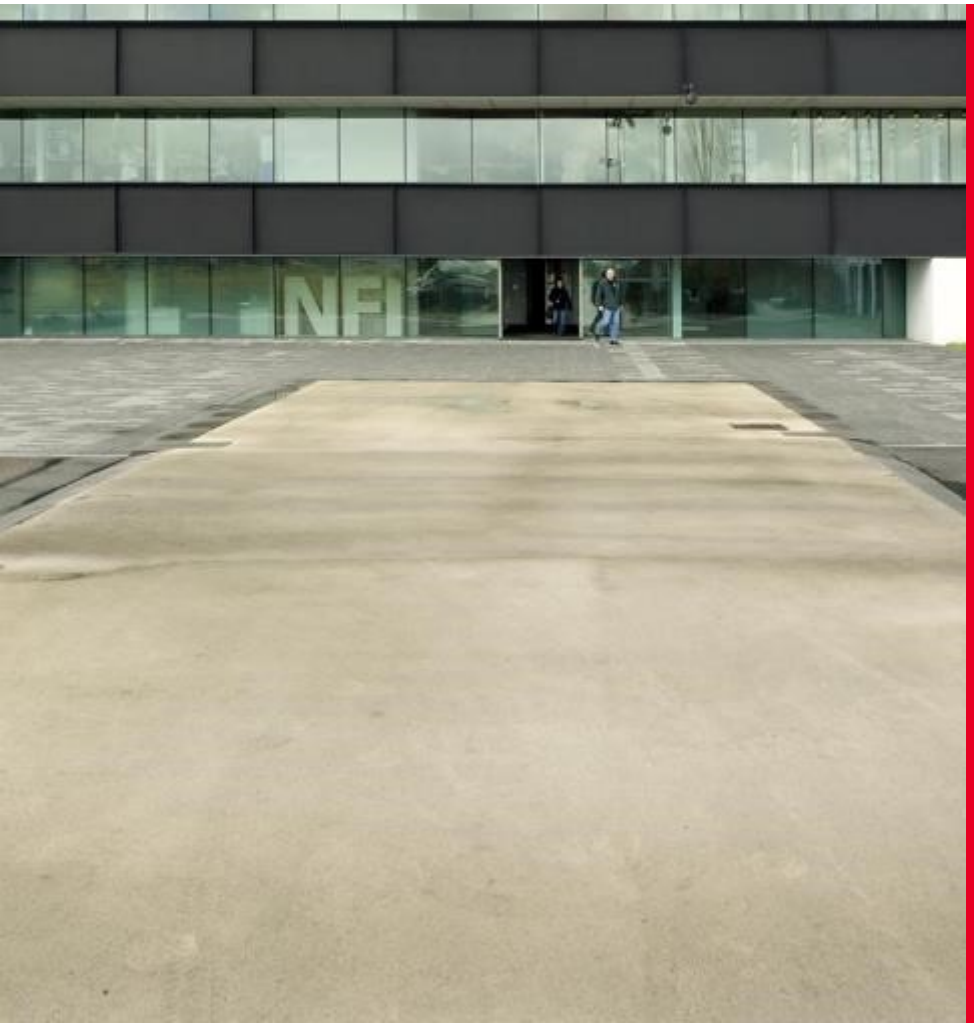
Netherlands Forensic Institute
Ministry of Security and Justice



Artificial Intelligence en digital forensich onderzoek, risico of oplossing ?

prof. dr. ing. Zeno Geradts
Senior forensic scientist /
Special Chair Forensic Data Science
Digital Technology and Biometrics /
University of Amsterdam

Symposium E-discovery 2019

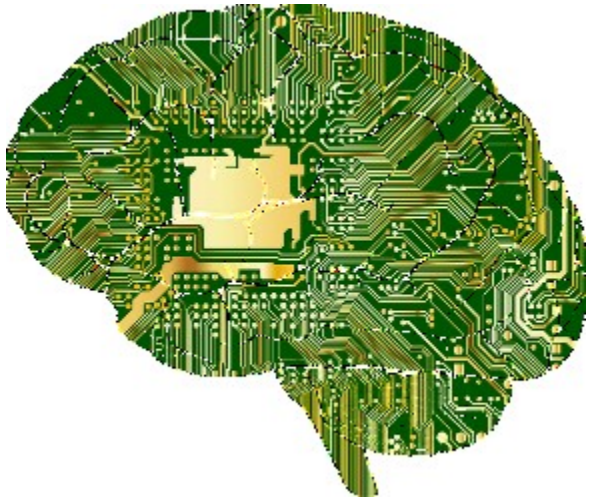
COST Project DigForAsp

DigForAsp (Digital forensics: evidence analysis via intelligent systems and practices) – CA17124 is funded by the European Cooperation in Science and Technology (COST). DigForAsp activities were launched on 10th September 2018 for 4 years.

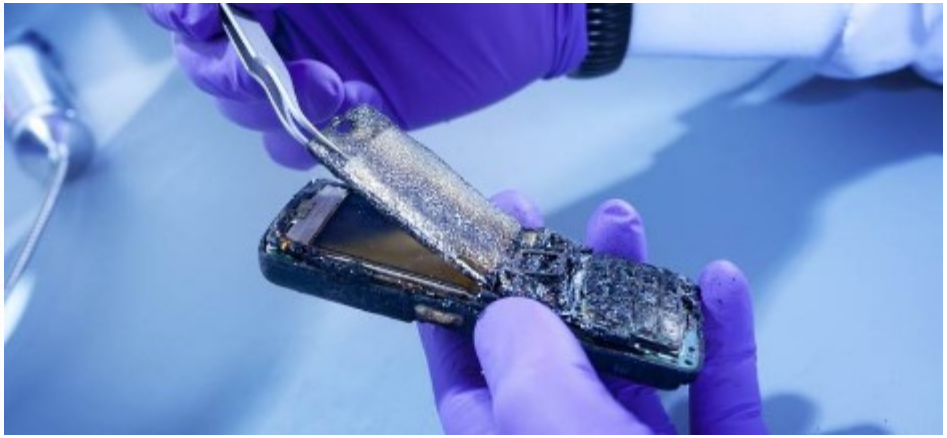
Funded by the Horizon 2020 Framework Programme of the European Union

Outline

- Introduction
- Deep learning and neural networks
- Examples
- Issues
- Outlook and conclusion



