AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

LARGE SCALE AUDIO-VISUAL VIDEO ANALYTICS PLATFORM FOR FORENSIC INVESTIGATIONS OF TERRORISTIC ATTACKS

Alexander Schindler
Martin Boyer
Andrew Lindley
David Schreiber
Thomas Philipp

Alexander Schindler
Scientist
Information Management
Center For Digital Safety & Security

AIT Austrian Institute Of Technology Gmbh
Alexander.schindler@ait.ac.at

LARGE SCALE AUDIO-VISUAL VIDEO ANALYTICS PLATFORM

FOR

FORENSIC INVESTIGATIONS

OF

TERRORISTIC ATTACKS

Alexander Schindler

Martin Boyer

Andrew Lindley

David Schreiber

Thomas Philipp

Alexander Schindler
Scientist
Information Management
Center For Digital Safety & Security

AIT Austrian Institute Of Technology Gmbh
Alexander.schindler@ait.ac.at

FORENSIC INVESTIGATIONS
OF
TERRORISTIC ATTACKS

- **Context**
  - Forensic Investigation
  - Investigating video data after a terroristic attack

- **Objectives**
  - Spot suspects
  - Follow hints by civilian witnesses
  - Collect and secure evidence
  - Prevent immediate or subsequent attacks

FORENSIC INVESTIGATIONS OF TERRORISTIC ATTACKS

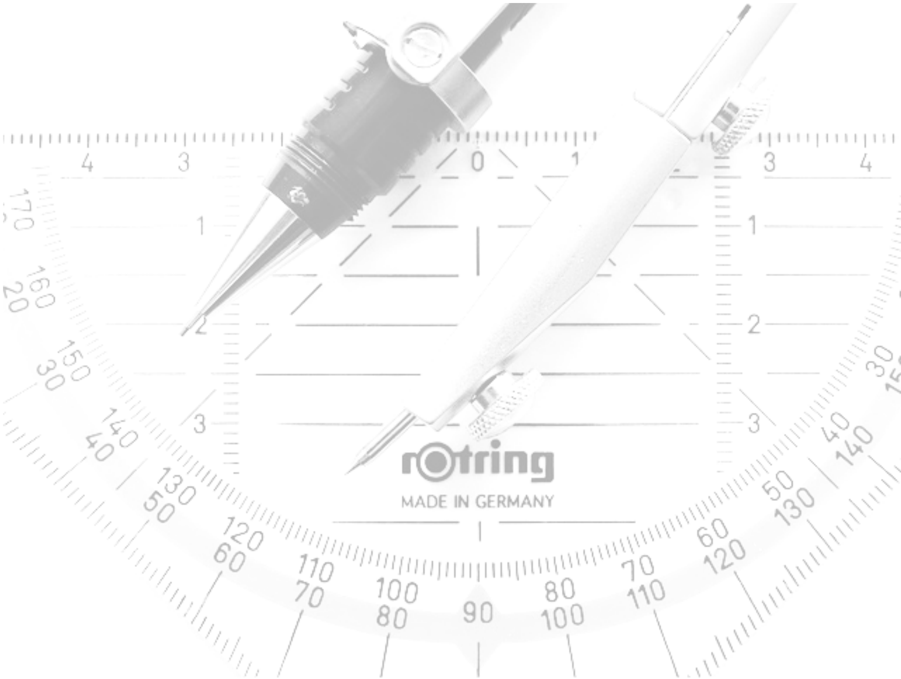
- **Obstacles**
  - Great increase in the number of public and private cameras
  - Massively increasing volume of video data to be analysed
    - Boston Marathon Bombing 5.000h
    - Toulouse and Montauban:10.000h (35TB)
  - Time pressure
    - Timely content evaluation of video mass data is of considerable importance

FORENSIC INVESTIGATIONS OF TERRORISTIC ATTACKS

- **Initial Situation (before project)**
  - manual viewing/processing of the video material
  - Personnel-intensive: time span from several hundred to several thousand hours
  - **Technical, supporting tools necessary**
- **Projects Goals and Outcomes**
  - 2 Projects
    - FLORIDA (Bi-Lateral funding Austria/Germany) => intial research
    - VICTORIA (H2020) => TRL 6 - 10
  - Large-scale computing platform
  - Analytical modules

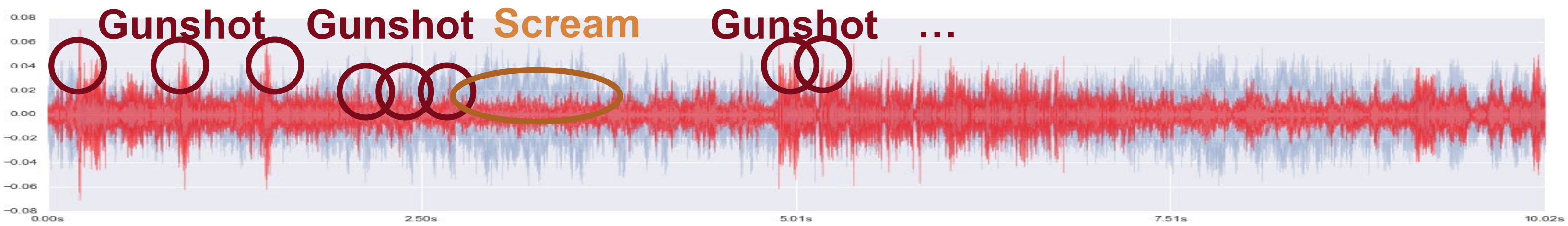
AUDIO EVENT DETECTION

Analytic Modules



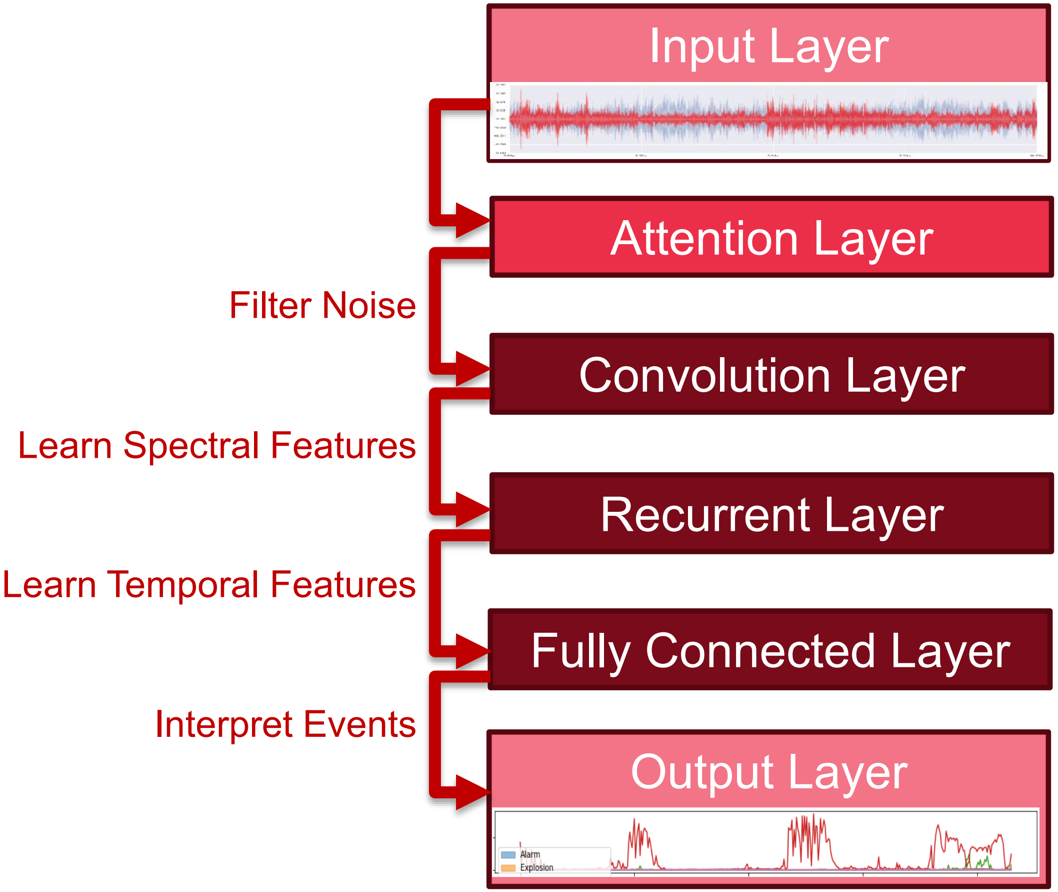
AUDIO EVENT DETECTION

- **Task**
  - Detect and predict audio-events into predefined categories
    - Gunshots, explosions, emergency vehicles, scream, speech, Alarm
- **Use-Case**
  - Content filter in mass video-data
  - Example: attack with firearms
    - => initiate search by filtering all videos which contain *Gunshots* (sorted by confidence)



AUDIO EVENT DETECTION

- **Applied Technology**
  - Recurrent Convolutional Neual Networks
  - With Attention Layer



