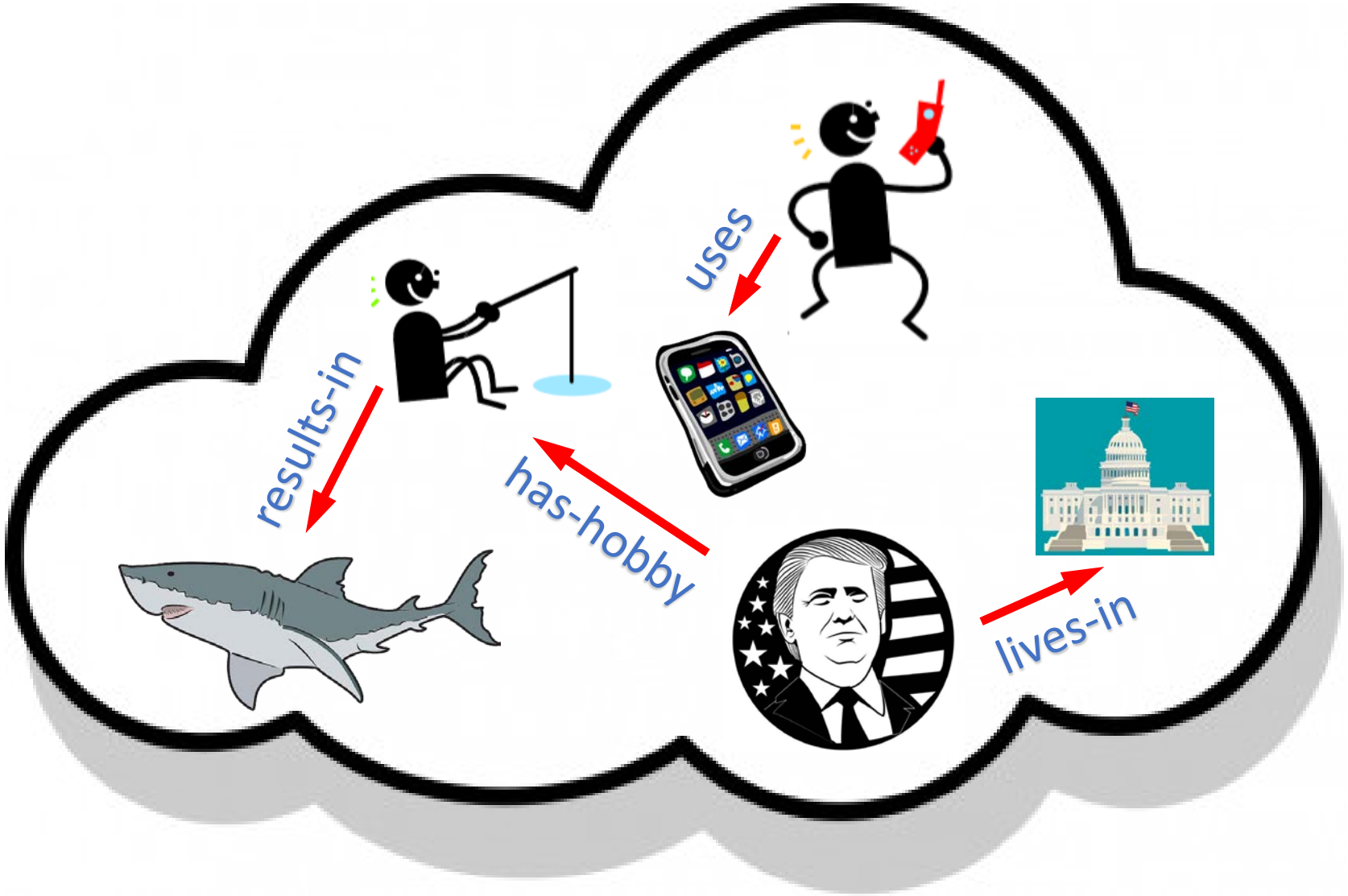
Towards Building Large-Scale Multimodal Knowledge Bases

