Data used in AI for e-evidence

4th webinar on the use of Artificial Intelligence in the justice field

Forensic AI in e-Evidence

27 April 2022

Prof. Dr. Marc Engelhart

Present Situation

File management:

- Often still: Paper, paper, paper
- Extensive files with several hundred/thousand pages

Investigation Phase

- Tradition rule in procedural codes: one offender and one act
- Laws are based on a physical world (account records, telephone surveillance, Seizures abroad with mutual legal assistance, ...).

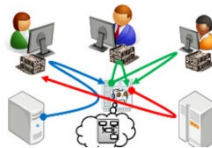# (Secret) Technical Investigation Measures

New Sources

- Mobile Devices

- Cyber-Physical Systems (IoT – Internet of Things)

- Internet/Darknet

- Social Media

- Cloud Server

Challenges for Investigations

- Number of Devices

- Number of Data Sources (Databanks, Google Docs, Shared Services, Workflow-Systems, …)

- Access, logs, IP-addresses

Digitalisation Police/Judiciary

Changing File management

- *From*:   Paper, paper, paper

- *By*:   Digitalization of paper files

- *To*:   Files, Files, Files

Extensive case files with several
hundred/thousand digital (pdf-/word-
/jpg-/mp4-) files

Data Analysis 1.0

- Still following the classic criminal procedure
- Many proceedings with limited use of IT
- Traditional handling of "data"
  (mostly manually by criminal police or external experts)
- Use of simple databases, transcriptions for telephone etc. surveillance
- Printouts of Bank statements and accounting exports
- Visual inspection of mobile phones or screenshots of communications
- Often difficulties (VoIP, messenger services, encryption, etc.)

Consequences

- **Problems**

  Especially in large-scale criminal proceedings, such as economic crime/ corruption/organized crime/terrorism cases due to

  – Amount of information

  – Number of Data sources

  – Complexity

  – Multiple Perpetrators/Group Structures

  – Intelligent perpetrators

  – Concealment (esp. money movements)

- **Data is an "unloved" child**

  – Often no/hardly any data analysis management

  – IT experts and external companies involved

  – Evaluation of data "by hand

  – Years of delays and deficiencies

  – Data carriers become broken or disappear

"The data analysis is still ongoing."

"Nothing could be found during the analysis of the data".

Large Amount of Data

- **Defence lawyers' concerns**
  - Access to seized data?
  - Doubting the chain of evidence and accusations of manipulation
  - Requests for evidence and involvement of (private) IT experts
  - Doubting completeness and accuracy

- **Dangers**
  - "Indictment-destroying" chain-of-evidence errors
  - "Indictment-destroying" document submissions
  - Massive delays in main trial due to IT evidence requests
  - Mitigation of sentences due to excessive length of proceedings

Data Analysis 2.0

Extensive Use of search software

- Enables the search of electronic files
- Enables the search of large files
- Enables the search of large number of files

- Better step-by-step reconstruction of the facts
- Better uncovering unknown connections
- Easier confirming or refuting investigative hypotheses

Data Analysis 2.0 – Problems

- **Text recognition/OCR**
    - Numerous files are not text recognised
      (typically Photos/scans from copy machines)
    - Search words do not work!
- **E-mail attachments**: Search function of e-mail programs does not search (all) attachments
- **Misspellings**, abbreviations, synonyms, code words, transposed numbers, etc.
- **Zip files**, encrypted data, True-Crypt containers, etc.
- **Special software** (e.g. self-programmed company software)
- **Unstructured data**
    - Little or no information about the data available (structure, storage location, content, etc.)
    - Forensic work necessary (deleted files, encryption, hidden TrueCrypt containers)
- **Volume of data**

Data Analysis 2.0

- **Preconditions**
  - Demands high discipline from the searcher
  - Requires strong organisational skills
  - Requires a lot of time/expertise to build up the mosaic of facts

- **Consequences**
  - Often overwhelming for criminal police
  - Incomplete results
  - Change of personnel

Bottom-up approach

Data Analysis 3.0

- Improves information processing

- Deals with large quantities of information

- Recognizes individual pieces of information

- Relates information to each other

- Discloses connections

New: Understanding content

Data Analysis 3.0

- **Normalisation**
  - "Unification" of the form
  - Different spellings of names, times and terms
    - Feb 09, 2.2.2009, 02/09, etc.
    - € 100, 100 EUR, 100,--, 100 Euro
  - Text recognition and spelling errors

Muster Semjce Bau

Service Bau

MSB Muster
Service Bau GmbH

Musterservi
ce

MSB GmbH

MSB Muster
GmbH

MSB Service

Musterservice GmbH

Muster Bau GmbH

- **Classifications**
  - Information type recognition
  - Catalogue of classification structures and document types (Invoices, bank statements, e-mail, calendars, travel documents, chats, etc.)
  - Both original electronic data and scans

