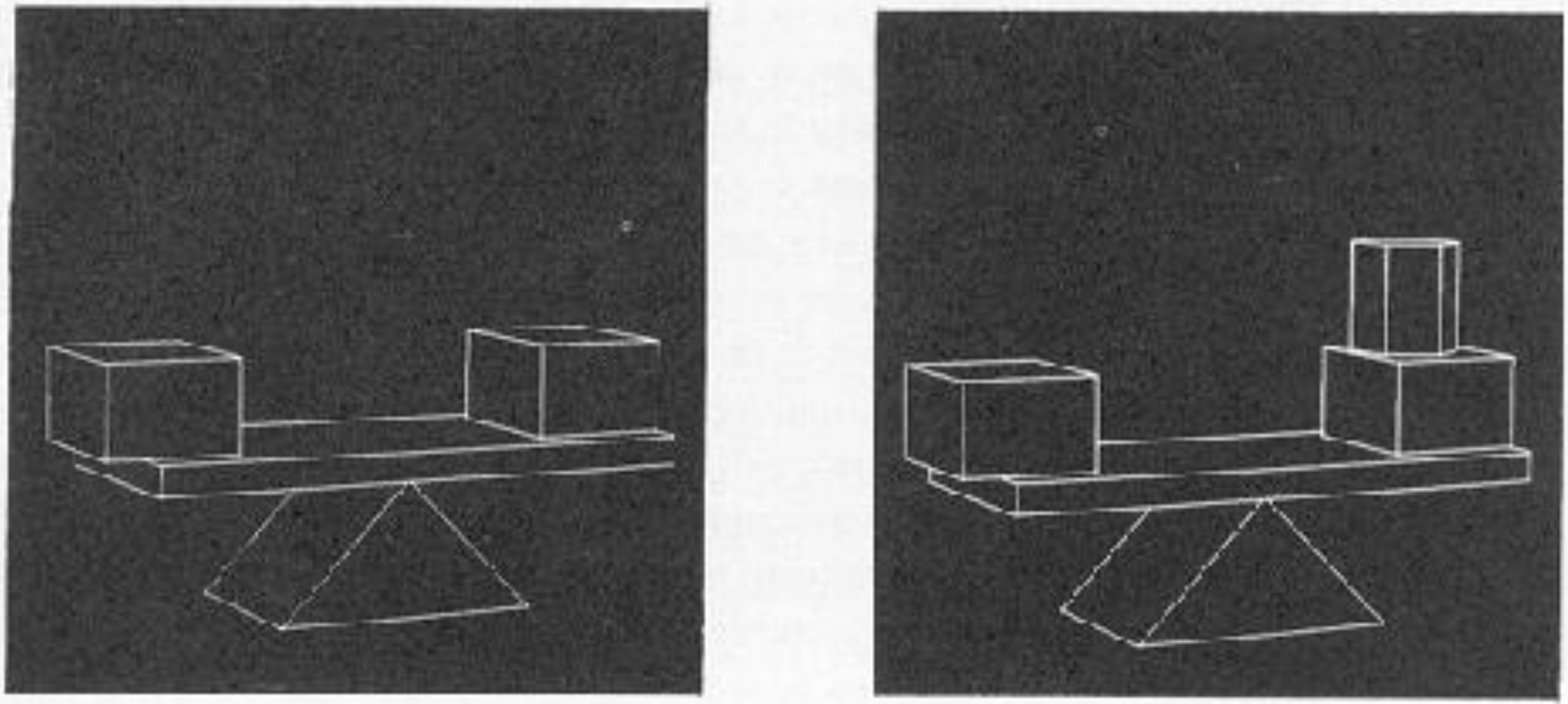
# Leveraging external knowledge in VQA

Anton van den Hengel,
Director, The Australian Centre for Visual Technologies
Professor, The University of Adelaide

Australian Centre for Visual Technologies
Innovation and education in visual information processing

