論文 / 著書情報
Article / Book Information

論題(和文)	コミュニケーションとしての映像とその検索
Title(English)	
著者(和文)	篠田浩一
Authors(English)	Koichi Shinoda
出典(和文)	第15回情報理論的学習理論ワークショップ(IBIS2012), , ,
Citation(English)	, , ,
発行日 / Pub. date	2012, 11
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2012 Institute of Electronics, Information and Communication Engineers.

第15回情報理論的学習理論ワークショップ(IBIS2012)

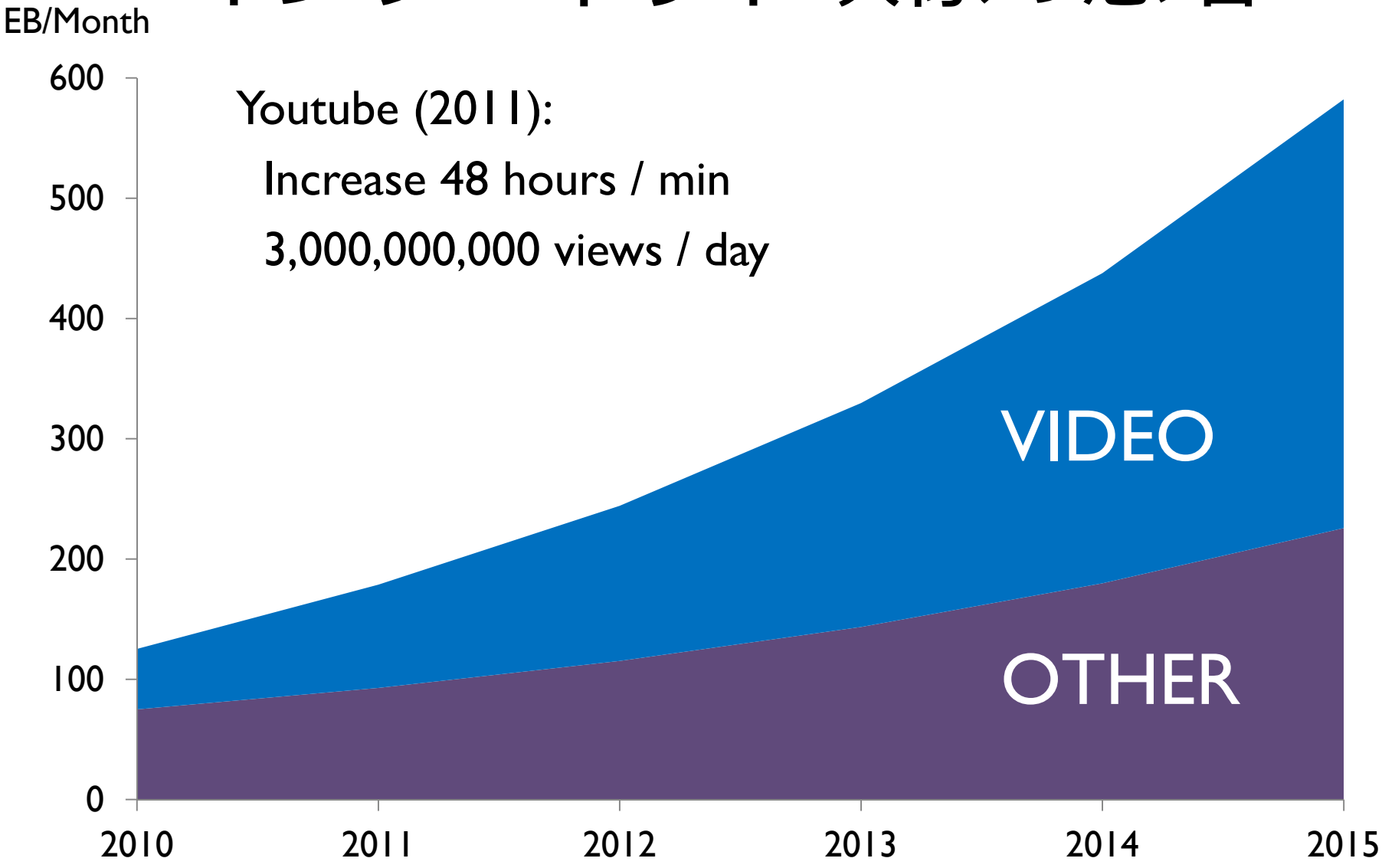
コミュニケーションとしての映像とその検索

篠田浩一
(東京工業大学)

講演の内容

1. 音声と映像
2. TRECVID Semantic Indexing (SIN)
3. SIN のための音声技術
4. TRECVID Multimedia Event Detection (MED)
5. まとめ

インターネット映像の急増



IP Traffic (Cisco Visual Networking Index 2010-2015)

No Meta data

Large Variety

Low quality

Mostly Useless

課題

インターネット映像からの Content-Based Video Retrieval (CBVR)