Dihong Gong

Advised by Dr. Daisy Zhe Wang

Knowledge Itself is Power

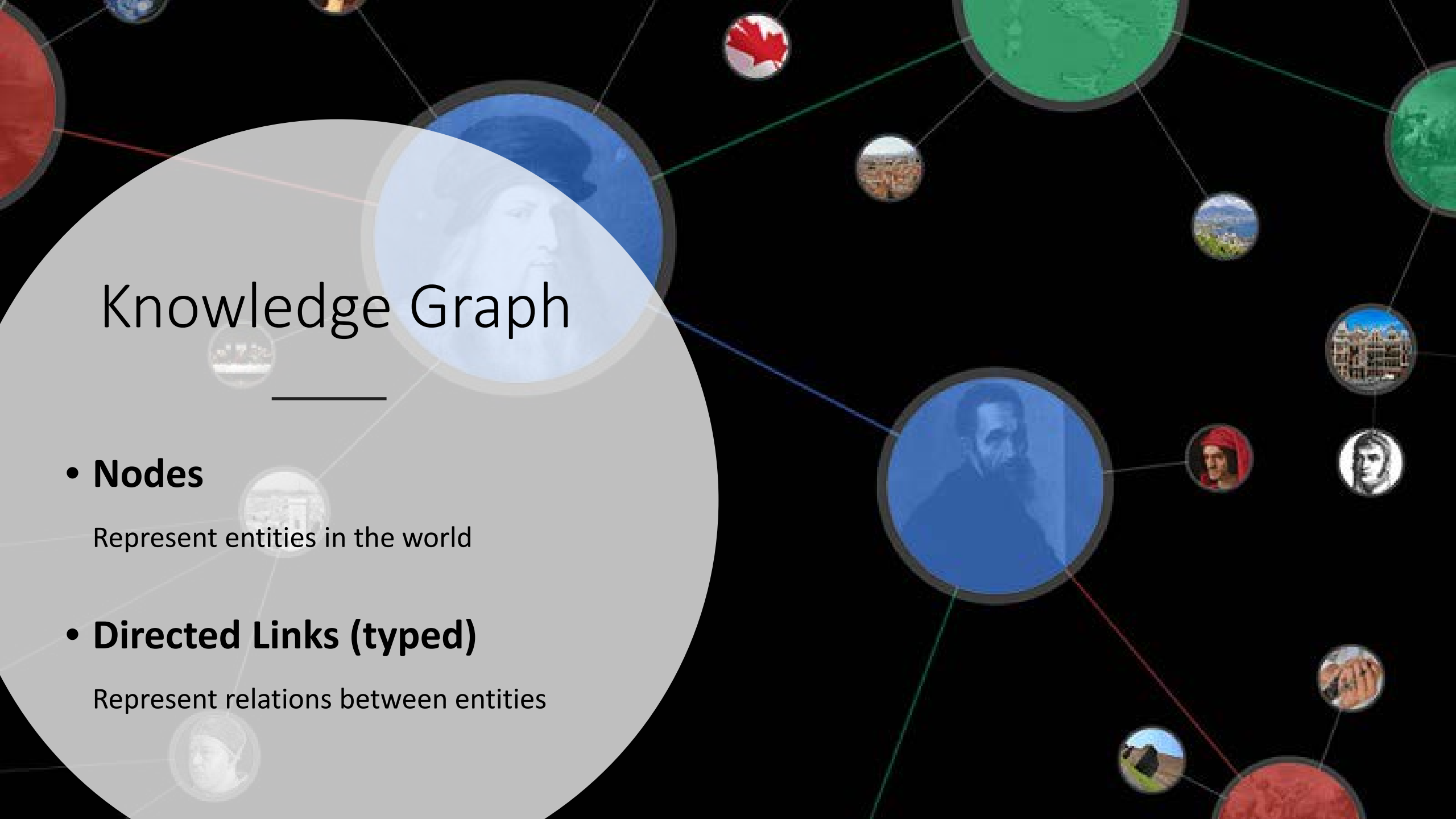
--Francis Bacon



Analytics

Social

Robotics



Knowledge Graph

- **Nodes**

  Represent entities in the world

- **Directed Links (typed)**

  Represent relations between entities

Given

1. A directed edge-typed graph G
2. Relation $R(\mathrm{x}, y)$
3. Source node s

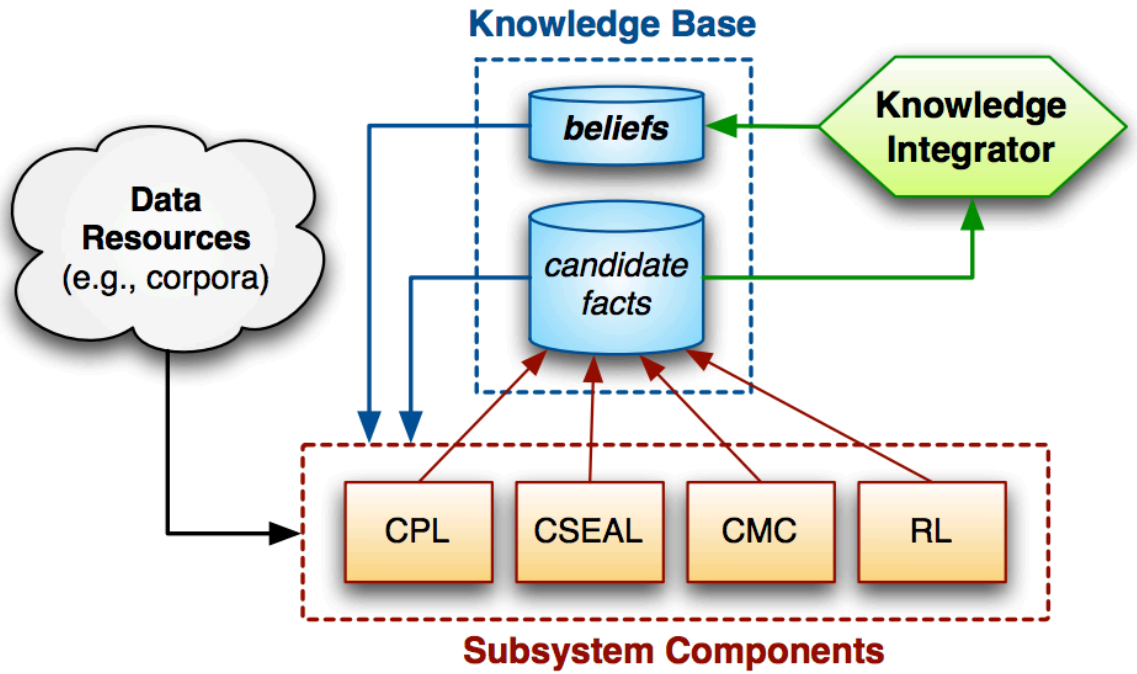
Find

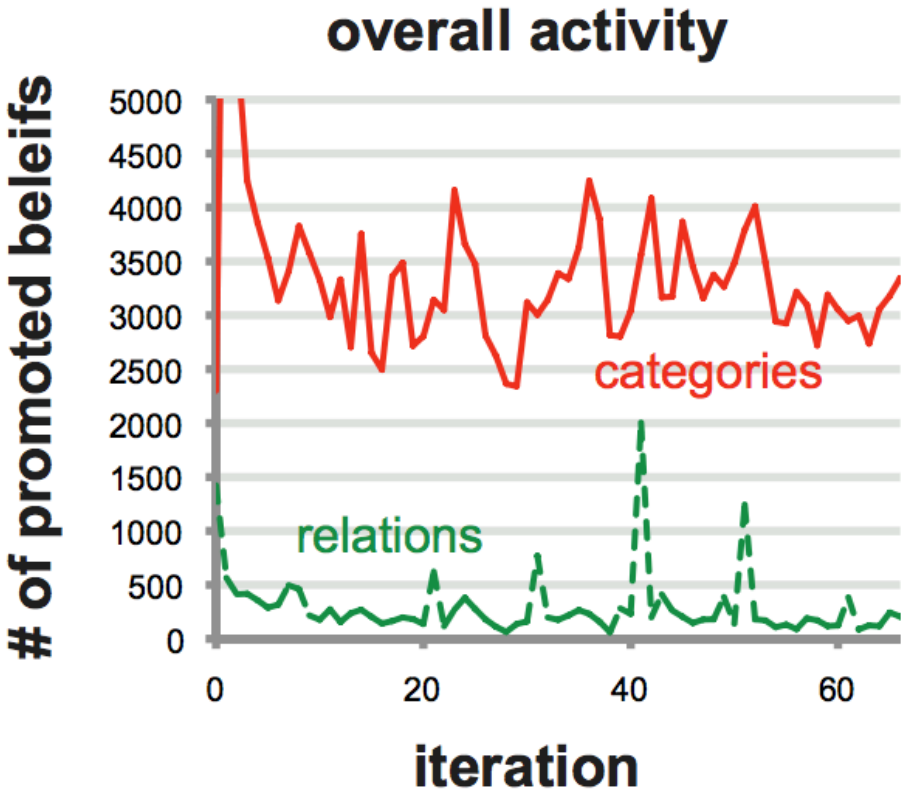
A set of nodes T in G: each $t \in T$ the relation $R(s, t)$ is true.

