# Activity Recognition Using Both Computer Vision and Audio Processing


**Sancharz Gore**
Harvard University

**Alyssa Lach**
Butler University

**Bella Murrer**
University of Notre Dame

**Regina Rex**
University of Wisconsin

**Fiona Ryan**
Indiana University

**Kara Schatz**
Xavier University

**Sophie Tian**
University of Washington

**Trang Tran**
Knox College

**Violet Xiang**
Indiana University

**David Crandall**
Indiana University

**Donald Williamson**
Indiana University


#Hello Research!

# Introduction & Motivation

- How do you recognize an acquaintance?

# Introduction & Motivation

- How do you recognize an acquaintance?
  - by voice and appearance

# Introduction & Motivation

- How do you recognize an acquaintance?
  - by voice and appearance
- What about machines?

# Introduction & Motivation

- How do you recognize an acquaintance?
  - by voice and appearance

- What about machines?
  - with our findings in the machine learning field and labeled data, will machines have the ability to recognize the speakers?

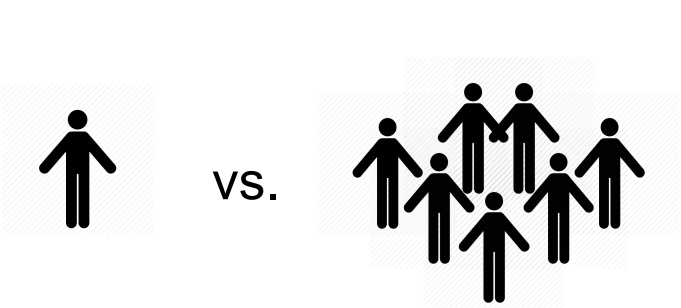
Sancharz

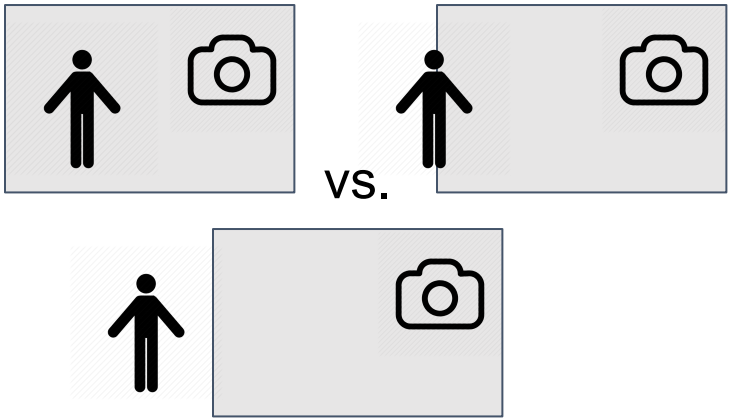
# Applications

- Personalized Smart Assistants (Alexa, Google Home,Siri…)
- Smart Baby Monitors
- Emotion Detection
- Image Captioning

# Task

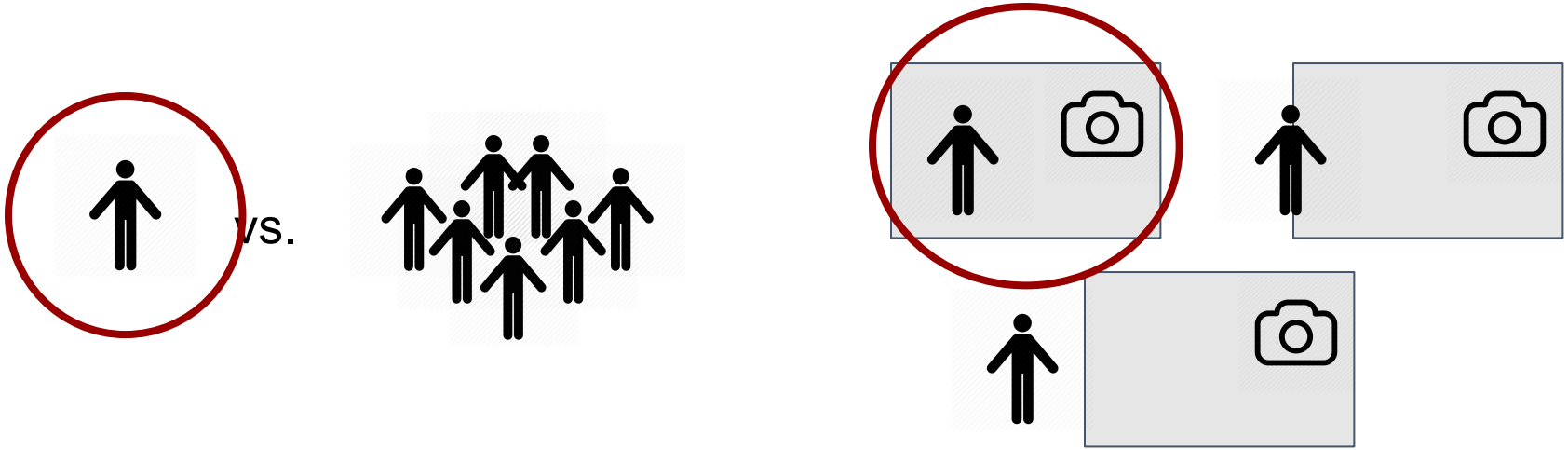
Given audio and visual data of a person talking, can we identify who it is?



