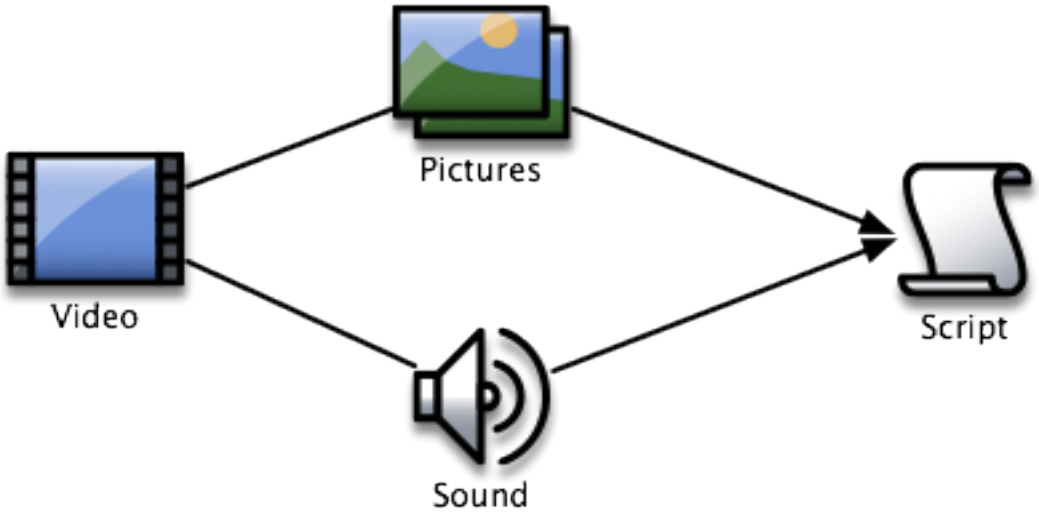
INTERNATIONAL COMPUTER SCIENCE INSTITUTE



Multimedia Event Detection for Large Scale Video

Benjamin Elizalde

Outline

- Motivation
- TrecVID task
- Related work
- Our approach (System, TF/IDF)
- Results & Processing time
- Conclusion & Future work
- Agenda

Motivation

- YouTube alone claims 72 hours uploaded per minute, with 3 billion viewers a day.

- Video need to be searched, sorted, retrieved based on content descriptions that are “higher-level”.

- High demand in industry, even higher demand in intelligence community.

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

Industry

Microsoft

CISCO

YouTube

Hrishikesh Aradhye, "Finding cats playing pianos: Discovering the next viral hit on YouTube".

TrecVID Multimedia Event Detection

- 17 teams.

- SRI AURORA
  - SRI International (SRI) (Sarnoff)
  - International Computer Science Institute, University of California, Berkeley (ICSI)
  - Cycorp
  - University of Central Florida (UCFL)
  - UMass Amherst

- ICSI contribution: **Acoustic**, visual, and multimodal methods for Video Event Detection.

Task: Multimedia Event Detection

- Given:
  - An **event kit** which consists of an event name, definition, explication, video example.
- Wanted:
  - A system that can search multimedia recordings for user-defined events.

What is Video Event Detection?

- An event:
  - is a complex activity occurring at **a specific place and time**;
    - involves **people interacting** with other people and/or objects;
    - consists of a number of human actions, processes, and activities that are loosely or tightly organized and that have significant temporal and **semantic relationships** to the overarching activity;
    - is **directly observable**.

INTERNATIONAL
COMPUTER SCIENCE
INSTITUTE

Sample Video 1: “Board Tricks”

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

Sample Video 2: “Board Tricks”

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

Test Video: “Board Tricks”

Related Work: TrecVid MED 2010

Making a cake

Batting a run in

Assembling a shelter

Related Work: TrecVid MED 2010

Human Action Concepts	*Scene Concepts*	*Audio Concepts*
▪ Person walking	▪ Indoor kitchen	▪ Outdoor rural
▪ Person running	▪ Outdoor with grass/trees visible	▪ Outdoor urban
▪ Person squatting	▪ Baseball field	▪ Indoor quiet
▪ Person standing up	▪ Crowd (a group of 3+ people)	▪ Indoor noisy
▪ Person making/assembling stuffs with hands (hands visible)	▪ Cakes (close-up view)	▪ Original audio ▪ Dubbed audio ▪ Speech comprehensible ▪ Music ▪ Cheering ▪ Clapping
▪ Person batting baseball		

Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, Subhabrata Bhattacharya, Dan Ellis, Mubarak Shah, Shih-Fu Chang: Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching, Proceedings of TrecVid 2010, Gaithersburg, MD, December 2010.

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

General Observations

Year	Number of classes
2010	3
2011	15
2012	30

- E010 Grooming an animal
- E011 Making a sandwich
- E014 Repairing an appliance
- E022 Cleaning an appliance
- E025 Marriage proposal
- E029 Winning a race without a vehicle

General Observations

- Single-model approach is problematic:
  - Too noisy
- Classifier ensembles problematic:
  - Which classifiers to build?
  - Training data?
  - Annotation?
  - Idea doesn't scale!

General Observations Audio

- Speech Recognition (incl. keyword spotting) mostly infeasible (50% of videos contain no speech, speech is arbitrary languages, quality varies.
- Acoustic event detection has same issues as visual object recognition.
- Music indicative of events but not with high confidence.

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

Theoretical Framework

- **Concepts** without **percepts** are empty; **percepts** without **concepts** are blind. (Kant)

- **Percepts** are an impression of an object obtained by use of the senses.

