TRECVID 2017

AD-HOC VIDEO SEARCH TASK : OVERVIEW

Georges Quénot
Laboratoire d'Informatique de Grenoble

George Awad
Dakota Consulting, Inc
National Institute of Standards and Technology

Table of contents

- Task Definition
- Video Data
- Topics (Queries)
- Participating teams
- Evaluation & results
- General observation

Ad-hoc Video Search Task Definition

- Goal: promote progress in content-based retrieval based on end user **ad-hoc queries** that include persons, objects, locations, activities and their combinations.

- Task: Given a test collection, a query, and a master shot boundary reference, return a ranked list of at most 1000 shots (out of 335 944) which best satisfy the need.

- Testing data: 4593 Internet Archive videos (IACC.3), 600 total hours with video durations between 6.5 min to 9.5 min.

- Development data: ≈1400 hours of previous IACC data used between 2010-2015 with concept annotations.

Query Development

- Test videos were viewed by 10 human assessors hired by the National Institute of Standards and Technology (NIST).
- 4 facet description of different scenes were used (if applicable):
  - Who : concrete objects and being (kind of persons, animals, things)
  - What : are the objects and/or beings doing ? (generic actions, conditions/state)
  - Where : locale, site, place, geographic, architectural
  - When : time of day, season
- In total assessors watched ≈35% of the IACC.3 videos
- 90 Candidate queries chosen from human written descriptions to be used between 2016-2018.

TV2017 Queries by complexity

- Person + Action + Object + Location
- Find shots of one or more people eating food at a table indoors
- Find shots of one or more people driving snowmobiles in the snow
- Find shots of a man sitting down on a couch in a room
- Find shots of a person talking behind a podium wearing a suit outdoors during daytime
- Find shots of a person standing in front of a brick building or wall

- Person + Action + Location
- Find shots of children playing in a playground
- Find shots of one or more people swimming in a swimming pool
- Find shots of a crowd of people attending a football game in a stadium
- Find shots of an adult person running in a city street

TV2017 Queries by complexity

- Person + Action/state + Object
- Find shots of a person riding a horse including horse-drawn carts
- Find shots of a person wearing any kind of hat
- Find shots of a person talking on a cell phone
- Find shots of a person holding or operating a tv or movie camera
- Find shots of a person holding or opening a briefcase
- Find shots of a person wearing a blue shirt
- Find shots of person holding, throwing or playing with a balloon
- Find shots of a person wearing a scarf
- Find shots of a person holding, opening, closing or handing over a box

- Person + Action
- Find shots of a person communicating using sign language
- Find shots of a child or group of children dancing
- Find shots of people marching in a parade
- Find shots of a male person falling down

TV2017 Queries by complexity

- Person + Object + Location
- Find shots of a man and woman inside a car

- Person + Location
- Find shots of a chef or cook in a kitchen
- Find shots of a blond female indoors

- Person + Object
- Find shots of a person with a gun visible

- Object + Location
- Find shots of a map indoors

- Object
- Find shots of vegetables and/or fruits
- Find shots of a newspaper
- Find shots of at least two planes both visible

Training and run types

Four training data types:

- ✓ A – used only IACC training data (0 runs)
- ✓ D – used any other training data (40 runs)
- ✓ E – used only training data collected automatically using only the query text (12 runs)
- ✓ F – used only training data collected automatically using a query built manually from the given query text (0 runs)

Two run submission types:

- ✓ Manually-assisted (M): Query built manually (19 runs)
- ✓ Fully automatic (F): System uses official query directly(33 runs)

Finishers : 10 out of 20

Team	Organization	M	F
INF	Renmin University; Shandong Normal University; Chongqing university of posts and telecommunications; Carnegie Mellon University	-	4
kobe_nict_siegen	Kobe University, Japan Center for Information and Neural Networks, National Institute of Information and Communications Technology (NICT), Japan Pattern Recognition Group, University of Siegen, Germany	3	-
ITI_CERTH	Information Technologies Institute, Centre for Research and Technology Hellas	-	4
ITEC_UNIKLU	Klagenfurt University	4	4
NII_Hitachi_UIT	National Institute of Informatics, Japan (NII); Hitachi, Ltd; University of Information Technology, VNU-HCM, Vietnam (HCM-UIT)	-	4
MediaMill	University of Amsterdam	-	4
Waseda_Meisei	Waseda University; Meisei University	4	4
VIREO	City University of Hong Kong	4	4
EURECOM	EURECOM	-	4
FIU_UM	Florida International University, University of Miami	4	-

Evaluation

Each query assumed to be binary: absent or present for each master reference shot.

NIST sampled ranked pools and judged top results from all submissions.

Metrics: *inferred average precision per query.*

Compared runs in terms of **mean** *inferred average precision* across the 30 queries.

