



**Real time network,
text, and speaker
analytics for
combating
organized crime**

Presentation of the Autocrime platform and its features

Joint work

presented by

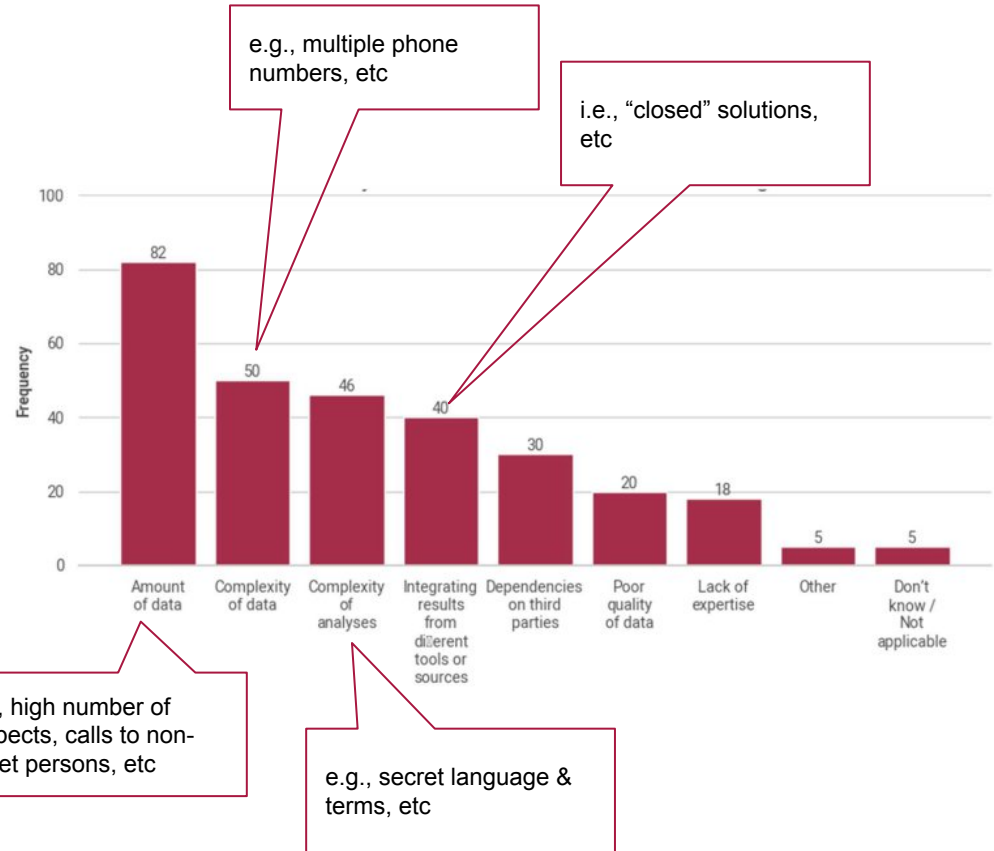
Costas Kalogiros (AEGIS),
Srikanth Madikeri, Jakub Tkaczuk (IDIAP),
Amanda Muscat (IDIAP & Malta Police),
Dawei Zhu (USAAR),
Denis Marraud (AIRBUS),
Zahra Ahmadi (LUH)



This project has received funding from the European Union's Horizon 2020 Work Programme for research and innovation 2018-2020, under grant agreement n°833635

LEAs' pain points

- ROXANNE run a survey on LEAs' requirements
- 121 responses were collected from 40 countries
- The amount of data to be processed and analysed is the main pain point of LEAs

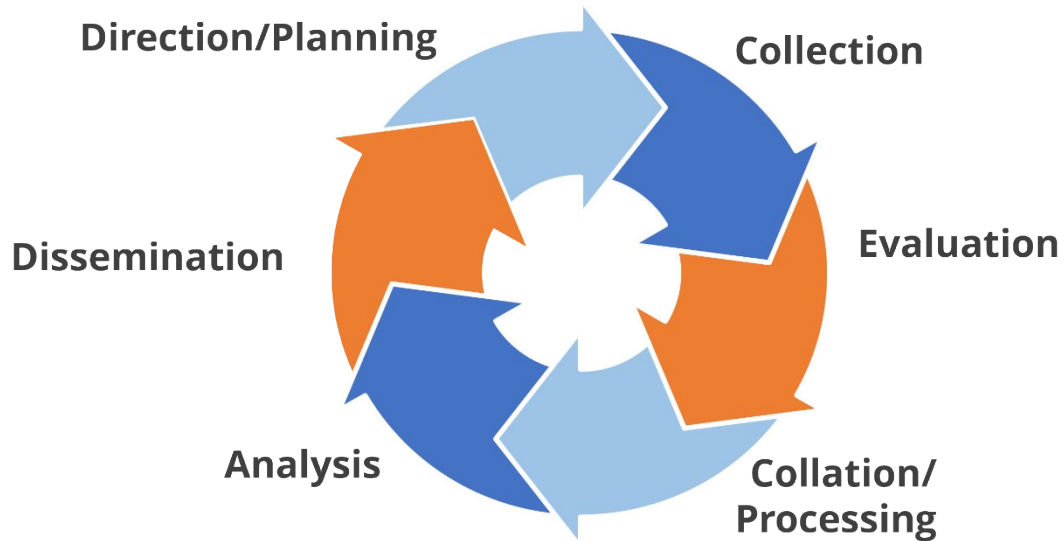


The Autocrime vision

The Autocrime vision is to:

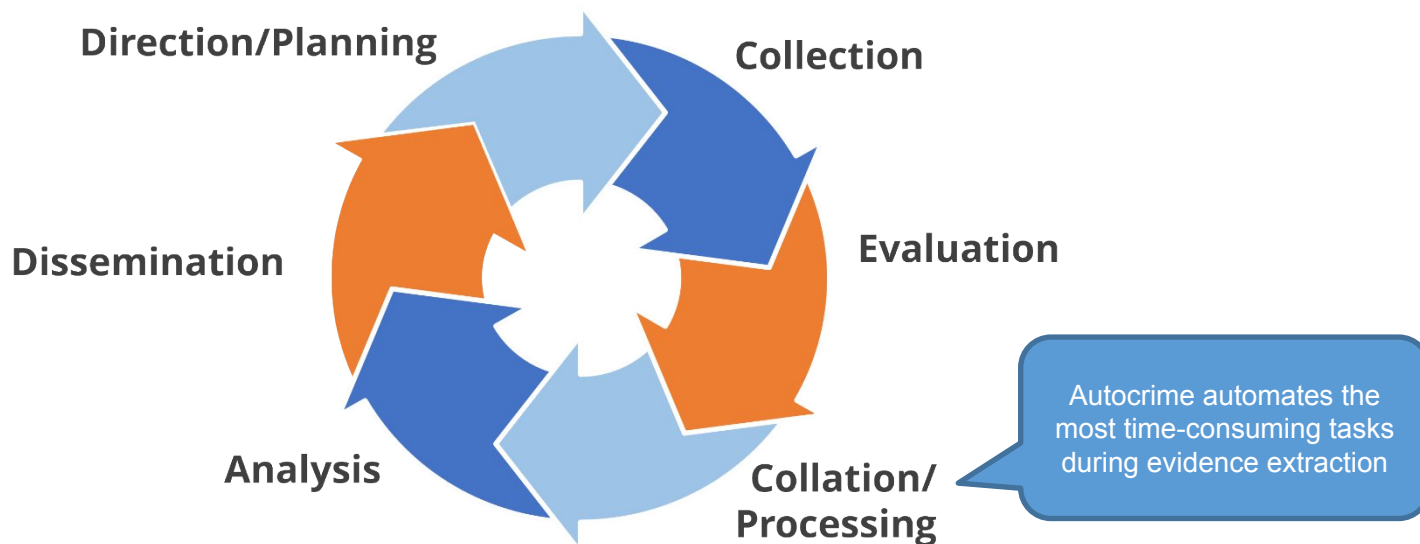
- **automate the time-consuming** activities during crime investigation and
 - provide practitioners with **actionable intelligence and support for exploratory analysis**
- in order to eventually prosecute the offenders.