Input Layer

Attention Layer

Filter Noise

Convolution Layer

Learn Spectral Features

Recurrent Layer

Learn Temporal Features

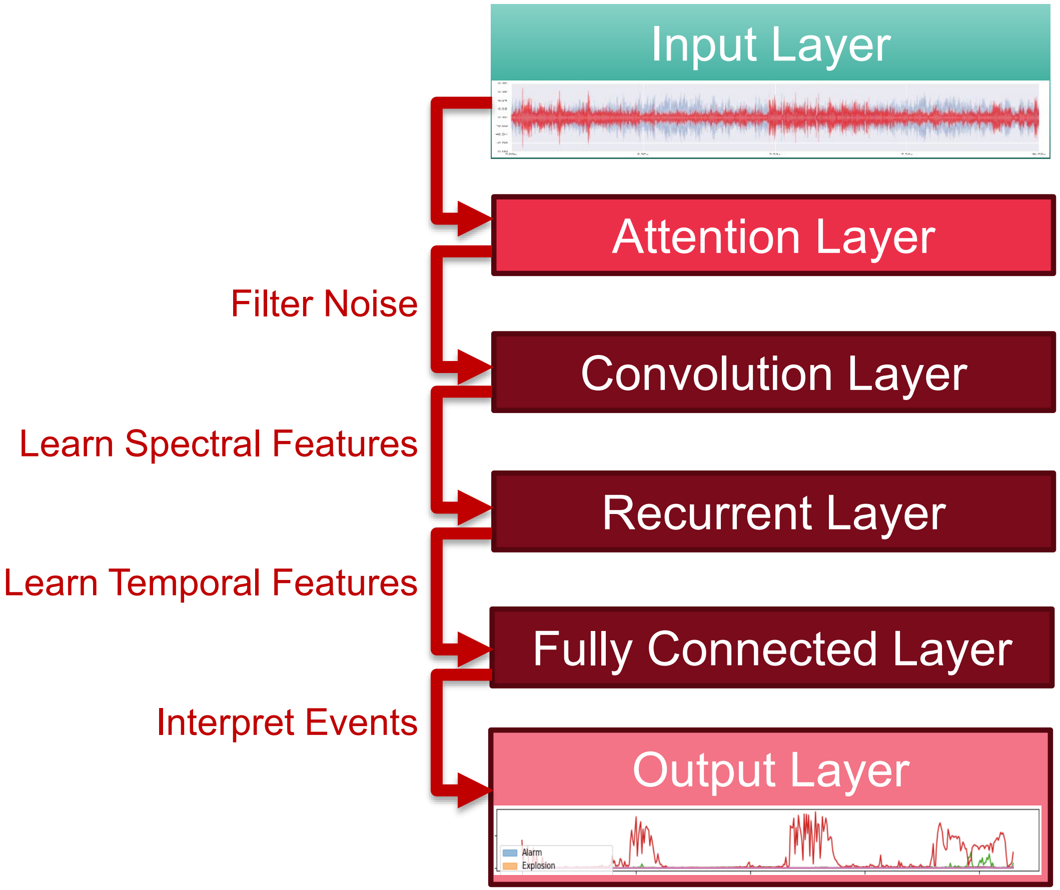
Fully Connected Layer

Interpret Events

Output Layer

AUDIO EVENT DETECTION

1. **Input representation**
   - **Common: Mel-Spectrograms**



Input Layer

Attention Layer

Filter Noise

Convolution Layer

Learn Spectral Features

Recurrent Layer

Learn Temporal Features

Fully Connected Layer

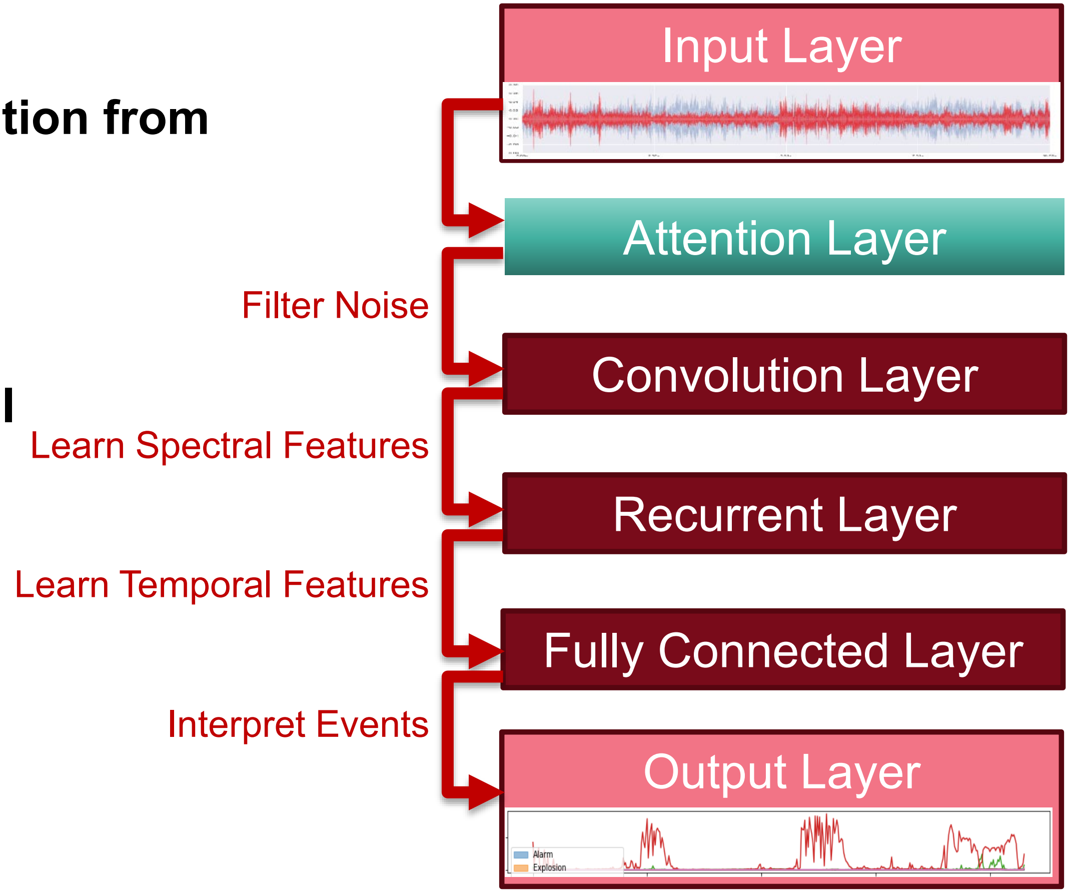
Interpret Events

Output Layer

AUDIO EVENT DETECTION

1. Input representation
   - Common: Mel-Spectrograms

- **Attention Layer**
  - **Filter non-relevant information from Input**
  - **Help to learn faster**
  - **Better convergence**
  - **Better generalization**
  - **Smoother prediction signal**



Input Layer

Attention Layer

Filter Noise

Convolution Layer

Learn Spectral Features

Recurrent Layer

Learn Temporal Features

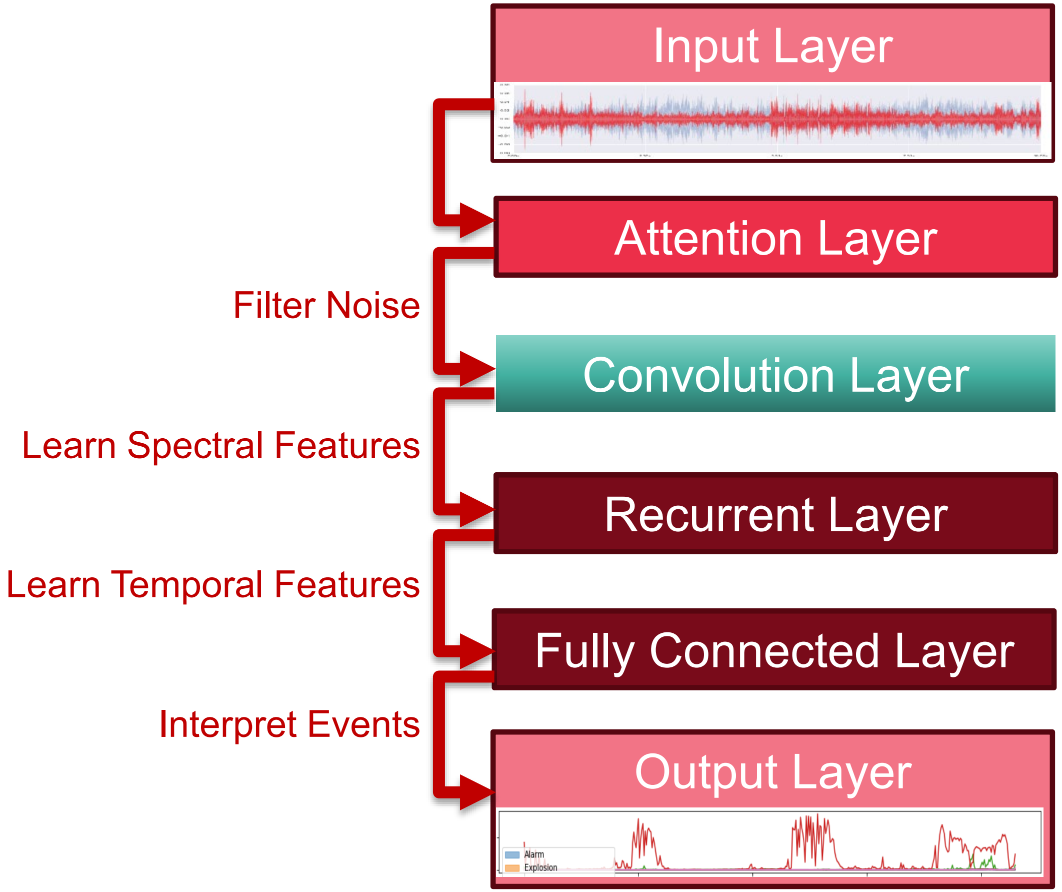
Fully Connected Layer

Interpret Events

Output Layer

AUDIO EVENT DETECTION

1. Input representation
   - Common: Mel-Spectrograms
2. **Convolutional Neural Network Block (CNN)**
   - **Learn audio embeddings**



