

# Video Classification and Localization

**Danna Gurari**

The University of Texas at Austin

Fall 2019



# Review

- Last week:
  - Copy right & images
  - Image captioning applications
  - Image captioning evaluation
  - Crowdsourcing image captions
- Assignments (Class Website & Canvas)
  - Lab 2 assignment due yesterday
  - Project pre-proposal due yesterday
  - Project Proposal due week
- Questions?

# Today's Topics

- Video classification and localization applications
- Evaluating video classification and localization
- Crowdsourcing video classification and localization
- Lab: video annotation & writing papers in latex

# Today's Topics

- Video classification and localization applications
- Evaluating video classification and localization
- Crowdsourcing video classification and localization
- Lab: video annotation & writing papers in latex

# Definitions

- Video **Classification**: tag key topical themes/activity/etc for a video
- Activity **Localization**: localize sub-clip of a video where activity occurs

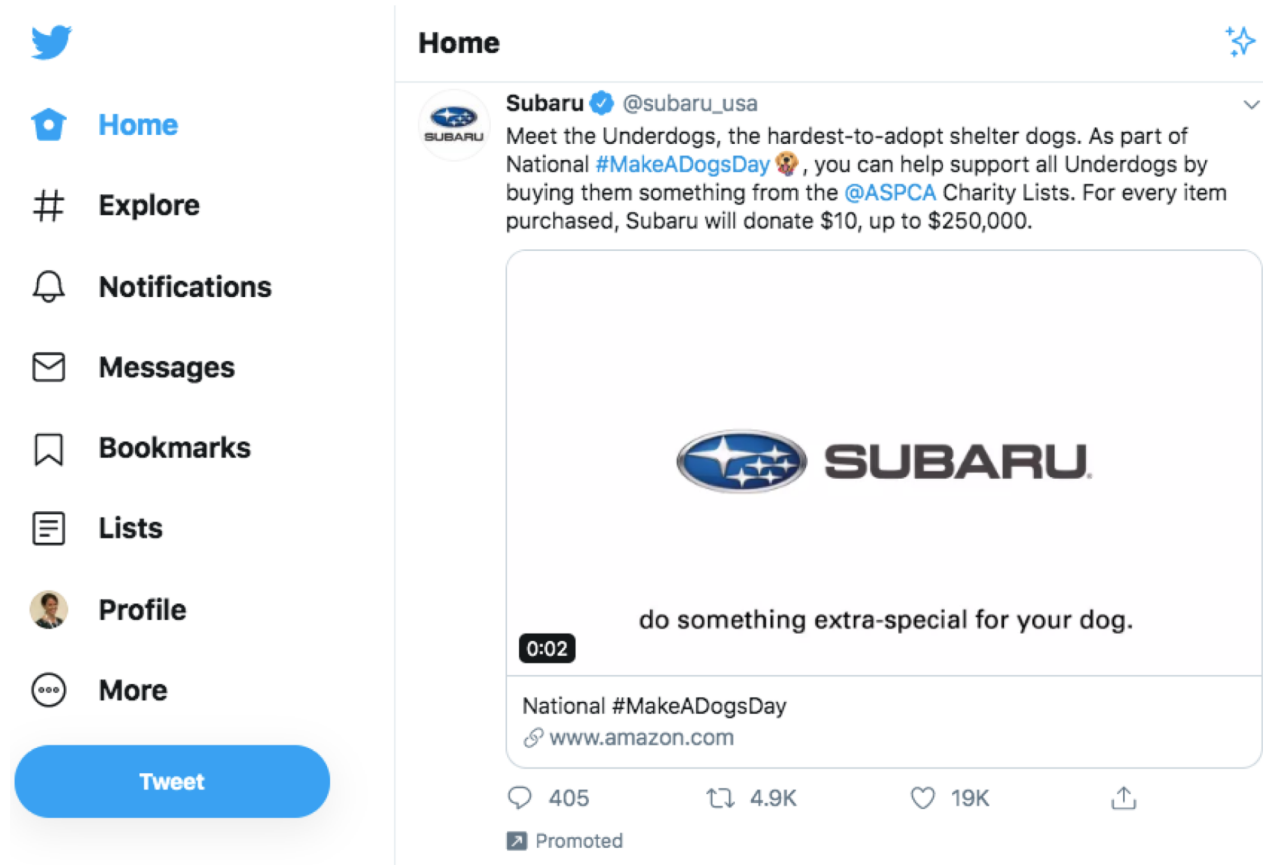
# Application: Video Search

The screenshot displays the YouTube homepage layout. At the top, there is a navigation bar with the YouTube logo, a search bar, and a 'SIGN IN' button. Below the navigation bar is a sidebar with menu items: Home, Trending, Subscriptions, Library, and History. The main content area features a large 'YouTubeTV' banner with a play button and the text 'Try it free'. To the right of the banner is a 'Watch the World Series' advertisement for YouTube TV with a 'TRY IT FREE' button. Below the banner is a 'Trending' section with five video thumbnails. Each thumbnail includes a video title, channel name, and view count.

Video Title	Channel	Views	Time
You Can't Con a Con Artist If You're Also a Con Artist - Ke...	Key & Peele	2M views	3:05
Patriots vs. Jets Week 7 Highlights   NFL 2019	NFL	1M views	8:10
TECHNIQUE CRITIQUE S1 • E14	WIRED	790K views	22:10
Unexpected Trick Shots   Dude Perfect	Dude Perfect	6M views	4:52
\$3.50 Soup Vs. \$29 Soup • Taiwan	BuzzFeedVideo	2.1M views	18:30

300 hours of video uploaded every minute (<https://merchdope.com/youtube-stats/>)

# Application: Social Media Recommendations



“An estimated 12 million micro-videos are posted to Twitter each day. The number of microvideos produced surpasses the total inventory of YouTube every 3 months”

- “The Open World of Micro-Videos; Nguyen et al.; [https://www.ics.uci.edu/~fowlkes/papers/nrfr\\_bigvision.pdf](https://www.ics.uci.edu/~fowlkes/papers/nrfr_bigvision.pdf)

# Application: Video Organization



Lists search results based on your collection of videos (spanning YouTube, news, movies, and more) in one list



# Application: Automatically Remove Objectionable Content



## Nudity or sexual content

YouTube is not for pornography or sexually explicit content. If this describes your video, even if it's a video of yourself, don't post it on YouTube. Also, be advised that we work closely with law enforcement and we report child exploitation. [Learn more](#)



## Harmful or dangerous content

Don't post videos that encourage others to do things that might cause them to get badly hurt, especially kids. Videos showing such harmful or dangerous acts may get age-restricted or removed depending on their severity. [Learn more](#)



## Hateful content

Our products are platforms for free expression. But we don't support content that promotes or condones violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, nationality, veteran status, or sexual orientation/gender identity, or whose primary purpose is inciting hatred on the basis of these core characteristics. This can be a delicate balancing act, but if the primary purpose is to attack a protected group, the content crosses the line. [Learn more](#)



## Violent or graphic content

It's not okay to post violent or gory content that's primarily intended to be shocking, sensational, or gratuitous. If posting graphic content in a news or documentary context, please be mindful to provide enough information to help people understand what's going on in the video. Don't encourage others to commit specific acts of violence. [Learn more](#)

And more listed here: <https://www.youtube.com/about/policies/#community-guidelines>

# Applications

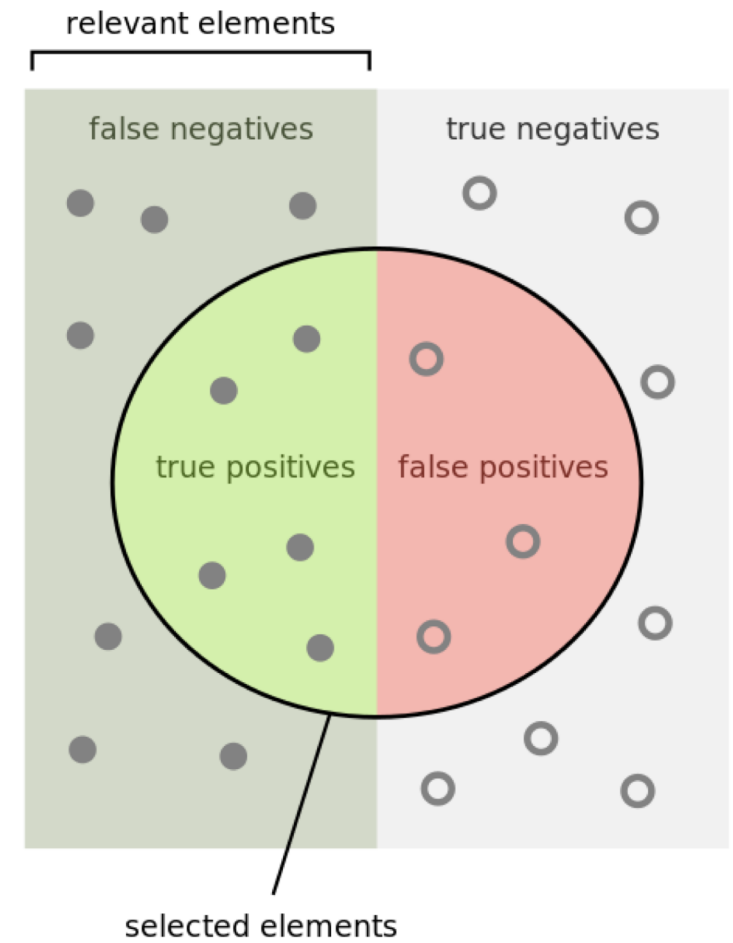
For what other applications might video classification and location be useful?