The intelligence cycle



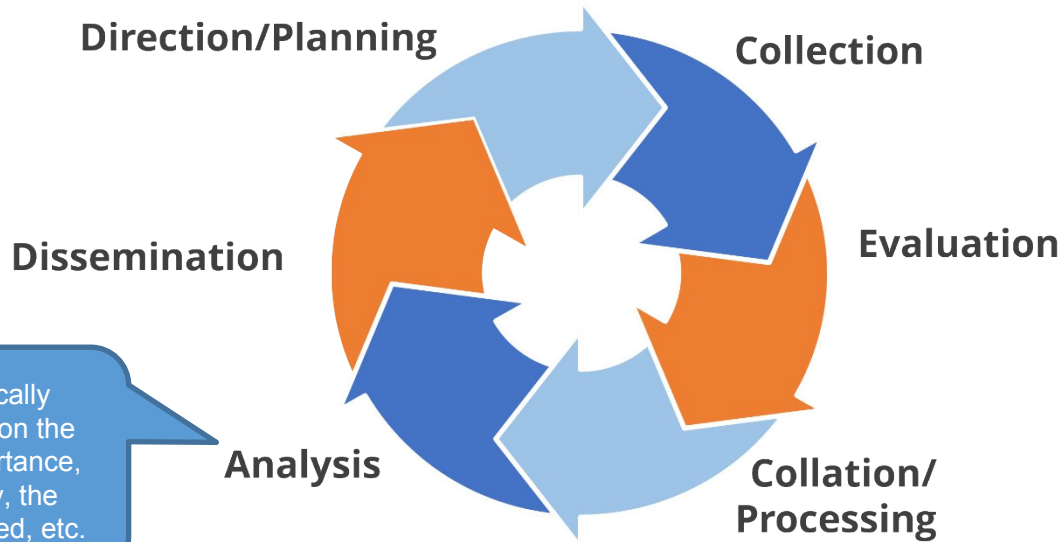
The intelligence cycle (adapted from *)

How Autocrime supports the intelligence cycle (1/3)



The intelligence cycle (adapted from *)

How Autocrime supports the intelligence cycle (2/3)



Autocrime automatically produces information on the entities and their importance, the network topology, the topics that are discussed, etc.

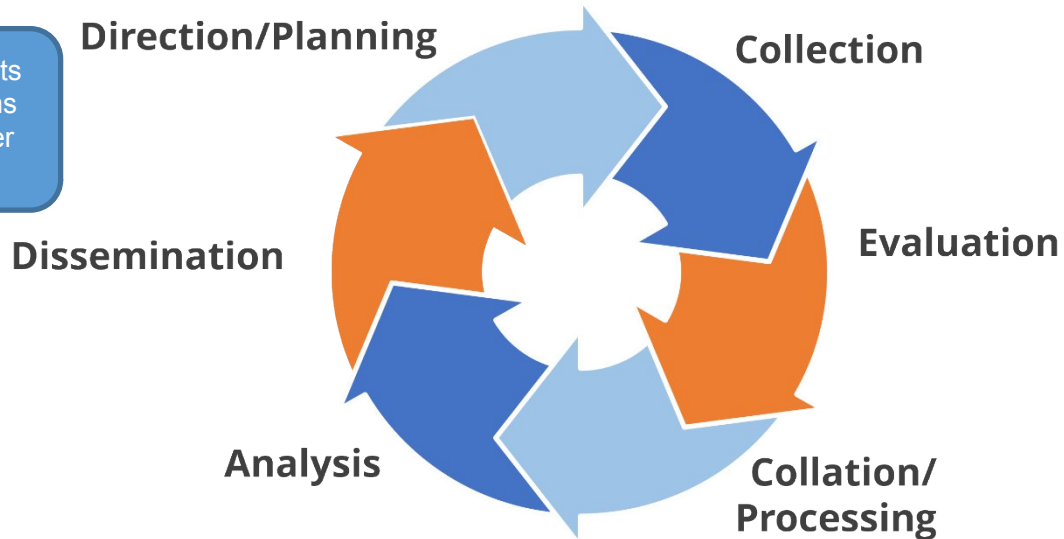
The intelligence cycle (adapted from *)



*"Criminal Intelligence for Front Line Law Enforcement", The United Nations Office on Drugs and Crime, 2010, available online from https://www.unodc.org/documents/organized-crime/Law-Enforcement/Criminal_Intelligence_for_Front_Line_Law_Enforcement.pdf

How Autocrime supports the intelligence cycle (3/3)

Autocrime exports key outputs in other commercial platforms that can be shared with other colleagues

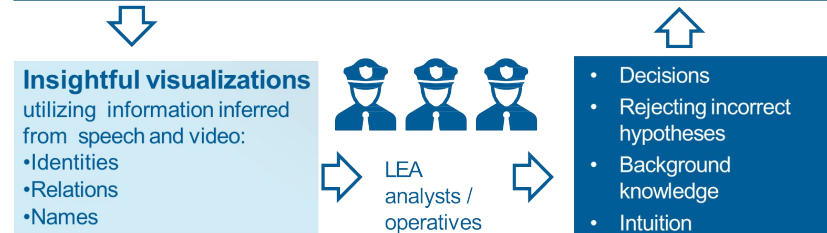
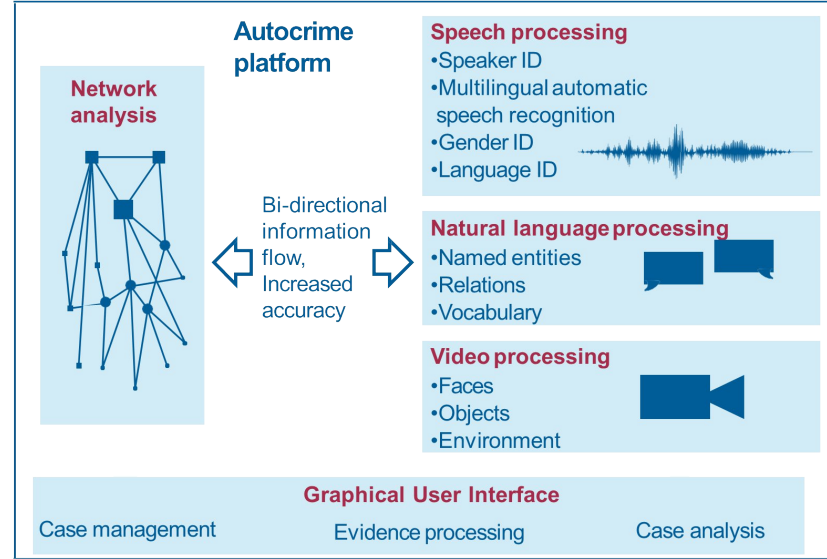


The intelligence cycle (adapted from *)

The Autocrime platform



Activities performed as part of investigations (based on Roxanne survey)



Autocrime vs Competition

	Standard* tools used by LEAs	Autocrime platform
General features		
Ability to combine 2 (or more) cases in the same time	Usually YES	YES
Incorporation of several different features/techniques	Usually YES	YES
Speech processing		
Voice detection	Usually NO	YES
Speaker Clustering and Identification	Usually NO	YES
Gender Identification	Usually NO	YES
Language and Dialect identification	Usually NO	YES
Automatic Speech Recognition	Usually NO	YES
Keyword-Spotting	Usually NO	YES
Natural Language Processing (NLP)		
Named Entity Recognition (location, organizations, ...)	Usually YES	YES
Topic Detection	Usually NO	YES
Identification of unknown 2nd party in calls	Usually NO	YES
Detection of mentions of 3rd parties	Usually NO	YES
Visual analysis		
Facial similarity search	Usually NO	YES
Scene similarity search	Usually NO	YES
Network analysis		
Link prediction	NO	YES
Social influence analysis	Usually NO	YES
Community detection	Usually NO	YES
Outlier detection	NO	YES
Cross network analysis	NO	YES
Exploratory analysis		
Timeline Analysis	Usually YES	YES
Geospatial analysis	Usually NO	YES (Soon)
Details on entities and events	Usually YES	YES
Advanced filters	Usually YES	YES

