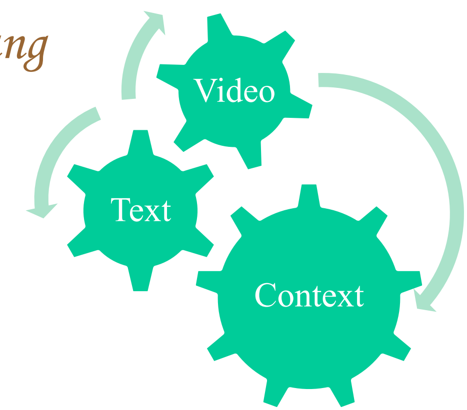
Zero-example Video Search

Chong-Wah Ngo

City University of Hong Kong

James Yi-Jie Lu, Maaike de Boer and Hao Zhang



VIREO Video Retrieval Group

Agenda

- Problem definition
- Quick overview of problems, methods, results
- Baseline to be shared

The talk will be delivered in ICMR 2017 as an 1-hour tutorial, along with few other talks: 1) Semantic Indexing, 2) Ad-hoc search, 2) Instance Search, 3) Multimedia Event Detection, 4) Video-to-Text, *with a number of baselines to be released*. **The series of talks is to encourage more TRECVid participation,** by providing tools/open sources to rapidly setup a decent system.

Problem

Given a textual query, find the relevant video clips from large video collection.

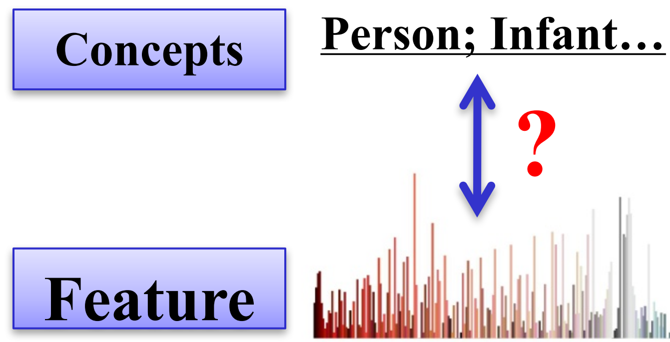
Query: Find shots of something burning with flames visible

explosion?

smoke?

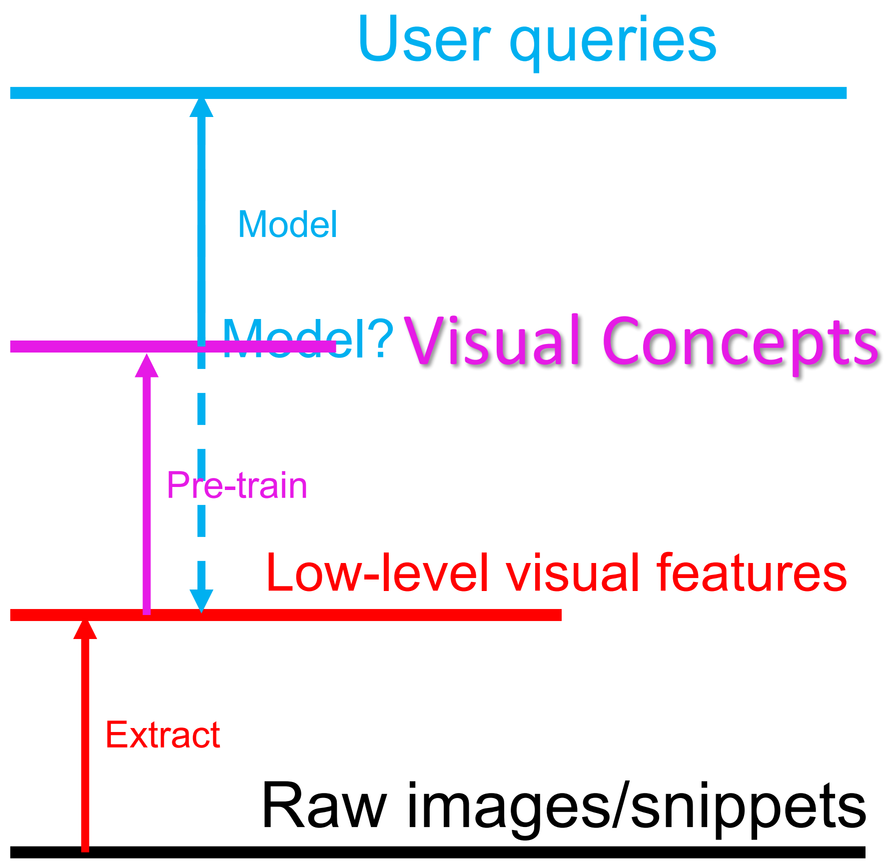
Semantic Gap

- Computers are not as *smart* as humans…
  - Between computable low-level features and high-level semantics

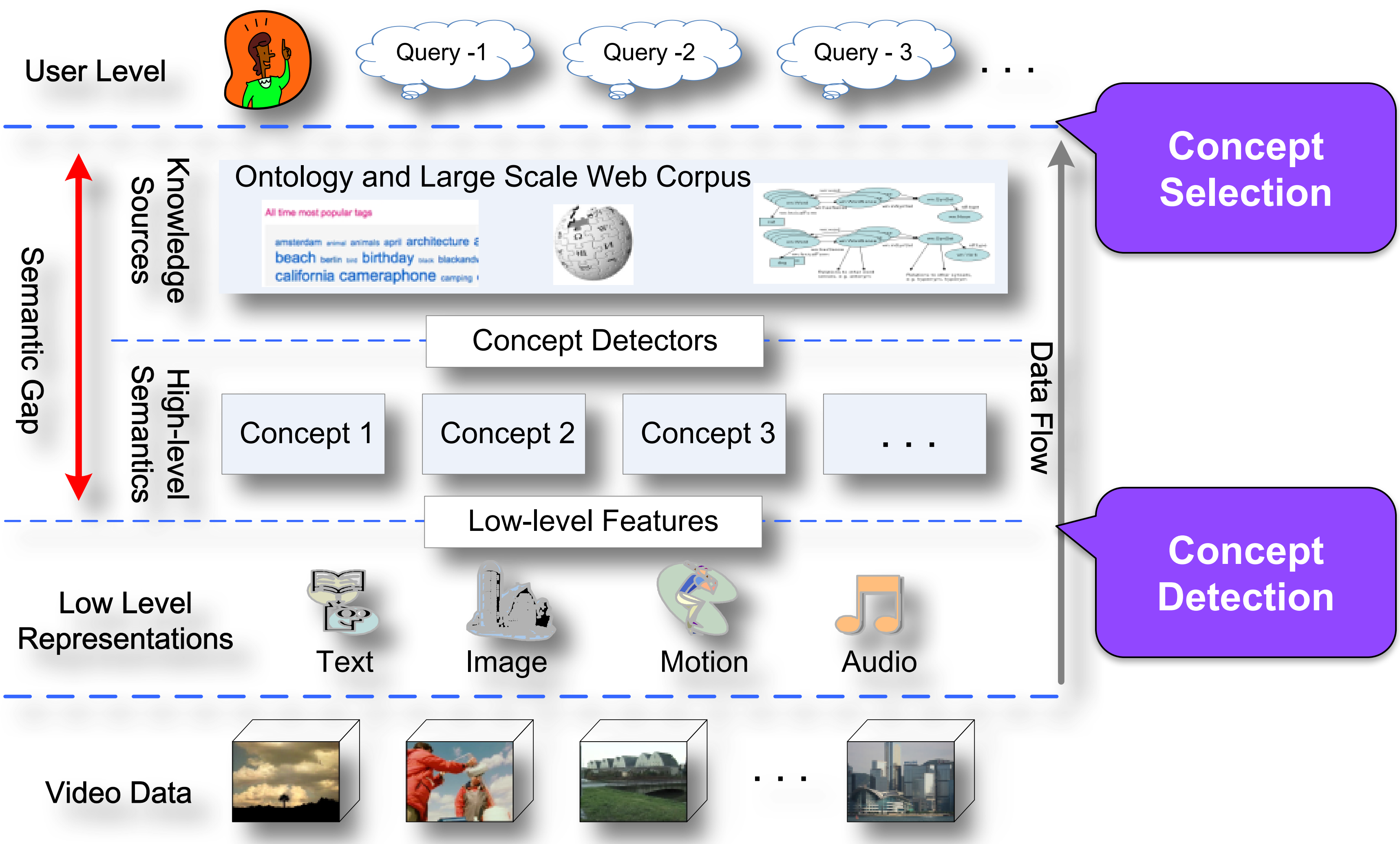


- Scalability (speed)
  - Thousands of semantic categories
  - Billions of images/videos on the web

Semantic Gap



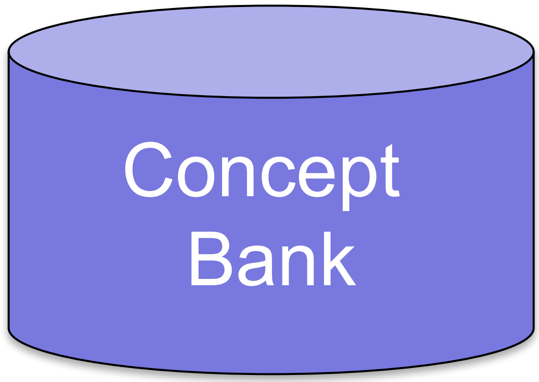
Concept-based video search



X. Y. Wei, C. W. Ngo, and Y. G. Jiang. Selection of concept detectors for video search by ontology enrich semantic space. IEEE. Trans on Multimedia, 2008.

System Pipeline

- Database indexing

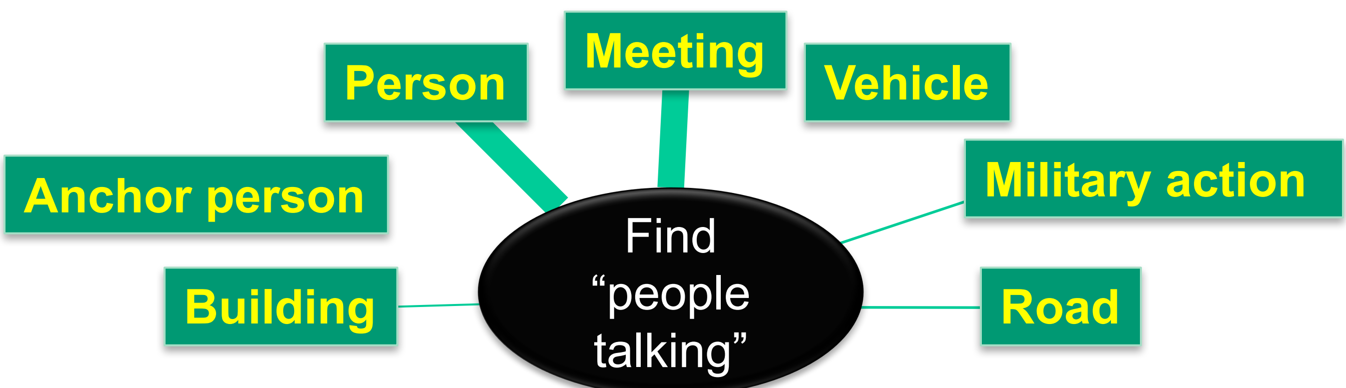


Anchor person

Person,
Meeting, …

Military action,
Vehicle, Road,
Building…

- Online search



TREC Video Retrieval Evaluation (TREVID)

- The idea of zero-example video search starts from TRECVID

- TRECVID
  - Sponsored by NIST, USA
  - Provide benchmark and evaluation annually for system evaluation

- TRECVID dataset
  - Broadcast News (NTV, CCTV, MSNBC, CNN…) *2003-2006*
  - Documentary Videos from the Netherlands. *2007-2008*
  - Web Videos*: 2010 and beyond*

TREC Video Retrieval Evaluation (TREVID)

	TV	Data domain	Devel. set (keyframe #)	Test set (keyframe #)	# of evaluated queries
Ad-hoc	05	Broadcast news	80h (43,873)	80h (45,765)	24
	06	Broadcast news	--	80h (79,484)	24
	07	Documentary	50h (21,532)	50h (18,142)	24
	08	Documentary	--	100h (35,766)	48
	16	Internet archive (AVS)	1,400h	600h (335,944)	30
Event	13-16	User-generated (MED)	416h	849h (MED14Test) 7,580h (full) 1,244h (sub)	20 (predefined) + 10 (ad-hoc)

Ad-hoc Query

Event and/or Person-Things

Query

Find shots with a person walking or riding a bicycle

Selected Concepts

Wordnet	Google search	Flickr context	transfer learning
Person, Bicycle, Individual,Walking, Running, Backpacker…	Walking, Walking_Running, Person, Bicycle, Car, Group, Motorcycle...	Bicycles, Person, Walking, Walking_Running, Horse ,Dog…	Bicycles, Person, Walking, Walking_Running, Daytime_outdoor…

Return top-1,000 shots

Wordnet

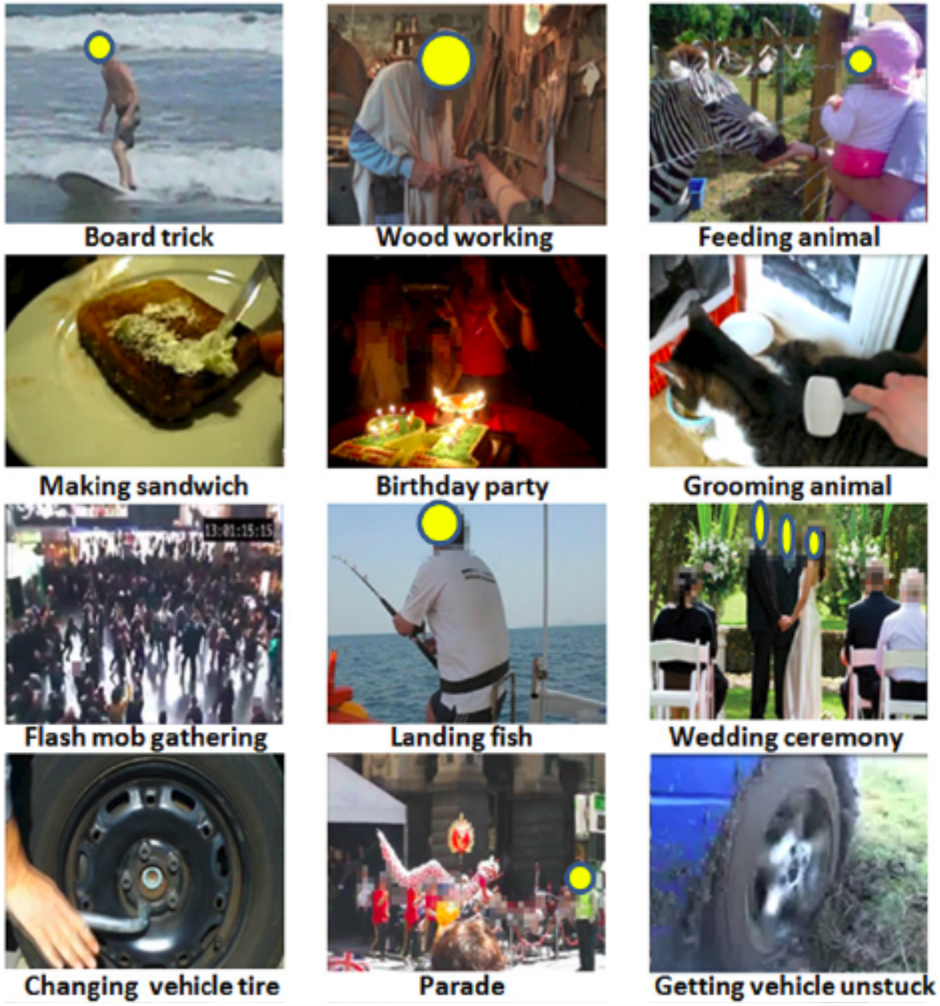
Google search

Flickr context

transfer learning

Multimedia Event Query

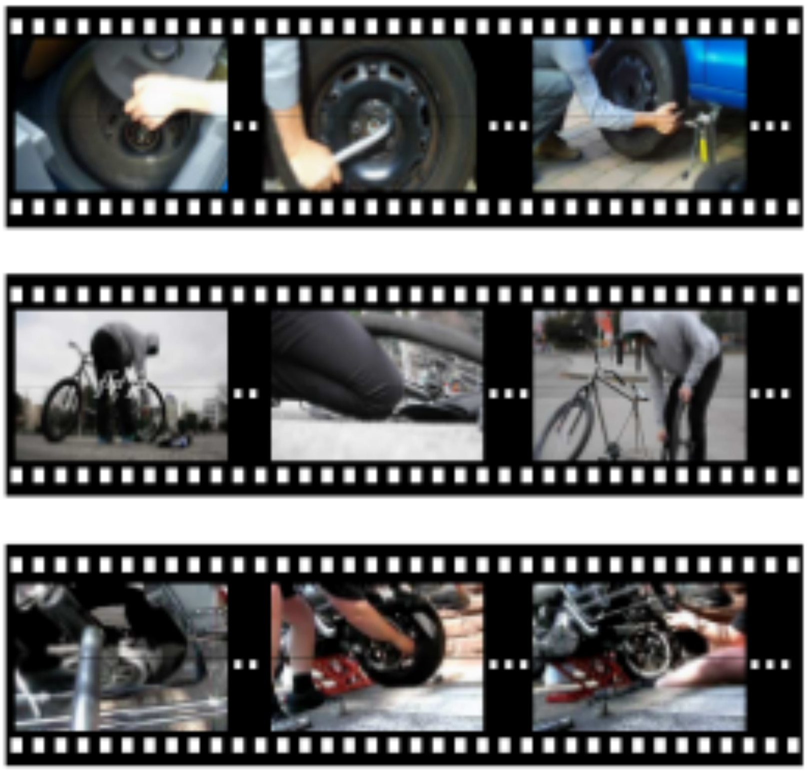
Complex and **generic** events occurring at a specific place and time involving people interacting with other people / objects.