# Today's Topics

- Video classification and localization applications
- **Evaluating video classification and localization**
- Crowdsourcing video classification and localization
- Lab: video annotation & writing papers in latex

# Video Classification Evaluation

- Precision & Recall



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

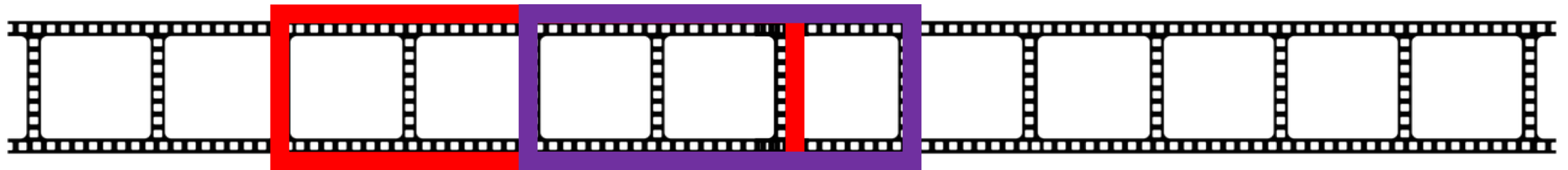
How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

[https://en.wikipedia.org/wiki/Precision\\_and\\_recall:](https://en.wikipedia.org/wiki/Precision_and_recall)

# Video Localization Evaluation

- Temporal intersection over union: checks if overlap of the **predicted frame selection** and **ground truth** exceeds a given threshold (e.g., 0.5)

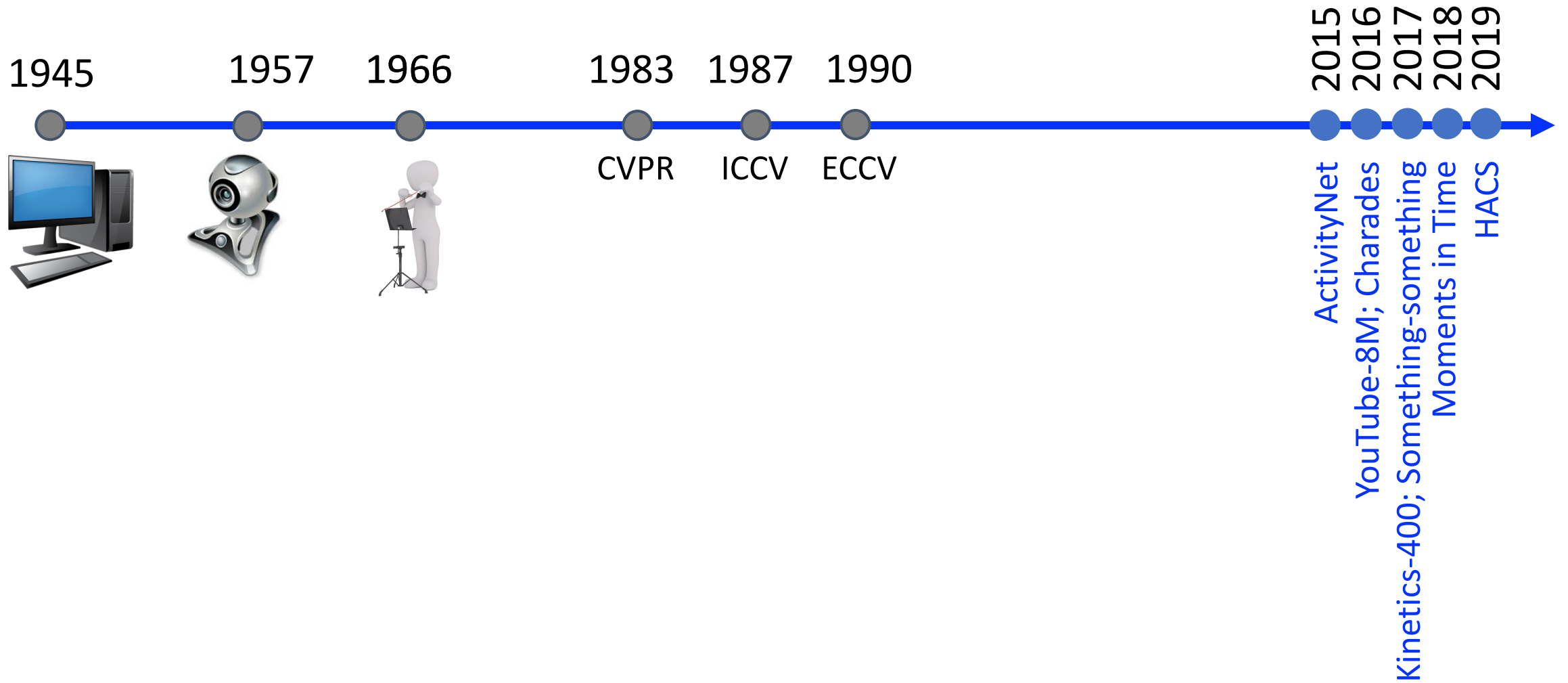


- What is the IoU score for this example?

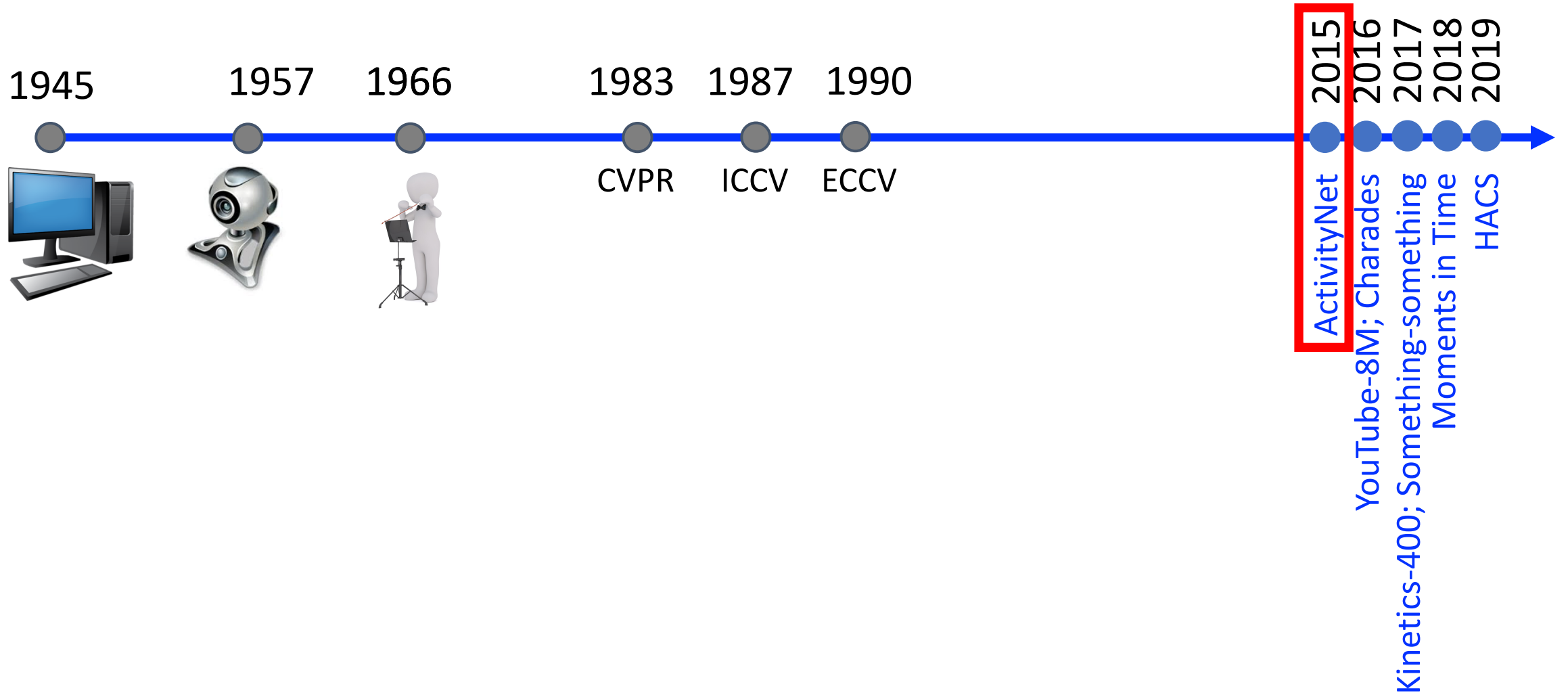
# Today's Topics

- Video classification and localization applications
- Evaluating video classification and localization
- **Crowdsourcing video classification and localization**
- Lab: video annotation & writing papers in latex

# Video Classification & Localization Datasets



# Video Classification & Localization Datasets





# ActivityNet

Focus on activities that humans spend most of their time doing in their lives

# ActivityNet

## 1. Category Selection

- \* American Time Use Survey (ATUS) created by the Department of Labor organizes activities according to:
  - social interactions
  - where activity usually occurs
- \* Authors selected 203 from the 2000+ activities in ATUS:
  - 7 top-level categories:  
*Personal Care, Eating and Drinking, Household, Working,...*
  - 4-level hierarchy



# ActivityNet

## 1. Category Selection

- \* American Time Use Survey (ATUS) created by the Department of Labor organizes activities according to:
  - social interactions
  - where activity usually occurs
- \* Authors selected 203 from the 2000+ activities in ATUS:
  - 7 top-level categories: *Personal Care, Eating and Drinking, Household, Working,...*
  - 4-level hierarchy

## 2. Video Collection



## 3. Video Verification

- \* “Expert” AMT workers verify presence of activity in each video
- \* Honey pot tasks introduced to assess trust of each crowd worker’s work

