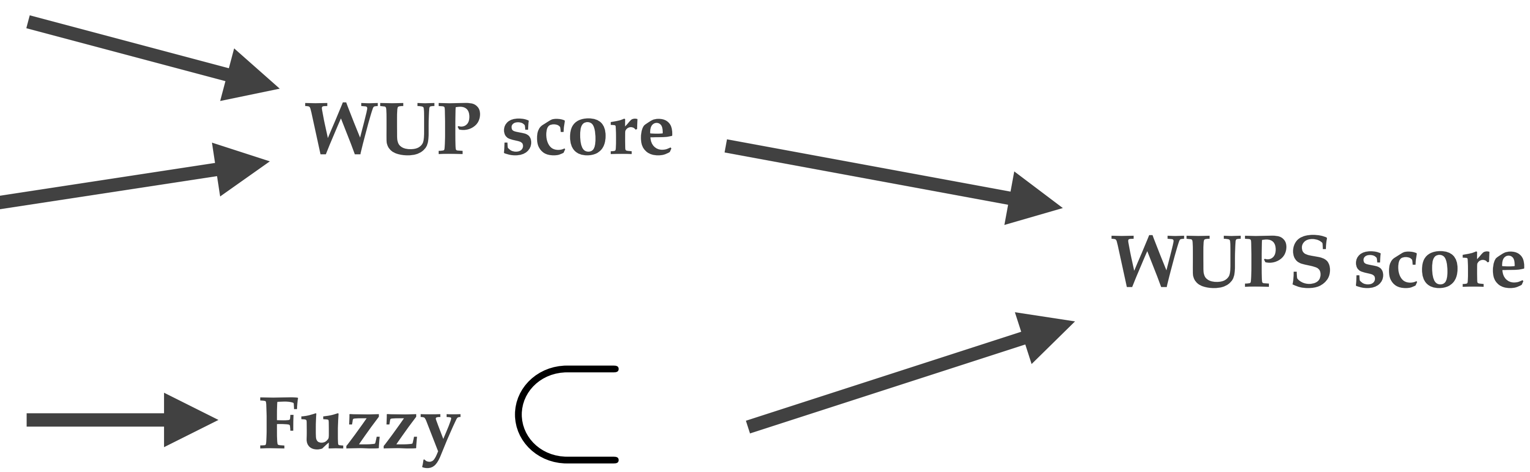
**Difficulties:**

- near-synonyms, e.g. *couch* vs *futon*
- hypernyms, e.g. *person* vs *woman* vs *skateboarder*
- comparing sets, e.g. *{pillow, book}* vs *{pillow}*

# Performance Measure

**Difficulties:**

- near-synonyms
- hypernyms
- comparing sets



# What's wrong with WUP?

$$WUP(a,b) = \frac{2 * depth(lca(a,b))}{depth(a) + depth(b)} \in [0,1]$$

$$WUP(lamp, table) = \frac{2 * depth(furniture)}{depth(lamp) + depth(table)} = 0.88$$

$$WUP(couch, futon) = 0.52$$

# What about distributed representations?

- generalize to multi-word answers like “red jacket” and “female tennis player”
- usually trained on huge text corpora, with no visual information

# Asymmetry

**Question:** Who is holding the racquet?
**GT Answer:** female tennis player
**Answer:** person

$d(person, female\ tennis\ player) \sim 0$

**Question:** Who is speaking?
**GT Answer:** person
**Answer:** female tennis player

$d(female\ tennis\ player, person) \sim \infty$

# Performance Measure

**Needs to:**

- include visual similarity
- be asymmetric

# Technical Approach

# Two sources of uncertainty

1. Vision: what is in the image? → semantic segmentation

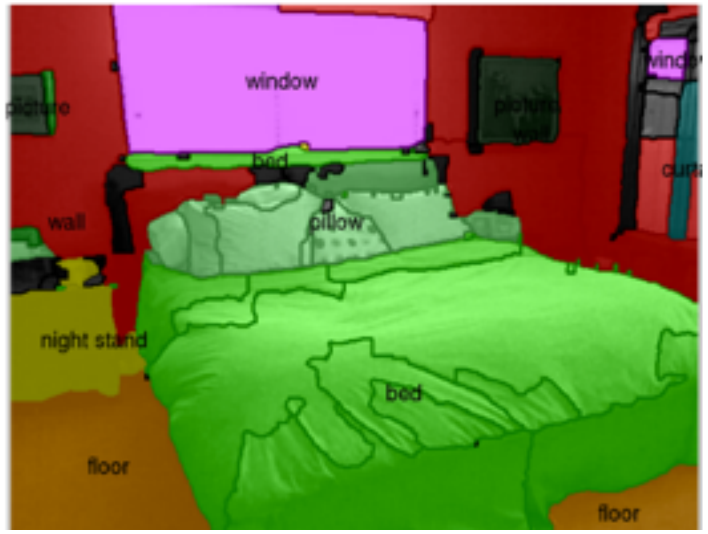



*Figure from "Perceptual Organization and Recognition of Indoor Scenes from RGB-D images", Gupta et al, 2013*