Board trick | Wood working | Feeding animal
Making sandwich | Birthday party | Grooming animal
Flash mob gathering | Landing fish | Wedding ceremony
Changing vehicle tire | Parade | Getting vehicle unstuck

High Variability

Changing a vehicle tire



Marriage proposal

Changing A Vehicle Tire

- Description

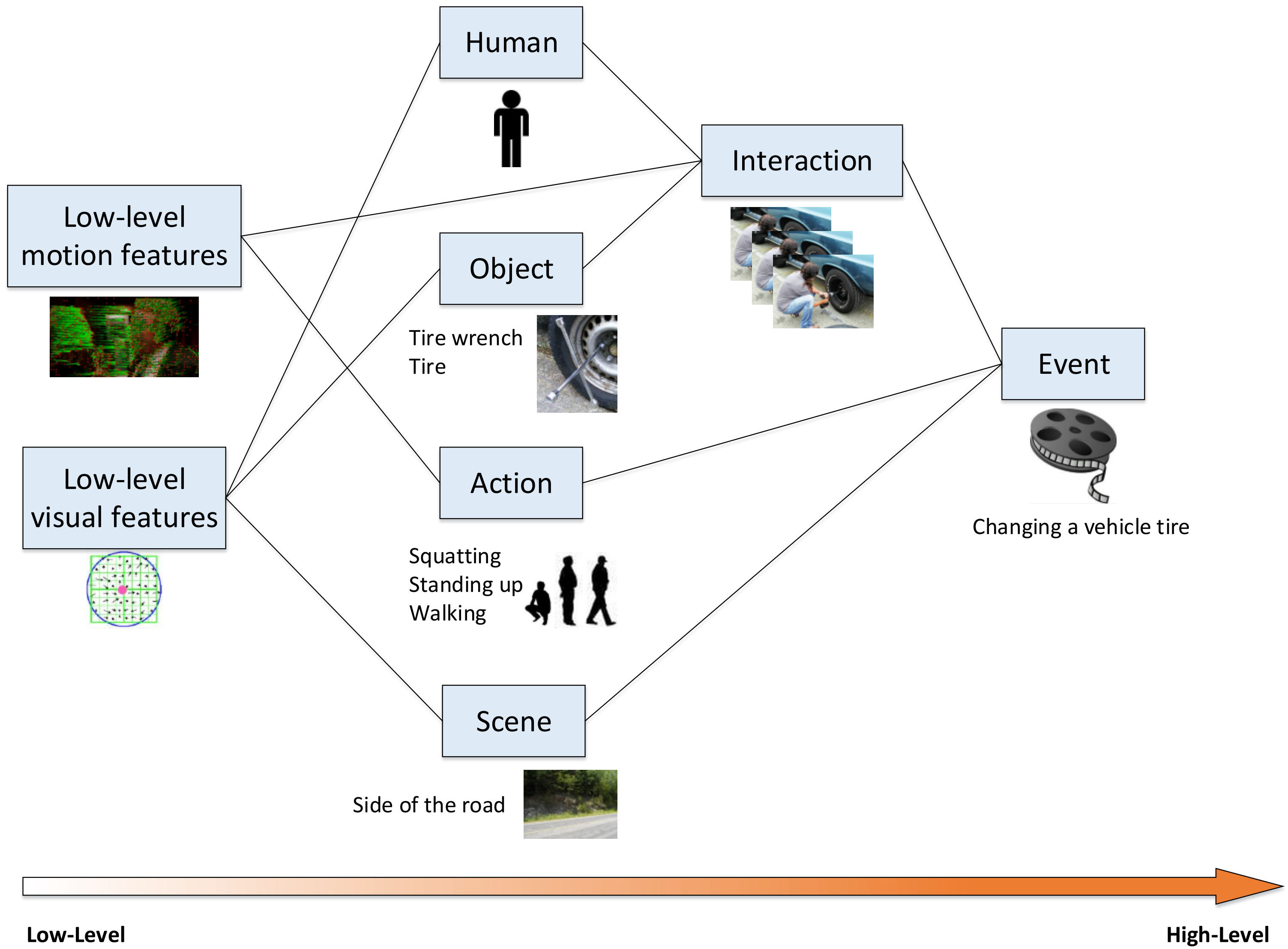
  One or more people work to replace a tire on a vehicle

- Explication

  The process for replacing a tire includes removing the existing tire and installing the new tire onto the wheel of the vehicle ......

- Evidential description
  - *Scene:* garage, outdoors, street, parking lot
  - *Objects/people:* tire, lug wrench, hubcap, vehicle, tire jack
  - *Activities:* removing hubcap, turning lug wrench, unscrewing bolts
  - *Audio:* sounds of tools being used; street/traffic noise

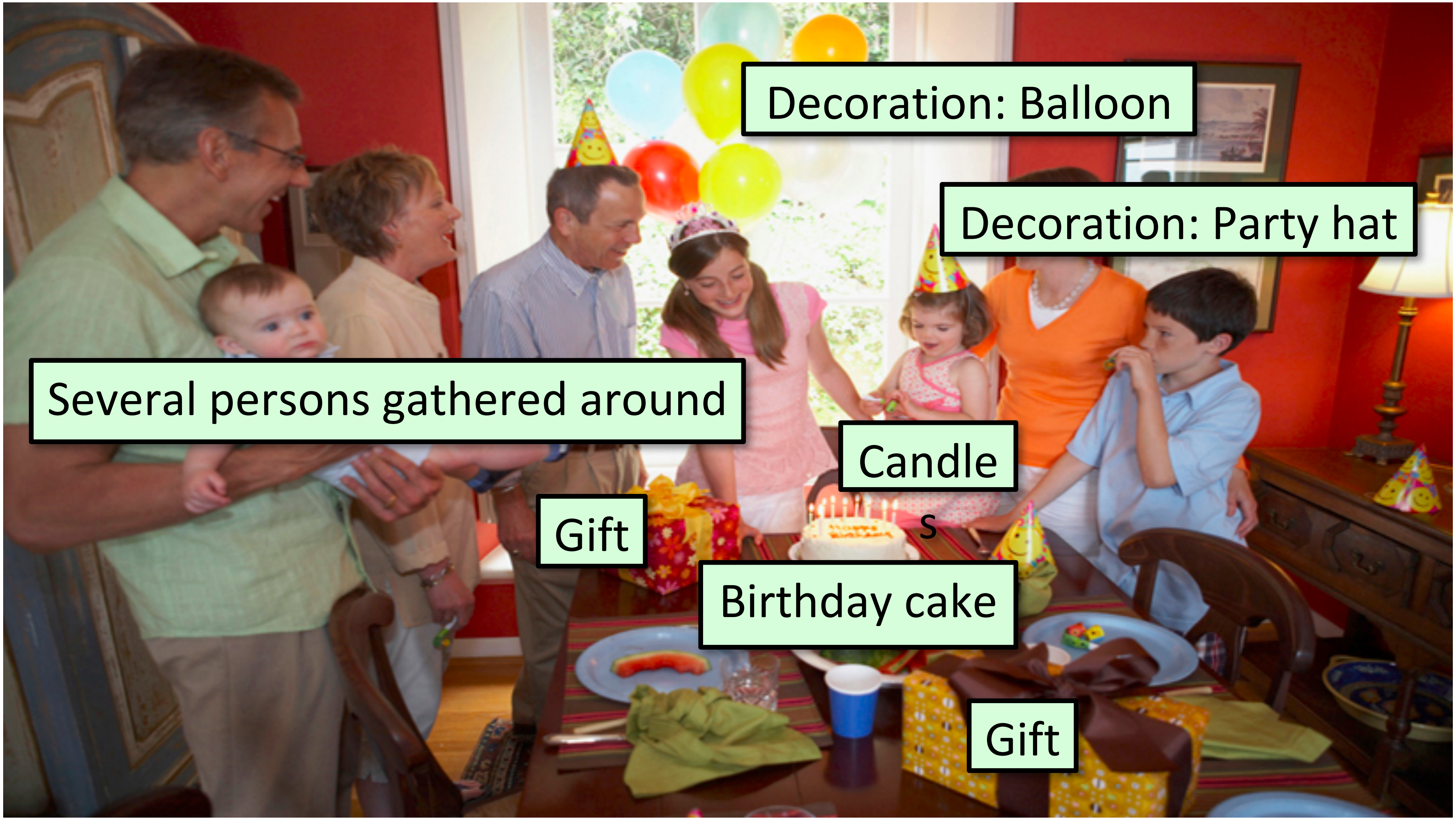
Changing A Vehicle Tire



Issue-A: Concept Bank

1. How to determine the list of concepts to index?
2. How large should the concept bank be?
3. What are the expected classification accuracy?

How many concepts required for "Happy Birthday"?



Issue-A: Concept Bank

General guidelines

- A mixed of general and specific concepts
- A mixed of concepts with different complexities (from event-oriented concepts to objects and scenes)
- General versus specific
  - ✓General concepts are more important than specific concepts
- Quantity versus quality
  - ✓Include more concepts than to improve the quality of individual concepts
- *Summary*: A large and diverse concept bank

Issue-A: Concept Bank

Higher level

Mid level

Lower level

FCVID 239

activities, events

Sports 487

activities, events

SIN 346 (AVS)

objects, activities

ImageNet 1000

objects

Places 205

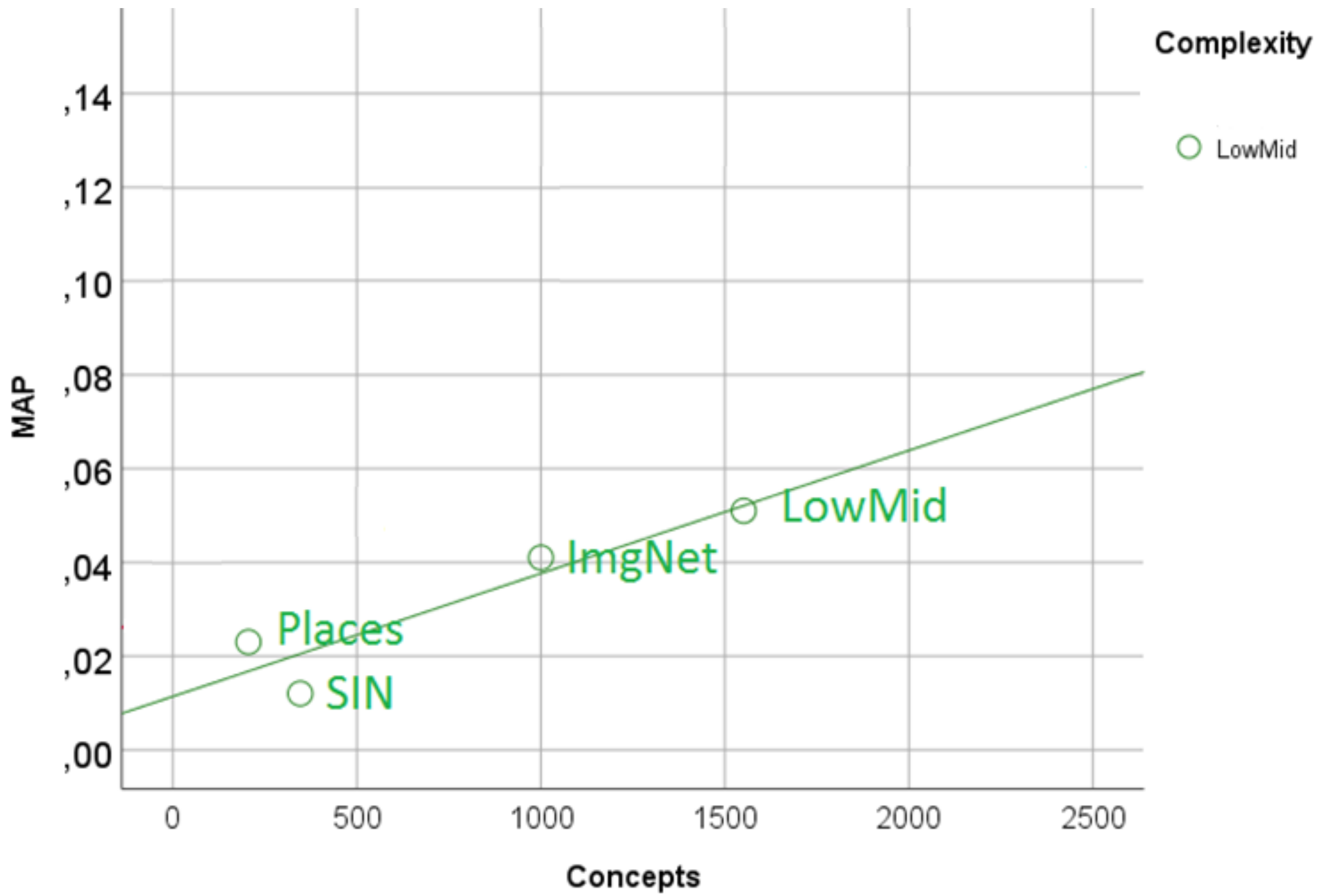
scenes

RC 497 (MED)

mixed

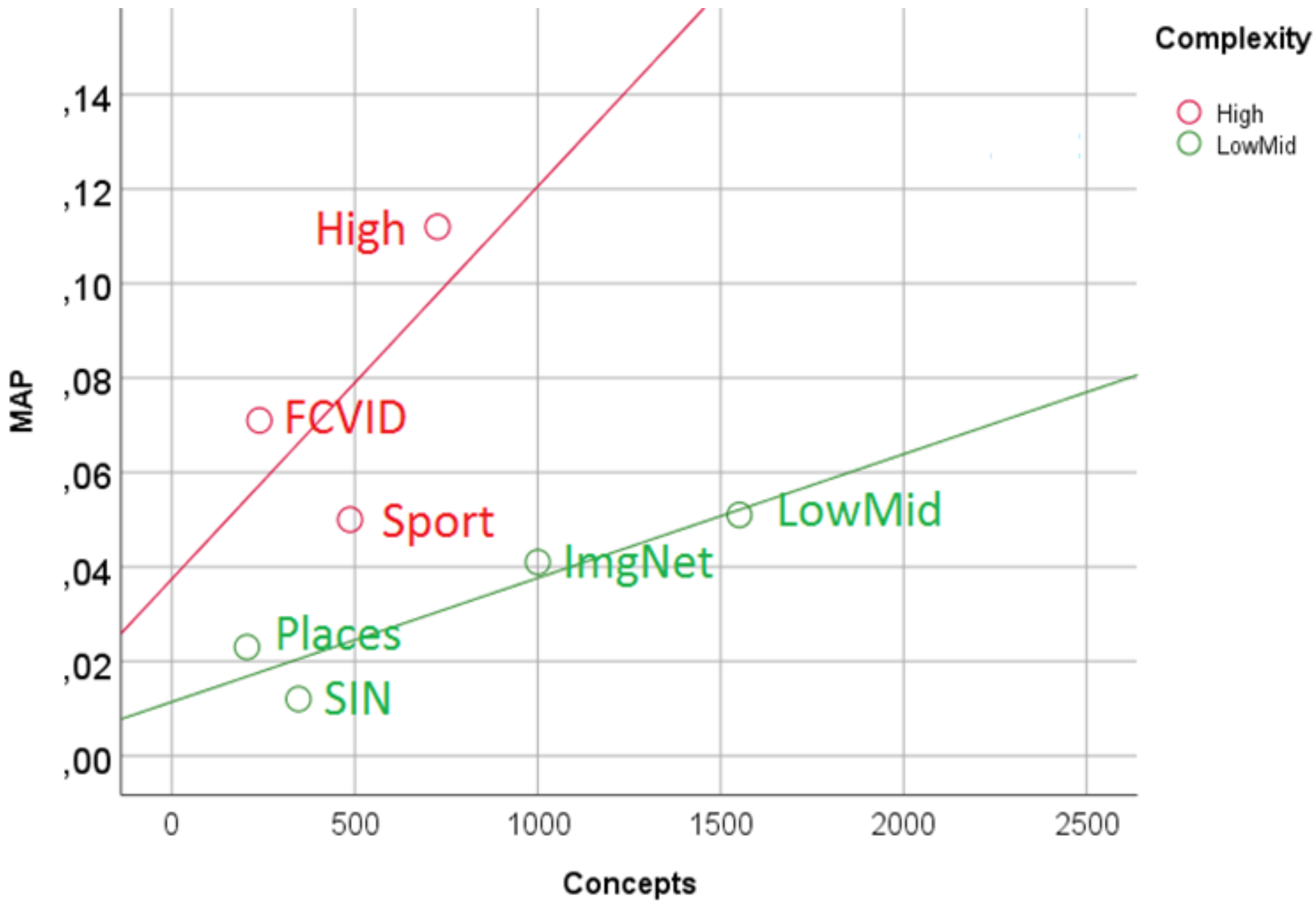
Number of concepts vs. MAP

MED14Test (oracle run of top-k concepts)



