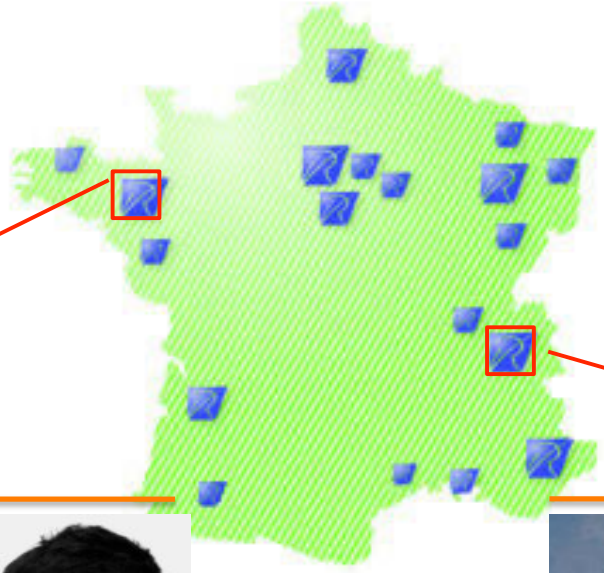
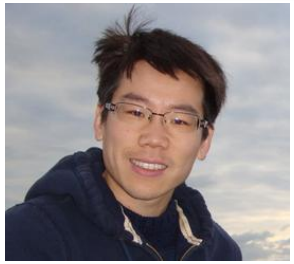
INRIA@TRECVID-CCD



TEXMEX

LEAR

Jonathan Delhumeau

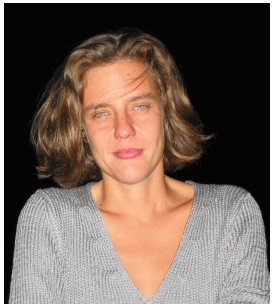


Jiangbo Yuan

Hervé Jégou

Jerome Revaud

Matthijs Douze



Cordelia Schmid

Inria

INVENTEURS DU MONDE NUMÉRIQUE

Conclusions and questions from last year

- What are the individual contributions of audio and video ?

- Audio weaker than video, apparently
  - But complementary to image
  - Further improvement possible ?

- Fusion step is critical
  - Is early fusion an option ?

- Scoring strategies to optimize NDCR looks critical
  - Keep maximum 1 result per query ?

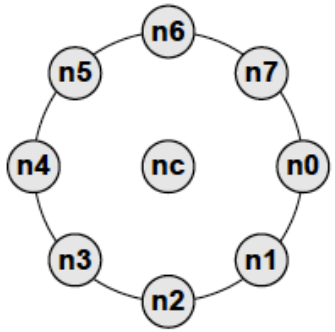
Our Runs at Trecvid

Run	Profile	Visual	Audio	Fusion	Cut
DEAF	balanced	yes	no	N/A	yes
AUDIOONLY	balanced	no	yes	N/A	yes
THEMIS	balanced	yes	yes	late	yes
ZOZO	balanced	yes	yes	late	no
DODO_{bal}	balanced	yes	yes	early	yes
TYCHE	nofa	yes	yes	late	yes
$\text{DODO}_{\text{nofa}}$	nofa	yes	yes	early	yes

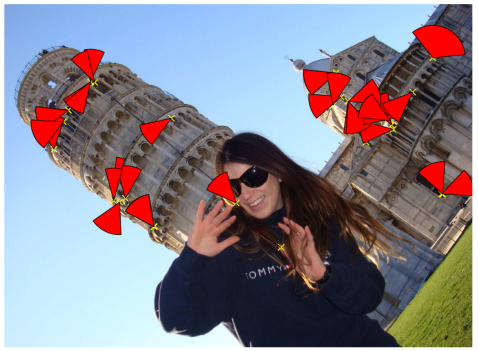
- 5 runs to measure the individual contributions of our system
- 2 runs designed for “best” search quality: the **DODO** runs

Video visual system: ingredients (same as in 2010)

