TRECVID 2016
AD-HOC VIDEO SEARCH TASK : OVERVIEW

Georges Quénot
Laboratoire d'Informatique de Grenoble

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

NIST
National Institute of Standards and Technology

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.
- New 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.

National Institute of Standards and Technology

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.

TV2016 Queries samples by complexity

- Person + Action + Object + Location

Find shots of a person playing guitar outdoors.

Find shots of a man indoors looking at camera where a bookcase is behind him.

Find shots of a person playing drums indoors.

Find shots of a diver wearing diving suit and swimming under water.

- Person + Action + Location

Find shots of the 43rd president George W. Bush sitting down talking with people indoors.

Find shots of a choir or orchestra and conductor performing on stage.

Find shots of one or more people walking or bicycling on a bridge during daytime.

TV2016 Queries by complexity

- Person + Action/state + Object

Find shots of a person sitting down with a laptop visible.
Find shots of a man with beard talking or singing into a microphone.
Find shots of one or more people opening a door and exiting through it.
Find shots of a person holding a knife.
Find shots of a woman wearing glasses.
Find shots of a person drinking from a cup, mug, bottle, or other container.
Find shots of a person wearing a helmet.
Find shots of a person lighting a candle.

- Person + Action

Find shots of people shopping.
Find shots of soldiers performing training or other military maneuvers.
Find shots of a person jumping.
Find shots of a man shake hands with a woman.

TV2016 Queries by complexity

- Person + Location

Find shots of one or more people at train station platform.

Find shots of two or more men at a beach scene.

- Person + Object

Find shots of a policeman where a police car is visible.

- Object + Location

Find shots of any type of fountains outdoors.

- Object

Find shots of a sewing machine.

Find shots of destroyed buildings.

Find shots of palm trees.

Training and run types

Four training data types:

- ✓ A – used only IACC training data (4 runs)
- ✓ D – used any other training data (42 runs)
- ✓ E – used only training data collected automatically using only the query text (6 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
- ✓ Fully automatic (F) – System uses official query directly

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 200) sampled at 100 %
- ✓ Bottom pool (ranks 201 to 1000) sampled at 11.1 %
- ✓ % of sampled and judged clips from rank 201 to 1000 across all runs (min= 10.5 %, max = 76 %, mean = 35 %)

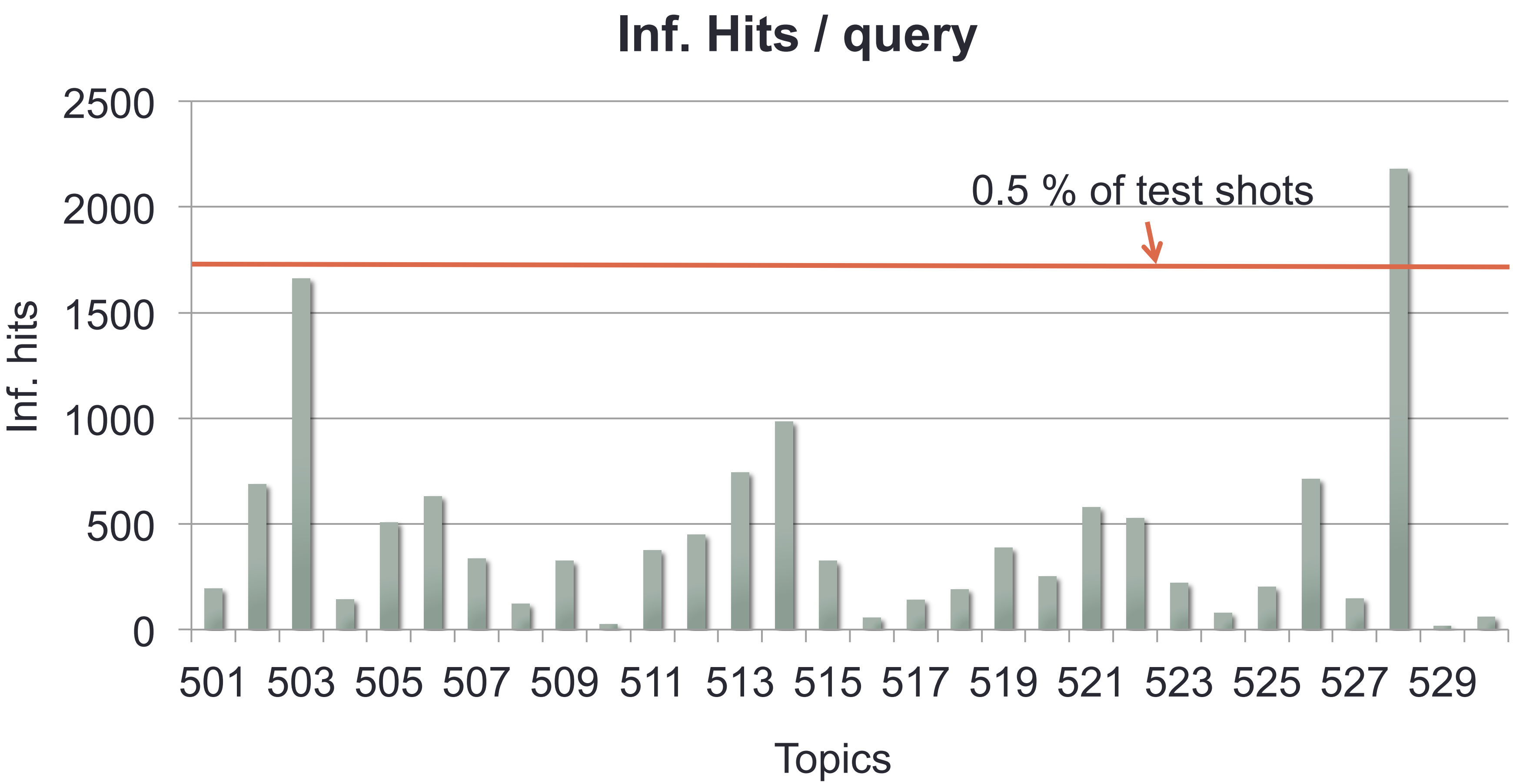
30 queries
187 918 total judgments
7448 total hits
4642 hits at ranks (1 to100)
2080 hits at ranks (101 to200)
726 hits at ranks (201 to 2000)