2. Language: what does the question ask? → semantic parse

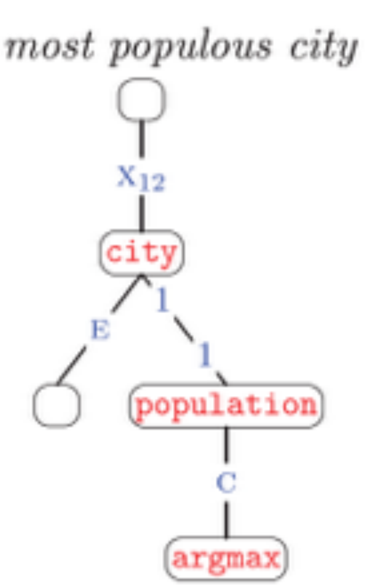
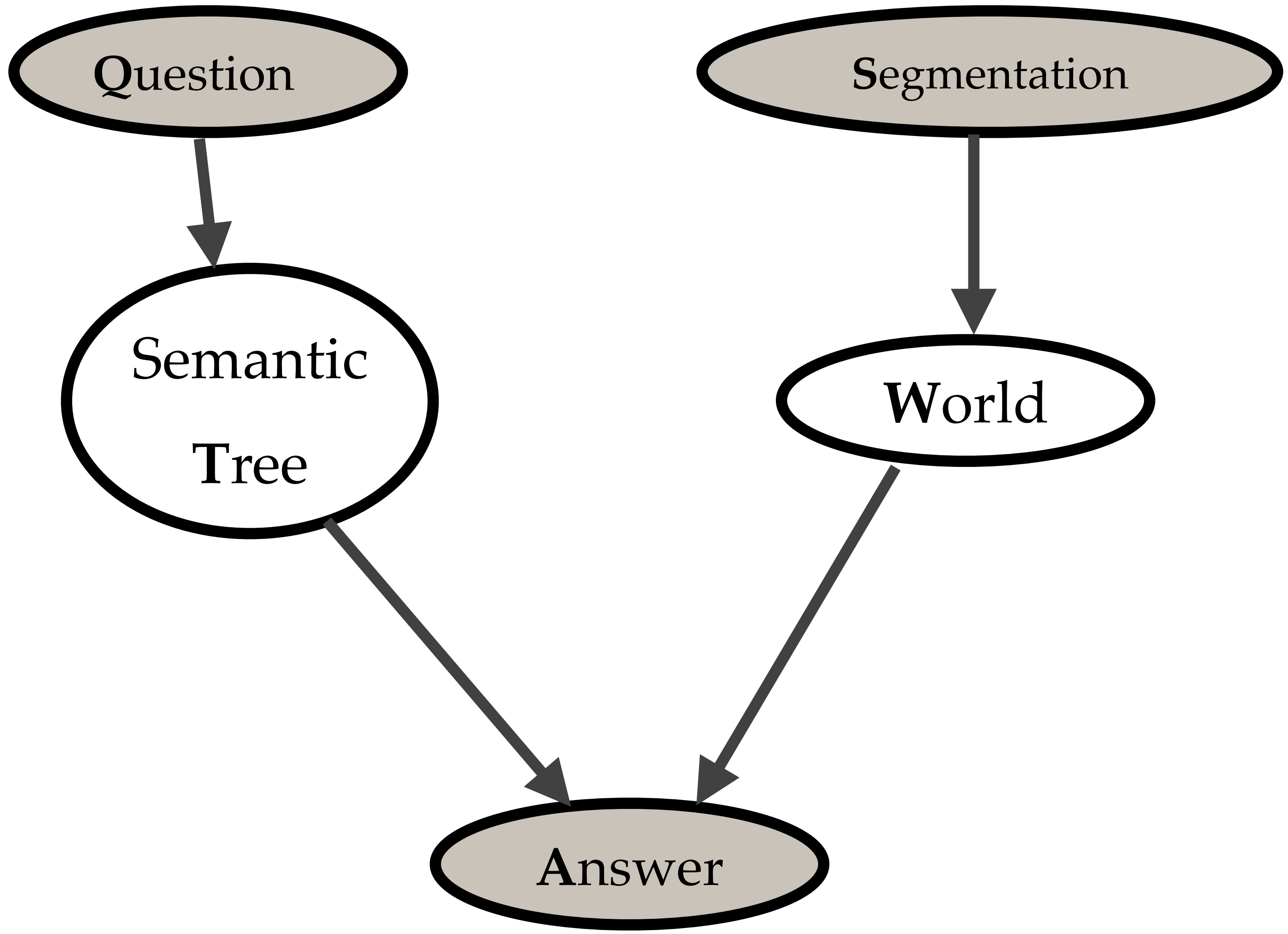