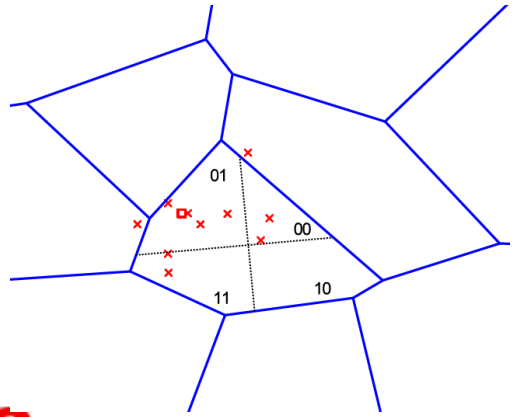
- Local descriptors: CS-LBP



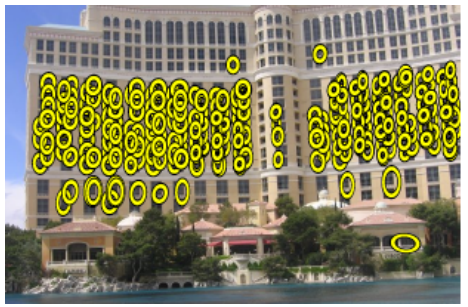
- Weak geometric consistency



- Hamming Embedding
  - Improve bag-of-features

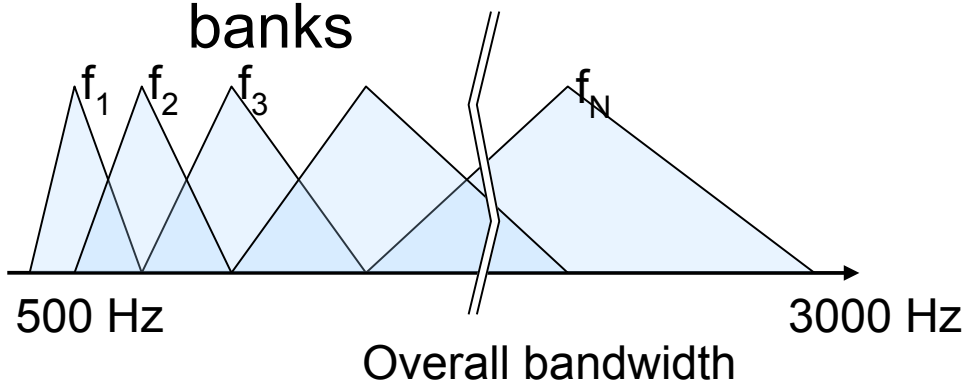


- Burstiness strategy + Multi-probe

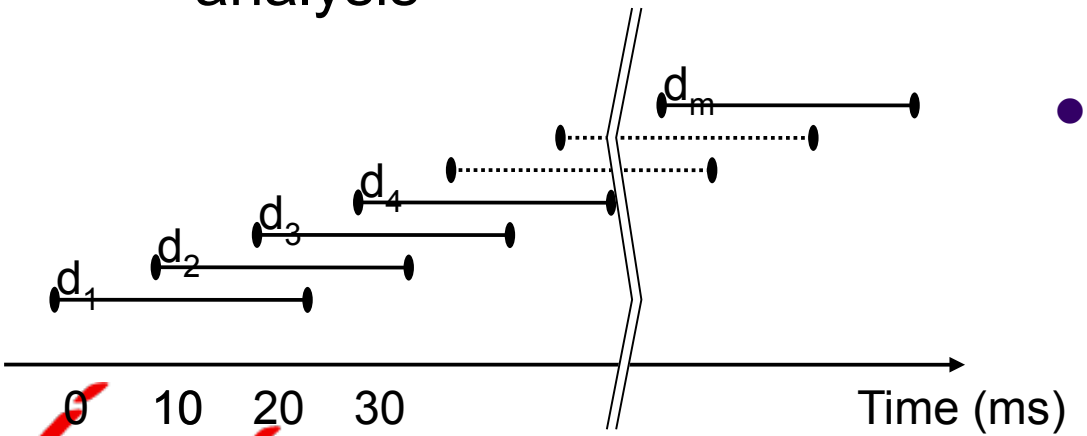


Audio system: basic ingredients (same as in 2010)