Input Layer

Attention Layer

Filter Noise

Convolution Layer

Learn Spectral Features

Recurrent Layer

Learn Temporal Features

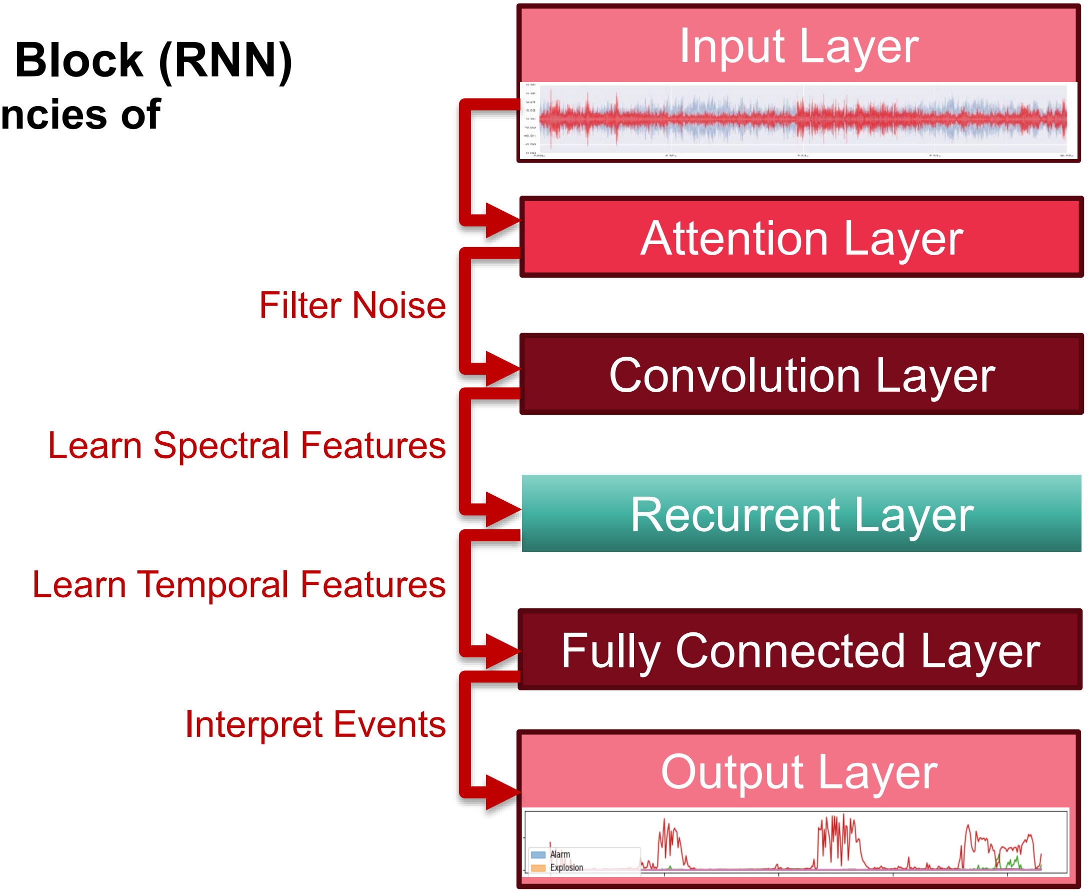
Fully Connected Layer

Interpret Events

Output Layer

AUDIO EVENT DETECTION

1. Input representation
   - Common: Mel-Spectrograms
2. Convolutional Neural Network Block (CNN)
   - Learn audio embeddings
3. **Recurrent Neural Network Block (RNN)**
   - **Learn Temporal dependencies of embeddings**



Input Layer

Attention Layer

Filter Noise

Convolution Layer

Learn Spectral Features

Recurrent Layer

Learn Temporal Features

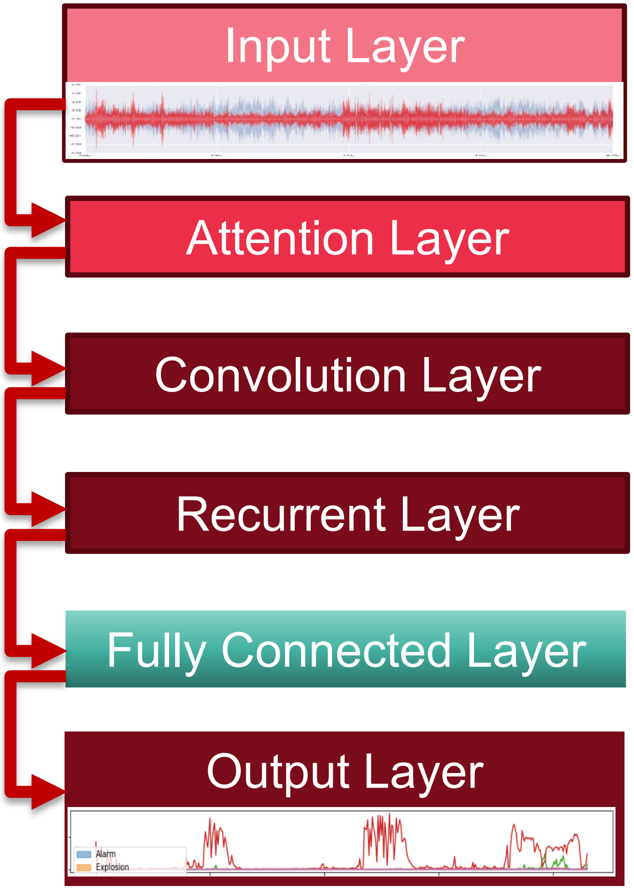
Fully Connected Layer

Interpret Events

Output Layer

AUDIO EVENT DETECTION

1. Input representation
   - Common: Mel-Spectrograms
2. Convolutional Neural Network Block (CNN)
   - Learn audio embeddings
3. Recurrent Neural Network Block (RNN)
   - Learn Temporal dependencies of embeddings
4. **Array of Fully Connected Layers**
   - **One Layer per temporal dimension (Time-Distributed)**
   - **Dimensionality of Layer = Number of classes**



Input Layer

Attention Layer

Convolution Layer

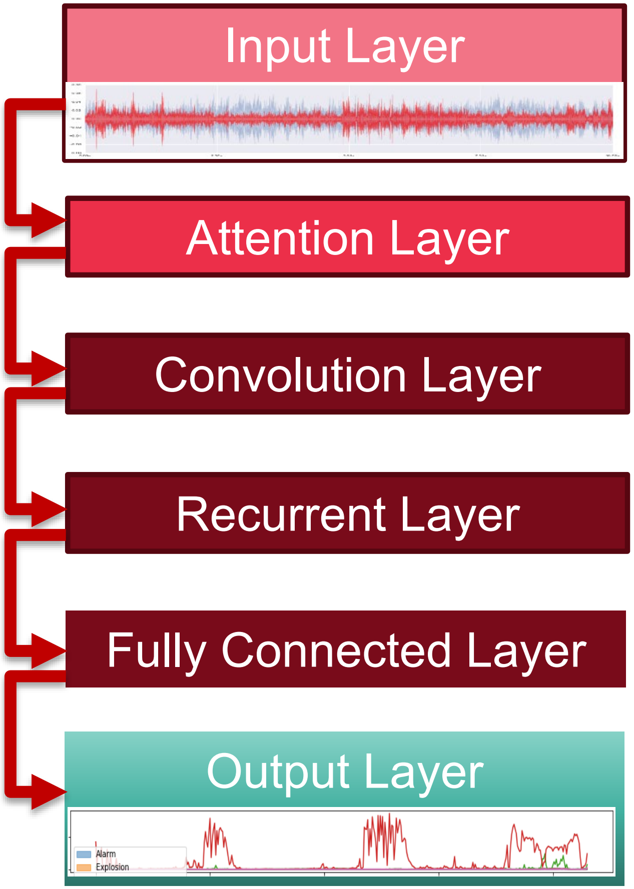
Recurrent Layer

Fully Connected Layer

Output Layer

AUDIO EVENT DETECTION

1. Input representation
   - Common: Mel-Spectrograms
2. Convolutional Neural Network Block (CNN)
   - Learn audio embeddings
3. Recurrent Neural Network Block (RNN)
   - Learn Temporal dependencies of embeddings
4. Array of Fully Connected Layers
   - One Layer per temporal dimension (Time-Distributed)
   - Dimensionality of Layer = Number of classes
5. **Outputs**
   - **Strong Labels – Training & Inference**
     - Output of Time-Distributed Fully Connected Layers
   - **Weak Labels - Training**
     - Output Layer aggregation (e.g. avg, max)
     - Multi label prediction