Link Prediction

A generic framework for knowledge expansion

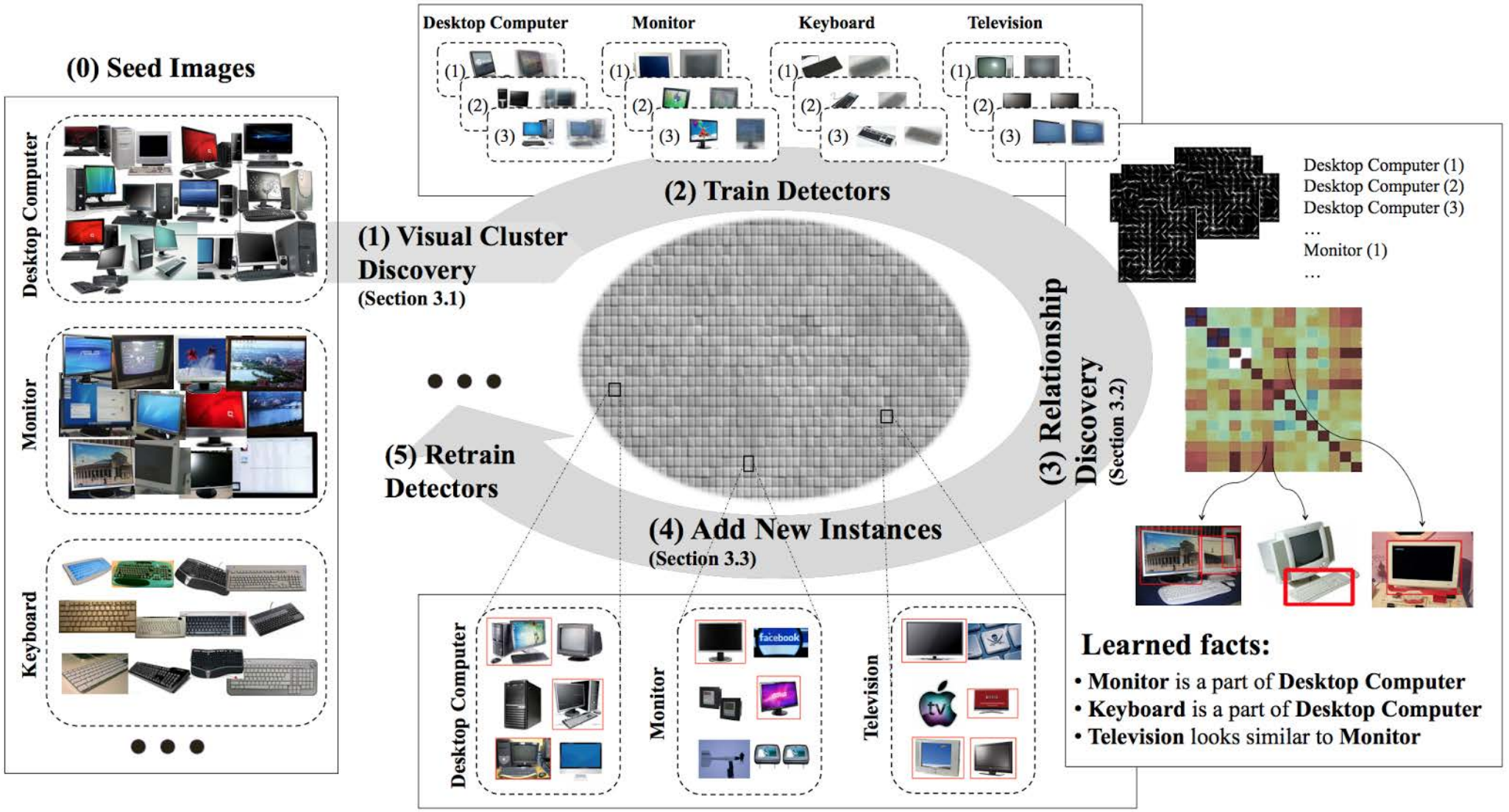
NELL: Text Knowledge Miner



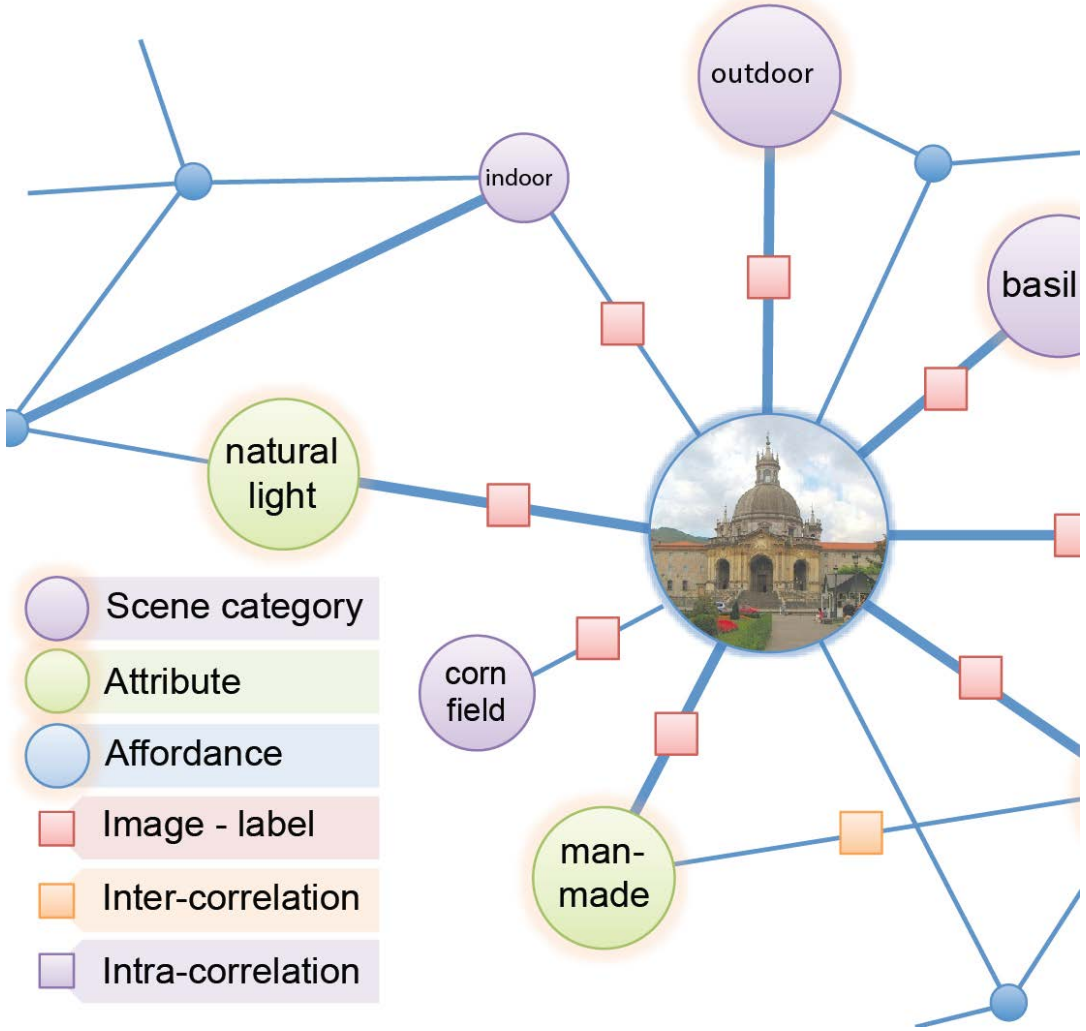


Tasks: (1) extract text knowledge from web. (2) focused on link prediction with type “is-a”.
Learning iteratively (24x7): **initial seeds** => **training extractors** => **extract instances** => **repeat**

NEIL: Image Knowledge Miner



Multimodal Knowledge as a Factor Graph



(a) **Find me pictures of a dog.**

(b)

Q: Where can I find similar cuisines in downtown Chicago?

Answers:

U. U. Grill
Chicago, IL 60642

S. C. Steak House
Chicago, IL 60657

Q: Find photos of me sea kayaking last Halloween in my photo album.

Answers: Saturday, October 31

Multimodal Knowledge Graph

- **Nodes**

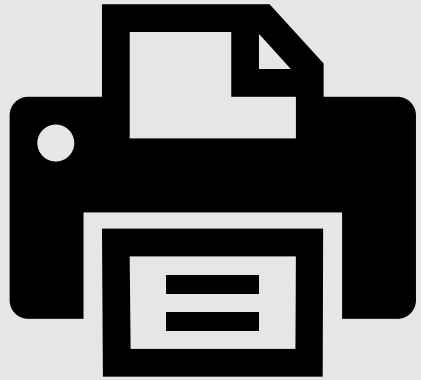
  Represent entities in the world

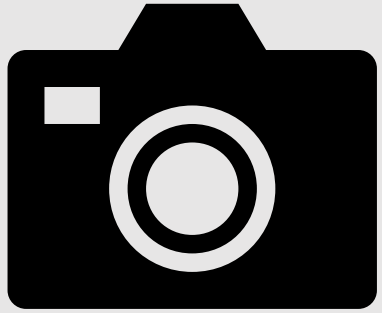
  Entities are multimodal: text, image, audio

- **Directed Links (typed)**

  Represent relations between entities

  Relations link entities within & across modalities





Example

Sea

Island

Maui

Beach

Scenery

castle

sunset

sky

Building

Nodes
- image
- text
- visual-concept:
- textual-concept

Edge Types
- is-a
- has-visual-object
- meta-of
- has-tag
- co-locate-with

Multimodal Link Prediction

- Physical Links

  The links that represent **physical relations** between nodes.

  These links are directly accessible (e.g. by document parsing) without further logical reasoning.

  Examples: co-locate-with, has-tag, meta-of, etc.

- Logical Links

  The links that represent **logical relations** between nodes.

  These links are obtained by logical reasoning, or link prediction.

  Examples: is-a (text), has-visual-object (image).

Why is Multimodal Appealing?

1. Encode relations between multimodal objects.

   Visual question answering (VQA).

   Complicated scene understanding, e.g. autonomous driving.

2. Multimodal information is complementary.

   Information from different independent sources complement each other.

3. Multimodal information is correlated.

   Information tends to correlate, which provides additional redundancy for better robustness.