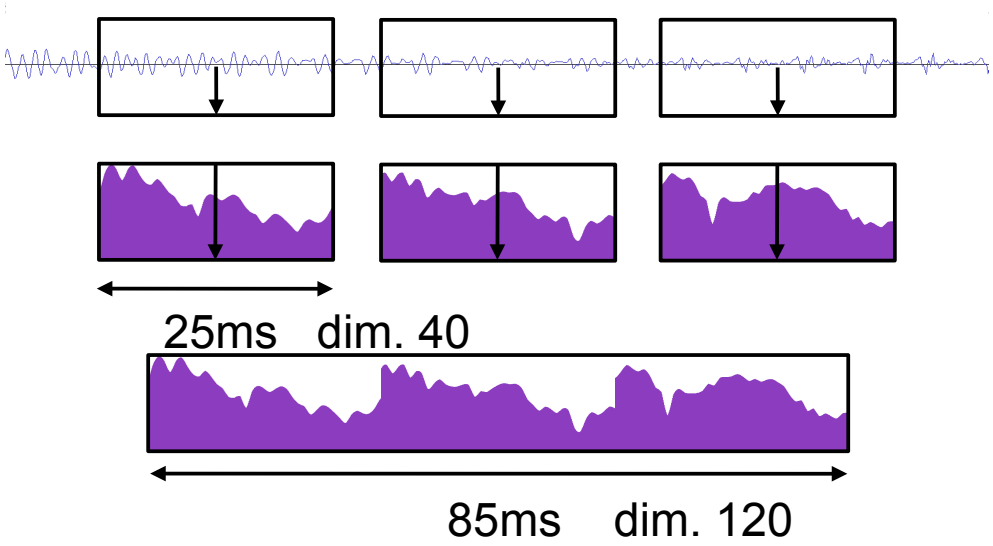
- Base descriptor: Filter banks



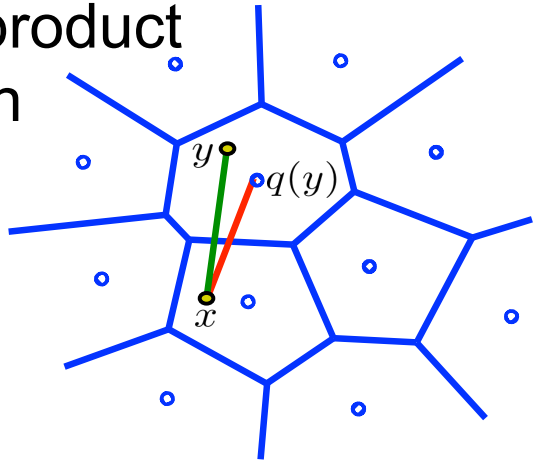
- Overlapping temporal analysis



- Compounding



- Matching: product quantization

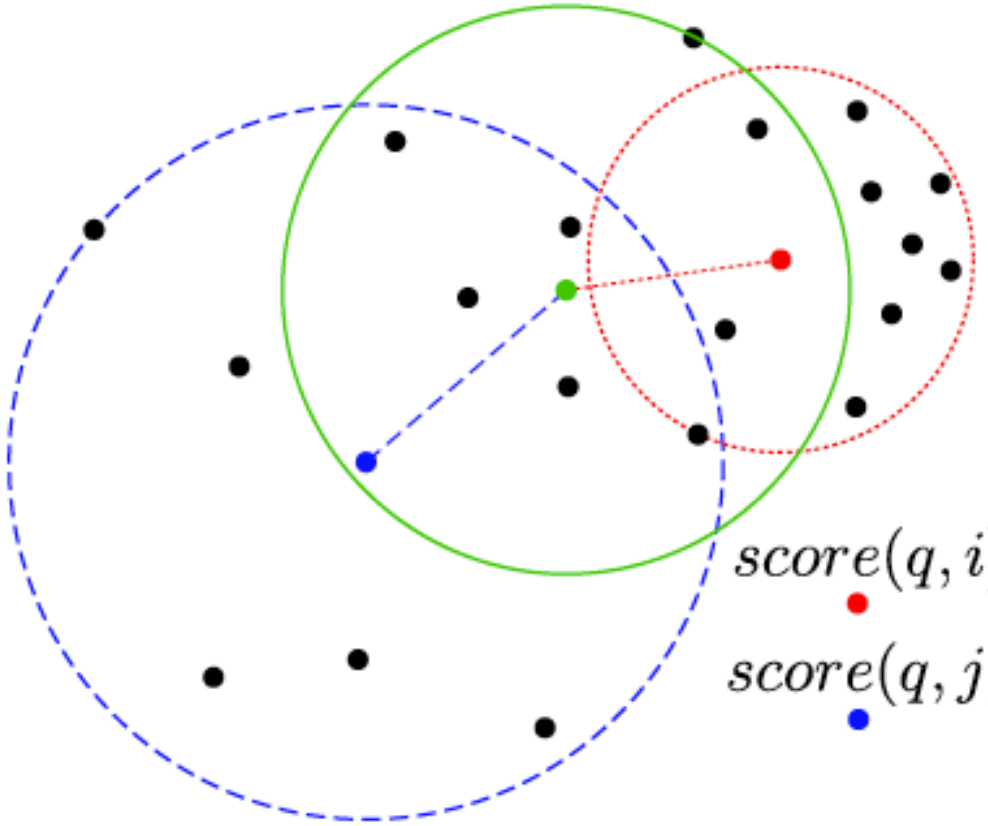


New ingredient 1: temporal shift

INVENTEURS DU MONDE NUMÉRIQUE

New ingredient 2: reciprocal nearest neighbors

- Audio matches: k-nearest neighbors
- Pb: if X neighbor of Y, Y not necessarily neighbor of X
- Weighted **Reciprocal nearest neighbors**