# Vision used to be closer to AI

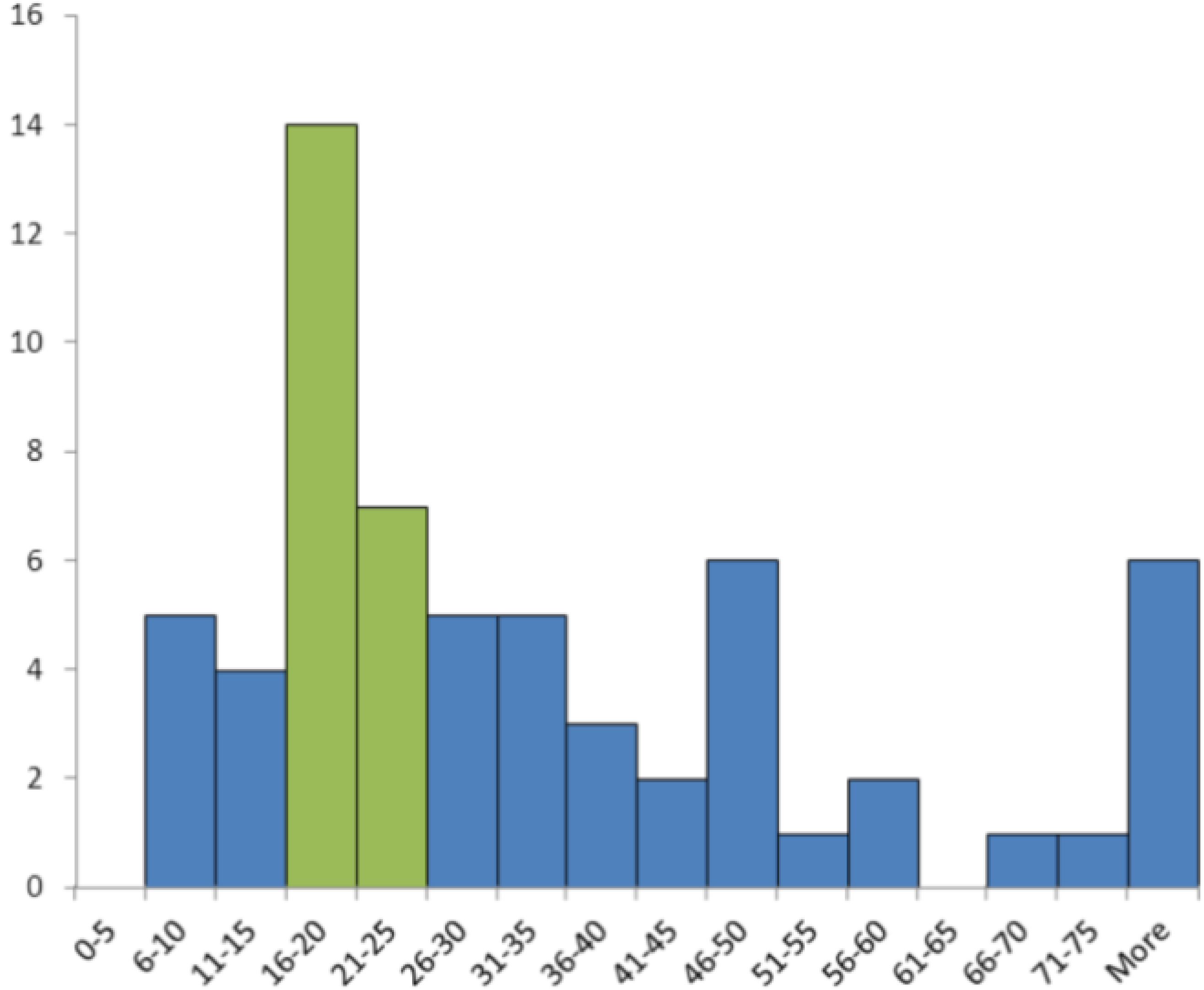


(e) See-saw. (f) With hexagonal prism.

# Vision used to be closer to AI

- The idea was to start simple and slowly add complexity
  - 1965, H. A. Simon: "machines will be capable, within twenty years, of doing any work a man can do."
  - 1967, Marvin Minsky: "Within a generation ... the problem of creating 'artificial intelligence' will substantially be solved."
  - 1970, Marvin Minsky: "In from three to eight years we will have a machine with the general intelligence of an average human being."
- It didn't work

# Expert predictions of years until AI




Armstrong, Stuart, and Kaj Sotala. "How we're predicting AI–or failing to."
*Beyond Artificial Intelligence*. Springer International Publishing, 2015. 11-29.

# Who was the most famous person to fly a plane like this?

Image: http://visualqa.csail.mit.edu/

# Who was the most famous person to fly a plane like this?

**Answer (http://visualqa.csail.mit.edu/):**

- **yes** (score: 12.88 = 3.87 [image] + 9.01 [word])
- **no** (score: 12.82 = 3.77 [image] + 9.05 [word])
- **pilot** (score: 8.83 = 4.95 [image] + 3.88 [word])