Mean Extended Inferred Average Precision (XInfAP)

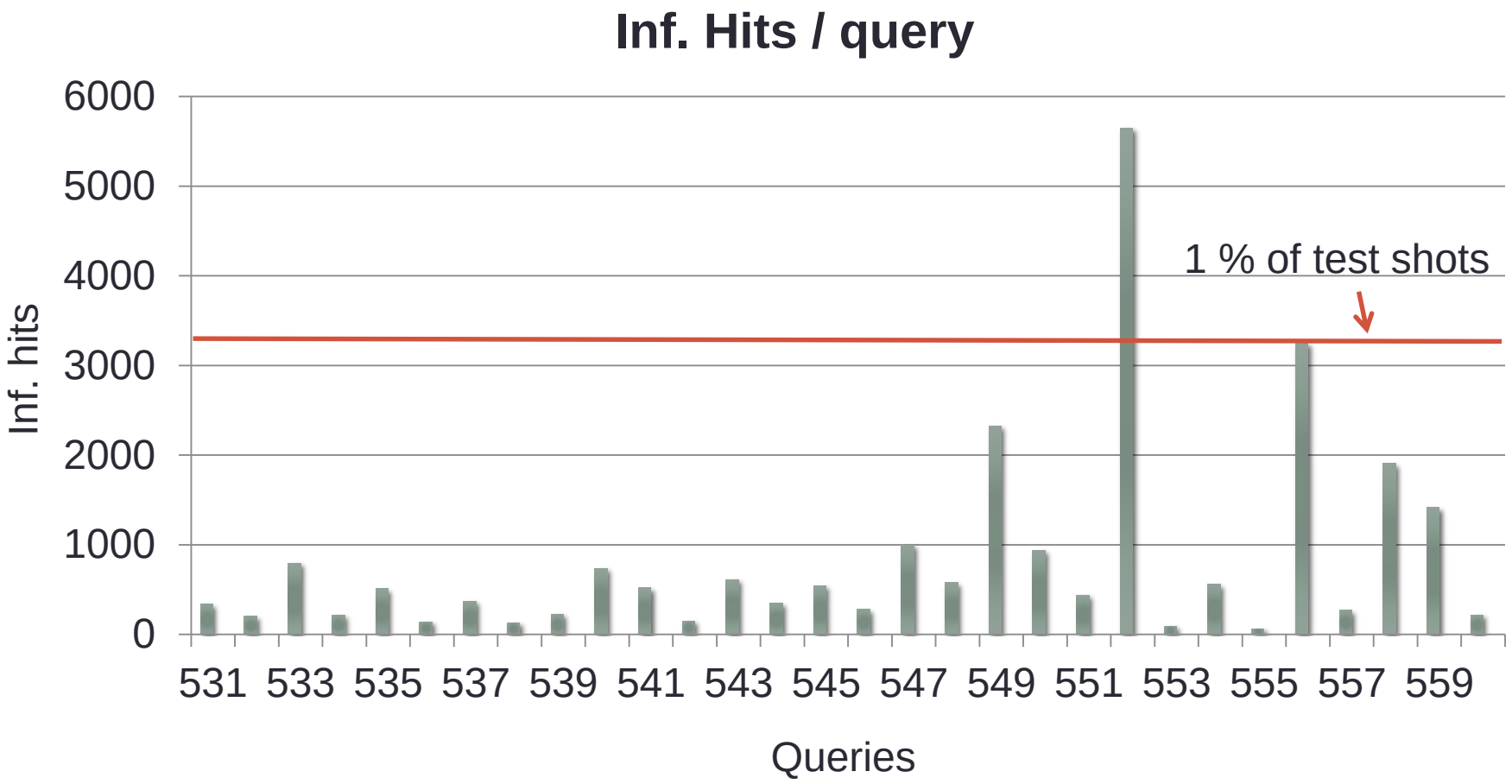
2 pools were created for each query and sampled as:

- ✓ Top pool (ranks 1 to 150) sampled at 100 %
- ✓ Bottom pool (ranks 151 to 1000) sampled at 2.5 %
- ✓ % of sampled and judged clips from rank 151 to 1000 across all runs and topics (min= 2 %, max = 64.4 %, mean = 29 %)