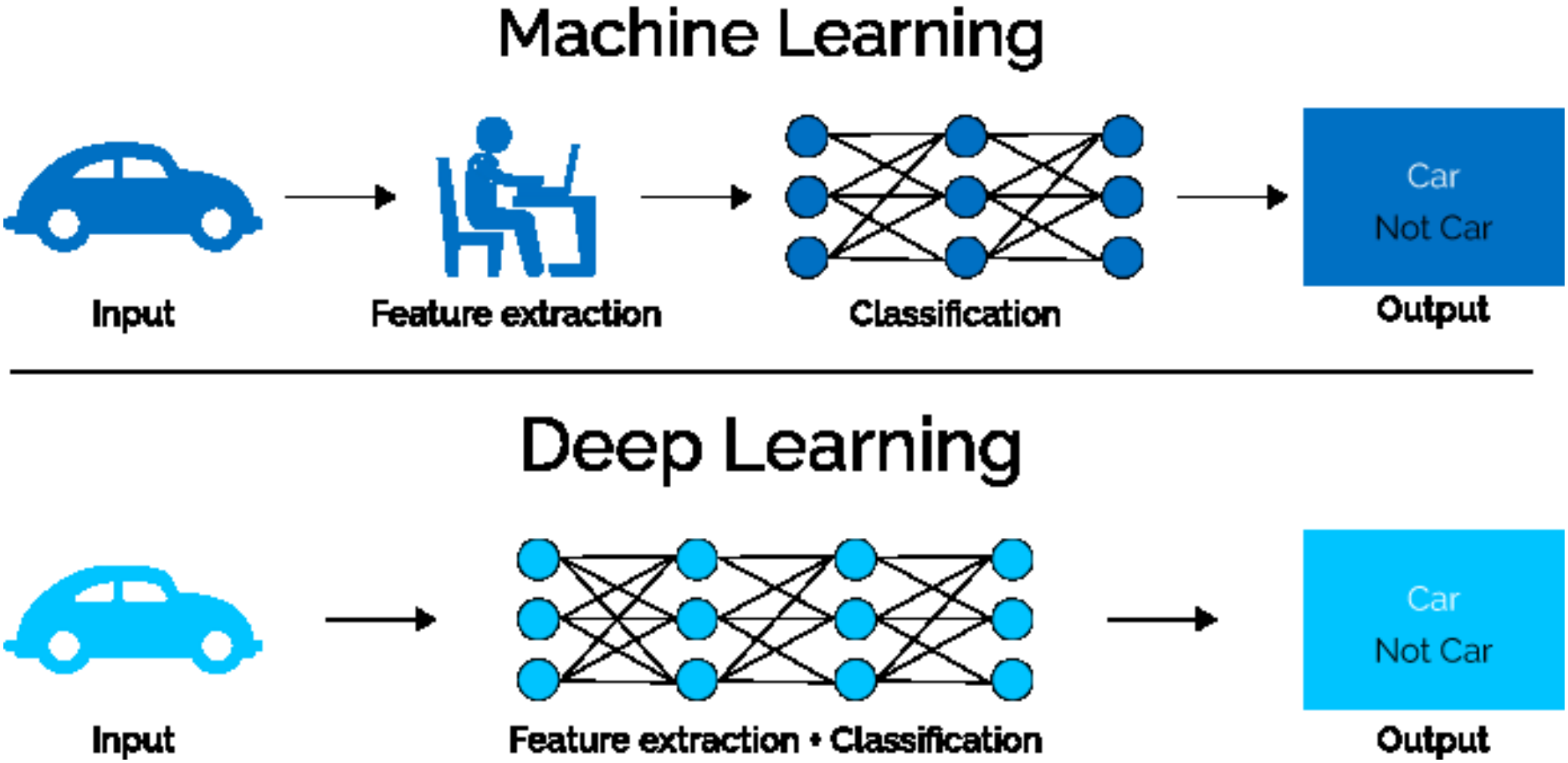
Netherlands Forensic Institute



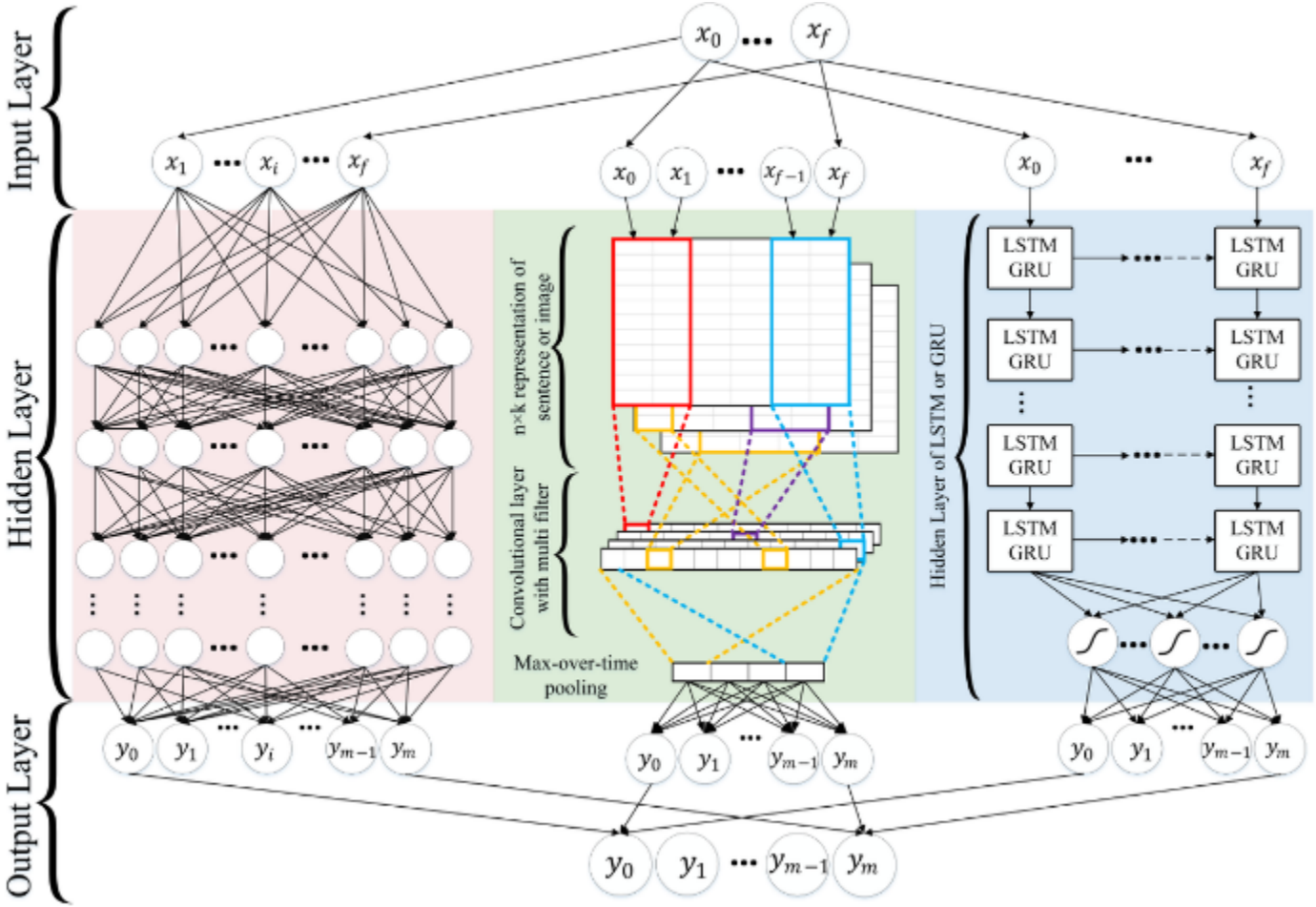
University of Amsterdam
Chair Forensic Data Science

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

Machine learning vs deep learning



Neural network multilayer



Calculation speed with Digital Evidence

Digital Evidence

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

```
000025c0  0e 4a 5b fc 74 d6 21 a2  fb d3 5d bf 59 45 11 9a  |.J[.t.!...].YE..|
000025d0  fd d6 00 28 d6 6f a1 b0  60 59 6c ce c6 d0 4c 5c  |...(.o..`Yl...L\|
000025e0  61 10 8d cf 95 41 c1 3e  b3 f3 62 ff 1b b0 fc dc  |a....A.>..b.....|
000025f0  ea 5b fb 07 95 27 28 59  9a 05 e0 06 27 7b 2a 59  |.[...'(Y....'{*Y|
00002600  0e 43 72 1b ce 4b 1f 59  e2 ce d9 f3 86 34 5e f9  |.Cr..K.Y.....4^.|
00002610  38 d1 4a 0f 06 2e 70 66  c9 49 01 00 7b ca 93 c2  |8.J...pf.I..{...|
00002620  6d 70 02 ab b6 78 90 e1  5b ca 1c 14 29 13 77 93  |mp...x..[...).w.|
00002630  9f 29 a4 d1 1f 1f 3f 20  69 29 c4 ae fd c3 01 bf  |.)....? i).....|
00002640  76 c4 bd a8 cc 99 0b e3  93 74 82 b8 1e cc 2e da  |v........t......|
00002650  64 eb 74 64 5c 6c d7 91  78 5a 58 5b 59 c5 9a 82  |d.td\l..xZX[Y...|
00002660  4d e0 2c 58 1b 5c 83 c7  7e 98 3e 37 b2 93 99 90  |M.,X.\..~.>7....|
00002670  fd 00 e0 3a 8e 4f 13 e5  1f 23 bb b5 f8 b0 a3 85  |...:.O...#......|
00002680  86 74 b9 1b 18 b7 5f 03  4b a1 6a c5 7c c4 46 1e  |.t...._.K.j.|.F.|
00002690  6b 09 51 77 6b 3b 0d 9c  17 36 31 71 07 f4 9a bb  |k.Qwk;...61q....|
```