**Based on image only**: jet, plane, airport,

**Based on word only**: no, yes, filter,

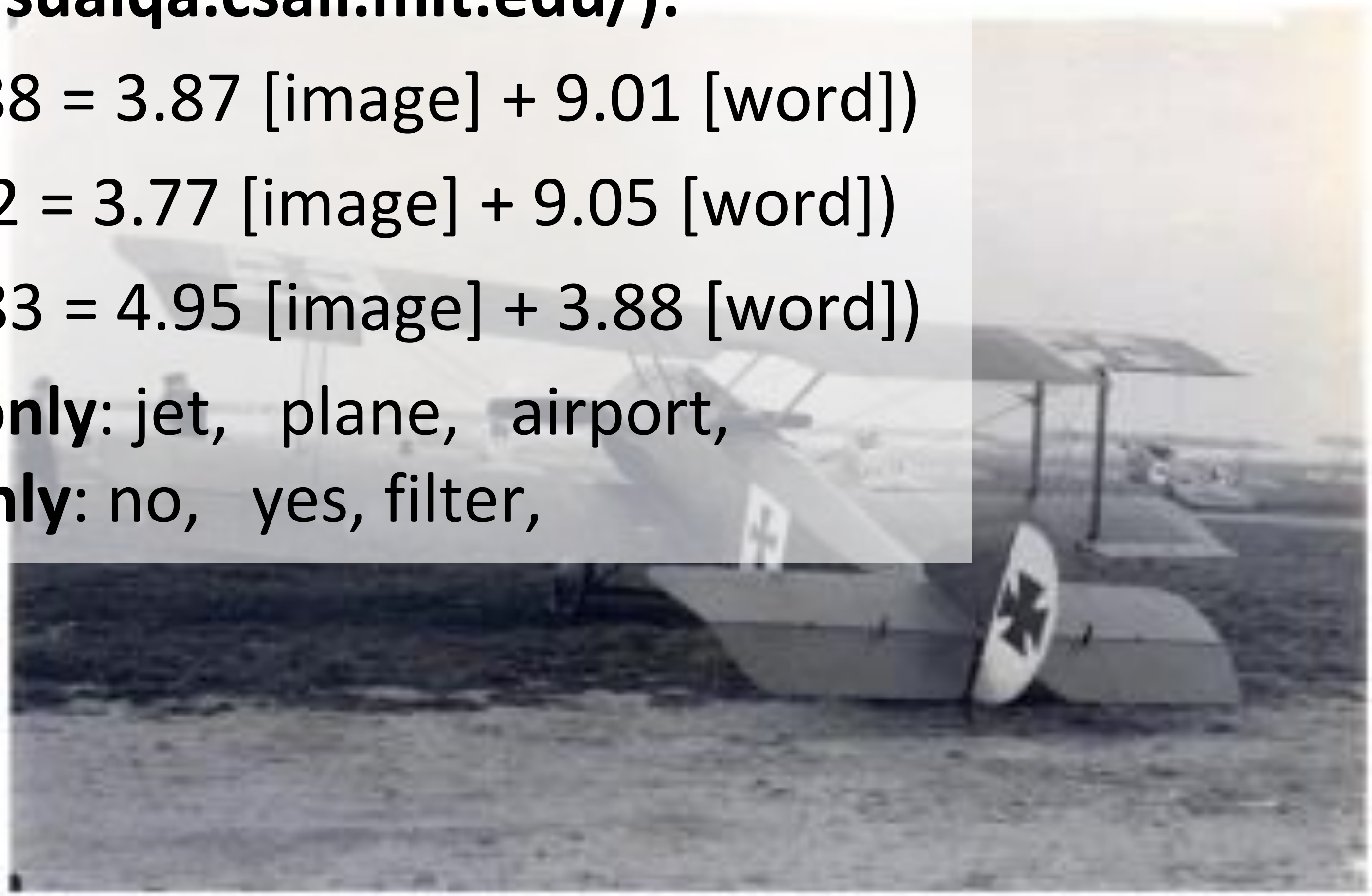


Image: https://commons.wikimedia.org/w/index.php?curid=32905


Did this player win the point?
Straße 12
Tel. 03628/6180-0
Fax 03628/6180-10
PAULANER
MÜNCHEN
PAUL

# Did this player win the point?


- **yes** (score: 8.66 = 3.47 [image] + 5.18 [word])
- **tennis court** (score: 8.17 = 6.66 [image] + 1.51 [word])
- **no** (score: 7.62 = 2.74 [image] + 4.88 [word])
- Based on image only: tennis court, net, tennis,
- Based on words only: before, yes, no,
- From http://visualqa.csail.mit.edu/