* We compared general tools used by different LEAs, including i2 Analyst's Notebook, i2 iBase, Tovek, Hansken, SIVE, BATVOX, ACU - EXPERT LAB, ADOBE AUDITION, Watson NLU, Google Cloud Natural, VoiceGain, Speechmatics, Newton Technologies

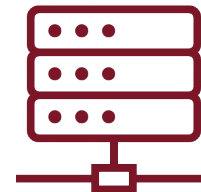
3 modes of operation



Single user
&
low HW specs



Multi-user
&
moderate HW specs
(in progress)



Multi-user
&
high HW specs
(in progress)

15 technologies integrated



Speech

- Voice Activity Detection
- Voiceprint Extraction
- Open-set speaker recognition
- Speech-to-text (English, German)
- Language Identification
- Gender Identification



Text

- Named Entity Recognition
- Topic Detection
- Mention Network
- Relation Extraction



Network Analysis

- Link Prediction
- Community Detection
- Outlier Detection
- Network Merging (manual)



Image & Video

- Scene Analysis
- Face ID

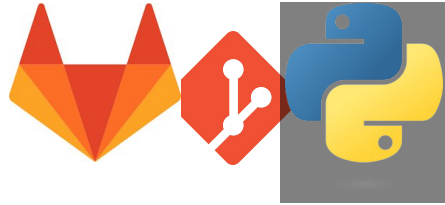
Evaluation of technologies

Technology	Partner(s)	Method	Performance (Other datasets)	Performance (Roxsdv3)
Speaker Diarization	BUT & Phonexia	Energy-based VAD + VBx	DER 5.91% (which data)	14.6% DER
Speaker ID	BUT & Phonexia	ResNet architecture	99.95% speaker accuracy	98.9%
Open set Speaker ID	BUT	Same as SID	90% accuracy	89.7%

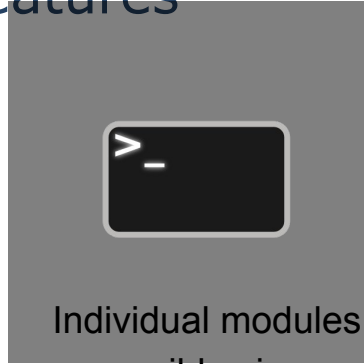
Two-pronged evaluation: modules are evaluated on both **benchmark** datasets and our simulated **Roxsdv3 dataset**

Reproducible evaluations: Code available to generate the numbers claimed

Backend design features



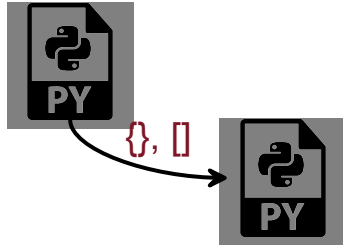
Python codebase
(Pytorch, Hydra, TF,
Pandas, ...)



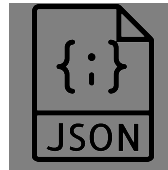
Individual modules
accessible via
command line



Common audio formats
supported (e.g. mono,
stereo, ..)