NETHERLANDSFORENSICINSTITUTE

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.

```
0e 4a 5b fc 74 d6 21 a2  fb d3 5d bf 59 45 11 9a  |.J[.t.!...].YE..|
fd d6 00 28 d6 6f                                 ce c6 d0 4c 5c  |...(.o..`Yl...L\|
61 10 8d cf 95 41                                 ff 1b b0 fc dc  |a....A.>..b.....|
000025f0  ea 5b fb 07 95 27                       06 27 7b 2a 59  |.[...'(Y....'{*Y|
00002600  0e 43 72 1b ce 4b                       f3 86 34 5e f9  |.Cr..K.Y.....4^.|
00002610  38 d1 4a 0f 06 2e                       00 7b ca 93 c2  |8.J...pf.I..{...|
00002620  6d 70 02 ab b6 78                       14 29 13 77 93  |mp...x..[...).w.|
00002630  9f 29 a4 d1 1f 1f                       ae fd c3 01 bf  |.)....? i).....|
00002640  76 c4 bd a8 cc 99                       b8 1e cc 2e da  |v.........t.....|
00002650  64 eb 74 64 5c 6c                       5b 59 c5 9a 82  |d.td\l..xZX[Y...|
00002660  4d e0 2c 58 1b 5c                       37 b2 93 99 90  |M.,X.\..~.>7....|
00002670  fd 00 e0 3a 8e 4f                       b5 f8 b0 a3 85  |...:.O...#......|
00002680  86 74 b9 1b 18 b7 5f 03  4b a1 6a c5 7c c4 46 1e  |.t...._.K.j.|.F.|
00002690  6b 09 51 77 6b 3b 0d 9c  17 36 31 71 07 f4 9a bb  |k.Qwk;...61q....|
```

rapid/short development cycles

increasing streaming data volume

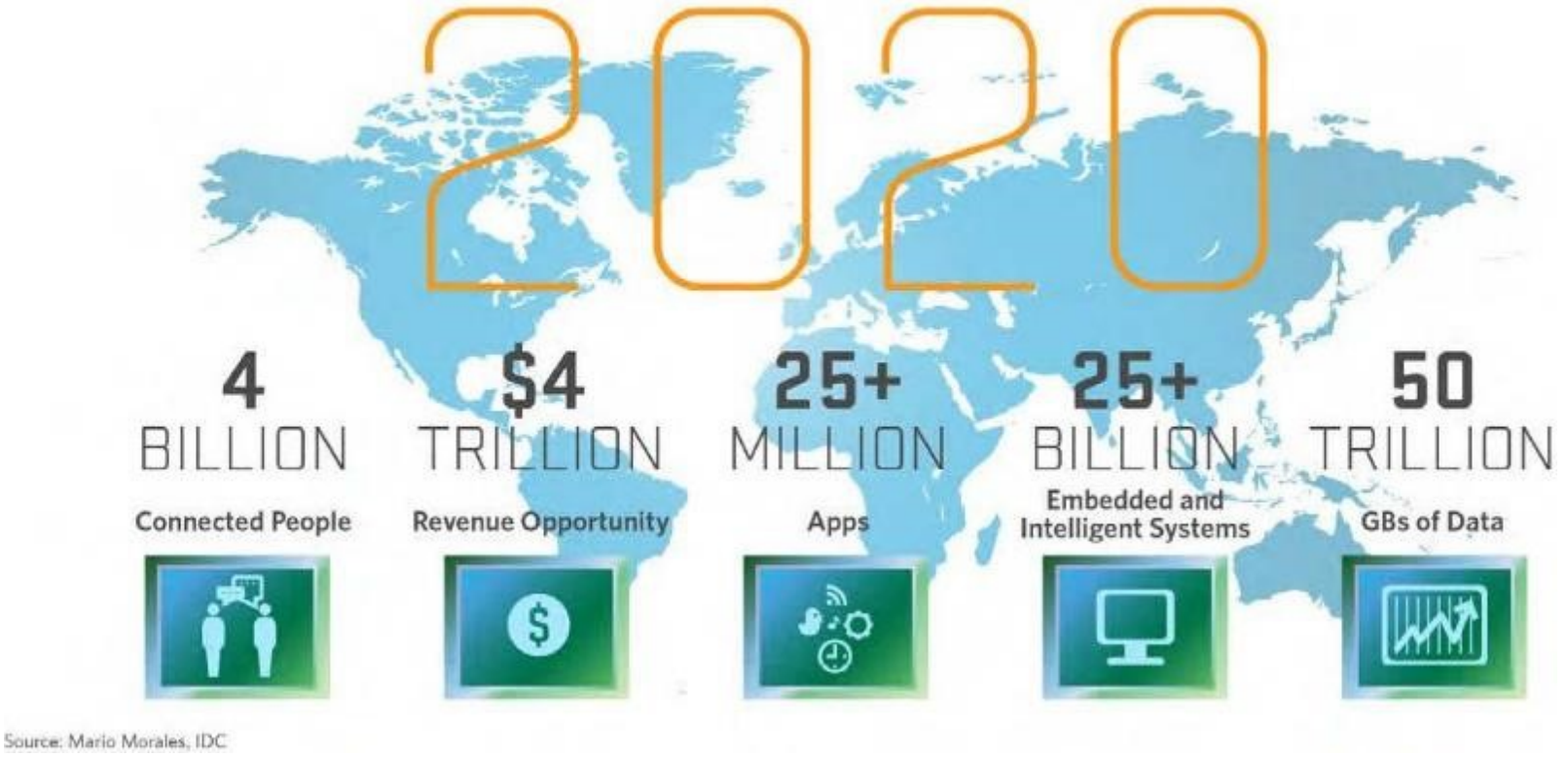
time spend online

consumer prices for devices+data falling rapidly

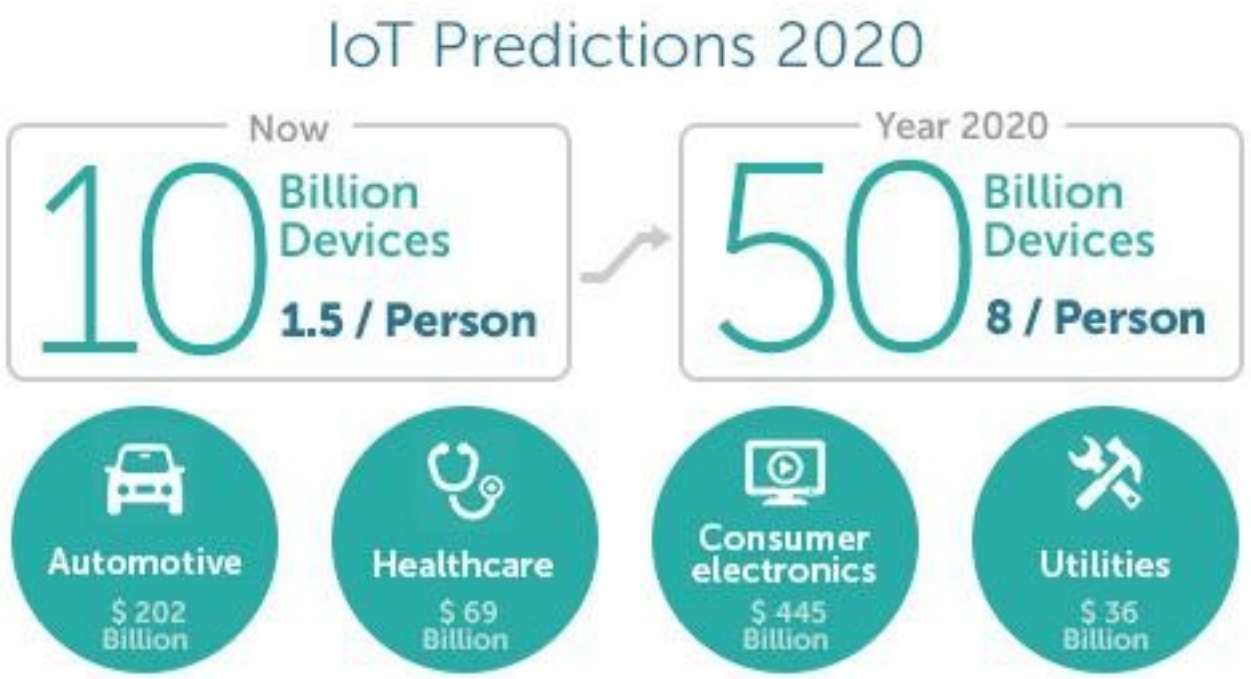
fast global expension of bandwidth
57% per year

digital behaviour

Internet of things



Internet of things 2020 Gartner



40 kilometers queue of trucks filled with paper!!!

