Forensic Challenges on Multimedia analytics, Big Data and the Internet of Things

Zeno Geradts
Senior forensic scientist
Digital Technology and Biometrics / University of Amsterdam

ICETE Porto 2018
Outline

- Introduction NFI
- Internet of Things
- Digital Forensics and big data
- Big data in Multimedia
- Summary
First case to testify on ballpoint case

- 1991 Woman found dead with ballpoint in her eye
- Suspect crossbow and ballpoint
- 1995- "Battle" of experts
- 2 weeks for Ballistics research with high speed video camera and gelatin / marks on pen
- Many scenarios possible
- Complex crime scene
- Forensic resources in 1995 limited
Expert of Defense University of Groningen
Chair Forensic Data Science

store and process
understand and decide
analyse and model
Report and visualize
Higher efficiency
Data-intensive
Evidential strength big data?

?
Focus on data

- Extract data
- Make data readable
- Organize data
- Interpret data

Police does 97% of the work
Challenge: many formats, old & new, non-standard

• Tool and library development

• Reverse engineering
Discover the technological principles of a system (e.g. software or communication protocol) through analysis of its function and operation
Trace Recovery & Analysis

Trace-analysis is the expertise to conserve, detect, repair, undelete, decrypt, find, structure and interpret data and traces on any case related digital medium.
rapid/short development cycles

increasing streaming data volume

time spend online

consumer prices for devices+data falling rapidly

fast global expansion of bandwidth 57% per year

digital behaviour
Internet of things

4 BILLION
Connected People

$4 TRILLION
Revenue Opportunity

25+ MILLION
Apps

25+ BILLION
Embedded and Intelligent Systems

50 TRILLION
GBs of Data

Source: Mario Morales, IDC
Internet of things 2020 Gartner

IoT Predictions 2020

Now
10 Billion Devices
1.5 / Person

Year 2020
50 Billion Devices
8 / Person

Automotive
$ 202 Billion

Healthcare
$ 69 Billion

Consumer electronics
$ 445 Billion

Utilities
$ 36 Billion
Internet of things

Insulin pumps webinterface for hospital remote control
40 kilometers queue of trucks filled with paper!!!

8 Terabyte?

1600 hours HD Video
A few years back, a multinational research team based out of Europe made a startling announcement: They had developed a process for storing massive amounts of digital data in microscopic DNA strands.

Theoretically, according to the research, the process could store up to 300,000 terabytes of data in a fraction of an ounce of DNA — which could last for thousands of years. By comparison, today’s most powerful desktop hard drives hold around 6 terabytes of data, and might last 50 years.

Big Data issues
16
Hype Cycle digital (source Gartner)

Source: Gartner (July 2015)
Faulty examples big data
Google Flu Trends: does not work anymore

FEVER PEAKS
A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.

Google’s algorithms overestimated peak flu levels this year

Google’s flu trend map
Everything big data claims to know about you could be wrong

By Yasmin Anwar, Media Relations | JUNE 18, 2018

When it comes to understanding what makes people tick — and get sick — medical science has long assumed that the bigger the sample of human subjects, the better. But new research led by UC Berkeley suggests this big-data approach may be wildly off the mark.

That’s largely because emotions, behavior and physiology vary markedly from one person to
The good news: many examples were it works well credit card fraud detection and casework. VISA states they save billions of euros a year.
Big Data at NFI

- Text Mining
- Data Profiling
- Financial Data Analysis
- Social Network Analysis
Prof. Sloan suggested that the digital forensics landscape resembles the Wild West.

"Increasingly, people who recover and process digital evidence seldom have adequate training in computer science, information systems, or digital forensics. And, there's no agreed-upon standard protocol for retrieving or analyzing digital evidence -- unlike with DNA evidence."
Software Designer Reports Error in Anthony Trial

By LIZETTE ALVAREZ  JULY 18, 2011

MAMI — Assertions by the prosecution that Casey Anthony conducted extensive computer searches on the word “chloroform” were based on inaccurate data, a software designer who testified at the trial said Monday.
What are we looking for?

- more & more innovations with huge data impact
- wide range of digital devices

Product Highlights:
- Thin layers of plastic and other materials are placed on top of each other to create a 3D object based on a computer generated image.
- Provides an alternative for inexpensive and quick production of items. Items such as clothes, drugs, food and jewelry are being printed with this technology.
- Mobile phones are expected to be printable in the next 10 years.
- Prosthetic body parts are being created using this technology.
How to identify relevant digital traces?

By smart automation of our data factories!

smart search+find - and smart analysis solutions

smart broadband infra and smart scalable storage
What are digital traces?