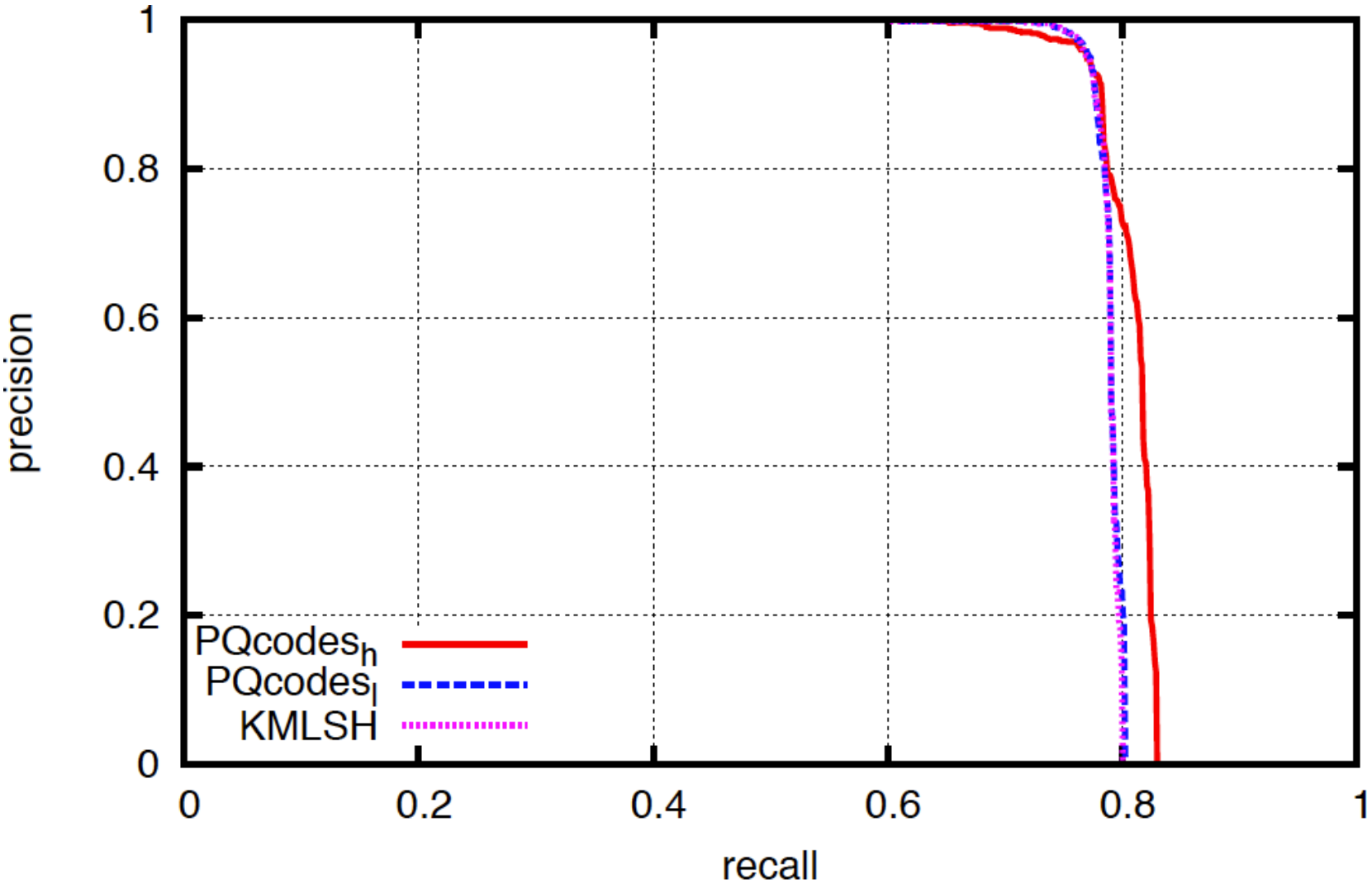


$$score(q,i) = d_k(q) + d_k(i) - 2 \cdot d(q,i)$$

$$score(q,j) = d_k(q) + d_k(j) - 2 \cdot d(q,j)$$

Audio: improvement over last year

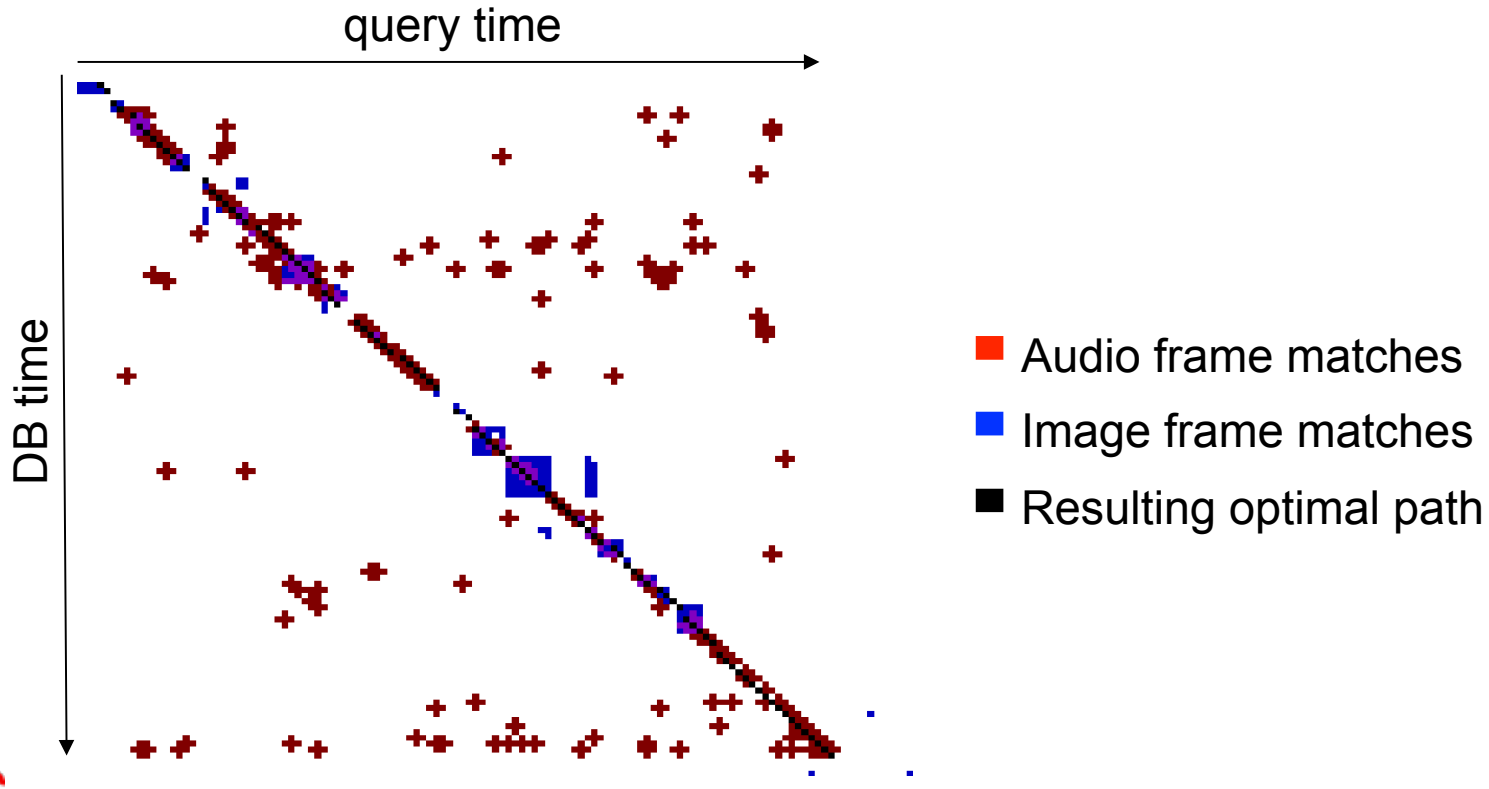
- 6 times slower in total, for a limited improvement