Input Layer

Attention Layer

Convolution Layer

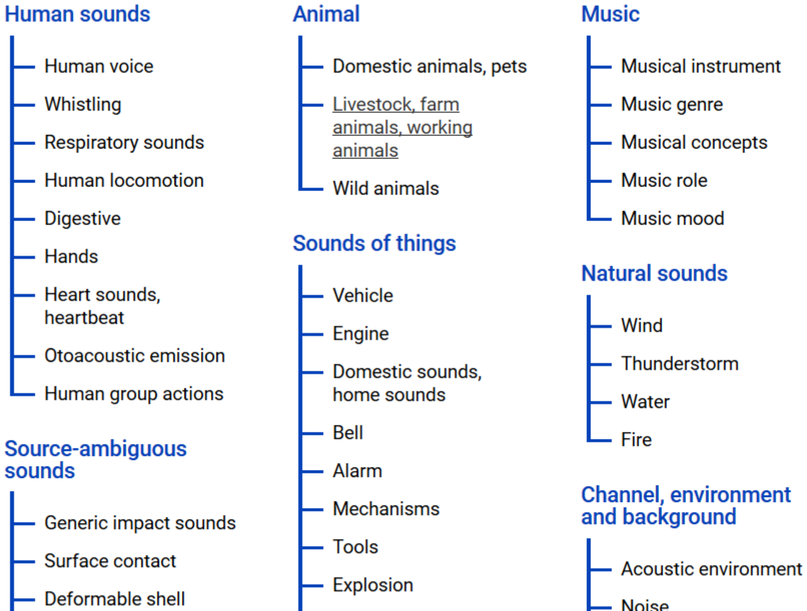
Recurrent Layer

Fully Connected Layer

Output Layer

Google Audio Set

- 2M Videos
- 632 audio events
- annotaded according acoustic categories
- Weakly labelled (10s)
- Currently largest source of data



Human sounds
- Human voice
- Whistling
- Respiratory sounds
- Human locomotion
- Digestive
- Hands
- Heart sounds, heartbeat
- Otoacoustic emission
- Human group actions

Source-ambiguous sounds
- Generic impact sounds
- Surface contact
- Deformable shell

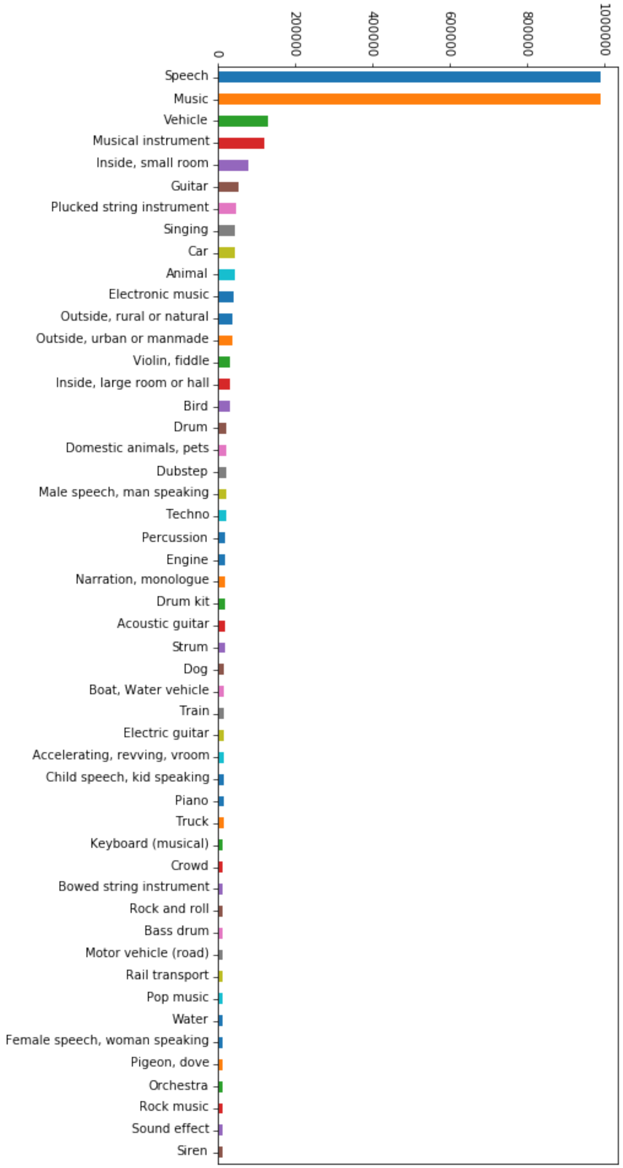
Animal
- Domestic animals, pets
- Livestock, farm animals, working animals
- Wild animals

Sounds of things
- Vehicle
- Engine
- Domestic sounds, home sounds
- Bell
- Alarm
- Mechanisms
- Tools
- Explosion

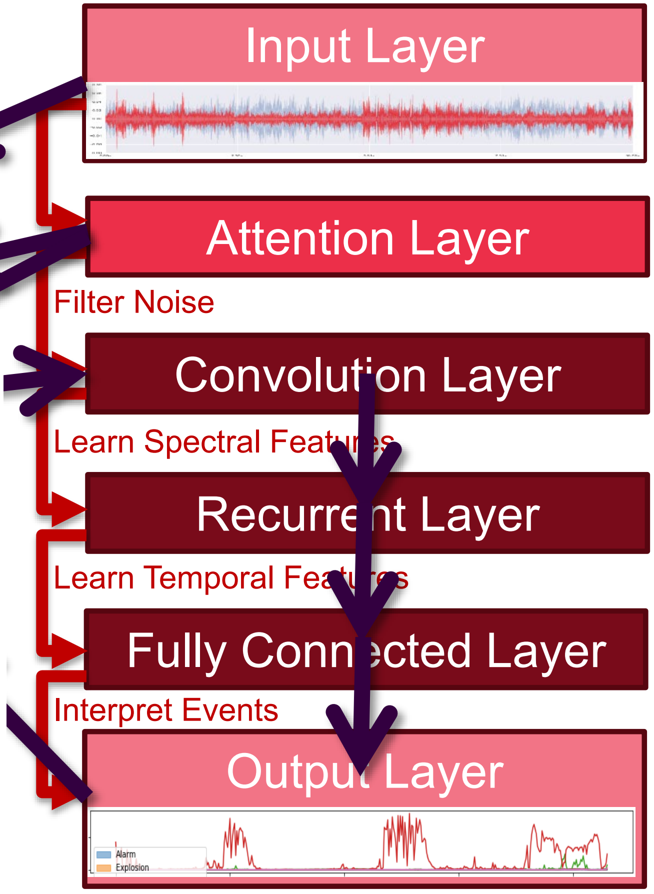
Music
- Musical instrument
- Music genre
- Musical concepts
- Music role
- Music mood