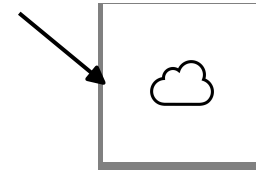


Simple inter-module
communication



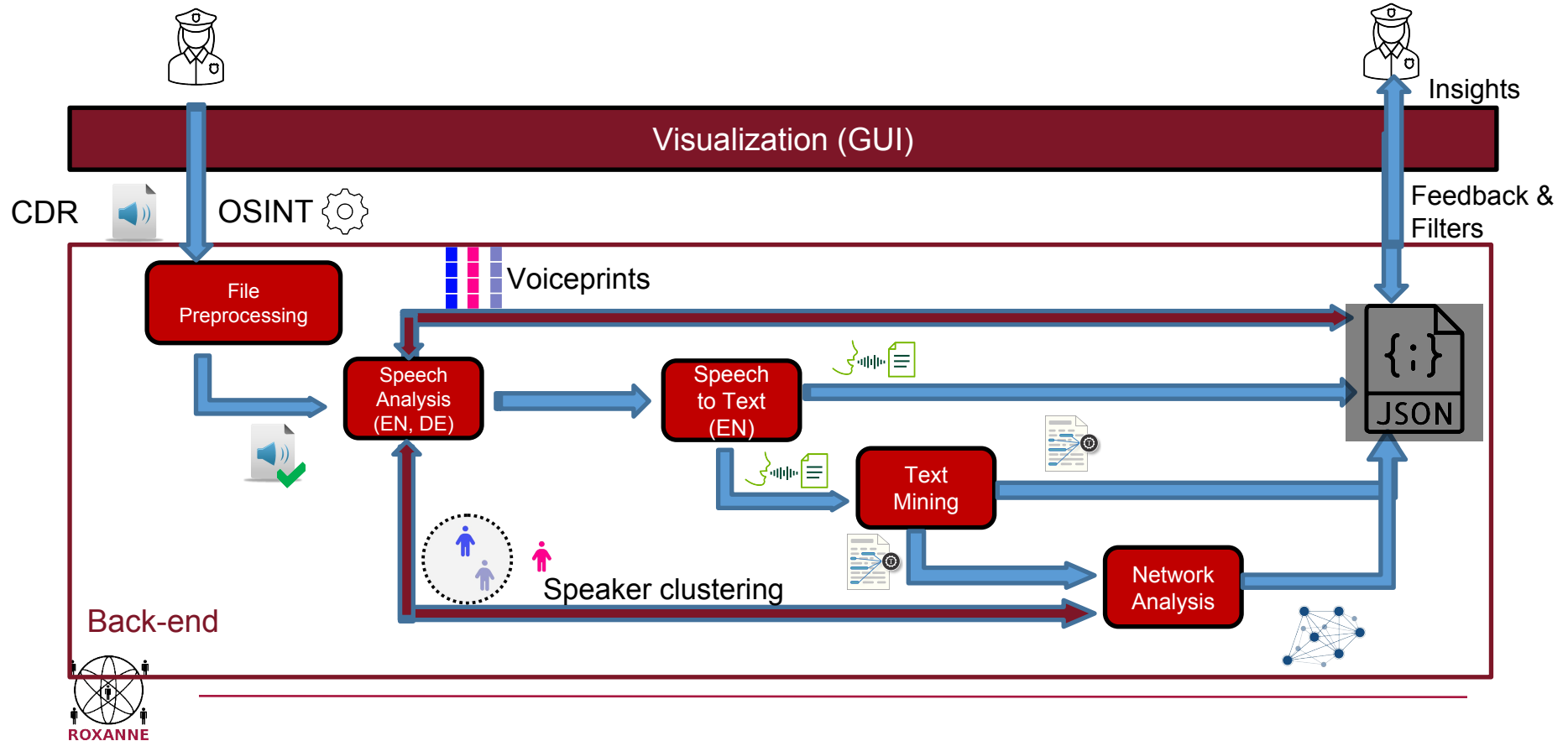
Human-readable
results

Autocrime



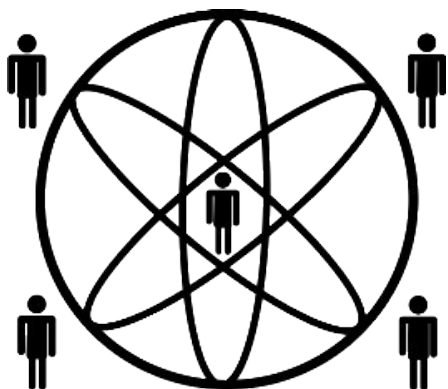
Simple API exposes
core functionalities

The (simplified) analysis workflow



Start the case with a CSV

FROM	TO	DATE	TIME	DURATION	AUDIO	LEFT	Enroll_LEFT	RIGHT	Enroll_RIGHT	Language	Transcription
+420 736 98828	+420 702 90329	16/02/2020	20:06:32	00:00:22	RE54260404ef626cec2d882859a4a930db.wav	Krystof	yes	ru01M_T	no	English	yeah will you come to no i have to clear okay call the guy okay i will call him thanks bye
+420 702 90329	+420 736 98828	06/12/2019	14:32:58	00:01:55	RE5a63bc89e183634e9e8d08aa0c46d1fe.wav	ru01M_T	no	Krystof	no	English	hi where are you im already in hrou its weird the who im going from du what about the d i filled out a ques what happens if h it sucks me so its so we need about i hope that everyt yeah yeah yeah h what are we goin look i have to pai well i can do it in then i have to go ill try it as soon a okay see you
+43 664 24 9095	+420 680 71230	20/12/2019	11:54:18	00:00:18	RE18beae7b701cfe83ae82d0b960fc1dd6.wav	de01M_T	no	Krystof	no	English	hi where are you im on česká so wherere we gc okay i can be the ill be there
+420 699 31024	+420 738 61659	02/01/2020	11:41:08	00:03:50	RE1e2835870f6d735b133550b80afd5f7.wav	cs15M_NT	no	Kristyna	yes	English	hello hello ciao ciao kristýna yes sure go on hc yeah yeah yeah y yeah yeah like like nott yeah yeah yeah y all all the same ye airight listen kika okay sure sure w i know yeah its going to okay



ROXANNE

**Real time network,
text, and speaker
analytics for
combating
organized crime**

Speech Technologies

Speaker Recognition & Diarization



This project has received funding from the European Union's Horizon 2020 Work Programme for research and innovation 2018-2020, under grant agreement n°833635

Speaker recognition

- The task of recognizing the identity of speaker(s) from their voice by means of computer algorithms
- An integral part of the ROXANNE platform because the identity of the speaker(s) in the phone calls are often not known
- Phone numbers can generally not be used to determine the speaker identity because
 - People may change their phone number
 - More than one person may use the same phone number

Speaker recognition tasks

Speaker Diarization (SD)

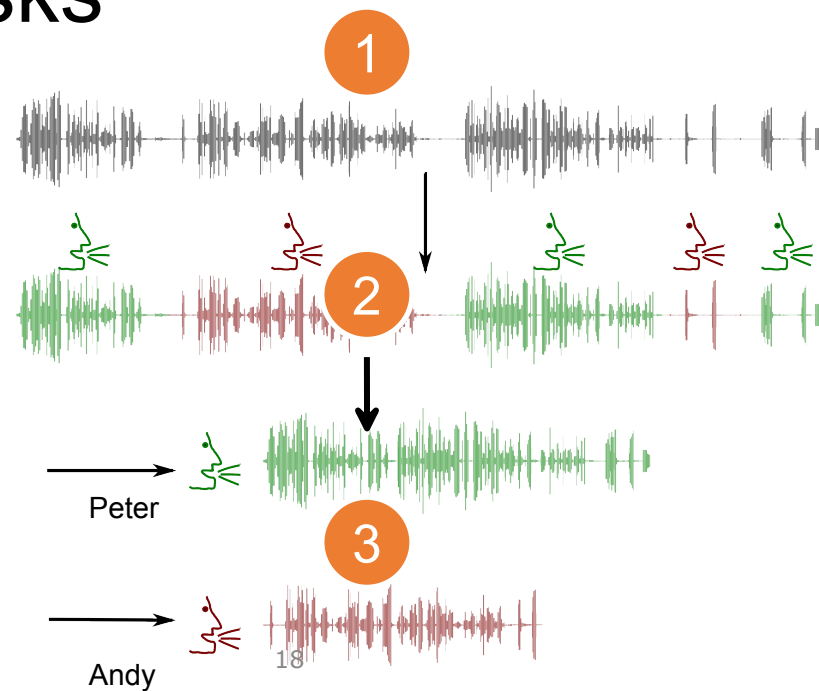
Separate speakers within one mono channel recording

Speaker Clustering

Creates clusters of similar voiceprints

Speaker Identification

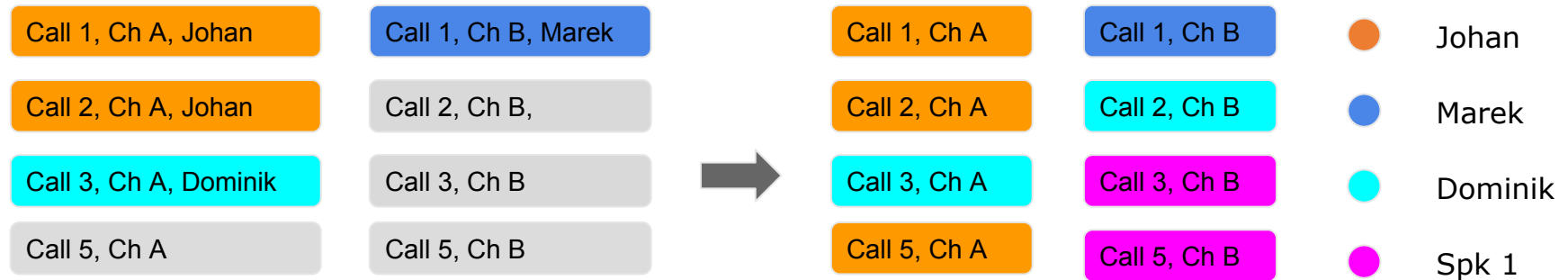
Compare database of enrolled speakers against recordings containing unknown speakers.



Speaker Diarization and recognition

ROXANNE generic scenario: Speaker clustering with enrollment

- Intercepted telephone recordings
- Known IDs (names) for some recordings
- One speaker ID for each recording



- Different side (A,B) in one call cannot be the same speaker
- Data with same ID must be grouped together
- Data with different ID must not be grouped together

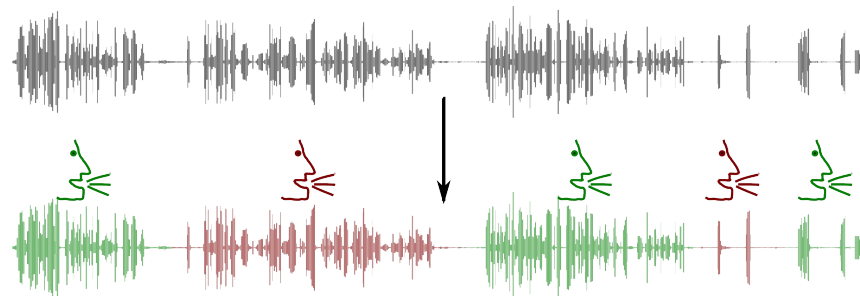
Speaker recognition results

- ROXSDv3: 481 calls - mostly stereo
- 13 enrolled speakers
- Speaker clustering
 - All speakers that were not enrolled are being placed in clusters
 - **Accuracy 89.7%**
 - Same results with and without enrollment of the 13 speakers
- Closed set identification (13 speakers, ~460 recordings)
 - **Accuracy: 98.9%**

20

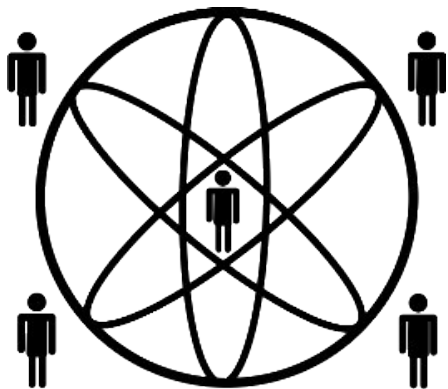
Diarization

- Segment and detect a recording into speaker regions
- Needed if
 - input data is mono
 - more than one speaker per channel, e.g. if phone is handed over
- We can constrain max and min expected speakers



Speaker clustering results

- **Diarization error rate 14.6%**
 - Considers all the following errors
 - Detecting speech where there is no speech
 - Fail to detect speech
 - Assigns speech to the wrong speakers
- For better intuition, see how it affects the application
 - Clustering accuracy if stereo (diarization not needed): 89.7%
 - **Clustering accuracy if mono** (diarization needed): **84.1%**