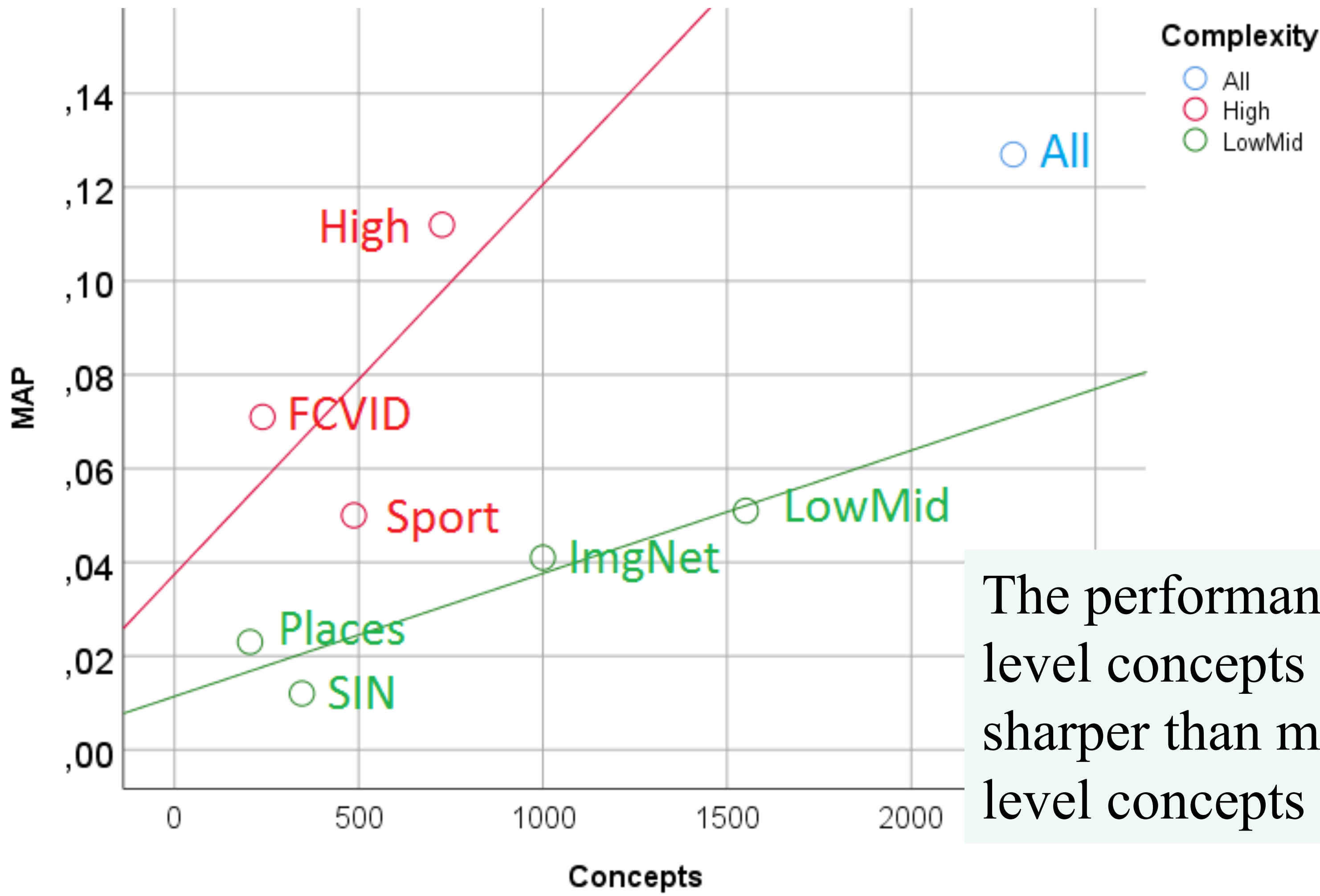
Number of concepts vs. MAP

MED14Test (oracle run of top-*k* concepts)



Number of concepts vs. MAP

MED14Test (oracle run of top-*k* concepts)



The performance of high level concepts increases sharper than mid and low level concepts

Issue-B: Concept Selection

1. How to select the most appropriate concept detectors?
2. How many concepts are enough?
3. How to combine concepts?

Find shots of something burning with flames visible

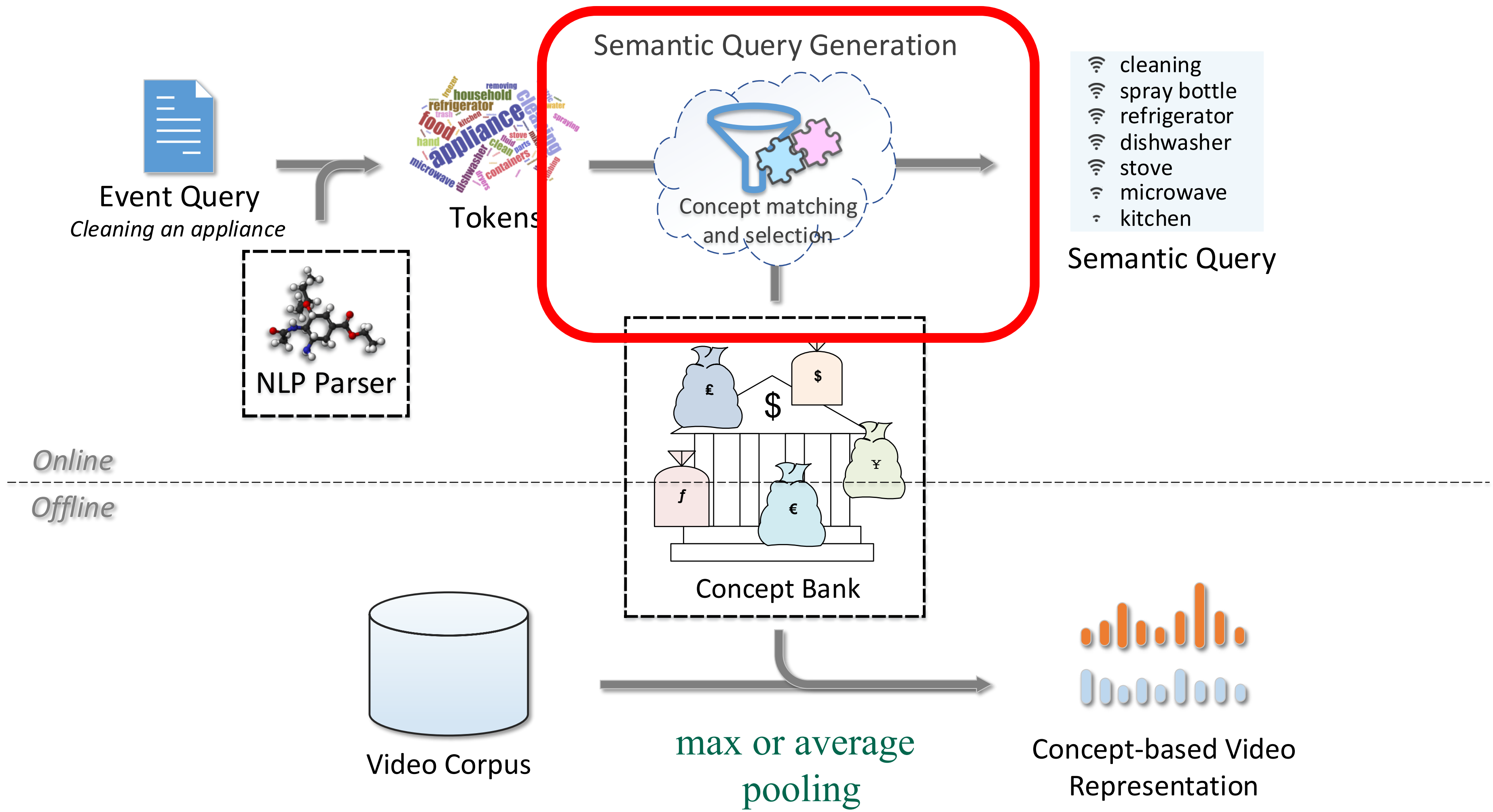
explosion?

smoke?

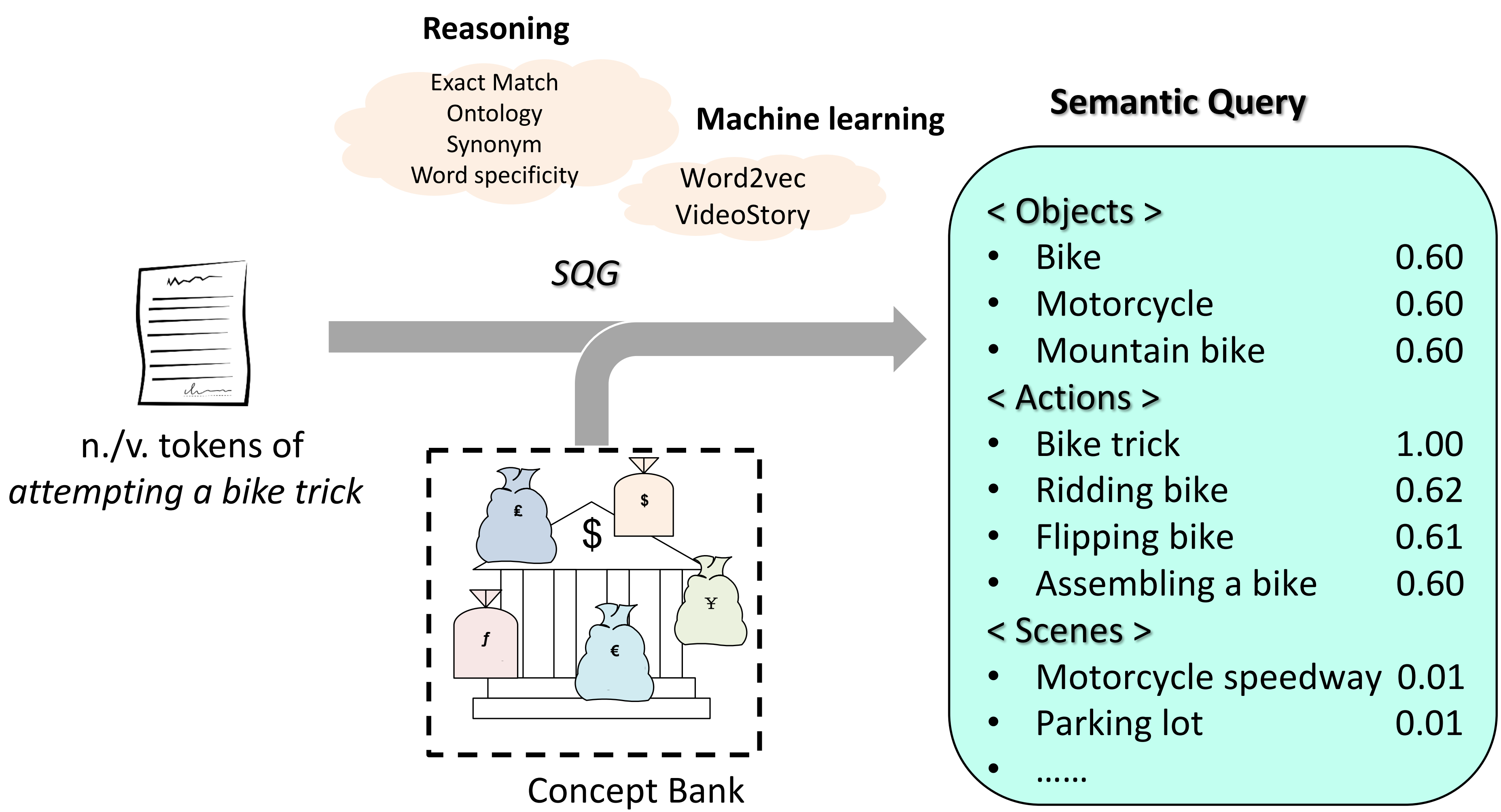
fire?

screaming?

Issue-B: Concept Selection

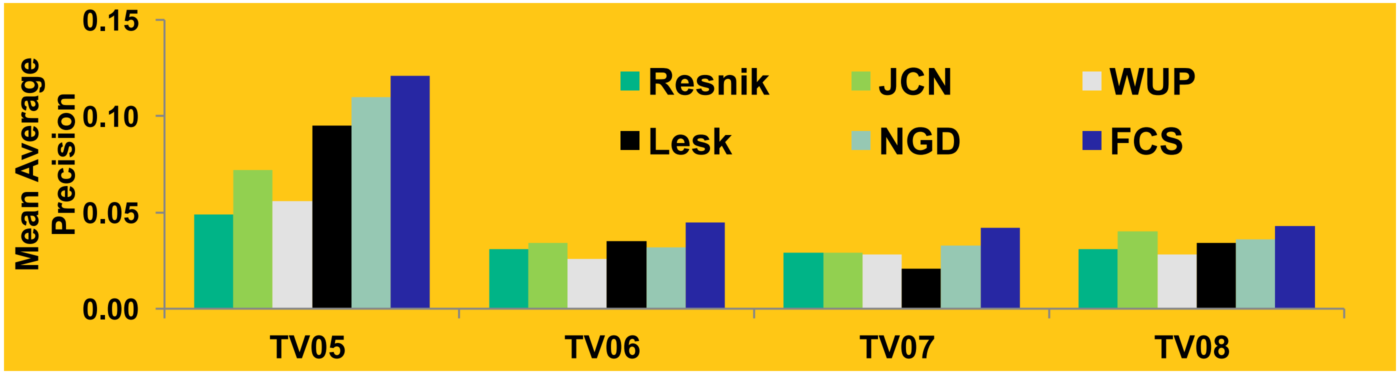


B1: Semantic Query Generation



B1: Ad-hoc Query

Query	WordNet (WUP)	Google Search (NGD)	Flickr Context (FCS)
A **goal** in a soccer match	Striking	Sports	Soccer
Something burning with **flames** visible	Sky	Soldiers	Smoke
Scenes with snow	Landscape	Person	Urban_Scenes
A **train** in motion	Vehicle	Car	Railroad



Y. G. Jiang, C. W. Ngo, S. F. Chang, "Semantic context transfer across heterogeneous sources for domain adaptive video search," in *ACM Multimedia*, 2009.

B1: Flickr Context Similarity

Query #199: Find shots with a person walking or riding a bicycle

flickr Context Similarity (FCS)

Couple walking
by Selestadienne
1 fave
Tagged with bicycle, cuba, cardenas, couplewalking ...
Taken on January 22, 2009, uploaded January 22, 2009
See more of Selestadienne's photos, or visit her profile.

Family Bicycle Ride
by AGeekMom
Tagged with family, bicycle, cycling, ride ...
Taken on September 7, 2006, uploaded September 7, 2006
See AGeekMom's photos or profile.

Candid - Girl walking dog
by willterry64
Tagged with park, dog, photoshop, walking ...
Taken on November 11, 2008, uploaded November 11, 2008
See more of willterry64's photos, or visit his profile.

Selected Concepts:

Bicycles (0.31), Person (0.31), Walking (0.12), Walking_Running (0.10), **Horse** (0.06), **Dog** (0.05) …

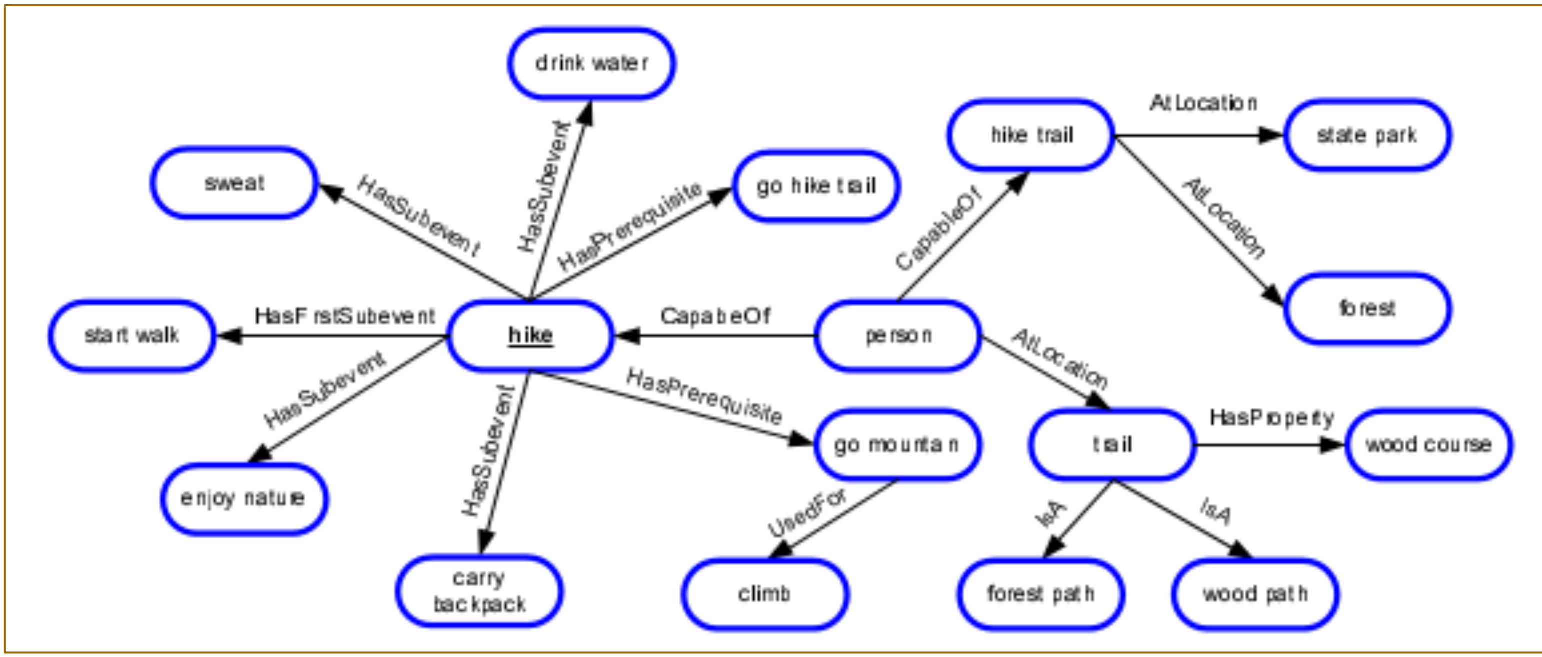
Top 10 ranked shots
AP=0.323

Video Data Set

B1: Event Query

B1: Concept Selection

"Hiking"



How to convert "context" into a searchable query?

