# Privacy-Preserving Video Classification with Convolutional Neural Networks

**Sikha Pentyala**
University of Washington Tacoma
sikha@uw.edu

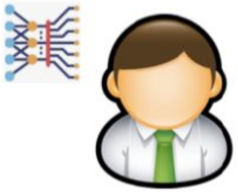
Rafael Dowsley
Monash University
rafael.dowsley@monash.edu

Martine De Cock
University of Washington Tacoma
mdecock@uw.edu

SCHOOL OF ENGINEERING & TECHNOLOGY
UNIVERSITY of WASHINGTON | TACOMA

MONASH University

# Video Classification - Applications




## Surveillance

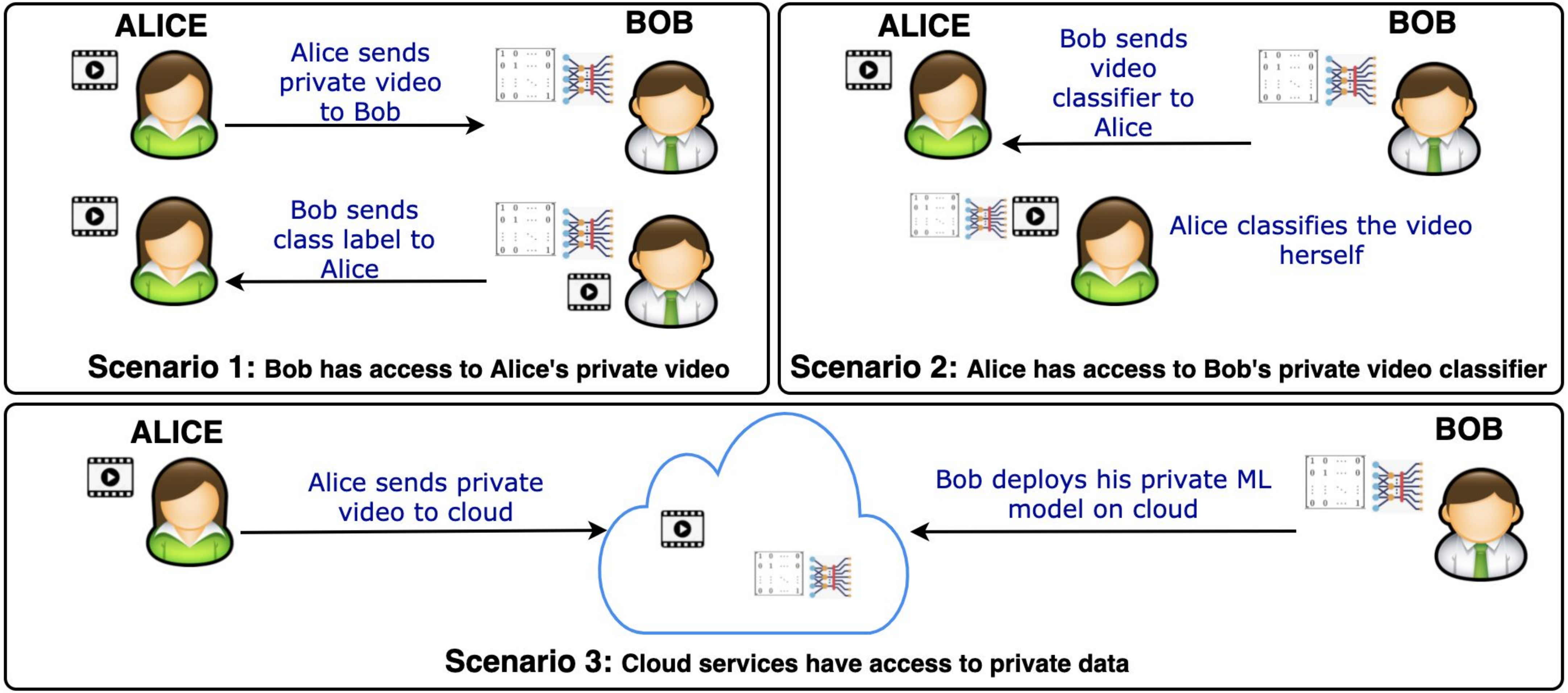
- Security
  - identify strangers, identify threatful actions, home monitoring systems,
  - facial recognition, masked face detection and recognition
- Retail – identify shoplifting
- Detecting concentration of students in online courses
- Activity recognition in care centers – baby monitoring systems, detection of abusive activities

