A Concept for a Comprehensive Understanding of Communication in Mobile Forensics

Jian Xi, Michael Spranger and Dirk Labudde

Jian Xi, University of Applied Sciences, Germany
xi@hs-mittweida.de

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Forensic Science Investigation Lab (FoSIL)

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Agenda

- Introduction of FORMOBILE project
- Evaluation of big data by semantic analysis
  - Problem in data analysis in mobile forensic
  - Proposed concept for understanding the communication data
- Outlook of semantic analysis

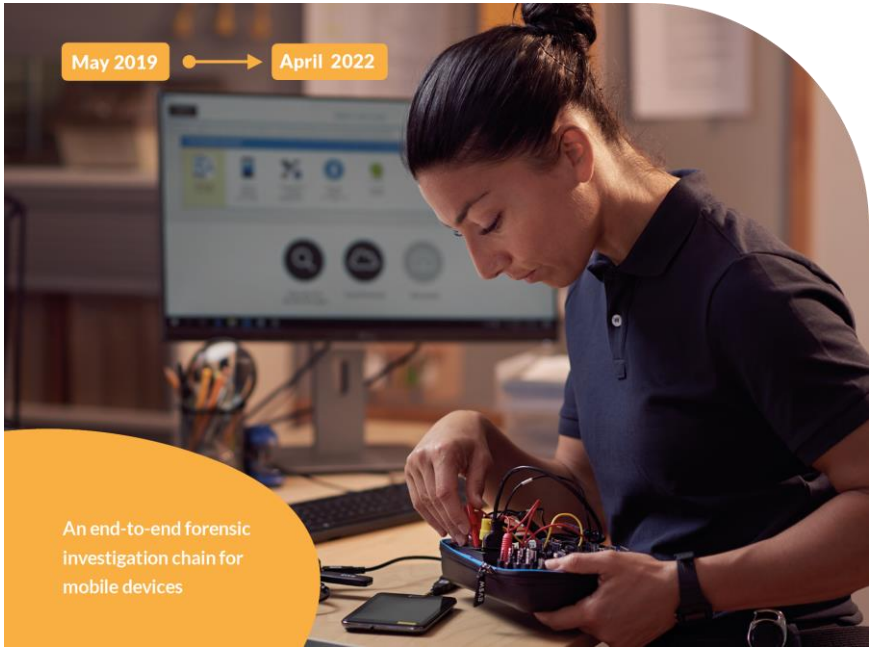
This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

FORMOBILE

HOCHSCHULE MITTWEIDA
University of Applied Sciences

From Mobile Phones to Court

An EU Project to Help Keep Citizens Safe



An EU project aiming to create an **end-to-end mobile forensic investigation chain**, striving to improve digital safety, and security in the EU

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

FORMOBILE

Project Objectives

Tool

- Innovative new tools available to Law Enforcement Agencies (LEA)s allowing for the rapid retrieval, storage and analysis of mobile phone data

Standard

- Define European standard for the forensic investigation of mobile phones

Training

- Training course helps LEAs to effectively use the new tools and follow the standardized procedures

Advisory Boards

Ethics

Security

Scientific

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

WP6.1 Evaluation of big data by semantic analysis

Goal

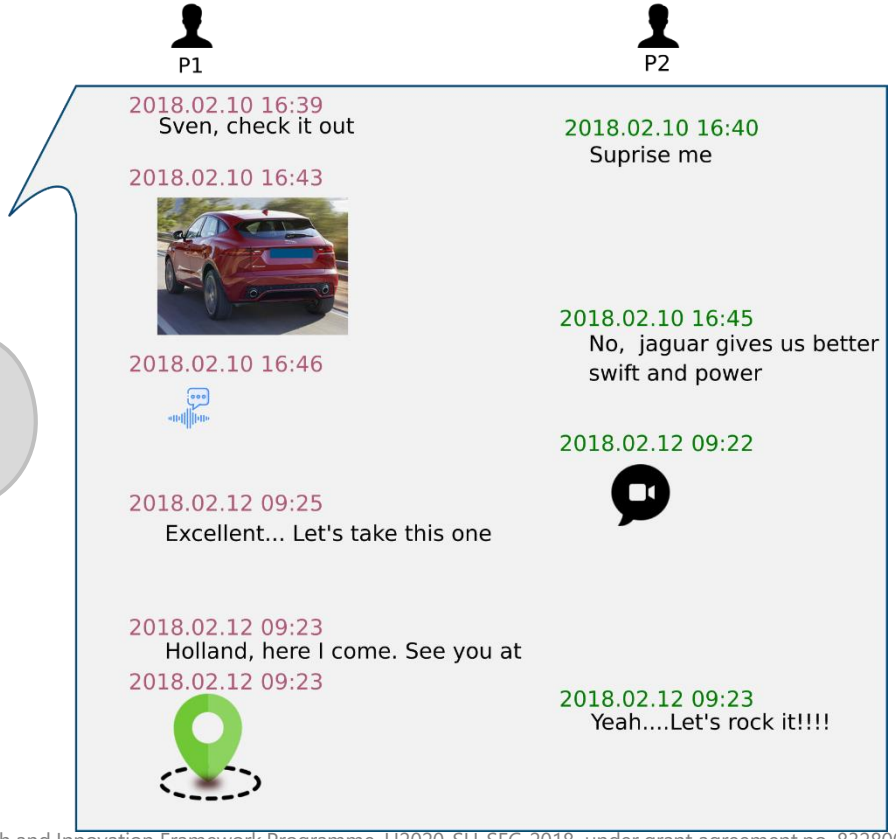
- Analysis and assessment of all communication for the purpose of forensic information retrieval and extraction
- Develop new algorithms to analyze each data type and integrate all retrieved information in one knowledge map as an extension to a common communication network
    - Jointly considering all possible data modalities and communication channels in investigations

Understanding the data properly to support investigation

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Jointly Understanding of Communication in Mobile Forensics

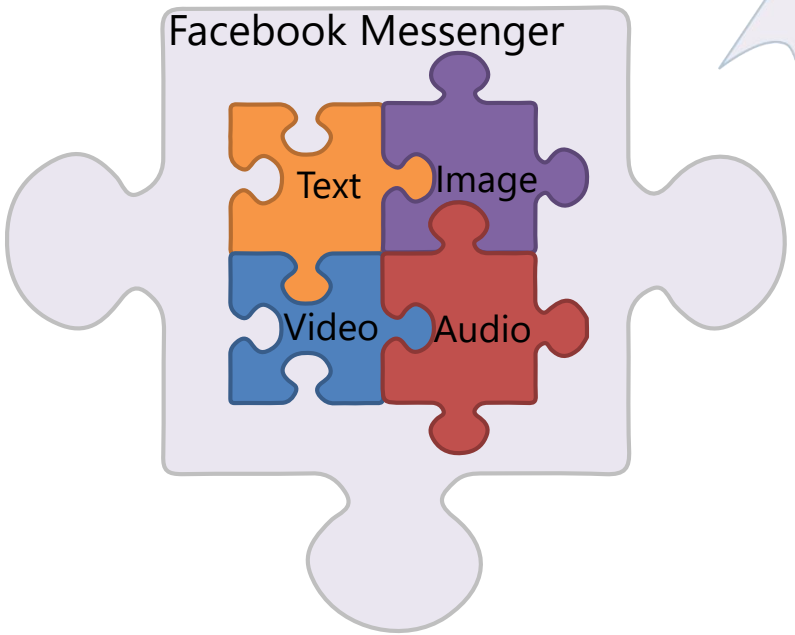
Why is jointly understanding important?