C. C. Tan, and C. W. Ngo. On the use of commonsense ontology for multimedia event recounting. Int. Journal of Multimedia Information Retrieval, 2015.

B1: Concept Selection

Exact or partial string matching

+

Ontology reasoning
Wordnet, Conceptnet
Context inference
Word2Vec, Flickr, Wikipedia

(hope for better precision)

"dog show" to "dog show" (1.0)
"dog show" to "dog" (0.5)

(hope for better recall)

Context understanding is difficult, and can easily cause query drift when context is interpreted wrongly

B1: Why Wordnet reasoning is risky?

Dog Show

Concept "dog"

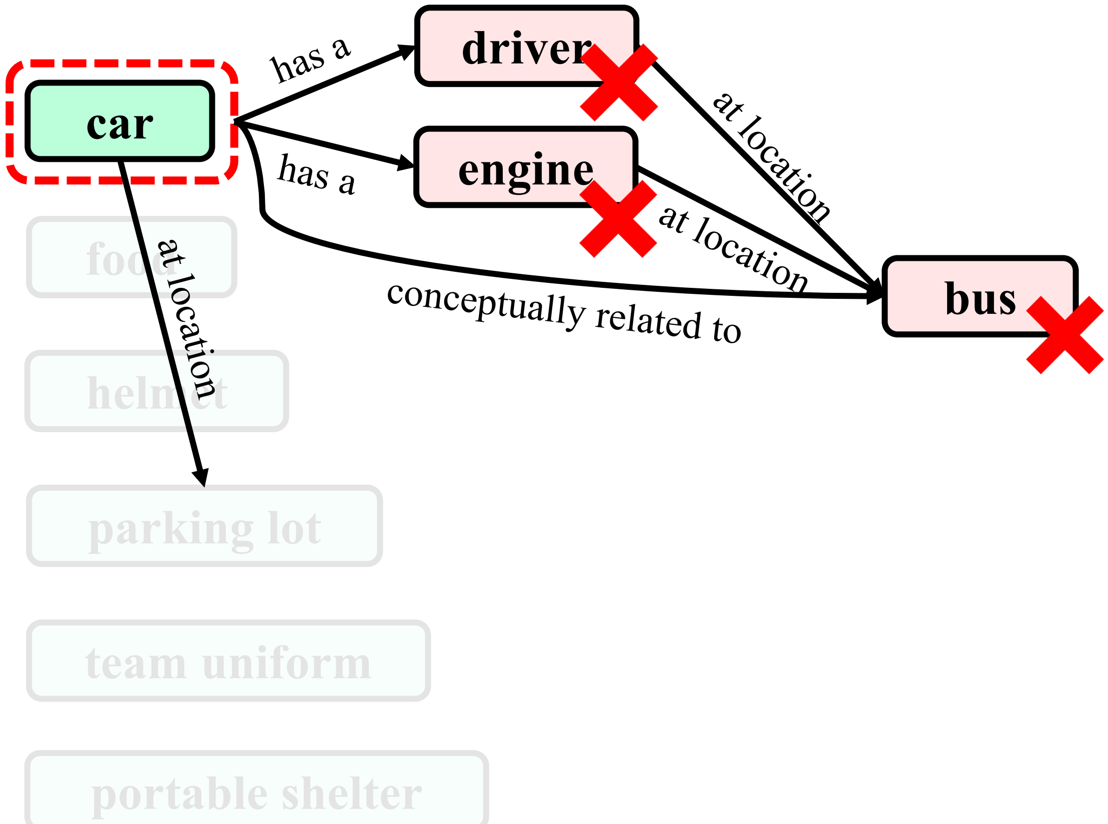
- red wolf
- kit fox
- cat
- horse
- mammal
- carnivore
- animal

ImageNet

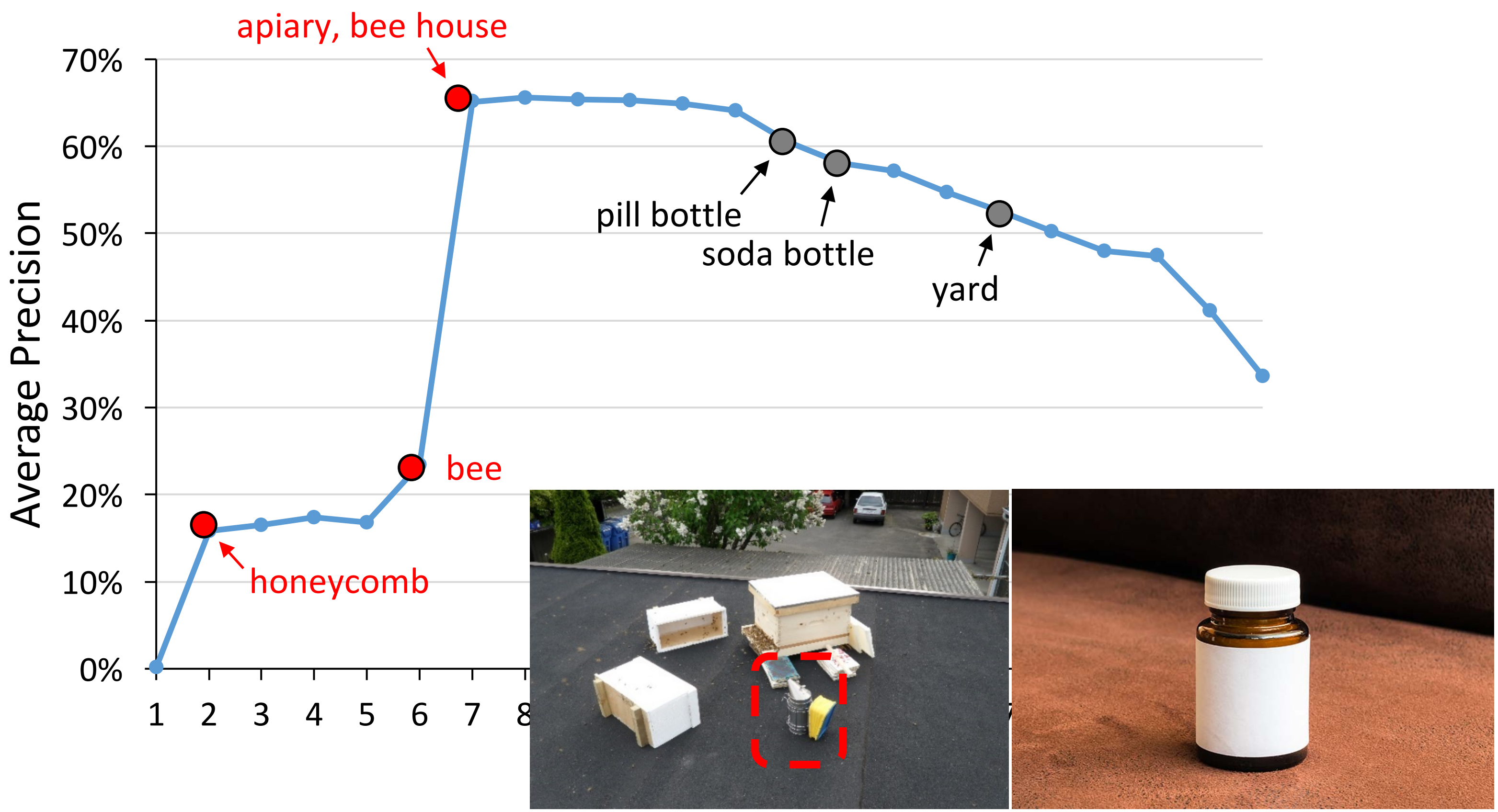
SIN

B1: Why Conceptnet is risky?

Tailgating

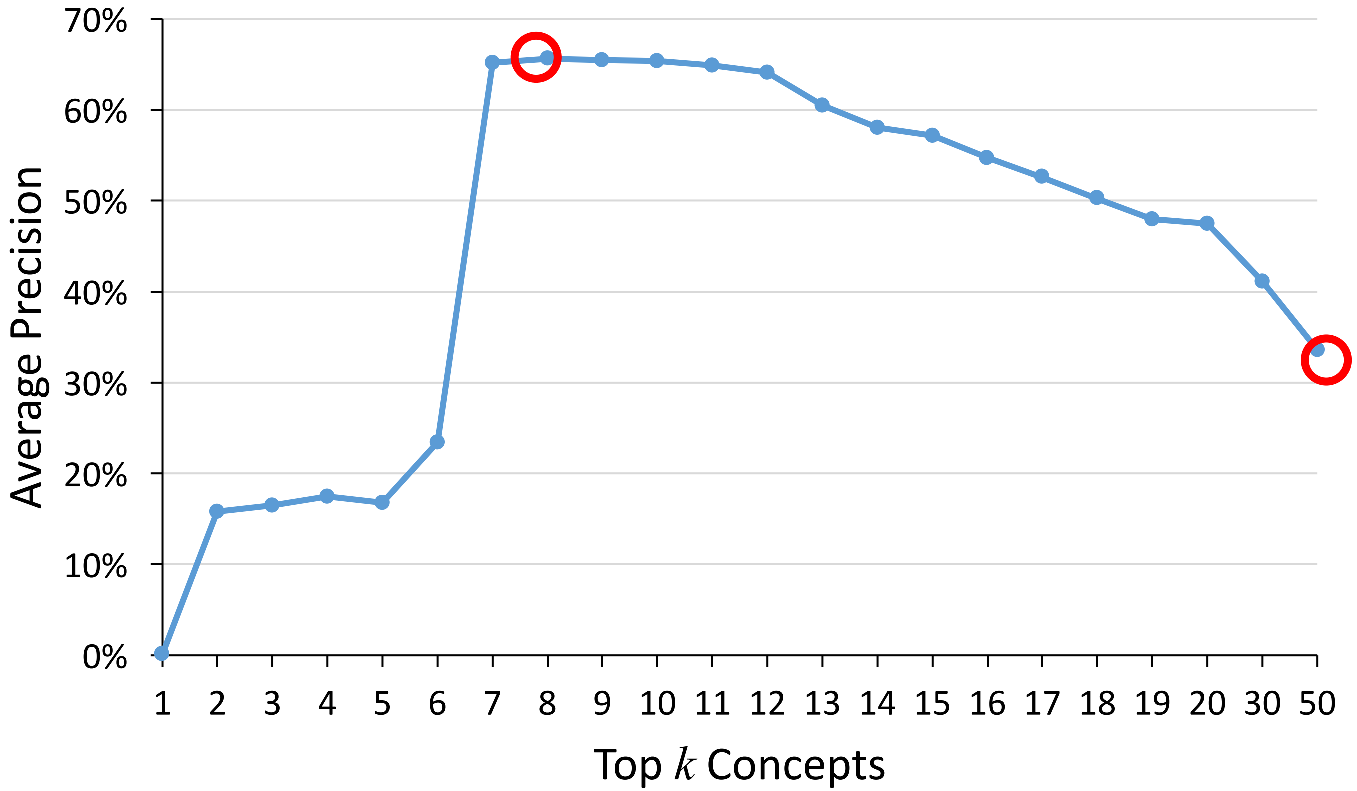


B2: How many concepts are enough?



Event 31: Beekeeping

B2: How many concepts are enough?



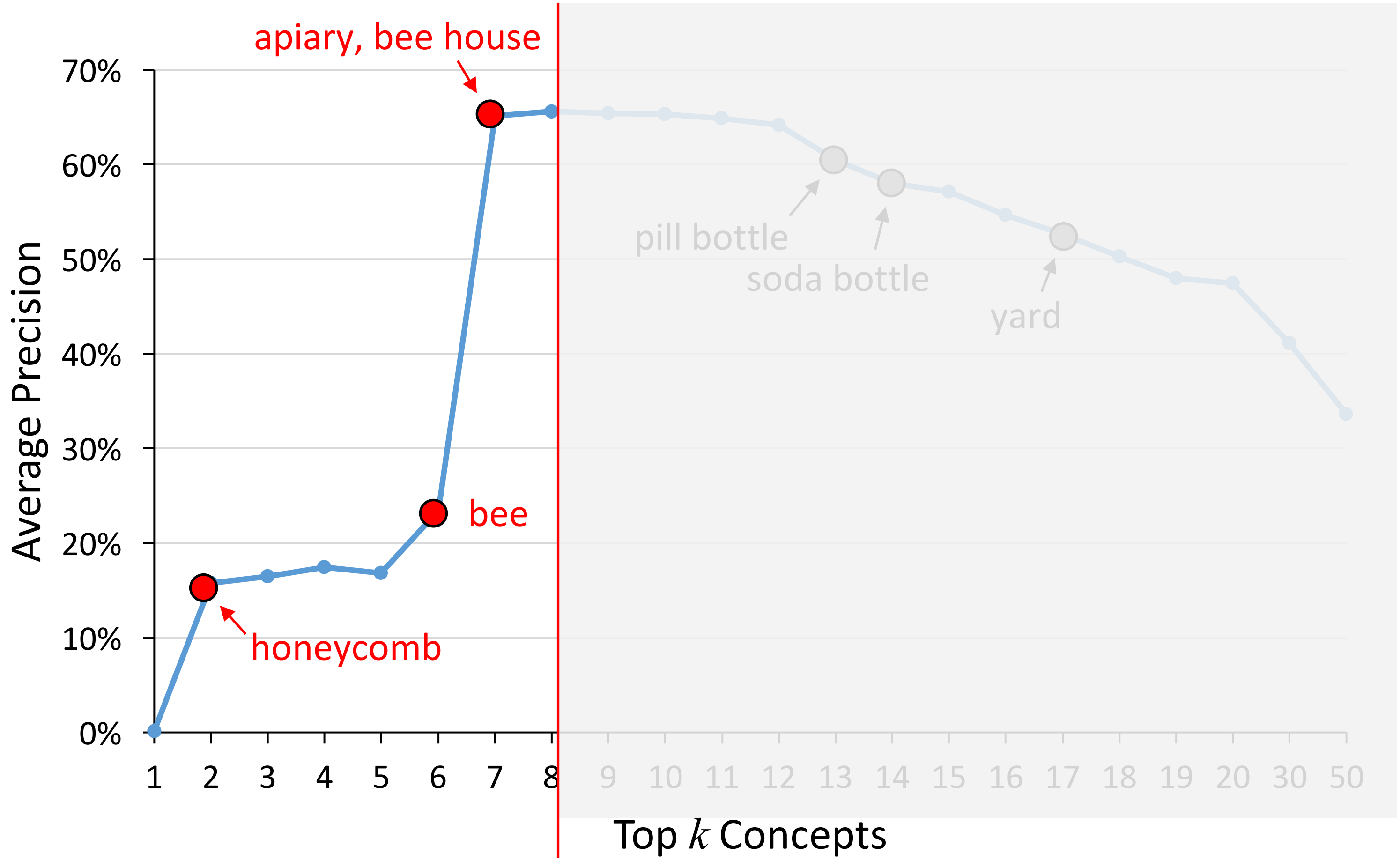
Event 31: Beekeeping

B2: How many concepts are enough?

Approaches

- Thresholding – simple fix
- Manual screening
- Evidential pooling
- Incremental Word2Vec

B2: A simple fix – Naïve cutoff



Event 31: Beekeeping

B2: A simple fix – Naïve cutoff

Concept Bank	#Concepts	Optimum k	MAP
Sports	487	10	0.103
FCVID	239	1	0.071
Research Collection	497	2	0.053
ImageNet	1,000	3	0.049
Places	205	2	0.020
SIN	346	5	0.014
Concept Bank (ALL)	2,774	9	0.129
AutoSQGSys [1]	4,043	--	0.115

** The MAP is reported on MED14Test

[1] L. Jiang, S.-I. Yu, D. Meng, T. Mitamura, and A. G. Hauptmann. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In 5th ICMR, pages 27-34, 2015.