(bits)rows: 0’s en 1’s:
010101001010010010101110101001001111011011011110010101
010011001001001001101010010100101001101111100100101111
1111001100101011010100100101001101011011111010110101

...with a meaning (after interpretation)

Interpretation difficult because of:
Undocumented storage formats
Deleted files
Files partly overwritten
Encryption

10 kB 10.000 bytes
100 kB 100.000 bytes
2 MB 2.000.000 bytes
Data analysis at the NFI

*Specifications available?*
- Yes? Use the specs
- No? Reverse Engineering and Carving

*Add results to Forensic libraries*
- File systems (Snorkel)
- File formats (Traces)
- RAM memory (Mammal)

*Process data based on libraries*
- Create trace index (Data model)
- Investigate using GUI or API (Query model)
Conventional digital investigation

Several months
Digital investigation using XIRAF

Several weeks (1 TB in 24/hrs)
Future of digital investigation: HANSKEN

Some hours (1Tb/20 min) – direct results at start
HANSKEN: secure big data investigation platform
Evolution forensic analysis – automation, speed & coverage

<table>
<thead>
<tr>
<th>Method</th>
<th>Throughput</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual import and manual processing</td>
<td>Conventional: throughput months</td>
<td>50% 50%</td>
</tr>
<tr>
<td>Manual import and automated processing</td>
<td>XIRAF: throughput weeks</td>
<td>70% 30%</td>
</tr>
<tr>
<td>Automated import and automated massive-parallel processing</td>
<td>HANSKEN: throughput hours</td>
<td>85% 15%</td>
</tr>
</tbody>
</table>
Many new techniques
Digital Forensics as a Service: A game changer

R.B. van Baar*, H.M.A. van Beek, E.J. van Eijk

Netherlands Forensics Institute, Laan van Ypenburg 6, 2497 GB The Hague, The Netherlands

ABSTRACT

How is it that digital investigators are always busy and still never have enough time to actually dig deep into digital evidence? In this paper we will explore the current implementation of the digital forensic process and analyze factors that impact the efficiency of...
Examples hypotheses in digital forensic science

- has the computer been hacked or not?
- has the email been send or not?
- has the USB been plugged in or not?
- was the phone in this location or at the location presented by the defence?
- has the child pornography been send by the computer of the suspect or not?
- is the child porn photographed with this camera or another camera?
The FBI’s second thoughts — and a second chance for privacy

A customer tried out an iPhone 6S at an Apple store in Chicago.

By Hiawatha Bray | GLOBE STAFF MARCH 23, 2016
US innocence project

Contributing Causes of Wrongful Convictions (first 325 DNA exonerations)
Total is more than 100% because wrongful convictions can have more than one cause.

- Eyewitness Misidentification (235 cases)
- Unvalidated / Improper Forensics (154 cases)
- False Confessions / Admissions (88 cases)
- Informants / Snitches (48 cases)
Forensic flaws

To date, 24 people arrested or convicted on bite mark evidence have been exonerated by DNA.

FBI examiners gave flawed forensic testimony in **257 of those 268 trials**, or more than **95 percent**.

**268 trials** in which hair evidence was used against criminal defendants.

**257 trials** with flawed forensic testimony.

**32 death-penalty cases** with flawed forensic testimony.

NOTE: The FBI is completing reviews of about 900 lab reports.
Source: National Association of Criminal Defense Lawyers and Innocence Project analysis of FBI and Justice Department data as of March 2015
THE WASHINGTON POST
Deputy Attorney General Sally Yates is expected to propose expanding the Justice Department’s review of forensic testimony by the FBI Laboratory beyond hair matching to widely used techniques such as fingerprint examinations. (Evelyn Hockstein/For The Washington Post)

By Spencer S. Hsu March 21

The Justice Department on Monday proposed expanding its review of forensic testimony by the FBI Laboratory beyond hair matching to widely used techniques such as fingerprint examinations and bullet-tracing.
Lawful internet interception
Interpret data

Challenge: data is not self-explaining

Add models and analysis to support interpretation

- Scenario analysis
- Timeline analysis
- Geographical models: e.g. location of cell phones
- Analysis of images / video / audio
  - Size
  - Speed
  - Face recognition
  - Speech recognition
- Author recognition
Problem: data derive meaning from relationships to each other → find patterns

Intelligent Data Analysis
Structure, analyze, link and interpret data with data/text/process mining, (social) network analysis, profiling and classification methods
About Forensic Big Data Analysis

- Our data come from confiscated phones, hard drives, licence plate cameras, telephone providers, and so on...
What if...

- An ATM machine is blown up
- A prepaid cell phone is found on the scene
- The police have their eyes on a suspect
- **Research question: is the suspect the user of the prepaid phone?**
What information do we have?

- You know the phone number of the prepaid phone and that of the suspect’s private phone.

- The telephone provider provides the police with usage data for both phones.

- Every time a phone connects to a cell tower, you know when it happened.

- You know the location of each cell tower.
Problem 1: cell tower location data are not precise and depend on...

- Theoretical range: 35km
- Direction of transmission
- Distance
- Obstacles (tall buildings)
- Weather conditions
- Network load
To summarize...

