# Video Search when examples are scarce

Dennis Koelma and Cees Snoek

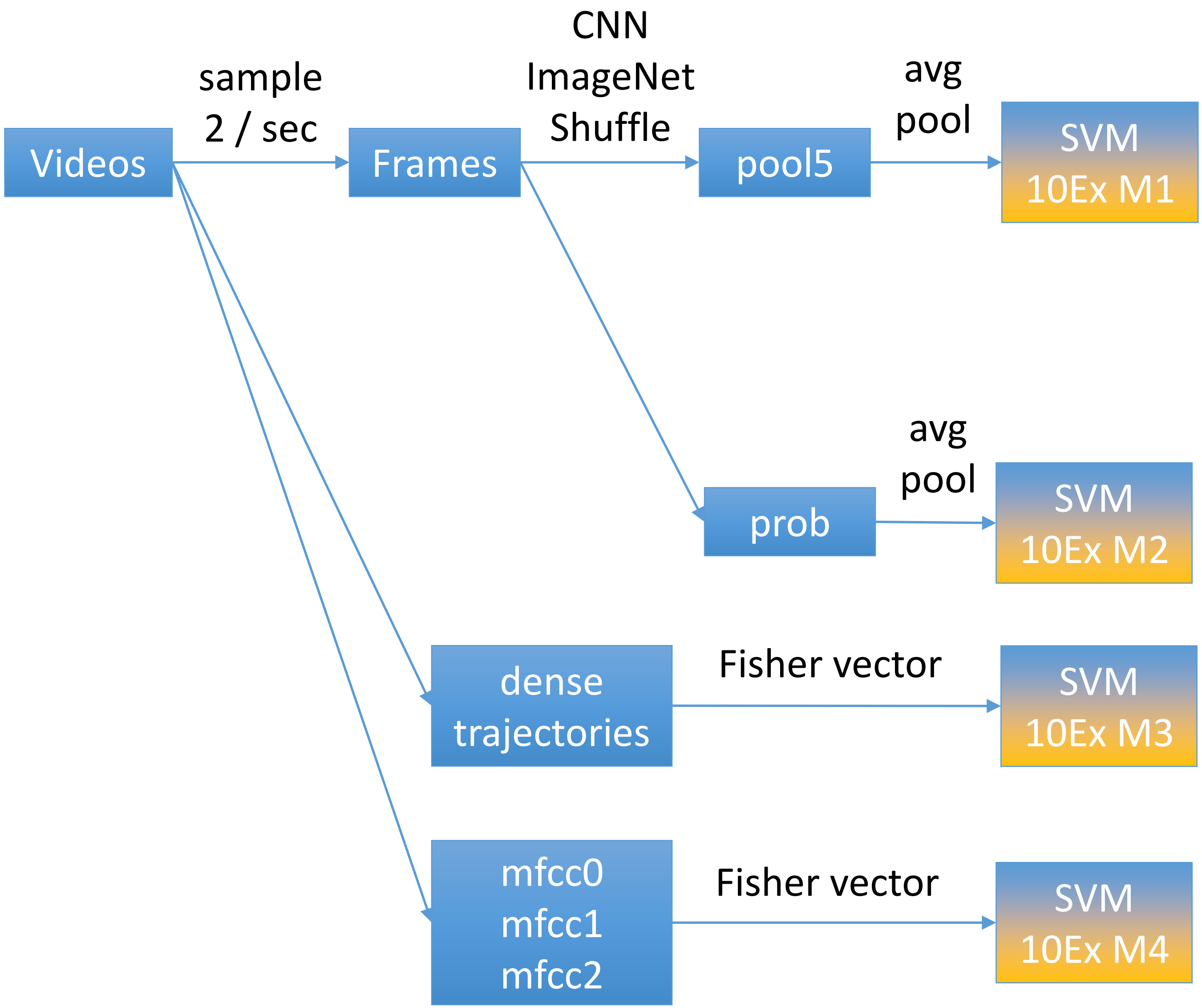
University of Amsterdam

The Netherlands

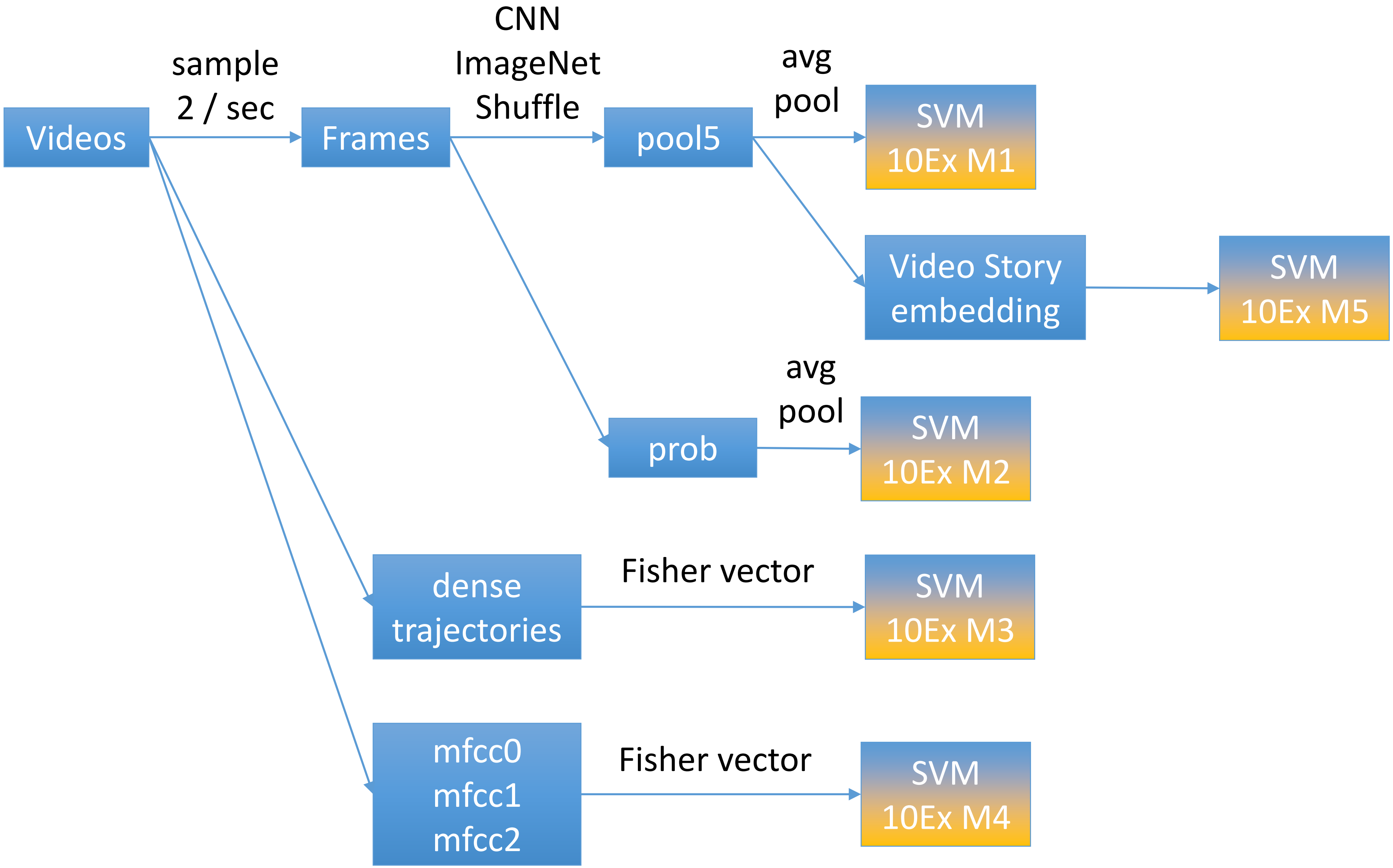
# Overview

- Pipeline 10ex
  - ImageNet shuffle
  - Video Story
  - Results
- Pipeline 0ex
  - Video Story
  - Concepts