## 4. Activity Localization

- \* Multiple “expert” AMT workers annotate the start/end times for activity in each video
- \* A single ground truth is created by clustering the multiple annotations

# ActivityNet

## 1. Category Selection

- \* American Time Use Survey (ATUS) created by the Department of Labor organizes activities according to:
  - social interactions
  - where activity usually occurs
- \* Authors selected 203 from the 2000+ activities in ATUS:
  - 7 top-level categories: *Personal Care, Eating and Drinking, Household, Working,...*
  - 4-level hierarchy

## 2. Video Collection



## 3. Video Verification

- \* "Expert" AMT workers verify presence of activity in each video
- \* Honey pot tasks introduced to assess trust of each crowd worker's work

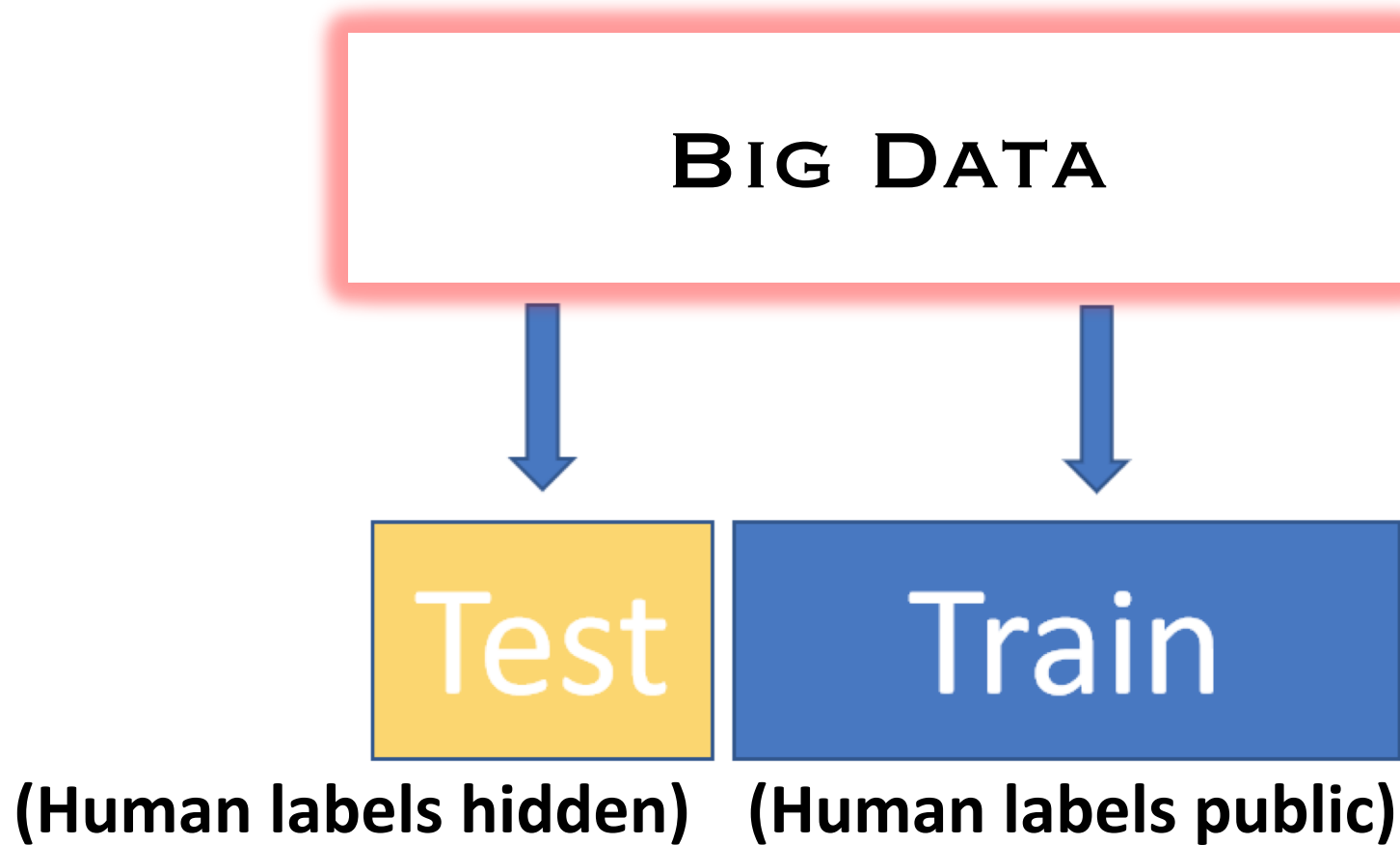
## 4. Activity Localization

- \* Multiple "expert" AMT workers annotate the start/end times for activity in each video
- \* A single ground truth is created by clustering the multiple annotations

# Activity Classification: ActivityNet Challenge



# Activity Classification: ActivityNet Challenge



**Winner: highest scoring method on the hidden test set**

# Activity Classification: ActivityNet Workshop



[HOME](#) [PEOPLE](#) [CHALLENGE](#) [PROGRAM](#) [DATES](#) [EVALUATION](#) [CONTACT](#)



## International Challenge on Activity Recognition (ActivityNet)

### Introduction

This challenge is the 4th annual installment of International Challenge on Activity Recognition, previously called the ActivityNet Large-Scale Activity Recognition Challenge which was first

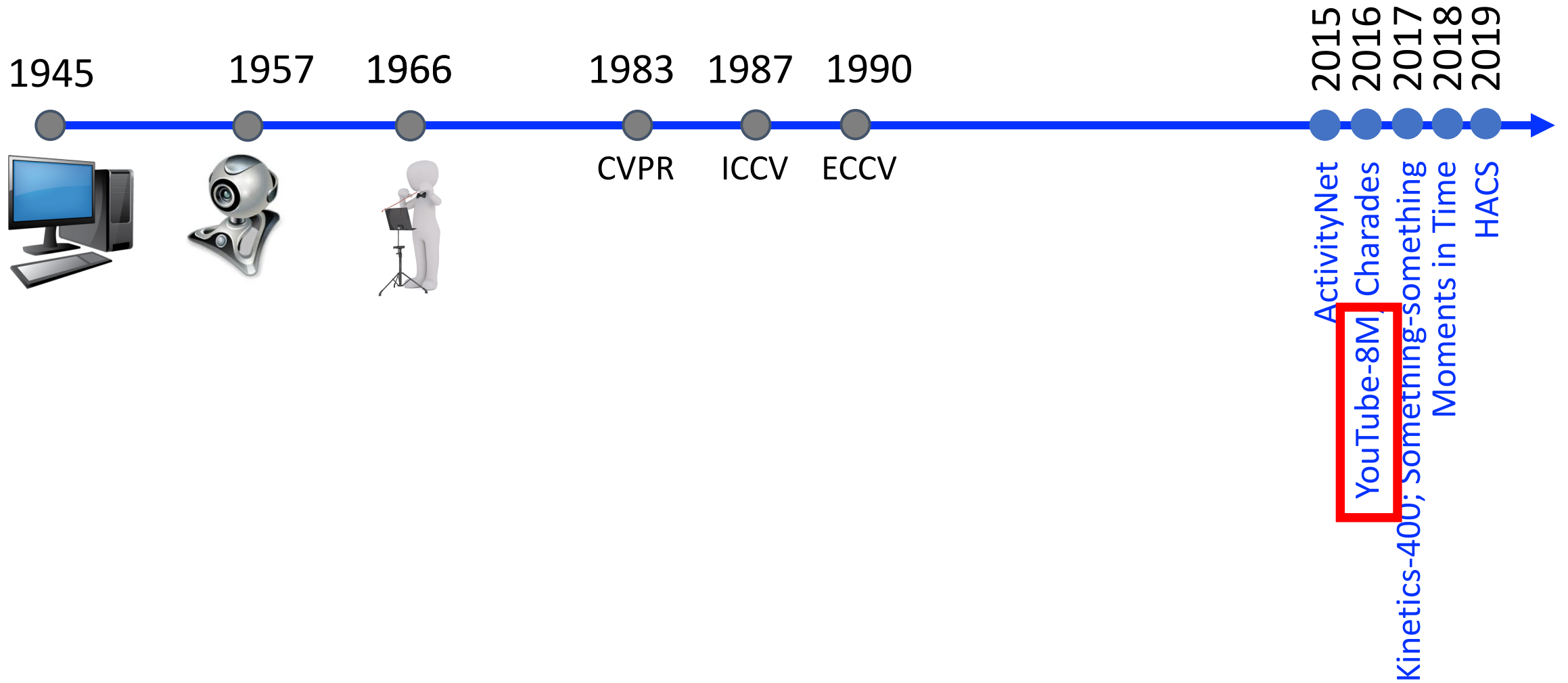
### News

10-June  
2019

If evaluation server is unresponsive, send your

<http://activity-net.org/challenges/2019/>

# Video Classification & Localization Datasets





# YouTube-8M

## 1. Category Selection

\* Began with 50,000 video topics from YouTube's "Knowledge Graph"

\* Reduced to ~10,000 topics that most of 3 human raters indicate is distinguishable by visual information alone

\* Kept YouTube categories that have 1,000+ views, > 120 secs, < 500 secs, and >= 200 videos

Entity Name	Entity URL	Entity Description
Thunderstorm	<a href="http://www.freebase.com/m/0jb2l">http://www.freebase.com/m/0jb2l</a>	A thunderstorm, also known as an electrical storm, a lightning storm, or a thundershower, is a type of storm characterized by the presence of lightning and its acoustic effect on the Earth's atmosphere known as thunder. The meteorologically assigned cloud type associated with the thunderstorm is the cumulonimbus. Thunderstorms are usually accompanied by strong winds, heavy rain and sometimes snow, sleet, hail, or no precipitation at all...