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

Finishers : 13 out of 29

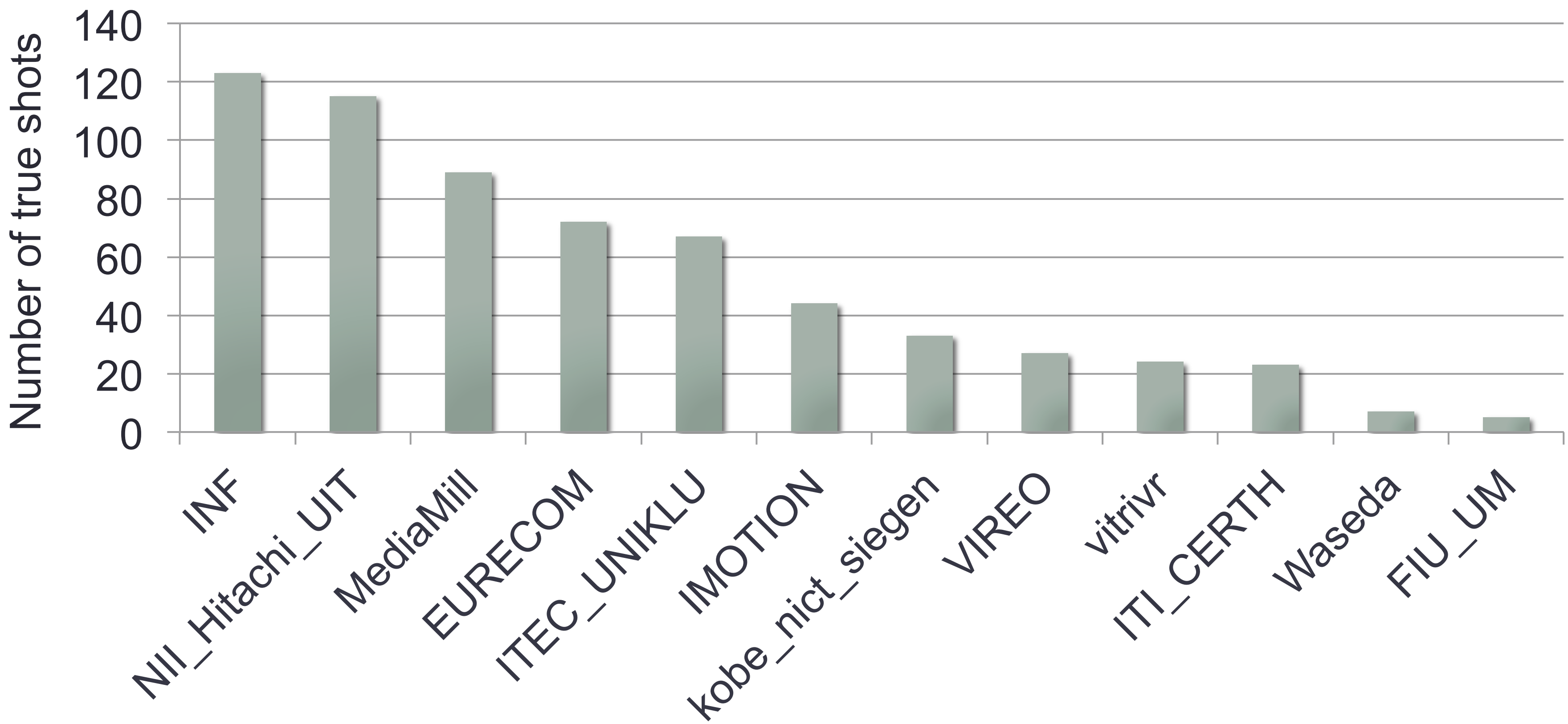
		M	F
INF	CMU; Beijing University of Posts and Telecommunication; University Autonoma de Madrid; Shandong University; Xian JiaoTong University Singapore	-	4
kobe_nict_siegen	Kobe University, Japan; National Institute of Information and Communications Technology, Japan; University of Siegen, Germany	3	-
UEC	Dept. of Informatics, The University of Electro-Communications, Tokyo	2	-
ITI_CERTH	Inf. Tech. Inst., Centre for Research and Technology Hellas	4	4
ITEC_UNIKLU	Klagenfurt University	-	3
NII_Hitachi_UIT	Natl. Inst. Of Info.; Hitachi Ltd; University of Inf. Tech. (HCM-UIT)	-	4
IMOTION	University of Basel, Switzerland; University of Mons, Belgium; Koc University, Turkey	2	2
MediaMill	University of Amsterdam Qualcomm	-	4
Vitrivr	University of Basel	2	2
Waseda	Waseda University	4	-
VIREO	City University of Hong Kong	3	3
EURECOM	EURECOM	-	4
FIU_UM	Florida International University, University of Miami	2	-