- We want to know if the suspect is the user of a prepaid phone that can be linked to a crime.
- We know when and where the prepaid phone was used.
- We know when and where the suspect’s phone was used.
- But our data are sparse and imprecise...
Likelihood Ratio
Prosecutor’s fallacy

- Suppose a fingerprint found on a crime scene matches a person in a large fingerprint database.

- Probability of a random match is 1 in 10 thousand.

- **Prosecutor’s fallacy**
  “The likelihood of the evidence is 1 in 10 thousand for an innocent person, therefore, the probability of innocence is 1 in 10 thousand.”

  - This forgets about the prior odds and other evidence.
  - Most people aren’t murderers. Prior odds are very low.
  - There may be other evidence supporting innocence.
CSI?

Who is this person?

Comparative Height/Volume Analysis

Height: 5' 7"

weight: 120Lbs
Example 1 Known person
Unknown persons (Burton)
http://media.leidenuniv.nl/legacy/burton-talk-leider
Super recognizer
Forensic Comparison

Hypothesis 1: The suspect is the same person as the person on the CCTV-images of the robbery
Hypothesis 2: The suspect is another person with similar generic facial features as the person on the CCTV-images of the robbery (and is also not a direct family-member).
The findings of the investigation are far more probable if the suspect is the same person at the person on the CCTV-images than if this is another person with similar general facial features.

<table>
<thead>
<tr>
<th>Verbal equivalent</th>
<th>Order of magnitude of evidential strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>approximately equally probable</td>
<td>1-2</td>
</tr>
<tr>
<td>slightly more probable</td>
<td>2-10</td>
</tr>
<tr>
<td>more probable</td>
<td>10-100</td>
</tr>
<tr>
<td>appreciably more probable</td>
<td>100-10.000</td>
</tr>
<tr>
<td>far more probable</td>
<td>10.000-1.000.000</td>
</tr>
<tr>
<td>extremely more probable</td>
<td>&gt;1.000.000</td>
</tr>
</tbody>
</table>
A video file can be unplayable because:

- you don’t have the proper player,
- it is broken, or
- it is deleted.

More on this on Friday.
CSI?

What is the license plate number of the passing cars?
Image improvement: super resolution

New methods are tested and validated
Image improvement: super resolution

New methods are tested and validated
Description of the man in the red rectangle

Does this man have a mustache?
Description of the man in the red rectangle
Description of the man in the red rectangle

Be aware of artifacts
Which ear?

A

B
The answer:

Different angles can change the image dramatically.
Comparison of video images

Reference images from the suspect/object:
Under similar circumstances, in front of the same camera, at the same spot in the same pose.
Visual comparison

Which features are depicted consistently?
If a feature looks different in the different images, does that mean a different person is depicted?
If a feature looks the same in different pictures, is that a clue that the same person is depicted?
More on this topic tomorrow.

questioned

Reference image with suspect
What is the color of the car?

A. Blue  B. Red  C. Neither, it’s...
What is the color of the car?

A. Blue  B. Red  C. Neither, it’s...
The car was illuminated by a traffic light
Distance from the train

A. 30 m
B. 90 m
C. 160 m
Answer:

C. 160 m
Forensic Event Viewer

Synchronized video streams

Forensic Event viewer

GSM cell site data

3D model

Annotations and report
Solution: Repair NFI Defraser: open source repair video carving
http://sourceforge.net/projects/defraser/
Challenge:

- Who? Attribution can be difficult in cyberspace
- What? Is the data reliable, protocols
- When? At what time
- Where? Jurisdictions etc.
- Combining evidence

Extract data
Digital Camera Identification

The process of

Linking images to the source camera

Linking images to images in a database to determine a common source
Casework links
Casework

- Example where it worked
PRNU Compare
Bayesian

Question: were the images made with the seized camera?

Conclusion

The findings of the investigation are:

- Equally likely
- Somewhat more likely
- More likely
- Much more likely
- Very much more likely

if H1 is true, than if H2 is true.

The findings are very much more likely if the Seized Camera took the child pornographic image, than if another camera took the image.
Large Scale Camera Identification

- Sorting photos by source
- Identify photos from the same source (camera)
- New valuable information and insight
Sorting Images by Source

**Scan** → Extract → Compare → Cluster → Explore

Sorted by resolution and directory

4320x3240

1024x768
Sorting Images by Source

Scan → Extract → Compare → Cluster → Explore

PRNU noise patterns (fingerprints)
Sorting Images by Source

Scan → Extract → **Compare** → Cluster → Explore

Images compared to all images
Sorting Images by Source

Scan → Extract → Compare → Cluster → Explore

Images grouped by source

threshold = 0.001
Sorting Images by Source also GPU / social networks

Scan → Extract → Compare → Cluster → Explore
Image Integrity

Has there been tampered with this image?
North Korea 28 Mar 2013
North Korea
Critical systems
• Shut down?
• Copy?
• Ignore?

Virtualization storage, networks, servers
• Locate the data?
• Whose jurisdiction?
• Who is the owner?

Shielding
• Anti forensic software
• Encryption
• Darknets / bitcoins or other crypto coins
Questions