  - Results
- Conclusions

# Pipeline 10Ex 2015



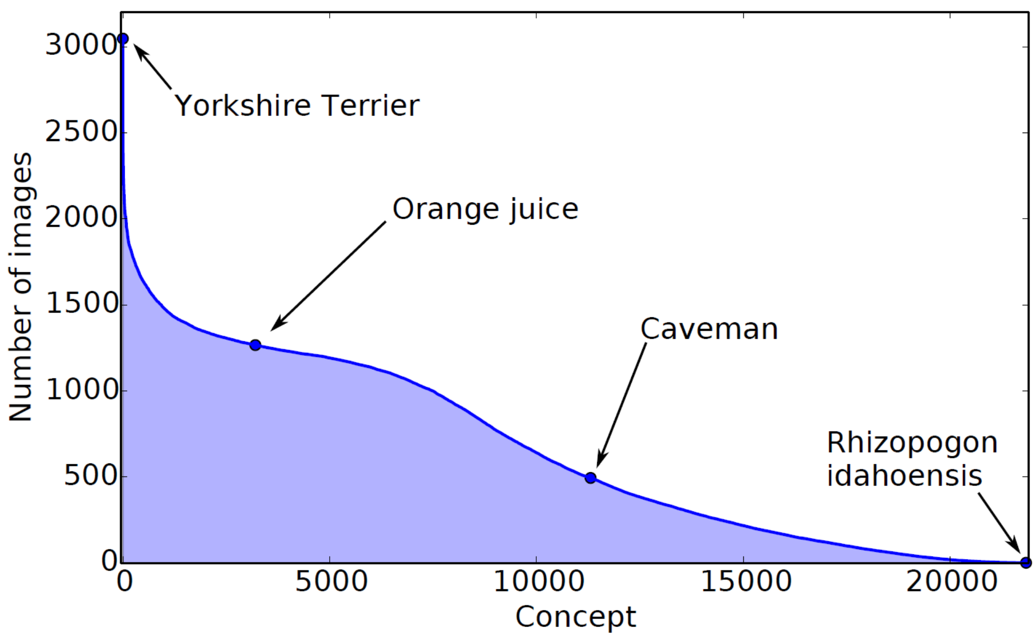
# Pipeline 10Ex 2016



# 22k ImageNet classes

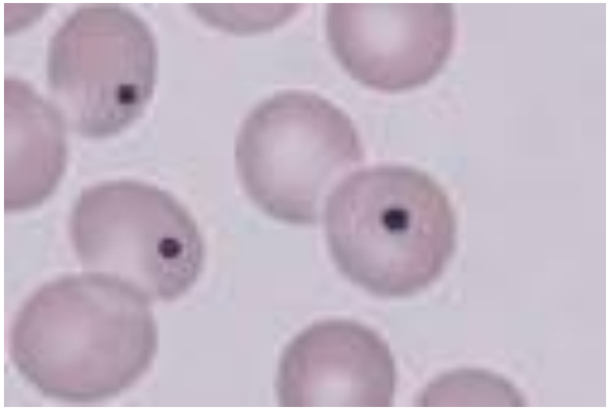
- Use as many classes as possible
- Find a balance between level of abstraction of classes and number of images in a class