- Concepts = Events
- Percepts = Observations

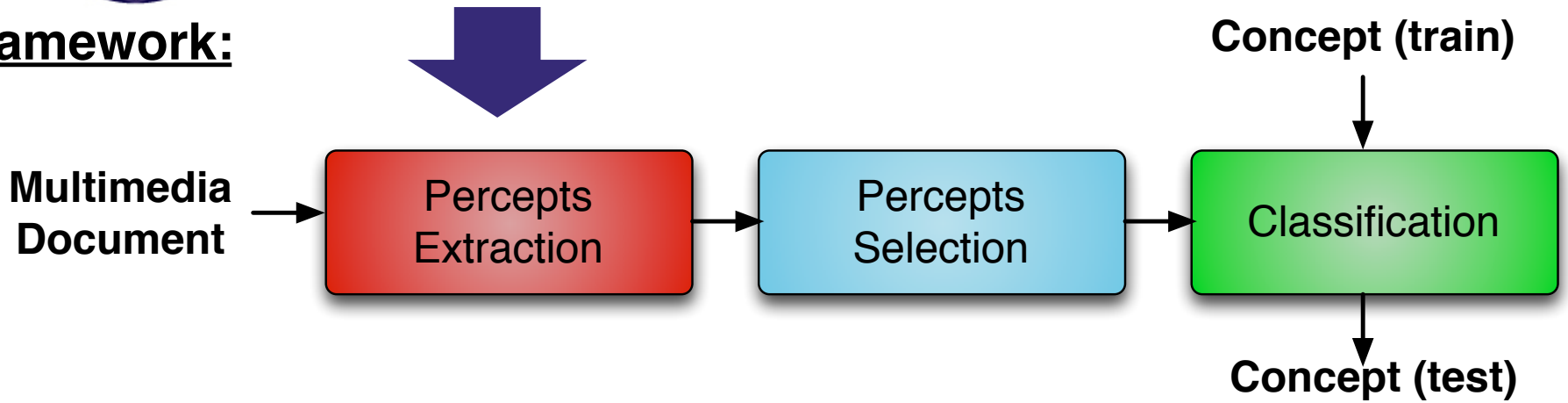
Elizalde Benjamin, Gerald Friedland et al. There is No Data Like Less Data: Percepts for Video Concept Detection on Consumer-Produced Media. ACM Multimedia 2012 Workshop.

Contribution

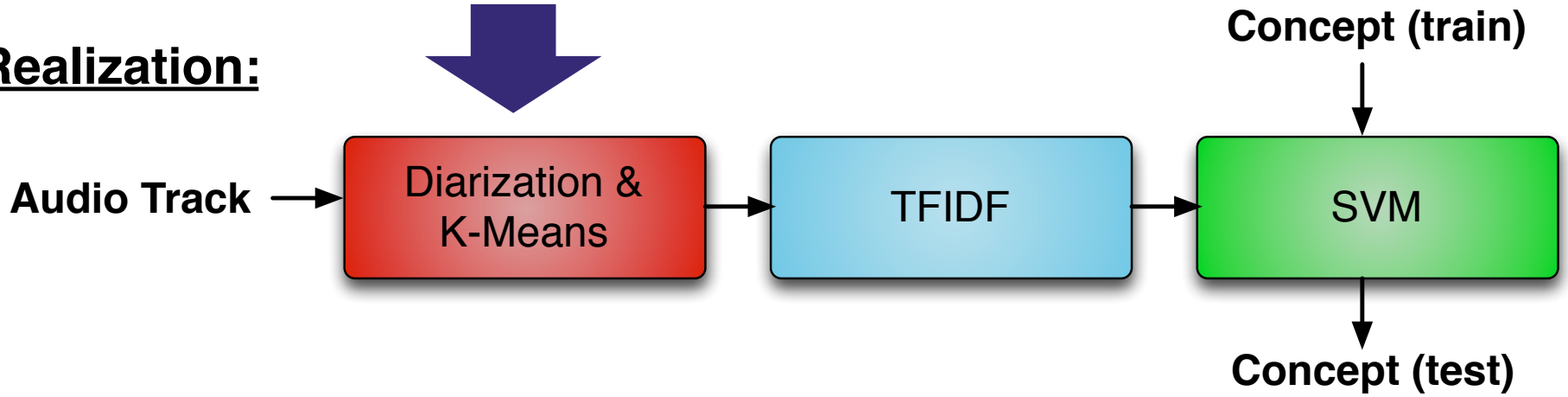
- Extract audio percepts.
- Determine which percepts are uncommon across concepts but common to the same concept.
- Detect video concepts by detecting common percepts.

Conceptual System Overview

Framework:



Realization:



Diarization or Percept Extraction

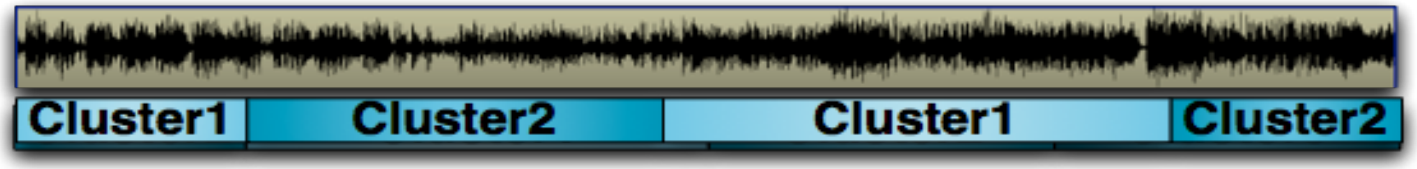
- Based on ICSI Speaker Diarization System…
  - Who spoke when?

