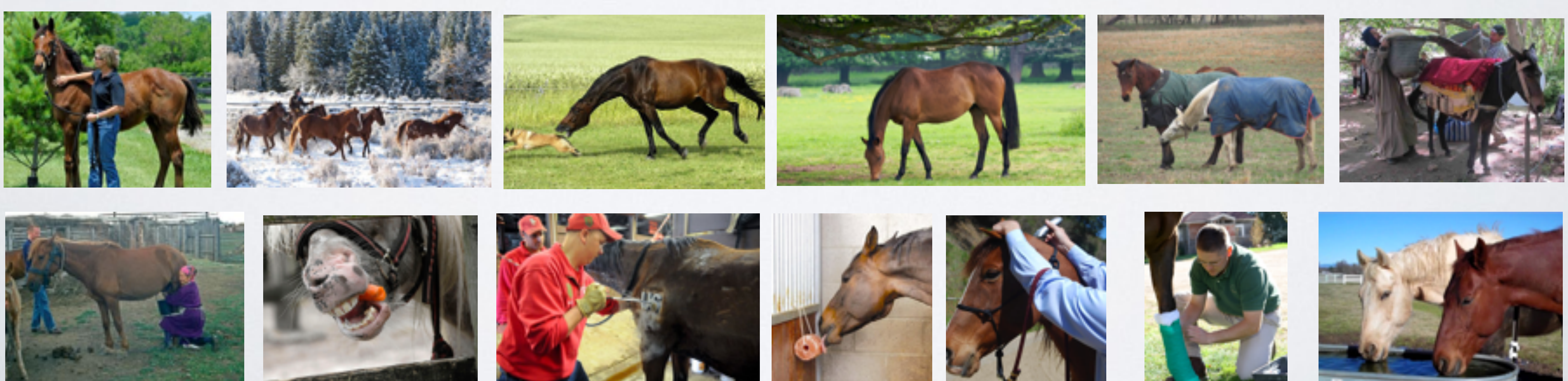
# Situation Recognition

## Visual Semantic Role Labeling for Image Understanding

Mark Yatskar

in collaboration w/ Luke Zettlemoyer, Ali Farhadi

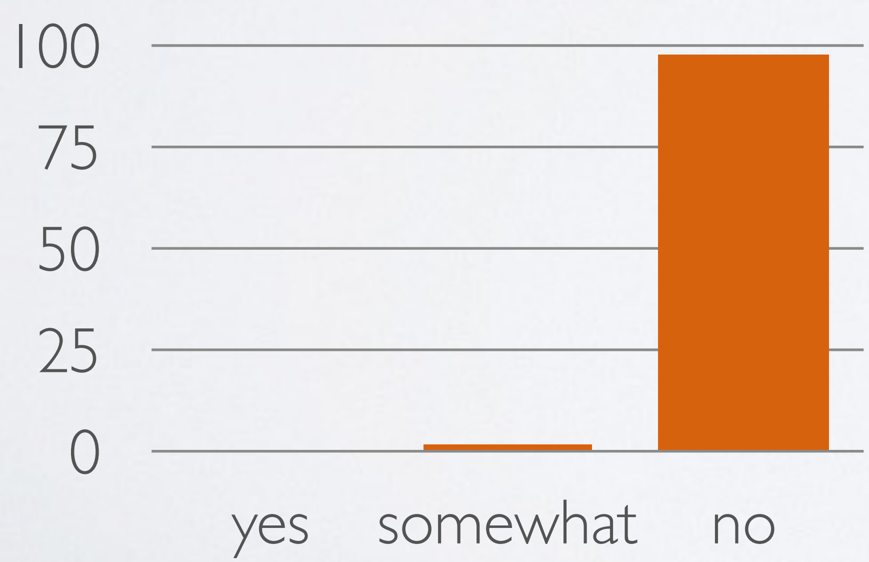


# How can we summarize what is happening in an image?

| LOADING | | | | |
|---|---|---|---|---|
| **AGENT** | **ITEM** | **DESTINATION** | **TOOL** | **PLACE** |
| WOMAN | HORSE | TRAILER | ROPE | OUTDOORS |

# Is the same thing happening in two images?

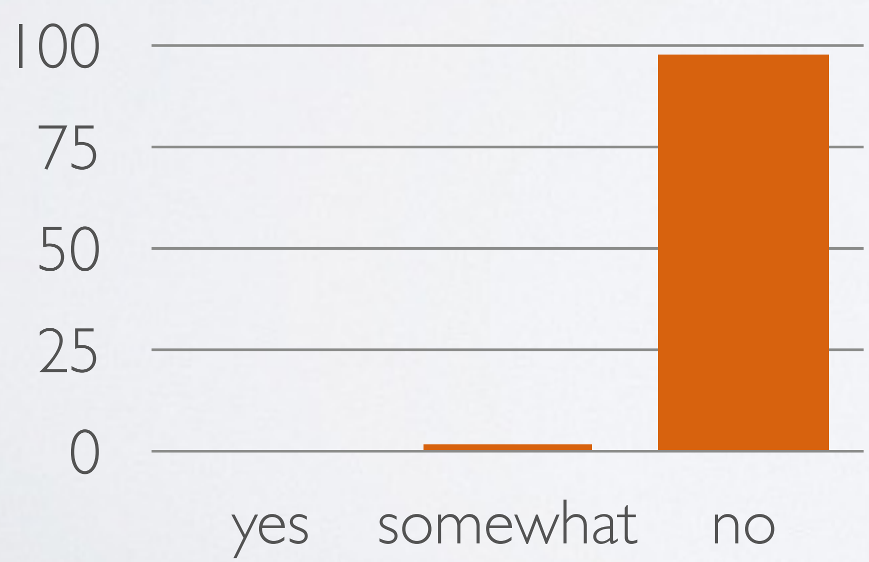
turkers say…



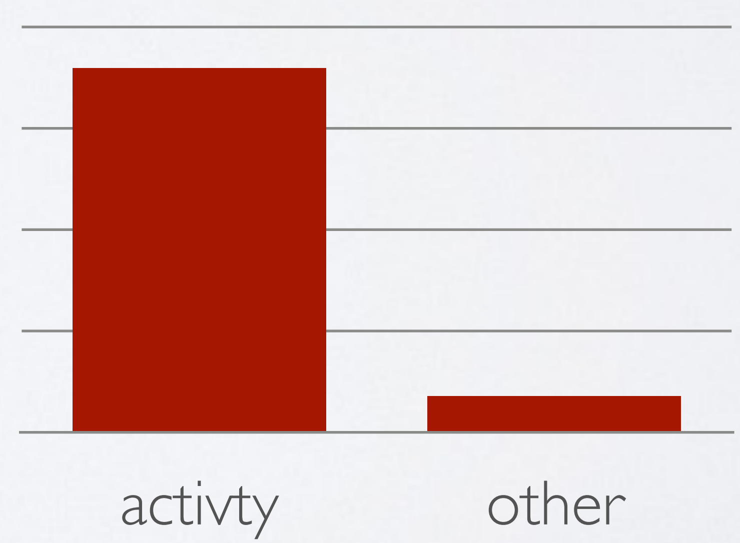
# Is the same thing happening in two images?

Activity

turkers say…




why no?

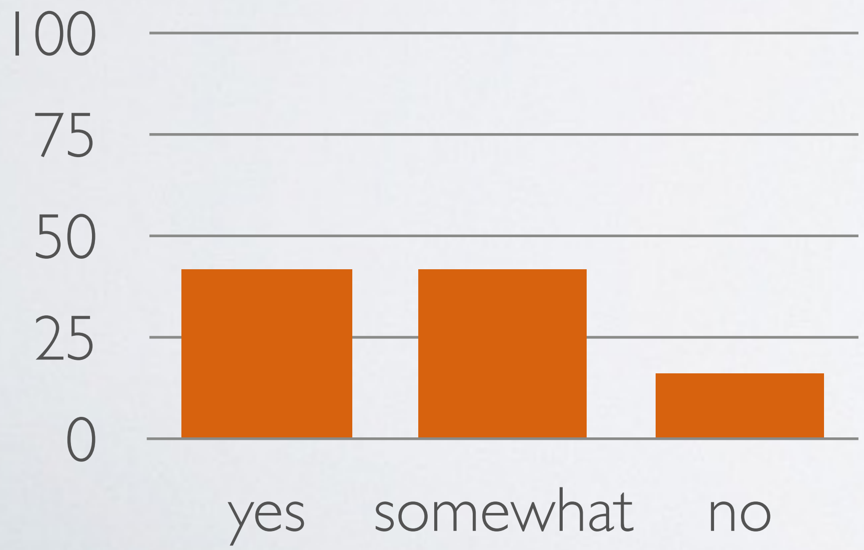


# Is the same thing happening in two images?

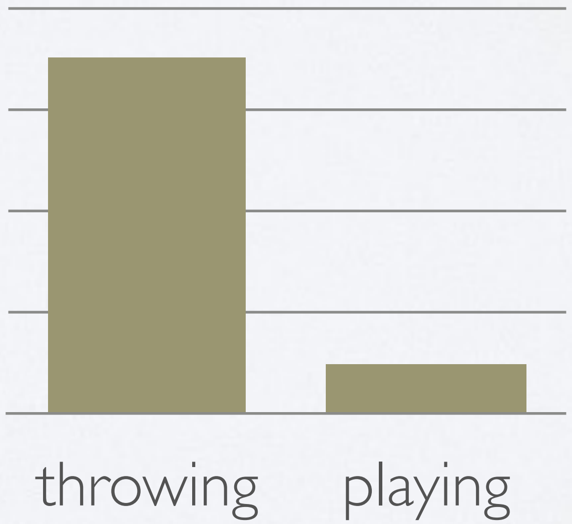


Activity

turkers say…




why yes?



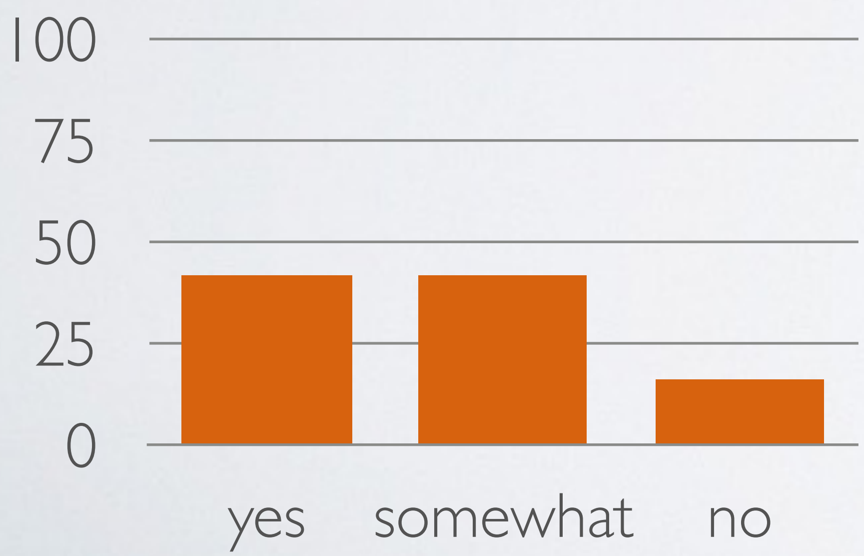
# Is the same thing happening in two images?