これまでの研究対象

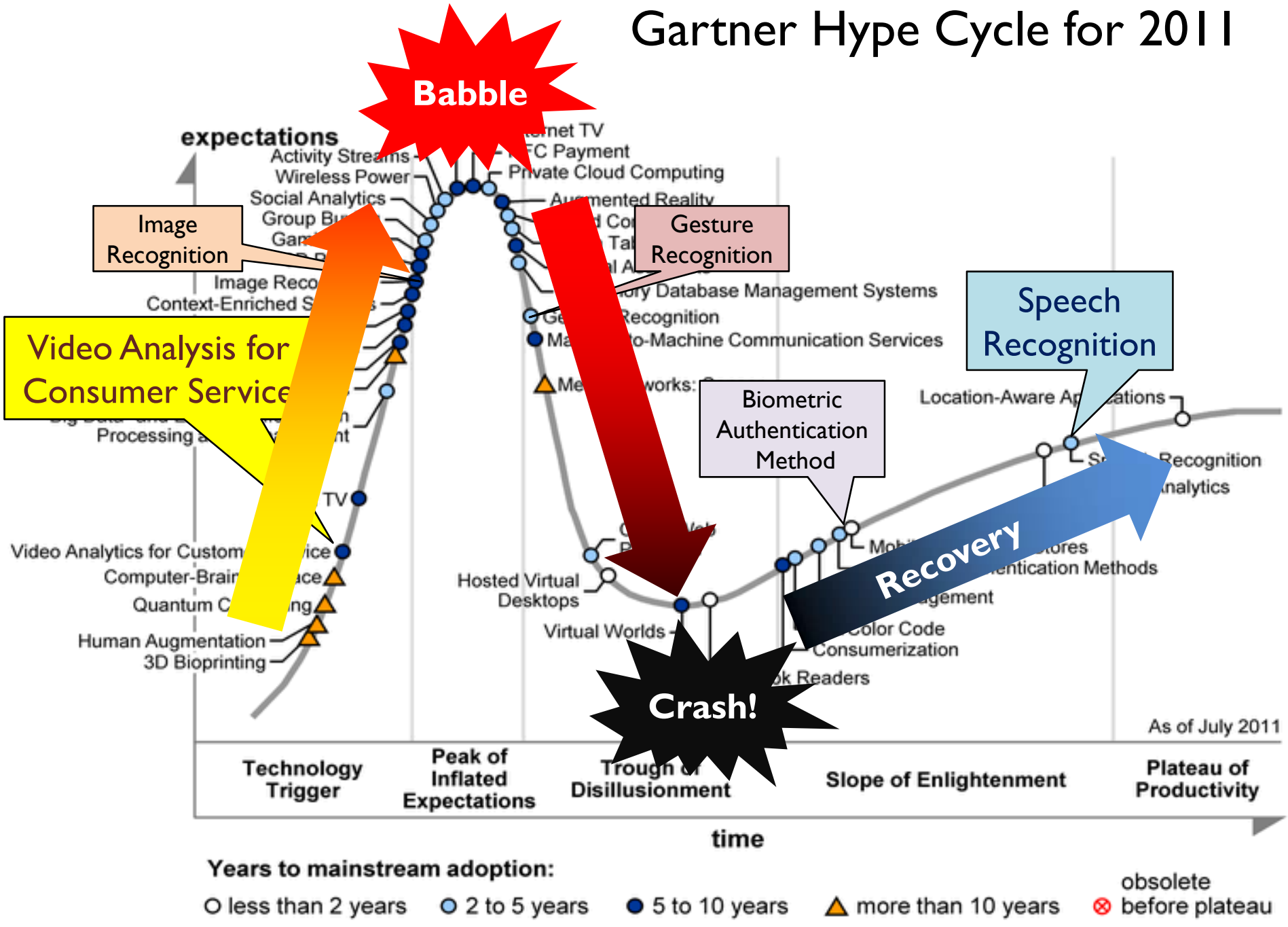
TV ドラマ，映画，ニュース，スポーツなど

- ジャンルが特定
- 高品質
- プロによる編集
- メタデータが豊富

インターネット映像とは明らかに異なる

どのような方法論をとるべきか？

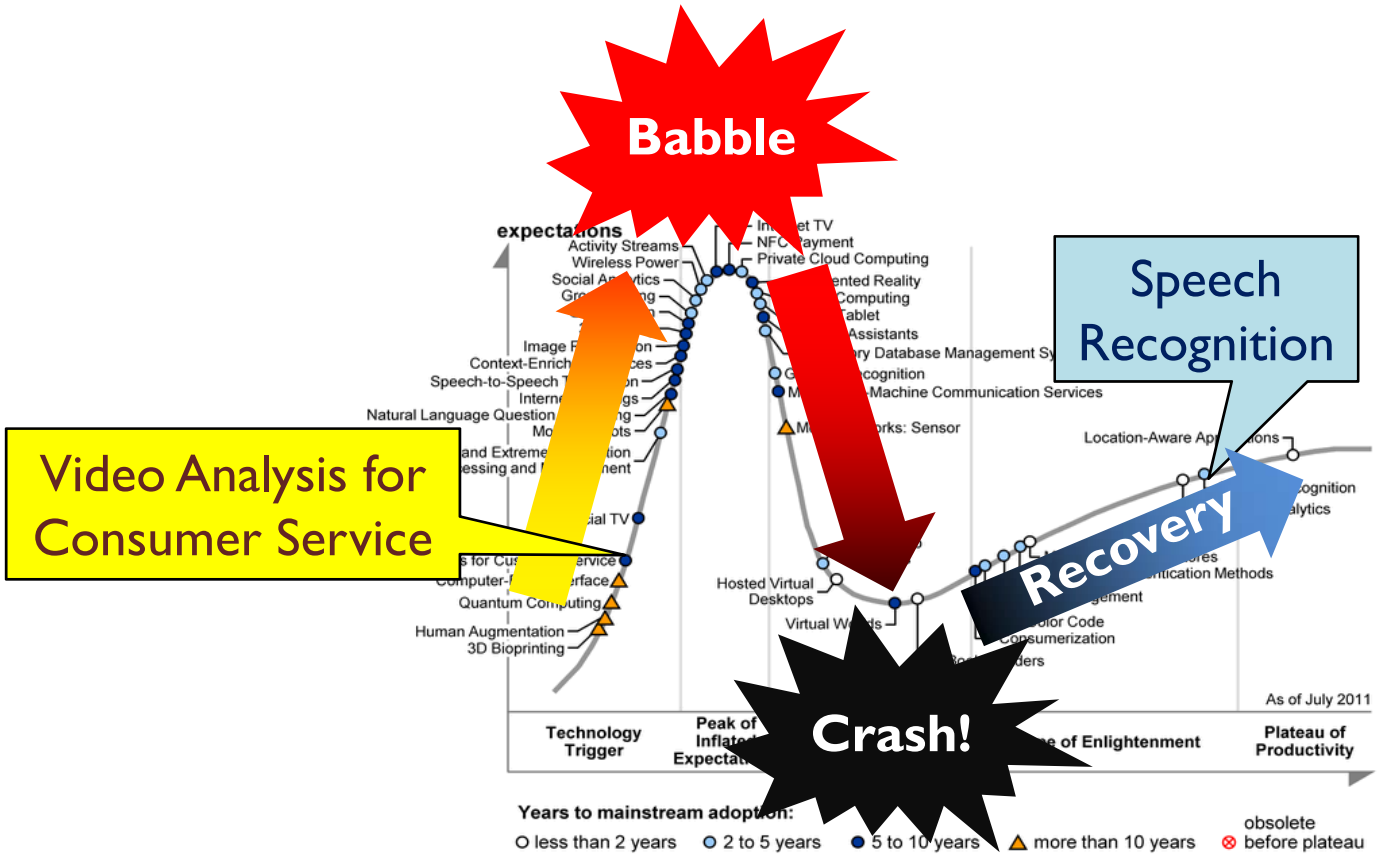
Gartner Hype Cycle for 2011

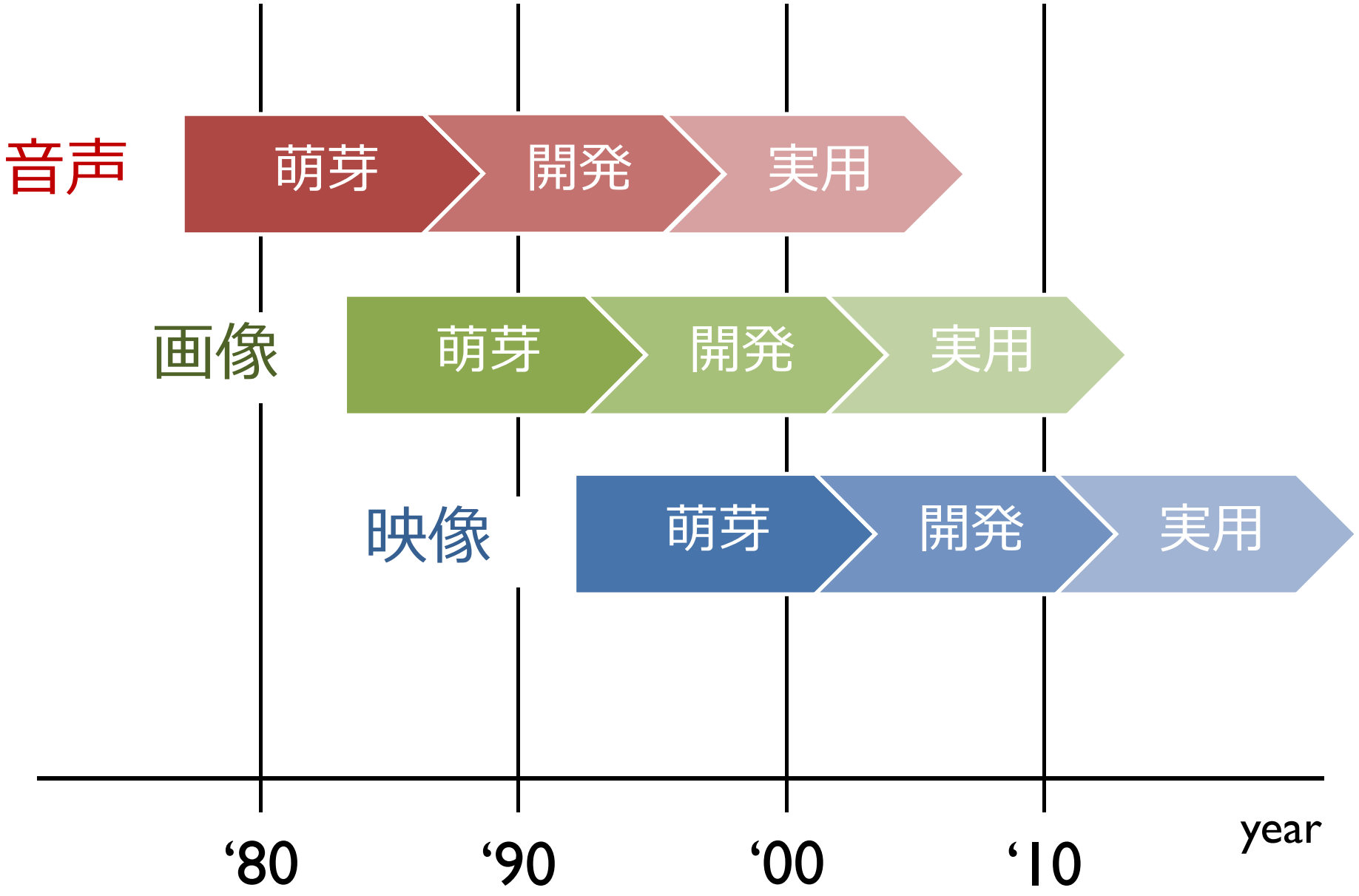


音声研究から学べないか？

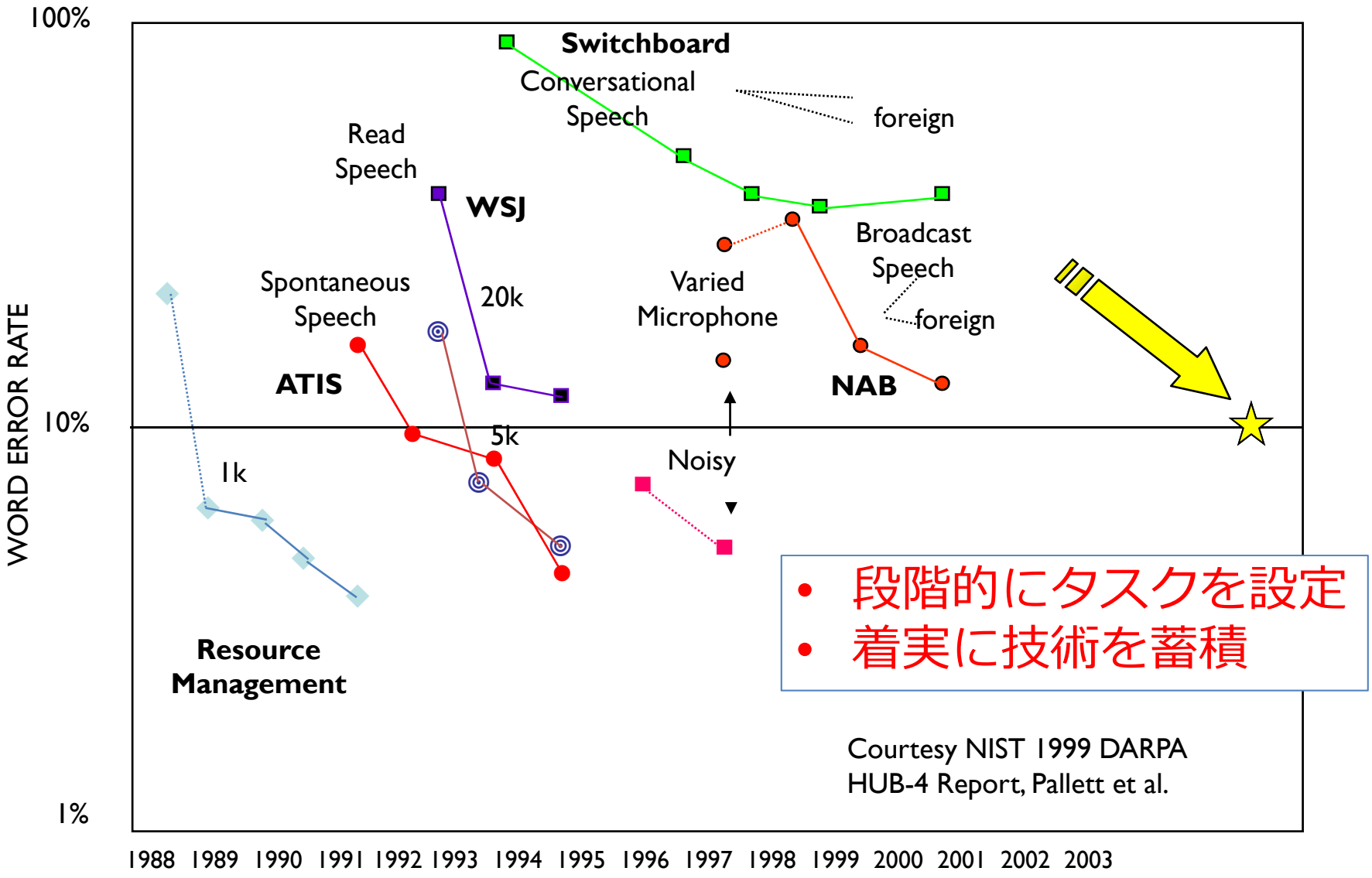
映像研究はこれから “Babble”

音声研究は “Babble Crash”から生き延びた




音声
萌芽
開発
実用
画像
萌芽
開発
実用
映像
萌芽
開発
実用
‘80
‘90
‘00
‘10
year

米国国防省(DARPA)音声認識ベンチマーク



- 段階的にタスクを設定
- 着実に技術を蓄積

Courtesy NIST 1999 DARPA
HUB-4 Report, Pallett et al.

音声と映像は違う?

- 音声は 1 次元、映像は 3 次元
- 音声には “Semantic Gap” がない(?)
- 音声はコミュニケーションの道具であるが映像は違う (?)
- 映像は、音声のような明確な構造がない(?)
  (音素→形態素→単語→文、文法)

音声と映像は同じ

送り手

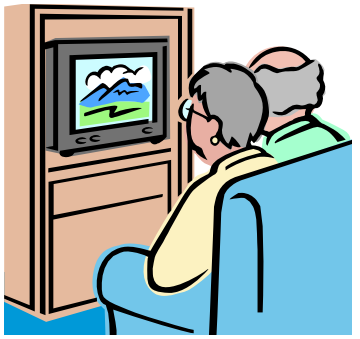
Audio Channel

メッセージ

メッセージ

Video Channel

受け手



映像はコミュニケーションの手段

(非明示的な)語彙や文法をもつ

※音声にもSemantic Gapは存在する

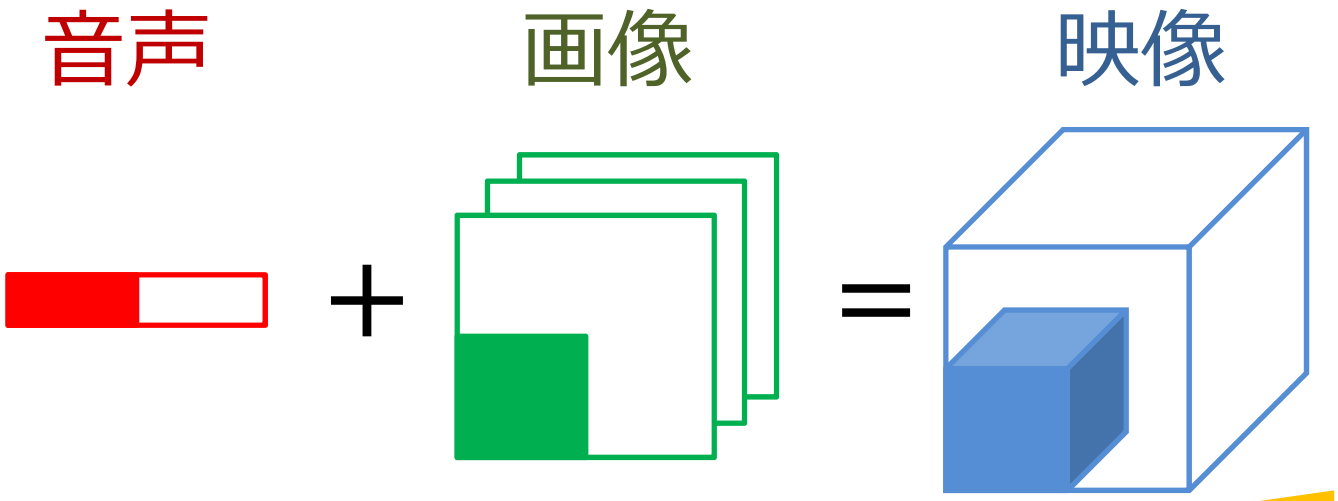
映像検索のための音声技術

1. 「送り手」をモデル化するための生成モデル

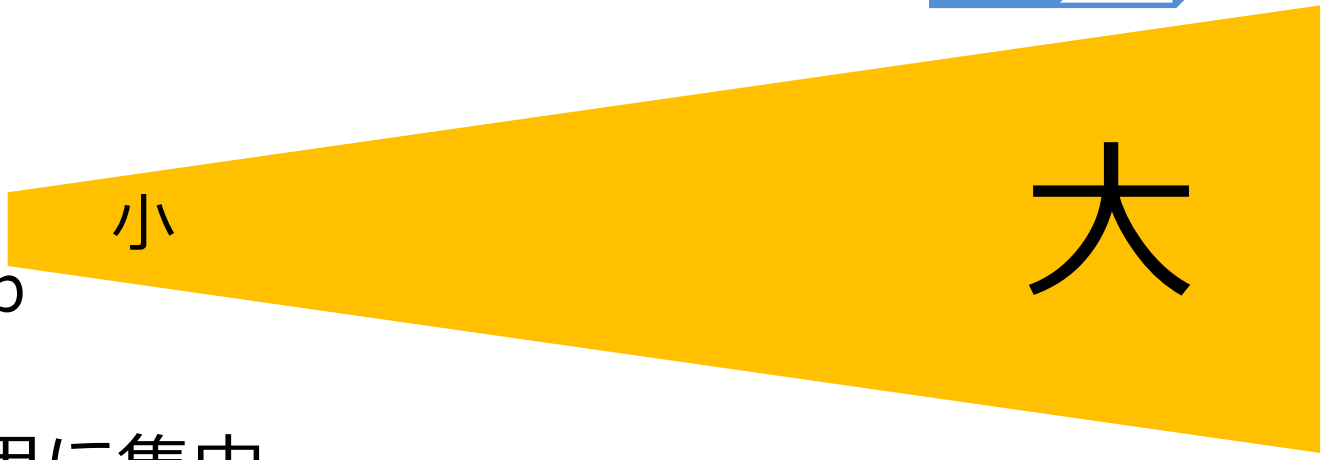
2. 低品質・多様性・データ不足に頑健な
   確率的フレームワーク

3. 高速計算手法

機械学習によるアプローチ（共通）



データ量
計算量
Semantic gap



- 特定の応用に集中
- 他機関との協働

TRECVID
Semantic Indexing

TRECVID
(TREC Video Retrieval Evaluation)

2001年に Text REtrieval Conference (TREC) から独立

NIST(National Institute of Standard and Technology) が主催

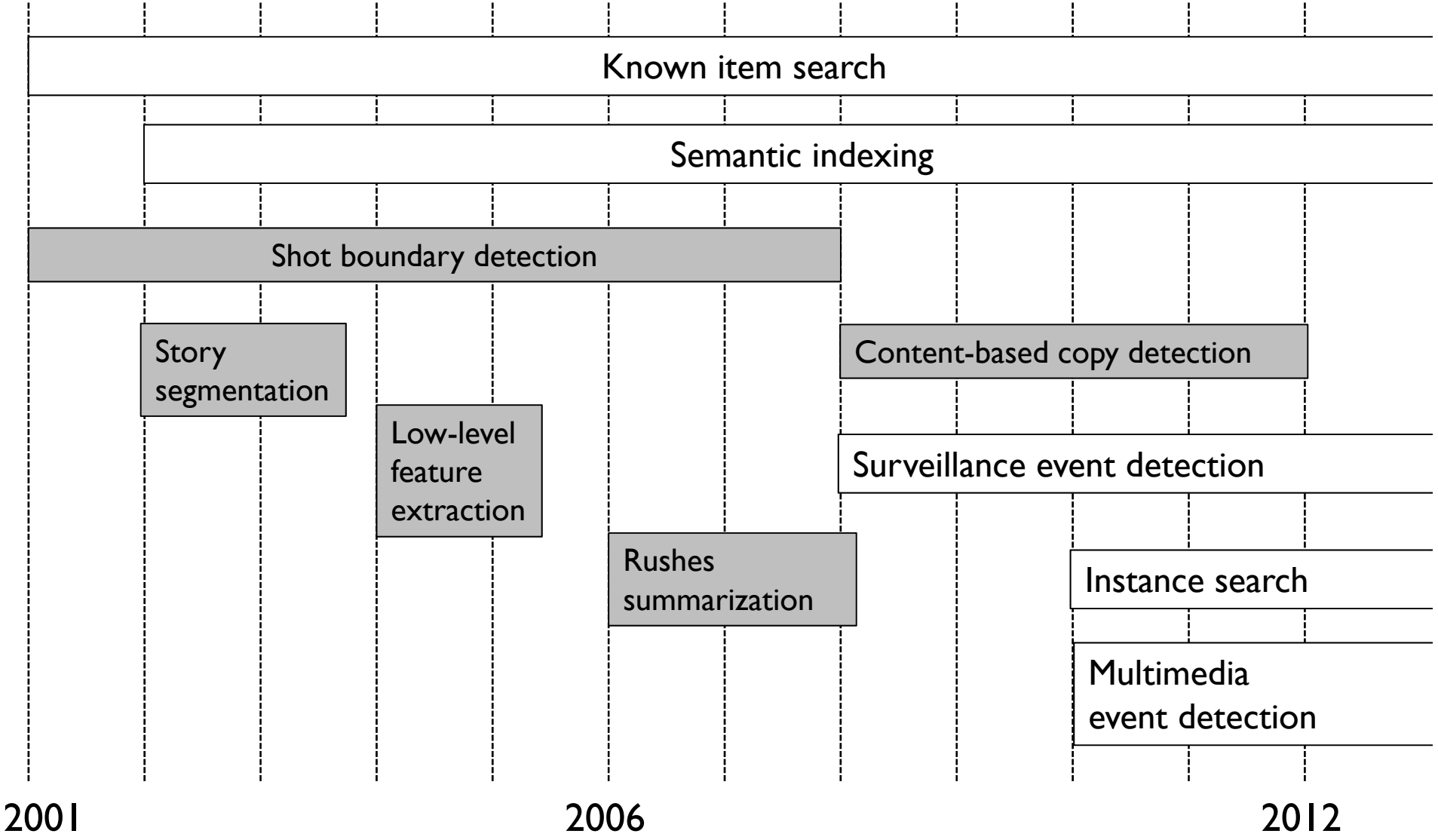
目的：映像コンテンツ分析・検索研究の促進

クローズドな国際競争型ワークショップ

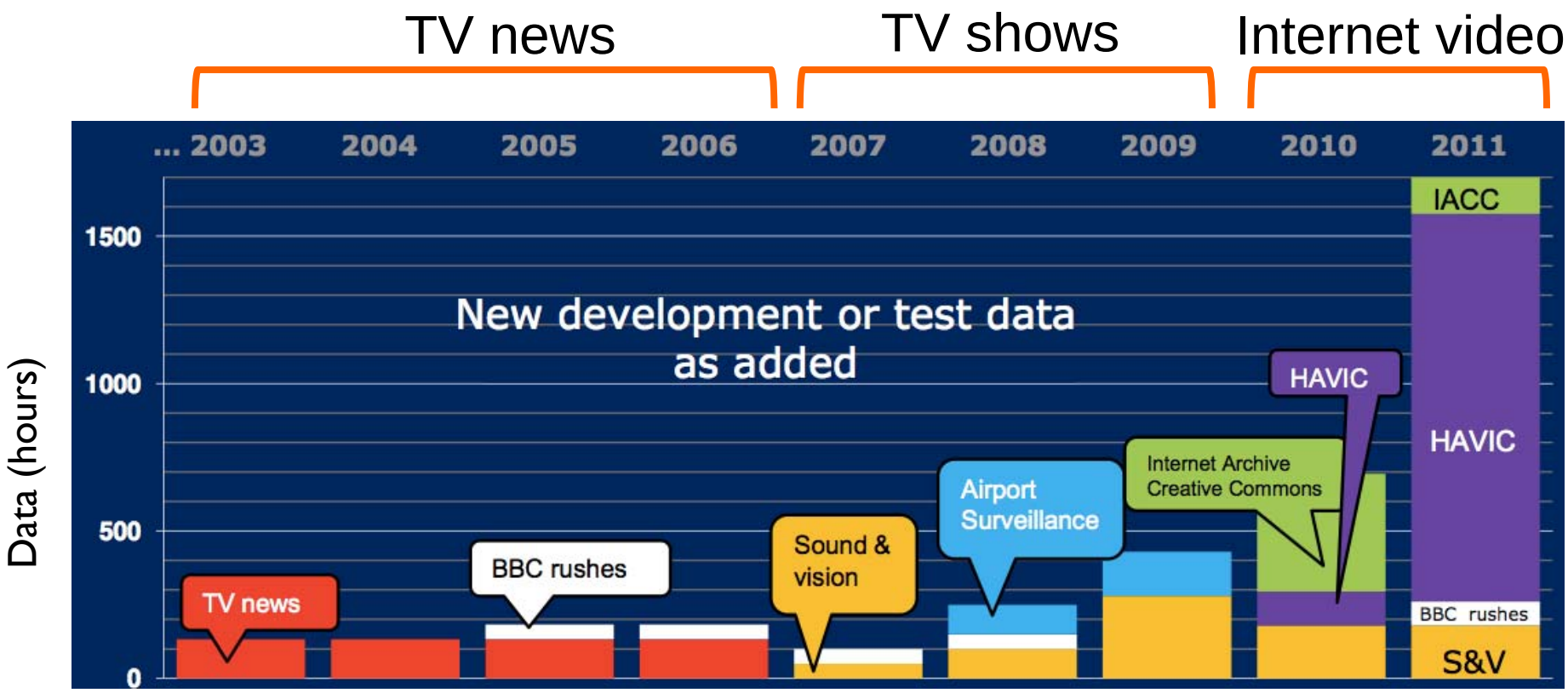
ホームページ: http://trecvid.nist.gov

- 大規模データが使える (著作権等の問題をクリア)
- 手法の比較が容易、そのため進歩が速い
- ラベル付け作業を分担
- 勝ち負けがはっきりする

TRECVID タスクの歴史



TRECVID データセット



http://www-nlpir.nist.gov/projects/tvpubs/tv11.slides/tv11.intro.slides.pdf

2011年 TRECVID

66 チームが参加（日本からは12チーム)

5 tasks:

- Semantic indexing (SIN)
- Multimedia event detection (MED)
- Known item search (KIS)
- Instance search (INS)
- Surveillance event detection (SED)

Semantic Indexing (SIN)

目的

ビデオショットからの Concept を検出

Concepts: objects, scenes, ...