Natural sounds
- Wind
- Thunderstorm
- Water
- Fire

Channel, environment and background
- Acoustic environment
- Noise



RECURRENT CONVOLUTIONAL NEURAL NETWORKS



Input Layer

Attention Layer

Filter Noise

Convolution Layer

Learn Spectral Features

Recurrent Layer

Learn Temporal Features

Fully Connected Layer

Interpret Events

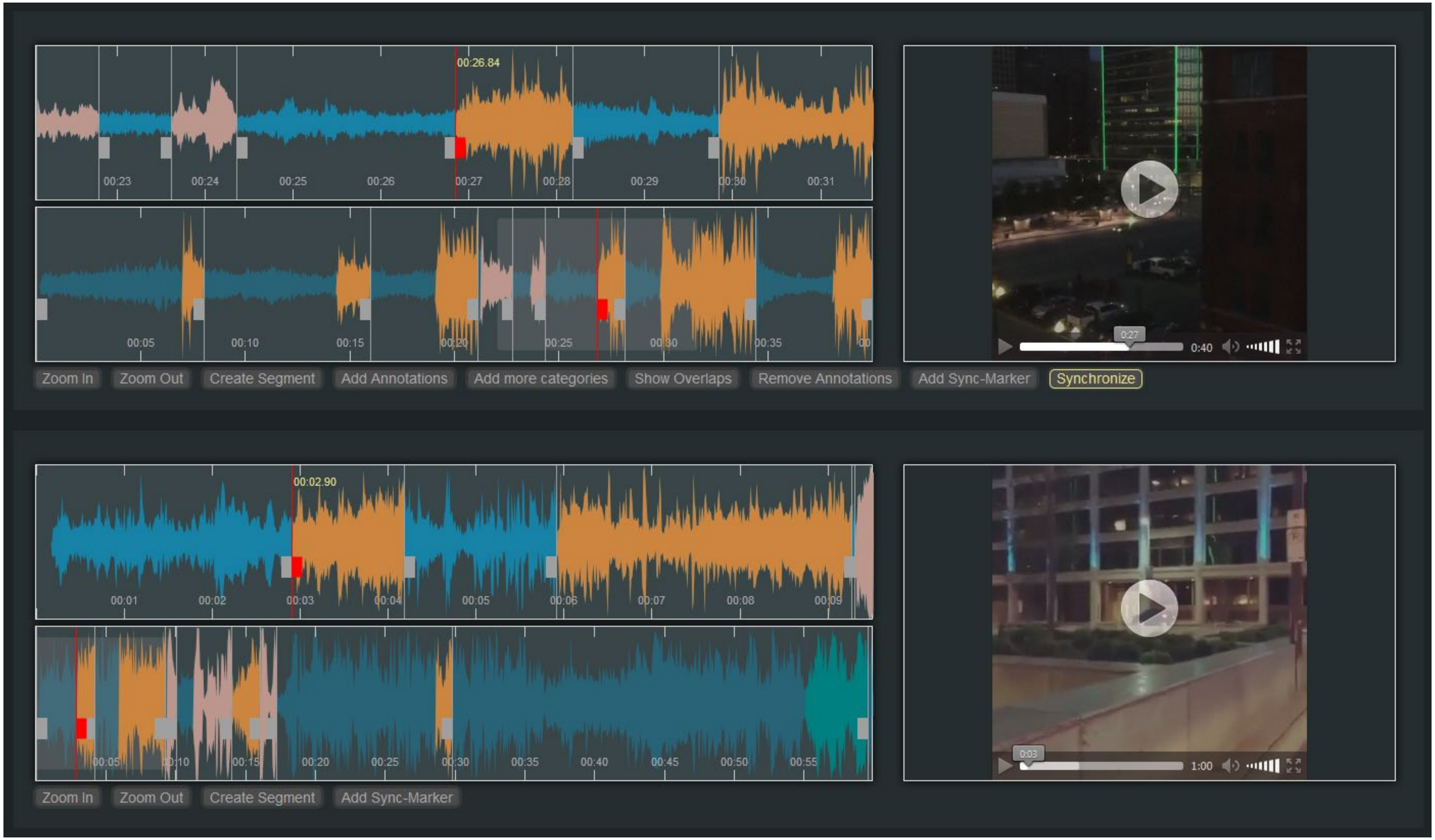
Output Layer

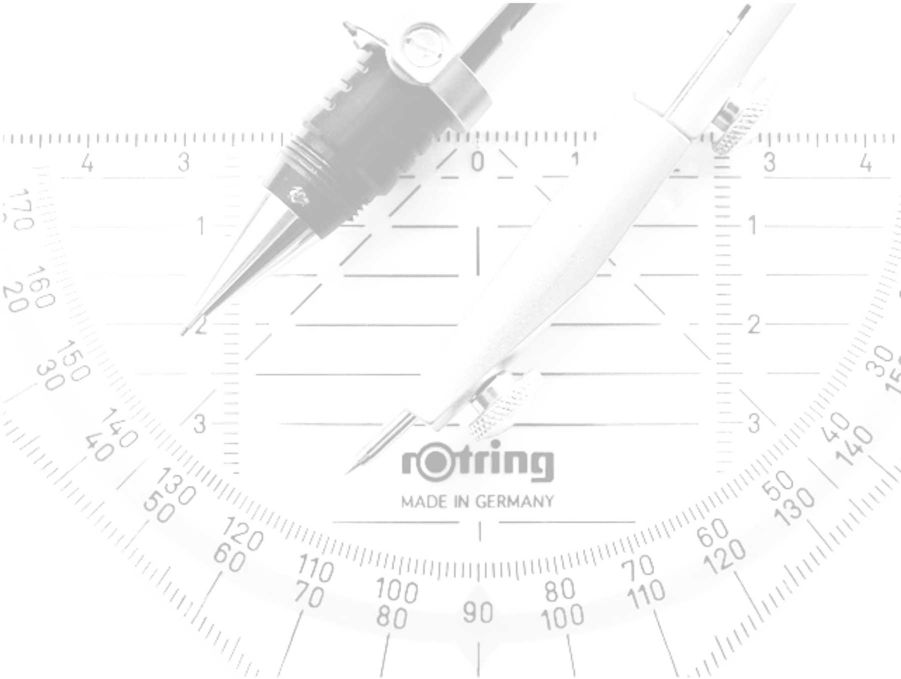
High signal to noise ratio
Due to attention layer

Smoothing Function
Temporal Segmentation per Event Category

AUDIO EVENTS VISUALIZED

User Interface Prototyp





AUDIO SIMILARITY

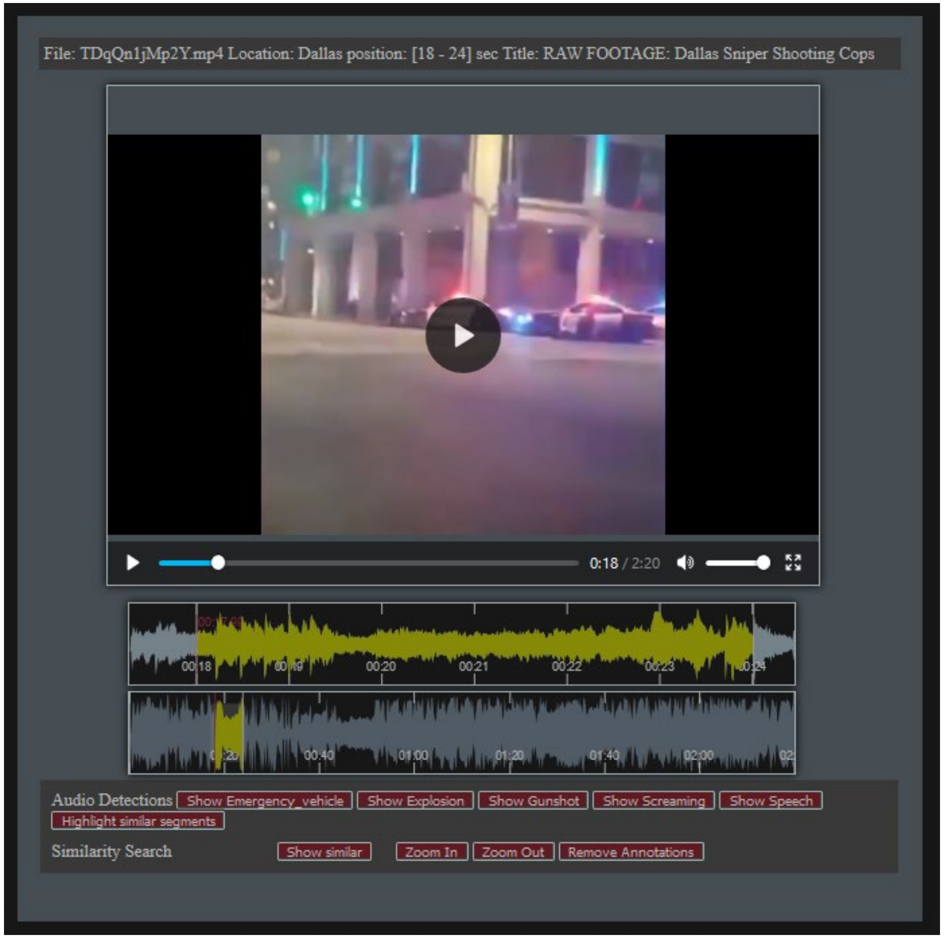
Analytic Modules

AUDIO SIMILARITY SEARCH

- **Task**
  - Searching for video-segments with similar audio-signature
  - Sub-Segment video-search

- **Use-Case**
  - Suspect could not be identified in one video
  - Select segment and search for others using audio-signature
  - Instant localization (videos close to audio source)

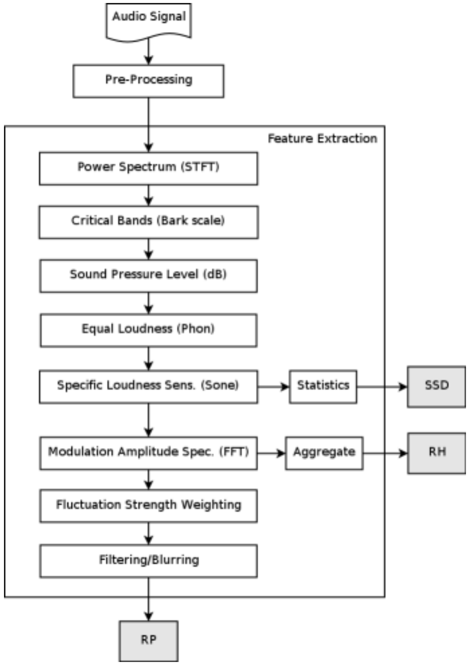
- **Solution**
  - Select range in video
  - Retrieve a list of similar sounding video segments
  - Sorted by simlarity