How difficult is it to identify this entity in images or videos (without audio, titles, comments, etc)?

- 1. Any layperson could
- 2. Any layperson after studying examples, wikipedia, etc could
- 3. Experts in some field can
- 4. Not possible without non-visual knowledge
- 5. Non-visual

# YouTube-8M

## 1. Category Selection

- \* Began with 50,000 video topics from YouTube's "Knowledge Graph"
- \* Reduced to ~10,000 topics that most of 3 human raters indicate is distinguishable by visual information alone
- \* Kept YouTube categories that have 1,000+ views, > 120 secs, < 500 secs, and  $\geq 200$  videos

## 2. Video Collection



# YouTube-8M Challenge & Annual Workshop

## Updated Dataset

YouTube-8M Segments was released in June 2019 with segment-level annotations. Human-verified labels on about 237K segments and 1000 classes are collected from the validation set of the YouTube-8M dataset. Each video will again come with time-localized features so classifier predictions can be made at segment-level granularity.

YouTube-8M was updated in May 2018 to include higher-quality, more topical annotations, and to clean up the annotation vocabulary. A number of low-frequency or low-quality labels and associated videos were removed, resulting in a smaller but higher-quality dataset (5.6M videos, 3862 classes). Additionally, the video IDs in the TensorFlow Record files have been anonymized, and the mapping to the real YouTube IDs will be periodically updated to exclude any videos that have been subsequently deleted (while preserving their anonymized features).

Dataset versions:

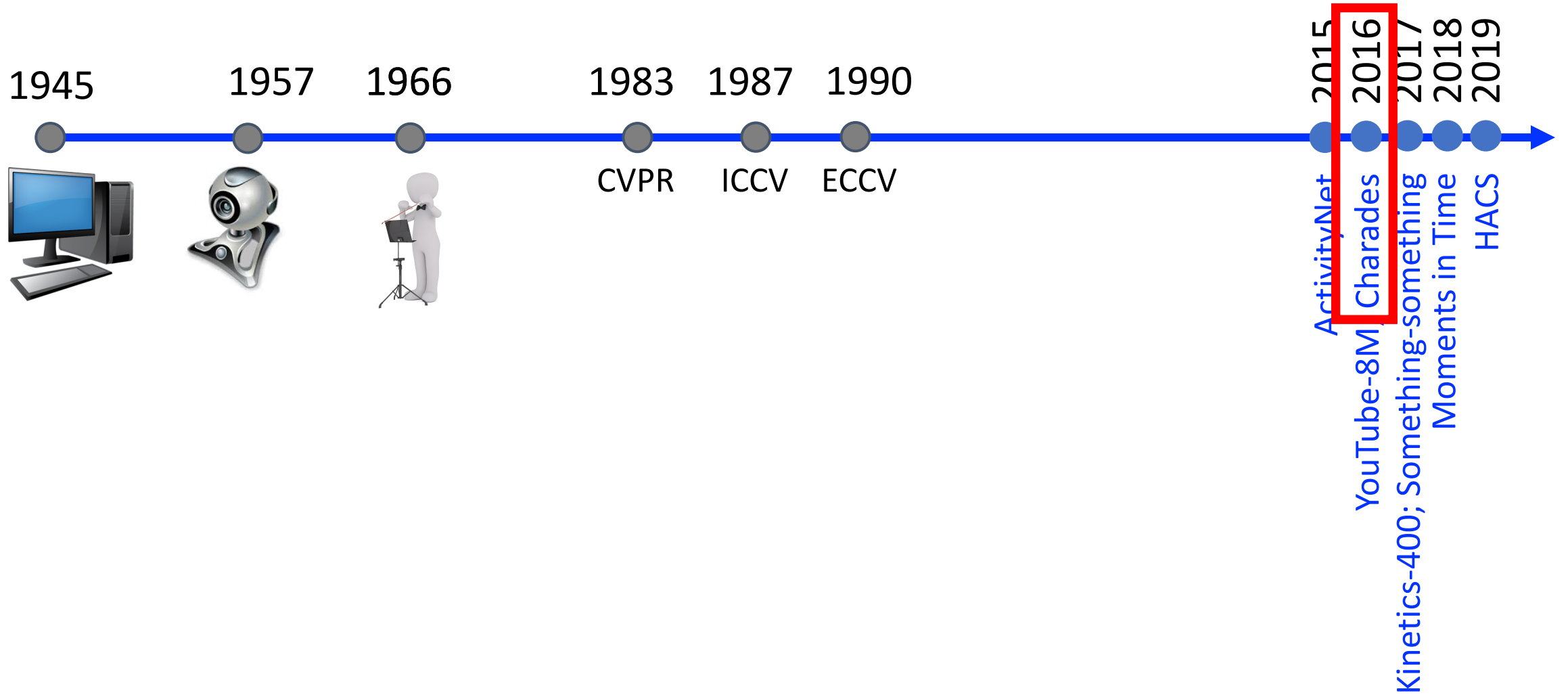
1. Jun 2019 version (current): 230K human-verified segment labels, 1000 classes, 5 segments/video
2. May 2018 version (current): 6.1M videos, 3862 classes, 3.0 labels/video, 2.6B audio-visual features
3. Feb 2017 version (deprecated): 7.0M videos, 4716 classes, 3.4 labels/video, 3.2B audio-visual features
4. Sep 2016 version (deprecated): 8.2M videos, 4800 classes, 1.8 labels/video, 1.9B visual-only features

2019

2018

2017

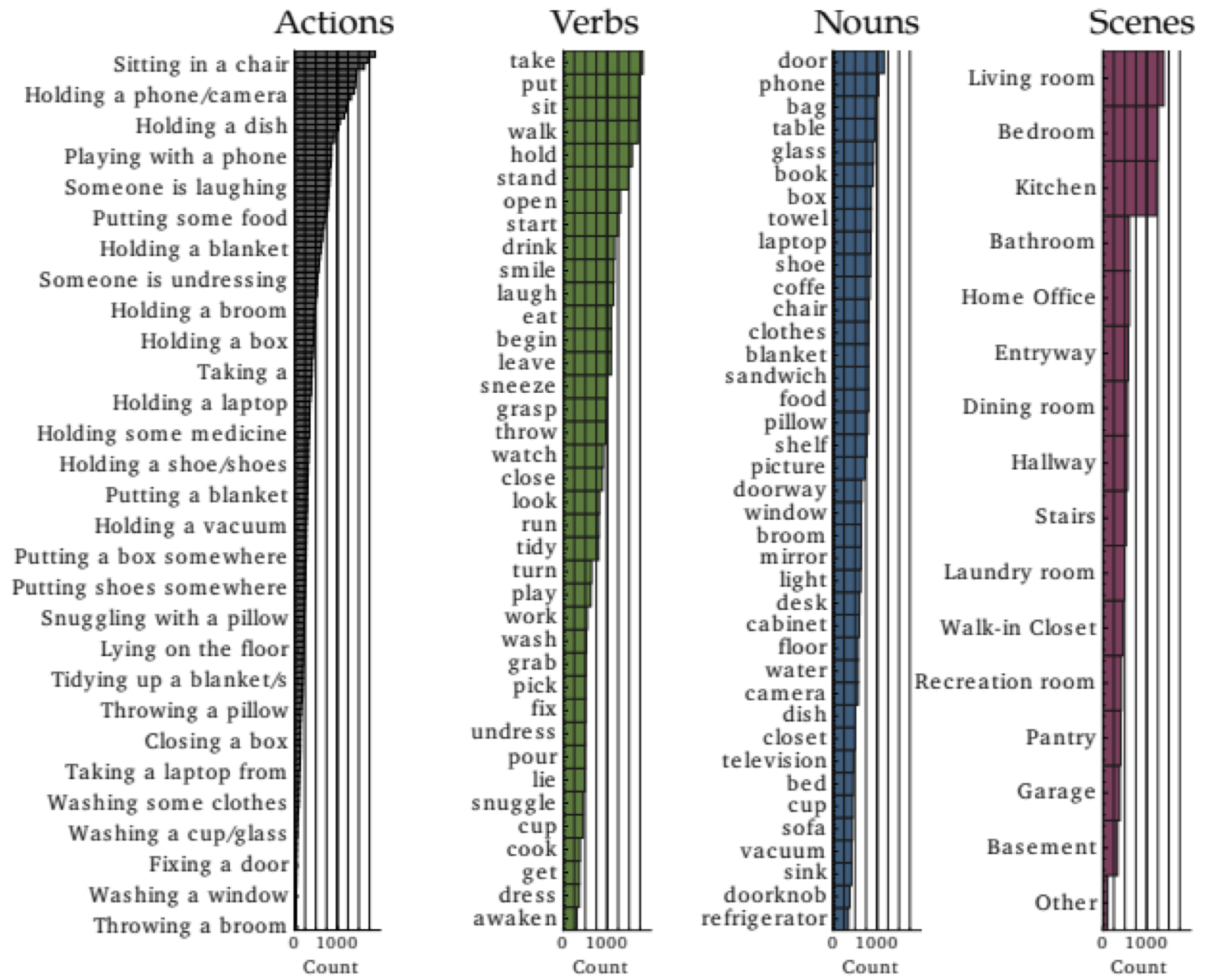
# Video Classification & Localization Datasets



# Charades

## 1. Video Script Generation

\* Authors identified 15 indoor scenes in residential homes (e.g., living room, home office)  
 \* Most common nouns and verbs in these scenes analyzed from 549 movie scripts resulting in 40 objects and 30 actions  
 \* Crowd workers generated scripts describing commonplace, realistic activities that involve 2 objects & 2 actions (given a scene, 5 objects, & 5 actions)