TRECVIDの中核的タスク

静止画の一般物体認識に対応

Multimedia Event Detection (MED)

目的

ビデオクリップからのイベント検出

e.g. Batting a run in

Making a cake

SINより高次の対象

スポーツ番組からのハイライト検出

をインターネット映像まで延長

Instance Search (INS)

目的

　特定の人物、場所、ロゴを検出

対象は明確、学習データは少ない

データベース：BBC rushes

Known Item Search (KIS)

目的

詳細なテキスト記述に合致する映像シーンを検出

例：赤いシャツの男が犬にりんごをあげている

学習データなし

SINタスクで得られた

コンセプトを利用

Surveillance Event Detection (SED)

目的

監視カメラからのイベント検出

イベント: PeopleRuns, Pointing, PeopleMeet, など

混雑状況、固定カメラ

データベース

イギリス・ガトウィック空港における５台の監視カメラ映像(145時間)

Semantic Indexing (SIN)

タスク設定

データベース：IACC

(Internet Archive videos with Creative Commons licenses) ： **600 h**

コンセプト数: **346**

Run: 各々のコンセプトについて上位**2000**個のショットのリストを提出

※各々のチームは最大４つのRunを提出できる

評価基準: Average Precision (AP)

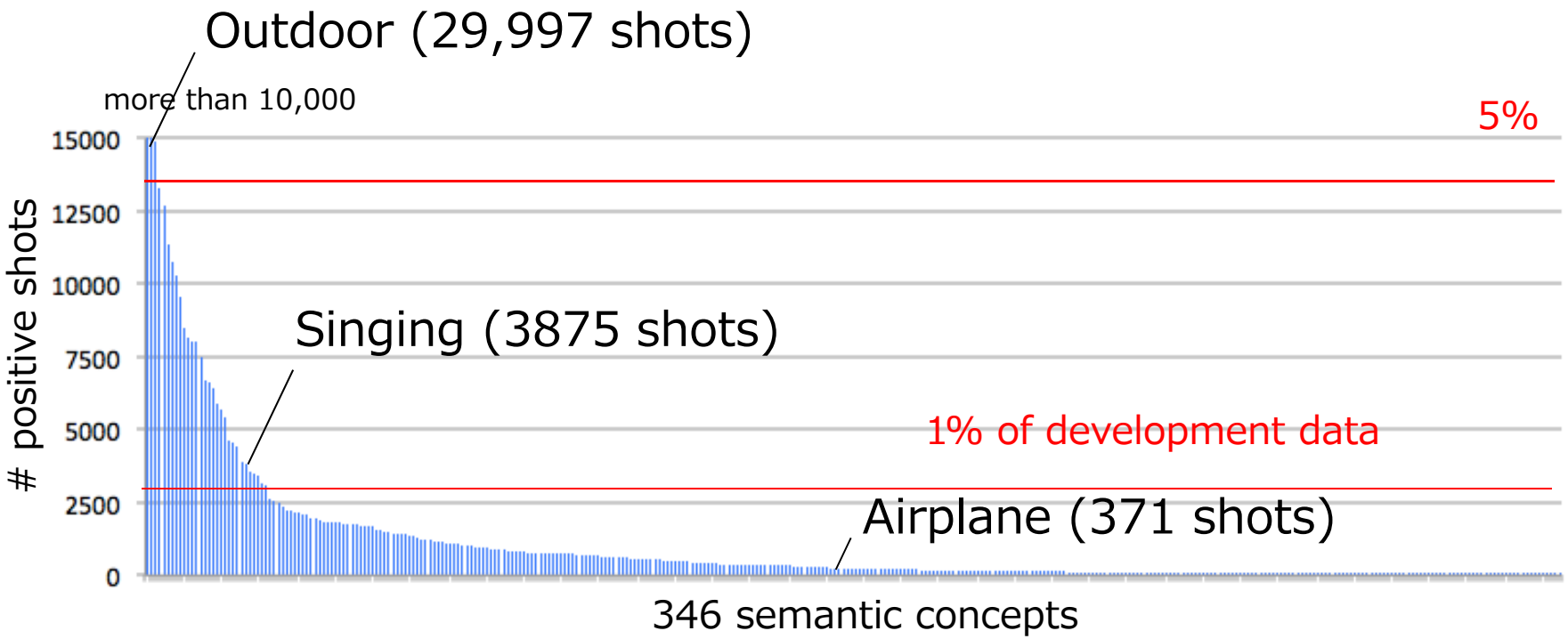
$$\frac{1}{K}\sum_{k=1}^{K}\frac{p(k)}{k}$$

k: Rank

$p(k)$: Number of true shots from 1st to k-th

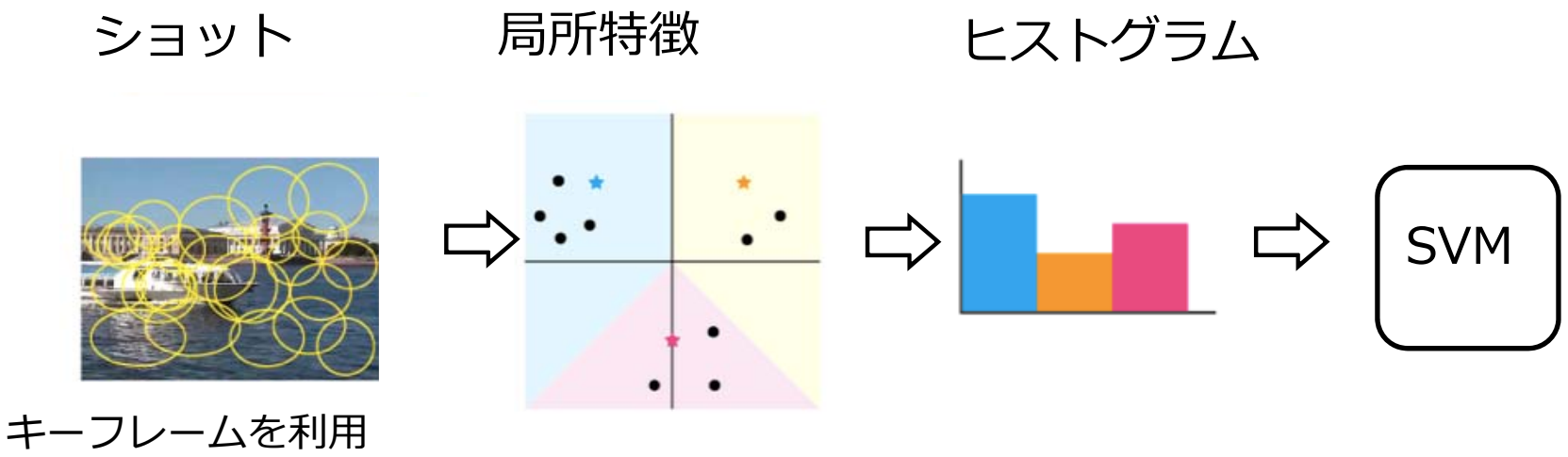
コンセプトの出現頻度

Number of positive samples in 264,673 training video shots



Bag of Words (BoW)

静止画における一般物体認識で主流



- 計算量が比較的少ない
- 量子化誤差が大きい

新たな動き (1)： 頑健性

低品質、多様性、データ不足に対応

- More features
  SIFT, Color SIFT, SURF, HOG, GIST, Dense features
- Multi-modal
  音声の利用： Singing, Dance, Car, etc.
- Multi-frame
  キーフレーム以外を利用
- Soft clustering
  量子化誤差の低減

新たな動き (2)： 高速化

参加58チーム中28チーム(半分)しか
結果を提出できなかった

- 近似アルゴリズム
- 分散処理
- Graphical Processing Unit (GPU)の利用

期待したが効果のなかったもの

- 大局特徴(色ヒストグラムなど)

  局所特徴だけで十分(相補的な関係にない)

- 音声認識，OCR

  それら自体の性能が低い

- 物体の位置検出

  位置検出性能が低い、位置のないConceptも

- Concept間のコンテキスト

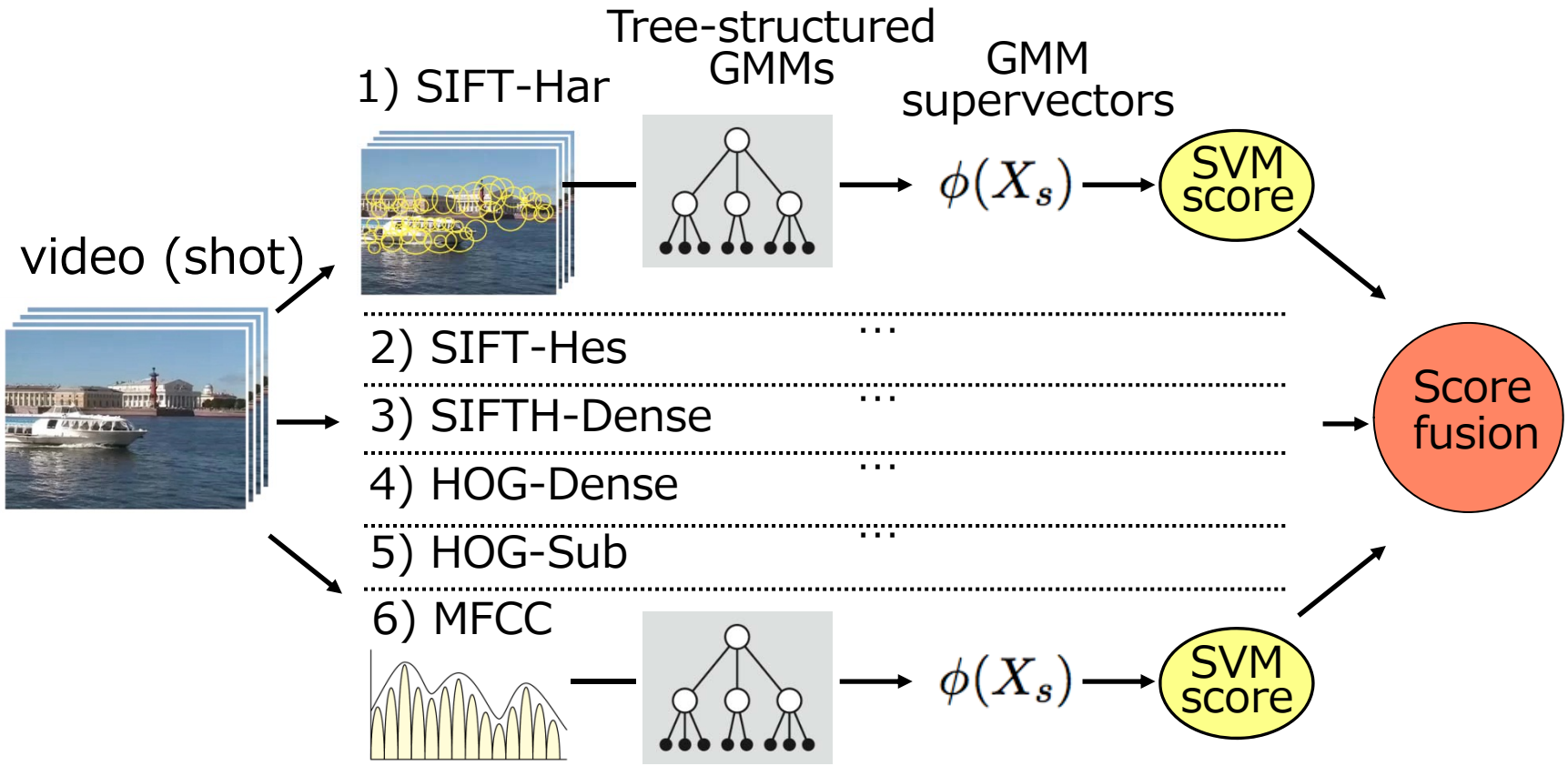
  データ量が少なすぎる

Semantic Indexingのための音声技術

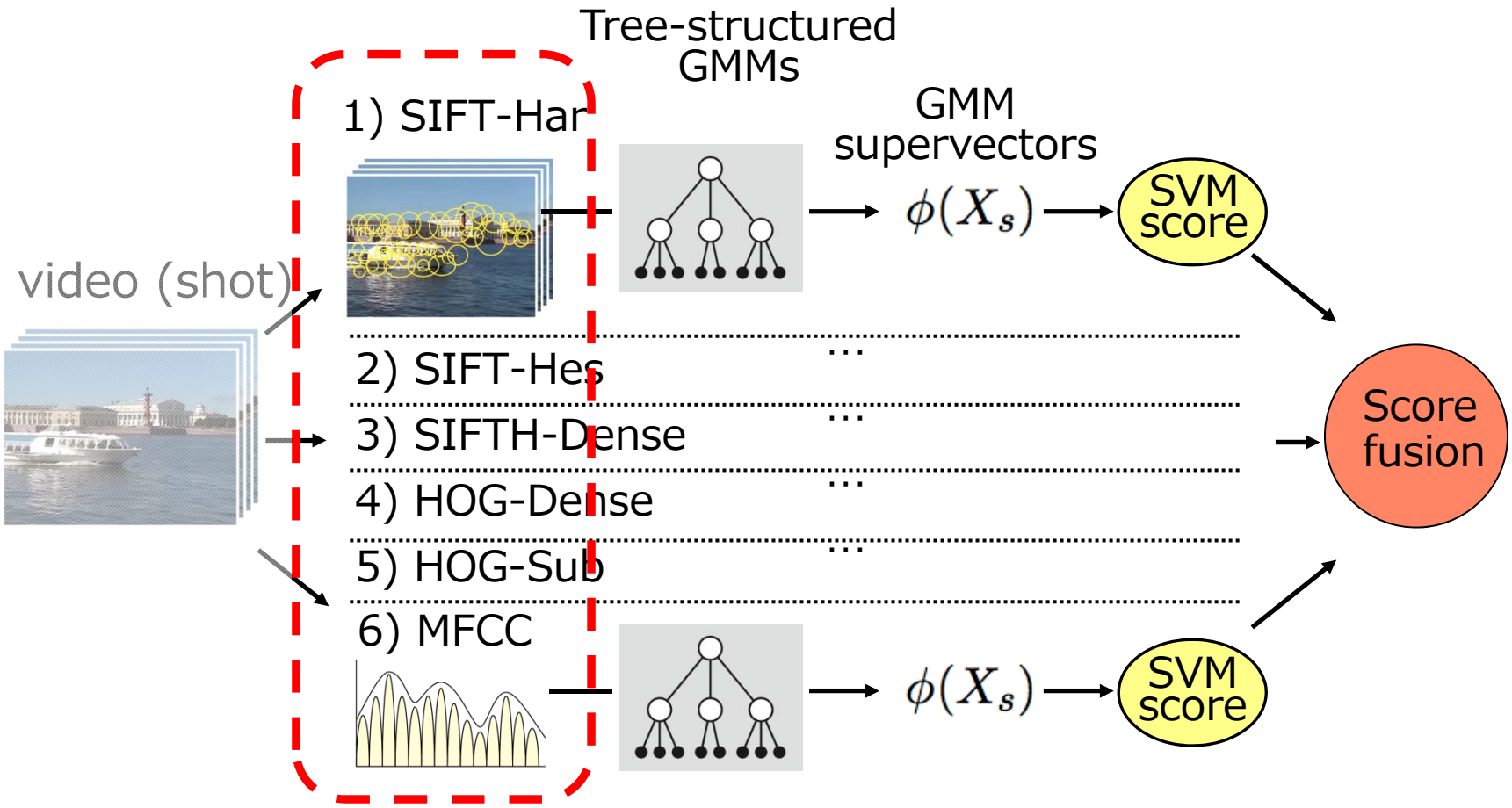
３つの音声技術

1. 多様性、低品質
   Gaussian Mixture Models (GMM)
2. データ不足
   MAP 適応
3. 高速化
   木構造サーチ

フレームワーク


Tree-structured GMMs
GMM supervectors
1) SIFT-Har
video (shot)
$\phi(X_s)$
SVM score
2) SIFT-Hes
3) SIFTH-Dense
4) HOG-Dense
5) HOG-Sub
6) MFCC
Score fusion

特徴抽出


Tree-structured GMMs
GMM supervectors
video (shot)
1) SIFT-Har
2) SIFT-Hes
3) SIFTH-Dense
4) HOG-Dense
5) HOG-Sub
6) MFCC
$\phi(X_s)$
$\phi(X_s)$
SVM score
SVM score
Score fusion

低次特徴

- 6つの画像特徴

  SIFT-Har，SIFT-Hes，SIFTH-Dense，HOG-Dense，HOG-Sub
- Multi-modal

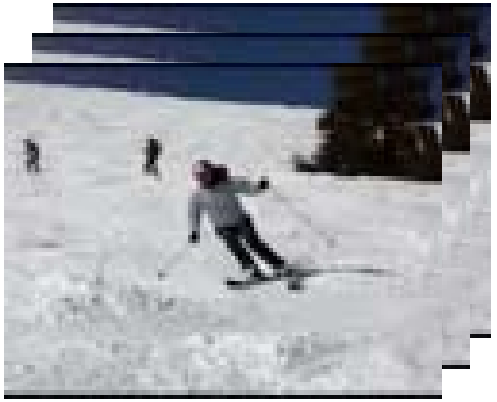
  音響特徴: Mel-Frequency Cepstral Coefficient (MFCC)
- Multi-frame