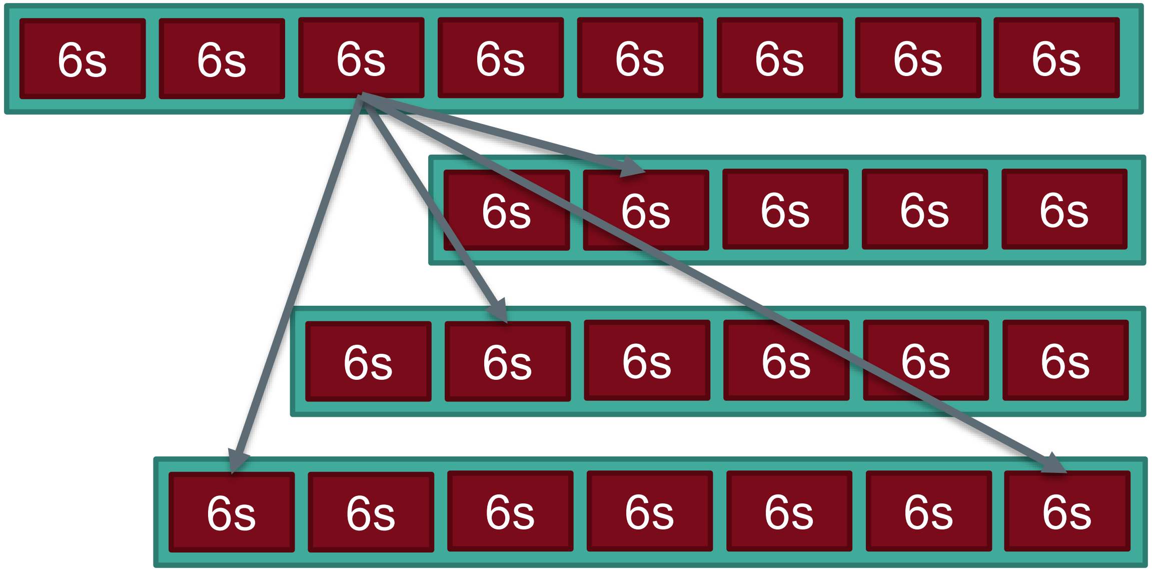


File: TDqQn1jMp2Y.mp4 Location: Dallas position: [18 - 24] sec Title: RAW FOOTAGE: Dallas Sniper Shooting Cops

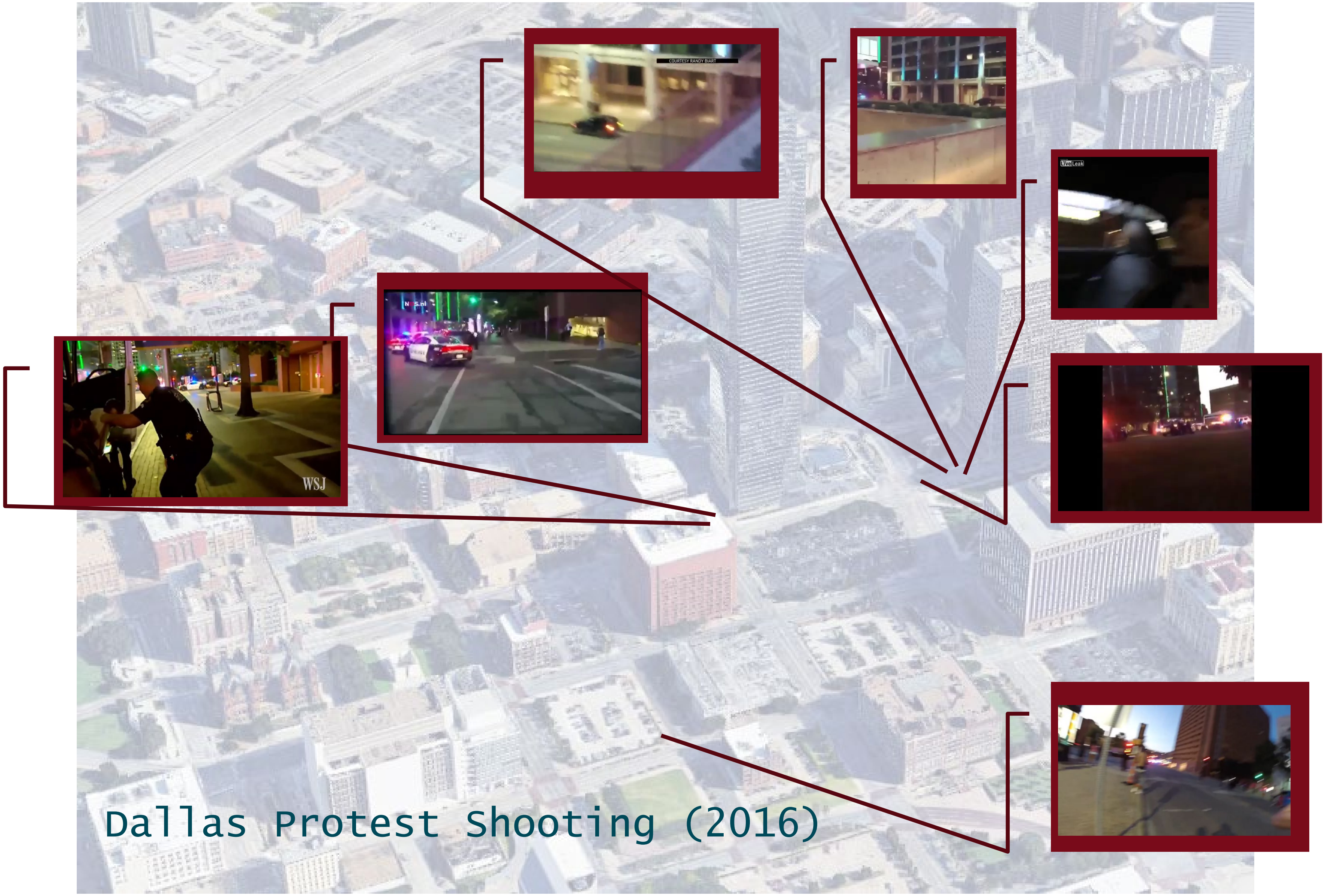
AUDIO SIMILARITY SEARCH

- **Audio features extracted for each 6s segment**
  - Rhythm Patterns (repetitiveness in audio)
  - Statistical Spectrum Descriptors
- Nearest Neighbor search using late fusion in a normalized feature space


Audio Signal
Pre-Processing
Feature Extraction
Power Spectrum (STFT)
Critical Bands (Bark scale)
Sound Pressure Level (dB)
Equal Loudness (Phon)
Specific Loudness Sens. (Sone)
Statistics
SSD
Modulation Amplitude Spec. (FFT)
Aggregate
RH
Fluctuation Strength Weighting
Filtering/Blurring
RP



AUDIO SIMILARITY – INSTANT LOCALIZATION (WITHOUT GPS)

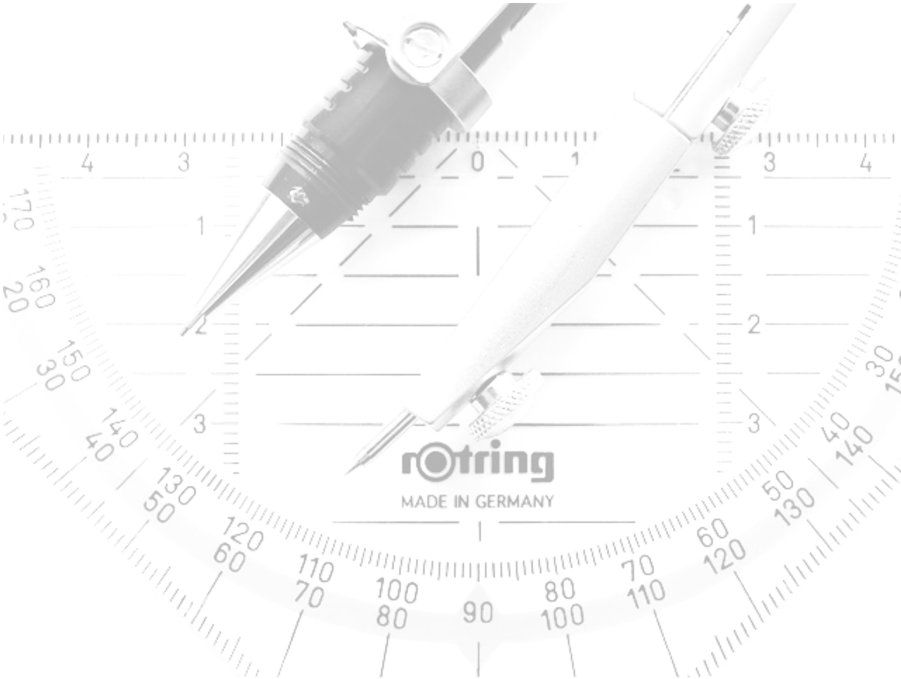


Dallas Protest Shooting (2016)