Data Analysis 3.0

- **Pattern recognition and entity extraction**
  - Recognition of the information itself (Person, address, organisation, email address, telephone number, account number, date)
  - Pattern recognition (Signatures, letterheads, footers, etc.)
  - Supported by classification (E-mail; Invoices; Meeting minutes)
  - Similarity search
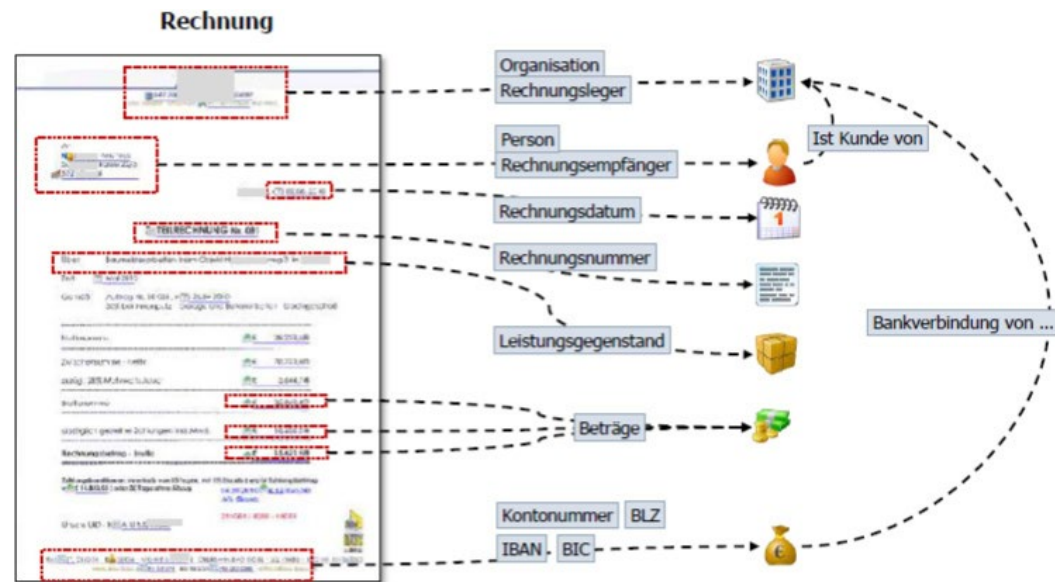
Data Analysis 3.0

- **Ontology building and population**
  - Classification in structures
  - Any semantic ontology possible (is managing director of ...; is in business relationship with...; knows document...)
  - Provides weighting for relevance recognition of the software

Data Analysis 3.0

- Methods
  - **Heuristics** (Rule-based recognition)
    - E.g. many amounts in document is probably invoice
    - E.g. Long number with DE in front is probably IBAN
  - **Machine Learning**
    - Automated decision making by software based on probabilities/confidences
    - Training based on data sets by users
    - Corrections by users lead to an improvement in recognition in theoverall dataset

Prof. Dr. Marc Engelhart – Data used in AI for e-evidence – 27 April 2022

What is new?

- New approach on data analysis
- Construction of the knowledge model ("state of investigation") by software
- Filtering instead of search queries by keywords
  - Faster review and editing of the dataset
  - Queries do not require "programming knowledge" (search syntax)
  - More complete results
  - Knowledge documentation integrated in software
  - Data evaluation is accelerated and motivates investigators

Top-down approach

Prof. Dr. Marc Engelhart – Data used in AI for e-evidence – 27 April 2022

Data Analysis 4.0

- Robo-Prosecutor/Robo Judge?
  - **Complex Case Analysis**
    - Relevant Offenses
    - Scope and Quality of Evidence
  - **Decision Proposal**
    - Proposing the suitable decision (dropping the case because of lack of evidence, non-prosecution agreement subject to imposition of conditions, indictment …)
    - Proposal for specific text of decision (indictment, Non-prosecution agreement …)

Main problem: lack of sufficient training data

EU Regulations?

- DIRECTIVE (EU) 2016/680 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA

- Similar: Article 22 GDPR (with Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, wp251rev.01)

Article 11

Automated individual decision-making

1.   Member States shall provide for a decision based solely on automated processing, including profiling, which produces an adverse legal effect concerning the data subject or significantly affects him or her, to be prohibited unless authorised by Union or Member State law to which the controller is subject and which provides appropriate safeguards for the rights and freedoms of the data subject, at least the right to obtain human intervention on the part of the controller.

2.   Decisions referred to in paragraph 1 of this Article shall not be based on special categories of personal data referred to in Article 10, unless suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

3.   Profiling that results in discrimination against natural persons on the basis of special categories of personal data referred to in Article 10 shall be prohibited, in accordance with Union law.

New Rules?

EUROPEAN COMMISSION

Brussels, 21.4.2021
COM(2021) 206 final

2021/0106 (COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE
(ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION
LEGISLATIVE ACTS

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

- Few specific rules
- Concentrates on real-time remote biometric identification systems

What we need …

More detailed rules on automated decision making in criminal proceedings

- What **decisions** are meant?

  - Types of legal decisions included/excluded (e.g. final decisions)?

  - Certain legal aspects included/excluded (vague legal terms, discretion)

- What **means** automated?

  - Automated perperations of human decision making?

  - Automated support of human decision making?

  - What about partial automatization?

  …

Thank you for you attention!

Prof. Dr. Marc Engelhart

Goethe University Frankfurt am Main

Law Faculty

Entlastungsprofessur Strafrecht

marcengelhart@googlemail.com