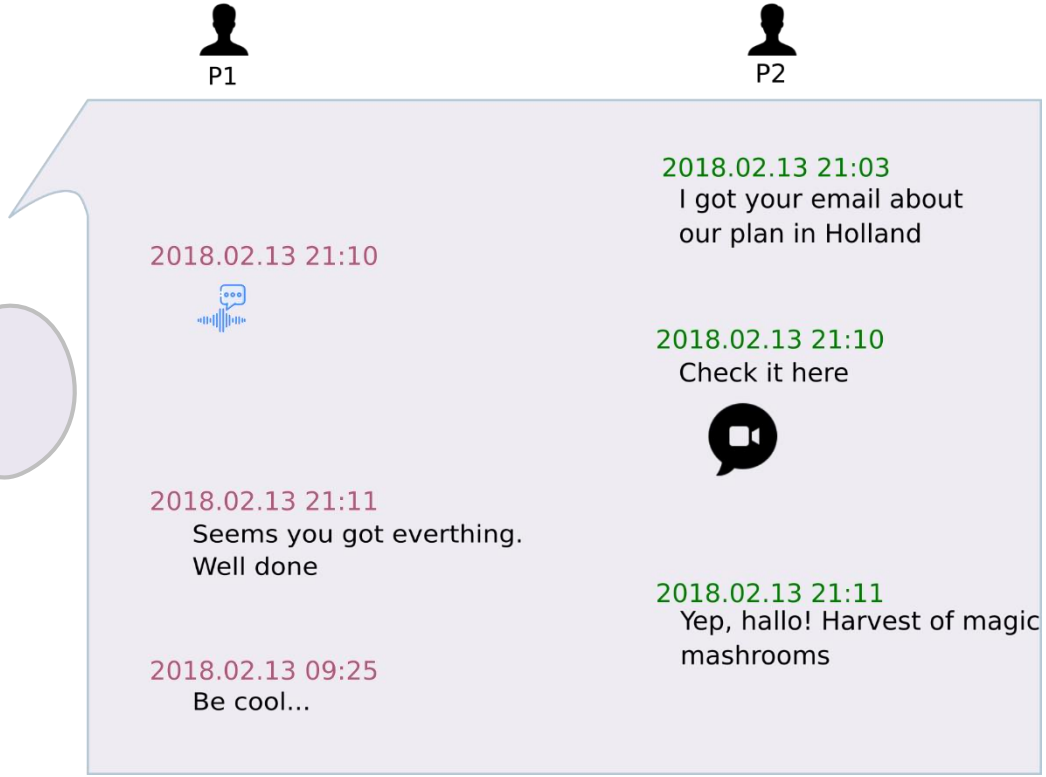


This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Jointly Understanding of Communication in Mobile Forensics

Why is jointly understanding important?

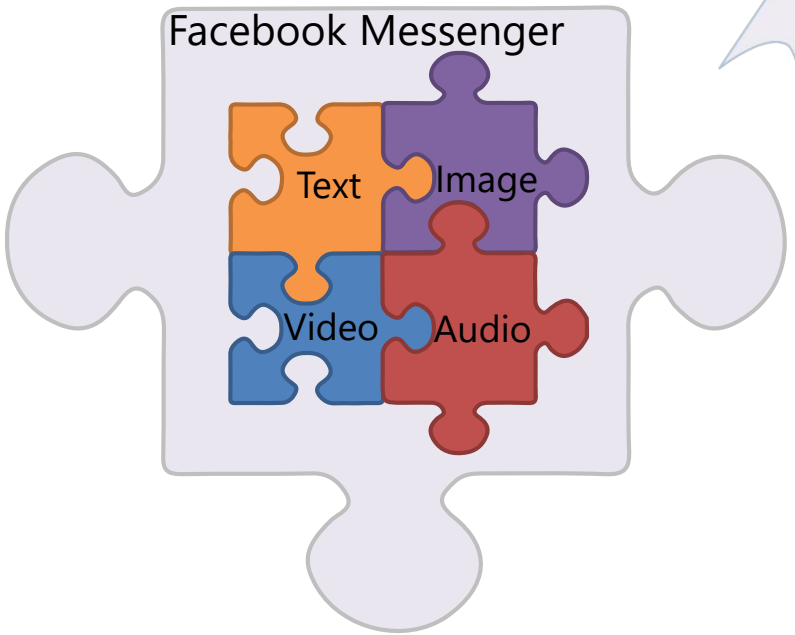


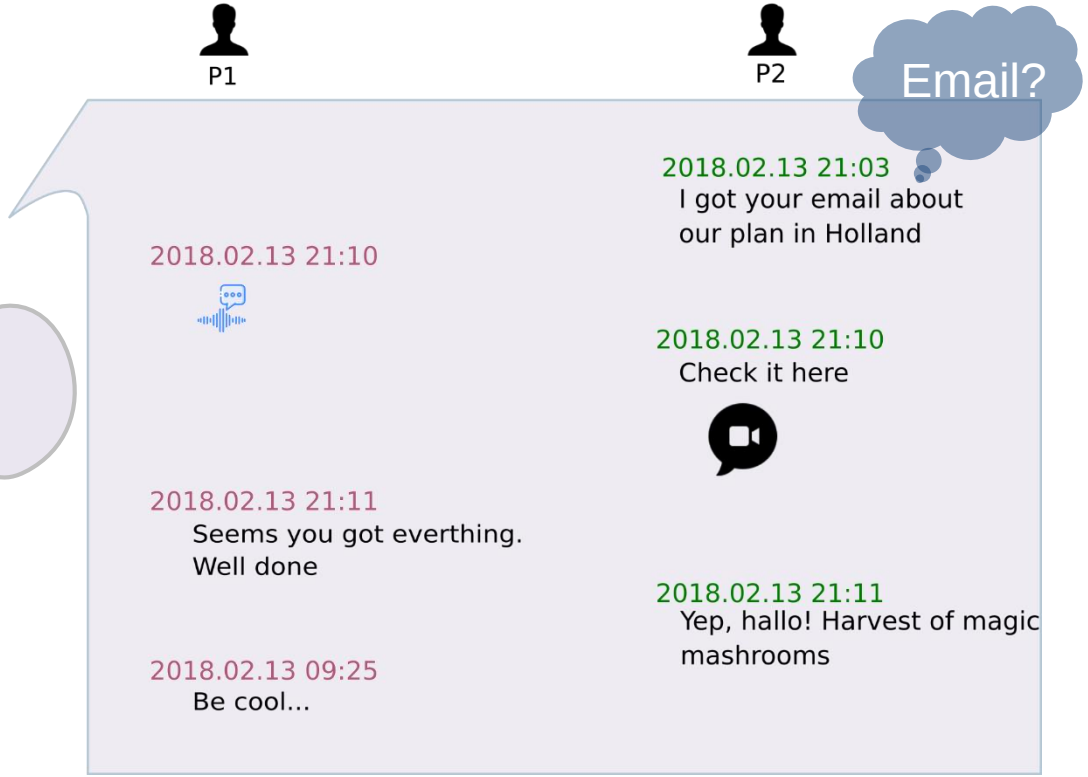


This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Jointly Understanding of Communication in Mobile Forensics

Why is jointly understanding important?