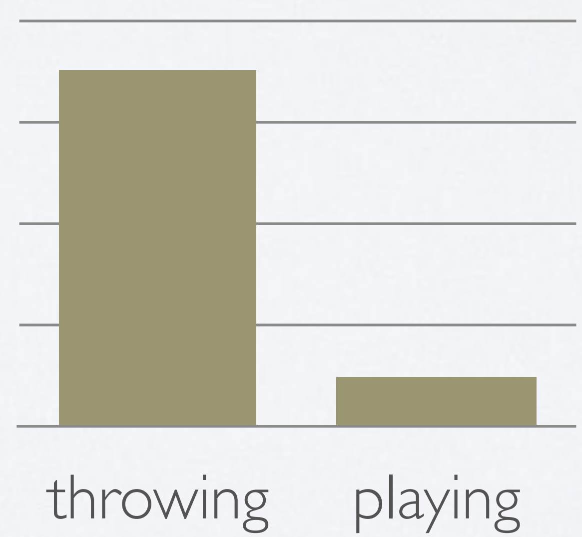
Activity

Object

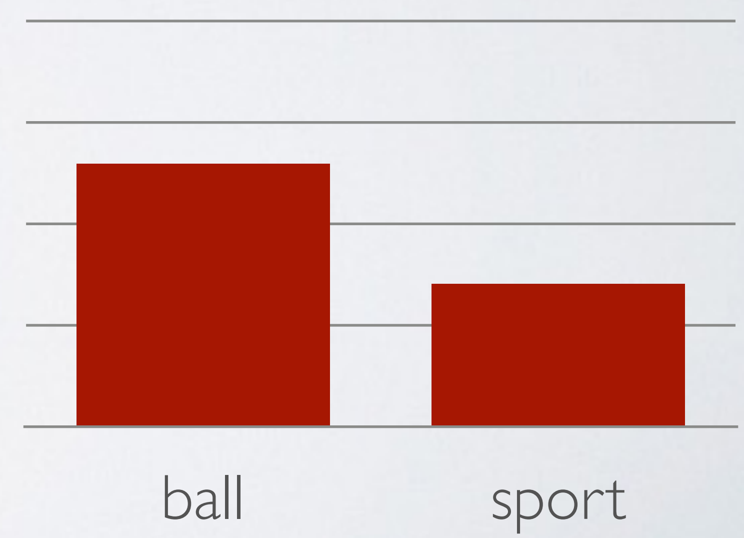
turkers say…



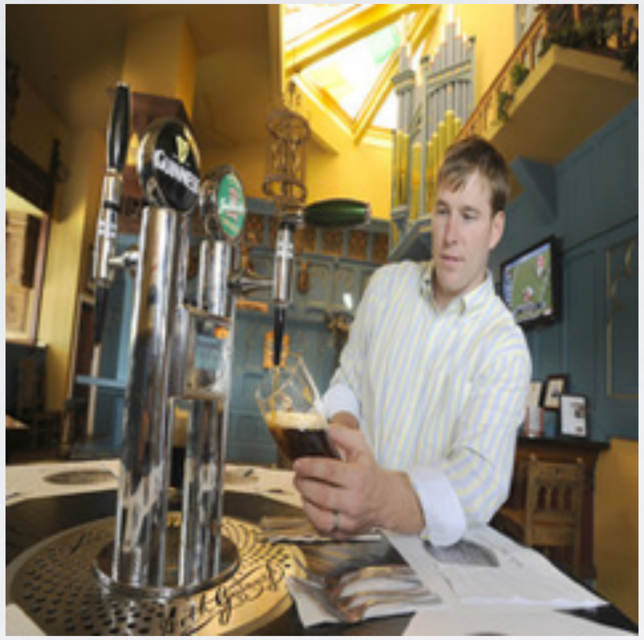

why yes?




why no?



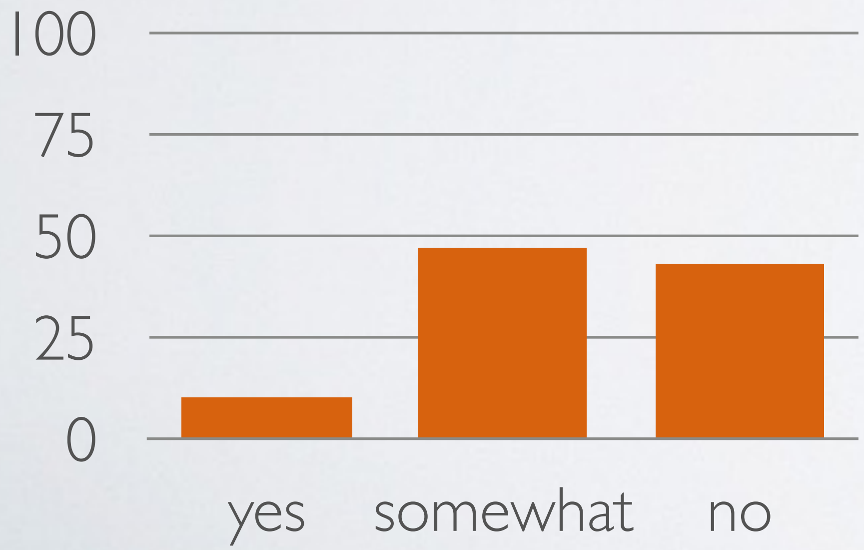
# Is the same thing happening in two images?



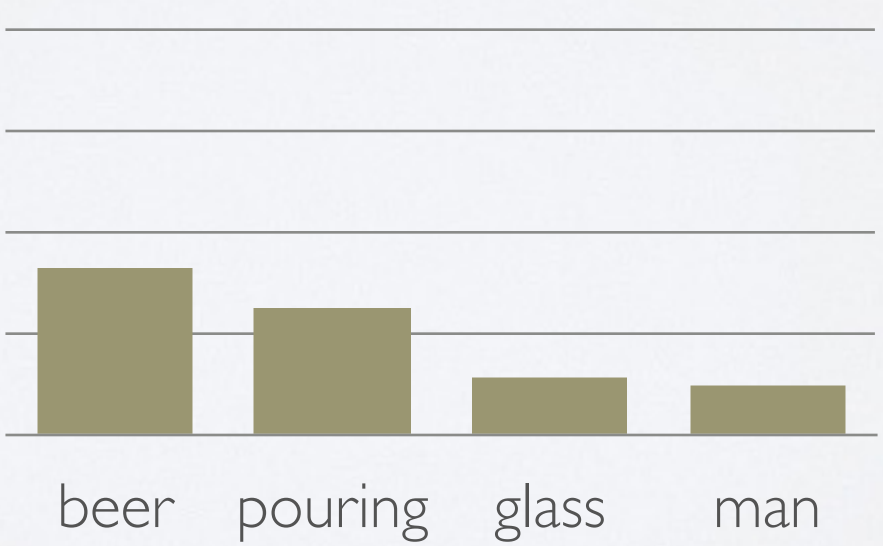
Activity

Object

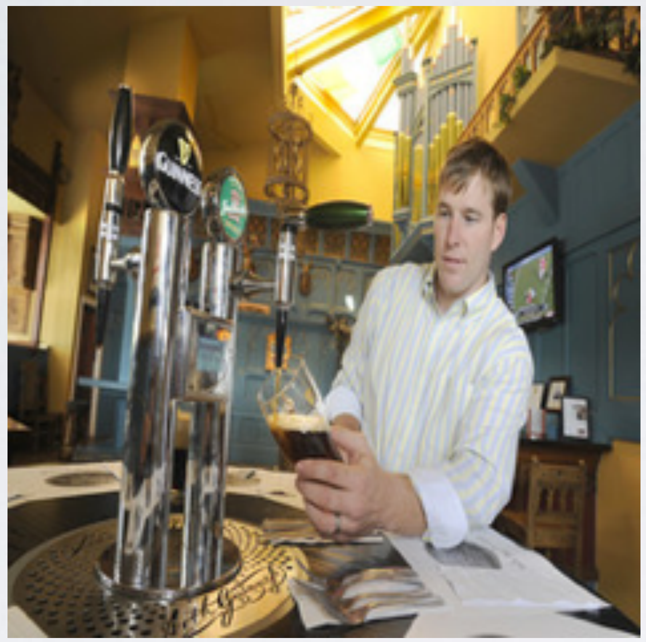
turkers say…




why yes?