Contributions

Multimodal Named Entity Recognition (MNER)

Utilize visual relations in the multimodal knowledge graph for enhanced text “is-a” link prediction.

D. Gong, D. Wang, Y. Peng, “Multimodal Learning for Web Information Extraction”, ACM Multimedia, 2017.

(Rank A+ conference)

Visual Text-Assisted Knowledge Extraction (VTAKE)

Employ textual relations to improve precision of “has-visual-object” link prediction.

D. Gong, D. Wang, “Extracting Visual Knowledge from the Web with Multimodal Learning”, IJCAI, 2017.

(Rank A+ conference)

Multimodal Convolutional Neural Network (MCNN)

Improve from VTAKE, by a unified end-to-end CNN model, for “has-visual-object” link prediction.

D. Gong, D. Wang, “Multimodal Convolutional Neural Networks for Visual Knowledge Extraction”, AAAI, 2018.

(Under review)

MNER

Goal: extract text instances of predefined categories (e.g. predicting “is-a” links).

Textual Rules

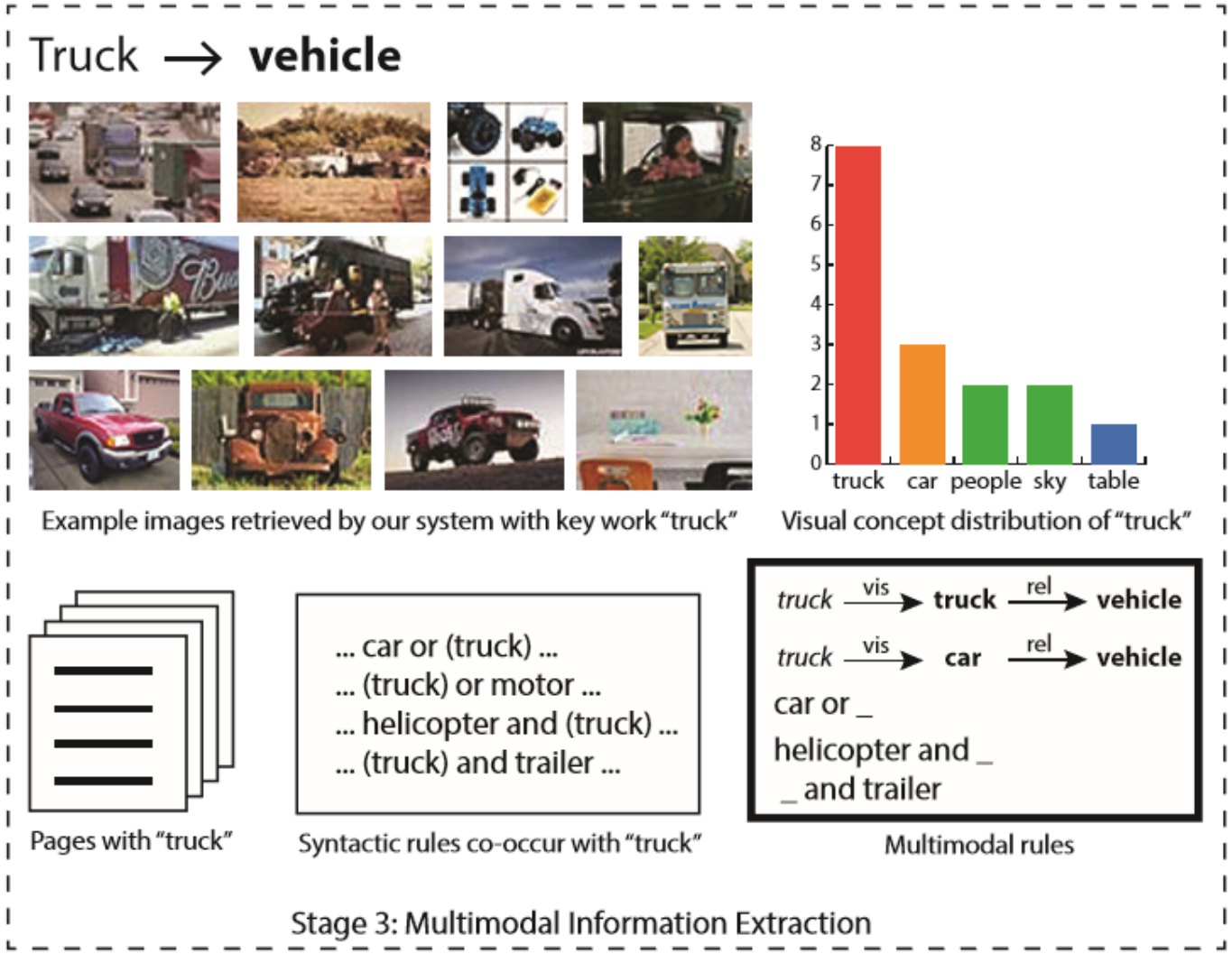
- Companies such as _
- _ is river

Visual Rules

- acura->car->automobilemaker
- adam_sandler->person>director

```
mysql> select * from TCB where iter = 1 limit 20;
+-----------------+-----------------+--------+------+
| entity          | category        | prob   | iter |
+-----------------+-----------------+--------+------+
| acura           | automobilemaker | 0.9999 |    1 |
| adam_sandler    | director        | 0.3935 |    1 |
| all             | musicinstrument | 0.3935 |    1 |
| amazon          | company         | 0.7210 |    1 |
| andres_iniesta  | athlete         | 0.3935 |    1 |
| angelina_jolie  | actor           | 0.3935 |    1 |
| animals         | animal          | 0.3935 |    1 |
| ants            | insect          | 0.3935 |    1 |
| apricots        | fruit           | 0.5115 |    1 |
| argentina       | athlete         | 0.3935 |    1 |
| arjen_robben    | athlete         | 0.3935 |    1 |
| atlantic        | river           | 0.3935 |    1 |
| audi            | automobilemaker | 0.9999 |    1 |
| banana          | fruit           | 0.4160 |    1 |
| banco_santander | bank            | 0.3935 |    1 |
| bank            | bank            | 0.8647 |    1 |
| bering_sea      | river           | 0.3935 |    1 |
| berries         | fruit           | 0.4119 |    1 |
| blackberries    | fruit           | 0.3956 |    1 |
| blueberries     | fruit           | 0.6321 |    1 |
+-----------------+-----------------+--------+------+
```