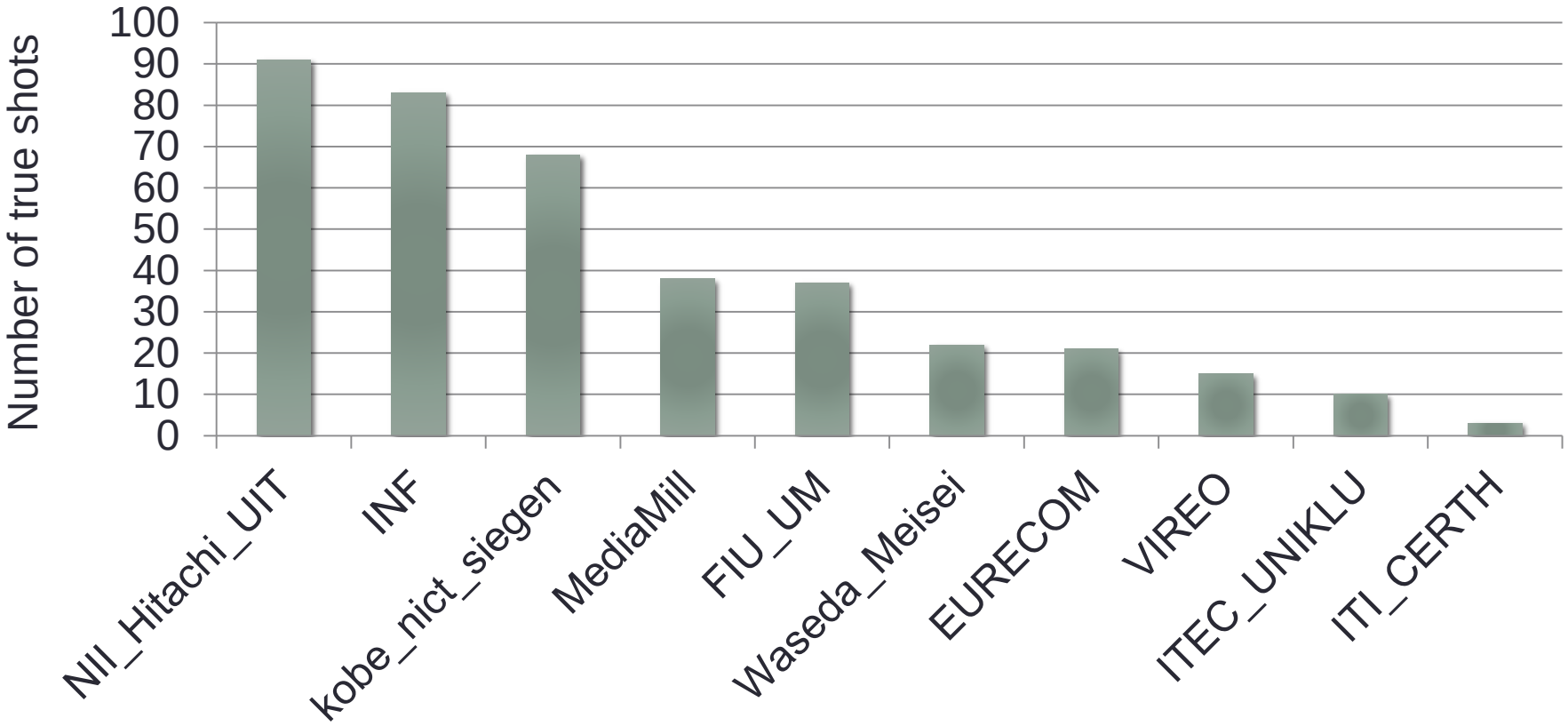
30 queries	
89 435 total judgments	
9611 total hits	>> TV2016
7209 hits at ranks (1 to100)	> TV2016
2013 hits at ranks (101 to 150)	
389 hits at ranks (151 to 1000)	

Judgment process: one assessor per query, watched complete shot while listening to the audio. infAP was calculated using the judged and unjudged pool by sample_eval tool

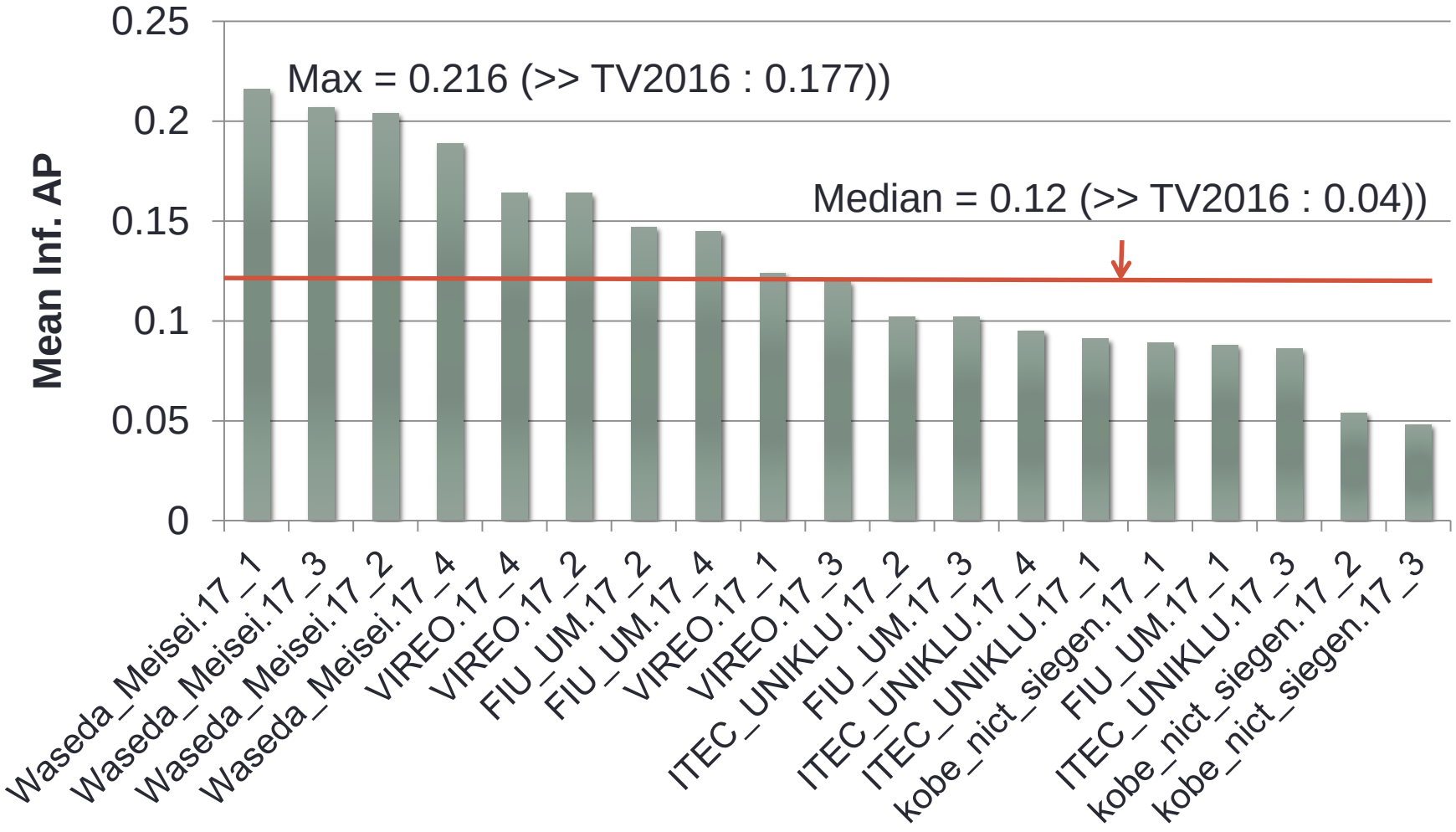
Inferred frequency of hits varies by query