Inferred frequency of hits varies by query

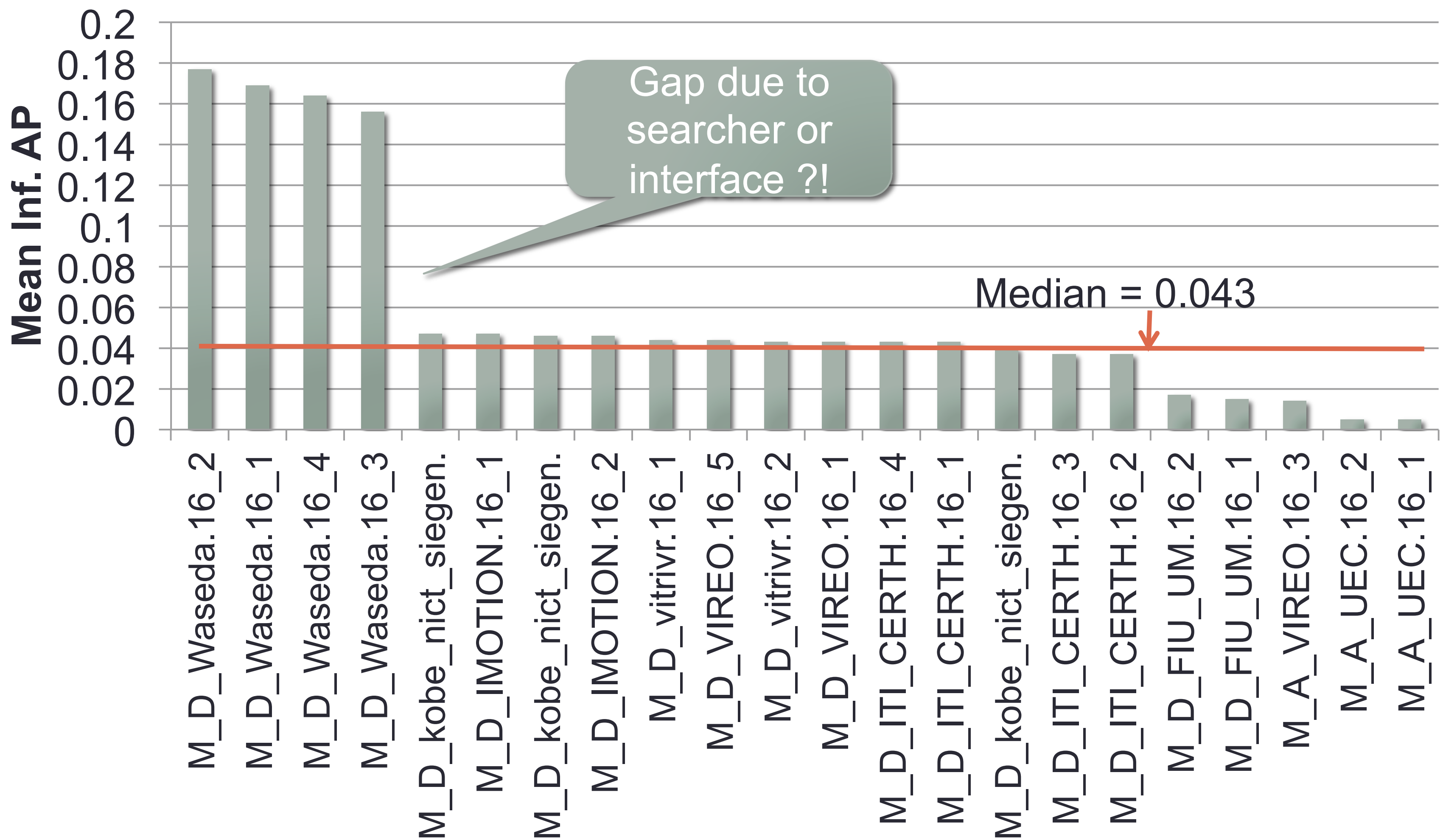


NIST National Institute of Standards and Technology