Overview

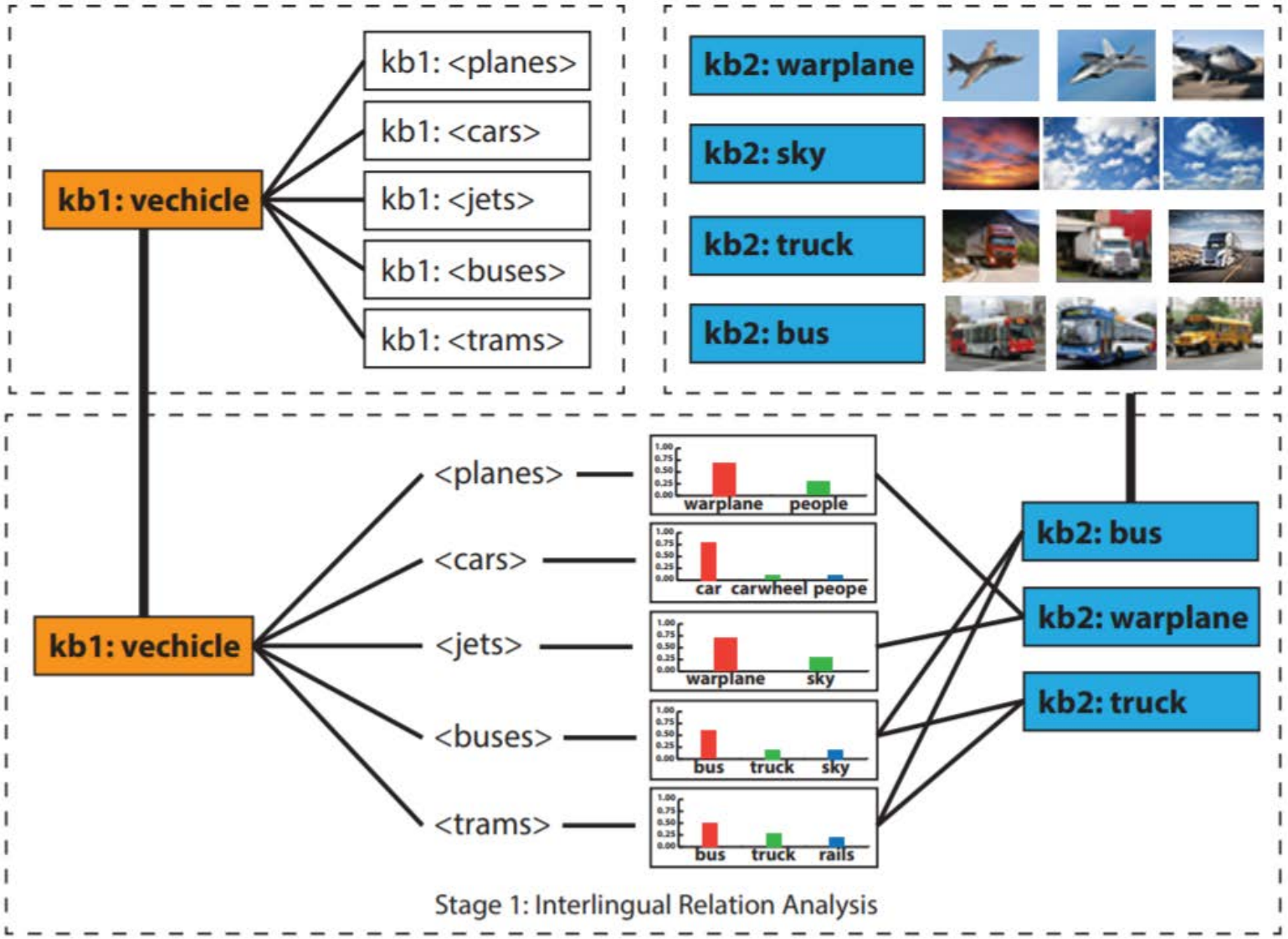


Three Stages

1. Multimodal Relation Analysis
2. Learn Multimodal Rules
3. Multimodal Information Extraction

Category	Related Visual Concepts
bridge	bridge, suspension bridge
coach	man's clothing, musician
fish	pizza, striped bass, salmon
clothing	skirt, woman's clothing, jersey
automobilemaker	car, beach wagon, sports car
city	people, sky, window
lake	seashore, ship, cliff, sky
actor	man's clothing, sunglasses
vehicle	warplane, bus, motorcycle, ship
beach	seashore, beach, musical instrument
bird	wading bird, female child, loon
company	computer screen, laptop
hotel	building complex, bed
fruit	orange, fringe tree
airport	warplane, bed, kitchen

Stage #1: Multimodal Relation Analysis



Relations:

1. Vehicle – Bus
2. Vehicle – Warplane
3. Vehicle – Truck

Stage #2: Learn Multimodal Rules

Rules Template: $r(x, V_c, T_c) : x \xrightarrow{vis} V_c \xrightarrow{rel} T_c,$

Rules Measurement: $Precision(r) = \frac{count(r, s_p)}{count(r)}$

$Recall(r) = \frac{count(r, s_p)}{count(s_p)}$

Example Visual Rules

$r(x, bus, vehicle) : x \xrightarrow{vis} \mathbf{bus} \xrightarrow{rel} \mathbf{vehicle}$

$r(x, warplane, vehicle) : x \xrightarrow{vis} \mathbf{warplane} \xrightarrow{rel} \mathbf{vehicle}$

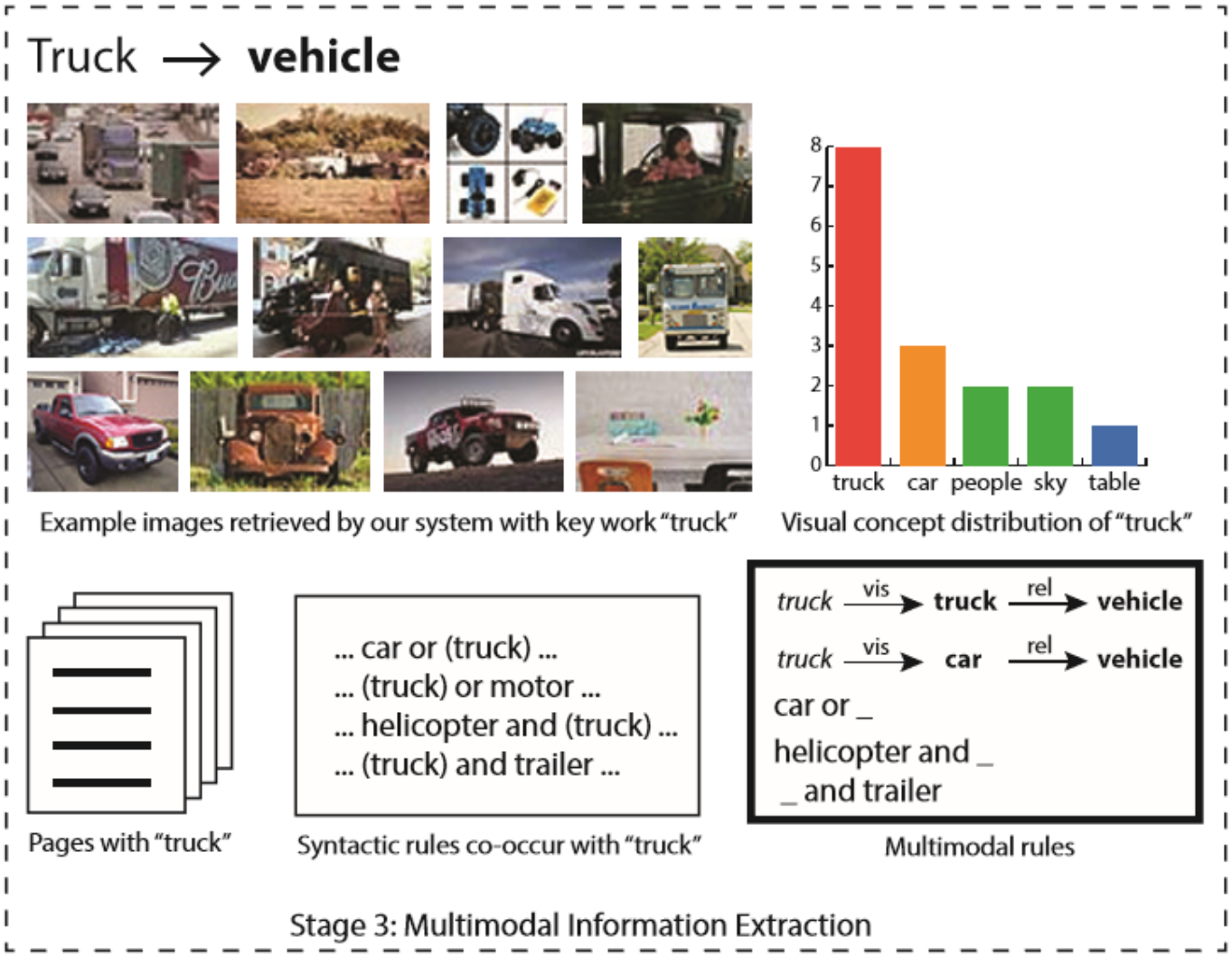
$r(x, truck, vehicle) : x \xrightarrow{vis} \mathbf{truck} \xrightarrow{rel} \mathbf{vehicle}$