B2: Manual concept screening

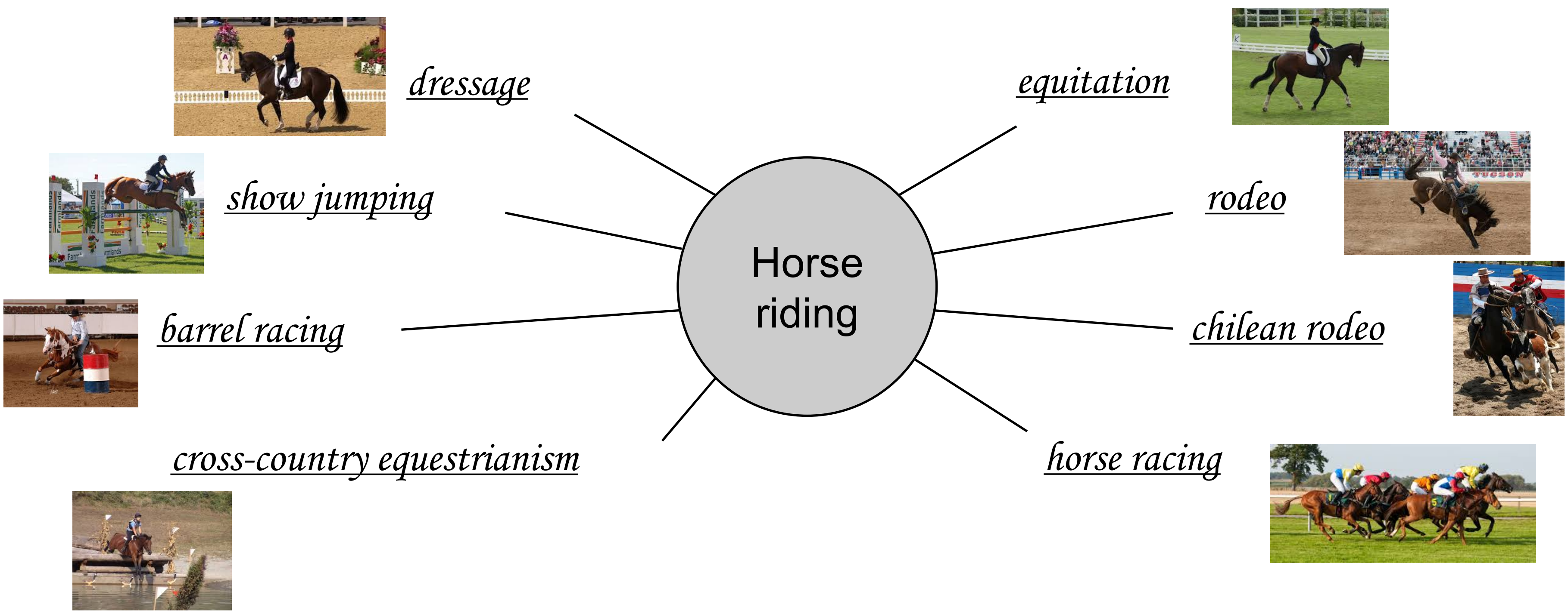
– Remove *false positives* by screening the names of concepts
– Only include concepts that are *distinctive* to an event if we find a concept detector *semantically matches* the event
– Remove concepts for which training videos may appear in *very different context* based on human's common sense

Rock climbing

- *Rock climbing, bouldering, sport climbing, artificial rock wall*	**Relevant**
- *Rope climbing, climbing, rock*	**Non-distinctive**
- *Rock fishing, rock band performance*	**False positive**
- *Stone wall, grabbing rock*	**Different context**

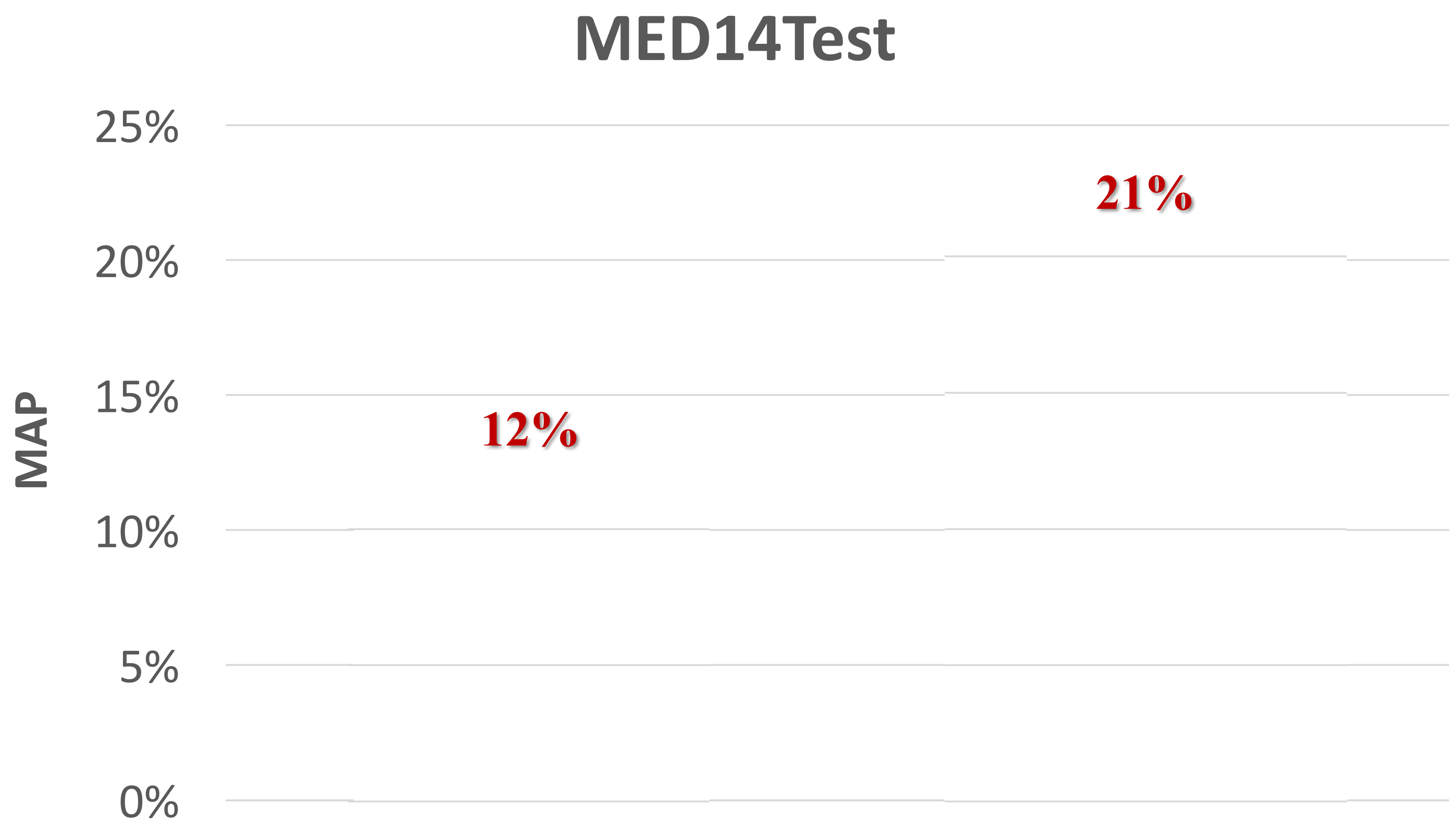
Concept Bank (Sport)

Manual search is particularly useful!

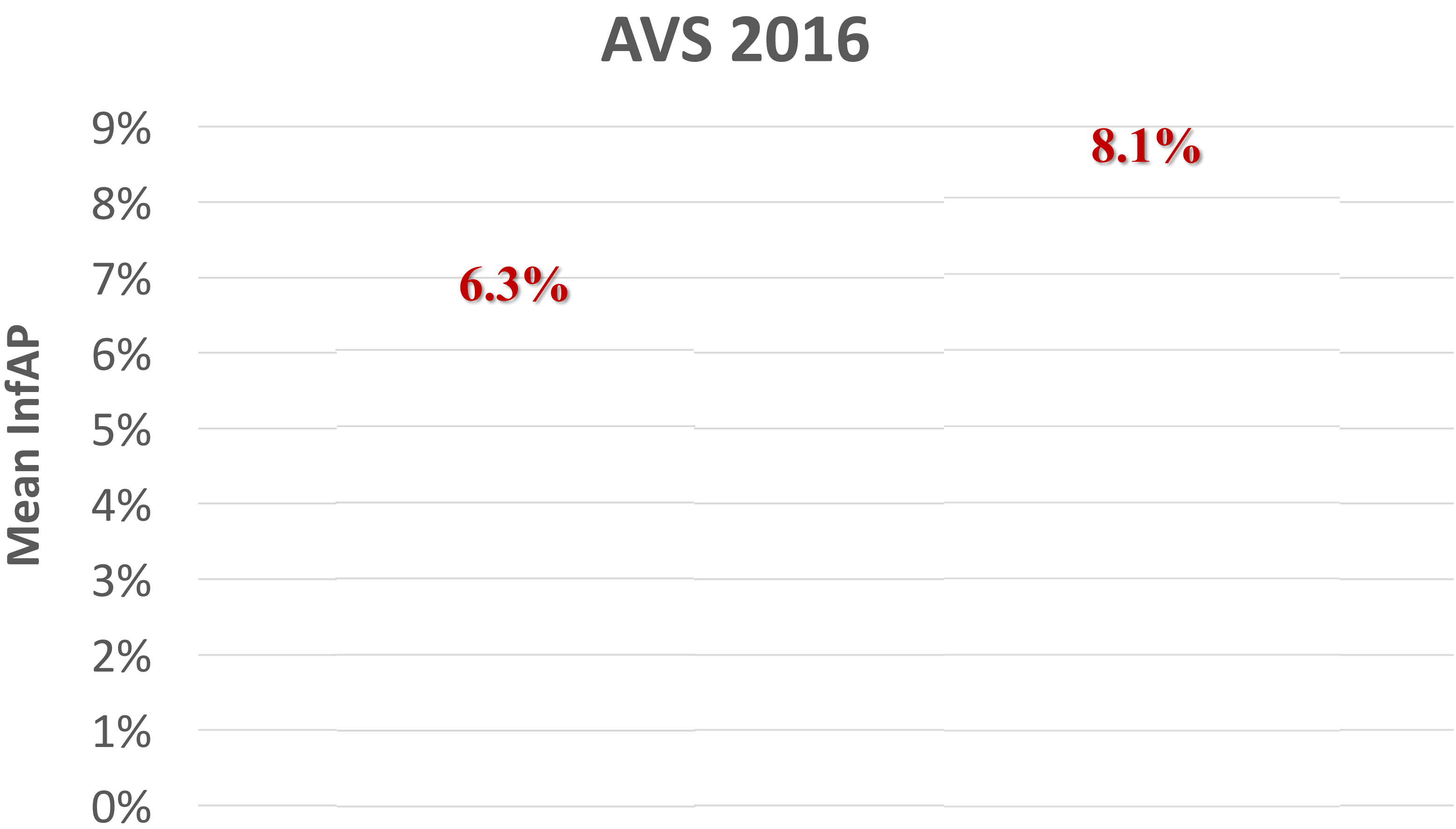


L. Jiang, S.-I. Yu, D. Meng, T. Mitamura, and A. G. Hauptmann, "Bridging the ultimate semantic gap: A semantic search engine for internet videos," in *International Conference on Multimedia Retrieval*, 2015.

Auto versus manual search

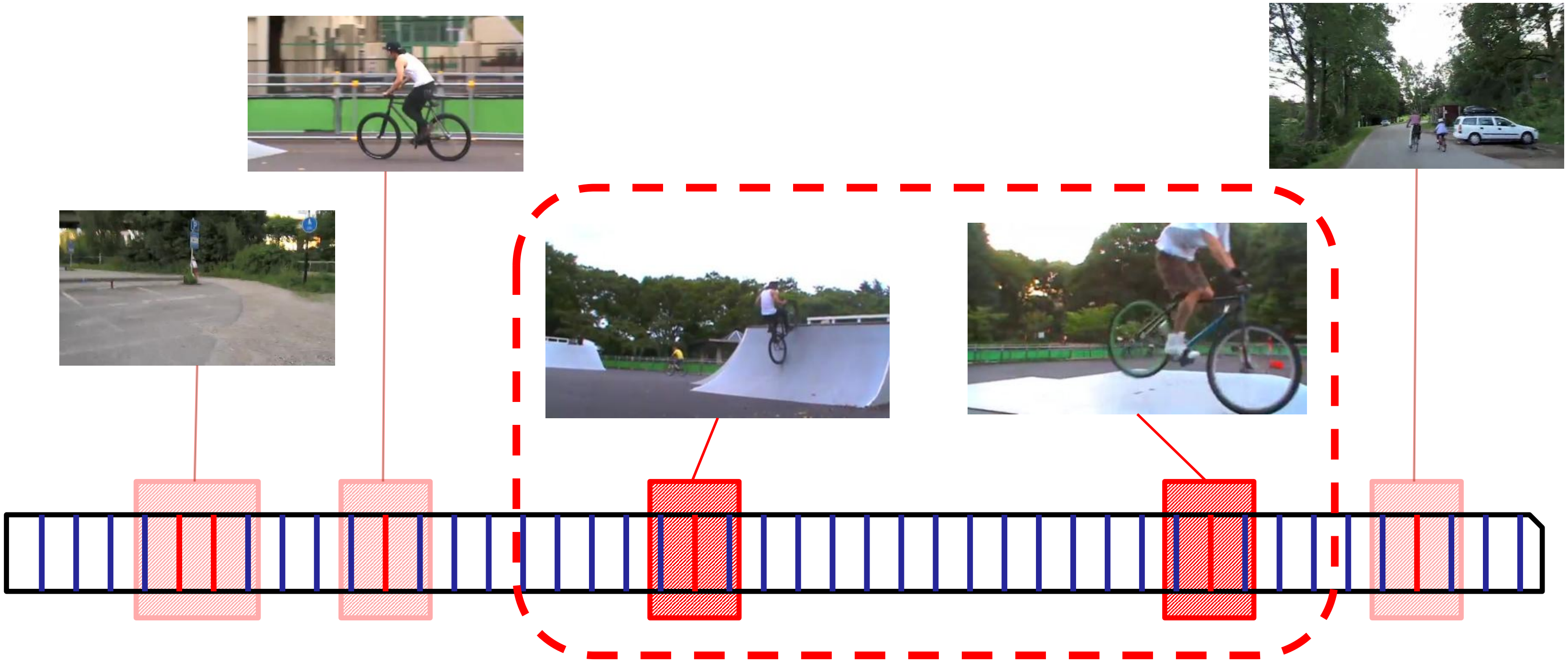


Auto versus Manual search



B2: Evidential pooling

The ***evidences*** that justifies a detection result only resides in a few shots

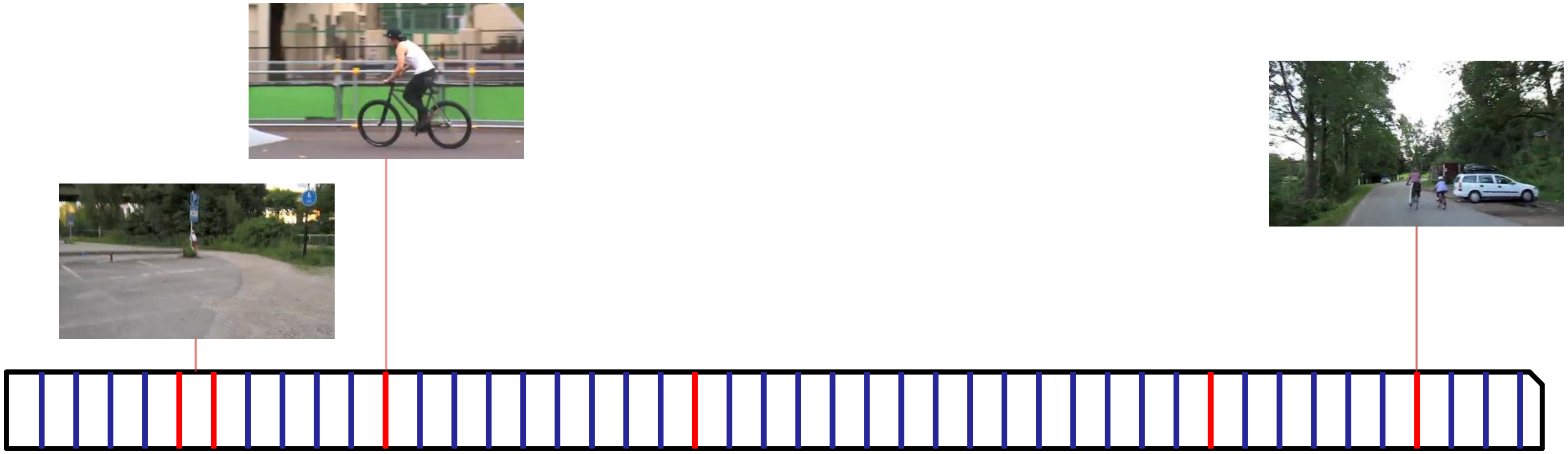


An example video of "attempting a bike trick"