# Charades

## 1. Video Script Generation

- \* Authors identified 15 indoor scenes in residential homes (e.g., living room, home office)
- \* Most common nouns and verbs in these scenes analyzed from 549 movie scripts resulting in 40 objects and 30 actions
- \* Crowd workers generated scripts describing commonplace, realistic activities that involve 2 objects & 2 actions (given a scene, 5 objects, & 5 actions)

## 2. Video Collection

- \* Crowd workers recruited to record 30s videos of them executing the scripts

### Demo of videos

<https://www.youtube.com/watch?v=x9AhZLDkbyc>

# Charades

## 1. Video Script Generation

- \* Authors identified 15 indoor scenes in residential homes (e.g., living room, home office)
- \* Most common nouns and verbs in these scenes analyzed from 549 movie scripts resulting in 40 objects and 30 actions
- \* Crowd workers generated scripts describing commonplace, realistic activities that involve 2 objects & 2 actions (given a scene, 5 objects, & 5 actions)

## 2. Video Collection

- \* Crowd workers recruited to record 30s videos of them executing the scripts

## 3. Category Selection

- \* AMT workers recruited to watch each video and create a description
- \* Automatically identified “interacted objects” mentioned both in script & description
- \* 150 actions chosen following crowd worker verification

## 4. Activity Localization

- \* Crowd workers then annotated the start/end times for each activity in each video

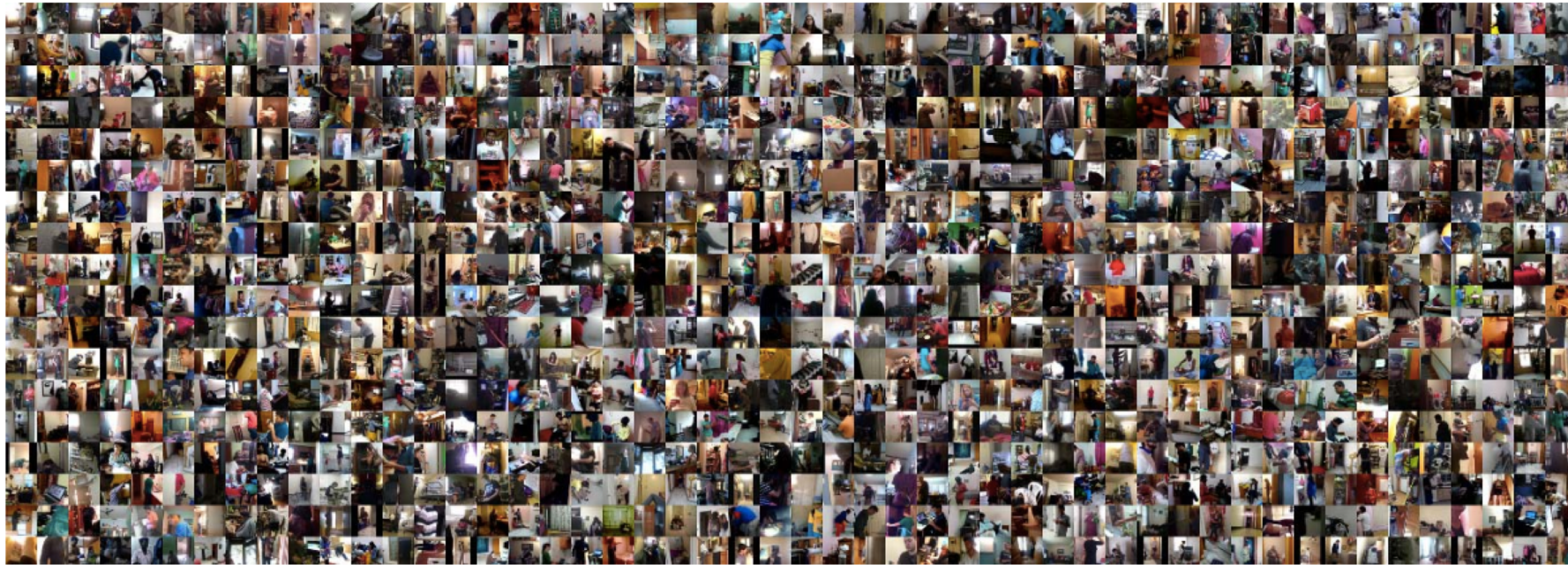
# Charades Challenge & Annual Workshop

CVPR 2017 Workshop on Visual Understanding Across Modalities

Home THOR Charades TQA

## Charades Challenge

Recognize and locate activities taking place in a video

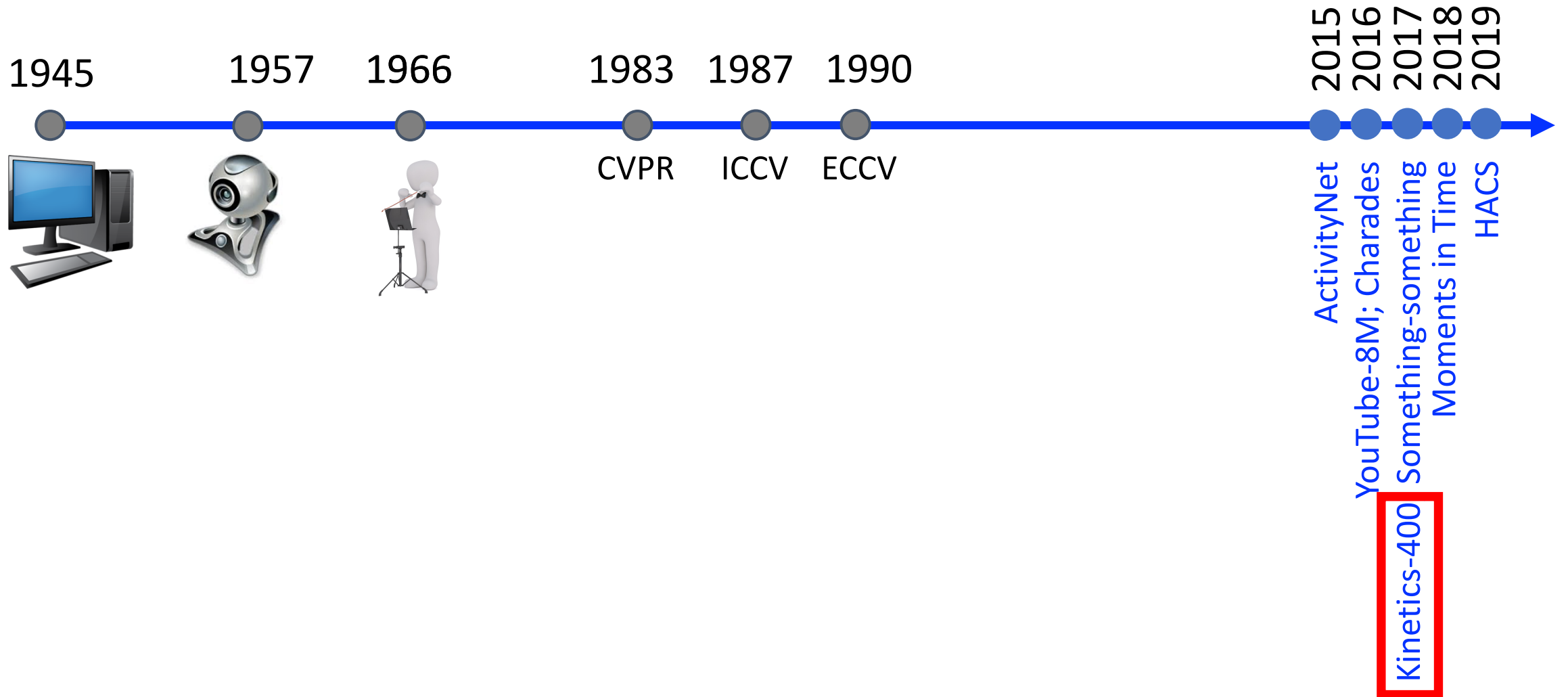


The Charades Activity Challenge aims towards automatic understanding of daily activities, by providing realistic videos of people doing everyday activities. [The Charades dataset](http://vuchallenge.org/charades.html) is collected for an unique insight into daily tasks such as drinking coffee, putting on shoes while sitting in a chair, or snuggling with a blanket on the couch while watching something

<http://vuchallenge.org/charades.html>



# Video Classification & Localization Datasets



# Kinetics-400

## 1. Category Selection

\* Categories taken from existing action datasets, motion capture datasets, and AMT workers feedback about more suitable categories

## 2. Video Collection



\* Action recognition classifiers applied to identify relevant clips (5 seconds before plus 5 seconds after image where activity is recognized)


## 3. Video Verification

\* AMT worker verifies action is present in the video for 20 videos per HIT  
\* Honeytrap videos used to prompt a warning to workers when their accuracy falls below 50%  
\* Label determined by majority of 5 workers

# Kinetics-400

## Evaluating Actions in Videos



Can you see a  human performing the action  
**riding mule?**



### Instructions

We would like to find videos that contain real humans performing actions e.g. scrubbing their face, jumping, kissing someone etc.

Please click on the most appropriate button after watching each video:



Yes, this contains a true example of the action



No, this does not contain an example of the action



You are unsure if there is an example of the action



Replay the video



Video does not play, does not contain a human, is an image, cartoon or a computer game.



We have turned off the audio, you need to judge the clip using the visuals only.

# Kinetics Challenge & Annual Workshop

## Task A – Trimmed Action Recognition

Challenge 2019 → Task A – Trimmed Action Recognition

The goal of the Kinetics dataset is to help the computer vision and machine learning communities advance models for video understanding. Given this large human action classification dataset, it may be possible to learn powerful video representations that transfer to different video tasks.