# Is the same thing happening in two images?

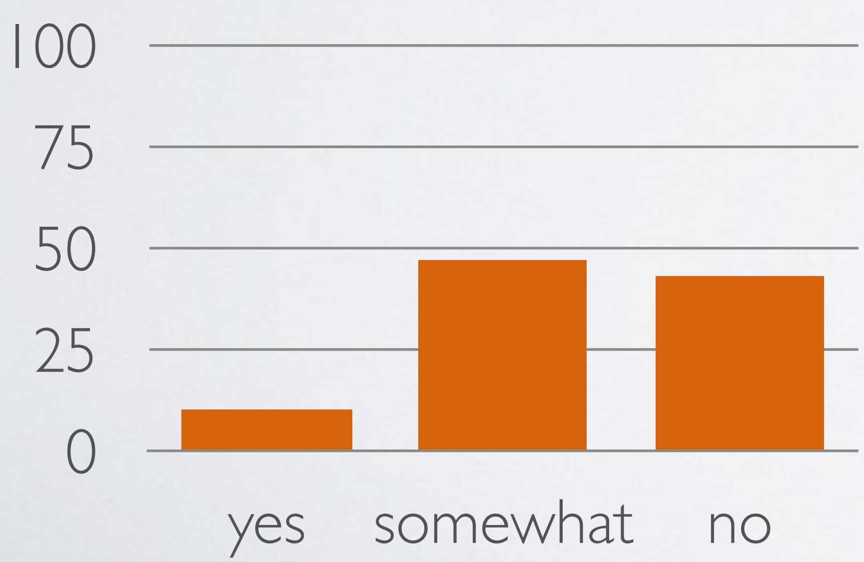


Activity

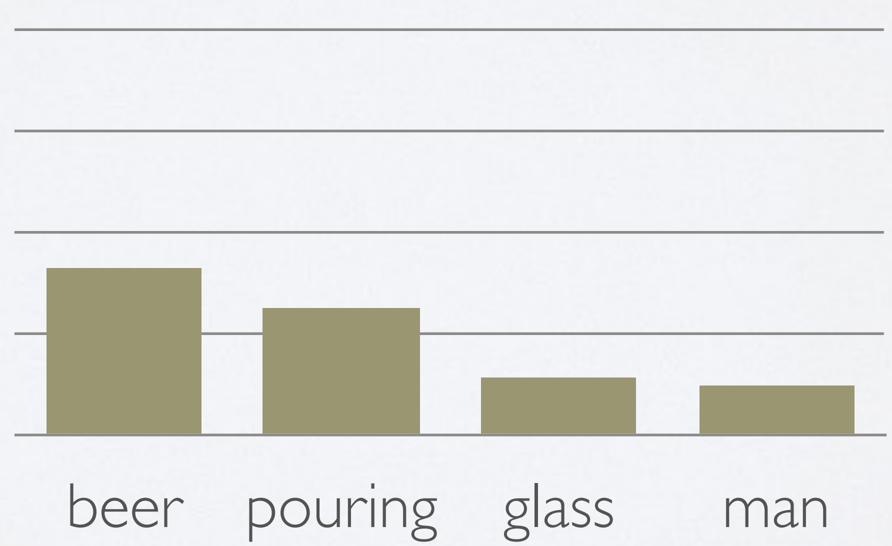
Object

**Role**

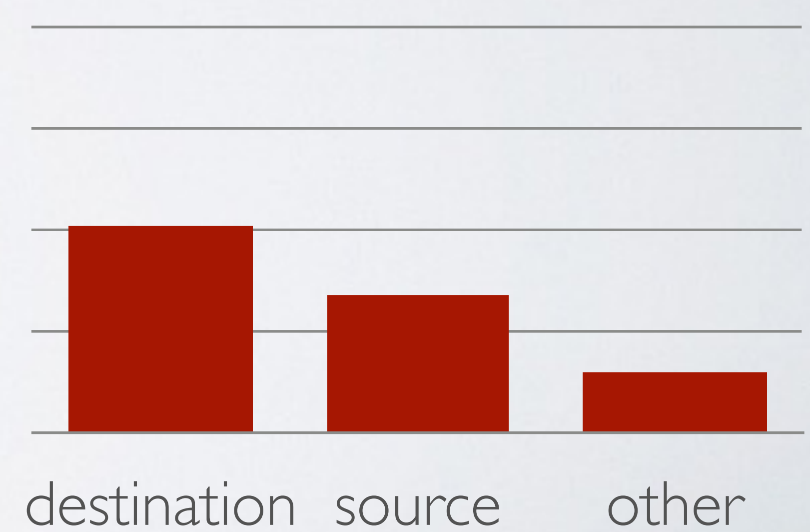
turkers say…




why yes?




why no?



# Systematically describe how objects participate in activities through **roles**

| LOADING | | | | |
|---|---|---|---|---|
| **AGENT** | **ITEM** | **DESTINATION** | **TOOL** | **PLACE** |
| WOMAN | HORSE | TRAILER | ROPE | OUTDOORS |

# Situation Recognition

| FIXING | | | | |
|---|---|---|---|---|
| **AGENT** | **OBJECT** | **PART** | **TOOL** | **PLACE** |
| BOY | CAR | TIRE | TIRE IRON | OUTDOORS |

# Situation Recognition

| POURING | |
|---|---|
| AGENT | MAN |
| SUBSTANCE | BEER |
| SOURCE | TAP |
| DESTINATION | GLASS |
| PLACE | BARROOM |

Same

Different

| POURING | |
|---|---|
| AGENT | MAN |
| SUBSTANCE | BEER |
| SOURCE | GLASS |
| DESTINATION | MOUTH |
| PLACE | BACKYARD |

# Situation Recognition

**What is the space of possible situations?**

| POURING | |
|---|---|
| AGENT | MAN |
| SUBSTANCE | BEER |
| SOURCE | TAP |
| DESTINATION | GLASS |
| PLACE | BARROOM |

Same

Different

| POURING | |
|---|---|
| AGENT | MAN |
| SUBSTANCE | BEER |
| SOURCE | GLASS |
| DESTINATION | MOUTH |
| PLACE | BACKYARD |

# imSitu

## A Large Scale Situation Dataset

120k+ images, 500+ verbs, 100k+ situations

# Natural Language Processing: Semantic Role Labeling

Activity

Object

**Role**

A boy is fixing a car tire with a tire iron outdoors.

# Natural Language Processing: Semantic Role Labeling

| |
|---|
| Activity |
| Object |
| Role |

A boy is **fixing** a car tire with a tire iron outdoors.

# Natural Language Processing: Semantic Role Labeling

Activity

Object

Role

A boy is **fixing** a car tire with a tire iron outdoors.

# Natural Language Processing: Semantic Role Labeling

Activity

Object

Role

A boy is **fixing** a car tire with a tire iron outdoors.

| FIXING | | | | |
|---|---|---|---|---|
| **AGENT** | **OBJECT** | **PART** | **TOOL** | **PLACE** |
| BOY | CAR | TIRE | TIRE IRON | OUTDOORS |

# Natural Language Processing: Semantic Role Labeling

Activity

Object

Role

A jockey **falling** from a horse onto the ground at a racetrack.

| FALLING | | | |
|---|---|---|---|
| **AGENT** | **SOURCE** | **DESTINATION** | **PLACE** |
| JOCKEY | HORSE | GROUND | RACETRACK |

# FrameNet for Verb and Role Inventory

semantic role labeling ontology:
FrameNet (8000 verbs)

| FIXING | | | | |
|---|---|---|---|---|
| AGENT | OBJECT | PART | TOOL | PLACE |
| | | | | |

# Visualness

## filter verbs, semantic roles

~1000 visual verbs
~3.5 roles/verb

semantic role labeling ontology:
FrameNet (8000 verbs)

| FIXING | | | | |
|---|---|---|---|---|
| AGENT | OBJECT | PART | TOOL | PLACE |

# WordNet for Noun Inventory

semantic role labeling ontology: FrameNet (8000 verbs)

| FIXING | | | | |
|---|---|---|---|---|
| AGENT | OBJECT | PART | TOOL | PLACE |
| | | | | |

values from noun ontology: WordNet (80, 000 nouns)

## creating imSitu

FrameNet
Visualness
WordNet

Filter Images

Web N-grams

Google Images Search

semantic role labeling ontology: FrameNet (8000 verbs)

| FIXING | | | | |
|---|---|---|---|---|
| AGENT | OBJECT | PART | TOOL | PLACE |
| | | | | |

values from noun ontology: WordNet (80, 000 nouns)

creating imSitu

FrameNet

Visualness

WordNet

Filter Images

# Fill Values

semantic role labeling ontology: FrameNet (8000 verbs)

| FIXING | | | | |
|---|---|---|---|---|
| AGENT | OBJECT | PART | TOOL | PLACE |
| BOY | CAR | TIRE | TIRE IRON | OUTDOORS |

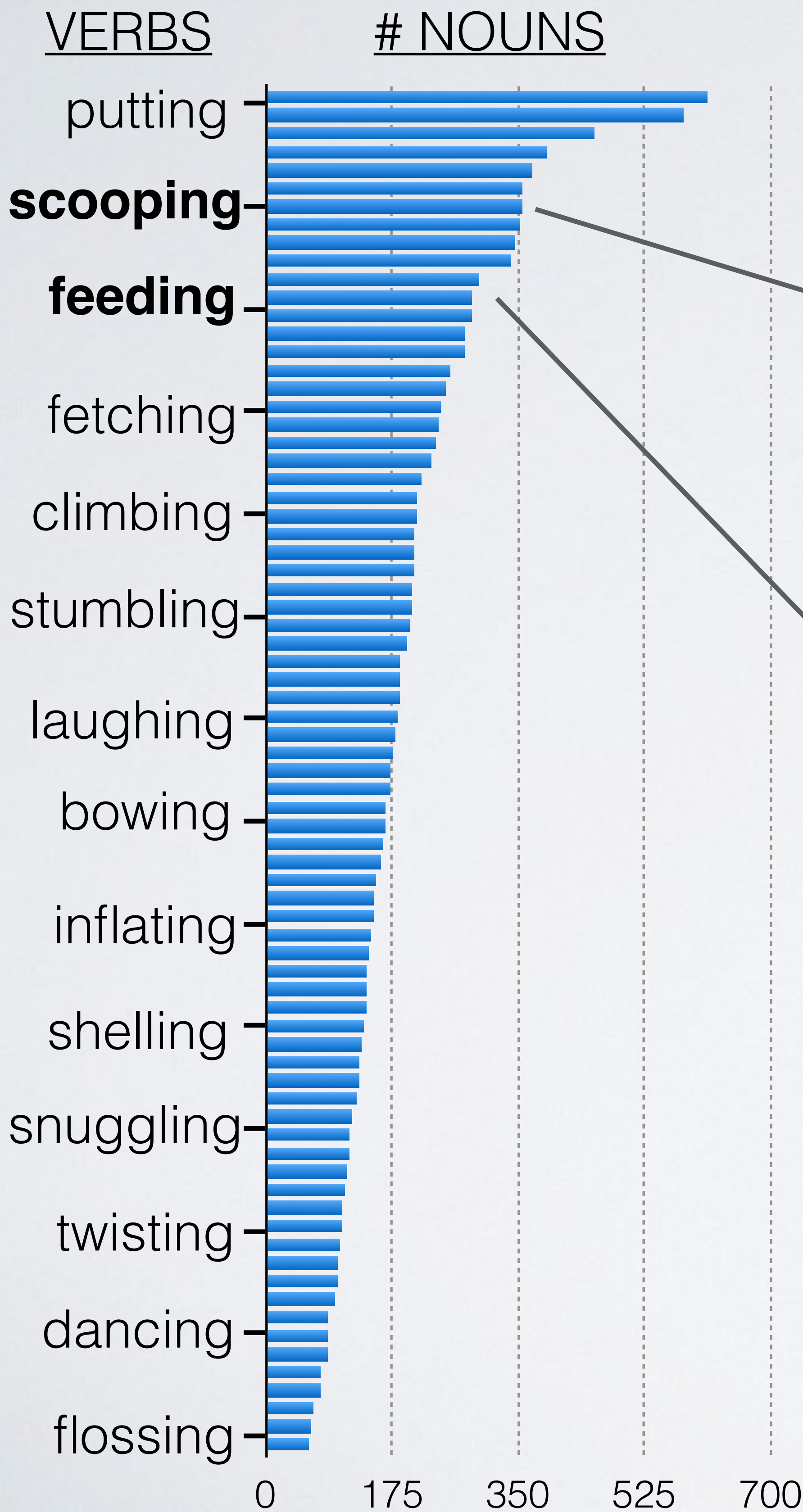
values from noun ontology: WordNet (80, 000 nouns)