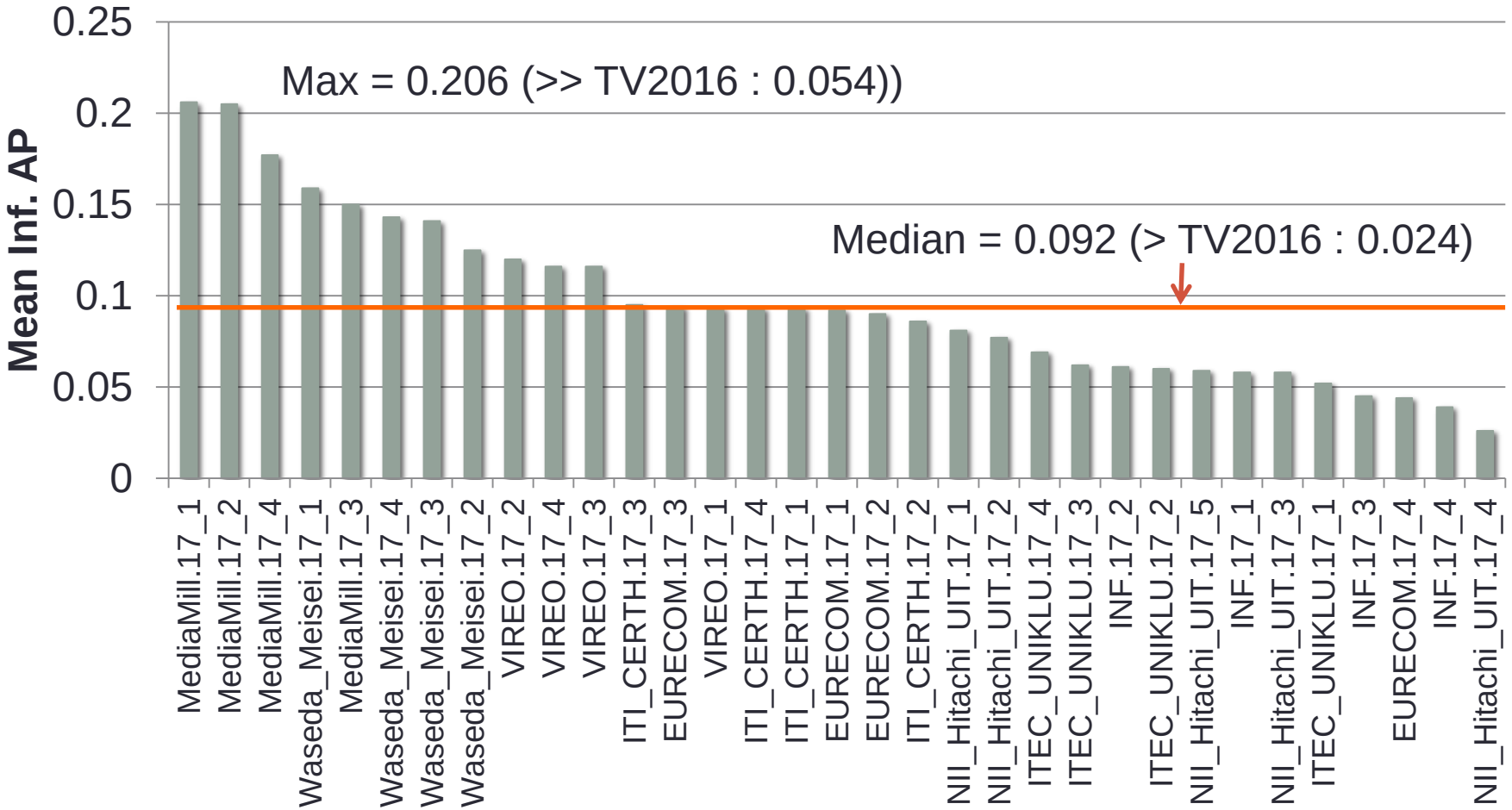
Total true shots contributed uniquely by team