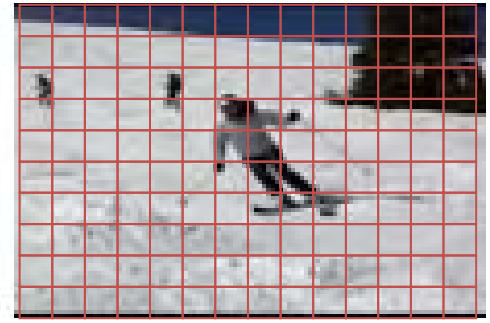
  毎フレーム、１フレームおき、２秒に１フレームなど

画像特徴

1) SIFT-Har
  - Harris-affine detector
  - Multi-frame (every two frame)
2) SIFT-Hes
  - Hessian-affine detector
  - Multi-frame (every two frame)
3) SIFTH-Dense
 - SIFT + Hue histogram
 - 30,000 samples in a key frame
4) HOG-Dense
 - 32 dim HOG feature
 - 10,000 samples in a key frame
5) HOG-Sub
 - Temporal differential of HOG
 - Detect movement

※ Reduce dim to 32 by PCA

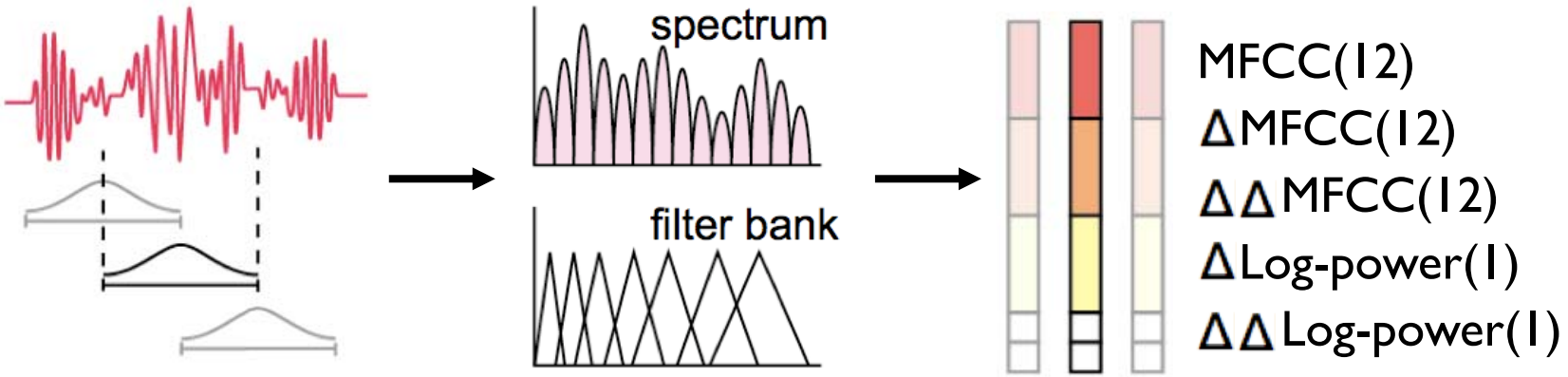




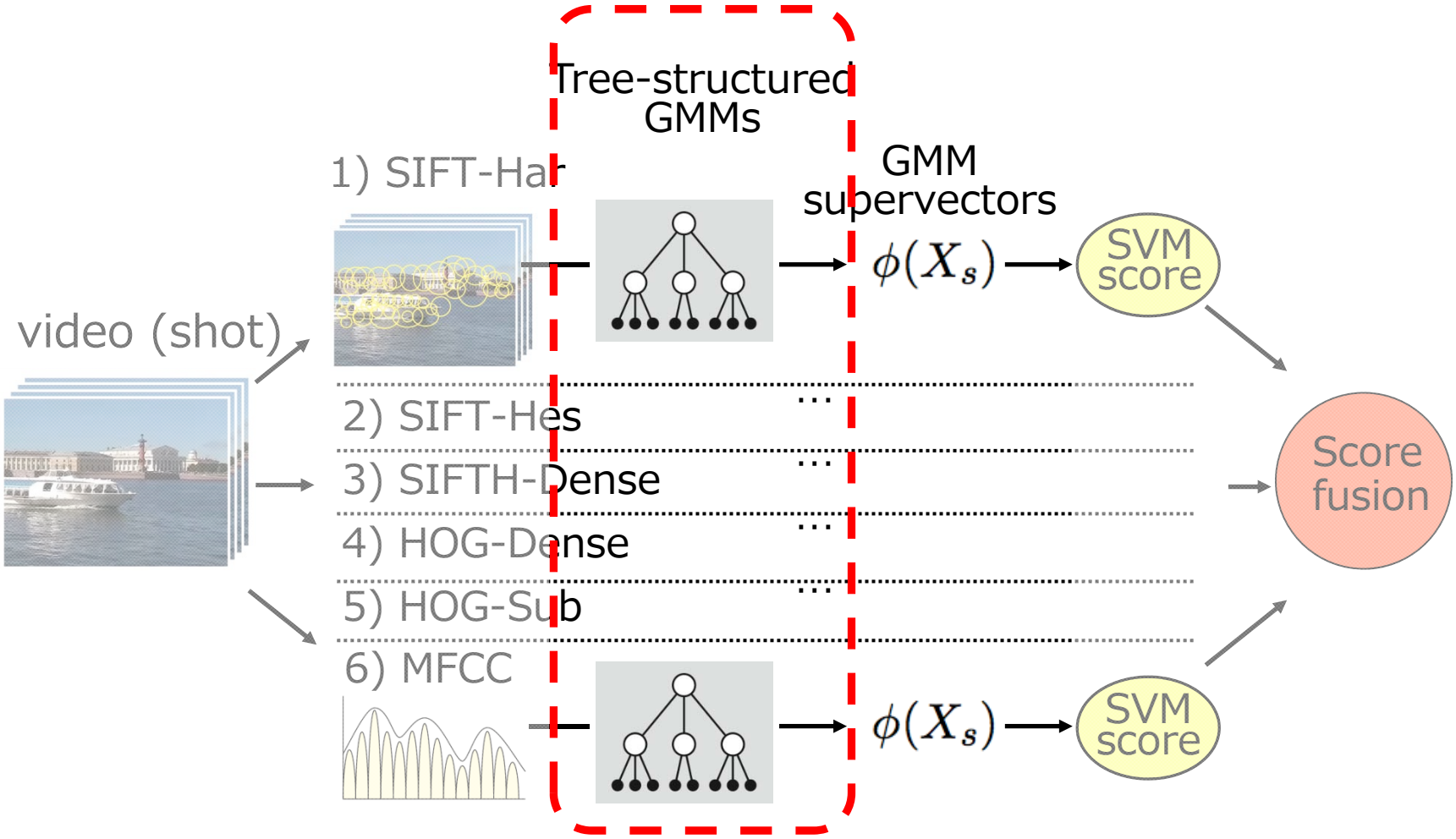
音響特徴: MFCC

Mel-frequency cepstral coefficients

音声認識、音響イベント認識でよく用いられる



コンセプトのモデル


Tree-structured GMMs
GMM supervectors
video (shot)
1) SIFT-Har
2) SIFT-Hes
3) SIFTH-Dense
4) HOG-Dense
5) HOG-Sub
6) MFCC
$\phi(X_s)$
SVM score
$\phi(X_s)$
SVM score
Score fusion

音声技術その１

Gaussian Mixture Model (GMM)

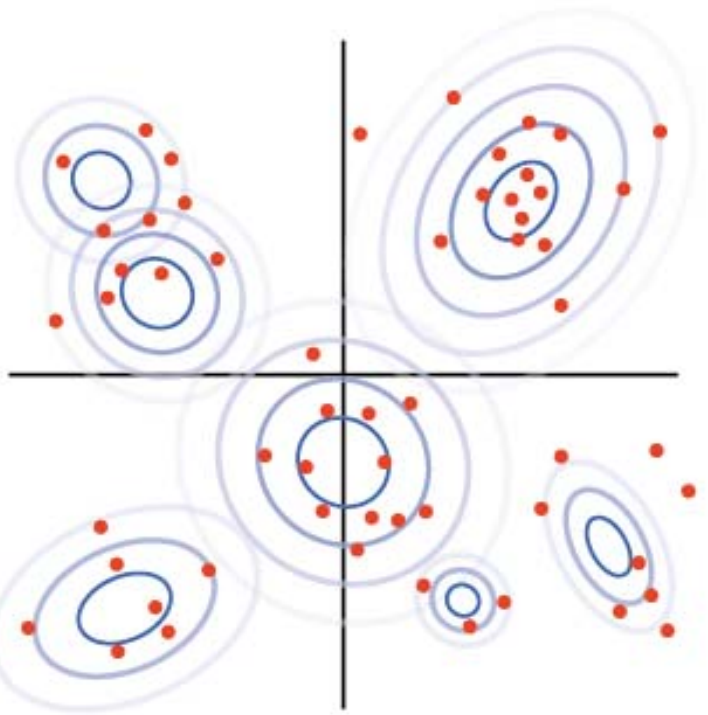
ガウス分布の重み付け和

$X_{\mathrm{F}} = \{x_i\}_{i=1}^{n}$　：　入力特徴

$$p(x|\theta) = \sum_{k=1}^{K} w_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

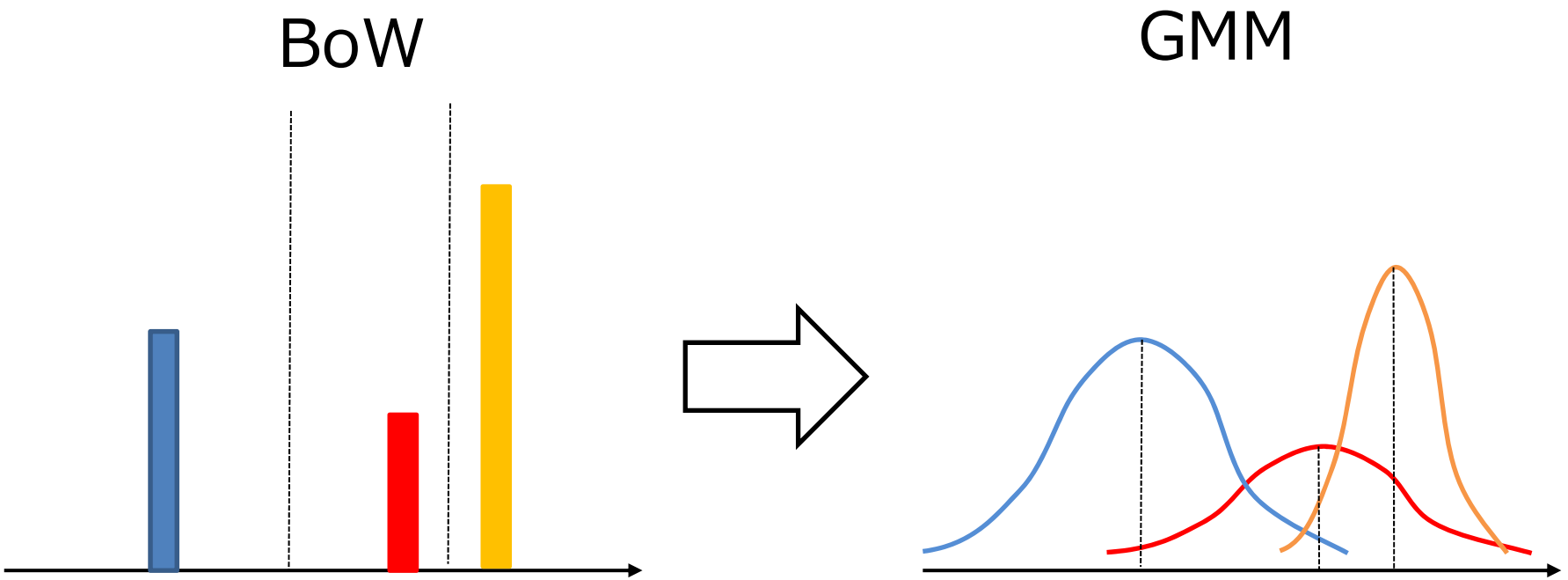
w_k:　　混合成分 k の重み($\sum w_k = 1$)

μ_k, Σ_k:　混合成分 k の平均と分散



各々のショットをGMMでモデル化

GMM は BoW の拡張



Code vector ⇒ Gaussian mean

Histogram ⇒ Weight distribution

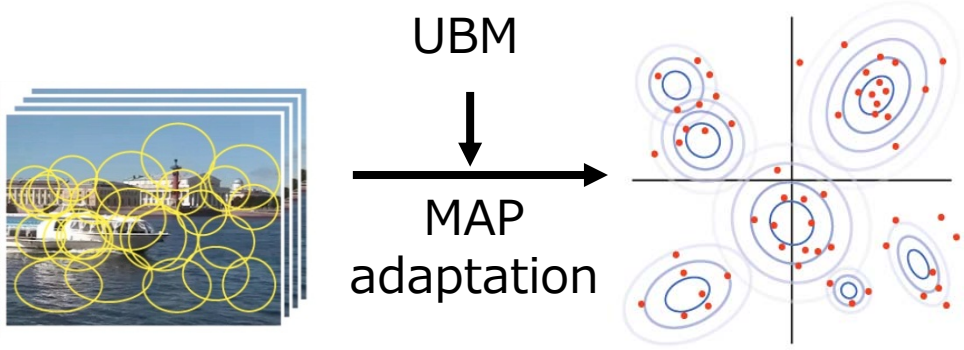
Red Color: ショットごとに推定

音声技術その2
Maximum A Posteriori (MAP) 適応

- 転移学習の一手法
- GMMの平均ベクトルに対し、その事前分布を仮定

1. すべての学習データを用いて Universal background model (UBM) を推定
2. UBMを初期モデルとして、MAP適応により、GMMの平均ベクトルを推定する。

事前分布：UBMにおける、対応する分布



より少ないデータ量で高精度な推定

MAP 適応

入力: $x_1, \ldots, x_n$

$\mu_k^{(U)}, \Sigma_k^{(U)}$: UBMにおける混合成分k の平均と分散

$\hat{\mu}_k$: 混合成分 k の平均のMAP推定量

τ : 制御パラメータ

$$\hat{\mu}_k = \frac{\tau \mu_k^{(U)} + \sum_{i=1}^{n} c_{ik} x_i}{\tau + C_k}$$

where

$$c_{ik} = \frac{w_k \mathcal{N}(x_i | \mu_k^{(U)}, \Sigma_k^{(U)})}{\sum_{k=1}^{K} w_k \mathcal{N}(x_i | \mu_k^{(U)}, \Sigma_k^{(U)})}, \quad C_k = \sum_{i=1}^{n_s} c_{ik}$$

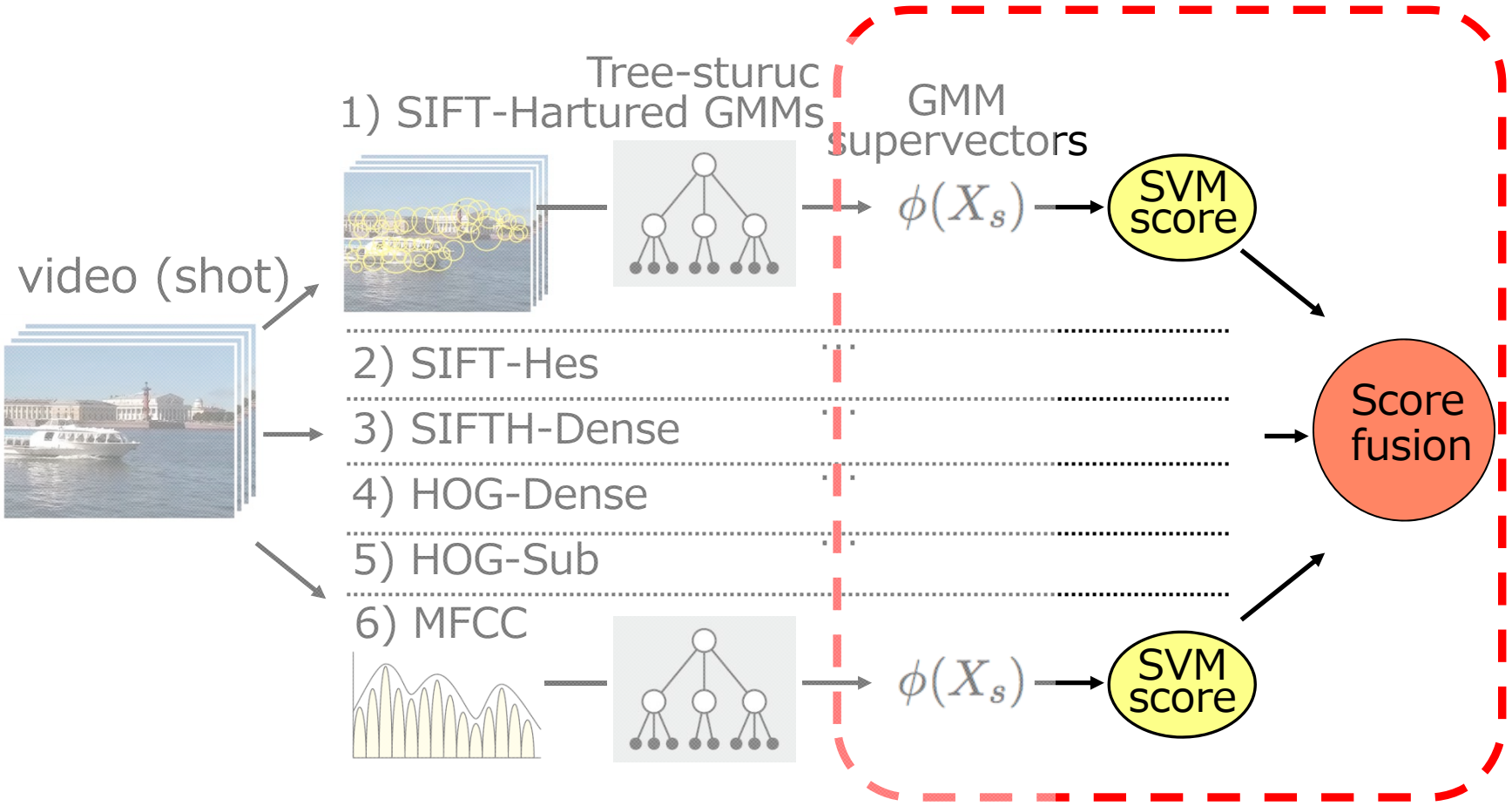
入力 x_i に対する混合成分 k の負担率

漸近的な性質をもつ

データが少なくなると、$\hat{\mu}_k \to \mu_k^{(U)}$

データが多くなると、$\hat{\mu}_k$ は最尤推定量に近づく

識別器


Tree-sturuc
1) SIFT-Hartured GMMs
GMM
supervectors
video (shot)
$\phi(X_s)$
SVM score
2) SIFT-Hes
3) SIFTH-Dense
4) HOG-Dense
5) HOG-Sub
6) MFCC
$\phi(X_s)$
SVM score
Score fusion