8 Terabyte?

1600 hours HD Video

Big Data issues

news.berkeley.edu/2018/06/18/big-data-flaws/

UC Berkeley

Berkeley News

Research People Campus & communit

MIND & BODY, RESEARCH, TECHNOLOGY & ENGINEERING

Everything big data claims to know about you could be wrong

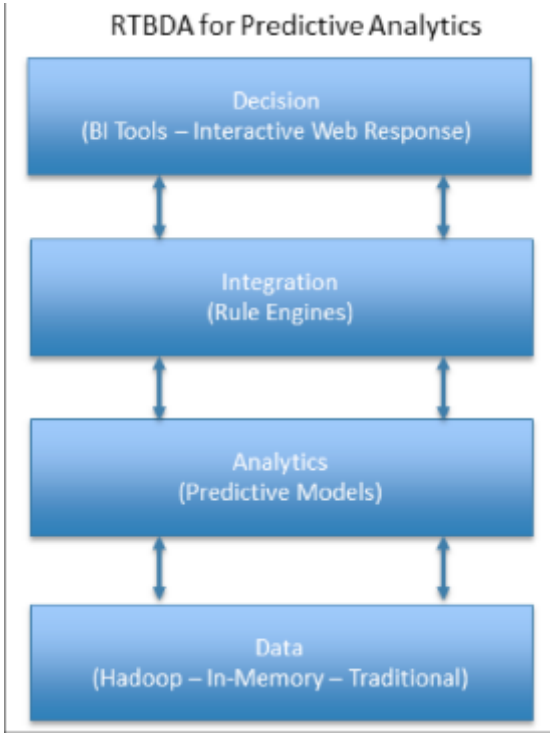
By Yasmin Anwar, Media Relations | JUNE 18, 2018

445

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.

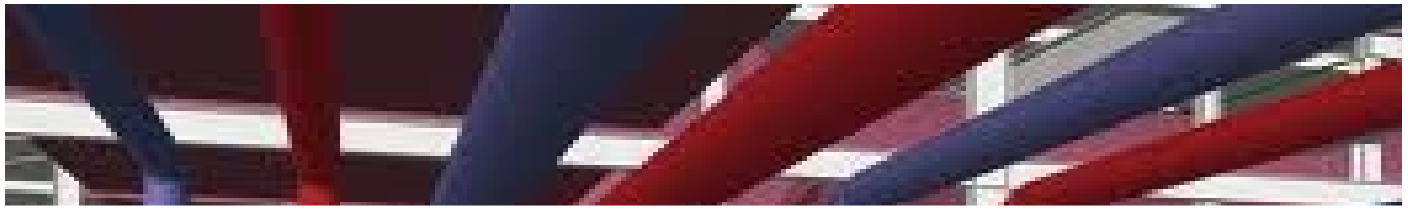
That's largely because emotions, behavior and physiology vary markedly from one person to

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



How to identify relevant digital traces?

smart search+find - and smart analysis solutions

By smart automation of our data factories!

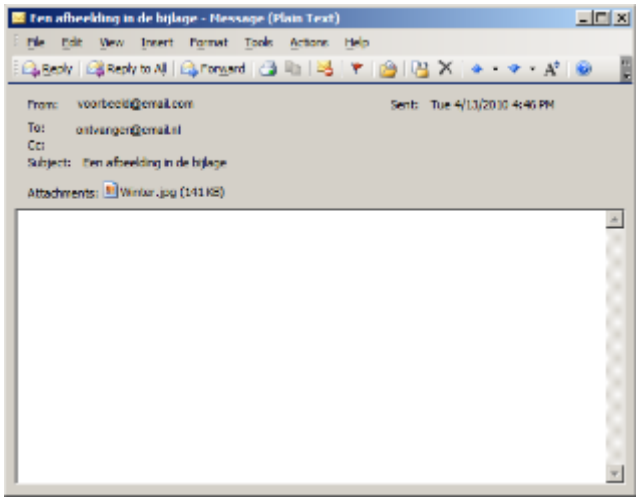
smart broadband infra and smart scalable storage

What are digital traces?

(bits)rows: 0's en 1's:

```
01010100100101001001011100101001110101001000111100101010
01001100100100100100110101010101001010010100001010111111
11111001001001101010101010010100101000010101111101010111
```

…with a meaning (after interpretation)



Interpretation difficult because of:

- Undocumented storageformats
- 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?

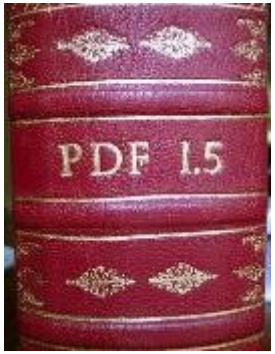
- Yes? Use the specs
- No? Reverse Engineering and Carving

Add results to Forensic libraries

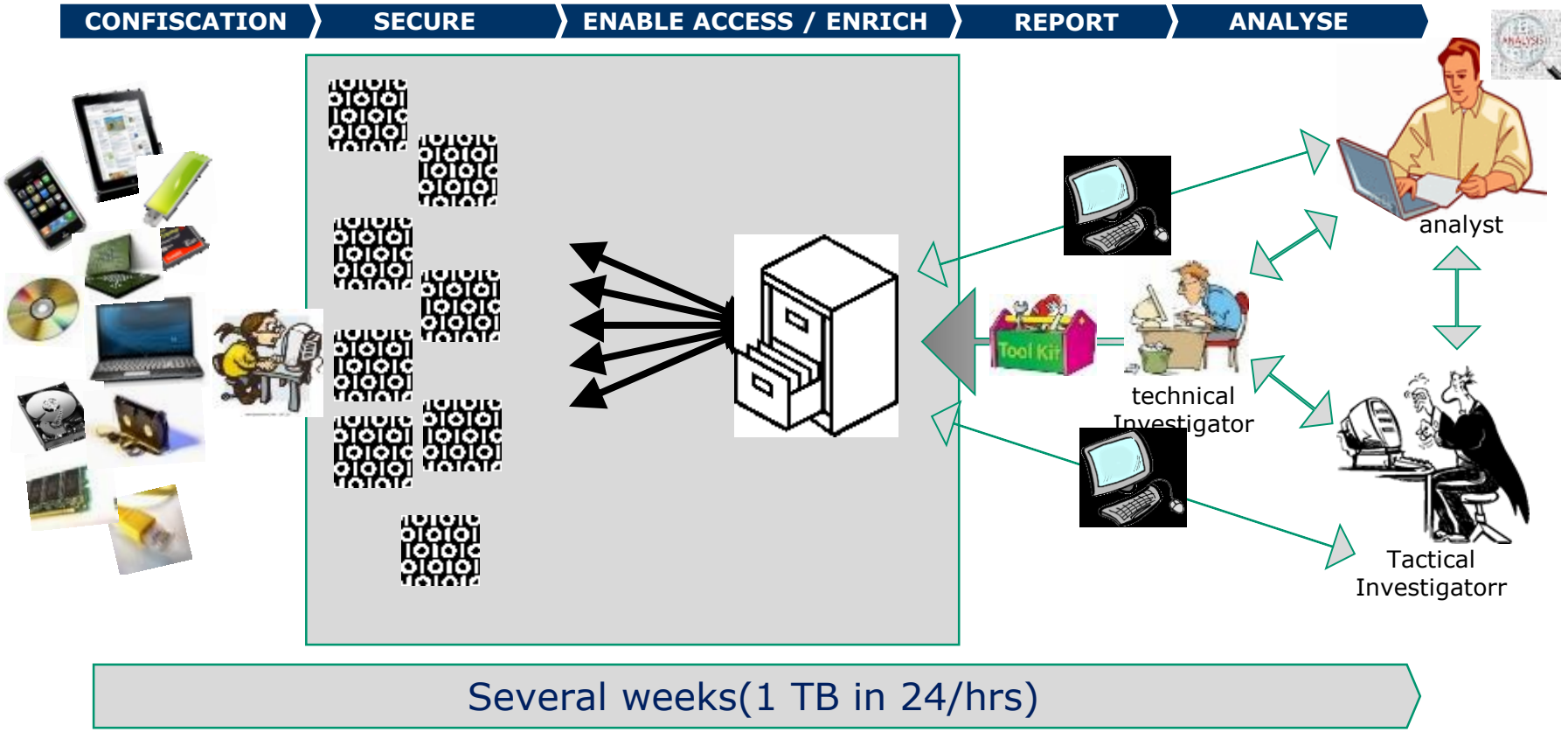
- File systems (Snorkel)
- File formats (Traces)
- RAM memory (Mammal)

Process data based on libraries

- Create trace index (Data model)
- Investigate using GUI or API (Query model)