For information related to this task, please contact: [enoland@google.com](mailto:enoland@google.com), [joaluis@google.com](mailto:joaluis@google.com)

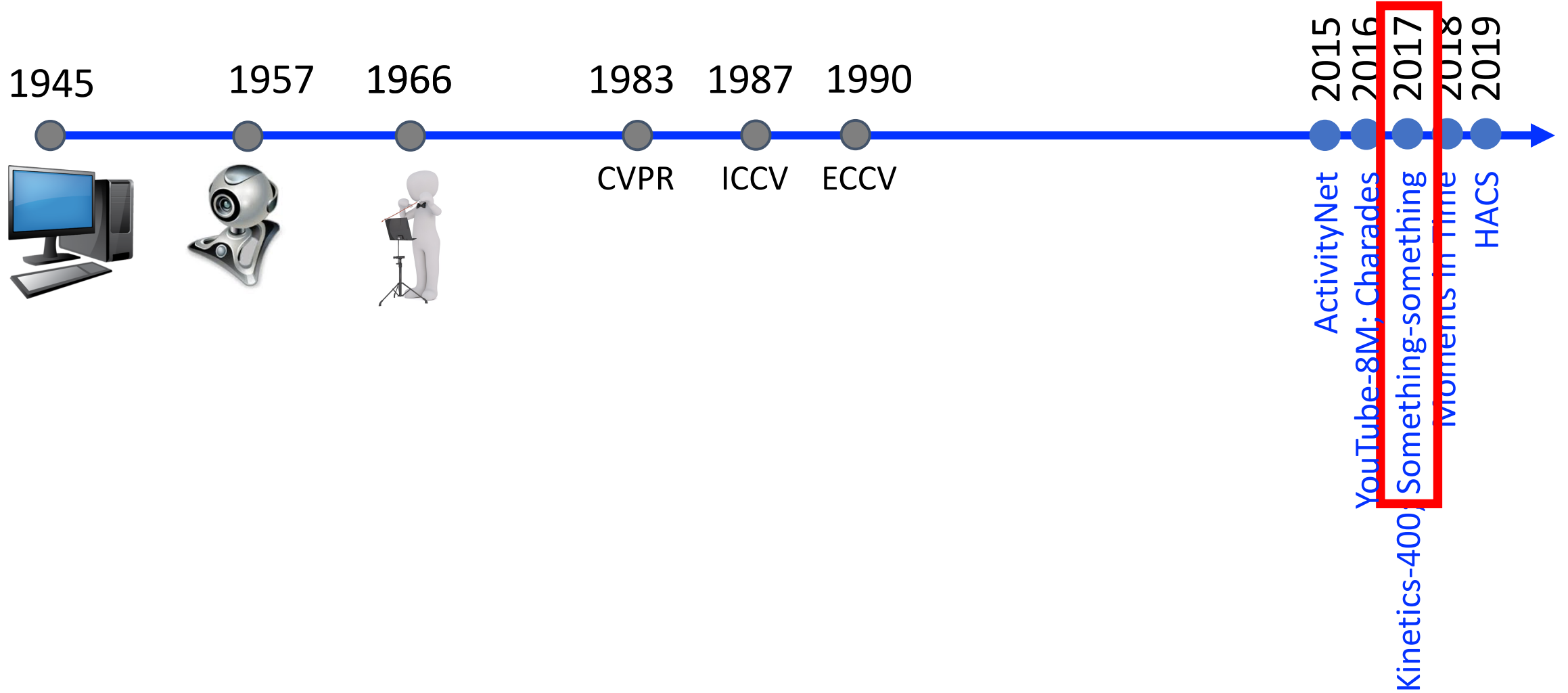
### Dataset

The **Kinetics-700 dataset** will be used for this challenge. Kinetics-700 is a large-scale, high-quality dataset of YouTube video URLs which include a diverse range of human focused actions. Our aim in releasing the Kinetics dataset is to help the machine learning community to advance models for video understanding. It is an approximate super-set of both Kinetics-400, released in 2017 and Kinetics-600, released in 2018.

The dataset consists of approximately 650,000 video clips, and covers 700 human action classes with at least 600 video clips for each action class. Each clip lasts around 10 seconds and is labeled with a single class. All of the clips have been through multiple rounds of human annotation, and each is taken from a unique YouTube video. The actions cover a broad range of classes including human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands and hugging.

[http://activity-net.org/challenges/2019/tasks/guest\\_kinetics.html](http://activity-net.org/challenges/2019/tasks/guest_kinetics.html)

# Video Classification & Localization Datasets



# Something-something

## 1. Category Selection

\* Authors created 175 something-something templates

e.g.,

<b>10 selected classes</b>
Dropping [something]
Moving [something] from right to left
Moving [something] from left to right
Picking [something] up
Putting [something]
Poking [something]
Tearing [something]
Pouring [something]
Holding [something]
Showing [something] (almost no hand)

# Something-something

## 1. Category Selection

\* Authors created 175 something-something templates →

## 2. Video Collection

\* Crowd workers submit videos of them recording an implementation of the template

UI

You have selected 5 of 10 descriptions

- ▶ Folding something (2)
- ▶ Stuffing/Taking out (1)
- ▼ Holding something (5)
  - Holding [something]
  - Holding [something] over [something]
  - Holding [something] next to [something]
  - Holding [something] in front of [something]
  - Holding [something] behind [something]
- ▶ Crowd of things (2)
- ▶ Shadows (1)
- ▶ Collisions of objects (3)
- ▼ Tearing something (3)
  - Tearing [something] into two pieces
  - Pretending to be tearing [something that is not tearable]
  - Tearing [something] just a little bit
- ▶ Lifting/Tilting objects with other objects on them (3)
- ▼ Moving two objects relative to each other (4)
  - Moving [something] closer to [something]
  - Moving [something] away from [something]
  - Moving [something] and [something] closer to each other (fix the camera and use both hands to move both objects)
  - Moving [something] and [something] away from each other (fix the camera and use both hands to move both objects)
- ▶ Attaching/Trying to attach (2)
- ▶ Spinning something (3)
- ▶ Something falling (2)
- ▶ Putting/Taking objects into/out of/next to/... other objects (19)
- ▼ Rolling and sliding something (10)
  - Letting [something] roll down a slanted surface
  - Letting [something] roll up a slanted surface, so it rolls back down
  - Letting [something] roll along a flat surface
  - Putting [something] on a flat surface without letting it roll
  - Putting [something] that can't roll onto a slanted surface, so it stays where it is
  - Putting [something] that can't roll onto a slanted surface, so it slides down
  - Lifting a surface with [something] on it until it starts sliding down
  - Lifting a surface with [something] on it but not enough for it to slide down
  - Putting [something] onto a slanted surface but it doesn't glide down
  - Rolling [something] on a flat surface
- ▶ Plugging something into something (2)

You have uploaded 0 of 10 videos

Submit Task

holding something over something

moving something away from something

lifting a surface with something on it but not enough for it to slide down

moving something away from something

tearing something into two pieces



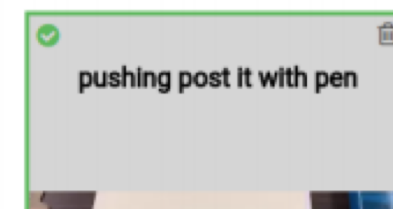
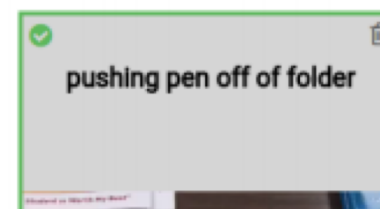
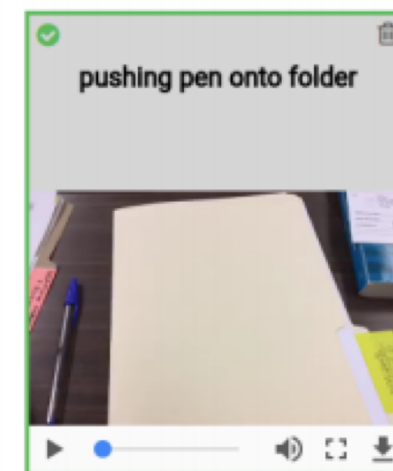
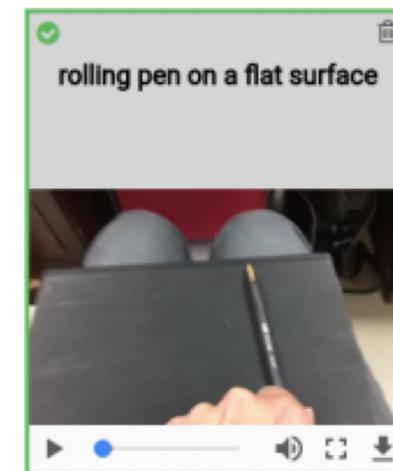
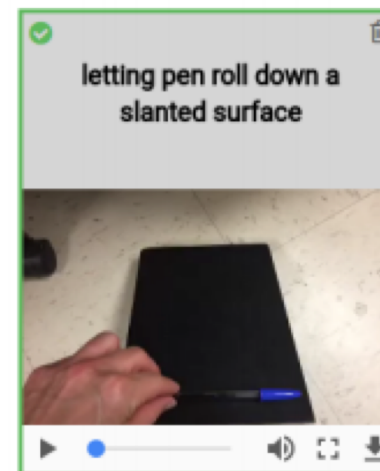
UI