GMM Supervector + SVM

1. GMMの平均ベクトルを連結 → GMM supervector

$$\phi(X_{\mathrm{F}}) = \begin{pmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \\ \vdots \\ \tilde{\mu}_K \end{pmatrix}$$

where

$$\tilde{\mu}_k = \sqrt{w_k^{(U)}}(\Sigma_k^{(U)})^{-\frac{1}{2}}\hat{\mu}_k$$

normalized mean

2. Support Vector Machine (SVM) with RBF kernel

$$k(X_{\mathrm{F}}, X'_{\mathrm{F}}) = \exp\left(-\gamma\|\phi(X_{\mathrm{F}}) - \phi(X'_{\mathrm{F}})\|_2^2\right),$$

GMMに対するFisher Kernelの近似

Score fusion

SVM スコアの重み付け和

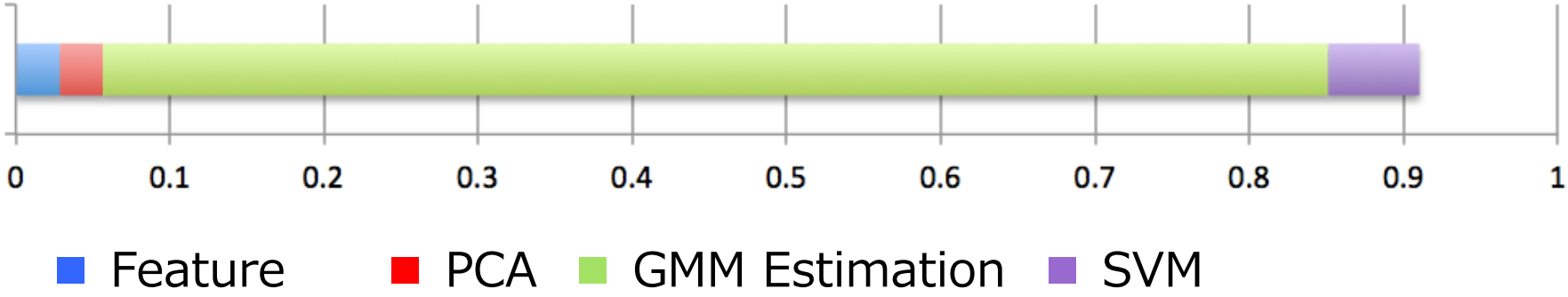
$$f(X) = \sum_{\mathrm{F} \in \mathcal{F}} \alpha_{\mathrm{F}} f_{\mathrm{F}}(X_{\mathrm{F}}), \quad 0 \leq \alpha_{\mathrm{F}} \leq 1, \quad \sum_{\mathrm{F}} \alpha_{\mathrm{F}} = 1$$

where $\mathcal{F} = \{\text{SIFT-Har}, \text{SIFT-Hes}, \text{SIFTH-Dense}, \text{HOG-Dense}, \text{HOG-Sub}, \text{MFCC}\}$

重みは Concept ごとに cross validation で決定

計算量

HOG-Dense 特徴を用いたときの計算時間 (sec)



GMMの推定が大部分

負担率の計算の高速化

High cost!

$$\hat{\mu}_k = \frac{\tau \hat{\mu}_k^{(\mathrm{U})} + \sum_{i=1}^{n} c_{ik} x_i}{\tau + C_k}$$

where

$$c_{ik} = \frac{w_k \mathcal{N}(x_i | \mu_k^{(U)}, \Sigma_k^{(U)})}{\sum_{k=1}^{K} w_k \mathcal{N}(x_i | \mu_k^{(U)}, \Sigma_k^{(U)})}, \quad C_k = \sum_{i=1}^{n_s} c_{ik}$$

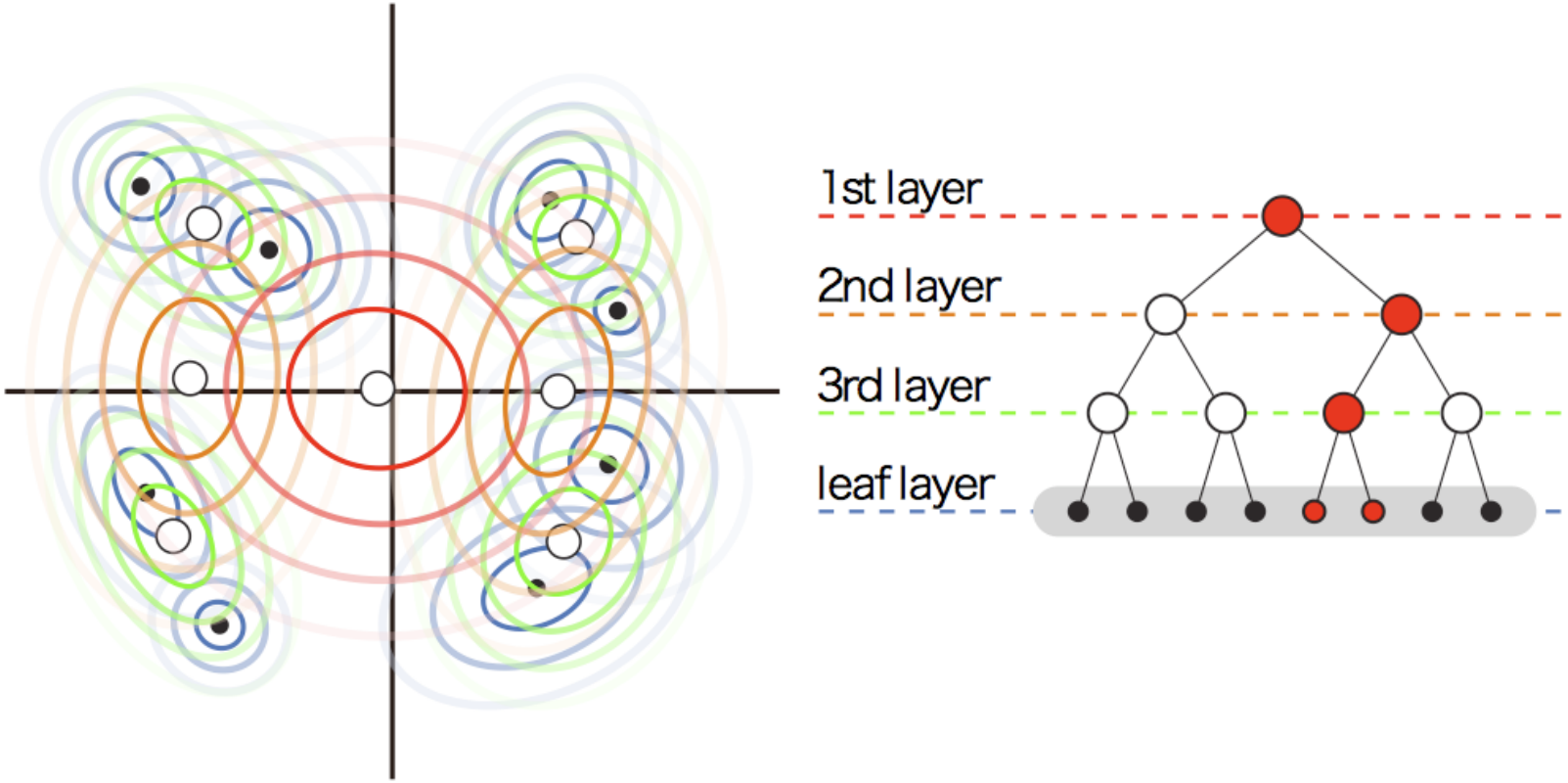
入力 x_i に対する混合成分 k の負担率

音声技術その3

木構造GMMを用いた高速計算

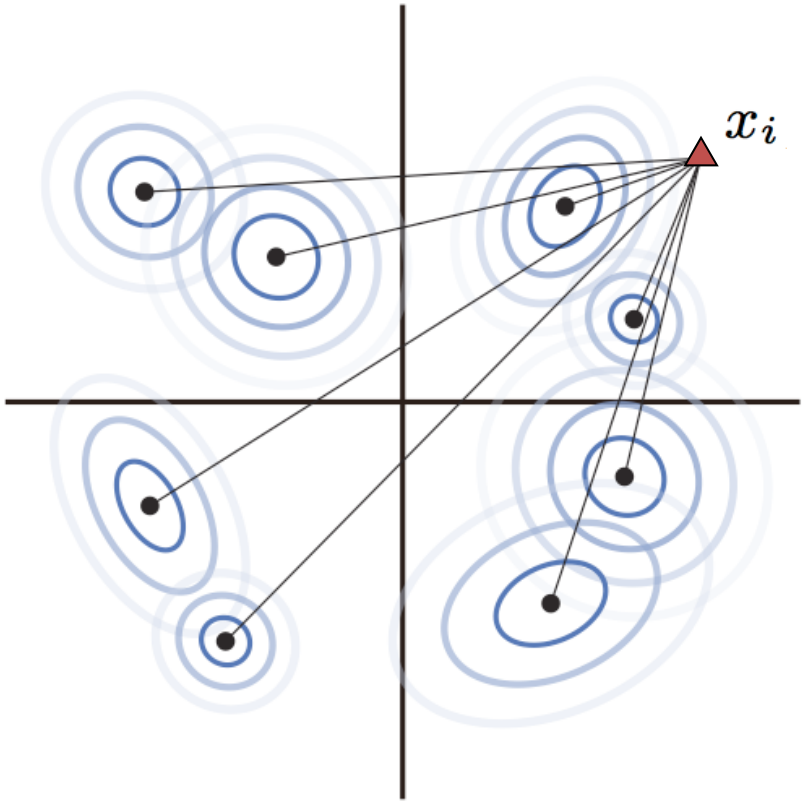
入力がどの混合成分に属するか?

(BoW:　入力をどのコードに割り当てるか?)

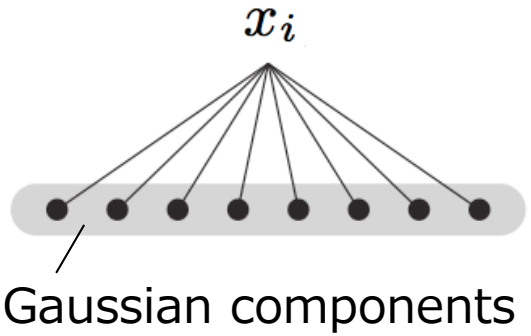


負担率

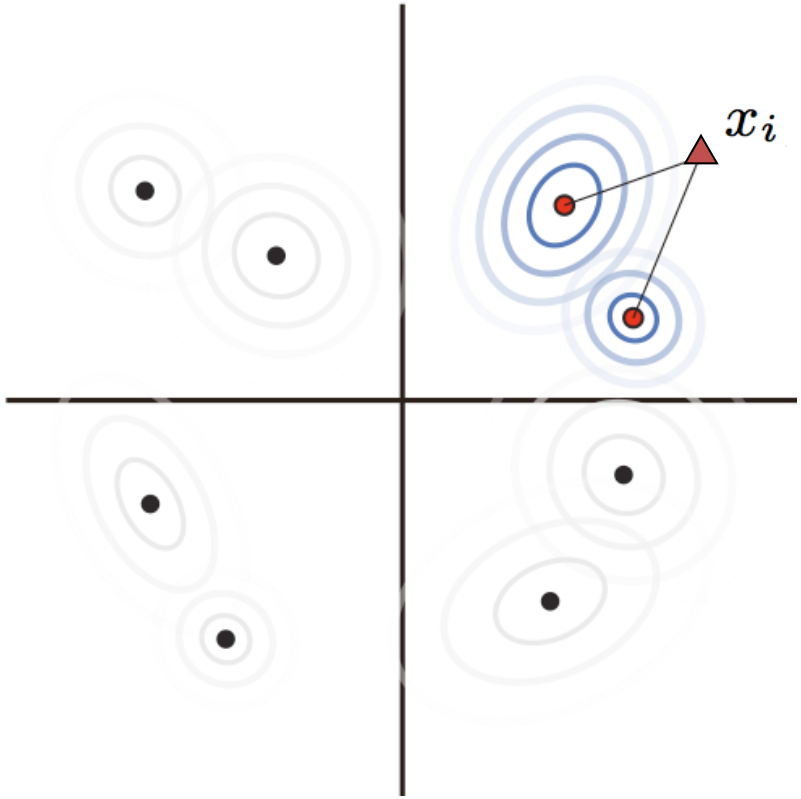
c_{ik} ： 局所特徴 x_i に対する混合成分 k の負担率



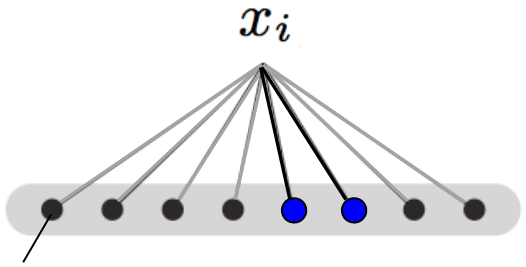
$$c_{ik} = \frac{w_k \mathcal{N}(x_i | \mu_k^{(U)}, \Sigma_k^{(U)})}{\sum_{k=1}^{K} w_k \mathcal{N}(x_i | \mu_k^{(U)}, \Sigma_k^{(U)})},$$



特徴量空間の一部でのみ計算



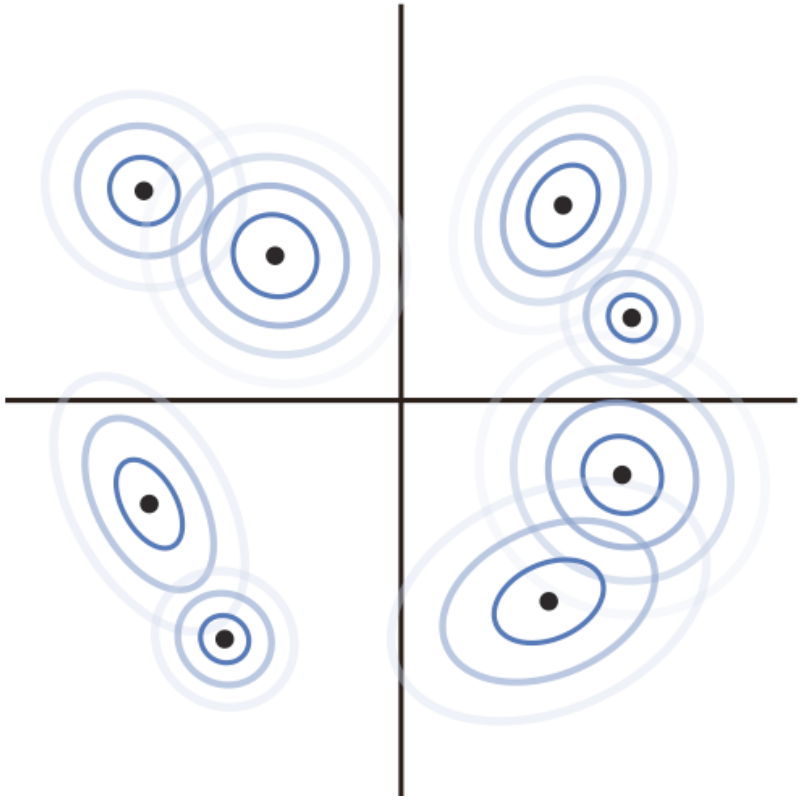
$$c_{ik} = \frac{w_k \mathcal{N}(x_i|\mu_k^{(U)}, \Sigma_k^{(U)})}{\sum_{k=1}^{K} w_k \mathcal{N}(x_i|\mu_k^{(U)}, \Sigma_k^{(U)})},$$