Stage 2: Learning Multimodal Rules (only visual rules are shown)

Stage #3: Multimodal Information Extraction

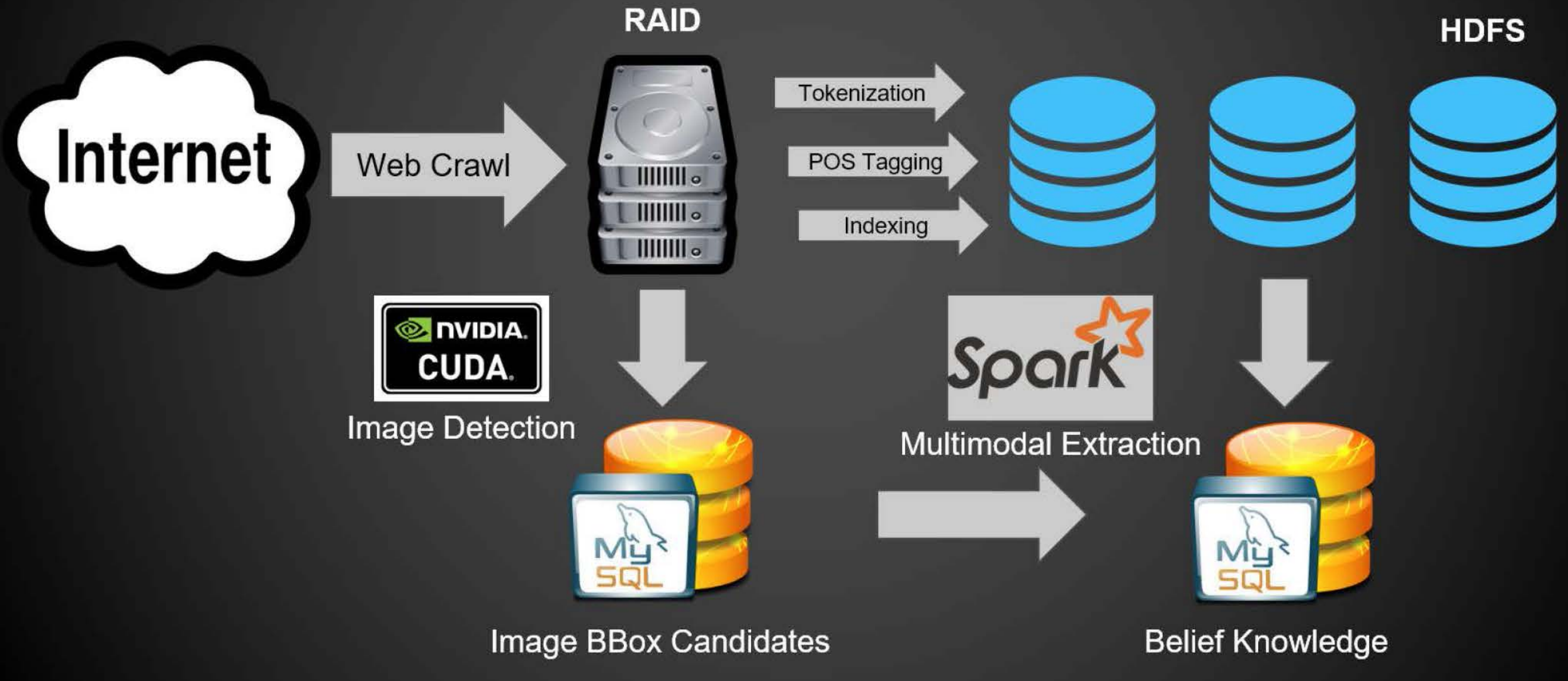
Ranking

$$P(e) = 1 - \prod_{x \in R_t(e)} \prod_{y \in R_v(e)} (1 - P(x))(1 - P(y))$$



Implementations

-- System Framework



500M English web pages.

20M images with title, name and alt meta information.

42 text categories, with each category having 15 seed instances.

116 image categories, with each category having 250 seed instances.

Evaluation

Results

Category	CPL	CPL-NMF	Proposed
vehicle	69.43	80.24	85.75
automobilemaker	86.23	90.11	95.16
fish	80.91	75.67	92.86
bird	68.24	71.28	80.22
bridge	42.57	48.62	52.81
hotel	68.47	78.04	76.45
clothing	80.41	91.72	94.11
airport	85.73	81.37	90.22
musicinstrument	88.14	87.41	91.75
consumerelectronicitem	65.20	70.84	75.23
beach	70.32	71.22	73.04
lake	59.66	63.97	62.90
river	79.64	81.08	89.22
company	96.41	94.54	97.65
plant	74.35	80.22	81.45
insect	71.03	76.25	83.82
city	95.88	91.57	94.47
coach	93.27	95.22	95.74
fruit	74.68	65.32	67.54
actor	95.73	98.65	98.17
athlete	90.05	92.11	94.11
governmentorganization	68.37	70.21	71.43
drug	98.22	96.74	97.83
ceo	79.68	79.24	77.21
Average	**78.44**	**80.48**	**84.13**

Comparative Approaches

CPL: Coupled Pattern Learner (text-only)
CPL-NMF: Fuse CPL score with visual score in a “naive manner”

Observations

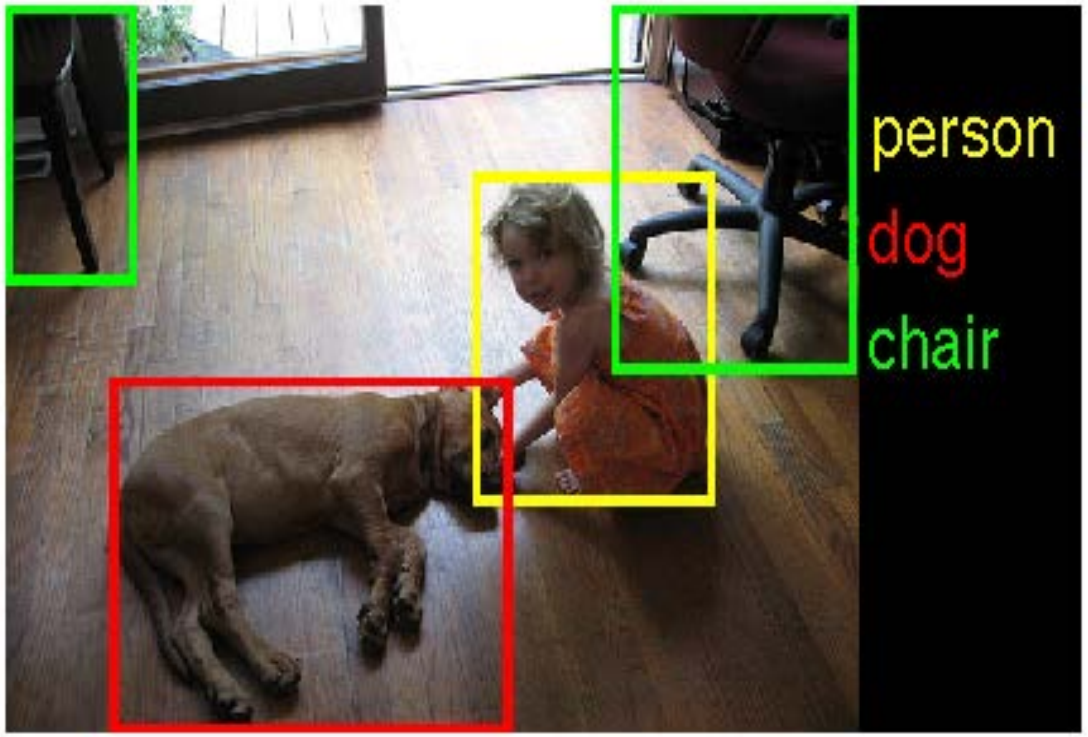
- Based on the testing categories, the proposed approach outperforms the text-only CPL by clear margin.
- However, for categories that do not have stable visual correspondence, the improvement is not significant.
- It remains a challenge to generalize the approach to general categories.