- …but:
  - Speech/Speaker specific components removed.
  - Tuned for generic event diarization.

Xavier Anguera, et al. Speaker Diarization: A Review of Recent Research. In IEE transactions on audio, and speech processing.

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

Diarization or Percept Extraction



Diarization or Percept Extraction

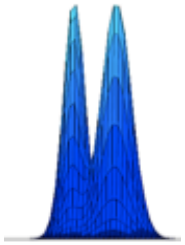
- Input:
  - Features: MFCC (19) + D + DD.
  - HTK format.
  - Window size 25ms every 10ms.
  - High number of initial segments.
  - Assumed 54 clusters/sounds.
  - Gaussian Mixture Models.

Diarization or Percept Extraction

- Output:
  - Segmentation file.

SPEAKER [label] [channel] [begin time] [length] < NA >< NA > [speaker ID] < NA > [word-based transcription]

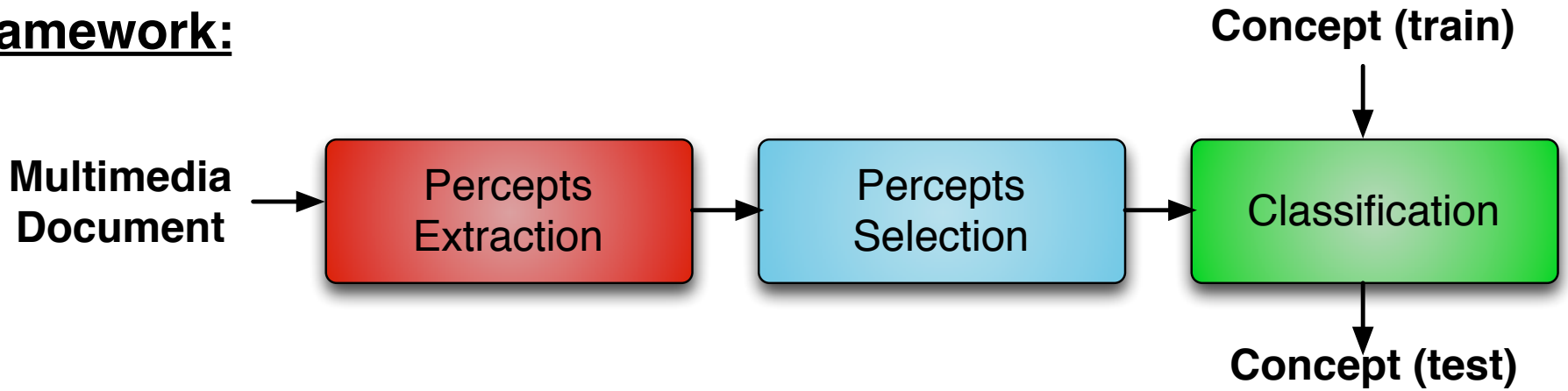
  - A GMM file with one GMM per segment.