ROXANNE

Real time network, text,
and speaker analytics
for combating organized
crime

ASR

Multilingual Automatic Speech Recognition

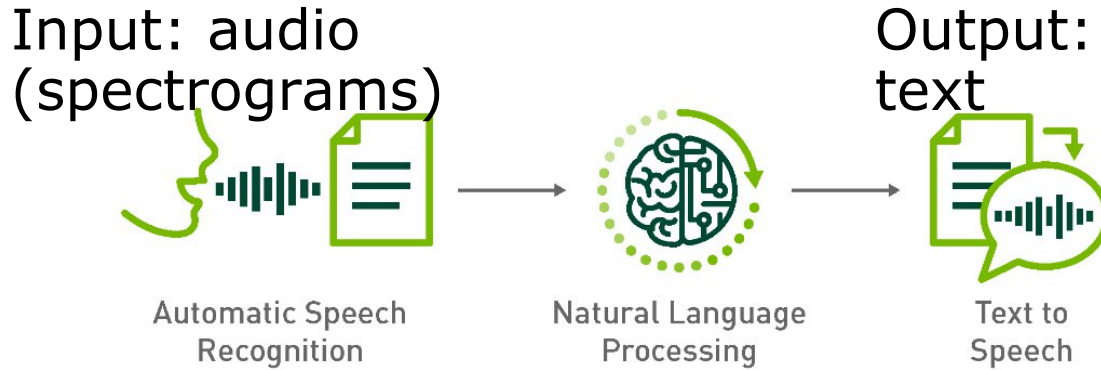
Jakub TKACZUK (Idiap)
Srikanth MADIKERI (Idiap)
Petr MOTLICEK (Idiap)
Ericc DIKICI (Hensoldt)



This project has received funding from the European Union's Horizon 2020 Work Programme for research and innovation 2018-2020, under grant agreement n°833635

Automatic Speech Recognition (ASR)

speech to text



Source: NVIDIA

Speech recognition technology converts spoken language (audio signal) into written text using neural networks with some sort of memory

Why ASR?

How the ASR task can be done another way?

Listening and transcribing the audio
manually...

Transcription can be done
automatically!

This process can also be long
but does not require human attention

Speech Recognition in Autocrime

model	training data	system requirements	languages
more accurate	56 languages (including low resource languages - few seconds of data only)	resource consuming	English German
faster	~ 1000 h of telephone calls	runs on any machine	English

Evaluation Results

evaluation case	speech time	Word Error Rate
English	5 hours	28.5%
German	4 hours	35.1%

English Hands-on case ASR **run time** (i7 CPU, 16GB RAM):

- More accurate ASR decodes 28 min of speech in 16.2 min
- Faster ASR decodes 28 min of speech in 3.8 min

Boosting the low probability words

ASR output is based on probabilities of character/word appearance in training data. User can specify words which they expect to find in audio:

- names
- places
- generic words used for by criminals

Why is it useful?

The investigator can listen to a few calls and catch meaningful words which are not correctly transcribed by ASR and add them manually.

Results

Two ASR engines work together and provide transcripts for English and German languages **based on CSV file inputs**

The results contain speaker ID outputs and ASR outputs

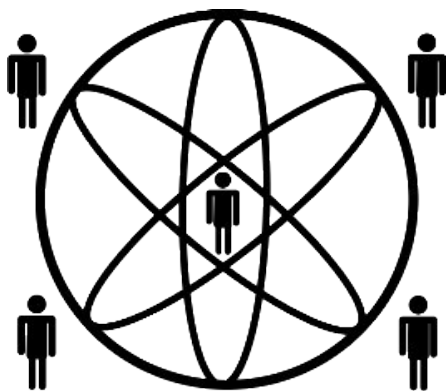
0.630000	1.825000	B 21:german	ja
1.970000	3.825000	A 38:german	hallo wie gehts dir
3.880000	7.865000	B 21:german	ja es geht so ich zieh gerade ist viel arbeit
8.030000	15.755000	A 38:german	okay würdest du zum karlsplatz kommen wenn du fertig bist du hugo hat dort ein sicher tschau
15.770000	21.385000	B 21:german	oho naja ich kanns versuchen aber ich kann dich versprechen ich bin relativ müde
21.720000	27.315000	A 38:german	du könntest auch diese drei bücher die ich bei dir gelassen habe mitnehmen
27.220677	29.155677	B 21:german	welche drei bücher
30.440000	38.915000	A 38:german	die müssen irgendwo liegen ich weiß nicht vielleicht wann ist dann nimm nimm hat alle
39.170000	41.815000	B 21:german	was du meinst diese kleinen
42.400000	50.455000	A 38:german	nein nein nein die anderen ich meine weißt eh myry mary poppins
50.830000	53.515000	B 21:german	ja aber ich habs nicht viel meine süße
54.520000	57.365000	A 38:german	hast du sie jemandem anderen ausgeborgt
57.610000	59.995000	B 21:german	nein die habe ich niemanden gegeben

Voice
Activity
Detection

Speaker
Identification

Automatic Speech
Recognition





ROXANNE

Real time network, text,
and speaker analytics
for combating organized
crime

NLP

Named-Entity Recognition,
Mention Disambiguation,
Relation Extraction,
Topic Detection



This project has received funding from the European Union's Horizon 2020 Work Programme for research and innovation 2018-2020, under grant agreement n°833635

Named-Entity Recognition (NER)

Automatically detecting entities in text data.

- **Direct the users' attention** to informative text pieces
- **Speed-up** reading comprehension during investigation

The model supports detecting:

- **Person Names**
- **Location**
- **Time and Date**



NER - Architecture

🦋 **SOTA Architecture** - Powerful Pre-trained Language Model **RoBERTa-Large**.

- Deep Neural Network with **354** Mio parameters
- Pre-trained on **160 GB** text data
- Fast during inference: **~0.4 s** to process the transcripts in an entire phone call

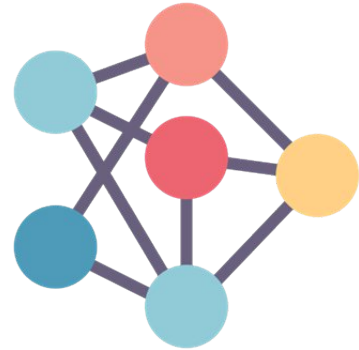
🔍 **Optimized for ROXANNE Cases:**

- **Challenge 1:** Transcripts in phone calls are not formal text

Solution: Model is optimized to perform robustly on informal text

- **Challenge 2:** ASR transcripts do not contain punctuations and casing

Solution: Model is optimized to perform consistently



NER - Evaluation

Evaluation case	ROXSDv3 English Audio	ROXHOOD Text Messages	ROXHOOD English Videos
Content	164 phone calls 4856 utterances 1278 entities	299 posts 110 entities	23 videos 107 utterances 91 entities
Entities	487 person entities 301 location entities 309 time entities	21 person entities 49 location entities 40 time entities	44 person entities 26 location entities 21 time entities
Performance	F1-Score: 82.60%	F1-Score: 77.76%	F1-Score: 92.27%

Mention Disambiguation

Automatically resolving the person names mentioned in the phone call.

- **Party**: either the caller or receiver.
- **Third Party**: other names, e.g., friends or both parties.

Providing additional information for network analysis.