Digital investigation using XIRAF

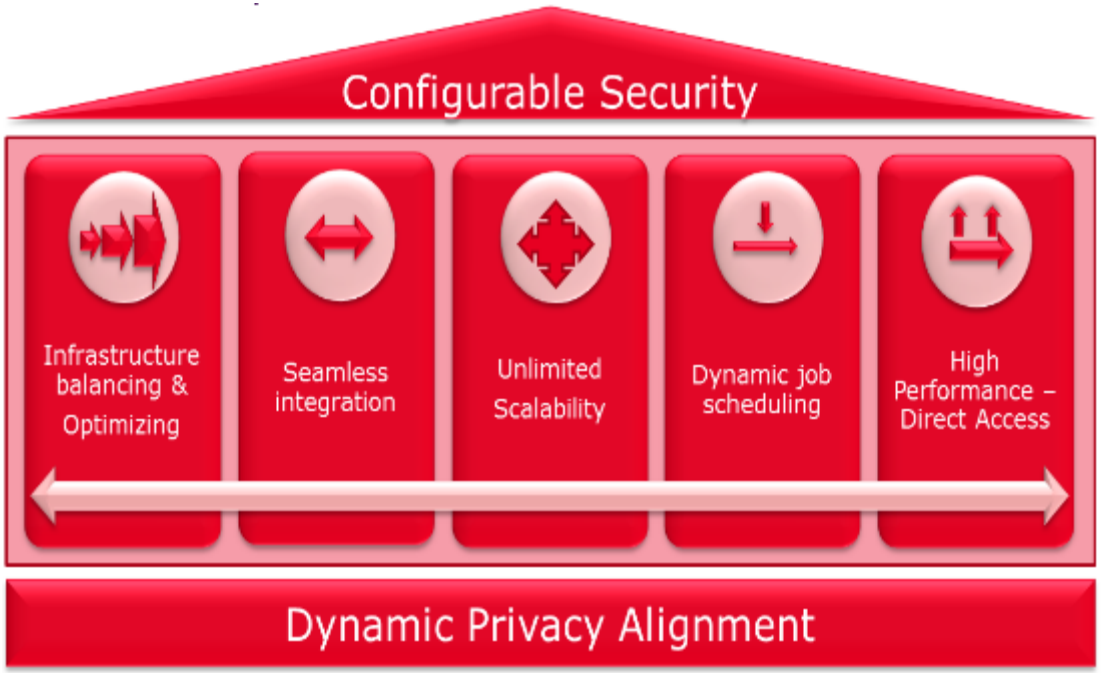


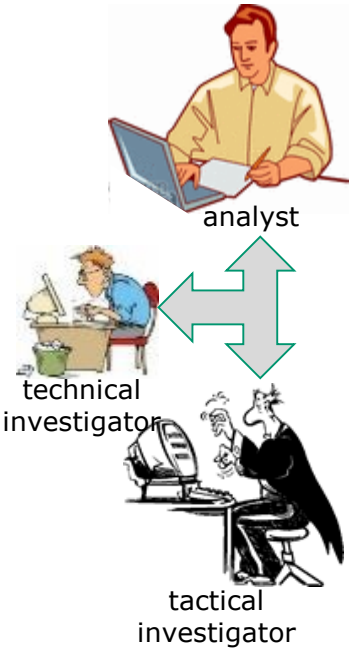
Future of digital investigation: HANSKEN

CONFISCATION | SECURE | ENABLE ACCESS / ENRICH | REPORT | ANALYSE

Datastream X

Datastream Y





Some hours (1Tb/20 min) – direct results at start

Evolution forensic analysis – automation, speed & coverage

manual import and manual processing

Conventional: throughput months

50% | 50%

manual import and automated processing

XIRAF: throughput weeks

70% | 30%

automated import and automated massive-parallel processing

HANSKEN:throughput hours

85% | 15%

https://www.dfrws.org/2014eu/proceedings/DFRWS-EU-2014-7.pdf

DFRWS-EU-2014-7.pdf 1 / 9

https://www.dfrws.org/2014eu/proceedings/DFRWS-EU-2014-7.pdf

Digital Investigation 11 (2014) S54–S62

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

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

Keywords:
Digital forensics
DFaaS

A B S T R A C T

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

is processed and of the manner in which the traces collected by this processing is analyzed.

Related work

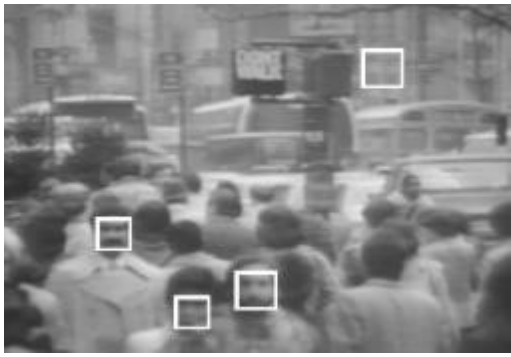
Examples hypotheses in digital forensic science

- 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 ?

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



About Forensic Big Data Analysis

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



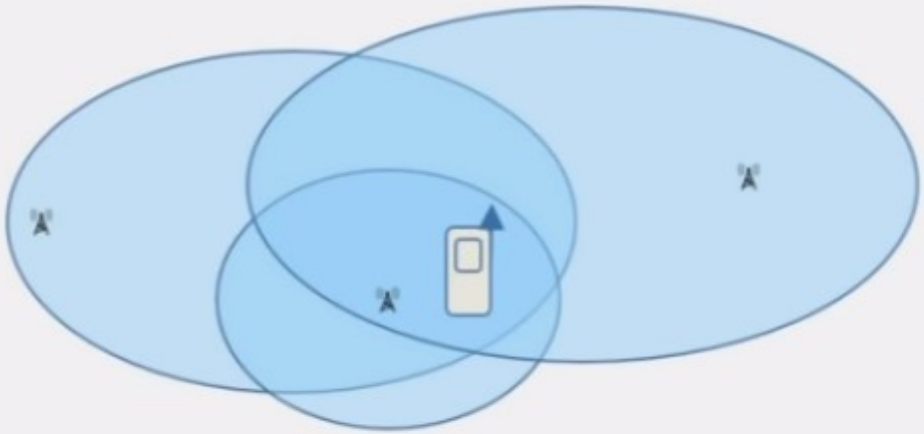
What if…

- 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?**

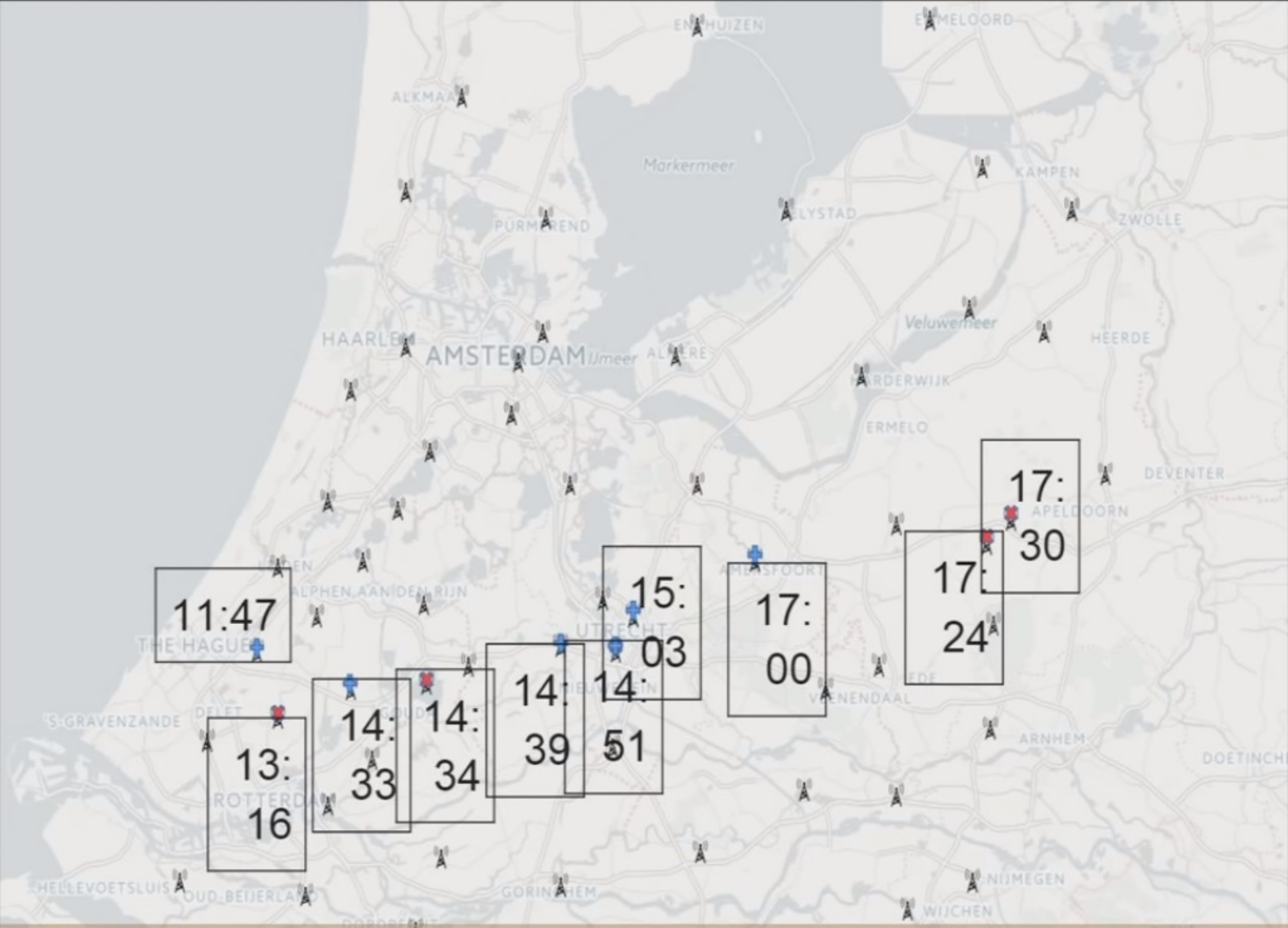
What information do we have?