Total true shots contributed uniquely by team

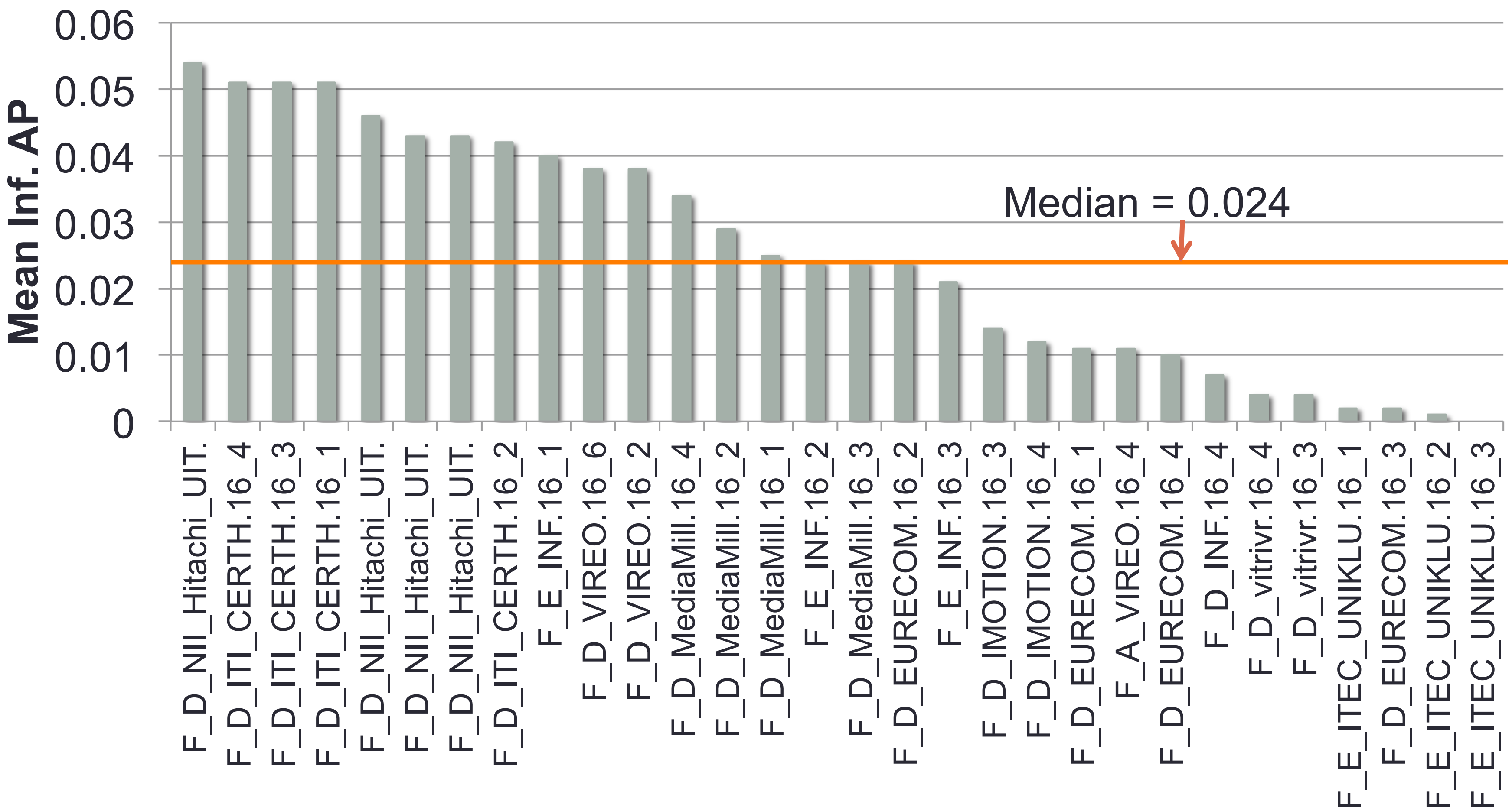


2016 run submissions scores (22 Manually-assisted runs)