Example imbalance




*296 classes with 1 image*

Irrelevant classes

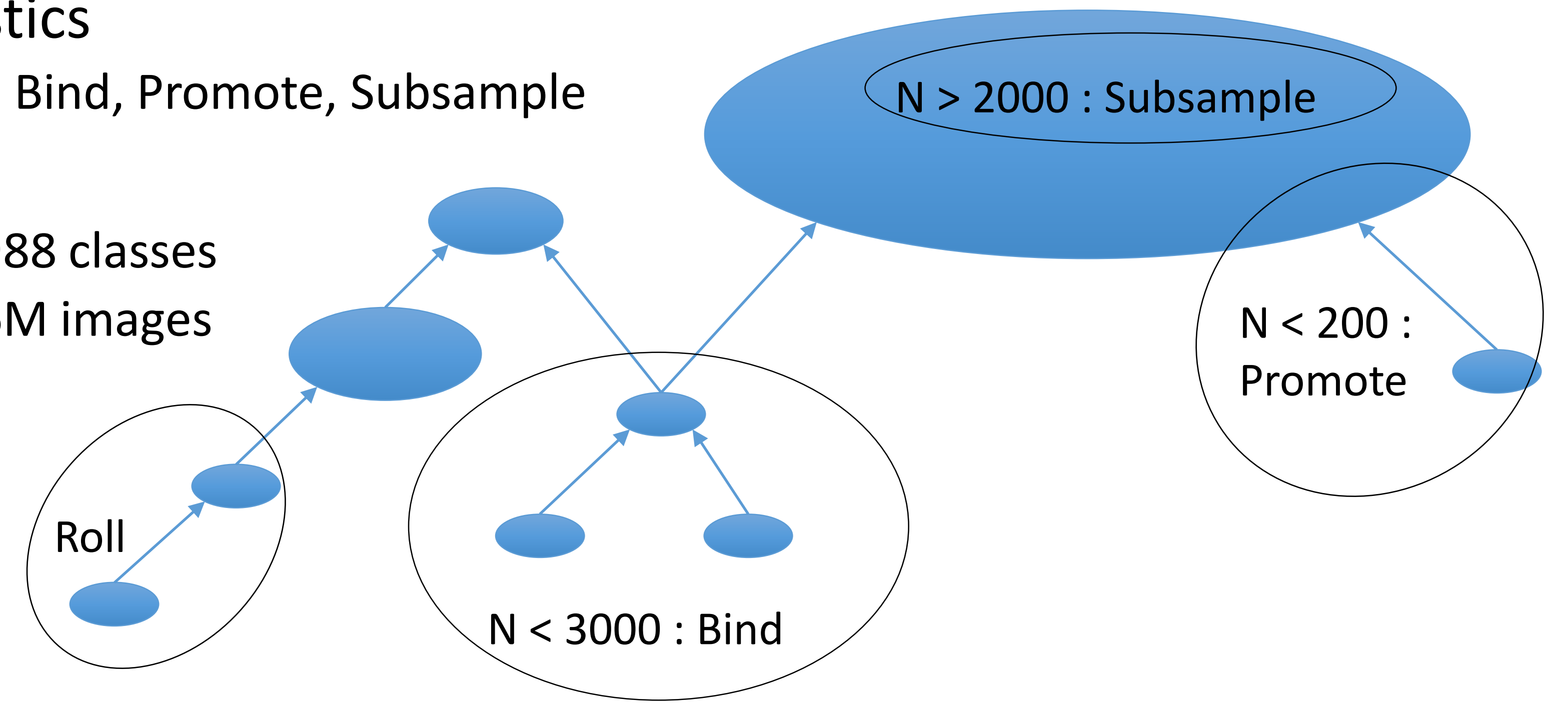


*Siderocyte*

*Gametophyte*

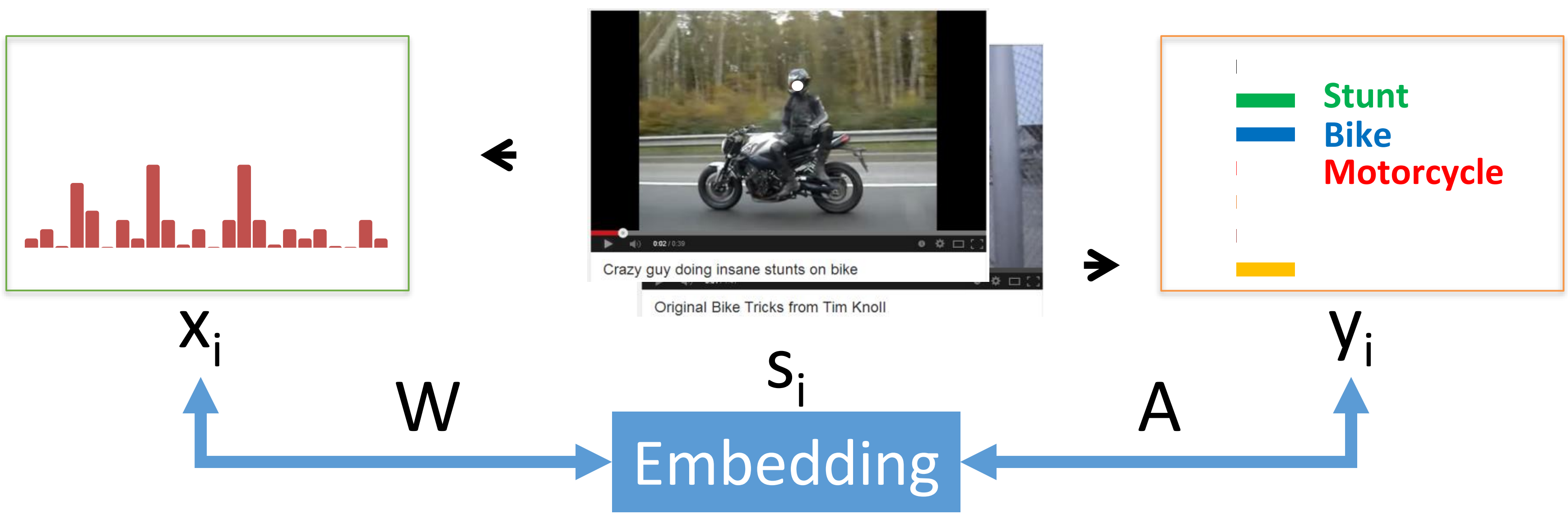
# CNN training on selection out of 22k ImageNet classes

- Idea
  - Increase level of abstraction of classes
  - Incorporate classes with less than 200 samples
- Heuristics
  - Roll, Bind, Promote, Subsample
- Result
  - 12,988 classes
  - 13.6M images




The ImageNet Shuffle: Reorganized Pre-training for Video Event Detection,
Pascal Mettes and Dennis Koelma and Cees Snoek,
International Conference on Multimedia Retrieval, 2016