- 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…



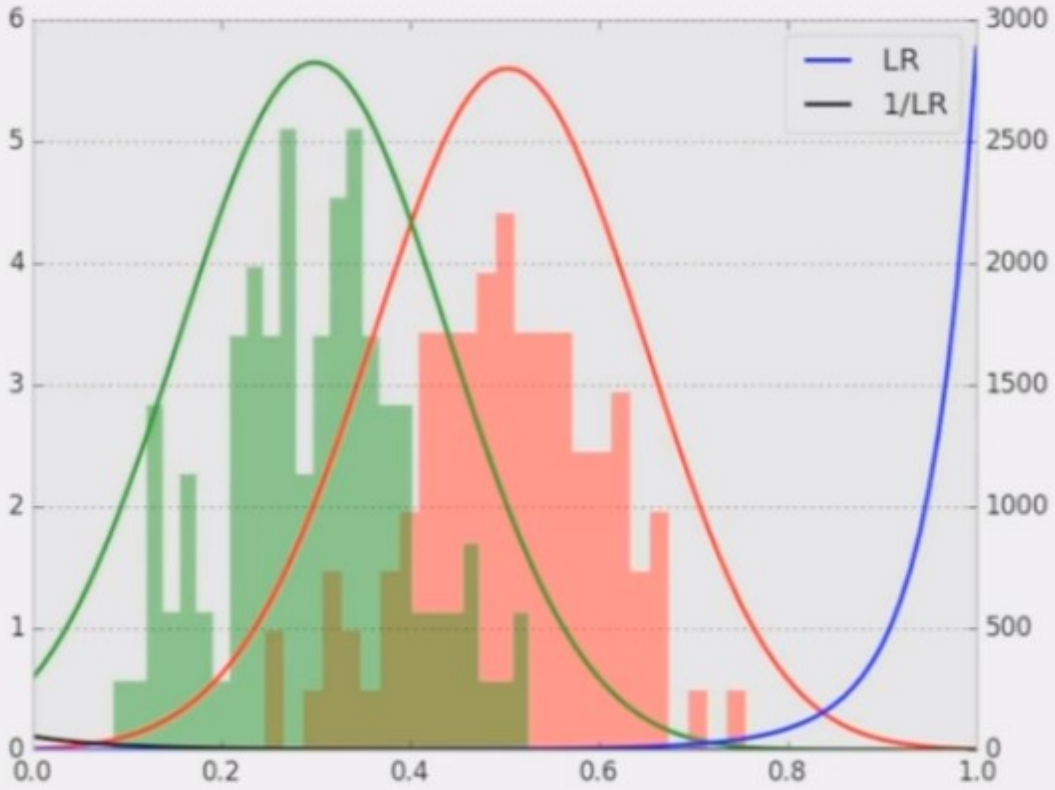
- Theoretical range: 35km
- Direction of transmission
- Distance
- Obstacles (tall buildings)
- Weather conditions
- Network load


11:47
13:
16
14:
33
14:
34
14:
39
14:
51
15:
03
17:
00
17:
24
17:
30
Markermeer
IJmeer
Veluwemeer
AMSTERDAM
HAARLEM
ALKMAAR
PURMEREND
LYSTAD
ENKHUIZEN
EMMELOORD
KAMPEN
ZWOLLE
HEERDE
HARDERWIJK
ERMELO
DEVENTER
APELDOORN
AMERSFOORT
UTRECHT
VENENDAAL
ARNHEM
NIJMEGEN
WIJCHEN
THE HAGUE
DELFT
ROTTERDAM
S-GRAVENZANDE
HELLEVOETSLUIS
OUD-BEIJERLAND
GORINCHEM
DORDRECHT