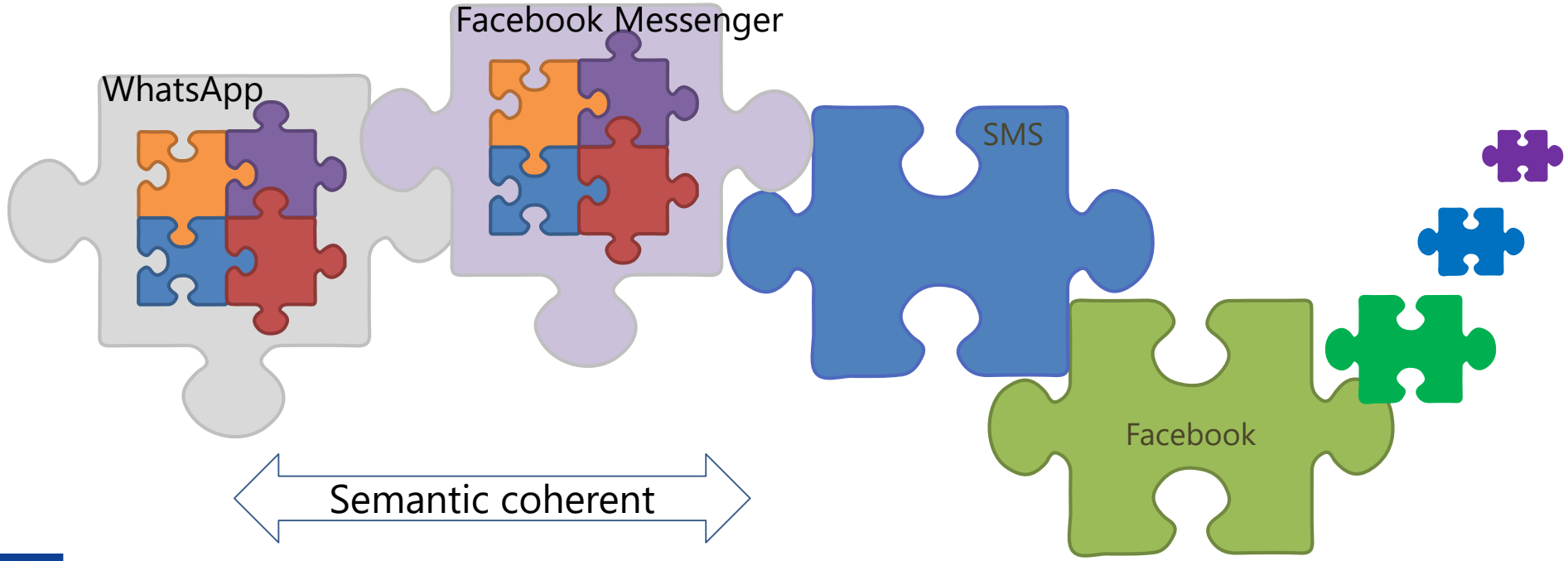




This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Jointly Understanding of Communication in Mobile Forensics

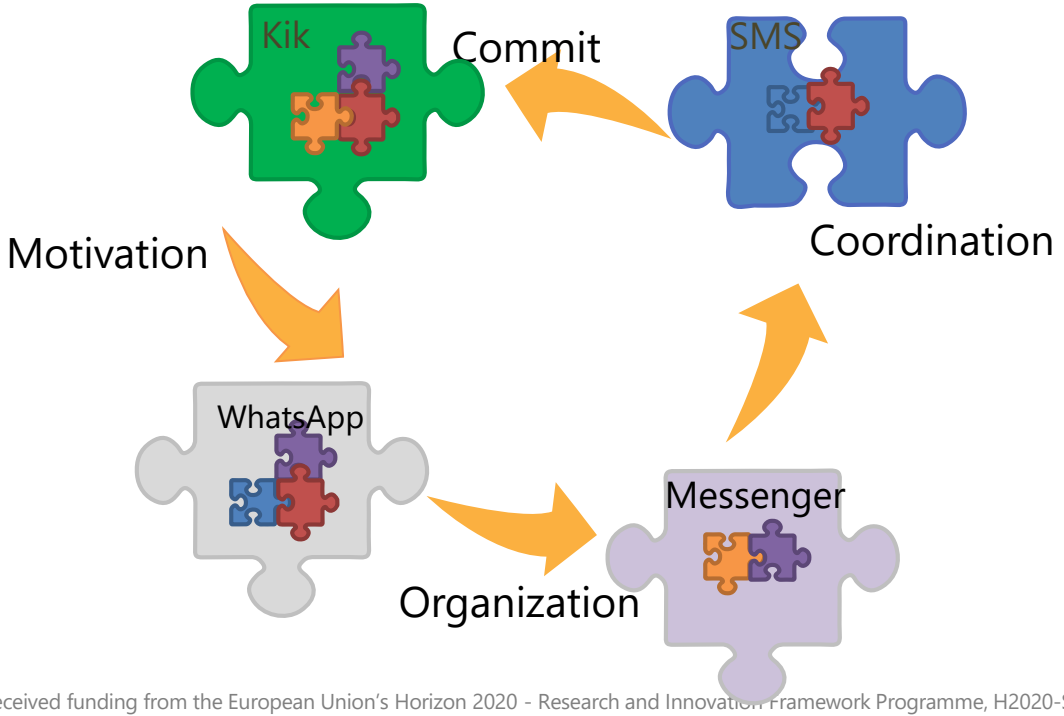
Why is jointly understanding important?



This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Jointly Understanding of Communication in Mobile Forensics