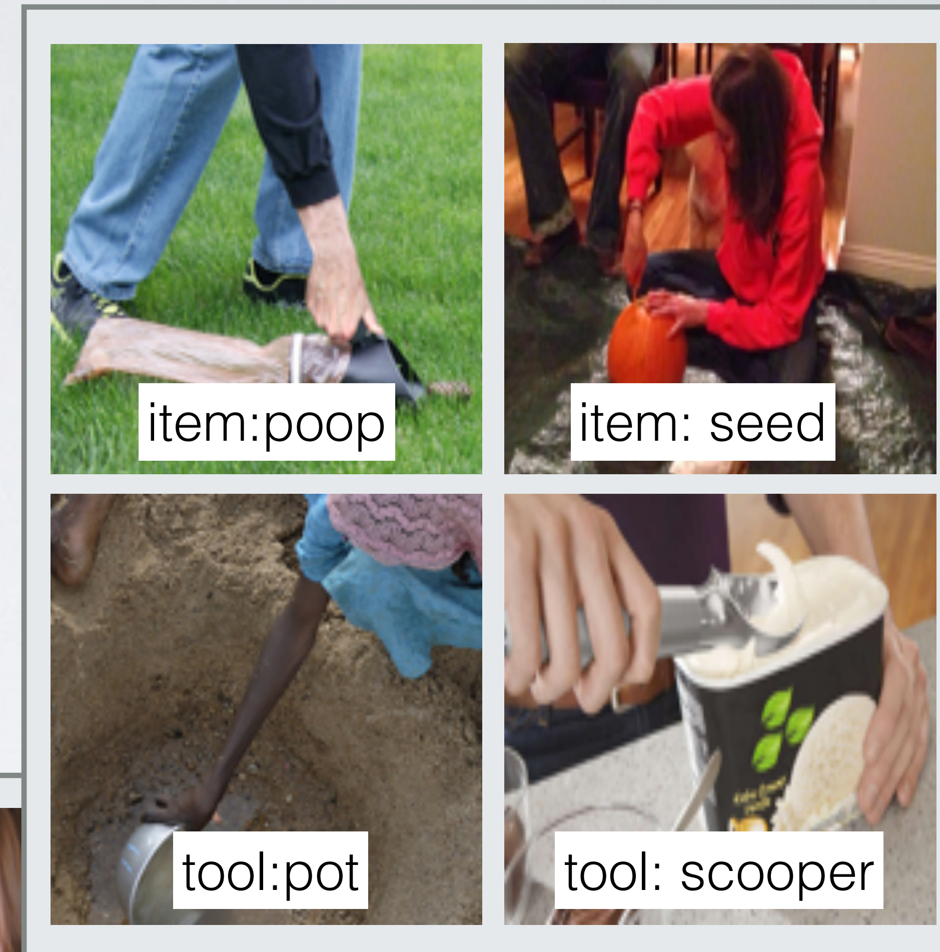
# imSitu: Dataset Statistics

| | |
|---|---|
| Verbs | 504 |
| Images | 126,102 |
| Situation/Image | 3 |
| Roles (types) | 1,788 (190) |
| Nouns ( >=3) | 11,538 (6,794) |
| Annotations | 1,481,851 |
| Images/Verb | 200-400 |
| Uniq. situations (>= 3) | 205,095 (21,505) |

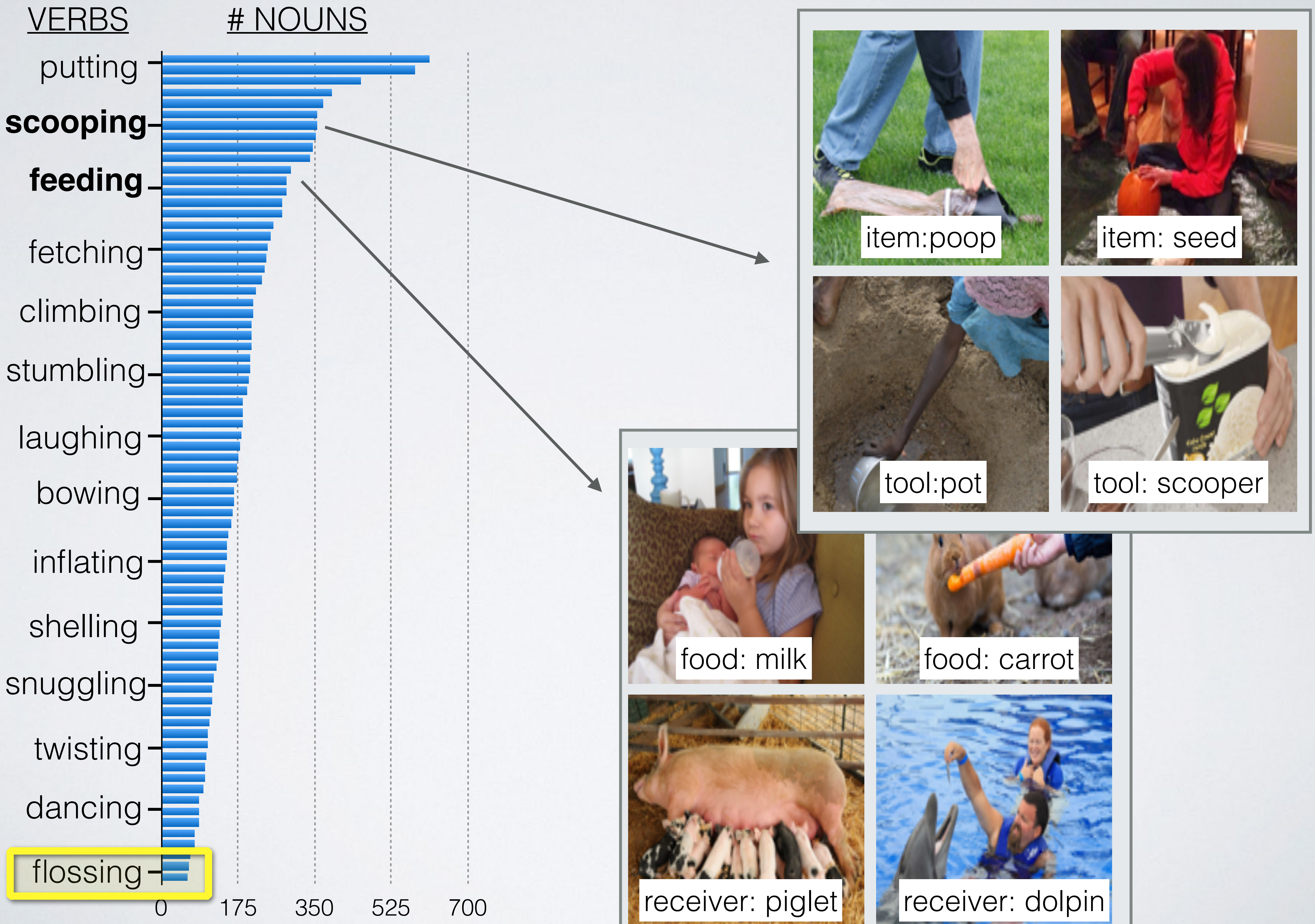
Despite 80,000 possible values, 2/3 annotators on 76.8% of role-value