# Who’s winning?

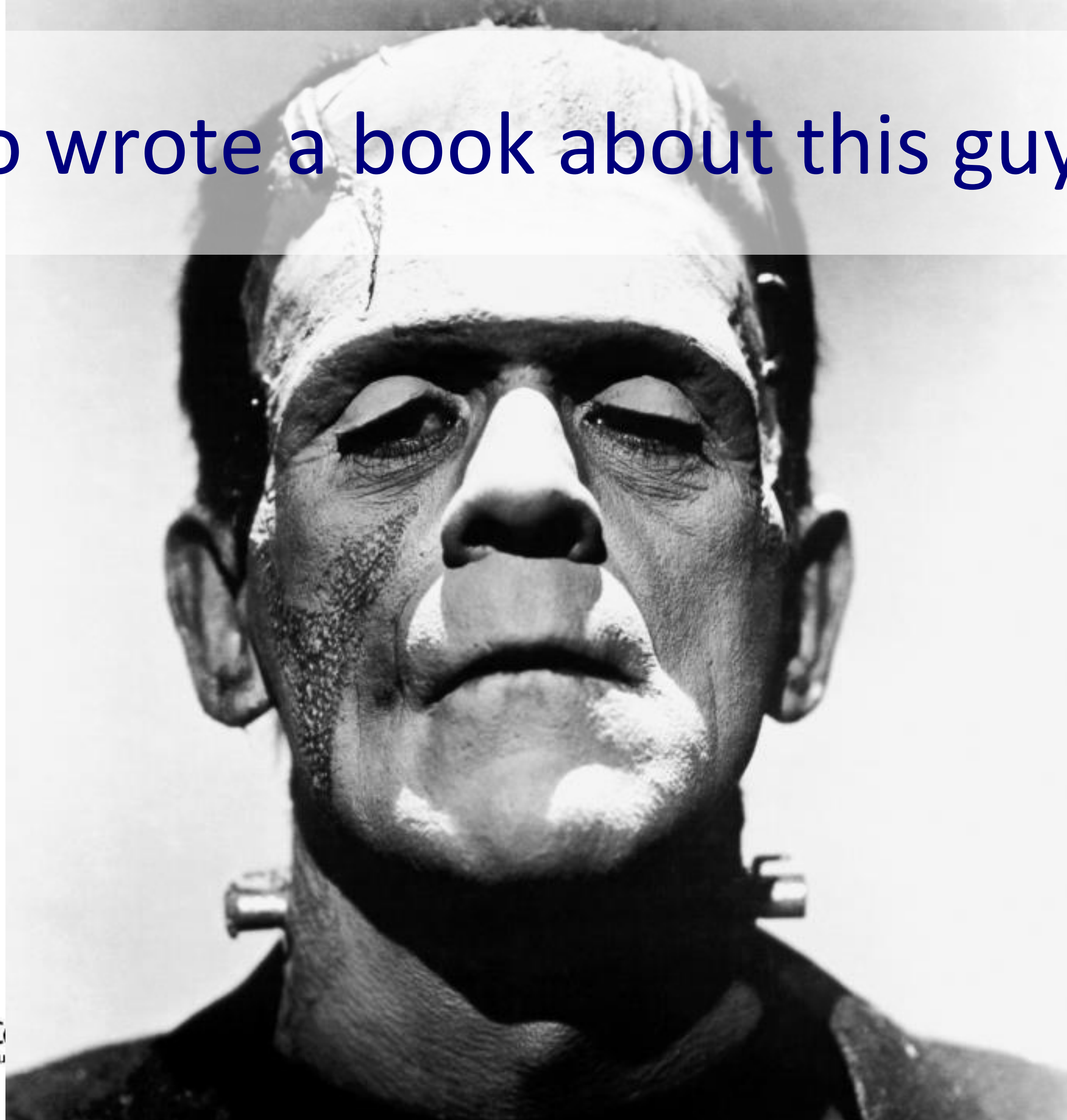
- Yes
- No
- Skiing


Image: http://visualqa.csail.mit.edu/

# NLP QA tackles harder questions

- Watson won Jeopardy
  - Q: William Wilkinson’s “An Account of the Principalities of Wallachia and Moldovia” inspired this author’s most famous novel
  - A: Bram Stoker

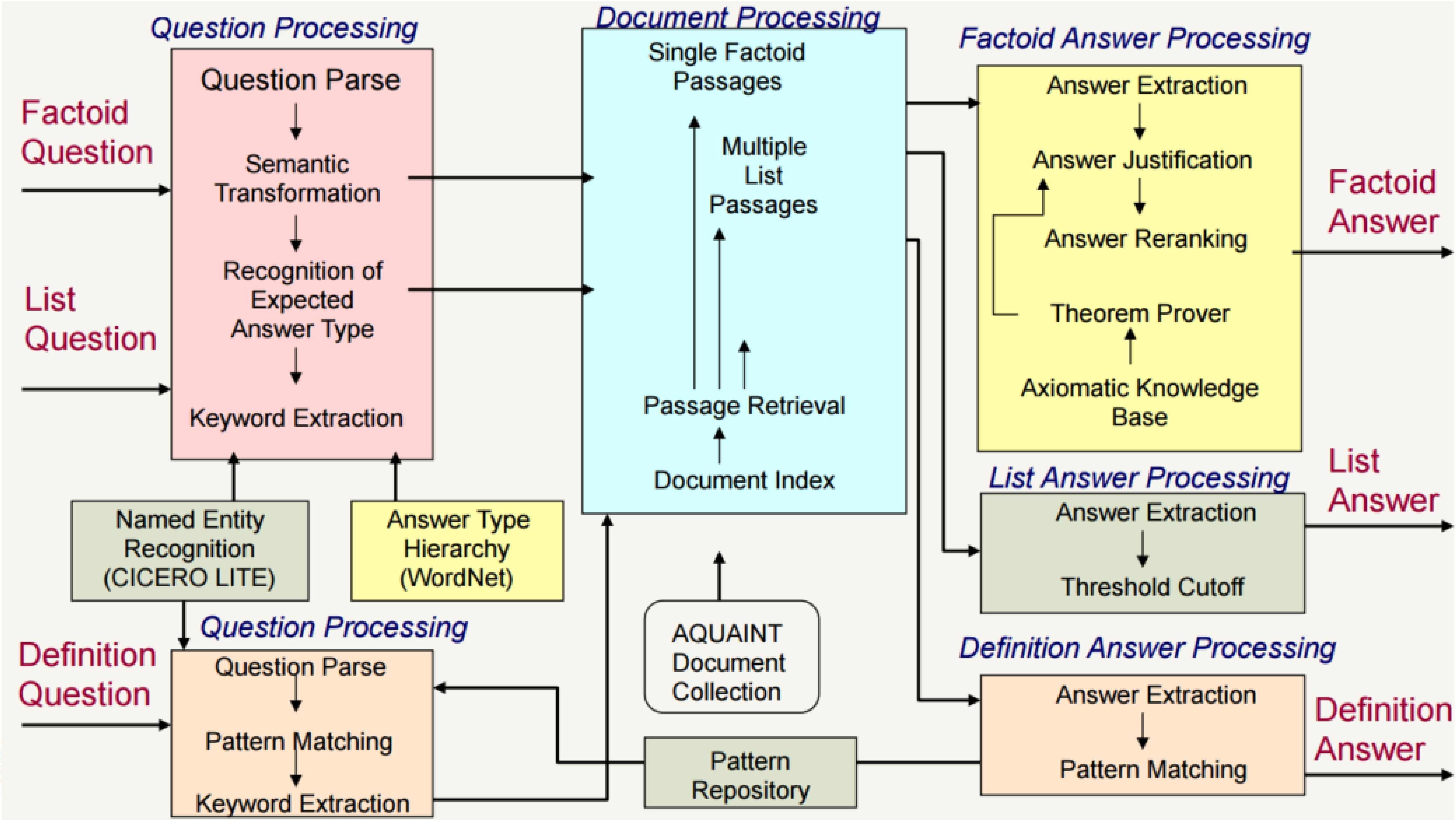
# Who wrote a book about this guy?