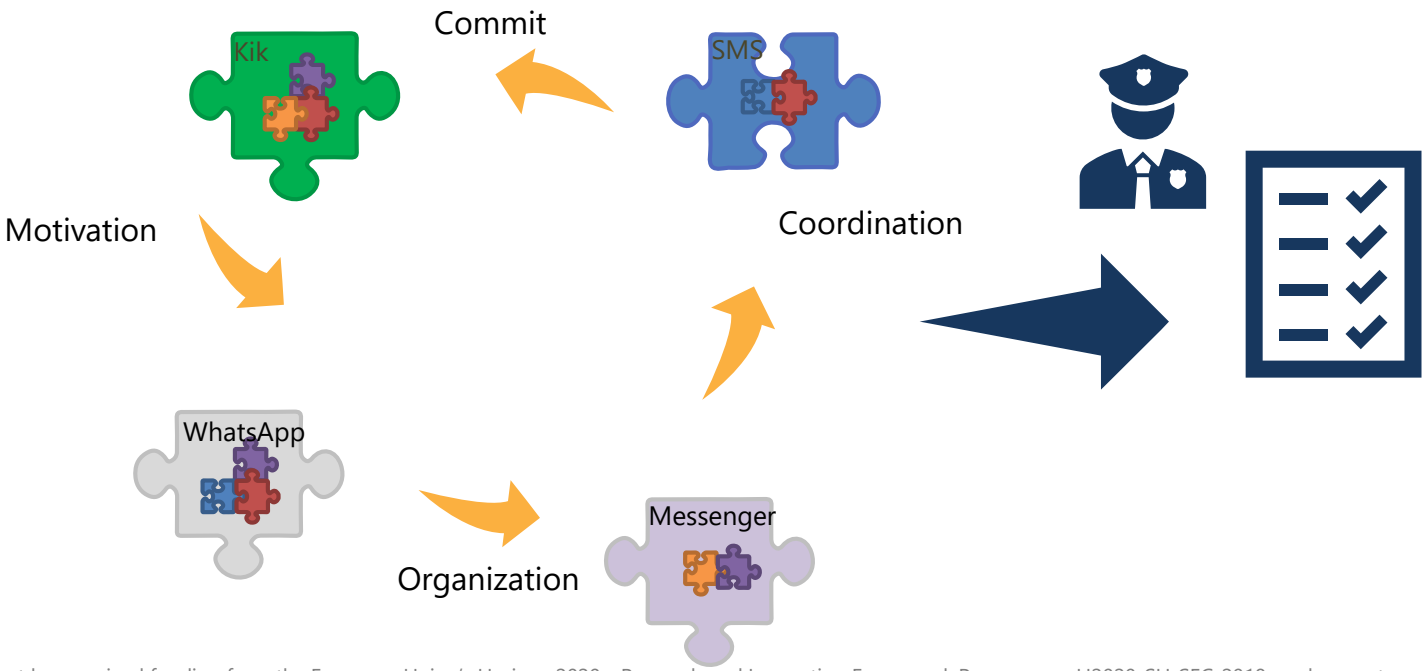
A well-organized criminal offense works



This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Jointly Understanding of Communication in Mobile Forensics

Detect evidence from communication



This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Jointly Understanding of Communication in Mobile Forensics

Problems in analysis of mobile communication data

- Different data types (modalities)
  - Difficulty in semantic alignment, e.g., object in image vs object in text, audio, video mentioned
  - Incomplete contextual information
    - Information loss or misunderstanding
      - Correct understanding data only if all modalities are considered[Hans Bucher]

Jerry, what do you think the idea we discussed yesterday?

Tom, what idea? I was trying to survive from your hunting all day…

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Jointly Understanding of Communication in Mobile Forensics

Problems in analysis of mobile communication data - multimodality

- Misunderstanding or misinterpretation of data if only text is considered[Spranger et al.]

2018.02.10 16:39
Sven, check it out

2018.02.10 16:40
Suprise me

2018.02.10 16:45
No, jaguar gives us better swift and power

2018.02.12 09:25
Excellent... Let's take this one

2018.02.12 09:23
Holland, here I come. See you at

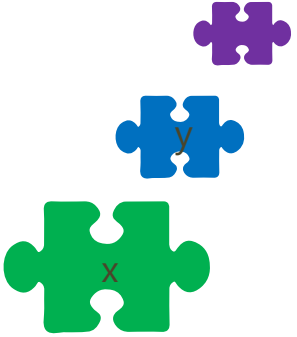
animal or vehicle?

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Jointly Understanding of Communication in Mobile Forensics

Problems in analysis of mobile communication data

- Various communication channels
  - Communication (as discourse) is semantic coherent [Jerry Hobbs]

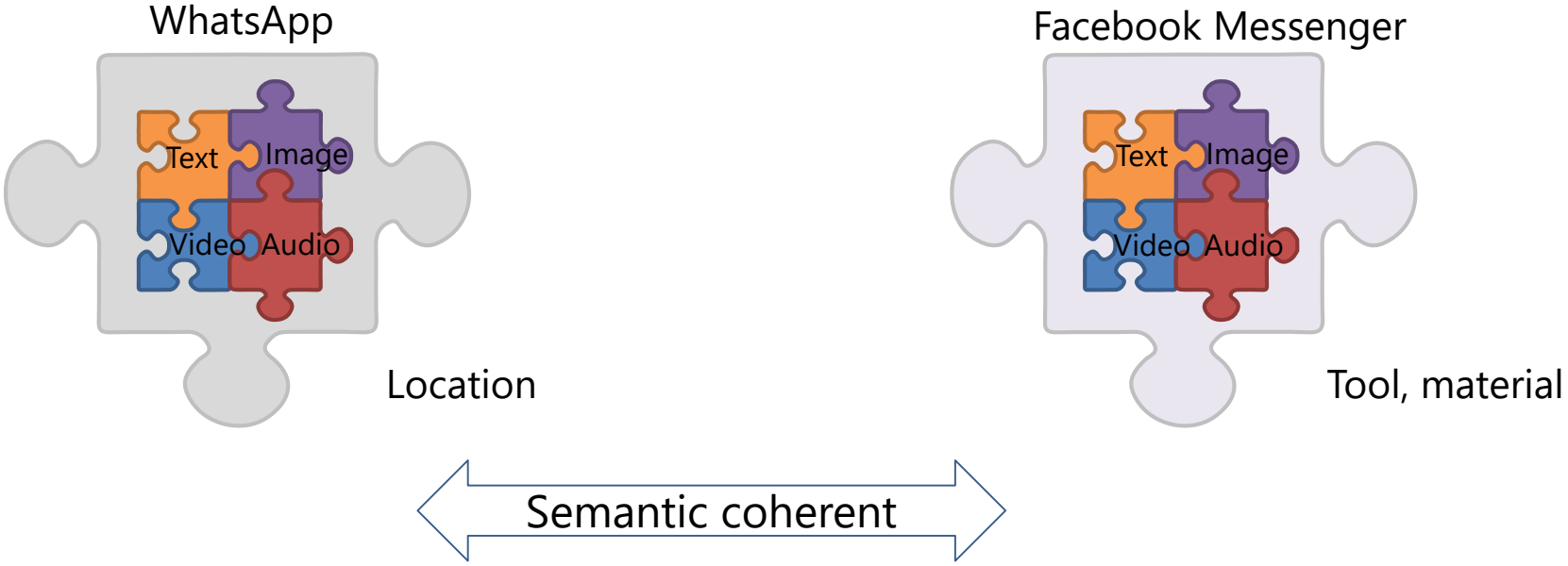


This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Jointly Understanding of Communication in Mobile Forensics