Significantly Improved (>10%)	Fairly Improved (<5%)	Declined
• Vehicle (+16%): bus(0.70,0.20), warplane(0.27,0.20), car(0.11,0.4) • Fish (+12%): striped_bass(0.55,0.36), game_fish(0.48,0.21), salmon(0.32,0.43) • Bird (+13%): wading_bird(0.65,0.20), crane(0.21,0.16), insect(0.10,0.07)	• Governmentorganization (3%): government_building(0.20,0.33) • Coach (+3%): player(0.12,0.2), man's clothing(0.08,0.64) • Athlete (2%): player(0.18,0.53), people(0.06,0.25), man's clothing(0.03,0.41)	• Drug (-2%): mouse(0.06,0.08) • City (-4%): window(0.09,0.17), sky(0.06,0.15), people(0.01,0.35) • Ceo (-1%): sunglasses(0.04,0.25), civilian_clothing(0.02,0.31), man's_clothing(0.01,0.43)

Example Visual Rules

VTAKE

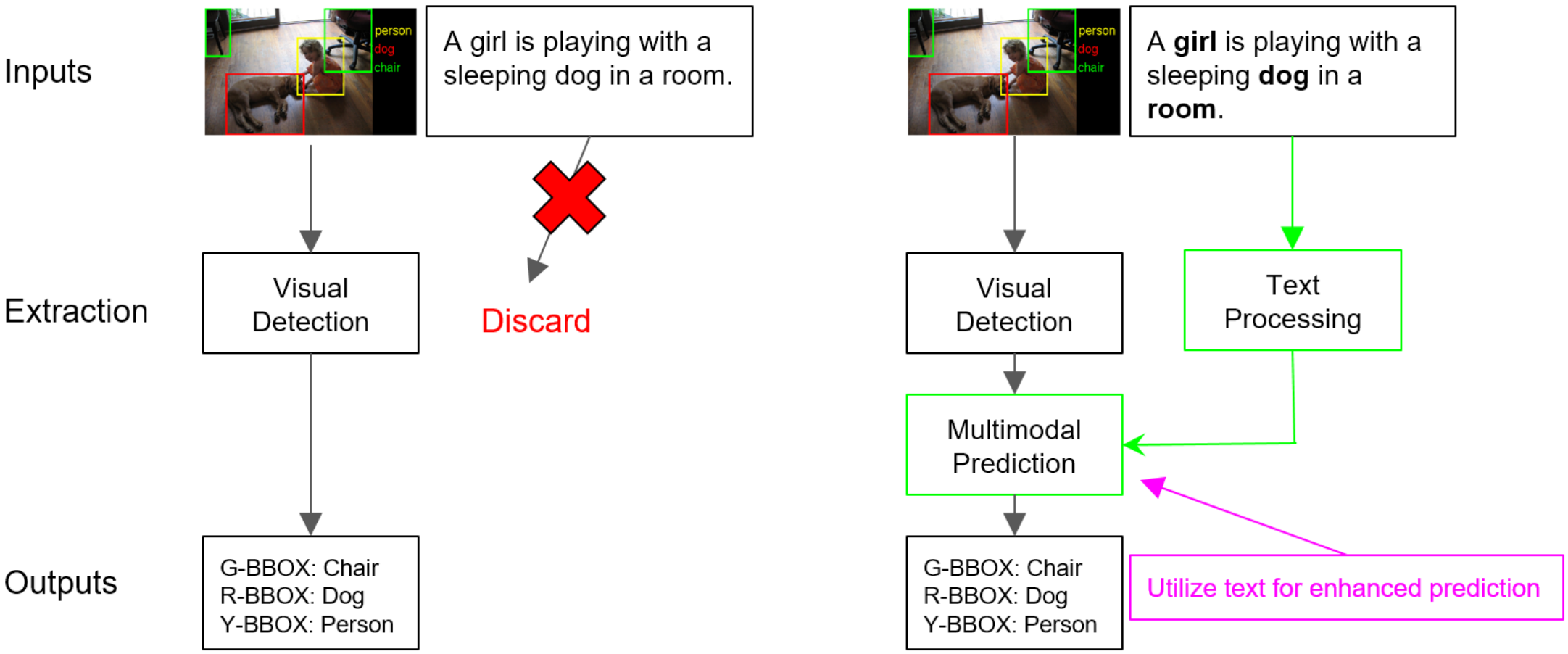
Goal: extract object bounding boxes from web images (e.g. predicting “has-visual-object” links).



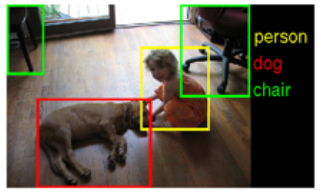
Challenge

As of 2016, the best visual object detection programs on ImageNet challenge has 66% mean average precision (200 categories).

Traditional Approaches vs. Our Approach



Overview



A **girl** is playing with a sleeping **dog** in a **room**.

Step #1: Multimodal Embedding

- Detection outputs: person, dog, chair.
- Text Processing outputs: girl, dog, room.
- Convert ALL 6 objects into 500-d vectors.
- Note: objects that are correlated stronger has closer vector representations.

Step #2: Prediction (sparse logistic regression)

$$p_{\boldsymbol{\theta}}(c|W_n) = \frac{e^{\theta_0 + \sum_{w_k \in W_n, w_k \neq c} \theta_k \vec{v}(w_k)^T \vec{v}(c)}}{1 + e^{\theta_0 + \sum_{w_k \in W_n, w_k \neq c} \theta_k \vec{v}(w_k)^T \vec{v}(c)}}, \quad (6)$$

Score is higher if there are more "words" correlated with [c] in W_n.
e.g. c = person, correlated "words" can be room, chair, girl.

e.g.,

W_n = {person, dog(v), chair, girl, dog(t), room}
c => a visual category such as person

Multimodal Embedding: Skip-Gram Model

Learn vector representation of multimodal objects,
by maximizing probability of predicting “nearby” words:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t), \qquad (1)$$

where $w_1, w_2, \ldots w_T$ is a sequence of training words in the corpus, and c is the size of the window around target w_t. In the basic Skip-Gram model, the conditional probability $p(w_{t+j}|w_t)$ is defined using the softmax function as

$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^T v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left({v'_w}^T v_{w_I}\right)}, \qquad (2)$$

e.g. {person, dog(v), chair, girl, dog(t), room} should predicts each other. Thus, their vector representation is closed to each other.

Learning Sparse Logistic Regression Model

- Train one model per image category.
- Training data: all image instances that detect objects in category C are positive samples, and the rest are negative samples.
- Sparsity: efficiently suppress unrelated multimodal words.

Maximize:

$$L(\boldsymbol{\theta}, c) = \sum_{n \in \mathcal{P}} \frac{\ln p_{\boldsymbol{\theta}}(c|W_n)}{\mathbf{card}(\mathcal{P})} + \sum_{n \in \mathcal{N}} \frac{\ln\left(1 - p_{\boldsymbol{\theta}}(c|W_n)\right)}{\mathbf{card}(\mathcal{N})} - \lambda|\boldsymbol{\theta}|_1$$

Evaluation Metric

Experimental Procedures We evaluate our approach by comparing the quality of the extracted visual knowledge. For each visual category, we rank all relevant images by scoring functions in Equation (8) and (9) respectively, and then retrieve the top-k images with the highest scores as output. The precision of output images of category c are estimated by

$$Precision(c,k) = \frac{\#\text{relevant}(S_k, c)}{\mathbf{card}(S_k)}. \qquad (10)$$

#relevant(Sk,c) => #samples in Sk that are relevant to image category [c].

card(Sk) := 1,000

We estimate #relevant(Sk,c) by randomly sampling 100 images which are reviewed by human judges.