Gaussian components

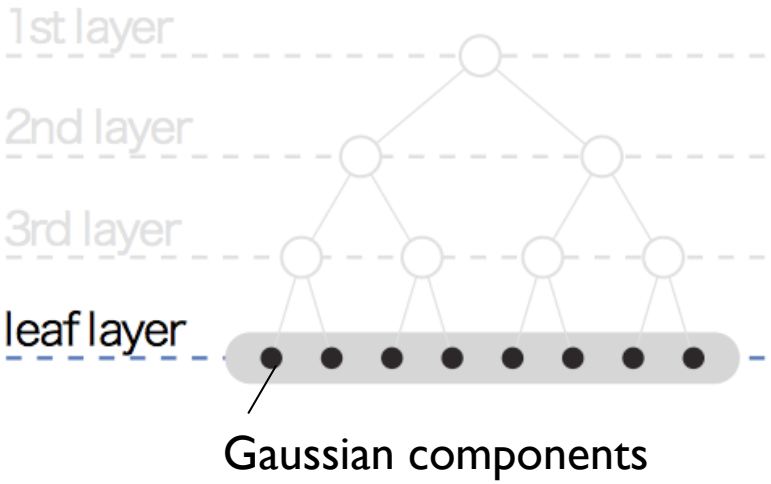
木構造GMM (1)

Leaf layer



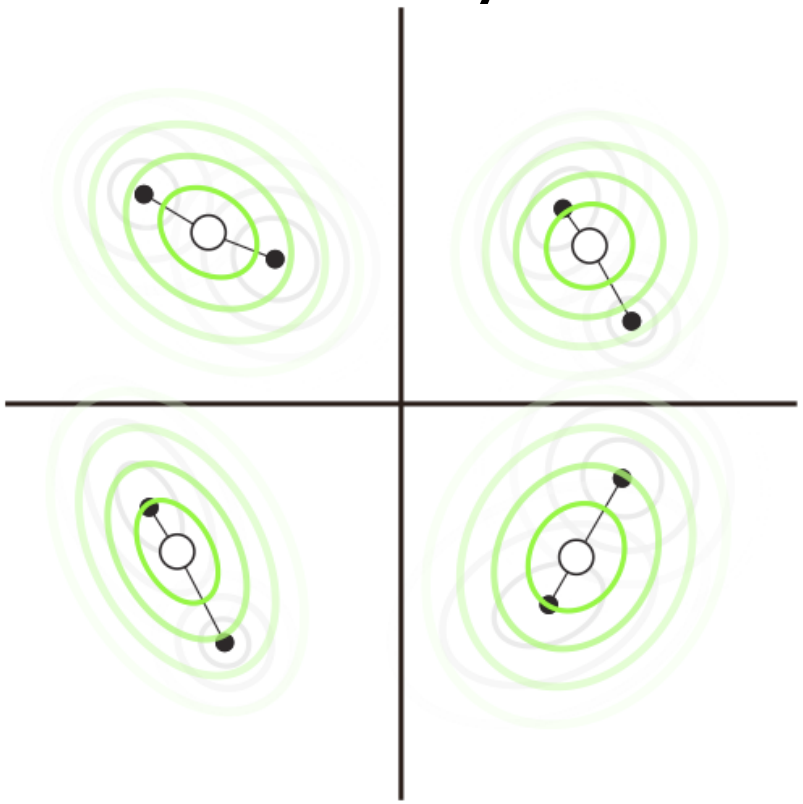
混合成分間の距離：

Symmetric KL divergence

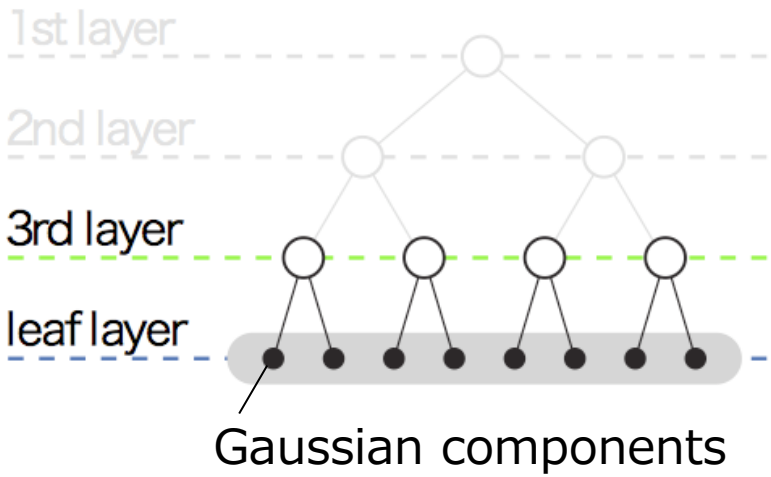


木構造GMM (2)

Non-leaf layers



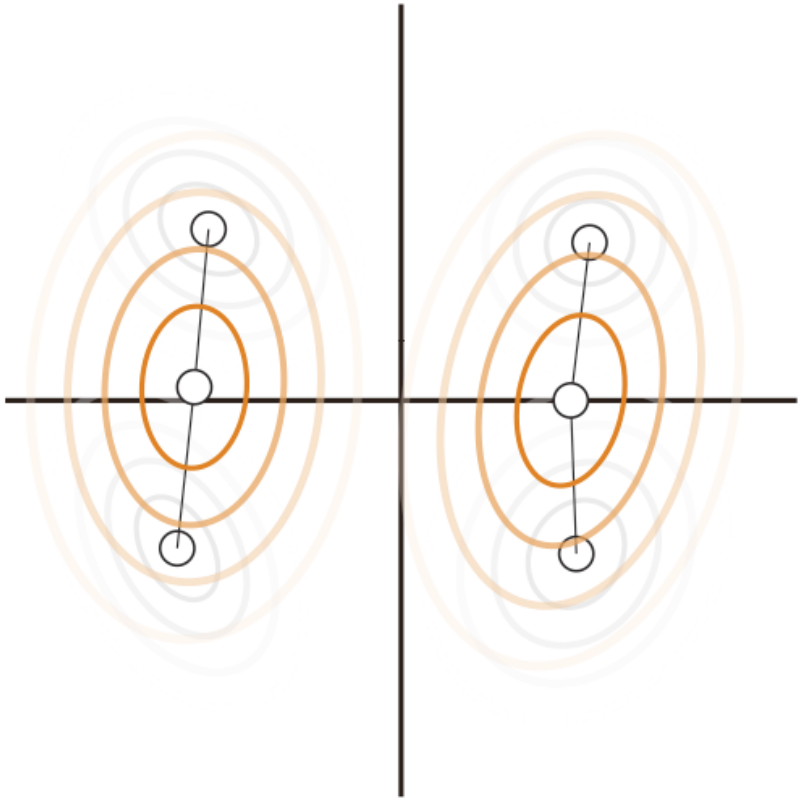
ノードのガウス分布は、リーフのガウス分布集合を近似

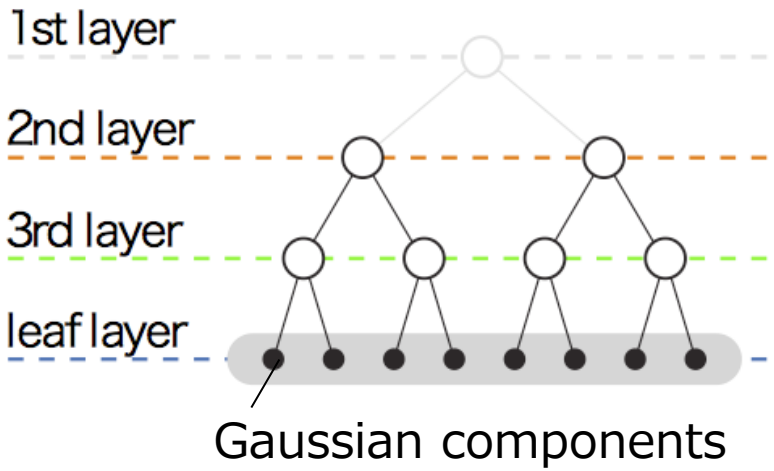


木構造GMM (3)

Non-leaf layers

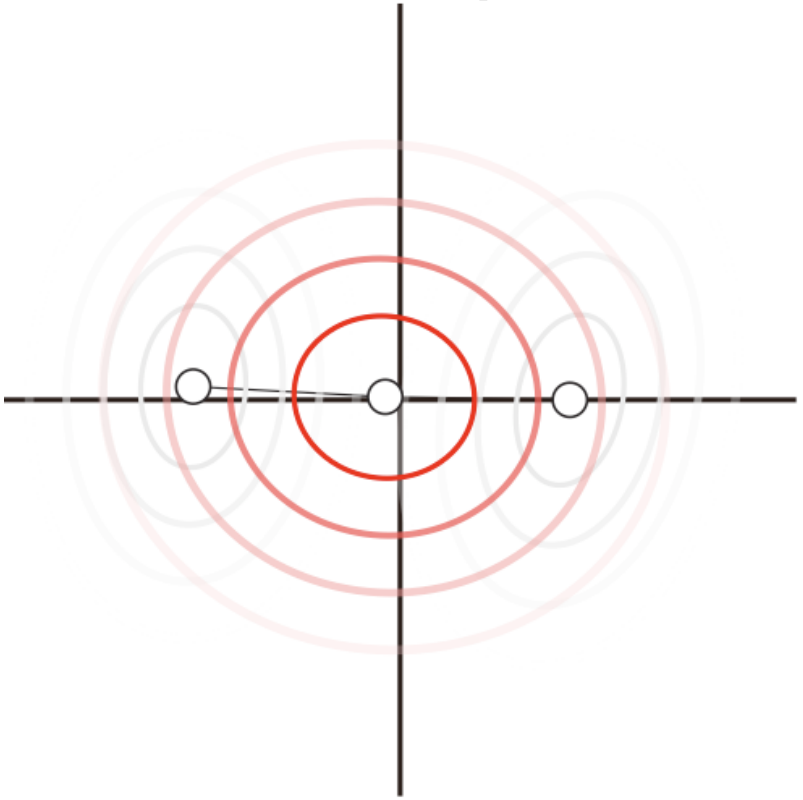
ノードのガウス分布は、リーフのガウス分布集合を近似



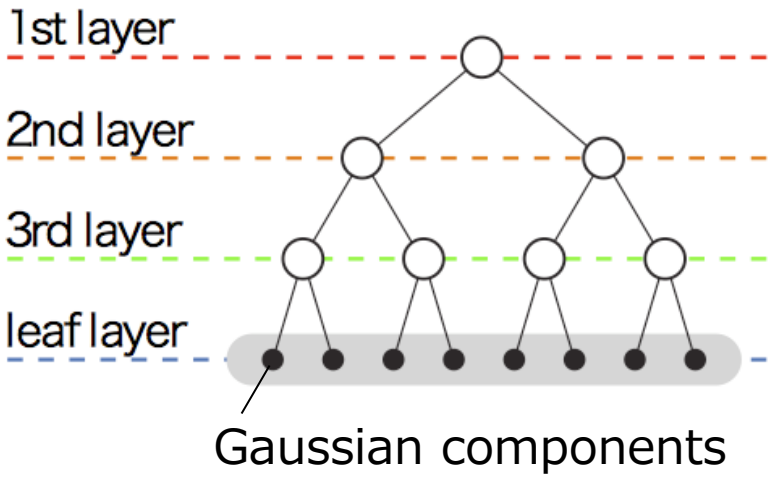


木構造GMM (4)