Problems in analysis of mobile communication data - multichannel

- Segmented semantic information

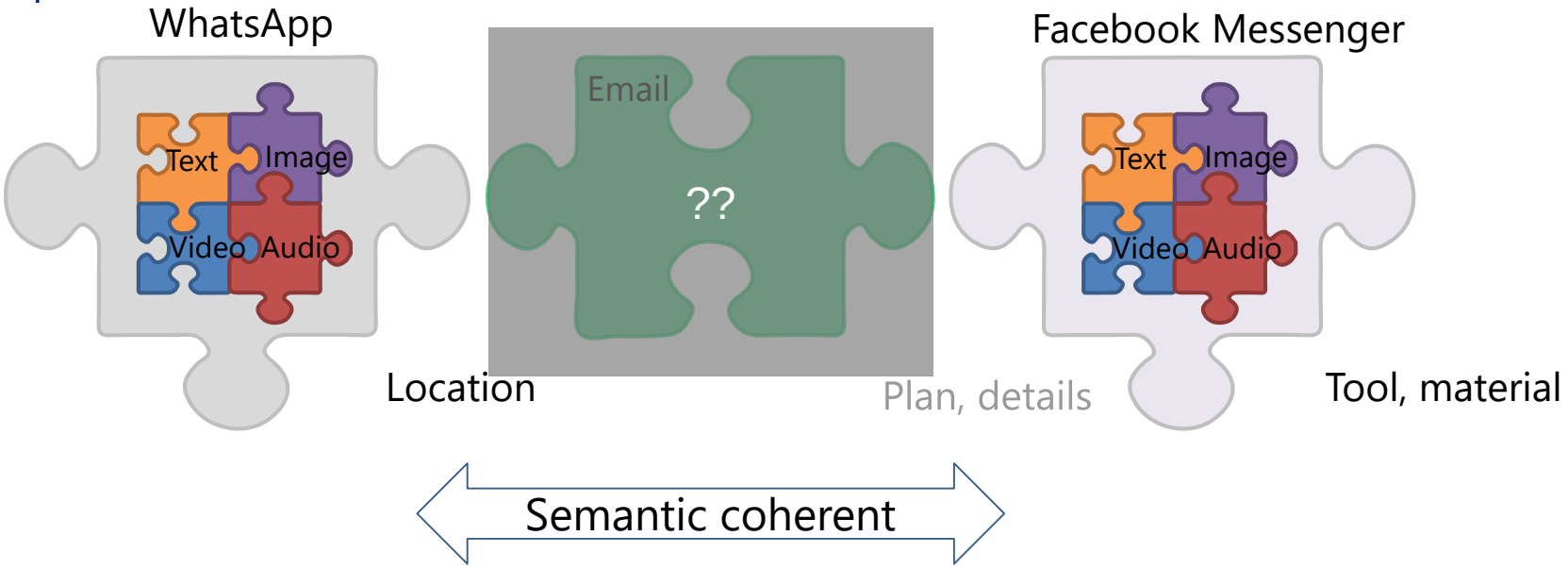


This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Jointly Understanding of Communication in Mobile Forensics

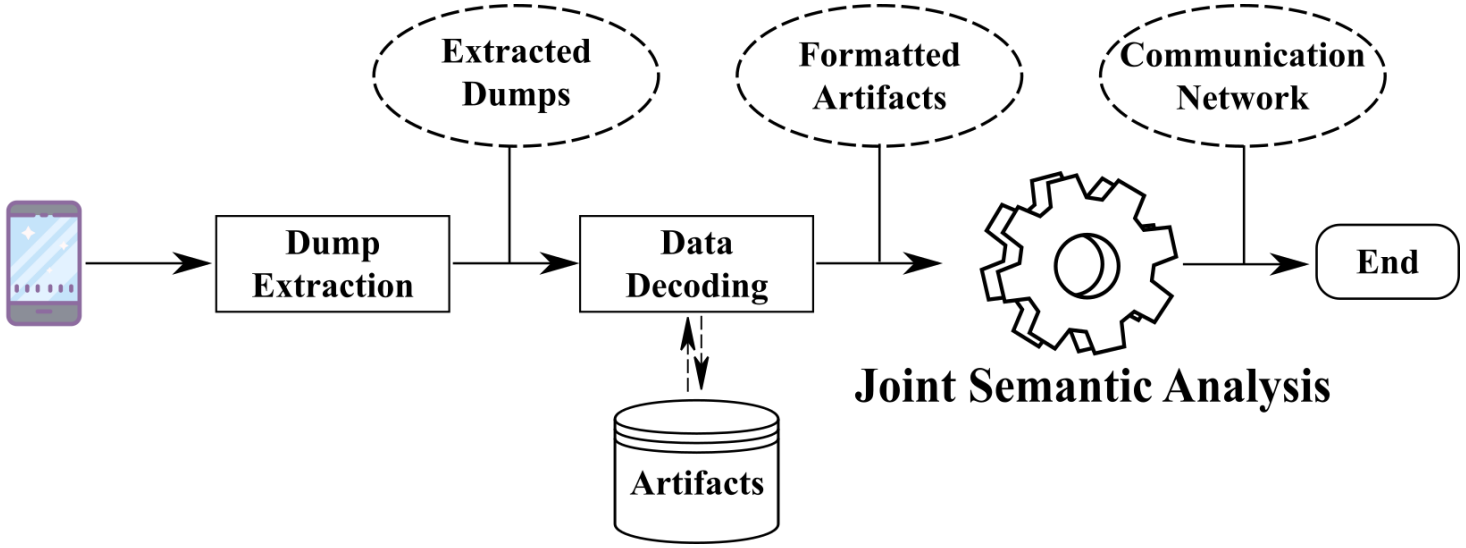
Problems in analysis of mobile communication data - multichannel

- Incomplete contextual information



This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Concept for joint semantic analysis in mobile forensic

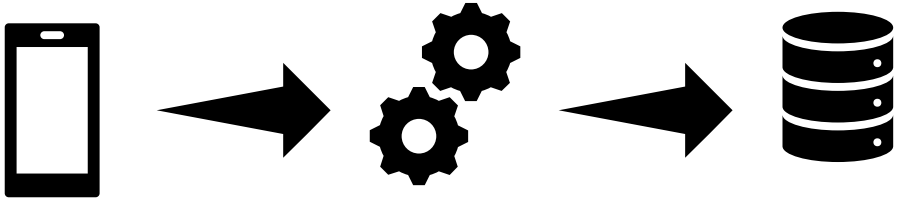


This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Concept for joint semantic analysis in mobile forensic

Dump Extraction

- Extracting data from suspicious mobile devices
  - Physically: extract a complete copy of the examined devices
  - Logically: extract specific parts of the file system, e.g., the logical partitions in a device or an SQLite database.
  - Conducted in compliance with the chain of custody



This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Concept for joint semantic analysis in mobile forensic

Data Decoding

- Decoding data extracted dump files
  - In specific artifact structure in order to analyze the data later
  - Necessary information:
    - Artifact identification: from where this artifact is taken, e.g., Calls
    - Unique identification: uniquely indexed id of artifact
    - Channel identification: concrete resource where this artifact was extracted from, e.g., iOS call
    - Attachment: information about where the attachment is stored locally
    - Information about the sender/caller/receiver
    - Time stamp of artifacts
    - Auxiliary information: whether artifact is read, sent, deleted etc.