## Behavioral analysis

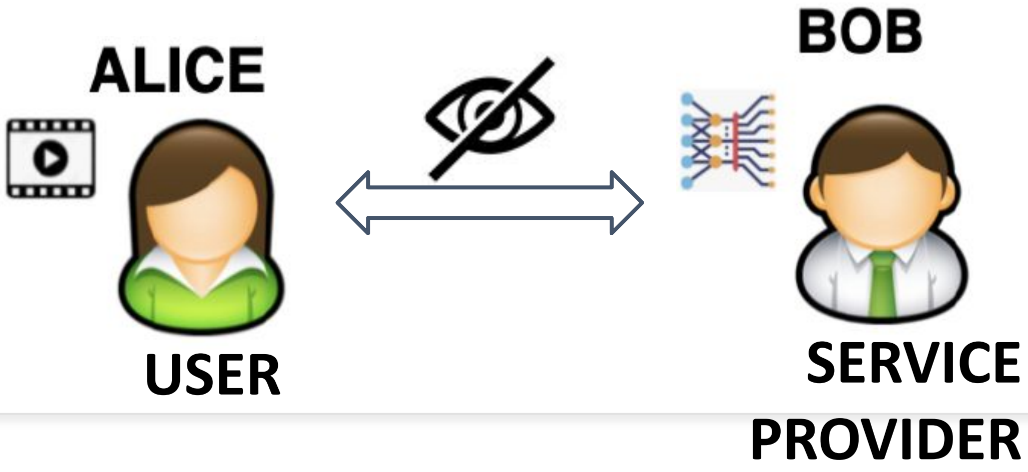
- Gesture analysis
- Sentiment and mood analysis
- Driver drowsiness
- Stress detection
- Eye gaze estimation
- Face, gesture and body analysis for monitoring intervention-measure compliance for COVID-19