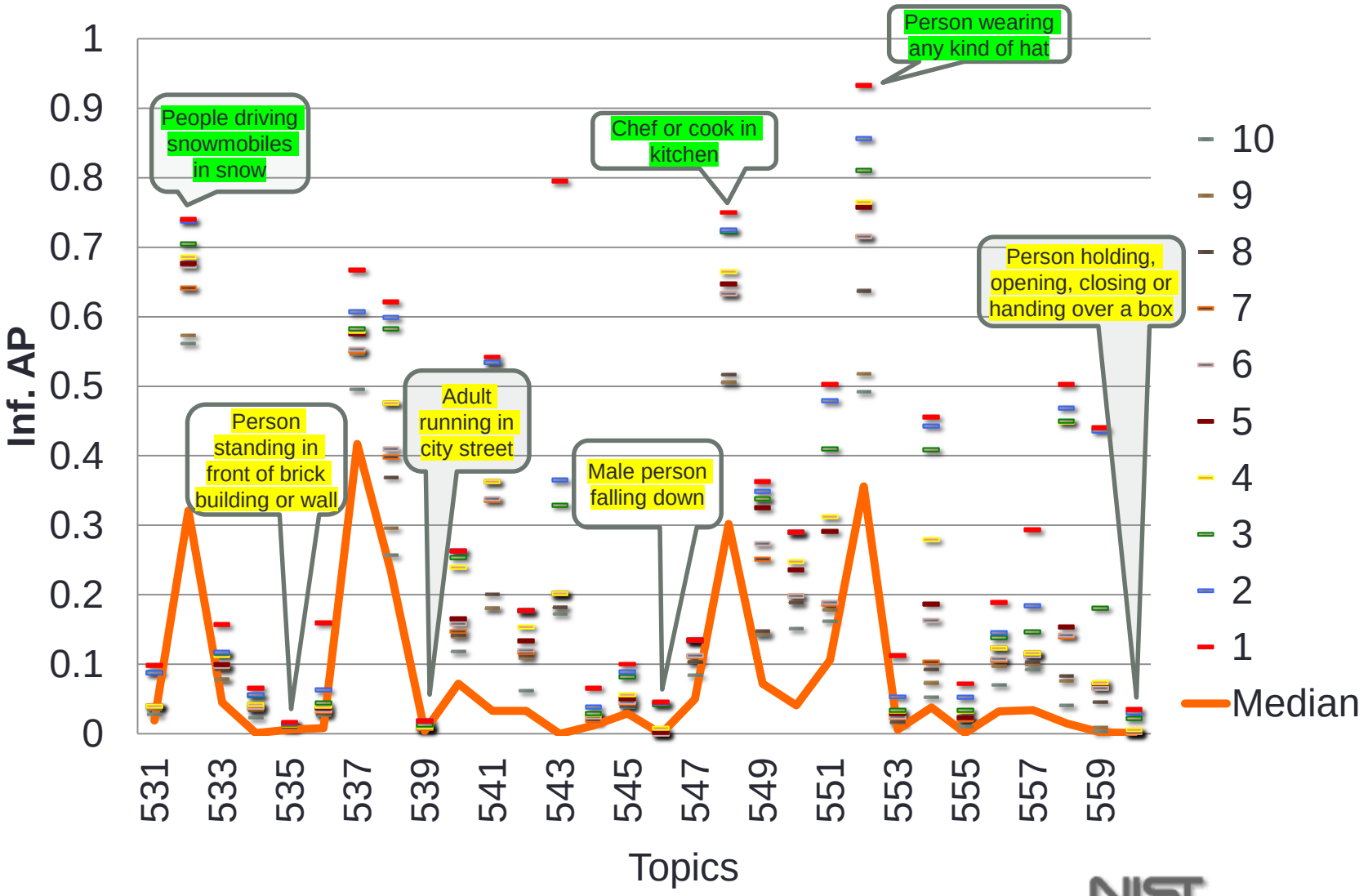
2017 run submissions scores (19 Manually-assisted runs)



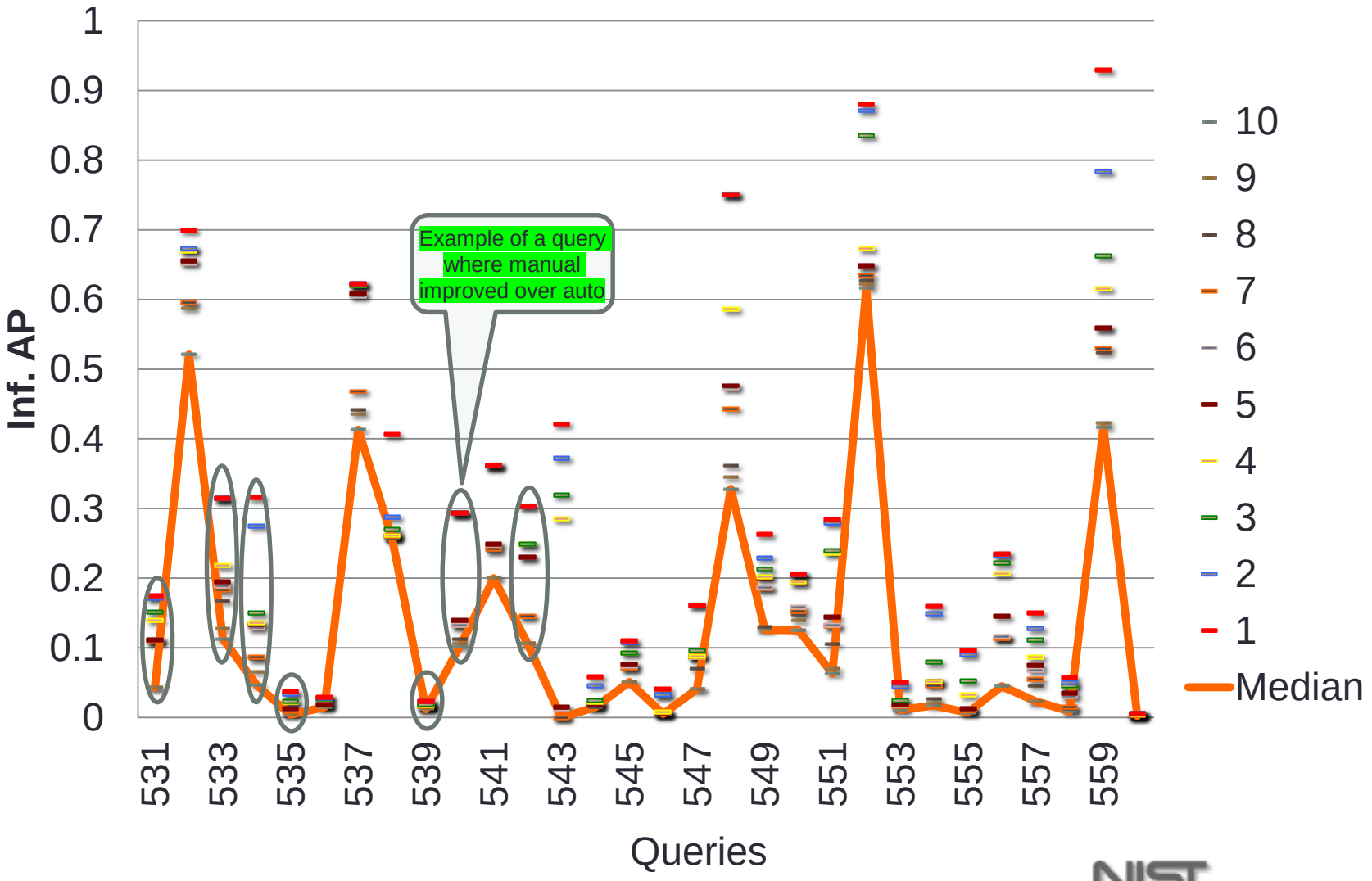
2017 run submissions scores (33 Fully automatic runs)