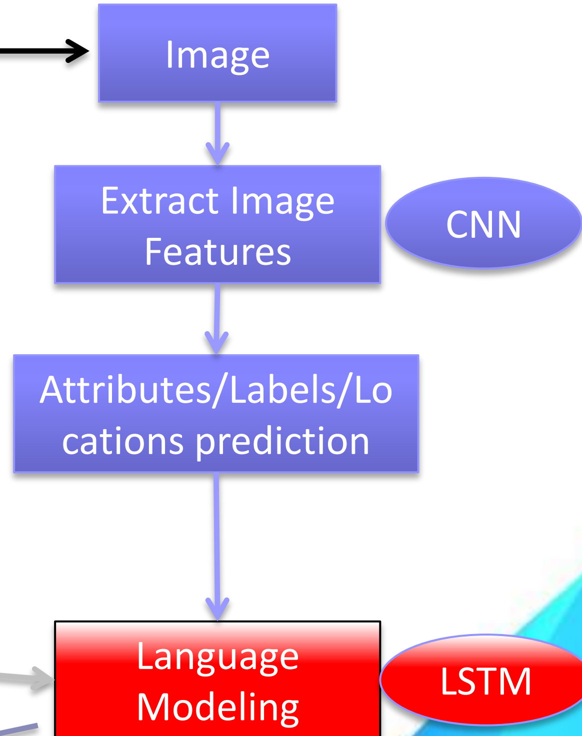
# NLP QA tackles harder questions

- TREC questions:
  - What was the monetary value of the Nobel Peace Prize in 1989?
  - What does the Peugeot company manufacture?
  - How much did Mercury spend on advertising in 1993?
  - What is the name of the managing director of Apricot Computer?
  - Why did David Koresh ask the FBI for a word processor?
- Average performance is about 70%