**Many more ...**

# Video Classification - Undesirable Scenarios



# Problem Statement




Find a solution to

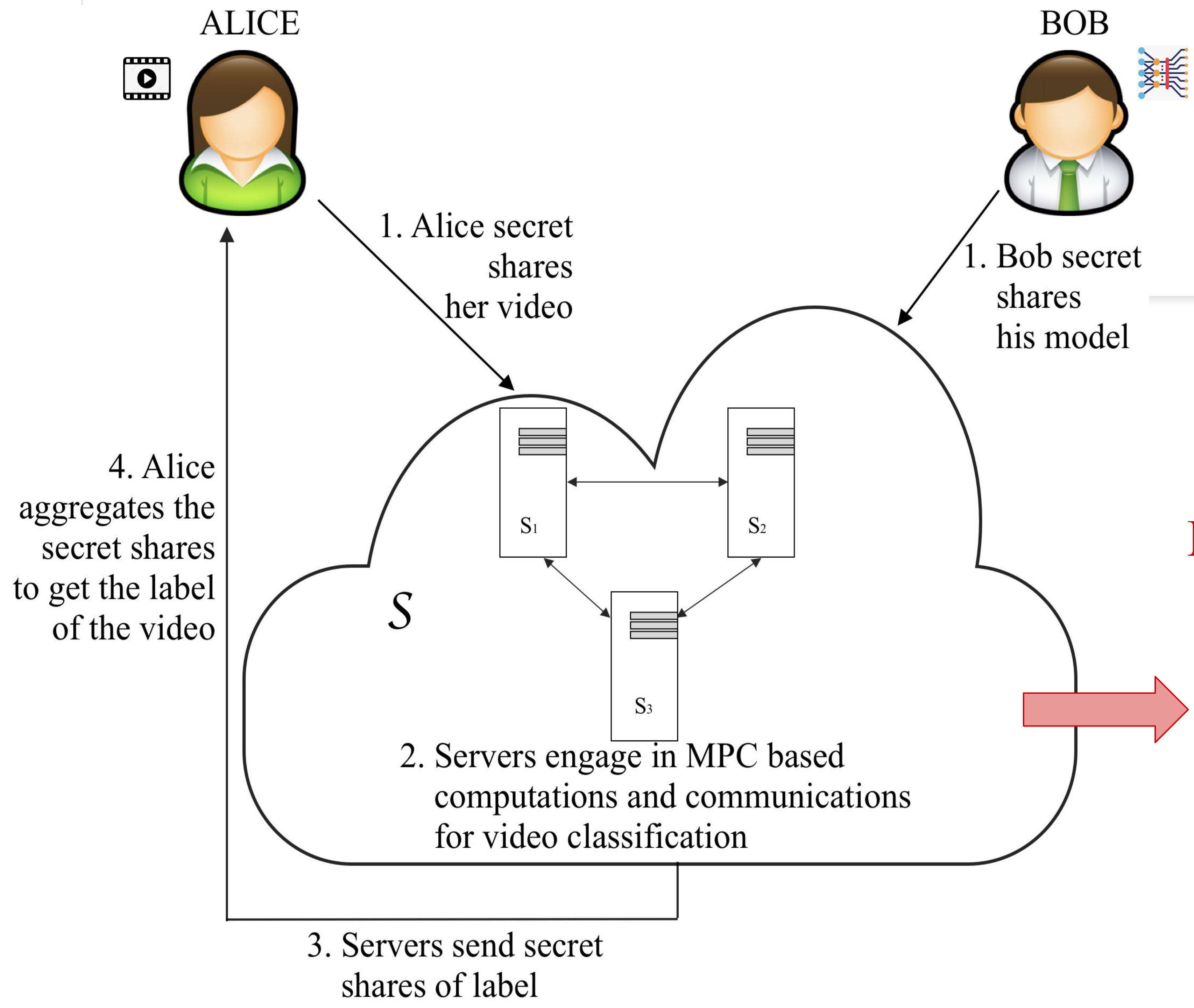
- **Classify** a video
- **Protect** Alice’s video
- **Protect** Bob’s video classifier

with

- **‘No’** information leakage
- **No** special hardware
- **Reduced** computational complexity