B2: Evidential Pooling

If we uniformly pool a video

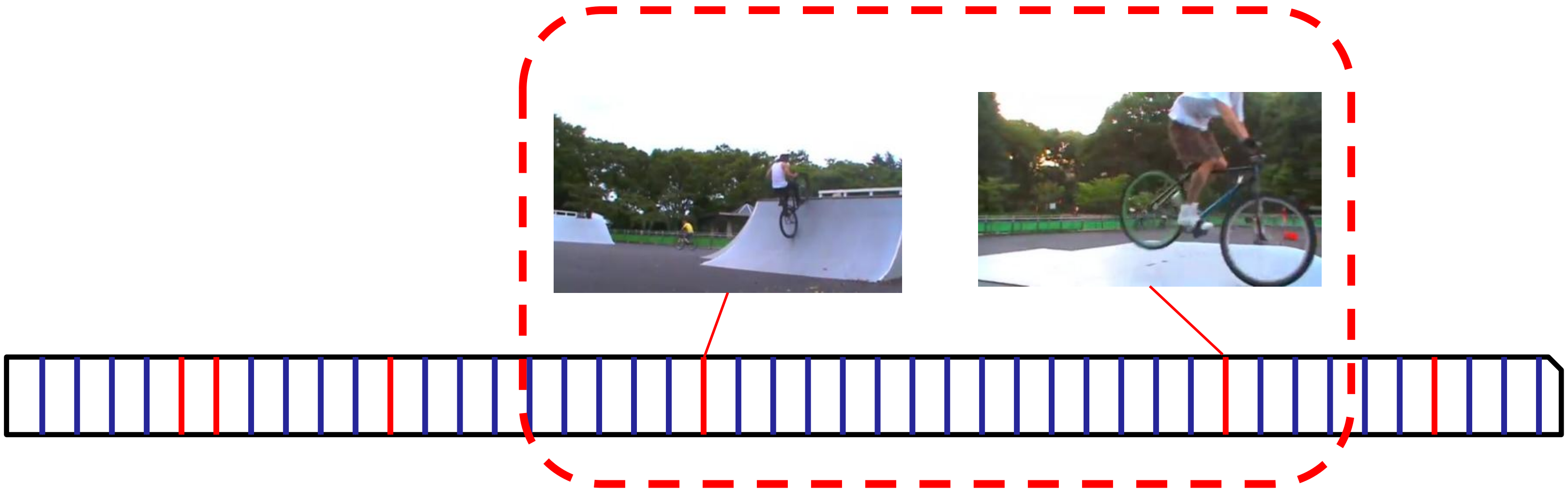
- We would collect the responses of many contextually relevant but indiscriminative concepts
- Lane/road, bike, riding bike, parking meter, traffic sign, guard rail, tree, cars, and pedestrians



B2: Evidential Pooling

If we pool only the query-related evidential shots

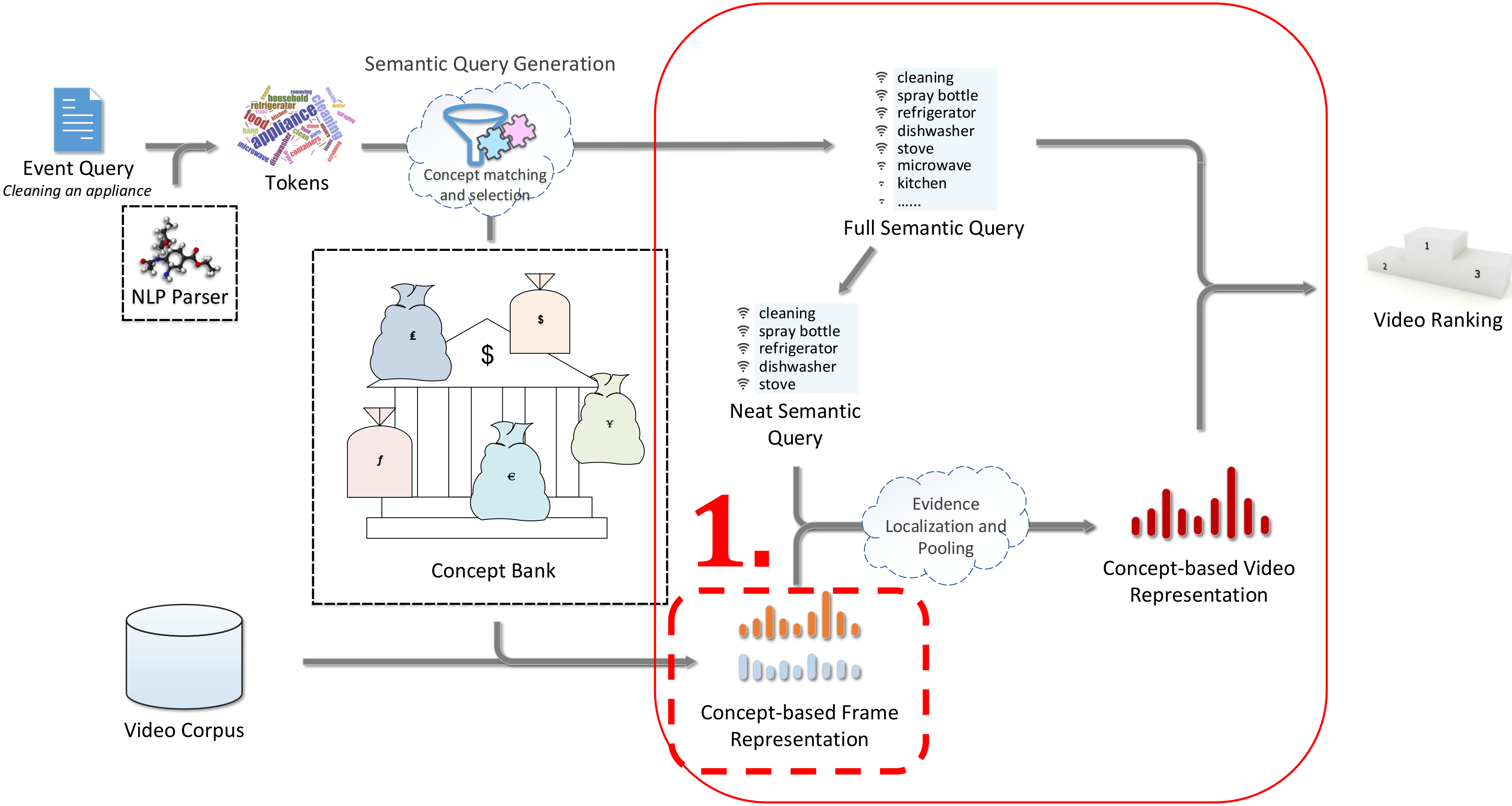
- The video representation is more focused (refined)
- Lane/road, tree, bike, riding bike, platform



Yi-Jie Lu, Hao Zhang, Maaike de Boer, Chong-Wah Ngo, "Event Detection with Zero Example: Select the Right and Suppress the Wrong Concepts," In ICMR, New York City, USA, Jun 2016.

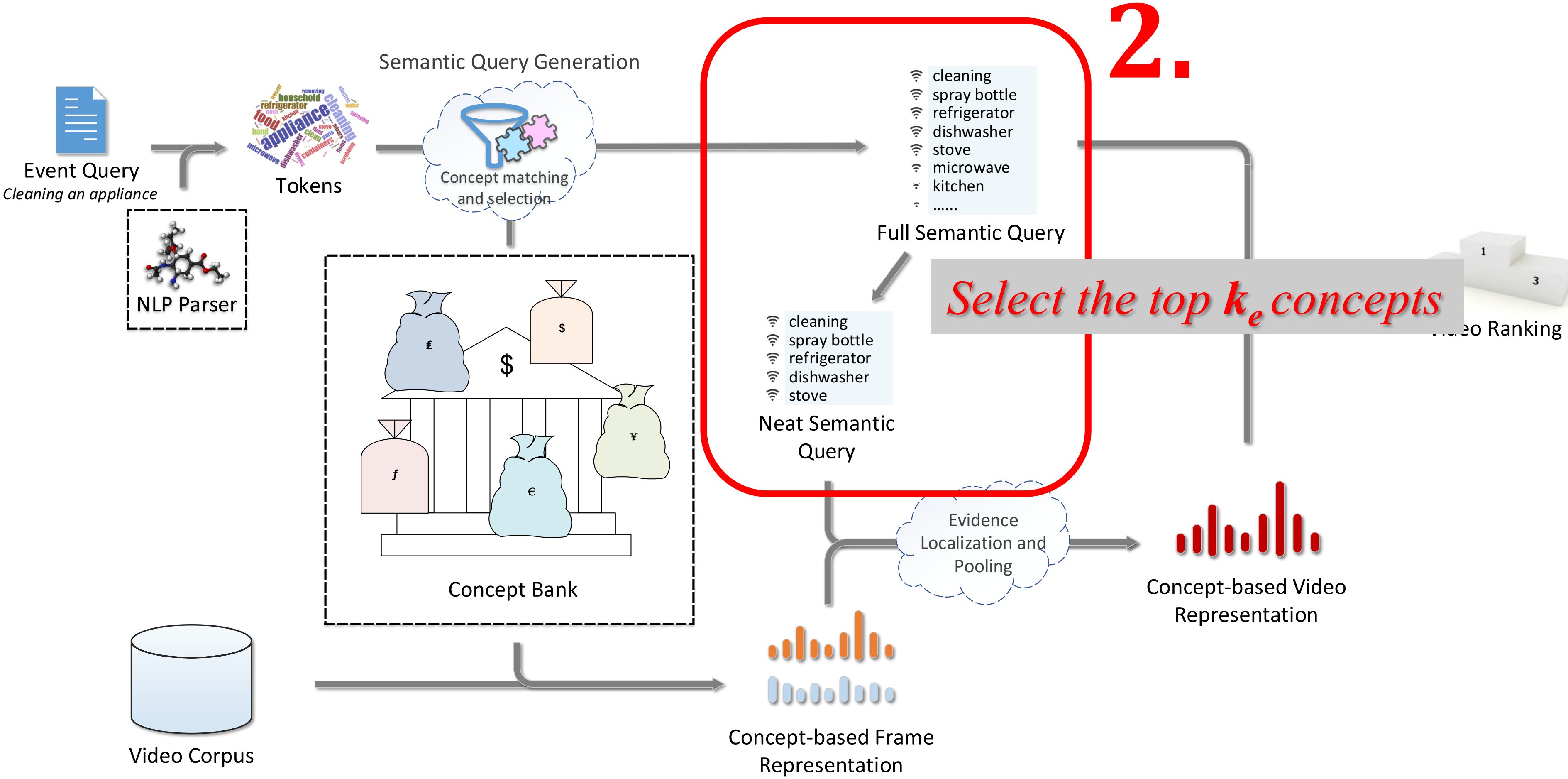
B2: Evidential Pooling (Algorithm)

Index every (key)frame of a video



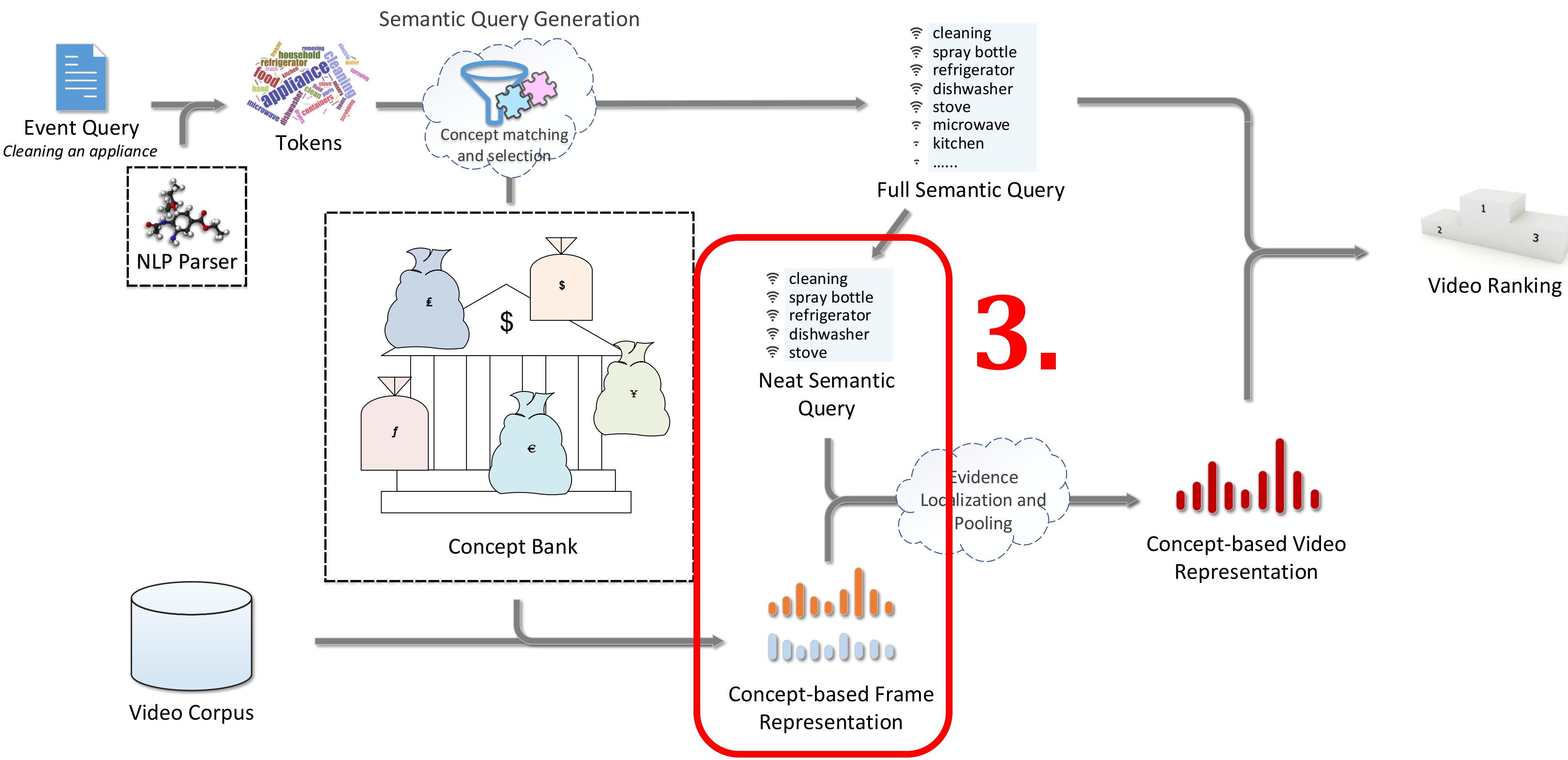
B2: Evidential Pooling (Algorithm)

Determine the most confident k_e concepts



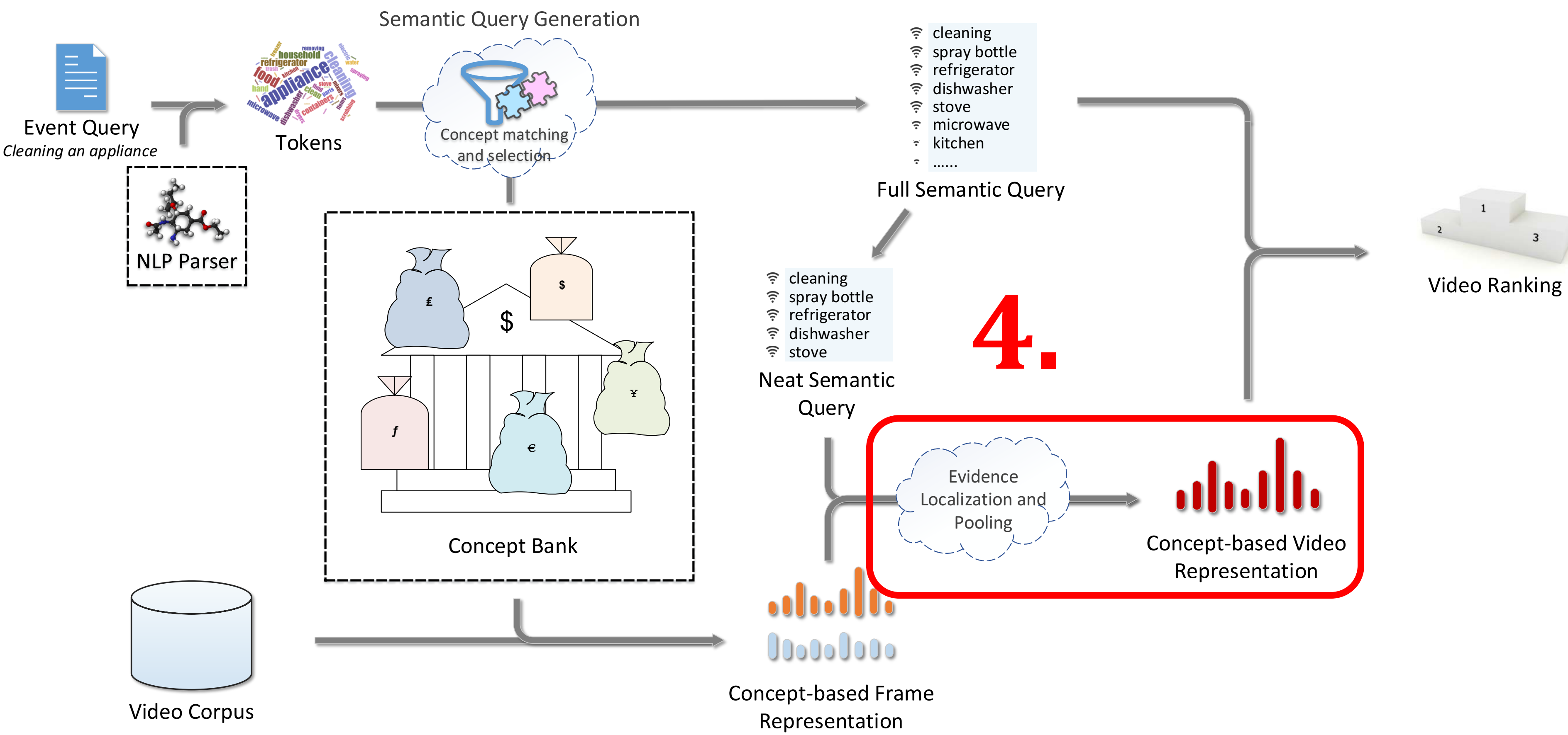
B2: Evidential Pooling (Algorithm)