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

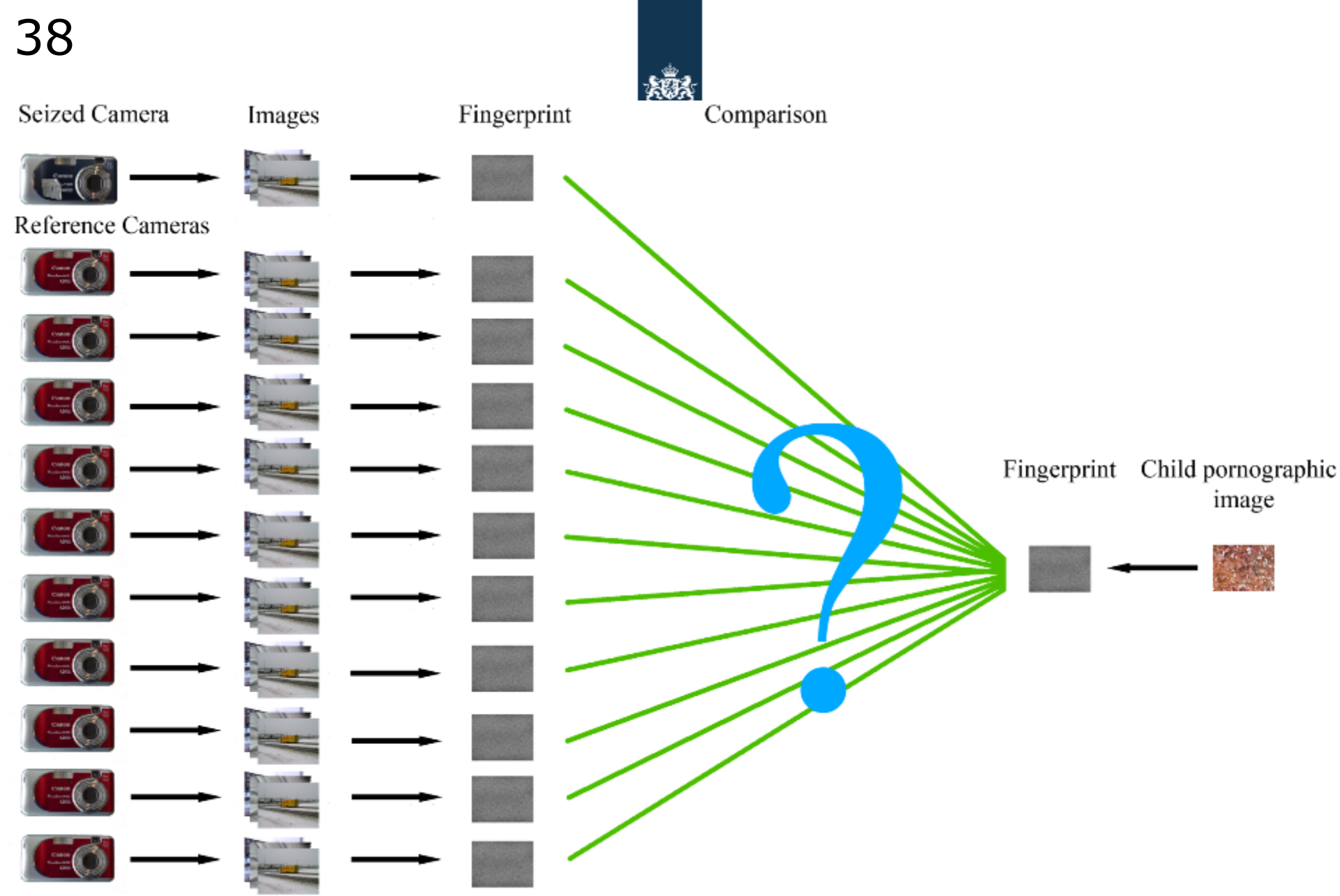


Digital Camera Identification

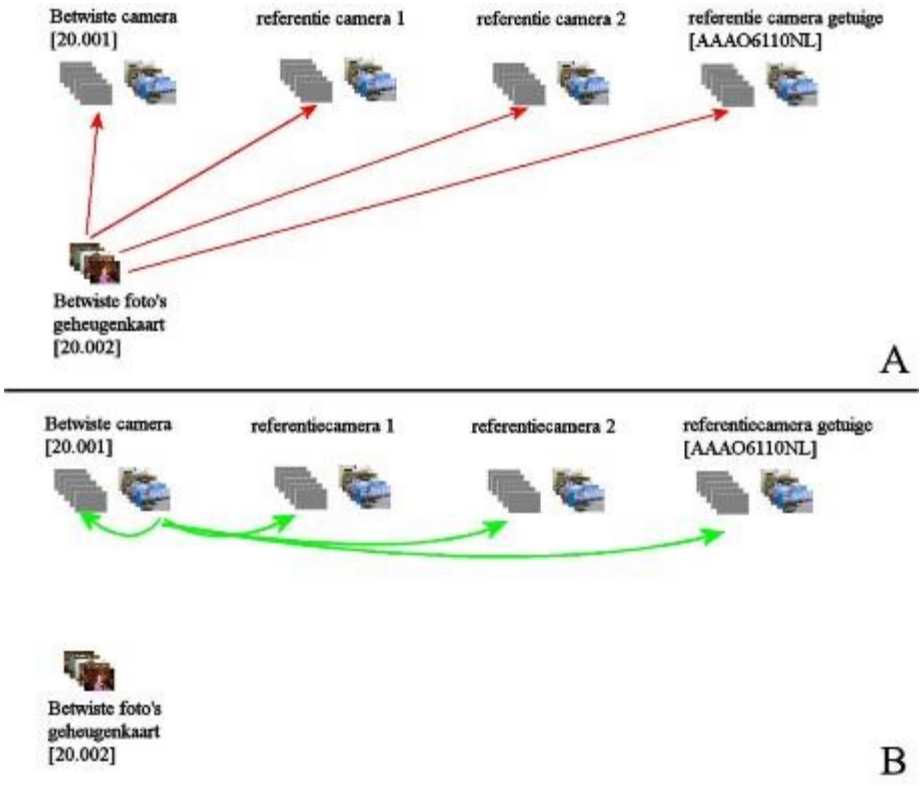
The process of

Linking images to the source camera

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

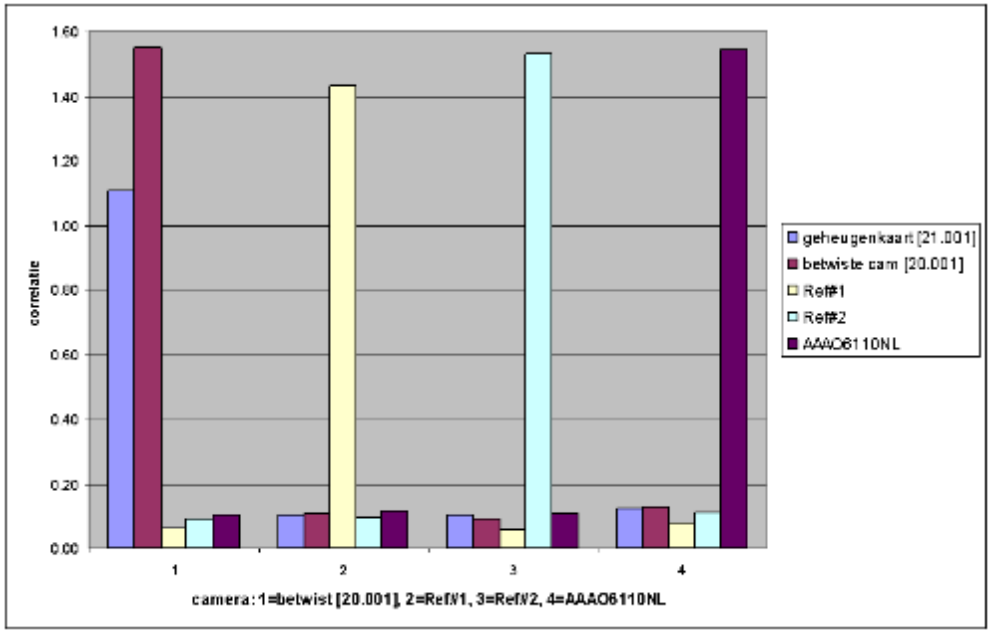

Seized Camera
Images
Fingerprint
Comparison
Reference Cameras
Fingerprint
Child pornographic image

Casework links



Casework

- Example where it worked



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

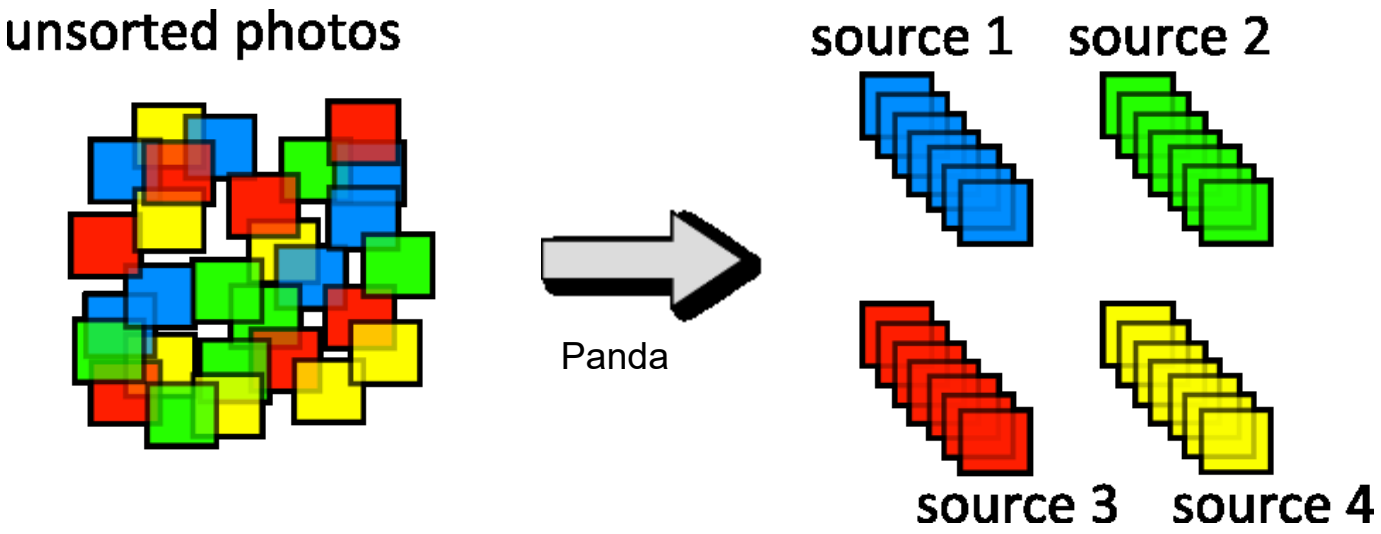
Camera	...	...	...	Sum
Comparing: Foto in kwestie (Onbekend)				
Verdachte Camera reference (Canon PowerShot)	..	..	...	0.131330
Camera 1 reference (Canon PowerShot)	..	..	...	0.008054
Camera 4 reference (Canon PowerShot)	..	..	...	0.007700
Camera 2 reference (Canon PowerShot)	..	..	...	0.007022
Camera 3 reference (Canon PowerShot)	..	..	...	0.006287