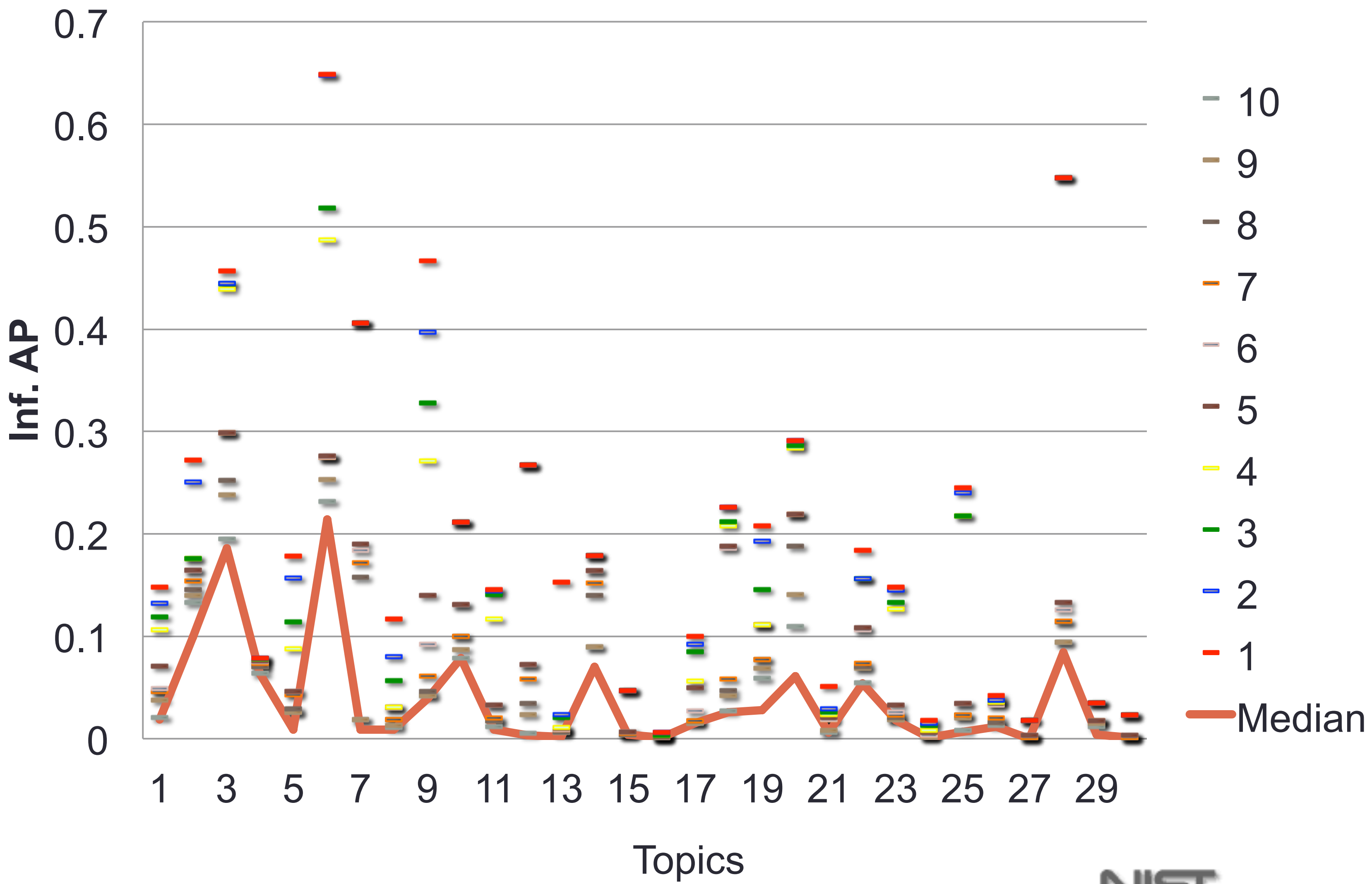


NIST National Institute of Standards and Technology

2016 run submissions scores (30 Fully automatic runs)