Results

seashore	113,937	68	82 (+14)
skirt	117,309	88	98 (+10)
sky	161,540	95	100 (+5)
suspension bridge	5,841	30	48 (+18)
table	150,542	84	90 (+6)
television	45,690	19	45 (+26)
truck	24,263	92	97 (+5)
vehicle	5,446	88	97 (+9)
wading bird	4,371	91	98 (+7)
warplane	32,506	95	99 (+4)
window	275,872	75	92 (+17)
Average	**81,696**	**72.95**	**81.43 (+8.48)**

#objects

Image Only

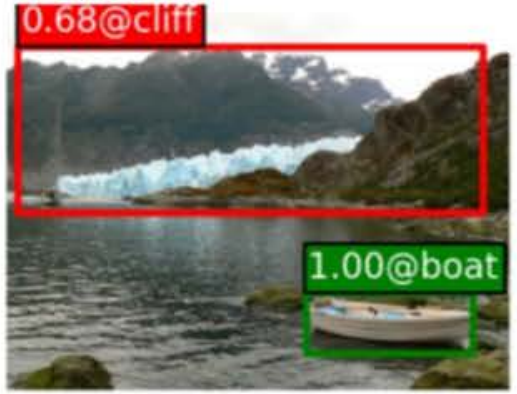
Proposed

Extraction Examples

Uni.@boat

tossa, mar

ukraine, chernobyl



south america, travel, park

thailand

Mul.@boat

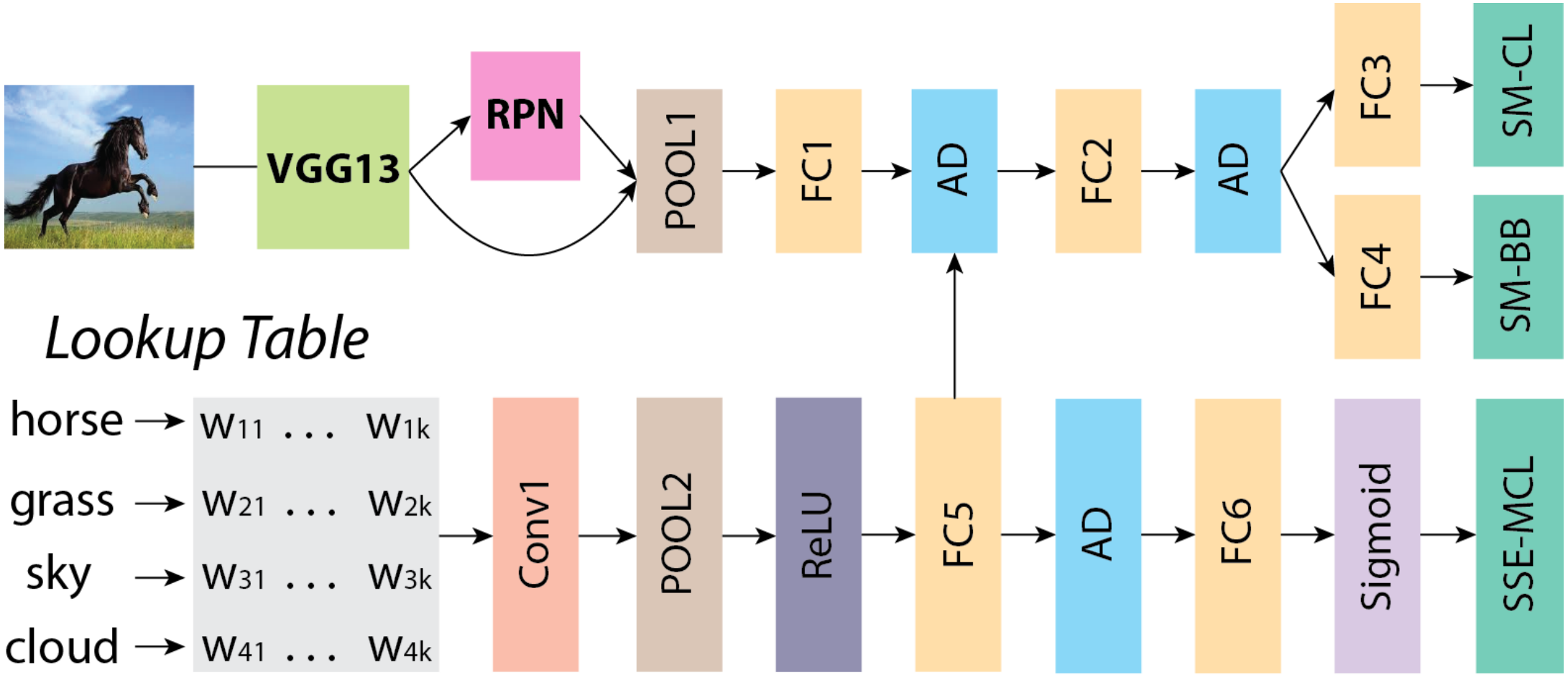
canoe

whitehall, maine, boat

boat

boat

MCNN


VGG13
RPN
POOL1
FC1
AD
FC2
AD
FC3
SM-CL
FC4
SM-BB
Lookup Table
horse → w11 ... w1k
grass → w21 ... w2k
sky → w31 ... w3k
cloud → w41 ... w4k
Conv1
POOL2
ReLU
FC5
AD
FC6
Sigmoid
SSE-MCL

MCNN vs. VTAKE

01

MCNN has two subnets: text CNN and image CNN.

02

MCNN is end-to-end trainable & predictable.

03

MCNN fuses at feature level, while VTAKE at prediction level.

Table 3: The comparison of precision by image category. For each image category, the blue color highlights the highest precision among the three comparison algorithms.

Category	Baseline	MM-LR	MCNN
beach	65	82	85
bed	91	95	95
boat	70	85	82
car	95	98	100
car mirror	17	15	14
civilian clothing	71	80	84
computer keyboard	25	30	32
fish	18	26	31
hand-held computer	91	89	91
helicopter	85	82	80
kitchen	79	90	88
lifeboat	49	58	61
microwave	79	82	90
musical instrument	12	9	13
people	96	99	100
pizza	69	59	66
riverbed	71	75	77
school bus	47	42	54
table	84	92	89
vehicle	75	83	87
Average	**64.45**	**68.55**	**70.70**

Table 5: The one-tail sign test

Comparison	#Wins	#Losses	p-value
MCNN **vs** Baseline	16	4	0.0059
MCNN **vs** MM-LR	14	5	0.0318

Evaluation Results

Illustrative Comparison

Table 6: MCNN vs MM-LR: example images of the highest rank improvement.

wiltshire, incident, scene

india, infantry, army

wheelchair, man, ottawa

trail rides, riders, trail

Table 7: MCNN vs Baseline: example false positive images.

delicious, olives, pasta

beans, vegan, side dish

hummus, tahini, dinner

potato, dinner, breakfast

Summary & Future Work

Summary

- We present a multimodal knowledge graph.
- We propose three different approaches to utilize multimodal information for knowledge expansion with multimodal link prediction.

Future Work

- Based on our current knowledge graph, with limited relations, extract richer knowledge by defining and predicting new multimodal links.
- Further improve link prediction precision by utilizing richer multimodal links.

References

- Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
- Zhu, Yuke, and Li Fei-Fei. "Building a large-scale multimodal knowledge base system for answering visual queries." *arXiv preprint arXiv:1507.05670* (2015).
- Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.
- Mitchell, T. M., Cohen, W. W., Hruschka Jr, E. R., Talukdar, P. P., Betteridge, J., Carlson, A., ... & Lao, N. (2015, January). Never Ending Learning. In *AAAI* (pp. 2302-2310).
- Chen, Xinlei, Abhinav Shrivastava, and Abhinav Gupta. "Neil: Extracting visual knowledge from web data." *Proceedings of the IEEE International Conference on Computer Vision*. 2013.