New ingredient 3: early fusion audio/video

- Early fusion:
  - Input: image & audio raw Hough hypotheses
  - Robust time warping to align query frames with DB frames

Example of time warping matrix:



INVENTEURS DU MONDE NUMÉRIQUE

Early fusion

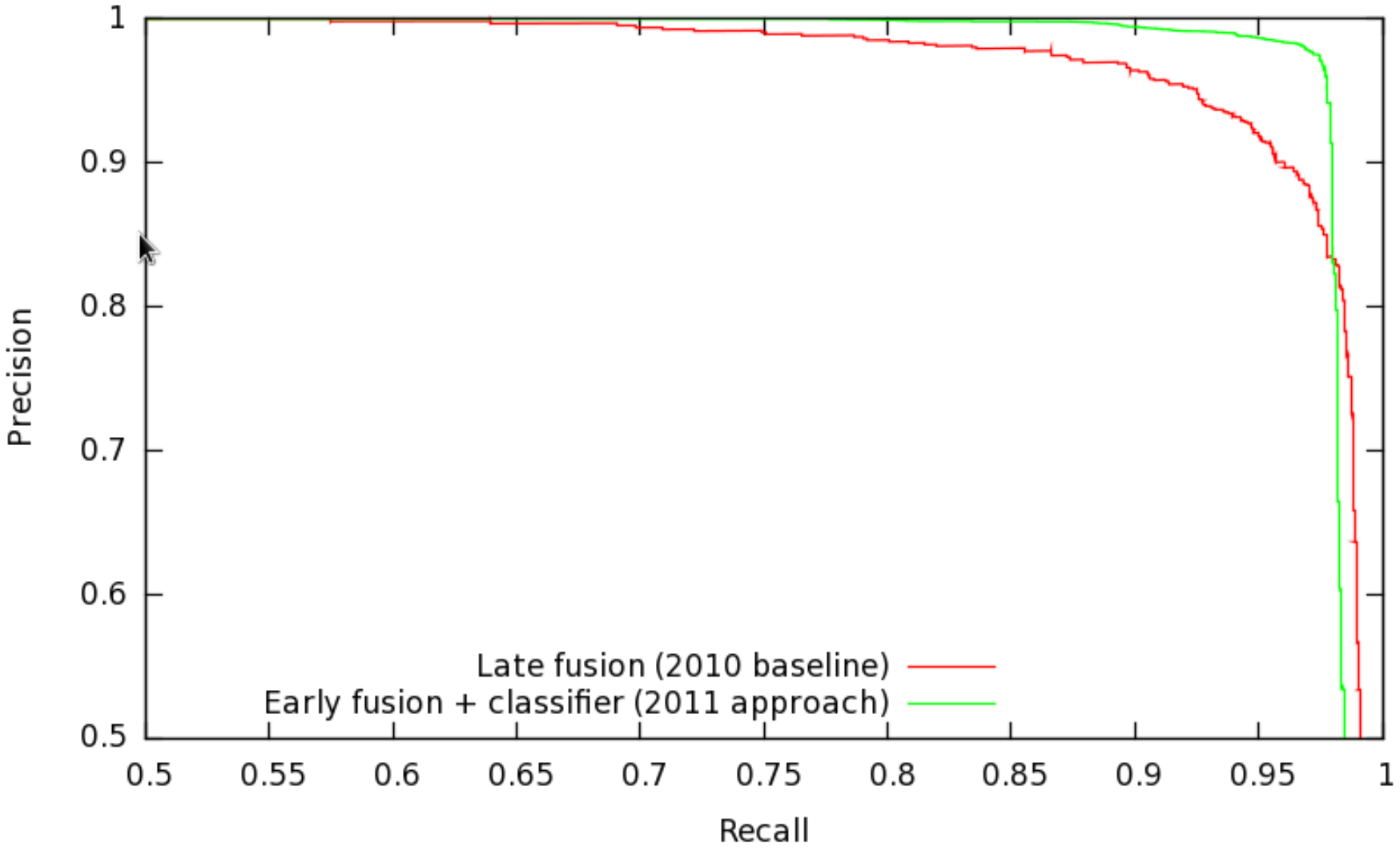
- Input: image & audio raw Hough hypotheses
- 1. Robust time warping - align query frames with db frames
- 2. Description of matching segments
  - segment length, number of audio/image frame matches, …
  - surface of the image recognized on the database side
  - KL-divergence between db keypoints distribution / matches distribution
  - relative support of image & audio for the hypothesis
  - etc.
- 3. Classifier produces a score