# NLP QA is complex



# Attributes for Visual Question Answering

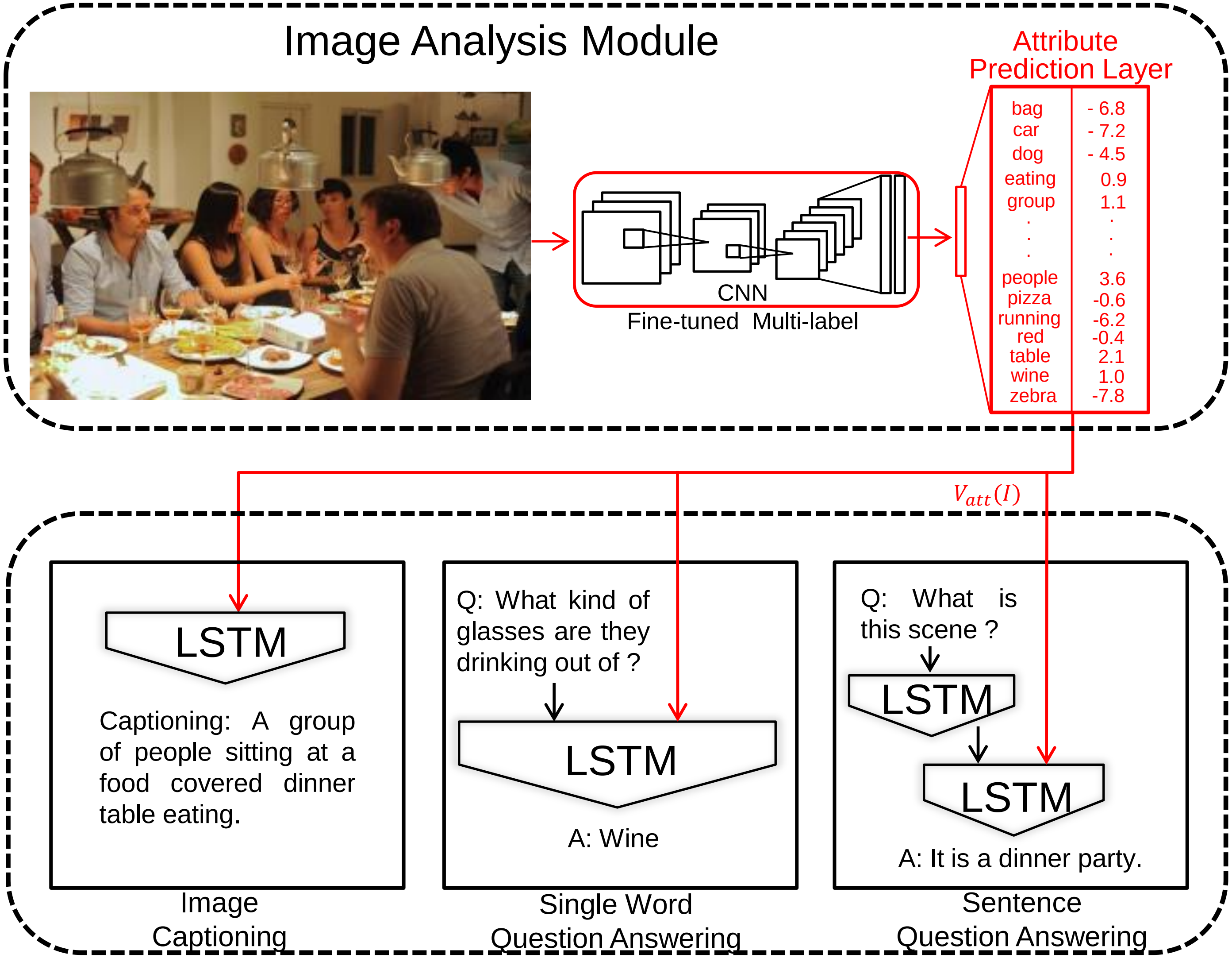



Q: What kind of glasses are they drinking out of ?

A: Wine

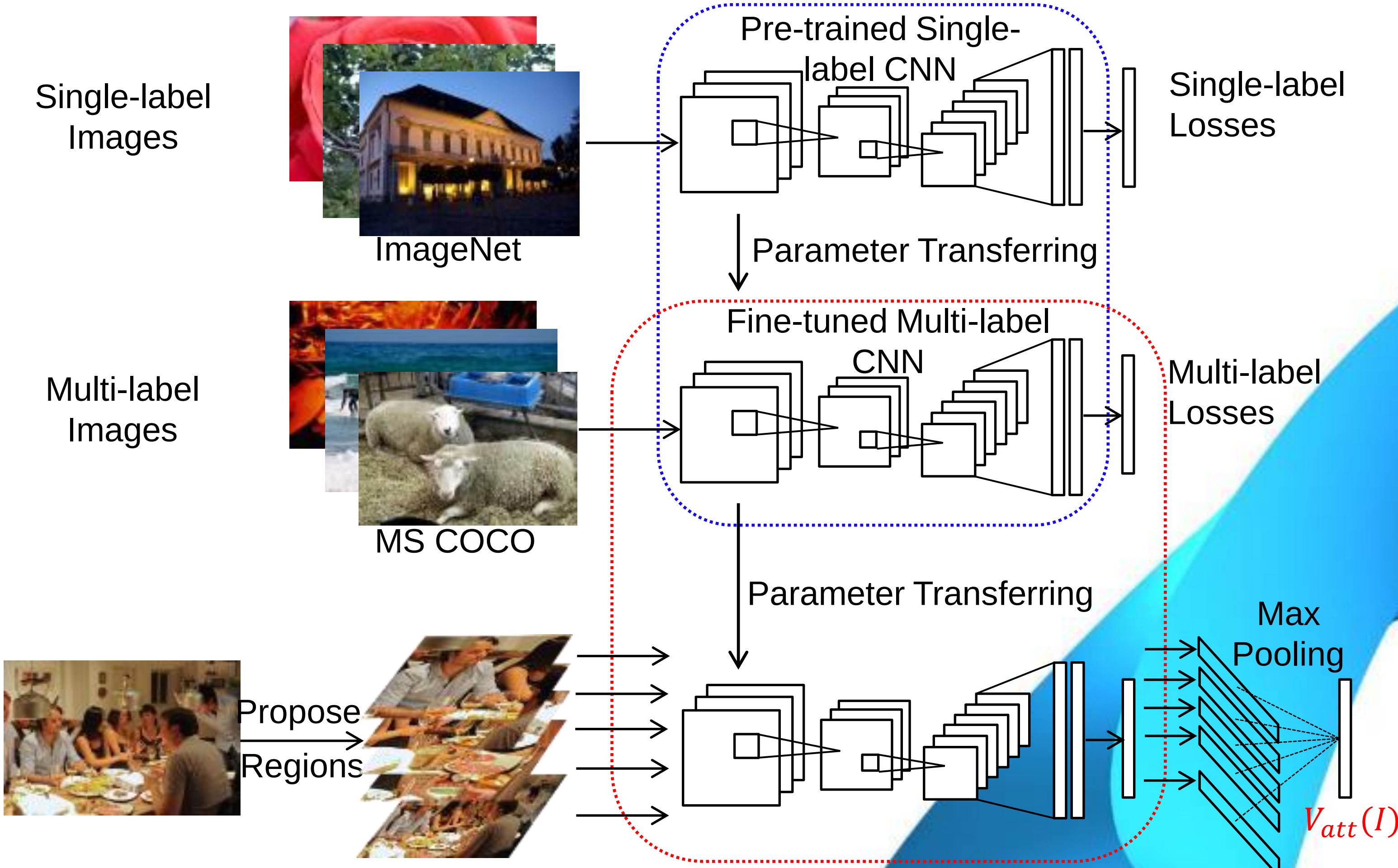
Australian Centre for Visual Technologies
Innovation and education in visual information processing

**What Value Do Explicit High Level Concepts Have in Vision to Language Problems?**
Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel, CVPR'16




**What Value Do Explicit High Level Concepts Have in Vision to Language Problems?**
Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel, CVPR'16

# Visual Concept Prediction CNN




Australian Centre for Visual Technologies
Innovation and education in visual information processing

**What Value Do Explicit High Level Concepts Have in Vision to Language Problems?**
Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel, CVPR'16