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Concept for joint semantic analysis in mobile forensic

Joint Semantic Analysis

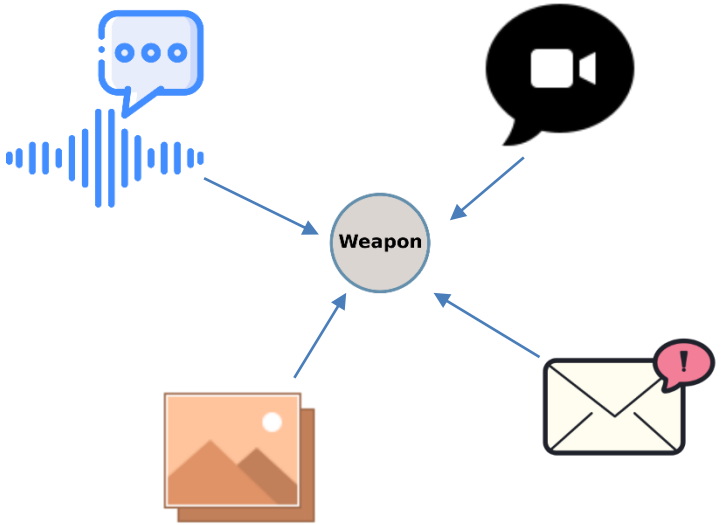
- Aims at explaining the coherent semantic content and hidden connections in a mobile communication in a consistent way
- Is formulated hypothetically as follows:
  - $\tilde{e} = \text{argmax}_{\boldsymbol{\theta}} \tilde{P}(e|d_{cm}; \boldsymbol{\theta}), d_{cm} \in D$
  - e the semantic context in the conversation data D
    - is mostly presented by a topic and possibly connected to a concrete crime
  - $d_{cm} \in D$ stands for a single artifact message spread via the communication channel $c \in D_c$ {WhatsApp, Telegram, Facebook Messenge, email etc.} and represented in the modality $m \in D_m$ {Text, Image, Audio, Video etc.}
  - D is time- and semantically-coherent and organized chronologically
  - $\boldsymbol{\theta}$ is the parameter set that captures the latent semantic in the data

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Concept for joint semantic analysis in mobile forensic

Joint Semantic Analysis

- Aims at retrieving forensic information in different data modalities
  - Correlating data to concept labels

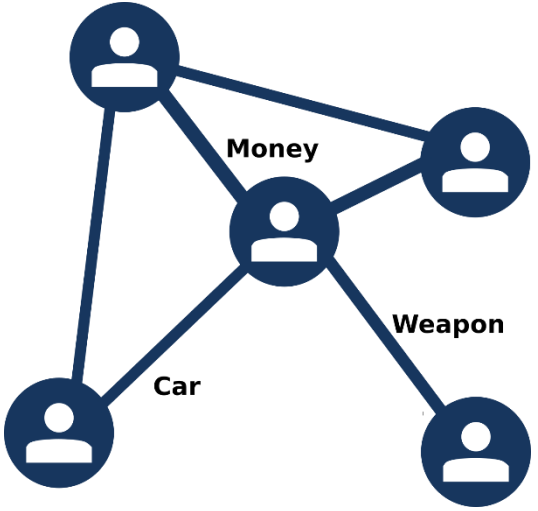


This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Concept for joint semantic analysis in mobile forensic

Joint Semantic Analysis

- Aims at building communication network with respect to the topics discussed in the communication
  - Need a common representation way that maps all data in the same representational form to extract topics



This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Joint Semantic Analysis - Audio

Audio Forensic:

- Refers to the acquisition, analysis, and interpretation of audio recordings as part of an official investigation [Rob Maher]
- Only focus on understanding (transcribing) the spoken data in this study currently
  - Automatic Speech Recognition (ASR)
    - Audio and video (with acoustic signal) can be transcribed in textual form
    - $\widetilde{w} = argmax_w \{p(Y|w)\ P(w))\}$, where Y the acoustic observations, w the sequence of tokens of given acoustic observation.

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Joint Semantic Analysis - Image

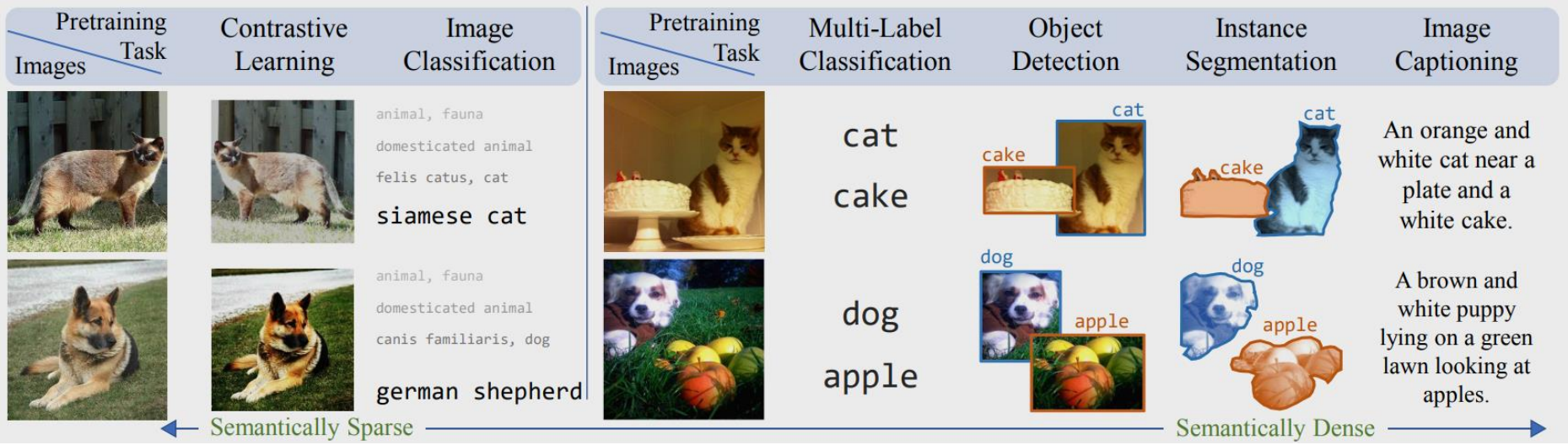
Image Forensic:

- Focus mainly on[Dixit et al.]
  - Image source identification: investigating which device (or class of device) captured or formed the image under investigation
  - Image forgery detection: investigating whether the image under question represents the unmodified captured scene, or has it been forged to deceive the viewer.
- Only focus on semantic information of image
  - Integrate the labels of semantic concepts into retrieving function as well as topic modelling
    - Feature-based approach in image retrieval system for querying contraband [Roussev et al.]
  - Image Classification & Captioning deliver different semantic representation

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Joint Semantic Analysis

Comparison of Semantic Density[Desai et al.]



This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Joint Semantic Analysis - Video

Video Forensic:

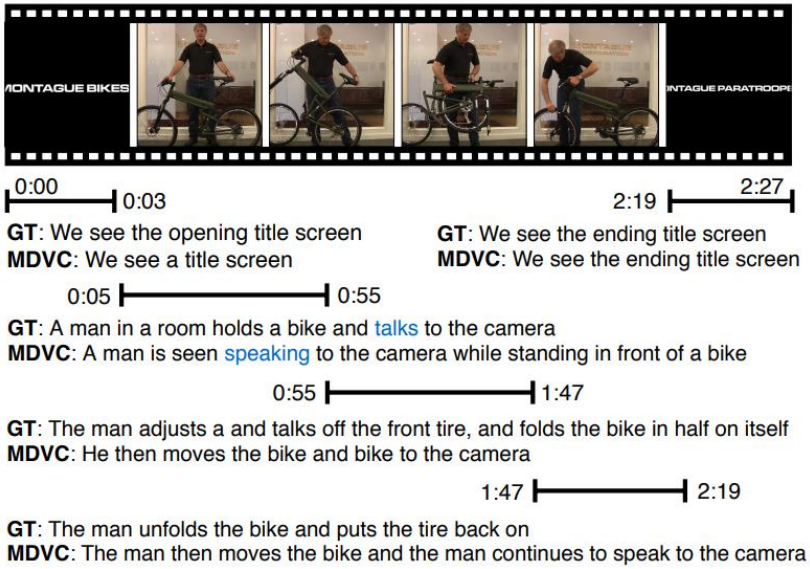
- Aims at showing images and videos used in court and media are verifiably true.
- Based on video content analysis
- Only focus on semantic information of video
  - What activity is going on in video?
    - Feature driven approach in illicit content detection [Rea et al.]
  - Integrate the labels of semantic concepts into retrieving function as well as topic modelling
  - Video Captioning provides descriptions in terms of content

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Joint Semantic Analysis - Video

Video Captioning

- Learning sequential features from frames
- Generating natural sentence with respect to content
  - Might not efficient for longer video
    - Video summarization
      - Detect key frames
    - Image or Video captioning



[Iashin et al.]