# Video Story: Embed the story of a video



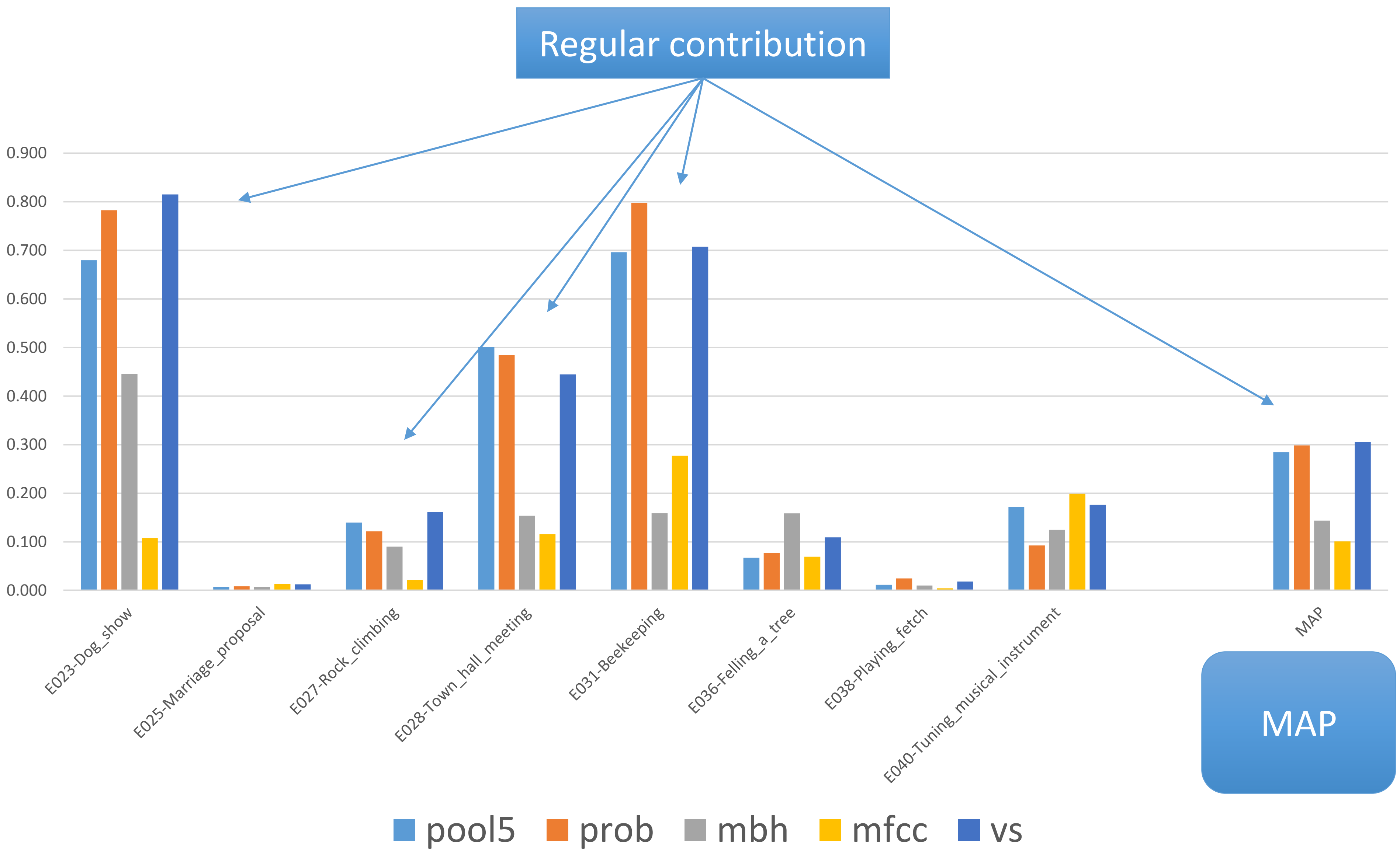

Joint optimization of W and A to preserve

*Descriptiveness:* preserve video descriptions : L(A,S)

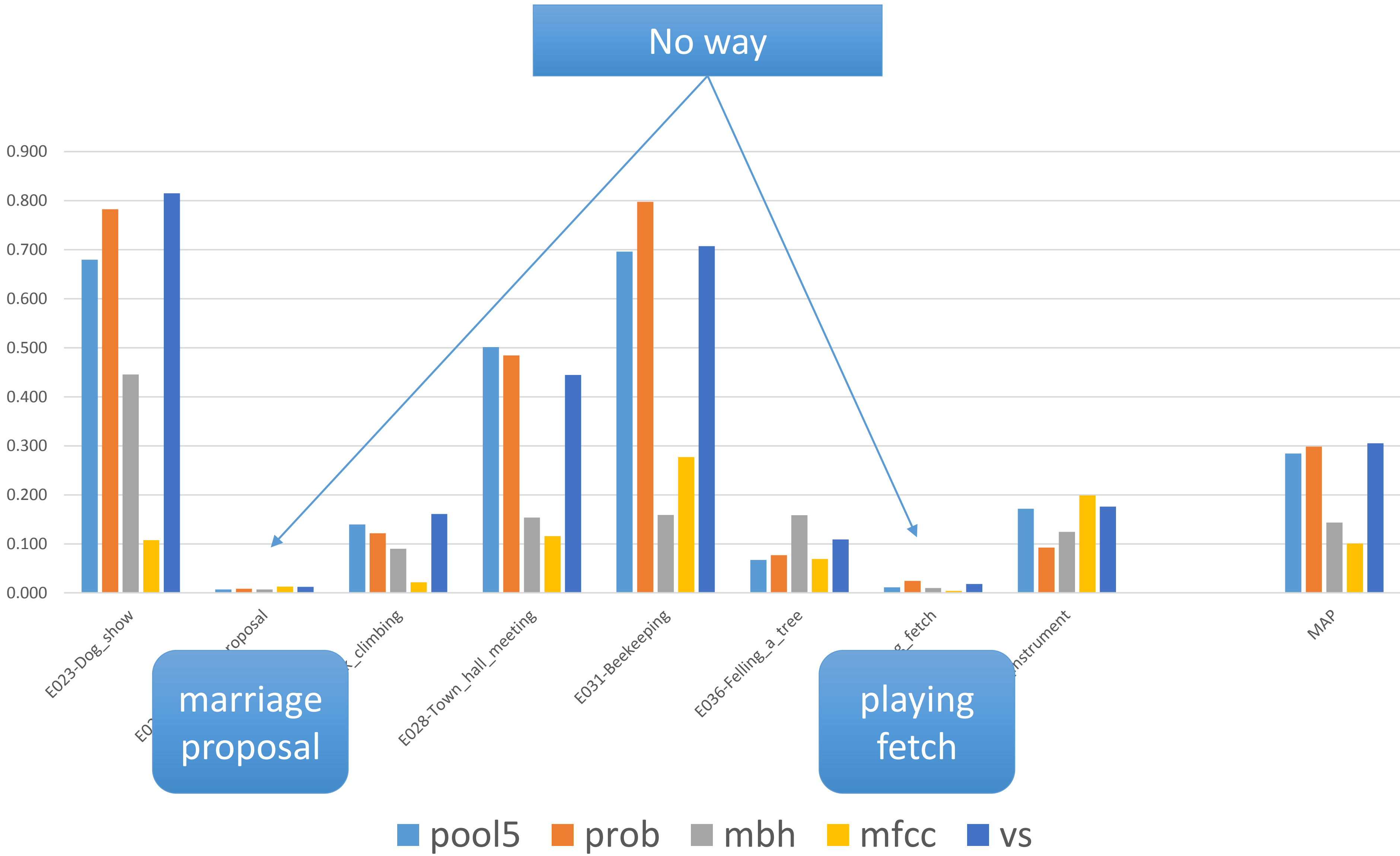
*Predictability:* recognize terms from video content : L(S,W)

Videostory: A new multimedia embedding for few-example recognition and translation of events,
Amirhossein Habibian and Thomas Mensink and Cees Snoek,
Proceedings of the ACM International Conference on Multimedia, 2014