Early fusion: training the classifier

- Boosting scheme:
  - Each iteration, addition of a new feature
    - Criterion: maximize AP on validation set
  - Classifier: Logistic regression (better than SVM here)
    - 40,000 positive samples
    - 150,000 negative examples

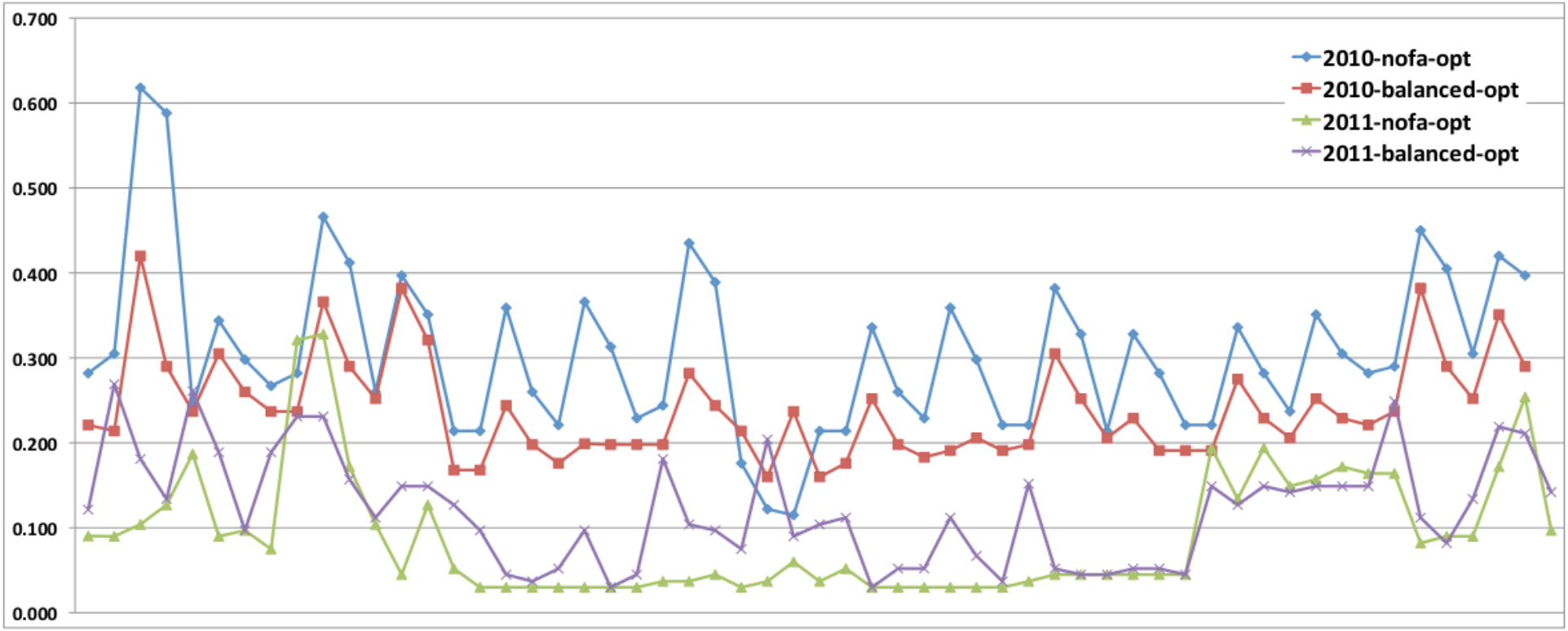
- Result: selected features (sorted)
  - Detected area
  - Nb of audio & image frame matches
  - KL divergence between keypoints distribution
  - Length of matching segment in seconds
  - Etc…