Top 10 infAP scores by query (Fully automatic)



Top 10 infAP scores by queries (Manually-Assisted)



Which topics where easy or difficult overall ?

Top 10 Easy (sorted by count of runs with InfAP >= 0.7)	Top 10 Hard (sorted by count of runs with InfAP < 0.7)
a person wearing any kind of hat	an adult person running in a city street
a chef or cook in a kitchen	person standing in front of a brick building or wall
one or more people driving snowmobiles in the snow	person holding, opening, closing or handing over a box
one or more people swimming in a swimming pool	a male person falling down
a man and woman inside a car	child or group of children dancing
a crowd of people attending a football game in a stadium	children playing in a playground
a newspaper	person talking on a cell phone
a person communicating using sign language	person holding or opening a briefcase
a person wearing a scarf	one or more people eating food at a table indoor
a person riding a horse including horse-drawn carts	person talking behind a podium wearing a suit outdoors during daytime

More action and dynamics in hard queries

Statistical significant differences among top 10 "M" runs (using randomization test, p < 0.05)

Run	Mean Inf. AP score
D_Waseda_Meisei.17_1	0.216 +
D_Waseda_Meisei.17_3	0.207 +
D_Waseda_Meisei.17_2	0.204 +
D_Waseda_Meisei.17_4	0.189 +
D_VIREO.17_4	0.164 !
D_VIREO.17_2	0.164 !
D_FIU_UM.17_2	0.147 #
D_FIU_UM.17_4	0.145 #
D_VIREO.17_1	0.124 *
D_VIREO.17_3	0.120 *

D_Waseda_Meisei.17_1
- D_VIREO.17_4
  - D_VIREO.17_1
  - D_VIREO.17_3
- D_VIREO.17_2
  - D_VIREO.17_1
  - D_VIREO.17_3
- D_FIU_UM.17_2
- D_FIU_UM.17_4

D_Waseda_Meisei.17_2
- D_VIREO.17_1
- D_VIREO.17_3
- D_FIU_UM.17_2
- D_FIU_UM.17_4

D_Waseda_Meisei.17_3
- D_VIREO.17_4
  - D_VIREO.17_1
  - D_VIREO.17_3
- D_VIREO.17_2
  - D_VIREO.17_1
  - D_VIREO.17_3
- D_FIU_UM.17_2
- D_FIU_UM.17_4

D_Waseda_Meisei.17_4
- D_VIREO.17_1
- D_VIREO.17_3
- D_FIU_UM.17_4

+!#* : no significant difference among each set of runs

- Runs higher in the hierarchy are significantly better than runs more indented.

Statistical significant differences among top 10 "F" runs (using randomization test, p < 0.05)

Run	Mean Inf. AP score
D_MediaMill.17_1	0.206 +
D_MediaMill.17_2	0.205 +
D_MediaMill.17_4	0.177
D_Waseda_Meisei.17_1	0.159
D_MediaMill.17_3	0.150
D_Waseda_Meisei.17_4	0.143 #
D_Waseda_Meisei.17_3	0.141 #
D_Waseda_Meisei.17_2	0.125
D_VIREO.17_2	0.120 *
D_VIREO.17_4	0.116 *
D_VIREO.17_3	0.116 *