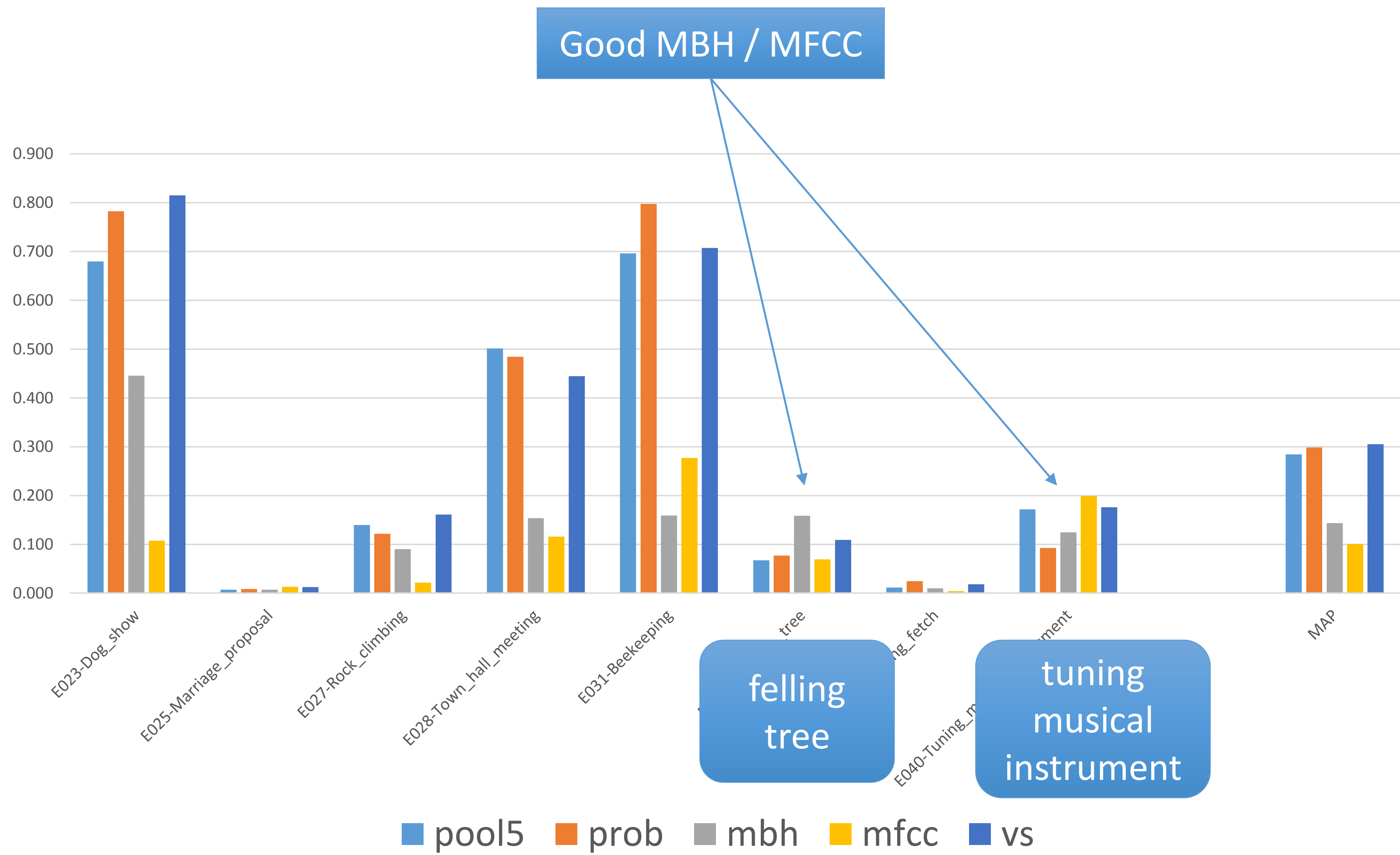
# Results 10Ex Individual Modalities on 2014 Test Set