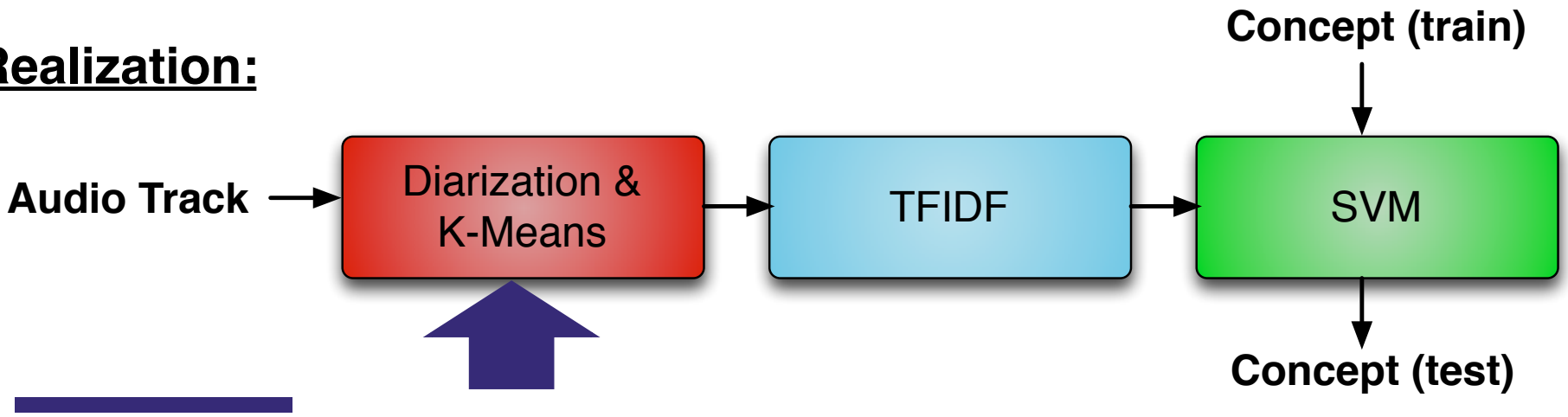
- Weight
- Variance
- Mean

Conceptual System Overview

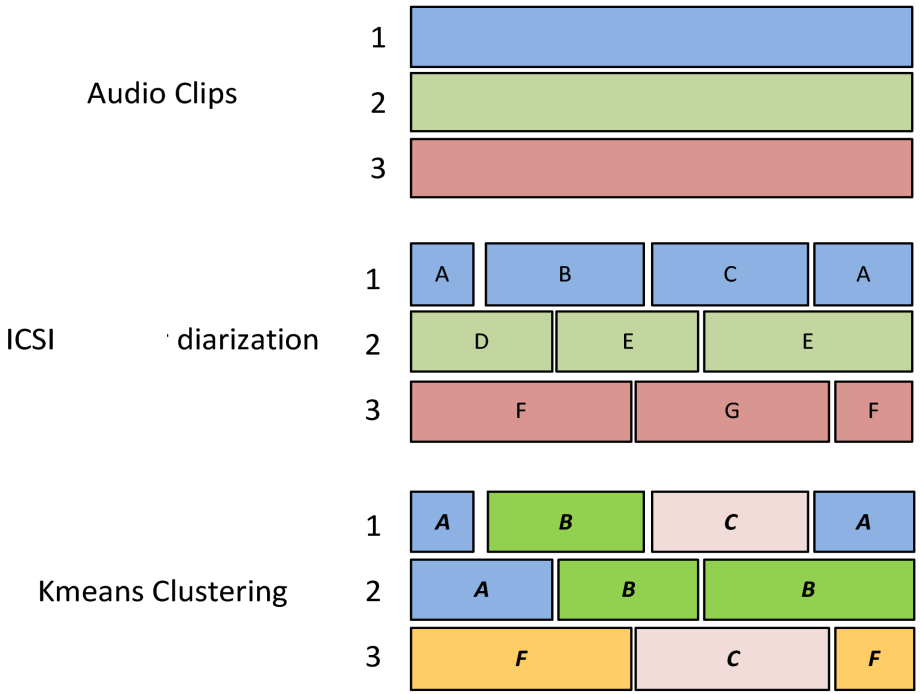
Framework:



Realization:

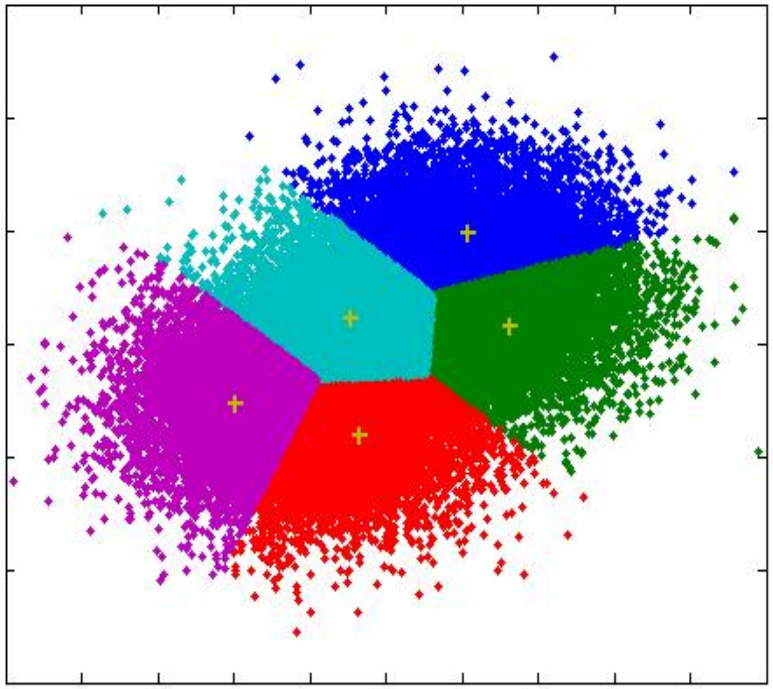


Percepts Dictionary



- Percepts extraction works on video-by-video basis.
- Use clustering to unify percepts across videos in one concept and build prototype percepts.