using

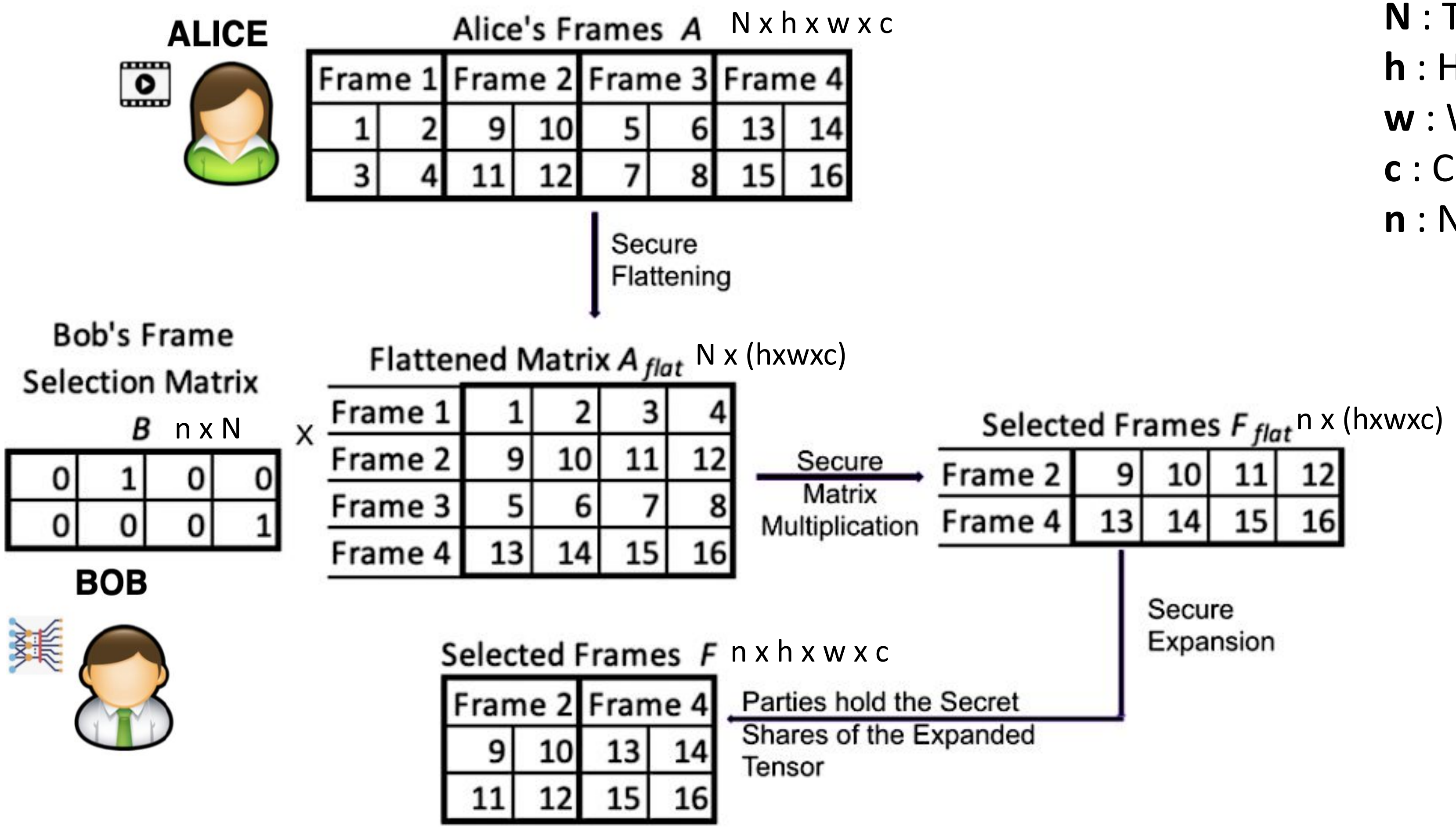
**Secure Multi-Party computation (SMC/MPC)**




**MPC protocols for:**

- **Frame selection**
- **Frame classification with CNN**
- **Label aggregation**

# Step 1: Oblivious Frame Selection

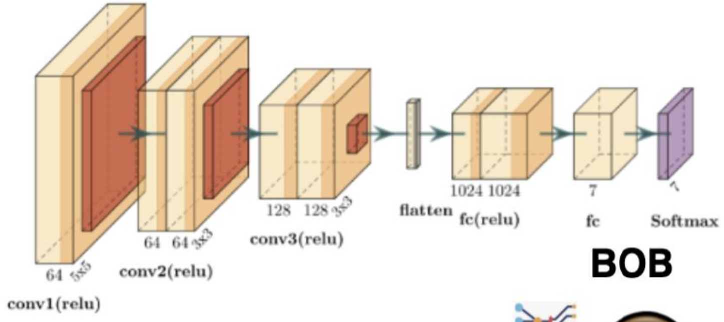



**N** : Total number of frames in video
**h** : Height of each frame
**w** : Width of each frame
**c** : Channels in the video
**n** : Number of frames to be selected