# Results 10Ex Individual Modalities on 2014 Test Set



# Results 10Ex Individual Modalities on 2014 Test Set

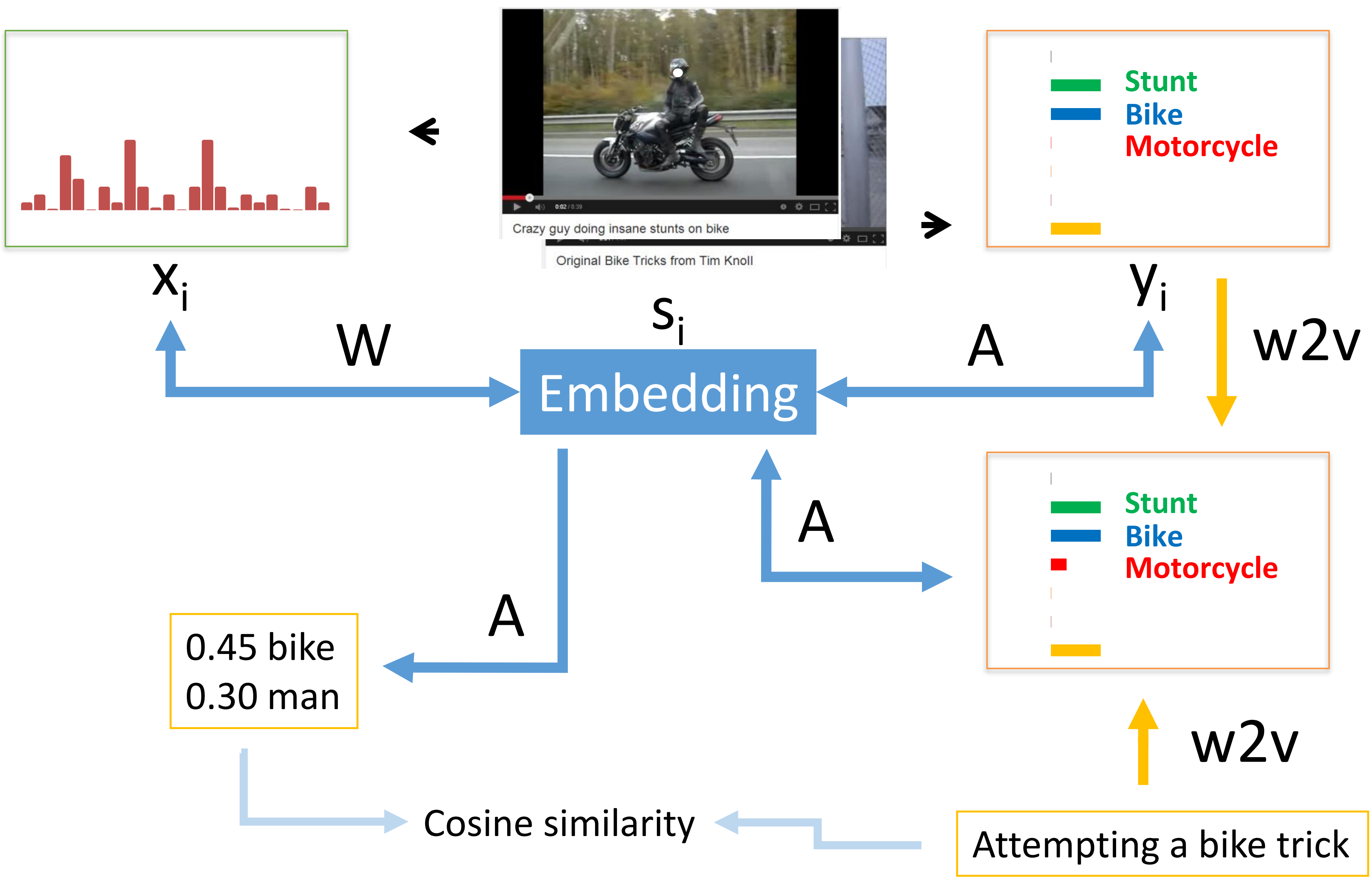


# 10Ex Results

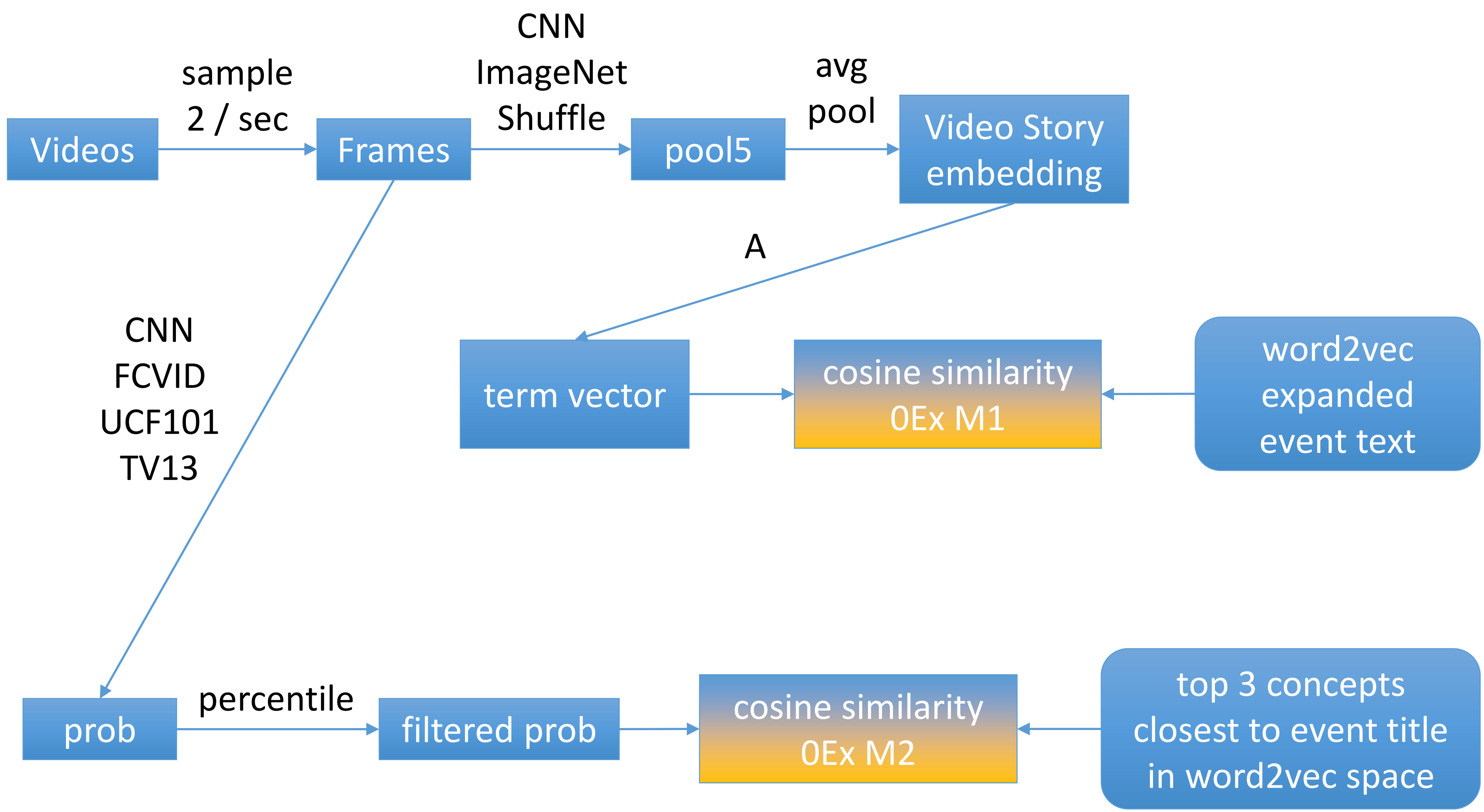
| | PS 2014 | PS 2016 | PS 2016 | AH 2016 | AH 2016 |
|---|---|---|---|---|---|
| | Test | EvalFull | EvalFull | EvalFull | EvalFull |
| | | Progress | | Progress | |
| | MAP | MAP | InfMAP | MAP | InfMAP |
| pool5 | 0.254 | | | | |
| prob | 0.256 | | | | |
| mbh | 0.127 | | | | |
| mfcc | 0.069 | | | | |
| vs | 0.258 | | | | |
| | | | | | |
| pool5 + prob | 0.272 | | | | |
| pool5 + prob + mbh + mfcc (2015) | 0.283 | | | | |
| pool5 + prob + vs | 0.279 | 0.283 | 0.368 | 0.179 | 0.445 |
| pool5 + prob + mbh + mfcc + vs | 0.290 | 0.290 | 0.394 | 0.187 | 0.463 |

- Top performance in 2015 and 2016
- Some progress but not a lot
- We shifted focus to 0Ex