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

AUDIO-BASED VIDEO-SYNCHRONIZATION

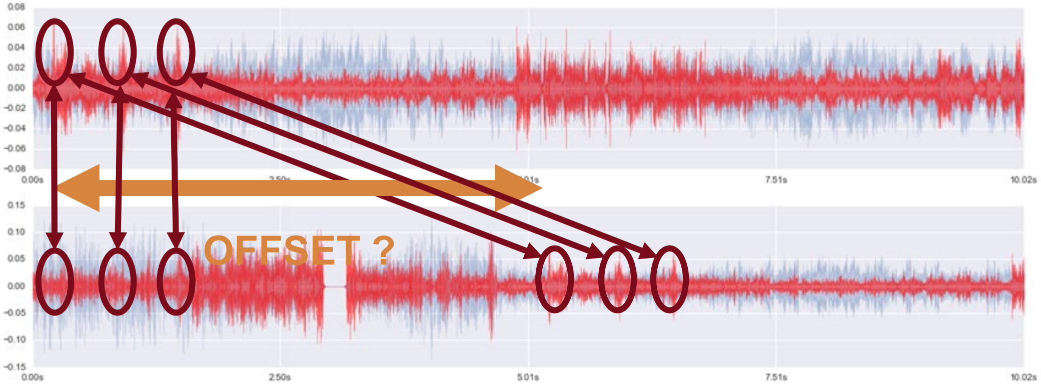
Analytic Module



AUDIO-BASED VIDEO-SYNCHRONIZATION

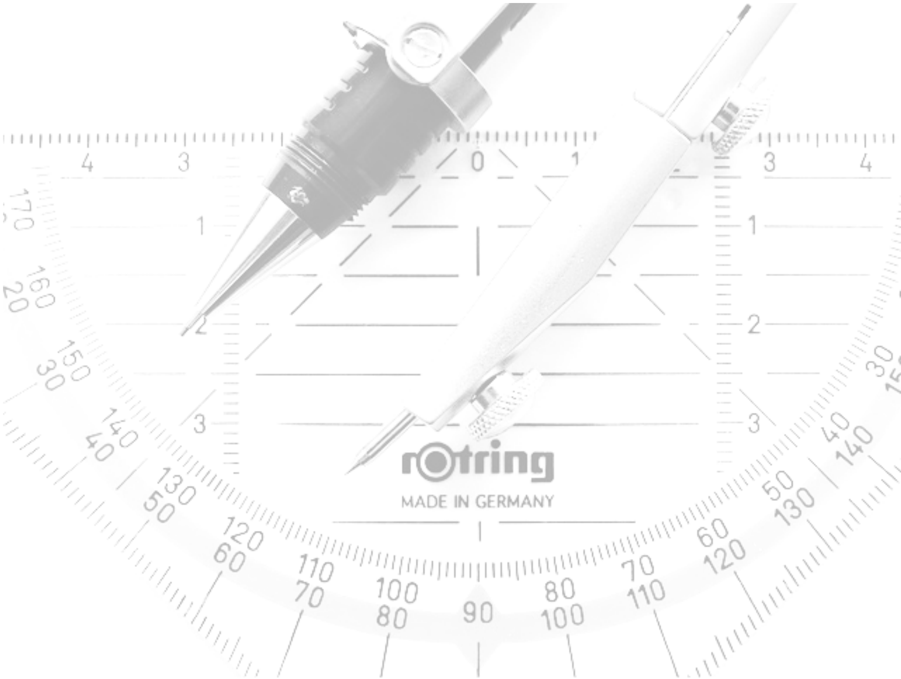
- **Task**
  - Synchronize various video files with unreliable time metadata
  - Use audio-signature to relatively align video files

- **Technology**
  - Audio-fingerprints (chromaprint)
  - Noise invariant



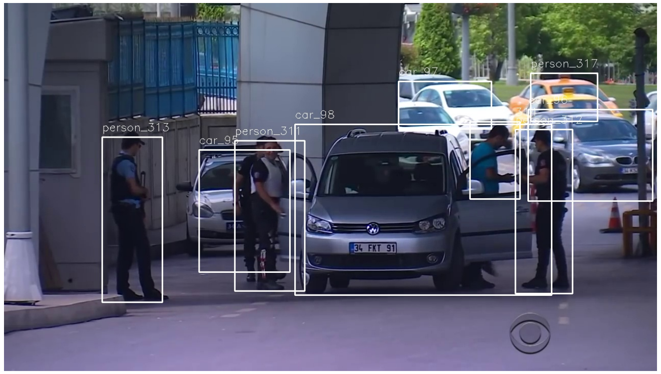
VISUAL ANALYTICS

Analytic Modules



VISUAL CONCEPT DETECTION

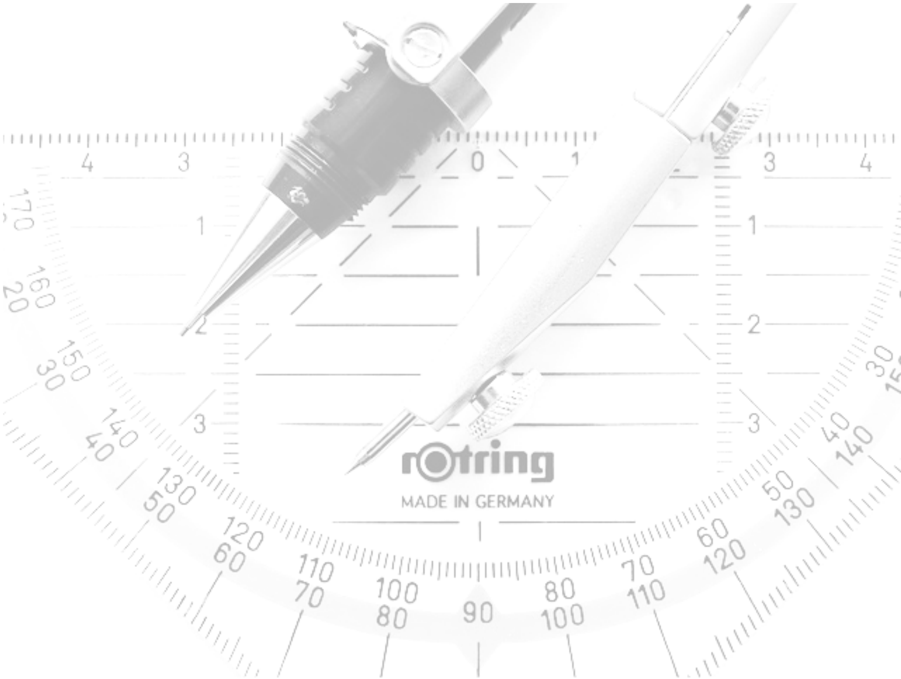
- YOLO
- License Plate detection
- Vehicle Color detection
- **Connected Vision** framework
  - Modular
  - Serviceoriented
  - Distributed
  - Scalable
- Rest Interface



AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

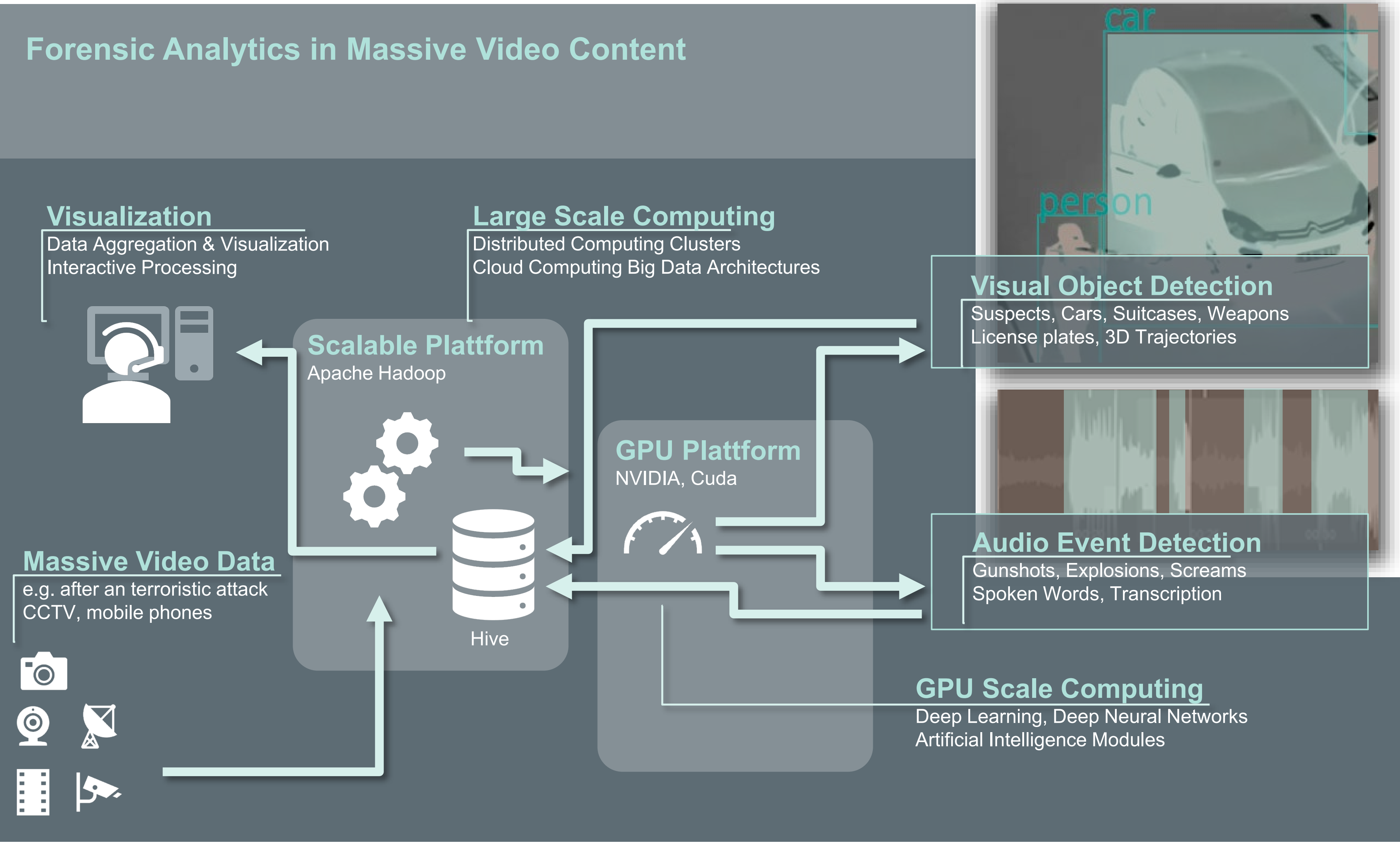
LARGE SCALE PLATFORM

Integration



LARGE SCALE PLATFORM

Forensic Analytics in Massive Video Content



Visualization
Data Aggregation & Visualization
Interactive Processing

Large Scale Computing
Distributed Computing Clusters
Cloud Computing Big Data Architectures

Visual Object Detection
Suspects, Cars, Suitcases, Weapons
License plates, 3D Trajectories

Scalable Plattform
Apache Hadoop

Hive

GPU Plattform
NVIDIA, Cuda

Audio Event Detection
Gunshots, Explosions, Screams
Spoken Words, Transcription

Massive Video Data
e.g. after an terroristic attack
CCTV, mobile phones

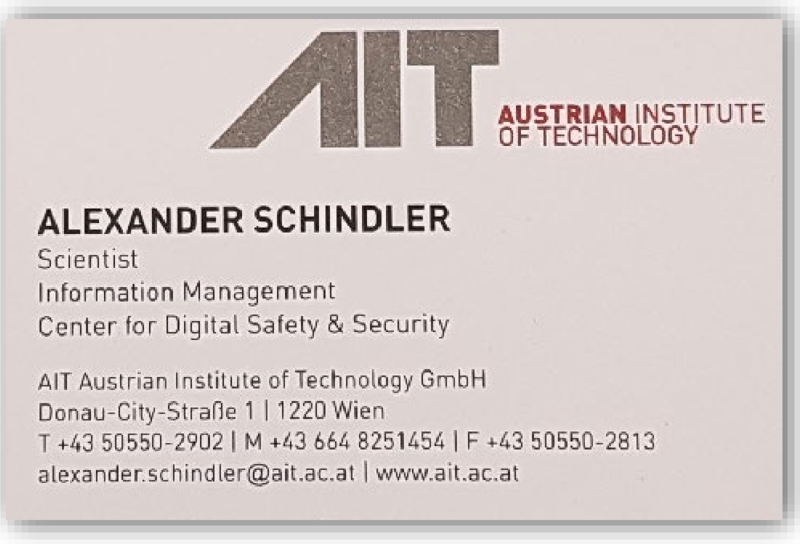
GPU Scale Computing
Deep Learning, Deep Neural Networks
Artificial Intelligence Modules

CONCLUSIONS

- Audio Modules facilitate a rapid start into an investigation
  - Audio Event Detection => Filter
  - Audio Simlarity => Search
- Visual Modules facilitate a broad search for certain objects
  - Follow hints
- Hadoop is not the best choice for multimedia processing
- Integration of audio algorithms into pre-existing visual analytics systems holds pitfalls

THANK YOU

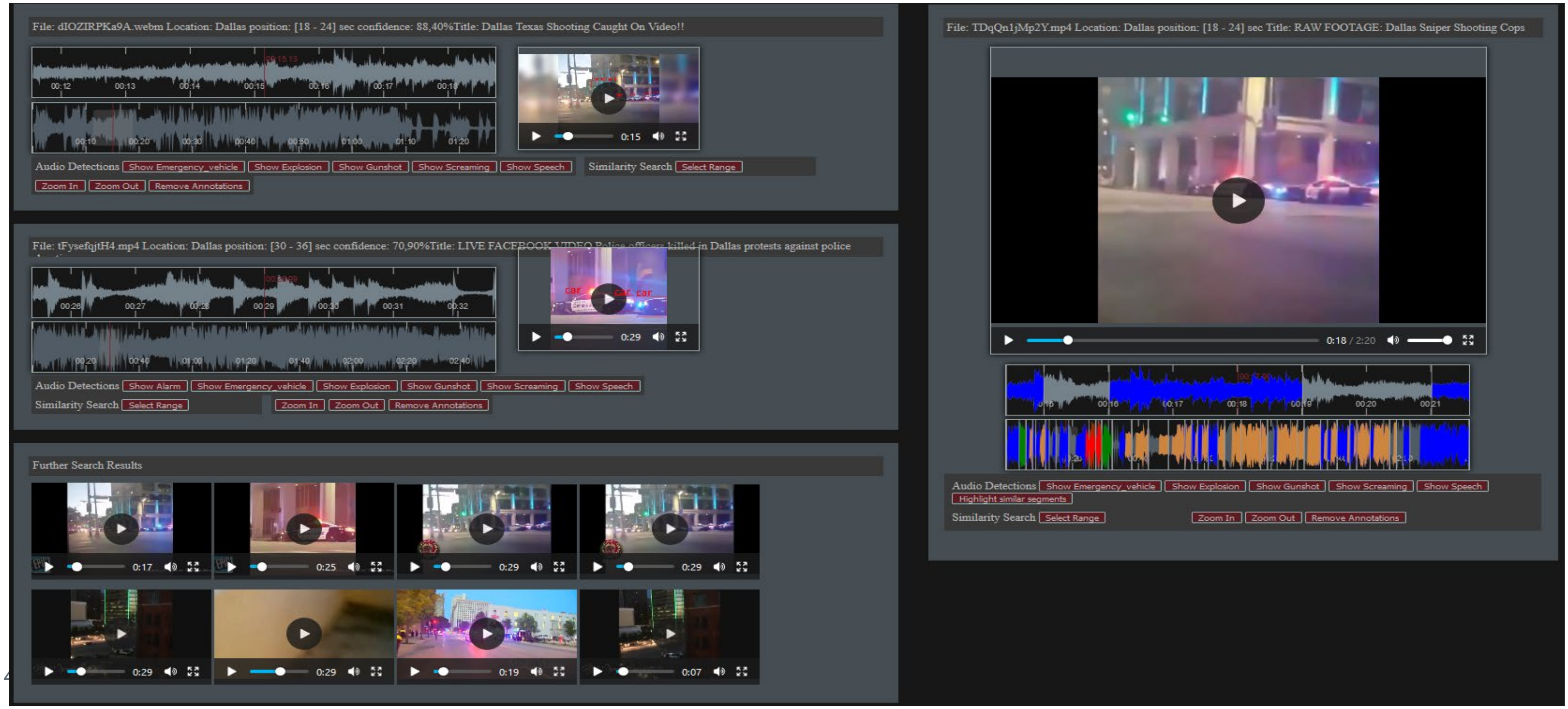
If you are interested
in the demo
ask me in the break!



AUSTRIAN INSTITUTE
OF TECHNOLOGY

ALEXANDER SCHINDLER
Scientist
Information Management
Center for Digital Safety & Security

AIT Austrian Institute of Technology GmbH
Donau-City-Straße 1 | 1220 Wien
T +43 50550-2902 | M +43 664 8251454 | F +43 50550-2813
alexander.schindler@ait.ac.at | www.ait.ac.at



File: dIOZIRPKa9A.webm Location: Dallas position: [18 - 24] sec confidence: 88,40%Title: Dallas Texas Shooting Caught On Video!!

File: tFysefqjtH4.mp4 Location: Dallas position: [30 - 36] sec confidence: 70,90%Title: LIVE FACEBOOK VIDEO Police officers killed in Dallas protests against police

Further Search Results

File: TDqQn1jMp2Y.mp4 Location: Dallas position: [18 - 24] sec Title: RAW FOOTAGE: Dallas Sniper Shooting Cops