Percepts Dictionary/ Clustering

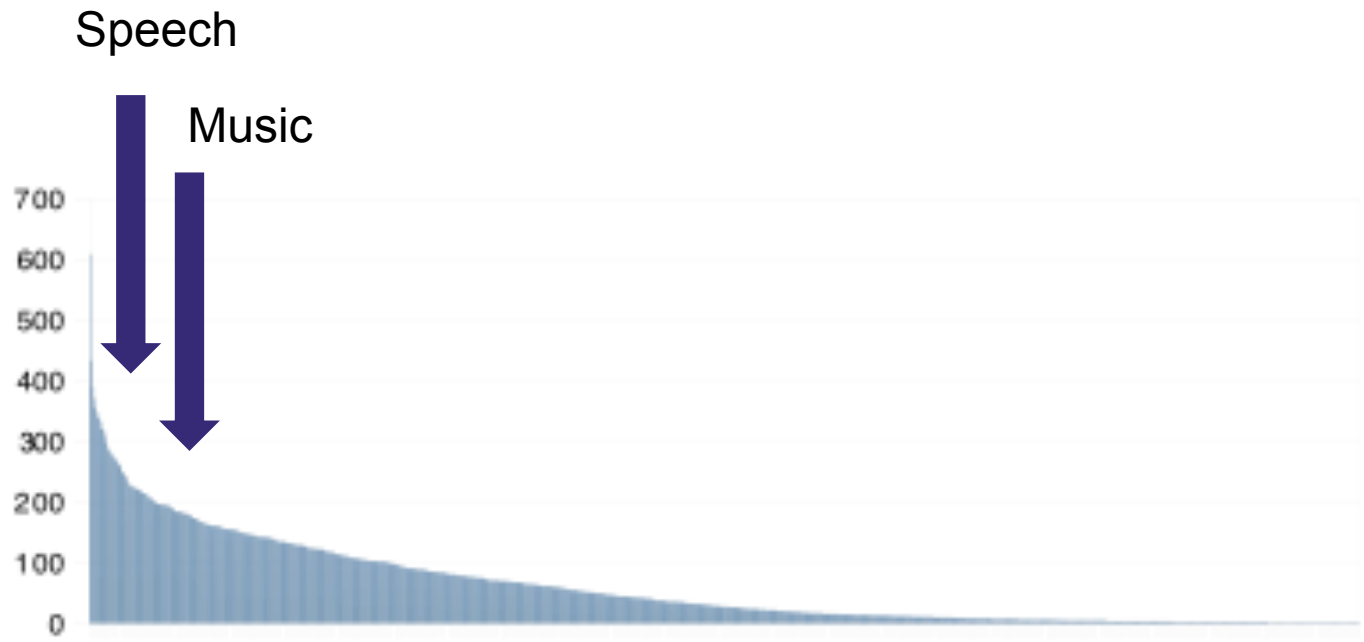


- 300 clusters
- Euclidean distance
- 10 iterations

- The GMMs' weight, mean and variance are combined to create simplified super vectors.
- Apply K-means to cluster the super vectors.
- Represent videos by (K) super vectors of prototype percepts = "words."

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

Distribution of "Words"



Histogram of top-300 "words".

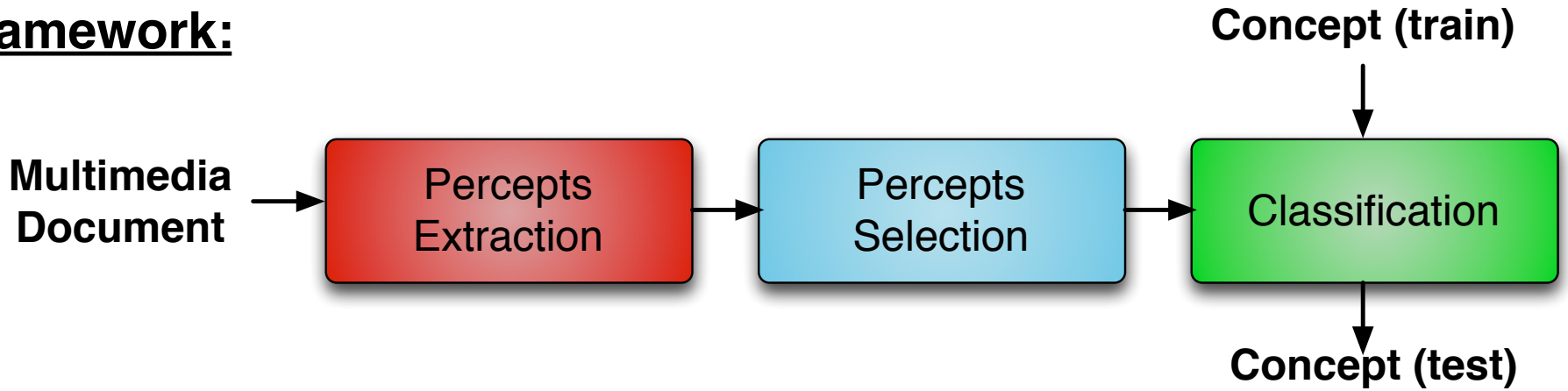
Near Zipfian Distribution!

Properties of "Words"

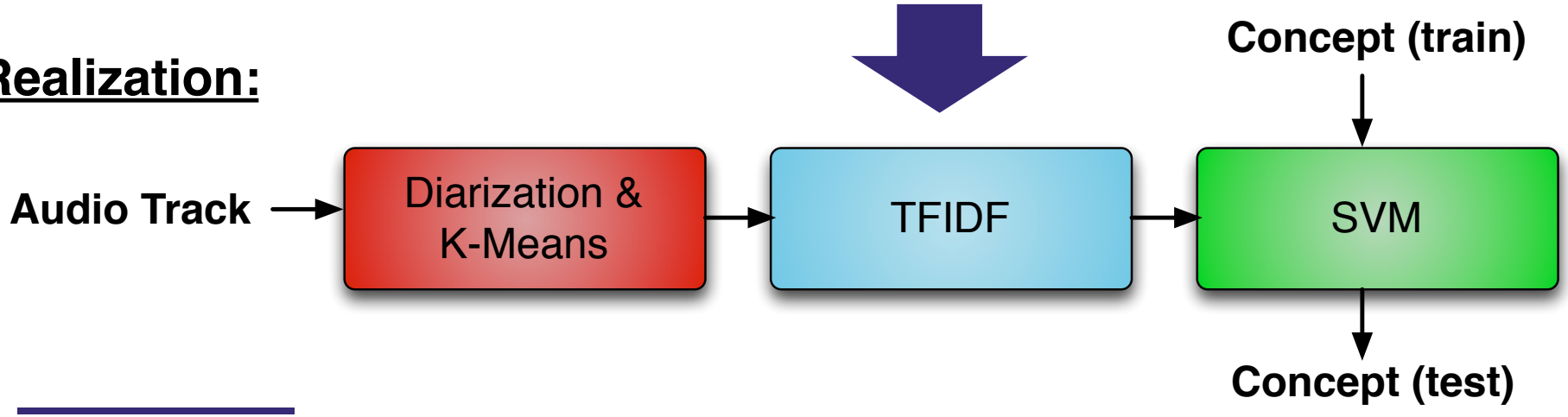
- Sometimes same "word" describes more percepts (homonym).
- Sometimes same percept is described by the different "words" (synonym).
- =>Problem?
  - (ant, aunt)
  - (smart, bright)

Conceptual System Overview

Framework:



Realization:



Term Frequency / Inverse Document Frequency

Rank	Word
1	the
2	be
3	to
4	of
5	and

Adjectives

1. good
2. new
3. first
4. last
5. long

- Reflects how important a word is to a document in a collection.
- Will the most common words in English help us finding a specific document?
  - Search for " the good fat grey kitty"

- Variations are often used by search engines for scoring and ranking documents.