# Skew - not all verbs are equal

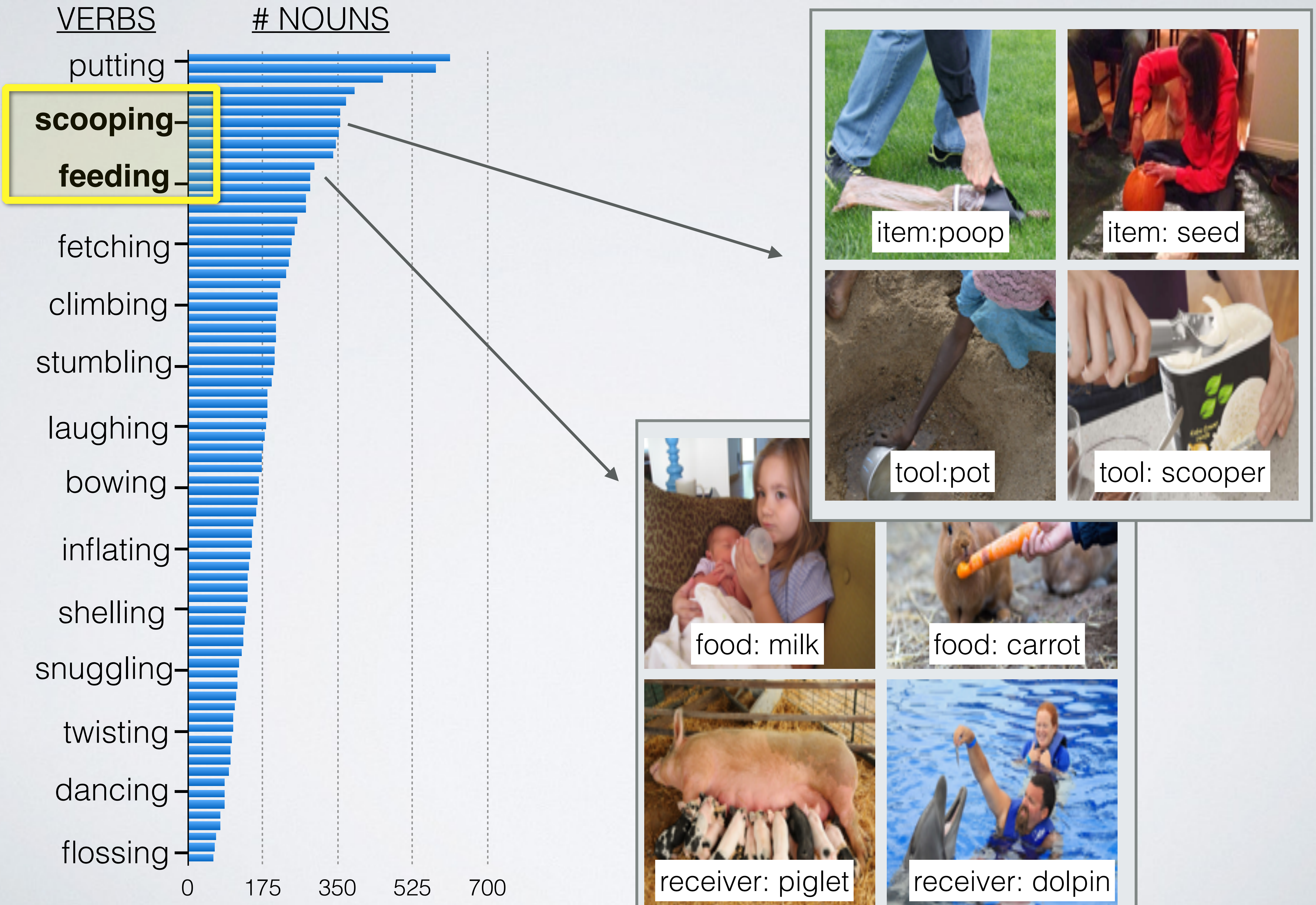


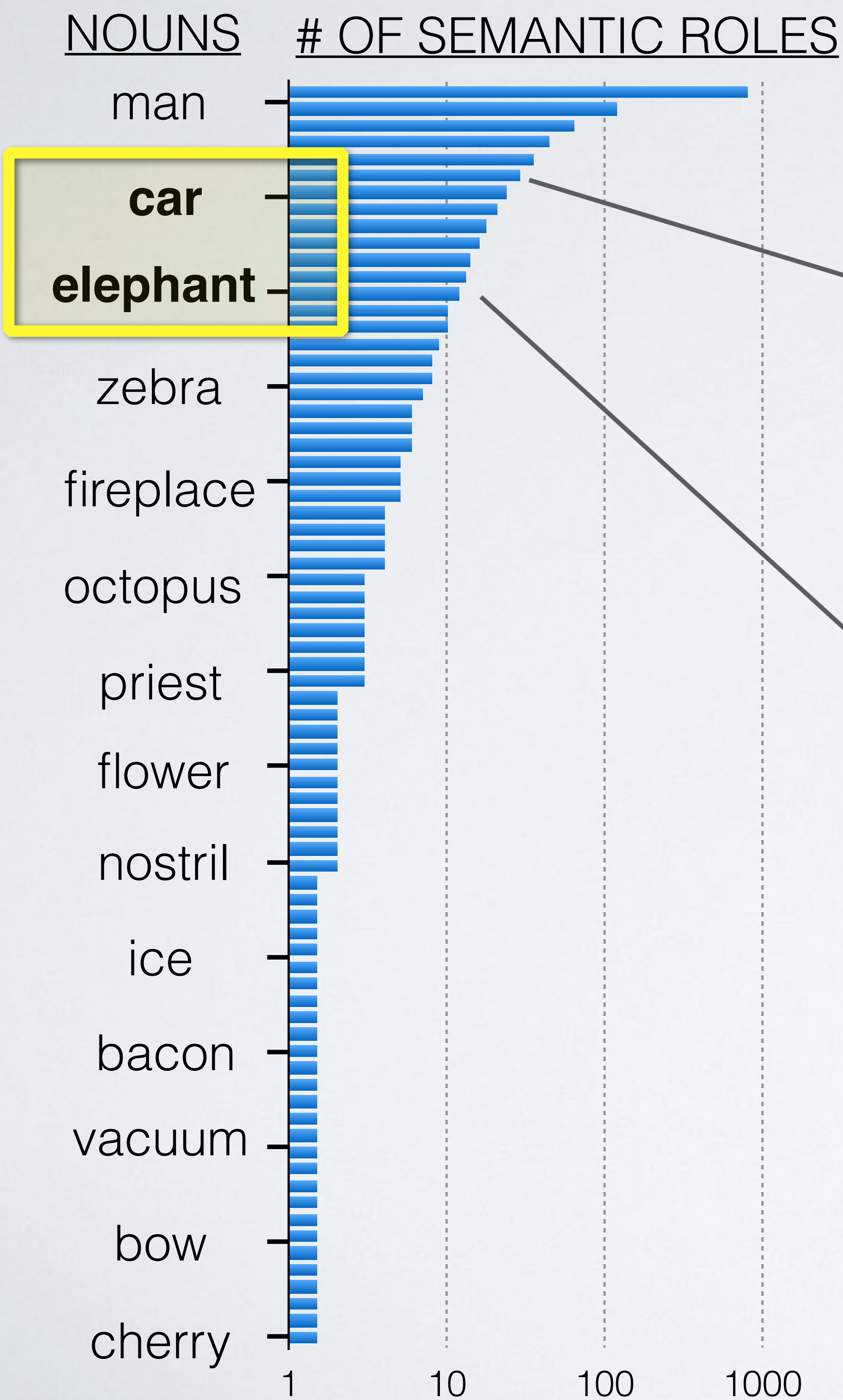

# Skew - not all verbs are equal

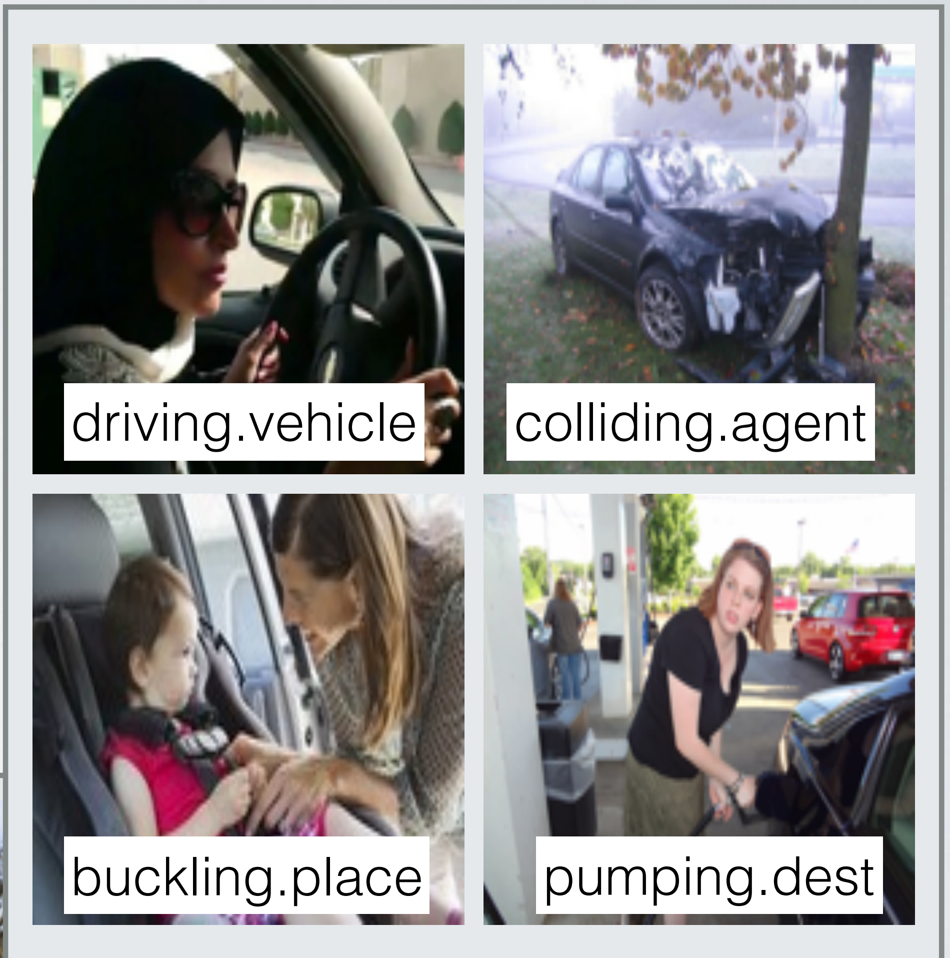


# Skew - not all verbs are equal

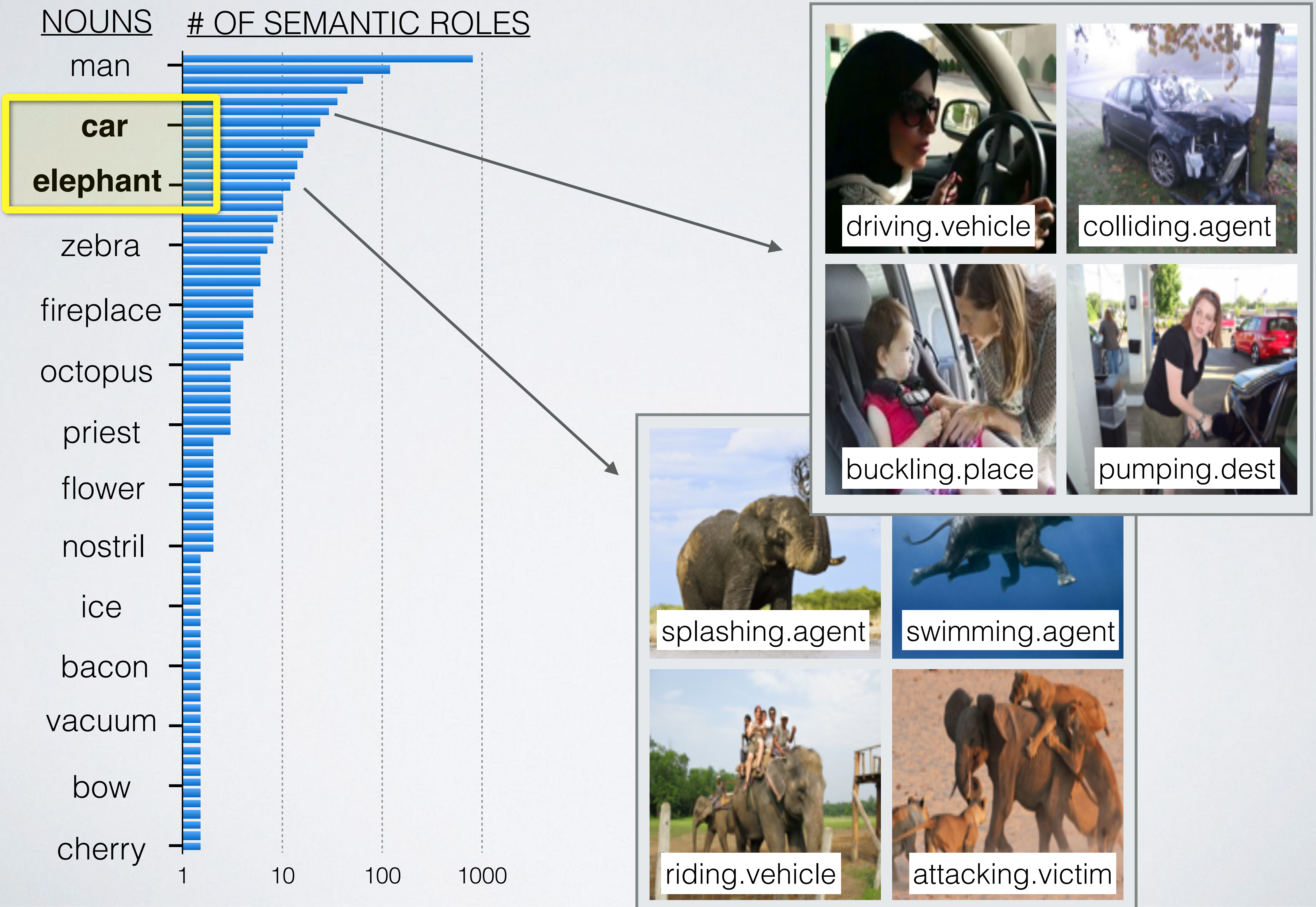


# Skew - not all nouns are equal




# Skew - not all nouns are equal

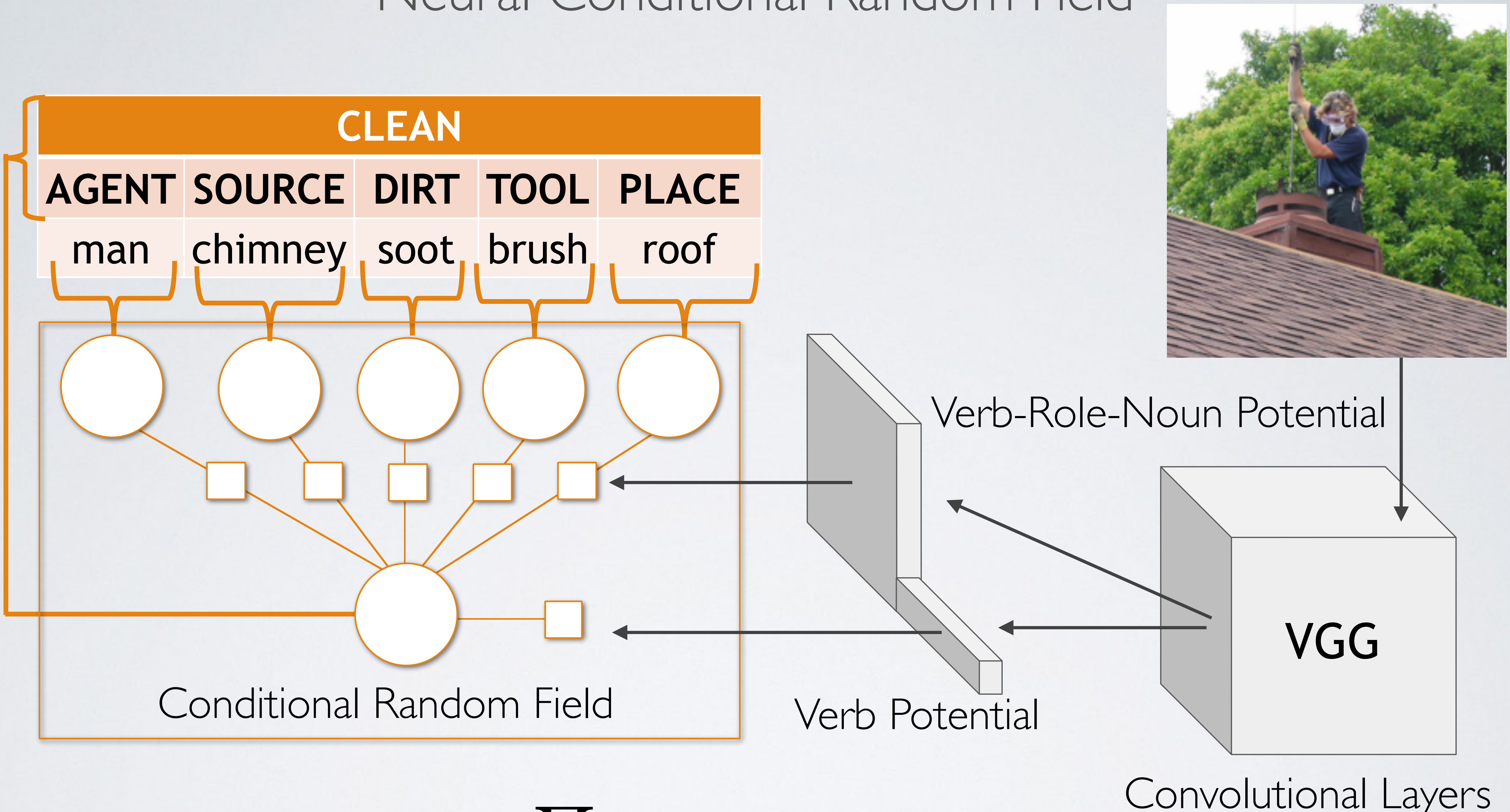


# Situation Recognition

## Models, Evaluation and Basic Results

structure matters
situation recognition improves object and activity recognition

# Neural Conditional Random Field




$$p(S|i;\theta) \propto \psi_v(v,i;\theta) \prod_{(r,n_r)\in F} \psi_e(v,r,n_r,i;\theta)$$

Backpropogate CRF loss through VGG

# Qualitative Examples

Gold Correct Incorrect

| SWIMMING | | |
|---|---|---|
| AGENT | SNAKE | SNAKE |
| PLACE | OCEAN | OCEAN |

| SPEARING | | |
|---|---|---|
| AGENT | PERSON | PERSON |