Mention Disambiguation - Architecture

Hybrid Model

- Deep Neural Networks (DNNs) + Rule-based Systems

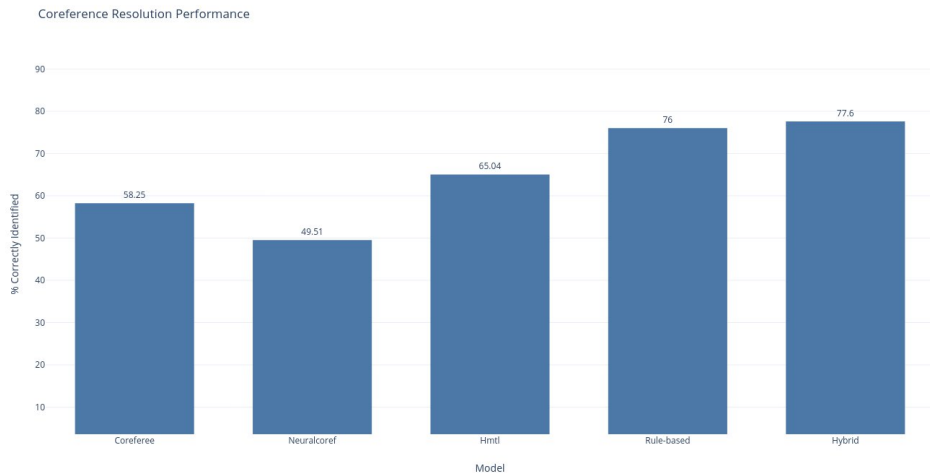
SOTA DNN

- Hierarchical Multi-Tasking LSTM-based Model (HMTL)



Flexibility

- We allow to disable DNNs to **speed-up** processing - trade-off between performance and speed

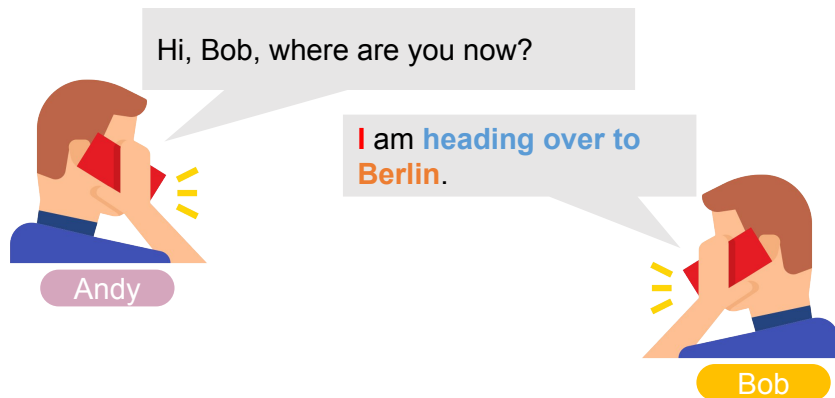


Mention Disambiguation - Evaluation

Evaluation Case	ROXSDv3 English audio	ROXHOOD English videos
Content	164 phone calls 4856 utterances 487 mentioned names	23 videos 107 utterances 44 mentioned names
Entity Distribution	330 parties 157 third parties	25 parties 19 third parties
Performance	Accuracy: 74.05%	Accuracy: 90.90%

Relation Extraction

- Automatically extracting relations from the sentence
- Two Relations are supported
 - **Go**: <subject, object, go>
 - **In**: <subject, object, in>
- Flexible design, easy to support new relation types

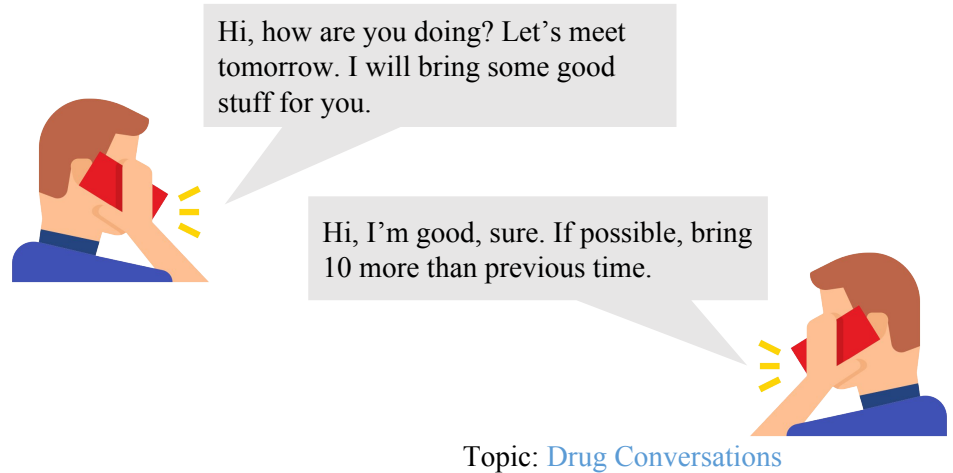


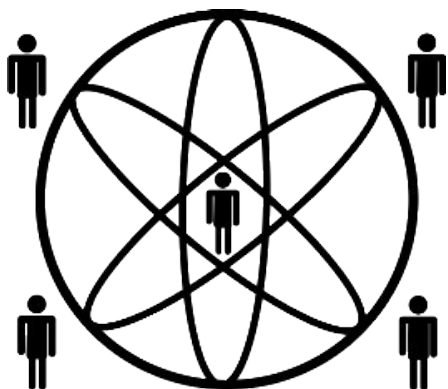
Relation Extraction - Evaluation

Evaluation case	ROXSDv3 English	ROXHOOD Text messages	ROXHOOD English videos
Content	164 calls, 4856 utterances	348 English posts	23 videos, 107 utterances
Accuracy	62.95% F1-Score	78.40 % F1-Score	90 % F1-Score

Topic Detection

- Automatically generates the topic for each phone call
- The aim is to get the gist of the call for investigators to identify if the call belongs to
 - Drug
 - Work
 - Family-Friend
 - Meeting
 - Money
 - Others





ROXANNE

Real time network, text,
and speaker analytics
for combating organized
crime

Video

Face & Scene similarity matching



This project has received funding from the European Union's Horizon 2020 Work Programme for research and innovation 2018-2020, under grant agreement n°833635

Video processing technologies

Objectives:

Leverage images and videos potentially available from several sources:

- seized phone or computer
- surveillance
- internet...

To **enrich speaker network** with additional edges / nodes derived from image or video content

Considered video technologies:

Faces:

- detection, clustering, cluster summarization, similarity matching

Scenes (whole image):

- clustering, cluster summarization, similarity matching

Face detection & matching

Face processing pipeline:

For each ingested image or video:

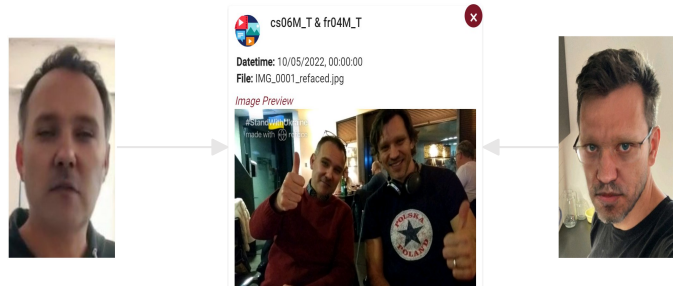
- Detect sufficiently resolved faces
- Characterize each face with a signature
- Cluster faces at video level (to gather all face observations of a same person)
- Summarize each cluster with 5

representative but different pictures of each person



Evaluations

- 54 enrolled pseudonymized faces (using SimSwap face swapping framework) and present in 98 images or videos
- 96/98 of these faces were indeed found (detection rate = 98%)



Scene characterization & matching

Scene / object processing pipeline:

For each ingested image or sampled frame in a video:

- Characterize the whole image with a signature
- Cluster scenes at video level (to gather all observations of a same location or object)
- Summarize each cluster with 1 **representative** picture of each scene



id_0_repr_0.jpg



id_1_repr_0.jpg

Evaluations

- Enrolment of 12 different indoor / outdoor scenes visible in 150 images or videos
- 105/150 scenes properly recognized (detection rate = 70%)
- 16 scenes wrongly found in some documents (precision = 86.8%)



Video technologies for autocrime

How it is used

Face or scene pictures can be enrolled and associated to speaker nodes

Enrolled faces or scenes are searched in all ingested document (image or video)

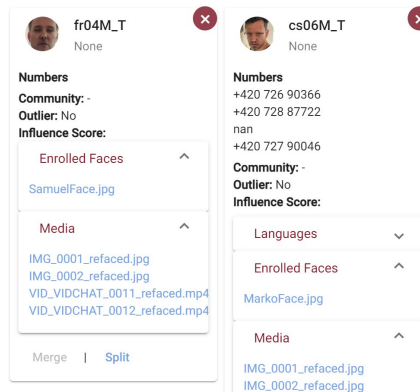
Ingested images or videos can be assigned an owner (speaker node)