# Video Story for 0Ex



# Pipeline 0Ex



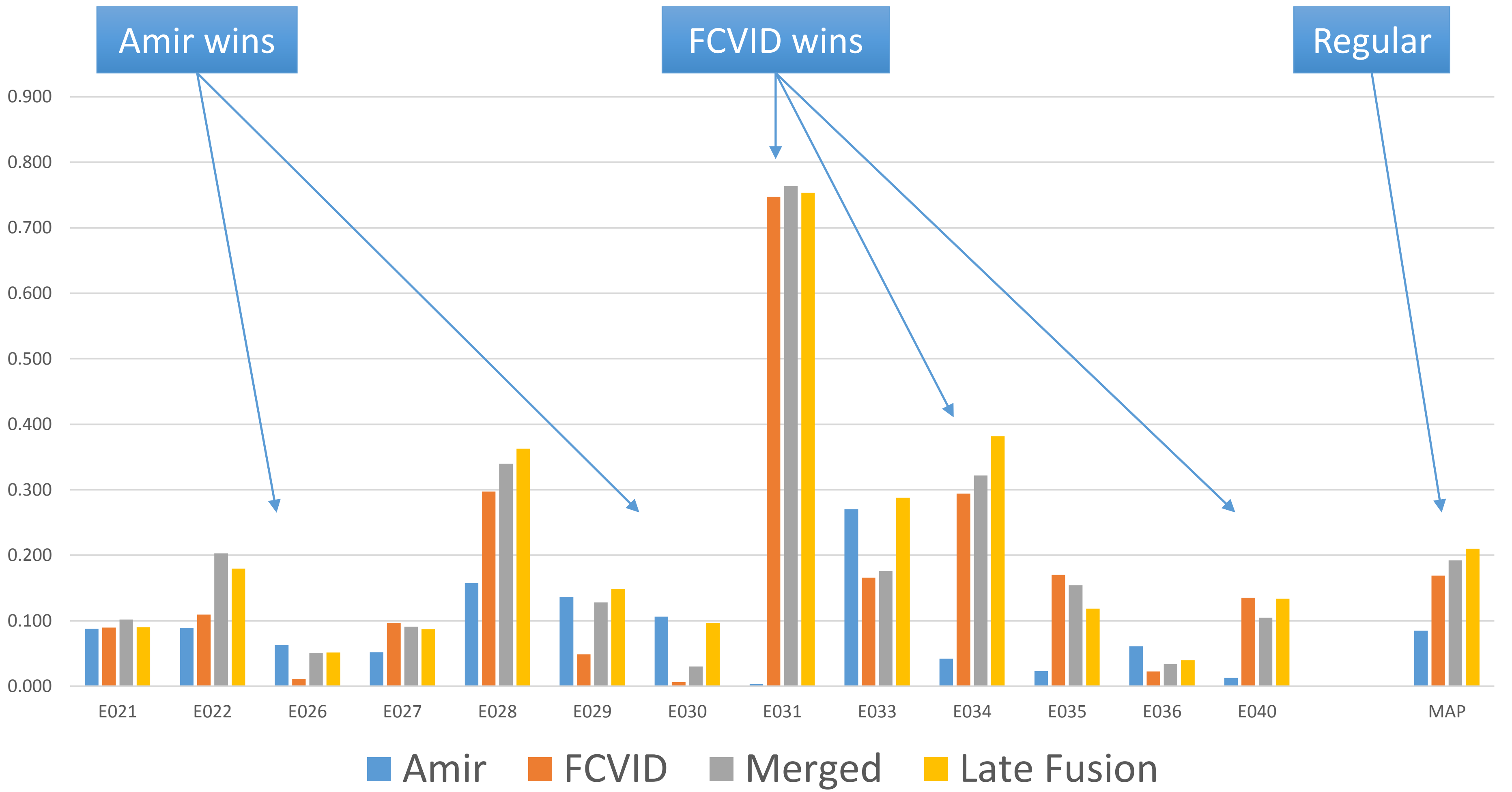
# Video Story Training Sets

- Amir's YouTube46k - www.mediamill.com
  - 45826 videos from YouTube based on 2013 MED research set terms
- FCVID: Fudan Columbia Video Dataset
  - 87609 videos
- Merged

- Video Story dictionary: Terms that occur more than 10 times in the dataset
  - Merged : 5587 terms
- Using vocabulary of stemmed terms that occur more than 100 times in Wikipedia dump
  - With stemming: Respect the Video Story dictionary
  - 267.836 terms
- Use word2vec to expand them per video

# Results 0Ex Video Story on 2014 Test Set

- Fails on 5 events (AP < 0.01), unusable on 2 events (AP < 0.05)