Identify top-$\underline{n}$ (key) frames with high responses to $\underline{k_e}$ concepts



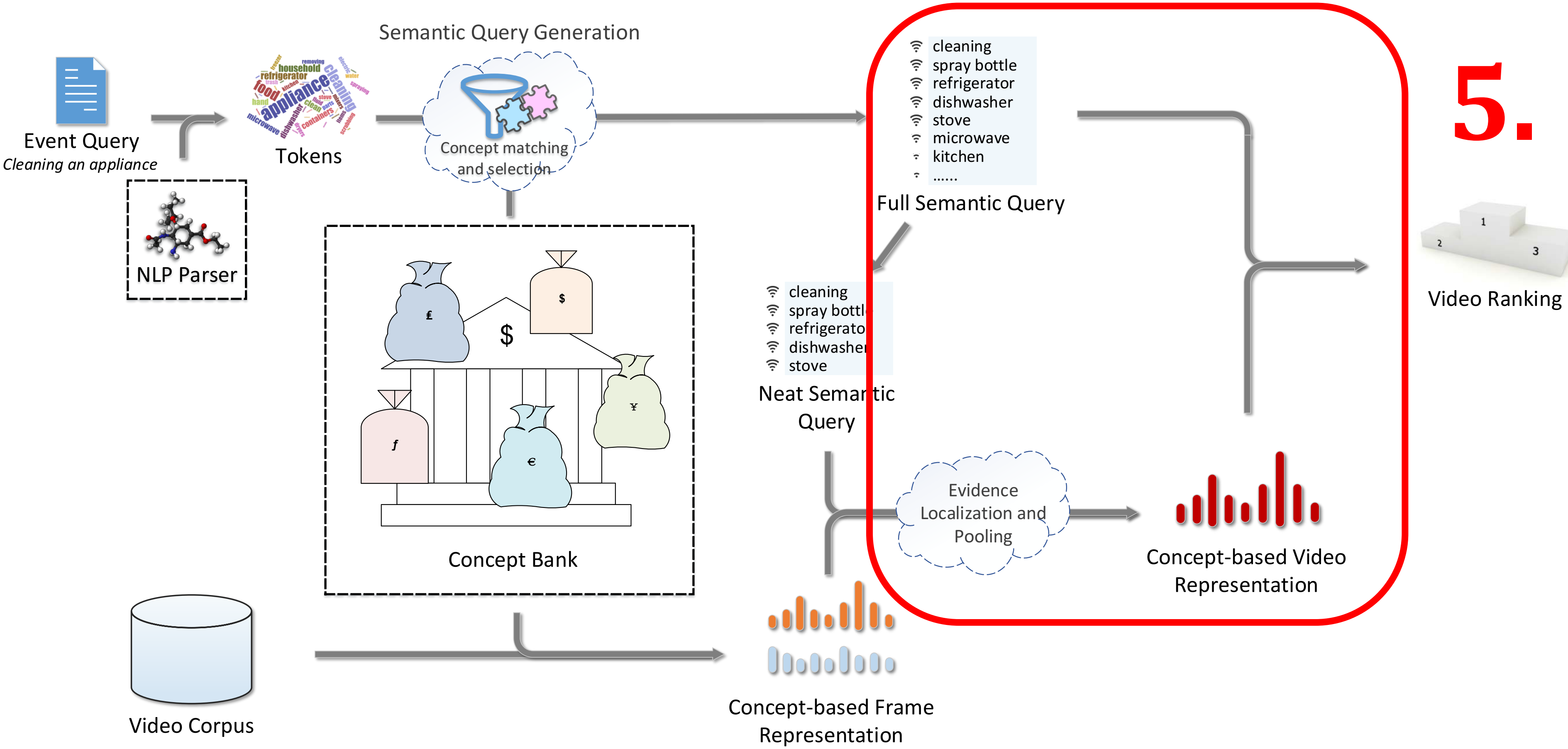
B2: Evidential Pooling (Algorithm)

Perform evidential pooling to generate video-level features



B2: Evidential Pooling (Algorithm)

Perform video search using full semantic query



Sensitivity of k_e

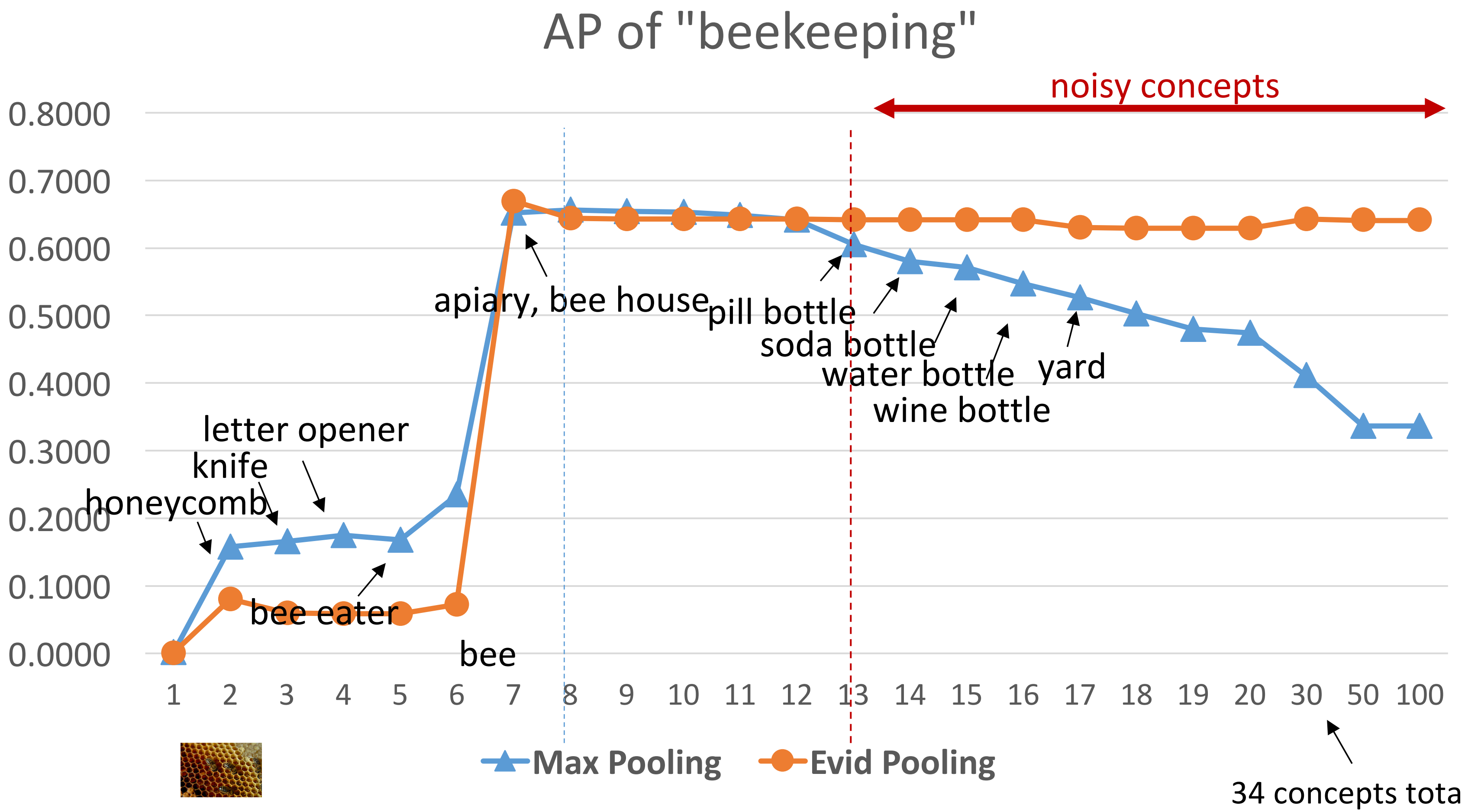
k_e	**2**	**4**	**8**	**16**
MAP	0.0725	0.0771	0.0775	0.0770

** The MAPs are based on **10 evidential shots**

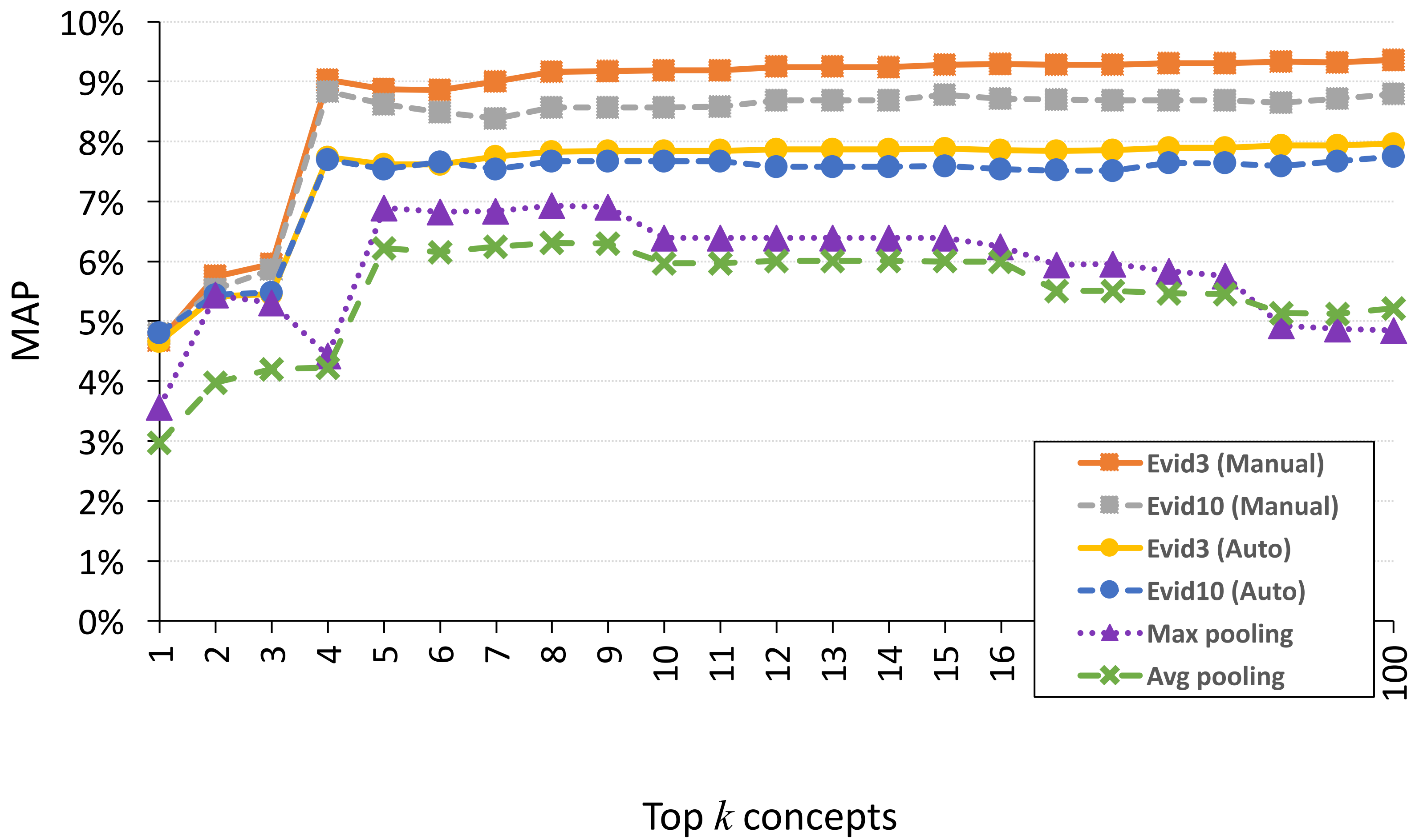
The user studies show that one can recognize a complex query **with very small number of shots** (1-3).

S. Bhattacharya, F. X. Yu, S. F. Chang, "Minimally needed evidence for complex event recognition in unconstrained videos," in *International Conference on Multimedia Retrieval*, 2014.

B2: Evidential Pooling



B2: Evidential Pooling



B2: Incremental Word2Vec

- Key idea
  - Vector composition using element-wise addition
  - Care about query drift more than concept similarity

felling tree ≈ fruit tree pruning + tree frog + tree farm

parking a vehicle ≈ vehicle + parking lot

M. H. T de Boer, Y. J. Lu, H. Zhang, K. Schutte, W. Kraaij and C. W. Ngo. Semantic Reasoning in Zero Example Video Event Retrieval. TOMM, to appear.

B2: Incremental Word2Vec (Algorithm)

1. Embed a query with Word2Vec, Q

   UQ = Word2Vec(changing) + Word2Vec(vehicle) + Word2Vec(tire)

2. Embed concepts to Word2Vec
3. Pick the most similar concept to query

   C = Word2Vec(most-similar-concept)

4. Pick the next most similar concept
   - C' = C+Word2Vec(next-most-similar-concept)
   - if Cosine(UQ,C') > Cosine(UQ,C)

     C = C'

Get rid of concepts which may cause query drift

B2: Incremental Word2Vec (Algorithm)

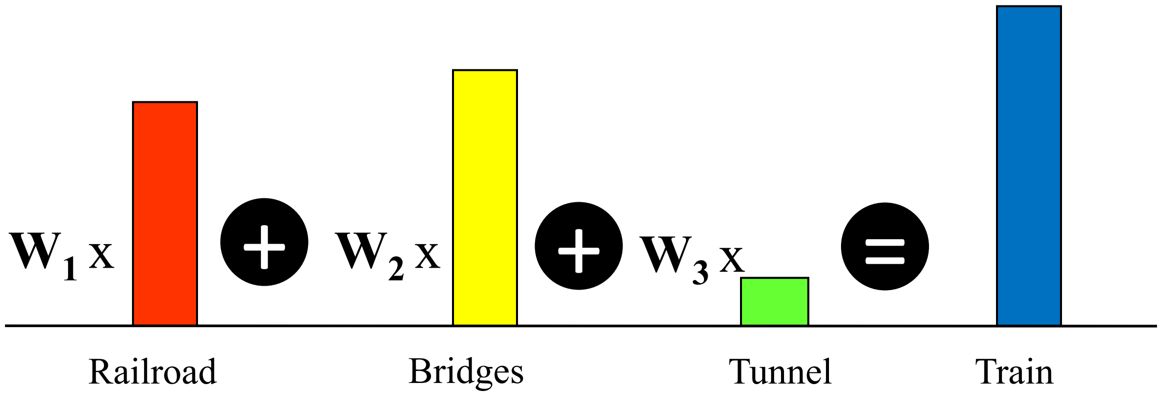
MED14Test

Cutoff at top-X%	MAP	Ave. # of concepts	Stdv
90%	0.142	1.9	1.3
85%	0.142	2.3	2.4
80%	0.142	3.0	4.6
75%	0.141	3.8	6.3
50%	0.137	7.2	12.1
25%	0.136	9.3	13.4
0%	0.136	9.4	13.4

MAPs are quite stable over different cutoff points

B3: How to combine concepts?

Weighted sum of concept scores



<u>Open issue</u>

- How to normalize the score of concepts from different datasets and trained using different classification methods?
- How to combine concepts of similar names, probably learnt with different training examples of different context, from different datasets?

B3: How to combine concepts?