| VICTIM | FISH | FISH |
| PLACE | OCEAN | OCEAN |

| FALLING | | |
|---|---|---|
| AGENT | PERSON | PERSON |
| SOURCE | HORSE | HORSE |
| DEST. | GRND. | GRND. |
| PLACE | FIELD | FIELD |

# Qualitative Examples

Gold    Correct    Incorrect

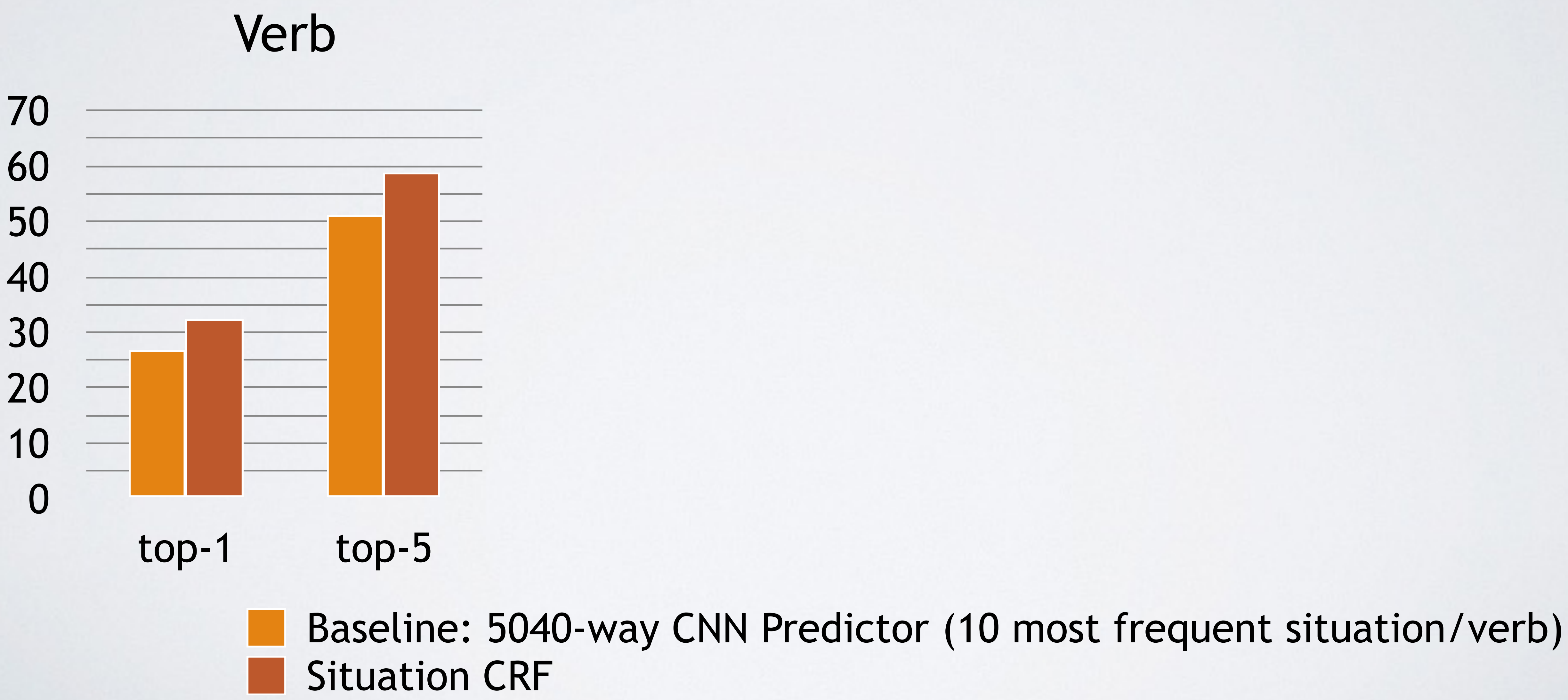
| SHAVING | | |
|---|---|---|
| AGENT | MAN | PERSON |
| CO-AGENT | MAN | MAN |
| BODYPART | HEAD | HEAD |
| SUBSTANCE | | S. CREAM |
| TOOL | RAZOR | RAZOR |
| PLACE | INSIDE | INSIDE |

| DETAINING | |
|---|---|
| AGENT | SOLDIER |
| VICTIM | MAN |
| PLACE | OUTSIDE |

| GIVING | |
|---|---|
| AGENT | SOLDIER |
| RECIPIENT | GIRL |
| ITEM | BAG |
| PLACE | OUTSIDE |

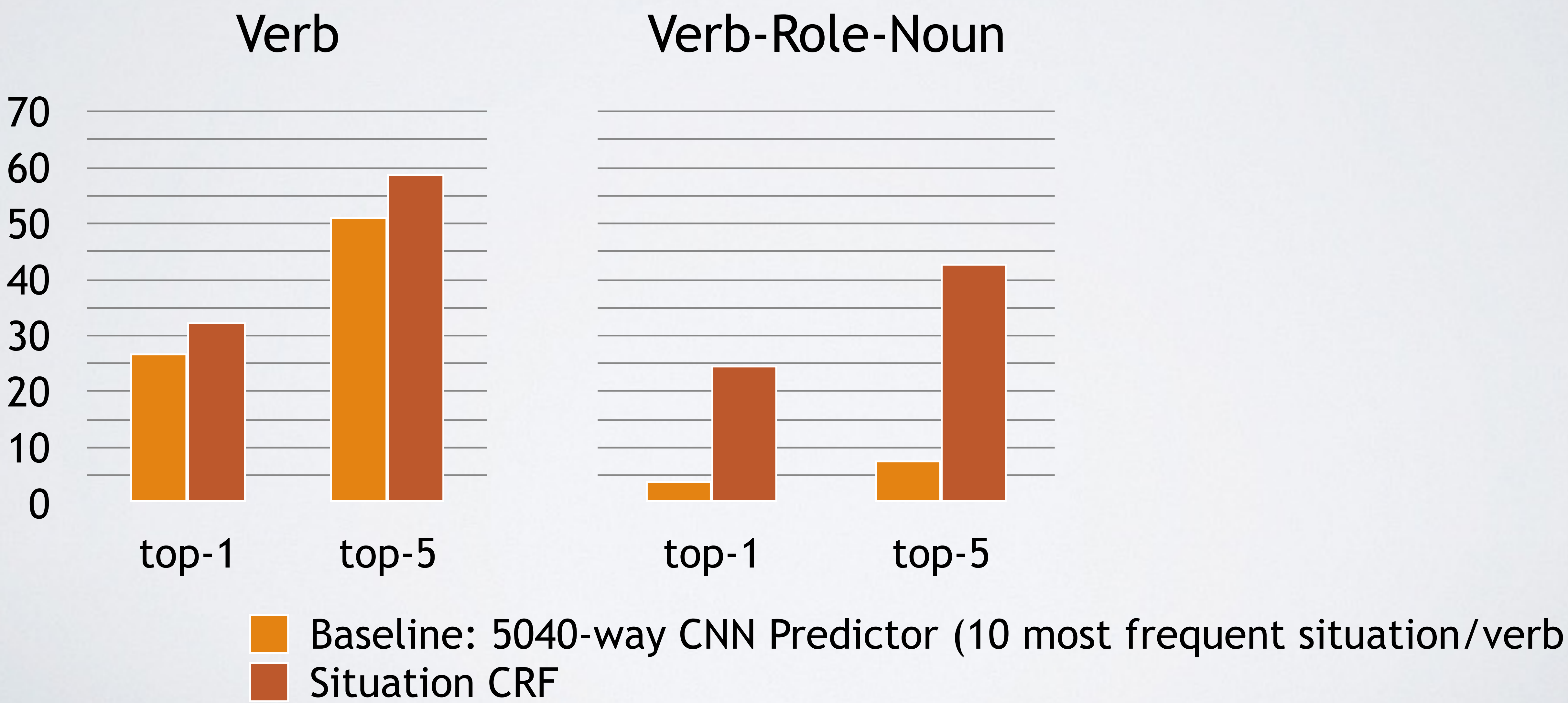
# Quantitive : Structured Prediction Crucial

| FIXING | | | | |
|---|---|---|---|---|
| AGENT | OBJECT | PART | TOOL | PLACE |
| BOY | CAR | TIRE | TIRE IRON | OUTDOORS |