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

Scan

4320x3240

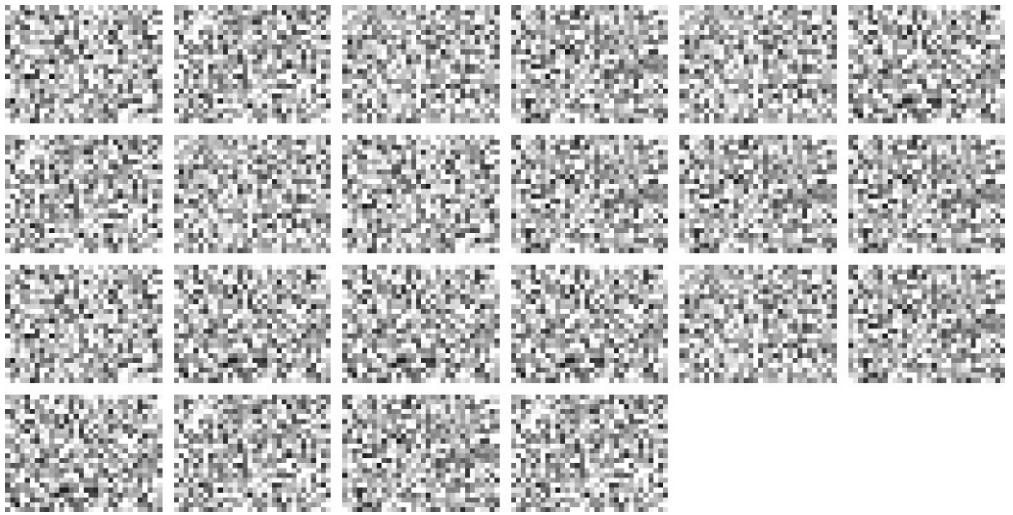
1024x768

Sorted by resolution and directory

Sorting Images by Source

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

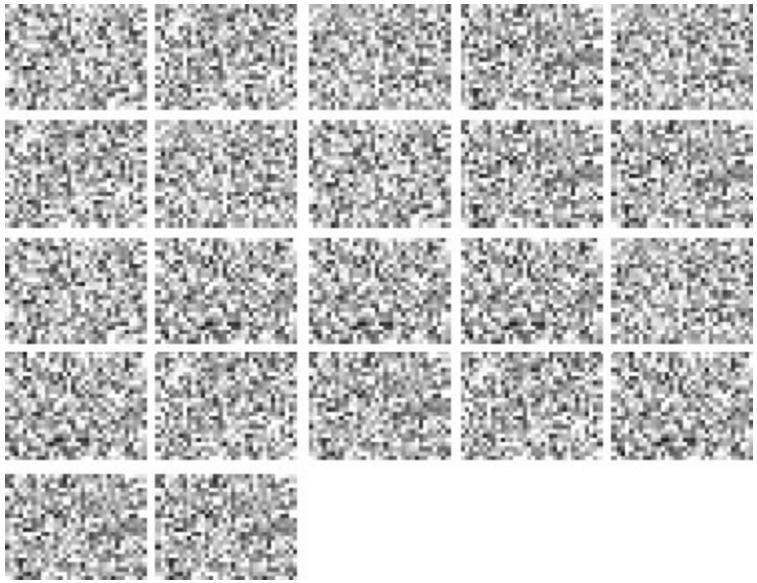
Extract



PRNU noise patterns (fingerprints)

Sorting Images by Source

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



Compare

Images compared to all images

Sorting Images by Source

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

threshold = 0.001

Cluster

Images grouped by source

Sorting Images by Source also GPU / social networks also deep learning applied

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

GP

Facial comparison

NO

YES

NIST test of faces in the wild

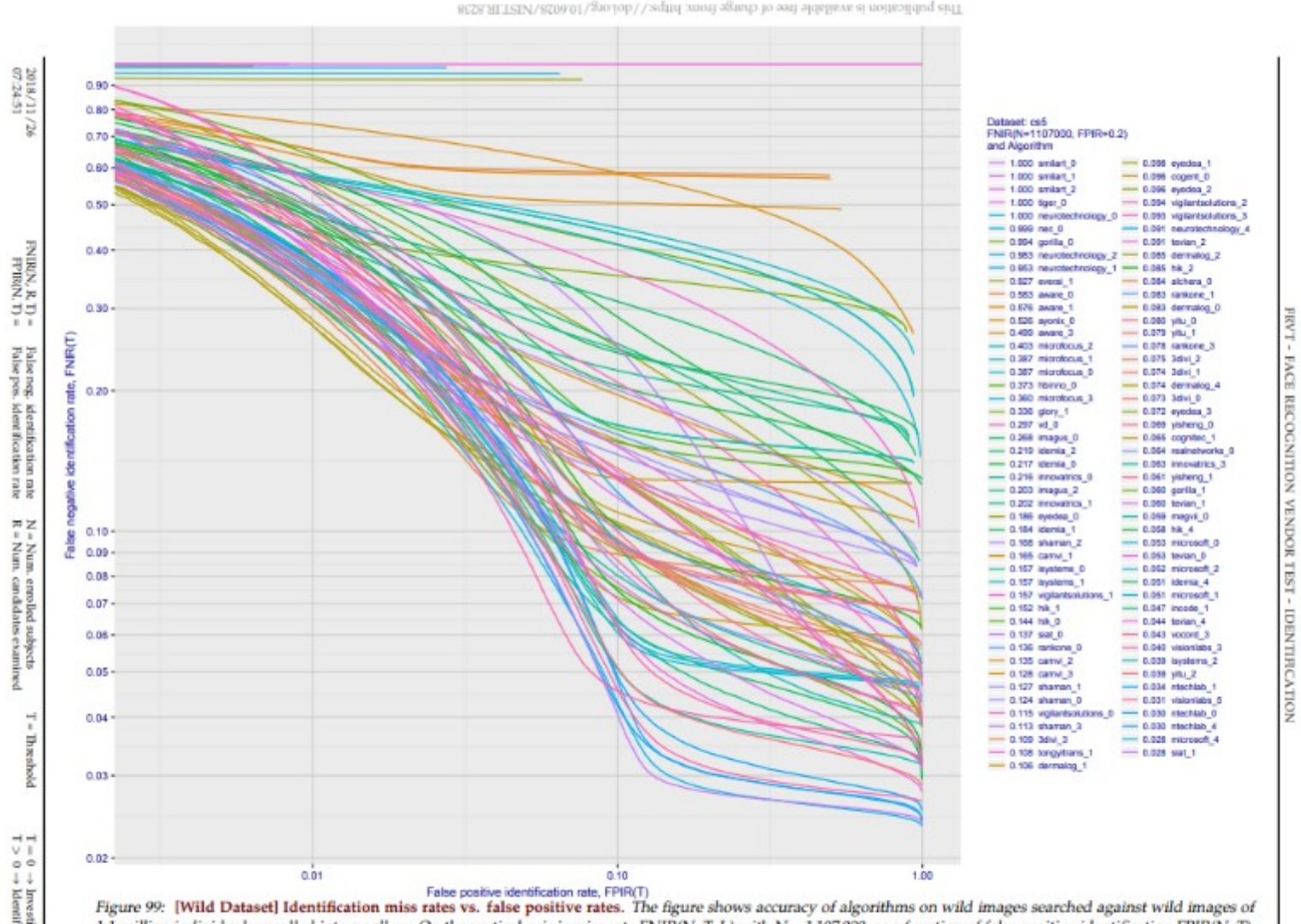


Figure 99: **[Wild Dataset] Identification miss rates vs. false positive rates.** *The figure shows accuracy of algorithms on wild images searched against wild images of*

Other examples of deep learning

- manipulation detection
- face morphing / deepfakes
- court findings finding irregularities

ELSEVIER

Science & Justice

Volume 58, Issue 6, November 2018, Pages 433-440

“I couldn't find it your honour, it mustn't be there!” – Tool errors, tool limitations and user error in digital forensics

Graeme Horsman

Show more

https://doi.org/10.1016/j.scijus.2018.04.001

Highlights

- An examination of tool errors, tool limitations and user error in digital forensics.
- Digital forensic end user license agreements are discussed and evaluated.
- Suggestions for improving testing and defining tool limitations are made.

Discussion

Rafferty said: “Cost-cutting and outsourcing has put the administration of justice at risk … I don’t think it’s bad faith by the police. They have been under-resourced. They are swamped. In some of my cases it’s the police who have revealed material that’s helpful to the defence.”

Collie, the head of Discovery Forensics in London who mainly works for defendants, said: “The odds are stacked against the defence in many ways. We rarely get access to the actual piece of equipment. In the past I could go to the police station and see a phone or a computer and physically check it’s the right piece. Now everything comes prepackaged and is handed over on a hard drive or USB stick.”

Collapsed rape prosecutions

December: Liam Allan
The first case to be abandoned due to the failure by police to hand over crucial digital evidence was that of London student Liam Allan, 22, in December. Allan was charged with 12 counts of rape and sexual assault, but his trial was abandoned after police were ordered to hand over phone records that should have already been provided to the defence.

December: Isaac Itiary
Shortly before Christmas, an alleged child rapist, Isaac Itiary, 25, was cleared at Inner London crown court when the prosecution offered no evidence. Material recovered from the phone of the complainant by police was only handed over to defence lawyers shortly before it was due to come to trial.

January: Oliver Mears
In January, Oliver Mears, 19, a student at Oxford University, was

- Explain Deep Learning in court
- Bias in Model
- Training of users
- Anti forensic software

Questions