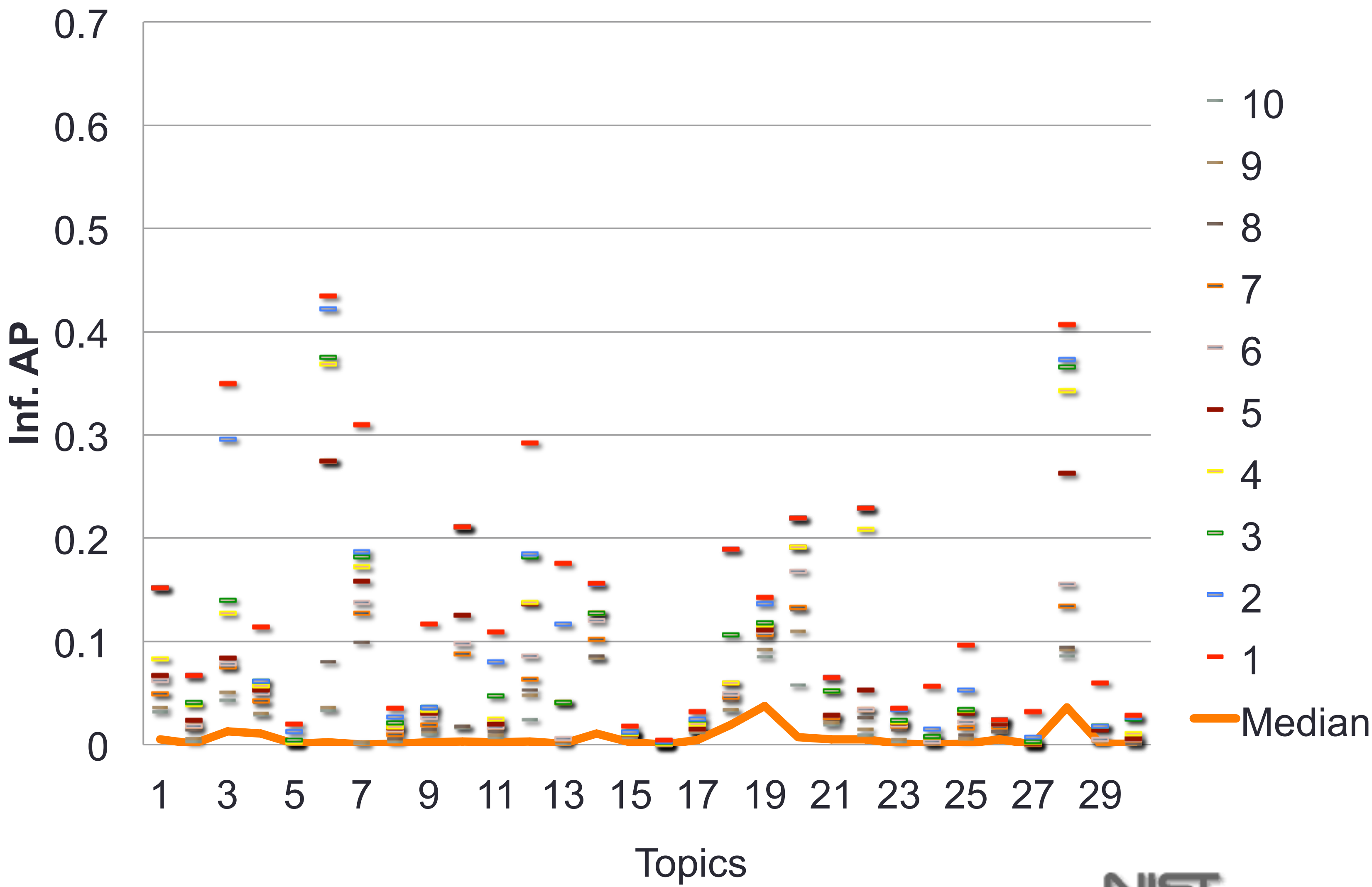


Top 10 infAP scores by query (Manually-assisted)



NIST National Institute of Standards and Technology

Top 10 infAP scores by query (Fully automatic)



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

D_Waseda.16_2
- D_Waseda.16_3
  - D_kobe_nict_siegen.16_3
  - D_kobe_nict_siegen.16_1
  - D_IMOTION.16_1
  - D_IMOTION.16_2
  - D_vitrivr.16_1
  - D_VIREO.16_5
- D_Waseda.16_4
  - D_kobe_nict_siegen.16_3
  - D_kobe_nict_siegen.16_1
  - D_IMOTION.16_1
  - D_IMOTION.16_2
  - D_vitrivr.16_1
  - D_VIREO.16_5

D_Waseda.16_1
- D_Waseda.16_3
  - D_kobe_nict_siegen.16_3
  - D_kobe_nict_siegen.16_1
  - D_IMOTION.16_1
  - D_IMOTION.16_2
  - D_vitrivr.16_1
  - D_VIREO.16_5