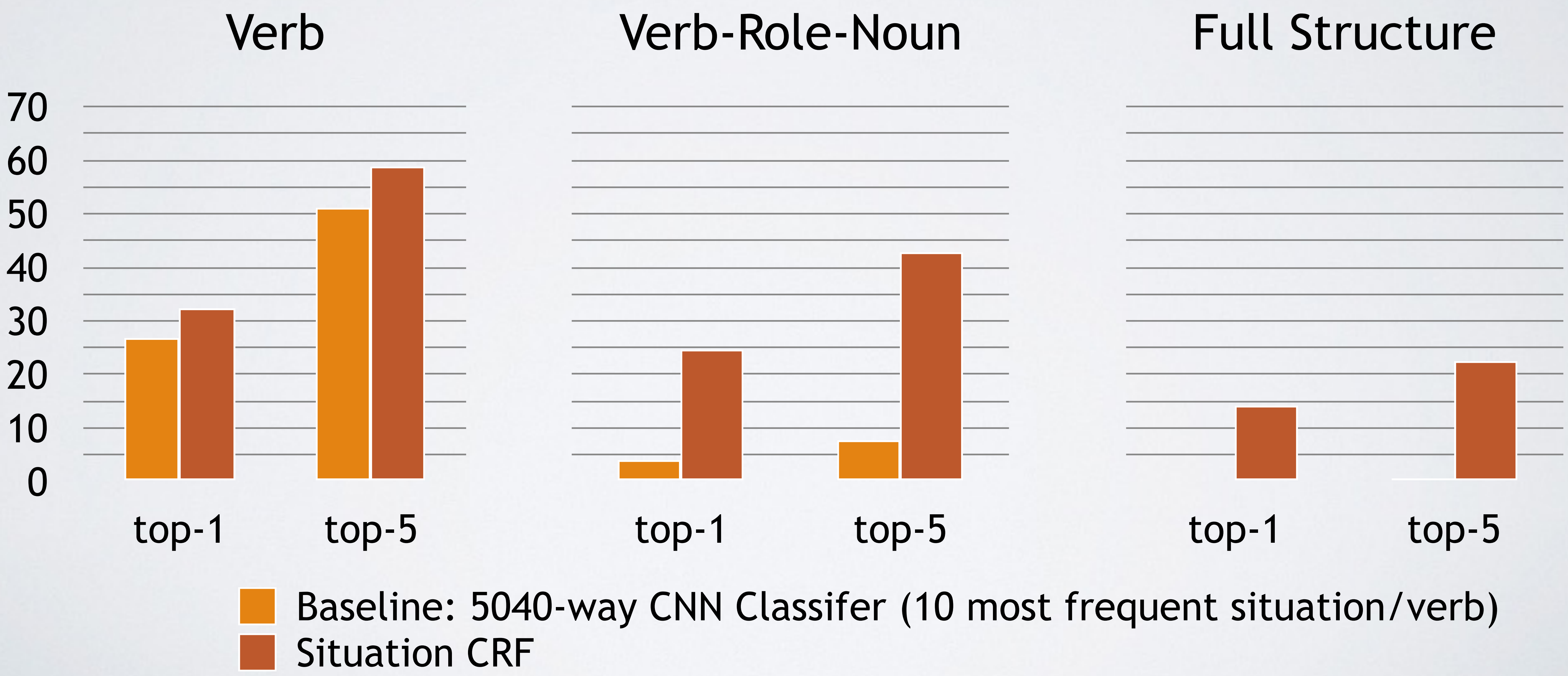
# Quantitive : Structured Prediction Crucial

| FIXING | | | | |
|---|---|---|---|---|
| AGENT | OBJECT | PART | TOOL | PLACE |
| BOY | CAR | TIRE | TIRE IRON | OUTDOORS |



# Quantitive : Structured Prediction Crucial

| FIXING | | | | |
|---|---|---|---|---|
| AGENT | OBJECT | PART | TOOL | PLACE |
| BOY | CAR | TIRE | TIRE IRON | OUTDOORS |



# Generalize to Unseen Combinations

## Train

| FEEDING | |
|---|---|
| AGENT | MAN |
| EATER | BABY |
| FOOD | MILK |
| SOURCE | BOTTLE |
| PLACE | ROOM |

Instances in train : 35

| FEEDING | |
|---|---|
| AGENT | GIRL |
| EATER | HORSE |
| FOOD | CARROT |
| SOURCE | HAND |
| PLACE | PEN |

Instances in train : 7

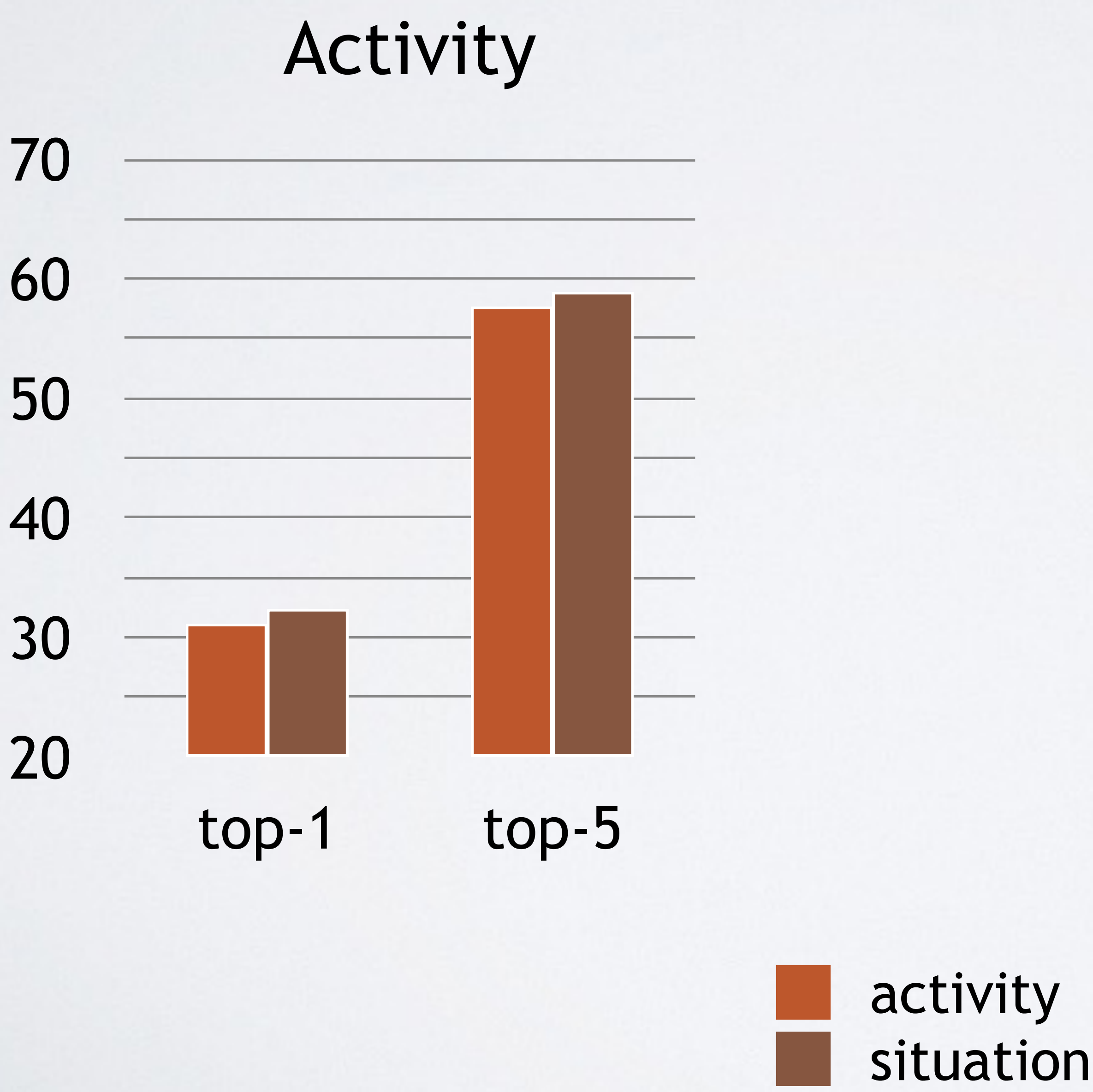
## Test

| FEEDING | |
|---|---|
| AGENT | WOMAN |
| EATER | HORSE |
| FOOD | MILK |
| SOURCE | BOTTLE |
| PLACE | BARN |

Instances in train : 0

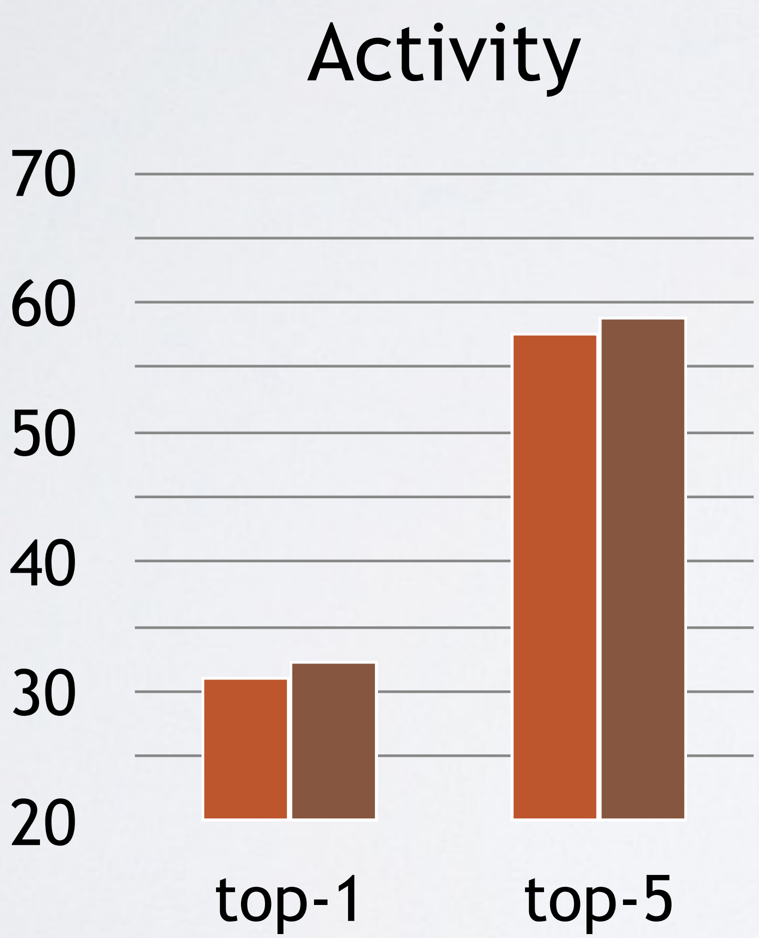
# Situations Improves Object and Activity Recognition