Faces or scene matches **create links** between:

- document owners and observed enrolled people or scenes
- between enrolled people when observed or heard in a same video



Result examples



fr04M_T
None

Numbers
Community: -
Outlier: No
Influence Score:

Enrolled Faces
SamuelFace.jpg

Media
IMG_0001_refaced.jpg
IMG_0002_refaced.jpg
VID_VIDCHAT_0011_refaced.mp4
VID_VIDCHAT_0012_refaced.mp4

Merge | Split

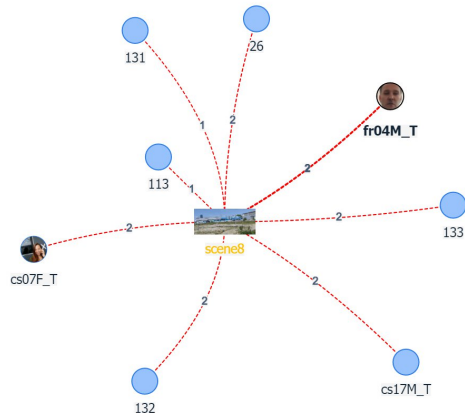
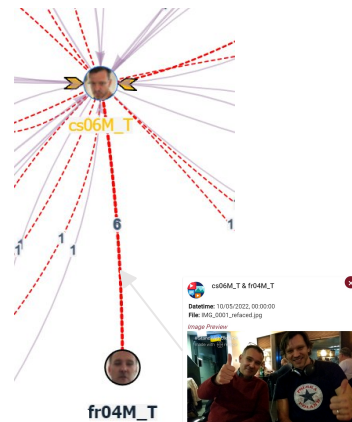
cs06M_T
None

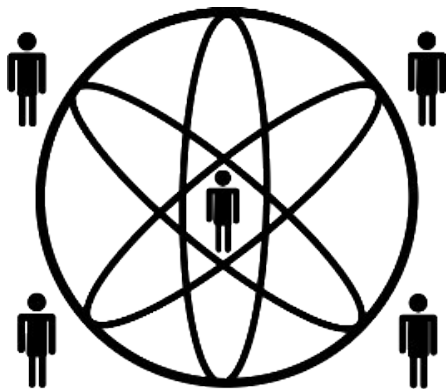
Numbers
+420 726 90366
+420 728 87722
nan
+420 727 90046
Community: -
Outlier: No
Influence Score:

Languages

Enrolled Faces
MarkoFace.jpg

Media
IMG_0001_refaced.jpg
IMG_0002_refaced.jpg





ROXANNE

Real time network, text,
and speaker analytics
for combating organized
crime

Network analysis

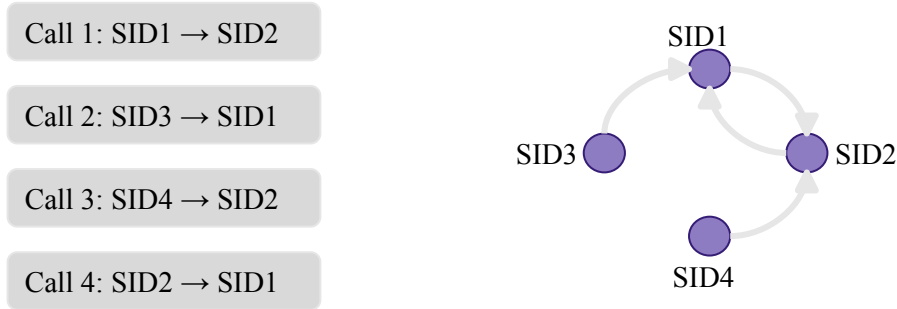
Social influence analysis, Outlier
detection, Community detection, Link
prediction, Cross-network Analysis



This project has received funding from the European Union's Horizon 2020 Work Programme for research and innovation 2018-2020, under grant agreement n°833635

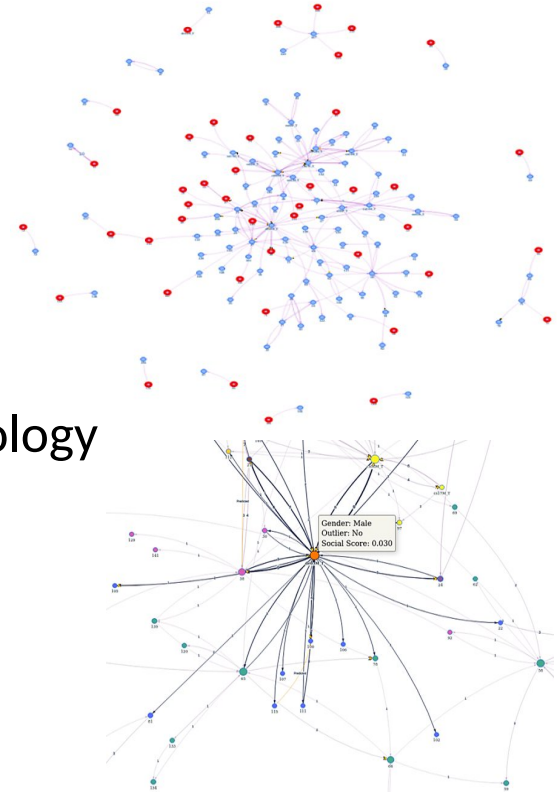
Network Construction

- Input: list of calls with the output of speaker identification module
- Output: network of caller-callees
 - Each node represents a speaker
 - Each edge indicates one call between two speakers



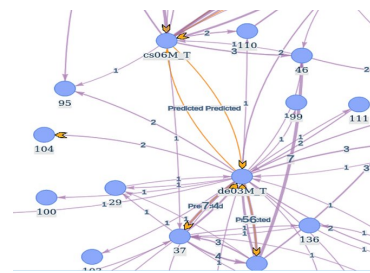
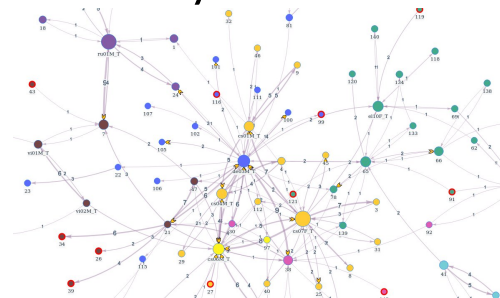
Network Analysis Motivation and Technologies (1)

- Reduce the workload of investigations by
 - Focusing on more influential speakers
 - **Social influence analysis technology**
 - Removing uninteresting speakers
 - **Outlier detection technology**
- Unsupervised methods based on network topology
- 30% network reduction on ROXSDv3



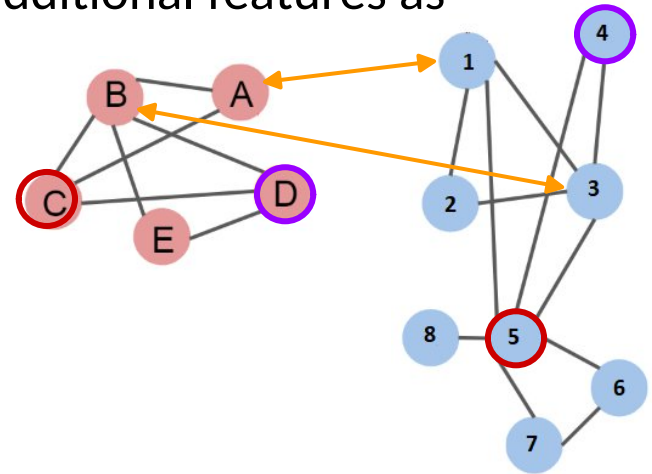
Network Analysis Motivation and Technologies (2)

- Exploitation of information from known and hidden connections by
 - Identification of cohesive groups that are not easy to see by human
 - **Community detection technology**
 - On ROXSDv3, detected communities based on merely network structure are inline with detected languages from ASR
 - Identification of possible missing links and the most likely outgoing calls for each speaker
 - **Link prediction technology**
 - Accuracy of 60% on ROXSDv3
- Unsupervised and based on network topology