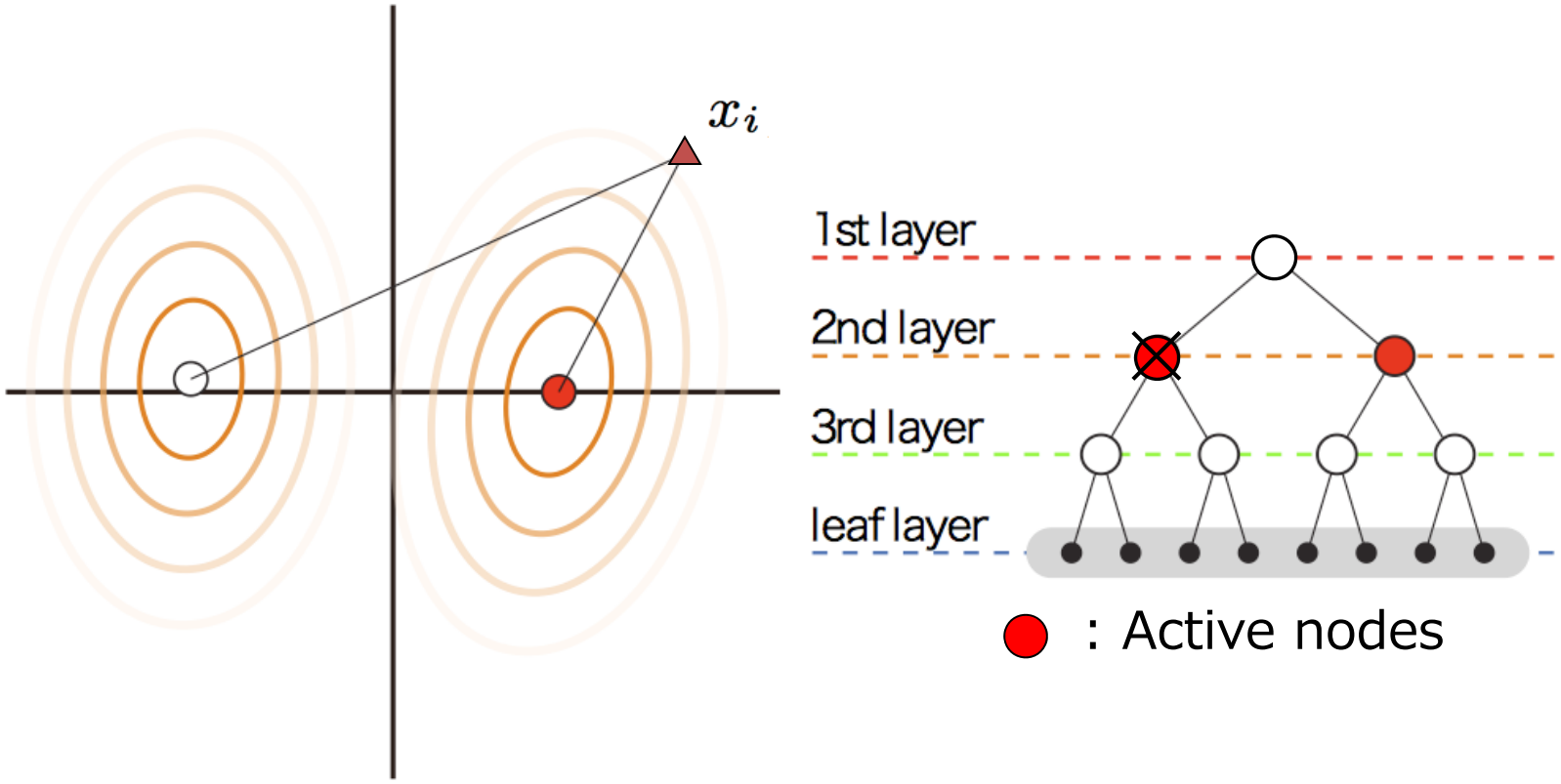
Non-leaf layers



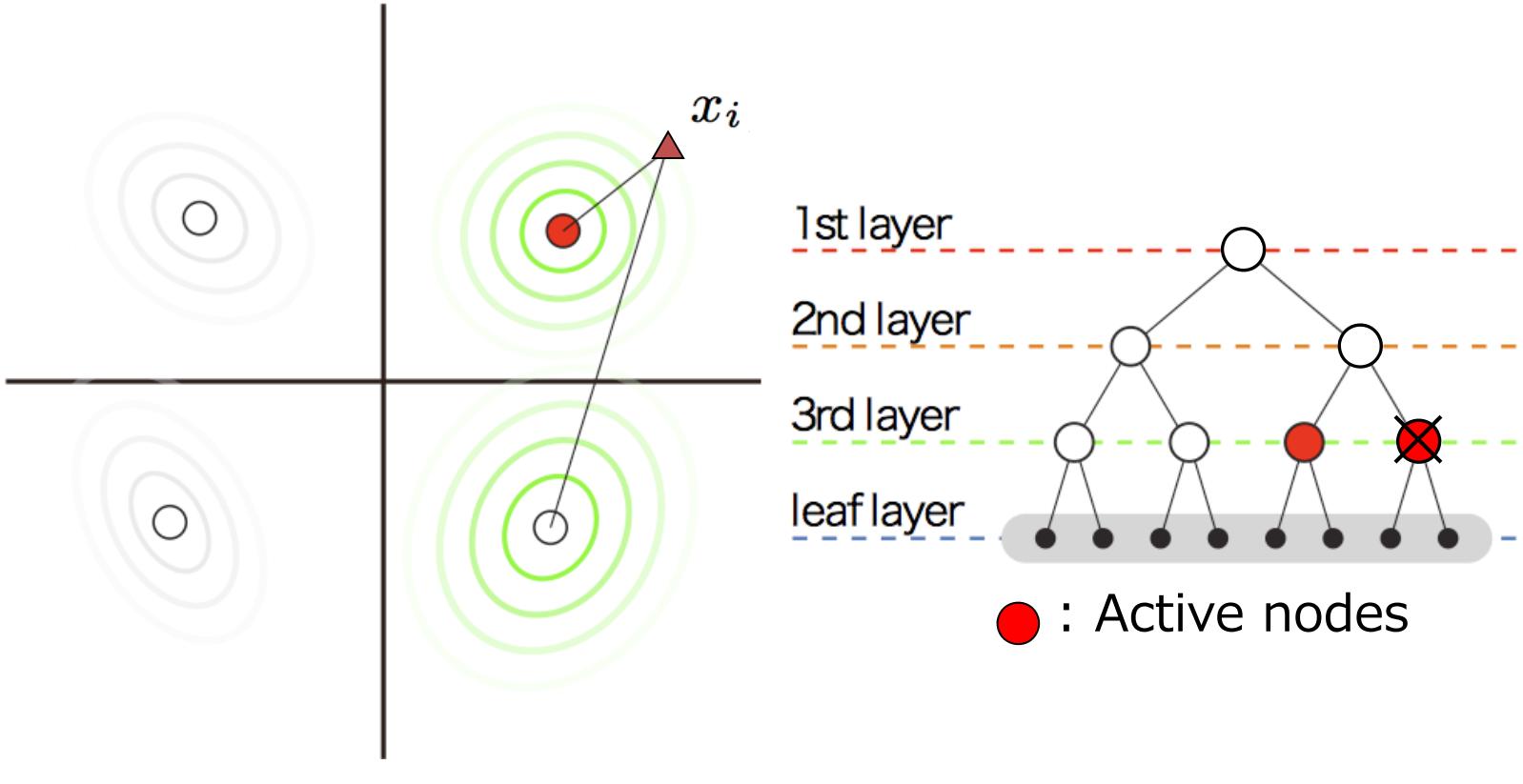
ノードのガウス分布は、リーフのガウス分布集合を近似



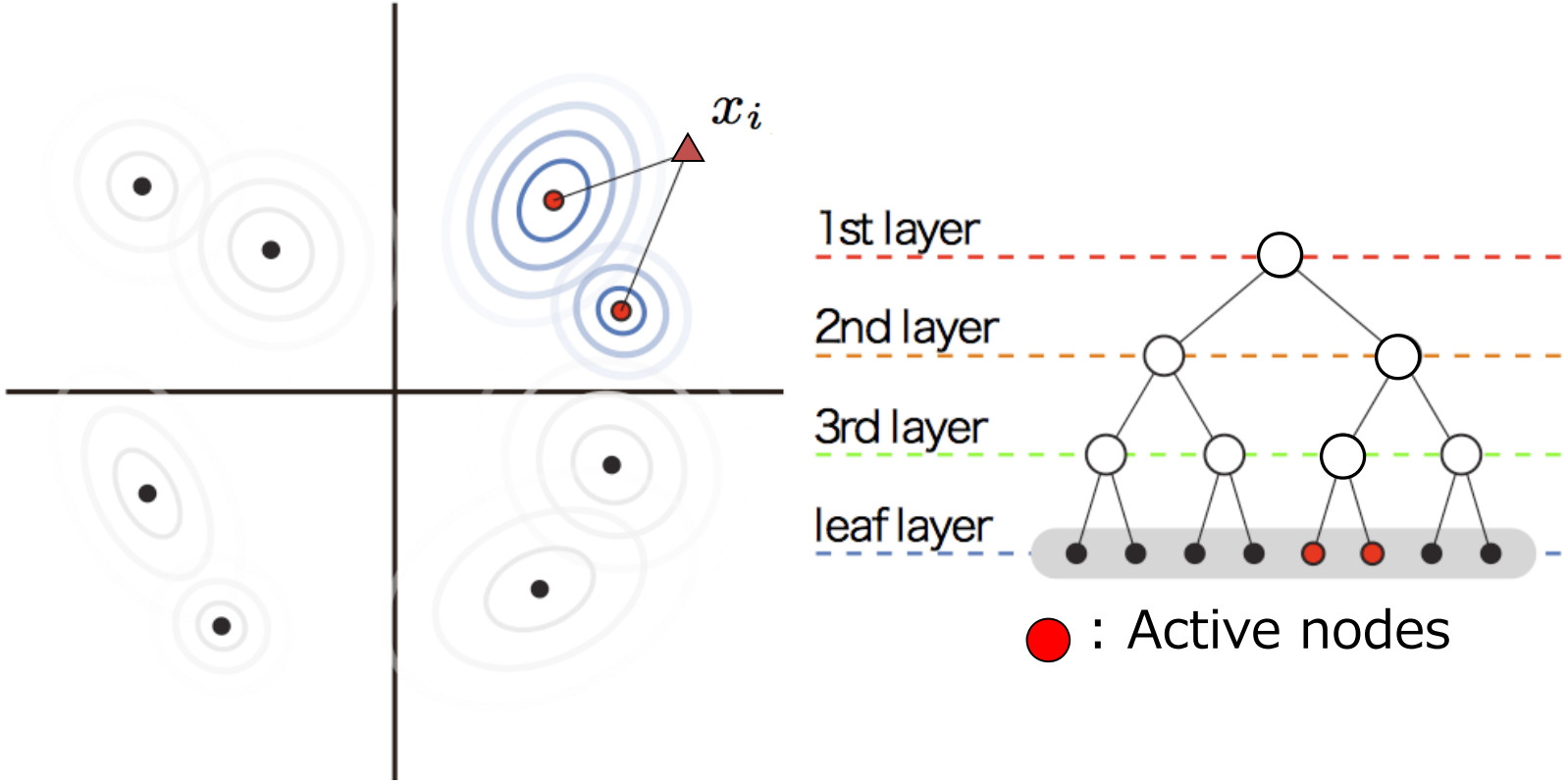
高速サーチ (1)



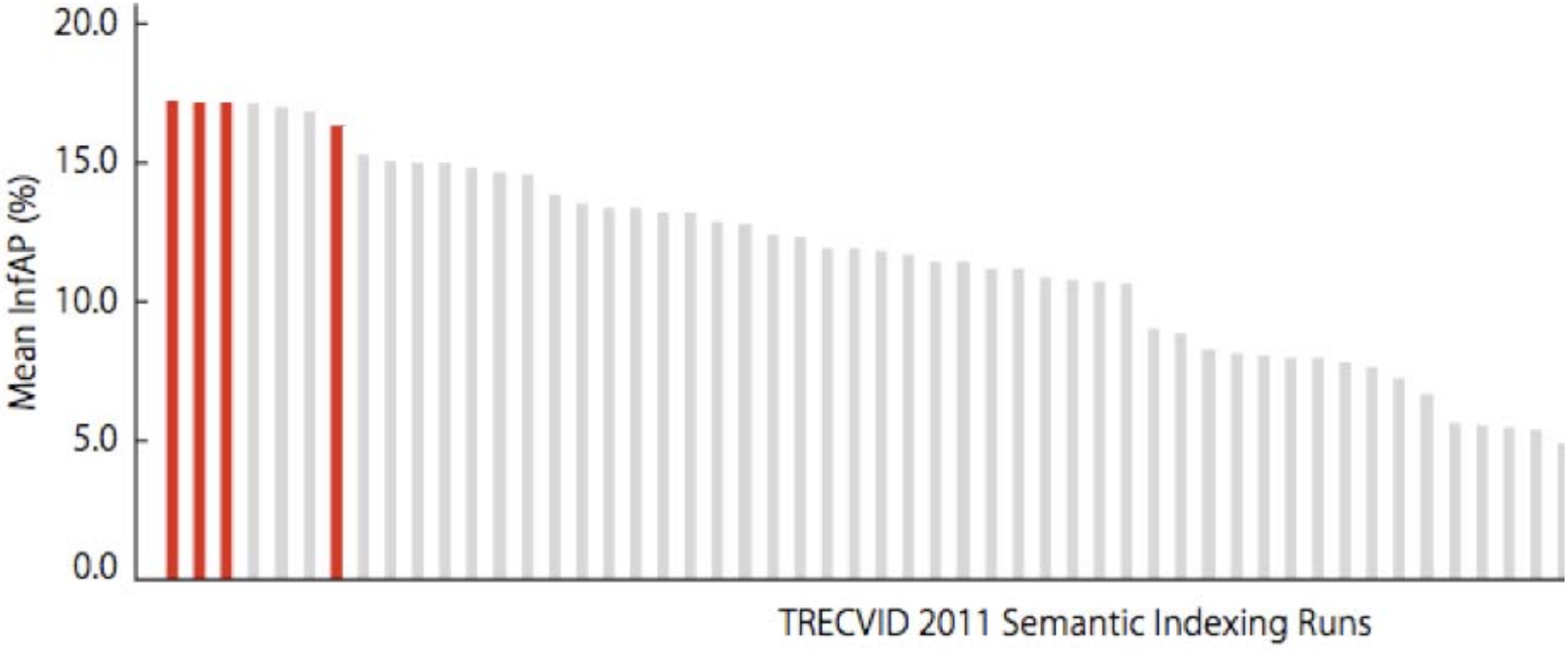
高速サーチ (2)



高速サーチ (3)