Term Frequency / Inverse Document Frequency

- **$TF(c_i, D_k)$** is the frequency of "word" c_i in concept D_k
  - $P(c_i = c_j | c_j \in D_k)$ is the probability that "word" c_i equals c_j in concept D_k

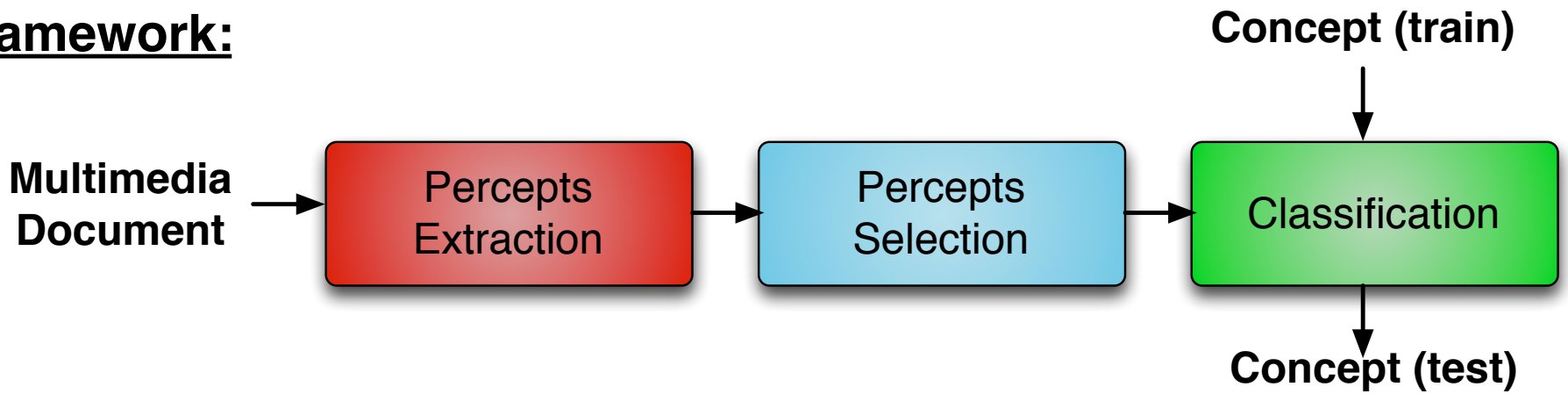
- **$IDF(c_i)$** and tells you whether a word is common or rare across the documents
  - $|D|$ is the total number of concepts
  - $P(c_i \in D_k)$ is the probability of "word" c_i in concept D_k

$$TF(c_i, D_k) = \frac{\Sigma_j n_j P(c_i = c_j \mid c_j \in D_k)}{\Sigma_j}$$

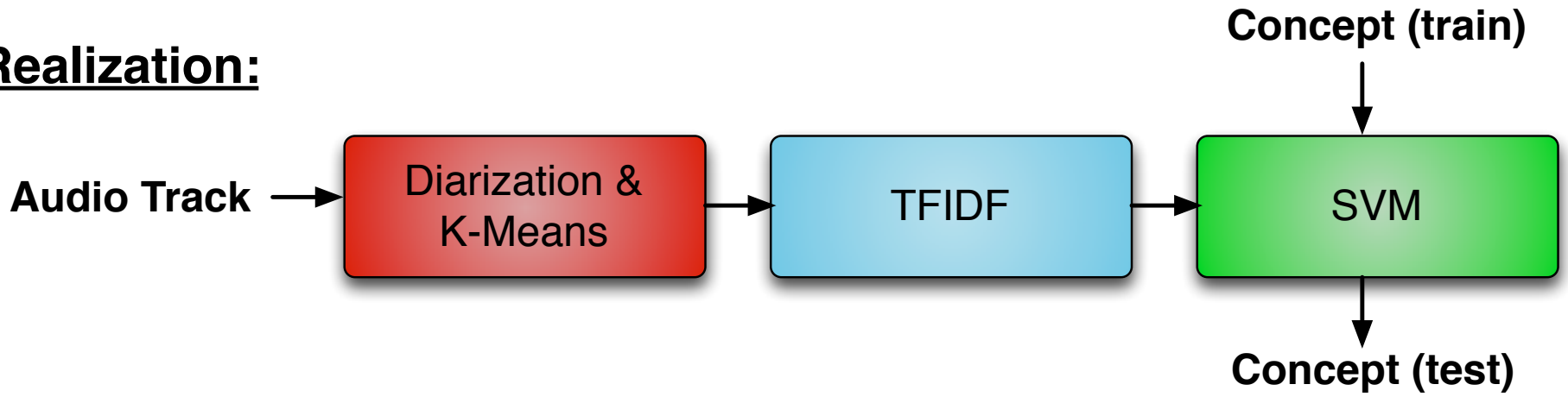
$$IDF(c_i) = \log \frac{|D|}{\Sigma_k P(c_i \in D_k)}$$