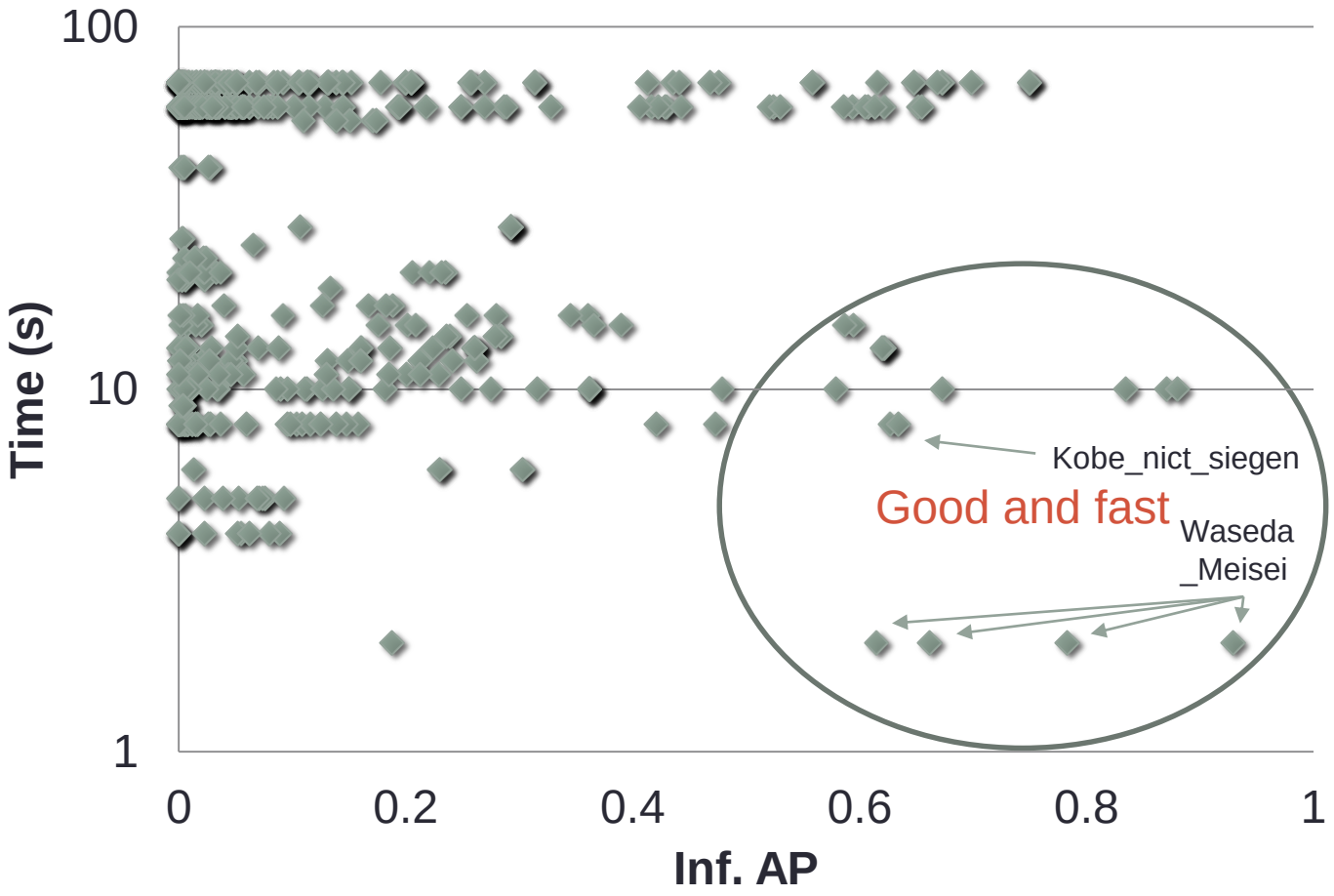
+#* : no significant difference among each set of runs

- Runs higher in the hierarchy are significantly better than runs more indented.