2010 vs 2011 approach



INVENTEURS DU MONDE NUMÉRIQUE

2010 vs 2011 approach



INVENTEURS DU MONDE NUMÉRIQUE

Analysis: balanced profile (opt-NDCR)

RUN	AUDIO	VIDEO	FUSION	NDCR (avg)
DEAF	no	yes	n/a	0.258
AUDIOONLY	yes	no	n/a	0.406
THEMIS	yes	yes	late	0.211
ZOZO	yes	yes	late	0.194
DODO	yes	yes	early	0.144

- One surprise: ZOZO > THEMIS
  - Keeping more than 1 result is better if scores are ties

Overview of results (opt-NDCR)

Balanced profile

RANK	INRIA	PKU	CRIM	NTT-CSL
1	5	31	21	1
2	16	23	8	3
3	9	2	9	5
4	19	0	10	7
5	4	0	4	4
6	1	0	4	13
7	2	0	0	12

NoFa profile

RANK	INRIA	PKU	CRIM	NTT-CSL
1	23	14	18	8
2	10	31	11	7
3	11	10	13	4
4	9	1	9	5
5	3	0	4	7
6	0	0	1	5
7	0	0	0	3

- **PKU and CRIM are much better with Actual-NDCR**
  - We don’t know how to set the threshold
  - This problem may be inherent to our system

Introduction of Babaz audio matching system

- Open source: http://babaz.gforge.inria.fr/

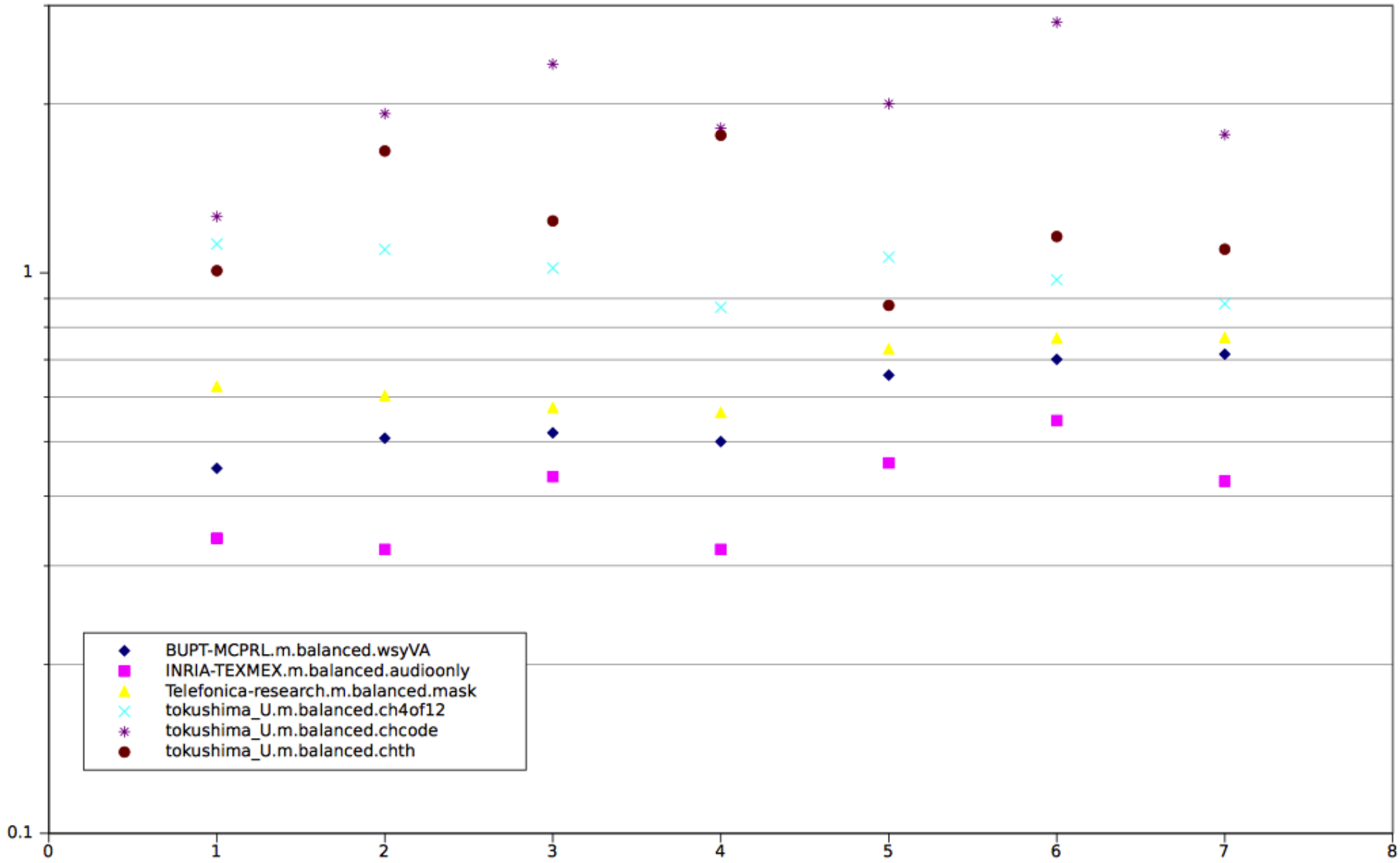
- Well… PQ-codes replaced by k-means LSH (licensing issue)
  - Requires more memory (40GB instead of 5GB) and slower
  - But PQ-codes Matlab implementation available