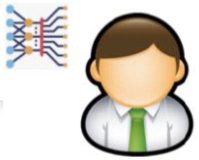
**This example:**
N : 4 frames in video
h : 2
w : 2
c : 1 (grayscale)
n : 2 frames selected

B selects Frame 2 and 4

# Step 2: Private Frame Classification




BOB



- Secure frame classification = secure image classification
- Efficient secure image classification protocols available
- Operations for frame classification*:
  - Convolution: $\pi_{\text{DMM}}, \pi_{\text{DM}}$
  - ReLU: $\pi_{\text{ReLU}}, \pi_{\text{LT}}$
  - Average Pooling: $\pi_{\text{DIV}}$
  - Fully Connected layers: $\pi_{\text{DMM}}$
  - Softmax: $\pi_{\text{SOFT}}$

**Approximated Softmax** :**

$$f(u_i) = \begin{cases} \dfrac{\mathrm{RELU}(u_i)}{\sum_{j=1}^{C} \mathrm{RELU}(u_j)}, & \text{if } \sum_{j=1}^{C} \mathrm{RELU}(u_j) > 0 \\ 1/C, & \text{otherwise} \end{cases}$$

* A. Dalskov, D. Escudero, and M. Keller. Secure evaluation of quantized neural networks. Proceedings on Privacy Enhancing Technologies, 2020(4):355–375, 2020.

** P. Mohassel and Y. Zhang. Secureml: A system for scalable privacy-preserving machine learning. In 2017 IEEE Symposium on Security and Privacy (SP), pages 19–38, 2017.

# Step 3: Secure Label Aggregation

**Protocol 3** Protocol $\pi_{\mathsf{LABELVIDEO}}$ for classifying a video securely based on the single-frame method

**Input:** A video $\mathcal{V}$ secret shared as a 4D-array $[\![A]\!]$, a frame selection matrix secret shared as $[\![B]\!]$, the parameters of the ConvNet model $\mathcal{M}$ secret shared as $[\![M]\!]$

**Output:** A secret share $[\![L]\!]$ of the video label

Let $[\![prob_{\mathsf{sum}}]\!]$ be a list of length $C$ that is initialized with zeros in all indices.
$[\![F]\!] \leftarrow \pi_{\mathsf{FSELECT}}([\![A]\!], [\![B]\!])$
**for all** $[\![F[j]]\!]$ **do**
  $[\![SM_{\mathsf{approx}}]\!] \leftarrow \pi_{\mathsf{FINFER}}([\![M]\!], [\![F[j]]\!])$
  **for** $i = 1$ **to** $C$ **do**
    $[\![prob_{\mathsf{sum}}[i]]\!] \leftarrow [\![prob_{\mathsf{sum}}[i]]\!] + [\![SM_{\mathsf{approx}}[i]]\!]$
  **end for**
**end for**
$[\![L]\!] \leftarrow \pi_{\mathsf{ARGMAX}}([\![prob_{\mathsf{sum}}]\!])$
**return** $[\![L]\!]$

$SM_{approx}$ for Frames

| Labels → | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Frame 1 | 0 | 0 | 0 | 0 | 0.28 | 0 | 0.72 |
| Frame 2 | 0 | 0 | 0 | 0 | 0.55 | 0.45 | 0 |
| Frame 3 | 0 | 0 | 0 | 0 | 0.83 | 0.17 | 0 |
| Frame 4 | 0 | 0.21 | 0 | 0 | 0.48 | 0.31 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $prob_{sum}$ | 0 | 0.21 | 0 | 0 | 2.14 | 0.93 | 0.72 |