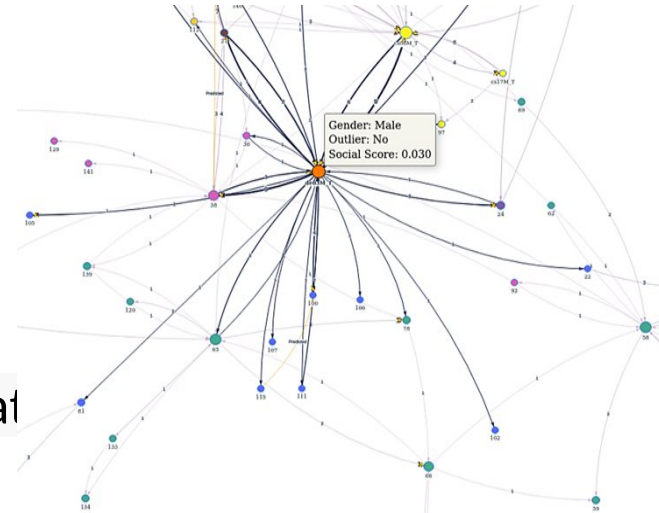
Network Analysis Motivation and Technologies (3)

- Improve identification of persons of interest among other networks
- Find matching nodes across two networks of different modalities
 - **Cross-network analysis technology**
- Can use network structure information and additional features as node embeddings
- Supervised learning
- Accuracy of 75% on ROXSDv3 and ROXHOOD



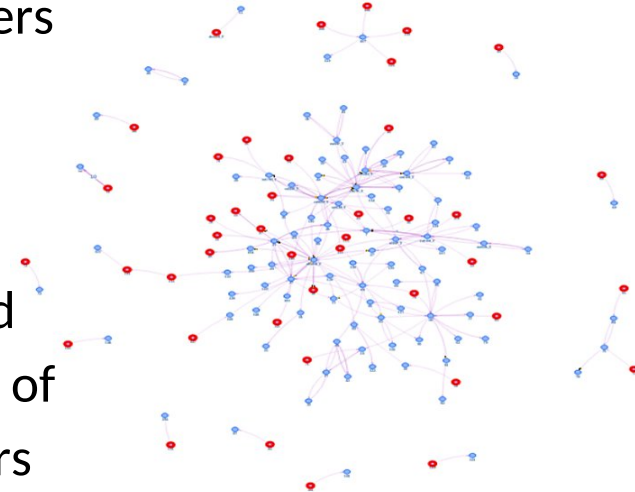
Social Influence Analysis

- Unsupervised learning
- Based on network topology
 - Each node in the network is assigned a score based on its number of incoming links
 - Important speakers whose reach extends beyond their direct connections
 - Probability that a random walker lands on that speaker node after taking some steps
 - Systematic assessment on the resulted scores



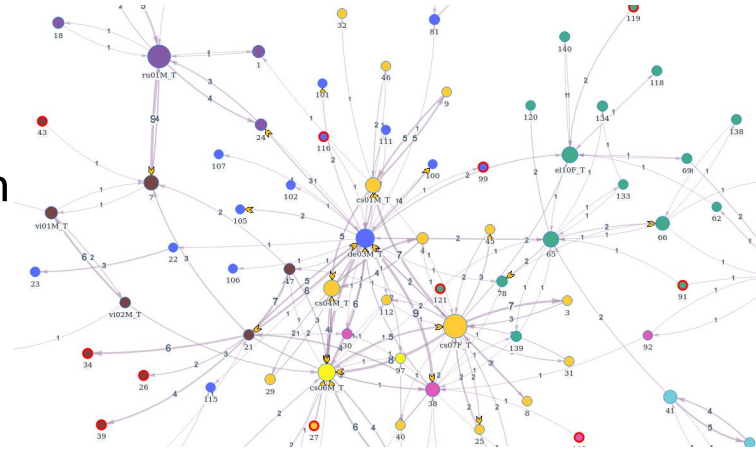
Outlier Detection

- Shrink network by removing uninteresting speakers for investigations
- Unsupervised learning
 - Based on social influence scores
 - Select speakers with influence below a threshold
 - 30% network reduction on ROXSDv3 while none of the removed speakers are among target speakers



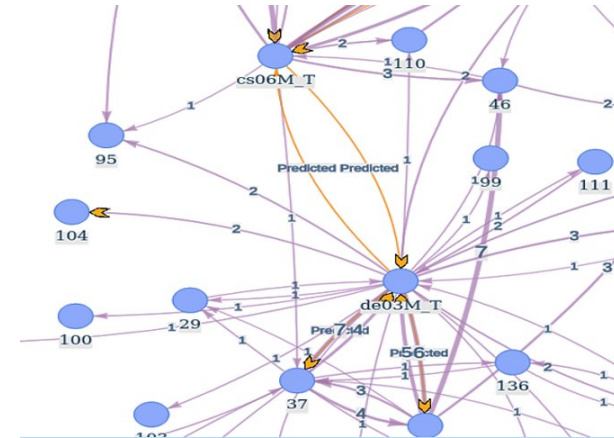
Community Detection

- Unsupervised learning
- Modularity maximization
- Follows the intuition that communities should have more internal interactions/relations than cross-communities
- May identify cohesive groups that are not easy to see by human
- Indicates if a group has only a few bad behaviors or is acting as a ring, which would be indicated by a higher relationship density than average



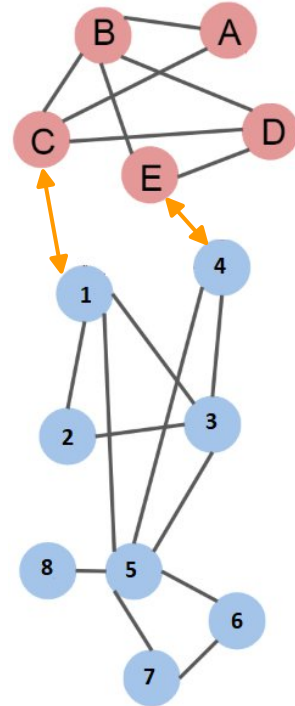
Link Prediction

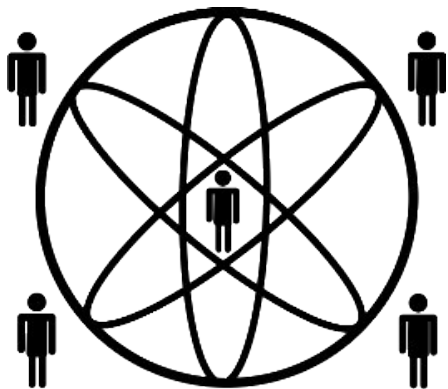
- Can assist in identifying possible missing links
- Predict the most likely outgoing calls for each speaker
- Unsupervised learning
- Based on network topology
 - Measures the proximity between each pair of nodes and returns the top-scored ones
 - The ratio between the two nodes shared interactions divided by their union interactions
 - Accuracy of 60% on ROXSDv3



Cross-network Analysis

- Supervised learning
- Find matching nodes across two networks of different modalities
- Uses the alignment of network structure and additional attributes as node embeddings
 - Tested ROXSD and ROXHOOD based on network structure alone and also embeddings from contents
 - results ...





ROXANNE

Real time network, text,
and speaker analytics
for combating organized
crime

Interpretation



This project has received funding from the European Union's Horizon 2020 Work Programme for research and innovation 2018-2020, under grant agreement n°833635

Case Study on Maltese data

- A total of 1161 audio recordings from prison in Maltese.
- All speech technologies were ran (speaker diarization, speaker clustering and speaker identification).

Malta results	Correct	Incorrect	Accuracy
Without merging	88	93	0.48
With adding the best cluster to each speaker	152	29	0.83
With merging	177	4	0.97

Usefulness for the Malta Police Force

- No other similar tool
- No pre-processing needed
- Timesaving
- Highlights the network topology
- Ability to export results to collate them with other results

Thank you