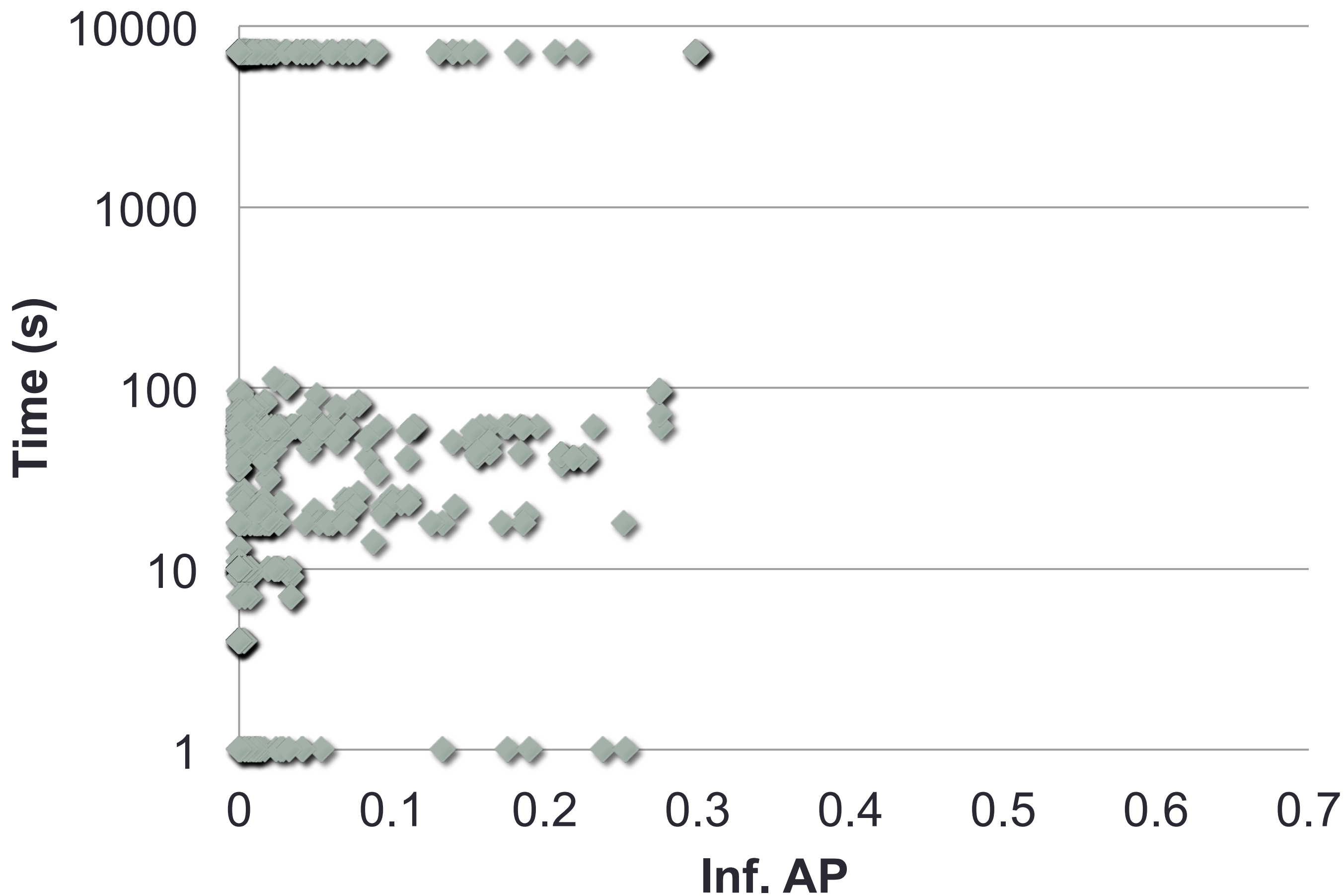
Run	Inf. AP score
D_Waseda.16_2	0.177 *
D_Waseda.16_1	0.169 *
D_Waseda.16_4	0.164 #
D_Waseda.16_3	0.156 #
D_kobe_nict_siegen.16_3	0.047 ^
D_IMOTION.16_1	0.047 ^
D_kobe_nict_siegen.16_1	0.046 ^
D_IMOTION.16_2	0.046 ^
D_vitrivr.16_1	0.044 ^
D_VIREO.16_5	0.044 ^