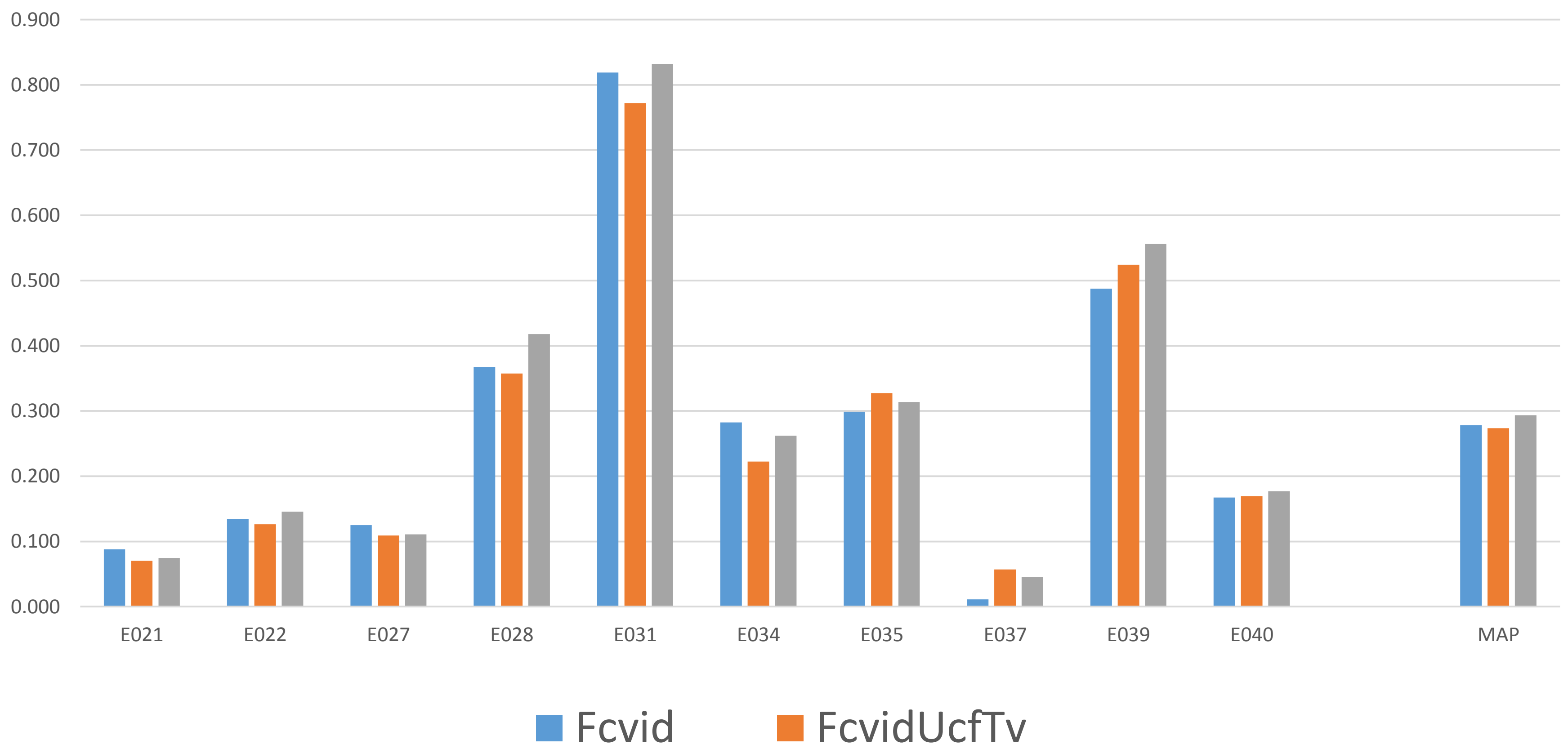


# Concept Bank

- Datasets
  - FCVID
    - 233 concepts
    - Shot segmentation -> max 5 keyframes / video -> max 3000 keyframes / concept (expand within shot if less than 3000)
  - UCF101
    - 101 concepts
    - Shot segmentation -> max 5 keyframes / video -> max 3000 keyframes / concept (expand within shot if less than 3000)
  - TV13 SIN concepts
    - 346 concepts
    - max 3000 keyframes / concept (expand within shot if less than 3000)
- CNN is finetune on ImageNet Shuffle network trained on 13k classes
- Two CNN's:
  - FCVID
  - FCVID + UCF101 + TV13

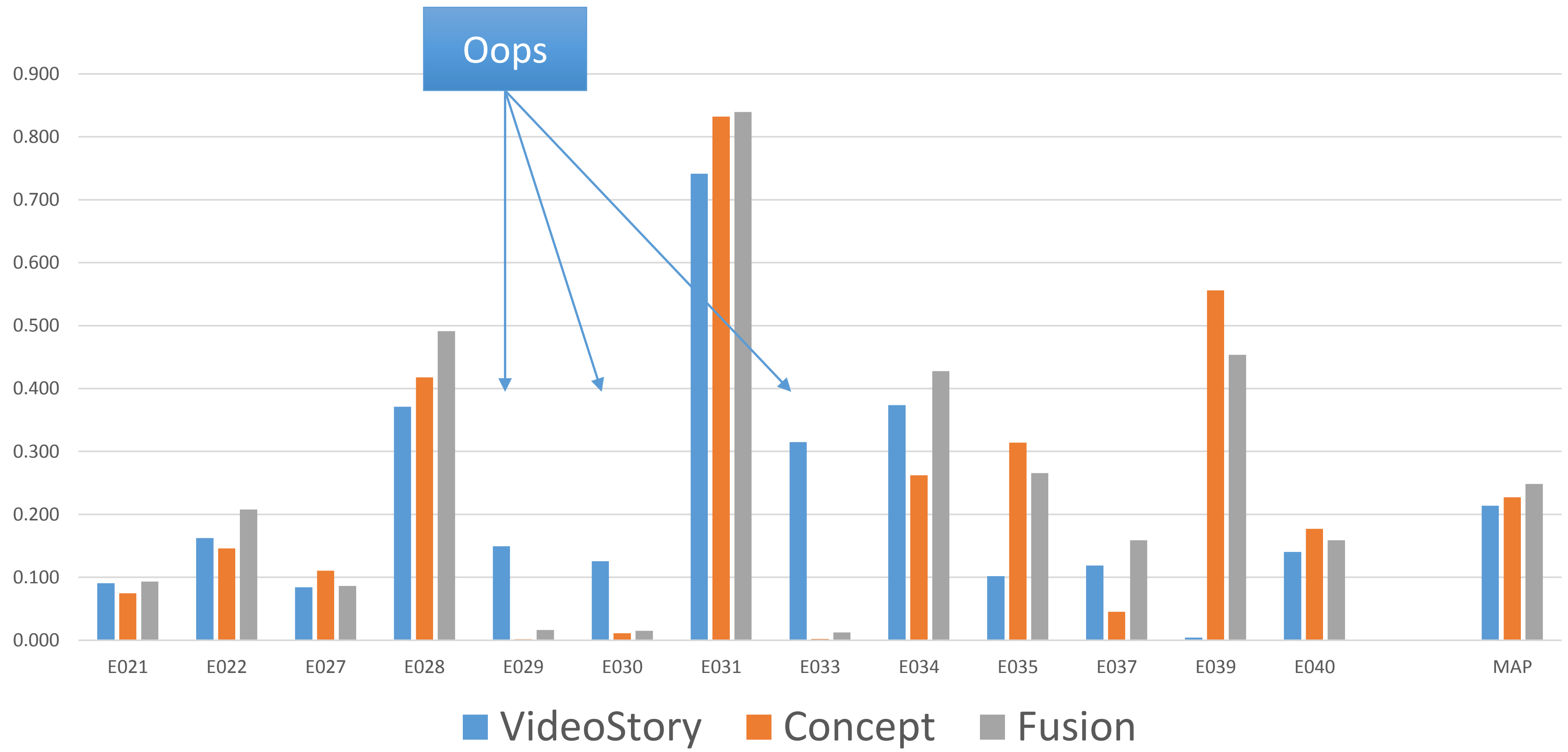
# Results 0Ex Concept Bank on 2014 Test Set

- Fails on 9 events (AP < 0.01), unusable on 1 event (AP < 0.05)
- FCVID CNN is main contributor
- FcvidUcfTv CNN is worse but fusion makes it a bit better overall



# Results 0Ex VideoStory Concept Fusion on 2014 Test Set

- Fails on 3 events (AP < 0.01), unusable on 7 events (AP < 0.05)
- Concepts are slightly better than Video Story but fusion is best



# 0Ex Results

| | 2014 | 2016 | 2016 |
|---|---|---|---|
| | Test | EvalFull | EvalFull |
| | | Progress | |
| | MAP | MAP | InfMAP |
| | | | |
| Video Story Amir | 0.060 | | |
| Video Story FCVID | 0.118 | | |
| Video Story Merged | 0.133 | | |
| concepts FCVID | 0.143 | | |
| concepts FCVID + UCF + TV | 0.140 | | |
| | | | |
| Video Story | 0.146 | | |
| concepts | 0.150 | 0.171 | 0.135 |
| concepts + Video Story | 0.167 (0.175) | 0.181 | 0.149 |

- The “trend” is the same
- Top performance with fully automatic search

# Conclusions

- Video Story for 0Ex benefits from “carefully” selected training material
- Concepts produce higher MAP than Video Story but Video Story is applicable to more events

- Fully automatic video search with just a few examples is becoming feasible
  - 0ex is doable when relevant concepts are present
    - You just have to find them
  - 10ex still makes a big difference

# Thank You

- Video Story - http://www.mediamill.nl
- ImageNet Shuffle CNN’s - http://tinyurl.com/imagenetshuffle