- All Trecvid queries: query times (16 cores), memory, mAP
  - Pqcodes – heavy: 20H 5GB mAP = 80.7 %
  - Pqcodes – light: 3H 5GB mAP = 78.9 %
  - K-means LSH: 25H 40GB mAP = 78.8 %

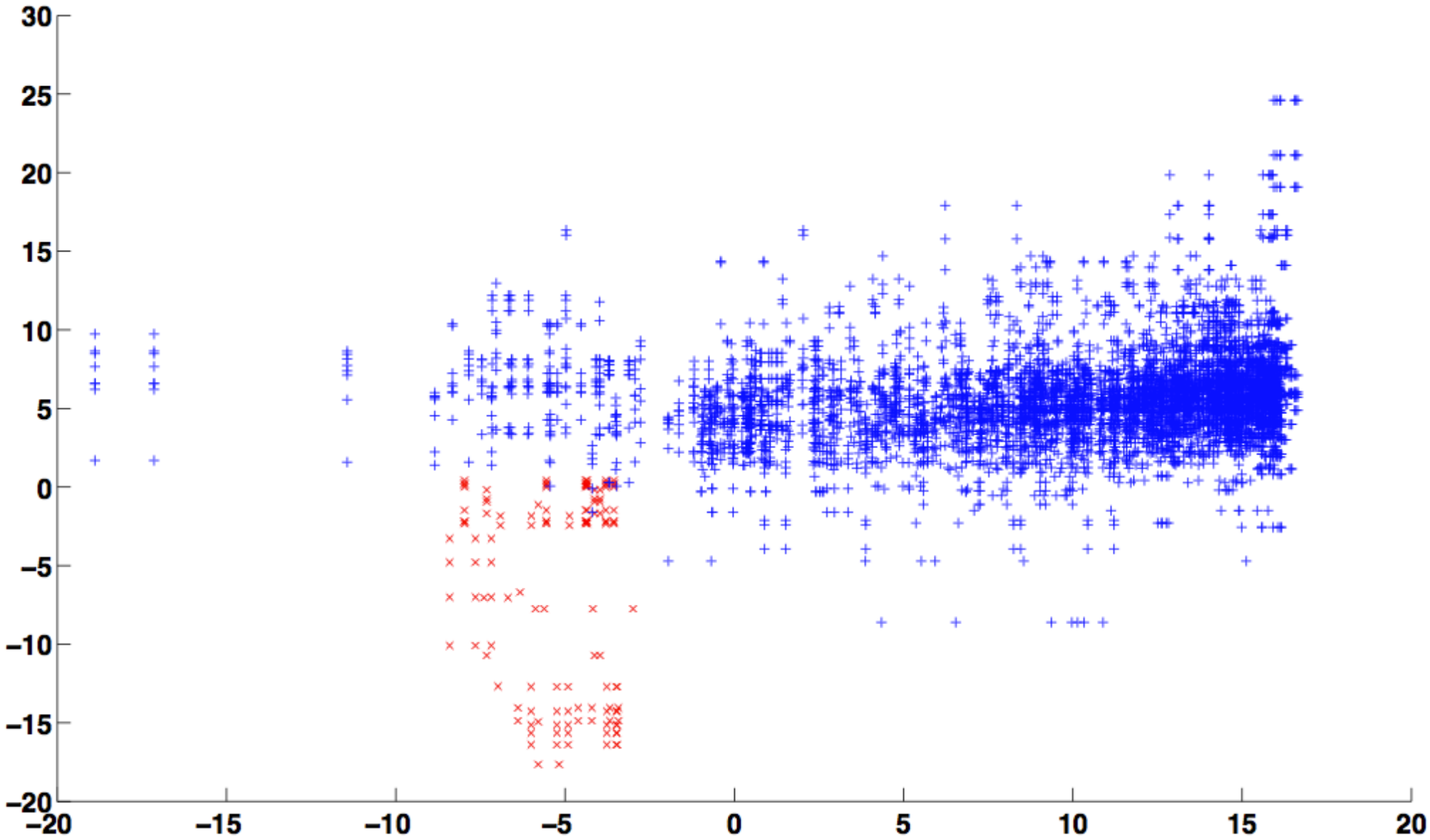
- Offline: Pqcodes-h: 69H, Pqcodes-l: 11H, KMLSH: 17H

Inria
INVENTEURS DU MONDE NUMÉRIQUE

Questions?



INVENTEURS DU MONDE NUMÉRIQUE



INVENTEURS DU MONDE NUMÉRIQUE