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

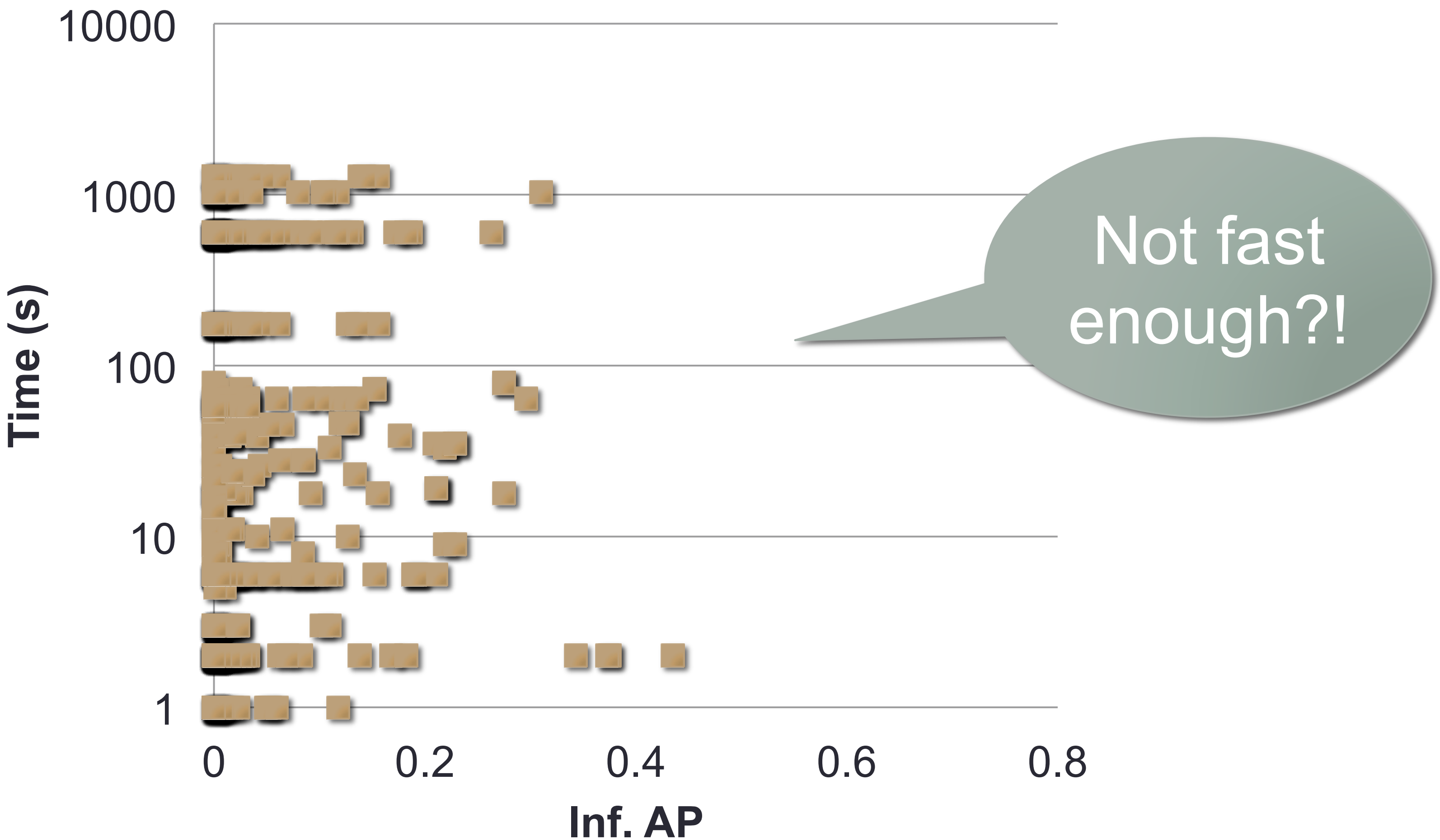
Run	Inf. AP score
D_NII_Hitachi_UIT.16_4	0.054
D_ITI_CERTH.16_4	0.051
D_ITI_CERTH.16_3	0.051
D_ITI_CERTH.16_1	0.051
D_NII_Hitachi_UIT.16_3	0.046
D_NII_Hitachi_UIT.16_2	0.043
D_NII_Hitachi_UIT.16_1	0.043
D_ITI_CERTH.16_2	0.042
E_INF.16_1	0.040
D_VIREO.16_6	0.038

No statistical significant differences among the top 10 runs

Processing time vs Inf. AP ("M" runs)



Processing time vs Inf. AP ("F" runs)



2016 Observations / Questions

- Most teams relied on intensive visual concept indexing, leveraging on past Semantic Indexing (SIN) task and similar like ImageNet, Scenes …
- Combined with manual or automatic query transformation
- Clever combination of concept scores (e.g., Waseda)

- Ad-hoc search is more difficult than simple concept-based tagging.
- Big gap between SIN best performance and AVS: maybe performance should be better compared with the "concept pair" task within SIN

- Manually-assisted runs performed better than fully-automatic.
- Most systems are not real-time (slower systems were not necessarily effective).
- Some systems reported 0 time!!!
- E and F runs are still rare compared to A and D

- Was the task/queries realistic enough?!
- Do we need to change/add/remove anything from the task in 2017 ?

Continued at MMM2017

- 10 Ad-Hoc Video Search (AVS) tasks, 5 of which are a random subset of the 30 AVS tasks of TRECVID 2016 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

NIST
National Institute of Standards and Technology

9:20 - 12:00 : Ad-hoc Video Search

- **9:20 - 9:40**, Task Overview
- **9:40 - 10:00**, NII_Hitachi_UIT (National Institute of Informatics; Hitachi; U. of Inf. Tech.)
- **10:00 - 10:20**, ITI_CERTH (Centre for Research and Technology Hellas)

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

- **10:40 - 11:00**, Waseda (Waseda University)
- **11:00 - 11:20**, kobe_nict_siegen (Kobe U.; Japan National Institute of Inf. and Communications Tech.;U. of Siegen)
- **11:20 - 11:40**, INF (Carnegie Mellon University, University of Technology Sydney, Renmin University of China, Shandong University)
- **11:40 - 12:00**, AVS discussion