Conceptual System Overview

Framework:



Realization:

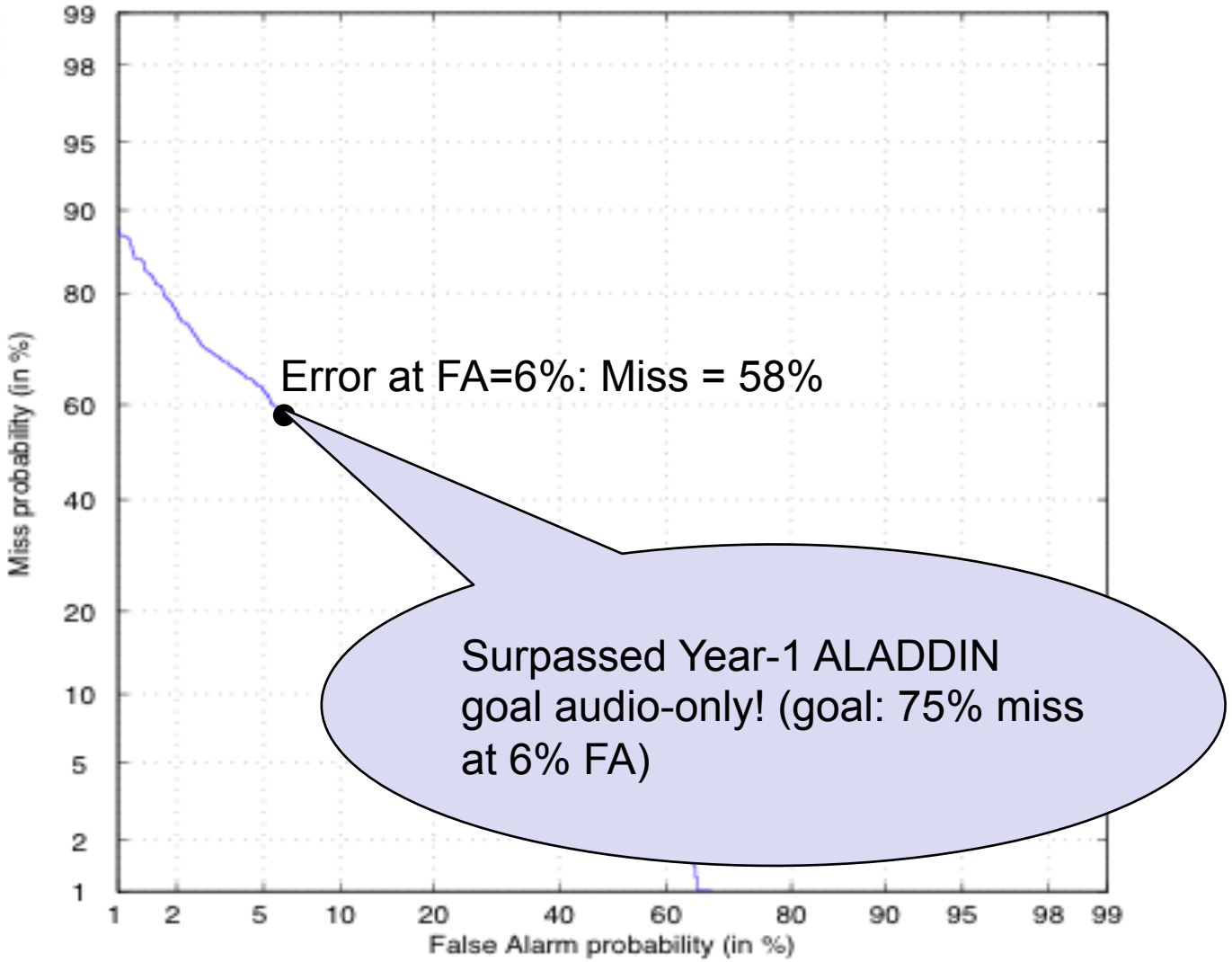


Support Vector Machine Classifier

- Input
  - A histogram per clip with TF/IDF weighted values.
  - Multiclass SVM
  - Intersection Kernel