You have selected 10 of 10 descriptions

- › Folding something (2)
- › Stuffing/Taking out (1)
- › Holding something (5)
- › Crowd of things (2)
- › Shadows (1)
- › Collisions of objects (3)
- › Tearing something (3)
- › Lifting/Tilting objects with other objects on them (3)
- › Moving two objects relative to each other (4)
- › Attaching/Trying to attach (2)
- › Spinning something (3)
- › Something falling (2)
- › Putting/Taking objects into/out of/next to/... other objects (19)
- › Rolling and sliding something (10)
- › Plugging something into something (2)
- › Twisting something (3)
- › Opening or closing something (4)
- › Pushing something (9)
- › Tipping something over (2)
- › Filming objects, without any actions (3)
- › Spilling something (3)
- › Turning something upside down (2)
- › Putting something somewhere (2)
- › Picking something up (2)
- › Hitting something with something (1)
- › Dropping something (5)
- › Poking something (9)
- › Throwing something (6)
- › Wiping something off of something (2)
- › Camera motions (8)
- › Showing objects and photos of objects (2)
- › Squeezing something (2)
- › Revolving something (3)

You have uploaded 10 of 10 videos

Submit Task



# Something-something Challenge

## The 20BN-something-something Dataset V2

### Introduction

The 20BN-SOMETHING-SOMETHING dataset is a large collection of densely-labeled video clips that show **humans performing pre-defined basic actions with everyday objects**. The dataset was created by a large number of crowd workers. It allows machine learning models to develop fine-grained understanding of basic actions that occur in the physical world. It is **available free of charge for academic research**. Commercial licenses are available upon request.

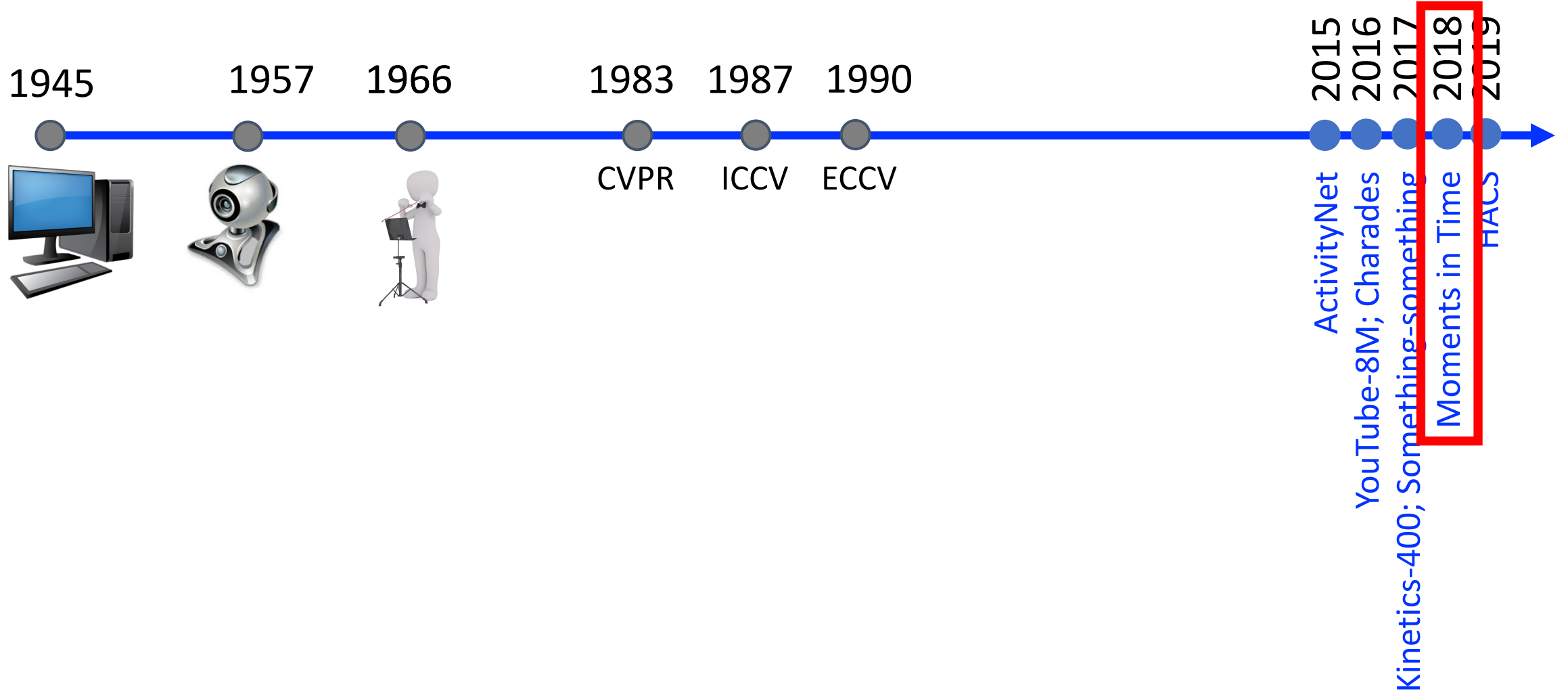
This is the second release of the dataset. The first release is also still available [here](#). The new release features the following updates:

- **Greatly increased number of videos:** With 220,847 videos (vs. 108,499 in V1) we release more than twice as many videos.
- **Object annotations and captioning:** For each video in the training and validation sets we now also provide object annotations in addition to the video label if applicable. For example, for a label like "Putting



<https://20bn.com/datasets/something-something>

# Video Classification & Localization Datasets



# Moments in Time

## 1. Category Selection

- \* Initial list of 4,500 most common verbs listed in VerbNet
- \* Automated clustering of verbs to distill them to 339 verbs that are both visual and unique

## 2. Video Collection


- \* Candidate videos collected from Youtube, Flickr, Vine, Metacafe, Peeks, Vimeo, VideoBlocks, Bing, Giphy, The Weather Channel, and Getty-Images for entire vocabulary
- \* Random 3 second clip used per video

## 3. Video Verification

- \* AMT worker verifies action is present in the video for 74 videos per HIT
- \* First 4 videos used solely for training
- \* 10 videos used for honeypot testing, with results used when 90%+ accuracy
- \* Use majority label from 3+ workers

Instructions   Action Definition   Submit (39 actions left)

In the following video

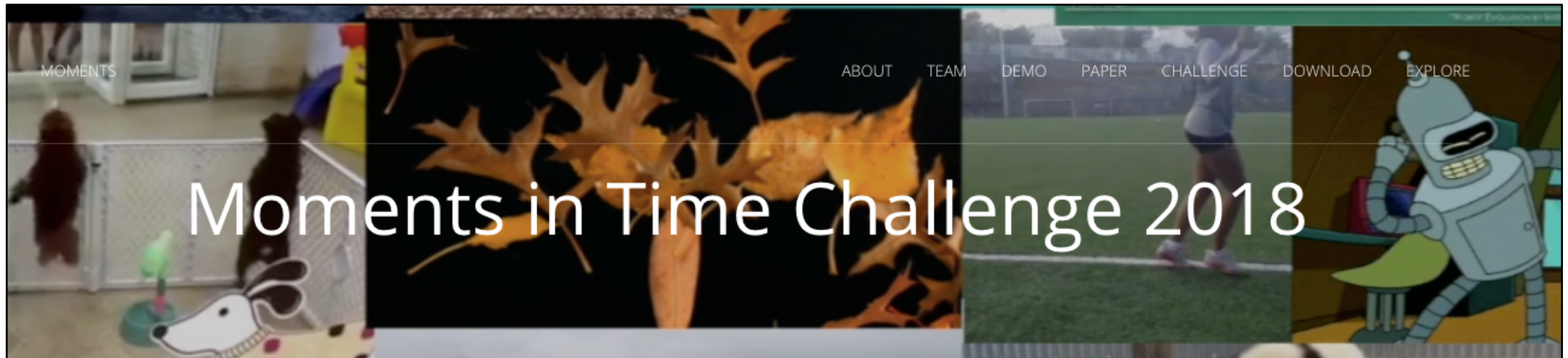


do you see or hear the action  
**cooking**  
(Prepare a hot meal.)  
happening within this video?

[Press 1]                      [Press 2]  
YES                                      NO

35/73

# Moments in Time Challenge & Workshop



The Moments in Time Recognition Challenge at [CVPR'18](#) was jointly held with the [ActivityNet Challenge 2018](#). The goal of this challenge was to identify the event labels depicted in a 3 second video. The video data came from the Moments in Time dataset, which could be downloaded [here](#).