*Figure from "Learning Dependency-Based Compositional Semantics", Liang et al, 2013*

# Graphical Model




$$P(A|Q,S) = \sum_W \sum_T P(A|T,W)P(T|Q)P(W|S)$$

# Representation

**What is a world?**

- Facts, e.g $chair(segment, color, X_{\{min,mean,max\}}, Y_{\{min,mean,max\}}, Z_{\{min,mean,max\}})$

- Relations, e.g. $above(A, B)$ , $inFront(A, B)$

| | Predicate | Definition |
|---|---|---|
| auxiliary relations | $closeAbove(A, B)$ | $above(A, B)$ and $(Y_{min}(B) < Y_{max}(A) + \epsilon)$ |
| | $closeLeftOf(A, B)$ | $leftOf(A, B)$ and $(X_{min}(B) < X_{max}(A) + \epsilon)$ |
| | $closeInFrontOf(A, B)$ | $inFrontOf(A, B)$ and $(Z_{min}(B) < Z_{max}(A) + \epsilon)$ |
| | $X_{aux}(A, B)$ | $X_{mean}(A) < X_{max}(B)$ and $X_{min}(B) < X_{mean}(A)$ |
| | $Z_{aux}(A, B)$ | $Z_{mean}(A) < Z_{max}(B)$ and $Z_{min}(B) < Z_{mean}(A)$ |
| | $h_{aux}(A, B)$ | $closeAbove(A, B)$ or $closeBelow(A, B)$ |
| | $v_{aux}(A, B)$ | $closeLeftOf(A, B)$ or $closeRightOf(A, B)$ |
| | $d_{aux}(A, B)$ | $closeInFrontOf(A, B)$ or $closeBehind(A, B)$ |
| spatial | $leftOf(A, B)$ | $X_{mean}(A) < X_{mean}(B))$ |
| | $above(A, B)$ | $Y_{mean}(A) < Y_{mean}(B)$ |
| | $inFrontOf(A, B)$ | $Z_{mean}(A) < Z_{mean}(B))$ |
| | $on(A, B)$ | $closeAbove(A, B)$ and $Z_{aux}(A, B)$ and $X_{aux}(A, B)$ |
| | $close(A, B)$ | $h_{aux}(A, B)$ or $v_{aux}(A, B)$ or $d_{aux}(A, B)$ |

*Figure from Malinowski and Fritz, 2014*

# Simplifying Assumptions

$$P(A|Q,S) = \sum_W \sum_T P(A|T,W)P(T|Q)P(W|S)$$

1. $P(T|Q) \propto \exp(w^T \phi(T,Q))$
2. $P(W|S) = P(s_1 = c_{f(1)}..., s_n = c_{f(n)}) = \prod_i P(s_i = c_{f(i)})$
3. Sample 25 possible worlds
4. $P(A|T,W) \sim$ 3-nearest neighbour "batch" approximation

# Results

# Results

Human question-answer pairs (HumanQA)

| Segmentation | World(s) | #classes | Accuracy | WUPS at 0.9 | WUPS at 0 |
|---|---|---|---|---|---|
| HumanSeg | Single | 894 | 7.86% | 11.86% | 38.79% |
| HumanSeg | Single | 37 | 12.47% | 16.49% | 50.28% |
| AutoSeg | Single | 37 | 9.69% | 14.73% | 48.57% |
| AutoSeg | Multi | 37 | 12.73% | 18.10% | 51.47% |
| Human Baseline | | 894 | 50.20% | 50.82% | 67.27% |
| Human Baseline | | 37 | 60.27% | 61.04% | 78.96% |

**My Baseline**

"how many" —> **2**

"what color" —> ***white***

else ——————> {***table***}

17.85%

# Error Analysis - Language

Q: How many red chairs are there?
H: ()
M: 6
C: blinds

Q: How many chairs are at the table?
H: wall
M: 4
C: chair

# Error Analysis - Vision

synthetic question-answer pairs (SynthQA)

| Segmentation | World(s) | # classes | Accuracy |
|---|---|---|---|
| HumanSeg | Single with Neg. 3 | 37 | 56.0% |
| HumanSeg | Single | 37 | 59.5% |
| AutoSeg | Single | 37 | 11.25% |
| AutoSeg | Multi | 37 | 13.75% |

# Summary

- Very interesting, high-level vision problem
- Very difficult, large dataset
- Unclear performance measure
- Authors’ approach doesn’t work