# Task

Given audio and visual data of a person talking, can we identify who it is?

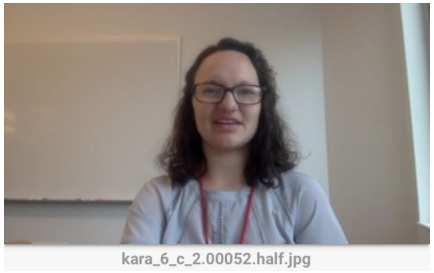


# Data Collection:

- **Background Variation:**
  - 5 different backgrounds
- **Sentence Variation:**
  - 10 different sentences
- **Clothing Variation:**
  - 2 different clothing (jacket/no jacket)
- **Total Data Collected (10 people):**
  - Video Samples: 188
  - Audio Samples: 188
  - Image Samples: 16,668

# Visual Recognition

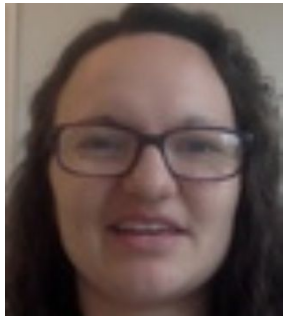
Fiona Ryan




Training data: sampled frames from videos labeled with correct person in video


```
{
  "faceAnnotations": [
    {
      "boundingPoly": {
        "vertices": [
          {
            "x": 130,
            "y": 27
          },
          {
            "x": 205,
            "y": 27
          },
          {
            "x": 205,
            "y": 113
          },
          {
            "x": 130,
            "y": 113
          }
        ]
      },
```


Output from Google Cloud Vision API



Standard size JPEG image for neural network

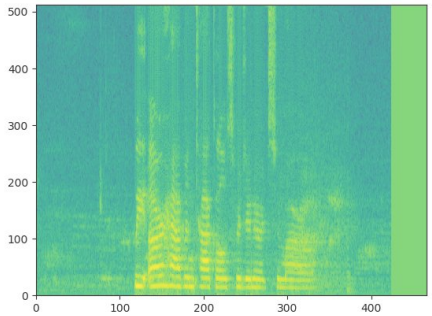
**Convolutional Neural Network**

Determines most likely person in a given picture by comparing it to the training data

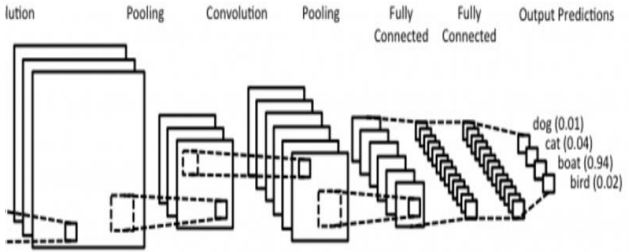
Bella Murrer and Trang Tran

# Audio Recognition

- Collecting data
- Converting mp3 to wav file
- Loaded wav file data to Google Colab
- Converting wav file to spectrograms
- Shuffle, batch and normalize the spectrograms
- Build Convolutional Neural Network
- Parameter tuning - Learning rate




Spectrogram of Bella's Voice




Convoluted Neural Network

# Preliminary Results

Regina Rex

- **Computer Vision:**
  - **Classification Accuracy:**
    - Frame Classification Model: **45%**
      - Classifies every frame in the video
    - Video Classification Model: **60%**
      - Groups the frames and weight the videos based on the classification of the Group

- **Audio Processing:**
  - **Training Model:** Hasn't been applied to the test data (set)
    - Minimum data loss: .286 from Audio Training Model

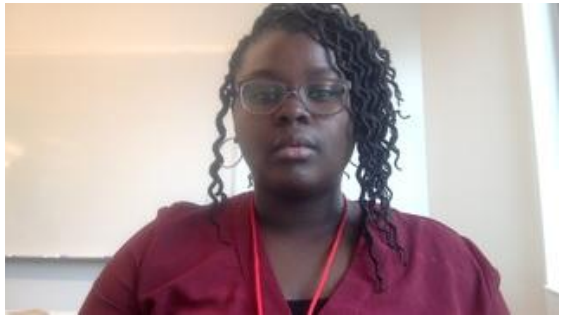
# Preliminary Results

Regina Rex

## Confusion(Error) Matrix:

| | Alyssa | Bardia | Bella | Fiona | Kara | Regina | Sancharz | Sophie | Trang | Violet |
|---|---|---|---|---|---|---|---|---|---|---|
| Alyssa | 2 | | | | | | | | | |
| Bardia | | 2 | | | | | | | | |
| Bella | | | | 2 | | | | | | |
| Fiona | | | | | 2 | | | | | |
| Kara | | | | | 2 | | | | | |
| Regina | | 2 | | | | | | | | |
| Sancharz | | 1 | | | | | 1 | | | |
| Sophie | | | | | | | | 2 | | |
| Trang | | | | 1 | | | | | 1 | |
| Violet | | | | | | | | | | 2 |

**Bardia (Graduate Student)**



**Me**

# Future work

1. **Improvement on Image Classification**
   a. Aggregate information from a sequence of frames, such as motion
   b. Method for improving classification accuracy, i.e. cropping faces
   c. Identify multiple people in images
2. **Improvement on Speaker Identification**
   a. Finish Training
   b. Audio to identify speakers from multi-speakers
3. **Combine audio and video information to get better results**
4. **Be able to identify strangers in video**