- AND-OR (Wasade@TRECVid2016)
  - OR: max operator
  - AND: sum or multiplying operator

Query: One or more people walking or bicycling on a bridge during daytime

people and (walking or bicycling) and bridge and daytime

0Ex Baseline

- For both ad-hoc and multimedia event queries
- Support more than 10,000 visual concepts
- Highly efficient: Can search a complex query within seconds on a laptop
- Support *concept screening* and *interactive search*
- Publicly available
  - ❖Open source
  - ❖Console for interactive search
  - ❖Concept features for MED14Test, IACC, TV08

All resources are available at http://vireo.cs.cityu.edu.hk/zeroex/ or https://github.com/iiedii/0-ex

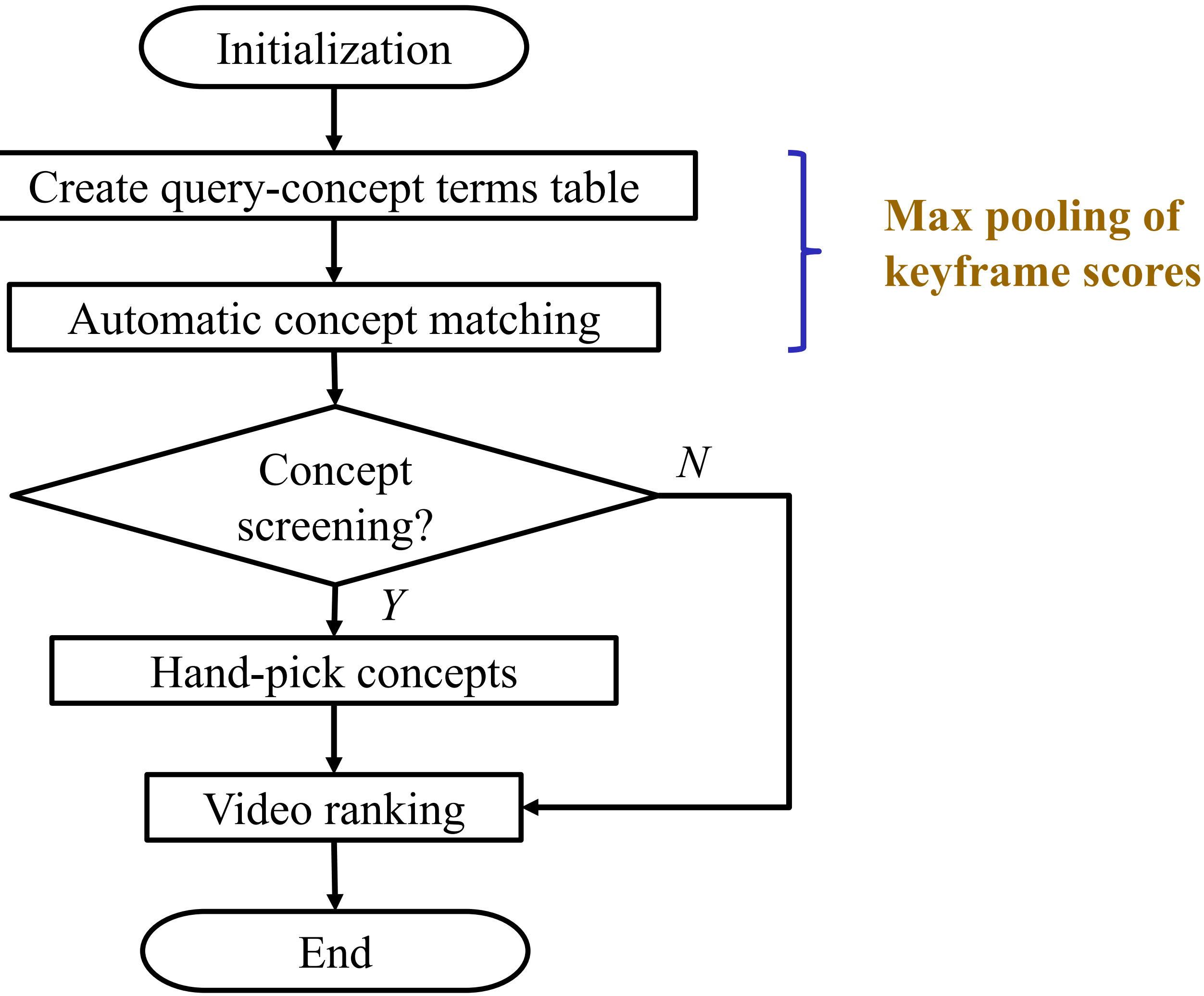
Baseline: Concept Bank

- Places 205
- SIN 346
- RC 497
- ImageNet 1,000

ResNet-50

- ImageNet 12,998 — **GoogLeNet** (ImageNet-Shuffle 13K)
- FCVID 239 — **FC7 (AlexNet) + SVM**
- Sport 487 — **3D CNN**

Baseline: Concept Selection



Baseline: Concept Selection

Query Tokens

- *bike*
- *motorized bike*
- *motorcycle*
- *bike jump*
- *street*
- *parking lot*
- *concrete floor*
- *bike ramps*
- *riding*
- *one wheel standing*
- *jumping*
- *spinning*
- *flipping*

Concept Names

…… ……
bicycle
…… ……
…… ……
…… ……
bike trick
…… ……
…… ……
…… ……
…… ……
riding bike
…… ……
freestyle bmx bike
…… ……
…… ……
…… ……
…… ……

1.

Each token votes for its most relevant concepts by a score metric concerning TF-IDF, word specificity, WordNet synonyms, etc. [1]

[1] Yi-Jie Lu, Hao Zhang, Maaike de Boer, Chong-Wah Ngo, "Event Detection with Zero Example: Select the Right and Suppress the Wrong Concepts," In ICMR, New York City, USA, Jun 2016.

Video Retrieval Group

Baseline: Concept Selection

Query Tokens

- *bike*
- *motorized bike*
- *motorcycle*
- *bike jump*
- *street*
- *parking lot*
- *concrete floor*
- *bike ramps*
- *riding*
- *one wheel standing*
- *jumping*
- *spinning*
- *flipping*

Concept Names

… …
bicycle
motorcycle
… …
… …
bike trick
… …
… …
… …
… …
riding bike
… …
freestyle bmx bike
… …
… …
motorcycle racing
… …

1.

Each token votes for its most relevant concepts by a score metric concerning TF-IDF, word specificity, WordNet synonyms, etc. [1]

[1] Yi-Jie Lu, Hao Zhang, Maaike de Boer, Chong-Wah Ngo, "Event Detection with Zero Example: Select the Right and Suppress the Wrong Concepts," In ICMR, New York City, USA, Jun 2016.

Baseline: Concept Selection

Query Tokens

- *bike*
- *motorized bike*
- *motorcycle*
- *bike jump*
- *street*
- *parking lot*
- *concrete floor*
- *bike ramps*
- *riding*
- *one wheel standing*
- *jumping*
- *spinning*
- *flipping*

Concept Names

……
bicycle ●
motorcycle ●
……
……
bike trick ●●
……
……
street ●
……
riding bike ●●
……
freestyle bmx bike ●●
……
……
motorcycle racing ●●
……

1.

Each token votes for its most relevant concepts by a score metric concerning TF-IDF, word specificity, WordNet synonyms, etc. [1]

[1] Yi-Jie Lu, Hao Zhang, Maaike de Boer, Chong-Wah Ngo, "Event Detection with Zero Example: Select the Right and Suppress the Wrong Concepts," In ICMR, New York City, USA, Jun 2016.

Baseline: Concept Weighting

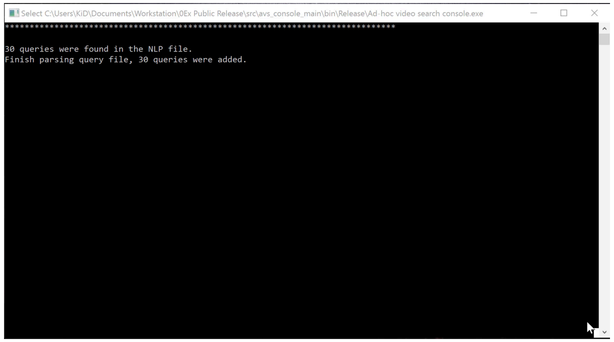
Query Tokens	Concept Names
bike	*bike trick* ●●
motorized bike	*riding bike* ●●
motorcycle	*freestyle bmx bike* ●●
bike jump	*motorcycle racing* ●●
street	*bicycle* ●
parking lot	*motorcycle* ●
concrete floor	*street* ●
bike ramps	……
riding	……
one wheel standing	……
jumping	……
spinning	……
flipping	……
	……
	……
	……
	……
	……
	……

2.

Weight the selected concepts according to the votes. [1]

[1] Yi-Jie Lu, Hao Zhang, Maaike de Boer, Chong-Wah Ngo, "Event Detection with Zero Example: Select the Right and Suppress the Wrong Concepts," In ICMR, New York City, USA, Jun 2016.

Baseline: Console




```
Ranking videos for all queries...
   - at Query 1: playing guitar outdoor
   - at Query 22: sitting laptop visible
   - at Query 8: person walking bicycling bridge daytime
   - at Query 15: jumping
   - at Query 16: man shake hand woman
   - at Query 9: crowd demonstrating city street night
   - at Query 2: man indoor camera bookcase
   - at Query 23: person opening door exiting
   - at Query 17: policeman polouse car visible
   - at Query 10: sewing machine
   - at Query 3: playing drum indoor
   - at Query 24: man beard wearing white robe speaking gesturing camera
   - at Query 11: destroyed building ruin
   - at Query 25: holding knife
   - at Query 4: diver wearing diving suit swimming water
   - at Query 18: person train station platform
   - at Query 12: palm tree
   - at Query 19: man beach scene
   - at Query 26: woman female wearing glass
   - at Query 13: military personnel interacting protester
   - at Query 5: holding poster street daytime
   - at Query 20: type fountain outdoor
   - at Query 27: drinking cup mug bottle container
   - at Query 21: man beard talking singing microphone
   - at Query 28: wearing helmet
   - at Query 14: soldier performing training military maneuver
   - at Query 6: 43rd president george w. bush sitting talking person indoor
   - at Query 29: lightening candle
   - at Query 7: choir orchestra conductor performing stage
   - at Query 30: person shopping
Done.
Video ranking time spent: 00:00:27.5176059.

Loading ground truth...
Writing ground truth info...

Calculating infAP for each query...
```

4s /query for a concept bank (14K), including query processing, searching, ranking

AVS 2016 Concept Bank

Basic

- Places 205
- SIN 346
- Research collection 497
- ImageNet 1,000

2,048 in total

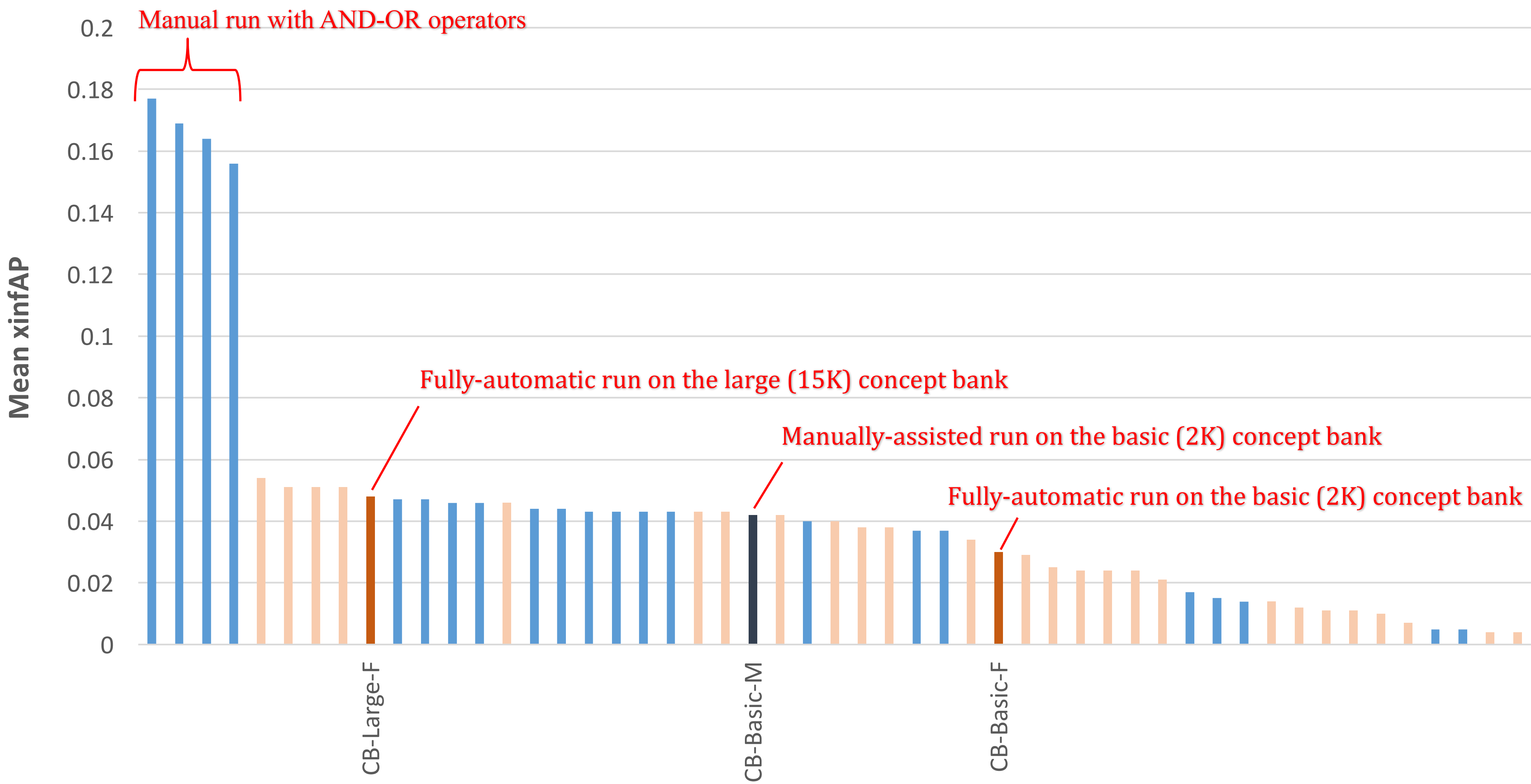
Large

- Places 205
- SIN 346
- Research collection 497
- ImageNet 12,998

14,046 in total

AVS 2016 Performance

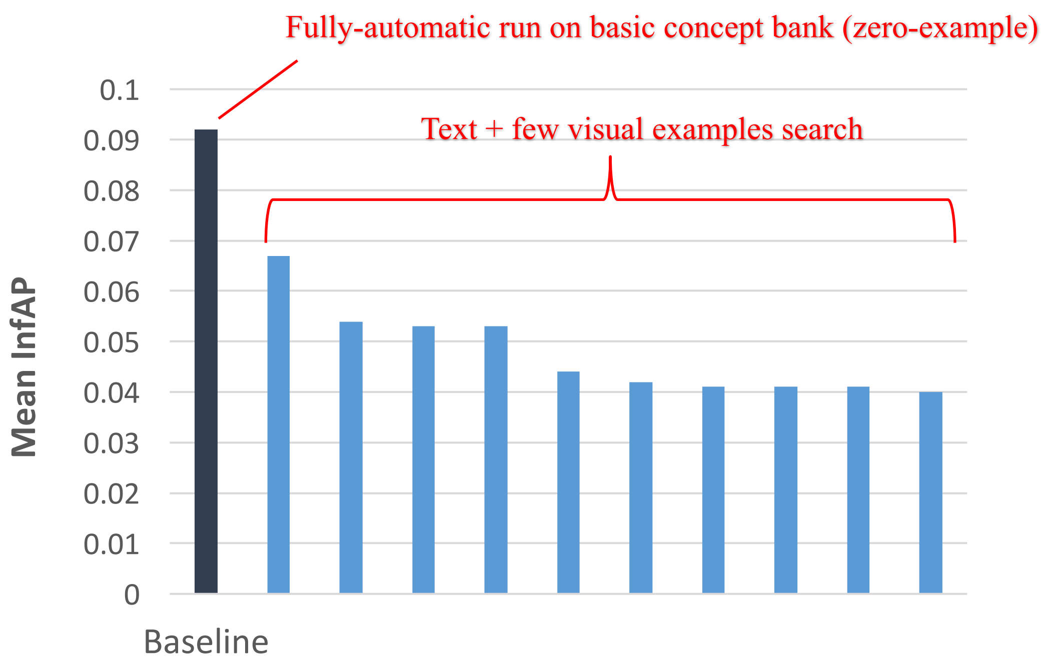
On TRECVID-2016 AVS Benchmark



** Scores of the other teams are from the official report of TRECVID-2016 AVS task.
The blue bars are manually-assisted runs and orange bars are fully-automatic runs.

Search Task 2008 Performance

On TRECVID-2008 Video Search Task Benchmark (Fully-Automatic)



** Scores of the other teams are from the official report of TRECVID-2008 search task.

MED14Test Concept Bank

Basic

- Places 205
- SIN 346
- Research collection 497
- ImageNet 1,000

2,048 in total

Large

- Places 205
- SIN 346
- Research collection 497
- ImageNet 1,000
- FCVID 239
- Sport 487

2,774 in total

Video Retrieval Group

MED14Test Performance

On MED14Test 0-Ex Benchmark

Concept Bank	#Concepts	Query	MAP
Baseline (basic)	2,048	Automatic	0.069
Baseline (basic)	2,048	Manual	0.110
Baseline (large)	2,774	Automatic	0.113
Baseline (large)	2,774	Manual	0.191
AutoSQGSys [1]	4,043	Automatic	0.115
Dynamic composition [2]	3,135	Automatic	0.134
Incremental Word2Vec [3]	2,277	Automatic	0.142
VisualSys [1]	4,043	Manual	0.176

[1] L. Jiang, S.-I. Yu, D. Meng, T. Mitamura, and A. G. Hauptmann. Bridging the ultimate semantic gap: A semantic search engine for internet videos. ICMR, page 27-34, 2015.

[2] Z. Chang, Y. Yang, G. Long, C. Zhang, and A. G. Hauptmann. Dynamic concept composition for zero-example event detection. arXiv, 2015.

[3] M. H. T de Boer, Y. J. Lu, H. Zhang, K. Schutte, W. Kraaij and C. W. Ngo. Semantic Reasoning in Zero Example Video Event Retrieval. TOMM, to appear.

Conclusion

What do we gain over the past one decade?

- Bigger dataset
- Bigger concept bank (more than 5,000 concepts!)
- Better and more variety of features
- Still a large room for performance improvement (MAP hardly > 0.2)
- Still, simple approach works better
- Some light on number of concepts to select for a query
- Context is still difficult

Try the baseline 😄😃😀😊
http://vireo.cs.cityu.edu.hk/zeroex/ or
https://github.com/iiedii/0-ex