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

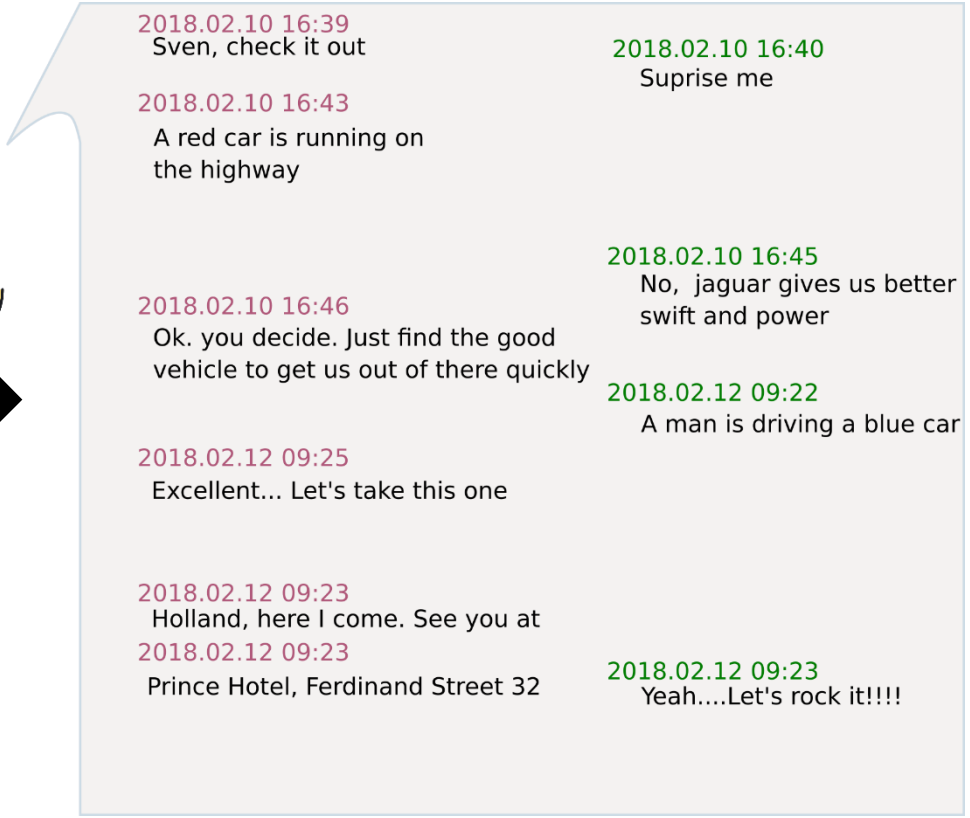
Joint Semantic Analysis

Topic modelling

- As so far, we assume that the data with various modalities is content coherent in the communication
- Textual semantic representation of multimedia data is now available

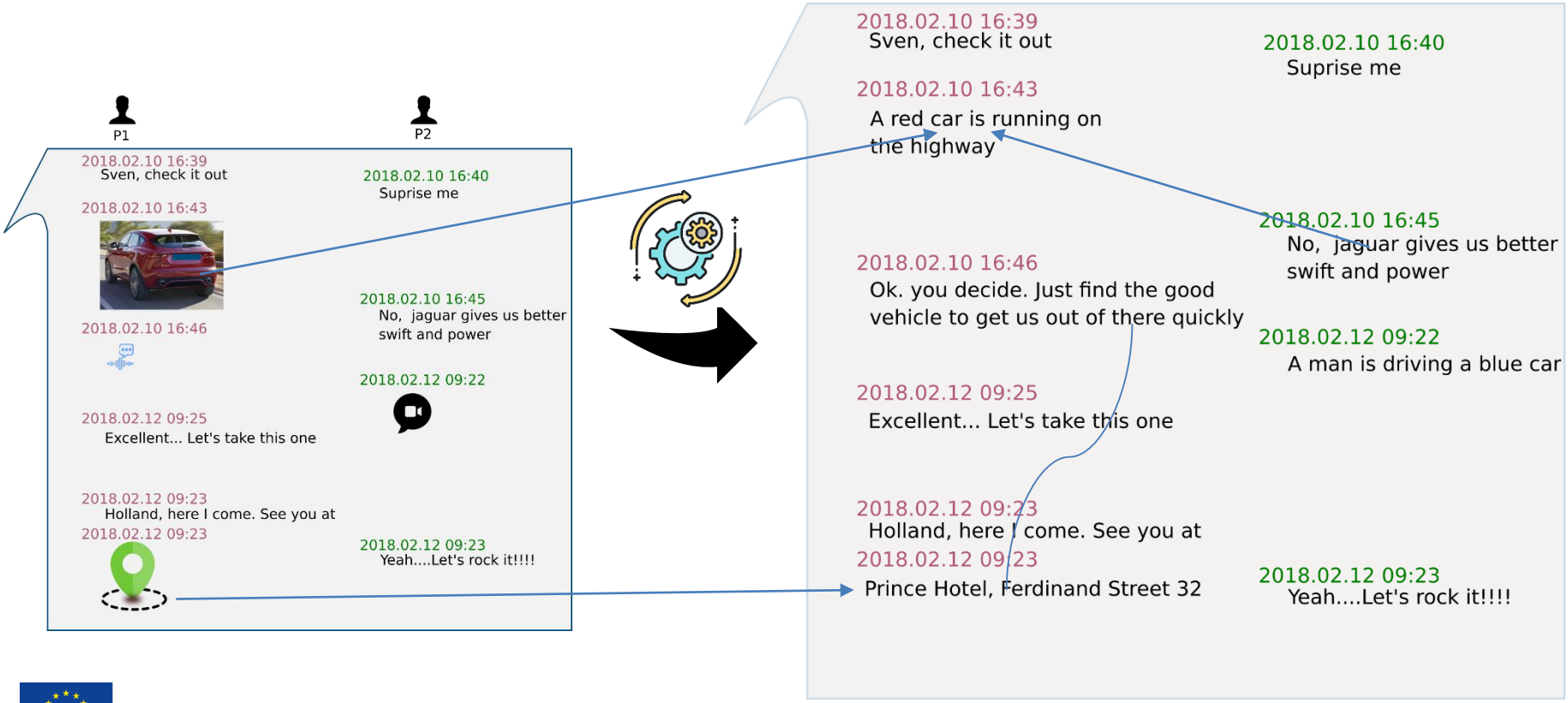
This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Joint Semantic Analysis - Textual semantic representation



This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Joint Semantic Analysis - Textual semantic representation