The challenge had two tracks:



## FULL TRACK

A classification task on the **entire** Moments in Time dataset:

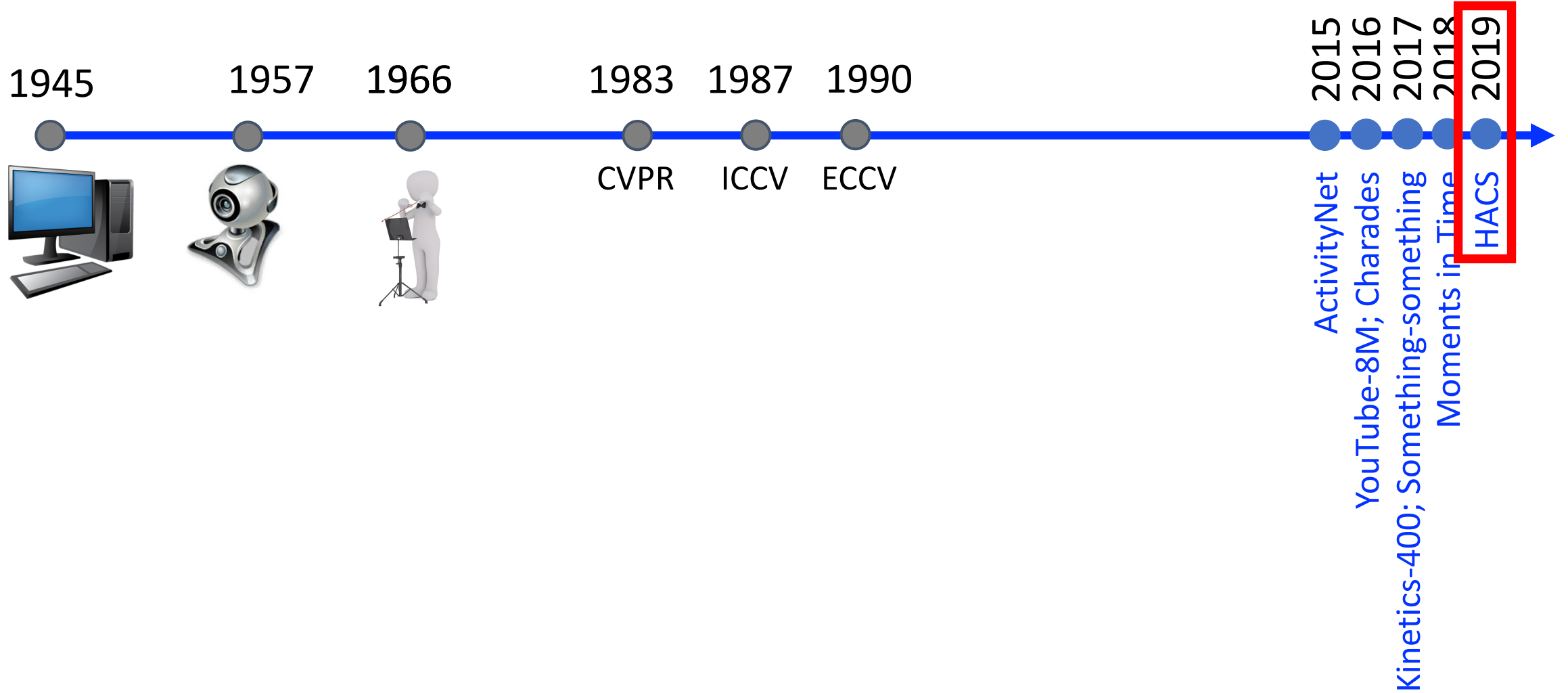


## MINI TRACK

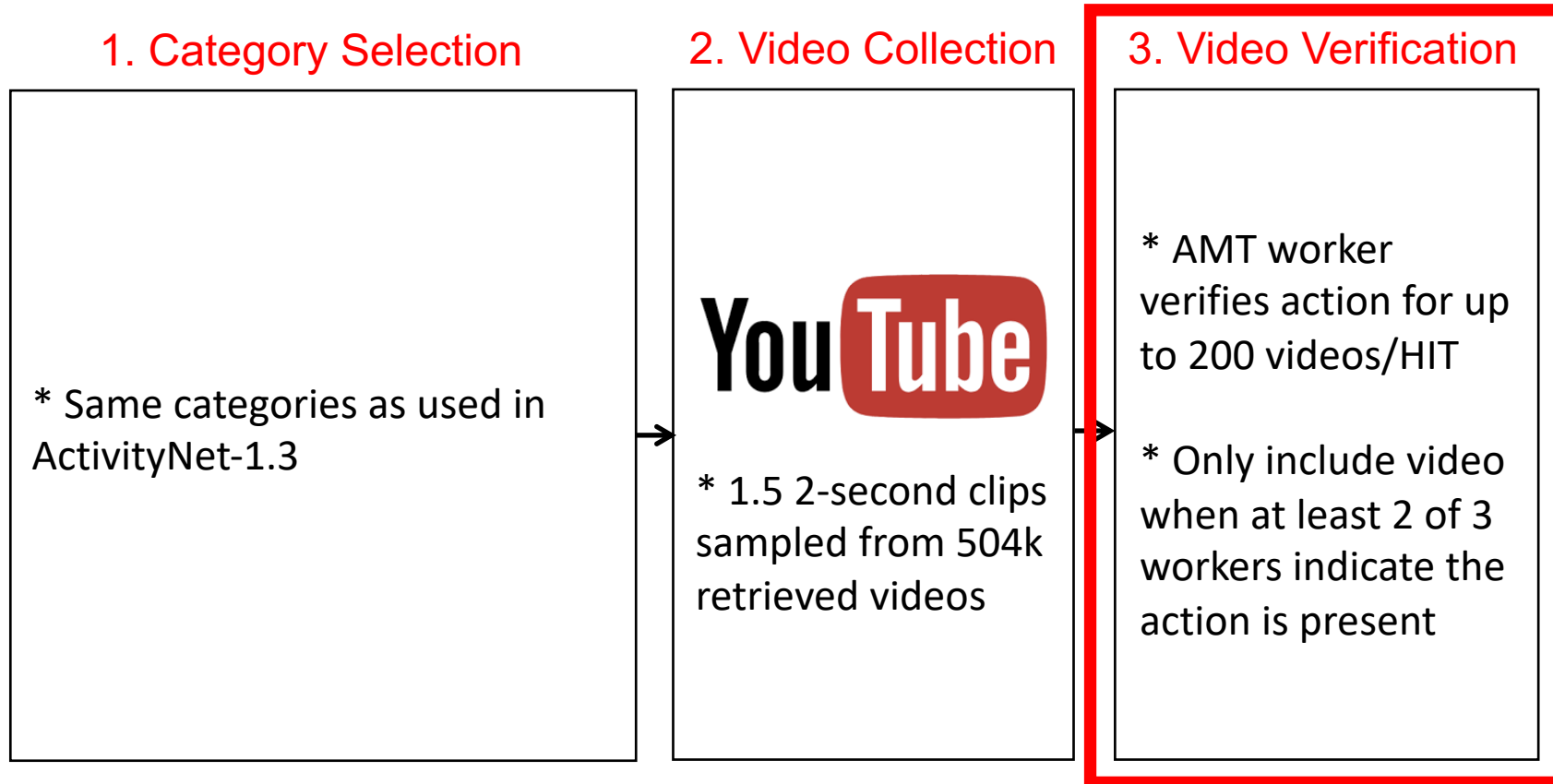
A classification task for students on a **subset** of Moments in Time dataset:

<http://moments.csail.mit.edu/challenge.html>

# Video Classification & Localization Datasets



# HACS




# HACS


High\_jump

10 annotator segments:


Start Time	End Time
24.47	28.38
41.44	45.73
59.44	64.13
72.15	77.37
92.81	97.74
105.56	110.78
128.90	134.24
150.70	155.56
166.98	173.38
181.38	186.96



Overview



Annotation



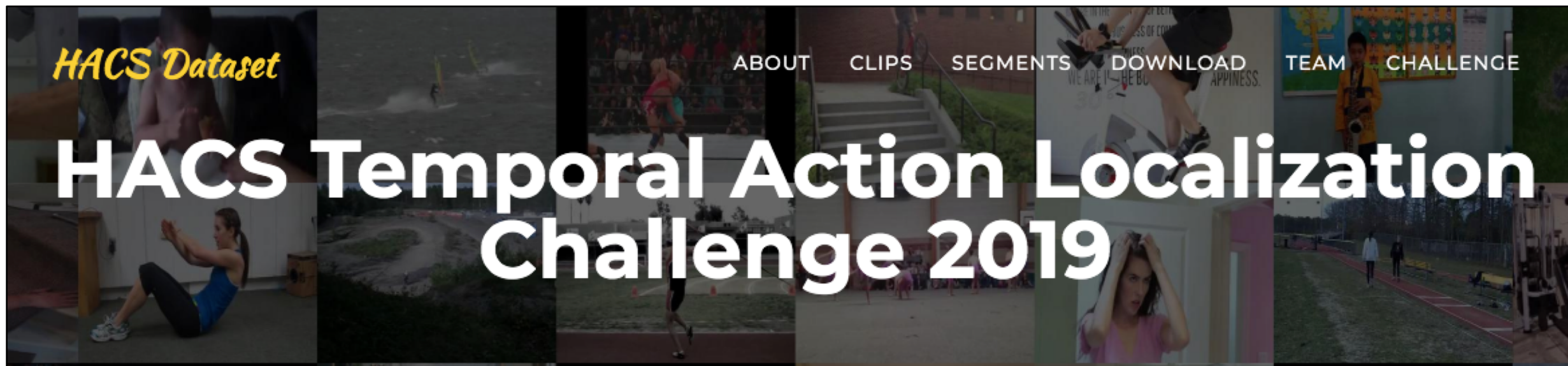
Submitted  
5/02/2018 at 1:00AM

< 29 out of 282 >

Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization. arXiv 2019.



# HACS Challenge and Workshop



We will host HACS Temporal Action Localization Challenge at [ICCV'19 Workshop on Multi-modal Video Analysis](#). The goal of this challenge is to detect actions in untrimmed videos. The action localization challenge uses HACS Segments dataset, which contains:

- 200 classes, 140K action segments
- 37.6K/6K training/validation videos, 6K testing videos
- \* Sparse annotations in HACS Clips dataset are NOT permitted. \*

<http://hacs.csail.mit.edu/>

# Discussion: Video Classification Costs

Assume the task is to classify the presence of 10 activities in 1,000,000 3-minute videos. **How much do you believe it will cost in US dollars to collect all the crowdsourced annotations for the datasets?**

# Today's Topics

- Video classification and localization applications
- Evaluating video classification and localization
- Crowdsourcing video classification and localization
- **Lab: video annotation & writing papers in latex**