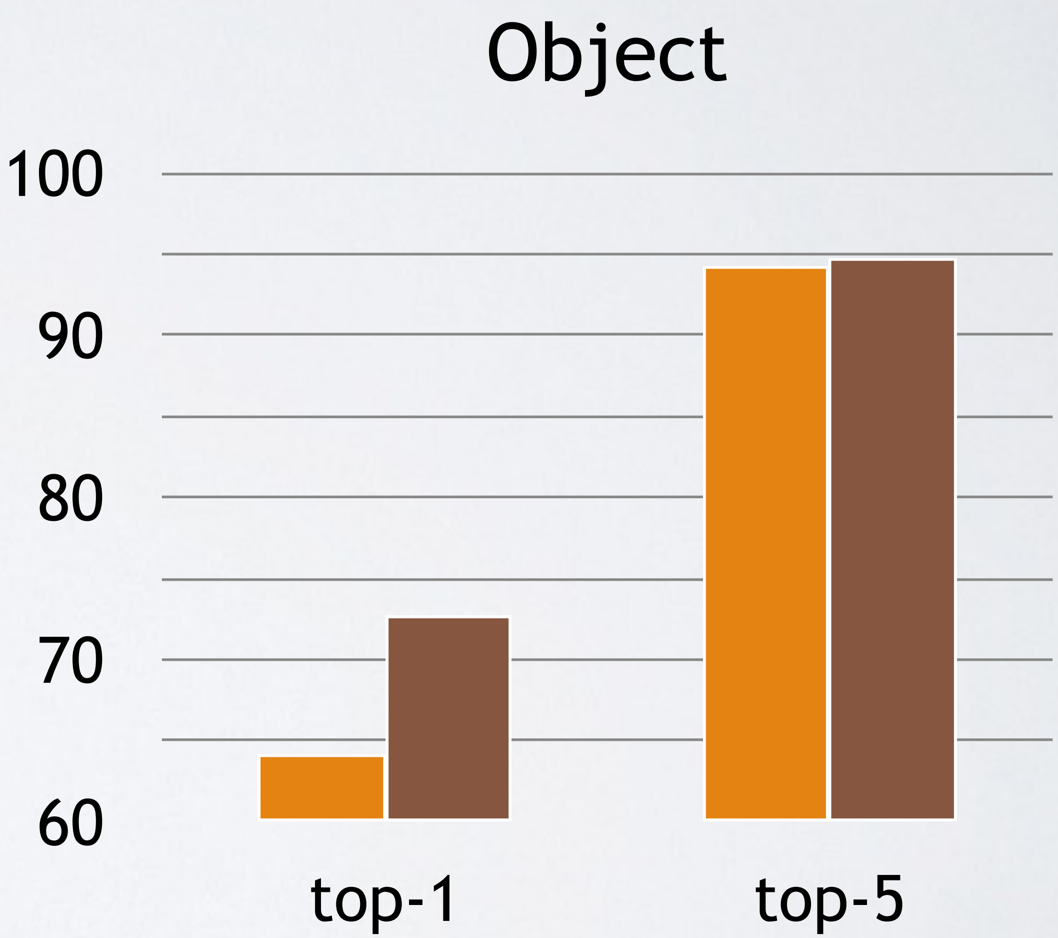
| FIXING | | | | |
|---|---|---|---|---|
| AGENT | OBJECT | PART | TOOL | PLACE |
| BOY | CAR | TIRE | TIRE IRON | OUTDOORS |



# Situations Improves Object and Activity Recognition

| FIXING | | | | |
|---|---|---|---|---|
| AGENT | OBJECT | PART | TOOL | PLACE |
| BOY | CAR | TIRE | TIRE IRON | OUTDOORS |

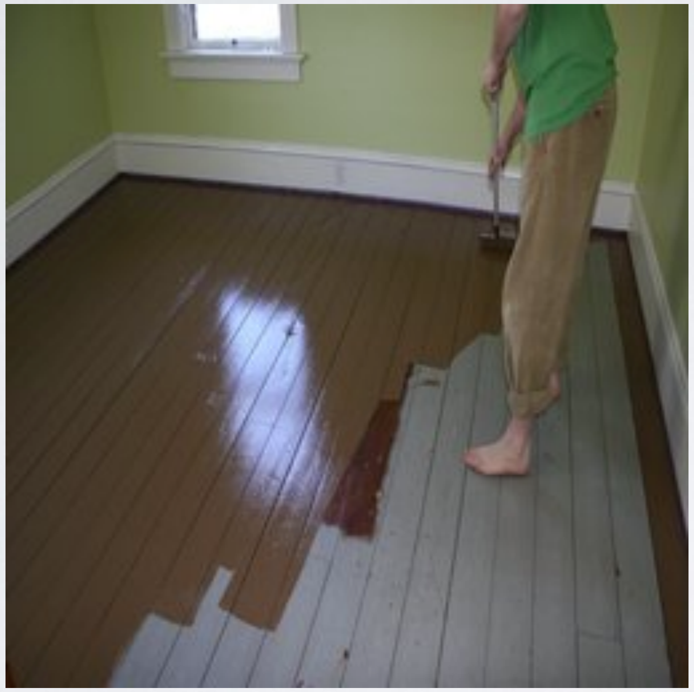



object
activity
situation

# Errors

PAINTING



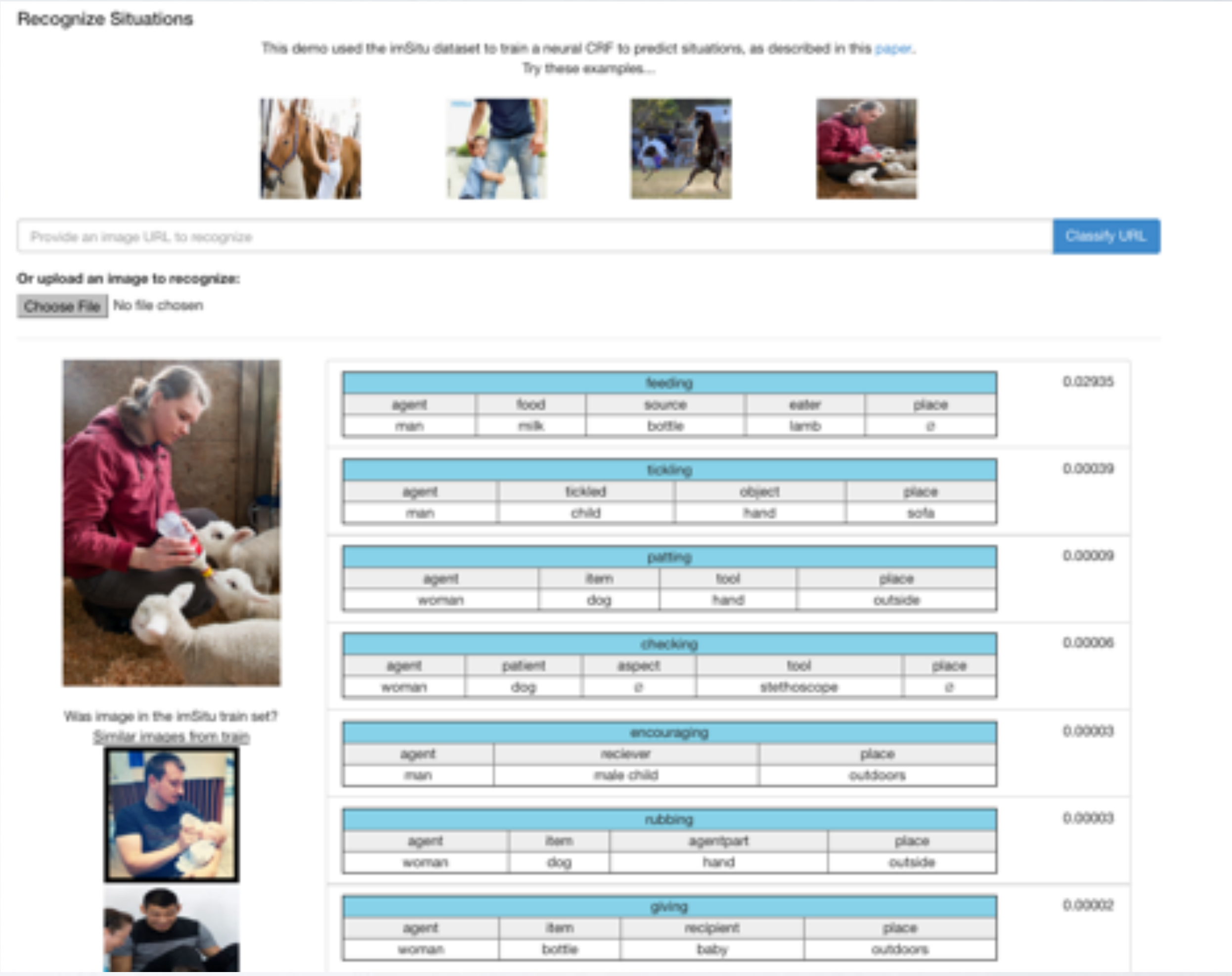
| PRYING | |
|---|---|
| AGENT | PERSON |
| ITEM | WOOD |
| SOURCE | FLOOR |
| TOOL | CROWBAR |
| PLACE | ROOM |

SPRAYING

| PUMPING | |
|---|---|
| AGENT | PERSON |
| ITEM | AIR |
| SOURCE | AIR |
| DESTINATION | WHEEL |
| TOOL | PUMP |
| PLACE | OUTSIDE |

# imsitu.org

data/browsing/demo/code



# Conclusion

Introduced **situation recognition**

**role**-centric structured representation of whats happening

Collected **imSitu**

120k+ images, 500+ verbs, 100k+ situations

Introduced simple model **neural CRF** for situation

structure matters

provides strong context for activity and object recognition

data/browsing/demo/code

imsitu.org