# Performance

| State-of-art | B-1 | B-2 | B-3 | B-4 | M | C | $\mathcal{P}$ |
|---|---|---|---|---|---|---|---|
| NeuralTalk [22] | 0.63 | 0.45 | 0.32 | 0.23 | 0.20 | 0.66 | - |
| Mind's Eye [6] | - | - | - | 0.19 | 0.20 | - | 11.60 |
| NIC [50] | - | - | - | 0.28 | 0.24 | 0.86 | - |
| LRCN [10] | 0.67 | 0.49 | 0.35 | 0.25 | - | - | - |
| Mao et al.[36] | 0.67 | 0.49 | 0.34 | 0.24 | - | - | 13.60 |
| Jia et al.[18] | 0.67 | 0.49 | 0.36 | 0.26 | 0.23 | 0.81 | - |
| MSR [11] | - | - | - | 0.26 | 0.24 | - | 18.10 |
| Xu et al.[53] | 0.72 | 0.50 | 0.36 | 0.25 | 0.23 | - | - |
| Jin et al.[21] | 0.70 | 0.52 | 0.38 | 0.28 | 0.24 | 0.84 | - |
| **Baseline-*CNN(I)*** | | | | | | | |
| VNet+LSTM | 0.61 | 0.42 | 0.28 | 0.19 | 0.19 | 0.56 | 13.58 |
| VNet-PCA+LSTM | 0.62 | 0.43 | 0.29 | 0.19 | 0.20 | 0.60 | 13.02 |
| GNet+LSTM | 0.60 | 0.40 | 0.26 | 0.17 | 0.19 | 0.55 | 14.01 |
| VNet+ft+LSTM | 0.68 | 0.50 | 0.37 | 0.25 | 0.22 | 0.73 | 13.29 |
| **Ours-$V_{att}(I)$** | | | | | | | |
| Att-GT+LSTM‡ | 0.80 | 0.64 | 0.50 | 0.40 | 0.28 | 1.07 | 9.60 |
| Att-SVM+LSTM | 0.69 | 0.52 | 0.38 | 0.28 | 0.23 | 0.82 | 12.62 |
| Att-CNN+LSTM | **0.74** | **0.56** | **0.42** | **0.31** | **0.26** | **0.94** | **10.49** |

Table 1. BLEU-1,2,3,4, METEOR, CIDEr and $\mathcal{PPL}$ metrics compared with other state-of-the-art methods and our baseline on MS COCO dataset. ‡ indicates ground truth attributes labels are used, which (in gray) will not participate in rankings.

# External information?

- Now operating at a higher semantic level
  - Use it to add explicit external information
- Explicit storage means less to store implicitly
  - It’s not feasible to store all relevant knowledge implicitly
- And why train a NN to do something it’s not good at

# Use a Knowledge Base

- Scraped or hand crafted
- RDF tuples
  - <Obama, President, United States of America>
  - But not <everything, gravity, everything>
- In a DBMS
  - Which does inference
  - Admits queries in SPARQL (which is like SQL)

# Reasoning in VQA

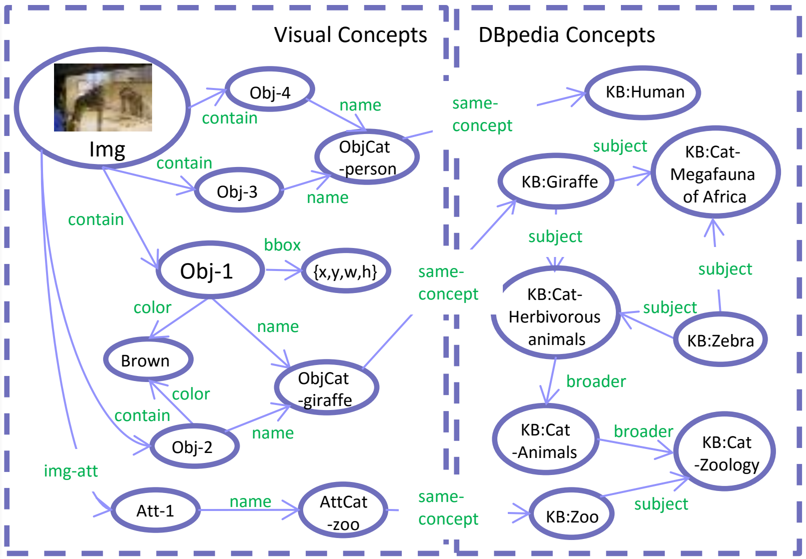
**Visual Concepts**

*Object*: Person, Giraffe
*Attributes*: glass, house, room, standing, walking, wall, zoo
*Scenes*: museum, indoor

DCNN models

**Input Image**

**Linked Knowledge Graph**




**Input Question**

What are the common properties between the animal in this image and the zebra?