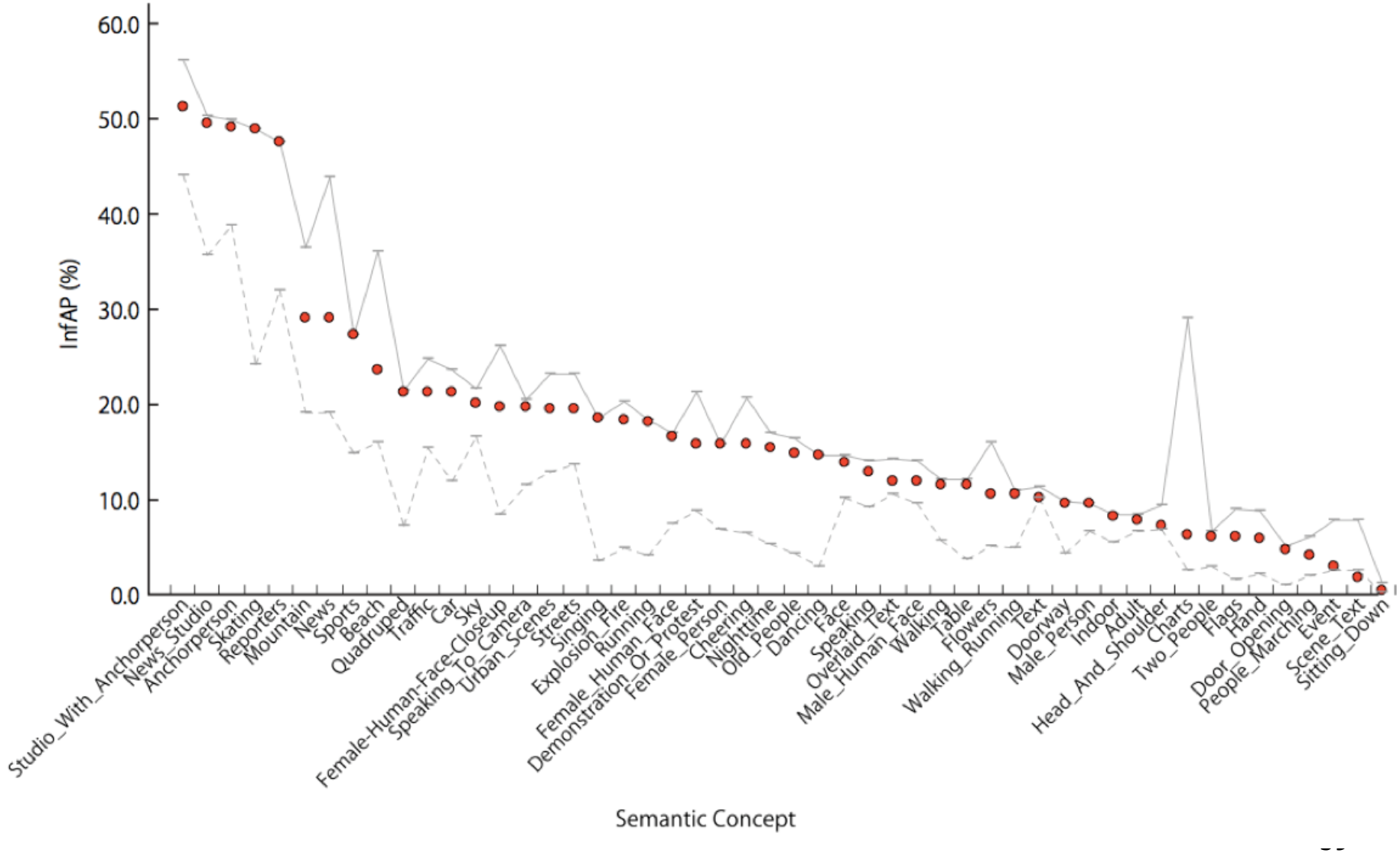


TRECVID2011 SIN の結果

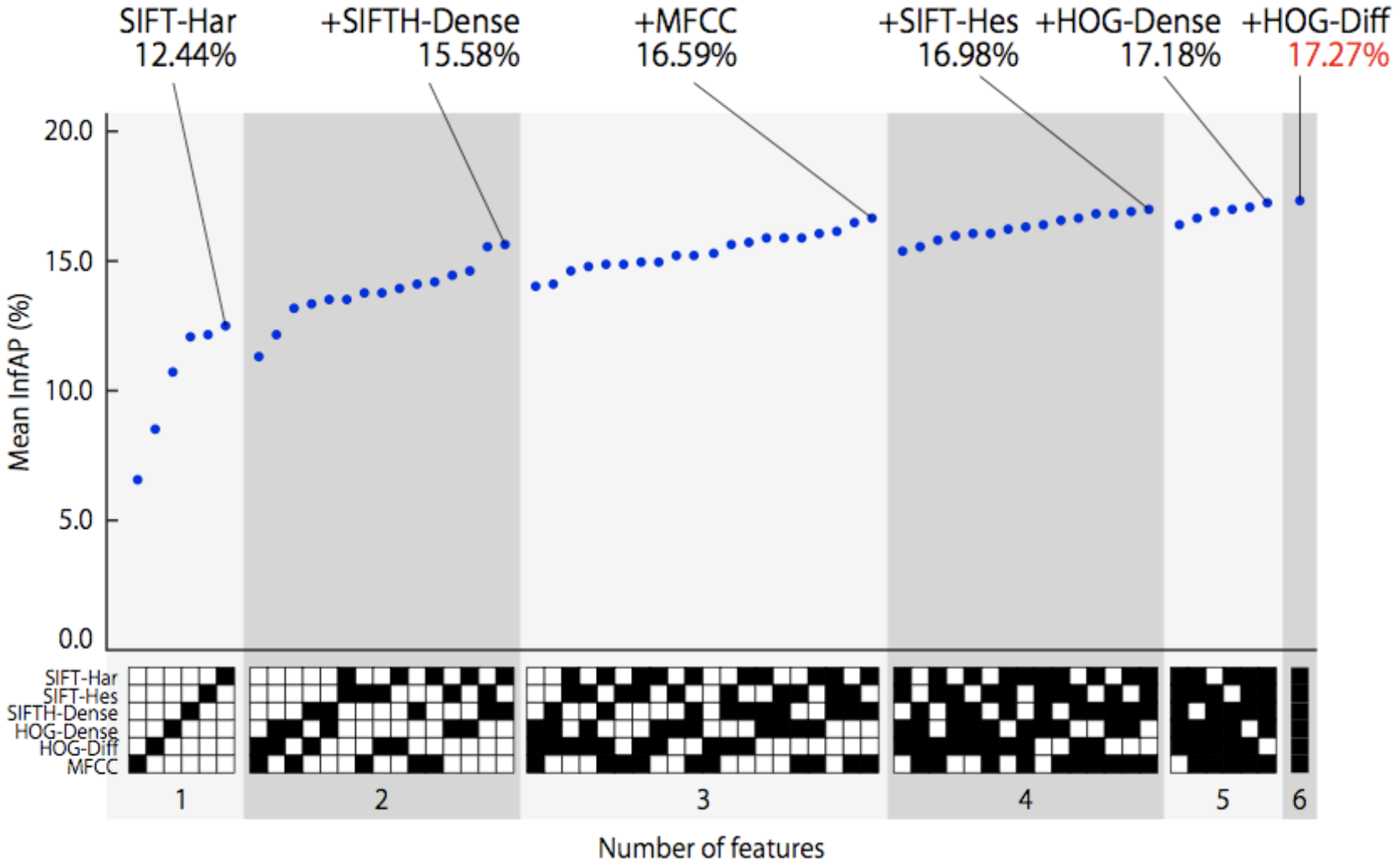


※ Mean InfAP: Inferred AP averaged over all concepts

コンセプト毎のAverage Precision

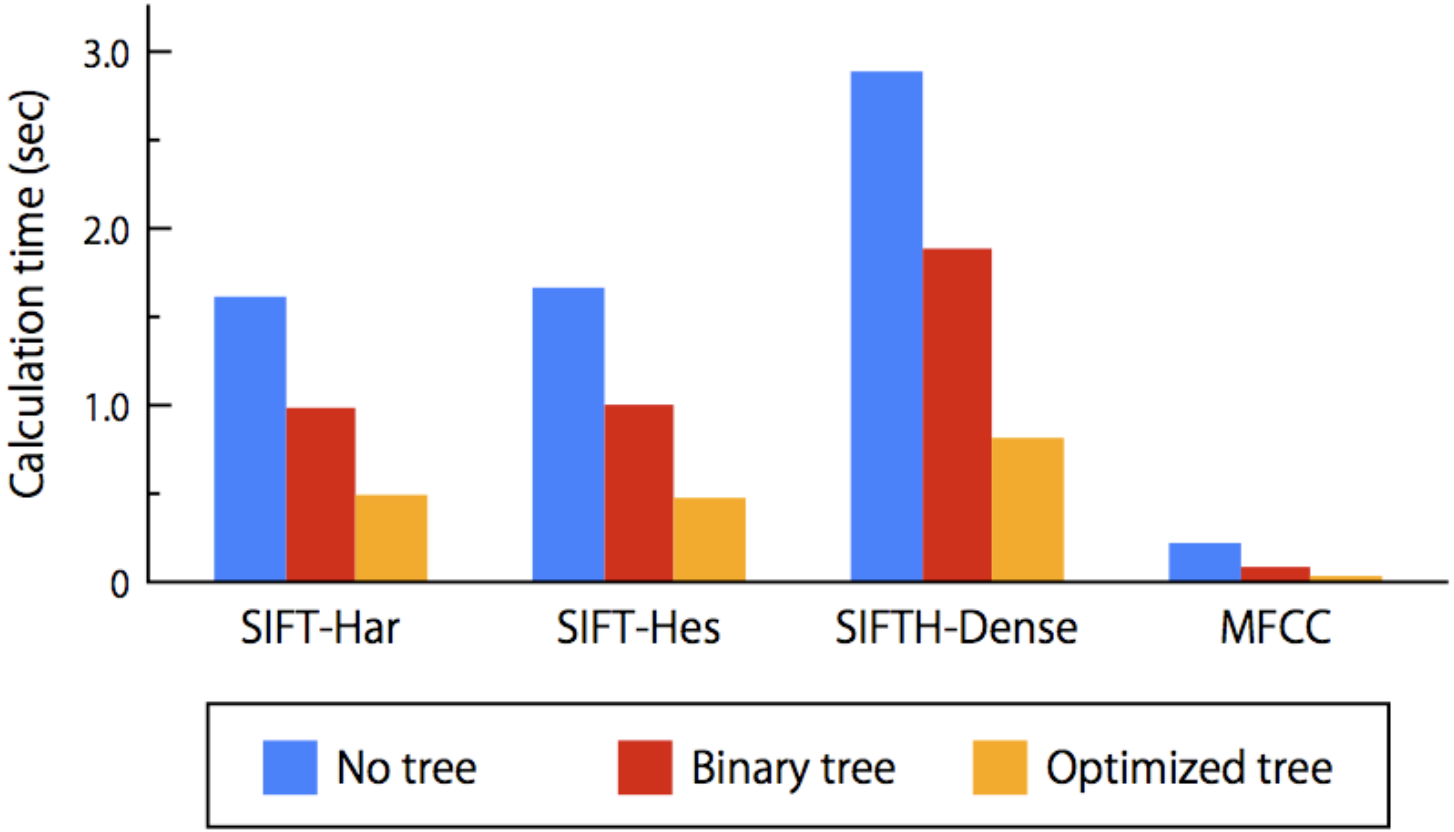


効果的な低次特徴は？


SIFT-Har
12.44%
+SIFTH-Dense
15.58%
+MFCC
16.59%
+SIFT-Hes
16.98%
+HOG-Dense
17.18%
+HOG-Diff
17.27%
Mean InfAP (%)
20.0
15.0
10.0
5.0
0.0
SIFT-Har
SIFT-Hes
SIFTH-Dense
HOG-Dense
HOG-Diff
MFCC
1
2
3
4
5
6
Number of features

GMM推定の計算時間

検出性能の劣化なしに、4.2 倍の高速化



Multimedia Event Detection (MED)

Multimedia Event Detection (MED)

目的

ビデオクリップからのイベント検出

e.g. Batting a run in

Making a cake

SINより高次の対象

スポーツ番組からのハイライト検出

をインターネット映像まで延長

データベース

HAVIC：2000時間のホームビデオ

Linguistic data consortium (LDC)が提供

MED (2)

- 2010に開始された新しいタスク
- 2011年は18チーム(日本からは5チーム)
- 米国情報省(IARPA)のAutomated Low-Level Analysis and Description of Diverse Intelligence Video (ALADDIN) プロジェクトが援助

HAVIC データベース

- ビデオクリップ(２分程度)： 3488個
- サンプル：各々のイベントに付き100個
  (半分が開発用、半分がテスト用)

2010 (3 events)	2011 (10 events)	
Assembling a shelter	Birthday party	Making a sandwich
Batting a run in	Changing a vehicle tire	Parade
Making a cake	Flash mob gathering	Parkour
	Getting a vehicle unstuck	Repairing an appliance
	Grooming an animal	Working on a sewing project

評価基準：

- Missed Detection Probability P_{miss}
  1 – Recall
- False Alarm Probability P_{FA}
  False Alarm / Clips with no events
- Normalized Detection Cost (NDC)
  上記２つを適当な重みで混合したもの

$$NDC = \frac{Cost_{miss}\, P_{miss} P_{target} + Cost_{FA}\, P_{FA}\left(1 - P_{target}\right)}{MIN\left(Cost_{miss} P_{target} + Cost_{FA}\left(1 - P_{target}\right)\right)}$$

$Cost_{Miss} = 80$
$Cost_{FA} = 1$
$P_{target} = 0.001$

トレンド

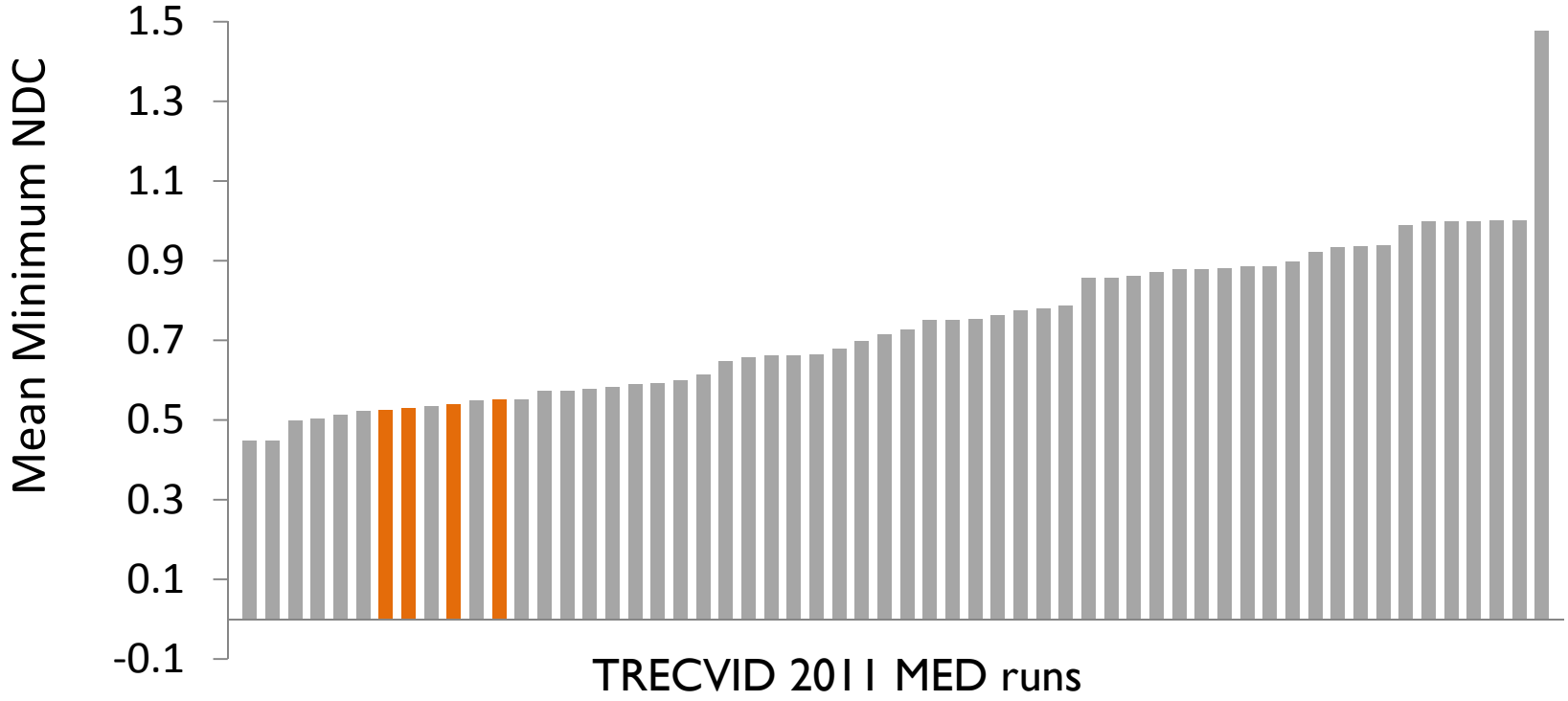
- SINの手法を応用
  多くの特徴＋BoW+SVM

- 時空間特徴
  STIP (Space-time interest point), etc.

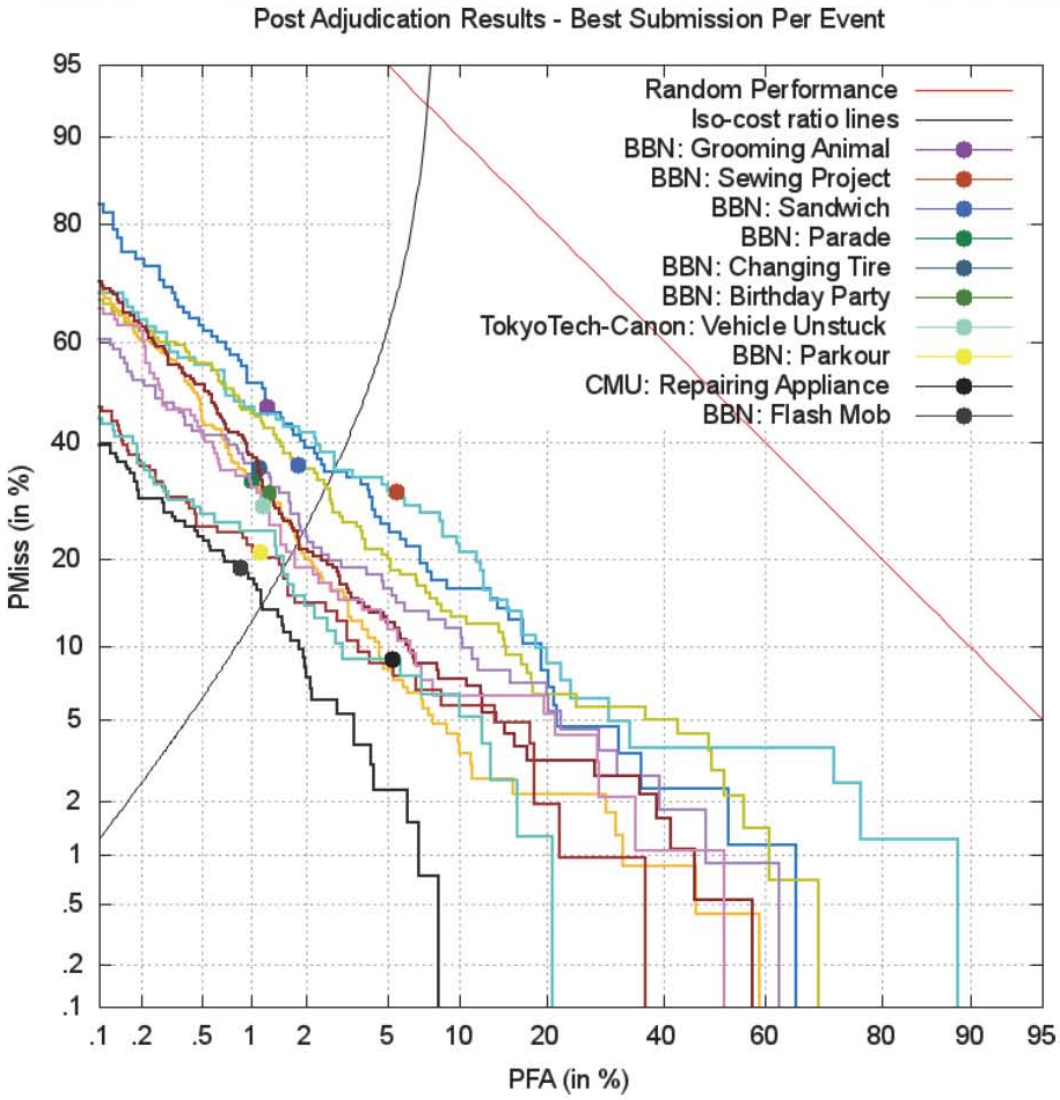
- コンテキストのモデル化（Semantic model)
  効果があまりない ← データが少ない？

- 音声認識、OCR
  効果なし、SINと同じ理由

TRECVID2011



	Mean MNDC	Mean ANDC
1st Team	0.448	0.465
2nd Team	0.499	0.522
3rd Our team	0.525	0.556

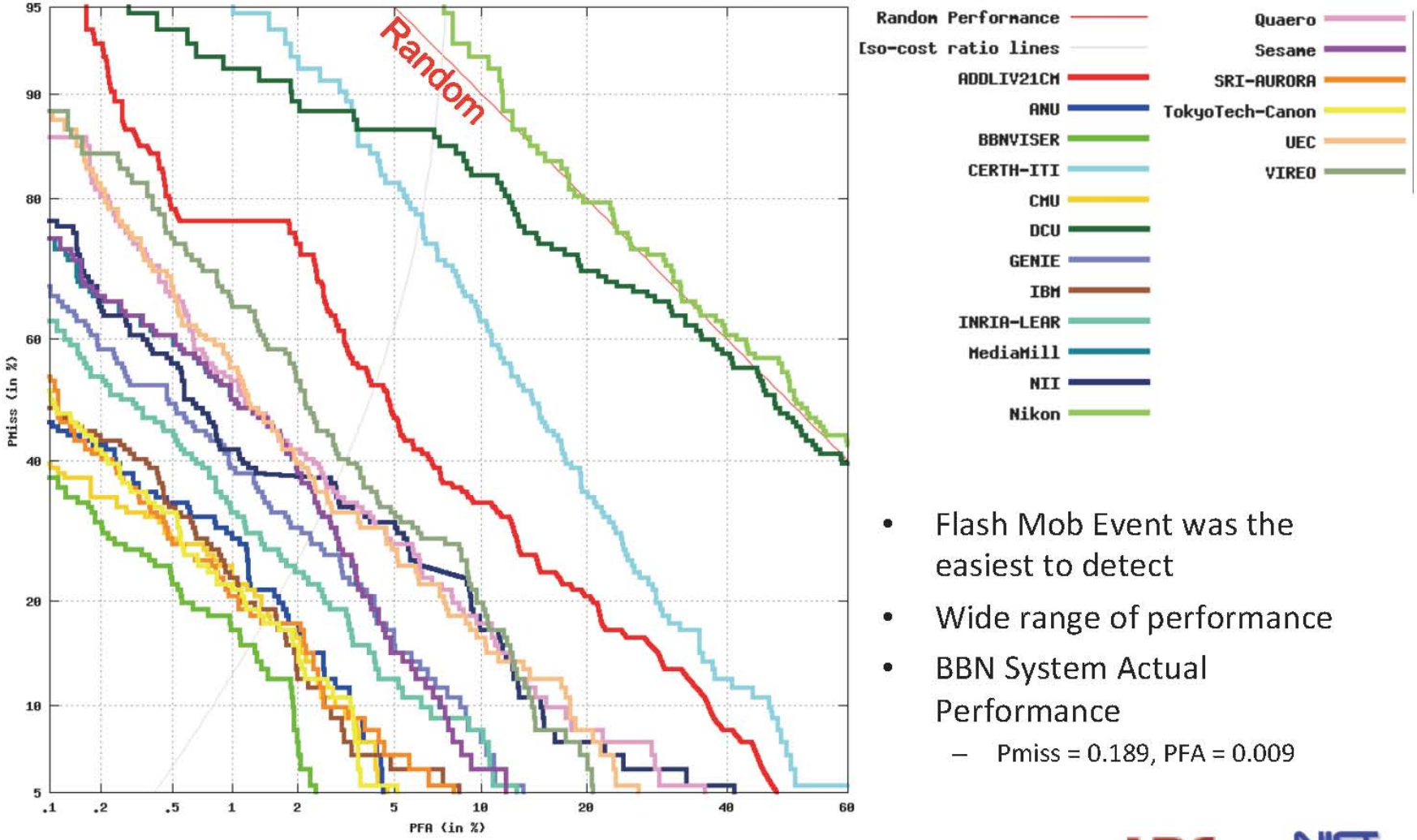


Lowest Error Primary System per Event

(Based on Iso-Ratio Line)

- Easiest: Flash mob gathering
  - PMiss = 0.1438, PFA = 0.0115
- Toughest: Grooming a animal
  - PMiss = 0.3445, PFA = 0.0275
- Error Rates more than double for both error types

Flash mob gathering (Primary systems)



- Flash Mob Event was the easiest to detect
- Wide range of performance
- BBN System Actual Performance
  - Pmiss = 0.189, PFA = 0.009

おわりに

- 頑健かつ高速な映像検索
  - 音声分野で開発された技術が性能向上に寄与
  - GMM, MAP適応, 木構造サーチ
- 単語レベル (SIN) から文レベル (MED) へ
  - 映像のコミュニケーションモデル
  - コンテキストの活用
- No data like more data
  - データ量にスケールする技術が重要
  - 計算の高速化がますます重要に
- 他に使える音声技術は?

  言語モデル、識別学習、Deep Learning, etc.