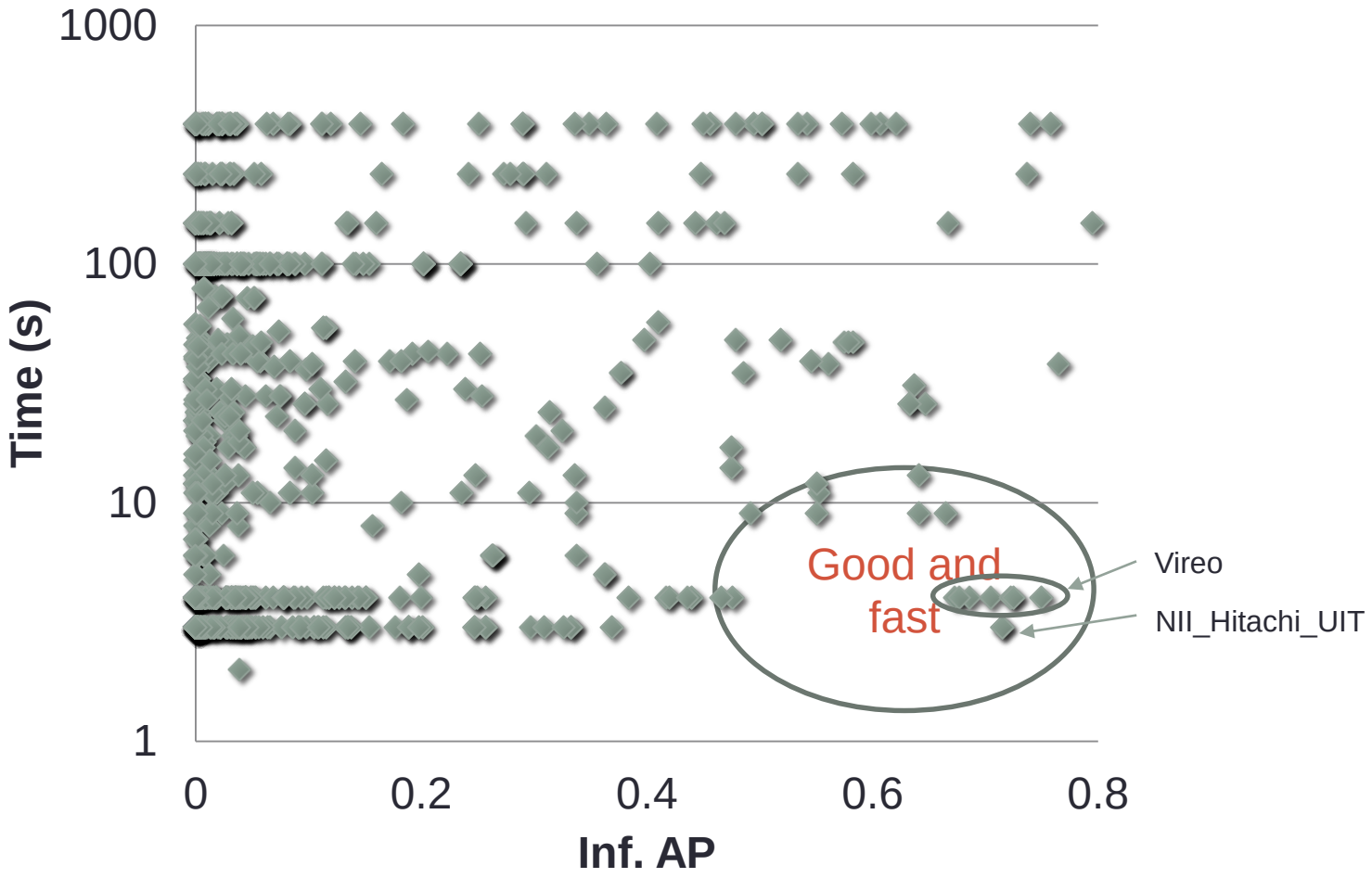
D_MediaMill.17_1
- D_MediaMill.17_4
- D_VIREO.17_2
- D_VIREO.17_3
- D_VIREO.17_4
- D_Waseda_Meisei.17_1
  - D_Waseda_Meisei.17_2
- D_Waseda_Meisei.17_3
- D_Waseda_Meisei.17_4

D_MediaMill.17_2
- D_MediaMill.17_4
- D_VIREO.17_2
- D_VIREO.17_3
- D_VIREO.17_4
- D_Waseda_Meisei.17_1
  - D_Waseda_Meisei.17_2
- D_Waseda_Meisei.17_3
- D_Waseda_Meisei.17_4

Processing time vs Inf. AP (“M” runs)
Across all topics and runs



Processing time vs Inf. AP (“F” runs)
Across all topics and runs



2017 Main Approaches

- Concept bank with automatic or manual mapping with query terms
- Combination of concept scores from Boolean operators
- Work on Query Understanding
- Rectified Linear Score Normalization
- Use of Video-To-Text techniques on shots
- Query expansion / term matching techniques
- Use of unified text-image vector space

2017 Observations

- Ad-hoc search is more difficult than simple concept-based tagging.
- Max and Median scores are better than TV2016 for both M and F runs.
- Manually-assisted runs performed slightly better than fully-automatic.
- Most systems are not real-time (slower systems were not necessarily effective).
- Some systems reported 0 time!!! (or didn't measure it!)
- There was 0 A and F runs submitted compared to D and E

Continued at MMM2018

- 10 Ad-Hoc Video Search (AVS) tasks, 5 of which are a random subset of the 30 AVS tasks of TRECVID 2017 and 5 will be chosen directly by human judges as a surprise. Each AVS task has several/many target shots that should be found.
- 10 Known-Item Search (KIS) tasks, which are selected completely random on site. Each KIS task has only one single 20 s long target segment.
- Registration for the task is now closed

9:20 - 11:40 : Ad-hoc Video Search

- **9:40 - 10:00**, Query understanding is key for zero-example video search (**MediaMill - University of Amsterdam**)

- **10:00 - 10:20**, Waseda_Meisei at TRECVID 2017: Ad-hoc video search (**Waseda_Meisei - Waseda University; Meisei University**)

10:20 - 10:40, **Break** with refreshments

- **10:40 - 11:00**, FIU-UM@TRECVID 2017: Rectified Linear Score Normalization and Weighted Integration for Ad-hoc Video Search (**FIU_UM - Florida International University, University of Miami**)

- **11:00 - 11:20**, Interactive Video Search at VBS (**ITEC_UNIKLU -Institute of Information Technology, Klagenfurt University**)

- **11:20 - 11:40**, AVS discussion

2017 Questions

- Was the task/queries realistic enough?!
- Do we need to change/add/remove anything from the task in 2018 ?
- Is there any specific reason why systems did not submit any "F" runs? (training data collected automatically using a query built manually from the given query text)
- Did any team run their 2017 system on TV2016 topics or 2016 system on this year's topics ?

- Should we consider new dataset in 2019 to continue working on Ad-hoc ? (e.g YouTube, Vimeo, etc)