This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Joint Semantic Analysis

Topic modelling

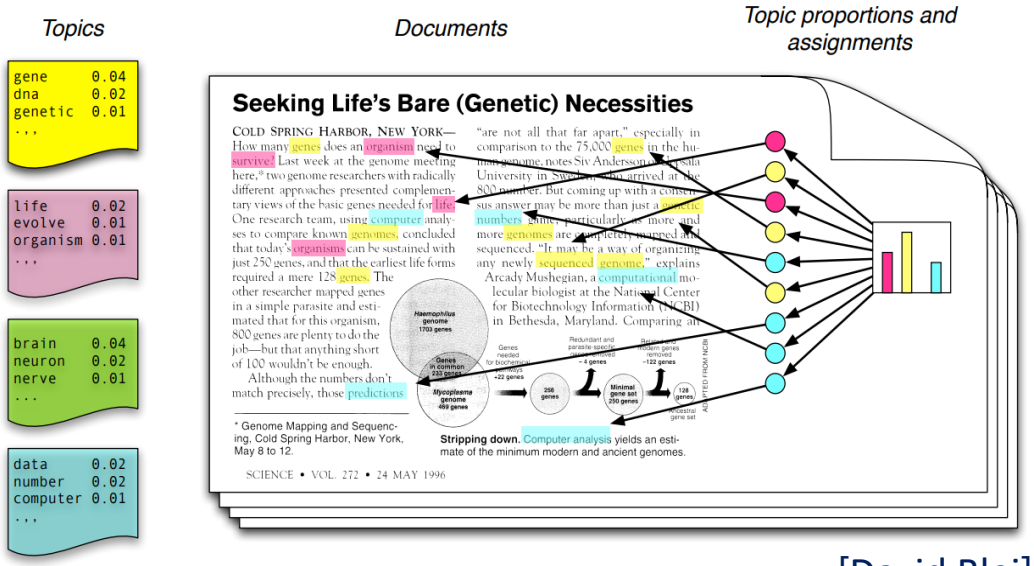
- Assume the communication is mixture of finite topics
- The words occur in similar contexts in text capture same topics
  - Mouse, cat, Tom & Jerry →Cartoons
  - Mouse, keyboard →Computer supplies

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Joint Semantic Analysis

Topic modelling

- Assume topics are distributions of words.
- Generating a document is modelled as a generative event model:
  - Choose a distribution over the topics
  - For each word, choose a topic assignment and choose the word from the corresponding topic
- Goal:
  - Inferring the latent topic structure from given data



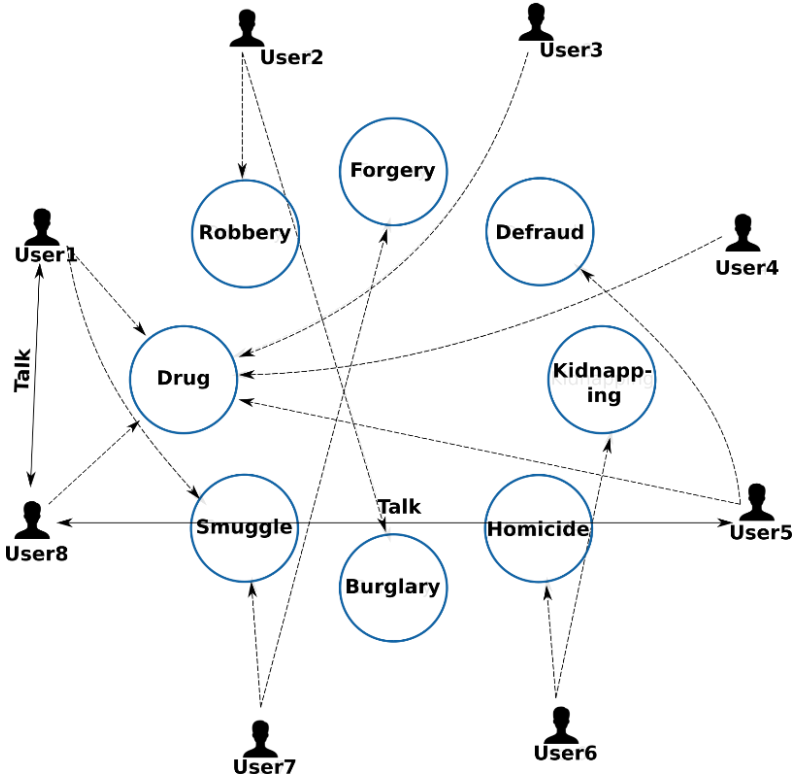
[David Blei]

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Joint Semantic Analysis

Outcomes

- Semantic annotation of each given artifact
- Semantic label of each given artifact
  - Captures the communication network with respect to the semantic topics in communication



This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Concept for joint semantic analysis in mobile forensic

Outlook

- Integrate expert knowledge in this pipeline
- Focus on developing a alias matching strategy in order to match the people who have different names in different communication channels as well as devices
- Taking time information in topic modelling

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Thank you for listening

Jian Xi
WP6.1 Leader
xi@hs-mittweida.de

Contact us

communication@formobile-project.eu
office@formobile-project.eu

Stay in touch

twitter.com/formobile2019

linkedin.com/company/formobile-project

www.formobile-project.eu

This project has received funding from the European Union's Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Reference

[Hans Bucher] H.-J. Bucher, “Multimodal understanding or reception as interaction Theoretical and empirical foundations of a systematic analysis of multimodality”, Bildlinguistik: Theorien – Methoden – Fallbeispiele, Erich Schmidt Verlag, Berlin, 2011

[Spranger et al.] M. Spranger, F. Heinke, L. Appelt, M. Puders, and D. Labudde, “Mona: Automated identification of evidence in forensic short messages,” International Journal On Advances in Security, vol. 9, no. 1&2, pp. 14–24, 2016.

[Jerry Hobbs] J. R. Hobbs, “Why is discourse coherent?” SRI International, November,1978

[Rob Maher] R. C. Maher, “Principles of Forensic Audio Analysis”, Modern Acoustics and Signal Processing, Springer, Cham, 2018

[Dixit et al.] A. Roy, R. Dixit, R. Naskar, R. S. Chakraborty, “Digital Image Forensics: Theory and Implementation”, Studies in Computational Intelligence 755, Springer Singapore, 2020

This project has received funding from the European Union’s Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800

Reference

[Roussev et al.] Y. Chen, V. Roussev, G. Richard, and Y. Gao, “Content-Based Image Retrieval for Digital Forensics,” in Advances in Digital Forensics. IFIP, vol. 194. Springer, Boston, MA, 2005, pp. 271–282, ISBN: 978-0-387-30012-2.

[Desai et al.] K.Desai and J. Johnson, "VirTex: Learning Visual Representations from Textual Annotations", CVPR, 2021

[Rea et al.] N. Rea, G. Lacey, R. Dahyotit, and R. Dahyot, “Multimodal periodicity analysis for illicit content detection in videos,” in The 3rd European Conference on Visual Media Production (CVMP 2006), 2006, pp. 106–114

[Iashin et al.] V. Iashin and E. Rahtu, “Multi-modal dense video captioning,” arXive-prints, Mar. 2020.

[David Blei] D. M. Blei, Introduction to probabilistic topic models, In Communications of the ACM, 2011

This project has received funding from the European Union’s Horizon 2020 - Research and Innovation Framework Programme, H2020-SU-SEC-2018, under grant agreement no. 832800