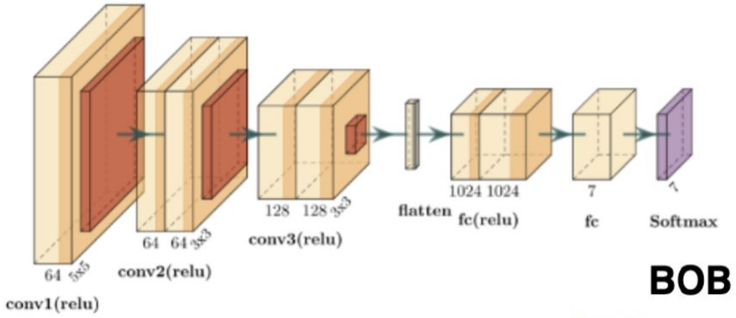
Output Label L is 5

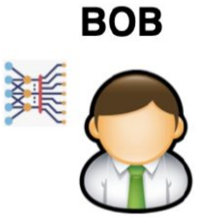
The probabilities for each class are summed up over all the frames.

Index with maximum probability is the class label

# Experiments


- Emotion detection in a video
- RAVDESS dataset*
  - 1,116 videos for train/validation; 132 videos for testing
  - 7 emotions: happy, sad, angry, fearful, surprised, disgust, neutral
- Bob has trained CNN model with 1.5 M parameters
  - video preprocessing: face detection, alignment, cropping, resizing, converting to grayscale, normalization




*S.R. Livingstone and F.A. Russo. The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS One, 13(5), 2018.

data61 / MP-SPDZ

Microsoft Azure

Thanks to Microsoft for donating cloud credits for this research

# Experimental Setup

- 2PC/3PC/4PC: 2, 3 or 4 computing parties (servers), one of which may be corrupted by an adversary

- **passive (semi-honest):** corrupted party follows protocol instructions but tries to learn information from the messages it sees
- **active (malicious):** corrupted party may deviate from protocol instructions

- F32s Azure VMs:  32 vCPUs, 64 GiB Memory, connected over up to 14 Gbps link

*https://github.com/data61/MP-SPDZ

# Results

**Accuracy over the test set** : 56.8% (same as that in-the-clear - without secure pipeline)

*Table 4.* Averages for classifying one RAVDESS video of duration 3-5 seconds. Average metrics are obtained over a set of 10 such videos with a number of frames in the 7-10 range on F32s VMs with n_threads=32 in MP-SDPZ. VC: time to classify one video ($\pi_{\mathsf{LABELVIDEO}}$); FS: time for frame selection for one video ($\pi_{\mathsf{FSELECT}}$); FI: time to classify a selected frame for one video averaged over all selected frames in the videos ($\pi_{\mathsf{FINFER}}$); LA: time taken for label aggregation (sum up all probabilities, $\pi_{\mathsf{ARGMAX}}$). Communication is measured per party.

| F32s V2 | VMs | Time VC | Time FS | Time single FI | Time LA | Comm. VC |
|---|---|---|---|---|---|---|
| Passive | 2PC | 302.24 sec | 12.95 sec | 35.38 sec | 0.00500 sec | 374.28 GB |
| | 3PC | **8.69 sec** | 0.07 sec | 0.26 sec | 0.00298 sec | 0.28 GB |
| Active | 2PC | 6576.27 sec | 393.57 sec | 759.211 sec | 0.00871 sec | 5492.38 GB |
| | 3PC | 27.61 sec | 0.94 sec | 2.05 sec | 0.00348 sec | 2.29 GB |
| | 4PC | **11.67 sec** | 0.15 sec | 0.57 sec | 0.00328 sec | 0.57 GB |

# Conclusion and Future Work

- First baseline end-to-end privacy-preserving solution to classify a video using MPC
- Novel baseline MPC protocols for
  - oblivious frame selection
  - secure label aggregation
- Demonstrated feasibility of our solution to detect emotions in a video
  - with no information leakage (mathematically provable)
  - with state-of-the-art accuracy: as accurate as in-the-clear (without encryption)
  - no special hardware

Future directions

- Use of machine learning for intelligence frame selection
- Develop MPC protocols for other state-of-the-art video classification methods beyond single-frame technique

Thank You