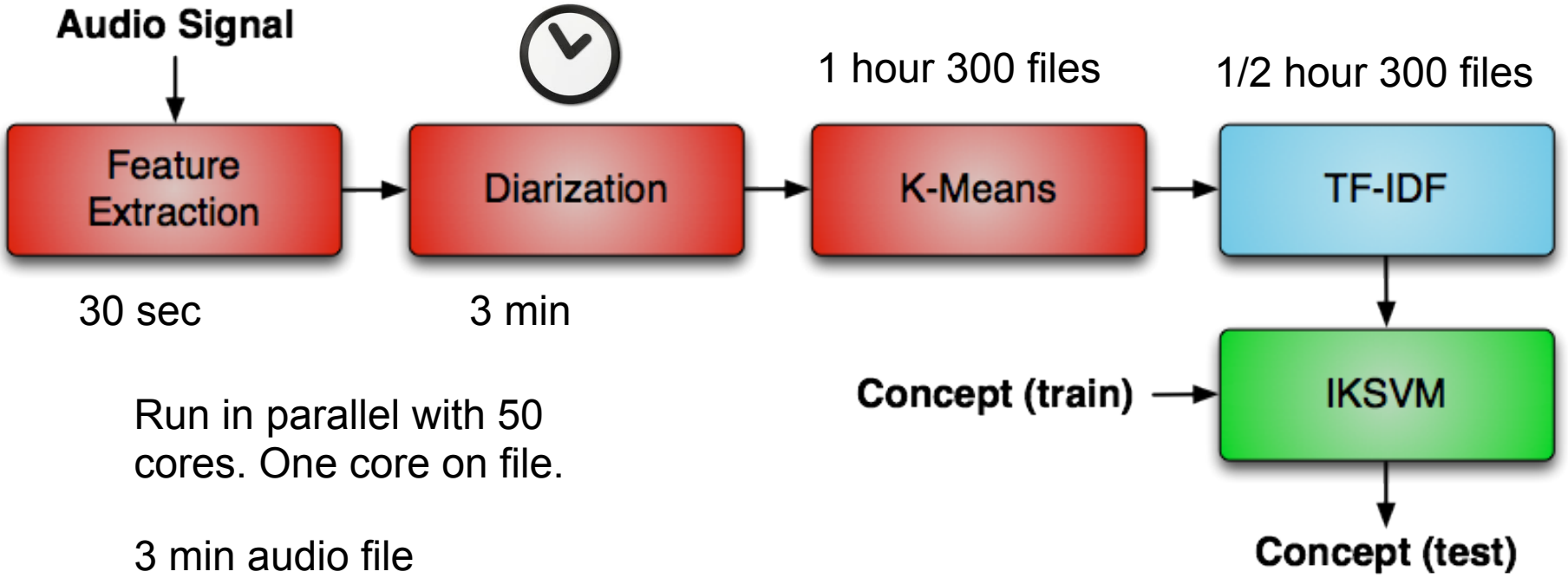
- Output
  - Score for each of the possible classes.

Audio-Only Detection on MED-DEV11



Processing Time

Database includes **150k** files with 3 min duration average.



Run in parallel with 50 cores. One core on file.

3 min audio file

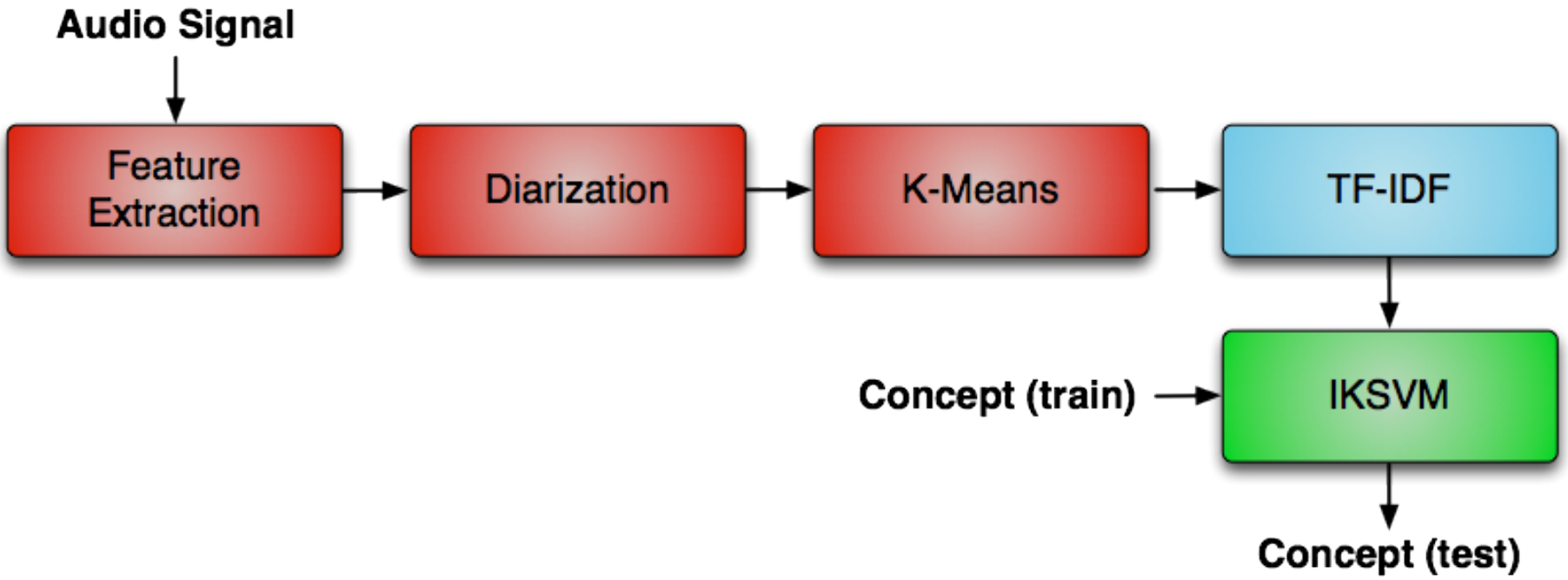
Diarization Time: 150,000*3/60/24/50 = 7 days

Conclusions

- 150k videos = no more looking at dataset...
- Teach computers to think, but not necessarily like a human.
- Event/Concept detection is still blooming.

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

What would you do?



Future Work

- Many knobs to tune.
- Reduce ambiguities in percepts extraction.
- Exploit temporal dimension better: (“sentences”, “paragraphs”?)
- Diarization using CUDA parallelization.

Thank You! Questions?

- Email: benmael@icsi.berkeley.edu

- Work together with: Gerald Friedland, Robert Mertens, Luke Gottlieb, and others.