QUEPY

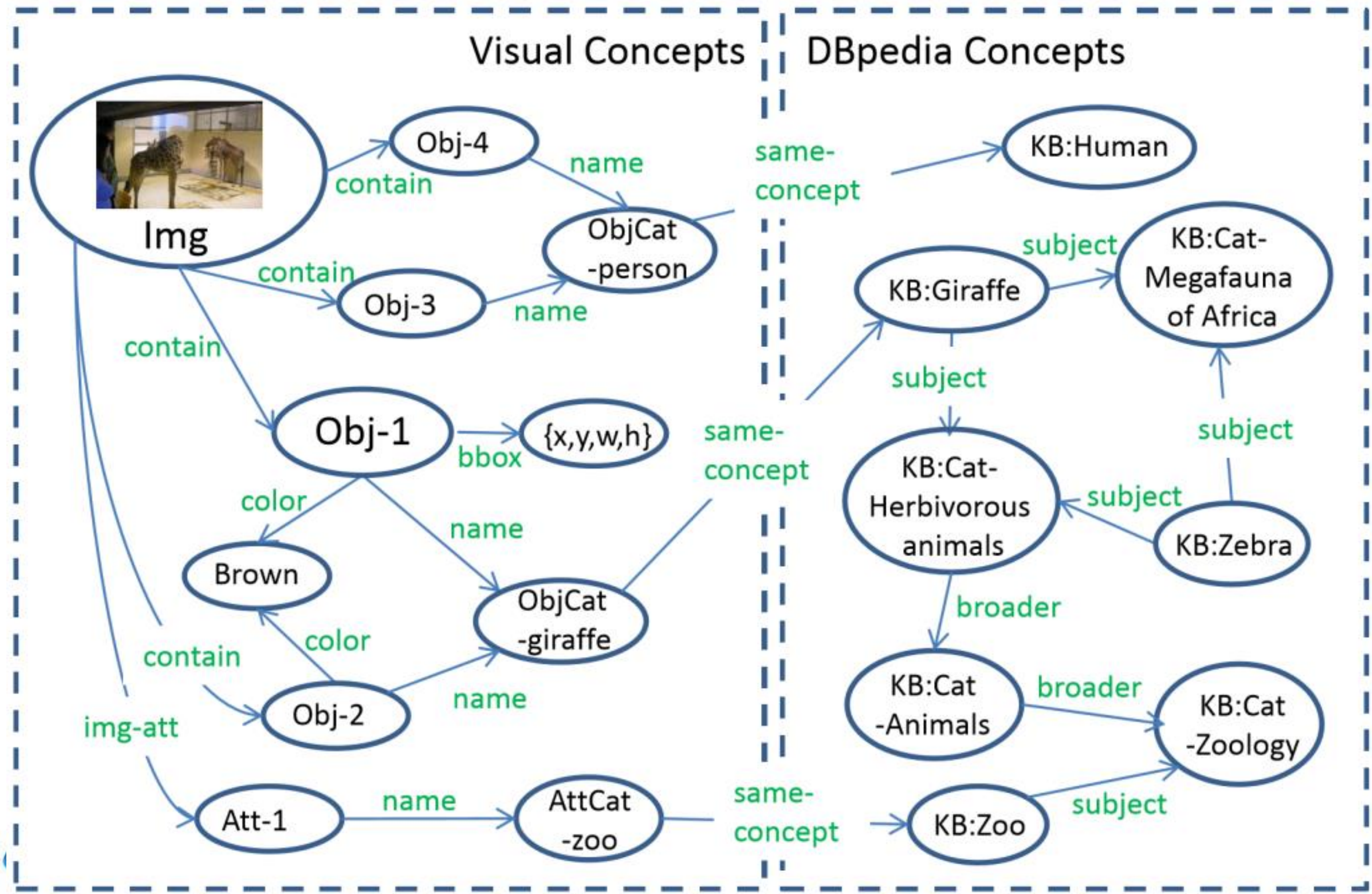
**Database Queries**

```
?x:((KB:Giraffe, subject/?broader, ?x) AND
    (KB:Zebra,   subject/?broader, ?x))
```

**Answer and Reason**

A: Herbivorous animals, Animals, Magafauna of Africa
R: Giraffe → Herbivorous animals → animals
Zebra → Magafauna of Africa

Australian Centre for Visual Technologies
Innovation and education in visual information processing


Visual Concepts
DBpedia Concepts
Img
contain
Obj-4
name
ObjCat
-person
same-
concept
KB:Human
contain
Obj-3
name
contain
Obj-1
bbox
{x,y,w,h}
color
name
Brown
ObjCat
-giraffe
color
contain
Obj-2
name
img-att
Att-1
name
AttCat
-zoo
same-
concept
same-
concept
KB:Giraffe
subject
KB:Cat-
Megafauna
of Africa
subject
subject
KB:Cat-
Herbivorous
animals
subject
KB:Zebra
broader
KB:Cat
-Animals
broader
KB:Cat
-Zoology
KB:Zoo
subject

# Traversing the Knowledge Base

**Q**: List close relatives of the animal.
**A**: Donkey, horse, mule, asinus, hinny

**Q1**: Which object in this image is most related to entertainment?
**A1**: TV.
**R1**: Television → Performing Arts → Entertainment.

**Q4**: How many road vehicles in this image?
**A4**: Three.
**R4**: There are two trucks and one car.

**Q2**: Is the image related to sleep?
**A2**: Yes.
**R2**: Attribute-bedroom → sleep; Object-bed → sleep.

**Q5**: Tell me the ingredient of the food in the image.
**A5**: Meat, bread, vegetable, sauce, cheese, spread.

Q: List common properties of these two images.

A: Background: snow;

Scene: ski slope, ski resort, mountain snowy

Object concepts: racing, winter sports, outdoor recreation;

Australian Centre for Visual Technologies
Innovation and education in visual information processing

Q: List common properties of these two images.

A: Scene concepts: transport infrastructure;

Australian Centre for Visual Technologies
Innovation and education in visual information processing